"Predicting Stunting in Toddlers Under Five: A Machine Learning Approach Using Height, Gender, and Age"

CSC 642: Statistical Learning with Applications

Jonathan Latim & Malena Fuentes

Spring 2024

1. Introduction

Childhood stunting, defined as a condition of impaired growth and development, continues to be a critical global health issue that affects children under the age of five, specifically in low-income and developed countries. Stunting not only reflects chronic malnutrition but also indicates inadequate physical and cognitive development, leading to long-term health and societal consequences. According to the World Health Organization (WHO), approximately 150 million children under the age of five worldwide were estimated to be stunted in 2021. It is turned into a universal problem because it is also related to an increased risk of morbidity and mortality. The burden of stunting is particularly severe in low‑ and middle‑income countries where 98% of stunted children reside (WHO, 2021).

Childhood stunting is a complex issue with multifaceted causes, including nutritional, environmental, and socioeconomic factors. Traditional statistical methods have been used to study and address stunting, but machine learning offers a powerful alternative approach. Machine learning techniques, such as those employed in this study, can analyze large datasets like the one based on the WHO z-score formula with greater efficiency and accuracy, allowing for the identification of subtle patterns and interactions among variables that may not be apparent using traditional methods.

In this study, we aim to predict stunting in toddlers under the age of five based on key factors such as height, gender, and age. The dataset used in this analysis is based on the WHO z-score formula, which provides a standardized measure of how a child's height compares to the heights of other children of the same age and gender. By leveraging machine learning techniques, we seek to identify the significant predictors of stunting and develop a model that can effectively classify toddlers in four categories, from severely stunted to tall. This research is crucial as it can help healthcare practitioners, policymakers, and researchers better understand the determinants of stunting and develop targeted interventions to prevent and address this pressing public health issue. By identifying children at risk of stunting early on, we can improve the effectiveness of nutritional interventions and ultimately enhance the health and well-being of children worldwide.

2. State-of-the-art

Childhood stunting, characterized by impaired growth and development, is a critical global health issue affecting children under the age of five. In Indonesia, the prevalence of stunting is particularly concerning, with approximately 37% of children affected. To address this issue, researchers have utilized innovative approaches, such as machine learning, to predict stunting in toddlers based on key factors like height, gender, and age. Haris, M. S., Anshori, M., & Khudori, A. N. (2023) employed the random forest algorithm to predict stunting prevalence in East Java Province, Indonesia, considering 20 potential factors, of which only 12 were found to be correlated with stunting. Despite this, the random forest model did not surpass multi-linear regression, yielding MAE and MSE error values of 1.02 and 1.64, respectively, on a dataset of 38 data points. Ndagijimana et al. (2023) developed a model to predict stunting in Rwandan children under five, using machine learning and data from the Rwanda Demographic and Health Survey 2019-2020. Their best model, based on the gradient boosting classifier algorithm, achieved an 80.49% training accuracy. The model identified predictors such as mother's height, television, child's age, province, mother's education, birth weight, and childbirth size. This study highlights the potential of machine learning for accurately predicting stunting in Rwandan children, offering valuable insights for public health interventions.

Rahman et al. (2021) investigated risk factors for stunting, wasting, and underweight in under-five Bangladeshi children, using machine learning to predict these outcomes. Analyzing data from the 2014 Bangladesh Demographic and Health Survey with 7079 children and 13 factors, logistic regression identified potential risks. Machine learning classifiers, including support vector machine, random forest, and logistic regression, were used for prediction. Results showed average prevalence rates of 35.4% for stunting, 15.4% for wasting, and 32.8% for underweight. Random forest achieved the highest accuracy in classification, demonstrating its effectiveness in identifying malnutrition risks and aiding policymakers in addressing malnutrition among Bangladeshi children. In a 2016 study, Mardani et al. used secondary data from the Indonesian Family Life Survey (IFLS) 2007 to develop a model for predicting and preventing stunting in Indonesian children under five years old. The study included 3589 children meeting specific inclusion criteria. Multiple logistic regression and Receiver Operating Characteristic (ROC) Curve analysis were employed for modeling and diagnostic testing,

respectively. The developed model was able to prevent and delay stunting in toddlers by 64%, with 61.9% sensitivity and 60.9% specificity, and an AUROC of 65.5%.

