
Implementation Of The Data Mining Cart Algorithm In The Characteristic Pattern Of New Student Admissions

Ahmad Syahban Rifandy Siregar^{1)*}, Yunita Sari Siregar²⁾, Mufida Khairani³⁾

¹⁾²⁾³⁾Universitas Harapan Medan, Indonesia

¹⁾ahmadsyahbanrifandysiregar2311@gmail.com, ²⁾yunitasarisiregar1990@gmail.com, ³⁾mufidakhairani@gmail.com

ABSTRACT

University of Harapan Medan is one of the private universities in North Sumatra which has an Informatics Engineering Study Program with Good Accreditation. With better accreditation, the number of students who register is also increasing. At the admission of new students, the committee has a huge pile of data, making it difficult in the process of whether the student passed or did not pass. Therefore, in this study, we will implement data mining with the CART (Classification And Regression Tree) algorithm. Data mining is a technique to determine the characteristic pattern of a variable or data criteria with a large amount. In the CART method, the data is first converted into testing data, which will then be used to form a classification tree by calculating the value of information gain, Gini index and goodness of split. From the results obtained, it will be re-determined terminal nodes, marking class labels and finally pruning the classification tree which produces a decision tree. In this study, the number of testing data was 75 with 3 criteria, namely the average value of report cards, CAT test scores, and interview scores. The results of testing data testing using RapisMiner 5.3 software produce 23 number of characteristic pattern rules, where node 1 is the CAT test score, level 1 branch node is the interview score criteria and level 2 branch node is the average report card value.

Keywords: Data Mining, CART (Classification And Regression Tree), Pattern, Student, Informatics Engineering Study Program

1. INTRODUCTION

The Informatics Engineering study program is one of the study programs at Harapan University Medan with B accreditation, where there are 1,366 active students until 2022. With a large number of students enrolled in the study program, it can be said that this study program has many enthusiasts. One of the quality assessments of a tertiary institution can be seen from the percentage of alumni in a tertiary institution. The more alumni and students who enter, the higher the quality of the tertiary institution. In order to produce and create quality, intellectual and ethical human resources, the university conducts a selection process in admitting new students. The large number of incoming new student data creates piles of files, making it difficult for the new student admissions committee to process the data and determine whether students have passed or not entered the study program. Therefore, to solve the problem of new student admissions, the CART (Classification And Regression Tree) algorithm data mining method can be used.

Data mining is a series of processes to explore the added value of a set of data in the form of knowledge that has not been known manually. Data mining actually has long roots in fields of science such as artificial intelligence, machine learning, statistics and databases. Some of the techniques that are often mentioned in the Data Mining literature include clustering, classification, association rule mining, neural network, and genetic algorithms (Siregar et al., 2021). Data mining is mining or discovering new information by looking for certain patterns or rules from a very large amount of data. Classification and Regression Tree (CART) is one of the methods or algorithms of the decision tree technique. CART is a nonparametric statistical method that can describe the relationship between the response variable (the dependent variable) and one or more predictor variables (the independent variable). If the response variable is continuous, the method used is the regression tree method, whereas if the response variable has a categorical scale, the method used is the classification tree method (Prasetya, 2020).

* Corresponding author



This is an Creative Commons License This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International (CC BY-SA 4.0).

2. LITERATURE REVIEW

Knowledge Discovery in Databases (KDD) is a set of processes for finding useful knowledge from data. KDD consists of a series of change steps, including data preprocessing as well as post processing. Data preprocessing is a step to convert raw data into a format suitable for the next stage of analysis. In addition, data preprocessing is also used to assist in the recognition of attributes and data segments that are relevant to data mining tasks (Karsito & Monika Sari, 2018). Knowledge Discovery In Database (KDD) is an activity that includes collecting, using historical data to find regularities, patterns or relationships in large data and relationships with integration techniques and scientific discoveries, interpretation and visualization of patterns in a number of data sets (Elisa, 2017). Knowledge discovery as a process consists of data cleaning, data integration, data selection, data transformation, data mining, pattern evaluation and knowledge presentation). Data mining refers to the process of mining knowledge from a very large set of data (Firdaus, 2017).

Data mining is the process of finding interesting patterns and knowledge from large data. Data sources include databases, data warehouses, the Web, other information repositories, or data that flows into systems dynamically. Data mining is a term used to describe the discovery of knowledge in databases (Prabawati et al., 2019). Data mining is a process that uses statistical, mathematical, artificial intelligence, and machine learning techniques to extract and identify large amounts of information (Siregar & Harliana, 2018a). Data mining is an iterative and interactive process to find new patterns or models that are perfect, useful and understandable in a massive database. Data mining contains a search for trends or patterns desired in a large database to help decision makers in the future, these patterns are recognized by certain devices that can provide a useful and insightful analysis of data that can then be studied more thoroughly, which may use other decision support tools (Sikumbang, 2018).

Characteristics of data mining as follows (Siregar & Harliana, 2018b):

1. Data mining relates to the discovery of something hidden and certain data patterns that were not known before.
2. Data mining usually uses very large data. Usually big data is used to make the results more believable.
3. Data mining is useful for making critical decisions, especially in strategy.

Schematically dividing the steps of implementing data mining into three activities, namely (Asparizal et al., 2016):

1. Data exploration, consisting of data cleaning activities, data transformation, dimension reduction, feature selection, and others
2. Creating a model and testing the validity of the model, is the maintenance of the models that have been developed that match the case at hand. In other words, competitive model selection is carried out.
3. Application of models with new data to produce estimates of existing cases. This stage is the stage that determines whether the model that has been built can answer the problems faced.

Modeling is the use of certain principles or techniques in a system design. The more complex the problems encountered today make the modeling process even more complex. Therefore, knowledge that is also more specific and detailed is needed. The stages are as follows (Asparizal et al., 2016):

1. Identification. This is the first stage in data mining modeling of a problem in the field. In identifying a problem, there are two contradictory approaches. The first approach is an approach that prioritizes prior knowledge of a case. In this case a priori knowledge is the mainstay of these supporters. The second approach is purely data-based identification. As far as possible, pre-judgment of a condition is avoided
2. Estimation and Matching. After the selection stage is complete, the next step is to make a numerical formulation of a model. This stage is known as the stage of matching the model with the data. While the conversion of the model into a numerical figure is called estimation
3. Testing. Testing is the final stage before the system is implemented. The system has been tested against other data that has never been owned and is not the data used to form the model. The success of a test depends on the output produced by a system being tested, whether it is in accordance with the existing reality or not.
4. Practical Application. The designed system is intended to solve existing problems in the field.

