Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2231−2239
ISSN 2684-8910 (print media)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.6670

Analysis of Stunting Prediction in Toddlers in Bekasi District Using *Random Forest* and *Naïve Bayes*

Chintya Annisah Solin*, Putu Harry Gunawan

Faculty of Informatics, Telkom University, Bandung, Indonesia
Email: [1,*]chintyaanisa@student.telkomuniversity.ac.id, [2]phgunawan@telkomuniversity.ac.id
Correspondence Author Email: chintyaanisa@student.telkomuniversity.ac.id
Submitted: **10/01/2025**; Accepted: **26/02/2025**; Published**: 01/03/2025**

Abstract−This study aims to compare the performance of the Random Forest and Naïve Bayes algorithms in predicting stunting in toddlers using data from the Bekasi District Health Office. The analysis process begins with data cleaning, normalization, and sampling using the Adaptive Synthetic Sampling (ADASYN) method to handle data imbalance, followed by validation with Stratified K-Fold Cross Validation. The implementation of the algorithm shows that Random Forest has the highest accuracy of 89.62% and an F1-Score of 89.09%. Naïve Bayes Gaussian produces an accuracy of 88.72% and an F1-Score of 88.81%, while Naïve Bayes Bernoulli has a lower performance with an accuracy of 67.83% and an F1-Score of 69.72%. Random Forest shows advantages in overcoming noise and imbalanced data, making it an optimal choice for stunting prediction. Meanwhile, the performance of Naïve Bayes is influenced by the characteristics of the data, where the Gaussian variation is more suitable for continuous data. The results of this study provide insight that choosing the right algorithm, especially on imbalanced data, is very important to improve prediction accuracy. This study also recommends more attention to data preprocessing to ensure optimal prediction quality, especially for minority classes.

Keywords: Stunting; Naïve Bayes; Random Forest; Adasyn; K-Fold

1. INTRODUCTION

Stunting is one of the chronic nutrition problems that is of global concern, especially in developing countries, including Indonesia. This problem occurs due to a lack of nutritional intake in the long term, which has an impact on the disruption of physical growth and development of children. Children who are stunted generally have a lower height than children their age, and are at risk of cognitive and health impairment in the future. This situation is very worrying, especially for children in the golden period of the first 1000 days of their lives (HPK), which is calculated from pregnancy to the age of two. The HPK period is a very critical time because it is during this period that the foundation for the development of the brain, immune system, and other vital organs is formed [1] , [2] .

According to data quoted from *upk.kemkes.go.id*, the government through the Ministry of Health announced that the prevalence of stunting in Indonesia has decreased significantly in recent years. In 2021, the stunting prevalence rate was at the level of 24.4%, and managed to decrease to 21.6% in 2022. This decline is the result of various government efforts and collaboration with various parties in stunting prevention programs involving nutrition education, supplementary feeding, and improving maternal and child health services [3].

Locally, local governments also continue to show commitment to overcoming the problem of stunting. One example is Bekasi Regency, which targets a reduction in stunting prevalence by 14% by 2024, as quoted from *prokopim.bekasikab.go.id*. This target is quite ambitious, considering the consistent reduction in stunting rates of 3% per year in the last three years. This shows that the stunting prevention program in this area has given positive results, as well as a motivation to continue to improve the quality of nutrition and health intervention programs [4].

Although various efforts have been made and the results are beginning to be seen, stunting is still a serious threat to the future of the nation. Therefore, further research on stunting is needed to understand the dynamics of its prevalence, evaluate the effectiveness of programs that have been running, and predict future trends. This prediction is important because it can help the government in designing more targeted policies, especially in the face of new challenges such as pandemics, climate change, or economic crises that can affect food security and public health.

In the era of information technology, the use of artificial intelligence (AI) and *machine learning* (ML) has become an innovative solution in dealing with various problems, including in the field of public health. This technology allows for more in-depth and accurate data analysis, resulting in relevant insights to support decision-making. In the context of stunting research, ML algorithms can be used to make predictions and classifications based on existing datasets, such as child anthropometric data, nutritional status, maternal health conditions, and environmental factors.

Several previous studies have shown the success of ML algorithms in predicting and classifying stunting-related data. According to Indah Pratiwi Putri and her colleagues. (2024), a comparison between three ML algorithms, namely *Naive Bayes*, *K-Nearest Neighbors* (KNN), and *Random Forest*, shows that *Random Forest* provides the best performance with an accuracy of 87.75%. This algorithm is followed by KNN with an accuracy of 84.5%, and *Naive Bayes* with an accuracy of 83.2%. These findings suggest that *Random Forest* has an advantage in capturing complex data patterns, although other algorithms also show quite good performance. Another study conducted by Fadellia Azahra and her colleagues corroborates these findings. They report that [5] *the Random Forest* model is able to achieve an accuracy of 97.88%, a very high number and shows the great potential of this algorithm in analyzing stunting data. Similar results were also found by Muhammad Ghiyaats Daffa, who compared the algorithms of [6]

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2231−2239
ISSN 2684-8910 (print media)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.6670

Random Forest, KNN, and *Boosted KNN*. In the study, *Random Forest* again showed the highest accuracy of 97.76%, with an F1 score of 97.70% [7].

