OPEN ACCESS

Volume 9 (4), October-December 2025, 1365-1375

E-ISSN:2580-1643

Jurnal JTIK (Jurnal Teknologi Informasi dan Komunikasi)

DOI: https://doi.org/10.35870/jtik.v9i4.3853

# Comparative Analysis of Machine Learning Models for Stunting Prediction in Jakarta

Ferdinand Marudut Tua Pane [1], Djarot Hindarto [2*]

*[1,2*] Study Program Informatics Engineering, Faculty of Communication and Information Technology, University Nasional, South Jakarta City, Special Capital Region of Jakarta, Indonesia.*

a r t i c l e i n f o

*Article history:*
Received 24 February 2025
Received in revised form
20 March 2025
Accepted 1 May 2025
Available online October
2025.

*Keywords:*
Stunting; Naive Bayes; Stunting Prediction; Data Mining; Machine Learning; Jakarta.

*Kata Kunci:*
Stunting; Naive Bayes; Prediksi Stunting; Data Mining; Pembelajaran Mesin; Jakarta.

a b s t r a c t

Stunting is one medical problem that inhibits a baby's growth. Prompt diagnosis is essential to prevent long-term harm. This study compares machine learning techniques, including Naïve Bayes, Decision Tree, Random Forest, SVM, and ensemble methodologies, in order to improve prediction accuracy. Information on 1,723 children in Jakarta, including age, height, gender, family health history, household income, access to health services, and hygienic circumstances, is included in this dataset, which was collected from Riskesdas and hospital and clinic medical records. To improve model performance, SMOTE, feature selection, and normalization techniques were used. The ensemble approach combined Naïve Bayes with Decision Trees via stacking. The assessment findings indicated that Random Forest had the best accuracy (98%), followed by ensemble technique and Decision Tree (97%), while Naïve Bayes and SVM had lesser accuracy (38% and 37%). This model can assist the government in early intervention to prevent stunting.

a b s t r a k

Stunting adalah salah satu masalah medis yang menghambat pertumbuhan bayi. Diagnosis yang cepat sangat penting untuk mencegah bahaya jangka panjang. Penelitian ini membandingkan teknik pembelajaran mesin, termasuk Naïve Bayes, Decision Tree, Random Forest, SVM, dan metode ensemble, untuk meningkatkan akurasi prediksi. Informasi mengenai 1.723 anak di Jakarta, termasuk usia, tinggi badan, jenis kelamin, riwayat kesehatan keluarga, pendapatan keluarga, akses terhadap layanan kesehatan, dan sanitasi, termasuk dalam dataset ini, yang dikumpulkan dari Riskesdas dan rekam medis rumah sakit dan klinik. Untuk meningkatkan kinerja model, teknik SMOTE, seleksi fitur, dan normalisasi digunakan. Pendekatan ensemble menggabungkan Naïve Bayes dengan Decision Trees melalui stacking. Temuan penilaian menunjukkan bahwa Random Forest memiliki akurasi terbaik (98%), diikuti oleh teknik ensemble dan Decision Tree (97%), sedangkan Naïve Bayes dan SVM memiliki akurasi yang lebih rendah (38% dan 37%). Model ini dapat membantu pemerintah dalam melakukan intervensi dini untuk mencegah stunting.

*Corresponding Author. Email: djarot.hindarto@civitas.unas.ac.id [2*].

ACM Computing Classification System (CCS)
EBSCOhost
Communication and Mass Media Complete (CMMC)

Copyright 2025 by the authors of this article. Published by Lembaga Otonom Lembaga Informasi dan Riset Indonesia (KITA INFO dan RISET). This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

## 1. Introduction

Stunting is a condition of growth failure in children under five due to chronic malnutrition, repeated infections, and inadequate psychosocial stimulation. This condition has a significant impact on brain development, intelligence, and potential productivity in adulthood. Based on the results of the Indonesian Nutrition Status Survey (SSGI) in 2022, the prevalence of stunting in Indonesia decreased from 24.4% in 2021 to 21.6% in 2022 (Tarmizi, 2023). Jakarta, as the nation's capital and economic center, plays a strategic role in efforts to reduce stunting. Although the prevalence of stunting in Jakarta is lower than the national average, challenges remain, especially in areas with high population density and low socioeconomic levels. Factors influencing children's nutritional status include access to healthcare, parenting practices, and sanitation (Nazella & Pujonarti, 2024). To combat stunting, the DKI Jakarta Provincial Government has implemented several initiatives, such as enhancing sanitation, promoting nutrition education, and increasing access to healthcare (Pemerintah Provinsi DKI Jakarta, 2023). Despite these efforts, early identification of children at risk for stunting remains a challenge. In this regard, machine learning, particularly through its ability to predict and detect risk factors with high precision, can offer valuable support.

