Journal of Applied Engineering and Technological Science
Vol 4(2) 2023: 955-968

CLASSIFICATION ACADEMIC DATA USING MACHINE LEARNING FOR DECISION MAKING PROCESS

Elin Haerani[1*], Fadhilah Syafria[2], Fitra Lestari[3], Novriyanto[4], Ismail Marzuki[5]
Informatics Engineering Department, Faculty of Science and Technology, UIN Sultan Syarif Kasim Riau, Pekanbaru, Indonesia[124]
Industrial Engineering Department, Faculty of Science and Technology, UIN Sultan Syarif Kasim Riau, Pekanbaru, Indonesia[3]
Interdigital, Canada[5]
Email: elin.haerani@uin-suska.ac.id[1*], fadhilah.syafria@uin-suska.ac.id[2], fitra.lestari@uin-suska.ac.id[3], novriyanto@uin-suska.ac.id[4], ismail.marzuki@interdigital.com[5]

Received : 15 April 2023, Revised: 28 May 2023, Accepted : 29 May 2023
**Corresponding Author*

ABSTRACT
One of the qualities of higher education is determined by the success rate of student learning. Assessment of student success rates is based on students' graduation on time. The university always evaluates the performance of its students to find out information related to the factors that cause students to become inactive so that they are more likely to drop out and what data affects students ability to graduate on time. The evaluation results are stored in an academic database so that the data can later be used as supporting data when the university makes decisions. The data was processed using the Decision Tree C4.5 method so as to produce a model in the form of a tree and rules. The data used in this study is the graduation data of Informatics Engineering students from 2011 to 2015, totaling 632 data records. Variables used are Nim, in-progress grades each semester, credit taken every semester, GPA, and graduation status. Tests were conducted using split data scenarios with comparison of training data: 90:10, 80:20, 70:30, 60:40, and 50:50. Based on the test results, it is known that the attribute that influences the success of student studies is the grade point average (GPA), where the accuracy of the maximum recognition rate is 88.19% is in the comparison of training data and test data (80%: 20%).
Keywords: *Data science, Decision Tree, Graduate on Time, Machine Learning*

1. Introduction

One of the factors that determine the quality of a university is a high student success rate and a low student failure rate, and vice versa. This statement is corroborated by the rate. The success rate of students is based on their graduation on time, according to D.Heredia in his junal while student failure is based on the status of students who drop out (Heredia, D., Amaya, Y., & Barrientos, 2015). The higher the success rate of students, the better the quality of higher education is evaluated by the study program accreditation assessment instrument from BAN PT, which states that the percentage of students who graduate on time is one element of the university's accreditation assessment (BAN-PT, 2011).

Inactive students or alpha studies can potentially cause problems for graduation if they are not on time and even drop out (Burgos et al., 2018), so that the quality of education and the assessment of higher education accreditation decrease (Yossy, E. H., Heryadi, Y., 2019). One of the most prevalent problems in education is dropout. The high rate of dropouts will negatively affect higher education (Utari, M., Warsito, B., & Kusumaningrum, 2020). The percentage of graduate students and the university's methods for preventing student dropouts are used to gauge a university's reputation (Dharmawan, T., Ginardi, H., & Munif, n.d.). To assess the degree of success of the technique at each university, initial projections of students who are at risk of dropping out are crucial (Asif et al., 2017) . The university always evaluates the performance of its students to find out information related to the factors that cause students to become inactive so that they are more likely to drop out and what data affects students ability to graduate on time, as stated in the journal (Sukhbaatar, O., Ogata, K., & Usagawa, 2018). The results of the evaluation are stored in an academic database for later use as supporting data when making university decisions (Osman Hegazi, M., & Abugroon, 2016). This large amount of data opens up

opportunities to produce useful information for universities (Heredia, D., Amaya, Y., & Barrientos, 2015).

Data science is the science or technique of exploring and extracting data sets or databases to find new models, shapes, patterns, and insights that can be used as decision-making tools (Baker, 2014). The concept of data science is more about how to extract (excavate) or predict data analysis that will be filtered so that the correct data is found to produce accurate data according to actual data, as stated in research (M. , A., & Rahman, 2016), (Campagni et al., 2015). The use of data mining in higher education and higher learning institutions is relatively new. In research (Baker, 2014; Osman Hegazi, M., & Abugroon, 2016), numerous studies have been conducted to show the value of "data mining" approaches in education, showing that this is a novel idea for the extraction of reliable and accurate data regarding behavior and learning efficacy. The decision tree provides good accuracy in the prediction of student dropout, according to the research in (Teli, S., & Kanikar, 2015b).

Extracting or digging up student information from large data sets cannot be done easily (M. Han, 2006). According to (Fernandes et al., 2019) explained that data mining technology is an interdisciplinary field of research whose essence is the intersection of machine learning, statistics, and databases. In the book Decision Support Systems and Intelligent Systems, it is stated that data mining is a process that uses statistical techniques, mathematics, artificial intelligence, and machine learning to extract and identify useful information and related knowledge from various large databases (Jain, A., Somwanshi, D., Joshi, K., & Bhatt, 2022; Turban, 2005). Machine learning can be defined as the result of using algorithms to use data, study it, and then predict it. Machine learning consists of statistical analysis and predictive analysis used to find patterns and capture hidden insights based on perceived data (D. , R., M. , & Giansiracusa, 2018).

To quantify timely completion, classification data mining can be applied to student value data (Natek & Zwilling, 2014; Yunus, M., Ramadhan, H., Aji, D. R., & Yulianto, 2021). In this instance, classification is the data mining technique that will be employed (Massulloh, I., 2020). Student grade data will be categorized based on a number of variables, including attendance, grade history, and other variables. The system will then use this classification to suggest solutions to help students graduate at the right time with the best values (Asroni, A., Masajeng Respati, B., & Riyadi, 2018). Student primary data and student academic data are the system's inputs (Khan et al., 2021). Decision trees, naive Bayes algorithms, neural networks, and nearest neighbor classifiers are a few classification techniques that can be utilized in data mining (Rasjid & Setiawan, 2017).

