The State of the Art on Educational Data Mining in Higher Education

Mohamed Osman Hegazi[#1], Mazahir Abdelrhman Abugroon[*2]

[#1]Department of Computer Science, College of Computer Engineering and Science, Prince Sattam Bin Abdulaziz University, Saudi, Arabia

[*2] Department of Computer, College of Science, Hafar elbatin University, Saudi Arabia

Abstract *Educational data mining (EDM) is a broader term that focuses on analyzing, exploring, predicting, clustering, and classification of data in educational institutions. EDM grows faster and covers many interdisciplinary such as education, e-learning, data mining, data analysis, intelligent system etc... The paper presents most relevant work in the area of EDM in higher education it covers course management systems, student behaviors, decision support system, and student retention and attrition. The paper also provide a comparison study between some of research work in such areas. Because of the growth in the interdisciplinary nature of EDM the paper, also try to provide boundary scope and definitions for EDM.*

Keywords *Data Mining , DM, Educational Data Mining, EDM, Knowledge Discovery, KDD, Decision Support System, DSS, Course Management Systems, CMS.*

I. EDUCATIONAL DATA MINING DEFININTION

Several definitions of educational data mining has been used. Most of these definitions has been concentrated on the relation of data mining with education, and some of them oriented to the analytical processing of the educational data, while EDM is grows faster and have been merge with several areas such as education, e-learning, data mining, data analysis, intelligent system and etc., boundary scope and definitions of EDM is needed. This section provides an over view of some difference definition of EDM in some related work, and finally try to provide scope and definitions of EDM

The Educational Data Mining community website, [1] defines educational data mining as: “Educational Data Mining is 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.”

Barnes and other [two] define education data mining as the application of data mining techniques for educational data to analyze this type of data in order to resolve issues of educational research.

Campbell and Oblinger [3] concede red the academic analytics as a sub-field of educational data mining, because they considered the academic analytics as the used of the statistical techniques and data mining in ways that will help faculty and advisors become more proactive in identifying at-risk students and responding accordingly. In this way, the results of data mining can be used to improve student retention. Academic analytics focuses on processes that occur at the department, unit, or college and university level. This type of analysis does not focus on the details of each individual course, so it can be said that academic analytics has a macro perspective.

Baker and Yacef [4] defined educational data mining 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”. Their definition does not mention data mining, leaving researchers open to exploring and developing other analytical methods that can be applied to educationally related data. Also, many educators would not know how to use data mining tools, thus there is a need to make it easy for educators to conduct advanced analytics against data that pertains to them (such as online CMS data, etc.).

From these definitions, we can conclude that educational data mining is a broader term that focuses on nearly any type of data in educational institutions, while academic analytics is specific to data related to institutional effectiveness and student retention issues. The discipline relies on several reference disciplines and in the future, there will be additional growth in the interdisciplinary nature of EDM. As the discipline grows, researchers will need to refine the scope and definitions of EDM.

As a conclusion, we can scope and define EDM as the process and the implementation of data mining concepts and techniques in educational data.

II. HISTORY OF EDUCATIONAL DATA MINING

The EDM literature draws from several reference disciplines including data mining, learning theory, data visualization, machine learning and psychometrics**.** Some of the earliest works are coming under the umbrella of Artificial Intelligence in Education

While educational data analysis itself is not a new practice, recent advances in educational technology, including the increase in energy account and the ability to record detailed data on the use of students to an environment-based learning on the computer, has led to increased interest in developing technologies

This interest appeared into a series of EDM workshops held from 2000-2007 as part of several international research conferences.[1] In 2008, a group of researchers established what has become an annual international research conference on EDM, the first of which took place in Montreal, Canada.[5]

Academic journal in 2009 and reached the attention of the administrative organization to increase, researchers EDM. In 2011, the researchers established the EDM "educational data mining of the International Society" to link researchers and the EDM continues to grow in the field.

In 2008 , public data sets made educational extract more data exists and feasible, to contribute to the growth, the introduction of public educational data warehouses, such as the Pittsburgh Science of Learning Center (PSLC) replaced the data and the National Center for Education Statistics (NCES).[4]

Recent literature on educational data mining is an emerging discipline that focuses on the application of tools and techniques for educationally relevant data and data extraction. Researchers within EDM focus on topics ranging from use of data mining to improve institutional effectiveness of application of data mining to improve learning for students for example [6], [7], [8], [9], [10], and [11]

III. EDUCATIONAL DATA MINING AND HIGHER EDUCATION

This section presents some **r**elated studies on educational data mining in higher education, the section categorized these studies in according to the educational areas, providing an over view of each application and techniques used in these areas.

A number of publications about EDM in higher education has grown exponentially over the past few years.

Romero & Ventura [12] mentioned the compression, in higher educational institutions to provide up – to date information on institutional effectiveness. Institutions are also increasingly held accountable for student success It also increases the responsibility of educational institutions for the success of the student [3].

One response to this pressure is to find new ways of applying the methods of analysis and data mining to educationally relevant data.

At an early stage, Baker & Yacef [4] mentioned that it would be helpful to have a more thorough taxonomy of the different areas of study within EDM even though researchers have already established a basic taxonomy. One drawback to Baker and Yacef's taxonomy is that it does not address aspects of the clustering data-mining task.

There are different ways that educational data mining can be used in higher education institutes. Such as Course management systems, Student behaviours, decision support system in higher education student retention and attrition

A. Course management systems

A large number of researchers within EDM focus directly on course management systems and how they can be improved to support student learning outcomes and student success.

