JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 10, No. 2, Juni 2025, Pp. 1111-1117

PREDICTION OF STUNTING NUTRITIONAL STATUS IN TODDLERS USING NAÏVE BAYES CLASSIFIER ALGORITHM

Rudi Hariyanto[1], Mohammad Zoqi Sarwani[*2], Yunita Nur Aprilia[3]

1. Informatika, Teknologi Informasi, Universitas Merdeka Pasuruan, Indonesia
2. Informatika, Teknologi Informasi, Universitas Merdeka Pasuruan, Indonesia
3. Informatika, Teknologi Informasi, Universitas Merdeka Pasuruan, Indonesia

Article Info

Keywords: Naïve Bayes Classifier; Nutrional Status; Predictions; Stunting

Article history:
Received 24 Oktober 2024
Revised 19 November 2024
Accepted 28 Desember 2024
Available online 15 March 2025

DOI :
https://doi.org/10.29100/jipi.v10i2.5930

* Corresponding author.
Mohammad Zoqi Sarwani
E-mail address:
zoqi.sarwani@unmerpas.ac.id

ABSTRACT

Stunting is a chronic nutritional problem in toddlers that affects children's physical growth and cognitive development. Early identification and prediction of toddlers' nutritional status are crucial for timely intervention. This study aims to predict the nutritional status of stunting in toddlers using the Naïve Bayes Classifier algorithm. The data used in this study is derived from community health surveys with variables such as age, weight, height, and parental nutritional status. The research process began with data collection and pre-processing to ensure high-quality data. Subsequently, the data was trained using the Naïve Bayes Classifier algorithm, known for its simplicity and efficiency in data classification. Prediction results were then evaluated using metrics of accuracy, precision, and recall to measure the model's performance. The study results indicate that the Naïve Bayes Classifier algorithm has high accuracy in predicting stunting status in toddlers, with an accuracy rate of 72,2%. Precision and recall also showed satisfactory results, at 94,1% and 76,1%, respectively. This model can be used as a tool for health workers to identify toddlers at risk of stunting, enabling earlier preventive measures. In conclusion, the use of the Naïve Bayes Classifier algorithm is proven effective in predicting the nutritional status of stunting in toddlers. The implementation of this model is expected to support child health programs and accelerate the reduction of stunting prevalence in the community.

I. INTRODUCTION

STUNTING is a widespread issue in Indonesia and other developing countries. Stunting in children, as illustrated here, refers to being below average height. This results from a prolonged mismatch between long-term nutrition intake and requirements. Delayed cognitive development, impaired learning capacity, and increased vulnerability to metabolic syndrome, hypertension, and obesity are potential outcomes of developmental delays [1] [2].

He stunting status in toddlers is determined according to the Minister of Health of the Republic of Indonesia Decree No. 13. 1995/MENKES/SK/XII/2010, which regulates the procedures for assessing children's nutritional status, using growth and development standards tailored to community nutrition assessment programs [3]. Assessment is based on height for age (HAZ) or length for age (LAZ) between -3 SD and -2 SD. If the HAZ or LAZ measurement is below -3 SD, it is considered as severely short stature. Maternal and infant mortality, child stunting, infectious diseases, and non-communicable disease control are the four main focuses of health development initiatives [4].

In 2018, the Ministry of Health of the Republic of Indonesia recorded that the percentage of stunting in Indonesia reached 30.8%, which is significantly higher compared to other ASEAN countries ranging between 4% to 17%. However, according to the Indonesian Nutrition Status Survey (SSGI) in 2022, the stunting rate in East Java Province was recorded at 19.2%. The national target for 2024 is to reduce the stunting rate to 14%. In Pasuruan Regency, the prevalence of stunting was recorded at 21.5% in 2020. This figure then decreased to 18.1% in 2021, and significantly dropped to 10.8% in 2022. In the city of Pasuruan, the prevalence of stunting has fluctuated over the past three years. In 2020, the stunting prevalence was around 19.06%. This number saw a notable increase in 2021 to 23.7%, but decreased again to 18% in 2022 [5].