Several studies related to data mining have been used as an effort to improve the quality of higher education (Yu & Zhang, 2018) and (M. , A., & Rahman, 2016). In this study, academic data analysis will be carried out to classify the continuity and success of student studies at UIN Suska Riau by utilizing the concept of data mining based on the C4.5 decision tree classification method (D. , P. , & Hegde, 2016). To create a decision based on a sample of data, data mining uses the C4.5 algorithm as a decision tree classifier (Yuda, 2022). The method creates a decision tree that may be used to categorize new data based on the properties or features of the data after being provided a set of data representing objects that are already categorized (Maulida, R., 2020). The C4.5 algorithm's decision tree can be used to categorize both discrete and continuous data (Muhammad, L. J., Besiru Jibrin, M., Yahaya, B. Z., Mohammed Besiru Jibrin, I. A., Ahmad, A., & Amshi, 2020). The technique, which can deal with the problem of inadequate data, incorporates a single-pass pruning procedure to prevent overfitting. The C4.5 method is a well-known decision tree program in machine learning and is widely applied in practice (Haryoto, P. P., Okprana, H., & Saragih, 2021).

This research is expected to obtain findings that can have a great potential for a university to determine strategic policies for UIN Suska Riau in anticipating students who are predicted to graduate not on time or with the possibility of dropping out and paying special attention to students who are predicted to graduate on time. The results of the study will also show the factors that most influence the success rate of student studies. This study also provides important information regarding strategic plans for universities based on the interpretation of the research results

2. Literature Review

Data mining is the process of looking for hidden patterns in a set of data, which can be found in databases, dataware, or other information storage medium, as well as previously undiscovered knowledge (Gorunescu, 2011). Large-scale databases can be mined for relevant information and related knowledge using statistical, mathematical, artificial intelligence, and machine learning techniques (Dol, S. M., & Jawandhiya, 2023).Data mining classification techniques are employed to divide a set of objects into specific target groups and to forecast the nature of an item or piece of data based on the classes of items that are currently accessible (Oluwaseun, A., & Chaubey, 2019). Data mining classification techniques are utilized in the healthcare sector to find hidden trends and make wise choices. For instance, heart disease and chronic renal disease have both been predicted using data mining categorization approaches (Kunwar, V., Chandel, K., Sabitha, A. S., & Bansal, 2016). One of the techniques used in data mining for categorizing items is the decision tree classification methodology, which has been used to evaluate student performance (Teli, S., & Kanikar, 2015). Data mining also employs Naive Bayes, neural networks, and multilayer perceptrons as additional categorization methods (Dol, 2021).

Data mining has several characteristics, which are (Dabas, P., & Singh, 2021):

1. Data mining is related to the discovery of something hidden and certain data patterns that were not known before.
2. Data mining usually uses very large data sets.
3. Data mining is useful for making important decisions, especially in making strategies and policies.

The stages of data mining are (Adekitan & Salau, 2019; Baker, 2014):

1. Data cleaning is the process of taking out errors and skewed data. In the cleaning process, duplicate data is eliminated, inconsistent data is checked, and data inaccuracies are corrected, among other things.
2. Several data sources must be combined or integrated in data integration.
3. Data selection: Obtain pertinent data for analysis from the database.
4. Operations involving data transformation, data summary transformation, or data aggregation
5. Data mining is a crucial procedure where particular strategies or approaches are employed to extract hidden data patterns. Data mining methodologies, techniques, and algorithms can be very different. The objectives and overall KDD process really influence the choice of the best method or algorithm.
6. The information patterns produced by the data mining process need to be presented in a way that is simple to understand by interested parties for interpretation and evaluation. Checking whether the pattern or information found conflicts with previously known facts or assumptions is a part of this stage.

There are five types of analytical techniques that can be classified as part of data mining (Turban, 2005). In this study, the method used is classification. The following are five types of analytical techniques (Ukwuoma, C. C., Bo, C., Chikwendu, I. A., & Bondzie-Selby, 2019) :

1. Association.
2. Classification.
3. Clustering.
4. Estimation.
5. Predictions

In order to determine the class of an object whose label is unknown, classification is the process of identifying models or functions that explain or distinguish concepts or data classes (Dol, S. M., & Jawandhiya, 2023). The model itself can be a decision tree, a mathematical formula, or a neural network that represents a "if-then" rule. Learning and testing are the typical divisions of the classification process (D. , R., M. , & Giansiracusa, 2018; Dabas, P., & Singh, 2021). A rough model is created during the learning phase using some data that is already known for the data class (Yossy, E. H., Heryadi, Y., 2019). The model that has been created is then tested with additional data during the testing phase to ascertain its accuracy. This model can be used to

forecast unknown data classes if the accuracy is satisfactory (Patil, R., Salunke, S., Kalbhor, M., & Lomte, 2018)

