Journal of Applied Intelligent System (e-ISSN : 2502-9401 | p-ISSN : 2503-0493)
Vol. 9 No. 2, 2024, pp. 227 – 237

# C4.5 Algorithm Based on Forward Selection and Particle Swarm Optimization for Improving Accuracy in Heart Disease Patient Classification

**Aji Awang Setiawan**
*Universitas Dian Nuswantoro, Semarang, Indonesia*
*E-mail : ajiwangsa@gmail.com*

**Abstract -** Early detection of heart disease is crucial given the high number of cases occurring in advanced stages and affecting individuals in their productive years. Utilizing data mining, the C4.5 Algorithm is one method capable of detecting the onset of heart disease, prompting timely awareness and early prevention. The dataset employed is the Heart Disease Cleveland UCI from Kaggle, featuring 13 input attributes and 1 target attribute. Using the Decision Tree method results in decision-making by constructing a decision tree. The test outcomes revealed an accuracy rate of 77.11% with the C4.5 algorithm, 83.69% with the C4.5 algorithm employing Forward Selection, and 84.73% with the C4.5 algorithm based on Forward Selection and Particle Swarm Optimization.

**Keywords -** C4.5 Algorithm, Forward Selection, Particle Swarm Optimization, Data Mining, Cardiovascular

## 1. INTRODUCTION

Currently, cardiovascular disease remains a global threat as a leading cause of death. Data from the World Health Organization indicates that more than 17 million lives are lost annually due to heart attacks and vascular issues [1]. Based on historical data of heart disease patients, effective heart disease prediction recommendations can be made using Data Mining techniques [2] which can assist healthcare professionals through data classification approaches, one of which is the implementation of Decision Tree models [3][4].

Data Mining is the process of collecting and processing old data to discover patterns and relationships within a dataset [5]. Data Mining is the process of extracting information from large datasets, with one of the applied models being classification [4][6]. Classification aims to obtain a training data model that distinguishes attributes into appropriate categories. This model will be used to classify attributes into classes that were previously unknown [3][7]. Decision tree is one of the most popular classification methods because it is easily interpretable, similar to the C4.5 Algorithm [8].

The Decision tree method is a machine learning algorithm that applies rules to make decisions with a decision tree-like structure that models utilities (possible outcomes) and risks (possible consequences) [9]. The concept involves presenting the algorithm with conditional statements, in the form of branches to indicate decision-making steps and obtain results [10]. Based on the description, the C4.5 method based on Forward Selection and PSO (Particle Swarm Optimization) is used in modeling and classifying data to determine the classification results of heart disease patients. It is hoped that the accuracy level in data classification will be more effective. This research is very useful as it will gather information about the classifications made to identify patients who are potentially suffering from heart disease.

Additionally, it can be used as a reference and source for further research related to the use of the C4.5 algorithm, as well as a data mining exploration tool to discover new patterns.

## 2. RESEARCH METHOD

In this research, calculations will be carried out on the dataset using the C4.5 algorithm and the use of feature selection and the PSO (Particle Swarm Optimization) algorithm. The research flow/chart is as follows in Figure 1.

