Critical Appraisal of Data Mining as an Approach to Improve Student Retention Rate

Leila Dadkhahan, Mohamad Anas Al Azmeh

Abstract— one of the main challenges and goals of today higher education institutions is increasing the student retention rate by providing high quality education. This can be achieved through discovering knowledge from the vast amount of historical data in institutions information systems through applying data mining techniques. Data mining techniques can be used to identify students at risks and predict their academic performance. This would help to provide appropriate intervention strategies to those students as early as possible. This paper i) justify the importance of high student retention rate to higher education institutions, ii) presents theoretical student retention models, and iii) critically appraisal the previous related works to show how data mining techniques can be applied to discover new knowledge in order to increase retention rate.

Index Terms—Data Mining Techniques, Retention Model, Student Retention Rate.

I. INTRODUCTION

In many Higher Education (HE) institutions, increasing the student retention rate is becoming an important long term goal and challenging tasks [4],[3]. It has become an indication of academic performance and enrolment management. The outcome of student retention is significant for both students and institutions. For universities, low student retention rate has potential negative impact on the institution’s reputation while a higher rate may lead to better academic programme and higher revenue [5],[3].

To increase student retention rate, institutions need to establish the educational conditions and perspectives which can be applied to all kind of students rather than some groups. They have to identify the factors which have impact on student non-retention with institution, situations they face, and character of the educational settings in which students learn. [4]

Data Mining (DM) techniques can be applied to higher education institutions Information Systems (IS) as an approach to improve student retention rate. The higher education institutions information systems consist of vast amount of novel and historical educational data. DM techniques help to gain personalized educational knowledge which help institutions to enhance decision making, maximize educational system efficiency, and reduce the cost of education processes. All these benefits lead to increase student’s retention rate and student’s learning outcome. [3], [17]

The aim of this research is to evaluate the importance of retention rate and critically appraised the DM techniques to improve the retention rate. This report presents how DM techniques were applied to collected data to identify factors affecting student retention in order to predict students at risk as early as possible. Moreover, this report would justify how these techniques helped to tailor the interventions strategies to increase student retention.

II. STUDENT RETENTION AND ITS IMPORTANCE IN HIGHER EDUCATION INSTITUTIONS

Student retention is becoming important for higher education institutions to manage enrolment, administration and academic performance. Institutions need to improve their student graduation rate to increase tuition revenue. This could be gained by preventing students from dropping out or transferring to another institution by applying proper retention strategies at the right time. In UK, ‘retention’ refers to students remaining in one HE institution and completing their programme of study within a specific timeframe. [1]

Institutions might use other indicators to track retention. Following is the different definitions of the most commonly used variables as retention indicators: [2]

- Attrition: a student who failed to enrol at the institution in the following semester.
- Dismissal: the institution does not allow a student to enrol.
- Drop-out: a student who did not complete their education but their initial educational goal was to complete the degree.
- Mortality: the failure of students being retained in the institution until graduation.
- Persistence: the desire and action of a student to continue education in institution from the first year to the completion of a degree.
- Retention: the ability of institution to retain a student from admission to graduation.
- Stop-out: s student who temporarily withdraws from institution.
- Withdrawal: departure of a student from the institution campus.

In this paper, retention rate is defined as data from the first-time, fulltime freshman students persisting with enrolling and completing the degree within given time frame in the same institution.

A report by American College Testing (ACT) illustrates that in 2012, the first-to-second year retention rate in public institution for a five year degree course is 65.2 percentages. On the other hand, the degree completion rate for same programme and institution type is only at 36.6 percent. [9]
In 2013, The University of Kentucky reported statistics for first-year full-time students’ performance over six years. The report shows that 4,118 students enrolled in 2006 for five-year undergraduate degree. 89.5 percent of these students persist to second semester, and 76.4 percent persist to sophomore year. This is while 30.4 percent completed their five years degree and graduated. [10]

A study showed that about one fourth of students dropped out of college after their first year [4]. As these reports also justify, first year of education is critical in order to increase the retention rate. The students who continued their sophomore year are more likely to graduate; however, data also shows many students may still drop out after completing their first year. [5]

There are many factors that have negative impact on the rate and institutions must consider as part of their responsibilities to prevent them. The tasks such as developing transition programme focusing on generic study skills, peer mentoring, and residential experiences, will help to improve university retention, progression and completion rates.[8] [7]

III. STUDENT RETENTION THEORETICAL MODEL

The major retention models are explained in this section. These models would help to get perspectives of required factors influencing student retention.

