

Original Article**Predictive Analytics for Student Academic Performance Using Educational Data**

Alsaboore Khalil Sayyed

UG Student, B.Tech CSE (AI and ML) Woxxsen University, Hyderabad

Manuscript ID:
CSJ-2026-020502

Volume 2

Issue 5

Pp. 7-12

October 2026

ISSN: 3067-3089

Submitted: 08 Sept. 2026

Revised: 16 Sept. 2026

Accepted: 05 Oct. 2026

Published: 31 Oct. 2026

Correspondence Address:

Alsaboore Khalil Sayyed
UG Student, B.Tech CSE (AI and ML)
Woxxsen University, Hyderabad
Email: sayyedalsaboore48n@gmail.com

Quick Response Code:

Web: <https://csjour.com/>

DOI: 10.5281/zenodo.23029748

DOI Link:

<https://doi.org/10.5281/zenodo.23029748>

Creative Commons

**Abstract**

Educational data mining (EDM) techniques, as used in learning analytics, have been widely embraced as means of predicting students' academic performance. Machine learning models that can predict such parameters as course grade, grade point average, and likelihood of dropping out can be generated using demographic, previous academic record, and LMS engagement information. Recent surveys and studies suggest that compared to conventional statistical approaches, machine learning methods, particularly the use of random forest, gradient boosting, support vector machine, and neural networks, show better results if provided with data from diverse sources. At the same time, systematic reviews reveal methodological problems, such as heterogeneity of approaches, weak external validity, and ethical issues not being addressed properly.

This paper suggests a framework for predicting end-of-semester academic performance in higher education institutions using institutional data.[1] It describes a quantitative non-experimental design where supervised machine learning models are trained using demographic, prior achievement, and LMS engagement features with careful preprocessing, cross-validation, and comprehensive metric assessment. The study will be aimed at (a) producing accurate predictions of students' academic performance, (b) identifying predictors that will help to develop strategies of teaching and supporting students, and (c) demonstrating how the predictive model will serve as an engine for early warning systems and intervention.[2] Combining recent research findings on EDM and learning analytics, the paper provides both the methodology and the conceptual basis for the implementation of predictive analytics in higher education.[3]

Searched academic databases for recent work on educational data mining and learning analytics to ground this paper in current evidence.

Keywords: Predictive Analytics, Educational Data Mining, Student Academic Performance, Machine Learning, Learning Analytics, Learning Management System (LMS), Early Warning System, Explainable AI, Academic Risk Prediction, Educational Data.

Introduction

Digital learning environment proliferation has changed the way educational institutions collect, process, and analyze data. Information systems for students, learning management systems, e-assessment systems, and e-government services altogether provide a detailed picture of learners' journey from admission to graduation. The domains of educational data mining and learning analytics are focused on application of the collected data for learning analysis and improved decision-making, with student academic performance prediction being among the best-developed subdomains of the former. machine learning methods. These surveys indicate the application of a variety of machine learning models, including logistic regression, decision trees, k-nearest neighbors, naïve Bayes, random forests, gradient boosting, support vector machines, and neural networks of different types. As a rule, non-linear models and ensembles demonstrate better results compared to linear models, especially when the feature set includes complex interactions obtained from LMS logs and behavioral features. Prediction of the students' academic performance is defined as the task of modeling the relation between the input variables (demographic parameters, grades, and behavioral factors) and academic outcomes (course grades, GPA, success/failure status, drop-out) by means of statistical and machine learning methods. Many researchers prove the efficiency of such models in identifying "at-risk" students and, consequently, providing efficient allocation of the support resources. Several recent surveys summarize two decades of research on prediction of student performance with the use of EDM and On the other hand, the existing literature notes the insufficiency of merely predictive accuracy for the achievement of educational outcomes. Systematic reviews of studies on the prediction of dropout and academic performance report inconsistencies in evaluation procedures, lack of external validation across populations, and lack of interpretability and considerations of fairness.

