NASKAH PUBLIKASI (*MANUSCRIPT*)

IMPLEMENTASI ALGORITMA NAÏVE BAYES DAN ALGORITMA ROUGH SET UNTUK MEMPREDIKSI TINGKAT PEMAHAMAN MAHASISWA TERHADAP MATA KULIAH

IMPLEMENTATION OF NAÏVE BAYES AND ROUGH SET TO PREDICT THE LEVEL OF STUDENT UNDERSTANDING OF THE COURSE

Siti Lailatus Soimah, Rofilde Hasudungan

DISUSUN OLEH :

SITI LAILATUS SOIMAH

1811102441091

PROGRAM STUDI S1 TEKNIK INFORMATIKA

FAKULTAS SAINS DAN TEKNOLOGI

UNIVERSITAS MUHAMMADIYAH KALIMANTAN TIMUR

SAMARINDA

2023

Naskah Publikasi (*Manuscript*)

Implementasi Algoritma Naïve Bayes dan Algoritma Rough Set untuk Memprediksi Tingkat Pemahaman Mahasiswa terhadap Mata Kuliah

Implementation of Naïve Bayes and Rough Set to Predict the Level of Student Understanding of the Course

Siti Lailatus Soimah, Rofilde Hasudungan

Disusun Oleh :

Siti Lailatus Soimah

1811102441091

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

HALAMAN PENGESAHAN

IMPLEMENTATION OF NAÏVE BAYES AND ROUGH SET TO PREDICT THE LEVEL OF STUDENT UNDERSTANDING OF THE COURSE

NASKAH PUBLIKASI

DISUSUN OLEH :

SITI LAILATUS SOIMAH
1811102441091

Dosen Pembimbing	Penguji
Rofilde Hasudungan, S.Kom., M.Sc	Wawan Joko Pranoto, S.Kom., M.TI
NIDN : 1107048601	NIDN : 1102057701
Dekan	Ketua Progam Studi
Prof. Ir. Sarjito, MT., Ph.D.	Asslia Johar Latipah, M.Cs
NIDN : 0610116204	NIDN : 1124098902

UNIVERSITAS MUHAMMADIYAH KALIMANTAN TIMUR
FAKULTAS SAINS DAN TEKNOLOGI

UNIVERSITAS MUHAMMADIYAH KALIMANTAN TIMUR
PROGRAM STUDI INFORMATIKA

Implementasi Algoritma Naïve Bayes dan Algoritma Rough Set untuk Memprediksi Tingkat Pemahaman Mahasiswa terhadap Mata Kuliah

Siti Lailatus Soimah[1*], Rofilde Hasudungan[2]

[1,2]Teknik Informatika, Universitas Muhammadiyah Kalimantan Timur, Jl. Ir. H. Juanda No.15, Sidodadi, Kec.Samarinda Ulu, Kota Samarinda, Kalimantan Timur 75124, Indonesia

**Email Corresponding Author*: Lailatussoimahsiti@gmail.com

ABSTRAK

Dalam proses belajar mengajar tingkat pemahaman mahasiswa terhadap mata kuliah merupakan salah satu hal utama yang penting bagi berjalannya proses kegiatan perkuliahan. Maka dari itu perlu adanya prediksi Tingkat Pemahaman Mahasiswa Terhadap Mata Kuliah menggunakan algoritma rough set dan algoritma naïve bayes tujuan peneliitian ini ingin mengetahui performa naive bayes dan rough set dalam memprediksi Tingkat Pemahaman Mahasiswa Terhadap Mata Kuliah dan mengkomparasi hasilnya dengan algoritma naive bayes saja. Jumlah data yang digunakan untuk proses pengujuan kinerja algoritma adalah 146 data mahasiswa dengan rasio 30% data testing 70% data training hasil pengujian algoritma rough set dan naïve bayes menghasilkan akurasi 67.14%, sedangkan metode naïve bayes tanpa rough set mengkasilkan akurasi 62.44%. Berdasarkan evaluasi diketahui bahwa penggunaan metode rough set dapat meningkatkan hasil prediksi pada klasifikasi naïve bayes dari hasil akurasi 62.79% menjadi 67.44% Sehingga penggunaan rough set dan naïve bayes sangat bagus dan dapat diterapkan dengan sangat baik, dan dapat digunakan dalam memprediksi tingkat pemahaman mahasiswa terhadap mata kuliah pemprogaman berbasis objek (PBO).

Kata kunci: Rough Set; Naïve Bayes; Tingkat Pemahaman Mahasiswa; Confusion Matrix; Accuracy.

Implementation of Naïve Bayes and Rough Set to Predict the Level of Student Understanding of the Course

Siti Lailatus Soimah^{1*}, Rofilde Hasudungan^{2}

[1,2]Teknik Informatika, Universitas Muhammadiyah Kalimantan Timur, Jl. Ir. H. Juanda No.15, Sidodadi, Kec.Samarinda Ulu, Kota Samarinda, Kalimantan Timur 75124, Indonesia

**Email Corresponding Author*: Lailatussoimahsiti@gmail.com

ABSTRACT

In the learning process the level of student understanding of the subject is one of the main things that is important for the course of the lecture activity process. Therefore it is necessary to predict the level of student understanding of the course using the rough set algorithm and the naïve Bayes algorithm. The purposes of this research is to determine the performance of naive Bayes and rough set in predicting the level of student understanding of the course and to compare the results with the naive Bayes algorithm only. The amount of data used for the process of testing the performance of the algorithm is 146 student data with a ratio of 30% data testing 70% data training the results of testing the rough set and naïve Bayes algorithms produce an accuracy of 67.14%, while the naïve Bayes method without rough set produces an accuracy of 62.44%. Based on the evaluation it is known that the use of the rough set method can increase the prediction results in the naïve Bayes classification from 62.79% to 67.44% accuracy. So the use of rough set and naïve bayes is very good and can be applied very well, and can be used in predicting students understanding of the eye object-based programming course.

Keywords*: Rough Set; Naïve Bayes; Student Understanding Level; Confusion Matrix; Accuracy*

1. Introduction

Through higher education at the Muhammadiyah University of East Kalimantan (UMKT). Students are guided to become experts, professionals in a science or scientific field, in situations of participating in lecture activities which include activities to listen to lecturers, think, argue, ask questions and various other activities [1]. In the teaching and learning process the level of student understanding of the subject is one of the main things that is important for the course 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 students can understand. Especially with regard to how a lecturer conveys the content of lecture material. Each lecturer who provides material has a different learning method for his students. differences in the way lecturers teach greatly affect the results that will be obtained by students when the lecture process takes place. In addition, several factors that affect the level of student understanding such as learning comfort, securing learning and so on are also very influential on student understanding. The presence 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. 2]. In previous research there were researchers who predicted the level of student understanding such as [3] rough set method, [4] rough set method, [5], [6] case based learning method, [2] C4.5 algorithm method, [7] k-means clutering algorithm method, [8], [9] quantitative method, [10], [1] using the naïve Bayes method.

Naive Bayes itself is a method that has advantages such as speed and a very accurate level of accuracy in classifying data. Naive Bayes is a classification method that is very effective and efficient in testing large datasets to determine patterns in the past and look for functions that will become patterns of assessing data in the future. This algorithm aims to classify data in certain classes (Patrimurti & Septiani, 2020). In previous studies, there were studies using the naïve Bayes method, including, [11] to predict student graduation on time, [12] to predict student achievement, [13] to predict students taking courses, [14] to predict student study period based on factors related to student academics, [15] for predicting student graduation on time, [16] for predicting graduation rates on time, [12] using naïve bayes for student data analysis. Naïve Bayes also has drawbacks, namely when certain parameters are empty or have no value and Naive Bayes excludes them, this affects the quality of the results issued, so a method is needed to select the best parameter, namely the rough set which can reveal hidden patterns in the data and help predict.