* Corresponding author



This is an Creative Commons License This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International (CC BY-SA 4.0).

5. Iteration. Iteration requires the designer to always think again about the model he makes. With iterations, it is hoped that a model that is robust and suitable for the situations and conditions that occur during implementation is obtained.

Data mining techniques are divided into several techniques. Some of the techniques and properties of data mining are as follows:

1. Classification Determines a new data record of one of several pre-defined categories (or classes).
2. Regression predicts the value of a given continuous variable based on the value of another variable, assuming a linear or non-linear dependency model. This technique is widely studied in statistics, the field of artificial neural networks.
3. Clustering partitions data sets into sub-sets or groups in such a way that elements of a particular group have a set of shared properties, with a high degree of similarity in one group and a low level of similarity between groups. Also called unsupervised learning.
4. Association Rules detect a set of attributes that appear together (co-occur) in frequent frequencies and form a number of rules of those sets. e. Sequential (Sequence Mining) Looking for a number of events that generally occur together (Siregar et al., 2021).

Classification is the process of finding a model or function that describes and differentiates data classes or concepts. Data classification has a two-step process consisting of a learning step where a classification model is built and a classification step where the model is used to predict the class label of the given data (Prabawati et al., 2019). Classification is the process of learning an objective function (target) f which maps each set of attributes x to one of the previously defined class labels y . The target function is also called a classification model. The model in classification has the same meaning as a black box, where a model receives input, then thinks about the input and provides answers as the output of the results of its thinking (Irmayani, 2020).

The classification process has four components (Prabawati et al., 2019):

1. Class. Class is a dependent variable that represents the label contained in the object.
2. Predictors. Predictors are independent variables represented by data attributes
3. Training Datasets. Training Dataset is a data set used to determine a suitable class based on a predictor
4. Testing Datasets. Testing Dataset is a new data set that will be classified by the model that has been created

A decision tree is a structure that can be used to divide large data sets into smaller record sets by applying a set of decision rules. With each set of divisors, the members of the result set become similar to one another. The data in a decision tree is usually expressed in the form of a table with attributes and records. Attributes declare a parameter that is made as a criterion in forming a tree. One of the attributes is an attribute that represents the solution data per data item called the target attribute. Attributes have values called instances. The process in the decision tree is to change the shape of the data into a tree model, change the tree model into a rule, and simplify the rule (Mardi, 2019). In the decision tree there are 3 types of nodes, namely:

1. Root Node, is the initial node that has no input and can have no output or have more than one output.
2. Internal Node, is a branching node. This node has only one input and has at least two outputs..
3. Leaf node or terminal node, is the final node, at this node there is only one input and has no output (Prabawati et al., 2019)

CART (Classification and Regression Trees) is a method or algorithm of one of the data exploration techniques, namely the decision tree technique. CART is a simple but powerful method. CART aims to obtain a group of data that is accurate as an identifier of a classification, besides that CART is used to describe the relationship between the response variable (dependent or dependent variable) with one or more predictor variables (independent or independent variables). The resulting tree model depends on the scale of the response variable, if the data response variable is continuous then the resulting tree model is a regression tree, whereas if the response variable has a categorical scale then the resulting classification trees are classification trees (Pratiwi & Zain, 2014).

CART will produce a classification tree if the response variable has a categorical scale and will produce a regression tree if the response variable is continuous data. The main objective of CART is to obtain an accurate group of data as an identifier for a classifier. The CART algorithm goes through three stages, namely (Sumartini & Purnami, 2015) :

1. Formation of a classification tree. This stage begins with determining the variables and thresholds to be used as separators for each node. The stages of forming a classification tree consist of:

* Corresponding author



This is an Creative Commons License This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International (CC BY-SA 4.0).

- a. **Sorting** Sorting. The data used is a sample of learning data. The subset resulting from the sorting process must be more homogeneous than the previous sorting. The heterogeneity function used is the Gini Index because it will always separate classes with the largest/most important class members in the node first. The Gini Index function is shown in equation 1.

$$i(t) = \sum_{i,j=1} p(j|t) p(i|t), i \neq j \tag{1}$$

Explanation :

- $i(t)$: Gini index at node t
- $p(j|t)$: Proportion of class j at node t
- $p(i|t)$: The proportion of class i at the t node.

The selected sorting will form a set of classes called nodes. The node will perform recursive sorting until terminal nodes are obtained. The next stage is to determine the goodness of split criteria to evaluate the splitter of the splitters at node t with equation 2.

$$\phi(s,t) = \Delta i(s,t) = i(t) - P_L i(t_L) - P_R i(t_R) \tag{2}$$

Explanation :

- t_L : Left branch of decision node t
- t_R : The right branch of the decision node t
- P_L : Proposition of the number of objects that enter the t_L
- P_R : Proposition of the number of objects that enter the t_R

The separator that produces higher $\phi(s,t)$ is the best separator because it is able to reduce heterogeneity more highly. The probability of the number of objects entering t_L in equations 2 and t_R can be shown in equations 3 and 4

$$P_L = \frac{\text{Candidate left node}}{\text{Exercise data}} \tag{3}$$

$$P_R = \frac{\text{Candidate right node}}{\text{Exercise data}} \tag{4}$$

- b. **Terminal Node Determination**. The development of the tree will stop if at the node there are observations totaling less than or equal to 5 ($n \leq 5$). In addition, the tree formation process will also stop when it reaches the limit on the number of predetermined levels or the maximum depth level in the tree. The process of splitting the tree at node t into t_R and t_L applies equation 5.

$$R(t) > R(t_R) + R(t_L) \tag{5}$$

Explanation :

- $R(t)$: Classification error at node t
- $R(t_R)$: Classification error on the left node of node t
- $R(t_L)$: Classification error on the right node of node t

- c. **Class Label Marking**. Determination of the class label at the terminal node is based on the highest number rule, which can be seen in equation 6

$$p(j_0|t) = \max_j \frac{N_j(t)}{N(t)} \tag{6}$$

Explanation :

- $p(j_0|t)$: The proportion of the number of classes at node t
- $N_j(t)$: The number of class j observations at node t
- $N(t)$: Number of observations at node t

The class label for the t terminal node is j_0 which gives the value of the smallest suspected misclassification at the t node $r(t) = 1 - \max_j p(j|t)$.

