Comparison of Random Forest and XGBoost Algorithms for Predicting On-Time Graduation of Scholarship Recipients at STMIK PN

Yulianus Zai, Andysah Putera Utama Siahaan

Abstract

On-time graduation of scholarship recipients is a critical indicator for evaluating the efficacy of higher education financial aid and achieving the Sustainable Development Goals (SDG 4: Quality Education). This study compares the predictive performance of Random Forest and Extreme Gradient Boosting (XGBoost) for an early warning system based on institutional academic records at the end of the sixth semester. The training dataset comprises 71 scholarship recipients (KIP-Kuliah and Foundation Scholarship) from the 2019–2022 cohorts at STMIK PN, of whom 50 graduated on time (70.4%) and 21 did not (29.6%). Nine academic features measured up to the end of the sixth semester were used: semester GPAs (Semesters 1 to 6), cumulative GPA, credits passed, and the number of courses graded D/E. Class imbalance was handled with SMOTE applied to the training folds, and performance was evaluated using Repeated Stratified 5-Fold Cross-Validation (10 repetitions). Random Forest achieved an Accuracy of 85.46%, Precision of 77.35%, F1-Score of 75.71%, and AUC-ROC of 0.9405, whereas XGBoost attained the highest Recall (80.30%) with an F1-Score of 74.51% and AUC-ROC of 0.9041. The corrected resampled t-test showed no statistically significant differences between the models, including against the Decision Tree baseline ($p > 0.05$). Tree SHAP analysis revealed that the sixth-semester cumulative GPA is the dominant predictor, followed by the GPAs of Semesters 3 to 5. Deployment on 21 active scholarship recipients from the 2023 cohort flagged 1 high-risk student (98.3% risk probability on XGBoost and 88.9% on Random Forest), enabling academic intervention before the final academic year.

Keywords: Random Forest, XGBoost, Graduation Prediction, Scholarship Recipients, Early Warning System, SHAP, SDG 4.

Yulianus Zai[1]
[1]Magister Teknologi Informasi, Universitas Pembangunan Panca Budi, Indonesia
e-mail: yulianuszai98@gmail.com [1]

Andysah Putera Utama Siahaan[2]
[2]Magister Teknologi Informasi, Universitas Pembangunan Panca Budi, Indonesia
e-mail: andiesiahaan@gmail.com [2]
2nd International Conference on Islamic Community Studies (ICICS)
Theme: History of Malay Civilization and Islamic Human Capacity and Halal Hub in the Globalization Era
https://proceeding.pancabudi.ac.id/index.php/ICIE/index

Introduction

Higher education plays an important role in sustainable development. The fourth Sustainable Development Goal (SDG 4) calls for equal access to higher education [1], and a systematic review shows that need-based grants quite consistently improve the completion rates of students from disadvantaged backgrounds [2]. In Indonesia, the government and private higher education institutions provide tuition assistance, such as the Indonesia Smart Card for College program (KIP-Kuliah) and the Foundation Scholarship, to students from economically vulnerable backgrounds; such programs are therefore also related to the tenth Sustainable Development Goal (SDG 10) on reduced inequalities. A study of the government Bidikmisi scholarship program showed that recipients had a higher grade point average (GPA), with the strongest association among students from the lowest-income families [3]. The effectiveness of scholarship programs is also measured by whether recipients complete their studies; research on private higher education institutions in Brazil showed that recipients of the Bolsa Permanência scholarship program had a higher probability of graduating and completed their studies in less time than students who did not receive it [4]. In this study, on-time graduation is defined as completing an undergraduate (S1) degree within a maximum of 8 semesters (4 years).

Delayed graduation among scholarship recipients has consequences for both students and institutions. For students, studying beyond the eighth semester ends their tuition subsidy. Scholarship status itself is related to study persistence: an analysis of data from a university in Portugal showed that receiving a scholarship reduced the odds of dropping out by almost 40% [5]. For institutions, study duration and on-time graduation are important criteria in the accreditation and quality assessment of study programs [6]. At STMIK PN, the graduation of scholarship recipients has so far been monitored only through an administrative review of transcripts shortly before the graduation assessment. This reactive mechanism comes too late: by the time students reach the seventh or eighth semester with an unfinished thesis and courses to retake, little time remains for remedial intervention. Yet the institutional data collected by universities can be used to build models that predict students' time to graduation [7].

Advances in Educational Data Mining (EDM) and Machine Learning (ML) enable universities to identify dropout risk factors so that preventive action can be taken earlier [5]. Several previous studies have applied ensemble algorithms to predict student graduation. Imron et al. [6] found that Random Forest was superior with default parameters, whereas XGBoost achieved higher accuracy after hyperparameter tuning. Sitio et al. [8] compared Random Forest and XGBoost and obtained almost equal accuracy, with XGBoost performing better on recall and F1-Score, while Nirmala et al. [9] showed that XGBoost with MissForest imputation significantly outperformed Random Forest in predicting study completion status. However, these studies were conducted on general student populations rather than specifically on scholarship recipients. In addition, some studies include demographic and socioeconomic variables as predictors, although the use of such features requires fairness considerations because it risks perpetuating existing inequalities [10].