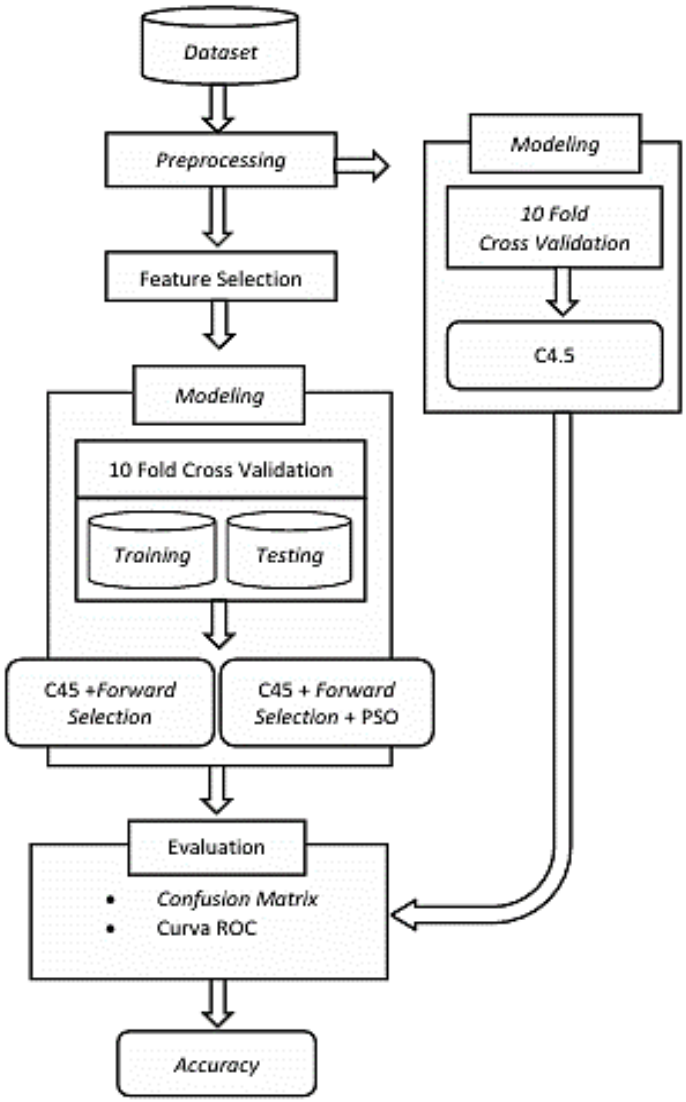


Figure 1. Proposed Method

The initial stage carried out was preparing a dataset that was published from Kaggle [11]. then data preprocessing is carried out to ensure the data used in the data mining analysis process is clean, relevant, and ready to be extracted patterns or run by the proposed algorithm. The stage carried out in data preprocessing is (data cleaning) to identify and resolve problems with incomplete, inconsistent or invalid data, such as missing or duplicate data. Then the next step is (data transformation) with data normalization. After data preparation is

complete, the next step is modeling. With the performance evaluation or model validation method using K-Fold Cross Validation with a percentage of 90% training data and 10% for test data. Model testing is carried out in three stages, namely Model testing is carried out in three stages, First, test the C4.5 Algorithm Model. Second, Testing the Model with the C4.5 Algorithm and Forward Selection. Third, testing the C4.5 Algorithm model based on Forward Selection and Particle Swarm Optimization (PSO). Evaluation results use a Confusion Matrix and ROC Curve which can be used to measure model performance, including metrics such as accuracy, precision, recall, and other evaluation metrics.

### *2.1. Datasets*

The data source used is (public data), namely the "Heart Disease Cleveland UCI" dataset from Kaggle. The dataset attribute table is as follows:

Table 1. Datasets Attributes

| Attribute | Description |
|---|---|
| *Age* | Age of the patient |
| *Sex* | Gender |
| *Cp* | Chest pain |
| *Trestbps* | Resting blood pressure |
| *Chol* | Cholesterol |
| *Fbs* | Fasting blood sugar >120mg/dl |
| *Restecg* | Resting electrocardiograph results |
| *Thalach* | Heart rate |
| *Exang* | Exercises with induced angina (1=yes, 0=no) |
| *Oldpeak* | Relative exercise-induced depression |
| *Slope* | ST segment peak slope |
| *Ca* | Number of colors of blood vessels |
| *Thal* | Type of blood vessel damage (2=temporary disability, 1=permanent disability, 0=normal |
| **condition** | Detected heart disease (1=yes, 0=no) |

### *2.2. Preprocessing*

There are three stages in data preprocessing. First, Identify and handle missing values. Identify attributes that have missing or incomplete values. The treatment for missing values is to delete rows or columns that have missing values. Second, Data Duplication. Handling duplication aims to ensure the integrity and quality of the dataset used in the data mining process. The treatment carried out is by deleting duplication. The next stage is normalization and standardization. Normalize and standardize data to change the scale or range of attributes so they can be compared fairly. Normalization is generally performed on numeric attributes.

### *2.3. Feature Selection*

The dataset will select the best attributes and delete attributes that do not contribute to accuracy. To carry out attribute selection, the (Forward Selection) method is used. Attributes will be tested by building a model. Then the model is tested to determine the resulting level of accuracy. Attributes with high accuracy results will be selected and then tested again on the remaining attributes. This process is repeated until the attribute being tested no longer provides a significant improvement.

### *2.4. Modeling*

The next step is to apply the data mining classification method using the C4.5 algorithm to build a decision tree. The model proposed in this research divides the data into (training dataset) and (testing dataset) then involves processing the dataset to obtain variables

that have been selected using the Forward Selection method, and optimization using the Particle Swarm Optimization (PSO) method.

