STUDENT DROPOUT ANALYSIS WITH APPLICATION OF DATA MINING METHODS

Mario Jadrić[*]
Željko Garača[**]
Maja Ćukušić[***]

One of the indicators of potential problems in the higher education system may be a large number of student dropouts in the junior years. An analysis of the existing transaction data provides the information on students that will allow the definition of the key processes that have to be adapted in order to enhance the efficiency of studying. To understand better the problem of dropouts, the data are processed by the application of data mining methods: logistic regression, decision trees and neural networks. The models are built according to the SEMMA methodology and then compared to select the one which best predicts the student dropout. This paper concentrates primarily to the application of the data mining method in area of higher education, in which such methods have not been applied yet. In addition, a model, useful for strategic planning of additional mechanisms to improve the efficiency of studying, is also suggested.

1. INTRODUCTION

Higher education is the essential element of knowledge based economy: universities are nowadays the most important source of basic research and are therefore crucial for the development of new technologies (Enders, 2001; Garača, 2007). By analysing the relation between higher education and society in Central, Southern, and Eastern Europe, numerous experts in higher education

---

[*] Mario Jadrić, BSc, Faculty of Economics Split, Matice hrvatske 31, 21000 Split, Croatia, Phone: +385 21 430 739, E-mail: jadric@efst.hr

[**] Željko Garača, PhD, Faculty of Economics Split, Matice hrvatske 31, 21000 Split, Croatia, Phone: +385 21 430 654, E-mail: garaca@efst.hr

[***] Maja Ćukušić, MSc, Faculty of Economics Split, Matice hrvatske 31, 21000 Split, Croatia, Phone: +385 21 430 758, E-mail: mcukusic@efst.hr

(13 of them in Pausits and Pellert, 2007) call for revision of the higher education system, especially in terms of institution management strategy.

Due to an increasing number of students and institutions, higher education institutions (HEIs) are becoming increasingly oriented to performances and their measurement and are accordingly setting goals and developing strategies for their achievements (Al-Hawaj, Elali and Twizell, 2008; Deem, Hillyard and Reed, 2007; Pausits and Pellert, 2007). The interest for performance indicators in the higher education sector has become extremely high in Europe (GFME, 2008; McKelvey and Holmén, 2009) and thus also in Croatia (NCVVO, 2009; Vašiček, Budimir and Letinić, 2007). The reason for this lies in the relevant political and social changes in the recent years (Orsingher, 2006; Al-Hawaj, Elali and Twizell, 2008; Knust and Hanft, 2009):

- In Europe, the government is progressively retreating from its position as university financier. Therefore, HEIs have to try and develop new ways of attracting students and finance.
- On the other hand, having allowed higher institutional autonomy for HEIs the government requires more transparency and responsibility from them.
- The process is also affected by a number of external factors such as the labour market, changes in the European higher education, and increased relevance of research for the society as a whole.

Consequently, it may be stated that the pronounced interest in the way of HEI functioning is primarily due to the need for useful information for presumptive students and their parents, the need for comparability of institutions in terms of curricula and their performance, and justification of government funds expended on higher education. Globalization in higher education entails increasing competition for students, faculty and financial resources. The EU member countries are currently focused on the fast adjustment of their curricula and education processes (Knust and Hanft, 2009). This is, in turn, stimulating higher mobility of students and instructors. The governments have to care about brain drain, research outcomes, higher education costs, and availability of education for all citizens, while financial resources play a significant role for almost any decision in higher education (Michael and Kretovics, 2005).

In the light of these new challenges, quality assurance is for many HEIs the main tool of planning, management, and control. Transparency, responsibility, legitimacy, and comparability between various European qualifications are only

some of the quality assurance process outcomes (Orsingher, 2006; Al-Hawaj, Elali and Twizell, 2008; Knust and Hanft, 2009). In some countries, quality assurance is the internal responsibility of each HEI based on the internal evaluation of its program, while in some other countries quality assurance implies external evaluation and accreditation (Deem, Hillyard and Reed, 2007).

An indicator of potential weaknesses in the higher education system may be a large number of dropouts in the first years of studies. The strategic goal of HEIs should therefore be planning, management and control of education processes with the purpose of improving the efficiency of studying. The dropout trends have to be recognized and the causes (course, previous knowledge, assessment) isolated. Also, the typical dropout student profile is to be determined in order to plan the number of potential students in lifelong learning programs or those that need additional motivation (Vranić, Pintar and Skočir, 2007). It is possible to follow the dropout trend throughout several years in order to check the effectiveness of corrective activities.

