The Journal of Asia Entrepreneurship and Sustainability RESEARCH ARTICLE

A Review of Machine Learning Techniques for Student Performance and Quality Enhancement in Higher Education

Mrs. Arati Patil*[1] , Mrs. Poonam Siddhanurle [2] Dr Prashant P Patil[3]

[1*]Research Scholar, Bharati Vidyapeeth Deemed To Be University, Pune
Email id: aratijadhavpatil@gmail.com
[2]Assistant Professor at D Y Patil Institute of MCA and Management Akurdi Pune.
Email id: poonamsiddhanurle@gmail.com
[3]Research Guide, Department Of Computer Applications, Bharati Vidyapeeth Deemed To Be University, Pune
Email id: Prashant.Patil@bharatividyapeeth.edu

Abstract
Higher education faces ongoing challenges in boosting student success rates and institutional quality amid vast data from learning management systems (LMS), assessments, and student interactions. Machine learning (ML) has become a cornerstone for transforming this data into predictive insights, enabling early detection of struggling students, personalized support, and strategic quality improvements. This systematic literature review synthesizes empirical studies (primarily 2015–2025) on ML for performance prediction and quality enhancement. It categorizes key methods classification, regression, clustering, and ensembles while evaluating their applications, advantages, and drawbacks, with special attention to skill-focused professional degrees like Master of Computer Applications (MCA), where coding proficiency, projects, and ongoing evaluations are pivotal. Recent trends show tree-based ensembles (e.g., Random Forest, XGBoost) achieving superior accuracy (often 85–95%) in predicting GPA and dropout risk, though interpretability issues persist. Major gaps include limited domain-specific models for professional curricula, scarce real-time systems, underutilization of explainable AI (XAI), and insufficient cross-institutional validation. Proposed future paths involve hybrid models, deep learning for multimodal data, XAI integration for transparency, and tailored analytics for programs like MCA to promote ethical, scalable, and impactful educational advancements.

Keywords: Machine learning (ML), learning management systems (LMS), Master of Computer Applications(MCA)

1. Introduction

Higher education institutions worldwide are under increasing pressure to improve student success rates, reduce dropout levels, and enhance overall institutional quality. These priorities have become even more critical in recent years, as global enrolment continues to rise while retention challenges persist. For instance, dropout rates in higher education often range from 20–40% depending on the region and institution type, leading to significant personal, economic, and societal costs such as lost tuition revenue, underutilized resources, and reduced graduate contributions to the workforce. In India, where the higher education system is one of the largest globally with millions of students enrolled across diverse programs, these issues are particularly pronounced. Challenges include high dropout rates (especially in the early semesters), uneven access to quality education (particularly in rural or underserved areas), faculty shortages, outdated curricula, inadequate infrastructure, and limited integration of modern teaching tools. Institutions often struggle with low student engagement, socioeconomic disparities affecting persistence, and difficulties in providing timely interventions for at-risk learners. Globally, similar patterns emerge: factors like financial pressures, academic unpreparedness, lack of support services, and mismatched expectations contribute to attrition. Recent studies highlight that early identification of struggling students can reduce dropout by enabling proactive measures, such as personalized tutoring or adaptive learning pathways. This systematic literature review synthesizes empirical research from 2015–2025 on ML applications for student performance prediction and quality improvement in higher education. It examines widely used techniques classification (e.g., for pass/fail or risk categorization),
Regression (e.g., for GPA forecasting), clustering (e.g., for student segmentation), and ensembles (e.g., for superior accuracy) while highlighting their strengths, limitations, and real- world applications. Particular emphasis is placed on professional programs like MCA, where technical and skill-oriented factors play a pivotal role in success. By addressing these elements, the review identifies persistent gaps (e.g., real-time analytics, XAI integration, and program- tailored models) and proposes future directions toward ethical, scalable ML solutions that empower educators, support students, and elevate institutional quality in an increasingly data- rich educational landscape.

2. Methodology

This systematic literature review (SLR) adopts a structured, transparent, and replicable protocol

inspired by the PRISMA 2020 guidelines (Page et al., 2021) and established SLR frameworks in computing/education domains (e.g., Kitchenham & Charters, 2007; Okoli, 2015). These standards ensure minimization of selection bias, comprehensive coverage, and clear reporting of methods. The process includes planning (protocol definition and research questions), execution (search, screening, extraction), and synthesis/reporting.

