A Systematic Review and Comparative Analysis of Educational Data Mining Techniques for Student Performance Prediction

Zahira Noor Quraishi[1], Dr. Atul Dattarya Newase[2]

[1]Research Scholar, [2]Research Supervisor
[1,2]Dr. A. P. J. Abdul Kalam University, Indore, India
zahira16sep@gmail.com , dr.atulnewase@gmail.com

Presented at International Conference on Engineering, Economics, Management and Applied Sciences (ICE2MAS-24), Bangkok, 21-24 December 2024, organized by Academy of Art, Science and Technology (AAST).

Published in IJIRMPS (E-ISSN: 2349-7300), ICE2MAS-24

Abstract
Educational institutions increasingly rely on data-driven approaches to enhance learning outcomes and reduce academic failure. Educational Data Mining (EDM) has emerged as a prominent research field that applies data mining and machine learning techniques to analyze educational data for predicting student performance. This study presents a systematic review and comparative analysis of EDM techniques used for student performance prediction across diverse educational contexts. Following a structured review process, recent empirical studies and review papers were analyzed with respect to prediction objectives, data sources, modeling approaches, evaluation metrics, and reported outcomes. The review highlights the widespread use of traditional machine learning models such as decision trees, random forest, naïve Bayes, neural networks, and support vector machines, along with emerging methods including causal modeling, fuzzy logic, and optimization-based approaches. Comparative analysis indicates that ensemble and tree-based models often achieve high predictive accuracy, while simpler models provide greater interpretability for educational decision-making. However, the findings also reveal key limitations, including overreliance on higher education datasets, limited focus on primary education, data quality challenges, and insufficient integration of predictive models with intervention mechanisms. By synthesizing trends, strengths, and gaps, this study provides a consolidated understanding of current EDM practices and offers guidance for developing robust, explainable, and context-aware student performance prediction systems.

Keywords: Educational Data Mining, Student Performance Prediction, Machine Learning, Systematic Review, Learning Analytics.

1. Introduction

The rapid digitalization of education has led to an unprecedented growth in educational data generated from learning management systems, online platforms, assessments, student information systems, and institutional databases. These data sources contain valuable information about students' academic progress, learning behavior, engagement patterns, and demographic characteristics. Effectively utilizing this data has become essential for improving educational quality, supporting early intervention, and enhancing student success. Educational Data Mining (EDM) has emerged as a specialized interdisciplinary field that applies data mining, machine learning, and statistical techniques to extract meaningful insights from educational data [1].

One of the most important applications of EDM is student performance prediction. Predicting academic outcomes enables educators and administrators to identify students at risk of failure or dropout, design personalized learning strategies, and allocate resources more effectively. Early and accurate predictions are particularly valuable in large-scale educational environments, open and distance learning systems, and institutions with high enrollment and dropout rates. As a result, student performance prediction has become a central research topic within the EDM community [2].

Over the past decade, numerous data mining techniques have been applied to student performance prediction. Traditional supervised learning algorithms such as decision trees, naïve Bayes, logistic regression, support vector machines, and k-nearest neighbors are widely used due to their simplicity and ease of interpretation. Ensemble methods, especially random forest and boosting-based models, have gained popularity for their robustness and ability to handle complex, high-dimensional datasets. In addition, neural networks and deep learning approaches have been explored to capture non-linear relationships in learning data [3].

Beyond predictive accuracy, recent studies have emphasized the importance of feature engineering, data preprocessing, and class imbalance handling. Feature selection techniques help identify influential academic, behavioral, and demographic variables, improving both model performance and interpretability. Resampling methods are often applied to address skewed class distributions commonly observed in educational datasets, particularly in dropout and failure prediction tasks [4].

Despite significant progress, the existing literature reveals several challenges. First, EDM research is heavily skewed toward higher education, while primary and secondary education contexts remain underrepresented. This imbalance is concerning because early educational stages play a critical role in shaping long-term learning trajectories. Second, many studies rely on data from a single institution or region, limiting the generalizability and scalability of proposed models. Third, while high predictive accuracy is frequently reported, fewer studies address explainability, causal reasoning, and practical deployment in real educational settings [5].

In response to these challenges, recent research has begun exploring advanced and hybrid approaches, such as causal modeling to improve interpretability, fuzzy logic to handle uncertainty in human judgment, and optimization techniques to enhance predictive performance. Additionally, there is growing interest in integrating prediction models with learning analytics and intervention mechanisms to move beyond prediction toward actionable educational support [6].