3. Research Methods

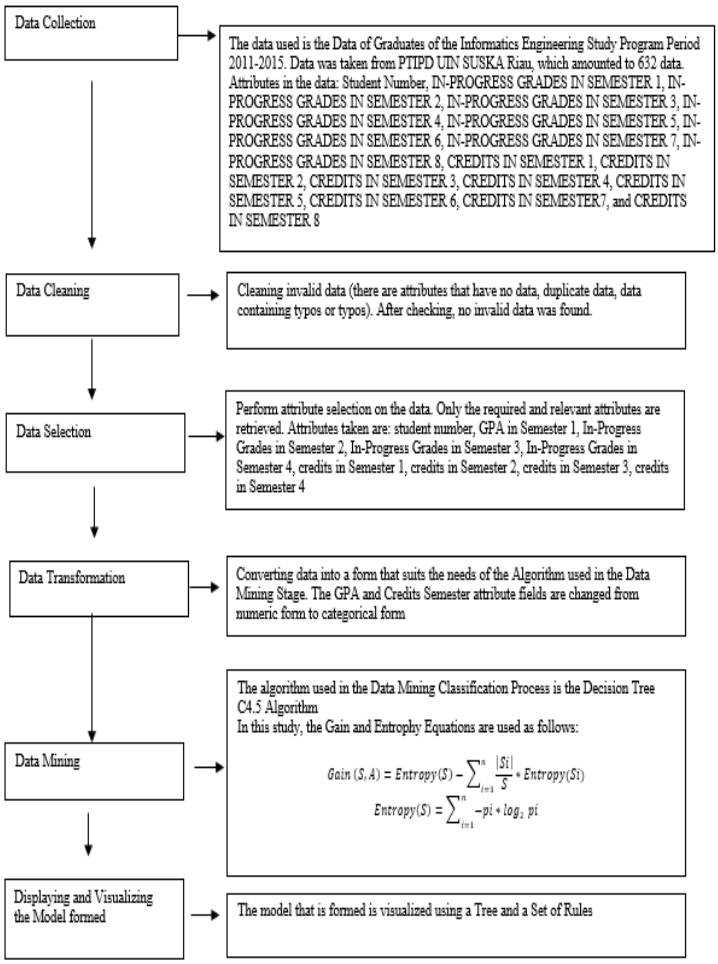


Fig. 1. Research methodology

The first step before building a model is collecting research data. The data collection will be carried out in the preprocessing phase (Sharma, A., Sharma, M. K., & Dwivedi, 2021; Syukri Mustafa, M., Rizky Ramadhan, M., & Thenata, 2017). The data will go through the stages of KDD: data cleaning, data integration, data selection, and data transformation. At the stage of KDD, data changes occur so that the attributes of data sets are good to be studied with the aim of understanding the content of records on academic data. Once this phase is completed, you will enter the data mining phase. In this phase, we will use the decision tree method C4.5. The preprocessed data will then be processed using the Decision Tree C4.5 method to produce a tree and rule model.

4. Results and Discussions

Research Data

The first step before making the model is collecting research data. The data that has been collected will be carried out in the pre-process stage. The preprocessed data will then be processed using the Decision Tree C4.5 method so as to produce a model in the form of a tree and rules. The data used in this study is the graduation data of Informatics Engineering students from 2011 to 2015, totaling 632 data records. The data was taken from the Center for Technology and Database of Sultan Syarif Kasim State Islamic University Riau. The data has the attributes of student number, in-progress grades in Semester 1, in-progress grades in Semester 2, in-progress grades in Semester 3, in-progress grades in Semester 4, in-progress grades in Semester 5, in-progress

grades in Semester 6, in-progress grades in Semester 7, in-progress grades in Semester 8, credits in Semester 1, credits in Semester 2, credits in Semester 3, credits in Semester 4, credits in Semester 5, credits in Semester 6, credits in Semester 7, and credits in Semester 8, GPA, and graduation time. These attributes will be managed and used as input and output classes in the Decision Tree method.

The data that has been collected will then be in the preprocessing stage and the data mining stage, where this stage is in the KDD stage. Data cleaning, data selection, and data transformation processes are carried out at the pre-process stage. Data transformation functions change data into a form that suits the needs at the data mining stage, so that it is easier to understand. Changes to data occur in In-Process Grades and GPA, which will be categorized into ranges in Table 1 and Table 2 below:

Table 1 - GPA Category

No	GPA	Transformation Result
1	> 3.50	High
2	3.00 - 3.50	Moderate
3	< 3.00	Low

Table 2 - Category of Length of Study

No	Credits	Transformation Results
1	> 17	Low
2	17-19	Moderate
3	20-21	High
4	22-24	Very High

After passing the pre-processing stage, the attributes that will be used are in-progress grades in semester 1, in-progress grades in semester 2, in-progress grades in semester 3, in-progress grades in semester 4, semester credit units for semester 1, semester credit units for semester 2, semester credit units for semester 3, semester credit units for semester 4, grade point average (GPA), and the output class attribute, namely the graduation attribute.

The results of the preprocessing stage will be used in the decision tree C4.5 method. The concept of the C4.5 decision tree is to convert the data into a decision tree that will produce decision rules. The C4.5 algorithm is a development of the ID3 algorithm. Here are the steps of the C4.5 algorithm:

a. Select attribute as root

The selection of attributes as roots is based on the highest Gain value of all existing attributes. To calculate the Gain value, the following Equation 1 is used:

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

Before obtain the Gain value, the Entropy value must be calculated. Equation 2 shows the equation to produce the Entropy value:

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

The following is the calculation results of Entropy and Gain:

Table 3 - Calculation Results of Entropy and Gain

	Attribute	Total	Punctual	Not Punctual	Entrophy	Gain
Total		568	56	512	0,46451494	
Credit Grade 1	Low	256	12	244	0,27297086	0,03498696
	Medium	203	18	185	0,43201646	
	High	109	26	83	0,7925907	
Credit Grade 2	Low	216	2	214	0,07584151	0,06738195
	Medium	280	33	247	0,52316976	
	High	72	21	51	0,87086447	

