International Journal of Computer, Network Security and Information System (IJCONSIST)
Vol. 5 No. 1, September 2023

Comparison of C4.5 Decision Tree and Naive Bayes Algorithms for Classification of Nutritional Status in Stunting Toddlers

Ishak Febrianto
Informatika, ilmu komputer
Universitas Pembangunan Nasional Jawa Timur
Surabaya, Jawa Timur, Indonesia
Ishakf735@gmail.com

Anggraini Puspita Sari
Informatika, ilmu komputer
Universitas Pembangunan Nasional Jawa Timur
Surabaya, Jawa Timur, Indonesia
Anggraini.puspita.if@upnjatim.ac.id

Abstract **- Stunting is a condition where growth and development of children under 5 years of age is impaired due to chronic malnutrition. Data mining with classification techniques on the nutritional status of stunting toddlers can be performed to help identify toddlers experiencing stunting and provide objective measurements of their nutritional status. There are several classification methods, but this research will compare the performance of the C4.5 decision tree algorithm, which is included in the decision tree approach, and naive Bayes, which uses a probability-based approach of class occurrence in classifying nutritional status of stunting toddlers, with discretization performed in the preprocessing stage. The data used in this research was obtained from Jagir Health Center, Surabaya, in the form of secondary data on toddler nutrition in 2021, totaling 2,801 records. The labeling of stunting or normal in the dataset uses the reference of child anthropometric standards in Indonesia as stated in the Republic of Indonesia Minister of Health Regulation number 2 of 2020. The best method based on the AUC (Area Under the Curve) value was the C4.5 decision tree with a value of 86% (good classification), while naive Bayes achieved 77% (fair classification) using a 70:30 training and testing data ratio.**

Keywords*: Classification, Decision Tree C4.5, Naive Bayes, Stunting, Data Mining.*

I. INTRODUCTION

Nutritional problems remain a crucial issue today. One of the many existing nutritional problems is stunting. According to WHO (2015), stunting is a condition of malnutrition during child growth and development, characterized by length or height below the standard. Based on studies conducted by Soliman et al. (2021), and also by Dewey & Begum (2011), stunting has serious short-term and long-term effects. A child's nutritional status can be determined using anthropometric measurements as stated in the Republic of Indonesia Minister of Health Regulation number 2 of 2020. If a child's anthropometric z-score is less than minus 2 standard deviations (SD) based on the length-for-age index (LAZ) or height-for-age index (HAZ), then the child is classified as stunting.

Data mining with classification techniques on the nutritional status of stunting toddlers can be performed to help identify toddlers experiencing stunting and provide objective measurements of their nutritional status. There are several classification methods, but this research will develop models using the C4.5 decision tree algorithm, which is included in the decision tree approach, and naive Bayes, which uses a probability-based approach of class occurrence in classifying nutritional status of stunting toddlers.

In research conducted by Herliansyah et al. (2021), the accuracy results of the naive Bayes algorithm method were 64.02% on 90:10 data partition testing and 64.36% accuracy on K-Fold Cross Validation testing in stunting toddler classification. This was achieved using attributes of gender, height, and toddler age. The C4.5 algorithm was also previously researched to identify toddler nutritional status by Firdausia Ismi Nurhayati et al. (2021), achieving 85% accuracy with a 70:30 training and testing data ratio. From these two previously researched methods, this study will analyze how the performance of both models compares in classifying nutritional status of stunting toddlers using toddler nutrition data from 2021 obtained from Jagir Health Center, Surabaya.

II. LITERATURE REVIEW

A. Stunting

Stunting is a condition where growth and development of children under 5 years of age is impaired due to malnutrition. This growth and development period is crucial as it occurs rapidly and only happens once in a lifetime (golden age) (Setyawati, 2012). Stunting is defined as a condition where a child's height or length falls below the threshold (z score) between -3 standard deviations (SD) to <-2 SD based on the Length-for-Age (LAZ) or Height-for-Age (HAZ) index (Agustina, 2022).

