International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys. , No. (Mon-20..)

http://dx.doi.org/10.12785/ijcds/XXXXXX

Comparative Analysis of Naive Bayes and K-NN Approaches to Predict Timely Graduation using Academic History

Imam Riadi[1], Rusydi Umar[2] and Rio Anggara[3]

[1]*Department of Information System, Universitas Ahmad Dahlan, Ringroad Selatan, Banguntapan, Bantul, Yogyakarta 55191, Indonesia*
[2]*Departement of Informatics, Universitas Ahmad Dahlan, Ringroad Selatan, Banguntapan, Bantul, Yogyakarta 55191, Indonesia*
[3]*Master Program of Informatics, Universitas Ahmad Dahlan, Yogyakarta, 55164 Indonesia*

Received Mon. 20, Revised Mon. 20, Accepted Mon. 20, Published Mon. 20

Abstract: Graduation is an important benchmark in higher education accreditation and community assessment. This research seeks to ensure timely graduation for all students, considering the important role of higher education in this regard. To achieve this goal, data sets that cover academic performance during undergraduate and graduate studies are essential. The dataset includes details of universities, study programs, undergraduate and master's GPAs, TOEFL scores, and duration of study. Utilizing classification techniques in data mining, specifically Naive Bayes and K-Nearest Neighbors (K-NN), a comparative analysis was conducted to accurately predict graduate students' on-time completion. The graduation prediction process begins with data preprocessing, transformation, and division into training and testing sets. Next, the method is applied, and analysis is carried out to predict graduation outcomes. Experimental findings show that the Naive Bayes technique achieves an accuracy rate of 80%, while K-NN reaches 73%. Notably, Naive Bayes demonstrated superior efficacy in predicting on-time graduation. Efforts to improve accuracy require expanding data sets and diversifying variables. The findings of this research can guide academic institutions in implementing proactive measures to support students in completing their studies within the expected timeframe.

Keywords: Prediction, Graduation, Naive Bayes, K-Nearest Neighbor, Academic History, Confusion Matrix

1. INTRODUCTION

Continuing education after undergraduate studies is a prevalent practice, with students aspiring to complete their respective academic programs. However, there remains a significant cohort of students who exhibit limited engagement in educational pursuits, characterized by irregular attendance and sporadic participation in classroom activities for various reasons[1]. Numerous factors contribute to disparities in students capabilities, potentially impeding their advancement in completing final projects. The aim is to mitigate these skill differentials and facilitate timely graduation for all students, a significant milestone eagerly anticipated by individuals in academia. Subsequent to graduation, post-graduation GPAs serve as a benchmark for incoming students, parents, and peers in evaluating a student's academic aptitude[2]. The quality of student service holds paramount importance for the sustainability of an educational institution. Students serve as the focal point of higher education management, representing the primary clientele[3].

Ahmad Dahlan University (UAD) situated in Yogyakarta, stands as one of the esteemed institutions within the Muhammadiyah organization. The timely graduation of UAD students holds significant relevance, considering its pivotal role in sustaining campus accreditation and meeting academic requisites. This research is predicated upon the cohort of UAD students who did not fulfill graduation requirements within the designated timeframe, thereby falling short of achieving a 100% graduation rate. The study delves into the academic trajectories of both undergraduate and postgraduate students to comprehensively analyze the factors contributing to untimely graduation. Persistence and retention are the student's and institution's view of continued enrollment; similarly, dropout and attrition are attitudes towards college leavers[4]. The presence of budgetary allocations plays a substantial role in evaluating students' opportunities for pursuing higher education[5]. The predominant proportion of graduating students fails to meet the expected timeline, necessitating a considerable demand for lecturers and a heightened requirement for facilities to accommodate the student population. This includes the provision of enhanced infrastructure, encompassing classrooms, laboratories, and scheduling arrangements for lectures. Prior to enrollment in the UAD Postgraduate Program, prospective students are required to furnish documentation pertaining to their originating university, original academic program, and undergraduate Grade Point Average

E-mail : imam.riadi@is.uad.ac.id, rusydi@mti.uad.ac.id, rio2107048006@webmail.uad.ac.id

(GPA). And graduation data for postgraduate students who have taken lectures include data on postgraduate GPA scores, TOEFL scores, and length of postgraduate study. The database is very redundant, because it is not used optimally as learning material[6][7].

The selection of variables for graduation prediction in this research was informed by previous studies conducted over the years. These studies have consistently identified certain key variables that are strong predictors of student graduation outcomes. Therefore, in selecting the variables for this research, we relied on the findings and recommendations from these previous studies. We considered variables such as the origin of education, program of origin, undergraduate GPA (S1 GPA), postgraduate GPA (S2 GPA), TOEFL scores, and duration of study (long study). These variables were chosen based on their demonstrated significance in predicting student graduation in numerous prior studies. By leveraging the insights gained from past research and the collective knowledge within the academic community, we aimed to construct a robust predictive model that incorporates the most influential factors for accurately forecasting student graduation outcomes.

Based on the description above, it can be concluded that to overcome the problems of students who have not graduated, it is necessary to conduct research with students who have graduated on time[8]. Imbalances between student enrollment and graduation can be addressed by predicting and identifying students at risk of not graduating early so that schools can develop and implement remedial policies. Properly restored and maintained [9]. Students graduating on time will be a rule that supports other students graduating on time. Student research can be carried out by applying data mining[10]. Research can also use the concept of random tree decision trees in forming rules [11]. However, the application of data mining uses a comparison method of several studies that have been carried out with the same object, namely students. This study addresses a research lacuna by investigating student graduation predictions through data mining-based predictive models, specifically employing the Naive Bayes and K-NN algorithms. While prior literature has examined various approaches to predicting student graduation, this study introduces novel contributions by broadening the spectrum of variables utilized and adopting a comprehensive modeling approach.

Earlier research has often been constrained to singular or limited variables, such as GPA or duration of study. Nonetheless, this study underscores the necessity of encompassing a wider array of variables, including university affiliation, program of study, TOEFL scores, among others, to enhance the precision and comprehensiveness of student graduation forecasts. Through an expansive approach and the incorporation of diverse variables, this study offers novelty in deciphering the determinants impacting student graduation. Consequently, the proposed model in this research harbors the potential to furnish profound and pertinent insights for stakeholders in higher education, facilitating enhanced decision-making and more efficacious intervention strategies to bolster student graduation rates.

