JTPED
Journal of Technological Pedagogy and Educational Development
Vol. 1, No. 1, 2024, pp. 30-43
https://ejournal.jtped.org/ojs/index.php/jtped

Comparing Machine Learning Models in Predicting On-Time Graduation with Emphasis on Feature Importance

Zulfa Safina Ibrahim [a,1,*], Rifqah Khairunnisa [b,2], Madiha Syarifah Balqis [c,3]

[a] Universitas Negeri Yogyakarta, Sleman, Yogyakarta 55281, Indonesia
[b] Zonguldak Bülent Ecevit University, İncivez Mahallesi, Zonguldak 67100, Turkey
[c] Al-Azhar University, Al-Azhar Street, Cairo 11651, Egypt
[1] zulfasafina.ibr00@gmail.com; [2] rkhairunnisa50@gmail.com; [3] madihabalqiskapo@gmail.com
* Corresponding Author

ARTICLE INFO

Article history
Received June 30, 2025
Revised September 21, 2025
Accepted October 23, 2025

Keywords
Machine Learning;
Student Performance;
Random Forest;
Educational Data Mining

ABSTRACT

Timely graduation remains a key performance indicator in higher education and is closely linked to institutional efficiency and student success. In Indonesia, many students fail to graduate on time, resulting in resource inefficiencies and delayed workforce entry. Previous studies have primarily used conventional statistical methods such as logistic regression to analyze factors influencing graduation, but these approaches are limited in capturing complex, non-linear interactions. This study addresses this gap by applying a machine learning (ML) approach to predict on-time graduation while integrating logistic regression to enhance interpretability. The main contribution of this research lies in developing a hybrid model that balances predictive accuracy and interpretability, providing actionable insights for higher education institutions and aligning with Sustainable Development Goal (SDG) 4. Three ML algorithms (Random Forest, Support Vector Machine, and Naïve Bayes) were applied to a dataset comprising 18 academic, demographic, and institutional variables from an Indonesian university. Model performance was evaluated using accuracy, sensitivity, specificity, AUC, and Kappa metrics. Logistic regression was used to test the significance of key predictors. Results show that Random Forest achieved the highest overall accuracy (75.13%) and AUC (0.7021), while SVM and Naïve Bayes exhibited complementary strengths in sensitivity and specificity. Feature importance analysis highlighted GPA, faculty affiliation, and total credits as key predictors. These findings demonstrate the potential of combining ML and statistical techniques to support data-informed decisions in higher education and align with SDG 4 objectives. However, this study is limited to a single institution, which may affect the generalizability of the findings. Future research could extend the model to multi-institutional datasets for broader validation.

1. Introduction

Timely graduation is widely recognized as a critical indicator of a well-functioning, inclusive, and sustainable higher education system [1], [2]. This notion is embedded in Sustainable Development Goal (SDG) 4, particularly Target 4.3, which promotes inclusive and equitable access to quality tertiary education for all by 2030 [3]. According to OECD's "Education at a Glance" report, countries with high on-time graduation rates often maintain responsive, data-informed education systems that

https://doi.org/10.59247/jtped.v1i1.7

continuously adapt to student needs [4]. In Indonesia and across Southeast Asia, timely graduation remains a persistent challenge, particularly among students in open and private universities [5], [6]. This issue contributes to inefficient resource use, delayed labor market entry, and growing concerns over academic management [7].

Numerous studies have examined a wide range of factors influencing student graduation outcomes, including GPA, credit accumulation, admission pathways, faculty affiliation, gender, and socioeconomic background [8], [9]. While these investigations have provided valuable insights, many rely on traditional statistical techniques such as logistic regression or discriminant analysis. These methods, although established, often fall short in detecting non-linear relationships or the intricate interplay among multiple variables. In contrast, recent advancements in big data analytics and artificial intelligence have opened new avenues for understanding graduation dynamics more comprehensively. Machine learning (ML), in particular, has been increasingly adopted to enhance predictive accuracy in higher education research [4], [10].

Among widely used ML models, Random Forest [11], Support Vector Machine (SVM) [12], and Naïve Bayes [13] have demonstrated strong performance in educational prediction tasks. Random Forest, in particular, is praised for its accuracy and ability to quantify the importance of features in classification tasks [14]. Accuracy rates exceeding 95% have been reported when combined with optimization techniques such as Particle Swarm Optimization (PSO) [9]. Similarly, parameter tuning has been shown to improve model reliability in Indonesian higher education settings [15]. These algorithms represent distinct methodological paradigms (ensemble learning, geometric margin classification, and probabilistic modeling, respectively) and their comparative evaluation provides valuable insights into educational data mining. However, despite their strong predictive performance, many prior studies using these algorithms lack interpretability, an essential aspect for informing educational policy. For example, although One-Class SVM achieves near-perfect accuracy in graduation classification [16], it offers limited insights into policy-relevant factors.

In recent developments in this field, some researchers have proposed integrating classification models with interpretative approaches to support evidence-based decision-making. A study by [6] demonstrated that integrating deep learning techniques with feature analysis (e.g., GPA and age) produced a robust model for predicting timely graduation. Similar results were also demonstrated by [17], who successfully modified Random Forest to improve prediction accuracy on student data. At the global level, ML approaches are also used to predict the achievement of Sustainable Development Goals (SDGs) through national education performance analysis [18].

Therefore, this study proposes a hybrid approach that combines ML classification and logistic regression to predict timely graduation among students at an Indonesian university. Three ML algorithms (Random Forest, SVM, and Naïve Bayes), were compared using classification metrics such as accuracy, AUC, and sensitivity. Feature importance was analyzed through Random Forest, while logistic regression was used to evaluate statistical significance and support policy interpretation. The main contribution of this research lies in building an interpretable and accurate model tailored to the Indonesian context, with practical relevance to educational decision-making and alignment with SDG 4.3 in developing countries. While previous studies have applied machine learning to predict graduation outcomes, few have emphasized model interpretability and its alignment with the SDG 4 framework, particularly within the Indonesian higher education context. The remainder of this paper is organized as follows: Section 2 presents the research methodology, Section 3 discusses the results and analysis, and Section 4 concludes with implications and future research directions.

