

Novel Techniques for Educational Data Mining

Riya Jain¹, Vivek Kumar¹, Vinay Kumar Mishra¹, Mayank Singh¹, Neeraj Kumar Pandey²

¹ Computer Science and Engineering, Teerthanker Mahaveer University, India

²Computer Science and Engineering, DIT University, India

Abstract- Educational data mining (EDM) is a branch of study that focuses on the application of data mining, machine learning, and statistics of data, generated in educational contexts. This research area has been popular and some related terms like academic analytics, institutional analytics, and teaching analytics. Data mining is crucial in the subject of education, especially when assessing behaviour in an online learning environment. Cluster analysis and decision trees were applied as data mining techniques, in this study. The limits of the current study are discussed, as well as suggestions for further research.

Keywords— educational data mining, decision tree, cluster analysis, academic analytics, teaching analytics

I. INTRODUCTION

Data mining is a commonly used way of extracting necessary or relevant information from large data sources. The main purpose of EDM is to evaluate the various types of education-based data to solve research issues in the field of education [1]. Extraction of hidden data patterns and detecting connections between parameters in a huge amount of data are the policies for applying for data mining. “Survey for data in education using data mining techniques is popularly known as educational data mining” [2]. Education data mine is a scientific field that extracts information from educational data. Predicting student performance is one of the things that is done, which helps teachers identify students who need extra help [3]. Academic Analytics (AA) and Institutional Analytics (IA) are associated with the gathering, collection, and observation of curriculum activities such as courses and degree programmes, as well as research, student financial income, course testing, resource allocation, and management, in order to build institutional understanding.

Teaching Analytics (TA) is the study of teaching activities and performance data, as well as the planning, implementation, and assessment of educational activities. This is focused on the educational challenges as seen through the perspective of educators [4]. Learning analytics (LA) is the analysis and representation of student data in order to improve education is known as learning analytics (LA) [5]. LA implementation is often associated with web-based platforms, which provide direct access to student information with minimal effort or optimization [24]. Educational Data Mining (EDM) is concerned with improved methods for evaluating various forms of educational data from different institutions. It can also be explained in terms of the application of data mining techniques (DM) in this specific type of educational data to answer important educational problems [4]. This technology is so beneficial that it can clarify patterns derived

from data analysis to understand hidden information and to facilitate decision-making [6]. The cloud computing is used for resource allocation. Cluster analysis is a systematic way of building a collection of patterns by groups based on their similarities of particular property or action because a collection analysis is used for various purposes in educational data mining, one of the most interesting areas in its use separating students to identify common patterns of behaviour [2].

The purpose of decision trees is direct identification object classes. Decision trees (DT) use a variety of attributes to distinguish different low-level items and do not use one attribute or set of prescribed symbols. The appeal of decision-trees is on its way to understanding and interpretation [2, 7]. DT is a branch structure made up of rules. Recursive apportioning is the process of creating a branch structure in a DT [8]. The different techniques used in EDM are:

1. Neural Network: Traditionally, the word "neural network" has been used to describe a network or circuit of biological neurons. The phrase is frequently used in modern usage. Artificial neural networks are made up of artificial neurons. Nodes, or neurons, are the building blocks of the brain. Other types of signaling exist in addition to electrical signaling. Signaling is caused by the diffusion of brain transmitters, which has an impact on electrical signaling. As a result, neural networks are quite useful. The applications are: radial basis function, networks, neural classification, bayesian confidence propagation neural networks.

2. Algorithm Architecture: To calculate a function, algorithm design is described as a finite list of well-defined instructions. Calculation, data processing, and automated reasoning are all done via algorithms, simply describes an algorithm as a computation technique that follows a set of steps. The various application are: gap statistic algorithms, chi-square automated interaction detection, models and algorithms, GRASP, OLAP,

k-means, clustering algorithms, decision forest algorithms, classification and regression trees, Euclidean distance, bagged clustering algorithms.

