KNOWLEDGE DISCOVERY USING DATA MINING TECHNIQUES: A CASE STUDY

OPREA Cristina[1], ZAHARIA Marian[1], ENĂCHESCU Daniela[1]
Petroleum&Gas University[1], Ploiesti, oprea_cris2005@yahoo.com,
marianzaharia53@gmail.com, denachescu22@yahoo.com

Keywords: data mining, association rules, decision tree, ID3 algoritm.

Abstract: *This paper is an attempt to use data mining techniques and methods for knowledge discovery that describe the performance of first year students at the end of the first semester. The study was performed at the Faculty of Economics of the Petroleum-Gas University in Ploieşti and targeted a total of 323 students in specialties: Bank and Finance and Economic Informatics. The methods used are the algorithm "ChiSquartAttributeEval" with Ranker search method for classification of attributes, and association with Apriori algorithm and ID3 classification with decision trees algorithm.*

1. INTRODUCTION

Data mining consists of an evolving set of techniques that can be used to extract valuable information and knowledge from massive volumes of data. [6].
Data mining process includes the following activities[12]:

- *data selection* aimed retrieval of massive data only for relevant data for analysis;
- *data cleaning* dealing with data cleaning and preparation of the activities that are necessary to ensure accurate results,
- *data transformation*, converts the data into a bidimensional table and eliminates unwanted fields so that the results to be valid;
- extracting patterns from data (*data mining*) is to analyze the data by a suitable set of algorithms to discover patterns and significant rules and to produce predictive models
- *data validation* which requires proper interpretation of the results of data mining and aims to select those models that are valid and useful in future decisions in different areas.

Data mining techniques to discover segments, clusters, subgroups to classify and better understanding of the phenomenon analyzed and implement details of its forecasts evolution .
While using data mining techniques in higher education is an area of recent research, there are many works in this field. Studies conducted between 2000 and 2009 led to the conclusion that educational data mining is a promising area of research and have specific requirements.

2. DATA PRESENTATION

Monitoring and support of first year students is very important for higher education. Many students are unable to accommodate to the specific requirements and university programs and leaves school after the first session or require transfer to other specialties or other institutions with part-time program. Also, there is always a potential group of students called "risk group" who can graduate, but who need extra attention or advice. Detection of this group of students at an early stage is essential for saving them from abandonment
The case study presented below was developed based on data collected from the University Petroleum-Gas in Ploiesti on the activity of 323 students from the specialization "Finance-Banks "and" Economic Information "in the first half of academic year 2009-2010.
At the Oil and Gas University in Ploieşti were identified two data sources that provide information about students. These data sources are:

1. the database of the Economics faculty of the University Petroleum-Gas in Ploieşti, which were taken data over marks obtained in examinations, tests unpromoted or were not submitted, the form of education that follows (with fee or no fee);
2. the database of candidates in 2009 at the Faculty of Economics who extracted data on the average admission marks;
3. student questionnaires.
   Data quality in terms of attributes used for completing the degree is 99%.
   We have identified 8 attributes, namely:

Table 1 The attributes model

Attribute	Description	Possible Values
Gender	Student Gender	0,1
Average_entrance_exam	The average college admission	1,2,3,4
Financial form	Student finance form	0,1
Job	If the student has a job	0,1
Specialisation	Specialization that follows at the university	0,1
Remaining_exams	The number of exams unpromoted	0,1,2,3,4
Semester_average	The average marks obtained in first exam session	0,1,2,3,4

Average_entrance_exam attribute values are average for admission to university is determined by the average obtained at the baccalaureate exam, the average obtained in high school and high school average obtained from one of the disciplines mathematics, economics, or geography.

