International Journal of Advance Research in Science and Engineering
Vol. No.6, Issue No. 03 , March 2017
www.ijarse.com
IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

# HIGHER EDUCATION SYSTEM USING DATA MINING METHODS

**Suganya S[1], Dr.V.Narayani[2]**

*Research Scholar, Bharathiar University, Coimbatore (India)*

*Director I/C- MCA, Karpagam College of Engineering, Coimbatore (India)*

## ABSTRACT

*Current years have shown a increasing importance and concern in many countries about quandary of school break down and the fortitude of its main causative factors [1]. The great pact of research [2] has been done on identifying the factors that affect the low concert of students (school failure and dropout) at different enlightening levels (primary, secondary and higher) using the large amount of information that current computers can accumulate in databases. Classifying and find useful information concealed in large databases is a complex task [3]. A very hopeful solution to realize this goal is the use of knowledge discovery in databases techniques or data mining in education, called educational data mining, EDM [4]. This new area of research focuses on the growth of methods to better identify with students and the settings in which they learn [5]. In fact, there are good examples of how to apply EDM techniques to create models that envisage plummeting out and student failure specifically [6]. These facility have shown potential results with respect to those sociological, financial, or didactic characteristics that may be more proper in the prediction of low scholastic performance [7]. It is also important to perceive that most of the research on the relevance of EDM to resolve the problems of student failure and drop-outs has been applied principally to the specific case of higher education [8] and more specifically to online or distance education [9]. However, very little information about specific research on elementary and secondary education has been found, and what has been found uses only statistical methods, not DM techniques [10].*

***Keywords: Classification Algorithms, Educational Institutions, Data Mining, Prediction Methods, Failure Analysis, Writing, Behavioral Science, Classification***

## I INTRODUCTION

A complete primary education is a basic human right and is necessary for enjoying many other rights. It is transformative and empowering and a means for accessing broad economic, social, political and cultural benefits. Education contributes to building more just societies through reducing poverty and inequalities. No country has ever climbed the human development ladder without steady investment in education. Primary education is a powerful driver for the realization of all the Millennium Development Goals (MDGs) and for sustainable development more

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

broadly. Universal primary education involves entering school at an appropriate age, progressing through the system and completing a full cycle.Today, there are over 30 million more children in school than in the beginning of the decade. There have been some remarkable success storiesThe School education structure in Tamil Nadu has four levels namely Primary, Upper Primary, Secondary and Higher Secondary. At operational level considerable variations are found in the patterns of schooling across the different states of India, at the elementary level. Several states follow patterns in which elementary schooling consistsof Seven Years, divided in to four years of primary followed by three years of upper primary. Student dropout prediction is an indispensable for numerous intelligent systems to measure the education system and success rate of any university as well as throughout the university in the world. Therefore, it becomes essential to develop efficient methods for prediction of the students at risk of dropping out, enabling the adoption of proactive process to minimize the situation. Thus, this research work propose a prototype machine learning tool which can automatically recognize whether the student will continue their study or drop their study using classification technique based on decision tree and extract hidden information from large data about what factors are responsible for dropout student. Further the contribution of factors responsible for dropoutrisk was studied using discriminant analysis and to extract interesting correlations, frequent patterns, associations or casual structures among significant datasets, Association rule mining was applied.

## II THE STATE OF EDUCATION IN INDIA

The money, time and focus it will take to get our primary education working and to create globally competitive institutions is no small matter. Since the Industrial Revolution, in India and around the world, the tradition of home or community schooling often centered on the teacher has gradually been transformed into a human supply chain schooling system centered on the educational institution. [11].As we move towards an information age, nations around the world are grappling with what the next transformation in education needs to be. The current challenge in India remains a 20th century challenge of quantity and quality for its primary and higher education systems. Private schools perform better than government schools, though there is substantial room for improvement. There is a significant variation among states in both quantity and quality.

## III WHAT OF THE HIGHER EDUCATION SYSTEM?

There is no equivalent independent and rigorous survey of higher education quality in India[11]. The department of higher education in the human resources development ministry puts out an annual survey—which is numerical and supply-side—called the All India Survey on Higher Education (AISHE). According to the latest survey results, there are 38,000 colleges that make up 767 universities in India. Approximately 33 million students are enrolled in these universities with 1.4 million teachers. The gross enrolment ratio, the percentage of population 18-23 enrolled in college, is 23.6% for India and varies widely from 12% for Bihar to 44% for Tamil Nadu. For higher education, the usual tosh about achieving access, equality, justice, quality, inclusiveness and employability, all the same time, is simply not possible. Excellent institutions cannot by definition be “equal”. There will have to be a two-pronged

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

approach—universal universities and stellar universities. For universality, there will have to be easy on and off ramps, building block education and portable academic credit, affordability, ubiquity, employability and signaling value. The money, time and focus it will take to get our primary education working and to create globally competitive institutions is no small matter. India will need to accelerate that process now so that when fundamental innovation becomes necessary as an engine of growth in the coming decades, India has at least a handful of institutions that can contribute.

