International Journal of Advances in Engineering Research http://www.ijaer.com
(IJAER) 2012, Vol. No. 4, Issue No.IV, October **ISSN:2231-5152**

DATA MINING IN EDUCATION DOMAIN

***DR. U.S.PANDEY, #MR. AVANEESH ANAND SINGH**

**ASSOSIATE PROFESSOR, SOL, UNIVERSITY OF DELHI*
#M.PHIL. MADURAI KAMRAJ UNIVERSITY

ABSTRACT

Conventional face-to-face educational systems are widely characterized by direct interaction between the students and the teacher. Although there are many categories and variations of such educational systems (Private, Public etc.) collectively they have been criticized for promoting passive learning and ignoring individual learner capacity instead of encouraging problem solving skills, critical thinking and personalized learning. In established educational systems, the student performance is mainly measured based on student progress records such as student attendance and grades. But in contrast to Web Based System With advance web technologies and the Internet, web based educational or widely known as e-Learning became very famous. An ideal e-Learning system is characterized by features enabling the learning take place without the intervention of space, location or time constraints. However, web based educational systems are often criticized for lack of proper feedback to students and poor quality of user interactions. In recent years, with the growth of data mining EDM is defined as the area of scientific inquiry centered on the development of methods for making discoveries within the unique kinds of data that come from education domain, and using those methods to better understand students and the domain which they learn in . Utilization of information and communication technologies in educational processes generates large amount of data as a side effect. Educational data may contain many interesting information and potential knowledge about students and their learning habits. However, such knowledge is hidden and their extraction is not trivial. Knowledge discovery from databases (also known as Data Mining) is a methodology for extraction of non-trivial, previously unknown, and potentially useful knowledge from data.

Keywords: *EDM (Educational data mining), Data Mining, Knowledge Discovery*

INTRODUCTION

In the education domain, a recommendation system is an intelligent agent that suggests different alternatives to students, having as starting point previous actions from other students with approximately the similar characteristics, such as academic performance and other personal information. It is known that before taking a course, the student have to enroll on the course; the most notorious of this process is not enrolment itself, but the previous decision that has to be taken, mainly related to how many and which course are going to be taken. In this work, we show a web based collaborative recommendation system based on data mining techniques applied to the educational environment. The aim of this work is to offer students and educators as key elements to take better decisions in variousacademic processes , using as basis the academic performance of other students with similar profiles, in order to obtain good decision in pertaining to its academic itinerary. To retain qualified in educational domain, a deep understanding of the knowledge hidden among the data is required. In today's higher education lack of deep and enough knowledge among the processes such as evaluation, counseling and etc, prevents management system from achieving this quality objective, so there has not been an efficient and effective use of their strategic resources yet. Data mining techniques can be used to extract unknown pattern from the set of data and discover useful knowledge, which would assist decision makers to improve the decision-making and policy-making procedures. It results in extracting greater value from the raw data set, and making use of strategic resources efficiently and effectively. It finally improves the quality of higher educationalprocesses.

The main objective of higher education institutes is to provide quality education to its students and to improve the quality of managerial decisions. One way to achieve highest level of quality in higher education system is by discovering knowledge from educational data to study the main attributes that may affect the students' performance. The discovered knowledge can be used to offer a helpful and constructive recommendations to the academic planners in higher education institutes to enhance their decision making process, to improve students' academic performance and trim down failure rate, to better understand students' behavior, to assist instructors, to improve teaching and many other benefits[1].

Automated learning environments collect large amounts of information on the activities of their students. Unfortunately, analyzing and interpreting these data manually can be tedious and requires substantial training and skill. Although automatic techniques do exist for mining data, the results are often hard to interpret or incorporate into existing scientific theories of learning and education. We therefore present a model for performing automatic scientific discovery in the context of human learning and education. We demonstrate, using empirical results relating the frequency of student self-assessments to quiz performance, that our framework and techniques yield results better than those available using human-crafted features.

One of the basic goals of scientific research is the modeling of event. In particular, we are interested in examining the data produced by students using an on-line course. Intuitively, researchers believe many interesting, and potentially useful trends and patterns are contained in these logs. Researchers usually begin with some general idea of the event they would like to understand, and then proceed to collect some observations of it. The scientist, for example, might have some prior belief, based on existing scientific theory and intuition, that the amount of time a student spends reading course notes will affect his performance on quizzes, but is not able to specify exactly what he means by ―time reading notes.‖ Is it the cumulative number of minutes, split into any number of sessions, conducted under any condition, prior to the evaluation that matters? Or is the intensity of the reading more important? Is it better to read the notes right before the quiz, for higher recall, or perhaps an earlier viewing helps prime the student for learning? Unfortunately, researchers have neither the time nor patience to go through all these logs, by hand, to find ideal instantiations of their features. In this work, we develop a partial solution to this problem of feature discovery that uses a computer to intelligently induce higher-level features from low-level data.

