Research & Reviews: Journal of Medical and Health Sciences

e-ISSN:2319-9865
p-ISSN:2322-0104

Prediction of Infant Mortality in Bangladesh Using SVMs

Rumana Rois*,Faruk Hasan and Mst. Nilufar Yasmin

Department of Statistics, Jahangirnagar University, Savar, Dhaka, Bangladesh

Research Article

Received date: 23/11/2021
Accepted date: 01/122021
Published date: 08/12/2021

***For Correspondence**
Rumana Rois, Department of Statistics, Jahangirnagar University, Savar, Dhaka, Bangladesh

E-mail: rois@juniv.edu

Keywords: Infant Mortality, Machine learning, Decision Tree, Random Forest, Support vector machine, Logistic Regression, Accuracy, Precision, Sensitivity, Specificity, ROC, K-fold cross-validation.

ABSTRACT

Infant mortality (IM) is one of the most important indicators for the development of a country. Therefore, identification of the risk factors associated with IM is essential to know to develop different effective public health interventions and educational programs. This study, therefore, investigates the determinants and prediction of IM in Bangladesh using support vector machines (SVMs), a popular machine learning (ML) algorithm. The study data based on IM of 43,772 children from the 2014 Bangladesh Demographic and Health Survey (BDHS). The SVM algorithm was used to extract important risk factors associated with IM. Prediction of IM was done using different ML models, for instance, decision tree (DT), random forest (RF), SVM and logistic regression (LR). Performance of these techniques were evaluated via different parameters of confusion matrix, receiver operating characteristics (ROC) curve and k-fold cross-validation. The study revealed that proportion of IM was 7.4% (3243 infants out of 43772 children). Husband/partner's education level and occupation, mother's age at first birth and body mass index, total children ever born, type of cooking fuel, wealth index, and division were selected as significant features of predicting IM using the SVM, whereas the conventional chi-square test showed different results. To predict IM in Bangladesh, the SVM with Gaussian kernel (Accuracy=0.840, Sensitivity =0. 861, Specificity=0.360, Precision=0.970, area under the ROC curve (AUC)=0.608, k-fold accuracy=0.808) performed better among examined machine learning models.

INTRODUCTION

Infant mortality (IM) is one of the most substantial indicators for the advancement of a country's economy and health sector. This reflects the apparent association between the causes of IM and other factors that are likely to influence the health status of whole populations such as their economic development, living conditions, social well-being, rates of illness, and the quality of the environment [1]. It is defined as the death of an infant before his or her first birthday. The IM rate is the number of infant deaths for every 1,000 live births [2]. One of the important millennium development goals (MDG) is to reduce child mortality, particularly infant mortality, all over the world [3]. Globally, the IM rate has reduced to 29 deaths per 1000 live births in 2018 from 65 deaths per 1000 live births [4].

Certainly, neonatal and child mortality is a vital indicator of the development of a developing country like Bangladesh. Though Bangladesh has achieved the Millennium Development Goal 4 (MDG 4) target by experiencing a significant reduction of child mortality over the past decades, IM is still reasonably high. For instance, the IM rate has been decreased from 87 (in 1993) to 38 (in 2014) in Bangladesh [5]. The reduction in Infant mortality by two-thirds of a country indicating the progress towards achieves the MDG-4 [6]. To meet the sustainable development goals (SDG) 3 target: "By 2030, end preventable deaths of new-borns and children under 5 years of age, with all countries aiming to reduce neonatal mortality to at least as low as 12 per 1,000 live births and under-5 mortality to at least as low as 25 per 1000 live births", reducing IM rates will make a significant contribution to improving children's health [7].

