# Data Mining and Statistical Analyses for High Education Improvement

M. Vranić, D. Pintar and Z. Skočir
Faculty of Electrical Engineering and Computing,
University of Zagreb
Unska 3, Zagreb, Croatia
Telefon: 01-6129 686 Fax: 01-6129 616 E-mail: mihaela.vranic@fer.hr, damir.pintar@fer.hr, zoran.skocir@fer.hr

**Abstract - Knowledge creation is essential for process improvement in every type of environment. This is also the case in lecturing and grading students at universities. In this paper a visual representation of data is combined with statistical analysis in order to be used for analyzing objectivity of student testing. Predictive data mining is used for timely recognition of students who require additional attention. Regression analysis and decision trees are methods used for this student selection. Comparison of these two methods, as well as a simple method of score border is also given.**

## I. INTRODUCTION

Driver for research and analyses described in this paper are ongoing changes in education process at Faculty of electrical engineering and computing.

This facility, along with other faculties in Croatia, is going through process of transformation based on Bologna declaration. In that process, big changes were made in education syllabus. First of all, studying is now divided in two stages. After three years of studying and collecting a certain number of ECTS points, students achieve the title of baccalaureus. After additional two years academic title of master is granted.

For the new programme, classes have been significantly altered and rearranged. Way of teaching students has been changed along with procedures of examination.

Professors previously relied on personal teaching experience when judging the minimal knowledge required for passing an exam. Verbal (oral) examination was a pretty good way to control minimal knowledge acquired during the whole studying process. In the new programme, verbal examination has become a rarity, especially in the first years where large number of students has to be examined in the same time, in same conditions. Students are now required to continually study during the whole semester and their knowledge is examined in multiple and various ways, depending on class organization. Each examination mode (for example homework, short tests, larger tests during semester and final exams) brings some points to the student. For every course professors must set up the border in advance – minimal sum of points needed for passing the course. Setting up this border is not an easy job. On one side, collected points have to reflect acquired knowledge and large peaks showing number of students who gathered particular sum of points should not appear (Gaussian distribution is expected). On the other hand, border has to be high enough to guarantee sum minimal acquired knowledge. Additional encumbrances emerge with tests which can vary in complexity in great extent, and with students who try to pass class with minimal effort.

With all described changes, it is evident that additional ways of controlling student's knowledge is needed to maintain the quality of students finishing their education at the faculty. With certain collected data at hand, process of combination gives university the ability to create new knowledge about processes of tutoring and examination. As Nonaka states [1], creation of knowledge at any organization is a driver for improvements. How knowledge creation is interconnected with data mining process is described in [2] and [3]. Also, competitiveness and increasing business needs for educated staff in this area (electrical engineering and computing) are drivers for better dynamics of educational process [4].

The ability to recognize student's weaknesses in timely fashion and enable him better understanding of subject is always welcome. In that way students would finish the faculty education with broader knowledge base and it would be achieved in shorter time (in some cases appropriate and timely move toward student would prevent him from dropping from the class).

Business intelligence tools could be very useful for that purpose. Implementation of additional statistical analyses and data mining algorithms on collected data can enable further understandings and improvements. As stated in [5], control over subject taught and the way of knowledge transfer should be established. Educational staff is responsible for the whole process. This paper is written following those ideas.

Goals of research presented in this paper are:

- to improve examination objectivity
- to show how data mining can be used for better understanding of student behaviour and timely prediction of those students who are likely to drop the course.

Regarding social context of privacy and data mining, all delicate data is encoded.

Paper is structured as follows. Chapter II refines the problem with presentation of available data and data mining techniques. Section III has testing objectivity in focus and presents conducted analyses and results. In section IV predictive analysis is introduced. Descriptions of constructed models are given and results are compared. In section V possible improvements and look at future directions are presented. Finally, section VI concludes the paper.

## II. PROBLEM REFINING

### Available data

The data available for research was data concerning one particular course at the first semester of new programme. That course resulted from comprising two courses in the previous faculty programme. Subject matter was customized to suite the new requirements. New lecture materials were made. Examination methods were changed and oral exams were excluded. Besides the data about success at each examination mode and group that student belongs to for attending lectures, data about previous schooling, gender and enrolment score was also retrieved as they were considered useful for analysis. All available attributes are shown in Table I.