Creative Commons (CC BY-NC-SA 4.0)

This is an open access journal, and articles are distributed under the terms of the Creative Commons Attribution-Non Commercial-ShareAlike 4.0 International Public License, which allows others to remix, tweak, and build upon the work no commercially, as long as appropriate credit is given and the new creations are licensed under the identical terms.

How to cite this article:

Sayyed, A. K. (2026). Predictive Analytics for Student Academic Performance Using Educational Data. *Computer Science Journal*, 2(5), 7–12. <https://doi.org/10.5281/zenodo.23029748>

Moreover, the majority of models are trained and tested on one dataset without any analyses of error distribution among different subgroups of students, thus raising concerns about the risks that predictive algorithms may unintentionally promote inequalities among the target population.

Recent empirical evidence tries to deal with the abovementioned issues by implementing (a) multiseason/multicourse databases; (b) fine-grained data about interaction within LMS, e.g., clickstream information, submission time of assignments, forum participation rate; (c) model-agnostic interpretable approaches (SHAP values, partial dependence plots), etc.[4] Thus, studies show that predictive models could be built achieving very high accuracy – often exceeding 0.80 AUC for binary classification problem – and providing a meaningful interpretation of what factors contribute to the prediction results.[5] Against the background of the abovementioned developments, the current paper offers to conduct a research project on the topic of "Predictive Analytics for Student Academic Performance Using Educational Data." It includes an analysis of major themes from the literature review and development of a methodological framework for developing, testing, and interpreting the predictive models.

Rationale of the Study

Although there is a fast-growing body of research on EDM and learning analytics, a number of gaps exist that justify the current study. Firstly, despite the existence of numerous surveys and systematic reviews, it can be noted that most prediction studies use rather narrow sets of features that mainly comprise prior grades and some basic demographic characteristics. At the same time, engagement with the LMS, temporal data, and institutional support services data are largely overlooked. In case when multiple data sources are used, it is also characteristic that the process of feature extraction and pre-processing is not described in a transparent way. Secondly, it should be noted that in case of existing literature on this topic, there is a high variability in the level of methodological rigor of conducted research. Most common problems are represented by the use of small and convenient samples, lack of measures aimed at solving class imbalance of at-risk and not-at-risk students, and non-adequate validation procedures due to reuse of the same cohort for model development and testing procedures. Thirdly, although the use of predictive models is usually associated with an attempt to improve retention and performance of students, there are only a few works where the use of predictions for the improvement of learning experiences is considered in detail. There are recent studies of LMS-based predictive frameworks that show how the integration of predictive models can have a positive impact on learning outcomes, but this approach is not widespread.

Finally, the problems of privacy, informed consent, algorithmic bias, and transparency are quite often discussed, but not taken into account when designing prediction models. In systematic reviews of dropout prediction studies, the lack of systematic evaluation of fairness is mentioned together with the absence of consideration of the process of communication of the status of at-risk label.

Therefore, this study is motivated by the necessity of creation of a rigorous, comprehensive, and ethical framework for predicting students' academic performance on the basis of institutional data. The description of design choices and emphasis on interpretability and fairness together with the connection of predictions with actions will allow contributing to the scholarly and practical aspects of the topic.

Objectives

The research objectives are as follows:

1. To build predictive models of students' academic performance based on multi-source educational data consisting of demographic characteristics, previous academic achievement and LMS engagement metrics.
2. To conduct comparative analysis of various machine learning algorithms (logistic regression, decision trees, random forests, gradient boosting, support vector machines and neural networks) with respect to their ability to predict the outcomes of the semester (course completion/ failure, GPA-based performance categories).
3. To identify and rank the predictors of academic performance based on model-based (feature importance) and model-agnostic (SHAP values, partial dependence plots) approaches to model interpretation.
4. To propose the conceptual model of the early warning and intervention system that will involve triggering certain support interventions (tutoring, advising, time management, etc.) on the basis of model predictions according to institutional policy and capabilities. (Agrusti et al., 2019; Althibyani, 2024). To consider the ethical, privacy, and fairness-related aspects of developing predictive models for student performance.

