

NASKAH PUBLIKASI (*MANUSCRIPT*)

**IMPLEMENTASI KOMBINASI ALGORITMA *NAÏVE BAYES* DAN
ALGORITMA *ROUGH SET* UNTUK MEMPREDIKSI INDEKS PRESTASI
MAHASISWA**

***IMPLEMENTATION OF NAÏVE BAYES ALGORITHM AND ROUGH SET
ALGORITHM TO PREDICT GRADE POINT AVERAGE (GPA)***

Muhammad Febri Maulana ¹, Rofilde Hasudungan ²



DISUSUN OLEH :

MUHAMMAD FEBRI MAULANA

1811102441059

**PROGRAM STUDI S1 TEKNIK INFORMATIKA
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS MUHAMMADIYAH KALIMANTAN TIMUR**

2023

Naskah Publikasi (*Manuscript*)

**Implementasi Kombinasi Algoritma *Naïve Bayes* dan Algoritma *Rough Set*
untuk Memprediksi Indeks Prestasi Mahasiswa**

***Implementation of Naïve Bayes Algorithm and Rough Set Algorithm to Predict
Grade Point Average (GPA)***

Muhammad Febri Maulana ¹, Rofilde Hasudungan ²



Disusun Oleh :

Muhammad Febri Maulana

1811102441059

**PROGRAM STUDI S1 TEKNIK INFORMATIKA
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS MUHAMMADIYAH KALIMANTAN TIMUR
2023**

HALAMAN PENGESAHAN

IMPLEMENTASI KOMBINASI ALGORITMA *NAÏVE BAYES* DAN ALGORITMA *ROUGH SET* UNTUK MEMPREDIKSI INDEKS PRESTASI MAHASISWA

NASKAH PUBLIKASI


DISUSUN OLEH:

MUHAMMAD FEBRI MAULANA

1811102441059

Telah melaksanakan ujian skripsi dan dinyatakan lulus,
Pada tanggal 6 Januari 2023

Dosen Pembimbing



Rofie Hasudungan, S.Kom., M.Sc
NIDN. 1107048601

Penguji



Wawan Joko Pranoto, S.Kom., M.TI
NIDN. 1102057701

Dekan



Prof. Dr. Sarjito, MT., Ph.D
NIDN. 0610116204

Ketua Program Studi



Asslia Johar Latipah, S.Kom., M.Cs
NIDN. 1124098902

Implementation of Naïve Bayes Algorithm and Rough Set Algorithm to Predict Grade Point Average (GPA)

Muhammad Febri Maulana ^{1*}, Rofilde Hasudungan ²

^{1,2} Universitas Muhammadiyah Kalimantan Timur, Samarinda, Indonesia

* Corresponding Email: febrimaulana073.fm@gmail.com

Abstract – Grade Point Average (GPA) that are below standard can cause various problems that cause a chain effect that causes a decrease in low GPA in the coming semesters. To maintain good student learning next GPA, it is necessary to have an approach to extract important information about the factors behind students. One way to dig up information is to make predictions by applying data analysis using data mining. In data analysis, this research used data that totaled 306 student data consisting of 240 students who had a GPA and 66 students who did not yet have a GPA. The use of the rough set algorithm method as an attribute selection method for classification produces 4 attribute choices out of 14 attributes. To implement the classification, divide the data into 70% data training and 30% data testing on the model using all attributes and the model using attribute selection. In its implementation, it uses two models, namely a model that uses all the attributes to produce an accuracy of 83,58%. Meanwhile, the accuracy of the model using attribute selection resulted in an increase of 88,06%. So that predictions can be made on student data that does not yet have a GPA which results in a high GPA 47 students, a moderate GPA 4 students, and a low GPA 7 students.

Keywords: Naïve Bayes; Rough Set; Grade Point Average

Submitted: 29 October 2017 - Revised: 23 November 2017 - Accepted: 19 December 2018

1. Introduction

Excellent and quality universities can be seen from the achievements of their students, one of which is in the form of the first semester student achievement index [1], [2]. Student achievement index that is below the standard can cause various problems that cause a chain effect causing a decrease in low performance in future semesters [3]. Low student performance index is caused by many factors behind these students such as mood, time management, relationships with family, lecturer explanations, living atmosphere, activities other than lectures, environmental adaptation, parental attention, socialization, class atmosphere, and ability to capture material [4]. So that to keep student achievement good, an approach is needed to explore important information about the factors behind students. To process information can use a new method that is useful for university management as monitoring the learning process, and taking the necessary policies to improve student achievement index [5]. Therefore, predicting student achievement index is important as for some studies using data mining techniques such as those conducted by Firdaus [6], Hasudungan [8], Tommy and Mahmud [7], Hasudungan and Pranoto [9], Timur and Beatrix [32], Rolansa, et al [21], Sonang, et al [27] Desiani, et al [3], Alverina, et al [1].

course is one of the main things that is important for the running of the lecture activity process. In addition to the high willingness to learn from students, lecturers also have an important role in delivering lecture material that can be understood by students. Especially with regard to how a lecturer conveys the content of lecture material.

Every lecturer who provides material has a different learning method for their students. Differences in the way lecturers teach greatly affect the results that students will get when the lecture process takes place. In addition, several factors that affect the level of student understanding such as learning readiness, learning order and so on also greatly affect student understanding, the existence of students who understand and do not understand greatly impacts the success of the learning process, therefore a prediction of the level of student understanding is very important.

