University of Nairobi

School of Mathematics

ISSN: 2410-1397

Master Project in Biometry

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, Number 15, 2019

Kennedy Sifuna Wanyonyi

June 2019

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

Master Project in Biometry **University of Nairobi**

June 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, Number 15, 2019

Kennedy Sifuna Wanyonyi

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

0.1 Background

Survival trees and Random Survival Forests are extensions of classification and regression trees and random forests in analyzing time to event data. These methods are alternatives to Cox Proportional hazards models when Proportional Hazard assumption is violated. Survival tree methods are flexible and can handle high dimensional covariate data as they are fully non-parametric. Random survival forests use the Brieman's approach, first, by employing a random selection of a bootstrap sample used for growing a tree then growing tree learners by splitting the nodes on the randomly selected predictors. The performance of the survival trees highly depended on the splitting method that is applied while growing the tree.

0.2 Methods

In our analysis, we compare the performance of random survival forests in variable selecting based on the following split rules; Log-rank splitting, Log-rank score splitting, and and Conditioned Inference Forests. Our outcome variable is the under-five child mortality in Kenya using 2014 DHS data. Covariates that were included in the models were chosen based on the existing literature.

0.3 Results

Findings from this study show that Log-rank split rule outperforms Log-rank score split rule. Both split rules analyze time to event data based on the bootstrap cross-validated estimates for integrated Brier scores.

0.4 Conclusion

As much as it is evident that Log-rank is the best, there is need to investigate other split rules and the nature of data that that best suit each split rule to be able to identify the best slitting method.

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.

Signature Date

KENNEDY SIFUNA WANYONYI
Reg No. I56/9315/2017

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

Signature Date

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

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

Figures and Tables

Figures

Tables

Acknowledgments

First, I thank the almighty God for enabling me to complete my studies.

Secondly, DELTAS Africa Initiative SACCAB for funding my studies.

A huge gratitude to mu supervisors, Dr. Nelson Owuor and Dr. Rachel Sarguta.

Finally, I'm grateful to all who supported me through different ways during my studies.

Kennedy Sifuna Wanyonyi

Nairobi, 2017.

1 Introduction

1.1 Survival Analysis

Survival analysis is a collection of statistical techniques that were developed for analyzing biological data, with a time to event outcome that is usually censored (Wang and Li, 2017; Weathers and Cutler, 2017). There are three main types of censoring;left censoring, interval censoring and right censoring, right censoring being the most common. A subject is right censored if they die or drop out before the end of the study or completes the study before experiencing the event of interest, the exact time at which the event occurs is not known. An example, in a HIV study where a new antiviral drug is believed to be good at viral suppression; if the subject's viral load is not suppressed at their time of death or at the end of the study then this subject is right censored. Interval censoring is the case where the event happens within the period under investigation but the exact time at which the event occur ed is not known. Left censoring is the least common, the exact time at which the event occur ed is not known but the event happens before the period of investigation. The survivor function, the probability that a subject survived beyond a certain time point, t, is an important function in survival analysis, i.e $S(t) = P(T > t) = 1 - F(t)$, where $F(t)$ is the cumulative distribution function of the lifetime, T. The analysis of the survival function can e either parametric or non-parametric based on the validity of model assumptions (Weathers and Cutler, 2017).

1.1.1 Kaplan-Meier Estimator

This is a non-parametric estimator of the survival function that estimates the number of events that have occurred for each unit of time and the time taken until the event occurs. Kaplan-Meier estimate can also be used in prediction of some specific survival outcomes, an example, the period taken to viral supression for a subject on a specific antiretroviral drug. The survival curves, that are made up of the Kaplan-Meier estimates, are of a stepwise form, such that the values of the survival probability estimate, y, only change for times at which we actually observe the occurrence of an event, or the censoring of a survival time. The value of the survival curve is held constant for the time between two observed events.

Assuming the outcome of interest is death; let t_i denote the time at death occurred for i^{th} individual where $i \subset 1, \ldots, n$, n is the total number of subjects at risk. Let d_i denote the be the number of individuals who died t_i and r_i be the number of subjects at risk at a time before t_i.

Then,

$$\hat{S}(t) = \begin{cases} 1 & if t_j < t_1 \\ \prod_{j=1}^{n} \left[1 - \frac{d_j}{r_j}\right] & \text{if t} \leq t_j \leq t_n \end{cases}$$

Alternatively written as:

$$\hat{S}(t) = \prod_{t_j \leq t} \left(1 - \frac{d_j}{r_j}\right)$$

The variance of the Kaplan-Meier estimator is given by:

$$\hat{V}[\hat{S}(t)] = [\hat{S}(t)]^2 \sigma_S^2(t) = [\hat{S}(t)]^2 \sum_{i=1}^{i} \frac{d_j}{r_j\left(r_j - d_j\right)}$$

While computing Kaplan-Meier survival curves for several groups, an example treatment A and treatment B, estimation of $\hat{S}(t)$ is done separately for each group then use the log-rank test to compare the survival functions for the two groups (Weathers and Cutler, 2017).

1.1.2 Cox Proportional Hazards Model

This is a semi-parametric regression technique for analysing survival data by relating several predictors, considered simultaneously, to survival time. Hazard rate, the risk of getting the outcome event (i.e dying) after a subject has survived to a specific time, is the measure of effect. The hazard represents the expected number of events per unit time. The component related to the predictor variables is parametric while the estimate of the survival function component is fully non-parametric thus making Cox model semi-parametric. The Cox model allows both categorical and continuous predictors. Categorical predictors with more than 2 levels have to be converted into a series of binary classes in order to perform the regression.

Before fitting a Cox model the following assumptions must be satisfied:

- Censoring must be statistically independent of the failure times

- The ratio of the hazards for any two groups should be constant over time, i.e, they are proportional. Kaplan-Meier curves can be used for checking this assumption.
- Nonlinear covariate relationships, assumes that each variable makes a linear contribution to the model.

Right censoring, the commonly used, the time to to the occurrence of the event is not obtained due to various reasons. The possible reasons being a subject getting to the end of the study before experiencing the outcome of interest, a subject dying, or in clinical trials, a subject can be withdrawn from the study due to severe or life-threatening adverse events. If a subject censored due to being discontinued from the study due to adverse events or protocol violations, this kind of a scenario does not meet the assumptions of the Cox model.

According to (Weathers and Cutler, 2017), survival analysis examines the association between predictors and the survival distribution is of particular interest. This be done by specifying a model for the log hazard (Crumer, A, 2008). The hazard function for a parametric model based on an exponential distribution can be represented,

$$logh_i(t) = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik}$$

this can also be expressed as ,

$$h_i(t) = exp(\alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik}),$$

$$\text{where=}\begin{cases}\text{Observation are denoted by i}\\ \text{Predictors are denoted by x's}\\ \alpha \text{ is a constant representing the log-baseline hazard for the model.}\end{cases}$$

A vector parameter β is estimated by the partial likelihood,

$$L(\beta) = \Pi_{i=1}^{D} \frac{exp[\beta x_i]}{\sum_{j \subset R(t_i)} exp[\beta x_j]}$$

, D is the total number of observed event times.
The unspecified baseline hazard function in Cox model $\alpha(t) = logh_i(t)$, having set all

predictors to 0, the log-hazard function becomes,

$$logh_i(t) = \alpha(t) + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik},$$

resulting to a hazard function,

$$h_i(t) = h_o(t) exp(\beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} k),$$

leasing to a survival function,

$$S_i^{cox}(t) = exp(h_o(t) exp(\beta_1 x_{i1} + \beta_2 x_{i2} + + \beta_k x_{ik} k)).$$

It can be shown that Cox model is a proportional hazards model, taking any pair of observations, i and j, in which the values of the predictors are different. Taking the linear covariates, for these two observations,

$$n_i = \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} k \text{ and } n_j = \beta_1 x_{j1} + \beta_2 x_{j2} + ... + \beta_k x_{jk} k,$$

the ratio of the two hazard functions is shown by

$$\frac{h_i(t)}{h_j(t)} = \frac{h_0(t) exp[n_i]}{h_0(t) exp[n_j]} = \frac{exp[n_i]}{exp[n_j]} = exp[n_i - n_j],$$

and is independent of time. Thus, the ratio of the hazard functions is strictly proportional to the (exponential of) differences in values of the predictor variables.