Data about two generations of students was included – 951 students in the academic year 2005./2006. and 855 students in the academic year 2006./2007.

Data about success is fully known for both populations. If some part of data is missing, it implies that student didn't attend specific testing.

General data (like unique identifier, name, surname and gender) is also known for all students. Data about enrolment score was available only for enrolment exams in years 2005. and 2006. so, as stated in Table I., for part of population the data wasn't available. That is also the case with previous schooling and county that high school belongs to. Mentioned missing data is connected with students who dropped out previous classes and had to or decided to follow new education Bologna program. Availability of data was the same for students transferred from other faculties. Large number of students in academic year 2005./2006. were students with status of non-regular (repeting course listening or transfers) because they deliberately chose to follow new programme. That has to be taken into account in further analysis and conclusions as well.

Final availability of data:

population of academic year 2005./2006.:

- previous schooling data: 72,3 %
- enrolment score: 72,3 %

population of academic year 2006./2007.:

- previous schooling data: 91,7 %
- enrolment score: 93,2 %

### Appropriate data mining techniques

Beside descriptive analysis that gives same general insights in new or already known regularities (described in [6]), predictive analysis must be used for timely recognition of possible dropouts. Due to the data types available and ability to interpret final results, regression analysis and decision trees were used.

There was no need for normalization or other transformations of achieved scores by students, because one point for homework is equally valued as one point achieved at midexam.

### Used tools

Data came from various different systems. It was cleared and transformed to the appropriate form using relational model (Oracle XE database was used).

For statistical analysis and data mining SAS 9.1 was used. It is a very powerful tool and enables user-friendly implementation of data mining algorithms through compounding the whole process from blocks. Parameters of each block could be tweaked easily to form the most suitable model.

TABLE I – AVAILABLE ATTRIBUTES

| **Attribute** | Abbreviation | Attribute type | Availability |
|---|---|---|---|
| LECTURE RELATED DATA | | | |
| Group | GROUP | nominal | full |
| Lecturer | LECT | | |
| Status | STATUS | | |
| Room for mid exam n | DV_MI_n | | |
| Grade | GRADE | | |
| Repetition of final exam | RepF | | |
| Mid exam n (n=1,2,3) | MI_n | numeric | |
| Final test | FT | | |
| Course Attending | CA =N | | |
| Homework | HW = DZ | | |
| Test n (n=1,2) | T_n | | |
| Total Score | TS = UK | | |
| GENERAL DATA | | | |
| Gender | GEN | nominal | full |
| State | STATE | | partial |
| High School | HS | | |
| High School Type | HS_T | | |
| Postal code | PBR | | |
| Unconditioned Enrollment | UnE | | |
| Enrollment Score | EnrS | numeric | |

## III. ANALYSIS OF TESTING OBJECTIVITY

Objectivity of testing and other ways of acquiring points is very important.

Of the most interest are short tests (three tests during semester valued 5 point each), midexams (three times during semester - valued 20 point each) and one final exam (valued 10 points). Midexams and final exam are held in the same time for the whole generation of students. On the other hand, students don't take short test simultaneously. Short test is taken by students belonging to different groups at the specific time in the test week.

### Short tests analysis

Emphasis given here is on

- the interrelation of real knowledge and test score
- acquired score through time.

Test should objectively reflect someone's acquired knowledge. That's why distribution similar to Gaussian distribution of collected points is expected.

Real distributions are presented in Figure 1. Histograms on the left present accomplishments of the first generation of students, while histograms on the right reflect success of the second generation. It is noticeable that scores are largely grouped at high border for first generation. It is clear that tests for second generation are much better. There was much broader base of test questions and variations in questions were introduced during testing week.

The most important characteristic of each test is that it should be objective, that it properly reflects acquired knowledge. Since no measure exist which would ideally reflect that knowledge, final grade could be used for comparison with scores at tests. It is expected that higher grade is associated with higher scores. Real situation is

presented at Figure 2. It makes clear that short tests for second generation were formed in a much better way.

