

Early Prediction of At-Risk MCA Students Using Machine Learning-Based Classification Models

Mrs Arati Patil¹, Dr Prashant P Patil², Dr S V Deshmukh³

¹Research Scholar, Bharati Vidyapeeth Deemed to Be University, Pune

²Assistant Professor, D Y Patil Institute of MCA and Management, Akurdi, Pune

³Research Guide, Department of Computer Applications, Bharati Vidyapeeth Deemed to Be University, Pune

³Research Expert, Department of Computer Applications, Bharati Vidyapeeth Deemed to Be University, Pune

Emails: aratijadhavpatil@gmail.com¹, Prashant.Patil@bharatividyaapeeth.edu²,
shital.deshmukh@bharatividyaapeeth.edu³

Abstract

Student academic performance prediction has become an important area of educational data mining due to its potential to improve learning outcomes through timely intervention. Early identification of academically at-risk students enables educational institutions to provide personalized academic support, thereby reducing failure rates and improving student retention. Master of Computer Applications (MCA) programmes present unique challenges because students come from diverse academic backgrounds and are expected to acquire advanced technical competencies within a limited duration. Consequently, identifying students who may experience academic difficulties at an early stage is essential for enhancing educational quality and student success. This paper proposes a machine learning-based methodology for the early prediction of MCA students' academic performance using multiclass classification. The proposed framework categorizes students into three performance groups: At-Risk, Average Performer, and High Performer. The methodology integrates student demographic information, academic history, attendance records, continuous assessment scores, laboratory performance, assignment completion, learning management system engagement, and classroom participation as predictive features. A systematic pre-processing pipeline consisting of data cleaning, feature encoding, normalization, missing value treatment, and feature selection is incorporated to improve data quality before model training. The proposed methodology recommends the implementation and comparative evaluation of several supervised machine learning algorithms, including Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, K-Nearest Neighbour, Naïve Bayes, Gradient Boosting, and Extreme Gradient Boosting (XGBoost). Model performance is assessed using Accuracy, Precision, Recall, F1-score, Confusion Matrix, and Receiver Operating Characteristic (ROC) analysis. Greater emphasis is placed on Precision, Recall, and F1-score since these measures provide more reliable evaluation in educational datasets where class imbalance frequently exists. The proposed framework is expected to support higher education institutions in establishing an intelligent early warning system capable of identifying academically vulnerable students before the final examinations. Such predictive systems can facilitate timely counselling, personalized mentoring, remedial instruction, and data-driven academic decision-making, ultimately contributing to improved student retention and academic excellence.

Keywords: Educational Data Mining, Learning Analytics, Academic Performance Prediction, Machine Learning, Student Classification, At-Risk Students, Supervised Learning, Precision, Recall, F1-score.

1. Introduction

The rapid digital transformation of higher education has significantly increased the availability of academic data generated through learning management systems, online assessments, attendance monitoring, digital libraries, laboratory activities, and student information systems. Educational institutions

now possess extensive datasets that can be analysed to understand student learning behaviour, identify academic challenges, and support evidence-based decision-making. The emergence of Educational Data Mining (EDM) and Learning Analytics (LA) has enabled researchers to extract meaningful

knowledge from educational data and develop intelligent systems capable of predicting student performance with considerable accuracy. Academic success is influenced by numerous interconnected factors including previous academic achievements, attendance, classroom engagement, assignment submission patterns, laboratory performance, socio-economic background, motivation, and learning behaviour. Traditional methods of identifying academically weak students primarily rely on semester-end examination results, which often delay interventions until it becomes difficult to improve student outcomes. Consequently, there is a growing need for predictive models capable of identifying at-risk students during the early stages of a semester. Master of Computer Applications (MCA) programmes present additional challenges due to their interdisciplinary curriculum that integrates programming, software engineering, cloud computing, artificial intelligence, cybersecurity, data analytics, and project-based learning. Students entering MCA programmes often possess diverse educational backgrounds, resulting in varying levels of programming knowledge and analytical skills. These differences can significantly affect academic performance, making early identification of struggling students particularly important. Recent advances in machine learning have provided powerful tools for analysing educational datasets. Supervised learning algorithms are capable of discovering hidden relationships among student attributes and accurately classifying students into predefined performance categories. Unlike conventional statistical methods, machine learning techniques can effectively model complex nonlinear relationships, handle multidimensional datasets, and continuously improve prediction accuracy as additional data become available. Although several studies have investigated student performance prediction, many focus primarily on binary classification (Pass/Fail) or regression-based grade prediction. Such approaches provide limited information regarding varying levels of academic performance. A multiclass classification framework categorizing students as At-Risk, Average Performer, and High Performer offers more informative insights

