

USING DATA MINING METHODS IN KNOWLEDGE MANAGEMENT IN EDUCATIONAL FIELD

Cristina Oprea¹, Marian Zaharia¹

Petroleum&Gas University¹, Ploiesti, oprea_cris2005@yahoo.com,
marianzaharia53@gmail.com

Keywords: educational data mining, ID3, CART, Random Forest, Nayve Bayes, MultilayerPerceptron

Abstract: In this paper we analyzed the performance of first year university graduate. The study was conducted at the Faculty of Economics of the Petroleum&Gas University of Ploiesti. We have targeted a number of 379 masters from the specialties "Microeconomic Systems Management" and "Public Sector Management". We applied the most commonly used classification algorithms and compared three algorithms in the category of decision trees (ID3, CART, Random Forest), a classification algorithm bayesian (Nayve Bayes) and neural networks (MultilayerPerceptron). Study results indicate that the ID3 algorithm and Random Forest algorithm, provides the highest accuracy and correctly classified 342 of 379 instances. The knowledge discovered by applying the best methods of classification have shown that, for the prediction of performance of graduates, the relevant characteristics are: graduate profile, optional courses followed, master age, scores on admission and the number of failed exams.

1. INTRODUCTION

In the recent years, universities are increasingly concerned with the higher education quality and the degree of public trust in the different types, levels and educational qualifications. Data mining techniques are suitable for marketing applications at the level of university and analysis on how to improve the quality of educational process.

Data mining refers to a set of techniques that are in continuous development and can be successfully used to extract useful information and knowledge from massive volumes of data, hitherto unknown information. In recent years there has been an increasing interest in data mining and education, making educational data mining software a new area of research.

Educational data mining is the process of transformation of data that come from educational systems into useful information that can contribute at the improvement of educational system. This is defined as being the scientific research domain focused on developing methods to discover unique data types that come from educational environments and on using these methods for understanding better the students, their behaviour and their learning habits. [8].

