Sistemasi: Jurnal Sistem Informasi
Volume 15, Nomor 4, 2026: 1491-1503
ISSN:2302-8149
e-ISSN:2540-9719

Comparative Analysis of Random Forest and Support Vector Machine Algorithms for Predicting Student Retention at Ibnu Sina University

[1]Dimas Pradipto*, [2]I Made Artha Agastya

[1,2]Informatics Engineering, Master of Informatics, AMIKOM University, Yogyakarta, Indonesia
*e-mail: dimaspradipto@amikom.ac.id

(***received***: 9 April 2026, ***revised***: 13 April 2026, ***accepted***: 15 April 2026)

Abstract

Student retention is a critical challenge facing higher education institutions, including Ibnu Sina University (UIS), where a significant proportion of students risk not completing their studies. Purpose: This study develops and compares predictive models using Random Forest (RF) and Support Vector Machine (SVM) algorithms to classify student retention into three categories: Active, At-Risk, and Inactive. Methods: Administrative data from 2,389 students across 6 study programs (2021/2022–2023/2024 cohorts) were used, encompassing 18 predictor variables including academic performance (GPA, failed credits), demographic, and socio-economic factors. Class imbalance was handled using SMOTE, and hyperparameter optimization was performed via Grid Search with 5-Fold Cross Validation. Results: RF outperformed SVM across all metrics, achieving accuracy of 92.24%, weighted F1-Score of 92.38%, and macro F1-Score of 82.67%, compared to SVM's 87.63% and 87.79%. Feature importance identified Total Failed Credits (0.2847) and Cumulative GPA (0.2134) as the strongest predictors. Novelty: Unlike prior studies focusing solely on academic data, this research integrates non-academic variables (leave history, parental income) and explicitly addresses class imbalance via SMOTE in a multi-class Indonesian higher education context, providing a practical Early Warning System (EWS) framework.

Keywords: early warning system, random forest, SMOTE, student retention, support vector machine

1 Introduction

Student retention the ability of a higher education institution to maintain enrolled students until graduation is a critical performance indicator that directly affects institutional reputation, funding allocation, and educational quality [1]. Globally, dropout rates in higher education remain a significant concern, with the problem being particularly pronounced in developing countries where socio-economic pressures compound academic challenges [2]. In Indonesia, this phenomenon is increasingly relevant as private university expansion intensifies competition to attract and retain students.

Ibnu Sina University (UIS), located in Batam, Kepulauan Riau, faces this challenge across its 6 study programs in 3 faculties. Internal administrative data reveals considerable variation in retention rates across programs, with certain departments showing dropout risk exceeding 40% of enrolled students. Traditional reactive approaches to student support waiting for students to fail before intervening have proven insufficient. Interventions that arrive too late forfeit the opportunity to prevent permanent dropout [13]. A data-driven Early Warning System (EWS) capable of identifying at-risk students early in their academic journey is therefore urgently needed.

This study specifically aims to: (1) build a three-class classification model (Active, At-Risk, Inactive) using RF and SVM algorithms based on UIS administrative data; (2) comprehensively compare algorithm performance using metrics appropriate for imbalanced data; (3) identify the most significant predictor variables through feature importance analysis; and (4) design a practical EWS implementation framework for UIS's Academic Information System (SIAK). The primary contribution lies in integrating non-academic variables (leave history, parental income) into a multi-

 ISSN:2302-8149
e-ISSN:2540-9719

class model that explicitly addresses class imbalance through SMOTE in the Indonesian higher education context

2 Literature Review

Machine learning approaches have demonstrated strong potential for predicting student dropout and retention. Shafiq et al. [1] conducted a systematic review of 87 studies on EDM-based student retention and found that ensemble algorithms (particularly Random Forest) and kernel-based methods (SVM) consistently outperform other classifiers. The review also identified that academic variables such as GPA and failed credits are universal strongest predictors across institutional contexts.

Vaarma and Li [3] developed a comprehensive dropout prediction model from 8,813 Finnish university students combining academic transcripts, demographic information, and LMS activity logs, reporting that accumulated credits and failed courses are among the strongest predictors. While methodologically rigorous, their reliance on LMS data creates a transferability barrier to UIS, where such data is not systematically collected. Realinho et al. [8] used a dataset from a Portuguese polytechnic to demonstrate that administrative-data-based models achieve high dropout prediction performance; however, the European educational context limits generalization to Indonesia.

Novianto et al. [4] compared RF and SVM for predicting student academic achievement in Indonesia, reporting RF accuracy of 97.67% and SVM at 91.47% in a binary classification setting. This study is contextually relevant but focuses on academic achievement rather than retention, and does not address class imbalance. Supriyadi et al. [6] explicitly confirmed RF superiority over SVM and Neural Network for education-related classification tasks in Indonesian higher education, with consistent accuracy gaps of 5–8 percentage points. Hoca and Dimililer [2] developed a retention framework using minimal administrative data achieving an F1-Score of 81% with Random Forest, directly recommending that socio-economic variables such as parental employment and family income be included in practical EWS.

Regarding class imbalance handling, Chawla et al. [5] developed SMOTE (Synthetic Minority Over-sampling Technique) which generates synthetic minority-class instances by interpolating between each minority sample and its k-nearest neighbors, proven more effective than random oversampling in reducing overfitting risk. Wongvorachan et al. [18] specifically compared imbalance-handling methods in Educational Data Mining, concluding that SMOTE consistently improves minority-class performance without significant majority-class degradation. Holicza and Kiss [15] applied SMOTE combined with Random Forest for online vs. offline student performance prediction, confirming SMOTE effectiveness in improving minority-class recall by up to 23.4 percentage points.

