Machine Learning Algorithm for Analysing Infant Mortality in Bangladesh

Atikur Rahman[1](✉), Zakir Hossain[2], Enamul Kabir[3],
and Rumana Rois[1]

[1] Department of Statistics, Jahangirnagar University, Dhaka, Bangladesh
{arahman,rois}@juniv.edu

[2] Department of Statistics, University of Dhaka, Dhaka, Bangladesh
zakir.hossain@du.ac.bd

[3] School of Sciences, University of Southern Queensland, Toowoomba, Australia
Enamul.Kabir@usq.edu.au

Abstract. The study aims to investigate the potential predictors associated with infant mortality in Bangladesh through machine learning (ML) algorithm. Data on infant mortality of 26145 children were extracted from the latest Bangladesh Demographic and Health Survey 2017–18. The Boruta algorithm was used to extract important features of infant mortality. We adapted decision tree, random forest, support vector machine and logistic regression approaches to explore predictors of infant mortality. Performances of these techniques were evaluated via parameters of confusion matrix and receiver operating characteristics curve. The proportion of infant mortality was 9.7% (2523 out of 26145). Age at first marriage, age at first birth, birth interval, place of residence, administrative division, religion, education of parents, body mass index, gender of child, children ever born, exposure of media, wealth index, birth order, occupation of mother, toilet facility and cooking fuel were selected as significant features of predicting infant mortality. Overall, the random forest (accuracy = 0.893, precision = 0.715, sensitivity = 0.339, specificity = 0.979, F1-score = 0.460, area under the curve: AUC = 0.6613) perfectly and authentically predicted the infant mortality compared with other ML techniques, including individual and interaction effects of predictors. The significant predictors may help the policy-makers, stakeholders and mothers to take initiatives against infant mortality by improving awareness, community-based educational programs and public health interventions.

Keyword: Machine learning, boruta algorithm, random forest, auc.

1 Introduction

Infant mortality is defined as the death of infants during the first year of their life [1]. Infant mortality is a key health indicator to assess the progress of child

S. Siuly et al. (Eds.): HIS 2021, LNCS 13079, pp. 205–219, 2021.
https://doi.org/10.1007/978-3-030-90885-0_19

health and the development of a country. One of the important millennium development goals (MDGs) is to reduce child mortality, particularly infant mortality, all over the world [2]. Globally, 4.0 million infants died within the first year of life in 2018 which accounts for 75% of all under-five deaths. The infant mortality rate has reduced worldwide to 29 deaths in 2018 from 65 per 1000 live births in 1990 [3].

In Bangladesh, the infant mortality rate has decreased to 38 in the year 2014 from 87 in 1993 [4]. The reduction in infant mortality by two-thirds of a country indicating the progress towards achieves the MDG-4 [5]. To meet the sustainable development goals (SDGs), reducing the infant mortality rate will make a significant contribution to improving children's health. Bangladesh is far away from achieving the target to get down infant mortality by 5 deaths per 1000 live births [4,6].

A number of studies examined the causes and investigated the potential factors affecting infant mortality. It was reported that the factors related to mothers like her education had a significant positive impact on infant mortality [7,8]. It was also noted that the infant mortality was significantly higher in mothers who did not attend antenatal care follow up during pregnancy [9,10]. A short birth interval between two pregnancies was a highly significant determinant for infant mortality [11]. Moreover, maternal age, multiple births, domestic violence, place of residence, preterm, and having metabolic disorders were identified as statistically significant determinants associated with infant mortality [12–16].

Infant mortality is the consequence of various socio-economic and demographic factors, including factors related to infants themselves. Low birth weight of infants was one of the most important determinants of infant mortality [9,17,18]. Infant mortality was significantly higher among mothers from low income families than those who belong to middle and rich families [11]. In Bangladesh, infant mortality was significantly higher among mothers who give their births at home instead of health centre and also belong to the class of lower income groups [4,19]. Antenatal care during pregnancy, wealth status, birth size at the time of delivery, and gender of child were potential risk factors for higher rate of infant mortality [4]. A recent study reported that the higher educational attainments of mothers were the protective factors of infant mortality in Bangladesh [18].

Most mortality-related studies analyzed the data by the logistic regression (LR) model, particularly for binary responses. The LR model requires to fulfil all the underlying unavoidable assumptions, predictors are independent of each other and having a significant association with the outcome variable, for estimating the parameters. Therefore, this commonly used prognostic modelling approach is sometimes challenging for estimating the model parameters correctly and also the incorrect estimation algorithm provides misleading information. Recently, the machine learning (ML) and data mining approaches are the further improvement of modelling health data that incorporate artificial intelligence and explore more hidden information from a large volume of data [20–24]. In the area of health research, ML generally aims to predict several clinical outcomes based on multiple predictors [25,26].

