Journal of Data Science,17(1). P. 195 - 218,2019
DOI:10.6339/JDS.201901_17(1).0009

MACHINE LEARNING ALGORITHMS TO PREDICT THE CHILDHOOD ANEMIA IN BANGLADESH

Jahidur Rahman Khan[1-2*], Srizan Chowdhury[3], Humayera Islam[3-4], Enayetur Raheem[2]

[1]*Centre for Research and Action in Public Health (CeRAPH), Health Research Institute (HRI), Faculty of Health, University of Canberra, Canberra, Australia*

[2]*Biomedical Research Foundation (BRF), Dhaka, Bangladesh*

[3]*Institute of Statistical Research and Training (ISRT), University of Dhaka, Dhaka, Bangladesh*

[4]*Health Management and Informatics Institute, University of Missouri- Columbia, Missouri, USA*

ABSTRACT

Anemia, especially among children, is a serious public health problem in Bangladesh. Apart from understanding the factors associated with anemia, it may be of interest to know the likelihood of anemia given the factors. Prediction of disease status is a key to community and health service policy making as well as forecasting for resource planning. We considered machine learning (ML) algorithms to predict the anemia status among children (under five years) using common risk factors as features. Data were extracted from a nationally representative cross-sectional survey- Bangladesh Demographic and Health Survey (BDHS) conducted in 2011. In this study, a sample of 2013 children were selected for whom data on all selected variables was available. We used several ML algorithms such as linear discriminant analysis (LDA), classification and regression trees (CART), k-nearest neighbors (k-NN), support vector machines (SVM), random forest (RF) and logistic regression (LR) to predict the childhood anemia status. A systematic evaluation of the algorithms was performed in terms of accuracy, sensitivity, specificity, and area under the curve (AUC). We found that the RF algorithm achieved the best classification accuracy of 68.53% with a sensitivity of 70.73%, specificity of 66.41% and AUC of 0.6857. On the other hand, the classical LR algorithm reached a classification accuracy of 62.75% with a sensitivity of 63.41%, specificity of 62.11% and AUC of 0.6276. Among all considered algorithms, the k-NN gave the least accuracy. We conclude that

* Corresponding author
Email: jkhan@isrt.ac.bd

ML methods can be considered in addition to the classical regression techniques when the prediction of anemia is the primary focus.

Keywords: Anemia prediction, children, machine learning, Bangladesh

1. Background

According to the World Health Organization (WHO), anemia is one of the most common and prevalent health concerns in the world [1]. Anemia, resulting from iron deficiency (ID), has been distinguished as one of the ten most important health risks contributing to the burden of diseases globally [1]. Moreover, anemia affects one-quarter of the world's population with a significant impact on preschool-aged children (6-59 months of age) and pregnant women [2]. As estimated, the global prevalence of anemia is 24.8%, and its prevalence in preschool-aged children and pregnant women are 47.4% and 41.8%, respectively.

Anemia is a condition which decreases the hemoglobin (Hb) concentration in blood, consequently impeding its capacity to transport oxygen. If it occurs among children, it can result in adverse effects on their cognitive developments and immunization abilities against diseases [3-5]. In Bangladesh, several surveys reported the severity of anemia among children under-five years of age. According to the Nutritional Surveillance Project (NSP), the prevalence of anemia among 6-59 months of children rose to 64% in 2004, from 47% in 2001 [6]. However, National Micronutrient Survey in 2011-12 showed an anemia prevalence among 6-59 months aged children to be only 33 % [7]. Khan et al. [8] reported that about 52 % of the children aged 6-59 months are anemic, based on a comprehensive analysis on childhood anemia using the nationally representative Bangladesh Demographic and Health Survey (BDHS) data. These studies emphasized the determination of the risk factors associated with such escalating childhood anemia prevalence in Bangladesh [8].

Although, the development of anemia among children can be genetic, or due to nutritional deficiencies, such as deficiency in iron, folate, vitamins A and B (12), and copper, ID is the most significant non-genetic determinant of the disease, among others [2]. However, demographic characteristics, socio-economic factors as well as maternal health factors are also associated with the incidence of anemia among children [8,9]. Proper diagnosis and intervention for anemia could reduce the risk of being anemic. A number of medical tools for anemia risk assessment could also be developed for diagnosis purpose. The focal point of many of the risk assessment techniques is the accurate prediction of the disease. Machine learning, an area that intersects statistical learning and artificial intelligence research, is the process of exploring large amounts of data to discover unknown patterns or relationships [10]. ML models can help to develop models for prediction purposes. These models have demonstrated high performance in solving classification problems as compared to the classical statistical models. Moreover, machine learning is becoming popular in the field of medical and health research, where classification technique (part of supervised machine learning) is the most important among all the ML algorithms [11, 12]. In medical settings, several machine learning techniques has been applied to predict disease [13-17]. Methods such as support vector machine, random forest, and artificial neural network have been used to classify status

of diseases like diabetes [13-15], acute appendicitis [16], multiple sclerosis [17], and many others, using the common risk factors. However, very few researches have considered the use of machine learning techniques to construct prediction models for childhood anemia [18, 19]. Machine learning could be effective to predict, more accurately, the risk of childhood anemia in addition to the ability of conventional clinical tools. To the best of our knowledge, this is the first study, in Bangladesh, for predicting childhood anemia using ML techniques based on cross-sectional health survey data.