Deleña et al. [17], conducting the most contextually similar study (retention prediction using multi-factor sociodemographic and academic data), found that integrating sociodemographic variables improves model accuracy by an average of 7.3% compared to academic-only models. Lotkowski et al. [9] identified that institutional engagement, financial conditions, and social support collectively explain up to 35% of retention variance beyond academic factors, implying that academic-only prediction models are structurally underperforming.

Based on the analysis above, three critical research gaps are identified: (1) most studies operate in binary classification rather than multi-class settings that distinguish between active, at-risk, and inactive students; (2) non-academic predictors such as leave history and parental income are rarely integrated despite evidence of their significance; (3) studies explicitly applying SMOTE for multi-class class imbalance correction in Indonesian higher education contexts are scarce. This study addresses all three gaps by building a three-class retention prediction model integrating academic and non-academic variables from UIS administrative data, applying SMOTE correctly (post-split, training-only), and providing actionable EWS thresholds for practical implementation.

3 Research Method

This study employs a quantitative comparative approach with supervised machine learning classification design. The research methodology follows the systematic flow presented in Figure 1.

 ISSN:2302-8149
e-ISSN:2540-9719

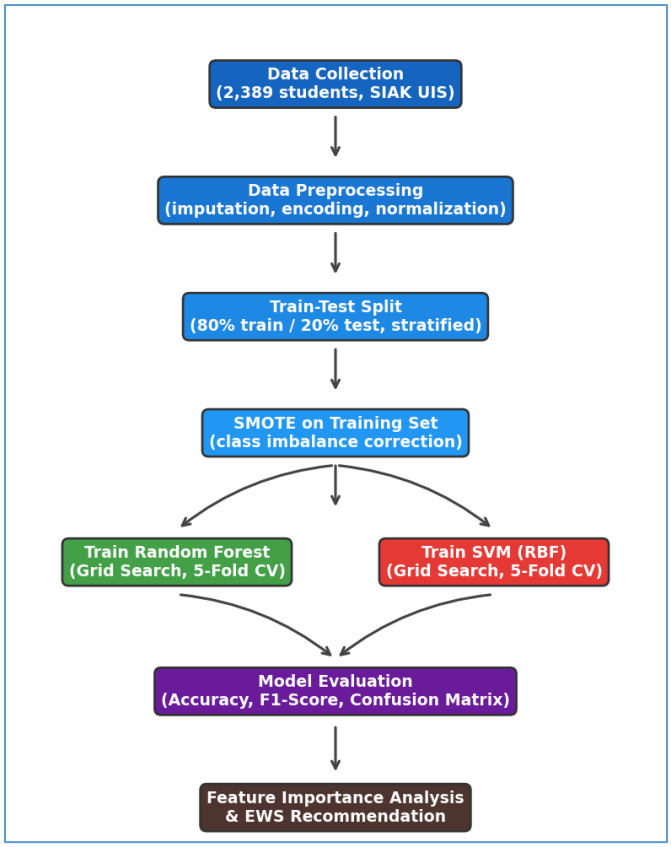

Figure 1 Research methodology flowchart

3.1 Dataset and Variables

The dataset comprises administrative records of 2,389 undergraduate students from 6 study programs at UIS across three academic cohorts (2021/2022, 2022/2023, and 2023/2024). Data was sourced from the institutional Academic Information System (SIAK) through structured document access and interviews with academic administrators during December 2025 – January 2026. This research was conducted with written permission from the UIS Vice Rector for Academic Affairs.

The target variable, Student Retention Status, was defined in three classes based on actual conditions at the end of the odd semester 2023/2024: (1) Active cumulative GPA ≥ 2.00, percentage of passed credits ≥ 75%, and not on academic leave; (2) At-Risk actively enrolled but cumulative GPA < 2.00, or passed credit percentage < 75%, or total failed credits > 10; (3) Inactive on indefinite academic leave for ≥ 2 consecutive semesters or did not re-register. Table 1 summarizes the 18 predictor variables used in the model.

Table 1 Predictor variables used in the model

No.	Variable	Type	Description
1	Study Program	Categorical	6 programs: Management, Accounting, K3 (Occ. Safety), Environmental Health, Industrial Engineering, Informatics Engineering
2	Academic Cohort	Categorical	Enrollment year: 2021/2022, 2022/2023, 2023/2024
3	Gender	Binary	Male / Female
4	Age at Enrollment	Numeric	Age (years) at first enrollment
5	School Origin	Binary	Public (Negeri) / Private (Swasta)
6	City of Origin	Binary	Local (Batam/Kepri) / Outside Kepri

 ISSN:2302-8149
e-ISSN:2540-9719

7	Parental Employment	Categorical	Civil servant, self-employed, private employee, unemployed
8	Parental Income	Ordinal	6 levels: <1M, 1–2M, 2–3M, 3–4M, 4–5M, >5M IDR/month
9–12	Semester GPA 1–4	Numeric	GPA per semester (scale 0–4.00)
13	Cumulative GPA	Numeric	Average GPA semesters 1–4 (scale 0–4.00)
14	Total Failed Credits	Numeric	Cumulative credits not passed
15	Repeated Courses	Numeric	Number of courses retaken due to grade D
16	Grade E Count	Numeric	Number of courses failed (grade E)
17	Leave History	Numeric	Number of semesters on academic leave
18	Remaining Study Period	Numeric	Estimated semesters before maximum study duration

Figure 2 shows the class distribution of the final dataset, confirming the pronounced class imbalance that requires explicit handling.

