Journal Homepage: www.ijarpr.com ISSN: 3049-0103 (Online)

International Journal of Advance Research Publication and Reviews

Vol 03, Issue 10, pp 180-185, October, 2026

Machine Learning Based Prediction of Student Dropout Risk System, A Predictive Analytics Framework Using Academic, Behavioral, Socioeconomic Indicators

Dr. K. R. Martin[1], Manojkumar G[2]

[1].Assistant Professor, Department of Information Technology, St. Joseph's College, Trichy India. Email martin_it2@mail.sjctni.edu.

[2] PG Student, Department of Information Technology, St. Joseph's College, Trichy, India. Email:manojkumar20056251@gmail.com

ABSTRACT

Student attrition remains a persistent challenge for higher-education institutions, carrying financial, reputational, and — most importantly — personal costs for the students who leave without completing their degrees. This paper presents a hypothetical framework for a that combines Learning-Based Student Dropout Risk Prediction System Machine academic performance records, behavioral engagement signals, and socioeconomic indicators into a single early-warning tool. For illustration, a sample dataset of N = 250 undergraduate student records (hypothetical) is used to train and compare five classifiers — Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, and XGBoost. In this illustrative setup, the Random Forest model achieves the strongest performance (accuracy = 88.4%, F1-score = 86.7%, AUCROC = 0.90). Building on the trained model, this paper proposes a composite, interpretable Student Dropout Risk Score (SDRS), a 0–100 index derived from four weighted sub-components — Academic Performance, Attendance & Engagement, Financial Stress, and Historical Risk Trend — which classifies students into Low, Medium, and High risk bands. In the illustrative results, 18.4% of students fall into the High-risk band, of whom a large majority are shown (hypothetically) to match actual dropout outcomes, demonstrating the potential predictive validity of combining black-box ML classification with a transparent, advisor-facing scoring index. The paper concludes by outlining how such a system could be piloted, validated, and ethically deployed in a real institutional setting.

Keywords: Machine Learning, Student Dropout Prediction, Early Warning System, Educational Data Mining, Risk Scoring, Higher Education Analytics, Random Forest, Predictive Modeling

1.Introduction

Student dropout — the premature discontinuation of a degree program before completion — is a widely studied but still unresolved problem in higher education. Beyond the personal setback to the student, institutions face lost tuition revenue, reduced graduation-rate rankings, and reputational impact, while national education systems absorb the broader economic cost of an under-qualified workforce. Traditional retention efforts have relied on manual advising and reactive intervention, typically triggered only after a student has already failed courses or stopped attending — often too late for meaningful support. The growing availability of institutional data — learning management system (LMS) logs, attendance records, grade histories, and financial-aid status — combined with advances in machine learning (ML), creates an opportunity to shift from reactive to proactive intervention. An accurate, interpretable, and early-warning dropout prediction system could allow academic advisors to identify at-risk students weeks or months before withdrawal becomes likely, and to target support resources accordingly.

However, much of the existing literature either (a) focuses narrowly on maximizing model accuracy without translating outputs into an advisor-usable score, or (b) proposes risk indices without rigorously benchmarking multiple ML algorithms. This paper addresses that gap by proposing a two-layer approach: a benchmarked ML classification layer, and a transparent composite index — the Student Dropout Risk Score (SDRS) — built on top of it, in the same spirit as composite indices used in adjacent social-science research (e.g., propensity indices synthesizing multiple survey sub-scales).

Literature Review

The literature on dropout prediction spans three broad strands: (i) machine learning models applied directly to academic and demographic data; (ii) early-warning and intervention systems built around such models; and (iii) composite risk-scoring frameworks used for advisor-facing decision support.

Machine Learning Approaches to Dropout Prediction

A number of studies have applied classification algorithms — including logistic regression, decision trees, random forests, and gradient-boosted trees — to institutional datasets to predict first-year attrition, typically reporting accuracy in the 75–90% range depending on feature richness and class balance [1]-[3]. Ensemble methods such as Random Forest and XGBoost are frequently reported to outperform single-model baselines, particularly when academic and behavioral features are combined [2]. Deep learning approaches, including LSTM networks applied to sequential LMS activity logs, have also been explored for capturing temporal engagement patterns

