International Journal of Statistics and Applications 2020, 10(5): 118-130
DOI: 10.5923/j.statistics.20201005.02

Determinants of Under Five Child Mortality from KDHS Data: A Balanced Random Survival Forests (BRSF) Technique

Hellen Wanjiru Waititu[1,*], Joseph K. Arap Koskei[1], Nelson Owuor Onyango[2]

[1]School of Physical and Biological Sciences, Moi University, Eldoret, Kenya
[2]School of Mathematics, University of Nairobi, Nairobi, Kenya

Abstract This study aimed at identifying the determinants of Under Five Child mortality (U5CM) based on Kenya Demographic and Health Survey (KDHS, 2014). One of the key challenges with Demographic and Health Survey datasets involves extreme imbalance between the mortality and non-mortality classes. In this particular research only 6.4% of children experienced under five years mortality while 94.6% survived beyond five years. To establish the determinants of U5CM, we opted to handle the class imbalance using four different balancing techniques: Random Under-sampling, Random Over-sampling, Both-sampling, and Synthetic Minority Over-sampling technique. We then did variable selection using Random Survival Forests following the four techniques. The variables selected from each of the four datasets were then used in a Cox-PH regression to determine the effect of select covariates on child mortality, after conducting appropriate model diagnostics. After the analysis, the variables which resulted in increased hazard of child mortality include V206 (Sum of demised sons), V207 (Sum of demised daughters), V203 (Sum of daughters living at home), V218 (Sum of existing children), V238 (Number of deliveries in the last 3 years), HW72 (Weight for height standard deviations) and interaction between B1 (Child's Month of birth) and V206. Based on model selection indices, Under-sampling balancing schemes performed well for identification of U5CM determinants. By grouping these variables, this study identified birth characteristics of the child (such as age at birth), reproduction factors of the mother (such as number of siblings born before), feeding conditions and anthropometric measurements as key determinants of U5CM.

Keywords Under five mortality, Balanced Random Survival Forests, Class Imbalance in data, Cox-PH regression in Survival analysis

1. Introduction

1.1. Background

The desire to understand the determinants of Under 5 Child Mortality (U5CM) poses a very important aspect of research, as countries aim to achieve the Millennium Development Goals (MDG 2015 – 2030). The Demographic and Health Surveys (DHS) program has been very instrumental for obtaining and disseminating authentic, national representative data on family planning, fertility, maternal and child health, among other health issues. The most recent DHS survey conducted in Kenya was KDHS 2014.

This study aims at identifying the determinants of U5CM in Kenya. Comparisons shall be made between mortality and non-mortality groups from the KDHS 2014 data. Mortality group composes a very minority class (less than 7%) of the entire population, while the non-mortalities constitute the majority class. Imbalanced classification is a common problem with most datasets including mortality data, fraud data, fraud detection, claim prediction, default prediction, spam detection among others. Handling imbalanced classification has received prominence in many studies ([1], [2], [3], [4], [5]).

The KDHS data is associated with 1,099 variables and 20,964 rows of data. Due to high dimensionality of the data, one needs to identify effective variable selection techniques in order to handle a problem such as to identify determinants of child mortality. Machine learning techniques (that require no distributional assumptions on data) such as Random Survival Forests, support vector machine among others have received wide application in studies involving high dimensional datasets ([6], [7], [8], [9], [10], [11], [12]). These machine learning techniques have been useful when dealing with problems such as missing data imputation, classification imbalance and variable selection.

* Corresponding author:
hwaititu@cuea.edu (Hellen Wanjiru Waititu)
Received: Sep. 11, 2020; Accepted: Sep. 30, 2020; Published: Oct. 15, 2020
Published online at http://journal.sapub.org/statistics

Besides, DHS data is often associated with missing data problem. This is often one of the main data analysis tasks before running the desired models. In this case, we did multiple imputation using RF algorithms, before proceeding with RSF classification. In this study however, we dwelt more on handling the challenge of imbalanced classification in mortality data.

The remaining part of the paper is laid out as follows: Section 2 discusses the methodology employed in this study, from description of the data, exploratory data analysis, effects of data imbalance, the theory behind Random Survival Forests, the structure of the COX-PH model used, and finally model selection criteria using concordance statistic. Section 3 summarizes the results of the study both from variable selection using RSF to the Cox-PH fit. Finally, section 4 offers a discussion of our results against other ongoing research on determinants of U5CM.

2. Methodology

2.1. Data Description and Ethical Approval

The data for this research was drawn from the 2014 Kenya Demographic and Health Survey (KDHS) data [13]. This is the sixth Demographic and Health Survey (DHS) conducted in Kenya since 1989. KDHS is a national research undertaking conducted every five years with an intention of collecting a wide range of data with a strong interest on indicators of reproductive health, fertility, mortality, maternal and child health, nutrition and self-reported health habits among adults [14]. It is a household sample survey data with a national representation where households are selected at random from Kenya National Bureau of Statistics (KNBS) sampling frame.

The survey procedures, instruments and sampling methods used in the KDHS 2014 acquired ethical recommendation from the Institutional Review Board of Opinion Research Corporation (ORC) Macro International Incorporated, a health, demographic, market research and consulting company situated in New Jersey, USA. We sought official registration on the DHS website and got permission to use the KDHS 2014 data. The data was downloaded in SPSS format and constituted 1,099 variables and 20,964 observations. Using package foreign, the data was imported to R software version 3.6 for analysis. Variables with 100% missing observations and those which were correlated were deleted from the data reducing the number of variables to 786. Survival time and status variables which are important considerations when analyzing survival data were calculated and included in the dataset.

2.2. Data Exploration and Analysis

The data was explored and analyzed using R software. This involved summarizing and visualizing characteristics of the variables within the dataset. The entire dataset was found to be highly imbalanced with the mortality class having 871 observations, constituting 4% of the overall data while the majority class had 20,093 observations constituting 96%. For this analysis, we singled out on the Nairobi dataset only from the KDHS (2014) data. Different covariates including region, residence, sex, level of education, wealth index, among others, were also found to have high class imbalance (between survivors and non survivors), with the minority class size ranging between 3% and 6%.

The aim of this research is to find an effective way of applying the variable selection technique called Random Survival Forest (RSF), to analyze data with imbalance. KDHS data is a national survey data which is classified into 8 regions, constituting former provinces in Kenya. For this work, we analyzed data only for Nairobi region, being a unique urban system in Kenya. It's a metropolitan region with improved health facilities and access, while also having high levels of socio-economic disparity among populations. Nairobi hosts some of the largest slam settlements of the world including Kibera, Mukuru, Mathare and Kangemi. However, majority of Nairobians are in the middle and upper class by socio-economic status classification enjoying sufficient access to proper health and nutrition for their children.

