International Journal of Multidisciplinary Research and Growth Evaluation.

The Random Forest algorithm for classifying stunting in toddlers based on anthropometric data

Aidil Zikri [1], Alwis Nazir [2*], Suwanto Sanjaya [3], Elin Haerani [4], Iis Afrianty [5]
[1-5] Faculty of Science and Technology, Sultan Syarif Kasim Riau State Islamic University, Pekanbaru, Indonesia

* Corresponding Author: **Alwis Nazir**

Article Info

ISSN (online): 2582-7138
Volume: 05
Issue: 03
May-June 2024
Received: 18-04-2024
Accepted: 20-05-2024
Page No: 931-937

Abstract
Stunting is a growth and developmental disorder in children that often occurs during the first 1,000 days of life, from conception to the age of two. Factors such as inadequate nutritional intake, recurrent infections, and a less clean environment can contribute to stunting. In this case, use the random forest algorithm. The goal is to categorize the stunting case. The variables used are gender, age, birth weight, birth height, weight, height, and Exclusive Breastfeeding. With the number of datasets used reaching 6,500, the experiment was performed with a combination of parameters, namely n_estimators 100, 200, and 500. Max_features = 5, 7, 10, and 13. As well as min_samples_split and min_sample_leaf = 20, 50, and 100. Based on the specified set of hyperparameters, 108 test results were obtained. The highest accuracy is 0.9651, with precision = 0.9603, recall = 0.9718, and F1-score = 0.9660. With n_estimators formed = 100, max_depth = 13, min_samples_split = 20, and min_samples_leaf = 20.

DOI: https://doi.org/10.54660/.IJMRGE.2024.5.3.931-937

Keywords: Data mining, random forest, stunting

Introduction

Stunting is a growth and developmental disorder experienced by a child and occurs early in life, especially in the first 1000 days from conception to the age of two. (World Health Organization, 2015) [20]. This condition is characterized by a child's height being lower than the average age. Factors such as poor nutritional intake, recurrent infections, and a less clean environment can contribute to stunting. In the long run, stunting has a serious impact on the health and ability of children to reach their potential. (Azahra *et al*., 2022) [2]. HAZ and WAZ are assessments for stunting measurements based on anthropometric data.

Data anthropometry is a science that studies the human body's morphology and various dimensions. It is a series of quantitative measurements of muscles, bones, and adipose tissue used to assess body composition. The process of determining a person's nutritional status generally involves the collection of important data that can be objective and subjective, then compared to existing criteria.(Permana Ratumanan *et al*., 2023) [15]. It is used to measure stunts and reduce the rate of prevalence that occurs. According to a report by the World Health Organization (World Health Organization, 2023) [21] in 2022, it was found that 148.1 million (22.3%) children under the age of five were characterized by too short heights compared to their age. (Stunting). Asia ranks first with 52 percent of the global prevalence, followed by Africa with 43 percent. According to the Indonesian Nutrition Status Study (SSGI) report, the prevalence of stunting will be 24.4% by 2021 in Indonesia itself. There was a decrease in 2022, to 21.6%. The government targets a reduction of 17.8% by 2023 and a reduction of 14% by 2024. (Kemenkes RI, 2023) [7].

Based on the above data, the attempt to decrease the stunting rate is done by determining the factors that influence stunting. To identify such factors, one can use the data mining method. Data mining is the process of digging and managing large databases to obtain new information or knowledge (knowledge discovery) (Gede Iwan Sudipa *et al*., 2023) [4]. Machine learning is one method used in data mining.

One important aspect of data mining is machine learning, which is the study of computer algorithms capable of identifying patterns in data without significant human intervention. (Situmorang, 2023) [18]. In the book "Introduction to Concepts and Machine Learning," written by (Wira Gotama Putra, 2020) [19], machine learning is described as a technique to create models

that reflect patterns in data. In machine learning, there are techniques called classification. Classification is a type of data analysis that can help people determine the label class of the sample they want to classify.

Classification is supervised learning, a method that tries to find a relationship between input and target attributes. (Hendrian, 2018) [5]. There are many algorithms that can be used for classification methods; one of them is Random Forest.

