**International Journal of Electrical and Computer Engineering (IJECE)**
Vol. 15, No. 6, December 2025, pp. 6001~6011
ISSN: 2088-8708, DOI: 10.11591/ijece.v15i6.pp6001-6011 

# Combination of rough set and cosine similarity approaches in student graduation prediction

**Ratna Yulika Go[1], Tinuk Andriyanti Asianto[1], Dewi Setiowati[1], Ranny Meilisa[1], Christine Cecylia Munthe[2], R. Hendra Kusumawardhana[3]**
[1]Departement of Informatic Engineering, Faculty of Computer Science, Universitas Esa Unggul, Jakarta, Indonesia
[2]National Research and Innovation Agency, Jakarta, Indonesia
[3]Department of Technology Information, Faculty Computer Science, Universitas Indonesia, Jakarta, Indonesia

**Article Info**

***Article history:***

Received Jan 28, 2025
Revised Jul 15, 2025
Accepted Sep 14, 2025

***Keywords:***

Case-based reasoning
Cosine similarity
K-fold
Rough set
Student graduation prediction

**ABSTRACT**

Higher education institutions must deliver high-quality education that produces graduates who are knowledgeable, skilled, creative, and competitive. In this system, students are a vital asset, and their timely graduation rate is an important factor to consider. In the Department of Computer Science, a challenge arises in distinguishing between students who graduate on time and those who do not. With a low on-time graduation rate of just 1.90% out of 158 graduates, this issue could negatively affect the institution's accreditation evaluation. This research employs the case-based reasoning method, enhanced with an indexing process using rough sets and a prediction process utilizing cosine similarity. The testing, conducted using k-fold validation with 60%, 70%, and 80% of the data, produced average accuracy rates of 64.2%, 66.3%, and 65.6%, respectively. The test results indicate that the highest average accuracy of 66.3% was achieved with 70% of the cases.



***Corresponding Author:***

Ratna Yulika Go
Department of Informatic Engineering, Faculty of Computer Science, Universitas Esa Unggul
Arjuna Utara Street 9, Duri Kepa, Kb. Jeruk, Jakarta 11510, Indonesia
Email: ratna.yulika@esaunggul.ac.id

## 1. INTRODUCTION

The highest educational institution is a university that organizes academic education for students. Students are often referred to as people with broader intellectual characteristics compared to their peers who are not students or other age groups below them [1]. With their intellectuality, students will be able to face and find problems systematically, which will later be applied in everyday life to compete in the world of work [2]. Universities are required to provide quality education for students to produce human resources that are knowledgeable, capable, creative, and competitive. In the education system, students are an important asset for an educational institution, so it is necessary to pay attention to the level of student graduation on time [3]. The percentage of ups and downs in students' ability to complete their studies on time is one of the elements of university accreditation assessment. For that, it is necessary to monitor and evaluate the tendency of students to graduate on time or not [4]. Monitoring or evaluation of performance will produce helpful information to help students, lecturers, administrators, and policymakers [5]. Thus, it is clear that predicting student graduation is important for education providers in determining strategies for their institutions. The percentage of on-time graduates at each university is generally smaller than that of late graduates. The current obstacle or problem in the computer science department is that it has been unable to classify alum data who graduate on time and late.

The inaccuracy of graduation time can be caused by several factors that can be identified by statistical analysis. One analysis that can be used is binary logistic regression analysis with response variables of on-time and late graduates [6]. The results of the binary logistic regression analysis indicate that the faculty of origin of the graduate, grade point average (GPA), gender of the graduate, scholarship, entry path, and minor affect the timeliness of graduation. This can be seen from the data obtained by researchers at the computer science department of University X, which has 2300 alums from the 2007-2024 class (computer science study program, 2017). Each class has a different study period. In the last five years (2019-2024), data from 260 alums shows that 25 people graduated on time (4 years) with a percentage of 1.90%. The study period was 4.5-5 years, with a percentage of 17.72% being 40 people. The study period of 5.5 years-6 years with a percentage of 21.52% was 85 people. Furthermore, the dominant study period in the Computer Science Department, with a total of 110 people, graduated with a study period of 6.5 years-7 years with a percentage of 58.86%. This shows that computer science students' graduation punctuality is still very low, with a percentage of 1.90%. This student graduation prediction system requires existing information to determine whether a student can graduate on time [7]. Suppose student graduation can be known early on; in that case, the academic party can implement a policy to minimize the number of students who do not graduate on time according to their study period. Based on these data, completing studies on time is important for both students and the Computer Science Department. Accreditation Department is an assessment of a department's eligibility. Alumni data, active students, and outstanding students are among the assessments in the accreditation of the computer science department. With the prediction of student graduation, it is hoped that it can be a reference for academics in setting strategies for their students so that they can complete their studies on time [8].

The right strategy for students still studying to complete their studies on time is to use the rough set method. The rough set method is a calculation method suitable for determining the level of student graduation [9]. Through the rough set method, it can be used to produce output in the form of student graduation predictions. The purpose of implementing this method is to help academics know the possibility of student graduation based on student data that has been stored. The benefits obtained are that the possibility of student graduation can be determined early on based on the knowledge obtained through the rough set method. After obtaining knowledge from the rough set method, the similarity process is continued to predict student graduation. In this study, researchers used cosine similarity to calculate the highest similarity value based on knowledge. These results produce a solution or prediction of student graduation [10].