Cause and predictors of IM are different demographic and socio-economic factors, including factors related to infants themselves. One of the independent determinants of IM was low birth weight of infants at birth [8,9,10]. Mothers from low-income families were significantly more likely to experience IM than those who belonged to the middle and rich wealth status [11]. In Bangladesh, mothers who did not delivered baby at institutions and belonged to poor class families were more likely to have infant death than others [12, 5]. Besides these, mothers antenatal care during pregnancy, and gender of child [5]. Conventionally, the chi-square test commonly used to detect the risk or protective factors of IM. However, we are motivated to use machine learning (ML) method, which is a scientific approach that involves artificial intelligence and explores more hidden information from a large volume of data [13], to detect the risk or protective factors of IM and to predict IM in Bangladesh. The application of ML in health research results in improved findings compared to conventional counterparts [14,15]. Therefore, the significant factors associated with IM have been identified using SVM. Furthermore, different well-known ML techniques, for instance, decision trees (DT), random forest (RF), support vector machines (SVM), and LR have been applied in this study for the prediction of IM in Bangladesh.

Methods and Materials

Data and variables

This study uses secondary data for investigating the potential factors of IM in Bangladesh were extracted from the nationally representative Bangladesh Demographic and Health Survey (BDHS) conducted in 2014 under the authority of the National Institute of Population Research and Training (NIPORT) [16]. The survey was conducted by Mitra and Associates from June to November 2014. Funding was provided by the United States Agency for International Development (USAID)/Bangladesh. ICF International provided technical assistance through The DHS Program, a USAID-funded project. The birth record file was used in this study, detailed information of this data is available at https://dhsprogram.com/data/available-datasets.cfm. The information related to IM was collected from reproductive mothers and 43772 infants were included in this study after removing all missing cases. The primary outcome variable of this study is “infant death” which was defined as the death of a live birth before the age of one (coded as 0=no, and 1=yes). 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 mother age, mother Age at 1st birth, highest education level, type of place of residence, division, wealth index, birth order number, sex of child, size of child at birth, access to media, type of cooking fuel, husband’s education level, toilet facilities shared with other households, body mass index, body mass index category, husband’s occupation, total children ever born, told about pregnancy complications, number of antenatal visits during pregnancy, age at death, death before the first birthday, mother’s weight, exposure to NGO activity, place of delivery and mother’s height.

Statistical Models

This study aimed to assess the potential predictors associated with IM and to predict IM in Bangladesh using different ML classification models, e.g., decision tree (DT), random forest (RF), support vector machine (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 (DT, RF, SVM, LR) in the training data set and evaluate the performance of these models on the test data set, and finally using the best performed model to predict IM based on the entire data set. The performances were evaluated using four performance parameters from the confusion matrix such as accuracy, sensitivity, specificity and precision, the area under the receiver operating characteristics (ROC) curve (AUC), and the K-fold cross-validation. The scikit-learn module in Python programming language was used to perform all ML models.

Decision Tree (DT)

One of the most simple and intuitive techniques in ML is DT which is based on the divide and conquer paradigm [17]. A DT, whose leaf nodes are categories (of patterns) and whose input nodes are tests (on input patterns), assigns a class number (or output) to an input pattern by filtering the pattern down through the tests in the tree [18]. Each test has mutually exclusive and exhaustive outcomes [18].

Random Forest (RF)

A RF algorithm has hyper-parameters specifying the number of trees and the maximum depth of each tree (effectively how many interactions are considered in the model), whereas the decision rules are the parameters [19]. An ensemble learning approach for classification using a large collection of decorrelated DT is RF [20]. To implement the RF algorithm in python, we have used 100 DT and Gini for impurity index.

Support Vector Machine (SVM)

A supervised learning methods that analyze data and recognize patterns is known as SVM [21,22]. For a two-class learning task, an SVM training algorithm makes a model or classification function that assigns new observations to one of the two classes on either side of a hyperplane, making it a non-probabilistic binary linear classifier. An SVM model uses the kernel trick to map the data into a higher-dimensional space before solving the ML task as a convex optimization problem [20-24]. New observations are then predicted to belong to a class based on which side of the partition they fall. Support vectors are the data points nearest to the hyperplane that divides the classes [20]. We examined SVM models using the SVM with different kernels.