that enable institutions to design customized academic interventions for different categories of learners. Furthermore, model evaluation in educational prediction often relies heavily on overall classification accuracy. While accuracy provides a general measure of performance, it may produce misleading conclusions when class distributions are imbalanced. Metrics such as Precision, Recall, and F1-score provide a more comprehensive assessment of classification effectiveness, particularly for identifying at-risk students where minimizing false negatives is critically important. The methodology proposed in this paper presents a systematic framework for predicting MCA students' academic performance using multiple supervised machine learning algorithms. The framework emphasizes robust data pre-processing, comprehensive feature engineering, multiclass classification, and comparative evaluation using multiple performance metrics. The proposed methodology aims to contribute towards the development of intelligent academic early warning systems that facilitate timely interventions and improve student success.

2. Literature Review

Educational Data Mining has emerged as an active interdisciplinary research domain combining machine learning, statistics, artificial intelligence, and educational sciences. Numerous studies have demonstrated the effectiveness of predictive analytics in identifying students who may require additional academic support. Early investigations primarily employed statistical regression models to estimate students' final grades based on attendance and examination scores. Although these models provided useful insights, their ability to capture nonlinear relationships among educational variables remained limited. The introduction of machine learning techniques significantly enhanced predictive performance. Decision Tree algorithms became popular because of their interpretability and ability to generate easily understandable classification rules. Researchers observed that Decision Trees effectively identified influential factors such as attendance percentage, internal assessment scores, and assignment completion. Random Forest algorithms further improved prediction accuracy by combining

multiple decision trees through ensemble learning. The ensemble approach reduced overfitting while increasing generalization capability across diverse educational datasets. Several comparative studies reported Random Forest as one of the most reliable classifiers for student performance prediction due to its robustness against noisy data. Support Vector Machines demonstrated strong predictive capability, particularly when educational datasets contained complex nonlinear relationships. By maximizing class separation margins, SVM models frequently achieved high classification accuracy despite limited sample sizes. However, computational complexity and parameter optimization remain practical challenges. Naïve Bayes classifiers have also been extensively applied because of their computational efficiency and simplicity. Despite assuming conditional independence among features, Naïve Bayes often performs competitively in educational datasets where predictor variables exhibit moderate correlation. More recently, ensemble learning methods such as Gradient Boosting and Extreme Gradient Boosting (XGBoost) have attracted considerable attention. These techniques sequentially improve prediction performance by minimizing residual errors produced by previous models. Their ability to capture complex interactions among student attributes has resulted in superior classification performance across numerous educational prediction studies. Learning Analytics has further expanded predictive capabilities by incorporating digital learning behaviour collected through Learning Management Systems (LMS). Variables including login frequency, course material access, quiz participation, discussion forum activity, assignment submission timing, and online engagement have demonstrated significant predictive value when combined with traditional academic indicators. Researchers have also emphasized the importance of feature engineering. Combining demographic variables with behavioural and academic features generally improves model performance compared to using examination scores alone. Feature selection techniques including Information Gain, Chi-Square, Recursive Feature Elimination, and correlation analysis help eliminate redundant variables while

improving computational efficiency. Although substantial progress has been achieved, several limitations remain evident in existing literature. Many studies employ binary classification schemes that distinguish only between successful and unsuccessful students. Such classification provides limited guidance for personalized intervention strategies. Furthermore, relatively few studies specifically investigate postgraduate MCA programmes, despite their unique curriculum and student diversity. Another limitation concerns evaluation methodology. Numerous studies primarily report overall accuracy without adequately considering Precision, Recall, or F1-score. Since educational datasets frequently exhibit class imbalance, accuracy alone may overestimate classifier performance. Comprehensive evaluation using multiple performance metrics is therefore essential for practical implementation. These observations indicate a continuing need for robust multiclass classification frameworks specifically designed for MCA students, incorporating comprehensive pre-processing techniques, multiple machine learning algorithms, and balanced evaluation metrics[1 – 10].

