Contents lists available at https://csi.unisi.my.id

Computer Science Insights

Journal Page is available to https://csi.unisi.my.id

Research article

Analysis of Student Attendance in Information Systems Class B Cohort 2025 Using the Naïve Bayes Algorithm

*M Isnaldi [1] , Yoga Apririandi [2], M Fikrian Nurhadi [3], Satria Eko Saputra

[1,2,3] Universitas Islam Indragiri
email: [1,*] isnaldi13102018@gmail.com , [2] yogaapriri@gmail.com, [3] fkry8332@gmail.com, [4] ekol6850@gmail.com
* Correspondence

ARTICLE INFO

Article history:
Received November 26, 2025
Revised December 22, 2025
Accepted December 23, 2026
Available online June 2, 2026

Keywords:
Naïve Bayes,
Student Attendance,
Classification,
Machine Learning,
Data Analys

ABSTRAK

Student attendance is an important indicator for evaluating discipline and engagement in the learning process; however, manual attendance systems remain vulnerable to recording errors and data inconsistencies. This study analyzes attendance patterns of Information Systems Class B students from the 2025 cohort by applying the Gaussian Naïve Bayes algorithm. The dataset consists of 71 attendance records collected from 10 students over four lecture days, including features such as lecture day, arrival time, login activity, and prior attendance frequency. The data were processed through cleaning, encoding, normalization, and stratified train–test splitting with an 80:20 ratio. The Naïve Bayes model achieved perfect accuracy (1.000) on the test dataset, with precision, recall, and F1-score values of 1.00 for both classes. However, this perfect performance suggests potential issues such as highly homogeneous data structure, small dataset size, or overly strong feature–label relationships, indicating possible overfitting or data leakage. This study highlights the potential of simple probabilistic classification methods for student attendance analysis while emphasizing the need for more rigorous validation using larger and more diverse datasets.

1. Introduction

Student attendance is a fundamental component of the academic evaluation process, as it reflects discipline, engagement, and commitment to learning. Numerous studies have shown that attendance patterns are closely correlated with academic achievement and the level of student participation in learning activities. Nevertheless, many educational institutions still rely on manual attendance systems that are prone to recording errors, inconsistencies, and potential manipulation of attendance data.

In line with advancements in information technology, data-driven approaches have increasingly been adopted to identify patterns of student behavior. Machine learning techniques offer the capability to recognize patterns from historical data and generate probability-based predictions. One of the most widely used algorithms for classification task**s** is Naïve Bayes**,** which operates by estimating posterior probabilities under the assumption of feature independence.

Previous related studies have applied the Naïve Bayes algorithm to predict student attendance using demographic factors. However, the use of questionnaire-based data in those studies limits prediction validity, as such data do not fully reflect actual student behavior. In contrast, the present study utilizes digital attendance data obtained directly from the academic system, making the analysis more representative and contextually grounded.

Algoritma Native Bayes

Naïve Bayes is one of the most effective and efficient inductive learning algorithms used in machine learning and data mining. Naïve Bayes is a classification method based on probabilistic and statistical principles proposed by the British scientist Thomas Bayes, which predicts future probabilities based on past observations and is therefore known as Bayes' Theorem.

According to Danandjaja in his book *Metode Penelitian Kepustakaan*, Naïve Bayes classification works by dividing a problem into several classes based on similarities and differences in characteristics, using statistical methods to predict the probability of each class (Danandjaja, 2014).

The foundation of this algorithm is Bayes' Theorem, which is mathematically expressed as follows:

$$P(X \mid Y) = \frac{P(Y \mid X) \times P(X)}{P(Y)}$$

Where:

$P(X \mid Y)$	is the posterior probability
$P(X \mid Y)$	is the likelihood
$P(X)$	is the prior probability
$P(Y)$	is the normalization factor

This theorem describes how the probability of a hypothesis is revised in light of new evidence or observed data.
In classification contexts, the theorem is rewritten as:

$$P(C \mid F_1, F_2, \dots, F_n) = \frac{P(C) \times P(F_1, F_2, \dots, F_n \mid C)}{P(F_1, F_2, \dots, F_n)}$$

Where:

C	: the class or category to be predicted (e.g., *"Present"* or *"Absent"*)
$F_1, F_2, \dots, F_n$	: features or attributes (e.g., lecture day, arrival time, present/absent status)
P(C)	: prior probability of class C
$P(F_1, F_2, \dots, F_n \mid C)$	: probability of the occurrence of features given that class C is true
$P(F_1, F_2, \dots, F_n)$	: total probability of the features (as a normalization constant)

