**JTAM (Jurnal Teori dan Aplikasi Matematika)**
http://journal.ummat.ac.id/index.php/jtam
p-ISSN 2597-7512 | e-ISSN 2614-1175
Vol. 8, No. 4, October 2024, pp. 1121-1132

# Comparative Analysis of Decision Tree and Random Forest Algorithms for Diabetes Prediction

**Aufar Faiq Fadhlullah[1], Triyanna Widiyaningtyas[1]**
[1]Department of Electrical Engineering and Informatics, Universitas Negeri Malang, Indonesia
aufar.faiq.2005356@students.um.ac.id

**Article History:**
Received : 10-06-2024
Revised : 11-09-2024
Accepted : 18-09-2024
Online : 01-10-2024

**Keywords:**
Diabetes Mellitus;
Prediction Algorithm;
Random Forest;
Decision Tree.

**ABSTRACT**

Diabetes Mellitus is a long-term medical disorder marked by high blood glucose levels that raise the risk of early mortality and organ failure. It has become an increasing global health problem, so making an accurate and timely diagnosis is urgently necessary. This study aims to diagnose people with diabetes mellitus by utilizing prediction techniques in data mining using experimental research. The prediction stage for diagnosing diabetes consists of four stages: dataset collection, data pre-processing, data processing, and evaluation. Data was obtained from Electronic Health Records (EHRs), namely the public "Diabetes Prediction Dataset". The pre-processing stage involves data filtering, attribute conversion, and class selection. The data processing utilizes random forests and decision tree models for diabetes prediction. The models were evaluated using accuracy, precision, and recall metrics. The results showed that the Random Forest algorithm produced an accuracy value of 93.97%, precision of 99.88%, and recall of 66.56%, with a computational time of 16s. Meanwhile, the decision tree algorithm produces an accuracy value of 93.89%, precision of 98.73%, and recall of 66.88%, with a computation time of less than 1s. Based on these results, it can be concluded that the Decision Tree algorithm is more effective because the difference in accuracy, precision, and recall values produced by the two algorithms does not have significant differences. However, the Decision Tree algorithm has the advantage of using computational time more effectively, which is needed in detecting diabetes because it is related to someone's life.

https://doi.org/10.31764/jtam.v8i4.24388



## A. INTRODUCTION

Diabetes mellitus is a chronic metabolic disease caused by increased blood glucose levels. When glucose is not absorbed by the body's cells adequately, it builds up in the blood and can cause several problems with organ function. Diabetes raises the chance of premature mortality and can lead to problems in numerous areas of the body. It is a global health issue that continues to escalate each year. Over the past several decades, the number of cases and prevalence of diabetes has increased. According to the Institute for Health Metrics and Evaluation (2019), diabetes ranks third as the leading cause of death in Indonesia, with a mortality rate of 57.42 deaths per 100,000 population. This high number is attributed to insufficient awareness among the public regarding the risk factors that contribute to diabetes (Sękowski et al., 2022). Diabetes is a chronic illness for which there is now no treatment. On the other hand, diabetes can be delayed or prevented from progressing into acute phases by being detected early. Therefore, it is imperative and extremely advantageous to conduct research on the speedy and accurate detection of diabetes.

Classification plays a key role in the diagnostic and predictive process (Stoleru & Iftene, 2022). Prediction is a system used to make decisions by extracting information from data available in the real world based on temporary attributes (Singh & Jaiswal, 2022). Prediction involves a systematic process of estimating what is most likely to happen in the future based on information from the past and present, following a structured approach (Mehraeen et al., 2023). Prediction systems require a classification approach to serve as a predictive method, guiding future decision-making. In this process, specific entities are grouped based on their characteristics or features (Ismail et al., 2022). One effective method for classification and prediction is data mining.

Data Mining is a process of sifting through information to identify significant patterns within large datasets in a database, leading to knowledge discovery (Durugkar et al., 2022). Data mining techniques allow us to process and classify data based on collected information (Garg, 2023). The functions in data mining are description, estimation, prediction, classification, clustering, and association (Kumari et al., 2021). Data mining consists of three stages: data collection, data transformation, and data analysis. Initially, pre-processing is performed by gathering data to produce raw data that can be processed for data mining. It may involve techniques such as filtering or aggregation. The data transformation results can then be utilized as knowledge using machine learning and information visualization (Moreno-Lumbreras et al., 2023).

