A Review of Educational Data Mining in Higher Education System

Balwinder Kaur
DCSA, Panjab University, Chandigarh, India

Received: Feb. 19, 2018 Accepted: March 14, 2018

Abstract: *The quality of education directly influences a nation's financial system and various industries, highlighting the importance of assessing and enhancing student performance in educational institutions. Educational Data Mining (EDM) offers valuable insights into student learning behaviors and academic achievements, aiding in the identification of at-risk students and improving overall educational outcomes. This review discusses various data mining methods such as linear and nonlinear regression, classification, clustering, and relationship mining, along with their applications in educational settings. Additionally, the review analyzes selected works from 2011 to 2017, presenting objectives, methodologies, findings, and limitations of each study. Challenges in EDM, including the lack of specialized tools, adaptation of algorithms, and privacy concerns, are also discussed. The review emphasizes the importance of integrating EDM tools into educational systems and the need for further research to generalize results and validate findings across diverse contexts.*

Keywords: *Educational Data Mining, Higher Education, Data Mining Techniques, Student Performance, Challenges.*

Introduction:

The country's financial system is directly influenced by the quality of education students receive, which in turn impacts various industries. The excellence of educational institutions is gauged by the success rates of their students, and the effectiveness of these institutions is assessed by the rate at which they retain students who are at risk. Assessing student academic performance involves considering various factors such as individual, social, and psychological aspects. This evaluation helps identify students who may be at risk, enabling timely intervention by management.

Student academic performance is evaluated based on socioeconomic background and previous academic achievements, utilizing educational data mining techniques. This process, known as supervised learning, involves determining classes before analysing the data. By classifying data into predefined sets, educational data mining facilitates decision-making processes, such as identifying at-risk students, reducing dropout rates, and enhancing learning outcomes [1].

Educational data mining encompasses a range of computational and psychological methods for understanding how students learn. It involves applying various data mining techniques such as clustering, rule mining, and neural networks to uncover hidden knowledge within educational databases. This knowledge extraction process enhances decision-making in education, enabling institutions to better support students and improve overall productivity [2].

By measuring student performance, placement officers can provide tailored guidance, ensuring students make informed career decisions. Awareness programs conducted by educational planners can mitigate risks for students and enhance productivity. Educational data mining techniques offer insights into diverse learning needs among student groups, facilitating targeted interventions and personalized support for students [3].

The primary objectives of this study have an insight on various educational dataming techniques, review of studies conducted between year 2011 to 2017 and to understand various challenges related to educational data mining discussed in sections below.

Educational Data Mining Techniques

EDM not apply only data mining techniques Classification, clustering, and association analysis, but also apply methods and techniques drawn from the variety of areas related to EDM (statistics, machine learning, text mining, web log analysis, etc.). There are so many methods of educational data mining but all kind of methods lie in one of following categories [4, 5, 6,7, 8]:

Linear Regression: Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables by fitting a linear equation to the observed data. It assumes a linear relationship between the independent and dependent variables. Linear regression is used for predicting continuous outcomes [8].

Nonlinear Regression: Nonlinear regression is a method used to model the relationship between a dependent variable and one or more independent variables when the relationship is not linear. In nonlinear regression, the relationship between the variables is modelled using a nonlinear function. This allows for more complex relationships to be captured, such as exponential, logarithmic, or polynomial relationships [8].

Classification: Classification is a type of supervised learning task where the goal is to classify data into predefined categories or classes based on input features. Some common algorithms used for classification include [8, 9]:

Decision Trees: Decision trees are a popular method for classification. They work by partitioning the feature space into regions and assigning a class label to each region based on majority voting of the training examples within that region.

Naive Bayes Classification: Naive Bayes is a probabilistic classifier based on Bayes' theorem with the assumption of independence between features. Despite its simplicity, it often performs well in practice, especially for text classification tasks.

