

Student Dropout Risk Prediction After Two Semesters: A Comparative Study Of Logistic Regression, Random Forest, and XGBoost with SHAP

¹Jepri Banjarnahor, ²Very Clevery

^{1,2}Information System Study Program, Universitas Prima Indonesia, Medan, Indonesia
PUI PT ITIK (INOVASI TEKNOLOGI ILMU KOMPUTER)

*Correspondence Author: jepribanjarnahor@unprimdn.ac.id

Submitted: Aug 21, 2026 | Accepted: Sept 24, 2026 | Published: Sept 28, 2026

ABSTRACT

Student dropout requires accurate and explainable early detection so that model outputs can support academic decisions responsibly within an early warning system. This study compares logistic regression, Random Forest, and extreme gradient boosting for student dropout prediction and examines how explainable artificial intelligence clarifies influential predictive factors through SHapley Additive exPlanations (SHAP). The Predict Students' Dropout and Academic Success dataset from the Machine Learning Repository contains 4,424 observations and 36 features. The Enrolled class was excluded because it is not a final outcome, producing a binary dataset of 3,630 observations comprising 1,421 Dropout and 2,209 Graduate records. Data were stratified into 60% training, 20% validation, and 20% test sets. Hyperparameters and classification thresholds were selected on validation data, while the test set was reserved for final evaluation. Logistic regression achieved the highest precision-recall area under the curve of 0.9728 and Dropout recall of 0.9754. Random Forest produced the highest F1-score of 0.8346 and balanced accuracy of 0.8710. Extreme gradient boosting achieved a precision-recall area under the curve of 0.9705, recall of 0.9613, precision of 0.7319, F1-score of 0.8311, and balanced accuracy of 0.8675. SHapley Additive exPlanations (SHAP) ranked study program, second-semester approved curricular units, tuition payment status, and early academic achievement among the dominant features. Extreme gradient boosting is competitive but not universally superior. Local validation, temporal evaluation, fairness assessment, and intervention testing are required before operational deployment.

Keywords: Early Warning System; Explainable Artificial Intelligence; SHapley Additive exPlanations (SHAP); Student Dropout Prediction; XGBoost.

ABSTRAK

Putus sekolah memerlukan deteksi dini yang akurat dan dapat dijelaskan agar hasil model dapat mendukung pengambilan keputusan akademik secara bertanggung jawab dalam sistem peringatan dini. Penelitian ini membandingkan regresi logistik, Random Forest, dan extreme gradient boosting untuk prediksi putus sekolah, serta mengkaji bagaimana kecerdasan buatan yang dapat dijelaskan menguraikan faktor-faktor prediktif yang berpengaruh melalui SHapley Additive exPlanations (SHAP). Dataset "Predict Students' Dropout and Academic Success" dari Machine Learning Repository berisi 4.424 pengamatan dan 36 fitur. Kelas "Enrolled" dikecualikan karena bukan merupakan hasil akhir, sehingga menghasilkan dataset biner dengan 3.630 pengamatan yang terdiri dari 1.421 catatan "Dropout" dan 2.209 catatan "Graduate". Data dibagi menjadi 60% untuk set pelatihan, 20% untuk set validasi, dan 20% untuk set pengujian. Hiperparameter dan ambang batas klasifikasi dipilih berdasarkan data validasi, sedangkan set pengujian disisihkan untuk evaluasi akhir. Regresi logistik mencapai area di bawah kurva presisi-recall tertinggi sebesar 0,9728 dan recall putus sekolah sebesar 0,9754. Random Forest menghasilkan skor F1 tertinggi sebesar 0,8346 dan akurasi seimbang sebesar 0,8710. Extreme Gradient Boosting mencapai area di bawah kurva presisi-recall sebesar 0,9705, recall sebesar 0,9613, presisi sebesar 0,7319, skor F1 sebesar 0,8311, dan akurasi seimbang sebesar 0,8675. Penjelasan aditif Shapley menempatkan program studi, satuan kurikulum yang disetujui pada semester kedua, status pembayaran uang kuliah, dan prestasi akademik awal sebagai fitur-fitur dominan. Extreme gradient boosting kompetitif namun tidak secara universal unggul. Validasi

lokal, evaluasi temporal, penilaian keadilan, dan pengujian intervensi diperlukan sebelum penerapan operasional.

Keywords: *Early Warning System; Explainable Artificial Intelligence; SHapley Additive exPlanations (SHAP); Student Dropout Prediction; XGBoost.*

INTRODUCTION

Machine learning has consequently been used to classify students according to dropout risk. Research has compared conventional classifiers, ensemble approaches, class-imbalance strategies, and optimized models, but no single algorithm performs best in every setting because performance depends on the target definition, feature distribution, institutional context, and validation design (Duro et al. 2026; Flores, Heras, and Julian 2022; Oqaidi, Aouhassi, and Mansouri n.d.). Indonesian studies also show that recall, class imbalance, and hyperparameter optimization can materially change model performance (Alfahrizi et al. 2026; Arthana et al. 2024; Hakim et al. 2026). This makes comparative evaluation important for an EWS intended to be integrated into an Information System rather than assuming that one algorithm is universally appropriate.

