Machine Learning Classification Analysis for Proactive Prevention of Child Stunting in Bojongsoang: A Comparative Study

1st Caesar Fannany
School of Computing
Telkom University
Bandung, Indonesia
caesarfanany@std.telkomunversity.ac.id

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

3rd Narita Aquarini
École Doctorale Science Economics
Université de Poitiers Intervenant Finance
La Rochelle, France
aquarinin@excelia-group.com

***Abstract*—Stunting is a health condition that needs attention from the Indonesian government. According to the World Health Organization (WHO), the prevalence of stunting in a country should be below 20%, but in Indonesia, the prevalence of stunting is 21.6% as of 2023. An unbalanced dietary intake is one of the factors influencing the prevalence of stunting. Children with stunting may experience disruptions in growth intelligence and have a higher vulnerability to diseases. Healthcare professionals and researchers need to take appropriate preventive measures to reduce stunting occurrences. In efforts to reduce the incidence of stunting, concrete steps are needed to identify and predict stunting conditions in children using machine learning. Predictions facilitated by machine learning can be executed more efficiently, reducing the need for manual computation. This study employs three different methods to assess the performance of each method in predicting stunting cases. The machine learning methods used in this analysis include Logistic Regression, Random Forest, and Naïve Bayes. These three methods have already been proven in classifying stunting in other research studies. Imbalanced data will affect accuracy and F1-score macro. Oversampling is used as a method to avoid bias in the model. Imbalanced data yields accuracy results of 99.25% for Logistic Regression with an F1-score of 46.80%, 99.25% for Random Forest with an F1-score of 41.76%, and 94.5% for Naıve Bayes with an F1-score of 37.42%. Balanced data results in increased F1-scores, namely 52.73% for Logistic Regression, 65.44% for Random Forest, and 60.12% for Naıve Bayes.**

***Index Terms*—Stunting, Machine Learning, Logistic Regression, Random Forest, Naïve Bayes**

I. Introduction

Health is one of the most crucial sectors in Indonesia that requires improvement, particularly addressing issues such as low birth weight or what is commonly referred to as stunting. Stunting is a disturbance in a child's growth and development due to chronic nutritional deficiencies and recurrent infections, characterized by their height falling below the standard (WHO, 2015). The Ministry of Health data reveals the Indonesian Nutrition Status Survey (SSGI) results, indicating a stunting prevalence of 24.4% in 2021, which decreased to 21.6% in 2023. According to the World Health Organization (WHO), the prevalence of stunting should ideally be below 20%, highlighting that Indonesia's figures remain high.

An unbalanced dietary intake is one of the factors influencing the prevalence of stunting. Parents frequently face difficulties managing their children's food intake, especially when they lack adequate nutrition and health knowledge [1]. Stunting limits their ability to grow physically makes them more susceptible to illness, and interferes with their cognitive development [2]. In efforts to reduce stunting occurrences, concrete steps are essential to identify and predict stunting conditions in children, and machine learning emerges as a potentially effective tool. Predictions facilitated by machine learning can be executed more efficiently, reducing the need for manual computation. Machine Learning is a subfield of artificial intelligence that has emerged from pre-existing computing methods. Its goal is to emulate human intellect through environmental learning. [3].

In 2021, Aditya Yudha Perdana, et al.[4] conducted research using the random forest method, achieving an accuracy of 97.82% after performing k-fold cross-validation. Subsequently, Retno Kusumaningrum et al.[5] conducted research on preventing stunting using Logistic Regression, with an average accuracy of 78%, and Naïve Bayes, with an average accuracy of 61.00%. In 2022, Obvious Nchimunya Chilyabanyama, et al. [6] conducted research on stunting cases in Zambia using the random forest and logistic regression methods. The accuracy obtained using logistic regression was 45.92%, and for random forest, it was 61.62%.

This research aims to analyze and compare three methods of Logistic Regression, Random Forest, and Naïve Bayes in predicting child stunting in the Bojongsoang area. This study seeks to better understand each method's capabilities in stunting prediction. Previous researchers did not employ oversampling techniques in their studies, this research utilizes oversampling to address the issue of imbalanced data and how it affects accuracy. Through the evaluation of each method, it is anticipated that the most effective and accurate approach to predicting stunting datasets will be identified. This research, in turn, can offer valuable information for healthcare professionals and researchers to take appropriate preventive measures to reduce stunting occurrences.