The Faculty of Economics in Split is the second largest higher education institution in the field of economics in Croatia. Its operation involves all education levels from undergraduate, graduate, and postgraduate university programs to occupational college programs. An ever increasing competition, the requirements of the Ministry of Education, Science and Sport, and the imminent accession of Croatia to the EU position this institution within a new framework. By analysing the existing transaction data on students, the aim is to collect additional information and define the crucial processes that have to be adjusted for the purpose of improving studying efficiency.

The paper is structured as follows: research settings are presented in the second section with data extraction and transformation procedures and research methodology given as separate subsections; third section states research findings while fourth section concludes the paper.

2. RESEARCH SETTINGS

Modern information systems collect daily large quantities of various data from different domains and sources of various forms and contents. In practice, there is a need for methods, techniques, and tools that can search vast data quantities, recognize patterns and present them on the level of concrete reports. In such complex requirements, the classic analytical approach is not sufficient as it is difficult to set a general mathematical model. A wide range of tools for

collecting, storing, analysing and visualising data is defined as business intelligence (Michalewicz, Schmidt, Michalewicz and Chiriac, 2007).

To comprehend better the student dropout, statistical data processing will be performed and some data mining methods will be applied. In the first segment, graphs will be used to present the basic information on the structure of students and obtain directions for a detailed analysis of dropouts. In the second segment, the analysis will be carried out by use of logistic regression, decision trees, and neural networks. Models will be built according to the SEMMA methodology and compared to select the one which best predicts the student dropout.

For pre-processing, the SQL language is used to perform a query over Sybase database, while particular data grouping are carried out in Microsoft Office Excel 2003. SPSS 13.0 is used for designing clustered bar graphs, while data mining is conducted in SAS 9.1 Enterprise Miner. The software choice is SAS, which in the area of business intelligence dominates in advanced analytical solutions.

2.1. Data extraction and transformation

Transaction data on students are collected through the Faculty of Economics Information System (ISEF) within the autonomous subsystem ISEF_SS (Student service). ISEF stores data in the Sybase database. There are two databases: the central database and the replicated database for the web. Updating of the replicated database is carried out when required, while the essential data, such as marks, are updated immediately. The system is based on the desktop and web components. The desktop section integrates all functions, while the web section is used within the system MojEFST (MyEFST), which contains personalized information for each student and instructor. Currently, there is no data warehouse, but it will be introduced in the near future to improve the reporting system and allow creation of ad-hoc reports. For web reporting, the replicated database is used while other reports are created from the central database. SQL and procedural programming are used for reporting at the moment.

To perform the analysis, the crucial tables in the database are STUDENT – the basic data on the student, UPISGO – data on the enrolled years, UPISPR – data on the courses enrolled in an academic year, and ISPIT – data on examinations. For this analysis, the following attributes are separated from the database: ID, Generation, Sex, Date of Birth, Status, Study Program¸ Points

obtained from the secondary school, Enrolment Rank, Father Qualifications, Mother Qualifications, Social Status, Housing Indicator, Secondary School, Last year of study, and Last year of enrolment. Besides these data collected at the enrolment, the analysis also includes attributes referring to the studying process. The number of exam takings, signature, and marks are selected for the following courses: Introduction to Economics, Information Technologies, Mathematics, Statistics, Mathematics in Economics, Microeconomics 1, Accounting, and Statistical Analysis. The analysis is carried out on the sample of 715 students.

Students are grouped into categories in terms of their rank at the enrolment and the school they attended. Based on the performed analysis in SPSS, graphs are selected which in the best way describe the structure of students and its correlation with the dropout. Table 1 shows grouped secondary schools. According to the stated categories, Table 1 allows the interpretation of Figure 1 which presents the student dropouts in terms of the grouped schools. It is obvious that after the first and the second year of study, most dropouts occur in category 3, i.e. students that have attended vocational schools (except commercial schools).