*2.5. Evaluation*

At this stage, testing of the proposed method will be carried out to obtain information about accurate modeling. Then evaluate and compare the accuracy results of various experiments involving the use of the C4.5 algorithm, the C4.5 algorithm with Forward Selection, and the C4.5 algorithm with Forward Selection and Particle Swarm Optimization. In this process, we will evaluate and measure the accuracy results obtained for the model used using evaluation metrics, namely the Confusion Matrix and ROC Curve to obtain Accuracy, Precision and Recall values. The aim of this stage is to test and validate the performance of the proposed prediction model. By utilizing appropriate evaluation methods, you can identify the strengths and weaknesses of the model and measure the extent of its ability to make accurate predictions.

## 3. RESULTS AND DISCUSSION

*3.1 Datasets*

This dataset consists of 297 data records with each record having 14 attributes with the target attribute is the "Condition" column as in Table 2.

Table 2. Heart Disease datasets

| Age | Sex | Cp | Trestbps | Chol | Fbs | Restecg | Thalach | Exang | Oldpeak | Slope | Ca | Thal | Condition |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 69 | 1 | 0 | 160 | 234 | 1 | 2 | 131 | 0 | 0.1 | 1 | 1 | 0 | 0 |
| 69 | 0 | 0 | 140 | 239 | 0 | 0 | 151 | 0 | 1.8 | 0 | 2 | 0 | 0 |
| 66 | 0 | 0 | 150 | 226 | 0 | 0 | 114 | 0 | 2.6 | 2 | 0 | 0 | 0 |
| 65 | 1 | 0 | 138 | 282 | 1 | 2 | 174 | 0 | 1.4 | 1 | 1 | 0 | 1 |
| 64 | 1 | 0 | 110 | 211 | 0 | 2 | 144 | 1 | 1.8 | 1 | 0 | 0 | 0 |
| 64 | 1 | 0 | 170 | 227 | 0 | 2 | 155 | 0 | 0.6 | 1 | 0 | 2 | 0 |
| 63 | 1 | 0 | 145 | 233 | 1 | 2 | 150 | 0 | 2.3 | 2 | 0 | 1 | 0 |
| 61 | 1 | 0 | 134 | 234 | 0 | 0 | 145 | 0 | 2.6 | 1 | 2 | 0 | 1 |
| 60 | 0 | 0 | 150 | 240 | 0 | 0 | 171 | 0 | 0.9 | 0 | 0 | 0 | 0 |
| 59 | 1 | 0 | 178 | 270 | 0 | 2 | 145 | 0 | 4.2 | 2 | 0 | 2 | 0 |
| 59 | 1 | 0 | 170 | 288 | 0 | 2 | 159 | 0 | 0.2 | 1 | 0 | 2 | 1 |
| 56 | 1 | 0 | 120 | 193 | 0 | 2 | 162 | 0 | 1.9 | 1 | 0 | 2 | 0 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 35 | 1 | 3 | 126 | 282 | 0 | 2 | 156 | 1 | 0 | 0 | 0 | 2 | 1 |

*3.2 Preprocessing*

In the process of identifying missing values and duplicate data, there are no empty values or duplicate data so that the dataset remains intact and there is no reduction in the amount of data. The next stage is data normalization to facilitate model formation. Data normalization has been carried out for several attributes, namely converting numeric data into discrete data, namely the age (8 categories), trestbps (3 categories), chol (3 categories), thalach (3 categories) and oldpeak (3 categories) attributes. Thus, there is a data transformation after data preprocessing is carried out as follows in Table 3.