This study aims at building several predictive models using the already established risk factors of anemia in children through ML approach based on the Bangladesh Demographic and Health Survey (BDHS) data. Specifically, five widely-used machine learning models such as- linear discriminant analysis (LDA), classification and regression trees (CART), k-Nearest Neighbors (k-NN), support vector machines (SVM), and random forest (RF) will be considered. These algorithms are a good mixture of simple linear (LDA), nonlinear (CART, k-NN) and complex nonlinear methods (SVM, RF). We further compare their predictive performances with the traditional logistic regression model-based approach. Accuracy, sensitivity, specificity, area under the curve (AUC) and Cohen's Kappa have been used to evaluate the predictive performance of the models.

2. Methods

2.1 Data source

Data for this study was extracted from the 2011 Bangladesh Demographic and Health Survey (BDHS). It was a nation-wide, cross-sectional, and probability sample survey on the Bangladeshi population. A two-stage cluster survey design was used, where a total of 600 clusters (urban: 207 clusters and rural: 393 clusters) were chosen in the first stage, and a systematic sample of 30 households (HHs) was selected on average per cluster in the second stage of sampling. This survey collected demographic, socio-economic, health and nutritional history, from participants in household interviews. Participants (women and children) were also invited for the blood test and anthropometric measurements that are performed by trained personnel [20]. Specifically, children aged 6-59 months from every third household in the BDHS sample were tested for blood hemoglobin level using HemoCue rapid testing methodology. A drop of capillary blood sample was taken from a child's fingertip or heel and was analyzed using the HemoCue photometer that displays the Hb concentration [20]. In this study, children from de jure households with no missing information on Hb or any of the other key predictors were considered. The final analysis considers 2013 children of 6-59 months of age from the 2011 BDHS survey. Details note about survey methodology, data collection and indicators can be found in the 2011 BDHS report [20].

2.2 Outcome variables

In this study, anemia was considered as the outcome variable which was categorized according to WHO's criteria, children of age 6-59 months with Hb $\leq$ 11.0 grams per deciliter (g/dl) are considered as "anemic", otherwise "non-anemic" [8].

2.3 Explanatory variables

We selected twenty-four variables associated with the risk for childhood anemia based on previous studies [8, 19, 21]. These are maternal age (years) ("<20", "20-29", "30-39" and "≥40"), maternal education ("no education", "primary", "secondary" and "higher"); paternal education ("no education", "primary", "secondary" and "higher"), maternal working status ("yes", "no"), child age (month) ("6-23", "24-59"), child stunting status ("yes", "no"), child breastfeeding status ("yes", "no"), maternal anemia ("yes", "no"), gender ("female", "male"), maternal underweight ("yes", "no"), household toilet facilities ("improved", "non-improved"), household water source ("improved", "non-improved"), child morbidity (fever) ("yes", "no"), child morbidity (diarrhea) ("yes", "no"), place of residence ("urban", "rural"), division ("Barisal", "Chittagong", "Dhaka", "Khulna", "Rajshahi", "Rangpur", "Sylhet"), number of living children, wealth index ("poorer", "poorest", "middle", "richer", "richest"), size of children at birth ("very small", "smaller than average", "average", "larger than average", "very large"), vitamin A within 6 months ("yes", "no"), iron with 7 days ("yes", "no"), any drug for parasites within 6 months ("yes", "no"), number of household members and number of under - children. Wealth index was calculated based on the principal component analysis (PCA) on the asset variables. Data selection procedure was the same as the previous study [8].

2.4 Ethical Approval

The BDHS 2011 survey was reviewed and approved by the institutional review board of the Bangladesh Medical Research Council (BMRC) and ICF Macro Institutional Review Board. Informed consent was obtained from each respondent in the survey before interviewing, and again, separately before taking weight, height and hemoglobin measurements [20]. The DHS Program also removed all personal information of the respondents in the database prior to making it publicly available.

2.5 Algorithms

The machine learning algorithms are model-free methods that provide efficient solutions to classification problems. Hence, the performances of these ML algorithms were compared with the statistical classifier Logistic Regression (LR). In the following, we briefly describe each of the methods considered in this study.

LR: Logistic regression (LR) is the most widely used statistical method in classification problems in the public health arena which provides the probability for predicting the classes of categorical outcome variable by using the given set of predictors. It measures the relationship between the response and the predictor variables [22].

LDA: Linear discriminant analysis (LDA) is a classical statistical approach for classification, which aims to find a linear combination of features that characterizes or separates two or more classes of objects or events using the Bayes' classifier [23, 24]. It assumes that the predictors are drawn from multivariate Gaussian (MVG) distribution given the category of observations, and to assign an observation to a class, LDA uses linear functions of the predictors for which the discriminant has the largest value [23, 24].