Generalized Linear Models (GLM): GLMs are a class of models that generalize linear regression to allow for non-normal error distributions and non-linear relationships between the dependent and independent variables. They include logistic regression for binary classification and multinomial logistic regression for multi-class classification.

Support Vector Machine (SVM): SVM is a powerful classification algorithm that finds the hyperplane that best separates the classes in the feature space. It can handle both linear and nonlinear classification tasks by using different kernel functions [8].

Clustering: In the clustering technique, the dataset is divided into various groups known as clusters. According to clustering phenomenon, the data points within one cluster should be more similar to each other and more dissimilar to data points in other clusters. There are two ways to initiate clustering algorithms: Firstly, to start the clustering algorithm with no prior assumptions, and secondly, to start the clustering algorithm with a prior postulate [8].

Relationship Mining: This technique is employed to discover relationships between variables in a dataset and encode them as rules for later use. Various types of relationship mining techniques exist, such as association rule mining (identifying any relationships between variables), sequential pattern mining (uncovering temporal associations between variables), correlation mining (identifying linear correlations between variables), and causal data mining (investigating causal relationships between variables). In EDM, relationship mining is utilized to identify connections between students' online activities and their final marks, as well as to model sequences of the learners' problem-solving activities [4][9].

Discovery with Models: The primary objective of this approach is to utilize a validated model of a phenomenon (using prediction, clustering, or knowledge engineering) as a component in further analyses, such as prediction or relationship mining. For example, it aids in identifying the relationships between students' behaviors and their characteristics [4][9].

Discussion on Selected Work

Several collections of reviewed papers have comprehensively covered crucial aspects of data mining in educational research [3, 4, 5, 6, 10, 11]. One review focused on the application of data mining techniques in educational systems from 1995 to 2005. Each system reviewed exhibited diverse data sources and objectives for knowledge discovery [10]. Another review centred on applying data mining techniques to e-learning problems. It explored the utilization of e-learning in assessing students' learning performance, their learning behavior, and evaluating learning materials. Additionally, there was a review conducted on the current trends in Educational Data Mining (EDM) and the shifts in paper topics over the years [3]. For a more in-depth review of each scholar's most significant studies and the type of educational tasks they were

addressing, one can refer to Romero and Ventura's work. In the subsequent table review of different studies carried in education data mining field between 2011 to 2017 have been presented:

Table 3: Review of Studies between 2011 and 2017

Ref	Year	Objective	Data Mining Technique/ Method	Findings	Limitations
12	2011	To justify the capabilities of data mining techniques in higher education, Study students' performance in courses using data mining methodologies, and help improve student division while identifying those needing special attention to reduce failure rates.	Classification Decision Tree	The main findings include the use of data mining techniques to study and improve students' performance in courses. The ability to identify students who may need special attention and reduce the failure rate in semester examinations.	The size of dataset used is very small. The study may lack generalizability due to the specific choice of only ID3 Decision Tree algorithm. Lack of pre-processing details.
13	2012	To compare the achievements of Computer Engineering Department students based on certain criteria using data mining techniques.	Multilayer Perceptron (MLP) with back propagation algorithm and Quinlan's C5 algorithm	Compared the achievements of Computer Engineering department students using data mining techniques, Decision tree algorithms showing higher prediction accuracy compared to artificial neural networks.	The lack generalizability due to the specific choice of data mining methodology (CRISP-DM). The paper does not thoroughly explore the impact of parameter tuning on model performance.
14	2013	The study objectives are academic objectives related to teaching and learning processes. Administrative objectives related to resource utilization and relationships with industry and academia.	Review Study	The main findings include the classification of EDM objectives. research trends from 1998 to 2012. A focus on behavioural identification of learners using the web.	Difficulty in maintaining the data warehouse due to exponential growth of data Issues with aligning and translating incremental educational data Challenges in optimal utilization of computing and human resources Critical need for scalable data management considering diverse storage locations and data platform heterogeneity Difficulty in assigning supervisors to students with similar research interests, impacting