II. Methods

A. Design Research

This research begins with a literature review related to the issues addressed in this study. The study is then continued by collecting a dataset from the Bojongsoang Health Center. The obtained data needs to be thoroughly understood. Afterward, features in the data are selected for use in the subsequent stages. The selected features will undergo data visualization to understand the conditions present in the dataset. The data will go through a data cleaning stage to reduce bias in the

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

dataset. The cleaned data will then undergo data conversion into compatible types. It is necessary to check whether the data is balanced; if not, data balancing will be conducted before applying the methods to the data. Once the data is balanced, the Logistic Regression, Random Forest, and Naïve Bayes methods will be used to classify the data. The results of each method will be compared to conclude. Fig. 1 represents the research flowchart.

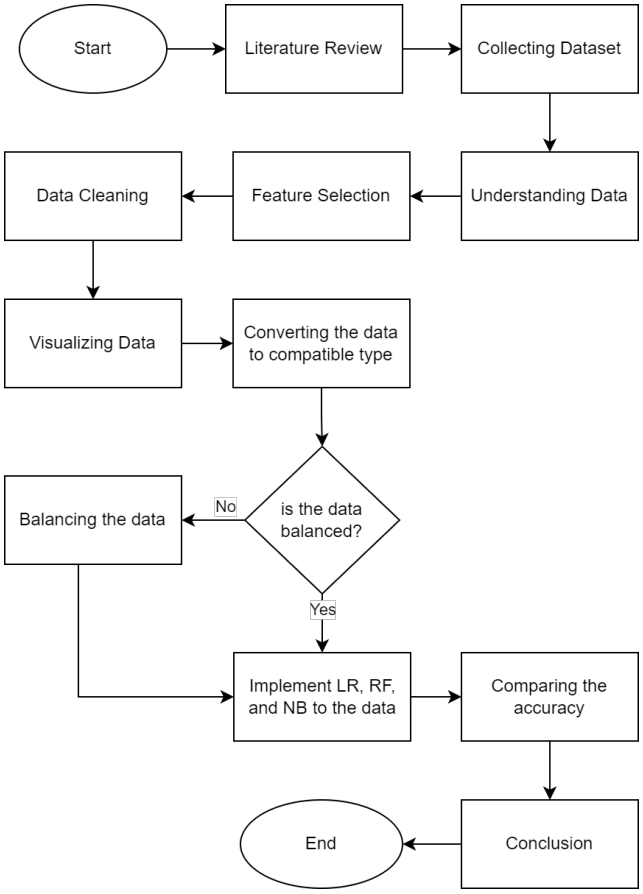

Fig. 1. Research Flowchart

B. Logistic Regression

Logistic Regression is a statistical method for analyzing data that describes the relationship between one or more predictor variables and a response variable [7]. Logistic Regression is a statistical method for predicting probabilities of categorical and classification outcomes. In this method, the sigmoid or Logistic function combines input values linearly and coefficients are assigned to predict the probability of outcomes. The sigmoid function assumes the most likely data and the predicted probability ranges from 0 to 1.[8].

Mathematically, the Logistic Regression algorithm is an extension of the Linear Regression algorithm, as illustrated below [9]:

Simple Linear Regression :

$$y = \alpha + \beta x \tag{1}$$

$$g(x) = \alpha + \beta x \tag{2}$$

Multiple Linear Regression :

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n \tag{3}$$

$$g(X) = a + \beta X \tag{4}$$

Logistic Regression :

$$g(X) = \text{sigmoid}(\alpha + \beta) \tag{5}$$

$$\text{sigmoid}(x) = \frac{1}{1 + \exp(-x)} \tag{6}$$

where :

$y =$ Response variable or outcome variable
$\alpha =$ Constant
$\beta =$ Regression coefficient (slope)
$X =$ Predictor variable or Independent variable

C. Random Forest

The random forest method employs a considerable number of decision trees that work together to form ensembles. Based on a random procedure, each tree casts a vote for the predicted class, and the class with the most votes is the model's prediction[5]. The Random Forest method has consistently outperformed the Single Decision Tree classifier [10]. Random forest can prevent overfitting and improve the accuracy of model predictions.[11].

