JIKO (Journal of Informatics and Computers) Accredited KEMENDIKBUD RISTEK, No.105/E/KPT/2022
Vol. 7, No. 2, August 2024, pp. 79-87 p-ISSN: 2614-8897
DOI: 10.33387/jiko.v7i2.8316 e-ISSN: 2656-1948

COMPARISON OF DECISION TREE AND RANDOM FOREST METHODS IN THE CLASSIFICATION OF DIABETES MELLITUS

Nofa Auliyatul Maulidiyyah[*1], Trimono[2] , Aviolla Terza Damaliana[3], Dwi Arman Prasetya[4]

[1,2,3,4]Universitas Pembangunan Nasional "Veteran" Jawa Timur

[*]Email: [1]20083010029@student.upnjatim.ac.id, [2]trimono.stat@upnjatim.ac.id,
[3]aviolla.terza.sada@upnjatim.ac.id, [4]arman.prasetya.sada@upnjatim.ac.id

(Received: 01 July 2024, Revised: 15 July 2024, Accepted: 26 July 2024)

Abstract

Diabetes mellitus is a deadly disease caused by the failure of the pancreas to produce enough insulin. Indonesia ranks fifth in the world with the number of people with diabetes in 2021 at around 19.47 million, and this number continues to increase. One of the main challenges in diabetes management is to make the right classification between type 1 and type 2 diabetes, as misdiagnosis can result in inappropriate treatment and worsen the patient's condition. This study uses a machine learning approach to compare Decision Tree and Random Forest methods in classifying type 1 and type 2 diabetes mellitus. The goal is to identify the most effective model in predicting the type of diabetes based on medical record data. The comparison was done using k-fold cross validation and confusion matrix. The results showed that Random Forest provided an average accuracy of 94%, while Decision Tree reached 93% during cross validation testing. Although both models were able to perform well in classification, Random Forest showed a more stable performance and a slight edge in accuracy over Decision Tree. Evaluation with the confusion matrix showed that the Decision Tree model achieved 93% accuracy compared to Random Forest's 91%. In addition, the Decision Tree model also had a lower number of prediction errors, 7, compared to 9 for Random Forest. The most influential variables in classification also differed between the two models, showing the unique advantages and characteristics of each approach.

Keywords: diabetes mellitus, decision tree, random forest, k-fold cross validation, confusion matrix

**Corresponding Author: Nofa Auliyatul Maulidiyyah*

1. INTRODUCTION

Diabetes mellitus is a chronic metabolic disorder when the patient's body does not produce enough insulin or the patient's body cannot utilize insulin properly, causing excessive blood sugar in the body, which often causes complications in the organs of the body [1]. Types of diabetes mellitus are classified into type 1, type 2, and gestational based on increased blood sugar levels caused by autoimmune and lifestyle factors. However, type 1 and type 2 are the most common [2].

In Indonesia, there were 19.47 million diabetic patients in 2021, showing an annual increase in the number of diabetics. Data collected by the International Diabetes Federation (IDF) shows that the number of diabetics in Indonesia is expected to increase to 23.32 by 2030 [3]. Recent research shows that 40% of cases of type 1 diabetes in adults over 30 are often misdiagnosed as type 2 diabetes. One of the main concerns is the assumption that adults are more likely to have type 2 diabetes, and some adults with type 1 diabetes may not need insulin at diagnosis, leading to symptoms similar to type 2 diabetes. This misdiagnosis results in improper treatment, negatively impacting the patient's quality of life and survival [4].

Prevention of diabetes mellitus is necessary with medical personnel and mathematical quantitative models, to prevent DM with early detection and promote a healthy lifestyle [5]. The purpose of this model approach is to use machine learning. Machine learning is a type of Artificial Intelligence (AI) that uses data and algorithms to mimic the way humans learn. The goal is to improve the accuracy and precision of predictions. One of the common techniques used in machine learning is classification.

Classification is the process of creating a function or model that describes a class on a data or concept to predict the class of an object whose label has not yet been obtained [1]. In this study, classification techniques were used to predict patients suffering from type 1 and type 2 diabetes. The classification process can be done with several algorithms, namely Decision Tree and Random Forest. Decision Tree is a prediction model for a decision using a tree structure or hierarchy and looking for a solution to a problem by using criteria as interconnected nodes to form a tree structure. Each tree has branches, these branches represent all the traits that must be met to grow to the next branch until those branches end up in the leaves [6]. Random Forest is a development of the *Classification and Regression Trees* (CART) that uses bootstrap techniques *aggregating (bagging)* and random selection of features to form a set of decision trees in classifying data [7]. This method is effective because it can overcome *overfitting* and works fast on large datasets, but tends to *overfitting* on unbalanced datasets [8].

