Handling Big Data in Education: A Review of Educational Data Mining Techniques for Specific Educational Problems

Yaw Boateng Ampadu[1,2,*]

1 Department of Educational Innovations in Science and Technology,
Kwame Nkrumah University of Science and Technology, Kumasi, Ghana
2 Visual Arts Department, Prempeh College, Kumasi, Ghana
*Correspondence: E-mail: ybampadu@st.knust.edu.gh; yawboatengampadu@gmail.com

Abstract

In the era of big data, where the amount of information is growing exponentially, the importance of data mining has never been greater. Educational institutions today collect and store vast amounts of data, such as student enrollment and attendance records, and their exam results. With the need to sift through enormous amounts of data and present it in a way that anyone can understand, educational institutions are at the forefront of this trend, and this calls for a more sophisticated set of algorithms. Data mining in education was born as a response to this problem. Traditional data mining methods cannot be directly applied to educational problems because of the special purpose and function they serve. Defining at-risk students, identifying priority learning requirements for varied groups of students, increasing graduation rates, monitoring institutional performance efficiently, managing campus resources, and optimizing curriculum renewal are just a few of the applications of educational data mining. This paper reviews methodologies used as knowledge extractors to tackle specific education challenges from large data sets of higher education institutions to the benefit of all educational stakeholders.

Keywords: big data, educational data mining, data mining techniques, machine learning, prediction

1. Introduction

In education, the rise of "big data" in combination with progress in technology through new extended instructional media [1] promises to improve learning processes in formal education and beyond. It has become increasingly important in education to use data mining to assist students in their data analysis, as it uses several factors and interprets it to deliver useful information [2]. The interaction of

students with education software and online learning are increasingly being made available with extremely huge data sets [3]. By analyzing the large amount of education data generated and collected during the course of teaching and learning, stakeholders, such as teachers, students, and managers, can gain a holistic view of the progress of learning and prescribe appropriate evidence-based interventions or recommendations based on personalized data. In the educational sector, educational data mining (EDM) uses data mining methods, some of which are used to predict results, such as classification, while others, such as clustering, are known to be descriptive [4]. Notwithstanding, various types of EDM techniques such as association-rule mining and clustering are used to discover student behaviour [5]. EDM is used for a variety of purposes, including identifying at-risk students, identifying priority learning requirements for various groups of students, increasing graduation rates, efficiently monitoring institutional performance, managing campus resources, and optimizing curriculum renewal [6].

This paper reviews methodologies used as knowledge extractors to tackle specific education challenges from large data sets of higher education institutions to the benefit of all educational stakeholders.

This paper intends to explore EDM techniques from the standpoint of Baker [7] on EDM techniques and applications. There are five sections in this study. Section 1 introduces the goals and organization of the paper. Section 2 looks at the development of EDM and its goals, while Section 3 looks at EDM methodologies and processes. Section 4 examines the use of EDM techniques in related publications. Section 5 ends this study and makes recommendations for future research in this field.

2. Background

2.1. Educational data mining

Educational data mining (EDM) refers to a sub-domain of data mining that focuses on extracting knowledge from the information in an academic database. The Educational Data Mining community website (educationaldatamining.org, [8]), defines EDM as: "an emerging discipline, concerned with developing methods for exploring the unique types of data that come from educational settings, and using those methods to better understand students, and the settings which they learn in."

EDM aims to create and enhance methods for analyzing educational data, which frequently contain several levels of meaningful structure, to uncover new insights into how students learn in such environments [9]. As a result, EDM has aided researchers in learning sciences in their investigations of learning theories [9, 10].

As a way to gather rich and multimodal data from students' learning activities in educational settings, EDM uses e-learning platforms like Learning Management Systems (LMS) and Intelligent Tutoring Systems (ITS), as well as Massive Open Online Courses (MOOC). These platforms, for example, keep track of when and how many times students access a particular learning resource, as well as whether the answer they provide to an exercise is correct.

A great deal of data is made available by the growing use of technology in education systems [9]. Data is recorded in an online learning environment each time a student uses a learning management system. Analyzing this data can help with various educational issues, such as generating recommendations and developing adaptive systems and providing automated grading for students' assignments. EDM utilizes this data to find relevant information on distinct types of learners and their learning, the structure of the field of knowledge, and the influence of teaching strategies incorporated into the different learning contexts [2].

