Kirana Alyssa Putri - Implementation of Data Mining to Predict Student Study Period with Decision Tree Algorithm (C4.5)

by Layanan Perpustakaan UHAMKA

Submission date: 13-Jan-2024 09:39AM (UTC+0700)
Submission ID: 2270246523
File name: irana_Alyssa_Putri_Revised_Version_-_Kirana_Alyssa_Putri_1.docx (317.47K)
Word count: 4441
Character count: 24040

Implementation of Data Mining to Predict Student Study Period with Decision Tree Algorithm (C4.5)

Kirana Alyssa Putri[1], Dimas Febriawan[2]*, Firman Noor Hasan[3]
Informatics Engineering, FTII Prof. DR. Hamka Muhammadiyah University [1], [2], [3]
Prof. DR. Hamka Muhammadiyah University
Jakarta, Indonesia
kiranaalyssa.putri@gmail.com[1], dimas.febriawan@uhamka.ac.id[2], firman.noorhasan@uhamka.ac.id[3]

***Abstract*— Graduating on time is what every student wants to accomplish in college. Students of Prof. Dr. Hamka Muhammadiyah University are one of those who have this dream. Based on 2020 graduates data from the Tracer Study, 60% said the university had a high enough impact on improving competence. This data indicates that university needs to evaluate improvement of academic quality. Often, students have difficulty finding information about important factors that support achieving timely graduation. A prediction analysis is needed to provide information about the student's graduation study period. For this analysis, data mining implemented using the classification function of the decision tree (C4.5) algorithm with RapidMiner tools. The methodology for implementing data mining follows the stages of *Knowledge Discovery In Database* (KDD), beginning with data collection, preprocessing, transformation, data mining, and evaluation. The research findings consist of visualization and decision tree rules that reveal GPA as the most influential factor in determining a student's study period.There is other information, namely, students graduated on time (less than equal to 4 years) amounted to 170 or 54.5% and students did not graduate on time (more than 4 years) amounted to 142 or 45.6%. Testing the performance of *decision tree (C4.5)* utilizing *confusion matrix* through RapidMiner tools, resulted in accuracy reaching 83.87%, with precision of 87.50% and recall of 91.18%. Provides evidence that the *decision tree algorithm (C4.5)* has optimal performance to provide valuable information about predicting student graduation in order to increase student enrollment with the right study period.**

Keywords*— Decision Tree, C4.5 Algorithm, Prediction, Study Period, RapidMiner

I. INTRODUCTION

The University of Muhammadiyah Prof. DR. Hamka is a private university (PU) that has the vision and mission to produce outstanding graduates in both spiritual and academic intelligence. Based on this vision and mission, the university always provides professionalism and quality in every lesson. The goal is for students to develop their excellent competencies in the subjects they study and to graduate in a very timely manner [1]. The study period for a Bachelor's degree (S1) according to SN Dikti has a maximum limit of 7 academic years, with students expected to complete at least 144 semester credit units (SKS). Students are considered to graduate on time if they can complete their undergraduate studies in no more than 5 academic years or 10 semesters, based on the ideal completion in 4 academic years [2]. Universities have an important role in providing a significant impact on preparing for graduation. Based on data from 2020 graduates derived from the Tracer Study PU, as many as 60% of students stated that universities have a fairly high influence in improving the competence of graduates. This data indicates that universities need to conduct continuous evaluation to improve academic quality. Especially during the study period, students often have difficulty in finding information about important factors that support the achievement of timely graduation. Because completing studies on time has important functions for students, such as saving tuition fees and getting ready to enter the workforce more quickly. In addition, the importance of graduating on time is also related to the accreditation evaluation of the university [3], this can motivate universities to evaluate the development of students during the college term through the academic system and the support of

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.xx.xx, Copyright ©2020
Submitted : , Revised : , Accepted : , Published :

lecturers. There is a need for information regarding student graduation based on academic data, which can be utilized to generate new knowledge, this will enable PU to better understand the conditions surrounding students' graduation [4]. As an educational institution, it should be able to manage data to enhance decision making, especially for the benefit of students [5].

According to the background explanation, data mining can be applied to predict the study period of PU students, particularly in the Faculty of Industrial Technology and Informatics (FTII). Data mining is the study of extracting useful knowledge from data, and the resulting insights have a significant impact on decision making [6]. The main goal of data mining is to identify meaningful data correlations and extract valuable new information using statistical techniques [7]. The application of data mining is able to reveal patterns and information hidden in large data sets [8]. Data mining is an element of *Knowledge Discovery in Databases* (KDD) methods. KDD refers to a series of steps to identify valuable, easily understood information from data that was previously large in scale and high in complexity. In general, the KDD process, which is the basis of data mining, is data collection, pre-processing/cleaning, transformation, data mining and evaluation [9]. Data mining is categorized into description, estimation, prediction, classification, and clustering. One technique that is always used for predicting student graduation is classification [10]. Classification is a data analysis technique applied to group data into categories based on certain attributes. The classification process includes the following stages, the learning phase (training phase) where classification algorithms are applied to the training data to create classification rules which are then used to classify new data. The second stage is classification, with testing data that serves for accuracy calculation [11]. Various analytical methods applied in classification can be an appropriate choice for prediction [12]. There are various classification algorithms in data mining such as *decision tree, naïve bayes, support vector machine, k-nearest neighbor* [13].