Previous studies have applied machine learning techniques such as Decision Tree, Naïve Bayes, and k-Nearest Neighbors to identify stunting prevalence (Byna, 2020). Moreover, the Random Forest algorithm has been used to predict stunting in toddlers with an accuracy of 97.87% using 10-fold cross-validation (Perdana *et al.*, 2023). An ensemble learning approach, which combines multiple machine learning algorithms, has also been shown to improve the accuracy of predicting stunting and the nutritional status of children under five years old (Husaini *et al.*, 2023). This research aims to develop a stunting risk prediction model for toddlers in Jakarta using various machine learning methods. The study will compare the performance of algorithms such as Naïve Bayes, Decision Tree, Random Forest, Support Vector Machine (SVM), and ensemble methods (Stacking). These algorithms were selected due to their capability to handle complex data characteristics, class imbalances, and their potential to provide interpretable results for policymakers.

## 2. Research Methodology

Stunting is a critical global health issue that impairs children's physical growth and cognitive development. Machine learning methods such as Naïve Bayes, SVM, Decision Tree, Random Forest, and Ensemble Techniques are employed to enhance prediction accuracy and identify key risk factors. These models offer robust analytical capabilities while ensuring adaptability in health data applications.

### Data Preprocessing

Before starting the analysis, several preparatory steps were undertaken to ensure the integrity and quality of the dataset. First, the data underwent cleaning using various approaches to enhance its reliability. Missing values in continuous numerical attributes, such as height (in cm), were addressed through mean imputation, where missing values were replaced with the average of the available data, preserving the dataset's statistical characteristics. For income data, which often exhibits skewed distributions, median imputation was applied as it is a more reliable measure of central tendency in such cases. Categorical variables, including gender and access to healthcare, were handled using mode imputation, where the most frequent value in each column was used to fill in missing entries. Additionally, duplicate data were detected and removed to reduce bias in the model's outcomes. To further improve model performance, new features were generated through feature engineering, ensuring that the dataset provided comprehensive information for the analysis. Body Mass Index (BMI) was calculated using the formula:

$$BMI = \frac{(Height(m))^2}{Weight(kg)}$$

Categorical variables, such as gender and access to healthcare, were encoded using Label Encoding to convert them into numerical representations suitable for machine learning models. This transformation allowed the categorical data to be processed effectively by the algorithms. Additionally, numerical

characteristics were scaled using Min-Max Normalization, which transformed the data into a range of [0, 1]. This normalization strategy was implemented to ensure that all features contributed equally to the machine learning models, preventing any particular feature from dominating the results and ensuring a fair comparison between variables.

$$X_{normalized} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

The dataset exhibited a class imbalance, with a significantly lower number of severely stunted children compared to those with normal nutritional status. This imbalance could lead to biased models that favor the majority class, potentially undermining the model's ability to accurately predict the minority class. To address this issue, the Synthetic Minority Over-sampling Technique (SMOTE) was applied to the entire training set. SMOTE works by generating synthetic samples through interpolation between existing instances of the minority class, thus balancing the class distribution without causing data leakage. The SMOTE settings were adjusted to create enough synthetic samples for the minority class, ensuring its representation was equal to that of the majority class, thereby improving the model's ability to predict both classes more accurately.

**Model Training and Hyperparameter Tuning**

In this study, several machine learning models were employed, including Naïve Bayes, Decision Tree, Random Forest, Support Vector Machine (SVM), and Ensemble Methods (Stacking). Hyperparameter tuning was performed to optimize the performance of each model. For Naïve Bayes (GaussianNB), no hyperparameters were adjusted due to its probabilistic nature and the assumption of feature independence. For the Decision Tree model, hyperparameters were set with a maximum depth of 10 and a minimum samples split of 2 to control tree growth and avoid overfitting. The Random Forest model was configured with 100 trees, a maximum depth of 10, and a minimum samples split of 2 to ensure diverse and effective tree building while managing overfitting. The SVM model utilized a Radial Basis Function (RBF) kernel, with a regularization parameter (C) set to 1.0 and the Gamma parameter set to 'scale' to balance bias and variance. Finally, the Ensemble Method employed a stacking approach, combining Naïve Bayes and Decision Tree models with Logistic Regression used as the meta-classifier to improve prediction accuracy by leveraging the strengths of each base model.

**Cross-Validation**

Cross-validation was employed to more accurately evaluate the performance of the stunting prediction models. This technique involves dividing the entire dataset into subsets, or folds, and using each fold for both training and testing. Specifically, the dataset was split into k equal parts as part of the k-fold cross-validation process. In each iteration, the model is trained on k-1 folds and tested on the remaining fold, ensuring that the model is assessed multiple times with different data combinations. This approach reduces the potential for bias that could arise from irregular data splits and enhances the model's ability to generalize across unseen data, improving its accuracy in predicting children's nutritional status. Through this iterative process, cross-validation helps ensure that the model's performance is not overly influenced by specific data subsets, providing a more robust evaluation of its predictive capability.