CART: Classification and regression trees (CART) is a non-parametric algorithm used for the purpose of classification of a set of data using the predictive structure of the problem under consideration. If the dependent variable is continuous, CART produces regression trees, and if the dependent variable is categorical, CART produces classification trees. For our study, CART has been used to classify a binary outcome variable, i.e. the status of anemia. CART divides the predictor space into non-overlapping regions, and assign each observation to a specific region depending on the proportion of observations belonging to that region for a specific class given the characteristics of the observations [25].

k-NN: k-nearest neighbors (k-NN) is a robust and versatile classifier which falls into the supervised learning family of algorithms. k-NN is non-parametric algorithm because it makes no explicit assumptions about the data distribution. This algorithm stores all available cases and classifies new cases based on a similarity measure. A case is classified by a majority vote of its neighbors, with the case being assigned to the class most common amongst its k nearest neighbors measured by a distance function [26, 27].

SVM: Support vector machines (SVM) is a kernel based supervised machine learning technique widely used in classification problems. The SVM algorithm constructs a hyperplane that separates the training observations perfectly according to their class labels by maximizing the margin among the classes. SVM assigns a test observation to a class depending on which side of the hyperplane it is located [28].

RF: Random forest (RF) is a classification technique, which is based on "growing" an ensemble of tree structured classifiers. To classify a new individual, features of this individual are used for classification using each classification tree in the forest. The grown trees are built randomly, and each tree gives a classification (or "voting") for a class label. The decision is based on the majority votes over all the trees in the forest [29, 30].

2.6 Model Evaluation

Model evaluation is important as it helps quantify a classifier's performance in serving as a general model. This means the relationships between input-output derived from the training

data set need to work equally well to test (or validation) data set as well.

To check the prediction accuracy of a ML classifier, the most novel way is to extensively test the classifier on a set of independent samples in a way to incorporate all possible sources of variability to be experienced. To estimate the predictive accuracy of the classifier from the training data, one way is to develop a k-fold cross-validation. This method provides "internal estimates" of predictive accuracy of the classification models. In *k*-fold cross-validation, dividing the data into k subsets of approximately equal size, the model is trained k times. In each of the k times of model training, one of the subsets is randomly set aside which in turn is used to evaluate the classifier's performance. By this means, all the possible cases of the whole dataset undergo training and testing, leading to a lower variance within the set estimator and less bias of the true rate estimator, which is the main advantage of using this method. However, in spite of being computationally both intensive and time consuming, this method ensures a more accurate prediction. In our study, we have relied on the 10-fold Cross Validation method, which has been used in several health care and medical related studies [30,32,33].

The performances of the algorithms are commonly evaluated using the criteria: confusion matrices and receiver operating characteristic (ROC). In a confusion matrix for a two-class case with classes "0" and "1", there are four possible outcomes of prediction, which are TR= true positives, TN=true negatives, FP=false positives, and FN=false negatives. Various performance measures such as accuracy, sensitivity, specificity are commonly computed using these four possible outcomes to evaluate the classifier, as defined by,

- Accuracy=TP+TN/(TP+TN+FP+FN)
- Sensitivity=TP/(TP+FN)
- Specificity=TN/(TN+FP).

The classification accuracy measured the proportion of cases correctly classified. Sensitivity measured the fraction of positive cases that were classified as positive, whereas the fraction of negative cases that were classified as negative are measured in specificity. The higher the values of these statistics, better the predictive performance of the algorithm.

The receiver operating characteristic (ROC) curves were calculated based on the predicted outcome and the true outcome. The AUC of the ROC was averaged for the test data sets to compare the discriminating powers of the algorithms [33].Theoretically, the AUC lies between 0 and 1, where a perfect classifier can take a maximum value of 1. However, the practical lower bound for random classification is 0.5, and classifiers with an AUC significantly greater than 0.5 have at least some ability to discriminate between cases and non-cases [30].

Cohen's kappa statistic is a very good measure to handle the multi-class and imbalanced class problems. It is a measure of the agreement between the predicted and the actual classifications in a dataset [34]. The value of Cohen's kappa is always less than or equal to 1.

Landis and Koch (1977) provide a way to characterize the values of this statistic: value < 0 is indicating “no agreement”, 0–0.20 as “slight”, 0.21–0.40 as “fair”, 0.41–0.60 as “moderate”, 0.61–0.80 as “substantial”, and 0.81–1 as “almost perfect agreement” [34].

In this paper, algorithms were developed using a sample of 80% of the individuals in each group (training dataset, n=1610) and validated in the remaining 20% (test dataset, n=403). All models were trained based on 10-fold cross validation. We designed the 10-fold cross validation not only to assess performance, but also to optimize prediction models using ML techniques. We used 10-fold cross validation on the training set, and the performance was measured on the testing set. We used open source statistical software R for the analysis purpose.

2.7 Results

Among the training set of selected samples (n=1610), 826 were anemic and 784 were non-anemic. Bivariate analyses of the outcome variable and characteristics of participants for training data and full data are shown in Table 1 and Table A1 (Appendix), respectively. Those variables (except Division) that showed statistically significant differences between the two groups of the outcome variables (anemia status), in full data, according to Pearson Chi-square test, were considered for modeling purposes. In total, fifteen factors were found to be significantly associated with anemia based on the BDHS (2011) data set. As shown in Table A1, factors based on parents’ demographic and health status, like mother’s age, underweight status, anemia status, education status of both parents, as well as factors like, child age, child stunting status, child breastfeeding status, child morbidity, along with other household characteristics, were used to develop the machine learning algorithms on the training dataset.

