International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys. 16, No.1 (Sep-24)

http://dx.doi.org/10.12785/ijcds/160185

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, Yogyakarta, Indonesia*
[2]*Departement of Informatics, Universitas Ahmad Dahlan, Yogyakarta, Indonesia*
[3]*Master Program of Informatics, Universitas Ahmad Dahlan, Yogyakarta, Indonesia*

Received 9 Sep. 2023, Revised 21 May 2024, Accepted 25 May 2024, Published 1 Sep. 2024

Abstract: Graduation is a pivotal moment in higher education, significantly impacting institutional accreditation and public perception. This study aims to ensure that all students graduate punctually, recognizing the critical role of higher education in achieving this goal. Central to this effort are comprehensive datasets that capture academic performance throughout both undergraduate and graduate studies. These datasets include details such as university and program specifics, undergraduate and master's GPAs, TOEFL scores, and the duration of study. By leveraging classification techniques within data mining, particularly K-NN and Naive Bayes, a comparative analysis was conducted to precisely predict the on-time completion of graduate students. The process of predicting graduation involves several stages, including data preprocessing, transformation, and the segmentation of data into training and testing sets. Subsequently, the selected methods are applied, and analyses are undertaken to accurately forecast graduation outcomes. Experimental findings reveal an 80% accuracy rate for Naive Bayes and 73% for K-NN. Notably, Naive Bayes demonstrates superior efficacy in predicting on-time graduation. However, to further refine accuracy, it is necessary to expand datasets and diversify the variables used in the analysis, such as incorporating additional academic and non-academic factors that may influence graduation timelines. The insights derived from this research hold significant implications for academic institutions, offering valuable guidance for implementing proactive measures to support students in completing their studies within the expected timeframe. By utilizing the findings of this study, educational institutions can develop tailored strategies and interventions to address potential barriers to timely graduation, such as enhancing academic advising, providing targeted support services, and optimizing course scheduling. These efforts will ultimately foster student success, improve institutional outcomes, and contribute to the overall excellence of higher education institutions.

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

1. Introduction

Continuing education beyond undergraduate studies is a common practice, but many students face challenges such as irregular attendance and sporadic participation, which hinder the timely completion of their academic programs[1]. These disparities can hinder students' progress towards completing their final projects. The aim of this study is to bridge skill gaps and facilitate timely graduation, which is eagerly anticipated by individuals in academia. Post-graduation GPAs serve as benchmarks for assessing students' academic aptitude, which is crucial for incoming students, parents, and peers[2]. The quality of student service is essential for the sustainability of educational institutions, with students being the primary clientele[3].

Ahmad Dahlan University (UAD), located in Yogyakarta, is renowned within the Muhammadiyah organization. Timely graduation of UAD students is crucial for maintaining accreditation and meeting academic requirements. This research focuses on UAD students who did not graduate within the designated timeframe, aiming to analyze factors contributing to untimely graduation. Persistence and retention are essential for both students and institutions, while dropout and attrition reflect attitudes towards college leavers. Budget allocations play a significant role in students' opportunities for higher education[4]. Despite the prevalent issue of students failing to graduate on time, there is a substantial demand for lecturers and facilities to accommodate the student population[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

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

undergraduate Grade Point Average (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, considering factors such as educational background, GPA, TOEFL scores, and study duration. Leveraging insights from past research, the goal was to construct a robust predictive model.

To address the issue of students failing to graduate, it is essential to conduct research with students who graduate on time[8]. By predicting and identifying students at risk of not graduating early, schools can implement remedial policies. This study fills research gaps by investigating graduation predictions using data mining-based predictive models, specifically Naive Bayes and K-NN algorithms[9][10][11]. The study introduces novelty by expanding the range of variables and adopting a comprehensive modeling approach.

