Integra: Journal of Integrated Mathematics and Computer Science

Journal Homepage: https://integrajimcs.com/index.php/integra/index

integra
Journal of Integrated Mathematics and Computer Science

Research Paper

Comparison of Naïve Bayes and Random Forest Models in Predicting Undergraduate Study Duration Classification at the University of Lampung

Shelvira Hestina P.[1*], Widiarti[1], Aang Nuryaman[1], Mustofa Usman[1]

[1] *Department of Mathematics, Faculty of Mathematics and Natural Science, Universitas Lampung, Lampung, 35141, Indonesia*
[*] **Corresponding author**: shelvirahestinap@gmail.com

Keywords
Naïve Bayes, Random Forest, Classification, Study Duration, On-time Graduation

Abstract
This study aims to compare the performance of the Naïve Bayes and Random Forest classification algorithms in predicting the study duration of undergraduate students in the Mathematics Study Program at the University of Lampung. The dataset consists of 537 graduation records from 2020–2024. The research steps include data preprocessing, data partitioning (train-test split and k-fold cross validation), model building, and evaluation using a confusion matrix. The results show that the Random Forest algorithm achieved the highest accuracy of 94.44%, outperforming Naïve Bayes which reached a maximum accuracy of 92.59%. These findings suggest that Random Forest is more effective for classifying student study durations. These findings suggest that Random Forest is more effective for classifying student study durations.

Received: 7 August 2024, Accepted: 13 October 2024
https://doi.org/10.26554/integrajimcs.20241317

1. INTRODUCTION

Timely graduation is essential for maintaining educational quality and institutional performance. According to Pêgo [1], extended study durations not only reflect academic inefficiencies but also highlight the need for proactive academic guidance, as prolonged time to degree completion may signal deeper structural or individual challenges within the academic journey. To address the growing concern of academic underperformance, many higher education institutions have begun leveraging machine learning techniques to predict student outcomes and implement early academic interventions. Oyedeji [2] demonstrate that the use of predictive models can effectively uncover at-risk students and support data-driven strategies for improving academic achievement. Moreover, driven by the growing availability of student data, predictive analytics and machine learning are increasingly being adopted in higher education to anticipate learning difficulties, identify academic risks, and provide personalized support systems [3].

Machine learning has become a key technology in Educational Data Mining (EDM), offering reliable tools for predicting student performance, identifying dropout risks, and modeling learning behaviors [4, 5]. Among various ML algorithms, Naïve Bayes and Random Forest are widely used due to their interpretability, efficiency, and strong empirical performance in educational datasets [6, 7].

Naïve Bayes, based on Bayes' Theorem, assumes feature independence and calculates class probabilities to make predictions [8], The classification rule for Naïve Bayes is defined as:

$$P(H_k|X_i) = \frac{P(X_i|H_k)P(H_k)}{P(X_i)} \tag{1}$$

Where:

- $P(H_k|X_i)$ is the posterior probability of class H_k given predictor X_i
- $P(X_i|H_k)$ is the likelihood of predictor X_i given class H_k
- $P(H_k)$ is the prior probability of class (H_k)
- $P(X_i)$ is the prior probability of predictor X

In Classification, the label assigned is the one with the highest posterior probability:

$$\hat{H} = argmax_{H_k}\left(P(H_k)\prod_{i=1}^{n}P(X_i|H_k)\right) \quad (2)$$

Despite its simplicity, Naïve Bayes has shown reliable performance in various educational prediction tasks due to its low computational cost and resilience to noise [9].

In contrast, Random Forest builds multiple decision trees and predicts outcomes via majority voting, offering strong performance on complex data [10], It is known for its robustness, resistance to overfitting, and ability to capture non-linear patterns in data.

Given a dataset $D = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, Random Forest builds T decision trees on bootstrap samples D_t, and the final class prediction is made by:

$$\hat{y} = \text{mode}\left(\{h_t(x)\}_{t=1}^{T}\right) \quad (3)$$

Where $h_t(x)$ is the prediction of the t–th decision tree. Random Forest is robust to overfitting, handles non-linear data well, and performs efficiently on imbalanced datasets [11, 12].