Previous research also reviewed the prediction of student graduation using the k-nearest neighbor (k-NN) algorithm, with the problem being the percentage of ups and downs of student's ability to complete their studies on time as one of the elements of university accreditation assessment. The result is that the level of accuracy of testing the student graduation model using the k-NN algorithm using the attributes of gender, marital status, employment status, and *indeks prestasi semester* (IPS) I-IV is influenced by the number of data clustering. The highest accuracy and area under the curve (AUC) value is by clustering the 5$^{th}$ data. The accuracy value is 85.15% and the AUC value is 0.888 [11]. The latest research by Pelima *et al.* [12] entitled "Predicting the level of student graduation on time using naive Bayes". The problem is that it was found that the number of new students is greater than the number of students who graduated and have not been able to produce knowledge about this condition. The attributes used in this study are gender, type of selection, father's income, mother's education, IPS I-IV, and semester credits I-IV. The result is that the accuracy of the data testing obtained in this study is 80.72% of the 1162 data used for training data and 587 data for testing. Based on the problems and previous research, this study was conducted to measure and predict the graduation rate of computer science students at University X, which later the data and recommendations produced can be used as a reference in taking strategic steps for the study program. This study uses a cosine similarity approach with a rough set that is different from several previous studies that use a lot of k-NN and naive Bayes.

## 2. METHOD

### 2.1. Data

Data from the 2019-2024 batch, as many as 260 datasets, were used in this study. Of the 260 datasets used, 70% of the data is training data, and 30% of the data is test data. The attributes used are gender, Grade points in semester 2, and GPA in semester 4. The data obtained is given codes. This is useful for simplifying the calculation stage. The following are the codes and descriptions used for grade points (GP) and GPA defined by the symbol $(p, q, r, s, t)$. The symbol is identified the range of GPA that shared by expert (dean of faculty computer science). The details are: $p =$ 3,00-4,00; $q =$ 2,50-2,99; $r =$ 2,00-2,49; $s =$ 1,51-1,99; $t =\leq$ 1,50. And for gender is separate to gender it is $p$ for woman and $q$ for man. For recommendation divide by 4 categories: $a =$ graduated $\leq$ 4,00 years; $b =$ graduated 4,5–5,00 years; $c =$ graduated 5,5–6,00 years; and $d =$ graduated 6,5–7,00 years.

### 2.2. Case-based reasoning

In general, case-based reasoning case-based reasoning (CBR) is a concept of reasoning in problem-solving through case handling records that an expert has carried out. Case-based reasoning has four stages, which include [13]:

a. Retrieve: Getting/retrieving the most similar/relevant cases to the new case.
b. Reuse*:* Modeling/reusing knowledge and information from old cases based on the most relevant similarity weights into new cases.
c. Revise: Reviewing the proposed solution and then testing it on real cases (simulation).
d. Retain: Integrating/saving new cases that have successfully obtained solutions so that they can be used by subsequent cases similar to the case.

### 2.3. Pre-processing

In this study, pre-processing is carried out using the outlier function before the data is processed. Based on the data obtained, the data is classified as an outlier [14]. This means that the observation data that appears has extreme values or values that are far from most of the other values in its group. So, the outlier data needs to be cleaned in order to get good results. As many as 70% of the training data or 182 datasets using the outlier function produced 181 clean datasets. The results of the outlier function are then calculated using the indexing method, namely rough set.

### 2.4. Rough set

A rough set is a mathematical technique developed by Pawlak in 1991. The steps in determining the reduction in equivalence classes are as [15]:

a. Data representation:

Rough set is represented by two elements, namely information systems (IS) and decision systems (DS). An $IS$ is defined as a pair $IS = \{U, A\}$, where $U = \{e_1, e_2, \ldots, e_{\mathrm{m}}\}$ represents a set of cases, and $A = \{a_1, a_2, \ldots, a_{\mathrm{n}}\}$ represents a set of attributes. The information system in the context of the system can be illustrated in a Table 1.

Table 1. Information systems

| GPA | Gender | GP | Recommendation |
|---|---|---|---|
| q | q | q | b |
| q | p | q | b |
| r | q | q | b |
| q | q | r | b |
| q | q | q | b |

The data is an example of 5 cases, evaluated using the parameters GPA, gender, and IP. In an information system, each row represents an object, while each column represents an attribute, consisting of $m$ objects:

$U = \{e_1, e_2, \ldots, e_{\mathrm{m}}\}$: cases 1, 2, 3, ..., 20
$A = \{a_1, a_2, \ldots, a_{\mathrm{n}}\}$: GPA, gender, IP

In many applications, an outcome or classification decision is known, which is represented by a decision attribute, $C = \{C_1, C_2, \ldots, C_{\mathrm{p}}\}$. Therefore, the information system becomes:

$$IS = (U, \{A, C\}).$$

Each object in the system is described by values of these attributes, providing a structured way to capture information. When a special attribute representing decisions or outcomes is added, the system becomes a decision system, which facilitates classification and decision-making processes. Decision systems link the condition attributes with decision attributes, enabling the analysis of how different attribute combinations influence specific outcomes. This structure is fundamental in rough set theory, where it helps in identifying patterns, dependencies, and rules within data for knowledge discovery and reasoning.