Table 3. Datasets after preprocessing

| age | sex | cp | trestbps | chol | fbs | restecg | thalach | exang | oldpeak | slope | ca | thal | condition |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 7 | 1 | 0 | 3 | 2 | 1 | 2 | 2 | 0 | 1 | 1 | 1 | 0 | 0 |
| 7 | 0 | 0 | 3 | 2 | 0 | 0 | 3 | 0 | 2 | 0 | 2 | 0 | 0 |
| 7 | 0 | 0 | 3 | 2 | 0 | 0 | 2 | 0 | 3 | 2 | 0 | 0 | 0 |
| 7 | 1 | 0 | 3 | 3 | 1 | 2 | 3 | 0 | 2 | 1 | 1 | 0 | 1 |
| 6 | 1 | 0 | 2 | 2 | 0 | 2 | 2 | 1 | 2 | 1 | 0 | 0 | 0 |
| 6 | 1 | 0 | 3 | 2 | 0 | 2 | 3 | 0 | 1 | 1 | 0 | 2 | 0 |
| 6 | 1 | 0 | 3 | 2 | 1 | 2 | 2 | 0 | 3 | 2 | 0 | 1 | 0 |
| 6 | 1 | 0 | 3 | 2 | 0 | 0 | 2 | 0 | 3 | 1 | 2 | 0 | 1 |
| 6 | 0 | 0 | 3 | 3 | 0 | 0 | 3 | 0 | 1 | 0 | 0 | 0 | 0 |
| 6 | 1 | 0 | 3 | 3 | 0 | 2 | 2 | 0 | 3 | 2 | 0 | 2 | 0 |
| 6 | 1 | 0 | 3 | 3 | 0 | 2 | 3 | 0 | 1 | 1 | 0 | 2 | 1 |
| 6 | 1 | 0 | 3 | 3 | 0 | 2 | 2 | 0 | 1 | 0 | 0 | 0 | 1 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 6 | 0 | 1 | 3 | 1 | 0 | 0 | 3 | 0 | 1 | 0 | 2 | 0 | 0 |

*3.3 Experiments with Algorithm C4.5*

To build a decision tree using the C4.5 algorithm, the initial step is to calculate the number of identified and unidentified classes affected by heart disease. Next, the Entropy calculation is carried out using equation (1).

$$Entropy(S) = \sum_{i=1}^{n} -pi * log_2(pi) \qquad (1)$$

*Entropy* (total) = ((-137/297) *$Log_2$ (137/297)+(-160/297) * $Log_2$ (160/297))
*Entropy* (total) = 0.995669658

To calculate the Gain for each attribute, the attribute entropy is calculated based on each case or representative attribute. The following is the entropy calculation for the Sex attribute.

Entropy (Sex, Women) = ((-25/96) *$Log_2$ (25/96)+(-71/96) * $Log_2$ (71/96))
Entropy (Sex, Women) = 0.827374478
Entropy (Sex, Men) = ((-112/201) *$Log_2$ (112/201)+(-89/201) * $Log_2$ (89/201))
Entropy (Sex, Men) = 0.990534146

After getting the entropy value for each attribute, the next step is to calculate the gain to determine which attribute is the most informative in predicting the target class as in equation (2).

$$Gain(S, A) = Entropy\ (\ S\ ) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} xEntropy(\ S_i) \qquad (2)$$

Gain (total,sex) = (0.995669658)-(((96/297)*( 0.827374478)+((201/297)* 0.990534146)))
Gain (total,sex) = 0.057873991

The overall calculation results are as follow in Table 4.

Table 4. Calculation of Gain and Entropy values

| Attribute | Value | Amount | Yes | No | Entropy | Gain |
|---|---|---|---|---|---|---|
| total | | 297 | 137 | 160 | 0.995669658 | |
| age | | | | | | 0.077538953 |
| | 1 | 0 | 0 | 0 | 0 | |
| | 2 | 0 | 0 | 0 | 0 | |
| | 3 | 3 | 0 | 3 | 0 | |
| | 4 | 50 | 13 | 37 | 0.826746372 | |
| | 5 | 85 | 29 | 56 | 0.92594006 | |
| | 6 | 118 | 75 | 43 | 0.946280454 | |
| | 7 | 39 | 19 | 20 | 0.999525689 | |
| | 8 | 2 | 1 | 1 | 1 | |
| sex | | | | | | 0.057873991 |
| | 0 | 96 | 25 | 71 | 0.827374478 | |
| | 1 | 201 | 112 | 89 | 0.990534146 | |
| cp | | | | | | 0.197203869 |
| | 0 | 23 | 7 | 16 | 0.886540893 | |
| | 1 | 49 | 9 | 40 | 0.688047624 | |
| | 2 | 83 | 18 | 65 | 0.754406204 | |
| | 3 | 142 | 103 | 39 | 0.848055283 | |
| trestbps | | | | | | 0.049652377 |
| | 1 | 0 | 0 | 40 | 0 | |
| | 2 | 97 | 37 | 65 | 0.91739616 | |
| | 3 | 200 | 100 | 39 | 0.959898524 | |
| Chol | | | | | | 0.007487692 |
| | 1 | 48 | 20 | 28 | 0.979868757 | |
| | 2 | 94 | 38 | 56 | 0.973385435 | |
| | 3 | 155 | 79 | 76 | 0.999729759 | |
| fbs | | | | | | 0.144279614 |
| | 0 | 254 | 117 | 137 | 0.995523003 | |
| | 1 | 43 | 20 | 23 | 0.99648599 | |
| Restecg | | | | | | 0.023473719 |
| | 0 | 147 | 55 | 92 | 0.953805105 | |
| | 1 | 4 | 3 | 1 | 0.811278124 | |
| | 2 | 146 | 79 | 67 | 0.995121443 | |
| thalach | | | | | | 0.128860335 |
| | 1 | 8 | 7 | 1 | 0.543564443 | |
| | 2 | 128 | 86 | 42 | 0.912999214 | |
| | 3 | 161 | 44 | 117 | 0.846148782 | |

