EDUCATIONAL DATA MINING: A COMPREHENSIVE STUDY

[1] Rajiv Kumar, [2] Er. Vani Kansal
[1] Student (M.Tech), [2] Assistant Professor
Department of Computer Science Engineering, GGSCET,
Guru Kashi University, Talwandi Saboo (Bathinda) Punjab, India.

Abstract: Research in the field of academics is developing expeditiously. Exploration and Experimentation in the domain of Educational Data Mining (EDM) is increasing on account of the various benefits achieved from the knowledge gained from the analysis of educational data with the help of Data mining (DM) processes implemented. The academic institutions can utilize EDM not only to examine the performance of students but also in improving teaching – learning process, resource utilization etc. EDM helps in creation of new methods for discovering knowledge from educational database and this helps to enhance decision making processes in educational systems at all levels. This study is an effort towards understanding the educational datamining concept and its various aspects. At the same time, the study tries to uncover number of issues that drives the need of the research community to furnish comprehensive reporting of methods and results. There is a need to increase efforts to validate the work and gives a direction for future work to be carried out in the field.

Index Terms: Data Mining, Educational Data Mining, Tools, Applications.

I. INTRODUCTION

Educational institutions impart good quality education, competent of generating and upgrading the extent of knowledge and awareness, and also the capability of the human mind. Like any other field educational institutes store huge volumes of data related to students, faculty, courses and many more. Recently, there is an increased interest in analysing this data for the improvement of academics and decision making at higher levels. Data Mining (DM), also known as knowledge discovery from databases, is the automatic procedure of extracting implicit, potentially useful information from data, for better decision-making [1]. The primary function of data mining is the application of specific methods to develop models and discover previously unknown patterns [2]. The use of DM in education is referred as Educational Data Mining (EDM). EDM brings into use various method to retrieve beneficial patterns and information from comprehensive databases of educational institutes. As per Ahmad et al. [3], Higher educational institutions store large amount of data in their databases comprising data about courses, students, etc. and this data keeps on multiplying with time but, no action is taken to utilize this data to acquire knowledge. Data mining is assemblage of various relevant techniques belonging to statistics, machine learning, visualization etc. which are being used for extracting and discovering knowledge in manner that humans can interpreted easily [1][2]. EDM is an interdisciplinary area inclusive of recommender system, information retrieval, domain-driven DM, visual data analytics, social network analysis (SNA), and many more. In fact, EDM can be seen as amalgamation of 3 areas as shown in Figure 1: computer science, statistics and education and combination or intersection of these main areas form subareas which are quite related to EDM [2].

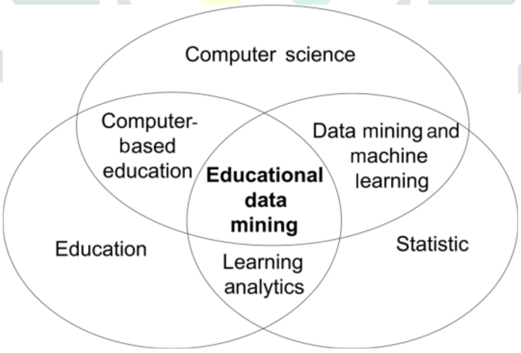

Fig. 1: Areas and subareas related to EDM [2]

In this study an effort is made to understand EDM, its concept, components, applications, and tools on the basis of various studies performed by the researchers in this area. Study also tries to uncover various issues which gives a forward direction for the future research work. The paper is organised in different section starting with introduction in Section 1, followed by Section 2: literature review, EDM objectives are covered in section 3 , followed by EDM process in Section4, Section 5: features used, Section 6: methodologies, Section 7: EDM tools and Section 8: Applications ,Section 9: issues and conclusion

