Predicting At-Risk Students from Administrative Data Using Machine Learning: A Multi-Factor Analysis at a Cambodian University

1st Haksrun Lao
Office of Academic Affairs
Paragon International University
Phnom Penh, Cambodia
hlao@paragoniu.edu.kh

2nd Lalin Duch
Content Writer and Transcriber
Aii Language Center
Phnom Penh, Cambodia
duchlalin97@gmail.com

***Abstract*—Early identification of students who fail to meet minimum academic standing allows universities to intervene before academic damage accumulates. This study evaluates whether routinely collected administrative records alone can identify such students at Paragon International University, Phnom Penh, Cambodia. The dataset covers 3,650 students with a recorded cumulative GPA (CGPA) from 15 departments in five faculties across 16 academic years (2010–2026), and includes currently enrolled, graduated, and withdrawn students. A student was labeled at-risk when their CGPA fell below 2.0, the university's minimum passing standard. Eight administrative features (gender, age, department, enrollment year, credits earned, scholarship status, scholarship amount, and grade level) were used to train four classifiers: Logistic Regression, Decision Tree, Random Forest, and Gradient Boosting. Gradient Boosting performed best on the held-out test set ($n = 730$), with 89.18% accuracy, an F1-score of 0.675, and an AUC-ROC of 0.928. Credits earned was the dominant feature (42.46% importance), followed by age (12.88%) and enrollment year (11.97%). Overall, 17.6% of students were at-risk, and scholarship recipients had a higher mean CGPA (2.76) than non-recipients (2.23). Because several features accumulate during enrollment, the models classify current academic standing rather than forecast it at entry; prospective early prediction is identified as the next step.**

***Index Terms*—at-risk students, academic performance prediction, machine learning, administrative data, higher education, Cambodia, Gradient Boosting**

I. Introduction

Academic underachievement affects individual students' progression and, collectively, an institution's graduation rates, accreditation standing, and educational quality [2]. Most universities still identify struggling students reactively: advisers act after a student has already failed courses or withdrawn. Machine learning (ML) applied to institutional student records offers a way to flag these students systematically and earlier [1], [2].

In this study, a student is considered *at-risk* when their cumulative GPA (CGPA) is below 2.0, the minimum passing standard applied by Paragon International University. Such a student is at risk of failing to maintain satisfactory academic standing and, if the situation persists, of not completing the degree. This definition follows the institution's own academic standard rather than an arbitrary cut-off.

This paper develops and evaluates ML classifiers for this target using administrative data from 3,650 students over 16 academic years. Its contributions are: (1) showing that administrative records alone, without learning management system (LMS) logs or surveys, can identify at-risk students with high discrimination; (2) identifying which administrative characteristics are most associated with academic risk; and (3) interpreting departmental differences in the context of program history, to inform academic affairs practice.

A. Research Objectives

1) To describe the distribution of academic performance across departments, grade levels, and demographic groups.
2) To develop and evaluate ML models that classify students' academic risk from administrative data.
3) To identify the most influential features through feature importance analysis.
4) To provide data-driven recommendations for academic affairs administrators.

B. Research Questions

1) What is the prevalence of academic risk across departments, grade levels, and demographic groups?
2) Which ML method achieves the best F1-score and AUC-ROC in classifying academic risk?
3) Which administrative features are most strongly associated with academic risk?
4) How are scholarship status and departmental affiliation associated with CGPA?

II. Literature Review

A. Predicting Academic Achievement in Higher Education

Predicting student performance has long been a central task in educational data mining [1]. A systematic review of studies from 2009 to 2021 found that ML methods are widely used to identify at-risk students and predict dropout, and that most

studies draw on either university databases or online learning platforms [2]. Ensemble methods such as Random Forest and Gradient Boosting frequently outperform single classifiers on such data because they capture non-linear relationships among student attributes [3], [4].

B. Recent Southeast Asian Studies by Discipline Area

Research on this problem in Southeast Asia has grown in recent years. Grouping it by discipline area clarifies where the present study fits.