The researchers employ one of the classification methods, specifically the *decision tree algorithm (C4.5)*. This algorithm proves to be an optimal choice for predicting students' study time due to its advantages, such as being easy to understand and interpret, handling incomplete data, and managing datasets with numerous attributes [14]. The research aims to generate information from data that has been collected, and become the basis for predictions on new data. The dataset utilized comprises the graduation records of FTII PU students. Analysis is needed through testing the performance of the previously formed *decision tree* model. Performance testing through the application of *confusion matrix*, which is generated through comparison of the predicted results of the model with the actual results of the test. Data processing and performance testing of the *decision tree algorithm (C4.5)* are carried out using the help of RapidMiner. This tool provides solutions in data mining analysis. In RapidMiner, various methods such as the *Decision Tree (C4.5) algorithm* are available and can be applied to a set of data including calculations and trials [15].

There is a similar previous research conducted by Endang Etriyanti, Dedy Syamsuar, and Yesi Novaria Kunang (2020) with the title “Implementasi Data Mining Menggunakan Algoritme Naive Bayes Classifier dan C4.5 untuk Memprediksi Kelulusan Mahasiswa”. The findings of the research analysis produced information that the accuracy level of the *C4.5 decision tree algorithm* was 79.08% while the naïve bayes algorithm was only 78.46% [16]. Further research by Lydia Yohana Lumban Gaol, M. Safii, and Dedi Suhendro (2021) with the title "Prediksi Kelulusan Mahasiswa Stikom Tunas Bangsa Prodi Sistem Informasi Dengan Menggunakan Algoritma C4.5”. The research shows that the algorithm used is very good with an accuracy rate of 90.00% with precision 91.38% and recall 98.15% [10]. Another research by Agung Wibowo and Abdul Rohman (2022) entitled “Prediksi Predikat Kelulusan Mahasiswa Menggunakan Naive Bayes dan Decision Tree pada Universitas XYZ”. The focus of this research is to provide information about the description of students who achieve certain graduation predicates and analyze the factors that have an impact on achievement [17].

The main difference between this research and

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.xx.xx, Copyright ©2020
Submitted : , Revised : , Accepted : , Published :

previous research is in the selection of algorithms and methodological approaches. In previous research by Endang Etriyanti, Dedy Syamsuar, and Yesi Novaria Kunang (2020) and Agung Wibowo and Abdul Rohman (2022) two classification techniques were used, namely the *naïve bayes* algorithm and *decision tree*. On the other hand, research conducted by Lydia Yohana Lumban Gaol, M. Safii, and Dedi Suhendro (2021) used a methodological approach without including the pre-processing stage (data cleaning). The current research specifically employs a singular classification method, namely the *decision tree algorithm (C4.5)*, and implements the research method from the pre-processing stage to the evaluation of the algorithm model.

II. RESEARCH METHODS

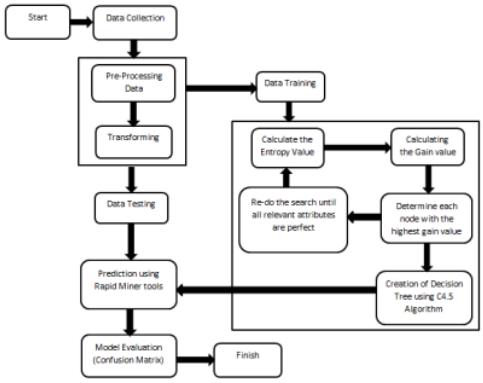

Figure 1. Stage Of The Research Method

1. *Data Collection*

This research uses the data of FTII PU Students Graduation in 2021-2022 as the source of information in the analysis. This is the main data obtained directly from the university database. The data collection process is done through access and official procedures set by the university. This data includes 312 data, which consists of 7 columns, containing students id, total credit courses, GPA, research, length of study and yudisium. This data will be used in the analysis to identify patterns and gain insights into the study periods of students.