Research conducted by [9] shows that the Decision Tree method in classifying best-selling products (private data) obtained a result of 90% and an AUC value of 0.709, this value is included in Good Classificataion. Research conducted by [10] indicates that *Random Forest* and Decision Tree in the classification of interphylity diseases obtained Random Forest results are a superior method of 1.3% when compared to Decision Tree C4.5, which is 87.20% with 85.90% in predicting accuracy in Fertility Dataset. After that, the research conducted by [11] in the classification of type 2 diabetes mellitus using the Logistic Regression, Decision Tree, and Random Forest showed the highest accuracy of 97.1%. research conducted by [12] in determining student achievement using the K-Means algorithm to group students into clusters based on certain characteristics, and implementing Decision Tree to classify student learning achievement. The test results showed a prediction accuracy of 71%.

This study compares Decision Tree and Random Forest in classifying diabetes mellitus. The purpose of this study was to find out to what extent Decision Tree and Random Forest can improve the accuracy and reliability of classification models when aiding in the diagnosis and management of diabetes mellitus.

2. RESEARCH METHODS

2.1 Decision Tree

A decision tree is a type of tree structure in which the test attribute is represented by each node, the test result is represented by each branch, and a particular class group is represented by each leaf node. A decision tree's root, which is often the property having the biggest impact on a certain class, is the highest level of nodes. Decision trees often employ a top-down search approach to locate answers. A new class is established in accordance with the findings of the attribute value testing procedure, which involves tracing the path from the root node to the leaf node [9]. A sample of data whose class is unknown is classified into existing classes using a decision tree.

The following are the steps involved in creating a decision tree [1]:

1. Selecting a root attribute
2. Make a branch for every attribute value.
3. Assigning cases to every branch
4. Choosing the attribute with the greatest gain value among all currently available attributes, repeat this procedure on each branch until all cases in each branch have the same class to designate the attribute as the root. The following formula is used to get this gain value:

$$Gain\ (S, A) = Entropi\ (S) - \sum_{i=1}^{n} - \left|\frac{S_i}{S}\right| \times Entropi\ S_i \quad (1)$$

Information:
S = declares the case set
A = declare the
n = expresses the number of partitions of attribute A
$|S_i|$ = states the number of cases in the ith partition
$|S|$ = states the number of cases in s

Meanwhile, to generate the Entropy value with the following formula:

$$Entropy(S) = \sum_{i=1}^{k} -P_I\ \log_2 P_i \quad (2)$$

Information:
S = declares the case set
k = expresses the number of partitions S
P_i = expresses the probability obtained from the total number of samples.

The following are some common characteristics of Decision Trees [13]:

1. Decision Tree is a nonparametric approach to building a classification model
2. The technique used to build the Decision Tree allows for the rapid creation of models from large training sets
3. Decision Trees with small tree sizes are relatively easy to interpret
4. Decision Tree provides an expressive overview of learning discrete value functions
5. Decision Tree is quite resistant to noise, especially for methods that can handle overfitting issues

2.2 Random Forest

Breiman introduced Random Forest in 2001. Random Forest has two main goals in solving problems, namely to perform classification and regression using multiple decision trees [14]. Random Forest is an ensemble method that improves

classification accuracy by combining multiple classification methods. A Random Forest consists of a set of Decision Trees, and the more decision trees used, the more powerful the Random Forest algorithm becomes. In each decision tree, the data starts from the root and moves to the leaf to determine the class or prediction value of the data. [15]. The number of decision trees in a Random Forest affects the accuracy of the overall random structure.

Random Forest works by building multiple decision trees and generating predictions in the form of class modes or averages from individual decision tree outcomes. The Random Forest concept combines multiple random decision trees into a model because the number of decision trees affects the accuracy and stability of the model as a whole. [16]. [17] said in the decision tree, the process begins by calculating the entropy to assess the level of impurity of the attributes, and then the information gain value is used to determine which attributes will be used to perform the data separation. The calculation of entropy refers to the formula as defined in equation (5), while the information gain is calculated using equation (6).

$$\text{Entropy (Y)} = -\sum_{i=1}^{n} p_i \log_2 (p_i) \quad (3)$$

Equation (5) is a set of cases, and is the proportion of the number of class samples to the total number of samples.Y p_ii

$$\text{information gain (Y, a)} = \text{Entropy (Y)} - \sum_{i=1}^{n} \frac{|Y_i|}{|Y|} \text{ entropy (Yi)} \quad (4)$$

Formula (6) The number of divisions that the attribute
a produces is n, and an is the attribute that is being assessed. Whereas |Y| represents the overall number of instances in Y, $|Y_i|$ represents the number of cases in partition i. The following is how the Random Forest algorithm works [10]:

1. Indicates how many trees (k) were chosen out of a total of m features, where k is smaller than m.
2. For every tree in the dataset, N random samples are selected.
3. A subset of predictors, m < p, where p is the number of predictor variables, is randomly selected from each tree.
4. Until there are as many trees as possible, the procedures from the second and third phases are repeated.
5. The prediction results come from the categorization results with the highest number of votes, up to a maximum of k trees.