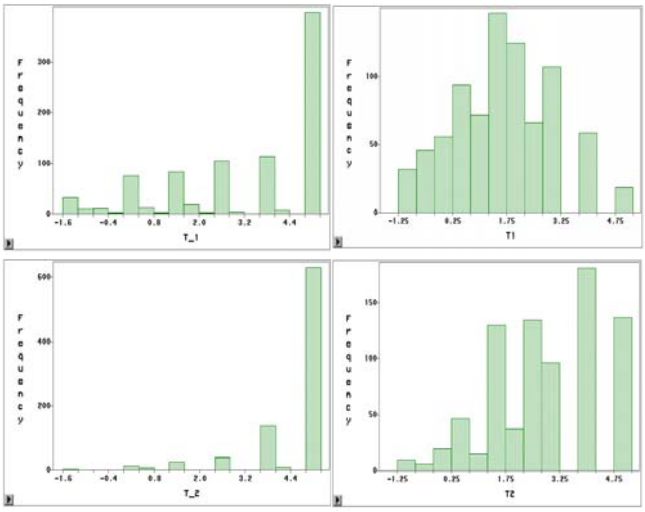


Figure 1: Histogram showing number of students acquiring certain level of points for first (left) and second generation (right) on first and second short test

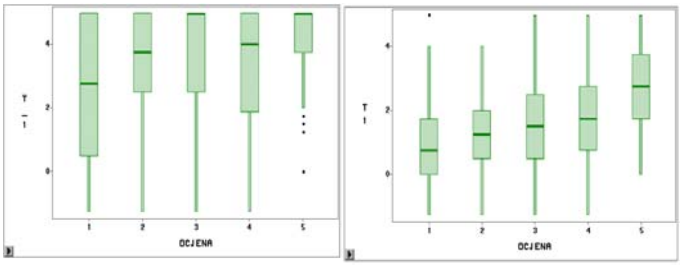


Figure 2: Box plot showing relationship between grade and points acquired at first short test for both generations

It is expected that there could be some irregularities caused by student communication about questions asked and that students taking test in later time would score better. Visual presentation of acquired points during week or even during one day could suggest some reasoning.

Time stamp connected with taking short tests was available only for second year population. Box plots can give good insight in trends. Figures 3 and 4 show how acquired scores changed over time – days of the week and hours of the day. Only two students took the test afterwards, and their score is not much better than the rest of the generation. No significant abnormalities could be noticed.

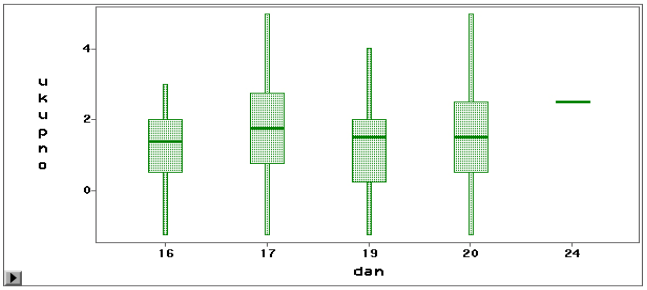


Figure 3: Box plots – scores acquired by students depending on the day of the week

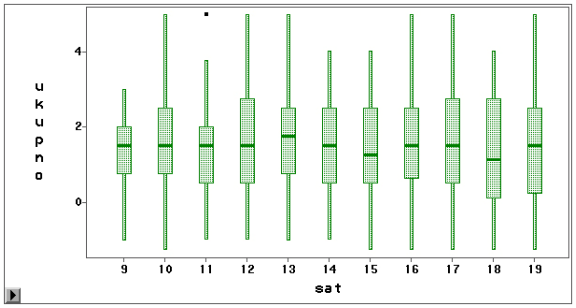


Figure 4: Box plots – scores acquired by students depending on the time of the day

**Midexam analysis**

Emphasis in this analysis is on conditions that students have when taking this type of exams.

Whole generation of students takes midexams and final exam at the same time. They previously don't have insight in asked questions, and they all have access to the same study materials. There are some differences between different groups taking classes held by different professors. However, presentation materials and covered topics are identical for all lecturers.