In this area, educational data mining (EDM) has been used for utilizing data mining techniques and research approaches for understanding how students learn. Interactive e-learning methods and tools have opened up opportunities to collect and scrutinize student data, to ascertain patterns and trends in these data, and to formulate new discoveries and test assumptions about how students learn. Technology enhanced learning relies heavily on learning management systems (LMS) or course management systems (CMS). These LMS/CMS automatically record the keystrokes of individual users as server logs. Mining these logs provide patterns for teachers to identify slow learners and can adjust teaching strategies.

Rabbany, Reihaneh, Mansoureh Takaffoli, and Osmar R. Zaïane [13] suggests an application for data mining, to study participation on-line courses. This article proposes and visualization EDM's toolbox influential to the formation of the community.

Course management systems (CMSs) can tender a great variety of procedure and workspaces to facilitate information sharing and communication among participants in a course.

The Course management systems (CMSs) also let educators distribute information to students, produce content material, prepare assignments and tests, engage in discussions, manage distance classes and enable collaborative learning with forums, chats, file storage areas, news services, etc. Some examples of commercial systems are Blackboard, WebCT and Top-class while some examples of free systems are Moodle, Ilias and Claroline. The most commonly used one is Moodle (modular object oriented developmental learning environment), a free learning management system, enabling the creation of powerful, flexible and charming online courses and experiences. These e-learning systems accumulate a large amount of information, which is very valuable for analyzing students' behaviour, and could create a goldmine of education data. Record any student activities involved, such as reading, writing, taking tests, performing various tasks, and even communicating with peers, they provide a database that stores all the system's information: personal information about the users (profile), academic results and users' interaction data. However, due to the vast quantities of data, these systems can generate daily, it is very difficult to manage manually. Instructors and course authors require tools to assist them in this task, ideally on a continual basis. Although some podium offers some reporting tools, it becomes hard for a tutor to extract useful information when there are a great number of students. They do not provide specific tools allowing educators to thoroughly track and assess all learners' activities while evaluating the structure and contents of the course and its effectiveness for the learning process [13], [14].

In the last few years, researchers have begun to investigate various data mining methods to help instructors and administrators to improve e-learning systems. Data mining or knowledge discovery in databases (KDD) is the automatic extraction of implicit and interesting patterns from large data collections. Data mining is a multidisciplinary area applying to many different branches of education in which several computing patterns converge: decision tree construction, rule induction, neural networks, instance-based learning, bayesian learning, logic programming, statistical algorithms, etc. In addition, some of the most useful data mining tasks and methods are statistics, visualization, clustering, classification and association rule mining. These methods uncover new, interesting and useful knowledge based on students' usage data. Some of the mains e-learning problems or subjects to which data mining techniques have been applied are dealing with the assessment of student's learning performance, provide course adaptation and learning recommendations based on the students' learning behaviour, dealing with the evaluation of learning material and educational web-based courses, provide feedback to both teachers and students of e-learning courses, and detection of atypical student's learning behavior. [17]

Most of the current data mining tools are too complex for educators to use and their features go further away the scope of what an educator might require. As a result, the CMS administrator is more likely to apply data mining techniques in order to produce reports for instructors who then use these reports to make decisions about how to improve the student's learning and the online courses. This knowledge, however, can be useful not only to the providers (educators) but also to the users themselves (students), as it can be oriented towards different ends for different partakers in the process [17]. It could be oriented towards students in order to recommend learners' activities, resources, suggest path pruning and shortening or simply links that would favour and improve their learning or to educators in order to get more objective feedback for instruction. Instructors can evaluate the structure of course content and its effectiveness in the learning process and classify learners into groups based on their needs for guidance and monitoring. Learners' regular and irregular patterns could be determined allowing the most frequently made mistakes to be identified and more effective activities to be elaborated. There could be more trends towards obtaining parameters and measures to improve site efficiency and adapt it to the behaviour of the users (optimal server size, network traffic distribution, etc.) and to organize better institutional resources (human and material) and educational offer. [17, 14]

Data mining has been applied to data coming from different types of educational systems. On one hand, there are traditional face-to-face classroom environments such as special education [39] and higher education [34]. On the other, there is computer-based education as well as web based education, such as well-known learning management systems [40], web-based adaptive hypermedia systems [41] and intelligent tutoring systems [42]. The main difference between one and the other is the data available in each system. Traditional classrooms only have information about student attendance, course information, curriculum goals and customized plan data. However, computer and web-based education has much more information available because these systems can record all the information about students' actions and interactions onto log files and databases.

One research team developed a simplified data mining toolkit that operates within the course management system and allows non-expert users to get data mining information for their courses [14]. In addition, a toolkit allows teachers to collaborate with each other and share results. This research is important because most data mining tools are complicated and require deep expertise in data mining tools, methods and processes, statistics, and machine learning algorithms. This study follows a typical data mining process, thus it is quantitative. The data mining process usually follows a pre-processing phase, then an application of specific data mining techniques, and then a post -processing phase. The research and application contributions will allow non- technical faculty to engage in educational data mining activities. It is clear that additional is needed in this area to make educational data mining tools more accessible to non-technical users. [14], [16]

Course management systems such as open source Moodle can be mined for usage data to find interesting patterns and trends in student online behaviour. A systematic method for applying data mining techniques to Moodle usage data was established [14]. The benefit to mining usage data is that it contains data about every user activity, such as testing, quizzes, reading, and discussion posts. The authors discuss the importance of pre-processing the data and then discuss specifics on how to apply data mining techniques to Moodle data. Their research results demonstrated how straightforward, it is to mine data, even if a reader does not have much experience in this area. The authors also use both Keel and Weka as their data mining software packages. These software programs are open source and build on the Java language, so they are extendable as well.

