Child Malnutritional Status In Bangladesh: Machine Learning Approaches

1st Sandamini Senaratne
Dept of Statistics & Data Science
University of Central Florida
Orlando, FL, USA
lo602443@ucf.edu

2nd Shahd Alnofaie
Dept of Statistics & Data Science
University of Central Florida
Orlando, FL, USA
sh467442@ucf.edu

3rd Md Mehedi Bhuiyan
Dept of Statistics & Data Science
University of Central Florida
Orlando, FL, USA
mdmehedihasan.bhuiyan@ucf.edu

***Abstract*—Childhood malnutrition remains a significant concern in developing countries. This study utilizes machine learning algorithms to predict the malnutrition status of households based on data collected through the Multiple Indicator Cluster Survey (MICS) conducted by UNICEF in 2019 in Bangladesh. The dataset consists of variables such as prenatal care, education levels, sanitation facilities, and economic status. Factor Analysis for Mixed Data (FAMD) combined with Logistic Regression, Lasso Logistic Regression, Random Forest, and Support Vector Machine algorithms have been used in the advanced analysis. Bivariate analysis, including t-tests and chi-square tests, examines the association between predictor variables and malnutrition outcomes. The Random Forest model gives out the best results with high sensitivity, specificity, and accuracy. This suggests that Random Forests can be use in predicting child malnutrition status for households in Bangladesh. Yet, because of other influencing factors and potential biases in survey data means we might need more ongoing research and validation to get to a cohesive decision. As such new variables and new qualitative studies can be introduced to make the predictions more accurate.**

***Index Terms*—malnutrition, Bangladesh, mixed data, random forest, logistic**

I. Introduction

Childhood malnutrition occurs due to deficiencies in minerals, vitamins, and lack of physical requirements for growth. This results in conditions such as underweight, stunting, wasting, and overweight. These condition adversely affects linear growth, both physically and mentally, and contributes to various health issues such as chronic diseases, cardiovascular problems, compromised immune systems, and increased mortality risk [1], [2].

Although measurements to reduce child malnutrition are implemented globally over the past decades, child malnutrition still remains a significant concern in developing countries such as Bangladesh. According to studies done in the past, globally, percentage of malnourished children seem to have been declined between the years 2000 and 2022. In the year 2000 the percentages of underweight, stunted, wasted, and overweight children are 25%, 33%, 8.7%, and 5.3% respectively. Whereas, by the year of 2022 the values has been changed to 13%, 22%, 6.8%, 5.6%. In the perspective of Bangladesh, it is not the case, according to a study that have been conducted within Bangladesh comparing the rates in 2013 and 2019. In 2013, the percentages of underweight, stunted, wasted, and overweight have been 21.9%, 42%, 9.6%, and 1.6% respectively. By the year of 2019 the rates have been changed up to, 16.6%, 28%, 9.8%, and 2.4% [3], [4], [5].

Child anthropometric (the study of human body measurements) measurements for ages 0-5 years involve continuous cases. Hence, according to the World Health Organization (WHO), Z-scores are used to assess malnutrition, with a Z-score lower than two standard deviations (SD) from the median indicating malnutrition. This includes underweight ($< -2SD$ weight for age), stunting ($< -2SD$ height for age), wasting ($< -2SD$ weight for height), and overweight ($> 2SD$ weight for height) [5] [6].

In this study we are trying to predict the child malnutrition status for each household. The report is structured into several sections. First we will be introducing the methodology that is used in the project. Then, the statistical analysis is done for the data set used using several machine learning algorithms. Finally, the discussion and the conclusion is provided.

II. Methodology

A. Data Collection

In collaboration with the Bangladesh Bureau of Statistics (BBS), UNICEF has carried out a survey named Multiple Indicator Cluster Survey (MICS) in 2019. The survey encompass 64,000 households across eight divisions, employing a two-stage stratified sampling technique. Among these households, 21,000 has been specifically selected due to the presence of children aged 0-5 years.