Credit Grade 3	Low	277	11	266	0,24096644	
	Medium	202	29	173	0,59349796	0,02944914
	High	89	16	73	0,6795852	
Credit Grade 4	Low	347	21		0,32950066	
	Medium	176	22		0,54356444	0,02607878
	High	45	13		0,86728162	
In-Process Grades in Semester 1	Low	4	0	4	0	
	Medium	564	56	508	0,466734338	
	High	0	0	0	0	0,00105849
	Very High	0	0	0	0	
In-Process Grades in Semester 2	Low	31	2	29	0,34511731	
	Medium	198	9	189	0,26676499	
	High	339	45	294	0,56491415	0,01552903
	Very High	0	0	0	0	
In-Process Grades in Semester 3	Low	7	0	7	0	
	Medium	60	2	58	0,2108423	
	High	200	12	188	0,32744492	0,01798613
	Very High	301	42	259	0,58301942	
In-Process Grades in Semester 4	Low	27	2	25	0,38094659	
	Medium	93	4	89	0,25593004	
	High	233	26	197	0,51948447	0,00653543
	Very High	225	24	201	0,48977901	
GPA	Low	147	0	147	0	
	Medium	360	36	324	0,46899559	0,06924171
	High	61	20	41	0,91273416	
GAIN MAX						0,06924171

Based on the table 1, the highest Gain value is obtained on the **GPA** attribute, which is **0.06924171.** For this reason, this attribute becomes the root of the tree that is formed.

b. Create a branch for each value

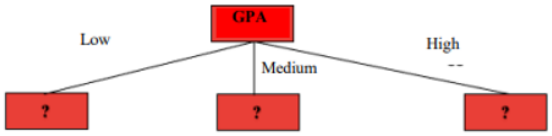


Fig. 2. Branch for each value

Branches are instances of the attributes that were selected as roots in the previous process. Because the attribute selected as the root is **GPA,** the branches are **Low, Medium**, and **High.**

c. Divide cases into branches

At this stage, the cases/data will be divided into branches according to the corresponding attributes and instances. List all cases on data with Low GPA, Medium GPA and High GPA.

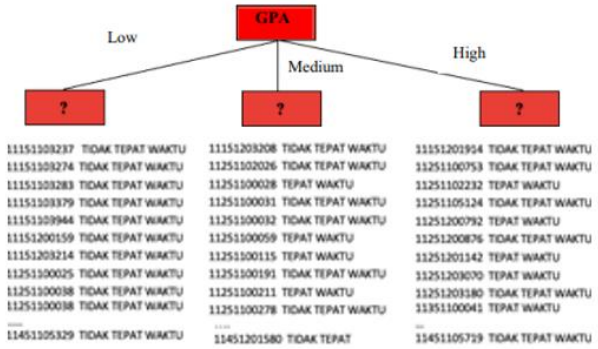


Fig. 3. Divide cases into branches (1)

Based on the results of the division of cases above in figure 3, students with low GPAs were declared to have passed all but not on time (all data had the same class), so the nodes in this branch were not developed anymore. Meanwhile, for Medium GPA and High GPA, there are students who graduated on time and not on time (data has different classes) so that the nodes in this branch need to be developed again in figure 4. Completion on time means that students graduate in the period up to semester 8 and are not fixed for more than semester 8.

Fig. 4. Divide cases into branches (2)

d. Repeat the process for each branch until all cases in the branch have the same class
The above process is repeated continuously so that all cases on the branch have the same class and a complete tree is formed. Here is the resulting Rule:

GPA = Low: NOT PUNCTUAL {PUNCTUAL=2, NOT PUNCTUAL=149}
GPA = Medium
| CREDIT GRADE 2 = Low
| | CREDIT GRADE 3 = Low: NOT PUNCTUAL {PUNCTUAL=1, NOT PUNCTUAL=66}
| | CREDIT GRADE 3 = Medium: NOT PUNCTUAL {PUNCTUAL=1, NOT PUNCTUAL=30}
| | CREDIT GRADE 3 = High
| | | CREDIT GRADE 1 = Low: PUNCTUAL {PUNCTUAL=1, NOT PUNCTUAL=1}
| | | CREDIT GRADE 1 = Medium: NOT PUNCTUAL {PUNCTUAL=0, NOT PUNCTUAL=2}
| CREDIT GRADE 2 = Medium
| | CREDIT GRADE 3 = Low
| | | CREDIT GRADE 1 = Low
| | | | IN-PROCESS GRADES 3 = Very High
| | | | | IN-PROCESS GRADES 4 = Very High: NOT PUNCTUAL {PUNCTUAL=0, NOT PUNCTUAL=5}
| | | | | IN-PROCESS GRADES 4 = Medium: PUNCTUAL {PUNCTUAL=1, NOT PUNCTUAL=1}
| | | | | IN-PROCESS GRADES 4 = High: NOT PUNCTUAL {PUNCTUAL=4, NOT PUNCTUAL=10}
| | | | IN-PROCESS GRADES 3 = Medium: NOT PUNCTUAL {PUNCTUAL=0, NOT PUNCTUAL=4}
| | | | IN-PROCESS GRADES 3 = High: NOT PUNCTUAL {PUNCTUAL=0, NOT PUNCTUAL=14}
| | | CREDIT GRADE 1 = Medium: NOT PUNCTUAL {PUNCTUAL=0, NOT PUNCTUAL=22}

