University of Nairobi

School of Mathematics

ISSN: 2410-1397

Master Project in Biometry

Feature selection methods and Resampling techniques in Survival data: Determination of risk factors of under-five child mortality

Research Report in Mathematics, Number XX, 2021

Chelangat Maureen Rono

July 2021

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

Feature selection methods and Resampling techniques in Survival data: Determination of risk factors of under-five child mortality

Research Report in Mathematics, Number XX, 2021

Chelangat Maureen Rono

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

Master Thesis

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

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

Abstract

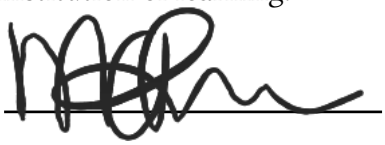
The main aim of the research was to identify the risk factors of under-five child mortality using Kenya Demographic Health Survey (KDHS) 2014 data set. Demographic Health Surveys are faced with three main challenges, acute class imbalance, high dimensionality and missing data. The KDHS 2014 data set is made up of 1129 variables and 20964 observations. In addition, the mortality class accounted for 4% of the data while the non-mortality class accounted for the remaining 96%. To determine the risk factors, first we dealt with the missing data by imputation. The class imbalance was handled using three balancing methods: both sampling, under-sampling and over-sampling. We then handled high dimensionality using three filter methods. Random survival forest was used to select highly predictive variables and parameter estimation was done using Cox-PH model. The variables that were found to be significant were child is twin, sex of child, births in the last five years, currently pregnant, wanted pregnancy, living children & current pregnancy, wanted last child, respondent slept under mosquito bed net, ideal number of children, disposal of youngest child's stools when not using toilet, received polio vaccine and weight for age standard deviation.

Master Thesis in Mathematics at the University of Nairobi, Kenya.
ISSN 2410-1397: Research Report in Mathematics

DISTRIBUTOR: School of Mathematics, University of Nairobi, Kenya

Declaration and Approval

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



August 20, 2021

Signature Date

Chelangat Maureen Rono
Reg No. I56/34764/2019

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

August 20, 2021

Signature Date

Dr Nelson Owuor
School of Mathematics,
University of Nairobi,
Box 30197, 00100 Nairobi, Kenya.
E-mail: onyango@uonbi.ac.ke

RJS

AUGUST 20, 2021

Signature Date

Dr Rachel Sarguta
School of Mathematics
University of Nairobi,
Box 30197, 00100 Nairobi, Kenya.
E-mail: rsarguta@uonbi.ac.ke

Dedication

This project is dedicated to my family and my supervisors, Dr. Nelson Owuor and Dr. Rachel Sarguta.

Contents

List of Tables

List of Figures

Acknowledgments

First, I thank the almighty God for enabling me to complete my studies. Secondly, Delta Africa Initiative SACCAB for funding my studies. A huge gratitude to my supervisors, Dr. Nelson Owuor and Rachel Sarguta. Finally, I'm grateful to all who supported me through various ways during my studies ....

Chelangat Maureen Rono

Nairobi, 2020.

1 Introduction

1.1 Background

Under-five child mortality (U5M) is defined as the likelihood of a child dying between age zero to 59 months. U5M is a key index of child health and overall development of a country and depicts the social, economic and environmental conditions in which the children live in as well as the health care status of the country. The world has documented significant reduction in under-five child mortality in the past three decades. Globally, the under-five mortality rate fell to 38 deaths per 1000 live births in 2019 compared to 93 deaths per 1000 live births in 1990, confirming a 59% reduction in child mortality (Unicef, 2019). In spite of this global advancement, sub-Saharan Africa carries on to experience the highest under-five child mortality. In 2019, sub-Saharan Africa accounted for more than half of the under-five mortality. Two-thirds of the deaths were from avoidable causes such as unsafe water, malnutrition, lack of education, health care and social service. With the current pattern, more efforts are needed to achieve the Sustainable Development Goals of reducing neonatal and under-five child mortality and achieving below 25 deaths per 1000 live births (Unicef-2019).

Demographic Health Surveys (DHS) are nationally representative samples that provide data on important health indicators. These health indicators include fertility, child mortality, family planning, health and nutrition. There has been a wide use of DHS surveys in many health studies by organizations and researchers. Several studies including Tagoe et al (2020), Zewudie et al (2020), Fikru et al (2019), Gebretsadik & Gabreyohannes (2016), Dejene (2016), Aheto (2019), Acheampong & Avorgbedor (2017) and Nasejje & Mwambi (2017) have put to use DHS data to study the risk factors of under-five child mortality in Sub-Saharan Africa. The risk factors established to be significant include wealth index, sex of the household head, total number of children ever born, number of children under five in a household, mother's birth in the last five years, mother's number of children living, family size, sex of the child, weight of the child, place of residence, religion, level of education of a mother, type of birth, birth order, family size, preceding birth interval, plurality, source of drinking water and mother's income. A number of the studies reviewed did not mention how they handled the high dimensionality and class imbalance in the data.

1.2 Survival analysis

Survival analysis are methods of modeling time to event data. The term event can be used in many ways. For example, an event can mean death, infection, marriage, divorce, tumor, remission, defective machine. Survival analysis is put to use in many fields including medicine, biology, economics, social and behavioral sciences and engineering.

Features of survival analysis

Survival time, t_i is the time until an event of interest happens.

Censored observation refers to an individual or subject who has not experienced the event of interest i.e. the survival time is not known. Censoring could be as a consequence of an individual leaving the study before end of study period, the study terminates before the subject experiences the event or a person is lost to follow-up. The above is a case of right censoring. Censoring happens in two ways, when the event of interest is experienced before the start of study period and when the event occurs after the study period.

Type of right censoring

Fixed type I censoring- This happens when a study is set to end after T years of follow up and each individual who does not have the event of interest during the study period is censored at time T years.

Random type I censoring- This happens when a study is set to end after T years but the censored individuals have different censored time.

Type II censoring- This happens when study subjects join the study at the same time but the study ends when a predetermined number of individuals have gotten the event of interest.

Type III censoring- This happens when subjects join the study at different times but the study period is fixed.

Survival function

The survival function is represented by $S(t)$ and it is the chance that the random variable T is greater than a specified time t i.e.

$S(t) = Pr(T > t)$

It gives the likelihood that a person survives past time t.

Properties of the survival function

1. The time t ranges as $t \in [0, \infty)$
2. $S(t)$ is a decreasing function i.e $S(t_1) \geq S(t_2)$ for $t_1 \leq t_2$
3. At time $t = 0$, $S(t = 0) = 1$ i.e. the chance of surviving past time 0 is 1.

The probability density function for $S(t)$ is obtained as follows

$S(t) = \int_t^\infty f(u)du$

We differentiate $S(t)$ to obtain

$f(t) = -\frac{dS(T)}{dt}$

And the mean life expectation is given by

$\mu = \mathbb{E}[T] = \int_0^\infty t f(t)dt$

Kaplan-Meier Estimator

The Kaplan-Meier(KM) estimator for the survival function is obtained as

$S_{KM}(T) = \prod_{i:ti<t} \frac{n_i - d_i}{n_i} = \prod_{i:t_i<t}(1 - \frac{d_i}{n_i}) \quad \forall t > 0$

Where n_i is the number at risk at time t_i and d_i is the number of events at time t_i.

Hazard function

The hazard function also known as hazard rate is an instantaneous rate of an event occurring and the formula is as below.

