Student Relationship in Higher Education Using Data Mining Techniques

GJCST Classification (FOR) H.2.8, F.4.2

Boumedyen Shannaq[1], Yusupov Rafael[2], V. Alexandro[3]

***Abstract*- The aim of research paper is to improve the current trends in the higher education systems to understand from the outside which factors might create loyal students. The necessity of having loyal students motivates higher education systems to know them well, one way to do this is by using valid management and processing of the students database. Data mining methods represent a valid approach for the extraction of precious information from existing students to manage relations with future students. This may indicate at an early stage which type of students will potentially be enrolled and what areas to concentrate upon in higher education systems for support. For this purpose the data mining framework is used for mining related to academic data from enrolled students. The rule generation process is based on the decision tree as a classification method. The generated rules are studied and evaluated using different evaluation methods and the main attributes that may affect the student's loyalty have been highlighted. Software that facilitates the use of the generated rules is built using VB.net programming language which allows the higher education systems to predict thestudent's loyalty (numbers of enrolled students) so that they can manage and prepare necessary resources for the new enrolled students.**

Keyword-Data mining ,decision tree , Exploratory data analysis, Adaptive System .

I. Introduction

Nowadays there is an evolution of educational systems and there is a great importance of the educational field. Modern educational organizations start developing and enhancing the educational system increasing their capability to help the decision makers obtain the right knowledge, and to make the best decisions by using the new techniques such as data mining methods [1]. Subsequently, a suitable knowledge needs to be extracted from the existing data. Data mining is the process of extracting useful knowledge and information including: patterns, associations, changes, anomalies and significant structures from a great deal of data stored in databases, data warehouses, or other information repositories. The data mining expediency is delivered through a series of functionalities such as outlier analysis, evolution analysis, association analysis, classification, clustering and prediction. Data mining is an integral part of Knowledge Discovery in Database (KDD) [2][3]. Student enrollment process in any higher education system is of great concern of the higher education managements. Severalfactors may affect the student enrollment process in a particular Institute. One of the biggest challenges that higher education faces today is predicting the paths of loyal students (enrolled students). Institutions would like to know, for example, which students and how many will enroll in particular institute.This research paper is an attempt to use the data mining processes, particularly predictive classification to enhance the quality of the higher educational system to increase numbers of loyal students (enrollment students) to evaluate student data to study the main attributes that may affect the student enrollment factors to plan for institutes resources (Instructors, classes, labs, etc.)from knowing how many students will be enrolled and to make a big effort to concentrate on all factors that play main role in motivating the new student to enrolls in a particular institute.

II. Related work

Higher education enrollment and admission departments are increasingly being asked to do more work in these aspects. Additionally, as described in [4] they recognize that each institution has different student relationship management, admissions and enrollment goals. In strengthening student relationship management for enterprise students projects, the project aims to improve the overall quality of the "entrepreneurial" student experience by mapping existing and proposing new service designs, this can lead to an increase in the efficiency and effectiveness of University or College's system supporting the management of relationships with students as well as in the interface between operational processes and the platforms and technical solutions used for it. As described in Student Relationship Management (SRM) article they introduce the topic of Student Relationship Management (SRM) in Germany. The concept has been derived from the idea of a Customer Relationship Management (CRM), which has already been successfully implemented in many enterprises [4]. The German university information system HIS ascertained that 65% of the first-year students decided to choose their university due to their place of residence. Good equipment of the university is an important criterion for 58% of high school graduates, as well as the reputation of the university or college (52%). Nevertheless, for 90% of the high school graduates it is above all important that the courses offered correspond to their specialized interests [5]. Furthermore, the use of information offered by university rankings becomes more and more frequent. However, not only German but also foreign students were used to select German university on the basis of the results of university

