Jurnal Ilmiah Multidisiplin Indonesia (JIM-ID)
Volume 4, Number 07, 2025, DOI 10.58471/esaprom.v4i07
ESSN 2828-9463 (Online)
https://ejournal.seaninstitute.or.id/index.php/esaprom

Comparative Analysis of the C4.5 and Random Forest Algorithms for the Prediction of Diarrheal Disease

Sipra Barutu[1], Muhammad Iqbal[2], Khairul[3], Darmeli Nasution[4]
Universitas Pembangunan Pancabudi, Medan, Indonesia

Article Info
Keywords:
diarrhea, toddlers, machine learning, C4.5 algorithm, random forest, disease prediction

ABSTRACT
Diarrhea remains one of the leading causes of death among infants in Indonesia, especially in areas with limited access to healthcare. Environmental pollution and unhealthy lifestyles are the main causes of its spread. This study aims to compare the performance of the C4.5 and Random Forest algorithms in predicting diarrhea cases among infants in the working area of the Parlilitan Subdistrict Health Center, Humbahas Regency, North Sumatra Province. Secondary data were obtained from medical records and health center reports, which were then analyzed using Python. Model performance evaluation was conducted using the metrics Accuracy, Precision, Recall, F1-Score, Specificity, False Positive Rate (FPR), and True Positive Rate (TPR). The test results showed that the C4.5 algorithm had superior performance with an Accuracy of 0.92; Precision, Recall, and F1-Score of 0.875 each; Specificity of 0.9412; and FPR of 0.0588. Meanwhile, Random Forest obtained an Accuracy of 0.88; Precision of 0.7778; Recall of 0.875; F1-Score of 0.8235; Specificity of 0.8824; and FPR of 0.1176. These findings indicate that C4.5 is more effective in maintaining a balance between prediction accuracy and detection capability, and is better at minimizing classification errors for negative classes.

Corresponding Author:
Sipra Barutu
Universitas Pembangunan Pancabudi, Medan, Indonesia
barutusipra@gmail.com

INTRODUCTION

According to data from the WHO and UNICEF, there are an estimated 2 billion cases of diarrhea each year, resulting in approximately 1.9 million deaths among children under five. Of these deaths, 78% occur in developing countries, particularly in Africa and Southeast Asia. (WHO and UNICEF, 2024).

In Indonesia, based on the 2018 Basic Health Research, the prevalence of diarrhea in toddlers was recorded at 12.3%, with a prevalence rate of 10.6% for infants. These figures indicate that diarrhea remains one of the leading causes of death among infants, particularly in areas with limited access to healthcare (Ministry of Health of the Republic of Indonesia, 2018) (Kemenkes, 2018).

Persistent diarrhea in toddlers can cause stunting, a condition of malnutrition that is dangerous for children's growth and development. Data from the 2020 Indonesian Health Profile shows that diarrhea contributes significantly to mortality in toddlers, with deaths due to diarrhea in the toddler group reaching 4.55% (Indonesian Health Profile, 2020).

Jurnal Ilmiah Multidisiplin Indonesia (JIM-ID)
Volume 4, Number 07, 2025, DOI 10.58471/esaprom.v4i07
ESSN 2828-9463 (Online)
https://ejournal.seaninstitute.or.id/index.php/esaprom

The detection rate of diarrhea cases among toddlers in Indonesia remains low. In 2021, the detection rate of diarrhea cases among toddlers only reached 22.18% of the set target, which is approximately 818,687 toddlers out of a target of 3,690,984. This indicates significant challenges in achieving effective health program targets, with various factors such as frequent changes in program managers, low accuracy and completeness of reports, and lack of data integration between programs (Ministry of Health of the Republic of Indonesia, 2022).

Environmental factors, such as inadequate clean water and sewage facilities, are the main causes of diarrhea. The combination of contaminated environmental factors and unhealthy human behavior, such as poor food and beverage hygiene, exacerbates the spread of this disease (Indonesian Health Profile, 2020).

