


RESEARCH



# An assessment of random forest technique using simulation study: illustration with infant mortality in Bangladesh

Atikur Rahman<sup>1\*</sup> , Zakir Hossain<sup>2</sup>, Enamul Kabir<sup>3</sup> and Rumana Rois<sup>1</sup>

## Abstract

We aimed to assess different machine learning techniques for predicting infant mortality (<1 year) in Bangladesh. The decision tree (DT), random forest (RF), support vector machine (SVM) and logistic regression (LR) approaches were evaluated through accuracy, sensitivity, specificity, precision, F1-score, receiver operating characteristics curve and *k*-fold cross-validation via simulations. The Boruta algorithm and chi-square ( $\chi^2$ ) test were used for features selection of infant mortality. Overall, the RF technique (Boruta: accuracy = 0.8890, sensitivity = 0.0480, specificity = 0.9789, precision = 0.1960, F1-score = 0.0771, AUC = 0.6590;  $\chi^2$ : accuracy = 0.8856, sensitivity = 0.0536, specificity = 0.9745, precision = 0.1837, F1-score = 0.0828, AUC = 0.6480) showed higher predictive performance for infant mortality compared to other approaches. Age at first marriage and birth, body mass index (BMI), birth interval, place of residence, religion, administrative division, parents education, occupation of mother, media-exposure, wealth index, gender of child, birth order, children ever born, toilet facility and cooking fuel were potential determinants of infant mortality in Bangladesh. Study findings may help women, stakeholders and policy-makers to take necessary steps for reducing infant mortality by creating awareness, expanding educational programs at community levels and public health interventions.

**Keywords:** Machine learning, Random forest, Boruta algorithm, Chi-square, Infant mortality

## Introduction

Infant mortality, death of infants less than 1 year, is the most important indicator for child health development in a country [1]. The reduction in child mortality particularly infants death is one of the most important Millennium Development Goals (MDGs) [2]. Globally, the infant mortality rate (per 1000 live births) decreased to 29 from 65 deaths in 2018 from 1990 [3]. In Bangladesh, this rate also declined to 38 in 2014 from 87 deaths in 1993 [4]. To meet the MDG-4 and sustainable development

goals (SDGs), the reduction of infant mortality can play an important role in improving child health [5]. Bangladesh is still far away to achieve the target of SDGs for the infant mortality rate (5 deaths per 1000 live births) [4, 6].

Several studies measured the consequences of infant mortality and also investigated the significant factors affecting infant mortality in different countries. For example, maternal education and antenatal care were potential predictors of infant mortality [7, 8]. The birth interval (duration of two subsequent pregnancies) was identified as a significant factor for infant mortality [9, 10]. Moreover, multiple and preterm births, mother's age, domestic abuse, place of residence, wealth index and metabolic abnormalities were statistically significant determinants of infant mortality [11, 12]. Higher educational attainment of mother's, low birth weight and birth size were the most important factors of infant mortality [13, 14]. In Bangladesh, antenatal care during the pregnancy period, place of delivery, wealth status, birth size,

This paper is an extended version of "Machine Learning Algorithm for Analysing Infant Mortality in Bangladesh" by Rahman et al. which appeared in the Proceedings of the 10th International Conference on Health Information Science (HIS 2021), Melbourne, Australia, October 2021.

\*Correspondence: arahman@juniv.edu

<sup>1</sup> Department of Statistics, Jahangirnagar University, Dhaka, Bangladesh  
Full list of author information is available at the end of the article

and sex of child were identified as significant predictors of infant mortality [4, 15].

Most of the mortality-associated previous studies analysed the data using logistic regression (LR), notably for the dichotomous outcome variable. However, the LR approach is often challenging and not suitable for the estimation of model parameters and predicting mortality. Recently, machine learning (ML) techniques have been developed as further advancement of modelling health data incorporating artificial intelligence and exploring the unknown relationships or patterns from a huge volume of data [16]. In health sciences and medical research, the ML approach is widely used for predicting numerous clinical responses based on multiple covariates [17, 18] and for others: fever detection from twitter [19], identification of diabetic eye diseases [20], epilepsy detection [21], cardiac arrhythmia detection [22] etc. Therefore, various well-known ML techniques: decision tree (DT), random forest (RF), support vector machine (SVM) and LR were assessed via simulations for the classification of infant mortality and identification of its predicting factors in Bangladesh. The systematic assessment of these ML techniques was done through simulation studies by comparing their accuracy, sensitivity, specificity, precision, F1-score, receiver operating characteristic (ROC) and *k*-fold cross-validation.

## Materials and methods

### Data and variables

This study used representative cross-sectional survey data of infant mortality extracted from the latest Bangladesh Demographic and Health Survey (BDHS), 2017–2018 [23]. A two-stage stratified random sampling design was used to collect data. In the first stage, enumeration areas (EAs) were selected using probability proportional to size (PPS) of EAs. A complete list of households was then carried out in all selected EAs to provide a sampling frame for the second-stage household selection process. This stage of sampling was then performed using a systematic sampling design which selected 30 households per EAs on average to provide statistically reliable estimates of key demographic and health variables for the country. Finally, the survey selected 20250 households in total and 20,100 women of reproductive age group 15–49 years were interviewed. The detailed information can be accessed at <https://dhsprogram.com/data/available-datasets.cfm>. In total, 26145 infant records were used after filtering out all missing cases in this study.

