Jurnal Sistem Informasi Bisnis 02(2025)
Copyright ©2025, JSINBIS, p-ISSN: 2502-2377, e-ISSN: 2088-3587
On-line: http://ejournal.undip.ac.id/index.php/jsinbis/article/view/71808

Increasing the Accuracy of Random Forest Algorithm Using Bagging Techniques in Cases of Stunting Toddlers

Amir Ali[a,b*], Purwanto[c], Mundakir[d]

[a]Doctoral Program of Information System, School of Post Graduate Studies, Diponegoro University,
Jl. Imam Bardjo S.H., No. 5, Pleburan, Semarang, Indonesia 50241
[b]Medical Record and Health Information, Dr. Soetomo Hospital Foundation College of Health Sciences,
Surabaya Indonesia, 60286
[c]Department of Information Systems, School of Post Graduate Studies, Diponegoro University,
Jl. Imam Bardjo S.H., No. 5, Pleburan, Semarang, Indonesia 50241
[d]Department of Health Faculty, Muhammadiyah Surabaya University, Surabaya, Indonesia 60113

Submitted: February 12th, 2025; Revised: March 16th, 2025; Accepted: April 5th, 2025; Available Online: May 31st, 2025
DOI: 10.14710/vol15iss2pp167-172

Abstract

Increasing the accuracy value can be increased by using other algorithms. Increasing the accuracy value of a classification algorithm, the level of success of the algorithm's prediction is more precise and appropriate in providing its label. The purpose of the research is looking for performance of accurate value for prediction with bagging algorithm. This research uses Random Forest and Bagging algorithm used for optimization. 12 data whose position is far from other data. 12 data deviates from the data pattern and are outliers. With z-score process, it will be processed to eliminate outlier data. After removing the outlier data, the data clean is 137 toddler data. After removing outliers and standardizing the data, the accuracy obtained was 71% up to 100th accuracy with random forest algorithm. Optimization of a bagging algorithm to predict stunting in a dataset of toddlers that has been acquired and assessed its performance. This can be seen from the optimization of prediction results up to the 100th iteration, where the prediction accuracy results were 80.67%. Using the Random Forest algorithm and bagging techniques, the prediction of stunting in toddlers works well. Optimization of prediction results up to the 100th iteration, where the prediction accuracy results were 80.67%.

***Keywords**: Random Forest; Bagging algorithm; Outlier removal; Stunting prediction; Accuracy Improvement.*

1. Introduction

Naturally, the database records the collection of toddler health examination data that is digitally recorded in the health data application, particularly the data on toddler examinations at each posyandu that is recorded in the electronic information system application for community-based nutrition recording and reporting (E-PPGBM). Naturally, a database has a plethora of data that can be utilised to obtain valuable insights. If proper data processing procedures are applied throughout the processing of relevant information, the results will be beneficial
One technique for handling data in a database is data mining. You can use this technique to look through data that is kept in a database. Classification is one of the data processing techniques used in data mining. The process of classification entails identifying models or functions that can be used to distinguish between distinct concepts or explain different types of data.

Predicting the classes of items whose labels are unknown is the aim of the investigation. Several data mining algorithms, like random forest, can be used for classification problems. An algorithm for classifying data might be deemed useful or not based on its accuracy value. Bagging is an algorithm that can be used to increase the accuracy of the classification algorithm. When a classification algorithm's accuracy value rises, the algorithm's prediction becomes more accurate and suitable for assigning a label.

In this study, the author uses a dataset of community-based nutrition recording and reporting data (E-PPGBM) to improve the accuracy value of the random forest classification algorithm using bagging optimization. The author will employ bagging optimization in this investigation since it has been shown in previous studies to improve the accuracy of a number of the methods. One demonstrates some methods to improve accuracy in the assessment of facial emotions. K-Nearest Neighbors (KNN) has shown accuracy of 76.3% from a facial emotion database. Support Vector Machine (SVM) has obtained accuracy of 90%, while the suggested AdaBoost based Random Forest classifier for Emotion Classification (ARFEC) model was able to demonstrate accuracy of 92.5% (Gubbala et al., 2023). In the other side, prediction accuracy for coronary heart disease was improved. The accuracy of the suggested Random Forest with Bagging model has

*) Corresponding author: amirali@students.undip.ac.id

obtained 84,07%, whereas the Random Forest exhibited 77,40% (Saifudin et al., 2020).