Previous research often focused on singular variables such as GPA or study duration. However, this study emphasizes the need for a broader array of variables to enhance the accuracy of graduation forecasts. Through an inclusive approach and diverse variables, this study provides insights for stakeholders, facilitating improved decision-making and intervention strategies to enhance 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. Earlier studies aimed to determine the mean GPA at the Arab International University, particularly amidst the persistent conflict in Syria. Data from six faculties were gathered from the university's registration system, encompassing variables like high school GPA, source of student certificates, and student gender (GR). Utilizing Multiple Linear Regression (MLR) and Artificial Neural Network (ANN) methods, it was revealed that male students notably impact GPA, highlighting their significance as a pivotal factor in academic outcomes [12]. Another study focused on secondary students' academic data in mathematics, evaluating accuracy, precision, recall, and f-1 score. XGBoost yielded the highest prediction scores, with an accuracy of 90.89% [13]. In the article predicting student timelines for students in Indonesia using data mining. Analyzing data from 132,734 students, including demographic and academic records, the research identifies GPA as the most crucial factor in predicting graduation timelines. The predictive model, employing random forest and neural network algorithms, achieves the highest accuracy compared to other models, with a classification accuracy of 76% and an AUC of 79%.Moreover, the study identifies GPA as the most influential variable in predicting graduation timelines, alongside six other significant factors [14]. The study employs alternative machine learning techniques, namely Artificial Neural Networks (ANN) and Naïve Bayes (NB), to distinguish prosperous first-year medical students by utilizing sociodemographic data and academic records. Compared to NB, the ANN model exhibits marginally superior accuracy, sensitivity, and specificity. Both models excel in sensitivity for typical students and specificity for atypical ones. Within the top 25 predictive factors identified by NB, the accuracy percentage on diagnostic exams emerges as the most impactful [15]. Article that tries 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 [16]. 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 [17]. The properties found in the document that can determine the prediction are GPA (Cumulative Performance Index) attributes. In another 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% [18].The research uses a filtering technique based on k nearest neighbors (K-NN) (a node is connected to its k nearest neighbors) with automatic, integrated parameter evaluation for all classifiers [19]. The study conducts a comparative assessment to gauge the effectiveness of the K-Nearest Neighbor (K-NN) algorithm and Support Vector Machine (SVM) as classification techniques for predicting the duration of studies among students enrolled in the Bachelor of Laws program at the Law Faculty. The findings indicate that the K-NN algorithm achieves its highest prediction accuracy, reaching 98.45%, when K=5. Conversely, the SVM model yields an accuracy of 96.90%. Hence, the research outcomes are deemed satisfactory, demonstrating high precision. Notably, the K-NN modeling approach exhibits superior accuracy, particularly with K=5, surpassing the performance of the SVM method [20].

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% [21]. 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[22]. 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 [23]. In another research investigation, the accuracy of three distance metrics—Euclidean Distance, Minkowski Distance, and Manhattan Distance—was evaluated in the context of the Chi-Square-based K-Means Clustering Algorithm to determine the similarity among objects. The goal is to identify the most effective multivariate methods. 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 [24].

In research on the use of 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[25]. 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[26]. 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[27]. The same research with student objects uses historical data before college. The results of research using K-NN clustering have the highest value of k=3 neighbors. The prediction accuracy test results reached 78%[28].

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. Through comparing the effectiveness of Naive Bayes and K-NN methods, this study aims to determine the optimal classification approach for forecasting postgraduate graduation in the UAD MTI program. By conducting this comparative analysis, the research endeavors to enhance the development of reliable predictive models, which can consequently enhance the precision of graduation forecasts and provide valuable insights for decision-making in academic settings.

3. Research Flow and Basic Concepts

Before discussing the design of the prediction system, the tools and materials used to facilitate this research to run successfully and obtain maximum results can be seen in TABLE I and TABLE II.

TABLE I. Research Tools

No	Device	Operating System	Purpose
1	Laptop Aspire ES-14 (model ES1-432-C6AH)	Windows 10	Visualizing Research
2	Mouse	-	Operating the Laptop
3	23-inch Monitor	-	Managing Research

The system design depicted in the ”Comparative Analysis of Naive Bayes and K-NN Approaches to Predict On-Time Graduation using Academic History” study is illustrated 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.

