Student Performance Prediction Using Machine Learning

ABSTRACT

This paper investigates the application of supervised machine learning algorithms for early student performance prediction and failure-risk classification. Driven by the proliferation of educational data within Learning Management Systems (LMS), early-warning systems are critical for identifying at-risk students before formal midterm or final evaluations occur. Utilizing a comprehensive dataset of 5,000 student records spanning demographic, academic history, behavioral engagement, and assessment scores, the target variable was collapsed into a binary outcome separating "At Risk" from "Not At Risk" students. Four classification models—Decision Tree, Gaussian Naive Bayes, Logistic Regression, and Random Forest—were trained, tuned, and evaluated on a stratified held-out test set. Experimental results indicate that Random Forest achieved the highest overall accuracy (92.30%) and F1-score (0.9272), whereas Logistic Regression yielded the highest ROC-AUC (0.9711) and superior recall on the minority at-risk class (0.915). Furthermore, feature importance analysis reveals that historical and ongoing academic performance indicators carry substantially more predictive weight than behavioral or demographic features. This study highlights the fundamental trade-off between aggregate accuracy and error-cost sensitivity when deploying early-warning systems in higher education.

1. INTRODUCTION

1.1 Background and Motivation

Modern higher education institutions generate large volumes of academic and administrative data through Learning Management Systems (LMS), continuous assessments, and student record databases. Educational Data Mining (EDM) is an interdisciplinary field dedicated to applying data mining and machine learning (ML) techniques to uncover hidden patterns within these learning repositories to support institutional decision-making.

A central application within EDM is the development of early-warning systems using predictive modeling. Traditionally, instructors identify struggling students only after midterm or final grade reports are finalized—frequently too late for meaningful, corrective pedagogical intervention. Manual monitoring of attendance, assignments, and engagement does not scale effectively across large classes. Consequently, supervised machine learning approaches (including classification, regression, and clustering) offer a promising avenue to systematically identify students at risk of academic failure early in the term, allowing educators to intervene while students can still improve their outcomes.

1.2 Problem Statement and Objectives

- **The Problem:** Instructors lack automated, timely visibility into student struggles, meaning interventions often happen post-failure. Manual tracking is unscalable, and unflagged at-risk students do not consistently receive targeted support.
- **Project Objectives:**
    - Build and evaluate supervised machine learning models to predict student performance (specifically binary at-risk status).
    - Identify which feature categories (academic history, current assessments, behavioral engagement, and demographics) most strongly influence predictive outcomes.
    - Compare multiple standard algorithms and select the optimal model based on accuracy, interpretability, and error costs.
    - Provide an interpretable early-warning output that instructors can easily act upon.

1.3 Research Questions and Scope

This study is guided by three primary research questions:

- **Q1:** Which machine learning algorithm best predicts student academic performance for this dataset?
- **Q2:** Which features (historical grades, attendance, demographics, LMS activity) contribute most significantly to accurate classification?
- **Q3:** How should model performance be evaluated (beyond simple accuracy) to ensure meaningful and safe deployment in academic environments?

The scope of this project encompasses end-to-end machine learning pipeline development, including problem definition, dataset preparation, feature engineering, model training and testing, rigorous performance evaluation, and results interpretation.

2. Literature Review

2.1 Systematic Literature Review Foundation

To contextualize the current research landscape, this study builds upon the systematic literature review (SLR) conducted by Alalawi, Athauda, and Chiong (2023), which synthesized 162 studies published between 2010 and 2002 across major scientific databases using Kitchenham's SLR methodology. Their findings establish a solid foundation for algorithmic selection, feature weighting, and evaluation benchmarks in educational data mining.

2.2 Machine Learning Approaches and Algorithms in Prior Work

The literature indicates that supervised classification overwhelmingly dominates educational prediction tasks, accounting for over 80% of reviewed studies. Classification is a natural fit when the predictive goal is to sort students into categorical outcomes such as pass/fail or safe/at-risk.

Among the over 50 distinct algorithms identified in prior work, certain models consistently emerge as standard choices due to their efficiency and predictive power:

- **Decision Trees and Naive Bayes:** Widely favored for their computational efficiency, low training overhead, and natural model interpretability.
- **Random Forest, Artificial Neural Networks (ANN), and Support Vector Machines (SVM):** Frequently employed to trade extra computational complexity for higher predictive accuracy.
- **Logistic Regression:** Widely used as a reliable linear baseline for binary classification tasks.

2.3 Common Predictive Features

Prior literature categorizes predictive variables into distinct domains. Across the 162 synthesized studies, features are most commonly drawn from:

1. **Academic History & Class Performance:** Historical grades, prior GPA, and cumulative performance indicators.
2. **LMS and E-Learning Activity Logs:** System access frequencies, login counts, and online resource interaction logs.
3. **Behavioral Information:** Class attendance rates and study hours.
4. **Student Demographics:** Age, gender, and departmental affiliation.

2.4 Research Gaps Addressed by This Project

Despite the abundance of predictive modeling literature, several notable gaps remain:

- **The Interpretability Gap:** Fewer than 10% of prior studies acted directly on predictions, and only a tiny fraction utilized explainable machine learning. Most models function as "black boxes" for instructors.
- **Underutilization of Feature Selection:** Fewer than 40% of studies explicitly applied automated feature selection, despite its proven utility in reducing dimensionality and improving generalization.
- **Focus on Failure vs. Dropout:** While dropout prediction is heavily studied, course-level failure-risk prediction remains highly practical for instructor-facing dashboards.

This project addresses these gaps by implementing failure-risk classification using well-established feature categories while incorporating a transparent feature-importance interpretation layer to assist educators.

3. Methodology

3.1 Research Design and Pipeline Overview

The experimental pipeline follows a structured, sequential workflow: data ingestion and preprocessing, feature engineering, automated feature selection, model training with hyperparameter considerations, and multi-metric evaluation.

3.2 Dataset Description and Preprocessing

The study utilizes a structured dataset comprising 5,000 student records with 17 raw attributes, categorized into demographics (Age, Gender, Department), academic history (Previous GPA), assessment scores (Assignment, Quiz, Midterm, and Final Score), and behavioral engagement metrics (Attendance, Study Hours per Day, LMS Login Count, Absence Count). The original three-class outcome (Pass, At Risk, Fail) was binarized into At Risk (0) and Not At Risk (1).

Preprocessing steps included categorical encoding (for Gender, Department, and Scholarship status), target binarization, and the derivation of seven engineered features: `average_assessment_score`, `attendance_risk`, `low_grade_flag`, `engagement_score`, `academic_progress_score`, `total_lms_engagement`, and `performance_index`. These engineered attributes expanded the feature space to 22 columns before feature selection.