The need for an EWS is therefore closely related to the practical role of Information Systems: collecting, processing, interpreting, and presenting information for timely organizational action. In a dropout-risk context, a useful EWS should provide more than a binary prediction. It should produce interpretable risk information, identify supporting predictive factors, and allow academic staff to verify the result before an intervention is considered. Explainable Artificial Intelligence can contribute to this requirement by showing which observed features contribute to a model output. However, an explanation can still accompany an incorrect prediction and therefore requires human verification (Buñay-guisñan et al. 2024; Nagy and Molontay 2024; Saqr and Pernas 2024).

From a System Information perspective, academic information systems, learning management systems, digital attendance platforms, and administrative services continuously generate data describing students' study trajectories. Variables such as completed curricular units, semester grades, learning activity, payment status, scholarships, and admission characteristics have been associated with dropout risk (Gutierrez-pachas et al. 2022; Nagy and Molontay 2024; Romero and Liao 2025). These data create an opportunity to develop an Early Warning System (EWS) in which predictive models convert historical and academic information into risk indicators that can be reviewed by academic advisers. Such a system should support, rather than replace, human decision-making because predictive relationships do not establish that a particular student characteristic causes dropout (Romero and Liao 2025; Saqr and Pernas 2024).

Student persistence is a major concern in higher education because student dropout affects students, institutions, and the efficiency of educational resources. Indonesian Higher Education Statistics for 2025 recorded 289,670 students who left higher education, an increase of 2.62% from 2024. Private higher education institutions accounted for 73.81% of these cases, while students aged 21–30 years formed the largest age group among those leaving their studies (Tinggi 2026). These figures indicate that dropout is not only an academic outcome but also an information-management challenge. Higher education institutions need mechanisms that can transform available student data into timely risk information so that academic staff can review and support students before dropout becomes a final outcome.

This study addresses four related gaps. First, it compares logistic regression, Random Forest, and Extreme Gradient Boosting (XGBoost) using an explicit training-validation-test design. Second, hyperparameters and classification thresholds are selected without using the test set. Third, the selected XGBoost model is explained with SHapley Additive exPlanations (SHAP) at the global feature level. Fourth, the findings are interpreted as a foundation for a human-in-the-loop Early Warning System rather than an automated decision mechanism. The contribution to Information Systems practice is a reproducible prediction-and-explanation pipeline that can be connected to an institutional workflow, where risk probabilities, categories, supporting factors, and verification status can inform academic review. The study therefore focuses not only on model performance, but also on how predictive and explainable outputs can be structured as decision-support information within an academic information system.

RESEARCH METHODOLOGY

Research Design

This study used a quantitative experimental design focused on reproducing a student dropout risk-prediction pipeline. A public dataset was selected so that the data source, processing logic, model parameters, and evaluation outputs could be inspected and reproduced. The experiment produced trained models, evaluation tables, test-set predictions, confidence intervals, and SHAP explanations. Dashboard usability, user acceptance, and intervention impact were not tested because those components were outside the executed experiment. The experimental workflow is summarized in Fig. 1.

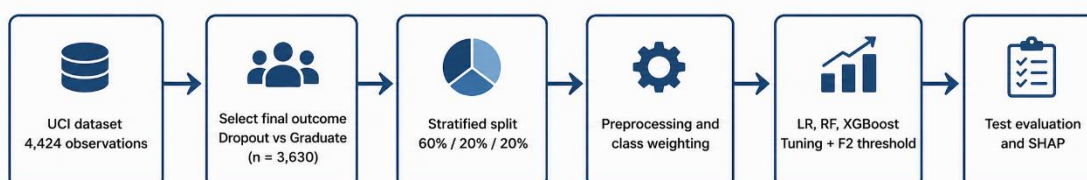


Fig. 1 Experimental workflow for student dropout risk prediction

Dataset and Target Construction

The Predict Students' Dropout and Academic Success dataset was obtained from the UCI Machine Learning Repository under DOI 10.24432/C5MC89 and a CC BY 4.0 license (UC Irvine Machine Learning Repository 2021). The data were collected from students in a Portuguese higher education institution and include multiple study programs (Uci 2026). Each row represents one student. The 36 predictors cover admission characteristics, demographics, socioeconomic information, macroeconomic indicators, and first- and second-semester academic achievement.