Nutritional deficiencies are caused by various factors, such as inadequate food consumption, poverty, inadequate environmental sanitation, and lack of nutritional knowledge. All of these factors work together to create ecological issues that result in malnutrition [6] [7].

Monitoring and data collection on stunting at various health centers in the Pasuruan City area play a crucial role in assessing the growth and development of fetuses and newborns. One of the frequent issues encountered at the Pasuruan City Health Office is the inaccuracy and lack of timeliness in collecting monthly stunting data, which is still done manually using Microsoft Excel. This results in large data sizes and heavier calculation processes, limiting public access to the data to once a month during integrated health post activities. Given this problem, there is an urgent need to develop a system to track the status of young children, especially stunting, using multiple indicators to support innovation and enhance public understanding of the importance of adequate nutrition for toddlers.

The method employed in this research is data mining technique using the Naive Bayesian classifier algorithm titled "Prediction of Stunting Nutritional Status in Toddlers Using Naive Bayes Classifier Algorithm". The Naive Bayes classification algorithm is a statistical procedure used to estimate the likelihood of group membership. When applied to large datasets, this Naive Bayesian classification technique achieves high accuracy and speed. Compared to other classification algorithms, Naive Bayes has significantly lower error rates [8].

The method applied in this research is a data mining technique using the Naive Bayesian classifier algorithm titled "Prediction of Stunting Nutritional Status in Toddlers Using the Naive Bayes Classifier Algorithm." This algorithm was chosen because when applied to large datasets, the Naive Bayesian classification technique achieves high accuracy and speed. Compared to other classification algorithms, the error rate of Naive Bayes is much lower [8] and the Naive Bayes method has a higher accuracy rate as in the study [9] where Naive Bayes achieved 70% accuracy while the support vector machine method only reached 61%. The Naive Bayes algorithm used in this study is Gaussian Naive Bayes because the data used in this study are continuous and the expected target is binary [10].

II. RESEARCH METHOD

Research methodology is the stage carried out in research used to solve existing problems. In this study, the research methodology includes several discussions covering the research flow, literature review, problem identification, data collection, data analysis, classification using Naïve Bayes, and finally, model evaluation.

A. Research Flow

In this study, there are several stages to be conducted. These stages start with conducting a literature review, identifying the problem, collecting data, analyzing data, modeling, optimizing, evaluating the model, and drawing conclusions. Figure 1 illustrates the flow of this research.

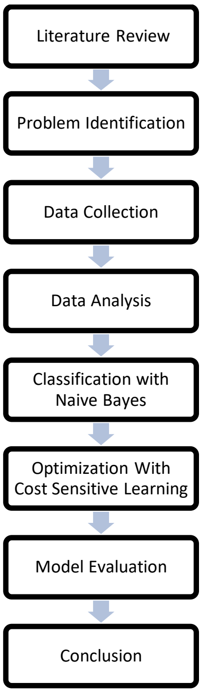

Figure 1. Research Flow

Literature review is a stage conducted by researchers to gather references used in addressing the problems selected by the researcher. The problem identification process is a stage to gain an understanding of the issues to be solved in this study. The data collection process is used to gather raw data, which will then be processed in the data analysis stage to determine the appropriate parameters. Additionally, the data analysis stage also includes feature engineering, which involves cleaning and improving data to ensure it can be processed by the Naive Bayes method or algorithm used in this research. The final process is model evaluation to assess how accurately the Naive Bayes algorithm resolves the issues in this study.

B. Literature Review

In this stage, the initial step of conducting research involves gathering information about the questions studied previously from various sources of references, beginning with searching for references in books, journals, reports, and articles related to developmental delays in early childhood.