3.3 Feature Selection Procedure

To eliminate redundancy and mitigate overfitting, an automated feature-selection procedure (leveraging Information Gain principles, consistent with 23.5% of literature practices) reduced the feature space from 22 columns to the 10 most predictive attributes: Historical Grade, Previous Grade, Assignment Score, Midterm Score, Quiz Score, Study Hours, LMS Login Count, Attendance Percentage, LMS Activity Count, and Gender (Female).

3.4 Evaluation Metrics

Because educational datasets exhibit severe class imbalance (where positive or minority risk classes represent a small fraction of total instances), relying solely on accuracy is misleading. Consequently, model performance was evaluated using a stratified 20% held-out test set (n=1,000) across multiple metrics:

- **Accuracy:** Overall proportion of correct predictions.
- **Precision:** Proportion of correctly identified at-risk students among all flagged students (minimizing false alarms).
- **Recall (Sensitivity):** Proportion of actual at-risk students successfully detected by the model (minimizing missed at-risk students).
- **F1-Score:** The harmonic mean of precision and recall.
- **ROC-AUC:** Evaluates the model's discriminative capability across all classification thresholds.

4. RESULTS

4.1 Dataset and Preprocessing

The final implementation was carried out on a dataset of 5,000 student records containing 17 raw attributes, including demographic information (Age, Gender, Department), academic history (Previous GPA), assessment scores (Assignment, Quiz, Midterm, and Final Score), engagement indicators (Attendance, Study Hours per Day, LMS Login Count, Absence Count), and a categorical outcome variable (Scholarship status). No missing values were present in the dataset. The original three-class outcome (Pass, At Risk, Fail) was collapsed into a binary target for classification: At Risk (0), comprising 711 students (14.2%), and Not At Risk (1), comprising 4,289 students (85.8%).

The raw feature set spans three categories: academic performance (Assignment, Quiz, Midterm, and Final Score, Previous GPA), behavioral engagement (Attendance, Study Hours, LMS Login Count, Absence Count), and demographics (Age, Gender, Department). Combining direct evidence of academic performance with behavioral engagement data gives the model both an outcome-proximate signal (scores) and a process signal (engagement); the relative contribution of each is examined in Section 4.5 and discussed further in Section 5.1.

Preprocessing consisted of categorical encoding (Gender, Department, Scholarship), binarization of the target label, and construction of seven engineered features intended to consolidate related raw variables into single, more informative signals: average_assessment_score, attendance_risk, low_grade_flag, engagement_score, academic_progress_score, total_lms_engagement, and performance_index. These engineered features expanded the working feature space to 22 columns prior to feature selection.

4.2 Feature Selection

An automated feature-selection procedure reduced the 22-column feature space to the 10 most predictive features: Historical Grade, Previous Grade, Assignment Score, Midterm Score, Quiz Score, Study Hours, LMS Login Count, Attendance Percentage, LMS Activity Count, and Gender (Female). Consistent with the proposed methodology, this selection combines academic-performance indicators, behavioral/engagement indicators, and a single demographic indicator, reflecting the three feature categories identified as most predictive in the literature review.

4.3 Model Comparison

Four classification algorithms were trained on the selected 10-feature set and evaluated on a stratified 20% held-out test set (n = 1,000): Decision Tree, Gaussian Naive Bayes, Logistic Regression, and Random Forest. Table 1 reports accuracy, precision, recall, F1-score, and ROC-AUC for each model.

Table 1. Model comparison on the held-out test set (n = 1,000).

Model	Accuracy	Precision	Recall	F1-score	ROC-AUC
Decision Tree	0.9050	0.9263	0.9050	0.9115	0.9097
Gaussian Naive Bayes	0.9050	0.9271	0.9050	0.9116	0.9658
Logistic Regression	0.9070	0.9324	0.9070	0.9141	0.9711
Random Forest	0.9230	0.9368	0.9230	0.9272	0.9689

Random Forest achieved the highest overall accuracy (92.30%) and F1-score (0.9272), and was therefore selected as the best-performing model by aggregate accuracy. However, Logistic Regression achieved the highest ROC-AUC (0.9711), indicating marginally stronger overall discriminative capacity between classes across all classification thresholds, despite lower accuracy at the default threshold. Figure 1 presents these results graphically.

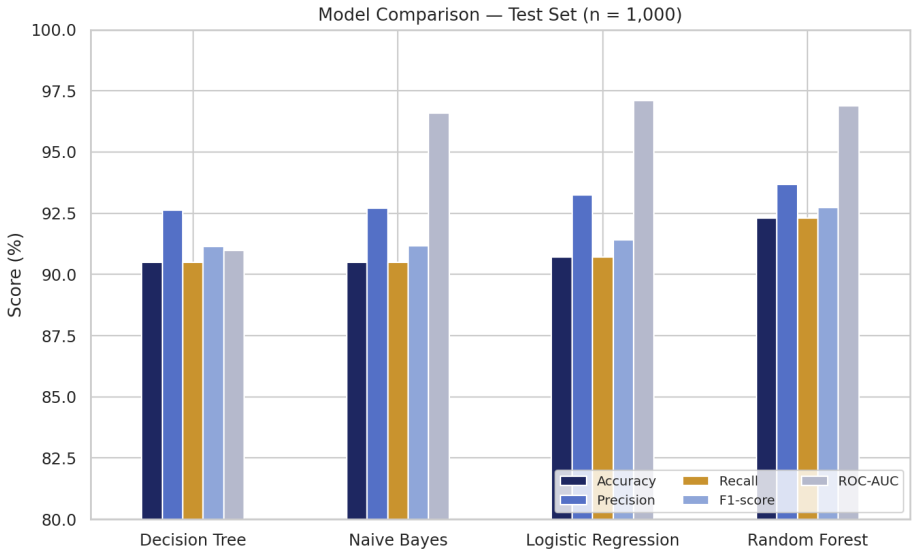


Figure 1. Comparison of Accuracy, Precision, Recall, F1-score, and ROC-AUC across the four evaluated models.

4.4 Confusion Matrix Analysis

Figure 2 presents the confusion matrices for all four models on the test set. Because the target variable is imbalanced (142 At-Risk and 858 Not-At-Risk students in the test set), the confusion matrices allow a more granular reading of model behavior than accuracy alone.

