Vol. 8, No. 4, August 2026

ISI Journal of Information Systems and Informatics

Published By Asosiasi Doktor Sistem Informasi Indonesia

Cross-Cohort Early Prediction of Cluster-Derived GPA Trajectories Using First-Year Academic Data and Machine Learning

Achmad Ridwan[1], Tole Sutikno[2], Imam Riadi[3]

[1,2,3]Doctoral Program in Informatics, Faculty of Industrial Technology, Universitas Ahmad Dahlan, Indonesia

[2]Information Systems Study Program, Faculty of Science and Technology, Universitas Muhammadiyah Kudus, Kudus, Indonesia

Received:
May 29, 2026
Revised:
June 22 6, 2026
Accepted:
July 30, 2026
Published:
September 12, 2026

Corresponding Author:
Author Name*:
Achmad Ridwan
Email*:
achmadridwan@umkudus.ac.id

DOI:
10.63158/journalisi.v8i4.1871

Abstract. Early identification of unstable academic development can support timely advising, yet models evaluated only with random splits may not transfer across cohorts. This study examines whether data available after semester-2 grades are finalized can classify two K-Means-derived GPA trajectory labels—stable-improving and fluctuating—observed in semesters 3–7. A retrospective dataset of 1,362 anonymized students was analyzed. Labels were generated from standardized semester 3–7 GPA profiles, while predictors were limited to admission attributes and semester 1–2 records. Train-only preprocessing, RFECV, validation-set threshold tuning, calibration analysis, and GaussianNB, logistic regression, Random Forest, and SVM-RBF were evaluated using a stratified random test and a 2024/2025 administrative-cohort holdout. Random Forest achieved 0.831 balanced accuracy and 0.916 ROC-AUC on the random test. In the holdout, where fluctuating-label prevalence increased from 25.5% to 69.3%, Random Forest had the highest threshold-based balanced accuracy (0.581), whereas SVM-RBF had the highest ROC-AUC (0.763). Drift was substantial in student type, admission semester, status, and study-program composition. In this dataset, random splitting substantially overestimated cross-cohort performance. The observed degradation represents a moderate generalization challenge; these cluster-derived labels require cohort-aware recalibration and educational validation, and predictions should be used only as a human-supervised screening aid.

Keywords: Academic risk, Cohort drift, Cross-cohort validation, K-Means trajectory labeling, Student success prediction

1. INTRODUCTION

Higher education institutions increasingly use learning analytics and educational data mining to identify students who may benefit from academic support before difficulties become persistent. Learning analytics can transform institutional records into actionable information for study-success interventions [1], while systematic reviews show that grades, assessment records, engagement traces, and demographic attributes are frequently used to predict student outcomes [2], [3], [4]. GPA and accumulated credits remain among the most consistently associated academic indicators [5]. In practice, early-warning outputs are typically intended to support advisors in prioritizing outreach, reviewing student context, and selecting appropriate assistance rather than replacing academic judgment.

A single final GPA cannot represent how achievement changes over time. Semester-level GPA can reveal stable, improving, declining, or fluctuating patterns. Marbouti et al. used semester GPA to capture temporal effects in engineering student success [6], Baron et al. showed that both GPA level and direction distinguish meaningful academic signatures [7], and Nawa et al. categorized repeated GPA measurements into academic trajectories [8]. In the present study, the outcome is not a continuous trajectory forecast and is not a directly observed dropout or graduation-risk label. It is a binary, cluster-derived trajectory label. The stable-improving label describes a centroid whose mean GPA rises consistently from semester 3 to semester 7, whereas the fluctuating label describes a centroid with alternating decreases and recoveries. These names are post-hoc descriptions of K-Means centroids; declining, consistently low, or consistently high groups were not manually merged because the selected two-cluster solution did not produce them as separate institutional segments. The upstream clustering retained K=2 because it was the interpretable solution satisfying the 5% minimum cluster-representation rule; higher-K solutions isolated a 0.44% micro-cluster.

The operational prediction point is immediately after semester-2 grades are finalized and before semester-3 instruction begins for students whose semester records follow the conventional chronology. The fluctuating class is interpreted as academic instability that may justify advisor review, not as a deterministic negative outcome. Prior work indicates

that admission information and first-year grades can support early prediction [9], clustering can provide data-driven performance groups [10], and comparative classifiers can support graduation-related analysis [11]. However, random train-test partitions can overestimate operational performance when curricula, admission policies, program composition, grading practices, or student populations change between cohorts [3], [12]. Cohort drift can reflect curriculum, admission policy, program composition, grading practice, recognition-of-prior-learning pathways, and post-pandemic behavioral changes.

This study addresses three gaps. First, it separates the predictor window (information available through semester 2) from the trajectory-outcome window (semester 3–7). Second, preprocessing, RFECV, model fitting, calibration, and threshold selection are confined to development data. Third, performance is compared between a conventional stratified random test and an administrative-cohort holdout, supplemented by drift, calibration, subgroup, ablation, and chronology-sensitivity analyses. The objective is to determine how reliably early academic information can classify cluster-derived GPA trajectories and to identify why performance changes when the data distribution shifts across cohorts.

2. METHODS

2.1. Study Design, Data Source, and Trajectory-Label Construction

The study used a retrospective quantitative predictive-analytics design. The source comprised 1,412 graduate/student academic records from multiple study programs at Universitas Muhammadiyah Kudus. Direct identifiers were removed and replaced with anonymous IDs. Upstream trajectory construction followed the same validated data pipeline used in the preceding clustering study: 48 records with invalid academic-sequence or level information and two records with incomplete semester 3–7 GPA histories were excluded, leaving 1,362 complete records. Missing semester 3–7 values were therefore not imputed for target construction; incomplete trajectories were excluded before clustering.

The K-Means target used GPA from semesters 3, 4, 5, 6, and 7 only. Each semester feature was standardized by Z-score so that differences in semester-specific dispersion would

not dominate Euclidean distance. Candidate cluster counts K=2–6 were evaluated with inertia, Silhouette Score, Davies–Bouldin Index (DBI), Calinski–Harabasz Index (CHI), and a minimum 5% cluster-representation rule. In the evaluation stage, K-Means used k-means++ initialization, random_state=42, and n_init=50; the final K=2 model used n_init=100 to reduce sensitivity to initialization. K=2 achieved Silhouette=0.2769, DBI=1.6001, and CHI=364.326. Although K=6 had a higher Silhouette (0.3197), its smallest group represented only 0.44% of observations; inertia also decreased monotonically without a uniquely decisive elbow. K=2 was therefore selected as the most representative and interpretable institutional segmentation rather than as the mathematically highest-scoring configuration. With scikit-learn 1.8.0, K-Means used algorithm='lloyd', max_iter=300, and tol=1e-4. Table 1 reports the final centroids, and Figure 1 depicts their mean semester profiles.