However, behind the success of the *Random Forest algorithm*, there are interesting questions about the potential of other, simpler algorithms, such as *Naive Bayes*. The algorithm has several variants, including *Gaussian Naive Bayes* and *Bernoulli Naive Bayes*, each of which has advantages in handling different types of data. *Gaussian Naive Bayes*, for example, is designed to handle continuous data, while *Bernoulli Naive Bayes* is more suitable for binary data. With this different approach, further research is needed to evaluate whether any of *the Naive Bayes variants* are able to compete with or even outperform *Random Forest* in certain contexts.

This study aims to compare two main algorithms, namely *Naive Bayes* and *Random Forest*, in predicting stunting. In the *Naive Bayes* algorithm, this study will explore two variants, namely *Gaussian Naive Bayes* and *Bernoulli Naive Bayes*. This comparison not only aims to find the model with the best accuracy, but also to understand the characteristics of each algorithm, including its advantages and disadvantages in the context of stunting data analysis.

The selection of *the Random Forest* algorithm is based on its ability to handle data with a large number of variables and high complexity. This algorithm uses an ensemble learning approach, where decisions are made based on the combined results of many decision trees. This advantage makes *Random Forest* very effective in overcoming overfitting and providing stable results.

On the other hand, *Naive Bayes* offers an edge in simplicity and computing efficiency. Assuming independence between variables, this algorithm is able to provide quite good results even with limited computing resources. This makes *Naive Bayes* an attractive option to apply to cases where the available data is relatively small or when processing speed is a priority.

In addition to the technical aspect, this research is also expected to make a practical contribution to stunting prevention efforts in Indonesia. By understanding the strengths and weaknesses of each algorithm, the results of this study can be used to develop a more effective decision support system in predicting stunting risk. This system can later be integrated with government programs, such as Posyandu, to monitor children's nutritional conditions in real-time and provide timely interventions.

The contribution of this research is not only limited to the technical aspects, but also covers a wider social impact. With more accurate predictions, it is hoped that stunting prevention programs can be more directed, so that available resources can be used optimally. Ultimately, this research aims to support the vision of a stunting-free Indonesia by 2045, in accordance with the government's target to create a healthy, intelligent, and productive golden generation.

Thus, this research not only provides added value in the academic realm, but also has significant practical implications. The combination of modern technology and community-based intervention is expected to be an effective solution in overcoming the stunting problem in Indonesia.

2. RESEARCH METHODOLOGY

2.1 Research Stages

This research was carried out through a series of stages that were systematically designed to ensure the validity and accuracy of the results obtained. These stages include initial steps such as literature review to final analysis using the selected algorithm. A detailed explanation of the research stages can be seen in Figure 1 and the following description.

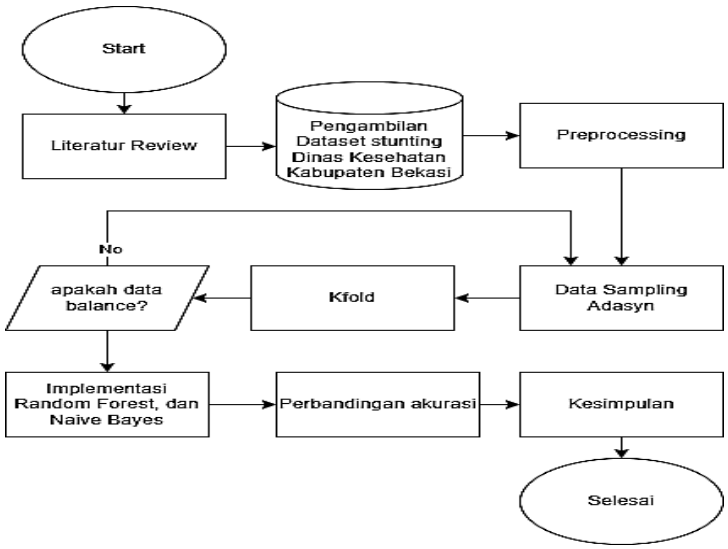


Figure 1. Research Stages

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2231−2239
ISSN 2684-8910 (print media)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.6670

The following is a detailed explanation of the stages in this study, which are listed in Figure 1, starting from the literature review to the final conclusion produced.

a. Literatur Review
The first stage in this study is to conduct a literature review or literature review. This process aims to understand the background of the problem to be researched as well as to find relevant methods and algorithms to be used in research. The literature reviewed includes journals, scientific articles, official reports, and other reliable sources that discuss the topic of stunting as well as the algorithms used for data classification. The results of this review literature are the basis for determining research steps and provide a strong theoretical foundation for further analysis.