The scientific novelty of this article lies in three main contributions. First, this study focuses specifically on the population of scholarship recipients (Foundation Scholarship and KIP-Kuliah) at a private computing higher education institution, a group with strict limits on study duration. Second, the predictive model is built solely on institutional academic records without demographic attributes. This approach is consistent with situations in which access to students' personal data is limited and raises privacy issues, so that institutions need to rely on academic data to build reliable prediction systems [11]. Third, the sixth-semester prediction approach previously applied by Daniawan et al. [12] to data on 564 students is applied specifically to the population of scholarship recipients, with all features measured exactly at the end of the sixth semester, and is combined with the SMOTE technique for imbalanced data

[13], Repeated Stratified 5-Fold cross-validation, and Tree SHAP (Shapley Additive Explanations) [14] explanations to reveal the academic features that most strongly determine graduation.

Based on the background and research gap described above, this study aims to compare the predictive performance of Random Forest and XGBoost, using Repeated Stratified 5-Fold cross-validation with SMOTE class balancing, in classifying the on-time graduation of scholarship recipients at STMIK PN based on academic records at the end of the sixth semester. This study also aims to analyze the importance and direction of contribution of each academic feature transparently and objectively using an Explainable AI approach based on Tree SHAP, in order to identify the most decisive factors behind delayed graduation. Furthermore, this study tests and demonstrates the model as a prototype Early Warning System for detecting the delay-risk profiles of an active scholarship cohort (the 2023 cohort), thereby producing targeted pedagogical and managerial intervention recommendations for university scholarship managers before students enter their final year of study.

Literature Review

Ensemble learning algorithms, particularly Random Forest and XGBoost, are widely applied in research on predicting student graduation and dropout [6], [8], [9], [10]. Ensemble learning combines the predictions of several base models to improve the generalization ability of the model.

Random Forest, introduced by Breiman [15], is a combination of decision trees, each grown from a random vector, with random feature selection at each split. Breiman [15] showed that the generalization error of a Random Forest converges as the number of trees increases, so that adding trees does not cause overfitting, and that this error depends on the strength of the individual trees and the correlation between them. XGBoost, developed by Chen and Guestrin [16], is a scalable tree boosting system that is widely used to achieve state-of-the-art results in many machine learning competitions. XGBoost minimizes a regularized objective function that includes a penalty on the number of leaves and an L2 penalty on leaf weights, and uses a second-order approximation of the loss function during training [16].

One of the challenges in modeling student graduation is class imbalance, that is, a condition in which students who graduate on time far outnumber those who graduate late or drop out. On imbalanced data, predictive accuracy is not an appropriate evaluation measure [13]. SMOTE (Synthetic Minority Over-sampling Technique) addresses this problem by generating synthetic minority-class samples between each minority sample and its k-nearest neighbors [13]. This technique has been applied to dropout prediction in higher education, where Random Forest was the most accurate and robust model after class balancing [17].

Beyond predictive accuracy, a challenge for adopting machine learning in educational institutions is model interpretability (Explainable Artificial Intelligence/XAI). Ensemble models such as XGBoost are generally regarded as black boxes that are difficult to understand [18], and explaining the predictions of tree-based models has received relatively little attention [14]. Lundberg et al. [14] developed Tree SHAP, which computes exact Shapley values in polynomial time with the theoretical guarantees of local accuracy and consistency. SHAP values have also been used to explain the predictions of ensemble models of student academic performance [18].

Research Methodology

The research framework follows standard data analytics stages: collection of the institutional dataset, preprocessing and engineering of academic features at the end of the sixth semester, class balancing of the training data with SMOTE, Repeated Stratified 5-Fold Cross-

Validation, multi-metric comparative evaluation and statistical significance testing, interpretation of SHAP values, and applied testing on an active cohort.

3.1 Data Source and Sample Description

This study uses secondary data in the form of academic archives of scholarship recipients (Foundation Scholarship and KIP-Kuliah) at STMIK PN, obtained from the collective Study Result Cards (Kartu Hasil Studi, KHS) and the study status records in the campus Academic Information System (SIAKAD). The training dataset covers 71 students from four cohorts (2019, 2020, 2021, and 2022) whose final study status is known. The inclusion criterion was that students were still registered as active at the end of the sixth semester, which is the checkpoint of the early warning system; students who left before the sixth semester were excluded because they cannot be predicted at that point. Of the 71 students, 50 (70.4%) graduated on time (<= 8 semesters), while 21 (29.6%) did not graduate on time, consisting of 9 students who had not graduated after the eighth semester, 2 students who graduated after the eighth semester, and 10 students who dropped out. In addition to the training dataset, a test dataset was prepared containing 21 scholarship recipients from the 2023 cohort who are currently active and do not yet have a graduation label. All data were anonymized using identity codes (IDs) without names or student registration numbers.

