Jurnal MATHOLOGI: Jurnal Pendidikan dan Riset Matematika
Volume 1, No. 1, Januari 2026 – Juni 2026, pp. 19 - 32

Application of Decision Tree Algorithms in Predicting the Risk of Stroke

Bagas Harmadi
Universitas Islam Negeri Maulana Malik Ibrahim Malang
e-mail: bagasharmadi29@gmail.com

Submitted : *13-01-2026* *Accepted* : *31-04-2026*
Revised : *11-03-2026* *Published* : *15-05-2026*

MATHOLOGI Jurnal Pendidikan dan Riset Matematika is licensed under a Creative Commons Attribution-Share Alike 4.0 International License

ABSTRACT

Health is the most important aspect of human life. Currently, many diseases are caused by germs, viruses, and bacteria, but the main cause is unhealthy habits or lifestyles. Stroke is one of these diseases. Therefore, an analysis of the prediction of a person's susceptibility to disease is needed, such as research related to the prediction of stroke. This study aims to determine the model and results of predicting the risk of stroke in humans using the Decision Tree algorithm. This method has a good level of accuracy and is effective in decision making. From the eight factors that cause a person to suffer from stroke, namely gender, age, hypertension, heart disease, type of residence, glucose levels, body mass index (BMI), and smoking status, the results of stroke risk prediction were obtained from 360 data using the C4 Decision Tree algorithm. 5 and the RapidMiner tool, with an accuracy percentage of 68.89%, precision of 68.68%, recall of 69.4%, and a ROC curve with an AUC of 0.726. These results indicate that the model in the Decision Tree method is classified as fair. Furthermore, using 360 data, a tree model was obtained with 28 decision rules, and the most dominant factor causing stroke was age above 65 years.

Keywords: Decision tree; Stroke Risk Prediction; RapidMiner

INTRODUCTION

Decision trees are one method that can be applied to perform prediction processes. This method produces a model that can predict data categories by studying the rules for determining categories based on the features possessed by the data (Ceballos, 2013). Decision trees also have a high level of accuracy when applied to large amounts of data compared to other methods. Therefore, in the health and medical industry, the accuracy of disease prediction is very important and requires effective decisions in analyzing and determining the accuracy of a disease suffered by a patient (Rifai, 2013).

One of the diseases that can be predicted is stroke. Stroke is a disease that occurs when the blood supply to the brain is disrupted due to a ruptured blood vessel or a blockage in the form of a blood clot. The effects of stroke can cause

difficulty in performing simple tasks, such as moving, walking, and loss of balance, loss of consciousness or fainting, and headaches without any apparent cause. The consequences of a stroke depend on the severity of the damage to the affected part of the brain. Even patients who experience a very serious stroke can die suddenly (American Stroke Association, 2020).

Stroke can be caused by several factors, including high blood pressure, a history of atrial fibrillation, cholesterol, diabetes mellitus, and heart disease (Hopkins, 2020). Stroke treatment is usually carried out manually. In practice, patients who have suffered a stroke visit a neurologist for examination, who asks them several questions about their symptoms and the possible causes of their stroke. The doctor will then diagnose the possible risk level experienced by the stroke patient. This treatment certainly poses its own problems, such as issues with financing and the very long time required. Therefore, an analysis or study is needed that can predict the likelihood of a person having a stroke, so that preventive measures can be taken before a stroke strikes.

Determining the right method for predicting the risk of stroke is very important because it can affect the results and conclusions obtained. Decision trees are an easy method for interpreting prediction results. Decision trees are often used in the process of classifying objects based on data with small differences and very close neighbors to the object. A decision tree aims to break down complex decision-making processes into simple ones using a decision tree (Kusrini & Luthfi, 2009). The general principle of the decision tree algorithm is to determine the label attributes and divide the data into training data and testing data. The data is then processed using the decision tree algorithm based on predetermined parameters. This method is therefore suitable for predicting the risk of stroke.