2.1 Systematic Literature Review (SLR) Protocol

The SLR protocol was predefined to guide the entire process and enhance reproducibility. The review is driven by the following focused research questions:

- **RQ1**: Which machine learning techniques (e.g., classification, regression, clustering, and ensemble methods) are most frequently applied for predicting student performance and supporting quality enhancement in higher education?
- **RQ2**: What types of datasets, features (academic, demographic, behavioural, LMS-derived), and evaluation metrics dominate these studies?
- **RQ3**: What are the reported strengths, limitations, and comparative performance of these ML approaches, with emphasis on applications in professional programs (e.g., MCA, engineering)?
- **RQ4**: What key research gaps (e.g., dataset diversity, model interpretability, real-time capabilities, and program-specific adaptations) persist, and what future directions emerge?

These RQs follow a PICOC framework (Population: higher education students/institutions; Intervention: ML techniques; Comparison: different ML categories; Outcomes: prediction accuracy, quality insights; Context: higher education, professional programs).

2.2 Data Extraction and Analysis

A standardized data extraction template (Excel-based) was used to systematically capture relevant information from each eligible study, ensuring consistency and traceability.

Extracted Elements:

- Bibliographic information: Authors, year, title, publication venue, DOI/link.
- ML techniques: Specific algorithms (e.g., Logistic Regression, Decision Trees, SVM, k-Means, Random Fore
- Dataset and features: Source (e.g., institutional LMS, open datasets like Kaggle), size/sample, feature types (academic: prior grades, attendance; demographic: age, gender, socioeconomic; behavioral: LMS logs, engagement time, assignment patterns; program-specific: coding logs for MCA).
- Performance metrics: Classification (accuracy, precision, recall, F1-score, AUC-ROC); Regression (MAE, RMSE, R^2); Clustering (Silhouette score, Davies-Bouldin); validation method (e.g., k-fold cross-validation).
- Application domain: General higher education vs. professional programs (e.g., MCA emphasis on technical skills).
- Key findings: Predictive accuracy, influential features, comparative advantages.
- Limitations and gaps: Issues like data quality, interpretability, real-time feasibility, bias, or single-institution focus.

Quality Assessment: An informal appraisal considered aspects such as dataset size (>500 samples preferred), use of cross-validation, handling of class imbalance, reporting of limitations, and ethical considerations (e.g., privacy). Studies with severe flaws were excluded during full-text review.

Data Synthesis and Analysis:

- Quantitative: Frequency counts (e.g., % of studies using ensembles), average performance metrics across categories, tabulated comparisons.
- Qualitative: Thematic narrative synthesis grouping trends (e.g., ensembles outperform singles), strengths/limitations, and gap identification (e.g., limited MCA-specific models).
- No formal meta-analysis due to heterogeneity in datasets/metrics; instead, comparative tables highlight efficacy and suitability.

This protocol ensures a balanced, evidence-based synthesis of ML applications in student performance prediction and quality enhancement, with targeted insights for professional programs like MCA.

3. Machine Learning Techniques in Education

Machine learning (ML) has become a cornerstone in educational data mining (EDM) and learning analytics, enabling institutions to transform vast amounts of student data such as grades, attendance, LMS interactions, demographics, and behavioural logs into predictive and actionable insights. These techniques support early identification of at-risk students, personalized interventions, tailored learning pathways, and overall quality enhancement in higher education. This section categorizes and elaborates on the primary ML paradigms applied: **supervised** (classification and regression for prediction), **unsupervised** (clustering for pattern discovery and segmentation), and **ensemble** methods (for improved robustness and accuracy). Recent systematic reviews and

empirical studies (2020–2025) highlight that ensembles (especially tree-based like Random Forest and XGBoost) dominate due to superior performance on imbalanced educational datasets, while clustering remains valuable for student profiling despite lower frequency in predictive tasks.

3.1 Classification Methods

Classification is a supervised learning approach that predicts discrete categorical outcomes, making it ideal for binary or multi-class problems in education, such as labelling students as "pass/fail," "low/medium/high risk," or categorizing engagement levels.

Popular Techniques:

- **Logistic Regression**: A probabilistic linear model that outputs probabilities for class membership; serves as a strong baseline due to its simplicity and interpretability.
- **Decision Trees**: Rule-based models that recursively split data based on feature thresholds, producing interpretable "if-then" rules.
- **Support Vector Machines (SVM)**: Finds an optimal hyperplane to separate classes, effective in high-dimensional spaces with clear margins.
- **Naïve Bayes**: A probabilistic classifier assuming feature independence; computationally efficient and performs well with categorical data.
- **K-Nearest Neighbors (k-NN)**: Instance-based lazy learning that assigns labels based on majority vote among the k closest training examples.