In the KDHS 2014 data, Nairobi region alone was associated with 788 covariates and 532 observations. Some variables in this subset of data were found to have 100% missing information and others were highly correlated. These variables were deleted leaving 757 variables. Some of the variables that were deleted from Nairobi data include variables related to medication for fever that are currently out of use, for example ML15A (time when the individual began malaria drugs), ML15B (days when child took malaria drugs), ML15C (first source of fansidar), ML23C (first source for other anti-malaria) among others. Other variables like V000 (country code), V024 (De facto region of residence), among others were also deleted from Nairobi data set.

The data was found to have high level of missing information. The algorithm "missForest," which is a random forest-based algorithm for missing data imputation [15] was applied to handle missing data.

Nairobi dataset equally showed high level of class imbalance. This imbalance between mortality and survivor classes is clearly shown on Table 1(a) with 6.4% minority class (mortality class) representation. Similarly, the variables in the data (covariates) show high imbalance in the mortality class. Table 1(b) shows the imbalance between mortality and survivor classes in one of the covariates – child sex.

Table 1(a). Imbalance in KDHS 2014 Nairobi region data

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

Table 1(b). Imbalance in the KDHS 2014 Nairobi region data by Covariate (Child Sex)

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

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

Imbalanced data has been seen to severely hamper the classification performance of learning algorithms, inclusive of Random Forests and other ensemble methods, since their opinions are determined from classification error [16]. In such imbalanced datasets, the classifiers often show biased behavior supporting the majority class and present the minority class lightly [17]. We are therefore interested in construction of classifiers that are skewed toward the minority class, while still maintaining the precision of the majority class.

2.3. Imbalance and Its Effects in Datasets

A dataset is said to be technically imbalanced if its class distributions are not equal. However, when there is a significant, or in some cases extreme, disproportion among the number of examples of each class of the problem, then the dataset is said to be imbalanced [18]. For instance, in a cohort of 1000 children, its often the case that mortality group over the study period composes of less than 50 children (representing less than 5%) or less, hence leaving an entire 95% plus as the non-mortality group.

Imbalanced data classes are common in many real-life situations including mortality data where the survivors greatly outnumbers the mortality, rare disease diagnosis data records where large number of patients do not have the disease, fraud detection, among others. In most of the imbalanced data situations, it is the underrepresented class which is of most interest, since despite its being rare, the minority class may carry important and useful knowledge required in prediction.

When dataset is imbalanced and one class dominates the other, machine learning algorithms such as random forests among others have issues classifying correctly. The algorithms are sensitive to proportions of different classes. They often show biased behavior supporting the majority class and present the minority class lightly [16], [19]. This leads to higher rate of misclassification in the minority class samples [20], [21] which in turn results in weak predictive accuracy of the minority class and misleading high predictive accuracies in the majority class, as a result of correct classification [22], [23], [24]. Thus, the performance of such algorithms is decreases significantly when it comes to predicting the minority class.

Many machine learning algorithms are designed to maximize overall accuracy. This can be misleading in imbalanced datasets because the minority class holds a small effect of this measure. However, when data is balanced, accuracy rates tend to decline [25]. This is attributed to the fact that balanced data reduces the training set size leading to degeneracy of the model through omission of cases encountered to the test set.

The machine learning algorithms aim at minimizing the overall error rate instead of paying attention to the minority class. Therefore, they do not make accurate prediction for the minority class if they don't get the necessary amount of information.

[25] in his research demonstrating problems encountered when unbalance data is used in data mining algorithms found that algorithms tend to degenerate by assigning all cases to the majority class when data is highly imbalanced and still achieve high accuracy scores. Hence, evaluating algorithm performance using predictive accuracy alone is inappropriate when data is imbalanced.

In order to overcome these issues it is important, when working with such machine learning algorithms to work with balanced classification. However, this is in most cases overlooked. We are therefore interested in construction of classifiers that are skewed toward the minority class, while still maintaining the precision of the majority class.

2.4. Data Balancing Techniques

Various techniques have been suggested to solve problems associated with class imbalance. We can group these techniques into four categories, subject to how they deal with imbalance. The categories includes data level (or external/ re-sampling techniques), algorithm level (or internal) techniques, cost-sensitive learning techniques and ensemble-based methods. There is no open directive that indicates the best strategy to use. However, many studies have shown that, external techniques greatly improve the ultimate performance of the classification in comparison with non-preprocessed data set for various types of classifiers [18]. In addition, re-sampling techniques are independent of the classifier, can be easily implemented for any problem and do not need adaptation of any algorithm to the dataset [26]. They are also able to effectively balance the dataset resulting in training sets that are suitable for satisfactory calibration of machine learning algorithms [27]. [28], [29] and [16] have proved the effectiveness of balancing class distributions using data level techniques.

In this research we apply the Data level Preprocessing (or external) techniques. The methods re-balance the sample space aiming to lessen the effect of the imbalanced class distribution in the learning process. The Data level techniques are further classified into three groups [30] which are: under-sampling methods, over-sampling methods and hybrids methods which combine both sampling methods. The Data level techniques used in this research are:

a) Random under-sampling

This aims at balancing dataset by randomly eliminating examples of the majority class up to when the dataset is balanced. The major drawback of this method is that there is

a high possibility of discarding potentially useful data pertaining to majority class leading to a possibility of information loss.

b) Random over-sampling

While the under-sampling method involves removal of samples from the majority group, over-sampling method generates new samples for the minority class. To balance the data using this method, the observations from the minority class are reduplicated. New instances are created from the existing ones; hence over-sampling does not increase information but raises the weight of the minority class by replication. One advantage of over-sampling methods is that there is no information loss. However, since over-sampling simply makes exact copies of the minority class observations, it increases the chances of over fitting due to replication. Therefore, even if there will be improvement in the training accuracy of the data the overall accuracy of the data may be worse. In addition, while dealing with large imbalanced data sets, over-sampling may increases computational work and execution time [31].

c) Both-sampling

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

d) Synthetic Minority Oversampling technique (SMOTE).

This is a hybrid method in re-sampling techniques where both under-sampling and over-sampling approaches are combined with an aim to overcome their drawbacks. SMOTE has become one of the most outstanding approaches in data balancing field [18]. The key idea in SMOTE proposed by [32] is to produce new samples of the minority class artificially. This helps to avoid over fitting brought about by reduplication of minority class instances. Additionally, the majority class examples are under-sampled, giving rise to a more balanced dataset.