In the context of Indonesian higher education, particularly at the University of Lampung, classification algorithms such as Naïve Bayes and Random Forest have already shown promising results. As demonstrated by Kurniasari [13], both models effectively categorized academic performance data, with Random Forest achieving higher predictive accuracy. This suggests the practical relevance of applying these algorithms to support academic decision-making at the institutional level.

Recent studies have demonstrated the applicability of these classifiers in academic settings. Bakri [12] and Hartanto [14] compared Random Forest with other models in predicting timely graduation and showed its superior accuracy. Nakhipova [15] and Farhana [16] applied Naïve Bayes for academic performance prediction, while Akanbi [17] discussed trends in predictive analytics for educational success. Other works [18, 19, 20] explored early risk detection, dropout prediction, and academic performance modeling using both Naïve Bayes and Random Forest, confirming their effectiveness and practicality.

This study aims to evaluate and compare the performance of Naïve Bayes and Random Forest algorithms in predicting undergraduate study duration classification. The dataset consists of 537 graduation records from the Mathematics Study Program at the University of Lampung, spanning 2020–2024. The outcome variable is whether a student graduates "on time" ($\leq$ 4 years) or "late" (> 4 years). The results are expected to support the development of academic early warning systems and enhance decision-making processes in higher education institutions.

2. METHODS

2.1 Data

The data was sourced from the graduation records of undergraduate students in the Mathematics Study Program at the University of Lampung between 2020 and 2024, totaling 537 entries. The class label represents the study duration ($\leq$ 4 years = on time, >4 years = delayed), with attributes including admission track, scholarship, organizational involvement, completed credits (SKS), and GPA.

2.2 Preprocessing Data

The preprocessing phase involved several key steps to prepare the data for modeling:

1. View and Handle Missing Values
   Missing value occurs when there is data or variables in the sample that are not observed. In this study, there were no variables that had missing values. So that data preprocessing can be continued to the next stage.
2. Attribute Selection
   Irrelevant attributes such as student identification numbers and personal information were removed to avoid noise and protect privacy.
3. Categorical Encoding
   Categorical variables were transformed into numerical representations using label encoding. For example, admission type (SNMPTN, SBMPTN, Mandiri) and organizational involvement (Yes/No) were encoded into numeric categories.
4. Feature Scaling
   Numerical attributes such as GPA and completed credits (SKS) were standardized using Scikit-learn's `StandardScaler` to normalize value ranges and improve model convergence.

2.3 Data Partitioning

To evaluate model performance and generalizability, two data partitioning strategies were applied:

1. Train-Test Split
   The dataset was divided into training subsets using four different ratios namely 60% data training and 40% data testing, 70% data training and 30% data testing, 80% data training and 20% data testing and 90% data training and 10% data testing.
2. K-Fold Cross Validation
   To further enhance robustness, k-fold cross validation was performed with three values of k, namely k=5, k=8 and k=10.

2.4 Model Construction

The models were built using Python (Scikit-learn) employing Gaussian Naïve Bayes and RandomForestClassifier.

2.5 Evaluation

Model performance was evaluated using the confusion matrix to compute accuracy, precision, recall, F-1 Score and conditional probabilities.

3. RESULTS AND DISCUSSION

This study evaluates the performance of Naïve Bayes and Random Forest classification models in predicting undergraduate

study duration at the University of Lampung. Validation was conducted using both Train-Test Split and K-Fold Cross Validation methods, and model performance was measured based on accuracy metrics. Accuracy metrics were computed using confusion matrix analysis.

3.1 Data Preprocessing

The preprocessing phase consisted of four major steps applied to the 537 student graduation records:

1. Missing Value Handling
   All variables were inspected for missing values. Table 1. shows that there were no missing values in any of the attributes, allowing the analysis to proceed without imputation procedures.

Table 1. Variable Missing Value

Variables	Missing Value
Study Duration (year)	0
Student ID	0
Name	0
Gender	0
Year of Entry	0
Year of Graduation	0
Admission Pathway	0
Scholarship Status	0
Organization	0
Total Credits Earned	0
GPA	0

2. Data Reduction
   Irrelevant attributes such as student ID (NPM), name, gender, admission year, and graduation year were excluded because they did not contribute significantly to the predictive modeling. Table 2 displays the data before and after reduction.