Although computers can produce copious log data, the unstructured, low-level nature of these data unfortunately makes it difficult to design an algorithm that can construct features and models the researcher and his communities are interested in and can understand. In fact, the complexity of a full search of the feature space, from a statistical point of view, would depend on the size of the sufficient statistics of the entire data set. Thus, for all real-world problems, brute force search is intractable. A more insidious problem, however, is that even if the space of features were able to be enumerated and searched efficiently, the number of possible models based on those features would be even larger, and any attempt at learning a true model would suffer from overfitting and the curse of dimensionality. Although techniques do exist for addressing this issue, many do not take into consideration the semantics of the features, instead relying on an estimation of complexity. It turns out that by carefully limiting the types of features that we can represent and search, we reduce our search and overfitting problems without, hopefully, cutting out too many expressive features. Certain techniques do exist for addressing feature selection. Principle component analysis (PCA) (e.g. Schölkopf, Smola, and Müller 1998 in [2]), for example, finds a projection of the data from a higher dimensional space to one with fewer dimensions. This projection reduces the number of features needed to represent the data. Unfortunately, these projections distort the original, presumably intuitive, definition of the features of the data into linear combinations of these features. In this process, much of the interpretability of the resulting models is sacrificed. This weakness is present in many of the other methods used for feature selection and dimensionality reduction, such as clustering and kernel methods (Jain, Duin, and Mao 2000 in [3]). All suffer from a sacrifice of interpretability which has been shown to be essential if computational techniques should ever have a serious impact on the progress of scientific research (Pazzini, Mani, and Shankle2001[4])[5].

In this paper, we will discuss several data mining approach to discuss enhancement of facility to student as well educators who are key elements of higher education system. The rest of the paper is organized as follows: In section 2 we present previously discussed Real world process mining concept .Next in section 3 we present a new non parametric data mining techniques MARS to discuss its implication on education. In section 4 we experiment knowledge discovery process with support level analysis. In Section 5 we present knowledge discovery from given data set using K-mode method. And in section 6 we present conclusion with givenframework.

REAL PROCESS MINING FRAMEWORK

Conventional data mining techniques have been rapidly applied to find interesting patterns, build various models from large volumes of data accumulated through the use of different information systems. Outcome of data mining can be used for getting a better understanding of the fundamental educational processes, for generating recommendations and advice to students, for improving management of learning objects, etc. However, most of the conventional data mining techniques focus on data dependencies or simple patterns and do not provide a visual representation of the complete educational process ready to be analyzed. To allow for these types of analysis where process is playing central role, a new line of data-mining research, called process mining, has been initiated. Process mining focuses on the development of a set of intelligent tools and techniques aimed at extracting process-related knowledge from event logs recorded by an information system. Therefore, we demonstrate the applicability of process mining, and the ProM framework in particular, to educational data mining context. We analyze assessment data from recently organized online multiple choice tests and demonstrate the use of process discovery, conformance checking and performance analysis techniques. Process mining has emerged from the field of Business Process Management (BPM). It focuses on extracting process-related knowledge from event logs recorded by an information system. It aims particularly at discovering or analyzing the complete (business, or in our case educational) process and is supported by powerful tools that allow getting a clear visual representation of the whole process. The three major types of process mining applicationsare

- *Harmony checking* - reflecting on the observed reality, i.e. checking whether the modeled behavior matches the observedbehavior

- Process model discovery - constructing complete and compact process models able to reproduce the observedbehavior

- Extension of Process model - projection of information extracted from the logs onto the model, to make the tacit knowledge explicit and facilitate better understanding of the process model.

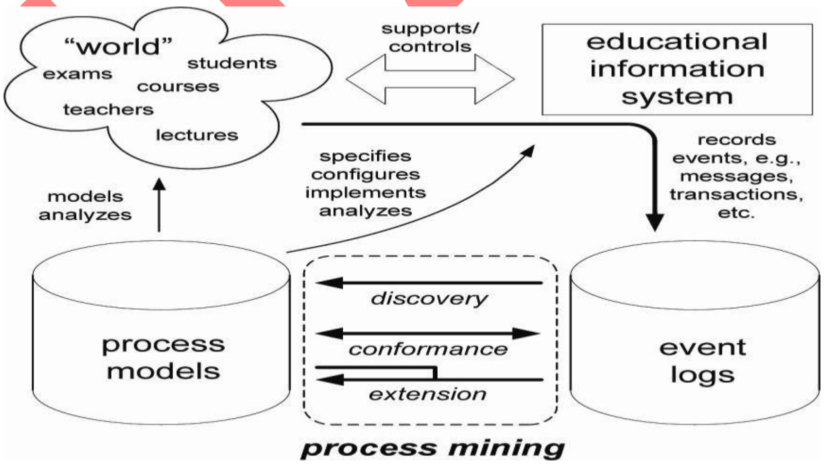

Figure 1 The Process mining spectrum supported by ProM

1. Identifying prognosticate of student retention:

One of the fundamental issues for all university policy makers due to the potential negative impact on the image of the university and the career path of the dropouts is the student retention. Although this issue has been thoroughly studied by many institutional researchers using parametric techniques, such as regression analysis and logitmodeling,