Table 1. Cluster-Derived GPA Centroids and Standard Deviations

Trajectory label	N	S3	S4	S5	S6	S7
Stable - Rising	888	3.534±0.262	3.581±0.171	3.630±0.113	3.665±0.149	3.734±0.184
Fluctuating	474	3.398±0.206	3.351±0.208	3.489±0.356	3.350±0.305	3.367±0.343

The stable-improving centroid rises from 3.534 in semester 3 to 3.734 in semester 7. The fluctuating centroid falls in semester 4, recovers in semester 5, declines again in semester 6, and ends at 3.367 in semester 7. These centroid shapes—not an expert-assigned risk score or a slope threshold—define the two target labels. Consequently, the labels indicate patterns of academic stability/instability and require educational interpretation before any intervention is attached to them.

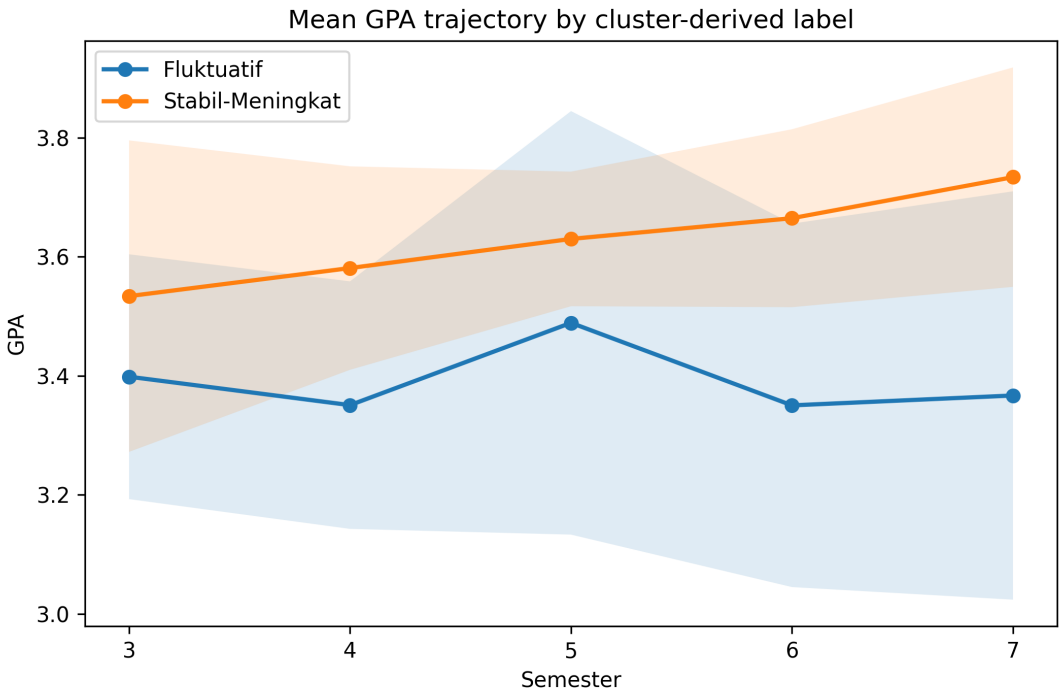

Figure 1. Mean semester 3–7 GPA profile of the two cluster-derived trajectory labels

2.2. Predictor Variables, Prediction Point, and Leakage Prevention

Predictors were restricted to information recorded no later than the end of semester 2: study program, program level, admission semester, initial status, student type, scholarship status, semester-1 GPA, semester-2 GPA, mean GPA across semesters 1–2, and GPA change from semester 1 to semester 2. Academic-entry cohort and anonymous ID were used only for splitting and record tracking. Semester 3–7 GPA, final GPA, accumulated credits, study duration, final semester, graduation-status variables, and the trajectory label were excluded from the feature matrix. Thus, the prediction task is classification of later trajectory labels from earlier information, not reconstruction of labels from the same variables that created them. The predictor groups, split-only variables, and excluded outcome fields are summarized in Table 2.

Table 2. Predictor Groups and Roles

Group	Variables	Role
Categorical	Study program, program level, admission semester, initial status, student type, scholarship	One-hot encoded predictors

Group	Variables	Role
Numeric	GPA S1, GPA S2, mean GPA S1–S2, GPA change S2–S1	Median imputation + Z-score
Split-only	Administrative academic-entry cohort	Cohort holdout definition
Excluded	GPA S3–S7, final GPA, credits, study duration, final status, target label	Leakage prevention

The 2024/2025 field refers to the administrative entry cohort stored in the institutional academic system, not the graduation year or data-extraction year. Importantly, this cohort contains special-program and recognition-of-prior-learning records for which semester-equivalent GPA histories were already present in the system at extraction. The 2024/2025 experiment is therefore interpreted as a retrospective cohort-transfer and distribution-shift test, not as a prospective prediction executed in 2024/2025 before those later records existed. To address the chronology concern directly, an additional sensitivity analysis used 2022/2023 as the latest chronology-compatible holdout and trained only on 2019/2020–2021/2022 cohorts.

2.3. Preprocessing, RFECV, and Classification Models

All preprocessing parameters were estimated from training data only. Missing categorical values were imputed with the training-mode category and encoded with one-hot encoding that ignored unseen categories. Missing numeric predictor values were imputed with the training median and standardized by Z-score. The random-split training pipeline produced 29 encoded features; the administrative-cohort development pipeline produced 28. RFECV retained 19 and 18 encoded features, respectively, for the optimized Gaussian Naive Bayes model.

RFECV used Random Forest because its nonlinear split structure can rank mixed numeric and one-hot-encoded predictors [13]. Feature selection remained inside development data, consistent with wrapper-selection principles [14]. RFECV was applied only to the optimized Gaussian Naive Bayes pipeline because this experiment specifically tests whether removing redundant or weak features improves a classifier whose conditional-independence assumption can be sensitive to correlated encodings. Logistic regression,

Random Forest, and RBF-SVM retained the full preprocessed feature matrix as benchmarking models rather than receiving the GaussianNB-specific wrapper selection. The comparison included linear, ensemble, kernel, and probabilistic classifiers [13], [15]. The complete RFECV and classifier hyperparameters, including class-weight settings and threshold-selection procedures, are reported in Table 3.

Table 3. Model and RFECV Hyperparameters

Component	Main settings	Class weighting	Selection/tuning
RFECV estimator	Random Forest; 120 trees; Gini; step=2; min_features=3; 3-fold stratified CV; macro-F1	balanced	Training data only
GaussianNB baseline	var_smoothing=1e-9	None	Threshold fixed at 0.50
Optimized GaussianNB	RFECV-selected features; var_smoothing=1e-9	None	Validation threshold
Logistic regression	C=1.0; lbfgs; max_iter=3000	balanced	Validation threshold
Random Forest	250 trees; Gini; max_depth=None; max_features=sqrt	balanced	Validation threshold
SVM-RBF	C=1.0; gamma=scale; probability=True	balanced	Validation threshold

Hyperparameters were fixed a priori for reproducibility rather than selected through a large nested hyperparameter search. The deliberate optimization target was the decision threshold. For models that produced probabilities, thresholds from 0.05 to 0.95 in 0.01 increments were evaluated on the validation set, and the threshold with the highest balanced accuracy was selected; F1 and recall were tie breakers. The GaussianNB baseline retained the conventional 0.50 threshold to preserve its role as an unoptimized baseline.