1.2 Random Survival Forests

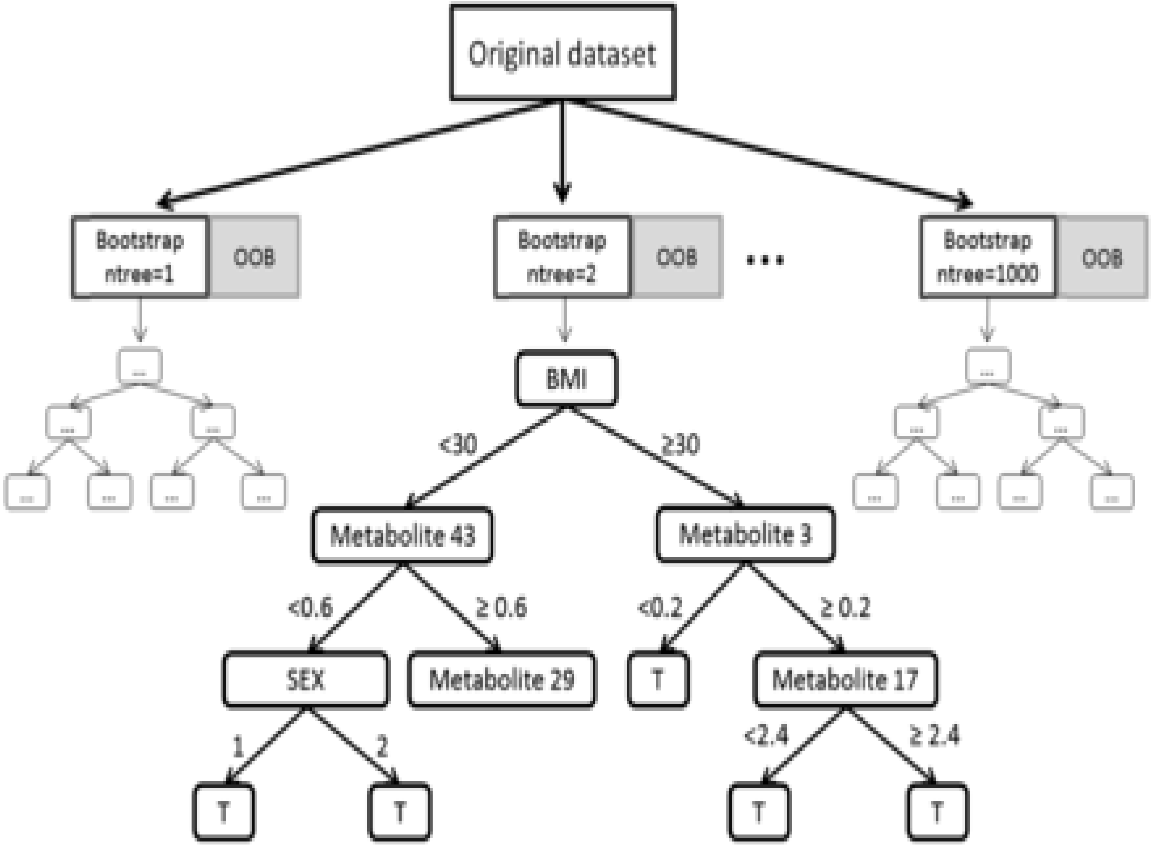
Cox proportional hazard is commonly applied in analyzing right censored data. This assumes proportional hazards which is often violated (Ishwaran and Udaya, 2007; Ng'andu, NH, 1997; Nasejje et al., 2017). Survival trees and Random Survival Forests borrow the concepts of classification and regression trees and random forests in analyzing survival data. These methods are alternatives to Cox Proportional hazards models when Proportional Hazard assumption is violated. Survival tree techniques are fully non-parametric thus making them flexible to handle high dimensional covariate data (Nasejje et al., 2017) Random survival forest is an ensemble tree method applied in analysis of right censored

survival data. Using trees as base learners in construction of ensembles can significantly improve the learning performance. Introducing randomization at the base learning stage can boast ensemble learning-Random forests (Breiman et al., 2017). Random survival forests use the Brieman's approach, first, by employing a random selection of a bootstrap sample used for growing a tree then growing tree learners by splitting the nodes on the randomly selected predictors. The performance of the survival trees highly depended on the splitting method that is applied while growing the tree (Nasejje et al., 2017)
Performance of RSF compares closely to methods like bagging and boosting. Notable features on RSF are (i) Easy to build as only three features are required- no. of selected parameters, number of trees and splitting criteria. (ii) they are highly adaptive thus overcoming restrictive assumptions in the semi-parametric models like proportional hazards (Ishwaran and Udaya, 2007).
The random survival forest algorithm is given by:

1. Draw ntree bootstrap samples from the original data.
2. Grow a tree for each bootstrapped data set. At each node of the tree randomly select mtry covariates for splitting on using a specified split rule. The nodes should be split in a way that increases between the node heterogeneity and increases within the node homogeneity thus maximizing survival differences across daughter nodes.
3. Grow the tree to full size under the constraint that a terminal node should have no less than node size unique deaths.
4. Calculate an ensemble cumulative hazard estimate by combining information from the ntree trees. One estimate for each individual in the data is calculated
5. Compute an out-of-bag (OOB) error rate for the ensemble derived using the first b trees, where b = 1, . . . ,***ntree***.

Diagrammatically, the above steps can be summarised as shown in the figure 4 below:

Figure 1. Illustration of the computation of Random Survival Forests using 1000 bootstrap samples



1.3 Statement of the problem

Random Survival Forests (RSF) is robust to many statistical assumptions. However the splitting criteria often used is the logrank test statistic. This splitting rule assumes proportional hazards in the survival data. Conditional Inference Forests (CIF) (Wei, Fu and Simonoff, S., 2017), has been used analyze determinants of time to even data, as a way of trying to overcome this challenge. Based on one unique dataset, do we gain efficiency by using CIF over and above RSF?

1.4 Objectives

1.4.1 Overall objective

The broad objective of this study is to apply survival analysis using machine learning on Demographic Health Survey to identify the risk factors to under-five child mortality in Kenya. Random Survival Forests machine learning technique under several split rules will be used.

1.4.2 Specific objects

1. Selection of predictors of under-five child mortality using log-rank, log-rank score split rules in RSF.
2. Apply conditional inference method in selecting the predictors.
3. Identify the best split rule based on the error rates of the three models above.