3. Research Gap

Based on the reviewed literature, the following research gaps have been identified:

- Most previous studies focus on undergraduate education, with comparatively limited attention given to postgraduate MCA programmes.
- Existing research predominantly employs binary classification (Pass/Fail), which does not adequately differentiate varying academic performance levels.
- Several predictive models rely mainly on examination scores while overlooking behavioural, engagement, and learning analytics variables.
- Few studies perform comprehensive comparisons of multiple supervised learning algorithms within a common experimental framework[11 – 15].
- Many investigations evaluate models primarily using accuracy, neglecting

Precision, Recall, and F1-score, which provide better assessment under class imbalance.

- Limited research proposes a generalized methodological framework that can be adapted by higher education institutions for early academic intervention[16 – 20].

4. Research Objectives

The proposed study is designed to achieve the following objectives:

- To develop a machine learning-based framework for the early prediction of academic performance among MCA students.
- To classify students into three performance categories: At-Risk, Average Performer, and High Performer.
- To identify significant academic and behavioural factors influencing student performance.
- To compare the predictive performance of multiple supervised machine learning algorithms.
- To evaluate classification models using Accuracy, Precision, Recall, F1-score, Confusion Matrix, and ROC analysis.
- To provide a methodological foundation for implementing intelligent academic early warning systems in higher education institutions[21].

5. Proposed Research Methodology

5.1.Overview

The proposed methodology aims to develop a machine learning-based framework for the early identification of MCA students who are academically at risk. Unlike conventional approaches that rely solely on end-semester examination results, the proposed framework utilizes multiple academic, behavioural, and engagement-related variables collected during the semester. The methodology follows a systematic pipeline comprising data collection, preprocessing, feature engineering, multiclass classification, model evaluation, and interpretation. The framework is designed to support higher education institutions in implementing an early warning system that enables faculty members to identify students requiring academic intervention

before the completion of the semester. Figure 1 presents the overall workflow of the proposed methodology.

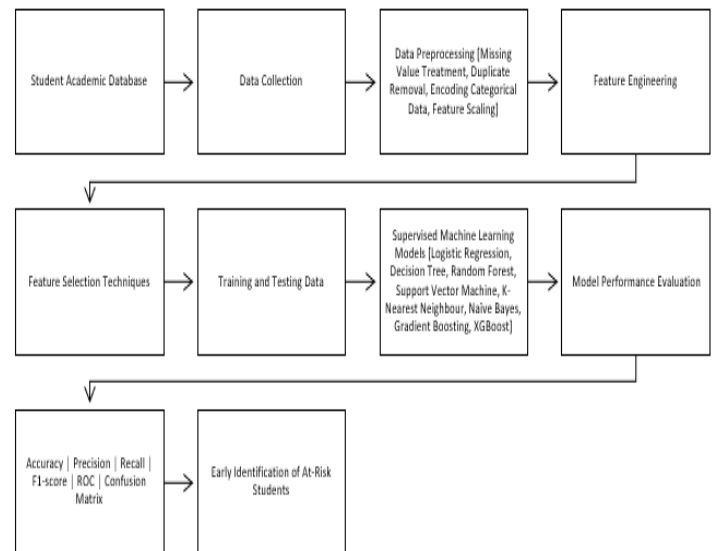


Figure 1 Proposed Machine Learning Framework for Early Prediction of Student Performance

5.2.Research Design

The study adopts a quantitative research design based on supervised machine learning techniques. Historical academic records of MCA students constitute the primary source of data. The methodology involves training classification algorithms using previously observed student performance data and evaluating their ability to classify new students into predefined performance categories.

The research workflow consists of six major phases:

- Data collection
- Data pre-processing
- Feature engineering
- Model development
- Model evaluation
- Prediction and interpretation

The proposed framework is generic and can be adapted by universities and higher education institutions irrespective of their academic management systems[22].

5.3.Dataset Description

The methodology assumes the availability of historical academic records of MCA students collected from institutional databases such as Student Information Systems (SIS), Learning Management Systems (LMS), examination records, attendance registers, and internal assessment reports. The dataset is expected to include three categories of variables[23].