3. Materials and Methods

3.1 Dataset

The "Stunting Toddler (Balita) Detection" dataset was obtained from Kaggle and contains no missing values. It is based on the z-score formula for determining stunting according to the World Health Organization (WHO) and focuses on stunting detection in children under five years old. The dataset contains 121,000 rows of data, including information on the age, sex, height, and nutritional status of toddlers. It is a valuable resource for researchers, nutritionists, and policymakers working in child health and nutrition, providing insights for the development of effective interventions and policies.

The dataset "Stunting Baby/Toddler Detection" was imported and cleaned for analysis. Initially, the dataset was read from a CSV file and examined to understand its structure. The dataset contained columns such as Age_month, Gender, Height_cm, and Nutritional_status, with some entries in Indonesian. The data was then relabeled in English, with columns renamed and certain values translated. Missing values were checked and found to be absent. One-hot encoding was applied to the Gender column to convert it into a numerical column. Age_month and Height_cm columns were then scaled to have them centered at zero with a standard deviation of one. This was done to avoid machine learning algorithms giving more importance to features with larger magnitudes. Scaling also helps some of the machine algorithms used in this project i.e. logistic regression and support vector machines to converge faster. Summary statistics were then calculated on the resulting dataset showing the distribution of data across different categories. Additionally, a bar plot was created to visualize the distribution of Nutritional_status categories in the dataset, as shown in Figure 1. Overall, the dataset was successfully imported, cleaned, and prepared for further analysis.

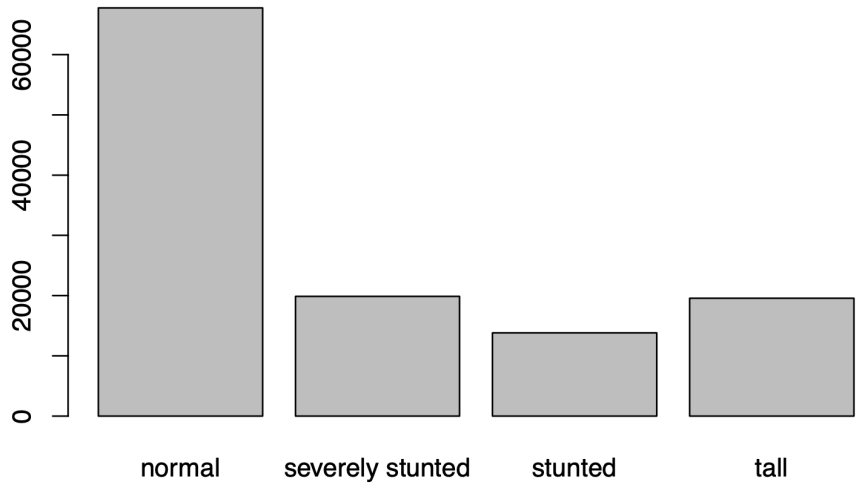

Figure 1: Distribution of Nutritional_status categories in the dataset, indicating the proportion of children classified as 'normal', 'severely stunted', 'stunted', and 'tall'.

Table 1 presents the distribution of gender among different nutritional status categories. The table shows the proportions of female and male children classified as normal, severely stunted, stunted, and tall. The proportions indicate that, overall, there is a relatively balanced distribution of genders across the nutritional status categories, with slight variations in some categories.