1.5 Justification of the study

The studies that have been done aimed at shrinking the rate of child mortality in Africa with various suggestions for improvement. According to UNICEF, malnutrition, infectious diseases such as tetanus, pneumonia, diarrhea, and meningitis resulted in many deaths in 2015. This situation can only be improved by employing public insights to take informed action to offer curative and preventive measures in Kenya that demands the utilization of large under-5 mortality data sets from 2014 DHS and these data sets comprises many variables as well as regression approaches. The nature of this information poses a statistical challenge such as handling for multicollinearity and correlation for multiple testing, among others. These bottlenecks have only been scarcely solved and thus limiting current statistical methods. This research, therefore, seeks to bridge the gap by employing multivariate classification method, mainly the random survival forest to identify the risk factors of under-five mortality.

2 Literature review

2.1 Child Mortality

According to (Khodaee et al., 2015), child mortality is a core indicator of both child health and well-being. Despite the improvement of the survival of under-five children between 1990, 2015, the goal of the fourth-millennium development goals target of at least a two-thirds decrease in the mortality rate of under 5 is yet to be achieved globally. Out of 5.9 million deaths of under-five children, 2.7 million took place in the neonatal period in 2015 were attributed to the preterm birth complications, pneumonia, and intrapartum-related events. According to WHO, however, this rate has decreased by 58% with regards to the estimation of 93 deaths per 1000 live births recorded in 1990 to about 39 deaths in 1000 lives births registered in 2017. In the pursuit towards the achievement of SDGs to end preventable deaths among the children under five years and newborns, the international community has developed a new framework that adopted by 117 member states and expected to achieve the target by 2030. According to (UNICEF, 2019), the elimination of preventable child deaths required concrete details about the current distribution of major causes of under-5 mortality that vary with time.

(UNICEF, 2019), estimation of the child mortality by cause between 2000-13 as well as the cause-specific scenarios to 2030 and 2035 established that sub-Saharan Africa will exhibit 33% of the births and 60% of the deaths different from 25% and 50% respectively in 2013. This region has the highest percentage of neonatal deaths, with 1.1 million newborns dying in the first month. In the entire African continent, the likelihood of a child under five years dying is about seven times higher compared to the WHO European Region. In addition to this, the existence of inequities in the under-five mortality between the low-income and high-income communities remain large in Africa. A significant difference exists in the distribution of the under-5 deaths by cause across different regions of Africa. West and Central Africa recorded the highest rate of one in seven children dying before five years in 2008. Despite the global progress towards the reduction of child mortality over the past decade, child mortality remains high concerning the millennium development goals. Kenya recorded 46.37 deaths in every 1000 live births in 2018, which represent a significant reduction of under-5 mortality from 164.34 deaths in 1000 live births in 1969. (King, B. E., and Rice, J., 2018). Their investigation of trends in child mortality shows a downward trend in the under-5 mortality rate, child mortality rate, and infant mortality rate between 1993 and 2008 in both urban and rural areas. This decline was more statistically significant in the rural areas, unlike in the metropolitan regions, posing a gap in the urban-rural differentials narrowed over time (King, B. E., and Rice, J., 2018). Similar trends are portrayed in slum areas between 2003 and 2010, revealing a decrease from 113 to 79, 33 to 24, and 83 to 57 for U5MR, CMR, and IMR respectively. According to Hastie et al. (2005), under-five mortality is majorly caused by neonatal sepsis. Pneumonia causes a more significant number of deaths among under-5 children globally, worse in developing countries where there is limited access to effective and low-cost alternatives and other clinical services. Over 15% of the newborn deaths in Africa are linked to the infections associated with the delivery process in 2010. Globally, 11% of the under-5 children die from diarrhea, with 90 percent of the deaths taking place in Sub-Saharan Africa.

Research by (Kanmiki et al., 2014) reveals that the under-5 mortality rate is dependent on the educational level of mothers, marital status, age, and presence of co-wives. Mothers with junior high school or primary education exhibit 43% lesser chances of encountering under-5 deaths contrary to mothers lacking formal education. Malaria accounted for over 500,000 deaths of under-5 children worldwide in 2011, where a more significant fraction is experienced in sub-Saharan Africa. In 2010, 6% of the deaths in Africa were linked to HIV, especially in countries where prevalence is high such as Swaziland and South Africa.

Under-five mortality is considered a critical indicator of the state of a society's public health, and thus, an array of studies has been done aimed at creating an optimal intervention for improving child survival by 2030. (Hastie et al., 2005), focused on generating high-resolution estimates for the neonatal and under-5 all-cause mortality across 46 African countries. The findings underscore a more vital benefit of tracking geospatially granular patterns in child survival. (Kanmiki et al., 2014), explored the relationship between demographic and socio-economic factors and the under-5 mortality in the impoverished areas in the rural northern Ghana which revealed that causes of a high rate of under-five death are complex and thus demands concerted efforts to enable clarifying the implications to boost child survival.

2.2 Variable Selection

Variable section is fundamentally crucial to high dimensional modeling that includes both regression and classification. The most suitable variable selection technique can improve the accuracy of estimation by precisely identifying the subset of relevant and vital predictors as well as enhancing the interpretability of the model with parsimonious representation (Liu et al., 2016). Different variable selection techniques are tailored to establish the minimal set of the strongest predictors linked to the under-5 mortality and are hence useful in the identification of potential diagnostic, prognostic or prognostic biomarkers of the risk factors. The commonly used variable selection techniques include SBRT, MS Prime, Stochastic gradient boosting, Random Forests, Linear regression, and Support Vendor Machine (Shi et al., 2018).

Linear regression utilizes forward, backward elimination, stepwise, R-squared, or All-possible subsets to identify the best subset from a set of many variables include within the mode. (Hitziger, M., and Ließ, M., 2014) describe gradient boosting model, (2014), as a technique in which boosting draws the bootstrap of predictor data samples, fits the tree, and eliminates the prediction from the original data. In (Yu, L., and Liu, H., 2003), SBRT technique is a hybrid feature selection algorithm that relies on Boruta and SVM-RFE and combines the benefits of both wrapper and filter that has proved most effective in the identification of tsAPA sites in rice.

Variable selection is continuously evolving from regression and classification of independent data structures to multilevel techniques capable of handling dependent data structures that enable better dissection of treatment effects (Yu, L., and Liu, H., 2003). The focus on the linear model started in the 1960s when vital developments occurred, and computing was costly. The traditional idea of best selection methods of a subset is computationally expensive for most modern statistical applications that have prompted incorporation of machine learning to cope with the high dimensionality. These techniques have been utilized widely in simultaneously selecting required variables and estimating their impacts in obtaining a high dimensional statistical inference. Today, statistical procedures are anchored on three critical pillars, including model interoperability, statistical accuracy, and computational complexity.

The commonly used variation selection techniques, including linear regression, suffers a significant drawback. Despite clear solutions to the issues of selection bias, more advanced solutions are yet to be developed to deal with variable selection using multiple model classes. Also, the problems increase with the demand for data mining of massive data sets. Availability of numerous variable selection procedures and various justifications poses the likelihood of getting misled. Random survival forest technique bridges this gap, especially when handling high or ultra-high dimensional data due to its ability to deal with a large number of predictor variables for time-to-event data. Compared with regression-based techniques, random survival forests technique is entirely data-driven and therefore independent of model assumptions. RSF employs a model which better explains the data and thus it is the most suitable tool in the exploratory analysis of under-5 mortality where information is limited (Hastie et al., 2005). Random forest, unlike regression techniques, is free from limitations such as overfitting, inflated standard errors, and unreliable estimation of the regression coefficients (King, B. E., and Rice, J., 2018). In the research aimed at developing a hybrid method for waveband selection and classification of the hyperspectral data, (Liu et al., 2016), demonstrated a higher classification accuracy when a union of RFE and Boruta selected wavebands are used in the analysis. As per (Liu et al., 2016), Random Forest has been successfully used in genetic, metabolomics, gene expression, proteomics, and methylation and hence suitable in predicting quantitative traits associated with under-five child mortality.