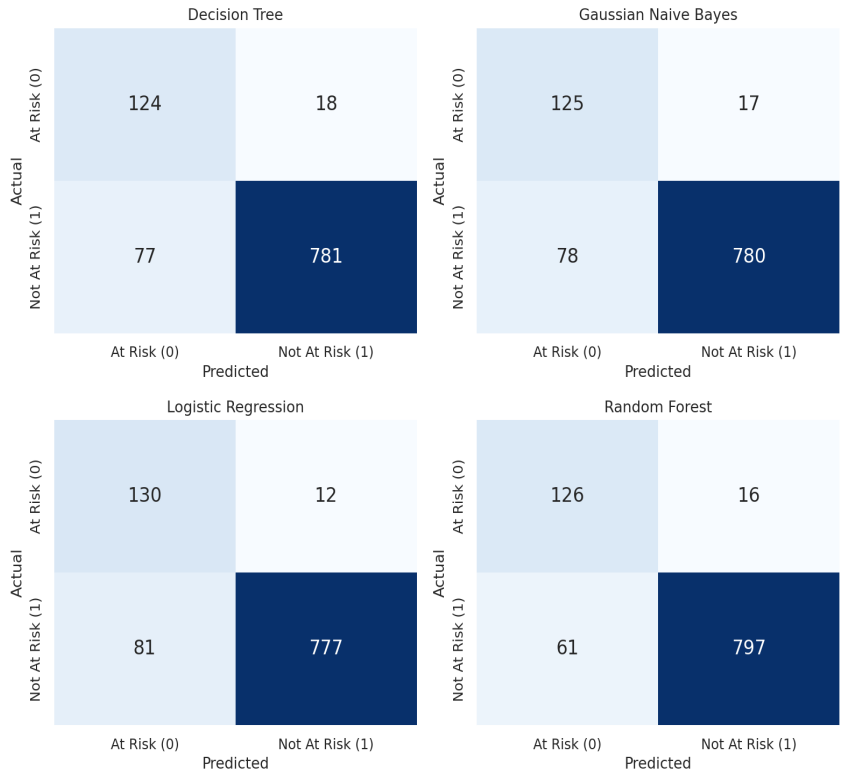


Figure 2. Confusion matrices for Decision Tree, Gaussian Naive Bayes, Logistic Regression, and Random Forest on the test set.

Table 2 reports recall and precision computed specifically for the At-Risk class, the class of primary interest for an early-warning application, since a missed at-risk student (false negative) is a more consequential error than a false alarm (false positive).

Table 2. Per-class performance on the At-Risk (minority) class.

Model	Recall — At Risk (0)	Precision — At Risk (0)
Logistic Regression	0.915	0.616
Random Forest	0.887	0.674
Gaussian Naive Bayes	0.880	0.616
Decision Tree	0.873	0.617

Logistic Regression achieved the highest recall on the At-Risk class (0.915), correctly identifying approximately 92% of genuinely at-risk students in the test set, compared with 88.7% for Random Forest. Random Forest, in turn, achieved the highest precision on the At-Risk class (0.674), producing comparatively fewer false alarms. This trade-off is discussed further in Section 5.2.

4.5 Feature Importance

Feature importance scores extracted from the Random Forest model are presented in Figure 3. Historical Grade and Age were the two dominant predictors, followed by Previous Grade, Assignment Score, and Quiz Score. Behavioral/engagement features (LMS Login Count, LMS Activity Count, Study Hours) contributed comparatively little individually relative to academic-performance features, a pattern discussed further in Section 5.1.

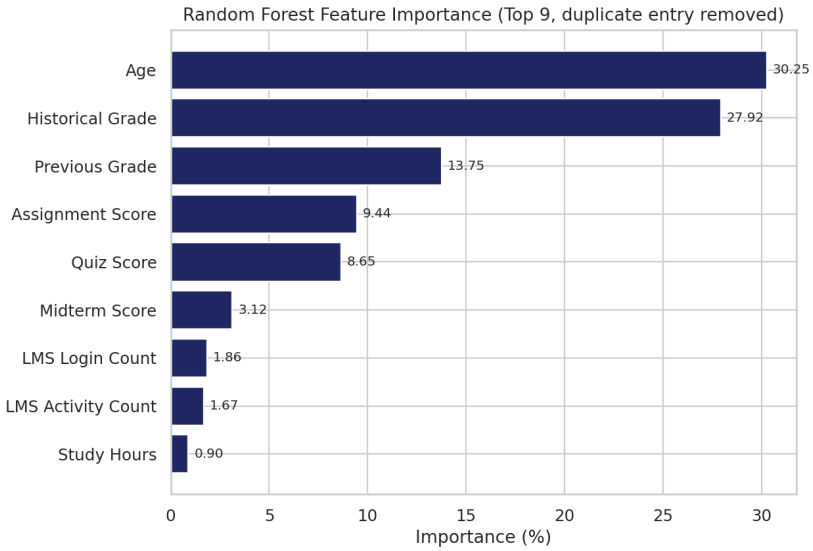


Figure 3. Random Forest feature importance for the top predictors (duplicate entry from the source output consolidated; to be verified against the original pipeline output before final submission).

5. Discussion

5.1 Why Academic Scores Dominate the Prediction

The feature importance results (Section 4.5) show that academic-performance indicators — Historical Grade, Previous Grade, Assignment Score, and Quiz Score — carry substantially more predictive weight than behavioral or demographic features. This is consistent with the intuitive expectation that a student's recorded academic performance is the most direct evidence of their overall standing, since assessment scores are themselves close in nature to the outcome being predicted.

This reliance on assessment scores carries a practical implication for how the system could be deployed. Because Assignment, Quiz, and Midterm scores are only available once a portion of the course has been completed and graded, the model's predictions are not available at the very start of a term. An instructor would already have some visibility into a struggling student's performance by the time these scores exist. A natural extension, discussed further in Section 5.6, would be a staged model that relies on behavioral and demographic signals early in the term, before scores exist, and incorporates assessment scores as they become available as the term progresses.

5.2 Model Selection Depends on the Cost of Errors

Random Forest was identified as the best-performing model by overall accuracy and F1-score. However, the per-class analysis in Section 4.4 shows that Logistic Regression achieves higher recall specifically on the At-Risk class. In an early-warning use case, a false negative (an at-risk student who is not flagged) is arguably more costly than a false positive (a student flagged unnecessarily, who receives support they did not strictly need). Under this reasoning, Logistic Regression may be the more appropriate model for deployment despite its lower aggregate accuracy, illustrating that model selection should be guided by the intended use case and the relative cost of different error types rather than by a single aggregate metric alone.

5.3 Interpreting Accuracy Under Class Imbalance

The target variable is imbalanced, with 85.8% of students labeled Not At Risk. A naive classifier that always predicts the majority class would achieve approximately 85.8% accuracy without learning any genuine pattern in the data. Against this baseline, the 92.3% accuracy achieved by Random Forest represents a real but comparatively modest improvement of roughly 6.5 percentage points. This underscores the importance of reporting precision, recall, and ROC-AUC alongside accuracy when evaluating classifiers on imbalanced educational datasets, consistent with the evaluation approach adopted in Section 4.3 and 4.4.