Generation of Synthetic samples takes the following steps:

- Randomly select a minority and its k nearest minority class neighbors. The value of k is determined by the amount of oversampling needed.
- Calculate the difference between the vector of selected minority and that of one of its nearest neighbors.
- The difference got is then multiplied by a random number between 0 and 1. The result is added to the selected minority vector. By so doing a new random point is added along the line joining the two vectors under consideration.

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

In the R environment, Package DMwR [33] and ROSE package [34] are used to enhance data balancing. ROSE package [34] is used to enhance data balancing using under-sampling, over-sampling and both-sampling methods. On the other hand, package DMwR [33], assists in data balancing using SMOTE. In SMOTE the parameters $perc.over$ and $perc.under$ respectively control the amount of over-sampling and under sampling to be done. If a completely balanced data set is required, the minority cases are doubled while the majority class is halved.

In this study, we used under-sampling, over-sampling, both-sampling and SMOTE methods to balance the Nairobi region data. The balanced data was analyzed using RSF algorithm.

2.5. Random Survival Forest Algorithm

The KDHS dataset has a total of 1099 variables that are possible candidates for predicting child mortality. After some data management exercise, the number of candidate covariates reduced to 757 possible covariates. Before fitting a regression type model in order to embark on the exercise of determining the effect of child mortality predictors, we needed to do a variable selection exercise in order to further reduce the variables of importance to a manageable subset of important variables. A Random Survival Forest technique, supplemented by our own intuition of sensible covariates for child mortality resulted into a reduced set of utmost 20 covariates for the regression steps that followed.

Random Survival Forest algorithm is described as follows [35]:

a) The procedure starts by randomly drawing $ntree$ bootstrap samples from the initial data consisting of G samples. On average, each bootstrap sample sets aside 37% of the data called out of bag (OOB) data with respect to the bootstrap sample and each sample has R predictors.

b) For each of the drawn samples, a survival tree is grown. Construction of survival tree begins with randomly selecting $mtry$ out of R possible predictors in x for splitting on. The value of $mtry$ depends on the number of available predictors and is data specific. All the $ntree$ bootstrap samples are designated to the top most node of the tree which is also referred to as the root node. This root node is then separated into two daughter nodes each of which is recursively split progressively maximizing survival difference between daughter nodes/ increasing within-node homogeneity.

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

d) After the tree is fully grown, cumulative hazard function (CHF) is computed as well as the mean over all CHFs for the $ntree$ trees. This is done to attain the

ensemble CHF.

e) By using out-of-bag (OOB) data only, the ensemble OOB error is calculated using the first b trees, where $b = 1, \ldots, ntree$.

2.5.1. Node Splitting

From the RSF algorithm, a forest originates from randomly drawn $ntree$ bootstrap samples. Each bootstrap sample becomes the root of each tree in the forest. There are R predictors in each bootstrap sample. From the R predictors, we randomly select $mtry$ predictors for splitting on. Suppose we take h to be the h^{th} node to be split into two daughter nodes. Within node h, let there be n observations each with survival time denoted by T_l, and censoring status given by

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

In right censored data, all details of developing a forest take into consideration the outcome. For right censored data, the outcome is survival time and censoring status [36].

The information at time t_i can be summarized as in Table 2 below.

Table 2. Summary table of information at time t_i

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

Where, $d_{i,j}$ stands for the number of events in daughter node $j = 1,2$ at time t_i, $d_i = d_{i,1} + d_{i,2}$

$Y_{i,j}$ represents individuals who are alive in daughter node j, $j = 1,2$ at time t_i, $Y_{i,1}$ is the number of $T_i \geq t_i, x_i \leq C$, where T_i is the duration of survival for the i^{th} individual and t_i the distinct event time in node h

$Y_{i,2}$ is the number of $T_i \geq t_i, x_i > C$

$$Y_i = Y_{i,1} + Y_{i,2}$$

From the $mtry$ predictors in node h, take any predictor x (for example age). Using predictor x, find a splitting value c (for example from predictor age, the splitting value could be 2 years). The splitting value c is chosen in such a way that the survival difference for predictor x between $x \leq c$ and $x > c$ is maximized. $x \leq c$ separates to the left node while $x > c$ goes to the right node. The survival difference between the two nodes is calculated using a predetermined splitting method. This procedure is repeated with another splitting value c until we get a value which results in maximum survival difference in predictor x. The same procedure is repeated for the remaining $mtry - 1$ predictors in node h. This is done until we get predictor x^* and split value c^* which results in maximum survival difference between the two daughter nodes [37]. This process is repeated at every node. When survival difference is maximum, unlike cases with respect to survival are pushed apart by the tree. Increase in the number of nodes causes dissimilar cases to separate more. This results in homogeneous nodes in the tree consisting of cases with similar survival.

Splitting criteria is one of the aspects of growing a tree. In this research, log rank splitting rule was used in splitting the node.

2.5.2. Log Rank Splitting Rule

The log-rank splitting rule separates the nodes by selecting the split that yields the largest log rank test. The log rank test is the most frequently used statistical test to compare two or more samples non-parametrically in censored data. PH assumption is the key requirement for the optimality of log rank test. For a split using covariate x and its splitting value c, the goodness of fit will be measured using log rank statistics represented as;

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

This equation measures the magnitude of separation between two daughter nodes. The best split is given by the greatest difference between the two daughter nodes which is given by the largest value of $|L(x, c)|$.

RSF gives a measure of variable importance (VIMP) which is totally nonparametric. In this study, using the RSF model, the highly predictive risk factors from the four balanced datasets were extracted. The extracted important predictors were then fitted in the Cox PH model in order to estimate the effect of statistically significant predictors.

2.6. Determining Predictors of Child Mortality

The Cox ph model [38] is frequently used to determine collectively the effect of various risk factors on survival duration. The formula for the Cox ph model is written as

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

This formula displays the risk at time t for an individual specified by a set of covariates X. In this case, X is a group of variables that are used in the model for prediction of the risk of the given observations. From the formula, the risk at time t is a product of $h_0(t)$, the baseline hazard function and $\exp\{\sum_{i=1}^{p}(\beta_i X_i)\}$, the exponential to the sum of the p predictor variables in X. the baseline hazard function indicates what the risk would be when there are no covariates. The coefficient β_i gives the magnitude of the influence of the covariates.