The binary indicator (yes, no) of infant mortality was considered as the outcome variable of interest. Based on literature various maternal, socio-economic, demographic and environmental factors: mother's age at first marriage, age at first birth, body mass index (BMI) of

mother's, birth interval, antenatal care visits, tetanus toxoid (TT) injection, administrative division, place of residence, religion, parents education, mother's occupation and empowerment, media-exposure, children ever born, gender of child, birth order, toilet facility, drinking water and type of cooking fuel were considered as covariates of infant mortality.

### Machine learning models

We adapted four different machine learning techniques: decision tree (DT), random forest (RF), support vector machine (SVM), and logistic regression (LR). The DT, commonly used machine learning technique, that develops prediction algorithm for the target variable. The total population is divided into branch-like sectors forming an inverted tree with roots, inner and leaf nodes in this approach. The algorithm can handle a large amount of data without imposing a complicated parametric structure [24, 25]. The RF is a tree-structured classifier technique that depends on a collection of random variables. The number of trees and the highest depth of each tree are identified using hyper-parameters in the RF technique [26]. The RF is a group-learning method for the classification of outcome variables utilizing a large number of decorrelated decision trees [27]. The SVM is a supervised machine learning algorithm that can be used to analyze the data for classification and regression problems [28, 29]. The SVM training process creates a model or classification function that assigns new observations to one group on any side of the hyper-plane, generating a non-probabilistic dichotomous linear classifier for a two-groups learning problem. In the SVM approach, the kernel trick is employed for mapping the data in a multi-dimensional space before addressing the machine learning job treated as an optimization problem [30]. The best performed sigmoid kernel was used in this study. The LR is a commonly used statistical technique for classification problems and predicting the probability of occurrence of an event (infant mortality). The association between infant mortality (binary response) and a set of covariates is modelled in the LR approach [27].

### Prediction performance parameters

We considered the prediction performance parameters of the confusion matrix, a graphical technique of true against predicted classifications in the form of true positive ( $T^{+ve}$ ), false positive ( $F^{+ve}$ ), true negative ( $T^{-ve}$ ) and false negative ( $F^{-ve}$ ). The different performance measures: accuracy (total data values correctly classified into the true groups), sensitivity (correctly classifying data values in a positive group), specificity (accurately classifying data points in a negative group) and precision (number of data values accurately classified from the positive group)

of a confusion matrix are calculated as: accuracy =  $(T^{+ve} + T^{-ve}) / (T^{+ve} + T^{-ve} + F^{+ve} + F^{-ve})$ , sensitivity =  $T^{+ve} / (T^{+ve} + F^{-ve})$ , specificity =  $T^{-ve} / (T^{-ve} + F^{+ve})$  and precision =  $T^{+ve} / (T^{+ve} + F^{+ve})$  [27]. The receiver operating characteristic (ROC) curve is a two-dimensional diagram used to visualize, organise, and select classifiers depending on machine learning model performances [31]. The true-positive data points are plotted against the false-positive values to reveal the sensitivity of the classifier. The area under the receiver operating characteristic curve (AUC) measures how well a performance parameter classifies data values in two diagnostic groups. The true positive rate increases and the AUC value is approximately 1 for the outstanding classifier [24].

**k-Fold cross-validation**

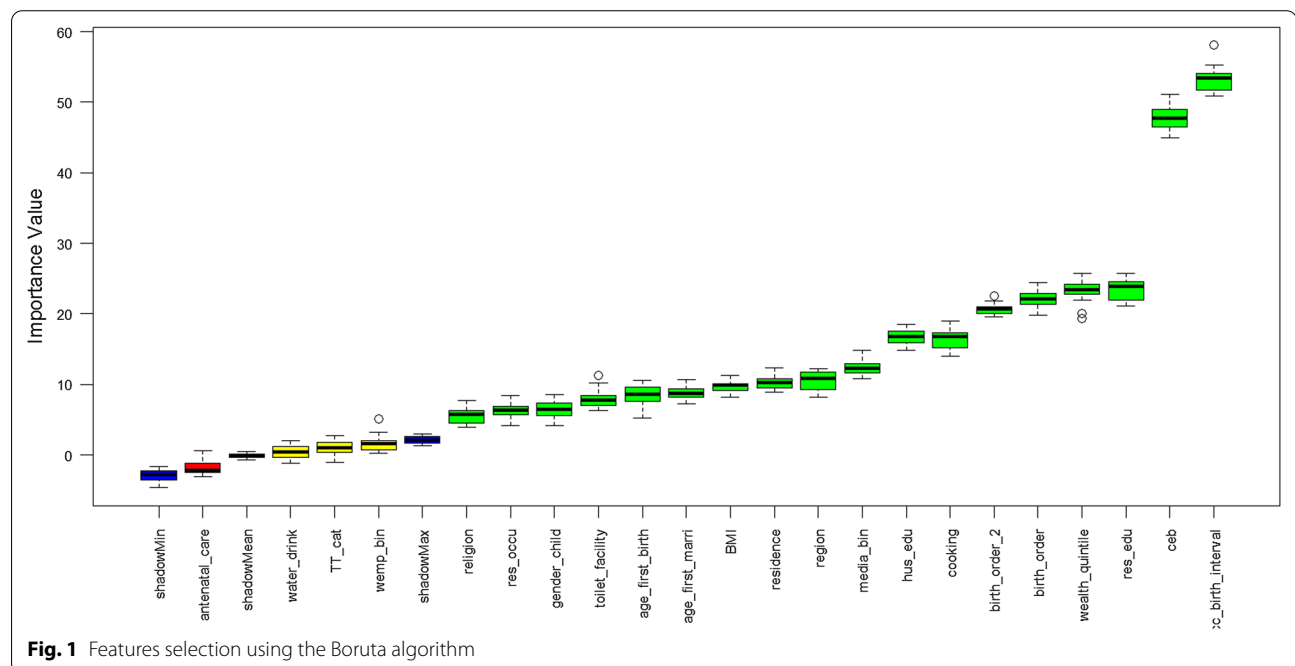
The cross-validation (verification approach) is used for evaluation of the generalization capability of a model to an independent data set [27]. The training data set is randomly subdivided into *k* sub-samples (mutually exclusive folds) of equal sizes in the *k*-fold cross-validation technique. The model is trained *k* times (or folds) iteratively, where *k* sub-samples are used for testing (cross-validation) in every iteration and the remaining (*k* - 1) sub-samples are then used for training the model. An average of *k* cross-validation results is used for estimating the accuracy of ML techniques [27]. The 3, 5, 10 and 20-folds cross-validation approaches were used to evaluate the predictive performance of classifiers for a large sample size (n = 26145) considered in this study.