3 Split rules

3.1 Split Rule Notations

Assume that we are node h in the process of growing a tree and we want to split node h into two daughter nodes. Assume that node h is of size n. Let σi be an indicator variable and Ti be the time variable. $\sigma[i] = 1$ is death occurred at T[i] and 0 if and individual is right censored.

The split of node h on a predictor is given in such a way that $x > c$ or $x <= c$, observations that satisfy the first criterion are placed into one sub-node and those which don't are grouped into a second sub-node. This creates two sub-nodes at every criterion that is evaluated until all the possible criteria are evaluated. This kind of splitting this works well only if the covariates are time independent.

In the scenario where X_{ij} changes with time, $X_{ij}(t) < C$ for $t < t*$ but $X_{ij}(t) > C$ for $t > t^*$. In this case it is impossible to assign i to a node as there is no clear distinction. To handle time-varying covariates, every observation is split several pseudo-subjects based on the criterion $x(t) <= C$, in that every pseudo-subject represents a non-overlapping time interval and either $x(t) > C$ or $x(t) <= C$ in the entire interval. For observation i, the procedure splits the record at time t into two pseudo-subjects, one with $X_j(t) < C$ (since $< t^*$) and one with $X_j(t) >= C$ (since $t >= t^*$). These two pseudo-subjects can then go to separate sub-nodes.

Let $t_1 < t_2 < \; < t_N$ be the distinct death times in the parent node h, and let $d_{i,j}$ and $Y_{i,j}$ the number of deaths and individuals at risk at time t_i in the daughter nodes $j = 1,2$. $Y_{i,j}$ is the number of individuals in daughter j who are alive at time t_i, or who have an event (death) at time t_i. Let $Y_i = Y_{i,1} + Y_{i,2}$ and $d_i = d_{i,1} + d_{i,2}. Let n_j$ be the total number of individuals in node j. This means that, n = $n_1 + n_2$ (Wei, Fu and Simonoff, S., 2017; Nasejje et al., 2017).

3.1.1 Log rank split method

The method uses the log rank statistic, the best split is one with the largest value of log rank statistic. The disadvantage of this method is that it favors predictors with a larger number of split points. An example, a dataset with an outcome y variable and independent variables x_1 and x_2 where $n2 < n1$, in this case x_1 will have a higher chance of having a split point with a larger effect on y. Introduction of bias at the point of split point variable extends this bias to other parameters like the variable of importance (Wright et al., 2017). According to (Nasejje et al., 2017; Ishwaran and Udaya, 2007; Bou-Hamad et al., 2011; Ciampi at al., 1987; Segal MR,, 1988), the log-rank statistic for a predictor x at a split value c is given by:

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

The Log-rank Survival Tree Algorithm

This algorithm was proposed by (Nasejje et al., 2017):

1. Select $\sqrt{p-}$ predictors randomly as candidates for splitting at every node and split into two sub-nodes, α and β, p is the total number of covariates.

2. Calculate the logrank value impurity measure for the sister nodes and at node h.

3. Select the covariate with the highest statistically significant value of the statistic obtained from one of the sub-nodes obtained from the splits. The predictor with the highest value of the statistic is partitioned into two daughter nodes.

4. Treat each sub-node as a root node and iterate steps 2 and 3.

5. A node is considered as the terminal node if it has no $d_0 > 0$ unique observed events.

3.1.2 Log-rank split score rule

This split utilizes the log-rank cores to improve on the log-rank split method. Suppose the rank vector of survival times given by $r = (r_1, r_2, ..., r_N)$, and their indicator variable $(T, \sigma) = ((T_1, \sigma_1), (T_2, \sigma_2)...(T_N, \sigma_N))$ and $a = a(T, \sigma) = (a_1(r), a_2(r), ..., a_N(r))$ represents the score vector that depends on ranks in vector r. Suppose the ranks order the predictors in that $x_1 < x_2 < ... < x_N$.

At time T_l an observation is scored by:

$$a_l = a_l(T, \sigma) = \sigma_l - \sum_{k=1}^{Y_l(T)} \frac{\sigma_k}{N - Y_k(T) + 1}, \qquad (1)$$

where

$$Y_k(T) = \sum_{l=1}^{N} x[T_l \leq T_k]$$

accounts for the number of subject who experienced the outcome of interest or censored before or at time T_k

$$i(x, s^*) = \frac{\sum_{x_j \leq s^*}\left(a_j - R_1 \bar{a}\right)}{\sqrt{R_1\left(1 - \frac{R_1}{N}\right) S_a^2}}$$

$\bar{a}$ is the mean and S_a^2 the variance of the scores $\{a_j : j = 1, 2...n\}$.The best partition optimizes $i(x, s^*)$ on all $x'_j s$ and possible partitions s^*. The tree algorithm for implementing this split is same as the one used in the log-rank test above.

3.1.3 Split based on Conditional Inference Forests approach

As opposed to the log-rank score and log-rank methods, conditioned inference forests (CIF) separates the process of picking the best predictor to split on from that of picking the optimum split point (Nasejje et al., 2017). In the first step, the best-split variable is established by doing tests of association of all the predictors with the survival outcome variable by us of linear rank test based on log-rank scores. Based on the permutation tests, the predictor with the highest association with the outcome variable, the one with the smallest p-value, is picked for splitting (Nasejje et al., 2017; Hothorn T, Hornik K, and Zeileis A, 2006). The second step grows a binary tree is grown.

The entire forest is then grown because single trees are considered not stable (Nasejje et al., 2017; Wright et al., 2017; Hothorn T, Hornik K, and Zeileis A, 2006). According to (Nasejje et al., 2017; Hothorn T, Hornik K, and Zeileis A, 2006) the algorithm for growing conditioned inference forests is given by:

1. For case weights w, test for independence between any of the p predictors and the outcome variable, K. Stop if the null hypothesis cannot be rejected otherwise select the j^*th predictors X_{j^*} with highest association to K.
2. Select a set $B^* \subset X_j^*$ inorder to partition X_j^* into two distinct sets i.e. B^* and $X_j^* B^*$. The weights $w\alpha$ and $w\beta$ determine the two subgroups with $\alpha_{w,i} = w_i I\left(X_{j^*,i^*} \subset B^*\right)$ and $w\beta_{w,i} = w_i I\left(X_{j^*i^*} \neq B^*\right)$ for all $i = 1,2,...,n$.
3. Repeat the above steps with modified weights $w\alpha$ and $w\beta$, respectively.

3.1.4 Maximally selected rank split rule

Maximally selected statistic rank statistic (MSR-RF) is implemented in the conditional inference forests environment (Wright et al., 2017). In the CIF, the optimal variable for splitting is established by a linear rank statistic, whereas the best split point is based on a binary split, MSR-RF overcomes this challenge. The optimal split variable is identified through a dichotomous-based split statistical test, factoring in adjustments for multiple testing for several possible split points. This reduces the bias in the selection of the optimal variable selected for splitting. There is no requirement for multiple testing for binary split variables like gender (Wang and Li, 2017).