Data mining procedures identify transferred hours, residency, and ethnicity as crucial factors to retention. Carrying transferred hours into the university implies that the students have taken college level classes somewhere else, suggesting that they are more academically prepared for university study than those who have no transferred hours. Although residency was found to be a crucial predictor to retention, one should not go too far as to interpret this finding that retention is affected by proximity to the university location. The geographical information system analysis indicates that non-residents from the east coast tend to be more persistent in enrollment than their west coast schoolmates. Universities with high attrition rates face the substantial loss of tuition, fees, and potential alumni contributions (DeBerard, Spielmans and Julka, 2004[6]), while the students themselves also face negative consequences. Despite the identified consequences of college dropout for universities and students, as well as concentrated efforts from all educational institutions on improving student retention, attrition rates remain relatively high across the United States. Data from the National Center for Public Policy and Higher Education reveal that only 73.6 percent of first-time, full-time freshmen (enrolled in 2002) returned for their second semester. Looking at college completion data from 2005, only 39.5 percent of undergraduate students enrolled in public institutions completed their degrees within five years. Tinto's (1975)[7] widely accepted model of student retention examines factors contributing to a student's decision to continue their higher education. The primary focus of this model is a student's academic and social integration into the university. Another model of student retention, developed by Bean in the 1980's, focuses on the psychological and behavioral factors related to student retention (Bean and Eaton, 2001) [8]. Despite the differences in their models of student retention, the commonality between Tinto and Bean's models is the broader concept of student integration. While Bean's focus is on the psychological factors contributing to student integration, a very difficult concept to measure, the goal of each model is to determine the influences on student retention, as is the case in this study. Although in this study neither direct variables relating to social integration nor a way of measuring the underlying psychological processes are available, efforts were devoted to collect proxy measures to social integration, such as residency, living locations, and online course enrollment. It is assumed that being a resident and living on campus could enhance social integration, and conversely, taking many online courses could make a student socially isolated. Although this issue has been thoroughly studied by many institutional researchers using parametric techniques, such as regression analysis and logit modeling, very few studies on retention yield the strong predictive power associated with data mining tools (Herzog, 2006)[9]. This article attempts to bring in a new perspective by exploring theissuewiththeuseofthreedataminingtechniques,namely,classificationtree,multivariate adaptive regression splines (MARS), and neural network. In discussing retention statistics, it is important to explore the definition and methods for calculating persistence and retention. Retention rates are generally calculated based on data from first-time, full-time freshman students who graduate within six years of their initial enrollment date (Hagedorn, 2005)[10], Freshman persistence is commonly defined as returning to regular enrollment status in the first semester of the sophomore year and is strongly associated with the likelihood of eventual graduation from the institution (Mallinckrodt and Sedlacek, 1987)[11]. However, major gaps exist in the literature on retaining students beyond their freshman year (Nara, Barlow and Crisp, 2005)[12], despite the importance of are still lost after completing their first year. If 73.6 percent of students persist to their sophomore year, but only 39.5 percent of students graduate within five years, then approximately 34.1 percent of students are lost after completing their freshman year. There is an overwhelming amount of research on freshman persistence; however, the purpose of this study is to examine the less-researched factors that lead to student persistence beyond the freshman year. In this study, retention rates will be studied with data from sophomore students who initially enrolled in the 2003 academic year, following these students through their junioryear.

2. Available DataSource:

In this study, a data set was compiled by tracking the continuous enrollment or withdrawal of 6690 sophomore students enrolled at Arizona State University (ASU) starting in 2003. The dependent variable is a dichotomous variable, retention. In this study, retention is defined as persisting enrollment within the given time frame (2003-2004 academic years, excluding summer). It is understandable that sometime students may take off one semester for various reasons. Thus, non-persisting enrollment is defined as being absent from two consecutive semesters. There are three sets of potential predictors:

- Demographic: This set of predictors includes gender, ethnic, residence (in state/out of state), and location (living on campus/off campus). An Arizona resident is an adult person (18 years or older) who physically resides in the state for twelve consecutive months immediately preceding the term for which resident classification is requested. Students who live on campus are those having a residential halladdress.

- Pre-college or external academic performance indicators: This set of variables includes high school GPA, high school class rank, Scholastic Aptitude Test (SAT) quantitative z-scores, SAT verbal z-scores, American College Testing (ACT) English z-scores, ACT reading z-scores, ACT mathematics z-scores, ACT science reason z-scores, transferred hours, and university mathematics placement test scores. High school class rank indicates a student’s academic ranking relative to his or her classmates. For example, a student ranked 10 out of a class size of 100 would have a class rank 10%. SAT and ACT are standardized tests administered by the US College Board and ACT, Inc., respectively. Although there isa widely used formula to convert SAT combined scores to ACT composite scores and ACT combined scores to SAT composite scores, this conversion scheme was not adopted because both SAT and ACT are composed of sub-tests specific to different cognitive abilities. Rather, different exams addressing different domains in SAT and ACT were used. All SAT and ACT scores were rescaled as z-scores in order to facilitate comparison. All applicants must take either SAT or ACT in order to apply for admission, but the SAT is more popular than the ACT and some students took both examinations (SAT: 73.7% vs. ACT: 45.9%). In this sense, ACT has more missing values than SAT. Nevertheless, when both variables are treated as academic performance indictors in terms of standardized exams, every student has this performance indictor in one way or the other. University math placement test is an internal examination. ASU requires all incoming freshmen to complete the Unified Placement Test (UPT) and enroll in the appropriate mathematics course as determined by theirscore.

- Online class hours as a percentage of total hours during the sophomore year: Online classes are courses operated in a completely online fashion and thus hybrid classes are excluded from thiscategory.