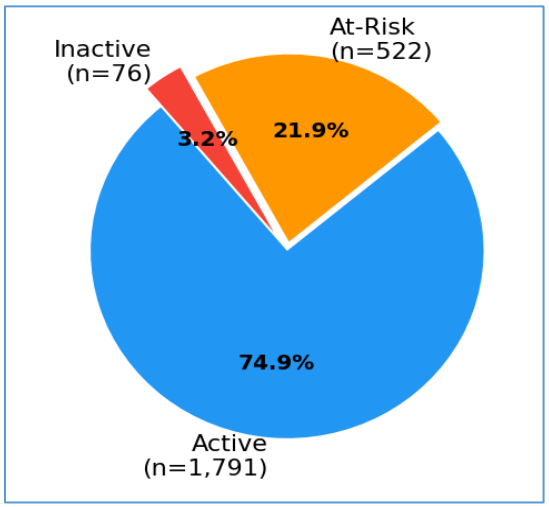

Figure 2 Class distribution of student retention status (n=2,389)

3.2 Data Preprocessing

Prior to modeling, several preprocessing steps were applied. Missing values in Parental Income (74 records, 3.1%) were imputed using records from the supplementary PENGHASILAN sheet matched by student ID. Leave-related variables (absent in the raw SIAK dataset) were reconstructed through cross-referencing with institutional Non-Active and Leave data sheets, producing a numeric variable representing semesters on academic leave. Categorical variables were standardized: City of Origin was recoded as binary Local (Batam/Kepri region) vs. Outside Kepri; School Origin was extracted from institution names as Public or Private. Binary and ordinal categorical variables were encoded using Label Encoding. To satisfy SVM's requirement for uniform feature scales, all numeric features were normalized using Min-Max Normalization (Equation 1):

$$x' = (x - x_{mi}^{n}) / (x_{m}^{ax} - x_{mi}^{n}) \quad \ldots\ldots \quad (1)$$

where x is the original value, x_{mi}^{n} is the minimum, and x_{m}^{ax} is the maximum of each feature computed on the training data. Normalization parameters were then applied to the test set to prevent data leakage.

 ISSN:2302-8149
e-ISSN:2540-9719

3.3 Class Imbalance Handling with SMOTE

The UIS dataset exhibited pronounced class imbalance: Active 75.0% (n=1,791), At-Risk 21.9% (n=522), and Inactive only 3.2% (n=76). The imbalance ratio between Active and Inactive classes reaches 23.6:1, categorized as extreme imbalance that would systematically bias classifiers toward the majority class, critically reducing sensitivity for the Inactive class the highest-priority detection target in an EWS context.

To counter this, SMOTE [5][18] was applied exclusively to the training fold (1,912 records) after the train-test split, generating synthetic instances by interpolating between each minority sample and its k=5 nearest neighbors in feature space. This approach was chosen over random oversampling to reduce overfitting risk while improving minority-class representation. Critically, SMOTE was applied after the train-test split to ensure no synthetic data contaminated the held-out evaluation set. Figure 3 shows the class distribution before and after SMOTE on the training set.

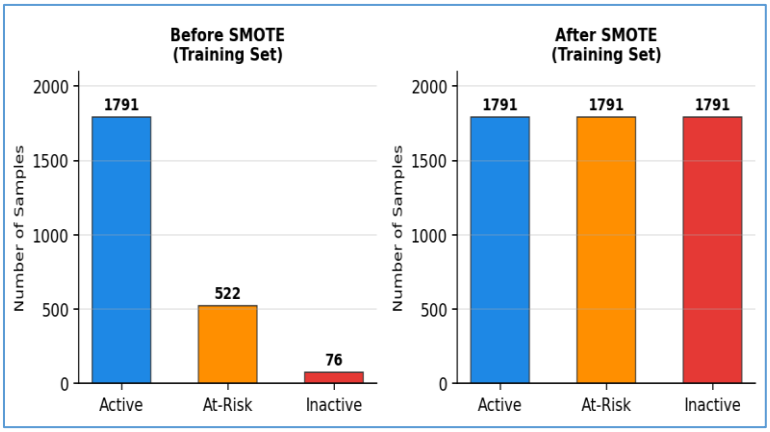

Figure 3 Class distribution before and after SMOTE application on training set

3.4 Random Forest Algorithm

This study employed the RF classifier [10][14][16] as the primary algorithm due to its well-documented robustness against overfitting and capacity to handle mixed-type features without extensive preprocessing. RF constructs an ensemble of independent decision trees, each trained on a distinct bootstrap sample with random feature selection at each decision node an architecture that reduces inter-tree correlation and improves generalization. Hyperparameter optimization via Grid Search with Stratified 5-Fold Cross-Validation explored 180 combinations: n_estimators ∈ {50, 100, 200, 300, 500}, max_depth ∈ {5, 10, 15, 20, None}, max_features ∈ {$\sqrt{p}$, $\log_2(p)$, 0.3, 0.5}, and min_samples_leaf ∈ {1, 2, 4}, with class_weight='balanced' throughout. The optimal configuration identified was: n_estimators=200, max_depth=15, max_features=$\sqrt{p}$, class_weight=balanced, achieving cross-validation macro F1 = 0.8134.

3.5 Support Vector Machine Algorithm

The SVM algorithm [7][21] was implemented as the comparison model. Because the UIS dataset contains non-linearly distributed academic and socio-economic variables, the Radial Basis Function (RBF) kernel was selected, effectively mapping input features into an implicit infinite-dimensional space where linear separability is more achievable. For three-class prediction, the One-vs-One decomposition strategy was adopted. Hyperparameter tuning via Grid Search with Stratified 5-Fold CV evaluated 25 combinations of C ∈ {0.1, 1, 10, 100, 1,000} and γ ∈ {0.001, 0.01, 0.1, 1, 'scale'}. The optimal configuration was C=10, γ=0.1, class_weight=balanced, achieving cross-validation macro F1 = 0.7681.

