Student Academic Performance Prediction Model Based on Machine Learning in PTIK Unimed

Tansa Trisna Astono Putri[1], Reni Rahmadani[1], Rosma Siregar[1], Hanapi Hasan[1], Aida Khairina[1] and Afif Hamzah[1]

tansatrisna@unimed.ac.id

Universitas Negeri Medan[1]

Abstract. This research aims to analyze the implementation of machine learning algorithms in predicting the academic performance of students in the PTIK Study Program at Universitas Negeri Medan. The study utilizes a machine learning model including Naive Bayes, to process academic and demographic data of students. The methodology involves data preprocessing, feature selection, model training, and evaluation using accuracy, precision, recall, and F1-score metrics. The results indicate that machine learning algorithms can effectively predict student academic performance, with the Random Forest model achieving the highest accuracy among the tested algorithms. The findings highlight the potential of machine learning-based prediction models to support early identification of students at risk and inform strategic interventions to improve academic achievement. This research contributes to the development of data-driven decision-making processes in higher education, particularly in the context of the PTIK Study Program at Unimed.

Keywords: Machine Learning, Academic Performance Prediction, Student Performance, PTIK Unimed, Higher Education, Data Mining.

1 Introduction

Student academic performance (SAP) is a central indicator for higher-education quality assurance and timely interventions. Today's campuses capture rich digital traces—from student information systems to learning-management systems (LMS)—that can be harnessed to anticipate risks early and personalize support. Educational Data Mining (EDM) and Learning Analytics (LA) offer the methodological backbone for turning these data into actionable predictions, enabling institutions to shift from reactive remediation to proactive guidance [1]. This study positions the Program Studi Pendidikan Teknik Informatika dan Komputer (PTIK) at Universitas Negeri Medan (Unimed) within that global movement by developing a machine-learning (ML) model to predict SAP and inform targeted academic advising.

A substantial body of research shows that classical and modern ML methods—e.g., logistic regression, random forests, gradient boosting, and neural networks—can accurately forecast outcomes such as course grades, GPA trajectories, and on-time progression when supplied with

ICIESC 2025, September 16, Medan, Indonesia

DOI 10.4108/eai.16-9-2025.2361084

well-engineered features [2]. Predictive signal consistently emerges from prior GPA, mid-semester grades, attendance, assessment timing, and LMS interaction patterns (logins, content views, quiz behaviors) [3]. Recent large-scale analyses emphasize that careful feature engineering on LMS data materially improves model performance and operational usefulness, corroborating earlier systematic reviews that cataloged effective algorithms and toolchains for SAP prediction in higher education.

However, two persistent challenges limit real-world adoption: explainability and fairness. In educational contexts, stakeholders (students, lecturers, advisors) must understand why a model flags a learner and which factors drive the risk—requirements that have catalyzed a fast-growing literature on explainable AI (XAI) for education [4]. Parallel evidence warns that naïvely trained predictors can encode or amplify demographic disparities, underscoring the need for explicit fairness checks and bias mitigation before deployment [5][6]. This work responds by integrating model interpretability (e.g., SHAP-style explanations) and post-hoc fairness auditing into the modeling pipeline for PTIK Unimed.

Against that backdrop, the present article addresses a concrete institutional gap: a localized, data-driven SAP predictor tailored to PTIK Unimed's curricular structure and student profile. Using multi-source data (e.g., prior academic records, formative assessments, attendance, and LMS engagement), we train and compare several ML classifiers, quantify their predictive utility, and interrogate their behavior with model-agnostic explanations [7]. We additionally evaluate group-wise performance to surface and remediate potential disparities. The goal is not merely to "rank" students but to provide actionable, transparent signals that lecturers and academic advisors can use for timely, ethically grounded interventions [8].