2.4. Validation, Drift, Calibration, Statistical Comparison, and Subgroups

The stratified random strategy used 60% training, 20% validation, and 20% untouched testing. The administrative-cohort strategy reserved all 2024/2025 records as the holdout

and split the remaining records into 80% training and 20% validation. The chronology sensitivity analysis trained/validated on cohorts through 2021/2022 and tested on 2022/2023. No test set was used for imputation, encoding, scaling, RFECV, fitting, or threshold selection. Sample sizes and class counts for both validation strategies are given in Table 4, while the leakage-controlled sequence of preprocessing, feature selection, calibration, and testing is depicted in Figure 2.

Table 4. Data Distribution in the Primary Validation Strategies

Strategy	Subset	N	Stable-improving	Fluctuating
Random	Training	816	532	284
Random	Validation	273	178	95
Random	Testing	273	178	95
Administrative cohort	Training	857	639	218
Administrative cohort	Validation	215	160	55
Administrative cohort	2024/2025 holdout	290	89	201

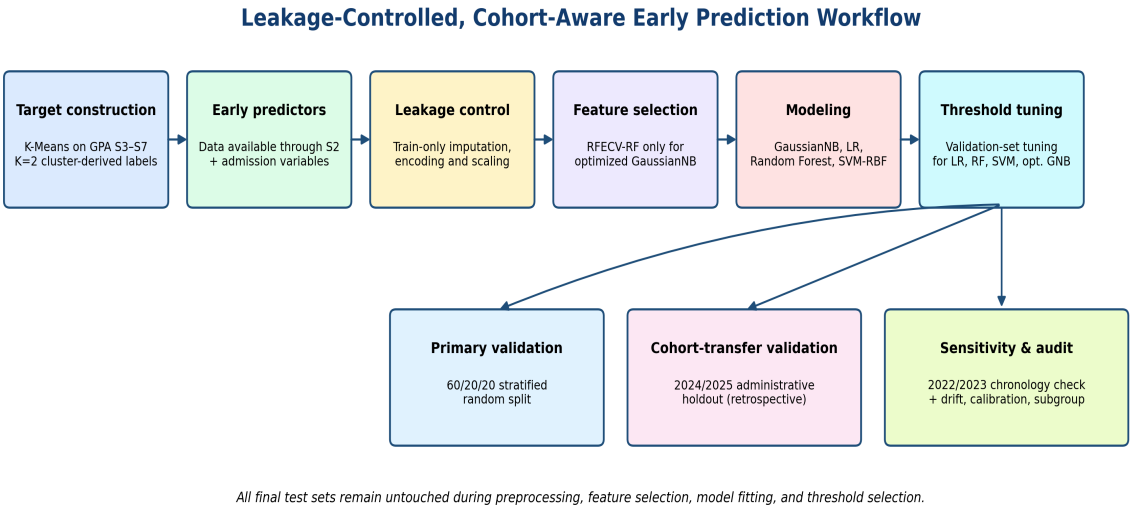

Figure 2. Leakage-controlled and cohort-aware research workflow

Evaluation included accuracy, balanced accuracy, fluctuating-class precision and recall, stable-class specificity, fluctuating-class F1, macro-F1, weighted-F1, Matthews correlation coefficient (MCC), ROC-AUC, and PR-AUC. ROC-AUC evaluates ranking discrimination [16], PR-AUC and class-specific measures are informative under class imbalance [17], and balanced accuracy reduces majority-class dominance [18]. Calibration was examined using

Brier score, expected calibration error (ECE; 10 probability bins), and reliability diagrams. Percentile bootstrap confidence intervals were calculated for principal metrics, and exact McNemar tests compared paired test errors between Random Forest and the two closest alternative models. Cohort drift was quantified using standardized mean differences and Kolmogorov–Smirnov tests for numeric predictors and Population Stability Index (PSI) for categorical distributions. Feature-group ablation compared first-year GPA variables only, admission variables only, study-program only, and the full feature set. Random Forest feature importance was reported together with a descriptive direction indicator based on the association between each encoded feature and predicted fluctuating probability. Subgroup performance was examined by study program and gender where both target classes were present; scholarship, student type, and program-level fairness could not be estimated in the 2024/2025 holdout because those attributes lacked meaningful variation.

2.5. Reproducibility and Data Governance

The analyses were reproduced in PyCharm using Python 3.13.5, scikit-learn 1.8.0, pandas 2.2.3, NumPy 2.3.5, SciPy 1.17.0, and Matplotlib 3.10.8. random_state=42 was fixed for all stochastic splitting and compatible model components. The implementation follows the pseudocode: (1) load anonymized data and audit forbidden outcome features; (2) construct development/test splits; (3) fit preprocessing on training data; (4) apply RFECV only within the optimized GaussianNB pipeline; (5) fit benchmark models; (6) tune eligible thresholds on validation data; (7) evaluate untouched tests; and (8) conduct drift, calibration, ablation, statistical, and subgroup audits. Use of the anonymized administrative dataset was institutionally authorized for research purposes; no direct identifiers were included in the analytical file. Student-facing deployment would additionally require transparent communication, governance, and consent procedures.

3. RESULTS AND DISCUSSION

3.1. Cohort Composition and Distribution Shift

To establish the magnitude of cohort shift before comparing classifiers, Table 5 contrasts class prevalence and key first-year characteristics between the development data and the 2024/2025 administrative holdout.

Table 5. Development Versus 2024/2025 Administrative Holdout Composition

Group	N	Fluct.	GPA1	GPA2	ΔGPA	Progsus	Non-scholar
Development	1072	25.5%	3.527	3.368	-0.159	15.5%	94.7%
2024/2025 administrative holdout	290	69.3%	3.527	3.543	0.017	100.0%	100.0%

The administrative holdout differed sharply from the development cohorts. The fluctuating-label prevalence increased from 25.5% to 69.3%. Mean semester-1 GPA was almost unchanged, but semester-2 GPA was higher in the holdout (3.543 versus 3.368), changing the average semester-2 minus semester-1 GPA difference from −0.159 to +0.017. All holdout students were coded as special-program (Progsus), S1, and non-scholarship, whereas the development data were more heterogeneous. This confirms that the holdout is not simply a later random sample; it represents a substantial administrative and program-composition shift. The variable-level magnitude of this shift is quantified in Table 6 using standardized mean differences, Kolmogorov-Smirnov statistics, and PSI values.

