VOL. 22 (1) (2025), 64-75 p-ISSN: 693-7554, e-ISSN:2654-3990

Analysis of Student Dropouts at SMA Negeri 6 Padangsidimpuan Using C5.0 Algorithm

Wahyudi Saleh Nasution[1], **Sriani**[2*]

[1,2] Study Program of Computer Science, Faculty of Science and Technology, State Islamic University of North Sumatra, 20353, Medan

Abstract

The issue of student dropouts in Indonesia, including at SMA Negeri 6 Padangsidimpuan City, poses a significant challenge in the field of education. This phenomenon hampers the achievement of compulsory education programs and impacts the quality of human resources. The dropout rate at SMA Negeri 6 Kota Padangsidimpuan reached 19.08% in the 2023/2024 academic year, highlighting the urgency of this problem. Factors such as economic conditions, geographical constraints, and individual motivation are the primary contributors to this issue. This study aims to analyze the factors contributing to student dropouts using the C5.0 decision tree algorithm method. The research data was collected through interviews, observations, and school archives, then processed using RapidMiner software. The research process included data preprocessing, classification model development, and model evaluation using a confusion matrix. The results showed that the C5.0 decision tree algorithm could identify significant relationships among these variables, achieving an accuracy rate of 87%. In conclusion, the C5.0 decision tree algorithm is effective in analyzing the factors causing dropouts and can serve as a basis for formulating more targeted prevention strategies.

Keywords: *dropout, C5.0 decision tree algorithm, education, contributing factors, prevention*

1. Introduction

In this era of globalisation, education is one of the most important sectors for a country in an effort to develop the quality of Human Resources (HR). Based on the National Education System Law it emphasises that the function of education is as a medium to form a dignified personality in educating and exploring the potential of each individual, in order to make the nation's generation obedient to God Almighty, behave well, be dignified and insightful. However, not all people understand the importance of seeking knowledge, and not a few of the nation's children decide to drop out of school based on various factors [1].

Education is one of the important factors in building quality human resources. One of the indicators of educational success is the low level of student dropout. A student dropout is an event in which a student stops attending school before completing his or her education [2]. The dropout rate of students in Indonesia is still relatively high, including at SMA Negeri 6 Padangsidempuan City.

Dropping out of school is an educational problem that must be addressed at the root. Especially at the secondary school level or equivalent, which is the final stage of the government scheme, it is mandatory to study for 12 years, because after that a person is considered fit to enter the world of work. Some of the factors that cause this include economic factors, geographical conditions, and the desire of the students themselves [3]

Therefore, in this study, the researcher plans to analyse the factors that cause dropout at SMA Negeri 6 Padangsidimpuan City by using the decision tree method built with the C5.0 algorithm. In the 2023/2024 school year, the number of dropout students at SMA Negeri 6 Padang Sidimpuan City reached 200 students, which is 19.08% of 1048 students.

One of the methods that can be used for dropout analysis is to use the C5.0 decision tree algorithm.

*Corresponding author. *E-mail address:* wahyudisaleh12@gmail.com
Received: 13 Januari 2025, Accepted: 20 January 2025 and available online 31 January 2025
DOI: https://doi.org/10.33751/komputasi.v19i2.5260

Decision tree algorithms are used to solve the uncertainty of a problem. By using the decision tree algorithm, the cause of the problem can be known so that a definite conclusion can be drawn. This is done because these algorithms can tolerate imprecise or uncertain data. It was created by J. Ross Quinlan in 1960 [4].

The research conducted by [5] with the title "Application of the C5.0 Algorithm in the Classification of Factors Causing Dengue Fever" aims to apply the C5.0 algorithm in classifying the factors that cause dengue fever (DHF) based on patient medical record data at Dr. H. Soemarno Sosroatmodjo Hospital, Bulungan Regency, as well as to test and evaluate the effectiveness of the algorithm in solving health problems practically in the area. The results showed that the age variable had the greatest influence on the incidence of dengue compared to other factors such as gender, history of comorbidities, and occupational status. The C5.0 algorithm was chosen because of its ability to build an accurate classification model based on the gain ratio [6]. These findings show the importance of algorithm-based data analysis to understand patterns and factors that affect a phenomenon. This research is expected to provide deeper insights into the factors that affect the incidence of dengue, as well as contribute to disease prevention and control efforts at the local level [7]

Meanwhile, the research to be studied focuses on the analysis of factors of dropout students at SMA N 6 Padangsidimpuan City using the C5.0 decision tree algorithm method. Thus, the main difference lies in the focus of the research, where previous research is related to the health field, while this research focuses on the field of education. This research is expected to be able to analyse and classify student data at SMA Negeri 6 Padangsidimpuan City to support measures to prevent student dropouts in the future. The results of this study are expected to support the school in identifying the factors that cause student dropout and developing appromalete steps to overcome the problem.