TABLE II. Research Support Tools

No	Software	Version	Purpose
1	Windows 10	Windows 10	Supporting Operating System
2	Python 3	3.11.3	Workstation Operating System
3	Visual Studio Code	4.7	Text Editor
4	Microsoft Excel	2019	Number Processing
5	Streamlit	1.25.0	Research Framework
6	Pandas	1.2.2	Libraries and Analysis
7	Sklearn	0.24.1	Libraries and Analysis
8	Nearest Neighbours	1.6.2	Libraries and Analysis
9	GussianNB	-	Libraries and Analysis
10	Numpy	1.20.0	Analysis Tools
11	Openpyxl	3.0.4	Hashing Tool
12	Browser Mozilla or Chrome	115.0.5790.171	Research Render

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. The undergraduate and postgraduate study history database used in this research was obtained from the system owned by the campus, through the administration of the informatics study program. Next, the first stage of this database is carried out, namely the loading process.

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[22]. 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[29]. Overall, 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[30].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.

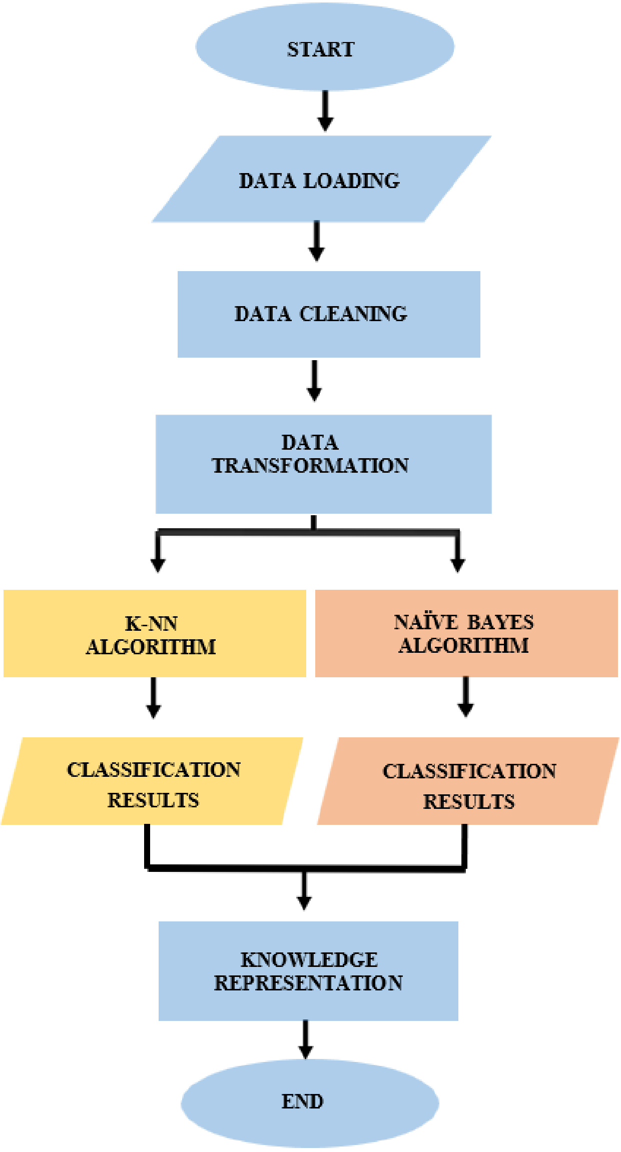


Figure 1. Research System Stages

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 phase, prediction algorithms are implemented into the prediction system. Two algorithms, namely the Naive Bayes Classifier and K-NN, are utilized in this research. The Naive Bayes method is a straightforward, probability-based prediction technique that operates on the assumption of independence, often referred to as ”naive” [31][32]. Its fundamental assumption is that attributes in the data are independent of each other. This algorithm applies Bayes' theorem to classify unknown items by calculating probabilities [33]. In the context of this study, Naive Bayes is employed to forecast graduation based on datasets of graduated students.