3.2 Definition and Engineering of Academic Features

To comply with the principle of an early warning system at the end of the sixth semester, all variables were restricted to academic performance records from the first to the sixth semester. Nine academic predictor features were defined, as summarized in Table 1. The cumulative GPA (IPK), credits passed (SKS), and the number of courses graded D/E were calculated from the best grade of each course up to the end of the sixth semester, while the semester GPA (IPS) of a semester the student did not attend was set to 0. All variables that only become known after the sixth semester (such as the final study grade and the SIAKAD status) were excluded from the feature matrix to prevent target leakage, which could produce spurious accuracy.

Table 1. Description of Academic Predictor Variables and Descriptive Statistics (N = 71)

Feature Code	Variable Name	Data Type	Mean ± SD	Operational Definition
IPS_Sem1	Semester 1 GPA	Float	3.38 ± 0.30	Early academic adjustment performance
IPS_Sem2	Semester 2 GPA	Float	3.43 ± 0.31	First-year academic stability
IPS_Sem3	Semester 3 GPA	Float	3.41 ± 0.64	Performance in core competency courses
IPS_Sem4	Semester 4 GPA	Float	3.45 ± 0.77	Mid-study evaluation (milestone)
IPS_Sem5	Semester 5 GPA	Float	3.22 ± 1.17	Performance in advanced & elective courses
IPS_Sem6	Semester 6 GPA	Float	3.20 ± 0.89	Third-year evaluation checkpoint (EWS)
IPK_Sem6	Cumulative GPA at Sem 6	Float	3.35 ± 0.53	Cumulative performance up to sem 6
SKS_Lulus_Sem6	Total Credits Passed at Sem 6	Integer	108.9 ± 16.5	Accumulated credits passed up to sem 6
MK_DE_Sem6	No. of Courses Graded D/E	Integer	2.11 ± 4.63	Courses graded D/E up to sem 6

3.3 Validation Scheme and Model Training

The training dataset was evaluated using Repeated Stratified 5-Fold Cross-Validation with 10 repetitions (50 folds in total) and controlled randomization (random_state = 42). Repetition was used to obtain more stable performance estimates on a small dataset. In each fold, the data were split into approximately 80% training data (56–57 samples) and 20% validation data (14–15 samples) while preserving the class distribution in each fold. SMOTE (k_neighbors = 3) was applied strictly to the training data (training fold) only, to prevent information leakage from the validation data, following the procedure of a previous study [10]. Three models were built with fixed hyperparameters and without optimization. A Decision Tree (CART) with a maximum depth of 3 was used as the baseline. The Random Forest Classifier was configured with n_estimators = 100, max_depth = 4, min_samples_split = 3, and class_weight = 'balanced', while the XGBoost Classifier was configured with n_estimators = 100, max_depth = 3, learning_rate = 0.08, and the weighting parameter scale_pos_weight = 50/21 to compensate for class imbalance. All modeling and evaluation were performed in Python in the Google Colab environment with scikit-learn 1.6.1, xgboost 3.4.1, imbalanced-learn 0.14.2, and shap 0.54.1. The library versions are reported because Random Forest results may differ between scikit-learn versions even with the same random_state.

Differences in performance between models were tested using the corrected resampled t-test [19] on the F1-Score and AUC-ROC of the 50 folds. This test corrects the variance of the score differences because the training sets of different folds overlap, so that it does not produce the overly optimistic significance of an ordinary paired t-test. The significance level was $\alpha = 0.05$ with 49 degrees of freedom.

3.4 Performance Evaluation Metrics

Model performance was evaluated using five metrics: Accuracy, Precision, Recall, F1-Score, and the Area Under the Receiver Operating Characteristic Curve (AUC-ROC). In the context of an early warning system for delayed graduation, Recall (sensitivity) for class 1 (Not On Time) is treated as the primary metric in order to minimize False Negatives (at-risk students wrongly predicted as safe), balanced by the F1-Score and AUC-ROC as indicators of overall discrimination ability. Because the data are imbalanced, a precision-based measure is also reported, as it is more informative than the ROC curve alone for evaluating binary classifiers [20].

Results

The comparative performance of the Decision Tree, Random Forest, and XGBoost classifiers under Repeated Stratified 5-Fold Cross-Validation is presented in Table 2. All scores are reported as mean ± standard deviation over 50 folds.