Given the rapid expansion and diversification of EDM research, systematic review and comparative analysis are necessary to consolidate existing knowledge, identify dominant trends, and highlight research gaps. While several review studies exist, many focus on specific domains such as programming education or admissions or emphasize descriptive trends without detailed methodological comparison. This study aims to address this need by systematically reviewing recent EDM literature on student performance prediction and conducting a comparative analysis of techniques, data types, and outcomes across educational levels [7].

The objectives of this study are threefold: (1) to identify commonly used EDM techniques and data sources for student performance prediction, (2) to compare their reported effectiveness, strengths, and limitations, and (3) to highlight open challenges and future research directions. By providing a structured and comprehensive synthesis, this work seeks to support researchers, educators, and policymakers in developing more effective, explainable, and context-aware EDM solutions for improving student performance.

2. Literature review

Tosun et al. (2024), Academic success prediction is crucial in open and distance education programs with mass enrollment, where dropout risk is often high. This study predicts student success (successful vs unsuccessful) for 26,708 Istanbul University learners enrolled between 2011 and 2017 using demographic data and course grades in several subjects. Using SPSS Modeler 18, the dataset was split into training (70%) and testing (30%), and multiple supervised classifiers were compared, including Random Forest, C&RT, C5.0, CHAID, naïve Bayes, logistic regression, neural nets, and SVM. The C&RT model achieved the best performance, notably the highest specificity (0.915). **[1]**

Choi et al. (2023), Programming education is essential but challenging for beginners, and EDM is increasingly used to understand learning behavior and improve outcomes in programming courses. This systematic literature review synthesizes research from the last five years on EDM-based performance prediction in programming education. It examines common data sources and influential features, predictive targets, modeling approaches, preprocessing steps, validation strategies, and evaluation metrics used to assess model quality. The review also discusses limitations and challenges across prediction approaches and proposes directions for future work, aiming to guide researchers toward more robust and meaningful prediction systems in programming learning contexts. **[2]**

Khairy et al. (2024), Forecasting student outcomes is important for higher education quality assurance, including accreditation, and supports efforts to reduce failure and improve persistence. This study predicts the performance of first-level undergraduate Computer Department students (2016–2021) using institutional records from Damietta University. After cleaning, 830 instances remained with six features (e.g., midterm, practical, written exam, total degree, grade), split into 70% training and 30% testing. Five ML algorithms were compared—Random Forest, Decision Tree, naïve Bayes, neural network, and k-nearest neighbors—using accuracy, precision, recall, F-measure, and confusion matrices. Random Forest and Decision Tree performed best, misclassifying only 3 of 253 test instances. **[3]**

Wang et al. (2024), While EDM and LA can support early warning and intervention, the literature still needs more empirical evidence on feedback interventions—especially in primary and secondary contexts. This study proposes a data-driven precision teaching intervention mechanism combining EDM and LA for prediction plus actionable support. A quasi-experiment with 142 seventh-grade students compared an experimental group receiving precision interventions against two control groups receiving traditional or experience-stratified group interventions. After three intervention rounds, the experimental group showed higher academic achievement, intrinsic motivation, self-efficacy, and metacognitive awareness than controls. The results suggest integrating prediction with targeted interventions can improve learning outcomes and personalization in secondary education. **[4]**

Silva Filho et al. (2023), To address a common limitation in EDM—weak causal reasoning—this study combines EDM techniques with theory-driven causal models to better interpret performance interventions. Using large-scale Brazilian assessment data, the authors map unobserved confounders with causal graphs and apply a two-way fixed-effects logistic regression to control for confounding. The model's predictive ability is evaluated and then explored via classification rules and decision trees to generate interpretable insights. Findings emphasize the influence of socio-economic factors and highlight the impact of faculty education policies, including variation across Brazilian states, demonstrating how causal modeling can strengthen the usefulness of EDM findings for decision-makers. **[5]**

Yang et al. (2024), A data-mining-based high-quality management method is proposed to improve higher education quality, student satisfaction, and employment outcomes. The approach first builds a higher-education quality evaluation system, then applies association rule mining to construct a management model and compute weights for key impact indicators. A fuzzy evaluation method is subsequently used to define an evaluation function and generate quality scores, which guide targeted improvement strategies. Reported experimental results claim very high outcomes, with student satisfaction reaching 99.3% and employment rate reaching 99.9%, positioning the method as a decision-support framework for quality enhancement. **[6]**