| | | | | | | |
|---|---|---|---|---|---|---|
| exang | | | | | | 0.132294981 |
| | 0 | 200 | 63 | 137 | 0.898861037 | |
| | 1 | 97 | 74 | 23 | 0.790206924 | |
| oldpeak | | | | | | 0.146909226 |
| | 1 | 161 | 46 | 115 | 0.863120569 | |
| | 2 | 77 | 41 | 36 | 0.996956252 | |
| | 3 | 59 | 50 | 9 | 0.616166193 | |
| slop | | | | | | 0.178437756 |
| | 0 | 139 | 36 | 103 | 0.825222696 | |
| | 1 | 137 | 89 | 48 | 0.934393576 | |
| ca | | | | | | 0.18463533 |
| | 0 | 174 | 45 | 129 | 0.824657833 | |
| | 1 | 65 | 44 | 21 | 0.907696165 | |
| | 2 | 38 | 31 | 7 | 0.689201985 | |
| | 3 | 20 | 17 | 3 | 0.609840305 | |
| thal | | | | | | 0.210233513 |
| | 0 | 164 | 37 | 127 | 0.770279262 | |
| | 1 | 18 | 12 | 6 | 0.918295834 | |
| | 2 | 115 | 88 | 27 | 0.786255747 | |

From the entropy and gain calculations in Table 4, it was found that the "thal" attribute had the highest gain value, namely 0.210233513. Threfore, it will be chosen as the root or first node in forming the decision tree. Next, it is evaluated with a Confusion matrix to analyze the performance of the classification model by calculating the Accuracy, Precision and Recall values.

Table 5. Confusion Matrix

| Classification | | Actual Value | |
|---|---|---|---|
| | | 1 (positif) | 0 (negatif) |
| Predicted value | 1 (poositif) | TP<br>True Positive | FP<br>False Positive |
| | 0 (negatif) | FN<br>False Negative | TN<br>True Negative |

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}} \text{ x } 100\% \qquad (3)$$

Accuracy = (TP+TN)/(TP+FP+FN+TN) x 100%

Accuracy = (125+97)/(125+30+36+97) x 100% = 77.08%

$$\text{Precission} = \frac{\text{TP}}{\text{TP} + \text{FP}} \qquad (4)$$

Precision = TP/(TP+FP)

Precision = 125/(125+30) = 0.81

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (5)$$

Recall = TP/(TP+FN)

Recall = 125/(125+36) = 0.776

Table 6. Accuracy, Precision and Recall Values.

| Evaluation of Results | Values |
|---|---|
| Accuracy | 77.08% |
| Precision | 0.81 |
| Recall | 0.776 |

The test results of the testing data for the C4.5 algorithm on the ROC Curve value are as follow in Figure 2.

