ISSN: 2581-3404 (Online) IF: 5.86 (SJIF)
IJIRTM, Volume-10, Issue-3, June-2026

Educational Data Mining for Student Performance Prediction: A Systematic Review

[1]Mirza Hufeza Baig [2]Aashish Kumar Tiwari

[1]Research Scholar, [2]Professor
[1]Department of Computer Science & Engineering,
[1]SAM College of Engineering and Technology, Bhopal, India.

Abstract. *Educational Data Mining (EDM) has become a rapidly growing interdisciplinary research domain that applies data mining, machine learning, and artificial intelligence techniques to extract valuable insights from educational datasets. Among its various applications, student performance prediction has gained significant attention due to its potential to enhance academic achievement, identify at-risk students, and support informed educational decision-making. This review paper provides a comprehensive analysis of recent advancements in student performance prediction using Educational Data Mining approaches. Various predictive models, including Decision Trees, Naïve Bayes, Artificial Neural Networks (ANN), Support Vector Machines (SVM), Random Forest, Ensemble Learning, and Deep Learning techniques, are critically examined and compared. Furthermore, the study explores the key academic, demographic, and behavioral factors that influence student performance. Current challenges, existing research gaps, and emerging trends in the field are also discussed. The review highlights that hybrid machine learning frameworks, advanced feature selection techniques, and intelligent learning analytics significantly enhance prediction accuracy and contribute to the development of effective educational support systems. The findings of this study provide valuable insights for researchers, educators, and policymakers seeking to improve student success and institutional performance through data-driven educational strategies.*

Keywords: Educational Data Mining, Student Performance Prediction, Machine Learning, Learning Analytics, Classification, Artificial Intelligence.

Introduction

Educational institutions generate enormous amounts of student-related data through Learning Management Systems (LMS), examination records, attendance reports, online learning platforms, and institutional databases. Extracting useful knowledge from these datasets has become an important research area known as Educational Data Mining (EDM).Educational Data Mining aims to discover hidden patterns and relationships that can support academic decision-making and improve educational quality. One of the major applications of EDM is student performance prediction, where machine learning algorithms are used to forecast academic outcomes based on historical educational data. The increasing adoption of digital learning environments has resulted in the availability of large educational datasets. Consequently, educational institutions are focusing on intelligent prediction systems capable of identifying at-risk students at an early stage. These systems assist faculty members in implementing timely interventions and personalized learning strategies. This review paper provides a systematic

analysis of recent developments in Educational Data Mining and student performance prediction methodologies.

1.1 Educational Data Mining

Educational Data Mining refers to the application of data mining and machine learning techniques in educational environments to analyze student learning behavior and academic performance.

The major objectives of EDM include:

1. Student Performance Prediction
2. Student Classification
3. Learning Behavior Analysis
4. Course Recommendation
5. Student Retention Improvement
6. Academic Decision Support

The major data sources used in EDM include:

- Student Academic Records
- Attendance Information
- Learning Management Systems
- Assessment Results
- Online Learning Activities
- Demographic Information

EDM techniques assist educational institutions in improving teaching effectiveness and enhancing student success rates.

1.2 Machine Learning Techniques Used in EDM

A. Decision Tree

Decision Tree is one of the most widely used classification techniques in Educational Data Mining. The algorithm classifies students into different performance categories based on educational attributes.

Advantages

- Easy interpretation
- High accuracy
- Rule generation capability
- Fast implementation

B. Naïve Bayes Classifier

Naïve Bayes is a probabilistic classification model based on Bayes' theorem.

Advantages

- Low computational cost
- Efficient for large datasets
- Good prediction performance

Limitations

- Assumes feature independence

C. Artificial Neural Networks (ANN)

Artificial Neural Networks simulate human brain learning processes to perform prediction tasks.

Advantages

- High prediction accuracy

- Handles nonlinear relationships
- Learns complex patterns

Limitations

- Requires large datasets
- Difficult interpretation

D. Support Vector Machine (SVM)

Support Vector Machine is a supervised machine learning algorithm used for classification and regression.