					the applicability of project results Inability of models to predict accurate results in student modeling or academic planning due to uncertain errors
15	2013	To find patterns in available data for predicting students' performance at the university based on personal and pre-university characteristics. To evaluate the potential usefulness of different data mining classification algorithms for achieving the project goal and objectives.	Decision trees (J48 classifier), Bayesian classifiers (NaiveBayes and BayesNet)	The prediction rates of the classification algorithms applied to the university sample data were not remarkable, ranging between 52-67%. The classifiers performed differently for the five classes, with varying levels of accuracy. Attributes such as University Admission Score and Number of Failures at the first-year university exams were identified as influential factors in the classification process.	The study is in the initial stages and further steps need to be defined for the university data mining project. Acknowledgment of the need for possible transformations of the dataset and tuning of classification algorithms' parameters for more accurate results and extraction of important knowledge.
16	2014	To explore the application of data mining techniques in educational databases. Predict career options for high school students. Predict potentially violent behavior among students.	Decision tree algorithms (ID3, C4.5, CHAID)	Showcases the use of data mining techniques, particularly the ID3 algorithm, to suggest career options for high school students. Predict potentially violent behavior among students. ID3 algorithm was found to outperform other decision trees in predicting violent behavior and the need for counseling. The paper emphasizes the significance of data mining in educational databases for valuable insights beyond academic performance.	Potential limitations in the preprocessing techniques used Limitation in the number of attributes considered Limitation in the techniques used for analysis
17	2015	To predict students' performance in semester exams, determine the grades students may obtain in their end	Naive Bayes classifier, Bayesian Classification	The experimental results demonstrate that predicting students' performance in semester exams can	Lack of new or innovative approaches Generalizability of results not discussed

		semester results. Assist educational institutions, teachers, and students in taking necessary actions to improve student outcomes.		be achieved by utilizing their previous semester marks and overall performance in various activities of the current semester. The prediction model implemented in the study can assist educational stakeholders in taking necessary actions to improve students' results.	
18	2015	To predict student academic performance using a classification model based on Deep Learning and to automatically learn multiple levels of representation to improve the prediction accuracy.	unsupervised learning algorithm and classification algorithms including NaiveBayes, Multilayer Perception (MLP), and SVM.	The study developed a classification model using Deep Learning to predict student performance effectively and demonstrated the applicability of the proposed method in academic pre-warning mechanisms.	Complexity of measuring academic performance due to diverse factors and variables. gather more training samples, and use temporal information.
19	2015	To develop a system for predicting students' course grades for the next enrollment term, assist students in choosing suitable majors and course schedules. To provide information to advisors and educators to identify students in need of additional attention.	Factorization Machine (FM)	The factorization machine (FM) model achieved the lowest prediction error and outperformed all other methods by a wide margin. The FM model was successful in accurately predicting student grades for both cold-start and non-cold-start scenarios. The study highlights the effectiveness of leveraging a combination of methods for next-term student grade prediction.	The system seems slow to adapt to shifting characteristics in the student population as a whole. Incorporating side information is suggested to address the problem. Experimenting with pass vs. fail grade prediction is planned for the future. Accurately classifying failing dyads is crucial for improving predictive performance.
20	2016	The study objectives are to utilize predictive modeling methods to identify at-risk students early in the semester, compare different predictive methods in a course using standards-based grading, and	Naive Bayes Classifier (NBC), K-Nearest Neighbor (KNN), and Support Vector Machine	The study emphasizes the persistent challenges in student retention despite past efforts. The Naive Bayes Classifier and Ensemble model were the most successful in identifying	Lack of generalizability, done related to specific questions only. Lack considering various factors influencing student performance, such

