Technology and Society Perspectives (TACIT) Vol. 4, No. 2, July-October 2026, pages 683–703

Predicting On-Time Graduation in Indonesian Private Higher Education: A Machine Learning Comparison for Institutional Decision Support

Achmad Zakki Falani[1*], Syaad Patmanthara[2], Didik Dwi Prasetya[3]

[1,2,3,4]*Universitas Negeri Malang, Jawa Timur, Indonesia*

ARTICLE INFO

Article history:
Received: 18 September 2026
Revised: 16 October 2026
Accepted: 20 October 2026

DOI:
10.61100/tacit.v4i2.386

Keywords:
Graduation Prediction, Educational Data Mining, Algorithmic Accountability, Random Forest, Higher Education Governance

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License

ABSTRACT

Timely graduation shapes accreditation and resource decisions in Indonesia's private higher education, so how predictive models are built and used is a social as much as a technical question. Prior machine learning comparisons in this setting rely on overall accuracy, run under differing conditions, and rarely test how well models detect late-graduating students or inform institutional action. This study compares Random Forest (RF) and K-Nearest Neighbor (KNN) for on-time graduation prediction on 656 student records (2018–2022) under identical conditions, using accuracy, precision, recall, F1-score, 10-fold cross-validation, feature importance (MDI), and the Wilcoxon Signed-Rank Test with a rank-biserial effect size. RF reached higher accuracy (80.30% versus 70.45%), but both models detected the minority late-graduating class poorly (recall 0.40 and 0.26), and the difference was not statistically significant ($p = 0.3477$; $r \approx 0.49$). Academic attributes accounted for 72.38% of feature importance. These results position such models as accountable decision support rather than a validated early-warning intervention, which would require evaluation in real academic governance.

ABSTRAK

Kelulusan tepat waktu turut menentukan akreditasi dan keputusan sumber daya di perguruan tinggi swasta Indonesia, sehingga cara model prediksi dibangun dan digunakan menjadi persoalan sosial sekaligus teknis. Perbandingan pembelajaran mesin pada konteks ini umumnya bertumpu pada akurasi keseluruhan, dijalankan pada kondisi yang berbeda-beda, dan jarang menguji seberapa baik model mendeteksi mahasiswa yang terlambat lulus atau menuntun tindakan institusi. Penelitian ini membandingkan Random Forest (RF) dan K-Nearest Neighbor (KNN) untuk prediksi kelulusan tepat waktu pada 656 rekam data mahasiswa (2018–2022) dalam kondisi identik, menggunakan akurasi, presisi, recall, F1-score, 10-fold cross-validation, feature importance (MDI), dan Wilcoxon Signed-Rank Test dengan ukuran efek rank-biserial. RF mencapai akurasi lebih tinggi (80,30% dibanding 70,45%), tetapi kedua model lemah mendeteksi kelas minoritas mahasiswa terlambat lulus (recall 0,40 dan 0,26), dan perbedaannya tidak signifikan secara statistik ($p = 0,3477$; $r \approx 0,49$). Atribut akademik menyumbang 72,38% feature importance. Hasil ini menempatkan model semacam ini sebagai alat bantu keputusan yang akuntabel, bukan intervensi peringatan dini yang tervalidasi, yang masih memerlukan evaluasi dalam tata kelola akademik nyata.

1. INTRODUCTION

Timely student graduation is an institutional performance indicator that directly affects higher education accreditation scores, as stipulated in Ministerial Regulation of the Ministry of Higher Education, Science, and Technology Number 39 of 2025 (Kementerian Pendidikan Tinggi Sains dan Teknologi, 2025). In the context of private higher education institutions (PHEIs), this dynamic presents more complex challenges. The heterogeneity of PHEI student profiles, ranging from age at matriculation and admission mechanisms to the flexibility of study schedules, contributes to the emergence of multidimensional clusters of graduation-delay risk (Santoso et al., 2024). Interventions initiated only near the end of the maximum study period tend to be

** Corresponding author, email address: achmad.zakki.2605349@students.um.ac.id*

suboptimal because students have limited opportunities to recover their academic performance, while institutional capacity to explore strategic solutions is also constrained (Psyridou et al., 2024). This condition highlights the urgent need for prediction systems capable of identifying risk signals from the early semesters and translating them into concrete intervention recommendations. The scale of this problem is tangible at the institutional level: among the 656 students examined in this study from the 2018–2022 cohorts, 174 (26.52%) did not graduate on time, indicating that delayed graduation is a routine rather than marginal outcome in this setting.

Beyond its methodological dimension, the deployment of machine-learning-based prediction in higher education is embedded in a broader technology-society relationship that is particularly salient in the Indonesian private higher education sector. Private higher education institutions operate under uneven resource conditions and sustained accreditation pressure, so the adoption of data-driven prediction is not merely a technical upgrade but an institutional and social process that reshapes how at-risk students are identified, prioritized, and supported. Because such systems mediate consequential academic decisions, their design raises questions of institutional responsibility, transparency, and fairness that extend beyond predictive accuracy. Situating graduation prediction within this technology-society perspective clarifies that the value of a predictive model depends not only on its statistical performance but also on how its outputs are interpreted and acted upon within institutional decision-making structures.

Educational Data Mining (EDM) approaches based on machine learning have demonstrated effectiveness in analyzing historical academic data to support institutional decision-making (Shafiq et al., 2022), with Random Forest (RF) and K-Nearest Neighbor (KNN) reported as high-performing algorithms for academic performance prediction tasks (Albreiki et al., 2021). Santoso et al. demonstrated that academic attributes, including cumulative GPA, semester grades, and study status, are closely associated with the time required for students to graduate in Indonesia (Santoso et al., 2024). These findings raise an important unresolved question: to what extent does each attribute contribute across semesters, and when can the earliest reliable risk signals be identified? RF was selected as the primary algorithm because of its ability to generate feature importance measures that can help translate predictive results into evidence-based intervention recommendations (Breiman, 2001; Yağcı, 2022; Nachouki et al., 2023). KNN was selected as the comparison algorithm because it employs a fundamentally different approach based on instance-level similarity without constructing an explicit structural model. Consequently, measurable differences in performance can more directly reflect algorithmic characteristics rather than bias arising from inconsistent experimental conditions (Dawar et al., 2024; Kabathova & Drlik, 2021).

The literature review reveals three methodological gaps that have not been adequately addressed. First, most studies on student graduation prediction in Indonesia, whether employing KNN (Dawar et al., 2024; Wati et al., 2023) or RF (Pelima et al., 2024), still rely primarily on accuracy as the main evaluation metric. However, in datasets with imbalanced class distributions, accuracy alone can produce misleading assessments because it does not adequately reflect a model's ability to accurately detect at-risk students (Cruz & Lumauag, 2025; Thabtah et al., 2020). Second, the validity of algorithmic comparisons remains a concern. Comparisons between RF and KNN in Indonesian EDM literature are generally conducted under different experimental conditions, including differences in datasets, preprocessing pipelines, and validation schemes. Consequently, reported claims of algorithmic superiority are difficult to disentangle from the effects of these differing experimental conditions (Yağcı, 2022; Kabathova & Drlik, 2021). Without consistent experimental conditions, comparative conclusions cannot serve as a reliable reference. From an institutional perspective, the most fundamental gap lies in the absence of feature importance analysis that examines the contribution of academic attributes across semesters. Existing studies generally report global attribute rankings (Nachouki et al., 2023; Pelima et al., 2024), but do not indicate when and to what extent risk signals emerge at each semester. Such information is needed by institutions to provide an operational basis for designing targeted early intervention programs (Psyridou et al., 2024; Zanellati et al., 2024).

This study is grounded in three philosophical foundations of science. Ontologically, on-time student graduation is regarded as an academic phenomenon that exists objectively and can be observed and represented through measurable student characteristics and institutional academic records. Epistemologically, knowledge about the patterns underlying graduation is constructed empirically from these records through an Educational Data Mining approach using Random Forest and K-Nearest Neighbor, and is justified through multi-metric evaluation, 10-Fold Cross-Validation, and statistical testing rather than through a priori reasoning. Axiologically, the resulting predictive knowledge is oriented toward practical and ethical value

for higher education institutions, supporting the early identification of students at risk of delayed graduation and the design of fair and responsible academic interventions. This study therefore aims not only to develop a predictive model but also to generate empirical knowledge that can serve as a basis for data-driven and value-conscious academic decision-making.

This study presents three methodological contributions that, to the best of our knowledge, have not previously been combined within a single student graduation prediction study in the Indonesian PHEI context: (1) a comparison of RF and KNN under fully identical experimental conditions, such that differences in performance primarily reflect algorithmic characteristics; (2) MDI-based feature importance analysis that examines attribute contributions across semesters to construct a progressive map of predictive signals for guiding early intervention; and (3) statistical validation using the Wilcoxon Signed-Rank Test, an approach that remains relatively uncommon in the Indonesian EDM literature, to ensure that comparative conclusions are supported by defensible statistical evidence.