b. Positive region

In rough set theory, the positive region represents the set of objects in the universe that can be certainly classified into specific decision classes based on the given condition attributes. It is formed by combining all the lower approximations of the decision attribute partitions. The lower approximation consists of objects whose equivalence classes, defined by condition attributes, are entirely included within a particular

decision class. By uniting these lower approximations, the positive region helps identify which data points can be definitively categorized without ambiguity. This concept is essential for evaluating the classification power of the attributes used in a decision system.

c. Equivalence class

In the positive region, cases with equivalent attribute values based on the decision attribute $C$. $C$ are grouped into equivalence classes. These classes consist of objects (cases) that share identical values for the condition attributes and are fully contained within a single decision class. The union of these equivalence classes that meet this criterion forms the positive region. This approach ensures that only those cases which can be certainly classified (without ambiguity) are included in the decision process. The equivalence class groups the same objects for attribute $A(U, A)$. The equivalence class table is shown in Table 2.

Table 2. Equivalence class

| Class | GPA | Gender | GP | Recommendation | Number of count (NOC) |
|---|---|---|---|---|---|
| Equiv_1 | q | q | q | b | 2 |
| Equiv_2 | q | p | q | b | 1 |
| Equiv_3 | r | q | q | b | 2 |
| Equiv_4 | q | q | r | b | 1 |
| Equiv_5 | q | q | q | b | 1 |
| Equiv_6 | q | q | q | b | 1 |
| Equiv_7 | r | q | s | c | 1 |
| Equiv_8 | r | p | r | c | 1 |
| Equiv_9 | q | p | p | d | 1 |
| Equiv_10 | r | q | s | d | 1 |

Compare each class; if there is a difference in any class attribute, record it in the discernibility matrix table. If all attributes are the same, mark it with a cross (Null). The attributes are represented as follows: GPA, gender, and IP. To evaluate the diagnostic performance of an algorithm, the best algorithm is one that not only demonstrates strong performance but also has the potential to accurately diagnose data.

d. Discernibility matrix

At the discernibility matrix stage, the data in the form of a table is processed by comparing and considering only the condition variables. From this stage, the process of selecting minimal variables from a set of condition variables is carried out using the prime implicant Boolean function. A prime implicant in a Boolean function is a fundamental concept used in the simplification of logic expressions. It refers to a group of one or more minterms that can be combined because they differ in only one variable, and this group represents a product term that cannot be combined any further without losing its ability to represent parts of the function. In the process of minimization, prime implicants are used to cover the output values of 1 (true) in a function. Among these, essential prime implicants are those that cover at least one minterm not covered by any other prime implicant, making them necessary components of the simplified expression. Prime implicants play a critical role in techniques like Karnaugh maps and the quine-McCluskey method, helping derive the most efficient logic circuit representation [16].

The result of knowledge is the indexing process used for the cosine similarity process. After the indexing process generates an index for each case in the case database, the case retrieval process is limited to only those cases that have the same index as the new case being tested. After the indexing process is completed, the next stage is retrieval. In this stage, the similarity values are calculated using the cosine similarity method for the old cases that share the same index as the new input case. The similarity values are computed based on (1), where the similarity is calculated one by one for each case with the same index. Once the similarity values have been calculated for all matching-index cases, a total of 8 similarity values is obtained.

e. Cosine similarity

Cosine similarity is a calculation of the similarity between two n-dimensional vectors by finding the cosine of the angle between them. It is often used to compare documents in text mining [17]. The formula for cosine similarity is as:

$$\text{Similarity}(\text{x}, \text{y}) = \cos(\Theta) = \frac{\text{x}.\,\text{y}}{||\text{x}||||\text{y}||} = \frac{\sum_{i=1}^{n} x_i .\, y_i}{\sqrt{\sum_{i=1}^{n} (x_i)^2} \cdot \sqrt{\sum_{i=1}^{n} (y_i)^2}} \quad (1)$$

where:
$x.y = vector\ dot\ product\ from\ x\ and\ y, calculated\ by\ \sum_{i=1}^{n} x_i.\ y_i$
$||x|| = length\ of\ vektor\ x\ calculated\ by\ \sum_{i=1}^{n}(x_i)^2$
$||y|| = length\ of\ vektor\ y\ calculated\ by\ \sum_{i=1}^{n}(y_i)^2$
$x_i = (x_1, x_2, x_3, \ldots\ldots, x_n) =\ values\ on\ new\ case$
$y_i = (y_1, y_2, y_3, \ldots\ldots, y_n) = values\ on\ old\ case$
$n\ = amount\ of\ value$

Tan and Kumar [18] explain that the greater the result of the similarity function, the more similar the two objects being evaluated are considered. If vice versa, the smaller the result of the similarity function, the more different the two objects are considered. In a function that produces values in the range [0...1], a value of 1 represents that the two objects are the same, while a value of 0 represents that the two objects are entirely different [19].

### 2.5. As-is analysis

In the current system, alum data still uses a manual system to summarize and store alum data. Alumni personal data is stored in alum files, while value data, graduation dates, class years, and academic information are stored in Microsoft Excel. So, it becomes a deficiency in predicting student graduation in the computer science department.

### 2.6. To-be analysis

The system to be built can predict student graduation. The system has 4 process stages: retrieve, reuse, revise, and retain. The way the system works, in general, is guided by the knowledge base owned by the system, which is sourced from alum data, which is then calculated for its similarity level with new cases entered by the user. Based on the similarity level of the case, the system will issue the results of student graduation predictions, which are expected to help the study program or system users predict student graduation.