3.1.5 R-Squared split rules

According to (Strasser H, and Weber C, 1999), R^2 statistic can be used as prediction measure in a nonlinear models for right-censored time-to-event data. R^2 split method has three advantages:

1. When there is no censoring, the statistic converges to the classical coefficient of determination.
2. When predicting the conditional mean response for correctly specified models, R^2 is an accurate in approximating the non-parametric coefficient of determination.

3. R^2 can be applied to a variety of right-censored survival data even if the model is not correctly specified.

4 Methodology

In this section, were are going to explore our data and the methods that we used in the evaluating the models.

4.1 Random Survival Forests

The Random Survival algorithm is as discussed in **section 1.2** of **chapter 1**. Log rank and Log rank score split rules were used in the RSF models as detailed in **sections 3.1.1** and **3.1.2** respectively in **chapter 3**.

4.2 Data

Demographic Health Surveys are done in different countries in the world. The DHS program if funded by USAID and PEPFAR among other donors. In our analysis we used data a specific sample for Kenya, 2014 DHS. DHS data has information on the household wealth index, family planning, malaria, child nutrition among others, immunizations. Our study focuses on a subset of Kenya DHS data on the child section. The dataset contains information on women of reproductive age, 15-49years. Our study picked the child information section that contains children between 1-59 months with a total of 1099 variables and 20943 records. In our analyses we only picked 63 variables that are as predictors of child mortality based on (Khaoya,M.,, 2018; Nuwasiima,A.,, 2018).

4.3 Missing data imputation

The *R* software packages *randomForestSRC* and *party* have inbuilt algorithms for imputing missing data. The imputation is usually done at the parent node when splitting the selected variables (Ishwaran et al.,, 2008). For p candidate variables at node m, the following steps are used to impute that:

1. Before splitting, impute the missing values at every node m. Let $X^0_{r,m}$ denote observed data values for the r^{th} coordinate in the training data of X-covariates at node m. Denote $f\left(X^0_{r,m}\right)$ as the posterior distribution of $X^0_{r,m}$.
2. Draw from $f\left(X^0_{r,m}\right)$ at node m to impute missing data for every case in the training data for the r^{th} coordinate.
3. Using the imputed data split node$m$ using the specified split criteria in the model.
4. Reset the imputed values to missing in the daughter node after splitting the parent node m.
5. Iterate steps 1 to 4 until the tree reaches the saturation point.
6. Impute OOB using the same rule.

4.4 Variable Importance

Best predictors were selected based on variable importance measure. To compute VarImp for x, OOB cases are dropped in their training survival tree. Daughter nodes are assigned in a randomly whenever x split. Both ensemble CHF and resulting predicting error are computed. The difference between the new and original prediction errors yields VarImp. Larger values of VarImp shows the covariates are highly predictive.

4.5 Model evaluation

4.5.1 Ensemble estimation

The ensemble cumulative hazard function (CHF) has to be computed to enable the comparison of the effectiveness and precision of the different split rules that were used in building the models. This starts by computing the CHF at each and every node h in every survival tree. According to (Ishwaran and Udaya, 2007; Hong Wang,Xiaolin Chen, and Gang LI, 2018): Suppose (t_1, h) represent definite death times in h. Assume d_l, h is the number of deaths and Y_l, h individuals exposed to the risk factor at time t_l, h. At node h the CHF is given by

$$\hat{H}_h(t) = \sum_{t_{1,h} \leq t} \frac{d_l, h}{Y_l, h}$$

Every tree has a series of $\hat{H}_h(t)$ estimates, the number of these estimates equals to the number of nodes in a tree. For a subject i with predictor x_i,$\hat{H}_h(t)$ is computed by dropping x_i down the tree. This applies to both out and in of bag data. The desired estimator for i is achieved by the terminal node. This is given by:

$$\hat{H}_h(t\backslash x_i) = \hat{H}_h(t), \text{if} x_i \subset h \tag{1}$$

This estimate is just for one tree 1, the ensemble, CHF, is calculated by getting the average for all the trees. Let, $\hat{H}_b(t\backslash x)$ represent the estimate of the cumulative hazard 2.1 for tree b = 1....*ntrees* (Ishwaran and Udaya, 2007).Let an indicator variable $I_{i,b}$=0 when i does not belong to an out of bag for b and $I_{i,b}$=1 when i belongs to an out of bag point for b. The estimator of OOB ensemble cumulative hazard (Ishwaran and Udaya, 2007) for i is given by:

$$\hat{H}_e^*(t\backslash x_i) = \frac{\sum_{b=1}^{ntrees} I_{i,b}\hat{H}(t\backslash x_i)}{\sum_{b=1}^{ntrees} I_{i,b}}$$

This estimate is only for the bootstrap samples whee i is an out of bag value. The total CHF uses all data points from all the samples and thus given by

$$\hat{H}_e(t\backslash x_i) = \frac{\sum_{b=1}^{ntrees} \hat{H}_b(t\backslash x_i)}{\sum_{b=1}^{ntrees} I_{i,b}}$$

4.5.2 Concordance error rate

According to (Ishwaran and Udaya, 2007), $\hat{H}_e(t\backslash x_i)$ is the estimator for the ensembles OOB cumulative hazard estimator. To calculate the error rate, Harrell's Concordance index was used (Ishwaran and Udaya, 2007; Harrell, C., Pryor, L., and Rosati, 1982). To compute the Concordance index it is required that we define what makes up a worse predicted outcome. Let $t_1^*, ..., t_N^*$ represent all distinct survival times in the data.A subject i is considered to has a worse outcome than j if

$$\hat{H}_e^*(t_k^*\backslash x_i) > \hat{H}_e^*(t_k^*\backslash x_j)$$

Computation of concordance error rate is done by:

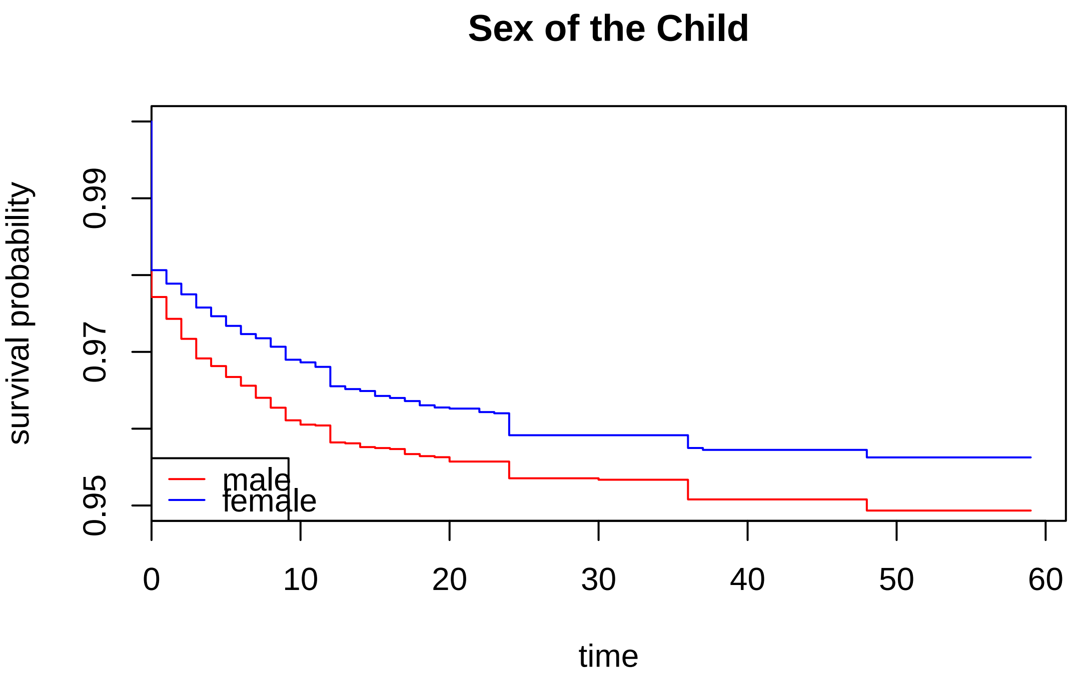
1. Check the data set and establish all possible pairs
2. Select the pairs with uncensored shorter survival time. If $T_i = T_j$, ignore the pairs i and j with exception of the case where one is an event(death). Let Permissible represent the number of permissible pairs.
3. In the event of a worse predicted outcome in the shorter event time, count 1 for every permissible pair. If there is a tie in the predicted outcome then count 0.5. Let Concordance be the total of all permissible pairs.
4. The concordance index C is defined as, $C = \frac{Concordance}{Permissible}$
5. Compute the error rate, Error=1-C. The error interval is between 0 and 1 with an error of 0 showing perfect accuracy while an error of 0.5 is just same as tossing the coin.

