IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, January 2026
Impact Factor: 7.67

Predictive Analysis of Student Dropout

Atharwa Baban Pawar[1], Aryan Sachin Gangawane[2], Saksham Deepak Choundiye[3], Prof. N. S. Kharatmal[4]
Students, Computer Science and Engineering[1,2,3]
Lecture, Computer Science and Engineering[4]
Matsyodari Shikshan Sanstha College of Engineering and Polytechnic, Jalna, India
pawaratharwa75@gmail.com[1], Aryan.gangwane0025@gmail.com[2],
sakshamchoundiye01@gmail.com[3], nanditakhartmal27@gmail.com[4]

Abstract: *Dropout rates have remained a major problem for institutions of higher learning, and performance at the institutional level and academic and socio-economic development of students. Early identification of dyslexicsat-risk students allows schools to go beyond simply reacting to problems and take proactive steps to effectively address the issues of these individuals. By doing so interventions. The paper outlines an extensive framework for the predictive analysis of dropout in students. using data-driven and machine learning-based methods. In particular, the research becomes significant with respect to a multidimensional data set, involving the following variables: attendance trends, demographic characteristics, and socio-economic conditions. To overcome the limitations of traditional statistical methods to capture the complex and non-linear relationships of machine learning models are used. Comparison of ensemble learning techniques, such as Random Forest, Gradient Boosting (XGBoost) and LightGBM are used to determine the efficacy of models such as students having a high possibility of attrition. Additionally, methods of feature selection and data preprocessing. are used to improve the performance of models.*

Owing to the natural class imbalance problem in the student datasets, the number of dropout cases is much lower compared to regular cases. This class imbalance problem successful completion of models, the Synthetic Minority Over-sampling Technique (SMOTE) is employed to enhance sensitivity of the statistical models used, and reliability of the research implementation process. Findings from the experiment show, attendance regularity, and financial viability factors are among the most significant predictors of student persistence. The most accurate models are capable of achieving high levels of prediction accuracy and F1 scores, thus identification of high-risk students.

Keywords: *Dropout rates*

I. INTRODUCTION

Student dropout has been identified to be a very significant issue prevalent in modern-day higher education institutions globally, impacting not just academic and professional paths of the students but also the efficiency and reputation of the educational institutions. Whenever students withdraw from learning before finishing a programme, it results in an Academic Loss for the individual level and inefficient use of available resources-economic development, thus bringing "retention of students" into the foreground. Ideally, a "for policymakers and educators.

There has been a great shift in the education industry worldwide with the application of technology such as learning management systems (LMS), assessment systems, and student information systems. These systems enable educators to curate and manage learning resources data-driven digital environments produce massive data sets about student performance, attendance, and other related activities. Behavioral patterns, and demographic background information. If analyzed systematically, it can be very beneficial Perspectives or insights into student learning behavior and factors that cause academic disengagement and attritionConventionally, the selection of students who require attention for risk of poor results is done by observation or counsel feedback, or through analysis of end-of-term academic performance. Nevertheless, these methods seem rather responsive rather than proactive because they don't offer enough proactive

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 1, January 2026 Impact Factor: 7.67

support to students to improve their performance time for effective intervention. Furthermore, student dropout is a complex issue because it is shaped by a Academic difficulties, socio-economic issues, financial instabilities, and lack of social integration. The traditional models of statistics and linear modeling are inadequate in coping with such complex and non-linear inter-relationships amongst various variables.

II. LITERATURE SURVEY

Dropout prediction for students has gained considerable attention from the education analytics domain, as it holds a very important academic and institutional impact. “With the increasing popularity of Educational Data Mining (EDM) and Machine Learning applications” there is growing research on this matter. This was presently carried domain: the domain itself has grown from descriptive analysis to modeling. There is current literature “Highlights that student attrition is a complex, multi-factor problem influenced by academic, socio-economic, psychological factors,” social, behavioral, and institutional.

2.1 Evolution of Predictive Models

Initial research attempts addressing student dropout prediction were carried out using traditional statistical analysis methods by Logistic Regression and Linear Discriminant Analysis (LDA). The regression and LDA proposed in this study targeted academic performance factors such as school grades and test scores. Despite achieving high levels of interpretability in their approaches, they was incapable of extracting the non-linear and dimensionally complex patterns existing in educational data, hence producing minimal predictive performance.

