Predicting Stunting in Toddlers Using KNN and Naïve Bayes Methods

1st Galih Atha Fayi Khansa
School of Computing
Telkom University
kfayikhansa@student.telkomuniversity.ac.id

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

***Abstract*—Stunting poses a significant health challenge in Indonesia, with a prevalence rate reaching 21.6% in 2022, exceeding the WHO's tolerance limit for stunting. This study aims to identify stunting in toddlers using the K-Nearest Neighbor (KNN) and Naive Bayes (NB) algorithms. The KNN method is chosen for its effectiveness in utilizing the similarity between the attributes of new data and the training dataset, while Naive Bayes is selected for its probabilistic approach, addressing attributes related to stunting in toddlers. The dataset, obtained from the Bojongsoang Community Health Centre, underwent preprocessing to address initial imbalances such as converting the age format to month and removing an irrelevant column like address details. The attributes used in this study include age-in-months, height, weight, Z-Score Weight-for-Age, Z-Score Height-for-Age, and Z-Score Weight-for-Height. This research uses an SMOTE algorithm for handling imbalance data. In the SMOTE algorithm, imbalanced data can be managed by using undersampling and oversampling techniques. The research reveals a significant improvement in both KNN and NB models after applying the oversampling technique. Particularly, the KNN model demonstrates superior performance, increasing the F1-Score from 67.20% to 95.62%, with an accuracy of 95.67%. The Naive Bayes model also experiences enhancement, raising the F1-Score from 71.22% to 95.62%, with an accuracy of 94%. This study contributes to effective stunting classification methods. Leveraging direct measurements of height and weight, the research aids in stunting identification and proposes methods, including oversampling, to enhance classification accuracy KNN and Naive Bayes.**

***Index Terms*—Stunting, Machine Learning, Classification, K-Nearest Neighbors, Naive Bayes.**

I. Introduction

Stunting remains a critical issue affecting the growth and development of toddlers, posing significant challenges to their health and well-being. Defined as a condition where children under five years of age experience impaired growth and development due to chronic malnutrition, stunting continues to be a major concern in Indonesia [1]. Data from the Indonesian Nutritional Status Survey (SSGI) conducted by the Ministry of Health in 2022 indicates that the prevalence of stunting in Indonesia is still at 21.6% [2]. However, this figure still exceeds the maximum tolerance limit set by the World Health Organization (WHO), which establishes that the maximum tolerance for stunting is 20% of the total number of toddlers [3]. Factors contributing to this concerning statistic are multifaceted, including inadequate nutrition, limited access to sufficient food, poverty, unsatisfactory environmental sanitation, and inadequate healthcare accessibility [4].

In a study titled 'Government Efforts in Combating Stunting in the Bangka Belitung Islands Province,' it was reported that various government programs have been implemented to address the issue of stunting. These programs include the provision of Supplementary Feeding (PMT) for toddlers and pregnant women, Iron Tablets (TTD) for adolescent girls and pregnant women, increased coverage of complete basic immunization for infants and toddlers, vitamin A supplementation, and zinc supplementation for diarrhea cases, especially in pregnant women and toddlers [5]. However, these programs still require time to show tangible results.

Machine Learning (ML) is one of the applications of Artificial Intelligence (AI) that focuses on developing a system capable of self-learning without the need for repetitive programming. ML requires training data as part of the learning process before achieving specific criteria or performance using a set of training data or past experiences [6]. In recent years, machine learning approaches have made significant contributions to the diagnosis of breast cancer [7]. The capabilities of machine learning are notably effective in performing classification tasks.

Previous research, particularly the work conducted by Rizky and Agung in 2022, has delved into classification methods such as web-based Naive Bayes and K-Nearest Neighbor to assess the nutritional status of toddlers. Their findings indicate promising results, with an accuracy of 91.79% using the K-Nearest Neighbor method and 80.60% using Naive Bayes in classifying the nutritional status of toddlers [8]. Building upon previous research, this study is closely related to the work conducted by Rizky and Agung. However, the focus of this research is on classifying stunting in toddlers using the K-Nearest Neighbor (KNN) and Naive Bayes methods, leveraging data obtained from direct measurements of height and weight in toddlers.