Logistic Regression (LR)

LR is a probabilistic statistical classification model that predicts the probability of the occurrence of an event [20]. LR models the relationship between a categorical dependent variable and a dichotomous categorical outcome or feature. It is used as a binary (multiple) model to predict binary (multiple) responses, the outcome of a categorical dependent variable, based on one or more independent variables [17].

Confusion Matrix Performance Parameters

A confusion matrix provides a visual representation of actual versus predicted class accuracies [20]. To visualize the performance of the classification algorithm, it compares the predicted classification against the actual classification in the form of false positive, true positive, false negative and true negative information [20]. Therefore, the performance parameters are: accuracy is the number of data points correctly classified by the classifier, sensitivity is a measure of how well a classification algorithm classifies data points in the positive class, specificity is a measure of how well a classification algorithm classifies data points in the negative class, and precision is the number of data points correctly classified from the positive class [20].

Receiver Operating Characteristic (ROC) Curve

ROC curves offer another useful graphical representation for classifiers operating on datasets. Fawcett [24] provided a comprehensive introduction to ROC analysis, highlighting common misconceptions. The ROC curve shows the sensitivity of the classifier by plotting the rate of true positives to the rate of false positives. If the classifier is outstanding, the true positive rate will increase, and the area under the curve (AUC) will be close to 1 [17].

K-fold Cross-Validation

Cross-validation is a verification technique that evaluates the generalization ability of a model for an independent dataset [20]. It evaluates the performance of various prediction functions. In k-fold cross-validation, the training dataset is arbitrarily partitioned into k mutually exclusive subsamples (or folds) of equal sizes. The model is trained k times (or folds), where each iteration uses one of the k subsamples for testing (cross-validating), and the remaining k-1 subsamples are applied toward training the model. The k results of cross-validation are averaged to estimate the accuracy as a single estimation [20]. For this large sample size, we applied 3-fold, 5-fold, 10-fold, and 30-fold cross-validation techniques to evaluate the performance of classifiers.

RESULTS

presents the selected demographic, socio-economic and anthropometric characteristics of reproductive mothers aged 15–49 years and 43772 infants. The study findings reveal that child deaths before at the first birthday was 7.41% in Bangladesh based on BDHS 2014. Infant death was comparatively higher in the rural areas (7.9%), in Sylhet division (8.9%), families with the poorest wealth index (9.1%), the male child (8%), the family has no access to media (8.7%), mothers with no exposure to NGO activity (7.5%), respondents who shared toilet facilities with other households (8%), respondents used agricultural crop as a cooking fuel (9%), respondents who don't know the complications of pregnancies (25%) and respondents who were underweight (8.5%) **Table 1**.

Table 1. Socio-demographic characteristics of infant mortality in Bangladesh based on BDHS 2014.

Characteristics	Death Before First Birthday		χ^2	*P-value*
	No $n =$ 40529 (92.59%)	Yes $n =$ 3243 (7.41%)		
Highest education level				
No education	13407 (90.1%)	1472 (9.9%)	313.868	<0.001*
Primary	13207 (92.4%)	1088 (7.6%)		
Secondary	11731 (94.9%)	630 (5.1%)		
Higher	2184 (92.6)	53 (2.4%)		
Type of place of residence				
Urban	12660 (93.7%)	855 (6.3%)	33.401	<0.001*
Rural	27869 (92.1%)	2388 (7.9%)		
Division				
Barisal	5046(92.7%)	397 (7.3%)	45.221	<0.001*
Chittagong	7097 (93.5%)	491 (6.5%)		
Dhaka	6596 (93.1%)	487 (6.9%)		
Khulna	5296 (93.4%)	376 (6.6%)		
Rajshahi	5165 (91.7%)	468 (8.3%)		
Rangpur	5514 (92.3%)	457 (7.7%)		
Sylhet	5815 (91.1%)	567 (8.9%)		
Wealth Index				
Poorest	8425 (90.9%)	844 (9.1%)	147.822	<0.001*
Poorer	8351 (91.6%)	770 (8.4%)		
Middle	8337 (92.2%)	704 (7.8%)		
Richer	7977 (93.5%)	558 (6.5%)		
Richest	7439 (95.3%)	367 (4.7%)		
Birth order number				
Total	40529 (92.6%)	3243 (7.4%)	120.676	<0.001*

