IMPROVED BALANCED RANDOM SURVIVAL FOREST FOR THE ANALYSIS OF RIGHT CENSORED DATA: APPLICATION IN DETERMINING UNDER FIVE CHILD MORTALITY

BY

WAITITU HELLEN WANJIRU

A thesis submitted in partial fulfillment of the requirements for the award of the degree of Doctor of Philosophy in Biostatistics in the School of Sciences and Aerospace Studies, Department of Mathematics, Physics and Computing of Moi University.

MOI UNIVERSITY

2021

Declaration

Declaration by Candidate

This thesis is my original work and has not been presented to any other examination body. No part of this thesis may be produced without prior permission of the author and/or Moi University.

Signature: ..... Date .....

Waititu Hellen Wanjiru

DPS/PHD/007/16

Declaration by the Supervisors

This thesis has been submitted with our approval as the university supervisors.

Signature: ..... Date .....

Prof. Joseph K. arap Koske

School of Sciences and Aerospace Studies,

Department of Mathematics, Physics & Computing,

Moi University.

Signature: ..... Date .....

Dr. Nelson Owuor

School of Mathematics,

University of Nairobi.

Dedication

This thesis is dedicated to my beloved husband Reuben and our children Violet, Victor and Vivian for their love, understanding, encouragement and tolerance during the long and tedious journey of my research.

Acknowledgments

Praises and thanks to God, the Almighty, for His showers of blessings throughout my research work. Indeed it has taken His Mighty hand to complete the research successfully.

I wish to thank all the people whose assistance was a milestone in the completion of this work. Specifically, I would like to extend my gratitudes to the following people.

My thanks goes to Prof. Joseph Koske for his professional guidance, organizational skills, encouragement and assistance throughout the study period.

I express my deepest gratitude to Dr. Nelson Owuor. He has taught me the methodology to carry out the research. His guidance, corrections and encouragement cannot be underestimated. He introduced me the area of machine learning, working with latex and running the R codes on my own. This has made me proud of the work that I have been able to accomplish. It was a great privilege and honor to work and study under his guidance. I am extremely grateful for what he has offered me. Without his persistent help, the goal of this research would not have been realized.

To my dear husband Dr. Reuben, it is whole-heartedly appreciated that your great advice, support and assistance for my study proved monumental towards the success of this study. I enjoyed your instant and continuous support and skills especially when working with latex and running the R codes. Your motivation to do this study and immense support cannot go unnoticed.

I wish to acknowledge the support and great love of my family; my husband, my children; Violet, Victor and Vivian for their acceptance, love, patience, understanding, prayers and continuing support to complete this research work. They kept me going on and this work would not have been possible without their input. I also express my thanks to my sisters and brothers for their encouragement, support and valuable prayers.

Finally, my thanks go to my colleagues, friends and all the people who have supported me to complete the research work directly or indirectly.

To all, may God bless you.

Abstract

The desire to understand the determinants of Under Five Child Mortality ($U5CM$) poses a very important aspect of research. One of the main challenges affecting the Low and Middle Income Countries ($LMIC$) is the aspect of child mortality. The Sustainable Development Goals target of at most 25 deaths per 1000 live births has not been met, despite the many interventions governments have put in place to avert child mortality. There is huge need to understand the determinants of child mortality, especially the U5CM. Most studies rely on household surveys such as the Kenya Demographic and Health Survey ($KDHS$) data, with $KDHS - 2014$ being the most recent household survey in Kenya. Some of the statistical challenges that come with DHS datasets are the presence of high imbalance in comparison classes, high dimensional problem, statistical selection of variables, and distributional assumptions among other factors. Random Survival Forests (RSF) have recently become a popular method for survival data analysis. However, statistical challenges such as imbalance between mortality and non mortality class and violation of Proportional Hazard (PH) assumption pose significant challenge(s) to RSF. This is due to its stopping criterion based on daughter node constraint which demonstrates bias towards predictors in a large population and use of log-lank splitting rule whose optimality is achieved when PH assumptions are satisfied. The main aim of this study was to develop a machine learning algorithm to handle the above mentioned statistical challenges that come with high dimensional survey data in identifying the determinants of U5CM. The specific objectives were: To analyze Balanced Random Survival Forests ($BRSF$) using specified balancing techniques; to analyze $BRSF$ using specified splitting rules; to develop an Improved Balanced Random Survival Forests ($IBRSF$) model and finally to apply the $BRSF$ to determine the U5CM. The study methodology involved data balancing using four specified external data balancing techniques: Random Under-sampling, Random Over-sampling, Both-sampling, and Synthetic Minority Oversampling technique. The balanced data

was integrated with RSF for variable selection and model selection done using concordance index to identify the model with the best balancing technique. The $BRSF$ was then analyzed using three specified splitting rules: log-rank, log-rank score and Bs.gradient splitting rules. Finally, an $IBRSF$ algorithm was developed by integrating balanced data with RSF while using optimal splitting rule. The study found that the model with random under-sampling balancing method produced the best fit with a concordance index of 0.90. The model using Bs.gradient splitting rule recorded a concordance of 0.87, and was the most optimal method when PH assumptions were violated. The final model, the $IBRSF$ model, integrated data balancing using random under-sampling method and Bs.gradient rule in splitting the nodes. Based on this model, $B7$ (age at death of the child) resulted as the highest determinant of $U5CM$ with the largest variable importance ($VIMP$) value of 0.0472. In conclusion, $IBRSF$ produced a good fit to the data and enabled data analysis that solved all the specified statistical challenges that come with $KDHS$ type of data. The study recommends the use of $IBRSF$ model for prediction of highly imbalanced right censored data in situations where PH assumption is violated.

Contents

List of Publications

1. Hellen Wanjiru Waititu, Joseph K. Arap Koskei, Nelson Owuor Onyango, Determinants of Under Five Child Mortality from KDHS Data: A Balanced Random Survival Forests (BRSF) Technique, International Journal of Statistics and Applications, Vol. 10 No. 5, 2020, pp. 118-130.
   doi: 10.5923/j.statistics.20201005.02.

2. Hellen Wanjiru Waititu, Joseph K. Arap Koske, Nelson Owuor Onyango, Analysis of Balanced Random Survival Forest Using Different Splitting Rules: Application on Child Mortality, International Journal of Statistics and Applications, Vol. 11 No. 2, 2021, pp. 37-49. doi: 10.5923/j.statistics.20211102.03.

DEFINITION OF TERMS

- **Algorithm** - An Algorithm is a sequence of instructions or a set of rules to be followed in solving a problem especially by a computer. In mathematics and computer science, an algorithm usually means a small procedure that solves a recurrent problem.

- **Classifier**- A classifier in machine learning is an algorithm that automatically orders or categorizes data into one or more of a set of classes. One of the most common examples is an email classifier that scans emails to filter them by class label: Spam or Not Spam. A classifier is the algorithm itself the rules used by machines to classify data.

- **Data mining** - Data mining is the process of finding anomalies, patterns and correlations within large data sets to predict outcomes. It implies analyzing data patterns in large batches of data using one or more software.

- **Ensemble** - is a data mining technique composed of a number of individual classifiers to classify the data to generate new instances of data. The goal of ensemble methods is to combine the predictions of several base estimators built with a given learning algorithm in order to improve robustness over a single estimator.

- **Ensemble learning** - is the process by which multiple models, such as classifiers or experts, are strategically generated and combined to solve a particular computational intelligence.

- **Machine learning** - Machine learning is a branch of artificial intelligence based on the idea that systems can learn from data, identify patterns and make decisions with minimal human intervention. It is a method of data analysis that automates analytical model building by identifying patterns in data. It is especially useful for diverse, high-dimensional data. Machine Learning algorithms learn from data. They find relationships, develop understanding, make decisions, and evaluate their confidence from the training data they are given. While it generally delivers faster, more accurate results in order to iden-

tify profitable opportunities or dangerous risks, it may also require additional time and resources to train it properly.

LIST OF ABBREVIATIONS AND ACRONYMS

- BRF- Balanced Random Forests
- BRSF- Balanced Random Survival Forests
- CIF- Conditional Inference Forests
- C-Index Concordance index
- CPH- Cox Proportional Hazard
- CHF- Cumulative Hazard Function
- C-Statistics - Concordance Statistics
- DHS- Demographic and Health Surveys
- IBRF- Improved Balanced Random Forests
- iRSF- Improved Random Survival Forests
- KDHS- Kenya Demographic and Health Surveys
- LMIC- Low and Middle Income Countries
- MDG- Millennium Development Goals
- OOB-data- Out Of Bag data
- PH- Proportional hazard
- RF- Random Forests
- RSF- Random Survival Forests
- SMOTE- Synthetic Minority Oversampling Technique
- U5CM- Under Five Child Mortality
- VIMP- Variable Importance
- WRF- Weighted Random Forests

List of Figures

List of Tables

Chapter 1

INTRODUCTION

1.1 Background

One of the main challenges affecting the Low and Middle Income Countries ($LMIC$) is the aspect of child mortality. The Sustainable Development Goals target of at most 25 deaths per 1000 live births has not been met, despite the many interventions governments have put in place to avert child mortality. There is huge need to understand the determinants of child mortality, especially the Under Five Child Mortality ($U5CM$). The desire to understand the determinants of $U5CM$ therefore poses a very important aspect of research, as countries aim to achieve the Sustainable Development Goals ($MDG2015-2030$).

Demographic and Health Surveys (DHS) program has been very instrumental for acquiring and distributing authentic, nationally representative data on fertility, family planning, maternal and child health, among other health issues. They are known to give very important source of information on the health of children and women in low- and middle-income countries and are appropriate for studies of health. The most recent DHS survey conducted in Kenya was Kenya Demographic and Health Surveys ($KDHS$) 2014. Some of the statistical challenges that come with

DHS datasets are the presence of high imbalance in comparison classes, missing data problem, high dimensional problem, statistical selection of variables, distributional assumptions among other factors.

This study aims to identify the determinants of $U5CM$ in Kenya. Comparisons shall be made between mortality and non-mortality groups from the 2014 $KDHS$ study. Mortality group which is of most interest in our study composes a very small minority class (less than 7% of the entire population), while the non-mortalities constitute the majority class. From the KDHS 2014 data, only 4.2% of children experienced under five years mortality while the rest 95.8% survived until after the fifth birthday.

Imbalanced classification is a common problem with most datasets including mortality data, fraud detection among others. High imbalance has been observed to suppress the effectiveness of machine learning algorithms such as Random forests (RF). When dataset is imbalanced and one class dominates the other, such machine learning algorithms have issues classifying correctly. The classifiers become inclined towards the majority class leading to poor representation of the minority class.

The 2014 $KDHS$ data is also associated with $1,099$ variables and $20,964$ rows of data. Due to high dimensionality of the data, we need to identify effective variable selection techniques in order to identify determinants of child mortality. Machine learning techniques (that require no distributional assumptions on data) such as Random Survival Forests and support vector machine, among others, have received wide application in studies involving high dimensional datasets. These machine learning techniques are useful when dealing with problems such as missing data imputation, classification imbalance and variable selection.

Besides many variables in the DHS data, there are missing observations. In this study however, we aimed at handling the challenge of imbalanced classification in mortality data and the best way of dealing with dataset with variables that violate PH assumptions. Our interest was to identify and use a model that takes into

account these challenges during the process of data analysis.

1.1.1 Survival Analysis Regression

Survival analysis is a collection of statistical procedures for data analysis for which the outcome variable of interest is time until an event occurs Kleinbaum and Klein (2005). It involves modeling of time to event data. While carrying out survival analysis, we may not have exact event times for all observations. The observations which experience an event yield complete observation while those that do not experience an event within the follow up period give rise to incomplete observations resulting to censored observations.

Censoring happens in situations where individuals under observation have not experienced an event by the end of the study period but some information about individual's survival is known. The fact that the event under consideration did not occur when the individual was under observation makes the survival time information incomplete. Unlike ordinary regression methods, survival methods correctly incorporate information from both uncensored and censored observations in estimating important model parameters. The observations that had not yet failed has a big impact in survival estimates. The important difference between survival analysis and other statistical analysis (eg the ordinary regression models) is the presence of censoring. This makes survival analysis important in writing down models.

There are generally three types of censoring. These are right censoring, left censoring and interval censoring. In this study, we dealt with right censoring which occurs when the study terminates before the event has occurred or when a subject leaves the study before the occurrence of an event.

Survival models can be analyzed using parametric, non-parametric and semi-parametric approaches. Parametric methods assume a survival distribution which affects the shape of the model's hazard function. On the other hand, non-parametric

methods analyze survival data without parametric assumptions about the form of distribution. In Semi-parametric methods, the functional form of covariates is parametric while the hazard function is estimated non-parametrically.

One of the popular semi-parametric model for survival data analysis is Cox proportional hazard (Cox PH) model Cox (1972). The model estimates survival curves when considering several explanatory variables simultaneously. Popularity of Cox PH model rises from its semi-parametric nature which does not require any distributional assumption of the baseline hazard function to estimate the regression coefficients. The fact that baseline hazard, is an unspecified function makes the Cox model a semi-parametric model. With Cox PH model, even though the baseline hazard is not specified, reasonably good estimates of regression coefficients, hazard ratios of interest, and adjusted survival curves can be obtained for a wide variety of data situations Kleinbaum and Klein (2005). However Cox model makes certain restrictive assumptions many of which do not hold in real life scenarios. One such assumption is a constant hazard ratio between any two observations at any time instant t. In addition, Cox PH model does not take into account the missing predictors, non-linearity of exponential factors and interdependence among observations. It is also known to have an inherent bias and high generalization error Pan (1998).

Recent studies has seen the development of non-parametric survival analysis techniques which deals with the challenges of the cox PH model. Non parametric techniques do not require the data to meet certain assumptions of parameters.

1.1.2 Random Survival Forests

Random Survival Forests (RSF) is an ensemble tree method for the analysis of right censored survival data Ishwaran et al. (2008). It was proposed as an extension of random forests (RF) for non-parametric survival analysis.

To grow a forest using the RSF procedure, $ntree$ bootstrap samples are randomly

selected from the initial data with G samples. From each bootstrap sample, a mean of 37% of the data is set aside. This is known as out-of-bag (OOB) data. A survival tree is then grown from each of the bootstrap samples. At each node of the tree, *mtry* predictors are randomly selected for node splitting. The node is split using the candidate variable that maximizes survival difference between daughter nodes. Then each tree is grown to full size under the constraint that the most extreme node should have no less than *nodesize* unique deaths. A Cumulative hazard function (CHF) for each tree is calculated and then averaged to obtain the ensemble CHF. Prediction error for the ensemble CHF is calculated using the OOB data.

In RSF each of the different trees in the forest gives a prediction. Further, the mean of the predictions from each tree gives the final prediction for the forest. All aspects of growing a forest in right censored survival data take into account the outcome which is survival time and censoring status Breiman (2003). The aspects of growing a tree includes the splitting criteria used in growing a tree, tree node impurity measuring effectiveness of a split in separating data, resulting ensemble predicted value from the forest and the measure of prediction accuracy.

Averaging over trees, and two way randomization while growing a tree, enables RSF to approximate complex survival functions while maintaining low prediction error Ishwaran et al. (2008). The two way randomization improves both bias and variance.

RSF have recently become a popular method for survival data analysis due to its ability to adaptively discovers nonlinear effects, automatically detect certain types of interactions without specifying them beforehand and effectively imputing missing data. However characteristic of survival data such as extreme imbalance between mortality and non mortality classes pose significant challenges to RSF. The presence of extreme imbalance between the censored and mortality class with as low as $2-10\%$ data in the minority class is commonly occurring in survival data Afrin et al. (2018). In addition, the stopping criterion in RSF is arbitrary and

demonstrates bias towards predictors with a large population Miao et al. (2018). This is because it is difficult for predictors with a smaller population to satisfy the criterion especially when *nodesize* is large.

Recently, improvement of RSF have emerged to further improve prediction. These includes the Improved Random Survival Forests ($iRSF$) Miao et al. (2018) which was proposed to improve RSF with a new split rule and stopping criterion and Balanced Random Survival Forests (BRSF) Afrin et al. (2018) which integrates synthetic minority oversampling technique with RSF.

In this research, we developed an Improved Balanced Random Survival Forest ($IBRSF$) algorithm for right censored data. This followed a unified three stage procedure which involved data balancing in the first stage. The balanced data was subjected to variable selection using RSF in the second stage and the selected variables were used for determination of U5CM in the last stage.

1.2 Statement of the Problem

One of the main challenges affecting the Low and Middle Income Countries (LMIC) is the aspect of child mortality. The Sustainable Development Goals targets of at most 25 deaths per 1000 live births has not been met, despite the many interventions governments have put in place to avert child mortality. There is huge need to understand the determinants of child mortality, especially the Under Five Child Mortality (U5CM). Most studies rely on household surveys such as the Kenya Demographic and Health Survey (KDHS) data, with KDHS-2014 being the most recent household survey in Kenya. Some of the statistical challenges that come with DHS datasets include the presence of high imbalance in comparison classes, missing data problem, high dimensional problem, statistical selection of variables, and distributional assumptions among other factors. For instance, in the 2014 KDHS data, the mortality

group which composes of minority class, constitutes the minority group (less than 7% of the entire population) while the non-mortalities constitute the majority class. Furthermore, the 2014 $KDHS$ data is associated with 1,099 variables and 20,964 rows of data. Some of these variables violated the proportional hazards assumption. Random Survival Forests (RSF) have recently become a popular method for survival data analysis. However, characteristics of survival data such as imbalance between the survival and mortality class sizes pose significant challenge(s) to RSF. This is due to its stopping criterion based on daughter node constraint which demonstrates bias towards predictors in a large population. In addition, RSF mainly uses log-lank test as the split rule that maximizes the survival difference between daughter nodes. Proportional hazard assumption is the key requirement for the optimality of log-rank test. However, there are many situations where the proportional hazard assumption is violated making the log-rank test insufficient. The main problem of this study was to develop a machine learning algorithm (in this case IBRSF) to handle the above mentioned statistical challenges that come with high dimensional survey data in identifying the determinants of U5CM.

1.3 Objectives

1.3.1 General Objective

Develop an Improved Balanced Random Survival Forests (IBRSF) model for the analysis of right censored data and use the model to identify the determinants of U5CM.

1.3.2 Specific Objectives

1. Analyse Balanced Random Survival Forests ($BRSF$) model under different balancing techniques to select the best balancing technique.

2. Analyse Balanced Random Survival Forests ($BRSF$) model under different splitting rules to select an optimum splitting rule.

3. Develop an Improved Balanced Random Survival Forests ($IBRSF$) model for right censored data.

4. Apply the Improved Balanced Random Survival Forests ($IBRSF$) model to identify the determinants of U5CM.

1.4 Justification of the Problem

The desire to have a model that can accurately identify predictors of mortality cannot be underestimated. The main problem of this study was to develop a machine learning algorithm to handle the statistical challenges that come with high dimensional survey data which include high imbalance between mortality and non mortality class, high dimensionality, and violation of PH assumption. Random Survival Forests (RSF) have recently become a popular method for survival data analysis. However, characteristics of survival data such as imbalance between the survival and mortality class sizes pose significant challenge(s) to RSF. This is due to its stopping criterion based on daughter node constraint which demonstrates bias towards predictors in a large population. In addition, RSF mainly uses log-lank test as the split rule that maximizes the survival difference between daughter nodes. Proportional hazard assumption is the key requirement for the optimality of log-rank test. However, there are many situations where the proportional hazard assumption is violated making the log-rank test insufficient.

This research has led to development of a new model; Improved Balanced Random Survival Forests ($IBRSF$) model for analysis of right censored data. This is an improvement to RSF model enhancing accuracy when the data is highly imbalanced and PH assumption is violated. This study is important in analysis of highly imbalanced right censored survival data in which PH assumption is violated. The model is able to address challenges with data imbalance, violation of PH assumption, and high dimensionality of data in identifying the determinants of U5CM. This has assisted in accurate identification of determinants of U5CM using the 2014 KDHS dataset. The research further helps in guiding clinical decisions geared towards reduction of mortality. The study also acts as a base for more research on other balancing techniques, splitting rules and censoring methods that can be integrated in IBRSF.

1.5 Scope of the Study

This research aims to identify the determinants of U5CM. This was done using a unified model which involved data balancing, variable selection and variable prediction. In data balancing stage, this study was limited to the external data balancing techniques only. Other data balancing techniques were not analyzed in this research. Variable selection was carried out using RSF technique. At this stage, different splitting rules were analyzed. We limited ourselves to analysis of log-rank, log-rank score and Bs.gradient splitting rule. In variable prediction stage, different methods were used which include Cox PH model, Cox Aalen's method and RSF variable importance (VIMP). The Kenya Demographic Health Survey (KDHS) 2014 data was used in this research to identify the determinants of U5CM. This dataset was found to be classified in regions which represents the former provinces in Kenya. In this research, we only analyzed the Nairobi region dataset. This dataset consisted of 532 observations and 757 variables after data cleaning.

Chapter 2

LITERATURE REVIEW

2.1 Introduction

This Chapter focuses on the review of most recent, relevant and significant studies related to our research. It illustrates what other researchers have done with respect to the gap in their area of study, the objectives achieved and the methodology used. The different areas covered in this chapter include U5CM, data balancing, improvements to RSF and splitting methods.

2.2 Under Five Child Mortality

A number of studies have explored determinants of child mortality using DHS data. Ayiko et al. (2009) used Uganda 1996, 2000, 2006 DHS dataset to assess the trends and levels of childhood mortality between 1990 and 2006 as well as the determinants of under-five mortality. The authors used a Cox PH regression model to explore region of residence, sex of the child, type of birth (multiple), birth interval (less than 24 months after the preceding birth), and mother's education in relation to an

increased risk of children mortality before their fifth birthday.

Ettarh and Kimani (2012) used the 2008 KDHS data to determine the U5CM in rural and urban areas . They used Cox PH regression to explore the effect of maternal, demographic and geographical factors on mortality. According to the findings of their study, the broad possibility of death in rural areas was significantly higher than that of urban areas with household poverty and influence of breast feeding being the highest risk factors for mortality.

Sreeramareddy et al. (2013) analyzed the data from complete birth histories of four Nepal Demographic and Health Surveys (NDHS) done in the years 1996, 2001, 2006 and 2011. In their study, they explored the effect of mothers education, child's sex, rural/urban residence, household wealth index, regions ecological zones and development.

Nasejje et al. (2015) used Uganda 2011 DHS. Using Cox-proportional hazard model, factors related to mother characteristics and previous births such as sex of the child, sex of the head of the household and the number of births in the past one year was found to be significant.

In this study, we have also tapped into the richness of KDHS (2014) dataset, to establish the determinants of U5CM. All the variables after data cleaning were viewed as potential determinants of U5CM.

2.3 Imbalanced Classification

Imbalanced classification is a common problem with most datasets including mortality data, fraud data, fraud detection, claim prediction, default prediction, spam detection among others. Handling imbalanced classification has received prominence in many studies including (Lessmann (2004); Tang et al. (2008); Lpez et al. (2012); Yan et al. (2019); Lin et al. (2020)). This is due to some of the challenges affecting survival analysis when data is highly imbalanced.

Krawczyk (2016) explores open challenges and future directions in learning from imbalanced datasets. The motivation for this study stems from the challenges facing imbalanced learning with the advent of big data in addition to the expansion of machine learning and data mining. The work discusses various forms of learning in which there was an issue with data imbalance. Further open challenges as well as those evolving from learning imbalanced datasets in real world applications are also highlighted.

Fernndez et al. (2018) Addressed the issues of learning from imbalanced datasets with the aim of offering general and comprehensible overview in the area of imbalance. In their work, they stressed on the challenges with standard classification tasks, introduced the main evaluation metrics to be considered in learning with imbalanced datasets, covered different approaches that have been traditionally applied to address imbalance as well as ensemble learning solutions.

2.4 Dealing with Imbalanced Data in Random Forests

When dealing with imbalanced data in Random Forests (RF), there is a high chance for a bootstrap sample to have few or no minority cases. This leads to a tree with poor prediction performance of the minority class. RF was intended to minimize the overall error rate. This makes it focus more on prediction accuracy of the majority class leading to poor accuracy of the minority class.

Chen et al. (2004) proposed two different ways of dealing with imbalanced data in RF. These are Balanced Random Forests (BRF) based on sampling technique and Weighted Random Forests (WRF) based on cost sensitive learning. In BRF, a bootstrap sample is drawn from the minority class and the same number of cases drawn with replacement from the majority class. A tree is then grown from the

resulting balanced dataset. In WRF, weights are assigned to each class with a higher weight given to the minority class since classification algorithms tend to be biased towards the majority class. The weights are incorporated during the tree growing procedure to weight the splitting criterion used and also in the terminal node. According to Chen et al. (2004) none of the two strategies can be regarded as being dominant over the other BRF is computationally more efficient with large imbalanced data and more noise tolerant while WRF has more effect on classifiers produced by decision tree learning method.

By combining BRF and WRF, Yaya et al. (2009) came up with Improved Balanced Random Forests ($IBRF$) and demonstrated its application to prediction of customers tendency in a given period to stop doing business with a given company. They were responding to the challenges brought about by learning from imbalanced datasets. Combination of the two methods was done by use of interval variables. Given the random variables d, the length of the interval, m, the midpoint of the interval together with the training data set $D = (X_1Y_1), ..., (X_nY_n)$ a distribution variable α is randomly generated within the interval between $m - \frac{d}{2}$ and $m + \frac{d}{2}$. $n\alpha$ samples are drawn with replacement from the majority class (negative training dataset D^- and $n(1 - \alpha)$ samples from minority class (positive training data set D^+). Weight w_1 is assigned to the negative class and w_2 to the positive class. Here, $w_1 = (1 - \alpha)$ and $w_2 = \alpha$. A classification tree is then grown to its full size. By introducing "interval variable", these two approaches alter the class distribution and put heavier penalties on misclassification of the minority class.

2.5 Random Survival Forests and Improvements to Random Survival Forests

Machine learning techniques (that require no distributional assumptions on data) such as Random Survival Forests (RSF), support vector machine, among others, have received wide application in studies involving high dimensional datasets (Nasejje et al. (2017), Sreeramareddy et al. (2013); Cassy et al. (2019), Liu (2019)). These machine learning techniques are useful when dealing with problems such as classification imbalance and variable selection.

The applications of Random Forests (RF) focused primarily on classification and regression problems and not survival analysis. In 2008, Ishwaran et al. (2008) extended RF to RSF. They introduced random survival forests, an ensemble tree method for analysis of right censored survival data. This was an extension of Breiman (2001a) RF method which focused on classification and regression problems. Their research was intended to give a solution to the challenges which faced analysis of survival data. These includes; analyzing survival data using methods that rely on restrictive assumptions like the PH assumption, methods that could not deal with non linear effects of the variables and methods that could not identify interactions. These challenges were handled automatically using forests. In their work, they gave a detailed description of RSF, illustrated several of its important features and investigated the use of variable importance for variable selection. In addition, they introduced new survival splitting rules for splitting survival trees, new missing data algorithm for imputing missing data and a measure of mortality that is simple and interpretable.

Various researchers have compared the performance of Cox model and RSF. Nasejje et al. (2017) in their research compared RSF model with Conditional Inference Forests (CIF). In their research, CIF were found to perform comparably similar to RSF model in data consisting of covariates with fewer split-points.

Recent works using RSF for survival data have shown improved results as compared to the cox PH models and are getting popular as a survival analysis tool. Nonetheless, the characteristics of the survival data pose significant challenges to RSF. Miao et al. (2015) found that RSF was weak at identifying predictors in relatively small populations. This is due to its stopping criterion based on daughter node size constraint that the terminal node should have no fewer than *nodesize* unique deaths. This demonstrates bias toward predictors with a larger population since it is difficult for predictors with a smaller population to satisfy the criterion, especially when *nodesize* is large.

Miao et al. (2018) developed a risk model for prediction of heart mortality using improved random survival forest ($iRSF$) with a new split rule and stopping criterion. According to their findings, the model was able to identify more accurate predictors which could separate survivors from non survivors in small populations improving discrimination ability. Weighted log-rank test was used to split the node where the adaptive weights were obtained using the model of Yang and Prentice (2005). They used split function decreasing as the stopping criterion in this model.

In 2018, Afrin et al. (2018) researched on BRSF for right censored data in extremely imbalanced situations. Their concern was on the limited accuracy of survival models due to sparcity of samples and extreme imbalance. In as much as RSF model has various strengths including overcoming proportion hazard assumption, imbalance in the dataset leads to underestimation of mortality class. In their research, a BRSF model was developed by integrating data balancing using SMOTE with RSF to address this gap. In addition, they gave theoretical results on the effects of balancing on prediction accuracy and conducted studies on different levels of imbalance. According to their findings, BRSF provided an improved discriminatory strength between survival and mortality classes. In comparison, BRSF was found to outperform RSF and optimized cox with and without balancing.

Fiorentini and Losa (2020) noted a significant trend in neglecting the aspect of

imbalance while using machine learning algorithms in predicting crush severity in acutely imbalanced datasets. They researched on handling imbalanced accidents datasets in order to provide a better prediction of the minority class while using machine learning algorithms. They balanced the data using the random under sampling majority class (RUMC) balancing scheme. A comparison of four different crash severity predictive models which included the random tree, k-nearest neighbor, logistic regression and random forest was done. After the assessment using accuracy, precision, confusion matrix , RUMC based model was found to be reliable and significantly more effective in recognizing the minority class. They stressed on the importance of using RUMC and machine learning algorithms in prediction of severity of a crash occurrence.

In their research on prediction of survival in patients with non small cell lung cancer, He et al. (2020) used the relative importance approach in survival prediction. They compared the Variable importance (VIMP) RSF model with Cox model where VIMP was found to be more robust. In their work, they suggested the use of RSF VIMP alongside with Cox model in order to advance the understanding of the roles of prognostic factors and improve on their precision and care efficiency.

Ishwaran and Lu (2019) proposed a sub-sampling approach to be used for estimation of variance of VIMP and construction of confidence intervals. This was motivated by the limitation that no systematic method existed for estimating the variance of VIMP. Using simulations, they demonstrated the effectiveness of the delete-d jackknife variance estimator under low sub sampling rates. They also described a general procedure for estimating the variance of VIMP and showed how to construct confidence intervals for VIMP using the estimated variance.

Morvan et al. (2020) proposed a model for prediction of progression-free survival (PFS). The model consisted of two stages both of which involved RSF and VIMP. In the first stage, feature selection was carried out to identify relevant variables while prognosis prediction using the selected features was done in the second stage. A

comparison between the model and conventional methods such as Lasso Cox and gradient-boosting Cox was done where evaluation using C-index showed a better performance for VIMP+RSF model. Vimp was found to select more stable variables and better results in identification of clinically relevant biomarkers in comparison with minimal depth and variable hunting.

This research addressed the problem of imbalance using the 2014 KDHS dataset in determination of U5CM in situations where PH assumption is violated. This was motivated by the desire to search for a model that leads to improved accuracy in predicting the determinants of mortality which will help in guiding clinical decisions geared towards reduction of mortality. Three challenges were addressed in this study. One problem involved trying to balance the dataset classes before making comparisons between mortality and non mortality cases. The other challenge was due to variable selection. One needs to conduct a proper variable selection exercise in order to identify the correct set of variables to use for the regression analysis. The third challenge is the use of a splitting rule that takes into account variables that violate PH assumption. To address these challenges we came up with an Improved Balanced Random Survival Forest (IBRSF) model for analysis of highly imbalanced right censored data sets in situations where PH assumption is violated. The model involves a 3 stage framework which integrates data balancing technique with RSF for both variable selection and prediction and use of BS.gradient splitting rule in situations where variables violate PH assumption. The model leads to improved accuracy in prediction of mortality.

Chapter 3

MATERIALS AND METHODS

3.1 Introduction

This chapter describes the data used in our research as well as the methods used to achieve the various objectives of the study. In the first section of this chapter, we describe and explore the overall 2014 $KDHS$ data from which a subset used in this research was extracted. The exploration was done to give a general view of the data with our main focus being on data imbalance and satisfaction of proportional hazard assumption. Thereafter,the methods through which the various objectives were achieved are described. These includes description of the different balancing methods which were later integrated with RSF model with different splitting rules. The methods of achieving the various objectives are given in sections 3.2, 3.3, and 3.4. Methods used to compare prediction accuracy of the different models are also given.