The initial audit identified 4,424 observations, no missing values, and no duplicated rows. The original target contained 2,209 Graduate, 1,421 Dropout, and 794 Enrolled records. The Enrolled class was excluded because it does not represent a final outcome. Assigning those records to either Graduate or Dropout would create label error. The binary experiment therefore used 3,630 observations, with Dropout encoded as 1 and Graduate encoded as 0. The predictor set contained 17 categorical and 19 numerical features. Categorical examples included marital status, application mode, course, previous qualification, debtor status, tuition-fee status, and scholarship status. Numerical examples included admission grade, age at enrollment, curricular-unit counts and grades, unemployment, inflation, and gross domestic product.

Because first- and second-semester features were retained, the resulting risk score is conceptually available after the end of the second semester. "Early" prediction in this study therefore means identifying risk after the first two semesters, not before university entry. The dataset does not provide cohort or timestamp variables that permit temporal validation.

Data Split and Leakage Prevention

The binary dataset was divided using stratified sampling with random seed 42. Twenty percent of the data was first reserved as the test set. The remaining 80% was divided again to obtain final proportions of 60% training, 20% validation, and 20% test data. Stratification maintained a Dropout proportion close to 39.1% in each subset. The training set was used to fit model parameters. The validation set was used to select hyperparameters and classification thresholds. The test set was used only once for final evaluation.

All imputation, standardization, and one-hot encoding operations were placed inside pipelines learned from the training set. This design prevented validation or test statistics from entering the training process. Because the dataset does not include cohort identifiers, stratified random splitting cannot test distribution shifts between academic intakes. Test performance therefore represents generalization to a random sample from the same source population rather than temporal generalization.

Preprocessing and Class Imbalance Handling

Numerical features were processed using median imputation. Logistic regression applied z-score standardization after imputation, whereas Random Forest and XGBoost used

numerical values without standardization because tree models do not require a common feature scale. Categorical features used mode imputation followed by one-hot encoding with unknown-category handling enabled.

Class imbalance was handled through weighting rather than oversampling. Logistic regression and Random Forest used balanced class weights. XGBoost used a positive-class weight calculated from the ratio of Graduate to Dropout cases in the training set, equal to 1.553. Weighting retained all original observations and avoided generating synthetic records. Previous studies show that SMOTE can be useful in some dropout settings, but its effect remains data-dependent and should be compared empirically (Arthana et al. 2024; Flores et al. 2022).

Models, Hyperparameter Optimization, and Classification Threshold

Three models were compared. Logistic regression served as an interpretable baseline. Random Forest represented a bagging ensemble. Extreme Gradient Boosting represented a regularized boosting ensemble suited to tabular data (Lin et al. 2025). Candidate hyperparameter combinations were specified before final testing. The best candidate for each model was selected using precision-recall area under the curve on the validation set, with the F2-score as a secondary criterion.

Table 1. Best Hyperparameters Selected on the Validation Set

Model	Selected Hyperparameters	Validation Threshold	F2
Logistic Regression	C=5.0; solver=lbfgs	class_weight=balanced; 0.1767	
Random Forest	n_estimators=500; min_samples_leaf=2; max_features=0.5	max_depth=14; 0.2103	
XGBoost	n_estimators=300; learning_rate=0.05; colsample_bytree=0.85	max_depth=3; subsample=0.85; 0.2280	

The classification threshold was not fixed at 0.50. For each selected candidate, the validation-set precision-recall curve was used to identify the threshold that maximized the F2-score. The F2-score gives more weight to recall than precision, which is appropriate for early screening when missing a truly at-risk student is considered more costly than reviewing an additional false positive. The selected threshold was then locked and applied to the test set.

Evaluation Metrics and Confidence Intervals

The primary metric was precision-recall area under the curve because the analysis focused on the Dropout class. Additional metrics were receiver operating characteristic area under the curve, recall, precision, F1-score, balanced accuracy, Brier score, and the confusion matrix. Recall measured the proportion of actual Dropout cases detected. Precision measured the proportion of Dropout alerts that were correct. Balanced accuracy gave equal importance to both classes. Brier score measured squared probability error, with lower values indicating better probability predictions.

Uncertainty in XGBoost performance was estimated using 1,000 bootstrap resamples of the test set. The 95% confidence interval was obtained from the 2.5th and 97.5th percentiles of the bootstrap distribution. Bootstrapping was performed only after the model and threshold had been locked.

Explainable Artificial Intelligence

The selected XGBoost model was explained using TreeSHAP after model, hyperparameter, and threshold selection were complete (Wang et al. 2024). SHAP values were computed on transformed test data. One-hot encoding creates several columns from a single categorical feature, so the absolute SHAP values of derived columns were summed back to the original feature. Mean absolute SHAP was then used to rank global feature contribution (Givisis, Kalatzis, and Christakis 2025; Kim, Khomidov, and Lee 2025).

SHAP values describe how strongly observed features contribute to model output;

they do not demonstrate causality. Sensitive or difficult-to-act-on variables, such as family characteristics or study program, should therefore be treated as signals for further review rather than grounds for automated treatment (Nti and Ramanayake 2026a; Romero and Liao 2025; Saqr and Pernas 2024).