Table 6. Predictor Drift Between Development and 2024/2025 Holdout

Variable	Type	SMD	KS	PSI
IPK_Smt_1	Numeric	-0.001	0.165	—
IPK_Smt_2	Numeric	0.574	0.287	—
Rata_IPK_Smt_1_2	Numeric	0.534	0.289	—
Perubahan_IPK_Smt_2_1	Numeric	0.444	0.183	—
Prodi	Categorical	—	—	1.633
Jenjang	Categorical	—	—	0.000
Semester_Masuk	Categorical	—	—	6.167
Status_Awal	Categorical	—	—	4.339
Tipe_Mahasiswa	Categorical	—	—	12.673
Beasiswa	Categorical	—	—	0.573

The largest categorical shifts occurred for student type (PSI=12.673), admission semester (6.167), initial status (4.339), and study program (1.633). Numeric drift was moderate for

semester-2 GPA (SMD=0.574), mean first-year GPA (0.534), and GPA change (0.444). These shifts provide an empirical explanation for the cross-cohort degradation rather than treating it as random model failure.

3.2. Classification Performance, Thresholds, and Calibration

Threshold-dependent classification results for the random test are reported in Table 7, and the complementary threshold-independent ranking and probability-calibration results are reported in Table 8.

Table 7. Extended Performance on the Stratified Random Test

Model	Th	Acc	BalAcc	Prec+	Rec+	Spec	F1+	MacroF1	WF1	MCC
Random Forest	0.31	0.824	0.831	0.704	0.853	0.809	0.771	0.814	0.827	0.638
SVM-RBF	0.27	0.813	0.820	0.690	0.842	0.798	0.758	0.803	0.817	0.617
Logistic Regression	0.54	0.788	0.786	0.667	0.779	0.792	0.718	0.774	0.791	0.554
Optimized GaussianNB	0.95	0.810	0.763	0.795	0.611	0.916	0.690	0.776	0.803	0.566
GaussianNB Baseline	0.50	0.465	0.575	0.389	0.937	0.213	0.549	0.446	0.414	0.195

Table 8. Ranking and Calibration Metrics on the Stratified Random Test

Model	ROC	PR	Brier	ECE	95% CI BalAcc	95% CI ROC
Random Forest	0.916	0.873	0.111	0.061	0.790-0.877	0.882-0.949
SVM-RBF	0.900	0.830	0.128	0.051	0.776-0.863	0.863-0.932
Logistic Regression	0.851	0.708	0.155	0.108	0.721-0.828	0.801-0.887
Optimized GaussianNB	0.776	0.693	0.203	0.194	0.708-0.820	0.716-0.843
GaussianNB Baseline	0.789	0.707	0.535	0.535	0.543-0.618	0.737-0.845

With validation-selected thresholds, Random Forest produced the strongest random-test thresholded performance: accuracy=0.824, balanced accuracy=0.831, fluctuating recall=0.853, specificity=0.809, F1=0.771, and MCC=0.638. Its ROC-AUC was 0.916 and PR-AUC 0.873. SVM-RBF was close in balanced accuracy (0.820) and ROC-AUC (0.900). Exact McNemar comparison between Random Forest and SVM-RBF was not significant

(p=0.728), indicating that their paired error counts were not demonstrably different on this test set. The same diagnostic framework is applied to the shifted administrative cohort: Table 9 reports threshold-dependent classification performance and Table 10 reports ROC-AUC, PR-AUC, Brier score, ECE, and confidence intervals.

Table 9. Extended Performance on the 2024/2025 Administrative Holdout

Model	Th	Acc	BalAcc	Prec+	Rec+	Spec	F1+	MacroF1	WF1	MCC
Random Forest	0.31	0.710	0.581	0.733	0.915	0.247	0.814	0.579	0.670	0.220
Logistic Regression	0.43	0.607	0.504	0.695	0.771	0.236	0.731	0.500	0.589	0.008
Optimized GaussianNB	0.14	0.693	0.500	0.693	1.000	0.000	0.819	0.409	0.567	0.000
SVM-RBF	0.09	0.693	0.500	0.693	1.000	0.000	0.819	0.409	0.567	0.000
GaussianNB Baseline	0.50	0.562	0.449	0.665	0.741	0.157	0.701	0.441	0.541	-0.112

Table 10. Ranking and Calibration Metrics on the 2024/2025 Administrative Holdout

Model	ROC	PR	Brier	ECE	95% CI BalAcc	95% CI ROC
Random Forest	0.689	0.813	0.212	0.133	0.544-0.631	0.630-0.772
Logistic Regression	0.623	0.816	0.228	0.131	0.445-0.564	0.532-0.685
Optimized GaussianNB	0.572	0.794	0.306	0.306	0.500-0.500	0.493-0.630
SVM-RBF	0.763	0.840	0.192	0.178	0.500-0.500	0.669-0.827
GaussianNB Baseline	0.496	0.725	0.438	0.438	0.408-0.505	0.437-0.560

The model ranking changed under the administrative cohort shift. Random Forest retained the highest threshold-based balanced accuracy (0.581) and classified 184 of 201 fluctuating students, but specificity for the stable-improving class fell to 0.247. SVM-RBF achieved the highest ROC-AUC (0.763) and PR-AUC (0.840), showing that its probability ranking still contained useful signal; nevertheless, the validation-selected threshold of 0.09 classified every holdout record as fluctuating and therefore yielded balanced accuracy=0.500. This pattern indicates a threshold/calibration failure under distribution shift rather than a complete loss of ranking discrimination. Random Forest is therefore described as the most balanced thresholded model, not universally the best model.

Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

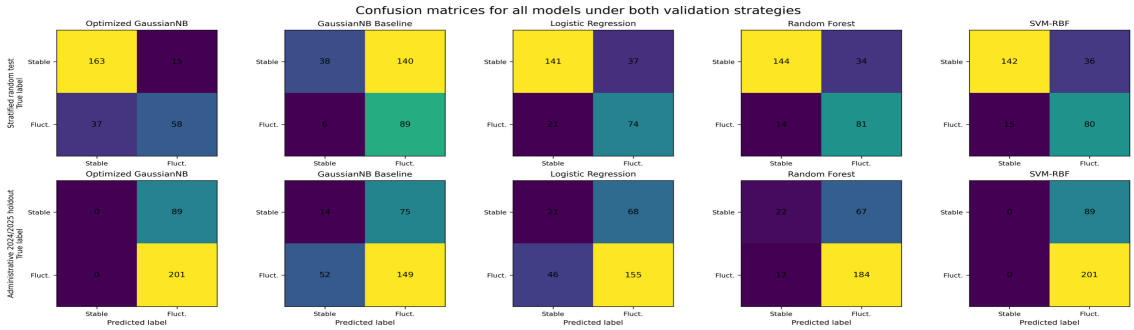

Figure 3. Confusion matrices for all models under random testing (left) and administrative-cohort testing (right)

Figures 3–7 summarize model-transfer diagnostics: Figure 3 shows confusion patterns, Figure 4 ROC curves, Figure 5 precision-recall curves, Figure 6 calibration, and Figure 7 threshold sensitivity.

