Terbit *online* pada laman web jurnal : https://jes-tm.org/index.php/jestm/index

Journal of Engineering Science and Technology Management

| ISSN (Online) 2828 -7886 |

Article

Implementation of the Naïve Bayes Algorithm for Predicting the On-Time Graduation of Informatics Engineering Students

Erlinda[1,a], Dwipa Junika Putra[2,b,*],Imam Andhika [3]
[1]Universitas Islam Kuantan Singingi, Kuantan Singingi
[2,3]Universitas Syiah Kuala, Banda Aceh

DOI: 10.31004/jestm.v6i2.445

E-mail: [a]erlinda120015@gmail.com, [b,*]dwipajunikaputra@usk.ac.id (Corresponding author), [b]imamandhika14@usk.ac.id

ARTICLE INFORMATION

Volume 6 Issue 2
Received: 30 May 2026
Accepted: 26 June2026
Publish *Online*: 29 June 2026
Online: at https:JESTM.org

Keywords

Timely Graduation
Naïve Bayes Algorithm
Data Mining
Knowledge Discovery in Database (KDD)
Student Academic Performance

A B S T R A C T

Timely graduation is one of the important indicators used to measure the quality of academic performance in higher education institutions. However, the Informatics Engineering Study Program at the Faculty of Engineering, Universitas Islam Kuantan Singingi, still faces challenges related to students who are unable to complete their studies on time. This study aims to predict students timely graduation using the Naïve Bayes algorithm with the Knowledge Discovery in Database (KDD) approach. The research process consists of several stages, including data selection, data preprocessing, transformation, data mining, and evaluation. The data used in this study were obtained from students who completed their studies in 2025 and included several academic attributes such as gender, study duration, numerical grades, letter grades, and graduation status. The Naïve Bayes algorithm was applied to classify and predict whether students would graduate on time based on the probability of previous academic data. The results show that students with good academic performance tend to have a higher probability of graduating on time. Model evaluation using cross-validation produced the best performance in the 3-fold scenario, achieving an accuracy of 88.73%, precision of 64.62%, recall of 45.72%, and a kappa value of 0.451. These findings indicate that the Naïve Bayes algorithm is sufficiently effective for predicting students timely graduation.

1. Introduction

Timely graduation is one of the important indicators used to evaluate the quality and effectiveness of higher education institutions. A high graduation rate within the expected study period reflects the success of academic processes, student guidance systems, and institutional management. In addition, graduation performance is often considered in accreditation assessments and institutional quality evaluations. Therefore, universities are encouraged to continuously monitor and improve factors that influence students' ability to complete their studies on time (Julkarnain & Yustiardin, 2024).

The Informatics Engineering Study Program, Faculty of Engineering, Universitas Islam Kuantan Singingi, has experienced a steady increase in student enrollment over recent years. Along with this growth, the study program faces challenges in ensuring that students complete their studies within the prescribed academic period. Delayed graduation remains an issue that may affect both student outcomes and institutional performance (Fitriani & Wibowo, 2023). Various factors may contribute to this condition, including academic achievement, learning progress, course completion rates, and other academic-related indicators recorded during the study period.

Currently, the evaluation of students' graduation potential is generally conducted through manual observation of academic records. Such an approach requires considerable time and effort, particularly when dealing with a growing number of students. Moreover, manual assessment may not always provide consistent and objective results. As a consequence, there is a need for a data-driven approach that can assist academic administrators in identifying students who may require additional academic support before graduation delays occur.

The increasing availability of academic data stored in university information systems provides an opportunity to utilize analytical techniques for educational decision-making. Through the analysis of historical student data, patterns related to graduation outcomes can be identified and used to support predictive models. Educational data mining has been widely applied to predict academic performance, dropout risk, and graduation success, enabling institutions to make more informed decisions based on empirical evidence (Satrio Junaidi et al., 2024).

Among various classification techniques, the Naïve Bayes algorithm has gained attention due to its simplicity, computational efficiency, and ability to perform well in educational data mining applications. Several previous studies have demonstrated the effectiveness of Naïve Bayes in predicting student performance, academic achievement, and graduation outcomes. However, most of these studies were conducted using datasets from different institutions, predictor variables, and educational environments. Furthermore, limited research has specifically investigated graduation timeliness prediction among students of the Informatics Engineering Study Program at Universitas Islam Kuantan Singingi using local academic data and a structured Knowledge Discovery in Databases (KDD) approach. Therefore, this study addresses this gap by applying and evaluating the Naïve Bayes algorithm to predict students' graduation timeliness based on institutional academic data. The contribution of this study lies in providing empirical evidence regarding the applicability of Naïve Bayes for graduation prediction in the context of Universitas Islam Kuantan Singingi and supporting the development of academic monitoring and early identification systems for students at risk of delayed graduation.