3.1.1 Data Description

The data used in this research was drawn from the 2014 Kenya Demographic and Health Survey (KDHS) data KNBS et al. (2014). This is the sixth Demographic and Health Survey (DHS) conducted in Kenya since 1989. KDHS is a national research undertaking conducted every five years with the intention to collect a wide range of data with a strong focus on indicators of fertility, reproductive health, maternal and child health, mortality, nutrition and self-reported health behaviors among adults Corsi et al. (2012). It is a national representative household sample survey where households are selected at random from the Kenya National Bureau of Statistics (KNBS) sampling frame.

The survey procedures, instruments and sampling methods used in the KDHS 2014 acquired ethical recommendation from the Institutional Review Board of Opinion Research Corporation (ORC) Macro International Incorporated, a health, demographic, market research and consulting company situated in New Jersey, USA. We sought official registration on the DHS website and got permission to use the 2014 KDHS data. The data was downloaded in SPSS format and constituted 1,099 variables and 20,964 observations. Using package foreign, the data was imported to R software version 3.6 for analysis.

Variables with 100% missing observations as wells as those that were similar but had different names were deleted. For instance variable V102 (Region of residence) and variable V024 (Region of residence). In such a case, one of the variables was deleted from the data reducing the number of variables to 786. Survival time and status variables which are important considerations when analyzing survival data were calculated and included in the dataset giving rise to 788 variables. The date variables in the data are given in century month codes (CMC) which implies the number of months since the start of the century. Time variable is given by the number of months from birth to final status (censored or dead). Calculation of

time variable was done by subtracting variable $B3$ (date of birth in CMC) from variable $V008$ (date of interview in CMC), if the child was living at the time of interview ($B5 = 1$). On the other hand, if the child was not alive at the time of interview ($B5 = 0$), survival time was given by age at death of the child in completed months ($B7$). If the age at death is less than a month, it is given a value of 0 months. Variable $B5$ indicates whether the child was alive or dead at the time of the interview with $B5 = 0$ if the child was dead and $B5 = 1$ if the child was alive. To get the status variable, the values of $B5$ were interchanged such that $status = 1$ when an event (death) was experienced and $status = 0$ indicating censored observation. Time from birth to date of interview was considered as follow-up time. Calculation and inclusion of time and status variable was successfully done using R codes and stata do file commands.

3.1.2 Exploratory Data Analysis for the 2014 KDHS Dataset.

Imbalance in the 2014 KDHS Dataset.

The data in use was explored and analyzed using R software. This involved summarizing and visualizing characteristics of the variables within the dataset. The dataset was found to be highly imbalanced with the mortality class having 871 observations, constituting 4.2% of the overall data while the majority class had 20,093 observations constituting 95.8%. This imbalance between survivors and non survivors in the overall data is demonstrated in table 3.1

Table 3.1: Imbalance in KDHS 2014 data.

Status	Total	Percentage
Survivors(Censored cases)	20093	95.8%
Mortality (No. of observed Events)	871	4.2%
Sum total	20964	100%

Exploration of imbalance within some of the covariates in the dataset was also carried out. Different covariates from the dataset which include region, residence, sex

of the child, level of education, wealth index were also found to have high imbalance between survivors and non survivors with the minority class size ranging between 3% and 6.4%. This imbalance is clearly shown in tables 3.2, 3.3,and 3.4. The categories with highest percentage of mortality are Nairobi region, urban residence, and male children as shown in the tables.

Table 3.2: Imbalance in KDHS 2014 data by Region.

Status/Region	Central	Coast	Eastern	Nairobi	N.Eastern	Nyanza	Rift Valley	Western	Total
Censored cases	1356	2531	2906	498	1538	2757	6618	1889	20093
Observed Events	64	119	109	34	56	169	232	88	871
Total	1420	2650	3015	532	1594	2926	6850	1977	20964
Percentage of events	4.5%	4.5%	3.6%	6.4%	3.5%	5.8%	3.4%	4.5%	4.2%

Table 3.3: Imbalance in KDHS 2014 data by Residence.

Status/Residence	Rural	Urban	Total
Censored cases	13561	6532	20093
Observed Events	575	296	871
Total	14136	6828	20964
Percentage of mortality class	4.1%	4.3%	4.2%

Table 3.4: Imbalance in KDHS 2014 data by Sex.

Status/Child sex	Female	Male	Total
Censored cases	9936	10157	20093
Observed Events	395	476	871
Total	10331	10633	20964
Percentage of Events	3.0%	4.5%	4.2%

Exploration of Proportional Hazard Assumption in the 2014 KDHS Dataset.

Proportional hazard assumption in different covariates of the dataset was explored using the Kaplan Meir curves. The curves are shown in figures 3.1, 3.2, 3.3 and 3.4

From the Kaplan Meir curves, there is evidence of violation of the proportional hazards assumption which is shown by crossing of curves in different categories of the given variables.

From the exploration of the 2014 $KDHS$ data, it is evident that the data is highly imbalanced and some of the variables violate PH assumption. We therefore proceed to the methods of achieving the various objectives of the study.

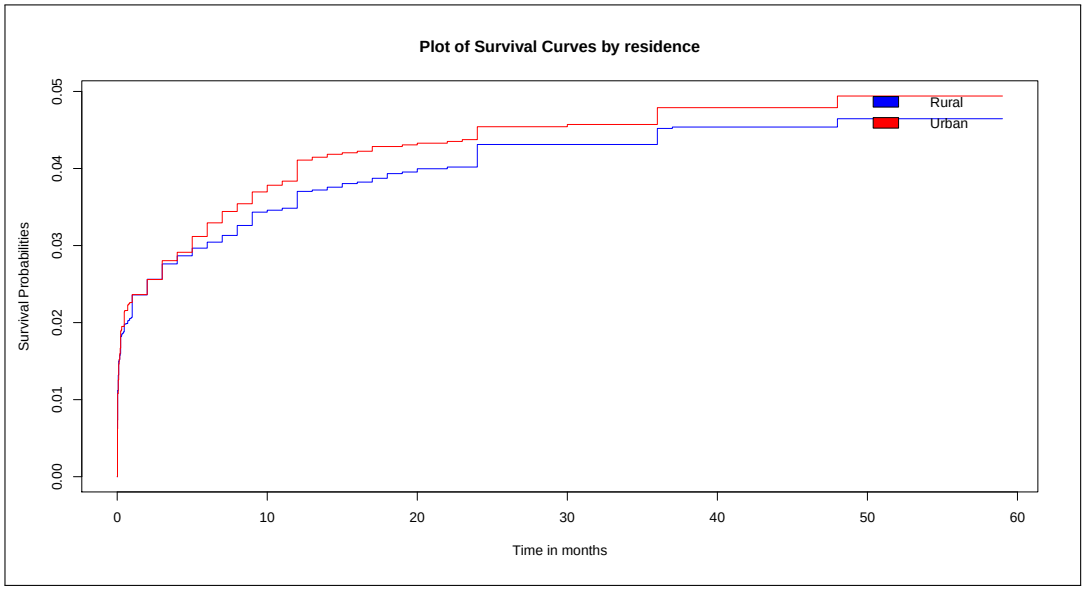

Figure 3.1: Survival Curves by Residence for 2014 KDHS Data.

The graphs shows curves that are crossing and not parallel during the first few months of follow up period which later became parallel towards the end of the follow up period. This is an indication of violation of PH assumption.

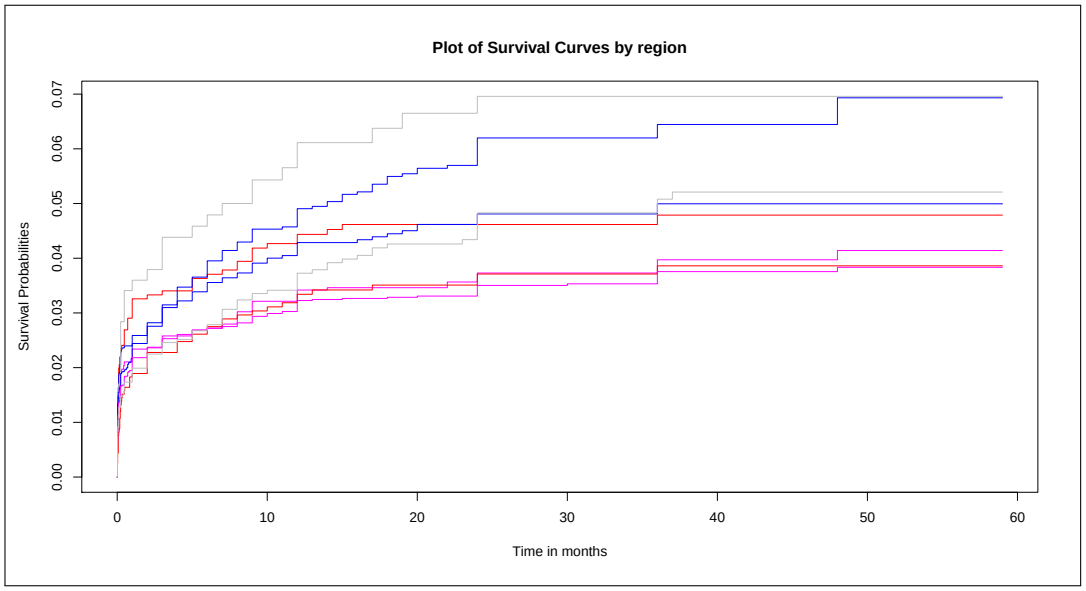

Figure 3.2: Survival curves by region for 2014 KDHS data.

The graphs shows crossing curves indicating violation of proportional hazards assumption.

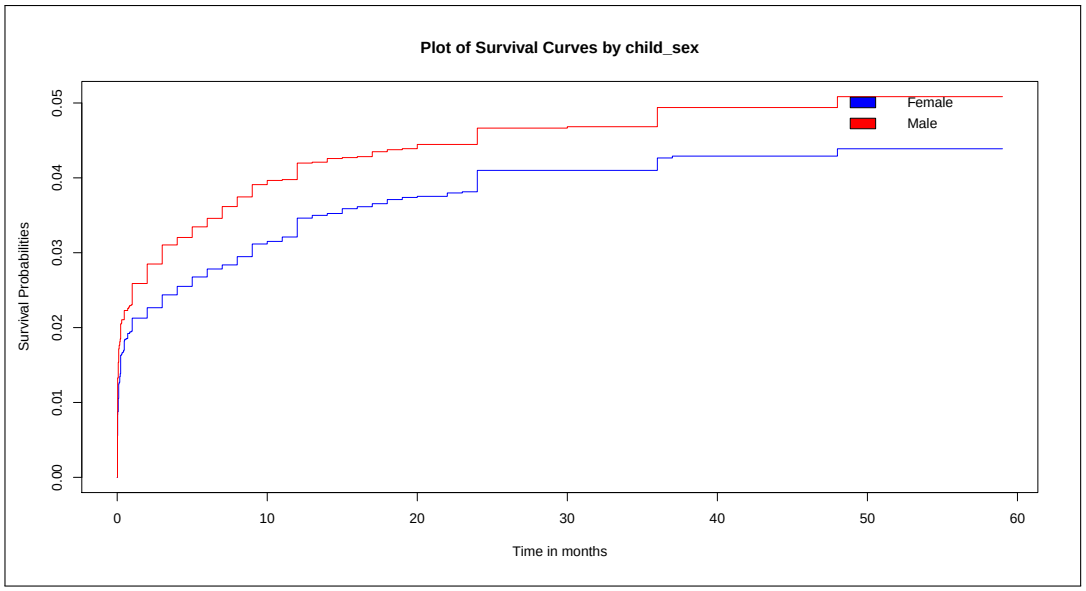

Figure 3.3: Survival curves by child sex for 2014 KDHS data.

The graphs shows curves that are not crossing and parallel after the beginning of follow up period indicating satisfaction of PH assumption.

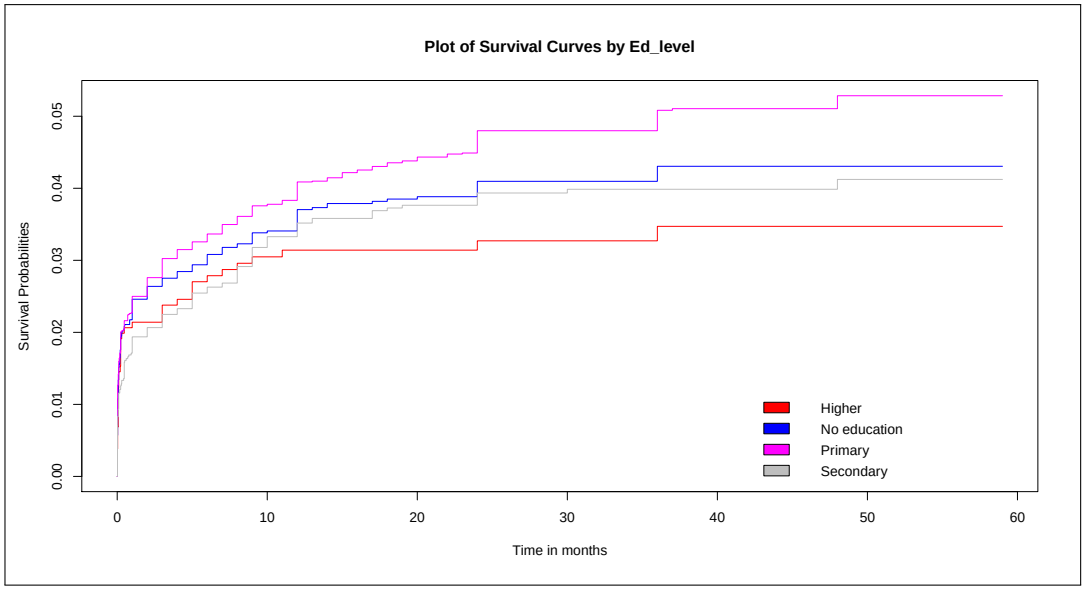

Figure 3.4: Survival curves by education level for 2014 KDHS data.

The graphs shows curves that are crossing during the first few months of the follow up period indicating violation of PH assumption.

3.2 Analysis of Balanced Random Survival Forest ($BRSF$) Model under Different Balancing Schemes.

3.2.1 Exploration of Imbalance in Nairobi Region Dataset.

The aim of this research was to find an effective way of applying the variable selection technique called Random Survival Forest (RSF), to analyze data with imbalance and relatively smaller in size. 2014 $KDHS$ data is a national sample data which is classified into 8 regions, constituting former provinces in Kenya as shown in table 3.2. For this work, we analyzed data for Nairobi region, being a unique urban system in Kenya. Nairobi is a metropolitan region with improved health facilities and access, while also having high levels of socio-economic disparity among populations. In the 2014 KDHS data, Nairobi region alone was associated with 532 observations and 788 covariates. Some variables in this region were found to have 100% missing information and were deleted. Other variables describing the region like V000 (country code), V024 (De facto region of residence), among others were also deleted from Nairobi region dataset. After the process of data cleaning, 757 variables remained.

The data was found to have high level of missing information. This is often one of the main data analysis tasks before running the desired models. In this case, we did multiple imputation using RF algorithms, missForest for missing data imputation Stekhoven and Bhlmann (2012). missForest is a nonparametric machine learning based algorithm for data imputation which apply the Random Forest algorithm. The method starts by imputing all the missing values using the mean imputation method. The missing values are then labeled as predict while the others are labeled as the training data. The data is then fed to Random Forest to predict the missing values and the generated prediction filled to give a new dataset. The process is repeated several times with each iteration giving rise to improved data. This con-

tinues until a stopping criterion is reached after a number of iterations. The method is capable of handling data with different types of variables at the same time, complex interactions, non linear relationships between variables and high dimensionality where the number of variables greatly outnumbers the observations giving good imputation results Stekhoven and Bhlmann (2012). In this study however, we dealt more on handling the challenge of imbalanced classification in mortality data.

Nairobi region dataset was found to be equally highly imbalanced with 6.4% minority class (mortality class) representation. Similarly, the variables in the data (covariates) showed high imbalance in the mortality class. Table 3.5 shows the imbalance between mortality and survivor groups as observed during the 2014 $KDHS$ survey within Nairobi region. The imbalance is also demonstrated in some of the covariates as shown in tables 3.6, 3.7 and 3.8.

Table 3.5: Imbalance in KDHS 2014 Nairobi region data.

Status	Total	Status Percentage
Survivors(Censored cases)	498	93.6%
Mortality (No. of observed Events)	34	6.4%
Sum Total	532	100%

Table 3.6: Imbalance in Nairobi region data by sex.

Status/Child sex	Female	Male	Total
Survivors(Censored cases)	254	244	498
Mortality (No. of observed Events)	17	17	34
Total	271	261	532
Proportion of Events	6.3%	6.5%	6.4%

Table 3.7: Imbalance in Nairobi region data by Education level.

Status/Education Level	Higher	No Education	Primary	Secondary	Total
Survivors(Censored cases)	102	7	203	186	498
Mortality (No. of observed Events)	8	0	13	13	34
Total	110	7	2165	199	532
Proportion of Events	7.3%	0%	6.0%	6.5%	6.4%

Table 3.8: Imbalance in Nairobi region data by wealth index.

Status/wealth index	1 (Poorest)	2 (Poorer)	3 (Middle)	Total
Survivors (Censored cases)	5	36	457	498
Mortality (No. of observed Events)	1	1	32	34
Total	6	37	489	532
Proportion of Events	16.7%	2.7%	6.5%	6.4%

Such imbalance may lead to lack of information and under-representation in the mortality class which is of great interest in our study.

3.2.2 Data Imbalance

A dataset is said to be technically imbalanced if its class distributions are not equal. However, when there is a significant, or in some cases extreme, disproportion among the number of examples of each class of the problem, then the dataset is said to be imbalanced Fernndez et al. (2018) . For instance, in a cohort of 1000 children, its often the case that mortality group over the study period composes of less than 50 children (representing less than 5%) or less, hence leaving an entire 95% plus as the non-mortality group. Imbalanced data classes are common in many real-life situations including mortality data where the survivors greatly outnumber the mortality, rare disease diagnosis data records where large number of patients do not have the disease, fraud detection, among others. The presence of extreme imbalance between the censored and mortality class with as low as $2-10\%$ data in the minority class is commonly occurring in survival data Afrin et al. (2018).

Classification of imbalanced datasets has been identified as a top problem in machine learning Yang (2006). This makes the class imbalance problem to be of crucial importance since it is encountered by a large number of domains in the real-world. Some of the applications that are known to suffer from this problem includes, fault diagnosis Yang et al. (2009), Zhu and Song (2010), medical diagnosis Mazurowski et al. (2008), disease prediction Khalilia et al. (2011) among others.

A dataset with only two classes is known as binary class, whereas the one with more than two classes is known as multi-class. Both the binary and multi-class datasets suffer from imbalanced data problems Haseeb et al. (2019). This research deals with imbalance in two-class problems. In two-class problems the minority (under-represented) class is usually referred to as the positive class, whereas the ma-

jority class is considered to be the negative one. These terms are used interchangeably in the literature Fernndez et al. (2018). Binary classification are common in various real life aplications like medicine (sick or healthy) among others. In such cases, one of the group becomes the minority while the other is the majority group.

Effects of Imbalance on Datasets

In most of the imbalanced data situations, it is the underrepresented class which is of most interest, since despite its being rare, the minority class may carry important and useful knowledge required in prediction. Such imbalance has been observed to seriously hinder the classification performance of learning algorithms, including Random Forests and other ensemble methods because their decisions are based on classification error Galar et al. (2012).

When a dataset is imbalanced and one class dominates the other, machine learning algorithms such as random forests among others have issues classifying correctly. The algorithms are sensitive to proportions of different classes. They often show biased behavior supporting the majority class and present the minority class lightly Garca et al. (2012), Zhao and Cen (2013), Haseeb et al. (2019). This leads to higher rate of misclassification in the minority class samples Datta and Das (2015), Ertekin et al. (2007) which in turn results in weak predictive accuracy of the minority class and misleading high predictive accuracies in the majority class, as a result of correct classification Cateni et al. (2014), He and Garcia (2009), Japkowicz and Stephen (2002). Thus, the performance of such algorithms decreases significantly when it comes to predicting the minority class.

Many machine learning algorithms are designed to maximize overall accuracy. This can be misleading in imbalanced datasets because the minority class holds a small effect of this measure. However, when data is balanced, accuracy rates tend to decline Olson (2005). This is attributed to the fact that balanced data reduces the training set size leading to degeneracy of the model through omission of cases

encountered to the test set.

Machine learning algorithms aim at minimizing the overall error rate instead of paying attention to the minority class. Therefore, they do not make accurate prediction for the minority class if they do not get the necessary amount of information. In his research demonstrating problems encountered when unbalance data is used in data mining algorithms, Olson (2005) found that algorithms tend to degenerate by assigning all cases to the majority class when data is highly imbalanced and still achieve high accuracy scores. Hence, evaluating algorithm performance using predictive accuracy alone is inappropriate when data is imbalanced.

The main concern in imbalanced problems is that usually, the underrepresented class is the class of interest of the problem from the application point of view Quinlan (1991). With imbalanced data sets, an algorithm does not get the necessary information about the minority class to make an accurate prediction. Due to poor representation and lack of information the minority class is low esteemed. We therefore need to somehow construct classifiers that are biased towards the minority class, without being harmful to the accuracy over the majority class. In order to overcome these issues it is important, when working with such machine learning algorithms to work with balanced classification. However, this is in most cases overlooked. We are therefore interested in construction of classifiers that are skewed toward the minority class, while still maintaining the precision of the majority class.

Data Balancing Techniques

Various techniques have been suggested to solve problems associated with class imbalance. The techniques can be grouped into four categories, subject to how they deal with imbalance. The categories includes:

1. Data level (or external) techniques/ Resampling techniques.

   These techniques involves balancing classes in the dataset before the classi-

fication process using machine learning algorithm. This preprocessing stage reduces the effect caused by imbalance. The balancing is done with an aim of either increasing the minority class or decreasing the majority class to make the two classes approximately equal. The techniques are simple and easy to process and can be used in collaboration with any learning algorithm Fernndez et al. (2018). The key strength of these techniques is that they are independent of the underlying classifier. Many studies in the specialized literature have shown that, for various types of classifiers, re-balancing the dataset remarkably improves the overall performance of the classification in comparison with a non-preprocessed dataset Fernndez et al. (2018).

2. Algorithm level(or internal) techniques.

   These approaches do not cause any change in data distributions. They focus on modifying the classifier learning procedure in order to relieve their bias towards majority class Krawczyk (2016). This necessitates a comprehensive understanding of the selected learning approach in order to identify the specific mechanism responsible for creating bias towards the majority class. In this approach, minority class is taken into consideration and the learner is not allowed to bias for the majority class to overcome the overall cost of misclassification Joshi et al. (2001).

3. Cost-sensitive learning

   Cost-sensitive learning refers to a specific set of algorithms that are sensitive to different costs associated with certain characteristics of considered problems. These costs can originate from various aspects related to a given real-life problem and be provided by a domain expert, or learned during the classifier training phase. A misclassification cost is introduced so as to minimize the conditional risk. By penalizing strongly misclassification cost of the minority class, the classifier tends to bias towards minority class leading to improved generalization on the class. For example, in medical diagnosis, if we declare

a sick person as positive class (minority) and a healthy person as negative class (majority), misclassifying the patient is called false negative meaning the patient was positive but classified as negative. This is a very sensitive and expensive case as compared to false positive because a delay in correct medical diagnosis and treatment can lead to loss of life. By assuming higher costs for the misclassification of minority class samples with respect to majority class samples, cost- sensitive learning can be incorporated both at the data level and at the algorithmic level Lopez et al. (2013)

4. Ensemble-based methods

   In Data Science, Ensemble based classifiers, that is, the combination of several classifiers into a single one, are known to improve the accuracy of a single classifier by training several classifiers and combining them to output a new classifier that outperforms every one of them. However, ensemble learning techniques cannot be able to solve class imbalance problem by themselves since they are designed to optimize accuracy. To deal with class imbalance, ensemble learning techniques are combined with any of the methods mentioned above to improve the final performance. For instance, use of costs in the ensemble learning process or preprocessing the data using a data level approach before learning each classifier. The most famous ensemble learning algorithms include Bagging Breiman (1996), Boosting and Adaboost. For classification of imbalanced data, a novel ensemble technique is also used, that convert an imbalanced dataset into many balanced subsets of original data and number of classifiers with specific classification algorithm are then applied on these multiple subsets.

There is no open directive that indicates the best strategy to use. However, many studies have shown that, external techniques greatly improve the ultimate performance of the classification in comparison with non-preprocessed data set for

various types of classifiers Fernndez et al. (2018). In addition, re-sampling techniques are independent of the classifier, can be easily implemented for any problem and do not need adaptation of any algorithm to the dataset Ofek et al. (2017) . They are also able to effectively balance the dataset resulting in training sets that are suitable for satisfactory calibration of machine learning algorithms Fiorentini and Losa (2020). Chawla et al. (2008), Estabrooks et al. (2004) and Garca et al. (2012) have proved the effectiveness of balancing class distributions using data level techniques.

In this research we apply the Data level/preprocessing (or external) techniques. The methods re-balance the sample space aiming to lessen the effect of imbalanced class distribution in the learning process. The Data level techniques are further classified into three groups Batista et al. (2004) which are: under-sampling methods, over-sampling methods and hybrid methods which combine both sampling techniques. The Data level techniques used in this research are:

- Random under-sampling:

  This aims at balancing dataset by randomly eliminating examples of the majority class up to when the dataset is balanced. The major drawback of this method is that there is a high possibility of discarding potentially useful data pertaining to majority class leading to a possibility of information loss.

- Random over-sampling:

  While the under-sampling method involves removal of samples from the majority group, over-sampling method generates new samples for the minority class. To balance the data using this method, the observations from the minority class are reduplicated. New instances are created from the existing ones; hence over-sampling does not increase information but raises the weight of the minority class by replication. One advantage of over-sampling methods is that there is no information loss. However, since over-sampling simply makes exact copies of the minority class observations, it increases the chances of over

fitting due to replication. Therefore, even if there will be improvement in the training accuracy of the data the overall accuracy of the data may be worse. In addition, while dealing with large imbalanced data sets, over-sampling may increases computational work and execution time Yen and Lee (2009).

- Both-sampling:

  This method combines both under-sampling and over-sampling methods by performing over-sampling with replacement on the minority class while the majority class undergoes under-sampling without replacement.

- Synthetic minority over-sampling technique (SMOTE)

  This is a hybrid method in re-sampling techniques where both under-sampling and over-sampling approaches are combined with an aim to overcome their drawbacks. SMOTE has become one of the most outstanding approaches in data balancing field Fernndez et al. (2018). The key idea in SMOTE proposed by Chawla et al. (2002) is to produce new samples of the minority class artificially. This helps to avoid over fitting brought about by reduplication of minority class instances. Additionally, the majority class examples are under-sampled, giving rise to a more balanced dataset. Generation of Synthetic samples takes the following steps:

- Randomly select a minority and its k nearest minority class neighbors. The value of k is determined by the amount of oversampling needed.

- Calculate the difference between the vector of selected minority and that of one of its nearest neighbors.

- The difference got is then multiplied by a random number between 0 and 1. The result is added to the selected minority vector. By so doing a new random point is added along the line joining the two vectors under consideration.

SMOTE is thus implemented as follows: Let x_i be the feature vector for the selected minority and x_j be the feature vector of a randomly chosen neighbor. A new synthetic minority x_s is generated in the feature space as: $x_s = x_i + \lambda(x_i - x_j)$ where $\lambda \sim Uniform(0; 1)$, is a uniform random variable. A random point is selected along the line segment between two specific features. Thus, the synthetically generated data can be interpreted as a randomly sampled point along the line segment between the two minority samples in the feature space.

In the R environment, Package DMwR Torgo (2010) and ROSE package Lunardon et al. (2013) are used to enhance data balancing. ROSE package Lunardon et al. (2013) is used to enhance data balancing using under-sampling, over- sampling and both-sampling methods. On the other hand, package DMwR Torgo (2010), provides a specific function (smote) to aid the estimation of a classifier in the presence of class imbalance. In SMOTE the parameters *perc.over* and *perc.under* respectively control the amount of over-sampling and under-sampling to be done. If a completely balanced dataset is required, the minority cases are doubled while the majority class is halved. In this study, we used under-sampling, over-sampling, both-sampling and SMOTE methods to balance the Nairobi region dataset. The balanced data was then analyzed using RSF algorithm.

3.2.3 Random Survival Forest Algorithm

The 2014 $KDHS$ dataset had a total of 1099 variables that are possible candidates for predicting child mortality. After some data management exercise, the number of candidate covariates reduced to 757 possible covariates. Before fitting a regression type model in order to embark on the exercise of determining child mortality predictors, we needed to do a variable selection exercise in order to further reduce the variables of importance. This was done using the RSF algorithm. RSF uses decision trees to predict and rank variables that are linked with time to event by

their importance.

RSF Algorithm

Ishwaran et al. (2008) describes the basic Random Survival Forest algorithm as follows:

1. The procedure starts by randomly drawing *ntree* bootstrap samples from the initial data Each bootstrap sample sets aside a mean of 37% of the data called out of bag (OOB) data with respect to the bootstrap sample. Each sample has R predictors (covariates).

2. For each of the drawn samples, a survival tree is grown. Construction of survival tree begins with randomly selecting $mtry$ out of R possible predictors in x for splitting on. The value of $mtry$ depends on the number of available predictors and is data specific. An increase in $mtry$ may tend to result in correlated trees Breiman (2003). All the *ntree* bootstrap samples are designated to the top most (root) node of the tree. The root node is then split into two daughter nodes each of which is recursively split progressively maximizing survival differences between daughter nodes/ increasing within-node homogeneity.

3. Trees are grown to full size until no new daughter nodes can be formed due to the stopping criterion that the end node (most extreme node in a saturated tree) should have larger than or equal to *nodesize* unique events.

4. For each grown tree, compute the cumulative hazard function (CHF). Calculate the mean over all $CHF's$ for the *ntree* trees to attain the ensemble CHF.

5. By using OOB data only, calculate the ensemble OOB error using the first b trees, where $b = 1, ..., ntree$.

By averaging over all trees, a reliable measure of importance of a variable regarding time to event can be obtained Ishwaran and Kogalur (2015). In right censored data, all details of developing a forest take into consideration the outcome. For right censored data, the outcome is survival time and censoring status Breiman (2003).

Node Splitting Process

From the RSF algorithm, a forest originates from randomly drawn *ntree* bootstrap samples. Each bootstrap sample becomes the root of each tree in the forest. There are R predictors in each bootstrap sample. From the R predictors, we randomly select *mtry* predictors for splitting on. The following notations were used in the node splitting precess.

Notations

- h: The h^{th} node of a tree.
- n: The number of individuals within node h.
- T_l: The survival time for the l^{th} individual where $l \in 1, ..., n$.
- δ_l: The censoring information for the l^{th} individual.
- censoring status:

$$\delta_l = \begin{cases} 0, & \text{if individual } l \text{ is censored} \\ 1, & \text{if individual } l \text{ experienced an event (death)} \end{cases}$$

- x: A candidate predictor for node splitting.
- c: A split value for predictor x.

- x^*: The predictor which maximizes survival differences between daughter nodes for predictor x.
- c^*: The split value which maximizes survival differences between daughter nodes for predictor x.
- x_l: Value of x for individual l, $l \in 1, ..., n$.
- j: Daughter node, $j \in 1, 2$
- t_N: The distinct event times in node h, $t_1 < t_2 < ..., < t_N$.
- $d_{i,j}$: Number of deaths at time t_i in the daughter node $j = 1, 2$.
- $d_i = d_{i,1} + d_{i,2}$
- $Y_{i,j}$: Number of individuals at risk (who are alive) at time t_i in the daughter node $j = 1, 2$.
- $Y_{i,1}$ = number of $(T_l \geq t_i, x_l \leq C)$
- $Y_{i,2}$ = number of $(T_l \geq t_i, x_l > C)$
- $Y_i = Y_{i,1} + Y_{i,2}$
- n_j: Total number of observations in daughter j such that $n = n_1 + n_2$ where n_1 =number of $(l : x_l \leq C)$ and n_2 = number of $(l : x_l > C)$

Suppose we take h to be the h^{th} node to be split into two daughter nodes. Within node h, let there be n observations each with survival time denoted by T_l, and censoring status given by

$$\delta_l = \begin{cases} 0, & \text{if individual } l \text{ is censored} \\ 1, & \text{if individual } l \text{ experienced an event (death)} \end{cases}$$

At node h of a tree, the information at time t_i can be summarized in the following table:

Time t_i	Event	Survivor	Risk Set
Node 1	$d_{i,1}$	$Y_{i,1} - d_{i,1}$	$Y_{i,1}$
Node 2	$d_{i,2}$	$Y_{i,2} - d_{i,2}$	$Y_{i,2}$
Total	d_i	$Y_i - d_i$	Y_i

Conditional on the four marginal totals, a single element (say $d_{i,1}$) defines the table. $d_{i,1}$ has the hypergeometric distribution with mean

$$E_i = \frac{Y_{i,1}}{Y_i} d_i$$

and variance

$$V_i = \frac{Y_{i,1}}{Y_i}(1 - \frac{Y_{i,1}}{Y_i})(\frac{Y_i - d_i}{Y_i - 1})d_i.$$

From the randomly selected $mtry$ predictors in node h, take any predictor x (for example age). Using predictor x, find a splitting value c (for example from predictor age, the splitting value could be 2 years). The splitting value c is chosen in such a way that the survival difference for predictor x between $x \leq c$ (criterion for an individual to be placed in daughter node 1) and $x > c$ (the criterion for an individual to be placed in daughter node 2), are maximized. $x \leq c$ separates to the left node while $x > c$ goes to the right node. The survival difference between the two nodes is calculated using a predetermined splitting method. This procedure is repeated with another splitting value c until we get a value which results in maximum survival difference in predictor x. The same procedure is repeated for the remaining $mtry - 1$ predictors in node h. This is done until we get predictor x^* and split value c^* which results in maximum survival difference between the two daughter nodes Weathers and Cutler (2017). The process is repeated at every node. When survival difference is maximum, unlike cases with respect to survival are pushed apart by the tree. Increase in the number of nodes causes dissimilar cases to separate more. This

results in homogeneous nodes in the tree consisting of cases with similar survival. Splitting criteria is one of the aspects of growing a tree. In this section, log rank splitting rule which is commonly used in RSF was used in splitting the node.