2. **Classification tree pruning**. The size of the tree formed by the sorting rules and goodness of split criteria is very large because the termination of the tree is based on the number of observations at the terminal nodes or the degree of homogeneity. The large size of the tree can lead to overfitting, but if tree observations are limited to certain limits, underfitting can occur. a decent tree size can be done by pruning the tree with a minimum cost complexity size which can be shown in equation 7.

$$R_\alpha(t) = R(t) + \alpha |\tilde{r}| \tag{7}$$

* Corresponding author



$R_{\alpha}(t)$ is a linear combination of cost and tree complexity formed by adding the cost penalty of complexity to the cost of misclassifying the tree. Next, a search for the part tree $T(\alpha) < T_{max}$ which minimizes $R_{\alpha}(t)$ can be shown in equation 8.

$$R_{\alpha}(t(\alpha)) = \min_{T < T_{max}} R_{\alpha}(t) \tag{8}$$

3. Optimum classification tree determination. A replacement estimator that is often used when the observations are not large enough is the Cross Validation Estimate. Observations in L are divided randomly into V separate parts of approximately the same size for each class. The $T(v)$ tree is formed from the v th learning samples with $v=1,2,\dots,V$. suppose $d^{(v)}(x)$ is the result of classification, then the test sample estimator for $R(T_i^{(v)})$ can be shown in equation 9.

$$R(T_i^{(v)}) = \frac{1}{N_2} \sum_{X_n, J_n \in L_v} X(d(v))(X_n) \neq J_n \tag{9}$$

Explanation :

$N_v \cong N/V$: Number of observations in L_v .

3. METHOD

In the analysis and design of this system will explain the framework in research used in solving thesis problems. The stages in the research framework are problem analysis and identification, literature study, data collection, implementation, analysis and model design, data testing, results and discussion. At this stage it also uses a flowchart model design to describe system data flows, so that it can help user understanding. This framework is the stages that will be carried out in solving the thesis problem. The stages in the research framework can be seen in Figure 1 below.

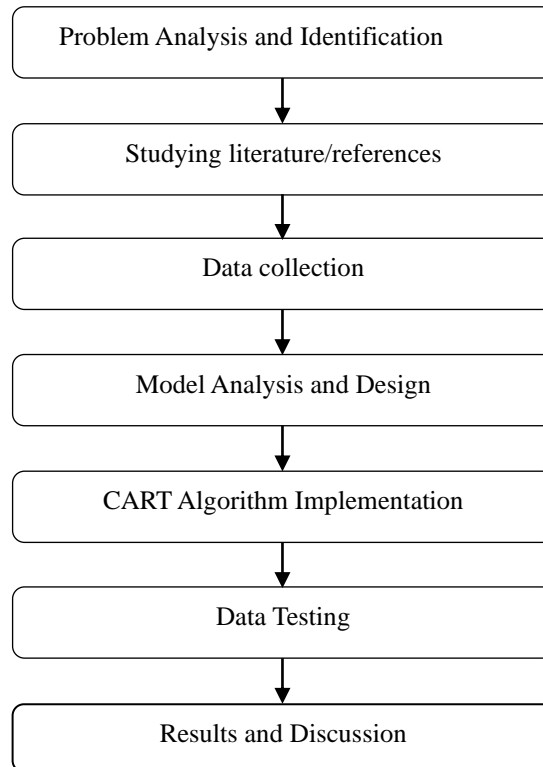


Figure 1. Research Framework

Based on figure 1, there are several explanations, including:

1. Problem Analysis and Identification. At this stage, the problem will be formulated, by determining the boundaries of the problem, as well as the scope that is the object of research, so as not to deviate from the research problem

* Corresponding author



This is an Creative Commons License This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International (CC BY-SA 4.0).

2. Studying Literature/References. This stage is carried out to complement the concepts and theories derived from journals that discuss data mining, decision trees, classification, CART (Classification And Regression Tree) algorithms, criteria / variables in the admission of new students.
3. Data Collection. The data used in this study came from the Informatics Engineering Study Program, University of Harapan Medan.
4. Analysis and Design. he analysis stage includes determining the decision model to be used, the inputs needed, and the expected outputs
5. Implementation of the CART (Classification And Regression Tree) Algorithm. At this stage will be described the hardware and software used in the study.
6. Data testing. After the process of analyzing and designing models with data mining methods or techniques is carried out, the next step is to test the data manually with the CART algorithm and test the data with the CART algorithm to the system using the RapidMiner 5.3 software.
7. Results and Discussion. At the stage of test results aims to find out whether the previous stage of testing has provided a solution to the problem as desired

4. RESULT

The data used in determining the characteristic pattern of new student admissions at the Informatics Engineering Study Program, Harapan University, Medan, totaled 75 samples. The criteria used in this study consisted of: average report card scores, CAT (Computer Assisted Test) exam scores, and interview scores. The data obtained will then be analyzed and grouped into several groups of data which will be processed by designing a decision tree. In the process of determining the pattern of characteristics of new student admissions, it will produce decision trees and rules by utilizing the data mining algorithm CART (Classification And Regreaaion Tree). The weight of the values of each criterion used in the study can be seen in table 1.