$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t} = \frac{f(t)}{S(t)}$

The cumulative hazard function is the total amount of risk that has been accumulated to time t and is given by

$H(t) = \int_0^t h(u)d(u)$

Survival regression

Cox proportional hazard model

This is a semi-parametric regression model that allows inclusion of predictors of the subject. The hazard function is represented as

$h(t,X) = h_0(t)\exp(\beta_1 X)$

Where $h_0(t)$ is the baseline hazard, X is a vector of predictors and β is a vector of regression coefficients

The hazard function can be generalized for p predictors as

$h(t,X) = h_0(t)\exp\{\sum_{i=1}^p \beta_i X_i\}$

1.3 Statement of problem

Demographic Health Surveys (DHS) are national representative surveys that provide data on a wide scope of health indicators. DHS has been used in many studies as a source of information on child mortality. However, just like any real life data set, DHS data has a challenge of high dimensionality and imbalanced class prevalence. The KDHS 2014 data constitutes 1129 variables and 20964 observations. The data set was found to have high imbalance with the mortality class constituting 4% and the non-mortality class 96%. Classification of data is negatively affected by imbalance data and high dimensionality

in a data set may influence the performance of the classifiers. The study seeks to explore and analyze the effect of different feature selection methods and re-sampling methods. Random forest which is a machine learning algorithm shall be used to check the efficacy of the methods. After resampling and feature selection, random survival forest is used for survival prediction and variable selection.

1.4 Objectives

1.4.1 Overall objective

The main objective of the study is to identify the risk factors of under-five child mortality using 2014 KDHS data set.

1.4.2 Specific objective

1. To identify determinants of under-five child mortality from a pool of independent variables.
2. To assess the effect of balancing techniques and feature selection methods in identification of determinants of under-five child mortality
3. To compare the performance of different balancing techniques and feature selection methods.

1.5 Significance of the study

Identifying the key determinants of under-five child mortality will help the government and health stakeholders to put in place measures and policies to reduce child mortality in the country.

2 Literature Review

2.1 Introduction

2.2 Risk factors of under-five child mortality

Several studies have been conducted to determine the factors of under-five child mortality. Tagoe et al (2020) set out to develop a predictive model of determinants of under-five mortality using 2008 and 2013 Sierra Leone Demographic Health Survey data. The study used Lasso regression technique, a machine learning algorithm to select the independent variables. The selected covariates were then used in the single-level and multi-level logistic regression model. The study showed that under-five child mortality is dependent on total number of children ever born,number of children under five in a household, mother's birth in the last five years, mother's number of children living, family size, contraceptive use and intention, number of eligible women in the household , type of toilet facility, sex of child and weight of the child at birth. However, the study noted high number of missing data. The study considered only the complete cases which might have led to loss of information.
According to Zewudie et al (2020), the risk factors of child mortality include place of residence, religion, level of education of a mother, breast feeding status, type of birth, sex of a child, birth order and family size. The study utilized 2016 Ethiopia Demographic Health Survey (EDHS) data and used both bi-variate and multivariate logistic regression. Similarly, Fikru et al (2019) utilized both bi-variate and multivariate logistic regression to determine predictors of under-five using EDHS 2016. The significant predictors were place of residence, preceding birth interval, plurality, size of child at birth and sex of the child. On the hand Gebretsadik and Gabreyohannes (2016) conducted a survival analysis to determine factors of under-five mortality using 2011 EDHS. The study used Kaplan Meier, log-rank test and Cox proportional hazard regression to select the factors. The following were found to be significant determinants: preceding birth interval, family size, birth type, breast feeding status, source of drinking water and income of mother. However, the study did not mention whether the proportional hazard assumption was met. Dejene (2016) assessed the levels, trend and determinants of under-five mortality in Amhara region, Ethiopia using EDHS (2000 to 2011). The study used both multivariate analysis and cox proportional hazard regression. The predictors found to be significant were: mother's education level, marital status, contraceptive use or intention and source of drinking water. The study however, did not highlight proportional hazard assumption and whether it was met.

Acheampong and Avorgbedor (2017)examined the trend of under-five mortality in Ghana using Ghana Demographic Health Survey (GDHS)data between 1988 to 2014. This was made possible by use of a logistic regression model. The study revealed that maternal age, marital status, breast feeding status, source of drinking water and the kind of assistance at birth were significant predictors of under-five mortality. Similarly, Iddrisu et al (2020) examined the relationship between under-five child mortality and drivers of child mortality using logistic regression under frequentist and Bayesian framework. The results for frequentist framework and Bayesian framework were in line. The study identified cesarean section, size of the child at birth and mother's education status as significant risk factors of under-five child mortality. Also, the study highlighted imbalance classification of the mortality class. On the other hand, Aheto(2019) set to identify a predictive model, assessed the determinants of under-five child mortality and heterogeneity in the household level using 2014 GDHS. The study used both single level binary and multilevel logistic regression model. The study showed no significant unobserved household level variations. Also, the study indicated number of total children ever born, number of births in the last five years, contraceptive use or intention, place of residence, multiple birth, maternal education and sex of children as significant drivers of child mortality.

Nasejje et al (2015) carried out a survival analysis to understand the risk factors of under-five mortality in Uganda and estimate heterogeneity in the household-level and community level using Uganda Demographic Health Survey (UDHS) 2011. The study utilized Cox proportional hazard model with frailty effects and drew inference using frequentist framework and Bayesian framework. The study results showed significant heterogeneity at household level and not at community level and the significant predictors as sex of household head, sex of the child and number of births in the past one year. All the variables that did not satisfy the proportional hazard assumption were excluded and this might have led to key risk factors being left out. Also, the study mentioned missing data as one of the limitation but did not detail how it was handled. Similarly, Nasejje and Mwambi (2017) implemented both Cox proportional hazard regression and Random Survival Forest on UDHS 2011 to identify the risk factors of under-five mortality and to compare the performance of the two models. Both models were compared in the presence of covariates that satisfied the proportional hazard assumption and those that violated the proportional hazard condition. The results showed Random Survival Forest to have a good predictive performance in presence of non-proportional hazard. In addition, the two models identified the significant variables as sex of the household head, sex of child and number of births in the past one year. Random Survival Forest further showed that the variables that had been eliminated because they violated the proportional hazard condition were significant predictors. These included number of children under five in the household, number of births in the past five years, wealth index, total number of children ever born and child order. The study noted the high missing data as a limitation and proposed combining multiple imputation and Random Survival Forest in future studies.

Waititu et al (2020) sought to identify the risk factors of under-five child mortality using Kenya Demographic Survey data (KDHS) 2014 by first handling the high imbalance problem that is notable in the data set. The balancing techniques employed in the study were re-sampling techniques namely random under-sampling, random over-sampling, both-sampling and synthetic minority over-sampling technique. This was followed by variable selection using Random Survival Forest as per the four techniques. The predictors from the four data sets were then used in the Cox proportional hazard model. Model diagnostics were then carried out to establish a good model fit. The study results showed that random under-sampling was a better technique and the significant determinants as number of dead sons, number of dead daughters, number of daughters living, number of children living, number of deliveries in the last three years, weight for height standard deviation and interaction between child's month of birth and number of dead sons.

3 Methodology

3.1 Introduction

The chapter outlines the three main aspects, namely, random survival forest modeling, feature selection methods and balancing techniques.

3.2 Data

The data for the study was extracted from the 2014 Kenya Demographic and Health Survey (KDHS) data. The data consisted of 1130 variables and 20964 observations. Summaries of the predictors indicated acute class imbalance. KDHS 2014 data is a national survey that is divided to 8 regions as per the former provinces. Our study focused on Nyanza region which recorded the highest under-five child mortality in the survey. A subset of 151 variables were considered in our analyses. The data was found to have high missing data. The *MissForest* algorithm was used to impute missing data. The Nyanza region data set also showed high data imbalance with 5.8% in the mortality class compared to 94.2% in the non-mortality group. Similarly, the predictors had high imbalance in the mortality class. Table 3.1 illustrate the class distribution on our data.