2.6.1. Checking the COX-PH Assumptions

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

These include:

- The proportional hazard assumption. Schoenfeld residuals were used to test this assumption.

- Functional relationship between the log hazard and the covariates. Martingale residual were used to assess this assumption.
- Possible presence of outliers or influential observations. Deviance residual was used to examine possible presence of influential observations.

2.7. Model Selection Criterion

Comparison of prediction accuracy of the different models was done based on concordance index. In survival analysis, a pair of observations is said to be concordant if for the individual that got the event fast the model predicts a higher risk of event. Harrell's concordance index (C-index) [39] is used to estimate prediction error. It estimates the likelihood that in a pair of cases selected at random, the case that came to have an event first had a worse predicted result. Suppose we have two observations whose outcome is predicted. If the observation predicted to have the worst outcome experiences an event first, then the two observations are said to be concordant (i.e. they have the appropriate practice). Computation of concordance error rate is as given below.

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

b) A pair is omitted if:

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

c) After the omissions are done, we remain with all the other pairs which are referred to as permissible pairs.

A score of value 1 is given to a permissible pair if:

- For all pairs having unequal survival durations resulting in prediction being worse for the observation with shorter survival duration.
- For all pairs having uniform survival durations resulting in similar prediction results.
- For all pairs having equal survival duration given that not both observations are events, the observation with event results in a worse prediction outcome.

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

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

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

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

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

3. Results

3.1. Balancing Schemes

The sample sizes obtained after different balancing methods are shown in Tables 3(a) and 3(b)

Table 3(a). Balanced KDHS 2014 Nairobi region data

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

Table 3(b). Balance in KDHS 2014 Nairobi survival data grouped by child sex

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

The different methods of data balancing resulted in different sample sizes. Under-sampling method resulted in the smallest sample size of 68 with both the mortality and survival classes each taking 34 observations. The two tables 3(a) and 3(b) show balance in mortality and non mortality classes in the overall data as well as in sample covariates.

The balanced data is then analyzed for variable selection using RSF algorithm. The results of running the RSF algorithm using balanced data are given in the Table 4.

3.2. Variable Selection Using RSF after Different Balancing Schemes

From the results in table 4, a forest of 1000 trees was grown for each data set. This was done by drawing 1000 bootstrap samples from the respective initial data with the sample sizes given in the table. The size of each bootstrap

sample drawn is given as resample size used to grow trees in table 4. The bootstrap samples are of different sizes depending on the sample size of the initial data and the balancing method used. Each of the 1000 bootstrap samples is designated to the root of the tree. To develop each tree, 28 out of the 757 possible predictors are selected at random for splitting. The root node is then split into two daughter nodes each of which is recursively split progressively maximizing survival difference between daughter nodes. Node splitting continues until each tree if fully grown. This is achieved when the most extreme node has no fewer than 15 different events. This implies that the samples with bigger number of events will form bigger trees. Hence, the more the number of events, the bigger the average number of terminal nodes and the smaller is the error rate. Over-sampling method with the biggest number of events has the smallest error rate while under-sampling method with the smallest number of events has the highest error rate. Even though the sample sizes are different, the number of variables in the four samples is the same. This explains why the number of variables tried at each split and the numbers of random split points are equal in the four samples.

The identified predictors based on balanced random survival forest (BRSF) using the different balancing methods are presented in table 5.

Table 4. Application of RSF in Balanced data sets

Description	Under-sampling	Over-sampling	Both sampling	SMOTE
Sample size	68	996	1000	136
No. of deaths	34	498	480	68
Number of trees	1000	1000	1000	1000
Forest terminal node size	15	15	15	15
Average no. of terminal nodes	2.518	20.461	19.867	5.41
No. of variables tried at each split	28	28	28	28
Total no. of variables	757	757	757	757
Resample size used to grow trees	43	629	632	86
No. of random split points	10	10	10	10
Error rate	13.33%	7.11%	7.5%	13.32%

Table 5. Important variables from the different balanced datasets (selected variables had a variable importance > 0.02. For variable names, refer to the Appendix)

	Balancing method							
	Under- sampling		Over-sampling		Both sampling		SMOTE	
	Variable	Importance	Variable	Importance	Variable	Importance	Variable	Importance
1	B7	0.0029	B7	0.0251	B7	0.0219	V206	0.0153
2	HW72	0.0086	HW71	0.0157	HW70	0.0105	V207	0.0103
3	HW70	0.0079	HW70	0.0124	HW73	0.0103	V219	0.0055
4	B12	0.0076	HW73	0.0111	HW72	0.0091	B7	0.0055
5	V219	0.0069	HW72	0.0111	HW71	0.0086	V218	0.0044
6	HW71	0.0057	B12	0.0075	V206	0.0063	V419	0.0038
7	HW73	0.0052	V206	0.0074	V214	0.0062	V238	0.0037
8	B8	0.0052	V214	0.0057	B12	0.0056	V203	0.0024
9	V206	0.0042	B8	0.0049	V207	0.0039	V417	0.0022
10	V207	0.0024	M1E	0.0035	B1	0.0035		
11			V419	0.0029	V218	0.0035		
12			H4M	0.0029	B8	0.0034		
13			V208	0.0028	V419	0.0031		
14			V218	0.0027	M1E	0.0026		
15			V418	0.0024	V219	0.0026		
16			V219	0.0024	HW1	0.0024		
17			HW1	0.0022	V208	0.0024		
18			B1	0.0020	V418	0.0023		
19			V230	0.0020	V417	0.0021		
20			V207	0.0020				

The bigger the importance value, the higher the predictive ability of the variable. Variables with VIMP exceeding 0.002 were considered predictive. From table 5, the oversampling method which resulted to 498 events, extracted the highest number of important predictors (20 predictors). Both-sampling method, which resulted into 480 events, extracted 19 important variables. SMOTE method extracted the smallest number of predictors (9 predictors) followed by under sampling method (10 samples).

3.3. Determining the Variable Effects

In order to measure the effects of the selected variables on child mortality, we fit a Cox PH model on the covariates from each variable selection exercise. Before the predictors are fitted in the Cox model, ph assumptions were tested.

3.3.1. Testing Cox Proportional Hazards (PH) Assumptions

Table 6 displays the results of proportional hazards assumption. The global test gives a general picture of proportional hazards violations among the variables in the model. Therefore, p.value < 0.05 suggests one or more violations. For variables that do not satisfy the assumption, interaction with time varying covariate is included. Variables that finally do not satisfy the assumption even after interaction with time varying covariate are not supposed to be included in the model.

