Available online at www.sciencedirect.com

ScienceDirect

Procedia - Social and Behavioral Sciences 147 (2014) 390 – 397

Procedia
Social and Behavioral Sciences

ICININFO

Improving Quality of Educational Processes Providing New Knowledge using Data Mining Techniques

Manolis Chalaris*, Stefanos Gritzalis, Manolis Maragoudakis, Cleo Sgouropoulou and Anastasios Tsolakidis

Technological Educational Institute of Athens, Ag. Spyridonos, 12210 Aigaleo, Athens, Greece
University of the Aegean, Department of Information and Communication Systems, Samos GR-83200, Greece

Abstract

One of the biggest challenges that Higher Education Institutions (HEI) faces is to improve the quality of their educational processes. Thus, it is crucial for the administration of the institutions to set new strategies and plans for a better management of the current processes. Furthermore, the managerial decision is becoming more difficult as the complexity of educational entities increase. The purpose of this study is to suggest a way to support the administration of a HEI by providing new knowledge related to the educational processes using data mining techniques. This knowledge can be extracted among other from educational data that derive from the evaluation processes that each department of a HEI conducts. These data can be found in educational databases, in students' questionnaires or in faculty members' records. This paper presents the capabilities of data mining in the context of a Higher Education Institute and tries to discover new explicit knowledge by applying data mining techniques to educational data of Technological Educational Institute of Athens. The data used for this study come from students' questionnaires distributed in the classes within the evaluation process of each department of the Institute.

Selection and peer-review under responsibility of the 3rd International Conference on Integrated Information.

Keywords: Data mining techniques; Higher Education Institutes; Educational P rocesses; Educational Data Mining; Decision support, CRISP-DM methodology

1. Introduction

Nowadays, where knowledge and quality are considered as critical factors in the global economy, Higher Education Institutes (HEI) as knowledge centers and human resource developers play a crucial role. Thus, it is important to ensure the quality of the educational processes and to identify the means by which they can be

* Corresponding author. Tel.: +30-2105385809; fax: +30-2105910975.
E-mail address: manoshlr@teiath.gr

Selection and peer-review under responsibility of the 3rd International Conference on Integrated Information.
doi:10.1016/j.sbspro.2014.07.117

validated and improved in order to provide quality education to students. Quality education is one of the key responsibilities of any University/ HEI to its stakeholders denoting not only the requirement for production of high level of knowledge, but also the need for efficient provision of education so that students achieve their learning objectives without any problem (S. Kumar Yadav, J. P. Nagar, 2012). One way to enhance the quality of educational processes is by improving the decision-making procedures on the various processes by providing the administration of an educational institute with useful knowledge, which is currently unknown to the decision makers. This knowledge can be discovered from data that reside in various databases of the organisation or in evaluation forms that collect data for evaluating related quality criteria (course assessment, lecturer assessment, student assessment, etc.) and can be extracted through data mining technology. The new research field concerned with methods for exploring the unique types of data that come from educational settings and their use to better understand learners and the settings, which they learn in is called Educational Data Mining (EDM International Educational Data Mining Society). Educational Data Mining is considered as one of the most appropriate technology in providing new knowledge about the behavior of lecturer, student, alumni, manager, and other educational staff and acting as an active automated assistant in helping them to make better decisions on their educational activities (N. Delavari, Alaa M. El-Halees, Dr. M. Reza Beikzadeh 2005).

In this paper our aim is to demonstrate the capability of data mining in improving the quality in education by supporting the administration of educational institutions in the decision-making process and in identifying more enhanced policies for educational practices. In addition to that, a main objective of this paper is to conduct some experiments in applying data mining techniques like clustering analysis, correlation analysis and association rules on the educational data of the Technological Educational Institute of Athens (TEIA) collected through the evaluation process of its academic units as well as to present the results and some conclusions.

The rest of the paper is organized as follows: In Section 2, we quote related work in the field of educational data mining and describe the capabilities it has as well as the factors that we claim as success factors of its application. In Section 3, we describe the application of data mining techniques in TEI of Athens and present the results of the experiments conducted. Finally, in Section 4, we conclude this paper and give an outlook of future work.

2. Using Data Mining in Higher Education Institutes

In this section we deal in more detail with the application of data mining technology in Higher Education Institutions. First, we present related work that has been conducted in this area. Subsequently we focus on the outcomes that derive from the application of data mining techniques on educational data and how can a HEI be supported in improving the quality of its educational processes.