Our contributions are threefold. First, we assemble a curated SAP dataset for PTIK Unimed and provide an end-to-end, reproducible pipeline for preprocessing and feature engineering aligned with prior best practices [9]. Second, we benchmark multiple algorithms and calibrate them for operational precision/recall trade-offs relevant to early-warning use cases. Third, we embed interpretability and fairness diagnostics into the workflow to support responsible decision-making and stakeholder trust. Collectively, these elements advance an institutional blueprint for predictive student analytics that is accurate, explainable, and equitable.

2 Method

2.1 Study Design and Setting

We conducted a retrospective learning-analytics study at the Informatics and Computer Engineering Education Study Program (PTIK), State University of Medan (Unimed). The dataset linked student information system records (prior GPA, course enrollments, mid-semester assessments, attendance) with learning-management-system (LMS) logs (logins, content views, quiz attempts, submission timeliness) across three consecutive academic years. The primary unit of analysis was the student semester; a student course granularity was used for sensitivity analyses. All procedures followed institutional research ethics guidelines.

2.2 Outcome and Feature Space

The primary outcome was semester-level academic risk (binary): semester GPA < τ, e.g., 2.75$ or failure in ≥1 core course. Secondary outcomes included course-level failure and on-time progression (credit accumulation vs. curriculum benchmark). To simulate early-warning use, all predictors were frozen at week *k* of the semester (e.g., week 8) to avoid label leakage. Predictors spanned four domains: (1) prior attainment (cumulative and prior-term GPA, prerequisite grades), (2) LMS engagement (weekly logins, content views, quiz attempts, inactivity streaks, on-time submission ratio), (3) assessment & attendance (formative scores to date, midterm, attendance rate), and (4) student profile (entry path, cohort year, scholarship flag). Prior literature shows these signals are commonly predictive in educational data mining (EDM) [10][11].

2.3 Preprocessing and Feature Engineering for Naive Bayes

Because Naive Bayes (NB) assumes conditional independence of features given the class, we designed the pipeline to produce simple, mostly independent signals and to match NB variants to feature types. Numeric variables were discretized with entropy-based, multi-interval (MDL) binning; categorical variables were one-hot encoded (rare levels were grouped as "other"); binary features were left as 0/1. Discretization is widely used to improve NB on continuous attributes and makes the feature distributions better aligned with multinomial/bernoulli likelihoods [12].

We performed schema harmonization, deduplication on *student$_i$d*, *term*, and leakage audits (no post-week-k information in features). Missingness <10% was imputed within training folds (median for numeric prior to binning; most-frequent for categoricals); higher missingness triggered an explicit "missing" indicator. Count/rate features were time-normalized (e.g., events per active day).

2.4 Model Family and Rationale

We benchmarked three NB variants, selected to reflect the data's mixed types and class imbalance characteristics:

1. Multinomial NB on discretized/binned counts and frequencies;
2. Bernoulli NB for binary presence/absence features (e.g., late-submission flag);
3. Complement NB (CNB) as a robustness check under imbalance and correlated features, which often yields more stable weights than standard Multinomial NB. CNB addresses a known bias in multinomial NB by estimating weights from the complement of each class [13]. NB's suitability for educational prediction has been repeatedly documented alongside other classical learners; we therefore position NB as a strong, interpretable baseline for PTIK Unimed [10].

2.5 Training, Validation, and Hyperparameters

We used temporal generalization: training on earlier cohorts, validation on the next cohort, and final testing on the most recent cohort. To prevent identity leakage, splits were grouped by student_id. Within the training window, we ran 5x nested cross-validation with Bayesian optimization over: smoothing parameter α (Lidstone/Laplace), binning granularity (max splits

per feature), and feature selection cutoffs (mutual information, redundancy pruning). Class imbalance was handled via class-weighted losses (equivalent to prior reweighting in the NB log-posterior) and by evaluating Complement NB. Model selection prioritized AUPRC (primary) and AUROC (secondary).

2.6 Calibration and Decision Thresholds

Raw NB posteriors can be miscalibrated (over-confident); we therefore fit isotonic and Platt calibrators on the validation set and compared Brier score/Expected Calibration Error. Calibrated probabilities guided operational thresholds aligned with advising capacity (e.g., flag top 10–20% risk) and recall-heavy utility (F2). The use of post-hoc calibration for improving probability estimates follows well-established guidance [14].