B. Variables

The quantification of malnutrition involves determining the total count of children aged 0-5 years falling into various malnourished categories within each household. Malnutrition is a dichotomous response variable. In this particular study, four separate data sets will be analysed. Each for underweight, stunted, wasted, and overweight respectively. Therefore, response variables for each data set consists of values 0, and 1, where 1 specifies a potential malnourished status of underweight, stunted, wasted, or overweight in a household.

Identify applicable funding agency here. If none, delete this.

The predictor variables consists of categorical, numerical, and continuous data. The variable descriptions are given in table I.

The variables ANC, $cdisability$, and $childbirthweight3$ were deleted since they had more than 50% missing values.

TABLE I
VARIABLES AND DESCRIPTIONS

Variable	Description
nochildbirth	Total number of childbirths given by a woman
ANC	Prenatal care (Yes/No)
homedelivery	Delivery status: Home/Hospital
disability	Mother's physical disability (Yes/No)
chage	Child's age in months
cdisability	Child's physical disability (Yes/No)
melevel	Mother's education level (No educ, Primary, Secondary, Higher)
illness	Child's illness last week (Yes/No)
antibiotic	Antibiotic taken during illness (Yes/No)
division	Geographical location
hhsize	Total members in a household
helevel	Father's education level (No educ, Primary, Secondary, Higher)
sex	Child's gender (Male/Female)
area	Living area (Urban/Rural)
sanitation	Sanitation facility in the household (Open/Flush toilet)
pwater	Drinking tubewell water (Yes/No)
iodin	Consumption of iodine from salt (No, 0-15 ppt, 15+ ppt)
wageatb	Woman's age during childbirth
windex3	Economic status (Poor, Middle, High)
childbirthweight3	Childbirth weight (Average, Underweight, Overweight)

C. Methods Applied

1) Factor Analysis for Mixed Data (FAMD): This model combines continuous and categorical variables, expressing observed variables X as a sum of latent factors F, factor loadings L, unique factors Ψ, and residuals ε:

$$X = LF + \Psi + \varepsilon$$

For continuous variables X_c:

$$X_c = L_c F + \Psi_c + \varepsilon_c$$

For categorical variables X_{cat}, where L_{cat} represents loadings for dummy variables:

$$X_{cat} = L_{cat} F + \Psi_{cat} + \varepsilon_{cat}$$

The factor loadings matrix L captures relationships, Ψ accounts for unique variance, and ε represents unexplained variance. FAMD aims to estimate these parameters for the best fit, typically using Maximum Likelihood Estimation (MLE) or Principal Component Analysis (PCA).

2) Logistic Regression: The multiple binary logistic regression model is represented as:

$$\pi(X) = \frac{e^{X\beta}}{1 + e^{X\beta}} = \frac{1}{1 + e^{-X\beta}}$$

where: π is the probability of an observation falling into a specified category, X represents the independent variables, β is the vector of coefficients.

The logistic model ensures that the estimated probabilities (π) are always between 0 and 1. The likelihood function for a sample of size n is given by:

$$L(\beta; y, X) = \prod_{i=1}^{n} \left(\frac{e^{X_i\beta}}{1 + e^{X_i\beta}} \right)^{y_i} \left(\frac{1}{1 + e^{X_i\beta}} \right)^{1-y_i}$$

This leads to the log likelihood:

$$\ell(\beta) = \sum_{i=1}^{n} \left[y_i X_i \beta - \log(1 + e^{X_i\beta}) \right]$$

The logistic regression model describes the probability of an event as a function of the independent variables (X), ensuring that the estimated probabilities are confined to the range of 0 to 1. The likelihood function quantifies the probability of observing the given outcomes in the sample, and the log likelihood simplifies the computations.