Y.-h. Wang & Liao, [18] mentioned that data mining can be used in such a way as to customize learning activities for each individual student. Data mining was used to adapt learning exercises based on students' progress through a course on English language instruction Instead of having static course content, the course adapts to student learning, taking him or her through the course at his or her own pace. This was an effort to create significant and optimal learning experiences for each student, and was a success. This research could be applied to other types of courses where students begin a course with varying levels of competency, e.g., a computer programming course. [18]

B. *Student behaviours*

Beck and Woolf [19] cited in their article, how educational data mining prediction methods can be used to develop student models. They use a variety of variables to predict whether a student will make a correct answer. This work has inspired a great deal of later educational data mining work – student modelling is a key theme in modern educational data mining, and the paradigm of testing EDM models' ability to predict future correctness

Data mining was used to assess complex student behaviours with respect to a three –week programming assignment. Blikstein [20] found results that showed different types of student programming behaviours in an online course. These log files contained different types of events as each student completed them. The events included coding and non - coding activities in the online course. This quantitative data mining research helped discover different programming strategies used by students, and developed three programming behaviour profiles: copy-and- pasters, mixed-mode, and self-sufficient [20].

In many online courses, discussion board posts are an important part of the learning experience. One research team used data mining as a strategy for assessing asynchronous discussion forums because it was challenging to manually assess the quality of the postings by each student [15]. Their research attempts to answer the question of what kind of information is embedded in online discussion groups. The data mining results were used to assess student progress in an online course. One drawback with this approach is that non-technical faculty would not know how to apply data mining to get results for their students, thus there is a need to create tools that are accessible to non-technical faculty members.

Like Blikstein [20], Dringus and Ellis [15] analyze student behaviour by applying data mining techniques. While the former examines programming activity, behaviour, the latter examines discussion board behaviour. The analysis is different based upon the type of task or activity. For example, the DM analysis programming tasks in a course management system is going to be different than the DM analysis for discussion boards. Each data mining task is usually very specific and is used with a specific data set However, may be more important to find ways of applying data mining to examine students' behaviour in a broader sense, rather than analyzing a single aspect of their behaviour within the CMS. In other cases, the data is less fine-grained. For example, a student's university transcript may contain a temporally ordered list of courses taken by the student, the grade that the student earned in each course, and when the student selected or changed his or her academic major. EDM, influence both types of data to discover meaningful information about different types of learners and how they learn, the structure of domain knowledge, and the effect of instructional strategies embedded within various learning environments. These analyses provide new information that would be difficult to discern by looking at the raw data. For example, analyzing data from an LMS may reveal a relationship between the learning objects that a student accessed during the course and their final course grade. Similarly, analyzing student transcript data may reveal a relationship between a student's grade in a particular course and their decision to change their academic major. Such information provides insight into the design of learning environments, which allows students, teachers, university administrators, and educational policy makers to make informed decisions about how to interact with, provide, and manage educational resources.

C. Decision support system in higher education

Other areas within EDM include analysis of educational processes including admissions, alumni relations, and course selections. Furthermore, applications of specific data mining techniques such as web mining, classification, association rule mining, and multivariate statistics are also key techniques applied to educationally related data [22]. These data mining methods are largely investigation techniques that can be used for prediction and forecasting of learning and institutional improvement needs. Also, the techniques can be used for modelling individual differences in students and provide a way to respond to those differences thus improve student learning. Although, one question is how institutions do adopt educational data mining to improve institutional effectiveness?

In order for educational data mining to be successful, it is critical to have a solid data warehousing strategy. Guan et al. [23] discussed how important it is to have meaningful information available for decision- makers within higher educational institutions. It is a defiance to get the information that decision makers need quickly and efficiently. Some of the primary drivers of initiating data warehouse projects include increased competitive landscape, and increased responsibilities of reporting to external stakeholders such as parents, board members, legislators and community leaders [23]**.**

Many researchers also have contributed to the field of educational data mining and decision support system in higher education.

Chau and Phung [24] in their study stated that education always plays an important role in building up every country around the world. According to them, educational decision making support is significant for students, educators, and educational organizations and the support will be more valuable if a lot of relevant data and knowledge mined from data are available for educational managers in their decision making process. They proposed a knowledge-driven educational decision support system. The knowledge-driven decision support is helpful for educational managers to make more appropriate and reasonable decisions about the student's study and further give support to students for their graduation.

A waste of effort, time, and money can be avoided accordingly for both students and educators through the proposed system.

Deniz and Ersan [25] presented different ways in which student performance data can be analyzed and presented for academic decision making. They demonstrated the usefulness of an academic decision-support system (ADSS) in evaluating huge amounts of student-course related data. In addition, they presented the basic concepts used in the analysis and design of a specific DSS software package called the Performance based Academic Decision Support System (PADSS).

The study conducted by Feghali, Zbib and Hallal [26] attempts to solve a technology-based “last mile” problem by developing and evaluating a web-based decision support tool – the Online Advisor, that helps advisors and students make better use of an already present university information. Their study showed that 79% of users stated that they were satisfied with the Online Advisor, 90% rated the Online Advisor as effective and efficient and more than 75% rated the Online Advisor as useful and helpful.

According to Kotsiantis [27], the use of machine learning techniques for educational purposes or educational data mining is an emerging field aimed at developing methods of exploring data from computational educational settings and discovering meaningful patterns. The stored data can be useful for machine learning algorithms. Students’ key demographic characteristics and their marks in a small number of written assignments can constitute the training set for a regression method in order to predict the student’s performance. A prototype version of software support tool for tutors has been constructed.