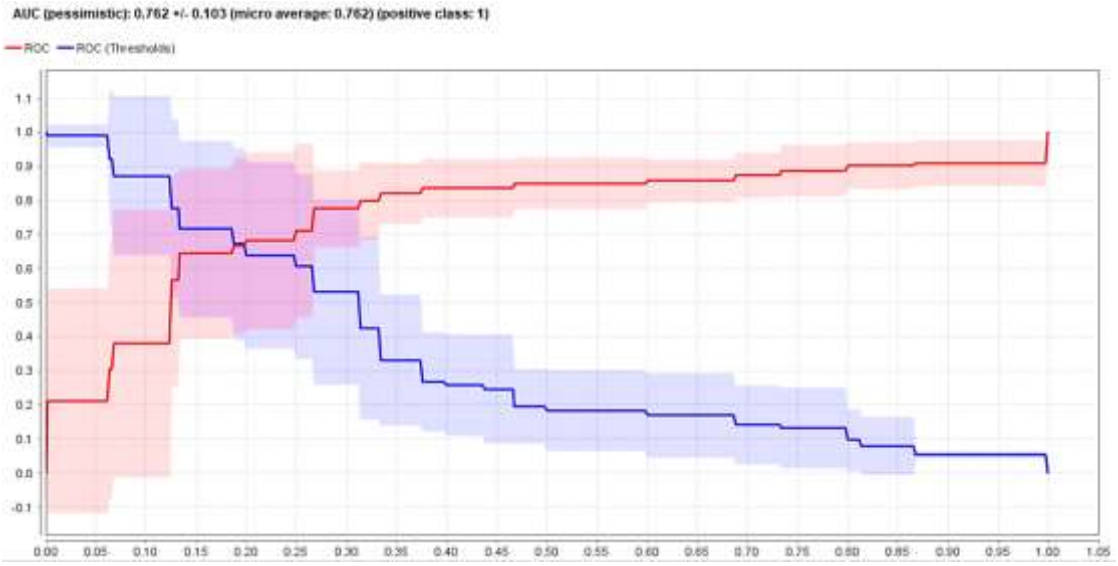


Figure 2. ROC Curve

*3.4 Experiment with C4.5 Algorithm With Feature selection*

In feature selection, the most relevant and informative attributes are selected to build a prediction model. This aims to reduce data dimensions and increase model efficiency and accuracy. The results of this experiment provide information about the extent of the influence of feature selection on the performance of the C4.5 algorithm in making predictions. The accuracy result is 83.69%, the precision value is 0.883 and the recall value is 0.568 and AUC (Area Under the Curve) value is 0.925.

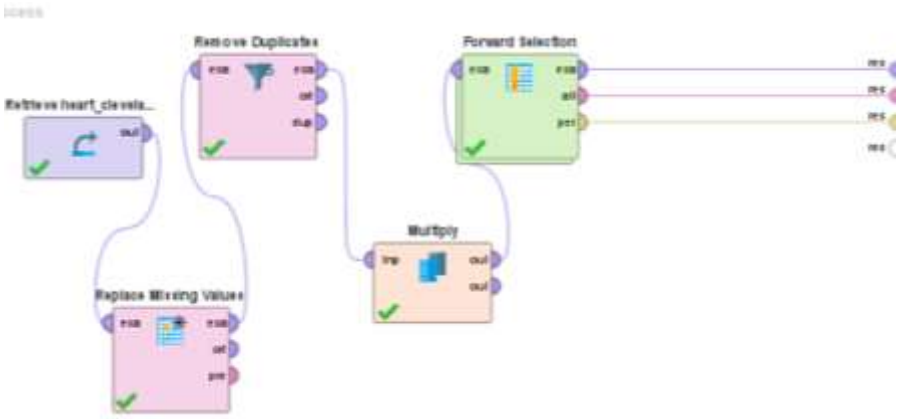


Figure 3. C.4.5 Classification Process with Forward Selection

Accuracy = (137+104)/(137+29+18+104) x 100% = 83.69%
Precision = 137/(137+18) = 0.883
Recall = 137/(137+104) = 0.568

*3.5 Experiment with C4.5 Algorithm With Feature selection and Particle Swarm Optimization.*

The C4.5 model was evaluated after going through a feature selection process using the forward selection and Particle Swarm Optimization (PSO) methods. Forward selection is used to iteratively add the most relevant attributes to the model, while PSO is used to find the combination of attributes that provides optimal results.