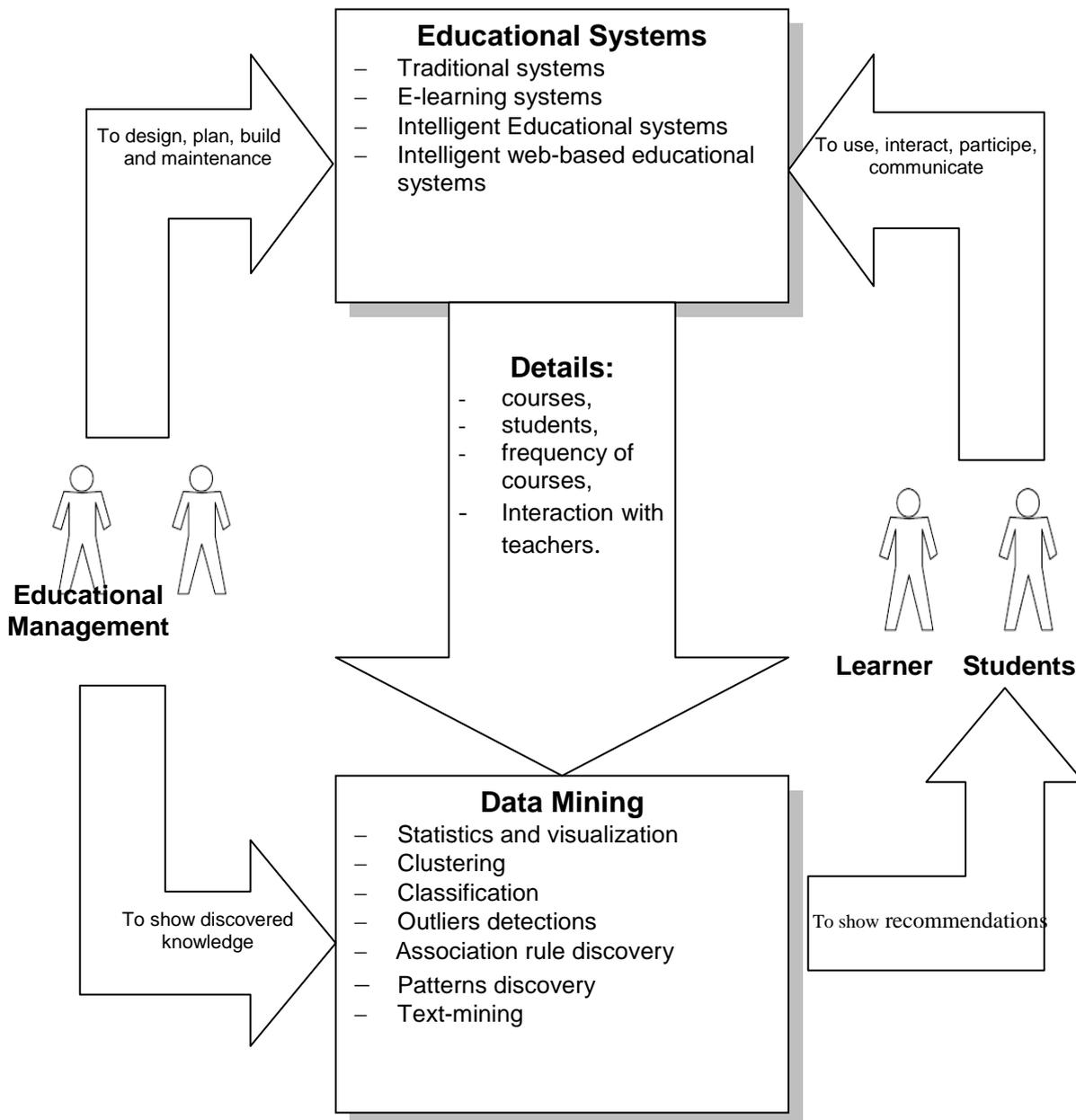
2. METHODS AND TECHNIQUES OF DATA MINING IN EDUCATION DOMAIN

Applying data mining techniques in the educational system is an iterative cycle of hypothesis formation, testing and refinement. Extracted and analyzed knowledge should enter in the system cycle and guide, facilitate and enhance the educational process as a whole. [9].

According to specific literature, data mining techniques are divided into two main categories:

- The category of *classical data mining techniques* incorporates the following techniques:
 - K-Nearest Neighbour classifier;
 - Clustering and
 - Statistics methods.
- The category of next-generation data mining techniques incorporates techniques such as:

- Bayesian classifiers;
- Decision tree;
- Neural networks;
- Association Rule;
- Genetic algorithms and
- Logistic regression



Sours: Adapted from Romero, Ventura 2006

Figure 1 The cycle of applying data mining in educational systems

Regarding the application of data mining techniques in education, the tasks performed by data mining can be grouped into the following categories:

- *Classification and regression* create models for predicting membership in a set of classes or values. Decision trees, Bayesian techniques, neural networks and K-nearest neighbor.

- *Clustering* seeks to identify a set of categories (clusters) to describe data. The purpose of clustering is to find different groups, whose members are similar components.
- *Association rules* seek to detect connections between data based on association or discovered sequence

Table 1 Applications of educational data mining in superior education in association with methods that can be used

Application of data mining in superior education	Methods that can be used	Algorithms that can be used
<ul style="list-style-type: none"> - Students classification based on their performances - Students performance estimation (obtained credits, grades); - Models identification regarding future students evolution (further studies: master, doctorate); - Detection of factors that contributed at the high study performance - Estimation of the correlation degree between the studied specialization and closed professional route . 	<p style="text-align: center;"><i>Classification and Prediction</i></p> <ul style="list-style-type: none"> - Decision and classification trees - Bayesian classifiers - Neural Networks - K-nearest neighbor classifier - Regression - Rule-based methods - Rough sets 	<ul style="list-style-type: none"> - ID3, C4.5 și C5.0, CART, SPRINT, THAID, CHAID - Naive Bayes, BayesNet - Single-Layer Perceptron, Multy-Layer Perceptron, RBF Network, SVM - K-NN, PEBLS - Linear Regresion, Simple Logistic - RIPPER, CN2, Holte's 1R, C4.5)
<ul style="list-style-type: none"> - Identification of student profile with graduation chances and with abandon studies chances. - Identification of the profile of the students care are tented to move to another university; - Identification of students typologies that do not make the object of motilities - Identification of the profile of the students the most credits 	<p style="text-align: center;"><i>Clustering</i></p>	<ul style="list-style-type: none"> - K-means - Kohonen networks
<ul style="list-style-type: none"> - Identification of classes that are preferred by students; - Identification of classes that are usually requested together - Identification of desired specializations by students. 	<p style="text-align: center;"><i>Association</i></p> <ul style="list-style-type: none"> - Association rule discovery - Patterns discovery 	<ul style="list-style-type: none"> - APRIORI, GRI, CARMA - Capri

Using data mining techniques in higher education is a relatively new area of research. Therefore, the implementation methodology is not transparent and is not yet clear what algorithms are preferred in this context. Studies conducted between 2000 and 2010 (for example, [2], [4], [5]) are still limited and is difficult to determine which approach should be favored or what methods should be used to analyze student performance.