These contributions apply established techniques rather than propose new algorithms; their novelty lies in their combination and context. Prior Indonesian EDM studies typically compare RF and KNN under non-identical conditions, report only global feature rankings, and rarely test comparative claims statistically. By holding experimental conditions constant, resolving feature importance at the semester level, and adding inferential testing with an effect size, this study contributes context-specific empirical evidence on when and how far academic signals matter in an Indonesian private higher education institution, rather than a methodological novelty in the algorithms themselves.

To address these three research gaps, this study aims to: (1) compare the performance of RF and KNN in predicting on-time student graduation using comprehensive multi-metric evaluation under identical experimental conditions; (2) examine the relative importance of academic attributes across semesters through feature importance analysis to characterize the distribution of predictive signals within the full feature set; and (3) analyze the performance characteristics and interpretability of RF and KNN as a basis for understanding their applicability to academic decision-support systems in Indonesian PHEIs. The study uses a real-world dataset comprising 656 academic records of students from Universitas Narotama from the 2018–2022 cohorts, consisting of 13 academic and non-academic predictor attributes. It should be emphasized that this study characterizes the relative importance of semester-level attributes within a model trained on the full feature set; it does not train or evaluate a model restricted to early-semester information. References to early risk signals therefore denote which attributes carry predictive weight, not a demonstrated capacity for early-semester prediction, which remains a direction for future work.

2. THEORETICAL FRAMEWORK AND HYPOTHESES

2.1 Literature Review

Graduation prediction has become an important application of Educational Data Mining (EDM) in higher education, enabling universities to identify academic patterns and assess factors associated with students' likelihood of completing their studies on time. By analyzing academic records, cumulative grade point averages, completed credits, course performance, attendance, and study duration, educational institutions can develop predictive models to identify students who may be at risk of delayed graduation. Among various machine learning techniques, the Random Forest algorithm offers a promising approach because it combines multiple decision trees to improve predictive stability and accommodate complex relationships among academic variables. Previous studies have demonstrated the potential of educational data mining and machine learning to support academic performance prediction and the early identification of at-risk students (Delen, 2010; Alyahyan & Düştegör, 2020). These predictive capabilities can help universities strengthen academic monitoring, develop targeted student support programs, and improve evidence-based decision-making.

However, the implementation of graduation prediction systems should extend beyond predictive accuracy to address the ethical and institutional implications of algorithm-driven decisions. The concept of Algorithmic Accountability emphasizes the importance of transparency, fairness, explainability, and institutional responsibility when machine learning models are used to evaluate student outcomes. Although Random Forest can identify influential predictors of graduation, its results may reflect historical inequalities or biases embedded in educational datasets. Consequently, predictive outcomes should serve as decision-support information rather than definitive judgments about students' academic potential. Within the framework of Higher Education Governance, integrating algorithmic accountability into educational data mining practices can strengthen institutional oversight, protect student information, and promote responsible data-

driven policies. As emphasized in the literature on learning analytics ethics and algorithmic accountability (Slade & Prinsloo, 2013; Kroll et al., 2017), responsible predictive systems require clear governance mechanisms, appropriate human oversight, and safeguards against unfair outcomes. Therefore, the integration of graduation prediction, Random Forest, and algorithmic accountability provides a foundation for improving academic management while ensuring that technological innovation remains aligned with educational equity, institutional responsibility, and sustainable higher education governance.

3. RESEARCH METHOD

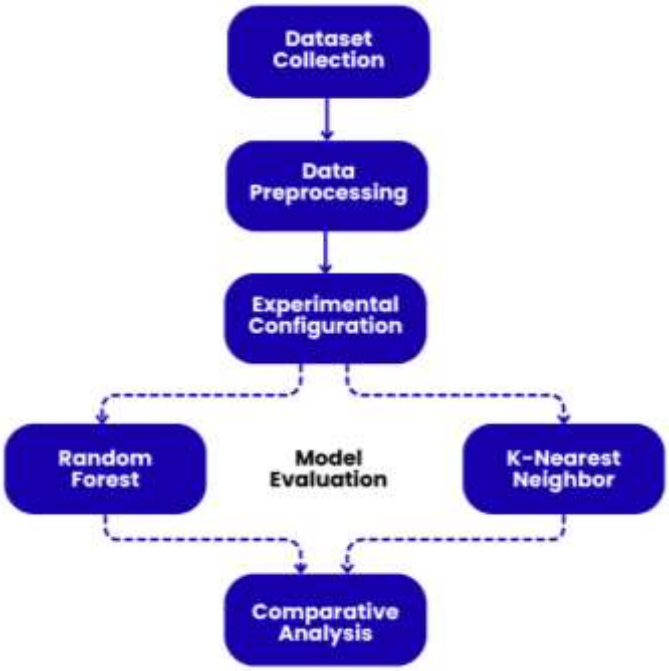


Figure 1. Research Stages

This study consists of several stages, namely data collection, data preprocessing, experimental configuration, application of the Random Forest and K-Nearest Neighbor (KNN) algorithms, model performance evaluation, and comparative analysis. The research stages were conducted systematically to obtain a student graduation prediction model with optimal classification performance.

From a philosophy-of-science perspective, the methodology of this study operationalizes its three foundational stances. Ontologically, the object of study is represented as a structured dataset of measurable academic and non-academic attributes. Epistemologically, predictive knowledge is derived from these data through a systematic and reproducible pipeline of preprocessing, model construction, and validation, so that the resulting claims rest on observable evidence. Axiologically, the methodological choices, including the use of an interpretable Random Forest model and the processing of de-identified data, are directed toward producing knowledge that is both practically useful and ethically responsible for institutional decision-making.

2.1 Research Dataset

This study uses academic data of students from Universitas Narotama obtained from the university's academic information system. The use of academic records obtained from higher education information systems is a common approach in Educational Data Mining (EDM) research for developing relevant and representative graduation prediction models (Santoso et al., 2024). The dataset consists of 656 student records from the 2018–2022 cohorts.

All students included in the dataset had completed at least four semesters of study at the time of data collection; therefore, Semester 1 through Semester 4 GPA values were available in full for all 656 academic records. Students who did not meet this criterion were excluded from the dataset during the data-cleaning stage to maintain the integrity of the predictor attributes and avoid bias arising from incomplete data.

The research dataset consists of several academic and non-academic attributes related to student

academic performance. The predictor variables include gender, Semester 1–4 GPA, age, unit code, cohort year, admission pathway, study system, cumulative GPA, completed credits, leave-of-absence status, and student graduation status as the target variable.

The academic records were obtained from the university's academic information system for research purposes and were processed in de-identified form, with personal identity attributes such as student names and identification numbers excluded from the modeling pipeline.

The predictor attributes differ in their temporal reference. Semester 1 to Semester 4 GPA are observed by the end of the fourth semester, whereas cumulative GPA, total earned credits, and leave-of-absence status reflect the student's record accumulated across the enrollment period as extracted from the academic information system. Because these accumulated attributes are not strictly separated in time from the graduation outcome, the models are best interpreted as a retrospective characterization of graduation-associated patterns rather than a prospective early-warning system. This temporal structure is a limitation of the present design, and future work should construct temporally gated predictors measured at a fixed early checkpoint to support genuinely prospective prediction.

Table 1. Description of Dataset Attributes

No.	Variable Category	Attribute	Data Type
1.	Student Characteristics	Gender, Age, Cohort Year, Admission Pathway	Categorical / Numerical
2.	Academic & Study Performance	Semester 1–4 GPA, Cumulative GPA, Completed Credits, Study System, Leave of Absence	Categorical / Numerical
3.	Administration	Study Program Unit Code	Numerical
4.	Target Variable	Graduation Status	Categorical

The target variable in this study consists of two classes: students who graduate on time (1) and students who do not graduate on time (0).

Table 2. Distribution of Student Graduation Target Classes

Target Class	Description	Row Count	Percentage (%)
1	On-Time Graduation	482	73.48
0	Delayed Graduation	174	26.52
Total		**656**	**100.00**

The class distribution in Table 2 indicates an imbalanced dataset, with a ratio of 73.48% to 26.52%. Therefore, model evaluation employs precision, recall, and F1-score in addition to accuracy (Thabtah et al., 2020).

2.2 Data Preprocessing

Data preprocessing was conducted to improve data quality before the dataset was used for machine learning model training. This stage aimed to reduce data inconsistencies and improve classification model performance (Dawar et al., 2024). Systematic preprocessing is a crucial step that directly affects the quality and generalization capability of the resulting models.

Based on the preprocessing results, the research dataset consists of 656 student records with 13 attributes comprising academic and non-academic variables. The preprocessing stages employed in this study are presented in Figure 2.