| | | CREDIT GRADE 1 = High: NOT PUNCTUAL {PUNCTUAL=1, NOT PUNCTUAL=2}
| | CREDIT GRADE 3 = Medium
| | | IN-PROCESS GRADES 4 = Low: NOT PUNCTUAL {PUNCTUAL=0, NOT PUNCTUAL=4}
| | | IN-PROCESS GRADES 4 = Very High
| | | | IN-PROCESS GRADES 3 = Very High
| | | | | CREDIT GRADE 4 = Low: PUNCTUAL {PUNCTUAL=6, NOT PUNCTUAL=6}
| | | | | CREDIT GRADE 4 = Medium
| | | | | | IN-PROCESS GRADES 2 = Medium: NOT PUNCTUAL {PUNCTUAL=0, NOT PUNCTUAL=7}
| | | | | | IN-PROCESS GRADES 2 = High
| | | | | | | CREDIT GRADE 1 = Medium: NOT PUNCTUAL {PUNCTUAL=3, NOT PUNCTUAL=16}
| | | | | | | CREDIT GRADE 1 = High: PUNCTUAL {PUNCTUAL=2, NOT PUNCTUAL=2}
| | | | | CREDIT GRADE 4 = High: NOT PUNCTUAL {PUNCTUAL=0, NOT PUNCTUAL=3}
| | | | IN-PROCESS GRADES 3 = High: NOT PUNCTUAL {PUNCTUAL=0, NOT PUNCTUAL=9}
| | | IN-PROCESS GRADES 4 = Medium: NOT PUNCTUAL {PUNCTUAL=3, NOT PUNCTUAL=13}
| | | IN-PROCESS GRADES 4 = High
| | | | CREDIT GRADE 4 = Low
| | | | | IN-PROCESS GRADES 2 = Medium
| | | | | | CREDIT GRADE 1 = Low: PUNCTUAL {PUNCTUAL=1, NOT PUNCTUAL=1}
| | | | | | CREDIT GRADE 1 = Medium: NOT PUNCTUAL {PUNCTUAL=0, NOT PUNCTUAL=6}
| | | | | IN-PROCESS GRADES 2 = High: NOT PUNCTUAL {PUNCTUAL=0, NOT PUNCTUAL=22}
| | | | CREDIT GRADE 4 = Medium: NOT PUNCTUAL {PUNCTUAL=4, NOT PUNCTUAL=15}
| | | | CREDIT GRADE 4 = High: NOT PUNCTUAL {PUNCTUAL=0, NOT PUNCTUAL=3}
| | CREDIT GRADE 3 = High: NOT PUNCTUAL {PUNCTUAL=1, NOT PUNCTUAL=38}
| CREDIT GRADE 2 = High
| | CREDIT GRADE 4 = Low
| | | CREDIT GRADE 1 = Low: PUNCTUAL {PUNCTUAL=2, NOT PUNCTUAL=1}
| | | CREDIT GRADE 1 = Medium: PUNCTUAL {PUNCTUAL=2, NOT PUNCTUAL=0}
| | | CREDIT GRADE 1 = High: NOT PUNCTUAL {PUNCTUAL=1, NOT PUNCTUAL=3}
| | CREDIT GRADE 4 = Medium: NOT PUNCTUAL {PUNCTUAL=1, NOT PUNCTUAL=17}
| | CREDIT GRADE 4 = High: NOT PUNCTUAL {PUNCTUAL=0, NOT PUNCTUAL=2}
GPA = High
| IN-PROCESS GRADES 2 = Low: PUNCTUAL {PUNCTUAL=2, NOT PUNCTUAL=1}
| IN-PROCESS GRADES 2 = Medium
| | IN-PROCESS GRADES 3 = Very High: NOT PUNCTUAL {PUNCTUAL=0, NOT PUNCTUAL=6}
| | IN-PROCESS GRADES 3 = Medium: NOT PUNCTUAL {PUNCTUAL=1, NOT PUNCTUAL=2}
| | IN-PROCESS GRADES 3 = High: PUNCTUAL {PUNCTUAL=1, NOT PUNCTUAL=1}
| IN-PROCESS GRADES 2 = High
| | CREDIT GRADE 3 = Low: NOT PUNCTUAL {PUNCTUAL=0, NOT PUNCTUAL=2}

| | CREDIT GRADE 3 = Medium
| | | CREDIT GRADE 4 = Medium
| | | | CREDIT GRADE 2 = Medium: PUNCTUAL {PUNCTUAL=2, NOT PUNCTUAL=2}
| | | | CREDIT GRADE 2 = High
| | | | | IN-PROCESS GRADES 3 = Very High: NOT PUNCTUAL {PUNCTUAL=1, NOT PUNCTUAL=5}
| | | | | IN-PROCESS GRADES 3 = High: PUNCTUAL {PUNCTUAL=1, NOT PUNCTUAL=1}
| | | CREDIT GRADE 4 = High: NOT PUNCTUAL {PUNCTUAL=0, NOT PUNCTUAL=2}
| | CREDIT GRADE 3 = High
| | | CREDIT GRADE 1 = Medium: NOT PUNCTUAL {PUNCTUAL=1, NOT PUNCTUAL=4}
| | | CREDIT GRADE 1 = High
| | | | CREDIT GRADE 4 = Medium
| | | | | IN-PROCESS GRADES 4 = Very High: NOT PUNCTUAL {PUNCTUAL=0, NOT PUNCTUAL=4}
| | | | | IN-PROCESS GRADES 4 = High
| | | | | | IN-PROCESS GRADES 3 = Very High: NOT PUNCTUAL {PUNCTUAL=1, NOT PUNCTUAL=3}
| | | | | | IN-PROCESS GRADES 3 = High: PUNCTUAL {PUNCTUAL=1, NOT PUNCTUAL=1}
| | | | CREDIT GRADE 4 = High: PUNCTUAL {PUNCTUAL=7, NOT PUNCTUAL=6}