3) Random Forests: Random Forest is an ensemble learning technique that combines the predictions of multiple decision trees to improve accuracy and robustness. It is widely used for both classification and regression tasks. The key idea behind Random Forest is to train a collection of diverse trees on different subsets of the data, and then combine their predictions through voting (for classification) or averaging (for regression). Random Forest is known for its ability to handle complex relationships in data, avoid overfitting, and provide feature importance insights.

For a given observation, let X represent the input features, and $T_1, T_2, ..., T_n$ be the individual decision trees in the Random Forest. The probability of class k for that observation is computed as follows:

$$P(Y = k|X) = \frac{1}{N} \sum_{i=1}^{N} P_i(Y = k|X)$$

where N is the number of trees in the forest, and $P_i(Y = k|X)$ is the probability predicted by the i-th tree.

The final predicted class for the observation is the one with the highest probability. In practice, Random Forest training involves constructing decision trees based on bootstrapped samples of the data and randomly selecting subsets of features at each split. The ensemble nature of Random Forest allows it to capture complex patterns in data while mitigating overfitting.

4) Support Vector Machines: This is a powerful supervised machine learning algorithm used for both classification and regression tasks. SVM aims to find a hyperplane in a high-dimensional space that best separates the data into distinct classes. The key idea is to maximize the margin, which is the distance between the hyperplane and the nearest data points (support vectors) from each class. SVM is effective in handling complex decision boundaries and is particularly useful in scenarios with high-dimensional feature spaces.

Given a dataset with input features X and corresponding class labels Y (1 or -1 for binary classification), the linear SVM seeks to find a hyperplane defined by the equation:

$$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n$$

where: $f(X)$ represents the decision function, β_0 is the intercept, $\beta_1, \beta_2, \ldots, \beta_n$ are the coefficients of the hyperplane, $X_1, X_2, \ldots, X_n$ are the input features.

The decision boundary is formed by points where $f(X) = 0$. The goal of SVM is to find β values that maximize the margin, subject to the constraint that data points are correctly classified.

SVM is a versatile algorithm with the flexibility to handle linear and non-linear classification problems. Its effectiveness lies in finding optimal decision boundaries that maximize the margin between different classes.

III. Statistical Analysis

A. Bivariate Analysis

In order to check for the association between the response variable and the predictor variables a bivariate analysis was done. To check for the relationship between response and the categorical variables a t-test was used, while a chi-square test was used to check for the association between the response and the continuous variables. The p-values for each of the tests for each bivariate comparison is specified in Table II.

p-Values: The entries in the table are p-values associated with each variable. These values are used to assess the statistical significance of the relationship between each variable and the response variable of each data set: underweight, stunted, wasted, and overweight.

Significance Levels: Values less than 0.05 suggest a statistically significant relationship between the variable and the outcome. The notation <2.2e-16 represents a very small p-value, indicating extremely high statistical significance.

From the results shown in Table II it can be seen that, there is no relationship between the response and the variables $disability$, $antibiotic$ since there p-values are greater than 0.05 in the dataset underweight. With respect to the dataset stunted, the variables $disability$, $illness$, $wageatb$, and $pwater$ seem to have no association with the response variable. Moroever, in the datatset named wasted, the response variable seems to be associated with all the variables. However, the variables $homedelivery$, $hhsize$, and $wageatb$ might not have a relationship with the dependent variable. These speculations need to be further analysed in order to come up with a proper conclusion.