2.1.1. Objectives of EDM

Using data mining (DM) techniques in education is primarily aimed at developing models that can predict the overall performance of students in specific courses [11]. EDM has been used to address a wide range of objectives, all of which are part of the overall goal of enhancing learning [12]. Many studies (including those by Romero and Ventura [9] Aldowah *et al.* [13]; Okewu *et al.* [14]; Safitri *et al.* [15]) have suggested categorizing EDM goals based on the final user's perspective (that of the learner), as well as the challenge to be solved and include the following:

- EDM can predict learners' behaviour by improving student models. Modeling is the process of describing and categorizing a student's knowledge, motivation, metacognition, and attitudes.
- Models of knowledge domain structure are being discovered or improved. There are concept models of the content being taught, as well as models that describe the interrelationships of knowledge within a domain.
- Learning systems are being used to research the best effective pedagogical support for student learning.
- Developing empirical data to support or define pedagogical theories, frameworks, and educational phenomena to identify fundamental influential learning components and create better learning systems.

Furthermore, EDM information is targeted to a variety of stakeholders [16]. Different groups of stakeholders examine educational data from different perspectives, each with their own purpose, vision, and goals for implementing EDM [9]. The four stakeholders are classified by Romero and Ventura [9] based on their EDM goals:

- Educators: Increasing teaching effectiveness by analyzing students' learning habits, obtaining the most supporting instructions, and anticipating student learning.
- Learners: Improving or suggesting individual learning methods, learning materials, and learning experiences.
- Organizations/Institution: Improving the efficiency and cost-effectiveness of decision-making processes in higher education institutions, such as admissions and the allocation of financial resources.
- Researchers and developers: evaluating learning materials, developing learning systems, and determining the efficiency of data mining approaches.

EDM gives useful information and a better view of students and their learning processes [13]. It uses DM methods to analyse educational data and find solutions to educational problems [9]. EDM extracts interesting, interpretable, valuable, and unique information from educational data in the same way that other DM methods do [17]. However, EDM is primarily designed to build methods using distinctive data types in educational systems [9]. These strategies are then employed to improve knowledge about educational phenomena, students and the environments in which they learn [18].

3. EDM techniques

The conventional approaches of DM cannot be readily applied to these types of data and challenges in educational environments [19]. As a result, different types of DM techniques are required for specific educational problems [20]. For a variety of purposes, there are a wide range of general DM methods. The problem is that these are not suited to handling educational data. Furthermore, these DM tools cannot be used by educators or teachers who do not have a basic understanding of DM concepts [19]. Methods for DM are derived from a wide variety of disciplines, including machine learning, statistical methods such as psychometrics, visualization techniques such as infographics, and computational modelling [21].

EDM goals have been achieved using most standard DM techniques, such as classification, clustering, and association analysis approaches, but these are by no means the only ones [9]. Educational systems, on the other hand, have unique characteristics that necessitate a unique approach to the mining problem [22]. Consequently, EDM researchers not only employ DM techniques, but also propose, develop, and employ approaches and techniques from a wide variety of EDM-related domains [9]. Baker's [7] categorization of these approaches is the most popular: prediction, clustering, connection mining, distillation for human assessment, and model discovery. In addition to Bienkowski *et al.* [23], Romero and Ventura [20] expanded this taxonomy. Work in educational data mining is classified into the following categories, according to Romero and Ventura:

- Statistics and visualization
- Web mining

  Classification, clustering, and outlier detection
  Association rule mining and sequential pattern mining
  Text mining

Logs of student-computer interaction are a primary source of educational data mining [24]. The web mining methods outlined by Romero and Ventura are widely used in EDM today, both in the mining of web data and other educational data.

Using Baker [7] as a guide, educational data mining can be looked at from a second perspective:

- *Prediction (Classification, Regression, Density estimation)*
- *Clustering*
- *Relationship mining (Association rule mining, Causal data mining, Correlation mining, Sequential pattern mining)*
- *Distillation of data for human judgment*
- *Discovery with models*

Baker's taxonomy of educational DM methods contains three familiar categories (the first set of sub-categories are directly derived from Moore's categorization of DM methods). Statistics and visualization are included in Romero and Ventura's definition of DM and have played an important role in both published EDM research [25] and theoretical discussions about EDM. Baker's EDM taxonomy has a fifth category that, from the standpoint of traditional DM, is the most unusual.

Prediction

Prediction is an educational DM technique that uses past data to anticipate and predict the future [26]. It is used to help teachers identify which students are most likely to succeed in various subjects, which students are most likely to need remediation in a subject, and which students are the most likely to fail their classes and drop out. The most common type of regression analysis in EDM is linear regression, which is a statistical technique that predicts a continuous value from one or more continuous or categorical input variables [27].

The objective of the predictive technique, according to Nithya and Ilango [28], is to develop a model that can infer a single aspect of the data (predicted variable) from a combination of other aspects of the data (predictor variables). Classification (when the predicted variable is a categorical value), regression (when the predicted variable is a continuous value), and density estimation are examples of prediction methods (when the predicted value is a probability density function).