Table 1 shows a series of applications of data mining in superior education, in combination with methods that can be used.

3. CASE STUDY:

In this paper we tried to analyze the performance of master students at the end of the last year. The study was made at the Economics sciences faculty from Petroleum&Gas University from Ploiesti, specializations “Management of microeconomics systems” and “Management of public sector” and had a number of 379 students. I identified a number of 12 attributes whose description was made in table 2.

Table 1 Attributes description and source of data analyzed

The Attribute	The description	The source
Gender	Master student gender	Current date base of UPG
Environment	Environment of origin of the master student: urban and rural	Admission database
Citizenship	Master student citizenship: Romanian or foreign.	Admission database
Age	Master student age	Current database of UPG
Specialization	Area of specialization	Current database of UPG
Financial form	Master student financial form	Current database of UPG
Average entrance	The average for admission to the Master	Admission database
Graduate_profil	High school study profile	Student questionnaire
Facultative course	The number of facultative courses followed by each master student	Current database of UPG
Job	If the master student have a job	Student questionnaire
Remaining_exames	The number of exams not passed	Current database of UPG
Performance	The average obtained for first year of study of each master student	Current database of UPG

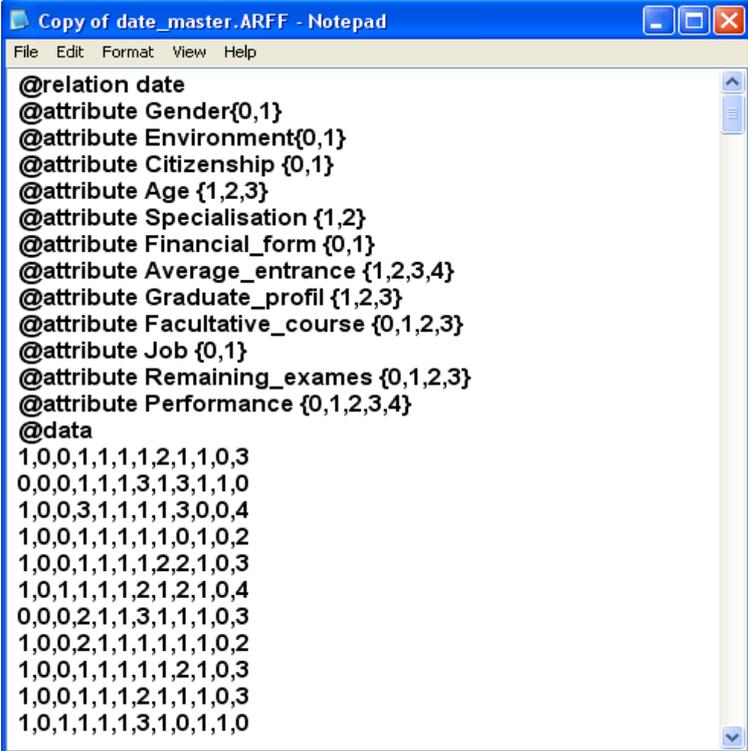
In the Petroleum-Gas University of Ploiești we identified two sources of data that provides information about candidates and students, namely:

- The database made annually by the admission committees of Petroleum-Gas University of Ploiești, where are entered the information about candidates in the entrance examination. The information is entered into the database via a form to be completed enrollment on a file.
- Current database of the university. The information in this database are placed throughout the academic year.

These two sources plus a questionnaire was sent to students

As the average data mining was used the platform Weka (the Waikato Environment for Knowledge Analysis), an application developed in Java under the GNU specialized department of the University of Waikato in Hamilton, New Zealand.

Whereas it used WEKA ARFF format data as input, the database was converted from Excel in CSV (comma-separated values), then was added header containing the description attribute types and their values (see Figure 2) .



```
@relation date
@attribute Gender{0,1}
@attribute Environment{0,1}
@attribute Citizenship {0,1}
@attribute Age {1,2,3}
@attribute Specialisation {1,2}
@attribute Financial_form {0,1}
@attribute Average_entrance {1,2,3,4}
@attribute Graduate_profil {1,2,3}
@attribute Facultative_course {0,1,2,3}
@attribute Job {0,1}
@attribute Remaining_exames {0,1,2,3}
@attribute Performance {0,1,2,3,4}
@data
1,0,0,1,1,1,1,2,1,1,0,3
0,0,0,1,1,1,3,1,3,1,1,0
1,0,0,3,1,1,1,1,3,0,0,4
1,0,0,1,1,1,1,1,0,1,0,2
1,0,0,1,1,1,1,2,2,1,0,3
1,0,1,1,1,1,2,1,2,1,0,4
0,0,0,2,1,1,3,1,1,1,0,3
1,0,0,2,1,1,1,1,1,1,0,2
1,0,0,1,1,1,1,1,2,1,0,3
1,0,0,1,1,1,2,1,1,1,0,3
1,0,1,1,1,1,3,1,0,1,1,0
```

Figure 2 Source file containing the input dataset

We applied the most commonly used classification algorithms and compared three algorithms in the category of decision trees (ID3, CART, Random Forest), a classification Bayesian algorithm (Nayve Bayes) and neural networks (MultilayerPerceptron).

