

## DATA MINING APPROACH TO STUDENT RETENTION

Ghazi Mohammed Zafaruddin<sup>1</sup> and Heena Jadhav<sup>2</sup>

<sup>1</sup>Assistant Professor, Millennium Institute of Management, Aurangabad, India  
Email: zafar.mohammed@outlook.com

<sup>2</sup>Assistant Professor, Shivchhatrapati College, Aurangabad, India  
Email: jadhavh.06jan@gmail.com

### ABSTRACT

*India's higher education system is the second largest in the world, after the United States. Our education system is often cited as one of the main contributors to the economic rise of India. Student retention still remains a big challenge to many institutions at large. This problem can be effectively addressed through simple changes to some basic academic procedures with the help of emerging and latest developments in the field of computer science and information technology. Data mining technologies can be used to monitor students and simultaneously analyzing their academic behavior, thus providing a basis for implementing necessary intervention procedures, if required. This paper discusses how data mining technologies can be used to help fish out risky students and evaluate them so as to implement tailor-made intervention strategies and reduce dropout rate in higher education.*

**Keywords:** Data Mining, Student Retention, KDD (Knowledge Discovery in Databases)

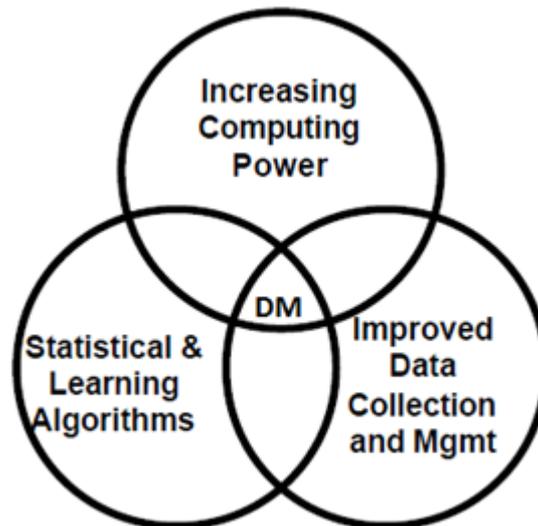
### INTRODUCTION

As the cost of processing power and storage is falling, data storage became easier and cheaper. Universities are facing the immense and quick growth of the volume of educational data [2].

With the enormous amount of data stored in files, databases, and other repositories, it is increasingly important, if not necessary, to develop powerful means for analysis and perhaps interpretation of such data and for the extraction of interesting knowledge that could help in decision-making. Data Mining, also popularly known as Knowledge Discovery in Databases (KDD), refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases. While data mining and knowledge discovery in databases (or KDD) are frequently treated as synonyms, data mining is actually part of the knowledge discovery process [3]. The actual data mining task is the automatic or semi-automatic analysis of large quantities of data to extract previously unknown interesting patterns such as groups of data records (cluster analysis), unusual records (anomaly detection) and dependencies (association rule mining). This usually involves using database techniques such as spatial indices. These patterns can then be seen as a kind of summary of the input data, and may be used in further analysis or, for example, in machine learning and predictive analytics. For example, the data mining step might identify multiple groups in the data, which can then be used to obtain more accurate prediction results by a decision support system. Neither, the data collection and data preparation nor result interpretation and

reporting are part of the data mining step, but do belong to the overall KDD process as additional steps [4].

Data mining is basically convergence of three key technologies [1]. The figure below depicts the same:



**Fig 1.** Data Mining-Convergence of Three Key Technologies

The kinds of patterns that can be discovered depend upon the data mining tasks employed. By and large, there are two types of data mining tasks: descriptive data mining tasks that describe the general properties of the existing data, and predictive data mining tasks that attempt to do predictions based on inference on available data [3].

Student retention is an indicator of academic performance and enrolment management of the university. Poor student retention could reflect badly on the university, and cause serious financial strains [2].