These algorithms can create accurate predictions by studying patterns and correlations in data. Predictive models may help educators, administrators, and policymakers make educated choices and distribute resources more efficiently. Predicting a student's academic success and behaviour is one application of EDM [29].

Clustering

Clustering, according to Ahuja *et al.* [30], is the process of grouping data items together based on their similarities and characteristics. Clustering is used in different fields to solve problems such as finding similar objects in a collection, identifying similar users based on their activities, and sorting objects based on their characteristics [31]. According to the data type, similarity measure, and theory used in clustering algorithms classifications can be made [32]. For example, if the data is categorical, then hierarchical clustering is the technique of choice. If the data is numeric, then K-means clustering is the clustering technique of choice [33]. Clustering techniques can be applied to a variety of educational data sources, including performance data from standardized tests, data from class discussions, student evaluations, and data from interviews with the students [22]. The goal of clustering in EDM is to identify groups of students that share similar patterns in their performance on a particular test [20]. Clustering techniques can also be used to identify groups of students within a class that share similar test scores, to identify groups of students that are likely to drop out of school, and to identify groups of students that are likely to perform well in the future [34].

Relationship mining

For relational databases, relation mining, also known as relational DM, is extensively utilized [35]. A relationship between different variables within a data collection is discovered using relationship mining. The relational DM algorithm searches for patterns among various patterns in a database. Two criteria must be met in a relationship between variables: interest and significance [7]. As a result, the goal of relationship mining is to discover connections between distinct variables in large data sets. This requires determining which variables are most linked to a certain variable of interest [36]. Relation mining also measures the strength of connections among various variables. Two requirements must be met in connection mining: statistical importance and interest [7]. Baker further explains that association rule mining (any connections between variables), sequential pattern mining (temporal associations between variables), correlation mining (linear correlations between variables), and causal data mining are all examples of relationship mining approaches (causal relationships between variables). The most popular EDM approach is association rule mining [37]. The basic objective of relationship mining is to discover whether one event causes another event in a dataset by looking at the

coverage of the two events or by looking at how an event is triggered [36]. Relationship mining is used in EDM to find correlations between students' online activities and final grades, as well as to model learners' problem-solving activity sequences.

Distillation of data for human judgment

Hicham *et al.* [38] describe this technique as a process of extracting the important aspects of data to make better decisions. Since the data is voluminous, it is necessary to distill the information into a manageable amount so that it can be easily analyzed [39]. In order to make the data more manageable, this technique is used. Visualization, summary, and interactive interfaces are used to highlight relevant information and aid in decision-making [36]. The classification or identification of data is used in educational data distillation for human judgment. Identification aims to display data in such a way that it is easily identifiable via well-known patterns that cannot be formalized [40]. Classification of data can be used as a pre-processing step in the development of a prediction model [41]. The technique of distillation is used in the field of education to help teachers make better decisions [38]. In a classroom, a teacher may want to know how many students in the class have high test scores, but don't have a lot of homework. In this case, the teacher distills the information by looking at the test scores and the amount of study time for a test is based on how much homework students have completed.

Discovery with models

"Discovery with Models" methodologies are becoming increasingly common in learning analytics and EDM studies [9]. In these studies, an existing model is used as a primary component of the analysis [42]. This is a methodology that consists of a collaborative process between teachers and students in which models are created as a visual representation of the knowledge that students are hoping to learn [20]. According to Bienkowski *et al.* [23], these models are created in a format that both students and teachers can understand, allowing teachers to keep tabs on their students' progress as they learn. Using this method, the teacher can also make improvements to the models.

Discovery models are based on clustering, prediction, or knowledge engineering using human reasoning rather than automated techniques [38]. As a result, the generated model is employed in other comprehensive models, such as relationship mining [36]. It is used, for example, to identify the relationships between the student's behavior and characteristics [9].

3.1. EDM process

Mehra and Agrawal [43] emphasized that the EDM process is the same as the DM process because it involves the same steps, which are preprocessing, data mining,