On the contrary, K-NN is a basic, non-parametric supervised learning algorithm. It relies on the idea that similar objects tend to be close to each other[34][35]. 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 III.

TABLE III. 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.

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

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

In the context of a classification study on predicting graduate school graduation over time based on graduate history, Naive Bayes and K-NN methods are used to determine which method and which outcome predict a true value of both methods. After processing the prediction system, we get the prediction results and correct values.

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[12]. 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.

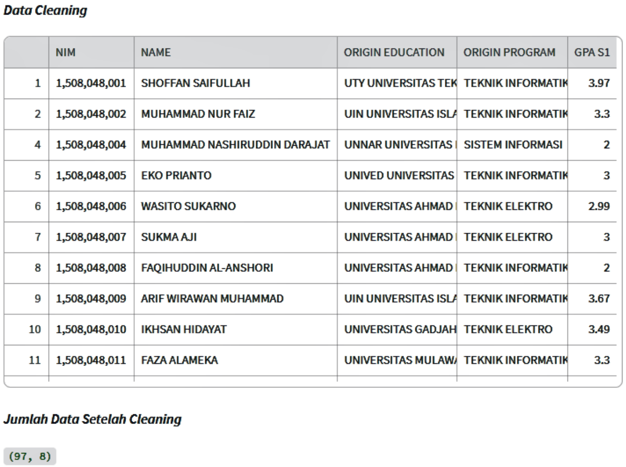
This data serves as training data for predicting postgraduate graduation by classifying graduates as either on time or not on time. The classification process employed in this research utilizes the Naive Bayes and K-NN algorithms.

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 V.

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.



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. Results of Research Data Cleaning with the Python Prediction System

2) Data Cleaning

Data cleansing is the first step in developing a prediction system using the Python programming language. This process involves removing student data from any attributes utilized in the study that contain missing values. Previously we discussed the concept of a prediction system built based on research. The cleaning process is carried out to obtain high accuracy values for both methods. The process of cleaning research data from a total of 229 datasets was carried out by the prediction system with results as 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 V.

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

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 VII. These attribute categories serve as a valuable guide during the data transformation process, making

TABLE IV. 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

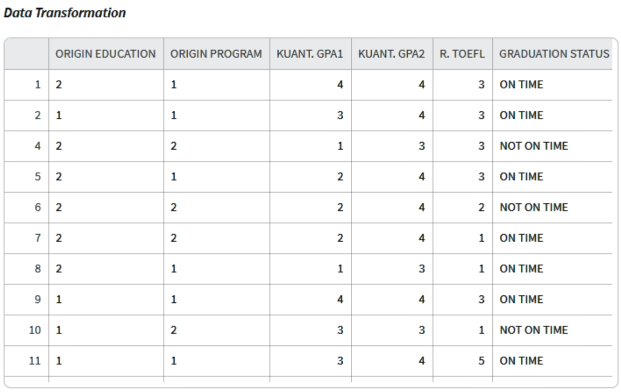
TABLE V. GPA (Grade Index) S1 & S2

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

TABLE VI. 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

the implementation of Naive Bayes and K-NN algorithms easier. They ensure that the data is organized in a way that facilitates the application of these algorithms. The results of the dataset transformation are processed into a graduation prediction system, following the criteria specified above, presented visually in Figure 3.