About[1]. Information Systems Department University of Nizwa , Sultanate of OmanEmail: boumedyen@unizwa.edu.om Email: boumedians@yahoo.com
About[2]. Director of Saint Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences Saint Petersburg, Russia Email: yusupov@iias.spb.su
About[3]- Academician of Academy of Natural Sciences, Russia Doc.Nat.Sci., St.Petersburg Institute Doc.Nat.Sci., St.Petersburg Institute E-mail: alexandr@iias.spb.su

rankings which are available for the people. Therefore, it is not amazing that faculties with good ranking results register more students in the following term in relation to the previous years. However, neither the university management nor education administrators are aware of this competition trend because they associate competition rather with areas of research and professors than with students. Good universities, though, are not only characterized by outstanding researchers but by excellent students as well [6]. The research has [7] proposed a model to represent how data mining is used in higher educational system to improve the efficiency and effectiveness of the traditional processes in this model a guideline was presented for higher educational system to improve their decision-making processes. In [8] a complete system implemented for parsing such repositories into a SQL database and for extracting from the database and repositories, various statistical measures of the code and version histories. The research by [9] is to use Rough Set theory as classification approach to analyze student data. They build a higher educational system to improve their decision-making procedures through data mining. Other papers discuss different AI technologies and compare them with genetic algorithm based induction of decision trees and discuss why the approach has a potential for developing into an alert tool. Many other related works can be found in [12, 13, 14, 15, 16, and 17].

III. Data pre-processing

In the entire data mining process, the data cleaning process is utilized in order to eliminate irrelevant items. The discovery of patterns will be only useful if the data represented in files offer a real representation of the enrolment process and the actions or decisions taken by the past student [16]. Here we study the enrollment factors and decisions taken by students for enrollment in a particular institute. Subsequently if the student is enrolled in a particular institute then he is considered as a loyal student. Initially, the institute provided us a database equal to 19012. After filtering process of the data 2106 records were left. The data supplied is from a particular institute, we cannot mention its name, but it is from the Arabic region. The main objective of our research is to discover patterns that will be used to predict a loyal or not loyal student previously to his enrollment or not enrollment to a particular institute. By taking information from other students with similar information, in this sense, we can know the role of each attribute and the implicit relations among them.

IV. Exploratory data analysis

To achieve the goals, we analyze the behavior over time of students who first register and pay fees. We did that to eliminate possible effects in case of changing the structure of the institute. We first considered all students that have entered into the students database between 2003 and 2007, the total number is large, it is time consuming to analyze the whole data set, so we took a sample and analyzed that. We selected the same number of students from each time slot over the whole entry period; the sample contains a total of 2069 students. Generally it is not necessary to sample the data; the main reason for doing it here is the low quality of available data. After a long process of data management we obtained the variables, the features (variables) listed in the data set are the following:

Age: three classes for age between 18 – 28 (young) given number one in data set, for age between 29 -39 (middle) given number two and for age above 39 given number 3 for representation data in the data set.

Ch: number of children's

Sx : define female and male, codes of 0 and 1 for female and male.

Rel : define religion, codes of 0 and 1 for Muslim and other .Pob: place of birth, codes of 0 and 1 for original (residence) and foreign.

Nationality: codes of 0 and 1 for original(residence) and foreign.

SAvg: grade of secondary school.

Stype: type of study, codes of 0 and 1 for literal and scientific.

PGS: place of graduated secondary school, codes of 0 and 1 for original(residence) and foreign.

Numos: Numbers of different specialization selected by the new student.

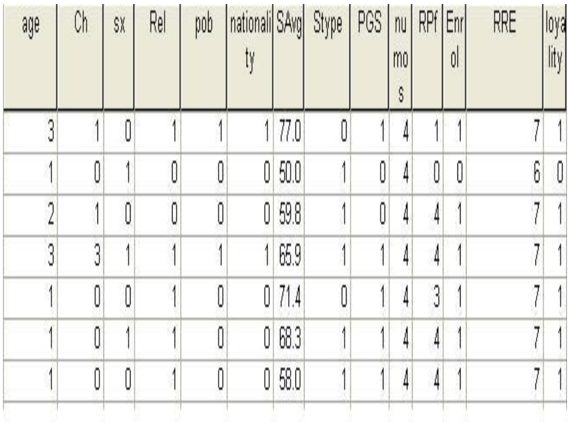
RPf: student information entered in student database and the student pays for some registered courses.