Applications:

- Dropout risk prediction (early flagging of vulnerable students for interventions).
- Binary/multiclass performance outcomes (e.g., pass/fail in courses).
- Classification of learning styles or engagement levels (e.g., active vs. passive learners based on LMS logs).

3.2 Regression Models

Regression is a supervised technique for predicting continuous numerical outcomes, such as final GPA, exam scores, or engagement metrics on a scale.

Examples:

- **Linear Regression**: Models linear relationships between features and the target; simple and interpretable but assumes linearity.
- **Support Vector Regression (SVR)**: Extends SVM to regression, using epsilon- insensitive loss to handle nonlinearities via kernels.
- **Regression Trees** (and ensembles like Gradient Boosting Regressors): Tree-based methods that partition data and fit constants in leaves, capturing complex interactions.

Insights:

- Linear Regression is straightforward, computationally efficient, and serves as a reliable baseline, but it often fails to capture nonlinear relationships common in educational data (e.g., diminishing returns from extra study time).
- SVR and tree-based regression excel at modeling nonlinearities and interactions, with tree methods particularly robust to outliers and mixed data types.
- Regression models are frequently used as baselines before advancing to ensembles, with recent studies showing tree-based regressors reducing RMSE by 10–20% over linear models.

Applications:

- Forecasting cumulative GPA or semester-end scores.
- Estimating performance after formative assessments or midterms to guide timely support.
- Predicting continuous engagement scores (e.g., hours logged in LMS).

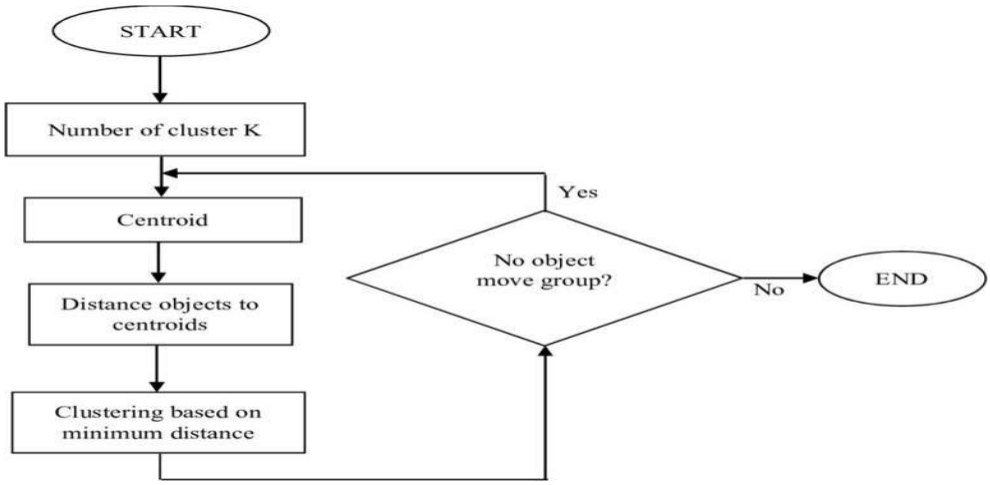
3.3 Clustering Techniques

Clustering is an unsupervised learning method that groups similar data points without predefined labels, revealing inherent structures in student data for segmentation and profiling.

Common Algorithms:

K-Means: Partition-based; iteratively assigns points to k centroids and updates centroids to minimize intra-cluster variance; requires specifying k (often via elbow method or silhouette score). The K-means clustering algorithm is one of the division-based clustering algorithms.

Figure 3.1: Flowchart/Overview of K-Means



Hierarchical Clustering: Builds a tree (dendrogram) of nested clusters; agglomerative (bottom-up, merging closest clusters) or divisive (top-down); no need to predefine k. Hierarchical clustering is a type of cluster analysis that builds a hierarchy of data points as they move into or out of a cluster.
There are two main types of strategies for this algorithm:

Agglomerative: With this clustering algorithm, we don't need to know ahead of time how many clusters there will be. Bottom-up algorithms start by treating each piece of data as a single cluster. They then group pairs of clusters together until all of the clusters have been combined into a single cluster that contains all of the data.

Divisive - another name for "top-down" method. A fixed number of clusters is not needed to run this algorithm. To perform top-down cluster analysis, one must first devise a strategy for decomposing a cluster that contains all of the data, and then continue decomposing clusters in a recursive fashion until each piece of data has been reduced to a standalone cluster.

