IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 9, Issue 9, September 2020

DOI 10.17148/IJARCCE.2020.9903

Nuts and Bolts of Educational Data Mining

Dr. Abhay Kumar Srivastava[1], Dr. Jaya Srivastava[2]

Associate Professor, Institute of Public Enterprise Hyderabad, India[1]

Lecturer, Mahila Degree College Lucknow[2]

Abstract: An overview of Educational Data mining is discussed in this paper**.** It introduces the concepts and importance of Data Mining, which has led to evolve educational data mining and its applications. It also illustrates nature of data used in Educational Data Mining, nature of Patterns for Educational Data Mining based on the kinds of knowledge to be mined, the kinds of technologies used, and popular tools used in various kinds of applications that are targeted. Finally, major challenges in the field of educational data mining are also highlighted. While taking a general overview of educational data Mining, a brief focus on technologies used in educational data mining are also covered.

Keywords: Educational Data Mining (EDM), Statistics, Data Mining, Prediction, Technologies in EDM, Challenges in EDM, Applications.

I. INTRODUCTION

Millions of gigabytes of data are generated everyday due to increasing computer application in almost all areas of society. Though we claim that current era is an information era but truly speaking this is a data era where a lot of data is created every day in different formats and left unattended and un-extracted. It is just like oil wells which were hidden in the earth for thousands of years before it could be explored and made available for commercial use. In such scenario of abundance data availability, there is a growing need to use data information for the benefit of society. Hence the time demands to look it further for usage of masses.

The explosive growth of available data volume is due to rapid penetration of computers in our society and the fast development of powerful data collection and storage tools. Not only Businesses but education sector is globally generating gigantic data sets, including discussion forums, social media, assignment and project descriptions, study materials, lecture videos and students' feedback.

Educational institutes have started using various techniques to analyze this data in the form of educational report stored with it such as enrollment data, students' performance, teachers' evaluations, gender differences, and many others. "Data mining techniques" provide an institute the needed information to better plan a number of students' enrollment, students drop out, early identification of weak students, and to efficiently allocate resources with a precise approximation. EDM is little different from Data mining. Where Data Mining deals with huge datasets, it may not be true with EDM where researchers often have to deal with small data sets. General unsupervised or semi-supervised learning has a more direct influence on EDM. However, EDM share some useful features with data mining in general.

Though conventional techniques of analyzing data are available like Statistics, which is commonly used for summarizing, visualizing and analyzing data but it has some limitations. The type of data used in statistics is mostly metric in nature with lot of complexities in designing an experiment. Thus, it creates some limitations on relying totally on statistical techniques for analysis data. Also, as the volume of data becomes very large, analysis with complex and tedious. Thus, Data mining can be termed as an extension of traditional data analysis and statistical approaches which incorporates various analytical techniques and also not limited to,

- Numerical analysis,
- Pattern matching and areas of artificial intelligence such as machine learning,
- Neural networks and genetic algorithms.

Data mining is a non-trivial process of analyzing large volume data. Mining is not only done in data, but text mining and image mining are also becoming very popular. Data mining find huge applications in almost all spheres of life including Business, Medical, Social Sciences, and Behavioral Sciences etc. There are two types of approaches in data mining. It differs in whether they seek to build models or to find patterns. The first approach is based on model building which is similar to conventional exploratory statistical methods. The objective is to produce an overall summary of a set of data to identify and describe the main features of the shape of the distribution [Hand 1998]. Some model building examples used in data mining are Cluster analysis for partition of a set of data, a regression for prediction, and a tree-based classification rule.

The second type of data mining approach, pattern detection, seeks to identify small abnormalities from the normal patterns of behavior. Generally business database often creates a problem for pattern extraction because of their complexity. Complexity arises from anomalies such as discontinuity, noise, ambiguity, and incompleteness (Fayyad, Piatetsky-Shapiro, and Smyth, 1996).

Recently Data mining finds a lot of application in education sector where it is helpful in predicting the performance of students on the basis of several attributes. These attributes comprises the background of students, its past academic performance, region, gender, experience etc. All these attributes can be used by educational institutes to lay great emphasis on the weak areas of the students so that they can meet or even surpass their professional goals. Proper classification of students can enable educational institutes to focus on their weak areas so that they can excel in their professional goals.

EDM is defined 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 they learn in” (Baker, 2009) defined by international consortium on educational data mining . EDM focuses on analyzing data generated in an educational setup by the various stakeholders in the system to develop model for improving teaching-learning experience and institutional effectiveness.

II. OBJECTIVE OF THE STUDY

It has been dedicated to understand Data Mining and it’s the role of data mining especially in education sector. Today’s world is often called as a knowledge world. Not only government agencies but ordinary man is also investing heavily in education. In the present era education level for any nation is considered to be the wealth of that nation as compared to the natural resources which was earlier considered as nation’s wealth.

Educational Data Mining (EDM) uses sophisticated data mining techniques for solving problems in education. EDM is a powerful tool to optimize student learning process.

There are three broad objectives of this chapter:

1. To explore tools and techniques used in data mining and EDM. We will also focus on those techniques that are suitable for segmenting students and predicting their performances.
2. To discuss applications of educational data mining
3. To discuss various issues and challenges in data mining and Educational data mining in context of selection and employability of students.

III. RESEARCH METHODOLOGY

The research focuses on extensive literature review from research paper, books, and articles taken from renowned publishers, journals and conferences proceedings in the area of not only Data Mining and Educational Data Mining but also from Statistics. The use of primary data or experimental research is excluded due to the nature of scope of study. Thus mainly exploratory research based on information available is used to comprehend the entire paper. Since data mining is used as an integrative approach in Educational data mining so it begins with a brief introduction of Data Mining. 55 research papers published of last two decades from national and international journals are reviewed. These papers are related to data mining, applications, educational data mining and its variety of applications. The citations of most of these articles referred are very high.

Data Mining Definition

Data mining is an interdisciplinary subject which is a combination of Machine Learning, Statistics and Business Intelligence, can be defined in many different ways. Data mining can also be termed as “knowledge mining from data,”, however, the shorter term; *knowledge mining* may not reflect the emphasis on mining from large amounts of data.

Data mining is often treated as a synonym for another popularly used term, knowledge discovery from data, or KDD, (Fayyad 1996). Data mining is also viewed as an essential step in the process of knowledge discovery.

The knowledge discovery process is an iterative sequence of the following steps:

1. Data cleaning: It is used for removing noise, anomaly and inconsistent data from a large data set.
2. Data integration: The process when multiple data sources, of same format or different format, are combined and brought to same plain
3. Data selection: Relevant data for the analysis task are retrieved from the database.
4. Data transformation: data are transformed and consolidated into forms appropriate for mining by performing summary or aggregation operations.
5. Data mining: It is an essential process where different tools & techniques are applied to extract useful patterns in data.
6. Pattern evaluation: Used to identify some really interesting and meaningful patterns in the databases.
7. Knowledge presentation: visualization and knowledge representation techniques like pictures, graphs and tables are used to present extracted knowledge from data to users.