Enroll: the student enrolled has codes 1 and for student who did not enroll with 0.

RRe: reason of enrollment in a particular institute codes of 1,2,3,4,5,6,7,8,9 for Public institute, much specializations ,accept low average after secondary school, much (prof, dr.), private institute , near place, much graduated students from this institute gets a job and offered grants .

Loyalty : codes 1 and 0 for loyal and not loyal.

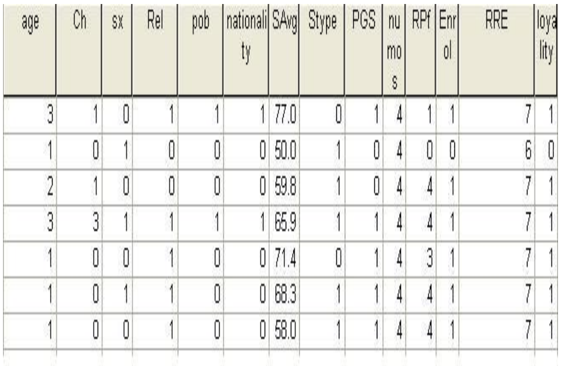
We assigned descriptive value labels for each value of a variable. This process is particularly useful if your data file uses numeric codes to represent non-numeric categories (for example, codes of 1 and 0 for male and female).Value labels are saved with the data file. You do not need to redefine value labels each time you open a data file, the value labels process illustrated in figure4.1 and figure4.2.



age	Ch	sx	Rel	pob	nationality	SAvg	Stype	PGS	numos	RPf	Enrol	RRE	loyality
3	1	0	1	1	1	77.0	0	1	4	1	1	7	1
1	0	1	0	0	0	50.0	1	0	4	0	0	6	0
2	1	0	0	0	0	59.8	1	0	4	4	1	7	1
3	3	1	1	1	1	65.9	1	1	4	4	1	7	1
1	0	0	1	0	0	71.4	0	1	4	3	1	7	1
1	0	1	1	0	0	68.3	1	1	4	4	1	7	1
1	0	0	1	0	0	58.0	1	1	4	4	1	7	1

Figuer4.1: Data set Sample in label view

INTRODUCTION



age	Ch	sx	Rel	pob	nationality	SAvg	Stype	PGS	numos	RPf	Enrol	RRE	loyality
3	1	0	1	1	1	77.0	0	1	4	1	1	7	1
1	0	1	0	0	0	50.0	1	0	4	0	0	6	0
2	1	0	0	0	0	59.8	1	0	4	4	1	7	1
3	3	1	1	1	1	65.9	1	1	4	4	1	7	1
1	0	0	1	0	0	71.4	0	1	4	3	1	7	1
1	0	1	1	0	0	68.3	1	1	4	4	1	7	1
1	0	0	1	0	0	58.0	1	1	4	4	1	7	1

Figure4.2 : Data set Sample in value label view

V. SYSTEM DESCRIPTIONS

What we want to do is to model a function – predict a loyal or not loyal student as a function of age, ch, sx, Rel, pob, etc. We are trying to build an adaptive system that will model an input-output relationship from our data file as Illustrated in figure 5.1