Nagy, Aly and Hegazy [28] proposed a “Student Advisory Framework” that utilizes classification and clustering to build an intelligent system. The system can be used to provide pieces of consultations to a first year university student to pursue certain education track where he/she will likely to succeed in. According to them, one of the main reasons for the high failure rate is the incorrect selection of the student’s department/section. The framework acquires information from the datasets that stores the academic achievements of students before enrolling in a certain department. After acquiring all the relevant information, a new student can challenge the intelligent system to receive a recommendation of a certain department in which he/she would likely succeed.

Vinnik and Scholl [29] proposed a methodology for assessing educational capacity and planning its distribution and utilization in universities. They integrated educational data mining and knowledge discovery in their proposed method. The DSS supports the administrative task of planning the university’s educational capacity in terms of the number of students its courses can accommodate under the specified constraints. Decision-makers were able to evaluate various strategies and generate forecasts by means of simulating with the input data. According to them, when the policy makers applied the system, it has resulted in significant acceleration in planning procedures, raised the overall awareness with respect to the underlying methodology and ultimately enabled more efficient academic administration.

Some decision support systems that use data mining have already been developed and introduced in the literature. The system was designed to support tactical decisions of a basketball coach during a basketball match through suggesting tactical solutions based on the data of the past games. The decision support system only supports the association rule data mining method and uses the association rule algorithm called *Apriori* algorithm combined with the *Decision* *query* algorithm. The decision support system enables the coach to submit data about his tactical strategies and data about the game and the rival team. After that, the system provides the coach with an opinion about the chosen strategies and with suggestions. The system is not designed to support other domains; it only supports the basketball domain.

Bose and Sugumaran introduced the Intelligent Data Miner (IDM) decision support system [30]. IDM is a Web-based application system intended to provide organization-wide decision support capability for business users. Besides data, mining it also supports some other function categories to enable decision support: data inquiry and multidimensional analysis through enabling OLAP views of multidimensional data. In the data mining part of IDM it supports the creation of models, manipulation of models and presentation of models in various presentation techniques of, among others, the following data mining methods: association rules, clustering and classifiers (classification). The system also performs data cleaning and data preparation and provides necessary parameters data mining algorithms. An interesting characteristic of IDM is that it makes a connection to an external data mining software tool, which performs data mining model creation. The system enables predefined and ad-hoc data mining model creation. The authors state that the disadvantage of IDM is the fact that nontechnical users (business users) need to have a fair amount of understanding of data mining and that the use of data mining and the creation of data mining models still needs to be clearly directed by the user, especially with ad-hoc model creation.

Lee and Park presented the Customized Sampling Decision Support System (CSDSS), which uses data mining [31]. CSDSS is a web-based system that enables the user to select a process sampling method that is most suitable according to his needs at purchasing semiconductor products. The system enables the autonomous generation of the available customized sampling methods and provides the performance information for those methods. CSDSS using a clustering data mining method within the generation of sampling methods. The system is not designed to support other domains; it only supports the domain mentioned.

D. Student Retention and Attrition

Some of educational data mining focuses on how data mining is used for improving student success and processes directly related to student learning. Here are some applications that covered in some research

Luan [34] applied data mining as a way to predict what types of students would drop out of school, and then return to school later on. He 1applied classification and regression trees (C&RT) – a specific data mining technique – to educational data in order to predict which students are unlikely to return to school. In this case, study, Luan applied both quantitative and qualitative research techniques to uncover student success factors. This research is important because it demonstrated the successful application of data mining tools to assist in student retention efforts. As noted earlier, the case study method for EDM may often produce results that are not generalizable. However, the process by which researchers apply the data mining can be generalized and used in other contexts. It is simply the results of the data mining models that may not be generalized. [34]

In a related study, Lin (2012) applied data mining as a way to improve student retention efforts. Lin (2012) was able to generate predictive models based on incoming students' data. The models were able to provide short -term accuracy for predicting which types of students would benefit from student retention programs on campus. He found that certain machine learning algorithms could provide useful predictions of student retention Lin (2012).

Researchers at Bowie State University developed a system based on data mining that supports and improves retention [35]. Their system helps the institution identify and respond to at - risk students. Their research contributes meaningfully to the EDM literature because it demonstrates a successful implementation and use of data mining. Their work is highly representative of the discipline in that it follows a strict data mining process and is quantitative. Chacon et al.'s (2012) research supports other work done in applying data mining to student retention issues, such as Lin (2012) and Luan [34], all with successful results. The work by Chacon et al. goes one-step further than Lin and Luan, because the researchers were able to develop and implement their solution in a production environment. [34]

Yeats, Reddy, Wheeler, Senior, & Murray, [36], mentioned that Data mining was used to assess the efficacy of a writing centre in an effort to analyze student achievement and student progress to the next grade. Their work demonstrated the ability to assess a specific educational support process, i.e., the writing centre, in an effort to improve institutional effectiveness. Their research approach used a combination of quantitative work and case study analysis. The mixed- methods approach to data mining was helpful in understanding much more about the ways data mining can be used in an actual implementation. Their research results were not surprising in that it found students who attend writing centres tend to do better in their classes. The drawback of this study did not make the link to student retention issues.

In another study, three different data mining techniques were used to determine predictors of student retention. Yu, DiGangi, Jannesch-Pennell and Kaprolet [37] applied classification trees, multivariate adaptive regression splines (MARS), and neural networks to educational data, which resulted in finding transferred hours, residency, and ethnicity as critical elements in retention efforts.