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. Data Transformation Results by Graduation Prediction System

TABLE VII. Graduation Status

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

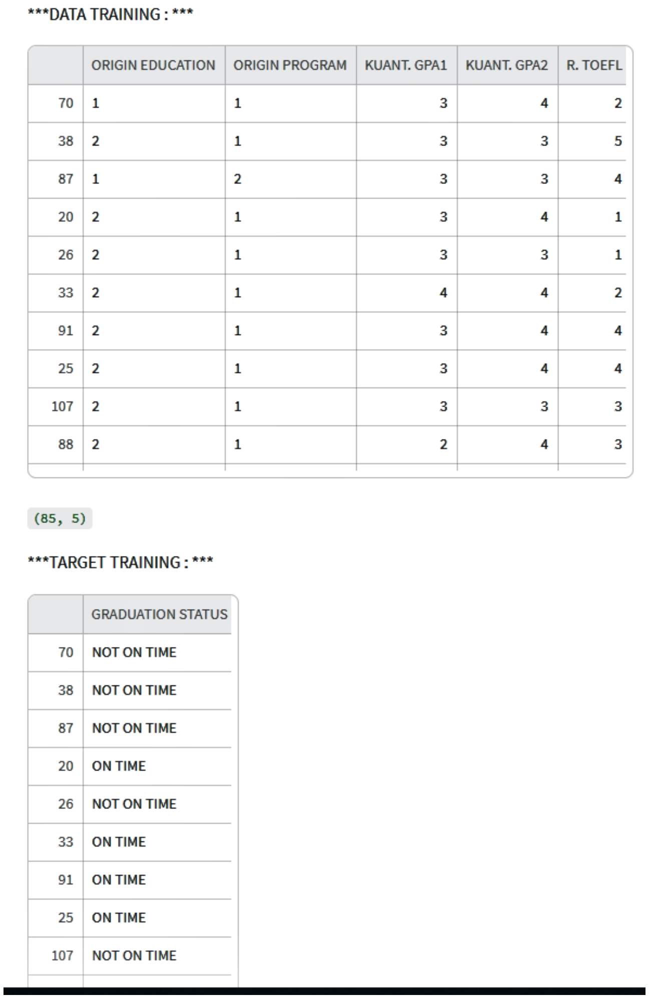
C. Naive Bayes Classification Technique

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 phase elaborates on the implementation of the second technique in the study, which is the K-NN method, to deploy the algorithm to the dataset under examination. 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 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



***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

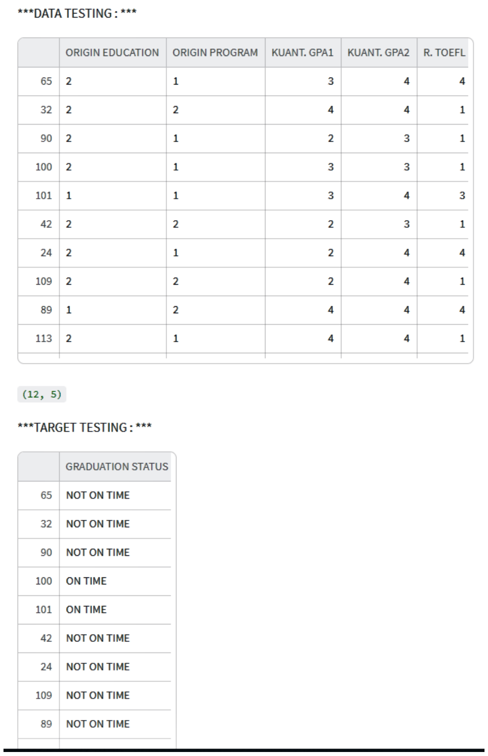
Figure 4. Results of Naive Bayes Method Training Data Selection by Graduation Prediction System

77% and the f1 score is 83%. The prediction system results are as shown in the visualization in the Figure 8.

