MedCrave
Step into the World of Research

Biometrics & Biostatistics International Journal

Research Article

Open Access

CrossMark

Classification tree and random forest model to predict under-five malnutrition in Bangladesh

Abstract

Malnutrition is one of the leading causes of morbidity and mortality in children under the age of five in most developing countries like Bangladesh. The main objective of this study is to design a model that predicts the nutritional status of under-five children using tree based model and classical approach. This study used secondary data from Bangladesh Demographic and Health Survey 2014 for 7,886 children. Decision tree based model like classification tree, random forest and classical model like multiple binary logistic regression model are fitted to assess the association of malnutrition of children with potential socioeconomic and demographic factors. In this study, predictive model is developed using random forest having an accuracy of 70.1% & 72.4% and area under receiver operating characteristic curve of 69.8% and 70% for stunting and underweight respectively. The prevalence of stunting and underweight are found 36.5% and 33% respectively among under-five children and higher in rural setting than in urban areas. Similarly, wealth index, exposure of mother to the mass media, age of child, size of child at birth, and parents' education are significantly associated with stunting and underweight of children.

Keywords: predictive modeling, data mining, random forest, classification tree

Volume 10 Issue 3 - 2021

Sabbir Ahmed Hemo, Md. Israt Rayhan
Researcher, Institute of Statistical Research and Training (ISRT), University of Dhaka, Bangladesh

Correspondence: Md. Israt Rayhan, Professor, Institute of Statistical Research and Training (ISRT), University of Dhaka, Dhaka-1000, Bangladesh, Tel +880-1924752885, Email israt@isrt.ac.bd

Received: August 16, 2021 | **Published:** September 14, 2021

Introduction

This study examines the effect of different socio-economic factors upon children's nutritional status. Under-five malnutrition is widespread in both Sub-Saharan Africa and South Asia.[1] Toma et al.[2] found from their study in Bangladesh that more educated and wealthier mothers have better nourished children at birth. Sarma et al.[3] used multiple binary logistic regression analysis and found significant association between the stunting of under-five children in Bangladesh and the covariates, such as: wealth index, exposure of mother to the mass media, age of child, size of child at birth, and parents' education. South Asia has the highest 49.3 percent of under-five underweight,[4] despite a better economic development than Sub-Saharan Africa. This enigma is due to the low status of women in South Asia (Ibid). Association between child malnutrition and socio-economic covariates were analyzed by Khare et al.[5] from the Indian DHS data, their study revealed that machine learning approach identifies some important features over the classical models. Talukder and Ahammed[6] claimed that the random forest algorithm was moderately superior to any other machine learning algorithms to predict malnutrition status among under-5 children in Bangladesh.

Batterham et al.[7] found that data mining methods can provide a more accuracy for weight loss trial study compare to the conventional assumptions. Amongst the data mining methods, a decision tree provided them the most accuracy. Bath[8] stated that data mining method can act as a knowledge discovery technique in analyzing health medica1 data, also added that data mining method can identify high risk covariates with common features whereas Cox's regression can provide an estimate of the strength of these risk factors. Rayhan and Khan[9] analyzed BDHS 1999-2000 data with Cox's logistic regression model and concluded that major contributing factors for under five malnutrition were birth interval, size at birth, mother`s body mass index at birth and parent`s education. Alom et al.[10] utilized BDHS 2007 data and through a multilevel regression analysis they found that under-five malnutrition in Bangladesh was significantly associated with child's age, mother's education, father's education, father's occupation, family wealth index, currently breast-feeding, place of delivery and division. Rahman et al.[11] analyzed BDHS 2011 data and stated that children's low birth weight was significantly associated with malnutrition by controlling other confounders. Sultana et al.[12] used a multilevel generalized linear regression model for BDHS 2014 data and found that stunting was significantly associated with mother's age at birth, previous birth interval, mother's BMI, parent's education, wealth index. Therefore this study is thriving towards the data mining process that would cluster the risk factors of under-five malnutrition. Decision tree and random forest model are acquired as unconventional methods to analyze the BDHS 2014 data with the conventional covariates suggested by the previous literature.