To simplify the computation, Naïve Bayes assumes that each feature is conditionally independent given the class. Therefore, the above equation can be rewritten as:

$$P(C \mid F_1, F_2, \dots, F_n) = \frac{P(C) \times \prod_{i=1}^{n} \quad P(F_i \mid C)}{P(F_1, F_2, \dots, F_n)}$$

Since the denominator $P(F_1, F_2, \dots, F_n)$is a constant for all classes, in the classification process the equation can be expressed as:

$$P(C \mid F_1, F_2, \dots, F_n) \propto P(C) \times \prod_{i=1}^{n} \quad P(F_i \mid C)$$

In the context of this study, the Naïve Bayes algorithm is used to analyze and classify attendance patterns of Information Systems students, specifically Class B of the 2025 cohort. Each student attendance record is represented as a set of features, such as:

- Lecture day
- Arrival time
- Attendance status (Excused, Absent, Sick)
- Number of previous absences

The Naïve Bayes model is then applied to calculate the probability that a student belongs to a certain attendance category, such as *"frequently present," "moderately present,"* or *"rarely present."*

Research Methods

This study employs a descriptive quantitative approach with the application of the Naïve Bayes algorithm to analyze student attendance data. The objective of the study is to identify patterns of student attendance behavior and to classify attendance status based on several influential features. The general stages of the research are described as follows.

1. Data Collection

The data were obtained from the attendance records of 10 Information Systems Class B students from the 2025 cohort over four lecture days. A total of 71 entries were collected, with each record consisting of the following attributes:

- Lecture day (Monday–Thursday)
- Arrival time (decimal time format)

- Login activity (1 = logged in, 0 = not logged in)
- Previous attendance frequency
- Attendance label (*"Present"* or *"Absent"*)

2. Data Preprocessing

The preprocessing stage was conducted to ensure that the data were suitable for processing by the algorithm. The preprocessing steps are as follows:

1. DataCleaning
   Checking for duplicate records and ensuring data completeness.
2. Categorical Variable Encoding
   The lecture day feature was converted into numerical representation.
   The attendance label was encoded in binary form (0 = *Present*, 1 = *Absent*).
3. Numerical Feature Normalization
   Normalization was applied to arrival time, login activity, and attendance frequency to prevent certain features from dominating the classification process.
4. Dataset Splitting
   The dataset was divided into training data (80%) and testing data (20%) using stratified sampling to preserve class proportions.

3. Gaussian Naïve Bayes Algorithm

The Gaussian Naïve Bayes algorithm was employed because the dataset contains numerical variables that approximately follow a normal distribution. The algorithm calculates the probability of each class based on the input features and assigns the class with the highest posterior probability.

4. Model Training and Testing

The model was trained using the training dataset to estimate the distribution parameters of each feature. Model performance was evaluated using the testing dataset based on the following metrics:

- Accuracy
- Precision
- Recall
- F1-score
- Confusion matrix

This stage aims to assess the model's ability to recognize simple patterns within the dataset rather than to produce a final predictive model.

2. Results and Discussion

1. Data collection

Data collection was conducted by retrieving attendance records of Information Systems Class B students from the 2025 cohort**.** The data were collected from 10 students only, due to limited data collection time, which consequently restricted the number of records used in this study.

The attendance data from these 10 students were organized in a spreadsheet format and categorized into several attributes, including student name, lecture day, arrival time, login activity, course, attendance frequency, and attendance status, as illustrated in Figure 1**.**

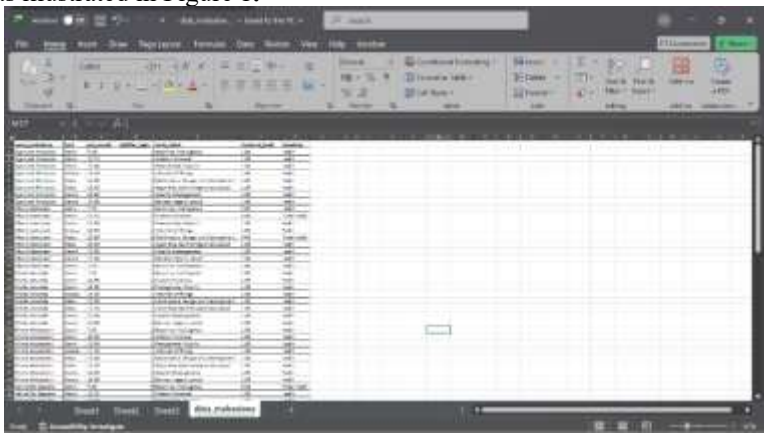

Figure 1 Data from 10 system information students class B

1. Dataset Preparation

The data were subsequently converted into CSV format, and each column was assigned an appropriate data type. The student name field was treated as a string, the lecture day as a numerical variable, arrival time as a numerical value, login activity as a numerical variable, course name as a string, attendance frequency as a numerical variable, and attendance status as a string. This conversion was necessary to ensure that the dataset could be properly processed by the Python programming environment used in the subsequent analysis.