Science, technology, engineering, and mathematics (STEM). In Thailand, Satmunee *et al.* [5] compared eight ML models on 2,104 science and technology undergraduates with 36 features and found Random Forest best, at 84.32% accuracy and an F1-score of 0.85.

Business and management. In Malaysia, Zaki *et al.* [6] used demographic attributes of pre-diploma students and found that gender, family income, region, and national secondary examination results predicted academic performance. In Cambodia, Pang *et al.* [7] predicted failure in a statistics course for 1,074 business students, finding midterm and continuous-assessment scores to be the strongest predictors.

Multi-faculty institutions. Chan *et al.* [8] predicted student CGPA at a Malaysian university using survey data on background, prior grades, and lifestyle, comparing classical classifiers with neural networks.

Health, humanities, and social sciences. Among the regional studies identified, none focused specifically on health, humanities and arts, or social and behavioral science programs, which limits cross-disciplinary comparison.

C. Predictive Attributes in Student Data

Credits earned is a widely used indicator of academic progress and engagement [9]. Demographic variables such as gender and age show moderate and discipline-dependent predictive value [6], [10]. Financial aid has been associated with better retention and engagement [11]. Most of the studies above rely on course-level assessments [7], survey responses [8], or rich feature sets of 30 or more variables [5].

D. Research Gap

Few studies in the region use *only* the administrative fields that academic affairs offices routinely hold, and fewer still cover every faculty of an institution over a long period. No such study was found for a Cambodian private university spanning 16 academic years. This study addresses that gap.

III. Methodology

A. Research Design

This quantitative study follows the CRISP-DM (Cross-Industry Standard Process for Data Mining) framework: data understanding, data preparation, modeling, evaluation, and interpretation.

B. Institutional Context and Population

All data come from a single institution, Paragon International University in Phnom Penh, Cambodia. The university is organized into five faculties. Table I lists the departments that appear in the analysis.

The records were exported by the Office of Academic Affairs and include every undergraduate student enrolled between academic years 2010–2011 and 2025–2026, regardless of status: currently enrolled students, graduates, and students who withdrew or dropped out before graduating. Including withdrawn students avoids the survivorship bias that would arise from analyzing graduates only.

TABLE I
Faculties and Departments in the Dataset

Faculty	Code	Department
Economics and Administrative Sciences	BAF	Banking and Finance
	BUS	Business Administration
	ECON	Economics
	IR	International Relations
	ITL	International Trade & Logistics
Information and Computer Tech.	CS	Computer Science
	MIS	Mgmt. of Information Systems
	DAD	Digital Arts and Design
Engineering	ARC	Architecture
	AE	Architectural Engineering
	CE	Civil Engineering
	IE	Industrial Engineering
Arts, Humanities and Languages	CM[a]	Communication
	ELT[a]	English Language Teaching
Mathematics and Sciences	MATH[b]	Mathematics

[a]Closed program (CM closed in 2020).
[b]Newly opened; first-year students only.

C. Dataset

The raw export contained 4,167 student records. Records without a recorded CGPA, i.e., students with no graded coursework, were removed, leaving an analytical dataset of 3,650 students. The export listed 17 department codes; 15 of these are represented in the analytical dataset and reported in this paper. Table II summarizes the dataset.

TABLE II
Dataset Characteristics

Characteristic	Value
Institution	Paragon International University
Total students (raw export)	4,167
Analytical dataset size	3,650
Faculties / departments analyzed	5 / 15
Academic years covered	16 (2010–2026)
Student status included	Enrolled, graduated, withdrawn
Gender (raw export)	Male: 2,290 / Female: 1,877
Mean CGPA	2.66 (SD = 0.99)
At-risk students (CGPA < 2.0)	641 (17.6%)
Scholarship recipients (raw export)	3,340 (80.2%)

D. Target Variable

The target is binary: at-risk (1) if CGPA < 2.0, otherwise not at-risk (0). The threshold of 2.0 is the university's minimum passing standard; a CGPA below it means the student is not meeting satisfactory academic standing. This yielded 641 at-risk (17.6%) and 3,009 not-at-risk (82.4%) students, a moderately imbalanced problem.