Reproducibility and Data Ethics

The experiment used random seed 42 and was executed with Python 3.13.5, pandas 2.2.3, NumPy 2.3.5, scikit-learn 1.8.0, XGBoost 3.1.3, SHAP 0.50.0, SciPy 1.17.0, and joblib 1.5.3. The Google Colab notebook, Python script, dataset, trained model, test predictions, selected parameters, confidence intervals, and SHAP outputs were retained as reproducibility materials.

The public dataset does not contain names or direct identity numbers. Any application to institutional data should require institutional authorization, pseudonymization, purpose limitation, access control, and appropriate safeguards under Indonesian Personal Data Protection Law No. 27 of 2022 (Republik 2022). Risk scores should not be used as an automatic basis for sanctions or service termination

RESULT AND DISSCUSSIONS

Data Characteristics and Sample Split

The original data audit confirmed 4,424 observations and 36 predictors with no missing values or duplicated rows. Graduate comprised 2,209 observations (49.93%), Dropout comprised 1,421 observations (32.12%), and Enrolled comprised 794 observations (17.95%). After Enrolled records were removed, the Dropout proportion in the binary dataset was 39.15%. Stratification maintained a similar class distribution in the training, validation, and test subsets. The test set contained 726 students, including 284 Dropout and 442 Graduate cases.

Table 2. Data Distribution in Each Subset

Subset	Total	Dropout	Graduate	Dropout Proportion
Training	2,178	853	1,325	39.16%
Validation	726	284	442	39.12%
Test	726	284	442	39.12%

Model Performance

All three models produced high discrimination on the test set. Logistic regression achieved the highest precision-recall area under the curve at 0.9728 and the highest Dropout recall at 0.9754. It detected 277 of 284 Dropout cases, but precision was 0.6707 because 136 Graduate cases were classified as Dropout. Random Forest produced the highest F1-score at 0.8346, the highest balanced accuracy at 0.8710, and the lowest Brier score at 0.0609. XGBoost was close to Random Forest with an F1-score of 0.8311, balanced accuracy of 0.8675, and Brier score of 0.0614.

Table 3. Model Performance on the Test Set

Model	PR-AUC	ROC-AUC	Recall	Precision	F1	Balanced Acc.	Brier
Logistic Regression	0.9728	0.9755	0.9754	0.6707	0.7948	0.8338	0.0628
Random Forest	0.9679	0.9726	0.9683	0.7333	0.8346	0.8710	0.0609
XGBoost	0.9705	0.9723	0.9613	0.7319	0.8311	0.8675	0.0614

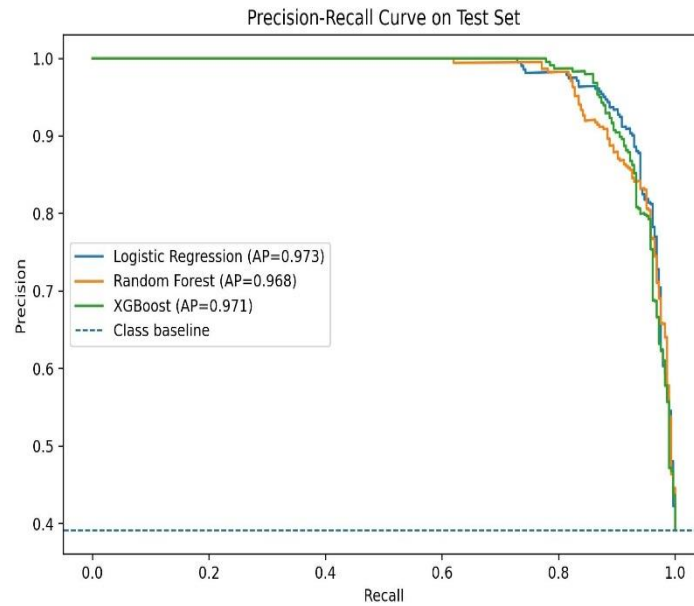


Fig. 2 Precision-recall curves for the three models on the test set

The precision-recall curves in Fig. 2 remained well above the positive-class baseline of approximately 0.391 across most of the recall range. The three curves were close to one another, consistent with their similar precision-recall area under the curve values.

XGBoost Confusion Matrix

XGBoost was retained for explainability analysis because it matched the study objective and has an efficient TreeSHAP implementation. At the validation-selected threshold of 0.2280, the test-set confusion matrix contained 342 true negatives, 100 false positives, 11 false negatives, and 273 true positives. Overall accuracy calculated from the confusion matrix was 84.71%. Dropout recall was 0.9613, while Dropout precision was 0.7319.