The data mining clustering technique was used to place students into four main groups based on their preferences and computer experience [32]. Data mining was used in this study as a way to analyze users' preferences in interactive multimedia learning systems. Although the researchers used student preferences as a variable and determined that computer experience as a factor that influences preferences, it is unknown what other types of factors might influence preferences in an online learning environment

Data mining was used in another study to provide learners with many recommendations to help them learn more effectively and efficiently. A methodology called frequent itemset mining was used to mine learner behavior patterns in an online course and subsequently, provide learners with different levels of recommendations rather than single ones that are produced from other recommender systems [21]. This system assisted learners by providing them with highly individualized recommend decisions for improved learning efficiency.

Su, Tseng, Lin, and Chen [38] present a newer stream of research focuses on mobile learning environments. the study applied data mining to help provide fast, dynamic, personalized learning content to mobile users Mobile devices have very different requirements for managing content than standard PCs and web browsers (Su, et all, 2011). They use data such as network conditions, hardware capabilities, and the user's preferences from their device. While this particular study is extremely technical, it demonstrates how mobile learning environments can benefit from data mining.

There are many applications or tasks in educational environments that have been resolved through DM. For example, Baker [4] suggests four key areas of application for EDM: improving student models, improving domain models, studying the pedagogical support provided by learning software, scientific research into learning and learners; and five approaches/methods: prediction, clustering, relationship mining, distillation of data for human judgment and discovery with models. Castro [14] suggests the following EDM subjects/tasks: applications dealing with the assessment of the student's learning performance, applications that provide course adaptation and learning recommendations based on the student's learning behaviour, approaches dealing with the evaluation of learning material and educational web based courses, applications that involve feedback to both teacher and students in e-learning courses, and developments for detection of atypical students' learning behaviours.

Table 1: Comparative studies of techniques and applications used in EDM

Task	Author	EDM tech- app.	Algorithm and DM tech.	Result
Course management systems	**García, Romero, Ventura, & de Castro [14]**	Authors established systematic method for applying data mining techniques to Moodle usage data. The advantage of mining usage data is that it contains data about every user activity, such as testing, quizzes, reading, and discussion posts.	Association rule, classification, and clustering	This research results demonstrated how straightforward, it is to mine data, even if a reader does not have much experience in this area. The authors also use both Keel and Weka as their data mining software packages.
	Rabbany et al., [13]	The researcher suggests an application for data mining, using it to study online courses.	Generating their won toolbox	Provide a visualization toolbox for discovering relevant structures in social networks.
	Dringus & Ellis [15]	They used data mining in representations of the data underlying asynchronous discussion forums.	Text mining	They do not provide specific tools allowing educators to thoroughly track and assess all learners' activities while evaluating the structure and contents of the course
	Castro, Vellido,Nebot, & Mugica [16]	Authors mentioned that, some of the most useful data mining tasks and methods are statistics, visualization, clustering, classification and association rule mining. These methods uncover new, interesting and useful knowledge based on students' usage data. Some of the mains e-learning problems or subjects to which data mining techniques have been applied.	Fuzzy Inductive Reasoning methodology	Authors are dealing with treating the issue of the assessment of a student's learning performance, provide course adaptation and learning recommendations based on the students' learning behaviour, dealing with the evaluation of learning material and educational web-based courses, provide feedback to both teachers and students.
	Zorrilla et al., 2005 [17]	The researchers found that CMS administrator is more likely to apply data mining techniques in order to produce reports for instructors who then use these reports to make decisions about how to improve the student's learning and the online courses. This knowledge, however, can be useful not only to the providers (educators) but also to the users themselves (students).	Classification	Instructors can evaluate the structure of course content and its effectiveness in the learning process and classify learners into groups based on their needs for guidance and monitoring Learners' regular and irregular patterns
	Y.-h. Wang & Liao [18]	Data mining was used to adapt conform learning exercises based on students' Progress through a course on English language instruction Instead of having static course content, the course adapts to student learning, taking him or her through the course at his or her own pace.	Artificial neural network (ANN)	This was an effort to create significant and optimal learning experiences for each student, and was a success. This research could be applied to other types of courses.
Student behaviours	**Beck and Woolf [19]**	cited in their article , how educational data mining prediction methods can be used to develop student models. They use a variety of variables to predict whether a student will make a correct answer.	Machine learning	This work has inspired a great deal of later educational data mining work with ability to predict future correctness.
	Blikstein [20]	Author provide an automated technique to assess, analyze and visualize students learning computer programming. I logged hundreds of snapshots of students' code during a programming assignment, and I employ different.	Quantitative techniques (clustering)	The author found results that showed different types of student programming behaviours in an online course.
	Huang, Chen, & Cheng [21]	The authors present methodology called frequent itemset mining was used to mine learner behaviour patterns in an online course and subsequently, provide learners with different levels of recommendations rather than single ones that are produced from other recommender systems	Mining frequent itemsets	This system assisted learners by providing them with highly individualized recommendations for improved learning efficiency.
Decision support system	**Calders & Pechenizkiy [22]**	The researchers mentioned Other areas within EDM include analysis of educational processes, including admissions, alumni relations, and course selections.	Classification, association and rule mining.	EDM information system
	Guan, Nunez, & Welsh [23]	The authors discussed how important it is to have meaningful information available for decision- makers within higher educational institutions. It is a defines to get the information that decision makers need quickly and efficiently.	Data warehouse	They find that some of the primary drivers of initiating data warehouse projects include increased competitive landscape, and increased responsibilities of reporting to external stakeholders such as parents, board members, legislators and community leaders
	Chau and Phung [24]	According to them, educational decision making support is significant for students, educators, and educational organizations and the support will be more valuable if a lot of relevant data and knowledge mined from data are available for educational managers in their decision making process.	knowledge-driven decision support	They proposed a knowledge-driven educational decision support system. The knowledge-driven decision support is helpful for educational managers to make more appropriate and reasonable decisions about the student's study and further give support to students for their graduation.
	Deniz and Ersan [25]	Researchers presented different ways in which student performance data can be analyzed and presented for academic decision making. They demonstrated the usefulness of an academic decision-support system (ADSS) in evaluating huge amounts of student-course related data. In addition, they presented the basic concepts used in the analysis and design of a specific DSS software package, which is called the Performance, based Academic Decision Support System (PADSS).	Performance based Academic Decision Support System	The Researchers demonstrated the usefulness of an academic decision-support system (ADSS) in evaluating huge amounts of student-course related data. Then they present specific DSS software package which is called the Performance based Academic Decision Support System (PADSS
	Feghali, Zbib and	The authors conducted study by attempts to solve a	Technology-	Their study showed that 79% of users stated