In another research by Anis Zubair and Moch Muksin, the application of the Naive Bayes method for classifying nutritional status at Bromo Clinic in Malang was examined. The research aimed to comprehend nutritional status based on five categories, employing a data mining approach and implementing the Naive Bayes Classifier as the primary classification method. The results of the study demonstrated a remarkable accuracy level of 98 percent, underscoring the reliability of the Naive Bayes method in the context of nutritional status classification [9]. This reference makes a significant contribution to understanding the application of Naive Bayes in similar contexts and serves as a valuable foundation for further research, including the current study focusing on the classification of stunting in toddlers using a similar methodology.

Other related research by Sandra Yoseba Sibi and Anastasia Rita Widiarti focused on the classification of obesity levels using the K-Nearest Neighbor (KNN) algorithm. The research aimed to develop a simple system for assessing an individual's obesity status by incorporating essential information

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

such as weight and age. Experiments with various values of k in KNN and cross-validation tests indicated that the obesity classification model achieved a maximum accuracy of 79.96% [10]. From this study, it was concluded that KNN can be effectively utilized for the classification of obesity levels, where the attributes used include weight and age.

This study aims to contribute to the existing body of knowledge by focusing on the specific identification of stunting using the classification methods KNN and Naive Bayes. Additionally, a performance comparison between the two methods is conducted to determine which method is more accurate in classifying stunting in toddlers, whether it be KNN or Naive Bayes. Thus, this research is expected to provide deeper insights into the effectiveness and reliability of both classification methods in the context of stunting conditions in toddlers in Indonesia.

II. Methodology

A. Methodology Flow

The research methodology employed for classifying the stunting status of toddlers is outlined in Fig 1.

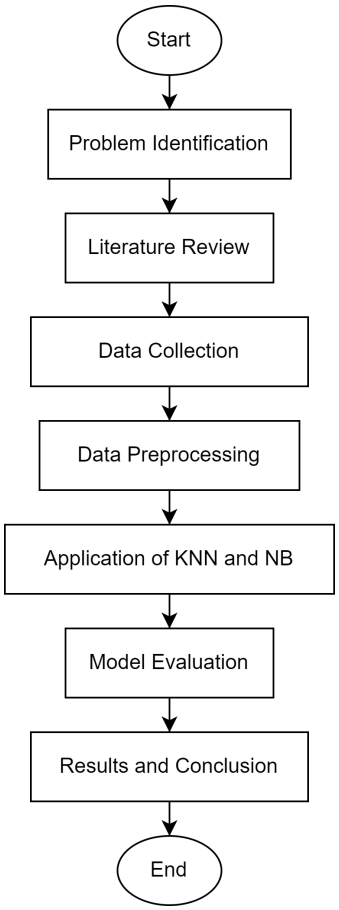

Fig. 1: Phases in Research

This research begins by identifying the problem of stunting in toddlers as the main focus. Subsequently, a comprehensive exploration of relevant literature is conducted, encompassing studies related to stunting in Indonesia, government programs, and classification methods. The purpose of this research is to find out the accuracy resulting from the KNN and Naive Bayes algorithms in predicting stunting in toddlers.

B. Stunting

Stunting is a chronic nutritional problem that affects the growth and development of children under five years old, characterized by shorter height than their age standard. In addition to resulting in short stature, stunting can cause physical and cognitive impairments that affect children's intelligence, work productivity, and a high risk of heart disease and diabetes mellitus [1].

According to the World Health Organization (WHO), a child is considered stunted when their body length or height falls more than two standard deviations (Z-Score ≤ -2 SD) below the established median length or height of the reference international population [11]. To calculate the Z-Score for Height-for-Age (H/A) explained in (1).

$$\text{ZS H/A} = \frac{\text{Child's Height} - \text{Median Height}}{\text{Standard Deviation of Height}} \quad (1)$$

Child's Height is the height of a child to be measured in the same unit as the Median Height, which represents the average height. Standard Deviation of Height is a measure of how spread out the heights are in the reference population.

