Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 1, January 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15703
e-ISSN : 2541-2019
p-ISSN : 2541-044X

From Prediction to Targeting: Comparative ML Models and Threshold-Based Re-Enrollment Segmentation in Higher Education Marketing

Akto Hariawan[1)], Arif Mu'amar Wahid[2)], Sultan Ananda Haikal[3)], Prayoga Pribadi[4)]
[1,3)]Informatics Department, Computer Sciences Faculty, Universitas Amikom Purwokerto, Indonesia, [2)] Graduate School of Natural and Science Technology, Kanazawa University, Japan, [4)] Digital Business Department, Business and Social Sciences Faculty, Universitas Amikom Purwokerto, Indonesia
[1)] akto.85@gmail.com, [2)] arifmuamar@stu.kanazawa-u.ac.jp, [3)] 24sa11a241@students.amikompurwokerto.ac.id, [4)] yoga@amikompurwokerto.ac.id

Submitted :Dec 14, 2025 | **Accepted** : Dec 29, 2025 | **Published** : Feb 05, 2026

Abstract: Student re-enrollment is a critical strategic concern for higher education institutions, directly impacting financial stability, capacity planning, and the effectiveness of retention programs. This study develops a decision-support approach to predict non–re-enrollment using institutional records from a private university in Central Java. The dataset includes 2,673 student records across three academic years, which were expanded into 1,099 engineered features through a unified preprocessing pipeline involving missing-value imputation, scaling, and one-hot encoding. Model development utilized a fixed train–test split with 5-fold cross-validation. Results demonstrate that tuned tree ensembles significantly outperform the linear baseline. While Logistic Regression yielded limited discrimination (test ROC-AUC = 0.5602, F1 = 0.7010), tuned Random Forest improved classification quality (test ROC-AUC = 0.7571, F1 = 0.8052). Tuned XGBoost achieved the strongest ranking performance (test ROC-AUC = 0.7606) and was selected for deployment due to its superior risk-ordering capability. SHAP-based interpretation identifies parental income as the dominant driver of non–re-enrollment risk, followed by program-choice indicators and demographic variables. Finally, threshold analysis supports risk-tier segmentation, translating predicted probabilities into practical outreach policies aligned with institutional capacity constraints—addressing two underexplored gaps in applied re-enrollment prediction: rigorous cross-validated ensemble modeling and the integration of predictive scores into actionable marketing segmentation, and highlighting that ranking quality—not classification alone—is essential for operational targeting.

Keywords: Student Re-enrollment; Machine Learning; Logistic Regression; Random Forest; XGBoost

INTRODUCTION

Student enrollment and re-enrollment decisions are increasingly strategic concerns for higher education institutions. Accurate enrollment prediction supports evidence-based planning for academic staffing, class capacity, budgeting, and student-support services, particularly when demand patterns and applicant profiles shift over time. Recent studies emphasize that prediction systems help institutions make timely admission and retention decisions and anticipate resource constraints that emerge when enrollment growth strains institutional capacity (Maphosa et al., 2023; Czibula et al., 2022). Beyond operational planning, predictive analytics can contribute to quality improvement by enabling early identification of students who may require targeted support and by informing institutional policies that strengthen student success outcomes (Rafique et al., 2021; Pelima et al., 2024; Alwarthan et al., 2022). In this sense, enrollment prediction is not only an administrative forecasting task but also an analytical foundation for improving institutional effectiveness.

A broad body of work in educational data mining and machine learning shows that predictive performance depends strongly on the availability and quality of features used to model student behavior and decision-making.

name of corresponding author

e-ISSN : 2541-2019
p-ISSN : 2541-044X

Common predictors include academic history, quizzes and examinations, demographic variables, learning management system (LMS) usage, and indicators of engagement or social interaction (Mengash, 2020). Machine learning classifiers—such as random forests, support vector machines, and neural networks—are often adopted because they can learn nonlinear relationships and interactions among heterogeneous student attributes (Rafique et al., 2021; Pelima et al., 2024). In addition, time-series and hybrid forecasting techniques have been proposed for short-term enrollment and load predictions when enrollment exhibits nonstationary patterns (Wan, 2024; Yi et al., 2021; Yang & Yang, 2020). Collectively, these studies motivate the use of modern predictive modeling pipelines that combine robust preprocessing, model comparison, and interpretable outputs to support institutional decision-making.

However, enrollment and re-enrollment prediction remains challenging due to the complex and heterogeneous factors that shape prospective students' decisions. Individual-level attributes (e.g., prior achievement, behavioral engagement, and personal background) and socio-economic factors (e.g., family income, parental education, affordability, and district-level conditions) materially influence enrollment intentions and program choice (Chen et al., 2023; Singh & Alhulail, 2022; Melak & Singh, 2021; Berges et al., 2021; Setiawan et al., 2024). These influences are frequently under-represented or only partially captured in modeling, which can reduce predictive validity and limit actionable interpretation (Mengash, 2020; Chen et al., 2023). In practical settings, additional modeling difficulties commonly arise, including class imbalance (when one outcome dominates), overfitting in high-dimensional feature spaces, and inadequate validation procedures that overstate real-world performance (Ghorbani & Ghousi, 2020; Roy & Farid, 2024; Charte, 2020; Alhazmi & Sheneamer, 2023). As a result, rigorous evaluation protocols and careful model selection are essential to ensure that predictive systems generalize to unseen cohorts.

In parallel, effective marketing strategies play a critical role in influencing student conversion and re-enrollment, and predictive models are most valuable when they translate into actionable targeting. Data-driven marketing in higher education can improve outreach efficiency by segmenting prospective students into meaningful groups and tailoring messaging and channels accordingly (Rasool et al., 2020). Marketing effectiveness is also strengthened when promotional narratives align with student motivations such as career relevance, labor-market value, or program fit, and when decision-support systems reduce friction in the application pathway (Fahim et al., 2021; Fernández-García et al., 2020; Mengash, 2020). Moreover, affordability and access-related concerns can meaningfully shape conversion rates, implying that institutional offers, scholarship communication, and pricing narratives should be aligned with segment-specific constraints (Setiawan et al., 2024). Therefore, an integrated approach that connects predictive modeling outputs with segmentation and targeting decisions can help institutions deploy marketing resources more efficiently and ethically. While ML-based enrollment prediction is widely studied, prior work often reports classification performance without translating predicted probabilities into actionable segmentation and threshold policies for marketing outreach. Accordingly, this study moves beyond prediction by demonstrating how risk ranking and threshold trade-offs can be used to prioritize retention actions under institutional capacity constraints.

This study proposes an integrated ML–Marketing Strategy Framework that combines rigorous predictive modeling with actionable enrollment segmentation and threshold-based targeting to support data-driven admissions and re-enrollment campaigns. Concretely, a standardized preprocessing pipeline is implemented for mixed-type, high-dimensional inputs, including imputation, scaling for numerical variables, and encoding for categorical and ordinal attributes. Model performance is then assessed using a stratified train–test split to preserve the observed class distribution, while extended hyperparameter optimization for Random Forest and XGBoost is conducted via randomized search with internal cross-validation on the training data to support fair comparison and reduce overfitting. To keep trade-offs transparent for operational use, results are summarized in a model-comparison table reporting Accuracy, Precision, Recall, F1-score, ROC-AUC, and training time. Predictive probabilities are subsequently translated into marketing action through ROC/threshold analysis, enabling threshold-based targeting policies for outreach under different budget constraints, and explanatory insight is derived from model coefficients and feature-importance rankings to highlight influential predictors of re-enrollment outcomes.