High class imbalance in the data leads to high accuracy by predicting the majority class but fail to predict the rare event which is always the aim of modeling. Imbalanced data has been observed to affect machine learning algorithms including random survival forests as they are designed to maximize accuracy and reduce errors.

Table 1. Target class distribution on Nyanza region KDHS data 2014

No. of features	151	
No. of samples	2926	
Target output	mortality	
Class	No-mortality	Mortality
Frequency	2757	169
Percentage	94.22%	5.78%
Imbalance ratio (IR)	0.06	

3.3 Class imbalance and its effects

Data is said to be imbalanced if one or some of the classes have more samples than the others. The most frequent class is referred as the majority class while the infrequent class as the minority class. Class imbalance is observed in many real-life data sets. These include mortality data, fraud detection in a transaction data, defects in a software data, disease screening data, advertising click-through data,natural disasters data among others.

Several machine learning algorithms are devised to perform best when the sample size of the classes are equal. This is because they are devised to maximize accuracy and minimize error. When a data set is imbalanced, the machine learning classifiers provide best classification of the numerous class while the rare classification is perverted. The prediction accuracy is biased to the prevalent class while the minority class remains unknown in spite of the high precision in the prediction model.

Since the minority class is a rare occurrence, there is high likelihood of it being regarded as noise by classifiers while the noise may be erroneously regarded as the minority class as both are rare events. This will result to miss-classification of the model.

3.4 High dimensionality and its effects

Classification involve building models to predict the response variable based on several predictor variables (features). Many real life data used for classification is usually made up of redundant, useless or misleading features that may adversely affect the accuracy of classifiers and therefore, feature selection is a crucial task in solving classification problems.

The KDHS 2014 data set faces the challenge of high dimensionality which results in high computational costs and complexity in interpreting data. High class imbalance combined with high feature dimensionality can cause the models not to detect infrequent cases. In order to overcome the problem of class imbalance, it is crucial to have equal sample sizes when working with machine learning techniques. Our focal point in our study is to develop models that favor the infrequent class but also retain the precision of the prevalent class.

3.5 Data balancing techniques

Various machine learning techniques have been proposed to solve class imbalance classification problem. The are categorized into four. These include data preprocessing techniques, algorithm level techniques, cost sensitive learning and ensemble methods. Our study shall focus on the preprocessing methods.

3.5.1 Preprocessing techniques

These are executed before building classifiers to achieve better data input. These methods include resampling methods and feature selection methods.

Re-sampling techniques

Re-sampling methods are popular ways of handling class imbalance puzzle. They are used to re-balance the sample space inequality in the class distribution. The methods are classified into three.

1. Oversampling methods

   The method removes the effect of imbalance by generating new infrequent class samples. The popular method that is used is random over-sampling. Random over-sampling methods involve randomly duplicating the infrequent class until they are equal with the most prevalent class. This method may lead to over-fitting as it makes identical copies of the rare class.

2. Under-sampling methods- it involves removing samples from the frequent class until equality is achieved in the two classes. The popular method used is random under-sampling which involve elimination of the prevalent class randomly. This method can lead to loss of information as useful data can be discarded.

3. Hybrid methods- this is a combination of over-sampling methods and under-sampling methods. Our study shall utilize both sampling.

Feature selection

Feature selection involve decreasing the dimensionality of the data by choosing a subset of k features from the original pool of features that will achieve best performance of the classifier. The focus of feature selection is to pick out the best subset of input variables that explain the target variable. The feature selection techniques are divided into filters, wrappers and embedded methods. These are machine learning techniques. Our study will focus on the filter methods.

Filter methods

The method checks the importance of features by ranking approach and the features with low scores are eliminated. The evaluation of features is independent of the classifier. The following filter methods shall be considered in our study.

(i) **Chi-squared statistics**

The method evaluates the dependency of two categorical variables. In a data set, the method calculates the chi-squared statistic between the feature and response variables and checks for presence of association. The chi-squared statistic is obtained as:

$$\chi^2 = \sum_{i=1}^{m} \sum_{j=1}^{k} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Where m is the number of intervals, k is the number of classes, O_{ij} is the observed value in the i^{th} interval j^{th} class and E_{ij} is the expected value in the i^{th} interval j^{th} class. A high chi-squared value means two features are dependent. A high score indicates that the feature is highly important.

(ii)**Information gain**

This is a measure based on information theory of entropy. Entropy measures the amount of information in a random variable. Information gain measures decrease in entropy before and after adding features (Uguz, 2012). The information gain about random variable X provided by random variable Y is obtained as:

$IG(X|Y) = H(X) - H(X|Y)$
Where
$H(X) = -\sum_{i=1}^{k} P(x_i) \log_2 P(x_i)$
is the entropy of a random variable X and

$H(X|Y) = -\sum_i P(y_i) \sum_i P(x_i|y_i) \log_2 (P(x_i|y_i))$
is the entropy of X after observing Y. Each variable will be ranked based on the information gain value and the higher the score, the more the importance of the feature.

(iii)**ReliefF attribute evaluation**
It is a supervised feature quality estimation which assess an attribute by how well its value identifies samples that are from different groups but are similar to each other. The ReliefF algorithm selects random samples R and then searches for k of its nearest neighbors from the same class(called nearest hit) and also k nearest neighbor from each of different class(called nearest miss). Feature x is valued as the sum of weighted differences in nearest miss and nearest hit. The formula for updating the quality estimation W_x for each attribute x is as follows.

$$W_x = W_x - \frac{diff(F,R,H)^2}{m} + \frac{diff(F,R,M)^2}{m}$$

Where W_x is the weight for attribute x, R is a random sample instance, M is the nearest miss, H is the nearest hit and m is the number of random sample instance.

3.6 Random Forests

A random forest (RF) is a classifier comprising of a collection of tree-structured classifiers $\{g(X,\Theta), i = 1,\ldots\}$ where Θ_i are iid random vectors and each tree casts a unit votes for the most popular class at input X (Breiman, 2001).

RF consists of many decision trees based on random selection of data and features. The predictor variables are selected randomly into sets and are used for building a decision tree. The random decision trees create a forest. The large number of decision trees provide high prediction accuracy.

3.7 Performance measure

Comparison in performance of classification models is checked using confusion matrix to find accuracy, specificity and sensitivity. A confusion matrix is a table of predicted class and actual class.

Confusion matrix

	Predicted No	Predicted Yes
Actual No	True Negative	False Positive
Actual Yes	False Negative	True Positive

True Positive (TP) are the positive class classified correctly as positive. True Negative (TN) are the negative class predicted corrected as negative. False Positive (FP) are the negative class incorrectly predicted as positive. False Negative (FN) are the positive class incorrectly predicted as negative.

Accuracy, sensitivity and specificity are performance metrics of a classification model.

Accuracy measure the chance that a classifier predicts the positive and negative elements correctly.

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

Sensitivity measure the chance of predicting the positive class correctly.

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

Specificity measure the chance of predicting the negative class correctly.

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

Precision is a measure of positive elements that were actually correct.

$$Precision = \frac{TP}{TP+FP}$$

3.8 Random survival forest algorithm

Random Survival Forest (RSF) is an extension of Breiman's RF techniques to survival settings allowing efficient non-parametric analysis of time to event data (Ishwaran et al. 2008). Its used to analyze right-censored survival data.

The Random Survival Forest algorithm is as outlined below (Ishwaran et al. 2008):

1. The method begins by randomly drawing B bootstrap samples from the original data of size n. The sample size of the bootstrap is approximately two-third of the original

data and the remaining sample which is about 37%, called out-of-bag (OOB)data is excluded from the sample.