Table 1
Weight of Criterion Values

Average Score Report Card	Weight Value
100 – 81	Very High
80 -71	High
70 – 61	Medium
60 – 51	Low
< 50	Very Low
CAT Exam Scores	Weight Value
100 – 81	Very High
80 -71	High
70 – 61	Medium
60 – 51	Low
< 50	Very Low
Average Score Report Card	Weight Value
> 71	High
70 - 51	Medium
< 50	Low

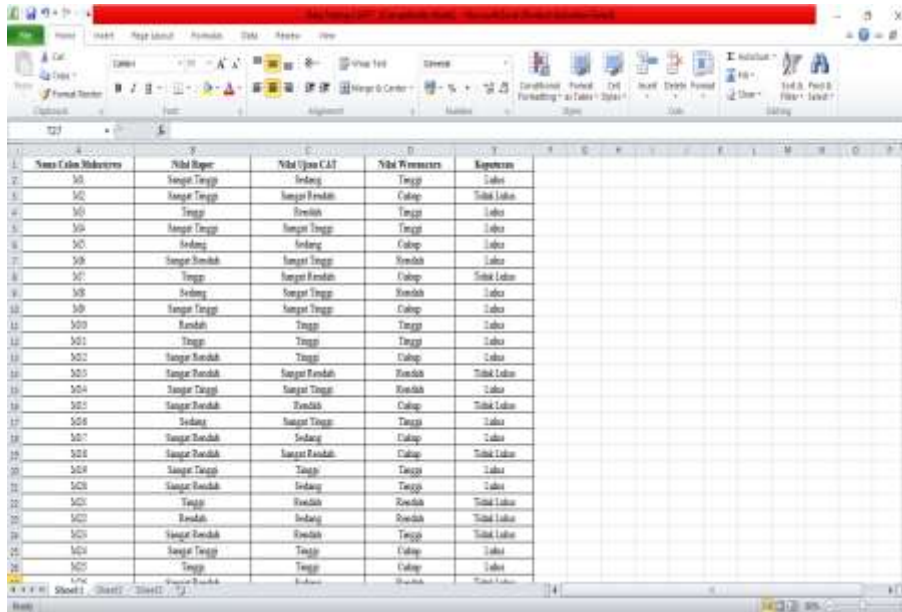
* Corresponding author



This is an Creative Commons License This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International (CC BY-SA 4.0).

The stages of implementation using RapidMiner 5.3 software are as follows:

1. All variables consisting of attributes (average report card scores, CAT test scores, interview scores) and their classification (very high, high, medium, low, very low) along with the decision results (pass and fail) used in determining The characteristic pattern of new student admissions for the Informatics Engineering Study Program, Harapan University, Medan, which amounts to 75 testing data, is stored in excel format (.xls) with the data file name of the candidate Data Testing CART.xls. Can be seen in figure 2.



Nama Calon Mahasiswa	Nilai Rapor	Nilai Ujian CAT	Nilai Wawancara	Keputusan
S01	Sangat Tinggi	Indah	Tinggi	Lulus
S02	Sangat Tinggi	Sangat Baik	Cukup	Tidak Lulus
S03	Tinggi	Baik	Tinggi	Lulus
S04	Sangat Tinggi	Sangat Tinggi	Baik	Lulus
S05	Indah	Indah	Cukup	Lulus
S06	Sangat Baik	Sangat Tinggi	Baik	Lulus
S07	Tinggi	Sangat Baik	Cukup	Tidak Lulus
S08	Indah	Sangat Tinggi	Baik	Lulus
S09	Sangat Tinggi	Sangat Tinggi	Cukup	Lulus
S10	Baik	Tinggi	Tinggi	Lulus
S11	Tinggi	Tinggi	Tinggi	Lulus
S12	Sangat Baik	Tinggi	Cukup	Lulus
S13	Sangat Baik	Sangat Baik	Baik	Tidak Lulus
S14	Sangat Tinggi	Sangat Tinggi	Baik	Lulus
S15	Sangat Baik	Baik	Cukup	Tidak Lulus
S16	Indah	Sangat Tinggi	Tinggi	Lulus
S17	Sangat Baik	Indah	Cukup	Lulus
S18	Sangat Baik	Sangat Baik	Cukup	Tidak Lulus
S19	Sangat Tinggi	Tinggi	Tinggi	Lulus
S20	Sangat Baik	Indah	Baik	Lulus
S21	Sangat Baik	Sangat Baik	Cukup	Tidak Lulus
S22	Sangat Tinggi	Tinggi	Tinggi	Lulus
S23	Sangat Baik	Indah	Baik	Lulus
S24	Sangat Baik	Baik	Cukup	Tidak Lulus
S25	Sangat Tinggi	Sangat Tinggi	Cukup	Lulus
S26	Tinggi	Tinggi	Cukup	Lulus
S27	Sangat Baik	Baik	Baik	Tidak Lulus

Figure 2. Data Testing CART.xls

2. After entering into RapidMiner 5.3, the next step is to select the New Process menu. In Parameters Read Excel there will be commands to enter and select the excel file to use. then a display like figure 3 will appear.

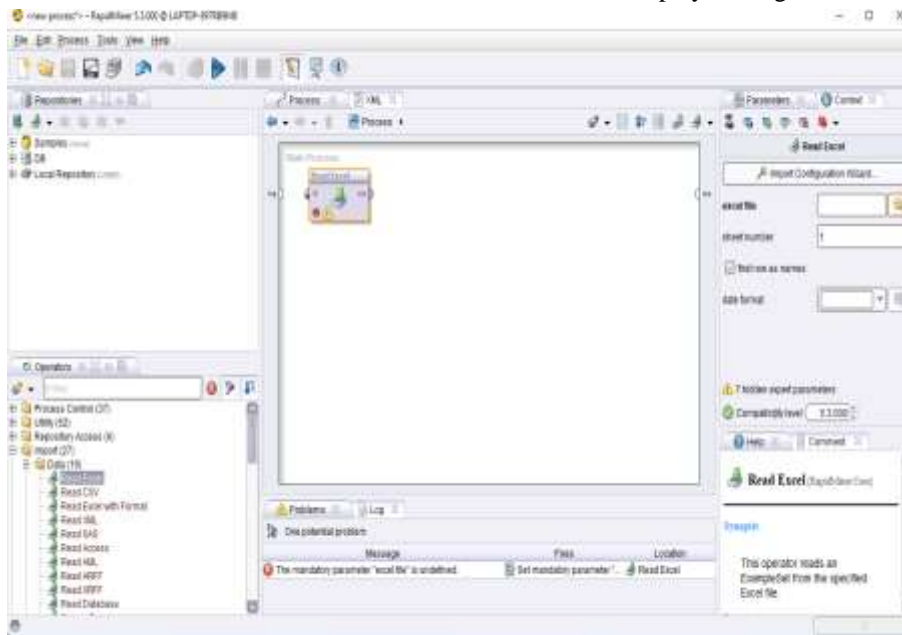


Figure 3. Excel File Selection

* Corresponding author



This is an Creative Commons License This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International (CC BY-SA 4.0).