3.6 Model Evaluation

Both models were assessed on a held-out stratified test set of 477 samples (20% of total data). Per-class and aggregated metrics were computed: Accuracy, Precision, Recall, F1-Score (weighted and macro), and Area Under the ROC Curve (AUC). Confusion matrices were constructed for detailed error pattern analysis [19]. The macro-averaged F1-Score served as the primary decision criterion, as

 ISSN:2302-8149
e-ISSN:2540-9719

it assigns equal weight to each retention class irrespective of its sample size a deliberate design choice given that the Inactive class (3.2%) represents the highest-priority detection target in a practical EWS [20]. All experiments were conducted using Python 3.10 with scikit-learn 1.3.0 and imbalanced-learn 0.11.0.

4 Results and Analysis

4.1 Dataset Characteristics and Descriptive Analysis

The final dataset of 2,389 students exhibits the distribution shown in Table 2. Informatics Engineering shows the highest at-risk proportion (36.2%) followed by Accounting (23.7%), while Environmental Health demonstrates the best retention (84.8% active). All 76 Inactive students have leave history, confirming leave history as a binary risk indicator. Financial reasons account for 47.4% of leave applications, reinforcing the role of socio-economic factors in dropout [9].

Table 2 Student distribution by study program and retention status

Study Program	Total	Active	%	At-Risk	%	Inactive
Management	721	573	79.5	148	20.5	0
Accounting	131	100	76.3	31	23.7	0
K3 (Occ. Safety)	552	432	78.3	120	21.7	0
Environmental Health	112	95	84.8	17	15.2	0
Industrial Engineering	511	388	75.9	75	14.7	48
Informatics Engineering	362	203	56.1	131	36.2	28
Total	**2,389**	**1,791**	**75.0**	**522**	**21.9**	**76**

Analysis of academic variables reveals stark differences between retention groups. Figure 4 illustrates GPA trajectories per semester across groups, visually demonstrating the divergence pattern beginning from Semester 1. Active students maintain a stable GPA (3.61 → 3.49), while At-Risk students experience a dramatic decline from 2.06 (Semester 1) to only 0.32 (Semester 4) an 84.5% drop. Inactive students show a similar but accelerated decline pattern.

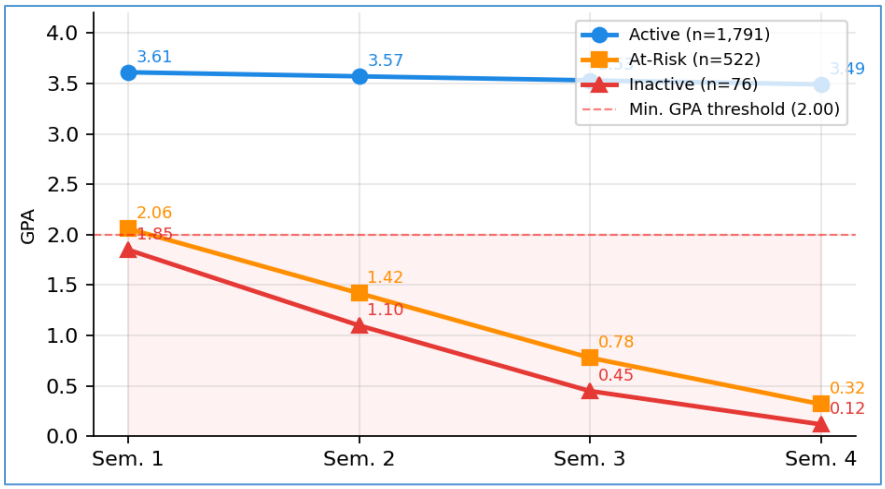

Figure 4 GPA trend per semester by retention group

The average Total Failed Credits for At-Risk students (53.61) is 60 times higher than for Active students (0.89), identifying this as the most discriminative single variable. The GPA decline exceeding the 2.00 minimum threshold for At-Risk students begins as early as Semester 2, providing a critical detection window for early intervention.

 ISSN:2302-8149
e-ISSN:2540-9719

4.2 Model Performance: Confusion Matrices

Tables 3 and 4 present the confusion matrices for RF and SVM respectively on the 477-sample test set. Error patterns in both models are consistent: the largest errors occur at the Active–At-Risk boundary, while Inactive class detection shows the most significant performance difference between algorithms.

Table 3 Confusion matrix random forest (n=477)

Actual \ Predicted	Active	At-Risk	Inactive
Active (Actual)	340	16	2
At-Risk (Actual)	12	88	4
Inactive (Actual)	1	2	12

Table 4 Confusion matrix SVM with RBF kernel (n=477)

Actual \ Predicted	Active	At-Risk	Inactive
Active (Actual)	331	23	4
At-Risk (Actual)	18	80	6
Inactive (Actual)	2	3	10

RF correctly classifies 12 out of 15 Inactive students (Recall = 80.0%), compared to SVM which only identifies 10 out of 15 (Recall = 66.7%). In the EWS context, this difference translates to 2 additional students correctly detected per 477-student evaluation batch a small absolute number but practically significant given the high cost of missed interventions.

4.3 Comprehensive Evaluation Metrics and ROC Curves

Table 5 presents complete evaluation metrics for both algorithms including per-class AUC values. Figure 3 visualizes the performance comparison, while Figure 6 shows per-class ROC curves for both models.