Based on these considerations, this study applies the Naïve Bayes classification algorithm within the Knowledge Discovery in Databases (KDD) framework to predict student graduation timeliness. The results of this research are expected to assist the study program in identifying students who are at risk of delayed graduation, support academic decision-making processes, and contribute to efforts aimed at improving educational quality and graduation performance.

2. Literature Review

Educational Data Mining (EDM) has become an important research area in higher education, aiming to extract meaningful knowledge from educational data to support academic decision-making and improve student outcomes. One of the major applications of EDM is the prediction of student performance and graduation outcomes, enabling institutions to identify students who may require academic support at an early stage.

Several recent studies have explored the use of machine learning techniques for predicting student graduation and academic success. Naïve Bayes has been widely

applied due to its simplicity, low computational cost, and effectiveness in handling classification tasks. Previous studies reported that Naïve Bayes can provide competitive performance in predicting academic achievement, graduation status, and student success when compared with other classification algorithms. In addition, machine learning approaches such as Decision Tree, Random Forest, Logistic Regression, and Support Vector Machine have also been utilized to develop academic early-warning systems and graduation prediction models.

Recent studies in educational data mining emphasize the importance of model evaluation using metrics beyond accuracy, particularly when dealing with small and imbalanced datasets. Precision, Recall, F1-Score, and Cohen's Kappa are commonly recommended to provide a more comprehensive assessment of classification performance. Furthermore, the use of cross-validation techniques has become a standard practice to improve the reliability and generalizability of predictive models.

Although previous studies have demonstrated the potential of Naïve Bayes for student performance prediction, limited research has specifically investigated graduation timeliness prediction among students of the Informatics Engineering Study Program at Universitas Islam Kuantan Singingi using institutional academic data. Therefore, this study contributes by evaluating the applicability of the Naïve Bayes algorithm within the KDD framework for predicting students' graduation timeliness and supporting academic monitoring activities.

2.1. Data Mining

Data mining is the process of discovering patterns within existing data in an understandable way so that it can produce meaningful information (Rama et al., 2023)(Nangi & Rinaldi Hadistio, 2025)(Meiriza et al., 2020). In this study, the data mining technique used is the Naïve Bayes algorithm, which is a simple probabilistic classification method. The basic concept of Bayes' theorem is to solve predictive problems by creating classifications to distinguish objects into certain categories (Julkarnain & Yustiardin, 2024). Data mining is also known as Knowledge Discovery in Database (KDD), which refers to activities that include data collection, data cleaning, data integration, data mining, and the presentation of information from large datasets (Rama et al., 2023)(Nangi & Rinaldi Hadistio, 2025).

2.2. Definition of Classification

Classification is one of the methods used in data mining. It is performed by predicting an unknown class based on existing data. Classification can be described as a method used to determine whether a data object belongs to a particular category that has been previously defined (Gigih Putra Kawani, 2019)(Fitriani & Wibowo, 2023).

2.3. Naïve Bayes Algorithm

The Naive Bayes Classifier is a classification algorithm based on Bayes' Theorem and utilizes the concept of probability in the decision-making process. This algorithm calculates the probability of a data item belonging to a particular class based on information obtained from previous data. This theorem, introduced by Thomas Bayes, allows predictions of events based on existing experience or historical data, making it widely applied in various fields of data mining and machine learning (Ihsan A. Abu Amra, 2017)(Felicia Watratan et al., 2020).

In Bayes' theorem, probability can be expressed as follows (Alfa Saleh, 2015)(Etriyanti et al., 2020):

$$P\,(H|X) = \frac{P\,(X|H)P\,(H)}{P(X)}$$

2.4. Knowledge Discovery in Database (KDD)

Knowledge Discovery in Database (KDD) is defined as a method and process of obtaining information from available databases. The KDD process in data mining consists of the following stages (Mujib Ridwan, 2013)(Bagus et al., 2017):

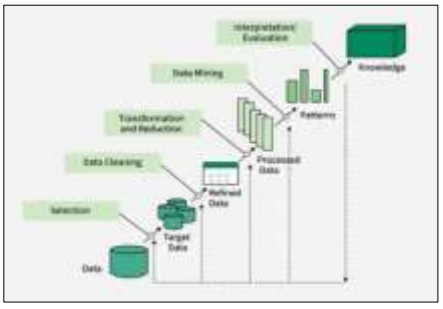


Figure 1. Knowledge Discovery in Database

1. Data Selection
   Data selection is the process of retrieving data from a dataset that will later be processed.

2. Data Preprocessing
   Data preprocessing is the initial stage of data processing. At this stage, the data are prepared to eliminate noise and inconsistent data.
3. Transformation
   Transformation is the stage of converting data into a format suitable for the model or algorithm used in the data processing stage.
4. Data Mining
   This stage involves the process of searching and extracting knowledge to obtain a model that can provide useful and valuable information.
5. Evaluation

Evaluation is performed to represent the results of the obtained model and to test its accuracy and suitability with the related data.