**Results**

**Features selection of infant mortality**

We first conducted chi-square ( $\chi^2$ ) tests for the feature selection of infant mortality [32]. Based on literature, initially twenty one exposure variables were considered to examine the association between these variables and infant mortality. The statistically significant associations were observed between infant mortality and covariates: mother’s BMI ( $\chi^2 = 52.295$ ,  $p < 0.001$ ), birth interval ( $\chi^2 = 957.434$ ,  $p < 0.001$ ), region ( $\chi^2 = 27.940$ ,  $p < 0.05$ ), religion ( $\chi^2 = 7.723$ ,  $p < 0.05$ ), maternal education ( $\chi^2 = 82.227$ ,  $p < 0.001$ ), father’s education ( $\chi^2 = 63.747$ ,  $p < 0.001$ ) mother’s occupation ( $\chi^2 = 10.889$ ,  $p < 0.001$ ), mass-media exposure ( $\chi^2 = 9.662$ ,  $p < 0.05$ ), wealth index ( $\chi^2 = 25.853$ ,  $p < 0.001$ ), gender of child ( $\chi^2 = 46.344$ ,  $p < 0.001$ ), birth order ( $\chi^2 = 12.302$ ,  $p < 0.05$ ), children ever born ( $\chi^2 = 252.454$ ,  $p < 0.001$ ), households toilet facility ( $\chi^2 = 14.479$ ,  $p < 0.001$ ), and households cooking fuel ( $\chi^2 = 8.757$ ,  $p < 0.001$ ).

In addition, the Boruta algorithm, implemented in R-package: Boruta [33], was used to find out potential predictors of infant mortality based on the mean decreased accuracy [34]. The output is shown in Fig. 1. The selected potential predictors with the green box, and the tentative and rejected ones (which hold less mean decrease accuracy) with the yellow and red boxes, respectively. Hereafter, the Boruta algorithm identified a total of seventeen features including three additional predictors (mother’s age at first marriage, age at first birth



**Fig. 1** Features selection using the Boruta algorithm

and place of residence) with fourteen features extracted from the  $\chi^2$  test.

**Evaluation of models**

The selected important possible predictors (using the Boruta algorithm and chi-square test) of infant mortality were considered to assess the performances of various ML techniques using the parameters of confusion matrix and area under the ROC curve (Fig. 2). The performance parameters: accuracy scores, sensitivity, specificity, precision, and F1-score calculated from the confusion matrix are summarized in Table 1. It was observed that the LR approach performed better compared to other machine learning techniques based on its higher accuracy and specificity scores for the extracted predictors using both of the methods (Boruta and  $\chi^2$  test). Although, the LR model produces the highest accuracy (accuracy = 0.903) and specificity score (specificity = 1.00) in both scenarios, but it completely failed to assess the precision of the test. The sensitivity and F1-score were also zero for the LR