Using a collection of datasets from nutritional recording and reporting, researchers attempted to examine the performance of bagging optimization with random forest algorithms before and following optimization with bagging approaches based on the literature review above.

2. Literature Review/Related Works

2.1. *Data*

Dataset used in this study was taken from *antropometri toodler*. The attribute of the Dataset is listed in Table 1.

Table 1. Dataset attribute from antropometri toodler

Attribute Name	Type	Explanation of Information
NIK	Numeric	Population Register Number
JK	Category	Category gender
BB_Birth	Numeric	Birth Weight
TB_Birth	Numeric	Birth Height
Age Toodler	Numeric	Age at measurement
BB	Numeric	Weight
TB	Numeric	Height
LiLA	Numeric	Upper arm circumference
TB/U	Category	height for age

2.2. *Analysis Technique*

In this study, bagging optimization techniques will be used to carry out random forest optimization classification algorithms. The accuracy values of the results before and after optimization with bagging on random forest algorithms will be compared in this study. The degree to which the optimization technique deepens the algorithm's accuracy value will be examined, and you can see which algorithm has the best accuracy value of the comparison results obtained.

This study employs percentage split (70%), one of the modes used. It entails assessing the accuracy results of an algorithm that uses 30% of the dataset for testing and 70% of the dataset for training. From the research results, Table 2 shows accuracy results were obtained.

Table 2. Accuracy comparison before & after optimization

Algorithm	Non Optimization	Bagging
Random Forest	71%	80,67%

2.3. *Random Forest*

Leo Breiman was the one who initially introduced Random Forest (RF). Random Forest (RF) is a technique that can improve accuracy by randomly generating attributes for every node. Decision trees are gathered together in RF, and this collection of decision trees is utilized for data classification. A few of random forest's benefits are its ability to withstand outliers, boost accuracy in the event of missing data, and store data effectively. In order to enhance the performance of the classification model, Random Forest also has a feature selection mechanism that can select the best features.

The Gini index is used to select features at each internal node of the decision tree. The Gini Index value can be calculated using Equation 1:

$$Gini\ (S_i) = 1 - \sum_{i=0}^{c-1} \quad p_i^2 \tag{1}$$

where p_i is the relative frequency of class C_i in the set C_i is n classes for i = 1, c – 1, and c is the number of classes that have been determined. The split quality of feature k into subset S_i is the number of samples belonging to class C_i, then calculated as the number of considerations indicating the Gini of the resulting subset. Data can be calculated using Equation 2. Where n_i is the number of samples in the subset S_i after splitting and n is the number of samples at the given node.

$$Gini_{Split} = \sum_{i=0}^{k-1} \quad \left(\frac{n_i}{n}\right) \ Gini\ (S_i) \tag{2}$$

2.4. *Bagging*

Bagging is a machine learning ensemble technique that has been shown to improve classification performance in difficult real-world situations. It is an easy-to-use yet effective method (L. Liu et al., 2019).

To create an aggregated predictor with a higher accuracy, the bagging ensemble predictive algorithm technique is utilized to integrate the predictive performance of many copies of a base predictor. The bootstrap approach, which is one of the most often used data resampling techniques in statistical research, creates several versions of the base predictor (Lin et al., 2021). Predictive model bagging frameworks tend to reduce variance and avoid overfitting (Lin et al., 2022). Bagging aids in mitigating the effects of variation and noisy data (C. L. Liu et al., 2024).

2.5. *Confusion Matrix*

One way to measure the performance of a model that produces binary classification, such as the concept of true positive (TP), true negative (TN), false positive (FP), and false negative (FN) which are mapped in Confusion Matrix as shown in Table 3.