## IV DATA MIINING

Data mining is a means of automating part of process to detect interpretable patterns; it helps us see the forest without getting lost in the trees.Discovering information from data takes two major forms: description and prediction. At the scale we are talking about, it is hard to know what the data shows. Data mining is used to simplify and summarize the data in a manner that we can understand, and then allow us to infer things about specific cases based on the patterns we have observed. Of course, specific applications of data mining methods are limited by the data and computing power available, and are tailored for specific needs and goals. However, there are several main types of pattern detection that are commonly used. Data mining, in this way, can grant immense inferential power. If an algorithm can correctly classify a case into known category based on limited data, it is possible to estimate a wide-range of other information about that case based on the properties of all the other cases in that category. This may sound dry, but it is how most successful Internet companies make their money and from where they draw their power.

Anomaly detection : in a large data set it is possible to get a picture of what the data tends to look like in a typical case. Statistics can be used to determine if something is notably different from this pattern. For instance, the IRS could model typical tax returns and use anomaly detection to identify specific returns that differ from this for review and audit.

Association learning: This is the type of data mining that drives the Amazon recommendation system. For instance, this might reveal that customers who bought a cocktail shaker and a cocktail recipe book also often buy martini glasses. These types of findings are often used for targeting coupons/deals or advertising. Similarly, this form of data mining (albeit a quite complex version) is behind Netflix movie recommendations.

Cluster detection: one type of pattern recognition that is particularly useful is recognizing distinct clusters or sub-categories within the data. Without data mining, an analyst would have to look at the data and decide on a set of categories which they believe captures the relevant distinctions between apparent groups in the data. This would risk missing important categories. With data mining it is possible to let the data itself determine the groups. This is one of the black-box type of algorithms that are hard to understand. But in a simple example - again with purchasing behavior - we can imagine that the purchasing habits of different hobbyists would look quite different from each

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

other: gardeners, fishermen and model airplane enthusiasts would all be quite distinct. Machine learning algorithms can detect all of the different subgroups within a dataset that differ significantly from each other.

Classification: If an existing structure is already known, data mining can be used to classify new cases into these pre-determined categories. Learning from a large set of pre-classified examples, algorithms can detect persistent systemic differences between items in each group and apply these rules to new classification problems. Spam filters are a great example of this - large sets of emails that have been identified as spam have enabled filters to notice differences in word usage between legitimate and spam messages, and classify incoming messages according to these rules with a high degree of accuracy.

Regression: Data mining can be used to construct predictive models based on many variables. Facebook, for example, might be interested in predicting future engagement for a user based on past behavior. Factors like the amount of personal information shared, number of photos tagged, friend requests initiated or accepted, comments, likes etc. could all be included in such a model. Over time, this model could be honed to include or weight things differently as Facebook compares how the predictions differ from observed behavior. Ultimately these findings could be used to guide design in order to encourage more of the behaviors that seem to lead to increased engagement over time.

## V EDUCATIONAL DATA MINING

Educational Data Mining is a promising regulation, concerned with developing methods for exploring the unique types of data that come from educational settings, and using those methods to better understand students, and the settings which they learn in. Whether educational data is taken from students' use of interactive learning environments, computer-supported collaborative learning, or administrative data from schools and universities, it often has multiple levels of meaningful hierarchy, which often need to be determined by properties in the data itself, rather than in advance. Issues of time, sequence, and context also play important roles in the study of educational data." (Educational Data Mining Society home page, retrieved Jan 17, 2014)

"Educational Data Mining is an emerging discipline, concerned with developing methods for exploring the unique types of data that come from educational settings, and using those methods to better understand students, and the settings in which they learn." (JEDM - Journal of Educational Data Mining, retrieved Jan 17, 2014) .Educational data mining is rooted in general data mining. However, there are specifics: "Data mining, also called Knowledge Discovery in Databases (KDD), is the field of discovering novel and potentially useful information from large amounts of data [Witten and Frank 1999]. It has been proposed that educational data mining methods are often different from standard data mining methods, due to the need to explicitly account for (and the opportunities to exploit) the multi-level hierarchy and non-independence in educational data [Baker in press].