Log-rank Splitting Rule

Logrank test is a large-sample chi-square test which makes use of observed versus expected cell counts over categories of outcome. It is the most frequently used statistical test to compare two or more samples non-parametrically with data that are subject to censoring. It's popularity is due to the fact that no modeling assumptions are needed regarding the form of survival distributions. In addition, under proportional hazards, the log rank statistic is optimal among the class of linear rank statistics. The log-rank splitting rule separates the nodes by selecting the split that yields the largest log rank test. PH assumption is the key requirement for the optimality of log rank test. For a split using covariate x and its splitting value c, the goodness of fit is measured using log rank statistics which is represented as;

$$L(x,c) = \frac{\sum_{i=1}^{N}\left(d_{i,1} - \frac{d_i}{Y_i}Y_{i,1}\right)}{\sqrt{\sum_{i=1}^{N}\frac{Y_{i,1}}{Y_i}\left(1-\frac{Y_{i,1}}{Y_i}\right)\left(\frac{Y_i-d_i}{Y_i-1}\right)d_i}} \tag{3.1}$$

This equation measures the magnitude of separation between two daughter nodes. The best split is given by the greatest difference between the two daughter nodes which is given by the largest value of $|L(x,c)|$. The larger the value, the greater is the difference between the two daughter nodes and the better is the split. Hence the best split at node h is determined by finding the predictor x^* and its value at the cut point c^* such that $|L(x^*,c^*)| \geq |L(x,c)|$ for all x and c. This process is repeated at every node.

Survival Tree Estimators

At the initial stage of the tree growing process *ntree* bootstrap samples are randomly selected from the original data. Each bootsrap sample sets aside on average 37% of the data called OOB data while the remaining 63% is called in-bag data. The in-bag data is used to grow the tree and gives estimators which are used for prediction. On the other hand, the OOB data is not involved in the growth of the tree but used for cross-validation purposes. RSF estimates cumulative hazard function (CHF) and survival function based on the terminal nodes using the in-bag and out-of-bag estimators

In-Bag Estimators.

Let $d(t)$ be the number of deaths and $Y(t)$ the individuals at risk at a given time t. The hazard function estimate $H(t)$ at time t with the Nelson Aalen estimator can be expressed as

$$H(t) = \frac{d(t)}{Y(t)} \tag{3.2}$$

The estimate of the CHF for each of the trees grown is accomplished by grouping cumulative hazard estimates by terminal nodes. Suppose:

h denote the terminal node of a tree,

$t_{1,h} < t_{2,h} <, ..., t_{m(h),h}$ denote the distinct event times within node h,

$d_{j,h}$ denote the number of deaths at time $t_{j,h}$ and

$Y_{j,h}$ denote the number of individuals at risk at time $t_{j,h}$.

The CHF for node h is estimated using the bootstrapped NelsonAalen estimator;

$$H_h(t) = \Sigma_{t_{j,h} \leq t} \frac{d_{j,h}}{Y_{j,h}} \tag{3.3}$$

This implies that for a given tree, the hazard estimate for node h is the ratio of events to individuals at risk summed across all unique event times. Each terminal node

of a tree provides a sequence of such estimates and each individual in node h has the same CHF. The survival function for node h is estimated using bootstrapped Kaplan Meier estimator;

$$S_h(t) = \prod_{t_{j,h} \leq t} (1 - \frac{d_{j,h}}{Y_{j,h}}) \tag{3.4}$$

This gives the estimates for the individuals in node h at a given time t. To estimate the CHF for a given predictor X, $H(t \setminus X)$ and the survival function of a given predictor X, $S(t \setminus X)$, X is dropped down the tree and ends up in a unique terminal node due to the binary nature of the tree. This implies that the CHF for x is the same as that of the terminal node it belongs to. That is

$$H(t \setminus X) = H_h(t) \tag{3.5}$$

and

$$S(t \setminus X) = S_h(t), if X \in h \tag{3.6}$$

This defines the CHF and survival function for all individuals in the data and the estimates for the tree. Due to bootstrapping (sampling with replacement) an observation can be found in various bootstrap samples and hence in various trees. The in-bag ensemble estimators for the b^{th} survival tree are computed by averaging the trees estimators. Hence the in-bag ensemble CHF and survival estimators are respectively given as

$$\bar{H}_e(t \setminus X) = \frac{1}{ntree} \Sigma_{b=1}^{ntree} H_b(t \setminus x) \tag{3.7}$$

and

$$\bar{S}_e(t \setminus X) = \frac{1}{ntree} \Sigma_{b=1}^{ntree} S_b(t \setminus x) \tag{3.8}$$

Out-Of-Bag (OOB) Estimators

Let I_i be an indicator pointing to whether case i is in-bag or out of bag such that

$$I_{i,b} = \begin{cases} 1, & \text{if i is an OOB point} \\ 0, & \text{Otherwise} \end{cases}$$

The OOB estimators are determined by whether the case is present in the terminal node. To determine the CHF and survival estimators for an OOB case i, the case is dropped down the tree to a terminal node h. The OOB CHF and survival estimators for i respectively becomes

$$H^*(t \setminus X) = H_h(t), if X_i \in h, I_i = 1 \tag{3.9}$$

and

$$S^*(t \setminus X) = S_h(t), if X_i \in h, I_i = 1 \tag{3.10}$$

The OOB ensemble estimators are calculated by averaging the OOB tree estimators. Hence the OOB ensemble estimators are given as

$$\bar{H}^*_e(t \setminus x_i) = \frac{\Sigma_{b=1}^{ntree} I_{i,b} H_b(t \setminus x_i)}{\Sigma_{b=1}^{ntree} I_{i,b}} \tag{3.11}$$

and

$$\bar{S}^*_e(t \setminus x_i) = \frac{\Sigma_{b=1}^{ntree} I_{i,b} S_b(t \setminus x_i)}{\Sigma_{b=1}^{ntree} I_{i,b}} \tag{3.12}$$

3.2.4 Determining Predictors of Child Mortality

RSF gives a measure of variable importance ($VIMP$) which is totally nonparametric. VIMP has been found to be effective in many applied settings for filtering variables Breiman (2001b), Lunetta et al. (2004), Bureau et al. (2005), Diaz-Uriarte and

Alvarez de Andres (2006), Ishwaran et al. (2009). In this study, using the RSF model, the highly predictive risk factors from the four balanced datasets were extracted. The extracted important predictors were then fitted in the Cox PH model in order to estimate the effect of statistically significant predictors.

The Cox PH model Cox (1972) is frequently used for modeling censored survival data. The model is able to determine collectively the effect of various risk factors on survival duration. It does not presume any shape or distribution of the survival function. In the Cox PH model the instantaneous hazard rate is modeled as a function of time and risk factors as in the equation 3.13.

$$h(t, X) = h_0(t) \exp\left[\Sigma_{i=1}^{p}(\beta_i X_i)\right] \tag{3.13}$$

This equation displays the risk at time t for an individual specified by a set of covariates X. In this case, X is a group of variables that are used in the model for prediction of the risk of the given observations. From the formula, the risk at time t is a product of $h_0(t)$, the baseline hazard function and $\exp\left[\Sigma_{i=1}^{p}(\beta_i X_i)\right]$, the exponential to the sum of the p predictor variables in X. The baseline hazard function is a function of time which indicates what the risk would be when there are no covariates (all covariate values are zero). The coefficient β_i gives the magnitude of the influence of the covariates. One of the assumptions of the Cox model is that the variables should satisfy the proportional hazard assumption.

Proportional Hazard (PH) Assumption in Cox Model

In survival analysis, hazard is the likelihood of an event happening at any given time point given that the event had not occurred. If we are concerned with only one predictor, the hazard is given by

$$h(t, x) = h_0(t) \exp\left[\beta x\right] \tag{3.14}$$

In most cases, we compare subjects or groups with respect to their hazards by using the hazard ratio. If we are concerned with two subjects with covariates x_1 and x_2, the respective hazards are given by

$$h(t, x_1) = h_0(t) \exp [\beta x_1] \tag{3.15}$$

and

$$h(t, x_2) = h_0(t) \exp [\beta x_2] \tag{3.16}$$

The hazards for the two subjects can give a hazard ratio as;

$$HR = \frac{h(t, x_2)}{h(t, x_1)} = \frac{h_0(t) \exp [\beta x_2]}{h_0(t) \exp [\beta x_1]} = \exp [\beta (x_2 - x_1)] \tag{3.17}$$

If we take $x_2 = x_1 + 1$ to represent one unit increase on the risk of event in covariate x_1, the hazard ratio becomes

$$HR = \exp [\beta (x_1 + 1 - x_1)] = \exp \beta \tag{3.18}$$

Hence the hazard ratio is constant or proportional while the hazard rate varies over time which is the assumption of proportional hazards. We therefore say that a variable has a time-varying effect when the HR is not constant over time. For example, the effect of a treatment may be strong at the beginning of the treatment but this may change with time. This is different from time-varying covariate which means that the value of a variable is not fixed over time. For example age, weight, smoking status, among others. It is however possible to have a time-varying covariate whose effect changes with time.

The coefficient β is the log of the hazard ratio. Hence, a value of $\beta > 0$ implies that $HR > 1$ which is an indication that the risk of event for subjects with covariate x_2 is higher in comparison with subjects with covariate x_1. On the other hand, a value of $\beta < 0$ implies that $HR < 1$ which is an indication that the risk of event for

subjects with covariate x_2 is lower in comparison with subjects with covariate x_1.

The HR in the cox model is estimated by taking into consideration each event time instead of follow up time. For the overall follow-up period, the HR is estimated by giving the same weights to the HR at the beginning of the follow up period which affects almost all the observations and also to the HR at the end of the follow up period which affects only a few observations who are still at risk. The HR is then averaged over the event times. If the HR changes over time (the hazard rates are not proportional), equal weighting may not give the actual representation of the HR which may lead to biased results O'Quigley and Pessione (1991).

Checking the Cox-PH Assumptions

For appropriate use of the Cox proportional hazards regression model, there are several important assumptions that need to be checked. These include:

- PH assumptions. These were checked graphically using Kaplan Meir curves and schoenfeld residuals. The graphs of Schoenfeld residuals are shown in the appendix in figures E.17 and E.29. However, these methods do not provide formal diagnostic tests Bellera et al. (2010).It is therefore important to test the PH assumption using statistical tests where proportionality of the hazard is equivalent to testing if the variable is not significantly different from zero.

- Functional relationship between the log hazard and the covariates. Martingale residuals were used to assess this assumption. The graphs of Martingale residuals are shown in the appendix in figures E.20 and E.24

- Possible presence of outliers or influential observations. Deviance residual was used to examine possible presence influential observations. The graphs of Deviance residuals are shown in the appendix in figures E.19, E.31, E.28, E.23, E.18, E.22, E.26 and E.30

3.2.5 Model Selection Criterion

Comparison of prediction accuracy of the different models was done based on concordance index. In survival analysis, a pair of observations is said to be concordant if, for the individual that got the event first, the model predicts a higher risk of event. The concordance probability is the frequency of concordant pairs among all pairs of subjects. Harrells concordance index (C-index) Harrell et al. (1982) is used to estimate prediction error. It estimates the likelihood that in a pair of cases selected at random, the case that came to have an event first had a worse predicted result. Suppose we have two observations whose outcome is predicted. If the observation predicted to have the worst outcome experiences an event first, then the two observations are said to be concordant (i.e. they have the appropriate practice). Computation of concordance error rate is as given below.

1. The procedure begins by forming all potential pairs of observations from the entire data.

2. A pair is omitted if:

   - The observation with shorter duration of survival is censored.
   - Duration of survival is equal for the pair but one or both observation is censored.

3. After the omissions are done, we remain with all the other pairs which are referred to as permissible pairs. A score of value 1 is given to a permissible pair if:

   - For all pairs having unequal survival durations resulting in prediction being worse for the observation with shorter survival duration.
   - For all pairs having uniform survival durations resulting in similar prediction results

- For all pairs having equal survival duration given that not both observations are events, the observation with event results in a worse prediction outcome.

A score of value 0.5 is given to a permissible pair if:

- For all pairs having unequal survival duration, the prediction outcome is equal.
- For all pairs having equal survival duration, prediction outcomes are not equal.
- For all pairs having equal survival duration given that not both observations are events, prediction outcome is worse for the observation with censored results.

If we denote the sum of all the permissible pairs as Concordance, then the concordance index, C is defined as:

$$C = \frac{concordance}{permissible}$$

The error rate, E is given by $E = 1 - C$ where $0 \leq E \leq 1$. $E = 0$ indicates perfect accuracy while $E = 0.5$ is equivalent to random guessing.

3.3 Analysis of $BRSF$ Model under Different Splitting Rules.

3.3.1 Data Description

In this section, we used the data balanced using under-sampling method, the method that performed best after different balancing methods were compared. The data

originated from the 2014 KDHS data from which the Nairobi region subset was extracted. The balancing process is as described in section 3.2.2. This dataset had a total of 68 observations with the mortality and censored classes each having 34 observations which represent 50% of the sample. The number of variables in the dataset was 757.

3.3.2 Exploration of the Data

Exploration of the Survival Trends in the Balanced Data

In order to get the general view of the survival trends in the data set, the table of survival estimates was generated. This is shown in table 3.9 and the survival curve in figure 3.5. Computation of the estimated survival function in the presence of right censoring, was done using the Kaplan Meier estimator as shown in table 3.9. The table shows the survival estimates at the event times for children under five years of age from the time they were born to the time of interview (end of the follow up period). The table also gives the number at risk of death, survival probabilities with their associated standard errors as well as the upper and lower confidence intervals for the respective survival probabilities.

The survival curves shown in figure 3.5 gives the probability of survival with the bands giving approximate confidence intervals. The horizontal axis indicates time in months starting from 0 to 60 months while the y axis indicates the survival probabilities or the proportion of individuals surviving (at risk). From table 3.9 and figure 3.5, the highest number of the deaths (18) occurred before the first month was over. The survival probability at time 0 months is estimated as 0.735 due to the deaths experienced before the end of the first month. The curve then drops gradually with each step downwards indicating death of one or more individuals with the last death occuring during the 24^{th} month.

Table 3.9: Survival Estimates for Under-sampled Nairobi Dataset.

Time in months	n.risk	n.event	Survival	std.err	lower95%$C.I$	Upper 95%$C.I$
0	68	18	0.735	0.0535	0.638	0.848
1	50	1	0.721	0.0544	0.621	0.836
2	49	1	0.706	0.0553	0.605	0.823
3	48	3	0.662	0.0574	0.558	0.784
5	44	1	0.647	0.0580	0.542	0.771
6	43	1	0.632	0.0586	0.527	0.758
7	42	1	0.617	0.0591	0.511	0.744
9	39	2	0.585	0.0601	0.478	0.716
11	37	1	0.569	0.0606	0.462	0.701
12	36	2	0.538	0.0612	0.430	0.672
17	32	1	0.521	0.0615	0.413	0.656
19	30	1	0.503	0.0619	0.396	0.641
24	28	1	0.485	0.0622	0.378	0.624

Sample size	No. of events	Median	lower95%$C.I$	Upper 95%$C.I$
68	34	24	9	NA

3.3.3 Exploration of the Proportional Hazards (PH) Assumption in the Balanced Data

The PH assumption implies that for any two categories of a variable of interest, the ratio of the hazard is unchanging or constant over time. It is essential to verify that the predictor variables in the model satisfy the PH assumptions. More about PH assumption is given in section 3.2.4. We used Kaplan Meier curves to explore the PH situation. The curves plots the estimated proportion at risk (survival probability) against time giving the estimated survival functions Clark et al. (2003). The curves are in form of step functions with each vertical drop pointing out one or more deaths happening Bewick et al. (2004). If the variables satisfy PH assumption, the survival curves should be parallel. If for two or more categories of a variable of interest do not result to parallel curves or the curves cross, then it is an indication that the PH assumption is violated. The figures 3.6, 3.7 and 3.8 shows the Kaplan Meir curves for some of the categorical variables in the data.

The curves show the probability of survival for children under five years. The horizontal axis indicates time in months starting from 0 to 60 months while the y

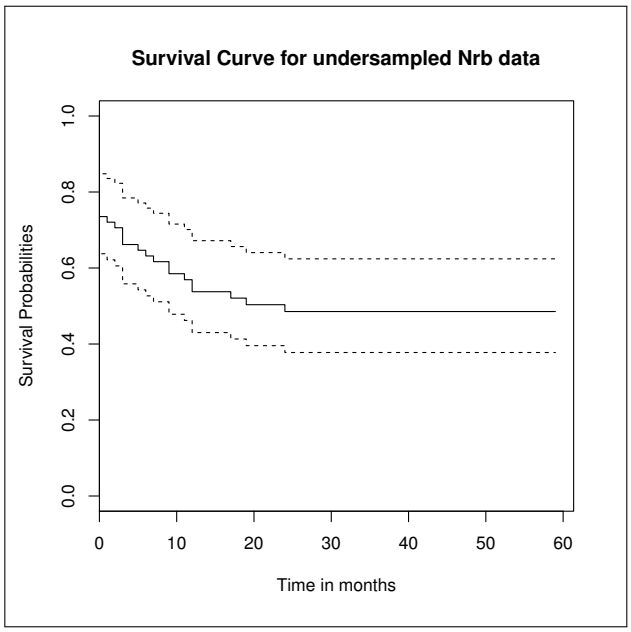

Figure 3.5: Kaplan Meir plot for the Undersampled Nairobi data.

axis indicates the survival probabilities or the proportion of individuals surviving (at risk). From figure 3.6, the highest level of education attained is classified into four categories (Higher, secondary, primary and no education). The curve for observations in the "no education" category is a horizontal line with survival probability of 1. This is as a result of the presence of observations with no education but none of their children encountered an event during the follow up period. This is also indicated in table 3.7. The curve for "primary education" category remained consistently higher than that of "Higher" and "secondary education". There is evidence of crossing curves between "secondary" and "higher education" categories indicating violation of the PH assumption. We can read from the curves that children from parents with no education have better survival prognosis than those who acquired the other levels of education. Similarly, children whose parents acquired primary education level have better survival prognosis than those with secondary and higher education levels.

From figure 3.7, the curves of survival by sex seems to be proportional over time with the KM curve for the female children being consistently higher than that of the

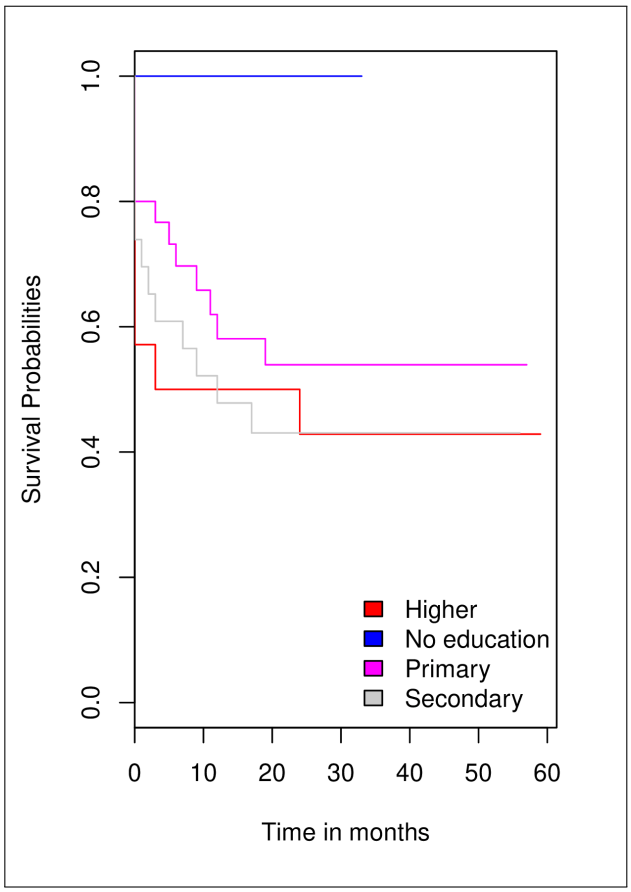

Figure 3.6: Survival curves by education for under-sampled Nrb data.
The presence of crossing curves in the figure indicates violation of PH assumption.

male children. This implies that the female children have better survival prognosis than male children. In a similar way, survival curves by wealth index show violation of PH assumption due to the presence of crossing curves between the poorer category and the middle level category. From the KDHS data, Wealth index was categorized into 5 groups; poorest, poorer, middle, richer and richest. However, in the under-sampled Nairobi region data, we only had the 3 categories (poorest, poorer and middle) as shown in figure 3.8. The vertical red line represents individuals in the poorest wealth index category who did not survive (got an event) hence the survival probability is zero. The blue line indicates individuals in the poorer wealth category. Some did not survive having zero probability of survival while those who survived had a constant probability of survival. In the middle level wealth index category, we have some events with zero probability. The curve for the poorer level crosses with

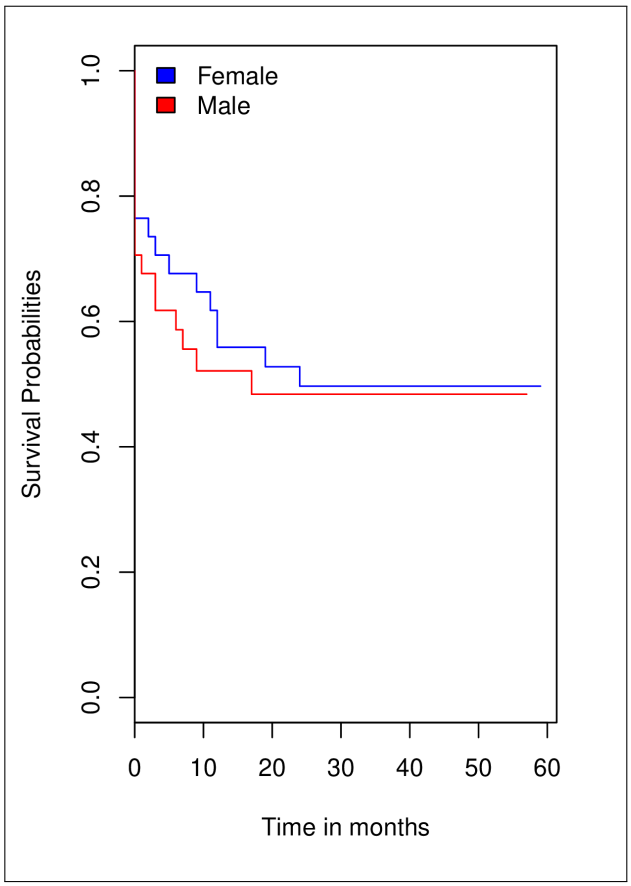

Figure 3.7: Survival curves by sex for the Under-sampled Nairobi data.

that for the middle level showing violation of PH assumption.

The PH assumptions were also checked using graphical diagnostics based on the scaled schoenfeld residuals. The Schoenfeld Residuals Test is analogous to testing whether the slope of scaled residuals on time is zero or not. If the slope is not zero then the proportional hazard assumption has been violated. Graphs of the scaled schoenfeld residuals are shown in figure E.17 in the appendices section.

3.3.4 Random Survival Forests using Different Splitting Rules

After exploration of the PH assumptions, the data was analyzed using the RSF algorithm given in section 3.2.3. The generation of the tree was also done as indicated in section 3.2.3.

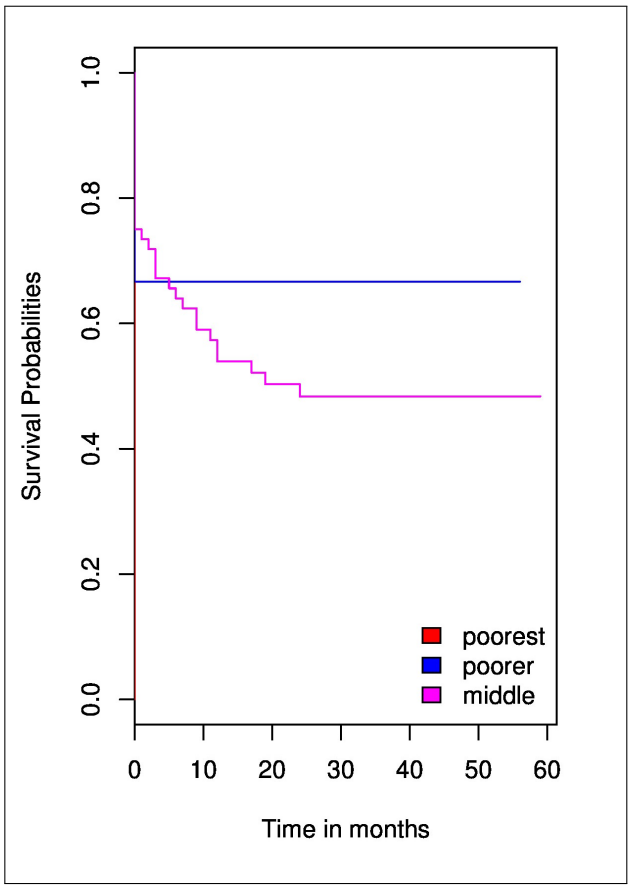

Figure 3.8: Survival by wealth for under-sampled Nairobi data.
The curves are crossing and not parallel indicating violation of PH assumption.

3.3.5 Splitting Rules

The choice of splitting rule is very important to the performance of a growing tree. Below we give the different splitting rules that can be used in splitting the node in addition to the log rank splitting rule given in section 3.2.3.

Weighted Log-rank Test

When the proportional hazards assumption is violated, logrank test has been shown to lose its power making it inefficient and a weighted version more suitable. Weighted log-rank tests with various fixed and adaptive weight functions have been proposed in the literature to increase the power of a test when non-proportional hazards are expected. The weighted Log-rank test is used when we want to compare groups but

wish to give more importance ("weight") to certain events. This makes it a very useful test when hazards are not proportional. For a split using covariate x and its splitting value c, the measure of node separation using weighted log rank statistics is represented as

$$L(x,c) = \frac{\sum_{i=1}^{N} w_{(t)} \left(d_{i,1} - \frac{d_i}{Y_i} Y_{i,1} \right)}{\sqrt{\sum_{i=1}^{N} w_{(t)}^2 \frac{Y_{i,1}}{Y_i} \left(1 - \frac{Y_{i,1}}{Y_i} \right) \left(\frac{Y_i - d_i}{Y_i - 1} \right) d_i}}$$

If $w_{(t)}$is constant over j we get the standard log-rank test statistic.

The Log-rank Approximation Splitting Rule

This is an approximation to the log rank split rule and is therefore named as approximate logrank splitting. It splits the nodes by using an approximation of the log-rank test to reduce computations. Approximating the numerator of $L(x,c)$, a revision is done using the Nelson-Aalen cumulative hazard estimator for the parent node. The Nelson-Aalen Estimator is given as

$$\hat{H}t = \Sigma_{t_i \leq t} \frac{d_i}{Y_i}$$

The numerator of $L(x,c)$ can be rewritten as

$$\sum_{i=1}^{N} \left(d_{i,1} - \frac{d_i}{Y_i} Y_{i,1} \right) = \sum_{i=1}^{N} d_{i,1} - \sum_{i=1}^{N} \frac{d_i}{Y_i} Y_{i,1} = D_j - \sum_{i=1}^{N} I[x \leq c] \hat{H}(T_l)$$

where

$$D_j = \sum_{i=1}^{N} d_i, j, j \in 1, 2$$

As suggested by Leblanc and Crowley (1993), the denominator can be simplified by approximating the variance of $L(x,c)'s$ numerator by letting $D = \sum_{i=1}^{N} d_i$. This

leads to the approximation of the logrank test given by

$$L(x,c) = \frac{D^{\frac{1}{2}}\left(D_j - \sum_{l=1}^{n} I\{x_l \leq c\}\hat{H}(T_l)\right)}{\sqrt{\left(\sum_{l=1}^{n} I\{x \leq c\}\hat{H}(t_l)\right)\left(D - \sum_{l=1}^{n} I\{x \leq c\}\hat{H}(T_l)\right)}}$$

The Log-rank Score Splitting Rule

Log-rank score splitting rule Hothorn and Lausen (2003) was developed from log-rank split rule. The method splits the nodes using a standardized log-rank statistic. The ranks for each survival time T_l are computed given an ordered predictor x such that $x_1 \leq x_2 \leq, ..., \leq x_n$. The rank for each survival time T_l is calculated as

$$a_l = \delta_l - \sum_{k=1}^{\Gamma_l} \frac{\delta_k}{n - \Gamma_l + 1}$$

Where Γ_l =the number of $(t : T_t \leq T_k)$. Let $\bar{a}$ and S_a^2 be the sample mean and sample variance of a_l for $l \in 1, ..., n$. The formula for the log-rank score test is given by:

$$S(x,c) = \frac{\sum_{x_l \leq c} a_l - n_1\bar{a}}{\sqrt{n_1\left[1 - \frac{n_1}{n}\right]S_a^2}}$$

This split rule defines the measure of node separation by $|S(x,c)|$ where the best split is given by the maximum value over x and c.

The Conservation-of-events Splitting Rule

The conservation-of-events splitting rule splits the nodes by finding daughter nodes closest to the conservation-of-events principle. This principle states that the sum of the estimated cumulative hazard function over the observed time points (deaths and censored values) must equal the total number of events. This is done by using an altered version of the Nelson- Aalen estimator which is now computed for each daughter node rather than the parent node. The Nelson-Aalen cumulative hazard

estimator for daughter j is given by

$$\hat{H}_j t = \Sigma_{t_{i,j} \leq t} \frac{d_{i,j}}{Y_{i,j}}$$

Where $t_{i,j}$ is the ordered event times for daughter j. The total number of events for each daughter j can be retained by using

$$\sum_{l=1}^{n_j} \hat{H}_j(T_{l,j}) = \sum_{l=1}^{n_j} \delta_{l,j}$$

The total number of deaths is conserved in each daughter node. Order the time points within each daughter node such that $T_{(1),j} \leq ..., \leq T_{(n_j),j}$ Let $\delta_{l,j}$ be the censoring indicator function for the ordered value $T_{(l),j}$. In order to get a measure of the accuracy of the conservation of events, we define

$$\mu_{k,j} = \sum_{l=1}^{n_j} \hat{H}_j(T_{l,j}) - \sum_{l=1}^{n_j} \delta_{(l),j}$$

The measure of conservation of events for the split on x at the value c is

$$conserve(x, c) = \frac{1}{Y_{1,1} + Y_{1,2}} \sum_{k=1}^{n_j - 1} Y_{1,j} \sum_{k=1}^{n_j - 1} |\mu_{k,j}|$$

In other words, for each daughter j, the magnitude of $\mu_{k,j}$ are summed and weighted by the number of individuals at risk within each daughter node. This value is small if two groups are well separated since the level of separation between the two daughter nodes increases as the test statistics decreases. In order for us to obtain the "best" split, we have to minimize this value or maximize the transformed value, $\frac{1}{conserve(x,c)}$. Since we want to maximize survival difference due to a split, we use the transformed value $\frac{1}{conserve(x,c)}$ as our measure of node separation. This statistic can be very time-consuming to compute because it sums over all the survival times within each daughter node. Fortunately, the computation time can be severely decreased by

using only the event times which can be expressed by the following formula:

$$conserve(x,c) = \frac{1}{Y_{1,1}+Y_{1,2}} \sum_{j=1}^{2} Y_{1,j} \sum_{k=1}^{N-1} [N_{k,j} T_{k+1,j} \sum_{l=1}^{k} \frac{d_{l,j}}{Y_{l,j}}]$$

Where $N_{i,j} = Y_{i,j} - Y_{i+1,j}$ is the amount of observations within daughter j with observed time falling within the interval $[t_i, t_{i+1})$ for $i = 1, ..., N$ where $t_{N+1} = \infty$. The two formulas for $conserve(x,c)$ can be shown to be equivalent.