2. For each bootstrap sample drawn, grow a full size survival tree depending on the chosen splitting criterion without pruning. At each node, randomly select *mtry* candidate predictors out of all P predictors. The node is split using the candidate predictors that maximize survival difference between the daughter nodes. The default splitting rule is log-rank and the candidate variable size $mtry = \sqrt{p}$.

3. The survival tree is grown to full size under the restraint that the terminal node should have no less than $d_0 > 0$ unique deaths.

4. Calculate the cumulative hazard function (CHF) for each tree. The CHF for each terminal node is approximated by Nelson-Aalen Estimator.
   $\hat{H}_h(t) = \sum_{t_{l,h} \leq t} \frac{d_{l,h}}{Y_{l,h}}$

   Where $d_{l,h}$ and $Y_{l,h}$ are number of deaths and number of individuals at risk at time $t_{l,h}$. The CHF for all observations within same node are equal. Calculate the mean of CHFs for all B trees to obtain the ensemble CHF of the forest.

   $\hat{H}_E(t|x) = \frac{1}{B} \sum_{i=1}^{B} \hat{H}_i(t|x)$

   Where $\hat{H}_i(t|x)$ represents the CHF of the tree grown in the i-th bootstrap sample.

5. By using out-of-bag (OOB) data, compute the predictor error for the ensemble CHF.

Node splitting

Binary survival trees are developed by repeated splitting of tree nodes. A tree develops at the root node which is the top of the tree consisting all the original data. Based on the set survival criterion, the root node splits to two daughter nodes, a left and right daughter nodes. Successively, each daughter nodes split into left and right daughter nodes. The process is repeated in each successive node.
The best split is one that gives the greatest survival difference between the daughters. This best split is obtained by searching over all possible p variables and split value c, and selecting x^* and c^* that gives the greatest survival difference. By ensuring the greatest survival difference, the survival tree moves apart the differing cases. In the end, as the number of nodes rise, and heterogeneous cases are separated, every node in the tree

becomes homogeneous and is made up of cases with similar survival. At the end, the survival tree reaches a saturation point when no new daughter nodes are formed as the criterion that every node must have no less than $d_0 > 0$ unique deaths has been reached. The uttermost nodes in a saturated tree are called terminal nodes.
Let $(T_{1,h}, \delta_{1,h}), \vdots, (T_{n(h),h}, \delta_{n(h),h})$ be the survival time and $0-1$ censoring status of an individual in terminal node h. When $\delta_{i,h} = 0$, it signifies that an individual i is right censored at time $T_{i,h}$ and $\delta_{i,h} = 1$ signifies that an individual i experienced event of interest at time $T_{i,h}$.
Log rank splitting rule was used as the splitting criterion in this study.

Log rank splitting criterion

The log rank statistics is obtained as below:

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

This measures the extent of separation between two daughter nodes. The maximum value of $|L(x,c)|$ gives the best split.

RSF provides a measure of variable importance (VIMP). The highly predictive covariates are chosen from the balanced data sets. The chosen variables are then fitted in a Cox-PH regression model to determine the effect of each variable.

3.9 Determination of risk factors of under-five child mortality

Cox- PH regression model is often used to determine the effect of risk factors on survival duration. The model is represented as

$$h(t,X) = h_0(t)\exp\{\sum_{i=1}^{p}\beta_i X_i\}$$

This models shows the hazard of an individual at time t given the set of predictors. $h_0(t)$ is the baseline hazard, X is a vector of predictors and B is a vector of regression coefficient. When $X = 0$, the risk equals to the baseline hazard i.e. $h(t,X) = h_0(t)$

3.9.1 Checking Cox-PH regression assumptions

Before fitting a Cox-PH regression models, certain assumption have to be met. These include

- Proportional hazard assumption- use Schoenfeld residuals to test
- Non-linearity relationship between log hazard and predictors- use Martingale residuals to check
- Outliers or influential observations- use dfbeta values to assess

4 Results

4.1 Balancing techniques

The sample sizes of our study after different balancing techniques are shown in table below.

Balancing method	Status	Total	Percentage
Under-sampling	Censored	169	50%
	Uncensored	169	50%
	Total	338	100%
Over-sampling	Censored	2747	50%
	Unsensored	2747	50%
	Total	5494	100%
Both sampling	Censored	1469	50.22%
	Unsensored	1456	49.78%
	Total	2925	100%

The different balancing techniques resulted in different sample sizes. The under-sampling method had the smallest sample size of 338 where both the mortality and non-mortality class had 169 samples. The table illustrate the overall balance in the mortality and survival classes after the different balancing methods.

4.2 Feature selection

Three filter methods were implemented in our study: chi-squired, information gain and ReliefF attribute selection to select the best subset of the original data that has 151 covariates. The number of subsets selected were 80, 100, 120 and 130 variables. The imbalanced class was then balanced using random over-sampling, random under-sampling and both sampling. Random Forest was used to check the performance of each subset for every filter methods and resample methods. Table 4.1 shows the results of accuracy, specificity and sensitivity for the different number of features selected in the imbalance and balanced classes.

Table 2. Accuracy, Sensitivity and Specificity in implementing Random forest classification for three filter methods on different number of subsets comparing the imbalanced class and the balanced class using under-sampling, over-sampling and both sampling

		Accuracy				Sensitivity				Specificity			
No. of features	Feature selection method	Imbalanced	under-sampling	over-sampling	both sampling	Imbalanced	under-sampling	over-sampling	both sampling	Imbalanced	under-sampling	over-sampling	both sampling
80	Chi-squired	96.93%	90.78%	98.07%	98.52%	40%	97.78%	66.67%	80%	100%	90.41%	99.76%	99.52%
80	Information gain	96.93%	92.83%	97.72%	98.41%	40%	100%	60%	82.22%	100%	92.45%	99.76%	99.28%
80	Relief F	96.59%	87.71%	97.72%	96.93%	35.56%	100%	66.66%	80%	99.88%	87.05%	99.4%	97.84%
100	Chi-squired	96.59%	91.35%	97.27%	98.07%	33.33%	97.78%	51.11%	80%	100%	90.77%	99.76%	99.04%
100	Information gain	96.47%	89.76%	97.61%	98.07%	31.11%	97.78%	57.78%	75.56%	100%	89.33%	99.76%	99.28%
100	Relief F	96.47%	88.96%	97.5%	97.95%	31.11%	100%	62.22%	82.22%	100%	88.37%	99.4%	98.8%
120	Chi-squired	96.02%	91.24%	96.13%	96.93%	22.22%	97.78%	35.56%	62.22%	100%	90.89%	99.4%	98.8%
120	Information gain	96.13%	88.74%	96.36%	97.16%	26.67%	100%	40%	57.78%	99.88%	88.13%	99.4%	99.28%
120	Relief F	96.25%	88.62%	96.7%	97.27%	26.67%	100%	46.67%	71.11%	100%	88.01%	99.4%	98.68%
130	Chi-squired	95.9%	86.92%	96.13%	96.93%	20%	97.78%	35.56%	62.22%	100%	86.33%	99.4%	98.8%
130	Information gain	95.68%	90.33%	96.02%	96.59%	17.78%	100%	33.33%	57.78%	99.88%	89.81%	99.4%	98.68%
130	Relief F	95.68%	89.42%	96.25%	96.47%	15.56%	100%	37.78%	53.33%	100%	88.85%	99.4%	98.81%