The Rough set method is a method that can deal with vague and inconsistent data. Rough sets are widely used, especially in selecting attributes such as Hasudungan and Wawan (2021). In previous studies, there were several studies that used the rough set to select attributes for naïve Bayes, such as (Rofile Hasudungan, Wawan joko Pranoto 2021) which used the rough set to select attributes for predicting student achievement. The results of the analysis show that the proposed model has an accuracy level of 77 .5%, and a lower yield of 69%. (Devi Silvia Siltonga, et al. 2019) Predicting the level of student understanding on the test results showed an accuracy of 88.24%, namely 8 respondents stated they did not understand and 60 respondents stated they understood the level of student understanding of the subject based on their sitting position. With class precision, the prediction of not understanding has a value of 0%, while the prediction of understanding has a value of 88.24%. Class recall on true does not understand has a value of 0%, while on true understand has a value of 100%. (Hajering, 2021) predicts factors that affect the level of student understanding. The results of this study indicate that learning methods have a positive and significant effect on course understanding. Therefore in this study the authors 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 students' level of understanding of the course.

2. Related Works

The level of student understanding is the degree or level of someone's response to things that are very important in learning something. The level of understanding possessed by a student is very influential in accepting a course material that is being followed. The level of student understanding is strongly influenced by many factors such as learning readiness, learning order and so on [1]. Therefore the importance of an analysis in predicting the level of student understanding makes many researchers conduct research on this matter using various algorithms such as the Rough set algorithm, the C4.5 algorithm, the Naive Bayes algorithm and so on. The following is a related research table that discusses predictions of student understanding levels listed in table 2.1

Table 1 Previous research

Author	Information
Nurul Rofiqo, Dkk[2]	Applying the C4.5 algorithm to predict the level of student understanding of the course. The accuracy obtained is 87.10%.
Algoritma et al[4]	Using the C4.5 algorithm to determine the classification level of student understanding of programming language courses. The accuracy obtained is 84.38%.
Raharjo & windarto[3]	Predict the level of student understanding of the course. The accuracy obtained is 53%.
Mutmainnah & Infokam[6]	Using Naïve Bayes to predict student study period based on factors related to student academics. The accuracy obtained is 85.17%.
Astuti et al[17]	Naïve Bayes to predict the level of student understanding of the data structure algorithm course. The accuracy obtained is 69.23%.
Siltonga & Dewi[1]	Analysis of the Naïve Bayes method predicts the level of student understanding of the subject based on sitting position. The accuracy obtained is 88.24%.
Eka Sabna, Muhardi [18]	Using the Decision Tree algorithm to predict academic achievement based on socioeconomic, motivation, lecturer role, discipline and learning outcomes. The accuracy obtained is 65%.
Abdul Rohman, Sri Mujiyono [19]	Using Decision Tree C4.5 in order to get a decision tree model with variables or grade point attributes that affect student graduation predicates. The accuracy obtained is 71.67%.
Riski Annisa dan Agung Sasongko [20]	Using Naïve Bayes to predict student academic scores by utilizing probability calculations and past data statistics to predict future data based on previous data. The accuracy obtained is 96.24%.

Ahmad Fauzi dan Tukiyat [21]	Using the Decision Tree and Naïve Bayes the results of the accuracy of the Naive Bayes method remain the greatest, even though the increase in accuracy after optimization is lower than the Decision Tree method. The accuracy obtained is 94.47%.
Aspiah & Tagfirul Azhima Yoga Siswa[22]	Implementation of correlation based feature selection (CFS) to increase the accuracy of the C4.5 algorithm in predicting student academic performance based on learning management systems. The accuracy obtained is 97.22%.

3. Rough Set

Rough Set theory was first introduced by Pawlak, who stated that Rough set is a mathematical method for dealing with inconsistent and ambiguous data (Pawlak, 1982). 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 (Pawlak, 1991). And Pawlak proposes that gross set theory is founded on the assumption that with every member of the universe of discourse we relate some information. The concept of a rough set is a new mathematical technique for dealing with obscurity, imprecision, and uncertainty (Pawlak & Skowron, 2007. The following flowchart for solving the rough set algorithm can be seen in Figure 1.

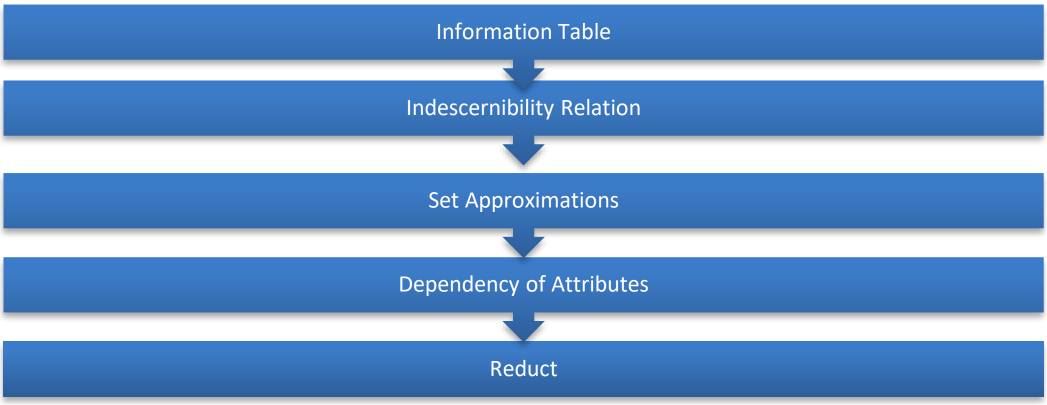


Figure 1. Rough set finishing flow

The following is a description based on the rough set algorithm completion flow as a solution : [23]

1. Information Table is a table consisting of columns and rows containing data, where the columns are labeled with attributes, and the rows are filled with the values of the attributes. With system information like S = (U,A,V, f), where U is the set of objects, A is the attribute set which cannot be empty, V=UaЄAVa, Va is the domain attribute A, f:U×A → V is a total function such that f(u,a) Є Va , for every f(u,a) Є U×A, is called the information or knowledge function. The table must have one decision attribute (Decision information system) which cannot have an empty value. With system information as follows D = (U, A U {d}, V, f, where U, A, V and f correspond to D and {d} are decision attributes where {d} ∩ A ≠ Ø).
2. *Indescernibility Relation* is an idea between objects that can be defined, have similarities so that they can be put together. By definition S = (U,A,V,f) becomes an information system and B will become part of A two elements x,y Є U is said to be B-indescernible (cannot be distinguished by the set of attributes B □ A in S) if only f (x,a) = f (y,a) for every a Є B.