The five different algorithms were applied to classify the children in the test dataset as “anemic” and “non-anemic” based on the risk factors found to be significantly associated in the bivariate analysis. The predictive performances of the algorithms used are compared based on the performance parameters such as accuracy, sensitivity and specificity. Moreover, the discriminative accuracy of the algorithms is compared using the AUC and Cohen’s kappa statistic.

The prediction results with performance parameters for each of the machine learning algorithms are presented in Table 2, for both the training and the test datasets. Using the logistic regression, the accuracy in the test dataset is found out to be 62.75% with a sensitivity of 63.41% and specificity of 62.11%. The LDA showed an accuracy of 63.15% in prediction of the anemia status of the test observations, with a sensitivity of 64.23%, specificity of 62.11%. This accuracy is followed by the accuracy of 62.35% by the CART algorithm and the accuracy of 62.75% by the SVM with linear kernel. The CART algorithm attains a relatively higher sensitivity of 71.54% compensating with a relatively lower specificity of 53.52%; whereas SVM (linear) lies parallel with the LDA having shown a sensitivity of 64.23% and a

specificity of 61.33%. Among the 5 classifiers, the best results have been achieved with the random forest algorithm having shown an accuracy of 68.53%, a sensitivity of 70.73% and a specificity of 66.41%. However, the k-NN algorithm has shown the relatively poorest performance with accuracy, sensitivity and specificity of 61.95, 65.85and 58.20 respectively.

The AUC value as found using the LR for classification of anemia status is 62.8% (0.6276), with a value of 0.2551 of the Cohen's kappa statistic, suggesting a "fair" discriminative power. However, the other algorithms including LDA, CART, k-NN and SVM show discriminative abilities similar to the traditional LR algorithm as shown by the Cohen's Kappa statistic values of 0.2632, 0.2496, 0.2401 and 0.2553, respectively. However, the Cohen's kappa statistic shows the greater similarity between exact class and predicted class in the RF algorithm (0.3831) with maximum discriminative ability of 68.57% as also shown by the AUC values in Table 2. Thus, the RF algorithm performed the best in predicting the anemia status of the cases as indicated by all the performance indicators among all the other algorithms used in our study.

Identification of important features is also crucial in machine learning prediction. Feature importance rates shows how important each feature is for the decision a tree makes. The random forest (best algorithm for anemia prediction in our study) give a high importance to the "child morbidity regarding fever" feature, it also chooses "household toilet facilities" and "children age" to be the 2nd and 3rd most informative features overall. Any drug for parasites with 6 months, maternal underweight and place of residence also seem to be some of the influential factors, followed by child stunting status, maternal age, child breastfeeding status, maternal anemia, maternal education and paternal education. The random forest forces the algorithm to consider many possible explanations due to the randomness in model building and it captures a much broader picture of the data (Figure 1).

3. Discussion

Integrating machine learning techniques in predicting patient survival and disease status has become increasingly popular in healthcare and public health research [11, 15-17, 35] resulting in a positive impact on the improvement of health care planning. However, till date, very little research has been done on the use of machine learning algorithms to predict the disease status using cross-sectional demographic and health survey data [36, 37]. Moreover, no research has explored the potential of ML in predicting anemia status of children under-five years in Bangladesh. We found that childhood anemia can be predicted fairly accurately using a set of socio-demographic and health characteristics of the population routinely collected in health surveys.

In this study, an attempt was made to explore the 2011 BDHS dataset to provide an initial insight into the potential applicability of machine learning techniques in predicting anemia

status of under-five children based on socio-demographic and health characteristics. The findings clearly suggested that most of the attributes related to maternal and child health such as child age, maternal anemia, child malnutrition, maternal malnutrition, breastfeeding status of the children etc., have strong relation with anemia status of the children. Moreover, household characteristics such as toilet and water facilities and the socio-economic condition of the household as reflected through the wealth index, show a clear association with the incidence of anemia. Several models were built during experimentation that could predict the risk of childhood anemia based on these significantly related attributes.

We trained five different machine learning algorithms using the training dataset for predicting the incidence of anemia adjusting for multiple risk factors. The predictions were compared with those obtained from the logistic regression, which is the most widely used classifier in predicting disease status. The logistic model classifier showed an accuracy of 62.75% in predicting the anemia incidence in the test dataset. Among the predictive models built using machine learning techniques, random forest showed the best prediction accuracy result of 68.6% and also showed the highest discriminating power, as shown by a ROC area of 68.57%. However, the other predictive models showed similar predictive capability as the traditional logistic regression, with k-NN showing the least in performing predictions. Anemia is a life-threatening disease which affects the hemoglobin production and is especially (potentially) fatal in children. Hence, the use of machine learning in this study showed that the probability of childhood anemia can be minimized substantially by intervening in certain socio- demographic and health-related factors. These algorithms can also be used not only as a guide to monitor future undertakings to control anemia but also for formulating child nutrition programs and health policies. Moreover, this can help in building a knowledge-based system for predicting childhood anemia incidence for children residing in Bangladesh, but it cannot replace the physician’s intuition and interpretive skills.

