BOHR International Journal of Smart Computing and Information Technology
2026, Vol. 7, No. 1, pp. 22–28
DOI: 10.54646/bijscit.2026.52
www.bohrpub.com

REVIEW

# Data mining in education: a comprehensive review

**C. M. M. Mansoor and S. L. Fathima Ruksana**[*†]

Department of Information Technology, Faculty of Arts and Culture, South Eastern University of Sri Lanka, Oluvil, Sri Lanka

[*]**Correspondence:**
S. L. Fathima Ruksana,
fathimaruksanasl@gmail.com

[†]**ORCID:**
S. L. Fathima Ruksana
0009-0003-6956-973X

**Received:** 21 January 2026; **Accepted:** 27 May 2026; **Published:** 01 July 2026

Schools and colleges collect vast amounts of data from learning management systems, online learning platforms, assessment systems, and administration databases. Educational Data Mining (EDM) is a newly developing research area that is associated with the application of data mining techniques to educational data, the extraction of useful information from it, and the use of the information to enhance both instruction and learning as well as the decision-making processes of an educational institution. This review paper aims to present a comprehensive survey of the area of data mining applications, techniques, problems, and future directions in the educational sector. The current research focused on secondary data analysis of past studies grounded in the EDM and learning analytics (LA) field. Sources such as various published papers, journal articles, and conference papers were reviewed to find commonly used techniques, applications, and new trends in the field. We review some techniques like classification, clustering, association rule mining, predictive modeling, and LA. The results show that the role of this EDM in predicting students' learning abilities, dropout rate detection, recommendation systems for student learning, and tools in curriculum designing has an important role. Moreover, data mining is used in the field of intelligent tutoring systems and in the analysis of the performance of institutions. The paper also identifies some significant issues in the field of EDM, such as data quality, privacy and ethical issues, and the requirement for understandable and accurate algorithms. In addition, new areas of research and development in educational technology are identified, including real-time analytics, multimodal data analysis, artificial intelligence (AI), and explainable AI. To summarize, by converting raw educational data into insightful information, EDM can be helpful to enhance the teaching, learning, and decision-making processes in educational activities and policies. The study underscores the need to use ethical and safe education data that could continue in the future for educational development while, at the same time, using effective data mining techniques.

**Keywords:** educational data mining, learning analytics, student performance, dropout prediction, classification, clustering, recommender systems, higher education

## Introduction

According to the new advancement in technology in education, we are able to learn about the learners as well as the environment related to the learning. Nowadays, students' behavior and their performance in learning are also the interactions that are recorded and traced from where they used to learn the platform, such as learning management systems, online quizzes, e-portfolios, massive open online courses (MOOCs), and student information systems (1, 2). According to all these data sources, it is able to provide a variable opportunity to mine the data easily, rather than using traditional methods. The complex data set can be transformed with the help of data mining as well as machine learning and can also provide meaningful information as well as help the

process of teaching and learning and also decision-making related to education.

According to the educational data, mining can be used in various fields such as education, computer science, psychology, and statistics. And also, education-related important questions can be addressed by educational data mining (EDM) that uses the technique on the data sets in machine learning (1). Also, for the purpose of understanding the learners as well as the environment where the learning occurs, we are able to do EDM by using some methods that are used in a disciplined way to use the unique data (3). With the help of learning analytics (LA), we can collect data, analyze it, and also make reports about the learners to improve their learning process (2, 4).

Algorithm innovation, as well as methodological innovation, is emphasized by EDM, and it also helps to provide the data that is related to educational insights. According to these, both EDM and data analytics are used for common aims and objectives and also help to utilize the data that help to improve teaching, as well as teaching in an environment that is too data-rich. According to this study, this helps to illustrate the beginning as well as the development of EDM and also types of EDM that now also include the techniques that are related to data mining and applications of data mining related to educational institutional analytics recommendation, clustering, and prediction. It also provides the challenges that are related to EDM, as well as issues related to ethics, and provides future research directions that are helpful to make advancements in EDM in the future.

# Aims and objectives

## Aim of the study

This study aims at giving a detailed overview of data mining in the field of education and at exploring the potential of data mining techniques for enhancing teaching and learning, student performance, and educational decision-making.

## Objectives of the study

1. To find and recognize the primary data mining techniques in education.
2. To understand the applications of EDM, including student performance prediction, dropout detection, personalized learning, and intelligent tutoring systems.
3. To explore the application of LA and data mining in institutional decision-making and curriculum development.
4. To gain insight into the problems related to EDM, such as data quality, privacy, ethical issues, and limitations in the algorithm.
5. To explore recent trends and future directions in EDM, including artificial intelligence (AI), explainable AI, real-time analytics, and multimodal data analysis.
6. Assess the impact of learning outcomes and EDM on the enhancement of teaching effectiveness and the enhancement of learning outcomes.