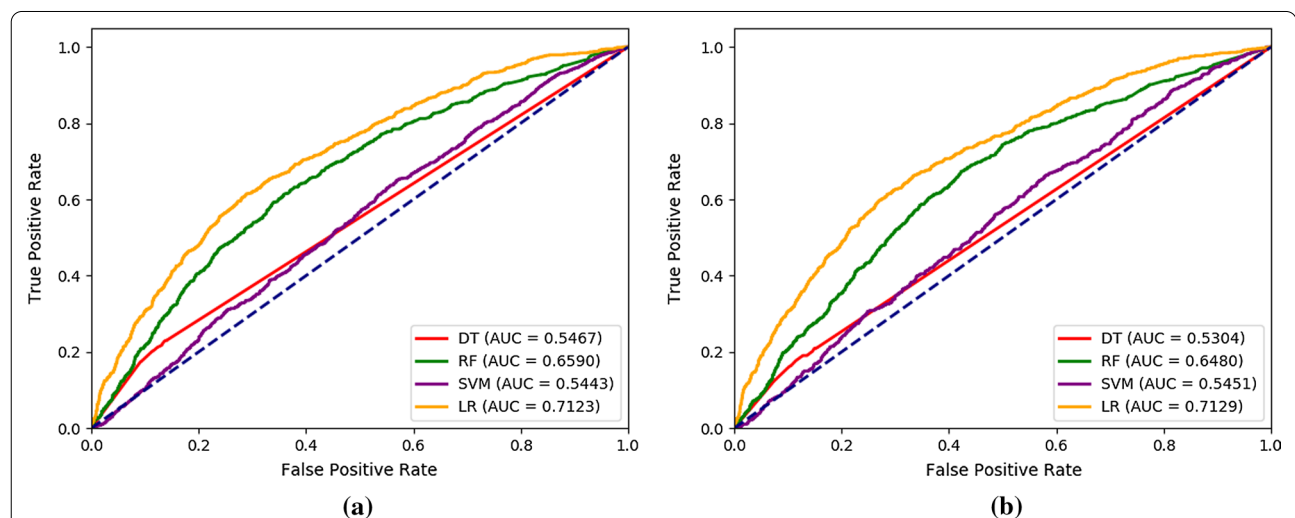
model in both cases. However, the RF model performed better for the selected important predictors associated with infant mortality using the Boruta algorithm and chi-square test. For example, when the important determinants were extracted using the Boruta algorithm, the RF technique showed 89.3% of precise predictions (accuracy = 0.893), 33.9% of true positive cases (sensitivity=0.339), 97.9% of true negative cases (specificity=0.979), 71.5% of correct positive predictions (precision = 0.715), and 46.0% of reasonable precision and recall (F1-score = 0.460). The RF model also performed better when the potential predictors associated with infant mortality were identified using the  $\chi^2$  test.

Figure 2 shows the predicted AUC of different machine learning techniques: DT, RF, SVM, and LR, which were performed in Python (version 3.7.3) using the random seed 1119 for the selected risk factors (Boruta algorithm and  $\chi^2$  test). It was observed that the LR model showed the maximum AUC compared to other ML techniques, but this model failed for the classification of true positive

**Table 1 Accuracy, sensitivity, specificity, precision and F1-score of different machine learning models**

Features selection	Models	Accuracy	Sensitivity	Specificity	Precision	F1-score
Boruta	DT	0.837	0.164	0.909	0.163	0.164
	RF	0.893	<b>0.339</b>	0.979	<b>0.715</b>	<b>0.460</b>
	SVM	0.823	0.103	0.900	0.099	0.101
	LR	<b>0.903</b>	0.000	<b>1.000</b>	N/A	0.00
$\chi^2$ Test	DT	0.849	0.122	0.927	0.152	0.135
	RF	0.886	<b>0.434</b>	0.977	<b>0.168</b>	<b>0.169</b>
	SVM	0.824	0.103	0.902	0.101	0.102
	LR	<b>0.903</b>	0.000	<b>1.000</b>	N/A	0.000

Highest values are indicated in bold, N/A not applicable



**Fig. 2** ROC curves to predict infant mortality of Bangladesh using DT, RF, SVM and LR based on the selected potential factors using Boruta algorithm in (a) and  $\chi^2$  test in (b)

cases of infant mortality. Thus, the performance of the RF model was relatively higher in all scenarios. However, these performances were evaluated based on a single realization (run with the random seed 1119). Hence, a simulation study is more desirable to explore the overall performances of the ML models.

**Simulation study**

A simulation study and the *k*-fold cross-validation were conducted to assess the overall performances of different ML techniques for the prediction of infant mortality in Bangladesh. In each run, we simulated 1000 times with the random seed 1111 to 2111 using 70% training data and 30% test data. The five performance measures of the confusion matrix with their uncertainty estimates based on 1000 simulations are organized in Table 2. Based on the simulation results, the RF technique performed better when the associated factors were identified using the Boruta algorithm as well as the chi-square test.

The summary results of *k*-fold cross-validation for various ML techniques with 10, 20 and 30-folds repetitions

with random seed number 1 and shuffle argument *True* are given in Table 3. It was observed that the mean accuracy values from each fold were higher and standard errors of accuracy scores were also minimum for the LR model than other ML techniques. But, the LR model was unable to adequately predict the true positive cases of infant mortality, as the sensitivity score was 0 and the precision could not be calculated. Thus, to predict infant mortality, the RF (avoiding the LR model performances) was found to be the most suitable modelling technique compared to other machine learning approaches considered in this study.