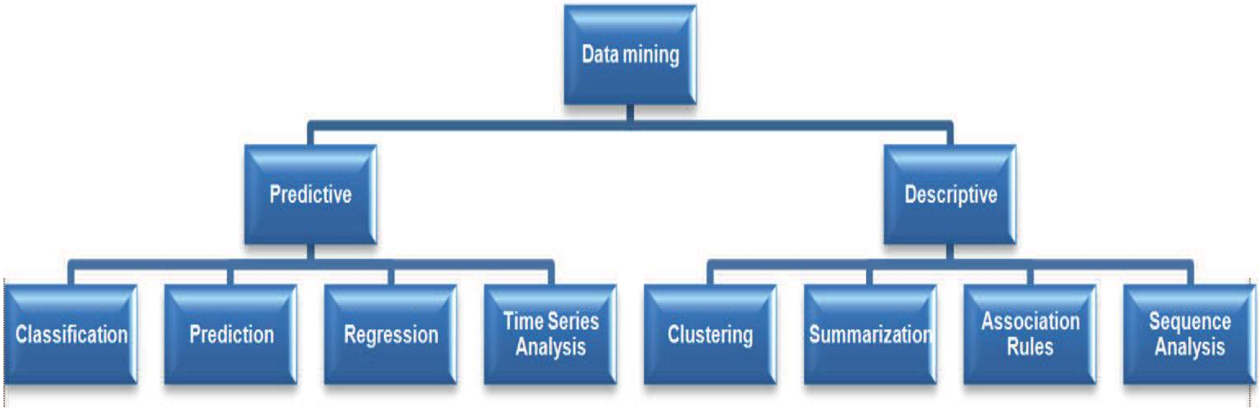

Figure 1: Data Mining Applications (Source: Siraj 2011)

Educational Data Mining

Educational Data Mining (EDM) is upcoming field in Knowledge discovery. Due to widespread growth of higher education, predictions related to student's performance can be accurately done through EDM. Not only predictions, classification, associations and grouping can also be done with perfection using statistical and software tools. The Education system can be equipped with more information relating to future drop out of students and their success in enrolled courses. Both students and stake holders could be benefitted by EDM. Nowadays interactive e-learning methods and tools like Moodle, have opened an opportunity to collect and scrutinize student data through online quizzes and online board discussions.

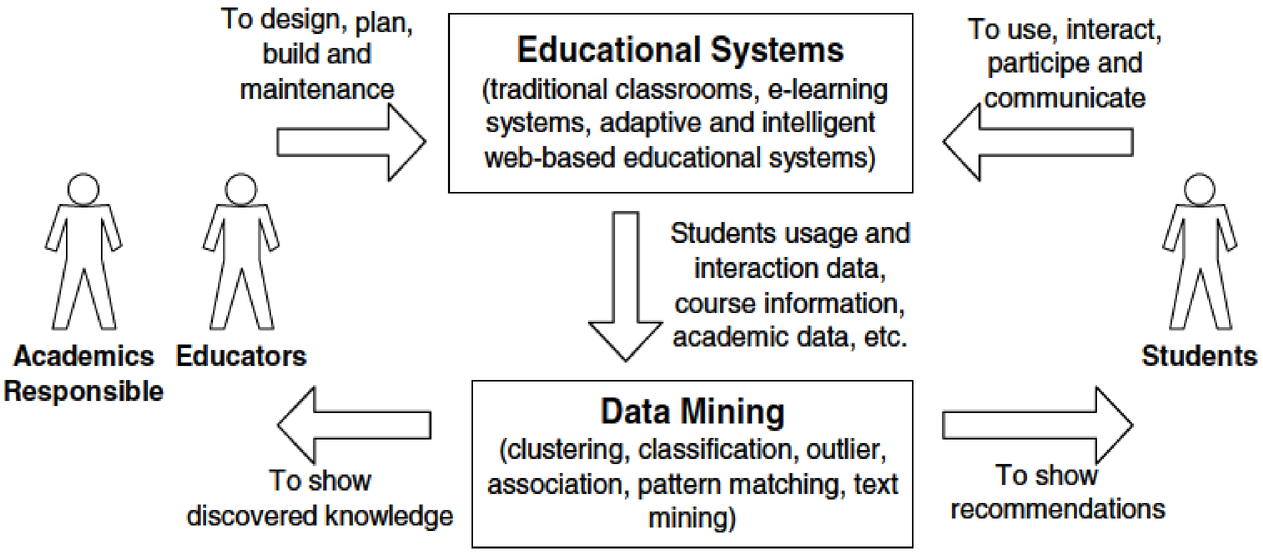
To design, plan, build and maintenance
Educational Systems
(traditional classrooms, e-learning systems, adaptive and intelligent web-based educational systems)
To use, interact, participe and communicate
Academics Responsible
Educators
Students usage and interaction data, course information, academic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

Figure 2 : Relationship between Data Mining and Educational Data Mining (Source: C. Romero, S. Ventura,(2007)

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 9, Issue 9, September 2020

DOI 10.17148/IJARCCE.2020.9903

In the educational field, data mining techniques can generate useful patterns that can be used both by educators and learners. EDM provide recommendations to learners to improve their learning and to create individual learning environments and assist educators to improve their learning environment. Not only this, educational data mining can be used to enhance employability index of students. The interconnection between Data Mining and Education Systems is very well explained by the figure1.

Educational data mining is emerging as a research area for understanding how students learn. New computer-supported interactive learning methods and tools—intelligent tutoring systems, simulations and games have opened up opportunities to collect and analyze student data, to discover patterns and trends in those data, and to make new discoveries and test hypotheses about how students learn (Baker, 2007). Data collected from online learning systems can be aggregated over large numbers of students and can contain many variables that data mining algorithms can explore for model building.

Initially in Education data mining, researchers used to capture website log data to understand behavior of learners specially students (Amershi, 2009). But now more integrated, instrumented, and sophisticated online learning systems provide different kinds of data. Educational data mining generally emphasizes reducing learning into small components that can be analyzed and then influenced by software that adapts to the student (Baker, 2010).

Student learning data collected by online learning systems are being explored to develop predictive models by applying educational data mining methods that classify data or find relationships. These models play a key role in building adaptive learning systems in which adaptations or interventions based on the model's predictions can be used to change what students experience in future (Amershi et al, 2006).

Educational data is hierarchical in nature. Data collected at various levels like session level, student level, teacher level and institute level are nested into one another. Time is important to capture data, such as length of practice sessions or time to learn is also key feature in EDM.

Similarly sequence and context also plays an important role in Educational data. Sequence represents how concepts build on one another and how practice and tutoring should be ordered. Context is important for explaining results and knowing where a model may or may not work. Thus hierarchical data mining methods and longitudinal data modeling are important developments in mining educational data (Merceron et al, 2008).