Astin introduced a model which claims student involvement in the institution has direct relation to student retention. The model is called input-environment-output (I-E-O). Inputs are related to characteristics of a student at the time they first enrolled at institution. Environment is learning settings such as programmes, faculty and educational experiences where student learns. Outcomes are related to characteristic of student after they are placed into the environment settings [11].

Tinto’s model is the most widely accepted model. The model focuses on predicting the student retention by identifying the level of student’s academic and social integration within institution. According to Tinto, these factors have strong effect on students’ decision to drop out from enrolled institution. [4] [11]

Bean developed a model which focuses on psychological and behavioural factors of student integration. The psychological processes are very difficult to measure [5]. The model is based on the student’s intention. The intention is assessed by attitudes and behaviour that have an impact on level of the student integration. [11]

Thomas developed a model known as ‘Institutional Habitus’ based on Tinto’s theory. The model divides factors into academic and social integration. The academic factor could be staff attitudes, teaching and learning assessment, and learning styles. The social factor is about friendship, mutual support and social networks.

Seidman introduced a formula for student retention:

Student Retention = Early Identification + (Early + Intensive + Continuous) Intervention

As the formula shows, early identification of students at risk and providing early proper intensive continuous intervention to them would result in increasing student retention.

Following table shows examples related to mentioned factors affecting the student retention rate:

<p>| TABLE I. influence factors on retention [21] |</p>
<table>
<thead>
<tr>
<th>Factors Influencing Retention</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Student Level</strong></td>
<td></td>
</tr>
<tr>
<td>Academic Performance</td>
<td>Institution GPA and academic Performance, High School GPA, course Load and credits earned</td>
</tr>
<tr>
<td>Attitudes and Satisfaction</td>
<td>Commitment to college, positive attitude about academics, sense of social integration with institution</td>
</tr>
<tr>
<td><strong>Institutional Level</strong></td>
<td></td>
</tr>
<tr>
<td>Academic Engagement</td>
<td>Research activities, the size of institution, clubs</td>
</tr>
<tr>
<td><strong>Social and External Level</strong></td>
<td></td>
</tr>
<tr>
<td>Social and Family support</td>
<td>Faculty and staff support, family support, sense of belonging and community</td>
</tr>
</tbody>
</table>

Through these models, it could be realised that academic and social integration would have major influence on retention. Thus, identifying the attributes related to level of these factors would help to predict student academic performance and identify those who are at risk of non-retention. These result in enhancing intervention programme and apply them to those students at risk as early as possible.

IV. DATA MINING AS AN APPROACH TO INCREASE STUDENT RETENTION RATE

In order to achieve high retention rate, institutions must utilize their resources effectively. They must identify what study settings should look like, what conditions would promote student retention, and predict how and when they should apply it, and to whom [4].

Data Mining (DM) techniques can be applied on higher education institution’s Information System as an approach to find the answer to all these questions. The institution’s information systems consist of vast amount of educational data. This large repository is growing rapidly and includes useful information, from student’s academic results, background, financial and demographic to module descriptions and learning outcomes. It has novel and historical data of previous students as well as features which can be related to the current and prospective students with similar attributes. [18]

DM which is known as Knowledge Discovery in Databases (KDD) has techniques that help to find relationships and existing, but hidden, patterns among the huge amount of data.
Data mining has been widely used for companies and their customer analysis to identify those who are at risk of leaving the company. This is closely related to the problem of student retention. Data mining techniques can be used to monitor students and simultaneously analyse their academic performance. Through Data mining’s clustering and predicting approach, predictive models can be generated from previous students. This model helps to predict each single student’s behaviour and academic performance and identify those more likely to drop out of the university. These would provide great opportunities for universities to act before students dropped out and to plan a proper retention strategy.