3. *Set Appromaximations* is grouping the results of the Indescernibility relation which is used to define approximations as a basic concept in the rough set algorithm, to determine the lowest estimate and the top estimate in a set can be defined as follows S = (U,A,V, f) becomes an information system and B will being a part of A, X will be a part of U. The B-lower approximation of X can be denoted as $\underline{\mathrm{B}}$□X), and the B-upper approximation of X can be denoted as $\overline{\mathrm{B}}$(X). So it can be defined by equation 1.

$$\underline{\mathrm{B}}(\mathrm{X}) = \{x \in \mathrm{U} \dashv |[x]\ \mathrm{B} \subseteq \mathrm{X}\}\ \mathrm{dan} \qquad (1)$$

$$\overline{\mathrm{B}}(\mathrm{X}) = \{x \in \mathrm{U} \dashv |[x]\mathrm{B} \cap \mathrm{X}\ \neq\ \emptyset\}$$

Another Another important problem is looking for or finding dependencies between attributes, with the definition S = (U,A,V, f) being an information system, D and C being part of A. Attribute D will functionally depend on attribute C, so it can be denoted C □ D, if each value of D (decision) is exactly related to the value of C. Dependency of Attributes is a step to calculate the consistency of each attribute with the following definition S = (U,A,V,f) to be a system information, D and C become part of A. D's dependency on C is in level k (0≤k≤1), with the notation C □k D. Then it can be defined by equation 2.

$$k\ =\ \frac{\sum \mathrm{x} \in \mathrm{U/D}\ \ |\underline{\mathrm{C}}(\mathrm{X})|}{|\mathrm{U}|} \qquad (2)$$

4. *Reduct* is the process of minimizing the set of attributes. By recalculating using the previous steps to be applied to each existing attribute, so as to get the best attribute and not reduce the attribute's consistency value. With the following definition S = (U,A,V,f) being an information system, and B being part of A, if B has an effect on attribute consistency it becomes excessive, it can be discarded with the notation B if U / (B – {b}) = U / B, if it doesn't affect the consistency of the attributes then it is very necessary. The following is a table of previous research using roughset as an attribute selection feature as shown in table 2.

Table 2 Previous Rough Set Research

Author	Information
[24]	Applying the naïve Bayes model for student data analysis. The accuracy results obtained are 68.09%.
[3]	Application of matching learning with the concept of data meaning roughset to predict the level of student understanding of courses. The results obtained are 90 rules.
[25]	Implementation of the rough set algorithm with Rosetta software for predicting learning outcomes. The accuracy results obtained are 14 rules.
[26]	Implementation of the Dana Naïve Bayes rough set algorithm to obtain rules in selecting applicants for houses of worship facilities. Accuracy results obtained 92%
[27]	Implementation of rough set algorithm in predicting children's intelligence. The accuracy results obtained are 13 rules.

4. Maximum Dependency of Attributes

The Maximum dependency attributes method is a rough set algorithm based on attribute selection that can find dependencies between attributes and can reduce redundant attributes. In reducing redundant attributes, you can use a method by calculating the dependence between one attribute and another based on the maximum dependency value of the attribute on the data [28]. As for the steps for implementing the maximum dependency attributes, it requires several stages of completion as shown in Figure 2 below.

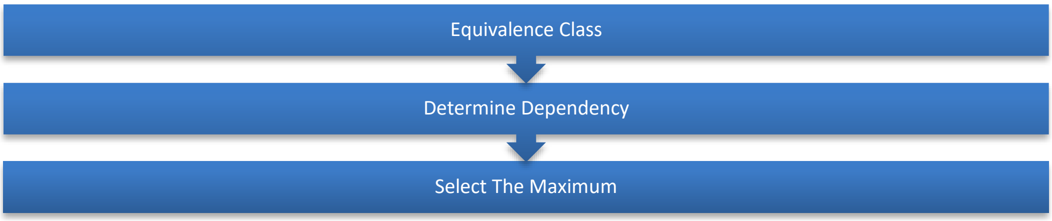


Figure 2. MDA Solution Flow

The following is a description of the maximum dependency attribute completion scheme as a method of calculating attribute dependency: [29].

1. *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 S = (U,A,V,f) being an information system, D and C being part of A If D is completely dependent on C, then αB (X) ≤ αC (X), for all members X ⊆ U. Based on this definition, IND(C) ⊆ IND(D) can therefore be applied to equation 2.4.

$$D(X) \subseteq C(X) \subseteq X \subset \bar{C}(X) \subseteq \bar{D}X \quad (3)$$

2. *Determine dependency is the next step in determining the maximum dependence of the attribute α^j with respect to all attributes α_i, but $\alpha^j \neq \alpha_i$. As for the application, you can use equation 4.*

$$\mathrm{D}\left(\underline{\mathrm{R}}(\mathrm{X}), \overline{\mathrm{R}}(\mathrm{X})\right) = 1 - \frac{|\underline{\mathrm{R}}(\mathrm{X}) \cap \overline{\mathrm{R}}(\mathrm{X})|}{|\underline{\mathrm{R}}(\mathrm{X}) \mathrm{U}\, \overline{\mathrm{R}}(\mathrm{X})|}, = 1 - \frac{|\underline{\mathrm{R}}(\mathrm{X})|}{|\overline{\mathrm{R}}(\mathrm{X})|}, = 1 - \alpha R(X) \quad (4)$$

3. *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 a dependency value. By definition S = (U,A,V, f) becomes an information system, S = (U,A,V, f) becomes an information system and $C1$, $C2$, … , Cn so that D becomes part of A. If $C1 \Rightarrow k1$ D, $C2 \Rightarrow k2$, ... $Cn \Rightarrow k$ ($\alpha C2$ (X) ≤ $\alpha C1$ (X) For every $X \subseteq$ U. As for Equation 5.*

αD (X)≤ αCn (X) | kn ≤ kn-1 ≤ ⋯ ≤ k2 ≤ k1 |
[x]Cn ⊆ [x]Cn-1
(5)

5. Naïve Bayes

Naïve Bayes is a classification algorithm based on the Bayesian theorem in statistics and can be used to predict the probability of class membership. Naïve Bayes calculates the value of the posterior probability P(H|X) using the probabilities of P(H), P(X), and P(X|H) where the value of X is testing data whose class is unknown. The value of H is the hypothesis of data X which is a more specific class. The value of P(X|H) or also called likelihood, is the probability of hypothesis X based on condition H. The value of P(H) or also called prior probability is the probability of hypothesis H. Meanwhile, the value of P(X) is also called predictor prior probability, is the probability of X [30].

$$P(H|X) = \frac{P(X|H).P(H)}{P(X)} \quad (6)$$

Information:

X :Data with an unknown class

H :The data hypothesis is a specific class

P(H|X) :Probability of hypothesis H based on condition X (posteriori probability)

P(H) :Probability hypothesis H (probability prior)

P(X|H) :The probability of X is based on the conditions in the H hypothesis

P(X) :The probability of X

Naïve Bayes algorithm has the advantage that it is considered fast and strong, especially when dealing with big data. And naïve Bayes considers all attributes to be the same, and that's why naïve Bayes is called naïve.

6. Evaluation

Evaluation is a process in data analysis to measure the model that has been produced. There are many tools that can be used to measure the performance of an algorithm, one of which is using accuracy, evaluation measurements on the role of classification data maining are measuring accuracy and calculating accuracy based on the confusion matrix. Confusion matrix is one way that is often used in the evaluation process of classification data mining models by predicting the truth of objects. The testing process utilizes the confusion matrix which places the prediction class at the top of the matrix then the observed sources are placed on the left of the matrix. Each matrix cell contains a number that displays the actual number of cases of the class being observed [31]. Table 3 describes an example of a classification process confusion matrix. To measure the accuracy of the model, you can apply equation 7 which is used to calculate the results of accuracy, while to calculate the error rate you can define it with equation 8, and to calculate the precision, measure the data that has been predicted positively with the reality that correct and incorrect can use equation 9. Lastly, to calculate the sensitivity (recall) of many successful data when predicted with a comparison of all data which is in fact positive, you can use equation 10.