A random forest algorithm is defined as a group of regression tree classifications trained on training data using random feature choices in the resulting tree process. Once a number of trees have been produced, each tree chooses to get the most popular class. For this technique of classification, two parameters are required: the number of trees and the number of attributes used. In a random forest algorithm, the decision tree does not leave because it is a majority vote group. (Perdana *et al*., 2021) [14]. In a study by (Miftahusalam *et al*., 2022) [10], a comparison of Random Forest, Support Vector Machine, and Naïve Bayes' algorithms in Twitter's sentimental analysis of public opinion about the removal of honorable officials showed the following results: Random forest had an accuracy of 66.67%, support vector machine (65.33%), and naive bayes (64%).

Several related studies on random forest classification methods and stunting case studies. According

Materials and Methods

Stunting is a condition in which a child's growth is hindered so that his or her height is lower than the standard of his or her peers (Rosarita Niken Widiastuti, 2019) [16]. According to research (Wulandari Leksono *et al*., 2021) [22], stunting is a condition of impaired growth and development in children due to malnutrition, infection frequency, and a lack of adequate psychosocial stimulation. According to research (Nugroho *et al*., 2021) [12], the factors that influence the incidence of stunting in early childhood are energy intake, birth weight, maternal education level, family income level, parenting practices, and dietary diversity. Meanwhile, according to research (Firrahmawati *et al*., 2023) [3], the factors that influence stunting are parental income and maternal education.

The process of classifying stunts can be done using data mining methods. Data mining, according to the book "Data Mining Algorithm C4.5," written by (Muslim *et al*., 2019) [11], to research (Juwariyem & Sriyanto, 2023) [6], based on the results of random forest testing to see the accuracy of the prediction of the success of the data tested, they obtained an accurate result of 85.86%. Testing was carried out through a confusion matrix evaluation. This model can be used to predict young people at risk of stunting. According to research conducted by (Adzhima, 2023) [1] using support vector machines to classify news stunting status, The test was performed with y = 0.01 parameters, using the attributes age, gender, premature lactation index (IMD) weight, and height, reaching an average accuracy of 98.99%.

According to the study (Setiawan, 2023) [17], a web- based stunting classification system using the Naïve Bayes method using the same data set obtained accuracy from stunting calculations of 91%, accuracy of 89%, and recall of 95%. According to a study (Pahlevi *et al*., 2024) [13], optimization of the Naïve Bayes-based particle swarm optimization algorithm for the classification of stunting status. Using the same data and variables, the test obtained an accuracy of 80.69%, and the model was evaluated with a confusion matrix and a ROC curve or under curve (AUC).

Therefore, based on previous studies on the classification of stunts, this study attempts to classify stunts using data that includes gender type, age, birth height, birth weight, weight, body heights, exclusive breastfeeding, and the source status of Kaggle. This study will use the Random Forest algorithm by finding its accuracy with Python is a process in which significant patterns and knowledge are found in large amounts of data. Data sources can be databases, warehouses, the web, repositories, or data that flows into the system. According to the book "Data Mining and the Application of Methods," written by (Liantoni, 2022) [9], data mining is the steps taken to extract useful information from a vast database. It needs extraction to generate new insights that can help or contribute. One algorithm that can be used to classify stunts is random forest. Random forest is a method that employs a number of decision trees. The maximum number of voices that appear from the entire decision tree will be used to determine the class of a data input. In general, the use of many decision trees can provide optimum global accuracy values (Kusumarini *et al*., 2021) [8]. Here are the stages of the research process, which can be seen in Fig *1*.

Fig 1: Research Process flow

Data Collections

In this study, the researchers identified an emerging problem, namely the classification of stunting status on news based on anthropometric data (Wulandari Leksono *et al*., 2021) [22]. The necessary data collection is carried out to be able to determine the status of stunting on the news by measuring the values of the variables that affect stunting.