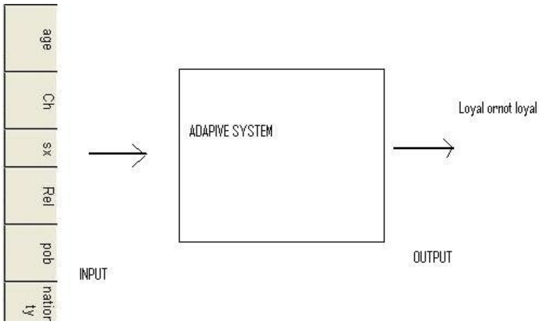
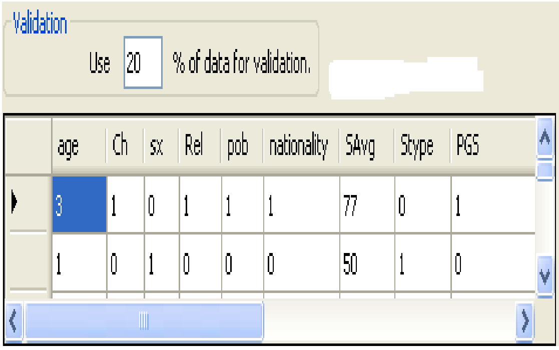


Figure5. 1: Adaptive System

We assigned 20% of data for validation; we did that because the Adaptive systems learn from data we supply it. But how can we be sure that they will work with other data – data that they haven't seen? The answer to that is validation. Instead of using all of the data available for training the system, we leave some aside with which we later test the system. This makes sure that we know how well the system is capable of generalization i.e. how well it works on data it hasn't been trained on. The 20% here refers to how much data should be put aside for validation figure 5.2.



Validation

Use 20 % of data for validation.

age	Ch	sx	Rel	pob	nationality	SAvg	Stype	PGS
3	1	0	1	1	1	77	0	1
1	0	1	0	0	0	50	1	0

Figure5.2: assignment validation data

1) *Attribute Selection*

We needed to find a minimal set of attributes that preserve the class distribution. Attribute relevance is with respect to the class, i.e. relevant attribute and not relevant attribute. This will help us to select the best attributes from 14 attributes that have been collected. This will rank the attributes, according to the effectiveness of the features starting with the most significant feature, as follows in table 5.1.Before selecting the best attributes The figure 5.1.1 graphically depicts the instances in the dataset by using various combination of attributes as x/y axis values .

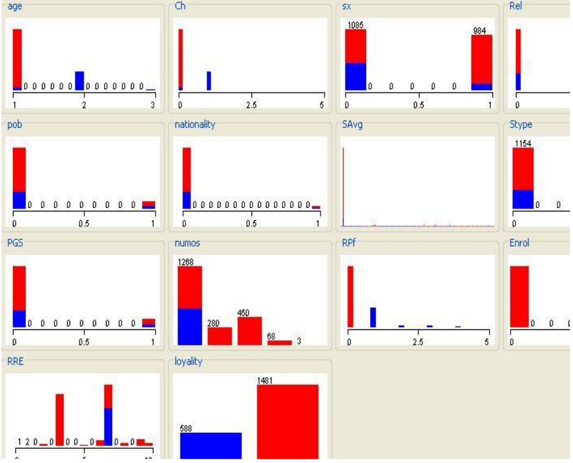


Figure 5.1.1 various combinations of attributes as x\y axis values.

Table 5.1.1: Ranked attributes, Attribute Selection on all input data

Instances: 2069	Evaluation mode: evaluate on all training data	Search Method: Best first.	Search direction: forward	Attribute Subset Evaluator (supervised, Class (nominal): 14 loyality)	Selected attributes: 1,4,11,12,13 age Rel RPf Enrol RRE

To select the best attributes we evaluated the worth of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them. Subsets of features that are highly correlated with the class while having low inter correlation are preferred. For more information about calculation the worth of a subset of attributes in [17].For Search method we apply Best First, its Searches the space of attribute subsets by greedy hill climbing augmented with a backtracking facility.

Setting the number of consecutive non-improving nodes allowed controls the level of backtracking done. Best first may start with the empty set of attributes and search forward, or start with the full set of attributes and search backward, or start at any point and search in both directions (by considering all possible single attribute additions and deletions at a given point).