		determine the best prediction method for identifying at-risk students while showcasing the accuracy and usability of course-specific standards-based prediction models.	(SVM)	at-risk students among the tested methods. The Ensemble model was the best at identifying at-risk students.	as motivation, personal circumstances, and learning styles, which are not fully captured by the predictive models.
21	2017	Evaluate the effectiveness of EDM techniques to early identify students likely to fail Assess if data preprocessing can increase the effectiveness of EDM techniques Determine if fine-tuning of algorithms can further increase the effectiveness of EDM techniques Identify the most effective EDM techniques for early identification of students likely to fail	Neural Networks, Decision Tree, Support Vector Machine (SVM), Naive Bayes	EDM techniques are effective in early identification of students likely to fail, with SVM showing the highest effectiveness. Data preprocessing and algorithms fine-tuning are crucial for improving effectiveness.	Results are not generalizable as they are based on data from only one university The study used a specific measure (f-measure) which may not capture all aspects of effectiveness Manual fine-tuning of techniques could introduce bias and impact effectiveness

Challenges in EDM

Educational Data Mining (EDM) encompasses various fields such as Intelligent Tutoring Systems (ITS), e-learning, web mining, and data mining. However, to effectively utilize EDM, it must consider both the pedagogical aspects of learners and the system [10]. In educational systems, different users interpret results and data differently based on their objectives, presenting challenges within the field of EDM. While progress has been made in resolving some of these challenges, more work remains to be done. Some of the key challenges include:

1. Lack of tools specifically designed for EDM research.
2. Need for user-friendly EDM tools accessible to both educators and non-expert users, featuring powerful yet flexible interfaces with intuitive visualization capabilities [11, 10].
3. Adaptation of existing data mining (DM) tools to handle data from educational environments, requiring standardization of data and preprocessing and post-processing stages [11, 10].
4. Modification of DM algorithms for educational data, which can significantly enhance pedagogical decisions and instructional design [11, 10].
5. Development of tools for improving open educational resources and creating generalized tools usable by experts and non-experts alike [22].
6. Requirement for tools to protect individuals' privacy [22].
7. Challenges in managing and monitoring the incremental nature of educational data, and in optimizing human and computing resources [14].
8. Uncertainty arising from unseen errors, necessitating the acknowledgment that no model can predict outcomes with 100% accuracy [14].

While significant work is underway, there is room for further improvement. EDM should not only be utilized by researchers for theoretical purposes but also implemented practically by educators and educational institutions. Integration of EDM tools into computer-based educational systems would enable educators to select DM algorithms and apply them with appropriate parameters. These tools should facilitate all stages of the EDM process for educators. Additionally, it is crucial to generalize results obtained from EDM research and conduct studies to validate these results across broader contexts.

Conclusion and Future Work

Educational Data Mining (EDM) holds immense potential in revolutionizing the higher education system by providing valuable insights into student learning behaviors and academic performance. By utilizing various data mining techniques such as classification, clustering, and relationship mining, educational institutions

can identify at-risk students, reduce dropout rates, and enhance learning outcomes. The discussion on selected works showcases the diverse applications of EDM in predicting student performance, suggesting career options, and early identification of students likely to fail.

However, the field of EDM also faces several challenges, including the lack of specialized tools, adaptation of existing algorithms, and ensuring privacy protection. Despite these challenges, significant progress has been made, and there is a clear path forward for further improvement. Future directions in EDM research should focus on developing user-friendly tools accessible to both educators and non-experts, standardizing data processing stages, and enhancing privacy protection measures.

Moreover, it is essential to validate the findings of EDM research across broader contexts and integrate EDM tools into educational systems to facilitate practical implementation by educators. By addressing these challenges and focusing on future directions, EDM can significantly contribute to improving the quality of education and student outcomes in higher education institutions.