b. Dataset Capture
The dataset that became the object of the research was taken from the Bekasi Regency Health Office. This dataset contains information related to stunting conditions in children, such as age, weight, height, and other relevant indicators. This stage is very important because the quality and completeness of the dataset will greatly affect the results of the analysis. The dataset obtained must reflect the actual population so that the research results can describe the conditions in the field. In addition, this process also involves administrative management to ensure that the data taken is in accordance with research ethics and privacy protection.

c. Preprocessing Data
After the dataset is obtained, the next step is to preprocess the data to clean the data from problems that can affect the results of the analysis, such as removing missing values, which are missing or unfilled values, which can be handled by deleting incomplete data or filling in the missing values using the imputation method. In addition, empty (null) or Not a Number (NaN) values need to be removed or repopulated in order for the dataset to be clean. Furthermore, normalization and standardization of data are carried out to ensure that all variables have the same scale, which is important so that algorithms such as Naive Bayes and Random Forest are not affected by differences in variable scales. This preprocessing process ensures that the dataset used is of high quality, structured, and ready for the next stage of analysis.

d. Data Sampling Using ADASYN
The next stage is to deal with data imbalances in the dataset using the Adaptive Synthetic Sampling (ADASYN) method. Data imbalance occurs when the amount of data in a certain class, such as stunting, is much smaller or larger than other classes, such as normal. The goal of ADASYN is to produce synthetic data for minority classes so that the distribution of data becomes more balanced. This is important because data imbalances can cause machine learning algorithms to be more inclined to predict majority classes, resulting in decreased accuracy for minority classes. The ADASYN process works by analyzing the distribution of data and creating synthetic samples based on distance and density of data on minority classes, resulting in a more balanced and representative dataset.

e. K-Fold Cross Validation
After the data is balanced, the next step is to divide the data into several parts (folds) using the K-Fold Cross Validation method. This process aims to ensure that the data is used equally for training and testing. In the way K-Fold works, the dataset is divided into K equal parts. In each iteration, one part is used as testing data, while the rest is used as training data. This process is repeated K times so that each piece of data is used as testing data once. The advantage of K-Fold is that it ensures that all data is used fairly in the training and testing process. This method also helps prevent overfitting by ensuring the model is tested on diverse data. If the K-Fold results show that the distribution of data is still unbalanced, then resampling is carried out to correct the imbalance.

f. Algorithm Implementation
At this stage, the Random Forest and Naive Bayes algorithms are applied to the dataset that has been processed. These two algorithms were chosen because they have their own advantages. Random Forest works by building multiple decision trees and combining the results to make predictions. This algorithm is known for its high accuracy, tolerance for unbalanced data, and resistance to overfitting. Meanwhile, Naive Bayes is based on Bayesian probability theory and is often used for classification. In this study, two Naive Bayes variants were used, namely Gaussian Naive Bayes for continuous data and Bernoulli Naive Bayes for binary data. The implementation process is carried out by training the data using both algorithms, and the prediction results of each model are recorded for further evaluation.

g. Comparison of Algorithm Results
Once the algorithm is applied, the results of the two models are compared to determine which algorithm provides the best performance. Some of the metrics used for evaluation include accuracy, which measures the percentage of correct predictions compared to actual data; precision and recall, which is used to evaluate the performance of the model on a particular class, especially if the dataset is unbalanced; and F1-Score, which is a combined metric that considers precision and recall, providing a more complete picture of the model's performance. By comparing the results of Random Forest and Naive Bayes, this study is expected to provide the most suitable algorithm recommendations for stunting prediction based on the dataset used.

h. Conclusion
The last stage in this study is to draw conclusions from the entire process that has been carried out. The conclusion includes the final results of the performance comparison of the two algorithms, which algorithm has the best accuracy for stunting prediction, as well as the implications of the research results on stunting prevention efforts,

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2231−2239
ISSN 2684-8910 (print media)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.6670

such as providing effective model recommendations for future stunting data analysis. In addition, the conclusion also includes suggestions for further research, such as exploring other algorithms or using larger and varied datasets to obtain more general results.

2.2 *Random Forest*

Random Forest is one of the Machine Learning methods commonly used to classify and regress by producing the final result in the form of a decision tree based on the results of the votes made [8]. *Random Forest* is a type of ensemble learning that uses the bagging method (*Bootstrap Aggregating*) to improve accuracy performance. This algorithm has the advantages of high accuracy, reduced overfitting, tolerance to noise caused by irrelevant data and variables so that the prediction results are stable because variations in the dataset do not affect the final result too much. Here's an example of [9] , [10] *a Random Forest implementation* [11]:

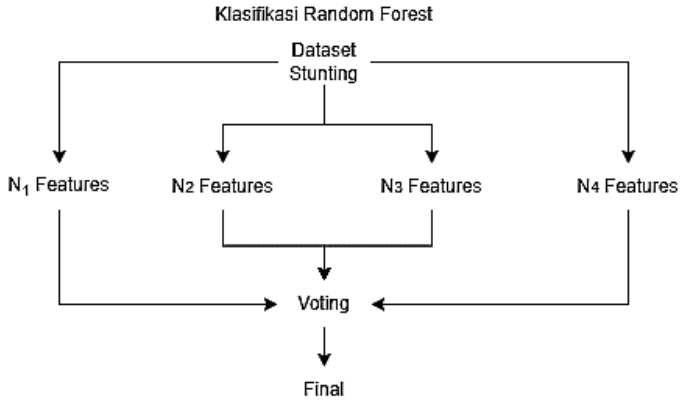


Figure 2. *Random Forest*

In figure 1 above, the Stunting dataset will be broken down into 4N Features that are randomly obtained from the Stunting dataset and generate 4 Decision Trees, the next stage of each decision tree generated will provide predictions for each class according to the data input. Once the decision tree generates the next prediction, the Random Forest algorithm will vote to determine the final prediction [12].

2.3 *Naïve Bayes*

Naïve Bayes is a classification algorithm based on Bayes' Theorem assuming feature independence. In classification, Naïve Bayes calculates the probability of each class based on the features that exist, by utilizing Bayes' Theorem which connects conditional probabilities between classes and features. The algorithm assumes that each feature used to describe the observation is independent, given the given class label. Although this assumption of independence is considered 'naïve' or simple, Naïve Bayes has proven to be effective and is often used in many classification applications, especially on big data and text. Mathematically, Naïve Bayes can be formulated with the following equation [13] , [14]:

$$P(A|B) = \frac{P(A|B)\,P(A)}{P(B)} \quad (1)$$

The probability of an event A, which is often referred to as the symbol P(A), is the probability of event A. the same thing also happens to the probability of event B symbolized by P(B). the probability of an event occurring by considering other events that have already occurred. For example, P(A|B) indicates the probability of event A occurring by assuming that event B has occurred first. That is, we calculate the likelihood of A happening, but only after knowing that B is definitely happening.

In contrast, P(B|A) refers to the probability of event B occurring assuming that event A has already occurred. This means that we are looking for the chance of B occurring, with additional information that A has already occurred.

The naïve bayes used in this study are naïve gaussian bayes and naïve bernoulli bayes. *Naive Bayes Gaussian* (GNB) is a classification method that relies on probability approaches and Gaussian distributions suitable for continuous data and Naive Bayes Bernoulli with boolean data [15].

2.4 *Evaluation Matrix*

Matrix evaluation is used to measure the performance of an algorithmic model based on a specific goal. In this study, metric evaluation was used to assess how well the model predicted data by classifying the data into two categories, namely 'Stunting' and 'Normal'. This evaluation is important to know the extent to which the algorithm can accurately identify these classes. Some of the metrics commonly used in classification evaluation include accuracy, precision, recall, and F1-score, which provide a complete picture of the model's performance in separating the two classes. Here is a systematic evaluation matrix [16]:

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2231−2239
ISSN 2684-8910 (print media)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.6670

$$Akurasi = \frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

$$Presisi = \frac{TP}{TP+FP} \quad (3)$$

$$Recall = \frac{TP}{TP+FN} \quad (4)$$

$$F1 - Score = 2 * \left(\frac{presisi*recall}{presisi+recall}\right) \quad (5)$$

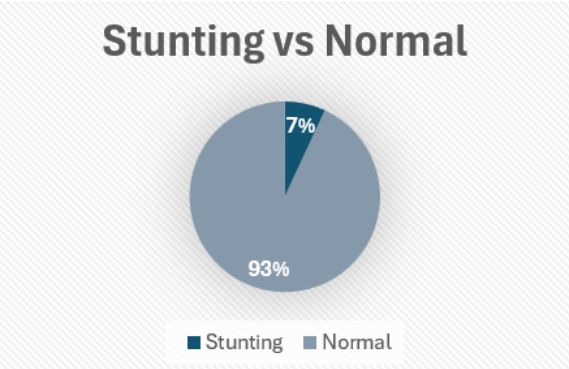
Formulas 2 to 5 are formulas used for the metric evaluation method in assessing the performance of an algorithm model. TP (True Positive) is a case identified as stunting, TN (True Negative) is a case identified as normal or not stunted, FP (False Positive) is an error in predicting stunting and FN (False Negative) is an error in predicting normal or non-stunting cases.

3. RESULTS AND DISCUSSION

3.1 *Explore Data Analyst*

The dataset used in this study is stunting data from the National Unity and Political Agency of Bekasi Regency. This dataset has data of 2255 rows of data collected in April 2024. In this dataset, the information available is in the form of the child's identity, gender, date of birth, weight of the child at birth, height of the child at birth, parent's name, health center, posyandu, age at the time of measurement, date of measurement, ZZ BB/U, TB/U, ZS TB/U, BB/TB, ZS BB/TB, and others.