2. Method

This section describes the dataset structure, the machine learning algorithms used in this study, variable selection process, and evaluation procedures. The aim is to compare predictive performance and interpretability across models. Data preprocessing involved handling missing values using

Journal of Technological Pedagogy and Educational Development
Vol. 1, No. 1, 2024, pp. 30-43

mean/mode imputation, encoding categorical variables with one-hot encoding, and addressing class imbalance through stratified sampling.

The research flowchart (Fig. 1) illustrates the overall process of this study, beginning with problem identification and literature review on timely graduation, machine learning, and SDG 4.3. It continues with data collection from university records, followed by preprocessing steps such as handling missing values, encoding categorical variables, and splitting data into training and testing sets. Three machine learning algorithms (Random Forest, SVM, and Naïve Bayes) were developed and evaluated using accuracy, sensitivity, specificity, AUC, and Kappa through 10-fold cross-validation. The best-performing model was then analyzed for feature importance and further interpreted using logistic regression to identify key predictors of timely graduation and derive policy-relevant insights.

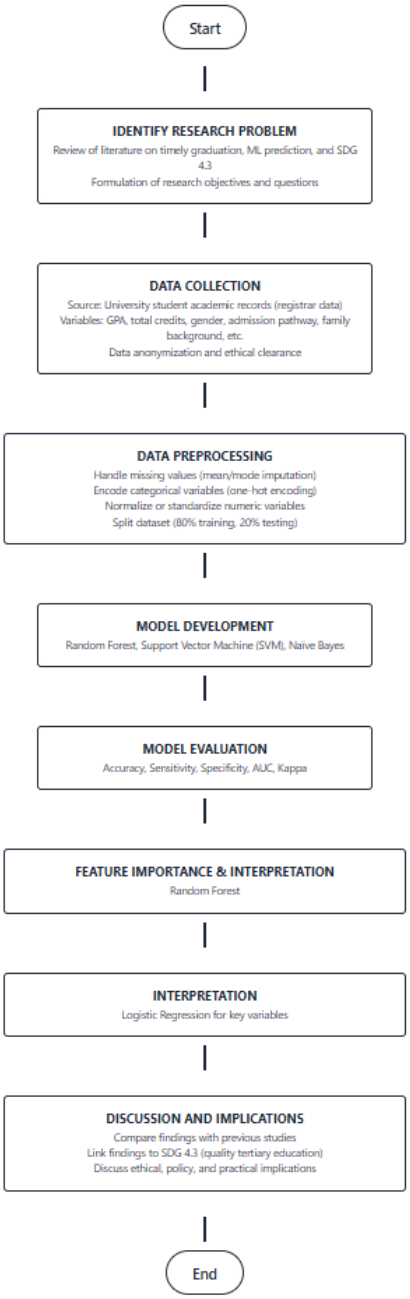

Fig. 1. Research Flowchart

2.1. Dataset and Feature Description

This dataset consists of academic and demographic records of students collected from a university in Indonesia. All student data were anonymized and used in accordance with institutional ethical approval and privacy regulations. The main predictor variables include GPA, total credits, admission pathway, student type, and faculty affiliation. There are a total of 18 variables considered, including numerical and categorical features. The target variable is binary: 1 for graduating on time and 0 for graduating late. Table 1 provides the operational definitions of the variables used.

Table 1. Operational Definitions of Variables

Variable	Operational Definition	Type and Scale
Graduated	Description of student study success	Categorical (1 = Graduated on time, 0 = Graduated late)
Admission Path	Entry route into the university	Categorical (e.g., SNBT, SNBP, etc.)
Age	Student's age at the time of university enrollment	Numeric
Cooperation	Enrollment through cooperation or institutional partnership	Categorical (1 = Yes, 0 = No)
Faculty	Faculty to which the student belongs	Categorical (e.g., FBSB, FEB, etc.)
Father's Education	Highest education level attained by the father	Categorical (ordinal)
Father's Income	Monthly income level of the student's father	Categorical (ordinal income brackets)
Father's Occupation	Type of occupation held by the father	Categorical
Gender	Student's biological sex	Categorical (1 = Male, 2 = Female)
GPA Mean	Average grade point from semesters 1 to 3	Numeric
Group	Scholarship status or financial support group	Categorical (1 = Bidikmisi, 2 = Non)
Mother's Education	Highest education level attained by the mother	Categorical (ordinal)
Mother's Income	Monthly income level of the student's mother	Categorical (ordinal income brackets)
Mother's Occupation	Type of occupation held by the mother	Categorical
National Exam Score	Total score from high school national examination	Numeric
Region Group Code	Classification of student's origin region	Categorical
Special Needs	Indicates whether the student has special needs	Categorical (1 = Yes, 0 = No)
Student Type	Type of student program	Categorical (e.g., Regular, Transfer, etc.)
Total Credits	Total academic credits taken in semesters 1 to 3	Numeric

2.2. Machine Learning Models

To accurately predict student graduation status in a timely manner, this study applies three machine learning algorithms with different approaches, namely Support Vector Machine (SVM), Random Forest (RF), and Naive Bayes (NB). Each model represents a distinct methodological family (geometric, ensemble, and probabilistic respectively). The SVM uses a radial basis function kernel, RF uses Gini impurity to construct decision trees, and NB assumes conditional independence among features. These three models are used to classify the target variable based on several predictor features such as semester GPA average (ipk_means), total credit hours (sks_total), as well as demographic and administrative attributes such as Faculty, Admission Pathway, Student Type, Group, and Gender.