In this study, we adapted four different well-known ML techniques: decision tree (DT), random forest (RF), support vector machine (SVM), and LR for the classification and prediction of the significant factors associated with infant mortality in Bangladesh. Moreover, systematic performances of these ML approaches are investigated by comparing accuracy, sensitivity, specificity and precision values.

2 Methods and Materials

2.1 Data and Variables

Infant mortality data were extracted from the latest country-wise representative survey, Bangladesh Demographic and Health Survey (BDHS) 2017–18 [27]. A two-stage stratified random sampling design was used for collecting data in this survey and the detailed information is available at https://dhsprogram.com/data/available-datasets.cfm. The data related to infant mortality were collected from reproductive mothers and 26145 infants were included in this study after excluding missing cases. The binary outcome variable: infant death (death of a live birth before the age of one year) is considered in this study. Infant mortality is the consequence of a variety of multiple factors. The various maternal, socio-economic, demographic and environmental factors were considered as exposure variables such as maternal age at first marriage, maternal age at first birth, mother's body mass index, birth interval between two subsequent pregnancies, antenatal care service during pregnancy, receiving a tetanus toxoid (TT) injection during pregnancy, administrative regions, place of residence, religion, educational attainment of both mother's and father's, occupational status of mother's, women empowerment, exposure of media, total children ever born, child sex, birth order number, sources of drinking water, type of toilet facilities and type of cooking fuel.

2.2 Models

This study aimed to assess the potential predictors associated with infant mortality and to predict infant mortality in Bangladesh using different ML classification models: DT, RF, SVM and LR. Our methodology involves accordingly data pre-processing, feature (the risk factors) selection using Boruta algorithm, splitting the entire data set into training and test data sets applying ML models in the training data set and evaluate the performance of these models on the test data set, and finally predicting infant mortality based on the entire data set using the best performed model. The performances were evaluated using five performance parameters (accuracy, sensitivity, specificity, precision, and F1-score) obtained from the confusion matrix, and the area under the receiver operating characteristics (ROC) curve (AUC). All ML models were performed using the scikit-learn module in Python programming language version 3.7.3. Only the Boruta algorithm was implemented to select the risk factors using the Boruta package in the R programming language [28].

2.3 Boruta Algorithm

Boruta algorithm was performed to extract the relevant risk factors for infant mortality in Bangladesh. This is a wrapper build algorithm around the RF classifier to find out the relevance and important features with respect to the outcome variable [29].

2.4 Decision Tree (DT)

The DT is one of the most simple and intuitive techniques in ML, based on the divide and conquer paradigm [30]. In a DT technique, tests (on input patterns) and categories (of patterns) are used as inner and leaf nodes, respectively. This technique also assigns a class number to an input array by filtering the array down via the tests in the tree [31].

2.5 Random Forest (RF)

The RF algorithm consists of taking hyper-parameters identifying the number of trees and the maximum depth of each tree [32]. The RF is an ensemble learning approach for classification using a large collection of decorrelated DT [33]. In this experiment, we have used 501 DT and Gini for impurity index to implement the RF algorithm in Python.

2.6 Support Vector Machine (SVM)

The SVM is a supervised ML technique used for analyzing data and recognizing patterns [34,35]. A model or classification function is constructed in the SVM training algorithm in order to assign new values into one class on either side of a hyper plane, building it a non-probabilistic binary linear classifier for the two-class learning task. The kernel trick is used in a SVM technique to map the data into a high-dimensional space prior solving the ML task as a convex optimization problem [33–36]. New values are then predicted belonging to a group on the basis of the side of the partition in which these values fall. The nearest data points to the hyper plane that divides the classes are considered as support vectors [33]. We examined SVM models using the sigmoid kernel (the best performed kernel for BDHS 2017–18 infant mortality data set) for this analysis.

2.7 Logistic Regression (LR)

The LR, a probabilistic model, is used for classification problem and predicting the likelihood of the incidence of an event [33]. The association between a categorical response variable and a dichotomous categorical outcome or feature is modelled by the LR. It is used as a binary (multiple) model to predict binary (multiple) responses, the outcome of a categorical response variable, based on one or more exposure variables [30].

2.8 Confusion Matrix Performance Parameters

The graphical representation of real versus predicted class accuracies is obtained by a confusion matrix [33]. To visualize the performance of the classification algorithm, the confusion matrix is used for the comparison of predicted versus real classifications in the form of true positive, false positive, true negative and false negative [33]. Therefore, the performance parameters: accuracy (number of data points correctly classified by the classifier), sensitivity (how well a classification algorithm classifies data points in the positive class), specificity (how well a classification algorithm classifies data points in the negative class) and precision (number of data points correctly classified from the positive class) are measured [33].

2.9 Receiver Operating Characteristic (ROC) Curve

The ROC curve is an alternative and useful visualization technique for classifiers operating on datasets. Fawcett [37] provided a complete and informative introduction about the ROC analysis, emphasising usual misconceptions. The ROC curve reveals the sensitivity of the classifier considering the true positives and false positives rates. When the classifier is outstanding, the true positive rate will increase, and area under the ROC curve (AUC) will be close to 1 [30].

