Early Detection of Stunting in Indonesian Toddlers: A Machine Learning Approach

1st Herjanto Janawisuta
School of Computing
Telkom University
Bandung, Indonesia
herjanto@student.telkomuniversity.ac.id

2nd Putu Harry Gunawan
CoE HUMIC, School of Computing
Telkom University
Bandung, Indonesia
phgunawan@telkomuniversity.ac.id

3rd Indwiarti
CoE HUMIC, School of Computing
Telkom University
Bandung, Indonesia
indwiarti@telkomuniversity.ac.id

***Abstract*—Stunting, or growth retardation, is a condition in which a child's physical growth is hindered due to prolonged inadequate nutrition, often accompanied by illness. Between 2000 and 2020, the global prevalence of stunting decreased from approximately 30.3% to 22%. Despite a significant reduction, in 2019, an estimated 21.3% of children under five were still projected to experience stunting worldwide. One contributing factor to stunting in young children is the need for more parental supervision and attention. Therefore, it is crucial to conduct research to predict the likelihood of stunting and, ideally, prevent it in toddlers. The use of accurate prediction models can aid in identifying at-risk toddlers, enabling timely interventions. Random Forest is commonly chosen for classification tasks due to its widespread popularity and robust performance. On the other hand, Logistic Regression is often preferred as a straightforward classification algorithm, being the simplest among its counterparts. Its appeal lies in its ability to achieve high accuracy without extensive hyperparameter tuning efforts. This study utilized a stunting dataset from the Bojongsoang Health Center, comprising 6677 entries collected in August 2022. The research shows that the Random Forest model consistently outperforms Logistic Regression, even in the presence of class imbalance. While Logistic Regression demonstrates strong performance, the Random Forest model stands out with near-perfect accuracy, particularly in oversampling scenarios where Logistic Regression is 89.4%, and Random Forest is an impressive 99.8%. This underscores the robustness of Random Forest in situations where achieving a balanced representation of classes is feasible.**

***Keywords*— Stunting, Machine Learning, Random Forest, Logistic regression**

I. Introduction

Stunting, or growth retardation, is caused by a lack of adequate nutrition over a long period and is often accompanied by illness [1]. One of the significant causes is malnutrition in children. Almost half of the deaths of children in developing countries are directly or indirectly related to malnutrition [2]. Malnutrition has been proven to be one of the significant problems that are now or concurrently responsible for more than half of all deaths worldwide, particularly among children under the age of five [3].

In 2019, about 21.3% of children under the age of five were estimated to experience stunting worldwide [4]. According to WHO, there are 144 million children under five years old who suffer from stunting. Globally, 2.6 million children die each year due to malnutrition [5]. A lack of supervision and attention from parents to their children is one of the reasons for the occurrence of stunting in children under age. Therefore, research is needed to predict the likelihood of stunting and hopefully prevent stunting in toddlers.

Furthermore, developing accurate and reliable prediction models to identify toddlers at high risk of stunting is also essential to this research. Using data related to stunting status, this research can create a predictive tool that can provide valuable information about the likelihood of stunting in each individual.

A similar research on the use of Machine Learning in the problem of stunting in East Java in 2023 has been conducted by M. Syauqi Haris, Mochammad Anshori, and Ahsanun Naseh Khudori. They used two machine learning methods: random forest (RF) and multilinear regression (MLR). They found that the Multi Linear Regression model provides the best prediction accuracy for stunting in children in East Javaprovince [6].

Other research on stunting in toddlers has also been shown in the city of East Aceh in 2022 by Eva Darnila, Maryana, Khalid Mawardi, Marzuki Sinambela, and Iwan Pahendra. The results of their research show that the machine learning classification algorithm considered by Random Forest can effectively predict stunting status in the East Aceh administrative area [2].

An alternative perspective reveals another example of a deep learning approach to predicting malnutrition status in Bangladesh by Md Mehrab Shahriar, Mirza Shaheen Iqubal, Samrat Mitra, and Amit Kumar Das. This study demonstrates the ANN's superior accuracy in classifying wasting, underweight, and stunting conditions, showcasing its potential as a scientific tool for policymakers and clinicians dealing with child malnutrition in developing nations [7].