2.2 Key Predictors of Student Attrition

Academic achievement has regularly appeared in the literature as the strongest predictor of students dropping out of school. Several studies have shown that low grades or poor performance in key or “bottleneck courses” in the early years of higher education predict low completion rates. semesters are known to be strong predictors of risk of attrition. However, behavioral cues based on Learning Management Systems (LMS), such as login rate, time spent on learning content, and patterns of assignment submissions, have received growing importance. Asselman et al. (2024) emphasized that decreased use of digital engagement may actually antecede dropout even when the academic trajectory is seemingly stable

2.3 Handling Class Imbalance in Dropout Prediction

A big technological challenge cited throughout the literature is related to class imbalance. Here, the challenge is posed by the abundance of dropout cases is much smaller than in successful completion instances Attempts to remedy this problem started with the development of" were increasingly applying data balancing methods such as the Synthetic Minority Over-sampling Technique (SMOTE) and cost-sensitive learning methods. Martins et al. (2024) proved that the class balance is much more crucial for the early warning system performance relative to algorithm selection

2.4 Interpretability, Ethics, and Explainable

There is recent interest in the literature regarding the roles of interpretability and ethics in predictive analytics. Sometimes, these models require openness and transparency to justify predictions even to educators and students. In this regard, techniques for Explainable AI (XAI), such as SHAP (SHapley Additive exPlanations) and LIME, become They are being increasingly incorporated into predictive models.

III. EXISTING MODELS AND THEIR LIMITATIONS

3.1 Manual and Reactive Monitoring Systems

Most learning institutions employ a manual observation process by members of the teaching staff or academic coordination team to establish students at risk of dropping out. It basically looks at the results for examinations, internal assessments, and attendance records. Because students only appear in the records after much academic deterioration has already occurred, these methods cannot offer much in terms of early support to those students who need warnings thus tend to limit the efficacy of interventions.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 1, January 2026 Impact Factor: 7.67

3.2 Dependence on Limited Academic Indicators

Current models have mainly centered on academic achievement variables. Other variables of a non-academic nature aspects such as socio-economic status, economic pressures, behavioral trends, and individual challenges are generally left out. This method has a tendency of decreasing the precision of the prediction of dropping out.

3.3: Usage of Traditional Statistical Models

Statistical methods such as Logistic Regression and Linear Discriminant Analysis are preferred because of their simplicity and interpretability. Nevertheless, these types of models assume that the variables are related linearly, which is challenged by capture complex, non-linear interactions that exist in typical educational data.

3.4 Limitations of Basic Machine Learning Techniques

Some institutions use machine learning models such as Decision Trees, Naïve Bayes, SVM Machines, and k-Nearest Neighbors. Though such models enhance the precision of predictions, there are problems such as overfitting, Outlier sensitivity, high computation complexity, and violation of independence among features.

3.5 Class Imbalance Problem

Data from the students is imbalanced, where the number of dropouts constitutes a small percentage of the entire population. Several models that currently exist are tuned for accuracy, and they may produce a poor "at-risk" detection for students. effectiveness in early warning systems.

IV. WORKING MODEL AND METHODOLOGY (SYSTEM ARCHITECTURE)

The proposed system is a data-driven prediction system that aims to detect students at risk of dropping out becauseanalysis of historical and current academic performance data. This approach adopts a structured pipeline for guaranteed accuracy prediction, scalability, and usability.

4.1 System Architecture Overview

Architecture of the System: The architecture of the system has a modular pipeline design. This consists of stages connected in a manner such that It encompasses activities such as data acquisition, preprocessing, feature development, modeling, evaluation, and decision-making. Moreover, it has a modular structure which makes it simpler to incorporate a variety of new sources of information such as attendance systems or Learning Management Systems. Learning Management Systems (LMS) with minimal adaptations.

4.2 Data Acquisition

Data is gathered from administrative sources, for example: Student records Records of classroom and/or building attendance Student records systems, and LMS systems. The dataset consists of performance (grades/marks, GPA), attendance percentage, demographic information, and enrollment details. Continuous data gathering helps ensure the system is updated.