Methodology

This study utilizes the data from Bangladesh Demographic and Health Survey (BDHS), 2014. The dependent variable is considered as malnutrition status of the children under- five years of age. Malnutrition is measured by the two different anthropometric indices, named as: stunting (less height for age) and underweight (less weight for age). The indices are expressed as the number of standard deviations (SD) above or below the median height of healthy children in the same age group of the reference population.[13] Household's socio-economic status, exposure of mother to the mass media, age of child, size of child at birth, parents' education, birth order, birth interval, mother's BMI, area of residence, administrative region, total number of children are considered to be associated with the child's malnutrition.[9-12]).

Anthropometric measures

According to WHO[13] height-for-age is considered to measure a child's stunting or acute malnutrition, caused by inadequate nutrition over a long period of time. A child who is two standard deviations below the median (-2 SD) group of the WHO reference population, in terms of height-for-age, is considered as stunted. If it is below three standard deviations (-3 SD) from the reference median, considered as severe stunting. A child whose weight-for-age is below two standard deviations (-2 SD) from the median of the reference group is classified as underweight. If it is below three standard deviations (-3 SD) from the

reference median, considered as severe underweight. Underweight is an overall indicator of a child's nutritional health.

Hosmer and Lemeshow[14] has described logistic regression for the binary dependent variable, this study considered a child's nutritional status as stunted or not, and underweight or not. A brief discussion of logistic model is as follows:

$$E(Y \mid x) = \pi(x)$$

be the conditional mean π of dependent variable Y, given explanatory variable x.

Then the logistic regression model $\pi(x)$ as,

$$\pi(x) = \frac{e^{(\beta_0 + \sum_{i=1}^{p} \beta_j x_{ij})}}{1 + e^{(\beta_0 + \sum_{i=1}^{p} \beta_j x_{ij})}} \quad (1)$$

where β_0 and β_j are the model parameters. The logistic function $\pi(x)$ ranges between 0 to 1 which is the major reason of the popularity of this model. This study has used this model as a classical approach.

Classification tree

Tree-based methods are simple and significant tools for grouping the covariates.[15] The prediction for a test observation based on such model is the modal class of y in the region to which it belongs. The shape of the splitting rules of the predictor space looks like a tree, so this type of approach is known as decision tree method.

There are two steps for building a classification tree:

1. First split the predictor space, the set of possible values for X_1, X_2,.... , X_p into J distinct and non-overlapping regions R_1, R_2, , R_J.

$$E = 1 - \max_k \left(\hat{p}_{mk} \right) \quad (2)$$

Here, $(\hat{p}_{mk})$ represents the proportion of training observations in the m^{th} region that are from the k^{th} class in terms of intra-correlation.

2. For the observations that fall into the region R_j, the same prediction is made, which is simply the mode of the values for the training observationsin $\sqrt{p}$. A classification tree is very close to a regression tree, it deals with qualitative response instead of quantitative one.

Random forest

Random forest chooses the variables to split in a group using an algorithm that minimizes error. Random forest uses the algorithm in the way that within sub-trees have more correlation but between sub-trees have less correlation.

A common choice of m is $\sqrt{p}$ where p is the total number of predictors in the data set. For each bootstrap sampling from the training data, a few samples left behind that were not included. The performance of each model on its drop out samples from an average can provide an estimated accuracy of the bagged models. This estimated performance is often called the out of bag (OOB) estimate of performance. The OOB performance measures are considered as the cross validation estimates. As the Bagged decision trees are constructed, this study can calculate how much the error function drops for a variable at each split point.

Cross Validation and parameter tuning

A procedure for tuning model parameters is cross validation. There are many approaches to perform this task. In *validation set approach*, the set of data at hand is randomly partitioned into two parts: the training data set and the test data set. The model to be trained is built on the training set with different combinations of values for the associated parameters, and applied on the test set. In *K-fold cross validation approach*, the set of data at hand are randomly divided into K partitions. For each partition or fold, the model is trained on the rest of the partitions with different parameter settings and applied on the unused partition. Checking out of the bag error is the ultimate extension if of the above two approaches, where the model is trained 'n' times (n being the total number of observations in the data set, each time one of the observations is left out while training the model and the model makes prediction on it.) Accuracy rate is the proportion of correctly classified observations in the test set. Sensitivity is the proportion of true positives that have been classified as positives. Specificity is the proportion of true negative that have been classified as negatives. Receiving operating curve (ROC) curve is a graph of 'sensitivity' over '1-specificity'. As the threshold value of the predicted class probability is moved from 0.5 to the both extremes, the sensitivity and the specificity of the classifier changes. The area under the ROC curve (AUC) is a measure of the discriminating ability of the classifier.