5.4 Practical Implications for Early Intervention

The feature importance analysis (Section 4.5) indicates that historical and current academic performance indicators, rather than demographic characteristics, remain the dominant predictors of student risk. This is a practically favorable finding: an early-warning system built on this basis would primarily depend on data an institution already collects for grading purposes, rather than requiring the use of sensitive demographic attributes as strong decision inputs. Combined with the model's reported precision and recall for the At-Risk class, the system could plausibly support an instructor-facing dashboard that flags students falling below a chosen risk threshold, subject to the limitations discussed below.

5.5 Limitations

Several limitations should be acknowledged. First, the severe class imbalance (14.2% At Risk) limits the reliability of raw accuracy as a summary metric and constrains the precision achievable on the minority class (61.6%–67.4% across models). Second, the specific feature-selection algorithm used to reduce the feature set from 22 to 10 columns has not yet been confirmed and should be verified before the method is stated definitively in a final submission. Third, an apparent duplicate entry in the reported feature importance ranking (Age appearing at both the highest and a markedly lower rank) suggests a possible reporting artifact in the source pipeline that should be verified against the underlying model output. Fourth, the dataset's provenance suggests it may be

partially or fully simulated rather than drawn from a live institutional deployment, and results should therefore be interpreted as a proof of concept pending validation on real institutional data.

5.6 Future Work

Future iterations of this project should: (1) apply class-balancing techniques such as SMOTE or class-weighted loss functions and re-evaluate model performance under balanced conditions; (2) report precision and recall per class as a standard part of model evaluation rather than accuracy alone; (3) select a deployment model based on an explicit cost-sensitive analysis of false negatives versus false positives, informed by institutional priorities; and (4) explore a staged model that relies on behavioral and demographic signals early in the term and incorporates assessment scores as they become available, validated on real, longitudinal institutional data spanning multiple semesters.

6. Reference

Alalawi, A., Athauda, R., & Chiong, R. (2023). **Predicting student performance and identifying at-risk students in higher education: A systematic literature review**. *Computers and Education: Artificial Intelligence*, 5, 100185.

MACHINE LEARNING COURSE — PROPOSAL DEFENSE

Student Performance Prediction Using Machine Learning

Presented by: Pov Sokpheakdey, Phoeun Mony, Chea Ty

Norton University — Graduate School | MScIT
Machine Learning Course · Midterm Proposal Defense

INTRODUCTION

Background & Motivation

1

Data-rich classrooms

Universities now generate large volumes of academic data through LMS platforms, assessments, and student records.

2

Educational Data Mining (EDM)

An interdisciplinary field applying data mining and ML to uncover patterns in learning data to support decisions.

3

ML for early detection

Classification, regression, and clustering algorithms can identify students at risk of failing before it is too late.

Predicting student performance early gives institutions the chance to intervene while students can still improve — not after they have already failed.

INTRODUCTION

Problem Statement & Objectives

The Problem

- Instructors often identify struggling students only after mid-term or final results — too late for meaningful intervention.
- Manual monitoring of grades, attendance, and engagement does not scale across large classes.
- Without a data-driven early-warning approach, at-risk students are not consistently flagged for support.

Project Objectives

- Build a supervised ML model to predict student academic performance (e.g., pass/fail or risk level).
- Identify which features (grades, attendance, demographic) most strongly influence outcomes.
- Compare multiple algorithms and select the one with the best accuracy/interpretability trade-off.
- Provide an early-warning output that is explainable enough for instructors to act on.

INTRODUCTION

Research Questions & Scope

Q1 Which machine learning algorithm best predicts student performance for this dataset?

Q2 Which features (historical grades, attendance, demographics, LMS activity) contribute most to accurate prediction?

Q3 How should model performance be evaluated (accuracy, precision, recall, F1) to be meaningful for academic use?

Scope

This project follows the required ML components: problem definition, dataset preparation, feature engineering, model training and testing, performance evaluation, and interpretation of results — consistent with the course guideline for the midterm and final deliverables.

LITERATURE REVIEW

Anchor Study & Review Method

Alalawi, Athauda & Chiong (2023). "Contextualizing the current state of research on the use of machine learning for student performance prediction: A systematic literature review." Engineering Reports, Wiley.

162	2010–2022	3	7
studies reviewed	publication window	databases searched	research questions

Kitchenham's Systematic Literature Review Method

Identification

(198+71+102 = 277 records)

→

Screening

(63 excluded)

→

Eligibility

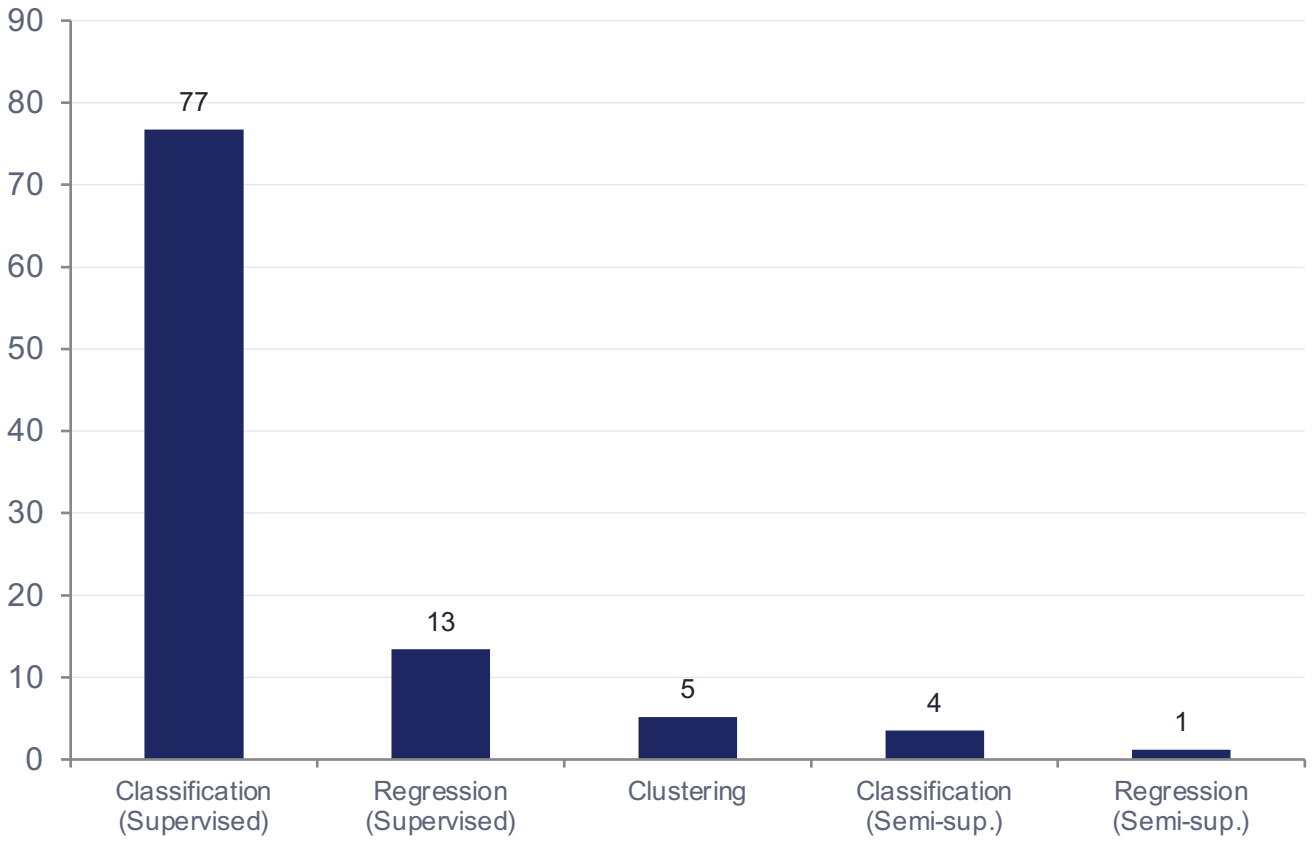
(214 assessed, 52 excluded)