C. Dataset

The data collection process involved obtaining a dataset of toddlers from the Bojongsoang Community Health Center. The dataset comprises twenty-eight columns, including Number, National Identity Number (NIN), Name, Gender, Birth Date, Birth Weight, Birth Height, Parents' Name, Province, City, District, Public Health Center Location, Village, Healthcare Center Name, Address, Age at Measurement, Measurement Date, Weight, Height, Mid Upper Arm Circumference (MUAC), Z-Score Height-for-Age (ZS HFA), and Z-Score Weight-for-Height (ZS WFH). The obtained dataset is substantial, encompassing a total of 6677 toddler records, as summarized in Table I.

Before performing the classification process with the KNN and Naive Bayes methods, the preprocessing stage is first carried out on the toddler data. This preprocessing process aims to prepare the data before entering the classification stage. The process begins by removing irrelevant columns such as identity and address information. Subsequently, the age data, originally presented in years, months, and days, is converted into numeric form, and the age in months is calculated. Additionally, handling is applied to missing or "-" character-containing data and normalization is performed on the TB/U column, where the "Normal" label is transformed to 1, while "Short" and "Very Short" labels are assigned a value of 2. Following that, the data is split into training and testing sets. Finally, oversampling is conducted on the imbalanced data using SMOTE. From the initial data consisting of 6677 toddlers, after preprocessing, 6578 toddler data remained that met the criteria and were consistent to be used in further analysis.

In the dataset, the attributes used in this study include age in months (Age_Months), height (Height), weight (Weight), Z-Score Weight-for-Age (ZS WFA), Z-Score Height-for-Age (ZS HFA), and Z-Score Weight-for-Height (ZS WFH). Subsequently, for data with normal status, there are 6534 instances, while only 54 toddlers fall into the stunting category. Table II provides a summary of the dataset used in the study.

A comparison of stunting and normal status in the data can be seen in Fig 2. From the entire dataset, the data labeled as

TABLE I: Toddler Dataset

No.	Gender	...	Age at Measure	Measurement Date	Height	...
1	M	...	0 Year-2 Month-14 Day	2022-08-02	58 cm	...
2	M	...	4 Year-10 Month-14 Day	2022-08-02	109 cm	...
3	F	...	0 Year-5 Month-5 Day	2022-08-02	68,5 cm	...
4	F	...	3 Year-9 Month-9 Day	2022-08-02	95 cm	...
5	M	...	4 Year-5 Month-19 Day	2022-08-02	106 cm	...
6	F	...	0 Year-8 Month-20 Day	2022-08-02	71,8 cm	...
7	F	...	4 Year-11 Month-16 Day	2022-08-02	107 cm	...
8	M	...	0 Year-8 Month-16 Day	2022-08-02	71 cm	...
9	F	...	2 Year-11 Month-5 Day	2022-08-02	95 cm	...
10	F	...	2 Year-10 Month-5 Day	2022-08-02	91,1 cm	...
...	...	...	...	...	...	...
6677	M	...	0 Year-0 Month-0 Day	2022-08-15	48 cm	...

[1]Source: Dataset obtained from Bojongsoang Community Health Center.

TABLE II: Cleaned Toddler Dataset

No.	Gender	Age_Months	Height	Weight	Result
1	M	27	90.4	12.9	1
2	M	28.0	80.0	9.0	2
3	M	39.0	89.0	14.8	2
4	M	28.0	90.4	12.9	1
5	F	52.0	105	17.0	1
6	F	53.0	101.0	15.0	1
7	M	30.0	84.7	12.0	2
8	M	37.0	86.7	14.5	2
9	M	21.0	82.4	10.2	1
10	F	40.0	80.0	10.0	2
...	...	...	...	...	...
6578	F	39.0	98.6	15.0	1

normal constitutes 99.2%, while the data labeled as stunting accounts for 0.8% of the total dataset.