In several studies, numerous experiments have been conducted using data mining to predict and classify diseases using various algorithms such as Decision Tree and Random Forest. Pyayt, Khan, and Gubanov use these two algorithms to classify some bacteria. The results revealed that the Random Forest algorithm achieved 90.7% precision, 94.4% recall, and 92.5% f-measure, then the Decision Tree algorithm achieved 96% precision, 100% recall, and 95.2% f-measure (Pyayt et al., 2020). In addition, Yilmaz and Yagin conducted a performance comparison of the Random Forest, Logistic Regression, and SVM algorithms for classifying coronary heart disease data. The Random Forest's algorithm had the highest average accuracy rate of 92.9% with a Specificity of 92.9%, Sensitivity of 92.8%, F1-score of 92.8%, Negative Predictive Value of 92.9%, and Positive Predictive Value of 92.8% (YILMAZ & YAĞIN, 2022). Maulana et al. utilize ZGBoost algorithm to detect diabetes. Their research result in an accuracy of 82.68.

Based on the high accuracy of the Random Forest and Decision Tree algorithms from previous research, these methods can be applied to predict an individual's risk of developing diabetes. However, this approach must be supported by data that can be classified using the Random Forest and Decision Tree algorithms, allowing us to obtain accuracy rates. Subsequently, a comparison between the two algorithms can reveal their efficiency in the early detection of diabetes patients. This study aims to compare the Random Forest and Decision Tree models in diagnosing people with diabetes mellitus.

## B. RESEARCH METHOD

This study compares the Random Forest and the Decision Tree algorithm to predict people with diabetes mellitus. The stages of our research consist of data collection, pre-processing data, processing data, and evaluation. The details of each stage can be explained below.

### 1. Data Collection

The data collection phase aims to gather the necessary data for research. This study's algorithm implementation requires a dataset containing patients' medical history and demographic information. Based on the researcher's criteria, the chosen dataset for this study is the "Diabetes Prediction Dataset." The dataset can be utilized to construct machine-learning models for diagnosing diabetes based on demographic information and medical history. Healthcare practitioners can use this dataset to identify people who may be at risk of diabetes and to create individualized treatment programs. Researchers may also utilize the dataset to investigate associations between other demographic and medical characteristics and the risk of getting diabetes.

This dataset is secondary and sourced from www.kaggle.com, a platform that provides publicly available datasets for researchers in predictive systems. The Diabetes Prediction Dataset contains patients' medical and demographic information and their diabetes status (positive or negative). The data includes 100,000 diabetes patient data containing nine variables: age, gender, body mass index (BMI), blood pressure, heart disease, smoking history, HbA1c level, blood glucose levels, and diabetes class.

The average age of the patients was 48.05 years, with an age range of 21 to 81 years. The distribution of the patients includes 52.4% aged 21-44 years, 12.2% aged 45-54 years, 16.8 % aged 55-64 years, and 18.6% aged >=65 years. Based on gender, the distribution of the patients includes 59% female and 41% male. Body mass index (BMI) averaged 31.99, with a BMI range from 18.2 to 67.1. The mean blood glucose level was 120.89 mg/dL, from 44 to 394 mg/dL. Blood pressure averaged 69.11 mm Hg, ranging from 24 to 122 mm Hg. The mean HbA1c level was 8.21%, from 4.3 to 14.8%. In addition, this dataset has two target categories: diabetes positive (8,500 patients) and diabetes negative (91,500 patients). Notably, all values and matched data are in the utilized dataset, ensuring the validity of the entire dataset.

### 2. Data Pre-Processing

Data pre-processing involves filtering, modifying attributes, and selecting relevant classes based on the researcher's requirements. Figure 1 illustrates the operators used in this research to assist the data mining before data is processed through model performance evaluation, as shown in Figure 1.

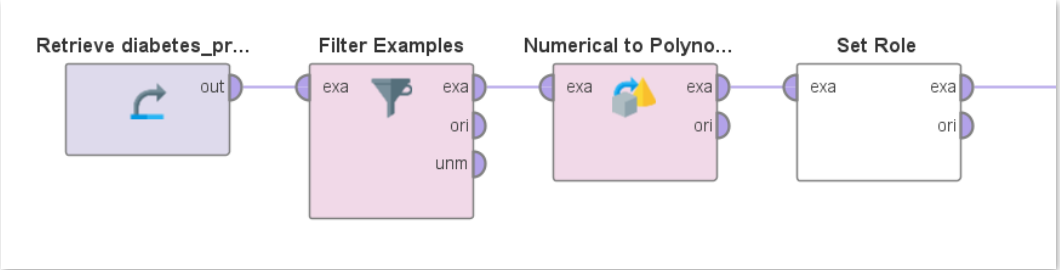


**Figure 1.** Pre-Processing Stages

The pre-processing stages in this study are as follows:

a. Filter Example

   The researcher utilizes this operator to determine the lower threshold of blood glucose levels for patients with diabetes. This threshold value functions to classify patients who suffer from diabetes or not. This lower limit is employed because there is a minimum blood sugar concentration in patients at risk of developing diabetes. When the blood glucose level reaches 100 mg/dL, the patient is considered to be in the prediabetes phase (Khan et al., 2019). The lower threshold the researcher uses for the blood glucose level is above 90 mg/dL as a warning for patients with blood glucose levels approaching 100 mg/dL. Additionally, the lower threshold is applied to the Body Mass Index (BMI) variable with a value greater than or equal to 26, as this falls within the average range for both men and women before entering the obesity condition, which can increase the risk of developing diabetes (Beulens et al., 2019). In addition, the researcher also filters the smoking history variable using the terms current, ever, and former. This is because a smoking history can influence the likelihood of disease complications after a patient is diagnosed with diabetes (Huh et al., 2022).

b. Numerical to Polynomial

   The Numerical to Polynomial operator is employed to convert numeric attributes into the desired polynomial attributes. This transformation aids in implementing algorithms that require non-numeric labels. Specifically, the gain ratio criteria selected for Random Forest and Decision Tree algorithms cannot be applied directly to numeric labels. Therefore, the researcher utilizes this operator to transform the numeric attribute related to diabetes prediction into a polynomial form. Doing so makes each numerical value a nominal value for the new attribute, facilitating the algorithm's implementation.

c. Set Role

   The Set Role operator assigns roles to attributes within a dataset, such as designating an attribute as a label or target. In this context, the researcher employs the Set Role operator to specify that the diabetes attribute serves as the label for the model.

## 3. Data Processing

This stage involves applying classification algorithms to the pre-processed dataset. In this context, data processing includes utilizing machine learning algorithms, specifically Decision Tree and Random Forest, to obtain relevant analysis results aligned with the research objective—such as predicting diabetes in patients. The Random Forest algorithm is a powerful tree-based ensemble model that can be used for regression and classification tasks (Dumitrescu et al., 2022). This model can achieve accurate predictions by randomly combining several decision trees from training data and performing feature selection (Qorib et al., 2023). Random Forest creates a 'Forest' from a collection of decision trees trained using the 'bagging' method (Govindan & Balakrishnan, 2022), combining several learning models to enhance overall capacity (Mohamed et al., 2023). As an illustration, Figure 2 depicts the concept of a Random Forest, as shown in Figure 2.

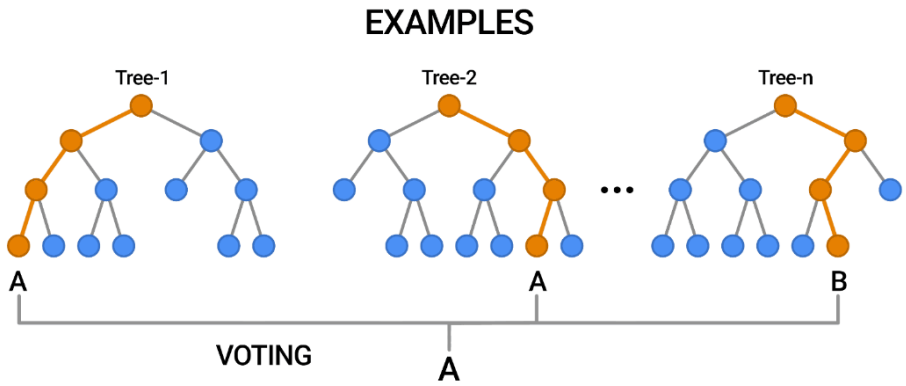


**Figure 2.** Random Forest Concept
Source: Hands-on Random Forest with Python - medium.com

A Decision Tree is an algorithm an algorithm that implements a classification rule as a decision tree for a specific dataset. This algorithm requires loading all data into memory and has the advantage of handling missing attributes (Maji & Arora, 2019). The Decision Tree algorithm undergoes an initiation and termination process that begins at the root node and ends at the leaf nodes, also known as terminal nodes (Tangirala, 2020). In a decision tree, internal nodes are located between the root node and the leaf nodes, which are used to test the characteristics of data points (Aldahiri et al., 2021). The concept of the Decision Tree algorithm is explained in the following Figure 3.