Putri et al. (2024), To address underutilized academic data in Indonesia and support preventive action against low grades and expulsion risk, this study predicts student performance using sociodemographic variables and semester grade averages. Using data from 643 vocational high school students, Decision Tree C4.5 and Naïve Bayes were implemented in RapidMiner, achieving accuracies of 78.12% and

76.88%, respectively. “Gender” emerged as the most influential factor in this setting. The resulting classification rules are positioned as actionable guidance for schools to identify students likely to struggle with minimum grade requirements and intervene earlier. **[7]**

Nagarajan et al. (2024), This research targets student performance prediction as a sustainability-linked measure of learning quality, using supervised machine learning to forecast grades and marks. A regression framework and a Decision Tree classifier are trained on labeled academic history with 30 selected characteristics, and the study proposes a Genetic Algorithm (GA)-enhanced decision tree to improve predictive output. The reported results argue that the improved decision tree provides more accurate and simpler prediction for student achievement, reinforcing EDM’s value for planning and long-term educational development using large institutional datasets. **[8]**

Arief et al. (2024), Using GPA and contextual factors (e.g., parents’ job/education, address, gender, extracurriculars), this study predicts academic performance for Information Systems students at the University of Jember. Multiple machine learning models were compared, including Decision Tree, Random Forest, KNN, SVC, Naïve Bayes, and Gaussian methods. Results show the Decision Tree achieved the highest accuracy (0.9264), followed by Random Forest and KNN, and the study notes these top models produced consistent prediction outputs. The work supports using EDM as an institutional tool for understanding student success patterns and improving academic decision-making. **[9]**

Chytas et al. (2023), An interactive system is proposed to assess and improve learning processes using data generated by online university services, analyzed across periods before, during, and after the COVID-19 outbreak at a Greek university. By examining learning paths, online presence, and service participation, the system derives performance insights and predicts future learning progression. The study argues such analytics can help universities refine learning design, adjust online and in-person delivery, and strengthen strategic planning. Overall, it positions institutional service data as a resource for improving quality, supporting students, and enabling more targeted teaching practices. **[10]**

Gök et al. (2023), This study applies data mining to understand factors influencing primary teachers’ mathematics teaching anxiety and motivation, using Random Forest for prediction and K-Means clustering to define profiles. Survey data from 485 Turkish teachers included demographic variables alongside standardized anxiety and motivation scales, with outcomes transformed into low/high categories. Across both models, “grade level taught” had the highest predictive importance, followed by “length of service.” The work demonstrates how EDM methods can reveal educator profiles and key predictors, potentially informing targeted professional support and interventions for mathematics teaching. **[11]**

Liu et al. (2024), Curriculum reform for physical education in China is examined using data mining methods combined with literature review, questionnaires, and analytical techniques to characterize “innovative ability” in PE majors. The study conceptualizes innovation as both thinking and practice abilities and proposes a structured indicator system with five primary indicators and eight secondary indicators tailored to PE contexts. It distinguishes subjective influences (e.g., innovative consciousness, motivation, knowledge) from objective conditions (e.g., teaching content, evaluation, environment, incentives). These findings are framed as guidance for curriculum design and educational reform aimed at developing innovative PE talent. **[12]**