3. Research Methodology

3.1. Research Object

The object of this study is the prediction of timely graduation of students in the Informatics Engineering Study Program, Faculty of Engineering, Universitas Islam Kuantan Singingi. The classification was conducted on students who graduated in 2025.

3.2. Research Methodology

The research methodology used in this study is Knowledge Discovery in Database (KDD). The stages of the KDD methodology include data selection, data preprocessing, transformation, data mining, and evaluation (Felicia Watratan et al., 2020). The following is a picture of the research flow:

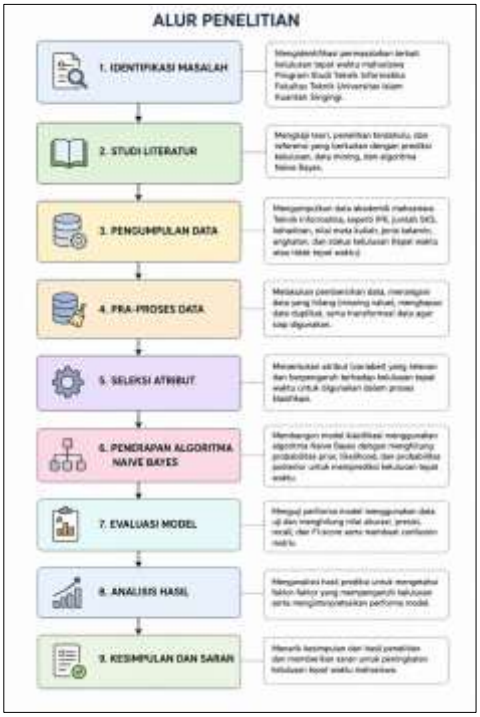


Figure 2. Research Flow

3.3. Research Design

To solve the problem related to the implementation of the Naïve Bayes algorithm, the following Knowledge Discovery in Database (KDD) steps were applied (Gigih Putra Kawani, 2019)(Alghifari & Juardi, 2021):

1. Data Collection
   At the data collection stage, the researcher collected academic data from students of the Informatics Engineering Study Program, Faculty of Engineering, Universitas Islam Kuantan Singingi. The data were obtained from the academic department or academic information system related to students study processes. The collected data included several attributes such as gender, year of enrollment, Grade Point Average (GPA), total credit units, course grades, attendance level, and students graduation status, whether timely or not timely. These data served as the basis for the student graduation prediction process.
2. Data Selection
   The data selection stage is the process of selecting relevant data to be used in the study. The academic data that had been collected were then selected based on attributes considered to influence students timely graduation. The selected attributes were further used in the preprocessing stage so that the

data processing could become more effective and aligned with the research objectives.

3. Data Preprocessing
   Data preprocessing aims to produce high-quality data suitable for analysis and classification processes. At this stage, data cleaning was carried out by removing incomplete, duplicate, and inconsistent data. In addition, data containing errors were checked to ensure that the data used in the research process became more accurate and valid.
4. Transformation
   Transformation is the stage of converting data into a form suitable for the data mining process. In this study, a discretization technique was applied, which converts numerical attributes into categorical attributes. For example, GPA values can be grouped into high, medium, and fair categories. Data transformation was performed to optimize the classification process using the Naïve Bayes algorithm.
5. Data Mining
   At the data mining stage, the processed data were used for student graduation classification using the Naïve Bayes algorithm. This method was applied to predict whether students would graduate on time or not based on previous academic data. The Naïve Bayes algorithm works by reading the prepared training data, calculating the number of classes, determining the probability of each attribute for a particular class, and multiplying all attribute probabilities to obtain the final probability value. The classification result is determined based on the highest probability value, allowing the prediction of students graduation status. The following is a picture of the Naïve Bayes Algorithm Flowchart below :

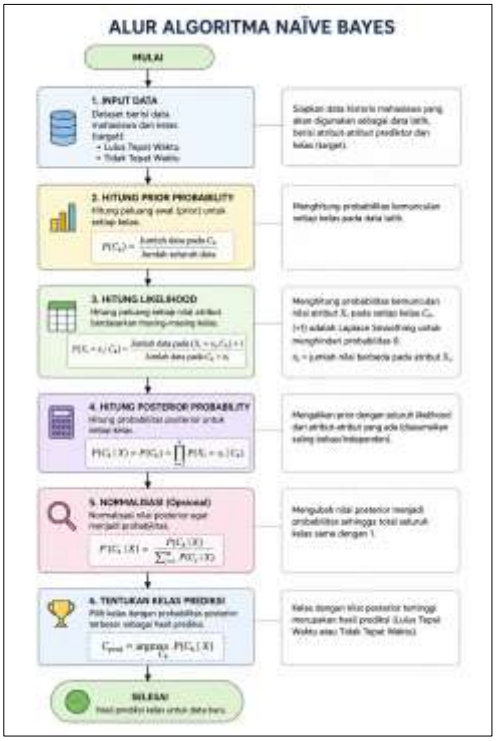


Figure 3. Naïve Bayes Algorithm Flowchart

6. Evaluation
   At this stage, evaluation was conducted to measure the validity of the classification model using cross-validation. The research results were evaluated through accuracy calculations obtained from the confusion matrix and Kappa value (Anggraini et al., 2020).