5.3.1. Demographic Variables

- Gender
- Age
- Previous Degree
- Graduation Percentage
- Admission Category

5.3.2. Academic Variables

- Attendance Percentage
- Internal Assessment Marks
- Mid-Semester Examination Marks
- Practical Examination Scores
- Laboratory Performance
- Assignment Marks

- Seminar Performance
- Programming Test Scores
- Project Evaluation

5.3.3. Behavioural Variables

- LMS Login Frequency
- Assignment Submission Timeliness
- Online Quiz Participation
- Discussion Forum Participation
- Library Usage
- Classroom Participation

These variables collectively represent students' academic engagement and learning behaviour, enabling the machine learning models to identify complex relationships influencing academic performance[24].

5.4. Performance Categories

The proposed methodology adopts multiclass classification by dividing students into three categories based on their academic performance. As shown in Table 1 Student Performance Categories.

Table 1 Student Performance Categories

Category	Description
At-Risk	Students requiring immediate academic intervention due to poor academic performance or low engagement
Average Performer	Students demonstrating satisfactory academic progress but requiring continuous monitoring
High Performer	Students consistently achieving high academic performance and demonstrating strong engagement

This classification provides more actionable information than binary classification because it enables institutions to implement differentiated intervention strategies.

5.5. Data Pre - processing

Educational datasets frequently contain incomplete records, inconsistent formats, missing values, and

redundant information. Effective pre-processing is therefore essential before model development. The pre-processing stage consists of the following steps.

5.5.1. Data Cleaning

The collected dataset is examined to identify duplicate records, inconsistent entries, typographical

errors, and irrelevant attributes. Duplicate student records are removed to ensure data integrity.

5.5.2. Missing Value Treatment

Missing values may arise due to absent students, delayed assessment records, or incomplete data entry. Depending on the nature of the variable, missing values can be handled using:

- Mean imputation
- Median imputation
- Mode imputation
- Removal of records with excessive missing information

Appropriate treatment minimizes information loss while improving model reliability.

5.5.3. Encoding Categorical Variables

Machine learning algorithms require numerical input. Therefore, categorical variables such as gender, graduation stream, and admission category are transformed into numerical representations using:

- Label Encoding
- One-Hot Encoding

The choice of encoding depends on the specific classifier and the characteristics of the categorical variables.

5.5.4. Feature Scaling

Several classification algorithms are sensitive to differences in variable scales. Feature scaling ensures that all numerical variables contribute proportionately during model training.

Common scaling techniques include:

- Standardization (Z-score normalization)
- Min-Max Normalization

Scaling is particularly beneficial for algorithms such as Support Vector Machine and K-Nearest Neighbour.

5.5.5. Outlier Detection

Extreme values may adversely influence classification performance. Statistical methods such as the Interquartile Range (IQR), Z-score analysis, or boxplot analysis can be employed to identify potential outliers. Depending on their relevance, outliers may be retained, transformed, or removed.

5.6.Feature Engineering

Feature engineering plays a critical role in improving predictive performance. In addition to using raw

variables, new derived features can be created to capture meaningful academic patterns.

Examples include:

- Attendance Consistency Index
- Assignment Completion Ratio
- Continuous Assessment Average
- Practical-to-Theory Performance Ratio
- LMS Engagement Score
- Overall Academic Engagement Index

These derived variables provide richer representations of student learning behaviour.

5.7.Feature Selection

Not all variables contribute equally to prediction. Redundant and irrelevant features may reduce model performance and increase computational complexity. The proposed methodology recommends applying feature selection techniques such as:

- Correlation Analysis
- Information Gain
- Chi-Square Test
- Mutual Information
- Recursive Feature Elimination (RFE)
- Random Forest Feature Importance

Selecting the most informative features enhances model interpretability while reducing overfitting.

5.8.Data Partitioning

Following pre-processing and feature selection, the dataset is divided into training and testing subsets. The training dataset is used for learning the predictive model, while the testing dataset evaluates its generalization performance on previously unseen data. A commonly adopted strategy is an 80:20 or 70:30 train-test split. To further improve the robustness of the evaluation, k-fold cross-validation may be employed, wherein the dataset is partitioned into k subsets and the model is trained and validated iteratively. This approach provides a more reliable estimate of predictive performance by reducing the dependence on a single train-test split.

6. Machine Learning-Based Classification Models

To accurately identify students who are academically at risk, the proposed framework recommends evaluating multiple supervised machine learning algorithms. Since educational datasets exhibit varying characteristics such as nonlinear

relationships, class imbalance, and mixed data types, comparing different classification models helps identify the most suitable algorithm for early prediction.