2. RELATED RESEARCH

This research refers to previous studies which are used as basic materials in conducting research in order to produce better information than some previous studies. Not only previous studies, this section also explains the theoretical basis that supports this research, including as described below. Research with the same audience predicts graduation rates of students who have more effective assessment time in their studies, especially in Computer Engineering, University of Widyagama. Research using association method and Apriori algorithm. This method calculates the support value, which is the support value of a large rule item that accounts for 60% of the course score data. The results of the study found information about the graduation rate based on the subject and student grades[12]. This article attempts to systematically review the literature on Naive Bayes data mining methods in predicting college graduates on time. It was found that naive Bayesian data mining can make predictions about on-time graduation considering the properties of the university database used. As for the accuracy level of the three documents, it produces more than 90% accuracy even when using several different data mining properties and applications[13]. In a study focusing on various Indonesian batik objects, encompassing a wide array of types and patterns, both the Naive Bayes and Random Forest methods were employed for classification. The Random Forest method achieved the highest accuracy, reaching 97.91%, whereas the Naive Bayes method yielded an accuracy rate of 96.66% when applied to the same dataset. The findings from this research indicate that the Random Forest method excelled in classifying different types of batik motifs compared to the Naive Bayes method[14]. The properties found in the document that can determine the prediction are GPA (Cumulative Performance Index) attributes. In this study, image processing methods, along with segmentation and feature extraction techniques, were employed on images of pistachio samples. A sophisticated classifier, built upon the K-NN method, known for its simplicity and effectiveness, was applied to the dataset. Furthermore, principal component analysis was implemented for dimension reduction.This research introduced a multi-stage system encompassing feature extraction, dimension reduction, and dimension weighting. The experimental outcomes indicate that this proposed approach achieved an impressive classification accuracy of 94.18%[15]. The study uses a filtering technique based on k nearest neighbors (K-NN) (a node is connected to its k nearest neighbors) with automatic parameter evaluation, unified for all classifiers[16]. Selecting input from this system is sample data in the 2014-2015 school year student data format. Two tests, namely test data and training data, were used in the tests in this study. The criteria used in this study are GPA semester 14, credit performance, and graduation status. The output of this system is in the form of graduation predictions which are divided into two areas, namely on time and not on time. The test results show the best performance in predicting graduation with the k-Nearest Neighbor method using a quarter performance index with accuracy = 98.46% and accuracy = 99.53%, based on the application of k = 14 and kfold = 5[17].

Research on digital image processing using Naive Bayes and k-Nearest Neighbor (K-NN) methods. The object of the study is the grain of 3 types of teak Semarang, Blora and Sulawesi with an accuracy rate of over 70%. However, the best classification for Sulawesi teak is using Naive Bayes method, the best classification for teak is using Naive Bayes method, with an accuracy rate of 82.7%[18]. This article attempts to systematically review the literature on Naive Bayes data mining methods in predicting college graduates on time. It was found that naive Bayesian data mining can make predictions about on-time graduation considering the properties of the university database used. As for the accuracy level of the three documents, it produces more than 90% accuracy even when using several different data mining properties and applications. The properties found in the document that can determine the prediction are GPA (Cumulative Achievement Index) attributes[19]. In a different research study, the accuracy of three distance metrics—Euclidean Distance, Minkowski Distance, and Manhattan Distance—was compared within the context of the Chi-Square-based K-Means Clustering Algorithm for assessing the similarity between objects. The aim was to identify the most effective multivariate method for schools such as SMP Atmaja Wacana, SMKN 3 Tegal, SMAS Muhammadiyah, SMAS Pancasakti Tegal, SMKS Muhammadiyah 1 Tegal City, and SMP IC Bias Assalam. The findings from this investigation demonstrated high accuracy levels, specifically 84.47% for the Euclidean distance method, 83.85% for both the Manhattan distance method and the Minkowski method[20].

The study's objective is to utilize the K-NN algorithm with both Euclidean and Manhattan distance metrics to assess graduation accuracy. This algorithm was executed through the Rapidminer software, and it was tested on a dataset comprising 380 training data points and 163 test data points. The findings reveal that employing the Euclidean and Manhattan distance methods for classifying graduating students yielded the highest accuracy rate of 85.28% when the value of k was set to 7[21]. Research Application of Data Mining in Classifying Student Data Based on Academic Data Before College and Study Period Using the K-Medoids Method with the variables average math scores[22]. This research explores English scores, computer grades, school origins, and district origins within various engineering study programs. Notably, in the electrical engineering, industrial engineering, and informatics engineering datasets, specific data points (9, 57, and 64) with math scores exceeding 82 were identified. Conversely, the chemical engineering dataset comprised 35 data points with an average math score of 73.89. The research achieved substantial accuracy, as evidenced by Silhouette Coefficient values 0.52 for electrical engineering, 0.67 for industrial engineering, 0.35 for chemistry, and 0.65 for informatics. Furthermore, the study involved a comparative analysis using the same method K-NN, employing sengon wood as the research material. It encompassed 135 training data points and 10 testing data points, experimenting with varying k values (1, 2, 3, 4, and 5). Impressively, this research achieved a 70% accuracy rate with a corresponding 30% error rate. The research utilized PHP programming tools in conjunction with a MySQL database and the Laravel framework. Variables such as diameter, width, height, and wood sawing results constituted the research object[23].

Based on the presentation of studies related to research using student objects. This research is also to predict the graduation of postgraduate students in the UAD MTI program using classification techniques, especially the Naive Bayes and K-NN methods. Based on the insights of related studies, which highlight the efficacy of data mining in accurately classifying large data sets, the goal is to leverage this technique to achieve greater accuracy in predicting student graduation outcomes. By comparing the performance of the Naive Bayes and K-NN methods, the research seeks to identify the most effective classification approach for predicting postgraduate graduation in the UAD MTI program. Through this comparative analysis, this research aims to contribute to the development of robust predictive models that can help improve the accuracy of graduation predictions and inform decision-making processes in academic contexts.

3. RESEARCH FLOW AND BASIC CONCEPTS

System design in the study "Comparative Analysis of Naive Bayes and K-NN Approaches to Predict On-Time Graduation using Academic History" can be seen in Figure 1 flow as follows:

Explanation of the above system design namely:

A. *Data Loading*

Data loading is a stage in data processing where data is retrieved or loaded from a predefined dataset according to identified needs for prediction. This process involves accessing available datasets and extracting the necessary information or values for analysis, processing, or other uses. The data loading process can involve various steps such as reading from a database, identifying relevant columns or attributes, and formatting the data in a suitable format for further analysis. The primary objective of data loading is to make the required data available and ready for use in subsequent applications or data analysis.

B. *Data Cleaning*

Data cleaning is a crucial step in the data preprocessing phase, where data is meticulously reviewed and then either removed or retained based on predefined criteria[24]. The primary goal of data cleaning is to eliminate irrelevant or noisy data. This process includes tasks such as removing duplicate records and handling missing values[25]. Overall,