3. Methodsused:

As all we know data mining, as a form of exploratory data analysis, is the process of automatically extracting patterns and relationships from immense quantities of data rather than testing pre-formulated hypotheses (Han and Kamber, 2006; Larose, 2005; Luan, 2002)[13][14][15]. In addition, typical data mining techniques include cross-validation, which is considered a form of re-sampling (Yu, 2007) [16]. The major goal of cross-validation is to avoid random error, which is a common problem when modelers try to account for every structure in one data set. As a remedy, cross-validation double-checks whether the alleged fitness is too good to be true (Larose, 2005) [17]. Hence, data mining can be viewed as an extension of both EDA and re-sampling. But unlike EDA that passes the initial finding to confirmatory data analysis (CDA), data mining tools, with the use of re-sampling, can go beyond the initial sample to validate the findings (Cuzzocrea, Saccardi, Lux, Porta and Benatti, 1997) [18]. In this study three data mining tools, namely, classification trees, neural networks, and multivariate adaptive regression splines (MARS) were employed. Several researchers conducted comparisons between traditional parametric procedures and data mining techniques, and also within the domain of data mining procedures. It is not surprising to learn that on some occasions one technique outperforms others in terms of prediction accuracy while in a different setting another technique seems to be the best. In this study we argue that using data mining is more appropriate to the study of retention and other forms of institutional analysis than its classical counterpart. First, using such large sample sizes as are found in institutional research will cause the statistical power for any parametric procedures to be 100%. On the contrary, data mining techniques are specifically designed for large data sets (Shmueli, Patel and Bruce, 2007)[19]. Second, institutional research data elements represent multiple data types, including discrete, ordinal, and interval scales. Traditional techniques, such as logistic or linear regression or discriminated function analysis, cannot handle this kind of complexity of data types in one single analysis unless tremendous data transformation, such as converting categorical variables to dummy codes, is used (Streifer and Shumann, 2005)[20]. Further, certain data mining techniques are robust against outliers and also can handle missing data without having to delete outliers, observations with missing values or perform data imputation (Shmueli, Patel and Bruce, 2007)[19]. Since tedious data cleaning is not necessary, it is especially convenient for institutional researchers to employ data mining for handling a huge dataset.

More importantly, most conventional procedures do not adequately address two important issues, namely, generalization across samples and under-determination of theory by evidence (Kieseppa, 2001)[21]. It is very common that in one sample a set of best predictors was yielded from regression analysis, but in another sample a different set of best predictors was found (Thompson, 1995)[22]. In other words, this kind of model can provide a post hoc explanation for an existing sample (insample forecasting), but cannot be useful in out-of-sample forecasting. Further, even if a researcher found the so-called best fit model, there may be numerous possible models to fit the same data. Nevertheless, data mining procedures have built-in features that can counteract the preceding problems. In most data mining procedures cross-validation is employed based on the premise that exploratory modeling using the training data set inevitably tends to over-fit the data. Hence, in the subsequent modeling using the testing data set, the overfitted model will be revised in order to enhance its generalizability. Specifically, the philosophy of MARS is built upon balancing the overfitted local models and the underfitted global model. MARS partitions the space of input cases into many regions in which local models fitting with cubic splines are generated. Later MARS adapts itself across the input space to generate the best global model. While there are more than one theory or model that can adequately fit the data, this problem is known as the problem of under-determination of theory by data. To remediate the problem of under-determination of theory by data, neural networks exhaust different models by the genetic algorithm, which begins by randomly generating pools of equations. These initial randomly generated equations are estimated to the training data set and prediction accuracy of the outcome measure is assessed using the test set to identify a family of the fittest models. Next, these equations are hybridized or randomly recombined to create the next generation of equations. Parameters from the surviving population of equations may be combined or excluded to form new equations as if they were genetic traits inherited from their ‖parents.‖ This process continues until no further improvement in predicting the outcome measure of the test set can beachieved.

IMPLICATION OF MARS ON EDUCATION

MARS is a data mining technique (Friedman, 1991; Hastie, Tishirani and Friedman, 2001)[23][24] for solving regression-type problems. Like EDA, MARS is a nonparametric procedure, and thus no functional relationship between the dependent and independent variables is assumed prior to the analysis. MARS accepts the premise that most relevant variables affect the outcome in a complex way.Thus,MARS–learns‖abouttheinter-relationshipfromasetofcoefficientsandbasisfunctions inadata-drivenfashion.MARSadoptsa–divideandconquer‖strategybydividingtheinputspace into regions, in which a local model is built with its own regression equation. When MARS considers whether to add a variable, it simultaneously searches for appropriate break points, namely, knots. In this initial stage MARS tests variables and potential knots, resulting in an overfit model. In the next stage MARS eliminates redundant variables that do not hold themselves under rigorous testing based upon the criterion of lowest generalized mean square errors in generalized cross validation (GCV). Because a global model tends to be biased but have low variance while local models are more likely to have less bias but suffer from high variance, the MARS approach could be conceptualized as a way to balance between bias andvariance.

Unlike conventional statistical procedures that either omit missing values or employ data imputation, MARS generates dummy variables when encountering variables that have missing values. These dummy variables represent the absence or the presence of data for the predictors in focus and are used to develop surrogate sub-models. This approach is useful in the analysis of epidemiological studies (e.g. Chou, Lee, Shao and Chen, 2004; Kuhnert, Do and McClure, 2000)[25] because when the focal variables have many missing values that invalidates use of logistic regression, epidemiologists can still see how the inversed variables compete equally with other variables for entry into the model. However, it is not strongly relevant in the setting of educational research. For clarity of interpretation, direct variables rather than new variables generated by missing values will be discussed in the results section. Last, in this analysis the software module named MARS with five-fold cross-validation was employed. Teachers and parents need to take a number of factors into account when they consider making a request for a student to be retained. Teachers can use the characteristics of students who are more likely than average to be retained to identify at risk students in need of an intervention. Interventions based on variables that negatively impact student academic performance such as low reading ability and low math scores can help reduce the risk of student grade retention. For example, Gatti (2004) conducted a study to provide empirical evidence of the effectiveness of several classroom activities designed to help students with low scores in reading and mathematics. One hundred seventy-seven students in grades 2, 3, and 4 were placed in the eight-week intervention based on teacher nominations of students they believed would be referred for grade retention within the next month. Students placed in the intervention received a series of assessments three times a week in 20-25 minute sessions. Assessments included teacher nominations for retention which were used to increase the accuracy of referrals to special education services.