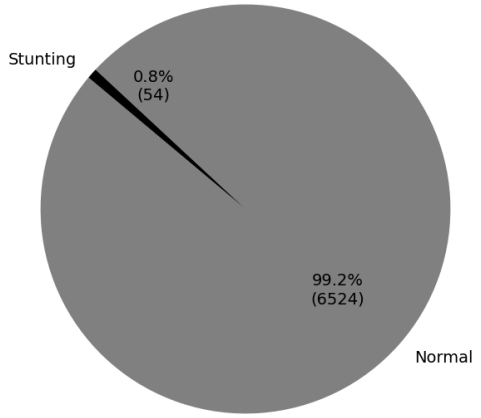

Fig. 2: Comparison of Stunting Status in the Data

D. Classification

Classification is a process of categorization used to determine the class of data with unknown class labels [12]. This process involves two main stages: the learning stage and the classification stage. In the learning stage, training data is analyzed using an appropriate classification algorithm. Subsequently, in the classification stage, class predictions are made for testing data by leveraging the outcomes of the learning phase. Through this classification process, unlabeled data can be categorized into the appropriate classes based on the analysis using a classification algorithm [13].

The obtained test results are compared with the ground truth values to generate a Confusion Matrix, providing insights into recall, accuracy, and precision. Accuracy is calculated by summing the correctly categorized data (TP and TN) and dividing it by the total number of data points, as outlined in (2).

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{Total Data}} \tag{2}$$

Precision is employed to measure the extent to which positive patterns are accurately predicted (TP) among the total positive prediction patterns (TP and FP), as indicated in (3).

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{3}$$

Meanwhile, Recall assesses the success of positive predictions by dividing TP by the sum of false negatives (FN) and TP, as explained in (4) [14].

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{4}$$

The F1-score, commonly referred to as the F-measure, is a measurement method in Machine Learning that represents the harmonic mean between the values of recall and precision. To calculate the F1-Score, can utilize (5).

$$F_1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5}$$

The formula for the F1-score involves dividing the product of precision and recall by the sum of precision and recall, and then multiplying the result by two[15].

E. K-Nearest Neighbors

The K-Nearest Neighbor is a data mining algorithm employed for classification by analyzing the proximity among K neighbors based on the attributes of training data. Consequently, when classifying new data, a comparison is made based on the majority similarity within the training data [16].

The K-Nearest Neighbor algorithm operates by classifying data according to their similarity with other data. The closer the similarity between K neighbors and the training data, the higher the likelihood that the training data is more similar to the nearest neighbors in the training dataset [17]. To calculate this proximity, the algorithm typically utilizes the Euclidean distance, which is a formula for determining the distance between two points in the training dataset. To calculate using the Euclidean distance, the (6) is required as follows:

$$d_i = \sqrt{\sum_{i=1}^{p} (x_{2i} - x_{1i})^2} \tag{6}$$

Where x_{2i} is the input of the 1st data from the testing dataset, x_{1i} is the input of the 1st data from the training dataset, and d_i is the Euclidean distance [18].

F. Naive Bayes

Naive Bayes applies Bayes' Theorem by probabilistically analyzing data and assumes the presence or absence of certain features in a class and their independence from features in other classes. This logarithm emphasizes probability estimation, hence termed "Naive" for assuming the independence of specific features from the occurrence of other features. Analyzing with Naive Bayes involves employing the (7).

$$P(A|B) = \frac{P(A) \cdot P(B|A)}{P(B)} \tag{7}$$

In this formula, B represents the class to be determined from the data, while A is the hypothesis class. $P(A|B)$ denotes the probability of A occurring given the condition B, $P(A)$ is the probability of A occurring, and $P(B)$ is the probability of B occurring [16].