### 2.7. System testing method

System testing utilized the k-fold cross validation method, where the dataset was randomly divided into K partitions (folds). Subsequently, K iterations of experiments were conducted. In each iteration, one distinct fold was used as the testing set, while the remaining K – 1 folds were used for training. This approach ensures that every data point is used for both training and testing, thereby enhancing the robustness and generalizability of the model evaluation.

## 3. RESULTS AND DISCUSSION

### 3.1. Preprocessing

Before the data is processed, preprocessing is carried out by performing outliers. To find out the truth of the system's accuracy results, researchers use the outlier function to find out what percentage of data is biased. Because the more outlier data there are, the less biased data there will be, and vice versa. The results of the outliers obtained 1 data that was significant enough so that the data was cleaned or deleted. Of the 182-training data after using outliers, only 1 was lost, so the clean data obtained was 181. The results of this study obtained indexing of 10 rules/knowledge with the stages of predicting graduation:

a. Retrieve

At this stage, the similarity value is calculated using the cosine similarity method on the old data according to the index that has similarities with the new data entered.

$$Similarity\ (x,y) = cos\ (\theta) = \frac{x.y}{||x||||y||}$$

$x.y = 8$
$||x|| = 3.16$
$||y|| = 2.83$

$$sim\ (x,y) = \frac{8}{3.16.2.83} = 0.89$$

And the result of every case can be seen at Table 3. A selection stage will be carried out after getting the similarity calculation value between the cases in the case base and the new cases. In the selection stage, the similarity values will be sorted from the highest to the lowest value, and the highest value will be sought.

From Table 3, it can be seen that case no. 5 has the highest similarity value with the new case entered. The CBR system will recommend case no. 5 as a solution. To evaluate the diagnostic performance of an algorithm, the best algorithm is one that not only demonstrates strong performance but also has the potential to accurately diagnose data.

Table 3. Similarity results

| No | GPA | Gender | GP | Recommendation | Sim (x, y) |
|---|---|---|---|---|---|
| 1 | q | q | q | b | 0.89 |
| 2 | q | p | q | b | 0.84 |
| 3 | r | q | q | b | 0.96 |
| 4 | q | q | r | b | 0.95 |
| **5** | **q** | **q** | **q** | **b** | **0.99** |
| 6 | q | q | q | b | 0.89 |
| 7 | r | q | s | c | 0.83 |
| 8 | r | p | r | c | 0.83 |
| 9 | q | p | p | d | 0.80 |
| 10 | r | q | s | d | 0.80 |

b. Reuse

Reuse is a stage in case-based reasoning where, at this stage, the cases stored in the case base are retrieved to be used as a solution [20], [21]. The criteria for selecting a case are cases with the highest calculation results carried out in the previous stage, namely retrieve. Based on the calculations in Table 3, one old case was obtained that had the highest level of similarity to the new case compared to the other cases, namely case no. 5, with a similarity value of 0.99. So, the predicted results obtained for the new case with GPA=3.50 are graduating with a study period of 4.5-5 years.

c. Revise

The revision stage is the process of reviewing the case and the solutions provided. If there are errors, then improvements will be made to overcome the errors that occur. Experts carry out the revision process. Experts can use the re-instantiation method to adapt cases/revisions. If the solution or system prediction results are correct with the actual graduation time, then there is no need to revise, and the new solution can be directly stored in the knowledge base. However, if the prediction results do not match the graduation time, a revision of the graduation time is carried out and then stored in the knowledge base [14], [22].

d. Retain

After the revision process is complete and a genuinely correct solution has been found, the expert will add the new case data that has been found to the knowledge base, which will be stored as student record data. However, if the solution to the new case already exists in the knowledge base, it does not need to be added because there will be duplication of data/solutions [23], [24]. New case data not yet in the knowledge base can be used later for the next case. This process is called the retain process. 260 datasets were divided into 181 training data, or 70% and 78 test data, or 30%. The attributes used are GPA semester 4, gender, and GPA semester 2. Through the indexing process, 10 rules were obtained that will be used for the retrieval process using cosine similarity. New cases will be calculated for similarity based on the indexing results. For example, in the previous chapter, students with a GPA of semester 2=2.30 and a GPA of semester 4=3.50 with male gender were calculated using the cosine similarity method. The attribute values, namely GPA and GPA, were converted to facilitate the calculation process, and the gender attribute value was first converted into an actual value. The results of calculating the case similarity value were obtained at 0.99 or 99% with a recommended study period of 4.5-5 years. Based on these results, the accuracy or similarity of new and old cases is very high.

Furthermore, from this temporary solution, revising student graduation predictions cannot be used because it waits for the student to graduate. If the prediction result is correct, then the prediction result can be stored in the case base [25], [26]. However, if the prediction results do not match the length of the student's study, the expert revises the prediction results and stores them on a case basis. The highest result of k-fold 60% 727% and the average is 63.505%. For the highest result of k-fold 70% is 78.788% with 10 epochs, and the average is 65.169%. And for k-fold 80% is 81.818% with the average is 64.447%

From the results of the system accuracy test by applying different training data, namely 60%, 70%, and 80% at k-10, the largest k results were obtained at k-7 with 72.727%, 78.788%, and 81.818%, respectively, with the largest k value in training data 82%. Based on the data above, the average system accuracy was obtained by applying different training data and the effect of the number of cases on the average accuracy. The results can be seen in the graph shown in Figure 1.