Table 2. Comparative Performance of the Classification Algorithms (Repeated Stratified 5-Fold CV with SMOTE)

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC-ROC
Decision Tree (Baseline)	81.70 ± 8.81	72.27 ± 17.73	72.50 ± 24.21	68.74 ± 16.29	0.8298 ± 0.1232
Random Forest	85.46 ± 7.56	77.35 ± 16.95	79.10 ± 19.84	75.71 ± 12.77	0.9405 ± 0.0537
XGBoost	83.79 ± 8.76	73.49 ± 17.02	80.30 ± 18.69	74.51 ± 12.99	0.9041 ± 0.0788

As shown in Table 2, Random Forest achieved the highest Accuracy (85.46%), Precision (77.35%), F1-Score (75.71%), and AUC-ROC (0.9405), while XGBoost achieved the highest Recall (80.30%), meaning that it detected about 4 out of every 5 scholarship recipients at risk of delayed graduation. On average, both ensemble models outperformed the Decision Tree baseline on all metrics. However, the corrected resampled t-test showed that these differences were not statistically significant. The comparison between Random Forest and XGBoost yielded $p = 0.711$ for F1-Score and $p = 0.194$ for AUC-ROC; the comparison between Decision Tree and Random Forest yielded $p = 0.412$ (F1-Score) and $p = 0.061$ (AUC-ROC) and the comparison between Decision Tree and XGBoost yielded $p = 0.492$ (F1-Score) and $p = 0.224$ (AUC-ROC). Thus, with the sample size of this study, the three models perform equivalently in statistical terms ($p > 0.05$), and the advantages of Random Forest and XGBoost should be read as tendencies rather than definite differences.

These tendencies are partly consistent with previous studies. Sitio et al. [8] also obtained almost equal accuracy for Random Forest and XGBoost, with XGBoost performing better on recall; a similar pattern is seen in this study, where XGBoost achieved the highest Recall, although the highest F1-Score was obtained by Random Forest. The accuracy advantage of Random Forest is consistent with the finding of [6] that Random Forest outperformed XGBoost with default parameters, given that this study did not perform hyperparameter tuning, and with the finding of [17] that Random Forest was the most accurate model after class balancing with SMOTE. In contrast, [9] reported that XGBoost significantly outperformed Random Forest, whereas in this study the difference between the two models was not statistically significant. These differences should be interpreted with the data size in mind: some of the comparison studies used thousands of student records, for example 4,424 records in [8] and 4,365 academic records in [17], whereas this study used 71 scholarship recipients, which limits its statistical power.

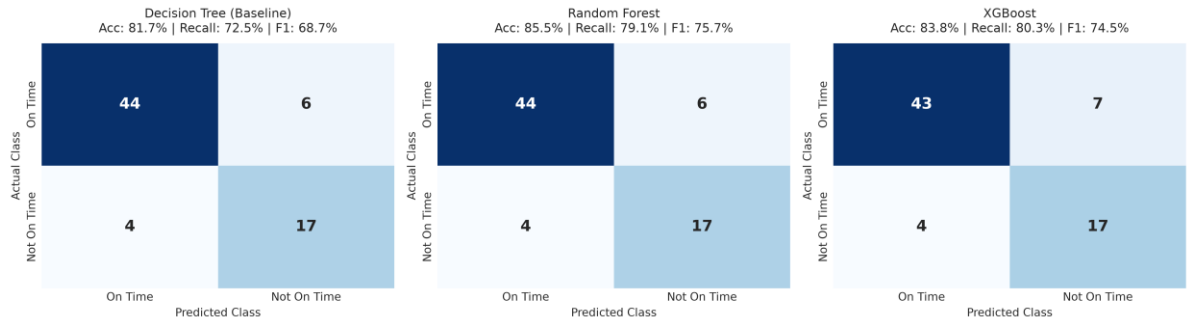


Figure 1. Comparison of Out-of-Fold Confusion Matrices for Decision Tree, Random Forest, and XGBoost

The confusion matrices in Figure 1 show the out-of-fold classification based on the probabilities averaged over 10 repetitions. All three models detected 17 of the 21 Not On Time students, with 4 False Negatives. The difference lies in the majority class: Random Forest and Decision Tree both correctly predicted 44 of the 50 on-time students (6 False Positives), while XGBoost correctly predicted 43 students (7 False Positives).