**Predicting infant mortality using the random forest (RF)**

For the full data set, the best performing ML algorithm, the RF technique, was fitted to predict infant mortality using the extracted potential factors based on the Boruta algorithm and the  $\chi^2$  test. Model interpretation is an important part of the evaluation process [35]. The explanation of the association between model variables and output is quite challenging for “black

**Table 2 The mean estimated value of performance parameters with their uncertainty estimates of different ML models based on 1000 simulations**

Features selection	Models	Accuracy		Sensitivity		Specificity		Precision		F1-score	
		MAcc	SE	MAcc	SE	MAcc	SE	MAcc	SE	MAcc	SE
Boruta	DT	0.8367	0.0040	0.1627	0.0132	0.9088	0.0042	0.1602	0.0126	0.1613	0.0121
	RF	0.8890	0.0029	0.0480	0.0068	0.9789	0.0020	0.1960	0.0248	0.0771	0.0103
	SVM	0.8277	0.0055	0.0932	0.0115	0.9062	0.0057	0.0962	0.0130	0.0946	0.0117
	LR	0.9034	0.0028	0.000	0.000	1.000	0.000	N/A	N/A	0.000	0.000
$\chi^2$ Test	DT	0.8458	0.0037	0.1462	0.0124	0.9206	0.0038	0.1644	0.0133	0.1546	0.0121
	RF	0.8856	0.0029	0.0536	0.0074	0.9745	0.0021	0.1837	0.0226	0.0828	0.0107
	SVM	0.8274	0.0044	0.0936	0.0114	0.9058	0.0046	0.0961	0.0118	0.0947	0.0112
	LR	0.9034	0.0028	0.000	0.000	1.000	0.000	N/A	N/A	0.000	0.000

MAcc Mean accuracy scores from each fold, SE standard error of accuracy scores

**Table 3 Results of *k*-fold cross-validation for different machine learning models**

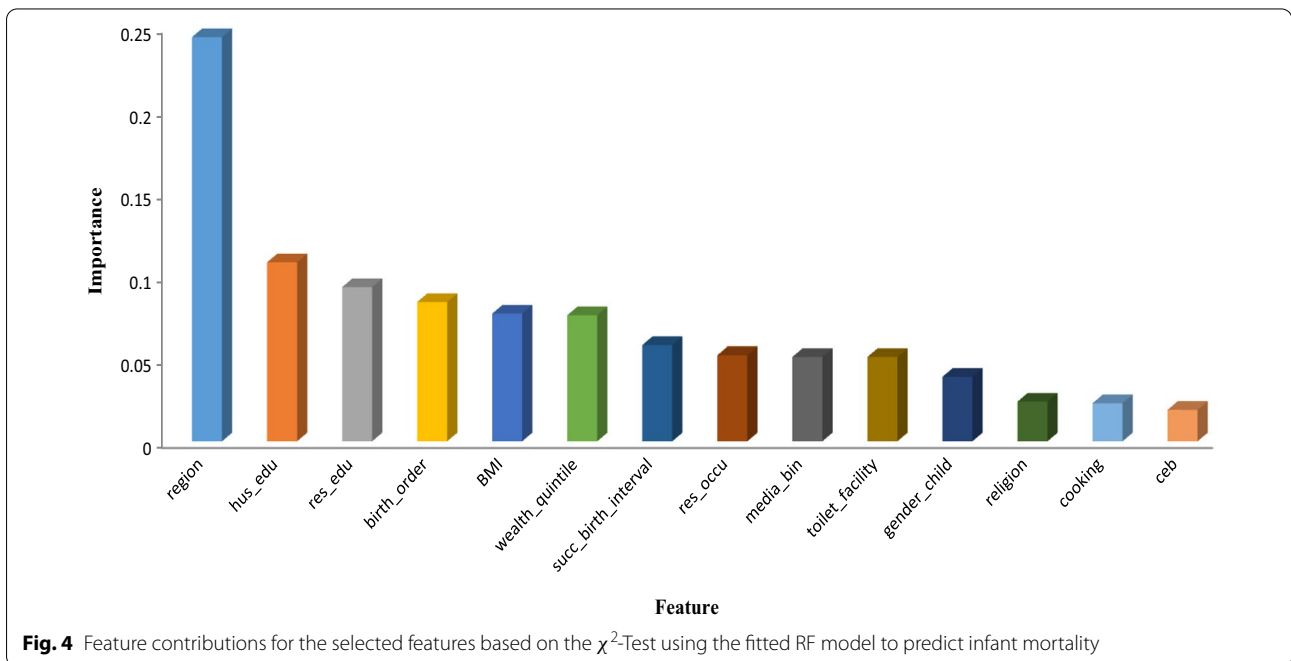
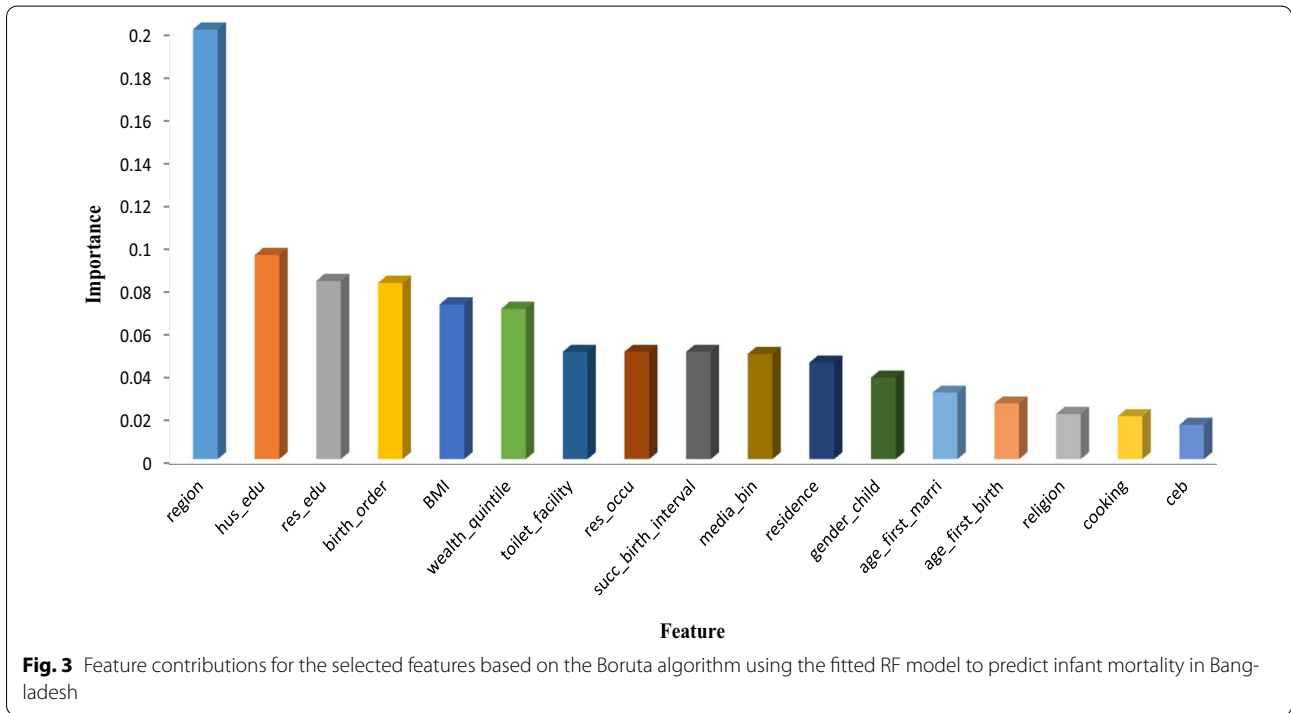
Features selection	Models	10-fold		20-fold		30-fold	
		MAcc	SE	MAcc	SE	MAcc	SE
Boruta	DT	0.8389	0.0073	0.8378	0.0095	0.8405	0.0110
	RF	0.8878	0.0049	0.8882	0.0086	0.8877	0.0093
	SVM	0.8474	0.0056	0.8471	0.0079	0.8480	0.0095
	LR	0.9035	0.0045	0.9035	0.0075	0.9035	0.0091
$\chi^2$ Test	DT	0.8489	0.0072	0.8499	0.0100	0.8513	0.0128
	RF	0.8843	0.0070	0.8848	0.0108	0.8847	0.0113
	SVM	0.8413	0.0056	0.8409	0.0084	0.8399	0.0107
	LR	0.9034	0.0045	0.9035	0.0075	0.9035	0.0091