Naive Bayes itself is one of the methods that has advantages such as speed and accuracy in classifying data. Naive Bayes is a classification method that is very effective and also efficient in testing on large datasets to determine past patterns and find functions that will become future data assessment patterns. So this algorithm aims to classify data in certain classes.

Naïve Bayes also has a drawback, namely when certain parameters are empty or have no value and Naïve Bayes excludes them, this affects the quality of the results issued, so a method is needed to select the best parameters, namely rough sets that can reveal hidden

patterns in the data and help predict [1].

Rough set method is a method that can deal with vague and inconsistent data. In Hasudungan, et al's research [2] previously used rough sets for attribute selection on naïve bayes which used rough sets to select attributes for predicting student achievement, the analysis results showed that the proposed model had an accuracy rate of 77.5%, and a lower result of 69%. The results of this study indicate that learning methods have a positive and significant effect on course understanding.

Therefore, in this research the author will use rough sets to improve the accuracy of naïve bayes to select the best features and eliminate redundant features, and use this method to improve the performance (accuracy) of naïve bayes in predicting the level of student understanding of the course.

2. Related Works

Applying the C4.5 algorithm to predict the level of student understanding of courses [6]. Using the C4.5 algorithm to determine the classification of students' level of understanding of programming language courses [4]. Predicting the level of student understanding of the course [8]. Using Naïve Bayes to predict student study period based on factors related to student academics [9]. Naïve Bayes to predict the level of student understanding of the data structure algorithm course [3]. Analysis of the naïve bayes method to predict the level of student understanding of courses based on sitting position [2]. Using the Decision Tree algorithm to predict academic achievement based on socio-economics, motivation, the role of lecturers, discipline and learning outcomes [6]. Using Decision Tree C4.5 to get a decision tree model with variables or attributes of the achievement index that affect the predicate of student graduation [8]. Using Naïve Bayes for predicting student academic grades by utilizing probability calculations and statistics of previous data to predict future data based on previous data [8]. Using Decision Tree and Naïve Bayes the accuracy results of the Naive Bayes method remain the largest, although the increase in accuracy value after optimization is lower than the Decision Tree method [9]. Implementation of correlation based feature selection (CFS) to increase the accuracy of the C4.5 algorithm in predicting student academic performance based on a learning management system [10].

3. Naïve Bayes

Naïve Bayes is a classification algorithm based on the Bayesian theorem in statistics and can be used to predict the probability of belonging to a class. Naïve Bayes calculates the posterior probability value $P(H|X)$ using the probabilities $P(H)$, $P(X)$, and $P(X|H)$ where the X value is the testing data whose class is unknown. The H value is the hypothesis of X data that is a more specific class. The value of $P(X|H)$, also called likelihood, is the

probability of hypothesis X based on condition H . The value of $P(H)$, also called prior probability, is the probability of hypothesis H . While the value of $P(X)$, also called predictor prior probability, is the probability of X [8].

$$P(H | X) = \frac{P(X | H).P(H)}{P(X)} \quad (1)$$

X is the data with unknown class, H is the hypothesis that the data is a specific class $P(H|X)$ is the probability of hypothesis H based on the condition of X (a posteriori probability), $P(H)$ is the probability of hypothesis H (prior probability), $P(X|H)$ is the probability of X based on the condition in hypothesis H , and $P(X)$ is the probability of X .

4. Rough Set (MDA)

Rough Set theory was first introduced by Pawlak, who stated that Rough set is one of the mathematical methods for handling inconsistent and vague data[18]. In addition, the advantage of this method is that it does not require parameters or input because the information related to the data is taken from the data itself [18] As well as Pawlak proposed that rough set theory is founded on the assumption that with each member of the universe of discourse we connect some information. The rough set concept is a new mathematical technique to overcome vagueness, imprecision, and uncertainty [18].

Rough set has various forms of development, one of which is Maximum dependency attributes is an attribute selection-based rough set that can find dependencies between attributes and can reduce excessive attributes. In reducing excessive attributes can use a way to calculate the dependency between attributes with other attributes based on the maximum dependency value of attributes on data [19].

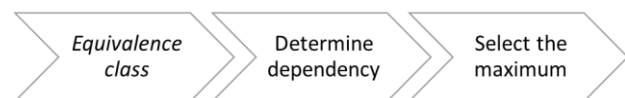


Figure 1. Rough set MDA solving scheme

The following information is based on the maximum dependency attributes completion scheme as a method of calculating attribute dependencies:

4.1. Equivalence class

Equivalence class is the first stage in applying the MDA rough set algorithm to find the equivalence class on each attribute of the set U by using the indiscernibility relation on each attribute with the definition of $S = (U, A, V, f)$ being an information system, D and C being part of A . If D is fully dependent on C , then $\alpha B(X) \leq \alpha C(X)$, for all members $X \subseteq U$. Based on that definition, $IND(C) \subseteq IND(D)$ can therefore be applied to equation 2.

$$D(X) \subseteq C(X) \subseteq X \subseteq C(X) \subseteq DX \quad (2)$$

4.2. Determine dependency

Determine dependency is the next stage in determining the maximum dependency of attribute α^j in relation to all attributes α_i , but $\alpha^j \neq \alpha_i$. The application can use equation 3..