Data in this study is obtained from a secondary source taken from Kaggle. Data is obtained from the Kaggle account https://www.kaggle.com/datasets/muhtarom/stunting/data, accessed at 00:12 GMT on March 13, 2024. The data set used is from the year 2023, with the total amount of data reaching 6500 records, with the number of variables being 7 variables (X) with 1 class (Y).

Transformation Data

The data obtained can still not be directly used in the classification model, so it requires a data processing effort, including data transformation. Once the data processing phase is completed, you get a ready-to-use dataset.

Build a Random Forest

After the data processing phase, the next step is to implement the data into a random forest algorithm to perform a

classification of the stunting status of the news. The stages of the analysis of the classification model in this study can be seen on the flowchart below in Fig *2*.

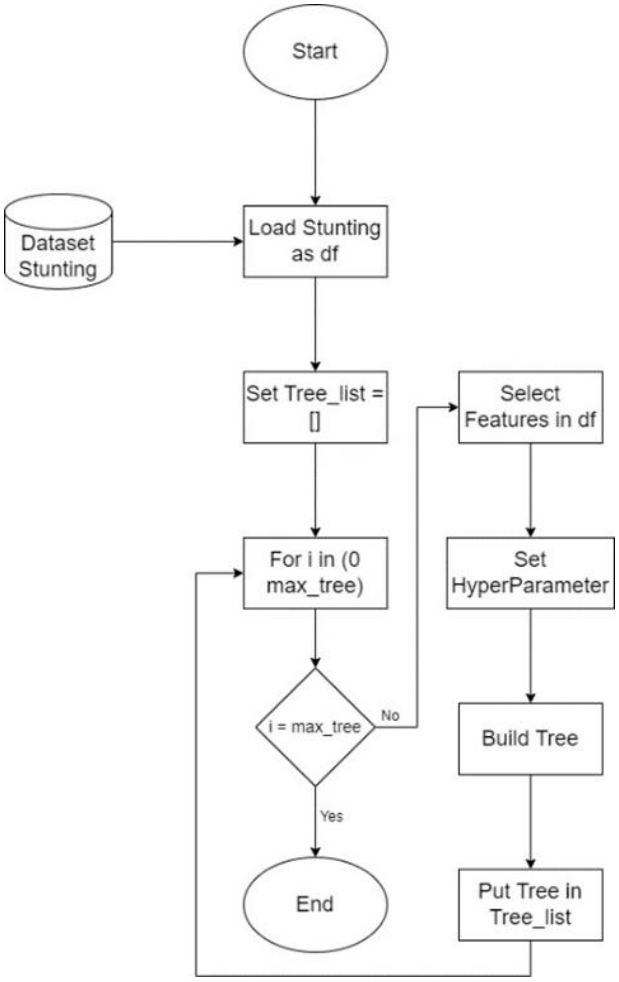

Fig 2: Random Forest Classification Flowchart

According to the flowchart above, the common random forest stages are:

1. According to the ratio used, divide the data into training data and test data.
2. Forms a prediction tree with as many trees to be built.
3. Each extinction tree has a random predictor based on the hyperparameter used.
4. Next, a random forest will make predictions by combining the results of each decision tree with a majority vote for classification.

Model Testing

The next step is to evaluate the model that has been built using some tests such as accuracy, precision, recall, and f1-score. Accuracy is a common and simple parameter for evaluating the performance of a classification algorithm, i.e., by showing how much of the percentage of the truth of the prediction. Precision is the overall prediction's degree of accuracy. Recall is the level of prediction of the total data that actually occurs. Whereas f1-score is how well the model can identify the true case while ensuring that there are not too many errors by comparing accuracies and recalls weighted.

Results

At this stage, analysis and implementation are carried out to understand the need to conduct a study aimed at finding out the problem in more detail. This is done by analyzing the stunting datasets obtained and analyzing the random forest classification algorithm. After that, implementation is carried out, i.e., applying the results of the analysis to the objective of accuracy values to classify the status of stunting.

At the initial stage is data analysis, data obtained from secondary sources taken from Kaggle. The data set used is from the year 2023, with the total amount of data reaching 6500 records, with the number of variables being 7 variables (X) with 1 class (Y). In this study, the scale or range of data used is for the age variable, measured from 0 to 60 months. Yes or no for the Exclusive Breastfeeding variable scale values, and yes or no for the Stunting class. The table can be seen in Table 1.