E. Feature Selection

The administrative export contained eight usable student-level attributes, and all eight were used (Table III). No other variables, such as prior-school grades, entrance scores, attendance, or socioeconomic indicators, were available in the export. The feature set therefore reflects what an academic affairs office can access directly, which is the scenario this study targets.

TABLE III
FEATURE DESCRIPTION AND ENCODING

Feature	Type	Encoding
Gender	Categorical	Label encoding
Age	Numerical	Raw value
Department code	Categorical	Label encoding
Enrollment year	Numerical	Extracted from AY
Credit earned	Numerical	Raw value
Has scholarship	Binary	0/1 flag
Scholarship amount	Numerical	Percentage (0–100)
Grade level	Categorical	Label encoding

F. Data Preprocessing

Preprocessing comprised: (1) removal of records without a CGPA; (2) median imputation of missing Credit Earned values; (3) label encoding of categorical variables; (4) extraction of enrollment year from the academic-year string; and (5) binary encoding of scholarship status. The proportion of missing Credit Earned values was not recorded during preprocessing; this is acknowledged as a limitation in Section V-C. StandardScaler was applied for Logistic Regression only; tree-based models used unscaled features.

G. Model Selection and Training

Four supervised classifiers were compared:

- **Logistic Regression**: interpretable linear baseline; scaled features; max_iter = 1000.
- **Decision Tree**: rule-based, non-linear; max_depth = 10.
- **Random Forest**: bagging ensemble; 100 estimators.
- **Gradient Boosting**: sequential boosting ensemble; 100 estimators.

The data were split into 80% training ($n = 2{,}920$) and 20% testing ($n = 730$) with stratification on the target. Five-fold cross-validation was applied on the training set to assess generalizability. All experiments used `random_state=42`.

H. Evaluation Metrics

Models were evaluated with accuracy, precision, recall, F1-score, and AUC-ROC. Because of the class imbalance, F1-score and AUC-ROC were the primary criteria.

IV. RESULTS

A. Descriptive Analysis

Fig. 1 shows the distribution of performance categories and of CGPA. Of the 3,650 students, 1,544 (42.3%) are Passing, 987 (27.0%) Good, and 466 (12.8%) Excellent. The At-Risk bar contains 653 students (17.9%) because the category bins in Fig. 1 include the boundary value: 12 students with a CGPA of exactly 2.0 are counted there. Under the model target (CGPA strictly below 2.0), these 12 students are not at-risk, giving the 641 (17.6%) used throughout the modeling.

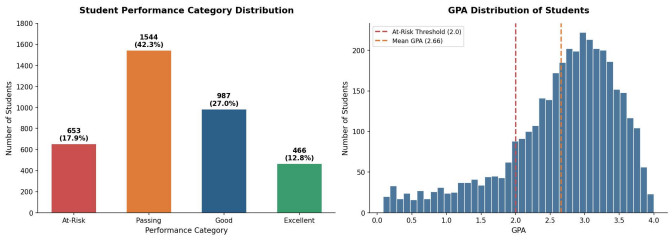


Fig. 1. Performance categories (left) and CGPA distribution (right). Category bins include the upper boundary, so the At-Risk bar contains 12 students with CGPA = 2.0. The red dashed line marks the at-risk threshold.

Female students have a higher mean CGPA than male students (2.77 vs. 2.57). Scholarship recipients have a higher mean CGPA than non-recipients (2.76 vs. 2.23).

Fig. 2 shows mean CGPA by department. Mathematics (MATH) has the highest mean (3.15), while Architectural Engineering (AE, approx. 2.12), English Language Teaching (ELT, 2.24), and Communication (CM, approx. 2.36) have the lowest. These extremes must be read in light of program history (Table I). MATH opened recently and contains only first-year students, whose grades come from introductory courses and who have not yet faced upper-level coursework. CM and ELT are closed programs: their records consist of earlier cohorts, and students who remained after closure may have been those progressing slowly. Departmental means for new or closed programs therefore reflect cohort composition at least as much as curriculum or teaching.