Table 3. Confusion Matrix Concept

Prediction	Positive	Negative
Positive	TP	FP
Negative	FN	TN

This table is called a confusion matrix because the terms used can create confusion for people who read it. TP is the model successfully predicts positive (yes), because in reality it is positive (yes). TN is the model successfully predicted negative (no), because in reality it was negative (no), FP is the model predicts positive (yes) but is wrong because the reality is negative (no), FN is the model predicts negative (yes) but is wrong because the reality is positive (yes).

3. Method

The antropometri toddler dataset is comprised of 149 sample. consisting of 73 male toddlers and 76 female toddlers. The dataset has a total of 9 attributes (representing toddler anthropometric data) as listed in Table 4.

Table 4. Toddler antropometric dataset

NIK	Sex	Birth Weight	Birth Height	Age At measurement	Weight	Height	Upper Arm Circumference	Height For Age
351511460722xxxx	P	3	49	1	6.6	71	0	Very Short
351509300522xxxx	L	2.3	50	1	8.8	77.5		Short
351509270522xxxx	L	3	50	1	8.5	77.5		Short
351509040323xxxx	L	3	50	0	7.2	66		Very Short
…………………	…	……	…..	….	……	…….	……	………….
351502170123xxxx	L	3.3	53	0	9.4	70.5		Short
351502431222xxxx	P	3.1	50	1	7.7	68.9		Short
351502570423xxxx	P	3.3	50	0	6.2	62.5		Very Short
351502150620xxxx	L	3.1	49	3	13.5	91		Short
351502230919xxxx	L	5.1	50	4	16	95		Short

The attributes in the dataset including NIK, Sex, Birth Weight, Birth Height, Age at measurement, Weight, Height, Upper arm circumference, and Height for age.

The following steps make up this section: the description of the data, the preprocessing method, and the computation of the classification algorithm's accuracy value using the random forest algorithm. This accuracy value will be compared with the algorithm for increasing the accuracy value, namely the bagging algorithm as one of the accuracy increasing algorithms which is part of the ensemble algorithm. This research uses anthropometric data on toddlers approach. This research uses random forest algorithm. Increasing accuracy values using the bagging algorithm. An ensemble algorithm is used to design and implement the suggested model. The bagging method is one of the ensemble algorithms. Figure 1 displays the proposed model flowchart.

Data preparation is required before the accuracy algorithm calculation method is applied to the toddlers' data. In order to do data preprocessing, transaction data is retrieved by random sampling and partial data cleaning. When data has a value of NaN, data cleansing is done. Up to 149 complete sets of data are gathered, of which 73 are related to toddlers who are male and 76 are related to toddlers who are female.

In addition, there is a method known as imputation for replacing lost data. Correlation-based imputation is one method of imputation technique (Curioso et al., 2023).

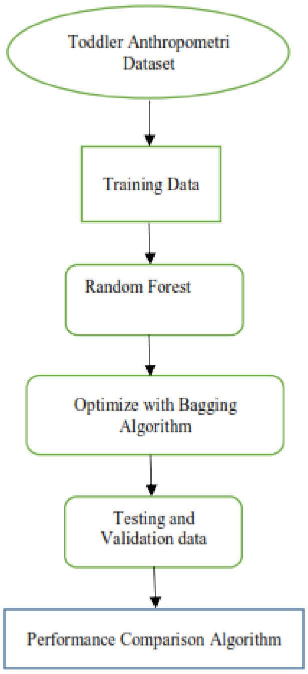

Toddler Anthropometri Dataset
Training Data
Random Forest
Optimize with Bagging Algorithm
Testing and Validation data
Performance Comparison Algorithm

Figure 1. Proposed algorithm bagging random forest method

A one-time approach to imputation called Nullify the Missing Values before Imputation (NMVI) divides the data into complete and incomplete subsets first. For each class in which there is missing data, an upper limit is then set to help the model estimate missing values more closely to the actual values (Bhagat & Singh, 2022).

4. Results and Discussion

This research optimizes the accuracy value of the random forest algorithm by using a bagging algorithm in an area in order to obtain an overview of which areas experience cases of babies suffering from stunting in the Sidoarjo Regency.

4.1. Preprocessing Data

Preprocessing data starts from removing features in the dataset that are not used. For example, the date of birth feature. A part from that, changing the data type format from object data type to numeric data type needs to be done, such as changing the data type of birth weight and birth height features.