Table 1: Original Data

No	Sex	Age	Birth Weight	Birth Length	Body Weight	Body Length	Exclusive Breastfeeding	Stunting
0	F	56	2.9	50	11.0	90.0	Yes	No
1	F	20	3.3	49	11.1	80.5	No	No
2	M	4	2.8	48	6.5	63.0	No	No
...	...	...	...	...	...	...	...	...
6498	M	11	2.9	49	7.7	66.0	No	Yes
6498	F	14	2.9	49	6.5	66.0	No	Yes

After the data is analyzed, the next step is the data processing phase. This research performs the transformation of data that is of the of the object or string type. On the data obtained, there are three variables that are still using the data type string but have been changed to the data type numeric. This is done to run the classification process that uses the type of data (numerical or float). The variable that is transformed is gender; the value "F" will be represented by value 1 and "M" by value 0. Once the data processing phase is completed, you get a ready-to-use dataset. The measurement is taken from eight columns of seven variables (x) and one class (y). The variables used are gender, age, birth weight, birth height, weight, height, Exclusive Breastfeeding, and category of the stunting class. Data used as much as 6500 records. Table 2 can be seen here.

Table 2: Transformation

No	Sex	Age	Birth Weight	Birth Length	Body Weight	Body Length	Exclusive Breastfeeding	Stunting
0	1	56	2.9	50	11.0	90.0	1	0
1	1	20	3.3	49	11.1	80.5	0	0
2	0	4	2.8	48	6.5	63.0	0	0
3	1	14	2.0	49	7.0	71.0	1	0
...	...	...	...	...	...	...	...	...
6498	0	11	2.9	49	7.7	66.0	0	1
6498	1	14	2.9	49	6.5	66.0	0	1

The analysis and pre-processing phases of the data have been completed, followed by the random forest formation phase. The ratios used are (70; 30), (80; 20), and (90; 10). Based on this ratio, any amount of data used can be seen in Table 3.

Table 3: Number of Data

Ratio	Number of Training Data	Number of Testing Data
70;30	4550	1950
80;20	5200	1300
90;10	5850	650

Based on the ratio that has been determined, the next step is to perform a model test. This research uses random forest classification algorithms. A total of 108 experiments were obtained based on the hyperparameters used. At the initial stage, the model is called, and the hyperparameter is determined. That is, n_estimators is the number of trees that will be built to form a forest. The criterion is used to divide the node when building a tree, and max_features is the maximum number of variables used to separate the split from the decision tree. Max_depth is the maximum amount of depth of the tree to be built. Min_samples_split is the minimum amount of data to be used in a node, and min_amples_leaf is the minimal number of data in a sheet. Based on its hyperparameters, the criterions that will be used are 'entropy', max_features'sqrt', and min_samples_leaf, which will be equated with min_sample_split. Here's a combination of experiments that can be seen in Table 4.

Table 4: Hyperparameter

N_Estimators	Max_Depth	Min_Samples_Leaf
100	5	20
200	7	50
500	10	100
	13	

Once the hyperparameter is determined, the next step is to perform a manual entropy calculation. Entropy is used to measure impurity in data. This process entails assessing how variable class labels are present in the datasets. Approaching 1 (one) indicates that the data has a high level of impurity, while the lower the data, the more regular it is. Here's the calculation.

$$\text{Entropy} = \left(\frac{3.312}{6.500} * \log 2 \left(\frac{3.312}{6.500}\right)\right) - \left(\frac{3.188}{6.500} * \log 2 \left(\frac{3.188}{5.500}\right)\right)$$

$$= (0.509 * \log 2 (0.509) + (0.491 * \log 2 (0.491)))$$

$$= (-0.498 + -0.497)$$

$$= 0.995$$

Then based on the manual calculations above, it can be assured that the data has a high level of impurity. The next step is to test a model against a data set that has some combination of parameters already defined. We got 108 test results. From each ratio and hyperparameter used, the highest accuracy can be seen in Table 5.