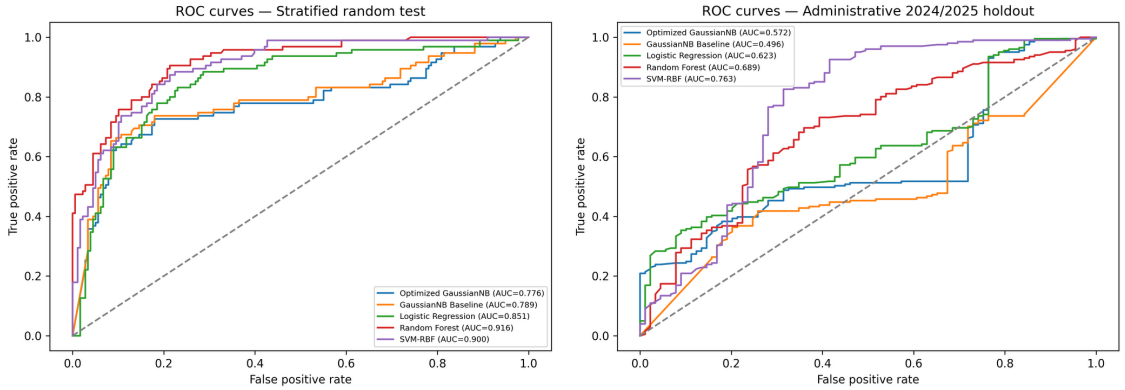

Figure 4. ROC curves for all models under both validation strategies

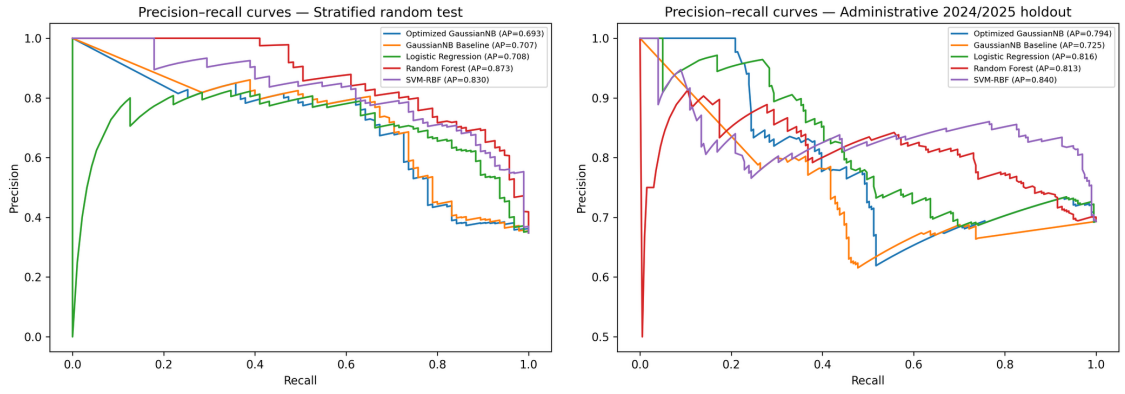

Figure 5. Precision–recall curves for all models under both validation strategies

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

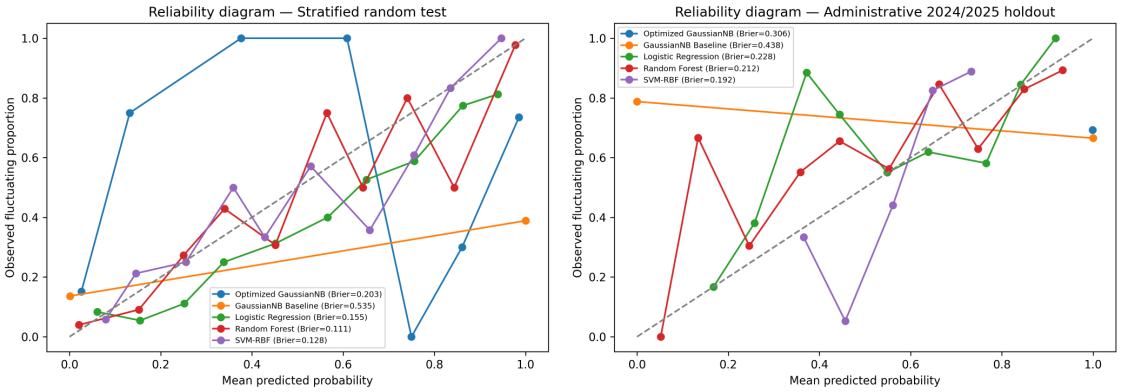

Figure 6. Reliability diagrams and Brier scores under both validation strategies

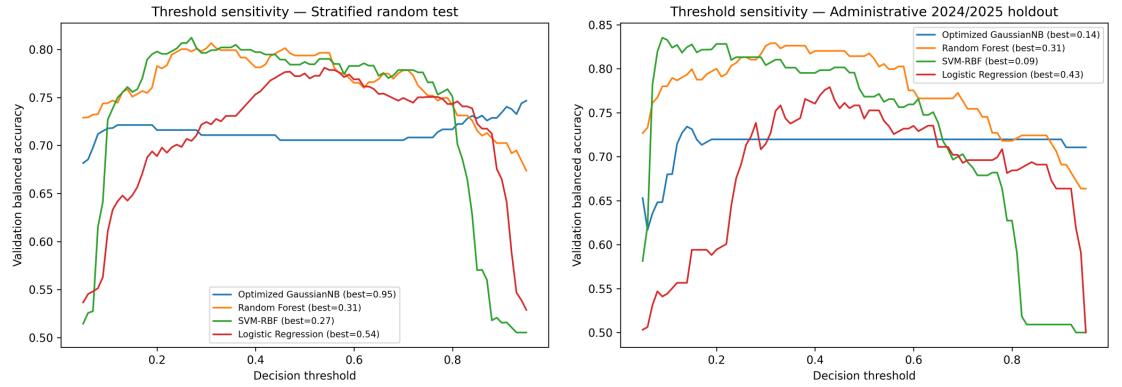

Figure 7. Validation balanced accuracy across probability thresholds

Thresholds changed materially across development settings. The optimized GaussianNB threshold moved from 0.95 in the random experiment to 0.14 in the administrative-cohort experiment; SVM-RBF moved from 0.27 to 0.09, while Random Forest remained at 0.31. The low thresholds in the administrative setting were selected because the development cohort and holdout differed strongly in class and program composition. The calibration plots and ECE values confirm that probability calibration was less reliable under shift. Operational use would therefore require cohort-specific recalibration or a monitored updating policy rather than a fixed threshold carried forward indefinitely.