To operationalize these contributions, several classifiers are evaluated under a unified preprocessing pipeline, including Logistic Regression as a baseline and tuned tree-based ensembles (Random Forest and XGBoost). Hyperparameter optimization for the ensemble models is carried out using randomized search with cross-validation on the training data, while final generalization performance is assessed on a held-out test set created using a stratified split. Model selection prioritizes balanced performance with emphasis on identifying did not re-enrolled cases for risk-focused outreach, while ROC-AUC is used to assess ranking quality across thresholds. Threshold metrics are then used to formalize marketing operating points that reflect explicit trade-offs between false positives (unnecessary outreach) and false negatives (missed at-risk candidates).

name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 1, January 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15703
e-ISSN : 2541-2019
p-ISSN : 2541-044X

LITERATURE REVIEW

Student Enrollment Prediction and Educational Data Mining

Student enrollment prediction is increasingly important for higher-education planning because institutions must make decisions about admissions capacity, staffing, facilities, and retention support under changing demand (Maphosa et al., 2023; Czibula et al., 2022). Much of the recent work treats enrollment-related outcomes as predictive modeling problems within educational data mining (EDM). Common predictors include academic records, assessment results, demographics, learning management system (LMS) activity, and social interaction signals, while pre-admission test information is sometimes underused despite potential value (Mengash, 2020). To improve predictive performance and support earlier identification of at-risk learners, studies frequently apply classifiers such as random forests, support vector machines, and neural networks, with increasing emphasis on explainable outputs that can be translated into interventions (Rafique et al., 2021; Pelima et al., 2024; Alwarthan et al., 2022).

Logistic regression remains a common baseline because it is simple and interpretable for estimating enrollment likelihood (Esquivel & Esquivel, 2021). However, when relationships are nonlinear or involve interactions, tree-based models and ensembles often perform better. Decision trees can capture rule-like patterns but may overfit, motivating random forests and other ensembles that improve robustness and provide feature-importance estimates (Raghavendran et al., 2021; Wu et al., 2023; Couronné et al., 2018). Gradient boosting methods such as XGBoost are also widely used due to strong predictive performance when appropriately tuned and validated (Farhood, 2024).

For short-term projections of cohort size, course demand, or resource load, time-series and hybrid forecasting approaches (including LSTM-based and ensemble pipelines) can improve predictions when integrated with EDM workflows (Wan, 2024; Yi et al., 2021; Yang & Yang, 2020). Performance can still be undermined by applicant heterogeneity and nonstationarity, where historical patterns shift due to demographic or economic changes; the literature therefore highlights adaptive modeling, continuous updating, and rigorous evaluation (Parmezan et al., 2022). Practical issues such as class imbalance, overfitting, and limited external validity are also common, motivating resampling strategies, feature selection, and systematic validation protocols (Ghorbani & Ghousi, 2020; Roy & Farid, 2024; Charte, 2020; Alhazmi & Sheneamer, 2023).

Across studies, individual and socioeconomic factors repeatedly emerge as key drivers of enrollment decisions. Individual attributes (e.g., academic achievement and behavioral indicators) and socioeconomic determinants (e.g., family income, parental education, affordability constraints, district-level conditions, and pricing) are frequently linked to program choice and enrollment demand (Chen et al., 2023; Singh & Alhulail, 2022; Melak & Singh, 2021; Berges et al., 2021; Setiawan et al., 2024). As a result, contemporary systems increasingly integrate diverse covariates and emphasize interpretable outputs to support evidence-based admissions and retention decisions (Rafique et al., 2021; Pelima et al., 2024; Alwarthan et al., 2022).

College Choice, Pricing, and Attrition as a Conceptual Bridge

Classic higher-education theory explains enrollment outcomes as the result of a staged decision process shaped by information, perceived fit, and constraints. The three-phase college-choice model frames choice as moving from predisposition to search and finally choice, implying that signals encountered during recruitment (information access, perceived program fit, and perceived value) can shift decisions across stages rather than only at the end (Hossler & Gallagher, 1987). From a marketing perspective, educational institutions are also positioned as service organizations that must understand prospective students as segmented audiences and communicate value propositions through purposeful strategy, messaging, and channel design (Kotler & Fox, 1995). Together, these foundations support treating recruitment and conversion as a decision-support problem in which institutions diagnose barriers and tailor interventions, not merely estimate a single enrollment probability.

Affordability and persistence models further justify why socioeconomic capacity should be central in enrollment analytics and retention-oriented outreach. Pricing-policy research emphasizes that tuition and financial considerations systematically shape college choice, making affordability a structural determinant rather than a secondary attribute (Chapman, 1979). In parallel, attrition theory explains non-continuation as a process in which persistence is influenced by the alignment between the student and the institution and by accumulating academic and social integration experiences, meaning early constraints and frictions can compound into withdrawal or non–re-enrollment decisions (Tinto, 1993). This conceptual lens motivates an analytics-to-action framing: predictive modeling can be used to rank risk and segment students by dominant barriers (e.g., affordability, administrative friction, or perceived fit), while targeted outreach becomes an institutional mechanism to reduce constraints and improve persistence outcomes.

Data-Driven Marketing Strategies in Higher Education

In parallel with predictive analytics, higher-education marketing has shifted toward data-driven practices that use digital channels and audience insights to improve recruitment and conversion. Digital marketing and social

name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 1, January 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15703
e-ISSN : 2541-2019
p-ISSN : 2541-044X

media marketing are commonly described as essential for expanding reach, shaping institutional brand perception, and engaging prospective students with tailored content (Budnikevych, 2023; Belostecinic, 2023). E-marketing tactics such as targeted advertising, email campaigns, webinars, and virtual open houses are also reported to reduce informational barriers and enable personalized communication at scale (Rajasekar & Aithal, 2022).

Recent work also emphasizes that effective recruitment depends on matching prospective students to programs through segmentation, personas, and decision-support systems. Precision marketing approaches that combine analytics (e.g., feature selection or sentiment-informed methods) can help identify high-potential prospect segments, while recommender-system concepts can reduce friction in program selection and application pathways (Rasool et al., 2020; Fernández-García et al., 2020; Mengash, 2020). Marketing messages are increasingly aligned with labor-market value signals such as expected salary and Industry 4.0 relevance, and institutions often highlight employer partnerships to differentiate programs and signal career readiness (Fahim et al., 2021; Trần, 2016; Melak & Singh, 2021).

Affordability and perceived value remain decisive for many applicants, making pricing, financial support, and transparent communication integral to enrollment strategy. Addressing affordability and accessibility can improve enrollment intentions, while credibility and ethical marketing practices strengthen trust and institutional reputation (Setiawan et al., 2024; Suleiman, 2021). Overall, the literature supports an integrated view in which predictive modeling informs segmentation and targeting, and explainable outputs guide practical marketing actions by clarifying which factors drive enrollment likelihood for different applicant groups (Rafique et al., 2021; Pelima et al., 2024; Alwarthan et al., 2022).

METHOD

Data preparation, preprocessing, and data splitting

This research frames re-enrollment prediction as a binary classification task that maps each student record to one of two outcomes: Daftar Ulang (re-enrolled) or Tidak Daftar Ulang (did not re-enroll). The operational objective is early identification of students with higher risk of `did not re-enroll` so that outreach and retention resources can be prioritized under limited capacity. For decision-support reporting, `did not re-enroll` class is treated as the target outcome. Consequently, precision, recall, and F1-score are interpreted in terms of how effectively the modeling workflow identifies non–re-enrolling students while controlling unnecessary outreach.

The dataset consists of institutional enrollment records from academic years 2022/2023–2024/2025. Before modeling, the target field is cleaned and converted into a binary format. Feature engineering adds two derived variables—age and gap_year—to capture maturity and time gap between prior graduation and application. A single preprocessing pipeline transforms raw predictors into a model-ready matrix. Numerical variables are scaled, categorical variables are one-hot encoded, and ordinal variables (e.g., requirement completion) are encoded to preserve ordering. Missing values are handled within the pipeline using median imputation for continuous fields and a constant “missing” category for categorical fields. This design ensures that identical transformations are applied during model training, cross-validation, and final test evaluation.

A fixed train–test split is used to preserve an untouched test set for final performance reporting. Model development and selection are performed on the training set using 5-fold cross-validation, producing stability estimates across folds rather than relying on a single split. After cross-validation, each selected configuration is refit on the full training data and evaluated once on the held-out test set. This separation keeps model selection independent from the final test evaluation and supports a more reliable estimate of real-world generalization.

Models development and hyperparameter optimization

Model comparison focuses on a small set of classifiers aligned with the study goals and the latest experimental code. Logistic Regression is used as an interpretable linear baseline. Two nonlinear tree-ensemble methods are evaluated as primary candidates: Random Forest and XGBoost. To ensure a fair comparison, RandomizedSearchCV is used to tune Random Forest and XGBoost under the same 5-fold cross-validation protocol. The search explores key complexity and regularization controls (e.g., tree depth, number of estimators, feature/row subsampling, and regularization terms) to balance bias–variance trade-offs. Class-imbalance handling is incorporated through weighting strategies supported by the algorithms so that the training objective remains sensitive to the minority outcome.