5 Results

5.0.1 Exploratory analysis using Kaplan-Meir plots

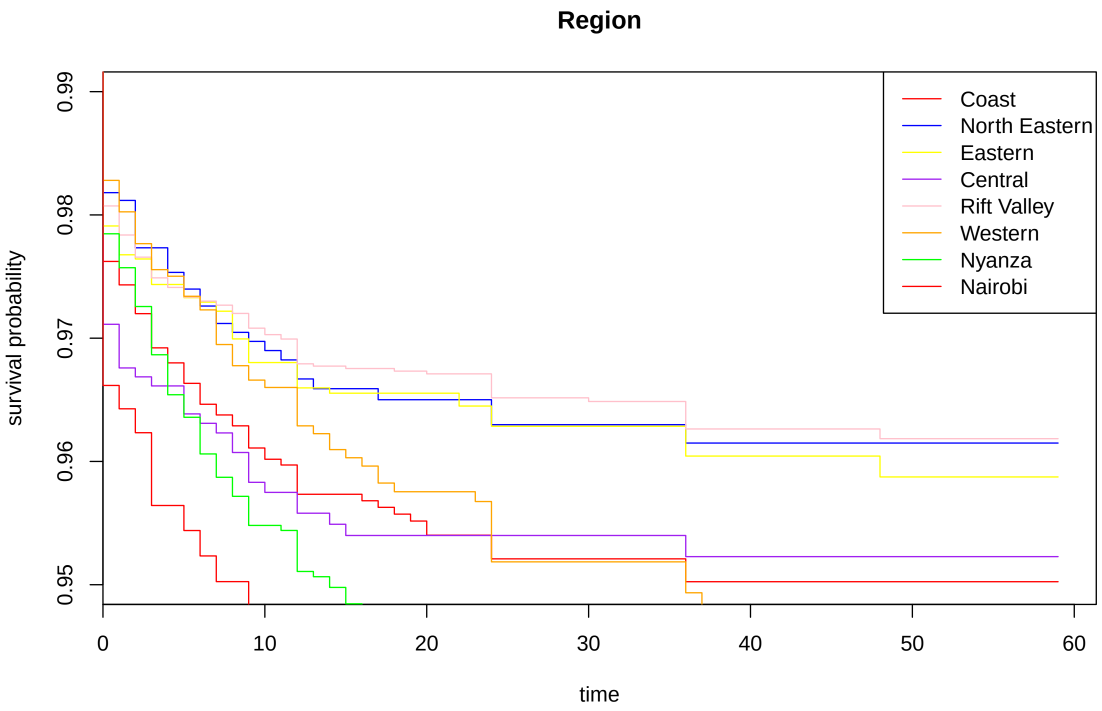
In order to explore the risk of death for the under-five children, we used Kaplan-Meir survival curves to explore the time to death for the children under two predictors, sex of the child and the region where the children live. Time to death was recorded in years but we converted it into months in our analysis.

Figure 2. Survival probabilities of the child by gender



Survival probabilities of female children is higher than that of male children across all the ages as shown in figure **4** above. This could be attributed to physiological and genetic disparities between girls and boys.

Figure 3. Survival probabilities of the child by region



As shown in figure 3, children in the coastal region of Kenya have the lowest survival probabilities. The children in the sample died before the age of 10 years. This could be as a result of high malaria prevalence in this region.

5.1 Summary results from log-rank and log-rank score models

Table 1. Summary characteristics of the fitted split rules

	Log-rank	Log-rank score
Sample size	20964	20964
Number of deaths	871	871
Was data imputed	yes	yes
Number of trees	1000	1000
Forest terminal node size	15	15
Average no. of terminal nodes	321.605	809.989
No. of variables tried at each split	8	8
Total no. of variables	63	63
Resampling used to grow trees	swor	swor
Resample size used to grow trees	13249	13249
Analysis	RSF	RSF
Family	surv	surv
Number of random split points	10	10
Error rate	5.35%	14.72%

Table1 shows the summary results from the three models. Log-rank and log-rank score split rules applied permutation importance to estimate the importance of risk factors in determining the under-five child mortality in Kenya. As discussed in **Section 3.3**, based on the error rate Log-rank model outperformed the Log-rank score model.

Figure 4. Prediction error rate and variable importance rankings for determinants of under-five child mortality in the log-rank model