This study has some limitations. As the predictive models used in this study were established using cross-sectional demographic and health survey data, additional information on potentially other relevant clinical and dietary variables were unavailable. Incorporating those variables would likely to have improved the predictive accuracy. Since some attributes (i.e. diarrhea and fever status of the children within last two weeks from survey interview date) were self-reported, there were chances of recall bias. Also, due to lack of data availability post year 2011, any change in the scenario of anemia over time could not be reported in this study. Finally, out of the numerous ML algorithms that could have been applied in this context, the algorithms were chosen on subjective judgment. However, this study provides evidence that ML algorithms can be used to predict anemia based on the common risk factors, which can assist in the development of interventions in preventing anemia among children.

4. Conclusion

We compared several machine learning prediction models for predicting whether a patient has anemia given the risk factors. Among the models considered, the random forest performed the best with the highest classification accuracy to predict anemia in the Bangladeshi population. This study highlights not only the utility of ML algorithms but also the importance of using common socio-demographic and health related characteristics to predict the disease status. Moreover, our findings would be useful for identifying children who are at risk of anemia in the future giving the policymakers and healthcare providers a tool to implement necessary interventions and improve care practices. Thus, a model built on the common risk factors would assist in the prevention and control of childhood anemia.

Acknowledgments

The authors acknowledge the contributions of NIPORT, MEASURE DHS and ICF International teams for their efforts to collect data and allowing access to the data set.

Competing Interests

The authors declared that they have no competing interests.

Authors' Contributions

JRK: idea + data analysis + manuscript preparation + revision
SC: idea + preliminary analysis + manuscript revision
HI: manuscript preparation + revision
ER: critical review + revision
All authors are approved the revised version of the manuscript.

References

[1] World Health Organization. The World Health Report 2002: Reducing risks, promoting healthy life. World Health Organization; 2002.

[2] McLean E, Cogswell M, Egli I, Wojdyla D, De Benoist B. Worldwide prevalence of anaemia, WHO vitamin and mineral nutrition information system, 1993–2005. Public health nutrition. 2009 Apr;12(4):444-54.

[3] World Health Organization. Iron deficiency anemia. assessment, prevention, and control. A guide for programme managers. Geneva, World Health Organization, 2001.

[4] Brabin BJ, Premji Z, and Verhoe F. An analysis of anemia and child mortality. The Journal of nutrition. 2001 Feb 1;131(2):636S-48S.

[5] McCann JC, Ames BN. An overview of evidence for a causal relation between iron deficiency during development and deficits in cognitive or behavioral function–. The American journal of clinical nutrition. 2007 Apr 1;85(4):931-45.

[6] Rashid M, Flora MS, Moni MA, Akhter A, Mahmud Z. Reviewing Anemia and iron folic acid supplementation program in Bangladesh-a special article. Bangladesh Medical Journal. 2010;39(3).

[7] International Centre for Diarrhoeal Diseases Research, Bangladesh (icddr,b), Global Alliance for Improved Nutrition (GAIN), The United Nations Children's Fund (UNICEF). The National Micronutrients Status Survey 2011–12. Dhaka, Bangladesh: International Centre for Diarrhoeal Diseases Research, Bangladesh; 2013.

[8] Khan JR, Awan N, Misu F. Determinants of anemia among 6–59 months aged children in Bangladesh: evidence from nationally representative data. BMC pediatrics. 2016 Dec;16(1):3.

[9] Ayoya MA, Ngnie-Teta I, Séraphin MN, Mamadoultaibou A, Boldon E, Saint-Fleur JE, Koo L, Bernard S. Prevalence and risk factors of anemia among children 6–59 months old in Haiti. Anemia. 2013;502968.

[10] Witten IH, Frank E, Hall MA, Pal CJ. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann; 2016 Oct 1.

[11] 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. 2017 Jul 24;12(7): e0179805.

[12] Kotsiantis SB, Zaharakis I, Pintelas P. Supervised machine learning: A review of classification techniques. Emerging artificial intelligence applications in computer engineering. 2007; 160:3-24.

[13] Meng XH, Huang YX, Rao DP, Zhang Q, Liu Q. Comparison of three data mining models for predicting diabetes or prediabetes by risk factors. The Kaohsiung journal of medical sciences. 2013 Feb 1;29(2):93-9.

[14] Choi SB, Kim WJ, Yoo TK, Park JS, Chung JW, Lee YH, Kang ES, Kim DW. Screening for prediabetes using machine learning models. Computational and Mathematical Methods in Medicine, 2014, 2014: 618976.

[15] Yu W, Liu T, Valdez R, Gwinn M, Khoury MJ. Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes. BMC medical informatics and decision making. 2010;10(1):16.

[16] Hsieh CH, Lu RH, Lee NH, Chiu WT, Hsu MH, Li YC. Novel solutions for an old disease: diagnosis of acute appendicitis with random forest, support vector machines, and artificial neural networks. Surgery. 2011;149(1):87-93.

[17] Zhao Y, Healy BC, Rotstein D, Guttmann CR, Bakshi R, Weiner HL, Brodley CE, Chitnis T. Exploration of machine learning techniques in predicting multiple sclerosis disease course. PloS one. 2017;12(4):e0174866.