B. Model Performance

Table IV reports test-set performance for the four models. Figs. 3 and 4 show the same results graphically.

TABLE IV
MODEL PERFORMANCE ON THE TEST SET ($n = 730$)

Model	Acc.	Prec.	Rec.	F1	AUC
Log. Reg.	87.53%	0.667	0.578	0.619	0.909
Dec. Tree	87.53%	0.646	**0.641**	0.643	0.839
Rand. Forest	88.08%	0.681	0.602	0.639	0.916
Grad. Boost.	**89.18%**	**0.713**	**0.641**	**0.675**	**0.928**

Best value per column in bold. Recall is tied between Dec. Tree and Grad. Boost.

Gradient Boosting performed best on every metric, tying with the Decision Tree on recall (0.641). It achieved 89.18%

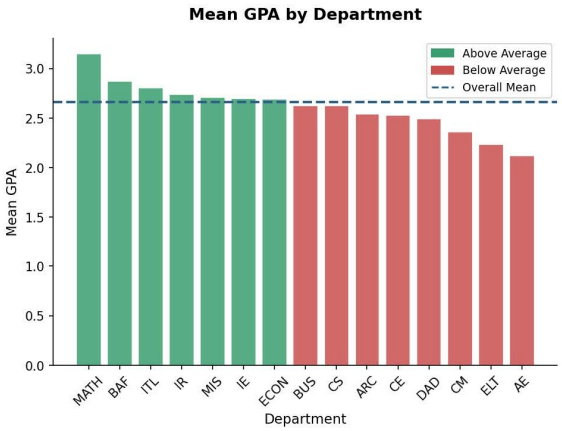


Fig. 2. Mean CGPA by department. Green bars are above the overall mean (dashed line); red bars are below it.

accuracy, an F1-score of 0.675, and an AUC-ROC of 0.928. Random Forest ranked second by AUC-ROC (0.916), followed closely by Logistic Regression (0.909). The Decision Tree had the lowest AUC-ROC (0.839), indicating less reliable ranking across thresholds.

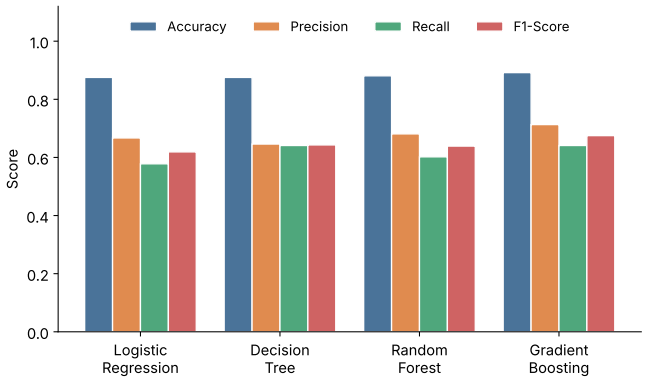


Fig. 3. Accuracy, precision, recall, and F1-score by model.

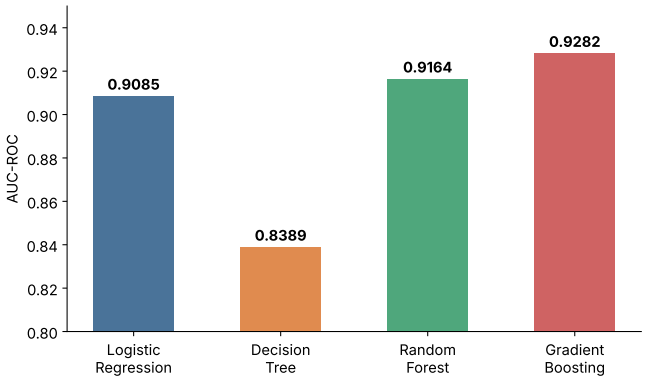


Fig. 4. AUC-ROC by model.

C. Feature Importance

Fig. 5 shows Random Forest feature importances (mean decrease in impurity). Credit Earned dominates at 42.46%. Age (12.88%) and Enrollment Year (11.97%) follow, then Grade Level (10.50%) and Department (10.38%). Scholarship Amount (6.90%), Has Scholarship (2.59%), and Gender (2.32%) contribute least once the other variables are included.