Analysis and results

Among the children in rural area, 38.4% were stunted and 35.3% were underweight which is higher than the urban area. Prevalence of stunting and underweight are highest in Sylhet division which are 49.9% and 40.5% respectively and lowest in Khulna division which are 28.1% and 25.9% respectively. Prevalence of stunting and underweight are highest among the poorest which are 49.5% and 46.4% respectively. Mother's education level plays a significant role in stunting and underweight as the prevalence increases as mother's education level decreases. Percentage of malnutrition increases as the age of child increases. And underweight mothers are more likely to have a malnourished children. Mothers who watch TV or read the newspaper once a week have a lower percentageof malnourished children than those mothers never watches TV or read the newspaper. There has been significant association of child malnutrition and Sex of child, Size at birth, Mother's age, Order of birth, Type of place of residence, Division, Wealth index, Mother's highest level of education, Exposure to media, Had fever recently, Had diarrhea recently, Currently breastfeeding, Age of child and Mother's BMI as the p-value from chi-square test statistic is very small.

Table 1 Prevalence of malnutrition among different background characteristics (inweighted percentage)

		Proportion of stunting		Proportion of underweight	
Covariates		**Stunted**	**P-value**	**Underweight**	**P-value**
Sex of child					
	Female	35.8	0.06	33.5	0.52
	Male	37.2		32.6	

Citation: Hemo SA, Rayhan MI. Classification tree and random forest model to predict under-five malnutrition in Bangladesh. *Biom Biostat Int J.* 2021;10(3):116–123. DOI: 10.15406/bbij.2021.10.00337

Table Continued...

		Proportion of stunting		Proportion of underweight	
Covariates		**Stunted**	**P-value**	**Underweight**	**P-value**
Size at birth					
	Very small	45.3	0	51.8	0
	Smaller than average	43.6		43.4	
	Average	31.6		27.9	
	Larger than average	22.1		16.9	
	Very large	26.1		19.2	
Mother's age					
	15-19	34.3	0.01	32	0.01
	20-24	35.8		32.9	
	25-29	37		31.9	
	30-34	37.1		35.2	
	35-39	38.9		34.6	
	40-44	42.7		30.2	
	45-49	69.4		60.4	
Order of birth					
	1	32.1	0	29	0
	2	35.9		33.3	
	3	37.3		34.2	
	4+	48.2		41.4	
Type of place of residence					
	Urban	30.9	0	26.3	0
	Rural	38.4		35.3	
Division					
	Barisal	40.2	0	36.7	0
	Chittagong	38.2		36.1	
	Dhaka	34.5		28.8	
	Khulna	28.1		25.9	
	Rajshahi	31.2		32.4	
	Rangpur	37		38	
	Sylhet	49.9		40.5	
Wealth index					
	Poorest	49.5	0	46.4	0
	Poorer	43.2		39.1	
	Middle	36.6		32.3	
	Richer	31.5		27.5	
	Richest	19.6		17.4	
Mother's education					
	No education	47.7	0	42.5	0
	Primary	44.2		39.5	
	Secondary	31.4		28.8	
	Higher	19.8		18.2	

Citation: Hemo SA, Rayhan MI. Classification tree and random forest model to predict under-five malnutrition in Bangladesh. *Biom Biostat Int J.* 2021;10(3):116–123. DOI: 10.15406/bbij.2021.10.00337

Table Continued...