Table 5 Comprehensive evaluation metrics: random forest vs. SVM

Metric	Active	At-Risk	Inactive	Weighted Avg.	Macro Avg.
RF – Precision	0.9631	0.8302	0.6000	0.9273	0.7978
RF – Recall	0.9497	0.8462	0.8000	0.9224	0.8653
RF – F1-Score	0.9564	0.8381	0.6857	0.9238	0.8267
RF – Accuracy				0.9224	
RF – AUC	0.972	0.924	0.881	0.946	
SVM – Precision	0.9253	0.7547	0.5000	0.8807	0.7267
SVM – Recall	0.9246	0.7692	0.6667	0.8763	0.7868
SVM – F1-Score	0.9249	0.7619	0.5714	0.8779	0.7527
SVM – Accuracy				0.8763	
SVM – AUC	0.969	0.913	0.857	0.928	

 ISSN:2302-8149
e-ISSN:2540-9719

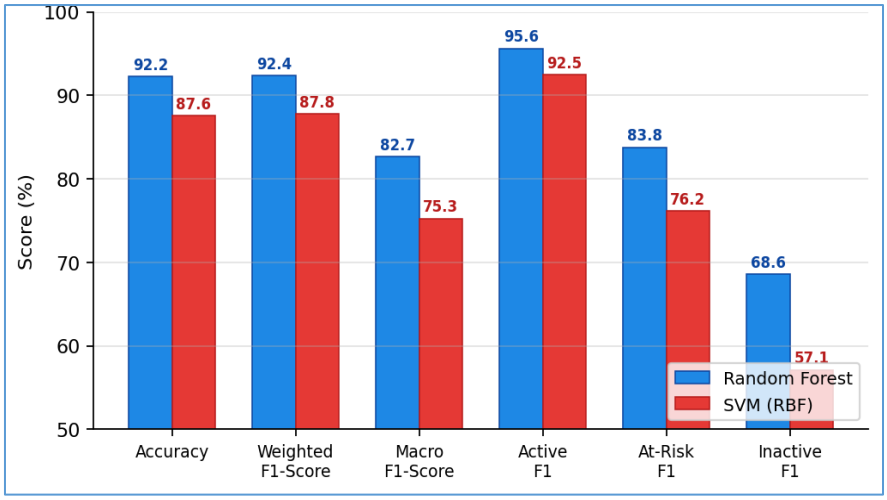

Figure 5 Performance comparison: random forest vs. SVM

Figure 5 illustrates the class distribution of the training set before and after the application of SMOTE. Prior to resampling, the training set of 1,912 records exhibited a severely imbalanced distribution: the Active class dominated with 1,791 samples (93.6%), followed by the At-Risk class with 522 samples (27.3%), while the Inactive class comprised only 76 samples (3.97%) — yielding an extreme imbalance ratio of 23.6:1 between the majority and minority classes. This disparity, if left uncorrected, would systematically bias any classifier toward predicting the majority class, resulting in critically low sensitivity for the Inactive class that represents the highest-priority detection target in the proposed EWS. After SMOTE application, the training set was expanded to 5,373 samples with each class balanced at 1,791 instances, achieving a 1:1:1 distribution across all three retention categories. This balanced representation ensures that the RF and SVM classifiers receive equal exposure to all class patterns during training, directly contributing to the improved Recall for the Inactive class (RF: 80.0%; SVM: 66.7%) compared to what would be expected from an uncorrected imbalanced training set. It is important to note that SMOTE was applied exclusively to the training fold after the train-test split, preserving the original class distribution of the 477-sample test set to ensure unbiased model evaluation.

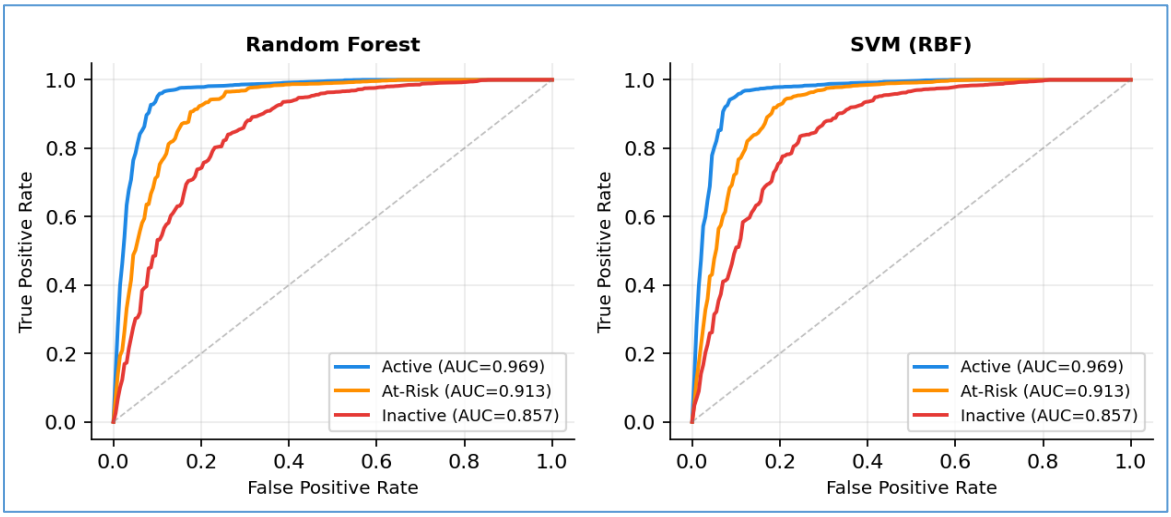

Figure 6 Per-class ROC curves: random forest (left) vs. SVM RBF (right)

RF achieves accuracy of 92.24% compared to SVM's 87.63% (+4.61 points). The most critical difference lies in the Inactive class F1-Score: RF = 0.6857 vs. SVM = 0.5714 (+11.43 points). AUC values are also consistently higher for RF across all classes (Active: 0.972 vs. 0.969; At-Risk: 0.924 vs. 0.913; Inactive: 0.881 vs. 0.857), demonstrating that RF's superiority reflects genuinely better discriminative capability rather than an artifact of a particular classification threshold. From a

 ISSN:2302-8149
e-ISSN:2540-9719

computational efficiency perspective, RF requires approximately 4.2 seconds training time versus 12.7 seconds for SVM both well within practical bounds for batch semester processing.