2.1. Related work

Despite the fact that using data mining in HEIs is a recent research area, a considerable amount of work has been conducted during the last years. Romero and Ventura (2007) have conducted a survey for the years between 1995 and 2005, where they present a review of different types of educational systems and how data mining can be applied to each of them. In addition, they describe the data mining techniques that have been applied to educational systems grouping them by task. Al-Radaideh et al. (2006) applied decision tree as classification method to evaluate student data in order to find which attributes affect their performance in a course. Mohammed M. Abu Tair & Alaa M. El-Halees (2012) use educational data mining to discover knowledge that may affect the students' performance and Baradwaj and Pal (2011) applied the decision tree as a classification method for evaluating students' performance. Furthermore, in Karel Dejaeger et al. (2011) investigated the construction of data mining models to identify the main attributes of students satisfaction and to support the management in the decision making process while Dursun Delen (2010) examines the institution-specific nature of the attrition problem through models developed by educational data using machine learning techniques.

2.2. Benefits and Success Factor of EDM

As mentioned, Higher Education Institutions face the challenge of providing efficient, effective and qualitative learning experiences to their students. In addition, there is a large amount of educational data about students, alumni, courses, academic staff, etc. that is "hidden" in various databases and files of an institution. These data can be proved as a strategic resource for the institutions in order to enhance the quality of their educational processes. This can be achieved by using various data mining techniques on data stemming from the educational activity i.e. by extracting useful knowledge that will support the administration of each institution in making the appropriate decision for improving the quality of the educational processes.

There are a lot of data mining techniques that can be applied on educational data and each of them can provide useful outcomes and results that assist in addressing many issues and problems in the educational domain. Data mining tasks such as clustering can reveal comprehensive characteristics of students, while prediction (classification and regression) and relationship mining (association, correlation, sequential mining) can help the university to formulate policies and initiatives for decreasing student's drop-out rate or increasing student retention rate, success and learning outcome achievement. It could help in providing more personalized education, maximize educational system efficiency, and reduce the cost of education processes (Y. Zhang, S. Oussena, T. Clark, H. Kim 2010).

An important factor for the success of the application of DM in higher education is the existence of an appropriate infrastructure that supports the institution in finding and collecting all educational data in a centralized system. This system could be a Quality Assurance Information System, which monitors, analyses and reports all factors related to assessing and improving the quality of services provided by the institution. In our case, the Quality Assurance Unit of TEIA has recently developed such a system (M. Chalaris, An Tsolakidis, C. Sgouropoulou, I. Chalaris. 2011) that supports all departments of the institution in the evaluation process and, eventually, in improving the educational processes by the application of data mining techniques on the educational data stored in its repositories.

Beside the importance of an appropriate infrastructure, it is essential for the efficient application of EDM in Higher Education to use an analysis model as a roadmap for the institution in order to identify which part of the educational processes can be improved through data mining and how to achieve each strategic goal. Our intention is to propose such a model in further work. In the context of the current research we conducted some experiments using a number of data mining techniques on the educational data of the TEIA in order to extract some outcomes that provide a first knowledge about indicators of the educational process or to examine the student behavior concerning their performance in each Faculty of the institution.

3. Application of data mining techniques in TEI of Athens

In this section, we describe the experiments conducted on the educational data of TEIA and present the results. More specifically, the experiments have been conducted with data collected by the Quality Assurance Unit of TEIA (MODIP TEIA) for the evaluation process of all departments of the institution for the spring semester of the academic year 2011 - 2012. In specific, unsupervised learning techniques such as clustering and association discovery were applied to data collected from questionnaires which are delivered to student between the 8^{th} and 10^{th} week of courses. For the application of the data mining techniques the Rapid Miner software is used, which is the world-leading open-source system for data mining.

The methodology that is used is based on CRISP-DM methodology- the CRoss Industry Standard Process for Data Mining, a very common methodology used by data miners. The process consists of six steps or phases, as illustrated in Figure 1 (Dr. M. North, 2012).