The SVM model is used with a radial kernel (Radial Basis Function) to handle potential non-linearity between features. The kernel function used is defined as:

$$K(x_i, x_j) = \exp\left(-\gamma \left|\left|x_i - x_j\right|\right|^2\right) \quad (1)$$

In this formulation, x_i and x_j represent the feature vectors of the i-th and j-th student respectively. The term $\left|\left|x_i - x_j\right|\right|^2$ refers to the squared Euclidean distance between the two data points. The

parameter γ (gamma) controls the width of the RBF kernel and determines how much influence a single training example has. The exponential function exp is used to convert this distance into a similarity measure; smaller distances produce values closer to 1, indicating higher similarity.

To construct the optimal classification boundary, the SVM algorithm aims to find a hyperplane that maximizes the margin between classes. This is achieved by minimizing the objective function:

$$\min_{w,b,\xi} \frac{1}{2}||w||^2 + C\sum_{i=1}^{n}\xi_i \quad (2)$$

In this equation, w is the weight vector that defines the orientation of the hyperplane, and b is the bias term that determines its offset. The variable ξ_i represents a slack variable for each data point, allowing for some misclassification or tolerance in the margin. The constant C is a regularization parameter that controls the trade-off between maximizing the margin and minimizing classification errors. The optimization is subject to the constraint:

$$y_i(w.\emptyset(x_i) + b) \geq 1 - \xi_i \quad (3)$$

Here, y_i denotes the class label of the i-th sample, which typically takes values of either +1 or –1. The function $\emptyset(x_i)$ represents a transformation of the input features into a higher-dimensional space where a linear separation may become possible. The inner product $w.\emptyset(x_i)$ calculates the projection of x_i onto the hyperplane's normal vector. The inequality ensures that each data point is either correctly classified or within an acceptable margin of error defined by ξ_i.

This mathematical formulation allows the SVM to perform robust classification, particularly when dealing with complex, non-linear data patterns commonly found in educational datasets. This method is grounded in the principle of maximum-margin classification, as outlined by [19] and [20]. On the other hand, the Random Forest (RF) algorithm adopts a fundamentally different approach from SVM by constructing an ensemble of decision trees. Each tree is trained on a randomly selected subset of the data and features, a process known as bootstrap aggregation or bagging. The final prediction is determined through a majority voting mechanism among all individual trees, which enhances robustness and reduces the risk of overfitting.

At each decision node within a tree, the algorithm selects the best feature to split the data by minimizing Gini impurity, a metric that reflects the degree of class impurity. The Gini impurity at node ttt is calculated using the formula:

$$G(t) = 1 - \sum_{i=1}^{C} p_i^2 \quad (4)$$

In this expression, p_i denotes the proportion of samples belonging to class i in node t, and C represents the total number of classes. A lower Gini value indicates a purer node, i.e., a node dominated by a single class, which is desirable for improving classification accuracy. Once all trees in the forest are constructed, the model generates its final prediction by aggregating the outputs of each tree. This aggregation is commonly performed using a majority voting scheme, formalized as:

$$\hat{y} = mode(h_1(x), h_2(x), \ldots, h_B(x)) \quad (5)$$

Here, $h_b(x)$ represents the prediction of the b-th decision tree in the ensemble for input sample x, and B is the total number of trees. The function $mode$ selects the class label that appears most frequently among all predictions. This ensemble strategy enhances generalization and leverages the diversity of trees to improve overall model accuracy and stability.

As noted in Machine Learning with R [21], Random Forest (RF) is particularly valued for its robustness against overfitting and its capability to assess the relative importance of input features. In

contrast, the Naïve Bayes (NB) classifier adopts a probabilistic framework grounded in Bayes' theorem to estimate the likelihood of a data point belonging to a given class. It operates under the simplifying assumption that all predictor features are conditionally independent given the class label. This leads to the following general formulation:

$$P(C_k|x_1, x_2, \dots, x_n) \propto P(C_k). \prod_{i=1}^{n} P(x_i|C_k) \quad (6)$$

In this expression, $P(C_k|x_1, x_2, \dots, x_n)$ is the posterior probability that an observation with features $x_1, x_2, \dots, x_n$ belongs to class C_k. $P(C_k)$ represents the prior probability of class C_k, and $P(x_i|C_k)$is the likelihood of feature x_i given class C_k. The product across all features reflects the independence assumption, which significantly simplifies computation.

For numerical features such as GPA and total credits, NB assumes that feature values follow a Gaussian (normal) distribution within each class. Under this assumption, the likelihood $P(x_i|C_k)$ is computed using the formula:

$$P(x_i|C_k) = \frac{1}{\sqrt{2\pi\sigma_{ik}^2}} exp\left(-\frac{(x_i - \mu_{ik})^2}{2\pi\sigma_{ik}^2}\right) \quad (7)$$

Here, μ_{ik} and σ_{ik}^2 are the mean and variance of feature x_i within class C_k, respectively. This allows the model to generate continuous probability estimates based on observed values, which is particularly useful for academic data containing real-valued indicators. Despite its strong assumption of independence among features, Naïve Bayes is widely used due to its computational efficiency, scalability, and surprisingly good performance in many practical applications, especially when the dimensionality of the data is high. Its simplicity makes it a strong baseline model for classification problems, including those involving educational datasets.

This model is widely used because of its simplicity and efficiency, as discussed in [22] and [21]. The combination of these three models not only allows for the evaluation of prediction accuracy, but also enables the comparison of geometric (SVM), ensemble (RF), and probabilistic (NB) approaches in the context of higher education data classification.

3. Results and Discussion

3.1. Feature Importance Analysis

Random Forest analysis indicated that GPA Mean, Faculty, and Total Credits were the most influential variables. Mean Decrease Accuracy and Mean Decrease Gini were used to rank the features. These findings are aligned with academic literature highlighting GPA as a strong early predictor of academic success.

In the analysis of predictors of graduation using the Random Forest algorithm (Table 2), two important metrics are used to assess the contribution of each variable: Mean Decrease Accuracy (MDA) and Mean Decrease Gini (MDG). MDA measures how much the model's accuracy decreases when the values of a variable are randomized. The higher the MDA value, the more important the variable is for prediction accuracy. Meanwhile, MDG indicates how much a variable helps separate target classes in the decision tree, where the higher the MDG value, the more frequently and effectively the variable is used in optimal data separation.