MAcc Mean accuracy scores from each fold, SE standard error of accuracy scores

box” models like RF, since the data is hidden within the model structure. Feature contributions were calculated independently for each selected factor, providing precise information about the association between variables and the predicted value of infant mortality. These results are organized in Figs. 3 and 4 , where the

selected factors were identified using the Boruta algorithm and the  $\chi^2$  test, respectively.

Figure 3 illustrates that among the seventeen selected factors using the Boruta algorithm for the prediction of infant mortality, feature contributions were higher for administrative division (20.0%), father’s education (9.5%),





mother's education (8.3%), birth interval (8.2%), mother's BMI (7.2%), and wealth index (7.0%). Conversely, feature contributions were lower for mother's age at first marriage (3.1%), age at first birth (2.6%), religion (2.1%), cooking fuel (2.0%), and total children ever born (1.6%).

Figure 4 explains that among the fourteen selected variables using the  $\chi^2$  test, feature contributions were higher for administrative division (24.4%), father's education (10.8%), mother's education (9.3%), birth order (8.4%), mother's BMI (7.7%), and wealth index (7.6%). Whereas, feature contributions were lower for households toilet facility (5.1%), child's gender (3.9%), religion (2.4%), cooking fuel (2.3%), and total children ever born (1.9%) to predict infant mortality in Bangladesh.

## Discussion and conclusion

Globally, infant mortality is a major public health concern. In our study, four different well-known machine learning techniques: DT, RF, SVM and LR were adapted to assess the potential predictors of infant mortality in Bangladesh. These techniques were evaluated by using 70% values as the training data and remaining 30% observations as test data. All the machine learning techniques were evaluated based on the selected important features extracted by using both the Boruta algorithm and chi-square test. The predictive performances of all models were compared using the different performance parameters (accuracy, sensitivity, specificity, precision, and F1-score) of the confusion matrix, the AUC, and the  $k$ -fold cross-validation techniques via simulations.