Nutritional Status	Female	Male
Normal	0.5070032	0.4929968
Severely Stunted	0.4778801	0.5221199
Stunted	0.5071299	0.4928701
Tall	0.5188650	0.4811350

Table 1: Gender Distribution by Nutritional Status, Showing Proportions of Female and Male Children in Each Category

3.2 Methods

Multinomial Logistic Regression

In this study, multinomial logistic regression was utilized as a method for predicting the nutritional status of children based on various features. Unlike binomial logistic regression, which is suitable for predicting two categories, multinomial logistic regression allows for the direct modeling of probabilities for each category, enabling a clearer interpretation of the effects of different features on the probabilities of each nutritional status category. Logistic regression was also chosen to be used as a starting point for our exploration of the data because it is an interpretable linear model. Its limitations set the stage for exploring more complex models. The model was trained using the ‘multinom’ function from the ‘nnet’ package in R, with the ‘Nutritional_status’ variable as the response variable and all other available variables as predictors. The model was fitted to the training data to predict the nutritional status of children based on their age, gender, and height. The 'car' package in R was used to conduct an ANOVA test on a multinomial logistic regression model fitted to the dataset. Type III ANOVA assessed the significance of each predictor variable. The results indicated that all predictors were statistically significant in predicting the nutritional status of children based on their age, gender, and height. Gender was found to be significant only when both male and female where kept together in the same column i.e. before one-hot encoding. After one-hot encoding either male or female columns were not found to be significant. From the “caret” package, confusionMatrix function was used to summarize how logistic regression (and every other machine learning algorithm used) performed in predicting the nutritional class of each child in the test dataset.

Decision Trees

Decision Trees are powerful models for predicting because they can capture complex, non-linear relationships between predictors and the outcome. Unlike linear models, Decision Trees create a tree-like structure that explicitly shows how predictions are made based on the predictor values. This method is particularly useful when the relationship between predictors and stunting status is not straightforward. This machine learning algorithm was chosen in order to capture any non-linear relationships that logistic regression may have missed. This was important in case the boundaries between the nutritional classes were not being captured using the chosen linear model (logistic regression).

Bagging

Bagging, short for Bootstrap Aggregating, is a powerful ensemble learning technique that improves the performance of decision trees by addressing overfitting and instability. It works by creating multiple decision trees, each trained on a randomly sampled subset of the original data with replacement. This means that some instances may be included multiple times in a subset, while others may not be included at all. By averaging the predictions of these individual trees (for regression) or using a majority vote (for classification), bagging reduces variance and improves the overall stability of the model. This approach helps overcome the limitations of a single decision tree and leads to a more robust and accurate prediction model. We applied bagging to our dataset by utilizing the randomForest package in R, which implements the bagging algorithm. This involved creating an ensemble of decision trees, each trained on a different bootstrap sample of the original data. The randomForest function was used to build the bagging model, specifying the number of trees (ntree) and the number of variables randomly sampled as candidates at each split (mtry). These parameters were tuned to optimize the model's performance. Once the bagging model was trained, we used it to predict the nutritional status of children in our testing dataset. The predictions were then aggregated to make a final prediction, providing a more stable and accurate classification compared to a single decision tree.

Random Forest

Random forest model was used as it extends the concepts of bagging by introducing additional randomness in the feature selection process in order to reduce overfitting. In a random forest model, at each split in a decision tree, only a random subset of features is considered, further diversifying the individual trees. This randomness enhances the reduction in variance compared to bagging, making the model more robust and less sensitive to the impact of specific features. We used the ‘randomForest’ function in R to build our random forests model, specifying 500 trees and a square root of the number of features (mtry = sqrt(m1)) as the number of variables randomly sampled at each split. The accuracy of the random forests model on the testing data was calculated to evaluate its performance, which demonstrated the effectiveness of the random forests approach in our analysis.

Support Vector Machines (SVMs)

In this project, the Support Vector Machine (SVM) algorithm with a radial kernel was applied to classify the nutritional status of children based on various features in the dataset. SVMs are a type of supervised learning model used for classification and regression analysis. The radial kernel in SVMs is particularly useful for handling non-linear relationships between variables, making it suitable for complex datasets where the decision boundary is not linear. The SVM model was trained using the 'svm' function from the 'e1071' package in R, with the 'cost' parameter and the 'gamma' parameter set to 1.