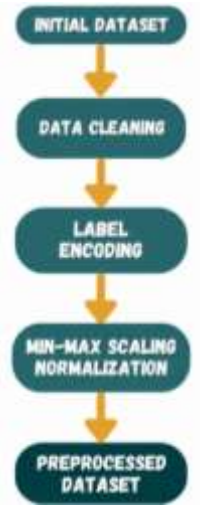


Figure 2. Data Preprocessing Stages

2.2.1 Data Cleaning

Data cleaning was performed to address incomplete data, duplicate records, and inconsistencies in the student academic dataset. This stage was conducted to ensure that the data used for model training were of adequate quality (Dawar et al., 2024). Based on dataset inspection, no missing values or duplicate records were identified; consequently, all 656 records were retained for the analysis. In addition, consistency checks were performed on each academic attribute, including semester GPA, cumulative GPA, completed credits, and leave-of-absence status, to ensure that no values fell outside their expected ranges.

2.2.2 Label Encoding

Categorical attributes were transformed into numerical representations using an encoding procedure within the preprocessing pipeline. In this study, the categorical predictor attributes, including gender and study system, were encoded before being used by the classification algorithms. The encoder was fitted using the training data and subsequently applied to the testing data to prevent information leakage. Unknown categories encountered during transformation were handled by the encoder configuration. The graduation target was represented as a binary class, with On-Time Graduation = 1 and Not On-Time Graduation = 0.

2.2.3 Data Normalization

Data normalization was performed using Min-Max Scaling to place numerical attributes within a comparable numerical range, particularly for the KNN classification process. Min-Max Scaling was applied within the KNN preprocessing pipeline because KNN calculates Euclidean distances between observations and is therefore sensitive to differences in feature scales (Ahmed et al., 2022). The scaling transformation was fitted using the training data only and subsequently applied to the testing data. In contrast, Random Forest does not require feature normalization because decision-tree splitting is based on threshold comparisons and is not affected by the scale of numerical features (Breiman, 2001).

Min-Max Scaling transforms the data range to between 0 and 1, preventing attributes with larger numerical scales from dominating the classification process due to differences in measurement scales. The Min-Max Scaling formula is presented in Equation (1).

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}} \quad \ldots\ldots\ldots\ldots \quad (1)$$

where:
X=original data value,
Xmin=minimum value of the attribute,
Xmax=maximum value of the attribute, and
X'=normalized value.

2.2.4 Data Splitting

Before model training, the dataset was first divided into training and testing sets using an 80:20 stratified split. The preprocessing transformations were then fitted using the training data only and subsequently applied to the testing data. This procedure was implemented to prevent information from the testing data from influencing the preprocessing stage and model development. Given the imbalanced class distribution of 73.48% versus 26.52%, stratified random sampling was employed using the stratify parameter of the train_test_split function in Scikit-Learn. This approach maintains a similar class distribution in both the training and testing datasets. The data-splitting process resulted in 524 training records and 132 testing records.

Table 3. Data Split Distribution

Data Type	Total	Class 1 (Graduated)	Class 0 (Not Graduated)
Training Data	524	385 (73.48%)	139 (26.52%)
Testing Data	132	97 (73.48%)	35 (26.52%)
Total Data	656	482 (73.48%)	174 (26.52%)

The training data were used to train the Random Forest and K-Nearest Neighbor models, whereas the testing data were used to evaluate the performance of the resulting classification models. In addition, 10-Fold Cross-Validation was conducted as an additional evaluation procedure to assess model performance consistency across different data-partitioning scenarios.

2.3 Experimental Configuration

This study uses the Python programming language with the Scikit-Learn library in the Google Colaboratory environment. The use of Python and Scikit-Learn is common in comparative machine learning studies within the EDM context because of their support for reproducible experimentation and consistent implementation (Dawar et al., 2024). All preprocessing, model training, and evaluation procedures were implemented using Python-based machine learning libraries.

Table 4. Experimental Configuration

Component	Configuration
Programming Language	Python
Platform	Google Colaboratory
Library	Scikit-Learn
Data Split	Stratified 80:20
Cross-Validation	10-Fold Cross-Validation
Random Forest (n_estimators)	100
KNN (n_neighbors)	5
Distance Metric	Euclidean
Random State	42

The Random Forest algorithm was implemented using the Scikit-Learn library with n_estimators=100, criterion='gini', and random_state =42 to ensure reproducibility. Meanwhile, the K-Nearest Neighbor algorithm was configured with n_neighbors = 5 and the Euclidean distance metric, as justified in Section 2.5.

2.4 Random Forest Algorithm

Random Forest (RF) is an ensemble learning method that constructs multiple decision trees independently using bootstrap sampling and subsequently aggregates the predictions of all trees through a majority-voting mechanism (Breiman, 2001). RF was used in this study because it can accommodate heterogeneous attributes comprising academic and non-academic dimensions, while also providing feature importance values that can be used to examine the relative importance of attributes in the classification model.

The RF configuration in this study consists of 100 decision trees (n_estimators = 100), constructed using 524 training records with 13 predictor attributes. The model employs Gini Impurity as the node-splitting criterion (criterion = 'gini') and random_state = 42 to ensure reproducibility (Breiman, 2001). Feature

importance values were obtained using the Mean Decrease in Impurity (MDI) mechanism through the feature_importances_ attribute in Scikit-Learn (Nayak et al., 2023). A higher importance value indicates a greater relative importance of an attribute in the classification model, based on the MDI calculation across the decision trees (Nachouki et al., 2023). The RF process is illustrated in Figure 3.

$$H(x) = \arg\max_y \sum_{i=1}^{n} I\left(h_i(x) = y\right) \quad \ldots\ldots. (2)$$

where:
$H(x)$ = final RF prediction,
$h_i(x)$ = prediction of the *i*-th decision tree,
I = indicator function,
y = target class,
n = number of decision trees.

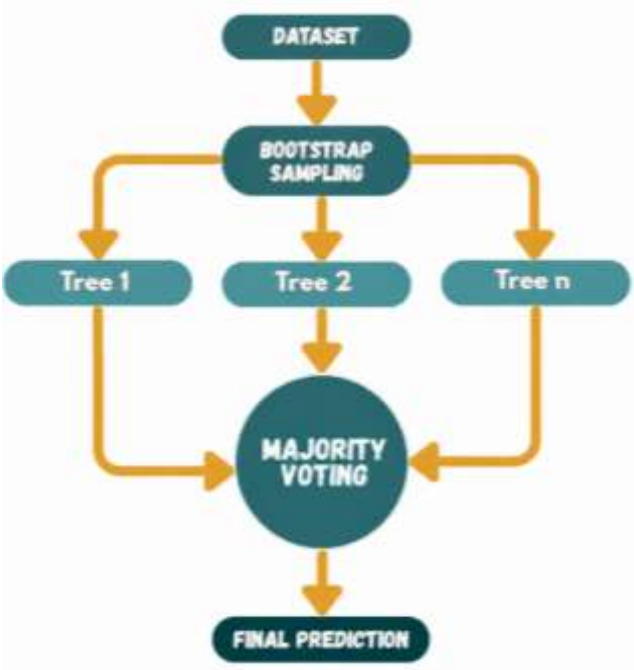


Figure 3. Random Forest Process

2.5 K-Nearest Neighbor Algorithm

In this study, KNN was used as a comparative algorithm to evaluate classification performance based on the similarity of student characteristics with respect to graduation status. Unlike RF, which constructs a structural model, KNN does not construct an explicit parametric model and classifies each testing instance based on its K nearest neighbors in the feature space. The neighbors are determined using Euclidean distance based on the preprocessed training data.

The predicted class is determined by the majority class among the K nearest neighbors in the feature space, making KNN a memory-based algorithm that is highly responsive to the local distribution of the data (Kabathova & Drlik, 2021). In this study, KNN was used as the comparison algorithm because its underlying mechanism differs from that of RF, allowing the performance characteristics of the two algorithms to be compared under the same experimental conditions.

The selection of the K value is a critical stage in KNN configuration because a value that is too small is susceptible to noise, whereas a value that is too large may produce an overly smooth decision boundary in a medium-sized dataset and reduce sensitivity to local patterns.

In this study, K = 5 was adopted as a conventional default configuration for K-Nearest Neighbor in binary classification tasks, chosen a priori as a fixed baseline rather than through data-driven parameter search, in order to keep the experimental setup simple and to allow the comparison with Random Forest to be conducted under fixed and reproducible conditions. The value of K governs the bias-variance behavior of

the classifier: an excessively small K (e.g., K = 3) is more sensitive to noise in the training data, whereas an excessively large K in a medium-sized dataset tends to produce overly smooth decision boundaries and may reduce sensitivity to local patterns. It should be noted that K was not optimized in this study; systematic tuning of K through dataset-driven parameter search, as advocated by Inyang et al. (2023) is left for future work and may further affect KNN performance.