Based on these two metrics, seven variables were selected because they had the highest MDA values and conceptually reasonable interpretations. These variables are: GPA Mean, Faculty, Total Credits, Gender, Group (Bidikmisi), Admission Path, and Student Type. GPA Mean is the most dominant predictor, indicating that initial academic performance strongly influences graduation.

Faculty and total credits are also strongly correlated, reflecting the influence of institutional academic background and academic readiness. Gender and scholarship group provide insights into the social background that influences academic performance. Meanwhile, admission path and student type reflect the selection process and administrative differentiation that impact students' learning experiences. The selection of these seven variables not only considers statistical strength but also policy relevance and interpretive clarity, making them a strong foundation for building an accurate and practically applicable graduation prediction model in higher education settings.

Table 2. Variable Importance Rankings from Random Forest Model

Variable	Mean Decrease Accuracy	Mean Decrease Gini
GPA Mean	93.02	1264.83
Faculty	62.10	540.70
Total Credits	53.25	713.69
Gender	31.87	152.25
Group	21.99	82.67
Admission Path	20.57	339.67
Student Type	9.54	98.25
National Exam Score	9.08	950.26
Mother's Income	9.06	186.98
Cooperation	9.12	9.39
Region Group Code	9.85	143.82
Father's Education	8.12	468.37
Mother's Education	5.51	452.42
Father's Occupation	3.74	680.94
Father's Income	3.80	298.71
Age	2.18	274.84
Mother's Occupation	1.21	546.04
Special Needs	–0.98	5.59

3.2. Interpretation through Logistic Regression

Logistic regression confirmed the significance of several predictors, especially GPA, Faculty, and Admission Path. Some categories within Admission Path and Student Type also showed statistically significant contributions, offering nuanced insights for policy formulation.

Fig. 2 is a visualization of the logistic regression results to see the effect of each predictor on the probability of graduating on time. Based on the results of the logistic regression analysis, several variables show different effects on the probability of students graduating on time. The Admission Path variable exhibits highly variable effects; some admission paths have a strong negative influence on graduation (e.g., Admission Path 6), while others have a significant positive influence (e.g., Admission Path 9). This underscores that the type of admission pathway can be a critical determinant of student academic success.

The Student Type variable also contributes differently depending on its category. Some student types, such as regular or transfer students, show significant effects, both positive and negative, on graduation. Meanwhile, the GPA Mean consistently has a positive and significant effect on the likelihood of graduating on time. This indicates that academic performance during the early stages of study is a strong predictor of academic success at the end of the program.

The Gender variable appears to have a small but significant effect, depending on the interpretation of the confidence interval. Similarly, some categories within the faculty variable show significant effects on graduation, indicating differences in characteristics or academic demands between faculties. Meanwhile, the Total Credits variable does not show a significant effect, as its estimated value is close to zero and has a narrow confidence interval. Finally, the Group variable also shows a very small and statistically insignificant effect because its confidence interval cuts through zero. Overall, these results indicate that academic variables such as GPA and administrative

characteristics such as admission pathway and student type have a greater contribution in predicting timely graduation compared to other administrative variables such as group or number of credits.

Based on the results of logistic regression analysis (Table 3), several variables showed a significant influence on the likelihood of students graduating on time. First, the average GPA (GPA Mean) had a positive and highly significant influence on graduating on time (estimate = 0.817, p < 2e-16), meaning that the higher a student's GPA, the greater their chances of graduating on time. Conversely, the total number of credit hours (Total Credits) has a negative effect (estimate = -0.015, p = 6.82e-07), indicating that students with a heavier credit load tend to have a lower likelihood of graduating on time.

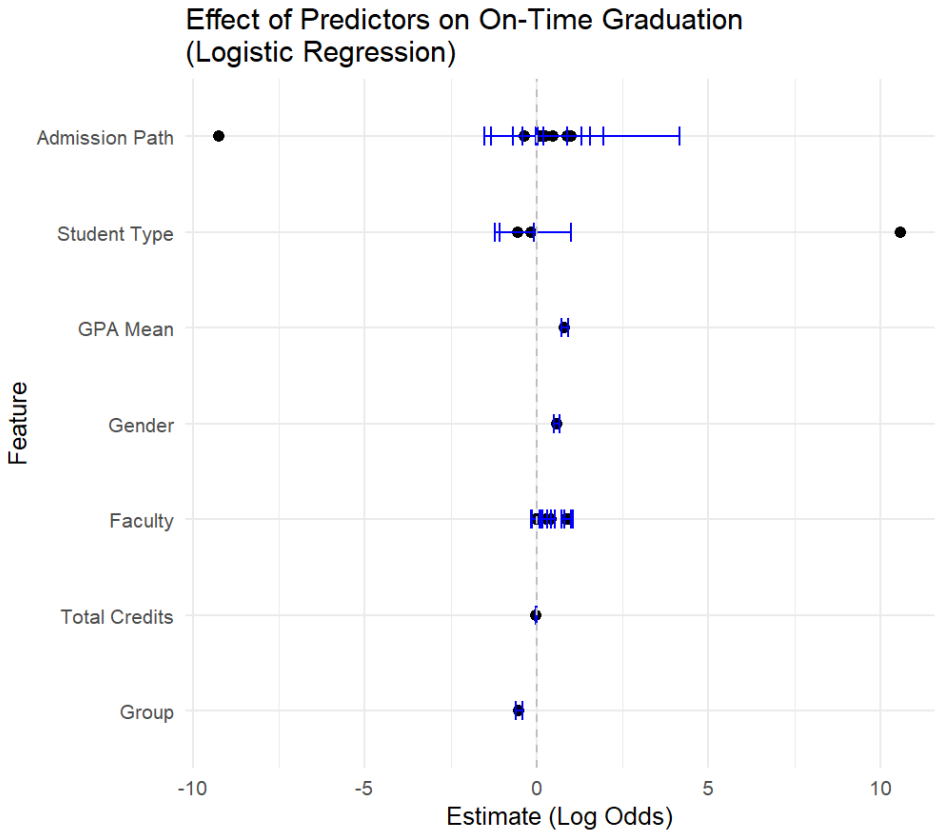

Fig. 2. Predictor Effects on On-Time Graduation (Logistic Regression)