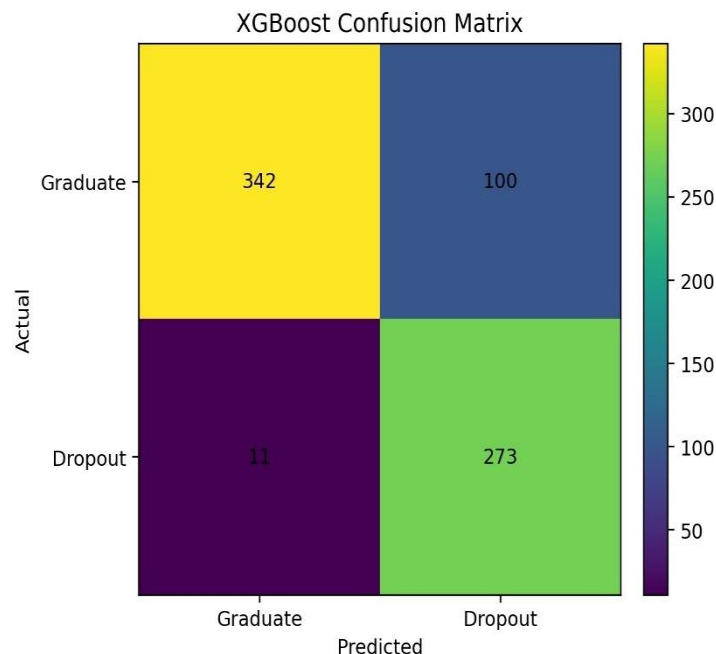


Fig. 3 XGBoost confusion matrix on the test set

Performance Uncertainty

Bootstrap resampling showed that XGBoost performance remained high within the test sample. The 95% confidence interval for precision-recall area under the curve was 0.9578–0.9813, and the interval for Dropout recall was 0.9387–0.9836. The corresponding intervals for the other reported metrics are presented in Table 4.

Table 4. XGBoost 95% Bootstrap Confidence Intervals

Metric	Estimate	Lower Bound	Upper Bound
PR-AUC	0.9705	0.9578	0.9813
ROC-AUC	0.9723	0.9596	0.9839
Dropout Recall	0.9613	0.9387	0.9836
Dropout Precision	0.7319	0.6857	0.7735
Dropout F1	0.8311	0.7994	0.8600
Balanced Accuracy	0.8675	0.8427	0.8895
Brier Score	0.0614	0.0513	0.0726

Dominant Features Based on SHAP

Mean absolute SHAP ranked Course as the largest aggregated contributor, followed by Curricular units 2nd sem (approved), Tuition fees up to date, Father's occupation, Father's qualification, and Curricular units 1st sem (approved). Other highly ranked variables included Mother's occupation, first-semester evaluations, second-semester grade, application mode, scholarship status, admission grade, and second-semester evaluations. The fifteen highest-ranked original features and their mean absolute SHAP values are shown in Table 5 and Fig. 4.

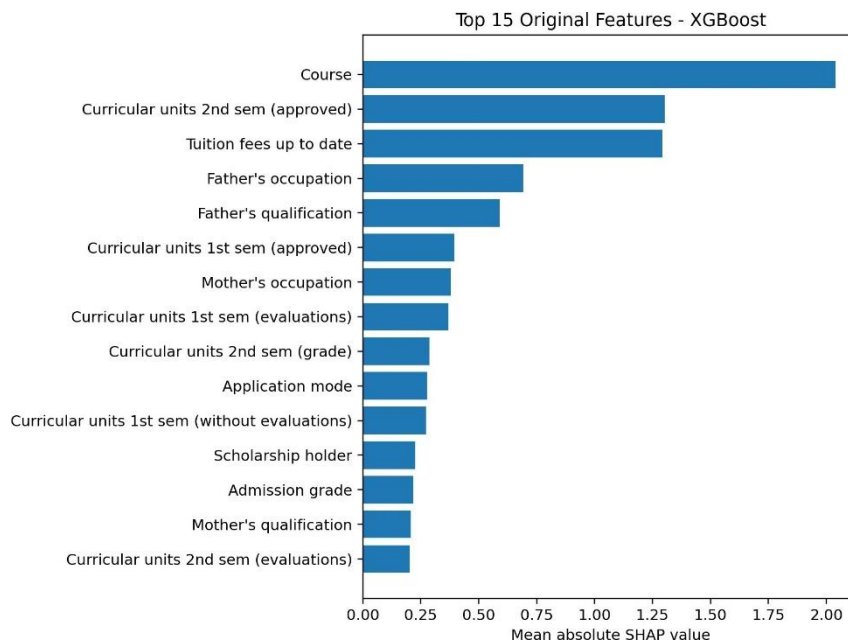


Fig. 4 Fifteen original features with the highest mean absolute SHAP values

Table 5. Fifteen Dominant Features Based on Mean Absolute SHAP

Rank	Feature	Mean SHAP
1	Course	2.0388
2	Curricular units 2nd sem (approved)	1.3022
3	Tuition fees up to date	1.2919
4	Father's occupation	0.6927
5	Father's qualification	0.5900
6	Curricular units 1st sem (approved)	0.3944
7	Mother's occupation	0.3808
8	Curricular units 1st sem (evaluations)	0.3700
9	Curricular units 2nd sem (grade)	0.2878