Table 2. Data Before and After Variable Reduction

Variables	Variables After Reduction
Study Duration (year)	Study Duration (year)
Student ID	Admission Pathway
Name	Scholarship Status
Gender	Organization
Year of Entry	Total Credits Earned
Year of Graduation	GPA
Admission Pathway	
Scholarship Status	
Organization	
Total Credits Earned	
GPA	

3. Categorical Encoding
   In the categorical encoding stage, a labeling process will be applied to the data. The categorical encoding technique used is **LabelEncoder**, which automatically sorts the categories alphabetically and then assigns numerical labels starting from 0, 1, 2, and so on. Table 3 displays the categorial encoding.
4. Feature Scaling
   To normalize the range of continuous features, `StandardScaler` from Scikit-learn was applied. This standardization method ensures that each numeric attribute (e.g., GPA, SKS) has a mean of 0 and standard deviation of 1, which is particularly important for algorithms sensitive to feature scale such as Naïve Bayes.

3.2 Model Performance Overview

The results of the classification experiments are summarized in Table 4. Each model was evaluated across four data partition ratios (60:40, 70:30, 80:20, and 90:10) for the train-test split method, and three different k-values (k=5, 8, and 10) for k-fold cross-validation.

Table 4 shows that Random Forest consistently outperforms Naïve Bayes in terms of accuracy across both Train-Test Split and K-Fold Cross Validation methods. The highest accuracy was recorded by Random Forest with a 90:10 train-test split (94.44%) and 10-fold cross validation (93.14%).

3.3 Comparison of Naïve Bayes and Random Forest Models

Naive bayes and random forest methods have the highest accuracy value when using the splitting method compared to the k-fold cross validation method. The highest accuracy value of the naive bayes method is 92.59% and random forest is 94.44%. Figure 1 displays the Comparison the models with Splitting and K-Fold Cross Validation Methods.

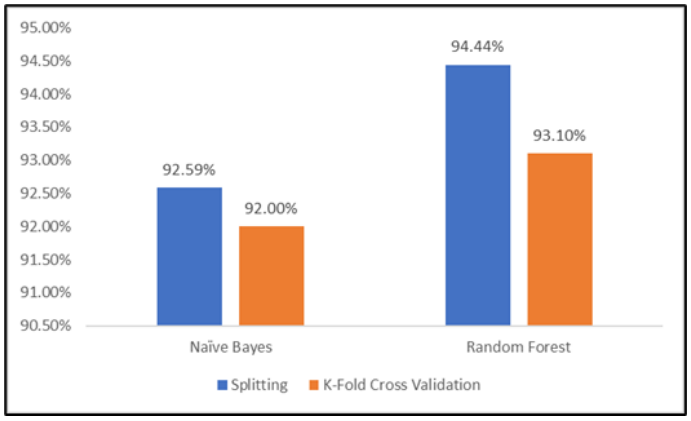

Figure 1. Comparison the Models with Splitting and K-Fold Cross Validation Methods

Random Forest excels due to its ensemble nature that combines multiple decision trees to handle complex data interactions. Meanwhile, Naïve Bayes performs optimally when feature independence is assumed, which may not hold in this case. Nevertheless, Naïve Bayes still demonstrates stable and competitive performance.

Table 3. Categorical Encoding

Variables	Categories	Encoded Values
Study Duration (Y)	≤4 years, ≥4 years	0 or 1
Admission Pathway (X_1)	SNMPTN, SBMPTN, Mandiri, PMPAP	0, 1, 2, or 3
Scholarship Status (X_2)	No, Yes	0 or 1
Organization (X_3)	No, Yes	0 or 1
Total Credits Earned (X_4)	<144, ≥144	0 or 1
GPA (X_5)	≤3.00, 3.00-3.49, ≥3.50	0, 1, or 2

Table 4. Testing Results of Naïve Bayes and Random Forest Models with Splitting and K-Fold Cross Validation Method

Validation Methods	Naïve Bayes (%)	Random Forest (%)
Train-Test Split 60:40	87.90	87.90
Train-Test Split 70:30	89.50	86.41
Train-Test Split 80:20	91.66	90.74
Train-Test Split 90:10	92.59	94.44
K-Fold Cross Validation (k=5)	89.81	91.81
K-Fold Cross Validation (k=8)	90.16	92.67
K-Fold Cross Validation (k=10)	91.08	93.14

4. CONCLUSIONS

Random Forest achieved higher accuracy than Naïve Bayes in predicting undergraduate study duration at the University of Lampung. The highest accuracy was obtained using a 90:10 train-test split and 10-fold cross validation.