Table 1. Grouped secondary schools

Category number	School category
1	High schools in Split
2	High schools from other towns
3	Vocational schools in Split
4	Vocational schools from other towns
5	Commercial schools in Split
6	Commercial schools from other towns

It is assumed that after the first year students drop out of their own free will (due to different reasons), while the students dropping out after the second year mostly give up due to the exam failure. Voluntary dropout after the first year is lowest for the students that have attended commercial schools away from Split. Such students are more motivated because they have come to Split mostly because of their studies. The graph shows that after the first year the students who have attended commercial schools drop out less than the students who have attended high schools. This relation changes after the second year of study when more students who have attended commercial schools drop out due to the failure to pass the exams.
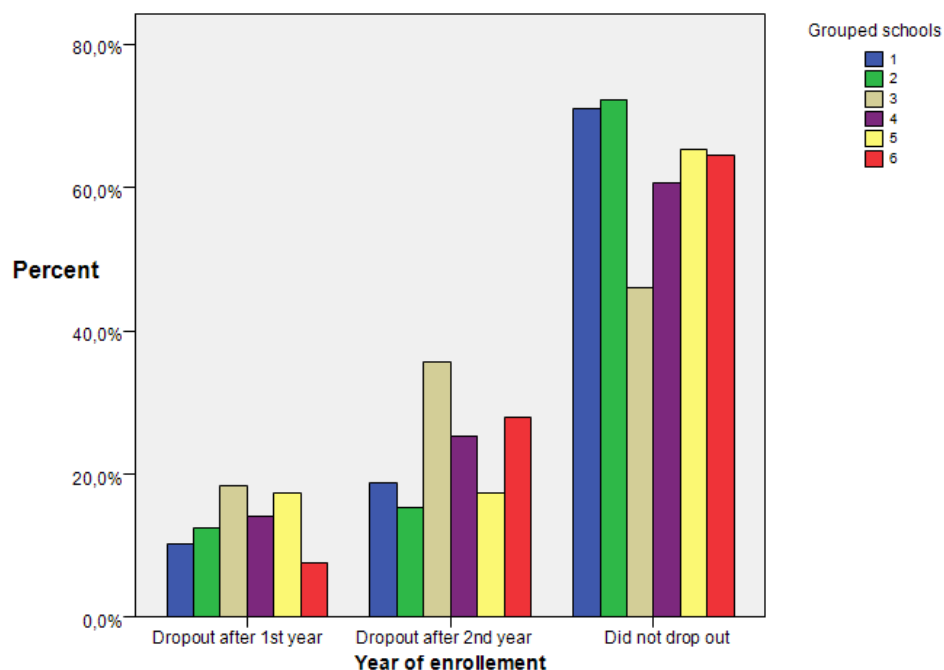
Figure 1. Students grouped by schools and their dropout

Students are also grouped in categories according to their rating obtained at the entrance examination. The way in which the grouping is conducted is shown in Table 2, while the distribution of students grouped in terms of entrance ranking and their dropout is illustrated in Figure 2.

Table 2. Grouping in terms of entrance examination rank

Program – Business Economics		Program – Economics, Tourism	
1-100	A	1-20	A
101-250	B	21-50	B
201 and below	C	51 and below	C
No entrance exam	D	No entrance exam	D

Figure 2 confirms the thesis that students who are better ranked at the entrance examination drop out less. Figures 1 and 2 show the correlation of the entrance examination rank, the secondary school attended and the dropout. When setting the goals of analysis, it was stressed that the dropout causes have to be isolated. Therefore, other elements also have to be investigated that could

affect the dropout. It is assumed that in addition to the previous knowledge, success achieved in particular courses affects the dropout to the greatest extent.
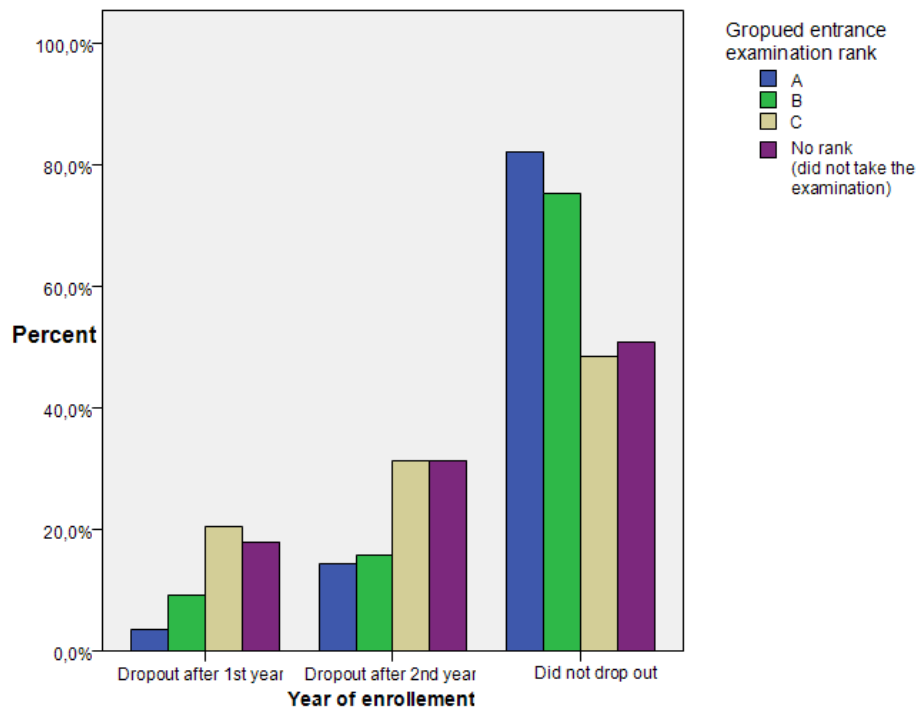