The hyperparameters for both algorithms were set to standard default values rather than tuned: Random Forest used 100 trees with the Gini criterion, and KNN used K = 5 with the Euclidean distance. These fixed, untuned settings were chosen deliberately so that the comparison reflects default-configuration behavior under identical conditions; a consequence is that the reported figures may understate the peak performance each algorithm could reach after dedicated hyperparameter optimization. To keep the comparison fair and reproducible, both algorithms were trained and evaluated on the same stratified 80:20 partition and the same ten cross-validation folds, with random_state fixed at 42. Categorical predictors such as gender and study system were label-encoded, and both the encoder and the Min-Max scaler were fitted on the training partition only and then applied to the test partition. Min-Max scaling was applied for KNN, which relies on distance, but not for Random Forest, whose tree-based splits are invariant to monotonic feature scaling.

The Euclidean distance is mathematically formulated as follows:

$$d(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad \ldots\ldots\ldots (3)$$

where:
$d(x, y)$ = distance between the testing and training data,
x_i = attribute value of the testing data,
y_i = attribute value of the training data, and
n = number of attributes.

The equation is used to calculate the distance between observations, such that data points with the shortest distances are used as the basis for determining the predicted class. The KNN classification process is illustrated in Figure 4.



Figure 4. KNN Algorithm Classification Process

For KNN, the preprocessing transformations were fitted using the training data and subsequently applied to the testing data before classification, as described in Section 2.2.3.

2.6 Model Evaluation

Model evaluation was conducted to compare the performance of the Random Forest and K-Nearest Neighbor algorithms. Classification performance was evaluated using accuracy, precision, recall, F1-score, and the confusion matrix based on predictions generated by each algorithm on the testing data. Precision, recall, and F1-score were calculated for each class and summarized using weighted averages to account for the class distribution. Rainio et al. (2024) noted that using multiple evaluation metrics provides a more comprehensive assessment of binary classification performance, particularly when class distributions are imbalanced.

The equations for the evaluation metrics are as follows:

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \quad \ldots\ldots\ldots (4)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \quad \ldots\ldots\ldots (5)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \quad \ldots\ldots\ldots (6)$$

$$\text{F1} - \text{Score} = \frac{\text{2 x (Precision x Recall)}}{\text{(Precision + Recall)}} \quad \ldots\ldots\ldots (7)$$

where:
TP = True Positive,
TN = True Negative,
FP = False Positive,
FN = False Negative.

In addition to evaluation on the testing data, this study applied 10-Fold Cross-Validation to assess model performance consistency across multiple data partitions. For the 10-Fold Cross-Validation procedure, the dataset was divided into 10 folds, with each fold alternately used as the validation fold while the remaining folds were used for training. The preprocessing pipeline was fitted separately within each training fold and subsequently applied to the corresponding validation fold to prevent information leakage during cross-validation. K-fold cross-validation is used to assess model performance consistency across multiple data partitions and to obtain a more robust estimate of performance than a single data partition (Yağcı, 2022).

The 10-Fold Cross-Validation scheme is shown in Figure 5.

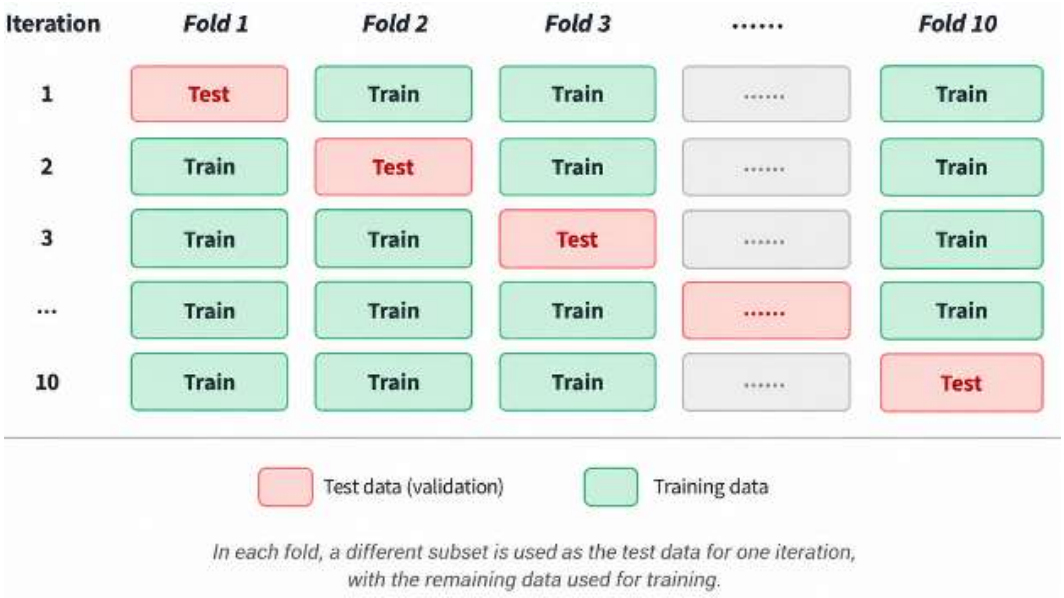


Figure 5. 10-Fold Cross-Validation Scheme

The 10-fold cross-validation accuracy scores for Random Forest and K-Nearest Neighbor were subsequently compared using the Wilcoxon Signed-Rank Test to assess whether the paired fold-level performance differences were statistically significant. The Wilcoxon Signed-Rank Test was selected as a non-parametric test for comparing paired observations without requiring the assumption of normally distributed differences between the paired scores (Rainio et al., 2024).

The Wilcoxon Signed-Rank Test equation used in this study is presented as follows:

$$W = \min(W^+, W^-) \quad \text{……. (8)}$$

where:
W^+ = sum of positive ranks,
W^- = sum of negative ranks, and
W = Wilcoxon test statistic.

The Wilcoxon Signed-Rank Test is used to compare two paired samples based on the ranks of their differences. The final decision was based on a significance level of $\alpha = 0.05$. A p-value < 0.05 indicates a statistically significant difference between the paired performance scores, whereas a p-value ≥ 0.05 indicates that the null hypothesis is not rejected.

Two statistical caveats should be acknowledged. First, the ten paired fold-level accuracies are not fully independent, because the cross-validation training folds overlap and therefore share data; this dependence can underestimate the variance of the differences and reduce the power of the Wilcoxon test, so its non-significant result is interpreted descriptively alongside the effect size rather than as evidence of equivalence. Second, the effect size reported with the test is the matched-pairs rank-biserial correlation, computed as $r = (W+ - W-) / [n(n+1)/2] = (41 - 14) / 55 \approx 0.49$ for $n = 10$ paired folds, assuming no fold produced identical RF and KNN accuracy. Future work should adopt repeated or nested stratified cross-validation with a correction for the dependence between folds, such as a corrected resampled t-test, to obtain more reliable comparative inference.

2.7 Comparative Analysis

All experimental results were subsequently analyzed comparatively to examine the performance characteristics and relative feature importance of the Random Forest and K-Nearest Neighbor algorithms under the same experimental conditions. The comparative analysis considered testing-set performance, 10-fold cross-validation results, feature importance patterns, and the Wilcoxon Signed-Rank Test results to provide a comprehensive interpretation of the two algorithms in the context of academic decision support within Indonesian PHEIs.

4. DATA ANALYSIS AND DISCUSSION

This section presents the results and discussion organized according to the three research objectives: comparison of RF and KNN performance (Sections 3.1–3.3), feature importance analysis as a progressive predictive signal map (Section 3.4), and statistical validation using the Wilcoxon Signed-Rank Test (Section 3.5), which are subsequently synthesized in an integrated discussion in Section 3.6.

These results can also be read through the study's three philosophical foundations. Ontologically, the measurable phenomenon of on-time graduation is manifested in the model outputs and performance results. Epistemologically, knowledge about its predictive structure is established empirically through evaluation metrics, feature importance analysis, and statistical testing rather than assumption. Axiologically, the institutional and ethical implications discussed in Section 3.7, including fairness in the use of demographic attributes and the responsible treatment of predictions, reflect the value orientation that guides the interpretation of these findings.

3.1 Random Forest Evaluation Results

As an ensemble learning method, Random Forest (RF) constructs a classification model by combining the majority voting results of multiple decision trees trained independently using bootstrap sampling (Breiman, 2001). The model evaluation results on the 132 testing samples are presented in Tables 5 and 6.

Table 5. Confusion Matrix of the Random Forest Algorithm

Actual / Predicted	Not On-Time Graduation	On-Time Graduation	Total
Not On-Time Graduation (0)	14 (TP_0)	21 (FN_0)	35
On-Time Graduation (1)	5 (FP_0)	92 (TN_0)	97

Total	19	113	132

Table 6. Classification Report of the Random Forest Algorithm

Class	Precision	Recall	F1-Score	Support
0 – Not On-Time Graduation	0.74	0.40	0.52	35
1 – On-Time Graduation	0.81	0.95	0.88	97
Accuracy			0.8030	132
Weighted Average	0.79	0.80	0.78	132

Based on Tables 5 and 6, the RF model achieved an accuracy of 80.30% on the testing data. For Class 1 (On-Time Graduation), RF recorded a precision of 0.81, a recall of 0.95, and an F1-score of 0.88, with 92 of the 97 observations in this class correctly classified. For Class 0 (Not On-Time Graduation), the model recorded a precision of 0.74, a recall of 0.40, and an F1-score of 0.52, correctly identifying 14 of the 35 observations in this class. The weighted precision, recall, and F1-score were 0.79, 0.80, and 0.78, respectively.