EXPERIMENTING KNOWLEDGE DISCOVERY PROCESS WITH SUPPORT LEVELANALYSIS

The main characteristic of formal education is teaching within an administrative framework. Therefore a lot of teaching and administrative activities must be carried out during different education processes, from library services evaluation to building cognitive student models. Efficient development of such education processes usually rest either on intuition or on performing data analysis techniques in order to reveal key concepts and their relationships. We investigate the application of data mining techniques within the education framework. Data mining techniques are used to automatically extract knowledge, in the form of relationships and patterns, from large databases. We do not aim at an exhausted evaluation of all application areas within education. Instead, we rather aim at interesting the reader in such an idea by presenting some practical cases. Data Mining is a process through which one can extract valuable knowledge from a large database. The necessity for the development of data mining evolved due to the immense and quick growth of the volume of stored corporate data. Ordinary querying methods could no longer produce results showing hidden patterns in such vast amounts of data. Using advanced methods derived from artificial intelligence, pattern recognition and statistics, data mining can construct a comprehensively descriptive model on input data. The data model can be produced in various forms and serves the purpose of describing and predicting behavior of the data object.

On the other hand, teaching and administrative activities that must be carried out during different education processes usually rest on performing data analysis techniques in order to reveal key concepts and their relationships. Data mining techniques are best suited for extracting from data such key concepts along with their relationships. We try to investigate the application of data mining techniques within the education framework, aiming at interesting the reader in such an idea, presenting some practical cases.
The process of knowledge discovery involves several steps [26] one of which is applying the data mining technique that we have chosen, according to the nature of the data and the kind of knowledge we would like to extract. These steps are shown in the followingdiagram.

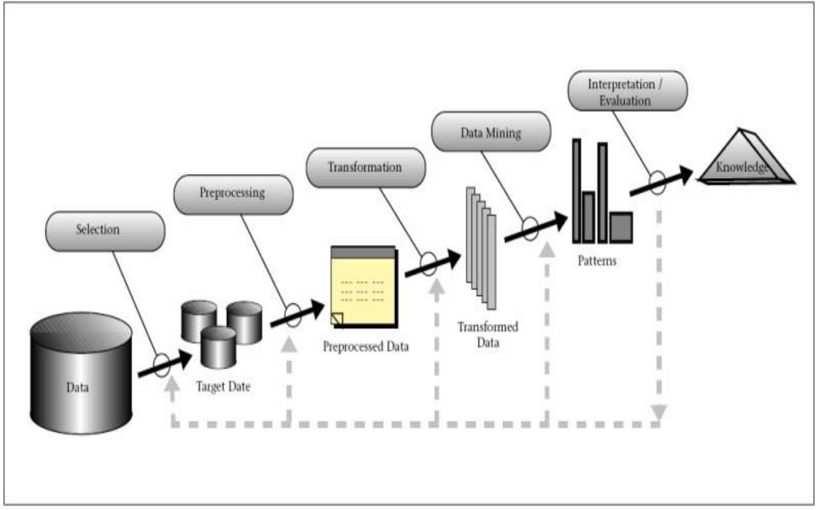

Figure 2: The Knowledge DiscoveryProcess

The first three steps in the Figure 2 involve data handling. The data are rarely stored in a form suitable for data mining. They have to be selected from various sources and then combined into one dataset, upon which various transformations may be applied. For example we may have numerical values which are generally better to be transformed into categorical values, thus producing more comprehensive results. These steps are highly important for the overall success of the process, perhaps equally important to the actual data mining step. The data mining step is the application of the actual algorithm on the pre-processed and probably transformed dataset. The analyst that performs the process has to carefully examine which algorithm to apply and how to specify its parameters, factors that can dramatically affect the quality of the results.

Pre-processing steps have a strict technical context, as well as the specification of the critical parameters of data mining algorithms. Due to them, there is an impression that data mining is not an automatic process, in most of cases. Some users may argue that data mining is mostly a model driven exercise, since the researcher needs to have a clear understanding of the domain, the semantics of the dataset (in order to perform data preprocessing and transformation) and finally to have a clear objective for the data mining process in order to drive the knowledge extraction. Such arguments do not hold true if we refer to an integrated data mining system (eg. Clementine, Keppler, etc), with an appropriate userinterface.