B. Data Mining

According to Ian H. Witten et al. (2011), data mining is the process of extracting information from large data using techniques that enable the discovery of previously unknown patterns. Data mining encompasses various algorithms and techniques such as classification, clustering, association rule mining, decision trees, neural networks, and more. Data mining can be applied in various fields such as business, healthcare, education, finance, and others. Several studies show that data mining can be useful in everyday matters. For example, a study by Al Haromainy et al. (2022) shows that data mining techniques can be used to accurately classify Javanese characters. Meanwhile, a study by Saputra et al. (2022) shows that data mining can be used to classify vehicle images into three classes: bus, car, and truck.

C. Decision Tree

According to Han (2011), a decision tree is a machine learning model used to make decisions based on attributes present in a dataset. Decision trees work by dividing the dataset into increasingly smaller subgroups based on predetermined rules. Each subgroup is generated from selecting the most informative predictor variables in distinguishing target variable values in the data. This process is repeated until it's no longer possible to divide the data into smaller subgroups. In the medical field, the main advantage of decision trees is their ability to produce decision criteria that are easily understood by humans. The decision-making process when using decision trees can also be easily validated by experts. For these reasons, decision trees are suitable for use in the medical field (Podgorelec et al., 2002).

D. Algoritma C45

This algorithm was developed by Ross Quinlan in 1993 as an improvement of its predecessor, ID3. C4.5 has the ability to produce optimal decision trees by considering various factors such as attribute selection effectiveness, tree pruning, and handling missing data. The C4.5 algorithm works by splitting the dataset based on attribute values present in the dataset. The splitting process is done by considering information gain from each attribute. The attribute with the highest information gain will be chosen as the attribute used in splitting. The splitting process is repeated until all instances in the sub-dataset belong to one class or all attributes in the dataset have been used in splitting. Here are the formulas used in the C4.5 algorithm to calculate entropy (Han et al., 2011):

$$S) \ = \ \sum (-pi) \ * \ log2 \ pi \ n \ i = 1 \quad (1)$$

Where :

S : Set of cases
n : Number of S partitions
pi : Conditional probability

While to calculate gain in the C4.5 algorithm, the formula is (Han et al., 2011):

$$(S, A) = \ (S) - \sum \ \ |Si| / \ |S| \ * \ (Si) \ n \ i = 1 \quad (2)$$

S : The initial entropy of the dataset before being split by attribute **A**
A : Attribute
n : Number of partitions of attribute **A**
Si : Number of cases in the **i-th** partition
S : Total number of cases in the dataset

E. Naïve Bayes

Naive Bayes is a classification algorithm based on probability theory. This algorithm predicts the class or label of data by calculating the probability of each class based on the features present in the data by assuming all features are independent. This class-conditional independence is made to simplify computation, hence the algorithm is called "naive". The theoretical foundation of Naive Bayes is Bayes' theorem. Bayes' theorem states that the probability of a hypothesis (class or label) can be calculated based on the probability of the data and the conditional probability of the hypothesis against that data. The general equation of Bayes' theorem is (Han et al., 2011):

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)} \quad (3)$$

Where:

X :Data class that is not yet known.
H :Hypothesis of the data.
P(H|X) :Probability of hypothesis **H** given condition **X** (posterior probability).
P(H) :Initial probability of hypothesis **H** (prior probability).
P(X|H) : Probability of **X** given the condition of hypothesis

H(likelihood).

P(X) : Probability of data **X** (evidence).

F. Confusion Matrix

The confusion matrix is one of the methods that can be used to measure the effectiveness of a classification method. The confusion matrix contains information comparing classification results by comparing actual values with the classification results produced by the model or system. The confusion matrix can be used to calculate various evaluation metrics such as accuracy, precision, recall, F1-score, and others (Han et al., 2011).