TABLE II
CHI- SQUARE AND T-TEST RESULTS

	underweight	stunted	wasted	overweight
nochildbirth	<2.2e-16	<2.2e-16	<2.2e-16	2.412e-07
homedelivery	1.291e-08	3.243e-10	<2.2e-16	0.8895
disability	0.9738	0.9013	0.02209	0.003402
chage	< 2.2e-16	< 2.2e-16	0.004655	<2.2e-16
melevel	< 2.2e-16	< 2.2e-16	< 2.2e-16	< 2.2e-16
illness	0.005732	0.5834	4.771e-09	9.528e-09
antibiotic	0.7485	0.01047	1.826e-05	4.278e-05
division	< 2.2e-16	< 2.2e-16	6.695e-08	< 2.2e-16
sex	0.004614	0.001823	3.137e-09	1.239e-05
area	< 2.2e-16	4.959e-14	0.0001063	< 2.2e-16
hhsize	7.254e-09	9.595e-11	4.484e-13	0.06889
helevel	< 2.2e-16	< 2.2e-16	< 2.2e-16	1.4e-12
sanitation	< 2.2e-16	< 2.2e-16	< 2.2e-16	< 2.2e-16
pwater	0.0006436	0.6844	0.0001169	< 2.2e-16
iodin	4.889e-12	2.2e-16	7.794e-08	< 2.2e-16
wageatb	0.0243	0.4165	1.615e-09	0.4087
windex3	<2.2e-16	<2.2e-16	<2.2e-16	< 2.2e-16

B. Model Analysis

The machine learning algorithms explained in section II are implemented on the dataset separately.

The table III presents sensitivity, specificity, and accuracy metrics for different models used in the analysis. Each model is evaluated based on its performance for predicting outcomes related to underweight, stunted, wasted, and overweight conditions.

Each row of the table include the different models used in the analysis, including FAMD with Logistic Regression, Lasso Logistic Regression, Random Forest, and Support Vector Machine.

The evaluation metrics used are,

Sensitivity: The proportion of true positive predictions among all actual positive instances. It measures the model's ability to correctly identify positive cases.

Specificity: The proportion of true negative predictions among all actual negative instances. It measures the model's ability to correctly identify negative cases.

Accuracy: The overall correctness of the model, measuring the proportion of correctly predicted instances (both true positives and true negatives) among all instances.

If we consider how each model performed for used datasets separately it can be seen that the FAMD with Logistic Regression model has performed moderately for underweight and stunted conditions but has lower accuracy for wasted and overweight. Furthermore, the Lasso Logistic Regression model has shown a balanced performance across different outcomes, with moderate sensitivity, specificity, and accuracy. The Random Forest model has demonstrated high performance across all outcomes, with strong sensitivity, specificity, and accuracy values. The performance of the Support Vector Machine varies across outcomes, with moderate sensitivity, specificity, and accuracy.

IV. Conclusion and Discussion

According to the results of the model analysis given in Table III, it can be said that the Random Forest would be a good option in predicting the malnutrition status of a household in Bangladesh. This model shows good performance across different malnutrition categories of datasets, making it better in identifying vulnerable households. The accuracy values were 86%, 86%, 97%, and 98% for each case of underweight, stunted, wasted, and overweight. Moreover, the sensitivity values were, 86%, 85%, 96%, and 97% and the specificity

TABLE III
SENSITIVITY, SPECIFICITY, AND ACCURACY FOR THE MODELS USED

Model	Sensitivity	Specificity	Accuracy
FAMD with Logistic Regression			
Underweight	0.64	0.41	0.53
Stunted	0.54	0.52	0.53
Wasted	0.77	0.27	0.52
Overweight	0.74	0.23	0.48
Lasso Logistic Regression			
Underweight	0.53	0.60	0.56
Stunted	0.53	0.58	0.55
Wasted	0.45	0.61	0.53
Overweight	0.45	0.67	0.56
Random Forest			
Underweight	0.86	0.87	0.86
Stunted	0.85	0.83	0.86
Wasted	0.96	0.98	0.97
Overweight	0.97	0.99	0.98
Support Vector Machine			
Underweight	0.5481	0.5764	0.5623
Stunted	0.5656	0.5720	0.5688
Wasted	0.5216	0.5420	0.5319
Overweight	0.5994	0.5631	0.5812

results were 87%, 83%, 98%, and 99% for underweight, stunted, wasted, and overweight respectively.

Moreover, a similar study has been done on the Bangladesh malnutrition dataset the results we attained show better results. [8] has also achieved best results with a Random Forest algorithm, but with an accuracy of 68.51%, a sensitivity of 94.66%, and a specificity of 69.76%. Where as the study we conducted gave us better accuarcaies for all the separate conditions.