Educational data mining (Baker, 2009, 2010) view the following as the goals for its research:

1. Predicting students' future learning behavior by creating student models that incorporate such detailed information as students' knowledge, motivation, meta-cognition, and attitudes (Bharadwaj & Pal, 2011).
2. Discovering or improving domain models that characterize the content to be learned and optimal instructional sequences.
3. Studying the effects of different kinds of pedagogical support that can be provided.
4. Advancing scientific knowledge about learning and learners through building computational models that incorporate models of the student and the domain.

List of top 5 cited papers in Educational Data Mining

Table 1: List of top 5 cited papers

Article/book	**Author name and Publication year**	**Area of Contribution**	**No. of Citations**
The state of Educational Data Mining in 2009: A review and future vision	Baker, R.S.J.D. ,and Yacef, K, 2009	Comprehensive review of papers that cover the important aspects of data mining in educational research	1590

Knowledge Discovery in Databases: An overview	Frawley, W.J., Piatetsky-Shapiro, G., & Matheus, C.J., 1992	Discovery in databases including inductive learning, bayesian statistics, semantic query optimization, knowledge acquisition for expert systems, information theory, and fuzzy sets	2853
From Data Mining to Knowledge Discovery in Databases	Fayyad, U. Piatetsky-shapiro G., Smyth. P., 1996.	An overview of DM field, clarifying how data mining and knowledge discovery in databases are related both to each other and to related fields, such as machine learning, statistics, and databases	2869
Educational Data Mining: A review of the state of the Art	Romero, C., and Ventura, S., 2010	Reviewed 306 papers on the application of educational data mining and mined e-learning data	1977
Educational Data Mining: A survey from 1995 to 2005	Romero, C., and Ventura, S., 2007	Comprehensive Review on the application of data mining techniques in educational system from the year 1995 until 2005	1877

Nature of Data used in Educational Data Mining

The most basic forms of data for educational mining applications are data from student databases and academic records, data warehouses, and student transactions. The best results are obtained when data is non-trivial in nature. Educational Data mining can also be applied to other forms of data like multimedia data, text data, pictorial/graphical data, moodle data, spatial data and network data. EDM will certainly continue to embrace new data types as it emerges. It is important that in many applications, multiple types of data are present. Mining multiple data sources of complex data often leads to fruitful findings due to the mutual enhancement and consolidation of such multiple sources. But it is also possess some challenges because of the difficulties in data cleaning and data integration, as well as the complex interactions among the multiple sources of such data. The educational data could be generated from the offline as well as online sources

***Offline Sources of data*.**

Offline Data are generated from traditional classroom interactive environments, learner / educator's information, student's attendance and evaluation record, subject related information [18] etc.

***Online Sources of data*.**

Online Data are generated from MOOCs, distance educations and web based education contents, data generated through social networking sites and online discussion forum like Text data, publication databases, E-mail, Spreadsheets, and Telephonic Conversations etc.

Popular Tools and Technologies Used in Educational Data Mining

As a highly application-driven domain, data mining has incorporated many techniques from other domains such as statistics, machine learning, pattern recognition, database and data warehouse systems, information retrieval, visualization, algorithms, high performance computing, and many application domains. In this section, Some popular educational data mining tools which provide multiple data mining functions and multiple knowledge discovery techniques are listed below

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 9, Issue 9, September 2020

DOI 10.17148/IJARCCE.2020.9903

Table 2: Popular Tools and Technologies Used in Educational Data Mining

Name of Tool and Developer	**Source (Open/free/ Commercial)**	**Function/Features**	**Techniques/Tools**	**Environments**
WEKA	Open/free	Provides machine learning algorithms for data mining tasks. Well-suited for developing new machine learning schemes.	Data pre-processing, Classification, regression, clustering, association rules, and visualization.	Windows, Linux
ALPHA MINER	Open/free	Provides the best cost and performance ratio for data mining applications	Versatile data mining Functions	Windows, Linux, Mac
SPSS Clementine	Commercial	Provides an integrated data mining development environment for end users and developers.	Association Mining, Clustering, Classification, Prediction and visualization tools	Windows, Solaris, Linux
Enterprise Miner	Commercial	Provides variety of statistical analysis tools	Association Mining, Classification, Regression, Time series analysis, Statistical analysis, Clustering	Windows, Solaris, Linux
Oracle Data Mining	Commercial	Provides an embedded DWH infrastructure for multidimensional data analysis	Association Mining, Classification, Prediction, Regression, Clustering, Sequence similarity search and analysis.	Windows, Mac, Linux
Intelligent Miner	Commercial	Provides tight integration with IBB's DB2 relational db system, Scalability of Mining Algorithm	Association Mining, Classification, Regression, Predictive Modeling, Deviation detection, Clustering, Sequential Pattern Analysis	Windows, Solaris, Linux
MSSQL Server 2005	Commercial	Provides DM functions both in relational db system & Data Warehouse system environment.	Integrates the algorithms developed by third party vendors and application users.	Windows, Linux

CART	Commercial	Provides binary splitting and post pruning for Classification (Decision Tree) and for Prediction (Regression Trees).	Classification-Decision and Regression Tree	Windows, Linux
Random Forests	Commercial	Provides high levels of predictive accuracy and an innovative set of graphical displays to reveal unexpected patterns in data	Clustering	Windows, Linux

Nature of Patterns for Educational Data Mining

As in Data Mining, EDM also deals with different patterns. These include characterization and discrimination; frequent patterns mining, associations and correlations; classification and regression; clustering analysis and outlier analysis. It is used to specify the kinds of patterns to be found in data mining tasks. In general, such tasks can be classified into two categories: descriptive and predictive. Descriptive mining tasks characterize properties of the data (central tendencies and measures of dispersion) in a target data set. Predictive mining processes the current data to make predictions of student's progress and achievements.

Data characterization is a summarization of the general characteristics or features of a target class of data. There are several methods for effective data summarization and characterization. The output of data characterization can be presented in various forms like pie charts, bar charts, histograms, Box-plot, curves and multidimensional tables, including crosstabs and contingency tables.

Researchers have applied mining techniques to ITS (Intelligent Tutoring System) and CMS (Course Management System) data. For example (Romero et al, 2010) applied data mining were techniques to data collected with the Moodle CMS. This system allows students to both view and submit various assignments, and records detailed logs of students' interactions. Association rule mining was used to generate some useful patterns in these interaction logs to identify fringe behaviors exhibited by students. Similarly (Mostow, 2011) applied data mining techniques were applied to interaction logs taken from an ITS. This system teaches young students as they learn to read by listening to them read stories aloud and providing feedback. A system was developed which automatically identified meaningful features from these logs which were then used to train classifiers to predict students' future behavior with the system. In Educational Data Mining researchers often seek to find patterns that best distinguish students who do and do not perform well in the course. Thus, there is a need for novel pattern mining techniques aimed at differentiating between two databases of sequences.