2. Methods

2.1 Planning

At this stage, it is the beginning to determine the problem before conducting research on the research object. By looking for sources of problem information on the object of research to find solutions related to problems. So that it can describe the problem and facilitate the steps in solving the problem and in this planning the researcher has also planned the goal [8]

2.2 Data Collection

Data collection in this study was carried out through three main techniques. First, literature studies were conducted to obtain a theoretical foundation and an in-depth understanding of the topic being researched by referring to various sources such as books, journals, and other scientific works. Second, direct observation was carried out at SMA N 6 Padangsidimpuan City to collect primary data in the form of student academic data and related attributes. Finally, interviews were conducted to get more information about the problem of student dropout, including the handling efforts that have been made. The combination of these three techniques aims to obtain comprehensive and accurate data to support analysis in research.

2.3 Data Analysis

In the preparation of this thesis, there are several stages that need to be completed. Starting from the preparation of attributes which will later go through several stages beginning with data description, data cleaning, then at the dataset classifier stage will be classified using the C5.0 decision tree algorithm model. Finally, in the last stage, the data will be evaluated using a confusion matrix to obtain accuracy, precision, and recall values.

2.4 Design

Based on the analysis carried out, the author will dig up the data for grouping with predetermined criteria (variables). The criteria taken from the study were gender, parental education, parental employment, family members, and family income. The software used to complete this stage is Rapidminer [9].

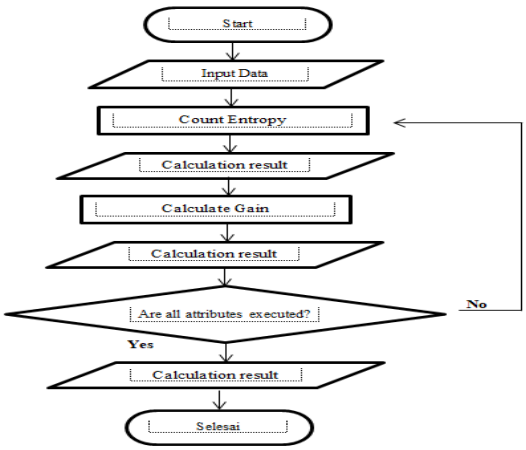

Figure 1. Flowchart Algorithm C5.0

2.5 Testing

At this stage, the research conducted tests after obtaining data to find out whether the data that had been processed was in accordance with expectations. The data tested was entered into data processed using RapidMiner.

3. Results and Discussion

3.1 Data Representation

The results and discussion in this study are designed to provide an understanding of the application of the C5.0 Decision Tree algorithm in analysing the factors that cause students to drop out at SMA Negeri 6 Padangsidimpuan City. This study uses student data for the 2023/2024 school year as many as 500 students with four main variables, namely gender, distance from home to school, parental work, and extracurricular participation. The analysis process begins by determining relevant variables, preprocessing data, and applying the C5.0 Decision Tree algorithm to build a decision tree. The important stages include calculating entropy and gain to determine the main attributes as tree roots. Branches are created based on the value of the maximum gain ratio until all classes on the branch have a consistent label.

3.2 Data Reprocessing

The next process after the dataset is successfully entered is in the preprocessing stage. As in the design that has been explained in the previous chapter 3, in this preprocessing stage it is divided into two, namely the transformation of numerical data into categorical data and dividing the data into training data and testing data. The next process after the dataset is successfully entered is in the preprocessing stage [10].

a. Gender

The gender variables of students are arranged into two types of categories, namely female and male.

b. Distance

The distance variable represents the distance between students' homes and the school. To optimize data processing in RapidMiner, it was categorized as follows: 0–5 km ("Medium"), 5.1–10 km ("Keep"), and >10 km ("Far").

c. Parents' work

The parental work variable refers to the occupation of the students' fathers, categorized as Farmer, Self-Employed, Civil Servants, Private Employees, and Not Working.

d. Extracurricular Participation

The extracurricular participation variable is a variable that shows student participation in extracurricular activities at SMA Negeri 6 Padang Sidimpuan City. The type of extracurricular participation Use active and inactive categories.