II. LITERATURE REVIEW

C. Romero et al. [1], in their paper performed a detailed survey of usage of data mining in various educational systems such as traditional, and in particular web-based courses, LMS and so forth. These educational systems have distinctive objectives, sources of data and features, for the process of knowledge discovery. They elaborated various aspects in their broad survey and discussed that the success various EDM research work needs specialized work. Acc. to Baker and Yacef [4], the EDM researchers are mostly attracted by the areas like use of educational software in individual learning, computer-supported learning, factors effecting student performance, drop-out and so forth. Peña Ayala [5], in his analysis, observed that student performance modelling is favourite area of this domain approach. The researcher point outs various performance indicators that deserve to be analysed and modelled these are namely; time elapsed, evaluation, competence, efficiency, achievement, and resource consuming [5]. Yadav et al. [6] uses DT classifiers, to find out the student drop out of an institution. He was able to predict by using EDM. DM methods were implemented by Al-Radaideh et al. [7]to analyse and evaluate student data for identification of main or key features influencing the performance of students. In [8], DM technique were implemented to analyse the Bulgarian university student data. The dataset collected was contained pre-admission and the personal information students. Different Classifiers like k-Nearest Neighbour, D Tree, Naïve Bayes classifiers, Bayesian, the JRip, and One R were implemented to extract knowledge with of accuracy 67%. The result obtained showed that the most influential features were no. of course failed in 1st year and the admission score. Yassein et al. [9], studied that the student performance depends on numerous diverse aspects namely; personal, demographical, socio-economic and various environmental factors. The analysis of these aspects and knowledge obtained can influence the entire educational system of an institute. Acc. to Pratiyush and Manu [10], EDM of data helps in organising the resource usage associated with student performance, placement prediction and evolving trends in educational sector. The research considers placement data of students and uses SVM as classification method. Patil et al. [11], selected student data of engineering course due to its increasing popularity, but due to several factors related to education system in India there is increase in the dropout rate. With the usage of DM techniques, the students' dropout rate as well as grade in a subject can be predicted. Author uses Naives bayes classifier in this research. Aderibigbe Israel Adekitan [12], in their study, carried out predictive analysis to find the extent up to which the final Cumulative Grade Point Average (CGPA) and 5th year marks of engineering students of Nigerian University can be found using the study, using GPA and first three years of study as input to 6 DM algorithms implemented using KNIME. The maximum 89.15% accuracy was achieved. Balwinder Kaur et al. [13], have performed a comprehensive study on EDM and its environment and present's an overview of its applications, Tools, techniques and issues. Dakić Dušanka et al [14]. performed a detailed comparative analysis of various free and open source contemporary DM tools and are compared against platforms, visual capabilities, programming Languages and so on. The study helps to learn about various tools which helps in understanding their capabilities and hence selection.

III. OBJECTIVES OF EDM

In EDM, the objectives can be divided into two categories these are [15]:

i) Applied research objectives: such as enhancing student learning process as well as supervising and managing students' learning,

ii) Pure research objectives: like attaining a comprehensive insight into educational phenomena.

From the study of various papers, the objectives of EDM based on perspectives of various researchers can be divided on the basis of the type of user/ stakeholder [1,2, 4,13]-

Learners: To help or assist a student/ learner's and personalize the e-learning experience, to provide constructive feedback or to suggest interesting learning paths to learner, to recommend courses, to enhance learning process and so forth.

Educators: To get feedback about the educator, to understand and analyse student learning behaviour and pattern which can reflect on the teaching methods of educators, to classify students and predict their performance, to improve courses and so on.

Researchers: To construct student and course models, to develop or build and compare DM techniques, to recommend the appropriate and relevant data mining technique to a particular task and many more.

Administrators: To assess the most suitable way to arrange institutional resources like material and human resource, and to utilize these available resources in an effective and efficient manner.

IV. EDM PROCESS

The EDM process is an iterative process. To build reliable models, Knowledge Discovery (KD) and DM process are considered. The KD process consists of hypothesis formulation, testing and refinement [16]. The process of knowledge discovery also includes the following steps: data collection, pre-processing, data mining, and interpretation of the results. The application process of DM techniques to educational systems can be explained from different perspectives [15] as shown in figure 2.