**Prediction**

Prediction refers to the ability of a machine learning model to classify a child's nutritional status based on factors contributing to stunting. The model aims to predict whether a child falls into the normal, stunted, or severely stunted category, using input data such as age, height, family medical history, household income, access to healthcare, and sanitation conditions (Husaini *et al.*, 2023). The models utilized in this study, including Naïve Bayes, Decision Tree, Random Forest, SVM, and ensemble methods, were trained using a dataset of children in Jakarta. These models learn the association patterns between the various factors in the dataset and are then applied to analyze new data to assess the risk of stunting in children (Husaini *et al.*, 2023).

**Machine Learning**

Machine learning plays a pivotal role in evaluating health data and associated variables to determine the risk of malnutrition in children. The models are trained on datasets that include various health, environmental, and socio-economic factors, helping

to identify patterns that can predict stunting risks. The goal is to provide early intervention opportunities and support more effective public health initiatives by offering accurate and reliable prediction tools (Tarmizi, 2023).

**Data Normalization**

Data normalization is the process of adjusting the dataset so that each feature has a comparable range of values. This is typically done by scaling the data to a range of [0, 1] or ensuring the data has a mean of 0 and a standard deviation of 1. The primary purpose of data normalization is to enhance the performance of machine learning algorithms by ensuring that all features contribute equally during model training, preventing any single feature from dominating the learning process (Indrajit, 2023).

**Data Cleaning**

Data cleaning ensures the dataset is of optimal quality before analysis. This process involves handling missing data, such as incomplete height or age information, removing duplicate entries that could introduce bias into predictions, and standardizing the format of the data, such as how gender and nutritional status are presented (Dazki, 2024). Proper data cleaning is essential as it enhances the ability of machine learning models to detect patterns accurately, leading to more reliable predictions of stunting risk and better decision-making in public health interventions (Indrajit, 2023).

**Stunting**

Stunting is a result of chronic malnutrition that affects millions of children globally. It leads to impaired growth and long-term developmental problems, highlighting the importance of early intervention strategies to address and prevent the condition (Pratama *et al.*, 2024).

**Data Mining**

Data mining techniques are instrumental in extracting valuable insights from large datasets. The use of algorithms such as Naïve Bayes (NB), Decision Tree (DT), Random Forest (RF), and Support Vector Machine (SVM) makes the data analysis process more effective in identifying trends and patterns. Without proper data mining techniques, vast amounts of data could remain underutilized, so applying the correct methods ensures the data produces meaningful insights for stunting prediction and intervention strategies (Saputra *et al.*, 2023).

**Classification**

Classification is an effective approach for identifying a child's nutritional condition using various algorithms, including Random Forest (RF), Support Vector Machine (SVM), Decision Tree (DT), and Naïve Bayes (NB). Each algorithm offers distinct advantages. Naïve Bayes excels in handling data with many features, assuming feature independence. Decision Trees are easy to interpret due to their simple structure, while Random Forest enhances accuracy by combining multiple decision trees. SVM is effective for high-dimensional data and provides optimal separation margins (Saputra *et al.*, 2023).

**Naïve Bayes**

Naïve Bayes is a probabilistic classification method based on Bayes' theorem. It assumes that all features in the dataset are independent, enabling fast and efficient probability calculations. Although Naïve Bayes is highly effective for tasks like text classification and works well with small datasets, its assumption of independence between features may not hold true in real-world data, potentially limiting its performance in more complex scenarios (Indrajit, 2023).

**Ensemble Methods**

Ensemble methods combine multiple base models to improve prediction accuracy and robustness compared to individual classifiers. By reducing bias and variance, ensemble learning promotes model stability and generalization. This approach leverages the diversity of multiple weak learners to create a more powerful and resilient prediction framework, making it particularly useful for handling complex and high-dimensional data. Ensemble learning is effective in enhancing the overall performance of models, ensuring better classification results, and increasing adaptability (Aziz, 2021).

**Decision Tree**

The Decision Tree algorithm, introduced by J. Ross Quinlan in 1975, is a widely used classification technique due to its interpretability and systematic

decision-making structure. Also known as Classification and Regression Tree (CART), it transforms data into decision rules, simplifying complex problems into more manageable, structured formats. The tree consists of a Root Node (the starting point), Internal Nodes (decision points that lead to multiple outputs), and Leaf Nodes (final outcomes). Its strength lies in breaking down complex decisions into simple binary choices, making it particularly effective for categorization, risk assessment, and predictive modeling (Indrajit, 2023).

**Support Vector Machine (SVM)**

Support Vector Machine (SVM), introduced by Boser, Guyon, and Vapnik in 1992, is a classification system that seeks to find the optimal hyperplane that maximizes the margin between distinct classes. The hyperplane serves as a boundary, with support vectors being the data points closest to the decision boundary. For non-linear data, the Kernel Trick is used to map the data into a higher-dimensional space, facilitating better class separation. While SVM is highly effective at managing high-dimensional data and achieving strong classification accuracy, its main limitations include the difficulty of designing appropriate kernel functions and the computational inefficiency when dealing with large datasets (Indrajit, 2023).