The lower recall for the minority class should be interpreted in the context of the class distribution in the testing data. The testing set consisted of 132 observations, with the minority class comprising 35 observations (26.52%). Consequently, each correctly classified or misclassified observation in this class affects the recall value by approximately 2.86 percentage points. In the testing set, 14 of the 35 Class 0 observations were correctly identified, while 21 were classified as Class 1. In practical terms, the Random Forest model missed 21 of the 35 students who did not graduate on time, so its overall accuracy overstates its usefulness for identifying at-risk students.

Villar & de Andrade (2024) emphasized that, in imbalanced datasets, class-level evaluation metrics provide important information beyond overall accuracy. Therefore, this study reports class-level precision, recall, and F1-score together with overall accuracy to provide a more comprehensive description of the RF classification results (Holicza & Kiss, 2023).

Table 7. 10-Fold Cross-Validation Results of the RF and KNN Algorithms

Statistic	RF	KNN (K=5)
Mean Accuracy	76.86%	74.71%
Standard Deviation	4.92%	4.54%
Minimum Accuracy	71.21%	65.15%
Maximum Accuracy	86.15%	81.82%

The results of the 10-Fold Cross-Validation for both algorithms are presented in Table 7. RF achieved a mean accuracy of 76.86% with a standard deviation of 4.92%, whereas KNN recorded a mean accuracy of 74.71% with a standard deviation of 4.54%. The accuracy values across the folds ranged from 71.21% to 86.15% for RF and from 65.15% to 81.82% for KNN. The lower standard deviation of KNN indicates slightly lower variation across folds, whereas RF recorded a higher mean accuracy in the cross-validation results. These results describe the observed performance consistency of the two algorithms across the 10 validation folds.

Although KNN exhibited a slightly lower standard deviation, indicating slightly lower variation across the validation folds, RF recorded a higher mean accuracy across the 10-fold cross-validation results. The RF evaluation results were subsequently compared with the KNN evaluation on the testing data, as discussed in Section 3.2.

3.2 Results of the K-Nearest Neighbor Algorithm Evaluation

The KNN model was configured with K = 5 as a fixed default baseline rather than a tuned value, so that the comparison with Random Forest was conducted under fixed and reproducible conditions. A value of K that is too small, such as K = 3, is more susceptible to noise, whereas an excessively large K in a medium-sized dataset tends to produce overly smooth decision boundaries and may reduce sensitivity to local patterns.

Before classification, the preprocessing transformations were fitted using the training data and subsequently applied to the testing data. For KNN, Min-Max Scaling was used to place numerical attributes on a comparable scale so that differences in feature magnitude would not disproportionately affect the Euclidean distance calculation (Ahmed et al., 2022). The KNN evaluation results on the testing data are presented in Tables 8 and 9.

Table 8. Confusion Matrix of the K-Nearest Neighbor Algorithm (K=5)

Actual / Predicted	Not On-Time Graduation	On-Time Graduation	Total
Not On-Time Graduation (0)	9 (TP_0)	26 (FN_0)	35
On-Time Graduation (1)	13 (FP_0)	84 (TN_0)	97
Total	22	110	132

Table 9. Classification Report of the K-Nearest Neighbor Algorithm (K=5)

Class	Precision	Recall	F1-Score	Support
0 – Not On-Time Graduation	0.41	0.26	0.32	35
1 – On-Time Graduation	0.76	0.87	0.81	97
Accuracy			0.7045	132
Weighted Average	0.67	0.70	0.68	132

The KNN model achieved an accuracy of 70.45% on the testing data. For Class 1 (On-Time Graduation), KNN recorded a precision of 0.76, a recall of 0.87, and an F1-score of 0.81, with 84 of the 97 observations correctly classified. For Class 0 (Not On-Time Graduation), the model recorded a precision of 0.41, a recall of 0.26, and an F1-score of 0.32, correctly identifying 9 of the 35 observations in this class. The weighted precision, recall, and F1-score were 0.67, 0.70, and 0.68, respectively. For KNN the shortfall is larger, with 26 of the 35 late-graduating students misclassified as on-time, a false-negative rate that would undermine any early-warning use.

The class-level results show that KNN identified Class 1 more effectively than Class 0 in the testing data. For Class 0, 9 of the 35 observations were correctly identified, while 26 were classified as Class 1. The lower recall for Class 0 should be interpreted in the context of the class distribution, in which Class 0 represented 26.52% of the testing observations. These results describe the observed behavior of KNN under the experimental conditions of this study. RF and KNN employ different classification mechanisms: RF aggregates predictions from multiple decision trees constructed using bootstrap samples and majority voting, whereas KNN determines predictions based on the classes of neighboring observations. The observed differences in class-level performance therefore describe the behavior of the two algorithms under the experimental conditions of this study.

It should also be noted that no explicit class-imbalance handling technique was applied in this study. Therefore, the observed class-level performance should be interpreted as an empirical result under the experimental conditions rather than as evidence that either algorithm inherently resolves class imbalance (Villar & de Andrade, 2024; Holicza & Kiss, 2023). A comparison of the two algorithms across the complete set of evaluation metrics is presented in Section 3.3.

3.3 Comparative Analysis of RF and KNN

A comprehensive comparison of the two algorithms is presented in Table 10 to provide an overall description of the differences in performance across all evaluation metrics used.

Table 10. Comprehensive Performance Comparison of RF and KNN

Evaluation Metric	Random Forest	KNN (K=5)	Difference (RF−KNN)
Accuracy	80.30%	70.45%	+9.85pp
Weighted Precision	79.37%	66.96%	+12.41pp
Weighted Recall	80.30%	70.45%	+9.85pp
Weighted F1-Score	78.14%	68.01%	+10.13pp
Precision – Class 0	0.74	0.41	+0.33
Recall – Class 0	0.40	0.26	+0.14

F1-Score – Class 0		0.52	0.32	+0.20
F1-Score – Class 1		0.88	0.81	+0.07
Cross Validation Mean Accuracy	76.86%	74.71%	+2.15pp	

Note: pp = percentage points

Based on Table 10, RF recorded higher observed values than KNN across the reported testing-set metrics. The largest observed difference was found in weighted precision, with a difference of 12.41 percentage points, followed by weighted F1-score with a difference of 10.13 percentage points. The testing-set accuracy difference was 9.85 percentage points, while the difference in cross-validation mean accuracy was smaller at 2.15 percentage points. These values describe the observed performance differences between the two algorithms under the experimental conditions of this study.

The observed differences are broadly consistent with the comparative pattern reported by Adilah et al. (2024), although differences in datasets, preprocessing procedures, and experimental settings should be considered when interpreting cross-study comparisons.

For Class 0, RF recorded a higher F1-score than KNN (0.52 versus 0.32), with differences also observed in precision (0.74 versus 0.41) and recall (0.40 versus 0.26). These results indicate that, under the experimental conditions of this study, RF identified the minority class more effectively than KNN on the testing data. However, this difference should be interpreted as an observed empirical result rather than evidence that the RF mechanism inherently resolves class imbalance.

The two algorithms employ different classification mechanisms. RF aggregates predictions from multiple decision trees constructed using bootstrap samples and majority voting, whereas KNN assigns predictions based on the classes of neighboring observations in the feature space. These algorithmic differences provide a descriptive context for interpreting the observed class-level performance, but they do not by themselves establish the causal source of the performance differences.

From an academic monitoring perspective, identifying students in the minority class may be particularly relevant because misclassification of students who do not graduate on time may affect the identification of cases requiring further academic attention (Pelima et al., 2024; Villar & de Andrade, 2024).

In terms of cross-validation results, KNN produced a slightly lower standard deviation than RF (4.54% versus 4.92%), indicating slightly lower variation across the validation folds. However, RF recorded a higher mean accuracy (76.86% versus 74.71%). The observed mean difference of 2.15 percentage points is smaller than the differences observed on the testing-set metrics. These results suggest that the two algorithms showed relatively similar average accuracy across the cross-validation folds, although their testing-set performance values differed more substantially.

The dataset contains heterogeneous attributes, including continuous numerical attributes such as GPA, semester GPA, and earned credits, as well as categorical attributes such as gender, admission pathway, and study system. This characteristic provides a relevant context for comparing RF and KNN because the two algorithms process feature relationships differently. KNN relies on distance calculations in the feature space, whereas RF uses decision-tree-based splitting. The extent to which these characteristics contributed to the observed performance differences cannot be established directly from the present experiment.