Hypothesis

Drawing on prior empirical findings in EDM and learning analytics, the study formulates the following hypotheses:

- **H1** (*Predictive value of multi-source data*). Machine learning models trained on integrated demographic, prior achievement, and LMS engagement data will significantly outperform baseline models that rely solely on prior GPA or a limited set of static variables, as measured by AUC and F1-score.

- **H2** (*Superiority of ensemble and non-linear models*). Ensemble methods (e.g., *random forests, gradient boosting*) and selected neural network architectures will achieve higher predictive performance than linear models (e.g., *logistic regression*) and simple decision trees for classifying students into performance or risk categories.

- **H3** (*Added value of behavioral engagement features*). Behavioral and temporal engagement features derived from LMS data—such as login frequency, recency of activity, assignment submission timeliness, and forum interactions—will contribute significantly to model performance beyond what can be achieved using demographic and prior achievement variables alone.

- **H4** (*Consistency of key predictors*). Across different algorithms and cohorts, a stable subset of predictors (e.g., *prior GPA, timely assignment submission, sustained LMS engagement*) will consistently emerge as highly influential in determining academic success, as indicated by feature importance rankings and interpretability analyses.

Methodology

Research Design

The study adopts a quantitative, non-experimental, explanatory research design. Historical institutional data will be used to model the relationship between predictor variables and academic outcomes without manipulating any instructional or environmental conditions. This design mirrors common approaches in *EDM and learning analytics*, where observational data from past cohorts serve as the basis for building and validating predictive models.

Supervised machine learning methods will be applied to construct models that predict end-of-semester academic performance. The study is explanatory in that, beyond forecasting outcomes, it seeks to identify and interpret the factors most strongly associated with success and risk.

Population and Sampling

Target population is the students attending the university of interest in various academic years. Sampling frame would comprise of all students for whom data is available for the selected semesters and courses. Stratified sampling would be used for the study to guarantee that there would be representation of the population from different faculties, different academic years, and other relevant factors such as admission categories (regular, special entry etc.).

Sample sizes have varied greatly in recent EDM and dropout prediction research from hundreds to tens of thousands of students; larger sample sizes can help in better estimation of models and also in performing detailed sub-group analysis. The study of interest will be targeted towards achieving a minimum effective sample size of 1,000 to 2,000 students.

Data Sources and Variables

Data sources would include the following:

1. **Demographic factors (Student Information System):** Age, Gender, Admission Type, Residency, and Socio-economic Status (Status in financial aid program).
2. **Measures of previous academic performance:** High school or entry test scores, Initial GPA at the beginning of the semester, Failure Rate, and Academic Standing.
3. **Measures of course performance:** Test scores, Mid-term exam scores, and Scores for Tasks and Quizzes during the semester.

Learning Management System (Moodle) use: Frequency of weekly logins, Time spent in course pages, Number of on-time submissions of assignments, Use of Forums, and Frequency of passing quizzes and engaging with learning prior to deadlines.

Measures of utilization of support services (if any): Frequency of visits to the tutoring center and scheduling an appointment with an academic counselor or another support service from the institution. Data can be aggregated.

Dependent variable (Outcome variable): Academic Achievement at the end of the semester is a variable with two measures as follows:

Continuous: Final Course Score/GPA at the end of the semester.

4. **Categorical:** Level of achievement: "High", "Medium", "Low"; "At Risk"/"Not At Risk". According to institutional criteria (GPA < 2.0).

Data Preprocessing

Data preprocessing would include the following EDM standard procedures:

- **Data cleaning:** elimination of duplicate data, handling of inconsistent identifiers, and treatment of missing data (*for example, imputation of occasional missing data and deletion of extensive missing data*).

- **Feature encoding:** coding of categorical variables (such as the field of study, admission type) to numeric form by techniques like one-hot encoding.

- **Feature scaling:** normalization and standardization of the features depending on the algorithms needed (*such as SVM, K-nearest neighbors*).

- **Temporal summarization:** creation of time-based features using LMS logs (*such as weekly counts of activities, early-semester indicators of involvement*).[7]

- **Class imbalance handling:** in case the number of at-risk students is very low, class imbalance handling techniques such as Synthetic Minority Over-sampling Technique (SMOTE), class weighting of loss functions can be used to prevent bias towards the majority class.