		Proportion of stunting		Proportion of underweight	
Covariates		**Stunted**	**P-value**	**Underweight**	**P-value**
Exposure to media					
	Not at all	42.7	0	40.2	0
	Less than once a week	43.1		40.3	
	At least once a week	30		25.5	
Currently breastfeeding					
	Yes	35	0	32.6	0.02
	No	38.9		33.7	
Age of child					
	0-11	18.4	0	21.6	0
	23-Dec	41		32.8	
	24-35	41.4		37.5	
	36-47	44.4		35.6	
	48-59	39.5		39.6	
Mother's BMI					
	Underweight	43.8	0	46.2	0
	Normal	38.3		33.7	
	Overweight	31.9		24.6	
	Obese	18.6		15.7	

In this paper, the approaches to finding best subset of models are kept to be as simplistic and intuitive as possible. This also opens up scope for further investigation and refinement of this study. A random forest model is built using the data and its parameters are tuned so that the out-of-the-bag error rate is minimized. Setting the number of trees to 5000, the optimum number of predictors to be considered at each split is found to be 4 which is the square root of number of total predictor. With parameters $n_{tree} = 5000$ and $m_{try} = 4$, a random forest model is trained and cross-validated (10-fold) on the dataset. A variable importance plot is constructed, where importance of a variable is calcul ated as the average amount of decrease in Gini index resulting from splitting a node with respect to that variable. The largest mean decrease in Gini index for predicting stunting and underweight is accounted for variables Division, Wealth index, Age of child, Mother's age, Mother's BMI, Order of birth, Size at birth, Mother's highest level of education, Mother's age at first marriage, Exposure to media, Sex of child, etc. These variables are used to train classification tree.

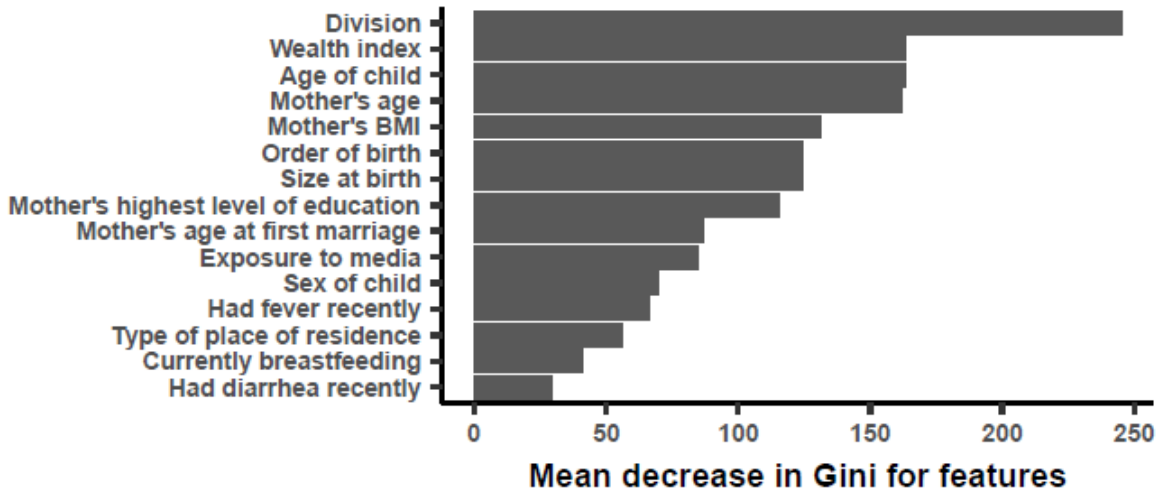

Figure 1 Variable importance plot by mean decrease in gini using random forest for stunting.