References:

[1] K. R. Kavyashree, Lakshmi Durga – "A Review on Mining Students Data for Performance Prediction" - International Journal of Advanced Research in Computer and Communication Engineering Vol. 5, Issue 4, April 2016

[2] Tripti Dwivedi, Diwakar Singh – "Analyzing Educational Data through EDM Process: A Survey" - International Journal of Computer Applications (0975 – 8887) Volume 136 – No.5, February 2016

[3] Abdulmohsen Algarni - "Data Mining in Education" - (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 7, No. 6, 2016

[4] R. S. J. D. Baker, K. Yacef., "The State of Educational Data Mining in 2009: A Review and Future Visions", Journal of Educational Data Mining, Vol. 1, Issue.1, pp. 3–17, 2009,

[5] S.J Ryan, D. Baker, "Learning analytics and educational data mining", Proceedings of the 2nd International Conference on Learning Analytics and Knowledge - LAK '12, Year-2012

[6] M. Anoopkumar, A. M. J. M. Z. Rahman, "A Review on Data Mining Techniques and Factors Used in Educational Data Mining to Predict Student Amelioration", Proceedings of 2016 International Conference on Data Mining and Advanced Computing, SAPIENCE, pp. 122-133,2016.

[7] G. Siemens, R.S.J.D Baker. "Learning analytics and educational data mining: towards communication and collaboration." In Proceedings of the 2nd international conference on learning analytics and knowledge, pp. 252-254, ACM, 2012.

[8] J. Han, J. Pei, M. Kamber, "Data mining: concepts and techniques", Elsevier, 2011.

[9]D Baker, S.J. Ryan, "Mining data for student models.", In Advances in intelligent tutoring systems, Springer, Berlin, Heidelberg, pp. 323-337, 2010.

[10] Romero, C., & Ventura, S. "Educational data mining: A survey from 1995 to 2005". Expert Systems with Applications, 33(1), 135- 146, 2007.

[11] Romero, C., & Ventura, S. "Educational data mining: a review of the state of the art". Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on, 40(6), 601-618, 2010.

[12] Brijesh K. Baradwaj and Saurabh Pal, "Mining Educational Data to Analyze Student Performance", International Journal of Advanced Computer Science and Applications, Volume. 2, Issue No. 6, 2011.

[13] Sen, Baha, and Emine Ucar. "Evaluating the achievements of computer engineering department of distance education students with data mining methods." Procedia Technology 1 (2012): 262-267.

[14] R. Jindal, B.M. Dutta, "A Survey on Educational Data Mining and Research Trends", International Journal of Database Management Systems (IJDMS), Vol. 5, Issue-3, pp. 53-73, 2013.

[15] Dorina Kabakchieva, "Predicting Student Performance by Using Data Mining Methods for Classification.", Cybernetics and Information Technologies, Volume. 13, Issue no. 1, pages: 61-72, 2013.

[16] Elakia, Gayathri, and Naren J. Aarthi. "Application of data mining in educational database for predicting behavioural patterns of the students." Elakia et al,/(IJCSIT) International Journal of Computer Science and Information Technologies 5.3 (2014): 4649-4652.

[17] Shaziya, Humera, Raniah Zaheer, and G. Kavitha. "Prediction of students performance in semester exams using a naïve bayes classifier." International Journal of Innovative Research in Science, Engineering and Technology 4.10 (2015): 9823-9829.

[18] Guo, Bo, et al. "Predicting students performance in educational data mining." 2015 international symposium on educational technology (ISET). IEEE, 2015.

[19] Sweeney, Mack, Jaime Lester, and Huzefa Rangwala. "Next-term student grade prediction." 2015 IEEE International Conference on Big Data (Big Data). IEEE, 2015.

[20] Marbouti, Farshid, Heidi A. Diefes-Dux, and Krishna Madhavan. "Models for early prediction of at-risk students in a course using standards-based grading." Computers & Education 103 (2016): 1-15.

[21] Costa, Evandro B., et al. "Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses." Computers in human behavior 73 (2017): 247-256.

[22] J. Kumar, "A Comprehensive Study of Educational Data Mining", International Journal of Electrical Electronics & Computer Science Engineering, Special Issue-TelMISR, pp. 58-63, 2015.