Chi-squired and information gain begun with an equal accuracy of 96.93% for the group with 80 features and reduced to 95.9% and 95.68% respectively for the groups with 130 features. ReliefF on the other hand had an accuracy of 96.59% for the subset with 80 features which was lower compared to the other two methods and reduced at a lower pace until 95.68% for the subset with 130 features. Balancing techniques applied seemed to have a significant effect on the accuracy. Over-sampling and both sampling increased the accuracy while under-sampling decreased the accuracy. This is because resampling methods increased minority class by over-sampling and decreased the majority group by under-sampling. Sensitivity on the imbalanced class was very low and decreased as the variables were added. Under-sampling gave the highest sensitivity and was the same for every subset. Information gain and ReliefF recorded 100% sensitivity for each subset.Sensitivity in both sampling ranked second. Chi-squired had equal sensitivity of 80% for the groups with 80 and 100 variables and decreased to 62.22% for both 120 and 130 features. On the other hand, sensitivity for information gain decreased from 82.22% for 80 variables to 75.56% for 100 variables then to 57.78% for both 120 and 130 variables. Over-sampling gave lower sensitivity and decreased as more variables were added. Specificity was high for the imbalanced class. Resampling also gave high specificity. Over-sampling gave the highest followed by both sampling then under-sampling. Specificity in the three re-sampling methods for each subset decreased for the 80 and 100 variables but was constant for the 120 and 130 variables. The three filter methods returned the best results with

fewer variables. This was because they repeatedly searched for the best subsets that gave best performance with less predictors.

4.3 Variable selection using RSF

The filter methods applied in the study had subsets with varying variables. Since the subset with 80 variables gave the best results, each subset from the three methods were used in the Random Survival Forest. The results are in the Table 4.2, 4.3, 4.4 and 4.5. For each RSF model, 1000 trees were grown for every data sets. Bootstrap samples were drawn for each original data set with the sample sizes as shown in the table. The size of each bootstrap is the re-sample size shown in the table and varies depending on the sample size of the original data and the balancing method used.

The 1000 bootstrap sample form the root node of the tree. At each split point, 15 out of 79 variables were selected to grow the tree. The root nodes spit to two daughter nodes and each daughter node split repeatedly thus maximizing survival difference between daughter nodes. Node splitting progresses until every tree is fully grown and the terminal nodes have no less than 15 varying events. If a sample size has larger number of events, the trees will be larger. This implies that the larger the number of events, the higher the average number of terminal nodes and the smaller the error rate. Over-sampling method which has the highest sample size records the smallest error rate while under-sampling which has the smallest number of occurrence has the largest error rate. The sample sizes for each group varies but the number of variables are equal.

The error rate in all the imbalance data set was smaller compared to the other balanced data set. It was also noted that the error rate in the imbalanced data set before feature selection was smaller and increased after feature selection. RSF model after information gain feature selection resulted in the highest error rate. Comparison of error rate in the balanced data set after the three feature selection methods, chi-squired methods recorded the highest error rates, followed by information gain method. ReliefF method resulted in the smallest error rates.

Table 4.6, 4.7 and 4.8 give the identified variables and the variable importance after feature selection and balancing methods. The higher the variables importance, the higher the association between the variable and the target response and the higher the predictive power of the variable. Under-sampling methods identified the highest number of variables in the three tables. Both sampling and over-sampling which had a higher sample size resulted in few significant variables. It is notable that the significant variables identified in both sampling and over-sampling were similar. Over-sampling method following ReliefF feature selection did not yield significant variables.

Table 3. Application of RSF on balanced data sets

Description	Imbalanced dataset	Under-sampling	Over-sampling	Both sampling
Sample size	2047	248	3846	1923
Number of deaths	124	124	1923	946
Number of trees	1000	1000	1000	1000
Forest terminal node size	15	15	15	15
Average no. of terminal nodes	92.315	10.965	162.562	83.972
No. of variables tried at each split	13	13	13	13
Total no. of variables	151	151	151	151
Resample size used to grow trees	1294	157	2431	1215
No. of random split points	10	10	10	10
Error rate	21.77%	42.89%	32.98%	33.37%

Table 4. Application of RSF on balanced data sets after chi-squared feature selection

Description	Imbalanced dataset	Under-sampling	Over-sampling	Both sampling
Sample size	2047	248	3846	1923
Number of deaths	124	124	1923	946
Number of trees	1000	1000	1000	1000
Forest terminal node size	15	15	15	15
Average no. of terminal nodes	93.848	11.195	159.432	82.538
No. of variables tried at each split	9	9	9	9
Total no. of variables	79	79	79	79
Resample size used to grow trees	1294	157	2431	1215
No. of random split points	10	10	10	10
Error rate	26.88%	50.65%	41.96%	41.33%

Table 5. Application of RSF on balanced data sets after ReliefF feature selection

Description	Imbalanced dataset	Under-sampling	Over-sampling	Both sampling
Sample size	2047	248	3846	1923
Number of deaths	124	124	1923	946
Number of trees	1000	1000	1000	1000
Forest terminal node size	15	15	15	15
Average no. of terminal nodes	91.09	10.606	159.556	82.589
No. of variables tried at each split	9	9	9	9
Total no. of variables	79	79	79	79
Resample size used to grow trees	1294	157	2431	1215
No. of random split points	10	10	10	10
Error rate	23.31%	42.16%	35.71%	36.93%

Table 6. Application of RSF on balanced data sets after Information gain feature selection

Description	Imbalanced dataset	Under-sampling	Over-sampling	Both sampling
Sample size	2047	248	3846	1923
Number of deaths	124	124	1923	946
Number of trees	1000	1000	1000	1000
Forest terminal node size	15	15	15	15
Average no. of terminal nodes	93.528	11.217	158.969	82.54
No. of variables tried at each split	9	9	9	9
Total no. of variables	79	79	79	79
Resample size used to grow trees	1294	157	2431	1215
No. of random split points	10	10	10	10
Error rate	27.19%	50.02%	41.97%	42.42%

Table 7. Important variables selected after different balancing techniques following chi-squared feature selection (selected variables had variable importance of > 0.002)

	Balancing methods					
	Under-sampling		Over-sampling		Both sampling	
	Variable	Importance	Variable	Importance	Variable	Importance
1	HW70	0.0163	HW70	0.0025	V207	0.0037
2	HW71	0.0094	V206	0.0023	HW70	0.0035
3	V207	0.0251	V207	0.0023	V206	0.0035
4	B1	0.0056	HW71	0.0020	HW71	0.0023
5	V137	0.0050	B4	0.0020	B4	0.0020
6	H33	0.0050				
7	V467F	0.0050				
8	H4	0.0041				
9	H6	0.0034				
10	H7	0.0030				
11	V467B	0.0028				
12	B4	0.0025				
13	V219	0.0024				
14	M71	0.0022				
15	H8	0.0022				
16	H2	0.0021				
17	V414S	0.0020				
18	V151	0.0020				
19	V106	0.0020				

Table 8. Important variables selected after different balancing techniques following Relief feature selection (selected variables had variable importance of > 0.002)

	Balancing methods					
	Under-sampling		Over-sampling		Both sampling	
	Variable	Importance	Variable	Importance	Variable	Importance
1	V206	0.0525			V207	0.0032
2	V207	0.0375			B4	0.0029
3	V137	0.0158			V137	0.0027
4	B4	0.0154			V206	0.0022
5	V210	0.0116				
6	H7	0.0079				
7	M17	0.0078				
8	B0	0.0070				
9	V461	0.0059				
10	V613	0.0056				
11	V202	0.0051				
12	V413	0.0048				
13	V414S	0.0047				
14	V414N	0.0043				
15	V411	0.0038				
16	V467B	0.0035				
17	V417	0.0031				
18	V463A	0.0024				
19	V414L	0.0024				
20	V414V	0.0023				

Table 9. Important variables selected after different balancing techniques following Information gain feature selection (selected variables had variable importance of > 0.002)

	Balancing methods					
	Under-sampling		Over-sampling		Both sampling	
	Variable	Importance	Variable	Importance	Variable	Importance
1	V206	0.0417	HW70	0.0030	V207	0.0048
2	V207	0.0372	V206	0.0024	V206	0.0035
3	HW70	0.0124	V207	0.0022	HW70	0.0029
4	V467F	0.0087	HW71	0.0022	B4	0.0023
5	B4	0.0086			HW71	0.0023
6	V137	0.0083				
7	H11	0.0080				
8	HW71	0.0059				
9	V465	0.0049				
10	V203	0.0047				
11	V233	0.0044				
12	B0	0.0043				
13	B1	0.0040				
14	V213	0.0032				
15	H3	0.0029				
16	V128	0.0029				
17	B15	0.0023				
18	M14	0.0022				
19	V367	0.0020				
20	V208	0.0020				

4.4 Determining the variable effects

To determine the effect of the variables, we fitted a Cox-PH regression model. The assumptions of the model were first tested before fitting the model.