Brier Score Gradient (Bs. gradient) Splitting Rule

Brier Score (BS) is the most frequently used scalar summary of correctness for probability predictions for binary events. Let $y_i, i = 1, 2, , n$ be the i^{th} likelihood prediction in a series of n such predictions. The paired observation $x_i = 1$ if the event of interest occurs on the i^{th} occasion, and $x_i = 0$ otherwise. The BS is then simply the meansquared error over the n forecast observation pairs,

$$BS = \frac{1}{n} \sum_{i=1}^{n} (y_i - x_i)^2$$

Suppose we have a pair of predictor-response, say (X_i, Y_i) for $i = 1, 2, ..., n$. The usual regression techniques attach the conditional mean of the response variable Y to a given set of predictors X. Meinshausen (2006) introduced Quartile Regression Function (QRF) which connects between an empirical cumulative distribution function and the outputs of a tree. Let D_0 be a group of randomly selected variables to be split into two daughter nodes D_1 and D_2. Suppose the homogeneity of each group is defined by

$$v(D_j) = \sum_{Y \in D_j} [Y - \bar{Y}(D_j)]^2$$

where $\bar{Y}(D_j)$ is the sample mean in D_j. For an optimal splitting selection, comparison is done between the homogeneities of $v(D_1)$ and $v(D_2)$ with that of

$v(D_0)$. The splitting value s is the one that maximizes

$$H(D_1, D_1) = \max_{s \in \epsilon^*}[v(D_0) - v(D_1) - v(D_2)]$$

Where ϵ^* is a randomly selected sample of predictors from the predictor space ϵ. The resulting nodes are recursively split until the stopping criterion is reached. The terminal node gives the predicted value. Athey et al. (2016) suggested that instead of maximizing variance heterogeneity of the daughter nodes, one maximizes the criterion

$$\Delta(D_1, D_2) = \sum_{j=1}^{2} \frac{-1}{i : Y_i \in D_j}(\sum_{i:Y_i \in D_j} \rho_i)^2$$

where

$$\rho_i = 1(Y_i \geqslant \hat{\theta}_q, D_0)$$

is an indicator function which is equal to one when Y_i is greater than the q^{th} quantile θ_q, D_0 of the observations of node D_0. The choice of ρ_i is linked with a gradient based approximation of the quantile function

$$\psi_{\hat{\theta}_q, D_0} = q1(Y_i > q) + (1 - q)1(Y_i \leq q)$$

hence the term gradient forest. The order for each split is chosen among given orders $(0.1, 0.5, 0.9)$

Weighted Log-rank Test with Adaptive Weights.

Yang and Prentice (2010) showed that log rank test can be improved by using weighted log rank statistics with adaptive weights. Adaptive weights are obtained by fitting the data to the model of Yang and Prentice (2005) which contains proportional hazard models and proportional odds model. The model accommodates a variety of non proportional hazard situations.

Model of Yang and Prentice (2005)

Assuming that the event times are absolutely continuous, Yang and Prentice (2005) proposed a model in which

$$H_R(t) = \frac{\theta_1\theta_2}{\theta_1 + (\theta_2 - \theta_1)Y_L(t)} H_L(t), t < \tau_0$$

where

$H_R(t)$ is the hazard function of the right branch.

$H_L(t)$ is the hazard function of the left branch.

$Y_R(t)$ is the survival function of the right branch.

$Y_L(t)$ is the survival function of the left branch.

$\tau_0 = supt : Y_L(t) > 0$ The hazard ratio between the two branches under this model are not constant. At time t, the hazard ratio is given by

$$\frac{H_R(t)}{H_L(t)} = \frac{\theta_1\theta_2}{\theta_1 + (\theta_2 - \theta_1)Y_L(t)}$$

which depends on $\theta_1, \theta_2 and Y_L(t)$

θ_1, θ_2 are constants.

If $\theta_2 > \theta_1$, the ratio is monotonically increasing.

If $\theta_1 > \theta_2$, the ratio is monotonically decreasing.

$$\theta_1 = \lim_{t \downarrow 0} \frac{H_R(t)}{H_L(t)}$$

which can be interpreted as short term hazard ratio.

$$\theta_2 = \lim_{t \uparrow \tau_0} \frac{H_R(t)}{H_L(t)}$$

which can be interpreted as long term hazard ratio. The model contains proportional hazard model corresponding to $\theta_1 = \theta_2$ and proportional odds model corresponding

to $\theta_2 = 1$ as two submodels.

When $\theta_1 = \theta_2$, $H_R(t) = \theta_1 H_L(t)$

When $\theta_2 = 1$, $H_R(t) = \frac{\theta_1}{\theta_1 + (\theta_2 - \theta_1) Y_L(t)} H_L(t)$. Various combinations of θ_1 and θ_2, give different non proportional hazard patterns such as

$\theta_1 = 1$ for no initial effect.

$\theta_1 < 1$ and $\theta_2 > 1$ or $\theta_1 > 1$ and $\theta_2 < 1$ for crossing survival functions. A χ^2 test using the two estimating functions of the right and left branches is used to test the hypothesis of significant difference Yang and Prentice (2010).

In the process of analysing BRSF using different splitting rules, we only worked with three different splitting rules which are logrank, logrank score and Bs.gradient splitting rules. The survival tree estimators are as in section 3.2.3

Following variable selection with RSF using 3 different splitting rules, the respective selected variables were subjected to variable predicton.

3.3.6 Prediction of Child Mortality

In the previous section, variables selected using RSF were fitted in Cox PH model for prediction. Some of the variables that did not satisfy the assumption were removed from the model. This could lead to removal of highly predictive variables in the model. In this section, we worked with the Cox-Aalen's model which is an appropriate alternative to the Cox PH model when PH assumptions are violated.

COX-Aalen's Model

In the Cox PH model, the effect of covariates is assumed to act multiplicatively on the baseline hazard rate and the ratio of the hazards is constant over time. When the PH assumption is not satisfied, the Cox model can lead to biased results. In some datasets, some of the covariate effects may be constant while others may not

be constant. In such situations, the Cox Aalen model which combines the two types of covariates in the same model is a better alternative.

The Cox-Aalen regression model proposed by Scheike and Zhang (2002) is a combination of additive and multiplicative model. In this model, the covariates are partitioned into two parts in which some act additively on the intensity while others work multiplicatively. The model is defined by,

$$h(t\backslash x) = Y(t)[X(t)^T\alpha(t)]exp(Z(t)^T)\beta \tag{3.19}$$

where

$Y(t)$ is the indicator of the risk,

$X(t)$ is the additive non parametric time varying covariate,

$Z(t)$ are the covariates with constant multiplicative effects,

$\alpha(t)$ is a $(p \times 1)$ vector of time varying regression coefficients, and

β is a $(q \times 1)$ vector of relative risk regression coefficients.

$X((t)andZ(t))$ are $(p + q) \times 1$ vectors of covariates.

Fitting the selected variables in the Cox Aalen's model resulted to 3 different sets of determinants of U5CM. Model selection using concordance index was done as in section 3.2.5 to evaluate the models.

3.4 Developing an Improved Balanced Random Survival Forest ($IBRSF$) Algorithm for Right Censored Data in Situations where PH Assumptions are Violated.

In this section we develop an $IBRSF$ model for highly imbalanced right censored data in situations where PH assumption is violated. We came up with a unified model for data balancing, variable selection and survival analysis. The model follows a three stage progression to establish the determinants of U5CM.

3.4.1 Data Balancing Stage

The first stage involves data balancing using under sampling method which was found to do well among the data level balancing techniques. Data balancing is covered in section 3.2.2.

3.4.2 Variable Selection and Prediction Stages

The dataset in use is associated with 757 variables after data cleaning which are candidate determinants of U5CM. There is therefore need for proper variable selection exercise in order to identify the correct set of variables to use for survival analysis. All the 757 available variables were used in the selection process. The balanced data is integrated with the RSF algorithm in the variable selection stage. This serves as a good starting point for identification of potential predictors from a dataset with a large number of variables. In the RSF algorithm, BS.gradient splitting rule is used for splitting the nodes. Using RSF VIMP, the most predictive variables were selected. These are the variables with importance level greater than of equal to 0.002.

The variables selected during the variable selection stage are then subjected to RSF VIMP a second time for identification of determinants of U5CM. In application of the RSF algorithm the node splitting process and survival tree estimators was done as in sections 3.2.3 and 3.2.3 respectively. We also constructed confidence intervals using sub-sampling approach proposed by Ishwaran and Lu (2019) for the important variables selected in the final stage.

3.4.3 Calculation of Variable Importance (VIMP)

The most commonly used measure of importance is known as permutation importance. This measure assumes a prediction based perspective by using prediction error on account of the variable. It estimates error by making use of OOB cases. To calculate VIMP, all values of the j^{th} variable are randomly permuted in the OOB cases for a tree. The new covariate value is put down the tree and a new internal error rate computed. The importance for the j^{th} variable in the tree is given by the difference between the new error and the original OOB. VIMP is then got by averaging over the forest. Given the learning data

$$L = (X_1, Y_1), ..., (X_n, Y_n) \tag{3.20}$$

where Y is the response and X a set of p-dimensional predictor variables. We need to estimate the function $h(x)$ of the response given $X = x$

Tree VIMP

Let $L^*(\theta_m)$ be the m^{th} bootstrap sample and $L^{**}(\theta_m)$ be the correspondin OOB data. We can write $X = (X^{(1)}, ..., X^{(j)}..., X^{(p)})$ where $X^{(j)}$ is the j^th variable coordinate. Denote the permuted value of the j^th coordinate of X by $\hat{X}^{(j)}$ Substituting this into the j^{th} cordinate of X gives $\hat{X}^{(j)} = (X^{(1)}, ..., X^{(j-1)}, \hat{X}^{(j)}, X^{(j+1)}..., X^{(p)})$ The

difference between the prediction error under the original X and the permuted $\hat{X}^{(j)}$ results in the tree Vimp. Suppose we denote the Vimp for $X^{(j)}$ for the m^{th} tree by $I(X^{(j)}, \theta_m, L)$, Then,

$$I(X^{(j)}, \theta_m, L) = \frac{\sum_{i \in L^{**}(\theta_m)} l(Y_i, h(\hat{X_i^j}, \theta_m, L)}{\sum_{i \in L^{**}(\theta_m)} I} - \frac{\sum_{i \in L^{**}(\theta_m)} l(Y_i, h(X_i, \theta_m, L)}{\sum_{i \in L^{**}(\theta_m)} I} \quad (3.21)$$

This can be written as

$$I(X^{(j)}, \theta_m, L) = \frac{1}{N(\theta_m)} \sum_{i \in L^{**}(\theta_m)} [l(Y_i, h(\hat{X_i^j}, \theta_m, L)) - l(Y_i, h(X_i), \theta_m, L)] \quad (3.22)$$

Forest VIMP

Averaging the tree VIMP over the forest results in VIMP which is given as

$$I(X^{(j)}, \theta_1, ..., \theta_m, L) = \frac{1}{M} \sum_{m=1}^{M} I(X^{(j)}, \theta_m, L) \quad (3.23)$$

A $100(1-\alpha)$ confidence region for the true VIMP can be defined as

$$\hat{\theta}_n^{(j)} \pm \frac{\alpha}{2} \sqrt{\hat{v_n^{(j)}}} \quad (3.24)$$

where $Z_{\frac{\alpha}{2}}$ is the $1 - \frac{\alpha}{2}$ quantile from a standard normal,$Pr[N(0,1) \leq Z_{\frac{\alpha}{2}}] = 1 - \frac{\alpha}{2}$

3.4.4 IBRSF Algorithm

1. The procedure starts with data balancing using under sampling method to make the mortality and non mortality classes approximately equal. The balanced data is then subjected to the tree growing process.

2. *ntree* bootstrap samples are randomly drawn from the balanced dataset. From each bootstrap sample a mean of 37% of the data called out of bag (OOB)

data is set aside with respect to the bootstrap sample. Each bootstrap sample has R covariates.

3. For each of the drawn samples, a survival tree is grown. Construction of survival tree begins with randomly selecting $mtry$ out of R possible predictors in x for splitting on. The value of $mtry$ depends on the number of available predictors and is data specific. All the $ntree$ bootstrap samples are designated to the root (top most) node of the tree. The root node is then split into two daughter nodes each of which is recursively split progressively maximizing survival differences between daughter nodes. Bs.gradient splitting rule is used to split the node.

4. Trees are grown to full size until no new daughter nodes can be formed due to the stopping criterion that the end node should have larger than or equal to $nodesize$ unique events.

5. For each grown tree, compute the cumulative hazard function (CHF). Calculate the mean over all $CHFs$ for the $ntree$ trees to attain the ensemble CHF.

6. By using out-of-bag (OOB) data only, calculate the ensemble out of bag error using the first b trees, where $b = 1, ..., ntree$.

Chapter 4

RESULTS

4.1 Introduction

This chapter presents the findings of our research which include; the outcome of balancing, the resulting important variables after variable selection and the resulting predictors of U5CM.

4.2 Results for Analysis of BRSF using Different Balancing Methods.

4.2.1 Data Balancing using Different Balancing Schemes

The Nairobi region dataset consisted 757 variables and 532 observations. This data set was found to be highly imbalanced with 498 majority instances and 34 minority instances as demonstrated in table 3.5. The dataset was successfully balanced using four different balancing methods. The result of balancing the mortality and non mortality classes using different balancing methods in the region are shown in table

4.1.

Table 4.1: Balanced Nairobi Region data with Different Balancing Methods

Balancing method	Status	Total	Percentage
Under-samplig	Censored	34	50%
	Uncensored	34	50%
	Total	68	100%
Over-samplig	Censored	498	50%
	Uncensored	498	50%
	Total	996	100%
Both-samplig	Censored	520	52%
	Uncensored	480	48%
	Total	1000	100%
SMOTE	Censored	68	50%
	Uncensored	68	50%
	Total	136	100%

From table 4.1 the datasets are balanced with the mortality and non mortality classes having equal or approximately equal representation. Different balancing methods resulted in different sample sizes. In oversampling technique, the minority class was oversampled until the number of minority instances became equal to the number of majority instances which is 498. This resulted to a total of 996 observations in the sample. Similarly, in under-sampling method, instances were randomly removed from the majority class until the total number of observations in the majority class became 34 instances. This led to a sample with 68 observations in total. In both-sampling method, both over-sampling and under-sampling on the imbalanced data took place. In this case, over-sampling with replacement was conducted on the minority class while under-sampling without replacement was performed on the majority class. From this combination, we expect cases of repeated observations due to over-sampling and removal of some information from the original data due to under-sampling. SMOTE balancing involved doubling of the minority cases while the majority class were halved. This was done by use of some parameters which led to complete balance of the data set.

Balancing of the overall dataset resulted to balance in the covariates. The results of data balance in some of the covariates is shown in table 4.2 and 4.3. These covariates includes education level and child sex. The graphs showing balance within

the covariates are shown in the appendices in section D.

Table 4.2: Balanced Nairobi Region data grouped by Education Level

Balancing method	Status	Higher	No education	Primary	Secondary	Sum
Under-samplig	Censored	6	1	17	10	34
	Uncensored	8	0	13	13	34
	Sum	14	1	30	23	68
Over-samplig	Censored	102	7	203	186	498
	Uncensored	131	0	172	195	498
	Sum	233	7	375	381	996
Both-samplig	Censored	74	11	232	203	520
	Uncensored	117	0	195	168	480
	Sum	191	11	427	371	1000
SMOTE	Censored	14	0	30	24	68
	Uncensored	11	0	29	28	68
	Sum	25	0	59	52	136

Table 4.3: Balanced Nairobi Region data grouped by child sex.

Balancing method	Status	Female	Male	Sum
Under-sampling	Censored	17	17	34
	Uncensored	17	17	34
	Sum	34	34	68
Over-sampling	Censored	254	244	498
	Uncensored	242	256	498
	Sum	496	500	996
Both-sampling	Censored	275	245	520
	Uncensored	248	232	480
	Sum	523	477	1000
SMOTE	Censored	28	40	68
	Uncensored	33	35	68
	Sum	61	75	136

From these tables and graphs, class balance is evident in the overall data as well as in the covariates. This gives almost an equal representation of the data classes. More importantly, the mortality class does not suffer lack of information and bias.

4.2.2 Results of Variable Selection using RSF after Balancing with Different Balancing Methods

The balanced datasets were then analyzed using RSF algorithm for variable selection. The results of application of the RSF algorithm using balanced data are given in the table 4.4.The graphs of the OOB error rates are also given in figures 4.1, 4.2, 4.3 and 4.4

Table 4.4: Application of RSF in Balanced datasets.

Description	Under-sampling	Over-sampling	Both-sampling	SMOTE
Sample size	68	996	1000	136
No. of deaths	34	498	480	68
Number of trees	1000	1000	1000	1000
Forest terminal node size	15	15	15	15
Average no. of terminal nodes	2.49	20.294	20.117	5.232
No. of variables tried at each split	28	28	28	28
Total no. of variables	757	757	757	757
Resample size used to grow trees	43	629	632	86
No. of random split points	10	10	10	10
Error rate	13.27%	7.33%	7.69%	9.12%

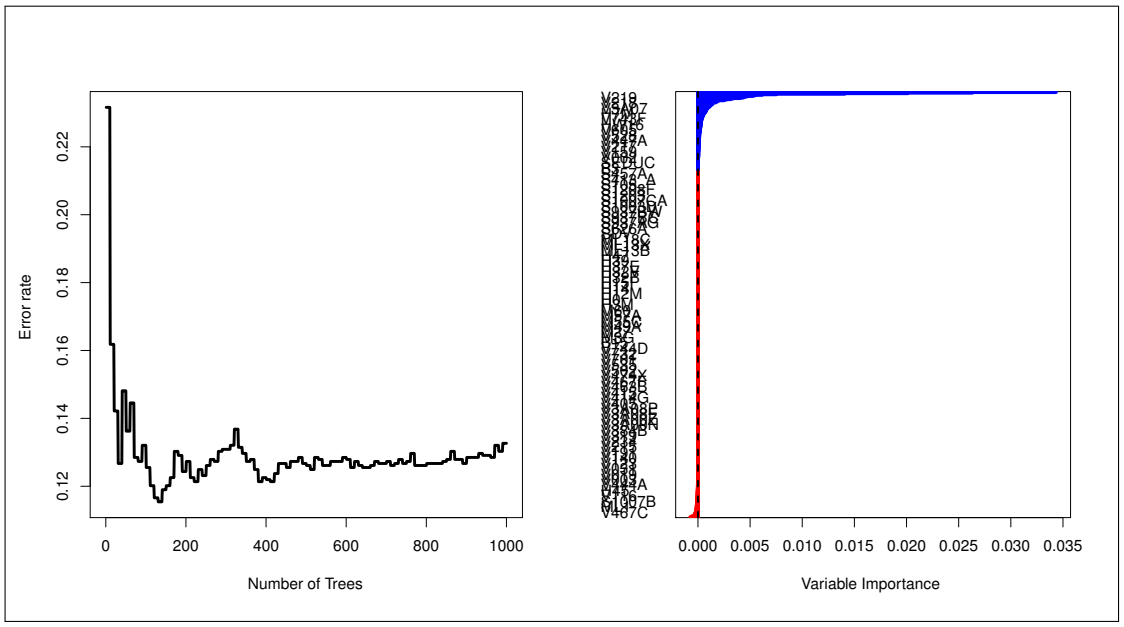

Figure 4.1: Under-sampling BRSF Nairobi error rate

For each of the four balanced dataset, a forest of 1000 trees was grown. This was done by drawing 1000 bootstrap samples from the respective initial data with the sample sizes given in table 4.4. The size of each bootstrap sample drawn is given as re-sample size used to grow trees in table 4.4. The bootstrap samples are of different sizes depending on the sample size of the initial data and the balancing method used. Each of the 1000 bootstrap samples is designated to the root of the tree. To develop each tree, 28 out of the 757 possible predictors were selected at random for splitting. The root node is then split into two daughter nodes each of which is recursively split progressively maximizing survival difference between daughter nodes. Node splitting continues until each tree if fully grown. This is

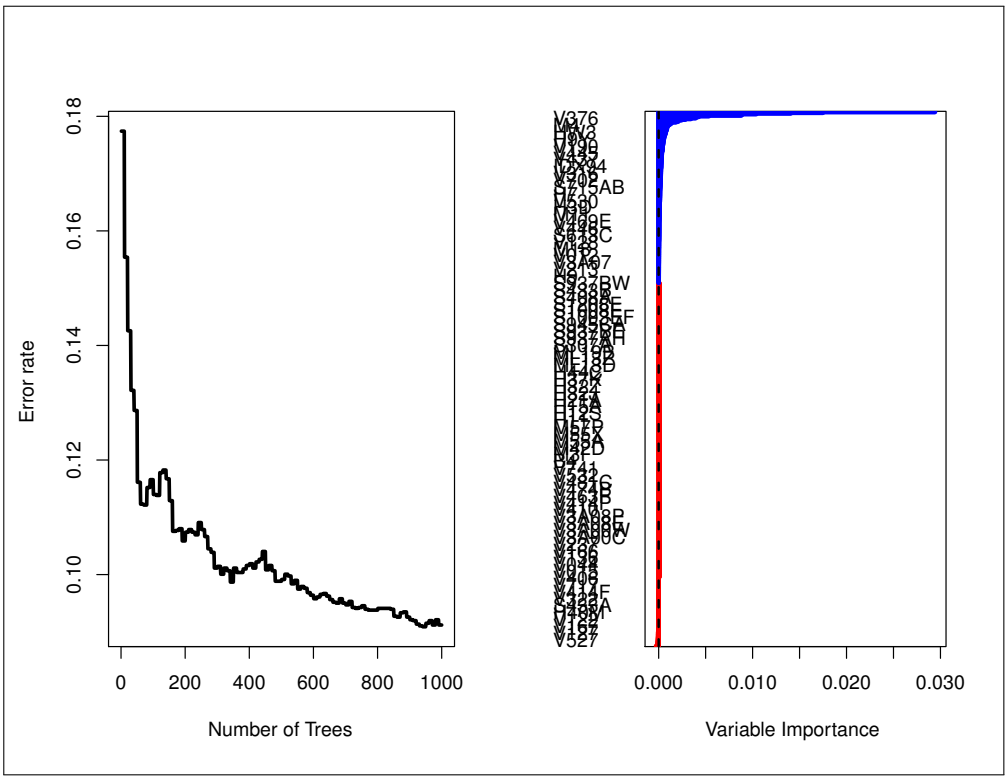

Figure 4.2: SMOTE BRSF Nairobi error rate

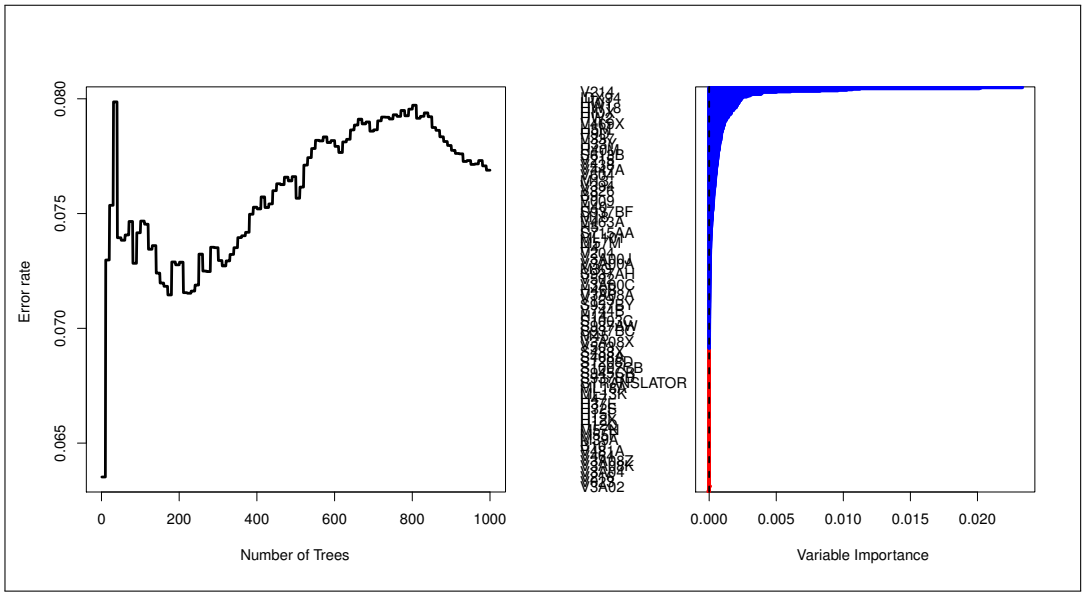

Figure 4.3: Both-sampling BRSF Nairobi error rate

achieved when the most extreme node has no fewer than 15 unique events. This implies that the samples with bigger number of events will form bigger trees. Hence, the more the number of events, the bigger the average number of terminal nodes and the smaller is the error rate. Over-sampling method with the biggest number

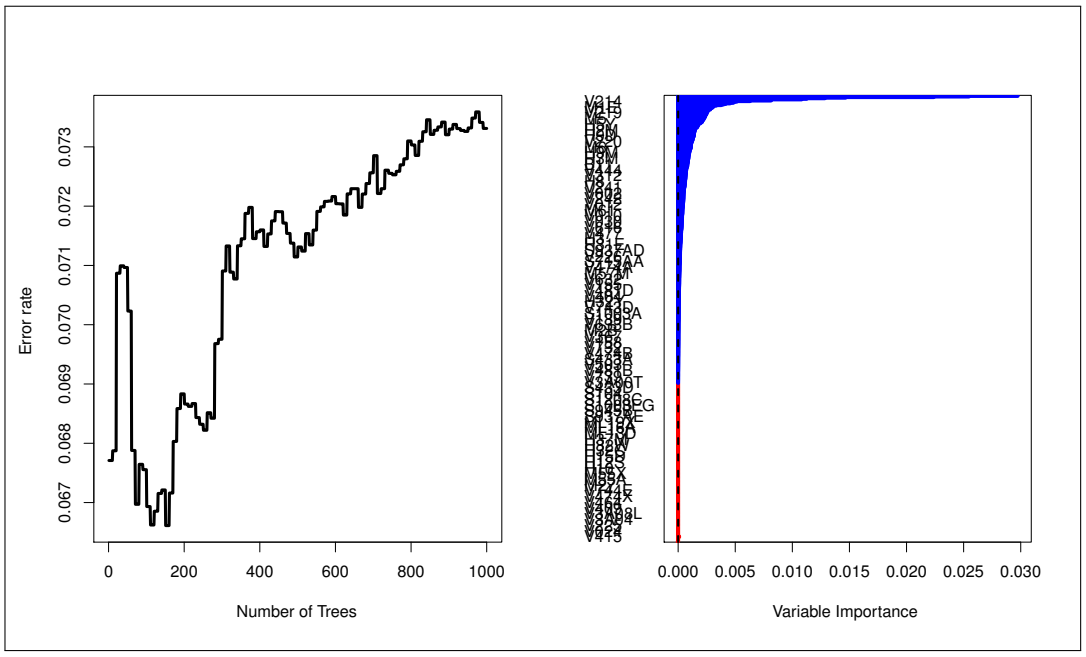

Figure 4.4: Oversampling BRSF Nairobi error rate

of events had the biggest average number of terminal nodes and smallest error rate while under-sampling method with the smallest number of events had the smallest average number of terminal nodes and highest error rate. Even though the sample sizes are different, the number of variables in the four samples is the same. This explains why the number of variables tried at each split and the numbers of random split points are equal in the four samples.

The selected variables based on balanced random survival forest ($BRSF$) using the different balancing methods are presented in table 4.5. From the Random Survival Forests, function $vimp()$ extracts variable importance (VIMP) information Ishwaran (2007). Using VIMP, we were able to identify the importance of the various variables in the different datasets as shown in table 4.5.

The $KDHS$ dataset has a total of 1099 variables that are possible candidates for predicting child mortality. After some data management exercise, the number of candidate covariates reduced to 757 possible covariates. Before fitting a regression type model in order to embark on the exercise of determining child mortality predictors, we needed to do a variable selection exercise in order to further reduce the

Table 4.5: Important Variables using Different Balancing Methods.

	Balancing method							
	Undersampling		Over sampling		Both sampling		SMOTE	
	Var	Importance	Var	Importance	Var	Importance	Var	Importance
1	B7	0.0263	B7	0.0255	B7	0.0201	B7	0.0294
2	B12	0.0141	HW70	0.0147	HW70	0.0112	HW73	0.0104
3	HW70	0.0075	HW71	0.0129	HW72	0.0107	HW72	0.0089
4	HW72	0.0062	HW72	0.0113	HW71	0.0102	HW71	0.0088
5	HW73	0.0055	HW73	0.0108	HW73	0.0096	HW70	0.0053
6	B8	0.0053	V206	0.0082	B12	0.0083	V206	0.0043
7	B16	0.0047	B12	0.0081	V206	0.0061	B12	0.0040
8	V219	0.0045	V214	0.0063	V214	0.0050	V376	0.0039
9	HW71	0.0044	B8	0.0051	V231	0.0040	V214	0.0034
10	V206	0.0041	B16	0.0048	V207	0.0039	M19A	0.0021
11	V220	0.0033	M19A	0.0045	B16	0.0039	HW7	0.0014
12	M1E	0.0031	V376	0.0042	B8	0.0039		
13	H3Y	0.0027	B6	0.0040	M19A	0.0032		
14	H9Y	0.0025	HW12	0.0038	M1E	0.0031		
15	V207	0.0020	HW7	0.0033	B6	0.0030		
16			HW3	0.0031	HW11	0.0029		
17			M1E	0.0031	V218	0.0028		
18			HW11	0.0030	V417	0.0025		
19			V218	0.0030	V219	0.0024		
20			H4M	0.0028	HW6	0.0024		
21			HW5	0.0026	B1	0.0024		
22			V207	0.0026	V506	0.0023		
23			HW10	0.0026	V3A07	0.0023		
24			V418	0.0026	V220	0.0022		
25			V219	0.0025	HW10	0.0022		
26			HW9	0.0025	HW1	0.0022		
27			HW6	0.0025	H4M	0.0022		
28			V419	0.0025	HW5	0.0021		
29			V506	0.0024	HW9	0.0021		
30			V231	0.0024	V478	0.0021		
31			HW18	0.0024	M4	0.0021		
32			V230	0.0023	V469F	0.0020		
33			V3A07	0.0023	V208	0.0020		
34			M5	0.0023	V376	0.0020		
35			V417	0.0022	HW18	0.0020		
36			M4	0.0022				
37			HW4	0.0022				
38			HW1	0.0021				
39			HW8	0.0020				

variables of importance. A combination of data cleaning and Random Survival Forest technique resulted into a reduced set of utmost 39 covariates for the regression steps that shall follow.

The extracted variables are shown in table 4.5.They are arranged in ascending order of importance in the respective datasets. The bigger the importance value,

the higher the predictive ability of the variable. Variables with $VIMP$ exceeding 0.002 were considered predictive. From table 4.5, over-sampling and both-sampling methods which had the highest number of events (498 and 480 events respectively), extracted the highest number of important predictors (39 and 35 predictors respectively). On the other hand, under-sampling and $SMOTE$ models with fewer number of events (34 and 68 events respectively) extracted the fewer number of predictors (15 and 11 predictors respectively).

4.2.3 Determination of Variable Effects

The effect of the selected variables on child mortality was measured by fitting the variables in Cox PH model which assess simultaneously the effect of several risk factors on survival time. One of the requirements of the cox model is that the covariate effect is proportional over time. Therefore, before the predictors are fitted in the Cox model, PH assumptions were tested. To validate PH assumption, Kaplan Meir curves and schoenfeld residuals were used and the results are shown in figures E.29 and E.17. Since schoenfeld residuals are independent of time, violation of the PH assumption is suspected when the Schoenfeld residual plot presents a relationship with time. However a Schoenfeld residual may fail to guarantee sufficient evidence due to lack of statistical hypothesis testing process Junyong and Dong (2019). Statistical tests were therefore conducted to validate the assumptions.