6.1.Logistic Regression (LR)

Logistic Regression is a widely used statistical classification algorithm that estimates the probability of a student belonging to a particular performance category. Although originally developed for binary classification, it can be extended to multiclass classification using multinomial logistic regression. Logistic Regression provides interpretable coefficients that help identify the influence of individual predictor variables on academic performance.

Advantages

- Simple and computationally efficient
- Easy to interpret
- Suitable for baseline model comparison

Limitations

- Assumes a linear relationship between predictors and the outcome
- May underperform when relationships are highly nonlinear

6.2.Decision Tree (DT)

Decision Tree is a rule-based classification algorithm that recursively partitions the dataset based on feature values. The model generates an easily interpretable tree structure where internal nodes represent decision rules and leaf nodes correspond to predicted performance categories.

Advantages

- Highly interpretable
- Handles numerical and categorical variables
- Requires minimal data preparation

Limitations

- Prone to overfitting
- Sensitive to small changes in the dataset

6.3.Random Forest (RF)

Random Forest is an ensemble learning algorithm that constructs multiple decision trees and combines their predictions using majority voting. Random Forest generally achieves higher predictive accuracy and better generalization than individual decision trees by reducing variance and minimizing overfitting.

Advantages

- High prediction accuracy
- Robust against noisy data
- Provides feature importance scores

Limitations

- More computationally intensive
- Reduced interpretability compared to a single decision tree

6.4.Support Vector Machine (SVM)

Support Vector Machine identifies optimal hyperplanes that maximize the separation between different performance classes. Kernel functions enable SVM to model complex nonlinear relationships within educational datasets.

Advantages

- Effective in high-dimensional datasets
- Good generalization capability
- Handles nonlinear boundaries effectively

Limitations

- Computationally expensive for large datasets
- Requires careful parameter tuning

6.5.K-Nearest Neighbour (KNN)

KNN is an instance-based learning algorithm that classifies students based on the performance of their nearest neighbours in the feature space. Classification depends on the majority class among the selected neighbouring instances.

Advantages

- Simple implementation
- No explicit training phase
- Effective for smaller datasets

Limitations

- Sensitive to irrelevant features
- Computational cost increases with dataset size

6.6.Naïve Bayes (NB)

Naïve Bayes is a probabilistic classifier based on Bayes' theorem and assumes conditional independence among predictor variables. Despite this assumption, it performs remarkably well in many educational prediction problems.

Advantages

- Fast and computationally efficient
- Works well with high-dimensional data

- Performs well even with limited training samples

Limitations

- Independence assumption may not hold in educational datasets

6.7.Gradient Boosting (GB)

Gradient Boosting constructs an ensemble of weak learners sequentially, where each subsequent model attempts to correct the errors of its predecessor. The algorithm optimizes prediction performance through gradient-based loss minimization.

Advantages

- High predictive accuracy
- Handles complex relationships
- Reduces prediction errors effectively

Limitations

- Longer training time
- Sensitive to parameter selection

6.8.Extreme Gradient Boosting (XGBoost)

XGBoost is an optimized implementation of gradient boosting that incorporates regularization techniques, parallel processing, and efficient tree construction. It has demonstrated superior performance across numerous machine learning applications, including educational data mining.

Advantages

- High accuracy
- Efficient handling of missing values
- Regularization reduces overfitting
- Fast computation

Limitations

- Requires hyper parameter optimization
- Less interpretable than simpler models

7. Model Evaluation

The proposed methodology recommends evaluating each classification model using multiple performance metrics rather than relying solely on accuracy. This comprehensive evaluation is particularly important for educational datasets, where class distributions may be imbalanced.

7.1.Confusion Matrix

The confusion matrix summarizes the classification outcomes by comparing predicted classes with actual classes. It enables the calculation of various evaluation metrics and provides detailed insights into classification performance. For multiclass

classification, the confusion matrix indicates the number of correctly and incorrectly classified students across all performance categories.

7.2.Accuracy

Accuracy represents the proportion of correctly classified instances among all observations.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Although accuracy provides an overall measure of classification performance, it may produce misleading results when class distributions are uneven.

7.3.Precision

Precision measures the proportion of students predicted as belonging to a particular class who actually belong to that class.

$$Precision = \frac{TP}{TP + FP}$$

Higher precision indicates fewer false positive predictions.