G. Synthetic Minority Oversampling Technique (SMOTE)

The SMOTE algorithm applies an oversampling approach to rebalance the original training dataset. The oversampling system employed by the SMOTE method involves using synthetic data, where it duplicates the smallest quantity of data and then creates random data while maintaining the same data distribution [19].

III. Results and Discussion

In the initial analysis, the dataset exhibits imbalance, with an unequal distribution of instances across each class. This imbalance has the potential to impact the performance of the classification models. This issue is evident in the following Table III, where the KNN and NB models achieve very high accuracy, but the F1-Score values are relatively low.

TABLE III: Model performance on Imbalanced Datasets

Model	Accuracy	F1-Score
KNN	99.24%	67.20%
NB	99.19%	71.22%

Based on the results presented in the table, it can be concluded that the limitation in the number of samples labeled as stunting significantly impacts the performance of the classification process in the utilized model. To address this issue, additional data labeled as stunting was introduced through the oversampling technique using SMOTE.

Before applying the oversampling technique, undersampling was initially performed on the data labeled as normal, reducing its quantity from 6534 to 500 instances. This action was taken to prevent overfitting and ensure that the model does not become overly specific to the training data, thereby enhancing its generalization to new data. Subsequently, the stunting-labeled data was transformed through oversampling, increasing its quantity from 54 to 500 instances to achieve a balance with the data labeled as normal.

A. K-Nearest Neighbors (KNN) Result

The division between the training and testing datasets is carried out with a ratio of 70% for training data and 30% for testing data. The value of K (number of nearest neighbors) employed is 3. Following the execution of tests on the dataset using the KNN model, the results of the confusion matrix can be found in the following Table IV.

TABLE IV: KNN Confusion Matrix Result

		Class Predicted	
		Stunting	**Normal**
Class Actual	**Stunting**	145	11
	Normal	2	142

After obtaining the results from the confusion matrix for the KNN model, where there are 142 True Positives (TP), 11 False Positives (FP), 145 True Negatives (TN), and 2 False Negatives (FN), the analysis proceeds with calculations using the appropriate formulas. Subsequently, values such as Accuracy, Precision, Recall, and F1 Score are computed to evaluate the performance of the classification model. The calculated results are presented in Table V.

TABLE V: KNN Result

Accuracy	**Precision**	**Recall**	**F1-Score**
95.67%	92.81%	98.61%	95.62%

Based on these results, the implementation of oversampling techniques significantly improved the performance of KNN. Previously, the F1-Score was 67.20%. However, after applying the oversampling technique, the F1-Score increased to 95.62%. Additionally, the obtained accuracy is 95.67%. The difference in results produced by the KNN model before imbalance and after balancing can be observed in Fig 3.

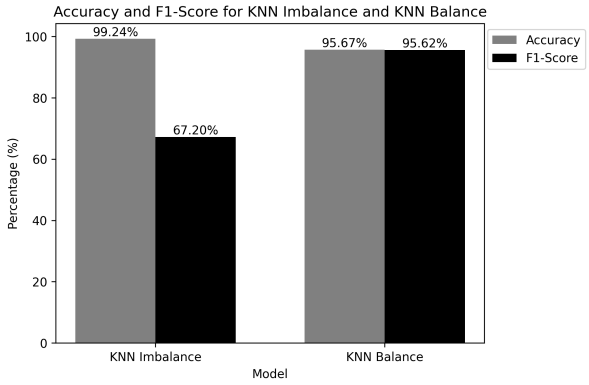

Fig. 3: Result for Balance and Imbalance KNN model

B. Naive Bayes (NB) Result

For testing using the NB model, the data division remains consistent with KNN, where 70% is allocated for training and 30% for testing. The intention is to maintain the same specifications for comparison between the two models. After conducting the tests on the dataset using the NB model, the results from the confusion matrix can be found in Table VI.

After obtaining the results from this confusion matrix, it can be observed that there are 149 True Positives (TP), 15 False Positives (FP), 133 True Negatives (TN), and 3 False Negatives (FN) in the NB model. The results from this confusion matrix are further utilized to calculate the values

TABLE VI: NB Confusion Matrix Result

		Class Predicted	
		Stunting	**Normal**
Class Actual	**Stunting**	133	15
	Normal	3	149

for Accuracy, Precision, Recall, and F1 Score. The results of these calculations can be observed in Table VII.