	Hallal [26]	technology-based "last mile" problem by developing and evaluating a web-based decision support tool – the Online Advisor, that helps advisors and students make better use of an already present university information.	based "last mile" problem	that they were satisfied with the Online Advisor, 90% rated the Online Advisor as effective and efficient and more than 75% rated the Online Advisor as useful and helpful.
	Kotsiantis [27]	As the author, the use of machine learning techniques for educational purposes or educational data mining is an emerging field aimed at developing methods of exploring data from computational educational settings and discovering meaningful patterns. The stored data can be useful for machine learning algorithms. Students' key demographic characteristics and their marks in a small number of written assignments can constitute the training set for a regression method in order to predict the student's performance	Machine learning	The author constructed A prototype version of software support tool for tutors.
	Nagy, Aly and Hegazy [28]	The researchers proposed a "Student Advisory Framework" that utilizes classification and clustering to build an intelligent system. The system can be used to provide pieces of consultations to a first year university student to pursue certain education track where he/she will likely to succeed in.	Classification and clustering	Results have proven the efficiency of the suggested framework.
	Vinnik and Scholl [29]	Authors proposed a methodology for assessing educational capacity and planning its distribution and utilization in universities. They integrated educational data mining and knowledge discovery in their proposed method. The DSS supports the administrative task of planning the university's educational capacity in terms of the number of students its courses can accommodate under the specified constraints.	Integrated educational data mining and knowledge discovery	As a result of this study, Decision-makers were able to evaluate various strategies and generate forecasts by means of simulating with the input data. According to them, when the policy makers applied the system, it has resulted in significant acceleration in planning procedures, raised the overall awareness.
	Bose and Sugumaran [30]	Introduced the Intelligent Data Miner (IDM) decision support system. IDM is a Web-based application system intended to provide organization-wide decision support capability for business users. Besides data, mining it also supports some other function categories to enable decision support: data inquiry and multidimensional analysis through enabling OLAP views of multidimensional data. In the data mining part of IDM it supports the creation of models, manipulation of models and presentation of models in various presentation techniques of, among others, the following data mining methods: association rules, clustering and classifiers (classification). The system also performs data cleansing and data preparation and provides necessary parameters data mining algorithms.	Association rules, clustering and classifiers	.The authors state that the disadvantage of IDM is the fact that nontechnical users (business users) need to have a fair amount of understanding of data mining and that the use of data mining and the creation of data mining models still needs to be clearly directed by the user, especially with ad-hoc model creation.
	Lee and Park [31]	Presented the Customized Sampling Decision Support System (CSDSS), which uses data mining. CSDSS is a web-based system that enables the user to select a process sampling method that is most suitable according to his needs at purchasing semiconductor products. The system enables the autonomous generation of the available customized sampling methods and provides the performance information for those methods.	Clustering	Internet-based prototype of CSDSS which had an architecture based on intelligent agent technology and also the successful integration of data mining process for the generation of optimal sampling method into DSS framework
Student Retention and Attrition	**Luan [34]**	Author applied data mining as a way to predict what types of students would drop out of school, and then return to school later on. He 1applied classification and regression trees (C&RT) – a specific data mining technique – to educational data in order to predict which students are unlikely to return to school. In this case study, Luan applied both quantitative and qualitative research techniques to uncover student success factors.	Classification and regression trees	This research is important because it demonstrated the successful application of data mining tools to assist in student retention efforts, the case study method for EDM may often produce results that are not generalizable. However, the process by which researchers apply the data mining can be generalized and used in other contexts.
	Chacon, Spicer, & Valbuena [35]	Researchers at Bowie State University developed a system based on data mining that supports and improves retention. Their system helps the institution identify and respond to at - risk students. Their research contributes meaningfully to the EDM literature because it demonstrates a successful implementation and use of data mining.	Data-Driven	Their work is highly representative of the discipline in that it follows a strict data mining process and is quantitative.
	Yeats, Reddy, Wheeler, Senior, & Murray, [36]	Mentioned that Data mining was used to assess the efficacy of a writing centre in an effort to analyze student achievement and student progress to the next grade. Their work demonstrated the ability to assess a specific educational support process, i.e., the writing centre, in an effort to improve institutional effectiveness. Their research approach used a combination of quantitative work and case study analysis..	Most of DM algorithms	Their research results were not surprising in that it found students who attend writing centers tend to do better in their classes. The drawback of this study did not make the link to student retention issues.