These class-level results should be read as the central practical finding rather than a footnote to overall accuracy. Both models identify the majority group well but detect the minority late-graduating group poorly (RF recall 0.40; KNN 0.26), so neither is adequate, on accuracy alone, for flagging students who actually need support. Strengthening minority detection would require handling the class imbalance explicitly, for example through resampling such as SMOTE or class-weighted training, together with decision-threshold adjustment and recall-oriented evaluation. These analyses were outside the scope of the present fixed-configuration comparison and are identified as priorities for future work.

3.4 Feature Importance Analysis

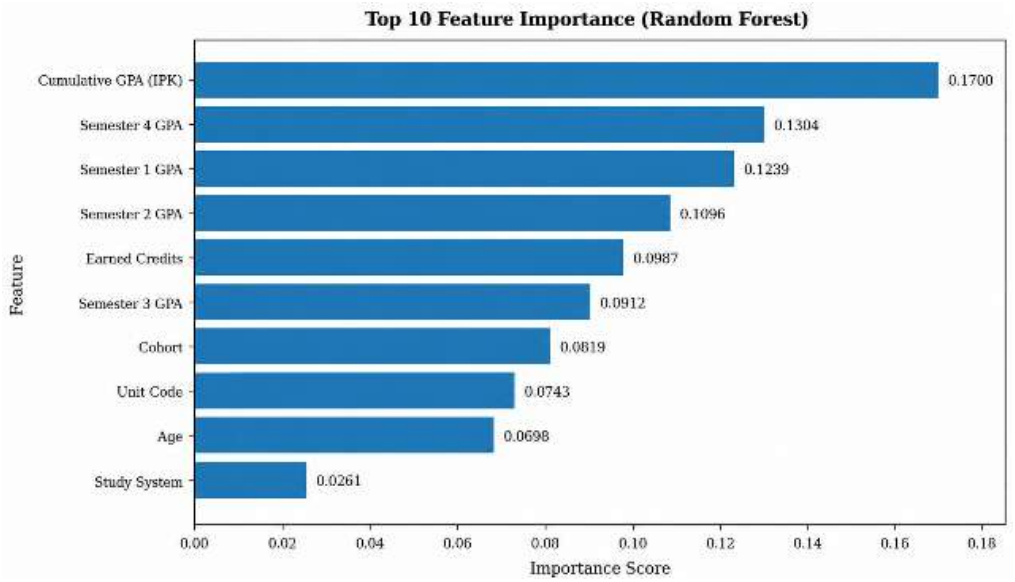


Figure 6. Visualization of Feature Importance for All Attributes

RF provides model-based feature importance measures that can be used to examine the relative importance of the predictor attributes in the classification model. The importance values were calculated using the Mean Decrease in Impurity (MDI) mechanism through the feature_importances_ attribute in the Scikit-Learn library (Nayak et al., 2023). A higher importance value indicates greater relative importance of an attribute in the RF model based on its contribution to impurity reduction across the decision trees (Nachouki et al., 2023).

The cumulative GPA (IPK) attribute ranked first, with an importance value of 0.1700, corresponding to approximately 17.00% of the total MDI importance. It was the most important individual predictor in the RF model, followed by Semester 4 GPA (13.04%) and Semester 1 GPA (12.39%). These values indicate their relative importance within the fitted RF model rather than a causal effect on graduation outcomes. The relatively high importance of cumulative GPA is consistent with findings reported by Santoso et al. (2024), who identified GPA-related variables as important predictors in student graduation analysis. It is also consistent with Nachouki et al. (2023), who reported substantial importance for cumulative academic performance variables in RF-based student prediction. However, the feature importance values in the present study describe model-based predictive importance and should not be interpreted as evidence of a causal relationship.

The group comprising Semester 1–4 GPA and earned credits accounted for approximately 55.38% of the total MDI importance. The individual importance values were 13.04% for Semester 4 GPA, 12.39% for Semester 1 GPA, 10.96% for Semester 2 GPA, 9.12% for Semester 3 GPA, and 9.87% for earned credits. Together with cumulative GPA, these academic attributes accounted for approximately 72.38% of the total feature importance. This indicates that academic attributes received substantial relative importance within the fitted RF model.

Semester 4 GPA had the highest importance among the semester-level GPA variables, at 13.04%, followed by Semester 1 GPA at 12.39%, Semester 2 GPA at 10.96%, and Semester 3 GPA at 9.12%. This ranking reflects the relative MDI importance of these variables within the fitted RF model and does not establish why one semester has greater predictive importance than another. The importance of semester-level academic attributes indicates that academic performance variables are relevant predictive signals within the fitted model. This finding may provide information for academic monitoring, although the present model does not specifically evaluate prediction using early-semester information alone. Cohort had an importance value of 8.18%, followed by Unit Code at 7.44% and Age at 6.98%. Study System had a lower importance value of 2.65%. These values represent their relative importance within the RF model. Gender, Admission Pathway, and Leave Status had relatively low importance values of 1.85%, 0.36%, and 0.16%, respectively. These results indicate that these attributes had relatively lower MDI importance than the other predictors in the fitted RF model. From an academic monitoring perspective, the relatively high importance of cumulative GPA and

semester-level GPA variables indicates that academic performance attributes may provide useful information for prioritizing monitoring activities. However, feature importance should be interpreted as model-based predictive information rather than as evidence of causal effects or direct intervention priorities. The RF feature importance analysis therefore provides an empirical basis for examining which attributes receive greater importance within the classification model (Nachouki et al., 2023; Pelima et al., 2024).

Two interpretive cautions apply to these importance values. First, Mean Decrease in Impurity is sensitive to the scale, cardinality, and correlation of predictors and can inflate the apparent importance of continuous or high-cardinality attributes, so the rankings reported here should be read as model-based and provisional; complementary methods such as permutation importance or SHAP are recommended to test their robustness in future work. Second, the relatively high importance of Semester 1 GPA within the full feature set does not establish an early-prediction capability, because the model was trained on all semesters jointly rather than on Semester 1 information alone; demonstrating genuine early-semester prediction would require training and evaluating a model restricted to early-semester features.

3.5 Wilcoxon Signed-Rank Test Results

The Wilcoxon Signed-Rank Test was conducted to assess whether the paired 10-fold cross-validation accuracy scores of RF and KNN differed statistically. The test was selected as a non-parametric method for comparing paired observations without requiring the assumption of normally distributed differences between the paired scores (Rainio et al., 2024).

Table 11. Results of the Wilcoxon Signed-Rank Test between RF and KNN

Parameter	Value
W Statistic	14.0
p-value	0.3477
Significance Level (α)	0.05
Decision on H_0	Fail to Reject H_0
Statistical Conclusion	No statistically significant difference

The test produced W = 14.0 and p = 0.3477. Because the p-value is greater than the significance level of α = 0.05, the null hypothesis is not rejected. Therefore, the results do not indicate a statistically significant difference between the paired cross-validation accuracy scores of RF and KNN.

To complement the significance test with a measure of effect magnitude, the matched-pairs rank-biserial correlation was computed from the Wilcoxon signed-rank results. With ten paired fold-level observations, the total rank sum is 55; given W = 14, the sum of ranks favoring Random Forest is 41 and the sum favoring KNN is 14, yielding a rank-biserial correlation of $r \approx (41 - 14)/55 = 0.49$. This value indicates a moderate effect size in the direction of Random Forest, meaning that across the ten folds the fold-level accuracy differences more often favored RF than KNN. However, this moderate effect magnitude was not accompanied by statistical significance (p = 0.3477), a pattern consistent with the limited statistical power afforded by only ten paired observations. The effect size and the non-significant p-value should therefore be interpreted together: the descriptive and rank-based evidence points toward a consistent RF advantage, whereas the inferential test does not provide sufficient evidence to treat this advantage as statistically established. This rank-biserial estimate assumes that no fold produced identical RF and KNN accuracy; if such ties occurred, the total rank sum would be adjusted accordingly.

The absence of a statistically significant difference should not be interpreted as evidence that the two algorithms have identical performance. The inferential result should also be considered in the context of the 10 paired observations used in the Wilcoxon analysis. With a limited number of paired folds, the statistical inference should be interpreted together with the descriptive performance results reported in Tables 7 and 10.

Although RF recorded higher observed values for several descriptive performance measures, including testing-set accuracy and cross-validation mean accuracy, these differences were not statistically significant according to the Wilcoxon Signed-Rank Test. Therefore, the descriptive differences and the inferential result should be considered together when interpreting the comparative performance of the two algorithms. Future studies may consider repeated K-fold cross-validation to obtain a larger number of paired performance estimates and provide additional evidence for statistical comparison between algorithms.

3.6 Discussion

The results of this study show higher observed values for RF than KNN across the reported testing-set metrics. RF achieved an accuracy of 80.30%, compared with 70.45% for KNN, while the weighted F1-scores were 78.14% and 68.01%, respectively. In the 10-fold cross-validation results, RF recorded a mean accuracy of 76.86%, compared with 74.71% for KNN. These descriptive differences provide a basis for comparing the performance characteristics of the two algorithms under the same experimental conditions.