While the examples drawn from diverse studies offer valuable insights into the classification of stunting using machine learning, There are more things to explore and develop. Certain studies have utilized only a single machine learning model, limiting the opportunity for robust model comparisons. Additionally, some research needs a thorough Exploratory Data Analysis (EDA), a pivotal step in understanding and visualizing data patterns. Considering these factors, This study aims to contribute to existing knowledge by using two machine learning models to classify stunting. Moreover, the study endeavors to address the challenge of imbalanced data, a prevalent issue in stunting classification, further enhancing the model's ability to make accurate predictions under such conditions. Through this comprehensive approach, the study aspires to fill gaps in the

979-8-3503-6535-1/24/$31.00 ©2024 IEEE

current literature and advance the methodologies employed in stunting classification research.

Using an accurate prediction model to help identify toddlers at risk of stunting so that timely interventions can be made. Using the Random Forest and Logistic Regression methods, predictions can be made by considering various risk factors for stunting, such as nutritional status, environment, and socioeconomic factors. In addition, the predicted data can also be used to determine more effective and targeted stunting prevention policies and programs [8].

II. Methodology

A. Research Structure

This study initiates a literature review of relevant research to establish the background of the issue and to gather methodological insights for addressing the problem. Subsequently, data on toddlers will be collected from the Bojongsoang Community Health Center. Prior to implementing the selected method on the data, it will undergo cleaning and visualizing to ensure compatibility with processing. Following this, the data will be evaluated for balance, and if found to be imbalanced, resampling techniques will be applied to achieve balance. After pre-processing, logistic regression (LR) and random forest (RF) algorithms will be employed to classify the data. The performance of each method will be compared to draw conclusive insights. For further understanding, a flowchart depicting the research methodology will be presented in Figure1.

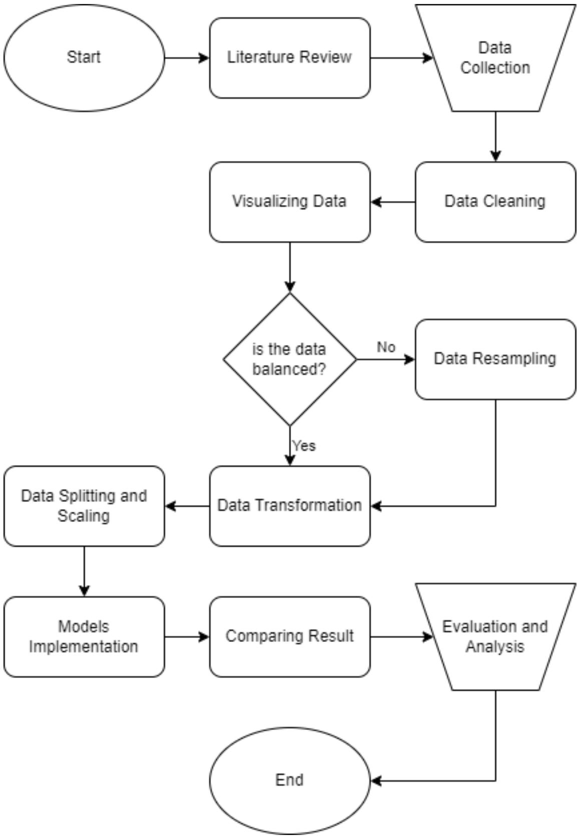

Fig. 1. Flowchart

B. Data Sourced

This study uses a stunting dataset in Table I collected from the Bojongsoang Health Center. The data covers measurements in August 2022 and consists of 6,677 entries. The information contained in this dataset includes various details, such as child identity, age, weight, height, and stunting status.

TABLE I
Child's Growth Table

Name	Gender	Age	Weight	Height	Status
F. F. R.	M	2,4	4,9	58	Normal
M. H. A. S.	M	58,4	17	109	Normal
K. A.	M	14,5	10	73	Stunting

Based on the sample dataset above, it can be seen that the Status column indicates whether the child is stunted or not.

C. Data Preprocessing

Many decisions affecting a model's predictive behavior are made during data preprocessing [9]. This research on predicting stunting in children employs comprehensive data processing techniques to ensure the reliability and accuracy of machine learning models. Libraries such as Pandas, NumPy, and TensorFlow are utilized to handle and process large datasets efficiently.