3.3 Algorithmic Process

The first step in forming a classification tree for the C5.0 algorithm is to determine the root node. Next, determine which branch each node enters into. Class separation is then repeated on the processed branches until each branch has a class. For example, the data used in the process of creating a classification tree consists of 80% of the total data, 838 samples (learning data), and the remaining 20%, 210 samples, are used as test data. The classification tree that is formed. The first step in the process of forming a classification tree is to calculate the entropy value. The first step is to calculate the entropy on each category on the status attribute, i.e. the entropy on the active and dropout categories [11].

1. Sharing of training data and data testing

The first step before running the classification process is to separate the training data and test data and

then randomise it so that all data has the same possibility to become training data and test data. The following is an example of a calculation to determine the amount of data that goes into the training data using the 80 : 20 division.

$$\text{Jumlah data } training = \frac{80}{100}\ x\ 500 = 400$$

The following is a calculation to determine the amount of data that goes into the testing data:

$$\text{Number of training data} = \frac{20}{100}\ x\ 500 = 100$$

Based on the results of the calculation above, it can be seen that the data entered into the training data for the 80:20 porpoise was 838 and the remaining 210 data was entered into the testing data.

2. C5.0 algorithm calculation

Proportion of each class:

The proportion of each class is the amount of data in that class divided by the total amount of data..

$Class\ proportions$ "Active" $= \frac{432}{500} = 0.864$

$Class\ proportions$ "$Dropout$" $= \frac{68}{500} = 0.136$

After knowing the proportions of each class, the next step is to calculate the entropy value of each class.

Entropy class "Active" :

$Active = -\left(0.864x\ log_2(0.864)\right) = -\left(0.864\ x\ (-0.209)\right) = 0.181$

Entropy kelas "*Dropout*" :

$$Dropout = -\left(0.136\ x\ log_2(0.136)\right) = = -\left(0.136\ x\ (-2.878)\right) = 0.391$$

Next is to calculate the total entropy value of each class with the equation below.

$Entropy\ total = E_{aktif} + E_{dropout} = 0.181 + 0.391 = 0.572$

Table 1. Total Entropy Results

Total Amount	Status "Active"	Status "*Dropout*"	*Entropy*
500	432	68	0.573

After obtaining the total entropy value, the next step is to calculate the entropy value of each variable, namely gender, home distance, parental work and extracurricular participation. The first step is to calculate the entropy value for gender.

a. Gender

Proportion to gender "Male".

Status "Active" : 216

Status "*dropout*" : 46

Total : 262

$Proporsi$ "Active" $= \frac{216}{262} = 0.824$

$Proporsi$ "$Dropout$" $= \frac{46}{262} = 0.176$

Entropy total to gender "Male".

$E_{Pria} = -(0.824x\ log_2(0.824) + (-(0.176x\ log_2(0.176)\ ) = \mathbf{0.668}$

Proportion to gender "Female".

Status "Active" : 216

Status "*dropout*" : 22

Total : 238

$Proporsi$ "Active" $= \frac{216}{238} = 0.907$

$Proporsi\ "Dropout"\quad = \frac{22}{238} = 0.092$

Entropy to gender "Female".
$E_{female} = -(0.907x\ log_2(0.907) + (-(0.092x\ log_2(0.092)) = \mathbf{0.44}$

After obtaining each entropy from the gender attribute, the next step is to calculate the weight entropy value using the equation below.

$E_{gender} = \frac{262}{500}\ x\ 0.668 + \frac{238}{500}\ x\ 0.44 = 0.559$

After the weight entropy is obtained, the next step is to calculate the gain value for gender attributes using the equation below.

$Gain_{gender} = E_{Total} - E_{gender} = 0.572 - 0.559 = \mathbf{0.013}$

After the gain value is obtained, the next step is to calculate the gain ratio value using the equation below.