7.4.Recall

Recall, also known as sensitivity, measures the proportion of actual positive instances correctly identified by the classifier.

$$Recall = \frac{TP}{TP + FN}$$

In educational prediction, recall is particularly important for identifying at-risk students because missing a vulnerable student may delay necessary academic intervention.

7.5.F1-Score

The F1-score combines precision and recall into a single performance metric.

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

A higher F1-score indicates a balanced trade-off between precision and recall.

7.6.Receiver Operating Characteristic (ROC) Curve

The ROC curve illustrates the relationship between the True Positive Rate and False Positive Rate across different classification thresholds. The Area Under the Curve (AUC) provides a numerical measure of classifier discrimination capability. For multiclass classification, ROC analysis may be performed using

one-versus-rest evaluation. As shown in Table 2 Performance Metrics Used for Model Comparison.

Table 2 Performance Metrics Used for Model Comparison

Metric	Purpose
Accuracy	Overall prediction performance
Precision	Reliability of positive predictions
Recall	Ability to detect actual positive cases
F1-score	Balance between Precision and Recall
ROC-AUC	Overall discrimination capability
Confusion Matrix	Detailed class-wise performance

8. Expected Outcomes

The proposed methodology is expected to achieve the following outcomes:

- Early identification of academically vulnerable MCA students.
- Development of a robust multiclass classification framework.
- Comparative evaluation of multiple supervised learning algorithms.
- Identification of significant academic and behavioural predictors influencing student performance.
- Establishment of an intelligent decision-support system for academic administrators.
- Facilitation of timely counselling and personalized academic interventions.

The framework is also expected to support institutional efforts aimed at improving student retention, reducing academic failures, and enhancing overall educational quality.

9. Advantages of the Proposed Framework

The proposed methodology offers several advantages over conventional approaches:

- Utilizes multiple academic and behavioural indicators instead of relying solely on examination scores.

- Supports multiclass classification, enabling differentiated academic interventions.
- Compares multiple machine learning algorithms within a unified framework.
- Employs comprehensive evaluation metrics beyond overall accuracy.
- Adaptable to different institutional environments.
- Supports evidence-based academic decision-making.
- Encourages proactive rather than reactive intervention strategies.

10. Limitations

Despite its potential advantages, the proposed methodology has certain limitations:

- Prediction performance depends on the quality and completeness of institutional data.
- Student behaviour may change during the semester, affecting prediction accuracy.
- Different institutions may use varying assessment policies, limiting direct generalization.
- Behavioural variables collected from Learning Management Systems may not be uniformly available across institutions.
- Hyper parameter tuning and feature selection require computational expertise.

Future implementations may overcome these limitations through larger multi-institutional datasets and continuous model updating.

Conclusion

Early identification of academically at-risk students is essential for improving student success, reducing dropout rates, and enhancing institutional effectiveness. This paper has presented a comprehensive machine learning-based methodology for predicting the academic performance of MCA students using multiclass classification. The proposed framework integrates demographic, academic, and behavioural variables within a structured data pre-processing and feature engineering pipeline to support accurate student classification. Unlike many existing approaches that focus primarily on binary classification or overall prediction accuracy, the proposed methodology emphasizes multiclass classification into At-Risk,

Average Performer, and High Performer categories. It further recommends evaluating multiple supervised learning algorithms using comprehensive performance metrics such as Precision, Recall, F1-score, Confusion Matrix, and ROC-AUC, thereby providing a more balanced assessment of model effectiveness. The methodology provides a practical foundation for developing intelligent early warning systems capable of supporting timely academic interventions, personalized mentoring, and data-driven educational planning. Its adaptable design enables implementation across diverse higher education institutions while encouraging future empirical validation using real-world datasets.

Future Scope

Future research may extend the proposed methodology in several directions:

- Validation using real-world datasets from multiple universities and affiliated colleges.
- Integration of deep learning models, including Artificial Neural Networks and Long Short-Term Memory (LSTM) networks, for temporal analysis of student performance.
- Incorporation of explainable artificial intelligence (XAI) techniques, such as SHAP and LIME, to enhance the interpretability of model predictions.
- Development of real-time predictive dashboards integrated with institutional ERP and Learning Management Systems.
- Exploration of hybrid ensemble learning techniques that combine multiple classifiers to improve predictive accuracy and robustness.
- Investigation of additional behavioural and psychological factors, including student motivation, learning preferences, and engagement patterns, to further enhance prediction performance.