$$D(\underline{R}(X), \bar{R}(X)) = 1 - \frac{|\underline{R}(X) \cap \bar{R}(X)|}{|\underline{R}(X) \cup \bar{R}(X)|}$$

$$= 1 - \frac{|\underline{R}(X)|}{|\bar{R}(X)|}, = 1 - \alpha R(X) \quad (3)$$

4.3. Select the maximum

Select the maximum is the stage of selecting the maximum dependency of each attribute. The maximum attribute dependency level can be determined based on the more attributes that have the same value will get the dependency value. With the definition of $S = (U, A, V, f)$ being an information system, $S = (U, A, V, f)$ being an information system and $C1, C2, \dots, Cn$ so that D becomes part of A . If $C1 \Rightarrow k1 D, C2 \Rightarrow k2 D, \dots, Cn \Rightarrow kn D$, where $kn \leq kn-1 \leq \dots \leq k2 \leq k1$, so that $\alpha D(X) \leq \alpha Cn(X) \leq \alpha Cn-1(X) \leq \dots \leq \alpha C2(X) \leq \alpha C1(X)$ For every $X \subseteq U$. As contained in equation 4.

$$\alpha D(X) \leq \alpha Cn(X)$$

$$kn \leq kn-1 \leq \dots \leq k2 \leq k1 \quad (4)$$

$$[x]Cn \subseteq [x]Cn-1$$

5. Evaluation

Before conducting the evaluation, first prepare the data processing. Data processing is a stage in data mining that is used to process data so that it can be run in the classification process. Data processing has the aim of reducing data, finding relationships between data, normalizing data, removing outliers and extracting knowledge to do this requires several techniques, namely Data cleaning, Data integration, Data transformation, and Data reduction. After that, the evaluation stage is ready to run.

Evaluation is the measurement of an algorithm's performance against data. The evaluation process assesses interesting patterns or prediction models whether they have fulfilled the initial hypothesis or not. Evaluation of classification-type data mining is done by testing the prediction process of object truth. Confusion matrix is one way that is often used in the process of evaluating classification data mining models by predicting object truth. The testing process utilizes a confusion matrix that places the prediction class at the top of the matrix then the observed source is placed on the left side of the matrix. Each cell of the matrix contains a number that displays the actual number of cases of the class being observed (Muslim et al., 2019).

Table 2.5 describes an example of a classification process confusion matrix.

Table 1
Confusion Matrix

	Action True	Action False
Predict True	TP (True Positive)	FP (False Positive)
Predict False	FN (False Negative)	TN (True Negative)

To measure the accuracy of the model, equation 5 is used to calculate the accuracy results.

$$\frac{\text{Jumlah Prediksi yang Benar}}{\text{Jumlah Prediksi yang dilakukan}} = \frac{TP+TN}{TP+FP+FN+TN} \quad (5)$$

As for calculating the error rate, it can be defined by equation 6 as follows.

$$\frac{\text{Jumlah Prediksi yang Salah}}{\text{Jumlah Prediksi yang dilakukan}} = \frac{FP+FN}{TP+FP+FN+TN} \quad (6)$$

And to calculate the accuracy (precision) of measuring data that has been predicted to be positive with the correct and incorrect reality can use equation 2.9. Meanwhile, to calculate the sentivity (recall) of many successful data when predicted by the comparison of all data that is in fact positive can use equation 2.10 as follows.

$$\text{Ketepatan} = \frac{TP}{TP + FP} \quad (7)$$

$$\text{Sentivitas} = \frac{TP}{TP + FN} \quad (8)$$

6. Metodology

6.1. Flowchart

This research will focus on comparing or comparing the prediction of student achievement index using the naïve bayes algorithm using rough sets and without using rough sets with the flow stages as shown in Figure 1.

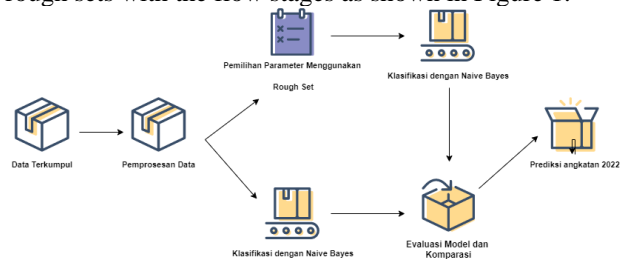


Figure 1. Research of flowchart

In the initial stage of the research, processing the collected data will be carried out data processing where this step performs data cleaning and data transformation, data cleaning is carried out to remove incomplete or

empty data and duplicated data that has the same content, while data transformation will group numerical data and convert it to categorical so that it can be processed when running attribute selection and modeling for classification. classification modeling will be divided into two, namely, the first will use all data with the naïve bayes algorithm and then evaluate the final results, the second by running the rough set algorithm first to select the best attributes to be used. then classify with the naïve bayes algorithm, then evaluate the final results. Then perform classification with the naïve bayes algorithm, then evaluate the final results. The last step is to compare the final results of the first model without selecting attributes and the second model that uses attribute selection with the rough set algorithm.

6.2. Data collection

In this study, data collection used a questionnaire distributed to students of the 2020 and 2021 faculties of science & technology majoring in informatics engineering at Universitas Muhammadiyah Kalimantan Timur (UMKT) using google form which can be seen in appendix 1. The data obtained uses attributes regarding student background factors obtained from [3][8].