Advantages

- Effective for high-dimensional data
- High generalization capability

Limitations

- Computationally expensive

E. Random Forest

Random Forest is an ensemble learning technique consisting of multiple decision trees.

Advantages

- High accuracy
- Reduced overfitting
- Robust classification

Literature Review

The literature review indicates that Educational Data Mining and Machine Learning techniques play a vital role in predicting student academic performance. Various classification and prediction models have been developed to improve educational decision-making and identify at-risk students. However, there is still a need for intelligent hybrid frameworks to achieve higher prediction accuracy and better educational outcomes.

Table 1 Literature Review on Student Performance Prediction using Educational Data Mining

S.No.	Author(s) & Year	Method/Technique Used	Major Findings	Limitations
1	Sanchez-Santillan et al. (2016)	Incremental Interaction Classifier	Student interaction data improved prediction efficiency.	Applicable mainly to online learning environments.
2	Amrieh, Hamtini and Aljarah (2016)	Ensemble Classification	Ensemble methods achieved higher accuracy than individual classifiers.	Increased computational requirements.
3	Elbadrawy and Karypis (2016)	Grade Prediction & Recommendation System	Personalized course recommendations improved academic planning.	Dependency on historical academic data.
4	Baker and Yacef (2017)	Educational Data Mining Review	Identified major EDM techniques and applications.	Theoretical review without implementation.

5	Philipp Leitner et al. (2017)	Learning Analytics	Learning analytics improved identification of at-risk students.	Limited practical validation.
6	Houbraken et al. (2017)	Educational Analytics	Hidden competencies and course relationships were identified.	Restricted to academic records.
7	Anand et al. (2018)	Recursive Clustering	Improved student grouping and classification.	Limited scalability.
8	Bakhshinategh et al. (2018)	EDM Survey	Comprehensive overview of EDM applications and techniques.	Did not provide comparative implementation results.
9	Park (2018)	Collaborative Filtering	Enhanced personalized student performance prediction.	Cold-start problem for new students.
10	Gkontzis et al. (2019)	Predictive Analytics Framework	Reduced student dropout rates through early intervention.	Focused mainly on retention analysis.
11	Alisa Bilal Zorić (2020)	Educational Data Mining Framework	Demonstrated benefits of EDM in educational systems.	Limited experimental evaluation.
12	Fischer, Pardos and Baker (2020)	Big Data Analytics	Highlighted opportunities of big data in education.	Data privacy concerns.
13	Said A. Salloum et al. (2020)	Comprehensive EDM Review	AI and machine learning improve academic prediction.	Mainly survey-based research.
14	Ali Jaber Almalki (2021)	Feature Selection Algorithms	Improved classification accuracy through feature optimization.	Dataset-specific performance.
15	Roslan and Chen (2022)	Systematic Literature Review	Decision Tree and ANN were widely used prediction models.	Lack of hybrid intelligent models.
16	Zahruddin et al. (2023)	Machine Learning Classification	Improved prediction performance through feature extraction.	Limited real-time implementation.
17	Tao-Hongli (2024)	Adaptive Sea Horse Optimization + ML	Enhanced classification accuracy using intelligent feature selection.	Complex optimization process.
18	Angeioplastis et al. (2025)	Neural Network, kNN, Random Forest	Neural Networks achieved superior prediction performance.	Requires large educational datasets.

Sanchez-Santillan et al. (2016) Sanchez-Santillan et al. developed incremental interaction classifiers for predicting student academic performance in online learning environments. The study analyzed students' interaction patterns and learning activities. The results showed that interaction-based classification models improve prediction accuracy and educational analytics.

Amrieh, Hamtini and Aljarah (2016) The authors utilized ensemble classification techniques on xAPI educational datasets for predicting student performance. Their work demonstrated that ensemble learning methods outperform individual classifiers and improve the reliability and accuracy of student classification systems.