There are still features with NaN data in its column, namely the Upper Arm Circumference feature. This data needs to be filled with values. by using imputation techniques, this can be done. for example, by giving a null (0) value to the data contents for that column.

Table 5. Processed dataset

NIK	Sex	Birth Weight	Birth Height	Age At measu rement	Weight	Height	Upper Arm Circumference	Height For Age
351511460722xxxx	0	3	49	1	6.6	71	0.0	Very Short
351509300522xxxx	1	2.3	50	1	8.8	77.5	0.0	Short
351509270522xxxx	1	3	50	1	8.5	77.5	0.0	Short
351509040323xxxx	1	3	50	0	7.2	66	0.0	Very Short
351509110921xxxx	1	3	50	2	11.0	81.1	0.0	Short

The results are as shown in Table 5, an example of a small dataset whose data has been processed. The existence of this outlier data will cause deviations from the results of data analysis. An outlier is an observation whose point of observation deviates far from the data pattern. Therefore, the outlier data needs to be removed.

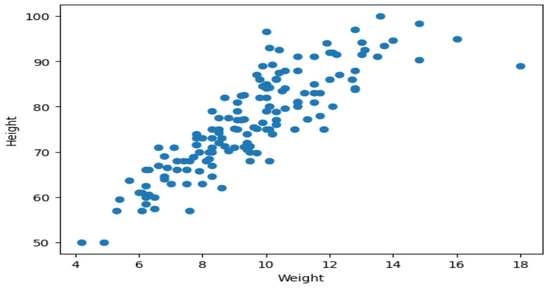


Figure 2. Outlier data visualization

From Figure 2 namely 12 data whose position is far from other data. These 12 data deviates from the data pattern and are outliers. With a certain process, it will be processed to eliminate outlier data. the process of eliminating outlier data using the z-score formula.

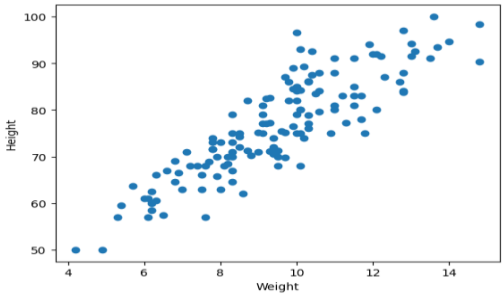
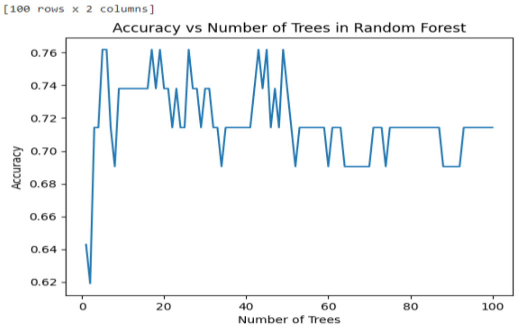


Figure 3. Data visualization after removing the outlier data

From Figure 3, information is obtained, there is no data that deviates far from the data pattern. So that after removing the outlier data, the data is 137 toddler data.

4.2. Calculate the Accuracy Random Forest Algorithm

The data can be normalized using data scaling techniques. To facilitate statistical analysis, data normalization is the process of ensuring that several values, none of which are excessively large variables have the same range of v or tiny.



```
[100 rows x 2 columns]
```

Figure 4. The result of accuracy vs number of trees in random forest

Calculating the accuracy of the random forest algorithm from the existing dataset. After removing outliers and standardizing the data, the accuracy value obtained was 71% up to 100th accuracy.

4.3. Optimization of Accuracy random forest algorithm with bagging algorithm

The Random Forest Bagging procedure is applied using the model as shown in Figure 1. For this dataset, optimization used the ensemble bagging approach. The performance of the suggested model is evaluated using a dataset that was acquired and used to predict stunting in toddlers. Table 2 displays the outcomes of evaluating the suggested model's performance.