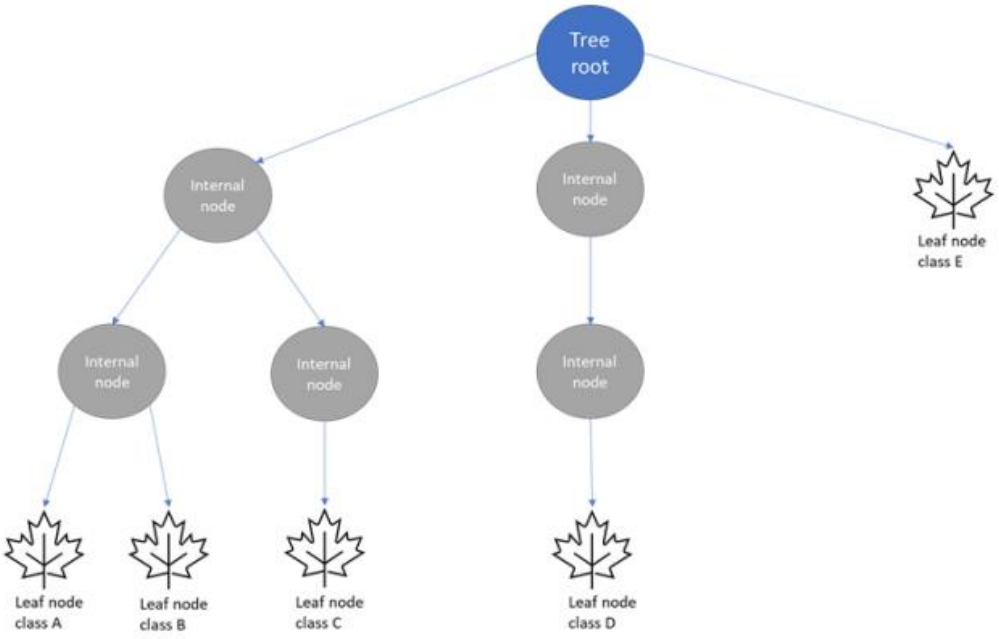


**Figure 3.** Decision Tree Concept. Source: Trends in Using IoT with Machine Learning in Health Prediction Systems

## 4. Evaluation

This stage consists of two processes as follows:

a. Model Training

This research employs k-fold cross-validation during model training to evaluate the model's performance by repeatedly dividing the data into training and testing subsets. This research employs 10-fold cross-validation with automatic sampling by RapidMiner10-fold cross-validation, which divides the dataset into ten equally sized parts, which is considered beneficial for enhancing decision tree performance (Malakouti et al., 2023).

b. Model Evaluation

Figure 4 and Figure 5 demonstrate the use of Decision Tree and Random Forest algorithms for measuring accuracy. These metrics assess how well the model accurately predicts outcomes based on pre-processed data. In addition to accuracy, this research evaluates precision and recall using a confusion matrix.