4. Results and Discussion

4.1 4.1. Data Collection

The data collection stage was conducted to obtain academic records of students from the Informatics Engineering Study Program, Faculty of Engineering, Universitas Islam Kuantan Singingi. The data were collected from the academic administration office and the university academic information system.

The dataset consists of student information including student identification number, enrollment year, Grade Point Average (GPA), thesis examination grade, academic achievement category, and graduation status. Graduation status was categorized into two classes : On Time and Not On Time. These attributes were selected because they are considered relevant factors that may influence students' graduation outcomes.

A total of 53 graduate records from the 2025 graduation period were collected and used as the initial

dataset for the classification process using the Naïve Bayes algorithm.

Table 1. presents the data of students who completed their thesis examination in 2025 :

No	Student ID	G	Enr. Year	GPA	Thesis Grade	Grad. Status
1	1802100xx	M	2018	3.14	A	Not On Time
2	2002110xx	M	2020	3.17	A	Not On Time
3	2002110xx	M	2020	3.28	A	Not On Time
4	1802100xx	M	2018	3.10	A	Not On Time
5	2102100xx	M	2021	3.42	A	On Time
6	2102100xx	M	2021	3.49	A	On Time
7	2102100xx	M	2021	3.61	A	On Time
8	2102100xx	M	2021	3.72	A	On Time
9	2102100xx	F	2021	3.55	A	On Time
10	2102100xx	M	2021	3.34	A	On Time
11	2102100xx	M	2021	3.38	A	On Time
12	2102100xx	M	2021	3.64	A	On Time
13	2102100xx	F	2021	3.66	A	On Time
14	2102100xx	F	2021	3.36	A	On Time
15	2002100xx	M	2020	3.19	A	Not On Time
16	2002100xx	M	2020	3.31	A	Not On Time
17	2102100xx	M	2021	3.02	A	On Time
18	2102100xx	M	2021	3.37	A	On Time
19	2102100xx	F	2021	3.38	A	On Time
20	2102100xx	M	2021	Null	Null	Null
21	2102100xx	F	2021	3.49	A	On Time
22	2102100xx	F	2021	3.68	A	On Time
23	2102100xx	F	2021	3.54	A	On Time
24	2102100xx	F	2021	2.95	B	On Time
25	2102100xx	F	2021	3.24	A	On Time
26	2102100xx	F	2021	3.28	A	On Time
27	2102100xx	F	2021	3.60	A	On Time
28	2102100xx	F	2021	3.63	A	On Time
29	2102100xx	M	2021	3.52	A	On Time
30	2002100xx	M	2020	2.78	B	Not On Time
31	2002100xx	M	2020	3.13	A	Not On Time
32	2002100xx	M	2020	3.01	A	Not On Time
33	2002100xx	M	2020	2.91	B	Not On Time
34	2102100xx	M	2021	3.73	A	On Time
35	2102100xx	F	2021	3.04	A	On Time
36	2002100xx	F	2020	3.56	A	Not On Time
37	2102100xx	M	2021	3.32	A	On Time
38	2102100xx	F	2021	3.44	A	On Time
39	2102100xx	F	2021	3.82	A	On Time
40	2102100xx	M	2021	3.29	A	On Time
41	2102100xx	F	2021	3.37	A	On Time
42	2102100xx	M	2021	3.60	A	On Time
43	2102100xx	M	2021	3.36	A	On Time
44	2102100xx	F	2021	3.19	A	On Time
45	2002100xx	M	2020	3.00	A	Not On Time
46	1902100xx	M	2019	3.05	A	Not On Time
47	2102100xx	M	2021	3.68	A	On Time
48	2102100xx	M	2021	3.28	A	On Time
49	2102100xx	F	2021	3.27	A	On Time
50	2102100xx	M	2021	3.23	A	On Time
51	2102100xx	F	2021	3.25	A	On Time
52	2102100xx	F	2021	3.12	A	On Time
53	2102100xx	M	2021	3.47	A	On Time

After the students academic data were obtained, the next stage was to describe the data in order to gain a deeper understanding of the attributes that would be used in the classification process for predicting the timely graduation of students in the Informatics Engineering Study Program, Faculty of Engineering, Universitas Islam Kuantan Singingi, using the Naïve Bayes algorithm. The attribute description was carried out so that the function and role of each data attribute used in the data processing stage could be clearly identified and to facilitate the analysis process in the subsequent stages. The following is a description of the data attributes used in this study.