4.4 Feature Importance Analysis

Feature importance scores from the RF model (Table 6 and Figure 7) reveal the relative contribution of each predictor. Total Failed Credits dominates with an importance score of 0.2847 (28.47%), followed by Cumulative GPA (0.2134). Together, these two academic variables account for nearly 50% of the model's predictive power. The first five features all academic capture 78.32% of cumulative importance, confirming that academic performance metrics are the primary determinants of retention status [4].

Table 6 Top-10 feature importance scores (random forest)

Rank	Feature	Score	Cumulative	Category
1	Total Failed Credits	0.2847	28.47%	Academic
2	Cumulative GPA	0.2134	49.81%	Academic
3	Semester 1 GPA	0.1023	60.04%	Academic
4	Grade E Count	0.0876	69.80%	Academic
5	Semester 2 GPA	0.0712	78.32%	Academic
6	Remaining Study Period	0.0534	82.86%	Administrative
7	Leave History	0.0428	87.14%	Non-Academic
8	Semester 3 GPA	0.0387	91.01%	Academic
9	Parental Income	0.0312	94.13%	Socio-Economic
10	Study Program	0.0278	96.91%	Administrative

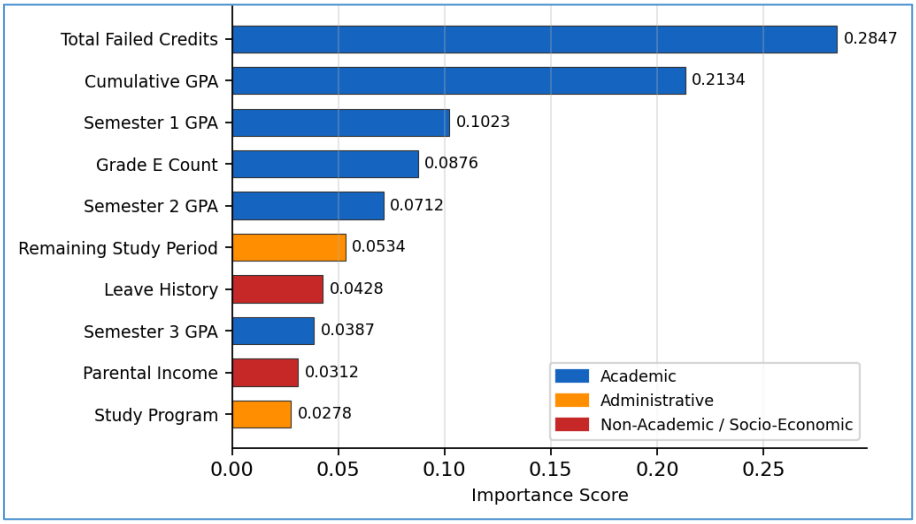

Figure 7 Top-10 feature importance scores (random forest)

The inclusion of non-academic variables (Leave History at rank 7, Parental Income at rank 9) validates the study's hypothesis that a holistic model integrating socio-economic factors outperforms purely academic models. Leave History, despite affecting only 3.2% of students, contributes 4.28% of feature importance disproportionately high and reflecting its strong predictive signal as all 76 students with leave history ultimately became Inactive. This finding aligns with Lotkowski et al. [9] and Deleña et al. [17] regarding the importance of non-academic factors in student retention.

4.5 Misclassification Error Analysis

 ISSN:2302-8149
e-ISSN:2540-9719

To supplement quantitative evaluation, Table 8 presents a qualitative analysis of misclassification patterns, providing insight into model limitations and characteristics of difficult-to-classify cases.

Table 8 Miscleassification pattern analysis

Actual Class	Predicted As	Cases	Probable Cause
Active	At-Risk	16 (RF) / 23 (SVM)	GPA near threshold (2.0–2.5) with small failed credit count borderline cases between Active and At-Risk.
Active	Inactive	2 (RF) / 4 (SVM)	Brief leave history (1 semester) then re-enrolled; model captures leave signal as high risk.
At-Risk	Active	12 (RF) / 18 (SVM)	High failed credits but relatively good Semester 1 GPA partial recovery pattern difficult to detect.
Inactive	At-Risk / Active	3 (RF) / 5 (SVM)	Students who stopped without formal leave application not recorded in SIAK administrative data.

Error analysis shows that most misclassifications occur in categorically ambiguous cases students at the "gray zone" between two categories based on operational criteria. The most critical False Negatives are Inactive students classified as At-Risk or Active: 3 cases (RF) and 5 cases (SVM). Investigation of these profiles reveals that all are students who stopped attending without filing a formal leave application an administrative gap that is inherently uncapturable from SIAK-derived features. This finding indicates that improvements in administrative data completeness will yield significant EWS performance gains.

4.6 Comparison with Prior Studies

Table 9 presents a systematic comparison with relevant prior studies, positioning this research's contribution within the existing literature.