Table 2 The value of Average_entrance_exam attribute

Average_entrance_exam values	Average
1	[9;10]
2	[8;9)
3	[7;8)
4	<7

Financial form attribute model the student funding and has two possible values, namely 0 for funding from the budget and 2 for funding of their studies (with tax).
Job attribute takes the value 0 if the student does not have a job and 1 if it has a job
Specialisation attribute indicates specialization that the student follows, and 1 for specialization "Finance Banks" and 2 for "Economic Informatics".
The number of exams that the student has not passed in the first session, *Remaining_exams* attribute, is given and receives value 0 if the student has no outstanding exams, 1 if not passed a single exam and for the students who have more than 3 exams outstanding, takes the value 4.
End attribute is *Semester_average* which is the average marks obtained by students in the first session of exams and may take the following values:

Table 3 The value of Semester_average attribute

Semester_average values	Average
0	If the student has passed all the exams
1	[9;10]
2	[8;9)
3	[7;8)
4	<7

3. THE PROPOSED MODEL

As data mining software it is used WEKA platform. Weka stands for Waikato Environment for Knowledge analysis (Waikato Environment Knowledge Analysis) software and is a University of Waikato product, New Zealand.

Weka is a collection of automatic learning algorithms for data mining in Java. Algorithms can be applied directly on a data set or can be called from code written by the programmer [9, 13]. Weka contains tools for data preprocessing, classification, regression, clustering, association rules and visualization. Contains a collection of visualization tools and algorithms for data analysis and predictive modeling associated with graphical user interfaces to offer easy access to it's tools.

Weka strengths of this package are:

1. is available under GNU (General Public License)
2. it is very portable as it is implemented in Java programming language, language that runs on any platform;
3. contains a collection of techniques for preprocessing and data modeling;
4. is easy to use even by a beginner because it uses graphical user interfaces.

Weka uses as input formats CSV or ARFF data. Therefore, Excel tables with the necessary data processing was converted into ARFF format

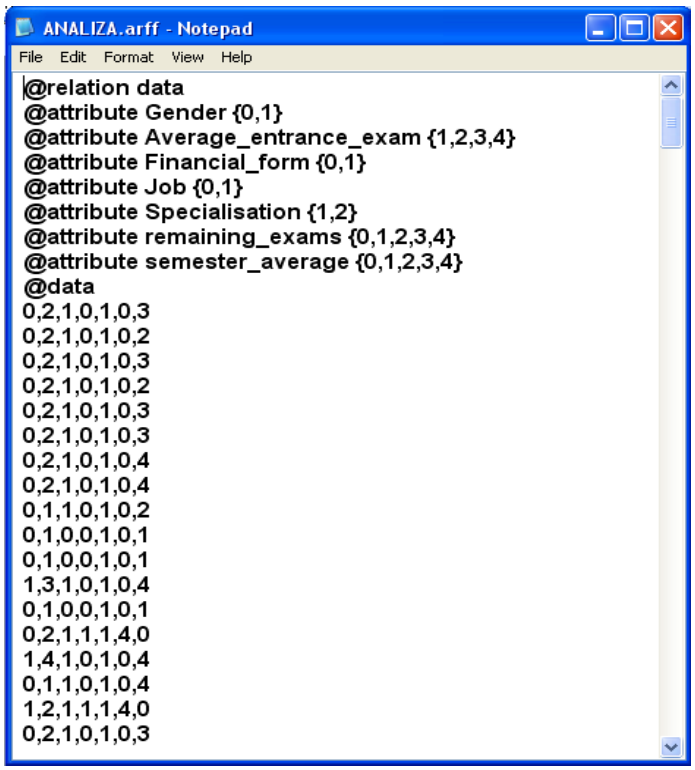

Figure 1 Source file containing the input dataset

Analyzing the average distribution obtained at the end of the first semester according to the seven attributes we observed the following aspects:

- From 323 students, 128 students have not passed all the exams, of which 61 students have over four exams unpromoted and only 13 with one test remaining.
- Number of students holding a job is 58.
- The student average mark for admission to college is quite high, over half of the students has average marks over 8.
- The results of the students in the specialization "Finance banks" were better than those of specialization" Economic Informatics"'.