The observed class-level differences between KNN and RF may be interpreted in the context of their different classification mechanisms. KNN relies on distance relationships in the feature space, whereas RF uses decision-tree-based splitting and ensemble aggregation. The dataset also contains heterogeneous numerical and categorical attributes and an imbalanced class distribution. These characteristics provide relevant context for interpreting the observed results, although their individual causal contributions to the performance differences were not directly tested in this study.

To position the findings of this study within the broader literature, Table 12 presents a descriptive comparison between the findings of this study and relevant previous studies in the domain of machine-learning-based student performance and graduation prediction. The comparison focuses on dataset context, algorithms compared, and the main findings reported by each study.

Table 12. Descriptive Comparison of the Findings of This Study with Previous Studies

Study	Dataset Context	Algorithms Compared	Relevant Findings
This study	656 academic records, Indonesian private higher education institution, graduation prediction	RF vs. KNN	RF achieved higher performance across all metrics; accuracy of 80.30% versus 70.45%;
Yağcı (2022)	1,854 students, Türkiye, prediction of final examination scores from midterm examination scores	RF, KNN, SVM, LR, NB, and NN	RF and KNN demonstrated competitive performance among the six compared algorithms; overall accuracy ranged from 70–75%
Dawar et al. (2024)	Student academic data, India, academic performance prediction	RF, KNN, DT, NB	RF achieved higher predictive performance than KNN and the other algorithms
Kim et al. (2023)	University student academic data, Korea, dropout prediction	Hybrid model Hybrid model (RF, GB, SVM, K-Means)	RF achieved high precision and recall in identifying students at risk of dropout
Pelima et al. (2024)	Systematic review of 18 studies on student graduation prediction	RF (dominant)	RF consistently emerged as the best-performing algorithm across various contexts of student graduation prediction

Table 12 shows that RF has been examined in several studies involving student academic prediction and related educational outcomes. The findings of the present study are directionally consistent with several comparative studies that report competitive RF performance, although differences in datasets, target variables, preprocessing procedures, and experimental configurations limit direct cross-study comparison.

Beyond restating these numbers, the study contributes to educational data mining in two ways. Under identical experimental conditions, it shows that a measured accuracy gap between RF and KNN need not correspond to a statistically meaningful difference, which cautions against the accuracy-only algorithm claims common in the Indonesian EDM literature. It also pairs a semester-resolved view of feature importance with an explicit account of where the models fail, namely minority detection, rather than reporting headline accuracy alone.

The Wilcoxon Signed-Rank Test produced $W=14.0$ and $p=0.3477$, indicating that the null hypothesis was not rejected at $\alpha = 0.05$. Therefore, the test did not indicate a statistically significant difference between the paired 10-fold cross-validation accuracy scores of RF and KNN. This result should not be interpreted as evidence that the two algorithms have identical performance. The inferential result should be considered together with the descriptive differences observed in the testing-set and cross-validation results, particularly because the Wilcoxon analysis was based on 10 paired fold-level observations.

Descriptively, RF recorded higher values for the reported testing-set metrics, including a 0.20 difference in minority-class F1-score between RF and KNN (0.52 versus 0.32). This difference represents an observed performance gap in the present testing data, but it should not be interpreted as a formal measure of practical significance.

An institutionally relevant finding concerns the distribution of feature importance across academic attributes. Cumulative GPA (IPK), Semester 1–4 GPA, and Earned Credits collectively accounted for approximately 72.38% of the total MDI importance. This indicates that academic attributes received substantial relative importance within the fitted RF model compared with the other predictor groups. Among the semester-level GPA variables, Semester 4 GPA had the highest importance at 13.04%, followed by Semester 1 GPA at 12.39%, Semester 2 GPA at 10.96%, and Semester 3 GPA at 9.12%. The importance of Semester 1 GPA indicates that this variable receives substantial relative importance within the full feature set; however, the present study does not evaluate a prediction model using Semester 1 information alone. Therefore, the result should not be interpreted as evidence that the model can already provide early-semester prediction.

Gender, Admission Pathway, and Leave Status had relatively low MDI importance values of 1.85%, 0.36%, and 0.16%, respectively. These values indicate relatively lower model-based importance than the major academic predictors in this dataset. Nevertheless, low feature importance should be interpreted as a property of the fitted model rather than as evidence that these attributes have no relevance to student outcomes.

Three limitations of this study should be acknowledged. First, no explicit class-imbalance handling technique was applied, and minority-class recall remained limited (RF: 0.40; KNN: 0.26). Therefore, the models did not identify all observations belonging to the minority class in the testing data. Second, the RF feature importance values were calculated using MDI. Although MDI provides a model-based measure of feature importance, its values may be affected by characteristics of the predictors, including their distribution and cardinality. Future studies may therefore complement MDI with permutation importance or SHAP-based analyses to further examine the stability of the identified importance patterns. Future studies may also evaluate explicit class-imbalance strategies, such as oversampling methods including SMOTE, and assess their effects on minority-class performance. Third, the dataset was obtained from a single institution and covered the period 2018–2022; therefore, generalization to other private higher education contexts requires validation using datasets from multiple institutions and settings.

From an academic monitoring perspective, the findings may provide information for prioritizing the review of academic performance indicators. Cumulative GPA and semester-level GPA received relatively high feature importance in the fitted RF model and may therefore be considered as variables for further academic monitoring. Student academic profiles may also be reviewed periodically as additional semester-level information becomes available. Academic advising may use model outputs as supplementary information alongside existing academic assessments rather than relying on demographic attributes alone. These implications require validation in an actual academic monitoring setting before being translated into operational intervention procedures.

3.7 Institutional and Ethical Implications

Beyond the comparative performance results, the findings carry implications for how predictive systems are situated within institutional decision-making in private higher education. First, the deployment of graduation-risk prediction shifts part of the institutional monitoring process toward algorithmic mediation. Because model outputs can influence which students are flagged for academic attention, institutions bear responsibility for how such flags are generated, communicated, and acted upon. Predictions should function as supplementary information supporting human academic judgment rather than as deterministic labels, particularly given the limited minority-class recall observed in this study (RF: 0.40; KNN: 0.26), which indicates that a non-trivial proportion of at-risk students may not be detected by the model.

Second, the feature importance results have a fairness dimension that merits explicit consideration. Although demographic attributes such as gender (1.85%) and age (6.98%) received relatively low importance in the fitted RF model, their presence in the feature set raises the question of whether demographic characteristics should inform academic intervention decisions at all. From a technology-and-society standpoint, using demographic attributes as a basis for prioritizing intervention could reproduce or amplify existing inequities, even when their statistical contribution is small. The results therefore support prioritizing modifiable academic indicators, such as cumulative and semester-level GPA and earned credits, over demographic attributes when translating model outputs into institutional action.

Third, the value of such systems in the private higher education context depends on institutional capacity. The uneven resources across private institutions imply that the same predictive tool may yield different social outcomes depending on whether an institution can act on early-warning signals with adequate advising and support structures. These considerations position graduation prediction not merely as a technical artifact but as a socio-technical intervention whose responsible use requires institutional safeguards, transparency toward students, and validation in an actual academic-monitoring setting before operational adoption.

A distinction should be drawn between the safeguards this study proposes and the governance mechanisms it actually evaluates. The recommendations on accountability, fairness, and the limited use of demographic attributes are normative proposals; no institutional governance process, intervention, or fairness audit was empirically tested here. Because the dataset comes from a single private institution and the 2018 to 2022 period, these implications may not generalize to other institutions or cohorts and should be validated with multi-institution data before any operational use.

5. CONCLUSION, IMPLICATION, SUGGESTION, AND LIMITATIONS

This study compared Random Forest and K-Nearest Neighbor under the same experimental conditions, examined feature importance using the RF model, and applied the Wilcoxon Signed-Rank Test to the paired cross-validation accuracy scores. Together, these analyses provide a descriptive and statistical basis for interpreting the comparative performance of the two algorithms and the relative importance of predictor attributes in student graduation classification.

RF recorded higher observed values than KNN across the reported testing-set metrics. RF achieved an accuracy of 80.30%, compared with 70.45% for KNN, while the weighted F1-scores were 78.14% and 68.01%, respectively. The 10-fold cross-validation mean accuracy was 76.86% for RF and 74.71% for KNN. However, the Wilcoxon Signed-Rank Test produced $W = 14.0$ and $p = 0.3477$, so the null hypothesis was not rejected at $\alpha = 0.05$. Therefore, the observed descriptive differences were not accompanied by a statistically significant difference in the paired cross-validation accuracy scores. The rank-biserial correlation derived from the Wilcoxon comparison ($r \approx 0.49$) indicated a moderate effect in favor of RF, but the accompanying non-significant p-value ($p = 0.3477$) reflected the limited power of ten paired observations; the two results should be read together rather than in isolation. It should be emphasized that the higher RF figures are not statistically significant ($p = 0.3477$), and that both models recorded low recall for the delayed-graduation class (RF 0.40; KNN 0.26); neither model is therefore suited, on its own, to high-stakes academic decisions about individual students.