3.3 Evaluation

Model performance was evaluated using two key metrics: accuracy and kappa. Accuracy measures the proportion of correctly classified instances out of the total instances i.e. how often a model correctly classifies a child's nutritional status across all classes, providing a general overview of the model's performance. Kappa, on the other hand, considers the agreement between the predicted and actual classifications while accounting for the possibility of the agreement occurring by chance. It provides a more nuanced evaluation of the model's performance especially when there is class imbalance within the dataset as in this case.

At a more nuanced and practical level, we wanted to measure the performance of each model at how well they particularly predicted whether a child was stunted (i.e. severely stunted or stunted) or not as these are the groups that need help. We used sensitivity and specificity as the metrics to use in evaluating how well each model performed at predicting whether a child was stunted (stunted or severely stunted) or not.

Sensitivity measures how well a model performs at correctly predicting a child's specific nutritional status. High sensitivity in predicting both stunted and severely stunted classes has two practical importances i.e. it allows for early intervention to the children that need urgent nutritional support and also minimizes the long term consequences of stunting is identified and addressed early.

Specificity measures how well a model performs at predicting children who don't belong in a particular nutritional class. This has two practical importances i.e. it prevents uncalled for worrying among misidentified children and their families and most importantly, given these children are located in third world countries where resources are limited, high specificity ensures that limited resources and interventions are going to the children who truly need them.

4. Results

In this section, we show and discuss how each machine learning algorithm usedin this project performed at predicting all the nutritional classes and how each algorithm performed at specifically identifying both stunted nutritional classes i.e. severely stunted and stunted.

Table 2 summarizes the overall performance of the different machine learning models used in this study across all nutritional classes using two key metrics i.e. accuracy and kappa. The table provides a comprehensive overview of the predictive capabilities of each model,

highlighting their effectiveness in classifying nutritional status in children under five years old across all the nutritional classes.

The accuracy metric explains the overall proportion of correct predictions. Kappa measures agreement between predictions and true labels, adjusted for chance. Logistic Regression achieved an accuracy of 0.7767 and a kappa value of 0.6206, indicating moderate performance. Decision Trees performed slightly better, with an accuracy of 0.7857 and a kappa value of 0.6446. Bagging, a method that combines multiple models to improve accuracy, demonstrated exceptional performance with an accuracy of 0.9993 and a kappa value of 0.9989. Random Forest, another ensemble method, achieved an accuracy of 0.9647 and a kappa value of 0.9413, showcasing its effectiveness. Support Vector Machines (SVM) also performed well, with an accuracy of 0.9897 and a kappa value of 0.9832. These results suggest that Bagging and SVM are particularly effective in predicting nutritional status, followed by Random Forest, Decision Trees, and Logistic Regression.

	Accuracy	**Kappa**
Logistic Regression	0.7767	0.6206
Decision Trees	0.7857	0.6446
Bagging	0.9993	0.9989
Random Forest	0.9647	0.9413
Support Vector Machines	0.9897	0.9832

Table 2: Performance Metrics of Different Models

Machine Learning Algorithms	**Sensitivity**	**Specificity**
Logistic Regression	0.8571	0.9384
Decision Trees	0.8498	0.9469
Bagging	0.9995	0.9998
Random Forest	0.9985	0.9985
Support Vector Machines	0.9919	0.9972

Table 3: Comparison of Sensitivity and Specificity Across Machine Learning Algorithms for Severely Stunted Classification.

Machine Learning Algorithms	Sensitivity	Specificity
Logistic Regression	0.19261	0.98009
Decision Trees	0.33161	0.97539
Bagging	0.9963	0.9998
Random Forest	0.80148	0.99967
Support Vector Machines	0.9627	0.9975

Table 4: Comparison of Sensitivity and Specificity Across Machine Learning Algorithms for Stunted Classification.