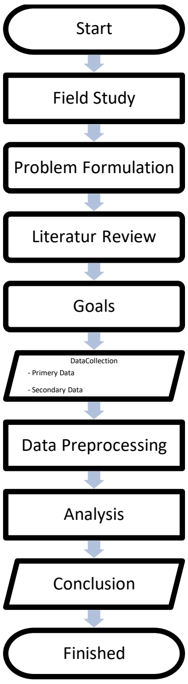

Figure 2. Literature Review

C. Problem Identification

In this study, the problem identification process is used to obtain a detailed overview of the issues addressed in this research. Detecting toddlers suffering from stunting is crucial as it relates to child health. The issues caused by toddlers suffering from stunting can contribute to mortality. Lack of knowledge about the symptoms of this condition is prevalent in society, so educational efforts should not only focus on raising awareness but also on providing a platform accessible to the general public to prevent or detect early signs of stunting in toddlers.

D. Data Collection

The data sources used in this research are from the Health Office and the nearest health centers (puskesmas). Discussions with medical staff responsible for dietary services were conducted to obtain additional information. These interviews aimed to gather information about the data collection procedures to be followed and the types of data that can be used in this research. The nutrition officers obtained data in the form of Excel files, which include the number of stunting cases from August to December 2022 in Pasuruan City.

The data collected includes nutritional status data of toddlers, totaling 110 entries across 16 columns, including "Name, Gender, Village/Subdistrict, Integrated Health Post (Posyandu), RT (neighborhood association), RW (residential association), Weight, Height, Upper Arm Circumference, Weight for Age (WFA), WFA Z-score, Height for Age (HFA), HFA Z-score, Weight for Height (WFH), and WFH Z-score."

TABLE 1.
CHILD STUNTING DATASET STUNTING

No.	…	Weight	Height	Upper Arm Circumference	WFA	WFA Z-score	HFA	HFA Z-score	WFH	WFH Z-score
1.		12,4	90,3	12	Normal	-1,86	stunted	-2,6	adequate nutrition	-0,55
2.		8,2	79	14	severely malnourished	-3,2	stunted	-2,61	malnutrition	-2,34
3.		7,5	68	13	Normal	-1,82	stunted	-2,3	adequate nutrition	-0,74
4.		10,3	84,5	13	Normal	-2,12	stunted	-2,36	adequate nutrition	-1,03
5.		9,6	78	14	Normal	-1,69	stunted	-2,58	adequate nutrition	-0,59
…										
110		12,5	76	14	undernourished	-2,98	stunted	-2,14	malnutrition	-2,3

The data will undergo preprocessing by transforming the values of the BB/U, TB/U, and BB/TB features. The transformation of each feature is carried out in several ways:

- WFA Feature:

Transform the categorical values into ordinal numbers: the "severely underweight" category is changed to 1, the "underweight" category is changed to 2, and the "normal" category is changed to 3. This transformation is called ordinal encoding, where the "normal" category has the highest value.

- HFA Feature:

Transform the categorical values into binary numbers: the "short" category is changed to 0 and the "tall" category is changed to 1. This transformation is called one-hot encoding, which is the process of converting categorical values into several binary columns, with one column for each unique category.

- WFH Feature:

Transform the categorical values into binary numbers: the "well-nourished" category is changed to 1 and the "malnourished" category is changed to 0. This transformation is also called one-hot encoding, which is the process of converting categorical values into several binary columns, with one column for each unique category.

E. Data Analysis

Data analysis is a step used by researchers to determine the parameters that are used as input for the Naive Bayes algorithm and to perform feature engineering improvements. In this study, the data analysis process was carried out in several stages, starting from analyzing the correlation between parameters, detecting data balance based on the target parameter, and handling missing values.

The first step is to analyze the correlation between parameters. This analysis is performed using statistical methods such as the Pearson correlation coefficient to measure the strength and direction of the linear relationship between two parameters. For example, if parameter A and parameter B have a high correlation value, it indicates that changes in parameter A are likely followed by changes in parameter B.

Next, we detect whether the data we have is balanced or not based on the target parameter. A common method used for this is class distribution analysis, where we count the number of instances in each target class and compare them. Balanced data is data where the distribution of target classes is almost equal or proportional.

The next step is to handle missing values. Several techniques commonly used to handle missing values include deleting rows or columns containing missing values, imputing with mean or median values, and using more advanced algorithms such as k-nearest neighbors to impute missing values.