- Next a display will appear for open file, then select the excel file that will be used to make the decision tree and click open. Select the folder with the thesis name + select the excel file with the name Data Testing CART.xls Can be seen in Figure 4.



Figure 4. Excel File Input

- After that, in the Parameters view, click Import Configuration Wizard. In the Import Configuration Wizard there will be 4 steps that must be passed. Next, the step 1 display will appear. On the step 1 display, we will select a file with the name Data Testing CART.xls. Then click the Next button. After that, the step 2 display will appear. The step 2 display will select which sheet in the excel file to use for testing. Then click the Next button to proceed to step 3. Can be seen in Figure 5.

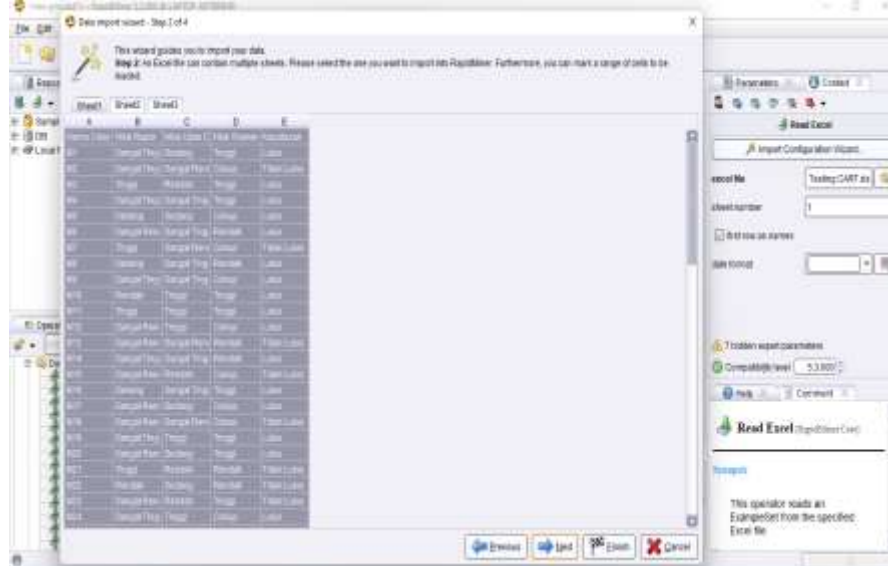


Figure 5. Display Step 2

- After that the step 3 display will appear. Where in step 3 the contents of sheet 1 will be shown again in the excel file selected in step 2. If the selected file is correct then click the Next button to continue step 4. Can be seen in figure 6

* Corresponding author



This is an Creative Commons License This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International (CC BY-SA 4.0).

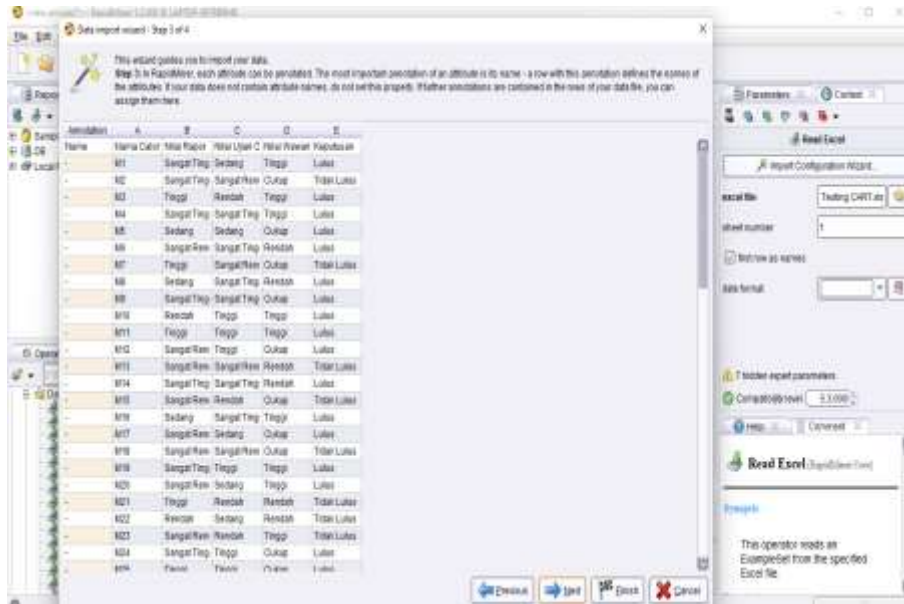


Figure 6. Display Step 3

6. Next, the display of the last step will appear, namely step 4. Step 4 is the selection of which criteria are used as the target attribute and attribute and the type of value to be used in the target attribute and attribute. There are 2 types of attributes, namely binominal (the classification consists of only 2 types of data) and polynomial (the classification consists of more than 2 types of data). For the prospective student's name column, it will be used as an attribute and the value type is polynomial. In the average column, the report card value will be used as an attribute with a polynomial value type (very high, high, medium, low, very low). The CAT (Computer Assisted Test) test scores column will be used as an attribute with a polynomial value type (very high, high, medium, low, very low). In the interview value column, it will be used as an attribute with a polynomial value type (high, medium, low). And in the Decision Result column, it will be used as a target attribute with conditions as a label with a binominal value type (passed, failed). After making the selection, click the Finish button. Can be seen in figure 7.