In the logistic regression analysis, we explored the relationship between various factors (age, height, and gender) and the nutritional status of children. The multinomial logistic regression model was employed due to the presence of more than two categories in the target variable (Nutritional_status). The model successfully converged, indicating a successful fitting process. The coefficients of the model indicate how a one-unit change in each feature affects the log-odds of a child belonging to a specific nutritional status category relative to the baseline category (normal). To aid interpretation of the logistic's regression model summary, we exponentiated the coefficients to obtain odds ratios since the coefficients can't be meaningfully interpreted directly from the summary. Using the Anova test, we found all the predictors to be significant however gender was only significant before one-hot encoding split the gender roles into unique columns. The significance of all the predictors justify their use in predicting the nutritional classes of each child. This model showed for every one cm increase height, the odds of being classified as 'severely stunted' become extremely low compared to being normal, holding other features constant. This is an expected result as one can't be tall and stunted at the same time. We also found an unlikely association of gender to being severely stunted i.e. we found out thru this model that females have a lower likelihood of being classified as 'severely stunted' compared to males. This wasn't something we were expecting and can't explain, however, it would be a useful insight to find out why males are more likely than women to be classified as severely stunted and targeted interventions enforced to erase stunting in males. These findings underscore the importance of age, height, and gender in determining a child's nutritional status and also highlight how using models can help craft effective policies and interventions.

In evaluating the logistic regression model, several key metrics were used to assess its performance. The accuracy metric, which measures the overall proportion of correct predictions, yielded a value of 0.7767, indicating that the model correctly predicted the nutritional class about 77.67% of the time. The Kappa statistic, which measures agreement between predictions and true labels adjusted for chance, was found to be 0.6206, suggesting substantial agreement between the

model's predictions and the actual labels. Sensitivity, or recall, which measures the proportion of true positives within each class that the model correctly identifies, varies across classes. The model performed best in detecting the 'normal' class, with a sensitivity of 0.89, followed by 'severely stunted' at 0.8571. However, sensitivity was notably lower for the 'stunted' class, at 0.19261, indicating that the model struggled more with identifying these cases. This is attributed to class imbalance in the data and logistic regression being poor at handling such an issue. On the other hand, specificity, which measures the proportion of true negatives within each class that the model correctly identifies, was generally high, indicating that the model was effective at correctly identifying individuals who did not belong to a particular class.

The decision tree model, known for its ability to generate a clear, interpretable structure and capture complex, non-linear patterns, was employed to predict stunting status. This method is advantageous when the relationship between predictors and stunting status is not linear. The decision tree model achieved an accuracy of 0.7857, indicating that it correctly predicted nutritional class about 78.57% of the time. The Kappa statistic, which measures agreement adjusted for chance, was found to be 0.6446, indicating substantial agreement between predictions and true labels. Sensitivity varied across classes, with the model performing best in detecting the 'normal' class (sensitivity = 0.8498) and 'severely stunted' class (sensitivity = 0.9469). This model however had a sensitivity of 0.33161 when predicting the stunted class, a poor performance, because of the lack of enough data for this class. Such imbalance in the classes is something decision trees do not handle well. Specificity was generally high, indicating that the model was effective at correctly identifying individuals who did not belong to a particular class. The confusion matrix further illustrates the model's performance across different nutritional status categories, providing valuable insights into its predictive capabilities.

Both bagging and random forests are ensemble methods that use decision trees as their base model. Decision trees, while effective, can suffer from overfitting and sensitivity to small changes in the data. Bagging, short for bootstrap aggregating, was chosen for its ability to mitigate these issues by creating an ensemble of trees. Each tree is trained on a randomly sampled subset of the original data, with replacement. Predictions are then made by averaging (for regression) or majority voting (for classification) across all the trees, resulting in a more stable and less prone-to-overfitting model than a single decision tree. The random forest model, an extension of bagging, further enhances performance by introducing randomness in the variable selection process for each split in the tree. This randomness helps to decorrelate the trees, leading to improved generalization and predictive performance. The random forest model achieved an out-of-bag (OOB) estimate of the error rate of 0.1%, indicating high accuracy in predicting nutritional status categories. The importance plot, shown in Figure 2 illustrates the variables that most significantly influence the model's predictions, providing insights into the key factors driving stunting classification, where variables with higher Mean Decrease Accuracy values are more influential in the classification process.