TABLE I. CONFUSION MATRIX

	Actual Positive	*Actual Negative*
Predicted Positive	TP	FP
Predicted Negative	FN	TN

In Table 1, there are several terms that represent classification results, namely True Positive (TP), True Negative (TN), False Negative (FN), and False Positive (FP). From the confusion matrix, other test metrics can be determined, including accuracy, precision, recall, F1-score, True Positive Rate (TPR), and False Positive Rate (FPR) to construct the Receiver Operating Characteristic (ROC) curve. The AUC (Area Under the Curve) value from the ROC graph can also be used as a measure of the classification system's performance. The larger the AUC value, the better the performance of the classification system. The quality of classification can be categorized based on the AUC value as follows: 90%-100%:Excellent, 80%-90%:Good, 70%-80%:Fair, 60%-70%:Poor, <60%:Failure. (Divva Meuthia Zulma & Chamidah, 2021).

III. METHOD

Figure 1 illustrates the process flow to be undertaken to achieve the main objective of this study, which is a comparative analysis of the C4.5 and Naive Bayes methods in determining the nutritional status of stunted toddlers. The system design stages in this research are divided into several phases, which are explained in the research procedures and will later be implemented in various data partitioning scenarios.

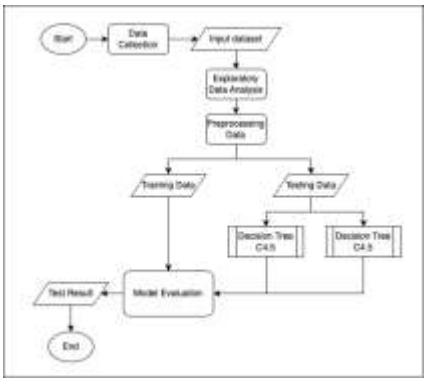

Fig. 1. Research Flowchart

A. Data Collection

The data collected comes from Puskesmas Jagir, Surabaya. The data obtained consists of secondary data on toddler nutrition from 2021, totaling 2,801 records. Based on the data collected, the research variable attributes were determined, including gender, age, weight, height, upper arm circumference, and birth weight, guided by the Indonesian Ministry of Health Regulation Number 2 of 2020 on child anthropometric standards, as well as several previous studies.

B. Explaratory Data Analysis (EDA)

At this stage, analysis and exploration are conducted to understand the condition of the dataset obtained, enabling further analysis to be carried out using various diagrams to visualize the state of the data. The main goal of Exploratory Data Analysis (EDA) is to reveal patterns, trends, and anomalies within the dataset, as well as to ensure the reliability and accuracy of the data.

C. Preprocessing

Data handling for imbalanced data will involve resampling. There are two types of resampling that will be used: oversampling and undersampling. The oversampling method that will be used is Synthetic Minority Over-sampling Technique (SMOTE). The undersampling method that will be used is random undersampling. Data transformation will also be performed using label encoder and discretization techniques. Label encoder is used to convert categorical variables into numeric variables for the Y variable in the dataset. Discretization is performed to convert continuous variables into discrete variables by grouping possible values into well-defined intervals or categories.

D. Data Splitting Scenarios

In this research, three scenarios of training and test data splitting will be created for model evaluation purposes. The training-to-test data split ratio for scenario 1 is 70:30, scenario 2 is 80:20, and scenario 3 is 90:10.

E. Modelling

At this stage, the data has been confirmed to be ready for processing into the modeling phase. The dataset will be used for training different models, namely the C4.5 decision tree and naive Bayes. Both models are created manually by implementing equations 1 and 2 in C4.5 to calculate entropy and information gain, while equation 3 in naive Bayes is used to calculate posterior probability.

F. Model Evaluation

The evaluation of both models will be conducted by analyzing the results from the classification model evaluation metrics explained in section 2 point E, with Area Under the Curve (AUC) as the main reference. Computation time records will also be considered as additional supporting information.

IV. Results and Discussion

The results and discussion phase of the conducted research will be elaborated according to the research methodology stages.