Table 2
Attribute table used in research

Simbol	Nama Atribut	Deskripsi	Keterangan
A1	Student Name	Contains the identity of the student's name	Full name of student
A2	NIM	Contains student identification number	Student identification number
A3	Gender	Nominal data, contains 2 categories, male and female	Student gender
A4	Place of Residence	Ordinal data, contains 4 categories, with family, with parents, boarding house, and contract with friends.	Student status of residence
A5	Wedding	Nominal data, contains 2 categories, married and unmarried	Marital status of students
A6	Jobs	Nominal data, containing 2 categories, yes, and no	Status of students working while studying
A7	Previous Education	Nominal data, Contains 3 categories, SMA, SMK, and MA	Type of education of students before entering college
A8	Previous Education Status	Nominal data, Contains 2 categories of public, and private	Previous student's school status
A9	Marital Status of Parents	Nominal data, Contains 2 categories, married and separated	Marital status of parents
A10	Father's		Educational

	Education	Ordinal Data, Contains 7 categories of not in school, completed elementary / MI, completed junior high school / MTs, completed vocational / high school / MA, and completed college S1, S2, S3	status of student's father
A11	Mother's Education	Ordinal Data, Contains 7 categories of not in school, completed elementary / MI, completed junior high school / MTs, completed vocational / high school / MA, and completed college S1, S2, S3	Educational status of student's mother
A12	Father's occupation	Nominal data, contains 5 categories: not working, laborer, private, civil servant, and non-civil servant.	Employment status of student's father
A13	Mother's Occupation	Nominal data, contains 5 categories of laborers, private sector, civil servants, non-civil servants, and housewives.	Employment status of student's mother
A14	Many Family Members	Fill in the number of family members	Number of student family members in one house
A15	Provisional Grade Point Average	Ratio data, student achievement index	Semester student achievement index

Table 3
Student achievement index category table

GPA	Category
< 2.5	Rendah
2.5 < 3	Cukup
3 >	Tinggi

Table 4
Family size category table

Number of family members	Category
< 4	Kecil
5 < 6	Sedang
7 >	Besar

7. Hasil dan pembahasan

7.1. Research data

In this study, we will use 306 total informatics engineering student data obtained from 240 informatics

engineering student data from the class of 2020 to 2021 which will be used to calculate the accuracy of algorithm performance and 66 data on students in the class of 2022 majoring in informatics engineering, faculty of science and technology at Muhammadiyah University of East Kalimantan which will be used as predictions

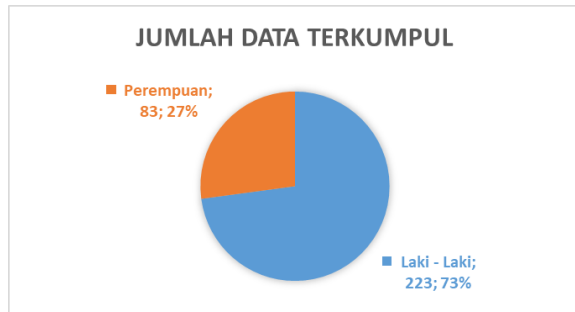


Figure 2. Total Data Collected

Data collection uses a google form questionnaire distributed to each informatics student whatsapp class group, and also collects data directly to each class using a printed questionnaire. The questionnaire distributed uses 14 attributes which can be seen in table 3.1. While there is 1 attribute, namely the first semester grade point average of informatics engineering students totaling 435 student data consisting of the 2020 batch of 195 data and 2021 batch of 240 data, which is obtained from the Bureau of Academic Administration (BAA) following the example in table 5.

Table 5
GPA data collection

Nama	Indeks Prestasi
Ahlada	3.3750
Salsabila	3.4750
Suryadi	3.5000
...	...
Anton	2.9750

7.2. Pemrosesan data

After the collected data is ready, the next step is to process the data so that it can be used in the attribute selection process and the classification process. Because the data must be in a state where there must be no empty data and categorical data types. The way to process data goes through three stages, namely data integration, data cleaning and data transformation.

At this stage, it will combine the student achievement index data which will be combined with the data obtained from the questionnaire into one unitary data referring to the Nim attribute column. So that it can be combined as an example in the 2020 and 2021 batch data seen in table 4.2 and the 2022 batch data does not have an achievement index attribute can be seen in table 6

Table 6
Data collection from google form

Nim	Jenis Kelamin	Tempat Tinggal	Status Mahasiswa	...	Jumlah Keluarga
2211102*****	Laki - Laki	Bersama	Belum Menikah	...	1
2211102*****	Laki - Laki	Bersama	Belum Menikah	...	6
2211102*****	Laki - Laki	Bersama	Belum Menikah	...	1
2211102*****	Laki - Laki	Bersama	Belum Menikah	...	3
...
2211102*****	Laki - Laki	Kos	Belum Menikah	...	4

In the data cleaning stage, we will delete data with missing value status, incomplete data, and duplicated data so that later it can be used for the attribute selection and classification process. The next step is to delete the data whose results are in table 7.