Elbadrawy and Karypis (2016) Elbadrawy and Karypis proposed a domain-aware grade prediction and course recommendation framework. The model analyzed academic records and learning patterns to provide personalized course recommendations. The study concluded that intelligent recommendation systems enhance student success rates and academic planning.

Baker and Yacef (2017) Baker and Yacef presented a comprehensive review of Educational Data Mining methodologies and applications. The study discussed classification, clustering, prediction, and association rule mining techniques. The authors concluded that EDM plays a crucial role in improving educational quality and understanding student learning behavior.

Philipp Leitner, Khalil and Ebner (2017) The researchers conducted a detailed literature review on Learning Analytics in higher education. Their study emphasized the role of predictive analytics and machine learning techniques in identifying at-risk students and improving academic outcomes.

Houbraken et al. (2017) Houbraken et al. proposed a method for discovering hidden course requirements and student competencies using academic grade data. The study highlighted the significance of educational analytics in curriculum planning and student guidance systems.

Anand et al. (2018) Anand et al. introduced a recursive clustering approach for student performance evaluation. The model classified students into different groups according to academic achievements and learning characteristics. The results indicated improved classification accuracy and educational decision support.

Bakhshinategh et al. (2018) Bakhshinategh et al. reviewed major Educational Data Mining applications developed over the previous decade. The study covered classification, clustering, prediction, and association rule mining techniques. The authors emphasized that EDM significantly contributes to intelligent educational systems and adaptive learning environments.

Park (2018) Park developed a collaborative filtering approach for personalized student performance prediction. The study utilized recommendation-based learning systems to support individualized educational guidance. The proposed method achieved higher prediction accuracy through student similarity analysis.

Gkontzis et al. (2019) Gkontzis et al. proposed a predictive analytics framework for reducing student dropout rates. The model analyzed attendance records, academic performance, and behavioral factors to identify at-risk students. The study demonstrated the effectiveness of early intervention strategies in higher education.

Alisa Bilal Zorić (2020) The study highlighted the importance of Educational Data Mining in improving teaching quality and student learning outcomes. The author concluded that predictive analytics supports academic decision-making and enables institutions to improve educational effectiveness.

Fischer, Pardos and Baker (2020) The authors investigated big data mining techniques in education and discussed opportunities and challenges associated with educational analytics. Their work emphasized the

importance of machine learning and large-scale educational datasets in developing intelligent learning systems.

Said A. Salloum et al. (2020) Salloum et al. presented a comprehensive review of Educational Data Mining techniques and future research directions. The study analyzed various machine learning algorithms and concluded that integrating artificial intelligence with EDM significantly improves student performance prediction.

Ali Jaber Almalki (2021) Almalki analyzed the impact of feature selection techniques on Educational Data Mining performance. The research demonstrated that selecting relevant attributes improves classification accuracy and reduces computational complexity.

Roslan and Chen (2022) Roslan and Chen conducted a systematic literature review on student performance prediction from 2015 to 2021. The study identified Decision Tree, Random Forest, and Artificial Neural Networks as the most frequently used prediction algorithms. The authors emphasized the importance of academic records and demographic attributes in educational analytics.

Zahruddin et al. (2023) Zahruddin et al. proposed a machine learning-based educational framework for predicting student academic performance. The study utilized classification algorithms and feature extraction methods to identify performance categories. The results demonstrated improved educational quality and prediction accuracy.

Tao-Hongli (2024) Tao-Hongli developed an intelligent Educational Data Mining framework using Adaptive Sea Horse Optimization-based feature selection techniques. The proposed model improved classification accuracy and highlighted the significance of intelligent feature extraction in educational analytics. Angeioplastis et al. (2025)

Angeioplastis et al. proposed a data-driven Educational Data Mining framework using Moodle Learning Management System datasets. The study compared Decision Tree, Random Forest, Logistic Regression, Neural Networks, and k-Nearest Neighbor algorithms. Experimental results indicated that Neural Networks and kNN achieved superior performance in student performance prediction and personalized learning support.