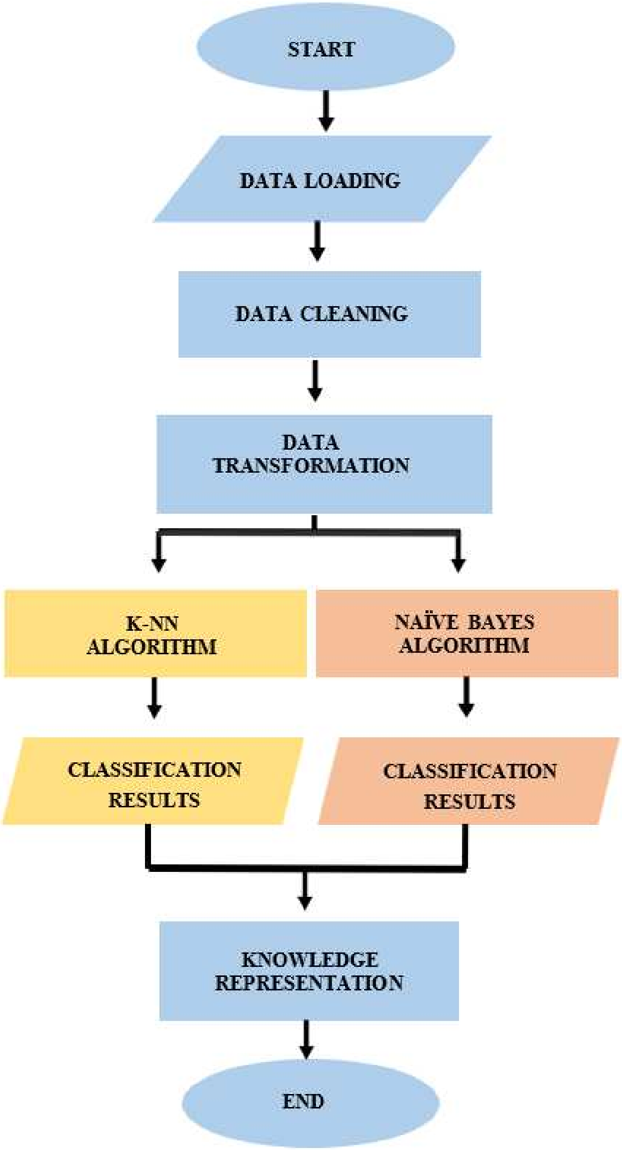


Figure 1. Research System Stages

data cleaning ensures that the dataset is accurate, consistent, and suitable for predictive analysis. This is highly important in data-driven analyses to ensure meaningful and reliable outcomes.

C. Data Transformation

Data transformation is a technique used to convert data into a different format that is assumed to be more suitable for statistical analysis or categorization based on specific categories. In this research, a data transformation technique known as the ”One Hot Encoding Technique” is employed in the K-NN algorithm. The implementation of the ”One Hot Encoding Technique” is a method for converting categorical data or data with non-numeric values into a format that is more appropriate for analysis, especially by the K-NN algorithm[26].This technique converts each category into a separate binary variable, where each category represents a new variable that only has a value of 0 or 1, indicating whether or not that category is present in the data.

In the Naive Bayes algorithm, a classification technique is utilized as the data transformation method. This involves categorizing the data into labels that are appropriate for processing using the Naive Bayes algorithm. Through the application of existing data transformation techniques, data originally presented in non-numerical formats and varied textual representations are converted into a numeric format corresponding to labels for further analysis. This process enables more effective and accurate analysis in predicting graduation outcomes.

D. Naive Bayes and K-NN Algorithm

In this stage, prediction algorithms are applied to the prediction system. This research employs two algorithms Naive Bayes Classifier and K-NN. Naive Bayes is a simple, probability-based prediction technique. It applies Bayes' theorem with a strong assumption of independence, often referred to as ”naive”[27][28]. The primary assumption of Naive Bayes classifiers is that attributes in the data are independent of each other. This algorithm applies Bayes' theorem to categorize unknown items by calculating probabilities[29]. In the context of this research, Naive Bayes is used to predict graduation based on graduate student datasets.

On the other hand, K-NN is a simple, non-parametric supervised learning algorithm. It relies on the idea that similar objects tend to be close to each other[30][31]. K-NN works by identifying the k nearest neighbors of an object and then using information from these neighbors to classify the object. In this research, K-NN is used to predict whether a student will graduate on time or not based on a predefined value of k. By applying both of these algorithms, this graduation prediction research aims to develop a prediction system that provides information about student graduation. This information can then be used to improve student management and provide relevant recommendations.

E. Classification Results

The clustering results in this context come from the previous stage which involves applying the Naive Bayes and K-NN algorithms to the data. These results categorize or label data based on the level of similarity between individual data points. In the context of the Naive Bayes algorithm, clustering refers to how student data is categorized based on the graduation probability produced by this algorithm. Students with the same probability of passing fall into the same category. On the other hand, in the context of K-NN, clustering is related to how student data is categorized based on the graduation prediction categories provided by the K-NN algorithm. Students whose graduation predictions are similar or in the same category are classified together.

The classification results obtained from this research can be useful for further analysis or decision-making processes, especially in student affairs management or in developing strategies to increase graduation rates.

F. Knowledge Representation

This final stage is the conclusion phase based on the results of testing the system on students who have been classified according to their respective classifications. Test results are obtained from the calculation of the Confusion Matrix table, such as accuracy, precision, and recall values[9]. Accuracy serves as an indicator of how effectively a classification system can accurately differentiate data from the entire dataset subjected to testing. In the context of this research, it serves to assess the overall success rate of the prediction system. A higher accuracy value reflects superior system performance. Conversely, Precision quantifies the degree to which the positive outcomes provided by the prediction system are accurate. In simpler terms, it evaluates how precisely the graduation prediction system categorizes positive data. Precision aids in determining the precision level when the system generates positive predictions. Recall quantifies the prediction system's ability to locate all positive instances that ought to be identified. Within this research, it evaluates the achievement rate of the prediction system in recognizing all positive cases and contributes to assessing the system's capacity to "recall" all existing positive cases.

In summary, accuracy offers an overall assessment of the prediction system's performance, while precision and recall contribute to comprehending the accuracy level and the system's competence in correctly identifying positive cases. Typically, a trade-off exists between precision and recall enhancing one may decrease the other due to the system's decision-making process in classifying data. These metrics are simplified by four main values. Firstly, True Positive (TP) is the case where students are predicted to graduate on time (Positive) and indeed graduate on time (True). Secondly, True Negative (TN) represents cases where students are predicted not to graduate on time (Negative) and indeed do not graduate on time (True). Thirdly, False Positive (FP) occurs when students are predicted to graduate on time (Positive) but do not graduate on time in reality (False). Lastly, False Negative (FN) encompasses situations where students are predicted not to graduate on time (Negative) but actually graduate on time (False).

To calculate its accuracy, you can use the following formula and the TABLE I.

TABLE I. Confusion Matrix

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

Using the Confusion Matrix table it can be seen the value of accuracy, precision and recall. Accuracy is a value that displays knowledge of how accurately the system performs data classification using the data. To obtain the recall value, you can use equation 1.

$$Acuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \tag{1}$$

Precision is a metric that measures the accuracy of positive predictions made by a classification model. It is calculated by dividing the number of actual positive predictions by the total number of events classified as positive. To obtain precision values, another important metric in classification evaluation. You can use the following equation: 2

$$Precision = \frac{TP}{TP + FP} \times 100\% \tag{2}$$

Recall is a value that represents data with a positive value that is correctly classified by the system. To obtain the recall value, you can use equation 3

$$Recall = \frac{TP}{TP + FN} \times 100\% \tag{3}$$

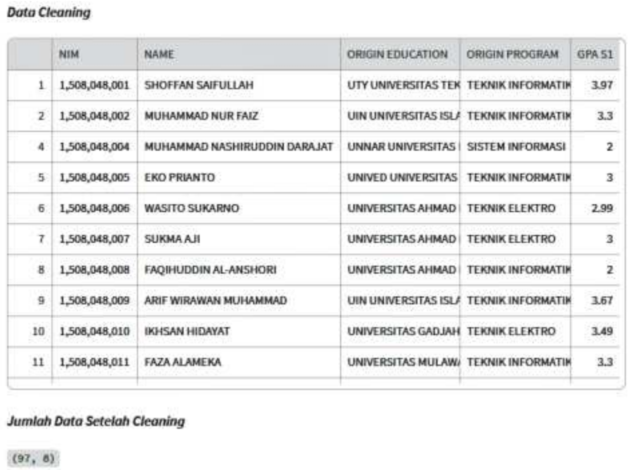
4. Result and Discussion

The discussion on the classification research on the prediction of graduate student graduation on time-based on undergraduate academic history uses the Naive Bayes and K-NN methods to determine the way with the highest accuracy value and predictive results of the two methods. Predictive results and accuracy values are obtained after processing the prediction system.