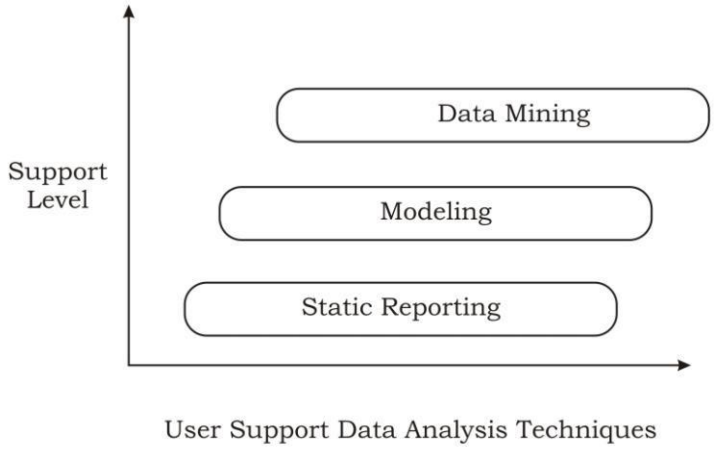

Figure 3 Support level analysis

After the data mining step is completed, the output may be subjected to a filtering step during which it is evaluated. Since data mining is an automated, data-driven process, it may produce results that could appear evident or even naïve. Therefore, it is compulsory to filter out only the knowledge that is truly useful and understandable. In some cases the process may be repeated by transforming the data in a different way and altering the parameters specified. Data mining is a revolutionary technique in terms of the level of support it provides to the user. As shown in the given diagram, it exceeds other popular data analysis techniques such as modeling. When we refer to the level of support in the diagram we mean the degree to which each process is automated. Static reporting is the process where the user foresees the patterns among the data and issues pre-designed queries to get the actual report. Modeling is performed using statistical software such as SPSS. It is more advanced than static reporting but is still user-driven since the user must make a work hypothesis upon which the system operates. Data mining on the other hand is completely data-driven. The user has to make no assumptions in regard to the results that he wants to receive. He simply feeds the data to the process and the results are produced depending solely on the form and values of the input data. Data mining techniques are divided into various categories. The main categories are *classification*, *clustering* and *association*. These kinds of techniques are preferred because they generally produce the best results. Classification is a process that aims at defining a model used for classifying data cases into one class of a set of predefined classes. Each unclassified data case can then be classified by the model, according to its values in certain fields. Classification algorithms are applied on –training‖datasetsthatcontaincasesthathavealreadybeenclassified.Themodeltheyproducecan be in several forms such as trees or sets of classification rules. Clustering also divides input data into groups. The main difference is that there are no predefined classes in this case. Clustering algorithms examined the data with the purpose of finding similarities among the different cases justifying their grouping into *clusters*. Finally, association aims at discovering dependencies among the values of different attributes. It produces rules that have the form $X => Y$, meaning that where there is X there is probably Y as well. There have been found uses for data mining in a significant number of business and research activities. Marketing is an area that has been greatly assisted by data mining. Throughthedataminingprocess,customer-relateddatacanbe–minedupon‖inordertodiscover patterns in customer behavior. By taking advantage of that knowledge, marketing policies can be redesigned for increased efficiency. In retail sales, data mining can be used for –market basket‖ analysis. By applying an association algorithm on retail data we can see which products are usually purchased together. We can use that information for advertising purposes or for better design of market shelves. Other areas have benefited from data mining such as manufacturing, finance and medicine. In general, we can apply data mining wherever data can be shaped or interpreted as various instances of one concept, enabling us to discover knowledge for the behavior of thatconcept.

K-M ODEEXPLANATION

In the following case study, we will demonstrate how one can efficiently go through all the steps of the knowledge discovery process, in order to receive quality results from data mining procedures. The input data are student records from the Ovrya High School, Achaia, Greece recorded during the school periods 1998-1999 and 1999-2000. The data were given in two Microsoft Access□ databases, one for each school period. Since we did not have various data sources to choose from, there was no need for a selection step in the process. There was, however, a great deal of tasks to perform in the preprocessing and transformation steps in order to bring the information included in the data in the best possible form for data mining.

The data were in identical structures in both databases so the procedure was repeated for both databases up to the point where we unified data from the two school periods into one dataset, as described later in this section. Initially, there were separate tables for student data, course data and trimester grades. These tables had to be combined into one dataset using multiple queries. That brought us to a point where we had a dataset with records including all three trimester grades per course, per student. That is to say, there was one record for student A's trimester grades in course A, another for the same student's grades in course B etc. Afterwards, we calculated an extra field containing the final grade. However, that was hardly the dataset form required to conduct data mining procedures. Dependencies and associations in data that data mining is used to extract lie among different fields of the dataset, not among different rows. So what we needed to do was, in a way, to sum up the information in that dataset into another dataset where each record would contain all the final grades for each student. Since that required complicated data manipulation scripts, we transferred our datasets (one for every school period) to an Oracle□ database where data manipulation is quite easier. After creating the new summarized datasets, we had to categorize all the grade attributes, since these were obviously numerical ones. Finally, we combined the two datasets into one and eliminated null records and fields. The null fields were courses that although recorded in the original courses table, it turned out there were no grades for them. In the end, we had one summarized and categorized dataset, containing 428 records free of null values, ready to be mined upon.

Applying the data mining algorithms on our dataset was a relatively simple process. The fields that interested us were obviously the ones related with the grades of each student. We also conducted a few runs to find dependencies among the sex of the students and their performance in various courses. We will now present the results of each algorithm on the data after having fulfilled the evaluation step. The results presented are a minor selection of the full amount of the data mining results, and their purpose is to demonstrate what kind of knowledge we are looking for out of this process. We applied the classification algorithm C4.5 to mine classification rules from the data. C4.5 firstly mines a decision tree by dividing the record set in each node according to an attribute's values. The rules are extracted by considering each path of the tree as a classification rule. Several runs were performed for classification since each time one field has to be the class field, the field, that is, whose values determine the class to which each record belongs. Thus, we can examine what kind of students do well in History, for example, by selecting History as a class field and other courses as fields on whose values the rules will be based. Some of the results where more or less expected, easy for one to assume. The following classification rules that were mined can serve as anexample.