From table 6, the test is observed to be statistically insignificant for each of the predictors in the Under-sampling method (p.values >0.05). The global test is also statistically insignificant in Under-sampling method. This is after removal of variable B7 from the model which had a p.value less than 0.05 showing statistical significance hence did not meet the requirements of PH assumption and was deleted from the model. In SMOTE method, two variables did not meet the PH assumptions and are not included in table 6.

In over-sampling and both sampling methods, quite a number of variables as well as the global p.value resulted in statistically significant test. Only a few which are given in table 6 satisfy the PH assumption which is supported by a non significant test of hypothesis result. We therefore assume proportional hazard assumption is met for the variables in table 6. Column "Rho" represents the Pearson product moment correlation between the scaled Schoenfeld residuals and log (time) for each predictor.

In the Schoenfeld residuals graphs shown in Fig 1, the broken lines representing a standard error band around the fit while the continuous line represents a smoothing spline fit to the plot. The line of fit is expected to stay close to the horizontal axis within the whole expanse of time, in order to conclude that the PH assumption holds. This is the case for all covariates selected from the Under-sampling scheme.

The pattern of the deviance residuals shown in Fig 2 looks fairly symmetric around zero. The positive values represent individuals who died too soon compared to the expected survival times while the negative values represent individuals who lived too long. The very large or very small values are the outliers which are poorly predicted in the model. In general, we have symmetry along the zero – line and have no fear for presence of outliers or influential observations in the data.

Table 6. Statistical tests

Under sampling method				**SMOTE**			
Covariate	Rho	Chi-square	P.Value	Covariate	Rho	Chi-square	P.Value
V206	0.1609	0.872	0.350	V206	-0.0166	0.0136	0.9073
V207	0.1956	0.888	0.346	V207	0.0740	0.1944	0.6593
V219	-0.1931	0.856	0.355	V203	0.2104	2.8585	0.0909
B8	-0.2369	1.075	0.300	V218	-0.2186	3.5175	0.0607
B12	-0.2004	2.322	0.128	V219	0.1316	1.0958	0.2952
HW70	-0.0496	0.177	0.674	V238	0.1183	0.9276	0.3355
HW71	-0.1515	1.285	0.257	V419	0.0684	0.2812	0.5959
HW72	0.0529	0.153	0.695	Global	NA	12.8034	0.0770
HW73	0.1674	1.870	0.171				
Global	NA	10.882	0.284				
Over sampling method				**Both sampling method**			
Covariate	Rho	Chi-square	P.Value	Covariate	Rho	Chi-square	P.Value
HW72	-0.1667	0.9343	0.334	V206	0.101	3.05	0.0807
H4M	-0.0355	0.0474	0.828	V207	0.136	4.85	0.0277
B1:V206	-0.0248	0.2277	0.633	Global	NA	5.54	0.0627
Global	NA	1.0825	0.781				