Table 1. Systematic literature review

Ref	Author (First author et al.)	Year	Title	Methods	Result	Advantage	Limitation	Data Used
[1]	Tosun et al.	2024	Data mining approach for prediction of academic success in open and distance education	Supervised classifiers (Random Forest, C&RT, C5.0, CHAID, Naïve Bayes, Logistic Regression, Neural Nets, SVM) using SPSS Modeler	C&RT achieved best performance with highest specificity (0.915)	Large-scale dataset; comprehensive model comparison	Limited to open and distance education context	Demographic data and course grades of 26,708 students
[2]	Choi et al.	2023	A systematic literature review on performance prediction in learning programming	Systematic literature review (last 5 years)	Identified common data sources, features, models, and challenges	Comprehensive synthesis guiding future research	No empirical validation	Prior EDM studies in programming education
[3]	Khairy et al.	2024	Prediction of student exam performance using data mining classification algorithms	Random Forest, Decision Tree, Naïve Bayes, Neural Network, KNN	RF and DT misclassified only 3 test instances	High prediction accuracy; robust evaluation metrics	Small institutional dataset	Institutional records of 830 undergraduate students
[4]	Wang et al.	2024	A data-driven precision teaching intervention mechanism	EDM + Learning Analytics; quasi-experimental design	Precision intervention group showed higher achievement and motivation	Integrates prediction with actionable interventions	Limited sample size	Academic and behavioral data of 142 Grade-7 students
[5]	Silva Filho et al.	2023	Leveraging causal reasoning in educational data mining	Causal graphs; two-way fixed-effects logistic regression; decision trees	Socio-economic factors and policies significantly influence performance	Improves interpretability and policy relevance	Complex modeling requirements	Large-scale Brazilian secondary education assessment data
[6]	Yang et al.	2024	High quality management of higher education based on data mining	Association rule mining; fuzzy evaluation model	Student satisfaction reached 99.3%, employment rate 99.9%	Decision-support framework for quality management	Results may be overly optimistic	Higher education quality evaluation indicators

[7]	Putri et al.	2024	Application of data mining to predict student learning outcomes	Decision Tree C4.5; Naïve Bayes (RapidMiner)	DT achieved 78.12% accuracy; gender most influential	Actionable classification rules for schools	Moderate accuracy; context-specific	Data of 643 vocational high school students
[8]	Nagarajan et al.	2024	Predicting academic performance using modified decision tree based GA	Regression; Decision Tree; GA-enhanced Decision Tree	Improved accuracy and model simplicity	Combines optimization with interpretability	Requires parameter tuning	Labeled academic history with 30 attributes
[9]	Arief et al.	2024	Educational Data Mining for Student Academic Performance Analysis	DT, RF, KNN, SVC, Naïve Bayes, Gaussian models	Decision Tree achieved highest accuracy (0.9264)	Consistent and interpretable predictions	Single-institution study	GPA and demographic/contextual student data
[10]	Chytas et al.	2023	Educational data mining in the academic setting	Learning analytics; performance prediction	Identified patterns to improve learning design	Uses real institutional service data	Context limited to one university	Online service usage and LMS data
[11]	Gök et al.	2023	Variables affecting primary teachers’ anxiety and motivation	Random Forest; K-Means clustering	Grade level taught most influential factor	Identifies teacher profiles for interventions	Focus on teachers, not students	Survey data of 485 primary teachers
[12]	Liu et al.	2024	Cultivation of innovative ability of PE students using data mining	Data mining + questionnaires + indicator system	Identified key subjective and objective innovation factors	Structured guidance for curriculum reform	Conceptual, not predictive	Survey and curriculum-related data

3. Research Gap

Although existing studies demonstrate the effectiveness of educational data mining techniques for predicting student performance and supporting educational decision-making, several critical gaps remain. As shown in Table 1, the literature is dominated by empirical and applied studies, while systematic literature reviews are limited, indicating a lack of consolidated frameworks that integrate findings across diverse educational contexts. Most research focuses on higher education, with primary education receiving minimal attention and secondary education only moderately represented, despite the importance of early-stage interventions for long-term academic success. Furthermore, while traditional machine learning methods such as decision trees, random forest, and naïve Bayes are widely used, advanced approaches incorporating causal reasoning, fuzzy logic, optimization algorithms, and intervention-oriented analytics remain underexplored and weakly integrated into unified models. Many studies also rely on single-institution or context-specific datasets, limiting model generalizability and scalability. Additionally, high predictive accuracy is often emphasized without sufficient focus on interpretability, real-time intervention mechanisms, and policy-level applicability, highlighting the need for comprehensive, explainable, and school-focused EDM frameworks.