$Ratio = \frac{Gain_{gender}}{E_{male} + E_{female}} = \frac{0.013}{0.668 + 0.44} = \mathbf{0.0117}$

b. Distance From Home to School

Proportions for "Close" distance
Status "Active" : 138
Status "*dropout*" : 22
Total : 160

$Proporsi\ "Active"\quad = \frac{138}{160} = 0.8625$

$Proporsi\ "Dropout"\quad = \frac{22}{160} = 0.1375$

Total entropy for "Close" distance.
$E_{near} = -(0.8625x\ log_2(0.8625) + (-(0.1375x log_2(0.1375)) = \mathbf{0.57}$

Proportion to "Medium" distance.
Status "Active" : 142
Status "*dropout*" : 23
Total : 165

$Proporsi\ "Active"\quad = \frac{142}{165} = 0.8606$

$Proporsi\ "Dropout"\quad = \frac{23}{165} = 0.1394$

Entropy for "Medium" distances.
$E_{medium} = -(0.8606x\ log_2(0.8606) + (-(0.1394x\ log_2(0.1394)) = \mathbf{0.576}$

Proportion to "Away" distance.
Status "Active" : 152
Status "*dropout*" : 23
Total : 175

$Proporsi\ "Active"\quad = \frac{152}{175} = 0.8686$

$Proporsi\ "Dropout"\quad = \frac{23}{175} = 0.1314$

Proportion to "Away" distance.
$E_{far} = -(0.8686x\ log_2(0.8686) + (-(0.1314x\ log_2(0.1314)) = \mathbf{0.5569}$

After obtaining each entropy from the distance attribute, the next step is to calculate the weight entropy value using the equation below.

$E_{distance} = \frac{160}{500}\ x\ 0.57 + \frac{165}{500}\ x\ 0.576 + \frac{175}{500}\ x\ 0.5569 = 0.567$

After the weight entropy is obtained, the next step is to calculate the gain value for the distance attribute using the equation below.

$Gain_{distance} = E_{Total} - E_{jarak} = 0.572 - 0.5688 = \mathbf{0.0032}$

After the gain value is obtained, the next step is to calculate the gain ratio value using the equation below.

$Ratio = \frac{Gain_{distance}}{E_{near}+E_{medium}+E_{far}} = \frac{0.0032}{0.57+0.576+0.5569} = \mathbf{0.00188}$

c. Parents Work

Proportion to the work of "Farmer".
Status "Active" : 212
Status "*dropout*" : 33
Total : 245
$Proporsi$ "Active" $= \frac{212}{245} = 0.8637$
$Proporsi$ "$Dropout$" $= \frac{33}{245} = 0.1347$

Total entropy for the "Farmer" job.
$E_{farmer} = -(0.8637x\ log_2(0.8637) + (-(0.1347\ x\ log_2(0.1347)) = \mathbf{0.5594}$

Proportion of "Self-employed" jobs.
Status "Active" : 186
Status "*dropout*" : 30
Total : 216
$Proporsi$ "Active" $= \frac{186}{216} = 0.8611$
$Proporsi$ "$Dropout$" $= \frac{30}{216} = 0.1389$

Total entropy for 'Self-employed' jobs.
$E_{self\ employed} = -(0.8611x\ log_2(0.8611) + (-(0.1389\ x\ log_2(0.1389)) = \mathbf{0.5756}$

Proportion for "civil servant" jobs.
Status "Active" : 27
Status "*dropout*" : 2
Total : 29
$Proporsi$ "Active" $= \frac{27}{29} = 0.931$
$Proporsi$ "$Dropout$" $= \frac{2}{29} = 0.0689$

Total entropy for "civil servant" jobs.
$E_{civil\ servant} = -(0.931x\ log_2(0.931) + (-(0.0689\ x\ log_2(0.0689))\ = \mathbf{0.3616}$

Proportion to "Private Employees" jobs.
Status "Active" : 4
Status "*dropout*" : 2
Total : 6
$Proporsi$ "Active" $= \frac{4}{6} = 0.667$
$Proporsi$ "$Dropout$" $= \frac{2}{6} = 0.334$

Total entropy for "Private Employee" jobs.
$E_{\text{Private Employee}} = -(0.667x\ log_2(0.667) + (-(0.334x\ log_2(0.334)) = \mathbf{0.917}$

Proportion to "Not Working" jobs..
Status "Active" : 3
Status "*dropout*" : 1
Total : 4

$Proporsi$ "Active" $= \frac{3}{4} = 0.75$
$Proporsi$ "$Dropout$" $= \frac{1}{4} = 0.25$

Total Entropy for "Not Working" jobs.
$E_{\text{Not Working}} = -(0.75\ x\ log_2(0.75) + (-(0.25\ x\ log_2(0.25)) = \mathbf{0.812}$