A. Data Collection

This research data was obtained through formal requests for information from the Informatics Postgraduate study program. The data submitted for research purposes comprises university data, the origin of the study program, undergraduate GPA, master's GPA, TOEFL scores, and the length of study[32]. In comparative analysis, this research aims to identify the relative influence of each variable on predicting student graduation. For instance, GPA variables at the undergraduate and graduate levels can offer insights into a student's academic performance, while TOEFL scores can indicate their proficiency in the English language, which may be a crucial factor in successfully completing their studies. Additionally, the university of origin and study program can provide insights into the academic environment and curriculum that affect a student's academic progress.

This data serves as training data for predicting postgraduate graduation by classifying graduates as either on time or not on time. The classification process employed in this research utilizes the Naive Bayes and K-NN algorithms.



Data Cleaning

	NIM	NAME	ORIGIN EDUCATION	ORIGIN PROGRAM	GPA S1
1	1,508,048,001	SHOFFAN SAIFULLAH	UTY UNIVERSITAS TEK	TEKNIK INFORMATIK	3.97
2	1,508,048,002	MUHAMMAD NUR FAIZ	UIN UNIVERSITAS ISLA	TEKNIK INFORMATIK	3.3
4	1,508,048,004	MUHAMMAD NASHIRUDDIN DARAJAT	UNNAR UNIVERSITAS	SISTEM INFORMASI	2
5	1,508,048,005	EKO PRIANTO	UNIVED UNIVERSITAS	TEKNIK INFORMATIK	3
6	1,508,048,006	WASITO SUKARNO	UNIVERSITAS AHMAD	TEKNIK ELEKTRO	2.99
7	1,508,048,007	SUKMA AJI	UNIVERSITAS AHMAD	TEKNIK ELEKTRO	3
8	1,508,048,008	FAQIHUDDIN AL-ANSHORI	UNIVERSITAS AHMAD	TEKNIK INFORMATIK	2
9	1,508,048,009	ARIF WIRAWAN MUHAMMAD	UIN UNIVERSITAS ISLA	TEKNIK INFORMATIK	3.67
10	1,508,048,010	IKHSAN HIDAYAT	UNIVERSITAS GADJAH	TEKNIK ELEKTRO	3.49
11	1,508,048,011	FAZA ALAMEKA	UNIVERSITAS MULAWA	TEKNIK INFORMATIK	3.3

Jumlah Data Setelah Cleaning

```
(97, 8)
```

Figure 2. Research Data Cleaning

B. Data Processing

1) Load Data

The research data was collected in one excel file ”Dataset” of 229 postgraduate student data. This data is the initial data before entering the prediction system for the cleaning and transformation process, as shown in TABLE III.

Dataset processing in the prediction system involves the initial stage, namely uploading the dataset into the prediction system. After the dataset is uploaded in Excel format, the prediction system will automatically start the dataset cleaning process.

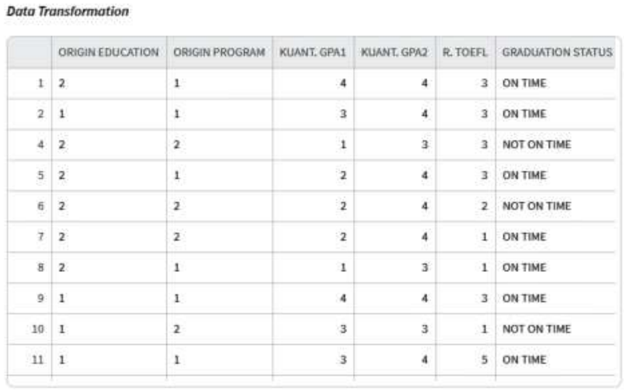
2) Data Cleaning

Data cleaning is a process because student data on one of the attributes used for calculations has an empty value. This process is carried out to get a high accuracy value in both methods. The process of cleaning research data carried out by the system with the results as shown in Figure 2.

3) Data Transformation

The transformation process requires the use of the Naive Bayes classification method for certain classifications. This classification includes several attributes, including university of origin, the study program of origin, GPA during undergraduate (S1), GPA during postgraduate (S2), TOEFL, and graduation status. The university of origin is categorized into two categories state college and private college. Origin Study Program category into two linear study program categories and non-linear study program. The linear study program category is Informatics Engineering and Computer Informatics, while the Non-Linear Study Program category is a study program other than the linear study program category. Category GPA S1 and GPA S2 classified 4, less, sufficient, good, and excellent can be seen in TABLE III.

The TOEFL attribute classification is organized into 11 different ranges, with each range determining certain score limits, as described in TABLE IV.



Data Transformation

	ORIGIN EDUCATION	ORIGIN PROGRAM	KUANT. GPA1	KUANT. GPA2	R. TOEFL	GRADUATION STATUS
1	2	1	4	4	3	ON TIME
2	1	1	3	4	3	ON TIME
4	2	2	1	3	3	NOT ON TIME
5	2	1	2	4	3	ON TIME
6	2	2	2	4	2	NOT ON TIME
7	2	2	2	4	1	ON TIME
8	2	1	1	3	1	ON TIME
9	1	1	4	4	3	ON TIME
10	1	2	3	3	1	NOT ON TIME
11	1	1	3	4	5	ON TIME

Figure 3. Results of The Data Transformation System

In contrast, the Graduation Status attribute classification is divided into two primary graduation categories ”on time” signifies graduation within two years. While ”not on time” denotes graduation occurring after two years, as illustrated in TABLE V. These attribute categories serve as a valuable guide during the data transformation process, making the application of Naive Bayes and K-NN algorithms easier. They ensure that the data is organized in a way that facilitates the implementation of these algorithms. The outcomes of the dataset transformation in the graduation prediction system, following the prescribed criteria described above, are visually presented in the Figure 3.

C. Naive Bayes Classification Technique

This research focuses on the graduation prediction system using the Naive Bayes method with total of 229 student datasets. After preprocessing the dataset, the system proceeds to conduct algorithmic calculations, as illustrated in the Figure 4. This stage involves applying the prepared data to the selected algorithms, namely Naive Bayes for further analysis. To assess how well this system can predict graduation, an evaluation stage is conducted using training data consisting of total of 15 student records selected from the previously mentioned dataset. The details of this stage are explained in the Figure 5. In evaluating the predictive performance of this system, the method employed is the Confusion Matrix. Data from the graduation prediction system is collected and processed. The research results indicate that the system’s accuracy rate reaches 80%, the precision value is 92%, the recall value is 85% and the f1 score is 88%. The prediction system results are as shown in the visualization in the Figure 6.