3 Statistical Results: Univariate and Bivariate Analysis

The frequency and percentage distributions of exposure variables, and the prevalence of infant mortality are presented in Table 1. As shown in the table, more than three quarters of mothers (82.6%) married before their legal recommended age of first marriage (at least 18 years) in Bangladesh. The enormous percentage of mothers (88.0%) had their first birth at 20 years or below. Approximately one-third (32.8%) of the mothers were overweight or obese, over half (55.3%) were normal and 11.9% were thin. The birth interval between two subsequent pregnancies for the majority (76.1%) of live births were more than two years. Only 0.4% and 0.3% of the mothers received ANC services and TT-injection, respectively during their pregnancy period. In this study, 70.0% of children were selected from rural and 30.0% from urban areas. The vast majority of participants were Muslim (92.1%), while only 7.9% of children were non-Muslim. With regards to mothers education, 27.0% had no education and only 4.8% of the mothers had higher education. Besides, 32.5% of fathers were illiterate, and 9.4% had higher education.

Table 1 shows that more than fifty percent (58.1%) of the mothers were employed, while 41.9% were unemployed. The proportion of children from low-income families was higher (45.9%) than rich (34.1%) and middle class (20.0%) families. The majority (82.4%) of respondents' mothers were not empowered, while the rest (17.6%) were fund to be empowered. Almost fifty percent (44.9%) of the children's mothers had mass-media exposure in Bangladesh. Half of the children were male and half were female. 46.7% of the children were the first ranked children, 27.7% were the second ranked children and the rest were third or higher ranked children. More than three quarters (80.6%) of mothers had 3 or more number of children, 42.4% of the mothers used to defecate in the unhygienic places, only 13.5% children were from households with less polluted cooking fuel and only 7.1% had no facilities of safe drinking water.

The prevalence of infant mortality was higher among mothers who were married before their legal age of 18 years and had their first child before 20 years (Table 1), although age at first marriage and age at first birth were found to be statistically insignificant factors for infant mortality. The percentage of infant mortality was higher for mothers who were underweight (12.2%) and had less than 2 years preceding birth interval (19.7%) in comparison with their counterparts. Mothers BMI ($p<0.001$) and birth interval ($p<0.001$) were found to be significantly associated with infant mortality in Bangladesh. The variables antenatal care and TT injection during pregnancy were also found to be statistically insignificant determinants with infant mortality. The administrative divisions showed significant association ($p<0.001$) with infant mortality and the prevalence varied from 8.0% (Chittagong) to 11.1% (Mymensingh). Infant mortality was comparatively lower in urban areas (9.2%) than rural areas (9.8%), though the place of residence was statistically insignificant ($p = 0.144$). Infant mortality was found to be significantly higher among non-Muslim (11.4%) as compared with Muslim (9.5%) and the religion was strongly associated with ($p<0.001$) infant mortality.

The educational attainment of both mothers ($p<0.001$) and fathers ($p<0.001$) showed a significant association with infant mortality. The decreasing trend of infant mortality was observed with mothers and fathers increasing levels of education. However, the proportion of infant mortality was found to be higher among working (10.2%) mothers in comparison to non-working (8.9%) mothers. The occupational status of mothers ($p<0.001$) was also significantly associated with infant mortality. The proportion of infant mortality was decreasing significantly with the increasing levels of wealth index ($p<0.001$). Mass-media exposure of mothers was found to be statistically significant ($p = 0.002$) for their infant mortality and this mortality was lower (9.0%) among mothers who were exposed to mass media than their counterparts (10.2%). The percentage of infant mortality was higher for the male children (10.9%), first order birth (10.3%), and mothers having 3 or more children (11.1%). All these variables: child of sex ($p<0.001$), birth order ($p = 0.002$), total children ever born ($p<0.001$) were significantly associated with infant mortality. Households with unhygienic toilet facility (10.5%) and polluted (9.9%) cooking fuel showed a significant ($p<0.001$) higher prevalence of infant mortality than those with hygienic toilet facility and less polluted cooking fuel.

Table 1. Background characteristics and chi-square (χ^2) test statistic associated with corresponding p-value.