	Yu, DiGangi, Jannesch-Pennell and Kaprolet [37]	The research team use three different data mining techniques to determine predictors of student retention, they applied classification trees, multivariate adaptive regression splines (MARS), and neural networks to educational data .	Classification trees, multivariate adaptive regression	The use of these techniques resulted in finding transferred hours, residency, and ethnicity as critical elements in retention efforts.
	Chrysostomou, Chen, & Liu [32]	The atauthors used data mining clustering technique to place students into four main groups based on their preferences and computer experience. Data mining was used in this study as a way to analyze users' preferences in interactive multimedia learning systems. Although the researchers used student preferences as a variable and determined that computer experience as a factor that influences preferences,	Clustering	The drawback of this study is unknown what other types of factors might influence preferences in an online learning environment
	Huang, Chen, & Cheng [21]	In this study Data mining was used to provide learners with many recommendations to help them learn more effectively and efficiently. A methodology called frequent itemset mining was used to mine learner behaviour patterns in an online course and subsequently, provide learners with different levels of recommendations rather than single ones that are produced from other recommender systems.	Mining frequent itemsets	This system assisted learners by providing them with highly individualized recommendations for improved learning efficiency.
	Su,Tseng, Lin, and Chen [38]	Authors Present A newer stream of research focuses on mobile learning environments. the study applied data mining to help provide fast, dynamic, personalized learning content to mobile users. They use data such as network conditions, hardware capabilities, and the user's preferences from their device.	Clustering and decision tree	This particular study is extremely technical; it demonstrates how mobile learning environments can benefit from data mining.

IV. COMPARATIVE STUDIES OF TECHNIQUES AND APPLICATIONS OF EDM IN HIGHER EDUCATION

This section provides a comparative studies between some of researches and practical works that mentioned in the previous section. As it shown on table (1) four educational areas has been covered, include Course management systems, Student behaviours, Decision support system, and Student Retention and Attrition. Under each area many research work have been studies and compare.

V. CONCLUTION

This paper surveys the state of art in EDM, it try to provides a boundary definition to the term Education Data Mining, hence we found that EDM covers broader area and many interdisciplinary such as education, e-learning, data mining, data analysis, intelligent system and so on,

The paper covers most relevant work in the area of EDM in course management systems, student behaviours, decision support system, and Student Retention and Attrition. The paper also provide a comparison study between some of research work in such areas.

The paper concludes that EDM discipline is growing fast and many new ideas and technologies can be merge in this discipline, in addition well-established research and application have provide a considerable contribution. Thus, we expect that EDM will become more useful, and fully operative and available even for external users.

Reference:

[1] The Educational Data Mining community website, www.educationaldatamining.org

[2] Barnes, T., Desmarais, M., Romero, C., Ventura, S. (2009). Educational Data Mining 2009: 2nd International Conference on Educational Data Mining, _Proceedings. Cordoba, Spain.

[3] Campbell, J., & Oblinger, D. (2007). Academic analytics. Washington, DC: Educause.

[4] Baker, R., & Yacef, K. (2009). The State of Educational Data mining in 2009: A Review and Future Visions. Journal of Educational Data Mining, 1

[5] Berson, A., Smith, S., & Thearling, K. (2011). An Overview of Data Mining Techniques Retrieved November 28, 2011, from

[6] Arockiam, L., S. Charles, and M. Amala Jayanthi. "An Impact of Emotional Happiness and Personality in Students' Learning Environment." Data Mining and Knowledge Engineering 7.2 (2015): 69-74.

[7] Chalaris, Manolis, et al. "Examining students' graduation issues using data mining techniques-The case of TEI of Athens." INTERNATIONAL CONFERENCE ON INTEGRATED INFORMATION (IC-ININFO 2014): Proceedings of the 4th International Conference on Integrated Information. Vol. 1644. AIP Publishing, 2015.

[8] Foltz, Peter W., and Mark Rosenstein. "Analysis of a Large-Scale Formative Writing Assessment System with Automated Feedback." Proceedings of the Second (2015) ACM Conference on Learning@ Scale. ACM, 2015.

[9] Ocumpaugh, Jaclyn, et al. "Population validity for Educational Data Mining models: A case study in affect detection." British Journal of Educational Technology 45.3 (2014): 487-501.

[10] Patidar, Preeti, Jitendra Dangra, and M. K. Rawar. "Decision Tree C4. 5 algorithm and its enhanced approach for Educational Data Mining." (2015).

[11] Yamamoto, Yukiko, et al. "Increasing the Sensitivity of a Personalized Educational Data Mining Method for Curriculum Composition." Emerging Issues in Smart Learning. Springer Berlin Heidelberg, 2015. 201-208.

[12] Romero, Cristóbal, and Sebastián Ventura. "Educational data mining: a review of the state of the art." *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on* 40.6 (2010): 601-618.

[13] Rabbany, Reihaneh, Mansoureh Takaffoli, and Osmar R. Zaïane. "Analyzing participation of students in online courses using social network analysis techniques." *Proceedings of educational data mining*. 2011.

[14] Romero, Cristóbal, Sebastián Ventura, and Enrique García. "Data mining in course management systems: Moodle case study and tutorial." *Computers & Education* 51.1 (2008): 368-384.

[15] Dringus, Laurie P., and Timothy Ellis. "Using data mining as a strategy for assessing asynchronous discussion forums." *Computers & Education* 45.1 (2005): 141-160.

[16] Castro, Félix, A. Nebot, and Francisco Mugica. "EXTRACTION OF LOGICAL RULES TO DESCRIBE STUDENTS'LEARNING BEHAVIOR."*Proceedings of the sixth conference on IASTED International Conference Web-Based Education*. Vol. 2. 2007.

[17] Zorrilla, Marta E., et al. "Web usage mining project for improving web-based learning sites." *Computer Aided Systems Theory–EUROCAST 2005*. Springer Berlin Heidelberg, 2005. 205-210.

[18] Wang, Ya-Huei, and Hung-Chang Liao. "Data mining for adaptive learning in a TESL-based e-learning system." *Expert Systems with Applications* 38.6 (2011): 6480-6485.