3.3. Feature Importance, Ablation, and Subgroup Audit

To examine whether model transfer depended on particular predictors, Table 11 lists the leading Random Forest features and their descriptive directions; Figure 8 then compares feature-importance profiles between the random and administrative-cohort experiments.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

Table 11. Top Random Forest Features in the Administrative-Cohort Experiment

Encoded feature	Importance	Descriptive direction
IPK_Smt_1	0.169	Higher → Stable-Improving
IPK_Smt_2	0.164	Higher → Fluctuating
Rata_IPK_Smt_1_2	0.159	Higher → Stable-Improving
Perubahan_IPK_Smt_2_1	0.156	Higher → Fluctuating
Tipe_Mahasiswa_Reguler	0.065	Present → Stable-Improving
Tipe_Mahasiswa_Progsus	0.056	Present → Stable-Improving
Prodi_Keperawatan	0.041	Present → Fluctuating
Status_Awal_Pindahan	0.032	Present → Fluctuating
Status_Awal_Baru	0.029	Present → Fluctuating
Prodi_Farmasi	0.029	Present → Stable-Improving

The four numeric first-year GPA descriptors occupied the top importance positions in the administrative-cohort Random Forest. GPA change from semester 1 to semester 2 was positively associated with fluctuating probability in the holdout, while higher semester-1 GPA was associated with lower fluctuating probability. Student-type and study-program indicators also contributed, which supports the reviewer concern that the model may partially learn institutional composition rather than a fully transportable academic mechanism. Because impurity importance is not causal, the direction column is descriptive and should not be interpreted as an intervention effect.

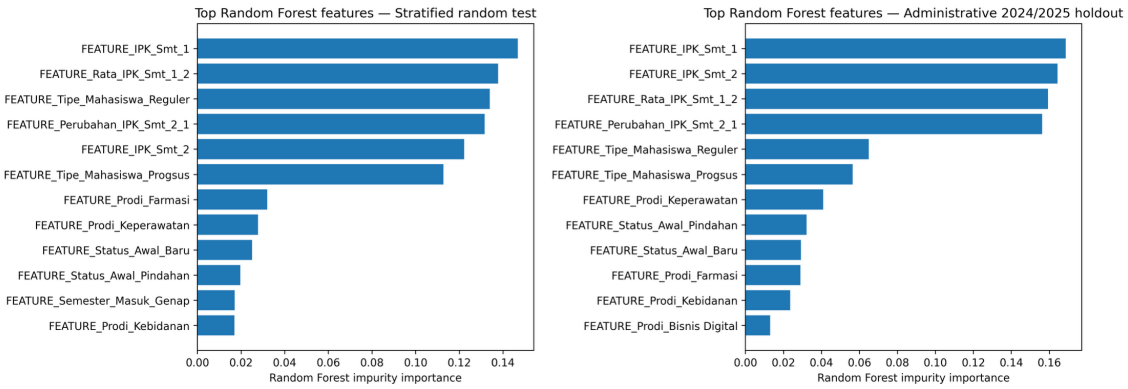

Figure 8. Random Forest feature importance under random and administrative-cohort evaluation

Feature-group dependence is tested explicitly in Table 12 by comparing first-year GPA, admission variables, study-program indicators, and the full feature set under both validation settings.

Table 12. Random Forest Feature-Group Ablation

Setting	Feature group	BalAcc	ROC	PR	MCC
Random	First-year GPA only	0.736	0.798	0.667	0.453
Random	Admission variables only	0.742	0.745	0.609	0.488
Random	Study-program only	0.581	0.589	0.409	0.157
Random	Full feature set	0.831	0.916	0.873	0.638
Admin holdout	First-year GPA only	0.503	0.532	0.698	0.006
Admin holdout	Admission variables only	0.449	0.447	0.671	-0.112
Admin holdout	Study-program only	0.575	0.496	0.729	0.140
Admin holdout	Full feature set	0.581	0.689	0.813	0.220

Ablation confirms that the full feature set was strongest in the random experiment (balanced accuracy=0.831; ROC-AUC=0.916), whereas first-year GPA alone reached 0.736 balanced accuracy. Under the 2024/2025 administrative shift, the full set retained the highest ROC-AUC (0.689) but balanced accuracy fell to 0.581. Study-program-only performance was weak in the random test (ROC-AUC=0.589) and did not provide reliable transfer by itself, but its interaction with cohort composition remains a plausible shortcut pathway. This supports retaining study program only with explicit drift and subgroup monitoring. Because aggregate performance can conceal subgroup failures, Table 13 reports the corresponding administrative-holdout audit by study program and gender for groups with adequate representation.

Table 13. Random Forest Subgroup Performance in the 2024/2025 Holdout

Attribute	Group	N	Fluct.	Recall+	Spec	BalAcc	ROC
Study_Program	Pharmacy	23	2	0.500	0.524	0.512	0.500
Study_Program	Nutrition	30	14	0.786	0.000	0.393	0.446
Study_Program	Midwifery	119	94	0.904	0.400	0.652	0.781
Study_Program	Nursing	118	91	0.956	0.037	0.497	0.554
Gender_Fairness	Male	44	35	0.971	0.333	0.652	0.649
Gender_Fairness	Female	246	166	0.904	0.237	0.571	0.692

Subgroup behavior varied substantially. Balanced accuracy ranged from 0.393 in Nutrition to 0.652 in Midwifery among the four study programs represented in the administrative holdout. Gender-specific balanced accuracy was 0.652 for men and 0.571 for women, but these estimates have unequal sample sizes and should not be interpreted as evidence of discrimination without formal fairness testing. Scholarship status, student type, and program level could not be meaningfully compared in this holdout because all records were non-scholarship, special-program, and S1. The absence of subgroup variation is itself a limitation and motivates fairness evaluation on future cohorts with broader representation.

3.4. Discussion

The random-test results confirm that early academic information contains signal about later cluster-derived GPA patterns, consistent with studies using early assessments and institutional attributes [19] and with trajectory-oriented research that models repeated academic performance [20]. However, the cross-cohort analysis changes the interpretation: model quality depends not only on discrimination but also on whether the feature and class distributions remain stable. This finding aligns with learning-analytics work showing that changing course and cohort contexts can produce performance drift and unstable feature relevance [36].

The 2024/2025 holdout should not be interpreted as a prospective future cohort in the strict chronological sense. It is an administrative entry cohort in a retrospective dataset and is dominated by special-program records whose semester-equivalent academic histories were already stored in the system. This composition explains part of the large PSI values and class shift. To test whether the main conclusion depends entirely on that administrative artifact, the chronology-compatible sensitivity analysis held out 2022/2023 and trained only on cohorts through 2021/2022. Random Forest reached ROC-AUC=0.699 and balanced accuracy=0.608, whereas SVM-RBF reached ROC-AUC=0.760 but lower thresholded utility. The sensitivity result does not establish deployment readiness, but it supports the broader conclusion that cohort transfer is harder than random-split evaluation.

The SVM result is particularly informative. Its administrative-holdout ROC-AUC was higher than Random Forest, yet thresholded predictions collapsed to the fluctuating class. This distinction shows why ranking, calibration, and decision thresholds must be evaluated separately. Broader reviews of student-performance prediction emphasize that validation design and dataset characteristics can materially influence apparent model performance [3], [4], [32], [35]. Annual deployment should therefore estimate drift before scoring, recalibrate probabilities or thresholds on labeled recent data when available, and suspend automated alerts when calibration deteriorates. Because ROC-AUC is threshold independent, SVM-RBF could retain AUC=0.763 while failing operationally. In the holdout, its validation-selected threshold of 0.09, together with ECE=0.178 and Brier=0.192, shifted predictions toward the fluctuating class and produced zero specificity. This indicates threshold/calibration failure under cohort shift rather than complete loss of ranking discrimination.