Previous researchers have conducted studies related to predictions using decision tree algorithms. First, research using the decision tree method with the application of the C4.5 algorithm was conducted by Susi Mashlahah in 2013. The results of this study explain that the test results using 60 sample data records produced an accuracy rate of 65.51%, 79 sample data records produced an accuracy rate of 70.96%, and 90 sample data records produced an accuracy rate of 82.79% (Mashlahah, 2013). Second, research conducted by Isa Iskandar et al. in 2019. The results of this study using a student graduation model produced an accuracy value of 73.19% with an AUC value of 0.80 (Rohman & Rufiyanto, 2019). Based on this, the researchers wanted to predict the risk of humans suffering from stroke using the Decision Tree algorithm.

METHOD

Data and Data Source

The dataset used was obtained based on secondary data sourced from the World Health Organization (WHO) and retrieved through the Kaggle repository (Han, 2006). This dataset was used to train and test classification models, specifically for stroke prediction. This dataset consists of 12 attributes, including ten independent variables, one ID variable, and one dependent variable or class

label used to predict stroke. Table 1 illustrates the feature information in the dataset used.

Table 1. Dataset

No	Attribute	Value	Description
1	Id	Nominal	Stroke patient identification code
2	Gender	Male	Patient gender
		Female	
		Other	
3	Age	Nominal	Continuous values with patient age ranges from toddlers to the elderly
4	Hypertension	1	Hypertensi
		0	Not hypertensive
5	Heart Disease	1	heart disease
		0	Not heart disease
6	Ever Maried	Yes	Get Merried
		No	Not Get Merried
7	Work Type	Children	
		Govt Job	
		Never Worked	
		Private	
		Self Employed	
8	Resident Type	Urban	
		Rural	
9	Avg Glucose Level	Nominal	Continuous glucose levels of patients
10	BMI	Nominal, N/A	Continuous body mass index values of patients
11	Smoking Status	Formerly Smoked	
		Never Smoked	
		Smokes	
		Unknown	
12	Stroke	1	Stroke
		0	Not Stroke

Ten dependent variables include gender, age, hypertension, heart disease, ever married, work type, resident type, average glucose level, BMI, and smoking status. The target attribute is stroke, which has two values: 0 indicates no indication of stroke, while 1 indicates the presence of stroke.

Data Analysis

This study has a sequence of steps in solving problems. Starting from selecting the dataset to be used, then performing the dataset preprocessing stage. To perform classification, data cleaning and transformation stages are carried out, as well as a data testing stage by dividing the data into training data and testing data. After dividing the data, the classifier can be implemented. The final step is to see the prediction results. The complete steps can be explained as shown in the following figure (Han, 2006).

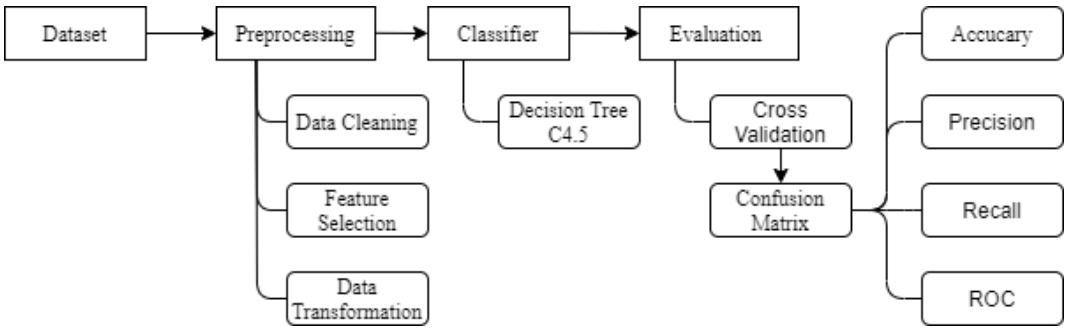

Figure 1. Classification Process Steps

1. Preprocessing
   a. Data Cleaning, is the process of removing data that has incomplete, empty, or missing values, or eliminating inconsistencies that appear in the dataset, as well as balancing unbalanced data.
   b. Feature Selection, which is the process of selecting variables to minimize the amount of data used in the mining process while still representing the original data.
   c. Data Transformation, which is the process of transforming selected data so that it is suitable for the mining procedure. In this process, changes are made to attribute values and discretization (Junaedi et al., 2011). The discretization process is carried out because continuous data is difficult to interpret due to having too many classes.
2. Classifier