Figure 2. Students grouped in terms of entrance ranking and their dropout

To investigate the dropout predictors and create the best possible models, the data mining methods will be used.

2.2. Research methodology

A number of analytical methods are used for data mining. The standard types include regression (normal regression for prediction, logistic regression for classification), neural networks and decision trees (Olson and Delen, 2008). Different data mining methodologies show that the set of activities performed by the analyst can be presented as a series of logical steps or tasks. SEMMA (Sample, Explore, Modify, Model, Assess) was developed by the SAS Institute which is also the producer of the data mining platform that uses the same methodology - SAS Enterprise Miner (SAS Institute, 2004; Matignon, 2007).

The acronym SEMMA signifies: Sampling, Exploring, Modifying, Modelling and Assessment. Starting with the statistically representative data sample, SEMMA allows the application of statistical and visualisation techniques, selection and transformation of most significant predictor variables, modelling of variables to predict output, and eventually confirmation of model credibility.

To create classification models of neural networks, decision trees and logistic regression, it is necessary to follow the steps of SEMMA methodology. The model, developed in SAS Enterprise Miner, based on SEMMA methodology, is illustrated by Figure 3.
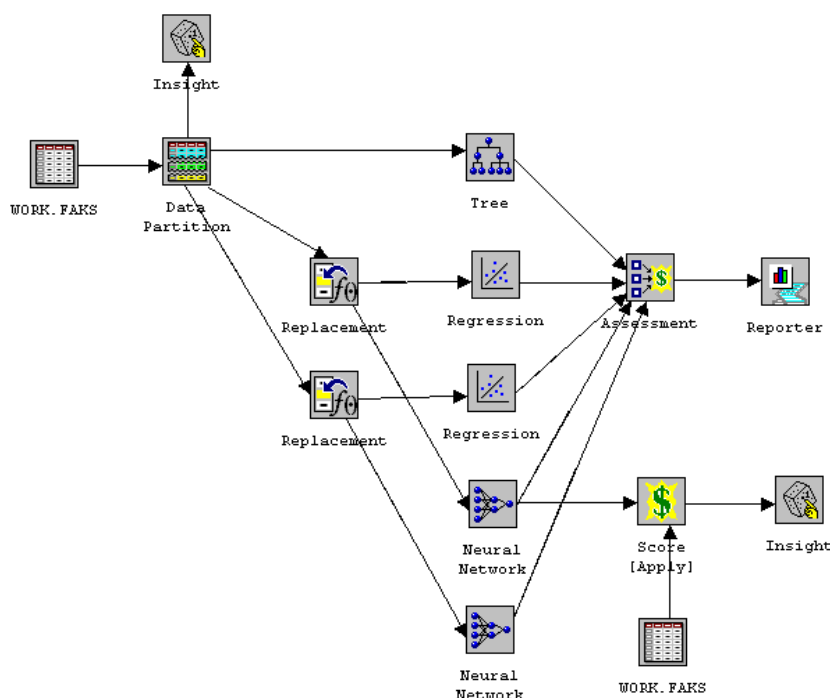

Figure 3. Model developed in SAS Enterprise miner based on SEMMA methodology

First, the input data are imported from the excel database and selected attributes that are analysed in more detail. The dropout attribute is marked as the target variable that can assume two conditions: 1 for the students who drop out and 2 for the students who do not. The selected input variables are all those that are mentioned in the section on data extraction. In the Sample step, besides defining the role of variables in the model, it is also necessary to define one of the measuring scales (nominal, ordinal, interval). This step also allows the