Table 3 Confusion matrix

	Action True	Action False
Predict True	TP	FP
Predict False	FN	TN

Information :

1. TP *(True Positive)* are the correct class observations and the correct predictions.
2. TN (True Negatif) is a correct class observation with a wrong prediction.
3. FP (False Positive) is an incorrect class observation with a correct prediction.
4. FN (False Negatif*)* is the wrong class observation with the wrong prediction.

$$Accurasi = \frac{\mathrm{TP + TN}}{\mathrm{TP + FP + FN + TN}} \quad (7)$$

$$Error = \frac{\mathrm{FP + FN}}{\mathrm{TP + FP + FN + TN}} \quad (8)$$

$$Accurasi = \frac{\mathrm{TP}}{\mathrm{TP + FP}} \quad (9)$$

$$Sentivity = \frac{TP}{TP + FN} \qquad (10)$$

7. Methodology

To solve the research problem, we designed the research stages as shown in Figure 3.1. The figure shows the stages of research carried out using 2 methods, namely method A: using Rough Set and Naïve Bayes, method B: using Naïve Bayes only. In general, the research stage for method A is to eliminate attributes that are not useful in processing data using Rough Set. Then the attributes that have been eliminated are continued to the Naïve Bayes stage to classify by predicting opportunities, and evaluation is carried out to determine the results of accuracy. As for method B does not use the rough set, after going through the Naïve Bayes process it proceeds to Evaluation to determine the results of accuracy. Both will be compared in Comparison to determine which method produces the most perfect accuracy value.

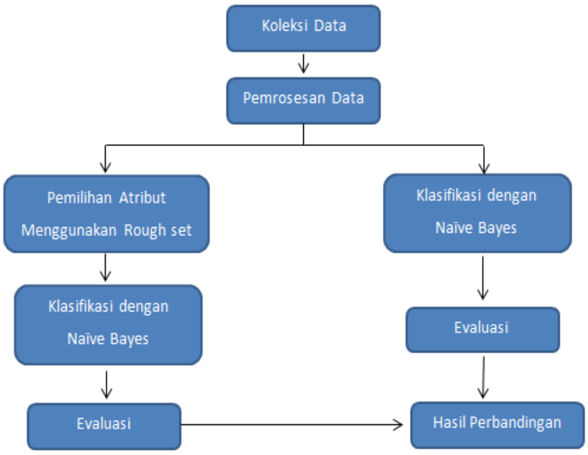

Koleksi Data
Pemrosesan Data
Pemilihan Atribut Menggunakan Rough set
Klasifikasi dengan Naïve Bayes
Klasifikasi dengan Naïve Bayes
Evaluasi
Evaluasi
Hasil Perbandingan

Figure 3. Research Stages

7.1. The data in this study were obtained from Informatics Engineering students class of 2021, Faculty of Science and Technology. In the questionnaire, the researchers used a Likert scale as respondents. The Likert scale is a scale used to measure attitudes and opinions. In the Likert scale there are 5 choices with gradations from very good, good, fair, bad, and very bad [32]. Questionnaires with a Likert scale will be distributed to students in the form of a Google Form, with the attributes used obtained from [33]. The following attributes are shown in Table 4.

Table 4 Attribute Collection

NO	Question	Grades/Answer Choices
A1	Name	Student's full name
A2	NIM	Student ID Number
A3	Gender	Student gender
Competency items pendagogik A4 – A20		
A4	Readiness to give lectures and/or practice/practicum	1. Very Good 2. Fine

A5	Regularity and order in the administration of lectures	3. Enough 4. Bad 5. Very Bad
A6	The ability to liven up the classroom atmosphere	
A7	Clarity in conveying material and answers to questions in class	
A8	Utilization of learning media and technology	
A9	Diversity of ways of measuring learning outcomes	
A10	Providing feedback on assignments	
A11	Appropriateness of exam material and/or course assignments	
A12	The suitability of the value given to the learning outcomes of professional competency items is	
A13	The ability to explain the subject matter or topic appropriately	
A14	Ability to provide relevant examples of the concepts being taught	
A15	The ability to explain the relationship between the fields/topics being taught and other fields/topics	
A16	The ability to explain the relationship between the fields/topics being taught and the context of life	
A17	Mastery of the latest issues in the field being taught	
A18	The use of research results to improve the quality of lectures	
A19	Involving students in research/study and/or development/engineering/design carried out by lecturers	
A20	Ability to use a variety of international communication technology item competence is	
Professional competency items A21 – A26		
A21	Authority as a personal lecturer	1. Very Good 2. Fine 3. Enough 4. Bad 5. Very Bad
A22	Wisdom in making decisions	
A23	Be an example in attitude and behavior	
A24	One word and action	
A25	The ability to control oneself in various situations and conditions	

A26	Fair in treating students the social competency item is	
Professional competency item A27 – A31		
A27	Ability to express opinions	1. Very Good 2. Fine 3. Enough 4. Bad 5. Very Bad
A28	Ability to convey criticism, suggestions, and opinions of others	
A29	Get to know the students who attend the course well	
A30	Easy to get along with colleagues, employees and students	
A31	Tolerance for student diversity	
A32	Object-oriented programming course grades	Course grades

8. Results and Discussion

8.1. Data Penelitian

The data taken was obtained from a Google Form questionnaire through students taking PBO (object-based programming) courses in Informatics Engineering study program class of 2021. Data collection was carried out in two ways, namely distributing questionnaires via the class WhatsApp group and distributing them directly to students when conducting offline learning process in class Students who participated in filling in the data totaled 146 students, with attributes on the questionnaire such as very good, good, fair, bad and very bad. In this study, 146 student data from Muhammadiyah University of East Kalimantan (UMKT) will be used as predictions.

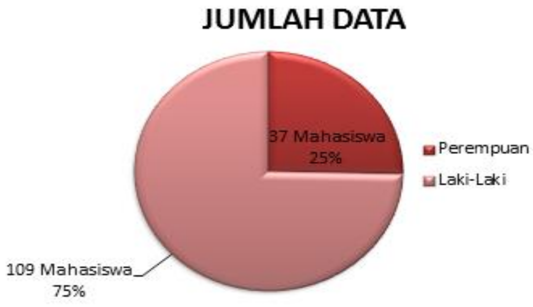


Figure 4. The number of student data for PBO courses class of 2021

8.2. Data Processing

Data The data that has been collected from the results of the questionnaire will then be processed so that it can be used in the attribute selection process and the classification process. The data cannot be empty or of categorical data type. In data processing carried out several stages, namely data cleaning and data transformation

8.2.1. Integrasi Data

The data integration stage is combining student data that has been obtained from the questionnaire with value data from object-oriented programming (PBO) lecturers into one unified data based on name. So it can be combined as in example 5.

Tabel 5 Combined table of student data and value data

No	A2	A3	A4	….	A32
1	2111102441032	L	Very good	….	90
2	2111102441108	L	Very good	…	70
3	2111102441074	P	good	….	40
…	…	…	….	….	….
146	2111102441149	L	Very good	….	65

A32 attribute data was obtained from object-oriented programming (PBO) lecturers, sample data can be seen in table 6 as follows.