Classifier or mining process is the most important process when the method is applied in the prediction process. In this process, the selection of methods to be used to find new and hidden knowledge or patterns from data is carried out, for example characterization, classification, regression, clustering or association. In this process, the selection of appropriate techniques, methods or algorithms is also carried out because it is highly dependent on the objectives and the overall process. At the classifier stage, researchers use one of the techniques in classification data mining, namely the C4.5 decision

tree algorithm. In this process, data processing is carried out with the help of RapidMiner software. At this stage, a decision tree model is also formed by calculating the total entropy value, the entropy of each value in the attribute, and the gain of each attribute. The highest gain value will become the first root and then the branch. The process is carried out on all attributes so that the decision tree is formed. The following are the entropy and gain formulas (Kusrini & Luthfi, 2009).

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

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

Keterangan:

S = case collection

N = Many values exist in the target attribute (number of classification classes)

pi = proportion of Si to S

A = Attribute

$|S_i|$ = number of cases in the i-th partition

$|S|$ = number of cases in S

3. Evaluation

The evaluation process is the process of translating patterns generated from data mining to test or evaluate the accuracy and performance of the methods used. This process uses a cross-validation technique with 10 tests. Cross-validation is a test method for evaluating the performance of the C4.5 decision tree algorithm, which will produce a confusion matrix. In measuring model performance in this study, there are four focal points, namely True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN). The following are some of the evaluations used (Indriani, 2014):

a. Accuracy is the level of accuracy in performing classification

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (3)$$

b. Accuracy is the level of accuracy in performing classification

$$Precision = \frac{TP}{TP + FP} \quad (4)$$

c. Recall or sensitivity is the effectiveness of classification in identifying positive values

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

RESULTS AND DISCUSSION

Preprocessing

1. Data Cleaning: In this process, data cleaning was performed on 201 empty values and 1,482 unknown values in the BMI variable. After the deletion process, there were 3,427 empty and unknown values remaining. Following data cleaning, class balancing was performed due to dataset imbalance. The following is an overview of the data before and after the preprocessing process.

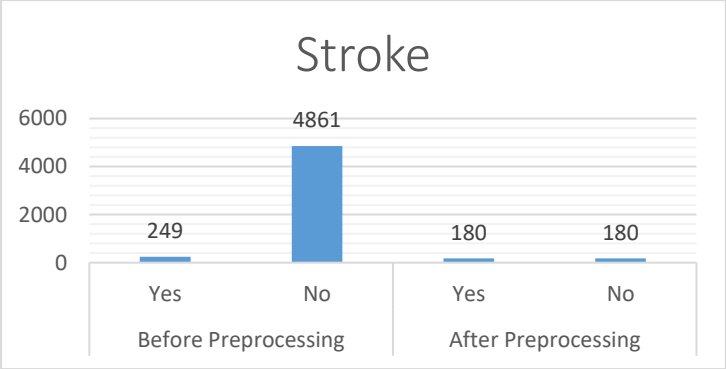

Figure 2. Dataset before and after preprocessing

2. Feature Selection, in the feature selection process, variables to be used in the study are selected. Variable selection is based on the causes of stroke (Hopkins, 2020). The variables used in the study are gender, age, hypertension, heart disease, resident type, average glucose level, BMI, smoking status, and stroke. The attributes to be removed are ID, ever married, and work type.
3. Data Transformation, in the data transformation process, data in the dataset is changed according to the format that can be processed by the software used. In this process, there are two transformation processes, namely changes in the attributes of hypertension, heart disease, and stroke, where the attribute values are changed from nominal to numerical, and the discretization process by changing the attribute values from continuous to categorical. The complete discretization process from continuous to categorical types is shown in Table 2 below.

Table 2. Type Dataset

No	Atribut	Row Data Type	Ready Data Type
1	Gender	Kategorik Nominal	Kategorik Nominal
2	Age	Numerik Rasio	Kategorik Ordinal
3	Hypertensi	Kategorik Nominal	Kategorik Nominal
4	Heart Disease	Kategorik Nominal	Kategorik Nominal
5	Resident Type	Kategorik Nominal	Kategorik Nominal

No	Atribut	Row Data Type	Ready Data Type
6	Avg Glucose Level	Numerik Rasio	Kategorik Ordinal
7	BMI	Numerik Rasio	Kategorik Ordinal
8	Smoking Status	Kategorik Nominal	Kategorik Nominal
9	*Stroke*	Kategorik Nominal	KategorikNominal

Classifier

In this process, the technique to be used is selected and a model is formed. The technique used in this study is the C4.5 decision tree. In the first stage of model formation, the root node is determined, followed by determining the branches of each node. Next, class division is performed on the branches obtained, and this process is repeated until each branch has a class.