D. K-Nearest Neighbors Classification Technique

This stage explains the use of the second method in the research, namely the K-Nearest Neighbors (K-NN) method, to apply the algorithm to the dataset being used. This dataset consists of 229 student data and 15 test data taken from the dataset. The process of testing the accuracy of the prediction system using the K-NN method is slightly different from the Naive Bayes method. As the name suggests, K-NN requires

TABLE II. Research Dataset

NO	NAME	ORIGIN EDUCATION	ORIGIN PROGRAM	S1 GPA	S2 GPA	TOEFL	LONG STUDY
1	VENDI RINANTO	UAD AHMAD DAHLAN UNIVERSITY	INFORMATICS ENGINEERING	3.3	4.0	430	2 Years 3 Months 27 Days
2	MUHAMMAD NASHIRUDDIN	UNNAR NAROTAMA SURABAYA UNIVERSITY	INFORMATION SYSTEMS	2	3.26	433	3 Years 9 Months 13 Days
3	EKO PRIANTO	DEHASEN BENGKULU UNIVERSITY	INFORMATICS ENGINEERING	3	3.51	440	2 Years 8 Months 29 Days
4	WASITO SUKARNO	UAD AHMAD DAHLAN UNIVERSITY	ELECTRICAL ENGINEERING	2.99	3.69	413	3 Years 3 Months 26 Days
5	SUKMA AJI	UNIVERSITY AHMAD DAHLAN	ELECTRICAL ENGINEERING	3	3.92	400	2 Years 2 Months 29 Days
6	FAQIHUDDIN	UNIVERSITY AHMAD DAHLAN	INFORMATICS ENGINEERING	2	3.46	363	2 Years 11 Months 22 Days

TABLE III. GPA (Grade Index) S1 & S2

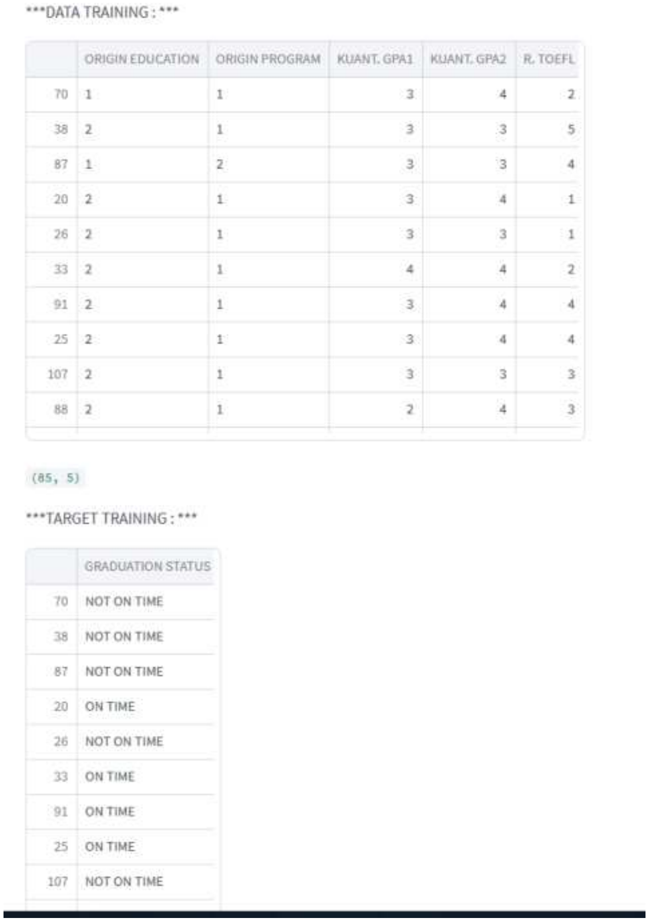
Category	Range
LESS	0.00 – 2.75
SUFFICIENT	2.76 – 3.00
GOOD	3.01 – 3.50
EXCELLENT	3.51 – 4.00

TABLE IV. TOEFL Score

Category	Range
RANGE 1	≤400
RANGE 2	401 – 420
RANGE 3	421 – 440
RANGE 4	441 – 460
RANGE 5	461 – 480
RANGE 6	481 – 500
RANGE 7	501 – 520
RANGE 8	521 – 540
RANGE 9	541 – 560
RANGE 10	561 – 580
RANGE 11	581 – 600

TABLE V. Graduation Status

Category	Information
ON TIME	≤ 2 YEARS
NOT ON TIME	> 2 YEARS



***DATA TRAINING : ***

	ORIGIN EDUCATION	ORIGIN PROGRAM	KUANT. GPA1	KUANT. GPA2	R. TOEFL
70	1	1	3	4	2
38	2	1	3	3	5
87	1	2	3	3	4
20	2	1	3	4	1
26	2	1	3	3	1
33	2	1	4	4	2
91	2	1	3	4	4
25	2	1	3	4	4
107	2	1	3	3	3
88	2	1	2	4	3

```
(85, 5)
```

***TARGET TRAINING : ***

	GRADUATION STATUS
70	NOT ON TIME
38	NOT ON TIME
87	NOT ON TIME
20	ON TIME
26	NOT ON TIME
33	ON TIME
91	ON TIME
25	ON TIME
107	NOT ON TIME

Figure 4. Naive Bayes Method Training Data

the determination of the ’k’ value. Which represents the number of neighbors considered in the prediction, as a requirement of this method. The ’k’ value significantly influences the accuracy in this research. To discover the ’k’ value that yields high accuracy, experiments are necessary. Testing is conducted by trying various ’k’ values ranging from 2 to 10, as seen in the Figure 7. The goal of these experiments is to determine the most suitable ’k’ value to optimize the accuracy of the prediction system.

From the test results, the correct ’k’ value in this study based on the highest accuracy is ’k’ = 4. The use of ’k’ = 4 in the graduation prediction system results in an accuracy rate of 73%, the precision value is 91%, the recall value is 77% and the f1 score is 83%. The prediction system results are as shown in the visualization in the Figure 8.

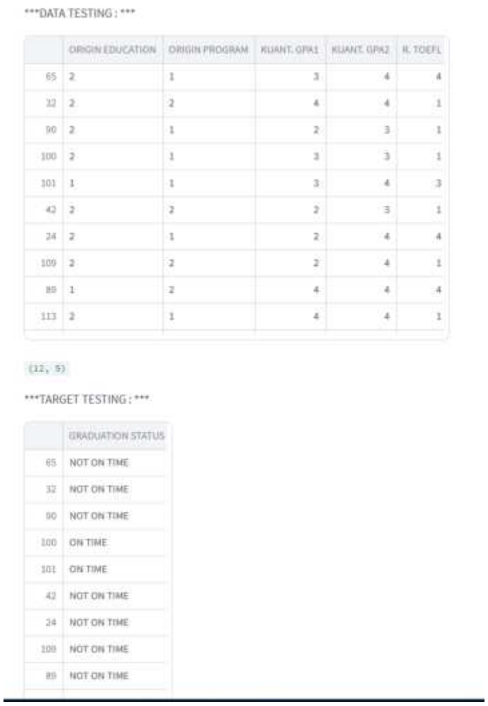
E. Data Testing

In this final stage, experiments were conducted in the research using 5 student records who had not graduated, as displayed in TABLE VI. These test data were input into the prediction system using both methods, Naive Bayes and K-Nearest Neighbor.