2) *Attribute Selection (ranking)*

Our objectives are to build the classification model using a decision tree method. The decision tree is a very good method since it is relatively fast, it can be converted to simple rule, and accuracy level is high. The decision tree function in this project is focusing on classify the data in suites, to reach our data mining goal, which will be used to find the relationship between a specific features. By using the ID3 algorithm .In this project ID3 is applied on Vb.net programming language. The basic idea of ID3 algorithm is to construct the decision tree, by using the metric information gain, where attribute that is most useful, and by measuring which questions provide, the most balanced splitting the depth of the tree. The following rules will be applied with Vb.net language. Before discussing rules we used the metric information gain to reduce the entropy related to specified attribute when splitting decision tree node. The gain ratio evaluates the worth of an attribute by measuring the gain ratio with respect to the class.GainR(Class, Attribute) = (H(Class) - H(Class | Attribute)) / H(Attribute).The Ranker ,Ranks attributes by their individual evaluations. Use in conjunction with attribute evaluators (ReliefF, GainRatio, Entropy etc). .We used the gain to rank attributes and to build our decision tree where each node is located the attribute with greatest gain, among the other attribute.

Table 5.2.1 , Attribute Selection (ranking)

Instances: 2069	Attributes: 14 Age, Ch, sx, Rel, pob, nationality, SAvg, Stype, PGS, numos, RPf, Enrol, RRE , loyality	Evaluation mode: evaluate on all training data	Search Method : Attribute ranking.	Attribute Evaluator (supervised, Class (nominal): 14 loyality): Gain Ratio feature evaluator	Ranked attributes: 1 11 RPf 1 12 Enrol 0.80082 1 age 0.80082 2 Ch 0.31307 13 RRE 0.24924 4 Rel 0.16957 10 numos 0.10321 3 sx 0.07551 7 SAvg 0.00546 5 pob 0 9 PGS 0 6 nationality 0 8 Stype	Selected attributes:11,12,1,2,13,4,10,3,7,6,8:13

See figure below after filtering the selected attributes

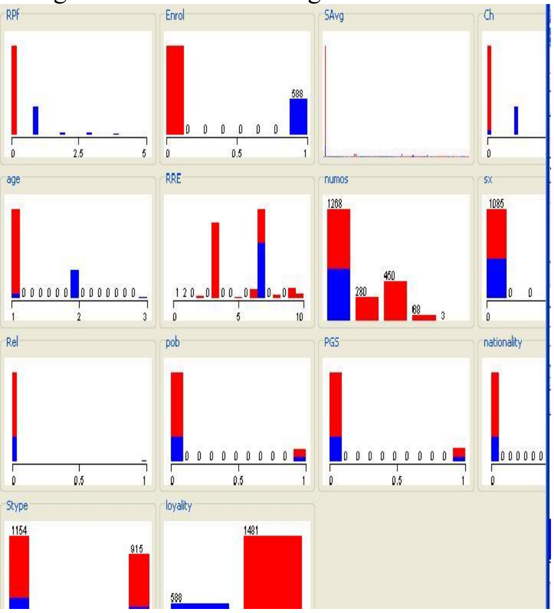


Figure 5.2.1 various combination of attributes as x\y axis values after filtering.

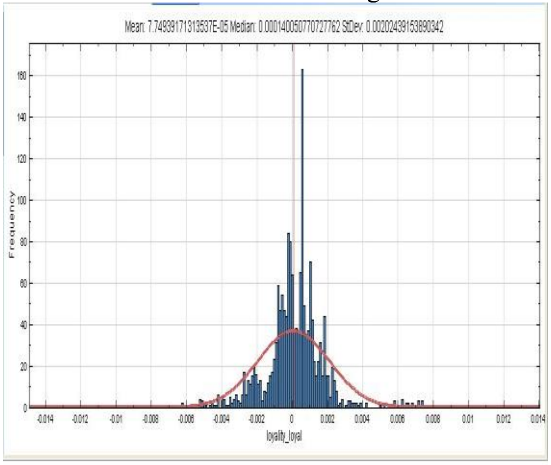
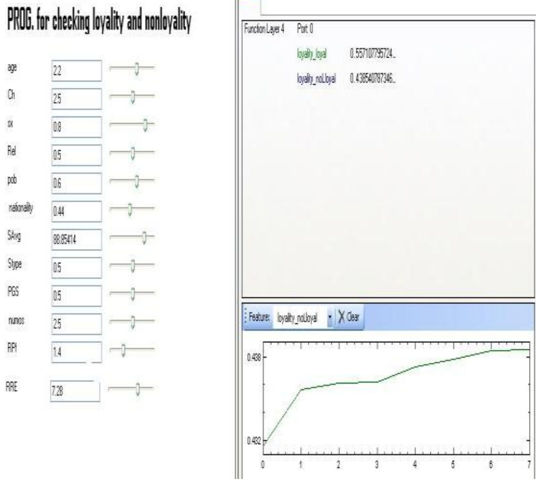