# Materials and methods

## Study design

The aim of this study was carried out as a review-based research study related to the EDM and LA field. This review was designed to find, summarize, analyze and categorize the knowledge, techniques, applications, challenges, and future trends of data mining in education. These were the EDM and LA literature that was collected from academic journals, conference proceedings, research papers, books, and online academic literature databases. Keyword strategy was used in the literature search, and the following keywords were used: EDM, LA, data mining in education, student performance prediction, intelligent tutoring systems, machine learning in education, and analytics articles were chosen for educational application in using data mining techniques.

The study included:

- Peer-reviewed articles published in conferences and journals
- EDM and LA -related studies.
- Education-related research on data mining methods, applications, problems, and future research needs
- English language publications

The qualitative review and analysis emerged for selected studies. The analysis analyzed the most common data mining applications and corresponding technologies in the areas of teaching and learning, student performance, and institutional decision-making, including applications of classification, clustering, association rule mining, and predictive analytics.

# Data mining and educational data mining

## Data mining overview

Data mining provides meaningful knowledge from large databases, which helps to provide meaningful information.

Also, the things that we had unknown earlier, we are also able to get patterns from the last data sets that can be used usefully (5). There are many fields that are getting advantages from the use of data mining, such as finance, web analytics, bioinformatics, and marketing, which are using tasks such as classification, regression, clustering, association rule mining, sequential pattern mining, and anomaly reduction. According to that, the data that are related to education differ from other datasets like financial transactions. According to the educational data, having multi-level hierarchical structures and also having the sequences in a temporal order that are related to learning helps to reflect the activities. According to this special analytical method, it should have these unique features and should have modeling with careful approaches that are helpful to the educational field (3).

### Definition and evolution of educational data mining

With the help of an intelligent tutoring system and computer-assisted content, EDM has begun with the purpose of analyzing the files that are related to the log to get insights about the students who interact with digital learning. According to that, the field of EDM has developed and has become established by the Journal of EDM and also by the Annual International Conference on EDM. By the comprehensive review of EDM, it is able to perform a lot of tasks, such as pain analysis techniques can also to use for many applications, such as predicting the performance of students, identifying groups of students, and learn the behaviors of students and learners (1). In 2009, the state of EDM was formed and has separated EDM from traditional data mining because of some challenges (3). Also, according to the reason for the advancement of technology, many broader scopes have been brought by many scholars related to EDM in an environment that has big data, MOOCs, and dashboards for LA that are merged with LA (2, 6).

## Educational data and data mining techniques

### Types of educational data

According to the environment with an educational background, having a variety of data that can be gained from various sources helps to analyze the challenges of the complexity of learners (1, 2). According to clickstream logs, page views, and time spent on task attempts, the problem is hint requests and posts on the forums. Also, the check messages are the types of data that can be included in the records. Also, the data that are related to the exam result grades as well as the scores of the quiz can be included as data. Also, other data related to the personal information about the students, such as gender; prior grade point average (GPA) status related to the socioeconomic status; and data related to the programs that are related to the academics. Also, some data that are related to the administrative works can be included, like the histories related to the student and enrollment; also, there are student attendance reports, their registrations, and, of course, the progress that is related to the degree, which can be considered as data related to the administration. And also, according to the vast improvement of the technology, we are able to find much data with the help of self-reports and can also use the sensors to get the data as well as use the measures related to the video. Also, data can be in three structures, like table-wise, and can be structured as logs as well. It is called semi-structured, and it can also be unstructured. Text data, video-related data, and audio-related data can be considered as unstructured data. Also, the help of EDM techniques can enable working with many types of data that are important to working with the challenges related to the information that is missing, as well as noise and the bias of potential.

### Core data mining tasks in education

There are many categories in the tasks of EDM that are important to understand about the learning process and make analyses. Prediction is the first category. According to that, we can predict data related to the ranks of the students and also the risks that are related to the dropouts, as well as the students' action data related to the future. It helps to build the regression models as well to get outcomes. Next, according to the method of clustering, it helps to identify the groups that are naturally formed by the students who have experiences in their behaviors and levels of performance. Next, according to the third category of relationship mining, we are able to use the techniques of association rule mining, correlation analysis, and frequent share pattern mining to gain deep insights related to the connections and integration between the variable examples and mistakes that are commonly patterned. The next category is discovery with models. Also, it has models like knowledge-tracing algorithms. It helps to make modern analysis and get deep information regarding the behavior related to learning. Finally, according to the distillation of data for human judgment, it helps to make translations of complex data by the use of tools like the visualization dashboard and other tools, through which one is able to interpret the information effectively and able to make decisions in an effective manner (1).