As seen in Figure 9, the results of testing the prediction system using the Naive Bayes method on these 5 student records indicate that one student was predicted to graduate late. While four students were predicted to graduate on time. Conversely, when the K-NN method was employed to test the same data in the graduation prediction system, Figure 10 reveals that two students were predicted to graduate late. While three students were predicted to graduate on time.

TABLE VI. Graduation Prediction System Test Data

No	Name	Origin Education	Origin Program	S1 GPA	S2 GPA	TOEFL
1	Shoffan Saifullah	UTY Universitas Teknologi Yogyakarta	Informatics Engineering	3.97	4.00	433
2	Muhammad Nur Faiz	UIN Universitas Islam Negeri Sunan Kalijaga	Informatics Engineering	3.30	3.90	440
3	Alfiansyah Imanda Putra	Darmajaya Institute of Informatics and Business	Informatics Engineering	3.08	3.42	376
4	Ikhsan Zuhriyanto	UIN Universitas Islam Negeri Sunan Kalijaga	Informatics Engineering	3.30	3.83	433
5	Muhammad Irwan Syahib	Halu Oleo University	Informatics Engineering	2.95	3.58	413



***DATA TESTING : ***

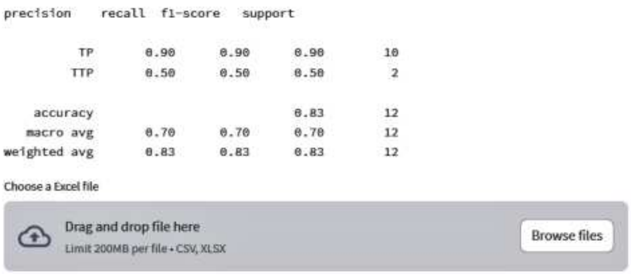
	ORIGIN EDUCATION	ORIGIN PROGRAM	KUANT. GPA1	KUANT. GPA2	R. TOEFL
65	2	1	3	4	4
32	2	2	4	4	1
90	2	1	2	3	1
100	2	1	3	3	1
101	1	1	3	4	3
42	2	2	2	3	1
24	2	1	2	4	4
109	2	2	2	4	1
89	1	2	4	4	4
113	2	1	4	4	1

```
(12, 5)
```

***TARGET TESTING : ***

	GRADUATION STATUS
65	NOT ON TIME
32	NOT ON TIME
90	NOT ON TIME
100	ON TIME
101	ON TIME
42	NOT ON TIME
24	NOT ON TIME
109	NOT ON TIME
89	NOT ON TIME

Figure 5. Data Testing Method Naive Bayes



```
precision    recall  f1-score   support

         TP       0.90      0.90      0.90        10
        TTP       0.50      0.50      0.50         2

   accuracy                           0.83        12
  macro avg       0.70      0.70      0.70        12
weighted avg      0.83      0.83      0.83        12
```

Choose a Excel file

Figure 6. Confusion Matrix Results of the Naive Bayes Method

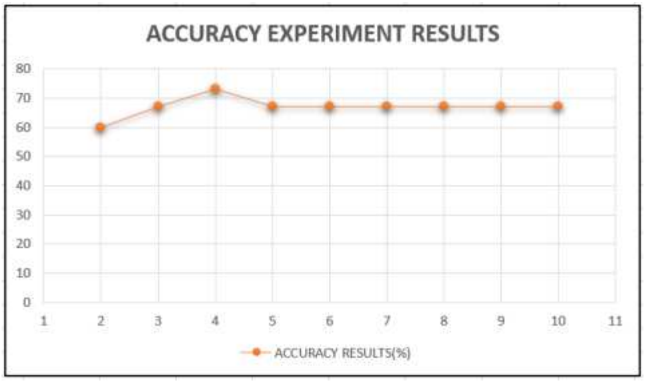
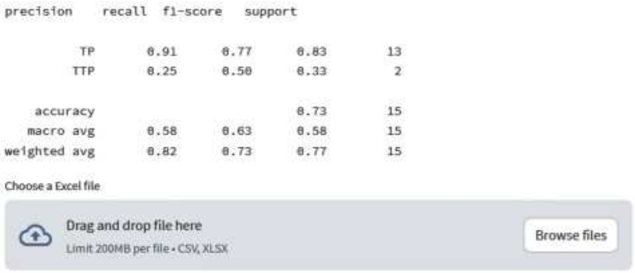


Figure 7. Accuracy Experiment Result

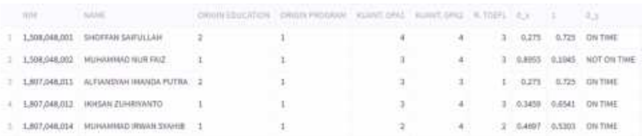


```
precision    recall  f1-score   support

         TP       0.91      0.77      0.83        13
        TTP       0.25      0.50      0.33         2

   accuracy                           0.73        15
  macro avg       0.58      0.63      0.58        15
weighted avg      0.82      0.73      0.77        15
```

Choose a Excel file

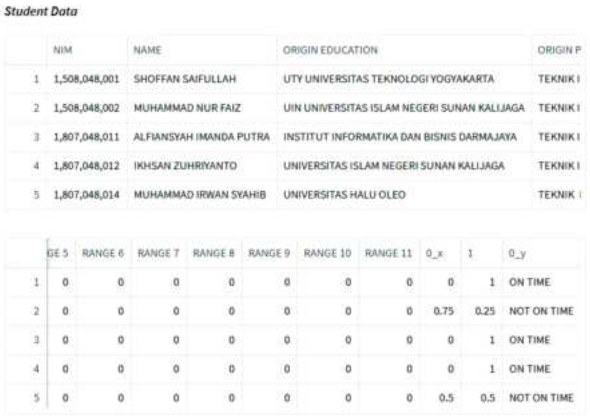
Figure 8. Confusion Matrix Results of The K-Nearest Neighbor Method

It's important to note that the accuracy of the prediction system generated by the system is influenced by the size of the dataset and the diversity of variables used in the research. A larger dataset and greater variability in variables will significantly impact the accuracy and precision of the prediction system.