This dataset consists of parameters not needed in stunting prediction, such as name, National Identification Number, and parents' names. It starts with data cleaning, which involves removing missing values, correcting inconsistencies, and normalizing the data to a standard format. This process is essential to eliminate potential biases and errors in the dataset.

D. Handling Imbalance

After preparing the dataset, there was one problem in this dataset. Out of 6676 data, only 64 children are classified as stunted, and 6612 are expected, which means there is an imbalance in the dataset in Fig. 2. Given the nature of medical datasets, an uneven distribution of cases is expected, especially in conditions like stunting, where the number of affected individuals may be significantly lower than the unaffected ones.

Rebalancing the data sets is done either by under sampling, i.e., removing majority class instances, or oversampling, i.e., by adding new minority instances to the datasets [10]. In oversampling, instances of the minority class are artificially augmented in the dataset to achieve a more balanced representation. This approach helps prevent the model from developing a bias towards the majority class and enhances its capacity to generalize effectively on diverse, unseen data.

Additionally, undersampling is employed to address the class imbalance by reducing the number of instances in the majority class [11]. This technique ensures the training dataset maintains a harmonized distribution across different classes. By combining oversampling and undersampling, the model is better equipped to learn from the minority class while preventing the dominance of the majority class, ultimately contributing to improved model performance and generalization capabilities.

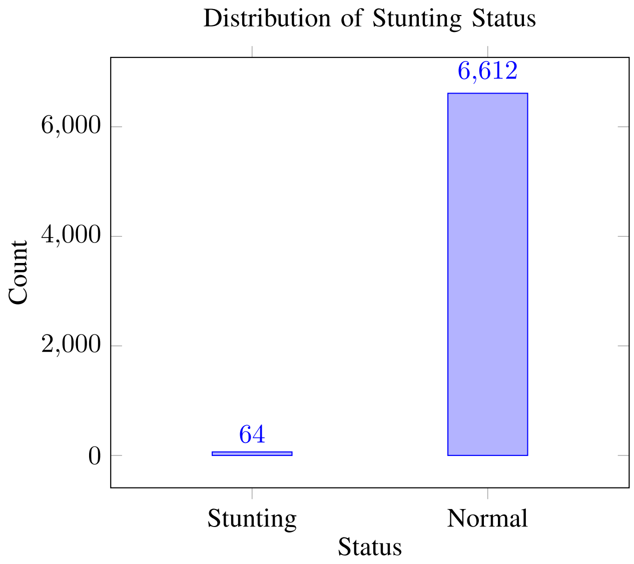

Fig. 2. Distribution of Stunting Status

E. Data Transformation

A critical step in preparing this study's dataset for stunting classification involves data transformation to enhance the model's capacity to learn from the available information effectively. This transformation involves a two-fold strategy: feature transformation and label encoding. The feature transformation leverages a ColumnTransformer with the OneHotEncoder to encode categorical features appropriately. Specifically, the first column, representing a categorical variable, undergoes one-hot encoding, preserving the remaining features unchanged. Simultaneously, label encoding is applied to the target variable using the LabelEncoder, ensuring a numerical representation of the stunting labels. To further mitigate class imbalance, an inversion is performed on the encoded labels, systematically converting instances of stunting (previously labeled as 1 to 0 and vice versa).

F. Data Splitting and Scaling

Data splitting is partitioning a dataset into distinct subsets, typically training and testing sets. The training set is employed to train the machine learning model, while the testing set is utilized to assess the model's performance on data it has not encountered during training. Splitting the data aims to determine how well the model generalizes to new, unseen examples. In this study, the dataset is divided into an 80-20 ratio, with the more significant portion designated for Training.

Following data splitting, the next step involves Feature Scaling. Feature scaling is a preprocessing step that standardizes or normalizes the numerical features of the dataset. It ensures that all features contribute equally to the learning process, preventing features with larger scales from dominating those with more minor scales. Standardization is used to rescale the input parameters to have mean to be zero (μ=0) and variance to be unit (σ=1) [12].

G. Classification Models Implemented

The process began with splitting the dataset into training and test sets. This separation allows for the evaluation of the model on unseen data, ensuring the robustness and generalizability of the predictions. Following the data split, a Logistic Regression and Random Forest model was instantiated.