### Common algorithms and techniques

There are many algorithms used in EDM. All the algorithms are related to machine learning, but according to their unique characteristics, they can be used for each purpose (1, 6, 7).

According to that, classification and regression trees (CART), Random Forest, Logistic Regression, Naive Bayes, Support Vector Machines, K-Nearest Neighbors (KNN), Artificial Neural Networks, and Gradient Boosting Models can be used for the task of prediction, can also be used for classification, and can also be used for regression. Also, according to the clustering, it mostly depends on the algorithms, such as k-means, hierarchical clustering, and density-based methods, used by the students to align them with the behaviors of those who have the same profiles, for the purpose of identifying the relationship as well as the patterns of sequential behaviors that can be captured by the techniques of a priori frequent pattern (FP) growth and sequence mining used for the behaviors that are occurring repeatedly. For the purpose of time series and for the model learning process, the methods of hidden Markov models, Bayesian knowledge tracing, and deep recurrent neural networks can be commonly used. Furthermore, for the purpose of analyzing the discussion forums, the use of methods like sentiment analysis, topic modeling, and text classification can be used for text mining and natural language processing purposes. Being able to write feedback through these methods, as well as these algorithms, is important to EDM (1, 6, 7).

## Applications of educational data mining

There are many ways of having the EDM applied in the educational setting to identify the performance of the students as well as the learners at risk by the help of the personal information as well as the grades of the students and the data that are related to the learning management system (LMS) related to the interaction, by the use of online assessments (7–10). According to the main studies that are illustrating that, with the help of the algorithms related to the classification, such as decision trees, logistic regression, support vector machine (SVM), and ensemble methods, they are providing an important role to predict the grades of the code as well as the risk related to the dropouts of the learners. Learning outcomes can be predicted through the classification algorithm, and also with the help of early engagement patterns as well as temporal features, extra information can be included (7–9). These analysis can predict the timely risk that is related to the students, as well as get the feedback, and make adjustments to the courses according to these warnings that the students' system can understand. With the help of supervised learning, it can predict the behaviors that are related to gaming the system as well as off-task activity and states of emotions like confusion or frustration (3).

By the use of recommender systems and adaptive learning, we can use data mining in the educational field in a customized manner. There are many techniques that can be used for these purposes, like collaborative filtering, content-based recommendation, and hybrid models; suggesting videos, reading, or exercise can match these requirements of the students as well as their preferences in learning (1, 6). With the help of adaptive sequencing of content and by the use of clustering, we can estimate the knowledge tracing as well as mastery models. With the help of the clustering method, we can find the groups that have the same strategies. And also, the data mining can be used at the institutional level for mining the data related to the evaluation related to the curriculum, design the assessment as well as plan the academic activities with the help of course progress as well as trends in enrollment and patterns of program completion. By the use of association rule mining, it is able to combine the course that has a link to student success as well as help with student success and help to enhance the curriculum guidance. According to the large institutions that are related to the online learning environment, such as MOOCs, they are able to use LA with EDM (2, 4), and also those who use the facilities of the dashboard as well as visualization are able to learn the process and can monitor the process.

## Challenges and ethical issues

According to EDM, there are many challenges to using data mining in the field of education. According to that, there is a major problem with the quality of data and its integration. According to educational data derived from platforms like LMS and student information systems, as well as other tools, student information can be incomplete and sometimes appear messy, scattered across various systems. The main reason for this is that the system uses different types of formats. Also, the use of identifiers that are inconsistent can make it more challenging to combine the data and may lead to errors. Also, the problem of generalization is one of the major problems in data mining because having a model in one institution cannot be able to be used in another place, and this may be one of the challenges to EDM (2, 6).

Also, there is another challenge, which is interpretability. It means that by the use of some machine learning, one can be able to predict accurate outputs. They can function like a black box. It means if some students are at risk normally, it will tell them they are at risk, but it won't tell them the action needed to take for that risk. Also, according to these things, explainable AI techniques and models can be able to use the theories related to the educational concept. According to that, the learners and professionals and researchers and teachers can be able to get a full understanding (3, 6). According to that, there are some challenges related to privacy, security, and ethical concerns because there is some data related to the students in the datasets that can be misused; and violate policies like family educational rights and privacy act (FERPA) and general data protection regulation (GDPR) for data governance and privacy (4).