4.4.1 Test for proportional hazard assumption

Table 4.9 shows the results for proportional hazard test. The global test show the overall state of proportional hazard violation. The variables V106, V137, V202, V206, V414S, H11, H33, V207, V151, V467F, H33 and M14 had a p.value < 0.05. The low p.value indicated significance of the test hence the variables did not meet the proportional hazard assumption and thus, were omitted from the model. The variables in the table satisfy the proportional hazard assumption as the high p.value > 0.05 support non significance of the test.

In Fig 4.1, the dashed line represent the ± 2 standard error band around the fit while the continuous line is a smoothing line fit to the plot. There are no systematic pattern around the horizontal line which indicates proportion hazard as there is no pattern with time. All the variables in the table thus supported proportional hazard assumption.

4.4.2 Test for influential observations

Comparing the magnitude of largest dfbeta values in the index plot in Fig 4.2 in relation to the regression coefficients indicated that none of the variables were highly influential than others. Nonetheless, the dfbeta values for V465 and V213 were high compared to the rest.

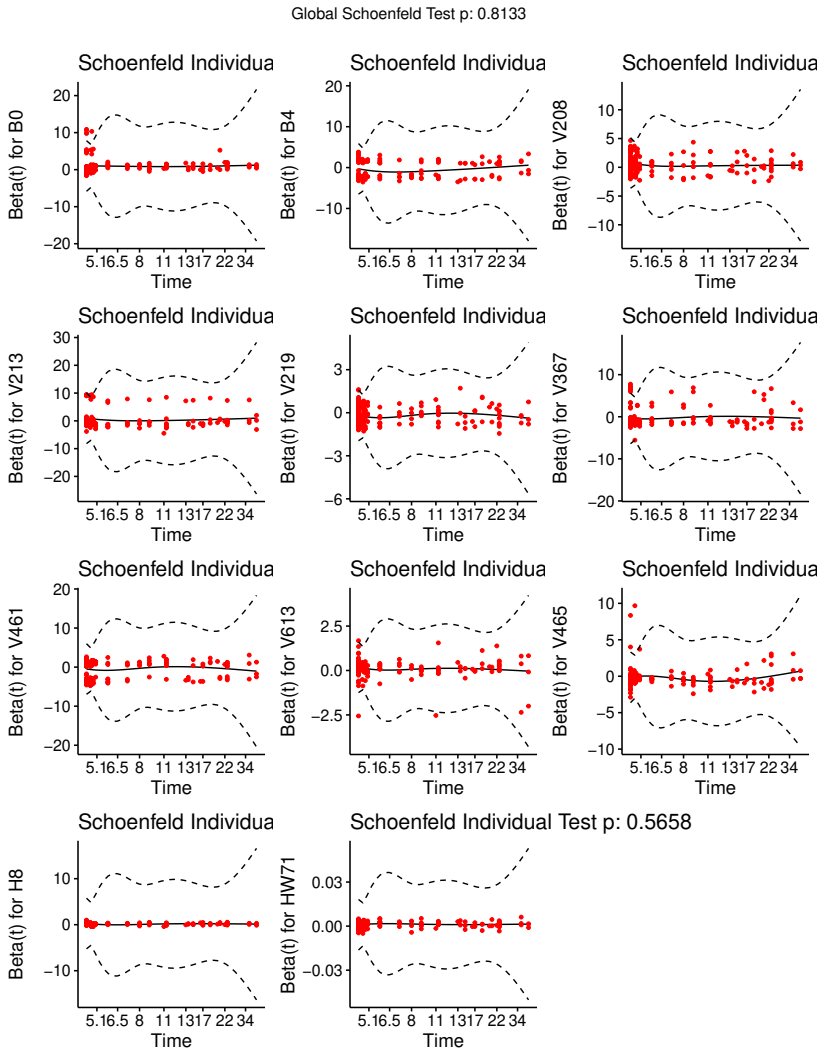
4.4.3 Parameter estimates

After different diagnostics test on Cox-PH regression model, the predictors were fitted to test the effect of each variable all at once on survival time. The results of the fit are shown on Table 4.10. The column named "Coefficient" give the estimated logarithm of the hazard ration while the column named "Exp(coefficient)" gives the estimate of the actual hazard ratio. The column "Exp(coefficients)" is thus important in interpretation. The positive coefficients increase the risk of death and thus, reduction in the mean survival time while negative coefficients decrease the risk of death and increases the mean survival time. The positive coefficients are crucial in determining the risk factors of under-five child mortality since they are positively related to the mortality occurrence and thus determine the mean survival time. All the predictors in Table 4.10 have positive coefficients. The Table 4.10 show that 11 predictors increase the risk of death. A HR > 1 increases risk of death, HR < 1 decreases risk of death while HR $= 1$ implies that the covariate has no influence on the survival time. The column named "Pr($> |z|$)" shows the value of the

Table 10. Test for proportional hazard assumption

No.	Variable	Chi-square	P-value	No.	Variable	Chi-square	P-value
1	B0	0.0103	0.919	16	B1	1.71	0.190
2	B4	0.0160	0.899	17	B15	0.813	0.367
3	V128	0.241	0.624	18	V130	1.19	0.276
4	V203	0.961	0.327	19	V208	1.60	0.206
5	V213	0.00676	0.934	20	V219	0.00731	0.932
6	V233	1.77	0.183	21	V367	0.330	0.566
7	V411	0.516	0.472	22	V413	0.000753	0.978
8	V414L	2.74	0.098	23	V414N	1.31	0.253
9	V461	0.08	0.777	25	V613	1.01	0.314
10	V465	0.381	0.537	26	V467B	0.498	0.481
11	H2	0.00153	0.969	26	H3	0.306	0.580
12	H4	0.362	0.547	27	H6	0.55	0.459
13	H7	0.174	0.676	28	H8	0.244	0.622
14	M17	0.376	0.540	29	M18	0.00651	0.936
15	HW70	0.152	0.697	30	HW71	0.125	0.724
	GLOBAL	0.354	0.229				