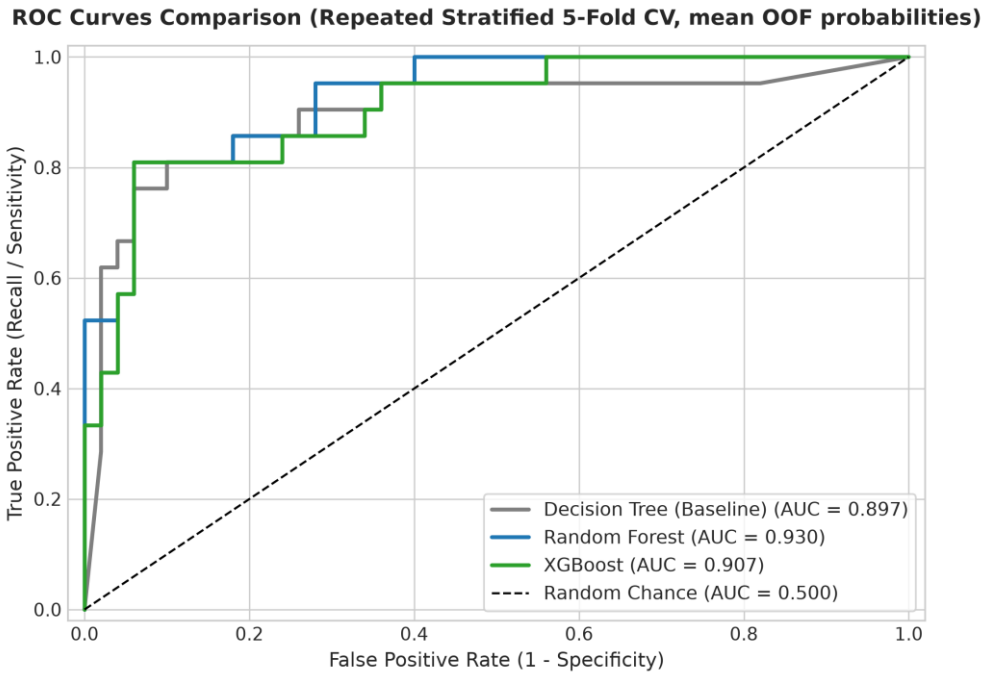


Figure 2. Receiver Operating Characteristic (ROC) Curves and Comparison of Model AUC-ROC Values

The ROC curves in Figure 2 were constructed from the averaged out-of-fold probabilities. Random Forest had the highest discrimination ability (AUC = 0.930), followed by XGBoost (AUC = 0.907) and Decision Tree (AUC = 0.897). The curves of both ensemble models lie close to the upper-left corner, indicating that their risk-probability rankings can be used by scholarship managers to set early warning thresholds.

4.1 Feature Interpretability Analysis Using Tree SHAP

To examine the decision-making mechanism of the ensemble models, a Tree SHAP analysis was performed on the final models trained on the entire training dataset. Figures 3 and 4 show the SHAP summary (beeswarm) plots for Random Forest and XGBoost, respectively, Table 3 presents the feature rankings based on the mean absolute SHAP value (mean |SHAP|), and Figure 5 compares the built-in feature importance of the two models as a reference.