Table 2. Data Attribute Descriptions :

No	Attribute	Description
1	G	Student gender
2	GPA	GPA Students
3	Grad. Status	On Time or Not On Time

4.2. Data Selection

The students academic data that had been obtained were then processed in the data selection stage. The purpose of the data selection stage was to choose attributes considered to have an influence on the classification process for predicting students timely graduation. Attribute selection was carried out to ensure that the data used were more relevant and in accordance with the research objectives, thereby making the data processing stage more effective. The following table presents the selected data attributes.

Table 3. Selected Data Attributes

Attribute	Role
GPA	Predictor
Grad. Status	Class Label

The object of this research is the prediction of timely graduation of students in the Informatics Engineering Study Program, Faculty of Engineering, Universitas Islam Kuantan Singingi. The prediction was conducted based on the academic data of students who had completed their studies. The classification process was performed to determine the categories of students who graduated on time and those who did not using the Naïve Bayes algorithm.

4.3. Data Preprocessing

In the selected students academic data, several incomplete, inconsistent, and incorrectly written data entries were found, making them irrelevant for use in the research process. Invalid data can affect classification results and reduce the prediction accuracy of the Naïve Bayes algorithm. Therefore, a data cleaning process was carried out to correct or remove problematic data so that the data used would become more accurate and of higher quality. The following table presents the irrelevant data.

Table 4. Irrelevant Data

Student ID	Problem
2102100xx	Missing GPA Category, Thesis Grade, Grad. Status

The removal of missing values was conducted to prevent interference with the data mining process, ensuring that the data used became more valid and ready for the classification process. To ensure data quality, the incomplete record was removed from the dataset. After the cleaning process, 52 valid student records remained and were used in subsequent stages of analysis. The following table presents a sample of the cleaned dataset.

Table 5. Sample of cleaned dataset

No	G	GPA	Grad. Status
1	M	3.14	Not On Time
2	M	3.17	Not On Time
3	F	3.66	On Time
4	F	3.36	On Time
5	M	3.19	Not On Time
6	M	3.31	Not On Time
7	F	3.63	On Time
8	M	3.52	On Time
9	M	2.78	Not On Time
10	M	2.91	Not On Time

The results of the data cleaning process indicate that data containing empty values had been corrected or removed, allowing the dataset to be ready for the transformation and classification stages using the Naïve Bayes algorithm.

Students' academic data obtained from the previous stage were still represented in several numerical attributes, particularly the Grade Point Average (GPA). To facilitate the classification process using the Naïve Bayes algorithm, the numerical GPA values were transformed into categorical data. This initialization process was performed to simplify probability calculations and improve the effectiveness of the classification model.

The data initialization process was conducted using the following criteria :

1. The GPA categorization was based on the institutional academic regulations regarding graduation predicates. Students with GPA ≥ 3.51 were categorized as High, GPA 3.01–3.50 as Medium, and GPA < 3.00 as Fair. These thresholds reflect the academic performance classifications commonly applied in Indonesian higher education institutions
2. Students with a GPA between 3.01 and 3.50 were categorized as Medium academic achievement.
3. Students with a GPA below 3.00 were categorized as Fair academic achievement.
4. Students who completed their studies within the standard study period were categorized as On Time graduates.
5. Students who exceeded the standard study period were categorized as Not On Time graduates.

The following is the student dataset that has undergone the data initialization process. The following table presents a sample of the initialized dataset after the categorization process.

Table 6. Sample of the initialized dataset

No	G	GPA	Grad. Status
1	M	Medium	Not On Time
2	M	Medium	Not On Time
3	F	High	On Time
4	F	Medium	On Time
5	M	Medium	Not On Time
6	M	Medium	Not On Time
7	F	High	On Time
8	M	High	On Time
9	M	Fair	Not On Time
10	M	Fair	Not On Time

4.4. Transformation

Transformation is the process of converting data into a format suitable for the data mining stage. At this stage, numerical and textual data were transformed into categorical forms to facilitate the classification process using the Naïve Bayes algorithm. The transformation technique applied in this study was discretization, which involves grouping continuous numerical values into predefined categories.

In this study, the Grade Point Average (GPA) attribute was transformed into three categories : High, Medium, and Fair. Students with a GPA of 3.51 or higher were categorized as High, students with a GPA between 3.01 and 3.50 were categorized as Medium, and students with a GPA below 3.00 were categorized as Fair. In addition, the gender attribute was represented as Male (M) and Female (F), while thesis examination results were represented using letter grades (A and B).

The target variable used in this study was Graduation Status, which consisted of two classes: On Time and Not On Time. These categories were determined based on the students graduation records obtained from the academic database.