Sex of child				
Male	20597 (92.0%)	1799 (8.0%)	26.017	<0.001*
Female	19932 (93.2)	1444 (6.8)		
Size of child at birth				
Very large	101 (97.1%)	3 (2.9%)	19.956	<0.001*
Larger than average	481 (93.9%)	31 (6.1%)		
Average	3089 (97.0%)	95 (3.0%)		
Smaller than average	599 (96.5%)	22 (3.5%)		
Very small	287 (93.5%)	20 (6.5%)		
Access to media				
No	17689 (91.3%)	1683 (8.7%)	82.865	<0.001*
Yes	22840 (93.6%)	1560 (6.4%)		
Exposure to NGO activity				
No	35275 (92.5%)	2851 (7.5%)	2.051	0.157
Yes	5254 (93.1%)	392 (6.9%)		
Toilet facilities shared with other households				
No	26937 (92.9%)	2072 (7.1%)	17.667	<0.001*
Yes	10912 (92.0%)	954 (8.0%)		
Not a dejure resident	1468 (99.1%)	86 (5.5%)		
Type of cooking fuel				
Electricity	115 (92.7%)	9 (7.3%)	90.029	<0.001*
LPG	617 (96.0%)	26 (4.0%)		
Natural gas	4175 (94.3%)	252 (5.7%)		
Biogas	61 (98.4%)	1 (1.6%)		
Kerosene	27 (96.4%)	1 (3.6%)		
Coal, lignite	111 (91.7%)	10 (8.3%)		
Charcoal	122 (94.6%)	7 (5.4%)		
Wood	21356 (92.9%)	1634 (7.1%)		
Straw/shrubs/grass	413 (92.4%)	34 (7.6%)		
Agricultural crop	8852 (91.0%)	871 (9.0%)		
Animal dung	3139 (91.1%)	305 (8.9%)		
No food cooked in house	1 (100%)	0 (0.0%)		
Other	72 (91.1%)	7 (8.9%)		
Not a dejure resident	1468 (94.5%)	86 (5.5%)		
Husband/partner's education level				
No education	14490 (91%)	1441 (9%)	166.168	<0.001*
Primary	11487 (92.4%)	949 (7.6%)		
Secondary	10001 (93.8%)	665 (6.2%)		
Higher	4548 (96%)	188 (4%)		
Told about pregnancy complications				
No	1813 (97.7%)	43 (2.3%)	8.934	0.093
Yes	1626 (97.7%)	39 (2.63%)		
Don't know	3 (75%)	1 (25%)		
Body mass index				
Underweight	7807 (91.5%)	727 (8.5%)	40.928	<0.001*

*Statistically significant at the 0.05 level.

Table 1. also illustrates that mother's Highest education level, Body mass index (BMI), Access to media, Type of place of residence, Division, Wealth Index, Birth order number, Sex of child, Size of child at birth, Toilet facilities shared with other households, Type of cooking fuel, and Husband/partner's education level were significantly associated with IM based on the *p-value* of the chi-square test. However, we were interested to explore the risk factors associated with IM using ML techniques. Therefore, SVM was used to identify those determinants of IM.