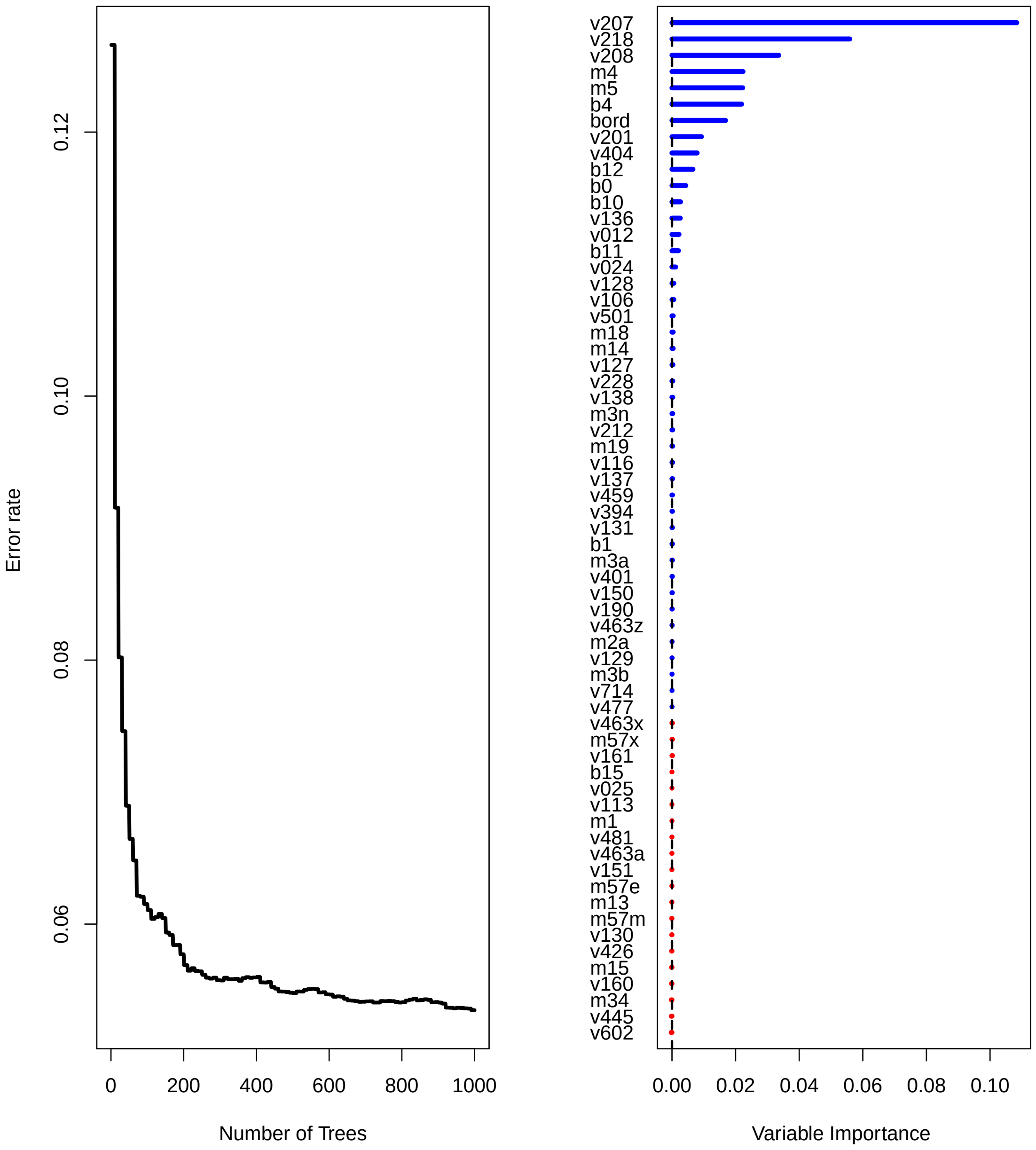
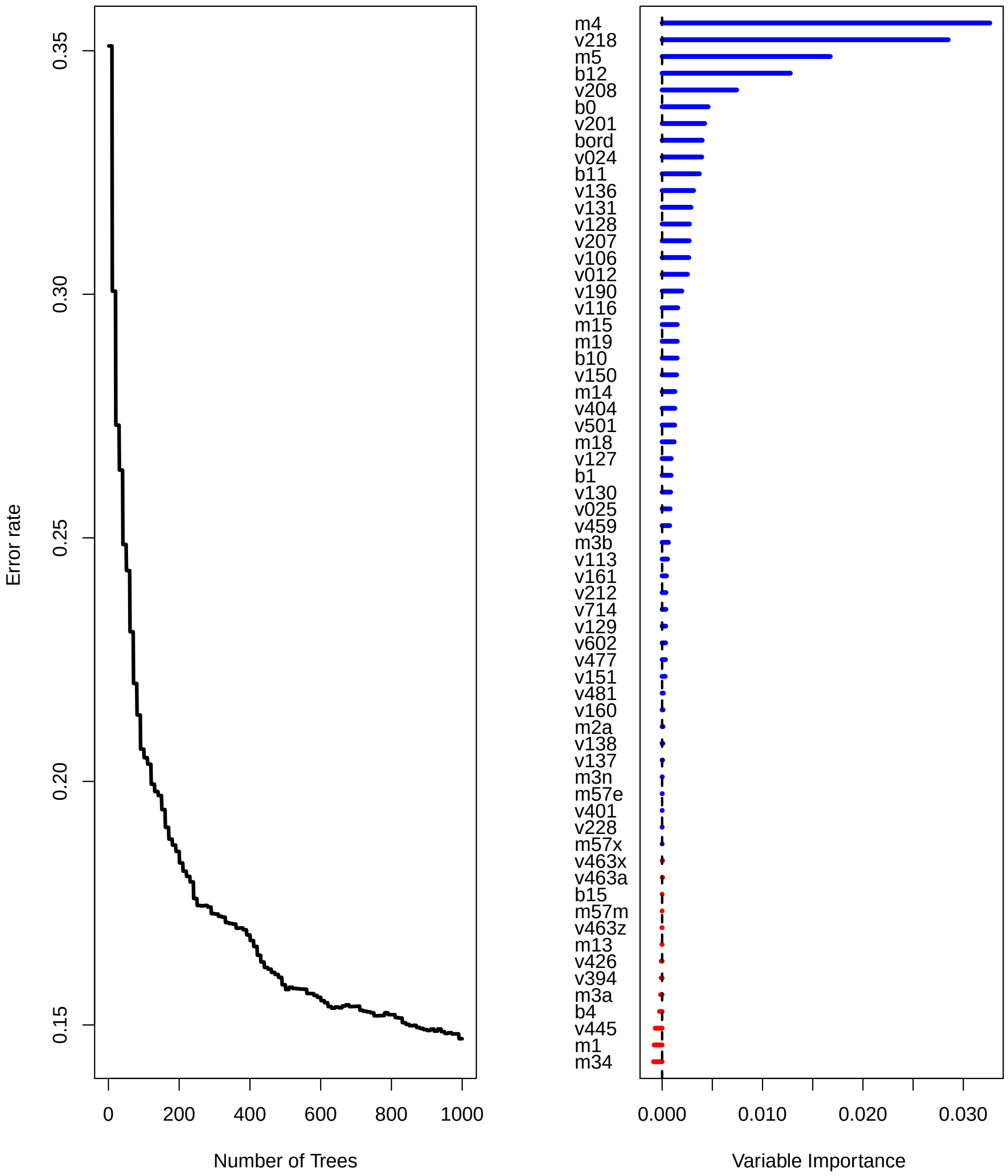


Figure 4 above shows the prediction errors and variable importance rankings from the log-rank score model. The error rate significantly drop after after the first 150 trees. Daughters who have died, number of children living and Births in the last five years are identified as the most influential predictors of under-five mortality in Kenya.

Figure 5. Prediction error rate and variable importance rankings for determinants of under-five child mortality in the log-rank score model



As shown in Figure 5 above, log-rank score split rule has a higher error rate than log-rank. The results from the variable importance with the log-rank model on Duration of breastfeeding as one of high risk factors of child survival in Kenya.

Table 2. Variable importance scores for under-five risk factors in the log-rank model

Risk factor	VarImp score
v207-Daughters who have died	0.1084
v218-Number of children living	0.0559
v208-Births in the last five years	0.0336
m4-Duration of breastfeeding	0.0223
m5-Months of breastfeeding	0.0222
b4-Sex of the child	0.0219
bord-Birth order number	0.0168
v201-Total children ever born	0.0093
v404-Currently breastfeeding	0.0079
b12-Succeeding birth intervals	0.0066
b0-Child is twin	0.0043
b10-Completeness of information	0.0027
v024-Region	0.0012
v128-House wall material	0.0006
v106-Highest level of education	0.0006
v501-Current marital status	0.0004
m18-Size of the child at birth	0.0004

Table 3. Variable importance scores for under-five risk factors in the log-rank score model

Risk factor	VarImp score
m4-Duration of breastfeeding	0.0326
v218-Number of children living	0.0285
m5-Months of breastfeeding	0.0167
b12-Succeeding birth intervals	0.0128
v208-Births in the last five years	0.0074
b0-Child is twin	0.0046
v201-Total children ever born	0.0042
bord-Birth order number	0.0040
v024-Region	0.0039
b11-Preceding birth interval in months	0.0037
v136-Number of household members	0.0031
v131-Ethinicity	0.0029
v128-House wall material	0.0027
v207 -Daughters who have died	0.0027
v106-Highest level of education	0.0027
v012-Current age of the mother	0.0025
v190-Wealth index	0.0020

Tables 2 and 3 above shows the first 17 top ranked under-five mortality risk factors. According to (Ishwaran et al.,, 2008) predictors with $VarImp < 0.002$ are important in predicting the outcome event. Log-rank model has 7 while log-rank score model has 3 highly predictive risk factors.