Variables	n(%)	Infant death n = 2523 (9.7%)	χ^2-value	p-value
Age at first marriage			6.155	0.013
<18	21586 (82.6)	2128 (9.9)		
18 and above	4559 (17.4)	395 (8.7)		
Age at first birth			2.951	0.086
≤20	22998 (88.0)	2246 (9.8)		
>20	3147 (12.0)	277 (8.8)		
Mother's BMI			52.295	<0.001*
Normal	14471 (55.3)	1456 (10.1)		
Underweight	3101 (11.9)	379 (12.2)		
Overweight and obesity	8573 (32.8)	688 (8.0)		
Birth interval			957.434	<0.001*
≤2 years	6256 (23.9)	1234 (19.7)		
>2 years	19889 (76.1)	1289 (6.5)		
Antenatal care during pregnancy			0.106	0.745
No	26051 (99.6)	2513 (9.6)		
Yes	94 (0.4)	10 (10.6)		
TT-Injection during pregnancy			0.914	0.339
No	26076 (99.7)	2514 (9.6)		
Yes	69 (0.3)	9 (13.0)		
Region			27.940	<0.001*
Dhaka	3276 (12.5)	316 (9.6)		
Barisal	2936 (11.2)	265 (9.0)		
Chittagong	4244 (16.2)	341 (8.0)		
Khulna	2826 (10.8)	259 (9.2)		
Mymensingh	3122 (11.9)	345 (11.1)		
Rajshahi	2859 (10.9)	311 (10.9)		
Rangpur	3273 (12.5)	334 (10.2)		
Sylhet	3609 (13.8)	352 (9.8)		
Place of residence			2.139	0.144
Urban	7855 (30.0)	726 (9.2)		
Rural	18290 (70.0)	1797 (9.8)		
Religion			7.723	0.005*
Non-Muslim	2074 (7.9)	236 (11.4)		
Muslim	24071 (92.1)	2287 (935)		
Maternal education			82.227	<0.001
No education	7055 (27.0)	831 (11.8)		
Primary	10534 (40.3)	1030 (9.8)		
Secondary	7292 (27.9)	594 (8.1)		
Higher	1264 (4.8)	68 (5.4)		
paternal education			63.747	<0.001*
No education	8496 (32.5)	929 (10.9)		
Primary	9162 (35.0)	942 (10.3)		
Secondary	6031 (23.1)	493 (8.2)		
Higher	2456 (9.4)	159 (6.5)		
Mother's occupation			10.889	<0.001
Not working	10961 (41.9)	980 (8.9)		
Working	15184 (58.1)	1543 (10.2)		
Wealth index			25.853	<0.001
Poor	11994 (45.9)	1265 (10.5)		
Middle	5226 (20.0)	504 (9.6)		
Rich	8925 (34.1)	754 (8.4)		
Women empowerment			0.077	0.781
No	21533 (82.4)	2013 (9.7)		
Yes	4612 (17.6)	440 (9.5)		
Exposure of media			9.662	0.002*
Non-exposure	14406 (55.1)	1464 (10.2)		
Exposure	11739 (44.9)	1059 (9.0)		
Child sex			46.344	<0.001
Female	13083 (50.0)	1100 (8.4)		
Male	13062 (50.0)	1423 (10.9)		
Birth order			12.302	0.002*
One	12215 (46.7)	1262 (10.3)		
Two	7094 (27.1)	637 (9.0)		
Three and more	6836 (26.1)	624 (9.1)		
Total children ever born			252.454	<0.001*
1 or 2	5071 (19.4)	190 (3.7)		
3 and more	21074 (80.6)	2333 (11.1)		
Toilet facility			14.479	<0.001*
Hygienic	15065 (57.6)	1364 (9.1)		
Unhygienic	11080 (42.4)	1159 (10.5)		
Cokking fuel			8.757	<0.001*
Less polluted	3526 (13.5)	292 (8.3)		
Polluted	22619 (86.5)	2231 (9.9)		
Drinking water			0.496	0.481
Safe water	24294 (92.9)	2353 (9.7)		
Unsafe water	1851 (7.1)	170 (9.2)		

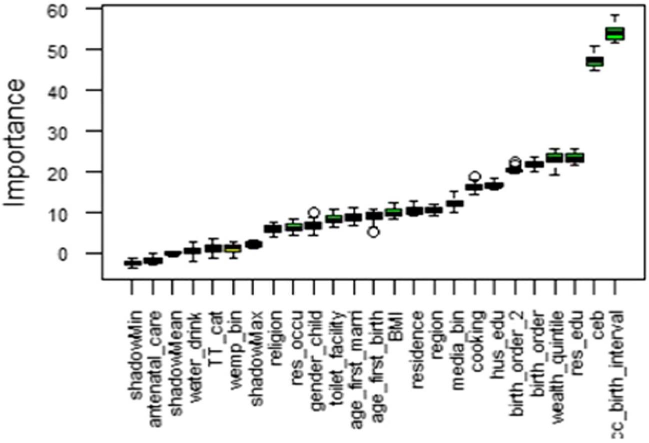

Fig. 1. Features selection using the Boruta algorithm

3.1 Machine Learning (ML) Results

3.1.1 Features Selection

Figure 1 reveals that with the aid of the Boruta algorithm, seventeen variables i.e., age at first marriage, age at first birth, birth interval, place of residence, administrative division, religion, education of parents, BMI, gender of child, children ever born, exposure of media, wealth index, birth order, occupation of mother, toilet facility and cooking fuel were selected among all surveyed variables as the risk factors to predict infant mortality in Bangladesh. Hereafter, these seventeen variables were used to evaluate the performance of ML algorithms.