The average accuracy of each trained data is 60%=63.505%, 70%=65.169%, and 80%=64.447%. It can be concluded that the highest average value is in the number of cases 70%. This indicates that as much as 70% of the trained data has a stability value compared to the number of other trained cases. So, the number of cases does not affect the average accuracy value. Based on the hypothesis using the confusion matrix calculation, the system accuracy hypothesis is ≥70%, while the highest system accuracy result is 65.169%. This shows that the system accuracy results after testing the percentage value are lower than the initial hypothesis, which is 65.169% less than 70%.

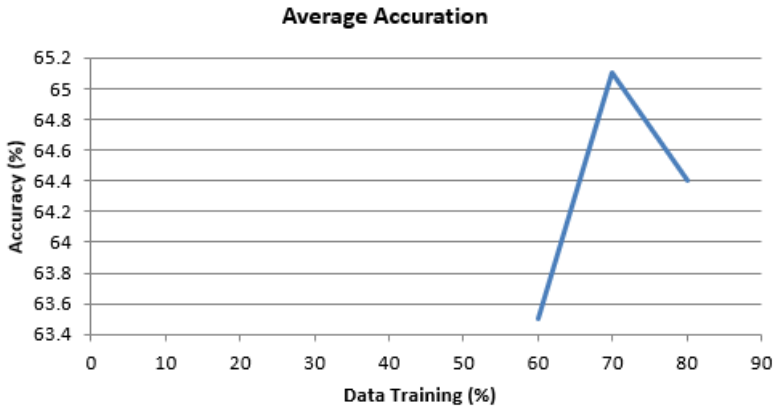


Figure 1. System test graph

## 4. DISCUSSION

In general, previous researchers have carried out research using the rough set method. In the research entitled predicting students' on-time and untimely graduation with k-means, the problem is that many previous research algorithms have been carried out with results and accuracy levels that are not yet close to perfect in predicting students' on-time graduation, so this study aims to improve the results and accuracy levels. From the results of the research that has been carried out, the creation of the k-means algorithm and neural network using student graduation data with attributes of class, gender, and IPS (Semester achievement index) I-IV, the resulting models are compared to determine the best method in predicting students' on-time and untimely graduation. To measure the model's performance, the confusion matrix shows that the k-means algorithm method produces an accuracy value of 84.43%. The neural network method produces an accuracy value of 90.41% [27].

Furthermore, research entitled comparing classification algorithm of data mining to predict the graduation of students in time with the problem being the low percentage of students who graduate on time. The method compares the decision tree algorithm, naive Bayes, ANN, support vector machine (SVM), and logistic regression (LR) with the attributes used: faculty, gender, age, and GPA semester I-IV. The results are the accuracy of the decision tree algorithm 80.01%, naive Bayes 75.16%, ANN 100%, SVM 100%, and LR 100% [28]–[30]. In the following year, a study entitled Predicting on-time graduation for new students with data mining with the problem being the number of students who were able to complete their studies on time in the 2019-2024 period was less than 10% so early efforts were needed to find out what parameters influenced a student to be able to complete their studies on time. From the test results using attributes of gender, religion, NEM, major, and profession by applying the k-NN method and using sample data from alums of the 2017-2022 graduation years for old cases and alums data of the 2023 graduation years for new cases, an accuracy level of 83.36% was obtained [31]. In the same year, a study entitled Predicting Student Graduation using the k-NN method was conducted by [32]. The problem was the low percentage of students who graduated on time, so this study aimed to determine the percentage value of student graduation using the k-NN method. In the k-NN method, the data used in the prediction is 167 data and 7 attributes, namely gender, residence status, transportation status, marital status, regional origin, school type, and Undana entrance route. The accuracy value using the k-NN method is 80% [33], [34].

Further research entitled k-NN algorithm model for student graduation prediction with the problem being the percentage of ups and downs in students' ability to complete their studies on time, which is one of the elements of university accreditation assessment. The result is the level of accuracy of testing the student graduation model using the k-NN algorithm using gender, marital status, employment status, and IPS I-IV attributes influenced by the number of data clustering. The highest accuracy and AUC value are obtained by clustering the 5th data. The accuracy value is 85.15% and the AUC value is 0.888 [35], [36].

The latest research was conducted by Atmaja *et al.* [37], titled predicting the level of student graduation on time using naive Bayes. The problem is that it was found that the number of new students is more than the number of students who graduated and have not been able to produce knowledge of this condition. The attributes used in this study are gender, type of selection, father's income, mother's education, IPS I-IV, and semester credits I-IV. The result is the accuracy of testing the data obtained in this study, which is 80.72%, from 1162 data used for training data and 587 data for testing [38].

## 5. CONCLUSION

Based on the collected datasets of 260, with 70% training data from 182 datasets and 30% test data from 78 datasets, the average accuracy of each training data was 60%=63.505%, 70%=65.169%, and 80%=64.447%. It can be concluded that the highest average value is in the number of cases of 70%. This indicates that the trained data of 70% has a stability value compared to the number of other cases trained. The researcher's analysis of the low accuracy of the system is caused by the few attributes used, namely semester 2 GP, semester 4 GPA, and gender, as well as data imbalance or accumulation of the most considerable student study period at 6.5-7 years with a percentage of 58.86% or 110 people so that the category of study period ≤ 4 years with a percentage of 1.90% or 25 people were eliminated. This causes biased data so that the system's accuracy is not optimal. Suggestions that can be given for further development are to increase the weight of the attributes to be used so that the results of the prediction process and system testing have good values.