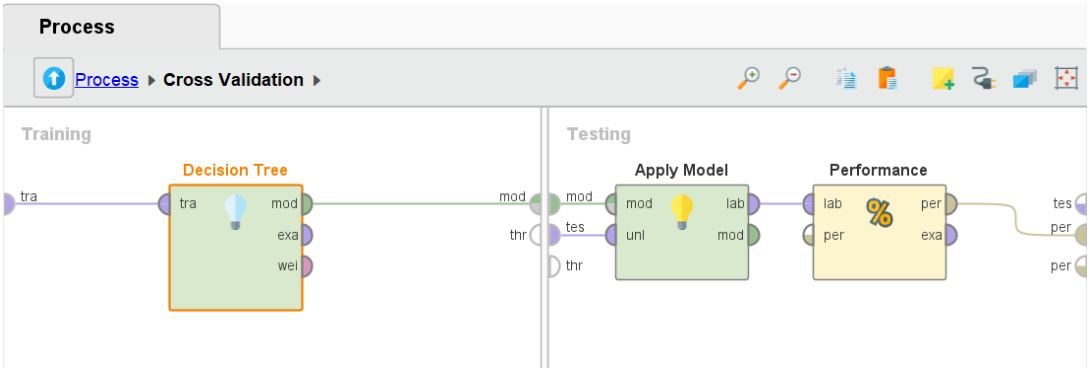


**Figure 4.** Decision Tree Model Evaluation

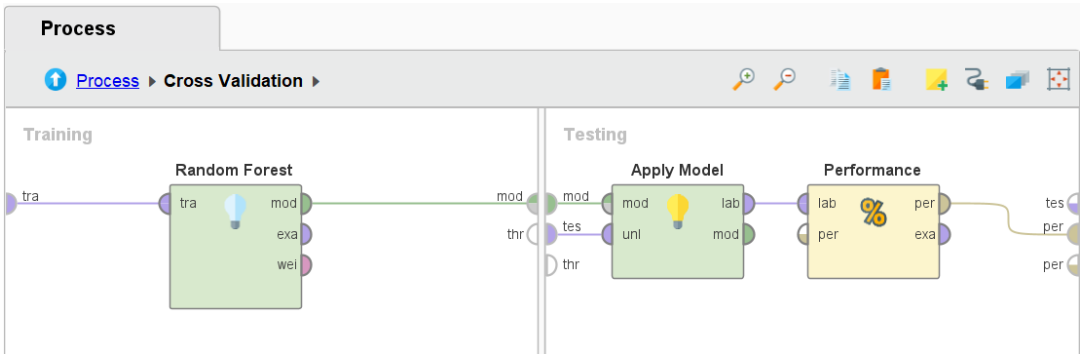


**Figure 5.** Random Forest Model Evaluation

A Confusion Matrix is a table created to describe the performance of a classification model on a dataset (Heydarian et al., 2022). During the evaluation process, the accuracy of predictions can be determined by calculating statistical measures, including True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) (Segala et al., 2023). Table 1 shows the components of the Confusion Matrix.

**Table 1.** The Confusion Matrix

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

In the Confusion Matrix, three evaluations determine the algorithm's performance when processing a dataset: accuracy, precision, and recall (Hasnain et al., 2020). Accuracy is the final value obtained by a prediction model, representing the overall performance of the dataset (Vives et al., 2024). Precision is the ratio of positive predictions to all predicted positive results. It is a crucial metric for assessing a system's ability to process data (Sun et al., 2023). Sensitivity (recall or true positive rate) is the ratio of correct positive predictions to the total number of positive instances. It quantifies how well the system recognizes positive outcomes in data that should indeed be positive (Schulte & Nissen, 2023). In this research, we consider high accuracy and high precision for detecting diabetes. The formula of three evaluation metrics refers to Equation 1, Equation 2, and Equation 3 (Mubin et al., 2023).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \text{ X } 100 \quad (1)$$

$$Presision = \frac{TP}{TP+FP} X\ 100 \quad (2)$$

$$Recall = \frac{TP}{TP+FN}\ X\ 100 \quad (3)$$

## C. RESULT AND DISCUSSION

### 1. Data Pre-Processing Result

a. Data Filtering Result

The first stage in the data pre-processing phase is data filtering. The filter example operator is employed to prioritize data with higher urgency. It involves filtering specific variables associated with an increased risk of diabetes in patients. These variables include blood glucose level (Khan et al., 2019) filtered for values above 90 mg/dL, BMI (Beulens et al., 2019) filtered for values greater than or equal to 20, and smoking history (Huh et al., 2022) including current, ever, and former smokers. As a result, the research dataset is reduced from 100,000 to 14,204 data points. The distribution of the target variable (diabetes class) after filtering is 2,557 data (18% data) for the positive diabetes class and 11,647 data (82% data) for the negative diabetes class.

b. Numerical to Polynomial Result

This study converts the diabetes class from numerical to polynomial. This step is necessary because if the label used is still in numeric form, the gain ratio criteria in the Random Forest and Decision Tree algorithms cannot be executed. After applying the "Numerical to Polynomial" operator, the dataset can proceed to the Set Role operator.

c. Set Role Result

The Set Role function serves as the determinant for the label used in research. In this study, the label chosen is diabetes. Once the label is specified using the Set Role parameter, the diabetes variable is automatically employed as the research label. Figure 6 illustrates the result of the Set Role operator.

| Row No. | diabetes | gender | age |
|---|---|---|---|
| 1 | 0 | Female | 36 |
| 2 | 0 | Male | 76 |
| 3 | 0 | Female | 54 |
| 4 | 0 | Female | 78 |
| 5 | 0 | Male | 37 |
| 6 | 0 | Female | 72 |
| 7 | 0 | Female | 41 |
| 8 | 1 | Male | 50 |
| 9 | 1 | Male | 73 |
| 10 | 1 | Female | 53 |

**Figure 6.** The result of Set Role Operator

## 2. Data Processing and Model Evaluation Result

After performing data pre-processing, model evaluation is conducted to measure the performance of the algorithms used in a prediction system. The algorithms employed are Random Forest and Decision Tree, utilizing 10-fold cross-validation. This technique aids in dividing the dataset into ten equally sized portions, which is beneficial for enhancing the algorithm's performance (Malakouti et al., 2023).

Table 2 shows the performance of the Random Forest algorithm, which is illustrated in the confusion matrix table. Based on Table 2, the random forest algorithm generated 1,700 True Positives (TP), representing the examples correctly predicted as positive by the model. False Negative (FN): 2 cases incorrectly predicted as negative when actually positive. False Positive (FP) of 854 Cases that were incorrectly predicted as positive when they were actually negative. True Negative (TN) has as many as 11,648 cases, which are examples that were correctly predicted as negative.