TABLE I. FTII PU Student Graduation (2021-2022)

Attribute Name	Data Type
ID	*Integer*
Total Credit Courses	*Integer*
GPA	*Float*
Research	*Text*
Length Of Study	*Integer*
Yudisium	*Text*

2. *Pre-processing/Cleaning*

This stage is the process of initial preparation and data cleaning. In the pre-processing stage, it is important to deal with incomplete data and to ensure the completeness of the data to maintain the integrity and stability of the data [18]. The purpose of this process is to clean the dataset by identifying and addressing potential errors and inconsistencies in the data that may exist. This process involves several steps that are performed in pre-processing [19]:

a. Data Validation

This step is used to detect and remove noisy data. Data validation is employed to identify inconsistent and anomalous data. If data is found that could interfere with the subsequent analysis process, it will be deleted. This process is performed manually using Microsoft Excel software.

b. Data Cleaning

The second stage involves the data cleaning process using RapidMiner tools to enhance speed and efficiency. Various operators are employed such as the read excel operator for reading data files in xlsx format. The missing value operator to replace missing values with zeros or, alternatively, with the average value of the entire dataset. Additionally, the remove duplicates operator for removes duplicate data. And select attributes operator is applied in this process to selectively eliminate attributes considered less influential in the subsequent analysis of student graduation data, specifically the Student ID and Yudisium attributes [20]. The data cleaning process with RapidMiner is outlined as follows:

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.xx.xx, Copyright ©2020
Submitted : , Revised : , Accepted : , Published :

Figure 2. Cleaning and Select Attribute data process with RapidMiner

3. ***Transforming***

Data transformation is an optional stage in the data mining process, which means that this stage is implemented if necessary. This stage will classify the data to categorize the data based on certain attributes [21]. In this research, transforming is used to change the data class with a certain criteria classification to facilitate understanding of a large dataset of 312 data. The data type in table 1 which is the initial data is converted into binomial and polynomial data types according to data mining rules.

TABLE II. Data Transformation

Attribute Name	Data Type	Class Of Data Used
Total Credit Courses	*Binomial*	≤ 150 Credit Courses, > 150 Credit Coruses
GPA	*Polynomial*	Statisfactory, Very Statisfactory, Cumlaude
Research	*Binomial*	Thesis, Publication
Length Of Study	*Binomial*	≤ 4 Years, > 4 Years

4. **Data Mining**

The data mining method applied in this study is the *Decision Tree*, a technique with the capability to transform complex datasets into a hierarchical series of decisions, resembling a tree with branches and leaves that signify decisions and class labels. Each node in this structure represents an attribute that has been tested to make a decision [22]. The application of data mining can also be used to understand the relationship between different attribute values [23]. In this research, using *Decision Tree* with the *C4.5 algorithm*. The *C4.5 algorithm* was invented through the development of the *ID3 algorithm* by Ross Quinlan. With a number of significant improvements, including the ability to handle missing values and noise in the dataset, simplify the interpretation of results, and produce more informative rules from the tree model formed [24]. In generating a *decision tree* with the *C4.5 algorithm*, several steps need to be taken. These steps are as follows [25]:

Calculate the *Entropy* of each attribute to determine the root node and other nodes that continue the formation of the *decision tree*, using the formula:

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

Description:

S : Dataset

n : Number Of Classes S

pi : Sum Of The Proportion of Si to S

Next, obtain *entropy* for all attributes, with the process of calculating the *gain* value using the formula:

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

Description:

S : Dataset

A : Features

n : Number Of Attribute Partitions

A |Sil : Number Of Data in the i-th partition

|Sl : Number Of Data S

5. **Result Evaluation**

Confusion Matrix is a method implemented in the process of testing classification patterns to gain an understanding of the performance of the prediction model. *Confusion Matrix* is a table containing various evaluation metrics where the predicted results of the model will be compared with the actual class of the input data. Using information from the *confusion matrix*, researchers can measure accuracy, precision, recall F1-score and other evaluation parameters [26] [27].

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.xx.xx, Copyright ©2020
Submitted : , Revised : , Accepted : , Published :

TABLE III. *Confusion Matrix* Table

Predicted Classification	Actual Classification	
	Positive	Negative
Positive	True Positive (TP)	False Positive (FP)
Negative	False Negative (FN)	True Negative (TN)

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

$$Precision = \frac{TP}{TP+FP} 100\% \quad (4)$$

$$Recall = \frac{TP}{TP+FN} 100\% \quad (5)$$

Description [28]:

TP = The amount of positive data correctly categorized by the system.

TN = The amount of negative data correctly classified by the system.

FP = The amount of negative data misclassified by the system as positive.

FN = The amount of positive data misclassified by the system as negative.

III. Results And Discussion

In this section, the data is ready to enter the data mining stage, with the application of *decision tree (C4.5)* for pattern formation. The calculation of the *decision tree algorithm (C4.5)* uses the *entropy* and *gain* formulas to find the root node and the next node. In this research, the calculation uses Microsoft Excel software.