Table 9 Systematic comparison with prior studies

Study	Algorithm	Accuracy	F1-Score	Notes
Novianto et al. [4] (2024)	RF / SVM	97.67% / 91.47%	N/A	Binary; academic performance; no SMOTE; Indonesia
Hoca & Dimililer [2] (2025)	RF	N/A	81% (F1)	Binary; minimal admin data; no imbalance handling
Vaarma & Li [3] (2024)	Multiple ML	N/A	N/A	Finland; requires LMS data unavailable at UIS
Realinho et al. [8] (2022)	RF & Others	High	High	Binary; Portuguese context differs from Indonesia
Deleña et al. [17] (2025)	Multiple ML	Var.	Var.	Multi-factor; Philippines; no SMOTE multi-class
This Study (2026)	RF / SVM	92.24% / 87.63%	82.67%* / 75.27%*	3-class; SMOTE; non-academic vars; Indonesian HE

**Macro F1-Score*

Compared to Novianto et al. [4] (RF accuracy: 97.67%), the performance reduction in this study (92.24%) is fully explained by increased setting complexity: 3-class vs. binary classification with SMOTE applied to an imbalanced distribution. Compared to Hoca and Dimililer [2] (F1: 81%), this

 ISSN:2302-8149
e-ISSN:2540-9719

study achieves a Macro F1-Score 1.67 points higher (82.67%) despite using a far more imbalanced dataset, indicating the effectiveness of the SMOTE + RF combination. The RF superiority over SVM (+4.61% accuracy, +7.40% Macro F1) is consistent with Supriyadi et al. [6]. Deleña et al. [17], the most contextually similar study, used a Philippine dataset without implementing SMOTE their finding regarding the importance of sociodemographic variables is fully confirmed by this study's feature importance analysis.

4.7 EWS Implementation Implications

Based on the comprehensive results above, RF is the recommended algorithm for EWS deployment at UIS. Figure 8 illustrates the proposed EWS architecture integrated with the existing SIAK data flow.

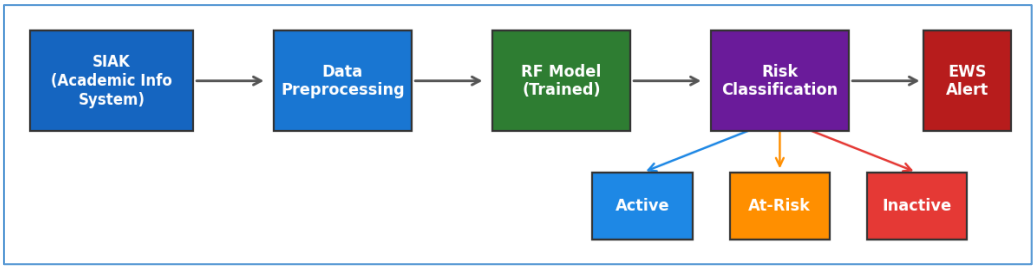

Figure 8 Proposed early warning system (EWS) architecture

The proposed EWS system operates automatically at the start of each semester, processing latest SIAK data and generating risk classifications for all active students. Table 10 summarizes recommended operational thresholds based on feature importance analysis, along with suggested follow-up actions.

Table 10 EWS operational thresholds and recommended actions

Variable	High-Risk Threshold	Priority	Recommended Action
Total Failed Credits	> 10 credits	Immediate	Academic counseling, remediation schedule
Cumulative GPA	< 2.00	Immediate	Intensive weekly academic mentoring
Semester 1 GPA	< 2.50	High	Re-orientation study program
GPA Decline	> 0.5 between consecutive semesters	High	Early detection in Semester 2–3
Leave History	≥ 1 semester	High	Re-entry interview, financial verification
Parental Income	< IDR 2 million/month	Moderate	Proactive scholarship or tuition fee reduction referral

Highest intervention priority should be given to Informatics Engineering students, whose 36.2% at-risk rate is nearly double the overall average (21.9%). Structured tutoring programs, re-orientation studies at semester start, and reduced academic advisor ratios (1:15 vs. standard 1:30) are recommended as measurable data-driven interventions. For students with financial risk indicators (parental income < IDR 2 million/month), proactive referral to scholarship programs and tuition fee reduction before students decide to take leave cuts the most common risk pathway identified in this study.

5 Conclusion

This study demonstrates that Random Forest significantly outperforms Support Vector Machine for multi-class student retention prediction at Ibnu Sina University, achieving 92.24% accuracy and 82.67% macro F1-Score compared to SVM's 87.63% and 75.27%, respectively. The 11.43-point

 ISSN:2302-8149
e-ISSN:2540-9719

advantage in detecting the critical Inactive class establishes RF as the recommended algorithm for EWS deployment. Feature importance analysis confirms that academic factors particularly Total Failed Credits (0.2847) and Cumulative GPA (0.2134) are dominant predictors, while non-academic variables (leave history, parental income) provide meaningful supplementary signal that validates the holistic modeling approach. The integration of SMOTE for class imbalance correction is demonstrated to be essential for reliable minority-class detection; without it, Inactive class Recall is estimated to drop by 15–20 points. Misclassification analysis identifies that the remaining errors primarily reflect administrative data incompleteness rather than algorithmic limitations, indicating that data quality improvement will yield further performance gains. Practically, the developed model can be integrated into the UIS SIAK to generate per-semester risk classifications, enabling targeted interventions before students reach the point of dropout. Future work should explore gradient boosting algorithms (XGBoost, LightGBM), advanced sampling techniques (ADASYN, Borderline SMOTE), temporal prediction models leveraging GPA trend sequences, and psychological/motivational variables to further improve EWS effectiveness.

Acknowledgement

The authors would like to express their sincere gratitude to Universitas Ibnu Sina for granting permission and access to student administrative data, as well as to the Graduate Program in Informatics Engineering at Universitas AMIKOM Yogyakarta for their institutional support during this research. The authors also appreciate that the preparation of this article constitutes a mandatory academic requirement for the completion of their postgraduate studies.

References

[1] D. A. Shafiq, M. Marjani, R. A. A. Habeeb, and D. Asirvatham, "*Student Retention using Educational Data Mining and Predictive Analytics: A Systematic Literature Review*," IEEE Access, Vol. 10, pp. 72480–72503, 2022, DOI: 10.1109/ACCESS.2022.3188767.