3. Dynamic prediction based approach: The dynamic prediction based approach is a mathematical model for stochastic dynamics that is used to study molecules as well as other domains such as the stock market. The presence of a gaussian random noise is the most significant property of Langevin dynamics. A work[9] explores and implements the notion of temporal localization to various disciplines of computer science in 1968. It notices that a thread's transactions do not reach a completely random collection of addresses. The various applications are: ophthalmic oncology, vehicle fault diagnosis, grid computing, dyadic wavelet, pre-fetching, fault restoration prediction models, fault prediction models, financial distress prediction models, Vlasov–Maxwell equations, forecasting, anomaly detection.

4. Analysis of system architecture: A conceptual model is used to analyze system architecture that describes a system's structure, behavior, and other characteristics. Both software and hardware elements are used in system architecture. A good architecture can be thought of as a 'partitioning method,' or algorithm, that partitions all of the system's current resources and future requirements into a manageable set of clearly defined requirements subsystems with a bounded area. The various applications are: regression analysis, statistical analysis, discriminative analysis, association analysis, penalized discriminative analysis, process parameter analysis, cluster analysis, decision making, decision support systems, decision tree based models.

5. Intelligence agent systems: An intelligent agent system (IAs) is a self-contained entity that monitors and reacts to its surroundings in the field of artificial intelligence. To attain their objectives, intelligent agents can also study or utilise knowledge. They might be quite simple or extremely complex. A reflex machine, such as a thermostat, is an intelligent agent, as is a human being, or a group of humans working toward a common objective. Goal-directed behaviour as the essence of intelligence, thus they adopted the phrase "rational agent" from economics to describe it [10]. The various applications are: computer interface design, multiuser database systems, intelligent analysis, manufacturing intelligence, intelligent tutoring systems, support vector machines, program diagnostics systems, supervisory and specialist systems, supervisory and specialist systems.

6. Modeling: The process of creating a model in software engineering is known as modeling. Data model are created by describing formal data models with the help of approaches for data modeling using inductive logic programming. Modeling technology can give quantitative ways for data processing,

representing, or acquiring expert knowledge so that AI, cognitive science, and other research fields can benefit from algorithms. The various applications are: model-based diagnosis, forest fire proliferation modeling, model output statistics, intonation modeling, XML document modeling, Cox proportional hazard modeling, load damage exponents, polynomials, similar waveforms, simple additive weight.

7. Knowledge-based systems: Artificial intelligence techniques that are built on knowledge are known as knowledge-based systems work in a specific domain to make wise conclusions justification because KBS has its roots in the realm of artificial intelligence and the most popular term is human-centered. Various knowledge representation strategies, rules, frames, and scripts are used to acquire and represent knowledge. The documentation of data is one of the primary benefits provided by such systems - intelligent decision support, self-learning, and reasoning are all words that come to mind when thinking about knowledge as well as an explanation. The various applications are: knowledge representation, digital libraries and information gain theory data mining.

II. LITERATURE REVIEW

The method of extracting useful knowledge and patterns from a large educational database is known as educational data mining [11]. This survey represents how educational data mining is used for educational data. The research area has been popular and some related terms like academic analytics, institutional analytics, teaching analytics, etc are used in reference list. A work by Romero gives reviews about main publications, free available datasets and future trends in this research area [4]. This paper shows how data mining techniques can be used to create a data mining model for the higher education system, which can be used to discover information for predicting student success, enrolment numbers, and other factors. Decision tree method in classification task is used to evaluate the students' performance in the end semester examination[7]. In a classification task, the decision tree approach is used to measure a student's success in the end-of-semester test. It explains the use of data mining techniques on educational data from a Croatian educational institution. The research employs cluster analysis and decision tree techniques. Cluster analysis organises a set of patterns into clusters, while Decision Tree is a method of creating a decision-making representation [2]. Meta-analyses show that researchers benefit from greater predictive success during study. However, predicting performance before the start of the course requires special attention [12]. This survey aims to show the importance of data mining in educational domain. The research in this area is to find the methods that can improve prediction for better result. It also help to find out techniques, metrics and attributes used and define the current statements[13]. Various machine learning algorithms and research tools of Data

Mining are used to analyze and predict various types of educational data. Some of the tools used are – RapidMiner, KNIME, Orange, SPSS, KEEL, The EDM Workbench, SPARK MLlib, D3js, PSLC Datashop and CloudSCAN [1] [18].