Rank	Feature	Mean SHAP
10	Application mode	0.2762
11	Curricular units 1st sem (without evaluations)	0.2716
12	Scholarship holder	0.2270
13	Admission grade	0.2169
14	Mother's qualification	0.2065
15	Curricular units 2nd sem (evaluations)	0.2032

Discussions

Comparative Interpretation

The results show that model choice depends on the operational objective rather than on a single ranking. Logistic regression was strongest when the priority was to maximize detection because it produced the highest Dropout recall and precision-recall area under the curve. Random Forest produced the best balance between precision and recall, as indicated by the highest F1-score and balanced accuracy. XGBoost was highly competitive but did not dominate every metric. This pattern is consistent with Putra et al. (2025), who found that Random Forest could slightly outperform XGBoost on the same public dataset. Direct numerical comparison remains inappropriate because the present experiment used a binary target, a separate validation set, class weighting, hyperparameter optimization, and an F2-based threshold. Broader reviews likewise report that model performance changes with target construction, features, class distribution, and evaluation design (Duro et al. 2026; Saqr and Pernas 2024).

The XGBoost confusion matrix illustrates the operational trade-off. Only 11 of 284 Dropout cases were missed, but 100 Graduate cases were flagged as Dropout. A false-positive rate of this magnitude may be acceptable when the first response is a low-cost academic review, such as a check by an academic adviser. It would be less appropriate if every alert triggered an expensive intervention or a decision that could stigmatize students. Explainability should therefore be shown together with model uncertainty and error patterns. Saqr and (Saqr and Pernas 2024) emphasize that an apparently reasonable explanation can accompany a wrong prediction, so human verification remains necessary.

Interpretation of SHAP Features

The SHAP analysis placed Course at the top of the global ranking, but this result requires caution because the reported contribution aggregates multiple one-hot encoded categories. A high aggregated contribution does not mean that a particular study program causes dropout. Course may act as a proxy for differences in curriculum, admission selection, academic load, or student composition that are not fully represented by other predictors. The same caution applies to parents' occupation and qualification. These variables should not be used as a basis for restricting services or assigning adverse treatment.

More actionable predictors included approved curricular units in the first and second semesters, second-semester grades, tuition payment status, evaluation activity, and scholarship status. This pattern is consistent with studies that identify early academic achievement and financial conditions as important predictive signals (Bettahi, Belouadha, and Harroud 2025; Nagy and Molontay 2024; Nti and Ramanayake 2026b; Romero and Liao 2025). SHAP can therefore support interpretation by showing which observed features increase or decrease a risk score, but the explanation remains predictive rather than causal. Any mapping from a feature to a specific intervention should be validated with academic managers and students.

Implications for an Early Warning System

In this workflow, the model functions as an information-processing component within the SIA: student data are collected from institutional sources, the model generates a risk probability, the threshold determines whether an alert is created, SHAP provides supporting explanatory information, and authorized staff perform verification and decide whether a support action is appropriate. The present experiment, however, did not build or test an actual dashboard, user-acceptance process, or intervention workflow. Therefore, the proposed SIA integration is a practical implementation framework derived from the experimental results, not evidence that such a workflow has already improved student retention. Any operational deployment should

first be validated with local pseudonymized data and evaluated for fairness, probability calibration, explanation stability, user acceptance, and intervention outcomes.

The alert generated by the SIA should be routed to an authorized academic adviser or designated student-support officer who is responsible for reviewing the case. The system can present the student's risk probability together with the main supporting predictive factors identified by SHAP, relevant academic indicators, and the date of the latest available data. The receiving staff member should then verify the alert against contextual information that may not be represented in the model, such as recent academic circumstances or other institutional information. The verification result can be recorded in the SIA as a status such as reviewed, confirmed for follow-up, requires additional information, or not confirmed, together with an appropriate follow-up note. This workflow preserves the human-in-the-loop approach described in the study and avoids treating a model prediction as proof that a student will drop out.

For the XGBoost model, the validation-selected classification threshold of 0.2280 can be operationalized as the alert rule: a predicted dropout probability at or above 0.2280 is flagged for academic review, while a probability below 0.2280 is not flagged under this threshold. The threshold was selected on the validation set by maximizing F2-score and was then locked before evaluation on the test set. In an operational SIA, this rule should be implemented in the prediction service or scoring layer, with the model producing a probability, risk category, supporting factors, and verification status rather than only a Dropout label. The threshold should not be interpreted as a universal institutional cutoff because adviser capacity, intervention cost, local validation results, and the consequences of false positives and false negatives may differ across institutions.