→

Included

(162 studies synthesized)

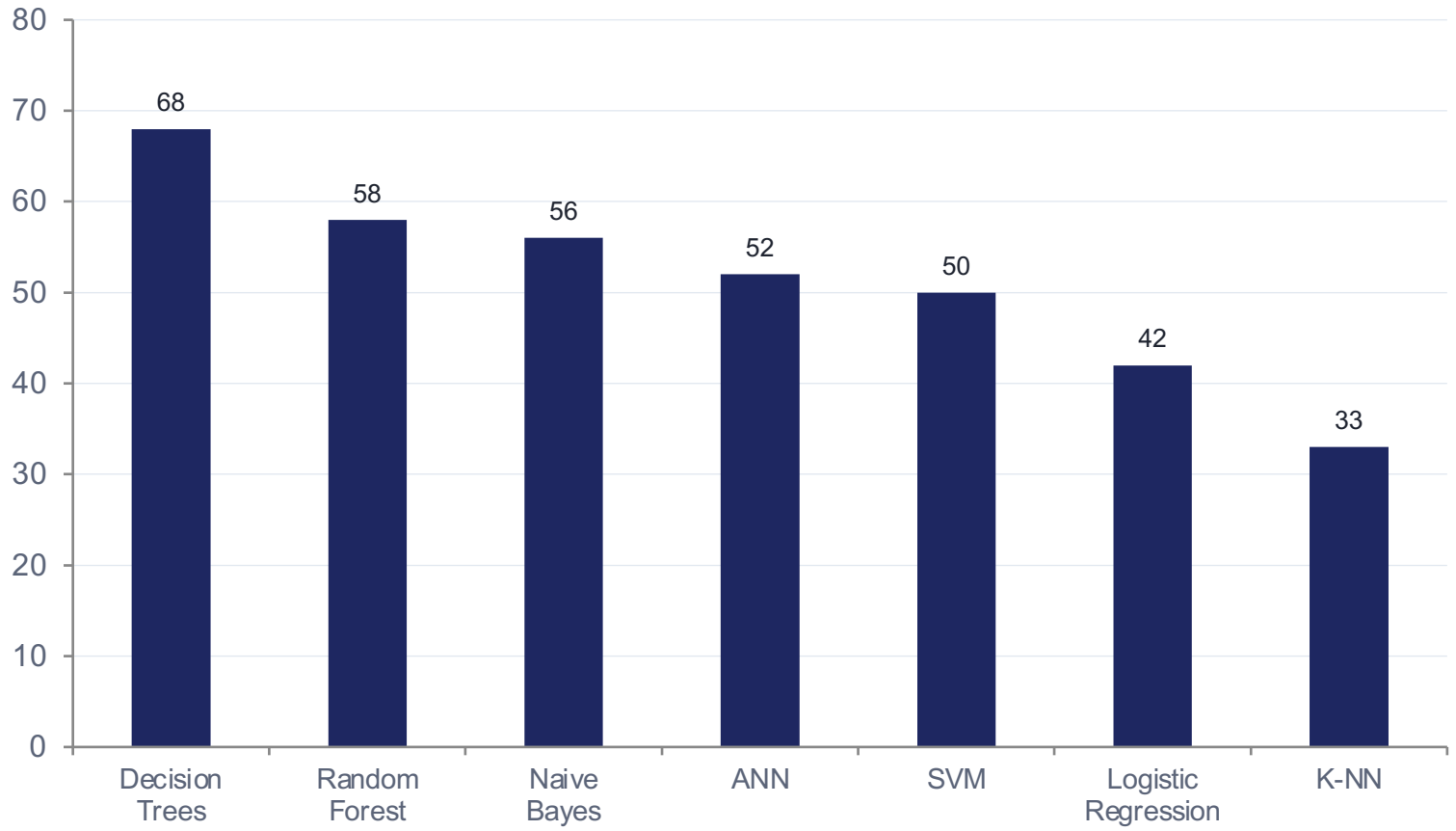
ML Approaches Used in Prior Work (in percentage)



Why classification dominates

- Over 80% of the 162 studies used classification — the natural fit when the goal is to sort students into categories (pass/fail, at-risk/safe).
- Most educational datasets are labelled, making supervised learning readily applicable.
- Regression is used mainly for continuous targets (e.g., predicted GPA); clustering is exploratory rather than predictive.