Evaluation metrics, interpretation and threshold-based targeting

Model quality is assessed using a complementary set of metrics. ROC-AUC summarizes ranking performance across thresholds, while precision, recall, and F1-score quantify the quality of identifying the target outcome (did not re-enroll) under a chosen decision threshold. Accuracy is reported for completeness but is not treated as the primary decision criterion under class imbalance. To support actionable interpretation, the selected tree-based model is analyzed using SHAP to identify predictors that most strongly contribute to higher predicted non–re-enrollment risk. Global SHAP summaries are used to characterize overall driver patterns, providing decision-

name of corresponding author

e-ISSN : 2541-2019
p-ISSN : 2541-044X

support insight beyond a single performance score. Because institutional interventions are typically threshold-driven, predicted probabilities are also examined through ROC/threshold analysis. This enables selection of operating points that match campaign constraints (e.g., maximizing coverage of at-risk students versus minimizing unnecessary outreach) and supports segmentation into risk tiers for targeted follow-up actions.

Data Governance, Cohort Validity, and Drift Considerations

The dataset spans three academic years (2022/2023–2024/2025), which introduces the possibility of cohort-level variation arising from changes in admission policies, marketing strategies, pricing, or student composition. To ensure data validity across cohorts, the study applies basic data governance checks, including consistent schema enforcement, harmonized coding of categorical variables, and verification of label consistency for re-enrollment status. Summary statistics and feature distributions are examined across cohorts to identify potential noise or systematic shifts in key academic, demographic, and socioeconomic variables. To support sustainable deployment, a drift monitoring plan is defined. In an operational setting, feature distributions (data drift) and model performance metrics (e.g., ROC-AUC, F1-score, calibration error) can be tracked by intake year. Substantial deviations may trigger model re-training or recalibration, ensuring that predictions remain reliable as enrollment conditions evolve over time.

Model Calibration and Probability Reliability

Because the proposed framework relies on probability thresholds for segmentation and outreach prioritization, probability calibration is explicitly considered. Calibration diagnostics (e.g., calibration curves and Brier score) are evaluated on held-out data to assess whether predicted probabilities correspond to observed non–re-enrollment frequencies. When miscalibration is detected, post-hoc calibration techniques such as Platt scaling or isotonic regression can be applied using cross-validated calibration on the training set and evaluated on the untouched test set. This step improves the reliability of threshold-based targeting and reduces the risk of systematically over- or under-estimating non–re-enrollment risk.

Fairness Considerations and Responsible Use

The study acknowledges that several influential predictors—such as parental income, region, and demographic attributes—may reflect structural inequalities rather than individual intent. Accordingly, the model is designed for supportive decision-making (prioritizing outreach and assistance) rather than exclusionary actions. Fairness is assessed by examining model performance across relevant subgroups (e.g., gender, region, income bands, admission wave), with particular attention to recall and false-negative rates for the target class (did not re-enroll). If meaningful disparities are observed, mitigation strategies such as threshold adjustment, reweighting, or sensitivity analysis can be considered. SHAP-based explanations are interpreted as descriptions of model behavior rather than causal claims and are used to guide equitable and transparent intervention design.

External Validity and Generalization Strategy

Beyond the primary fixed train–test split with 5-fold cross-validation, the study outlines a plan for assessing external validity in multi-cohort settings. A recommended extension is cohort-based temporal validation, where models trained on earlier academic years are evaluated on a later cohort to assess out-of-time generalization. Where feasible, the same workflow can be applied to future intake data or to data from another institution to test transferability. These strategies complement internal validation and strengthen confidence that the proposed ML–Marketing Strategy Framework remains robust under changing enrollment conditions.

RESULT

Data Overview and Class Balance

The analysis used institutional enrollment records from academic years 2022/2023–2024/2025, totaling 2.673 student records with 22 original fields. The target feature was cleaned and encoded as a binary outcome, where `Daftar Ulang` denotes re-enrollment and `Tidak Daftar Ulang` denotes non–re-enrollment. The class distribution was imbalanced, with 74.97% re-enrolled and 25.03% did not re-enrolled. This skew is operationally important because a naive classifier can appear accurate by favoring the majority class, while still failing to identify students at risk of non–re-enrollment. After preprocessing and feature engineering, the modeling matrix expanded to 1099 engineered features. This expansion was primarily driven by one-hot encoding of categorical variables (e.g., school origin, region, admission wave, parental occupation) and inclusion of derived numerical fields (age and gap_year). A fixed split produced 2.004 training instances and 669 test instances, and model development was conducted with 5-fold cross-validation on the training set, followed by a single final evaluation on the untouched test set. Cross-validation was used to obtain a more stable estimate of performance and to reduce sensitivity to any single partition of the training data.

name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 1, January 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15703
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Exploratory patterns from EDA

Several descriptive patterns emerged before modeling and motivated the inclusion of both socioeconomic and administrative variables. Here is a detailed explanation of the EDA visualizations.

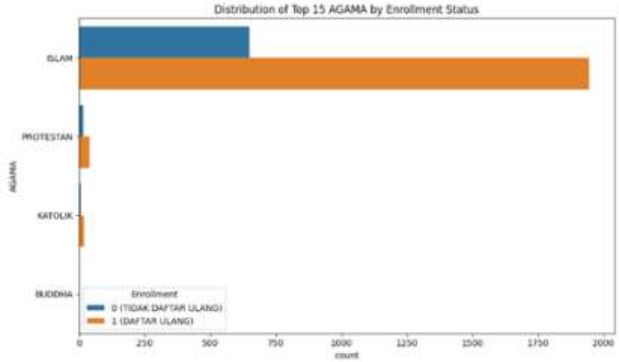

Figure 1. Distribution of Religion Groups

Figure 1 illustrates the distribution of students across different religious groups, including Islam, Protestant, Catholic, and others. Most students who re-enrolled belong to the Islam category, followed by smaller numbers from Protestant, Catholic, and other religions. This suggests that Islam is the dominant religion among students who re-enroll, though it is important to note that religious affiliation may not be the sole determinant of enrollment status.

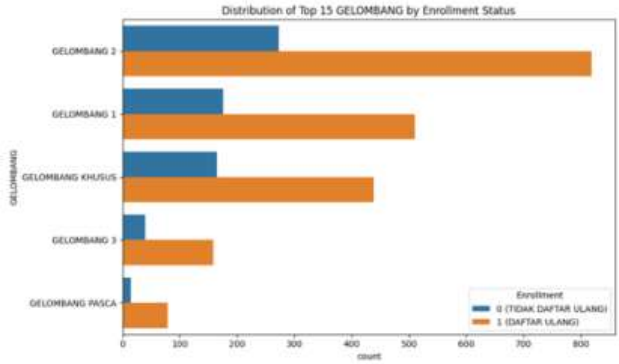

Figure 2. Distribution of Admission Groups

The admission pathway feature is shown to be a significant factor influencing re-enrollment, as shown in Figure 2. Gelombang 2 has the highest number of re-enrolled students, while Gelombang 3 and Gelombang Khusus show relatively fewer re-enrollments. This could imply that students in Gelombang 2 are more likely to stay enrolled compared to other groups, suggesting a potential area for targeted outreach or intervention.