A. Calculation to Find Entropy

In this study, the label is the length of study with the criteria ≤4 years and >4 years. The *Entropy* formula is as follows.

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

Means:

$$Entropy(\text{Total}) = ((-\frac{sum(\leq 4 \text{ Years})}{total} * log_2 (\frac{sum(\leq 4 \text{ Years})}{total})) + ((-\frac{sum(> 4 \text{ Years})}{total} * log_2 (\frac{sum(> 4 \text{ Years})}{total}))$$

$$Entropy(\text{Total}) = ((-\frac{170}{312} * log_2 (\frac{170}{312})) + ((-\frac{142}{312} * log_2 (\frac{142}{312}))$$

$$= 0{,}994182507$$

TABLE IV. Total *Entropy* Calculation Of Data (Label)

Total Data	Length Of Study		Entropy
312	≤ 4 Years	> 4 Years	0.994182507
	170	142	

From table 4, it is known that students who graduated on time (≤4 years) amounted to 170, meaning 54.5% and students who did not graduate on time (>4 years) amounted to 142 or 45.6%. Then the *entropy* calculation is carried out from the entire number of cases, namely 312 data with the length of study label which results in 0.994182507. *Entropy* of length of study will be the basis for calculating *gain* on other attributes. Furthermore, looking for *entropy* in attributes such as GPA, Research and Total Credit Courses.

1. Entropy Calculation for GPA

GPA is grouped into 3 categories: Satisfactory for students with GPA from 2.61 to 3.00, Very Satisfactory for students with GPA from 3.01 to 3.50 and Cumlaude for students with GPA from 3.51 to 4.00.

$$Entropy(\text{GPA, Statisfactory}) = ((-\frac{21}{118} * log_2 (\frac{21}{118})) + ((-\frac{97}{118} * log_2 (\frac{97}{118})) = 0.675607358$$

$$Entropy(\text{GPA, Very Statisfactory}) = ((-\frac{111}{153} * log_2 (\frac{111}{153})) + ((-\frac{42}{153} * log_2 (\frac{42}{153})) = 0.847861745$$

$$Entropy(\text{GPA, Cumlaude}) = ((-\frac{38}{41} * log_2 (\frac{38}{41})) + ((-\frac{38}{41} * log_2 (\frac{3}{41})) = 0.377646321$$

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.xx.xx, Copyright ©2020
Submitted : , Revised : , Accepted : , Published :

2. Entropy Calculation for Research

The research is categorized into two groups, namely Thesis for students who graduate with thesis research and Publication for students who graduate with publication (journal) research.

Entropy(Research, Thesis) = $((-\frac{165}{307} * log_2\ (\frac{165}{307})) + ((-\frac{142}{307} * log_2\ (\frac{142}{307})) = 0.995947431$

Entropy(Research, Publication) = $((-\frac{5}{5} * log_2\ (\frac{5}{5})) + ((-\frac{0}{5} * log_2\ (\frac{0}{5})) = 0$

3. Entropy Calculation for Total Credit Courses

Total Credit Courses are grouped into 2 categories, namely ≤150 credit courses for students who complete less than or equal to 150 credits. Meanwhile, the >150 credit courses category is for students who complete more than 150 credits.

Entropy(Total Credit Courses, ≤150 Credit Courses) = $((-\frac{29}{64} * log_2\ (\frac{29}{64})) + ((-\frac{35}{64} * log_2\ (\frac{35}{64})) = 0.993650712$

Entropy(Total Credit Courses, >150 Credit Courses) = $((-\frac{141}{248} * log_2\ (\frac{141}{248})) + ((-\frac{107}{248} * log_2\ (\frac{107}{248})) = 0.9863991$

B. Calculation to Find Gain

Gain calculation formula.

$$Gain(S,A) = Entropy(S) - \sum_{i=1}^{n} \frac{|si|}{|s|} * \text{Entropy(Si)}$$

1. Gain Calculation for GPA

$Gain(\text{Total, GPA}) = \text{Entropy(Total)} - \sum_{i=1}^{n} \frac{|\text{GPA}|}{|\text{Total}|} * \text{Entropy GPA}$

$= 0.994182507 - ((\left(\frac{118}{312}\right) * 0.675607358) + (\left(\frac{153}{312}\right) * 0.847861745) + (\left(\frac{41}{312}\right) * 0.377646321))$

$= 0.273259384$

2. Gain Calculation for Research

Gain(Total, Research)

$= Entropy(Total) - \sum_{i=1}^{n} \frac{|\text{Research}|}{|Total|} *$ Entropy Research

$= 0.994182507 - ((\left(\frac{307}{312}\right) * 0.995947431) + (\left(\frac{5}{312}\right) * 0))$

$= 0.014195772$

3. Gain Calculation for Credit Courses

Gain(Total, Total Credit Courses)