Table 3. Logistic Regression Estimates for On-Time Graduation Predictors

Predictor	Estimate	p-value
GPA Mean	0.817	< 2e-16
Total Credits	-0.015	6.82e-07
Student Type4	-0.552	0.029
Admission Path2	0.112	0.007
Admission Path5	0.880	0.078
Faculty2	0.289	1.97e-05
Faculty3	0.869	< 2e-16
Faculty4	0.413	7.96e-13
Faculty6	0.929	< 2e-16
Group2	-0.512	< 2e-16
Gender2	0.584	< 2e-16

From the students' background perspective, Student Type 4 (Transfer/etc.) shows a significant negative effect on graduation (estimate = -0.552, p = 0.029), indicating that students with this

background are more likely to fail to graduate on time compared to regular subsidized students. Regarding admission pathways (Admission Path), the SNBP pathway (code 2) contributes significantly positively to timely graduation (estimate = 0.112, p = 0.007), while the SM-SNBT pathway (code 5) shows a positive trend though not yet statistically significant (estimate = 0.880, p = 0.078).

Furthermore, from the faculty perspective, certain faculties such as FIKK (code 3), FIPP (code 4), and FMIPA (code 6) have a significantly positive influence on timely graduation compared to the reference faculty (FBSB, code 1), with estimated coefficients of 0.869, 0.413, and 0. 929, respectively. This indicates that students from these faculties have a higher tendency to complete their studies on time. For the group variable (Group), non-bidikmisi students (code 2) have a lower likelihood of graduating on time compared to bidikmisi students (estimate = -0.512, p < 2e-16), which may be related to differences in financial or administrative support.

Finally, gender is also a significant factor, with female students (code 2) showing a higher likelihood of graduating on time compared to males (estimate = 0.584, p < 2e-16). Overall, academic variables (GPA, credit hours), demographic variables (gender, group), and institutional variables (faculty, admission pathway, and student type) significantly contribute to the likelihood of students graduating on time.

3.3. Predictive Accuracy Comparison

A comparative analysis of the performance of Support Vector Machine (SVM), Random Forest (RF), and Naïve Bayes (NB) models reveals several important findings in predicting timely graduation (Table 4). Of the three, the Random Forest model recorded the highest accuracy of 75.13%, followed by SVM at 73.81%, and Naïve Bayes at 71.12%. However, accuracy alone is not sufficient to describe the overall quality of a model, especially in cases of imbalanced data where the majority class (e.g., students who graduate on time) dominates. Therefore, additional metrics such as sensitivity, specificity, and balanced accuracy are important for a fairer evaluation.

Table 4. Comparison of Classification Metrics

Metric	SVM	Random Forest	Naïve Bayes
Accuracy	0.7381	0.7513	0.7112
Sensitivity	0.9940	0.9577	0.1752
Specificity	0.0316	0.1815	0.9054
Balanced Accuracy	0.5128	0.5696	0.5403
Kappa	0.0367	0.1794	0.0982
McNemar's Test P-Value	<2e-16	<2.2e-16	<2e-16

The SVM model shows very high sensitivity (99.40%), indicating its excellent ability to identify students who graduate on time. However, this model has very low specificity (3.16%), meaning it is less effective at detecting students who do not graduate on time. Conversely, Naïve Bayes shows the opposite pattern, with the highest specificity (90.54%) but very low sensitivity (17.52%), making it more effective in detecting students who do not graduate on time but missing many who do. Random Forest offers more balanced performance with a sensitivity of 95.77% and specificity of 18.15%, and records the highest balanced accuracy of 56.96%.

When viewed from the Positive Predictive Value (PPV) score (Table 5), Random Forest also excels with a score of 76.36%, indicating that this model's predictions of timely graduation are more reliable. On the other hand, Naïve Bayes has the highest Negative Predictive Value (NPV) score of 75.19%, indicating greater confidence in predicting students who will not graduate on time. This complementary predictive strength indicates that each model has specific advantages depending on the institution's objectives. Whether the goal is to minimize false negatives (failing to detect students who graduate late) or false positives (incorrectly classifying students as graduating late when they are not).

Table 5. PPV and NPV Comparison among ML Algorithms

Metric	SVM	Random Forest	Naïve Bayes
Positive Predictive Value (PPV)	0.7392	0.7636	0.4015
Negative Predictive Value (NPV)	0.6557	0.6085	0.7519

In terms of Kappa statistics (Table 6), which measure the agreement between predictions and actual data after accounting for the possibility of random agreement, Random Forest again recorded the highest value of 0.1794, indicating a better level of agreement than the other two models. All models also produced significant McNemar test values ($p < 0.05$), indicating an imbalance in the distribution of predictions across classes, which is common in binary classification with unequal class prevalence.

Table 6. Kappa Statistics and McNemar's Test Results for Classification Models

Metric	SVM	Random Forest	Naïve Bayes
Kappa	0.0367	0.1794	0.0982
McNemar's Test P-Value	<2e-16	<2.2e-16	<2e-16

Overall, these results suggest Random Forest models timely graduation predictions most reliably because it balances specificity and sensitivity predicting accurately well. SVM as well as Naïve Bayes, conversely, can be suitable choices that are for specific purposes. These purposes include maximizing timely graduation detection or identifying at-risk students.