The advantages of the Random Forest Algorithm are that it can produce relatively low errors, has good classification performance, and is suitable for big data.

2.3 K-Fold Cross Validation

The k-fold cross validation method, which generalizes, divides the data into equal-sized sections. One partition is chosen for the test data during the procedure, while the remaining partitions are utilized for the training data. To ensure that every partition is utilized for the test data precisely once, this procedure is done several times. The size of the data set is set to k = N via the k-fold cross validation procedure. The benefit of this strategy is that it uses the most data feasible for training. All of the data set is essentially covered by the test set. This approach's main downside is the amount of computing required to repeat the process N times. K-fold cross-validation is one method for assessing a model's correctness [18]. The steps of k fold cross validation, namely:

1. There are k parts created from the overall data.
2. The first fold occurs when the first portion is converted to testing data and the remaining portion to training data. Next, using the section of the data, determine how close, how accurate, or how comparable a measurement result is to the real number or data. The accuracy is calculated using the following formula.

$$akurasi = \frac{\sum data\ uji\ benar\ klasifikasi}{total\ data\ uji} \times 100 \quad (5)$$

3. The second fold occurs when the remaining portion becomes training data and the second half becomes testing data. Next, compute the accuracy using the relevant data segment.
4. Until it reaches the K fold, and so on. Determine the k accuracy fruit's average accuracy. The ultimate accuracy is determined by this average accuracy.

2.4 Confusion Matrix

The confusion matrix table displays four different combinations of predicted and actual values. Each row of the matrix shows the actual data classification and the prediction classification, or vice versa [19].

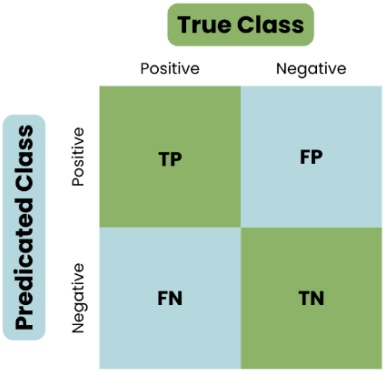

Figure 1. Confusion Matrix

The confusion matrix has four terms as a result of classification. True Positive (TP) is the amount of correctly classified positive data, True Negative (TN) is the amount of correctly classified negative data; False Positive (FP) is negative data but is considered positive data; and False Negative (FN) is positive data but is considered negative. There are five indicators: accuracy, precision, recall, and F1-Score.

$$accuracy = \frac{TP+TN}{TP+FP+FN+TN} \quad (6)$$

$$recall = \frac{TP}{TP+FN} \quad (7)$$

$$Precision = \frac{TP}{TP+FP} \quad (8)$$

$$f1 - score = 2.\frac{1}{\frac{1}{recall}+\frac{1}{precision}} \quad (9)$$

2.5 Data Description and Research Steps

The data used in this study came from patients suffering from diabetes mellitus at the Bungah Health Center, Bungah District, Gresik Regency. The data consisted of 513 with 10 variables. From this data, there are 341 patients with type 2 diabetes and 172 patients with type 1 diabetes.

Table 1. Research Data

Variable	Data	Information
X1 (Age)	1 : <20 Years 2 : 20-40 Years 3: >40 years	Information about the patient's age
X2 (GDA)	Numerical	Random blood sugar
X3 (GDP)	Numerical	Patients' blood sugar levels during fasting
X4 (Blood Sugar 2 JamPP)	Numerical	Post-prandial or post-meal blood sugar levels
X5 (Hba1C)	Numerical	Blood HbA1c levels (a long-term indicator for blood sugar control)
X6 (BMI)	Numerical	Patient's body mass index
X7 (Physical Activity)	0 : Light 1 : Weight	The patient's level of physical activity or lifestyle
X8 Systolic	Numerical	Blood pressure when the heart pumps blood
X9 Diatable	Numerical	Blood pressure at heart relaxation
Y	1 : Type 1 diabetes 2 : Type 2 diabetes	Diagnosis results based on doctors

Figure 2 shows the flow of the research methodology used to achieve the research analysis objectives. The process begins with data collection, *pre-processing*, the use of SMOTE to handle unbalanced data, and then classification using *Random Forest*.