The feature importance analysis showed that academic attributes received substantial relative importance in the fitted RF model. Cumulative GPA (IPK), Semester 1–4 GPA, and Earned Credits collectively accounted for approximately 72.38% of the total MDI importance. Semester 4 GPA had the highest importance among the semester-level GPA variables at 13.04%, followed by Semester 1 GPA at 12.39%, Semester 2 GPA at 10.96%, and Semester 3 GPA at 9.12%. Gender, Admission Pathway, and Leave Status had relatively low MDI importance values of 1.85%, 0.36%, and 0.16%, respectively. These values represent model-based relative importance and should not be interpreted as causal effects.

Although Semester 1 GPA received substantial relative importance within the full feature set, the present study did not evaluate a model using early-semester information alone. Therefore, the findings provide information for academic monitoring but do not by themselves establish early-semester prediction capability. The findings may serve as a basis for further validation using data from multiple institutions, explicit class-imbalance strategies, and alternative feature-importance methods such as permutation importance or SHAP. From a technology-and-society perspective, these findings suggest that predictive models can support institutional early-warning processes in private higher education, but their responsible use requires treating predictions as decision support rather than determinations, prioritizing modifiable academic indicators over demographic attributes, and validating the approach within institutional monitoring practice before operational adoption. Identifying which attributes carry predictive weight is not the same as demonstrating an operational early-warning system; because no early-semester-only model or real intervention was evaluated, these results should not be read as evidence that the approach is ready for institutional deployment.

Finally, from the perspective of the philosophy of science, this study reaffirms its three foundations: ontologically, on-time graduation was treated as a measurable academic phenomenon; epistemologically, knowledge about it was constructed empirically through machine learning and statistical validation; and

axiologically, the value of this knowledge lies in its responsible institutional use, supporting early and equitable intervention while treating predictions as decision support rather than determinations.

This study has several limitations that bound its conclusions. The data come from a single private institution over the 2018 to 2022 period; the class distribution is imbalanced and was not corrected; the model configurations were fixed rather than tuned; some predictors accumulate over the enrollment period and may not be strictly separated in time from the outcome; and the statistical evidence rests on ten dependent cross-validation folds. Future work should prioritize external validation across multiple institutions, temporally appropriate prediction using early-semester features only, improved detection of the minority class, and explicit evaluation of fairness and institutional implementation before any operational adoption.

REFERENCES

Ahmed, H. A., Muhammad Ali, P. J., Faeq, A. K., & Abdullah, S. M. (2022). An Investigation on Disparity Responds of Machine Learning Algorithms to Data Normalization Method. *ARO-THE SCIENTIFIC JOURNAL OF KOYA UNIVERSITY*, *10*(2 SE-Review Articles), 29–37. https://doi.org/10.14500/aro.10970

Albreiki, B., Zaki, N., & Alashwal, H. (2021). A Systematic Literature Review of Students' Performance Prediction Using Machine Learning Techniques. *Education Sciences*, 11(9), 552. https://doi.org/10.3390/educsci11090552

Alyahyan, E., & Düştegör, D. (2020). Predicting academic success in higher education: Literature review and best practices. *International Journal of Educational Technology in Higher Education*, 17, Article 3. https://doi.org/10.1186/s41239-020-0177-7

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Cruz, M. M., & Lumauag, R. (2025). *Comparative Analysis of Machine Learning Algorithms for Predicting Student Academic Performance in Higher Education*. https://doi.org/10.1109/ICUIS64676.2024.10866086

Dawar, I., Negi, S., Lamba, S., & Kumar, A. (2024). Enhancing Student Academic Performance Forecasting: A Comparative Analysis of Machine Learning Algorithms. *SN Computer Science*, *5*(6), 758. https://doi.org/10.1007/s42979-024-03118-3

Delen, D. (2010). A comparative analysis of machine learning techniques for student retention management. *Decision Support Systems*, 49(4), 498–506. https://doi.org/10.1016/j.dss.2010.06.003

Holicza, B., & Kiss, A. (2023). Predicting and Comparing Students' Online and Offline Academic Performance Using Machine Learning Algorithms. *Behavioral Sciences*, 13(4), 289. https://doi.org/10.3390/bs13040289

Inyang, U. G., Ijebu, F. F., Osang, F. B., & Afolorunso, A. A. (2023). A Dataset-Driven Parameter Tuning Approach for Enhanced K-Nearest Neighbour Algorithm Performance. *International Journal on Advanced Science Engineering Information Technology*, *13*(1), 380–391. https://doi.org/10.18517/ijaseit.13.1.16706

Kabathova, J., & Drlik, M. (2021). Towards Predicting Student's Dropout in University Courses Using Different Machine Learning Techniques. *Applied Sciences*, 11(7), 3130. https://doi.org/10.3390/app11073130

Kementerian Pendidikan Tinggi Sains dan Teknologi. (2025). *Peraturan Menteri Pendidikan Tinggi, Sains, dan Teknologi Nomor 39 Tahun 2025 tentang Penjaminan Mutu Pendidikan Tinggi*.

Kim, S., Choi, E., Jun, Y.-K., & Lee, S. (2023). Student Dropout Prediction for University with High Precision and Recall. *Applied Sciences*, 13(10), 6275. https://doi.org/10.3390/app13106275

Kroll, J. A., Huey, J., Barocas, S., Felten, E. W., Reidenberg, J. R., Robinson, D. G., & Yu, H. (2017). Accountable algorithms. *University of Pennsylvania Law Review*, 165(3), 633–705.

Nachouki, M., Mohamed, E. A., Mehdi, R., & Abou Naaj, M. (2023). Student course grade prediction using the random forest algorithm: Analysis of predictors' importance. *Trends in Neuroscience and Education*, *33*, 100214. https://doi.org/10.1016/j.tine.2023.100214

Nayak, P., Vaheed, Sk., Gupta, S., & Mohan, N. (2023). Predicting students' academic performance by mining the educational data through machine learning-based classification model. *Education and Information Technologies*, *28*(11), 14611–14637. https://doi.org/10.1007/s10639-023-11706-8

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

Psyridou, M., Prezja, F., Torppa, M., Lerkkanen, M.-K., Poikkeus, A.-M., & Vasalampi, K. (2024). Machine

learning predicts upper secondary education dropout as early as the end of primary school. *Scientific Reports*, *14*(1), 12956. https://doi.org/10.1038/s41598-024-63629-0

Rainio, O., Teuho, J., & Klén, R. (2024). Evaluation metrics and statistical tests for machine learning. *Scientific Reports*, *14*(1), 6086. https://doi.org/10.1038/s41598-024-56706-x

Santoso, A., Retnawati, H., Kartianom, K., Apino, E., Rafi, I., & Rosyada, M. (2024). Predicting Time to Graduation of Open University Students: An Educational Data Mining Study. *Open Education Studies*, *6*(1). https://doi.org/10.1515/edu-2022-0220

Shafiq, D. A., Marjani, M., Habeeb, R. A. A., & Asirvatham, D. (2022). Student Retention Using Educational Data Mining and Predictive Analytics: A Systematic Literature Review. *IEEE Access*, *10*, 72480–72503. https://doi.org/10.1109/ACCESS.2022.3188767

Slade, S., & Prinsloo, P. (2013). Learning analytics: Ethical issues and dilemmas. *American Behavioral Scientist*, 57(10), 1510–1529. https://doi.org/10.1177/0002764213479366

Thabtah, F., Hammoud, S., Kamalov, F., & Gonsalves, A. (2020). Data imbalance in classification: Experimental evaluation. *Information Sciences*, *513*, 429–441. https://doi.org/10.1016/j.ins.2019.11.004

Villar, A., & de Andrade, C. R. V. (2024). Supervised machine learning algorithms for predicting student dropout and academic success: a comparative study. *Discover Artificial Intelligence*, *4*(1), 2. https://doi.org/10.1007/s44163-023-00079-z

Wati, E. F., Perangin-angin, E. S., & Sari, A. P. (2023). Prediction of Student Graduation using the K-Nearest Neighbors Method. *International Journal of Information System and Technology*, *7*(3), 211–216. https://doi.org/10.30645/ijistech.v7i3

Yağcı, M. (2022). Educational data mining: prediction of students' academic performance using machine learning algorithms. *Smart Learning Environments*, *9*(1), 11. https://doi.org/10.1186/s40561-022-00192-z

Zanellati, A., Zingaro, S. P., & Gabbrielli, M. (2024). Balancing Performance and Explainability in Academic Dropout Prediction. *IEEE Transactions on Learning Technologies*, *17*, 2086–2099. https://doi.org/10.1109/TLT.2024.3425959