When conducting exams, authors noticed that room where the exam is held and alertness of person managing it could have impact on appearing abnormalities. Abnormalities include different concentration of students and possible cheating – copying someone else's answers.

Small abnormalities statistically could not be recognized. However, box plots reflecting students score (Figure 5) for each room reveal that there could have been abnormalities at larger scale. After inspection of few graphs, it is noticed that students in room C12 gathered much higher scores than the rest of the generation (second midexam, academic year 2005./2006.).

Some statistical tests could be used here although not all conditions are satisfied. Students are not randomly spread through different rooms. Their allocation depends on alphabetical order of their surname – which should not be related to their acquired knowledge and their ability to present it. That's why test will be made as it was random generation sample. Of course, final results (because all assumptions are not fully satisfied) could not be regarded as final proof for some irregularities but could suggest directions for improving test managing.

Sample distributions show some regularities. In case of random sample, independent observations and number of sample observations not exceed 10% of population, sample distributions show certain characteristics. Those characteristics should be applicable to students from room C12.

One-sample t test will be used. Certain neighbourhood of sample mean should include known mean of the whole population.

Hypotheses:

Null: *There is no significant difference between the sample mean and the population mean.*

Alternate: *There is a significant difference between the sample mean and the population mean.*

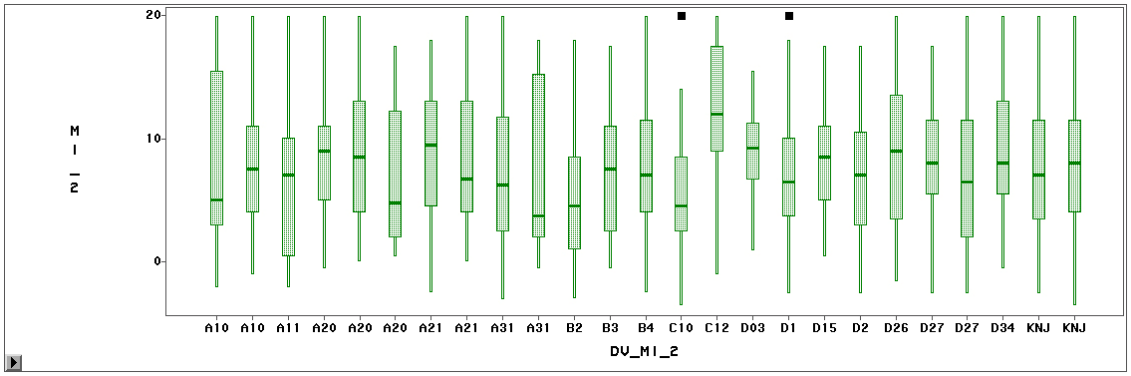


Figure 5: Box plots reflecting gathered points for students taking test in different rooms for second midexam - academic year 2005./2006.

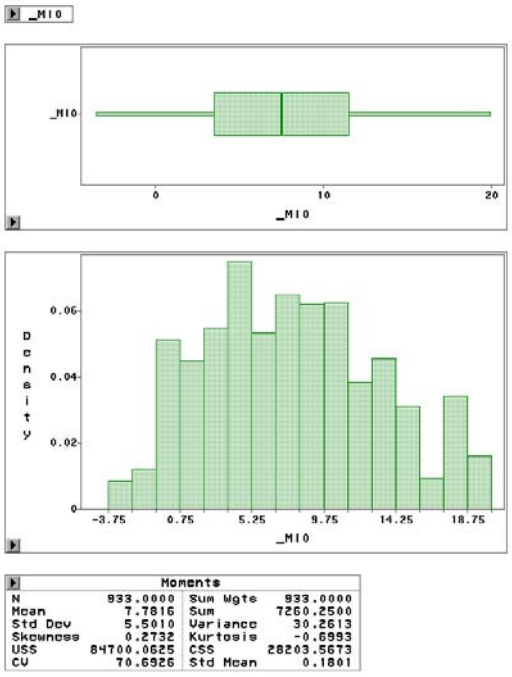


Figure 6: Distribution for score at second midexam – whole population

95% confidence interval was chosen. Known population mean is: 7.78. Distribution for whole population is shown in Figure 6. Sample mean is 11,69 and there were 37 students in room C12.