V. RELATED DATA MINING APPLICATIONS IN HIGHER EDUCATION

In this section, the literature review of data mining works is presented. These works have been done based on the student academic and social integration and aimed to improve the student retention rate in Higher Education Institutions by predicting student at risk in the early stage.

1) Luan used clustering techniques to specify the needs of various groups of students. This application could identify and group these students who could pile up their subjects and take those subjects for longer duration. The developed work resulted in enhancing strategies of offering subjects and curriculum, credit hours and so on.

2) Minaei-Bidgoli used classification and prediction techniques to predict students’ final grades and identify those at risk. This could help institution to provide effective academic advice to the students in the right time.

3) Superby used classification and prediction techniques and divided students into three categories: ‘low-risk’ those with high probability of degree completion, ‘medium-risk’ those who might complete the degree by applying proper academic advice and ‘high-risk’ those who have high probability of non-retention. The techniques tried to identify the level of risk for the first year.

4) Delavari applied classification and prediction as main techniques to educational data. Through these techniques, she could discover pattern of successful student in a subject as well as student success rate for individual lecturer and predict the rate. The knowledge gained from these models could help institution for better decision making in setting up new strategies or improving the current strategies to increase the student success rate.

5) Thames Valley University developed the Mining Course Management Systems (MCMS) project to improve student retention strategies by analysing student behaviour and early identification of those at risk. In this project, association, clustering, classification and prediction techniques were applied on collected data from current university information systems (library, student administration, online learning system, online resource system, online test system and so on) and integrated into data warehouse. The project could build models to predict each single student performance and their behaviour.

6) Kovacic applied classification and feature selection techniques to identify the influencing factors at the enrolment stage. The data was collected from the Student Management System of open Polytechnic of New Zealand. The research showed that identification of students at risk even before they start their study could help the university to take proper actions to improve the academic success.

7) Quadril and Kalyankar applied classification and prediction techniques. These techniques helped them to build student drop out model and predict each student’s academic performance by measuring student’s Cumulative Grade Point Average (CGPA).

8) Kabakchieva applied classification techniques on the data collected from University of National and World Economy (UNWE) system. The work focused on predicting students’ performance at enrolment stage by finding patterns from data related to students personal and pre-university information.

Each mentioned work has its own characteristics and approach to predict student performance and identify those at risk at early stage. This literature justifies that data mining is becoming popular in almost all applications of real world. It also shows Classification and Prediction techniques of Data Mining are the interesting topics to the researchers. These techniques helped to accurately classify the data for knowledge discovery and predict the risks. As a result, institutions became able to build proper intervention programme, therefore, student’s academic performance improved and institution student retention and graduation rates increased.

Following table summarises all the mentioned works to show the used techniques as well as the built data set and model. It provides the achieved results and findings of each work. It also justifies the benefits of Data Mining techniques to higher education institutions.
### TABLE II. overview of the works, algorithm used, design models and data sets, and results

<table>
<thead>
<tr>
<th>ID</th>
<th>Algorithm</th>
<th>Data Set</th>
<th>Model</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Neural Network, C5.0, C&amp;RT</td>
<td>Student personal and academic background</td>
<td>Persistence non-persistence</td>
<td>Neural Network resulted in best accuracy of 61.3%.</td>
</tr>
<tr>
<td>2</td>
<td>C5.0, CART, QUEST, CRUISE, Bayes, 1NN, kNN, Parzen, MLP, CMC</td>
<td>student database: Educational resources, User Information Activity Log</td>
<td>High, Middle, low</td>
<td>Pass, fail; kNN has the highest accuracy of 82.3% accuracy, CMC achieved 86.8% accuracy.</td>
</tr>
<tr>
<td>3</td>
<td>ID3, CART, Neural Networks, Linear discriminant analysis</td>
<td>Student personal and scholastic characteristic and background</td>
<td>Risk level (low-risk, medium-risk, high risk)</td>
<td>Student scholastic and socio-familial background have the highest factor on risk while are not related to success.</td>
</tr>
<tr>
<td>4</td>
<td>SLIQ, Neural Network, RBF</td>
<td>Lecturer demographic and academic knowledge</td>
<td>Success rate</td>
<td>Student: decision tree has the accuracy of 86.04% and neural network has the accuracy of 86.8%.</td>
</tr>
<tr>
<td>5</td>
<td>Naive Bayes, Support Vector Machine, Decision Tree</td>
<td>Student personal background</td>
<td>Drops out intervention</td>
<td>Student drop is more related to their academic activities rather than their background profile.</td>
</tr>
<tr>
<td>6</td>
<td>CART, CHAID</td>
<td>Student personal background and demography</td>
<td>Success (pass, fail)</td>
<td>CART accuracy of 60.5%, CHAID accuracy of 59.5%.</td>
</tr>
<tr>
<td>7</td>
<td>C4.5 (J48)</td>
<td>Student personal background</td>
<td>Drop out (yes, no)</td>
<td>Classification is not accurate using enrolment data only.</td>
</tr>
</tbody>
</table>