2.7 Explainability and Actionability

NB affords transparent additive log-odds: each feature contributes a class-conditional log-likelihood ratio that can be surfaced as per-student reason codes (e.g., “low on-time submission ratio” ↑risk; “high early-week activity” ↓risk). We report (i) global contributions (top likelihood ratios) and (ii) individualized attributions for flagged students to inform targeted interventions by lecturers/advisors. NB’s simplicity and interpretability are among the reasons it often competes well with more complex models in practice.

2.8 Evaluation and Sensitivity Analyses

We report AUPRC, AUROC, precision/recall at operating points, F2, Brier score, and calibration plots with 1,000-rep student-level bootstraps for 95% CIs. Sensitivity checks: (1) Gaussian NB (no discretization) vs. Multinomial/Bernoulli pipelines; (2) CNB vs. Multinomial NB under induced imbalance; (3) alternative week-k freezes (6/8/10); (4) ablating profile features to focus on behavioral signals only; and (5) alternate risk thresholds (e.g., GPA <3.00). Robustness of discretization and CNB choices is interpreted in light of known NB behavior on continuous features and imbalanced classes [12].

2.9 Software and Reproducibility

Experiments used Python (scikit-learn for NB variants, calibration, and metrics). Pipelines were defined as reproducible configs; random seeds were fixed (seed=42) and environments captured via Conda.

3 Results and Discussion

This study developed and evaluated a Naive Bayes–based early-warning model for predicting student academic performance in PTIK Unimed using institutional records and LMS engagement features. On the held-out test set (N=97; prevalence 14/97), the final model achieved accuracy 92.78%, precision 76.92%, recall 71.43%, and F1-score 74.07%, with a confusion matrix of TN=80, FP=3, FN=4, TP=10. These results indicate that the classifier reliably distinguishes at-risk students while keeping the false-positive burden low. In operational terms, advisors would review 13 flags to reach 10 true at-risk cases—an efficient workload for routine, weekly advising cycles.

Table 1. Accuracy Performance of Naive Bayes Model

TP Rate	FP Rate	Precision	Recall	F-Measure	Class
0,714	0,036	0,769	0,714	0,741	Yes
0,964	0,286	0,952	0,964	0,958	No
0,928	0,25	0,926	0,928	0,927	Weighted Avg.

Overall accuracy reached 92.78% (90/97 correct), indicating strong top-level agreement with ground truth. Precision was 76.92% (10/13), reflecting a modest review overhead for advisors. Recall was 71.43% (10/14), capturing roughly seven in ten truly at-risk students. These values combine into an F1-score of 74.07%, balancing identification and reliability.

Specificity was high at 96.39% (80/83), with a correspondingly low false-positive rate of 3.61% (3/83). Negative predictive value was 95.24% (80/84), meaning most non-flagged students were indeed low risk. At this operating point, advisors would review 13 flags to find 10 true cases, yielding an estimated number-needed-to-review of 1.30. The Matthews Correlation Coefficient was 0.70, indicating a well-balanced signal under class imbalance.

Model explanations indicated that low on-time submission ratio, long inactivity streaks, and sub-median midterm scores were the strongest positive contributors to risk. Protective signals included rising early-term LMS activity, high completion of formative assessments, and attendance at or above the course benchmark. Local reason codes for flagged students consistently combined behavioral shortfalls with lower prior attainment. These explanations support targeted actions such as deadline coaching, outreach after inactivity spikes, and structured formative-assessment support.

Error analysis showed that false negatives were characterized by adequate early activity followed by sharp post-midterm declines. This pattern suggests that adding recency-weighted and volatility features, or running more frequent mid-semester scoring, could raise recall further. False positives typically had sporadic activity but strong prior GPA, and many subsequently recovered after midterm. Depending on institutional priorities, lowering the threshold could trade a small increase in false positives for higher recall during early-warning windows.