Figure 1. Schoenfeld residuals for variables

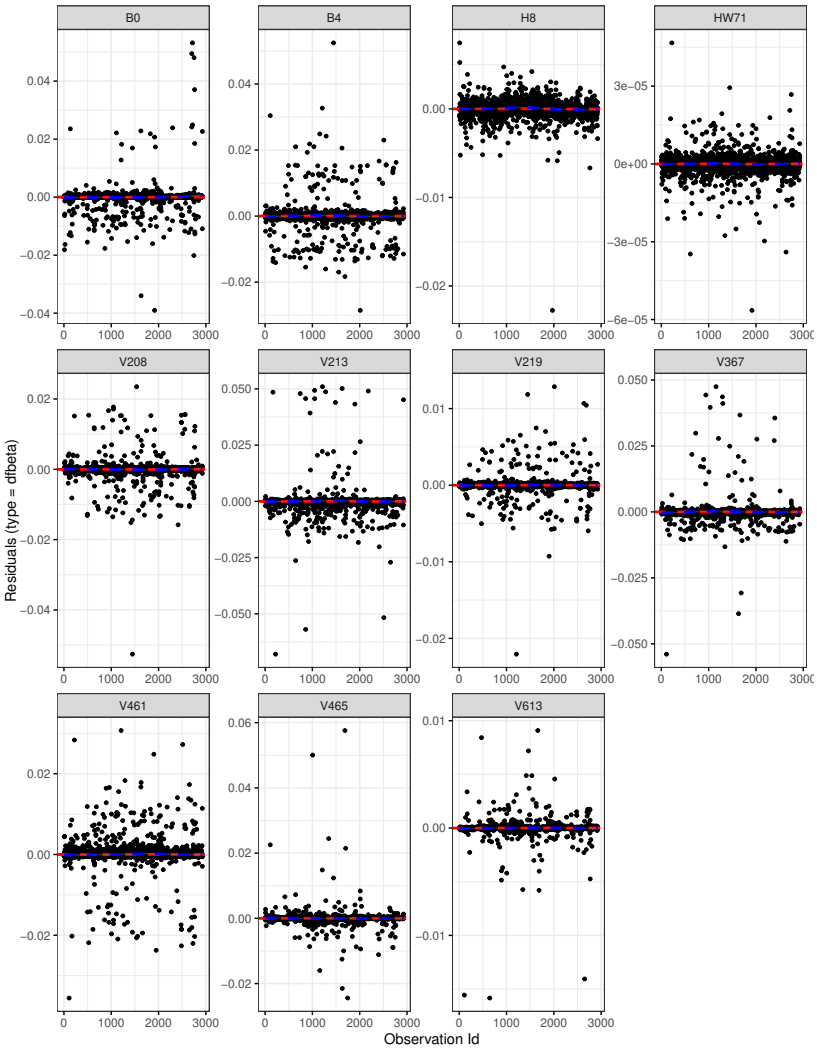


Wald statistic. Wald statistic examines whether the predictors in a model are significant. A p.value < 0.05 implies that a variable is significant.

Table 11. Results of fitting the covariates on Cox-PH regression model

Covariate	Coefficient	Exp(coefficient)	Se(coefficient)	Pr($>|z|$)
B0	0.9049	2.4717	0.1865	1.22e-06
B1	-0.0422	0.9587	0.0239	0.0772
B4	-0.4778	0.6202	0.1720	0.0055
B15	1.4326	4.1896	1.0888	0.1883
V128	-0.0140	0.9861	0.0207	0.4992
V130	-0.0159	0.9842	0.2161	0.9414
V203	0.0641	1.0662	0.0916	0.4843
V208	0.4395	1.5520	0.1262	0.0005
V213	0.6249	1.8680	0.2602	0.0163
V219	-0.2350	0.7905	0.0591	6.98e-05
V233	0.0377	1.03845	0.1759	0.8300
V367	-0.3430	0.7097	0.1676	0.0408
V411	0.3053	1.3570	0.2110	0.1480
V413	0.1521	1.1643	0.1814	0.4018
V414L	-0.0431	0.9578	0.2109	0.8380
V414N	0.4588	1.5821	0.2649	0.0833
V461	-0.5765	0.5618	0.1782	0.0012
V613	0.1036	1.1091	0.0386	0.0072
V465	-0.2384	0.7879	0.1009	0.0181
V467B	-0.4232	0.6549	0.3831	0.2693
H2	-0.2553	0.7747	0.3210	0.4264
H3	-0.8146	0.4428	0.5162	0.1146
H4	-0.9328	0.3934	0.5993	0.1196
H6	-0.0382	0.9625	0.5156	0.9409
H7	0.4007	1.4930	0.3367	0.2339
H8	0.6618	1.9382	0.2821	0.0190
M17	0.2511	1.2854	0.3729	0.5007
M18	0.0652	1.0673	0.1327	0.6232
HW70	-0.0008	0.9992	0.0004	0.0890
HW71	0.0014	1.0014	0.0005	0.0045

Figure 2. Index plot of dfbeta values



5 Discussion

The research main aim is to identify the risk factors of under-five child mortality. The Kenya DHS 2014 data set which is a great resource on health indicators was used in the study. A subset of 151 variables were selected for study. The key advancement in our study is the handling of the three main components. Imputation of missing data using *MissForest* algorithm, Feature selection to remove redundant variables and reduce computational cost and removal of class imbalance. Several studies such as Tagoe et al (2020), Zewudie et al (2020) and Nasejje & Mwambi (2017) among others have made use of the DHS data set. The study results showed that the significant risk factors for under-five child mortality were anthropometry of the child(HW70,HW71), reproduction history(V202, V203, V206, V207, V208, V210, V233, V219), birth history (B0, B1, B4), mother's characteristic (V106, V128, V137, V151), maternity and feeding (V411, V413, V414L, V414N, V414S, V414V, V417, V461, V463A, V467B, V467F), immunization and child health (H2, H3, H4, H6,H7, H11) and maternal health (M14, m71). The results were consistent with (Waititu et al 2020) where the dependent variables identified were number of dead sons, number of dead daughters, number of daughters living and number of children living.

Besides variable selection using RSF, feature selection methods were explored and the performance analyzed. The three filter methods considered were chi-squared, ReliefF and information gain. The three methods gave good results with fewer variables as they repeatedly searched for the subsets that performed best. This was not in line with (Khaldy & Kambhampati 2017) as our study showed that information gain feature selection performs best as more features are added.

The balancing techniques used were under-sampling, over-sampling and both sampling. Under-sampling method seemed to perform best as it identified more significant variables compared to over-sampling and both sampling. This was not matching (Waititu et al 2020) study where over-sampling and both sampling identified the majority of the significant variables. Given that under-sampling included reduction of the feature space by decreasing the prevalent class, this did not lead to loss of information in the rare class which was the focal point.

6 Conclusion

The study details a framework for determining risk factor of under-five child mortality. This include feature selection, class balancing methods, variable selection using RSF and parameter estimation using Cox-PH model. The problems of high dimensionality and class imbalance were examined along with the feature selection methods and balancing methods. The factors of under-five child mortality included anthropometry of children (such as height for age standard deviation), reproduction factors of the mother (such as number of births in the last five years), birth history (such as child is twin), mother's characteristics (such as highest education level of the mother), maternity and feeding (such as sleeping under a mosquito net), immunization and child health (such as whether a child received BCG immunization) and maternal health (such as cesarean session).

Appendix

Table 12. Variable names of the significant variables

Category	Variable	Description
Birth history	B0	Child is twin
	B1	Month of birth
	B4	Sex of child
	B15	Live birth between births
Maternal health	M14	Number of antenatal visits during pregnancy
	M17	Delivery by caesarean section
	M71	Time after delivery postnatal check took place
Immunization	H2	Received BCG
	H3	Received DPT 1
	H4	Received POLIO 1
	H6	Received POLIO 2
	H7	Received DPT 3
	H8	Received POLIO 3
	H11	Had diarrhea recently
	H33	Received Vitamin A1 (most recent)
Mother's history	V106	Highest educational level
	V128	Main wall material
	V137	Number of children 5 and under in household (de jure)
	V151	Sex of household head
Reproduction history	V202	Sons at home
	V203	Daughters at home
	V206	Sons who have died
	V207	Daughters who have died
	V208	Births in last five years
	V210	Births in month of interview
	V213	Currently pregnant
	V233	Months when pregnancy ended
Contraception use	V367	Wanted last child
	V411	Gave child tinned, powdered or fresh milk
	V413	Gave child other liquid
	V414L	Gave child any other fruits
	V414N	Gave child fish or shellfish

Bibliography

[1] Afrin, K., Illangovan, G., Srivatsa, S. S., & Bukkapatnam, S. T. (2018). Balanced random survival forests for extremely unbalanced, right censored data. arXiv preprint arXiv:1803.09177.