In this study, the 'TB/U' feature includes categories that describe the height status of children, with values divided into four categories, namely: 'Height', 'Normal', 'Short', and 'Very Short'. Meanwhile, the 'ZS TB/U' feature is the result of calculating the Z-Score score which shows whether the child is stunted or not, based on guidelines and formulas developed by WHO to assess the nutritional status of children. Figure 3 presents the results of data exploration used to analyze the distribution between stunting and normal conditions in the dataset. This analysis aims to understand the data distribution patterns, whether the stunting and normal data are well distributed, and to ascertain whether there are imbalances in the dataset that can affect the performance of the model in the further classification process. A good data balance is very important in obtaining more accurate prediction results.



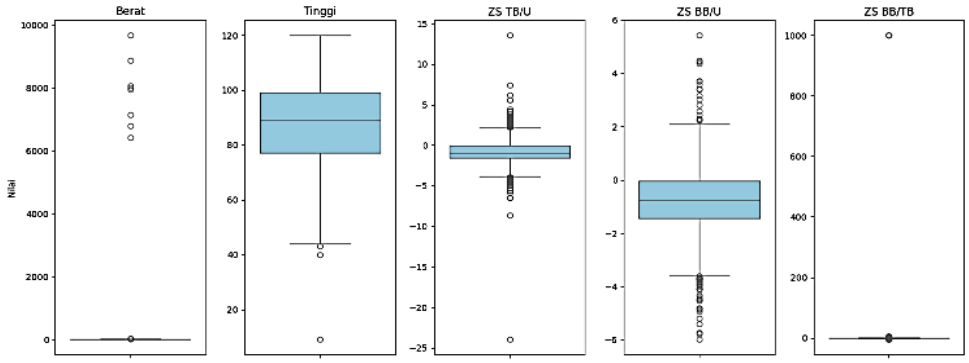
Gambar 3. Explore Data Analyst Stunting vs Normal

Based on Figure 3, it can be seen that the distribution of data between the stunting and normal categories is 7% each. This distribution indicates the presence of data imbalances, which can affect the performance of the prediction model, especially if the algorithm used is sensitive to the distribution of the data. Therefore, steps such as sampling or adjusting the data distribution will be carried out to ensure the data is balanced before being applied to the model. In addition, this exploratory analysis also aims to identify potential anomalies or outliers in the data that can affect the preprocessing process and the final result of the model.

3.2 *Data Preprocessing*

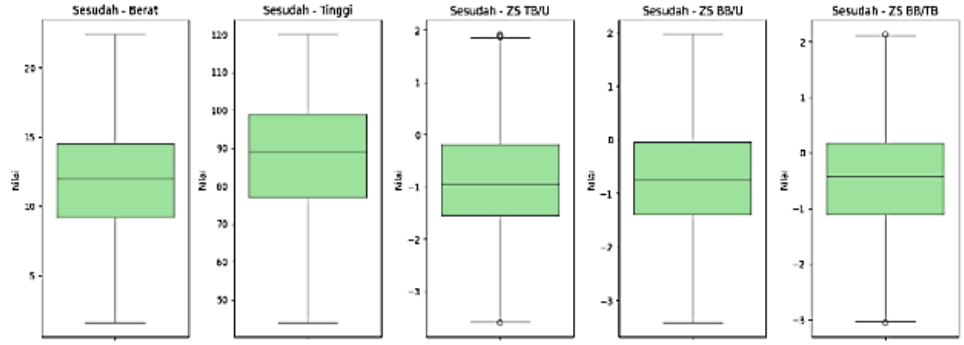
The data preprocessing carried out in this study involves several important stages to prepare data before being used in Machine Learning model training. The first step is the selection of relevant features to train the model, where only those features that are considered significant for this study are selected. In addition, the data cleaning process is also carried out by removing NaN values and handling missing values so that the dataset becomes more consistent and ready for further processing. Furthermore, data labeling was carried out for the stunting and normal categories, which was then used to define independent (X) and dependent (Y) variables. Although there are 33 columns in the dataset, only 5 columns were selected as the main relevant features in this study, namely: 'Weight', 'Height', 'ZS TB/U', 'ZS BB/U', and 'ZS BB/TB'. The following is a visualization of the distribution of numerical data after the preprocessing process, which illustrates the distribution of values on each of the features used.

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2231−2239
ISSN 2684-8910 (print media)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.6670



Gambar 4. *Outliner Preprocessing*

It can be seen in Figure 4 that the distribution of data before preprocessing is very uneven, with some data categories appearing to be more dominant than others. This imbalance can affect the performance of the model in predicting minority classes, so preprocessing steps are needed to correct this condition. After the preprocessing process is carried out, the initially unbalanced data becomes more evenly distributed and more suitable for model training. The following is a visualization of the distribution of data after preprocessing, which shows a more balanced distribution of data and is ready for further processing.