4.3 Data Preprocessing and Cleaning

Missing, inconsistent, and noisy data can be found in the raw educational data. Missing data handling is included in this phase. Data preprocessing is a crucialhandling data with appropriate imputation methods, normalization of numeric columns to have equal scale, andEncoding of categorical variables. A proper preprocessure increases data quality and credibility.

4.4 Handling Class Imbalance

Datasets of student dropouts tend to be unbalanced, featuring dropouts as a minority class. In order to handle this problem, The “Synthetic Minority Over-sampling Technique” (SMOTE) is introduced. In order for SMOTE to generate

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, January 2026
Impact Factor: 7.67

new synthetic instances of minority class, so it can learn interesting patterns in relation to at-risk individuals instead of being biased toward the majority class.

4.5 Feature Selection and Engineering

Not all predictors affect a prediction equally. This necessitates techniques in feature selection for determining which predictors are most relevant factors, like trends of early academic performance, attendance patterns, and financial factors. This step lowers computational cost and promotes precise prediction.

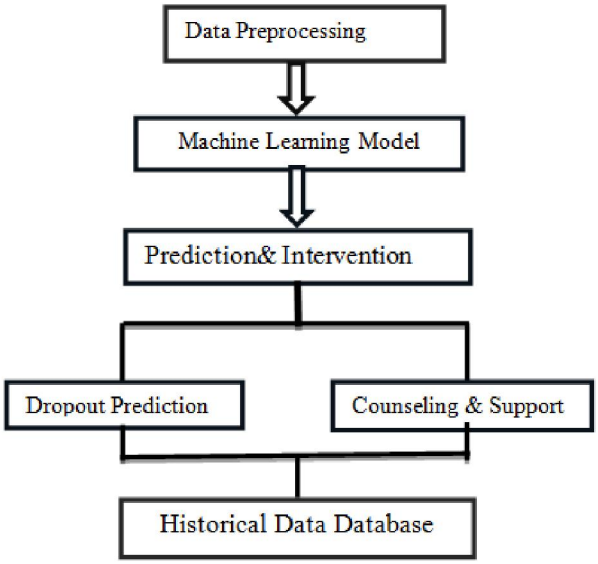

Fig 1.System Architecture

V. ALGORITHM USED IN EXISTING SYSTEM

Category	Existing System Algorithms / Techniques	Limitations in Existing System	Proposed System Algorithms / Techniques
Academic Performance Analysis	Simple percentage calculation and grade-based evaluation	Considers only final marks; ignores performance trends	Machine learning–based performance trend analysis using historical data
AttendanceMonitoring	Fixed threshold-based attendance checking	Late detection; no early warning mechanism	Predictive attendance behavior analysis using classification models
Dropout Identification Logic	Rule-based systems (fail count, low attendance rules)	Rigid rules, lack adaptability	Adaptive prediction using Logistic Regression and Decision Trees
Statistical Analysis	Mean score analysis, pass–fail ratio, basic statistics	Cannot model complex, multi-factor relationships	Multivariate data analysis with feature selection techniques
Machine Learning Usage	Limited use of basic classifiers	Low accuracy and poor generalization	Ensemble learning models such as Random Forest and XGBoost

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 1, January 2026 Impact Factor: 7.67

Class Imbalance Handling	Standard classification without sampling	Fails to identify minority dropout cases	SMOTE-based data balancing for improved dropout detection
Feature Selection	Focused mainly on GPA and exam scores	Ignores socio-economic and behavioral factors	Advanced feature selection including socio-economic and behavioral attributes
Data Handling	Manual data entry or end-of-semester batch processing	Delayed identification of at-risk students	Automated pipeline with continuous data integration
Risk Identification Method	Reactive detection after academic decline	No early-stage prediction	Early dropout risk prediction using historical and real-time data
Interpretability	Complex statistical outputs or black-box models	Difficult for large institutions	Explainable AI using SHAP/LIME for transparent decision support
Scalability	Suitable only for small datasets	Not effective for large institutions	Scalable models handling large and evolving student datasets
Decision Support	Human-dependent decision making	Inconsistent and delayed interventions	Data-supported intervention recommendations for educators