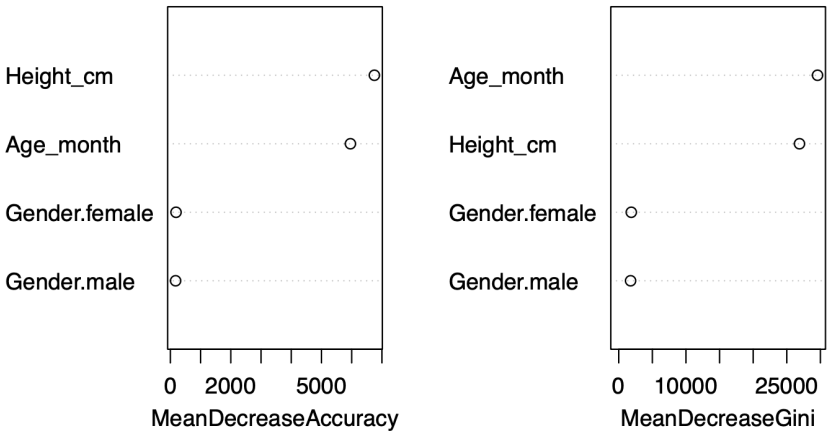

Figure 2: Variable Importance Plot: Bagging model showing the relative importance of variables in predicting nutritional status.

Random forests, chosen for their enhancement of bagging, introduce an additional level of randomness to further diversify the individual trees. This method selects only a random subset of features at each split, ensuring that the trees are even more distinct from each other. This diversity leads to improved variance reduction compared to bagging and enhances the model's robustness against the influence of strong individual features. The random forests model, trained with 500 trees and using the square root of the number of features at each split, achieved an out-of-bag error rate of 3.92%. This indicates that, on average, the model misclassified 3.92% of the cases. The confusion matrix shows that the model performs well in predicting the 'normal' class but struggles more with the 'stunted' and 'tall' classes, as evident from the higher misclassification rates in these categories. The variable importance plot in Figure 3, highlights the most influential features in predicting nutritional status, providing insights into which factors have the greatest impact on the model's predictions.

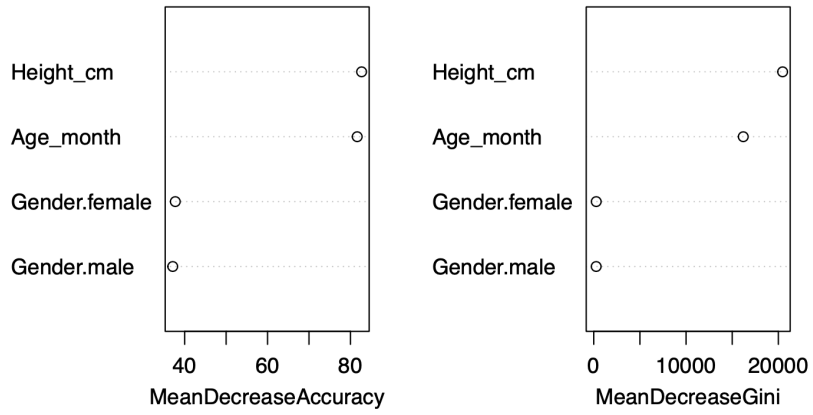

Figure 2: Variable Importance Plot: Random Forest model showing the relative importance of variables in predicting nutritional status.