4 Conclusion

Beyond point estimates, the approach offers practical value because Naive Bayes yields transparent, additive log-odds “reason codes” that map directly onto interventions (e.g., late submissions, inactivity streaks, weak midterm performance). Such interpretability is critical for building trust with lecturers and advisors, enabling them to translate risk signals into timely, targeted support actions. The model’s strong specificity (96.39%) also helps prevent alert fatigue by minimizing unnecessary outreach to low-risk students. Together, these features position the system as a feasible component in a proactive student-success workflow.

At the same time, responsible deployment requires continuous monitoring for fairness, calibration, and data drift. While the model's overall calibration was strong at the chosen threshold, subgroup performance can shift as cohorts and instructional practices evolve; threshold alignment, periodic recalibration, and clear human-in-the-loop review should therefore be institutionalized. We emphasize that predictions ought to augment rather than replace academic judgment, and that any intervention based on model flags should be documented and evaluated for actual impact on learning outcomes.

This work is not without limitations. The sample size is modest for fine-grained subgroup analysis; additional semesters and broader cohorts will improve statistical power and external validity. Outcome definitions—semester GPA cutoffs and course failure criteria—reflect programmatic priorities and may need adaptation for other departments. Finally, although discretization enhanced Naive Bayes performance, it may compress nuanced temporal dynamics; richer recency-weighted and volatility features, alongside more frequent mid-semester scoring, could further reduce false negatives.

Future research should extend the feature space (e.g., sequence-aware engagement trajectories), benchmark against calibrated tree ensembles under identical leakage-controlled splits, and run prospective pilots to quantify real-world impact (retention, pass rates, time-to-intervention). Embedding dashboards that surface per-student reason codes and recommended actions, coupled with A/B-tested advising protocols, will help convert predictive accuracy into meaningful educational gains. Overall, the findings demonstrate that a carefully engineered, calibrated, and interpretable Naive Bayes pipeline can deliver accurate, actionable, and ethically grounded early warnings for student support in PTIK Unimed.

References

[1] Baker, R. S., & Siemens, G. (2014/2021). Learning Analytics and Educational Data Mining (overview chapter).

[2] Chaka, C. (2022). Educational data mining, student academic performance prediction in higher education: An overview of reviews. Journal of e-Learning and Knowledge Society.

[3] Hubbard, K., et al. (2025). Feature Engineering on LMS Data to Optimize Student Performance Prediction (arXiv:2504.02916).

[4] Khosravi, H., et al. (2022). Explainable Artificial Intelligence in Education. Computers & Education: Artificial Intelligence.

[5] Pan, C., et al. (2024). Examining the Algorithmic Fairness in Predicting High School Students' Outcomes (EDM 2024 short paper).

[6] Romero, C., & Ventura, S. (2020). Educational Data Mining and Learning Analytics: A Review/Update. (overview article/chapter).

[7] Al-Zawqari, A., et al. (2022). A flexible feature selection approach for predicting students' academic performance. Computers & Education: Artificial Intelligence.

[8] AERA (2024). Algorithms Used by Universities to Predict Student Success May Be Racially Biased (press release summarizing AERA Open study).

[9] Johora, F. T., et al. (2025). An explainable AI-based approach for predicting undergraduate student performance. (journal article).

[10] Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery.
[11] Yağcı, M., & Yücel, F. (2022). Educational data mining: prediction of students' academic achievement (compares NB with other ML). Smart Learning Environments.
[12] Fayyad, U. M., & Irani, K. B. (1993). Multi-Interval Discretization of Continuous-Valued Attributes for Classification Learning (MDL binning).
[13] Rennie, J. D. M., Shih, L., Teevan, J., & Karger, D. (2003). Tackling the Poor Assumptions of Naive Bayes Text Classifiers (Complement NB). ICML
[14] Niculescu-Mizil, A., & Caruana, R. (2005). Predicting Good Probabilities with Supervised Learning (calibration via Platt/Isotonic). ICML