1) Logistic Regression: In this study, the modeling approach employed Logistic Regression, a generalized linear model with a continuous-valued output. The sigmoid function is used in logistic regression to convert the pre-criterion value of linear regression into a binary value [13]. It is widely used for binary classification problems since the output is between 0 & 1, each serving as one of the two classes [14]. This method was chosen due to its effectiveness in handling datasets with a mix of categorical and continuous variables, as is often the case in medical and health-related research.

The model was then fitted to the training data, adjusting its parameters to minimize prediction errors. Providing the model to the training set allows it to 'learn' from the data, identifying patterns and relationships between the input features and the target variable, which in this case is the stunting status of the children.

2) Random Forest: Alongside Logistic Regression, the study also incorporated Random Forest, an ensemble learning method, for predicting stunting in children. Random Forest is known for its high accuracy, robustness, and ability to handle large datasets with numerous input variables. It is an ideal choice for complex predictive tasks like stunting prediction.

The Random Forest model operates by constructing many decision trees during training and outputting the class that is the mode of the classes of the individual trees [15]. This methodology improves the predictive accuracy and controls over-fitting, which is common in decision tree models.

The same dataset was used to implement the Random Forest model. Initially, the dataset was preprocessed using steps similar to the Logistic Regression model, ensuring consistency and reliability in comparing the two models' performances. The Random Forest model was then trained on the preprocessed training data with selected Hyperparameter (Criterion = entropy, Max depth = 25, Max features = log2, N estimators = 150) [16]. This step involves building multiple decision trees and combining their predictive capabilities to improve overall accuracy and handle the complexity of the data effectively.

III. Result Analysis and Comparison

Analyzing the results from the Logistic Regression and Random Forest models provided critical insights into the efficacy of each method in predicting stunting in children. Both models were evaluated based on key performance metrics such as accuracy, precision, recall, F1 score, and confusion Matrix. These metrics provide insights into how well the model can predict stunting in children and highlight areas where the model may need further refinement or adjustment.

In comparing the two models, attention was paid to their ability to handle the complexity of the dataset, including the handling of imbalanced classes and the interpretation of the results.

The study investigated the impact of undersampling and oversampling techniques on the performance of Logistic Regression and Random Forest models in the presence of imbalanced datasets. The dataset exhibited significant class imbalance, prompting exploring various sampling strategies. Table II presents the accuracy results for the Logistic Regression model, highlighting its performance in predicting stunting. Similarly, Table III provides the accuracy results for the Random Forest model, offering a comprehensive view of its predictive capabilities. These tables serve as valuable references for understanding the quantitative outcomes of each model's performance.

TABLE II
LOGISTIC REGRESSION MODEL ACCURACY RESULTS

Undersampling	**Oversampling**
0.7692	0.8877
0.8077	0.8941
0.8077	0.8972
0.8077	0.8964
0.6923	0.9009
0.7308	0.8915
0.7308	0.8972
0.7692	0.8911
0.8077	0.8922
0.8462	0.8888

TABLE III
RANDOM FOREST MODEL ACCURACY RESULTS

Undersampling	**Oversampling**
0.8846	0.9992
0.8846	0.9974
0.8462	0.9985
0.7692	0.9989
0.8077	0.9985
0.7692	0.9977
0.8462	0.9974
0.8462	0.9981
0.7692	0.9992
0.7308	0.9981

Under undersampling, the Logistic Regression model consistently achieved an average accuracy of 77.7%, while the Random Forest model outperformed with an average accuracy of 81.5%. The undersampling technique, aimed at addressing the class imbalance, improved the models' ability to generalize to the minority class. Notably, the Random Forest model consistently surpassed Logistic Regression in the undersampled scenarios, highlighting the efficacy of ensemble methods in handling imbalanced data.