$= Entropy(Total) - \sum_{i=1}^{n} \frac{|\text{Total Credit Courses}|}{|Total|} *$ Entropy Total Credit Courses

$= 0.994182507 - ((\left(\frac{64}{312}\right) * 0.993650712) + (\left(\frac{248}{312}\right) * 0.9863991))$

$= 0.006295897$

Determining the root node involves selecting the *gain* with the largest value. In this case, the root node is GPA, which exhibits the highest gain of 0.273259384. The GPA attribute is divided into three partitions namely Satisfactory with *entropy* 0.675607358, Very Satisfactory with *entropy* 0.847861745, and Cumlaude with *entropy* 0.377646321. The stopping criterion is set at zero. In other words, this attribute no longer provides further information. Since none of the three partitions of the GPA attribute have reached zero, the *decision tree* process continues. However, if all *entropy* values on the attribute become zero, the *decision tree* process will stop [29]. Similarly, with the selection of the highest *entropy* value, which is 0.847861745 for the very satisfactory category, this value is used as the basis for calculating the total *entropy* in finding node 1.2. The calculation is performed in the same manner. Below is the initial display of the *decision tree* formation with GPA as the root node.

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.xx.xx, Copyright ©2020
Submitted : , Revised : , Accepted : , Published :

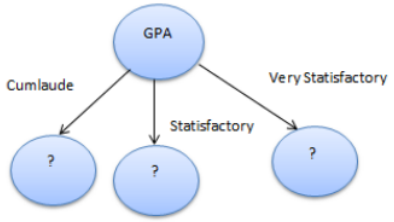

Figure 3. Initial Decision Tree

First calculate the node for GPA (Very Satisfactory) which gets the largest gain result, namely Research and is selected to be the next node 1.2. Then, the partition with the largest *entropy* is Thesis to continue the decision tree node 1.3. The last node is Total Credit Courses, so the calculation for node 1 is complete.

TABLE V. Node 1.2 Calculation Results

Node 1.2	Category	Data	≤4 Years	>4 Years	Entropy	Gain
GPA	Very Statisfactory	153	111	42	0.847861745	
Research						0.006099049
	Thesis	151	109	42	0.852911871	
	Publication	2	2	0	0	
Total Credit Courses						0.00228158
	≤150 Credit Courses	31	22	9	0.869137581	
	>150 Credit Courses	122	89	33	0.842169458	

TABLE VI. Node 1.3 Calculation Results

Node 1.3	Category	Data	≤ 4 Years	> 4 Years	Entropy	Gain
Research	Thesis	151	109	42	0.852911871	
Total Credit Courses						0.000420544
	≤150 Credit Courses	30	21	9	0.881290899	
	>150 Credit Courses	121	88	33	0.845350937	

Here are the decision trees for nodes 1, 1.2, and 1.3:

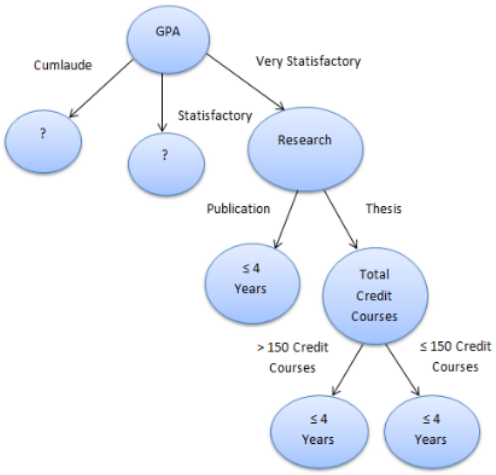

Figure 4. Node 1 Decision Tree

The calculation of *entropy* and *gain* will be done in the same way, satisfactory partition will be used in getting node 2 and cumlaude partition to get node

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.xx.xx, Copyright ©2020
Submitted : , Revised : , Accepted : , Published :

3. Below is a calculation table and a picture of the completed *decision tree*:

TABLE VII. Node 2 Calculation Results

Node 2	Category	Data	≤ 4 Years	> 4 Years	Entropy	Gain
GPA	Statisfactory	118	21	97	0.675607358	
Research						0
	Thesis	118	21	97	0.675607358	
	Publication	0	0	0	0	
Total Credit Courses						5,61756E-07
	≤150 Credit Courses	28	5	23	0.67694187	
	>150 Credit Courses	90	16	74	0.67519144	

TABLE VIII. Node 3 Calculation Results

Node 3	Category	Data	≤ 4 Years	> 4 Years	Entropy	Gain
GPA	Cumlaude	41	38	3	0.377646321	
Research						0.008342604
	Thesis	38	35	3	0.398459274	
	Publication	3	3	0	0	
Total Credit Courses						0.118317676
	≤150 Credit Courses	4	2	2	1	
	>150 Credit Courses	37	36	1	0.179256067	