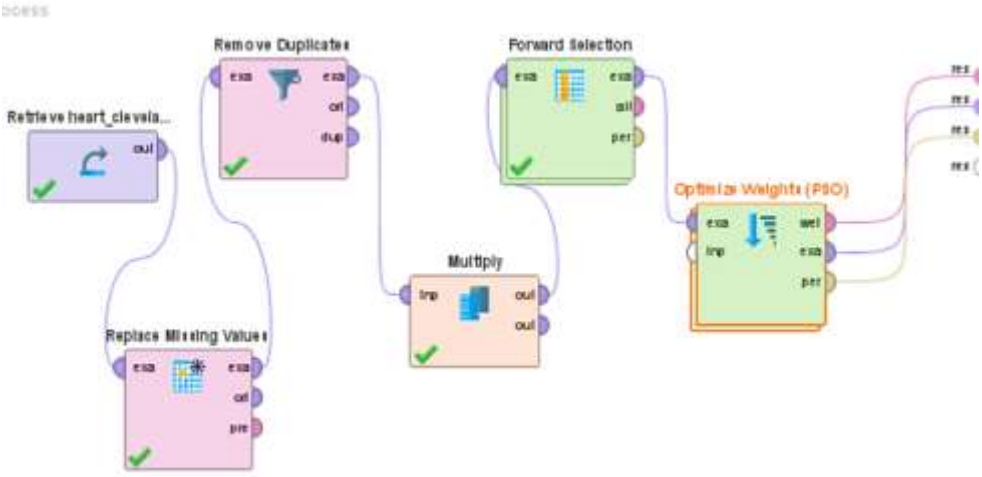


Figure 4. C.4.5 Classification Process with Forward Selection and PSO


Figure 5. C.4.5 Accuracy Results with Feature Selection and PSO

Accuracy = (139+105)/(139+28+16+105) x 100% = 84.73%
Precision = 139/(139+16) = 0.896
Recall = 139/(139+105) = 0.569

The accuracy result is 84.73%, the precision value is 0.896 and the recall value is 0.569 and and AUC (Area Under the Curve) value is 0.921. Table 7 shown the comparison results of accuracy values with the C4.5 algorithm, the C4.5 algorithm with feature selection, and the C.45 algorithm with feature selection and Particles Swarm Optimization.

Table 7. Comparison of model accuracy.

| Model | Accuracy | AUC |
|---|---|---|
| C4.5 algorithm | 77.11% | 0.793 |
| C4.5 algorithm + Feature Selection | 83.69% | 0.925 |
| C4.5 algorithm + Feature Selection + PSO | 84.73% | 0.921 |

## 4. CONCLUSION

The test results with the C4.5 Algorithm have an accuracy value of 77.11%. The C4.5 algorithm with Forward Selection has an accuracy value of 83.69%. The C4.5 algorithm based on Forward Selection and PSO (Particle Swarm Optimization) in predicting heart disease has an accuracy value of 84.73%. Based on these results, the application of C4.5 optimization techniques based on Forward Selection and Particle Swarm Optimization (PSO) is able to select attributes in the C4.5 algorithm, resulting in a better level of accuracy compared to the conventional C4.5 algorithm method. Based on the testing process and conclusions that have been carried out, suggestions for future research are to carry out exploration using alternative methods that are considered more optimal, Application of other optimization methods such as Genetic Algorithms or other optimization methods, Carry out development using other attribute selection methods such as backward elimination and so on to perfect attribute selection, Carry out tests on different datasets or use primary (private) data from hospitals.

## *REFERENCES*


[1] WHO (World Health Organization), "WHO." https://www.who.int/health-topics/cardiovascular-diseases

[2] M. A. Muslim and Dkk., *Data Mining Algoritma C4.5 Disertai Contoh Kasus dan Penerapannya dengan Program Computer*, 1st ed. 2019.

[3] J. Han, *Data Mining : Concepts and Techniques*. 2015.

[4] G. S. Santhana Krishnan J., "Prediction of Heart Disease Using Machine Learning Algorithms," *IEEE*, 2019, doi: 10.1109/ICIICT1.2019.8741465.

[5] W. S. Dharmawan, "Komparasi Algoritma Klasifikasi SVM-PSO Dan C4.5-PSO Dalam Prediksi Penyakit Jantung," *Informatika*, vol. 13, no. 2, 2021, doi: 10.36723/juri.v13i2.301.

[6] R. Rino, "Comparison of Data Mining Methods Using C4.5 Algorithm and Naive Bayes in Predicting Heart Disease," *Tech-E*, vol. 4, no. 2, 2021, doi: 10.31253/te.v4i2.543.

[7] Z. Maisat, E. Darmawan, and A. Fauzan, "The Implementation of GridSearchCV Hyperparameter Optimization in Heart Attack Prediction System Using SVM," *J. Ilm. Sist. Inf.*, vol. 13, no. 1, 2023.

[8] Y. Widiastiwi and I. Ernawati, “Klasifikasi Penyakit Batu Ginjal Menggunakan Algoritma Decision Tree C4.5 Dengan Membandingkan Hasil Uji Akurasi,” *IKRA-ITH Inform.*, vol. 5, no. 2, 2021.

[9] E. M. Z. Anirma Kandida Br Ginting, Maya Silvi Lydia, “Reduksi Atribut Menggunakan Chi Square untuk Optimasi Kinerja Metode Decision Tree C4.5,” *JEPIN*, vol. 9, no. 1, 2023.