F. Classification with Naïve Bayes

Naive Bayes is a simple probabilistic classification technique designed under the assumption that explanatory variables are independent. This algorithm focuses on learning in probability estimation. The advantage of Naive Bayes algorithm is that it exhibits lower error rates, higher accuracy, and better processing time for large datasets compared to smaller ones [11].

The Naive Bayes algorithm has three types: Multinomial Naive Bayes, Bernoulli Naive Bayes, and Gaussian Naive Bayes. This study uses the Gaussian Naive Bayes algorithm because the data used in this research is continuous, and the expected target is binary [10]. Gaussian Naive Bayes is characterized by its ability to handle continuous (non-discrete) feature values. Each feature is assumed to follow a Gaussian distribution, which is also known as a normal distribution. When the data is plotted, it produces a bell-shaped curve that is symmetrical around the mean. This bell curve represents the average value of the feature and shows how the data is distributed around the mean, with most values clustering near the center and fewer values appearing as you move away from the center in either direction. The symmetry of the curve indicates that the data is evenly distributed on both sides of the mean, with equal probabilities for values occurring above and below the average [12].

The formula for Gaussian Naive Bayes is based on the normal (Gaussian) distribution to estimate the probability of continuous features [12]. The formula for the normal distribution is:

$$P(x|C) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{1}$$

Where:
$P(x|C)$ is the probability of feature x given class C
μ is the mean of the feature for class C
σ is the standard deviation of the feature for class C
π is the constant pi (approximately 3,14159).
exp is the exponential function.

In the context of Gaussian Naive Bayes, the total probability of class C given features $x_1, x_2, x_3, \dots, x_n$ is calculated as:

$$P(C|x_1, x_2, x_3, \dots, x_n) = P(C) \prod_{i=1}^{n} P(x_i|C) \tag{2}$$

Where:
(P(C) is the prior probability of class C.
(P($x_i|C$) is the probability of feature (x_i) given class C, calculated using the normal distribution formula above.

G. *Model Evaluation*

A confusion matrix is a technique used to assess the effectiveness of a model. It provides a comparison between the output produced by the system and the actual values [13].. The confusion matrix allows for a more detailed evaluation of how the model performs classification by breaking down the prediction results into four categories:
- True Positives (TP): Cases where children suffering from stunting are correctly predicted as stunting.
- True Negatives (TN): Cases where children not suffering from stunting are correctly predicted as not stunting.
- False Positives (FP): Cases where children not suffering from stunting are incorrectly predicted as stunting.
- False Negatives (FN): Cases where children suffering from stunting are incorrectly predicted as not stunting.

From the confusion matrix, several important metrics can be calculated, such as accuracy, precision, and recall. Accuracy measures the proportion of correct predictions out of the total predictions made by the model. It provides a general idea of how well the model works in predicting the nutritional status of stunting. Accuracy is important for getting an overall view of the model's performance, especially if the data is balanced. High accuracy indicates that the model is good at correctly classifying children as either stunting or not stunting. However, if the data is imbalanced (the number of stunting cases is much fewer or greater than non-stunting cases), accuracy alone may not adequately reflect the true performance of the model. The formula for accuracy is:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \times 100\% \tag{3}$$

Precision measures the proportion of true positive predictions out of all positive predictions made by the model. This is important when the cost of false positives is high. In a medical context, precision helps ensure that when the model predicts a child as suffering from stunting, the prediction is correct with a high proportion. High precision means the model rarely incorrectly identifies children who are not suffering from stunting as stunting. This is important to avoid unnecessary concern or intervention. The formula for precision is:

$$Precision = \frac{TP}{TP+FP} \times 100\% \tag{4}$$

Recall measures the proportion of true positive cases that are correctly identified out of all actual positive cases. This is important when the cost of false negatives is high. In the context of stunting, recall is important because every missed case of stunting can mean a failure to provide necessary intervention to the child who needs it. High recall indicates that the model is effective in detecting all children suffering from stunting, ensuring that children requiring medical attention are not overlooked. The formula for recall is:

$$Recall = \frac{TP}{TP+FN} \times 100\% \qquad (5)$$

III. RESULTS AND DISCUSSION

A. Data Analysis

In this study, data from Integrated Health Posts (Posyandu) were used to predict stunting. Several stages were involved in the data analysis process, including feature selection and data preprocessing or feature engineering. The results of the data analysis conducted in this study include:

- Feature Selection

The selection process resulted in parameters used for this study, such as age, height, weight, gender, and nutritional status. Features such as name and address were not used because they are unique to each data entry.