[18] Sanap SA, Nagori M, Kshirsagar V. Classification of anemia using data mining techniques. In International Conference on Swarm, Evolutionary, and Memetic Computing 2011 (pp. 113-121). Springer, Berlin, Heidelberg.

[19] Abdullah M, Al-Asmari S. Anemia types prediction based on data mining classification algorithms. Communication, Management and Information Technology –Sampaio de Alencar (Ed.). 2017.

[20] National Institute of Population Research and Training (NIPORT), Mitra and Associates, ICF International. Bangladesh Demographic and Health Survey 2011. Dhaka: Bangladesh and Calverton, Maryland, USA: NIPORT, Mitra and Associates, ICF International; 2013.

[21] Singh BP, Maheshwari S, Gupta PK. Anemia in Married Females of Uttar Pradesh and Its relation to Body Mass Index: Application of Poisson Regression. Journal of Data Science. 2017;15(2):267-74.

[22] Le Cessie S, Van Houwelingen JC. Ridge estimators in logistic regression. Applied statistics. 1992; p. 191–201.

[23] McLachlan GJ. Discriminant Analysis and Statistical Pattern Recognition. New York: Wiley Interscience; 2004.

[24] Izenman AJ. Linear discriminant analysis, Modern Multivariate Statistical Techniques, Springer, 2008; pp.237-280.

[25] Breiman L, Friedman JH, Olshen RA, Stone CJ: Classification and regression trees. 1984, Monterey, Calif., USA: Wadsworth, Inc.

[26] Cover T, Hart P. Nearest neighbor pattern classification. IEEE transactions on information theory. 1967 Jan;13(1):21-7.

[27] Gaber T, Hassanien AE, El-Bendary N, Dey N, editors. The 1st International Conference on Advanced Intelligent System and Informatics (AISI2015), November 28-30, 2015, Beni Suef, Egypt. Springer; 2015 Nov 9.

[28] Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. Machine Learning. 2002;46:389–422.

[29] Breiman L. Random forest. Machine Learning. 2001;45:5–32.

[30] Liaw A, Wiener M. Classification and regression by randomForest. R news. 2002;2(3):18–22.

[31] Kang J, Cho J, Zhao H. Practical issues in building risk-predicting models for complex diseases. Journal of biopharmaceutical statistics. 2010 Mar 19;20(2):415-40.

[32] Liu B, Fang L, Liu F, Wang X, Chen J, Chou KC. Identification of real microRNA precursors with a pseudo structure status composition approach. PloS one. 2015; 10(3):e0121501. https://doi.org/10. 1371/journal.pone.0121501 PMID: 25821974

[33] Liu B, Fang L, Liu F, Wang X, Chou KC. iMiRNA-PseDPC: microRNA precursor identification with a pseudo distance-pair composition approach. Journal of Biomolecular Structure and Dynamics. 2016; 34(1):223-235.

[34] Yu W, Liu T, Valdez R, Gwinn M, Khoury MJ. Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes. BMC medical informatics and decision making. 2010 Dec;10(1):16.

[35] Meng XH, Huang YX, Rao DP, Zhang Q, Liu Q. Comparison of three data mining models for predicting diabetes or prediabetes by risk factors. The Kaohsiung journal of medical sciences. 2013;29(2):93–99. pmid:23347811

[36] Landis JR, Koch GG. The measurement of observer agreement for categorical data. biometrics. 1977:159-74.

[37] Khare S, Kavyashree S, Gupta D, Jyotishi A. Investigation of Nutritional Status of Children based on Machine Learning Techniques using Indian Demographic and Health Survey Data. Procedia Computer Science. 2017;115:338-49.

[38] Sahle G. Ethiopic maternal care data mining: discovering the factors that affect postnatal care visit in Ethiopia. Health information science and systems. 2016;4(1):4.

Table 1: Bivariate analyses of selected variables and anemia status of the children (based on training data, sample size=1610).

Variables	Non-anemic N=784	Anemic N=826	Odds Ratio (OR)	p-value
Maternal age (year)				<0.001
<20	8.16%	17.10%	Ref.	
>=40	2.55%	2.54 %	0.48	
20-29	64.70%	57.60%	0.43	
30-39	24.60%	22.80%	0.44	
Maternal education				≤0.05
Higher	8.55%	4.72%	Ref.	
No education	19.30%	20.10%	1.88	
Primary	30.90%	32.70%	1.91	
Secondary	41.30%	42.50%	1.86	
Paternal education				0.163
Higher	14.20%	10.70%	Ref.	
No education	28.70%	28.70%	1.33	
Primary	29.50%	32.30%	1.46	
Secondary	27.70%	28.30%	1.36	
Maternal working status				0.999
No	90.10%	90.00%	Ref.	
Yes	9.95%	10.00%	1.01	
Child age (month)				<0.001
24-59	79.70%	54.80%	Ref	
6-23	20.30%	45.20%	3.23	
Child stunting status				<0.001
No	62.90%	53.40%	Ref	
Yes	37.10%	45.20%	1.48	
Child breastfeeding status				<0.001
No	44.40%	26.40%	Ref.	
Yes	55.60%	73.60%	2.22	
Maternal anemia				<0.001
No	63.30%	48.20%	Ref.	
Yes	33.70%	51.80%	2.12	
Gender				≤ 0.05
Female	51.40%	45.60%	Ref	
Male	48.60%	54.40%	1.26	
Maternal underweight				≤ 0.01