38

testing of distribution for each variable. For example, Figure 4 shows the distribution of students for the target variable "dropout". The possibility to explore descriptive statistics is also available so that the percentage of missing values can be determined; and for the interval variables, it is possible to determine the minimal, maximal, and medium values, standard deviation, etc.
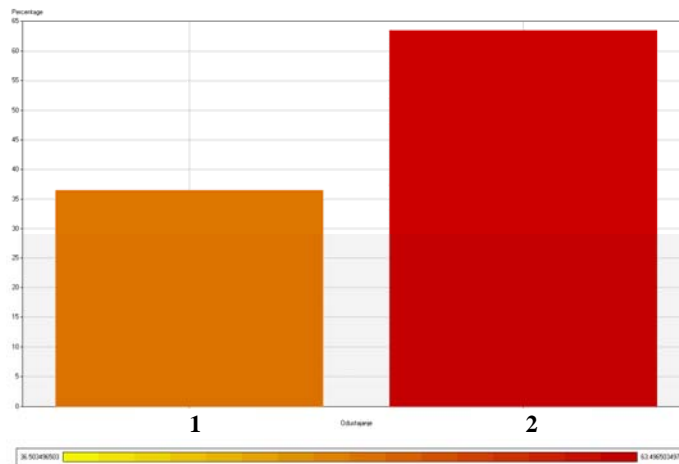

Figure 4. Distribution of students for the target variable dropout

Enterprise Miner in the Data partition node samples the input data and distributes them into training, validation and testing data (SAS Institute, 2003). The chosen method is Simple Random, in which each observation has the same possibility to be sampled.

In the step Explore, the node Insight is included. According to the settings for this analysis node, the sample size is 2000, and if it is less than that as in this case, the entire set of data is included in the analysis. In this example, 286 rows are selected for training, or 40% of the entire data set. This node offers numerous possibilities of data visualisation. From the step Modify, according to the SEMMA methodology, two Replacement nodes are inserted to explore the effect of various data replacement methods on the logistic regression and neural networks models. In the upper node Replacement in Figure 3, the missing values replacement method is selected by Mode, or the most frequent values, while in the lower node, the selected method is Tree replacement which adjusts replacement values by using the decision tree.

Decision trees are a highly flexible modelling technique. For instance, to build regression models and neural networks models, the missing values have to be inserted into training data while decision trees can be built even with missing values. Decision trees are intended for the classification of attributes regarding the given target variable (Panian and Klepac, 2003). Decision trees are attractive because they offer, in comparison to neural networks, data models in readable, comprehensible form – in fact, in the form of rules. They are used not only for classification but also for prediction (Gamberger and Šmuc, 2001).

Logistic regression is an analysis of asymmetric relations between two variable sets of which one has the predictor status and the other criterion status (Halmi, 2003). The dependent variable is dichotomous and marked by values 0 and 1, while the independent variables in logistic regression may be categorical or continuous (Hair, Anderson and Babin, 2009).

Neural networks behave very well in more complex classification problems. Their disadvantage, in comparison to simpler methods, is the relatively slow and demanding process of model "learning" (optimization of weight factors) (Gamberger and Šmuc, 2001). Neural networks are a powerful tool in trend prognostics and predictions based on historical data. In data mining, neural networks are often combined with other methods because if used alone, they can hardly guarantee a good interpretation of results (Panian and Klepac, 2003).

3. RESEARCH FINDINGS

Application of basic statistical methods is used to study the student population. It is found that women drop out comparatively less than men, that students who have attended high schools drop out less in comparison to those who have attended other schools, and that students with better entrance ranking drop out less. To carry out a more detailed dropout analysis and separate the attributes which have the highest effect, decision trees are used.

The decision tree in Figure 5 shows that in the training sample, 98 students or 34.3% drop out, while 188 students or 65.7% continue their studies after the second year. Of the 98 that drop out, 79 have a failing mark in Mathematics. Nineteen students drop out due to other reasons: seven of them due to the failing mark in Statistics, even though they have a pass in Mathematics. Twelve students drop out even though they have a pass both in Mathematics and Statistics. Eight of them drop out because their mark in Introduction to