E. Data Testing

Successful research using a database of students who have graduated, found that the naive Bayes dominance method has a higher value than the K-NN method. starting from the values of accuracy, precision, recall, and f1 score. The classification results of the two methods are visualized on a graph as in the Figure 9

The experiment in this research involved testing the predictive ability of the Naive Bayes and K-Nearest Neighbor (K-NN) methods on a dataset consisting of five students who had not yet graduated, as shown in TABLE VIII. The student data is entered into a graduation prediction system, and both methods are applied to assess whether the student is likely to graduate on time or not. Assessments are carried out based on new exam data for those students who have not yet graduated. By utilizing a database of



***DATA TESTING : ***

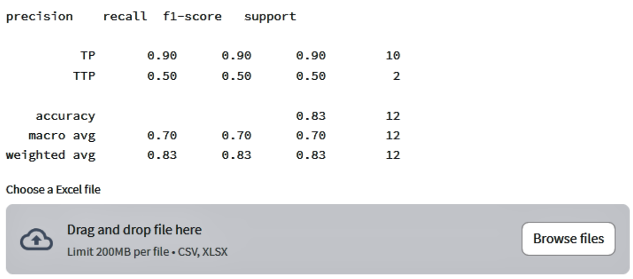
	ORIGIN EDUCATION	ORIGIN PROGRAM	KUANT. GPA1	KUANT. GPA2	R. TOEFL
65	2	1	3	4	4
32	2	2	4	4	1
90	2	1	2	3	1
100	2	1	3	3	1
101	1	1	3	4	3
42	2	2	2	3	1
24	2	1	2	4	4
109	2	2	2	4	1
89	1	2	4	4	4
113	2	1	4	4	1

```
(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. GUI Display of Naive Bayes Method Test Data in Graduation Prediction Systems



```
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

Drag and drop file here

Limit 200MB per file • CSV, XLSX

Figure 6. Confusion Matrix Results of the Naive Bayes Method

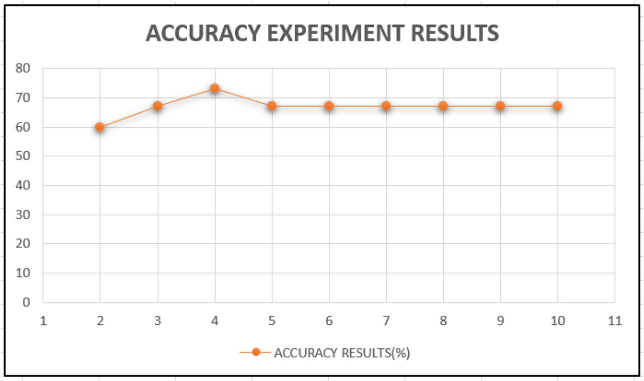
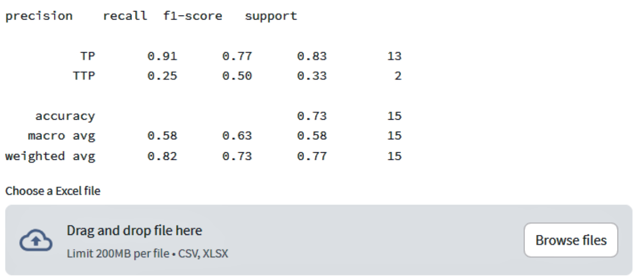