Conversely, oversampling yielded more varied results. Logistic Regression and Random Forest models achieved higher average accuracies, with Logistic Regression at 89.4% and Random Forest at an impressive 99.8%. Oversampling facilitated improved classification performance by increasing the representation of the minority class. Notably, the Random Forest model showcased near-perfect accuracy, underscoring its robustness in handling imbalanced datasets by integrating multiple decision trees. This was supported by the common occurrence of previous research that obtained high accuracy by using the random forest method for classification. For instance, a study about Disease Prediction by A.N.V.K Swarupa achieved a 93% accuracy score using Random Forest Algorithm. From "Disease prediction: Smart disease prediction system using random forest algorithm," by A.N.V.K Swarupa, V. H. Sree, S. Nookambika, Y. K. Sai Kishore, and U. R. Teja, Result and discussions section, Table IV [15].

Comparing the two sampling strategies, oversampling generally led to higher accuracies than undersampling. That implies for this specific dataset, addressing class imbalance by augmenting the representation of the minority class is more effective than reducing the majority class. The choice between Logistic Regression and Random Forest depends on task-specific requirements, with Random Forest demonstrating superior performance in most instances.

Further detailed information is provided in Table IV, which encompasses all resampling methods with 100 iterations, including no resampling, employed in this study for each classification model, along with a focus on metrics such as F1-Score. F1 0 corresponds to class 0 (normal), while F1 1 pertains to class 1 (stunting). The table reveals noteworthy insights, particularly in cases where no resampling method was applied. Both algorithms encountered challenges in predicting the F1 score for class 1 in this scenario. Logistic Regression achieved a mere 9.8%, while Random Forest demonstrated a higher, albeit still challenging, F1-score of 33.6%. These findings underscore the importance of resampling strategies to enhance the models' performance, especially when dealing with imbalanced classes, as evident from the comparative analysis in Table IV.

These findings emphasize the importance of considering both the choice of sampling strategy and the algorithm when dealing with imbalanced datasets. The results provide valuable insights for practitioners, guiding the selection of appropriate techniques to enhance the performance of classification models in the presence of class imbalance.

TABLE IV
OVERALL AVERAGE METRICS FOR THE CLASSIFIERS

Resampling	**Logistic Regression**			**Random Forest**		
Method	**ACC**	**F1 0**	**F1 1**	**ACC**	**F1 0**	**F1 1**
None	0.991	0.995	0.098	0.992	0.996	0.336
Undersampling	0.777	0.771	0.774	0.815	0.821	0.804
Oversampling	0.894	0.892	0.895	0.998	0.998	0.998

IV. CONCLUSION

In conclusion, this research explains how class imbalance, sampling techniques, and algorithm choices are interconnected. Undersampling and oversampling each play a crucial role in mitigating the challenges posed by imbalanced datasets, but their impact varies depending on the specific characteristics of the data at hand.

The consistently higher performance of the Random Forest model compared to Logistic Regression suggests that ensemble methods, with their ability to capture complex relationships in data, are well-suited for imbalanced datasets. The near-perfect accuracy of Random Forest in the oversampling scenarios, where Logistic Regression achieved 89.4% and Random Forest at an impressive 99.8% underscores its robustness in situations where achieving a balance in class representation is possible.

Practitioners confronted with imbalanced datasets can leverage these findings to make informed decisions about sampling strategies and algorithmic choices. The study provides practical guidance for improving classification model performance in real-world scenarios where class imbalance is a prevalent challenge. Future research may delve deeper into understanding the nuanced dynamics between class imbalance and various machine learning techniques, further refining strategies for handling imbalanced datasets in diverse domains.

References

[1] M. F. Rizal and E. van Doorslaer, "Explaining the fall of socioeconomic inequality in childhood stunting in indonesia," *SSM - Population Health*, vol. 9, p. 100469, 2019. [Online]. Available: https://doi.org/10.1016/j.ssmph.2019.100469

[2] E. Darnila, Maryana, K. Mawardi, M. Sinambela, and I. Pahendra, "Supervised models to predict the stunting in east aceh," *International Journal of Engineering Science & InformationTechnology*, vol. 2, no. 3, pp. 34–39, 2022. [Online]. Available: https://doi.org/10.52088/ijesty.v1i4.280

[3] R. Joseph, V. Sawant, S. Shenai, M. Paryani, and G. Patil, "Machine learning based factors affecting malnutrition and anemia among children in india," in *2022 6th International Conference on Intelligent Computing and Control Systems (ICICCS)*. IEEE, 2022. [Online]. Available: https://ieeexplore.ieee.org/document/9788386