VI. OUTPUT/RESULT AND DISCUSSION

Category	Parameter / Task	Technique / Tool Used	Discussion
Student Academic & Attendance Data	Institutional database & dataset	CSV dataset	Student academic and attendance data collected from institutional records
Data Input	Handling missing values and normalization	Noise-free standardized dataset	Removes missing values and biases present in raw data
Data Processing	Handling and cleaning of relevant attributes	Noise-free and cleaned dataset	Ensures selection of useful and reliable data
Feature Selection	Identification of relevant attributes	Key features selected	Reduces noise and improves model performance
Model Training	Separating training dataset	Selected key features	Ensures accurate learning from historical data
Dropout Prediction	Training predictive model	Prediction-based ML model	Predicts whether a student is at risk of dropout
Accuracy Evaluation	Trained dropout prediction model	Performance metrics	Measures accuracy of the prediction model
Precision &	Training predictive model	Evaluation metrics	Evaluates model reliability

DOI: 10.48175/568

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 1, January 2026 Impact Factor: 7.67

Recall			and correctness
Comparative Analysis	Trained dropout prediction model	Feature-based comparison	Compares different models and feature sets
Output Visualization	Displaying results using graphs and charts	Data visualization tools	Provides clear and understandable results
Output Visualization	Student insights based on prediction results	Key result formatting and analysis	Visual representation of student risk levels
Decision Support	Student prediction model	Final selected model	Helps educators make informed intervention decisions

VII. CONCLUSION

Powerhouse rates are a complex miracle that affect both pupil achievement and institutional effectiveness. Conventional ways homemade monitoring, fixed rules, and introductory statistical analysis may not inescapably offer an early accurate identification of threat scholars. This study shows that prophetic analysis via advanced machine literacy models, Random Forest, and XG Boost provides a better result within the span of contemporaneously assaying colorful factors like academic achievement, class attendance, geste and socio- profitable pointers.

The proposed system, bettered with effective data processing ways similar as SMOTE, overcomes class- imbalance problems. Reaches high recall situations, that a significant number of scholars who are potentially at threat are linked. By incorporating the vaticination model, colorful scholars who may be at threat are transubstantiating it into an Early Warning System(EWS), associations are suitable to make targeted interventions, similar as personalized comforting, mentoring, and fiscal backing. Prophetic analytics is thus feasible, scalable, and data- grounded way through which the problem of dropping out of educational institutions could be averted effectively.

REFERENCES

[1]B. Roychowdhury, S. Sarkar, and R. Bandyopadhyay, " Predicting Student " Data Mining ways, " International Journal of Computer Applications, vol. 180, no. 25, pp. 12- 18, 2002

[2] A. Kaur, S. Kaur," A Review on Student Dropout Prediction Using Machine Learning Algorithms," International Journal of Advanced Research in Computer Science, vol. 10, no. 3, pp. 45- 52, 2019.

[3]S. Sharma and R. Kumar," Educational Data Mining Student Dropout Prediction Using Bracket Algorithms" International Journal of Engineering and Technology, vol. 9, no. 2, pp. 124- 130, 2022

[4]S. Bhatia, P. Singh, and M. Gupta, " operation of Machine literacy in Predicting Student Performance and Early Warning Systems, " International Dropout Risk, " Journal of Educational Technology Systems, vol. 49, no. 3, pp. 382 – 400, 202

[5]R. Kumar, Sharma A. "Predictive Analytics in Education A Case Study on Student Retention," Procedia Computer Science. Computer Science, vol. 167, pp. 1948-

[6]T. Kotsiantis," Educational Data Mining Review of the State- of- the- Art," Expert Systems with operations Vol. 33, No. 1, pp. 135

[7]J. Romero and S. Ventura, "Educational Data Mining Survey from 1995 to 2005," Expert Systems with Applications, vol. 33, no. 1, pp. 135

[8]P. Singh & R. Jain, " Machine literacy ways for Predicting Student Dropouts A relative Study, " International Journal of Innovative exploration in Science, Engineering and Technology, vol. 8, no. 12, pp. 14500 – 14508,

[9]A. Pandey and R. Mishra, "Predicting Student Dropout Using Random Forest Classifier," International Journal of Computer Applications, Volume 175, Issue 24(2020), pp. 1-22

[10]S. Kumar and A. Bansal, "A Study on Predictive Modeling for Student Dropout in Higher Education," Journal Of Information Technology Education Research, vol. 19, pp. 1- 20.