Based on the information presented, it is necessary to treat diarrhea in toddlers using technology, one of which is through predictive analysis using machine learning algorithms. The C4.5 and Random Forest algorithms are two techniques that are widely used in disease classification. Both have been proven effective in various previous studies, such as the classification of heart disease and other diseases. A related study conducted by Sepharni (2022) used C4.5 in heart disease classification, achieving an accuracy rate of 75.92%. Meanwhile, research by Depari, Widiastiwi, and Santoni (2022) showed that the Random Forest algorithm had the highest accuracy in heart disease classification, at 75%.

Diarrhea is a serious health problem worldwide, including in Indonesia, with high prevalence among toddlers and significant mortality rates, especially in areas with limited access to healthcare. Polluted environmental factors and unhealthy behaviors exacerbate the spread of this disease. Despite various prevention efforts, the detection rate of diarrhea cases among toddlers remains low, indicating major challenges in achieving health program targets. To enhance the effectiveness of management, the application of technology, such as predictive analysis using machine learning algorithms, is necessary to more accurately estimate the occurrence of diarrhea. Although the C4.5 and Random Forest algorithms have been widely applied in disease classification, their application for diarrhea in infants remains limited. Therefore, this study aims to compare the two algorithms in predicting diarrhea cases in infants in the UPT Puskesmas Kec. Parlilitan Kab. Humbahas, North Sumatra Province.

Although the C4.5 and Random Forest algorithms are used for disease prediction, their application in predicting diarrhea in infants is still limited. Therefore, this study was conducted to fill this gap by analyzing the comparison between the two algorithms in predicting the incidence of diarrhea in infants in the working area of the Parlilitan District Health Center, Humbahas Regency, North Sumatra Province. The results of this study are expected to provide important contributions to efforts to prevent and control diarrhea in infants through more accurate and targeted predictions. The objectives of this study are as follows: To present data that will be used for predicting diarrhea. To apply the C4.5 algorithm to determine the prediction pattern of diarrhea. To apply the Random Forest algorithm to determine the prediction pattern of diarrhea. To apply precision, recall, and F1-score techniques as part of the comparison and evaluation of the algorithms used.

Jurnal Ilmiah Multidisiplin Indonesia (JIM-ID)
Volume 4, Number 07, 2025, DOI 10.58471/esaprom.v4i07
ESSN 2828-9463 (Online)
https://ejournal.seaninstitute.or.id/index.php/esaprom

METHOD

The research method used is a quantitative approach with an experimental design. The data used is secondary data obtained from medical records and health reports available at the Community Health Center. Next, the data will be processed and analyzed using the C4.5 and Random Forest algorithms to build a prediction model that can provide information about the factors that influence the incidence of diarrhea in toddlers.

The framework of this study is explained through the design flow depicted in Figure 1, which illustrates the stages that will be followed in the study to achieve the established objectives. This design flow includes systematic steps covering problem identification, data collection, data pre-processing, algorithm application, model performance evaluation, and analysis of the results produced.

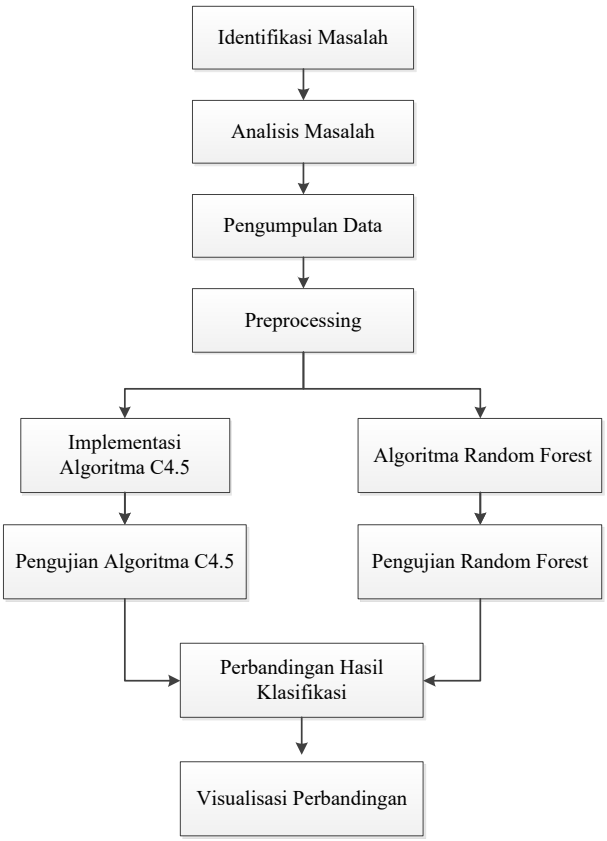

Figure 1 Research Framework

Application of C4.5 and Random Forest Algorithms

At this stage, the steps in solving the problem are arranged logically and in an easy-to-understand manner.