Mining Frequent Patterns, Associations, and Correlations:

The discovery of frequent patterns, associations, and correlation relationships among huge amounts of data is useful in selective marketing, decision analysis, business management and other sectors. A popular area of application is market basket analysis, which studies students' performance behavior by searching for undergraduate background that frequently occurs with the gender of student.

Association rule mining consists of first finding frequent item sets. Association rules are if/then statements that help uncover relationships between seemingly unrelated data in a relational database or other information repository. An example of an association rule would be "If a student is good in Analytics, it is very likely that he also sound in Statistics." In data mining, association rules are useful for analyzing and predicting students' behavior. It plays an important part in basket data analysis, product clustering, catalog design and Library layout.

Many efficient and scalable algorithms have been developed for Market Basket Analysis also called as frequent item set mining, from which association and correlation rules can be derived. The commonly used algorithms for frequent item set mining is (Chen et al, 2012) the *Apriori algorithm.* The Apriori algorithm is an important algorithm for mining frequent item sets for Boolean association rules. It uses bottom-up approach where frequent subsets are extended as on item at a time in transactional databases.

Association rule mining can be used for finding co-occurrence of student problems, associating pedagogies with different student to build recommendations for methodology that is likely to be interesting, or for making changes to teaching

DOI 10.17148/IJARCCE.2020.9903

approaches (Merceron and Yacef, 2010). These techniques can be used to associate student activity, in a learning management system or discussion forums, with student grades or to investigate such questions as why students' use of practice tests decreases over a semester of study.

Sequential pattern mining builds rules that capture the connections between occurrences of sequential events, for example, finding temporal sequences, such as student mistakes followed by help seeking. This could be used to detect events, such as students regressing to making errors in mechanics when they are writing with more complex and critical thinking techniques, and to analyze interactions in online discussion forums.

Classification and Regression for Predictive Analysis*:*

Classification places an object into one class or category, based on its other characteristics. In education, teachers and instructors are all the time classifying their students for their knowledge, motivation, and behavior. Classification is the process of deriving a suitable model (or function) that describes and distinguishes data classes or labels are known. The model is used to predict the class label of objects like student's background, their gender, region etc.

While Classification predicts categorical (discrete, unordered) labels, regression models continuous-valued functions. It implies, linear regression is used to predict missing or unavailable *numerical data values (also termed as dependent variable)* rather than (discrete) class labels. The term *prediction* refers to both numeric prediction and class label prediction. Linear Regression analysis is a statistical methodology that is most often used for numeric prediction, although other methods exist as well (Giudici & Paolo, 2013). Logistic Regression is found to be very useful in prediction of categorical variables. Binary Regression, which is a class under Logistic regression, is used for dichotomous outputs like yes/no. In Educational Mining, prediction has two key uses.

Prediction methods can be used to study model features that are important for prediction, giving information about the underlying construct. This is generally used to predict student educational outcomes (Romero et al, 2008) without predicting intermediate or mediating factors first. The second issue is that Prediction methods are used in order to predict what the output value would be in contexts where it is not desirable to directly obtain a label for that construct. Predictive models have been used for understanding student's behaviors in an online learning environment and also their performance in classroom. Prediction shows promise in developing domain models, such as connecting procedures or facts with the specific sequence and amount of practice items that best teach them, and forecasting and understanding student educational outcomes, such as success on posttests after tutoring (Aksenova et al 2006).

Logistic Models (Binary Regression) are used prediction techniques in higher education research. Though linear regression is commonly used for prediction, it poses certain limitations. Firstly one has to understand the difference between Linear and Binary regression. The primary difference between linear and binary logistic regression is that later is better suited for the dichotomous outcomes, which are categorical in nature such as selection versus no selection of students in a recruitment process (Berry J.A.et al, 2004).

Linear regression predicts the value of dependent variable from one or more independent variables that is based on the model with certain assumptions: normality of the variables, normal distribution and homoscedasticity of residuals. This problem is not faced in case of logistic regression where the variable can be both continuous and dichotomous in nature as it does not require the variables to meet the above mentioned assumptions of normality and homoscedasticity. Binary regression calculates the odds of either of the two outcomes of a categorical variable from one or more independent variables (C. Romero et al 2007). Binary regression uses the logistic response function (Baker et al 2008) which is given by

$$P\ (y = 1) = 1/\ (1 + e - (b_0 + b_1x_1 + \ldots + b_nx_n)$$

It is a non-linear transformation of the linear regression equation to predict a number in the range of 0 to 1(as shown in Figure 4.1) which will equal the probability of the outcome y=1. Using the probability of the outcome y=1, one can calculate the probability of y=0 as

$$P(y=0) = 1 - P(y=1)$$

Classification and regression may need to be preceded by relevance analysis, which attempts to identify attributes that are significantly relevant to the classification and regression process. Such attributes will be selected for the classification and regression process. Other attributes, which are irrelevant, can then be discarded for further consideration.

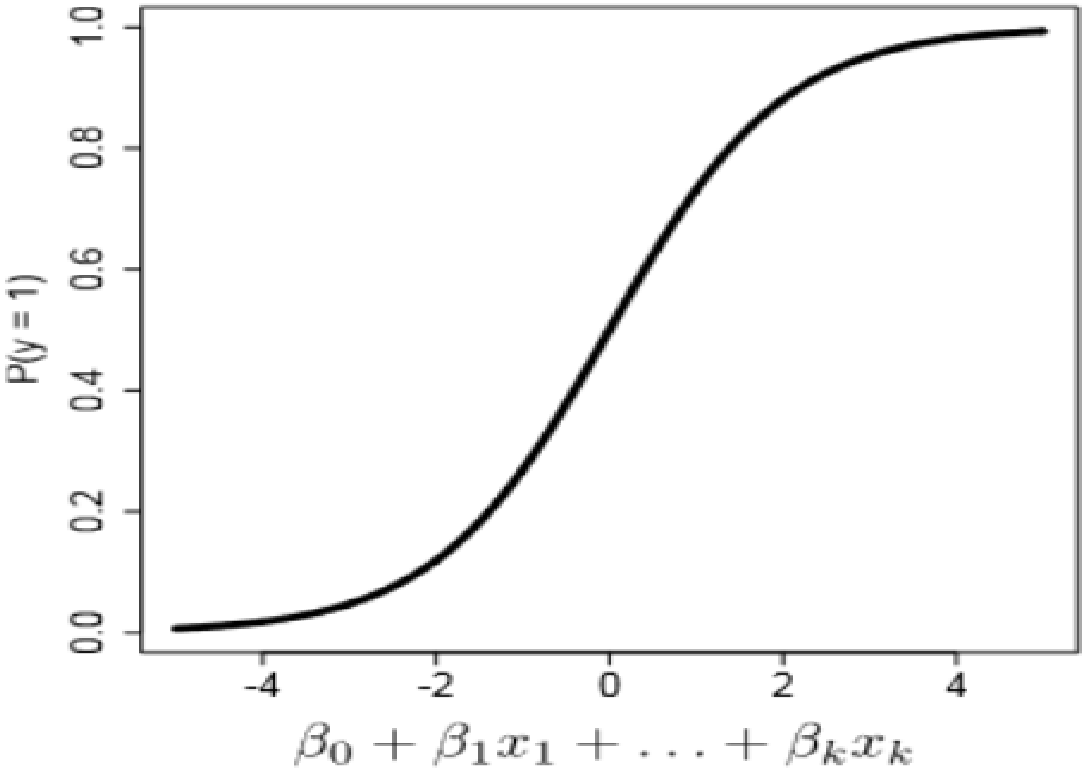

Figure 3: Graph depicting the logistic response function

Cluster Analysis:

It identifies natural grouping of the data points by splitting the entire data into naturally occurring clusters. Clustering is also called as unsupervised learning. It analyzes data objects without consulting class labels as opposed to classification where class labelled data sets are analyzed (Friedman, 2009) Clustering is useful for exploring data. If there are many cases and no obvious groupings, clustering algorithms can be used to find natural groupings. Cluster analysis works on clustering algorithms that can identify clusters automatically. Two most used defined classes of clustering algorithms are Hierarchical and Partitioning clustering algorithms.

Hierarchical clustering procedures are characterized by the tree-like structure established in the course of the analysis. Most hierarchical techniques fall into a category called agglomerative clustering. In this category, clusters are consecutively formed from objects. Initially, this type of procedure starts with each object representing an individual cluster. These clusters are then sequentially merged according to their similarity. First, the two most similar clusters (i.e., those with the smallest distance between them) are merged to form a new cluster at the bottom of the hierarchy. In the next step, another pair of clusters is merged and linked to a higher level of the hierarchy, and so on. This allows a hierarchy of clusters to be established from the bottom up. A cluster hierarchy can also be generated top-down. In this divisive clustering, all objects are initially merged into a single cluster, which is then gradually split up.

Another important group of clustering procedure are partitioning methods. As with hierarchical clustering, there is a wide array of different algorithms; of these, the k-means procedure is the most important one. The k-means algorithm follows an entirely different concept than the hierarchical methods. This algorithm is not based on distance measures such as Euclidean distance or city-block distance, but uses the within-cluster variation as a measure to form homogenous clusters. Specifically, the procedure aims at segmenting the data in such a way that the within-cluster variation is minimized. Consequently, a distance measure in the first step of the analysis is not done in this algorithm.

The clustering process starts by randomly assigning objects to a number of clusters. The objects are then successively reassigned to other clusters to minimize the within-cluster variation, which is basically the (squared) distance from each observation to the center of the associated cluster. If the reallocation of an object to another cluster decreases the within-cluster variation, this object is reassigned to that cluster. With the hierarchical methods, an object remains in a cluster once it is assigned to it, but with k-means, cluster affiliations can change in the course of the clustering process. Consequently, k-means does not build a hierarchy as described before, which is why the approach is also frequently labeled as non-hierarchical.

Md. Hedayetul Islam Shovon (2012) predicted of student academic performance by applying K means clustering algorithm. Class quizzes, mid and final exam marks were used as an evaluation parameter. The study aimed to help the teachers to reduce the drop out ratio to a significant level and improve the performance of students. Oyelade, (2010) predicted the Students' Academic Performance using k-means clustering. He reviewed different clustering techniques that could be applied for educational data mining to predict academic performance of students and its implications.

***Outlier Analysis**:*

A data set may contain objects that do not comply with the general behavior or model of the data. It is a process of finding anomalies in the data set. These anamolies are outliers (Hodge, 2004). Many data mining & EDM methods discard outliers as noise, abnormalities, deviants or exceptions. However, in some applications (e.g., extra ordinary performer in the batch) the rare events can be more interesting than the more regularly occurring ones.

The analysis of outlier data is referred to as outlier analysis or anomaly mining (Barnes, 2005). In EDM, outlier analysis can detect learning issues of students and irregular learning processes by using the learners response time data for e-learning data (Chan, 2007). It can also detect irregularities and deviations in the learners or educators actions by comparing with other learners (Muehlenbrock, 2005). As an example of outlier detection, if any student who hadn't score well in exams got decent placement is an outlier.

Text mining

Text Mining is one of the most critical ways of analyzing and processing unstructured data that contributes roughly 80% of the total data generated today in entire world. It has becomes a huge challenge for organizations to store, process, and analyze vast amounts of textual data generated every minute from multiple sources. Role of text mining becomes very crucial in such situations.

According to Wikipedia, "*Text mining, also referred to as text data mining, roughly equivalent to text analytics, is the process of deriving high-quality information from text."* Text mining probe into unstructured data to extract meaningful patterns and insights required for exploring textual data sources. It combines usage of data mining, machine learning, statistics, and computational linguistics.

Text mining techniques can be used in organizing online information in such a way that helps researchers to easily access immense amount of repositories available online. It can work with unstructured or semi-structured datasets like text documents, HTML files, emails, etc. (Ayesha, 2010). Various text mining tools and techniques are used in EDM to analyze the student's interest in specific field and employment ratio and trends in education in a given geographic region (R. Al-Hashemi, 2010).

Social Network Analysis (SNA)

SNA of sites social networking sites like twitter and Facebook can be very effectively used in EDM to bring out the pattern of student's interaction on these sites. It is a field of study that attempts to understand and measure relationships of users in networked information. In EDM, the approaches can be used for extracting information of group activities (P. Reyes & P. Tchounikine, 2005).

SNA techniques not only focus on social networks, but also focus on other fields, such as marketing (customer and supplier networks) or public safety, although it is not much used in Education domain but its application in e-learning context is tremendous (Palazuelos et al, 2013).

Discovery with Models

Model discovery is a phenomenon of developing a suitable model using prediction, clustering and knowledge engineering (Technical, scientific and social aspects involved in building and using knowledge based systems). This model is then used as a component in another analysis, such as prediction or relationship mining. In the prediction case, the created model's predictions are used as predictor variables in predicting a new variable.

While developing complex constructs like gaming, the system within online learning depends on the assessments of the probability that the student has the current knowledge (Baker et al, 2008; Walonoski & Heffernan, 2006). Hence Discovery with models can be used for conducting analyses of learning behaviors that are difficult to study with more traditional methods (Pardos et al, 2011). A key benefit of discovery with models is being able to study behavioral constructs in a non-disruptive fashion that is both scalable/longitudinal and fine-grained (Pardos, Z.A., Baker, R.S.J.d., Gowda, S.M., & Heffernan, 2011)

These established models has also become a powerful tool for making scientific discoveries about learning and learners, through studying the contexts in which a learning-related behavior occurs or construct emerges, and through examining its relationships with other constructs.