Educational data mining (EDM) is an area full of exciting opportunities for researchers and practitioners. This field assists higher educational institutions with efficient ways to improve institutional effectiveness and student learning. Data mining is a significant tool for helping organizations enhance decision making and analyzing new patterns and relationships among a large amount of data [9].

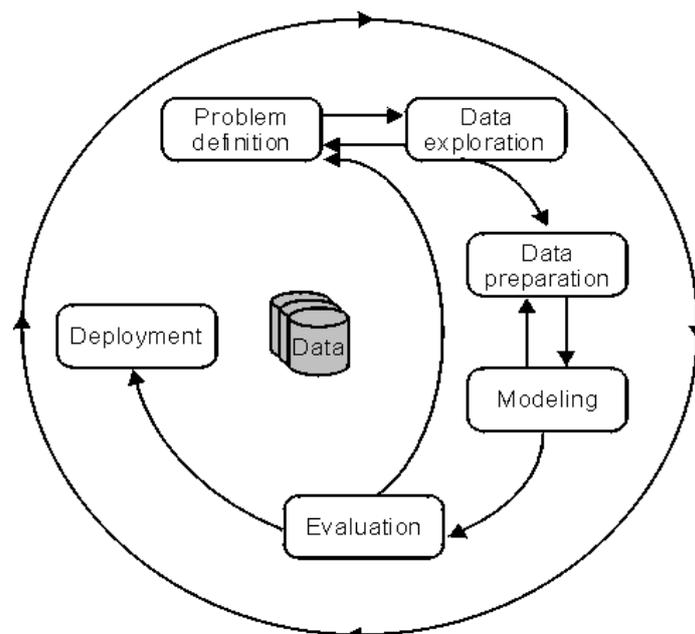
### **Student Retention**

One of the biggest challenges that higher education faces is to improve student retention. In general, more students remaining in the university means better academic programs and higher revenue [2].

University student retention, sometimes referred to as persistence, is of increasing importance to college administrators as they try to improve graduate rates and decrease loss of tuition revenue from students that either drop out or transfer to another school. Universities are now creating a number of new programs for students that help keep them engaged in their classes and involved on campus. This includes campus funded tutoring, freshman seminar courses, and intramural sports among many other things. These programs

are important when it comes to campus life because it has been shown that student involvement is directly related to student success. When a student participates, he or she forms both social and emotional ties to the university that both encourage the student to do well academically and reduce the chance that the student will drop out of school entirely or leave for another university [5].

Data mining uses algorithms with predictive capabilities which can be used to find patterns and correlations in underlying data sets. Data mining is already widely used across the private sector and customer churn analysis is the activity which is most closely related to the problem of student retention, i.e. identification of customers who are at risk of leaving the company. This is important since the cost of retaining a customer is far less than acquiring a new one.



**Fig 2.** Steps to Student Data Mining

On the basis of the above model a system can be designed to help fish out the risky students and provide proper guidance to them, thus reducing student drop out ratio in higher education. The initial steps have already been implemented as the problem has been defined and the data is ready with a vast number of universities and institutions storing and maintaining the student related data in their databases.

The data needs to be cleansed and prepared to be used for analysis and modeling. On the basis of cleansed data some future predications can be made on the basis of the past experiences and outcomes.

### **Data Mining and Student Retention**

Colleges and universities squeezed by budget cuts stabilize revenue streams by improving student retention. The surprising connections these institutions uncovered through data mining and predictive models have informed improvements to student retention strategies [6]:

**South Texas College**

Through data mining and analysis, administrators at South Texas College discovered that students who register late for a course are more likely to fail or withdraw. Realizing that this has a negative impact on student retention and time to completion, South Texas decided to eliminate late registration [6].

**University of Alabama**

Students in a graduate course built a predictive model of student retention using SAS software. One of the group's findings was that commuter students are more likely to drop out of the university. Consequently, the university developed student retention strategies including requiring all freshmen to live on-campus and recruiting at-risk students to participate in specialized seminars and other programs [6].