and post-processing. An important part of the DM process is the transformation of raw data (information that has not been analysed) into useful information (knowledge) [44]. The steps of the data mining process for extracting knowledge are shown in figure 1.
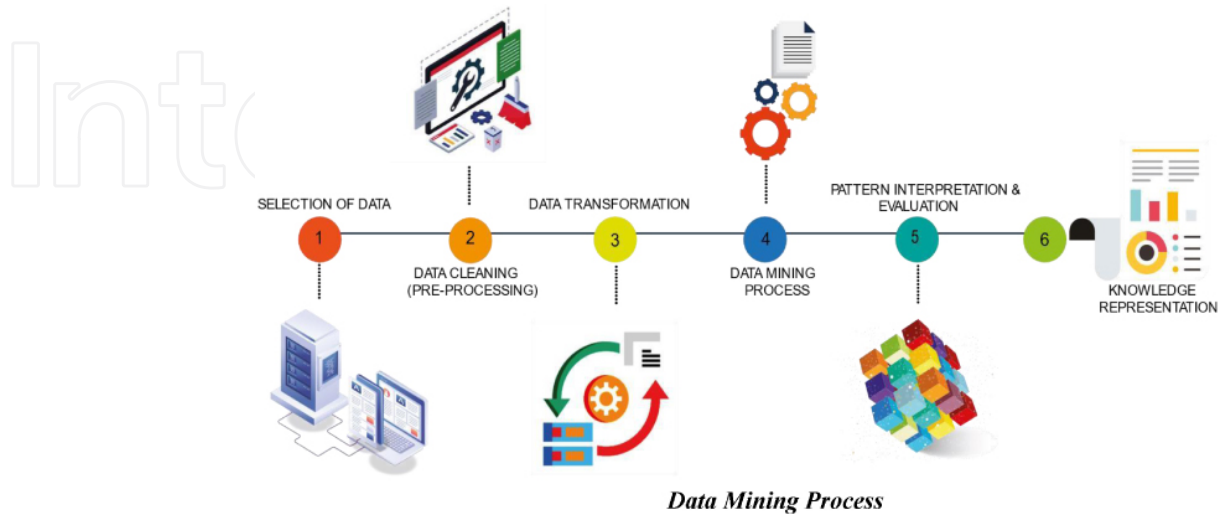

Figure 1. Steps of the data mining process for extracting knowledge.

- *Data Selection:* In this step, a pre-processed data set is selected or retrieved for use in the discovery process [45].
- *Data Pre-Processing:* To improve the reliability of the data, this step involves removing unnecessary information from the data set, as well as identifying missing values [46].
- *Data Transformation:* Transformation and categorization of the data for mining, such as classification and clustering are carried out in this phase of the process [46].
- *Data Mining:* To extract useful patterns, techniques and tools must be applied at this point in the process. Classification, clustering, regression, and other techniques are all part of data mining algorithms [47].
- *Pattern Evaluation:* In this stage, we can identify specific patterns and evaluate them to arrive at the desired goals [19].
- *Knowledge Representation:* Knowledge gained in previous phases is visually presented in this final phase. This stage makes use of visual techniques to assist users in interpreting the results and gaining a complete and an overall picture [45].

4. EDM application

Researchers in the field of education are increasingly relying on DM techniques to delve deeper into the academic performance and habits of their students. It is possible to use various DM techniques (such as decision trees, association rules, nearest neighbors, neural networks, genetic algorithms, exploratory factor analysis and regression) to analyse large amounts of educational data in order to help students improve their performance. These methods assist teachers in identifying students who require special advice or academic counselling. This provides a high-quality education.

4.1. Prediction of students' performance

The metric prediction, in other words, is called "regression." Regression can be used to represent the relation between one or more independent and dependent variables. In prediction, records are classified according to some predicted future behaviour [48]. These predictions use numerous DM techniques, like some classification techniques (such as support vector machines, backpropagation, and k-nearest neighbour classifiers) that can be used for prediction [49]. DM techniques can be used to improve academic performance in educational institutions, according to Pal and Pal [50]. These researchers looked into and compared the educational applications of DM based on the personal, social, psychological, and other environmental characteristics of their subjects. It was their goal to use the information they gleaned from the student database to help students improve their performance. A rule learner (OneR), a common decision tree algorithm (C4.5) (J48), a neural network (Multi-Layer Perceptron), and a Nearest Neighbor algorithm (IB1) were some of the data mining techniques employed. They achieved their goal of evaluating student performance using the four Weka-based classification algorithms. Based on the placement data, the best algorithm was IB1 Classification, which had an accuracy rate of 82.00% and a build time of 0 seconds. IB1's average error is only 0.20, which is significantly lower than those of other classifiers. Based on these findings, it appears that IB1 classifier, among the machine learning algorithms evaluated, has the greatest potential to improve on the performance of conventional classification methods. Student performance was found to be more strongly influenced by factors such as SSG (Senior Secondary School Grade), HSG (High School Grade), Mqual (Mother's Qualification), and FAIn (Family Annual Income). By analysing the data from previous students, we were able to generate a short but precise list of predictions for each new student.