Figure 7. Accuracy Experiment Result



```
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

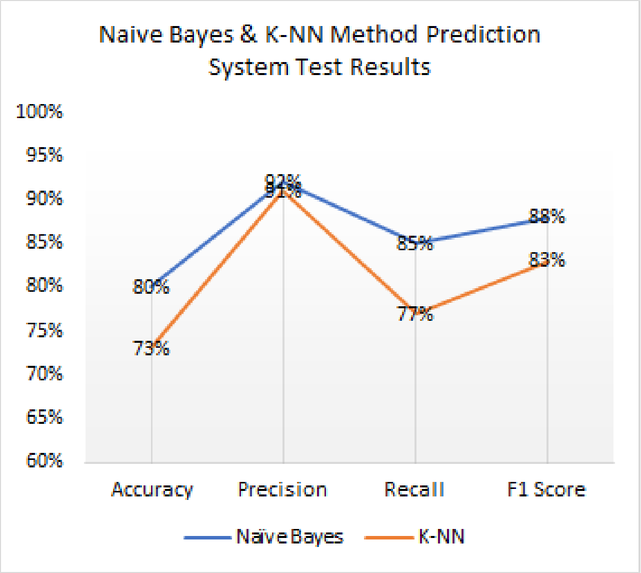
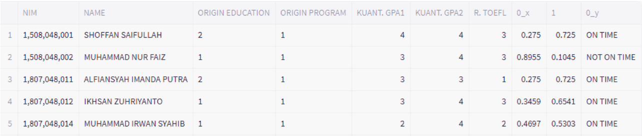
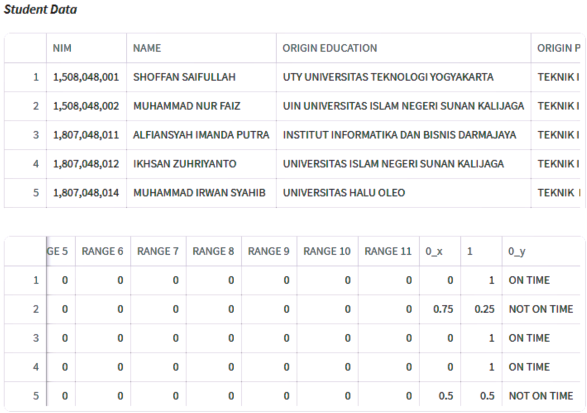


Figure 9. Naive Bayes and K-NN Method Prediction System Test Results



	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 10. GUI Display of the Pass Prediction System Test Results of the Naive Bayes Method



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 11. GUI Display of the Pass Prediction System for K-NN Method Test Results

students who have graduated, this research aims to evaluate the performance of the prediction system in determining the graduation schedule accurately. Next, the Naive Bayes and K-NN methods are used to calculate the probability that the five students will graduate on time. This process allows comparison of the prediction results produced by both methods, providing insight into the method's effectiveness in predicting graduation schedules for students who have not yet completed their studies. Overall, this experimental procedure allows assessing the performance of the prediction system in accurately determining whether students will graduate on time or not, based on their academic data and the prediction models generated by the Naive Bayes and K-NN methods.

As seen in Figure 10, the outcomes of evaluating the prediction system using the Naive Bayes method on these 5 student records reveal that one student was projected to graduate later than expected. 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 11 reveals that two students were predicted to graduate late. While three students were predicted to graduate on time.

It's crucial to acknowledge that the accuracy of the prediction system produced by the model is affected by both the size of the dataset and the diversity of variables utilized in the study. A larger dataset and more diverse variables will have a substantial impact on the accuracy and precision of the prediction system.

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. Research has succeeded in developing a graduation prediction system using the Python programming language which begins with cleaning 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

TABLE VIII. Graduation Prediction System Test Data

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

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 10.

Meanwhile, testing the accuracy of the graduation prediction system using the K-NN algorithm produced an accuracy rate of 73% and a classification error rate of 27%. The test used 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. The results of the accuracy test and prediction value show that the Näıve Bayes method has higher accuracy and prediction from 5 student data, four graduated on time and one did not graduate on time. Meanwhile, the K-NN method has lower accuracy, from 5 predictive data, three people passed on time and two people did not pass on time as seen in Figure **??**. The misclassification value in research is caused by several explanatory factors. Naive Bayes relies on the assumption of feature independence, meaning that each feature is considered independent of each other. When the features in a dataset show significant independence, Naive Bayes tends to perform well. In addition, Naive Bayes uses a simpler model than K-NN, making it easier to understand and interpret, especially when the data has a relatively simple and linear structure. The implications and importance of this research are twofold. First, these findings explain the potential limitations and advantages of the K-NN and Naive Bayes algorithms in predicting student graduation schedules. Although K-NN may be prone to overfitting, especially smaller values of k. Naive Bayes offers a simple model that is more resistant to overfitting, especially with smaller or noisy datasets. Additionally, the relative performance of these algorithms can vary based on the size of the data set, and larger data sets can potentially require more computational resources for K-NN due to its distance-based approach. Second, this research underscores the importance of selecting appropriate algorithms for predictive modeling tasks, such as predicting student graduation schedules. The