The traditional  $\chi^2$  test identified fourteen factors: administrative division, birth interval [9, 10], religion, body mass index (BMI), education of parents [7, 13], occupation of mother, children ever born, gender of child [4, 15], exposure of media, birth order [36], wealth index, toilet facility and cooking fuel as significant predictors of infant mortality in Bangladesh, while the Boruta algorithm also suggested three more additional predictors: age at first marriage, age at first birth and place of residence [37]. Therefore, infant mortality was significantly higher among mothers who got married in their teens (before 18 years) and gave their first birth before 20 years [38, 39]. Interval of pregnancy was significantly associated with infant mortality and hence the short intervals between two pregnancies ( $\leq 2$  years) can increase the risk of infant mortality [9, 10, 36]. Along with birth interval, infant mortality was higher among women from rural than urban areas [9]. Educational attainment of both mothers and fathers was also found to be a significant factor for infant mortality. Similar to other studies, a lower risk of infant death has been observed among higher educated mothers and fathers compared with their counterparts [7, 8, 13]. The prevalence of infant

mortality decreased for mothers who had mass-media exposure, higher birth order and better sanitation facility [40, 41]. The poor economic status of the family also leads to high infant mortality [42].

Though accuracy and specificity scores were found to be higher for the LR model (Table 1), but this model failed to accurately classify true positive cases of infant mortality. Therefore, the LR was not capable to compute the precision and hence provided 0 (zero) for both the sensitivity and F1-score. On the other hand, the random forest technique predicted the infant mortality better compared to other machine learning approaches. Based on simulation results, the RF technique was found to be superior among all different machine learning models considered in this study with an accuracy of 88.90%, a sensitivity of 4.80%, a specificity of 97.89%, a precision of 19.6%, and a F1-score of 7.71% when the important features were extracted using the Boruta algorithm. Moreover, the highest mean of accuracy scores for 10-fold (MAcc = 88.78%), 20-fold (MAcc = 88.82%) and 30-fold (MAcc = 88.77%) repetitions appeared in the RF classification technique.

### Acknowledgements

The authors would like to acknowledge the Bangladesh Demographic and Health Survey (BDHS) authority for the freely available secondary data.

### Funding

There is no funding for this work.

### Data availability

The secondary data are available at <https://dhsprogram.com/data/available-datasets.cfm>.

### Declarations

#### Conflict of interest

No conflict of interest exists among the authors.

#### Ethical approval

The National Research Ethics Committee of the Bangladesh Medical Research Council and ICF Macro Institutional Review Board approved BDHS, 2017–2018. The interview was conducted with the prior written approval of survey participants.

#### Consent for publication

Not applicable.

#### Author details

<sup>1</sup>Department of Statistics, Jahangirnagar University, Dhaka, Bangladesh.

<sup>2</sup>Department of Statistics, University of Dhaka, Dhaka, Bangladesh. <sup>3</sup>School of Sciences, University of Southern Queensland, Toowoomba, Australia.