After obtaining each entropy from the parent's occupational attributes, the next step is to calculate the weight entropy value using the equation below.

$$E_{work} = \frac{245}{500}\ x\ 0.5594 + \frac{216}{500}\ x\ 0.5756 + \frac{29}{500}\ x\ 0.3616 + \frac{6}{500}\ x\ 0.917 + \frac{4}{500}\ x\ 0.812 = \mathbf{0.5599}$$

After the weight entropy is obtained, the next step is to calculate the gain value for gender attributes using the equation below.
$Gain_{work} = E_{Total} - E_{work} = 0.572 - 0.5599 = \mathbf{0.0121}$

After the gain value is obtained, the next step is to calculate the gain ratio value using the equation below.

$$Ratio = \frac{Gain_{work}}{E_{Farmer}+E_{PNS}+E_{Self-employed}+E_{Private\ Employes}+E_{far}} = \frac{0.0121}{0.5594+0.5756+0.3616+0.917+0.812} = \mathbf{0.00375}$$

d. Esktrakulikuler

Proportion to "Active" jobs.
Status "Active" : 232
Status "*dropout*" : 6
Total : 238
$Proporsi$ "Active" $= \frac{232}{238} = 0.9748$
$Proporsi$ "$Dropout$" $= \frac{6}{238} = 0.0252$

Total entropy for 'Active' extracurriculars.
$E_{aktif} = -(0.9748x\ log_2(0.9748) + (-(0.0252x\ log_2(0.0252)) = \mathbf{0.1697}$

Proportion for "inactive" extracurriculars.
Status "Active" : 200
Status "*dropout*" : 62
Total : 262
$Proporsi$ "Active" $= \frac{200}{262} = 0.7634$
$Proporsi$ "$Dropout$" $= \frac{62}{262} = 0.2366$

Total entropy for "inactive" extracurricular.
$E_{\text{tidak Active}} = -(0.7634xlog_2(0.7634) + (-(0.2366x\ log_2(0.2366)) = \mathbf{0.7888}$

After obtaining each entropy from the extracurricular attribute, the next step is to calculate the weight entropy value using the equation below.

$$E_{extracurricular} = \frac{238}{500}\ x\ 0.1697 + \frac{262}{500}\ x\ 0.7888 = \mathbf{0.4941}$$

After the weight entropy is obtained, the next is the calculation of the gain value for extracurricular attributes using the equation below.

$$Gain_{extracurricural} = E_{Total} - E_{ekstrakulikuler} = 0.572 - 0.4941 = \mathbf{0.0779}$$

After the gain value is obtained, the next step is to calculate the gain ratio value using the equation below.
$Ratio = \frac{Gain_{extracurricural}}{E_{active}+E_{inactive}} = \frac{0.0779}{0.1697+0.7888} = \mathbf{0.0812}$

After calculating the entropy value and gain value of all attributes, namely gender, parental occupation, distance from home to school and extracurricular participation, the results of the entire calculation can be seen in Table 2 below.

Table 2. C5.0 method calculation results for Node 1

No	Attribut		*Entropy*	*Gain*	*Ratio*
1	Gender	Male	0.8	0.013	0.0117
		Female	0.14		
2	Distance from home to school	near	0.58	0.0032	0.00188
		keep	0.6		
		far	0.6		
3	Parent's work	farmer	0.6	0.0121	0.00375
		Self employed	0.59		
		Civil servants	0.51		
		Private Employees	0.92		
		Not Working	0.76		
4	Extracurricular Participation	Active	0.25	0.0779	**0.0812**
		Inactive	0.8		

From Table 2 above, it can be seen that the highest entropy value is found in the extracurricular participation attribute, the attribute that has the highest entropy value is used as the root node. Branch nodes are taken from each Category on the Extracurricular Participation Variable. Since each branch node has its own sample in its own class, the tree calculation has not stopped and is still continuing on the following branch nodes

The calculation of entropy, gain, and gain ratio values is still continued to determine the branch at node 2. The equation for calculating the entropy, gain, and gain ratio values is still the same as the previous calculation of the root node. The data used to calculate node 2 is the data that is not used in the calculation of node 1. The attributes used are still the same as the calculation process of node 1, the results of the entropy, gain, and gain ratio calculations can be seen in the following Table 3.