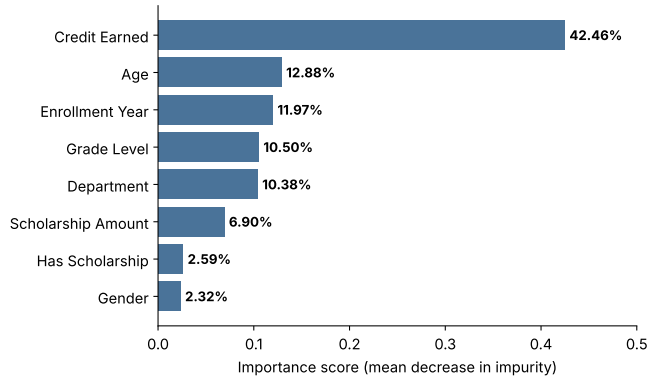


Fig. 5. Random Forest feature importance (percentages sum to 100%).

D. ROC Curve and Confusion Matrix

Fig. 6 shows the ROC curve and confusion matrix for Gradient Boosting. Of the 128 at-risk students in the test set, 82 were correctly identified (recall 64.1%) and 46 were missed (35.9%). Of the 602 not-at-risk students, 569 were correctly classified (94.5%) and 33 were falsely flagged (5.5%), giving a precision of 0.713.

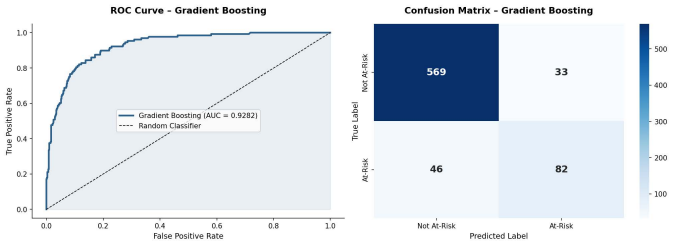


Fig. 6. ROC curve (left) and confusion matrix (right) for Gradient Boosting (AUC-ROC = 0.928).

V. Discussion

A. Interpretation

Administrative records alone, without LMS logs, behavioral tracking, or surveys, separated at-risk from not-at-risk students well (AUC-ROC 0.928). This matters for institutions where only a basic student information system is available. Performance is comparable to regional studies that used much richer inputs, such as the 84.32% accuracy obtained with 36 features in a Thai STEM cohort [5].

The dominance of Credit Earned is expected but needs careful interpretation. Credits accumulate as students pass courses,

so a low count partly reflects the same failures that lower CGPA. Age, grade level, and enrollment year likewise change during enrollment. The models therefore classify a student's *current* academic standing from administrative data, rather than forecasting it at admission. Their practical value lies in screening the whole student body each semester to prioritize advising, not in predicting outcomes for new entrants.

In an early-warning setting, a missed at-risk student (false negative) costs more than an unnecessary advising contact (false positive). At the default threshold, 46 of 128 at-risk test students were missed. Lowering the decision threshold would raise recall at some cost in precision, and the right operating point depends on the advising capacity available.

The scholarship gap (2.76 vs. 2.23) is consistent with prior findings on financial aid [11]. However, scholarships at many institutions depend partly on prior or continuing academic performance, so the association may reflect selection rather than a causal effect.

B. Comparison Across Discipline Areas

Mapping the departments to broad areas, Business and Management departments (BAF, ITL, IR, ECON) sit at or above the institutional mean, apart from BUS. STEM departments are mixed: MATH is highest, IE and MIS are near the mean, and CS, ARC, CE, and AE are below it. Humanities and language programs (CM, ELT) are among the lowest, though both are closed programs. Regional studies cover STEM [5] and business [6], [7] cohorts separately; this study adds a single-institution, cross-faculty view. Because the dataset contains no health or medicine programs, comparison with that area is not possible.