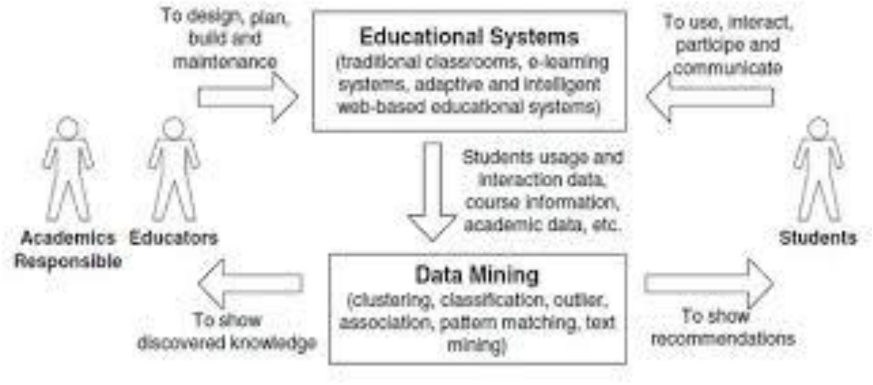
To design, plan, build and maintenance
Educational Systems
(traditional classrooms, e-learning systems, adaptive and intelligent web-based educational systems)
To use, interact, particpe and communicate
Students usage and interaction data, course information, academic data, etc.
Academics Responsible
Educators
Students
To show discovered knowledge
Data Mining
(clustering, classification, outlier, association, pattern matching, text mining)
To show recommendations

Fig. 2: EDM process with its environmental components.[1]

According to experimental and educational perspective: it is a repetitive process of formulating hypothesis, testing, and refinement as shown in Figure 3 [2]. The process is implemented with an objective to transform data into detailed knowledge, and to select or filter the knowledge mined from the process, for decision-making in educational system, to enhance overall student's learning as well as educational system. It is a kind of formative evaluation with the goal of continually improving the academic programs. Analysis of how the learners or students are using these systems is one of the ways to evaluate instructional design and it may assist the educational designers to upgrade instructional contents and materials. For instance, EDM methods like discovery with models or patterns can be brought into usage to help educational designers to initiate a pedagogical basis for decisions making at the time of designing or refining pedagogical approach [2].

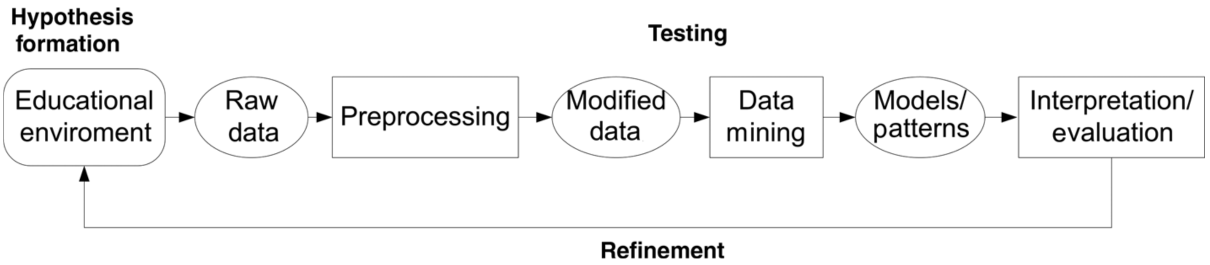

Fig. 3: knowledge discovery and DM process in Education [2]

V. FEATURES

EDM analyse the data collected and produced by different kinds of education and information system used in academics and supporting learning. The education system can be traditional like in schools and colleges etc or informal like e-learning etc. The data produced is not confined to activities of an individual students like quiz, interaction with e-learning or navigation behaviour, but also include data from group activities, administrative data, demographic data (like student gender, DOB, etc.), academic data (school marks or percentage), student affectivity (e.g., motivation, emotional states), and so on [9]. These features have distinctive characteristics namely; multiple levels of hierarchy, context, granularity and longitudinal. Broadly, the features or the data used can be divided into following categories [9][17]:

i. Pedagogical or Scholastic (e.g. secondary or senior secondary school performances, subject/course performance).
ii. Demographic (e.g. gender, locality, gender).
iii. Personality (e.g. self-motivation, regulation, self-efficacy).
iv. Institutional (e.g., quality of high-school, teaching-learning approach).
v. Behavioral (e.g. Social data, LMS data).