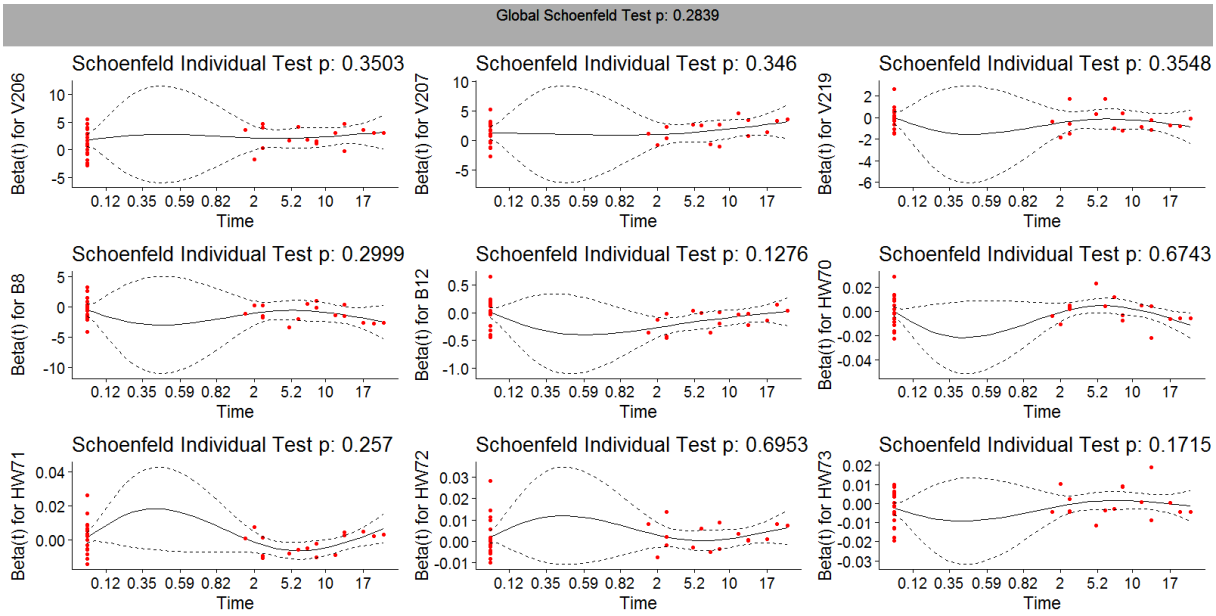

Figure 1. Schoenfeld residuals for variables in under sampling method

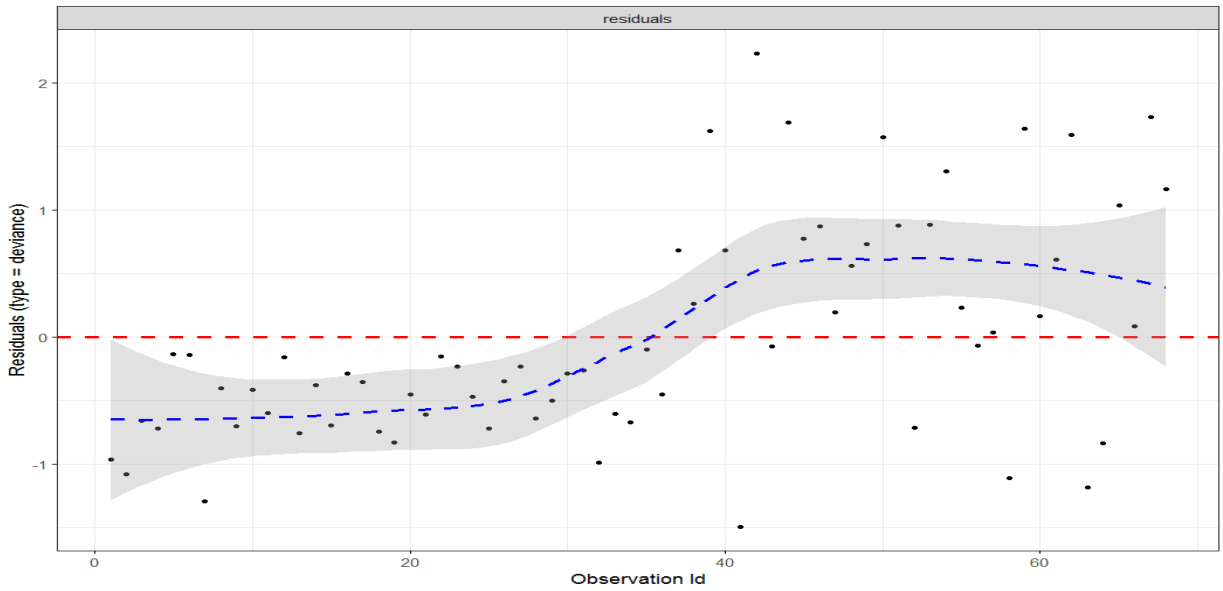

Figure 2. Deviance residuals for under sampling method

3.3.2. Parameter Estimates

From the previous section, we noted that the different balancing methods yielded different sample sizes and different predictors from the RSF classification. After diagnostic tests on Cox PH models, the respective predictors were fitted to the parsimonious Cox PH model [37] in order to check concurrently the effect of different risk factors on survival time.

The results of fitting the Cox model are shown in Table 7. The regression coefficient column marked "Coefficient" gives estimates of the logarithm of the hazard ratio between the two groups. From the estimates, a positive coefficient is said to increase the risk of death (hazard) and thus decrease the expected (average) survival time. On the other hand, a negative coefficient reduces the risk of death and thus raises the expected survival span.

In explaining the determinants of child mortality, one therefore is interested in the variables with positive coefficient, which are positively related with the event (mortality) probability, and consequently negatively related with the length of survival. From table 7, under-sampling method resulted in 9 predictors, out of which only 3 were likely to increase the risk of death. Similarly, SMOTE returned 5 predictors that are likely to increase the risk of death out of 7 important variables which satisfy PH assumptions. Over-sampling and both-sampling method had 3 and 2 predictors respectively all of which had positive coefficients.

Table 7. Result of fitting the respective predictors in Cox PH model

Under sampling method					**SMOTE**				
Predictor	Coefficient	Exp(coefficient)	Se(coefficient)	$pr(>\|z\|)$	Predictor	Coefficient	Exp(coefficient	Se(coefficient)	$pr(>\|z\|)$
V206	2.0637	7.8753	0.3988	2.29e-07	V206	2.2819	9.7956	0.3499	6.94e-11
V207	1.5189	4.5675	0.3728	4.61e-05	V207	1.8688	6.4805	0.3111	1.88e-09
V219	-0.1912	0.8259	0.2032	0.3466	V203	0.0922	1.0966	0.2903	0.7509
B8	-0.8111	0.4444	0.3636	0.0257	V218	0.3171	1.3732	0.4721	0.5017
B12	-0.0589	0.9428	0.3246	0.0697	V219	-0.1723	0.8418	0.4972	0.7289
HW70	-0.0002	0.9998	0.0014	0.8667	V238	0.6561	1.9273	0.2286	0.0041
HW71	-0.0005	0.9995	0.0011	0.6490	V419	-0.6068	0.5451	0.3061	0.0474
HW72	0.0022	1.0022	0.0010	0.0340					
HW73	-0.0013	0.9987	0.0010	0.2124					
Over sampling method					**Both sampling method**				
Predictor	Coefficient	Exp (coefficient)	Se(coefficient)	$pr(>\|z\|)$	Predictor	Coefficient	Exp (coefficient	Se (coefficient)	$pr(>\|z\|)$
HW72	0.0001	1.0000	2.152e-05	4.4e-09	V206	1.8300	6.2339	0.0763	<2e-16
H4M	0.0244	1.025	0.02115	0.25	V207	1.5285	4.6112	0.0730	<2e-16
B1:V206	0.1854	1.025	0.01161	<2e-16					

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

From table 6 for example, variable V206, in under-sampling method has $(coefficient) = exp(2.0637) = 7.8753$. HR value which is clearly greater than 1 implies that variable V206 increases the hazard by a factor 7.8753. This is deduced from the fact that a predictor is related with increased risk when the value of HR>1, and decreased risk when HR<1. When the HR value is close to 1, the predictor has no impact on survival. From our results, there are 2 predictors in under-sampling method associated with increased risk, 0 in over-sampling, 2 in both-sampling and 4 in SMOTE (refer to Table 6 above).

The column marked $pr(>|z|)$ gives the value of the Wald statistic. Wald statistic evaluates whether the explanatory variables in a model are significant. A variable is said to be statistically significant when its p.value is less than 0.05.

3.3.3. Model Goodness of Fit Statistic

The concordance statistic was used to analyze the performance of the models on prediction of mortality. Concordance values are given in Table 8 below.

High values of concordance indicate that for higher observed survival duration, the model predicts higher probabilities of survival. Concordance values ranges from 0 to 1. A perfect Concordance results in a value of 1 while 0.5 is as good as random guessing. All our models gave high concordance values above 0.7 with standard errors less than 0.02 as shown in Table 7. Hence all the models represent a good fit according to the concordance Index. Under-sampling method gives largest concordance value of 0.91 indicating the best model fit while over-sampling had the smallest concordance value. SMOTE and both-sampling methods have almost equal concordance value.

Table 8. Model fit statistics: Concordance measure

Description/ Method	Under-sampling	Over-sampling	Both-sampling	SMOTE
Sample size	68	996	1000	136
Concordance	0.91	0.781	0.8644	0.8645
Standard error	0.0262	0.01206	0.0091	0.0243
Discordant	1386	248084	257769	5325
Concordant	137	69549	26991	830
Tied.x	0	0	31815	13
Tied.y	158	33849	23690	471
Tied.xy	0	3621	10434	6

4. Discussions

The study attempts to understand the determinants of under five mortality using survey data from DHS. In this case, Kenya DHS survey 2014 dataset was used for the analysis. The dataset (after variable cleaning) is composed of 757 variables that are candidate determinants of Under five Child mortality. This poses a problem of variable selection from such high dimensional datasets preceding a proper analysis in which the intention is to explain variable effects. Besides, there is too much class imbalance in the datasets particularly where interest is to compare mortality and non mortality groups. For instance, 6.4% of children experience mortality while 93.6% survived up to the age of 5 years. This imbalance is too huge that a direct comparison (before balancing) between two such groups is likely to yield biased

results.