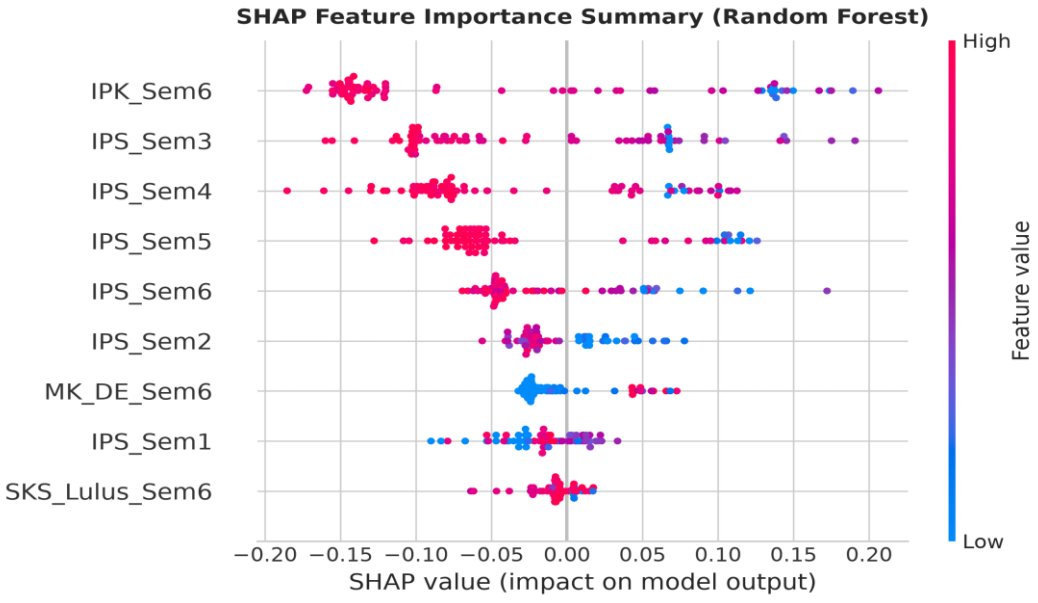


Figure 3. SHAP Summary (Beeswarm) Plot for the Random Forest Model

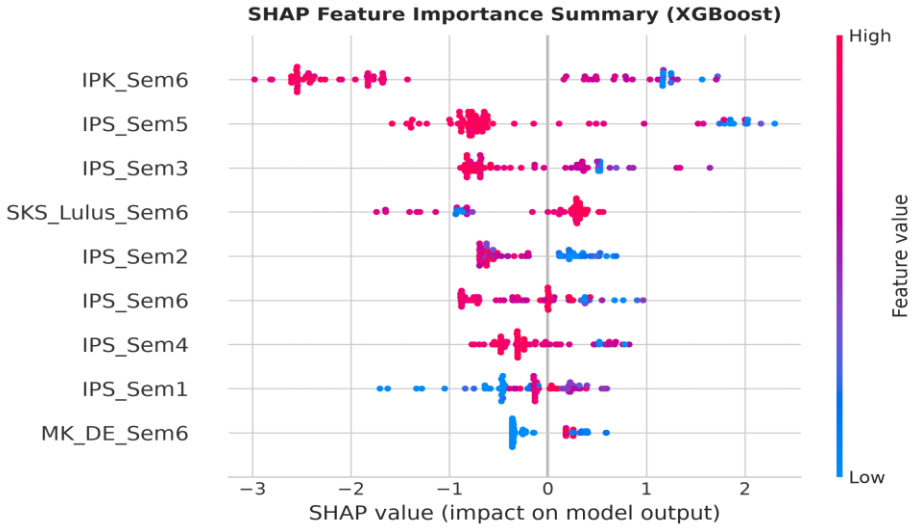


Figure 4. SHAP Summary (Beeswarm) Plot for the XGBoost Model

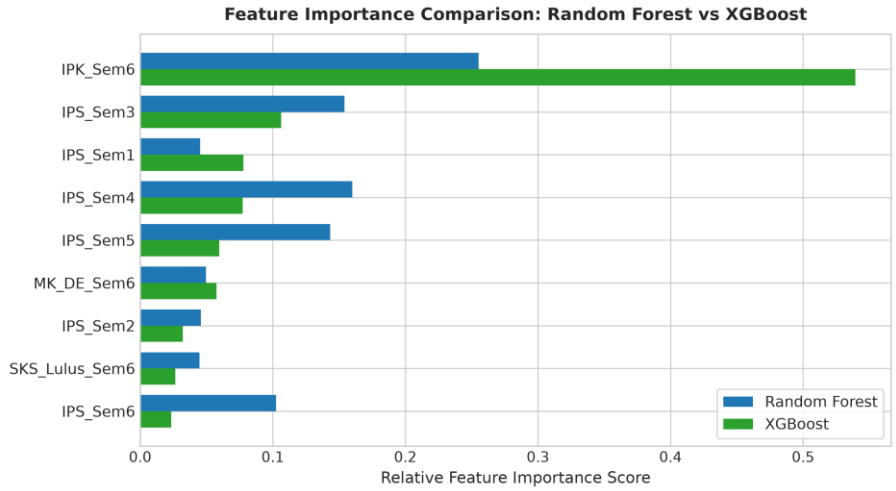


Figure 5. Comparison of Built-in Feature Importance (Gini/Gain) between Random Forest and XGBoost

Table 3. Ranking of Academic Features by Mean Absolute SHAP Value (Mean |SHAP|)

Academic Feature	Mean |SHAP| Random Forest	RF Rank	Mean |SHAP| XGBoost	XGB Rank
Cumulative GPA (Semester 6)	0.1243	1	1.7138	1
Semester 5 GPA	0.0737	4	1.0138	2
Semester 3 GPA	0.0838	2	0.6294	3
Credits Passed (Semester 6)	0.0119	9	0.5185	4
Semester 2 GPA	0.0268	6	0.4978	5
Semester 6 GPA	0.0479	5	0.4668	6
Semester 4 GPA	0.0830	3	0.3974	7
Semester 1 GPA	0.0236	8	0.3786	8
Courses Graded D/E (Semester 6)	0.0264	7	0.3198	9

The SHAP analysis (Table 3 and Figures 3 and 4) provides several pedagogical insights. First, the cumulative GPA at the end of the sixth semester ranks first in both models. As the SHAP plots show, low GPA values have high positive SHAP values; in XGBoost, a sixth-semester cumulative GPA of about 3.43 or lower pushes the model toward predicting a delay. Second, the semester GPAs in the middle of the study period, namely Semesters 3 to 5, rank next (the Semester 5 and Semester 3 GPAs in XGBoost; the Semester 3 and Semester 4 GPAs in Random Forest). The mean semester GPA of the Not On Time group is almost the same as that of the On Time group in the first semester (3.35 and 3.39), then falls to 1.99 in the fifth semester, while the On Time group remains stable at around 3.5–3.7; 10 of the 21 Not On Time students obtained a semester GPA below 2.00 in at least one of Semesters 3 to 6. Monitoring semester GPAs in the middle of the study period is therefore key to early detection. Third, the number of courses graded D/E ranks low, because its information is largely captured by the cumulative and semester GPAs, which decline in the same semesters.