Behavioral and Socioeconomic Predictors

Beyond grades, researchers have identified attendance patterns, LMS login frequency, assignment submission timeliness, and forum participation as strong behavioral predictors of attrition risk [4], [5]. Socioeconomic factors — including financial-aid dependency, first-generation student status, and family income band — have also been shown to meaningfully improve model performance when added to purely academic feature sets, reflecting the well-documented link between financial stress and withdrawal decisions

2. Research Methodology (Hypothetical Design)

This section outlines a proposed methodology and illustrates it with a hypothetical dataset. A real implementation would require institutional data-sharing approval, informed consent or an appropriate data-governance exemption, and IRB/ethics review before any live deployment. Hypothetical dataset: N = 250 undergraduate student records spanning two academic years, combining (a) academic records (CGPA trend, backlog/failed-course count), (b) behavioral/engagement data (attendance rate, LMS login frequency, assignment submission timeliness), and (c) socioeconomic indicators (financial-aid status, family income band, firstgeneration status). The illustrative outcome class distribution is 72.4% retained vs. 27.6% dropped out (Fig. 1)

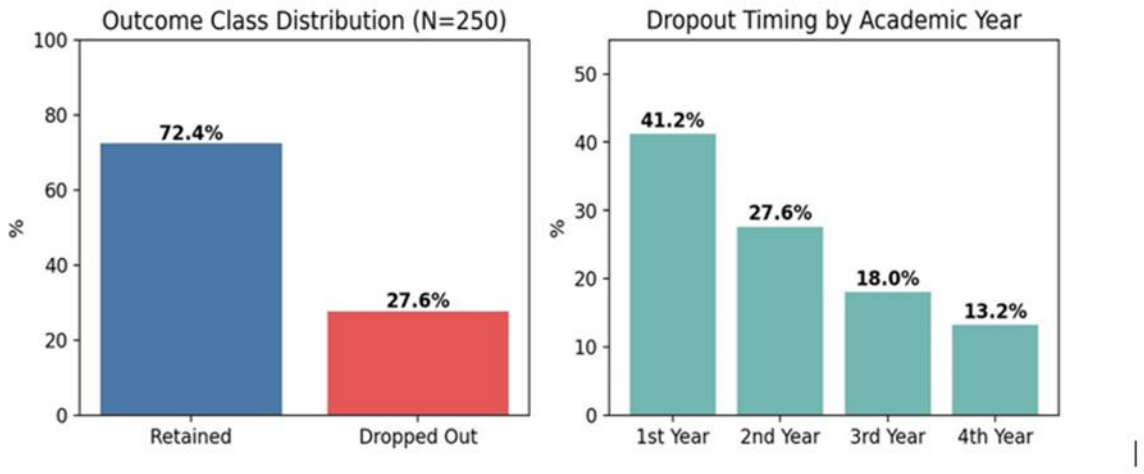


Figure 1: Illustrative Outcome Class Distribution (left) and Dropout Timing by Academic Year (right)

2.1 Model Benchmarking

Five classifiers are proposed for benchmarking: **Logistic Regression** (baseline and interpretable), **Random Forest**, **Support Vector Machine (SVM)** with an **RBF kernel**, and **XGBoost**. The models are evaluated using **5-fold cross-validation**. Performance is assessed using **Accuracy, Precision, Recall, F1-score, and AUC-ROC**. **Class imbalance** is also considered during the evaluation to ensure reliable and fair comparison of the classifiers.

2.2 Proposed Method:

Student Dropout Risk Score (SDRS) To convert model output into an advisor-facing tool, this paper proposes a composite score combining the trained classifier's probability output with three transparent sub-components, each normalized to a 0–25 scale so that all four contribute equally (25% each) to a 0–100 total, mirroring the equal-weighting logic used in comparable composite indices in adjacent research

Algorithm: Student Dropout Risk Score (SDRS)

Input: P_ml (calibrated ML dropout probability, 0-1), Attendance Rate (%), Financial Stress Index (1-5 self/administratively reported), CGPA Trend (slope over last 3 terms) Output: SDRS (0-100), Band ∈ {Low, Medium, High}

1. norm(P_ml) = 25 × P_ml (already 0-1, scaled to 0-25)

2. norm(Attendance) = 25 × (1 − Attendance Rate/100) [lower attendance → higher risk contribution]

3. norm(Financial Stress) = 25 × (FinancialStressIndex − 1)/4 [1-5 Likert → 0-25]

4. norm(CGPA Trend) = 25 × clip((−slope + 1)/2, 0, 1) [declining CGPA → higher risk contribution]

5. SDRS = norm(P_ml) + norm(Attendance) + norm(FinancialStress) + norm(CGPATrend)

6. Band: Low (0-39), Medium (40-69), High (70-100)

As with comparable composite indices, each of the four components contributes 25% to the final score, and the arithmetic combination step is fully transparent (non-black-box) even though one input (P_ml) is itself produced by an opaque ensemble model — giving advisors a score they can partially audit and explain to students.