Two challenges were addressed in this study. One problem involved trying to balance the dataset classes before making comparisons between mortality and non mortality cases. The other challenge was due to variable selection. One needs to conduct a proper variable selection exercise in order to identify the correct set of variables to use for the regression analysis.

Most studies explore determinants of child mortality using DHS survey data. [6] used Uganda 1996, 2000, 2006 DHS dataset, [7] used Uganda 2011 DHS, [8] analyzed the data from complete birth histories of four Nepal Demographic and Health Surveys (NDHS) done in the years 1996, 2001, 2006 and 2011, among many other studies. In this study, we have also tapped into the richness of KDHS (2014) dataset, to establish the determinants of U5CM. The key improvement over many studies that have used DHS data to answer the same question lies in our choice to ensure the following remedies are done: (i) class imbalance is eliminated before comparisons are done, (ii) imputation for missing data is done using a machine learning approach (the missForest package in R software used), (iii) variable selection is accomplished again using a machine learning algorithm (RSF). In most studies, researchers often use self intuition or previous studies to determine which covariates to add to their regression models. All these remedies were done before applying a Cox PH regression on the data to reduce chance of reporting biased findings.

Many studies commonly employed regression techniques to explore the determinants of U5CM. Cox PH regression was used by [6], [7], [8] among others. Although we also used the Cox PH model, we preceded it diagnostics including multiple imputation, classification balancing, variable selection, and Cox PH assumptions tests, to ensure that the results from the Cox PH are more reliable.

Our findings show that child mortality is associated with variables related to: child characteristics at birth (such as age at birth), reproduction factors of the mother (such as number of siblings born before), feeding characteristics and anthropometric measurements. This is in line with other findings such as [6] who used Cox PH regression and established that region of residence, sex of the child, type of birth (multiple), birth interval (less than 24 months after the preceding birth), and mother's education were related with an increased risk of children mortality before their fifth birthday. [7] also established that factors related to mother characteristics and previous births such as sex of the child, sex of the head of the household and the number of births in the past one year was found to be significant. [8] explored the effect of mother's education, child's sex, rural/urban residence, household wealth index, regions ecological zones and development.

It's worth to note that even though most of the studies that rely on DHS datasets ([6], [7], [8]) are challenged with high dimensional data and a variable selection dilemma, there is no mention of any statistical form of variable selection. DHS datasets typically are composed of over 700 variables that are candidate determinants of child mortality and one need to carefully select which variables to include in the resultant regression type models. Majority of the studies explore the effect of a predetermined, select group set of covariates, based on self intuition or variables explored from previous studies. We attempted to do a variable selection using a machine learning algorithm, before subjecting the selected variables to Cox PH regression.

Other than finding the determinants of under five mortality, different data balancing methods were used and model selection done using concordance index. In their research [40] used SMOTE to balance data before integrating it with RSF. In this research, under-sampling method resulted in a better model with a concordance index of 0.91 as compared to other balancing methods used. SMOTE generates synthetic samples along the line segment joining two minority samples. By so doing there is a tendency of generating a decimal value in factor or numeric variables which are not meant to be in decimal form. In as much as under-sampling method may discard potentially useful data in majority class there is no loss of data in the minority class which is our main class of interest.

5. Conclusions

In this research, we presented a framework for determination of under five child mortality using the 2014 KDHS data. The framework involved data balancing, variable selection using RSF method and variable prediction using Cox PH model. Various challenges and effects of working with imbalanced data are discussed in this research as well as the various data balancing methods. Analysis of four data balancing methods; over-sampling, under-sampling both-sampling and SMOTE techniques was conducted where under-sampling model emerged the best with a concordance index of 0.91. Based on this research, child mortality is associated with variables related to child characteristics at birth (such as age at birth), reproduction factors of the mother (such as number of siblings born before), feeding characteristics and anthropometric measurements.

Appendix

Table 9. Description of Important variables

Category	Variable	Description
Child characteristics at birth	B1	Month of birth of child.
	B7	Age at death of the child in completed months.
	B8	Current age of the child in single years for all living children.
	B12	Succeeding birth interval is calculated as the difference in months between the current birth and the following birth, counting twins as one birth.
Reproduction (siblings information)	V203	Total number of daughters living at home.
	V206	Total number of sons who have died.
	V207	Total number of daughters who have died.
	V208	Total number of births in the last five years is defined as all births in the months 0 to 59 prior to the month of interview, where month 0 is the month of interview.
	V214	Imputed duration of the current pregnancy.
	V218	Total number of living children.
	V219	Total number of living children including current pregnancy.
	V230	Year of the last pregnancy termination.
	V238	Total number of births in the last three years.
Maternity and Feeding	V417	Number of entries in the pregnancy and postnatal care history.
	V418	Number of entries in the immunization history.
	V419	Number of entries in the height and weight table.
Height and Weight and Hemoglobin	HW70	Height for age standard deviation (according to WHO).
	HW71	Weight for age standard deviation (according to WHO).
	HW72	Weight for height standard deviations (according to WHO).
	HW73	BMI standard deviations (according to WHO).
	HW1	Age in months of the child.
Maternity	M1E	Last tetanus injection before last pregnancy (CMC).

REFERENCES

[1] Lessmann, S. (2004). Solving Imbalanced Classification Problems with Support Vector Machines. In *IC-AI* (Vol. 4, pp. 214-220).

[2] Tang, Y., Zhang, Y. Q., Chawla, N. V., & Krasser, S. (2008). SVMs modeling for highly imbalanced classification. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 39*(1), 281-288.

[3] López, V., Fernández, A., Moreno-Torres, J. G., & Herrera, F. (2012). Analysis of preprocessing vs. cost-sensitive learning for imbalanced classification. Open problems on intrinsic data characteristics. *Expert Systems with Applications, 39*(7), 6585-6608.

[4] Yan, Y., Liu, R., Ding, Z., Du, X., Chen, J., & Zhang, Y. (2019). A parameter-free cleaning method for SMOTE in imbalanced classification. *IEEE Access, 7*, 23537-23548.