VI. METHODOLOGIES OF EDM

EDM methods/ Techniques come from different fields like computational modelling, statistics, visualization, so forth. Romero and Ventura [1], in their work categorize EDM methods into the following groups:

a. Classification and Clustering
b. Statistics and Visualization
c. Sequential Pattern Mining, Association rule mining and sequential pattern mining
d. Text mining

Baker[4], defined a new topology according to which the EDM methods were divided into Prediction (Regression, Density estimation, and Classification,), Clustering, Outlier Detection, Relationship mining (Association rule mining, Sequential pattern mining, Correlation mining, and Causal data mining), Discovery with model and Distillation of data for human judgment. Further, from the literature these used methodologies can be split into following [1][2][4][13][18]:

Prediction: in prediction, the model infers a target label or attribute known as predicted variable from the combination of some other predictor variables. The prediction is further divided into following types: classification, regression (used when predicted variable is continuous), or density estimation [2][13].

Classification: It is a supervised learning technique. The input to classification is a data set with known class labels called training set. On the basis of training set classification technique creates a model which is capable to assign the labels to unlabelled data, e.g., Decision Tree, Rule, Naive Bayes, and so forth [18].

Clustering: it is unsupervised learning technique. In this data is grouped into categories based on the similarity between the objects, such that objects within a group share common properties. In clustering while making groups the intra -cluster distance should be minimum and inter- cluster distance should be maximum. It constitutes major class of DM [18].

Outlier detection: it is to identify the outlying points of the dataset and to identify them as noise or impurity. It helps in fraud detection, etc.

Statistical: Support vector machine is learning theory which is based on statistical theory and is being frequently used these days. It uses the concept of hyper-planes and margins.

Relationship mining/ Sequential Mining: it is concerned with mining the sequential data to find the linkage or relationship between the various features of the data set. There are various other types of methods under this category namely; association mining (association rules the association between items is found irrespective of their order of occurrence), Sequential pattern mining (temporal linkage among data attributes), Correlation mining, and Causal data mining [2].

Discovery with Model: a model is build using techniques such as prediction, regression, classification, and so on. The created model is then used as one on the component in developing another model [2].

Distillation of Data for Human Judgment: The method aims to present information using techniques like visualization, summarization, and so forth to support DSS and to highlight important information [2].

Classification is one of the most frequently used and compared technique. Under classification Decision trees are the frequently used due ease to understand and its simplicity. It has a flowchart type structure where each internal node is a test attribute and the branch indicate the result. The class labels are the terminal nodes [6][18]. Neural networks (NN) are also one of the essential classification tools. NN are self-adaptive and data driven, they are able to adjust themselves to data. Naïve Bayes (NB) is a probabilistic classifier. Capable of performing well in real world applications despite Its assumption of feature independence.

VII. EDM TOOLS

Due to wide spread use of data analysis and datamining, there are a lot of open source software's available under GNU (general public licence) and commercial tools. According to the categorisation of various DM methods, these tools are also categorised into different categories like Visualization and Statistical tool, Classification and clustering tool, text mining tool, Association and sequential mining tool and others (the tools that don't fall under these specified heads). A few tools are listed in table 1 below [13][14]:

Table 1: Commonly used free/ open source software tools

Tool Name	Description
WEKA	Built in Java, it is an assemblage of various ML algorithms for data mining tasks.
RapidMiner	Is a powerful visual workflow designer for building predictive analytic workflows
SAS (Statistical Analysis System)	SAS can mine data, manage data and alter data, from different sources and perform statistical analysis.
R language	R/ R-studio is a free/ open source software environment capable of performing statistical analysis and graphics.
Python	Python is an open source software under open source license, making it freely usable and distributable. It delivers a vast library support.
Konstanz Information Miner (KNIME)	KNIME (Konstanz Information Miner) is a popular Java based, modular DM application which provides interactive, visual, easy assembling, testing, and executing data mining pipelines.