3.1.2 Machine Learning (ML) Models Evaluation

The performance of different ML models were evaluated using five performance parameters of the confusion matrix (Table 3) and the realized confusion matrices of different ML models using a single run (Table 2), and the area under the ROC curve (Fig. 2). Considering 70% observations as the training data and 30% observation as the test data with the random seed 1119, using the scikit-learn module, we estimated accuracy, sensitivity, specificity, precision, and F1-score of DT, RF, SVM, and LR algorithms to predict infant mortality in Bangladesh and the results are illustrated in Table 2 and 3.

Table 2 illustrates various realized confusion matrices of different ML models using random seed 1119. The confusion matrix compares the actual target positive 761 and negative 7083 cases with those predicted by the different ML models. The DT model has True Positive (TP) = 125, True Negative (TN) = 6440, False Positive (FP) = 643, and False Negative (FN) = 636, i.e., 125 positive and 6440 negative classes data points were correctly classified by the DT, and 643 negative and 636 positive classes data points were incorrectly classified by the DT model. The maximum 371 positive and 6935 negative classes data

Table 2. Realized confusion matrices of different machine learning models

Label	Actual	DT (predicted)		RF (predicted)		SVM (predicted)		LR (predicted)	
		+ve	−ve	+ve	−ve	+ve	−ve	+ve	−ve
Positive	761	125	636	371	390	78	683	0	761
Negative	7083	643	6440	148	6935	707	6376	0	7083

+ve = Positive label, −ve = Negative label

Table 3. Accuracy, sensitivity, specificity, precision, and F1-score of different machine learning models

Models	Accuracy	Sensitivity	Specificity	Precision	F1-score
DT	0.837	0.164	0.909	0.163	0.164
RF	0.893	0.339	0.979	0.715	0.460
SVM (sigmoid kernel)	0.823	0.103	0.900	0.099	0.101
LR	0.903	0.000	1.000	NA	0.00

NA: Not applicable

points were correctly classified by the RF model, whereas the LR model failed to correctly classify any positive class data points though classified correctly all negative class data points.

Table 3 shows that the RF model was the efficient one to predict infant mortality based on the higher value of the performance parameters in all cases. For instance, the RF model provided 89.3% of accurate predictions (accuracy = 0.893), 33.9% of positive cases that were predicted as positive (sensitivity = 0.339), 97.9% of negative cases that were predicted as negative (specificity = 0.979), 71.5% of positive predictions that were correct (precision = 0.715), and 46.0% of F1-score indicating moderate precision and recall (F1-score = 0.460). Though, commonly used LR model provides the highest accuracy score (accuracy = 0.903) and specificity score (specificity = 1.00), but completely failed to estimate the precision of the test. Furthermore, the sensitivity and F1-score were also zero in that case. Figure 2 illustrates the estimated AUC of DT, RF, SVM, and LR models, which were run using the scikit-learn module in Python 3.7.3 by considering 70% observations as training data and 30% observation as test data with the random seed 1119. To predict infant mortality in Bangladesh the estimated AUC scores were 0.5416, 0.6613, 0.5443, and 0.7123 for the ML models DT, RF, SVM with the sigmoid kernel, and LR, respectively. Although the LR algorithm showed the maximum AUC among all examined ML models, however it completely failed to classify the positive cases. Therefore, performance of the RF model is comparatively better among all situations. Consequently, to predict infant mortality, the RF algorithm performed better based on the precision, sensitivity, specificity and accuracy measures, and the ROC approaches.

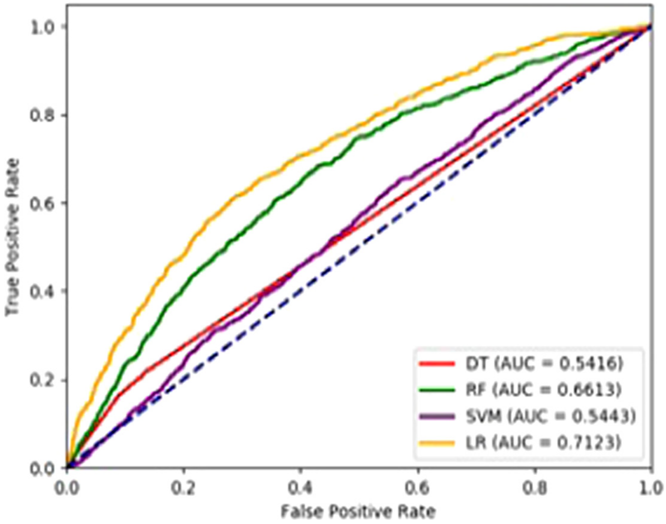

Fig. 2. The ROC curves to predict infant mortality in Bangladesh using DT, RF, SVM, and LR models