**Random Forest**

Random Forest is an excellent method for classification in your research due to its ability to improve accuracy, reduce overfitting, and handle data sets with multiple features. Using this technique, you can obtain more reliable classification results compared to single methods such as Decision Tree.

**Dataset**

The use of appropriate datasets provides significant benefits in machine learning research. With quality datasets, the model training process can be optimized, improving classification accuracy using algorithms such as Naive Bayes, Decision Tree, Random Forest, and Support Vector Machine. This research illustrates that proper selection, preparation, and processing of datasets can be a solution in overcoming data limitations in various application areas.

**Machine Learning**

Machine learning operates by analyzing historical data, recognizing patterns, and constructing models capable of predicting outcomes from new data. These models learn from past experiences in the data, allowing them to generalize and make predictions about unseen instances. In the context of predicting stunting in Indonesia, various machine learning techniques, such as Random Forest Regression, Linear Regression, and Decision Tree Regression, are employed.

These methods help forecast the likelihood of stunting in children by identifying patterns and relationships within the data that relate to nutritional status, socio-economic factors, and healthcare access, among other variables. Developing a prediction model involves using different algorithms, each with its own strengths and working mechanisms. Random Forest, however, has demonstrated the highest performance due to its ability to capture non-linear correlations and interactions among features. This machine learning approach is particularly well-suited for evaluating complex public health data, as it can independently manage large datasets and provide accurate predictions. These capabilities make Random Forest a powerful tool for guiding the development of health policies by offering reliable, data-driven insights into factors like stunting and malnutrition (Aziz, 2021).

**Stacking**

Stacking is an ensemble learning strategy that aims to improve prediction accuracy by combining multiple machine learning algorithms. The method involves training several base learners on the same dataset and then utilizing a meta-learner model to integrate the predictions from these base models. This approach capitalizes on the strengths of each algorithm while mitigating their individual weaknesses, thus enhancing the overall performance of the model. Stacking is particularly beneficial when different algorithms bring diverse strengths to the table, resulting in a more robust and accurate prediction model.

## 3. Results and Discussion

### Results

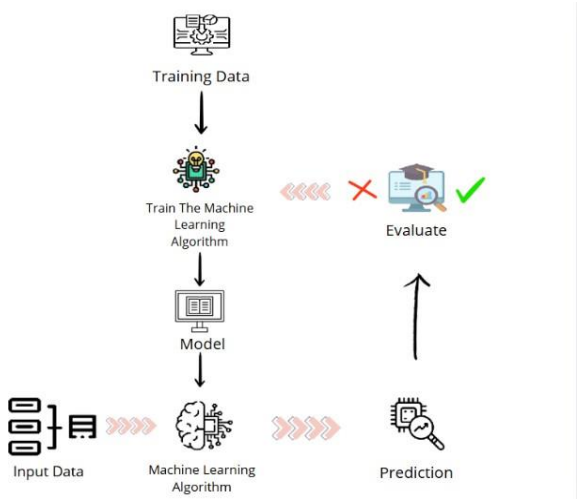


Figure 1. Flowchart

### Evaluation

Evaluation metrics are critical to assess the performance and effectiveness of a machine learning model. Several key metrics are commonly used to evaluate classification models:

1) True Positive (TP)
   This refers to the number of instances that are correctly identified as belonging to the positive class. It represents the successful identification of relevant instances.
2) False Positive (FP)
   This refers to the number of instances that are incorrectly classified as positive, when they actually belong to the negative class. It indicates a false alarm or misclassification in the positive class.
3) False Negative (FN)
   This represents the number of instances that are positive but were incorrectly identified as negative. False negatives are particularly important in applications where missing a positive instance (e.g., a stunted child) is critical.
4) Recall
   Recall is a metric that quantifies the model's ability to correctly identify all positive instances.
5) F1-score
   The F1-score is the harmonic mean of precision and recall, providing a single metric that balances both concerns. It is particularly useful in cases where there is an uneven class distribution or when the costs of false positives and false negatives are different.
6) Accuracy
   Accuracy is a fundamental metric that quantifies the overall correctness of a classification model. It is defined as the ratio of correctly predicted instances (both positive and negative) to the total number of instances in the dataset.