Testing Cox Proportional Hazards (PH) Assumptions.

Table 4.6 displays the results of testing proportional hazards assumption for the different important variables from the four different balancing methods. The global test gives a general picture of proportional hazards violations among the variables in the model. A p.value < 0.05 in the global test suggests one or more violations. From table 4.6, the global p.value from all the models had a $p.value < 0.05$ showing

statistical significance. This is an implication that all models had one or more variables violating the PH assumption. Similarly, a good number of variables from each model had a p value less than 0.05. In under-sampling model for example, 3 variables violated the assumption while in both-sampling method, almost all the variables (17) violated the assumption as shown in the table 4.6.

Table 4.6: Statistical tests.

Under-sampling				SMOTE			
Variable	rho	chisquare	P-value	Variable	rho	chisquare	P-value
V206	0.1953	1.0421	0.307	B7	0.6167	80.1502	$3.47e-19$
V207	-0.1144	0.2972	0.586	B12	-0.3528	12.6505	0.0004
V219	-0.1862	1.0546	0.304	HW70	0.1312	2.8738	0.0900
B8	-0.1492	0.2627	0.608	HW71	-0.2377	8.1717	0.00426
B12	-0.3266	6.3360	0.0118	HW72	0.2241	8.1670	0.00427
HW70	-0.0795	0.4568	0.499	HW73	-0.0886	1.3810	0.240
HW71	-0.1491	1.2483	0.264	V206	-0.0157	0.0247	0.875
HW72	0.2499	3.4646	0.0627	V214	-0.3799	19.0250	$1.29e-05$
HW73	0.0462	0.1168	0.733	Global	NA	96.4291	$2.29e-17$
M1E	0.2448	3.8905	0.0486				
H9Y	0.0106	0.0019	0.965				
H3Y	0.0791	0.1105	0.740				
B7	0.5807	25.0266	$5.65e^{-07}$				
Global	NA	43.9004	$3.19e^{-05}$				
Over-sampling				Both-sampling			
Variable	rho	chisquare	P-value	Variable	rho	chisquare	P-value
B7	0.6785	507	$2.69e^{-112}$	B1	-0.1924	20.176	$7.06e^{-06}$
B8	-0.2200	7.31	0.00686	B7	0.6318	561.967	$3.14e^{-124}$
B12	-0.3444	85.8	$1.98e^{-20}$	B8	-0.3470	25.269	$4.99e^{-07}$
V206	-0.0745	1.57	0.210	B12	-0.3457	75.486	$3.68e^{-18}$
V207	-0.2781	24.6	$7.04e^{-07}$	HW70	0.1326	19.143	$1.21e^{-05}$
V214	-0.3022	5.44	$1.60e^{-13}$	HW71	-0.2198	49.115	$2.41e^{-12}$
V218	0.0106	0.0689	0.7931	HW72	0.1954	33.532	$7.01e^{-09}$
V219	-0.0654	2.56	0.110	HW73	-0.1342	14.081	$1.75e^{-04}$
V230	0.0470	0.992	0.319	V206	-0.1316	8.118	0.00438
V417	0.0004	$8.48e^{-05}$	0.993	V207	-0.2350	18.797	$1.45e^{-05}$
HW70	0.0702	4.95	0.0262	V208	0.0979	6.295	0.0121
HW71	-0.2362	54.0	$1.96e^{-13}$	V214	-0.2392	31.429	$2.07e^{-08}$
HW72	0.2059	38.6	$5.22e^{-10}$	V218	-0.1359	16.527	$4.80e^{-05}$
HW73	-0.0128	0.138	0.710	V219	0.0932	7.072	$7.83e^{-03}$
HW1	0.1556	3.97	0.0463	V478	-0.0517	1.453	0.228
HW18	-0.2073	3.77	0.0521	V506	-0.2090	19.400	$1.06e^{-05}$
H4M	-0.0435	0.116	0.733	HW1	0.2593	15.270	$9.32e^{-05}$
M1E	0.2487	43.7	$3.86e^{-11}$	HW18	-0.0572	0.447	0.504
Global	NA	698	$1.55e^{-136}$	M1E	0.2671	42.597	$6.72e^{-11}$
				Global	NA	731.214	$8.67e^{-143}$

For variables that did not satisfy the PH assumption, interaction with time, functions of time or time varying covariate was included. Variables that finally did

not satisfy the assumption were deleted from the model. After the exercise, the variables that remained for use in Cox PH model for prediction are shown in table 4.7.

Table 4.7: Statistical tests after removal and interaction of violating variables.

Undersampling				SMOTE			
Variable	rho	chisquare	P-value	Variable	rho	chisquare	P-value
V206	0.1376	0.6315	0.427	B12	-0.2063	4.2168	0.0400
V207	0.1554	0.4991	0.480	HW70	0.1173	2.3004	0.1293
V219	-0.2325	1.8569	0.173	HW71	0.1272	2.2532	0.1333
B8	-0.1602	1.2131	0.271	HW72	0.0207	0.0560	0.8129
$\log(B12+1)$	-0.1841	1.9335	0.164	HW73	-0.0172	0.0508	0.8217
HW70	-0.0044	0.0015	0.969	V206	0.0378	0.1455	0.7029
HW71	-0.1139	0.7760	0.378	V214	-0.1854	3.4545	0.0631
HW72	0.0476	0.1319	0.716	Global	NA	14.7385	0.0395
HW73	0.0799	0.4599	0.498				
H9Y	-0.0864	0.1506	0.698				
H3Y	0.0896	0.1105	0.678				
B8:M1E	0.1582	1.1827	.0277				
Global	NA	12.5046	0.406				
Oversampling				Both sampling			
Variable	rho	chisquare	P-value	Variable	rho	chisquare	P-value
HW72	-0.1667	0.9343	0.334	V206	0.1428	7.085	0.0078
H4M	-0.0355	0.0474	0.828	V207	0.2041	9.834	0.0017
B1:V206	-0.0248	0.2277	0.633	V208	-0.0747	2.648	0.1037
Global	NA	1.0825	0.781	B1:HW70	0.0939	0.938	0.3328
				Global	NA	13.268	0.0099

Parameter Estimates

From the previous section, we note that the different balancing methods yielded different sample sizes and different predictors from the RSF classification. After diagnostic tests on Cox PH models, the respective predictors were fitted to the parsimonious Cox PH model in order to check concurrently the effect of different risk factors on survival time. The results of fitting variables in Cox PH model are given in table 4.8

From table 4.8, the regression coefficient column marked "Coef" gives estimates of the logarithm of the hazard ratio between the two groups. From the estimates, a positive coefficient is said to increase the risk of death (hazard) and thus decrease

Table 4.8: BRSF Cox ph model predictors.

Undersampling					SMOTE				
Var	coef	E(coef)	Se(coef)	P-value	Var	coef	E(coef)	Se(coef)	P-value
V206	2.1030	8.187	0.4348	$1.3e^{-6}$	B12	-0.1178	0.8889	0.0234	$4.6e^{-7}$
V207	1.4770	4.378	0.4120	0.0003	HW70	0.0017	1.0017	0.0011	0.1019
V219	-0.2018	0.8173	0.2246	0.3688	HW71	-0.0002	0.9998	0.0010	0.8179
B8	-16.34	$8.0e^{-8}$	14.70	0.2663	HW72	0.0026	1.0026	0.0008	0.0007
log(B12)	-1.198	0.3017	0.6671	0.0725	HW73	-0.0039	0.9960	0.0010	$9.8e^{-5}$
HW70	$-1.3e^{-5}$	1.000	0.0014	0.9926	V206	0.08385	2.3129	0.2877	0.0036
HW71	-0.0003	0.9997	0.0012	0.7917	V214	-1.3991	0.2466	0.2696	$2.1e^{-7}$
HW72	0.0022	1.002	0.0011	0.0383					
HW73	-0.0016	0.9984	0.0011	0.1521					
H9Y	-0.209	0.8109	0.8245	0.7992					
H3Y	-0.359	0.6983	0.7951	0.6515					
B8:M1E	0.0114	1.011	0.0109	0.2971					
Oversampling					Both sampling				
Var	coef	E(coef)	Se(coef)	P-value	Var	coef	E(coef)	Se(coef)	P-value
HW72	0.0001	1.000	$2.2e^{-5}$	$4.4e^{-9}$	V206	1.675	5.338	0.0834	$< 2e^{-16}$
H4M	0.0244	1.025	0.0212	0.25	V207	1.439	4.251	0.0725	$< 2e^{-16}$
B1:V206	0.1864	1.205	0.0116	$< 2e^{-16}$	V208	0.4126	1.5111	0.0799	$2.4e^{-16}$
					B1:HW71	$1.5e^{-5}$	1.000	$4.9e^{-6}$	0.0017

the expected (average) survival time. On the other hand, a negative coefficient reduces the risk of death and thus raises the expected survival span. In explaining the determinants of child mortality, one therefore is interested in the variables with positive coefficient, which are positively related with the event (mortality) probability, and consequently negatively related with the length of survival. From table 4.8, under-sampling method resulted in 4 predictors, which are likely to increase the risk of death. Similarly, SMOTE returned 3 predictors that are likely to increase the risk of death. Over-sampling and both-sampling method had 3 and 4 predictors respectively all of which had positive coefficients.

Its often useful for interpretation to look at the "E(coef)" column, which indicates the actual hazard ratio (HR) associated with the covariates. A value of regression coefficient greater than zero is equivalent to a hazard ratio greater than one, which shows that as the value of the i^{th} predictor increases (for continuous type covariates), the event hazard increases and thus the length of survival decreases.

From table 4.8 for example, variable V206, in under-sampling method resulted in $(coefficient) = exp(2.1030) = 8.187$. The HR value which is clearly greater than 1 implies that variable V206 increases the hazard by a factor 8.187. This is deduced from the fact that a predictor is related with increased risk when the value

of $HR > 1$, and decreased risk when $HR < 1$. When the HR value is close to 1, the predictor has no impact on survival. From our results, there are 3 predictors in under-sampling method associated with increased risk. These are V206, V207 and HW72. In a similar way,3,4 and 3 predictors in over-sampling, both sampling and SMOTE respectively are associated with increased risk. (refer to Table 4.8). The column marked $p - value$ gives the value of the Wald statistic. Wald statistic evaluates whether the explanatory variables in a model are significant. A variable is said to be statistically significant when its p.value is less than 0.05.

From the predictors got from the Cox model, we are interested in the relationship between death which is the event of interest and the resulting predictors $x_1, x_2, ..., x_n$.

The resulting specific Cox model for the respective groups are;

$$\begin{aligned} H_x(t) &= h_0(t)exp(2.1030 * V206 + 1.4770 * V207 - 0.2018 * V219 - 16.34 * B8 \\ &\quad -1.198 * log(B12)... + 0.0114 * B8 : M1E \end{aligned}$$

for under-sampling model and the same case applies to the other models. It is important to note that quite a number of variables did not go through the prediction stage since they did not satisfy the PH assumptions.

4.2.4 Concordance Measure of Model Fit

The concordance statistic was used to analyze the performance of the models on prediction of mortality.

High values of concordance indicate that for higher observed survival duration, the model predicts higher probabilities of survival. Concordance values ranges from 0 to 1. A perfect Concordance results in a value of 1 while 0.5 is as good as random guessing. All our models gave high concordance values above 0.7 with standard errors less than 0.02. Hence all the models represent a good fit according to the

Table 4.9: Model fit statistics: Concordance measure.

Description	Under-sampling	Over-sampling	Both-sampling	SMOTE
Sample size	68	996	1000	136
Concordance	0.9048	0.781	0.8768	0.8596
Standard error	0.0269	0.0121	0.0086	0.0222
Discordant	1378	248084	277577	5492
Concordant	145	69549	38998	897
Tied.x	0	0	0	0
Tied.y	158	33849	30686	329
Tied.xy	0	3621	3438	1

concordance Index. Under-sampling method gives the largest concordance value of 0.90 indicating the best model fit while over-sampling had the smallest concordance value of 0.78. SMOTE and both-sampling methods have concordance values of 0.86 and 0.87 respectively. According to these results, under-sampling model gave the best fit.

4.3 Results for Analysis of BRSF using Different Splitting Methods.

In this section, we present the results of analysing data balanced using under-sampling method while growing trees using different splitting methods. The dataset was explored and found to have variables that violated the PH assumption.

4.3.1 Results for Application of BRSF in Different Splitting Rules.

Random survival forest was used to analyse this data set. Using the randomForestSRC package, the following results were obtained. Table 4.10 shows the results of applying the RSF algorithm in the dataset while figures 4.5, 4.6 and 4.7 show the OOB error rates.

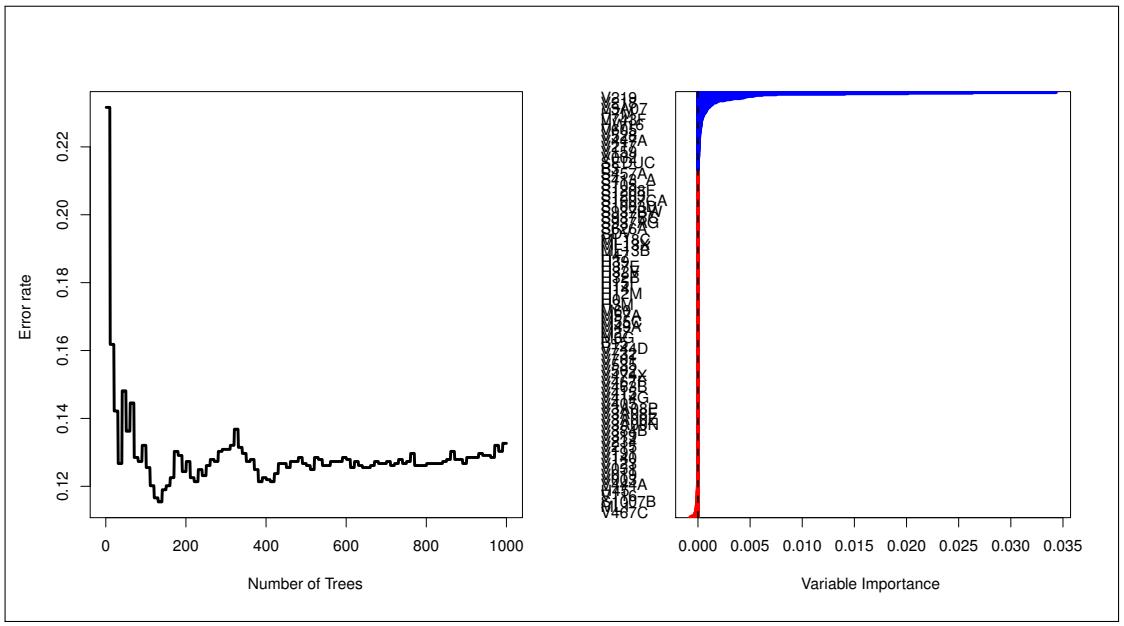

Figure 4.5: Under-sampling BRSF Nairobi log-rank error rate

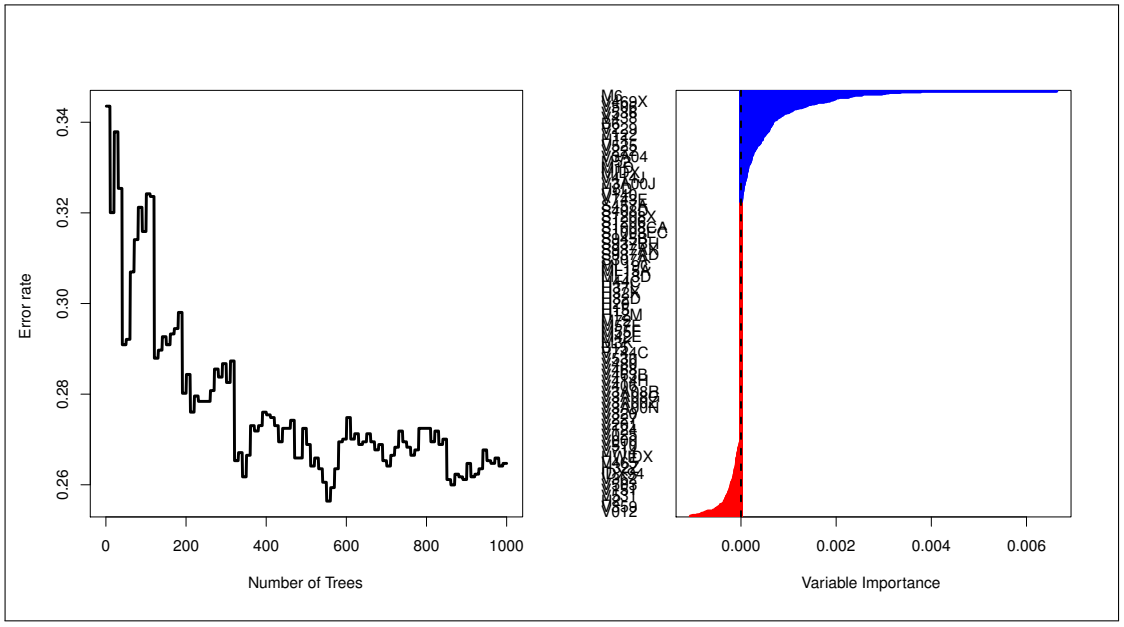

Figure 4.6: Under-sampling BRSF Nairobi log-rank score error rate

From the RSF output, a forest of 1000 trees was generated. This was done by randomly selecting 1000 bootstrap samples from the initial dataset. These bootstrap samples are designated to the root of the trees in the respective models. The dataset consisted of 757 variables and 68 observations. Out of the 68 observations, 34 were censored and 34 had acquired an event. According to the results in table 4.10, most of the output in various descriptions is the same since we used the same balanced

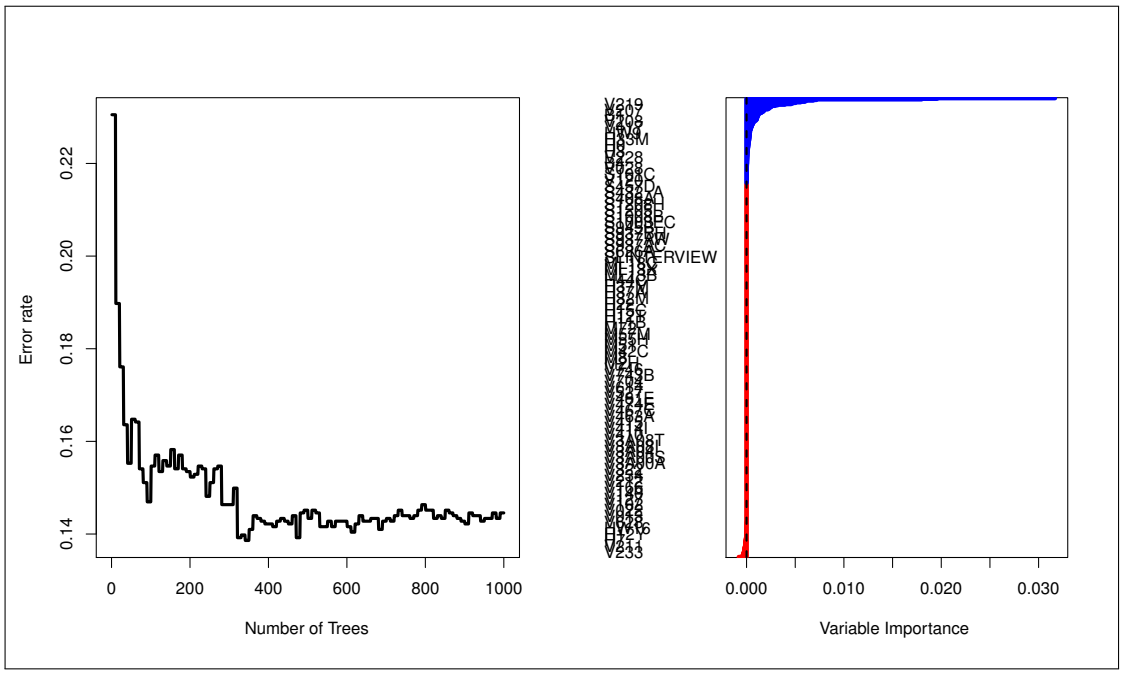

Figure 4.7: Under-sampling BRSF Nairobi Bs.gradient error rate

Table 4.10: Application of RSF in different splitting rules.

Description	Logrank	Logrank score	Bs.gradient
Sample size	68	68	68
No. of deaths	34	34	34
Number of trees	1000	1000	1000
Forest terminal node size	15	15	15
Average no. of terminal nodes	2.49	3.367	2.488
No. of variables tried at each split	28	28	28
Total no. of variables	757	757	757
Resample size used to grow trees	43	43	43
No. of random split points	10	10	10
Error rate	13.27%	26.47%	14.46%

dataset with equal number of covariates. The only difference observed is in the error rate and average number of terminal nodes. This difference is most likely brought about by the different splitting rules used. Log-rank splitting rule splits the nodes with greater accuracy returning an OOB error rate of 13.27% followed by Bs.gradient rule while Log-rank score yielded the highest error rate of 26.47%. Each tree is considered fully grown when each terminal node has no fewer than 15 unique events. From the average number of terminal nodes which were generated, Bs.gradient produced the least followed by log-rank test while log-rank score had the most number of terminal nodes. The selected variables based on BRSF are presented

in table 4.11.

Important Variables using Different Splitting Rule.

Table 4.11: Important Variables using Different Splitting Rule.

	Split rule					
	Log-rank		Log-rank score		BS.gradient	
	Variable	Importance	Variable	Importance	Variable	Importance
1	B7	0.0263	HW71	0.0038	B7	0.0180
2	B12	0.0141	B16	0.0033	HW73	0.0072
3	HW70	0.0075	H5Y	0.0031	HW72	0.0069
4	HW72	0.0062	HW70	0.0030	B12	0.0061
5	HW73	0.0055	B12	0.0026	HW70	0.0060
6	B8	0.0053	H6Y	0.0025	B8	0.0055
7	B16	0.0047	H2Y	0.0024	HW71	0.0050
8	V219	0.0045	M6	0.0023	V219	0.0047
9	HW71	0.0044	H4Y	0.0023	B16	0.0046
10	V206	0.0041	H3Y	0.0023	V506	0.0039
11	V220	0.0033	H8Y	0.0020	H9Y	0.0032
12	M1E	0.0031	ML1	0.0020	H7Y	0.0029
13	H3Y	0.0027	V626	0.0020	V321	0.0026
14	H9Y	0.0025	V218	0.0020	H5Y	0.0025
15	V207	0.0020			V220	0.0024
16					V3A07	0.0023
17					V214	0.0023
18					V207	0.0022

Table 4.11 shows the result of variable selection using the same dataset with different splitting rules. Variables with importance value above 0.002 are considered predictive according to Ishwaran et al. (2008). The different splitting rules led to extraction of different number of predictive variables. 18 variables were selected as predictive for U5CM based on Bs.gradient methods, 14 based on log-rank score while log-rank splitting rule resulted to 15 predictive variables. It is also seen that most of the extracted predictors using log-rank and Bs.gradient methods are similar. However, the importance measure of these variables are slightly different in the models.

4.3.2 Parameter Estimates

The extracted important variables were initially fitted to a Cox regression model and results presented in table 4.12.

Table 4.12: BRSF Cox ph model predictors for different splitting rules .

Logrank					Logrank score				
Var	coef	E(coef)	Se(coef)	P-value	Var	coef	E(coef)	Se(coef)	P-value
B7	-0.1561	0.8554	0.0589	0.0081	B12	-0.1016	0.9033	0.0276	0.0002
B8	-1.1113	0.3291	0.5771	0.0541	HW70	-0.0012	0.9988	0.0011	0.2835
B12	-0.0161	0.9840	0.0348	0.6450	HW71	0.0012	1.0012	0.0011	0.3034
HW70	-0.0010	0.9989	0.0013	0.4641	V218	-0.6343	0.5303	0.2218	0.0042
HW71	0.0008	1.0008	0.0012	0.4869	H2Y	2.2652	9.6328	1.1337	0.0457
HW72	0.0019	1.0019	0.0010	00517	H3Y	-0.2564	0.7738	1.8227	0.8881
HW73	-0.0015	0.9985	0.0009	0.1284	H4Y	-1.2277	0.2929	1.6461	0.4558
V206	2.0731	7.9490	0.4044	$2.95e^{-7}$	H5Y	1.7696	5.8682	1.6285	0.2772
V207	1.6683	5.3034	0.4317	0.0001	H6Y	-3.3369	0.0355	1.9417	0.0857
V219	-0.2936	0.7456	0.2184	0.1789	H8Y	1.5015	4.4884	1.0579	0.1558
M1E	0.0164	1.0165	0.0280	0.5584	ML1	0.5677	1.7642	0.2723	0.0371
H3Y	-0.2797	0.7560	0.7231	0.6989					
H9Y	-0.2311	0.7937	0.7495	0.7579					
Bs.gradient									
Var	coef	E(coef)	Se(coef)	P-value					
B7	-0.1124	0.8937	0.0514	0.0287					
B8	-0.7243	0.4846	0.5041	0.1508					
B12	-0.0281	0.9723	0.0332	0.3968					
HW70	-0.0014	0.9985	0.0012	0.2353					
HW71	0.0013	1.0013	0.0013	0.3291					
HW72	0.0011	1.0011	0.0011	0.3410					
HW73	-0.0008	0.9992	0.0009	0.4121					
V207	0.6904	1.9945	0.3385	0.0414					
V214	-0.4302	0.6504	0.2736	0.1158					
V219	-0.4122	0.6622	0.2647	0.1194					
V321	-0.0567	0.9449	0.0362	0.1171					
H5Y	1.4966	4.4664	1.6199	0.3555					
H7Y	-1.8359	0.1595	1.6952	0.2788					
H9Y	0.0634	1.0655	0.5892	0.9143					

However, proportionality of covariate effect is not satisfied for some of the variables according to statistical tests as shown in table 4.13. The global test gives a general picture of proportional hazards violations among the variables in the model. A p-value < 0.05 suggests one or more violations. According to table 4.13, the global test in log-rank and Bs.gradient models indicates violation of PH assumption (p-value= $3.19e^{-5}$ and 0.0001). The variables that violated the PH assumption in these models includes B7, B12 in both logrank and Bs.gradient models and M1E in logrank model.

Since the assumption that the relative risks are constant over time did not hold for all the variables, analyses that take into account time-varying effects is required.

Table 4.13: Statistical tests.

Logrank				Logrank score			
Variable	rho	chisquare	P-value	Variable	rho	chisquare	P-value
B7	0.5807	25.03	$5.65e-07$	B12	-0.2683	2.72	0.0991
B8	-0.1492	0.2627	0.608	HW70	0.0587	0.297	0.5856
B12	-0.3266	6.3360	0.0118	HW71	-0.0643	0.352	0.5529
HW70	-0.0795	0.4568	0.499	V218	-0.2177	3.19	0.0743
HW71	-0.1491	1.2483	0.264	H2Y	0.1192	0.292	0.589
HW72	0.2499	3.4646	0.0627	H3Y	0.0841	0.329	0.5660
HW73	0.0462	0.1168	0.733	H4Y	-0.0339	0.043	0.8358
V206	0.1953	1.0421	0.307	H5Y	0.1198	0.518	0.4717
V207	-0.1144	0.2972	0.586	H6Yl	-0.2206	1.76	0.1841
V219	-0.1862	1.0546	0.304	H8Y	0.1929	0.581	0.4460
M1E	0.2448	3.8905	0.0486	ML1	-0.0004	$2.96e^{-6}$	0.9986
H3Y	0.0791	0.1105	0.740	Global	NA	11.4	0.4130
H9Y	0.0106	0.0019	0.965				
Global	NA	43.9004	$3.19e^{-05}$				

Bs.gradient			
Variable	rho	chisquare	P-value
B7	0.6730	28.78	$8.12e^{-8}$
B8	0.0378	0.0182	0.893
B12	-0.3653	7.0356	0.00799
HW70	-0.0074	0.0037	0.952
HW71	-0.0531	0.2170	0.641
HW72	0.0512	0.1794	0.672
HW73	0.0108	0.0064	0.936
V207	0.0959	0.3536	0.552
V214	-0.0982	0.3273	0.567
V219	-0.1198	0.5868	0.444
V321	-0.059	0.1487	0.700
H5Y	0.0198	0.0136	0.907
H7Y	0.0530	0.0989	0.753
H9Y	-0.2158	0.7076	0.400
Global	NA	42.52	0.0001

In the previous section, variables that violated the PH assumption were interacted with time varying covariates while others were removed from the model before fitting in the cox PH model. To avoid removal of variables which could be predictors of mortality, we worked with Cox-Aalen's model in this section.

Cox-Aalen's model proposed by Scheike and Zhang (2002) is one of the tools for handling the problem of non-proportional effects in the Cox model. The model provides a simple way of including time-varying covariate effects. The extracted important variables were therefore fitted to a Cox-Aalen's model and results presented in table 4.14.

Table 4.14: Cox Aalen Model.

Logrank				BS.gradient			
Variable	coef	SE	P-value	Variable	coef	SE	P-value
V206	1.8400	0.2990	$3.47e-10$	V207	0.6750	0.2820	0.0129
V207	1.6000	0.3020	$3.81e-07$	V214	-0.2250	0.1680	0.2840
V219	-0.3400	0.1890	0.0578	V219	-0.5000	0.2360	0.0208
B7	-0.1490	0.0884	0.0482	B7	-0.1170	0.0973	0.1410
B8	-0.7300	0.2320	0.0022	B8	-0.4480	0.1340	0.0170
B12	-0.0567	0.0393	0.0725	B12	-0.0519	0.0404	0.1150
M1E	0.0145	0.0313	0.503	HW70	0.0001	0.0015	0.954
HW70	-0.0004	0.0016	0.771	HW71	0.0019	0.0015	0.129
HW71	0.0019	0.0017	0.180	HW72	-0.0011	0.0025	0.589
HW72	-0.0003	0.0028	0.902	HW73	-0.0008	0.0014	0.535
HW73	-0.0012	0.0015	0.350				
Logrank score							
Variable	coef	SE	P-value				
V218	-0.7100	0.2890	0.0040				
B12	-0.0874	0.0252	$2.37e-05$				
ML1	0.5330	0.1640	0.00821				
HW70	-0.0013	0.0014	0.231				
HW71	0.0013	0.0014	0.200				

From this table, variables that turned to be statistically significant with p value ≤ 0.05 include V206 (Total number of sons who have died), V207 (Total number of daughters who have died), B7 (Age at death of the child at completed months), and B8 (Current age of the child in single years for all living children) in the logrank mode, V207 (Total number of daughters who have died), V219 (Total number of living children including current pregnancy),and B8 (Current age of the child in single years for all living children) in Bs.gradient model and V218 (Total number of living children), B12 (Succeeding birth interval) and ML1 (Times the mother took SP/Fansidar during pregnancy) in the logrank score model.

4.3.3 Model Selection using Concordance Measure of Model Fit.

To compare the different models, concordance index was used in order to determine the effect of the various splitting methods and results shown in table 4.15.

According to the results in table 4.15, the three models resulted in good fit with

Table 4.15: Concordance measure in different splitting rules.

Description	Logrank	Logrank score	Bs.gradient
Sample size	68	68	68
Concordance	0.916	0.8536	0.8674
Standard error	0.0196	0.03306	0.02936
Discordant	1395	1300	1321
Concordant	128	223	202
Tied.x	0	0	0
Tied.y	158	158	158
Tied.xy	0	0	0

reference to the concordance index whereby all resulted in concordance above 0.79. The log-rank test resulted to the best model fit with concordance of 0.92 followed by Bs.gradient with a concordance of 0.86. Log-rank score had the lowest concordance among the three models. As indicated earlier, the optimality of log-rank is achieved when the model variables satisfy the PH assumptions. Hence in instances when the PH assumption is violated, Bs.gradient is the better option.

4.4 Results for Development of an IBRSF when PH Assumption is Violated.

In this section, we used the data balanced using under-sampling method. To arrive at determinants of U5CM, two stages after data balancing were followed which are variable selection and prediction stage. The results of variable selection and prediction using IBRSF method are given in this section.