References

- [1]. Dabhade, P., Agarwal, R., Alameen, K. P., Fathima, A. T., Sridharan, R., & Gopakumar, G. (2021). Educational data mining for predicting students' academic performance using machine learning algorithms. *Materials Today: Proceedings*, 47(15), 5260–5267. <https://doi.org/10.1016/j.matpr.2021.05.646>
- [2]. Fernández-García, A. J., Preciado, J. C., & Sánchez-Figueroa, F. (2021). Predicting student performance using educational data mining: A systematic review. *Applied Sciences*, 11(22), 10599.
- [3]. Kaur, P., Singh, M., & Josan, G. S. (2015). Classification and prediction based data mining algorithms to predict slow learners in education sector. *Procedia Computer Science*, 57, 500–508.
- [4]. Kumar, M., Singh, A. J., & Handa, D. (2017). Literature survey on student's performance prediction in education using data mining techniques. *International Journal of Education and Management Engineering*, 7(6), 40–49.
- [5]. Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *IEEE Transactions on Learning Technologies*, 13(1), 1–18.
- [6]. Shahiri, A. M., Husain, W., & Rashid, N. A. (2015). A review on predicting students' performance using data mining techniques. *Procedia Computer Science*, 72, 414–422.
- [7]. Lakshmi, B. N., & Rani, K. U. (2019). Student academic performance prediction using machine learning techniques. *International Journal of Engineering and Advanced Technology*, 8(6), 3625–3630.
- [8]. Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2017). *Data mining: Practical machine learning tools and techniques* (4th ed.). Morgan Kaufmann.
- [9]. Zhang, Y., Yun, Y., An, R., Cui, J., Dai, H., & Shang, X. (2021). Educational data mining techniques for student performance prediction: Method review and comparison analysis. *Frontiers in Psychology*, 12, 698490. <https://doi.org/10.3389/fpsyg.2021.698490>
- [10]. Pandey, M., & Taruna, S. (2016). Towards the integration of multiple classifier paradigms for improving prediction accuracy

- in educational data mining. *Computers & Electrical Engineering*, 59, 200–210.
- Aggarwal, C. C. (2015). *Data mining: The textbook*. Springer.
- [11]. Ramesh, V., Parkavi, P., & Ramar, K. (2013). Predicting student performance: A statistical and data mining approach. *International Journal of Computer Applications*, 63(8), 35–39.
- [12]. Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques* (3rd ed.). Morgan Kaufmann.
- [13]. Hastie, T., Tibshirani, R., & Friedman, J. (2021). *The elements of statistical learning* (2nd ed.). Springer.
- [14]. Pandey, M., & Taruna, S. (2017). A comparative study of ensemble methods for students' performance modeling. *International Journal of Intelligent Systems and Applications*, 9(8), 34–45.
- [15]. Pal, S. (2017). Framework for predicting students' performance: Review of techniques. *Indian Journal of Computer Science*, 2(2), 1–10.
- [16]. Kotsiantis, S. B. (2012). Use of machine learning techniques for educational purposes: A decision support system for forecasting students' grades. *Artificial Intelligence Review*, 37(4), 331–344.
- [17]. Parveen, Z., & Quadri, M. P. (2019). Classification and prediction of student academic performance using machine learning: A review. *International Journal of Computer Sciences and Engineering*, 7(3), 607–614.
- [18]. Yadav, S. K., & Pal, S. (2012). Data mining: A prediction for performance improvement of engineering students using classification. *World of Computer Science and Information Technology Journal*, 2(2), 51–56.
- [19]. Al Breiki, B., Zaki, N., Alashwal, H., & Alajmi, A. (2020). A systematic literature review of student performance prediction using machine learning techniques. *Education and Information Technologies*, 25(6), 5521–5555.
- [20]. Baker, R. S., & Inventado, P. S. (2014). Educational data mining and learning analytics. In J. A. Larusson & B. White (Eds.), *Learning analytics: From research to practice* (pp. 61–75). Springer.
- [21]. Bakhshinategh, B., Zaiane, O. R., ElAtia, S., & Ipperciel, D. (2018). Educational data mining applications and tasks: A survey of the last decade. *Education and Information Technologies*, 23(1), 537–553.
- [22]. Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.
- [23]. Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM*, 55(10), 78–87.