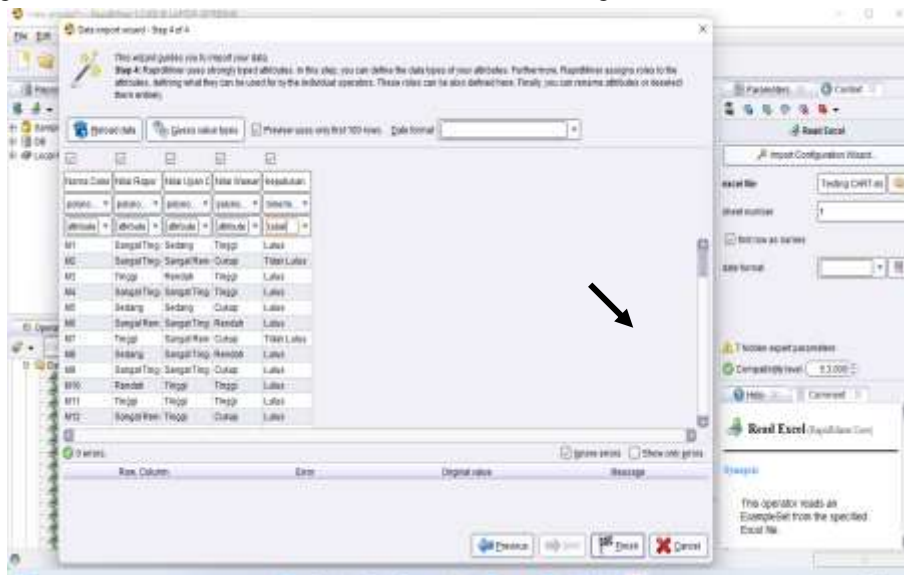


Figure 7. Display of Step 4

* Corresponding author



This is an Creative Commons License This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International (CC BY-SA 4.0).

- Next, in the Process view, connect the Read excel output (out) to the Decision Tree transformation (tra), then connect the Decision Tree model (mod) to the Result Process (res). After all processes between Read Excel and Decision Tree are connected, then click the Run button. In the decision treecriterion menu select gini_index. to Will appear like the picture 8.

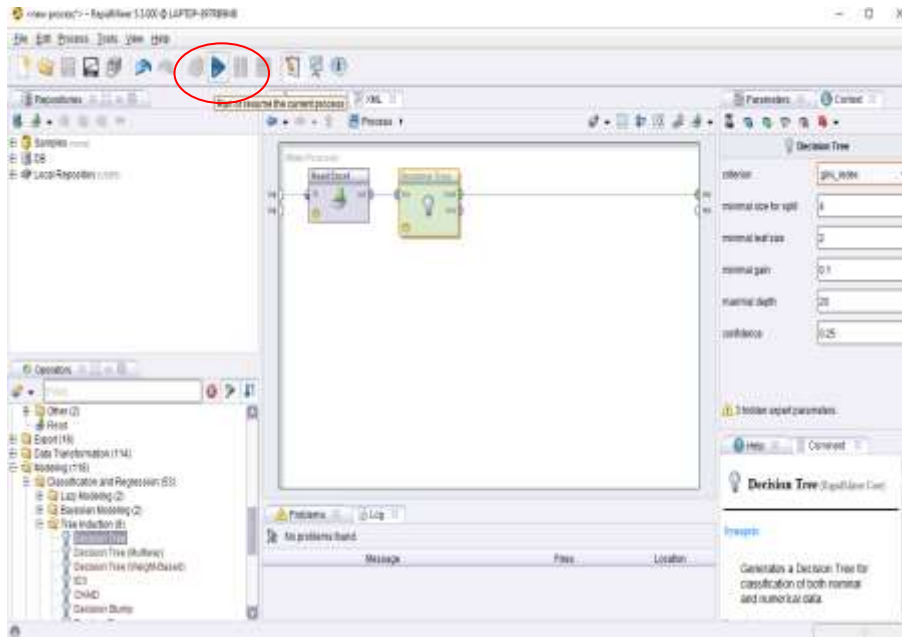


Figure 8. Run Process

- The results of the Run will appear. Run results that can be seen are Graph View and Text View. In the Graph View item, the final decision tree results will appear based on the data that has been selected in the application test. Can be seen in figure 9.

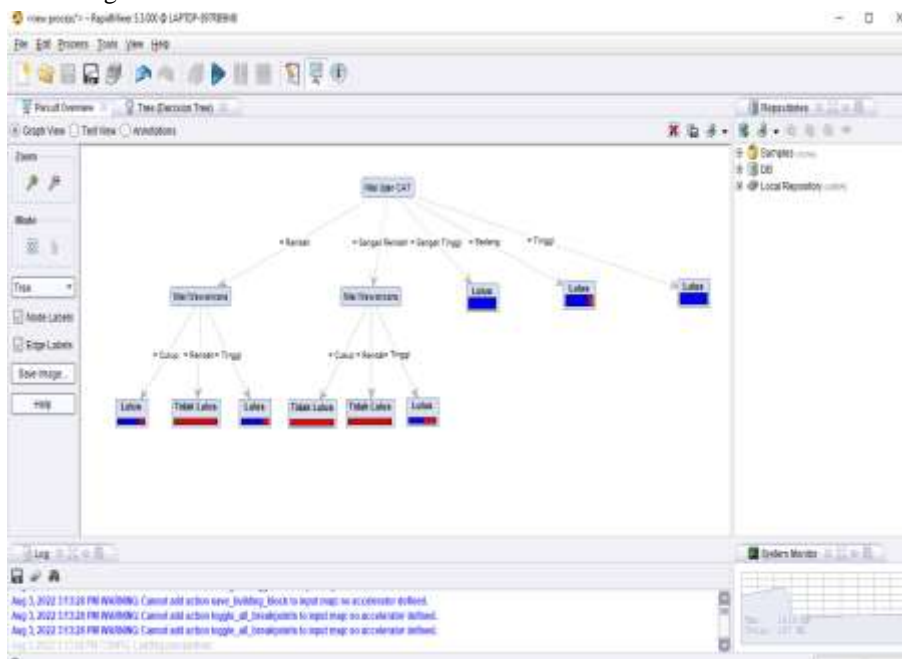


Figure 9. Graph View results

* Corresponding author



This is an Creative Commons License This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International (CC BY-SA 4.0).