No	73.70%	67.70%	Ref.	
Yes	26.30%	33.30%	1.34	
Household toilet facilities				0.066
Improved	56.60%	51.90%	Ref.	
Non-improved	43.40%	48.10%	1.21	
Household water source				0.366
Improved	98.30%	97.60%	Ref.	
Non-improved	1.66%	2.42%	1.47	
Child morbidity (fever)				≤ 0.05
No	62.10%	56.30%	Ref.	
Yes	37.90%	43.70%	1.27	
Child morbidity (diarrhea)				0.478
No	94.80%	93.80%	Ref.	
Yes	5.23%	6.17%	1.19	
Place of residence				0.341
Rural	69.30%	71.50%	Ref.	
Urban	30.70%	28.50%	0.90	
Division				0.073
Barisal	8.93%	11.86%	Ref.	
Chittagong	18.60%	17.80%	0.72	
Dhaka	16.50%	16.10%	0.74	
Khulna	10.70%	11.60%	0.82	
Rajshahi	13.80%	11.40%	0.62	
Rangpur	12.40%	15.60%	0.95	
Sylhet	19.10%	15.60%	0.62	
Number of living children	2.30 (1.06)	2.29 (1.11)	0.99	0.828
Wealth index				<0.01
Middle	17.50%	16.70%	Ref.	
Poorer	18.10%	21.70%	1.25	
Poorest	20.40%	26.90%	1.38	
Richer	22.20%	16.80%	0.79	
Richest	21.80%	17.90%	0.86	
Size of children at birth				0.683
Average	66.80%	66.90%	Ref.	
Larger than average	12.00%	14.00%	1.17	
Smaller than average	13.10%	11.90%	0.90	
Very large	1.79%	1.69%	0.95	
Very small	6.25%	5.45%	0.87	
Vitamin A within 6 months				0.037

No	37.50%	42.70%	Ref.	
Yes	62.50%	57.30%	0.80	
Iron with 7 days				0.512
No	97.30%	97.90%	Ref.	
Yes	2.68%	2.06%	0.84	
Any drug for parasites within 6 months				<0.001
No	42.60%	56.50%	Ref.	
Yes	57.40%	43.50%	0.57	
Number of household members	0.82 (0.66)	0.83 (0.65)	0.73	0.727
Number of under 5 children	0.39 (0.49)	0.40 (0.49)	0.63	0.632

Ref.: Reference category

Table 2: The performance indicators of all the six machine learning algorithms.

	Algorithms					
	LDA	**CART**	**k-NN**	**SVM (linear)**	**RF**	**LR**
Training set						
Accuracy (%, 95% CI)	63.84	62.14	67.73	66.73	96.16	63.99
	(61.69-65.95)	(59.98-64.27)	(65.63-69.77)	(64.62-68.79)	(95.22-9696)	(61.84-66.09)
Kappa	0.2749	0.2331	0.3524	0.3321	0.923	0.2777
Sensitivity (%)	65.50	73.18	69.95	69.38	96.59	65.88
Specificity (%)	62.00	49.89	65.26	63.79	95.68	61.89
AUC (95% CI)	0.6375	0.6154	0.6761	0.6659	0.9614	0.6389
	(0.6164-0.6586)	(0.5946-0.6361)	(0.6556-0.6966)	(0.6452-0.6865)	(0.9529-0.9698)	(0.6178-0.6599)
Test set						
Accuracy (%, 95% CI)	63.15	62.35	61.95	62.75	68.53**	62.75
	(58.76-67.38)	(57.95-66.60)	(57.55-66.22)	(58.35-66.99)	(64.26-72.57)	(58.35-66.99)
Kappa	0.2632	0.2496	0.2401	0.2553	0.3710**	0.2551
Sensitivity (%)	64.23	71.54**	65.85	64.23	70.73	63.41
Specificity (%)	62.11	53.52	58.20	61.33	66.41**	62.11
AUC (95% CI)	0.6317	0.6253	0.6203	0.6278	0.6857**	0.6276
	(0.5894-0.6740)	(0.5836-0.6670)	(0.5779-0.6627)	(0.5854-0.6701)	(0.645-0.7263)	(0.5852-0.6700)