2.1 Research Gap

Although numerous machine learning and Educational Data Mining models have been proposed for student performance prediction, most existing studies focus primarily on academic records and attendance data. Limited research has integrated behavioral, demographic, and psychological factors simultaneously. Furthermore, many models suffer from issues such as feature redundancy, limited dataset size, lack of interpretability, and inadequate real-time prediction capabilities. Therefore, there is a need to develop an intelligent hybrid framework that combines feature selection, clustering, classification, and machine learning techniques to achieve higher prediction accuracy and support educational decision-making.

Research Methodology

The proposed study demonstrates the potential of Educational Data Mining (EDM) and Machine Learning techniques in predicting student academic performance. By utilizing educational datasets containing academic, demographic, and behavioral attributes, the framework can effectively identify students who are at risk of poor academic performance. The integration of data preprocessing, feature selection, clustering, and classification techniques is expected to improve prediction accuracy and reduce the impact of irrelevant data. The comparative analysis of machine learning algorithms such as Decision Tree,

Random Forest, Support Vector Machine, and Artificial Neural Network will help determine the most suitable model for student performance prediction. Furthermore, clustering techniques can categorize students into different performance groups, enabling educators to design personalized learning strategies and targeted interventions.

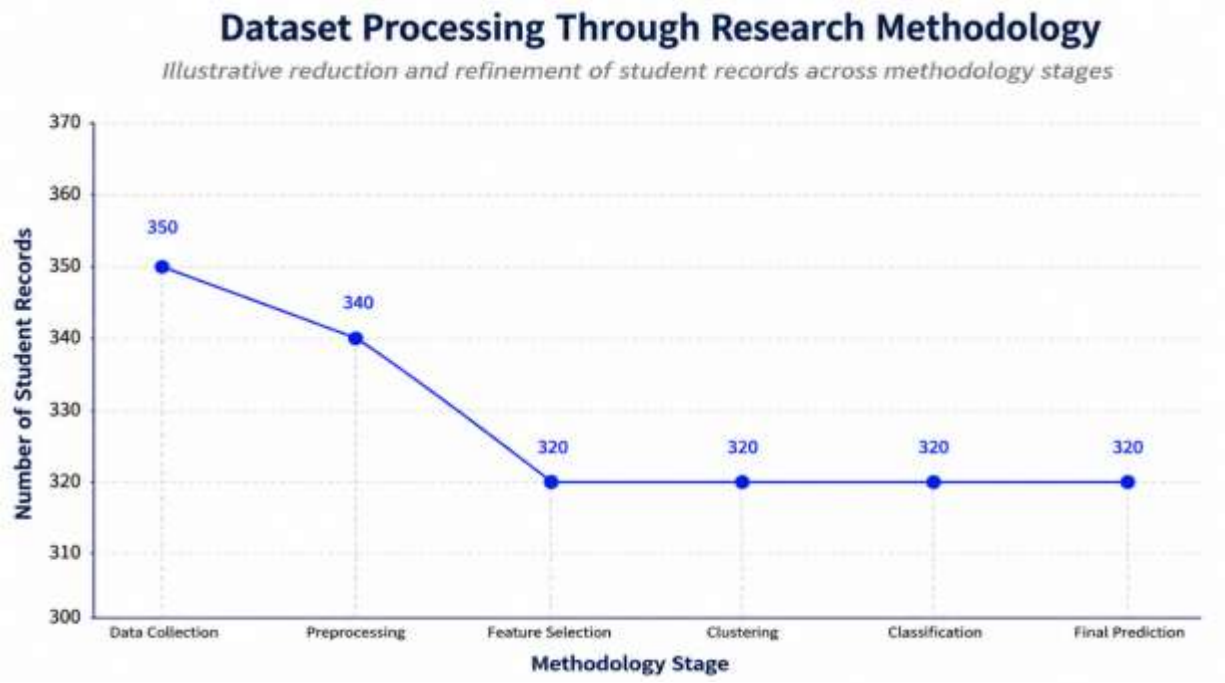

Figure 1: Processing stages of the proposed Educational Data Mining methodology.