4. Systematic result analysis

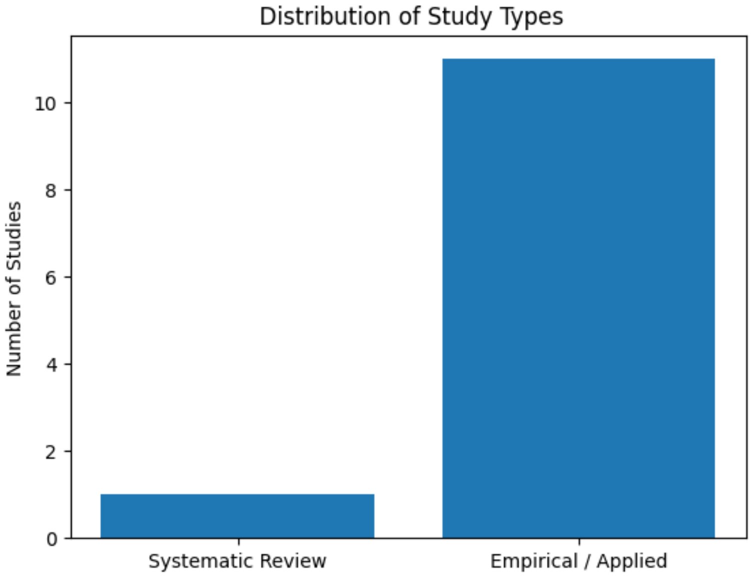

Figure 1: Distribution of Study Types

Figure 1 illustrates the distribution of study types in the reviewed literature. The results show that empirical and applied studies overwhelmingly dominate, while systematic literature reviews are scarce, indicating a need for more integrative and theory-driven syntheses of EDM research.

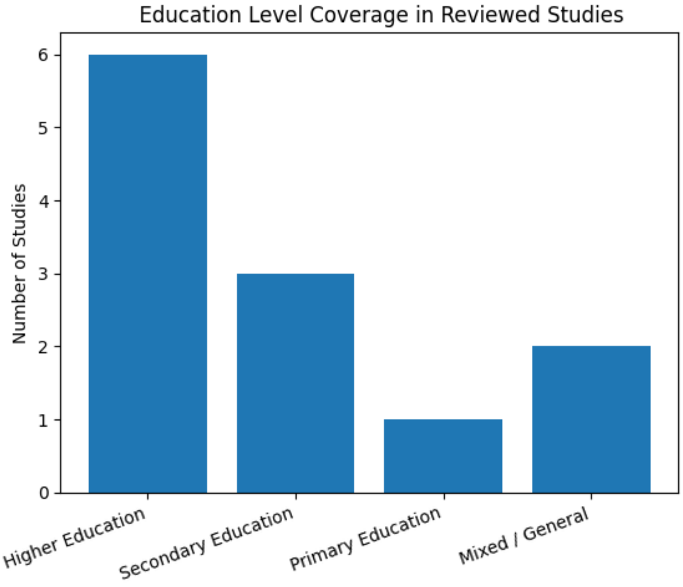

Figure 2: Education Level Coverage in Reviewed Studies

Figure 2 presents the education levels addressed by the selected studies. The majority of research focuses on higher education, with significantly fewer studies targeting secondary education and only one study addressing primary education. This imbalance reveals a substantial research gap in applying EDM techniques at early educational stages.

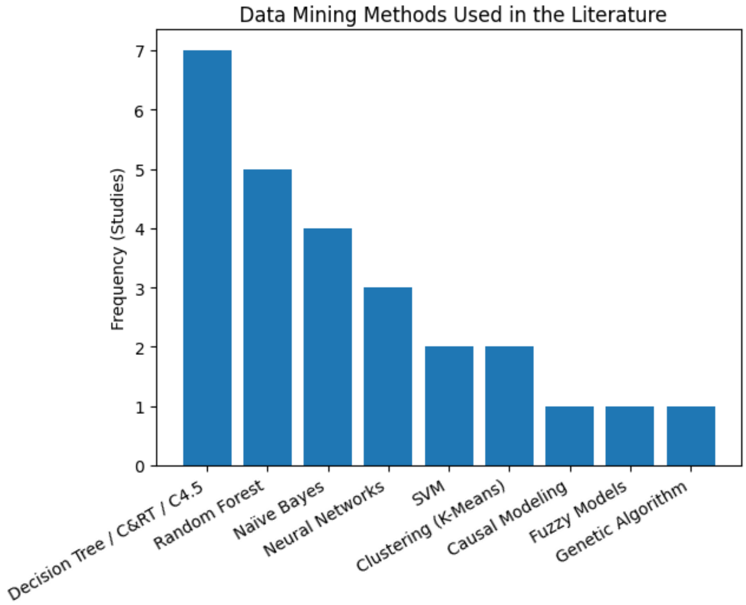

Figure 3: Data Mining Methods Used in the Literature

Figure 3 shows the frequency of data mining methods applied across studies. Decision tree-based models are the most commonly used, followed by random forest and naïve Bayes. More advanced techniques such as causal modeling, fuzzy models, and genetic algorithms appear infrequently, indicating opportunities for methodological innovation and hybrid model development.