3.4. Model Performance Evaluation

The evaluation of the performance of the timely graduation prediction model compared three classification algorithms commonly used in machine learning: Support Vector Machine (SVM), Random Forest (RF), along with Naïve Bayes (NB). Researchers applied these three models onto a dataset researchers divided into training data and test data composed of 75% and 25%. For each model, classifying students in a timely manner based upon graduation status was the main purpose with this comparison. Receiver Operating Characteristic or ROC curve analysis was used for carrying out performance evaluation. The Area Under the Curve (AUC) calculation served as the main indicator of success.

The ROC curve is an evaluation tool that can evaluate the model's ability for distinguishing between those two target classes as it displays that sensitivity value otherwise known as true positive rate against 1 – specificity otherwise known as false positive rate. The Random Forest curve (Fig. 3) is consistently above the SVM together with Naïve Bayes curves in the ROC graph. It happens in most specificity areas. Random Forest identifies students for graduation on time more consistently than other models as this indicates. The SVM curve exists just slightly below RF, and also the Naïve Bayes curve exists at the bottom because it indicates the weakest class separation performance among the three.

Furthermore, the AUC value measures classification quality for each model (Table 7). AUC serves as an integral measure of the ROC curve showing a value range from 0 up to 1. Values near to 1 indicate superb classification performance, while values near to 0.5 indicate performance equivalent to random guessing. Based on the calculations, Random Forest has the highest AUC value of 0.7021. This value falls into the moderate to good category, meaning that this model is quite reliable in distinguishing students who graduate on time from those who do not. SVM produced an AUC of 0.6762, also in the moderate category, though slightly lower than RF. Meanwhile, Naïve Bayes showed an AUC value of 0.667, making it the model with the lowest classification performance among the three, though still better than random guessing.

These performance differences can be explained by the characteristics of each method. Random Forest, as a decision tree-based ensemble method, excels at handling both categorical and numerical

predictor variables and has an internal mechanism to reduce overfitting [23], [24]. SVM works by forming an optimal hyperplane that separates classes, but its performance is highly dependent on kernel selection and parameter tuning. Therefore, optimizing kernel parameters, such as penalty values and gamma, is crucial for maximizing classification performance [25]. Naïve Bayes, although simple and fast, assumes that all features are conditionally independent, an assumption that is often unrealistic in real-world data such as complex educational data [26].

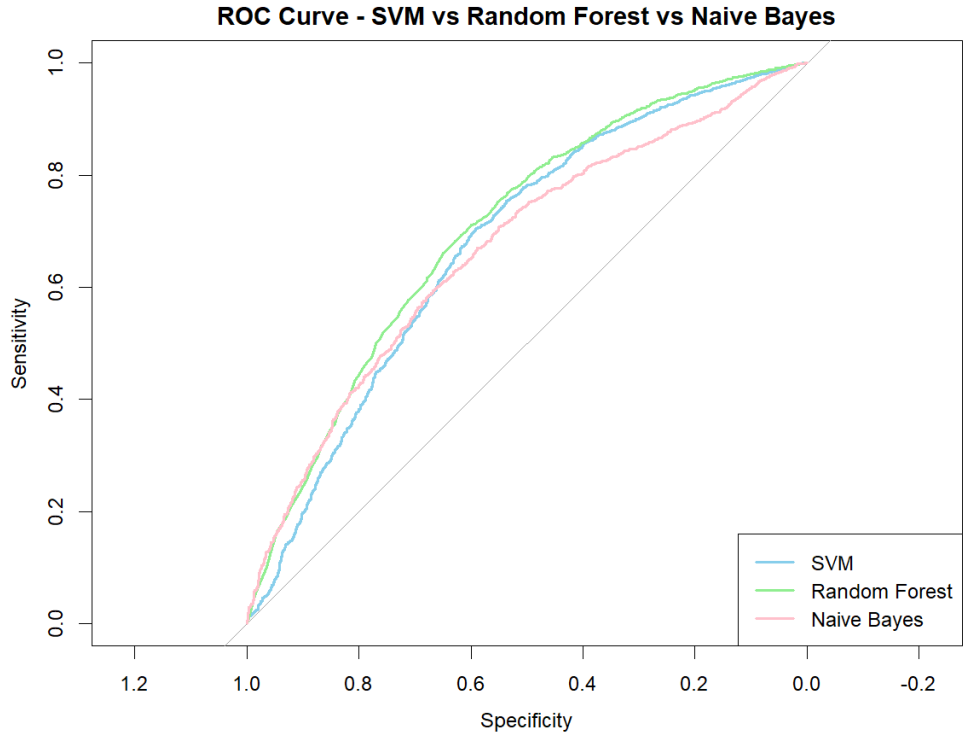

Fig. 3. Comparative ROC Analysis of ML Models

Table 7. ROC-AUC Values for Classification Models

	SVM	Random Forest	Naïve Bayes
Area Under the Curve (AUC)	0.6762	0.7021	0.667

Considering both the ROC curve as well as the AUC value, one can conclude that the Random Forest model is the most optimal algorithm to use for predicting timely graduation in this data. This model provides not only the most stable consistent classification results but also shows the highest AUC score since it indicates effectiveness in recognizing important patterns in the data. Meanwhile, SVM, with its moderate performance, remains as a viable alternative, and Naïve Bayes is more suitable as a baseline model or it is in conditions with computational limitations. According to this study, Random Forest has the best performance, which the highest AUC value and most stable ROC curve indicate versus the other two models. Earlier research supports these conclusions. For example, studies done by [27] and also [28] both found Random Forest generalized the best when predicting student academic performance.

The advantages of Random Forest in this study do align with algorithmic strengths that are identified in various literature sources. As [29], [30], together with [31] explained, this algorithm is accurate so strong when it deals with complex, high-dimensional datasets that contain missing values. Random Forest is known for effectiveness in handling data with many features also automatically selecting influential features to reduce dataset complexity without accuracy loss. In this context, predicting graduation in a timely way is highly relevant. It is influenced by numerous

multidimensional variables like GPA, credit hours, student type, admission pathway, and faculty. Furthermore, this algorithm can flexibly handle categorical data and numerical data without complex pre-processing, as [32] and [33] found.