Table 6. Performance accuracy of the model

Model	Accuracy Result	Iteration
Random Forest	71,00%	100
Random Forest + Bagging	80,67 %	100

Table 6 demonstrates that the Random Forest algorithm with bagging techniques outperforms the Random Forest algorithm without optimization when it comes to predicting stunting in toddlers.

4.3.1. Remove Ouliers

Because outlier identification has so many important applications, it is a crucial technique in data mining. It can be applied to remove noise or analyze individual observations in data that are not consistent with the surrounding data points (Souiden et al., 2022). To eliminate noise or examine individual observations in data that differ from those in the surrounding area, outlier identification can be employed (Aggarwal et al., 2019). To eliminate outliers, numerous methods are employed. In the field of outlier detection, the deep learning model proves to be a successful method (Abhaya & Patra, 2023). The degree of outlierness of each item can be calculated using the two outlier detection techniques mentioned above (Jiang et al., 2016), outlier detection simultaneously (Gan & Ng, 2017), adaptive removal of outliers from the standard Tukey rule (Shrifan et al., 2022), Z-Score (Jamshidi et al., 2022), and semi-supervised outlier detection (Zhao et al., 2022).

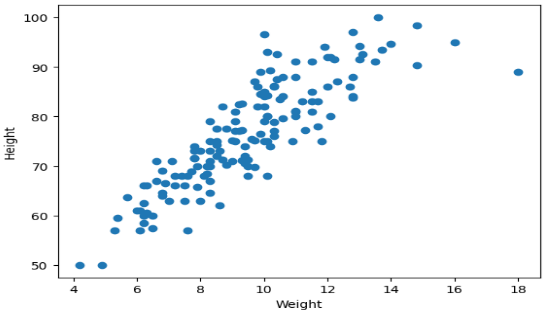


Figure 5. Outlier data toddler antropometric.

We can infer from the preceding image that objects classified as outliers are those that are found to deviate from other objects in the dataset. Information is gleaned from the image above, namely 12 data points whose locations are apart from one another. These 12 data are outliers because they depart from the overall pattern. It will be processed in a particular way to get rid of the outlier data. removing anomalous data by applying the z-score formula. Data outliers have long been identified using the z-score test (Aggarwal et al., 2019).

4.3.2. Optimization of accuracy with bagging algorithm

The purpose of this observation is to improve the accuracy and precision of toddlers' predictions on stunting, with the ultimate goal of reaching superior accuracy. The study's suggested framework is shown in Figure 1. To train on data samples, the Random Forest algorithm is utilized. The ensemble bagging technique is utilized to improve classification accuracy and reduce misclassification in predictions of stunting toddlers (Huda et al., 2018).

Optimizing the accuracy of predicting cases of stunting under five can be improved by using a bagging algorithm. This can be seen from the optimization of prediction results up to the 100th iteration, where the prediction accuracy results were 80.67%. The results of this prediction accuracy are calculated using the confusion matrix table.

Table 7. Confusion matrix random forest with bagging

Actual	Prediction (Yes)	Prediction (No)
Actual (Yes)	125	35
Actual (No)	23	117

Table 7 shows that: True Positive (TP) is 125, True Negative (TN) is 117, False Positive (FP) is 23, False Negative (FN) is 35. The accuracy value of (TP+TN)/(TP+TN+FP+FN) is (125+117)/(125+117+23+35) equal to 0,80666 which means that the accuracy value obtained is 80,67%.

5. Conclusions

The results of this research carried out bagging optimization on the random forest algorithm on the anthropometric dataset of stunted toddlers, which can be concluded that The number of iterations given to the parameter can influence the value of accuracy of optimization using bagging applied to the random forest algorithm. In the random forest algorithm, bagging optimization can increase the accuracy value. The accuracy has been improved from 71% to 80.67%.

Acknowledgement

The authors sincerely thank Diponegoro University for providing the necessary resources and facilities and Sidoarjo district health office for their invaluable assistance in data collection and analysis.