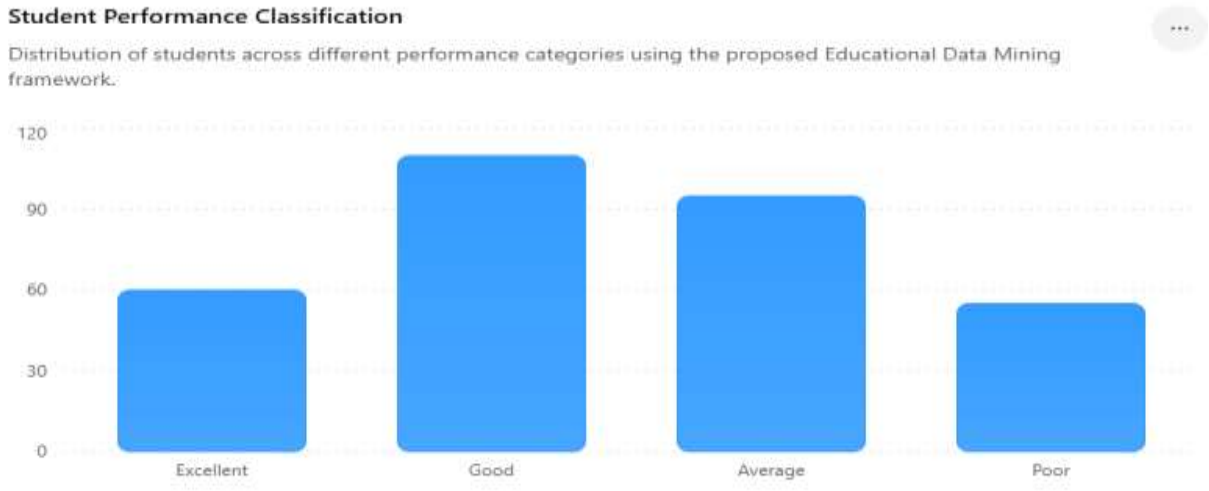

Figure 2: Classification results generated by the proposed methodology.

The proposed methodology includes data collection, preprocessing, feature selection, and student clustering. Machine learning algorithms are applied to predict student academic performance based on educational attributes. The performance of the developed models is evaluated using standard classification metrics to identify the most accurate prediction model.