VIII. EDM APPLICATIONS

There are many applications of educational datamining. Different methods are used in these applications depending upon the data, and specific domain. EDM is capable of addressing different types of problem related to academic / educational environment. EDM has a vast application area out of which a few applications are listed below in the table2 [2][4][5][13][17]:

Table 2: A few applications of EDM

Application	Description
Predicting academic performance	To predict the final grades, knowledge or any other type of learning outcomes which important for student and institutes.
Placement of Students	To predict the placement of the students, the job profile and so forth
Analysing and interpreting the student information	The aim is to highlight hidden yet important information and to support the decision-making task.
Recommendation System	To give recommendations to students with respect to contents, tasks, problems or courses which are most appropriate.
Personalizing the learning experience of student	To alter/redesign automatically; navigation, content presentation, learning and so on, with respect to each and every student.
Student modelling	The motive is to develop student model taking into consideration user characteristics and their learning behaviour.
Grouping Students	It creates groups or batches of students on the basis of common features like learning patterns, personal characteristics and so forth
Domain Modelling	It aims at describing the domain of instruction with regard to skills, learning items and their interrelationships, concepts etc.
Courseware/ curriculum construction	The aim is to automate and help faculty / teachers to develop the course contents according to the need of students and requirement of society or industry.

IX. ISSUES

EDM is a well-established research area consisting of e-learning, learning management system, adaptive system and so forth. EDM has specific requirements it has to take into consideration the pedagogical needs of not only the learner but the entire system. There are present some issues which are required to be tackled and they provide a future direction also. A few of the issues are discussed below [8][13][17][19]:

1. From the review study it is found that there exist number of ethical issues present in various research in EDM like absence of inclusion, lack of consent, unethical practices while gathering data, and many other related issues.
2. The usage of student behavioural data in a course (log data), is gaining attention within the EDM and education research domain, but still is comparatively used less.
3. The greater part of the studies in EDM are focusing on a individual subject, course or an individual institution, having no discrete data or population with which the work would have been replicated and compared with.
4. Standardization of methods, models and data is also one of the issues which is required to dealt with. Current tools for mining data are mainly useful to developers only. There is lack of general tools or reusable tools or techniques that can be easily applied to any academic system. So, there is a need for standardization of data, pre-processing, discovering and postprocessing tasks.

Currently large amount of work is being carried out but there is scope of further improvement. DM tools should be able to facilitate all the EDM process to the educators. Further, the results obtained from EDM research are to be generalised and there is a need to carry out studies to test the models on broader areas to validate the model.

X. CONCLUSION

Analysis of educational data as well as student performance prediction are crucial factor in educational environment. In this paper an effort is made to provide a comprehensive study about the EDM, its objectives, methods, tools and techniques, the type of feature used and the process. Educational data mining helps us in analysing the educational data related to students, courses etc using various data mining techniques like classification, clustering, Prediction and so forth. EDM is a research area which has wide applications related to education and its environment. New researches are being carried out in EDM by various researchers, but at same time, the study tries to uncover number of issues that are required to be handled and they also directs the need for the society to furnish more elaborated reporting of techniques or methods as well as results and to boost the efforts to approve and accept the work and gives a direction for future work to be carried out in the EDM field.

XI. ACKNOWLEDGEMENT

I would like to extend my sincere thanks to Er. Vani Kansal, Assistant Professor, Guru Kashi University, Talwandi Saboo, Bathinda, Punjab for her guidance and help in my study.

REFERENCES:

[1] C Romero, and S Ventura, “Data mining in education”, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, Vol. 3, Issue. 1, pp. 12-27, 2013.

[2] C. Romero, and S. Ventura, “Educational Data Mining: A Survey from 1995 to 2005”, Expert Systems with Applications, Vol. 33, pp. 135-146, 2007.