Learning analytics is a process that predicts the performance of learners and identifies their problems as EDM translates data into meaningful actions. These practices are used in online teaching more than in offline teaching. This paper designs a model to implement LA and EDM in the offline teaching environment. The said system is implemented in R and Shiny [14]. A study conducted on predicting student performance in order to improve their progress. Data mining techniques may be used to improve the quality and efficiency of education. Multiple graphical, mathematical, and quantitative methods can be used to achieve this. This paper proposes a systematic method for selecting an effective ensemble learner from a set of six possible machine learning algorithms based on Gini-index and P-value [15]. Learning statistics were described as one of the most relevant learning and educational methods established technologically in the NMC Horizon 2013 study [16]. Advances in communication technology, especially the Internet, mobile devices, computers, digital sensors, and recording devices, have resulted in a significant increase in the amount and complexity of data that can be analyzed [17]. However, EDM / LA has a wide range of potential targets or educational issues. In reality, in terms of a last resort customer, there are several more precise objectives. The following are some of the latest hot topics or more interesting problems in the field: 1) Analyzing educational theories, 2) Examining how learning concepts and statistics can be applied to academic study, 3) Analyzing pedagogical strategies, 4) Analyze and evaluate the application of teaching effect strategies for EDM / LA strategies, 5) Analyzing programming code using EDM / LA techniques to analyze code from software classes/ program presentations/ submissions, or other sources, 6) Collaborative education and teamwork group, 7) Analyzing collaborative learning and predicting group distance in collaboration groups, 8) Curriculum mining/ analytics, 9) Analyze program structure, curriculum and management, 10) Dashboards and visual learning analytics, and 11) Incorporating the process of recognition for proper assessment and understanding user leads collected locally and optimized.

Neural network architectures with multiple layers of processing units are being built in the EDM / LA research field. Early warning systems, such as finding the possibility of intervening early to facilitate student achievement by accurately predicting student performance and at-risk students, and discovery of causal relationships, such as finding the causal relationship between symbols in the

educational database. Emotional learning analytics, such as study, has an impact on the amount of time spent studying and the role of emotions in reading, as well as evaluating the intervention's effectiveness, including measuring the intervention's efficiency, data-driven student responses, and future guidance, among other things.

Incorporating data mining techniques and viewing strategies into player engagement in critical games is what game learning analytics is all about. Improve the rendering, descriptive, usable, and high-quality "white box" models by using interpretable and explanatory learner models. Learning a foreign language using EDM and LA improves the ability to read in another language. EDM/LA techniques are being used to assess the self-taught learning function as well as student behavior. Multimodal learning analytics- Using machine learning and low-cost hearing technology to provide new types of understanding of what's going on in the majority of cases [19]. In ongoing LA processes, for academic adoption, organize learning analytics means learning to accept, what to do, and more. Personalized feedback- Automated or personalized responses to assist students in their learning. Sentiment discovery- Identify attitudes, feelings, and student humility as well as learning resources automatically. Prepared models can be transferred or used in other similar applications such as courses, institutions, and so on. Text mining software and analytics are being deployed to input data from organizations, forums, social media, reviews, essays, and other sources to writing analytics.

III. PROPOSED MODEL

A data analyzer is designed and developed during the course of this work. The proposed work contains task as the procedure. Methodology will adapt to carry out the proposed work in a sequence as shown in Fig 1.