Table 5: Accuracy Results

Ratio	The Highest Accuracy Obtained
70;30	0,9651
80;20	0,9623
90;10	0,9646

Based on the above table, it can be seen that the ratio (70;30) has the highest accuracy of 0.9652 with the hyperparameters n_estimators = 100, max_depth = 13, min_samples_split = 20, and min_samples_leaf = 20. Whereas the ratios (80, 20) have a maximum accuracy of 0.9623 with n_estimators = 500, max_deepth = 13, min_amples_split, and min_sample_leaf = 20. And the ration (90;10) has the greatest accuracy of 0,9646, with n_estimates = 500 and max_depth = 13.

After that, some of the trees were visualized, which can be seen in Fig 3 and Fig 4.

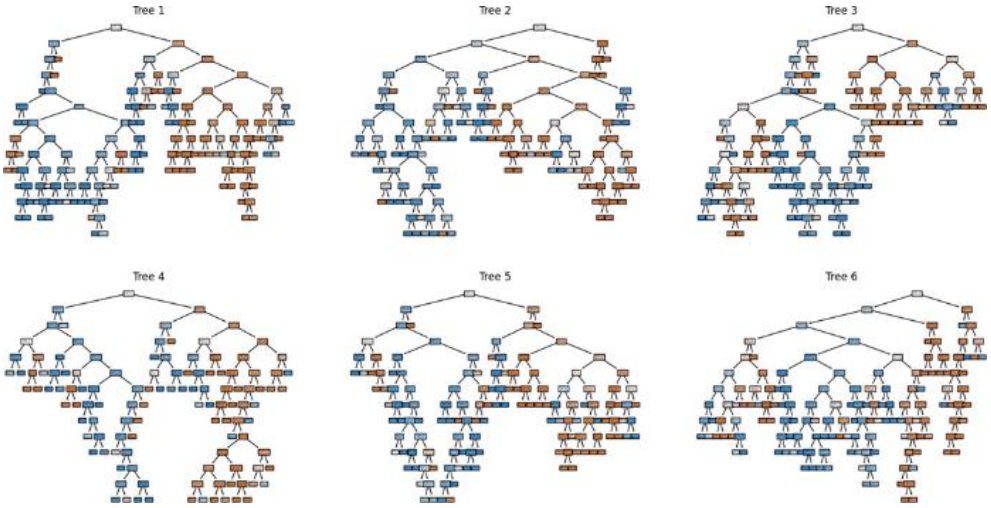

Fig 3 Six decision trees built from a random forest

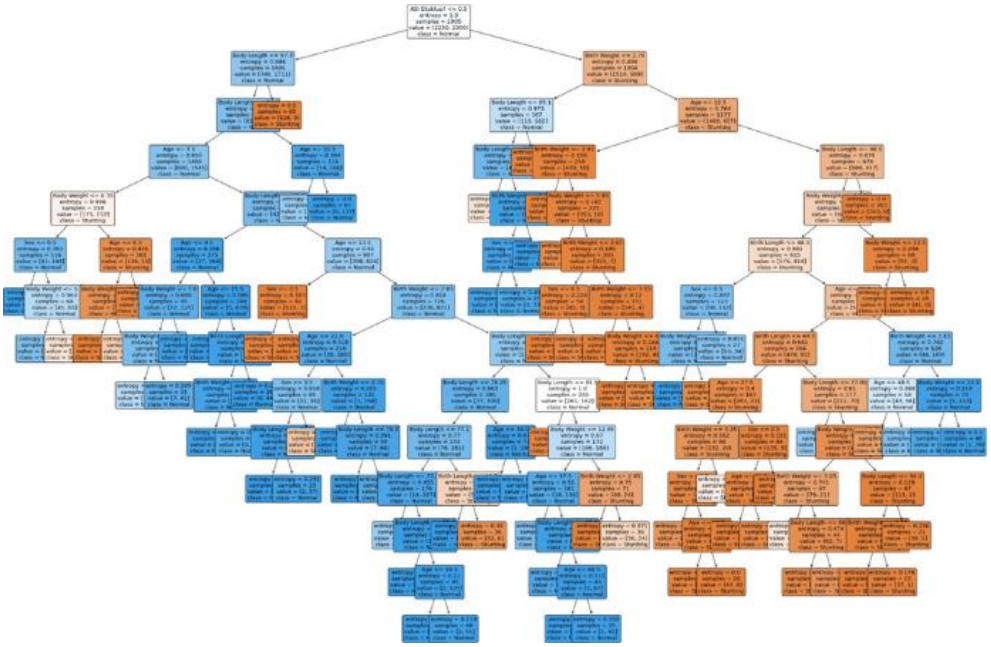

Fig 4 best random forest tree based on the highest accuracy

The visual tree from the random forest is chosen based on the highest accuracy obtained from the test data. That is, the ratio (70; 30) has a maximum precision of 0.9652 with the hyperparameters used, namely n_estimators = 100, max_depth = 13, min_samples_split = 20, and min_samples_leaf = 20. This decision tree visualization provides insight into how models make decisions based on the variables used to determine the stunting status. For example, a tree can start by dividing the data by variables such as gender or age. Then, each branch may represent more specific rules, such as birth weight or birth length. This process will continue until it reaches the leaf node.

The next step is the model evaluation, which is done to measure the accuracy or number of errors that occur in the random forest model of classification. In this study, we used accuracy, precision, recall, and the f1-score. As for the performance of the experiment with the Random Forest model, which has the highest accuracy concerns on each tree, the ratio will be measured, and the evaluation results can be seen in Table 6.

Table 6: Model Evaluation Results

Rasio	N_Estimators	Accuracy	Precision	Recall	F1-Score
70;30	100	0,9651	0,9603	0,9718	0,9660
	200	0,9620	0,9582	0,9678	0,9630
	500	0,9635	0,9620	0,9668	0,9644
80;20	100	0,9584	0,9580	0,9580	0,9580
	200	0,96	0,9567	0,9627	0,9597
	500	0,9623	0,9583	0,9658	0,9621
90;10	100	0,9630	0,9573	0,9691	0,9631
	200	0,96	0,9598	0,9598	0,9598
	500	0,9646	0,9602	0,9691	0,9646