- Feature Engineering

Feature engineering in this study involved improvements such as converting data types for parameters like age, height, and weight from object to numeric. Data cleaning involved removing strings or words such as 'kg' from the weight parameter

B. Naive Bayes Classifier

The test results conducted for this study provide information about the reliability of the results. The purpose of this study is to evaluate the effectiveness of the Naive Bayesian algorithm in predicting the nutritional status of toddlers at risk of stunting. With a total of 110 data points (88 for training and 22 for testing), this proportion is generally chosen to ensure that the training set is large enough for the model to learn and the testing set is large enough for accurate evaluation. The Naive Bayesian approach to predicting the nutritional status of stunted toddlers achieved an accuracy rate of 72.7%.

Furthermore, further analysis was conducted to understand how certain features affect the prediction results. Features such as weight, height, age, and maternal health status during pregnancy were identified as important variables in this model. The test results show that the Naive Bayesian algorithm has a fairly good ability to classify toddlers at risk of stunting based on the available data. Additionally, a confusion matrix was used to evaluate the model's performance in more detail. From the confusion matrix results, satisfactory precision and recall values were obtained, with each being 94.1% and 76.1%, respectively."

C. Confusion Matrix

In the model evaluation phase, confusion matrix calculations were used to determine how much of the previously processed test data fell into categories such as TP (True Positive), FP (False Positive), TN (True Negative), and FN (False Negative). The results are detailed in Table 2.

Table II
CONFUSION MATRIX RESULTS

Confusion Matrix		**Prediction**	
		Positive	*Negative*
Actual	*Positive*	TP = 16	FP = 1
	Negative	FN = 5	TN = 0

Table 2 shows that the Naive Bayes classifier model correctly predicts the status of developmental delay in young children with 16 True Positives, 5 False Negatives, 1 False Positive, and 0 True Negatives.

However, it is important to note that the imbalance in the test data can affect the model's results. In this case, the much smaller number of negative samples compared to positive samples can cause the model to have difficulty correctly predicting the negative class. This is evident from the True Negative value of 0, which indicates that the model did not correctly predict any negative samples.

Table III.
CONFUSION MATRIX RESULTS

Method	Accuracy	Precission	Recall
Naïve Bayes Classifier	72,7%	94,1%	76,1%

From Table 3, it can be seen that the precision value is 94,1%, recall is 76,1%, and accuracy is 72,7%. The impact of data imbalance can affect the interpretation of the evaluation metrics used. To address data imbalance, several methods can be applied, including resampling techniques such as oversampling for the minority class or under-sampling for the majority class.

IV. CONCLUSION

Data preparation, data preprocessing, Naive Bayesian classification process, and the use of Confusion Matrix to evaluate the performance of the Naive Bayesian method are all necessary steps in developing a model to predict developmental delays in young children. This information is derived from 2022 statistics on infant and child nutrition health. Both training and test data were incorporated into the Naive Bayesian classification procedure. High precision was achieved using the Confusion Matrix to accurately locate data. Regarding performance on delayed data, the evaluation indicates that partitioning data into 80% training and 20% testing yields the best results (accuracy 72,7%, precision 94,1%, and recall 76,1%). These findings form the basis for developing more sophisticated and effective prediction systems for toddler stunting, laying the groundwork for future system development. Furthermore, this study highlights the potential for further development considering data imbalance issues.

V. REFERENCES

[1] A. Rahayu *et al.*, "STUDY GUIDE-STUNTING DAN UPAYA PENCEGAHANNYA," 2018.