TABLE IX. Node 3.2 Calculation Results

Node 3.2	Category	Data	≤4 Years	>4 Years	Entropy	Gain
Total Credit Courses	>150 Credit Courses	37	36	1	0.17926067	
Research						0
	Thesis	33	33	0	0	
	Publication	4	4	0	0	

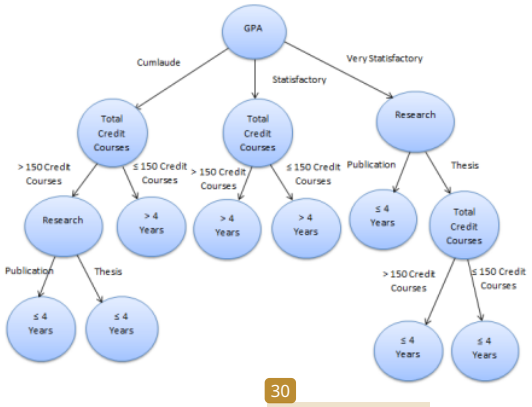

Figure 5. Final Result Of *C4.5 Algorithm Decision Tree*

The results of the *decision tree* model with RapidMiner for student graduation data are as

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.xx.xx, Copyright ©2020
Submitted : , Revised : , Accepted : , Published :

follows:

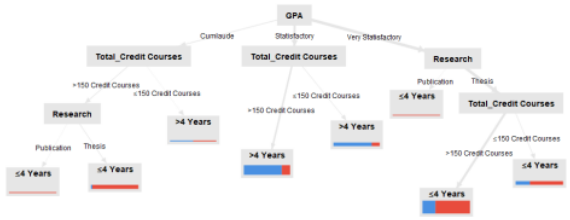

Figure 6. *Decision Tree (C4.5)* Results with RapidMiner

Calculations using Microsoft Excel software with *entropy* and *gain* formulas display the same results as the results of the *decision tree* using RapidMiner tools. This shows that the calculations that have been done are correct according to the rules of the *decision tree algorithm (C4.5)*.

Tree

```
GPA = Cumlaude
|   Total_Credit Courses = >150 Credit Courses
|   |   Research = Publication: ≤4 Years (>4 Years=0, ≤4 Years=3)
|   |   Research = Thesis: ≤4 Years (>4 Years=1, ≤4 Years=33)
|   Total_Credit Courses = ≤150 Credit Courses: >4 Years (>4 Years=2, ≤4 Years=2)
GPA = Satisfactory
|   Total_Credit Courses = >150 Credit Courses: >4 Years (>4 Years=73, ≤4 Years=16)
|   Total_Credit Courses = ≤150 Credit Courses: >4 Years (>4 Years=24, ≤4 Years=5)
GPA = Very Satisfactory
|   Research = Publication: ≤4 Years (>4 Years=0, ≤4 Years=2)
|   Research = Thesis
|   |   Total_Credit Courses = >150 Credit Courses: ≤4 Years (>4 Years=33, ≤4 Years=88
|   |   Total_Credit Courses = ≤150 Credit Courses: ≤4 Years (>4 Years=9, ≤4 Years=21)
```

Figure 7. Description Of Data Processing Results Of FTII PU Student Graduation

There is a textual description of the decision tree model formed from the *C4.5 algorithm*. This description shows that the *decision tree algorithm (C4.5)* is a good analytical technique for drawing conclusions from data that is easy to understand because it is visualized in a *decision tree*. Below is an explanation of the rules generated.

1. Rule 1 GPA = Cumlaude, Total Credit Courses = >150 Credit Courses and Research = Publication, then Length Of Study ≤ 4 Years.
2. Rule 2 GPA = Cumlaude, Total Credit Courses = ≤150 Credit Courses and Research = Thesis, then Length Of Study ≤ 4 Year.
3. Rule 3 GPA = Cumlaude dan Total Credit Courses = ≤150 Credit Courses, then Length Of Study > 4 Years.
4. Rule 4 GPA = Statisfactory and Total Credit Courses = >150 Credit Courses, then Length Of Study > 4 Years.
5. Rule 5 GPA = Statisfactory and Total Credit Courses = ≤150 Credit Courses, then Length Of Study > 4 Years.
6. Rule 6 GPA = Very Statisfactory and Research = Publication, then Length Of Study ≤ 4 Years.
7. Rule 7 GPA = Very Statisfactory, Research = Thesis and Total Credit Courses = >150 Credit Courses, then Length Of Study ≤ 4 Years.
8. Rule 8 GPA = Very Statisfactory, Research = Thesis and Total Credit Courses = ≤150 Credit Courses, then Length Of Study ≤4 Years.