Educational Raw Data

Data that has not been processed to be viewed in any visible form is known as raw data, also known as source data or atomic data. The raw data is never processed, and it could be in an improper format.

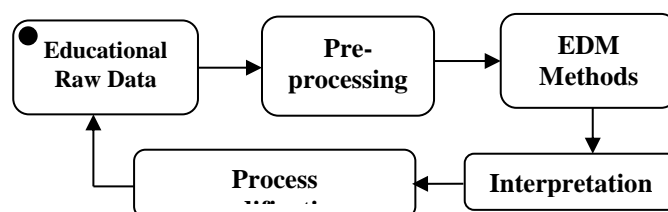


Fig. 1 Proposed Model for EDM Process

This type seems so different and almost meaningless without processing, but depending on the situation, it may also be in a form that others may understand.

Preprocessing

The first step in every data mining method is to process the data in advance. As a result, data is cleaned and converted into an executable format in this process. Pre-processing is the process of converting raw educational data into a format that can be used by a data mining algorithm to solve a particular educational problem.

Educational Data Mining (EDM) Methods

The third step is to implement specific EDM techniques in order to achieve the experiment's goal. EDM methods are used as data mining algorithms to process the preprocessed data. The techniques used for this purpose are Data mining; Clustering; Decision trees; Knowledge discovery.

Interpretation

The analysis of experiment results gives answers to educational questions and decision-making. Educational Data Mining (EDM) uses these algorithms to assess educational statistics to find patterns and predict data that show student performance. Various structural challenges such as accuracy, purpose and performance, as well as multiple heads where data is set are extremely large, etc.

Education Process Modification

The final step is to modify the educational process or postpone it until further testing is conducted for a better or more reliable outcome. EDM's key goal is to extract information from educational data in order to answer critical research questions and aid decision-making. The increased use of technology in education systems has resulted in the loss of a significant number of student records, making EDM a valuable tool for improving teaching and learning processes. EDM is useful in a variety of situations, including recognizing at-risk students and determining the priority learning needs of various classes. These students have successfully evaluated institutional efficiency, improved campus infrastructure, and improved curriculum redevelopment, resulting in higher graduation rates.

IV. CONCLUSION

The Educational data mining field has become demanding in recent years due to the advent of large data repositories containing records of student actions at e-courses, and educational institutions. In this article, the mining techniques for education data were examined. The research performed in this document uses the cluster analysis and the decision making process because educational data mining is part of the data mining industry. Cluster analysis identified groups of readers based on the frequency in which they accessed information, confirming the author's previous findings. The decision-tree process is used for collection results for in-depth

analysis student conduct in teaching and learning programs. Data mining and analysis are powerful in grade education. Generally, like most educational institutions, grade education will be enhanced by the increasing amounts of data currently available. There is the potential to increase the quality of learning materials, identify students at risk, and provide better support for both students and faculty. Students and teachers will benefit from this research as they work to change student segregation. This research will also look at students who need extra attention in order to lower their failure rate and take the necessary measures for further exams.