A. Data Collection

The data obtained by the author is secondary data on toddler nutrition under the scope of Jagir Health Center in 2021, with a total of 2081 records. The collected data contains the following information: work unit, posyandu (integrated health service post), district, sub-district, examination date, name, national ID number, toddler ID, address, date of birth, birth order, birth weight, residential address, gender ID, guardian name, guardian national ID, gender, blood type, age in months, weighing method, weight, height, MUAC (Mid-Upper Arm Circumference), height-for-age, height-for-age category, weight-for-age, weight-for-age category, weight-for-height, weight-for-height category, exclusive breastfeeding, vitamin A, early initiation of breastfeeding, taburia (micronutrient powder), and poverty status. The labeling refers to anthropometric z-scores based on the height-for-age index (HAZ). Therefore, the "height-for-age category" feature from the obtained data will serve as the data label (Y variable).

B. Explaratory Data Analysis (EDA)

The analysis step to understand data conditions can be performed by checking the number of samples in each class or label from the obtained dataset to determine the data balance condition. The distribution of sample numbers is displayed in the form of a histogram.

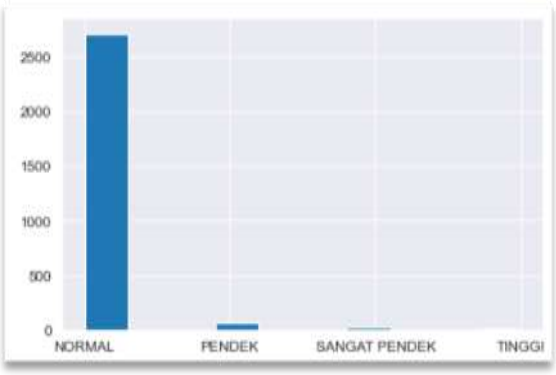

Fig. 2. Histogram Dataset

As shown in Figure 2, the distribution of the obtained data is highly imbalanced, necessitating further handling before the data can be processed.

C. Preprocessing

Classification in this research is divided into two classes: stunting or normal. Therefore, the dataset labeled as "tall" can be removed. Additionally, some data are indicated to have missing data where the MUAC (Mid-Upper Arm Circumference) value equals zero (0), these data should be removed.

- *Label Encoder*

  The Y variable labeling in the dataframe is changed from "normal", "short", "very short" to "0", "1", "2". Gender is changed from "male" to "0" and "female" to "1".

- *SMOTE*

  SMOTE is used to handle the significantly smaller number of samples compared to other classes. By utilizing the imblearn library, SMOTE can be implemented to generate new synthetic samples for the minority class

```
Before SMOTE:  Counter({0: 2631, 1: 64, 2: 21})
After SMOTE : Counter({0: 2631, 1: 2631, 2: 2631})
```

Fig. 3. Smote Result

- *Discretization*

After achieving data balance, the next step is the discretization or binning process by converting continuous numerical variables into discrete variables by grouping data into specific categories or value intervals using the pd.cut function from the pandas library. Target variables "short" and "very short" can be combined into "stunting" or "1". This is supported by the theoretical foundation that toddlers are considered stunting if their z-score is less than -2. As a result, the dataset will be completely transformed into categorical data as shown in Table 2.

TABLE II. BINNING RESULTS

cat_bblahir	cat_usiabulan	cat_bb	cat_tb	cat_lila	jk	kategori
3	7	2	5	1	0	0
3	4	1	3	1	0	1
4	6	2	5	1	0	0
4	5	2	5	1	0	0
3	7	2	6	1	0	0
3	7	2	6	1	0	0
4	7	3	7	1	0	0
3	6	2	7	1	0	0
4	7	2	5	1	1	0
3	7	2	5	1	1	0

- *Undersampling*

A new problem emerges after merging the dataframes labeled "short" and "very short" into one. The data that was previously balanced becomes imbalanced again, as shown in Figure 4.