## FUNDING INFORMATION

Authors state no funding involved.

## AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

| Name of Author | C | M | So | Va | Fo | I | R | D | O | E | Vi | Su | P | Fu |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Ratna Yulika Go | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | | ✓ | ✓ | ✓ | | | ✓ | |
| Tinuk Andriyanti Asianto | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | | ✓ | ✓ | ✓ | | | ✓ | |
| Dewi Setiowati | | ✓ | | | | ✓ | | ✓ | ✓ | ✓ | ✓ | ✓ | | |
| Ranny Meilisa | ✓ | | ✓ | ✓ | | | ✓ | | | ✓ | ✓ | | ✓ | ✓ |
| Christine Cecylia Munthe | ✓ | | ✓ | ✓ | | | ✓ | | | ✓ | ✓ | | ✓ | ✓ |
| R. Hendra Kusumawardhan | | ✓ | | | | ✓ | | ✓ | ✓ | ✓ | ✓ | ✓ | | |

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

## CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

## DATA AVAILABILITY

Data availability is not applicable to this paper as no new data were created or analyzed in this study.

## REFERENCES

[1] C. Romero and S. Ventura, "Guest editorial: special issue on early prediction and supporting of learning performance," *IEEE Transactions on Learning Technologies*, vol. 12, no. 2, pp. 145–147, Apr. 2019, doi: 10.1109/TLT.2019.2908106.
[2] F. C. Huang, H. Mohamadipanah, F. A. Mussa-Ivaldi, and C. M. Pugh, "Combining metrics from clinical simulators and sensorimotor tasks can reveal the training background of surgeons," *IEEE Transactions on Biomedical Engineering*, vol. 66, no. 9, pp. 2576–2584, Sep. 2019, doi: 10.1109/TBME.2019.2892342.
[3] Y. Nieto, V. Gacia-Diaz, C. Montenegro, C. C. Gonzalez, and R. Gonzalez Crespo, "Usage of machine learning for strategic decision