For Random Forest, handling imbalanced data is a meaningful advantage. This advantage is relevant in this research due to the distribution between the students. Random Forest, compared to other models such as SVM as well as Naïve Bayes, resists overfitting more, especially when scenarios have many features and data vary highly. This characteristic does crucially maintain stable model performance across various dataset conditions, as [27], [28], and [30] also highlight in their studies. Thus, both theoretically and empirically across studies, the selection of Random Forest as the best model in this research is not only methodologically valid but also consistent with the current direction of research in machine learning for academic performance prediction. While ML models offer valuable insights, their deployment in educational policy must consider fairness, interpretability, and data privacy to prevent bias or unintended consequences in student evaluation.

4. Conclusion

This study was carried out to address the ongoing issue of delayed graduation in Indonesian higher education by developing a machine learning-based predictive model. The research emphasizes the critical importance of timely graduation as a benchmark of educational effectiveness and aligns with Sustainable Development Goal (SDG) 4.3, which promotes inclusive access to quality tertiary education. By integrating logistic regression with three machine learning algorithms (Random Forest, Support Vector Machine (SVM), and Naïve Bayes) the study offers a hybrid modeling approach that balances predictive accuracy with interpretability. Among these, Random Forest achieved the most reliable results with the highest accuracy (75.13%) and AUC (0.7021), making it the most suitable model for practical institutional use. Feature importance analysis revealed that GPA, faculty affiliation, and total credits are the most influential variables, reinforcing the role of academic performance and institutional factors in determining graduation outcomes.

The findings validate the proposed hybrid framework and suggest that machine learning, when combined with statistical modeling, can offer actionable insights for academic policy. Logistic regression confirmed the significance of variables such as GPA, admission path, and student type, supporting data-driven decisions in academic advising. Future research should consider integrating these predictive models into real-time student information systems, allowing for longitudinal tracking and early intervention. Moreover, expanding the dataset across multiple institutions could enhance model generalizability. This research thus contributes both methodologically and practically to the fields of educational data mining and institutional performance evaluation, in alignment with global educational equity targets.

Declarations

Funding: This research received no external funding

Conflicts of Interest: The authors declare no conflict of interest.

References

[1] J. Kim, “Student Perspectives on Barriers to Timely Graduation,” *Int. Res. Educ.*, vol. 10, no. 1, p. 35, May 2022, https://doi.org/10.5296/ire.v10i1.19876.

[2] Y. Wu, “Sustainablity in Higher Education: Strategies, Performance and Future Challenges,” *Adv. Educ. Res. Eval.*, vol. 5, no. 1, pp. 264–266, Nov. 2024, https://doi.org/10.25082/AERE.2024.01.002.

[3] Z. Kilasonia, “Higher education and the Sustainable Development Goals,” *DAVID AGHMASHENEBELI Univ. Georg. Sci. J. „SPECTRI“*, Mar. 2023, https://doi.org/10.52340/spectri.2023.15.

[4] K. Okoye, J. T. Nganji, J. Escamilla, and S. Hosseini, “Machine learning model (RG-DMML) and

ensemble algorithm for prediction of students' retention and graduation in education," *Comput. Educ. Artif. Intell.*, vol. 6, p. 100205, Jun. 2024, https://doi.org/10.1016/j.caeai.2024.100205.

[5] M. A. S. Pawitra, H.-C. Hung, and H. Jati, "A Machine Learning Approach to Predicting On-Time Graduation in Indonesian Higher Education," *Elinvo (Electronics, Informatics, Vocat. Educ.*, vol. 9, no. 2, pp. 294–308, Dec. 2024, https://doi.org/10.21831/elinvo.v9i2.77052.

[6] A. Santoso, H. Retnawati, Kartianom, E. Apino, I. Rafi, and M. N. Rosyada, "Predicting Time to Graduation of Open University Students: An Educational Data Mining Study," *Open Educ. Stud.*, vol. 6, no. 1, Feb. 2024, https://doi.org/10.1515/edu-2022-0220.

[7] A. Desfiandi and B. Soewito, "Student Graduation Time Prediction Using Logistic Regression, Decision Tree, Support Vector Machine, And Adaboost Ensemble Learning," *IJISCS (International J. Inf. Syst. Comput. Sci.*, vol. 7, no. 3, p. 195, Oct. 2023, https://doi.org/10.56327/ijiscs.v7i2.1579.

[8] L. R. Pelima, Y. Sukmana, and Y. Rosmansyah, "Predicting University Student Graduation Using Academic Performance and Machine Learning: A Systematic Literature Review," *IEEE Access*, vol. 12, pp. 23451–23465, 2024, https://doi.org/10.1109/ACCESS.2024.3361479.

[9] A. Rahman, D. Mahdiana, and A. Fauzi, "Predicting Student On-Time Graduation Using Particle Swarm Optimization and Random Forest Algorithms," *Indones. J. Artif. Intell. Data Min.*, vol. 8, no. 1, 2025, https://doi.org/10.24014/ijaidm.v8i1.33577.

[10] A. Artyukhov, T. Wołowiec, N. Artyukhova, S. Bogacki, and T. Vasylieva, "SDG 4, Academic Integrity and Artificial Intelligence: Clash or Win-Win Cooperation?," *Sustainability*, vol. 16, no. 19, p. 8483, Sep. 2024, https://doi.org/10.3390/su16198483.

[11] M. Nachouki, E. A. Mohamed, R. Mehdi, and M. Abou Naaj, "Student course grade prediction using the random forest algorithm: Analysis of predictors' importance," *Trends Neurosci. Educ.*, vol. 33, p. 100214, 2023, https://doi.org/10.1016/j.tine.2023.100214.

[12] M. Bansal, A. Goyal, and A. Choudhary, "A comparative analysis of K-Nearest Neighbor, Genetic, Support Vector Machine, Decision Tree, and Long Short Term Memory algorithms in machine learning," *Decis. Anal. J.*, vol. 3, no. May, p. 100071, 2022, https://doi.org/10.1016/j.dajour.2022.100071.