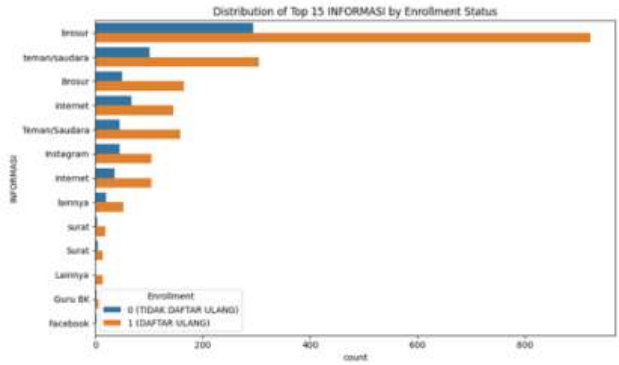

Figure 3. Distribution of Information Sources

Figure 3 reveals that most students who re-enrolled received information about the university through brochures, followed by recommendations from friends or family, and the internet. Students who learned about the institution via brochures tend to have higher re-enrollment rates. This highlights the importance of traditional information dissemination methods in influencing students' decisions to continue their studies, even in the digital age.

name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 1, January 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15703
e-ISSN : 2541-2019
p-ISSN : 2541-044X

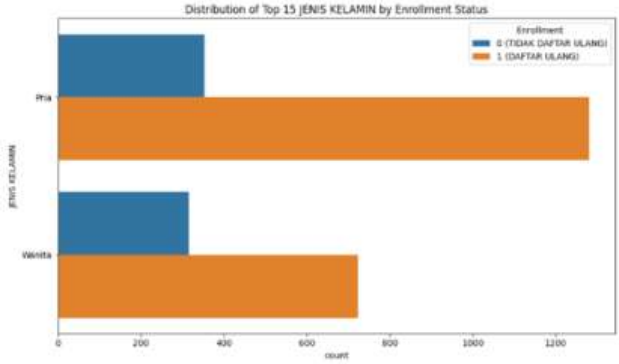

Figure 4. Distribution of Gender Groups

Figure 4 shows that male students make up a larger proportion of both the re-enrolled and did not re-enrolled groups. However, the difference in enrollment status between males and females is not significant, suggesting that gender does not play a major role in the likelihood of re-enrollment in this dataset.

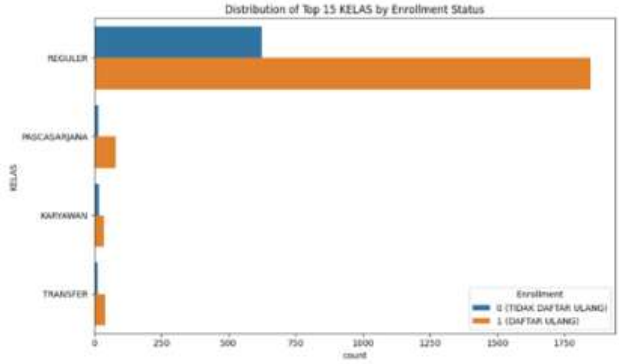

Figure 5. Distribution of Class Types

The class type feature indicates that regular students have the highest re-enrollment rates, as shown in Figure 5. Postgraduate and Employee categories also show notable re-enrollment, although they represent fewer students overall. This could suggest that regular students are more likely to stay enrolled, but those in postgraduate or employee categories also demonstrate significant retention rates, potentially due to different motivations or support systems.

Predictive performance comparison

Model performance was evaluated with metrics computed for the operational target outcome (did not re-enrolled treated as the positive class), using mean ± standard deviation from 5-fold cross-validation and a final report on the held-out test set. Logistic Regression provided a transparent baseline but showed limited discrimination, while tuned tree ensembles improved both classification quality and ranking ability. Accuracy was reported for completeness, but interpretation relied more heavily on F1-score (balance between precision and recall for the target class) and ROC-AUC (ranking quality across thresholds).

Table 1. Model performance (5-fold CV vs. test set)

Model	CV Accuracy (mean±std)	CV F1 (mean±std)	CV ROC-AUC (mean±std)	Test Accuracy	Test Precision	Test Recall	Test F1	Test ROC-AUC
Logistic Regression (baseline)	0.6088±0.0154	0.7174±0.0145	0.5827±0.0222	0.583	0.7587	0.6514	0.701	0.5602
Random Forest (tuned)	0.7181±0.0328	0.7952±0.0289	0.7526±0.0297	0.7294	0.8759	0.745	0.8052	0.7571
XGBoost (tuned)	0.7086±0.0290	0.7866±0.0270	0.7572±0.0278	0.7085	0.8699	0.7191	0.7874	0.7606

The baseline Logistic Regression achieved moderate F1 but weak ROC-AUC in both cross-validation and the test set. This indicates that the linear model produced limited separation in risk scores, which reduces usefulness

name of corresponding author

e-ISSN : 2541-2019
p-ISSN : 2541-044X

for ranking-based outreach. In contrast, both tuned ensembles delivered clear gains, supporting the presence of nonlinear relationships and interactions among socioeconomic, academic, administrative, and program-choice attributes. Random Forest achieved the strongest test-set F1 among the three models (0.8052), reflecting good balance between precision and recall for identifying did not re-enrolled under a specific decision threshold. XGBoost achieved the highest test-set ROC-AUC (0.7606), indicating the strongest overall ranking ability across possible thresholds. Because downstream use requires probability-based prioritization and policy-driven threshold selection, ranking quality was treated as especially important. Under this operational framing, XGBoost was retained as the primary model for interpretation and threshold policy design, while Random Forest remained an important reference for performance trade-offs.

Training and tuning cost

Extended tuning was applied only to Random Forest and XGBoost using randomized search within 5-fold cross-validation. The tuning effort added measurable computation, but the costs remained feasible for datasets of this scale and are compatible with periodic institutional re-training. The reported times summarize the wall-clock effort for hyperparameter search and the final refit of the selected configuration. The tuning results show that stronger models did not require prohibitive computational resources, even with cross-validated randomized search. This supports a practical workflow in which models can be periodically updated as enrollment conditions shift across cohorts, provided that comparable preprocessing and evaluation procedures are maintained.

Table 2. Wall-clock summaries for tuning and refitting

Model	Hyperparameter search	Search time (s)	Final refit time (s)
Logistic Regression (baseline)	No	—	0.87
Random Forest (tuned)	Yes	267.58	1.3
XGBoost (tuned)	Yes	181.42	1.94

SHAP-based interpretation of the selected model

SHAP analysis was applied to the tuned XGBoost model and reported for the did not re-enrolled outcome (shown in the plots as class0). Under this setup, positive SHAP values increase the model's predicted non–re-enrollment risk score, while negative values decrease it. Three complementary views were generated: (i) a global importance ranking (bar), (ii) a distributional summary with directionality (dot), and (iii) a dependence plot to inspect nonlinearity and feature interactions.

The SHAP summary bar plot (Figure 6) shows a clear dominance of `num__Penghasilan Ortu`, indicating that parental income is the strongest global driver of the predicted risk score in the selected model. The next most influential features are largely program-choice and demographic/context indicators, including `cat__Pilihan 1_Sistem Informasi S1`, `cat__Jenis Kelamin_Pria`, `cat__Pilihan 2_Bisnis Digital S1`, and parental-occupation context such as `cat__Pekerjaan Ayah_Buruh`. Cohort markers also appear among the top drivers (e.g., `cat__Tahun Akademik_2023/2024` and `cat__Tahun Akademik_2024/2025`), alongside engineered background variables (`num__Gap_Year` and `num__Umur`). This pattern suggests that the model's risk ranking reflects a joint effect of affordability capacity, program-fit signals, and cohort/administrative context.

name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 1, January 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15703
e-ISSN : 2541-2019
p-ISSN : 2541-044X

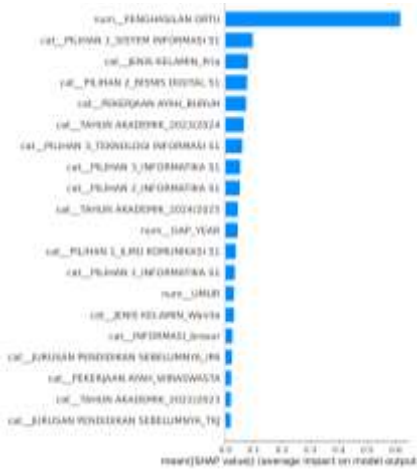

Figure 6. SHAP Summary Bar Plot

The SHAP summary dot plot (Figure 7) adds directionality and heterogeneity. For numeric variables (e.g., income, gap year, age), the color gradient reflects low-to-high values after preprocessing; for one-hot variables (e.g., program-choice indicators), the plot effectively contrasts presence (1) vs. absence (0). The largest spread occurs again for parental income, where contributions vary strongly across records, indicating nonlinear effects and interactions rather than a single monotonic relationship. Several program-choice indicators show small but consistent shifts around zero, meaning that their effect is typically conditional and depends on the broader attribute profile.