Presicion:

$$\text{Precision} = \frac{TP}{TP + FP} = \frac{118}{118 + 7} = 0.94$$

Recall:

$$\text{Recall} = \frac{\text{TP}}{TP + FN} = \frac{118}{118 + 0} = 1.00$$

F1-score:

$$F1 - \text{Score} = 2x\frac{\text{Precision x Recall}}{\text{Precision} + \text{Recall}} = 2x\frac{0.94 \text{ x } 1.00}{0.94 + 1.00} = 0.97$$

For the Normal Class, the Precision is calculated as the ratio of true positives (TP) to the sum of true positives (TP) and false positives (FP). In this case, the precision is 0.94, as there are 118 true positives and 7 false positives. The Recall for this class is 1.00, as there are 107 true positives and no false negatives. The F1-score is the harmonic mean of precision and recall, which in this case is calculated as 0.98, using the formula:

$$\text{F1-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2 \times 0.91 \times 1.00}{0.91 + 1.00} = 0.98$$

For the Stunted Class, precision is calculated in a similar manner, yielding a precision of 0.94, and the recall is 1.00, indicating perfect identification of stunted children. The F1-score for this class is also 0.97, calculated as:

$$\text{F1-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2 \times 0.94 \times 1.00}{0.94 + 1.00} = 0.97$$

Finally, Accuracy represents the overall correctness of the model, calculated by the sum of true positives (TP) and true negatives (TN) divided by the total

number of instances. This metric gives an overall performance score, but in cases of imbalanced data, precision, recall, and F1-score are often more insightful.

Table 1. Random Forest Model Results

| Status Gizi | *precision* | *recall* | *f1-score* | *Accuracy* |
|---|---|---|---|---|
| Normal | 0.98 | 1.00 | 0.99 | |
| *Severely Stunted* | 0.98 | 0.97 | 0.97 | 0.98 |
| *Stunted* | 0.99 | 0.97 | 0.98 | |

Table 2. SVM Model Results

| Status Gizi | *precision* | *recall* | *f1-score* | *Accuracy* |
|---|---|---|---|---|
| Normal | 0.40 | 0.25 | 0.31 | |
| *Severely Stunted* | 0.39 | 0.56 | 0.46 | 0.37 |
| *Stunted* | 0.32 | 0.29 | 0.30 | |

Table 3. Naïve Bayes Model Results

| Status Gizi | *precision* | *recall* | *f1-score* | *Accuracy* |
|---|---|---|---|---|
| Normal | 0.43 | 0.28 | 0.34 | |
| *Severely Stunted* | 0.35 | 0.45 | 0.39 | 0.38 |
| *Stunted* | 0.39 | 0.42 | 0.40 | |

Table 4. Decision Tree Model Results

| Status Gizi | *precision* | *recall* | *f1-score* | *Accuracy* |
|---|---|---|---|---|
| Normal | 0.94 | 1.00 | 0.97 | |
| *Severely Stunted* | 1.00 | 0.91 | 0.95 | 0.97 |
| *Stunted* | 0.96 | 1.00 | 0.98 | |

Table 5. Ensemble Methode Model Results

| Status Gizi | *precision* | *recall* | *f1-score* | *Accuracy* |
|---|---|---|---|---|
| Normal | 0.94 | 1.00 | 0.97 | |
| *Severely Stunted* | 1.00 | 0.91 | 0.95 | 0.97 |
| *Stunted* | 0.96 | 1.00 | 0.98 | |

The Random Forest model achieved the highest accuracy at 98%, followed closely by the Ensemble Method and Decision Tree at 97%. This performance improvement is largely attributed to the ensemble nature of Random Forest, which helps minimize variance and enhances generalization. Additionally, its automated feature selection process and ability to capture non-linear correlations significantly improve its predictive power. On the other hand, Naïve Bayes (38%) and SVM (37%) performed poorly. Naïve Bayes struggled due to its assumption of feature independence, which prevents it from capturing non-linear feature interactions. Meanwhile, SVM's underperformance is likely linked to the complexity of the dataset, its sensitivity to high-dimensional data, and the need for extensive hyperparameter tuning. A key consideration in this research is the trade-off between accuracy and memory. The Random Forest model demonstrated balanced performance across all metrics, making it particularly suitable for stunting prediction, where high recall is essential to identify at-risk children accurately. In contrast, Naïve Bayes and SVM, with their lower recall, are less useful for this purpose. The Ensemble Method, which combines Naïve Bayes and Decision Tree, offered balanced performance. However, its slightly lower accuracy compared to Random Forest suggests that further fine-tuning of the meta-classifier could improve its effectiveness. These findings underscore the importance of selecting machine learning methods that align with the specific characteristics of the dataset. Future research could explore hybrid ensemble techniques and deep learning models to further enhance prediction accuracy and support more effective public health interventions.