6 Discussion and Conclusion

6.1 Discussion

High rates of child mortality is a major public health challenge in developing countries especially in the Sub-Saharan region. Our research focuses on identifying the possible risk factors of deaths of children under the age of 5 years using log rank and log rank score split methods in Random Survival forests.

Random Survival Forests employs the concepts of classification and regression trees and random forests in analyzing survival data. The performance of the survival trees highly rely on the split rule used in growing survival trees (Nasejje et al., 2017). Daughters who have died, number of children living, births in the last five years, duration of breastfeeding, sex of the child and child birth order number were picked as the highly predictive factors of under-five child mortality in the log rank model while the log rank score model picked duration of breastfeeding, number of children living, months of breastfeeding and succeeding birth intervals.

The two models agree on duration of breastfeeding,number of children living and months of breastfeeding as the important predictors of under-five child mortality. This is consistent with the results from (Khaoya,M.,, 2018; Nasejje et al., 2017).

6.2 Conclusion

Findings from this study show that Log-rank split rule outperforms Log-rank score split rule. Both split rules analyze time to event data based on the bootstrap cross-validated estimates for integrated Brier scores.

6.3 Study Limitations

RSF algorithm does not run on missing data. It imputes data for all missing fields even in the cases where a skip pattern was used in collecting data.

6.4 Future Research

There is need to investigate other split rules and the nature of data that that best suit each split rule to be able to identify the best slitting method.

Bibliography

Hong Wang and Gang Li (2017). A Selective Review on Random Survival Forests for High Dimensional Data.Quant Biosci,36(2): 85–96.

Weathers Brandon and Cutler Richard (2017). Comparison of Survival Curves Between Cox Proportional Hazards.Random Forests,and Conditional Inference Forests in Survival Analysis". All Graduate Plan B and other Reports.

Crumer, A (2008).Comparison Between Weibull and Cox Proportional Hazards Models.

Hemant Ishwaran and Udaya B. Kogalur (2007).Random Survival Forests for R.

Ng'andu, NH (1997). An empirical comparison of statistical tests for assessing the proportional hazards assumption of cox's model. Stat Med,16(6):611–26.

Breiman L, Friedman J, Stone CJ, Olshen RA (2017). Classification and regression trees. Belmont: CRC press.

Wei Fu, Jeffrey S. Simonoff (2017). Survival trees for left-truncated and right-censored data, with application to time-varying covariate data. Statist. Med,(36)1272-1284.

Justine B. Nasejje, Henry Mwambi, Keertan Dheda2 ,Maia Lesosky (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.

Marvin N. Wright Theresa Dankowskia and Andreas Ziegler (2017). Unbiased split variable selection for random survival forests using maximally selected rank statistics.

Bou-Hamad I, Larocque D, Ben-Ameur H (2011). A review of survival trees.StatSurv,5:4471.

Ciampi A, Chang CH, Hogg S, McKinney S (1987). Recursive partition: A versatile method for exploratory-data. analysis in biostatistics. In: Biostatistics. New York: Springer, 23–50.

Segal MR (1988). Regression trees for censored data. Biometrics,35–47.

Hothorn T, Hornik K, Zeileis A (2006). Unbiased recursive partitioning: A conditional inference framework. J Comput Graph Stat,15:651–74.

Strasser H, and Weber C (1999). On the asymptotic theory of permutationistics. Math Methods Stat,8:220–50.

Hong Wang,Xiaolin Chen, and Gang LI (2018). Survival Forests with R-Squared Splitting Rules. Journal of Computational Biology,(25):4

Fabian Eifler and Mathias Schmid (2014). Introduction of AUC-based splitting criteria to random survival forests.Institute for Statistics,Universiy of Munich.

Harrell, C., Pryor, L., and Rosati (1982). Evaluating the yield of medical tests.JAMA, 247(18):2543-2546.

Ishwaran, Hemant; Kogalur, Udaya B.; Blackstone, Eugene H.; Lauer, Michael S (2008). Random survival forests. Ann. Appl. Stat,(3):841-860.

Khaoya Mutaki (2018). Application of Random Survival Forests and Accelerated Failure Time Shared Frailty Models in Understanding Under-Five Child Mortality in Kenya.University of Nairobi.

Afra Nuwasiima (2018). Multiple Imputation and Random Survival Forests:Application to Demographic and Health Survey Child Survival Data. University of Nairobi.

Data.unicef.org. (2019). [Online] Available at: https://data.unicef.org/wp-content/uploads/2016/11/UNICEF-Pneumonia-Diarrhoea-report2016-web-version.pdf (Accessed 13 Jul 2019).

Degenhardt, F., Seifert, S., Szymczak, S. (2017). Evaluation of variable selection methods for random forests and omics data sets. Briefings in bioinformatics, 20(2), 492-503.

Fan, J., and Lv, J. (2010). A selective overview of variable selection in high dimensional feature space. Statistica Sinica, 20(1), 101.

Golding, N., Burstein, R., Longbottom, J., Browne, A. J., Fullman, N., Osgood-Zimmerman, A., ... Dwyer-Lindgren, L. (2017). Mapping under-5 and neonatal mortality in Africa, 2000–15: a baseline analysis for the Sustainable Development Goals. The Lancet, 390(10108), 2171-2182.

Hastie, T., Tibshirani, R., Friedman, J., Franklin, J. (2005). The elements of statistical learning: data mining, inference, and prediction. The Mathematical Intelligencer, 27(2), 83-85.

Hitziger, M., Ließ, M. (2014). Comparison of three supervised learning methods for digital soil mapping: Application to a complex terrain in the Ecuadorian Andes. Applied and Environmental Soil Science, 2014.

Kanmiki, E. W., Bawah, A. A., Agorinya, I., Achana, F. S., Awoonor-Williams, J. K., Oduro, A. R., ... Akazili, J. (2014). Socio-economic and demographic determinants of under-five mortality in rural northern Ghana. BMC international health and human rights, 14(1), 24.

Khodaee, G. H., Khademi, G., Saeidi, M. (2015). Under-five Mortality in the World (1900-2015). International Journal of Pediatrics, 3(6.1), 1093-1095.

Kimani-Murage, E. W., Fotso, J. C., Egondi, T., Abuya, B., Elungata, P., Ziraba, A. K., ... Madise, N. (2014). Trends in childhood mortality in Kenya: the urban advantage has seemingly been wiped out. Health place, 29, 95-103.

King, B. E., Rice, J. (2018). Predicting Attendance at Major League Soccer Matches: A Comparison of Four Techniques. Journal of Computer Science, 6(2), 15-22.

Poona, N. K., Van Niekerk, A., Nadel, R. L., Ismail, R. (2016). Random forest (RF) wrappers for waveband selection and classification of hyperspectral data. Applied Spectroscopy, 70(2), 322-333.

Shi, L., Westerhuis, J. A., Rosén, J., Landberg, R., and Brunius, C. (2018). Variable selection and validation in multivariate modeling. Bioinformatics, 35(6), 972-980.

Yu, L., Liu, H. (2003). Feature selection for high-dimensional data: A fast correlation-based filter solution. In Proceedings of the 20th international conference on machine learning (ICML-03) (pp. 856-863).