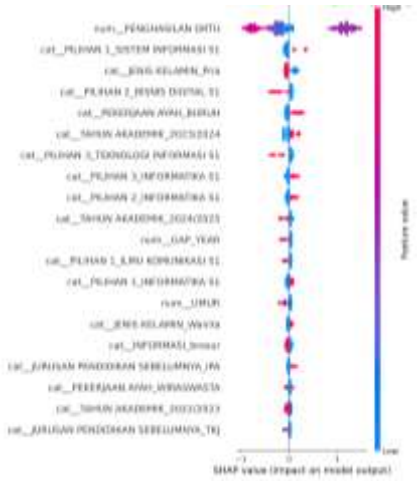

Figure 7. SHAP Summary Dot Plot

The SHAP dependence plot for `num__Penghasilan Ortu` clarifies this nonlinearity and highlights an interaction with `cat__Pilihan 1_Sistem Informasi S1` (used as the color feature). Because the numeric axis represents a standardized value, the horizontal scale should be interpreted relative to the cohort mean rather than in raw currency units. A distinct cluster at the standardized value near 0 contributes strongly in the positive direction, while moderately higher standardized values are associated with negative contributions for many observations. This pattern is compatible with (i) a genuinely non-monotonic relationship, and/or (ii) preprocessing artifacts where many records share an identical imputed or rounded income value. Regardless of cause, the dependence plot confirms that income acts as a primary gatekeeper feature, with its effect modulated by program-choice context.

name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 1, January 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15703
e-ISSN : 2541-2019
p-ISSN : 2541-044X

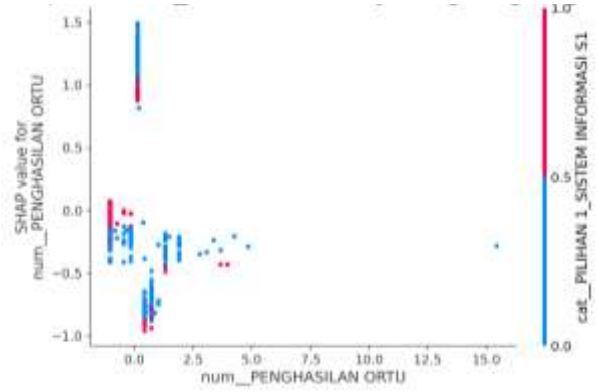

Figure 8. SHAP dependence plot for parental income

Threshold analysis and targeting-oriented outputs

Probability outputs from the tuned XGBoost model were analyzed through ROC/threshold behavior to support operational targeting. Instead of relying on a single default cutoff, threshold selection was treated as a policy decision that balances missed at-risk students (false negatives) against unnecessary outreach (false positives). This framing is aligned with capacity-limited interventions, where outreach volume is constrained by staffing, budget, and campaign timelines, and different campaigns may prefer different operating points. The threshold analysis supports segmentation by predicted risk, enabling tiered outreach strategies. A high-risk group can be prioritized for direct contact and individualized follow-up, a mid-risk group can receive lower-cost reminders or financial-administrative nudges, and a low-risk group can be maintained through general communications. Under this deployment logic, ranking quality (ROC-AUC) is central because it governs how effectively the risk list separates students across tiers, even when the final cutoff shifts across cohorts. As a result, the combination of competitive F1 and strong ROC-AUC supported selecting XGBoost as the primary model for threshold-based targeting and interpretation.

Statistical Significance Testing using McNemar

To assess whether the observed performance differences between the tuned ensembles are statistically meaningful at the operational decision level, McNemar's test was applied to Random Forest (tuned) and XGBoost (tuned) on the same held-out test set using a fixed threshold of 0.5. McNemar's test evaluates whether two classifiers exhibit different error patterns on paired instances by focusing on discordant outcomes. The discordant counts were b = 22 (Random Forest correct, XGBoost wrong) and c = 8 (Random Forest wrong, XGBoost correct), with a total of 30 discordant cases. The continuity-corrected McNemar chi-square statistic was 5.6333 with p = 0.0176, and the exact McNemar test yielded p = 0.0161. These results indicate a statistically significant difference in classification error patterns at alpha = 0.05, implying that the two models do not make mistakes on the same instances at the chosen operating threshold. The larger b than c further indicates that Random Forest correctly classified more cases where XGBoost erred than the reverse, meaning Random Forest produced fewer errors than XGBoost among the discordant test instances.

DISCUSSIONS

Model performance and selection rationale

The comparative evaluation shows that tuned tree ensembles provide clear gains over the linear baseline for predicting did not re-enrolled as the operational target class. Logistic Regression achieved only modest discrimination (test ROC-AUC = 0.5602) despite a moderate F1-score, indicating limited ability to rank students consistently by non–re-enrollment risk. In contrast, both tuned Random Forest and tuned XGBoost improved ranking quality substantially (test ROC-AUC = 0.7571 and 0.7606, respectively) and also delivered stronger classification performance under a fixed threshold (test F1 = 0.8052 and 0.7874). This pattern is consistent with the expectation that non–re-enrollment is shaped by nonlinear relationships and feature interactions across socioeconomic, academic, administrative, and program-choice attributes, which are better captured by ensemble trees than by linear decision boundaries.

Model selection was guided by the intended decision-support use case rather than by a single metric. Random Forest obtained a slightly higher test-set F1-score under the evaluated threshold, while XGBoost produced the highest test-set ROC-AUC, indicating the strongest overall ranking ability across thresholds. Because downstream targeting and segmentation rely on risk ordering and flexible threshold policies (rather than a single default cutoff), ranking quality was treated as especially important. Under this operational framing, tuned XGBoost provides a

name of corresponding author

e-ISSN : 2541-2019
p-ISSN : 2541-044X

strong balance of discrimination and stability while also supporting detailed explanation through SHAP, enabling both performance reporting and marketing-oriented interpretation in a single workflow.

Drivers of non–re-enrollment risk from SHAP and descriptive patterns

The SHAP results clarify which variables most strongly shape the non–re-enrollment risk score in the selected XGBoost model. The global importance ranking is dominated by parental income, indicating that affordability capacity is a primary driver of the risk ordering. Several additional features contribute meaningfully, including program-choice indicators, demographic markers, and family/household context. The dot and dependence plots add two important nuances. First, the income feature shows the widest spread in SHAP values, indicating heterogeneity and nonlinearity: the direction and magnitude of income's contribution vary across student profiles rather than following a single monotonic pattern. Second, the dependence plot highlights an interaction with program-choice context, implying that affordability capacity and program-selection signals jointly shape risk for some subgroups. Because income was standardized during preprocessing, the horizontal axis reflects relative position to the cohort mean rather than raw currency. In addition, the visible clustering around a standardized value near zero is compatible with repeated/typical values introduced by rounding or median imputation; therefore, interpretation benefits from caution and should treat the observed pattern as a model-explanation result that may reflect both behavioral structure and preprocessing artifacts. The descriptive analysis complements the SHAP findings by identifying practical, action-linked correlates observed before modeling. Patterns across admission waves, information sources, administrative completeness, academic readiness, and regional concentration indicate that continuation decisions reflect a mixture of affordability, preparedness, and procedural readiness. These variables are particularly relevant for intervention design because several of them can be influenced operationally (e.g., reminders for document completion, channel-specific outreach, or wave-specific follow-up schedules), even when the model itself remains non-causal.

Marketing and enrollment-operations implications

The findings support deployment as a risk-ranking and segmentation tool rather than a single yes/no classifier. In practice, the tuned XGBoost model produces a probability score that can be used to order students from highest to lowest predicted non–re-enrollment risk. The institution can then convert this ranked list into outreach tiers (for example, high, medium, and low) using a capacity-driven threshold or a top-K rule. This approach is operationally useful because it scales effort to available staff and budget while keeping attention on the students most likely to need intervention.

To make segmentation actionable, the SHAP patterns and descriptive variables can be used to derive persona-style profiles that describe why risk is elevated, not only how high it is. One recurring profile is an affordability-sensitive segment, where parental income and household context dominate the risk score; these students benefit most from financial clarity (total-cost transparency, payment timelines, installment options) and proactive support such as scholarship screening or short financial counseling calls. A second profile is an administratively stalled segment, suggested by signals related to requirement completion and admission-wave timing; for these students, the highest-impact action is reducing process friction through step-by-step checklists, document reminders, and staff-assisted completion when the predicted risk is high. A third profile is a program-fit/value-uncertain segment, where program-choice indicators and cohort markers appear strongly; here, outreach is more effective when it emphasizes program fit, career pathways, and value signals (alumni outcomes, curriculum highlights), delivered through targeted webinars, counselor chats, or program-specific Q&A. A final profile is a channel-dependent/low-touch segment, where information-source and regional patterns matter; in these cases, the main lever is channel matching—reinforcing the same financial and administrative guidance through the channel the student actually responds to (e.g., brochure follow-up calls, referral-based community/alumni reinforcement, or internet-first retargeting and email).