making at higher educational institutions," *IEEE Access*, vol. 7, pp. 75007–75017, 2019, doi: 10.1109/ACCESS.2019.2919343.
[4] D. Sobnath, T. Kaduk, I. U. Rehman, and O. Isiaq, "Feature selection for UK disabled students' engagement post higher education: a machine learning approach for a predictive employment model," *IEEE Access*, vol. 8, pp. 159530–159541, 2020, doi: 10.1109/ACCESS.2020.3018663.
[5] M. A. Prada *et al.*, "Educational data mining for tutoring support in higher education: a web-based tool case study in engineering degrees," *IEEE Access*, vol. 8, pp. 212818–212836, 2020, doi: 10.1109/ACCESS.2020.3040858.
[6] F. M. Almutairi, N. D. Sidiropoulos, and G. Karypis, "Context-aware recommendation-based learning analytics using tensor and coupled matrix factorization," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 5, pp. 729–741, Aug. 2017, doi: 10.1109/JSTSP.2017.2705581.
[7] A. J. Fernandez-Garcia, R. Rodriguez-Echeverria, J. C. Preciado, J. M. C. Manzano, and F. Sanchez-Figueroa, "Creating a recommender system to support higher education students in the subject enrollment decision," *IEEE Access*, vol. 8, pp. 189069–189088, 2020, doi: 10.1109/ACCESS.2020.3031572.
[8] J. Xu, K. H. Moon, and M. van der Schaar, "A machine learning approach for tracking and predicting student performance in degree programs," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 5, pp. 742–753, Aug. 2017, doi: 10.1109/JSTSP.2017.2692560.
[9] Y. Ouyang, T. Feng, R. Gao, Y. Xu, and J. Liu, "Prediction of graduation development based on hypergraph contrastive learning with imbalanced sampling," *IEEE Access*, vol. 11, pp. 89881–89895, 2023, doi: 10.1109/ACCESS.2023.3301878.
[10] D. Uliyan, A. S. Aljaloud, A. Alkhalil, H. S. Al Amer, M. A. E. A. Mohamed, and A. F. M. Alogali, "Deep learning model to predict students retention using BLSTM and CRF," *IEEE Access*, vol. 9, pp. 135550–135558, 2021, doi: 10.1109/ACCESS.2021.3117117.
[11] D. Y. Putri, R. Andreswari, and M. A. Hasibuan, "Analysis of students graduation target based on academic data record using C4.5 algorithm case study: information systems students of Telkom University," in *2018 6th International Conference on Cyber and IT Service Management (CITSM)*, Aug. 2018, pp. 1–6, doi: 10.1109/CITSM.2018.8674366.
[12] L. R. Pelima, Y. Sukmana, and Y. Rosmansyah, "Predicting university student graduation using academic performance and machine learning: a systematic literature review," *IEEE Access*, vol. 12, pp. 23451–23465, 2024, doi: 10.1109/ACCESS.2024.3361479.
[13] S. Wibowo, R. Andreswari, and M. A. Hasibuan, "Analysis and design of decision support system dashboard for predicting student graduation time," in *2018 5th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI)*, Oct. 2018, pp. 684–689, doi: 10.1109/EECSI.2018.8752876.
[14] A. A. Mohamed Elhadi Hussen and A. Saikhu, "Modeling of student graduation prediction using the naive Bayes classifier algorithm," in *2024 3rd International Conference on Creative Communication and Innovative Technology (ICCIT)*, Aug. 2024, pp. 1–8, doi: 10.1109/ICCIT62134.2024.10701117.
[15] P. Zdzisław, "Rough sets," *Polish Academy of Sciences*, pp. 1–51, 1991.
[16] O. Čepek, P. Kučera, and S. Kuřík, "Boolean functions with long prime implicants," *Information Processing Letters*, vol. 113, no. 19–21, pp. 698–703, Sep. 2013, doi: 10.1016/j.ipl.2013.07.001.
[17] J. Ye, "Cosine similarity measures for intuitionistic fuzzy sets and their applications," *Mathematical and Computer Modelling*, vol. 53, no. 1–2, pp. 91–97, Jan. 2011, doi: 10.1016/j.mcm.2010.07.022.
[18] P.-N. Tan and V. Kumar, "Mining indirect associations in web data," in *WEBKDD 2001 — Mining Web Log Data Across All Customers Touch Points*, Berlin, Heidelberg: Springer, 2002, pp. 145–166, doi: 10.1007/3-540-45640-6_7.
[19] B. de Ville, *Introduction to Data Mining*. Microsoft Data Mining, 2001, doi: 10.1016/b978-155558242-5/50003-6.
[20] S. Mehta, "Playing smart with numbers: predicting student graduation using the magic of naive Bayes," *International Transactions on Artificial Intelligence (ITALIC)*, vol. 2, no. 1, pp. 60–75, Nov. 2023, doi: 10.33050/italic.v2i1.405.
[21] S. MaryJohn Rukmony and S. Gnanamony, "Rough set method-cloud internet of things: a two-degree verification scheme for security in cloud-internet of things," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 13, no. 2, pp. 2233-2239, Apr. 2023, doi: 10.11591/ijece.v13i2.pp2233-2239.
[22] S. Kumar Mamdy and V. Petli, "Modified fuzzy rough set technique with stacked autoencoder model for magnetic resonance imaging based breast cancer detection," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 14, no. 1, pp. 294–304, Feb. 2024, doi: 10.11591/ijece.v14i1.pp294-304.
[23] S. T. Karamouzis and A. Vrettos, "Sensitivity analysis of neural network parameters for identifying the factors for college student success," in *2009 WRI World Congress on Computer Science and Information Engineering*, 2009, pp. 671–675, doi: 10.1109/CSIE.2009.592.
[24] Y. Elfakir, G. Khaissidi, M. Mrabti, D. Chenouni, and M. Boualam, "Combined cosine-linear regression model similarity with application to handwritten word spotting," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 10, no. 3, pp. 2367–2374, Jun. 2020, doi: 10.11591/ijece.v10i3.pp2367-2374.
[25] K. Lin, "Research on graduation destination prediction algorithm based on students' learning behavior data," in *2023 7th Asian Conference on Artificial Intelligence Technology (ACAIT)*, Nov. 2023, pp. 1073–1079, doi: 10.1109/ACAIT60137.2023.10528473.
[26] S. Al-Otaibi *et al.*, "Cosine similarity-based algorithm for social networking recommendation," *International Journal of Electrical and Computer Engineering*, vol. 12, no. 2, pp. 1881–1892, Apr. 2022, doi: 10.11591/ijece.v12i2.pp1881-1892.
[27] D. P. Ong and J. R. I. Pedrasa, "Student risk assessment: predicting undergraduate student graduation probability using logistic regression, SVM, and ANN," in *TENCON 2021 - 2021 IEEE Region 10 Conference (TENCON)*, Dec. 2021, pp. 105–110, doi: 10.1109/TENCON54134.2021.9707322.
[28] L. Cahaya, L. Hiryanto, and T. Handhayani, "Student graduation time prediction using intelligent k-medoids algorithm," in *2017 3rd International Conference on Science in Information Technology (ICSITech)*, Oct. 2017, pp. 263–266, doi: 10.1109/ICSITech.2017.8257122.
[29] C. Wirawan, E. Khudzaeva, T. H. Hasibuan, Karjono, and Y. H. K. Lubis, "Application of data mining to prediction of timeliness graduation of students (a case study)," in *2019 7th International Conference on Cyber and IT Service Management (CITSM)*, Nov. 2019, pp. 1–4, doi: 10.1109/CITSM47753.2019.8965425.
[30] N. Aldhafferi *et al.*, "Learning trends in customer churn with rule-based and kernel methods," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 12, no. 5, pp. 5364–5374, Oct. 2022, doi: 10.11591/ijece.v12i5.pp5364-5374.
[31] W. Prachuabsupakij and P. Doungpaisan, "Matching preprocessing methods for improving the prediction of student's graduation," in *2016 2nd IEEE International Conference on Computer and Communications (ICCC)*, Oct. 2016, pp. 33–37, doi: 10.1109/CompComm.2016.7924659.
[32] R. K. Halder, M. N. Uddin, M. A. Uddin, S. Aryal, and A. Khraisat, "Enhancing K-nearest neighbor algorithm: a comprehensive review and performance analysis of modifications," *Journal of Big Data*, vol. 11, no. 1, p. 113, Aug. 2024, doi: 10.1186/s40537-024-00973-y.
[33] Hartatik, K. Kusrini, and A. Budi Prasetio, "Prediction of student graduation with naive Bayes algorithm," in *2020 Fifth International Conference on Informatics and Computing (ICIC)*, Nov. 2020, pp. 1–5, doi: 10.1109/ICIC50835.2020.9288625.