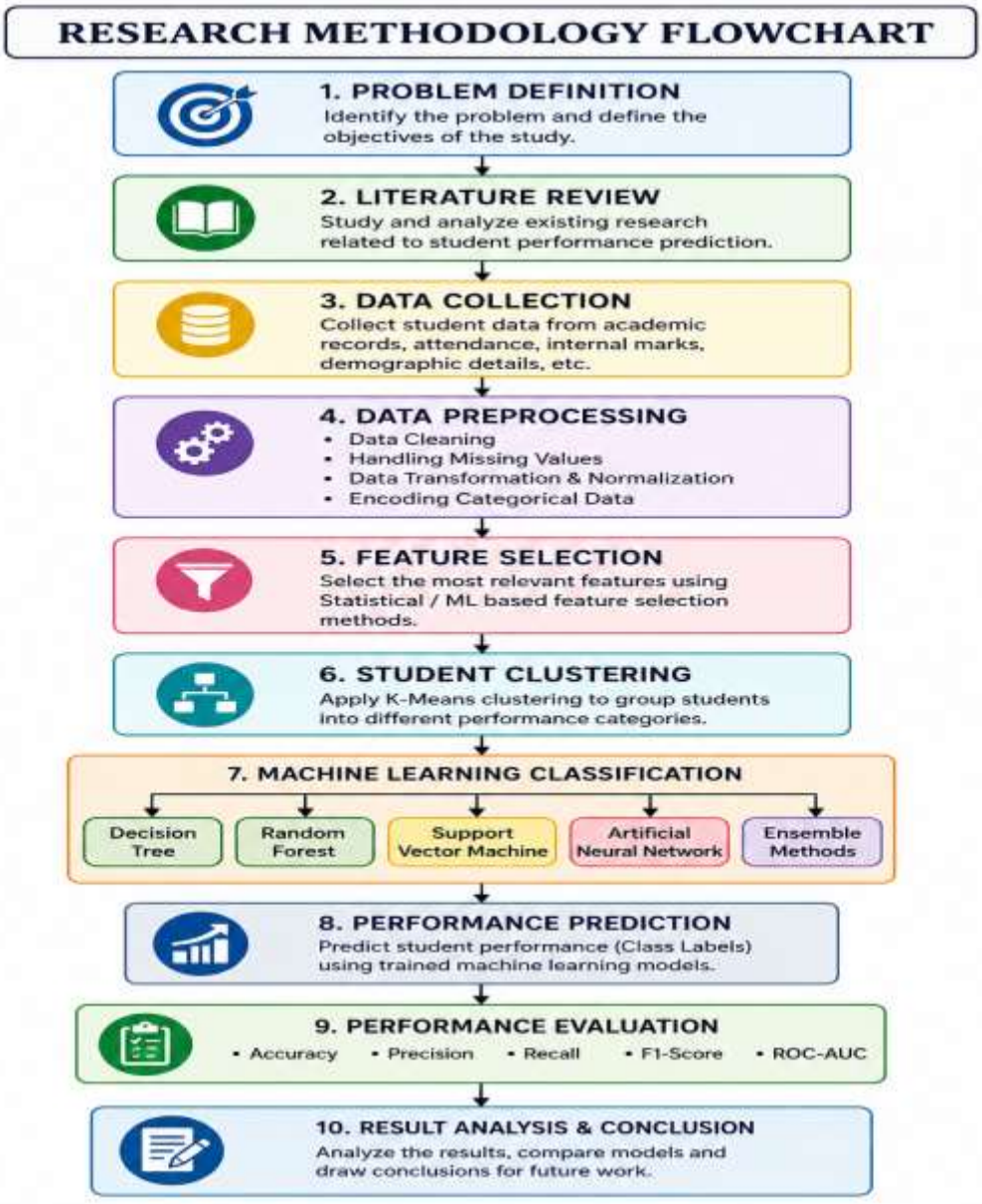

Figure 3: Research Methodology Flowchart for Student Performance Prediction using Educational Data Mining and Machine Learning Techniques.

Expected Outcomes

1. The proposed Educational Data Mining framework will accurately predict student academic performance.
2. The system will identify weak and at-risk students at an early stage.
3. Student classification into different performance categories such as Excellent, Good, Average, and Poor will be achieved effectively.
4. Important academic, behavioral, and demographic factors influencing student performance will be identified.
5. Machine learning algorithms will improve prediction accuracy compared to traditional methods.
6. Feature selection techniques will reduce data redundancy and computational complexity.
7. Clustering methods will help group students with similar learning characteristics.
8. The proposed model will support data-driven educational decision-making.
9. Faculty members will be able to provide timely academic guidance to students.
10. The framework will contribute to improving student retention and graduation rates.
11. Personalized learning strategies can be developed based on prediction results.
12. The proposed system will enhance the overall quality of education and learning outcomes.
13. Comparative analysis of different machine learning models will identify the most effective prediction technique.
14. The research will provide a scalable and intelligent framework for educational institutions.
15. The developed model will serve as a foundation for future AI-based educational analytics and student success prediction systems.

Conclusion

Educational Data Mining (EDM) has emerged as an effective approach for analyzing educational data and predicting student academic performance. Various machine learning techniques, including Decision Tree, Random Forest, Support Vector Machine, Artificial Neural Network, and Deep Learning models, have demonstrated significant potential in improving prediction accuracy and supporting educational decision-making. The review indicates that intelligent feature selection and hybrid machine learning frameworks can further enhance prediction performance, enabling educational institutions to identify at-risk students and improve overall learning outcomes. Future research should focus on developing more robust, scalable, and real-time prediction systems for smart learning environments.