[5] Lin, E., Chen, Q., & Qi, X. (2020). Deep reinforcement learning for imbalanced classification. *Applied Intelligence*, 1-15.

[6] Ayiko, R., Antai, D., & Kulane, A. (2009). Trends and determinants of under-five mortality in Uganda. *East African journal of public health*, 6(2), 136-140.

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

[8] Sreeramareddy, C.T., Kumar, H.N., & Sathian, B. (2013). Time Trends and Inequalities of Under-Five Mortality in Nepal: A Secondary Data Analysis of Four Demographic and Health Surveys between 1996 and 2011. *PLoS ONE, 8(11):* e79818. doi:10.1371/journal.pone.0079818.

[9] Gawande, R., Indulkar, S., Keswani, H., Khatri, M., & Saindane, P. (2019). Analysis and Prediction of Child Mortality in India. *International Research Journal of Engineering and Technology, 6(3),* 5071-5074.

[10] Zhang, X., Tang, F., Ji, J., Han, W., & Lu, P. (2019). Risk Prediction of Dyslipidemia for Chinese Han Adults Using Random Forest Survival Model. *Clinical Epidemiology,* 11, 1047.

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

[12] Liu, V. (2019). Predicting ovarian cancer survival times: Feature selection and performance of parametric, semi-parametric, and random survival forest methods. *Master Thesis, Simon Fraser University*.

[13] Kenya National Bureau of Statistics, Ministry of Health[Kenya], National AIDS Control Council [Kenya], Kenya Medical Research Institute, National Council for Population and Development [Kenya], ICF International. Kenya demographic and health survey 2014. Nairobi, Kenya, 2015.

[14] Corsi, D. J., Neuman, M., Finlay, J. E., & Subramanian, S. (2012). Demographic and health surveys: A profile. *International Journal of Epidemiology, 41*, 1602–1613.

[15] Stekhoven, D. J., & Bühlmann, P. (2012). MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics, 28(1),* 112–118.

[16] Ali, H., Salleh, M. N. M., Saedudin, R., Hussain. K., & Mushtaq, M. F. (2019). Imbalance class problems in data mining: a review. *Indonesian Journal of Electrical Engineering and Computer Science. 14(2),* 1560-1571.

[17] Galar, M., Ferńandez, A., Barrenechea, E., Bustince, H., & Herrera, F. (2012). A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches. IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART C: APPLICATIONS AND REVIEWS.

[18] Fernández, H. A., García, L. S., Galar, M., Prati, R. C., Krawczyk, B., & Herrera, F. (2018). Learning from Imbalanced Data Sets. *Springer,* Gewerbestrasse 11, 6330 Cham, Switzerland.

[19] Zhao, Y., Cen, Y. Data Mining Applications with R; Academic Press: Cambridge, MA, USA, 2013; ISBN 9780124115118.

[20] Datta, S., Das, S. Near-Bayesian support vector machines for imbalanced data classification with equal or unequal misclassification costs. Neural Netw. 70, 39–52 (2015).

[21] Ertekin, S., Huang, J., Bottou, L., Giles, C.L.: Learning on the border: active learning in imbalanced data classification. In: Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management, CIKM 2007, Lisbon, 6–10 Nov 2007, pp. 127–136 (2007).

[22] Cateni, S., Colla, V., Vannucci, M. A method for resampling imbalanced datasets in binary classification tasks for real-world problems. Neurocomputing 2014, 135, 32–41.

[23] He, H., Garcia, E.A. Learning from imbalanced data. IEEE Trans. Knowl. Data Eng. 2009.

[24] Japkowicz, N.; Stephen, S. The class imbalance problem: A systematic study. Intell. Data Anal. 2002.

[25] Olson, D.L. Data Set Balancing. In: Shi Y., Xu W., Chen Z. (eds) Data Mining and Knowledge Management. CASDMKM 2004. Lecture Notes in Computer Science, 3327, 71-80, (2005). Springer, Berlin, Heidelberg. https://doi.org/10.1007.

[26] Ofek, N., Rokach, L., Stern, R., Shabtai, A. Fast-CBUS: A fast clustering-based undersampling method for addressing the class imbalance problem. Neurocomputing 2017, 243, 88–102.

[27] Fiorentini, N.; Losa, M. Handling Imbalanced Data in Road Crash Severity Prediction by Machine Learning Algorithms. *Infrastructures* 2020, 5, 61.

[28] Chawla, N.V., Cieslak, D.A., Hall, L.O., Joshi, A.: Automatically countering imbalance and its empirical relationship to cost. Data Min. Knowl. Disc. 17(2), 225–252 (2008)

[29] Estabrooks, A., Jo, T., Japkowicz, N. A multiple resampling method for learning from imbalanced data sets. Comput. Intell. 20(1), 18–36 (2004).

[30] Batista, G.E.A.P.A., Prati, R.C., Monard, M.C.: A study of the behaviour of several methods for balancing machine learning training data. SIGKDD Explor. 6(1), 20–29 (2004).

[31] Yen, S.J., Lee, Y.S. Cluster-based under-sampling approaches for imbalanced data distributions. Expert Syst. Appl. 2009, 36, 5718–5727.

[32] Chawla, N.V., Bowyer, K.W., Hall, L.O., & Kegelmeyer, W.P. (2002). Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research,* 16, 321-357.

[33] Torgo, L. (2010). Data Mining using R: learning with case studies. *CRC Press* (ISBN: 9781439810187). http://www.dcc.fc.up.pt/~ltorgo/DataMiningWithR.

[34] Lunardon, N., Menardi, G., & Torelli, N. (2013). R package ROSE: Random Over-Sampling Examples (version 0.0-3). Università di Trieste and Università di Padova, Italia. http://cran.r-project.org/web/packages/ROSE/index.html. [p79].

[35] Ishwaran, H., Kogalurt, U. B., Blackstone, E. H., & Lauer, M.S. (2008). Random Survival Forests. *The Annals of Applied Statistics*, 2(3), 841-860.

[36] Breiman, L. (2003b). Setting up, using, and understanding random forests V4.0. https://www.stat.berkeley.edu/~breiman/Using_random_forests_v4.0.pdf.

[37] Weathers, W. & Cutler, R. (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, 927.* https://digitalcommons.usu.edu/gradreports/927.

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

[39] Harrell, F. E., Califf, R. M., Pryor, D. B., Lee, K.L. & Rosati, R.A. (1982). Evaluating the yield of medical tests. *Journal of American Medical Association, 247(18),* 2543—2546.

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