The findings suggest different use cases for the three models in an Academic Information System (SIA)-based Early Warning System. Logistic regression is suitable when the institution prioritizes sensitivity and has enough capacity to review more false positives. Random Forest offers the strongest balance between the two classes in this experiment. XGBoost provides nearly the same predictive performance while supporting efficient TreeSHAP explanations. From an SIA implementation perspective, the prediction should be treated as a decision-support alert rather than an automatic administrative decision.

Answers to the Research Questions

The experimental evidence answers the study questions as follows. The dominant predictors were study program, second-semester approved curricular units, tuition payment status, parental background variables, and early academic achievement, with academic, payment, evaluation, and scholarship variables being the most actionable. XGBoost was not universally better than the comparison models: logistic regression led on precision-recall area under the curve and recall, Random Forest led on F1-score and balanced accuracy, and XGBoost remained very close to the best model. SHAP provided a global ranking of feature contributions and can also support individual explanations, but its values must be interpreted as predictive contributions rather than causal effects. The pipeline can provide probabilities, labels, stored models, confusion matrices, and explanations as a technical foundation for a dashboard, while user acceptance, operational intervention, and retention impact remain untested.

Limitations and Threats to Validity

Several limitations constrain the interpretation of the findings. The dataset comes from one institution in Portugal, so its study-program distribution, academic policies, and socioeconomic conditions are not identical to Indonesian higher education. The absence of cohort or time markers means that stratified random splitting does not establish temporal generalization. The Enrolled class was excluded because it has no final outcome, so the binary model does not address three-class prediction or time-to-event analysis. The validation threshold was optimized for F2 rather than for measured intervention cost or adviser capacity. SHAP analysis did not test explanation stability, subgroup fairness, or user comprehension. The study also did not build and evaluate a dashboard, conduct user acceptance testing, or test whether model-supported intervention improves retention. Finally, Brier score was reported, but additional probability calibration methods such as isotonic regression or Platt scaling were not evaluated.

CONCLUSION

Importantly, this study has not demonstrated an improvement in student retention. The experiment evaluated model discrimination, classification performance, uncertainty, and feature contributions, but it did not implement or experimentally evaluate interventions, compare retention outcomes before and after deployment, or conduct a controlled evaluation of student-support services. Therefore, the evidence supports the conclusion that the proposed models can predict dropout risk under the study design, not that their use increases student retention. Before operational use in Indonesian higher education, the models should be validated on local pseudonymized data, tested across cohorts, assessed for fairness and probability calibration, and connected to interventions whose effects on actual retention outcomes are measured. Future work should also evaluate dashboard usability, user acceptance, explanation stability, and the effectiveness of different threshold policies.

SHAP identified study program, second-semester curricular completion, tuition payment status, and early academic achievement as major predictive contributors. These academic and financial signals can support review and referral to student services, but SHAP explanations do not establish causality and should not become a basis for automated adverse decisions. The main contribution is a reproducible predictive and explainability pipeline that can serve as a technical foundation for an Early Warning System with human verification.

This study compared logistic regression, Random Forest, and XGBoost on 3,630 binary observations derived from the UCI Predict Students' Dropout and Academic Success dataset. The training-validation-test design ensured that hyperparameters and classification thresholds were selected without using the final test set. All models achieved strong predictive performance, but their strengths differed by metric. Logistic regression achieved the highest precision-recall area under the curve and Dropout recall, while Random Forest achieved the highest F1-score and balanced accuracy. XGBoost remained highly competitive, with a precision-recall area under the curve of 0.9705, recall of 0.9613, precision of 0.7319, F1-score of 0.8311, and balanced accuracy of 0.8675. These findings demonstrate predictive performance under the experimental conditions; they do not establish that one model is universally superior or that the model itself improves student outcomes.

ACKNOWLEDGMENT

The study received no specific funding from public, commercial, or nonprofit agencies. The authors declare no conflict of interest. The main dataset is publicly available from the UCI Machine Learning Repository under DOI 10.24432/C5MC89 and CC BY 4.0. The reproducibility package described in the method section contains the notebook, scripts, processed data, trained model, test predictions, confidence intervals, and SHAP outputs.