## Result

The literature review shows that EDM has emerged as a significant research area to enhance teaching, learning, and educational decision-making via the analysis of educational data. The review found several core data mining techniques commonly used in education, such as classification, clustering, association rule mining, predictive analytics, regression, and text mining. These were mostly classification algorithms such as decision trees, random forest, logistic regression, SVM, Naïve Bayes, and KNN for predicting student academic performance and finding learners at risk of failure or dropout. The review also noted that EDM has many educational applications. Data mining techniques have been extensively applied in the field of education to identify students who require early academic intervention using predictive models. To facilitate personalized teaching, clustering techniques are used to group learners according to their learning habits and learning levels. Data mining techniques are used in recommendation systems and intelligent tutoring systems to provide adaptive learning resources and personalized feedback to the students, which improves student engagement and increases learning outcomes. Findings revealed that LA and EDM significantly contribute to institutional decision-making. Educational institutions use analytical insights to review curriculum, plan academics, develop student retention strategies, allocate resources, and assure quality. Interactive dashboards and visualization tools also serve to help administrators and educators track student progress and institutional performance. Across the reviewed studies, a number of challenges were persistently reported. These include data quality issues, challenges in data integration, privacy and ethical concerns, algorithmic bias, limited model interpretability, and the need to comply with data protection regulations. These problems are still major barriers to the widespread use of EDM. Finally, the review identified emergent trends impacting the future of EDM: AI, XAI, real-time learning analytics, multimodal data analysis, and human-centered intelligent educational systems. All in all, these developments suggest that EDM has huge potential to improve the effectiveness of learning, to personalize education, and to support evidence-based educational policies, while stressing the ethical and responsible use of educational data.

## Discussion

This review study examined the applications, techniques, challenges, and future directions of EDM. The findings reveal that educational institutions increasingly rely on data mining techniques to analyze large volumes of educational data generated through learning management systems, online platforms, and institutional databases. The study identified classification, clustering, association rule mining, and predictive analytics as the most commonly used techniques in educational environments.

One of the important findings of this review is that EDM significantly supports student performance prediction and dropout identification. Similar findings have been reported in previous studies, where predictive models were used to identify academically at-risk students and provide early interventions to improve learning outcomes. In addition, clustering methods help categorize students based on learning behaviors and performance patterns, enabling personalized learning approaches.

The review also highlights the importance of LA and intelligent tutoring systems in modern education. Recommendation systems and adaptive learning platforms support personalized education by providing customized learning materials according to students' needs and abilities. These findings are consistent with existing literature that emphasizes the role of data-driven educational systems in improving learner engagement and academic success.

Another important aspect identified in this study is the contribution of EDM to institutional decision-making. Educational institutions can use analytical insights for curriculum development, resource management, academic planning, and quality assurance processes. Therefore, EDM not only benefits students and educators but also supports administrative and strategic educational decisions.

Despite these advantages, several challenges related to EDM were identified. Data privacy, ethical concerns, and data quality issues remain major limitations in implementing effective EDM systems. In addition, some machine learning and AI algorithms lack transparency and interpretability, which may reduce trust in automated educational decisions. These findings raise important questions regarding how educational institutions can ensure fairness, privacy protection, and ethical use of student data.

The study also discusses emerging trends and future directions in EDM, including real-time analytics, multimodal LA, AI, and explainable AI. These technologies are expected to enhance adaptive learning systems and improve educational effectiveness in the future.

One of the strengths of this review is its comprehensive coverage of EDM techniques, applications, challenges, and future developments. However, the study has some limitations, as it is based only on secondary data from previously published literature and does not include experimental or quantitative analysis. Future research can focus on empirical investigations and the development of more secure, transparent, and explainable EDM models.

Overall, this study demonstrates that EDM has significant potential to improve teaching, learning, and educational management while supporting evidence-based educational policies and decision-making.

## Future research directions

According to future research on EDM, advanced technological improvement can be used for analysis, as well as for collecting data, and can also be used in real-world applications by the use of real-time adaptive analysis, we are able to respond according to the behaviors of the students immediately. According to this method, in the future, it will be able to provide an on-time response to the behaviors of the students. So, according to the advancement of technology, we are able to use multi-modal data, such as text, video, sensor-related data, and biometric information, which helps to make the student's data richer as well as helps to provide privacy to the personal data that is related to the student. Also, use of explainable and human-centered approaches that are important to the teachers as well as the educators to work collaboratively will enable to understand why the predictions are made. Also, it can be used as an interpretation tool that helps to make algorithms as well as helps to improve the usefulness of EDM and can also build trust (2, 8).