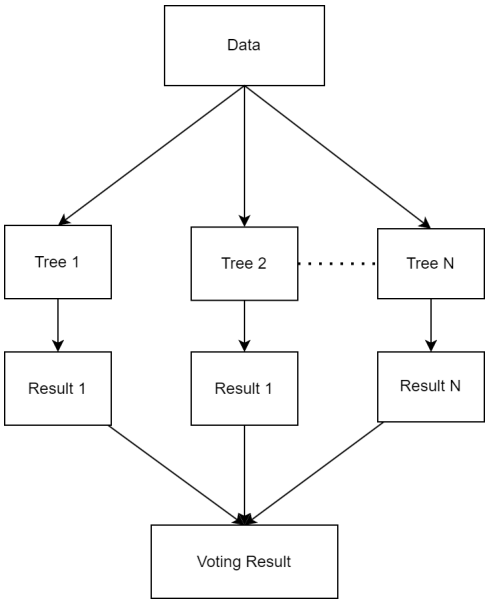

Fig. 2. Description of the Random Forest Algorithm

Fig. 2 illustrates random forest algorithm. The number of trees influences the effectiveness of the Random Forest it constructs. As the number of trees increases, the model's accuracy improves. The majority vote from the ensemble of trees determines the classification decision in Random Forest. The final prediction is based on the most frequently occurring result among the individual trees.

Random Forest begins by calculating the entropy value as a measure of attribute impurity. To compute the entropy value, the formula used is as follows in Equation (7) [12].

$$\text{Entropy} = -\sum_i p_i \log(p_i) \tag{7}$$

$$\text{IG} = E(\text{parent}) - \sum_i w_i \cdot E(\text{child}_i) \tag{8}$$

Entropy represents the likelihood of class I, whereas IG (Information Gain) is calculated by subtracting the parent's entropy from the leaf node's (child's) entropy. A random forest's outcome is determined by a majority voting process, in which the class with the highest count prevails[13].

D. Naïve Bayes

Naive Bayes is a data classification technique that employs the Bayes theorem[14]. The naive Bayes model used is the "independent feature model." Strong independence in features, as defined by Bayes, refers to a feature in the data that is unrelated to the presence or absence of other features in the same data[15].

Bayesian prediction based on Bayes' theorem has a general formula as follows:

$$P(H|E) = \frac{P(E|H) \cdot P(H)}{P(E)} \tag{9}$$

The explanation of equation (9) is as follows [15]:

- The conditional final probability (Posterior) of a hypothesis H occurring given the evidence E has occurred. The posterior formula is denoted as $P(H|E)$.
- The probability of evidence E occurring, influencing the hypothesis H (Likelihood). The likelihood formula is denoted as $P(E|H)$.
- The initial probability (Prior) of the hypothesis H occurring without considering any evidence. Prior is denoted as $P(H)$.
- $P(E)$ is the initial probability of evidence E occurring without considering other hypotheses/evidence.

This equation forms the fundamental idea of Bayes' rule, stating that the outcome of a hypothesis or event (H) can be estimated based on observed evidence (E).

III. Results and Discussions

A. Exploratory Data Analysis

The data used in this research was collected from the Bojongsoang Community Health Center, comprising 6,677 recorded children's data. The dataset contains personal information as well as information on the health status of toddlers. The selected datasets are Gender, Weight, Height, weight/age, height/age, weight/height, and Age.

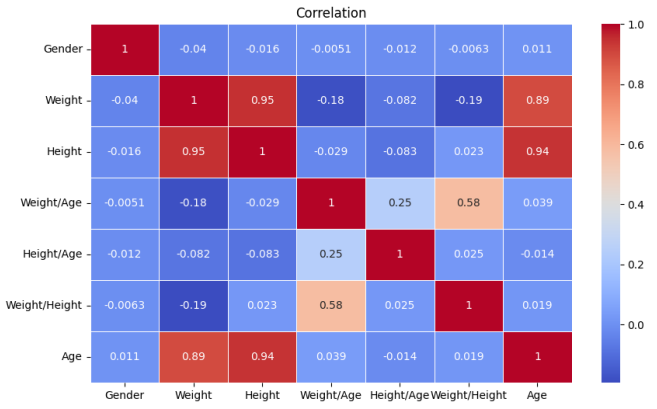

Fig. 3. Correlation data

Fig. 3 illustrates that the highest correlation is observed between weight and height, with a correlation coefficient of 0.95. Additionally, there is a strong correlation between age and height, with a coefficient of 0.94, and a correlation between age and weight, with a coefficient of 0.89. Consequently, the target variable height/age exhibits a high correlation with weight/age, amounting to 0.25.