**Table 2.** The Confusion Matrix Result of Random Forest

| | **Actual Positive** | **Actual Negative** |
|---|---|---|
| Predicted Positive | 1,700 | 854 |
| Predicted Negative | 2 | 11,648 |

Furthermore, Table 3 represents the performance of the Decision Tree algorithm illustrated in the confusion matrix table. Based on Table 3, the Decision Tree algorithm generated 1,708 True Positives (TP), 22 False Negative (FN), 846 False Positive (FP), and 11,628 True Negative (TN).

**Table 3.** The Confusion Matrix Result of Decision Tree

| | **Actual Positive** | **Actual Negative** |
|---|---|---|
| Predicted Positive | 1,708 | 846 |
| Predicted Negative | 22 | 11,628 |

After both algorithms (Random Forest and Decision Tree) calculate the number of TP, FN, FP, and TN, both algorithms' accuracy, precision, and recall are automatically calculated based on Equation 1, Equation 2, and Equation 3. Table 4 compares the two algorithms' performance in predicting diabetes diseases.

**Table 4.** The Comparison of two algorithms' performance in predicting diabetes

| **Algorithm** | **Accuracy** | **Precision** | **Recall** |
|---|---|---|---|
| Random Forest | 93.97% | 99.88% | 66.56% |
| Decision Tree | 93.89% | 98.73% | 66.88% |

Based on Table 4, the Random Forest algorithm's accuracy and precision values are greater than the Decision Tree algorithm's accuracy and precision values. Compared with the Decision Tree algorithm, the accuracy and precision of the Random Forest algorithm are 0.08 and 1.15 higher, respectively. This means that the Random Forest algorithm is superior in correctly predicting whether someone will have diabetes. However, the Random Forest algorithm's recall value is lower than the Decision Tree algorithm. The random forest algorithm produces a smaller recall value of 0.32 than the decision tree algorithm.

In addition to evaluating the accuracy, precision, and recall, this study also performed the computing time analysis of the Random Forest and Decision Tree algorithms. Table 5 shows the computing time for these two algorithms. Based on Table 5, the Random Forest algorithm requires a longer computing time of 15 seconds for the Decision Tree algorithm. This result occurred because the Random Forest Algorithm has a more complex algorithm than the Decision Tree algorithm. Based on this result, the Decision Tree algorithm provides the benefit of more efficiently using computing time, which is essential for diabetes detection, because diabetes affects a person's life.

**Table 5.** The Comparison of two algorithms' computing time in predicting diabetes

| **Algorithm** | **Computing Time** |
|---|---|
| Random Forest | 16s |
| Decision Tree | <1s |

## D. CONCLUSION

This study examines predictive data mining models, along with the provided operators, using the Random Forest and Decision Tree algorithms within the RapidMiner Studio software. The evaluation results indicate that the accuracy and precision of the Random Forest algorithm are greater than that of the Decision Tree algorithm. However, the Decision Tree algorithm demonstrates greater recall than the Random Forest. In addition, the Decision Tree algorithm also results in computational efficiency, especially in the context of health prediction with high urgency. Utilizing variables from the Diabetes Prediction Dataset (BMI, blood glucose, blood pressure, and HbA1c) proves more effective in enhancing accuracy. Based on the results, the Decision Tree is more effective, because it can result in high accuracy and highly computation speed. Although this research has produced high accuracy and computational speed, there is still room to explore similar research by considering the influence of data imbalance.

## REFERENCES

Aldahiri, A., Alrashed, B., & Hussain, W. (2021). Trends in Using IoT with Machine Learning in Health Prediction System. In *Forecasting* (Vol. 3, Issue 1, pp. 181–206). MDPI. https://doi.org/10.3390/forecast3010012

Beulens, J. W. J., Rutters, F., Rydén, L., Schnell, O., Mellbin, L., Hart, H. E., & Vos, R. C. (2019). Risk and management of pre-diabetes. *European Journal of Preventive Cardiology*, *26*(2_suppl) 47–54. https://doi.org/10.1177/2047487319880041

Dumitrescu, E., Hué, S., Hurlin, C., & Tokpavi, S. (2022). Machine learning for credit scoring: Improving logistic regression with non-linear decision-tree effects. *European Journal of Operational Research*, *297*(3), pp. 1–15. https://doi.org/10.1016/j.ejor.2021.06.053

Durugkar, S. R., Raja, R., Nagwanshi, K. K., & Kumar, S. (2022). Introduction to data mining. In *Data Mining and Machine Learning Applications*, vol. 4, 2022, pp. 1–19. https://doi.org/10.1002/9781119792529.ch1