Table 7
Total data collection

Data Mahasiswa	Tahun 2020-2021	Tahun 2022
<i>Terduplikasi</i>	8 data	1 data
<i>Missing value</i>	10 data	7 data
<i>Total</i>	18 data	8 data

Based on the total data of 306 student data, it is cleaned so that it becomes a total of 280 student data. Consisting of 222 data on students from 2020 to 2021, and 58 data on students from 2022.

The data transformation stage will change the numeric data type into categorical data so that it can be used in the attribute selection and classification process. Data transformation is carried out by changing the value of the family size attribute according to the rules for dividing the family size category in table 3.3 and the value of the first semester student achievement index attribute according to the rules of the university academic guide in table 4. So that it can be applied to sample data as in table 6. The data will be categorized through the Microsoft excel application found in appendix 5.

Table 9
Data integration from size family

Nim	Jenis Kelamin	Tempat Tinggal	Status Mahasiswa	...	Jumlah Keluarga
2211102*****	Laki - Laki	Bersama	Belum Menikah	...	Kecil
2211102*****	Laki - Laki	Bersama	Belum Menikah	...	Sedang

	Laki	Keluarga	Menikah		
2211102*****	Laki -	Bersama	Belum	...	Kecil
	Laki	Keluarga	Menikah		
2211102*****	Laki -	Bersama	Belum	...	Kecil
	Laki	Keluarga	Menikah		
...
2211102*****	Laki -	Kos	Belum	...	Kecil
	Laki		Menikah		

7.3. Pemilihan atribut dengan algoritma Rough set

In this study, based on 226 data from the 2020 and 2021 batch students, a consistent value equal to 1 was obtained, which stated that the data was very consistent. The value is calculated using the help of the rst-tools python library and the Google Colab web application which can be used to write programs, while the programming language used is python programming language.

Table 8
Tabel kategori jumlah anggota keluarga

Sign	Maximum Dependency
A5	1.0
A6	0.03982300884955752
A3	0.022123893805309734
A4	0.004424778761061947

Based on the dependency value obtained, attribute reduction is carried out so that the results of the attributes amount to 4 condition attributes, from the initial attributes totaling 14. The best condition attributes are (A5) Student marital status, (A6) Students studying while working, (A3) Gender, and (A4) Student residence status. The results of this attribute selection will be used in the classification while the remaining attributes will be deleted because they are not used.

7.4. Naïve bayes implementasi

In performing classification using the naïve bayes method using the help of data analysis applications, namely Rapiminer. In the process, it will divide into two models as in the research flow picture 3.1, where the first model will directly classify using all 15 attributes. While the second model will perform classification using attributes that have been selected using rough sets. Before classifying, the 2020 and 2021 batch student data will be divided into two parts with a percentage of 70% training data totaling 155 data and 30% testing data totaling 67 data.

After dividing the data, the next step is to calculate the probability value on the decision attribute labeled "High", "Fair", and "Low". Based on training data totaling 155 data. Obtained decision attributes labeled "High" as much as 120 data, "Enough" as much as 14 data, and "Low" as much as 21 data.

$$P(\text{Indeks} | \text{Tinggi}) = \frac{120}{155} = 0,774193548$$

$$P(\text{Indeks} | \text{Cukup}) = \frac{14}{155} = 0,090322581$$

$$P(\text{Indeks} | \text{Rendah}) = \frac{21}{155} = 0,135483871$$

The next step is to calculate the value of each supporting attribute in the training data using equation 2.1. One example of calculating the probability value on the gender attribute labeled "Male", and "Female" based on the student achievement index labeled "High", "Fair", and "Low".

$$P(\text{Laki} - \text{Laki} | \text{Tinggi}) = \frac{86}{120} = 0,716666667$$

$$P(\text{Laki} - \text{Laki} | \text{Cukup}) = \frac{10}{14} = 0,714285714$$

$$P(\text{Laki} - \text{Laki} | \text{Rendah}) = \frac{19}{21} = 0,904761905$$

$$P(\text{Perempuan} | \text{Tinggi}) = \frac{34}{120} = 0,283333333$$

$$P(\text{Perempuan} | \text{Cukup}) = \frac{4}{14} = 0,285714286$$

$$P(\text{Perempuan} | \text{Rendah}) = \frac{2}{21} = 0,095238095$$

The next step calculates all the values obtained on each attribute that will be used during classification using equation 2.1 which is applied to the 1st testing data.

$$P(\text{Tinggi}) = 0,716666667 \times 0,236263736 \times \dots \times 0,774193548 = 0,99689972548$$

$$P(\text{Cukup}) = 0,714285714 \times 0,401098901 \times \dots \times 0,090322581 = 0$$

$$P(\text{Rendah}) = 0,904761905 \times 0,115384615 \times \dots \times 0,135483871 = 0,00310027339$$

Based on the results of the above calculations, it can be seen that the probability value of the "High" achievement index is greater than the others. So that the prediction of the 1st testing data can be said to get a "High" index of achievement.

7.5. Evaluation

Evaluation processing uses the help of a confusion matrix table applied to 67 testing data and the first two models use all attributes and the second model uses attribute selection. In calculating accuracy can use equation 7. And for the comparison of the accuracy of the two models is in Figure 1.

Table 10
Confusion matrix table of the first model

Confusion Matrix	Facts
------------------	-------

		Tinggi	Rendah	Cukup
Prediksi	Tinggi	52	1	2
	Rendah	2	3	2
	Cukup	1	3	1

$$\text{Accuracy} = \frac{(52+3+1)}{(52+1+2+2+3+2+1+3+1)} \times 100\% = \frac{56}{67} = 0,8358 \times 100\% = 83,58\%$$

the first model uses all available attributes to perform classification without attribute selection.