References

Abhaya, A., & Patra, B. K. (2023). An efficient method for autoencoder based outlier detection. Expert Systems with Applications, 213, 118904. https://doi.org/10.1016/j.eswa.2022.118904

Aggarwal, V., Gupta, V., Singh, P., Sharma, K., & Sharma, N. (2019). Detection of spatial outlier by using improved Z-score test. In Proceedings of the International Conference on Trends in Electronics and Informatics (ICOEI 2019) (pp. 788–790). https://doi.org/10.1109/icoei.2019.8862582

Bhagat, H. V., & Singh, M. (2022). NMVI: A data-splitting based imputation technique for distinct types of missing data. Chemometrics and Intelligent Laboratory Systems, 223. https://doi.org/10.1016/j.chemolab.2022.104518

Curioso, I., et al. (2023). Addressing the curse of missing data in clinical contexts: A novel approach to correlation-based imputation. Journal of King Saud University - Computer and Information Sciences, 35(6). https://doi.org/10.1016/j.jksuci.2023.101562

Gan, G., & Ng, M. K. P. (2017). K-means clustering with outlier removal. Pattern Recognition Letters, 90, 8–14. https://doi.org/10.1016/j.patrec.2017.03.008

Gubbala, K., Kumar, M. N., & Sowjanya, A. M. (2023). AdaBoost based Random forest model for emotion classification of facial images. MethodsX, 11, 102422. https://doi.org/10.1016/j.mex.2023.102422

Huda, S., et al. (2018). An ensemble oversampling model for class imbalance problem in software defect prediction. IEEE Access, 6, 24184–24195. https://doi.org/10.1109/ACCESS.2018.2817572

Jamshidi, E. J., Yusup, Y., Kayode, J. S., & Kamaruddin, M. A. (2022). Detecting outliers in a univariate time series dataset using unsupervised combined statistical methods: A case study on surface water temperature. Ecological Informatics, 69. https://doi.org/10.1016/j.ecoinf.2022.101672

Jiang, F., Liu, G., Du, J., & Sui, Y. (2016). Initialization of K-modes clustering using outlier detection techniques. Information Sciences, 332, 167–183. https://doi.org/10.1016/j.ins.2015.11.005

Lin, E., Lin, C. H., & Lane, H. Y. (2021). Applying a bagging ensemble machine learning approach to predict functional outcome of schizophrenia with clinical symptoms and cognitive functions. Scientific Reports, 11, 1–9. https://doi.org/10.1038/s41598-021-86382-0

Lin, E., Lin, C. H., & Lane, H. Y. (2022). A bagging ensemble machine learning framework to predict overall cognitive function of schizophrenia patients with cognitive domains and tests. Asian Journal of Psychiatry, 69, 103008. https://doi.org/10.1016/j.ajp.2022.103008

Liu, C. L., et al. (2024). A bagging approach for improved predictive accuracy of intradialytic hypotension during hemodialysis treatment. Computers in Biology and Medicine, 172, 108244. https://doi.org/10.1016/j.compbiomed.2024.108244

Liu, L., Chin, S. P., & Tran, T. D. (2019). Reducing sampling ratios and increasing number of estimates improve bagging in sparse regression. In 2019 53rd Annual Conference on Information Sciences and Systems (CISS). https://doi.org/10.1109/CISS.2019.8692865

Saifudin, A., Nabillah, U. U., Yulianti, & Desyani, T. (2020). Bagging technique to reduce misclassification in coronary heart disease prediction based on random forest. Journal of Physics: Conference Series, 1477(3), 032009. https://doi.org/10.1088/1742-6596/1477/3/032009

Shrifan, N. H. M. M., Akbar, M. F., & Isa, N. A. M. (2022). An adaptive outlier removal aided k-means clustering algorithm. Journal of King Saud University - Computer and Information Sciences, 34(8), 6365–6376. https://doi.org/10.1016/j.jksuci.2021.07.003

Souiden, I., Omri, M. N., & Brahmi, Z. (2022). A survey of outlier detection in high dimensional data streams. Computer Science Review, 44, 100463. https://doi.org/10.1016/j.cosrev.2022.100463

Zhao, Y., Li, H., Yu, X., Ma, N., Yang, T., & Zhou, J. (2022). An independent central point OPTICS clustering algorithm for semi-supervised outlier detection of continuous glucose measurements. Biomedical Signal Processing and Control, 71, 103196. https://doi.org/10.1016/j.bspc.2021.103196