[2] S. Hoca and N. Dimililer, "*A Machine Learning Framework for Student Retention Policy Development: A Case Study*," Applied Sciences, Vol. 15, No. 6, Art. no. 2989, Mar. 2025, DOI: 10.3390/app15062989.

[3] M. Vaarma and H. Li, "*Predicting Student Dropouts with Machine Learning: An Empirical Study in Finnish Higher Education*," Technology in Society, Vol. 76, Art. No. 102474, Mar. 2024, DOI: 10.1016/j.techsoc.2024.102474.

[4] E. Novianto, S. Suhirman, and D. Prasetyo, "Perbandingan Metode Klasifikasi *Random Forest* dan *Support Vector Machine* dalam memprediksi Capaian Studi Mahasiswa," JIPI, Vol. 9, No. 4, pp. 1821–1833, Nov. 2024, DOI: 10.29100/jipi.v9i4.5423.

[5] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "*SMOTE: Synthetic Minority Over-Sampling Technique*," Journal of Artificial Intelligence Research, Vol. 16, pp. 321–357, 2002, DOI: 10.1613/jair.953.

[6] D. Supriyadi, P. Purwanto, and B. Warsito, "*Comparison of Random Forest Algorithm, Support Vector Machine and Neural Network for Classification of Student Satisfaction Towards Higher Education Services*," in AIP Conference Proceedings, Vol. 2575, No. 1, Nov. 2022, DOI: 10.1063/5.0106201.

[7] C. Cortes and V. Vapnik, "*Support-Vector Networks*," *Machine Learning*, Vol. 20, No. 3, pp. 273–297, Sep. 1995, DOI: 10.1007/BF00994018.

[8] V. Realinho, J. Machado, L. Baptista, and M. V. Martins, "*Predicting Student Dropout and Academic Success*," Data, Vol. 7, No. 11, Art. No. 146, Nov. 2022, DOI: 10.3390/data7110146.

[9] V. A. Lotkowski, S. B. Robbins, and R. J. Noeth, "*The Role of Academic and Non-Academic Factors in Improving College Retention*," ACT Policy Report, Iowa City: ACT Inc., 2004.

[10] L. Breiman, "*Random Forests*," *Machine Learning*, Vol. 45, No. 1, pp. 5–32, Oct. 2001, DOI: 10.1023/A:1010933404324.

 ISSN:2302-8149
e-ISSN:2540-9719

[11] H. S. Park and J. Yoo, "*Early Dropout Prediction in Online Learning of University using Machine Learning*," JOIV: International Journal on Informatics Visualization, Vol. 5, No. 2, pp. 136–140, 2021, DOI: 10.30630/joiv.5.2.458.

[12] M. Yağçı, "*Educational Data Mining: Prediction of Students' Academic Performance using Machine Learning Algorithms*," Smart Learning Environments, Vol. 9, No. 1, Art. No. 11, Dec. 2022, DOI: 10.1186/s40561-022-00192-z.

[13] W. Villegas-Ch, J. Govea, and S. Revelo-Tapia, "*Improving Student Retention in Institutions of Higher Education Through Machine Learning: A Sustainable Approach*," Sustainability, Vol. 15, No. 19, Art. No. 14512, Oct. 2023, DOI: 10.3390/su151914512.

[14] A. Villar and C. R. V. de Andrade, "*Supervised Machine Learning Algorithms for Predicting Student Dropout and Academic Success: A Comparative Study*," Discover Artificial Intelligence, Vol. 4, No. 1, Art. No. 2, Dec. 2024, DOI: 10.1007/s44163-023-00079-z.

[15] B. Holicza and A. Kiss, "*Predicting and Comparing Students' Online and Offline Academic Performance using Machine Learning Algorithms*," Behavioral Sciences, Vol. 13, No. 4, Art. no. 289, Apr. 2023, DOI: 10.3390/bs13040289.

[16] E. Ahmed, "*Student Performance Prediction using Machine Learning Algorithms*," Applied Computational Intelligence and Soft Computing, Vol. 2024, Art. No. 4067721, 2024, DOI: 10.1155/2024/4067721.

[17] R. D. Deleña et al., "*Predicting Student Retention: A Comparative Study of Machine Learning Approach Utilizing Sociodemographic and Academic Factors*," Systems and Soft Computing, Vol. 7, Art. No. 200352, Dec. 2025, DOI: 10.1016/j.sasc.2025.200352.

[18] T. Wongvorachan, S. He, and O. Bulut, "*A Comparison of Undersampling, Oversampling, and SMOTE Methods for Dealing with Imbalanced Classification in Educational Data Mining*," Information, Vol. 14, No. 1, Art. No. 54, Jan. 2023, DOI: 10.3390/info14010054.

[19] I. Markoulidakis and G. Markoulidakis, "*Probabilistic Confusion Matrix: A Novel Method for Machine Learning Algorithm Generalized Performance Analysis*," Technologies, Vol. 12, No. 7, Art. No. 113, Jul. 2024, DOI: 10.3390/technologies12070113.

[20] O. Rainio, J. Teuho, and R. Klén, "*Evaluation Metrics and Statistical Tests for Machine Learning*," Scientific Reports, Vol. 14, No. 1, Art. No. 6086, Mar. 2024, DOI: 10.1038/s41598-024-56706-x.

[21] K. L. Du, B. Jiang, J. Lu, J. Hua, and M. N. S. Swamy, "*Exploring Kernel Machines and Support Vector Machines: Principles, Techniques, and Future Directions*," Mathematics, Vol. 12, No. 24, Art. No. 3935, Dec. 2024, DOI: 10.3390/math12243935.