Features Selection using SVM

The important risk predictors of IM were explored using SVM. Once having the fitted SVM with linear kernel, then the important features can be determined by comparing the size of the classifier coefficients using .coef_ argument value. Figure 1 reveals those identified risk predictors with the blue bars and insignificant ones (which hold less variance) with the green bars. Hereafter,

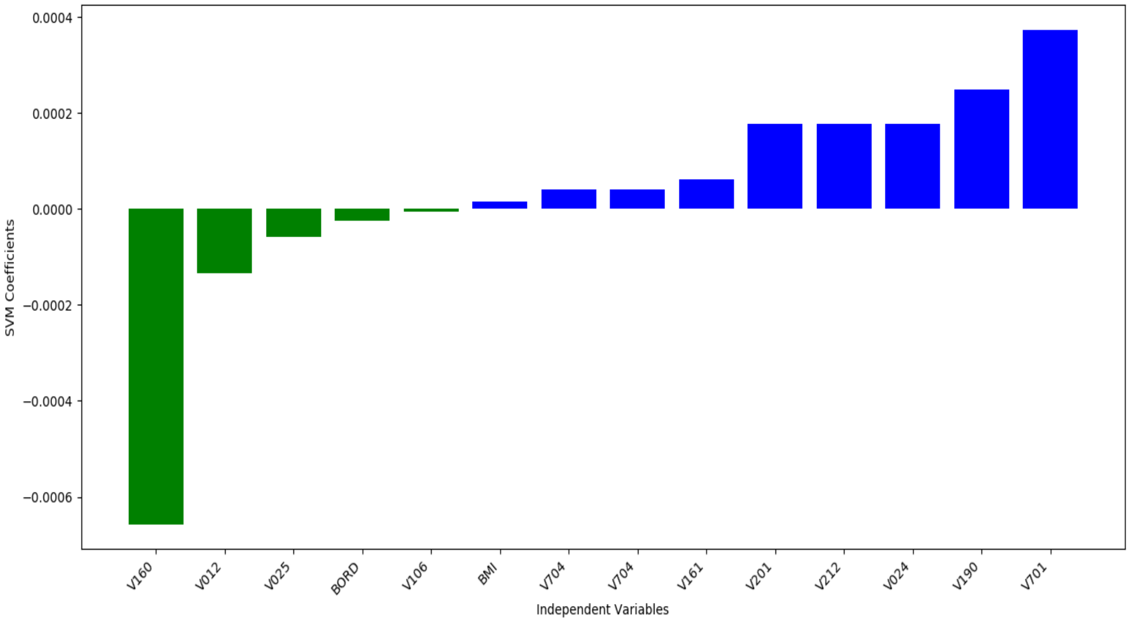

Figure 1. Features selection using SVM.

Table 2: Accuracy, sensitivity, specificity, and precision of different ML models

Models	Accuracy	Sensitivity	Specificity	Precision
DT	0.802	0.882	0.315	0.886
RF	0.836	0.869	0.352	0.953
SVM with Gaussian kernel	0.840	0.861	0.360	0.970
LR	0.854	0.854	N/A	1.00

eight variables were identified the main features (risk predictors) using SVM, for instance, V701 (Husband/partner's education level), V190 (Wealth index), V024 (Division), V212 (Mother age at first birth), V201 (Total children ever born), V161 (Type of cooking fuel), V704 (Husband/partner's occupation), and BMI (Body mass index) to predict IM in Bangladesh **Figure 1**.

Evaluation of Machine Learning Models

The performance of ML models such as DT, RF, SVM and LR were evaluated using four performance parameters of the confusion matrix (Table 2), the area under the ROC curve (Figure 2), and the k-fold cross-validation approaches (Table 3). Considering 70% observations as the training data and 30% observation as the test data with the random seed 100, using the scikit-learn module in python to predict IM in Bangladesh based on BDHS-2014 dataset.