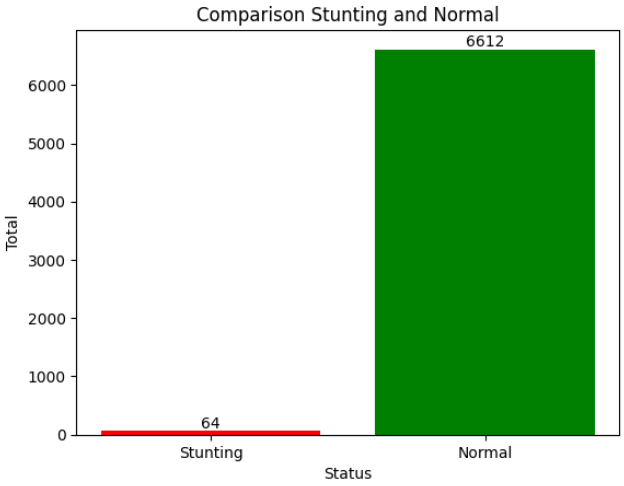

Fig. 4. Comparison of normal and stunting quantity

Fig. 4 illustrates that the stunting data consists of only 64 records, whereas there are 6,612 records for normal children. This information is extracted from the height/age feature column. The data indicates a quantity disparity between the two different classes. As a result, the dataset is considered imbalanced.

B. Preprocessing

The collected stunting dataset comprises 6,677 records with 33 feature columns. The data contain missing values in both the feature columns and individual records. Records with missing values cannot be utilized in the classification process, necessitating their removal to eliminate these gaps.

Unnecessary columns in the data will be deleted during the classification process. In columns such as gender, weight/age, height/age, and weight/height, the data type is a string, but the values in these columns are categorical. Therefore, the gender column will be converted, with "Male" transformed to 1 and "Female" to 2. For the weight/age column, "risk of having overnutrition" will be assigned a value of 1, "normal" 2, "underweight" 3, and "severely underweight" 4. In the height/age column, "normal" will be assigned 1, "short" 2, and "very short" 3. Regarding the weight/height column, "obesity" will be assigned 1, "overnutrition" 2, "risk of having overnutrition" 3, "good nutrition" 4, "undernutrition" 5, and "bad nutrition" 6. The target class in the height/age column will categorize "Stunting" with values 2 and 3, while "Normal" will have a value of 1.

Based on Fig.4, there is still an imbalance in the "Stunting" and "Normal" categories within the data. This complicates ordinary machine learning methods because they will be biased towards majority classes. In this case, they will demonstrate excellent accuracy on majority classes while performing poorly on minority classes[16]. To address this problem, the minority data must be increased in quantity through oversampling. Oversampling the dataset can be achieved by generating random data based on the distribution of the minority data. In the generation process, upper and lower

bounds are required as constraints to guide the generation process.

In Fig. 5, the boxplot is utilized to determine the first quartile as the lower limit for random sampling and the third quartile as the upper limit for random sampling. The generated samples will align with the stunting classification. The minority data will be generated to match the majority data, specifically in 6548 iterations. The generated iteration data will not exceed the upper limit or fall below the lower limit. The total stunting data will then amount to 6612 records, while the non-stunting data will also consist of 6612 records.

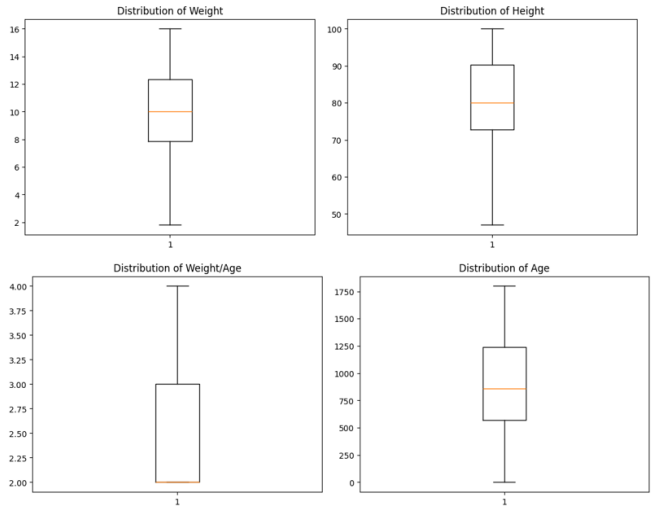

Fig. 5. Distribution of Stunting data

C. Logistic Regression Implementation

Logistic regression is employed for stunting classification in this research. The implementation of logistic regression utilizes the Python library called sci-kit-learn. The dataset is divided into training and testing sets with a ratio of 60:40. The logistic regression result is shown using a confusion matrix displayed in Table I.