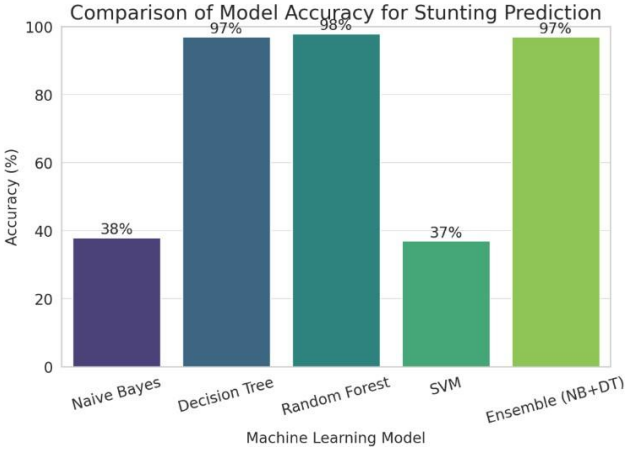


Figure 2. Comparison of Model

Figure 2 compares the accuracy of five machine learning models used for stunting prediction in Jakarta. The Random Forest model achieved the highest accuracy at 98%, making it the most reliable model for classifying the data. Both the Decision Tree and the Ensemble Method (Naïve Bayes + Decision Tree) performed very well, each attaining an accuracy of 97%, which suggests that combining models can significantly enhance prediction results. On the other hand, Naïve Bayes and Support Vector Machine (SVM) displayed much lower accuracy rates of 38% and 37%, respectively. These results highlight the

challenges these models face in handling the complexity of socioeconomic and health-related factors within the dataset, which may hinder their ability to produce reliable predictions.

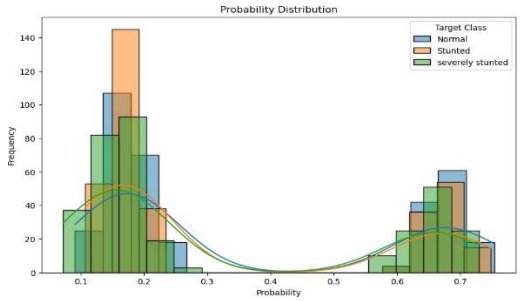


Figure 3. Probabiliy Distribution

Figure 3 shows the probability distribution of the model's predictions for the three target classes: Normal, Stunted, and Severely Stunted. The histogram, differentiated by colors, represents the probability distribution for each class, while the connecting lines display the Kernel Density Estimation (KDE), which offers a smoother visualization of the distribution. From the graph, it can be observed that the probability predictions for each class tend to form two peaks, indicating a bimodal distribution. This suggests that the model tends to assign higher probabilities to two specific ranges, which may reflect the underlying patterns in the training data distribution. This probability distribution is critical for evaluating the model's capability to classify the data accurately. For instance, in a given probability range, if the Normal class has a higher probability than the other classes, the model is more confident in classifying the data as Normal. This insight is important for understanding how well the model differentiates between the classes and its decision-making process when classifying new instances.

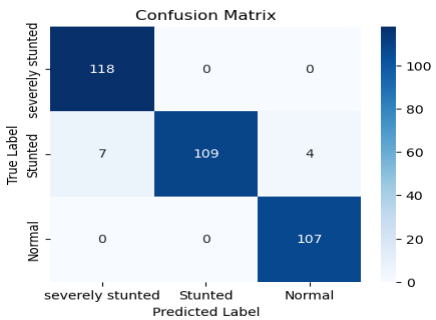


Figure 4. Confusion Matrix

Figure 4 presents the confusion matrix, which categorizes children into three groups: Severely Stunted, Stunted, and Normal. This matrix examines how well the classification model predicts the degree of stunting in children. The diagonal values indicate the correctly classified instances: 118 severely stunted children, 109 stunted children, and 107 normal children, showing that the model effectively distinguishes between the three categories. These results highlight the model's strong performance in correctly identifying each group. However, some misclassifications are observed, particularly within the Stunted category. Seven instances were incorrectly classified as Severely Stunted, and four instances were misclassified as Normal. These misclassifications suggest that there may be overlaps in the feature representation between the stunted and the other two categories. To improve the model's accuracy, enhancements in feature engineering, such as adding more relevant attributes, could help. Additionally, addressing class imbalance using methods like Synthetic Minority Over-sampling Technique (SMOTE) and optimizing hyperparameters through techniques such as fine-tuning the SVM kernel functions or adjusting the Decision Tree depth would further improve the model's performance. Overall, the confusion matrix analysis shows that the model achieves high classification accuracy with minimal misclassification. Future improvements in data preprocessing and model optimization are expected to enhance both the performance and robustness of the model, leading to more accurate predictions.

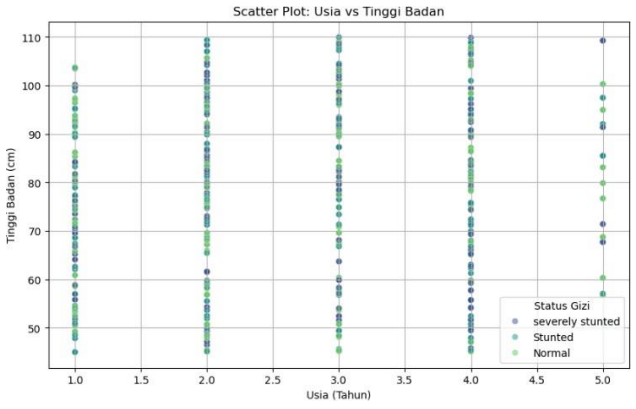


Figure 5. Scatter Plot: Age with Height

Figure 5, the scatter plot, illustrates the relationship between children's height and age. However, there is considerable variation in this relationship due to