[3] Ahmad, F., Ismail, N., and Aziz, A. A., “The prediction of students’ academic rerformance using classification data mining techniques”, Appl. Math. Sci., Vol. 129, pp :6415–6426, 2015.

[4] Baker R. S., and Yacef K., “The state of educational data mining in 2009: A review and future visions”, *Journal of Educational Data Mining (JEDM)*, Vol. *1, Issue.* 1, pp. 3-17, 2009.

[5] A. Peña-Ayala, “Educational data mining: applications and trends”, Vol. 524, Springer, 2013.

[6] Yadav SK, Bharadwaj B, Pal S.,” Data mining applications: A comparative study for predicting student's performance.”, arXiv preprint arXiv:1202.4815, 2012 Feb 22.

[7] Al-Radaideh, Qasem A., Emad M. Al-Shawakfa, and Mustafa I. Al-Najjar., "Mining student data using decision trees.", In proceedings *International Arab Conference on Information Technology (ACIT'2006), Yarmouk University, Jordan*, 2006.

[8] Dorina Kabakchieva, "Predicting student performance by using data mining methods for classification.", Cybernetics and Information Technologies, Vol. 13, no. 1, pp. 61-72, 2013.

[9] Yassein NA, Helali RG, Mohomad SB, “Predicting student academic performance in KSA using data mining techniques.”, Journal of Information Technology and Software Engineering, Vol 7, Issue 5, pp.:1-5, 2017.

[10] Pratiyush, G., and Manu, S., “Classifying educational data using support vector machines: A supervised data mining technique”. Indian Journal of Science Technology, Vol. 9, Issue. 34, 2016.

[11] Patil, V.R., Suryawanshi, S.A., Saner, M., Patil, V.C., & Sarode, B. “Student Performance Prediction Using Classification Data Mining Techniques.”, International Journal For Research In Emerging Science And Technology, Volume-4, Issue-4, Apr-2017

[12] Aderibigbe Israel Adekitan, Odunayo Salau., “The impact of engineering students’ performance in the first three years on their graduation result using educational data mining.” Heliyon 5, 2019, e01250.

[13] Kaur Balwinder, Gupta Anu, Singla R.K., “An Insight into Educational Data Mining.”, International Journal of Computer Sciences and Engineering, Vol.7, Issue.2, pp. 83-90, 2019.

[14] Dušanka, Dakić, Stefanović Darko, Sladojević Srdjan, Arsenović Marko, and Lolić Teodora. "A comparison of contemporary data mining tools.", In Proceedings XVII International Scientific Conference on Industrial Systems (IS'17), Novi Sad, Serbia, vol. 4, Issue no. 6. 2017.

[15] Romero, C., Ventura, S., “Educational data mining: a review of the state of the art.”, *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, Vol. *40, Issue.* 6, pp. 601-618, 2010.

[16] Kurgan, L., Musilek, P., “A survey of knowledge discovery and data mining process models.”, *The Knowledge Engineering Review, Vol. 21, Issue.* 1, pp. 1-24, 2006.

[17] Hellas, A, Ihantola, P, Petersen, A, Ajanovski, VV, Gutica, M, Hynninen, T, Knutas, A, Leinonen, J, Messom, C and Liao, SN, “Predicting Academic Performance: A Systematic Literature Review.” In *Proceedings Companion of the 23rd Annual ACM Conference on Innovation and Technology in Computer Science Education*, ITiCSE 2018 Companion, ACM, New York, NY, USA, pp. 175-199, 23rd Annual Conference on Innovation and Technology in Computer Science Education (ITiCSE 2018), Larnaca, Cyprus.

[18] J. Han, J. Pei, M. Kamber, “Data mining: concepts and techniques”, Elsevier, 2011.

[19] K. Venkatachalapathy, V. Vijayalakshmi, and V. Ohmprakash, "Educational Data Mining Tools: A Survey from 2001 to 2016", Second International Conference on Recent Trends and Challenges in Computational Models (ICRTCCM), 2017.