Figure 5.2.2 : Linear error analysis



PROG. for checking loyality and nonloyality

age	2.2
Ch	2.5
sx	0.8
Rel	0.5
pob	0.6
nationality	0.44
SAvg	88.85414
Stype	0.5
PGS	0.5
numos	2.5
RPf	1.4
RRE	7.28

Function Layer 4 Port 0

loyality_loyal 0.557107795724...

loyality_notloyal 0.438540787346...

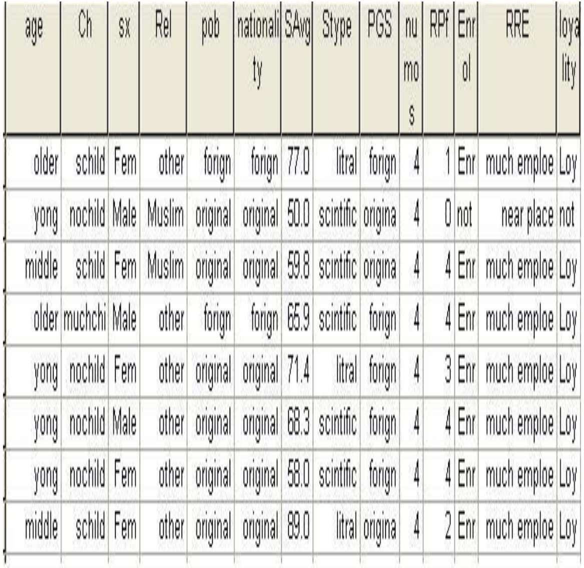
Feature: loyality_notloyal

Figure 5.5.3: Sample of Program Interface

After providing the loyal, non loyal program with previous attributes of this program will Generate 551 rules, these rules will be used for predicting loyalty of enrolled students in a particular institute.Sample of generated rules:

IF (AGE =3 AND CH=1 AND SX = 0 AND REL= 1 AND POB = 1 AND NATIONALITY = 1 AND SAVG = 77 AND STYPE =0 AND PGS = 1 AND NUMOS = 4 AND RPF= 1 AND RPE = 7)THEN

Loyal studentElseNon loyal studentEnd if



age	Ch	sx	Rel	pob	nationality	SAvg	Stype	PGS	numos	RPf	Enrol	RRE	loyality
older	schild	Fem	other	forign	forign	77.0	litral	forign	4	1	Enr	much emploe	Loy
yong	nochild	Male	Muslim	original	original	50.0	scintific	origina	4	0	not	near place	not
middle	schild	Fem	Muslim	original	original	59.8	scintific	origina	4	4	Enr	much emploe	Loy
older	muchchi	Male	other	forign	forign	65.9	scintific	forign	4	4	Enr	much emploe	Loy
yong	nochild	Fem	other	original	original	71.4	litral	forign	4	3	Enr	much emploe	Loy
yong	nochild	Male	other	original	original	68.3	scintific	forign	4	4	Enr	much emploe	Loy
yong	nochild	Fem	other	original	original	58.0	scintific	forign	4	4	Enr	much emploe	Loy
middle	schild	Fem	other	original	original	89.0	litral	origina	4	2	Enr	much emploe	Loy