Jurnal Ilmiah Multidisiplin Indonesia (JIM-ID)
Volume 4, Number 07, 2025, DOI 10.58471/esaprom.v4i07
ESSN 2828-9463 (Online)
https://ejournal.seaninstitute.or.id/index.php/esaprom

Implementation of C4.5

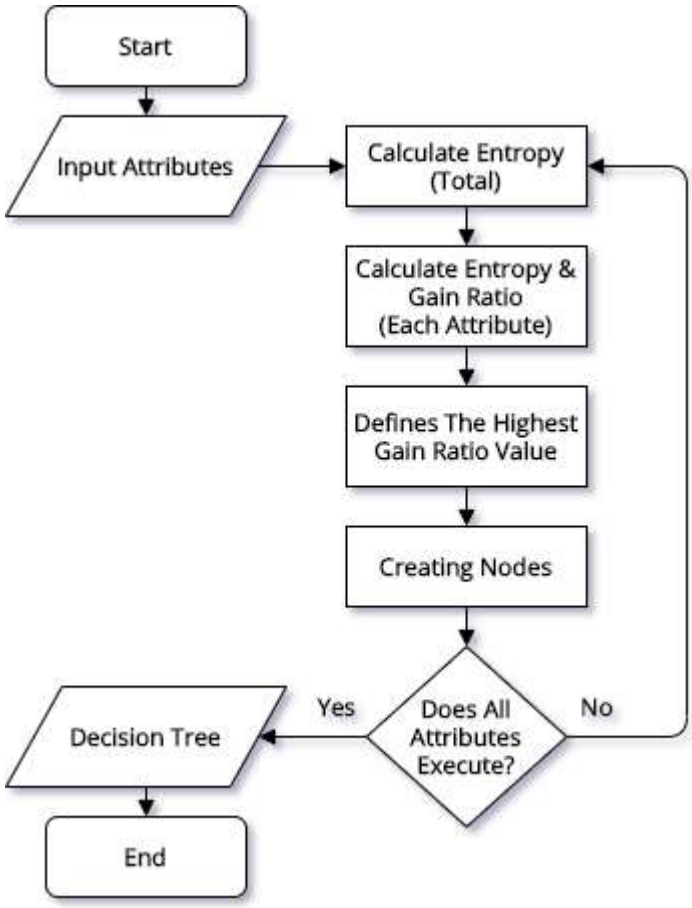

Figure 2. C45 flowchart

$$Enthropy(S) - (x+a)^n = \sum_{j=1}^{k} -\mathrm{P\ j\ Log2\ P\ j}$$

Explanation:
S : set of cases
A : attribute
n : number of partitions of attribute A
\Si\ : Proportion of Si to S
|Sr| : number of cases in S
And the second formula is to find the Entropy value, which is:

$$Enthropy(S) - (x+a)^n = \sum_{j=1}^{k} -\mathrm{P\ j\ Log2\ P\ j}$$

Explanation:
S = set of cases
n = number of partitions of S
Pi = proportion of Si to S

Jurnal Ilmiah Multidisiplin Indonesia (JIM-ID)
Volume 4, Number 07, 2025, DOI 10.58471/esaprom.v4i07
ESSN 2828-9463 (Online)
https://ejournal.seaninstitute.or.id/index.php/esaprom