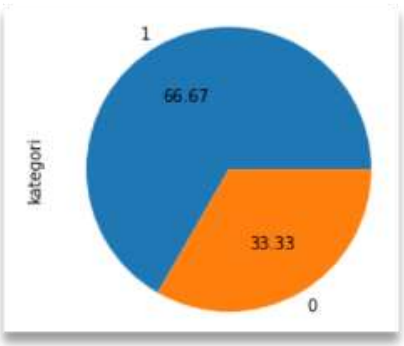

Fig. 4. Dataframe Returns to Imbalanced State

As shown, the "stunting" Dataframe becomes 66.67% and "normal" 33.33%. Considering that there is already a substantial amount of data, random undersampling (randomly removing data) can be performed on the larger class (stunting) to make it equal to the number in the smaller class (normal). The data distribution after random undersampling will be as shown in Figure 5.

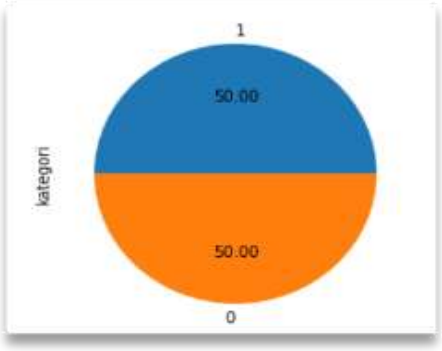

Fig. 5. Random Undersampling

D. Testing Scenarios

Testing scenarios implement all data splitting scenarios in model creation. Based on the equations, the C4.5 algorithm is arranged into two classes: the Node class which functions to store nodes in the decision tree, and the DecisionTreeClassifier class as the decision tree constructor that implements equations 1 and 2. The naive Bayes algorithm is composed of three functions to calculate posterior probability: "calculate_prior" to calculate prior probability, "calculate_likelihood_categorical" to calculate likelihood, and "naive_bayes_categorical" to calculate posterior probability.

With the created models, the best scenario was achieved in scenario 1 with a training-to-test data ratio of 70:30. Table 3 shows a summary of the test scenario metrics results.

TABLE III. TESTING RESULTS SUMMARY FOR DECISION TREE

	Decision Tree C4.5				
Skenario	*Accuracy*	*Precision*	*Recall*	*F1-Score*	AUC
1	86%	82%	92%	87%	0,86

TABLE IV. TESTING RESULTS SUMMARY FOR NAÏVE BAYES

	Naïve Bayes				
Scenario	*Accuracy*	*Precision*	*Recall*	*F1-Score*	AUC
1	77%	74%	85%	79%	0,776

Based on Tables 3 and 4, it can be determined that among the classification models created, the best performance was achieved by the C4.5 decision tree with an AUC value of 86%. Table 3 also shows excellent recall results of 92%. Recall becomes important if the goal of classification is to identify as many stunting toddlers as possible, even if this means some non-stunting toddlers are classified as stunting. This is because stunting is a condition that has negative long-term health impacts, and if stunting toddlers are not properly identified, they may not receive the necessary interventions to address their nutritional problems. Therefore, it is important to maximize the number of correctly identified stunting toddlers, even if some non-stunting toddlers might also be classified as stunting. With C4.5's superiority, it demonstrates that naive Bayes' weakness in dealing with correlated features is indeed true due to the assumption that all features are independent. The complex attributes or features that tend to be correlated in the case of stunting toddler classification make C4.5 excel in this regard. The computation time records of the models are also documented as shown in Table 5.

TABLE V. COMPUTATION TIME RECORDS (SECONDS)

	Decision Tree C4.5	***Naïve Bayes***
Scenario 1	0.184808	5.818802

As shown in Table 5, as the percentage of test data increases, the decision tree produces smaller computation times, while naive Bayes shows the opposite trend. This is influenced by the decision tree concept where the decision tree is built with a divide-and-conquer approach, so smaller data partitions become easier to process. On the other hand, naive Bayes' computation time is heavily influenced by probability calculations themselves. In the model implementation, there is a loop that iterates over each input test data, so the larger the percentage of test data, the greater the computation time required.