Despite the promising results, there is possibility of potential biases in survey data and the complexity of factors influencing malnutrition. Future work should explore additional variables, conduct in-depth qualitative studies, and validate models on diverse datasets for enhanced generalizability.

Addressing childhood malnutrition in Bangladesh requires a diverse approach that considers regional variations, socioeconomic factors, and healthcare access. Moreover, state-of-art machine learning techniques can be applied to get better predicted results.

To increase the rate of reduction of child malnutrition more, Government and non-government organisations could focus on improving education and household income-generating activities among poor households and raising awareness among women about the importance of receiving antenatal care during pregnancy. [9].

V. APPENDIX

Group 1: Project code

ACKNOWLEDGMENT

We express our gratitude to Professor Rui Xie for providing valuable advice and guidance with constructive comments that contributed towards the improvement of our project.

REFERENCES

[1] Jubayer F, Kayshar S, Arifin S, Parven A, Khan SI, Meftaul IM. Nutritional health of the Rohingya refugees in Bangladesh: Conceptualizing a multilevel action framework focusing the COVID-19. Nutrition and Health. 2023;0(0). doi:10.1177/02601060231169372

[2] Hossain MM, Tasnim S, Sultana A, Faizah F, Mazumder H, Zou L, McKyer ELJ, Ahmed HU, Ma P. Epidemiology of mental health problems in COVID-19: a review. F1000Res. 2020 Jun 23;9:636. doi:0.12688/f1000research.24457.1. PMID: 33093946; PMCID: PMC7549174.

[3] World Health Organization, United Nations Children's Fund (UNICEF) & International Bank for Reconstruction and Development/The World Bank, *Levels and trends in child malnutrition: UNICEF/WHO/World Bank Group joint child malnutrition estimates: key findings of the 2023 edition*. Publisher, 2023. ISBN: 978-92-4-007379-1.

[4] Akombi BJ, Agho KE, Hall JJ, Wali N, Renzaho AMN, Merom D. Stunting, Wasting and Underweight in Sub-Saharan Africa: A Systematic Review. Int J Environ Res Public Health. 2017 Aug 1;14(8):863. doi: 10.3390/ijerph14080863. PMID: 28788108; PMCID: PMC5580567.

[5] Hossain MM, Abdulla F, Rahman A. Prevalence and risk factors of underweight among under-5 children in Bangladesh: Evidence from a countrywide cross-sectional study. PLoS One. 2023 Apr 24;18(4):e0284797. doi: 10.1371/journal.pone.0284797. PMID: 37093817; PMCID: PMC10124832.

[6] Bloem M. The 2006 WHO child growth standards. BMJ. 2007 Apr 7;334(7596):705-6. doi: 10.1136/bmj.39155.658843.BE. PMID: 17413142; PMCID: PMC1847861.

[7] de Onis M, Onyango A, Borghi E, Siyam A, Blössner M, Lutter C; WHO Multicentre Growth Reference Study Group. Worldwide implementation of the WHO Child Growth Standards. Public Health Nutr. 2012 Sep;15(9):1603-10. doi: 10.1017/S136898001200105X. Epub 2012 Apr 12. PMID: 22717390.

[8] Talukder A, Ahammed B. Machine learning algorithms for predicting malnutrition among under-five children in Bangladesh. Nutrition. 2020 Oct;78:110861. doi: 10.1016/j.nut.2020.110861. Epub 2020 May 15. PMID: 32592978.

[9] Rahman MT, Jahangir Alam M, Ahmed N, Roy DC, Sultana P. Trend of risk and correlates of under-five child undernutrition in Bangladesh: an analysis based on Bangladesh Demographic and Health Survey data, 2007-2017/2018. BMJ Open. 2023 Jun 12;13(6):e070480. doi: 10.1136/bmjopen-2022-070480. PMID: 37308267; PMCID: PMC10277110.