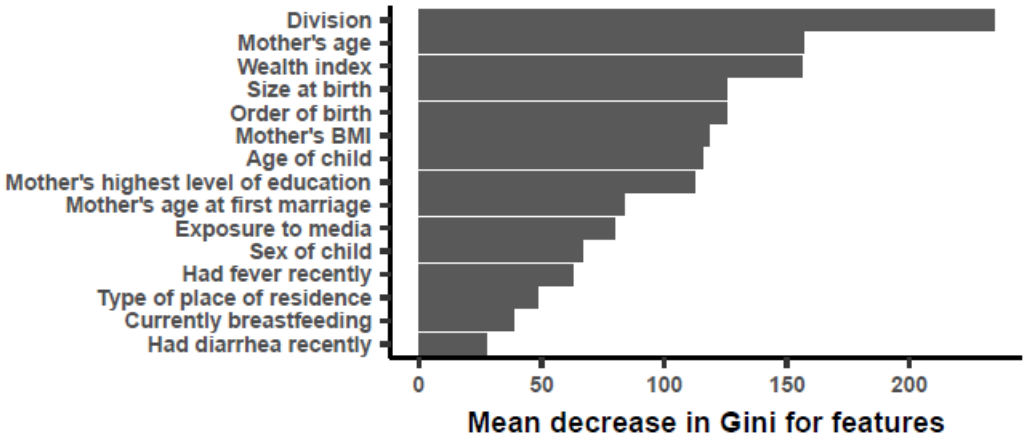

Figure 2 Variable importance plot by mean decrease in gini using random forest for predicting underweight.

Prediction using classification tree

A classification tree is first grown using twelve variables from the random forest model according to mean decrease in Gini index. Then the tree is pruned using the root mean square error and cross validation error. Figure below suggests the best number of splits and choice of the complexity parameter for pruning for which root mean square error and cross validation error are minimum.Then the final tree is fitted using the complexity parameter found. In this study,classification tree has predicted 45% of children as normal children whose father have secondary or higher education with probability 0.74. The second split happens whose father have no education or primary education and only 11% children below 1 year of age are predicted as normal. Again, those children aged 2 or more years and from Dhaka, Khulna, Rajshahi, Rangpur then low birth weight are predicted as stunted with a probability of 0.62. If the child from Barisal, Chittagong or Sylhet who are at poorest and poorer socio-economic group whose father have no education or primary education, 12% are predicted as stunted with probability 0.65. For predicting underweight, 40% children from the richer and richest group are predicted as normal with probability 0.78.Mother's age and mother's BMI play an important role in predicting underweight children. Among the poor family, whose children is 2-5 years old suffered from fever recently and whose mother is underweight, the children is predicted as underweight with a probability 0.75.

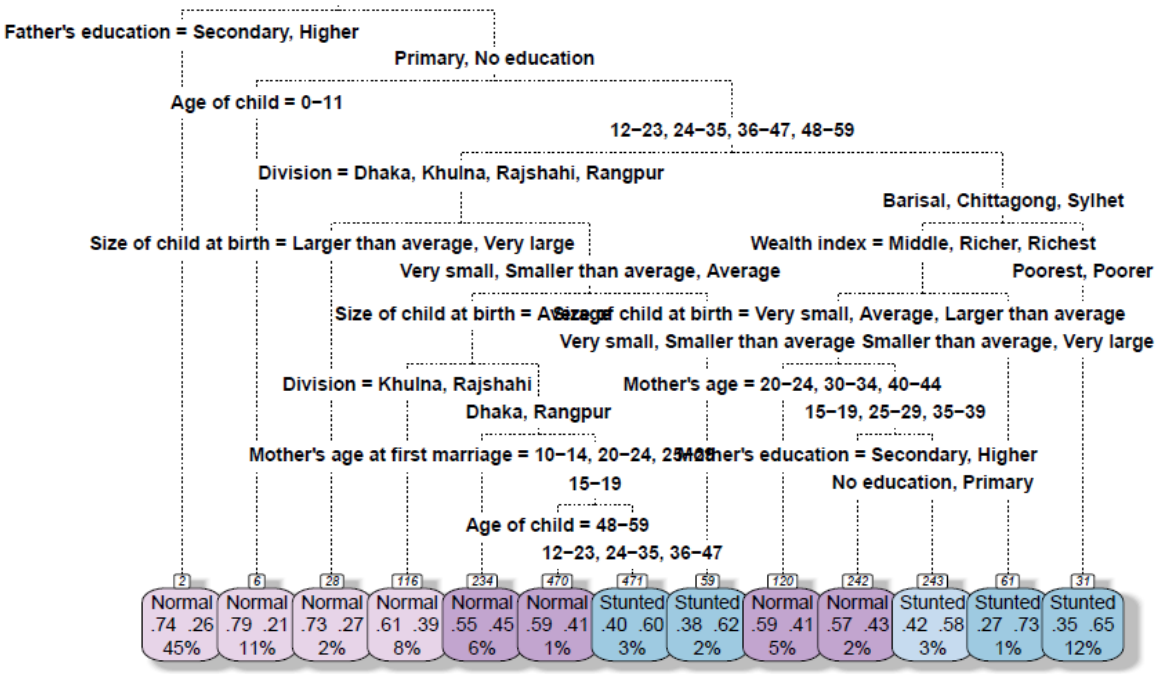

Figure 3 Classification tree for predicting stunting children.