REFERENCES

- Alfahrizi, Abiyya, Putra Arifiansyah, Muhammad Afandi, and Dodi Dwi Riskianto. 2026. "Prediksi Depresi Mahasiswa Menggunakan Algoritma Random Forest Berbasis Data Psikososial Depression Prediction Among University Students Using a Random Forest Algorithm Based on Psychosocial Data." 4(1):12–21.
- Arthana, I. Ketut Resika, I. Made Dendi Maysanjaya, Gede Aditra Pradnyana, and Gede Rasben Dantes. 2024. "Optimizing Dropout Prediction in University Using Oversampling Techniques for Imbalanced Datasets." 14(8):1052–60. doi:10.18178/ijiet.2024.14.8.2133.
- Bettahi, Abdelkarim, Fatima-zahra Belouadha, and Hamid Harroud. 2025. "A Modular and Explainable Machine Learning Pipeline for Student Dropout Prediction in Higher Education." 1–31.
- Buñay-guisñan, Pamela, Alfonso Lara, Alberto Cano, Rebeca Cerezo, and Cristóbal Ro-. 2024. "Easing the Prediction of Student Dropout for Everyone In- Tegration AutoML and Explainable Artificial Intelligence."
- Duro, Beatriz, Anabela Gomes, Fernanda Brito Correia, and Ana Rosa Borges. 2026. "Machine Learning and Deep Learning for Dropout Prediction in Higher Education : A Review." 1–26.
- Flores, Vaneza, Stella Heras, and Vicente Julian. 2022. "Comparison of Predictive Models with Balanced Classes Using the SMOTE Method for the Forecast of Student Dropout in Higher Education."
- Givisis, Ioannis, Dimitris Kalatzis, and Christos Christakis. 2025. "Comparing Explainable AI

- Models: SHAP , LIME , and Their Role in Electric Field Strength Prediction over Urban Areas." 1–34.
- Gutierrez-pachas, Daniel A., Germain Garcia-zanabria, Ernesto Cuadros-vargas, Guillermo Camara-chavez, and Erick Gomez-nieto. 2022. "Education Sciences Supporting Decision-Making Process on Higher Education Dropout by Analyzing Academic , Socioeconomic , and Equity Factors through Machine Learning and Survival Analysis Methods in the Latin American Context."
- Hakim, M. Faris Al, Siti Wahyuni, Kholiq Budiman, Aditya Marianti, and Bambang Eko Susilo. 2026. "Optimization of Machine Learning Model Using Grid and Random Search Algorithms for Predicting Student Dropout." 7(3):3012–24.
- Kim, Hojeong, Mavlonbek Khomidov, and Jong-ha Lee. 2025. "XGBoost and SHAP-Based Analysis of Risk Factors for Hypertension Classification in Korean Postmenopausal Women." 1–18.
- Lin, Yingying, Jingqi Gao, Linfang Chen, Yixiao Hong, Min Li, and Peiling Chen. 2025. "Informatics in Medicine Unlocked An Interpretable XGBoost Model for Risk Prediction of Progression from Sepsis-Associated Acute Kidney Injury to Chronic Kidney Disease." *Informatics in Medicine Unlocked* 58(August):101685. doi:10.1016/j.imu.2025.101685.
- Nagy, Marcell, and Roland Molontay. 2024. "Interpretable Dropout Prediction : Towards XAI - Based Personalized Intervention." *International Journal of Artificial Intelligence in Education* 34(2):274–300. doi:10.1007/s40593-023-00331-8.
- Nti, Isaac Kofi, and Selena Ramanayake. 2026a. "Discover Artificial Intelligence Article in Press Explainable Machine Learning for Student Dropout Prediction and Tailored Interventions in Online Personalized Education IN IN."
- Nti, Isaac Kofi, and Selena Ramanayake. 2026b. "Explainable Machine Learning for Student Dropout Prediction and Tailored Interventions in Online Personalized Education." 6.
- Oqaidi, Khalid, Sarah Aouhassi, and Khalifa Mansouri. n.d. "Towards a Students ' Dropout Prediction Model in Higher Education Institutions Using Machine Learning Algorithms." 17(18):103–17.
- Republik, Negara. 2022. "Data Pribadi." (016999).
- Romero, Stephan, and Xiyue Liao. 2025. "For Predicting University Dropout And." 1–19. doi:10.1371/journal.pone.0325047.
- Saqr, Mohammed, and Sonsoles López Pernas. 2024. "Why Explainable AI May Not Be Enough : Predictions and Mispredictions in Decision Making in Education." *Smart Learning Environments*. doi:10.1186/s40561-024-00343-4.
- Tinggi, Kementerian Pendidikan. 2026. "Statistik Pendidikan Tinggi." (June).
- UC Irvine Machine Learning Repository. 2021. "Predict Students' Dropout and Academic Success." <https://archive.ics.uci.edu/dataset/697/predict+students+dropout+and+academic+success>.
- Uci, Dataset. 2026. "Daftar File Data Dan Referensi Pdf." 11(3):2023–24. doi:10.24432/C5MC89.
- Wang, Huanjing, Qianxin Liang, John T. Hancock, and Taghi M. Khoshgoftaar. 2024. "Feature Selection Strategies: A Comparative Analysis of SHAP - Value and Importance - Based Methods." *Journal of Big Data*. doi:10.1186/s40537-024-00905-w.