Testing Stage

a. Testing to get Influential Attributes

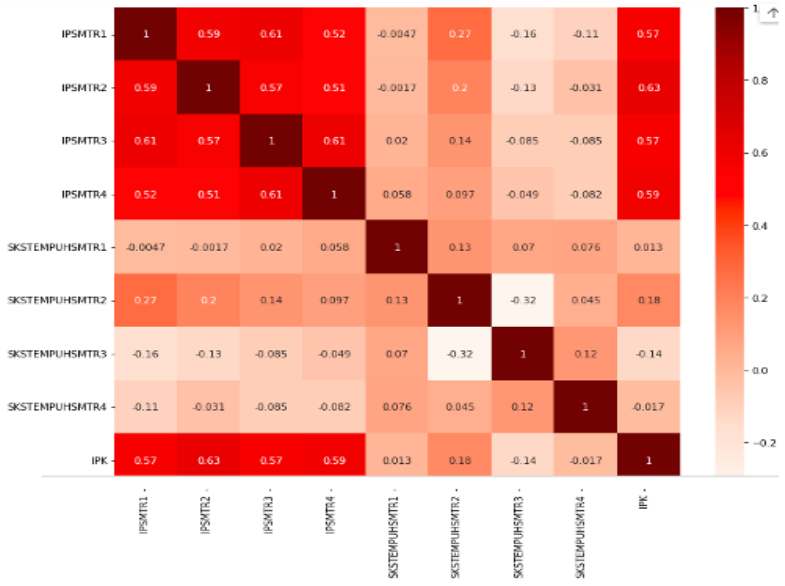
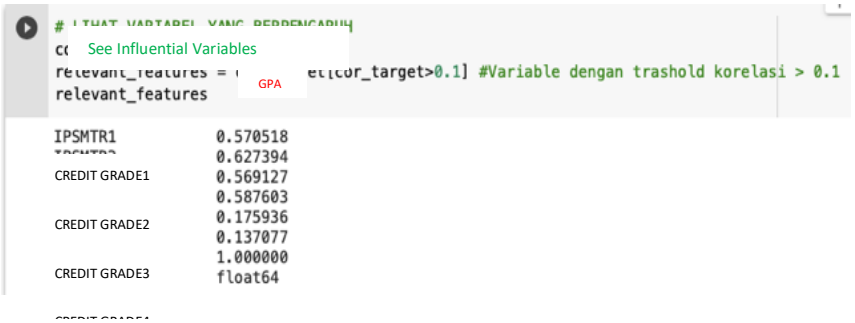


Fig. 5. Testing to get Influential Attributes

Based on the figure 5, the attributes that affect the continuity and success of the study of Informatics Engineering students at UIN SUSKA Riau are:



The attribute that most influences the continuity and success of the student study of Informatics Engineering UIN SUSKA Riau is the GPA attribute (Cumulative Achievement Index), followed by In-Progress Grades in Semester 2, In-Progress Grades in Semester 4, In-Progress Grades in Semester 1, In-Progress Grades in Semester 3, credits for Semester 2 and credits for semester 3.

b. Testing using comparison of training data and testing data
The test is carried out by changing the amount of training data and test data, namely 50% Training Data: 50% Test Data (50:50), 60:40, 70:30, 80:20, and 90:10. This test aims to see whether the amount of training data has an effect on the performance of the Decision Tree Method. The test results are as follows:

1. Comparison of 50:50

	Punctual	**Not Punctual**
Punctual	6	26
Not Punctual	28	256

$$Accuration = \frac{6 + 256}{6 + 26 + 28 + 256}\ x\ 100\% = 82{,}91\%$$

2. Comparison of 60:40

	Punctual	**Not Punctual**
Punctual	5	21
Not Punctual	12	215

$$Accuration = \frac{5 + 215}{5 + 21 + 12 + 215}\ x\ 100\% = 86{,}96\%$$

3. Comparison of 70:30

	Punctual	**Not Punctual**
Punctual	**1**	**20**
Not Punctual	**3**	**166**

$$Accuration = \frac{1 + 166}{1 + 20 + 3 + 166}\ x\ 100\% = 87{,}89$$

4. Comparison of 80:20

	Punctual	**Not Punctual**

Punctual	0	12
Not Punctual	3	112

$$Accuracy = \frac{0 + 112}{0 + 12 + 3 + 112} \; x \; 100\% = 88{,}19\%$$

5. Comparison of 90:10

	Punctual	**Not Punctual**
Punctual	0	5
Not Punctual	3	56

$$Accuracy = \frac{0 + 56}{0 + 5 + 3 + 56} \; x \; 100\% = 87{,}5\%$$

c. Conclusion of Testing

Based on the test scenarios that have been carried out, the following Tables and Graphs is obtained:

Table 4 - Testing Conclusion

Data Comparison	Accuracy
50 : 50	82,91 %
60 : 40	86,96 %
70 : 30	87,89 %
80 : 20	88,19 %
90 : 10	87,5 %

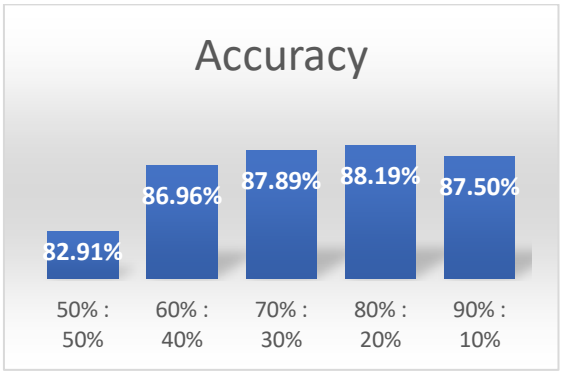


Fig. 6. The Graph of Accuracy Comparison

In the experiments, split data situations with the following ratios were applied: 90:10, 80:20, 70:30, 60:40, and 50:50. The accuracy of the maximum recognition rate when training data and test data are compared (80%:20%) is 88.19%. RapidMiner was utilized to test this research.

5. Conclusion

Understanding the findings of the research and the debate in relation to the study's goal The grade point average (GPA) is the subsequent factor that has the most impact on whether a student in informatics engineering graduates. Credits earned in semesters 1 and 2 do not have an impact on a student's ability to graduate from the Informatics Engineering study program. This is due to the fact that during semesters 1 and 2, the number of credits available to students is limited by the number of curricular packages. The following scenarios were used in the tests: 90:10, 80:20, 70:30, 60:40, and 50:50. When training data are compared to test data (80%:20%), the accuracy of the maximum recognition rate is 88.19%. Moreover, the comparison of training and test data

was found to have an effect on the accuracy of the introduction of the Decision Tree Method. In this case, the greater the use of training data, the better the recognition accuracy.