Most Common ML Algorithms(in percentage)

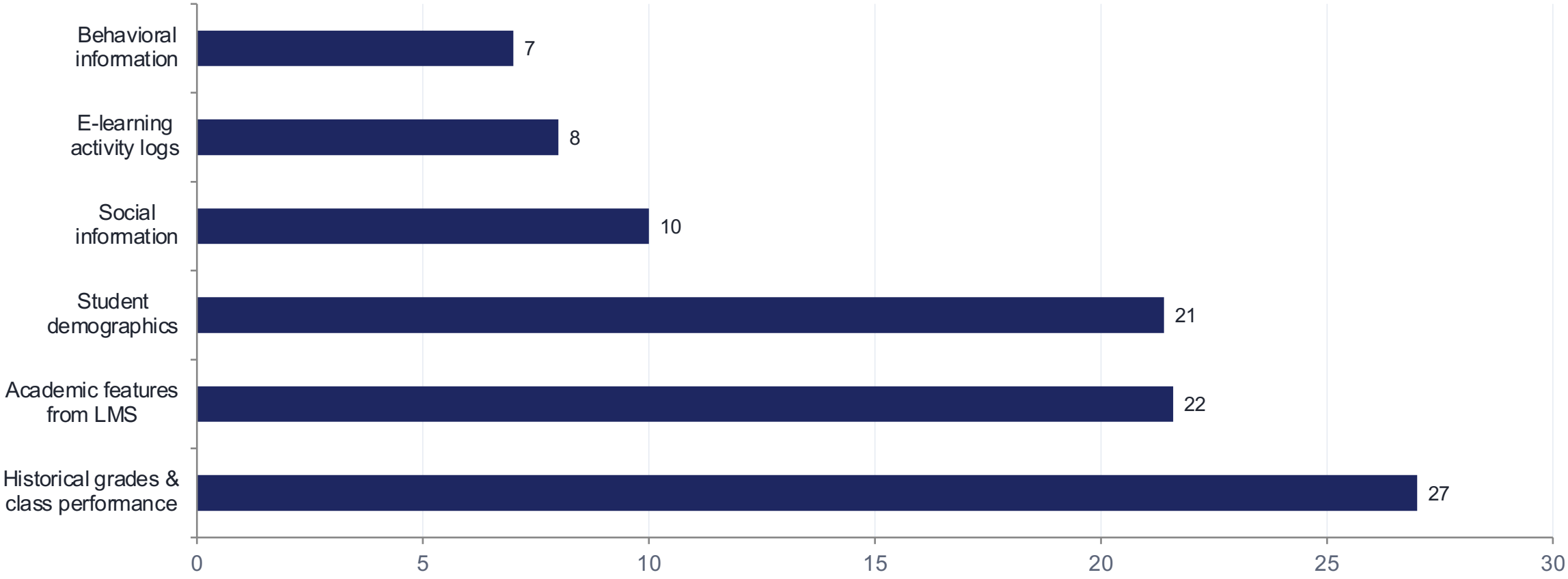


Key takeaway

Decision Trees and Naive Bayes are cheapest to train and remain top choices. Random Forest, ANN, and SVM trade extra computation for higher accuracy — a useful baseline-vs-advanced comparison for this project.

Over 50 distinct algorithms were used across the 162 studies reviewed.

Common Features Used for Prediction (in Percentage)



LITERATURE REVIEW

Gaps This Project Can Address

Interpretation is rare

Fewer than 10% of studies acted on predictions, and only 2 of 162 used explainable ML — most models remain "black boxes" for instructors.

Feature selection under-used

Under 40% of studies applied feature selection, even though it consistently improved model accuracy when tested.

Failure vs. dropout imbalance

129 studies focused on at-risk/failure prediction versus only 33 on dropout — failure-risk prediction is well-precedented and feasible for a course-level project.

Positioning: This project targets failure-risk classification using well-precedented features and algorithms, while adding a simple feature-importance explanation layer — directly addressing the interpretability gap noted above.

FROM EVIDENCE TO METHOD

Methodology & Algorithm Justification

Design Choice	Selected Option	Justification from Literature
Approach	Classification	Used in 80%+ of reviewed studies; matches a pass/fail or risk-level target variable.
Algorithms	Decision Tree, Random Forest, Naive Bayes (+ Logistic Regression baseline)	Top-5 most-used algorithms in the literature; DT/NB are computationally light (Table 2), RF typically improves accuracy over a single tree.
Features	Historical grades, LMS/academic data, demographics	The three most-used feature categories (≈70% combined) across 162 studies.
Feature selection	Information Gain-based selection	Most frequently used method (23.5% of studies) and shown to improve accuracy.
Evaluation	Accuracy, Precision, Recall, F1, Confusion Matrix	Standard classification metrics used consistently across the reviewed literature.

IMPLEMENTATION

Dataset Overview

5,000	17	3 → 2	0
student records	raw columns	outcome classes, collapsed to binary	missing values

Raw features: Age, Gender, Department, Year, Semester, Attendance, Assignment/Quiz/Midterm/Final Score, Previous GPA, Study Hours/Day, LMS Login Count, Absence Count, Scholarship

Three feature categories

The raw dataset combines academic performance (Assignment, Quiz, Midterm, Final Score, Previous GPA), behavioral engagement (Attendance, Study Hours, LMS Login Count, Absence Count), and demographics (Age, Gender, Department) — giving the model direct evidence of both what a student has scored and how they engage with the course.

Target variable: binary Risk flag — At Risk (0): 711 students (14.2%) | Not At Risk (1): 4,289 students (85.8%)

Original 3-class outcome (Pass / At Risk / Fail) was collapsed to binary for classification — this imbalance matters for interpreting the results (see Discussion).

IMPLEMENTATION

Preprocessing & Feature Engineering

1. Preprocessing

Cleaned and encoded the 5,000 × 17 raw dataset — categorical columns (Gender, Department, Scholarship) encoded, no missing values to impute, target binarized to At Risk / Not At Risk.

2. Feature Engineering — 7 new features derived

- average_assessment_score
- attendance_risk
- low_grade_flag
- engagement_score
- academic_progress_score
- total_lms_engagement
- performance_index

Why engineer these features?

- average_assessment_score — a single combined signal from Assignment + Quiz + Midterm, smoothing out noise from any one score
- attendance_risk / low_grade_flag — turn raw numbers into risk-oriented signals a model (and an instructor) can act on directly
- engagement_score / total_lms_engagement — combine login count and activity into one behavioral engagement summary
- performance_index — a composite score summarizing overall academic standing across features