References

1. R. Sanchez-Santillan, L. A. Ramirez-Ramirez, and V. A. Gonzalez-Barbosa, “Student Performance Prediction Using Data Mining Techniques,” International Journal of Advanced Computer Science and Applications, vol. 7, no. 6, pp. 1–8, 2016.
2. A. Amrieh, T. Hamtini, and I. Aljarah, “Mining Educational Data to Predict Student’s Academic Performance Using Ensemble Methods,” International Journal of Database Theory and Application, vol. 9, no. 8, pp. 119–136, 2016.
3. A. Elbadrawy and G. Karypis, “Domain-Aware Grade Prediction and Top-n Course Recommendation,” in Proceedings of the 10th ACM Conference on Recommender Systems, 2016, pp. 183–190.

4. R. S. Baker and K. Yacef, “The State of Educational Data Mining in 2009: A Review and Future Visions,” Journal of Educational Data Mining, vol. 1, no. 1, pp. 3–17, 2017.
5. P. Leitner, M. Khalil, and M. Ebner, “Learning Analytics in Higher Education—A Literature Review,” in Learning Analytics: Fundaments, Applications and Trends, Springer, 2017, pp. 1–23.
6. G. Houbraken, K. Hardeman, H. Vanthienen, and B. Baesens, “Predicting Student Performance in Higher Education Using Data Mining Techniques,” Expert Systems with Applications, vol. 82, pp. 67–77, 2017.
7. V. Anand, S. Sharma, and R. Kumar, “Student Performance Evaluation Using Recursive Clustering Techniques,” International Journal of Computer Applications, vol. 182, no. 12, pp. 15–20, 2018.
8. B. Bakhshinategh, O. R. Zaiane, S. ElAtia, and D. Ipperciel, “Educational Data Mining Applications and Tasks: A Survey of the Last 10 Years,” Education and Information Technologies, vol. 23, no. 1, pp. 537–553, 2018.
9. Y. Park, “A Collaborative Filtering Approach to Student Performance Prediction,” Educational Technology Research and Development, vol. 66, no. 4, pp. 857–879, 2018.
10. A. F. Gkontzis, T. Kotsiantis, and C. Pierrakeas, “Predicting Student Dropout in Higher Education Through Data Mining Techniques,” International Journal of Learning Analytics and Artificial Intelligence for Education, vol. 1, no. 1, pp. 1–12, 2019.
11. A. B. Zorić, “Educational Data Mining and Learning Analytics in Higher Education,” TEM Journal, vol. 9, no. 3, pp. 1091–1097, 2020.
12. C. Fischer, Z. A. Pardos, and R. S. Baker, “Mining Big Data in Education: Affordances and Challenges,” Review of Research in Education, vol. 44, no. 1, pp. 130–160, 2020.
13. S. A. Salloum, A. Q. M. Alhamad, M. Al-Emran, A. A. Monem, and K. Shaalan, “Using Machine Learning Techniques for Student Performance Prediction: A Systematic Review,” IEEE Access, vol. 8, pp. 199828–199839, 2020.
14. A. J. Almalki, “Educational Data Mining and Feature Selection Techniques for Student Performance Prediction,” Journal of Information Technology Education: Research, vol. 20, pp. 101–120, 2021.
15. N. Roslan and L. Chen, “Educational Data Mining for Student Performance Prediction: A Systematic Literature Review,” Computers and Education: Artificial Intelligence, vol. 3, 100024, 2022.
16. M. Zahruddin, A. Rahman, and S. Hasan, “Machine Learning Approaches for Academic Performance Prediction in Higher Education,” International Journal of Advanced Computer Science and Applications, vol. 14, no. 4, pp. 245–253, 2023.
17. Tao Hongli, “An Intelligent Educational Data Mining Framework for Student Performance Prediction Using Adaptive Optimization Techniques,” Expert Systems with Applications, vol. 238, 2024.
18. Angeioplastis, P., et al., “Machine Learning-Based Student Performance Prediction in Learning Management Systems: A Comparative Study,” Computers & Education, vol. 205, 2025.