The transformation process was carried out to standardize the dataset, reduce data complexity, and optimize the probability calculations performed by the Naïve Bayes algorithm. As a result, the transformed dataset became more suitable for classification and prediction of students graduation status.

4.5. Data Mining

The prepared students academic data were then processed through a data mining modeling stage using the Naïve Bayes algorithm to predict whether students would graduate on time or not. The Naïve Bayes algorithm was chosen because it is capable of performing classification processes based on probabilities derived from previous data.

The stages of the data mining process using the Naïve Bayes algorithm began with calculating the number of classes from the prepared training data, followed by calculating the number of cases in each class, and then calculating the probability of each attribute for a particular class. Afterward, all attribute probabilities were multiplied to obtain the final probability value for each class. The classification result was determined based on the highest probability value, enabling accurate prediction of students graduation status.

4.6. Calculating the Number of Classes

After completing the preprocessing and transformation stages, the dataset consisted of 52 valid student records. Based on the graduation status attribute, the data were divided into two classes, namely on time and not on time. A total of 39 students were categorized as graduating on time, while 13 students were categorized as graduating not on time. The prior probability calculations are presented as follows:

$$P(On\ Time) = \frac{39}{52} = 0.75$$
$$P(Not\ On\ Time) = \frac{13}{52} = 0.25$$

Table 7. Prior Probability Calculations

Class	Amount	Class Probability P(C)
On Time	39	0.75
Not On Time	13	0.25
Total	52	1.00

Note : The prior probability is represented by the notation P(Ci).

4.7. Calculating the Number of Cases in Each Class

After obtaining the prior probability values, the next step was to calculate the conditional probability of each academic achievement category for each graduation class.

The academic achievement attribute consisted of three categories : High, Medium, and Fair. The distribution of these categories in each class is shown in Table 8.

Table 8. Probabilities Based on Academic Performance

Academic Achievement	On Time	Not On Time
High	16	1
Medium	22	10
Fair	1	2
Total	39	13

Based on the table above, the conditional probabilities were calculated as follows :

High Academic Achievement

$$P(High \mid On\ Time) = \frac{16}{39} = 0.410$$
$$P(High \mid Not\ On\ Time) = \frac{1}{13} = 0.077$$

Medium Academic Achievement

$$P(Medium \mid On\ Time) = \frac{22}{39} = 0.564$$
$$P(Medium \mid Not\ On\ Time) = \frac{10}{13} = 0.769$$

Fair Academic Achievement

$$P(Fair \mid On\ Time) = \frac{1}{39} = 0.026$$
$$P(Fair \mid Not\ On\ Time) = \frac{2}{13} = 0.154$$

The probability values indicate that students with High academic achievement are predominantly found in the On Time class, whereas Fair academic achievement appears more frequently in the Not On Time class.

4.8. Multiplying All Class Variables

The prior probability and conditional probability values were then combined to determine the classification result using the Naïve Bayes algorithm.

As an example, suppose a student has the academic achievement category High.

A. Probability of High Academic Achievement in the On Time Class

$$P(High \mid On\ Time) = \frac{16}{39} = 0.410$$

B. Probability of High Academic Achievement in the Not On Time Class

$$P(High \mid Not\ On\ Time) = \frac{1}{13} = 0.077$$

C. Multiplication with Prior Probability for the On Time Class

$$P(On\ Time) \times P(High \mid On\ Time)$$
$$0.75 \times 0.410$$
$$= 0.308$$

D. Multiplication with Prior Probability for the Not On Time Class

$$P(Not\ On\ Time) \times P(High \mid Not\ On\ Time)$$
$$0.25 \times 0.077$$
$$= 0.019$$

The result probabilities are:

Table 9. Results Probabilities based on Class

Class	Class Score
On Time	0.308
Not On Time	0.019

The values 0.308 and 0.019 represent unnormalized Naïve Bayes class scores. The observation is classified into the "On Time" class

because its score is higher than that of the "Not On Time" class.

4.9. Comparing the Results of Each Class

The results of the testing data probability calculations are shown in the following Table The probability values for each class were calculated to predict student graduation status, and the final classification was assigned to the class with the highest probability value.

Table 10. Testing Data Probability Calculation Results :

Class	Class Score
P(On Time | High Academic Achievement)	0.308
P(Not On Time | High Academic Achievement)	0.019

The results indicate that the probability value for the On Time class is considerably higher than that of the Not On Time class. Based on these probability values, the testing data were classified into the On Time graduation category. This finding suggests that students with higher GPAs are more likely to complete their studies and graduate within the expected time frame.

4.10. Evaluation

The evaluation stage was conducted to assess the performance of the Naïve Bayes classification model in predicting students' graduation status. Due to the relatively small and imbalanced dataset consisting of 39 On Time and 13 Not On Time records, the model was evaluated using Stratified k-Fold Cross Validation. The "Not On Time" class was defined as the positive class because the objective of the study was to identify students at risk of delayed graduation.