**indicates the best in performance

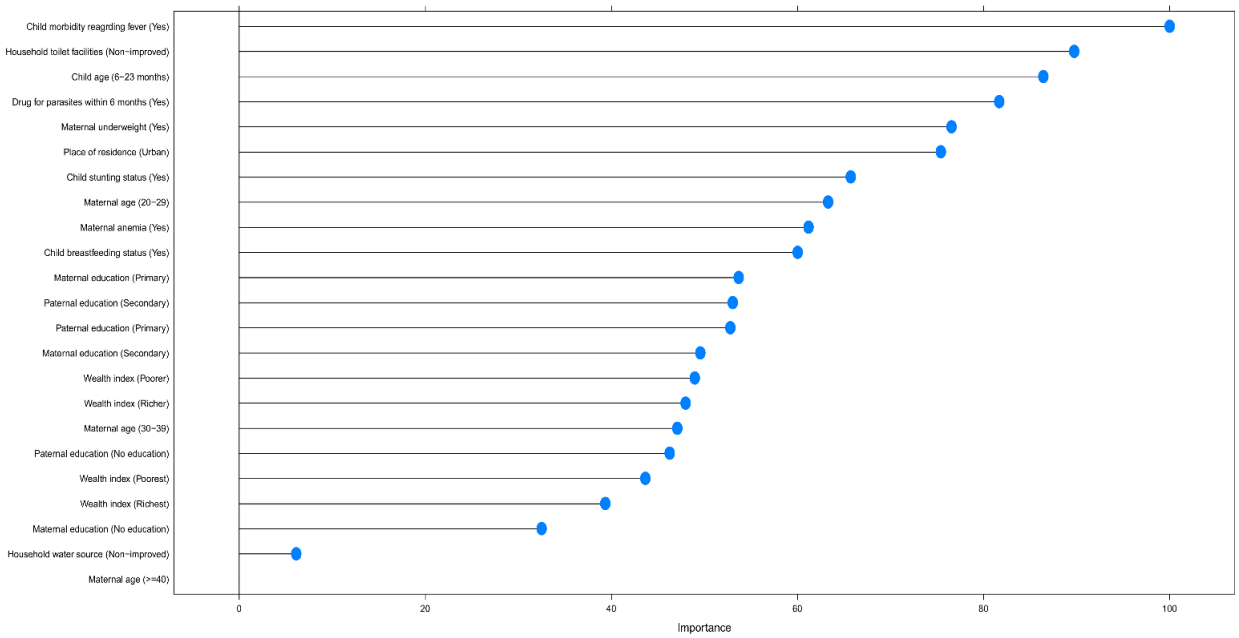

Figure 1: Variable importance from the best model (random forest).

Appendix

Table A1: Bivariate analyses of selected variables and anemia status of the children (based on full data, sample size=2013).

Variables	Non-anemic N=975	Anemic N=1038	p-value
Maternal age (year)			<0.001
<20	8.21%	15.20%	
>=40	3.08%	2.79%	
20-29	63.70%	59.50%	
30-39	25.00%	22.40%	
Maternal education			<0.01
Higher	8.82%	5.11%	
No education	19.40%	20.70%	
Primary	30.90%	34.00%	
Secondary	40.90%	40.20%	
Paternal education			<0.01
Higher	15.70%	10.20%	
No education	28.80%	28.90%	
Primary	28.00%	33.50%	
Secondary	27.50%	27.40%	
Maternal working status			0.911
No	89.80%	89.60%	
Yes	10.20%	10.40%	
Child age (month)			<0.001
24-59	79.90%	56.00%	
6-23	20.10%	44.00%	
Child stunting status			<0.001
No	63.00%	53.90%	
Yes	37.00%	46.10%	
Child breastfeeding status			<0.001
No	43.90%	27.40%	
Yes	56.10%	72.60%	
Maternal anemia			<0.001
No	64.80%	49.10%	
Yes	35.20%	50.90%	
Gender			0.186
Female	50.30%	47.20%	
Male	49.70%	52.80%	

Maternal underweight			<0.001
No	74.50%	66.90%	
Yes	25.50%	33.10%	
Household toilet facilities			≤ 0.01
Improved	56.30%	50.00%	
non-improved	43.70%	50.00%	
Household water source			≤0.05
Improved	98.60%	96.80%	
non-improved	1.44%	3.18%	
Child morbidity (fever)			≤0.01
No	62.80%	56.60%	
Yes	37.20%	43.40%	
Child morbidity (diarrhea)			0.397
No	95.10%	94.10%	
Yes	4.92%	5.88%	
Place of residence			≤0.05
Rural	67.70%	71.80%	
Urban	32.30%	28.20%	
Division			≤0.05
Barisal	8.92%	11.80%	
Chittagong	18.80%	17.90%	
Dhaka	18.10%	15.40%	
Khulna	10.70%	11.80%	
Rajshahi	13.10%	11.60%	
Rangpur	11.90%	15.40%	
Sylhet	18.60%	16.10%	
Number of living children	2.31 (1.07)	2.31 (1.10)	0.895
Wealth index			<0.001
Middle	18.10%	17.40%	
Poorer	18.10%	22.00%	
Poorest	19.90%	26.80%	
Richer	20.60%	16.60%	
Richest	23.40%	17.20%	
Size of children at birth			0.758
Average	68.30%	66.90%	
Larger than average	12.00%	13.80%	
Smaller than average	12.50%	12.50%	
Very large	1.85%	1.45%	
Very small	5.33%	5.39%	

Vitamin A within 6 months			0.086
No	38.10%	41.90%	
Yes	61.90%	58.10%	
Iron with 7 days			0.821
No	97.30%	97.60%	
Yes	2.67%	2.41%	
Any drug for parasites within 6 months			<0.001
No	43.40%	55.70%	
Yes	56.60%	44.30%	
Number of household members	0.83 (0.66)	0.81 (0.65)	0.457
Number of under 5 children	0.39 (0.49)	0.39 (0.49)	0.936