Table 3. C5.0 calculation results for Node 2

No	Attribut		*Entropy*	*Gain*	*Ratio*
1	Gender	Male	0.8	0.013	**0.0117**
		Female	0.14		
2	Distance from home to school	Near	0.58	0.0032	0.00188
		Keep	0.6		
		Far	0.6		
3	Parents' Work	Farmer	0.6	0.0121	0.00375
		Self employed	0.59		
		Civil servants	0.51		
		Private Employees	0.92		
		Not Working	0.76		

From Table 2 above, it can be seen that the highest entropy value is found in the gender attribute. The attribute that has the highest entropy value is used as the root node, the branch node is taken from each Category in the gender variable. Since each branch node has its own sample in each class, the tree calculation has not stopped and is still continuing on the following branch nodes.

The calculation of the entropy, gain, and gain ratio values is still continued to determine the branch at node 3, The equation for calculating the entropy, gain, and gain ratio values is still the same as the calculation of the previous root node. The data used to calculate node 3 is the data that is not used in the calculation of nodes 1 and 2. The attributes used are still the same as the calculation process for node 1 and node 2. The results of the entropy, gain, and gain ratio calculations can be seen in the following Table 4.

Table 4. C5.0 calculation results for Node 3

No	Attribut		*Entropy*	*Gain*	*Ratio*
1	Distance from home to school	Near	0.58	0.0032	0.00188
		Keep	0.6		
		Far	0.6		
2	Distance from home to school	Farmer	0.6	0.0121	**0.00375**
		Self employed	0.59		
		Civil servants	0.51		
		Private Employees	0.92		
		Not Working	0.76		

From Table 3 above, it can be seen that the highest entropy value is found in the attributes of parents' work. The attribute that has the highest entropy value is used as the root node, the branch node is taken from each Category on the parent work variable. Since each branch node has its own sample in each class, the tree calculation has not stopped and is still continuing on the following branch nodes.

3.3 RapidMiner Testing

In this study, the test was carried out using RapidMiner software to analyze the factors that affect students' decision to dropout (DO) at SMA Negeri 6 Padangsidimpuan City. The method used is the C5.0 Decision tree algorithm, which is one of the machine learning algorithms that is very effective in classifying data based on existing attributes. RapidMiner, as a powerful data analysis tool, enables efficient data processing and the application of precise prediction models [12].

In the training process, RapidMiner uses the C5.0 algorithm to build a decision tree model based on training data. This algorithm generates a model with rules that can explain what factors have a major influence on a student's decision to dropout. The model was then tested using test data to see its accuracy and effectiveness in predicting students at risk of dropout. [13]

Steps to test RapidMiner

a. *Import* Data
   The data import stage is the process where the data is uploaded into the RapidMiner software. After the data is imported, select the columns on the data to determine the variables to be taken into account and the labels that are the reference as shown in figure 4.7 below. The gender, distance, parental occupation and extracurricular participation columns will be used as variables, the status column will be used as a label and the number and name columns will be ignored. After arranging the columns on the data, then drag them into the blank project. Drag data 2 times for the process of forming a decision tree and classification [14].

b. Operator *Split Data*
   The split data operator is an operator used to regulate the amount of training data and test data to be used in the model. [15]

c. Operator *Decision tree*
   On the blank process start page, add a decision tree operator [16]

d. Operator *Apply Model*
   The apply model operator is used to apply a pre-trained model to a new dataset. In other words, the operator will use the model that has "learned" from the training data to make predictions on data that the model has never seen. [16]

e. Operator Performance
   The performance operator is used to evaluate the performance of the classification model that has been built. In other words, this operator will calculate various metrics or performance measures to find out how well our model is at predicting classes or categories from new data. [17]

f. Connection
   After entering the data and all the necessary operators into the blank process, then connect all the

entities in the blank process [18].

After running the RapidMiner scheme as above, a decision tree will be obtained that describes the results of the system analysis of the data used based on the C5.0 algorithm. The image of the resulting decision tree can be seen in the previous figure 4.5 [19].

The results of the test show that the C5.0 algorithm is able to identify critical factors that play a role in students' decision not to continue their education. Using the decision tree model, patterns related to the likelihood of students dropping out can be found. Based on the results of the decision tree obtained, it can be seen that the extracurricular participation variable is the most determining factor for student dropout at SMA Negeri 6 Padangsidimpuan City, then followed by the gender variable in second place which can be seen in the decision tree [20].