These personas can be combined with risk tiers to map "who to contact" and "what to say" in a consistent playbook. For a high-risk tier, the goal is to avoid missed at-risk students within a limited volume, so outreach should be direct and diagnostic: contact quickly (e.g., phone or WhatsApp), infer the likely barrier using persona cues (financial vs. administrative vs. program-fit), then route the student to the appropriate intervention (aid counseling, document support, or program consultation). For a medium-risk tier, the objective shifts to scaled nudges that reduce friction and reinforce commitment, such as deadline reminders, document-check prompts, and concise program/value messages, with invitations to targeted sessions when relevant; these can be delivered via WhatsApp/SMS/email broadcasts segmented by persona. For a low-risk tier, a lighter maintenance strategy is sufficient, focusing on timeline updates, confirmation of next steps, and broad engagement messages through general channels (website updates, social posts, and campus-wide notifications) so that students do not become "accidentally at risk" due to missing information.

Finally, threshold selection should be treated as a capacity-linked policy rather than a fixed default. When outreach capacity is tight, the institution can raise the threshold (or contact only the top-K highest-risk students)

name of corresponding author

e-ISSN : 2541-2019
p-ISSN : 2541-044X

to prioritize higher precision and reduce unnecessary contacts. When the goal is broader coverage, the threshold can be lowered to increase recall, accepting a larger outreach workload. Because the workflow is designed for probability-based prioritization, this policy framing complements the emphasis on ranking quality (ROC-AUC) and supports repeatable planning across cohorts.

Theoretical implications and academic contribution

This study contributes academically by formalizing affordability-driven heterogeneity in enrollment/retention risk modeling as a segmentation problem, not merely a binary classification task. The dominance of parental-income signals in global SHAP importance, together with the strong discrimination achieved by tuned tree ensembles, conceptually reinforces affordability as a latent constraint that shapes retention decisions through nonlinear interactions with program-choice signals and administrative readiness. This perspective extends prior enrollment prediction literature by framing non–re-enrollment risk as a ranked spectrum of vulnerability rather than a purely linear, additive relationship captured by a simple baseline.

A second contribution is positioning risk ordering as a decision-support mechanism for targeted marketing under capacity constraints. The probability scores are treated as an ordering that supports explicit resource allocation: limited outreach capacity can be prioritized toward the highest-risk students, while SHAP explanations provide a principled basis for differentiating affordability-related barriers from administrative friction or program-fit uncertainty. Conceptually, this bridges predictive enrollment modeling with decision-support and educational marketing scholarship by connecting model outputs to tiered interventions and policy-driven thresholds, emphasizing not only prediction quality but also explainability and operational logic.

Comparison with related work

The observed advantage of tuned ensembles over Logistic Regression aligns with educational data mining findings that nonlinear models often outperform linear baselines when outcomes are shaped by interactions among heterogeneous attributes (Esquivel & Esquivel, 2021; Couronné et al., 2018; Raghavendran et al., 2021; Wu et al., 2023). The strong discrimination achieved by XGBoost is consistent with reports that gradient-boosted trees can deliver robust performance in structured tabular data when tuning and validation are applied appropriately (Farhood, 2024). The use of cross-validation and explicit performance reporting across multiple metrics also addresses common concerns about overfitting and overstated results in applied enrollment prediction under class imbalance (Ghorbani & Ghousi, 2020; Charte, 2020; Roy & Farid, 2024; Alhazmi & Sheneamer, 2023).

Substantively, the prominence of affordability-related signals and household context is consistent with prior work emphasizing socioeconomic determinants in enrollment and retention decisions (Chen et al., 2023; Singh & Alhulail, 2022; Melak & Singh, 2021; Berges et al., 2021; Setiawan et al., 2024). The integration of SHAP further aligns with calls for explainable models that can support actionable interventions rather than opaque predictions (Rafique et al., 2021; Pelima et al., 2024; Alwarthan et al., 2022). By connecting risk scores to segmentation and threshold policies, the approach also matches recommendations from data-driven marketing literature on precision targeting and decision-support systems in higher education (Rasool et al., 2020; Fernández-García et al., 2020; Rajasekar & Aithal, 2022).

Limitations and future work

Several limitations constrain generalization. The dataset reflects one private university in Central Java across three academic years, so transferability to other institutions, regions, or admission policies is uncertain. Cohort shifts (policy changes, local economic conditions, competitor actions, or tuition/aid adjustments) may change feature–outcome relationships and reduce performance over time. Future work should therefore include cross-cohort drift analysis, reporting how discrimination and calibration evolve when models trained on earlier cohorts are tested on later cohorts, and establishing monitoring triggers for re-training.

A related methodological extension is to evaluate robustness under domain shift. Domain shift simulation can be used to stress-test the model by perturbing key covariate distributions (e.g., shifting income composition, changing admission-wave proportions, or altering information-channel mix) and measuring stability of ROC-AUC, PR-AUC, and calibration. This helps distinguish models that perform well on a single split from models that remain reliable under realistic operational changes.

Model explanations should also be interpreted cautiously. SHAP indicates how the trained model uses observed features, not causal effects. Patterns in income-related dependence may partially reflect preprocessing choices (standardization, median imputation, and repeated values). Practical improvements include adding missingness indicators for financial variables, comparing imputation strategies, and applying probability calibration so that score-to-action thresholds remain stable across cohorts and shifts.

Finally, ethical and fairness concerns are central when socioeconomic and demographic attributes influence predictions. Risk-based outreach should be designed as supportive intervention, not exclusion. Future work should explicitly evaluate counterfactual fairness in outreach, assessing whether recommended targeting decisions change

name of corresponding author

unfairly under plausible counterfactual changes to sensitive attributes while holding relevant non-sensitive information constant. Where feasible, causal-inference extensions can move beyond prediction by estimating the effect of specific interventions (e.g., financial counseling, document-support calls) using quasi-experimental designs, enabling evidence-based selection of outreach actions rather than only identifying high-risk students.

CONCLUSION

This study developed a reproducible machine-learning workflow to predict student non–re-enrollment using institutional records from a private university in Central Java. From 2673 records and 22 original variables, preprocessing and feature engineering produced 1099 engineered features, evaluated via a fixed train–test split with 5-fold cross-validation on the training set. Across models, tuned tree ensembles substantially outperformed Logistic Regression: the baseline showed limited discrimination (test ROC-AUC 0.5602), while Random Forest improved classification balance (test F1 0.8052) and XGBoost provided the strongest ranking ability (test ROC-AUC 0.7606, test F1 0.7874). Because outreach decisions depend on reliable risk ordering and flexible operating thresholds, XGBoost was selected as the primary model for decision support. This study contributes to enrollment analytics by demonstrating that ranking-oriented ensemble modeling can be systematically translated into threshold-based segmentation strategies for higher-education marketing — an area underrepresented in retention literature. SHAP interpretation of the selected XGBoost model indicates that parental income is the dominant driver of predicted non–re-enrollment risk, followed by program-choice indicators, demographic/context variables, cohort markers, and engineered background features (e.g., age and gap year). These explanations support an analytics-to-action workflow in which predicted probabilities are used to rank students and segment them into risk tiers, while threshold selection is treated as a policy choice that balances recall (coverage of at-risk students) against outreach capacity. The study is limited to one institution and three cohorts, and SHAP reflects model behavior rather than causality; future work should test cross-cohort stability, monitor drift, calibrate probabilities, and enrich features with support/engagement signals to improve robustness and operational usefulness.

ACKNOWLEDGEMENT

This research was funded by the Amikom Young Lecturer Research Scheme of Universitas Amikom Purwokerto in 2025. The authors wish to express their sincere gratitude to Universitas Amikom Purwokerto for the financial support and resources that made this study possible.