IMPLEMENTATION

Feature Selection

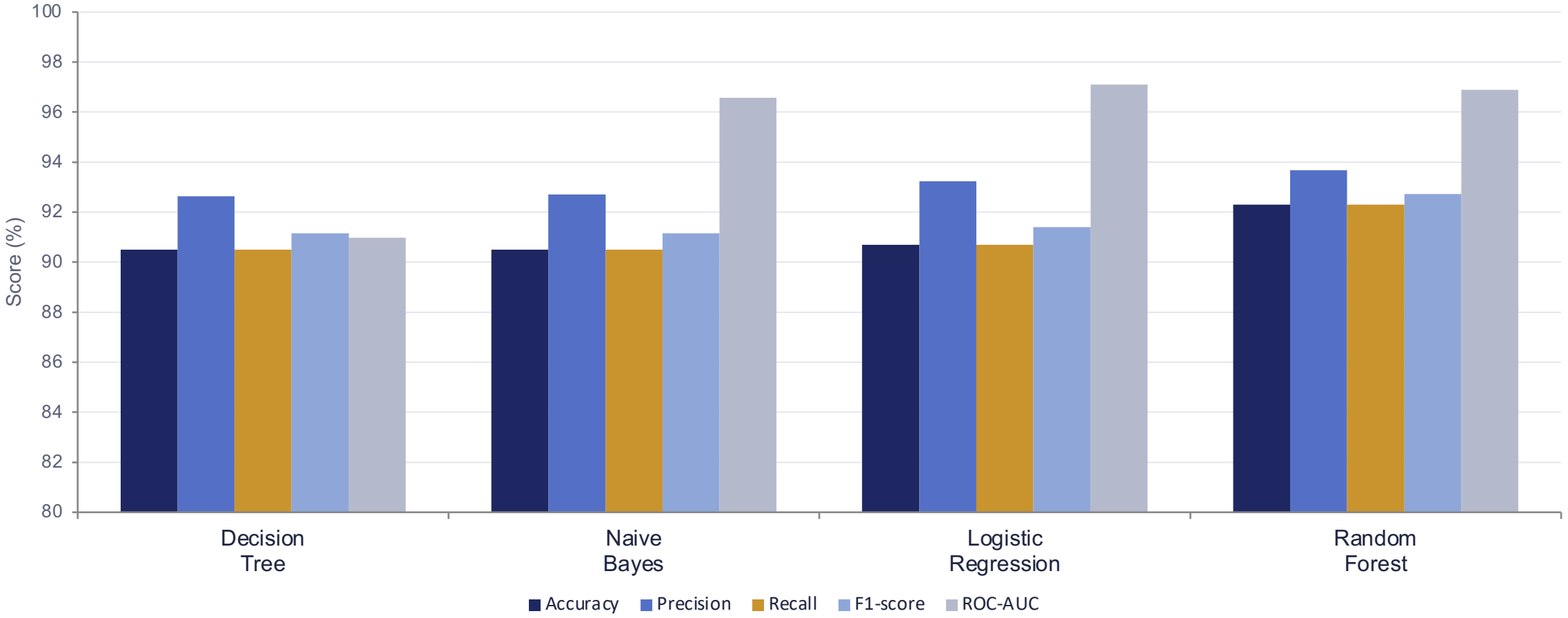
Reduced from 22 engineered/raw columns down to the 10 most predictive features

1	Historical Grade
2	Previous Grade
3	Assignment Score
4	Midterm Score
5	Quiz Score
6	Study Hours
7	LMS Login Count
8	Attendance Percentage
9	LMS Activity Count
10	Gender: Female

RESULTS

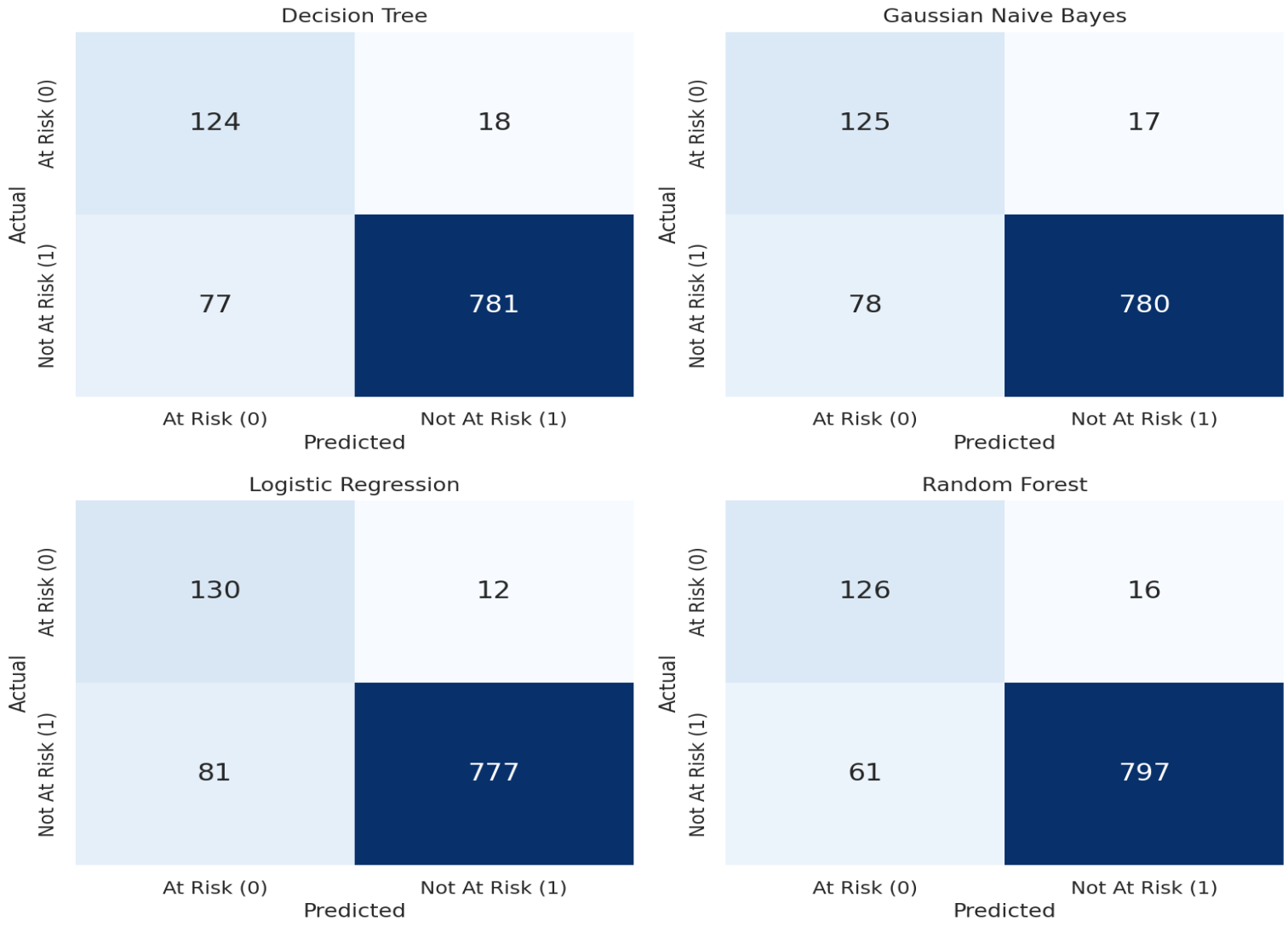
Model Comparison

Test set, n = 1,000 (20% held out) · Random Forest selected as best by overall accuracy



RESULTS

Confusion Matrices — All 4 Models



RESULTS — DEEPER LOOK

Accuracy Isn't the Whole Story

The target is imbalanced: 85.8% of students are Not At Risk. A model that always predicts "Not At Risk" would already score ~85.8% accuracy without learning anything.