The first step in the decision tree formation process is to calculate the total entropy value and the entropy of each attribute value in the data. The data is calculated using equation (1). The following is an example of the calculation of total entropy and entropy of the gender variable:

$$Entropy\ total = \left(\left(-\frac{180}{360}\right) * \log_2\left(\frac{180}{360}\right)\right) + \left(\left(-\frac{180}{360}\right) * \log_2\left(\frac{180}{360}\right)\right)$$

$$= 1$$

Then, the entropy value and gain value of the Gender variable are calculated.

- *Male*
  Number of cases : 145
  Yes : 75
  No : 70

$$Entropy\ Male = \left(\left(-\frac{75}{145}\right) * \log_2\left(\frac{75}{145}\right)\right) + \left(\left(-\frac{70}{145}\right) * \log_2\left(\frac{70}{145}\right)\right)$$

$$= 0.9991$$

- *Female*
  Number of cases : 215
  Yes : 105
  No : 110

$$Entropy\ Female = \left(\left(-\frac{105}{215}\right) * \log_2\left(\frac{105}{215}\right)\right) + \left(\left(-\frac{110}{215}\right) * \log_2\left(\frac{110}{215}\right)\right)$$

$$= 0.9996$$

Entropy calculations are performed on each class in the independent variable. After obtaining the entropy value for each class, gain calculations are performed for each variable. The following is an example of gain calculation.

$$Gain(s, A) = 1 - \left(\left(\frac{145}{360} entropy\ male \right) + \left(\frac{216}{360} entropy\ female \right) \right)$$

Using the same method, entropy and gain values were calculated for each of the other independent variables, namely Hypertension, Heart Disease, Smoking Status, Resident Type, Avg Glucose Level, and BMI. The next step was to calculate the gain value for each independent variable using equation (2). The results of the entropy and gain value calculations for the root node are presented in Table 3.

Table 3. Entropy and Gain Calculations

No	Variable	Value	Total (S)	Yes (Si)	No (Si)	*Entropy*	*Gain*
	Total		360	180	180	1	
1	*Gender*						0.00057
		Male	145	75	70	0.9991	
		Female	215	105	110	0.9996	
2	*Age*						0.19132
		6 - 11 th	2	0	2	0	
		12 - 16 th	2	0	2	0	
		17 - 25 th	20	0	20	0	
		26 - 35 th	16	1	15	0.3372	
		36 - 45 th	35	7	28	0.7219	
		46 - 55 th	54	24	30	0.9910	
		56 - 65 th	72	33	39	0.9949	
		Lebih dari 65 th	159	115	44	0.8509	
3	*Hypertension*						0.03073
		Yes	83	57	26	0.8968	
		No	277	123	154	0.9909	
4	*Heart Disease*						0.02886
		Yes	48	36	12	0.81128	
		No	312	144	168	0.99573	
5	*Smoking Status*						0.00772
		Formerly Smoked	104	57	47	0.9933	
		Never Smoked	181	84	97	0.9962	
		Smokes	75	39	39	0.9811	

No	Variable	Value	Total (S)	Yes (Si)	No (Si)	*Entropy*	*Gain*
6	*Resident Type*						0.00222
		Urban	178	94	84	0.9977	
		Rural	182	86	96	0.9978	
7	*Avg Glucose Level*						0.05424
		Kurang dari 101	187	77	110	0.9774	
		101 - 125	58	26	39	0.9039	
		Lebih dari 125	115	77	38	0.9153	
8	BMI						0.00698
		Kurang dari 18.6	5	1	4	0.7219	
		18.6 - 24.9	66	29	37	0.9893	
		25 - 29.9	120	64	56	0.9967	
		Lebih dari 29.9	169	86	83	0.9997	