Garg, M. (2023). Random Logistic Vector Analysis Based Opinion Mining For Identifying Best Product Using User Reviews in Ecommerce Applications. *2nd IEEE International Conference on Distributed Computing and Electrical Circuits and Electronics, ICDCECE 2023*. Ballar, India, 2023, pp. 1–6, https://doi.org/10.1109/ICDCECE57866.2023.10150493

Govindan, V., & Balakrishnan, V. (2022). A machine learning approach in analysing the effect of hyperboles using negative sentiment tweets for sarcasm detection. *Journal of King Saud University - Computer and Information Sciences*, *34*(8), pp. 5110–5120. https://doi.org/10.1016/j.jksuci.2022.01.008

Hasnain, M., Pasha, M. F., Ghani, I., Imran, M., Alzahrani, M. Y., & Budiarto, R. (2020). Evaluating Trust Prediction and Confusion Matrix Measures for Web Services Ranking. *IEEE Access*, vol. 8, pp. 90847–90861, 2020 . https://doi.org/10.1109/ACCESS.2020.2994222

Heydarian, M., Doyle, T. E., & Samavi, R. (2022). MLCM: Multi-Label Confusion Matrix. *IEEE Access*, vol. 10, pp. 19083–19095, 2022. https://doi.org/10.1109/ACCESS.2022.3151048

Huh, Y., Han, K., Choi, M. J., Kim, J. H., Kim, S. M., & Nam, G. E. (2022). Association of Smoking Status With the Risk of Type 2 Diabetes Among Young Adults: A Nationwide Cohort Study in South Korea. *Nicotine and Tobacco Research*, *24*(8), pp. 1234–1240. https://doi.org/10.1093/ntr/ntac044

Ismail, Mohmand, M. I., Hussain, H., Khan, A. A., Ullah, U., Zakarya, M., Ahmed, A., Raza, M., Rahman, I. U., & Haleem, M. (2022). A Machine Learning-Based Classification and Prediction Technique for DDoS Attacks. *IEEE Access*, vol. 10, no. 12, pp 21443–21454. https://doi.org/10.1109/ACCESS.2022.3152577

Khan, R. M. M., Chua, Z. J. Y., Tan, J. C., Yang, Y., Liao, Z., & Zhao, Y. (2019). From pre-diabetes to diabetes: Diagnosis, treatments and translational research. In *Medicina (Lithuania)* (Vol. 55, Issue 9, pp. 1–30). https://doi.org/10.3390/medicina55090546

Kumari, S., Vani, V., Malik, S., Tyagi, A. K., & Reddy, S. (2021). Analysis of Text Mining Tools in Disease Prediction. In Hybrid Intelligent Systems:20th International Conference on Hybrid Intelligent Systems (HIS 2020), December 14–16, 2020, 2021, pp. 546–564. Springer International Publishing. https://doi.org/10.1007/978-3-030-73050-5_55

Maji, S., & Arora, S. (2019). Decision Tree Algorithms for Prediction of Heart Disease. In *Lecture Notes in Networks and Systems* (Vol. 40, pp. 447–454). Springer. https://doi.org/10.1007/978-981-13-0586-3_45

Malakouti, S. M., Menhaj, M. B., & Suratgar, A. A. (2023). The usage of 10-fold cross-validation and grid search to enhance ML methods performance in solar farm power generation prediction. *Cleaner Engineering and Technology*, vol. 15, Art. no. 100664, 2023. Pp. 1-7, https://doi.org/10.1016/j.clet.2023.100664

Maulana, A., Faisal, F.A., Noviandy, T.R., Rizkia, T., Idroes, G.M., Tallei, T.E., El-Shazly, M., & Idroe, R. (2023). Machine Learning Approach for Diabetes Detection Using Fine-Tuned XGBoost Algorithm. *Infolitika Journal of Data Science*, vol. 1 (1), pp. 1-7, https://doi.org/10.60084/ijds.v1i1.72

Mehraeen, E., Pashaei, Z., Akhtaran, F. K., Dashti, M., Afzalian, A., Ghasemzadeh, A., Asili, P., Kahrizi, M. S., Mirahmad, M., Rahimi, E., Matini, P., Afsahi, A. M., Dadras, O., & Seyed Alinaghi, S. A. (2023). Estimating Methods of the Undetected Infections in the COVID-19 Outbreak: A Systematic Review. In *Infectious Disorders - Drug Targets* (Vol. 23, Issue 4, pp 1–20). https://doi.org/10.2174/1871526523666230124162103

Mohamed, E. S., Naqishbandi, T. A., Bukhari, S. A. C., Rauf, I., Sawrikar, V., & Hussain, A. (2023). A hybrid mental health prediction model using Support Vector Machine, Multilayer Perceptron, and Random Forest algorithms. *Healthcare Analytics*, Vol. 3, 100185, pp. 1–20. https://doi.org/10.1016/j.health.2023.100185