Results of one-sample t test are shown in Figure 7. We can say that our sample mean of 11,69 is significantly greater than the population mean of 3,9. Therefore null hypothesis is dismissed and alternative hypothesis is accepted.

As stated before, not all assumptions for t-test are fulfilled, but significant discrepancy between population mean and sample mean should suggest that observations in room C12 were not mutually independent.

Overview of gathered point's means for each scoring mode of the course for students that took exam in room C12

TABLE II – COMPARISON OF GATHERED SCORE – ANALYSED STUDENTS FOM ROOM C12 AND WHOLE POPULATION

| | 1.MI | 2.MI | 3.MI | ZI | T1 | T2 | NAS | DZA |
|---|---|---|---|---|---|---|---|---|
| 'C12' | 10,88 | 11,69 | 8,61 | 8,48 | 3,28 | 4,08 | 3,80 | 8,94 |
| all | 9,66 | 7,78 | 8,50 | 5,57 | 3,19 | 4,41 | 3,61 | 8,66 |

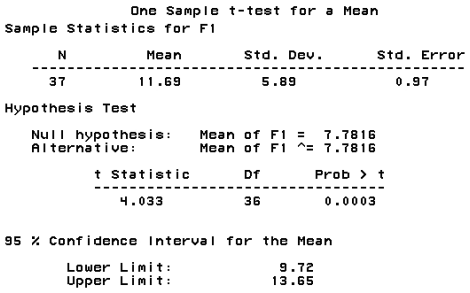


Figure 7: One-sample t test results

could give additional insight. Their score means could be compared with means for whole generation (Table II). Overview reveals that students from room C12 are above average students. However, since most of them took all midexams in room C12 – it could have impact on their collective success. Disregarding that fact and assuming that they are really above average, their score for second midexam significantly outstands success of whole population.

Conclusion is that special control over rooms where exams are held and staff managing exams should be established.

## IV. PREDICTIVE ANALYSIS

Timely recognition of students who are likely to drop the course could be very valuable. In that way opportune moves towards them could be made.

Formerly, if supplemental classes were planned, most frequently used method to determine to which students it should be directed to was setting up an arbitrary score border. Now, with tools and various algorithms at hand, this group of students could be recognized in more precise way.

For this purpose two methods were used: logistic regression and decision trees.

SAS Enterprise Miner was used to build up a model. Known results of third midexam and final test were discarded and not used to form a model. Model was formed on the bases of the first population (academic year 2005./2006.), and used to estimate drop outs in second population – generation of students in academic year 2006./2007.

Available data set was divided on training, validation and test data set in following proportions: 70%, 20% and 10% respectively. These proportions were chosen because

relatively small number of observations was at hand. New target binary variable named 'FAIL' was added. It was assigned value of '1' for all dropouts, and value '0' for all students who passed the course.

After model construction, program code was generated and used on second student population.

Logistic regression revealed that the biggest impact on final student success was made by scores collected at first and second midexam, homework points and first short test.

Application of model on second population gave, as expected, somewhat weaker results then application of model on test data of first population (Figure 8). Reasons for differences in model performance for different populations lie in different student structure – in second generation proportion of regular students was much higher. Also midexams were of different difficultness, and some scoring features were slightly changed. Results of model application on second population are given in Table III.

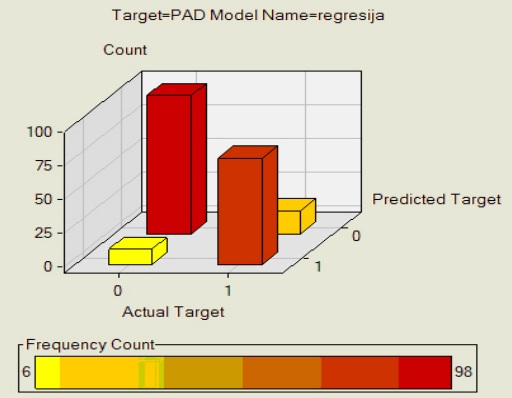


Figure 8: Model performance on test data

TABLE III - PREDICTIVE VALUE OF CONSTRUCTED LOGISTIC REGRESSION MODEL

| probability of student dropping the course | correct prediction | incorrect prediction | model performance |
|---|---|---|---|
| 80 – 100 % | 172 | 53 | 76,5 % |
| 60 – 100 % | 188 | 93 | 66,9 % |