Implementation of Random Forest

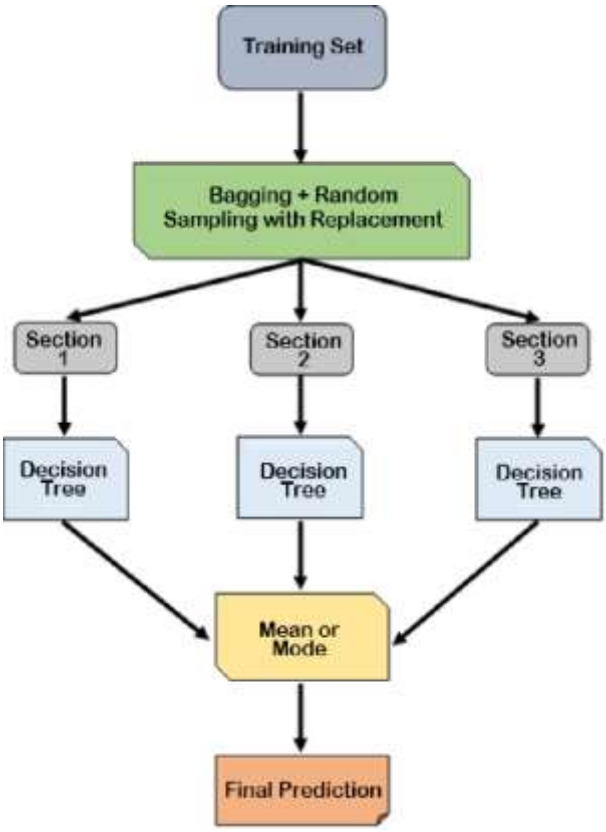

Figure 3. Random forest flowchart

$$Gini(s_i) = 1 - \sum_{i=0}^{c-1} p_1^2$$

where pi is the relative frequency of class Ci in the set. Ci is the class for i = 1, ..., c-1, and c is the number of classes that have been determined. The quality of the split on feature k into subset Si is the number of samples belonging to class Ci, then calculated as the sum of the Gini index considerations of the resulting subset. The data can be calculated using the formula

$$Gini_{Split} = \sum_{i=0}^{k-1} (\frac{n_i}{i})\ Gini(s_i)$$

where ni is the number of samples in subset Si after splitting and n is the number of samples in the given node.

Suppose {h(x, Θk), k= 1, ...} where { Θk} is an independent identically distributed (iid) random vector and each tree chooses the class that is most common on average (majority vote). For RF, the upper bound can be derived for generalization error in terms of two parameters that measure how strong the individual classifications are and the dependence between them (Breiman, 2001). The margin function for RF is

$$mr(X, Y) = P_\Theta(h(X, \Theta) = Y) - \max P_\Theta (h(X, \Theta) = j)\ j \neq Y$$

Jurnal Ilmiah Multidisiplin Indonesia (JIM-ID)
Volume 4, Number 07, 2025, DOI 10.58471/esaprom.v4i07
ESSN 2828-9463 (Online)
https://ejournal.seaninstitute.or.id/index.php/esaprom

and the power of the classifier set {h (X, Θ)} is

$$s = E_{X,Y}mr(X, Y)$$

Assuming s ≥ 0, Chebychev's inequality and the variance reduction mr of the margin function for the RF method, the upper bound of generalization error can be obtained as follows

$$PE \leq \frac{\bar{\rho}(1-s^2)}{s^2}$$

Where $\bar{\rho}$ is the average correlation value, namely:

$$\bar{\rho} = \frac{E_{\theta,\theta'}\left(\rho(\theta,\theta')sd(\theta)sd(\theta')\right)}{E_{\theta,\theta'}\left(sd(\theta)sd(\theta')\right)}$$

Classification Accuracy Evaluation

Confusion Matrix

A confusion matrix is a basic tool for evaluating the performance of a classification model. This matrix shows the number of correct and incorrect predictions in each class category. It consists of four main components, which are shown in the following table:

Tabel 1. Confusion Matrix

Confusion matrix	Description
True Positive (TP)	Data that is truly positive and predicted to be positive by the model.
True Negative (TN)	Data that is truly negative and predicted to be negative by the model.
False Positive (FP)	Data that is truly negative but predicted to be positive by the model.
False Negative (FN)	Data that is truly positive but predicted to be negative by the model

The matrix presented is the first step used to perform accuracy calculations with precision, recall, and F1-score, which can be calculated to provide a clearer picture of the model's performance.