Moreno-Lumbreras, D., Gonzalez-Barahona, J. M., & Robles, G. (2023). BabiaXR: Facilitating experiments about XR data visualization. *SoftwareX*, Volume 24, 2023, 101587, pp. 1-8, ISSN 2352-7110. https://doi.org/10.1016/j.softx.2023.101587

Mubin, M. N., Kusuma, H., & Rivai, M. (2023). Perspective Transformation Automation In Identification Of Parking Lot Status With Blob Detection. *JAREE (Journal on Advanced Research in Electrical Engineering)*, *7*(2), pp. 84–91. https://doi.org/10.12962/jaree.v7i2.364

Pyayt, A., Khan, R., Brzozowski, R., Eswara, P., & Gubanov, M. (2020). Rapid Antibiotic Susceptibility Analysis Using Microscopy and Machine Learning. *Proceedings - 2020 IEEE International Conference on Big Data, Big Data, 2020,* pp. 5804-5806. https://doi.org/10.1109/BigData50022.2020.9378005

Qorib, M., Oladunni, T., Denis, M., Ososanya, E., & Cotae, P. (2023). Covid-19 vaccine hesitancy: Text mining, sentiment analysis and machine learning on COVID-19 vaccination Twitter dataset. *Expert Systems with Applications*, vol. 212, 118715, pp. 1-14, 2023. https://doi.org/10.1016/j.eswa.2022.118715

Schulte, J., & Nissen, V. (2023). Sensitivity analysis of combinatorial optimization problems using evolutionary bilevel optimization and data mining. *Annals of Mathematics and Artificial Intelligence*, *91*(2–3), pp. 309–328. https://doi.org/10.1007/s10472-022-09827-w

Segala, F. V., Papagni, R., Cotugno, S., De Vita, E., Susini, M. C., Filippi, V., Tulone, O., Facci, E., Lattanzio, R., Marotta, C., Manenti, F., Bavaro, D. F., De Iaco, G., Putoto, G., Veronese, N., Barbagallo, M., Saracino, A., & Di Gennaro, F. (2023). Stool Xpert MTB/RIF as a possible diagnostic alternative to sputum

in Africa: a systematic review and meta-analysis. In *Frontiers in Public Health*, Vol. 11:1117709, pp. 1–9. https://doi.org/10.3389/fpubh.2023.1117709

Sękowski, K., Grudziąż-Sękowska, J., Pinkas, J., & Jankowski, M. (2022). Public knowledge and awareness of diabetes mellitus, its risk factors, complications, and prevention methods among adults in Poland—A 2022 nationwide cross-sectional survey. *Frontiers in Public Health*, 10: 1029358. pp. 1-28, https://doi.org/10.3389/fpubh.2022.1029358

Singh, B., & Jaiswal, R. (2022). Automation of prediction system for temporal data. *International Journal of Information Technology (Singapore)*, *14*(6), pp. 3165–3174. https://doi.org/10.1007/s41870-022-01065-x

Stoleru, G. I., & Iftene, A. (2022). Prediction of Medical Conditions Using Machine Learning Approaches: Alzheimer's Case Study. *Mathematics*, *10*(10), 1767, pp. 1–20. https://doi.org/10.3390/math10101767

Sun, D., Luo, R., Guo, Q., Xie, J., Liu, H., Lyu, S., Xue, X., Li, Z., & Song, S. (2023). A University Student Performance Prediction Model and Experiment Based on Multi-Feature Fusion and Attention Mechanism. *IEEE Access*, vol. 11, pp. 112307–112319, 2023. https://doi.org/10.1109/ACCESS.2023.3323365

Tangirala, S. (2020). Evaluating the impact of GINI index and information gain on classification using decision tree classifier algorithm. *International Journal of Advanced Computer Science and Applications*, 11(2), pp. 612–619. https://doi.org/10.14569/ijacsa.2020.0110277

Vives, L., Cabezas, I., Vives, J. C., Reyes, N. G., Aquino, J., Condor, J. B., & Altamirano, S. F. S. (2024). Prediction of Students' Academic Performance in the Programming Fundamentals Course Using Long Short-Term Memory Neural Networks. *IEEE Access*, vol. 12, pp. 5882–5898, 2024. https://doi.org/10.1109/ACCESS.2024.3350169

Yilmaz, R., & Yağin, F. H. (2022). Early Detection of Coronary Heart Disease Based on Machine Learning Methods. *Medical Records*, *4*(1), pp. 1–6. https://doi.org/10.37990/medr.1011924