40

Economics is 2 (sufficient) or lower, while in four students, this mark is higher than 2.
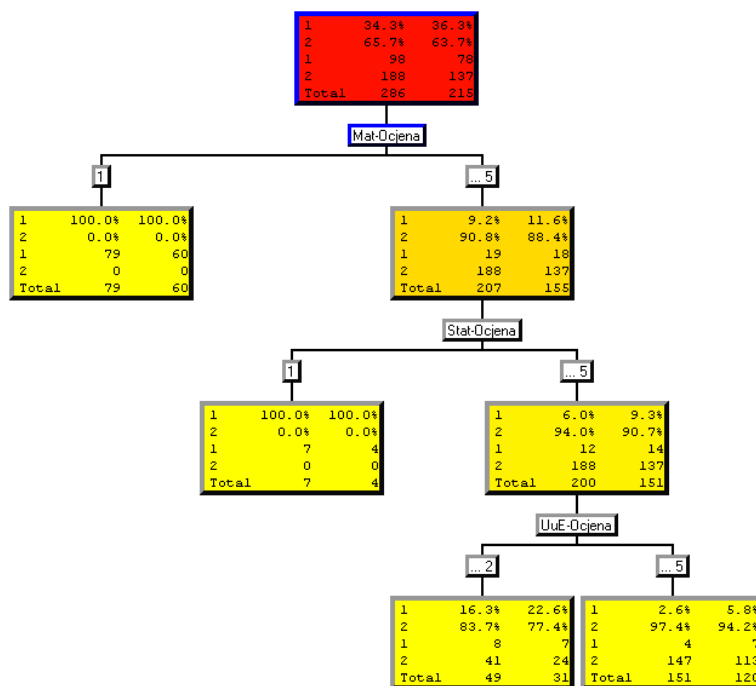

Figure 5. Analysis by decision tree

When including courses as input variables, only first semester courses are taken into account. In this way, the prediction model is obtained based on the students' previous knowledge and their performance in the very beginning of study. Such a model is more usable as it provides timely prediction on the number of dropouts and allows making decisions to improve the efficiency of studying. Despite the stated advantages of decision trees, neural networks are reputed to be extremely flexible for predictive modelling. To determine which data mining model and which technique will provide the best results, model evaluation is carried out. Missing values are inserted for logistic regression and neural networks models to allow the analysis on the same quantity of data. The reason for this is that in regression and neural networks such values are omitted, while in decision trees they are taken into account.

In the process of building different models, the number of variables is varied to achieve model reliability as high as possible. Eleven input variables are chosen: ID, sex, status, study program, father's qualifications, mother's

41

qualifications, social status, housing indicator, grouped entrance ranking, grouped secondary schools, and points from the secondary school, while the target variable is the dropout. The analysis does not take into account the effect of examination marks in order to state the reliability of models that predict dropouts based on variables that are not related to examination results. Figure 6 shows the evaluation and comparison of decision tree, logistic regression and neural networks models. The marks Neural and Reg denote neural network and logistic regression models which apply the method of missing values replacement called "Most frequent value", or Mode insertion, while the marks Neural-2 and Reg-2 denote models which apply the method of missing values replacement called "Tree imputation".
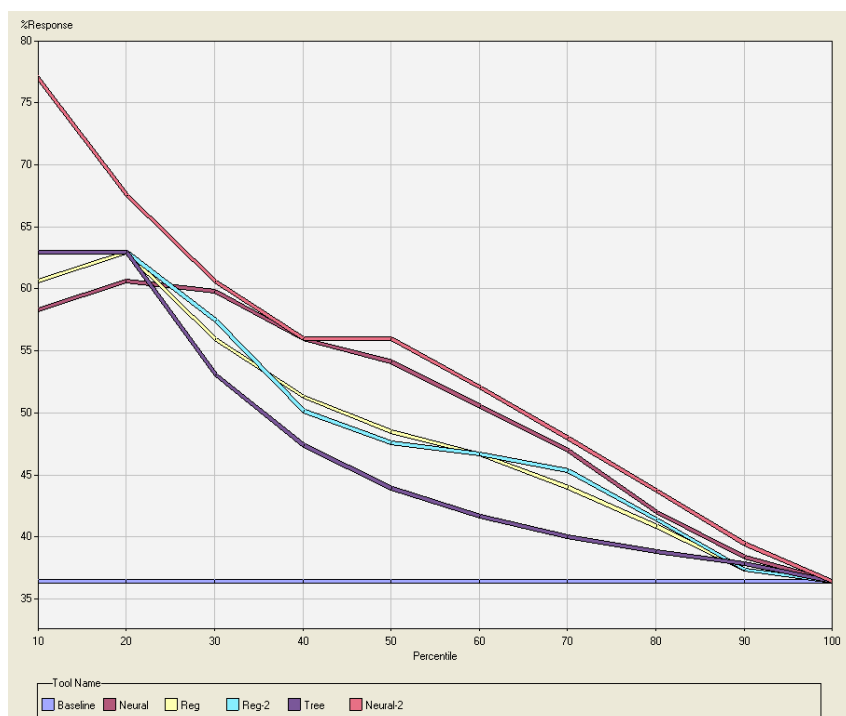