Citation: Hemo SA, Rayhan MI. Classification tree and random forest model to predict under-five malnutrition in Bangladesh. *Biom Biostat Int J.* 2021;10(3):116–123. DOI: 10.15406/bbij.2021.10.00337

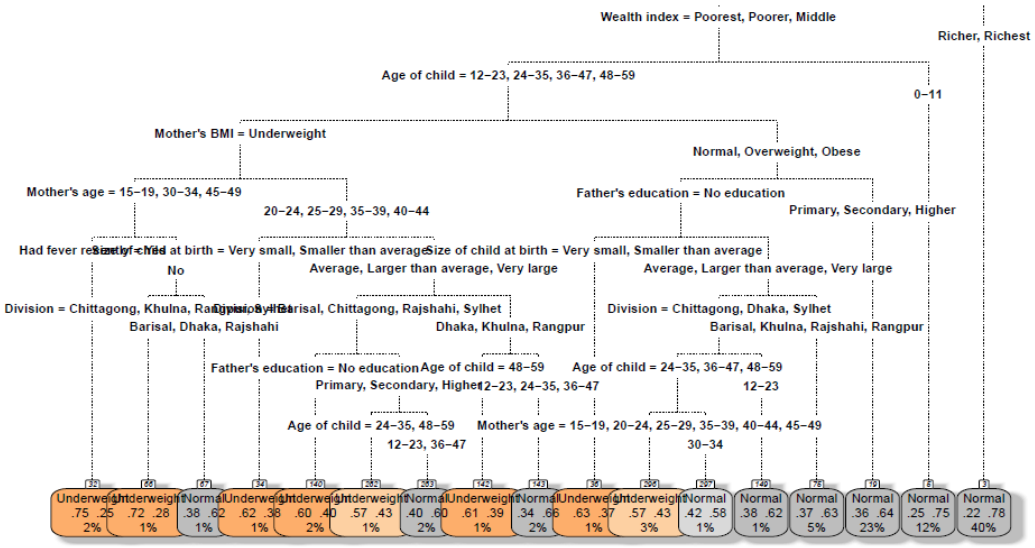

Figure 4 Classification tree for predicting underweight children.

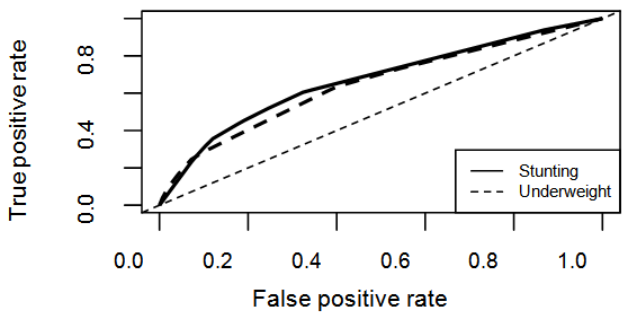

Figure 5 Area under curve of the ROC curve.

To assess the relationship of background characteristics in predicting the status of each of the dependent variable, a logistic regression model is fitted. This study illustrates that, children from poorest family has highest risk of being malnourished. They are almost 3 times as likely to be stunted and 2.3 times as likely to be underweighted than the children of richest family. This study also indicates that children aged 12-59 months are more than three times as likely to be stunted and more than two times as likely to be underweight than the children aged 0-11 months. Children whose mother is underweight has 84% more chance of being stunted and three times likely to be underweighted than the children of an obese mother. And, male children are almost at 11% risk of being stunted.