9. In the Text View the results will appear in the form of rules from the decision tree. Can be seen in figure 10.

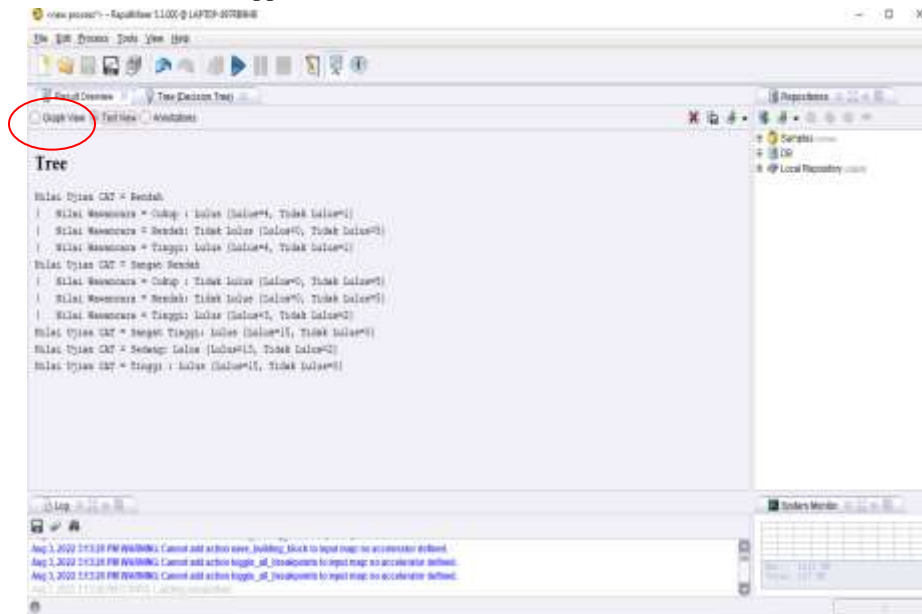


Figure 10. Text View results

Based on the results of testing the data mining algorithm CART (Classification And Regression Tree) which has been carried out by testing manual calculations and testing using RapidMiner 5.3 software, it can be concluded that the test results in the form of a decision tree and rules resulting from the two tests are very good. The rules generated using RapidMiner 5.3 software can be seen as follows:

1. If CAT Exam Score = Low, AND Interview Score = Medium, AND Average Report Card Score = Very High THEN Decision = Pass.
2. If CAT Exam Score = Low, AND Interview Score = Medium, AND Average Report Card Score = High THEN Decision = Pass.
3. If CAT Exam Score = Low, AND Interview Score = Medium, AND Average Report Card Score = Medium THEN Decision = Pass.
4. If CAT Exam Score = Low, AND Interview Score = Medium, AND Average Report Card Score = Low THEN Decision = Pass.
5. If CAT Test Score = Low, AND Interview Score = Medium, AND Average Report Card Score = Very Low THEN Decision = Failed.
6. If CAT Exam Score = Low, AND Interview Score = Low THEN Decision = Failed.
7. If CAT Test Score = Low, AND Interview Score = High, AND Average Report Card Score = Very High THEN Decision = Pass.
8. If CAT Test Score = Low, AND Interview Score = High, AND Average Report Card Score = High THEN Decision = Pass.
9. If CAT Test Score = Low, AND Interview Score = High, AND Average Report Card Score = Medium THEN Decision = Pass.
10. If CAT Test Score = Low, AND Interview Score = High, AND Average Report Card Score = Low THEN Decision = Pass.
11. If CAT Test Score = Low, AND Interview Score = High, AND Average Report Card Score = Very Low THEN Decision = Failed.
12. If CAT Examination Score = Very Low, AND Interview Score = Medium THEN Decision = Failed.
13. If CAT Test Score = Very Low, AND Interview Score = Low THEN Decision = Failed.
14. If CAT Test Score = Very Low, AND Interview Score = High, AND Average Report Card Score = Very High THEN Decision = Pass.

* Corresponding author



This is an Creative Commons License This work is licensed under a
Creative Commons Attribution-ShareAlike 4.0 International (CC BY-SA
4.0).

15. If CAT Test Score = Very Low, AND Interview Score = High, AND Average Report Card Score = High THEN Decision = Pass.
16. If CAT Test Score = Very Low, AND Interview Score = High, AND Average Report Card Score = Medium THEN Decision = Pass.
17. If CAT Test Score = Very Low, AND Interview Score = High, AND Average Report Card Score = Low THEN Decision = Failed.
18. If CAT Test Score = Very Low, AND Interview Score = High, AND Average Report Card Score = Very Low THEN Decision = Failed.
19. If CAT Test Score = Very High THEN Decision = Pass.
20. If CAT Test Score = Medium THEN Decision = Pass.
21. If CAT Test Score = Medium, AND Interview Score = Low, AND Average Report Card Score = Low THEN Decision = Failed.
22. If CAT Test Score = Medium, AND Interview Score = Low, AND Average Report Card Score = Very Low THEN Decision = Failed.
23. . If CAT Exam Score = High THEN Decision = Pass.

DISCUSSIONS

Based on calculations with the CART algorithm data mining method can be explained as follows:

1. The average (Very High) Report Card Score (High, Medium, Low, Very Low) obtained a goodness of split value of -0.08 at level 10 which is a non-terminal node.
2. The average (high) report card score (very high, medium, low, very low) obtained a goodness of split score of -0.08 at level 11 which is a non-terminal node.
3. Average Report Card Score (Medium) (Very High, High, Low, Very Low) obtained a goodness of split value of -0.08 at level 12 which is a non-terminal node.
4. The average (low) report card score (very high, high, medium, very low) obtained a goodness of split value of -0.0818 at level 13 which is a non-terminal node.
5. The average (Very Low) Report Card Score (Very High, High, Medium, Low) obtained a goodness of split value of -0.0658 at level 7 which is a non-terminal node.
6. CAT test scores (Very High) (High, Medium, Low, Very Low) obtained a goodness of split score of -0.044 at level 5 which is a non-terminal node.
7. CAT Test Scores (High) (Very High, Moderate, Low Very Low,) obtained a goodness of split score of -0.044 at level 6 which is a non-terminal node.
8. CAT Test Scores (Medium) (Very High, High, Low, Very Low,) obtained a goodness of split score of -0.0724 at level 9 which is a non-terminal node.
9. The value of the CAT Test (Low) (Very High, High, Low, Very Low,) obtained a goodness of split score of -0.0658 at level 8 which is a non-terminal node.
10. CAT Test Scores (Very Low) (Very High, High, Medium, Low,) obtained a goodness of split score of 0.052 at level 3 which is the terminal node.
11. Interview scores (High) (Medium, Low) obtained a goodness of split value of 0.0668 at level 2 which is the terminal node.
12. Interview Scores (Medium) (High, Low,) obtained a goodness of split value of 0.0428 at level 4 which is the terminal node.
13. Interview scores (Low) (High, Low,) obtained a goodness of split value of 0.0668 at level 1 which is the terminal node.