The importance of mid-study semester GPAs is consistent with the findings of [12], who also predicted delayed graduation using academic data up to the sixth semester and found the fourth- and fifth-semester GPAs to be the strongest predictors. Daniawan et al. [12] also found a tuned Random Forest to be the model with the most balanced performance, consistent with Random Forest achieving the highest F1-Score and AUC-ROC in this study.

The SHAP analysis also shows directions of contribution that should be interpreted with caution for two features. Higher Semester 1 GPAs and more credits passed tend to increase the predicted risk, which runs counter to academic intuition. This pattern is presumably a cohort and curriculum effect rather than a causal relationship: for example, almost all students of the 2021 cohort who graduated on time had a Semester 1 GPA of 3.00, while the credit load up to the sixth semester differs between study programs and cohorts. Therefore, these two features are not recommended as a basis for intervention without further analysis.

4.2 Application of the Early Warning System to the Active Cohort (2023 Cohort)

To test the practical usefulness of the model in actual academic operations, the trained ensemble models were applied to the active 2023 cohort dataset, comprising 21 scholarship recipients (12 Foundation Scholarship and 9 KIP-Kuliah). The predicted delay risks for the six students with the highest risk are presented in Table 4.

Table 4. Sample Early Warning Predictions for Scholarship Recipients of the 2023 Cohort

ID	Scholarship Type	GPA Sem 6	Courses D/E	RF Risk (%)	XGB Risk (%)	Predicted Status
MHS-23-18	KIP-Kuliah	3.23	2	88.9%	98.3%	At Risk of Delay
MHS-23-02	Foundation 100%	3.47	0	41.6%	37.3%	Safe (On Time)
MHS-23-21	KIP-Kuliah	3.46	0	12.4%	4.9%	Safe (On Time)
MHS-23-12	Foundation 50%	3.61	1	2.7%	3.9%	Safe (On Time)
MHS-23-13	KIP-Kuliah	3.72	0	7.6%	2.7%	Safe (On Time)
MHS-23-06	Foundation 50%	3.77	0	2.6%	2.6%	Safe (On Time)

The results in Table 4 show that, of the 21 scholarship recipients in the 2023 cohort, 20 students (95.2%) were predicted to be in the SAFE (On-Time Graduation) category, with XGBoost delay-risk probabilities ranging from 0.5% to 37.3%. The system detected 1 student with HIGH RISK OF DELAY, namely MHS-23-18 (a KIP-Kuliah recipient), with a risk probability of 98.3% on XGBoost and 88.9% on Random Forest. This student had the lowest sixth-semester cumulative GPA in the cohort (3.23), a lagging accumulation of credits passed (112 credits against the normal 118 credits), and 2 courses graded D/E that had not yet been

passed. In addition, MHS-23-02 had a moderate risk probability (37.3% on XGBoost and 41.6% on Random Forest) and therefore warrants monitoring even though it is below the 50% threshold. Early detection at the end of the sixth semester gives the student affairs office and academic advisors a window of time to provide intensive guidance promptly and to enroll students in remedial programs or the short semester in the seventh semester, before the eighth-semester graduation deadline.

Conclusion

This study developed and compared the performance of the Random Forest and XGBoost algorithms in predicting the on-time graduation of scholarship recipients at STMIK PN using academic records at the end of the sixth semester. The results of Repeated Stratified 5-Fold Cross-Validation with SMOTE balancing show that Random Forest achieved the highest Accuracy (85.46%), Precision (77.35%), F1-Score (75.71%), and AUC-ROC (0.9405), so its warnings are the most precise, while XGBoost achieved the highest Recall (80.30%), so it misses the fewest at-risk students. Although both ensemble models outperformed the Decision Tree on average, the corrected resampled t-test showed that the differences between the models were not statistically significant ($p > 0.05$). The Tree SHAP analysis revealed that the sixth-semester cumulative GPA is the most dominant factor, followed by the mid-study semester GPAs (Semesters 3 to 5).

The practical implication of these findings for private higher education institutions and scholarship fund managers in supporting SDG 4 and SDG 10 is a proactive capability for study program leaders to identify at-risk students from the end of the third year and promptly design targeted remedial interventions (such as academic clinics, short-semester fee waivers, and accelerated thesis supervision). This study has several limitations: the sample size is relatively small (71 students), which limits statistical power; the Not On Time class combines students who graduated late and those who dropped out; and there are indications of cohort effects and curriculum differences between study programs in the Semester 1 GPA and credits passed features. Future research is recommended to expand the training data across multiple cohorts and campuses, to separate the late-graduation and dropout categories, and to integrate digital learning interaction features (learning management system metrics) to test the generalizability of the model.

References

[1] M. Chankseliani and T. McCowan, “Higher education and the Sustainable Development Goals,” *High. Educ.*, vol. 81, no. 1, pp. 1–8, 2021, doi: 10.1007/s10734-020-00652-w.