Tabel 6 Table of Course Grades

No.	Nim	nilai
1	2111102441142	75
2	2111102441003	70
3	2111102441038	60
4	2211102441207	55
…	…	…
148	1911102441024	40

8.2.2. Integrasi Data

The data cleaning phase is carried out to remove incomplete or empty data, which has no value and duplicated data so that it can be used for the process of selecting attributes and classification. After checking the data, there were 146 student data for the 2021 batch and no duplicated or blank data was found in the data. So that it can be combined as in the example of object-based programming (PBO) student data that has gone through the cleaning stage.

8.2.3. Data Transformation

The data transformation stage was carried out to change the numeric type data to categorical, the transformation in this study was carried out so that it could be used for attribute selection and classification. By changing to adjust the table contained in table 7.

Table 7 Assessment Norms Based on Academic Programs

LETTER	NUMBER	FINAL SCORE	PREDIKATE	INFORMATION
A	4	≥80	Very Good	Graduated
AB	3,5	75-<80	Good	
B	3	70-<75		
BC	2,5	65-<70		

C	2	60-<65	Enough	
D	1	50-<60	Not Enough	Not Pass
E	0	<50	Fail	
T	0	Tertunda		

Table 8 Example of data that has been transformed

No	A2	A3	A4	…	A32
1	2111102441032	L	Very Good	…	Graduated
2	2111102441108	L	Very Good	…	Graduated
3	2111102441074	P	Good	…	not pass
…	…	…	…	…	…
146	2111102441149	L	Very Good	…	Graduated

8.3. Pemilihan Atribut dengan Rough Set

After processing the data, the data is ready to be processed using the rough set. In the 2021 class student data there are 32 attributes used consisting of 31 condition attributes and 1 student course value attribute. To perform attribute selection, the rough set algorithm can be applied. Because the use of many attributes will affect the results and computation time. The initial step in implementing the rough set algorithm requires a data consistency value that can be achieved through the completion scheme of Figure 1. The range of data consistency values ranges from 0 to 1, with the meaning 0 indicating inconsistent data and 1 indicating consistent data. Based on 146 student subject data (PBO), a consistent value equal to 1 was obtained, which stated that the data was consistent. The consistency value is calculated using the Google Colab web application and the python rst-tools library which can be used to write programs, while the programming language used is the python programming language.

Table 9 The result of the calculation of the MDA attribute dependency

Symbol	Maximum Dependency
A7	0.14383561643835616
A23	0.03424657534246575
A15	0.0273972602739729
A17	0.02054794520547945
A8	0.0136986301369863
A12	0.00684931506849315

Based on the data consistency value, attribute reduction is carried out so that the best attribute results are 6 condition attributes, from the initial attribute which totals 31. The 6 best condition attributes are clarity in conveying material and answers to questions in class (A7), utilization of media and learning technology (A8), the suitability of the value given with the learning outcomes of professional competency items is (A12), the ability to explain the relationship between the field or topic being taught with other fields or topics (A15), mastery of current issues in the field being taught (A17), becomes example in attitude and behavior (A23). The results of selecting this attribute will be used in classification while the remaining 25 attributes will be deleted because they are not used.

8.4. Classification with Naïve Bayes

At this stage the researcher will carry out the data classification process using the naïve Bayes algorithm. In carrying out the classification process, researchers used a data analysis application, namely rapid miner. In this process it will be divided into 2 models as shown in the research stages flowchart 3.1, the first model will classify using all 32 attributes. Whereas the second model will classify using the best attributes that have been selected by the rough set, so all of these experiments are carried out by dividing the data into two parts. Where the data is divided in half with a percentage of 70% data for training and 30% data for testing, totaling 103 data and testing 43 data. Then calculate the probability value using the naive Bayes algorithm on the decision attributes labeled "passed" and "failed" with training data of 103 data. The decision attribute obtained with the label “passed” was 108 data, “did not pass” was 38 data.

$$P\ (INilai = Lulus) = \frac{75}{103} = 0{,}72815534$$

$$P\ (INilai = Tidak\ Lulus) = \frac{28}{103} = 0{,}27184466$$

Next, calculate the supporting attribute values in the training data using formula 2.1. The following is an example of calculating the probability value of the A4 attribute with the labels "Very Good", "Good", "Enough", "Poor", "Very Bad". Subsequent calculations are based on A32 with the label "Passed", "Failed". Here's how to calculate the probability value of the A4 attribute::

$$P\ (Sangat\ Baik\ |\ Lulus) = \frac{15}{75} = 0{,}2$$

$$P\ (Baik|\ Lulus) = \frac{35}{75} = 0{,}46666667$$

$$P\ (Cukup|\ Lulus) = \frac{24}{75} = 0{,}32$$

$$P\ (Buruk|\ Lulus) = \frac{1}{75} = 0{,}01333333$$

$$P\ (Sangat\ Buruk|\ Lulus) = \frac{0}{75} = 0$$

$$P\ (Sangat\ Baik|\ Tidak\ Lulus) = \frac{8}{28} = 0{,}28571429$$

$$P\ (Baik|\ Tidak\ Lulus) = \frac{14}{28} = 0{,}5$$

$$P\ (Cukup|\ Tidak\ Lulus) = \frac{5}{28} = 0{,}17857143$$

$$P\ (Buruk|\ Tidak\ Lulus) = \frac{1}{28} = 0{,}03571429$$

$$P\ (Sangat\ Buruk|\ Tidak\ Lulus) = \frac{0}{28} = 0$$

Then calculate all the values obtained for each attribute that will be used in classification with formula 2.1 which is applied to the 1st testing data.

$$P\ (Lulus) = 0{,}46666667 \times \ldots \times 0{,}72815534 = 0.759$$

$$P\ (Tidak\ Lulus) = 0{,}5 \times \ldots \times 0{,}27184466 = 0{,}241$$

The results of the calculation above can be seen that the probability value of "Pass" is greater than the value of not passing. So that the prediction of the 1st testing data can be said to have passed.

8.5. All Attribute Classification Model

The classification stage with a model that uses all 31 condition attributes and 1 decision attribute, can be seen in table 9 and is applied to the testing data.

Table 10 Classification Results of All Attributes

No	A4	…	A32	Nilai	Hasil prediksi
1	Good	…	Graduated	0.759	Graduated
2	Good	…	not pass	0.217	not pass
3	Very Good	…	Graduated	0.634	Graduated
…	…	…	…	…	…
43	Good	…	Graduated	0.284	not pass

8.6. Classification Model with Attribute Selection

In the classification with the model using the best attribute selection, namely 6 condition attributes and 1 decision attribute contained in table 11 which will be applied to test data (testing).

Table 11 Attribute selection results

No	A4	…	A32	Nilai	Hasil prediksi
1	Good	…	Graduated	0.759	Graduated
2	Good	…	Tidak Lulus	0.577	Graduated
3	Very Good	…	Graduated	0.794	Graduated
…	…	…	…	…	…
43	Ba Good ik	…	Graduated	0.572	Graduated

8.7 Evaluation and Comparison

At this stage the results of the evaluation of the classification of all attributes use equation 2.4 to calculate accuracy using the cofusion matrix which produces an accuracy value of 62.79%. while the evaluation of naïve Bayes classification and attribute selection using the rough set algorithm uses equation 2.4 to calculate accuracy using the confusion matrix produces an accuracy value of 67.44%. Based on the evaluation it is known that the use of the rough set method can increase the prediction results in the naïve Bayes classification from 62.79% accuracy to 67.44 % 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 level of student understanding of object-oriented programming (PBO) courses.