4.4.1 Application of IBRSF

The results for application of IBRSF are given in the table 4.16. From the table, a sample of 68 observations was used from which half of them experienced the event while the other half survived. There were 757 variables in the dataset. From this dataset, a random forest of 1000 trees was grown as follows. 1000 bootstrap samples

were randomly selected from the dataset and assigned to the root of each tree to be grown. Each bootstrap sample had a total of 43 observations. From each bootstrap sample, 28 variables were randomly selected for splitting. Bs.gradient splitting rule was used to split the nodes. Using recursive partitioning of nodes, the trees were grown to full size where the most extreme node had no less than 15 observations. The average number of terminal nodes was 2.49 and the OOB error rate was 14.46%.

Table 4.16: Application of IBRSF in variable selection stage.

Description	
Splitting rule	Bs.gradient
Sample size	68
No. of deaths	34
Number of trees	1000
Forest terminal node size	15
Average no. of terminal nodes	2.49
No. of variables tried at each split	28
Total no. of variables	757
Resample size used to grow trees	43
No. of random split points	10
Error rate	14.46%

The graph for the OOB error rate is shown in figure 4.8

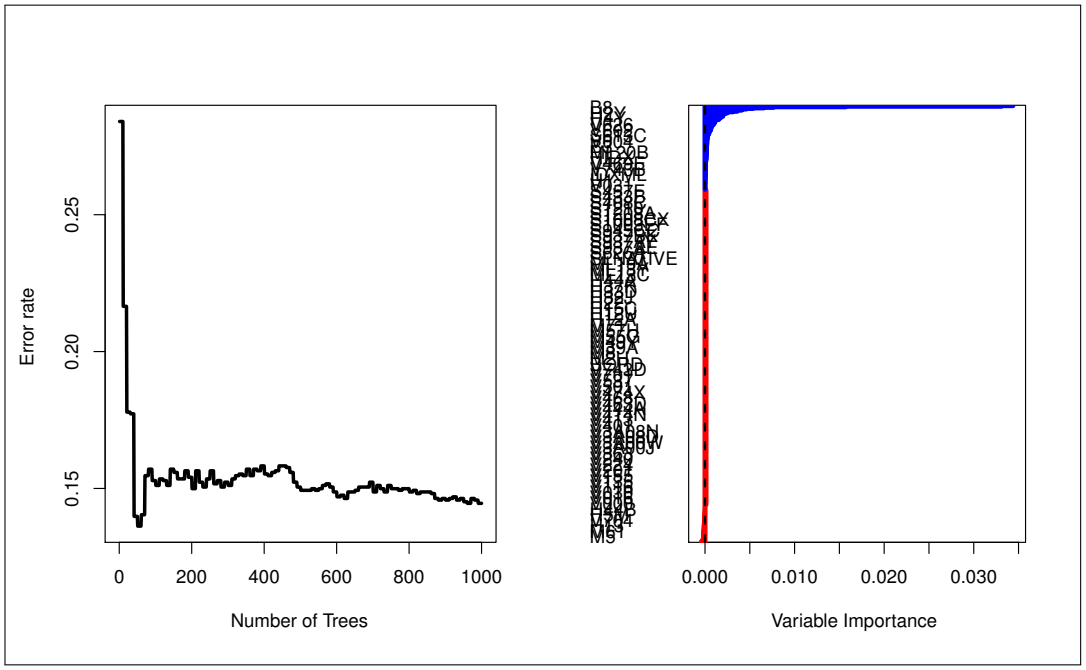

Figure 4.8: Under-sampling Vimp for IBRSF BS.gradient error rate

4.4.2 Variable Selection using VIMP

Table 4.17 gives the results for variables selected together with their importance measure. After performing variable selection using RSF VIMP, 18 variables out of a total of 757 variables were selected as shown in table 4.17. Only variables with

Table 4.17: IBRSF variable selection.

Variable	Description	Importance
B7	Age at death of child -completed months	0.0157
B8	Current age of child in single years	0.0077
HW70	Weight for age standard deviation	0.0066
HW72	Weight for height standard deviation	0.0054
HW73	BMI standard deviation	0.0046
B12	Succeeding birth interval	0.0042
H5Y	Oral Polio 2 year	0.0041
HW71	Age in months of the child	0.0036
H9Y	Measles 1 year	0.0036
V506	The rank of the respondent among the partner's wives	0.0031
H0Y	Oral Polio at birth year	0.0026
H2Y	BCG vaccination date - year	0.0025
V219	Total no.of living children plus current pregnancy	0.0023
V605	Desire for more children	0.0021
V218	Total number of living children	0.0020
V231	Century month code of the last pregnancy termination	0.0020
H3Y	DPT-HEP.B-HIB (PENTAVALENT	0.0020
B16	Child's line number in household	0.0020

importance measure greater ≥ 0.002 were selected for the final prediction stage.

4.4.3 Determinants of U5CM using IBRSF

The 18 selected variables were subjected to RSF VIMP a second time for identification of determinants of U5CM. Table 4.18 shows the application of RSF in the dataset.

From this table, the dataset used consisted of 68 observatitons and 18 variables. The dataset was balanced in that the number of events was equal to the number of survivors. A forest of 1000 trees was grown by randomly selecting 1000 bootstrap samples from the initial data. Each sample selected had 43 observations. From 18 variables, 5 variables were randomly selected for splitting on. Recursive partitioning

Table 4.18: Application of IBRSF in variable prediction stage.

Description	
Splitting rule	Bs.gradient
Sample size	68
No. of deaths	34
Number of trees	1000
Forest terminal node size	15
Average no. of terminal nodes	2.269
No. of variables tried at each split	5
Total no. of variables	18
Resample size used to grow trees	43
No. of random split points	10
Error rate	12.37%

of the nodes was done until the trees were fully grown. Each terminal node had no less than 15 observations. The average number of terminal nodes was 2.27 and the OOB error rate was 12.37%. Figure 4.9 shows the graph of error rate for the prediction model together with the predictor variables with their importance level.

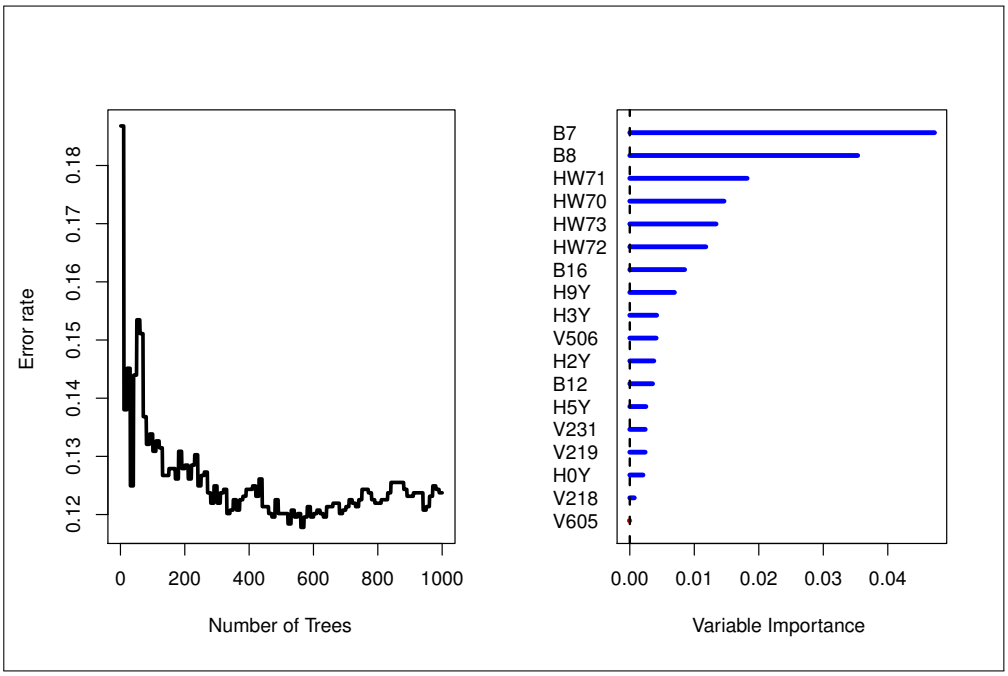

Figure 4.9: IBRSF prediction error rate

The result of variable prediction using VIMP is shown in table 4.19. Figure 4.10 shows the 95% confidence intervals for the predictive variables. For convenient interpretation as percentage, VIMP values have been multiplied by 100. The larger the positive VIMP value the higher the predictive ability of the variable. On the

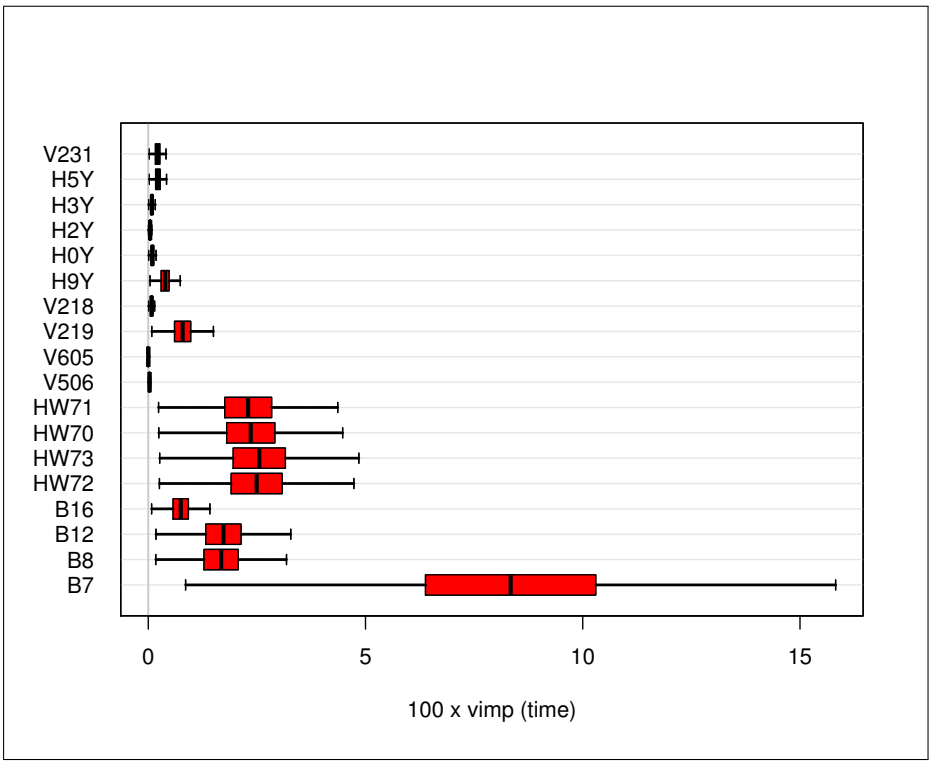

Figure 4.10: IBRSF variable prediction VIMP

other hand, negative values and zero indicate noise variables. We used subsumpling technique to approximate standard errors and confidence intervals for VIMP. Figure 4.10 shows the delete-d jacknife 95% asymptotic normal confidence intervals for the 18 selected variables.

The variable B7 (Age at death of the child in completed months) has the largest VIMP with confidence intervals well bounded away from zero. This gives us the highest predictor of U5CM. This is in line with the survival data where most of the children died before the end of the first month hence these results are not surprising. There are other variables with moderate VIMP size with their confidence intervals bounded away from zero. These are the interelations between age, height, weight and BMI of the children which are HW71(weight for age standard deviation), HW70(Height for weight standard deviation), HW73(BMI standard deviation) and HW72 (weight for height standard deviation). B12(succeeding birth interval) and B8(current age of child) also need to be considered as determinants of U5CM.

Table 4.19: IBRSF variable prediction.

Variable	Description	Importance
B7	Age at death of child -completed months	0.0472
B8	Current age of child in single years	0.0353
HW71	Age in months of the child	0.0182
HW70	Weight for age standard deviation	0.0146
HW73	BMI standard deviation	0.0134
HW72	Weight for height standard deviation	0.0118
B16	Child's line number in household	0.0085
H9Y	Measles 1 year	0.0069
H3Y	DPT-HEP.B-HIB (PENTAVALENT	0.0042
V506	The rank of the respondent among the partner's wives	0.0041
H2Y	BCG vaccination date - year	0.0037
B12	Succeeding birth interval	0.0035
H5Y	Oral Polio 2 year	0.0025
V231	Century month code of the last pregnancy termination	0.0024
V219	Total no.of living children plus current pregnancy	0.0024
H0Y	Oral Polio at birth year	0.0020

Chapter 5

DISCUSSION, CONCLUSION AND RECOMENDATIONS

5.1 Discussion

This study attempts to understand the determinants of under five mortality using survey data from DHS. In this case, Kenya DHS survey 2014 dataset was used for the analysis. The dataset (after variable cleaning) is composed of 757 variables that are candidate determinants of Under five Child mortality (U5CM). This poses a problem of variable selection from such high dimensional datasets preceding a proper analysis in which the intention is to explain variable effects. Besides, there is too much class imbalance in the datasets particularly where interest is to compare mortality and non mortality groups. For instance, 6.4% of children experience mortality while 93.6% survived up to the age of 5 years as shown in the 2014 KDHS data. This imbalance is too huge that a direct comparison (before balancing) between two such groups is likely to yield biased results. In addition, the commonly used methods in survival analysis are optimal when PH assumption is not violated which is not

always the case with many variables.

Three challenges were addressed in this study. One problem involved trying to balance the dataset classes before making comparisons between mortality and non mortality cases. The other challenge was due to variable selection. One needs to conduct a proper variable selection exercise in order to identify the correct set of variables to use for the regression analysis. And finally the use of an optimal splitting rule and prediction method in cases where not all variables satisfy the PH assumption.

Most studies explore determinants of child mortality using DHS survey data. Ayiko et al. (2009) used Uganda 1996, 2000, 2006 DHS dataset, Nasejje et al. (2015) used Uganda 2011 DHS, Sreeramareddy et al. (2013) analyzed the data from complete birth histories of four Nepal Demographic and Health Surveys (NDHS) done in the years 1996, 2001, 2006 and 2011. In this study, we have also tapped into the richness of KDHS (2014) dataset, to establish the determinants of U5CM. The key improvement over many studies that have used DHS data to answer the same question lies in our choice to ensure the following remedies are done:

1. Class imbalance is eliminated before comparisons are done.
2. Imputation for missing data is done using a machine learning approach (the missForest package in R software was used).
3. Variable selection is accomplished using a machine learning algorithm (RSF).
4. Splitting rule is done with a method that does not requires satisfaction of PH assumption.
5. Prediction using RSF VIMP which is a non-parametric ensemble method that does not require PH assumption to be satisfied.

In most studies, researchers often use self intuition or previous studies to determine which covariates to add to their regression models. In this research, we did variable

selection using RSF to select important variables from all the avalilable covariates in the dataset. All these remedies were done before moving to prediction stage to reduce chances of reporting biased findings.

Many studies commonly employed regression techniques to explore the determinants of U5CM. Cox PH regression was used by Ayiko et al. (2009), Nasejje et al. (2015), Sreeramareddy et al. (2013). Although we also used the Cox PH model, we preceded it with diagnostics including multiple imputation, classification balancing, variable selection, and Cox PH assumptions tests, to ensure that the results from the Cox PH are more reliable.

While using Cox PH model in prediction stage, it was realized that it is possible to have a variable that violates PH assumption and fail to include it in the model yet it is a determinant of U5CM. For instance, Variable B7 (Age at death of the child at completed months) was found to violate PH assumption yet it is a determinant of U5CM as seen using the other methods. Hence the importance of using other methods like Cox Aalen's method and RSF VIMP for prediction in the sections that followed.

In this research, we have worked with all the variables that are candidate determinants of U5CM. We then performed variable selection and predict using methods that do not require PH assumption.

Our findings show that child mortality is associated with variables related to: child characteristics (such as age at death of the child), reproduction factors of the mother (such as the number of siblings born before), feeding characteristics and anthropometric measurements. This is in line with other findings such as Ayiko et al. (2009) who used Cox PH regression and established that region of residence, sex of the child, type of birth (multiple), birth interval (less than 24 months after the preceding birth), and mother's education were related with an increased risk of children mortality before their fifth birthday. Nasejje et al. (2015) also established that factors related to mother characteristics and previous births such as sex of the

child, sex of the head of the household and the number of births in the past one year was found to be significant. Sreeramareddy et al. (2013) explored the effect of mothers education, child's sex, rural/urban residence, household wealth index, regions ecological zones and development. Its worth to note that even though most of the studies that rely on DHS datasets Ayiko et al. (2009), Nasejje et al. (2015), Sreeramareddy et al. (2013) are challenged with high dimensional data and a variable selection dilemma, there is no mention of any statistical form of variable selection.

DHS datasets typically are composed of over 700 variables that are candidate determinants of child mortality and one need to carefully select which variables to include in the resultant regression type models. Majority of the studies explore the effect of a predetermined, select group set of covariates, based on self intuition or variables explored from previous studies. We attempted to do a variable selection using a machine learning algorithm, before subjecting the selected variables to variable prediction using Cox PH regression, Cox Aalens model and RSF VIMP at different levels of this research.

Other than finding the determinants of under five mortality, different data balancing methods were used and model selection done using concordance index. In their research Afrin et al. (2018) used SMOTE to balance data before integrating it with RSF. In this research, we first compared the use of external data balancing techniques which include over-sampling, under-sampling, both sampling and SMOTE methods. Under-sampling resulted in a better model with a concordance index of 0.91 as compared to other balancing methods used. SMOTE which is a hybrid method generates synthetic samples along the line segment joining two minority samples. By so doing there is a tendency of generating a decimal value in factor or numeric variables which are not meant to be in decimal form hence distorting categorical variables in the data. In as much as under-sampling method may discard potentially useful data in majority class there is no loss of data in the minority class which is our main class of interest.

The other challenge addressed in this study was the selection of the best splitting method when the PH assumptions do not hold. A number of studies have explored the different splitting rules in RSF. Miao et al. (2018) proposed an improved RSF by using weighted log-rank test in splitting the node while using the model of Yang and Prentice (2005), Hong et al. (2018) used R-squared splitting rule in survival forests, Wanyonyi et al. (2019) compared RSF using different splitting rules. In this study, we considered a balanced dataset for maximum growth of the tree. The BRSF was then analyzed using logrank, logrank score and Bs. gradient splitting rules when the PH assumptions are violated. Log-rank test has been used for survival splitting as a means for maximizing survival difference between nodes Ciampi et al. (1988) ,Segal (1988) Segal (1988) [10], Segal (1995), LeBlanc and Crowley (1992) and LeBlanc and Crowley (1993). Since the use of log-rank splitting rule is optimal when PH assumption is met, we preffered the use of Bs.gradient splitting rule when PH assumptions are violated.

5.2 Conclusion

In this research, we have developed an IBRSF model for analysis of right censored data in situations where: data is highly imbalanced, high dimensional and variables do not satisfy the PH assumption. Based on the developed model, we have identified the determinants of U5CM in Nairobi region of Kenya using the 2014 KDHS data.

This was achieved through a unified process which involved three stages: data balancing, variable selection and variable prediction. The first stage attempted to balance the data using four different external data balancing techniques which are random under-sampling, random over-sampling, both-sampling and SMOTE methods. Based on concondance index, random under-sampling method emerged the best with the highest concordance of 0.9048.

In the second stage, the balanced data was integrated with RSF for variable

selection process. Using variable importance in RSF, the important variables were extracted from the dataset with 757 varables. This stage enabled us to identify the variables to use for prediction stage. During this stage, comparison of different splitting rules in the process of growing trees was carried out. The splitting rules used are log-rank, log-rank score and Bs.gradient splitting rules. Based on concordance index, log-rank scored the highest concordance of 0.916 followed by Bs.gradient splitting rule with a concondance of 0.8674. Since optimality of log-rank splitting rule is achieved when PH assumption is satisfied, Bs.gradient splitting rule was taken as the most optimal.

Variable prediction stage followed after variable selection. This stage involved the use of the selected variables in identification of determinants of U5CM. Variable prediction was carried out using three different methods at different satages. This was done as follows. In the first objective, Cox PH model was used . This was done after having performed model diagnostics to verify the adequacy of fitting the model. These showed that some of the variables which include B7, B12, V214, V207, V417 violated PH assumption. The involved variables were interacted with time variying covariates after which some of them did not satisfy the assumption leading to removal of the variables. This led to the challenge of falilure to integrate all potential determinants of mortality some of which could be highly predictive. In the second objective, we opted to work with Cox Aalen's model which is an appropriate alterantive for Cox model when PH assumptions are not satisfied inorder to avoid complete removal of predictor variables. In the third objective, we worked with RSF VIMP, a non parametric tree based method. The method is more superior since it overcomes the challenges with Cox and Cox Aalen's models.

The developed IBRSF model involved the use of the most optimal method from each stage. Hence, the IBRSF model involves working with balanced data where balancing is done using random under-sampling method and both variable selection and predictions are carried out using RSF VIMP. During the tree growing process,

Bs.gradient method was used to split the rule. The method is able to deal with challenges of extreme imbalance in datasets, high dimensionality of dataset and working with variables that violate PH assumption.

Using IBRSF model, we were able to identify the highest determinants of U5CM as age, interelations between age, height, weight and BMI and succeeding birth intervals.

5.3 Areas for Further Research

Identification of determinants of mortality is a very important area of reasearch for guiding clinical decesions. In this study, we have developed an IBRSF model for anlysis of right censored data in situations where PH assumption is violated. Our research mainly looked into the aspects of improving RSF model by using data balancing and node splitting rules for variable selection and prediction in the precense of censored data. Like in many other studies, our findings enables future efforts for studying determinants of mortality. There is still more that can be done to improve this study and direct future studies.

To begin with, our research concentrated on externl data balancing techniques. There is need to look into other data balancing techniques such as the algorithm level and cost sensitive learning and ensemble-based balancing methods for integration with RSF method.

Secondly, our study was limited to analysis of data that is right censored. There is need to look into other types of censoring such as left cesoring, interval censoring ,double censoring among others.

References

Afrin, K., Illangovan, G., Srivatsa, S. S., and Bukkapatnam, S. T. S. (2018). Balanced random survival forests for extremely unbalanced, right censored data. *Department of Industrial and Systems Engineering*, 108:246–257.

Athey, S., Tibshirani, J., and Wager, S. (2016). Solving heterogeneous estimating equations with gradient forests. Research Papers 3475, Stanford University.

Ayiko, R., Antai, D., and Kulane, A. (2009). Trends and determinants of under-five mortality in uganda. *East African journal of public health*, 6.

Batista, G., Prati, R., and Monard, M. (2004). A study of the behaviour of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter*, 6(1):20–29.

Bellera, C., MacGrogan, G., Debled, M., Tunon, L., and Brouste, V. (2010). Variables with time-varying effects and the cox model: Some statistical concepts illustrated with a prognostic factor study in breast cancer. *BMC Medical Research Methodology*, 10(20).

Bewick, V., Cheek, L., and Ball, J. (2004). Statistics review 12: survival analysis. *Critical Care (London, England)*, 8(5):389 –394.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24:123–140.

Breiman, L. (2001a). Random forests. *Machine Learning,*, 45(1):532.

Breiman, L. (2001b). Statistical modeling: The two cultures. *Statistical Science*, 6(199-231):206.

Breiman, L. (2003). *Manual-setting up, using and understanding random forests V4.0.* Available at ftp://ftp.stat.berkeley.edu/pub/users/breiman/Using_random_forests_v4.0.pdf.

Bureau, A., Dupuis, J., Falls, K., Lunetta, K. L., Hayward, B., Keith, T. P., and Eerdewegh, P. V. (2005). Identifying snps predictive of phenotype using random forests. *Genetic Epidemiology*, 28(171-182):206.

Cassy, A., Saifodine, A., Candrinho, B., do Rosrio Martins, M., da Cunha, S., Pereira, F. M., and Gudo, E. S. (2019). Care-seeking behaviour and treatment practices for malaria in children under 5 years in mozambique: a secondary analysis of 2011 dhs and 2015 imasida datasets. *Malaria journal*, 18(1):115.

Cateni, S., Colla, V., and Vannucci, M. (2014). A method for resampling imbalanced datasets in binary classification tasks for real-world problems. *Neurocomputing*, 135:32–41.

Chawla, N., Bowyer, K., Hall, L., and Kegelmeyer, W. (2002). Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357.

Chawla, N., Cieslak, D., Hall, L., and Joshi, A. (2008). Automatically countering imbalance and its empirical relationship to cost. *Data Mining and Knowledge Discovery.*, 17(2):225–252.

Chen, C., Liaw, A., and Breiman, L. (2004). Using random forest to learn imbalanced data. *University of California, Berkeley.*

Ciampi, A. andHogg, S., McKinney, S., and Thiffault, J. (1988). Recpam: A computer program for recursive partition and amalgamation for censored survival data. *Computer methods and programs in biomedicine*, 26(3):239–256.

Clark, T., Bradburn, M., Love, S., and Altman, D. (2003). Survival analysis part 1: basic concepts and first analyses. *British Journal of Cancer*, 89:232238.

Corsi, D. J., Neuman, M., Finlay, J. E., and Subramanian, S. (2012). Demographic and health surveys: a profile. *International Journal of Epidemiology*, 41(6):16021613.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187. 220.URL: http://www.jstor.org/stable/2985181.

Datta, S. and Das, S. (2015). Near-bayesian support vector machines for imbalanced data classification with equal or unequal misclassification costs. *Neural Networks*, 70:39–52.

Diaz-Uriarte, R. and Alvarez de Andres, S. (2006). Gene selection and classication of microarray data using random forest. *BMC Bioinformatics*, 7(3):205,206.

Ertekin, S., Huang, J., Bottou, L., and Giles, C. (2007). Learning on the border: active learning in imbalanced data classification. In CIKM 2007, Lisbon, . N. ., editor, *In: Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management*, number 127-136.

Estabrooks, A., Jo, T., and Japkowicz, N. (2004). A multiple resampling method for learning from imbalanced data sets. *Computer Intelligence*, 20(1):18–36.

Ettarh, R. and Kimani, J. (2012). Determinants of under-five mortality in rural and urban kenya. *Rural and Remote Health*, 12(1812).

Fernndez, A., Garca, S., Galar, M., Prati, R. C., Krawczyk, B., and Herrera, F. (2018). *Learning from Imbalanced Data Sets.* Springer, Gewerbestrasse 11, 6330 Cham, Switzerland.

Fiorentini, N. and Losa, M. (2020). Handling imbalanced data in road crash severity prediction by machine learning algorithms. *Infrastructures*, 5(61).

Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., and Herrera, F. (2012). A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches. *IEEE Transactions on Systems, Man, and Cybernetic-spart C: Applications and Reviews*, 42(4):463–484.

Garca, V., Snchez, J., and Mollineda, R. (2012). On the effectiveness of preprocessing methods when dealing with different levels of class imbalance. *Knowledge-Based Systems*, 25(1):13–21.

Harrell, F., Califf, R., Pryor, D., Lee, K., and Rosati, R. (1982). Evaluating the yield of medical tests. *JAMA Clinical Challenge*, 247(18):2543–2546.

Haseeb, A., Salleh, M., Rohmat, S., Hussain, K., and Mushtaq, M. (2019). Imbalance class problems in data mining: a review. *Indonesian Journal of Electrical Engineering and Computer Science.*, 14(3):1560–1571.

He, H. and Garcia, E. (2009). Learning from imbalanced data. *in IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263 –1284.

He, J., Zhang, J. X., Chen, C.-t., Ma, Y., De Guzman, R., Meng, J., and Pu, Y. (2020). The relative importance of clinical and socio-demographic variables in prognostic prediction in nonsmall cell lung cancer : A variable importance approach. *Medical Care*, 58(5).

Hong, W., Xiaolin, C., and Gang, L. (2018). Survival forests with r-squared splitting rules. *Journal Of Computational Biology*, 25(4):388–395.

Hothorn, T. and Lausen, B. (2003). On the exact distribution of maximally selected rank statistics. *Computational Statistics and Data Analysis*, 43(2):121–137.

Ishwaran, H. (2007). Variable importance in binary regression trees and forests. *Electronic Journal of Statistics*, Vol.1:519–537.

Ishwaran, H., Blackstone, E. H., Hansen, C. A., and Rice, T. W. (2009). A novel approach to cancer staging: Application to esophageal cancer. *Biostatistics*, 10(206):603–620.

Ishwaran, H. and Kogalur, U. (2015). *RandomForestSRC: Random Forests for Survival, Regression and Classification (RF-SRC).* https://cran.r-project.org/web/packages/randomForestSRC/randomForestSRC.pdf.

Ishwaran, H., Kogalurt, U. B., Blackstone, E. H., and Lauer, M. S. (2008). Random Survival Forests. *The Annals of Applied Statistics*, 2(3):841–860.

Ishwaran, H. and Lu, M. . (2019). Standard errors and confidence intervals for variable importance in random forest regression, classification, and survival. *Statistics in Medicine.*

Japkowicz, N. and Stephen, S. (2002). The class imbalance problem: A systematic study. *Intelligent Data Analysis*, 6(5):429 –449.

Joshi, M. V., Kumar, V., and Agarwal, R. C. (2001). Evaluating boosting algorithms to classify rare classes: Comparison and improvements. *IEEE International Conference on Data Mining*, pages 257–264.

Junyong, I. and Dong, K. L. (2019). Survival analysis: part ii applied clinical data analysis. *Korean Journal of Anesthesiology*, 72(5):441–457.

Khalilia, M., Chakraborty, S., and Popescu, M. (2011). Predicting disease risks from highly imbalanced data using random forest. *BMC Medical Informatics and Decision Making.*, 11:51.

Kleinbaum, D. G. and Klein, M. (2005). *Survival Analysis: A Self-Learning Text.* Springer, New York.

KNBS, MoH, NACC, KMRI, NCPD, and DHS, I. I. (2014). *Kenya Demographic and Health Survey 2014.*

Krawczyk, B. (2016). Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence*, 5:221–232.

LeBlanc, M. and Crowley, J. (1992). Relative risk trees for censored survival data. *International Biometric Society*, 48(2):411–425.

Leblanc, M. and Crowley, J. (1993). Survival trees by goodness of split. *J. Amer. Statist. Assoc.*, 88(422):457–467.

LeBlanc, M. and Crowley, J. (1993). Survival trees by goodness of split. *Journal of the American Statistical Association*, 88(422):457–467.

Lessmann, S. (2004). Solving imbalanced classification problems with support vector machines. *In IC-AI*, 4:214–220.

Lin, E., Chen, Q., and Qi, X. (2020). Deep reinforcement learning for imbalanced classification. *Applied Intelligence*, pages 1–15.

Liu, V. (2019). Predicting ovarian cancer survival times: Feature selection and performance of parametric, semi-parametric, and random survival forest methods. Master's thesis, Simon Fraser University.

Lopez, V., Fernndez, A., Garca, S., Palade, V., and Herrera, F. (2013). An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. *Information Sciences*, 250(29):113–141.

Lunardon, N., Menardi, G., and Torelli, N. (2013). *R package ROSE: Random Over-Sampling Examples (version0.0-3).* Universit di Trieste and Universit di Padova, Italia. http://cran.r-project.org/web/packages/ROSE/index.html. [p79].

Lunetta, K. L., Hayward, L. B., Segal, J., and Eerdewegh, P. V. (2004). Screening large-scale association study data: Exploiting interactions using random forests. *BMC Genetics*, 5(32):206.

Lpez, V., Fernndez, A., Moreno-Torres, J. G., and Herrera, F. (2012). Analysis of preprocessing vs. cost-sensitive learning for imbalanced classification. open problems on intrinsic data characteristics. *Expert Systems with Applications*, 39(7):6585–6608.

Mazurowski, M. A., Habas, P. A., Zurada, J. M., Lo, J. Y., Baker, J. A., and Tourassi, G. D. (2008). Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance. *Neural Networks.*, 21:427–436.

Meinshausen, N. (2006). Quantile regression forests. *The Journal of Machine Learning Research*, 7:983–999.

Miao, F., Cai, Y.-P., Zhang, Y.-T., and Li, C.-Y. (2015). Is random survival forest an alternative to cox proportional model on predicting cardiovascular disease? In Lacković, I. and Vacic, D., editors, *6^{th} European Conference of the International Federation for Medical and Biological Engineering*, volume 45. IFMBE Proceedings, Springer, Cham.

Miao, F., Cai, Y.-P., Zhang, Y.-X., Fan, X.-M., and Li, Y. (2018). Presdictive Modelling of Hospital Mortality for Patients with Heart Failure by Using an Improved Random Survival Forest. *Department of Industrial and Systems Engineering*, 6.

Morvan, L., Carlier, T., Bastien Jamet, Clment Bailly, C. B.-M. P. M. F. K.-B., and Mateus, D. (2020). Leveraging rsf and pet images for prognosis of multiple myeloma at diagnosis. *International Journal of Computer Assisted Radiology and Surgery*, 15:129–139.