Table 11
Confusion matrix table of the second model

Confusion Matrix		Facts		
		Tinggi	Rendah	Cukup
Prediksi	Tinggi	53	1	3
	Rendah	2	6	2
	Cukup	0	0	0

$$\text{Accuracy} = \frac{(52+6)}{(53+1+3+2+6+2+0+0+0)} \times 100\% = \frac{59}{67} = 0,8358 \times 100\% = 88,06\%$$

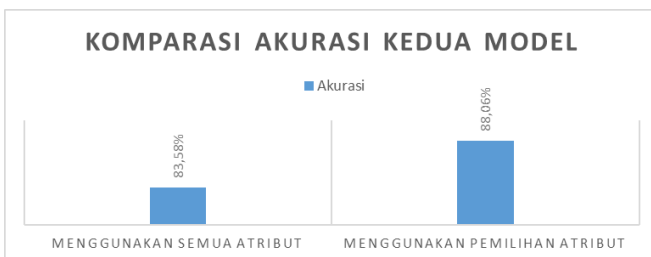


Figure 3. Classification model accuracy

Based on the evaluation, it is known that the use of the rough set method can improve the prediction results on naïve bayes classification from 83.58% to 88.06% accuracy. So that the use of rough sets and naïve bayes is very good and can be applied very well, and can be used in predicting the first semester grade point average of 2022 students.

7.6. Implementasi

In this stage, we will predict the grade point average of the 2022 batch of students who do not have decision attributes or classes using the naïve bayes classification method with the selection of rough set attributes so that the steps will be as in the previous stage, namely using the same attributes as found in table 8 then apply the naïve bayes classification with the results found in 12. This classification model is used because the results of the attribute selection accuracy comparison are higher than using all attributes to perform classification.

Table 12
Result of GPA 2022

Status	Status	Peluang	Peluang	Peluang	Hasil
Tempat	Menikah	Tinggi	Cukup	Rendah	Prediksi
Tinggal					
Laki -	Bersama	0,9794	0,0092	0,0112	Tinggi
Laki	Orang Tua				
Laki -	Bersama	0,9832	0,4397	0,0166	Tinggi
Laki	Keluarga				
Laki -	Bersama	0,9832	0,4397	0,0166	Tinggi
Laki	Keluarga				
Laki -	Bersama	0,8055	0,0040	0,1903	Tinggi
Laki	Keluarga				
...
Laki -	Kos	0,0884	0,7020	0,2094	Rendah
Laki					



Figure 4. Predicted results of the class of 2022

Based on the classification results using attribute selection, there are 47 students who have the potential to get a high achievement index, 4 students get a sufficient or standard achievement index, and 7 students get a low achievement index from a total of 58 data contained in Figure 4.

8. Kesimpulan

A student's grade point average can cause a chain of problems in the coming semester. Therefore, predicting student performance index is very important. In predicting student performance index, we can implement data analysis methods based on student background factors by utilizing rough set and naïve bayes algorithms in predicting student performance index. So that in the research analysis, the following conclusions can be obtained:

- 1) This research collected all data on informatics engineering students from 2020 to 2022 which amounted to 280 data, consisting of 222 data on students from 2020 to 2021 and 58 data on 2022 students. The data for this study were obtained from a

questionnaire using google form.

- 2) In selecting attributes using the rough set algorithm method, 4 attributes are obtained, namely gender, student residence status, student employment status, and student marital status.
- 3) Evaluation results in predicting student achievement index using data from 2020 to 2021 by dividing 70% training data into 155 data and 30% testing data into 67 data in the naïve bayes algorithm resulted in an accuracy of 83.58%. Meanwhile, the combination of the naïve bayes algorithm and the rough set algorithm increased by 88.06%.
- 4) The results of the implementation using a combination model of the naïve bayes algorithm and the rough set algorithm in predicting student achievement index on the data of informatics engineering students class of 2022 resulted in 47 students potentially getting a high achievement index, 7 students potentially getting a low achievement index, and 4 students potentially getting a standard or sufficient achievement index.