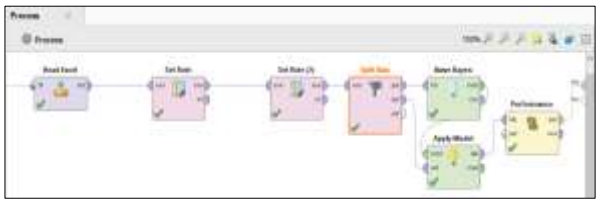
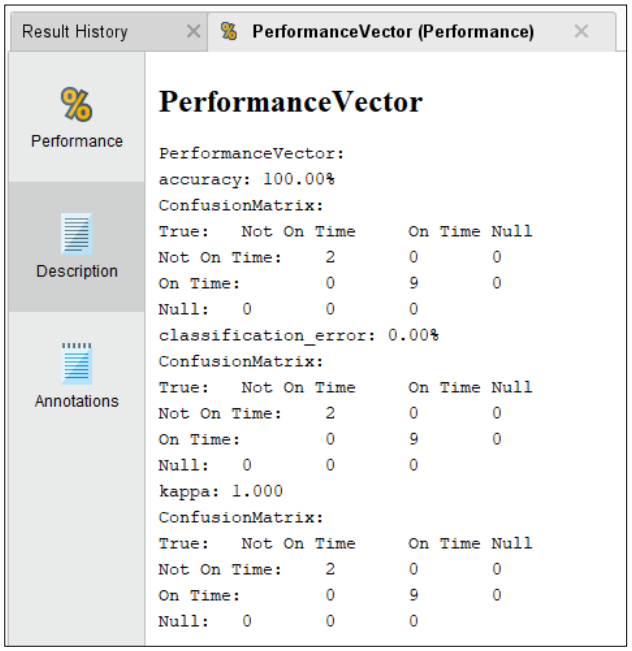


Figure 4. Process

The Naïve Bayes classification process implemented in RapidMiner. The process begins with data input and preprocessing, followed by the application of the Naïve Bayes algorithm for classification. Model performance is then evaluated using Cross Validation and Performance (Classification) operators to measure the accuracy and reliability of the classification results.

Figure 5. PerformancesVector (Table View)

The confusion matrix and classification performance results obtained from RapidMiner. The model achieved an accuracy of 100%, with class precision and class recall values of 100% for both the On Time and Not On Time classes, indicating that all testing instances were classified correctly.



Figur 6. Description

Detailed PerformanceVector results obtained from RapidMiner. The model achieved an accuracy of 100%, a classification error of 0.00%, and a kappa value of 1.000. The confusion matrix indicates that all instances were correctly classified into the On Time and Not On Time classes.

Table 11. Confusion Matrix

Predicted Class	Actual Not On Time	Actual On Time
Not On Time	2	0
On Time	0	9

Based on the confusion matrix, all instances were correctly classified by the Naïve Bayes model. Two records belonging to the Not On Time class and nine records belonging to the On Time class were predicted correctly.

Table 12. Performance Evaluation Results

Metric	Value
Accuracy	100.00%
Classification Error	0.00%
Kappa	1.000

The evaluation results indicate that the Naïve Bayes model achieved an accuracy of 100%, with no classification errors and a kappa value of 1.000, indicating perfect agreement between predicted and actual classes.

Table 13. Class-Specific Performance Metrics

Class	Precision	Recall	F1-Score
Not On Time (Positive Class)	100.00%	100.00%	100.00%
On Time	100.00%	100.00%	100.00%

The model achieved perfect precision, recall, and F1-score for both classes, indicating that all testing instances were classified correctly.

The accuracy value of 100% shown in Table 10 was obtained from a particular testing iteration in RapidMiner, where all testing records were classified correctly. However, this result does not represent the overall performance of the model. To obtain a more reliable evaluation, the model was assessed using Stratified k-Fold Cross Validation. The cross-validation results showed that the highest performance was achieved in the 3-fold scenario, with an accuracy of 88.73%, precision of 64.62%, recall of 45.72%, and a kappa value of 0.451. Therefore, the cross-validation results are used as the main performance indicators in this study because they provide a more representative estimate of the model's predictive performance.

5. Conclusion

This study applied the Naïve Bayes algorithm within the Knowledge Discovery in Databases (KDD) framework to predict students' graduation status in the Informatics Engineering Study Program, Faculty of Engineering, Universitas Islam Kuantan Singingi. The study utilized students' academic data, including Gender and GPA, as predictor variables to classify graduation status into On Time and Not On Time categories.