Distillation of Data for Human Judgment

Sometimes it is almost impossible for the automated data mining method to take decisions due to non-clarity of available data. Distillation of data for human judgment aims to make data more clear and understandable. If data is presented in different ways especially visualization then it is easier to understand by human brain. Distillation for human decision is a technique that involves presenting data in such a way that enables a human being to quickly identify and classify the features of the data. Hence the available Data is distilled for human judgment in educational data mining for two key purposes: classification and identification (Baker, 2006).

This helps in improving machine-learning models because humans can identify patterns in, or features of, student learning activities, student behavior's, or data involving collaboration among scholars. It is similar to visual data analytics (J. Kay et al, 2006). All of these above mentioned applications are summarized in the table 2.

TABLE 3: Educational Data Mining Applications

Category	**Objectives**	**Key Applications**
Clustering or Unsupervised learning	Natural grouping of data based on some characteristics. The number of cluster can differ and largely based on the model and objectives	Major applications are finding similarities and differences between students or educational institutions based on performance or any other characteristics important to the stake holders. Also useful in categorizing new student behavior
Prediction	Develop a model to predict one or more variables on the basis of other variables also known as predictor variables.	This is a useful technique to predict success rate, drop outs and learning outcomes in students.
Relationship Mining	Discover the relationship between two or more variables in the data set.	It is useful in Discovering effective pedagogical strategies that can lead to more effective/robust learning. Also useful in the discovery of curricular associations in course sequences
Model Discovery	To develop a model of a phenomenon with prediction, clustering, or knowledge Engineering and can be used as a one of the component in further prediction or relationship mining.	Useful in the discovery of relationships between student characteristics and student behaviour or attitude

Distillation of data for human judgment	To represent data in intelligible ways using summarization, visualization and interactive interfaces to enable a human to quickly identify or classify features of the data.	Helping instructors to visualize and analyze the ongoing activities of the students and the use of information. Human identification of patterns in student learning, behavior, or collaboration; Labeling data for use in later development of prediction model
Outlier detection	To identify extremeness in the given data set	Identifying students having below the average performance to initiate appropriate actions for their improvement in learning process.
Social Network analysis	To analyze the social relationships between entities in networked information.	Interpretation of the structure and relations in collaborative activities and interactions with communication tools.
Process mining	To discover process knowledge from event logs.	Obtaining knowledge of the process from event logs.
Text mining	To extract high-quality information from text	Analysing the contents of forums, chats, web pages and documents.

IV. MAJOR ISSUES IN EDUCATIONAL DATA MINING

Since EDM is still an emerging field, issues related to methodologies and techniques, multidimensionality of data, scalability of data, management of various data bases and privacy issues are very common**.** These issues in educational data mining research are divided into four groups as discussed below:

1. Methodologies in Mining:

There is a rapid development of new EDM methodologies as it creates a new area of research among researching fraternity. This involves the investigation of new kinds of knowledge, mining in multidimensional space, integrating methods from other disciplines, and the consideration of semantic ties among data objects. In addition, mining methodologies should consider issues such as data uncertainty, noise, and incompleteness. Some mining methods explore specific measures that can be used to assess the hidden patterns in students' record as well as guide the discovery process.

Diverse Analysis and Applications:
EDM covers a wide spectrum of data analysis and knowledge discovery tasks, from data characterization and discrimination to association and correlation analysis, classification, regression, clustering, outlier analysis, sequence analysis, and trend analysis. These tasks may use the same database in different ways and require the development of numerous data mining techniques. Due to the diversity of applications, new mining tasks continue to emerge, making EDM a dynamic and fast-growing field.

Multidimensional Data:
When searching for knowledge in large data sets, one can explore the data in multidimensional space. That is, interesting patterns can be searched among combinations of dimensions (attributes) at varying levels of abstraction. Such mining is known as *(exploratory) multidimensional data mining*.

Handling uncertainty, noise, or incompleteness of data:
Data often contain noise (random variations), errors, exceptions, or uncertainty, or are incomplete. Due to the presence of possible uncertainties, no model can predict hundred percent accurate results in terms of student modelling or overall academic planning. Data cleaning, data preprocessing, outlier detection and removal, and uncertainty reasoning are examples of techniques that need to be integrated with the data mining process.

To handle this problem in employability index measurement, following techniques are suitable:

- Regression: smooth by fitting the data into regression functions
- Clustering: detect and remove outliers

2. Scalability issues to problem of Data Compatibility

Scalability is a major performance related issue. Data management has become critical considering wide range of storage locations, data platform heterogeneity and a plethora of social networking sites (Deka, G.C, 2013), E.g.: Metadata Schema Registry is a tool to enhance to enhance Meta data interoperability. So there is a need to design a model to classify/ cluster the data or find relationships. Examples of clustering applications are grouped students based on their learning and interaction patterns used in (Amershi, 2006) and grouping users for purposes of recommending actions and resources to similar users.

Analytical routines required for exploration and modeling run faster on samples than on the entire database. Even the fastest hardware and software combinations have difficulty performing complex analyses such as fitting a stepwise logistic regression with millions of records and hundreds of input variables. Within a database, there can be huge variations across individual records. A few data values far from the main cluster can overly influence the analysis, and result in larger forecast errors and higher misclassification rates. These data values may have been miscoded values, they may be old data, or they may be outlying records. Had the sample been taken from the main cluster, these outlying records would not have been overly influential.

Alternatively, little variation might exist in the data for many of the variables; the records are very similar in many ways. Performing computationally intensive processing on the entire database might provide no additional information beyond what can be obtained from processing a small, well-chosen sample. Moreover, when the entire database is processed, the benefits that might have been obtained from a pilot study are lost.

3. Different types of Databases:

The wide diversity of database types brings about challenges to educational data mining. These include

Handling complex types of data:
Diverse applications generate a wide spectrum of new data types, from structured data such as relational and data warehouse data to semi-structured and unstructured data; from stable data repositories to dynamic data streams; from simple data objects to temporal data, hypertext data, multimedia data, software program code, Web data, and social network data. The construction of effective and efficient educational data mining tools for diverse applications remains a challenging and active area of research.

Mining dynamic, networked, and global data repositories:
Educational data is incremental in nature. Due to the exponential growth of data, the maintaining the data warehouse is difficult. To monitor the operational data sources, infer the student interest, intentions and its impact in a particular institution is the main issue. Another issue is the alignment and translation of the incremental educational data. It should focus on appropriating time, context and its sequence. Optimal utilization of computing and human resources (Deka, 2012) is another issue of incremental educational data.

DOI 10.17148/IJARCCE.2020.9903

4. Influence of Educational Data Mining in Education Society:

The major issues of social influences are privacy issues. With data mining penetrating our everyday lives, it is important to study the impact of educational data mining on student fraternity. Applications utilizing EDM technologies are becoming more and more prevalent in school systems (Simon, 2014), (Singer, 2014). However, the increase in EDM usage has raised concern in parents and students of how much data is being collected about students. The applications and companies that collect and use student data are coming under scrutiny, as parents, college authorities, and public officials grow concerned over student privacy.