Figure 6. Evaluation and comparison of model I

The blue horizontal line in Figure 6 represents the basis or the 36% of students from the original population that drop out. The graph shows that the model Neural-2 is evaluated as the best one in comparison to all the other models (it is positioned upwards and on the right). Other models are more difficult to interpret due to their overlapping. Figure 6 also shows that neural

42

networks provide better results when using the missing values replacement method "Tree imputation" rather than "Most frequent value". The graph can be read in the following way: the curve Neural-2 shows that in the top 10% model scoring, or in the first decile, there are approximately 77% of students that will drop out. The neural network which has used the missing values replacement by "Most frequent value" in the first decile has about 56% of students that will drop out. The higher the percentile, the results of all models are getting closer to each other and are finally equalized, predicting that 36% of the students will drop out. In further analysis, the number of input variables is increased to include the effect of examination results to create models based on students' previous knowledge and their achievements in the very beginning of their studies.

From Figure 7, it is obvious that all the models are better adjusted in comparison to the models in Figure 6. The curve Neural-2, here, also provides the best results, but in the first top 30% of model scoring, the neural network and decision tree will provide the same results, i.e. 100% of students that will drop out.
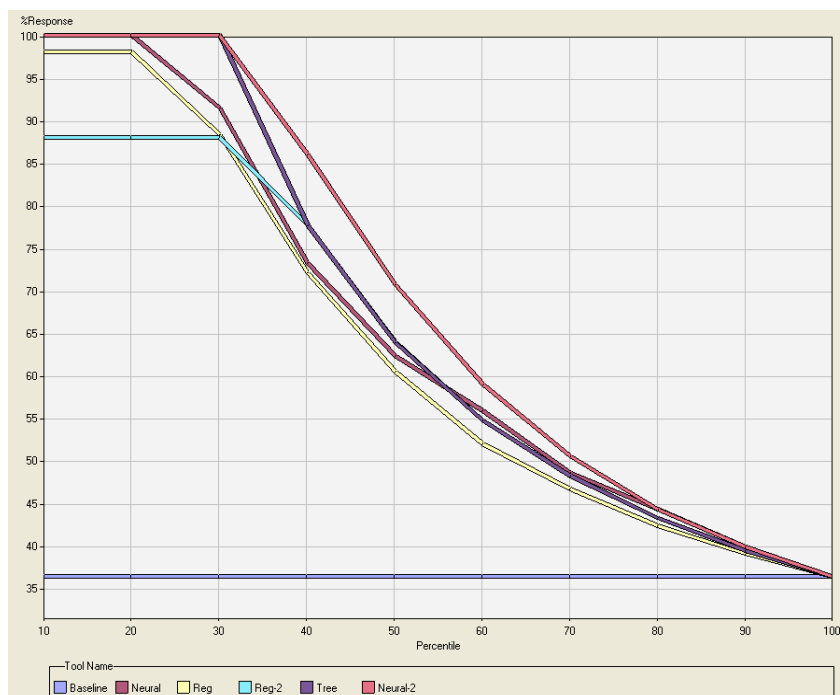

Figure 7. Evaluation and comparison of model II

4. CONCLUSION

This paper has explored the application of the data mining methods in higher education, where they are not usually applied. It is found out that this area abounds in unused data that, unfortunately, are not stored in an appropriate way. The analysis has separated the causes of students' dropout (e.g. previous knowledge, examination results). It has also determined the typical profile of the student inclined to drop out at the Faculty of Economics in Split. The obtained data should, in the earliest stage, be used to raise awareness on the possibilities and need to use the data mining models and methods at the institution in which this research has been carried out. The planned construction of data warehouse will allow support in strategic decisions and monitoring of the dropout trend. All this will also support the Bologna process, which also aims at enhancing the efficiency of studying.

The future research will be directed towards the design of an applicative solution to allow observation and classification of each student at the Faculty of Economics in Split into a particular dropout category depending on his/her characteristics. Such a solution will consist of two basic segments: the first is the crisp module based on the examination results in the first semester, while the second is the fuzzy module based on certain attributes of students' previous knowledge. In this way, an example of the application of a hybrid business intelligence system will be provided that combines the classic conception of expert systems and the conception based on fuzzy logic. Such a system may also have a practical application in decision-making support to the faculty administration in the early prediction of dropouts.