C. Evaluation Of Results

The performance testing analysis of the *decision tree algorithm (C4.5)*, formed from the graduation data of FTII PU students in 2021-2022, is conducted using a *confusion matrix*. Various RapidMiner operators, such as the split data, are utilized to divide the data into training data (80% of the total original data) and test data (20% of the original data) [30]. The apply model operator is then employed to apply models previously trained using the training data to unlabeled data (testing data) [31]. And the last operator is performance for accuracy testing with a *confusion matrix*. The subsequent section outlines the accuracy testing process using RapidMiner tools.

Figure 8. *Confusion Matrix* Accuracy Process

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.xx.xx, Copyright ©2020
Submitted : , Revised : , Accepted : , Published :

The following are the results of the *confusion matrix* for calculating the accuracy of the *decision tree algorithm (C4.5)*.

accuracy: 83.87%

	true >4 Years	true ≤4 Years	class precision
pred. >4 Years	21	3	87.50%
pred. ≤4 Years	7	31	81.58%
class recall	75.00%	91.18%	

Figure 9. Accuracy Value Of *C4.5 Decision Tree Algorithm*

The RapidMiner result *confusion matrix* table displays the predictions, with a summary of the number of correct and incorrect predictions by the processed *decision tree (C4.5) algorithm*. The matrix of predicted values is located in the rows, while the actual or true values are located in the columns. The dataset described has been divided into 80% training data and 20% testing data. The training data is used to build the model, while the testing data is used to validate the predictions of the new data.

Manual calculations are performed using the *confusion matrix* to generate accuracy, which describes how accurate the model is in classifying correctly. In other words, accuracy reflects the closeness of the predicted value to the true value. Calculation of accuracy can be done using the following formula:

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

$$= \frac{21+31}{21+31+3+7} \text{ x } 100\%$$

$$= \frac{52}{62} \text{ x } 100\%$$

$$= 0.8387 \text{ x } 100\%$$

$$= 83.87\%$$

The parameters that produce an accuracy of 83.87% involve True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). These values are derived from RapidMiner results, where True Positive (TP) is 21, indicating that the algorithm model classifies students with a study period >4 years and correctly predicts that the student's study period is >4 years. True Negative (TN) is 31, indicating that the algorithm model classifies students with a study period ≤4 years and predicts correctly that the student's study period is ≤4 years. False Positive (FP) is 3, indicating that the algorithm model classifies students with a study period ≤4 years, but the prediction result states that the study period is >4 years. And lastly, False Negative (FN) states that the algorithm model classifies students with a study period >4 years, but the prediction result states that the study period is ≤4 years.

The results of the comparison of manual calculations with RapidMiner show the accuracy level of the performance of the *decision tree algorithm (C4.5)* on the graduation data of FTII PU students reaches 83.87%. This proves that the information that has been known for predicting student graduation is good enough and can be used for decision making. Because, the algorithm model is considered a good model when it has a high True Positive (TP) and True Negative (TN), while False Positive (FP) and False Negative (FN) are low. Thus, it can be concluded that the implementation of the *decision tree algorithm (C4.5)* for predicting the study period of FTII PU students is included in the good enough category and can be used as a reference for decision making.

PerformanceVector

```
PerformanceVector:
accuracy: 83.87%
ConfusionMatrix:
True:   >4 Years        ≤4 Years
>4 Years:       21      3
≤4 Years:       7       31
```

Figure 10. *C4.5 Algorithm* Vector Performance Value

IV. CONCLUSION

The *decision tree algorithm (C4.5)* which is applied to predict the duration of time or study period for FTII students at Prof. Dr. Hamka Muhammadiyah University in 2021-2022, produces information or knowledge about the most influential factor in determining student study time, namely

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.xx.xx, Copyright ©2020
Submitted : , Revised : , Accepted : , Published :

GPA with an accuracy rate of 83.87%, precision reaching 87.50% and recall of 91.18%. Further information obtained, namely students who completed graduation on time (≤4 years) amounted to 170 or 54.5% and students who did not completed graduation on time (>4 years) amounted to 142 or 45.6%. In addition, there are 8 rules from the *decision tree* based on student graduation data that has been processed. Thus, the explanation of graduation prediction information will be easier to understand for universities and students. The conclusion of this research is that the prediction pattern formed through the *decision tree algorithm (C4.5)* has good performance in providing knowledge about student graduation. This can support universities in making decisions to increase the number of students with an on time study period and present this information to students through the academic system. Thus, students can know the factors that affect the study period and plan the lecture strategy from the beginning, aiming to achieve graduation according to the expected schedule.