REFERENCES

Alhazmi, E., & Sheneamer, A. (2023). Early Predicting of Students Performance in Higher Education. *Ieee Access*. https://doi.org/10.1109/access.2023.3250702

Alwarthan, S., Aslam, N., & Khan, I. U. (2022). An Explainable Model for Identifying at-Risk Student at Higher Education. *Ieee Access*. https://doi.org/10.1109/access.2022.3211070

Amare, M. Y., & Šimonová, S. (2021). Global Challenges of Students Dropout: A Prediction Model Development Using Machine Learning Algorithms on Higher Education Datasets. *SHS Web of Conferences*, *129*, 09001. https://doi.org/10.1051/shsconf/202112909001

Batool, S., & Liu, Z. (2021). Exploring the Relationships Between Socio-Economic Indicators and Student Enrollment in Higher Education Institutions of Pakistan. *Plos One*, *16*(12), e0261577. https://doi.org/10.1371/journal.pone.0261577

Belostecinic, G. (2023). Online Educational Marketing as a Means to Increase the Attractiveness of the University and Its Image. *Economica*, *3(125)*, 7–27. https://doi.org/10.53486/econ.2023.125.007

Berges, A., Bravo Ramirez, P. H., Pau, I., Tejero, A., & García-Crespo, Á. (2021). A Framework for Strategic Intelligence Systems Applied to Education Management: A Pilot Study in the Community of Madrid. *Ieee Access*. https://doi.org/10.1109/access.2021.3081734

Bieganek, C., Aliferis, C. F., & Ma, S. (2022). Prediction of Clinical Trial Enrollment Rates. *Plos One*, *17*(2), e0263193. https://doi.org/10.1371/journal.pone.0263193

Boumi, S., & Vela, A. (2021). Quantifying the Impact of Student Enrollment Patterns on Academic Success Using a Hidden Markov Model. *Applied Sciences*, *11*(14), 6453. https://doi.org/10.3390/app11146453

Bowen, J. C. (2024). Market Orientation and Performance: A Comparative Study of Private and Public Universities in Kenya. *Journal of Business and Entrepreneurship*, *6*(1), 1–12. https://doi.org/10.51317/ecjbms.v6i1.466

Budnikevych, I. (2023). Communication Component in the Formation of the Image of Higher Education Institutions Based on a Marketing Approach. *Proceedings of Scientific Works of Cherkasy State Technological University Series Economic Sciences*, *24*(4), 5–16. https://doi.org/10.62660/ebcstu/4.2023.05

name of corresponding author

Chapman, R. G. (1979). Pricing Policy and the College Choice Process. Research in Higher Education, 10(1), 37–57.

Charte, F. (2020). A Comprehensive and Didactic Review on Multilabel Learning Software Tools. *Ieee Access*. https://doi.org/10.1109/access.2020.2979787

Cheglakova, L., Devetyarova, I. P., Agalakova, O. S., & Kolesova, Y. A. (2020). Marketing Strategy of Quality Management During Reorganization of Regional Universities in the Process of Modernization of Education in the Conditions of Region's Transition to Industry 4.0. *International Journal for Quality Research*, *14*(1), 33–50. https://doi.org/10.24874/ijqr14.01-03

Chen, J. M. (2017). Three Levels of Push-Pull Dynamics Among Chinese International Students' Decision to Study Abroad in the Canadian Context. *Journal of International Students*, *7*(1), 113–135. https://doi.org/10.32674/jis.v7i1.248

Chen, Z., Cen, G., Wei, Y., & Li, Z. (2023). Student Performance Prediction Approach Based on Educational Data Mining. *Ieee Access*. https://doi.org/10.1109/access.2023.3335985

Couronné, R., Probst, P., & Boulesteix, A. (2018). Random Forest Versus Logistic Regression: A Large-Scale Benchmark Experiment. *BMC Bioinformatics*, *19*(1). https://doi.org/10.1186/s12859-018-2264-5

Czibula, G., Ciubotariu, G., MAIER, M.-I., & Lisei, H. (2022). <i>IntelliDaM</I>: A Machine Learning-Based Framework for Enhancing the Performance of Decision-Making Processes. A Case Study for Educational Data Mining. *Ieee Access*. https://doi.org/10.1109/access.2022.3195531

Durkin, M., McKenna, S., & Cummins, D. (2012). Emotional Connections in Higher Education Marketing. *International Journal of Educational Management*, *26*(2), 153–161. https://doi.org/10.1108/09513541211201960

Erpurini, W. (2024). The IMPACT OF TARGET MARKETS, UNIQUE RESOURCES &Amp; MARKET ATTRACTION ON DIGITAL MARKETING: THEIR SIGNIFICANCE FOR SELECTING POSTSECONDARY EDUCATION. *Jurnal Riset Bisnis Dan Manajemen*, *17*(2), 29–38. https://doi.org/10.23969/jrbm.v17i2.12093

Esquivel, J. A., & Esquivel, J. A. (2021). A Machine Learning Based DSS in Predicting Undergraduate Freshmen Enrolment in a Philippine University. *International Journal of Computer Trends and Technology*, *69*(5), 50–54. https://doi.org/10.14445/22312803/ijctt-v69i5p107

Fahim, A., Addae, B. A., Ofosu-Adarkwa, J., Tan, Q., & Bhatti, U. A. (2021). Industry 4.0 and Higher Education: An Evaluation of Barriers Affecting Master's in Business Administration Enrolments Using a Grey Incidence Analysis. *Ieee Access*. https://doi.org/10.1109/access.2021.3082144

Farhood, H. (2024). Evaluating and Enhancing Artificial Intelligence Models for Predicting Student Learning Outcomes. *Informatics*, *11*(3), 46. https://doi.org/10.3390/informatics11030046

Fernández-García, A. J., Rodríguez-Echeverría, R., Preciado, J. C., Conejero, J. M., & Sánchez-Figueroa, F. (2020). Creating a Recommender System to Support Higher Education Students in the Subject Enrollment Decision. *Ieee Access*. https://doi.org/10.1109/access.2020.3031572

Fu, L. D., & Aliferis, C. F. (2010). Using Content-Based and Bibliometric Features for Machine Learning Models to Predict Citation Counts in the Biomedical Literature. *Scientometrics*, *85*(1), 257–270. https://doi.org/10.1007/s11192-010-0160-5

Ghorbani, R., & Ghousi, R. (2020). Comparing Different Resampling Methods in Predicting Students' Performance Using Machine Learning Techniques. *Ieee Access*. https://doi.org/10.1109/access.2020.2986809

Gnoh, H. Q. (2024). Enhancing Business Sustainability Through Technology-Enabled AI: Forecasting Student Data and Comparing Prediction Models for Higher Education Institutions (HEIs). *PaperASIA*, *40*(2b), 48–58. https://doi.org/10.59953/paperasia.v40i2b.86

Hossler, D., & Gallagher, K. S. (1987). Studying Student College Choice: A Three-Phase Model and the Implications for Policymakers. College and University, 62(3), 207–221.

Jahir, A., Wahid, A. M., & Sufranto, T. T. (2024). Optimizing Higher Education Performance Through Data Integration Using the Zachman Framework: A Case Study on LAM Infokom Accreditation Criteria. *Jurnal Nasional Teknologi Dan Sistem Informasi*, *10*(3), Article 3. https://doi.org/10.25077/TEKNOSI.v10i3.2024.201-215

Kevin Mario Laura-De La Cruz. (2023). A Study on Higher Education Student Satisfaction and Educational Marketing in Peru. *Human Review International Humanities Review / Revista Internacional De Humanidades*, *21*(1), 1–10. https://doi.org/10.37467/revhuman.v21.5025

name of corresponding author

Khusnuliawati, H., & Putri, D. R. (2021). Hybrid Clustering Based on Multi-Criteria Segmentation for Higher Education Marketing. *Telkomnika (Telecommunication Computing Electronics and Control)*, *19*(5), 1498. https://doi.org/10.12928/telkomnika.v19i5.18965

Kotler, P., & Fox, K. F. A. (1995). Strategic Marketing for Educational Institutions (2nd ed.). Prentice Hall.