[2] E. Herbaut and K. Geven, “What Works to Reduce Inequalities in Higher Education? A Systematic Review of the (Quasi-) Experimental Literature on Outreach and Financial Aid,” *Res. Soc. Stratif. Mobil.*, vol. 65, no. 100442, 2020, doi: 10.1016/j.rssm.2019.100442.

[3] T. Mulyaningsih, S. Dong, R. Miranti, A. Daly, and Y. Purwaningsih, “Targeted scholarship for higher education and academic performance: Evidence from Indonesia,” *Int. J. Educ. Dev.*, vol. 88, no. 102510, 2022, doi: 10.1016/j.ijedudev.2021.102510.

[4] P. T. de F. e Silva and L. M. B. Sampaio, “Does student aid make a degree more likely? Evidence of the permanence scholarship program from survival models,” *Int. J. Educ. Dev.*, vol. 96, p. 102697, 2023, doi: 10.1016/j.ijedudev.2022.102697.

[5] S. Romero and X. Liao, “Statistical and machine learning models for predicting university dropout and scholarship impact,” *PLoS One*, vol. 20, no. 6, pp. 1–19, 2025, doi: 10.1371/journal.pone.0325047.

[6] S. Imron, A. Faizah, and Sugianto, “Prediksi Kelulusan Mahasiswa Menggunakan Algoritma XGBoost,” *SKANIKA Sist. Komput. dan Tek. Inform.*, vol. 9, no. 1, pp. 76–86,

2026, doi: 10.36080/skanika.v9i1.3647.

[7] J. M. Aiken, R. De Bin, M. Hjorth-Jensen, and M. D. Caballero, "Predicting time to graduation at a large enrollment American university," *PLoS One*, vol. 15, no. 11, pp. 1–28, 2020, doi: 10.1371/journal.pone.0242334.

[8] S. L. M. Sitio, Darmawati, and Y. Samudra, "Comparative analysis of bagging and boosting models in ensemble learning for graduation prediction," *JITK (Jurnal Ilmu Pengetah. dan Teknol. Komputer)*, vol. 11, no. 3, pp. 979–986, 2026, doi: 10.33480/jitk.v11i3.7579.

[9] I. Nirmala, H. Wijayanto, and K. A. Notodiputro, "Prediction of Undergraduate Student's Study Completion Status Using MissForest Imputation in Random Forest and XGBoost Models," *ComTech Comput. Math. Eng. Appl.*, vol. 13, no. 1, pp. 53–62, 2022, doi: 10.21512/comtech.v13i1.7388.

[10] F. E. Arévalo-Cordovilla and M. Peña, "Evaluating ensemble models for fair and interpretable prediction in higher education using multimodal data," *Sci. Rep.*, vol. 15, pp. 1–14, 2025, doi: 10.1038/s41598-025-15388-9.

[11] A. J. Fernández-García, J. C. Preciado, F. Melchor, R. Rodriguez-Echeverria, J. M. Conejero, and F. Sánchez-Figueroa, "A Real-Life Machine Learning Experience for Predicting University Dropout at Different Stages Using Academic Data," *IEEE Access*, vol. 9, pp. 133076–133090, 2021, doi: 10.1109/ACCESS.2021.3115851.

[12] B. Daniawan, Suwitno, A. Wijaya, A. R. K. Maranto, and Junaedi, "Machine Learning Based Early Warning Model for Delayed Student Graduation: A Sixth-Semester Prediction Approach," *JUITA J. Inform.*, vol. 14, no. 2, pp. 394–402, 2026, doi: 10.30595/juita.v14i2.30186.

[13] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002, doi: 10.1613/jair.953.

[14] S. M. Lundberg *et al.*, "From Local Explanations to Global Understanding with explainable AI for trees," *Nat. Mach. Intell.*, vol. 2, no. 1, pp. 1–72, 2020, doi: 10.1038/s42256-019-0138-9.

[15] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[16] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, pp. 785–794, 2016, doi: 10.1145/2939672.2939785.

[17] V. Flores, S. Heras, and V. Julian, "Comparison of Predictive Models with Balanced Classes Using the SMOTE Method for the Forecast of Student Dropout in Higher Education," *Electronics*, vol. 11, no. 3, p. 457, 2022, doi: 10.3390/electronics11030457.

[18] H. Sahlaoui, A. E. L. A. Alaoui, A. Nayyar, S. Agoujil, and M. M. Jaber, "Predicting and Interpreting Student Performance Using Ensemble Models and Shapley Additive Explanations," *IEEE Access*, vol. 9, pp. 152688–152703, 2021, doi: 10.1109/ACCESS.2021.3124270.

[19] C. Nadeau and Y. Bengio, "Inference for the Generalization Error," *Mach. Learn.*, vol. 52, no. 3, pp. 307–313, 2020, doi: 10.1023/A:1024068626366.

[20] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets," *PLoS One*, vol. 10, no. 3, p. e0118432, 2015, doi: 10.1371/journal.pone.0118432.