Classification and decision trees algorithms made the classification of court by covering the tree root node to leaf nodes. It starts from the root, its attribute testing, then through each tree according to attribute values of given data set.

Nayve Bayesian technique is a method of classification as potentially predictive, as well as descriptive. It allows analyzing the relationship between each independent variable and dependent variable, by calculating a conditional probability for each of these relationships. When a new instance is intended to be classified, the prediction is made by combining the effects of independent variables on the dependent variable.

Neural networks provide effective means for modeling complex and large problems. A neural network includes input layer, where each node corresponds to a predictive variable. Input nodes can be connected to a number of nodes in the hidden layer. Nodes of hidden layer nodes can be connected to another hidden layer or output layer. The latter consists of one or more response variables.

The performance of classification is evaluated using indicators Precision Correctly Classified Instances and Incorrectly Classified Instances (see table 3).

The number of correctly predicted values of the total number of predicted values is indicated by the Precision parameter that takes values between 0 and 1. Accuracy equal to 0 indicates that the model has no predictive power, is not conclusive.

Correctly Classified Instances is the total number of instances classified correctly, while Incorrectly Classified Instances is the total number of instances incorrectly classified.

Table 2 Accuracy of classification

Algorithm utilized Accuracy	ID3	CART	Random Forest	Naive Bayes	Multilayer Perception
Precision	0,906	0,534	0,912	0,691	0,892
Correctly Classified Instances	90,237%	73,087%	90,237%	73,614%	88,1%
Incorrectly Classified Instances	9,762%	26,912%	9,7625%	26,385%	11,9%

4. CONCLUSIONS

Study results indicate that the ID3 algorithm and Random Forest algorithm, provides the highest accuracy and correctly classified 342 of 379 instances. The knowledge discovered by applying the best methods of classification have shown that, for the prediction of performance of graduates, the relevant characteristics are: graduate profile, optional courses followed, master age, scores on admission and the number of failed exams.

References

1. Bodea, V., Roșca, I.Gh., - „Analiza performanțelor studenților cu tehnici de Data Mining: studiu de caz în Academia de Studii Economice București”, în Managementul cunoașterii în universitatea modernă, coord.: Bodea C.N. și Andone I., Editura ASE București, 2007, p.320.
2. 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
3. Gorunesu, F., - Data Mining. Concepte, modele și tehnici, Editura Albastră, 2006, Cluj-Napoca
4. Halees, Alaa El, „Mining Students Data To Analyze Learning Behavior:A Case Study”, The International Arab Conference on Information Technology (ACIT), 2008
a. <http://www.ijcaonline.org/archives/number22>
5. Herzog, S. Measuring determinants of student return vs. dropout/stopout vs. transfer: A first-to-second year analysis of new freshmen. In Proc. of 44th Annual Forum of the Association for Institutional Research (AIR), 2004.
6. Oprea Cristina, Zaharia, M., Gogonea, M., “ Analysis of student performance to license exam using data mining techniques”, The 14th IBIMA Conference on Global Business Transformation through Innovation and Knowledge Management, Istanbul, 2010, ISBN: 978-0-9821489-3-8
7. Oprea Cristina, Zaharia, M., Enăchescu, D., „Knowledge discovery using data mining techniques: A case study”, Annals of the Oradea University. Fascicle of Management and Technological Engineering vol. IX (XIX), 2010, pg.543 - 550 Editura Universitatii din Oradea 2010, ISSN 1583 - 0691
8. Ryan S.J.d. Baker, “Data Mining for Education”, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA, 2008
9. Romero, C., Ventura C., „Educational data mining: A survey from 1995 to 2005”, Expert Systems with Applications: An International Journal Volume 33 Issue 1, July, 2007
10. Quinlan, J.R. - Machine Learning, Induction of decision trees, vol.1, Springer Netherlands, 1986
11. Witten, I.H., Frank, E., - Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann Publishers, San Francisco, 2000