Gambar 5. *Outliner Preprocessing*

3.3 *Data Sampling and K-Fold*

To handle imbalanced data, it is necessary to conduct data sampling. The method used in data sampling is Adaptive Synthetic Sampling (ADASYN). This method generates synthetic data samples for data in minority classes in an unbalanced dataset. By generating synthetic data, the distribution of data between the majority and minority classes becomes more balanced, which helps to improve the performance of the model. After data sampling, the next process is Stratified K-Fold Cross Validation (SKCV), where this method divides the data into a number of folds evenly for each fold. With SKCV, each fold has a similar class distribution to the original dataset, which ensures that the model is tested on representative data. This process will later be used in the implementation of machine learning algorithm models, such as Random Forest and Naive Bayes, to improve classification performance [17], [18] , [19], [20].

3.4 Implementasi *Random Forest*

The machine learning algorithm used in this study is Random Forest, which is applied using the sklearn library with Python. The data used has gone through a folding process with parameters n_estimator = 10 and random_state = 42 to ensure the stability and accuracy of the prediction results. Random Forest is very suitable for handling unbalanced data, because when making predictions in minority classes, there is often interference or noise. However, Random Forest tends to be more resistant to noise because the predicted results are a combination of many decision trees. Thus, the resulting final decision becomes more stable and more accurate. The following are the results of the Random Forest algorithm applied to the data that has gone through the sampling and k-fold process, which is presented in the table below [21] , [22] .

Tabel 1. Confussion Matrix Random Forest

		Predict Values	
		Stunting	Normal
Actual Values	Stunting	591	103
	Normal	41	653

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2231−2239
ISSN 2684-8910 (print media)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.6670

Table 1. The Confusion Matrix Random Forest shows the results of the model classification of stunting and normal status data. This model successfully classified 591 stunting data correctly as True Positive, which shows that the model is effective in detecting stunting data. However, there were 103 stunting data that were incorrectly classified as normal, which was False Negative. On the other hand, the model managed to correctly classify 653 normal data as True Negative, which indicates that the model can well distinguish normal data. However, there were 41 normal data that were incorrectly classified as stunting, which were referred to as False Positives. These results provide an overview of the model's performance in identifying both classes (stunting and normal), and show that although the model has good performance, there are still misclassifications in both categories.

3.5 Implementation *of Naïve Bayes*

Rooted in Bayes' theorem, Naive Bayesian classification is a classification method that uses probability and statistics to predict categories from unknown data. In his application, Naive Bayes assumes that each feature is independent, which allows for simpler and more efficient probability calculations. Nonetheless, these models often yield good results even though the assumption of independence is not fully met in most cases. Naive Bayes is also a flexible algorithm, which allows merging with other methods to improve performance and make distribution more even and balanced. Therefore, this approach is often used in various fields, such as natural language processing and data analysis. Below, the table shows the predictions of the two Naive Bayes variants used in this study, to provide a comparative picture of their performance. Below is the prediction table of the two naïve bayes variants used [23], [24], [25].

Tabel 2. Confussion Matrix Naïve Bayes Gaussian

		Predict Values	
		Stunting	Normal
Actual Values	Stunting	381	77
	Normal	17	358

Table 2 shows the classification results obtained from the Naïve Bayes Gaussian model on data with stunting and normal status. The model successfully classified 381 stunting data correctly as True Positive, while 77 stunting data were incorrectly classified as normal, referred to as False Negative. For data with normal status, the model successfully identifies 358 data correctly as True Negative. However, the model also incorrectly classified 17 normal data as stunting, which was referred to as False Positive. The results of this classification provide an overview of the model's performance in distinguishing between the two categories, as well as show potential areas for improvement in data classifications that are more difficult to distinguish.

Tabel 3. Confussion Matrix Naïve Bayes Bernoulli

		Predict Values	
		Stunting	Normal
Actual Values	Stunting	326	196
	Normal	72	239

Table 3 illustrates the performance of the Naïve Bayes Bernoulli model in the classification of stunting and normal status. This model succeeded in correctly classifying as many as 326 stunting data as True Positive, but there were 196 stunting data that were incorrectly classified as normal, which was False Negative. On the other hand, the model managed to correctly classify 239 normal data as True Negative. However, there were 72 normal data that were incorrectly classified as stunting, which were referred to as False Positives. These results suggest that although the Naïve Bayes Bernoulli model can correctly identify most of the data, there are challenges in distinguishing the more difficult data or borderlines between the stunting and normal categories.

3.6 Model Evaluation

To evaluate the performance of the applied model, a comparison was made between the Random Forest algorithm and two Naïve Bayes variants, namely Naïve Bayes Gaussian and Naïve Bayes Bernoulli. Table 4 presents the accuracy and F1-Score results of the three models, which shows how well each model classifies data. This evaluation provides an overview of the effectiveness of each algorithm in completing classification tasks, as well as the performance differences between the two.