5. ACKNOWLEDGEMENT

The authors wish to thank Statistics Research Group Universitas Lampung for the support provided.

REFERENCES

[1] J. P. Pêgo, V. L. Miguéis, and A. Soeiro. Students' complex trajectories: Exploring degree change and time to degree. *International Journal of Educational Technology in Higher Education*, 21(1):5, 2024.

[2] A. O. Oyedeji, A. M. Salami, O. Folorunsho, and R. D. Ojerinde. Analysis and prediction of student academic performance using machine learning. *Journal of Computer Engineering and Intelligent Systems*, 11(2):21–29, 2020.

[3] R. Umer, T. Susnjak, A. Mathrani, and S. Hill. Current stance on predictive analytics in higher education: Opportunities, challenges and future directions. *Interactive Learning Environments*, pages 1–19, 2023.

[4] C. Romero and S. Ventura. Educational data mining and learning analytics: An updated survey. *arXiv preprint*, pages 49–56, 2024.

[5] Y. Sun, Y. Liu, J. Zhang, and H. Yu. Multi-source data fusion and ensemble learning model for early warning of college student dropout. *IEEE Access*, 8:149165–149177, 2020.

[6] Aurélien Géron. *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*. O'Reilly Media, 3rd edition, 2022.

[7] Sebastian Raschka and Vahid Mirjalili. *Python Machine Learning*. Packt Publishing, 3rd edition, 2022.

[8] Daniel Berrar. *Bayes' theorem and naive Bayes classifier*. The Open University, 2025. The Open University.

[9] K. Roy and D. M. Farid. An adaptive feature selection algorithm for student performance prediction. *IEEE Access*, 12:55678–55689, 2024.

[10] Matthias Schonlau and Renhai Y. Zou. The random forest algorithm for statistical learning. *The Stata Journal*, 20(1):3–29, 2020.

[11] M. Wang. *Stacking ensemble model for liver stiffness classification with imbalanced data*. Doctoral dissertation, ProQuest Dissertations Publishing, 2021.

[12] R. Bakri, N. P. Astuti, and A. S. Ahmar. Evaluating random forest algorithm in educational data mining: Optimizing graduation on-time prediction using imbalance methods. *ARRUS Journal of Social Sciences and Humanities*, 4(1):108–116, 2024.

[13] D. Kurniasari, R. N. Hidayah, Notiragayu, Warsono, and R. K. Nisa. Classification models for academic performance: A comparative study of naïve bayes and random forest algorithms in analyzing university of lampung student grades. *Jurnal Teknik Informatika (JUTIF)*, 5(5):1853–1861, 2024.

[14] M. B. Hartanto, T. Destanto, Y. Yuniarthe, and T. Winarko. Implementation of data mining for classifying student graduation levels using naïve bayes, decision tree, random forest, support vector machines and neural networks methods. *CCIT Journal*, 18(1):80–87, 2024.

[15] V. Nakhipova, Y. Kerimbekov, Z. Umarova, M. Abishev, and D. Tsoy. Use of the naive bayes classifier algorithm in machine learning for student performance prediction. *International Journal of Information and Education Technology*, 14(2):162–167, 2024.

[16] S. Farhana. Classification of academic performance for university research evaluation by implementing modified naive bayes algorithm. *Procedia Computer Science*, 192:1176–1185, 2021.

[17] O. B. Akanbi. Application of naive bayes to students' performance classification. *Asian Journal of Probability and Statistics*, 21(4):20–28, 2023.

[18] A. L. J. Martinez, K. Sood, and R. Mahto. Early detection of at-risk students using machine learning. In *Proceedings of the World Congress in Computer Science*, pages 45–57, 2025.

[19] E. Ahmed. Student performance prediction using machine learning algorithms. *Applied Computational Intelligence and Soft Computing*, 2024.

[20] J. M. Aiken, R. De Bin, M. Hjorth-Jensen, and M. D. Caballero. Predicting time to graduation at a large enrollment american university. *PLOS ONE*, 15(8):28–35, 2020.