The findings of a case study on educational data analytics that looked at the detection of undergraduate Systems Engineering (SE) students dropping out after six years of enrollment in a higher education institution are described by

Pérez *et al.* [51]. A feature engineering process was used to extend and enrich the original data.

Preliminary findings from a large dataset of student demographics and transcript records at various points in their degrees were presented by Pérez *et al.* for the prediction of student attrition. Our research revealed a correlation between the performance of physics and math courses and the performance of systems engineering courses. Dropout has a positive correlation with irregularity (standard deviation of term averages).

The findings of their experiment demonstrated that dropout predictors can be identified with dependable levels of accuracy using simple algorithms. To suggest the best choice, the results of Decision Trees, Logistic Regression, Naive Bayes, and Random Forest were compared.

4.2. Grouping/categorizing learners

Clustering is one of the most basic techniques for analysing the student data set. It is used in EDM to group students based on their characteristics. Clustering assists in classifying students into well defined clusters in order to identify students' behaviour and learning styles [52]. The objective is to organise students into groups based on shared traits, such as personality traits and interpersonal skills. As a result of this, an instructor or developer will create a custom learning framework that encourages productive community education, adaptive content, etc. [53]. According to Taha *et al.* [54], it provides a wealth of information on the subject of decision making and prediction in e-learning that can significantly improve them. In their study, Khasanah and Harwati [55] use two popular DM techniques (clustering and classification analysis) to predict student performance. The K-means algorithm was used because it is a well known and simple clustering algorithm. Because the attributes used in this study are numerical data, Linear Regression and Support Vector Machine (SVM) were used to predict the final GPA. In the classification analysis, the mean squared error (MSE) was compared between clustered and nonclustered data. They concluded from the study that all of the algorithms used could achieve the study's goal. After using K-means, the optimal number of clusters was determined to be four. Furthermore, they concluded that it is critical to cluster the data first before performing the classification analysis. This, they claim, can reduce the root mean squared error (RMSE). The results also revealed that Linear Regression outperformed SVM in predicting students' final GPA. For the purposes of classifying students with similar characteristics, Krpan and Stankov [56] used DM techniques to analyse real-world experience gained from an e-learning system. Students' Moodle database data and test scores were subjected to DM analysis. To find out if DM techniques can help us organise students, they conducted this study. A post-and-pre-test difference was used to determine the study's success

(to determine if students advanced or not). In theory, students with more free time and involvement in extracurricular activities should do better on tests, but the analysis revealed otherwise. Students in this category performed worse on tests and had lower course cores, despite the fact that they spent the majority of their class time in lessons with learning content. In their initial investigation, they noticed that something was amiss with the analysis, but they also double-checked the entire process. Because some of their students did not have access to a computer or the internet, they organised their learning time in a computer lab and had teaching assistants supervise it. It is possible that students were just browsing lesson pages in a controlled environment. They also mentioned that, despite spending a lot of time at home studying, some students' grades remained low. As with the correlation analysis, the cluster analysis produced similar outcomes. Students who scored lower had more actions, so they were all placed in the same cluster. The small number of cases and the smaller size of the dataset had an impact on the number of clusters (number of variables). Because students did not use all of the available activities, some variables were grouped together, while others were completely discarded. Unfortunately, they were forced to discard cases with missing data during the data cleaning stage. It has been found that a small data set allows for easier control and understanding of specific cases during EDM model design, especially in the early stages. Data collection processes (e.g., SQL queries) were easier to identify and validate because they were more specific. An important goal is preventing bad behaviour from affecting students' academic performance.

4.3. Decision making by educators

Processes involved in making decisions as a human thought behaviour include identifying issues, formulating goals, examining the materials involved, and finally, executing the chosen course of action [57]. For example, different domains may have different content for each step of this process, demonstrating fundamental differences in natural laws. Because it is based on educational principles and laws, the process exhibits decision-making characteristics. According to Lei *et al.* [58], a framework for educational decision making, which can be used to better understand some of the laws and characteristics of student development, has been proposed. In addition, they conducted a case study to demonstrate the framework's efficacy. According to Lei *et al.* [58], a framework for educational decision making, which can be used to better understand some of the laws and characteristics of student development, has been proposed. In addition, they conducted a case study to demonstrate the framework's efficacy.

As a common algorithm, decision trees were used to create a classifier that could predict one attribute or one aspect of the data from a combination of other aspects. Decision trees are widely used algorithms. In their research, Lei *et al.* [58] used the

Weka software to analyse the data of graduate students using the pruned C4.5 algorithm (with a collection of data mining algorithms, it contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization).