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

## VI DATA MINING IN HIGHER EDUCATION

Data mining is a powerful tool for academic intervention. Through data mining, a university could, for example, predict with 85 percent accuracy which students will or will not graduate.The university could use this information to concentrate academic assistance on those students most at risk. In order to understand how and why data mining works, it's important to understand a few fundamental concepts. First,data mining relies on four essential methods: Classification, categorization, estimation, and visualization. Classification identifies associations and clusters, and separates subjects under study. Categorization uses rule induction algorithms to handle categorical outcomes, such as "persist" or "dropout," and "transfer" or "stay." Estimation includes predictive functions or likelihood and deals with continuous outcome variables, such as GPA and salary level. Visualization uses interactive graphs to demonstrate mathematically induced rules and scores, and is far more sophisticated than pie or bar charts. Visualization is used primarily to depict three-dimensional geographic locations of mathematical coordinates.Higher education institutions can use classification, for example, for a comprehensive analysis of student characteristics,or use estimation to predict the likelihood of a variety of outcomes, such as transferability, persistence, retention, and course success.

## VII APPLICATIONS OF DATA MINING IN HIGHER EDUCATION

There are many application areas of data mining like customer analytics, Agriculture, banking, Security Applications, Educational data mining, Mass surveillance, Privacy preserving etc. The main concerned area is about data mining applications in educational systems.Educational Data Mining (EDM) is an emerging discipline, concerned with developing methods for exploring the unique types of data that come from educational settings, and using those methods to better understand students, and the settings which they learn in.[A key area of EDM is mining student's performance. Another key area is mining enrollment data. Key uses of EDM [10] include predicting student performance and studying learning in order to recommend improvements to current educational practice. EDM can be considered one of the learning sciences, as well as an area of data mining. The main applications of EDM are listed as follows

### Analysis and Visualization of Data

It is used to highlight useful information and support decision making. In the educational environment, for example, it can help educators and course administrators to analyze the students' course activities and usage information to get a general view of a student's learning. Statistics and visualization information are the two main techniques that have been most widely used for this task. Statistics is a mathematical science concerning the collection, analysis, interpretation or explanation, and presentation of data. It is relatively easy to get basic descriptive statistics from statistical software, such as SPSS. Statistical analysis of educational data (logs files/databases) can tell us things such as where students enter and exit, the most popular pages students browse, number of downloads of e-learning

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

resources, number of different pages browsed and total time for browsing different pages. It also provides knowledge about usage summaries and reports on weekly and monthly user trends, amount of material students might go through and the order in which students study topics, patterns of studying activity, timing and sequencing of events, and the content analysis of students notes and summaries. Statistical analysis is also very useful to obtain reports assessing how many minutes student worked, number of problems he resolved and his correct percentage along with our prediction about his score and performance level.Visualization uses graphic techniques to help people to understand and analyze data. There are several studies oriented toward visualizing different educational data such as patterns of annual, seasonal, daily and hourly user behavior on online forums. Some of such investigations are statistical graphs to analyze assignments complement, questions admitted, exam score, student tracking data to analyze student's attendance, results on assignments and quizzes, weekly information regarding students and group's activities.

### Predicting Student Performance

In this case, we estimate the unknown value of a variable that describes the student. In education, the values normally predicted are student's performance, their knowledge, score, or marks. This value can be numerical/continuous (regression task) or categorical/discrete (classification task). Regression analysis is used to find relation between a dependent variable and one or more independent variables. Classification is used to group individual items based upon quantitative characteristics inherent in the items or on training set of previously labeled items. Prediction of a student's performance is the most popular applications of DM in education. Different techniques and models are applied like neural networks, Bayesian networks, rule- based systems, regression, and correlation analysis to analyze educational data.

### Outlier Analysis

According to Grubbs [13] Outlier can be defined as "An outlying observation, or outlier, is one that appears to deviate markedly from other members of the sample in which it occurs". Outlier detection has been used to detect and, where appropriate, remove anomalous observations from data. Outlier detection can identify system faults and fraud before they escalate with potentially catastrophic consequences.There are three fundamental approaches for outlier detection.