Table 2 Odds ratios in Multiple Logistic Regressions Assessing the Impacts of Variables

Dependent Variable		**Stunting**			**Underweight**	
Coefficient	**OR**	**95% CI**	**Sig.**	**OR 95% CI**		**Sig.**
Age of Mother						
15-19	2.3	1.4-3.8	0	2.01	1.21-3.4	0.01
20-24	1.58	0.99-2.54	0.06	1.54	0.96-2.53	0.08
25-29	1.52	0.97-2.41	0.07	1.39	0.87-2.25	0.17
30-34	1.19	0.76-1.88	0.45	1.36	0.86-2.2	0.19
35-39	1.33	0.83-2.15	0.24	1.59	0.98-2.62	0.07
40-44	1			1		
45-49	2.46	1.05-6.04	0.04	2.38	1.03-5.59	0.04
Wealth Index						
Poorest	2.96	2.36-3.71	0	2.36	1.88-2.97	0
Poorer	2.31	1.86-2.86	0	1.92	1.54-2.39	0
Middle	2.08	1.71-2.53	0	1.7	1.39-2.08	0
Richer	1.6	1.33-1.94	0	1.28	1.05-1.56	0.01
Richest	1			1		

Citation: Hemo SA, Rayhan MI. Classification tree and random forest model to predict under-five malnutrition in Bangladesh. *Biom Biostat Int J.* 2021;10(3):116–123. DOI: 10.15406/bbij.2021.10.00337

Table Continued...

	Dependent Variable		**Stunting**		**Underweight**		
	Coefficient	**OR**	**95% CI**	**Sig.**	**OR 95% CI**		**Sig.**
Mother's education							
	No education	1.76	1.35-2.29	0	1.6	1.22-2.1	0
	Primary	1.55	1.23-1.97	0	1.52	1.19-1.95	0
	Secondary	1.16	0.93-1.44	0.19	1.18	0.94-1.48	0.16
	Higher	1			1		
Division							
	Barisal	1.48	1.18-1.86	0	1.22	0.97-1.54	0.09
	Chittagong	1.65	1.34-2.04	0	1.47	1.19-1.82	0
	Dhaka	1.46	1.18-1.81	0	1.15	0.93-1.44	0.21
	Khulna	1			1		
	Rajshahi	1.04	0.82-1.3	0.77	1.12	0.89-1.41	0.32
	Rangpur	1.21	0.97-1.52	0.1	1.11	0.89-1.4	0.35
	Sylhet	2.18	1.76-2.72	0	1.42	1.14-1.77	0
Age of child							
	0-11	1			1		
	23-Dec	3.45	2.88-4.13	0	2.04	1.7-2.44	0
	24-35	3.9	3.25-4.69	0	2.72	2.27-3.27	0
	36-47	4.3	3.57-5.18	0	2.65	2.2-3.2	0
	48-59	3.26	2.69-3.96	0	3.05	2.52-3.69	0
Mother's BMI							
	Underweight	1.91	1.49-2.47	0	3.22	2.47-4.22	0
	Normal	1.56	1.24-1.99	0	1.84	1.44-2.39	0
	Overweight	1.5	1.18-1.91	0	1.43	1.11-1.87	0.01
	Obese	1			1		
Order of birth							
	1	1			1		
	2	1.22	1.05-1.42	0.01	1.17	1.01-1.37	0.04
	3	1.28	1.05-1.56	0.02	1.23	1-1.51	0.05
	4+	1.5	1.19-1.9	0	1.22	0.96-1.54	0.11
Sex of child							
	Female	1			1.04	0.93-1.15	0.52
	Male	1.1	0.99-1.22	0.07	1		

Abbreviations: CI=confidence interval, OR=odds ratio

Performance measure

In order to compare different predictive models, some accuracy measure has been carried out to find the best performing models. For measuring model performance, we have calculated area under curve (AUC) of the receiving operating characteristic (ROC) curve. The comparison is presented below:

Table 3 Performance measure of random forest model

	Model using Random Forest	
Performance measure	**Stunting**	**Underweight**
Accuracy	70.10%	72.40%
95% CI	(68.6%, 71.5%)	(70.9%, 73.8%)
No Information Rate	80.10%	85.60%
Sensitivity	59.10%	56.10%
Specificity	72.80%	75.10%
AUC	69.80%	70.00%