2. Model Training Using the Naïve Bayes Algorithm

After the dataset format was defined, the next step was to train the Naïve Bayes algorithm. The training process began by importing the required libraries. Prior to importing the libraries, the computing environment was initialized; in this study, Google Colab was used as the execution platform. The libraries imported included Pandas for data handling and Naïve Bayes modules from the machine learning library, as illustrated in Figure 2.

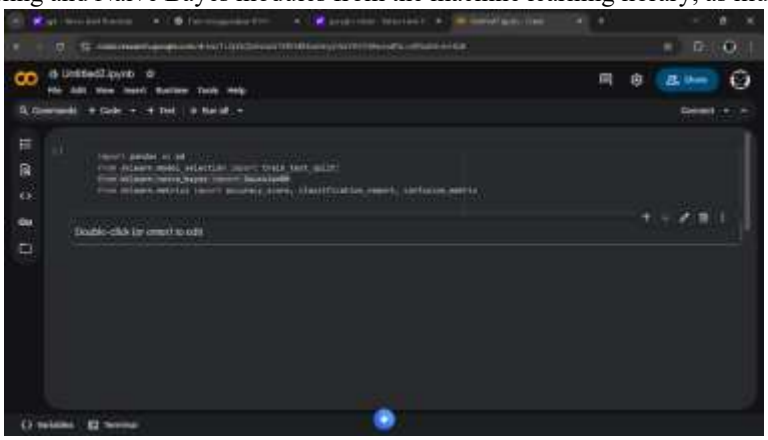

Figure 2 Apperance from *Google Collabs* and *import library*

4. Model Testing Using the Test Dataset

After importing the required libraries, the next step was to evaluate the model using the test dataset. The test data consisted of student attendance records that had been separated during the dataset splitting process. The model testing process and the resulting outputs are shown in Figures 3 and 4.

Figure 3 Testing model at *Naïve Bayes* algoritm

Figure 4 Testing model at *Naïve Bayes* algoritm

5. Evaluation Results and Analysis of Student Attendance
The results of the analysis are presented as follows.

A. Key Findings of Data Analysis

The initial dataset used in this study, namely *data_mahasiswa.csv*, consisted of 71 entries and 7 variables, with no missing values detected after the file was successfully loaded using a semicolon (;) as the delimiter. The target variable, attendance, was encoded during preprocessing, where *"Present"* was represented by the value 0, and *"Absent"* by the value 1.

Feature preprocessing included scaling of numerical features (arrival time, login activity, and attendance frequency), as well as the application of one-hot encoding to categorical features (lecture day and course). As a result of this preprocessing stage, a total of 15 features were generated and prepared for further analysis.

The dataset was then divided into two subsets: a training set (64 samples) and a testing set (17 samples), using an 80:20 ratio. Stratified sampling was applied to maintain class proportions, resulting in 56 "Present" and 8 "Absent" observations in the training set, and 15 "Present" and 2 "Absent" observations in the testing set.

1. Model Performance Evaluation

Evaluation of the Gaussian Naïve Bayes model produced an accuracy of 1.000, with precision, recall, and F1-score values of 1.00 for both classes. The confusion matrix indicated that all test samples—15 "Present" and 2 "Absent"—were correctly classified.

From a mathematical perspective, this result indicates that the feature distributions in the test dataset fall entirely within the range of patterns learned by the model from the training data. However, in the context of predictive modeling, such perfect performance should not be interpreted as a sign of robustness, but rather as an indication that further and more rigorous validation is required.

2. Examination of Potential Data Leakage

Ideal performance on a small dataset is often indicative of data leakage**,** a condition in which information from the target label unintentionally "leaks" into the feature set. An examination of the dataset structure revealed several potential sources of leakage:

1. Login Activity Feature
   If students who are marked as *"Present"* consistently log in to the system, while those marked as *"Absent"* do not, this feature effectively acts as a proxy for the class label. In such cases, the model is not learning underlying patterns, but merely reproducing a direct correlation.Distribusi

2. Arrival Time Distribution

   If arrival time values for the *"Absent"* class consistently fall into a single value range (for example, zero or missing values that have been replaced with a default value), the model can easily distinguish between classes without learning meaningful attendance behavior patterns.