N/A: Not applicable

Table 2 explores the accuracy, sensitivity, specificity and precision of different machine learning model based on BDHS-2014 dataset. Among these models, SVM with Gaussian kernel was the efficient model to predict IM in Bangladesh based on the higher value of the performance parameters in all cases. For instance, the SVM with Gaussian kernel provided 84% of accurate predictions (i.e., accuracy=0.840), 86.1% of positive cases that were predicted as positive (i.e., sensitivity=0.861), 36% of negative cases that were predicted as negative (i.e., specificity=0.360), 97% of positive predictions that were correct (i.e., precision=0.970). Though the logistic regression gives the highest accuracy score (85.4%) among the discussed machine learning models, it was unable to calculate the specificity score due to the convergence problem. Consequently, the SVM with Gaussian kernel performs better among all these models **Table 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 100 To predict IM in Bangladesh the estimated AUC was 0.5395, 0.6003, 0.6082, and 0.54515 using DT, RF, SVM with Gaussian kernel, and LR, respectively. In this figure the SVM with Gaussian kernel performed better with the maximum AUC among all examined ML models. Therefore, the SVM with Gaussian kernel model performances was considered as the better one among all situations **Figure 2**.

Highest values are indicated in bold represents the K-Fold cross validation of different ML models which was performed for 3-Fold, 5-Fold, 10-Fold and 30-Fold repeatedly. The results revealed that SVM (with Gaussian kernel) was performed better in 5-Fold, 10-Fold and 30-Fold cross validation. Consequently, to predict IM in Bangladesh for BDHS-2014, the SVM (with Gaussian kernel) algorithm performed better based on the precision, sensitivity, specificity and accuracy measures, the ROC, and the k-fold cross-validation approaches **Table 3**.

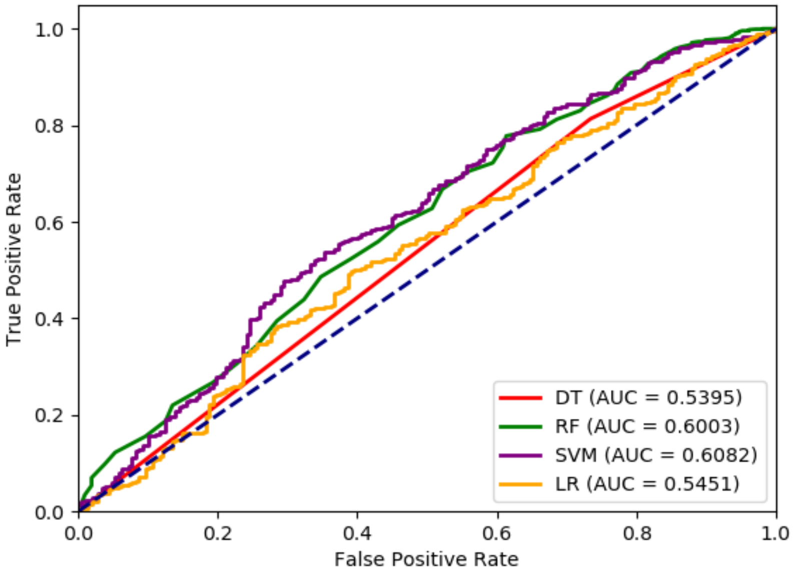

Figure 2. The ROC curves to predict IM in Bangladesh using DT, RF, SVM, and LR models.

Table 3. Result of K-Fold cross-validation of ML Models.

Models	Accuracy (%) – K-Fold			
	3-Fold	5-Fold	10-Fold	30-Fold
Decision Trees	0.705	0.711	0.708	0.702
Random Forests	0.805	0.806	0.806	0.805
SVM (with Gaussian kernel)	0.805	0.808	0.807	0.808
Logistic Regression	0.803	0.803	0.803	0.803

Discussion and Conclusion

This research was conducted to find the significant factors and prediction of IM in Bangladesh using different ML models. Conventional chi-square test revealed that mother's highest education level, BMI, access to media, type of place of residence, division, wealth index, birth order number, sex of child, size of child at birth, toilet facilities shared with other households, type of cooking fuel, and husband/partner's education level were significantly associated with IM in Bangladesh based on BDHS 2014 dataset. However, husband/partner's education level and occupation, mother's age at first birth and BMI, total children ever born, type of cooking fuel, wealth index, and division were selected as significant features of predicting IM using the SVM.