The Support Vector Machine (SVM) model, utilizing a radial kernel and trained on the dataset, achieved an accuracy of 98.97%, indicating that it correctly predicted nutritional classes nearly 99% of the time. The Kappa value of 0.9832 suggests substantial agreement between the model's predictions and the true labels, significantly better than random chance. In terms of sensitivity, the SVM model excels, with a score of 0.9919 for 'severely stunted', and 0.9627 for 'stunted'. This indicates the model's ability to correctly identify true positives within each class. Additionally, the model demonstrates high specificity, with scores of 0.9972 for 'severely stunted', and 0.9975 for 'stunted', indicating its proficiency in identifying true negatives. The bar plot of the dataset shown in Figure 1, reveals an imbalance in the classes, which can lead to bias towards the majority class in models. The sensitivity analysis across models shows that Logistic Regression and single Decision Trees have lower sensitivity for the 'stunted' class, indicating a potential bias issue due to the class imbalance. In contrast, Bagging, Random Forest, and SVMs exhibit higher sensitivity for the 'stunted' class, suggesting they are more robust to class imbalances.

5. Conclusions and future work

In conclusion, this study aimed to predict nutritional status in children using various machine learning models. Logistic Regression, Decision Trees, Bagging, Random Forest, and SVMs were employed, with SVMs demonstrating the best performance. SVMs achieved an impressive accuracy of 98.93% and a Kappa value of 0.9832, indicating substantial agreement with the true labels. Additionally, SVMs exhibited high sensitivity and specificity across all classes, suggesting its robustness in classifying nutritional status.

Future work could focus on optimizing the hyperparameters in each of these models in order to get the best performance of each model. This was not done in this case due to the fact fine tuning SVMs using cross validation was taking so long to compute (a high computational cost we could not afford) and because we couldn’t optimize it, we decided against optimizing the rest of the models to avoid comparing optimized models against those that aren’t. Additionally, exploring advanced techniques to address class imbalance, such as oversampling or synthetic data generation, could enhance the performance of models like Logistic Regression and Decision Trees. Replicating or synthesizing data points from minority classes to increase their representation in the training set to counter bias and improve decision boundaries using techniques such as Random Oversampling and SMOTE (Synthetic Minority Oversampling Technique. Furthermore, incorporating additional features, such as dietary information or socioeconomic factors, could provide a more comprehensive understanding of the factors influencing childhood malnutrition. On another hand, understanding Gender Differences: Investigate why females have a lower likelihood of being classified as 'severely stunted' compared to males for decision-making and lead to more targeted interventions. Overall, this study lays a foundation for future research to continue improving the accuracy and effectiveness of machine learning models in predicting and addressing childhood malnutrition.

References

- UNICEF, 2013. Improving Child Nutrition The achievable imperative for global progress. United Nations Publications Sales No.: E.13. XX.4.
- World health statistics 2021: Monitoring health for the SDGs, sustainable development goals. World Health Organization. Retrieved from World Health Organization website: https://apps.who.int/iris/bitstream/handle/10665/342703/9789240027053-eng.pdf
- Haris, M. S., Anshori, M. and Khudori, A. N. (2023) "PREDICTION OF STUNTING PREVALENCE IN EAST JAVA PROVINCE WITH RANDOM FOREST ALGORITHM", Jurnal Teknik Informatika (Jutif). Purwokerto, 4(1), pp. 11-13. doi: 10.52436/1.jutif.2023.4.1.614.
- Ndagijimana, S., Kabano, I. H., Masabo, E., & Ntaganda, J. M. (2023) 'Prediction of Stunting Among Under-5 Children in Rwanda Using Machine Learning Techniques', Journal of Preventive Medicine and Public Health = Yebang Uihakhoe Chi, 56(1), pp. 41–49. https://doi.org/10.3961/jpmph.22.388
- Rahman, S. M. J., Ahmed, N. A. M. F., Abedin, M. M., Ahammed, B., Ali, M., et al. (2021) 'Investigate the risk factors of stunting, wasting, and underweight among under-five Bangladeshi children and its prediction based on machine learning approach', PLOS ONE, 16(6), e0253172.
- Mardani, R., Wetasin, K., & Suwanwaiphatthana, W. (2015). THE PREDICTING FACTORS AFFECTING THE OCCURRENCE OF STUNTING IN CHILDREN UNDER FIVE YEARS OF AGE. Jurnal Kesehatan Masyarakat, 11(1), 1-7. doi:https://doi.org/10.15294/kemas.v11i1.3927