[34] P. T. and N. G. S., "Social-sine cosine algorithm-based cross layer resource allocation in wireless network," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 11, no. 1, pp. 458–470, Feb. 2021, doi: 10.11591/ijece.v11i1.pp458-470.
[35] E. Purnamasari, D. P. Rini, and Sukemi, "Prediction of the student graduation's level using C4.5 decision tree algorithm," in *2019 International Conference on Electrical Engineering and Computer Science (ICECOS)*, Oct. 2019, pp. 192–195, doi: 10.1109/ICECOS47637.2019.8984493.
[36] A. Salam, J. Zeniarja, and D. M. Anthareza, "Student graduation prediction model using deep learning convolutional neural network (CNN)," in *2022 International Seminar on Application for Technology of Information and Communication (iSemantic)*, Sep. 2022, pp. 362–366, doi: 10.1109/iSemantic55962.2022.9920449.
[37] K. J. Atmaja, I. P. Y. Indrawan, I. M. D. P. Asana, I. K. Wawan, and A. G. C. Udayanie, "Naïve Bayes-based student graduation prediction model: effectiveness and implementation to improve timely graduation," *Journal of Computer Networks, Architecture and High Performance Computing*, vol. 6, no. 3, pp. 1442–1450, Jul. 2024, doi: 10.47709/cnahpc.v6i3.4408.
[38] Laurentinus, O. Rizan, Sarwindah, Hamidah, R. Sulaiman, and P. Fuston, "Data mining using C4.5 algorithm in predicting student graduation," in *2022 5th International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, Dec. 2022, pp. 738–743, doi: 10.1109/ISRITI56927.2022.10052793.

## BIOGRAPHIES OF AUTHORS

**Ratna Yulika Go** received her M.T.I. degree in information technology from Universitas Indonesia. She is a lecturer at the Department of Informatics Engineering, Faculty of Computer Science, Universitas Esa Unggul, Jakarta, Indonesia. Her research interests include supply chain management, supply chain risk management, quality management systems, industrial quality, digital business information systems, e-learning, project management, ERP, e-business, engineering, and technology. She has been actively involved in various academic and research activities related to these fields. She is also committed to advancing knowledge in digital transformation and quality improvement through technology adoption in businesses and education. She can be contacted at email: ratna.yulika@esaunggul.ac.id.

**Tinuk Andriyanti Asianto** has over 24 years of experience in the ICT industry, with expertise in business development, government relations, corporate affairs, customer engagement, finance, technical operations, project and people management, and resource mobilization. She is also experienced in governance practices based on ISO 27001 and ISO 9001:2015 standards. Throughout her career, she has been involved in various strategic initiatives to support organizational growth and operational excellence, particularly in aligning technology with business objectives. Her strong background in both technical and managerial aspects enables her to contribute effectively to the success of ICT projects and corporate governance. She can be contacted via email at: tinuk.andriyanti@esaunggul.ac.id.

**Dewi Setiowati** received her M.Tr.Kom. degree in computer science from Politeknik Elektronika Negeri Surabaya. She is a lecturer at the Department of Informatics Engineering, Faculty of Computer Science, Universitas Esa Unggul, Jakarta, Indonesia. Her research interests include object-oriented programming, SQL, web development, software development, and C++. She is actively involved in teaching and research activities related to these fields and is dedicated to enhancing students' knowledge and skills in software engineering and information technology. She also participates in academic initiatives to support the development of digital competencies in higher education. She can be contacted via email at: dewi.setiowati@esaunggul.ac.id.

**Ranny Meilisa** received her M.Pd.T. degree in technic from Universitas Negeri Padang. She is a lecturer at the Department of Informatics Engineering, Faculty of Computer Science, Universitas Esa Unggul, Jakarta, Indonesia. Her research interests include informatics engineering, programming languages, vocational education, and instructional media. She is actively engaged in teaching and research activities related to these fields, focusing on enhancing the quality of education through innovative learning methods and the integration of technology. She also contributes to the development of instructional media to support effective learning in vocational and higher education. She can be contacted via email at: meilisa.rannya@gmail.com.

**Christine Cecylia Munthe** received her M.T.I. degree in information technology from Universitas Indonesia. She is a researcher at the National Research and Innovation Agency (Badan Riset dan Inovasi Nasional/BRIN). Her research interests include user experience (UX) and user interface (UI). She is actively involved in research projects focusing on improving digital product usability and enhancing user interaction with technology. Her work contributes to the advancement of human-computer interaction studies in Indonesia. She can be contacted via email: Chri010@brin.go.id.

**R. Hendra Kusumawardhana** received his M.T.I. degree in information technology from Universitas Indonesia. He is currently serving as a Data Center Infrastructure Manager at PT Taspen (Persero). His research interests include enterprise resource planning (ERP), data center management, and digital transformation. He is actively involved in managing and developing IT infrastructure to support organizational digitalization and operational efficiency. His expertise contributes to advancing technology adoption in the financial services sector, particularly in data center operations and enterprise systems. He can be contacted via email at: r.hendra.k@gmail.com.