V. CONCLUSION

From the processed data obtained from secondary toddler nutrition data in 2021 containing 2,801 records, conclusions can be drawn from the testing results and research of both C4.5 decision tree and naive Bayes methods for classifying stunting nutritional status in toddlers.

First, this research successfully implemented both methods to classify stunting nutritional status in toddlers. Second, the best scenario was observed in scenario 1 with the decision tree method recording the following metric values: accuracy: 86%, precision: 82%, recall: 92%, f1-score: 87%, AUC: 86%.

Additionally, scenario 1 produced faster computation time than other scenarios, taking only 0.185 seconds. From all the presented findings, it can be seen that the C4.5 decision tree is superior. Therefore, the C4.5 decision tree method can be a good option for classifying stunting nutritional status in toddlers.

While classification using C4.5 has already shown good classification results, future research could explore combining it with ensemble learning techniques such as Adaboost, which is expected to produce more optimal implementation results.

REFERENCES

[1] Agustina, N. (2022): *Apa itu Stunting*. https://yankes.kemkes.go.id/view_artikel/1516/apa-itu-stunting.

[2] Al Haromainy, M. M., Sihananto, A. N., Prasetya, D. A., Sari, A. P., Alfyando, M., & Purnomo, R. (2022): *Classification Of Javanese Script Using Convolutional Neural Network With Data Augmentation*. Proceeding - IEEE 8th Information Technology International Seminar, ITIS 2022, 288–291. https://doi.org/10.1109/ITIS57155.2022.10010262

[3] Dewey, K. G., & Begum, K. (2011): *Long-term consequences of stunting in early life*. Maternal and Child Nutrition, 7(SUPPL. 3), 5–18. https://doi.org/10.1111/j.1740-8709.2011.00349.x

[4] Divva Meuthia Zulma, G., & Chamidah, N. (2021): *Perbandingan Metode Klasifikasi Naive Bayes, Decision Tree Dan K-Nearest Neighbor Pada Data Log Firewall*. In Seminar Nasional Mahasiswa Ilmu Komputer dan Aplikasinya (SENAMIKA) Jakarta-Indonesia.

[5] Firdausia Ismi Nurhayati, Nugroho, B., & Yuniar Purbasari, I. (2021): *Implementasi Metode Decision Tree Pada Identifikasi Status Gizi Balita.* Jurnal Informatika Dan Sistem Informasi, 2(2), 204–213. https://doi.org/10.33005/jifosi.v2i2.326

[6] Han, J., Kamber, M., & Pei, J. (2011): Data Mining. Concepts and Techniques, 3rd Edition (The Morgan Kaufmann Series in Data Management Systems).

[7] Herliansyah, V., Latuconsina, R., & Dinimaharawati, A. (2021): Prediksi Stunting Pada Balita Dengan Menggunakan Algoritma Klasifikasi Naïve Bayes Stunting Prediction In Children Using Naïve Bayes Classification Algorithm. e-Proceeding of Engineering 8, 5:6642-6649

[8] Ian H. Witten, Eibe Frank, & Mark A. Hall. (2011): *Data Mining Third Edition*.

[9] Saputra, W. S. J., Puspaningrum, E. Y., Syahputra, W. F., Sari, A. P., Via, Y. V., & Idhom, M. (2022): *Car Classification Based on Image Using Transfer Learning Convolutional Neural Network*. Proceeding - IEEE 8th Information Technology International Seminar, ITIS 2022, 324–327. https://doi.org/10.1109/ITIS57155.2022.10010073

[10] Setyawati, V. A. V. (2012): Peran Status Gizi Terhadap Kecerdasan Kognitif Pada Masa Golden Age Period. Jurnal Visikes 11, 2.

[11] Soliman, A., De Sanctis, V., Alaaraj, N., Ahmed, S., Alyafei, F., Hamed, N., & Soliman, N. (2021): *Early and long-term consequences of nutritional stunting: From childhood to adulthood*. Acta Biomedica, 92(1). https://doi.org/10.23750/abm.v92i1.11346