3. Overly Deterministic Categorical Encoding

   The use of one-hot encoding on categorical features with very few distinct values may create rigid decision boundaries, causing the model to follow the dataset structure literally rather than learning generalizable patterns.

3. Class Imbalance Analysis

The dataset structure shows a significant imbalance between the *"Present"* and *"Absent"* classes. The minority class consists of only:

- 8 samples in the training dataset
- 2 samples in the testing dataset

This level of imbalance has several implications:

1. Stability of Evaluation Metric
   The recall value for the *"Absent"* class appears perfect; however, it is not statistically representative because it is evaluated on only two samples.

2. Overfitting to the Majority Class
   In small datasets, the model tends to focus on dominant attributes associated with the majority class, without adequately learning the behavioral variation of the minority class.

3. Limited Generalization Capability
   With a disproportionate class distribution, the model reflects patterns present only in the limited dataset rather than general student attendance behavior.

4. Limitations of Dataset Size and Variability
   The dataset consists of only 71 entries derived from observations of 10 students over four lecture days. This scale is insufficient to:
   - stably estimate Gaussian distributions,
   - capture variability in student attendance behavior,
   - assess model robustness against noise and outliers, or
   - represent patterns that can be generalized.

   An overly homogeneous sample also makes a simple algorithm such as Naïve Bayes appear to perform "perfectly," even though the dataset lacks sufficient complexity to meaningfully challenge the model.

5. Feature Relevance and Suitability
   A scientific evaluation of the selected features reveals several conceptual limitations:

a. Arrival Time
   The arrival time feature is theoretically relevant; however, if absent students are assigned identical default values, this feature effectively becomes a pseudo-label rather than an informative predictor.

b. Login Activity
   This feature should be statistically tested for its correlation with the target label. If the correlation coefficient exceeds 0.9, the feature should not be used in a predictive model, as it eliminates the essence of pattern learning by directly encoding the outcome.

c. Previous Attendance Frequency
   This feature may introduce a feedback loop, whereby students with a history of frequent absences continue to be predicted as absent regardless of other influencing factors.

d. Lecture Day
   With only four lecture days and limited variability, this feature contributes minimally to classification performance. Overall, the features were not statistically validated prior to model training, which limits the analytical quality of the study.

6. Model Validity and Potential Overfitting
   Based on the dataset structure and model performance, strong indications of overfitting arise from the following factors:
   - absolute agreement between training and testing results,
   - small dataset size,
   - overly simple and deterministic features,
   - imbalanced target distribution, and
   - potential information leakage.

   Although Naïve Bayes is generally considered resistant to overfitting, perfect performance on a highly homogeneous and limited dataset cannot be regarded as valid or generalizable.

7. Evaluation Limitations
   Model evaluation was conducted only once using a single train–test split. The study did not employ:
   - k-fold cross-validation,
   - repeated holdout validation,
   - residual analysis,
   - model variance testing, or
   - baseline comparisons (e.g., logistic regression or decision tree models).

   Without baseline comparisons, it cannot be determined whether Naïve Bayes outperforms alternative methods—or whether the dataset is so simple that any classifier would achieve perfect accuracy.

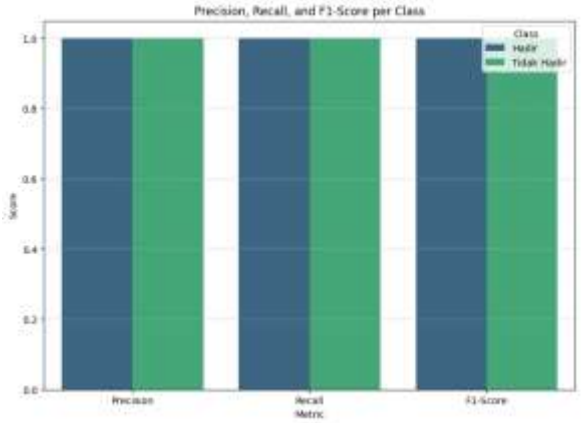

Figure 5 Model Testing Results of Student Attendance Data Using the Naïve Bayes Algorithm

B. Insights and Recommendations for Future Work

The model's performance, which achieved accuracy, precision, recall, and F1-score values of 1.00 on the test dataset, represents a result that is rarely observed in scientific research. Such perfect performance typically indicates non-ideal conditions in the modeling process, such as an excessively small dataset, a highly homogeneous data structure, or the presence of potential data leakage whereby label information is indirectly reflected in the features. Therefore, these results must be interpreted with caution.