3.2 Random Forest (RF) Model for Predicting Infant Mortality

For the entire dataset, therefore, the best performed ML model, the RF model, was fitted to predict infant mortality in Bangladesh using the selected risk factors: age at first marriage, age at first birth, birth interval, place of residence, administrative division, religion, education of parents, BMI, gender of child, children ever born, exposure of media, wealth index, birth order, occupation of mother, toilet facility and cooking fuel, and the top one tree from the forest is visualized in Fig. 3. All the nodes have five parts (feature's question, gini, samples, value and class) with a question based on a value of a feature, except the terminal leaf nodes have four parts (gini, samples, value and class) [38]. The part 'gini' indicates the Gini Impurity of the node, which is the average weighted Gini Impurity decreases as the path move down the tree, 'samples' is the number of observations in the node, 'value' is the number of samples in each class, and 'class' indicates the majority classification for points in the node ('class' is the prediction for all samples in the leaf node) [38].

Each feature's question has either a True (left nodes) or a False (right nodes) answer that splits the node. Based on the answer to the question, a data point moves down the tree and reaches a leaf node (the final decision). Moreover, the blue-type colored leaf indicates a prediction of infant mortality and the orange-type colored leaf indicates a prediction of not infant mortality as shown in Fig. 3. To predict any given respondent's data, simply move down the tree in Fig. 3, using the answer to the feature's question until arriving at a leaf node where the class is the prediction.

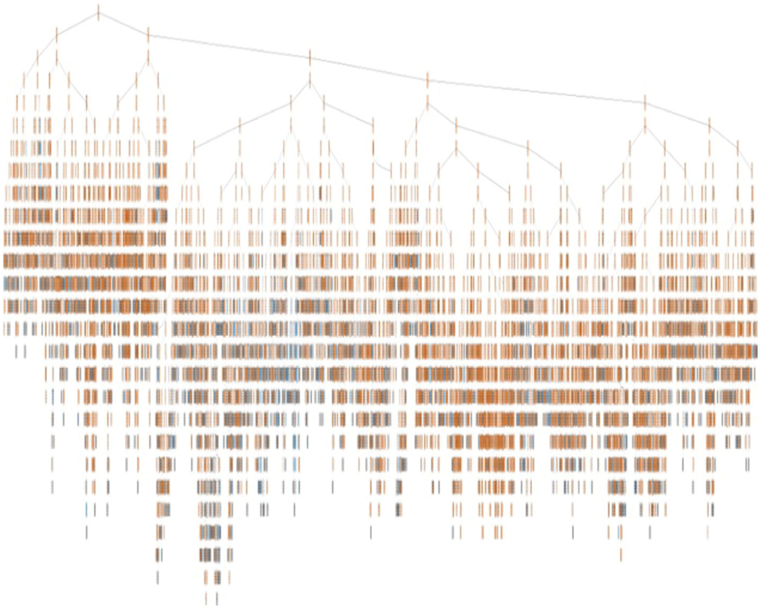

Fig. 3. Top one tree from the fitted RF model to infant mortality in Bangladesh

4 Discussion and Conclusion

Infant mortality is one of the most important public health problems universally; the problem is even more devastating in densely populated countries like Bangladesh. Motivated by such a noticeable public health concern, this research was conducted to find the significant factors and prediction of infant mortality in Bangladesh using different ML algorithms.

The study results reveal that age at first marriage, age at first birth, birth interval, place of residence, administrative division, religion, education of parents, BMI, gender of child, children ever born, exposure of media, wealth index, birth order, occupation of mother, toilet facility and cooking fuel were the major significant factors for predicting infant mortality using the ML features selection algorithm-Boruta. However, birth interval, administrative division, religion, education of parents, BMI, gender of child, children ever born, exposure of media, wealth index, birth order, occupation of mother, toilet facility and cooking fuel were only the significant factors for infant mortality using the conventional chi-squared test.

We evaluated the performance of different ML models: DT, RF, SVM and LR to predict infant mortality in Bangladesh using the performance parameters of the confusion matrix and the AUC. The RF model performed better to predict infant mortality in Bangladesh. The RF model considered the individual and interaction effects of all the selected factors to predict infant mortality in

Bangladesh. Following the path in Fig. 3, for any individual respondent with the given data, one may predict the infant mortality.

On the other hand, the LR model failed to correctly classify any infant mortality. As a result, the LR model failed to estimate the precision and concluded with zero values of the sensitivity and F1-score. This incomplete output is observed due to inappropriately estimating the LR model. As the LR model requires to satisfy all the underlying assumptions before estimating the model, among them predictors having a significant association with the outcome variable and their independence (to avoid the multicollinearity problem) are the foremost assumptions that need to satisfy. In this analysis, only the independent variable(s) among birth interval, administrative division, religion, education of parents, BMI, gender of child, children ever born, exposure of media, wealth index, birth order, occupation of mother, toilet facility and cooking fuel will be used as a predictor variable in estimating infant mortality using the LR model, as these variables were significantly associated (using the chi-squared test in Table 1) with infant mortality and may have a significant association between them.