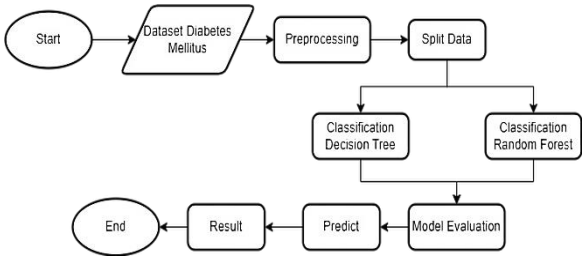

Figure 2. Research Flowchart

1. Dataset Diabetes Mellitus
   This study uses secondary data from the Bungah Health Center, Bungah District, Gresik Regency. The data included medical records of patients with diabetes mellitus. The variables analyzed included age, random blood sugar, fasting blood sugar, 2-hour PP blood sugar, HbA1C, BMI, physical activity, systolic blood pressure, and diastolic.
2. Preprocessing
   Data preprocessing includes checking blank and duplicate data, separating systolic and diastolic blood pressure, and encoding diagnoses from text to numerical.
3. Split Data
   The data is divided into two parts, namely training and testing data. In this study, training data was used for the classification process using the Random Forest method. After that, the testing data is used to evaluate the performance and final results of the model. In this study, the data was divided into 80% for training data and 20% for testing data.
4. Decision Tree and Random Forest Classification
5. Model evaluation
   At this stage, the performance of the Decision Tree and Random Forest algorithm classification models is evaluated. In this stage, it is seen based on the values of accuracy, precision, recall, and f1-score.
6. Predictions

At this stage, the trained model will generate predictions that state whether the patient falls under type 1 or type 2 diabetes.

7. Result
   At this stage, it determines how well the model predicts the status of diabetes in patients and to make decisions based on the predicted results.

3. RESULTS AND DISCUSSION

3.1 Dataset Diabetes Mellitus

This study used medical record data of diabetes mellitus patients from one of the health centers in Gresik. The data are classified into two types, namely type 1 and type 2 diabetes. A total of 513 data were collected with 9 dependent variables and 1 independent variable. The data consisted of 341 patients with type 2 diabetes mellitus and 172 patients with type 1 diabetes mellitus, as shown in Figure 3.

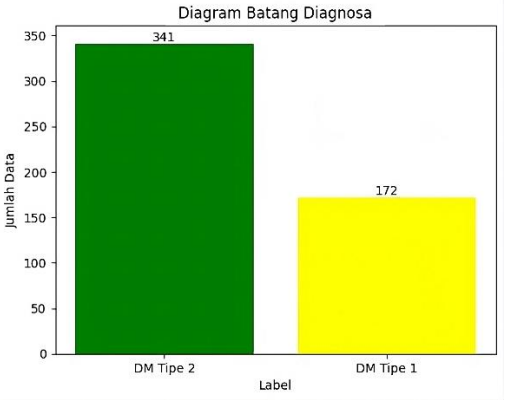

Figure 3. Data Diabetes Mellitus

3.2 Preprocessing

In this step, checking for blank, duplicate data, and the encoding process is carried out by using the LabelEncoder function from the Scikit-Learn library to convert diagnostic variables that were initially in text form into numerical, because Random Forest only accepts data in numerical form.

Table 2. Data After the Preprocessing Stage

No	Age	GDA	GDP	Blood Sugar 2 Hour PP	Systolic	Diastolic	HbA1c	BMI	Physical Activity	Diagnosa
1	64	93	102	220	175	96	8.5	33.6	1	1
2	49	108	115	180	150	70	7.0	26.6	1	0
3	56	186	126	300	174	96	9.0	23.3	1	1
4	57	255	137	250	110	60	8.0	28.1	1	1
5	42	292	148	190	104	63	7.5	43.1	1	1

After the *encoding process* is complete, in the Physical Activity column, the data of patients with physical activity (Yes) becomes 1, while the data of patients without Physical Activity (No) becomes 0. In the diagnosis column, type 1 diabetes mellitus is represented by 0, and type 2 diabetes mellitus is represented by 1.

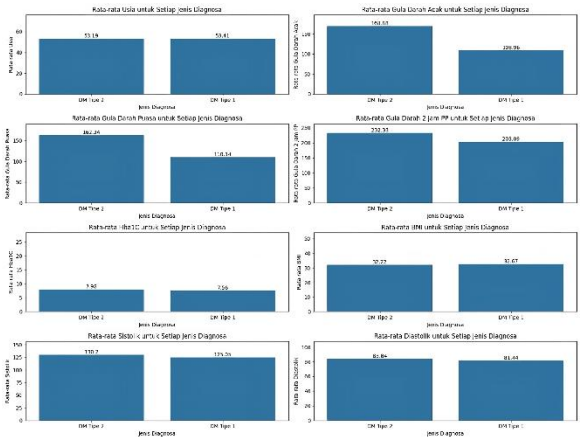

Figure 4. Feature Average Bar Chart per Diagnosis

Figure 4 shows that the average of each feature for type 2 diabetes mellitus is higher compared to type 1. In addition to bar charts, *heatmaps* of correlations between features and targets can also show the features that most influence the diagnosis. A visualization of *this heatmap* can be seen in Figure 5.