Nasejje, J., Mwambi, H., and Dheda, K. (2017). A comparison of the conditional inference survival forest model to random survival forests based on a simulation study as well as on two applications with time-to-event data. *BMC Medical Research Methodology*, 17(115).

Nasejje, J. B., Mwambi, H. G., and Achia, T. N. (2015). Understanding the determinants of under-five child mortality in uganda including the estimation of unobserved household and community effects using both frequentist and bayesian survival analysis approaches. *BMC public health*, 15(1):1003.

Ofek, N., Rokach, L., Stern, R., and Shabtai, A. (2017). A fast clustering-based undersampling method for addressing the class imbalance problem. *Neurocomputing*, 243(88-102).

Olson, D. (2005). Data set balancing. In Shi Y., Xu W., C. Z., editor, *Data Mining and Knowledge Management*, volume 3327, Springer, Berlin, Heidelberg.

O'Quigley, J. and Pessione, F. (1991). The problem of a covariate-time qualitative interaction in a survival study. *Biometrics*, 47:101–15.

Pan, W. (1998). Rank invariant tests with left truncated and interval censored data. *Journal of Statistical Computation and Simulation.*, 61(18):163–174.

Quinlan, J. (1991). Improved estimates for the accuracy of small disjuncts. *Machine Learning*, 6:93–98.

Scheike, T. and Zhang, M. (2002). An additive-multiplicative cox-aalen model. *Scandinavian Journal of Statistics*, 29(1):75–88.

Segal, M. (1988). Regression trees for censored data. *Biometrics.*, 15:35–47.

Segal, M. (1995). Extending the elements of tree-structured regression. *Statistical Methods in Medical Research*, 4(3):219–236.

Sreeramareddy, C., Kumar, H., and Sathian, B. (2013). Time trends and inequalities of under-five mortality in nepal: A secondary data analysis of four demographic and health surveys between 1996 and 2011. *PLoS ONE*, 8(11). e79818. doi:10.1371/journal.pone.0079818.

Stekhoven, D. J. and Bhlmann, P. (2012). Missforestnon-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112118.

Tang, Y., Zhang, Y. Q., Chawla, N. V., and Krasser, S. (2008). Svms modeling for highly imbalanced classification. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39(1):281–288.

Torgo, L. (2010). *Data Mining using R: learning with case studies.* Chapman and Hall/CRC.

Wanyonyi, K., Owuor, N., and Sarguta, R. (2019). Comparison of random survival forests split rules in selecting the determinants of under five mortality in kenya using 2014 dhs data. *Research report in Mathematics*, (15).

Weathers, B. and Cutler, R. (2017). Comparison of survival curves between cox proportional hazards, random forests, and conditional inference forests in survival analysis. Master's thesis, Utah State University. https://digitalcommons.usu.edu/gradreports/927.

Yan, Y., Liu, R., Ding, Z., Du, X., Chen, J., and Zhang, Y. (2019). A parameter-free cleaning method for smote in imbalanced classification. *IEEE Access*, 7:23537–23548.

Yang, S. and Prentice, R. (2005). Semi parametric analysis of short-term and long-term hazard ratios with two-sample survival data. *Biometrika.*, 92:1–17.

Yang, S. and Prentice, R. (2010). Improved logrank-type tests for survival data using adaptive weights. *Biometrics.*, 66(1):30–38.

Yang, Q. Wu, X. (2006). 10 challenging problems in data mining research. *International Journal of Information Technology and Decision Making.*, 5(04):597–604.

Yang, Z., Tang, W., Shintemirov, A., and Wu, Q. (2009). Association rule mining-based dissolved gas analysis for fault diagnosis of power transformers. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, 39(6):597–610.

Yaya, X., Xiu, L., Ngai, E., and Weiyun, Y. (2009). Customer churn prediction using Improved Balanced Random Forests. *Expert systems with Application*, 36:5445–5449.

Yen, S. and Lee, Y. (2009). Cluster-based under-sampling approaches for imbalanced data distributions. *Expert Systems with Applications*, 36(3):57185727.

Zhao, Y. and Cen, Y. (2013). Data mining applications with r. *Academic Press: Cambridge, MA, USA,*.

Zhu, Z. and Song, Z. (2010). Fault diagnosis based on imbalance modified kernel fisher discriminant analysis. *Chemical Engineering Research and Design*, 88(8):936–951.

APPENDICES

A Description of Important variables

Tables A.1 and A.2 gives the description of the variables found to be predictive in this research.

Table A.1: Description of Important variables.

Category	Variable	Description
Child Characteristics at Birth	B1	Month of birth of the child.
	B7	Age at death of the child at completed months.
	B 8	Current age of the child in single years for all living children.
	B 12	Succeeding birth interval.
Reproduction (siblings information)	B 16	Child's line number in household
	V 203	Total number of daughters living at home.
	V 206	Total number of sons who have died.
	V 207	Total number of daughters who have died.
	V 208	Total number of births in the last five years
	V 214	Imputed duration of the current pregnancy
	V 218	Total number of living children
	V 219, 220	Total number of living children including current pregnancy
	V 230	Year og the last pregnancy termination
	V 231	Century month code of the last pregnancy termination
	V 238	Total number of births in the last three years
Height and Weight and Hemoglobin	HW70	Height for age standard deviation (according to WHO).
	HW 71	Weight for age standard deviation (according to WHO).
	HW 72	Weight for height standard deviation (according to WHO).
	HW 73	BMI standard deviation (according to WHO).
	HW 1	Age in months of the child.
Maternity	M 1E	Last tetanus injection before last pregnancy.
Vaccination History	H2Y	BCG vaccination date-year.
	H3Y	DPT-HEP.B-HIB (PENTAVALENT) 1 year.
	H4Y	Oral Polio 1 year.
	H5Y	DPT-HEP.B-HIB (PENTAVALENT) 2 year.
	H6Y	Oral Polio 2 year.
	H7Y	DPT-HEP.B-HIB (PENTAVALENT) 3 year.
	H8Y	Oral Polio 3 month.
	H9Y	Measles 1 year.
	H10Y	Oral Polio at birth year.

Table A.2: Description of Important variables.

Category	Variable	Description
Contraceptive Use	V321	Marital duration at sterilization in 5-year groups with single women and those sterilized before marriage coded 0.
	V478	Reason the respondent does not intend to use a method of contraception in the future.
Maternity	M1E	Injections administered by a health worker.
Maternity and Feeding	V417	Number of entries in the pregnancy and postnatal care history.
	V 418	Number of entries in the immunization history.
	V 419	Number of entries in the height and weight table.
Marriage	V506	The rank of the respondent among the partner's wives.
Fertility Preferences	V605	Desire for more children.
Family Planning	V3A07	First source for current method.
Injections last 12 months	V478	Injections administered by a health worker.
Delivery care	M4	The duration of breastfeeding of the child in months.
	M5	The calculated months of breastfeeding.
	M6	The duration of postpartum amonorrhea after the birth of the child in months.
Malaria	ML1	The no. of times they took SP/Fansidar during pregnancy.
Height and Weight and Hemoglobin	HW1Y	Age in months of the child.
	HW3	Height in centimeters.
	HW4	Height for Age percentile.
	HW5	Height for Age standard deviations from the reference median.
	HW6	Height for Age percent of reference median.
	HW7	Weight for Age percentile.
	HW9	Weight for Age percent of reference median.
	HW10	Weight for Height percentile.
	HW11	Weight for Height standard deviations from the reference median.
	HW18	Month of measurement.
	HW19	Whether the weight at birth (variable M19) was recorded from a health card (code 1) or from the mothers recall (code 2). Children who were not weighed at birth are coded 0.

B Authorization Letter

Nov 15, 2018

Hellen Waititu
Moi University
Kenya
Phone: 0722668325
Email: hlnwaititu@gmail.com
Request Date: 11/15/2018

Dear Hellen Waititu:

This is to confirm that you are approved to use the following Survey Datasets for your registered research paper titled: "Improved balanced random survival forests for right censored data.":

Kenya

To access the datasets, please login at: https://www.dhsprogram.com/data/dataset_admin/login_main.cfm. The user name is the registered email address, and the password is the one selected during registration.

The IRB-approved procedures for DHS public-use datasets do not in any way allow respondents, households, or sample communities to be identified. There are no names of individuals or household addresses in the data files. The geographic identifiers only go down to the regional level (where regions are typically very large geographical areas encompassing several states/provinces). Each enumeration area (Primary Sampling Unit) has a PSU number in the data file, but the PSU numbers do not have any labels to indicate their names or locations. In surveys that collect GIS coordinates in the field, the coordinates are only for the enumeration area (EA) as a whole, and not for individual households, and the measured coordinates are randomly displaced within a large geographic area so that specific enumeration areas cannot be identified.

The DHS Data may be used only for the purpose of statistical reporting and analysis, and only for your registered research. To use the data for another purpose, a new research project must be registered. All DHS data should be treated as confidential, and no effort should be made to identify any household or individual respondent interviewed in the survey. Please reference the complete terms of use at: https://dhsprogram.com/Data/terms-of-use.cfm.

The data must not be passed on to other researchers without the written consent of DHS. Users are required to submit an electronic copy (pdf) of any reports/publications resulting from using the DHS data files to: archive@dhsprogram.com.

Sincerely,

Bridgette Wellington

Bridgette Wellington
Data Archivist
The Demographic and Health Surveys (DHS) Program

C Graphs Showing Balance in the 2014 KDHS

The graphs in this section shows the nature of balance in some of covariates after the overall 2014 KDHS data was balanced.

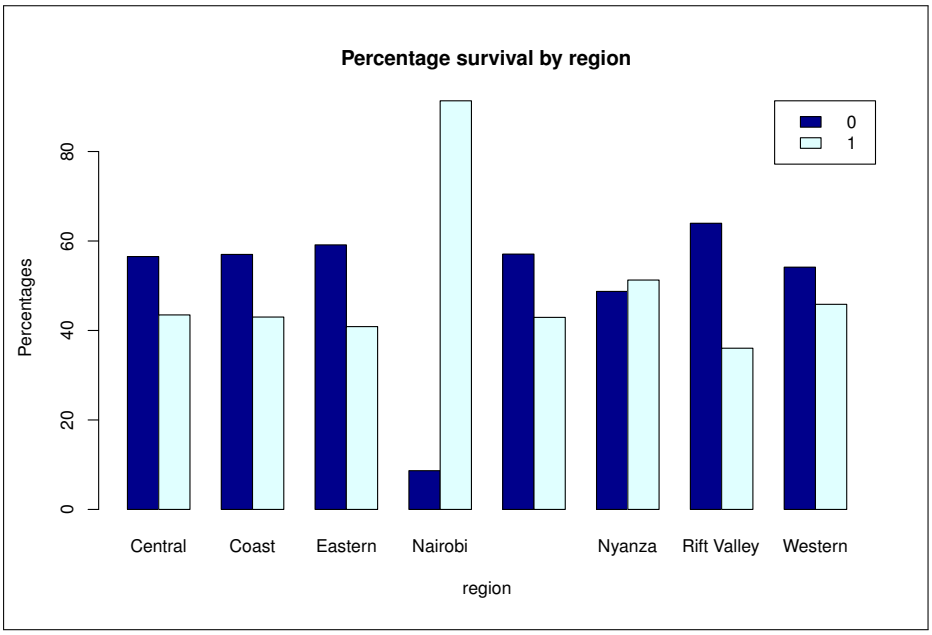

Figure C.1: Balance percentage survival by region in 2014 KDHS Data

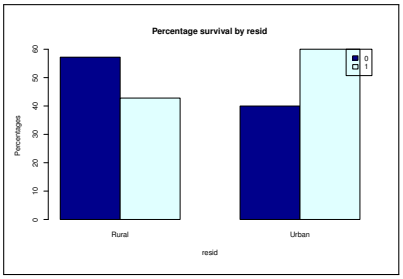

Figure C.2: Balanced percentage survival by residence 2014KDHS Data

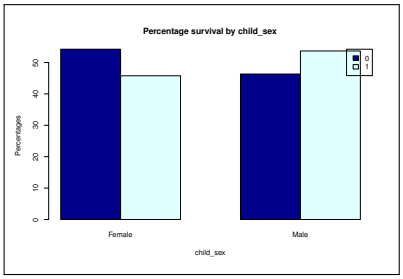

Figure C.3: Balanced percentage survival by sex 2014 KDHS

D Graphs showing Balance in Nairobi Region with Different Balancing Methods

The graphs in D shows the nature of balace after the Nairobi region dataset was balanced using the different balancing techniques as indicated.

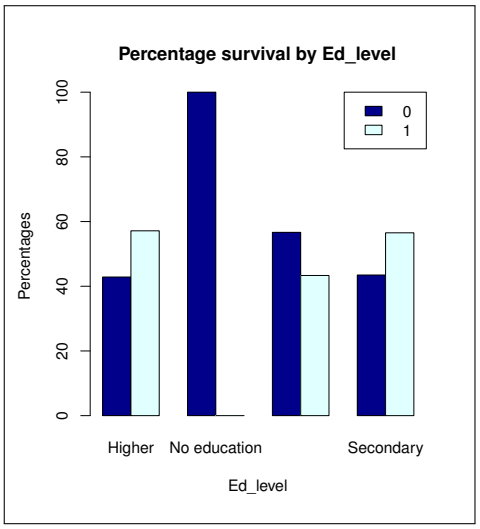

Figure D.4: Undersampling Balanced percentage survival by Education level

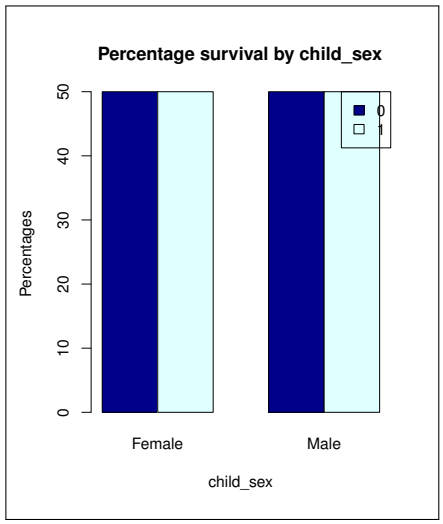

Figure D.5: Undersampling Balanced percentage survival by sex

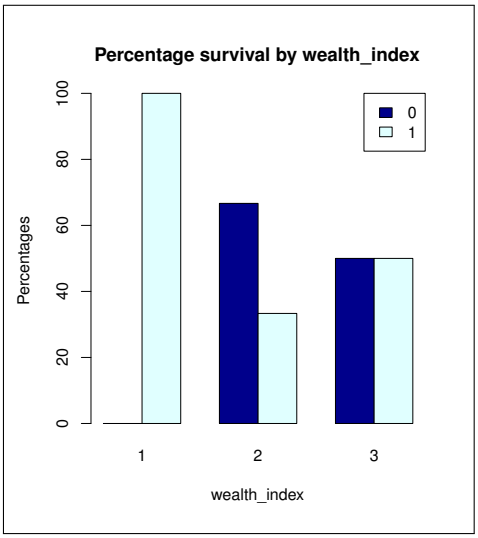

Figure D.6: Undersampling Balanced percentage survival by Wealth index

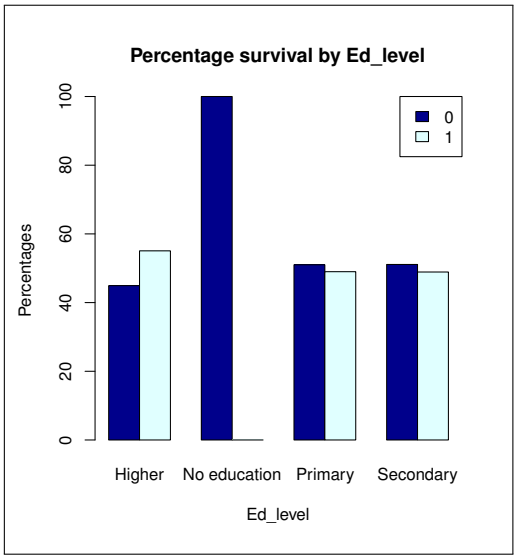

Figure D.7: Oversampling Balanced percentage survival by Education level

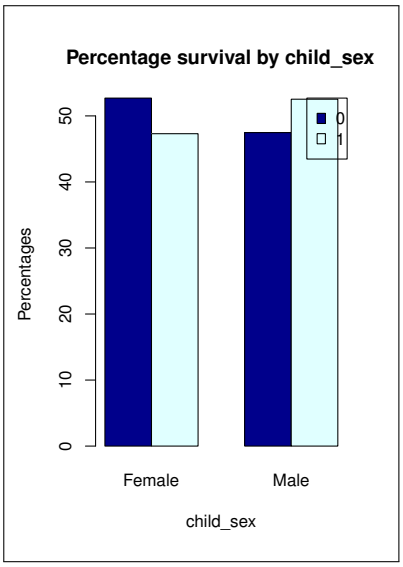

Figure D.8: Oversampling Balanced percentage survival by sex

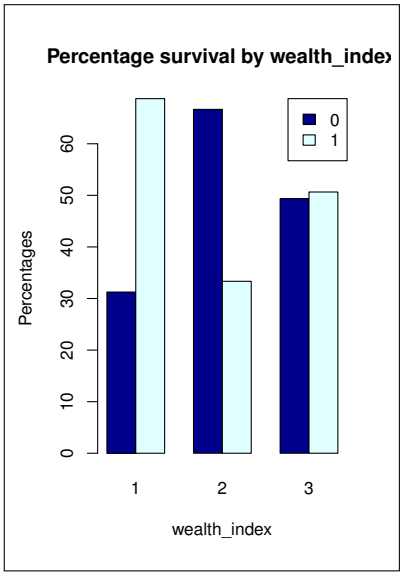

Figure D.9: Oversampling Balanced percentage survival by Wealth index

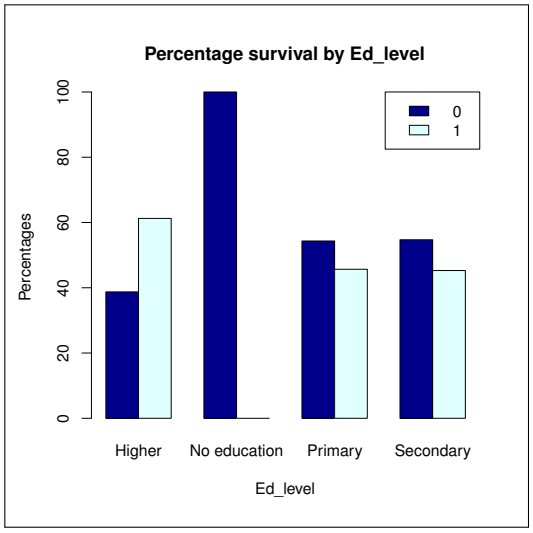

Figure D.10: Bothsampling Balanced percentage survival by Education level

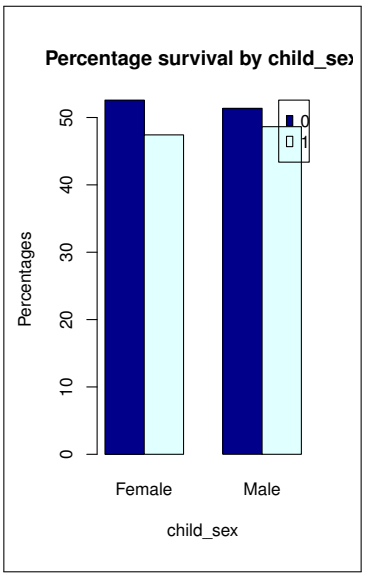

Figure D.11: Bothsampling Balanced percentage survival by sex

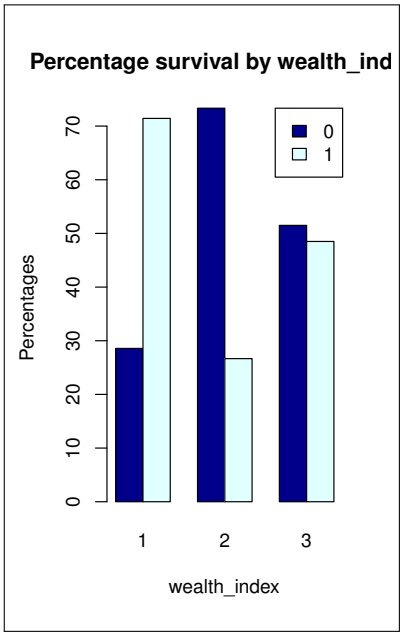

Figure D.12: Bothsampling Balanced percentage survival by Wealth index

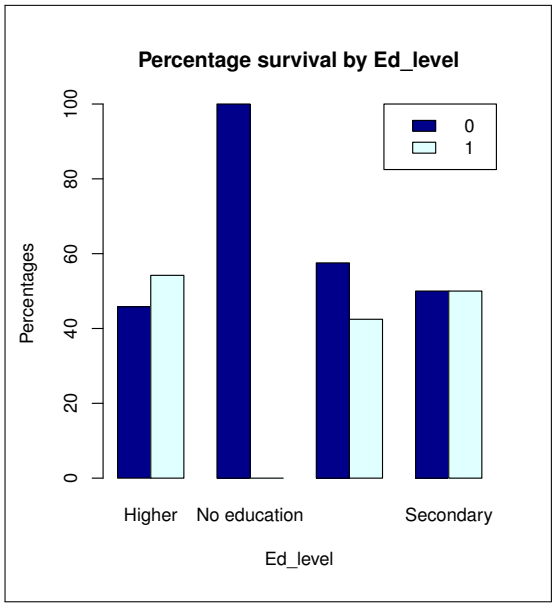

Figure D.13: ROSE sampling Balanced percentage survival by Education level

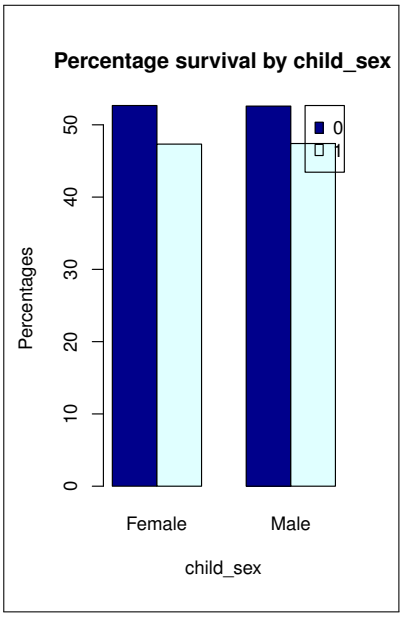

Figure D.14: ROSE sampling Balanced percentage survival by sex

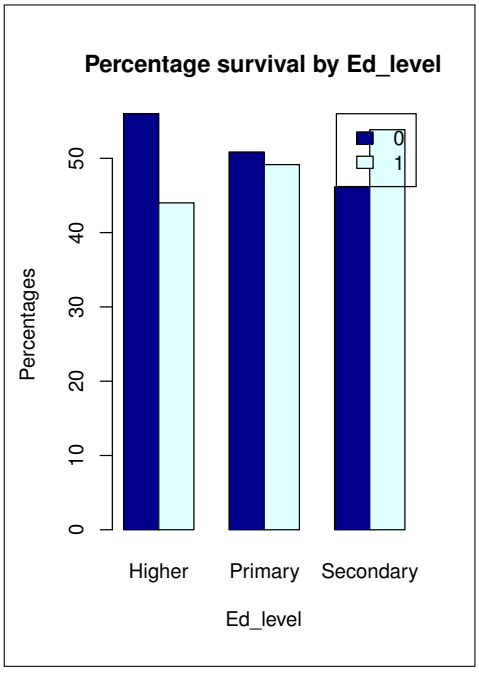

Figure D.15: SMOTE sampling Balanced percentage survival by Education level

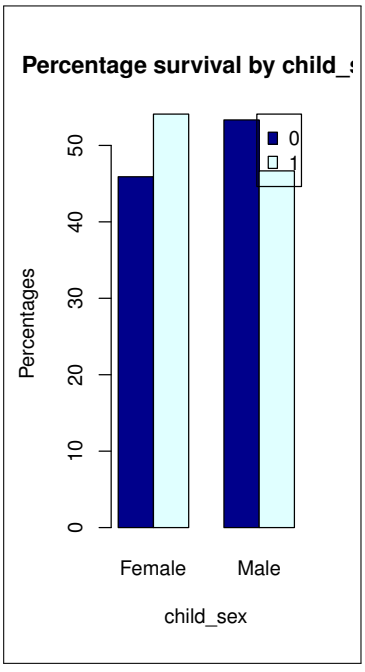

Figure D.16: SMOTE sampling Balanced percentage survival by sex

E Residuals for Predictors with Different Balancing Techniques

In this section, residuals for different predictors from different models are given as indicated.