[2] Batista, G. E. P. A., Prati, R. C, & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. SIGKDD Explor. Newsl. 6, 1 (June 2004), 20–29. DOI:https://doi.org/10.1145/1007730.1007735

[3] Ayiko, R., Antai, D. & Kulane, A. Trends and determinants of under-five mortality in Uganda. East Afr J Public Health. 2009 Aug;6(2):136-40. PMID: 20000017.

[4] Waititu, H. W., Koskei, J. K. & Onyango, N. O. 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

[5] Emmert-Streib, F. & Dehmer M. Introduction to Survival Analysis in Practice. Machine Learning and Knowledge Extraction. 2019; 1(3):1013-1038. https://doi.org/10.3390/make1030058

[6] Wang, H., & Li, G. (2017). A Selective Review on Random Survival Forests for High Dimensional Data. *Quantitative bio-science*, 36(2), 85–96. https://doi.org/10.22283/qbs.2017.36.2.85

[7] Breiman, L. Random Forests. Machine Learning 45, 5–32 (2001). https://doi.org/10.1023/A:1010933404324

[8] Khaldy, M. A. I. & Kambhampati, C. Resampling imbalanced class and the effectiveness of feature selection methods for heart failure dataset. Int Rob Auto J.2018;4(1):37?45. DOI: 10.15406/iratj.2018.04.00090

[9] Ehrlinger, J. (2016). ggRandomForests: Exploring random forest survival. arXiv preprint arXiv:1612.08974.

[10] Nasejje, J. B., Mwambi, H. Application of random survival forests in understanding the determinants of under-five child mortality in Uganda in the presence of covariates that satisfy the proportional and non-proportional hazards assumption. BMC Res Notes 10, 459 (2017). https://doi.org/10.1186/s13104-017-2775-6

[11] Tagoe, E. T., Agbadi, P., Nakua, E. K., Duodu, P.A., Nutor, J.J. & Aheto, J. M. K. A predictive model and socioeconomic and demographic determinants of under-five mortality in Sierra Leone. Heliyon. 2020 Mar 6;6(3):e03508. doi: 10.1016/j.heliyon.2020.e03508. PMID: 32181389; PMCID: PMC7063153.

[12] Zewudie, A. T., Gelagay, A. A. & Enyew, E. F. Determinants of Under-Five Child Mortality in Ethiopia: Analysis Using Ethiopian Demographic Health Survey, 2016. Int J Pediatr. 2020 Sep 18;2020:7471545. doi: 10.1155/2020/7471545. PMID: 33029153; PMCID: PMC7527934.

[13] Gebretsadik, S. & Gabreyohannes, E. (2016). Determinants of Under-Five Mortality in High Mortality Regions of Ethiopia: An Analysis of the 2011 Ethiopia Demographic and Health Survey Data. International Journal of Population Research. 2016. 1-7. 10.1155/2016/1602761.

[14] Acheampong, G. K. & Avorgbedor, Y. E. Determinants of under Five Mortality in Ghana; A Logistic Regression Analysis Using Evidence from the Demographic and Health Survey (1988-2014). American Journal of Public Health Research. 2017; 5(3):70-78. doi: 10.12691/ajphr-5-3-4

[15] Boerma, J. T. & Sommerfelt, A. E. Demographic and health surveys (DHS): contributions and limitations. World Health Stat Q. 1993;46(4):222-6. PMID: 8017081.

[16] Iddrisu, A. K., Tawiah, K. ,Bukari F. K. & Kumi, W. (2020). Frequentist and Bayesian Regression Approaches for Determining Risk Factors of Child Mortality in Ghana. BioMed Research International. 2020. 1-10. 10.1155/2020/8168479.

[17] Dejene, B.D. (2016). Levels, Trends and Determinants of Under-Five Mortality in Amhara Region, Ethiopia Using EDHS (2000 -2011). Journal of Health, Medicine and Nursing, 28, 73-83.

[18] Aheto, J. M. K. Predictive model and determinants of under-five child mortality: evidence from the 2014 Ghana demographic and health survey. BMC Public Health 19, 64 (2019). https://doi.org/10.1186/s12889-019-6390-4

[19] Fikru, C., Getnet, M., & Shaweno, T. (2019). Proximate Determinants of Under-Five Mortality in Ethiopia: Using 2016 Nationwide Survey Data. Pediatric health, medicine and therapeutics, 10, 169–176. https://doi.org/10.2147/PHMT.S231608

[20] Nasejje, J. B., Mwambi, H. G. & Achia, T. N. O. 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, 1003 (2015). https://doi.org/10.1186/s12889-015-2332-y

[21] Ayele, D. G., Zewotir, T. T. & Mwambi, H. Survival analysis of under-five mortality using Cox and frailty models in Ethiopia. J Health Popul Nutr 36, 25 (2017). https://doi.org/10.1186/s41043-017-0103-3

[22] Unicef (2020). Levels & Trends in Child Mortality. Estimates developed by the UN Inter-agency Group for Child Mortality Estimation. United Nations Children's Fund

[23] Leevy, J. L., Khoshgoftaar, T. M., Bauder, R. A. et al. A survey on addressing high-class imbalance in big data. J Big Data 5, 42 (2018). https://doi.org/10.1186/s40537-018-0151-6

[24] Sonak, A., & Patankar, R. (2015). A Survey on Methods to Handle Imbalance Dataset.

[25] Afrin, K., Illangovan, G., Srivatsa, S. S., & Bukkapatnam, S. T. (2018). Balanced random survival forests for extremely unbalanced, right censored data. arXiv preprint arXiv:1803.09177.

[26] Bichindaritz, I., & Quinn, T. P. (2017). Feature Selection for Survival Analysis in Bioinformatics. BAI@IJCAI.

[27] Yap, B. W., Ibrahim, N., Hamid, H. A., Rahman, S. A. & Fong, S. (2018). Feature selection methods: Case of filter and wrapper approaches for maximising classification accuracy. pertanika journal of science and technology, 26, 329-340.

[28] Lee, I. H., Lushington, G. H., & Visvanathan, M. (2011). A filter-based feature selection approach for identifying potential biomarkers for lung cancer. Journal of clinical bioinformatics, 1(1), 11. https://doi.org/10.1186/2043-9113-1-11

[29] Ramyachitra, D., & Manikandan, P. (2014). IMBALANCED DATASET CLASSIFICATION AND SOLUTIONS : A REVIEW.

[30] Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., Bing, G. Learning from class-imbalanced data: Review of methods and applications,Expert Systems with Applications,Volume 73, 201 7, Pages 220-239, ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2016.12.035. (https: www.sciencedirect.com/science/article/pii/S0957417416307175)

[31] Ishwaran, H., Kogalur, U. B., Blackstone, E., & Lauer, M. (2008). Random survival forests. The Annals of Applied Statistics, 2, 841-860, September 2008. https://doi.org/10.1214/08-AOAS169

[32] Uguz, H. (2012). A hybrid system based on information gain and principal component analysis for the classification of transcranial Doppler signals. Computer methods and programs in biomedicine, 107 3, 598-609 .