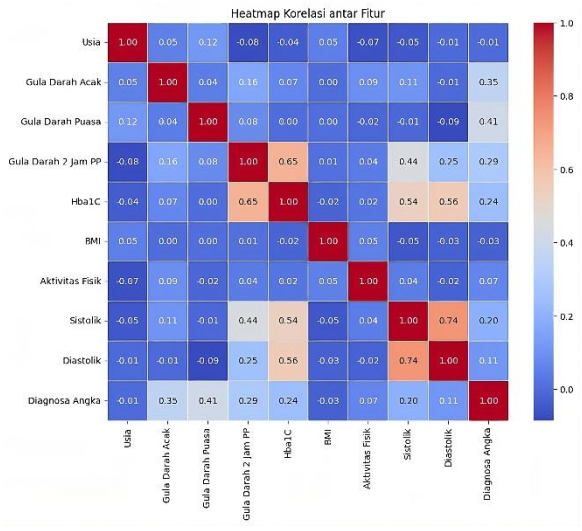

Figure 5. Correlation Between Features and Diagnosis

Figure 5 shows the correlation between features and diagnostics. The higher the correlation value, the greater the influence of this feature on the patient's diagnosis. The most influential feature can be seen in Figure 6

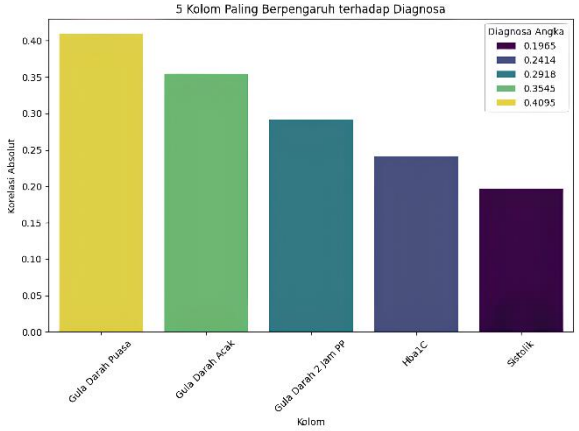

Figure 6. Features That Affect Diagnosis Most

Figure 6 shows the features that have the most influence on the patient's diagnosis, namely fasting blood sugar, random blood sugar, 2-hour PP blood sugar, Hba1C, and systolic.

3.3 Split Data

At this stage, the data is divided into *training* data and data *testing* with a ratio of 80% and 20%. Data sharing is done using the *train_test_split function* of the *Scikit-Learn library*. The results can be seen in Table 3.

Table 3. Decision Tree and Random Forest Data Sharing

Type of Diabetes	Data Decision Tree		Data Random Forest	
	Train	Test	Train	Test
Type 1	137	35	137	35
Type 2	273	68	273	68

3.4 Decision Tree Classification

At this stage, the training data is used to train the Decision Tree and Random Forest models, to compare their performance. At this stage, the decision tree model is created using the Scikit-Learn library with parameters 'max_depth' 5, 'min_samples_split' 4, and 'random_state' 42. 'max_depth' is used to control the maximum depth or complexity of a decision tree. After that 'min_samples_split' is used to set the minimum number of samples required in a node for a split to occur.

3.5 Random Forest Classification

Random Forest uses the Scikit-Learn library with a 'n_estimators' parameter of 100 and a 'random_state' of 42. 'n_estimators' is to determine the number of decision trees in the model. The more 'n_estimators', the more complex the Random Forest model becomes, improving its ability to learn complex patterns, but also extending training and prediction times. Increasing the 'n_estimators' generally improves the model's performance in terms of generalization, but should be carefully selected to avoid overfitting. random_state controls randomization when building the tree, ensuring reproducible results, which is important for sharing data and comparing models.

3.6 Model Evaluation

1. K-Fold Cross Validation

K-fold cross validation is a technique for assessing how well machine learning models work. The dataset is partitioned into K almost equal-sized portions (folds) for K-fold cross-validation. In this research, k equals 5. The K-1 part is used to train the model, while the remaining sections are used for testing. Each part is used as test data once during the K repetitions of this process. The average of the performance indicators calculated for each fold is the final outcome.

Table 4. K-Fold Cross Validation

K-n	Decision Tree	Random Forest
K-1	92%	94%
K-2	93%	96%
K-3	87%	92%
K-4	95%	94%
K-5	95%	94%
$\sum Avg$	93%	94%

The accuracy of the Decision Tree model shows greater variability compared to Random Forest, which is more consistent in its performance across each fold. Random Forest showed better performance or equivalent to a higher average accuracy of 94% to 93%.

2. Confusion Matrix

After training the model with training data, its performance was tested using data *testing* using *a confusion matrix*. The values of the *confusion matrix* are used to calculate *accuracy, precision, recall, and f1-score.* This performance calculation process uses the *Scikit-Learn library*.