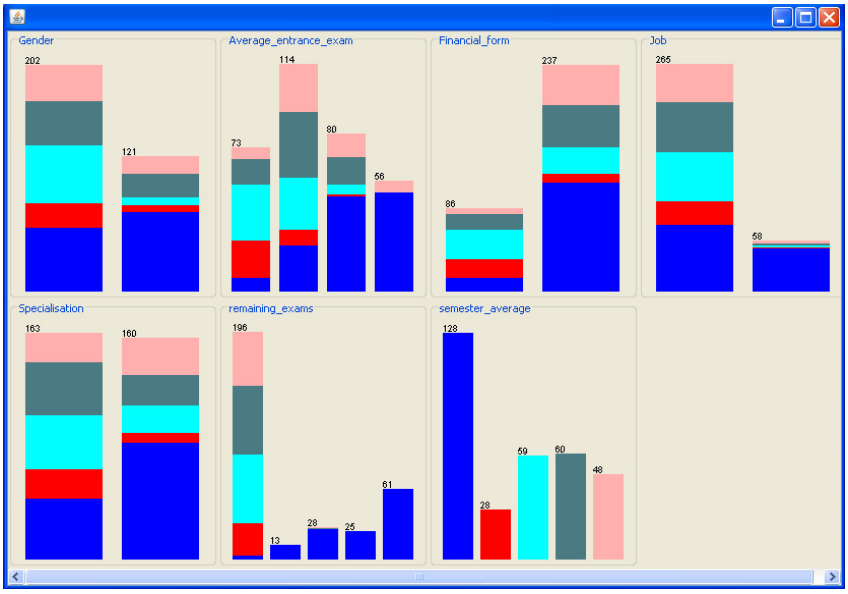

Figure 2 Distribution attributes

We will apply "*ChiSquartAttributeEval*" algorithm with Ranker search method to order the attributes according to their importance on purpose attribute. "*ChiSquartAttributeEval*" calculates the intensity of the relationship between variables using HI squared (C2) test. Evaluation is given by an integer on a scale of 1 to 6, where 1 is the value that is smaller and has a minor importance on the attribute purpose attribute.

```
=== Attribute Selection on all input data ===

Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 6 semester_average):
        Chi-squared Ranking Filter

Ranked attributes:
163.202   2 Specialisation
 65.678   3 Financial_form
 64.308   4 Job
 37.794   1 Gender
 30.61    5 Average_entrance_exam

Selected attributes: 2,3,4,1,5 : 5
```

Figure 3 Results of evaluation attributes

We observe that attribute Average_entrance_exam is the most important attribute for student performance after the first examination session, followed by *job* attributes and *Financial_form*.
After analyzing the associations with *Apriori* method there were discovered the following association rules for students who get good and very good results by the end of the first semester:

Table 4 Association rules

1.	Average_entrance_exam=2-Specialisation=1 75-Financial_form=1 74
2.	Average_entrance_exam=2 Job=0 - Specialisation=1 65-Financial_form=1 64
3.	Average_entrance_exam=4 56 - Financial_form=1 55
4.	Gender=0-Average_entrance_exam=2 Specialisation=1 55-Financial_form=1 54
5.	Average_entrance_exam=1 - Specialisation=1 51 - Gender=0 50
6.	Average_entrance_exam=4 Specialisation=2 49 - Financial_form=1 48
7.	Gender=0-Average_entrance_exam=2-Job=0-Specialisation=1 49-Financial_form=1 48
8.	semester_average=3 60-Job=0 58

It follows, therefore, a positive correlation between the followed specialization,
the form of funding, the fact that the student doesn't have a job , the admission average and good final results obtained by students.
The next step is to build the model of decision tree classification method. Decision tree method is a relatively quick and can be easily converted into simple classification rules. It also provides information on the most important attribute of the model.
The decicsion tree was generated usig the ID3 algorithm and the the folwing production rules were detected:

1. ***If*** *student's admission average grade is between 8 and 10* and *the student attends free of charge courses* and *he doesn't have a job* and *he's attending "Finance banks" specialization* then the average grade of the first semester will be between 8 and 9.
2. ***If*** *student's admission average grade is between 8 and 10* and *the student attends free of charge courses* and *he doesn't have a job* and *it's a girl* then the average grade of the first semester will be between 8 and 9.
3. ***If*** *student's admission average grade is between 5 and 7* and *the student attends charged courses* and then student doesn't pass all the exams on the first semester.