disparities in nutritional status. Children with average nutritional status tend to be taller, while severely stunted children are notably shorter. This variation highlights the impact of factors such as nutrition, family economics, access to healthcare, and sanitation on children's growth. The high number of stunted children suggests that these socio-economic and environmental factors significantly influence physical development. To combat stunting, early intervention is essential, including improving nutritional intake, promoting more effective parenting practices, and increasing access to healthcare. Addressing these factors is crucial in reducing the prevalence of stunting in children. The distinct differences observed between the various nutritional status groups provide valuable insights for building more accurate stunting prediction models. Algorithms such as Naïve Bayes, Decision Tree, Random Forest, and SVM can be employed to analyze these patterns further. A deeper investigation into the data patterns is expected to lead to the development of reliable and effective stunting prediction models, offering critical insights for interventions and the prevention of nutritional problems in children.

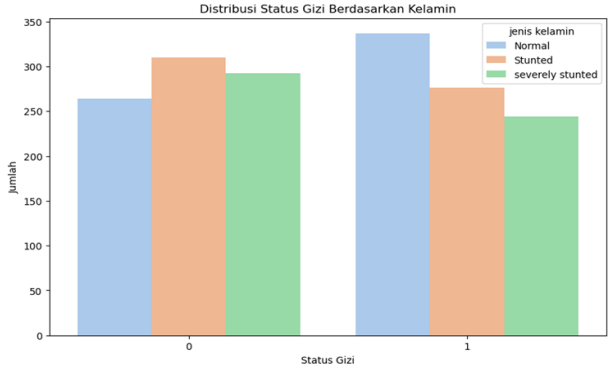


Figure 6. Distribution by gender

Figure 6 presents the distribution of nutritional status by gender, showing that there are more children with normal nutritional status than those who are stunted and severely stunted. In addition, a comparison between the sexes shows that both boys and girls have similar proportions in the nutritional status categories, although there is a slight variation in the number of children in each category. This indicates that gender does not have a significant influence on nutritional status in this dataset.

**Discussion**

This study focuses on the use of machine learning techniques to predict the risk of stunting in children in Jakarta. Based on the experimental results, the Random Forest model demonstrated the best performance with an accuracy of 98%, surpassing other models such as Naïve Bayes and SVM, which achieved accuracies of only 38% and 37%, respectively. The success of Random Forest in this study can be attributed to its ability to handle data with interacting and non-linear features. This is particularly relevant since factors influencing stunting are complex and interconnected, such as family nutrition status, access to healthcare, and sanitation conditions (Tarmizi, 2023; Pujonarti & Nazella, 2024). By utilizing ensemble techniques, this model also improved predictive accuracy by combining multiple decision trees, minimizing variance, and reducing overfitting. Meanwhile, the ensemble method that combines Naïve Bayes and Decision Tree also yielded good results with an accuracy of 97%. This stacking approach leverages the strengths of both models, where Naïve Bayes efficiently handles data with numerous features, and Decision Tree provides a clearer and simpler interpretation for decision-making. However, its performance is slightly lower than that of Random Forest due to its reliance on the combination of two simpler models, as compared to Random Forest's multiple decision trees that complement each other. On the other hand, Naïve Bayes and SVM showed unsatisfactory performance with very low accuracy.

This can be explained by the assumption of feature independence in Naïve Bayes, which does not hold in the context of health data where features are highly interdependent, as well as the sensitivity of SVM to high-dimensional data and class imbalance in the dataset. Stunting, being a health issue, is often not uniformly distributed, with fewer cases of severely stunted children compared to those with normal nutritional status (Government of DKI Jakarta, 2023). To address this class imbalance, this study applied the SMOTE (Synthetic Minority Over-sampling Technique) to generate synthetic samples to balance the classes. However, the effect of SMOTE on Naïve Bayes and SVM performance remained limited due to the models' inability to manage complex feature interactions effectively. The results highlight the

importance of selecting models appropriate to the characteristics of the dataset used. Random Forest, which excels at handling complex data with interacting features, shows that ensemble-based approaches are more effective in stunting prediction in Jakarta. This model is better at addressing class imbalance, while models like Naïve Bayes and SVM require further adjustment to perform better on such complex datasets (Byna, 2020; Perdana *et al.*, 2023). The success of Random Forest in predicting stunting risk opens the opportunity for implementing machine learning-based prediction models in public policy, particularly in preventing stunting in children in Jakarta. The DKI Jakarta Provincial Government has launched several initiatives to address stunting, such as improving sanitation, nutrition education, and healthcare access (Government of DKI Jakarta, 2023). Integrating machine learning-based prediction models could help accurately identify children at high risk, enabling faster and more targeted interventions. This study also suggests future development using deep learning models or hybrid ensemble methods to improve prediction accuracy, especially by utilizing longitudinal data that accounts for children's growth patterns over time (Husaini *et al.*, 2023). Incorporating more detailed information regarding food consumption, environmental conditions, and maternal health could further enhance the model's ability to detect factors contributing to stunting. In the future, these techniques could become powerful tools to support policies aimed at preventing stunting in Indonesia.