References

Adekitan, A. I., & Salau, O. (2019). The impact of engineering students' performance in the first three years on their graduation result using educational data mining. *Heliyon*, *5*(2), e01250. https://doi.org/10.1016/j.heliyon.2019.e01250

Asif, R., Merceron, A., Ali, S. A., & Haider, N. G. (2017). Analyzing undergraduate students' performance using educational data mining. *Computers and Education*, *113*, 177–194. https://doi.org/10.1016/j.compedu.2017.05.007

Asroni, A., Masajeng Respati, B., & Riyadi, S. (2018). Penerapan Algoritma C4.5 Untuk Klasifikasi Jenis Pekerjaan Alumni Di Universitas Muhammadiyah. *Semesta Teknika*, *21*(2), 158–165. https://doi.org/https://doi.org/10.18196/st.212222

Baker, R. S. (2014). Educational Data Mining: An Advance For Intelligent Systems In Education. *IEEE, Intelligent Systems.*, *29*(3), 78–82. https://doi.org/https://doi.org/10.1109/MIS.2014.42

BAN-PT. (2011). Buku II: Standar dan Prosedur. In *Badan Akreditasi Nasional Perguruan Tinggi …*. http://ppm.um-surabaya.ac.id/wp-content/uploads/2016/12/Buku-II-standar-prosedur.pdf

Burgos, C., Campanario, M. L., Peña, D. de la, Lara, J. A., Lizcano, D., & Martínez, M. A. (2018). Data mining for modeling students' performance: A tutoring action plan to prevent academic dropout. *Computers and Electrical Engineering*, *66*, 541–556. https://doi.org/10.1016/j.compeleceng.2017.03.005

Campagni, R., Merlini, D., Sprugnoli, R., & Verri, M. C. (2015). Data mining models for student careers. *Expert Systems with Applications*, *42*(13), 5508–5521. https://doi.org/10.1016/j.eswa.2015.02.052

D. , R., M. , & Giansiracusa, J. K. (2018). Machine Learning And Social Media To Mine And Disseminate Big Scientific Data. *2018 IEEE International Conference on Big Data*, 5312–5315. https://doi.org/https://doi.org/10.1109/BigData.2018.8622470

Dabas, P., & Singh, B. (2021). Analysis Of Data Mining Classification Techniques. *9th International Conference On Reliability, Infocom Technologies And Optimization, ICRITO 2021*, 1–3. https://doi.org/https://doi.org/10.1109/ICRITO51393.2021.9596174

Dharmawan, T., Ginardi, H., & Munif, A. (n.d.). Dropout Detection Using Non-Academic Data. *4th International Conference On Science And Technology (Icst). Yogyakarta, Indonesia.*, 2018. https://doi.org/10.1109/ICSTC.2018.8528619

Dol, S. M., & Jawandhiya, P. M. (2023). Classification Technique and its Combination with Clustering and Association Rule Mining in Educational Data Mining — A survey, Engineering Applications of Artificial Intelligence. *A Survey. Engineering Applications of Artificial Intelligence. Elsevier Ltd.*, *122*. https://doi.org/https://doi.org/10.1016/j.engappai.2023.106071

Dol, S. M. (2021). Use Of Classification Technique In Educational Data Mining. *International Conference on Nascent Technologies in Engineering, ICNET 2021 - Proceedings. Institute of Electrical and Electronics Engineers Inc.*, 1–7. https://doi.org/Doi: 10.1109/Icnte51185.2021.9487739

Fernandes, E., Holanda, M., Victorino, M., Borges, V., Carvalho, R., & Erven, G. Van. (2019). Educational data mining: Predictive analysis of academic performance of public school students in the capital of Brazil. *Journal of Business Research*, *94*(August 2017), 335–343. https://doi.org/10.1016/j.jbusres.2018.02.012

Gorunescu, F. (2011). Data Mining Concept, Models And Techniques. *Intelligent Systems Reference Library,Verlag Berlin Heidelb: Springer.*, *12*. https://doi.org/https://doi.org/10.1007/978-3-642-19721-5

Haryoto, P. P., Okprana, H., & Saragih, I. S. (2021). Algoritma C4.5 Dalam Data Mining Untuk Menentukan Klasifikasi Penerimaan Calon Mahasiswa Baru. *Terapan Informatika Nusantara*, *2*(5), 358–164.

Heredia, D., Amaya, Y., & Barrientos, E. (2015). Student Dropout Predictive Model Using Data

Mining Techniques. *IEEE Latin America Transactions*, *13*(9), 3127–3134. https://doi.org/Doi: 10.1109/Tla.2015.7350068.

Jain, A., Somwanshi, D., Joshi, K., & Bhatt, S. S. (2022). A Review: Data Mining Classification Techniques. *3rd International Conference On Intelligent Engineering And Management (Iciem)*, 636–642. https://doi.org/https://doi.org/10.1109/ICIEM54221.2022.9853036

Khan, A., Ghosh, S. K., Ghosh, D., & Chattopadhyay, S. (2021). Random wheel: An algorithm for early classification of student performance with confidence. *Engineering Applications of Artificial Intelligence*, *102*(July 2020), 104270. https://doi.org/10.1016/j.engappai.2021.104270

Kunwar, V., Chandel, K., Sabitha, A. S., & Bansal, A. (. (2016). Chronic Kidney Disease Analysis Using Data Mining Classification Techniques. *Proceedings of the 2016 6th International Conference - Cloud System and Big Data Engineering, Confluence 2016*, 300–305. https://doi.org/https://doi.org/10.1109/CONFLUENCE.2016.7508132

M. , A., & Rahman, A. M. (2016). *A Review On Data Mining Techniques And Factors Used In Educational Data Mining To Predict Student Amelioration.*

Massulloh, I., & F. (2020). (2020). Implementasi Algoritma C4.5 Untuk Klasifikasi Anak Berkebutuhan Khusus Di Ibnu Sina Stimulasi Center. *Eprosiding Sistem Informasi (Potensi)*, 136–144.