The practical costs of classification errors are asymmetric. A false negative can miss a student whose unstable trajectory warrants support, whereas a false positive can cause unnecessary outreach, labeling concerns, and inefficient use of advisor time. Because this study has no validated monetary or educational utility weights, it does not claim an optimal cost-sensitive threshold. Predictions should instead act as prompts for human review. Students also expect transparency and consent in institutional analytics [21], [22], and privacy requires governance beyond technical anonymization [23]. Federated approaches may support later multi-institutional validation without centralizing raw records [24]. A responsible advising workflow should use predictions only to prioritize review. Advisors should check current academic context before outreach, may override the model with documented rationale, and must avoid punitive action based solely on a prediction. Thresholds and calibration should be reassessed each academic cycle when recent labeled data are available.

The cluster-derived target is an analytical construct, not an independently validated educational-risk outcome. The related cluster-label framework demonstrates how unsupervised patterns can be transferred into supervised classification when predictor and label-construction features are separated [25], but educational validity must still be established through advisor review and intervention outcomes. Recent prediction studies

show the value of early institutional features [26], [27], [28], [29], feature-selection and ensemble comparisons [30], [31], [32], [33], [34], [35], model-stability analysis [36], explainable prediction [37], and fairness-aware evaluation [38]. Several author-involved studies also provide methodological continuity in feature selection [39], unsupervised pattern discovery [40], classification [41], and student-oriented intelligent-system development [42]; these are cited as methodological continuity rather than direct evidence for academic-trajectory prediction. For operational use, predictions should therefore remain human supervised and be interpreted together with transparent advising procedures [21], [22], [23]. Author-involved references are retained only as brief methodological continuity, not as evidence for the cross-cohort claims.

The main limitations are the single-institution sample, retrospective cohort definition, cluster-derived outcome, strong program-composition shift, absence of attendance/engagement/psychosocial variables, incomplete fairness coverage, and potential study-program shortcut learning. The institutional K-Means procedure was reproducible and transparent, but the two trajectory labels should not be interpreted as universal student types. External validation, prospective cohort follow-up, probability recalibration, subgroup fairness testing, explainability, and evaluation of actual advising outcomes are necessary before operational deployment.

4. CONCLUSION

This study developed a leakage-controlled framework for classifying two cluster-derived GPA trajectory labels from information available through semester 2. Random Forest showed strong within-distribution performance on the stratified random test (balanced accuracy=0.831; ROC-AUC=0.916), but its threshold-based balanced accuracy declined to 0.581 in the shifted 2024/2025 administrative holdout; SVM-RBF retained higher ranking discrimination in that holdout (ROC-AUC=0.763) while failing at the selected decision threshold. The findings demonstrate in this dataset that random splits substantially overestimate cohort-transfer performance and that ranking, calibration, class distribution, and threshold behavior must be evaluated separately. Because the target is a K-Means-derived representation of stable-improving versus fluctuating GPA patterns—not a directly observed dropout or graduation-risk label—the model should be used only

as a human-supervised screening aid and never as the sole basis for academic decisions. Current limitations include single-institution data, program-composition drift, possible shortcut learning, retrospective special-program records, and incomplete fairness testing. Before deployment, the framework requires prospective and external validation, recurring drift monitoring and recalibration, subgroup fairness analysis, explainability, student-facing transparency and consent procedures, and evaluation of whether advisor interventions improve real educational outcomes. Institutional use should include student-facing transparency on screening purpose, data use, human review, and routes for clarification or appeal.

ACKNOWLEDGMENT

The authors thank Universitas Muhammadiyah Kudus for facilitating institutionally authorized access to anonymized academic data for research purposes. The authors also acknowledge the Doctoral Program in Informatics, Universitas Ahmad Dahlan, for academic guidance during the research development.

REFERENCES

[1] D. Ifenthaler and J. Yau, "Utilising learning analytics to support study success in higher education: a systematic review," *Educ. Technol. Res. Dev.*, vol. 68, pp. 1961–1990, 2020.

[2] A. Namoun and A. Alshanqiti, "Predicting Student Performance Using Data Mining and Learning Analytics Techniques: A Systematic Literature Review," *Appl. Sci.*, vol. 11, no. 1, p. 237, 2021.

[3] E. A. Alyahyan and D. Düştegör, "Predicting academic success in higher education: literature review and best practices," *Int. J. Educ. Technol. High. Educ.*, vol. 17, pp. 1–21, 2020.

[4] N. Sghir, A. Adadi, and M. Lahmer, "Recent advances in Predictive Learning Analytics: A decade systematic review (2012--2022)," *Educ. Inf. Technol.*, vol. 28, pp. 8299–8333, 2023.

[5] Á. Kocsis and G. Molnár, "Factors influencing academic performance and dropout rates in higher education," *Oxford Rev. Educ.*, vol. 51, pp. 414–432, 2024.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

[6] F. Marbouti, J. Ulas, and C. Wang, "Academic and Demographic Cluster Analysis of Engineering Student Success," *IEEE Trans. Educ.*, vol. 64, pp. 261–266, 2021.

[7] T. Baron *et al.*, "Signatures of medical student applicants and academic success," *PLoS One*, vol. 15, 2020.

[8] N. Nawa *et al.*, "Associations between demographic factors and the academic trajectories of medical students in Japan," *PLoS One*, vol. 15, no. 5, p. e0233371, 2020.

[9] E. Alhazmi and A. M. Sheneamer, "Early Predicting of Students Performance in Higher Education," *IEEE Access*, vol. 11, pp. 27579–27589, 2023.

[10] A. F. Mohamed Nafuri, N. S. Sani, N. F. A. Zainudin, A. H. A. Rahman, and M. Aliff, "Clustering Analysis for Classifying Student Academic Performance in Higher Education," *Appl. Sci.*, vol. 12, no. 19, p. 9467, 2022.

[11] A. Ridwan, T. Sutikno, I. Riyadi, and W. C. Wahyudin, "On-Time Student Graduation Prediction Modeling: A Comparative Analysis of Naive Bayes Algorithm and Other Data Mining Classifications: Pemodelan Prediksi Kelulusan Mahasiswa Tepat Waktu: Analisis Komparatif Algoritma Naive Bayes Dan Klasifikasi Data Mining Lainnya," *JOINCS (Journal Informatics, Network, Comput. Sci.*, vol. 8, no. 2, pp. 128–135, 2025.

[12] C. Foster and P. Francis, "A systematic review on the deployment and effectiveness of data analytics in higher education to improve student outcomes," *Assess. Eval. High. Educ.*, vol. 45, pp. 822–841, 2020.

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

[14] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Mach. Learn.*, vol. 46, no. 1, pp. 389–422, 2002.

[15] C. Cortes and V. Vapnik, "Support-Vector Networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.

[16] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, 2006, doi: 10.1016/j.patrec.2005.10.010.