Model	Recall — At Risk	Precision — At Risk
Logistic Regression	91.5% (highest)	61.6%
Random Forest	88.7%	67.4% (highest)
Naive Bayes	88.0%	61.6%
Decision Tree	87.3%	61.7%

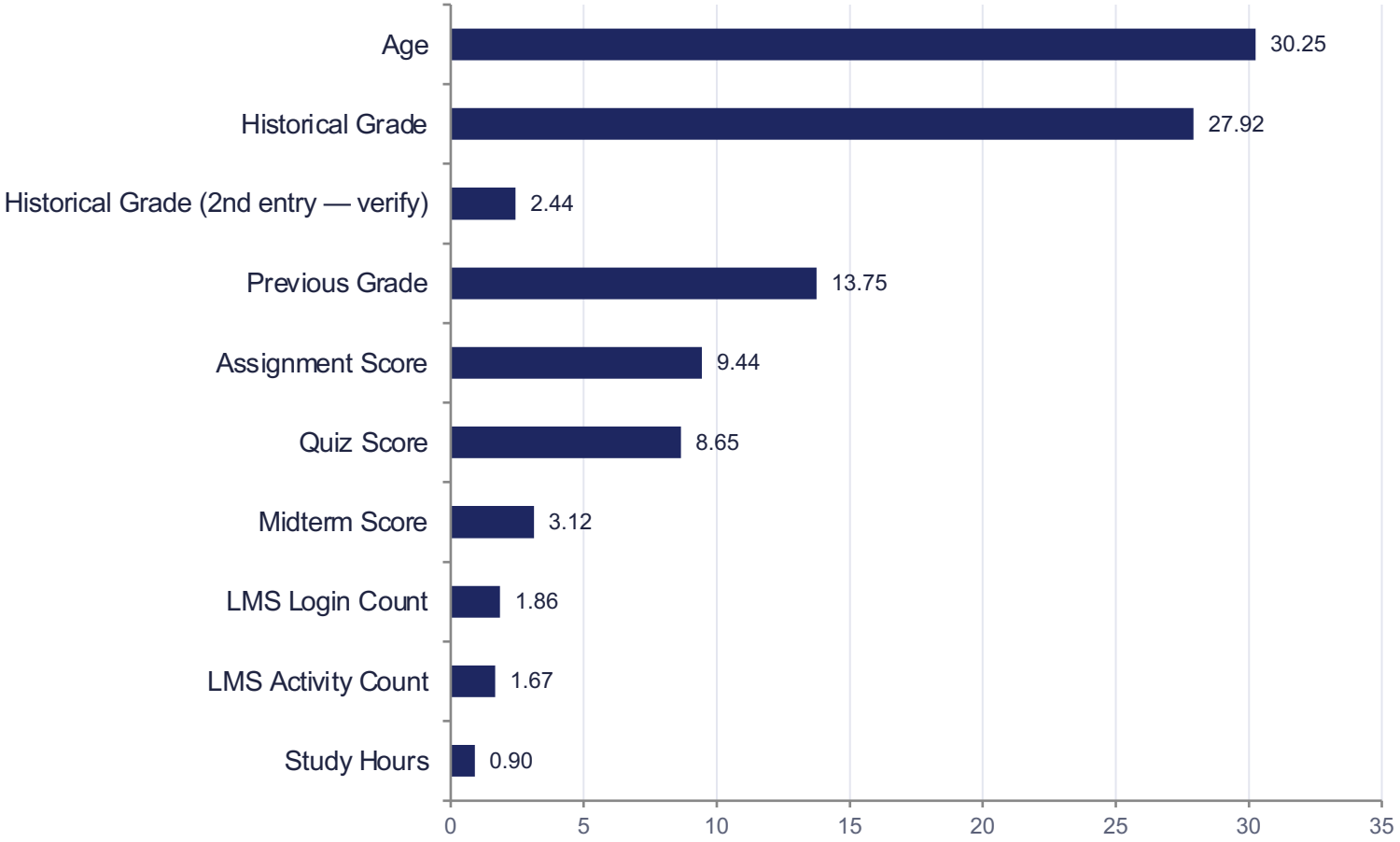
Why this matters

For an early-warning system, missing an at-risk student (a false negative) is more costly than a false alarm. By that standard, Logistic Regression — not Random Forest — catches the most at-risk students, even though it has lower overall accuracy. "Best model" depends on which error you care most about avoiding.

RESULTS — INTERPRETATION

Feature Importance

Random Forest, top 10 of 23 total features



DISCUSSION

Interpreting the Results

Scores drive the prediction

Historical Grade, Previous Grade, and Assignment/Quiz/Midterm Scores dominate the feature importance ranking. This means the model's strength comes primarily from academic performance data the school already records for grading — not from behavior or demographics.

Accuracy vs. recall trade-off

Random Forest wins on accuracy; Logistic Regression wins on catching at-risk students. Which one is "best" depends on whether the school prioritizes overall correctness or minimizing missed at-risk cases.

Imbalance changes the baseline

With an 85.8% majority class, all four models beat the naive baseline by only 5-7 points on accuracy. The ROC-AUC scores (0.91-0.97) are a fairer picture of true discriminative power than accuracy alone.

DISCUSSION

Limitations & Future Work

Limitations

- Severe class imbalance (14.2% At Risk) makes raw accuracy an unreliable headline metric
- Precision on the At-Risk class is only ~62-67% — a meaningful false-alarm rate

Future Work

- Apply class-balancing techniques (SMOTE, class weighting) and re-evaluate
- Report precision/recall per class as standard, not just overall accuracy
- Decide the deployed model based on the school's actual cost of a missed at-risk student vs. a false alarm
- Validate on real institutional data, not only simulated records

CONCLUSION

On 5,000 students with real academic scores, Random Forest reaches 92.3% accuracy — but Logistic Regression catches more at-risk students, and both must be read against an 85.8% majority-class baseline.

- Academic performance features (historical grades, assignment/quiz/midterm scores) are the dominant predictors of student risk.
- Model choice should follow the deployment goal: accuracy (Random Forest) or catching at-risk students (Logistic Regression).
- Reporting per-class metrics, not just accuracy, is essential given the class imbalance.

Thank you — Questions & Discussion

Reference

Alalawi, A., Athauda, R., & Chiong, R. (2023). **Predicting student performance and identifying at-risk students in higher education: A systematic literature review**. *Computers and Education: Artificial Intelligence*, 5, 100185.