The following were the conclusions that Lei *et al.* reached about the job type of master graduate students at this university based on the classification rules obtained from EDM:

- Graduate students' job preferences are clearly influenced by the schools they attend. Students in the fifth grade (S5), for example, have a preference for jobs in scientific institutions (J2), while students in the ninth grade (S9) concentrate on positions in the justice sector (J9) (other types of occupations).
- The job type is closely linked to the other outcome of the case student development system, which is the job location. As an example, students in School 4 appear to choose between jobs in state-owned enterprises (J1) in Beijing, Tianjin, and Shanghai, and jobs in scientific institutions (J2) in other districts (D2–D4) of China.
- Graduate students' job types are heavily influenced by the schools they attend, not just their academic achievements, such as their "Academic Score" or "National Prize," but also their "General Prize," "National Prize," and other student-engagement-related attributes like those mentioned above.

The interpretations of mining results may provide several decisions to policymakers or leadership in the university, considering the educational goal of this case study that helps make decisions to improve the distribution of graduate students' job types. The following were proposed;

- As a part of the university's student development system, the university should encourage schools to promote energy and initiative in order to encourage students to pursue their majors and find employment.
- Faculty and staff can encourage students to consider not only the type of job they want to pursue, but also the location of that job.
- In addition to academic achievements, such as a student's academic score and prize, educators should pay more attention to other involvement factors, such as the faculty student interaction.

5. *Conclusion*

A growing usage of technology creates enormous volume of data in education. The subject of data mining is expanding quickly in education and has the benefit that it contains new algorithms and technology created in several areas of data mining and

machine learning. In a range of domains, EDM may be used to detect students at risk, prioritize learning goals among the many groups of students, improve the number of graduates, maximize campus resources, and optimize curriculum innovation. This article presents some methods or techniques available and how they have been used in the EDM field. There are numerous ways that the EDM could benefit all educational stakeholders. Tools and techniques like this could help students succeed in academics, boost the performance of educators and institutions, and aid in decision-making. In this way, data mining in higher education could benefit both the educational institutions themselves and their faculty members. Researchers, education providers, educational decision makers, and others can use this review paper as a guide to better implement and promote EDM. It is worth mentioning that a growing variety of approaches are being utilized in EDM to analyze the various data produced in educational systems. The type of data provided the nature of the learning environment, and the study objectives all influence which approach to be used for extracting knowledge from educational data.

Conflict of interest

The author declares no conflict of interest.

References

1   Mayer M. Innovation at Google: the physics of data. PARC Forum [Internet]; 2009 [cited 2009 Aug 11].

2   Nguyen A, Gardner L, Sheridan D. Data analytics in higher education: an integrated view. *J Inf Syst Educ*. 2020;**31**(1):61–71.

3   Williamson B. *Big data in education: the digital future of learning, policy and practice*. London: SAGE; 2017. p. 1–256.

4   Ray S, Saeed M. Applications of educational data mining and learning analytics tools in handling big data in higher education. In: *Applications of big data analytics: trends, issues, and challenges*. Cham: Springer; 2018. p. 135–160.

5   Anjum N, Badugu S. A study of different techniques in educational data mining. In: *Advances in Decision Sciences, Image Processing, Security and Computer Vision: International Conference on Emerging Trends in Engineering (ICETE)*. vol. 2, Cham: Springer; 2020. p. 562–571.

6   Govindarajan M. Educational data mining techniques and applications. In: *Advancing the power of learning analytics and big data in education*. Hershey, PA: IGI Global; 2021. p. 234–251.

7   Baker RSJD. Data mining for education. *Int Encycl Educ*. 2010;**7**(3):112–118.

8   Educational Data Mining Consortium. (n.d.) Educational Data Mining [Internet]; 2022[Retrieved 2022 Sep 31]. Available from http://www.educationaldatamining.org/.

9   Romero C, Ventura S. Educational data mining and learning analytics: an updated survey. *Wiley Interdiscip Rev: Data Min Knowl Discov*. 2020;**10**(3):e1355.

10 Maclellan CJ, Harpstead E, Patel R, Koedinger KR. The apprentice learner architecture: closing the loop between learning theory and educational data. In: *9th International Conference on Educational Data Mining EDM '16, Raleigh, NC*. Washington, DC: ERIC; 2016.

11 Ahmad F, Ismail NH, Aziz AA. The prediction of students' academic performance using classification data mining techniques. *Appl Math Sci*. 2015;**9**(129):6415–6426.

12 Bakhshinategh B, Zaiane OR, Elatia S, Ipperciel D. Educational data mining applications and tasks: a survey of the last 10 years. *Educ Inf Technol*. 2018;**23**: 537–553.

13 Aldowah H, Al-Samarraie H, Fauzy WM. Educational data mining and learning analytics for 21st century higher education: a review and synthesis. *Telemat Inform*. 2019;**37**: 13–49.