REFERENCES

- [1] G. & S. B. Mahajan, "Educational Data Mining: A state-of-the-art survey on tools and techniques used in EDM.," *International Journal of Computer Applications & Information Technology*, vol. 12(1), pp. 310-316, 2020.
- [2] S. Krizanić, "Educational data mining using cluster analysis and decision," *International Journal of Engineering Business*, vol. 12, feb 2020.
- [3] W. W. P. N. & Z. A. Damopolii, "Educational Data Mining in Predicting Student Final Grades.," *International Journal*, 10(1), 2021.
- [4] C. S. Romero, "Educational data mining and learning analytics: An updated survey.," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 10, may 2020.
- [5] L. M. Oliva-Córdova, A. Garcia-Cabot and H. R. Amado-Salvatierra, "Learning Analytics to Support Teaching Skills: A Systematic Literature Review," *IEEE Access*, vol. 9, pp. 58351-58363, 2021.
- [6] N. D. & E. A. W. R. Lynn, "Using Data Mining Techniques to Predict Students' Performance. a Review.," in *In IOP Conference Series: Materials Science and Engineering*, 2021.
- [7] B. K. P. S. Baradwaj, "Mining educational data to analyze students' performance," *arXiv preprint arXiv:1201.3417*, 17 jan 2012.
- [8] M. N. S. P. Santosh and M. M. S. Ahmed, "A comprehensive analysis on educational data mining for students," *Journal of Xi'an University of Architecture & Technology*, vol. 12, no. 3, 2020.
- [9] J. Peter, "The working set model for program behavior," *Communications of the ACM*, vol. 11(5), pp. 323-333, 1968.
- [10] S. J. & N. P. Russell, *Artificial Intelligence: A Modern Approach*. Upper Saddle River, New Jersey: Prentice Hall. ISBN 0-13-790395-2., 2003.
- [11] A. M. & H. W. Shahiri, "A review on predicting student's performance using data mining techniques," *Procedia Computer Science*, vol. 72, pp. 414-422, 1 jan 2015.
- [12] A. & G. S. K. Khan, "Student performance analysis and prediction in classroom learning: A review of educational data mining studies.," *Education and information technologies*, vol. 26(1), pp. 205-240, 2021.
- [13] A. S. K. D. S. Dhankhar, "Predicting Students Performance Using Educational Data Mining and Learning Analytics," *A Systematic Literature Review. Innovative Data Communication Technologies and Application*, pp. 127-140, 2021.

- [14] A. S. J. R. Vaidya, A Framework for Implementation of Learning Analytics and Educational Data Mining in Traditional Learning Environment, In *ICT Analysis and Applications*, Springer, Singapore., 2021, pp. 105-114.
- [15] M. M. A. N. A. B. S. A. Injadat, "Systematic ensemble model selection approach for educational data mining," *Knowledge-Based Systems*, 200, 105992., 20 July 2020.
- [16] L. Johnson, S. Adams Becker, M. Cummins, V. Estrada, A. Freeman and H. Ludgate, "The NMC Horizon Report: 2013 Higher Education Edition," eric, Austin, 2013.
- [17] B. S. J. S. A. D. A. S. H. Wibawa, "Learning analytic and educational data mining for learning science and technology," in *In AIP Conference Proceedings*, AIP Publishing LLC., 2021.
- [18] V. Kumar and N. Singh, "CloudSCAN: A Smart Content Aggregation and Navigation Technique on Cloud," in *International Conference on Power Electronics & IoT Applications in Renewable Energy and its Control (PARC)*, Mathura, 2020.
- [19] I. P. N. & C. A. Kazanidis, "A Learning Analytics Conceptual Framework for Augmented Reality-Supported Educational Case Studies," *Multimodal Technologies and Interaction*, vol. 5(3), march 2021.
- [20] S. Križanić, "Educational data mining using cluster analysis and decision," 2020.
- [21] S. Križanić, "Educational data mining using cluster analysis and decision," *International Journal of Engineering Business Management*, vol. 12, feb 2020.
- [22] L. M. G.-C. A. & A.-S. H. R. Oliva-Córdova, "Learning Analytics to Support Teaching Skills, A Systematic Literature Review," *IEEE Access*, 9, pp. 58351-58363., april 2021.
- [23] M. N. S. P. & A. M. M. S. SANTHOSH, "A COMPREHENSIVE ANALYSIS ON EDUCATIONAL DATA MINING FOR STUDENTS.," *Journal of Xi'an University of Architecture & Technology*, vol. 12, no. 3, 2020.
- [24] N. Kumar Pandey, S. Chaudhary and N. K. Joshi, "Resource allocation strategies used in cloud computing: A critical analysis," *2016 2nd International Conference on Communication Control and Intelligent Systems (CCIS)*, 2016, pp. 213-216, doi: 10.1109/CCIntelS.2016.7878233.