1. ccuracy
   ccuracy measures the proportion of correct predictions compared to the total number of predictions using the formula
   $$Akurasi = \frac{TP + TN}{TP + TN + FP + FN}$$
2. Precision Theory
   Precision is important in situations where false positives (FP) are unacceptable. For example, in email spam detection, the model must minimize the number of emails that are incorrectly classified as spam.
   $$Presisi = \frac{TP}{TP + FP}$$
3. Recall
   Recall is important in situations where false negatives (FN) must be avoided. For example, in medical diagnosis, the model must ensure that as many cases of disease as possible are detected, even if that means increasing the likelihood of some negative cases being misclassified. The recall formula is as follows:

Jurnal Ilmiah Multidisiplin Indonesia (JIM-ID)
Volume 4, Number 07, 2025, DOI 10.58471/esaprom.v4i07
ESSN 2828-9463 (Online)
https://ejournal.seaninstitute.or.id/index.php/esaprom

$$Recall = \frac{TP}{TP + FN}$$

4. F1-Score Theory

   F1-score is a better metric when we want to maintain a balance between precision and recall. F1-score is very useful when there is imbalance in the data and helps reduce overfitting or underfitting issues by considering both aspects of performance (precision and recall) simultaneously with the formula

$$F1 - Score = 2x\frac{Presisi\ x\ Recall}{Presisi + Recall}$$

RESULTS AND DISCUSSION

The data used in the testing using Python consisted of 150 training data and 50 test data. Testing of the C4.5 and Random Forest models was carried out to evaluate the classification performance in detecting or predicting target conditions based on the available input variables. Both models were chosen because they represent decision tree approaches that are widely used in data analysis, particularly in the context of rule-based classification. The variables used in the study were systematically analyzed to support the clarity of the data analysis process. Each variable, ranging from water quality, environmental sanitation, hand hygiene, immunization, to diarrhea status.

Testing C.45 with Python

The resulting tree structure provides an overview of how combinations of attributes such as Environmental Sanitation, Drinking Water Quality, and Hand Hygiene of Mothers collectively influence the classification results. This demonstrates the effectiveness of the C4.5 algorithm in identifying patterns from data through the selection of the most informative attributes at each branching level.

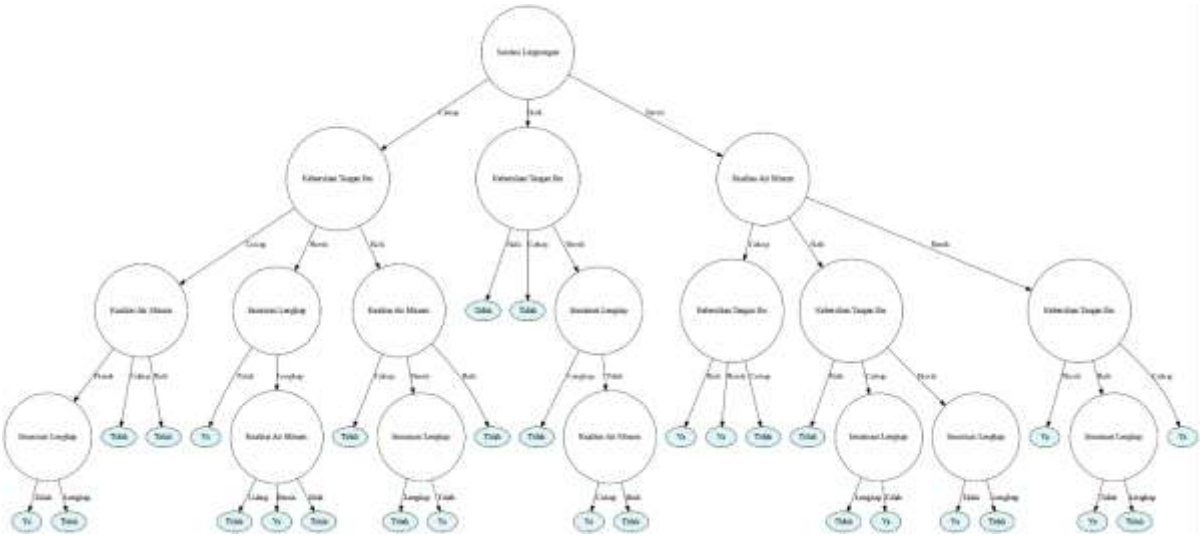

Figure 4. Decision Tree With c45

Jurnal Ilmiah Multidisiplin Indonesia (JIM-ID)
Volume 4, Number 07, 2025, DOI 10.58471/esaprom.v4i07
ESSN 2828-9463 (Online)
https://ejournal.seaninstitute.or.id/index.php/esaprom