Large institutions have been targeted for using student data in undesirable ways (Herold, 2014). This has resulted in the demand for stricter policy from privacy advocates have led to more than 100 bills being introduced in U.S. state legislatures to address issues of student privacy in 2014 (Trainor, 2015). The philosophy should be to observe data sensitivity and preserve people's privacy while performing successful data mining in education sector.

V. CONCLUSION

Educational data mining is a young research area. It is an emerging field related to several well-established areas of research including e-learning, adaptive hypermedia, intelligent tutoring systems, web mining, data mining, etc. Although the educational data mining is a very recent research area there are an important number of contributions published. EDM brings together researchers and practitioners from computer science, education, psychometrics, statistics, psychology and IT domain.

The paper discusses the evolutionary path of information technology, which has led to the need for data mining in education, and the importance of its applications. It examines the data types to be mined, including relational, transactional, and data warehouse data, as well as complex data types such as time-series, sequences, data streams, spatiotemporal data, multimedia data, text data, graphs, social networks, and Web data. It presents a general classification of data mining tasks, based on the kinds of knowledge to be mined, the kinds of technologies used, and the kinds of applications that are targeted specifically related to employability and selection of students. Finally, four major challenges in the field are discussed which includes choosing appropriate methodology, problems of scaling huge datasets, handling of wide diversity in databases and privacy issues in educational data mining.

The major applications of educational data mining are Predicting student performance, modeling students on various academic related parameters, Detecting undesirable behaviors of learners, analyzing and visualizing student's data, providing feedback support to teachers, planning and scheduling evaluations and curriculums, providing recommendations to students for their growth and career path. An important application of EDM is in MOOCs, where data from thousands of students can be employed to redesign courses for future students and automatic personalization according to student profiles (needs, objectives, background, country, learning style, etc.) and performance.

REFERENCES

[1]. Amershi, S., Conati, C. (2006) "Automatic Recognition of Learner Groups in Exploratory Learning Environments". Proceedings of ITS 2006, 8th International Conference on Intelligent Tutoring Systems.

[2]. Amershi, S., and Conati, C., (2009) "Combining unsupervised and supervised cassificationtobuild user models for exploratory learning environments" Journal of Educational Data Mining. Vol.1, No.1, pp. 18-71.

[3]. Anoop Kumar Jain, and Satyam Maheswari, (2012) "A Survey of Recent Clustering Techniques in Data Mining, International Archive of Applied Sciences and Technology" (ISSN: 0976-4828), vol. 3[2], pp: 68-75.

[4]. Asmita Yadav (2013), "A Survey of Issues and Challenges Associated with Clustering Algorithms," IJSETT: International Journal for Science and Emerging Technologies with Latest Trends (Online):2250-3641), vol.10 (1), pp.7-11.

[5]. B R Prakash, Dr.M. Hanumanthapp, Vasantha Kavitha (2014), "Big Data in Educational Data Mining and Learning Analytics", International Journal of Innovative Research in Computer and Communication Engineering, (An ISO 3297: 2007 Certified Organization), Vol. 2, Issue 12

[6]. Baker, R.S.J.d., Corbett, A.T., Aleven, V. (2008) "More Accurate Student Modeling Through Contextual Estimation of Slip and Guess Probabilities in Bayesian Knowledge Tracing". Proceedings of the 9th International Conference on Intelligent Tutoring Systems, 406-415

[7]. Baker, R.S.J.D.,and Yacef, K.(2009), "The state of Educational Data Mining in 2009:A review and future vision" Journal of Educational Data Mining, Vol.1,No. 1,pp.3-17

[8]. Baker, R.S.J.d., Corbett, A.T., Wagner, A.Z.: Human Classification of Low-Fidelity Replays of Student Actions. In: Proceedings of the Educational Data Mining Workshop at the 8th International Conference on Intelligent Tutoring Systems, 29--36 (2006).

[9]. Barnes, T., Bitzer, D., Vouk, M. (2005) "Experimental Analysis of the Q-Matrix Method in Knowledge Discovery" Lecture Notes in Computer Science 3488: Foundations of Intelligent Systems, 603-611.

DOI 10.17148/IJARCCE.2020.9903

[10]. C. C. Chan (2007), "A framework for assessing usage of web-based e-learning systems," in Innovative Computing, Information and Control, 2007. ICIC '07. Second International Conference on, pp. 147–147.

[11]. Beck, J.E., Mostow, J. (2008). "How who should practice: Using learning decomposition to evaluate the efficacy of different types of practice for different types of students". Proceedings of the 9th International Conference on Intelligent Tutoring Systems, 353-362.

[12]. Deka, G.C (2013), "A survey on Cloud Data Base", IEEE Transaction on IT professional, pp.99-106

[13]. David J. Hand (1998) "Data Mining: Statistics and More?" The American Statistician, Vol. 52, No .2.

[14]. David J.Hand (1999) Statistics and Data Mining: Intersecting Disciplines, copyright @ ACM SIGKDD, Vol.1, Issue1.pp.16-19

[15]. Friedman J.H. (1998) "Data Mining and Statistics- What's the Connection", 29th Symposium on the Interface. Houston, USA.

[16]. Fayyad, U. Piatetsky-shapiro G., Smyth. P. (1996). "From Data Mining to Knowledge Discovery in Databases". American Association for Artificial Intelligence, 17, 37-54.

[17]. Gupta, D., Jindal, R., Dutta Borah,M (2011), "A Knowledge Discovery based Decision Technique in Engineering Education Planning" in Proc. Int. Conf. On Emerging Trends & Technologies in Data management. Institute of Management Technology Ghaziabad, pp. 94-102.

[18]. Frawley, W. J., Piatetsky-Shapiro, G., & Matheus, C. J. (1992). Knowledge Discovery in Databases: An overview AI Magazine, 13(3), 57. https://doi.org/10.1609/aimag.v13i3.1011

[19]. Herold B.,(2005) "in Bloom to Shut Down Amid Growing Data Privacy Concerns," Education Week, 04-Feb-2014.

[20]. J. Mostow, J. Gonzalez-Brenes, and B. H. Tan. (2011) "Learning classifiers from a relational database of tutor logs." In Proceedings of the Fourth International Conference on Educational Data Mining.

[21]. J. Kay, N. Maisonneuve, K. Yacef, and P. Reimann (2006), "The big five and visualisations of team work activity," in Intelligent tutoring systems, pp. 197–206.