14 Okewu E, Adewole P, Misra S, Maskeliunas R, Damasevicius R. Artificial neural networks for educational data mining in higher education: a systematic literature review. *Appl Artif Intell*. 2021;**35**(13):983–1021.

15 Safitri SN, Setiadi H, Suryani E. Educational data mining using cluster analysis methods and decision trees based on log mining. *J RESTI (Rekayasa Sistem dan Teknologi Informasi)*. 2022;**6**(3):448–456.

16 Klose M, Desai V, Song Y, Gehringer E. EDM and privacy: ethics and legalities of data collection, usage, and storage. In: *International Educational Data Mining Society, Paper Presented at the International Conference on Educational Data Mining (EDM),* 13th, Online; 2020 Jul 10–13. Washington, DC: ERIC; 2020.

17 Ghorpade SJ, Patil SS, Chaudhari RS. Educational data mining: tools and techniques study. *Int J Res Anal Rev*. 2020;**7**: 520–525.

18 Fischer C, Pardos ZA, Baker RS, Williams JJ, Smyth P, Yu R, Slater S, Baker R, Warschauer M. Mining big data in education: affordances and challenges. *Rev Res Educ*. 2020;**44**(1):130–160.

19 Alshehri E, Alhakami H, Baz A, Alsubait T. A comparison of EDM tools and techniques. *Int J Adv Comput Sci Appl*. 2020;**11**(12):824–831.

20 Romero C, Ventura S. Data mining in education. *Wiley Interdiscip Rev Data Min Knowl Discov*. 2013;**3**(1):12–27.

21 Zoric AB. Benefits of educational data mining. In: *Economic and Social Development: Book of Proceedings*. Varazdin: Varazdin Development and Entrepreneurship Agency; 2019. p. 1–7.

22 Romero C, Ventura S, Pechenizkiy M, Baker RS, editors. In: *Handbook of educational data mining*. Boca Raton, FL: CRC Press; 2010.

23 Bienkowski M, Feng M, Means B. *Enhancing teaching and learning through educational data mining and learning analytics: an issue brief.* Office of Educational Technology, US Department of Education. Washington, DC: ERIC; 2012.

24 Melendez-Armenta R, Huerta-Pacheco N, Morales-Rosales L, Rebolledo-Mendez G. How do students behave when using a tutoring system? Employing data mining to identify behavioral patterns associated to the learning of mathematics. *Int J Emerg Technol Learn. (iJET)*. 2020;**15**(22):39–58.

25 Hartl K. The application potential of data mining in higher education management: a case study based on German universities [dissertation]. Germany: Karlsruher Institut für Technologie (KIT); 2019. 177 p.

26 Öztürk A. *Educational data mining: applications and trends*. Anadolu: Anadolu University; 2016.

27 Alyahyan E, Düştegör D. Predicting academic success in higher education: literature review and best practices. *Int J Educ Technol High Educ*. 2020;**17**(1):1–21.

28 Nithya B, Ilango V. Evaluation of machine learning based optimized feature selection approaches and classification methods for cervical cancer prediction. *SN Appl Sci*. 2019;**1**(6):1–16.

29 Nabil A, Seyam M, Abou-Elfetouh A. Predicting students' academic performance using machine learning techniques: a literature review. *Int J Bus Intell Data Min*. 2022;**20**(4):456–479.

30 Ahuja R, Jha A, Maurya R, Srivastava R. Analysis of educational data mining. In: *Harmony search and nature inspired optimization algorithms*. Singapore: Springer; 2019. p. 897–907.

31 Sajana T, Rani CS, Narayana KV. A survey on clustering techniques for big data mining. *Indian J Sci Technol*. 2016;**9**(3):1–12.

32 Sivogolovko E, Novikov B. Validating cluster structures in data mining tasks. In: *Proceedings of the 2012 Joint EDBT/ICDT Workshops*. New York: ACM; 2012 Mar. p. 245–250.

33 Ikotun AM, Ezugwu AE, Abualigah L, Abuhaija B, Heming J. K-means clustering algorithms: a comprehensive review, variants analysis, and advances in the era of big data. *Inf Sci*. 2023;**622**: 178–210.

34 Romero C, Ventura S. Educational data mining: a survey from 1995 to 2005. *Expert Syst Appl*. 2007;**33**(1):135–146.

35 Osman AS. Data mining techniques. *Int J Data Sci Res*. 2019 Jun;**2**(1):1–4.

36 Algarni A. Data mining in education. *Int J Adv Comput Sci Appl*. 2016;**7**(6):456–461.

37 Aleem A, Gore MM. Educational data mining methods: a survey. In: *2020 IEEE 9th International Conference on Communication Systems and Network Technologies (CSNT)*. Piscataway, NJ: IEEE; 2020 Apr. p. 182–188.