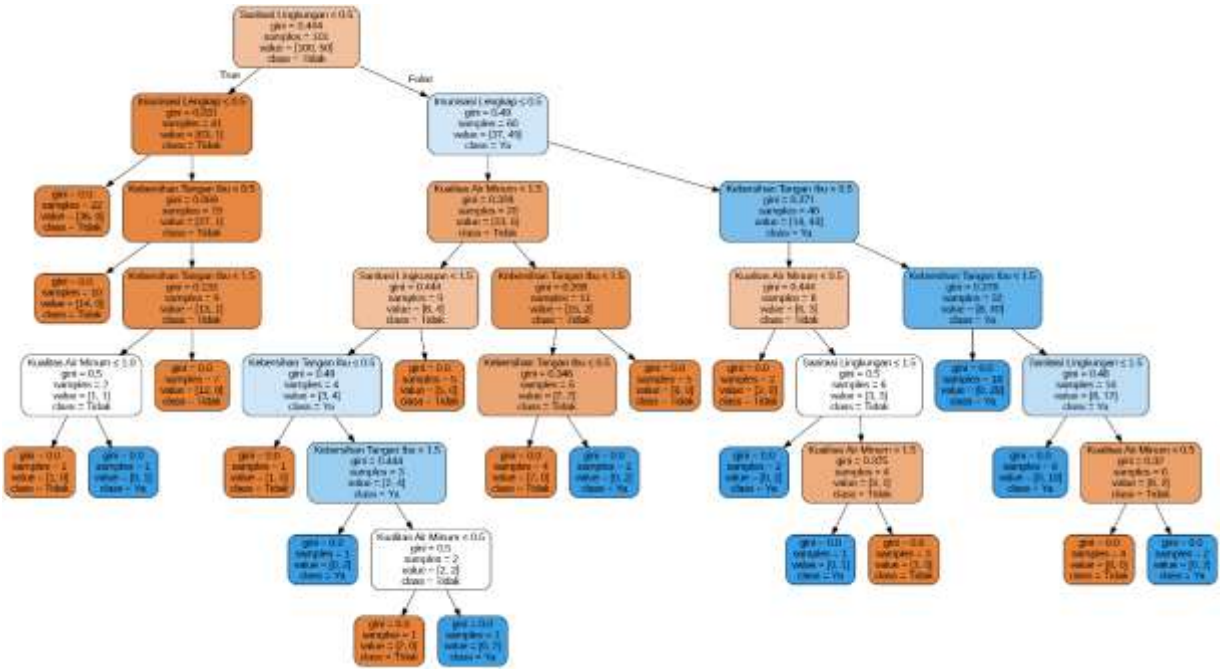

Figure 5. Decision Tree with Random Forest

Comparative Evaluation Analysis

The comparison was conducted using a number of evaluation metrics, including Accuracy, Precision, Recall (Sensitivity/TPR), F1-Score, Specificity (TNR), False Positive Rate (FPR), and True Positive Rate (TPR). Each metric provides a different picture of the model's performance, both in terms of its ability to identify positive and negative data.