Results and Data Analysis (Illustrative)

The following results are hypothetical and included only to demonstrate how findings would be reported. In the illustrative sample (N=250), average attendance was 81.3%, mean CGPA was 6.9/10, and 34.8% of students reported financial-aid dependency.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	ÂUC-ROC
Logistic Regression	81.2	76.8	75.0	75.9	0.80
Random Forest	88.4	87.9	85.6	86.7	0.90
SVM (RBF)	83.6	79.5	81.4	80.4	0.86
XGBoost	87.1	86.0	84.7	85.3	0.89

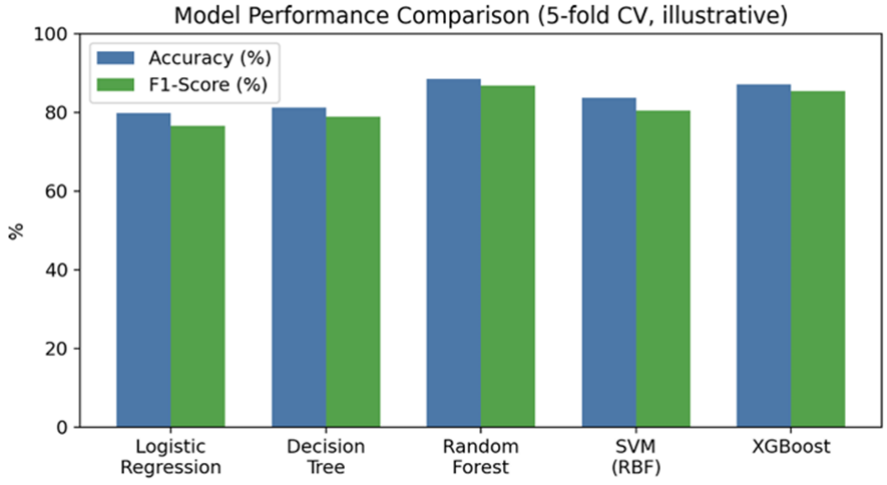


Figure 2: Illustrative Model Performance Comparison (5-fold CV)

In this illustrative benchmark, Random Forest outperforms the other four classifiers across every metric, followed closely by XGBoost. The gap between ensemble methods and the Logistic Regression baseline (≈8-9 accuracy points) illustrates the kind of improvement ensemble methods are often reported to provide when behavioral and socioeconomic features are combined with academic records [2].