[2] M. R. Nugroho, R. N. Sasongko, and M. Kristiawan, "Faktor-faktor yang Mempengaruhi Kejadian Stunting pada Anak Usia Dini di Indonesia," *Jurnal Obsesi : Jurnal Pendidikan Anak Usia Dini*, vol. 5, no. 2, pp. 2269–2276, Mar. 2021, doi: 10.31004/OBSESI.V5I2.1169.

[3] "Tumbuh Kembang Anak | PDF." Accessed: Jun. 11, 2024. [Online]. Available: https://id.scribd.com/document/659881609/Tumbuh-Kembang-Anak-1

[4] J. P. Masyarakat *et al.*, "Pemenuhan Pangan Lokal Sebagai Kebutuhan Gizi Bayi Dan Balita Umur 6 -24 Bulan Di Kabupaten Banyumas," *Jurnal Pengabdian Masyarakat - PIMAS*, vol. 1, no. 1, pp. 29–37, Feb. 2022, doi: 10.35960/PIMAS.V1I1.729.

[5] - Badan Penelitian dan Pengembangan Kesehatan, "Laporan Provinsi Jawa Timur Riskesdas 2018," *Kementerian Kesehatan RI*, p. 140, 2019.

[6] F. O. Aridiyah *et al.*, "Faktor-faktor yang Mempengaruhi Kejadian Stunting pada Anak Balita di Wilayah Pedesaan dan Perkotaan (The Factors Affecting Stunting on Toddlers in Rural and Urban Areas)," *Pustaka Kesehatan*, vol. 3, no. 1, pp. 163–170, Jan. 2015, Accessed: Jun. 11, 2024. [Online]. Available: https://jurnal.unej.ac.id/index.php/JPK/article/view/2520

[7] I. Colanus, R. Drajana, and A. Bode, "Prediksi Status Penderita Stunting Pada Balita Provinsi Gorontalo Menggunakan K-Nearest Neighbor Berbasis Seleksi Fitur Chi Square," *Jurnal Nasional Komputasi dan Teknologi Informasi (JNKTI)*, vol. 5, no. 2, pp. 309–316, Apr. 2022, doi: 10.32672/JNKTI.V5I2.4205.

[8] R. Wahyudi, M. Program Studi Ilmu Keperawatan Fakultas Keperawatan Universitas Syiah Kuala Banda Aceh, and S. Pengajar Bagian Keilmuan Keperawatan Anak Fakultas Keperawatan Universitas Syiah Kuala Banda Aceh, "PERTUMBUHAN DAN PERKEMBANGAN BALITA STUNTING THE GROWTH AND DEVELOPMENT OVERVIEW OF THE STUNTING TODDLER."

[9] J. Teknik Elektro dan Komputasi *et al.*, "Implementasi Algoritma Naïve Bayes dan Support Vector Machine (SVM) Pada Klasifikasi Penyakit Kardiovaskular," *Jurnal Teknik Elektro dan Komputasi (ELKOM)*, vol. 4, no. 2, pp. 207–214, Aug. 2022, doi: 10.32528/ELKOM.V4I2.7691.

[10] "Algoritma Naive Bayes: Definisi dan Contoh Penerapannya." Accessed: Jul. 17, 2024. [Online]. Available: https://blog.algorit.ma/algoritma-naive-bayes/

[11] T. D. J. P. SC Chu, "Identifying correctness data scheme for aggregating data in cluster heads of wireless sensor network based on naive Bayes classification [J]," *EURASIP J. Wirel. Commun. Netw.*, vol. 20, no. 1, pp. 963–982, 2020.

[12] "Gaussian Naive Bayes. Gaussian Naive Bayes merupakan sebuah… | by Michelle Ha | Medium." Accessed: Jul. 17, 2024. [Online]. Available: https://medium.com/@chellehdwjy/gaussian-naive-bayes-f05ec0b61d91

[13] D. Normawati and S. A. Prayogi, "Implementasi Naïve Bayes Classifier Dan Confusion Matrix Pada Analisis Sentimen Berbasis Teks Pada Twitter," 2021.