VI. CONCLUSION

Data mining is a powerful analytical tool that enables educational institutions to better allocate resources and staff, and proactively manage student outcomes. With the ability to uncover hidden patterns in large databases, community colleges and universities can build models that predict with a high degree of accuracy the behavior of population clusters. By acting on these predictive models, educational institutions can effectively address issues ranging from transfers and retention, to marketing and alumni relations. The goal of education is to help people, especially young people, to participate in the functions of society, to acquire knowledge and to develop skills that will help them to confront the needs of the future and to be productive and competitive in tomorrow's world. This research paper is intended to enhance the quality of the higher educational system by focusing on using the data mining techniques. Using this research, the University will have the ability to predict the students loyalty (numbers of enrolled students) so they can manage and prepare necessary resources for the new enrolled students. Moreover, the higher managements can use the classification model to enhance the University resources according to the extracted knowledge. It is certain that the future holds a lot of surprises. It is another major task for education to prepare all resources that give young people the qualities and the skills for the jobs that do not exist yet and we believe that this research paper can help considerably towards that. It should also be a major task of the educational system to provide these qualities and skills in an enjoyable and modern way. The result of our experiment can be used to obtain a deep understanding of student's enrollment pattern in a University where the faculty and managerial decision makers can utilize any action to provide extra basic course skill classes and academic counseling. In addition the management system can improve their policy, enhance their strategies and thereby improve the quality of that management system.

VII. References

1) Higher Education Enhancement Project (HEEP), 2007, http://www.heep.edu.eg.
2) Han J, Kamber M, Data Mining- Concepts and Techniques. Morgan KaufmannPublishers ,2001.
3) Luan J, "Data Mining and Its Applications in Higher Education" in A. Serban and J.
4) Luan (eds.) Knowledge Management: Building a Competitive Advantage for Higher Education. New Directions for Institutional Research, No. 113. San Francisco, CA: Jossey Bass (2002).
5) Nigel culkin, norbert morawetz, university of hertfordshire, centre for innovation and enterprise , www.hampp-verlag.de
   Hilbert, Andreas, Schnbrunn, Karoline, Schmode, Sophie."Student Relationship Management In Germany - Foundations And Opportunities", Management Revue, 2007.
6) http://www.britannica.com/bps/additionalcontent/18/25385295/Student-Relationship-Management-in-Germany--Foundations-and-Opportunities.
7) Delavari N, Beikzadeh M. R. A New Modfor Using Data Mining in Higher Educational System, 5th International Conference on Information Technology based Higher Education and Training: Istanbul, Turkey, May-2nd Jun 2004.
8) Mierle K, Laven K, Roweis S, Wilson G, Mining Student CVS Repositories for Performance Indicators, 2004.
9) Varapron P. et al. Using Rough Set theory for Automatic Data Analysis. 29th Congress on Science and Technology of Thailand. 2003.
10) Two Crows Corporation. Introduction to Data Mining and Knowledge Discovery, Two Crows Corporation. Third Edition, U.S.A, 1999.
11) Han J, Kamber M. Data Mining Concepts and Techniques. Morgan Kaufmann Publishers, 2001.
12) Mehta M, Agrawal R, Rissanen J. SLIQ: A Fast Scalable Classifier for Data Mining in proc. 1996 int. conf. extending database technology (EDBT'96),Avignon, France, Mar 1996.
13) Murthy K. Automatic Construction Of Decision trees from Data : Multi-Disciplinary Survey. Siemens Corporate Research, Princeton, NJ 08540 USA.
14) Peng W,Chen J, Zhou H, An Implementation of ID-Decision Tree Learning Algorithem, Universty of New South Wales, Australia.
15) Esposito F, Malerba, D & Semeraro G, A Comparative Analysis of Methods for Pruning Decision Trees. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 19, No. 5, pp. 476-491, 1997.
16) Chapman P, Clinton J, Kerber R, Khabaza T, Reinartz T, Shearer C, Wirth R.
17) M. A. Hall (1998). Correlation-based Feature Subset Selection for Machine Learning. Hamilton, New Zealand.