To ensure that the model possesses adequate generalization capability, subsequent efforts should focus on a more comprehensive validation process. The use of larger and more diverse datasets would allow the model to learn more representative patterns of student attendance behavior. In addition, the application of cross-validation techniques is strongly recommended to evaluate the stability of model performance across different data partitions.

Furthermore, the class distribution within the dataset exhibits a significant imbalance, with the *"Absent"* class represented by only a small number of samples (8 in the training set and 2 in the testing set). This imbalance may cause the model to optimize predictions toward the majority class while failing to properly capture the characteristics of the minority class. Consequently, further analysis of the impact of class imbalance is necessary. Approaches such as increasing the number of samples in the minority class, applying oversampling techniques (e.g., SMOTE), or adjusting class weights may be considered to achieve a more balanced class distribution and improve prediction quality.

Overall, although the initial results appear perfect, further validation stages and improvements in dataset quality are crucial to ensure that the Naïve Bayes model is truly capable of robustly recognizing student attendance patterns and can be reliably applied to broader and more complex data contexts.

3. Conclusion

This study aims to analyze and classify attendance patterns of Information Systems Class B students from the 2025 cohort using the Gaussian Naïve Bayes algorithm. By utilizing digital attendance data obtained directly from the academic system—including lecture day, arrival time, login activity, and prior attendance frequency—the classification model was successfully implemented. Evaluation results on the test dataset (17 samples) demonstrated exceptionally high performance, achieving accuracy, precision, recall, and F1-score values of 1.000 for both classes, *"Present"* and *"Absent"*. These results indicate that the Naïve Bayes model is highly effective in identifying attendance-related patterns within the dataset used in this study. The findings suggest that the Naïve Bayes model is capable of classifying simple and homogeneous attendance data. However, these results cannot be used as a basis to claim the model's effectiveness in other contexts or more complex datasets. Improvements in data quality, increased observational diversity, and the application of cross-validation techniques are required to enhance the robustness and reliability of the findings.

References

[1] S. González, S. García, F. Herrera, J. DelSer, and L. Rokach, "A practical tutorial on bagging and boosting based ensembles for machine learning: Algorithms, software tools, performance study, practical perspectives and opportunities," *Inf. Fusion*, vol. 64, 2020, doi: 10.1016/j.inffus.2020.07.007.

[2] K. Hameed, D. Chai, and A. Rassau, "A Progressive Weighted Average Weight Optimisation Ensemble Technique for Fruit and Vegetable Classification," in *16th IEEE International Conference on Control, Automation, Robotics and Vision, ICARCV 2020*, 2020, pp. 303–308, doi: 10.1109/ICARCV50220.2020.9305474.

[3] Z. Zhang, Y. Chen, J. Tang, and X. Luo, “An ensemble learning algorithm with Gaussian-based oversampling,” *Xitong Gongcheng Lilun yu Shijian/System Eng. Theory Pract.*, vol. 41, no. 2, pp. 513–523, 2021, doi: 10.12011/SETP2020-1790.
[4] G. Chicco, “Ant colony system-based applications to electrical distribution system optimization,” in *Ant Colony Optimization - Methods and Applications*, A. Otsfeld, Ed. Rijeka, Croatia: InTechOpen, 2011, pp. 237–262.
[5] Susanti Lusi, Daulay Khairani Nelly, Intan Bunga, “Sistem Absensi Mahasiswa Berbasis Pengenalan Wajah Menggunakan Algoritma YOLOv5” *JURIKOM (Jurnal Riset Komputer*) Vol. 10 No. 2, April 2023, DOI 10.30865/jurikom.v10i2.6032
[6] H. Hartatik, “Optimasi Model Prediksi Kelulusan Mahasiswa Menggunakan Algoritma Naive Bayes,” *Indones. J. Appl. Informatics*, vol. 5, no. 1, p. 32, Apr. 2021, doi: 10.20961/ijai.v5i1.44379.
[7] Rovidatul, Y. Yunus, and G. W. Nurcahyo, “Perbandingan algoritma c4.5 dan naive bayes dalam prediksi kelulusan mahasiswa,” *J. CoSciTech (Computer Sci. Inf. Technol.*, vol. 4, no. 1, pp. 193–199, Apr. 2023, doi: 10.37859/coscitech.v4i1.4755.