Table 4. Accuracy and F1-Score Model

Model	Accuracy	F1-Score
Random Forest	89,62%	89,09%
Naïve Bayes Gaussian	88,72%	88,81%
Naïve Bayes Bernoulli	67,83%	69,72%

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2231−2239
ISSN 2684-8910 (print media)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.6670

In Table 4, the results of the comparison between the Random Forest algorithm and two Naïve Bayes variants, namely Naïve Bayes Gaussian and Naïve Bayes Bernoulli, are displayed. The comparison results show that the Random Forest algorithm has the highest accuracy and F1-Score compared to Naïve Bayes, with an accuracy of 89.62% and an F1-Score of 89.09%. Meanwhile, when viewed from the results of the Naïve Bayes Gaussian evaluation, this model shows an excellent accuracy of 88.72% with an F1-Score of 88.91%, which indicates its relatively stable performance. In contrast, Naïve Bayes Bernoulli showed a lower performance with an accuracy of 67.83% and an F1-Score of 69.74%. This difference in performance can occur because each Naïve Bayes variant has different advantages and uses a different focus. Naïve Bayes Gaussian tends to be more effective on data that has a continuous distribution, while Naïve Bayes Bernoulli is better suited for binary or categorical data.

4. CONCLUSION

This study aims to compare two algorithms, namely Random Forest and Naïve Bayes, in finding the best performance with maximum accuracy and F1-Score values, using stunting data obtained from the Bekasi Regency Health Office. Before comparing algorithms, data management is carried out to overcome the imbalance in the dataset. The data that was originally unbalanced was then corrected using Adasyn and K-fold techniques to ensure the quality and balance of the data used in the model training process. After overcoming the data imbalance, a comparison of machine learning algorithms was carried out which resulted in an accuracy of 89.62% and an F1-Score of 89.09% for Random Forest. The study also compared the performance between Naïve Bayes Gaussian (NBG) and Naïve Bayes Bernoulli (NBB), with NBG accuracy of 88.72% and F1-Score of 81.81%, as well as NBB accuracy of 67.83% and F1-Score of 69.72%. The authors hope that further research, especially in terms of comparing algorithms using unbalanced data with minority class features, can pay more attention to the data preprocessing stage. Especially in ways to handle minority classes so that the data used becomes more feasible and optimal for the model training process.

REFERENCES

[1] H. Hatijar, “The incidence of stunting in infants and toddlers,” *Sandi Husada Health Scientific Journal*Vol. 12 No. 1 pp. 224–229, 2023, doi:10.35816/jskh.v12I1.1019.

[2] N. D. Yanti, F. Betriana, and I. R. Kartika, "Factors Causing Stunting in Children: A Literature Review," *Real In Nursing Journal*, vol. 3, no. 1, pp. 1–10, 2020, doi: 10.32883/rnj.v3i1.447.

[3] D. Husnaniyah, D. Yulyanti, and R. Rudiansyah, "The relationship between maternal education level and stunting incidence," *The Indonesian Journal of Health Science*, vol. 12, no. 1, pp. 57–64, 2020, doi: 10.32528/ijhs.v12i1.4857.

[4] T. A. E. Permatasari, Y. Chadirin, T. S. Yuliani, and S. Koswara, "Empowerment of Posyandu Cadres in Local Food-Based Organic Food Fortication as an Effort to Prevent Stunting in Toddlers," *Journal of Engineering Community Service*, vol. 4, no. 1, pp. 1–10, 2021, doi: 10.24853/jpmt.4.1.1-10.

[5] I. P. Putri, T. Terttiaavini, and N. Arminarahmah, "Comparative Analysis of Machine Learning Algorithms for Predicting Stunting in Children," *MALCOM: Indonesian Journal of Machine Learning and Computer Science*Vol. 4 No. 1 pp. 257–265, Jan. 2024, doi: 10.57152/malkam.v4I1.1078.

[6] Fadellia Azzahra, N. Suarna, and Y. Arie Wijaya, "Application of Random Forest and Cross Validation Algorithms for Stunting Data Prediction," *Kopertip : Scientific Journal of Informatics and Computer Management*, vol. 8, no. 1, pp. 1–6, Feb. 2024, doi: 10.32485/kopertip.v8i1.238.

[7] M. G. Daffa and P. H. Gunawan, “Stunting Classification Analysis for Toddlers in Bojongsoang: A Data-Driven Approach,” in *2024 2nd International Conference on Software Engineering and Information Technology (ICoSEIT)*, IEEE, 2024, pp. 42–46. doi: 10.1109/ICoSEIT60086.2024.10497515.

[8] R. Supriyadi, W. Gata, N. Maulidah, and A. Fauzi, "Application of Random Forest Algorithm to Determine the Quality of Red Wine," *E-Business: Scientific Journal of Economics and Business*, vol. 13, no. 2, pp. 67–75, 2020, doi: 10.51903/e-business.v13i2.247.

[9] L. Ratnawati and D. R. Sulistyaningrum, "Application of random forest to measure the severity of disease in apple leaves," *ITS Journal of Science and Art*, vol. 8, no. 2, pp. A71–A77, 2020, doi: 10.12962/j23373520.v8i2.48517.

[10] A. A. Santika, T. H. Saragih, and M. Muliadi, "Application of Likert Scale to the Classification of Brilink Agent Customer Satisfaction Levels Using Random Forest," *JUSTIN (Journal of Information Systems and Technology)*, vol. 11, no. 3, pp. 405–411, 2023, doi:10.26418/justin.v11i3.62086.