Table 4 Performance measure for classification tree

	Model using Classification Tree	
Performance measure	**Stunting**	**Underweight**
Accuracy	68.70%	70.50%
95% CI	(67.6%, 69.8%)	(69.4%, 71.6%)
No Information Rate	63.10%	67.30%
Sensitivity	35.60%	24.60%
Specificity	88.00%	92.60%
AUC	67%	65%

Conclusion

The demographic characteristics of the study including household socioeconomic status (poorest), and parents' education (no formal education) are appeared to influence the prevalence of malnutrition significantly. Random forest has an accuracy of 70.1% and 72.4% for predicting stunting and underweight, respectively. Classification tree has predicted 68.7% and 70.5% of children's stunting and underweight accurately, respectively. Classification tree has a higher specificity while random forest has higher sensitivity and AUC for predicting nutritional status. This study suggests that random forest has a better performance than the classification tree and multiple logistic regression model in predicting the stunting and underweight status of under-five children in Bangladesh.[6,16] Several target-based and fact-finding interventions can be taken to build a healthy future generation, today's child is tomorrow's leader. This study suggested a few pathways for future policymaking that are yet to be needed for development.

Conflicts of interest

Authors do not have any conflict of interest.

Funding information

Authors do not receive any kind of funding from any institutions throughout the study.

Contributor statement

The authors have read and approved the final manuscript. Both the authors have the equal contribution to perform the research, design the research study, contribute essential reagents or tools, analyze the data and write the paper.

References

1. Smith LC, Ramakrishnan U, Ndiaye, A. The importance of women's status for child nutrition in developing countries. *International Food Policy Research Institute (IFPRI) Research Report Abstract 131*. 2003;24(3):287–288.
2. Toma AS, Talukder A, Khan SS, et al. An assessment of the association between antenatal care and child malnutrition in Bangladesh. *Family Medicine & Primary Care Review*. 2018;20(4):373-378.
3. Sarma H, Khan JR, et al. Factors Influencing the Prevalence of Stunting Among Children Aged Below Five Years in Bangladesh. *Food Nutr Bull*. 2017;38(3):291–301.
4. Smith LC, Haddad LJ. Explaining child malnutrition in developing countries: *A cross-country analysis. International Food Policy Research Institute*. 2000.
5. Khare S, Kavyashree S, Gupta D, et al. Investigation of nutritional status of children based on machine learning techniques using indian demographic and health survey data. *Procedia Computer Science*. 2017;115:338–349.
6. Talukder A, Ahammed B. Machine learning algorithms for predicting malnutrition among under-five children in bangladesh. *Nutrition*. 2020;78:110861.
7. Batterham M, Tapsell L, Charlton K, et al. Using data mining to predict success in a weight loss trial. *J Hum Nutr Diet*. 2007;30(4):471–478.
8. Bath PA. Data mining in health and medical information. 2005.
9. Rayhan, Md. Israt M. Hayat Khan S. Factors causing malnutrition among under five children in bangladesh. *Pakistan Journal of Nutrition*. 2006;5:558–562.
10. Alom J, Quddus A, Amirul Islam M. Nutritional staus of under-five children in bangladesh: a multilevel analysis. *J Biosoc Sci*. 2012;44(5):525–535.
11. Rahman MS, Howlader T, Masud MS, et al. Association of low-birth weight with malnutrition in children under five years in bangladesh: do mother's education, socio-economic status, and birth interval matter? *PLoS ONE*. 2016;11(6):e0157814.
12. Sultana P, Rahman MM, Akter J. Correlates of stunting among under-five children in Bangladesh: a multilevel approach. *BMC Nutr*. 2019;5:41.
13. WHO working group. use and interpretation of anthropometric indicators of nutritional status. *Bull World Health Organ*. 1986;64(6):929–941.
14. Hosmer DW. Lemeshow S. applied logistic regression: hosmer/applied logistic regression. Hoboken, NJ, USA: John Wiley & Sons. 2000.
15. James G, Witten D, Hastie T, et al. An introduction to statistical learning. Vol. 103. Springer Texts in Statistics. New York, NY: Springer New York; 2013.
16. Markos Z. Predicting Under Nutrition status of under-five children using data mining techniques: The Case of 2011 Ethiopian Demo- graphic and Health Survey. *Journal of Health & Medical Informatics*. 2014;5(2).