Based on the table above, it can be seen that the variable with the highest gain is Age, which is 0.27169. Thus, Age will be the root node in the decision tree. There are 8 classes in the Age variable, namely 6-11 years, 12-16 years, 17-25 years, 26-35 years, 36-45 years, 46-55 years, 56-65 years, and over 65 years. Of these eight values, the variable value 6-11 years is classified as a No decision, the variable value 12-16 years is classified as a No decision, and the variable value 17-25 years is classified as a No decision, so no further calculations are needed. but for the values of the variables 26-35 years, 36-45 years, 46-55 years, 56-65 years, and Over 65 years, further calculations are still needed. Calculations are performed on all variables so that a decision tree is formed and rules are established. From the calculation results, a temporary decision tree can be illustrated as follows:

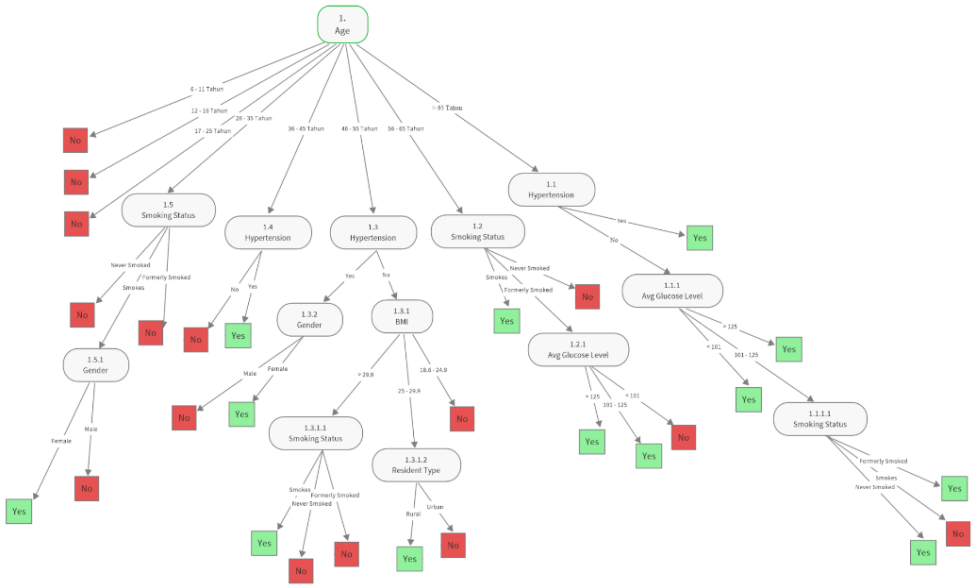

Figure 3. Tree Visualization

The decision tree generated 28 rules. The decision reached was Stroke with Yes and No classes. The rules are as follows:

Tree

```
Age = 12 -16 th: No {Yes=0, No=2}
Age = 17 - 25 th: No {Yes=0, No=20}
Age = 26 - 35 th
|   Smoking Status = formerly smoked: No {Yes=0, No=2}
|   Smoking Status = never smoked: No {Yes=0, No=10}
|   Smoking Status = smokes
|   |   Gender = Female: Yes {Yes=1, No=1}
|   |   Gender = Male: No {Yes=0, No=2}
Age = 36 - 45 th
|   Hypertension = No: No {Yes=6, No=27}
|   Hypertension = Yes: Yes {Yes=1, No=1}
Age = 46 - 55 th
|   Hypertension = No
|   |   BMI = 18.6-24.9: No {Yes=1, No=10}
|   |   BMI = 25-29.9
|   |   |   Resident Type = Rural: Yes {Yes=4, No=3}
|   |   |   Resident Type = Urban: No {Yes=4, No=5}
|   |   BMI = > 29.9
|   |   |   Smoking Status = formerly smoked: No {Yes=1, No=4}
|   |   |   Smoking Status = never smoked: No {Yes=2, No=4}
|   |   |   Smoking Status = smokes: Yes {Yes=5, No=2}
|   Hypertension = Yes
|   |   Gender = Female: Yes {Yes=6, No=0}
|   |   Gender = Male: No {Yes=1, No=2}
Age = 56 - 65 th
|   Smoking Status = formerly smoked
|   |   Avg Glucose Level = 101-125: Yes {Yes=4, No=0}
|   |   Avg Glucose Level = < 101: No {Yes=3, No=12}
|   |   Avg Glucose Level = > 125: Yes {Yes=7, No=0}
|   Smoking Status = never smoked: No {Yes=6, No=24}
|   Smoking Status = smokes: Yes {Yes=13, No=3}
Age = 6 - 11 th: No {Yes=0, No=2}
Age = > 65 th
|   Hypertension = No
|   |   Avg Glucose Level = 101-125
|   |   |   Smoking Status = formerly smoked: Yes {Yes=3, No=3}
|   |   |   Smoking Status = never smoked: Yes {Yes=6, No=2}
|   |   |   Smoking Status = smokes: No {Yes=0, No=2}
|   |   Avg Glucose Level = < 101: Yes {Yes=32, No=19}
|   |   Avg Glucose Level = > 125: Yes {Yes=32, No=9}
|   Hypertension = Yes: Yes {Yes=42, No=9}
```