[11] M. M. Mutoffar, M. Naseer, and A. Fadillah, "Classification of well water quality using random forest algorithm," *Narrative: National Journal of Research, Applications and Informatics Engineering*Vol. 4 No. 2 pp. 138–146, 2022, doi:10.53580/nartif.v4i2.160.

[12] I. Kurniawan, D. C. P. Buani, A. Abdussomad, W. Apriliah, and R. A. Saputra, "Implementation of Random Forest Algorithm to Determine Raskin Aid Recipients," *Journal of Information Technology and Computer Science*Vol. 10 No. 2 pp. 421–428, 2023, doi:10.25126/jatic.20231026225.

[13] J. Pratama, F. Fauziah, and I. D. Sholihati, "K-Nearest Neighbor and Naive Bayes Method in Determining the Nutritional Status of Toddlers," *Brahmin: Journal of the Application of Artificial Intelligence*, vol. 4, no. 2, pp. 214–221, 2023, doi: 10.30645/brahmana.v4i2.197.g196.

[14] A. F. Watratan and D. Moeis, "Implementation of Naive Bayes Algorithm to Predict the Rate of Spread of Covid-19 in Indonesia," *Journal of Applied Computer Science and Technology*, vol. 1, no. 1, pp. 7–14, 2020, doi: 10.52158/jacost.v1i1.9.

[15] R. Ramadhani and R. Ramadhanu, "Machine Learning Method for Classification of Toddler Nutrition Data with Naïve Bayes, KNN and Decision Tree Algorithms," *Symmetrical: Journal of Mechanical Engineering, Electrical and Computer Science*, vol. 15, no. 1, 2024, doi: 10.24176/simet.v15i1.10679.

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2231−2239
ISSN 2684-8910 (print media)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.6670

[16] B. Rahman, F. Fauzi, and S. Amri, "Perbandingan Hasil Klasifikasi Data Iris menggunakan Algoritma K-Nearest Neighbor dan Random Forest: Comparison of Iris Data Classification Results using the K-Nearest Neighbor and Random Forest Algorithms," *Journal Of Data Insights*, Vol. 1, No. 1, pp. 19–26, 2023, Yogurt: 10.26714/Jodi.V1I1.135.

[17] U. Ungkawa and M. A. Rafi, "Data Balancing Techniques Using the PCA-KMeans and ADASYN for Possible Stroke Disease Cases," *Informatics Online Journal*, vol. 9, no. 1, pp. 138–147, Jun. 2024, doi: 10.15575/join.v9i1.1293.

[18] C. G. Tekkali and K. Natarajan, "An advancement in AdaSyn for imbalanced learning: An application to fraud detection in digital transactions," *Journal of Intelligent & Fuzzy Systems*, vol. 46, pp. 11381–11396, 2024, doi: 10.3233/JIFS-236392.

[19] S. Prusty, S. Patnaik, and S. K. Dash, "SKCV: Stratified K-fold cross-validation on ML classifiers for predicting cervical cancer," *Frontiers in Nanotechnology*, vol. 4, Aug. 2022, doi: 10.3389/fnano.2022.972421.

[20] S. Szeghalmy and A. Fazekas, "A Comparative Study of the Use of Stratified Cross-Validation and Distribution-Balanced Stratified Cross-Validation in Imbalanced Learning," *Sensors*, vol. 23, no. 4, Feb. 2023, doi: 10.3390/s23042333.

[21] A. Nugroho and D. Harini, "Random Forest Techniques to Improve Unbalanced Data Accuracy," *JSTIK*, vol. 2, no. 2, 2024, doi: 10.53624/jsitik.v2i2.XX.

[22] Z. P. Agusta and Adiwijaya, "Modified balanced random forest for improving imbalanced data prediction," *International Journal of Advances in Intelligent Informatics*, vol. 5, no. 1, pp. 58–65, Mar. 2019, doi: 10.26555/ijain.v5i1.255.

[23] Y. Yusnida Lase *et al.*, "Bulletin of Information Technology (BIT) Predicting the Impact of Hybrid Learning Using Naive Bayes," vol. 4, no. 4, pp. 425–429, 2023, doi: 10.47065/bit.v3i1.

[24] N. S. Abd and D. A. Abdullah, "Diagnose of Chronic Kidney Diseases by Using Naive Bayes Algorithm," *Journal of Al-Qadisiyah for Computer Science and Mathematics*, vol. 13, no. 2, Jul. 2021, doi: 10.29304/jqcm.2021.13.2.819.

[25] I. Cholissodin *et al.*, "Development of big data app for classification based on map reduce of naive Bayes with or without web and mobile interface by RESTful API using Hadoop and spark," *Journal of Information Technology and Computer Science*, vol. 5, no. 3, pp. 302–312, 2020, doi: 10.25126/jitecs.202053233.