There is another use of cross-institutional and cross-cultural studies that are able to work in different educational systems and cultures, as well as with the population that needs to learn (2, 8, 10). And also, the integration of EDM with learning design or educational theory can work as a technical tool and provide analytic support to the goals that have to do with the EDM, as well as build the ethical framework and governance structure in the future to make the EDM private. It also helps to make decisions and can provide protection to the EDM, and by having all these future directions, one can get EDM to have features like accuracy, fairness, and ethics, as well as the ability to make an influence in real-world applications.

## Conclusion

With the help of EDM, it is able to provide a good development in the interdisciplinary field as well, and it also helps to provide an understanding of the tools that are valuable for EDM and also helps to improve learning skills. Also, with the help of providing the prediction related to the data in the field of education, we can predict the performance related to the students as well as the learners who are having difficulties in learning; see the patterns that students engage in; help to analyze the customized learning activities; and make decisions in a timely manner related to the educational scenarios. According to this, the EDM is still facing some challenges, like in the quality of data as well as the interpretability of models and the ethics related to the data of the students that are governed. Also, privacy as well as inequality are considered challenges related to EDM, and for the best technical development, it has used EDM, which means EDM with LA can get a good advantage in real-world applications and meaningful data.

EDM should be targeted to predict accuracy, as well as target the human-centered and, theory-driven, and be ethically responsible. Analytics should be targeted by helping to provide accurate predictions, and be able to provide a proper understanding as well as good, meaningful insight to the people. Teachers, learners, educators, and professionals, as well as data mining with strong governance, can build up a good learning environment that is customized and effective and ultimately achieve good outcomes.

## Funding

The authors declare that no financial support was received for the research, authorship, and/or publication of this article.

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## References

1. Romero C, Ventura S. Educational data mining: a review of the state of the art. *IEEE Trans Syst Man Cybern C Appl Rev.* (2010) 40(6):601–18. Available online at: https://scispace.com/papers/educational-data-mining-a-review-of-the-state-of-the-art-51aglgywl8?utm_source=chatgpt.com
2. Romero C, Ventura S. *Educational Data Mining and Learning Analytics: An Updated Survey*. Preprint (2020). Available online at: https://arxiv.org/abs/2402.07956?utm_source=chatgpt.com
3. Baker RSJD, Yacef K. The state of educational data mining in 2009: a review and future visions. *J Educ Data Min.* (2009) 1(1):3–17. Available online at: https://learninganalytics.upenn.edu/ryanbaker/BakerYacef.pdf?utm_source=chatgpt.com
4. Long P, Siemens G. What is learning analytics? *Proceedings of the 1st International Conference on Learning Analytics and Knowledge (LAK 2011)*. Banff, AB, Canada: Association for Computing Machinery (2011). Available online at: https://dl.acm.org/action/showFmPdf?doi=10.1145%2F2090116&utm_source=chatgpt.com
5. Witten IH, Frank E. *Data Mining: Practical Machine Learning Tools and Techniques*. San Francisco, CA: Morgan Kaufmann (2005). Available online at: https://dixieching.wordpress.com/2011/06/26/the-state-of-educational-data-mining-in-2009-a-review-and-future-visions-baker-r-s-j-d-yacef-k-2009/?utm_source=chatgpt.com
6. Luengo V. Educational assessment, educational data mining, and learning analytics. *Encyclopedia of Education and Information Technologies*. Cham, Switzerland: Springer (2020). Available online at: https://link.springer.com/rwe/10.1007/978-3-030-10576-1_43?utm_source=chatgpt.com

7. Kotsiantis SB, Patriarcheas K, Xenos M. A review on predicting students' performance using data mining techniques. *Procedia Comput Sci.* (2010) 1(2):143–50. Available online at: https://www.sciencedirect.com/science/article/pii/S1877050915036182?utm_source=chatgpt.com
8. Alammary A. Predicting student performance using data mining and learning analytics: a survey. *Appl Sci.* (2021) 11(1):237. Available online at: https://www.mdpi.com/2076-3417/11/1/237?utm_source=chatgpt.com
9. Albreiki B, Zaki N, Alashwal H. A systematic literature review of student performance prediction using machine learning techniques. *Educ Sci.* (2021) 11(9):552. Available online at: https://files.eric.ed.gov/fulltext/EJ1314372.pdf?utm_source=chatgpt.com
10. Lou Y, Colvin KF. Performance prediction using educational data mining techniques: a recent study. *Discov Educ.* (2025) 4:112. Available online at: https://link.springer.com/article/10.1007/s44217-025-00502-w?utm_source=chatgpt.com