C. Limitations

The study has several limitations. Data come from one private university, which limits generalizability. The proportion of missing Credit Earned values was not recorded before median imputation. Two of the 17 department codes are not represented in the analytical dataset. Some features change during enrollment, so the models are not prospective predictors. Only classical ML methods were evaluated, and no real-world validation of predictions against later outcomes has yet been carried out.

VI. Conclusion

This paper showed that a Gradient Boosting model trained on eight administrative fields can identify students below the university's minimum passing standard with 89.18% accuracy, an F1-score of 0.675, and an AUC-ROC of 0.928, using data from 3,650 students at Paragon International University over 16 years.

Key findings are: (1) 17.6% of students had a CGPA below 2.0; (2) Credit Earned was the most important feature (42.46%); (3) scholarship recipients had a mean CGPA 0.53 points higher than non-recipients; (4) female students had a mean CGPA 0.20 points higher than male students; and (5) departmental extremes were largely explained by program history, namely new (MATH) and closed (CM, ELT) programs.

Future work will: (1) build a prospective early-warning model using only features available at entry or in the first semester; (2) validate it in practice by flagging an incoming cohort and tracking its outcomes; (3) extend the dataset with course-level and attendance records and with data from other institutions; and (4) compare neural network models against the ensemble methods used here.

References

[1] C. Romero and S. Ventura, "Educational data mining: A review of the state of the art," *IEEE Trans. Syst., Man, Cybern. C*, vol. 40, no. 6, pp. 601–618, 2010. doi:10.1109/TSMCC.2010.2053532

[2] B. Albreiki, N. Zaki, and H. Alashwal, "A systematic literature review of student' performance prediction using machine learning techniques," *Educ. Sci.*, vol. 11, no. 9, p. 552, 2021. doi:10.3390/educsci11090552

[3] N. Mduma, K. Kalegele, and D. Machuve, "A survey of machine learning approaches and techniques for student dropout prediction," *Electron. J. Inf. Syst. Dev. Countries*, vol. 85, no. 2, p. e12079, 2019. doi:10.1002/isd2.12079

[4] D. Delen, "A comparative analysis of machine learning techniques for student retention management," *Decis. Support Syst.*, vol. 49, no. 4, pp. 498–506, 2010. doi:10.1016/j.dss.2010.06.003

[5] J. Satmunee, E. Naowanich, and T. Patikorn, "Analysis and prediction of undergraduate student dropouts using machine learning," *J. Sci. Technol. Mahasarakham Univ.*, vol. 45, no. 2, pp. 231–242, 2026. doi:10.14456/jscitech.2026.20

[6] S. M. Zaki, S. Razali, M. Z. Laton, and M. Ishak, "Predicting students' performance at higher education institutions using a machine learning approach," *Kybernetes*, 2024. doi:10.1108/K-12-2023-2742

[7] C. Pang, P. Hok, B. Khem, S. To, and S. Hoeung, "Education data mining: Students' performance in statistic class using machine learning," *AUB Res. Ser.*, vol. 4, pp. 110–125, 2025. doi:10.71215/ars.202545

[8] J. Y. Chan, H. Ng, T. T. V. Yap, and V. T. Goh, "Predictive modelling of student performance in MMU based on machine learning approach," in *Proc. Int. Conf. Comput., Inf. Technol. Intell. Comput. (CITIC 2022)*, Atlantis Press, 2022. doi:10.2991/978-94-6463-094-7_21

[9] S. M. Jayaprakash, E. W. Moody, E. J. M. Lauria, J. R. Regan, and J. D. Baron, "Early alert of academically at-risk students: An open source analytics initiative," *J. Learn. Anal.*, vol. 1, no. 1, pp. 6–47, 2014. doi:10.18608/jla.2014.11.3

[10] Z. J. Kovacic, "Early prediction of student success: Mining students enrollment data," in *Proc. Inf. Sci. Inf. Technol. Educ. Conf.*, 2010, pp. 647–665. doi:10.28945/1281

[11] L. D. Singell, "Come and stay a while: Does financial aid affect retention conditioned on enrollment at a large public university?," *Econ. Educ. Rev.*, vol. 23, no. 5, pp. 459–471, 2004. doi:10.1016/j.econedurev.2003.10.006