	NIM	NAME	ORIGIN EDUCATION	ORIGIN PROGRAM	KUANT. GPA1	KUANT. GPA2	R. TOEFL	0_x	1	0_y
1	1,508,048,001	SHOFFAN SAIFULLAH	2	1	4	4	3	0.275	0.725	ON TIME
2	1,508,048,002	MUHAMMAD NUR FAIZ	1	1	3	4	3	0.8955	0.1045	NOT ON TIME
3	1,807,048,011	ALFIANSYAH IMANDA PUTRA	2	1	3	3	1	0.275	0.725	ON TIME
4	1,807,048,012	IKHSAN ZUHRIYANTO	1	1	3	4	3	0.3459	0.6541	ON TIME
5	1,807,048,014	MUHAMMAD IRWAN SYAHIB	1	1	2	4	2	0.4697	0.5303	ON TIME

Figure 9. Test Results of The Naive Bayes Method Prediction System



Student Data

	NIM	NAME	ORIGIN EDUCATION	ORIGIN P
1	1,508,048,001	SHOFFAN SAIFULLAH	UTY UNIVERSITAS TEKNOLOGI YOGYAKARTA	TEKNIK I
2	1,508,048,002	MUHAMMAD NUR FAIZ	UIN UNIVERSITAS ISLAM NEGERI SUNAN KALIJAGA	TEKNIK I
3	1,807,048,011	ALFIANSYAH IMANDA PUTRA	INSTITUT INFORMATIKA DAN BISNIS DARMAJAYA	TEKNIK I
4	1,807,048,012	IKHSAN ZUHRIYANTO	UNIVERSITAS ISLAM NEGERI SUNAN KALIJAGA	TEKNIK I
5	1,807,048,014	MUHAMMAD IRWAN SYAHIB	UNIVERSITAS HALU OLEO	TEKNIK I

	GE 5	RANGE 6	RANGE 7	RANGE 8	RANGE 9	RANGE 10	RANGE 11	0_x	1	0_y
1	0	0	0	0	0	0	0	0	1	ON TIME
2	0	0	0	0	0	0	0	0.75	0.25	NOT ON TIME
3	0	0	0	0	0	0	0	0	1	ON TIME
4	0	0	0	0	0	0	0	0	1	ON TIME
5	0	0	0	0	0	0	0	0.5	0.5	NOT ON TIME

Figure 10. K-Nearest Neighbor Method System Testing Results

5. Conclusions

Based on the research that has been done, it can be concluded based on a postgraduate study of informatics program students. And analysis has collected student data from batch 6 (2015) to batch 13 (2021) as many as 229 graduate student data. The research has successfully developed a graduation prediction system, which begins with the cleaning of the dataset to remove noise or incomplete data for each student and variable. During this data cleaning stage, the system addresses any inconsistencies or missing values in the dataset resulting in complete and accurate data. As a result of this cleaning process, the system obtains 95 student records that are ready to enter the next stage of research. The graduation prediction system has been successful in doing so transformation process using the one hot encoding method, and This process is done to facilitate research to perform Naıve Bayes and K-NN calculation processes. Based on the processed dataset, the accuracy result from the Test Confusion Matrix in the prediction system using the Naïve Bayes algorithm shows an accuracy value of 80% and a misclassification rate of 20%. With the accuracy results, the prediction system was tested using predictive data from 5 students who had not graduated. The results of the system show that of 5 students, one student did not graduate on time, and four students did not graduate on time, with the results as shown in Figure 9.

Meanwhile, the accuracy testing of the graduation prediction system using the K-NN algorithm yielded an accuracy rate of 73% and a misclassification rate of 27%. The plan was tested with the same five student data using the K-NN method, giving results from 5 student data 2 students were predicted to graduate not on time. Test the accuracy and predictive value of 5 student data is very different, showing that Näıve Bayes method has higher accuracy and predictions of student graduation from 5 more student data, four graduating on time and one not graduating on time. Whereas the K-NN method has a lower accuracy value, from 5 predictive data, three graduate on time and two do not graduate on time with the results as shown in Figure 10. The classification error value obtained in the study can be attributed to several explanatory factors. Naive Bayes relies on the assumption of feature independence, meaning that each feature is considered independent of others. When the features in the dataset exhibit significant independence, Naive Bayes tends to perform well. Additionally, Naive Bayes employs a simpler model compared to K-NN, making it easier to understand and interpret, especially when the data has a relatively simple and linear structure. On the other hand, K-NN is susceptible to overfitting, particularly when a small value of k is used. In contrast, Naive Bayes typically maintains a lower level of complexity, which helps mitigate overfitting, especially on small or noisy datasets. Furthermore, the relative performance of Naive Bayes and K-NN may vary depending on the dataset's size. Larger datasets may require greater computational resources for K-NN due to its dependence on distance between data points. However, it's important to note that the dataset used in this research is not large. Based on the accuracy values obtained from the research, it can be concluded that the Naive Bayes method demonstrates higher accuracy in predicting student graduation. Delving deeper into the reasons why Naive Bayes may be more effective allows researchers to gain a better understanding of dataset characteristics and algorithm suitability for specific use cases. This understanding can inform the selection of the most suitable method for data analysis, considering various approaches and other data mining techniques available.

References

[1] H. Priyatman, F. Sajid, and D. Haldivany, "Clustering Using the K-Means Clustering Algorithm to Predict Student Graduation Time," *Journal of Informatics Education and Research(JEPIN)*, vol. 5, no. 1, p. 62, 2019.

[2] M. N. McCredie and J. E. Kurtz, "Prospective prediction of academic performance in college using self- and informant-rated personality traits," *Journal of Research in Personality*, vol. 85, p. 103911, 2020. [Online]. Available: https://doi.org/10.1016/j.jrp.2019.103911

[3] M. Siddik, Hendri, R. N. Putri, Y. Desnelita, and Gustientiedina, "Classification Of Student Satisfaction On Higher Education Services Using Naïve Bayes Algorithm," *Journal of Information Technology and Computer Science (INTECOMS)*, vol. 3, no. 2, 2020.

[4] E. L. Huerta-Manzanilla, M. W. Ohland, and R. d. R. Peniche-Vera, "Co-enrollment density predicts engineering students' persistence and graduation: College networks and logistic regression analysis," *Studies in Educational Evaluation*, vol. 70, no. October 2020, 2021.

[5] G. A. Agarkov, A. A. Tarasyev, and A. D. Sushchenko, "Optimization of Students' Graduation by the University Taking into Account the Needs of the Labor Market," *IFAC-PapersOnLine*, vol. 53, no. 2, pp. 17 399–17 404, 2020. [Online]. Available: https://doi.org/10.1016/j.ifacol.2020.12.2094

[6] Q. Guo, C. Yang, and S. Tian, "Prediction of Purchase Intention Among E-commerce Platform Users Based on Big Data Analysis," *Revue d'Intelligence Artificielle*, vol. 34, no. 1, pp. 95–100, 2020.

[7] J. Huang, I. Ul Haq, C. Dai, S. Khan, S. Nazir, and M. Imtiaz, “Isolated Handwritten Pashto Character Recognition using AK-NN Classification Tool Based on Zoning and Hog Feature Extraction Techniques,” *Complexity*, vol. 2021, 2021.

[8] L. Hannaford, X. Cheng, and M. Kunes-Connell, “Predicting nursing baccalaureate program graduates using machine learning models: A quantitative research study,” *Nurse Education Today*, vol. 99, no. December 2020, p. 104784, 2021. [Online]. Available: https://doi.org/10.1016/j.nedt.2021.104784

[9] “Classification algorithm accuracy improvement for student graduation prediction using ensemble model,” *International Journal of Information and Education Technology*, vol. 10, no. 10, pp. 723–727, 2020.

[10] A. Falade, A. Azeta, A. Oni, and I. Odun-ayo, “Systematic Literature Review of Crime Prediction and Data Mining,” *Review of Computer Engineering Studies*, vol. 6, no. 3, pp. 56–63, 2019.