[4] O. N. Chilyabanyama, R. Chilengi, M. Simuyandi, C. C. Chisenga, M. Chirwa, K. Hamusonde, R. K. Saroj, N. T. Iqbal, I. Ngaruye, and S. Bosomprah, "Performance of machine learning classifiers in classifying stunting among under-five children in zambia," *Children*, vol. 9, p. 1082, 2022. [Online]. Available: https://www.mdpi.com/2227-9067/9/7/1082

[5] S. M. J. Rahman, N. A. M. F. Ahmed, M. M. Abedin, B. Ahammed, M. Ali, M. J. Rahman *et al.*, "Investigate the risk factors of stunting wasting and underweight among under-five bangladeshi children and its prediction based on machine learning approach," *PLoS ONE*, vol. 16, no. 6, p. e0253172, 2021. [Online]. Available: https://doi.org/10.1371/journal.pone.0253172

[6] M. S. Haris, M. Anshori, and A. N. Khudori, "Prediction of stunting prevalence in east java province with random forest algorithm," *Jurnal Teknik Informatika (JUTIF)*, vol. 4, no. 1, pp. 11–13, 2023. [Online]. Available: https://doi.org/10.20884/1.jutif.2023.4.1.614

[7] M. M. S. Mirza, S. Iqubal, S. Mitra, and A. K. Das, "A deep learning approach to predict malnutrition status of 0-59 month's older children in bangladesh," in *2019 IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology (IAICT)*. Dhaka, Bangladesh: IEEE, 2019.

[8] M. Ohyver, J. V. Moniaga, K. R. Yunidwic, and M. I. Setiawan, "Logistic regression and growth charts to determine children," *Procedia Computer Science*, vol. 116, pp. 232–241, 2017. [Online]. Available: https://www.elsevier.com/locate/procedia

[9] C. V. Gonzalez Zelaya, "Towards explaining the effects of data preprocessing on machine learning," in *2019 IEEE 35th International Conference on Data Engineering (ICDE)*. IEEE, 2019, pp. 2086–2090. [Online]. Available: https://ieeexplore.ieee.org/document/8731532/

[10] A. Gosain and S. Sardana, "Handling class imbalance problem using oversampling techniques: A review," in *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*. IEEE, 2017, pp. 79–85. [Online]. Available: http://ieeexplore.ieee.org/document/8125820/

[11] S. Yadav and G. P. Bhole, "Handling imbalanced dataset classification in machine learning," in *2020 IEEE Pune Section International Conference (PuneCon)*. IEEE, 2020, pp. 38–43. [Online]. Available: https://ieeexplore.ieee.org/document/9362471/

[12] P. Sharma and J. Singh, "Machine learning based effort estimation using standardization," in *2018 International Conference on Computing, Power and Communication Technologies (GUCON)*. IEEE, 2018, pp. 716–720. [Online]. Available: https://ieeexplore.ieee.org/document/8674908/

[13] W. Li, C. Li, and L. Jiang, "Learning from crowds with robust logistic regression," *Information Sciences*, vol. 639, p. 119010, 2023. [Online]. Available: https://www.elsevier.com/locate/ins

[14] S. Jain, A. A. Khan, T. Khanam, and A. J. Abedi, "Efficient machine learning for malnutrition prediction among under-five children in india," in *2022 IEEE Delhi Section Conference (DELCON)*. IEEE, 2022. [Online]. Available: https://doi.org/10.1109/DELCON54057.2022.9753080

[15] A. N. V. K. Swarupa, V. H. Sree, S. Nookambika, Y. K. Sai Kishore, and U. R. Teja, "Disease prediction: Smart disease prediction system using random forest algorithm," in *2021 IEEE International Conference on Intelligent Systems Smart and Green Technologies (ICISSGT)*. IEEE, 2021. [Online]. Available: https://doi.org/10.1109/ICISSGT52025.2021.00021

[16] R. Mohammed, J. Rawashdeh, and M. Abdullah, "Machine learning with oversampling and undersampling techniques: Overview study and experimental results," in *2020 11th International Conference on Information and Communication Systems (ICICS)*. IEEE, 2020, pp. 243–248. [Online]. Available: https://ieeexplore.ieee.org/document/9078901/