TABLE VII: NB Result

Accuracy	**Precision**	**Recall**	**F1-Score**
94%	90.85%	98.02%	94.30%

Based on these results, the implementation of oversampling techniques in the performance of NB also successfully provided a significant improvement. Previously, the F1-Score was 71.22%; however, after applying the oversampling technique, the F1-Score increased to 95.62%. Additionally, the obtained accuracy is 94%. The difference in results produced by the NB model before imbalance and after balancing can be observed in Fig 4.

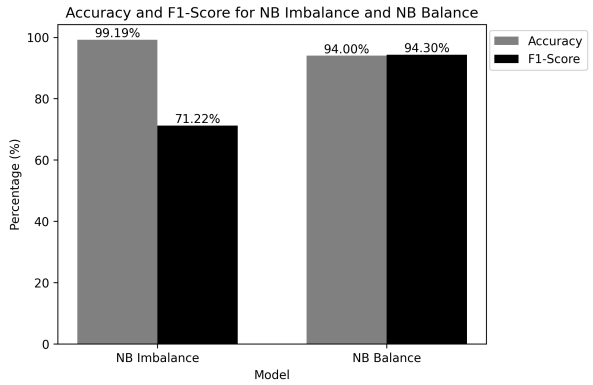

Fig. 4: Result for Balance and Imbalance NB model

IV. Conclusion

From the results obtained from the conducted tests, it can be concluded that both classification models, namely K-Nearest Neighbors (KNN) and Naive Bayes (NB), have demonstrated good performance in predicting cases of stunting in toddlers. Initially, the dataset exhibited an imbalance that affected the model evaluations, evident from the high accuracy but low F1-Score values for both models. However, after applying appropriate handling and preprocessing, both models experienced significant improvement. The KNN model showed an increase in F1-Score from 67.20% to 95.62%, with an accuracy of 95.67%. Meanwhile, the NB model also demonstrated an increase in F1-Score from 71,22% to 94,30%, with an accuracy of 94%. The comparative analysis between K-Nearest Neighbors (KNN) and Naive Bayes (NB) indicates that KNN exhibits slightly superior performance compared to NB. However, both models showcase commendable capabilities and can be utilized effectively as tools for predicting stunting status swiftly and accurately in toddlers. As a suggestion, future research may consider exploring the combination or ensemble of both classification models used (K-Nearest Neighbors and Naive Bayes). Furthermore, the study could delve deeper by utilizing a dataset that does not involve oversampling, aiming to provide a more accurate representation of stunting prediction in toddlers. Thus, future research is expected to contribute further to the understanding and implementation of classification methods for predicting stunting.

References

[1] T. Prasetiya, I. Ali, C. L. Rohmat, and O. Nurdiawan, "Klasifikasi status stunting balita di desa slangit menggunakan metode k-nearest neighbor," *INFORMATICS FOR EDUCATORS AND PROFESSIONAL: Journal of Informatics*, vol. 5, no. 1, p. 93, 2020.

[2] Kementerian Kesehatan Republik Indonesia. (2023) Prevalensi stunting di indonesia turun ke 21,6% dari 24,4%. Accessed on May 16, 2023. [Online]. Available: https://www.kemkes.go.id/article/view/23012500002/prevalensi-stunting-di-indonesia-turun-ke-21-6-dari-24-4-.html

[3] World Health Organization, *Reducing Stunting in Children: Equity Considerations for Achieving the Global Nutrition Targets 2025*. World Health Organization, 2018.

[4] H. Saleh, "Analisa faktor penyebab stunting menggunakan algoritma c4.5," *ScientiCO: Computer Science and Informatics*, vol. 3, pp. 11–17, 2020.