Hence, to overcome the multicollinearity problem only the independent variable(s) should involve in estimating the LR model, otherwise, the results will be misleading. Furthermore, the RF model does not require any assumptions in estimating the model. Therefore, considering the better performance, the RF model will be better and authentic (in terms of fulfilling the assumptions) to predict infant mortality in Bangladesh in this study.

Conventional chi-square test identified only fourteen variables as significant factors, whereas the ML framework identified seventeen variables as significant factors for predicting infant mortality in this analysis. Needless to say, this study introduces the application of different ML models in the prediction of infant mortality, for instance, DT and RF, which do not require any assumptions and very easy (available) to implement in any standard software. Furthermore, the RF model included all these seventeen significant variables to predict infant mortality using their individual and interaction effects. Considering the high accuracy in prediction, better performance, and assumptions-free feature, the RF model is found to be more authentic and informative to predict infant mortality in Bangladesh.

Acknowledgements. The authors are thankful to the authority of Bangladesh Demographic and Health Survey (BDHS) for making their data available for free. Authors would also like to express their gratitude to Department of Statistics, Jahangirnagar University, Savar, Dhaka, Bangladesh; Department of Statistics, University of Dhaka, Bangladesh; and Faculty of Health, Engineering and Sciences (HES) of University of Southern Queensland, Australia for the technical support.

Funding Information. There is no funding for this work.

Data Availibility Statement. We used secondary data from the Demographic and Health Surveys (DHS) Program. The data are available online at https://dhsprogram.com/data/available-datasets.cfm.

Conflicts of interest. No conflict of interest exits among the authors.

Patient Consent for Publication. Not applicable.

Ethics Statement. This article does not include any data of human participants conducted by any of the authors. The Bangladesh Demographic and Health Survey (BDHS) was approved by ICF Macro Institutional Review Board and the National Research Ethics Committee of the Bangladesh Medical Research Council. Written consent was given by participants in relation to this survey before the interview. All identification of the survey participants was dis-identified before publishing the data. In this study, we used the secondary data that are freely available on the DHS website: https://dhsprogram.com/data/available-datasets.cfm.

References

1. CDC: Infant Mortality. Centers for Disease Control and Prevention (2018). https://www.cdc.gov/reproductivehealth/MaternalInfantHealth/InfantMortality.htm. Accessed 14 July 2021
2. World Health Organization (WHO). Millennium development goals (MDGs) (2018). http://www.who.int/topics/millennium-development-goals/about/en. Accessed 14 July 2021
3. World Health Organization (WHO). The global health observatory (2018). https://www.who.int/data/gho/data/themes/topics/indicator-groups/indicator-group-details/GHO/infant-mortality. Accessed 14 July 2021
4. Vijay, J., Patel, K.K.: Risk factors of infant mortality in Bangladesh. Clin. Epidemiol. Global Health **8**, 211–214 (2020)
5. Hajizadeh, M., Nandi, A., Heymann, J.: Social inequality in infant mortality: what explains variation across low and middle income countries? Soc. Sci. Med. **101**, 36–46 (2014)
6. World Health Organization (WHO). Success factor for women's and child's health: Bangladesh (2015). www.who.int
7. Quansah, E., Ohene, L.A., Norman, L., Mireku, M.O., Karikari, T.K.: Social factors influencing child health in Ghana. PLoS One **11**(1), 1–10 (2016)
8. Kiross, G.T., Chojenta, C., Barker, D., Tiruye, T.Y., Loxton, D.: The effect of maternal education on infant mortality in Ethiopia: a systematic review and meta-analysis. PLoS One **14**(7), e0220076 (2019)
9. Dube, L., Taha, M., Asefa, H.: Determinants of infant mortality in community of Gilgel gibe field research center, Southwest Ethiopia: a matched case control study. BMC Public Health **13**, 401 (2013)
10. Leal, M.D., Bittencourt, S.D., Torres, R.M., Niquini, R.P., Souza, P.R., Jr.: Determinants of infant mortality in the Jequitinhonha valley and in the north and northeast regions of Brazil. Rev Saude Publica **51**(12), 1–9 (2017)