38 Hicham A, Jeghal A, Sabri A, Tairi H. A survey on educational data mining [2014–2019. In: *2020 International Conference on Intelligent Systems and Computer Vision (ISCV)*. Piscataway, NJ: IEEE; 2020 Jun. p. 1–6.

39 Chaturvedi M. Data mining and its application in EDM domain. In: *2017 International Conference on Intelligent Computing and Control Systems (ICICCS)*. Piscataway, NJ: IEEE; 2017 Jun. p. 829–834.

40 Corbett AT, Anderson JR. Knowledge tracing: modeling the acquisition of procedural knowledge. *User Model User-Adapt Interact*. 1994;**4**(4):253–278.

41 Baker RS, Corbett AT, Aleven V. More accurate student modeling through contextual estimation of slip and guess probabilities in Bayesian knowledge tracing. In: *International Conference on Intelligent Tutoring Systems*. Berlin, Heidelberg: Springer; 2008 Jun. p. 406–415.

42 Hershkovitz A, de Baker RS, Gobert J, Wixon M, Pedro MS. Discovery with models: a case study on carelessness in computer-based science inquiry. *Am Behav Sci*. 2013 Oct;**57**(10):1480–1499.

43 Mehra C, Agrawal R. Educational data mining approaches, challenges and goals: a review. *JIMS8I-Int J Inf Commun Comput Technol*. 2020;**8**(2):442–447.

44 Ali F, Bhatt D, Choudhury T, Thakral A. A brief analysis of data mining techniques. In: *2019 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE)*. Piscataway, NJ: IEEE; 2019 Dec. p. 752–758.

45 Shruthi P, Chaitra B. Student performance prediction in education sector using data mining. *Intl J Adv Res Comput Sci Softw Eng*. 2016;**6**(3):123-126.

46 Ibrahim FA, Shiba OA. Data mining: WEKA software (an overview). *J Pure Appl Sci*. 2019;**18**(3):54–58.

47 Jassim MA, Abdulwahid SN. Data mining preparation: process, techniques and major issues in data analysis. In: *IOP conference series: materials science and engineering*. vol. 1090, No. 1, Bristol: IOP Publishing; 2021 Mar. 012053 p.

48 Zhu R, Hu X, Hou J, Li X. Application of machine learning techniques for predicting the consequences of construction accidents in China. *Process Saf Environ Prot*. 2021;**145**: 293–302.

49 Boateng EY, Otoo J, Abaye DA. Basic tenets of classification algorithms K-nearest-neighbor, support vector machine, random forest and neural network: a review. *J Data Anal Inf Process*. 2020;**8**(4):341–357.

50 Pal AK, Pal S. Data mining techniques in EDM for predicting the performance of students. *Int J Compu Inf Technol*. 2013;**2**(06):1110–1116.

51 Perez B, Castellanos C, Correal D. Applying data mining techniques to predict student dropout: a case study. In: *2018 IEEE 1st Colombian Conference on Applications in Computational Intelligence (Colcaci)*. Piscataway, NJ: IEEE; 2018 May. p. 1–6.

52 Anuradha C, Velmurugan T. A comparative analysis on the evaluation of classification algorithms in the prediction of students performance. *Indian J Sci Technol*. 2015;**8**(15):1–12.

53 Huimin Q, Ming C, Mingming X. A personalized resource recommendation system using data mining. *2010 International Conference on E-Business and E-Government*. Piscataway, NJ: IEEE; 2010 May. p. 5365–5368.

54 Taha SA, Shihab RA, Sadik MC. Studying of educational data mining techniques. *Int J Adv Res Sci Eng Technol*. 2018;**5**(5):5742–5750.

55 Khasanah AU, Harwati H. Educational data mining techniques approach to predict student's performance. *Int J Inf Educ Technol*. 2019;**9**: 115118.

56 Krpan D, Stankov U. Analysis of real-life experience gained from e-learning system. In: *2012 Proceedings of the 35th International Convention MIPRO*. Opatija, Croatia: IEEE; 2012. p. 753–757. doi:10.1109/MIPRO.2012.6240302.

57 Faludi A. *Planning theory*. Hoboken, NJ: J Wiley; 2013.

58 Lei XF, Yang M, Cai Y. Educational data mining for decision-making: a framework based on student development theory. In: *2nd Annual International Conference on Electronics, Electrical Engineering and Information Science (EEEIS 2016)*. Amsterdam: Atlantis Press; 2016 Dec. p. 628–641.