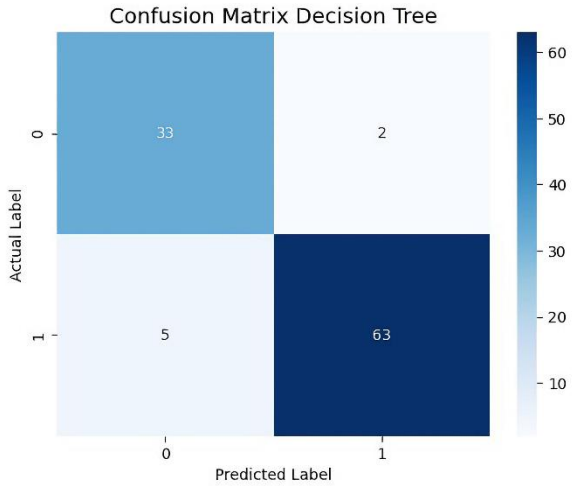

Figure 7. Confusion Matrix Decision Tree

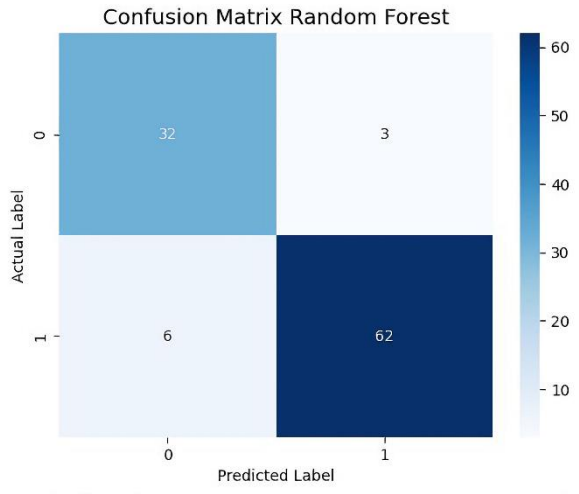

Figure 8. Confusion Matrix Random Forest

Based on the confusion matrix image above, the Decision Tree model shows better performance in predicting type 2 diabetes mellitus (1) compared to Random Forest. The Decision Tree model has a number of prediction errors for type 1 diabetes mellitus (0) as many as 2 cases of False Positive (FP) and 5 cases of False Negative (FN) for type 2 diabetes mellitus (1). Meanwhile, Random Forest has 3 cases of FP, and 6 cases of FN. Overall, Decision Trees have higher accuracy and a lower number of prediction errors. Based on the results of the confusion matrix, the following Table 4 shows the results of the performance calculations generated by the model.

Table 5. Model Performance Comparison

Metrics	Decision Tree	Random Forest
F1-Score	93%	91%
Accuracy	93%	91%
Recall	93%	91%
Precision	93%	91%

Models trained with Decision Tree showed better performance with a 2% increase in performance value compared to Random Forest. Table 6 below is an example of the data used in the test as well as the prediction results of each model.

Table 6. Model Prediction Results on Data Samples

No	Age	GDA	GDP	Blood Sugar 2 Hour PP
1	52	115	85	178
2	33	130	95	195
3	70	195	140	330
4	58	109	77	190
5	39	95	193	200

Hba1c	BMI	Physical Activity	Systolic	Diastolic
7.14	24	0	131	81
7.5	24.7	1	130	80
7	43.3	1	119	68
8.2	35.7	1	130	90
7	34.9	0	120	70

Decision Tree	Random Forest	Current
DM Type 2	DM Type 2	1
DM Type 2	DM Type 2	1
DM Type 1	DM Type 1	0
DM Type 1	DM Type 2	1
DM Type 2	DM Type 2	1

Models trained using Decision Tree and Random Forest are both good, with an accuracy of over 90%. Both models can correctly classify diabetic patients based on medical records. The results of the Decision Tree and Random Forest model training have the weight of each feature, as can be seen in the image