[22]. Jaya Srivastava, Abhay K Srivastava (2015) "Understanding Linkage between Data Mining and Statistics" International Journal of Engineering Technology, Management and Applied Sciences, Volume 3, Issue 10, ISSN 2349-4476

[23]. Jaya Srivastava, Abhay K Srivastava (2017) "An Overview of Educational Data Mining" International Journal of Advance and Innovative Research Volume 4 Issue 4(III)

[24]. M. Muehlenbrock (2005), "Automatic action analysis in an interactive learning environment," in Proceedings of the 12th International Conference on Artificial Intelligence in Education, pp. 73–80

[25]. Manpreet Kaur & Usvir Kaur (2013), "Comparison between K-Mean and Hierarchical Algorithm Using Query Redirection", International Journal of Advanced Research in Computer Science and Software Engineering. Vol. 3, Issue No. 7

[26]. Merceron, A., Yacef, K. (2008) "Interestingness Measures for Association Rules in Educational Data". Proceedings of the First International Conference on Educational Data Mining, 57-66.

[27]. Merceron, A., Yacef, K. (2005). Educational Data Mining: a Case Study. In International Conference on Artificial Intelligence in Education, Amsterdam, Netherlands, 1-8.

[28]. Mostow, J. (2008). "Experience from a Reading Tutor that listens: Evaluation purposes, excuses, and methods". In C. K. Kinzer & L. Verhoeven (Eds.), Interactive Literacy Education: Facilitating Literacy Environments through Technology, pp. 117-148. New York: Lawrence Erlbaum Associates, Taylor & Francis Group.

[29]. Nikhil Rajadhyax and Rudresh Shirwaikar (2012). "Analyzing Students' Performance Using Frequent item Set Mining, Clustering & Classification". International Journal of Management & Information Technology ISSN: pp 2278-5612 Volume 1, No 2.

[30]. P. Indira Priya and Dr. D. K. Ghosh (2013) . "A Survey on Different Clustering Algorithms in Data Mining Techniques", IJMER; International Journal of Modern Engineering Research (ISSN: 2249-6645), vol.3, Issue 1, pp. 267-274

[31]. P. Reyes and P. Tchounikine (2005), "Mining learning groups' activities in forum-type tools," in Proceedings of the 2005 conference on Computer support for collaborative learning: learning 2005: the next 10 years, pp. 509–513, International Society of the Learning Sciences.

[32]. Palazuelos, C., Garc´ıa-Saiz, D., Zorrilla, M. (2013) "Social network analysis and data mining: An application to the e-learning context". In: International Conference on Computational Collective Intelligence Technologies and Applications.

[33]. Pavlik, P., Cen, H., Wu, L., Koedinger, K. (2008) "Using Item-Type Performance Covariance to Improve the Skill Model of an Existing Tutor". Proceedings of the First International Conference on Educational Data Mining, 77-86.

[34]. R. Al-Hashemi (2010), "Text summarization extraction system (tses) using extracted keywords." International Arab Journal of e-Technology, vol. 1, no. 4, pp. 164– 168.

[35]. Romero , C. et al.,(2008), "Data Mining in course management systems: Moodle case study and Tutorial", Computer and Education, Elsevier publication. Vol. 51, No. 1, pp.368-384.

[36]. Romero, C., and Ventura, S. (2010), "Educational Data Mining: A review of the state of the Art", IEEE Transaction on System. Management and Cyber.-Part C: Appl. and rev., Vol.40, No.6, pp. 601-618.

[37]. Romero, C., and Ventura S. (2013), "DATA MINING IN EDUCATION". WIREs Data Mining and Knowledge. Discovery. Vol.3, pp. 12-27.

[38]. Romero, C., and Ventura, S. (2007), "Educational Data Mining: A survey from 1995 to 2005" Expert Systems with Applications. Vol. 33, pp.135-146.

[39]. R. Agrawal and R. Srikant (1995). "Mining sequential patterns in Data Engineering", Proceedings of the Eleventh International Conference, pp 3–14. IEEE.

[40]. Rama. B, Jayashree. P, and Salim Jiwani. (2010) "A Survey on clustering", IJCSE: International Journal on Computer Science and Engineering (ISSN: 0975-3397), vol. 02, No. 09, pp. 2976-2980.

[41]. Ramageri, M. Bharti (2011) "Data Mining Techniques and Applications". Indian Journal of Computer Sciences and Engineering, Vol.1 No. 4, 301-305.

DOI 10.17148/IJARCCE.2020.9903

[42]. Ramandeep Kaur, and Dr. Gurjit Singh Bhathal. (2013) "A Survey of Clustering Techniques". International Journal of Advanced Research in Computer Science and Software Engineering (ISSN: 2277 128X), vol. 3, Issue 5.

[43]. Ramesh V., Thenmozhi P. ,Ramar. K. (2012) ,"Study Of Influencing Factors Of Academic Performance of Students: A Data Mining Approach" International Journal Of Scientific And Engineering Research , Vol.No.3 Issue no.7 .

[44]. Ranjan J, Ranjan R., (2010), "Application of Data Mining Techniques in Higher Education in India", Journal of Knowledge Management Practice, Vol. 11, Special Issue 1.

[45]. S. Ayesha, T. Mustafa, A. R. Sattar, and M. I. Khan (2010), "Data mining model for higher education system," Europen Journal of Scientific Research, vol. 43, no. 1, pp. 24–29.

[46]. Siraj F., Abdoulha A. M. (2011) "Mining enrolment data using predictive and descriptive approaches" Applied Sciences, College of Art and Sciences, University Utara Malaysia.

[47]. Simon S., (2014) "Data Mining Your Children," Politico

[48]. Stephen M. Stigler (2002) "Statistics on the table: the history of statistical concepts and methods", Cambridge, Mass : Harvard university Press.

[49]. Sukanya M., Biruntha S., Karthik S. (2012), "Data Mining: Performance Improvement in Education Sector using Classification and Clustering Algorithm" International Conference on Computing and Control Engineering

[50]. Suma. V, Pushpavathi T.P, and Ramaswamy. V. (2012). "An Approach to Predict Software Project Success by Data Mining Clustering" . International Conference on Data Mining and Computer Engineering (ICDMCE'2012), Bangkok (Thailand).

[51]. Sunita B Aher and LOBO L.M.R.J (2011). "Data Mining in Educational System using WEKA" International Conference on Emerging Technology Trends (ICETT).

[52]. T. Hastie, R. Tibshirani, and J. Friedman (2009). "The Elements of Statistical Learning: Data Mining, Inference, and Prediction" (2nd ed.). Springer Verlag.

[53]. Trainor S., "Student data privacy is cloudy today, clearer tomorrow," Phi Delta Kappan, vol. 96, no. 5, pp. 13–18, 2015.

[54]. V. J. Hodge and J. Austin (2004), "A survey of outlier detection methodologies," Artificial Intelligence Review, vol. 22, no. 2, pp. 85–126, 2004.

[55]. Walonoski, J., & Heffernan, N. (2006). Detection and analysis of off-task gaming behavior in intelligent tutoring systems. Proceedings of the 8th International Conference on Intelligent Tutoring Systems, 382-391.