Predictive Methods and Tools

The research will explore various supervised learning methods used to predict students' academic performance:

- Logistic regression (*for classification problems*) and linear regression (*to predict students' GPA*).
- Decision trees as models based on rules that are easy to interpret.
- Random forest and gradient boosting algorithm, as ensemble methods based on decision trees, known for their effectiveness and reliability.
- Support vector machines with suitable kernels for non-linear decision boundaries.

Feed-forward neural networks with one or more hidden layers, where sample size and number of features make their use justified. Models will be developed using open-source platforms like Python (scikit-learn, XGBoost, TensorFlow/PyTorch for neural networks), or respective R platforms.

These platforms are in line with the existing EDM research practices where such libraries are commonly employed for experiments and practical applications.

Hyper-parameters tuning will be done using grid search or random search during cross-validation procedure.

Model Evaluation

Data will be partitioned into training and test sets (e.g., 70/30 or 80/20 split), stratified by outcome class where applicable. Within the training set, k-fold cross-validation ($k = 5$ or 10) will be used both to optimize hyper-parameters and to obtain robust estimates of model performance.[8]

Key evaluation metrics for classification tasks will include:

- Accuracy (overall correctness).
- Precision, recall, and F1-score, with a particular focus on recall and F1 for the at-risk class.
- Area under the ROC curve (AUC) to summarize discrimination ability.
- Confusion matrices to analyze error patterns, especially false negatives (at-risk students misclassified as not at risk).

For regression tasks (e.g., continuous GPA prediction), metrics such as mean absolute error (MAE), root mean squared error (RMSE), and coefficient of determination (R^2) will be computed.

Interpretability and Fairness Analysis

To support explanatory aims and ethical deployment, the study will conduct systematic interpretability and fairness analyses:

- Feature importance will be extracted from tree-based models (e.g., Gini importance, permutation importance) and, where appropriate, from linear models (standardized coefficients).
- Model-agnostic methods, such as SHAP values and partial dependence plots, will be used to visualize how changes in key features affect predicted outcomes, both globally and for illustrative individual cases.^[9]
- Subgroup analyses will compare model performance and error rates across demographic groups (e.g., gender, socio-economic status where available) to detect potential disparities that could indicate algorithmic bias.^[10]

Ethical and Privacy Considerations

The study will seek approval from the relevant institutional ethics committee. All data will be anonymized or pseudonymized prior to analysis, with direct identifiers removed and access restricted to authorized research personnel. Data handling will comply with applicable data protection regulations and institutional policies.

Information about the use of predictive analytics for educational purposes should be communicated transparently to stakeholders, including the intended uses of models, limitations of predictions, and mechanisms for contesting or contextualizing risk labels. The study will discuss these issues and propose principles for responsible governance of predictive analytics in higher education.

Analysis and Data Analysis

Exploratory Data Analysis

Initial analysis will focus on understanding the structure and distribution of the data:

- Descriptive statistics (means, standard deviations, medians, and frequency distributions) will be computed for all predictor and outcome variables.
- Visualization techniques such as histograms, boxplots, and density plots will be used to examine distributions, identify outliers, and assess the need for transformation.
- Correlation analyses (Pearson or Spearman, as appropriate) and heatmaps will explore associations among predictors and between predictors and outcomes, identifying potential multicollinearity and promising features.

These steps will inform feature selection and engineering decisions, such as combining highly correlated variables or introducing interaction terms where theoretically justified.

Model Training and Comparison

Each candidate algorithm will be trained on the preprocessed training data under comparable conditions. Cross-validation results will be used to select the best-performing configuration for each model type.

Comparative analysis will then be conducted on the hold-out test set, focusing on:

- Differences in predictive performance across models (e.g., which algorithms yield the highest AUC and F1-score for at-risk classification).
- Trade-offs between accuracy and interpretability (e.g., simpler but less accurate models vs. complex ensembles and neural networks).

The analysis will report both aggregate metrics and performance broken down by course, semester, and key demographic subgroups, providing a nuanced picture of model behavior.