5. Conclusion

This study presented a systematic review and comparative analysis of educational data mining techniques used for student performance prediction. The findings show that decision tree–based models and ensemble methods such as random forest dominate the literature due to their strong predictive performance and adaptability to educational data. Simpler models, including naïve Bayes and regression approaches, continue to be valued for their interpretability, while advanced methods such as causal modeling, fuzzy logic, and optimization techniques remain relatively underexplored. The review also highlights a strong concentration of studies in higher education, with limited empirical evidence from primary and secondary education contexts. Additional challenges include data quality issues, class imbalance, lack of standardized evaluation frameworks, and insufficient linkage between prediction models and actionable interventions. Addressing these limitations is essential for translating EDM research into meaningful educational impact. Future work should focus on developing explainable, generalizable, and intervention-oriented EDM frameworks validated across diverse educational levels and real-world learning environments.

References

1. Tosun, Selma, and Dilara Bakan Kalaycıoğlu. "Data mining approach for prediction of academic success in open and distance education." *Journal of Educational Technology and Online Learning* 7, no. 2 (2024): 168-176.
2. Choi, Wan-Chong, Chan-Tong Lam, and António José Mendes. "A systematic literature review on performance prediction in learning programming using educational data mining." In *2023 IEEE Frontiers in Education Conference (FIE)*, pp. 1-9. IEEE, 2023.
3. Khairy, Dalia, Nouf Alharbi, Mohamed A. Amasha, Marwa F. Areed, Salem Alkhalaf, and Rania A. Abougalala. "Prediction of student exam performance using data mining classification algorithms." *Education and Information Technologies* 29, no. 16 (2024): 21621-21645.
4. Wang, Yu-Jie, Chang-Lei Gao, and Xin-Dong Ye. "A data-driven precision teaching intervention mechanism to improve secondary school students' learning effectiveness." *Education and Information Technologies* 29, no. 9 (2024): 11645-11673.

5. Silva Filho, Rogério Luiz Cardoso, Kellyton Brito, and Paulo Jorge Leitão Adeodato. "Leveraging causal reasoning in educational data mining: an analysis of Brazilian secondary education." *Applied Sciences* 13, no. 8 (2023): 5198.
6. Yang, Lihui, Xiuhong Qin, and Wenhong Liu. "High quality management of higher education based on data mining." *International Journal of Business Intelligence and Data Mining* 25, no. 3-4 (2024): 424-450.
7. Putri, I.N., Maharani, P., Kurniawati, Y.E. and Putri, R.A., 2024, November. Application of Data Mining to Predict Student Learning Outcomes in Padang Panjang. In *2024 4th International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME)* (pp. 1-7). IEEE.
8. Nagarajan, Harikumar, Zaid Alsalami, Shweta Dhareshwar, K. Sandhya, and Punitha Palanisamy. "Predicting academic performance of students using modified decision tree based genetic algorithm." In *2024 Second International Conference on Data Science and Information System (ICDSIS)*, pp. 1-5. IEEE, 2024.
9. Arief, M. Habibullah, and Martiana Kholila Fadhil. "Educational Data Mining for Student Academic Performance Analysis." Jurnal Teknologi Informasi dan Terapan 11, no. 2 (2024): 83-90.
10. Chytas, Konstantinos, Anastasios Tsolakidis, Evangelia Triperina, and Christos Skourlas. "Educational data mining in the academic setting: employing the data produced by blended learning to ameliorate the learning process." Data Technologies and Applications 57, no. 3 (2023): 366-384.
11. Gök, B., Akkuş, E.B., Kavak, G. and KASAP, P.Y., 2023. Investigation of the Variables Affecting Primary School Teachers' State of Anxiety and Motivation in Mathematics Teaching through Data Mining Methods. Current Psychology, 42(31), pp.27678-27693.
12. Liu, Shasha, and Hua Jiang. "Research on the Cultivation of Innovative Ability of College Physical Education Students Based on Data Mining Technology." The Educational Review, USA 8, no. 5 (2024).