Table 1. Interpretation of the performance results of the two algorithms

Metrik	C4.5	Random Forest
Accuracy	0,92	0,88
Precision	0,875	0,7778
Recall (Sensitivity / TPR)	0,875	0,875
F1-Score	0,875	0,8235
Specificity (TNR)	0,9412	0,8824
False Positive Rate (FPR)	0,0588	0,1176
True Positive Rate (TPR)	0,875	0,875

Jurnal Ilmiah Multidisiplin Indonesia (JIM-ID)
Volume 4, Number 07, 2025, DOI 10.58471/esaprom.v4i07
ESSN 2828-9463 (Online)
https://ejournal.seaninstitute.or.id/index.php/esaprom

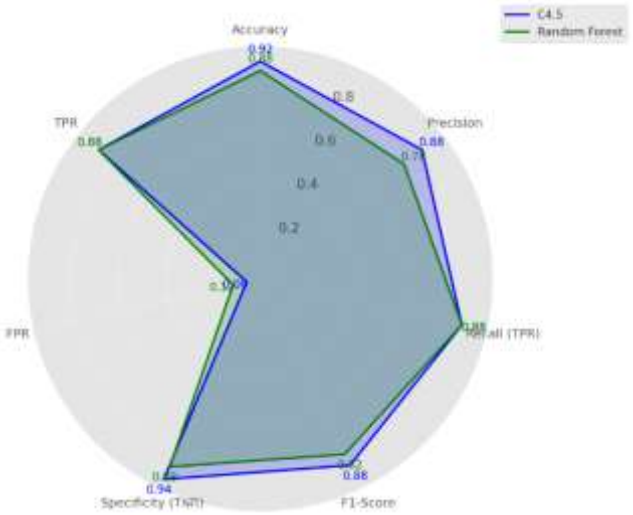

Figure 6. Comparison of C4.5 and Random Forest Model Performance

From the image, it can be seen that the C4.5 algorithm consistently performs better than Random Forest in most metrics. For example, C4.5 has an Accuracy value of 0.92, higher than Random Forest, which only reaches 0.88. A similar trend is observed for Precision (0.875 vs 0.7778), F1-Score (0.875 vs 0.8235), and Specificity (0.9412 vs 0.8824). Although both algorithms have the same Recall/TPR value (0.875), Random Forest has a weakness in False Positive Rate, which is 0.1176, much higher than C4.5 with an FPR of 0.0588. This shows that Random Forest tends to give false positives more often than C4.5.

Each point in the image represents one metric, and the values of each model are shown through colored lines that are interconnected to form a closed pattern. Blue is used to indicate the performance of the C4.5 algorithm, while green is used for Random Forest. In addition, each point is also accompanied by a numerical value to facilitate the reading of the actual values of each metric. The blue line pattern, which appears wider and covers a larger area, indicates that this model is more optimal in maintaining a balance between sensitivity (Recall) and prediction accuracy (Precision), and more effective in reducing the error rate.

CONCLUSION

Based on the evaluation results of the C4.5 and Random Forest algorithms using various measurement metrics, the following conclusions can be drawn: The C4.5 algorithm shows excellent classification performance with an accuracy value of 0.92. The precision, recall, and F1-score values, all at 0.875, indicate a balance between the ability to detect positive classes and the accuracy of the predictions provided. The high specificity of 0.9412 indicates that the algorithm is capable of effectively identifying negative data, supported by a low false positive rate of 0.0588. The Random Forest algorithm provides an accuracy of 0.88. Although the recall remains high at 0.875, the precision is only 0.7778, which results in a decrease in the F1-score to 0.8235. The specificity obtained is 0.8824, which is still quite good, although lower than C4.5, with a higher FPR of 0.1176. This shows that although Random Forest is quite reliable in detecting positive classes, there is a weakness in avoiding false predictions

of negative data. Overall, the C4.5 algorithm is superior to Random Forest in almost all evaluation metrics. Higher accuracy, precision, F1-score, and specificity values indicate that C4.5 is more efficient and consistent in the classification process. Although both algorithms show the same recall and TPR values, C4.5's ability to minimize classification errors for negative classes makes it a better choice in the context of the tested data. Based on the findings of this study, the following recommendations are suggested for consideration in future research. The performance of the Random Forest algorithm still has the potential to be improved through parameter adjustments, such as the number of trees (n_estimators), maximum tree depth, and random feature selection at each split. Additionally, the application of more optimal data preprocessing techniques can contribute to improved classification accuracy. Further research is recommended to test both algorithms on datasets with different characteristics, such as imbalanced data distribution or higher dimensions, to evaluate the stability and generalization of the model more comprehensively.