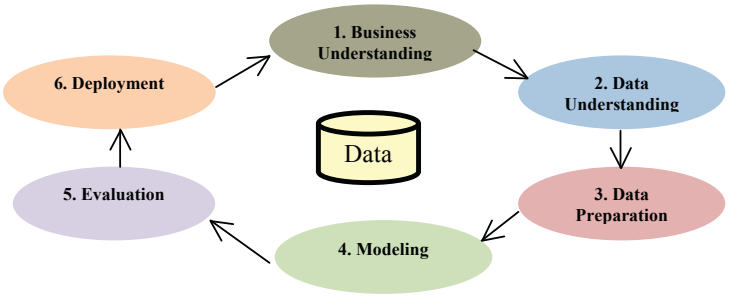

Fig. 1. CRISP-DM methodology

The first step in CRISP-DM is **Business (Organizational) Understanding**. This step focuses on understanding the project objectives and requirements from a business or organisational perspective. Next step is **Data Understanding** where initial data is collected, data quality problems are identified and/or interesting subsets to form hypotheses for hidden information are detected. The 3rd step is **Data Preparation**. In this phase all necessary tasks like data cleaning, data transformation and data selection are performed in order to construct the final dataset. Next we have the **Modeling** step where various modeling techniques are selected and applied. Modeling is followed by the **Evaluation** step, which determines how valuable the model is, and if it achieves the business objectives set. Final step is **Deployment**, which specifies the actions that should be carried out in order to use the developed models.

3.1. Business (Organizational) Understanding

Based on the 1st step of CRISP- DM methodology, we set as ultimate goal of TEIA, concerning the application of data mining techniques on its educational data, the quality improvement of the educational processes. Such a goal is very ambitious and cannot be fully achieved through the current work since it requires more tasks to be carried out in future work. In the context of this paper, the main objective is to conduct some experiments on institutional evaluation data using data mining techniques and derive knowledge that will be proved useful for the administration of the institution.

3.2. Data Understanding

The 2nd step, data understanding, is concerned with the collection of experimental data and the identification of interesting datasets. In our case the educational data come from student questionnaires that have been completed for evaluation purposes during the spring semester of the academic year 2011 - 2012 for the courses of all Departments and Faculties of TEIA. . There are two categories of questionnaires delivered, one for theoretical courses and one for laboratory practice. The questionnaire for theoretical courses contains 35 questions (attributes) and is organized in three sections-directions: course-centered items (e.g. is the course well structured? Is the main theme evident? Were the learning goals of the course clear?), instructor and teaching effectiveness (e.g. how s/he explains the content of the course? How open s/he is towards constructive criticism and suggestions?) and student-centered items (e.g. how often you attend the course? How much do you understand the concepts being taught?). The questionnaire for laboratory practice courses contains 22 questions (attributes) and is also organised in three sections-directions. The two are the same as in the theoretical course questionnaire, instructor/teaching effectiveness and student-centered items while the third one concerns laboratory work (e.g. Were notes/instructions provided?, Is the lab equipment sufficient?). The students make their evaluation using the typical five-level Likert scale; some questions are in yes/no format. It is important to mention that the Cronbach's Alpha value for the two questionnaires is very high, which proves their reliability.

It should be noted that in the described evaluation process all Faculties (five in total) and most Departments (27 of 33) of the institution participated. The total number of questionnaires filled out by the students of the TEI of Athens is approximately 10,000 for the theoretical courses and 16,000 for laboratory practice.

3.3. Data Preparation

In Data Preparation we constructed our final dataset that was used as input for the analysis phase. Recall that for all data mining disciplines, RapidMiner was incorporated. So, we gathered all data from all Departments and Faculties in one dataset, and after data integration we made the data cleansing. Highly incomplete records (a record corresponds to an entire questionnaire) the record was deleted from the dataset. In cases of few missing values these were replaced with the average of the rest of answers of the specific attribute. Finally, depending on the data-mining goal and technique, various attributes were selected in order to conduct the experiments and, thus, proceed to the modeling phase.

3.4. Modeling and Evaluation

As a first modeling technique we used cluster analysis conducted on the data derived from the theoretical courses questionnaire. We used the k-means algorithm, in order to find out if there is a faculty that has better averages in the attributes of the questionnaire concerning all three directions and compare it with the percentage of study duration for each Faculty (depicted in Table 1) (M. Chalaris, I. Chalaris, Ch. Skourlas, An. Tsolakidis, 2012). As number of clusters we chose k=3. For determining the optimal number of clusters we used Ward's algorithm, which is also used and proposed in P. Belsis, A. Koutoumanos, C. Sgouropoulou (2013). Concerning the distance measure, we used the Squared Euclidean Distance. As it is shown in Table 1 students from the Faculty of Health and Caring Professions graduate between 4 and 5 years of study in a percentage of 68,3% which is by far the best comparing it with the other Faculties.