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

[18] K. H. Brodersen, C. S. Ong, K. E. Stephan, and J. M. Buhmann, "The balanced accuracy and its posterior distribution," in *20th International Conference on Pattern Recognition*, 2010, pp. 3121–3124.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

[19] G. Feng, M. Fan, and Y. Chen, "Analysis and Prediction of Students' Academic Performance Based on Educational Data Mining," *IEEE Access*, vol. 10, pp. 19558–19571, 2022.

[20] H. Lim, S. Kim, K.-M. Chung, K. Lee, T. Kim, and J. Heo, "Is college students' trajectory associated with academic performance?," *Comput. Educ.*, vol. 178, p. 104397, 2022.

[21] K. M. L. Jones *et al.*, "We're being tracked at all times: Student perspectives of their privacy in relation to learning analytics in higher education," *J. Assoc. Inf. Sci. Technol.*, vol. 71, pp. 1044–1059, 2020.

[22] K. M. L. Jones *et al.*, "Transparency and Consent: Student Perspectives on Educational Data Analytics Scenarios," *portal Libr. Acad.*, vol. 23, pp. 485–515, 2023.

[23] P. Prinsloo, S. Slade, and M. Khalil, "The answer is (not only) technological: Considering student data privacy in learning analytics," *Br. J. Educ. Technol.*, vol. 53, pp. 876–893, 2022.

[24] C. Fachola, A. Tornaría, P. Bermolen, G. Capdehourat, L. Etcheverry, and M. Fariello, "Federated Learning for Data Analytics in Education," *Data*, vol. 8, p. 43, 2023.

[25] W. C. Wahyudin, T. Sutikno, and R. Umar, "A Cluster-Label-Based Framework for Water Quality Risk Pattern Classification Using Naive Bayes and Random Forest," *J. Ilm. Ilmu Terap. Univ. Jambi*, vol. 10, no. 3, pp. 1511–1522, 2026, doi: 10.22437/jiituj.v10i3.57126.

[26] M. Yakubu and A. Abubakar, "Applying machine learning approach to predict students' performance in higher educational institutions," *Kybernetes*, vol. 51, no. 2, pp. 916–934, 2021, doi: 10.1108/K-12-2020-0865.

[27] M. V Martins, D. Tolledo, J. Machado, L. M. T. Baptista, and V. Realinho, "Early Prediction of Student's Performance in Higher Education: A Case Study," in *Trends and Applications in Information Systems and Technologies, WorldCIST 2021*, 2021, vol. 1365, pp. 166–175, doi: 10.1007/978-3-030-72657-7_16.

[28] M. Yağcı, "Educational data mining: prediction of students' academic performance using machine learning algorithms," *Smart Learn. Environ.*, vol. 9, p. 11, 2022, doi: 10.1186/s40561-022-00192-z.

[29] Y. S. Balcıoğlu and M. Artar, "Predicting academic performance of students with machine learning," *Inf. Dev.*, vol. 41, no. 3, pp. 896–915, 2025, doi: 10.1177/02666669231213023.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

[30] K. Mahawar and P. Rattan, "Empowering education: Harnessing ensemble machine learning approach and ACO-DT classifier for early student academic performance prediction," *Educ. Inf. Technol.*, vol. 30, pp. 4639–4667, 2025, doi: 10.1007/s10639-024-12976-6.

[31] N. Butt, Z. Mahmood, K. Shakeel, S. Alfarhood, M. S. Safran, and I. Ashraf, "Performance Prediction of Students in Higher Education Using Multi-Model Ensemble Approach," *IEEE Access*, vol. 11, pp. 136091–136108, 2023, doi: 10.1109/ACCESS.2023.3336987.

[32] S. Alwarthan, N. Aslam, and I. U. Khan, "Predicting Student Academic Performance at Higher Education Using Data Mining: A Systematic Review," *Appl. Comput. Intell. Soft Comput.*, vol. 2022, p. 8924028, 2022, doi: 10.1155/2022/8924028.

[33] E. Ahmed, "Student Performance Prediction Using Machine Learning Algorithms," *Appl. Comput. Intell. Soft Comput.*, vol. 2024, p. 4067721, 2024, doi: 10.1155/2024/4067721.

[34] A. Harif and M. A. Kassimi, "Predictive Modeling of Student Performance Using RFECV-RF for Feature Selection and Machine Learning Techniques," *Int. J. Adv. Comput. Sci. Appl.*, vol. 15, no. 7, pp. 228–237, 2024, doi: 10.14569/IJACSA.2024.0150723.

[35] S. Batool, J. Rashid, M. W. Nisar, J. Kim, H.-Y. Kwon, and A. Hussain, "Educational data mining to predict students' academic performance: A survey study," *Educ. Inf. Technol.*, vol. 28, pp. 905–971, 2023, doi: 10.1007/s10639-022-11152-y.

[36] E. Tiukhova *et al.*, "Explainable Learning Analytics: Assessing the stability of student success prediction models by means of explainable AI," *Decis. Support Syst.*, vol. 182, p. 114229, 2024, doi: 10.1016/j.dss.2024.114229.

[37] R. Alamri and B. Alharbi, "Explainable Student Performance Prediction Models: A Systematic Review," *IEEE Access*, vol. 9, pp. 33132–33143, 2021, doi: 10.1109/ACCESS.2021.3061368.

[38] F. Arévalo-Cordovilla and M. Peña, "Evaluating ensemble models for fair and interpretable prediction in higher education using multimodal data," *Sci. Rep.*, vol. 15, art. 29420, 2025, doi: 10.1038/s41598-025-15388-9.

[39] W. C. Wahyudin, T. Sutikno, R. Umar, and A. Ridwan, "Comparison of Data Mining Model Performance in Heart Disease Detection with Feature Selection Application," *JOINCS (Journal Informatics, Network, Comput. Sci.*, vol. 8, no. 1, pp. 87–93, 2025, doi: 10.21070/joincs.v8i1.1669.

[40] W. C. Wahyudin, T. Sutikno, and R. Umar, "Identification of Bengawan Solo River Water Quality Patterns Using K-Means Clustering Based on Physicochemical and Environmental Parameters," *JOINCS (Journal Informatics, Network, Comput. Sci.*, vol. 9, no. 1, pp. 43–48, 2026, doi: 10.21070/joincs.v9i1.1710.

[41] L. N. Hakim, F. M. Hana, and W. C. Wahyudin, "Klasifikasi Komentar Toksik Berbahasa Indonesia di Media Sosial Berbasis Fine-Tuning IndoBERT," *JURIKOM (Jurnal Ris. Komputer)*, vol. 13, no. 1, pp. 202–209, 2026, doi: 10.30865/jurikom.v13i1.9449.

[42] S. P. Afrisia, F. M. Hana, and W. C. Wahyudin, "Implementasi Metode Long Short Term Memory (LSTM) pada Chatbot Kesehatan Mental Mahasiswa," *Sainteks*, vol. 21, no. 2, pp. 107–116, 2024, doi: 10.30595/sainteks.v21i2.23869.