3.4 Model Test Results

This section contains an explanation of the results of testing the model using the C5.0 algorithm. This test aims to determine the performance of the psychological disorder classification model that has been created. To test this, we use testing data with a total of 20% of the total data, namely 100 data. The test of this model uses the confusion matrix method. This confusion matrix method will produce accuracy, precision, recall and f1_score values from the calculation of True Positive (TP), False Positive (FP), True Negative (TN), False Negative (FN) values. The values of TP, FP, TN, and FN are obtained from the results of processing using a model that has been made on the test data.

The calculation of the model performance in the table above can be seen through the following equations.

a. Precision

$$\text{Precision} = \frac{TP}{(TP+FP)} x\ 100\ \% = \frac{85}{85+12} x\ 100\% = 87.63\%$$

b. Recall

$$\text{Recall} = \frac{TP}{(TP+FN)} x\ 100\ \% = \frac{85}{85+1} x\ 100\% = 98.84\%$$

c. F1-Score

$$\text{F1} - \text{Score} = \frac{2\ (Precision*Recall)}{Precision+Recall} x\ 100\ \% = \frac{2(87.63*98.84)}{87.63+98.84} x\ 100\% = 93\%$$

d. Accuracy

$$\text{Accuracy} = \frac{TP+FP}{TP+FP+TN+FN} x\ 100\ \% = \frac{85+2}{85+2+12+1} x\ 100\% = 87\%$$

By using 1048 student data obtained, the C5.0 algorithm can analyze and classify Active students and dropout students at SMA Negeri 6 Kota Padangsidimpuan by producing an accuracy of 87%.

4. Conclusion

Based on the research conducted regarding the analysis of dropout students at SMA Negeri 6 Padang sidimpuan City, the following conclusions can be obtained. Based on the results of a study that analyzed 4 variables regarding dropout cases at SMA Negeri 6 Padangsidimpuan City, it is known that the most influential variable is gender, then followed by the extracurricular participation variable in second place.

The C5.0 algorithm can be applied to classify and analyze the factors of dropout at SMA Negeri 6 Padangsidimpuan City by utilizing RapidMiner software. The C5.0 algorithm works effectively in generating patterns to identify the main causes of dropout at SMA Negeri 6 Kota Padangsidimpuan which can be proven based on the gain value obtained from each variable.

The results of the research carried out are still far from perfect. Therefore, development is needed to get more optimal results.

References

[1] A. Surip, M. Aji Pratama, I. Ali, A. Rinaldi Dikananda, and A. Irma Purnamasari, "Application of Machine Learning using C4.5 algorithm based on PSO in Analyzing Student Dropout Data," *INFORMATICS FOR EDUCATORS AND PROFESSIONALS*, vol. 5, no. 2, pp. 147–155, 2021, [Online]. Available: https://npd.kemdikbud.go.id/?

[2] L. Rajendra Haidar, E. Sediyono, A. Iriani, and J. O. Notohamidjojo Blotongan Sidorejo, "Prediction Analysis of Student Dropout Using Decision Tree Method with ID3 and C4.5 Algorithms,"

TRANSFORMATIKA, vol. 17, no. 2, pp. 97–106, 2020.

[3] Destiar A Maghfirah, "Data Mining in Analyzing Student Dropout Factors Using Decision Tree Algorithm," 2019.

[4] M. Azhari, H. Maulana, and F. Riza, "Data Mining in Analyzing Student Dropout Factors Using Decision Tree Algorithm," *JURNAL MEDIA INFORMATIKA BUDIDARMA*, vol. 8, no. 2, p. 1200, Apr. 2024, doi: 10.30865/mib.v8i2.7335.

[5] S. Syahdan, R. Rura, R. Dwi Christyanti, and J. Matematika, "Implementation of C5.0 Algorithm in Classifying Factors Causing Dengue Hemorrhagic Fever (DHF)," *Jurnal Sains Benuanta*, vol. 3, no. 1, pp. 1–7, 2024, doi: 10.61323/jsb.v3i1.105.

[6] Y. P. E. , & S. J. W. YUSMI NUR AINI, "Implementasi Decision tree Untuk Diagnosis Gangguan Kecemasan Umum," *Jurnal Informatika Dan Sistem Informasi*, vol. 2, no. 2, pp. 395–402, 2021.

[7] A. Rimadani, A. Pradjaningsih, I. Made Tirta, and U. Jember, "Classification of Dengue Fever Disease (DHF) Using C5.0 Algorithm Based on Binary Particle Swarm Optimization (BPSO)," *Jurnal Ilmiah Matematika*, vol. 9, no. 2, pp. 55–66, 2022, doi: 10.26555/konvergensi.v9i2.26089.