Type 1 - Determine the outliers with no prior knowledge of the data. This is essentially a learning approach analogous to unsupervised clustering. The approach processes the data as a static distribution, pinpoints the most remote points, and flags them as potential outliers.

Type 2 - Model both normality and abnormality. This approach is analogous to supervised classification and requires pre-labeled data, tagged as normal or abnormal.

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

Type 3 - Model only normality (or in a few cases model abnormality). This is analogous to a semi- supervised recognition or detection task. It may be considered semi-supervised as the normal class is taught but the algorithm learns to recognize abnormality.

## VIII CONCLUSION AND FUTURE SCOPE

Data analysis plays an imperative role for any type of pronouncement support irrespective of type of industry. Data warehousing and data mining methods for data analysis are explained in detail. This paper is to review role of data mining techniques in education system. Educational Data Mining has been introduced as an upcoming research area, Thus number of specific tools specially developed for applying DM algorithms in educational data/environments are emerging day by day.DM techniques in educational organizations help us to learn student performance, student behavior, carefully designing course curriculum, to motivate students and to group student depending upon various parameters. One possible solution is the development of tools that use a default algorithm for each task and parameter-free DM algorithms to simplify the configuration and execution for non-expert users. Secondly, the DM tool has to be integrated into the e-learning environment so that results obtained with DM techniques could be easily and directly applied. Moreover, current tools for mining data pertaining to a specific course/framework may be useful to their developers only. There are no general tools or reusing tools that can be applied to any educational system. Therefore, a standardization of input data and output model are needed.

## REFERENCES

**[1]**. L. A. A. Aldaco, "Comportamiento de la deserción y reprobación en el colegio de bachilleres del estado de baja california: Caso plantel ensenada", *Proc. 10th Congr. Nat. Invest. Educ.*, pp. 1-12, 2009.

**[2]**. F. Araque, C. Roldán, A. Salguero, "Factors influencing university drop out rates", *Comput. Educ.*, vol. 53, no. 3, pp. 563-574, 2009.

**[3]**. M. N. Quadril, N. V. Kalyankar, "Drop out feature of student data for academic performance using decision tree techniques", *Global J. Comput. Sci. Technol.*, vol. 10, pp. 2-5, Feb. 2010.

**[4]**. C. Romero, S. Ventura, "Educational data mining: A survey from 1995 to 2005", *Expert Syst. Appl.*, vol. 33, no. 1, pp. 135-146, 2007.

**[5]**. C. Romero, S. Ventura, "Educational data mining: A review of the state of the art", *IEEE Trans. Syst. Man Cybern. C Appl. Rev.*, vol. 40, no. 6, pp. 601-618, Nov. 2010.

**[6]**. S. Kotsiantis, K. Patriarcheas, M. Xenos, "A combinational incremental ensemble of classifiers as a technique for predicting students' performance in distance education", *Knowl. Based Syst.*, vol. 23, no. 6, pp. 529-535, Aug. 2010.

**[7]**. J. Más-Estellés, R. Alcover-Arándiga, A. Dapena-Janeiro, A. Valderruten-Vidal, R. Satorre-Cuerda, F. Llopis-Pascual, T. Rojo-Guillén, R. Mayo-Gual, M. Bermejo-Llopis, J. Gutiérrez-Serrano, J. García-Almiñana, E. Tovar-Caro, E. Menasalvas-Ruiz, "Rendimiento académico de los estudios de informática en algunos centros españoles", *Proc. 15th Jornadas Enseñanza Univ. Inf. Barcelona Rep. Conf.*, pp. 5-12, 2009.

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

**[8]**. S. Kotsiantis, "Educational data mining: A case study for predicting dropout—prone students", *Int. J. Know. Eng. Soft Data Paradigms*, vol. 1, no. 2, pp. 101-111, 2009.

**[9]**. I. Lykourentzou, I. Giannoukos, V. Nikolopoulos, G. Mpardis, V. Loumos, "Dropout prediction in e-learning courses through the combination of machine learning techniques", *Comput. Educ.*, vol. 53, no. 3, pp. 950-965, 2009.

**[10]**. A. Parker, "A study of variables that predict dropout from distance education", *Int. J. Educ. Technol.*, vol. 1, no. 2, pp. 1-11, 1999.

[11]. http://www.livemint.com/Opinion/PQmwXBdmg6FWPlPVPtmV7K/

[12]. https://www.theatlantic.com/technology/archive/2012/04

[13].F. E Grubbs, “Procedures for detecting outlying observations in samples,” pp.1–21, Technometrics 11.