**State University of New York at Buffalo**

The engineering school at SUNY-Buffalo rates incoming students on seven academic factors related to their preparation and test scores. If a student is substandard on five of these factors, then the school advises the student to enroll in specialized tutoring sessions for entry-level courses [6].

**Tiffin University**

Through predictive modeling, Tiffin found that academic, financial, and social factors all affected retention risk. The university's new student retention strategies include assigning high-risk students to personal mentors, designating a chief retention office to monitor students in the medium-risk group and sending an automated email message to students in the low-risk group to inform them of relevant extracurricular activities. Through these efforts, Tiffin improved its student retention strategies: Tiffin's one-year student retention rate increased from 51 percent to 63 percent in just five years [6].

**CONCLUSION**

Student retention is an important issue for all university policy makers due to the potential negative impact on the image of the university and the career path of the dropouts. As discussed in this paper data mining and analysis has helped many institutions improve student retention by making basic alteration to some of the basic academic procedures right from admission to transfer. The different colleges and universities already have a huge database of students and their performance which can be better utilized by researchers to analyze and predict. The latest development in the field of information technology will immensely boast this process and what is better than data warehousing and data mining when it comes to finding, analyzing and predicting.

Use of data mining would provide a university with a large number of benefits like understanding of the factors which influence student retention, understand how data which changes over time may influence a student's risk of withdrawal, generate a prediction of the risk associated that an individual student will withdraw from their course, enable them to intervene earlier with high risk students, design and implement appropriate intervention programmes, etc. Thus, data mining technologies can be used to help fish out risky students and evaluate them so as to implement tailor-made intervention strategies and reduce dropout rate in higher education.

**REFERENCES**

1. Thearling K. (2013), “An Introduction to Data Mining”, available online at [www.thearling.com](http://www.thearling.com)(accessed on 10th February 2013)
2. Zhang Y., Oussena S., Clark T., Kim H. (2013), “Use Data Mining to Improve Student Retention in Higher Education – A Case Study”, available online at [hssc.sla.mdx.ac.uk/.../%5BCam](http://hssc.sla.mdx.ac.uk/.../%5BCam)(accessed on 12th February 2013)
3. Zaiane O. R. (2013), “Introduction to Data Mining”, CMPUT690 Principles of Knowledge Discovery in Databases, Department of Computer Science, University of Alberta, available online at [www.cs.ualberta.ca/~zaiane/courses/cmput690/notes/.../ch1.pdf](http://www.cs.ualberta.ca/~zaiane/courses/cmput690/notes/.../ch1.pdf) (accessed on 15th February 2013)
4. Wikipedia (2013), “Data Mining”, From Wikipedia, The Free Encyclopedia available online at [http://en.wikipedia.org/wiki/Data\\_mining](http://en.wikipedia.org/wiki/Data_mining) (accessed on 15th February 2013)
5. Wikipedia (2013), “University Student Retention”, From Wikipedia, The Free Encyclopedia available online at [http://en.wikipedia.org/wiki/University\\_student\\_retention](http://en.wikipedia.org/wiki/University_student_retention) (accessed on 15th February 2013)
6. Higher Education Blog, “How Data Mining Helped 11 Universities Improve Student Retention Strategies”, available online at <http://www.hanoverresearch.com/2012/01/how-11-universities-will-improve-student-retention/> (accessed on 18th February 2013)
7. Chong Ho Yu et al. (2010), “A Data Mining Approach for Identifying Predictors of Student from Sophomore to Junior Year”, *Journal of Data Science* 8(2010), 307-325.
8. Eitel J. M. Lauria et al. (2012), “Mining Academic Data to Improve College Student Retention: An Open Source Perspective”, *Research in Progress* available online at <https://confluence.sakaiproject.org/.../LAK2012-CameraReady.pdf> (accessed on 19th February 2013)
9. Huebner R. A. (2013), “A Survey of Educational Data-Mining Research”, *Research In Higher Education Journal*, available online at [www.aabri.com/manuscripts/121328.pdf](http://www.aabri.com/manuscripts/121328.pdf) (accessed on 20<sup>th</sup> February 2013)