5. CONCLUSION

Based on the results of research that has been done by the author in determining the pattern of acceptance characteristics of new students in the Informatics Engineering Study Program, Harapan University, Medan with the CART (Classification And Regression Tree) data mining algorithm and data testing using RapidMiner 5.3 software, it can be concluded that the results of the decision tree in designing the characteristic pattern for accepting new student candidates are 23 rules where the highest factor of the 3 factors that influence the 75 data testing is the CAT (Computer Assisted Test) Score criterion. The second factor that also influences the outcome of the decision is the

* Corresponding author



This is an Creative Commons License This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International (CC BY-SA 4.0).

criteria for the value of the interview. And the last factor that also influences the outcome of the decision is the Average Report Card Score criterion. The results of the decision tree in the design of the characteristic pattern are 23 rules and implementation of the Data Mining CART (Classification And Regression Tree) algorithm by utilizing RapidMiner 5.3 software in determining the characteristic pattern of passing and failing admissions decisions to produce decision parameters in the form of a good decision tree. However, basically a prospective student can be said to have passed or not passed the acceptance of new students at the Informatics Engineering Study Program, Harapan University, Medan based on consideration of several established criteria

6. REFERENCES

- Aribowo, A., Kuswandhie, R., & Primadasa, Y. (2021). Penerapan dan Implementasi Algoritma CART Dalam Penentuan Kelayakan Penerima Bantuan PKH Di Desa Ngadirejo. *CogITO Smart Journal*, 7(1), 40. <https://doi.org/10.31154/cogito.v7i1.293.40-51>
- Asparizal, Yunita, P., & Ihsan, Z. (2016). SATIN – Sains dan Teknologi Informasi. *SATIN – Sains Dan Teknologi Informasi Journal*, 2(2), 90–99.
- Elisa, E. (2017). Analisa dan Penerapan Algoritma C4.5 Dalam Data Mining Untuk Mengidentifikasi Faktor-Faktor Penyebab Kecelakaan Kerja Kontruksi PT.Arupadhatu Adisesanti. *Jurnal Online Informatika*, 2(1), 36. <https://doi.org/10.15575/join.v2i1.71>
- Firdaus, D. (2017). Penggunaan Data Mining dalam Kegiatan Sistem Pembelajaran Berbantuan Komputer. *Jurnal Format*, 6(2), 91–97.
- Irmayani. (2020). Penerapan Algoritma Cart Klasifikasi Sosial Ekonomi Masyarakat Kelurahan Amessangeng. *Jurnal Ilmiah Information Technology d'Computare*, 10, 17–22.
- Karsito, & Monika Sari, W. (2018). Prediksi Potensi Penjualan Produk Delifrance Dengan Metode Naive Bayes Di Pt. Pangan Lestari. *Jurnal Teknologi Pelita Bangsa*, 9(1), 67–78.
- Mardi, Y. (2019). Data Mining : Klasifikasi Menggunakan Algoritma C4 . 5 Data mining merupakan bagian dari tahapan proses Knowledge Discovery in Database (KDD) . *Jurnal Edik Informatika. Jurnal Edik Informatika*, 2(2), 213–219.
- Prabawati, N. I., Widodo, & Ajie, H. (2019). Kinerja Algoritma Classification And Regression Tree (Cart) dalam Mengklasifikasikan Lama Masa Studi Mahasiswa yang Mengikuti Organisasi di Universitas Negeri Jakarta. *PINTER : Jurnal Pendidikan Teknik Informatika Dan Komputer*, 3(2), 139–145. <https://doi.org/10.21009/pinter.3.2.9>
- Prasetya, R. (2020). Penerapan Teknik Data Mining Dengan Algoritma Classification Tree Untuk Prediksi Hujan. *Jurnal Widya Climago*, 2(2), 13–23.
- Pratiwi, F. E., & Zain, I. (2014). Klasifikasi Pengangguran Terbuka Menggunakan CART (Classification and Regression Tree) di Provinsi Sulawesi Utara. *Jurnal Sains Dan Seni Pomits*, 3(1), D54–D59. http://www.ejournal.its.ac.id/index.php/sains_seni/article/view/6129
- Sikumbang, E. D. (2018). Penerapan Data Mining Penjualan Sepatu Menggunakan Metode Algoritma Apriori. *Jurnal Teknik Komputer AMIK BSI (JTK)*, Vol 4, No.(September), 1–4.
- Siregar, Y. S., & Harliana, P. (2018a). Algoritma Fuzzy C-Means Pada Aplikasi Matlab Dalam Menentukan Dosen Pembimbing Tugas Akhir. *Seminar Nasional Unisla*, 213–217. <http://semnas.unisla.ac.id/index.php/SAINS/article/download/198/20>
- Siregar, Y. S., & Harliana, P. (2018b). Analisis perancangan algoritma fuzzy c-means dalam menentukan dosen pembimbing tugas akhir. *Jurnal & Penelitian Teknik Informatika*, 3(1), 181–185.
- Siregar, Y. S., Sembiring, B. O., Hasdiana, H., Dewi, A. R., & Harahap, H. (2021). Algoritma C4.5 in mapping the admission patterns of new Students in Engineering Computer. *Sinkron*, 6(1), 80–90. <https://jurnal.polgan.ac.id/index.php/sinkron/article/view/11154>
- Sumartini, S. H., & Purnami, S. W. (2015). Penggunaan Metode Classification and Regression Trees (Cart) Untuk Klasifikasi. *Jurnal Sains Dan Seni Its*, 4(2), 211–216.

* Corresponding author



This is an Creative Commons License This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International (CC BY-SA 4.0).