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.xx.xx, Copyright ©2020
Submitted : , Revised : , Accepted : , Published :

Kirana Alyssa Putri - Implementation of Data Mining to Predict Student Study Period with Decision Tree Algorithm (C4.5)

ORIGINALITY REPORT

18%	14%	14%	10%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Universitas Brawijaya Student Paper	7%
2	Sisca Yuliantina, Betha Nurina Sari. "Application of C4.5 Classification in Improving Recitation Fluency in Students", Paradigma - Jurnal Komputer dan Informatika, 2023 Publication	1%
3	www.semanticscholar.org Internet Source	1%
4	Fitriana Harahap, Evri Ekadiansyah, Erwin Ginting, Nidia Enjelita Saragih, Robiatul Adawiyah, Ermayanti Astuti. "Factors of the Decrease in Student's Interest in Learning During the Covid-19 Pandemic Using the C4.5 Method", 2021 3rd International Conference on Cybernetics and Intelligent System (ICORIS), 2021 Publication	1%

5	scholarworks.sjsu.edu Internet Source	1 %
6	jurnal.atmaluhur.ac.id Internet Source	1 %
7	socj.telkomuniversity.ac.id Internet Source	1 %
8	article.sciencepublishinggroup.com Internet Source	1 %
9	journal.universitasbumigora.ac.id Internet Source	1 %
10	Submitted to Asia e University Student Paper	<1 %
11	"Prediction using C4.5 Method and RFM Method for Selling Furniture", International Journal of Engineering and Advanced Technology, 2019 Publication	<1 %
12	journal.pandawan.id Internet Source	<1 %
13	jurnal.sar.ac.id Internet Source	<1 %
14	hrcak.srce.hr Internet Source	<1 %
15	Submitted to Universitas Mercu Buana Student Paper	<1 %

16	Submitted to University of Sheffield Student Paper	<1 %
17	Dede Kurniadi, Shopi Nurhidayanti, Indri Tri Julianto, Teguh Wahyono, Yosep Septiana, Hetty Rohayani. "Classification of Television Programs Based on Public Opinion in Social Media Using Random Forest and Decision Tree", 2023 International Conference on Computer Science, Information Technology and Engineering (ICCoSITE), 2023 Publication	<1 %
18	www.slideshare.net Internet Source	<1 %
19	Dewi Ratnawati, Isnaini Handayani, Windia Hadi. "Pengaruh Model Pembelajaran Pbl Berbantu Question Card Terhadap Kemampuan Berpikir Kritis Matematis Siswa Smp", Edumatica : Jurnal Pendidikan Matematika, 2020 Publication	<1 %
20	dspace.umkt.ac.id Internet Source	<1 %
21	Budi Sunarko, Uswatun Hasanah, Ulfah Mediaty Arief, Feddy Setio Pribadi et al. "Prediction of Student Satisfaction with Academic Services Using Naive Bayes Classifier", 2022 6th International Conference	<1 %

on Information Technology, Information Systems and Electrical Engineering (ICITISEE), 2022
Publication

---

22 Muhammad Naufal Mukhbit Amrullah, Adiwijaya Adiwijaya, Widi Astuti. "Implementation of Modified Backpropagation with Conjugate Gradient as Microarray Data Classifier with Binary Particle Swarm Optimization as Feature Selection for Cancer Detection", Jurnal Sisfokom (Sistem Informasi dan Komputer), 2020 <1%
Publication

---

23 Rina Novita, Supratman Zakir, Agus Nur Khomarudin, Efmi Maiyana, Hamimah Hasyim. "Use of the C4.5 Algorithm in Determining Scholarship Recipients", Journal of Physics: Conference Series, 2021 <1%
Publication

---

24 ia903401.us.archive.org <1%
Internet Source

---

25 Rina Ayu Wulan Sari, Ivan Jaya, Marischa Elveny, Sri Melvani Hardi. "Classification of Beef Freshnesh With Color and Texture Feature Extraction Using Learning Vector Quantization Algorithm", 2021 International Conference on Data Science, Artificial <1%

Intelligence, and Business Analytics (DATABIA), 2021
Publication

26	ebin.pub Internet Source	<1%
27	K. C. Tan, Q. Yu, J. H. Ang. "A coevolutionary algorithm for rules discovery in data mining", International Journal of Systems Science, 2006 Publication	<1%
28	Roslaini Roslaini, Siti Ithriyah, Ika Purnama Sari, Makmur Harun. "THE UTILIZATION OF ISLAMIC-BASED TEXTS FOR CHARACTER INNOVATION: EFL STUDENTS' PERCEPTION", Jurnal As-Salam, 2023 Publication	<1%
29	Saruni Dwiasnati, Yudo Devianto. "Optimasi Prediksi Keputusan Calon Nasabah Potensial menggunakan Algoritma C 4.5 berbasis Particle Swarm Optimization", Jurnal Informatika, 2019 Publication	<1%
30	sistemasi.ftik.unisi.ac.id Internet Source	<1%
31	www.ijrte.org Internet Source	<1%

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	Off		