REFERENCES

Kementerian Kesehatan Republik Indonesia. (2018). Riset Kesehatan Dasar 2018. Jakarta: Kemenkes RI.

Kementerian Kesehatan Republik Indonesia. (2020). Profil Kesehatan Indonesia 2020. Jakarta: Kemenkes RI.

Kementerian Kesehatan Republik Indonesia. (2022). Laporan Kesehatan Nasional 2022. Jakarta: Kemenkes RI.

Sepharni. (2022). Klasifikasi Penyakit Jantung Menggunakan Algoritma C4.5. Jurnal Informatika, 10(2), 75-92.

Depari, Widiastiwi, & Santoni. (2022). Perbandingan Algoritma Machine Learning dalam Klasifikasi Penyakit Jantung. Jurnal Kesehatan Digital, 15(3), 70-85.

Munggaran, & Hidayatulloh. (2015). Penerapan Algoritma C4.5 untuk Diagnosa Penyakit Diare Pada Anak Balita Berbasis Mobile. Jurnal Sistem Informasi, 8(1), 55-67.

Ente, et al. (2020). Klasifikasi Faktor-Faktor Penyebab Penyakit Diabetes Melitus Di Rumah Sakit Unhas Menggunakan Algoritma C4.5. Jurnal Ilmu Komputer, 12(2), 98-112.

Afifuddin, & Hakim. (2023). Deteksi Penyakit Diabetes Mellitus Menggunakan Algoritma Decision Tree Model Arsitektur C4.5. Jurnal Teknologi Informasi, 19(1), 33-45.

Prabowo, et al. (2023). Komparasi Tingkat Akurasi Random Forest dan Decision Tree C4.5 Pada Klasifikasi Data Penyakit Infertilitas. Jurnal Kesehatan Digital, 17(4), 88-102.

Kalimah. (2022). Klasifikasi Penyakit Diabetes Menggunakan Metode Decision Tree dan Random Forest. Jurnal Informatika Medis, 14(3), 67-79.

Putra, & Handayani. (2024). Perbandingan Algoritma Decision Tree dan Random Forest Dalam Pengklasifikasian Penyakit Tiroid. Jurnal Data Science, 22(1), 55-68.

Aditya, et al. (2024). Prediksi Penyakit Hipertensi Menggunakan Metode Decision Tree dan Random Forest. Jurnal AI & Kesehatan, 16(2), 100-115.

Masriadi. (2017). Epidemiologi Penyakit Diare. Makassar: Universitas Hasanuddin Press.

Purnama. (2016). Penyakit Diare dan Faktor Risikonya. Jakarta: Pustaka Kesehatan.

Simatupang. (2004). Rotavirus dan Perannya dalam Diare pada Anak. Jurnal Kedokteran Indonesia, 10(2), 44-55.

Nikma Kumala Sari, & Almansyah Lukito. (2017). Faktor Penyebab dan Pencegahan Diare pada Balita. Jurnal Kesehatan Masyarakat, 12(1), 78-91.

Hassan, & Alatas. (1985). Patogenesis dan Pencegahan Diare pada Anak. Jakarta: Balai Pustaka.

Kliegman, Marcdante, & Jenson. (2006). Nelson Textbook of Pediatrics. Philadelphia: Elsevier.

Breiman, L. (2001). *Random Forests*. Machine Learning, 45(1), 5-32.

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann.

Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann.

Hamzah, I., & Sitorus, Z. (2024). Analisa Classification Decision Tree C45 dan Naïve Bayes Pada Indikasi Penyakit Diabetes Menggunakan Rapid Miner. Jurnal Nasional Teknologi Komputer, 4(1), 25-33.

Iqbal, M., & Efendi, S. (2023). Data-driven approach for credit risk analysis using C4. 5 algorithm. *ComTech: Computer, Mathematics and Engineering Applications*, *14*(1), 11-20.

Hamzah, I., & Sitorus, Z. (2024). Analisa Classification Decision Tree C45 dan Naïve Bayes Pada Indikasi Penyakit Diabetes Menggunakan Rapid Miner. *Jurnal Nasional Teknologi Komputer*, *4*(1), 25-33.