DBSCAN: Density-based; identifies clusters as dense regions separated by low-density areas, handling noise and arbitrary shapes effectively. This commonsense understanding of clusters and noise forms the foundation of the DBSCAN algorithm. Machine learning researchers can utilize DBSCAN, a clustering algorithm, to distinguish between dense and sparse clusters. The essential principle is that there must be at least some minimal number of points within a specific radius of each point in a cluster.

5. Research Gaps

Despite substantial progress in applying machine learning (ML) to student performance prediction and educational quality enhancement, several persistent gaps limit the field's maturity and real-world impact. These gaps are particularly pronounced in professional programs like Master of Computer Applications (MCA), engineering, and other skill-intensive curricula, where unique data sources (e.g., coding repositories, project artifacts, debugging logs, peer assessments) remain underexplored compared to general academic datasets. Recent systematic reviews and empirical studies (2023–2025) consistently highlight methodological, practical, ethical, and contextual shortcomings that hinder scalable, equitable, and trustworthy adoption in higher education institutions. This section elaborates on the major gaps, drawing from trends in educational data mining (EDM), learning analytics (LA), and related fields, while emphasizing implications for professional programs.

6. Future Directions

The rapid evolution of machine learning (ML) in educational data mining (EDM) and learning analytics opens promising avenues to overcome current limitations and deliver more impactful, ethical, and scalable solutions for student performance prediction and quality enhancement in higher education. Building on identified research gaps such as limited domain-specific models, static data reliance, lack of explainability, and poor integration with institutional frameworks future research should prioritize innovative, interdisciplinary approaches. These directions emphasize hybrid intelligence, advanced deep learning, transparency via explainable AI (XAI), real-time capabilities, and tailored applications for professional programs like Master of Computer Applications (MCA).

7. Conclusion

Machine learning has demonstrated substantial promise in predicting student performance and advancing educational quality in higher education. Classification techniques enable effective risk categorization, regression supports precise outcome forecasting, clustering reveals valuable

student profiles for segmentation, and ensemble methods (e.g., Random Forest and XGBoost) consistently deliver superior accuracy and robustness by mitigating bias and variance. Despite these strengths, notable gaps remain particularly in applying ML to professional programs such as MCA, where programming-intensive data (e.g., code submissions, project evaluations, and capstone metrics) remain underexplored. Additional limitations include over-reliance on static datasets, limited use of explainable AI (XAI) for transparency, single-institution bias, and weak integration with quality assurance frameworks like accreditation or outcome-based education standards. Addressing these gaps through domain-specific datasets, XAI-enhanced models, real-time streaming analytics from LMS and coding platforms, hybrid clustering-predictive approaches, and cross-institutional validations will enable more timely, ethical, and effective interventions. This will empower educators to provide personalized support, reduce dropout rates, enhance skill mastery in technical programs, and align institutional strategies with measurable quality improvements. Ultimately, by bridging these gaps with interdisciplinary collaboration and ethical focus, ML can transition from research prototypes to embedded tools that foster inclusive, adaptive, and high- performing higher education ecosystems especially in professional fields like MCA, where data-driven insights can directly boost technical competency and graduate employability.