Table 1. the percentage of graduate students per Faculty

Faculty	5 to 8 years	More than 8 years	Percentage of graduates – Length of study between 4 and 5 years
Faculty of Management and Economics (SDO)	14.35	14.61	52,86
Faculty of Technological Applications (STEF)	19.1	19.81	29.5
Faculty of Fine Arts and Design (SGTKS)	12.3	17.8	39.62
Faculty of Health and Caring Professions (SEYP)	8.75	7.4	68,3
Faculty of Food Technology & Nutrition (STETROD)	15.34	18.9	29.15

In this cluster analysis we decided not to include the Faculty of Fine Arts and Design and the Faculty of Food Technology and Nutrition since they had much less answers than the other three Faculties. The result of this cluster analysis (see Figure 2) confirms and explains why the percentage of graduates in the Faculty of Health and Caring Professions is by far the highest.

Specifically, we created three clusters for the students of the three Faculties of TEIA where cluster_2 includes 3516 items (answers), cluster_1 3586 items and cluster_0 1309 items. Cluster_2 is the cluster that has the highest averages in all attributes concerning the course, the lecturer and teaching effectiveness and the characteristics of the students. As we can see, this cluster contains mostly (47,8%) students of the Faculty of Health and Caring Professions ,which explains why they graduate earlier.

Attribute	cluster_0	cluster_1	cluster_2
SXOLH = ΣΤΕΦ	0.331	0.340	0.252
SXOLH = ΣΔΟ	0.320	0.276	0.269
SXOLH = ΣΕΥΠ	0.349	0.385	0.478
q1_theory_goals	2.840	3.893	4.681
q2_theory_material_goals	2.925	3.942	4.704
q3_theory_organization	2.626	3.873	4.725
q4_theory_material_understanding	2.534	3.699	4.636
q5_theory_books_on_time	3.012	3.798	4.511
q6_theory_book_quality	2.710	3.381	4.098
q8_theory_library_books	2.731	3.115	3.569
q20_teacher_transmissibility	2.428	3.852	4.757
q21_teacher_consistent	3.069	4.119	4.803
q22_teacher_cooperation	2.754	3.927	4.726
q23_teacher_material_organization	2.620	3.871	4.718
q24_teacher_interest	2.303	3.718	4.694
q25_teacher_encouragement	2.754	4.021	4.806
q26_teacher_informatics	2.498	3.407	4.292
q27_student_attendance	3.875	4.022	4.494
q28_student_understanding	2.891	3.701	4.316
q30_student_week_study_hours	1.969	1.968	2.225
q31_student_exam_study_hours	3.041	2.927	3.084
q32a_student_number_of_exams	0.565	0.444	0.332

Fig. 2. Centroid table of cluster analysis in theory questionnaire

Our next objective was to examine if the student concepts understanding (q28_student_understanding in theory questionnaire) interacts with other attributes since we claim that this attribute is an important indicator as outcome of the educational process. For this reason, we used correlation analysis, which is the appropriate technique for identifying trends in a data set. Figure 3 presents the correlation matrix, where we have the correlation coefficients between the attributes used in this experiment. As we can see the attribute q28_student_understanding has some positive correlation with attributes concerning the teacher effectiveness, such as teacher transmissibility or teacher interest, while there is no relationship with attributes that count the student attendance and study time and a small relationship with attributes of the course section.