[13] Y. Gu, "Exploring the application of teaching evaluation models incorporating association rules and weighted naive Bayesian algorithms," *Intell. Syst. with Appl.*, vol. 20, no. November 2022, p. 200297, 2023, https://doi.org/10.1016/j.iswa.2023.200297.

[14] H. Al Sagri and M. Ykhlef, "Quantifying Feature Importance for Detecting Depression using Random Forest," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 5, 2020, https://doi.org/10.14569/IJACSA.2020.0110577.

[15] R. Bakri, N. P. Astuti, and A. S. Ahmar, "Machine Learning Algorithms with Parameter Tuning to Predict Students' Graduation-on-time: A Case Study in Higher Education," *J. Appl. Sci. Eng. Technol. Educ.*, vol. 4, no. 2, pp. 259–265, Dec. 2022, https://doi.org/10.35877/454RI.asci1581.

[16] M. R. Julianti, Y. Heryadi, B. Yulianto, and W. Budiharto, "Performance Graduation Student Predicting Using One-Class Support Vector Machine Algorithm," *Int. J. Intell. Syst. Appl. Eng.*, vol. 2024, no. 4, 2024, https://ijisae.org/index.php/IJISAE/article/view/6208.

[17] C. Dewi, G. E. Laukon, H. J. Christanto, and S. A. Sutresno, "Modification of random forest method to predict student graduation data," *Mantik J.*, vol. 7, no. 4, 2024, https://iocscience.org/ejournal/index.php/mantik/article/view/4528.

[18] K. Chenary, O. Pirian Kalat, and A. Sharifi, "Forecasting sustainable development goals scores by 2030 using machine learning models," *Sustain. Dev.*, vol. 32, no. 6, pp. 6520–6538, Dec. 2024, https://doi.org/10.1002/sd.3037.

[19] S. Theodoridis, *Machine Learning: A Bayesian and Optimization Perspective, Second Edition*. Elsevier Ltd, 2020, https://doi.org/10.1016/C2019-0-03772-7.

[20] I. Kononenko and M. Kukar, *Machine learning and data mining: introduction to principles and algorithms*. Horwood Publishing, 2008, https://doi.org/10.5860/CHOICE.45-3834.

[21] B. Lantz, *Machine learning with R*. Packt Publishing, 2013,

https://books.google.co.id/books?id=iNuSDwAAQBAJ.

[22] Y. Zhang, *New Advances in Machine Learning*. InTech, 2012, https://doi.org/10.5772/225.

[23] E. Halabaku and E. Bytyçi, "Overfitting in Machine Learning: A Comparative Analysis of Decision Trees and Random Forests," *Intell. Autom. Soft Comput.*, vol. 39, no. 6, pp. 987–1006, 2024, https://doi.org/10.32604/iasc.2024.059429.

[24] A. Testas, "Random Forest Classification with Scikit-Learn and PySpark BT - Distributed Machine Learning with PySpark," Apress, 2023, https://doi.org/10.1007/978-1-4842-9751-3_9.

[25] V. S. Sahithi, I. V. M. Krishna, and M. V. S. S. Giridhar, "Analysing the Sensitivity of SVM Kernels on Hyperspectral Imagery for Land Use Land Cover Classification," *J. Image Process. Artif. Intell.*, vol. 8, no. 2, pp. 15–23, Jun. 2022, https://doi.org/10.46610/JOIPAI.2022.v08i02.003.

[26] R. Kumar, B. Krishna Goswami, S. Motiram Mhatre, and S. Agrawal, "Naive Bayes in Focus: A Thorough Examination of its Algorithmic Foundations and Use Cases," *Int. J. Innov. Sci. Res. Technol.*, pp. 2078–2081, Jun. 2024, https://doi.org/10.38124/ijisrt/IJISRT24MAY1438.

[27] Y. Chen and L. Zhai, "A comparative study on student performance prediction using machine learning," *Educ. Inf. Technol.*, vol. 28, no. 9, pp. 12039–12057, Sep. 2023, https://doi.org/10.1007/s10639-023-11672-1.

[28] E. Ismanto, H. A. Ghani, and N. I. B. Md Saleh, "A comparative study of machine learning algorithms for virtual learning environment performance prediction," *IAES Int. J. Artif. Intell.*, vol. 12, no. 4, p. 1677, Dec. 2023, https://doi.org/10.11591/ijai.v12.i4.pp1677-1686.

[29] M. D. Laddha, V. T. Lokare, A. W. Kiwelekar, and L. D. Netak, "Performance Analysis of the Impact of Technical Skills on Employability," *Int. J. Performability Eng.*, vol. 17, no. 4, p. 371, 2021, https://doi.org/10.23940/ijpe.21.04.p5.371378.

[30] G. Ben Brahim, "Predicting Student Performance from Online Engagement Activities Using Novel Statistical Features," *Arab. J. Sci. Eng.*, vol. 47, no. 8, pp. 10225–10243, Aug. 2022, https://doi.org/10.1007/s13369-021-06548-w.

[31] M. M. Tamada, R. Giusti, and J. F. de M. Netto, "Predicting Students at Risk of Dropout in Technical Course Using LMS Logs," *Electronics*, vol. 11, no. 3, p. 468, Feb. 2022, https://doi.org/10.3390/electronics11030468.

[32] M. V. Martins, L. Baptista, J. Machado, and V. Realinho, "Multi-Class Phased Prediction of Academic Performance and Dropout in Higher Education," *Appl. Sci.*, vol. 13, no. 8, p. 4702, Apr. 2023, https://doi.org/10.3390/app13084702.

[33] M. A. Arif, A. Jahan, M. I. Mau, and R. Tummarzia, "An Improved Prediction System of Students' Performance Using Classification model and Feature Selection Algorithm," *Int. J. Adv. Soft Comput. Appl.*, vol. 13, no. 1, 2021, https://www.i-csrs.org/Volumes/ijasca/2021.1.10.pdf.