REFERENCES

1. Al-Hawaj, A. Y., Elali, W., Twizell, E. H. (Ed.) (2008): *Higher Education in the Twenty-First Century: Issues and Challenges*, Taylor & Francis Group, London
2. Deem, R., Hillyard, S., Reed, M. (2007): *Knowledge, Higher Education, and the New Managerialism: The Changing Management of UK Universities*, Oxford University Press Inc., New York
3. Enders, J. (Ed.) (2001): *Academic Staff in Europe: Changing Contexts and Conditions*, Greenwood Press, Westport
4. Gamberger, D., Šmuc, T. (2001): Poslužitelj za analizu podataka [http://dms.irb.hr]. Zagreb, Hrvatska: Institut Rudjer Bošković, Laboratorij za informacijske sustave.
5. Garača, Ž. (2009): *ERP sustavi*, Ekonomski fakultet, Sveučilište u Splitu

6. GFME (2008): *The Global Management Education Landscape: Shaping the future of business schools*, Global Foundation for Management Education
7. Hair, J., Anderson, R., Babin, B. (2009): *Multivariate Data Analysis*, Prentice Hall
8. Halmi A. (2003): *Multivarijantna analiza u društvenim znanostima*, Alinea, Zagreb.
9. Klepac, G., Mršić, L. (2006): *Poslovna inteligencija kroz poslovne slučajeve*, Lider press
10. Knust, M., Hanft, A. (Ed.) (2009): *Continuing Higher Education and Lifelong Learning: An International Comparative Study on Structures, Organisation and Provisions*, Springer Science & Business Media, Heidelberg
11. Matignon, R. (2007): *Data Mining Using SAS Enterprise Miner TM*, John Wiley & Sons, Inc., Hoboken, New Jersey
12. McKelvey, M., Holmén, M. (Ed.) (2009): *Learning to Compete in European Universities: From Social Institution to Knowledge Business*, Edward Elgar Publishing, Inc., Massachusetts
13. Michael, S. O., Kretovics, M. A. (Ed.) (2005): *Financing Higher Education in a Global Market*, Algora Publishing, New York
14. Michalewicz, Z., Schmidt, M., Michalewicz, M., Chiriac, C. (2007): *Adaptive Business Intelligence*, Berlin Heidelberg, Springer-Verlag
15. NCVVO (2009): *Vodič za provedbu samovrjednovanja u osnovnim školama*, Nacionalni centar za vanjsko vrednovanje obrazovanja, Zagreb
16. Olson, D.L., Delen, D. (2008): *Advanced Data Mining Techniques*, Springer-Verlag, Berlin Heidelberg
17. Orsingher, Ch. (Ed.) (2006): *Assessing Quality in European Higher Education Institutions: Dissemination, Methods and Procedures*, Physica-Verlag: Springer, Heidelberg
18. Panian, Ž., Klepac, G. (2003): *Poslovna inteligencija*, Masmedia
19. Pausits, A., Pellert, A. (2007): *Higher Education Management and Development in Central, Southern and Eastern Europe*, WAXMANN Verlag, Munster
20. SAS Institute (2003): *Data Mining Using SAS Enterprise Miner, A Case Study Approach*, Second Edition, SAS Institute Inc., Cary
21. SAS Institute (2004): *Getting Started with SAS Enterprise Miner 4.3*, Second Edition, SAS Institute Inc., Cary
22. Vašiček, V., Budimir, V., Letinić, S. (2007): Pokazatelji uspješnosti u visokom obrazovanju, *Privredna kretanja i ekonomska politika, 17 (110):* str. 51 - 80.

23. Vranić, M., Pintar, D., Skočir, Z. (2007): The use of data mining in education environment, *Proceedings of the 9th International Conference on Telecommunications, ConTEL,* Zagreb, Croatia, pp. 243-250

ANALIZA ODUSTAJANJA OD STUDIJA PRIMJENOM METODA RUDARENJA PODATAKA

Sažetak

Jedan od indikatora potencijalnih problema u visokoškolskom sustavu može biti velik broj studenata koji odustaju od studija na nižim godinama. Analiza postojećih transakcijskih podataka pruža informacije o studentima, potrebne za definiranje ključnih procesa koje treba prilagoditi da bi se povećala učinkovitost studiranja. Radi boljeg razumijevanja problema odustajanja od studija, provedena je statistička obrada podataka te su na podatke primijenjene neke od metoda rudarenja podataka, i to: logistička regresija, stabla odlučivanja i neuralne mreže. Modeli su izrađeni prema SEMMA metodologiji, a zatim je izvršena usporedba i izbor modela koji najbolje predviđa odustajanje od studija. Ovaj se rad ponajprije koncentrira na istraživanje primjene metoda rudarenja podataka na područje visokog obrazovanja, u okviru kojeg one do sada nisu primjenjivane. Nadalje, predlaže se i model koristan u strateškom planiranju dodatnih mehanizama za povećanje učinkovitosti studiranja.