Attributes	**q4**	**q5**	**q6**	**q8**	**q20**	**q21**	**q22**	**q23**	**q24**	**q25**	**q26**	**q27**	**q28**	**q30**
q4_theory_material_understanding	1.0	0.489	0.456	0.277	0.620	0.505	0.533	0.645	0.600	0.540	0.401	0.177	0.461	0.078
q5_theory_books_on_time	0.489	1.0	0.510	0.265	0.408	0.376	0.390	0.430	0.387	0.391	0.242	0.155	0.302	0.036
q6_theory_book_quality	0.456	0.510	1.0	0.315	0.389	0.299	0.350	0.382	0.384	0.348	0.205	0.116	0.330	0.093
q8_theory_library_books	0.277	0.265	0.315	1.0	0.231	0.204	0.237	0.242	0.234	0.224	0.161	0.100	0.206	0.121
q20_teacher_transmissibility	0.620	0.408	0.389	0.231	1.0	0.629	0.684	0.722	0.781	0.690	0.33	0.200	0.534	0.081
q21_teacher_consistent	0.505	0.376	0.299	0.204	0.629	1.0	0.574	0.619	0.548	0.572	0.322	0.198	0.374	0.024
q22_teacher_cooperation	0.533	0.390	0.350	0.237	0.684	0.574	1.0	0.650	0.654	0.672	0.325	0.178	0.450	0.043
q23_teacher_material_organization	0.645	0.430	0.382	0.242	0.722	0.619	0.650	1.0	0.709	0.640	0.420	0.186	0.474	0.066
q24_teacher_interest	0.600	0.387	0.384	0.234	0.781	0.548	0.654	0.709	1.0	0.689	0.357	0.192	0.529	0.100
q25_teacher_encouragement	0.540	0.391	0.348	0.224	0.690	0.572	0.672	0.640	0.689	1.0	0.340	0.184	0.450	0.050
q26_teacher_informatics	0.401	0.242	0.205	0.161	0.33	0.322	0.325	0.420	0.357	0.340	1.0	0.082	0.231	0.100
q27_student_attendance	0.177	0.155	0.116	0.100	0.200	0.198	0.178	0.186	0.192	0.184	0.082	1.0	0.330	0.139
q28_student_understanding	**0.461**	**0.302**	**0.330**	**0.206**	**0.534**	**0.374**	**0.450**	**0.474**	**0.529**	**0.450**	**0.231**	**0.330**	**1.0**	**0.097**
q30_student_week_study_hours	0.078	0.036	0.093	0.121	0.081	0.024	0.043	0.066	0.100	0.050	0.100	0.139	0.097	1.0

Fig. 3. Correlation matrix

In addition, we extracted some association rules using the FP-Growth algorithm with focus on the student understanding (q17_student_understanding) of the laboratory practice courses. The most significant ones denote (Figure 4) that the student understanding relates, with a higher than 80% confidence threshold, to the teacher transmissibility as well as to the lab facilities and the lab facilities in ratio with the number of students attending the course. These rules express how important the laboratory facilities and teacher transmissibility are to students in order to achieve the learning outcomes of a course.

Premises	Conclusion	Support	Confidence
q1_teacher_transmissibility	q17_student_understanding	0.621	0.827
q1_teacher_transmissibility, q14_lab_facilities_vs_student	q17_student_understanding	0.302	0.834
q1_teacher_transmissibility, q13_lab_facilities	q17_student_understanding	0.378	0.855
q1_teacher_transmissibility, q14_lab_facilities_vs_student	q17_student_understanding	0.320	0.870
q1_teacher_transmissibility, q13_lab_facilities q14_lab_facilities_vs_student	q17_student_understanding	0.272	0.878

Fig. 4. Association Rules concerning q17_student understanding as conclusion

The next experiment concerned a cluster analysis on the laboratory questionnaire data in order to examine how students in the three Faculties (SEYP, SDO, STEF) assess the facilities of the labs per se as well as the facilities with regard to the number of attending students (q13_lab_facilities and q14_lab facilities_vs_students). As seen in Figure 5, there is a balance between the three Faculties concerning cluster_1, which has the highest averages in these attributes (STEF 0.318, SDO 0.334, and SEYP 0.347). But in the other two clusters and especially in cluster_0, which is the cluster with the lowest averages, the Faculty of Health and Caring Professions (SEYP) has by far the largest proportion (64, 2%).

Attribute	cluster_0	cluster_1	cluster_2
SXOLH=STEF	0.210	0.318	0.296
SXOLH=SDO	0.148	0.334	0.219
SXOLH=SEYP	0.642	0.347	0.485
q13_lab_facilities	1.745	4.448	3.285
q14_lab_facilities_vs_students	1.723	4.353	2.936

Fig. 5. Centroid table of Cluster Analysis in laboratory questionnaire

3.5. Deployment

What we have learned through the investigations on our educational data is firstly that student understanding, which is a very important outcome of the educational process, relates mainly with the instructor and teaching effectiveness especially in the theoretical courses, while in the laboratory practice courses, lab facilities are considered as a premise for the achievement of learning objectives. Based on this conclusion, HEI administration may invest on the organisation of seminars for the academic staff in order to improve the instructional methods they use in their courses and thus enhance their teaching effectiveness. Concerning the laboratory practice courses, emphasis should also be put on how to improve the lab facilities and reduce the student – equipment ratio.