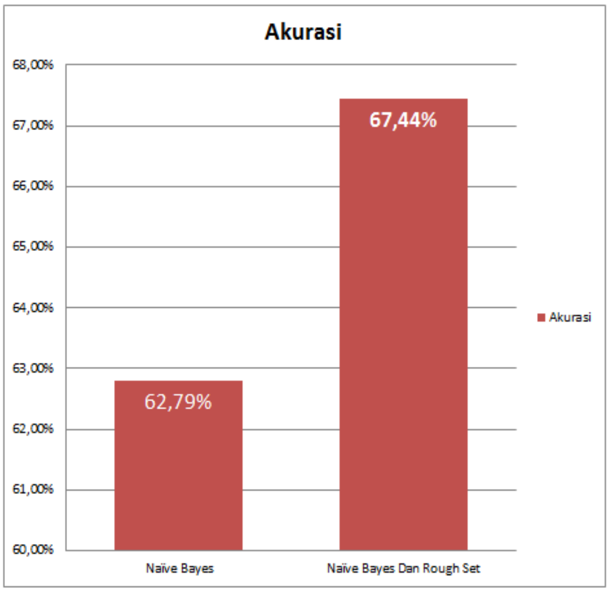


Figure 5. Comparison and accuracy chart

9. Conclusion

Based on the research that has been done the authors conclude as follows:

1. From the attribute collection process, 31 attributes were obtained that would be used in the implementation process using 2 methods, namely, Method A used naïve Bayes and roughset, and Method B only used naïve Bayes. Dari proses pengumpulan data menggunakan kuesioner yang dibuat dengan media google form dan sebarkan ke prodi teknik informatika angkatan 2021 mata kuliah pemprograman berorientasi objek (PBO) yang berjumlah 146 responden.
2. Initially 31 attributes were eliminated into 6 attributes which will be used or processed using the Naïve Bayes method. Dilakukan eksperimen dengan membagi data menjadi 2, yaitu data training dan data testing. Dimana data dibagi berdasarkan analisis statistik dengan rasio 70 : 30 untuk data training dan data testing yang dianggap sebagai rasio terbaik.
3. From the results of the comparison of methods that Method A is very influential and obtains high accuracy results compared to Method B. Berdasarkan eksperimen Metode A klasifikasi Naïve Bayes dengan atribut yang diperoleh dari hasil eliminasi menggunakan Rough Set, dengan 6 atribut mendapatkan nilai akurasi sebesar 67.44%. Pada Metode B klasifikasi Naïve Bayes dengan seluruh atribut, yaitu sebanyak 31 atribut mendapatkan hasil akurasi 62.79%. Maka dari hasil perbandingan 2 metode tersebut bahwa Metode A lebih unggul dari pada Metode B dari sisi Akurasi.
4. From the points above it can be concluded that the classification process of Method A using the naïve Bayes algorithm and rough set is superior in terms of accuracy compared to Method B which only uses the naïve Bayes algorithm.

References

[1] D. S. Siltonga dan R. Dewi, “Analisis Metode Naive Bayes dalam Memprediksi Tingkat Pemahaman Mahasiswa Terhadap Mata Kuliah Berdasarkan Posisi Duduk,” no. September, hal. 427–436, 2019.

[2] P. Seminar, N. Riset, N. Rofiqo, A. P. Windarto, dan E. Irawan, “Penerapan Algoritma C4 . 5 pada Penentuan Tingkat Pemahaman Mahasiswa Terhadap Matakuliah,” no. September, hal. 307–317, 2019.

[3] M. R. Raharjo dan A. P. Windarto, “Penerapan Machine Learning dengan Konsep Data Mining Rough Set (Prediksi Tingkat Pemahaman Mahasiswa terhadap Matakuliah),” *J. Media Inform. Budidarma*, vol. 5, no. 1, hal. 317, 2021, doi: 10.30865/mib.v5i1.2745.

[4] C. Algoritma, U. Menentukan, dan M. R. Lubis, “Algoritma c4.5 untuk menentukan klasifikasi tingkat pemahaman mahasiswa pada matakuliah bahasa pemrograman,” vol. 1, no. 3, 2022.

[5] E. Novi, “Analisis Tingkat Pemahaman Mahasiswa Akuntansi Terhadap Konsep Dasar Akuntansi Setelah Pemberlakuan Ifrs,” *J. Al-Iqtishad*, vol. 10, no. 1, hal. 1, 2017, doi: 10.24014/jiq.v10i1.3109.

[6] S. Mutamainnah dan A. P. INFOKAM, “Pengaruh Tingkat Pemahaman Mahasiswa Terhadap Perkuliahan Dari Penerapan Student-Centered Learning, Case Based Learning dan Cooperative Learning,” *Infokam*, vol. XV, no. II, 2019.

[7] F. Musharyadi, “Tingkat Pemahaman Mahasiswa Terhadap Norma Norma Agama Islam Menggunakan Algoritma K-Means Clustering,” *MENARA Ilmu*, vol. XI, no. 78, hal. 48–54, 2017.

[8] W. W. W. Dalam dan S. Sinarti, “Faktor-Faktor yang Mempengaruhi Tingkat Pemahaman Mahasiswa pada Mata Kuliah Auditing di Politeknik Negeri Batam,” *J. Appl. Account. Tax.*, vol. 4, no. 1, hal. 100–106, 2019, doi: 10.30871/jaat.v4i1.1110.

[9] D. A. Dewi, G. Sabaritha Nimaisa, S. Poetrie, dan C. Amalia, “ANALISIS PEMAHAMAN MAHASISWA PGSD UPI CIBIRU TERHADAP MATA KULIAH PEMBELAJARAN PKn DI SEKOLAH DASAR,” *J. Cakrawala Pendas*, vol. 8, no. 1, hal. 15–28, 2022.

[10] M. I. Shaufani, “ANALISIS PEMAHAMAN MAHASISWA TERHADAP MATA KULIAH PENGANTAR AKUNTANSI BERDASARKAN LATAR BELAKANG PENDIDIKAN DAN GENDER (Studi Empiris Pada Mahasiswa Program Studi Akuntansi Universitas Islam Kuantan Singingi),” *Univ. Islam Kuantan Singingi*, hal. 189–200, 2017.

[11] S. Salmu dan A. Solichin, “Prediksi Tingkat Kelulusan Mahasiswa Tepat Waktu Menggunakan Naïve Bayes : Studi Kasus UIN Syarif Hidayatullah Jakarta,” *Semin. Nas. Multidisiplin Ilmu 2017*, no. April, hal. 701–709, 2017.

[12] R. Hasudungan, “Naïve Bayes Model for Student Data Analysis,” *Int. J. Adv. Eng. Manag.*, vol. 3, no. 7, hal. 2931–2937, 2021, doi: 10.35629/5252-030729312937.

[13] I. K. Syahputra, F. A. Bachtiar, dan S. A. Wicaksono, “Implementasi Data Mining untuk Prediksi Mahasiswa Pengambil Mata Kuliah dengan Algoritme Naive Bayes,” *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 2, no. 11, hal. 5902–5910, 2018.

[14] M. winny Amelia, A. S. . Lumenta, dan A. Jacobus, “Prediksi Masa Studi Mahasiswa

dengan Menggunakan Algoritma Naïve Bayes," *J. Tek. Inform.*, vol. 11, no. 1, 2017, doi: 10.35793/jti.11.1.2017.17652.