8. References

1. Albreiki, B., Zaki, N., & Alashwal, H. (2021). A systematic literature review of student performance prediction using machine learning techniques. Education Sciences, 11(9), Article 552. https://doi.org/10.3390/educsci11090552
2. Namoun, A., & Alshanqiti, A. (2021). Predicting student performance using data mining and learning analytics techniques: A systematic literature review. Applied Sciences, 11(1), Article 237. https://doi.org/10.3390/app11010237
3. Kabathova, J., & Drlik, M. (2021). Towards predicting student's dropout in university courses using different machine learning techniques. Applied Sciences, 11(7), Article 3130. https://doi.org/10.3390/app11073130
4. Park, H. S., & Yoo, S. J. (2021). Early dropout prediction in online learning of university using machine learning. JOIV: International Journal on Informatics Visualization, 5(4), 347–353. https://doi.org/10.30630/joiv.5.4.732
5. Niyogisubizo, J., Liao, L., Nziyumva, E., Murwanashyaka, E., & Nshimyumukiza, P. C. (2022). Predicting student's dropout in university classes using two-layer ensemble machine learning approach: A novel stacked generalization. Computers and Education: Artificial Intelligence, 3, Article 100066. https://doi.org/10.1016/j.caeai.2022.100066
6. Sekeroglu, B., Abiyev, R., Ilhan, A., Arslan, M., & Idoko, J. B. (2021). Systematic literature review on machine learning and student performance prediction: Critical gaps and possible remedies. Applied Sciences, 11(22), Article 10907. https://doi.org/10.3390/app112210907
7. Alamri, R., & Alharbi, B. (2021). Explainable student performance prediction models: A systematic review. Computers and Education: Artificial Intelligence, 2, Article 100016.
8. Feng, G., Fan, M., & Chen, Y. (2022). Analysis and prediction of students' academic performance based on educational data mining. IEEE Access, 10, 19558– 19571. https://doi.org/10.1109/ACCESS.2022.3151234
9. Martins, M. V., Tolledo, D., Machado, J., Baptista, L. M., & Realinho, V. (2023). Early prediction of student's performance in higher education: A case study. In Á. Rocha et al. (Eds.), Trends and applications in information systems and technologies (Vol. 9, pp. 1–10). Springer. https://doi.org/10.1007/978-3-031-36957-5_1
10. Ahmed, W. (2024). Student performance prediction using machine learning algorithms. Applied Computational Intelligence and Soft Computing, 2024, Article 4067721. https://doi.org/10.1155/2024/4067721
11. Rebelo Marcolino, M., et al. (2025). Student dropout prediction through machine learning optimization: Insights from Moodle log data. Scientific Reports, 15, Article 9840. https://doi.org/10.1038/s41598-025-93918-1
12. Wang, Y. (2025). Artificial intelligence in student management systems to enhance academic performance monitoring and intervention. Scientific Reports, 15,
13. Article 35122. https://doi.org/10.1038/s41598-025-19159-4
14. Islam, M. M., Sojib, F. H., Mihad, M. F. H., Hasan, M. 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. https://doi.org/10.1016/j.teler.2025.100018
15. Turkmen, G. (2025). The review of studies on explainable artificial intelligence in educational research. Journal of Educational Computing Research.

https://doi.org/10.1177/07356331241310915
16. Balcioğlu, Y. S., & Artar, M. (2025). Predicting academic performance of students with machine learning. Information Development. https://doi.org/10.1177/02666669231213023
17. Turkmenbayev, A., Abdykerimova, E., Nurgozhayev, S., Karabassova, G., & Baigozhanova, D. (2025). The application of machine learning in predicting student performance in university engineering programs: A rapid review. Frontiers in Education, 10, Article 1562586. https://doi.org/10.3389/feduc.2025.1562586
18. Ahmed, W., Wani, M. A., Plawiak, P., et al. (2025). Machine learning-based academic performance prediction with explainability for enhanced decision-making in educational institutions. Scientific Reports, 15, Article 26879. https://doi.org/10.1038/s41598-025-12353-4
19. Chong, K. T., Ibrahim, N., Huspi, S. H., Wan Kadir, W. H. N., & Isa, M. A. (2025). A systematic review of machine learning techniques for predicting student engagement in higher education online learning. Journal of Information Technology Education: Research, 24, Article 5. https://doi.org/10.28945/5456
20. Shi, et al. (2025). Applications of machine learning for at-risk student prediction in online education: A 10-year systematic review of literature. Journal of Computer Assisted Learning. https://doi.org/10.1111/jcal.70058
21. Rabelo, A. M., & Zárate, L. E. (2025). A model for predicting dropout of higher education students. Data Science and Management, 8(1), 72–85. https://doi.org/10.1016/j.dsm.2024.07.001
22. Lyu, H., et al. (2025). Artificial intelligence for student performance prediction in blended learning: A systematic literature review. Neurocomputing. https://doi.org/10.1016/j.neucom.2025.1331
23. Yang, et al. (2025). Machine learning models for academic performance prediction: Interpretability and application in educational decision-making. Frontiers in Education. https://doi.org/10.3389/feduc.2025.1632315
24. Wang, J., et al. (2025). Machine learning approach to student performance prediction of online learning. PLoS ONE, 20(1), Article e0299018. https://doi.org/10.1371/journal.pone.0299018
25. Alamri, L. H., et al. (2020–2025 update; recent extension). Predicting student academic performance using support vector machine and random forest. In Proceedings of the 2020 3rd International Conference on Education Technology Management
26. Realinho, V., et al. (2021–2025 dataset). Predict students' dropout and academic success [Dataset]. UCI MachineLearning Repository. https://doi.org/10.24432/C5MC89