TABLE I
CONFUSION MATRIX FOR LOGISTIC REGRESSION METHOD

		Predicted	
		Stunting	**Normal**
Actual	**Stunting**	2633	652
	Normal	725	2595

According to Table I, the results indicate 2633 correct predictions for stunting and 2595 correct predictions for normal growth.

D. Random Forest Implementation

In addition to Logistic Regression, Random Forest is employed for stunting classification. The dataset is divided using the same ratio. The implementation of Random Forest utilizes the same library as before, generating 100 decision trees to determine classifications for the data. The results of Random Forest in the confusion matrix can be observed in Table II.

According to Table II, the results indicate 3240 correct predictions for stunting and 3247 correct predictions for normal growth.

TABLE II
CONFUSION MATRIX FOR RANDOM FOREST METHOD

		Predicted	
		Stunting	**Normal**
Actual	**Stunting**	3240	45
	Normal	73	3247

E. Naïve Bayes Implementation

After applying two methods, the Naïve Bayes method is also employed in this research for stunting classification. The implementation of Naïve Bayes utilizes the same library as before. The dataset is divided with the same ratio as previously. The results of Naïve Bayes are depicted in the confusion matrix, which can be seen in Table III.

TABLE III
CONFUSION MATRIX FOR NAÏVE BAYES METHOD

		Predicted	
		Stunting	**Normal**
Actual	**Stunting**	2485	734
	Normal	12	3299

According to Table III, the results indicate 2485 correct predictions for stunting and 3299 correct predictions for normal growth.

F. Accuracy Comparison

The methods employed for stunting classification yield different performances. The results will compare imbalanced data and balanced data. The experiment will be conducted over ten iterations. The results for the imbalanced data can be observed in Table IV.

TABLE IV
ACCURACY AND F-1 SCORE FOR IMBALANCED DATASET

Method	**Accuracy**	**F-1 Score Macro**
Logistic Regression	99.25%	46.80%
Random Forest	99.25%	41.76%
Naïve Bayes	94.5%	37.42%

Based on Table IV, the obtained results show a very high accuracy rate of up to 99%. However, the F1-score macro does not exceed 50%. Therefore, the classification results are ineffective due to the limited amount of Stunting data.

TABLE V
ACCURACY AND F-1 SCORE FOR BALANCED DATASET

Method	**Accuracy**	**F-1 Score Macro**
Logistic Regression	79.87%	52.73%
Random Forest	98.12%	65.44%
Naïve Bayes	86.66%	60.12%

Table V indicates that the accuracy results obtained are not higher than those of the imbalanced data. However, the F1-Score macro shows an improvement, exceeding 50% for each method. Therefore, the classification performed is more effective when using balanced data.

IV. Conclusion

This study demonstrates that imbalanced data significantly influences the performance of classification. Utilizing an imbalanced dataset leads to high accuracy but low F1-Score macro results. Conversely, when using a balanced dataset, the accuracy results may not surpass those of the imbalanced data, but the F1-Score macro indicates an improvement. The high accuracy obtained from imbalanced data can be attributed to the classification's compelling performance in the majority class, but it does not perform well in the minority class.

In Table IV, the results with imbalanced data show relatively high accuracy: LR 99.25% with F1-Score macro 46.80%, RF 99.25% with F1-Score macro 41.76%, and NB 94.5% with F1-Score macro 37.42%. However, in Table V, the balanced data results in lower accuracy: LR 79.87% with F1-Score macro 52.73%, RF 98.12% with F1-Score macro 65.44%, and NB 86.66% with F1-Score macro 60.12%.

The findings of this study indicate that using the Random Forest method yields the highest accuracy at 98.12% with an F1-Score macro of 65.44%. Therefore, Random Forest proves to be the most suitable method for classifying stunting cases in this research. The results of this study enable researchers and healthcare professionals to make decisions and provide information widely to enhance community awareness and sentiment towards stunting.

Based on the findings of this research, it is advisable that the study be conducted in a wider region and that the methods used be more varied in determining the conditions of stunting in children to enhance a more in-depth analysis.

Reference

[1] M. Barri et al. "Aksi Cegah Stunting Melalui Aplikasi SAGITA: Status Gizi Balita". In: *JMM (Jurnal Masyarakat Mandiri)* 1116.7 (2 2023), pp. 1–7. ISSN: 2598-8158. DOI: 10.31764/jmm.v7i2.13231. URL: https://journal.ubaya.ac.id/index.php/jmm/article/view/13231.