The results indicate that academic performance, represented by GPA, is associated with graduation timeliness, where students with higher GPA values tend to have a greater likelihood of graduating on time. The Naïve Bayes classification process also demonstrated its capability to classify graduation status based on the selected academic attributes.

The evaluation results showed that the highest performance was achieved in the 3-fold cross-validation scenario, producing an accuracy of 88.73%, precision of 64.62%, recall of 45.72%, and a kappa value of 0.451. These findings suggest that the Naïve Bayes algorithm has potential for predicting students' graduation timeliness. However, this study has several limitations. The dataset used in this research consisted of only 52 student records obtained from a single graduation period and a single study program, which may limit the generalizability of the prediction model. In addition, the classification process utilized a limited number of predictor attributes, primarily GPA, while other academic and non-academic factors that may influence graduation timeliness were not included in the analysis.

Future research should involve larger and more diverse datasets, additional academic and non-academic predictor variables, and comparative evaluations using other classification algorithms to improve predictive performance and support the development of a more reliable academic early-warning system.

References

Alfa Saleh. (2015). 18-6 Implementasi Metode Klasifikasi Naïve Bayes.

Alghifari, F., & Juardi, D. (2021). Fauzan Alghifari Penerapan Data Mining Pada Penerapan Data Mining Pada Penjualan Makanan Dan Minuman Menggunakan Metode Algoritma Naïve Bayes.

Anggraini, M., Ayuning Tyas, R., Ana Sulasiyah, I., & Aini, Q. (2020). Implementasi Algoritma Naïve Bayes Dalam Penentuan Rating Buku. www.kaggle.com

Bagus, I., Peling, A., Arnawan, N., Putu, I., Arthawan, A., & Janardana, I. (2017). Implementation of Data Mining To Predict Period of Students Study Using Naive Bayes Algorithm. In International Journal of Engineering and Emerging Technology (Vol. 2, Number 1).

Etriyanti, E., Kunang, Y. N., & Syamsuar, D. (2020). Implementasi Data Mining Menggunakan Algoritme Naive Bayes Classifier dan C4.5 untuk Memprediksi Kelulusan Mahasiswa. Telematika, 13(1), 56–67. https://doi.org/10.35671/telematika.v13i1.881

Felicia Watratan, A., Puspita, A. B., Moeis, D., Informasi, S., & Profesional Makassar, S. (2020). Implementasi Algoritma Naive Bayes Untuk Memprediksi Tingkat Penyebaran Covid-19 Di Indonesia. In JOURNAL OF APPLIED COMPUTER SCIENCE AND TECHNOLOGY (JACOST) (Vol. 1, Number 1). http://journal.isas.or.id/index.php/JACOST

Fitriani, U., & Wibowo, A. (2023). 3 rd Seminar Nasional Mahasiswa Fakultas Teknologi Informasi (SENAFTI) 30 Agustus 2023-Jakarta (Vol. 2, Number 2).

Gigih Putra Kawani. (2019). Implementasi Naive Bayes Untuk Menentukan Wadah Limbah B3 Sesuai Karakteristik. 1(2), 73–081. https://doi.org/10.20895/INISTA.V1I2

Ihsan A. Abu Amra, A. Y. A. M. (2017). Students Performance Prediction Using KNN and Naïve Bayesian. IEEE.

Julkarnain, M., & Yustiardin, M. (2024). Penerapan Algoritma Naive Bayes dalam Memprediksi Lulus Tepat Waktu Mahasiswa. Digital Transformation Technology, 4(2), 848–858. https://doi.org/10.47709/digitech.v4i2.4963

Meiriza, A., Lestari, E., Putra, P., Monaputri, A., & Lestari, D. A. (2020). Advances in Intelligent Systems Research (Vol. 172).

Mujib Ridwan, H. S. dan M. S. (2013). 18-6 Penerapan Data Mining Untuk Evaluasi Kinerja.

Nangi, J., & Rinaldi Hadistio, R. (2025). Penerapan Data Mining Untuk Memprediksi Penjualan Menggunakan Metode Knearest Neighbor (Studi Kasus: Thrifting Second 3). ANIMATOR, 3(3), 1–5.

Rama, K., Lubis, P., & Sitohang, S. (2023). Penerapan Data Mining Dengan Metode Naive Bayes Classifier Pada Penjualan Barang Untuk Optimasi Strategi Pemasaran. In JURNAL COMASIE. http://ejournal.upbatam.ac.id/index.php/comasiejournal

Satrio Junaidi, Valicia Anggela, R., & Kariman, D. (2024). Klasifikasi Metode Data Mining untuk Prediksi Kelulusan Tepat Waktu Mahasiswa dengan Algoritma Naïve Bayes, Random Forest, Support Vector Machine (SVM) dan Artificial Neural Nerwork (ANN). Journal of Applied Computer Science and Technology, 5(1), 109–119. https://doi.org/10.52158/jacost.v5i1.489