[15] I. G. I. Suwardika, "Prediksi Tingkat Kelulusan Mahasiswa Tepat Waktu Menggunakan Naive Bayes: Studi Kasus Fakultas Ekonomi Dan Bisnis Universitas Pendidikan Nasional," *J. Ilmu Komput. Indones.*, vol. 4, no. 2, hal. 37–44, 2019, doi: 10.23887/jik.v4i2.2775.

[16] S. Rahmatullah, "Prediksi Tingkat Kelulusan Tepat Waktu Dengan Metode Naïve Bayes Dan K-Nearest Neighbor," *J. Inf. dan Komput.*, vol. 7, no. 1, hal. 7–16, 2019, doi: 10.35959/jik.v7i1.118.

[17] Y. Astuti, I. R. Wulandari, A. R. Putra, dan N. Kharomadhona, "Naïve Bayes untuk Prediksi Tingkat Pemahaman Kuliah Online Terhadap Mata Kuliah Algoritma Struktur Data," *JEPIN (J. Edukasi dan Penelit. Inform.)*, vol. 8, no. 1, hal. 28–32, 2022.

[18] E. Sabna dan M. Muhardi, "Penerapan Data Mining Untuk Memprediksi Prestasi Akademik Mahasiswa Berdasarkan Dosen, Motivasi, Kedisiplinan, Ekonomi, dan Hasil Belajar," *J. CoreIT J. Has. Penelit. Ilmu Komput. dan Teknol. Inf.*, vol. 2, no. 2, hal. 41, 2016, doi: 10.24014/coreit.v2i2.2392.

[19] A. Rohman dan S. Mujiyono, "Permodelan Prediksi Predikat Kelulusan Mahasiswa Menggunakan Decision Tree C4 . 5," vol. II, no. 2, hal. 1–5, 2021.

[20] R.- Annisa dan A.- Sasongko, "Prediksi Nilai Akademik Mahasiswa Menggunakan Algoritma Naïve Bayes," *JST (Jurnal Sains dan Teknol.*, vol. 9, no. 1, hal. 1–10, 2020, doi: 10.23887/jst-undiksha.v9i1.19488.

[21] A. Fauzi dan Tukiyat, "Analisis Potensi Dana Retail pada Nasabah PT . Bank Tabungan Negara (Persero), Tbk . Dengan Metode Decision Tree dan Naive Bayes Berbasis Optimize Selection (Evolutionary)," *J. Adm. Dan Manjemen*, vol. 9, no. 1, hal. 30–36, 2019.

[22] R. Aspiah dan Taghfirul Azhima Yoga Siswa, "Implementasi Correlation Based Feature Selection (Cfs) Untuk Peningkatan Akurasi Algoritma C4.5 Dalam Prediksi Performa Akademik Mahasiswa Berbasis Learning Management System," *J. Ilm. Betrik*, vol. 13, no. 2, hal. 199–207, 2022, doi: 10.36050/betrik.v13i2.523.

[23] N. Senan, R. Ibrahim, N. Mohd Nawi, I. T. R. Yanto, dan T. Herawan, "Rough set approach for attributes selection of traditional Malay musical instruments sounds classification," *Commun. Comput. Inf. Sci.*, vol. 151 CCIS, no. PART 2, hal. 509–525, 2011, doi: 10.1007/978-3-642-20998-7_59.

[24] R. Hasudungan dan W. J. Pranoto, "Implementasi Teorema Naïve Bayes Pada Prediksi Prestasi Mahasiswa," *J. Rekayasa Teknol. …*, vol. 5, no. 1, hal. 10–16, 2021.

[25] S. Samaray, "Implementasi Algoritma Rough Set dengan Software Rosetta untuk Prediksi Hasil Belajar," *J. Eksplora Inform.*, vol. 11, no. 1, hal. 57–66, 2022, doi: 10.30864/eksplora.v11i1.498.

[26] A. Aziz, "Implementasi Algoritma Rough Set Dan Naive Bayes Untuk Mendapatkan Rule Dalam Menyeleksi Pemohon Bantuan Fasilitas Rumah Ibadah (Studi Kasus : Pemerintah Kabupaten Pringsewu)," *Jl. ZA. Pagar Alam*, vol. 03, no. 93, hal. 74–83, 2020.

[27] A. Putra, Z. A. Matondang, N. Sitompul, I. Pendahuluan, dan A. Prediksi,

"Implementasi Algoritma Rough Set Dalam Memprediksi Kecerdasan Anak," *J. Pelita Inform.*, vol. 7, no. 2, hal. 149–156, 2018.

[28] R. Hasudungan, W. J. Pranoto, dan Rudiman, "Using MDA to Improve Naïve Bayes Classification for Students Performance Prediction," *JSE J. Sci. Eng.*, vol. 1, no. 2, hal. 65–70, 2020.

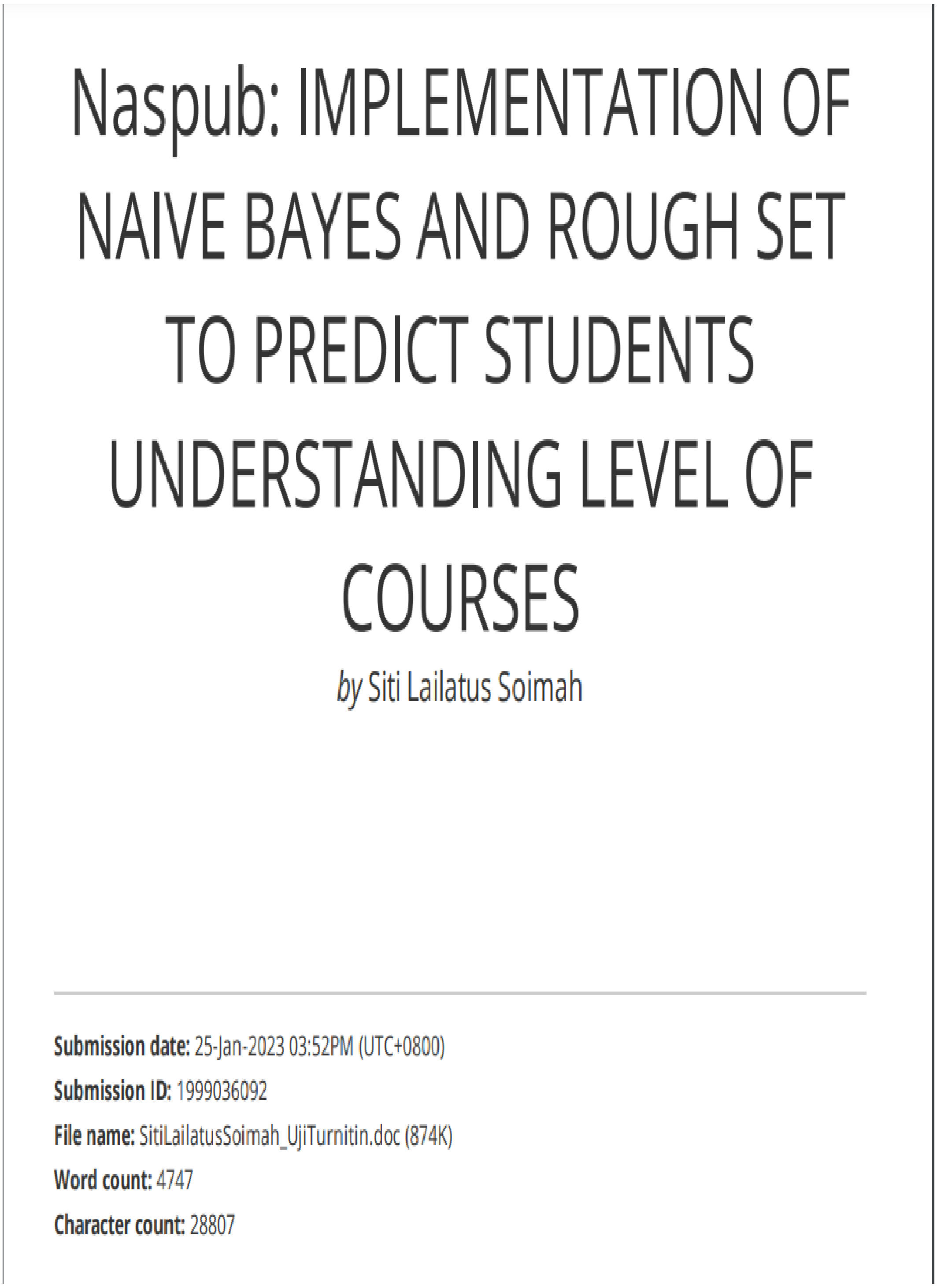
[29] T. Herawan, M. M. Deris, dan J. H. Abawajy, "A rough set approach for selecting clustering attribute," *Knowledge-Based Syst.*, vol. 23, no. 3, hal. 220–231, 2010, doi: 10.1016/j.knosys.2009.12.003.