## 4. Conclusion

This research proves that Random Forest surpasses other machine learning algorithms in predicting stunting risk in children, attaining the utmost accuracy of 98%. The Ensemble Method presents a practical alternative by blending complementing algorithms, while Naïve Bayes and SVM demonstrate limited performance due to their underlying assumptions and the dataset's complexity. The study's findings have significant ramifications for public health policy and provide a data-driven approach to enhance the early detection of stunting problems. Particularly in areas where stunting is very common, including prediction models into health information systems might support more efficient resource allocation and focused intervention tactics. Future research should concentrate on employing Long Short-Term Memory (LSTM) networks and Gradient Boosting Machines (GBM) to capture temporal growth patterns and increase prediction accuracy. Further supporting public health decision-making would be a more detailed risk assessment framework that includes information on food consumption, environmental variables, and maternal health indicators.

## 5. References

Afarini, N., & Hindarto, D. (2024). Forecasting airline passenger growth: Comparative study LSTM vs Prophet vs neural prophet. *Sinkron: jurnal dan penelitian teknik informatika*, *8*(1), 505-513. https://doi.org/10.33395/sinkron.v9i1.13237.

Aziz, F. (2021). Klasifikasi Aktivitas Manusia menggunakan metode Ensemble Stacking berbasis Smartphone. *Journal of System and Computer Engineering*, *2*(1), 106-111. https://doi.org/10.47650/jsce.v1i2.171.

Byna, A. (2020). Monograf analisis komparatif machine learning untuk klasifikasi kejadian stunting.

Hindarto, D. (2023). Enhancing Road Safety with Convolutional Neural Network Traffic Sign Classification. *Sinkron: jurnal dan penelitian teknik informatika*, *7*(4), 2810-2818. https://doi.org/10.33395/sinkron.v8i4.13124.

Hindarto, D. (2023). Use ResNet50V2 deep learning model to classify five animal species. *Jurnal JTIK (Jurnal Teknologi Informasi dan Komunikasi)*, *7*(4), 758-768.

Hindarto, D., & Hendrata, F. (2024). Development of Machine Learning Model for Breast Cancer Prediction from Ultrasound Images. *Sinkron: jurnal dan penelitian teknik informatika*, *8*(2), 1019-1028. https://doi.org/10.33395/sinkron.v8i2.13593.

Hindarto, D., & Santoso, H. (2021). Plat Nomor Kendaraan dengan Convolution Neural Network. *Jurnal Inovasi Informatika*, *6*(2), 1-12.

Hindarto, D., & Santoso, H. (2022). Performance Comparison of Supervised Learning Using Non-Neural Network and Neural Network. *Jurnal Nasional Pendidikan Teknik Informatika: JANAPATI*, *11*(1), 49-62. https://doi.org/10.23887/janapati.v11i1.40768.

Husaini, A., Hoeronis, I., Lumana, H. H., & Puspareni, L. D. (2023). Early detection of stunting in toddlers based on ensemble machine learning in Purbaratu Tasikmalaya. *Jurnal Sistem dan Teknologi Informasi (JustIN), 11*(3), 487. https://doi.org/10.26418/justin.v11i3.66465.

Mkungudza, J., Twabi, H. S., & Manda, S. O. (2024). Development of a diagnostic predictive model for determining child stunting in Malawi: a comparative analysis of variable selection approaches. *BMC Medical Research Methodology*, *24*(1), 175.

Pratama, M. A. E., Hendra, S., Ngemba, H. R., Nur, R., Azhar, R., & Laila, R. (2024). Comparison of machine learning algorithms for predicting stunting prevalence in Indonesia. *Jurnal Sisfokom (Sistem Informasi dan Komputer), 13*(2), 200–209. https://doi.org/10.32736/sisfokom.v13i2.2097.

Saragih, V. R., Arnita, A., Indra, Z., Taufik, I., & Sinaga, M. S. (2024). Comparison of supervised machine learning methods in predicting the prevalence of stunting in north sumatra province. *Journal of Soft Computing Exploration*, *5*(4), 370-379. https://doi.org/10.52465/joscex.v5i4.498.

Syahfitri, N. A. I., Juledi, A. P., & Muti'ah, R. (2024). Comparative Analysis of Machine Learning Algorithm Performance in Predicting Stunting in Toddlers. *Sinkron: jurnal dan penelitian teknik informatika*, *8*(3), 1452-1462. https://doi.org/10.33395/sinkron.v8i3.13698

Tarmizi, S. N. (2023). Prevalensi stunting di Indonesia turun ke 21, 6% dari 24, 4%. *Sehat Negeriku.*