[11] T. Tundo and S. Saifullah, “Mamdani’s Fuzzy Inference System in Predicting Woven Fabric Production Using Rules Based on Random Trees,” *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 9, no. 3, p. 443, 2022.

[12] I. Kurnawan, F. Marisa, and D. Purnomo, “Implementation Of Data Mining Using Apriori Algorithm To Predict Student Graduation Levels,” vol. 4, no. 1, 2018.

[13] F. S. Nugraha and H. F. Pardede, “Autoencoder for Baby Birth Weight Prediction System,” *Journal of Information Technology and Computer Science*, vol. 9, no. 2, p. 235, 2022.

[14] A. Fadlil, I. Riadi, and I. J. D. P. Putra, “Comparison of Machine Learning Performance Using Naive Bayes and Random Forest Methods to Classify Batik Fabric Patterns,” *Revue d’Intelligence Artificielle*, vol. 37, no. 2, pp. 379–385, 2023.

[15] I. A. Özkan, M. Köklü, and R. Saraçoğlu, “Classification of Pistachio Species using Improved k-NN Classifier,” *Progress in Nutrition*, vol. 23, no. 2, 2021.

[16] M. Ala’raj, M. Majdalawieh, and M. F. Abbod, “Improving binary classification using filtering based on k-NN proximity graphs,” *Journal of Big Data*, vol. 7, no. 1, 2020. [Online]. Available: https://doi.org/10.1186/s40537-020-00297-7

[17] E. S. Susanto, K. Kusrini, and H. A. Fatta, “Prediction of Graduate Students in Informatics Engineering, Amikom University, Yogyakarta Using the K-Nearest Neighbor Method,” *Respati*, vol. 13, no. 2, pp. 67–72, 2018. [Online]. Available: http://jti.respati.ac.id/index.php/jurnaljti/article/view/260/239.

[18] R. R. Waliyansyah and C. Fitriyah, “Comparison of image classification accuracy of teak wood using naive bayes and k-nearest neighbor (k-nn) methods,” *Journal of Informatics Education and Research (JEPIN)*, vol. 5, p. 157, 8 2019. [Online]. Available: http://testjurnal.untan.id/index.php/jepin/article/view/32473

[19] L. Setiyani, M. Wahidin, D. Awaludin, and S. Purwani, “Analysis of Timely Student Graduation Predictions Using Data Mining Methods,” *Faktor Exacta*, vol. 13, no. 1, p. 35, 2020.

[20] M. Nishom, J. T. Informatika, P. H. Bersama, and P. H. Bersama, “Comparison of Accuracy of Euclidean Distance, Minkowski Distance, and Manhattan Distance on Chi-Square Based K-Means Clustering Algorithm,” *Jurnal Informatika: Jurnal Pengembangan IT (JPIT)*, vol. 04, no. 01, pp. 20–24, 2019.

[21] N. Hidayati and A. Hermawan, “K-Nearest Neighbor (K-NN) algorithm with Euclidean and Manhattan in classification of student graduation,” vol. 2, no. 2, pp. 86–91, 2021.

[22] T. Jaringan, H. Kurnia, L. Zahrotun, and U. Linarti, “Grouping of Students Based on Academic Data Before Lecture and Study Period Using K-Medoids,” vol. 2, 2021.

[23] S. dan Agus Jaka Sri Hartanta Anton Yudhana, “K-nn algorithm with euclidean distance for prediction of sengon wood sawn results,” *UNDIP E-JOURNAL SYSTEM*, 2020.

[24] I. Riadi, Sunardi, and P. Widiandana, “Cyberbullying Detection on Instant Messaging Services Using Rocchio and Digital Forensics Research Workshop Framework,” *Journal of Engineering Science and Technology*, vol. 17, no. 2, pp. 1408–1421, 2022.

[25] R. Kondabala, V. Kumar, and A. Ali, “A machine learning prediction model for the affinity between glucose and binder,” *Revue d’Intelligence Artificielle*, vol. 33, no. 3, pp. 227–233, 2019.

[26] R. Anggara, “Klasifikasi Kelulusan Mahasiswa Berdasarkan Riwayat Akademik Sebelum Kuliah dan Nilai PMDK-Raport Menggunakan Metode K-NN(K-Nearest Neighbors),” Ph.D. dissertation, Ahmad Dahlan, 2021.

[27] A. Amin, A. Adnan, and S. Anwar, “An adaptive learning approach for customer churn prediction in the telecommunication industry using evolutionary computation and Naïve Bayes,” *Applied Soft Computing*, vol. 137, p. 110103, 2023. [Online]. Available: https://doi.org/10.1016/j.asoc.2023.110103

[28] S. Wang, J. Ren, and R. Bai, “A semi-supervised adaptive discriminative discretization method improving discrimination power of regularized naive Bayes,” *Expert Systems with Applications*, vol. 225, no. November 2022, p. 120094, 2023. [Online]. Available: https://doi.org/10.1016/j.eswa.2023.120094

[29] X. Zhao and Z. Xia, “Secure outsourced NB: Accurate and efficient privacy-preserving Naive Bayes classification,” *Computers and Security*, vol. 124, p. 103011, 2023. [Online]. Available: https://doi.org/10.1016/j.cose.2022.103011

[30] Q. Wang, S. Wang, B. Wei, W. Chen, and Y. Zhang, “Weighted K-NN Classification Method of Bearings Fault Diagnosis with Multi-Dimensional Sensitive Features,” *IEEE Access*, vol. 9, pp. 45 428–45 440, 2021.

[31] C. Gong, Z.-g. Su, X. Zhang, and Y. You, “Adaptive evidential K -NN classification : Integrating neighborhood search and feature weighting,” *Information Sciences*, vol. 648, no. August, p. 119620, 2023. [Online]. Available: https://doi.org/10.1016/j.ins.2023.119620

[32] A. M. Al-Swaidani and T. Al-Hajeh, “Estimation of GPA at Undergraduate Level using MLR and ANN at Arab International University during the Syrian Crisis: A Case Study,” *Open Education Studies*, vol. 5, no. 1, 2023.

Imam Riadi Holds the academic title of Professor in the field of Information System. He earned his Doctorate degree from Gadjah Mada University in 2014. He holds a Master's degree in Computer Science from Gadjah Mada University in 2004 and a Bachelor's degree in Electrical Engineering Education from Yogyakarta State University (UNY) in 2001. He has been a permanent lecturer at Universitas Ahmad Dahlan (UAD) since 2002. His courses are Information Security, Computer Networking and Digital Forensics.

Rusydi Umar He has the academic rank of Doctor in Computer Science. He is also an expert in computer engineering, informatics engineering, cloud & grid computing. Obtained Doctor of Computer Science degree from Hyderabad Central University India in 2014. Obtained Master of Informatics Engineering from Bandung Institute of Technology in 2003 and Bachelor of Electrical Engineering Study Program from Gadjah Mada University in 1998. Has been a permanent lecturer at Ahmad Dahlan University (UAD) since 2002 until now.

Rio Anggara He is a Master of Informatics student of the University of Ahmad Dahlan (UAD). He focuses on analytical research, data mining, machine learning and data science.