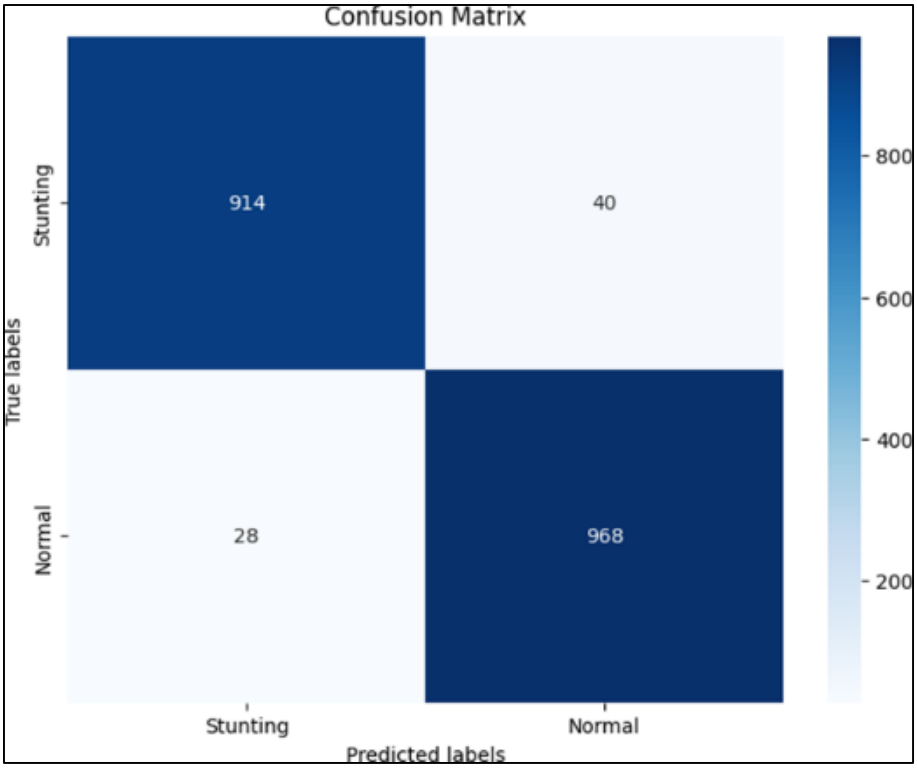

Fig 5: Confusion Matrix

Based on the above table, the ratio (70:30) with n_estimators = 100, max_depth = 13, min_samples_split and min_samples_leaf = 20 has an accuracy of 0.9651 with a precision of 0,9603, recall of 0.9718, and F1-score of 0.9660. whereas for the ratios (80; 20) with a number of N_estimators = 500, max_deepth = 13, and min_sample_splicit and mine_samples_leaf = 20, the accuracy is 0,9623 with a precedence of 0.9583, the recall is 0,9658, and the F1 score is 0.9621.

Meanwhile, Fig 5 shows the performance of a classification model in detecting stunting cases. Based on this figure, the true positive (TP) shows that the number of stunting cases actually identified as stunting by the model is 914. For the false positive (FP), the model evaluation shows that the number of cases is actually normal, but there is an error in identifying stunting with the number of cases being 40. For false negative (FN), the number of stunting cases that were predicted to be normal but were identified as stunting was 28. And for true negative (TN), the number of normal cases that were truly identified as normal by the model was 968.

Discussions

This study demonstrates that using random forest algorithms to classify stunting status on news produces excellent results. With the highest accuracy of 0.9651 on the training data ratio and 70:30 test data, this is significantly better. This advantage is achieved through optimum hyperparameter selection and Random Forest's ability to handle data complexity, allowing stunting identification with high accuracy and recall of 0.9603 and 0.9718, respectively.

The Random Forest model's high performance can help identify children at risk of stunting more accurately, allowing for more effective early intervention. This is important because stunting is a problem that can have long-term effects on children's physical and cognitive development. With this model, health workers can be more accurately targeted in providing treatment and nutritional interventions, so that efforts to reduce stunting numbers can be done more effectively.

Practically, the results of this research can be implemented in the youth health monitoring system. Using predictive models like Random Forest can help healthcare professionals be more proactive in identifying children at risk of stunting so that the necessary diagnosis can be made early. This implementation can be enhanced by training and educating healthcare professionals on how to use this predictive model, as well as integration into existing health systems. Thus, this model will not only help to reduce stunting rates, but it will also contribute to improving the overall quality of life of young people.

The research is expected to make a significant contribution to stunting efforts and form the basis for future development of machine learning-based diagnostic tools.

Conclusions

Based on the research that has been done, it can be confirmed that the random forest algorithm classification in stunting cases based on news data provides good performance. After implementation, we obtained a good accuracy result. The highest accuracy value is 0.9651 with a data ratio of 70–30, n_estimators = 100, max_depth = 13, min_samples_split = 20, and min_samples_leaf = 20. As well as precision 0.9603, recall 0.9718, and F1-score 0.9660. Based on the accuracy, precision, recall, and F1 scores obtained, the classification modeling was evaluated well in classifying stunting status on news using random forest methods. This research is expected to contribute to future stunting reduction strategies. Recommendations for further research include the addition of more complex variables such as midupper arm circumference (MUAC), body mass index (BMI), head circumference, and so on. As well, the random forest algorithm is expected to add feature selection.