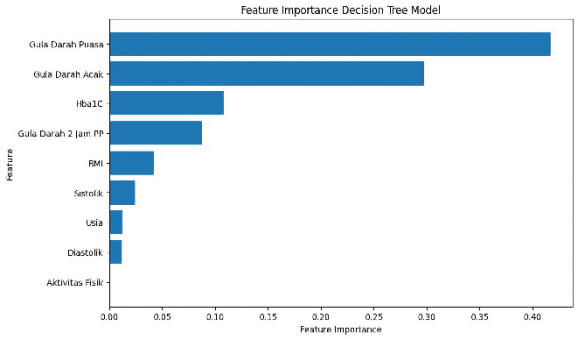

Figure 9. Feature Weight of Desicion Tree

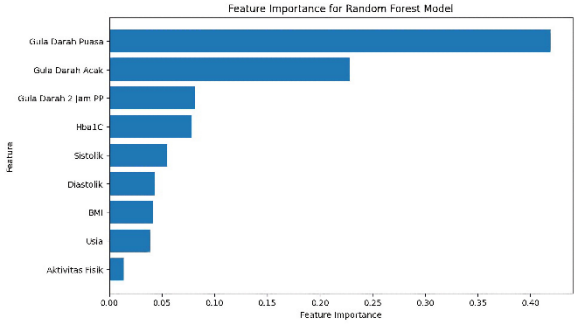

Figure 10. Random Forest Feature Weights

3.7 Discussion

Figure 11 shows the results of a performance comparison of two machine learning models, namely Decision Tree and Random Forest, using the K-fold cross-validation method. The evaluation was conducted using 5-fold cross-validation which resulted in the average accuracy of each fold for both models.

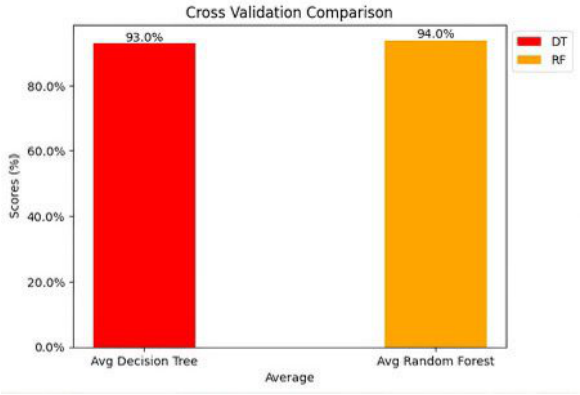

Figure 11. Perbanding Hasil 5-Fold Cross Validation

The Decision Tree model achieves an average accuracy of 93.0%. This shows that the model is quite good at classifying data with a low error rate. Meanwhile, Random Forest achieved an average accuracy of 94.0%. This shows that Random Forest has a slight advantage in terms of accuracy compared to Decision Tree. The Random Forest model is superior to the Decision Tree model in terms of accuracy and consistency of performance. Random Forest is able to reduce overfitting and generate more stable predictions.

In tests using the confusion matrix of the Decision Tree and Random Forest models, the model trained with the Decision Tree showed higher accuracy than the Random Forest model. The performance values of the Decision Tree and Random Forest models are 2% higher for each matrix, reaching 93% compared to 91%, as seen in Figure 12.

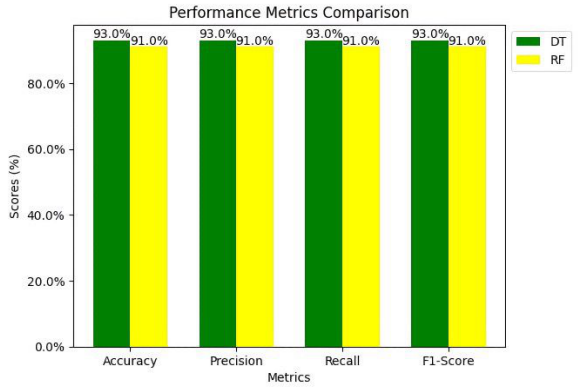

Figure 12. Comparison of Confusion Matrix Results

Data that uses the Decision Tree model results in better accuracy. The model trained with Drsicion Tree had fewer errors with 7 total predictions, 2 for type 1 diabetes mellitus and 5 for type 2 diabetes mellitus. While the Random Forest model has a total prediction error of 9, 3 for type 1 diabetes mellitus and 6 for type 2.

4. CONCLUSION

This study compares the performance of Decision Tree and Random Forest in classifying patients with diabetes mellitus based on medical record data from the Health Center in Gresik. The results show that:

1. Random Forest has an average accuracy based on cross validation of 94%, while Decision Tree reaches 93%. This shows that Random Forest is slightly superior in accuracy compared to Decision Tree. Based on the confusion matrix test, the Decision tree is superior to Random Forest, which is 93% with 91%.
2. Decision Tree showed higher accuracy in predicting type 2 diabetes mellitus, Random Forest was more consistent in its performance across each fold in K-fold cross-validation.
3. The performance of Decision Tree and Random Forest in predicting type 1 diabetes mellitus had a similar error rate, but Decision Tree was slightly better at avoiding type 2 prediction errors.
4. The factors that most affected the results of the Decision Tree classification (GDP, GDA, Hba1c, Blood Sugar 2 hours PP) were different from the results of the Random Forest classification (GDP, GDA, Blood Sugar 2 hours PP, Hba1c).

5. REFERENCE

[1] F. M. Hana, "Classification of Diabetic Patients Using the C4 Decision Tree Algorithm. 5," *J. Sist. Comput. Artificial Intelligence*, vol. 4, no. 2, 2020.

[2] R. Marzel, "Therapy in Type 1 DM," *J. Researcher. Nurse Prof.*, vol. 3, no. 1, pp. 51–62, 2021, doi: 10.37287/jppp.v3i1.297.