[2] M. S. Haris, A. N. Khudori, and W. T. Kusuma. "Perbandingan Metode Supervised Machine Learning untuk Prediksi Prevalensi Stunting di Provinsi Jawa Timur". In: *J. Teknol. Inf. dan Ilmu Komput.* 9.7 (2022), p. 1571. DOI: 10.25126/jtiik.2022976744.

[3] I. El Naqa and M. J. Murphy. "Machine Learning in Radiation Oncology". In: *Machine Learning in Radiation Oncology*. Springer, 2015, pp. 3–11. DOI: 10.1007/978-3-319-18305-3.

[4] A. L. Perdana, R. Latuconsina, and A. Dinimaharawati. "Prediksi Stunting Pada Balita Dengan Algoritma Randim Forest". In: *e-Proceeding of Engineering* Volume Number.Issue Number (2021), Page Range. ISSN: 2355-9365.

[5] R Kusumaningrum et al. "Benchmarking of Multi-Class Algorithms for Classifying Documents Related to Stunting". In: *Applied Sciences (Switzerland)* 10.23 (2020), pp. 1–13. ISSN: 2076-3417. DOI: 10.3390/app10238621.

[6] Obvious Chilyabanyama, Roma Chilengi, Michelo Simuyandi, et al. "Performance of Machine Learning Classifiers in Classifying Stunting among Under-Five Children in Zambia". In: *Children* 9.7 (2022), p. 1082. ISSN: 2227-9067. DOI: 10.3390/children9071082. URL: https://www.mdpi.com/2227-9067/9/7/1082.

[7] A. Bimantara and T. A. Dina. "Klasifikasi Web Berbahaya Menggunakan Metode Logistic Regression". In: *Annu. Res. Semin.* 4.1 (2019), pp. 173–177.

[8] F. Handayani et al. "JEPIN (Jurnal Edukasi dan Penelitian Informatika) Komparasi Support Vector Machine, Logistic Regression Dan Artificial Neural Network dalam Prediksi Penyakit Jantung". In: *JEPIN* 7.3 (2021), pp. 329–334.

[9] A. Shiddicky. "Analisis Sentimen Masyarakat Terhadap Kebijakan Vaksinasi Covid-19 Pada Media Sosial Twitter Menggunakan Metode Logistic Regression". In: *J. Ekon.* 18.1 (2022), pp. 41–49. DOI: https://doi.org/10.37859/coscitech.v3i2.3836.

[10] J. Ali et al. "Random forests and decision trees". In: *IJCSI Int. J. Comput. Sci. Issues* 9.5 (2012), pp. 272–278.

[11] M. Minnoor and V. Baths. "Diagnosis of Breast Cancer Using Random Forests". In: *Procedia Comput. Sci.* 218.2022 (2023), pp. 429–437. DOI: 10.1016/j.procs.2023.01.025.

[12] L. Andiani and D. Palupi Rini. "Analisis Penyakit Jantung Menggunakan Metode KNN Dan Random Forest". In: *Pros. Annu. Res. Semin.* 5.1 (2019), pp. 978–979.

[13] Vega Herliansyah, Roswan Latuconsina, and Ashri Dinimaharawati. "Prediksi Stunting Pada Balita Dengan Menggunakan Algoritma Klasifikasi Naïve Bayes". In: *e-Proceeding Eng.* 8.5 (2021), p. 6642.

[14] APurnomo A Rozaq. "Classification of Stunting Status in Toddlers Using Naive Bayes Method in the City of Madiun Based on Website". In: *Jurnal Techno Nusa Mandiri* 19.2 (2022), pp. 69–76. ISSN: 1978-2136. DOI: 10.33480/techno.v19i2.3337.

[15] N. D. Prayoga, N. Hidayat, and R. K. Dewi. "Sistem Diagnosis Penyakit Hati Menggunakan Metode Naïve Bayes". In: *J. Pengemb. Teknol. Inf. dan Ilmu Komput.* 2.8 (2018), pp. 2666–2671.

[16] J. Tanha et al. "Boosting Methods for Multi-class Imbalanced Data Classification: An Experimental Review". In: *Journal of Big Data* 7.1 (2020). ISSN: 2196-1115. DOI: 10.1186/s40537-020-00349-y.