[30] J. Suntoro, *Data Mining Algoritma Dan Implementasi dengan Pemograman PHP*, Kedua. Jakarta: PT Gramedia Jakarta, 2019.

[31] M. A. Muslim, B. Prasetiyo, E. L. H. Mawarni, dan A. J. Herowati, *Data Mining ALgoritma C4.5 Disertai contoh dan penerapanya dengan program komputer*. 2019.

[32] E. Suwandi, F. H. Imansyah, dan H. Dasril, "Analisis Tingkat Kepuasan Menggunakan Skala Likert pada Layanan Speedy yang Bermigrasi ke Indihome," *J. Tek. Elektro*, hal. 11, 2018.

[33] R. Hasudungan, "Analisis Indikator Kinerja Dosen Terhadap Prestasi Mahasiswa Semester Satu dengan Menggunakan Decision Tree," *J. Rekayasa Teknol. Inf.*, vol. 2, no. 2, hal. 192, 2018, doi: 10.30872/jurti.v2i2.1768.



Naspub: IMPLEMENTATION OF NAIVE BAYES AND ROUGH SET TO PREDICT STUDENTS UNDERSTANDING LEVEL OF COURSES

by Siti Lailatus Soimah

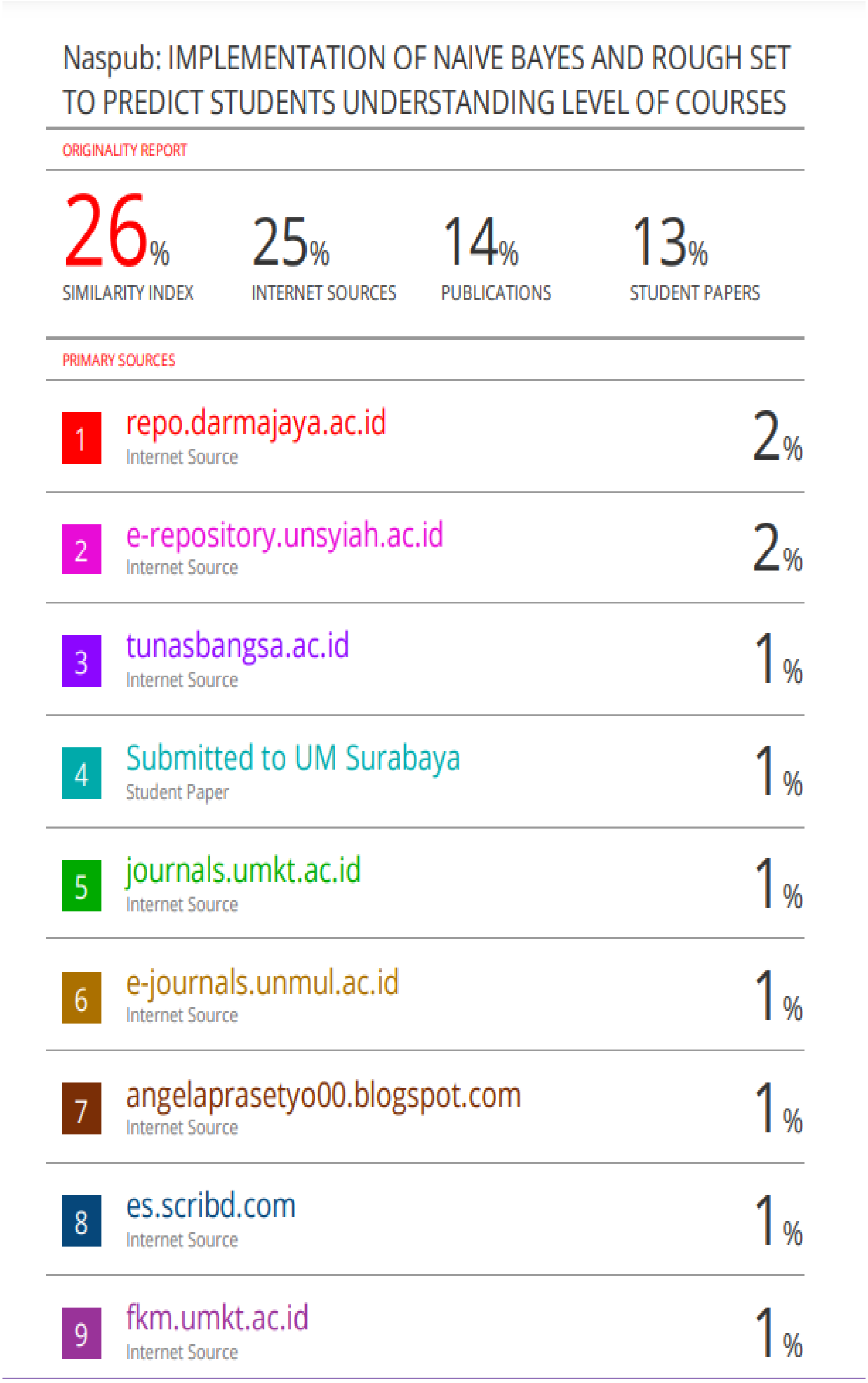
Submission date: 25-Jan-2023 03:52PM (UTC+0800)

Submission ID: 1999036092

File name: SitiLailatusSoimah_UjiTurnitin.doc (874K)

Word count: 4747

Character count: 28807



Naspub: IMPLEMENTATION OF NAIVE BAYES AND ROUGH SET TO PREDICT STUDENTS UNDERSTANDING LEVEL OF COURSES

ORIGINALITY REPORT

26%	25%	14%	13%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	repo.darmajaya.ac.id Internet Source	2%
2	e-repository.unsyiah.ac.id Internet Source	2%
3	tunasbangsa.ac.id Internet Source	1%
4	Submitted to UM Surabaya Student Paper	1%
5	journals.umkt.ac.id Internet Source	1%
6	e-journals.unmul.ac.id Internet Source	1%
7	angelaprasetyo00.blogspot.com Internet Source	1%
8	es.scribd.com Internet Source	1%
9	fkm.umkt.ac.id Internet Source	1%