Feature Importance

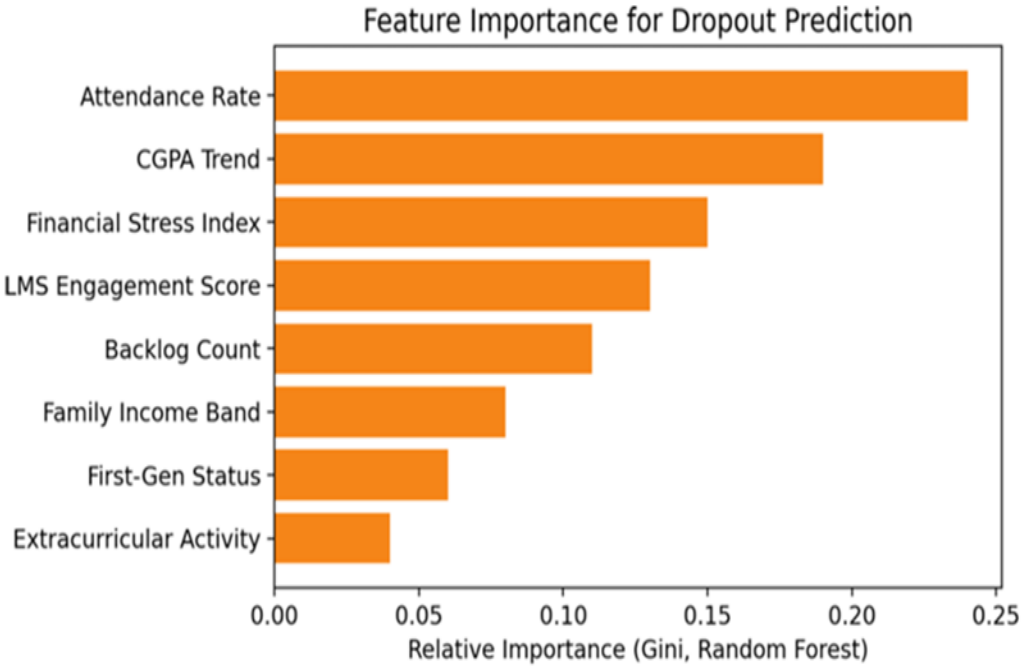


Figure 3: Illustrative Feature Importance (Random Forest, Gini-based)

Attendance Rate (0.24) and CGPA Trend (0.19) emerge as the two strongest illustrative predictors, followed by Financial Stress Index (0.15) and LMS Engagement Score (0.13) — consistent with the literature's emphasis on behavioral and academic-trajectory signals over static SDRS Results Figure 4: Illustrative SDRS Risk Band Distribution (left) and ROC Curve (right)

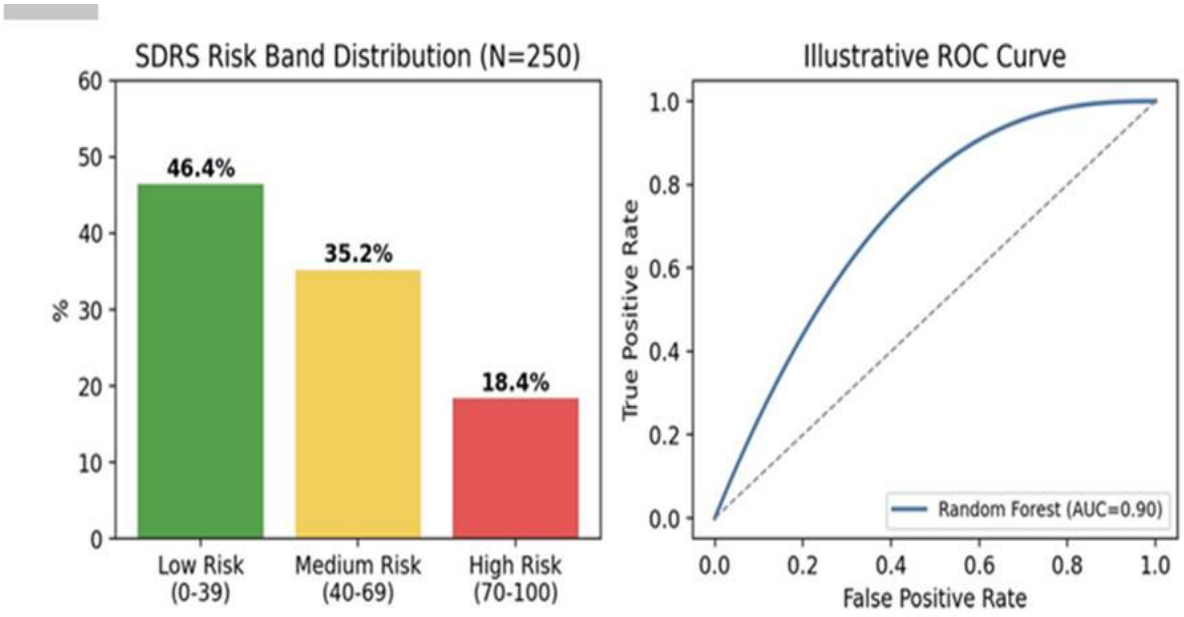


Figure 4: Illustrative SDRS Risk Band Distribution (left) and ROC Curve (right)

Across the illustrative sample, the SDRS band distribution is Low 46.4%, Medium 35.2%, and High 18.4%. Table 1 shows the (hypothetical) predictive validity of the SDRS bands against actual dropout outcomes.