Figure 4. Rule Tree

From Figure 4 above, it can be explained that the most influential factor on the first node is age, the second node is hypertension, the third node is average glucose level, and the fourth node is smoking status. For the stroke model, the best rule is when the patient is over 65 years old with high blood pressure and a total of 42 attributes. Meanwhile, for the non-stroke model, the best rule is when the patient is 36-45 years old with normal blood pressure and a total of 27 attributes.

Evaluation

The evaluation process was conducted to analyze the classification results. Data measurement was performed using a confusion matrix by evaluating the results of the C4.5 decision tree algorithm. The evaluation process was carried out using RapidMiner software to determine the accuracy of the calculations that had been performed. The attribute used as a label was Stroke. The author analyzed 360 data records that would be used as training data and testing data using the 10-fold cross-validation technique. The evaluation process uses a confusion matrix and ROC curve.

1. Confusion Matrix

The following is a confusion matrix table of the test results using the RapidMiner tool with 360 data points. The test results identified 248 data points as correct and 112 data points as incorrect.

Tabel 4. Confusion Matrix

	True Yes	**True No**
Pred. Yes	125	57
Pred. No	55	123

Table 4 is a confusion matrix table from testing data using 10-fold cross validation on RapidMiner software. There are 125 true yes (TY) records, 57 false yes (FY) records, 55 false no (FN) records, and 123 true no (TN) records. After testing with the confusion matrix, the accuracy, precision, and recall levels were calculated using equations (3), (4), and (5) as follows:

a. Accuracy

Accuracy was calculated by dividing the number of correctly classified data by the total sample data tested.

$$\text{Accuracy} = \frac{TY+TN}{TY+TN+FY+FN}$$

$$= \frac{125 + 123}{125 + 123 + 55 + 57}$$

$$= 0.6889$$

$$= 68.89\ \%$$

Accuracy is the degree of closeness between the predicted value and the actual value. The accuracy obtained from the model is 68.89%, which means that out of 360 patient data, 248 were successfully predicted.

b. Precision

The precision value is calculated by dividing the number of correct data with a value of yes (true yes) by the number of correct data with a value of yes (true yes) and incorrect data with a value of yes (false yes).

$$\text{Precision} = \frac{TY}{TY+FY}$$
$$= \frac{125}{125+57}$$
$$= 0.6868$$
$$= 68.68\ \%$$

Precision is the probability of a predicted positive case actually belonging to the positive category. In this case, the precision rate is 68.8%, which means that out of 182 stroke patient prediction data, the model successfully predicted 125 actual stroke cases correctly.

c. Recall

Recall is calculated by dividing the true yes data by the sum of the true yes data and the false no data.

$$Recall = \frac{TY}{TY+FN}$$
$$= \frac{125}{125+55}$$
$$= 0.694$$
$$= 69.4\ \%$$

Recall is the probability of correctly predicting positive cases. In this case, the recall rate is 69.4%, which means that from the actual data of 180 stroke patients, the model successfully predicted 125 stroke cases.

2. ROC Curve/AUC

The results of the ROC Curve analysis using the RapidMiner tool can be seen in Figure 5.