Received: 21 January 2022 Accepted: 1 June 2022

Published online: 21 June 2022

### 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). 2018. Millennium development goals (MDGs). <http://www.who.int/topics/millennium-development-goals/about/en/> accessed 14 July 2021.
3. World Health Organization (WHO). 2018. The global health observatory. <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 KK. Risk factors of infant mortality in Bangladesh. *Clin Epidemiol Glob Health*. 2020;8:211–4. <https://doi.org/10.1016/j.cegh.2019.07.003>.
5. Hajizadeh M, Nandi A, Heymann J. Social inequality in infant mortality: what explains variation across low and middle income countries? *Soc Sci Med*. 2014;101:36–46. <https://doi.org/10.1016/j.socscimed.2013.11.019>.
6. World Health Organization (WHO). 2015. Success factor for women's and child's health: Bangladesh. [www.who.int](http://www.who.int).
7. Quansah E, Ohene LA, Norman L, et al. Social factors influencing child health in Ghana. *PLoS ONE*. 2016. <https://doi.org/10.1371/journal.pone.0145401>.
8. Kiross GT, Chojenta C, Barker D, et al. The effect of maternal education on infant mortality in Ethiopia: a systematic review and meta-analysis. *PLoS ONE*. 2019;14:7. <https://doi.org/10.1371/journal.pone.0220076>.
9. Khadka KB, Lieberman LS, Giedraitis V, et al. The socio-economic determinants of infant mortality in Nepal: analysis of Nepal demographic health survey, 2011. *BMC Pediatr*. 2015;15:152. <https://doi.org/10.1186/s12887-015-0468-7>.
10. Santos SL, Santos LB, Campelo V, et al. Factors associated with infant mortality in a northeastern Brazilian capital. *Rev Bras Ginecol Obstet*. 2016;38(10):482–91. <https://doi.org/10.1055/s-0036-1584686>.
11. Baraki AG, Akalu TY, Wolde HF, 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*. 2020;20:299. <https://doi.org/10.1186/s12884-020-03002-x>.
12. Mohamoud YA, Kirby RS, Ehrenthal DB. Poverty, urban–rural classification and term infant mortality: a population-based multilevel analysis. *BMC Pregnancy Childbirth*. 2019;19:40. <https://doi.org/10.1186/s12884-019-2190-1>.
13. 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*. 2013;13:401. <https://doi.org/10.1186/1471-2458-13-401>.
14. Vilanova CS, Hiraoka VN, Buriol VCC, 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*. 2019. <https://doi.org/10.1186/s12963-019-0195-7>.
15. Dancer D, Rammohan A, Smith MD. Infant mortality and child nutrition in Bangladesh. *Health Econ*. 2008;17(9):1015–35. <https://doi.org/10.1002/hec.1379>.
16. Alghamdi M, Al-Mallah M, Keteyian S, et al. Predicting diabetes mellitus using SMOTE and ensemble machine learning approach: the Henry Ford Exercise Testing (FIT) project. *PLoS ONE*. 2017;12: e0179805. <https://doi.org/10.1371/journal.pone.0179805>.
17. Mateen BA, Liley J, Denniston AK, et al. Improving the quality of machine learning in health applications and clinical research. *Nat Mach Intell*. 2020;2(10):554–6. <https://doi.org/10.1038/s42256-020-00239-1>.
18. Sarki R, Ahmed K, Wang H, et al. Automated detection of mild and multi-class diabetic eye diseases using deep learning. *Health Inf Sci Syst*. 2020;8(1):1–9. <https://doi.org/10.1007/s13755-020-00125-5>.
19. Du J, Michalska S, Subramani S, et al. Neural attention with character embeddings for hay fever detection from Twitter. *Health Inf Sci Syst*. 2019. <https://doi.org/10.1007/s13755-019-0084-2>.
20. Sarki R, Ahmed K, Wang H, et al. Image preprocessing in classification and identification of diabetic eye diseases. *Data Sci Eng*. 2021;6:455–71. <https://doi.org/10.1007/s41019-021-00167-z>.
21. Supriya S, Siuly S, Wang H, et al. Automated epilepsy detection techniques from electroencephalogram signals: a review study. *Health Inf Sci Syst*. 2020;8(1):1–15. <https://doi.org/10.1007/s13755-020-00129-1>.
22. He J, Rong J, Sun L, et al. A framework for cardiac arrhythmia detection from IoT-based ECGs. *World Wide Web*. 2020;23(5):2835–50. <https://doi.org/10.1007/s11280-019-00776-9>.
23. National institute of population research and training (NIPROT). Bangladesh demographic and health survey 2017–2018. National institute of population research and training (NIPROT), Mitra and Associates, Dhaka, Bangladesh and ICF International, Calverton, Maryland, USA, 2019.
24. Igual L, Seguí S. Introduction to data science. Cham: Springer; 2017.
25. Nilsson NL. Introduction to machine learning; 1997.
26. Breiman L. Random forests. *Mach Learn*. 2001;45(1):5–32.
27. Awad M. Efficient Khanna, R machines learning. Berkeley: A press; 2015. <https://doi.org/10.1007/978-1-4302-5990-9-1>.
28. Burges CJ. A tutorial on support vector machines for pattern recognition. *Data Min Knowl Disc*. 1998;2(2):121–67.
29. Müller KR, Mika S, Rätsch G, et al. An introduction to kernel-based learning algorithms. *IEEE Trans Neural Netw*. 2001;12(2):181–201.
30. Vapnik VN. The nature of statistical learning theory. New York: Springer; 1995.
31. Fawcett T. An Introduction to ROC Analysis. *Pattern Recogn Lett*. 2006;27:861–74.
32. Rahman A, Hossain Z, Kabir E, et al. Machine learning algorithm for analysing infant mortality in Bangladesh. *International Conference on Health Information Science 2021*;205–219.
33. R Core Team. R: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing. <http://www.R-project.org/>; 2013.
34. Kursa MB, Rudnicki WR. Feature selection with the Boruta package. *J Stat Softw*. 2010;36(11):1–13.
35. Palczewska A, Palczewski J, Robinson RM, et al. Interpreting random forest classification models using a feature contribution method In *Integration of reusable systems*. Cham: Springer; 2014. p. 193–218.
36. Hajipour M, Taherpour N, Fateh H, et al. Predictive factors of infant mortality using data mining in Iran. *J Compr Ped*. 2021;12:1. <https://doi.org/10.5812/compreped.108575>.
37. de Bitencourt FH, Schwartz IVD, Vianna FSL. Infant mortality in Brazil attributable to inborn errors of metabolism associated with sudden death: a time-series study (2002–2014). *BMC Pediatr*. 2019;19:52. <https://doi.org/10.1186/s12887-019-1421-y>.
38. Hossain MM, Abdulla F, Banik R, et al. Child marriage and its association with morbidity and mortality of under-5 years old children in Bangladesh. *PLoS ONE*. 2022;17:2. <https://doi.org/10.1371/journal.pone.0262927>.
39. Finlay JE, Özaltın E, Canning D. The association of maternal age with infant mortality, child anthropometric failure, diarrhoea and anaemia for first births: evidence from 55 low- and middle-income countries. *BMJ Open*. 2011. <https://doi.org/10.1136/bmjopen-2011-000226>.
40. Rahman MM, Ara T, Mahmud S, et al. Revisit the correlates of infant mortality in Bangladesh: findings from two nationwide cross-sectional studies. *BMJ Open*. 2021. <https://doi.org/10.1136/bmjopen-2020-045506>.
41. Karmaker SC, Lahiry S, Roy DC, et al. Determinants of infant and child mortality in Bangladesh: time trends and comparisons across South Asia. *Bangladesh J Med Sci*. 2014. <https://doi.org/10.3329/bjms.v13i4.20590>.
42. Singh A, Pathak PK, Chauhan RK, et al. Infant and child mortality in India in the last two decades: a geospatial analysis. *PLoS ONE*. 2011;6(11):2011e26856.