Another result we extracted through the conducted data mining analysis was the fact that students of the Faculty of Health and Caring Professions (SEYP) are more consistent in their studies (attendance, studying, and understanding) than those of the other Faculties, which also confirms the highest percentages of normal student graduation for SEYP (between 4-5 years). On the other hand, SEYP has the biggest problem concerning the quality and the quantity of the lab facilities. This means that for the improvement of the quality of educational processes in TEIA, the administration of the institution should make decisions focusing on the improvement of the lab facilities in SEYP, while in the other Faculties the focus should be mainly on the educational process such as the organisation of the courses, the teaching effectiveness and student understanding.

4. Conclusion and future work

In this paper, we discussed how the use of data mining techniques on educational data can be proved a useful strategic tool for the administration of HEIs addressing the very difficult and crucial challenge of improving the quality of educational processes. Informed decisions can be made based on knowledge previously unknown and hidden inside the institutional resources. On this basis, decisions can be proved more accurate and correct for the benefit of all stakeholders involved in the educational setting. Furthermore, we presented the results of experiments conducted on educational data of TEI of Athens, as a first step of the application of data mining technology in the institution.

In our future work we aspire to present an integrated approach of applying data mining techniques in HEIs. More specifically, our research plans include the identification of all processes involved in providing education in a HEI, the definition of indicators that determine their quality as well as the proposal of an Analysis Model as a framework for the application of data mining on the educational data of HEIs. Subsequently, we will target the development of a strategic tool for supporting the decision making process for enhancing the quality of educational activities and practices in HEIs.

Acknowledgement

Research underlying this article has been supported by the national project “MODIP of TEI of Athens” – NRSF 2007-2013

References

Al-Radaideh, Al-Shawakfa and Al-Najjar, (2006), ‘Mining Student Data Using Decision Trees’, The 2006 International Arab Conference on Information Technology (ACIT'2006) – Conference Proceedings.

Baradwaj, B. and Pal, S. (2011) ‘Mining Educational Data to Analyze Student s’ Performance’, International Journal of Advanced Computer Science and Applications, vol. 2, no. 6, pp. 63-69.

D. Delen, "A comparative analysis of machine learning techniques for student retention management", (2010) Decision Support Systems, Vol. 49 Iss: 4 pp. 498 – 506

Dr. M. North, (2012) "Data Mining for the Masses"

International Educational Data Mining Society, http://www.educationaldatamining.org

K. Dejaeger, F. Goethals, A. Giangreco, L. Mola, B. Baesens, "Gaining insight into student satisfaction using comprehensible data mining techniques", (2011) European Journal of Operational Research, Vol. 218 Iss: 2 pp. 548 - 562

M. Chalaris, An Tsolakidis, C. Sgouropoulou, I. Chalaris. Developing an In-formation System for Quality Assurance in Higher Education using the Balanced Scorecard Technique - The case study of TEI-A, PCI 2011 in press. ISBN 978-0-7695-4389-5.DOI 10.1109/ PCI. 2011. 43.

M. Chalaris, I. Chalaris, Ch. Skourlas, An. Tsolakidis, (2012). “Extraction of rules based on students’ questionnaires”, Procedia - Social and Behavioral Sciences, Volume 73

M. M. Abu Tair, Alaa M. El-Halees, (2012). Mining Educational Data to Improve Students’ Performance: A Case Study

N. Delavari, Alaa M. El-Halees, Dr. M. Reza Beikzadeh. Application of Enhanced Analysis Model for Data Mining Processes in Higher Educational System, In Proceedings of 6th International Conference ITHET 2005 IEEE

P. Belsis, A. Koutoumanos, C. Sgouropoulou (2013). “PBURC: a patterns-based, unsupervised requirements clustering framework for distributed agile software development.”. Requirements Engineering © Springer-Verlag London 2013

Romero, C. and Ventura, S. (2007) ‘Educational data mining: A Survey from 1995 to 2005’, Expert Systems with Applications (33), pp. 135-146.

S. Kumar Yadav, J. P. Nagar, (2012). Data Mining Application in Enrollment Management: A Case Study

Y. Zhang, S. Oussena, T. Clark, H. Kim (2010). Use data mining to improve student retention in higher education – A case study. In Proceedings of ICEIS 2010.