If RELIGIOUS MATTERS=EXCELLENT and

MODERN GREEK=EXCELLENT

Then HISTORY=EXCELLENT

(70/8)

If RELIGIOUS MATTERS=FAIL and

COMPOSITION=FAIL and

ANCIENTGREEK-TRANSLATION=FAIL and

MODERN GREEK=FAIL

Then HISTORY=FAIL

(16/1)

The first rule states that 62 out of the 70 students who received excellent grades in Religious Matters and Modern Greek did the same in History. The second one states an opposite situation, that 15 out of the 16 students who failed in Religious Matters, Composition, Ancient Greek – Translation and Modern Greek, failed in History as well. Those rules represent knowledge that is easy to understand and one could rush to say that they be discarded as useless because they do not imply something new. However, these rules give us the chance to verify with real data what we previously knew only as a vagueassumption.
The next two rules might seem more useful to the sceptic eye.

If RELIGIOUS MATTERS=EXCELLENT and

MODERN GREEK=PASS

Then HISTORY=VERY GOOD

(3/0)

If MODERN GREEK=VERY GOOD and

ANCIENTGREEK=FAIL and

ANCIENTGREEK-TRANSLATION=VERY GOOD

Then COMPOSITION=VERY GOOD

(26/4)

The first rule states that all three students who were excellent in Religious Matters and merely passed Modern Greek were very good in History.That is something that could be explained perhaps if we think about the nature of the courses and how these are being taught. The second rule states that students who were very good in Modern Greek, Ancient Greek Translation and failed in regular Ancient Greek were very good in Composition. That clearly demonstrates that students who are very good in Modern Greek courses can face severe difficulties in Ancient Greek.

The clustering algorithm k-modes were applied so as to discover the different clusters that students can be grouped to. K-modes divide a record set into clusters by specifying the *center* record for each cluster. Its methodology is based on minimizing the average distance of a cluster's records to its center.

We instructed the algorithm to divide students into four clusters. The algorithm returned the center for each cluster that is the average description of the students who belong to that cluster. Here are the four centers that k-modes returned.

	1st CLUSTER	**2nd CLUSTER**	**3rd CLUSTER**	**4th CLUSTER**
Religious Matter	PASS	EXCELLENT	PASS	VERY GOOD
History	PASS	EXCELLENT	FAIL	VERY GOOD
English	VERY GOOD	EXCELLENT	PASS	EXCELLENT
French	PASS	EXCELLENT	FAIL	VERY GOOD
Mathematics	PASS	EXCELLENT	FAIL	VERY GOOD
Gymnastic	PASS	PASS	PASS	EXCELLENT
Ancient Greek Translation	PASS	EXCELLENT	FAIL	VERY GOOD
Ancient Greek	PASS	EXCELLENT	FAIL	VERY GOOD
Modern Greek	PASS	EXCELLENT	FAIL	VERY GOOD
Composition	PASS	EXCELLENT	FAIL	VERY GOOD
Music	PASS	PASS	FAIL	EXCELLENT
Artistic Matters	PASS	PASS	FAIL	EXCELLENT
Informatics	FAIL	EXCELLENT	FAIL	VERY GOOD

Table 1: K-mode

As was instructed, the algorithm divided the students into four clusters. Those who failed in almost every course, those with passing grades, very good grades and excellent grades. But it is interesting to see that excellent students merely passed gymnastics, artistic matters and music that require other talents except intellectuality and being good at studying. On the other hand, students who are characterized as very good in almost all other courses were excellent in these three. A priori tries to find the association among different values of the attributes by evaluating the number of appearances of the various combinations. The algorithm resulted with tremendously numerous rules, associating grades in some courses with grades in others. The majority of these rules associated good grades from various courses or failing grades or passing grades and so on, verifying our conclusions from the previous two techniques. For example:

History = EXCELLENT, Informatics = EXCELLENT and Religious Matter =EXCELLENT

English = PASS, Artistic M. = PASS and History = FAIL

There were, however, in this case as well the «interesting» exceptions, such as:

Mathematics = FAIL, Gymnastics = EXCELLENT and French = PASS

All of the above results were taken by using our own implementations of the corresponding algorithms, which were developed under the project .*Diogenis*. Sponsored by the General Secretariat for Research & Technology, Hellenic Ministry of Development. We need to stress once again that these results were not extracted based on any kind of user intervention. The outcome of this process was purely data-driven, and that is what could be referred to as the beauty of data mining. It was demonstrated that it offers the highest level of support to the user [27].

CONCLUSION

Nowadays, higher educational organizations are placing in a very high competitive environment and are aiming to get more competitive advantages over the other business competitors. They consider students and professors as their main assets and they want to improve their key process indicators by effective and efficient use of their assets. Today the important challenge that higher education faces, is reaching a stage to facilitate the universities in having more efficient, effective and accurate educational processes.

Data mining is considered as the most suited technology appropriate in giving additional insight into the lecturer, student, alumni, manager, and other educational staff behavior and acting as an active automated assistant in helping them for making better decisions on their educational activities like maximizing educational system efficiency, decreasing student's drop-out rate, increasing student's promotion rate, increasing student's retention rate, increasing student’s transition rate, increasing educational improvement ratio, increasing student's success, increasing student's learning outcome, and reducing the cost of system processes.