[8] F. Puspitaningrum, A. Haryoko, A. A. Suryanto, and A. Saputri, "Implementation of data mining to predict student graduation rates using the C5.0 decision tree algorithm.," *Curtina*, vol. 1, no. 1, pp. 31–39, Dec. 2020, doi: 10.55719/curtina.v1i1.181

[9] V. R. Prasetyo, H. Lazuardi, A. A. Mulyono, and C. Lauw, "Application of RapidMiner Application for Prediction of the Rupiah Exchange Rate Against the US Dollar Using the Linear Regression Method, National Journal of Technology and Information Systems, vol. 7, no. 1, pp. 8–17, May 2021, doi: 10.25077/teknosi.v7i1.2021.8-17.

[10] A. H. Nurdy, A. Rahim, and Arbansyah, "Stumble Guys Game Review Sentiment Analysis on Playstore Using Naïve Bayes' Algorithm," *Teknika*, vol. 13, no. 3, pp. 388–395, Sep. 2024, doi: 10.34148/teknika.v13i3.993

[11] S. Devi Asri and ainul Miftahul Huda, "Implementation of C5.0 Algorithm on Social Community Data Classification (Case Study: Eligibility for Direct Cash Assistance in Condong Village, Singkawang City)," 2023.

[12] Fahrullah, Y. Bintan, M. Ari Prayogo, and S. Artikel, "DETERMINATION OF PRIORITIES CRITERIA FOR ASSESSING THE SUCCESS OF FIELDWORK PRACTICE USING THE DECISION TREE METHOD INFO ABSTRACT ARTICLE", doi: 10.58290/jukomtek.v2i.

[13] M. E. Hadi, D. Arifianto, and Q. A'yun, "Classification of Learning Styles Using C5.0 Algorithm, 2023. [Online]. Available: http://jurnal.unmuhjember.ac.id/index.php/JST

[14] M. Zeno, L. Putra, and A. P. Wibowo, "IMPLEMENTATION OF APRIORI ALGORITHM FOR PURCHASE PATTERN ANALYSIS", [Online]. Available: http://journalbalitbangdalampung.org

[15] Apriyadi, M. Ridwan Lubis, B. Efendi Damanik, S. Informasi, and S. Tunas Bangsa Pematangsiantar, "Implementation of C5.0 Algorithm in Determining Student Comprehension Level of Online Learning," *KOMPUTA : Jurnal Ilmiah Komputer dan Informatika*, vol. 11, no. 1, 2022.

[16] N. Debi and A. Informatika, "Implementation of Decision Tree Algorithm in Classifying Korean Dramas."

[17] K. M. Dzatul, U. F. Rakhmat, and H. Ashaury, "Unemployment Prediction Using C5.0 Decision Tree Algorithm on Caringin District, Bogor Regency Population Data," 2022. [Online]. Available: https://e-journal.unper.ac.id/index.php/informatics

[18] A. U. Narestami, D. Suhartono, and T. Tarwoto, "Implementation of K-Means Algorithm to Determine Marketing Strategy," *JURNAL MEDIA INFORMATIKA BUDIDARMA*, vol. 8, no. 2, p. 1200, Apr. 2024, doi: 10.30865/mib.v8i2.7335.

[19] R. Pratiwi, M. N. Hayati, and S. Prangga, "BAREKENG: Jurnal Ilmu Matematika dan Terapan

COMPARISON OF C5.0 ALGORITHM CLASSIFICATION WITH CLASSIFICATION AND REGRESSION TREE (CASE STUDY: SOCIAL DATA OF FAMILY HEADS OF TELUK BARU VILLAGE, MUARA ANCALONG DISTRICT IN 2019) Comparison of C5.0 Algorithm Classification with Classification and Regression Tree (Case Study: Social Data of Family Head of Teluk Baru Village, Muara Ancalong District in 2019)," vol. 14, no. 2, 2020, doi: 10.30598/barekengvol14iss2pp267-278.

[20] P. B. N. Setio, D. R. S. Saputro, and B. Winarno, "PRISMA, Proceedings of the National Seminar on Classification Mathematics with Decision Trees Based on C4.5 Algorithm," vol. 3, pp. 64–71, 2020, [Online]. Available: https://journal.unnes.ac.id/sju/index.php/prisma/