4. CONCLUSIONS

In this paper we use data mining techniques to identify the profile of students who have difficulties adapting in first year of college. The results obtained indicate a moderate direct correlation between the average obtained but by students at the first session exams and the attributes analyzed. Most first year college students are not prepared to meet university requirements. This is aggravated by the fact that, following the payed course, they are forced to take a job. These students may be advised to attend the small frequency courses.
The model can be developed by introducing new attributes, such as average marks obtained in high school economics and mathematics disciplines and the financial situation of students.

REFERENCES

[1] Aligulizev, R., Clustering of document collection – A weighting approach, *Expert Systems with Applications*, vol. 36, nr. 4, 2009, pg. 7904-7916.

[2] Bodea, V., Roşca, I.Gh., "Analiza performanţelor studenţilor cu tehnici de data mining: Studiu de caz în Academia de Studii Economice Bucureşti", în Managementul cunoaşterii în universitatea modernă, coord.: Bodea C.N. şi Andone I., Editura ASE Bucureşti, 2007.

[3] Dekker, G., Pechenizkiy, M., Vleeshouwers, J, "*Predicting Students Drop Out: A Case Study"*, In *Proceedings of the International Conference on Educational Data Mining*, Cordoba, Spain, Educational data mining, 2009

[4] Gorunesu, F., *Data Mining. Concepte, modele şi tehnici, Editura Albastră,* Cluj-Napoca, 2006

[5] Halees, Alaa El, „Mining Students Data To Analyze Learning Behavior:A Case Study", The International Arab Conference on Information Technology (ACIT), 2008
http://www.ijcaonline.org/archives/number22

[6] Ioniţă, A., Asupra termenului de data mining, *Revista Română de Informatică şi Automatică*, vol. 15, nr. 2, 2005.

[7] Luan, J., *Data Mining Applications in Higher Education*, Executive Report SPSS, http://www.pse.pt/Documentos/Data%20mining%20in%20higher%20education.pdf

[8] Oprea, C., Zaharia, M., Gogonea, M., "*Analiza comportamentului studenţilor la licenţă utilizând tehnici de data mining*", The 14th IBIMA Conference on Global Business Transformation through Innovation and Knowledge Management, Istanbul, Turkey 23-24 June 2010.

[9] Oprea, M., Tudor, I., Cărbureanu, M. - *Prediction of Student Professional Evolution with Data Mining Techniques*, The eight international conference on informatics in economy, Bucharest, Romania, May 17-18, 2007

[10] Roşca, I., Bodea, C., *Managementul cunoaşterii în cadrul instituţiilor de învăţământul superior. Posibilităţi de îmbunătăţire a managementului universitar prin data mining*, în *Societatea cunoaşterii*, coord. Roşca I., Editura Economică, 2006.

[11] Qasem A. Al-Radaideh, Emad M. Al-Shawakfa, and Mustafa I. Al-Najjar, Mining Student *Data Using Decision Trees*, The 2006 International Arab Conference on Information Technology (ACIT'2006), 2006

[12] Tudor, I., Cărbureanu, M., *Tehnici de data mining în managementul cunoaşterii într-o universitate*, în Managementul cunoaşterii în universitatea modernă, coord.: Bodea C.N. şi Andone I., Editura ASE Bucureşti, 2007, p.293.

[13] http://www.cs.waikato.ac.nz/~ml/weka/, accesat pe 13 ianuarie 2010