[3] IDF, "Diabetes report 2000 — 2045," *Diabetes Atlas*, 2021. https://diabetesatlas.org/data/en/country/94/id.html

[4] S. P. Katongole, P. Akweongo, R. Anguyo, D. E. Kasozi, and A. Adomah-Afari, "Prevalence and Classification of Misdiagnosis Among Hospitalised Patients in Five General Hospitals of Central Uganda," *Clin. Audit*, vol. Volume 14, no. September, pp. 65–77, 2022, doi: 10.2147/ca.s370393.

[5] W. Nugraha and R. Sabaruddin, "Resampling Techniques to Overcome Class Imbalance in Diabetes Classification Using C4.5, Random Forest, and SVM," *Techno.Com*, vol. 20, no. 3, pp. 352–361, 2021, doi: 10.33633/tc.v20i3.4762.

[6] A. Tangkelayuk and E. Mailoa, "Classification of Water Quality Using the KNN, Naïve Bayes and Decision Tree Methods," vol. 9, no. 2, pp. 1109–1119, 2022.

[7] K. Siti, "CLASSIFICATION OF DIABETES USING THE DECISION TREE AND RANDOM FOREST METHOD," *repository.unsri.ac.id*, no. 8.5.2017, pp. 2003–2005, 2022.

[8] E. Rosta *et al.*, "Mental Health Data Classification in the Technology Industry Using the Random Forest Algorithm," vol. 1, no. 3, pp. 237–253, 2023.

[9] A. Husna Nasrullah, "IMPLEMENTATION OF

DECISION TREE ALGORITHM FOR CLASSIFICATION OF BEST-SELLING PRODUCTS," vol. 7, no. 2, pp. 45–51, 2021.

[10] A. Prabowo, S. Wardani, R. Wijaya Dewantoro, W. Wesly, and Leonardo, "Comparison of Random Forest and Decision Tree C4 Accuracy Levels. 5 On the Classification of Infertility Disease Data," vol. 4, no. 1, pp. 218–224, 2023, doi: 10.30865/klik.v4i1.1115.

[11] M. Sahebhonar and M. G. Dehaki, "A Comparison of Three Research Methods : Logistic Regression , Decision Tree , and Random Forest to Reveal Association of Type 2 Diabetes with Risk Factors and Classify Subjects in a Military Population," vol. 10, no. 2, pp. 9–11, 2022.

[12] N. R. Jevintya, U. Darusalam, S. Abdullah, and U. S. Asia, "APPLICATION OF THE K-MEANS AND DECISION TREE ALGORITHMS IN," vol. 7, no. 1, pp. 13–18, 2024, doi: 10.33387/jiko.v7i1.7580.

[13] L. Qadrini, A. Seppewali, and A. Aina, "DECISION TREE AND ADABOOST ON THE CLASSIFICATION OF RECIPIENTS OF SOCIAL ASSISTANCE PROGRAMS," *J. Inov. Researcher.*, vol. 2, no. 7, 2021.

[14] M. L. Suliztia, "APPLICATION OF RANDOM FOREST ANALYSIS TO THE PROTOTYPE OF THE USED CAMERA PRICE PREDICTION SYSTEM USING FLASK," *dspace.uii.ac.id*, 2020.

[15] M. Aqsha and N. Sunusi, "DATA CLASSIFICATION PERFORMANCE IS UNBALANCED WITH MACHINE LEARNING APPROACH (CASE STUDY: DIABETES INDIAN PIMA)," vol. 12, no. 2, pp. 176–193, 2023.

[16] F. Mu'alim and R. Hidayati, "Implementation of the Random Forest Method for Majors," vol. 14, no. 1, pp. 116–125, 2022.

[17] U. Azmi, "Detection of Dried Cannabis Aroma Using Random Forest Algorithm," *JITSI J. Ilm. Technol. Sist. Inf.*, vol. 4, no. 1, pp. 28–33, 2023, [Online]. Available: https://jurnal-itsi.org/index.php/jitsi/article/view/104%0Ahttps://jurnal-itsi.org/index.php/jitsi/article/download/104/82

[18] L. Mardiana, D. Kusnandar, and N. Satyahadewi, "DISCRIMINATION ANALYSIS WITH K FOLD CROSS VALIDATION FOR WATER QUALITY CLASSIFICATION IN PONTIANAK CITY," vol. 11, no. 1, pp. 97–102, 2022.

[19] A. Nurwalikadani, "Implementation of Smote Algorithm and Random Forest Classification on Imbalanced Lysine Protein Sequence Methylation Data," 2022, [Online]. Available: http://digilib.unila.ac.id/67956/%0Ahttp://digilib.unila.ac.id/67956/3/SKRIPSI FULL WITHOUT PEMBAHASAN.pdf