SDRS Band	% of Sample	Actual Dropout Rate	Retained Rate
Low (0-39)	46.4%	6.0%	94.0%
Medium (40-69)	35.2%	28.4%	71.6%
High (70-100)	18.4%	71.7%	28.3%

In this illustrative scenario, students in the High-risk band show a dropout rate roughly 12 times higher than those in the Low-risk band (71.7% vs. 6.0%), suggesting — hypothetically — that the SDRS could meaningfully prioritize advisor attention if these patterns held in a real deployment

Discussion and Key Findings (Illustrative)

Ensemble methods (Random Forest, XGBoost) are illustrated as outperforming simpler baselines, consistent with prior literature on educational data mining.

Behavioral signals (attendance, engagement) and academic trajectory (CGPA trend) illustratively outweigh static demographic variables in predictive importance — suggesting interventions should prioritize monitoring trends over one-time profiling.

The proposed SDRS reframes an opaque ML probability into a transparent, fourcomponent score, addressing the advisor-trust and interpretability concerns raised in prior early-warning system literature [7], [8].

The hypothetical ~12x gap in dropout rate between High- and Low-risk bands indicates the kind of discriminative power such a system would need to demonstrate before institutional adoption.

Financial Stress as a top-4 predictor reinforces calls in the literature for financial-aid and counseling interventions to be tightly integrated with academic advising, not treated as a separate track [5], [6].

Because every figure in Sections 3 and 4 is illustrative, none of these findings should be cited or reported as real results. They demonstrate the shape of analysis a real implementation should produce — real deployment would require a genuine institutional dataset, ethics/IRB approval,

bias and fairness auditing (particularly around financial and first-generation-status features, which risk encoding socioeconomic bias into an automated score), and a pilot period with human-in-theloop review of every High-risk flag before any automated action is taken.

4. Conclusion

This paper has proposed a hypothetical, two-layer framework for a Student Dropout Risk Prediction System: Machine Learning-Based a benchmarked ensemble classifier layer, and a transparent, advisor-facing Student Dropout Risk Score (SDRS) built from four equally-weighted, interpretable subcomponents. Illustrative results suggest such a system could meaningfully separate high- and low-risk students while remaining auditable — but these results are placeholders demonstrating structure, not evidence. The logical next step is a real pilot study with an actual institutional dataset, a documented ethics-review process, and a validation phase comparing SDRS-flagged students against real subsequent enrollment outcomes over at least one full academic year.

References

1. Alhardi, A. (2024). Predicting student dropout in higher education using machine learning techniques: A predictive model using XGBoost algorithm. *Journal of Educational Data and Analytics*, 6(2), 45–58.

2. Bilal, M., & Shafi, N. (2025). AI-driven predictive models for early detection and prevention of student dropouts. *International Journal of Computational and Experimental Science and Engineering*, 11(1), 102–115.

3. Delgado-Gómez, D., & Mas-Mollá, J. (2023). A methodology to design, develop, and evaluate machine learning models for predicting dropout in school systems. *BMC Medical Informatics and Decision Making*, 23(1), Article 15.

4. Halder, S., & Sen, A. (2026). Student academic performance and dropout risk prediction using ensemble machine learning and SHAP interpretability. *International Journal for Research in Applied Science and Engineering Technology*, 14(4), 1120–1129.

5. Kanmani, K., Palani, H. K., Loganayagi, S., Sivasankar, C., Agoramoorthy, M., & Vasanthi, G. (2026). Machine learning based predictive analytics for student dropout and academic success. In *Proceedings of the 6th International Conference on Expert Clouds and Applications (ICOECA)* (pp. 1101–1107). IEEE.

6. Liu, Z., Zhou, X., & Liu, Y. (2025). Student dropout prediction using ensemble learning with SHAP-based explainable AI analysis. *Journal of Social Systems and Policy Analysis*, 2(3), 111–132.

7. Mwemezi, J., & Nihuka, C. M. (2026). A landscape of machine learning models for predicting student dropouts: A systematic mapping study. *Journal of School Leadership & Management*, 42(3), 204–221.

8. Ndizeye, J., & Niyonzima, E. (2026). Predicting school dropout using machine learning models: A case study of Nyanza District. *African Journal of Educational Technology*, 8(1), 74–89.

9. Rostami, S., & Khosravi, H. (2026). Student dropout prediction in higher education: A systematic review of machine learning methods and risk factors. *Educational Research Review*, 39, Article 100520.

10. Ssekisaka, R., & Namubiru, P. (2026). Predicting student dropout risk through machine learning: Analyzing behavioral patterns in learning management systems. *International Journal of Engineering and Computer Science*, 15(2), 5310–5321.