| 8  | C4.5 (J48), NaiveBaye, BayesNet, IBk, OneR, JRip | Student personal background, Pre-University characteristic, Students in institution | Excellent, very good, good, average, bad | Predication rate was low as 52%-67% University Admission Score and Number of Failures at the first-year have high impact on classification process. |

### VI. CONCLUSION

This study was an attempt to show the importance of the student retention rate for higher education institutions. It was realized that the first year of the education is critical to increase the retention rate as the students who continue to their sophomore year are more likely to graduate. The discussed theoretical student retention models justified that student academic and social integration with institutions are

**techniques** have been successfully applied to developed models based on these theories to identify student at risks and predict their academic performance. It was found out that classification and prediction are interesting techniques compare to other mentioned techniques to achieve required knowledge. The achieved extract knowledge helped institutions in building proper intervention programme and apply it at the right time to those students in need of these programme. As results, students’ academic performance
improved and led to increase student retention and
graduation rate for institutions.

REFERENCES
[1] The Higher Education Academy, Retention and Success
States higher education: a synthesis of literature”, November
2010.
retention in Higher Education – A case study”, ICEIS. June
2010.
[4] V. Tinto, “Taking student retention seriously: Rethinking the
first year of college”, NACADA, vol.19, pp.5-9, fall 1999.
Kaprotec, “A Data Mining Approach for Identifying Predictors
of Student Retention from sophomore to Junior Year”, JDS,
Approach for Preventing undergraduate Student retention”,
IJCNN, pp.1 - 8, June 2012.
Commons Innovation, Universities, Science and Skills
[9] ACT, “National Collegiate Retention and Persistence to
Degree Rates”, 2012.
[10] University of Kentucky, Office of institutional Research,
February 2013.[online], available: http://www.uky.edu/IRPE/students/ret_grad/Section1.pdf
no cognitive risk factors, and student advising”, January 2012.
Punch, “Predicting Student Performance: An application of
data mining methods with the educational web-based system
Lon-CAPA”, IEEE Trans. International Conference on
“Determination of factors influencing the achievement of the
first-year university students using data mining methods”, In
Proceedings of the Workshop on Educational Data Mining at
the 8th International Conference on Intelligent Tutoring
[14] D. Kabakchieva, “Predicting Student Performance by Using
Data Mining Methods for Classification”, Cybernetics and
Mining Application in Higher Learning Institutions” ,
Section of educational Data Mining”, SIGKDD Exploration,
3-6, 2012.
[18] N. Thai-Nghe, T. Horvath, and L. Schmidt-Thieme,
“Factorization Models for Forecasting Student Performance”,
EDM International Conference on Educational Data Mining,
July 2011.
Students Enrolment Data”, Proceedings of Informing Science
& IT Education Conference (ISITE), June 2010.
Student Data for Academic Performance Using Decision Tree
Techniques”, Global Journal of Computer Science and
Education”, [online], available: http://www.ksbe/spi/PDFS/Retention_Brief.pdf , February
2011.

AUTHOR BIOGRAPHY
Leila Dadkhahan is Master student in Software Engineering in Asia
Pacific University of Technology and Innovation. She has wide experience
and interest in field of data mining, data base design and development, as
well as intelligent software tools.

Mohamad Anas Al Azmeh is Master student in Software Engineering in
Asia Pacific University of Technology and Innovation. His main interest is
in data base development, parallel database systems, and data mining.