Lan, Y. C., Tang, G., & Heitjan, D. F. (2018). Statistical Modeling and Prediction of Clinical Trial Recruitment. *Statistics in Medicine*, *38*(6), 945–955. https://doi.org/10.1002/sim.8036

Li, X. (2022). Sequence Model and Prediction for Sustainable Enrollments in Chinese Universities. *Sustainability*, *15*(1), 214. https://doi.org/10.3390/su15010214

Maphosa, M., Doorsamy, W., & Paul, B. (2023). Student Performance Patterns in Engineering at the University of Johannesburg: An Exploratory Data Analysis. *Ieee Access*. https://doi.org/10.1109/access.2023.3277225

Mbanga, S. (2023). Enhancing a Quality Teaching and Learning Environment in Large Classes in South African Universities: A Theoretical Exposition. *Annals of Social Sciences & Management Studies*, *9*(5). https://doi.org/10.19080/asm.2023.09.555773

Melak, A., & Singh, S. (2021). Factors Affecting Women's Choice of Learning Engineering and Technology Education in Ethiopia. *Ieee Access*. https://doi.org/10.1109/access.2021.3087548

Mengash, H. A. (2020). Using Data Mining Techniques to Predict Student Performance to Support Decision Making in University Admission Systems. *Ieee Access*. https://doi.org/10.1109/access.2020.2981905

Moogan, Y. J. (2011). Can a Higher Education Institution's Marketing Strategy Improve the Student-institution Match? *International Journal of Educational Management*, *25*(6), 570–589. https://doi.org/10.1108/09513541111159068

Mushketova, N. S., Bydanova, E., & Rouet, G. (2018). National Strategy for Promotion of Russian Universities in the World Market of Education Services. *International Journal of Educational Management*, *32*(1), 46–56. https://doi.org/10.1108/ijem-10-2016-0207

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

Plak, S., Cornelisz, I., Meeter, M., & Klaveren, C. v. (2021). Early Warning Systems for More Effective Student Counselling in Higher Education: Evidence From a Dutch Field Experiment. *Higher Education Quarterly*, *76*(1), 131–152. https://doi.org/10.1111/hequ.12298

Qin, L., Shanks, K., Phillips, G. A., & Bernard, D. (2019). The Impact of Lengths of Time Series on the Accuracy of the ARIMA Forecasting. *International Research in Higher Education*, *4*(3), 58. https://doi.org/10.5430/irhe.v4n3p58

Rafique, A., Khan, M. S., Jamal, M. H., Tasadduq, M., Rustam, F., Lee, E., Washington, P. B., & Ashraf, I. (2021). Integrating Learning Analytics and Collaborative Learning for Improving Student's Academic Performance. *Ieee Access*. https://doi.org/10.1109/access.2021.3135309

Raghavendran, Ch. V., Vamsi, Ch. P. V., Veerraju, T., & Veluri, R. K. (2021). *Predicting Student Admissions Rate Into University Using Machine Learning Models*. 151–162. https://doi.org/10.1007/978-981-15-9516-5_13

Rajasekar, D., & Aithal, P. S. (2022). Study on Cadet Satisfaction Towards E-Marketing Promotion in Private Maritime University (Study With Reference to Rural Cadets, Chennai). *International Journal of Case Studies in Business It and Education*, 489–495. https://doi.org/10.47992/ijcsbe.2581.6942.0212

Rasool, A., Tao, R., Kamyab, M., & Hayat, S. (2020). GAWA–A Feature Selection Method for Hybrid Sentiment Classification. *Ieee Access*. https://doi.org/10.1109/access.2020.3030642

Roy, K., & Farid, D. Md. (2024). An Adaptive Feature Selection Algorithm for Student Performance Prediction. *Ieee Access*. https://doi.org/10.1109/access.2024.3406252

Sabino Parmezan, A. R., A. Souza, V. M., & A. Batista, G. E. (2022). Time Series Prediction via Similarity Search: Exploring Invariances, Distance Measures and Ensemble Functions. *Ieee Access*. https://doi.org/10.1109/access.2022.3192849

Setiawan, R., Putranto, A., Wihendro, Princes, E., Geraldina, I., Julianti, E., Safitri, J., & Pannen, P. (2024). Tech-Driven Transformation: Innovative Pricing Strategies for E-Learning. *Ieee Access*. https://doi.org/10.1109/access.2024.3392489

Singh, H. P., & Alhulail, H. (2022). Predicting Student-Teachers Dropout Risk and Early Identification: A Four-Step Logistic Regression Approach. *Ieee Access*. https://doi.org/10.1109/access.2022.3141992

Suleiman, J. (2021). Credibility, Ethics and Sustainable Marketing in Higher Education Institutions. *International Journal of Innovative Research and Development*, *10*(12). https://doi.org/10.24940/ijird/2021/v10/i12/dec21016

name of corresponding author

Thomas, S. (2011). What Drives Student Loyalty in Universities: An Empirical Model From India. *International Business Research*, *4*(2). https://doi.org/10.5539/ibr.v4n2p183

Tinto, V. (1993). Leaving College: Rethinking the Causes and Cures of Student Attrition (2nd ed.). University of Chicago Press.

Trần, T. T. (2016). Enhancing Graduate Employability and the Need for University-Enterprise Collaboration. *Journal of Teaching and Learning for Graduate Employability*, *7*(1), 58–71. https://doi.org/10.21153/jtlge2016vol7no1art598

Vaitsis, C., Hervatis, V., & Zary, N. (2016). *Introduction to Big Data in Education and Its Contribution to the Quality Improvement Processes*. https://doi.org/10.5772/63896

Wahid, A. M., Afuan, L., & Utomo, F. S. (2024). ENHANCING COLLABORATION DATA MANAGEMENT THROUGH DATA WAREHOUSE DESIGN: MEETING BAN-PT ACCREDITATION AND KERMA REPORTING REQUIREMENTS IN HIGHER EDUCATION. *Jurnal Teknik Informatika (Jutif)*, *5*(6), Article 6. https://doi.org/10.52436/1.jutif.2024.5.6.1747

Wahid, A. M., Hariguna, T., & Karyono, G. (2025). Optimization of Recommender Systems for Image-Based Website Themes Using Transfer Learning. *Journal of Applied Data Sciences*, *6*(2), Article 2. https://doi.org/10.47738/jads.v6i2.671

Wan, S. J. (2024). A Denoising Time Window Algorithm for Optimizing LSTM Prediction. *Ieee Access*. https://doi.org/10.1109/access.2024.3404456

Wanjau, S. K., Okeyo, G., & Rimiru, R. (2016). Data Mining Model for Predicting Student Enrolment in STEM Courses in Higher Education Institutions. *International Journal of Computer Applications Technology and Research*, *5*(11), 698–704. https://doi.org/10.7753/ijcatr0511.1004

Watkins, A., & Kaplan, A. (2018). Modeling in R and Weka for Course Enrollment Prediction. *International Journal of Institutional Research and Management*, *2*(1), 1–17. https://doi.org/10.52731/ijirm.v2.i1.212

Wu, J., Lin, M.-S., & Tsai, C. (2023). A Predictive Model That Aligns Admission Offers With Student Enrollment Probability. *Education Sciences*, *13*(5), 440. https://doi.org/10.3390/educsci13050440

Yang, Y., & Yang, Y. (2020). Hybrid Method for Short-Term Time Series Forecasting Based on EEMD. *Ieee Access*. https://doi.org/10.1109/access.2020.2983588

Yi, J. C., Kang-Yi, C. D., Burton, F., & Chen, H. D. (2018). Predictive Analytics Approach to Improve and Sustain College Students' Non-Cognitive Skills and Their Educational Outcome. *Sustainability*, *10*(11), 4012. https://doi.org/10.3390/su10114012

Yi, X., Si, W., Zhu, J., Sun, Z., Zhao, J., Xu, M., & Xu, S. (2021). Multi-Model Fusion Short-Term Load Forecasting Based on Random Forest Feature Selection and Hybrid Neural Network. *Ieee Access*. https://doi.org/10.1109/access.2021.3051337

Zhang, K., Li, Z., Wang, H., & Wang, H. (2015). Fuzzy Time Series Prediction Model and Application Based on Fuzzy Inverse. *International Journal of Signal Processing Image Processing and Pattern Recognition*, *8*(10), 121–128. https://doi.org/10.14257/ijsip.2015.8.10.14

name of corresponding author