Maulida, R., & B. . (2020). Prediksi Kelulusan Mahasiswa Tepat Waktu Dengan Algoritma C4.5 Dengan Particle Swarm Optimization Pada Univeristas Xyz. *Journal Of Artificial Intelligence And Innovative Applications, Issn : 2716-1501.*, *1*(3), 138–144.

Muhammad, L. J., Besiru Jibrin, M., Yahaya, B. Z., Mohammed Besiru Jibrin, I. A., Ahmad, A., & Amshi, J. M. (. (2020). An Improved C4.5 Algorithm Using Principle Of Equivalent Of Infinitesimal And Arithmetic Mean Best Selection Attribute For Large Dataset. *10th International Conference On Computer And Knowledge Engineering (Iccke). Mashhad, Iran.*, 6–10. https://doi.org/https://doi.org/10.1109/ICCKE50421.2020.9303622

Natek, S., & Zwilling, M. (2014). Student data mining solution-knowledge management system related to higher education institutions. *Expert Systems with Applications*, *41*(14), 6400–6407. https://doi.org/10.1016/j.eswa.2014.04.024

Oluwaseun, A., & Chaubey, M. S. (2019). Data Mining Classification Techniques On The Analysis Of Student's Performance. *Global Scientific Journal*, *7*, 79–95.

Osman Hegazi, M., & Abugroon, M. A. (2016). The State Of The Art On Educational Data Mining In Higher Education. *International Journal Of Emerging Trends And Technology In Computer Science.*, *31*(1), 46–56.

Patil, R., Salunke, S., Kalbhor, M., & Lomte, R. (2018). Prediction System For Student Performance Using Data Mining Classification. Fourth. *International Conference On Computing Communication Control And Automation (Iccubea). Pune, India.*

Rasjid, Z. E., & Setiawan, R. (2017). Performance Comparison and Optimization of Text Document Classification using k-NN and Naïve Bayes Classification Techniques. *Procedia Computer Science*, *116*, 107–112. https://doi.org/10.1016/j.procs.2017.10.017

Sharma, A., Sharma, M. K., & Dwivedi, R. K. (2021). Improved Decision Tree Classification (IDT) Algorithm For Social Media Data. *Proceedings of the 2021 10th International Conference on System Modeling and Advancement in Research Trends, SMART 2021 (Pp. 155–157). Institute of Electrical and Electronics Engineers Inc.* https://doi.org/10.1109/SMART52563.2021.9676265

Sukhbaatar, O., Ogata, K., & Usagawa, T. (2018). Mining Educational Data To Predict Academic Dropouts: A Case Study In Blended Learning Course. *IEEE Region 10 Annual International Conference, Proceedings/TENCON, Institute of Electrical and Electronics Engineers Inc.*, 2205–2208. https://doi.org/10.1109/TENCON.2018.8650138

Syukri Mustafa, M., Rizky Ramadhan, M., & Thenata, A. P. (2017). Implementasi Data Mining untuk Evaluasi Kinerja Akademik Mahasiswa Menggunakan Algoritma Naive Bayes Classifier. *Citec Journal*, *4*(2), 151–162.

Teli, S., & Kanikar, P. (2015a). A Survey on Decision Tree Based Approaches in Data Mining. *International Journal Of Advanced Research In Computer Science And Software Engineering*, *5*(4), 613–617.

Teli, S., & Kanikar, P. . (2015b). A Survey On Decision Tree Based Approaches In Data Mining. International. *International Journal of Advanced Research in Computer Science and Software Engineering*, *5*(4), 613–617.

Turban, E. (2005). *Decision Support Systems and Intelligent Systems Edisi Bahasa Indonesia*. Andi.

Ukwuoma, C. C., Bo, C., Chikwendu, I. A., & Bondzie-Selby, E. (2019). Performance Analysis Of Students Based On Data Mining Techniques: A Literature Review. *4th Technology Innovation Management And Engineering Science International Conference (Times-Icon)*. https://doi.org/https://doi.org/10.1109/TIMES-iCON47539.2019.9024396

Utari, M., Warsito, B., & Kusumaningrum, R. (2020). Implementation Of Data Mining For Drop-Out Prediction Using Random Forest Method. *8th International Conference on Information and Communication Technology, ICoICT*, 1–5. https://doi.org/10.1109/ICoICT49345.2020.9166276

Yossy, E. H., Heryadi, Y., & L. (2019). Comparison Of Data Mining Classification Algorithms For Student Performance. *In TALE 2019 - 2019 IEEE International Conference on Engineering, Technology and Education. Institute of Electrical and Electronics Engineers Inc.*, 1–4. https://doi.org/https://doi.org/10.1109/TALE48000.2019.9225887

Yu, H., & Zhang, Z. Q. (2018). The Application of Data Mining Technology in Employment Analysis of University Graduates. *Proceedings - 17th IEEE/ACIS International Conference on Computer and Information Science, ICIS 2018*, 846–849. https://doi.org/10.1109/ICIS.2018.8466511

Yuda, O. &. (2022). Penerapan Penerapan Data Mining Untuk Klasifikasi Kelulusan Mahasiswa Tepat Waktu Menggunakan Metode Random Forest. *ATIN - Sains Dan Teknologi Informasi*, *8*(2), 122–131. https://doi.org/DOI: 10.33372/stn.v8i2.885

Yunus, M., Ramadhan, H., Aji, D. R., & Yulianto, A. (2021). Penerapan Metode Data Mining C4.5 Untuk Pemilihan Penerima Kartu. *Paradigma - Jurnal Komputer Dan Informatika*, *23*(2). https://doi.org/10.31294/p.v23i2.11395.