Feature Importance and Interpretability

For the selected best-performing models, detailed interpretability analyses will be conducted:

- Global feature importance rankings will highlight which variables most strongly influence predictions overall (e.g., prior GPA, cumulative LMS activity, timeliness of submissions).
- Partial dependence plots will illustrate how predicted risk changes as a function of specific features while holding others constant (e.g., how predicted probability of failure changes with increasing number of late submissions).
- SHAP value analyses will provide both global and local explanations, identifying for individual students which features contributed most to their predicted risk level.

These results will be synthesized into a set of interpretable findings that can guide instructors, advisors, and policy-makers (e.g., emphasizing the importance of early, consistent engagement and timely completion of tasks).

Robustness and Sensitivity Analyses

To assess robustness, several additional analyses will be performed

- **Temporal validation:** training models on earlier cohorts and testing on later cohorts to assess performance stability over time.
- **Program-specific validation:** evaluating models separately for different faculties or programs to examine whether performance generalizes across disciplines.
- **Sensitivity to preprocessing choices:** comparing results across alternative approaches to handling missing data, class imbalance, and feature scaling.

Where feasible, the study will also explore simplified models using a reduced set of top predictors, assessing how much predictive performance is lost relative to full models and whether such simplification enhances interpretability and deployability.

Linking Predictions to Interventions

Although the emphasis of the analysis is the development and evaluation of models, an important consideration is the establishment of an intervention framework from predictive results. By examining existing research where predictive analytics are included in either learning management system dashboards or institution-wide early warning systems, (Althibyani, 2024; Predicting Student Performance and Enhancing Learning... - MDPI, 2025),

The analysis will discuss the implementation of action plans based upon predicted risk scores including the following:

- Automatic notifications for instructors and academic advisors once student risk reaches certain thresholds.
- Feedback notifications suggesting desirable behavior for improved performance (such as attending weekly classes, working on
- Invitations to tutoring or study skills classes or advising meetings for high-risk students.

This analysis will include considerations on how these thresholds and actions can be determined while balancing sensitivity (the ability to detect as many at-risk students as possible) with the possibility of "alarm fatigue."

Conclusions

Educational data-driven prediction is a field which has shown significant potential in terms of increasing the quality of education and making educational organizations more effective. According to recent systematic reviews and surveys, educational machine learning algorithms can successfully predict students' grades, course completion rates, and risk of dropout especially when they take advantage of multi-source data which include demographics, previous achievements, and LMS usage data. At the same time, the existing research on the topic also highlights several weaknesses in the studies conducted in the field of predictive analytics such as lack of consistent methodology, external validation, insufficient interpretability and inadequate ethical frameworks for the use of predictions in the educational setting. In order to address these shortcomings, further research needs to be done in both prediction and application of the findings to actual practice.

In particular, the study described in this paper can make an important contribution to this end by introducing a detailed specification of a framework for predicting students' academic performance based on institutional data. Multi-algorithm comparison, robust evaluation protocol, and systematic interpretability and fairness analysis are among its main features along with implementation into early warning and intervention strategies.

Furthermore, the design of this framework takes into account current evidence from EDM and learning analytics, as well as ethical and governance issues.

Recommendations

Based on the synthesized literature and the proposed methodological framework, the following recommendations are advanced for researchers and practitioners:

1. Integrate heterogeneous data sources. Institutions should systematically combine demographic, academic history, LMS engagement, and, where appropriate, support services data to build richer predictive models, rather than relying solely on prior grades or basic background information.
2. Adopt rigorous evaluation standards. Studies should employ clear train-test splits, cross-validation, multi-metric evaluation, and, where possible, temporal and cross-program validation to assess model generalizability. Reporting should include detailed descriptions of datasets, preprocessing, and evaluation procedures.
3. Balance accuracy with interpretability. When selecting algorithms for deployment, practitioners should consider both predictive performance and the availability of interpretable explanations (e.g., feature importance rankings, SHAP analyses) that stakeholders can understand and act upon.
4. Embed models in early warning and support workflows. Predictive models should be integrated into institutional processes via dashboards, alert systems, and advising tools that translate risk scores into timely, targeted interventions, rather than functioning as standalone analytical exercises.
5. Implement robust ethical governance. Institutions should develop and enforce policies on data access, consent, transparency, and fairness monitoring. Regular audits of model performance across demographic subgroups and mechanisms for stakeholder feedback are essential to prevent and mitigate algorithmic bias.
6. Foster data literacy among staff and students. Effective and responsible use of predictive analytics requires that instructors, advisors, and administrators understand the capabilities and limitations of models. Training and clear communication can help ensure that predictions are interpreted cautiously and constructively. (Agrusti et al., 2019; Althibyani, 2024)
7. Encourage replication and cross-institutional collaboration. Sharing anonymized datasets, code, and methodological details—where permitted—can facilitate replication studies and cross-institutional comparisons, strengthening the evidence base for predictive analytics in education.

Acknowledgement

I would like to express my sincere gratitude to everyone who contributed to the completion of this research paper titled "Predictive Analytics for Student Academic Performance Using Educational Data." I am deeply thankful to my faculty members and academic mentors for their valuable guidance, encouragement, and constructive suggestions throughout the study.

Financial support and sponsorship

Nil.

Conflicts of interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

References

1. Batool, S., Rashid, J., Nisar, M. W., Kim, J., Kwon, H.-Y., & Hussain, A. (2023). Educational data mining to predict students' academic performance: A survey study. *Education and Information Technologies*, 28(1), 905–971. <https://doi.org/10.1007/s10639-022-11152-y>
2. Agrusti, F., Bonavolontà, G., & Mezzini, M. (2019). University dropout prediction through educational data mining techniques: A systematic review. *Journal of e-Learning and Knowledge Society*, 15(3), 161–182. <https://doi.org/10.20368/1971-8829/1135017>
3. Angeioplastis, A., Aliprantis, J., Konstantakis, M., & Tsimpiris, A. (2025). Predicting student performance and enhancing learning outcomes: A data-driven approach using educational data mining techniques. *Computers*, 14(3), Article 83. <https://doi.org/10.3390/computers14030083>
4. Althibyani, H. A. (2024). Predicting student success in MOOCs: A comprehensive analysis using machine learning models. *PeerJ Computer Science*, 10, e2221. <https://doi.org/10.7717/peerj-cs.2221>
5. Islam, M. M., Sojib, F. H., Mihad, M. F. H., Hasan, M., & Rahman, M. (2025). The integration of explainable AI in educational data mining for student academic performance prediction and support system. *Telematics and Informatics Reports*, 18, 100203. <https://doi.org/10.1016/j.teler.2025.100203>
6. Kesgin, K., Kiraz, S., Kosunalp, S., & Stoycheva, B. (2025). Beyond Performance: Explaining and Ensuring Fairness in Student Academic Performance Prediction with Machine Learning. *Applied Sciences*, 15(15), 8409. <https://doi.org/10.3390/app15158409>
7. Johora, F. T., Hasan, M. N., Rajbongshi, A., Ashrafuzzaman, M., & Akter, F. (2025). An explainable AI-based approach for predicting undergraduate students academic performance. *Array*, 26, 100384. <https://doi.org/10.1016/j.array.2025.100384>
8. Guevara-Reyes, R., Ortiz-Garcés, I., Andrade, R., Cox-Riquetti, F., & Villegas-Ch, W. (2025). Machine learning models for academic performance prediction: interpretability and application in educational decision-making. *Frontiers in Education*, 10. <https://doi.org/10.3389/feduc.2025.1632315>
9. Hou, N., & Chen, H. (2026). Enhancing Fairness and Explainability in Student Performance Prediction Using Bias Mitigation Techniques. *Information Technology and Control*, 55(1), 143–165. <https://doi.org/10.5755/j01.itc.55.1.39399>
10. El Jihaoui, M., Abra, O. E. K., & Mansouri, K. (2026). Towards a fair and transparent early warning system for identifying students at risk of academic failure. *Frontiers in Education*, 11, 1878686. <https://doi.org/10.3389/feduc.2026.1878686>