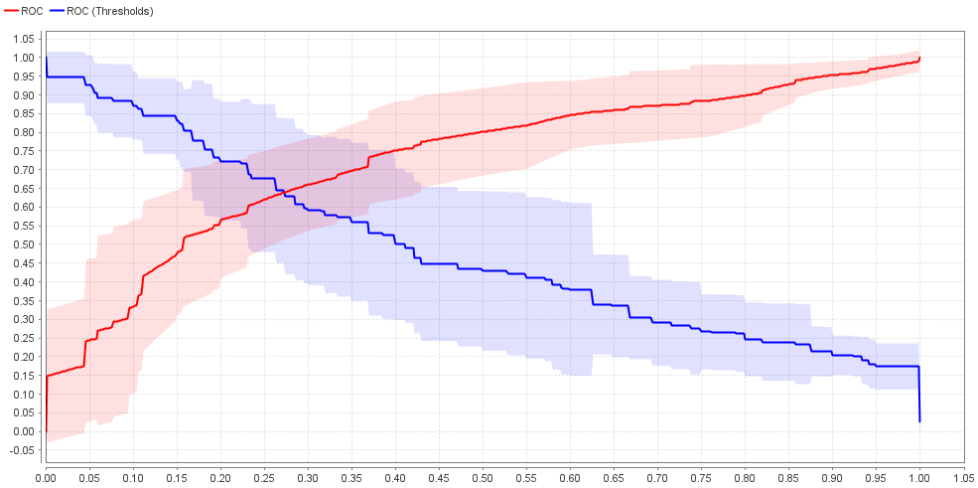

Figure 5. ROC Curve

AUC (Area Under Curve) is calculated to measure the accuracy and comparison of classification virtually with false positives as a horizontal line (blue line) and true positives as a vertical line (red line). The closer the ROC curve is to the Y(1,0) line, the better the model produced. When the prediction is correct for an example, the curve takes one step up (increased TP). If the prediction is incorrect, the curve takes one step to the right (increased FP). From the data above, analysis using the RapidMiner tool with Decision tree measurements shows a relationship between false yes and true yes of 0.726, which falls into the fair category (Fair Classification).

CONCLUSION

Based on the above problem formulation and discussion, it can be concluded that the stroke prediction model using 316 data points with the C4.5 Decision Tree algorithm successfully predicted stroke with a fairly high percentage, namely an accuracy of 68.89%, precision of 68.68%, and recall of 69.4%. The tree model produced 28 rules, with 13 predicting Yes and 15 predicting No. The most influential factor based on the tree formed was the Age variable. Then, using the ROC curve calculation, an AUC of 0.726 was obtained, which means that the prediction model using the Decision Tree C4.5 algorithm is classified as fair classification.

REFERENCES

American Stroke Association. (2020). *"About Stroke,."* Stroke.Org. https://www.stroke.org/en/about-stroke

Ceballos, F. (2013). *Scikit-Learn Decision Trees Explained*. Medium. https://towardsdatascience.com/scikit-learn-decision-trees-explained-803f3812290d

Han, J. (2006). *Data Mining:Concepts and Techniques*. Diane Cerra.

Hopkins, J. (2020). *Risk Factors for Stroke*. Hopkinsmedicien.Org. https://www.hopkinsmedicine.org/health/conditions-and-diseases/stroke/risk-factors-for-stroke

Indriani, A. (2014). Klasifikasi Data Forum dengan menggunakan Metode Naïve Bayes Classifier. *Seminasi Nasional Aplikasi Teknologi Informasi*, 5–10.

Junaedi, H., Budianto, H., Mariati, I., & Melani, Y. (2011). DATA TRANSFORMATION PADA DATA MINING. *Inovasi Dalam Desain Dan Teknologi*.

Kusrini, E. T., & Luthfi. (2009). *Algoritma Datamining,*. CV Andi Offset.

Mashlahah, S. (2013). *Prediksi kelulusan mahasiswa menggunakan metode decision tree dengan penerapan algoritma C4.5*. UIN Maulana Malik Ibrahim Malang.

Rifai, B. (2013). ALGORITMA NEURAL NETWORK UNTUK PREDIKSI PENYAKIT JANTUNG. *Jurnal Techno Nusa Mandiri*, *IX*(1).

Rohman, A., & Rufiyanto, A. (2019). PENERAPAN ALGORITMA DECISION TREE ID3 UNTUK PREDIKSI KELULUSAN MAHASISWA JENJANG PENDIDIKAN D3 DI FAKULTAS TEKNIK UNIVERSITAS PANDANARAN. *Jurnal Neo Teknika*, *5*(2).