[10] S. Kumar and G. Sahoo, “Enhanced Decision Tree Algorithm Using Genetic Algorithm for Heart Disease Prediction,” *Int. J. Bioinform. Res. Appl.*, vol. 14, no. 1–2, 2018, doi: 10.1504/IJBRA.2018.089199.

[11] “Heart Disease Cleveland UCI.” https://www.kaggle.com/datasets/cherngs/heart-disease-cleveland-uci

[12] I. P. E. P. I Made Agus Oka Gunawan, I Dewa Ayu Indah Saraswati, I Dewa Gede Riswana Agung, “Klasifikasi Penyakit Jantung Menggunakan Algoritma Decision Tree Series C4 . 5 Dengan Rapidminer,” *JTEKSIS*, vol. 5, no. 2, 2023, doi: 10.47233/jteksis.v5i2.775 Abstract.

[13] A. R. K. H. Warid Yunus, “Implementasi Algoritma C4.5 dalam Prediksi Penyakit Kanker,” *JIMIK*, vol. 4, no. 1, 2023, doi: 10.35870/jimik.v4i1.114.

[14] E. Nurlia and U. Enri, “Penerapan Fitur Seleksi Forward Selection Untuk Menentukan Kematian Akibat Gagal Jantung Menggunakan Algoritma C4.5,” *JUTIM*, vol. 6, no. 1, 2021.

[15] I. Ubaedi and Y. M. Djaksana, “Optimasi Algoritma C4.5 Menggunakan Metode Forward Selection Dan Stratified Sampling Untuk Prediksi Kelayakan Kredit,” *JSiI (Jurnal Sistem Informasi)*, vol. 9, no. 1. 2022. doi: 10.30656/jsii.v9i1.3505.

[16] A. M. Majid and M. N. Dwi Miharja, “Penerapan Metode Discretization Dan Adaboost Untuk Meningkatkan Akurasi Algoritma Klasifikasi Dalam Memprediksi Penyakit Jantung,” *Indones. J. Bus. Intell.*, vol. 5, no. 2, 2022, doi: 10.21927/ijubi.v5i2.2689.

[17] E. Prasetyo and B. Prasetiyo, “Increased Classification Accuracy C4.5 Algorithm Using Bagging Techniques In Diagnosing Heart Disease,” *J. Teknol. Inf. dan Ilmu Komput.*, vol. 7, no. 5, 2020, doi: 10.25126/jtiik.2020752379.

[18] R. Wajhillah, “Particle Swarm Optimization Untuk Prediksi Penyakit Jantung,” *SWABUMI*, vol. I, no. 1, 2014.

[19] H. W. N. Robby Anggriawan, “Komparasi algoritma c4.5 dan naive bayes dalam prediksi penderita penyakit gagal jantung,” *SIMADA*, vol. 5, no. 2, 2022.

[20] S. I. Novichasari and I. S. Wibisono, “Particle Swarm Optimization for Improved Accuracy of Disease Diagnosis,” *J. Appl. Intell. Syst.*, vol. 5, no. 2, 2020.

[21] E. L. Patrick Rim, “Optimizing the C4.5 Decision Tree Algorithm using MSD-Splitting,” *IJACSA*, vol. 11, no. 10, 2020.

[22] D. A. F. Ramdhan Saepul Rohman, Rizal Amegia Saputra, “Komparasi Algoritma C4.5 Berbasis PSO Dan GA Untuk Diagnosa Penyakit Stroke,” *CESS (Journal Comput. Eng. Syst. Sci.*, vol. 5, no. 1, 2020, doi: 10.24114/cess.v5i1.15225.

[23] H. El Hamdaoui, S. Boujraf, N. E. H. Chaoui, and M. Maaroufi, “A Clinical support system for Prediction of Heart Disease using Machine Learning Techniques,” *ATSIP*, 2020, doi: 10.1109/ATSIP49331.2020.9231760.

[24] D. Saputra, W. Irmayani, D. Purwaningtias, and J. Sidauruk, “A Comparative Analysis of C4.5 Classification Algorithm, Naive Bayes and Support Vector Machine Based on Particle Swarm Optimization (PSO) for Heart Disease Prediction,” *Int. J. Adv. Data Inf. Syst.*, vol. 2, no. 2, 2021, doi: 10.25008/ijadis.v2i2.1221.