Data mining technology assists higher educational institution by two following basic fundamental approaches:

- Prediction of trends andbehaviors
- Discovery of previously unknownpatterns

But, in order to apply these approaches we need to understand the limitations of various algorithms. To understand the limitation we need to explore the time and space complexity of the algorithms. For example, can these algorithms be completed in polynomial time? Are there any undesirable problems? If the problems are decidable what is the complexity of theproblems?

As we have seen in k-mode Data mining techniques is used to automatically extract knowledge, in the form of relationships and patterns, from large databases. We presented some practical cases of applying data mining techniques within the education framework. It seems that in whatever teaching or administrative activity, where a data analysis process is needed, data mining techniques can be used instead. Each application area that was previously mentioned constitutes a whole new research domain. In this way we have shown that data mining can be very useful techniques to improve higher education performance and assist students as well teachers to improve theirperformance.

REFERENCES

[1] Mohammed M. Abu Tair, Alaa M. El-Halees, Volume 2 No. 2, February‖ Mining Educational Data to Improve Students' Performance: A Case Study‖,2012.

[2] Schölkopf, B., Smola, A.J., Müller, K.-R.. Nonlinear component analysis as a kernel eigenvalue problem. Neural Computation, 10:1299-1319,1998.

[3] Jain, A., Duin, R., and Mao, J. Statistical Pattern Recognition: A Review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(1):4—37,2000.

[4] Pazzani, M. J., Mani, S., Shankle, W. R.. Acceptance of Rules Generated by Machine Learning among Medical Experts. *Methods of Information in Medicine*, 40:380—385,2001.

[5] Andrew Arnold, Joseph E. Beck, Richard Scheines‖ Feature Discovery in the Context of Educational Data Mining: An Inductive Approach‖,2002.

[6] DeBerard, M. S., Spielmans, G. I and Julka, D. C.. Predictors of academic achievement and retention among college freshmen: A longitudinal study. *College Student Journal* 38, 66-80,2004.

[7] Tinto, V. Dropout from higher education: A theoretical synthesis of recentresearch. *Review of Educational Research* 45, 89-125, 1975.

[8] Bean, J and Eaton, B. The psychology underlying successful retentionpractices. *Journal of College Student Retention: Research, Theory and Practice* 3, 73-89, 2001.

[9] Herzog, S. Estimating student retention and degree-completion time: Decision trees and neural networks vis-a-vis regression. *New Directions for Institutional Research* 131, 17-33,2006.

[10] Hagedorn, L. S. How to define retention. In *College Student Retention: Formula for Student Success*. (Edited by Alan Seidman, 89-106) Praeger Publishers,2005.

[11] Mallinckrodt, B and Sedlacek, W. E.. Student retention and the use of campus facilities by race. *NASPA Journal* 24, 28-32,1987.

[12] Nara, A., Barlow, E and Crisp, G. Student persistence and degree attainment beyond the first year in college: The need for research. In *College Student Retention* (Edited by Alan Seidman) 129-153. Praeger, 2005.

[13] Han, J and Kamber, M. *Data mining: Concepts and techniques* (2nd ed.). Elsevier, 2006.

[14] Larose, D. *Discovering Knowledge in Data: An Introduction to Data Mining*. Wiley-Interscience,2005.

[15] Luan, J. Data mining and its applications in higher education. In *Knowledge Management: Building a Competitive Advantage in Higher Education* (Edited by Serban and J. Luan), 17-36. Josey-Bass,2002.

[16] Yu, C. H. Resampling: A conceptual and procedural introduction. In *Best Practices in Quantitative Methods* (Edited by Jason Osborne), 283-298. Sage Publications,2007.

[17] Larose, D. *Discovering Knowledge in Data: An Introduction to Data Mining*. Wiley-Interscience,2005.

[18] Cuzzocrea, G., Saccardi, A., Lux, G., Porta, E. and Benatti, A. How many good fishes arethereinourNet?NeuralnetworksasadataanalysistoolinCDE-Mondadori’sdata

warehouse. Paper presented at the Annual meeting of SAS User Group International, San Diego, CA, 1997.

[19] Shmueli, G, Patel, N. R. and Bruce, P. *Data Mining for Business Intelligence: Concepts, Techniques and Applications in Microsoft Office Ecel with XLMiner*. Wiley-Interscience,2007.

[20] Streifer, P. A and Schumann, J. A. Using data mining to identify actionable information: breaking new ground in data-driven decision making. *Journal of Education for Students Placed at Risk* 10, 281-293,2005.

[21] Kieseppa, I. A. Statistical model selection criteria and the philosophical problem of underdetermination. *British Journal for the Philosophy of Science* 52, 761-794,2001.

[22] Thompson, B. Stepwise regression and stepwise discriminant analysis need not apply here: A guidelines editorial. *Educational and Psychological Measurement* 55, 525-534, 1995.

[23] Friedman, J. Multivariate adaptive regression splines. *Annals of Statistics* 19, 1-67, 1991.

[24] Hastie, T., Tishirani, R and Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer,2001.

[25] Chou, S. M., Lee, T. S., Shao, Y. E and Chen, I. F. Mining the breast cancer pattern using artificial neural networks and multivariate adaptive regression splines. *Expert Systems with Applications* 27, 133-142,2004.

[26] Usama M. Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, and Ramasamy Uthurusamy,–Advancesin Knowledge Discoveryand Data Mining‖, AAAIPress,1996.

[27] T. Gnardellis , B. Boutsinas‖ On Experimenting with Data Mining in Education‖, 1995.