We evaluated the performance of ML models such as DT, RF, SVM, and LR to predict IM in Bangladesh using four performance parameters of the confusion matrix, the AUC (ROC), and the k-fold cross-validation approaches. The SVM (with Gaussian kernel) model was performed better to predict IM with highest performance parameters, i.e., 84% of accuracy, 97% of precision, 86% of sensitivity, 36% of specificity, 60.8% of AUC, 80.8% of accuracy in all the 3, 5, 10-folds, and 30-folds cross-validation techniques. On the other hand, the RF and DT model performed with less performance parameter than SVM. LR model failed to estimate the specificity due to convergence problem. Considering the high accuracy in prediction and better performance, the SVM model will be more informative to predict IM in Bangladesh. Therefore, the findings may help the family members and health-policymakers to understand and prevent this major public health problem.

REFERENCES

1. Reidpath DD and Allotey P. Infant mortality rate as an indicator of population health. J Epidemiol Community Health. 2003;57(5):344-346.
2. CDC: Infant Mortality. Centers for Disease Control and Prevention. 2018.
3. World Health Organization (WHO). Millennium development goals (MDGs). 2018. http://www.who.int/topics/millennium-development-goals/about/en. Accessed 14 July 2021 Accessed 14 July 2021
4. World Health Organization (WHO). The global health observatory. 2018.
5. Vijay J and Patel KK. Risk factors of infant mortality in Bangladesh. Clin Epidemiology Glob Health. 2020;8(1): 211-214.
6. Hajizadeh M et al. Social inequality in infant mortality: What explains variation across low and middle income countries?. Social science & medicine. 2014;101:36-46.
7. World Health Organization (WHO). Success factor for women's and child's health: Bangladesh (2015).

8. Dube L, et al. Determinants of infant mortality in community of Gilgel Gibe Field Research Center, Southwest Ethiopia: a matched case control study. BMC public health. 2013;13(1),1-8.

9. Vilanova CS, 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. Population health metrics. 2019; 17(1),1-12.

10.Hajipour, M et al. Predictive factors of infant mortality using data mining in Iran. J. Compr Pediatr 2021;12(1).

11.Khadka KB ,et al. The socio-economic determinants of infant mortality in Nepal: analysis of Nepal Demographic Health Survey, BMC pediatrics. 2015;15(1),1-11.

12. Dancer D. Infant mortality and child nutrition in Bangladesh. Health economics, 2008;17(9):1015-1035.

13. Alghamdi M. Predicting diabetes mellitus using SMOTE and ensemble machine learning approach: The Henry Ford Exercise Testing (FIT) project. PloS one. 2017;12(7):e0179805.

14. Mateen BA. et al. Improving the quality of machine learning in health applications and clinical research. Nature Machine Intelligence. 2(10):554-556.

15. Rahman A, et al. Machine Learning Algorithm for Analysing Infant Mortality in Bangladesh. 2021;13079: 205-219.

16. National Institute of Population Research and Training (NIPORT), Mitra and Associates, and ICF International. Bangladesh Demographic and Health Survey 2011. Dhaka, Bangladesh and Calverton, MD: NIPORT, Mitra and Associates, and ICF International; 2013.

17. Igual L and Seguí S.Introduction to Data Science. Springer, Cham 2017.

18. Nilsson NL. Introduction to Machine Learning. 1997.

19. Breiman L. Random Forests. Machine Learning. 2001;45(1): 5-32.

20. Awad M and Khanna R. Efficient Learning Machines, 2015. Apress, Berkeley, CA. ss

21. Burges CJ. A tutorial on support vector machines for pattern recognition. Data mining and knowledge discovery, 1998;2(2):121-167.

22. Müller KR, et al. An introduction to kernel-based learning algorithms. IEEE transactions on neural networks. 2001;12(2):181-201.

23. Vapnik VN. The Nature of Statistical Learning Theory 1995. Springer-Verlag New York Inc.

24. Fawcett T. An Introduction to ROC Analysis. Pattern Recognition Letters. 2006;27:861–874.