11. Khadka, K.B., Lieberman, L.S., Giedraitis, V., Bhatta, L., Pandey, G.: The socio-economic determinants of infant mortality in Nepal: analysis of Nepal demographic health survey. BMC Pediatr. **15**(152), 1 (2015)
12. Santos, S.L., Santos, L.B., Campelo, V., Silva, A.R.: Factors associated with infant mortality in a northeastern Brazilian capital. Rev. Bras. Ginecol. Obstet. **38**(10), 482–491 (2016)
13. Baraki, A.G., et al.: Factors affecting infant mortality in the general population: evidence from the 2016 Ethiopian demographic and health survey (EDHS); a multilevel analysis. BMC Pregnancy Childbirth **20**, 299 (2020)
14. Varghese, S., Prasad, J.H., Jacob, K.S.: Domestic violence as a risk factor for infant and child mortality: a community-based case-control study from southern India. Natl. Med. J. India **26**(3), 142–146 (2013)
15. Mohamoud, Y.A., Kirby, R.S., Ehrenthal, D.B.: Poverty, urban-rural classification and term infant mortality: a population-based multilevel analysis. BMC Pregnancy Childbirth **19**, 40 (2019)
16. de Bitencourt, F.H., Schwartz, I.V.D., Vianna, F.S.L.: Infant mortality in Brazil attributable to inborn errors of metabolism associated with sudden death: a time-series study (2002–2014). BMC Pediatr. **19**, 52 (2019)
17. Vilanova, C.S., et al.: The relationship between the different low birth weight strata of newborns with infant mortality and the influence of the main health determinants in the extreme south of Brazil. Popul. Health Metrics **15**, 1–10 (2019)
18. Hajipour, M., et al.: Predictive factors of infant mortality using data mining in Iran. J. Comprehen. Pediatr. **12**(1), 1–8 (2021)
19. Dancer, D., Rammohan, A., Smith, M.D.: Infant mortality and child nutrition in Bangladesh. Health Econ. **17**(9), 1015–1035 (2008)
20. Alghamdi, M., Al-Mallah, M., Keteyian, S., Brawner, C., Ehrman, J., Sakr, S.: Predicting diabetes mellitus using SMOTE and ensemble machine learning approach: the Henry Ford Exercise Testing (FIT) project. PLoS One **12**, 1 (2017)
21. Supriya, S., Siuly, S., Wang, H., Zhang, Y.: Automated epilepsy detection techniques from electroencephalogram signals: a review study. Health Inf. Sci. Syst. **8**(1), 1–15 (2020). https://doi.org/10.1007/s13755-020-00129-1
22. Pandey, Y.Z.D., Yin, X., Wang, H.: Accurate vessel segmentation using maximum entropy incorporating line detection and phase-preserving denoising. Comput. Vision Image Underst. **155**, 162–172 (2017)
23. Sarki, R., Ahmed, K., Wang, H., Zhang, Y.: Image Preprocessing in Classification and Identification of Diabetic Eye Diseases. Data Sci. Eng. 1–17 (2021)
24. Supriya, S., Siuly, S., Wang, H., Zhang, Y.: EEG sleep stages analysis and classification based on weighed complex network features. IEEE Trans. Emerg. Topics Comput. Intell. **5**, 236–246 (2018)
25. Sarki, R., Ahmed, K., Wang, H., Zhang, Y.: Automated detection of mild and multi-class diabetic eye diseases using deep learning. Health Inf. Sci. Syst. **8**(1), 1–9 (2020). https://doi.org/10.1007/s13755-020-00125-5
26. Mateen, B.A., Liley, J., Denniston, A.K., Holmes, C.C., Vollmer, S.J.: Improving the quality of machine learning in health applications and clinical research. Nat. Mach. Intell. **2**(10), 554–556 (2020)
27. National institute of population research and training (NIPROT), Bangladesh demographic and health survey 2017–2018. Mitra and Associates, Dhaka, Bangladesh and ICF International, Calverton, Maryland, USA (2019)
28. R Core Team: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing. http://www.R-project.org

29. Kursa, M.B., Rudnicki, W.R.: Feature selection with the Boruta package. J. Statist. Softw. **36**(11), 1–13 (2010)
30. Igual, L., Seguí, S.: Introduction to Data Science. Springer, Cham (2017)
31. Nilsson, N.L.: Introduction to Machine Learning. Stanford University, Stanford, CA (1997)
32. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001)
33. Awad, M., Khanna, R.: Efficient Learning Machines. A press, Berkeley, CA (2015)
34. Burges, C.J.: A tutorial on support vector machines for pattern recognition. Data Mining Knowl. Disc. **2**(2), 121–167 (1998)
35. Müller, K.R., Mika, S., Rätsch, G., Tsuda, K., Schölkopf, B.: An introduction to kernel-based learning algorithms. IEEE Trans. Neural Netw. **12**(2), 181–201 (2001)
36. Vapnik, V.N.: The Nature of Statistical Learning Theory. Springer-Verlag, New York (1995)
37. Fawcett, T.: Introduction to ROC analysis. Pattern Recogn. Lett. **27**, 861–874 (2006)
38. Koehrsen, W.: An implementation and explanation of the random forest in Python. Towards Data Sci. **31**, 1 (2018)