References

- [1] Alverina, D., Chrismanto, A. R., & Santosa, R. G. (2018). Perbandingan Algoritma C4.5 dan CART dalam Memprediksi Kategori Indeks Prestasi Mahasiswa. *Jurnal Teknologi Dan Sistem Komputer*, 6(2), 76–83. <https://doi.org/10.14710/jtsiskom.6.2.2018.76-83>
- [2] Desember, J., Rizqiyani, V., Mulwindi, A., & Mahadji, D. (2017). Klasifikasi Judul Buku dengan Algoritma Naive Bayes dan Pencarian Buku pada Perpustakaan Jurusan Teknik Elektro. *Jurnal Teknik Elektro*, 9(2), 60–65.
- [3] Desiani, A., Yahdin, S., & Rodiah, D. (2020). Prediksi Tingkat Indeks Prestasi Kumulatif Akademik Mahasiswa dengan Menggunakan Teknik Data Mining. *Jurnal Teknologi Informasi Dan Ilmu Komputer*, 7(6), 1237. <https://doi.org/10.25126/jtiik.2020722493>
- [4] Elmanora, E., Muflikhati, I., & Alfiasari, A. (2012). Kesejahteraan Keluarga Petani Kayu Manis. In *Jurnal Ilmu Keluarga dan Konsumen* (Vol. 5, Issue 1, pp. 58–66). <https://doi.org/10.24156/jikk.2012.5.1.58>
- [5] Etriyanti, E., Syamsuar, D., & Kunang, N. (2020). Implementasi Data Mining Menggunakan Algoritma Naive Bayes Classifier dan C4.5 untuk Memprediksi Kelulusan Mahasiswa. *Telematika*, 13(1), 56–67. <https://doi.org/10.35671/telematika.v13i1.881>
- [6] Firdaus, Y. M. (2019). Penerapan Metode Naive Bayes Classifier Untuk Mengklasifikasi Tingkat Prestasi Akademik Santri Pondok Pesantren Mahasiswa (Ppm) Baitul Jannah Malang. *JATI (Jurnal Mahasiswa Teknik Informatika)*, 3(1), 327–336. <https://ejournal.itn.ac.id/index.php/jati/article/download/1398/1252>
- [7] Hasudungan, R. (2018). Analisis Indikator Kinerja Dosen Terhadap Prestasi Mahasiswa Semester Satu dengan Menggunakan Decision Tree. *Jurnal Rekayasa Teknologi Informasi (JURTI)*, 2(2), 192. <https://doi.org/10.30872/jurti.v2i2.1768>
- [8] Hasudungan, R. (2021). Naive Bayes Model for Student Data Analysis. *International Journal of Advances in Engineering and Management (IJAEM)*, 3(7), 2931–2937. <https://doi.org/10.35629/5252-030729312937>
- [9] Hasudungan, R., & Pranoto, W. J. (2021). Implementasi Teorema Naive Bayes Pada Prediksi Prestasi Mahasiswa. *Jurnal Rekayasa Teknologi ...*, 5(1), 10–16. <http://e-journals.unmul.ac.id/index.php/INF/article/view/4996>
- [10] Hasudungan, R., Pranoto, W. J., & Rudiman. (2020). Using MDA to Improve Naive Bayes Classification for Students Performance Prediction. *JSE Journal of Science and Engineering*, 1(2), 65–70.
- [11] Hendikawati, P. (2011). Analisis Faktor yang Mempengaruhi Indeks Prestasi Mahasiswa. *Kreano: Jurnal Matematika Kreatif-Inovatif*, 2(1), 27–35.
- [12] Herawan, T., Deris, M. M., & Abawajy, J. H. (2010). A rough set approach for selecting clustering attribute. *Knowledge-Based Systems*, 23(3), 220–231. <https://doi.org/10.1016/j.knsys.2009.12.003>
- [13] Juledi, A. P. (2022). Analisis Algoritma Roughset Pada Penerimaan Beasiswa. 3(4), 564–570. <https://doi.org/10.47065/josh.v3i4.1820>
- [14] Mahanggara, A., & Laksito, A. D. (2019). Prediksi Pengunduran Diri Mahasiswa Universitas Amikom Yogyakarta Menggunakan Metode Naive Bayes. *Simetris: Jurnal Teknik Mesin, Elektro Dan Ilmu Komputer*, 10(1), 273–280. <https://doi.org/10.24176/simet.v10i1.2967>
- [15] Maros, H., & Juniar, S. (2016). IMPLEMENTASI METODE K-NEARS PREDIKSI INDEKS PRESTASI. 1–23.
- [16] Muslim, M. A., Prasetyo, B., Mawarni, E. L. H., & Herowati, A. J. (2019). Data Mining ALgoritma C4.5 Disertai contoh dan penerapannya dengan program komputer.
- [17] Nofriansyah, D. (2015). Konsep Data Mining Vs Sistem Pendukung Keputusan. <https://books.google.co.id/books?id=PoJyCAAQBAJ&pg=PR6&ots=YWKmWjwQiW&dq=info%3AytJfhh7xVQYJ%3AScholar.google.com&lr&pg=PA12#v=onepage&q&f=false>
- [18] Pawlak, Z. (1998). Rough set theory and its applications. *Journal of Telecommunications and Information Technology*, 29(7), 7–10. <http://www.informaworld.com/openurl?genre=article&doi=10.1080/019697298125470&magic=crossref>
- [19] Pratama, R. O., Kartika, L., & Sayekti, A. (2018). Analisis Faktor-Faktor Yang Memengaruhi Prestasi Mahasiswa Di Perguruan Tinggi. *Perspektif Ilmu Pendidikan*, 32(2), 153–163. <https://doi.org/10.21009/pip.322.8>
- [20] Putra, A., Matondang, Z. A., Sitompul, N., Pendahuluan, I., & Prediksi, A. (2018). Implementasi Algoritma Rough Set Dalam Memprediksi Kecerdasan Anak. *J. Pelita Inform.*, 7(2), 149–156.
- [21] Rolansa, F., Yunita, Y., & Suheri, S. (2020). Sistem prediksi dan evaluasi prestasi akademik mahasiswa di Program Studi Teknik Informatika menggunakan data mining. *Jurnal Pendidikan Informatika Dan Sains*, 9(1), 75. <https://doi.org/10.31571/saintek.v9i1.1696>
- [22] Samaray, S. (2022). Implementasi Algoritma Rough Set dengan Software Rosetta untuk Prediksi Hasil Belajar. *Jurnal Eksplorasi Informatika*, 11(1), 57–66. <https://doi.org/10.30864/eksplorasi.v11i1.498>
- [23] Samuel, Y. T., Beatrix, C., & Nahuway, A. (2020). Prediksi Indeks Prestasi Mahasiswa Yang Berkuliah Sambil Bekerja Di Universitas Advent Indonesia Dengan Menggunakan Metode

Decision Tree C4 . 5 Dan SMOTE Predicting Student Grade Point Average Who Is Studying While Working At Adventist University Of Indon. 69–77.