References

1. Adzhima F. Klasifikasi Status Stunting Balita dengan Metode Support Vector Machine Berbasis Web; c2023.
2. Azahra FL, Rezeki S, Abrar M, Rizqiya F. Pengukuran Antropometri dan Edukasi Gizi pada Balita di Kelurahan Cipargi, Kecamatan Bogor Utara, Kota Bogor, Jawa Barat. Jurnal Kesehatan Masyarakat Andalas. 2022;14(2):3. DOI: 10.24893/jkma.v14i2.527.
3. Firrahmawati L, Wahyuni ES, Khotimah N, Munawaroh M. Analisis Faktor Penyebab yang Mempengaruhi Kejadian Stunting. Jurnal Kebidanan. 2023;12(1):28-38. Sudipa IGI, Putra TEA, Wahidin AJ, Syukrilla WA, Wardhani AK, Heryana N, et al. Data Mining. 1st ed; c2023. www.globaleksekutifteknologi.co.id,
4. Hendrian S. Algoritma Klasifikasi Data Mining Untuk Memprediksi Siswa Dalam Memperoleh Bantuan Dana Pendidikan. Faktor Exacta. 2018;11(3):266-274.
5. Juwariyem, Sriyanto. Prediksi Stunting pada Balita Menggunakan Algoritma Random Forest. Journal IndraTech. 2023;4(1):29-37.
6. Kemenkes RI. Hasil Survei Status Gizi Indonesia (SSGI) 2022. Badan KebijakanPembangunan Kesehatan. Kementerian Kesehatan Republik Indonesia. Jakarta; c2023.
7. Kusumarini AI, Hogantara PA, Chamidah N. Perbandingan Algoritma Random Forest, Naïve Bayes, dan Decision Tree dengan Oversampling untuk Klasifikasi Bakteri E. Coli. In: Seminar Nasional Mahasiswa Ilmu Komputer dan Aplikasinya (SENAMIKA). Jakarta-Indonesia; c2021.
8. Liantoni F. Data Mining dan Penerapan Metode; c2022.
9. Miftahusalam A, Nuraini AF, Khoirunisa AA, Pratiwi H. Perbandingan Algoritma Random Forest, Naïve Bayes, dan Support Vector Machine pada Analisis Sentimen Twitter Mengenai Opini Masyarakat terhadap Penghapusan Tenaga Honorer; c2022.
10. Muslim MA, Prasetiyo B, Nurzahputra A. Buku Data Mining; c2019.
11. Nugroho MR, Sasongko RN, Kristiawan M. Faktor-faktor yang mempengaruhi kejadian stunting pada anak usia dini di Indonesia. Jurnal Obsesi: Jurnal Pendidikan Anak Usia Dini. 2021;5(2):2269-2276.
12. Pahlevi O, Amrin A, Handrianto Y. Optimasi Algoritma Naïve Bayes Berbasis Particle Swarm Optimization Untuk Klasifikasi Status Stunting. Computer Science (CO-SCIENCE). 2024;4(1):37-43.
13. Perdana AY, Latuconsina R, Dinimaharawati A. Prediksi Stunting Pada Balita Dengan Algoritma Random Forest. eProceedings of Engineering. 2021, 8(5).
14. Ratumanan SP, Achadiyani, Khairani AF. Metode Antropometri untuk Menilai Status Gizi: Sebuah Studi Literaturl; c2023. Available from: https://myjurnal.poltekkes-kdi.ac.id/index.php/hijp.
15. Widiastuti RN. Bersama Perangi Stunting; c2019.
16. Setiawan BR. Sistem Klasifikasi Stunting Berbasis Web Menggunakan Metode Naive Bayes; c2023.
17. Situmorang S, Yahfizham Y. Analisis Kinerja Algoritma Machine Learning Dalam Deteksi Anomali Jaringan. Konstanta: Jurnal Matematika Dan Ilmu Pengetahuan Alam. 2023;1(4):258-269.
18. Putra W. Pengenalan Konsep Pembelajaran Mesin dan Deep Learning Edisi 1.4; c2020.
19. World Health Organization. Reducing stunting in children: equity considerations for achieving the Global Nutrition Targets; c2025.
20. World Health Organization. Levels and trends in child malnutrition; c2023.
21. Leksono AW, Prameswary DK, Pembajeng GS, Felix J, Dini MAS, Rahmadina N, *et al*. Risiko Penyebab Kejadian Stunting pada Anak. Jurnal Pengabdian Kesehatan Masyarakat: Pengmaskesmas. 2021;1(2):34-38. DOI: 10.31849/pengmaskesmas.v1i2/5747.