[5] R. A. Saputri, "Upaya pemerintah daerah dalam penanggulangan stunting di provinsi kepulauan bangka belitung," *Jurnal Universitas Abdurrab*, vol. 2, no. 2, pp. 152–168, Agustus 2019.

[6] R. Primartha, *Belajar Machine Learning: Teori dan Praktik*. Bandung: Penerbit Informatika, 2018, penerbit Informatika.

[7] C. Chazar and B. E. Widhiaputra, "Machine learning diagnosis kanker payudara menggunakan algoritma support vector machine," *INFORMASI (Jurnal Informatika dan Sistem Informasi)*, vol. 12, no. 1, Mei 2020.

[8] R. Setiawan and A. Triayudi, "Klasifikasi status gizi balita menggunakan naïve bayes dan k-nearest neighbor berbasis web," *JURNAL MEDIA INFORMATIKA BUDIDARMA*, vol. 6, no. 2, p. 777, 2022.

[9] A. Zubair and M. Muksin, "Penerapan metode naive bayes untuk klasifikasi status gizi (studi kasus di klinik bromo malang)," *senasif*, vol. 2, no. 1, pp. 1204–1208, September 2018.

[10] S. Y. Sibi and A. R. Widiarti, "Klasifikasi tingkat obesitas mempergunakan algoritma knn," *Prosiding Corisindo*, pp. 370–375, 2022.

[11] R. R. R. Arisandi, B. Warsito, and A. R. Hakim, "Aplikasi naÏve bayes classifier (nbc) pada klasifikasi status gizi balita stunting dengan pengujian k-fold cross validation," *Jurnal Gaussian*, vol. 11, no. 1, pp. 130–139, 2022.

[12] L. Andiani, Sukemi, and D. Palupi Rini, "Analisis penyakit jantung menggunakan metode knn dan random forest," in *Prosiding Annual Research Seminar*, vol. 5, 2019, pp. 1–5.

[13] A. D. Ghani, N. Salman, and Mustikasari, "Algoritma k-nearest neighbor berbasis backward elimination pada client telemarketing," in *Prosiding Seminar Ilmiah Sistem Informasi Dan Teknologi Informasi*, vol. 8, no. 2, 2019, pp. 141–150.

[14] R. R. Waliyansyah and C. Fitriyah, "Perbandingan akurasi klasifikasi citra kayu jati menggunakan metode naive bayes dan k-nearest neighbor (k-nn)," *JEPIN (Jurnal Edukasi dan Penelitian Informatika)*, vol. 5, no. 2, pp. 1–6, August 2019.

[15] D. K. Wardya, I. K. G. D. Putra, and N. K. D. Rusjayanthi, "Clustering artikel pada portal berita online menggunakan metode k-means," *JITTER : Jurnal Ilmiah Teknologi dan Komputer*, vol. 3, no. 1, pp. 985–993, 2022. [Online]. Available: https://ojs.unud.ac.id/index.php/jitter/article/view/84732

[16] A. P. Permana, K. Ainiyah, and K. F. H. Holle, "Analisis perbandingan algoritma decision tree, knn, dan naive bayes untuk prediksi kesuksesan start-up," *JISKA (Jurnal Informatika Sunan Kalijaga)*, vol. 6, no. 3, pp. 178–188, 2021.

[17] R. R. Sani, J. Zeniarja, and A. Luthfiarta, "Penerapan algoritma k-nearest neighbor pada information retrieval dalam penentuan topik referensi tugas akhir," *Journal of Applied Intelligent System*, vol. 1, no. 2, pp. 123–133, 2016.

[18] Wahyudi, M. Orisa, and N. Vendyansyah, "Penerapan algoritma k-nearest neighbors pada klasifikasi," *Jurnal Mahasiswa Teknik Informatika (JATI)*, vol. 5, no. 2, pp. 750–757, 2021.

[19] A. Fernández, S. García, F. Herrera, and N. V. Chawla, "Smote for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary," *Journal of Artificial Intelligence Research*, vol. 61, pp. 863–905, 2018.