higher accuracy demonstrated by the Naive Bayes method highlights its effectiveness in specific contexts. Understanding the reasons behind Naive Bayes' superior performance provides valuable insights into the characteristics of the data set and the suitability of the algorithm. Another finding is that campuses can implement proactive steps to support students in completing their studies within the expected time period. By leveraging these research findings, educational institutions can develop tailored strategies and interventions to address potential barriers to on-time graduation, thereby increasing student success and institutional excellence. Future research is expected to have more attributes and databases used. This understanding allows researchers to make informed decisions when choosing the most appropriate methods for research, taking into account the various available data mining approaches and techniques.

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.

[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.

[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.

[9] A. C. Lagman, L. P. Alfonso, M. L. I. Goh, J. A. P. Lalata, J. P. H. Magcuyao, and H. N. Vicente, "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," *Journal of Information Technology and Computer Science*, vol. 9, no. 3, p. 443, 2022.

[12] 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.

[13] S. A. A. Kharis, "Prediction of Student Graduation in Mathematics Subjects using Educational Data Mining," *Journal of School Mathematics Learning Research*, vol. 7, no. 1, pp. 21–29, 2023.

[14] A. Santoso, H. Retnawati, E. Apino, and M. N. Rosyada, "Predicting Time to Graduation of Open University Students : An Educational Data Mining Study," no. 2019, 2024.

[15] D. Monteverde Suarez, P. Gonzalez Flores, R. Santos Solorzano, M. Garcia Minjares, I. Zavala Sierra, V. L. de la Luz, and M. Sanchez Mendiola, "Predicting Students' Academic Progress and Related Attributes in First-Year medical students: An Analysis With Artificial Neural Networks and Naïve Bayes," *BMC Medical Education*, vol. 24, no. 1, pp. 7–8, 2024.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] E. Novianto and S. Suhirman, "Comparison of K-Nearest Neighbor Classification Methods and Support Vector Machine in Predicting Students' Study Period." *Journal of Education and Science*, vol. 33, no. 1, pp. 32–45, 2024.

[21] 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.

[22] 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.

[23] 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.

[24] M. Nishom, "Comparison of Accuracy of Euclidean Distance, Minkowski Distance, and Manhattan Distance on Chi-Square Based K-Means Clustering Algorithm," *Journal of Informatics: Journal of IT Development (JPIT)*, vol. 04, no. 01, pp. 20–24, 2019.

[25] 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.

[26] 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.

[27] S. H. Agus Jaka, A. Yudhana, and Sunardi, "K-nn algorithm with euclidean distance for prediction of sengon wood sawn results," *UNDIP E-JOURNAL SYSTEM*, 2020.

[28] I. Riadi, R. Umar, and R. Anggara, "Predicted Timely Graduation Based On Academic History Before College With K-Nearest Neighbor Method," *Journal of Information Technology and Computer Science (JTIIK)*, vol. 11, no. 2, pp. 249–256, 2024.

[29] 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.

[30] R. Anggara, "Classification of Student Graduation Based on Academic History Before College and PMDK-Report Scores using the K-NN (K-Nearest Neighbors) Method," Ph.D. dissertation, Ahmad Dahlan University, 2021.

[31] 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.

[32] S. Wang, J. Ren, and R. Bai, "A Semi-Supervised Aaptive Discriminative Discretization Method Improving Discrimination Power of Regularized Naive Bayes," *Expert Systems with Applications*, vol. 225, no. November 2022, p. 120094, 2023.

[33] X. Zhao and Z. Xia, "Secure Outsourced NB: Accurate and Efficient Privacy-Preserving Naive Bayes Classification," *Computers and Security*, vol. 124, p. 103011, 2023.

[34] 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.

[35] 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.

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.