[19] Beck, J.E., Woolf, B.P.: High-Level Student Modeling with Machine Learning. In: Gauthier, G., et al. (eds.): Intelligent Tutoring Systems, ITS 2000. Lecture Notes in Computer Science, Vol. 1839. Springer, Berlin Heidelberg New York (2000) 584-593

[20] Blikstein, P. (2011). Using learning analytics to assess students' behavior in open – ended programming tasks. Paper presented at the Proceedings of the 1st International Conference on Learning Analytics and Knowledge, Banff, Alberta, Canada.

[21] Huang, Jen-Peng, Show-Ju Chen, and Huang-Cheng Kuo. "An efficient incremental mining algorithm-QSD." *Intelligent Data Analysis* 11.3 (2007): 265-278.

[22] Calders, Toon, and Mykola Pechenizkiy. "Introduction to the special section on educational data mining." *ACM SIGKDD Explorations Newsletter* 13.2 (2012): 3-6.

[23] Guan, J., Nunez, W., & Welsh, J. (2002). Institutional strategy and information support: the role of data warehousing in higher education. Campus-Wide Information Systems,19(5)-164- 174.6. Chacon, F., Spicer, D., & Valbuena, A. (2012). Analytics in Support of Student Retention and Success (Research Bulletin 3, 2012ed.). Louisville, CO: Educause Center for Applied Research.

[24] Chau, Vo Thi Ngoc, and Nguyen Hua Phung. "A knowledge-driven educational decision support system." *Computing and Communication Technologies, Research, Innovation, and Vision for the Future (RIVF), 2012 IEEE RIVF International Conference on*. IEEE, 2012.

[25] Deniz, Dervis Z., and Ibrahim Ersan. "An Academic Decision Support System Based on Academic Performance Evaluation for Student and Program Assessment." *International Journal of Engineering Education* 18.2 (2002): 236-244.

[26] Feghali, Tony, Imad Zbib, and Sophia Hallal. "A web-based decision support tool for academic advising." *Journal of Educational Technology & Society*14.1 (2011): 82-94.

[27] Kotsiantis, Sotiris B. "Use of machine learning techniques for educational proposes: a decision support system for forecasting students' grades."*Artificial Intelligence Review* 37.4 (2012): 331-344.

[28] Nagy, Heba Mohammed, Walid Mohamed Aly, and Osama Fathy Hegazy. "An Educational Data Mining System for Advising Higher Education Students." *World Acad. Sci. Eng. Technol. Int. J. Inf. Sci. Eng* 7.10 (2013): 175-179.

[29] Vinnik, Svetlana, and Marc H. Scholl. *UNICAP: Efficient decision support for academic resource and capacity management*. Springer Berlin Heidelberg, 2005.

[30] Bose, Ranjit, and Vijayan Sugumaran. "Application of intelligent agent technology for managerial data analysis and mining." *ACM SIGMIS Database* 30.1 (1999): 77-94.

[31] Lee, Jang Hee, and Sang Chan Park. "Agent and data mining based decision support system and its adaptation to a new customer-centric electronic commerce." *Expert Systems with Applications* 25.4 (2003): 619-635.

[32] Chrysostomou, K., Chen, S. Y., & Liu, X. (2009). Investigation of Users' Preferences in Interactive Multimedia Learning Systems: A Data Mining Approach. Interactive Learning Environments, 17(2), 151-163.

[33] Lee, Jang Hee, and Sang Chan Park. "Agent and data mining based decision support system and its adaptation to a new customer-centric electronic commerce." *Expert Systems with Applications* 25.4 (2003): 619-635.

[34] Luan, J. (2002). Data Mining and Knowledge Management in Higher Education-potential Applications. Paper presented at the Annual Forum for the Association for Institutional Research, Toronto, Ontario, Canada.9. Lin, S.-H. (2012). Data mining for student retention management. J. Comput. Sci. Coll., 27 (4) 92-99.

[35] Chacon, Fabio, Donald Spicer, and A. Valbuena. "Analytics in support of student retention and success." *Research Bulletin* 3 (2012): 1-9.

[36] Yeats, Rowena, et al. "What a difference a writing centre makes: a small scale study." *Education+ Training* 52.6/7 (2010): 499-507.

[37] Ohri, Zinnia. "A Critical Analysis of Various Data Mining Techniques in Educational Assessment and Feedback." *IITM Journal of Information Technology*: 25.

[38] Su, Jun-Ming, et al. "A personalized learning content adaptation mechanism to meet diverse user needs in mobile learning environments." *User modeling and user-adapted interaction* 21.1-2 (2011): 5-49.

[39] Tsantis, L., Castellani, J. (2001). Enhancing learning environments through solution-based knowledge discovery tools. In Journal of Special Education Technology, 16,4, 39-52

[40] Pahl, C., Donnellan, D.: Data Mining Technology for the Evaluation of Webbased Teaching and Learning Systems. In: World Conference on e-Learning in Corp., Govt., Health., & Higher Education. (2002) 747-752

[41] Chu, K., Chang, M., Hsia, Y.: Designing a Course Recommendation System on Web based on the Students' Course Selection Records. In: World Conference on Educational Multimedia, Hypermedia and Telecommunications (2003) 14-21

[42] Chang, K., Beck, J., Mostow, J., Corbett, A.: A Bayes Net Toolkit for Student Modeling in Intelligent Tutoring Systems. In: Ikeda, M., et al. (eds.): 8th International Conference on Intelligent Tutoring Systems, ITS2006. LNCS Vol. 4053. Springer, Berlin Heidelberg New York (2006) 104-113