- [24] Saputro, I. W., & Sari, B. W. (2020). Uji Performa Algoritma Naïve Bayes untuk Prediksi Masa Studi Mahasiswa. *Creative Information Technology Journal*, 6(1), 1. <https://doi.org/10.24076/citec.2019v6i1.178>
- [25] Senan, N., Ibrahim, R., Mohd Nawi, N., Yanto, I. T. R., & Herawan, T. (2011). Rough set approach for attributes selection of traditional Malay musical instruments sounds classification. *Communications in Computer and Information Science*, 151 CCIS(PART 2), 509–525. https://doi.org/10.1007/978-3-642-20998-7_59
- [26] Setiyani, L., Wahidin, M., Awaludin, D., & Purwani, S. (2020). Analisis Prediksi Kelulusan Mahasiswa Tepat Waktu Menggunakan Metode Data Mining Naïve Bayes : Systematic Review. *Faktor Exacta*, 13(1), 35. <https://doi.org/10.30998/faktorexacta.v13i1.5548>
- [27] Sonang, S., Purba, A. T., & Sirait, S. (2022). Prediksi Prestasi Mahasiswa Dengan Menggunakan Algoritma Backpropagation. *Jurnal Teknik Informasi Dan Komputer (Tekinkom)*, 5(1), 67. <https://doi.org/10.37600/tekinkom.v5i1.512>
- [28] Suad A. Alasadi, & Wesam S. Bhaya. (2017). Review of Data Preprocessing Techniques in Data Mining. *Journal of Engineering and Applied Sciences*, 12(16), 4102–4107.
- [29] Syahputra, I. K., Bachtiar, F. A., & Wicaksono, S. A. (2018). Implementasi Data Mining untuk Prediksi Mahasiswa Pengambil Mata Kuliah dengan Algoritme Naive Bayes. *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer*, 2(11), 5902–5910. <http://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/3464>
- [30] 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.
- [31] Tommy, T., & Husein, A. M. (2021). Model Prediksi Prestasi Mahasiswa Berdasarkan Evaluasi Pembelajaran Menggunakan Pendekatan Data Science. *Data Sciences Indonesia (DSI)*, 1(1), 14–20. <https://doi.org/10.47709/dsi.v1i1.1168>
- [32] Universitas Muhammadiyah Kalimantan Timur. (2018). *Buku Panduan Akademik UMKT*.
- [33] Utari, S. (2022). Penerapan Algoritma Rought Set Untuk Memprediksi Jumlah Permintaan Produk. *Bulletin of Data Science*, 1(2), 73–79.
- [34] Wantono, S. (2013). PREDIKSI PENYELESAIAN STUDI MAHASISWA BARU DENGAN METODE FUZZY TSUKAMOTO. 27037, 6–26.
- [35] Widaningsih, S. (2019). Perbandingan Metode Data Mining Untuk Prediksi Nilai Dan Waktu Kelulusan Mahasiswa Prodi Teknik Informatika Dengan Algoritma C4,5, Naïve Bayes, Knn Dan Svm. *Jurnal Tekno Insentif*, 13(1), 16–25. <https://doi.org/10.36787/jti.v13i1.78>



UMKT
UNIVERSITAS MUHAMMADIYAH
Kalimantan Timur

Kampus 1 : Jl. Ir. H. Juanda, No.15, Samarinda
Kampus 2 : Jl. Pelita, Pesona Mahakam, Samarinda
Telp. 0541-748511 Fax 0541-766832

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ

SURAT KETERANGAN ARTIKEL PUBLIKASI

Assalamu'alaikum warahmatullahi wabarakatuh

Saya yang bertanda tangan dibawah ini:

Nama	:	Rofilde Hasudungan, S.Kom., M.Sc
NIDN	:	1107048601
Nama	:	Muhammad Febri Maulana
NIM	:	1811102441059
Fakultas	:	Sains dan Teknologi
Program Studi	:	SI Teknik Infomatika

Menyatakan bahwa artikel ilmiah yang berjudul "Implementasi kombinasi algoritma *naive bayes* dan algoritma *rough set* untuk memprediksi prestasi mahasiswa" telah di submit pada *Journal of Science and Engineering* pada tahun 2023.

<https://journals.umkt.ac.id>

https://drive.google.com/drive/folders/1fkUmuFyiep_ftoVS2sQatGOIramhdRCp?usp=sharing

Demikian surat keterangan ini dibuat untuk dapat dipergunakan sebagaimana mestinya.

Wassalamu'alaikum warahmatullahi wabarakatuh

Mengetahui

Muhammad Febri Maulana

Samarinda, 25 September 2023

Rofilde Hasudungan, S.Kom., M.Sc