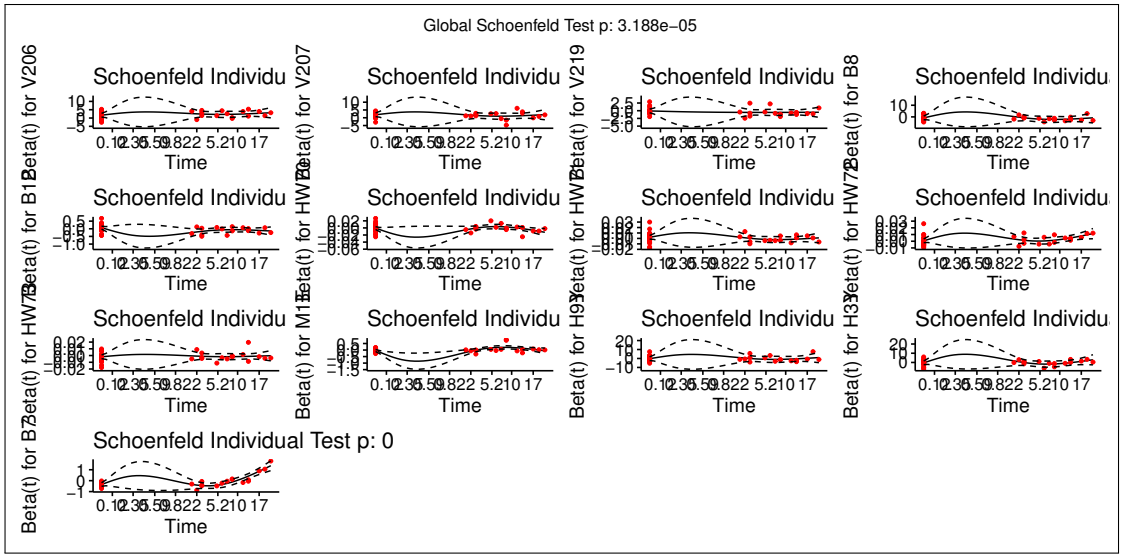

Figure E.17: Schoenfeld Residuals for BRSF with Undersampling

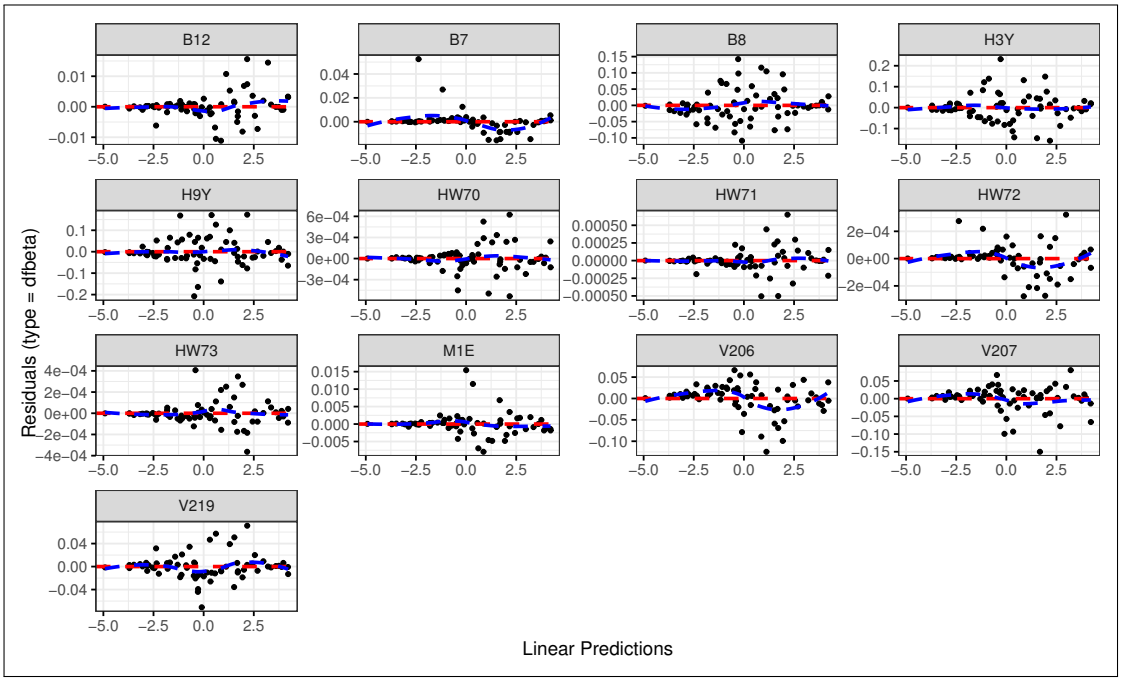

Figure E.18: dfbeta residuals for BRSF with Undersampling

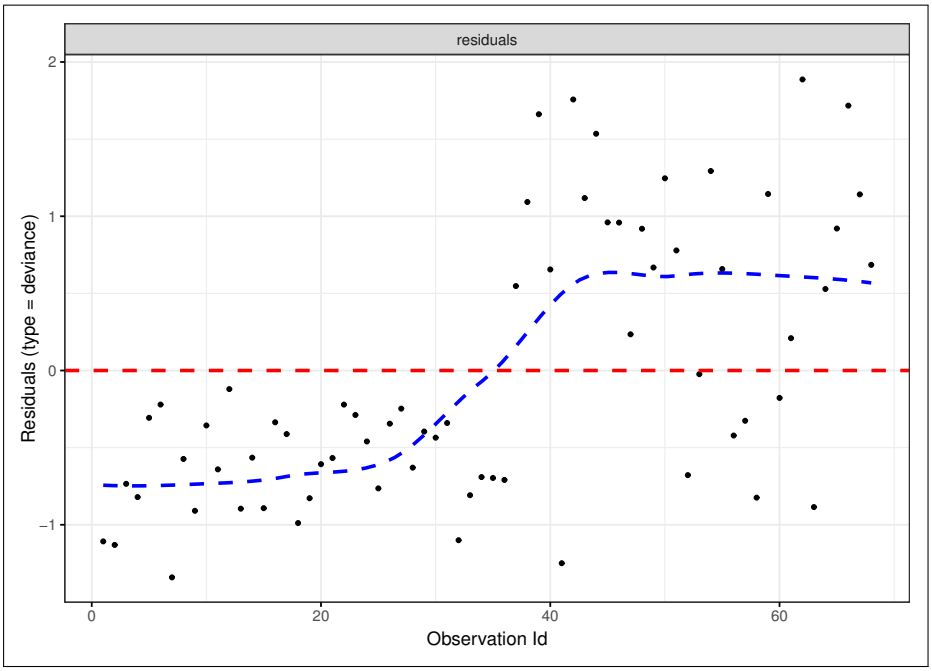

Figure E.19: Deviance Residuals for BRSF with Undersampling

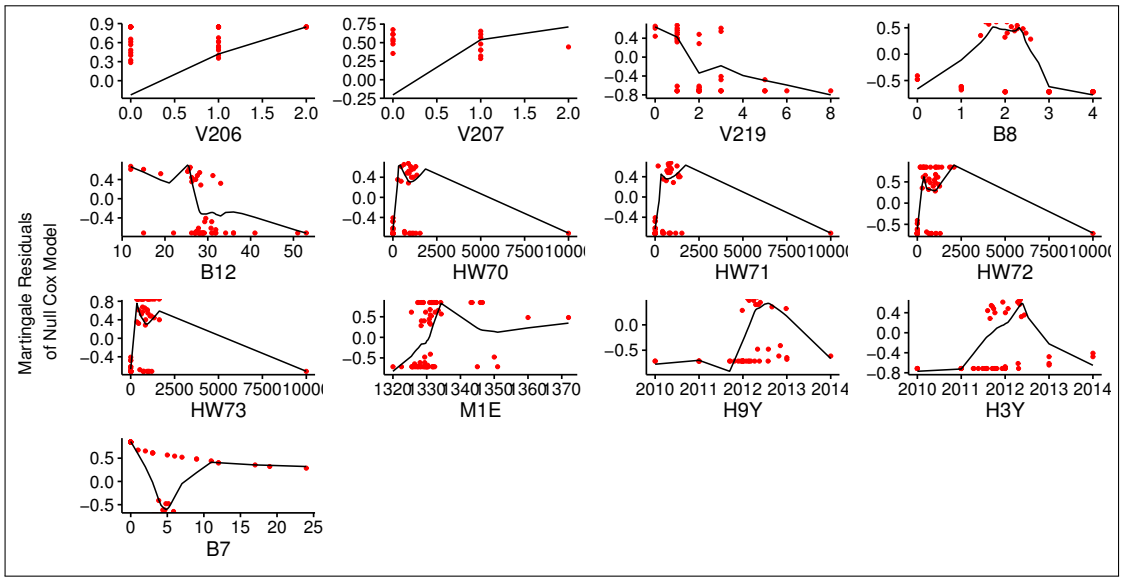

Figure E.20: Martingale residuals for BRSF with Undersampling

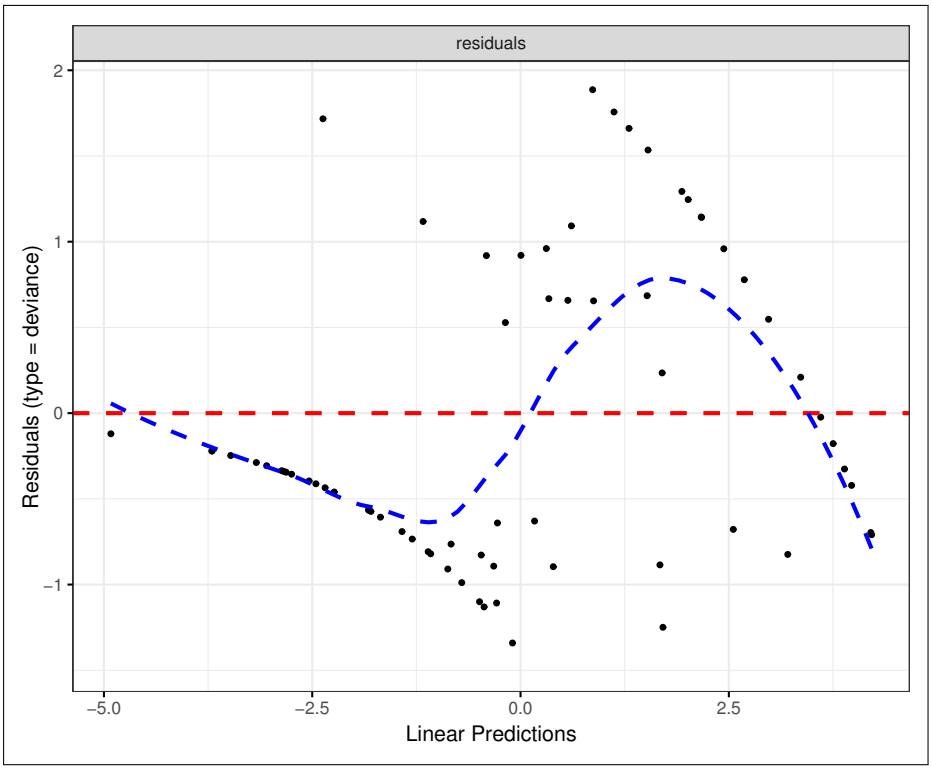

Figure E.21: Deviance Residuals for BRSF with Undersampling

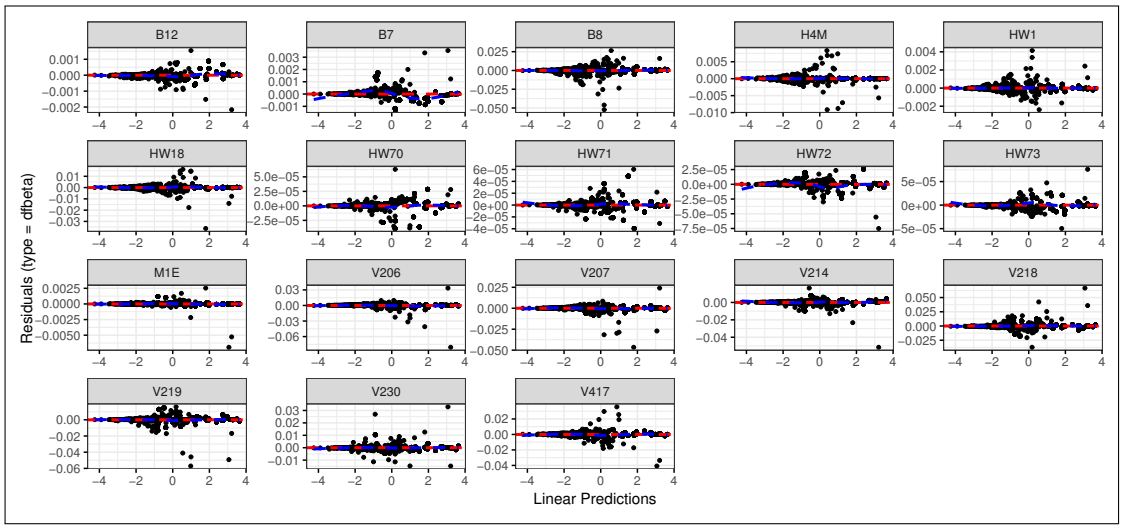

Figure E.22: dfbeta residuals for BRSF with Oversampling

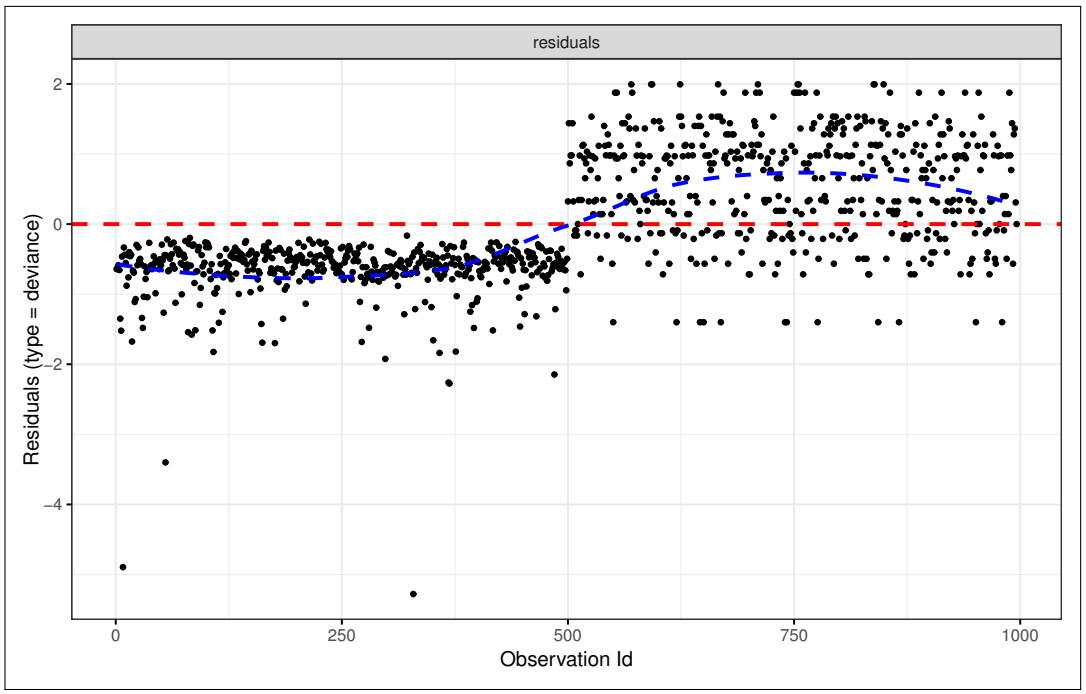

Figure E.23: Deviance Residuals for BRSF with Oversampling

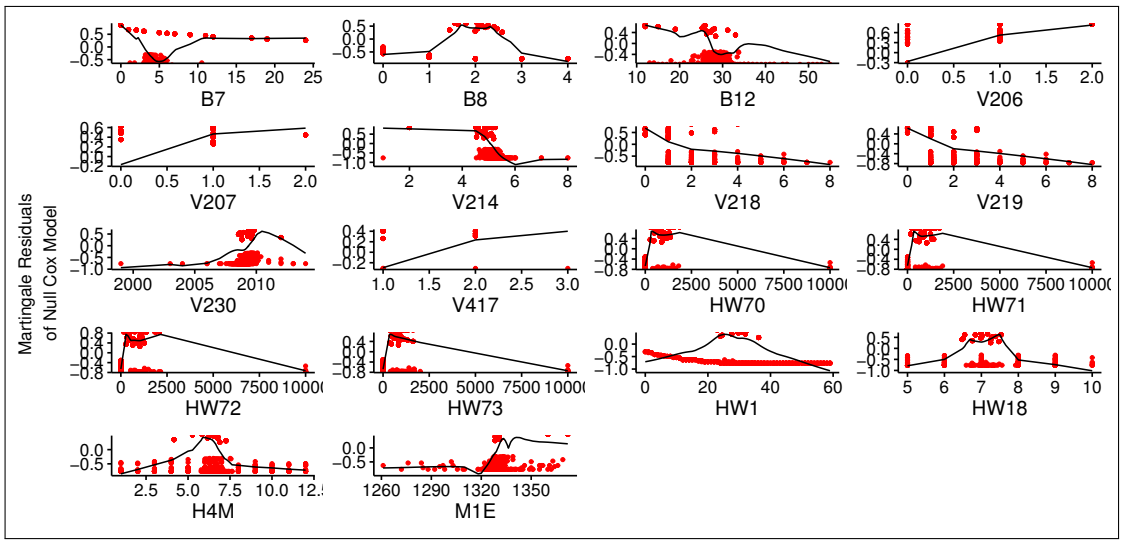

Figure E.24: Martingale residuals for BRSF with Oversampling

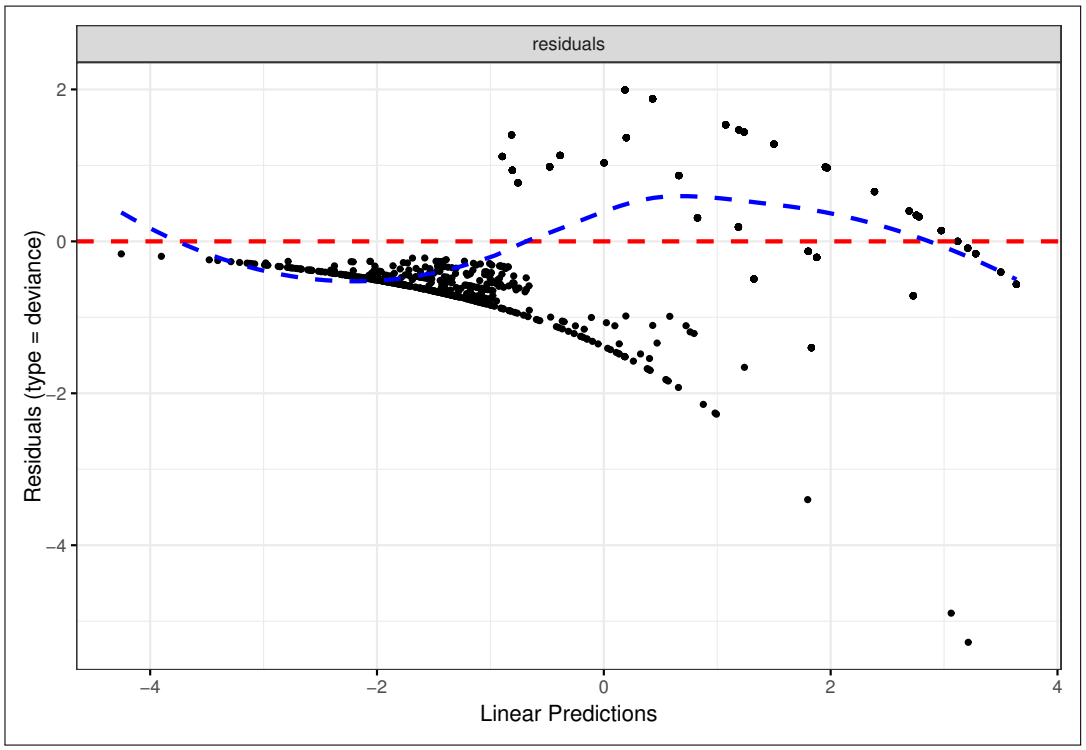

Figure E.25: Deviance Residuals for BRSF with Oversampling

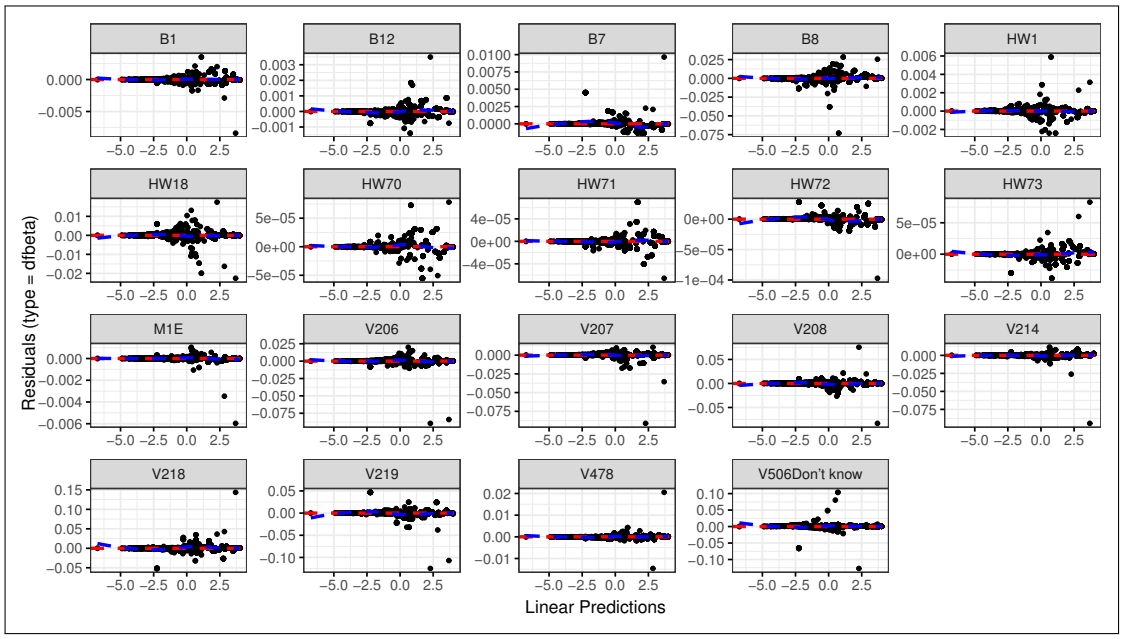

Figure E.26: dfbeta residuals for BRSF with Bothsampling

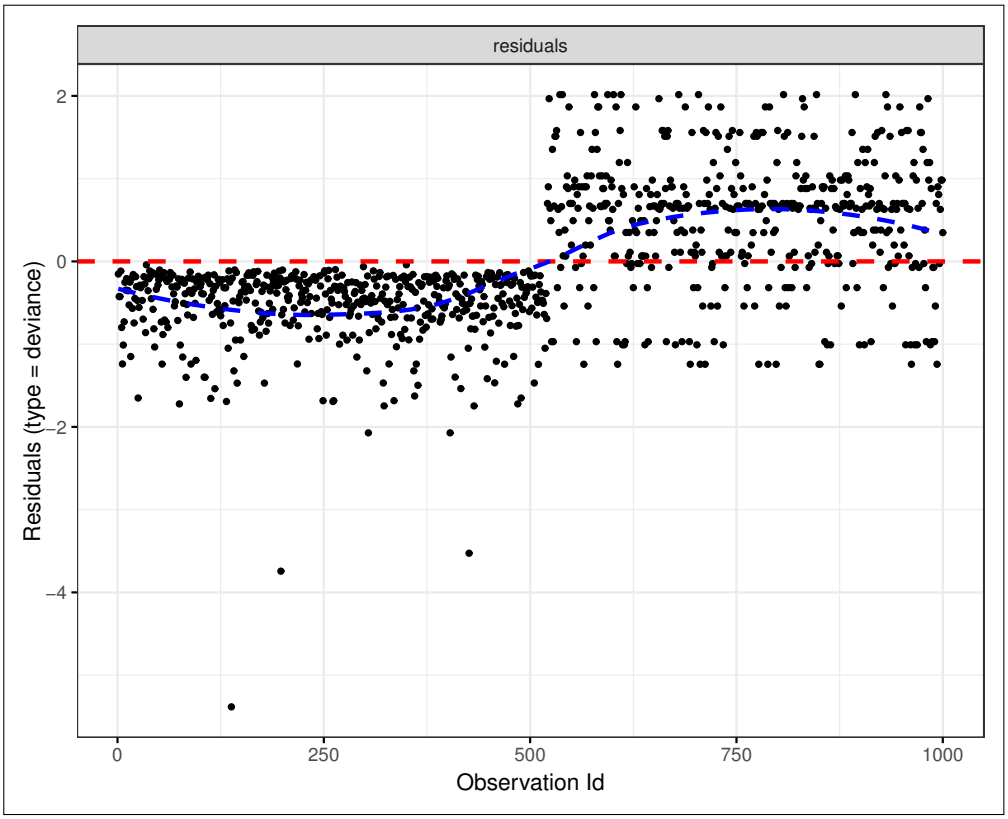

Figure E.27: Deviance Residuals for BRSF with Bothsampling

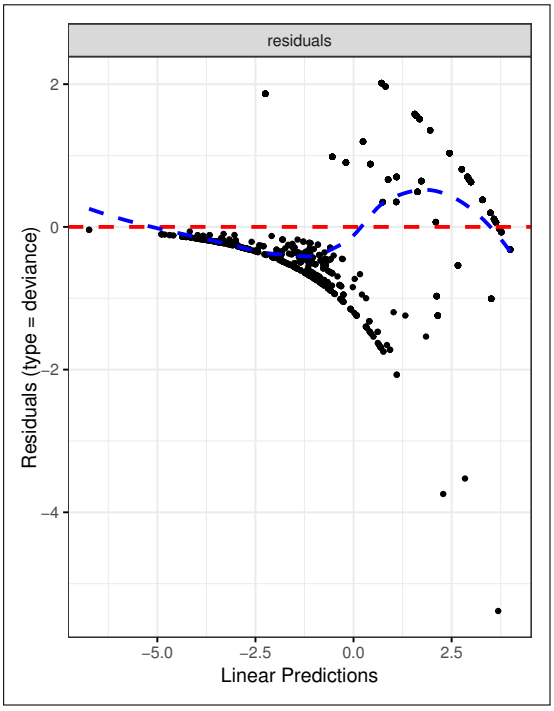

Figure E.28: Deviance Residuals for BRSF with Bothsampling

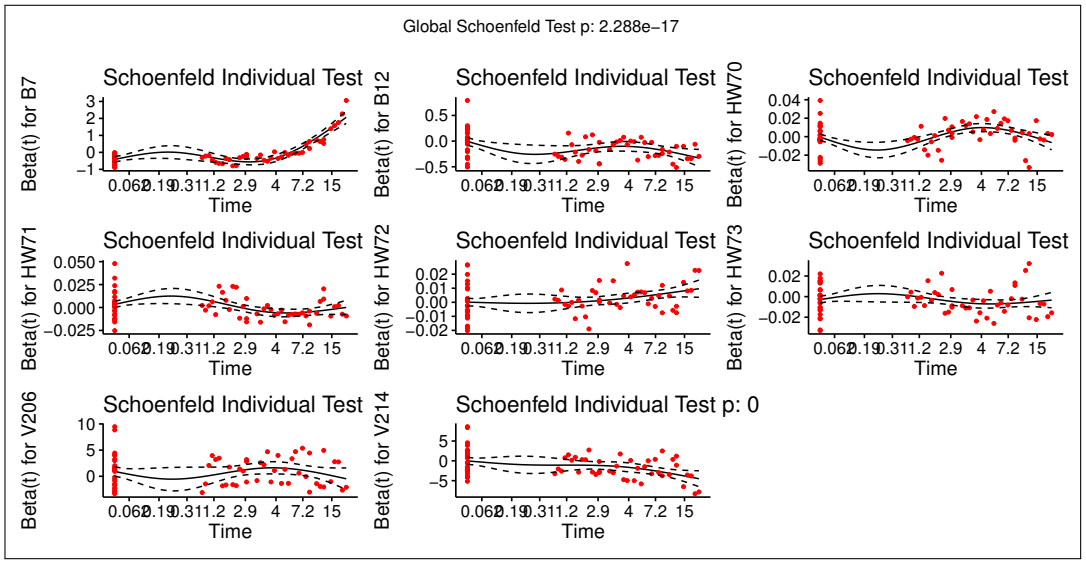

Figure E.29: Schoenfeld Residuals for BRSF with SMOTE sampling

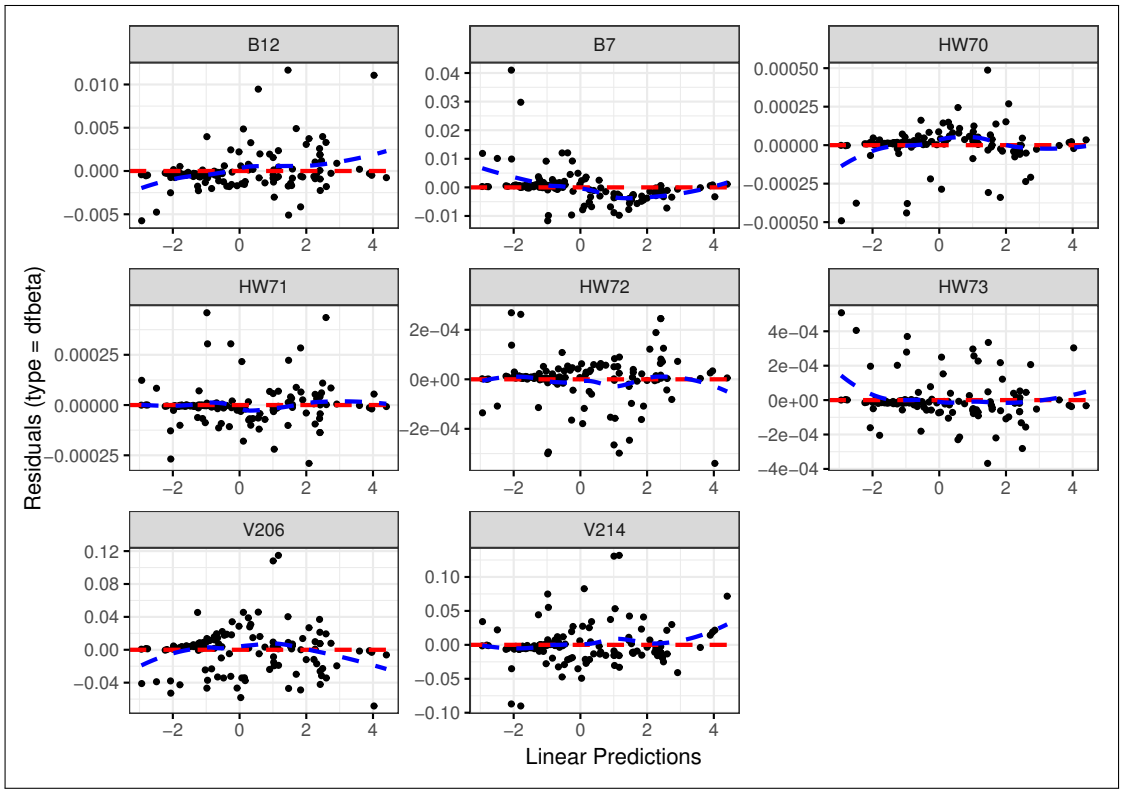

Figure E.30: dfbeta residuals for BRSF with SMOTE sampling

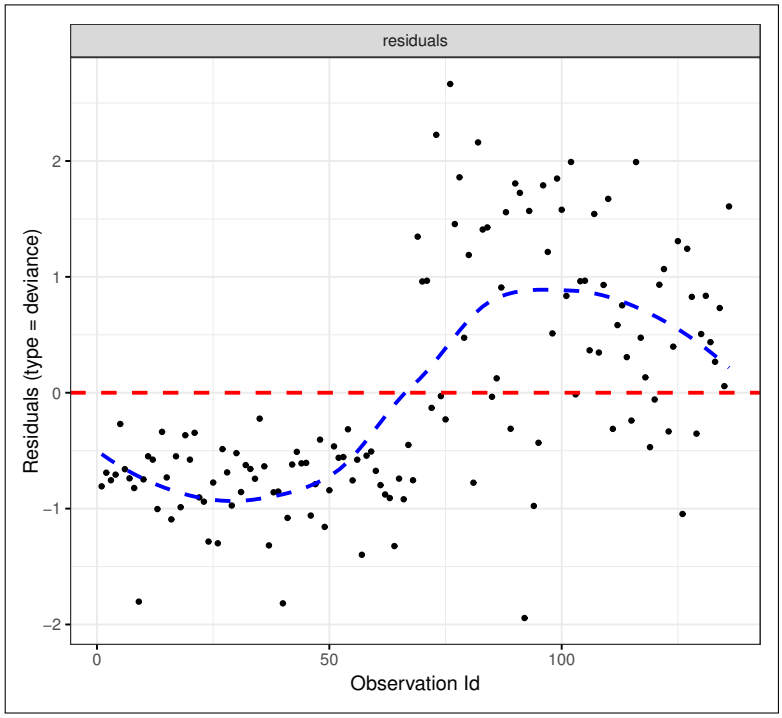

Figure E.31: Deviance Residuals for BRSF with SMOTE sampling

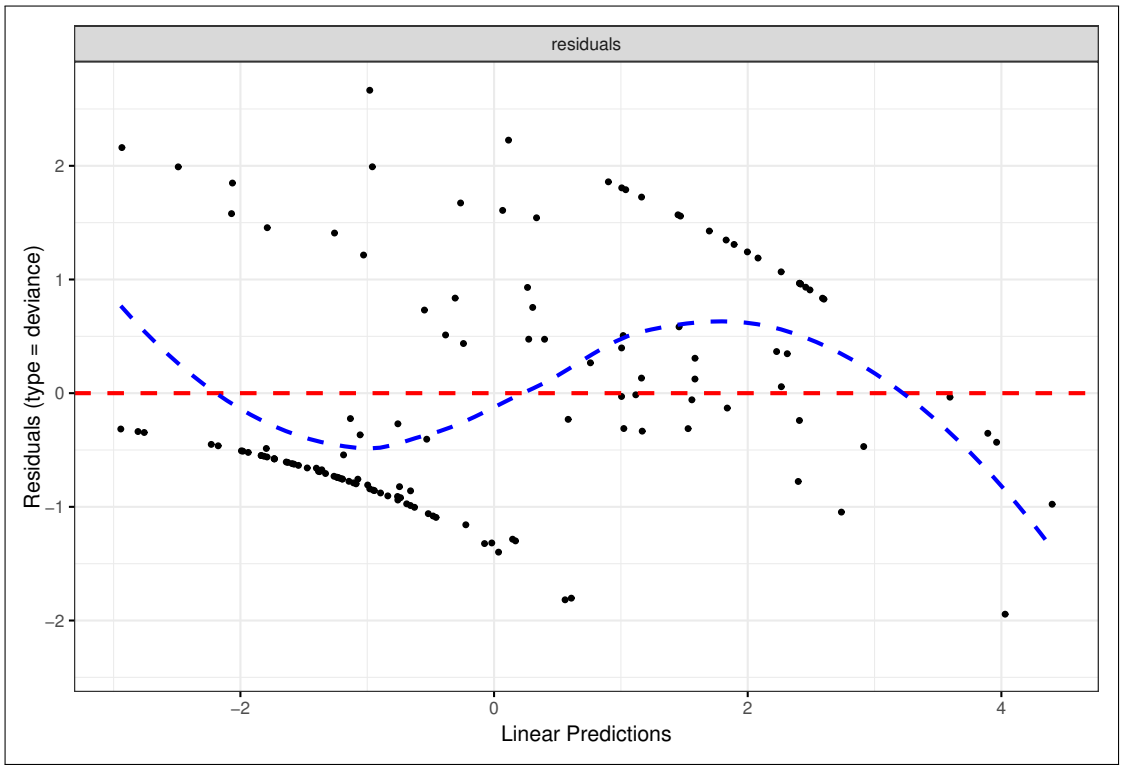

Figure E.32: Deviance Residuals for BRSF with SMOTE sampling

F Selected variables using different balancing techniques

The table in this section shows all the selected variables before the removal of variables that violated PH assumption for prediction using Cox PH model.

Table F.3: BRSF Cox ph model predictors with violation of PH assumptions.

Undersampling					SMOTE				
Var	coef	E(coef)	Se(coef)	P-value	Var	coef	E(coef)	Se(coef)	P-value
V206	2.0731	7.9490	0.4044	$2.95e^{-07}$	B7	-0.0897	0.9142	0.0490	0.0674
V207	1.6683	5.3034	0.4317	0.0001	B12	-0.1096	0.8962	0.0231	$2.21e^{-06}$
V219	-0.2936	0.7456	0.2184	0.1789	HW70	0.0015	1.0015	0.0011	0.1760
B8	-0.1113	0.3291	0.5771	0.0541	HW71	0.0005	1.0005	0.0011	0.6676
B12	-0.8249	0.9841	0.0349	0.6450	HW72	0.0021	1.0021	0.0008	0.0110
HW70	-0.0010	0.9989	0.0014	0.4641	HW73	-0.0039	0.9961	0.0010	0.0002
HW71	0.0008	1.0008	0.0012	0.4869	V206	0.8879	2.4301	0.2867	0.0019
HW72	0.0019	1.0019	0.0010	0.0517	V214	-1.2569	0.2845	0.2771	$5.73e^{-06}$
HW73	-0.0015	0.9985	0.0009	0.1284					
M1E	0.0164	1.0165	0.0280	0.5584					
H9Y	-0.2311	0.7937	0.7495	0.7579					
H3Y	-0.2797	0.7560	0.7231	0.6989					
B7	-0.1561	0.8554	0.0589	0.0081					
Oversampling					Both sampling				
Var	coef	E(coef)	Se(coef)	P-value	Var	coef	E(coef)	Se(coef)	P-value
B7	-0.1038	0.9014	0.0138	$6.60e^{-14}$	B1	-0.0652	0.9369	0.0174	0.0002
B8	-0.1247	0.8827	0.1704	0.4641	B7	-0.1069	0.8987	0.0162	$4.02e^{-11}$
B12	-0.0341	0.9665	0.0088	0.0001	B8	-0.4422	0.6426	0.1947	0.0231
V206	1.7171	5.5682	0.1087	$< 2e^{-16}$	B12	-0.0694	0.9329	0.0093	$9.46e^{-14}$
V207	1.3351	3.8005	0.1022	$< 2e^{-16}$	HW70	-0.0006	0.9993	0.0004	0.1167
V214	-0.1055	0.8999	0.0951	0.2673	HW71	0.0001	1.0001	0.0004	0.8623
V218	-1.1085	0.3300	0.1965	$1.07e^{-08}$	HW72	0.0019	1.0019	0.0003	$5.24e^{-11}$
V219	0.6849	1.9835	0.1822	0.0001	HW73	-0.0012	0.9988	0.0003	$3.04e^{-05}$
V230	-0.1491	0.8614	0.0866	0.0852	V206	1.5152	4.5505	0.1215	$< 2e^{-16}$
V417	0.3228	1.3809	0.1199	0.0071	V207	1.7935	6.0102	0.1193	$< 2e^{-16}$
HW70	0.0008	1.0008	0.0003	0.0207	V208	0.4238	1.5278	0.1452	0.0035
HW71	0.0002	1.0002	0.0003	0.5649	V214	-0.0964	0.9081	0.1039	0.3535
HW72	0.0019	1.0019	0.0003	$2.15e^{-11}$	V218	-0.7598	0.4678	0.2528	0.0027
HW73	-0.0027	0.9973	0.0003	$< 2e^{-16}$	V219	0.4049	1.4991	0.2295	0.0776
HW1	-0.0009	0.9992	0.0152	0.9534	V478	-0.1709	0.8429	0.0279	$1.03e^{-09}$
HW18	-0.0699	0.9324	0.0653	0.2837	V506	1.9114	6.7625	0.2877	$3.05e^{-11}$
H4M	0.0351	1.036	0.0333	0.2920	HW1	0.0441	1.0451	0.0175	0.0119
M1E	0.0077	1.0077	0.0066	0.2414	HW18	-0.1390	0.8701	0.0756	0.0657
					M1E	0.0122	1.0123	0.0065	0.0613