Decision tree method is also very suitable for predicting student's final score, that is, their ability to pass the course. Once again, binary target variable 'FAIL' is generated. Formulation of model rules is simple, just by following nodes and branches, or in this case circle segment, some regularities emerge. A handy feature of circle presentation is that size of each circle part reveals number of observations that satisfies certain rule. In this specific situation, to prevent overfitting, the optimal model has 5 leaf nodes (Figure 9). Decision trees reveal that the biggest impact on final result was made by second midexam, homework and first midexam (ordered by their relevance).

Cumulative lift chart gives good comparison of two constructed models (Figure 10). Regression model generally gives better results. Provided chart reveals also that random sample of chosen students would include 45% of real dropouts. That is actually percentage of dropouts for the whole population. By targeting 30% of population using the regression model, around 97% of them would certainly be those who actually are going to drop the course. One should however, have in mind that most of them are actually weaker students and harder 'to save' than other, undiscovered students.

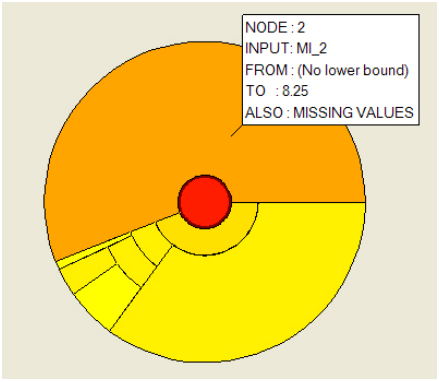


Figure 9: Constructed 5-leafs decision tree model

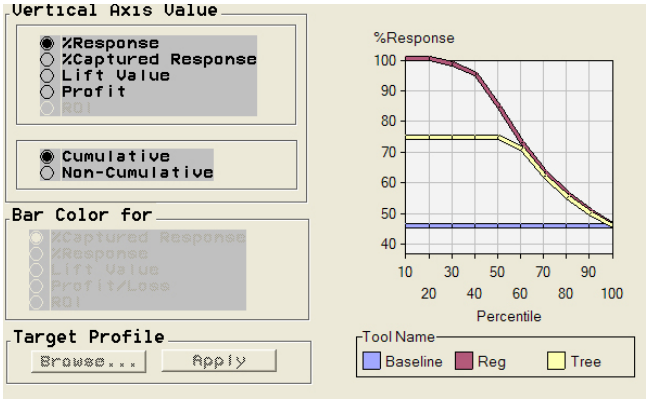


Figure 10: Lift chart gives good model performance comparison

Decision trees method gives much better results if some derived variables are added (like total sum of collected points to-date or total points on short tests, total points at midexams etc.). In that case performance of this model significantly grows and is much better than simple border setting of total points (as rule for such border could be included in whole model if detected as significant). In this case performance of the model with added derived variables grew, and if 30% of population is targeted, 94% of them are real dropouts. Regression, though, still remains the best model.

One of the benefits of using decision trees method is that rules are easily interpreted and students are more transparently directed for additional classes – especially because that population wants to have clear reasons for any action (similar behaviour was indicated in [7]).

## V. FURTHER WORK

In the future, better results of presented methods could be accomplished by using additional data about students – their personal characteristics on one side and data about their success at other courses on the other. In presented analyses it was impossible due to the unavailability of mentioned data.

Additional data would most probably enable detection of new and unknown regularities, and allow further improvements of the whole teaching process. Experienced professors and psychologists should be consulted to determine which additional data and in what form should

be stored and analysed. Extensions could be arranged in three groups:

- improvement in specific subject tracking
- collecting additional data about student success at different courses
- enrichment of students' general data.

Courses are interconnected. Knowledge collected at one course could be helpful or even necessary to successfully complete the other. Way of learning, problem handling or even some arbitrary characteristic like food consummation habits could reveal some impact on successfulness of learning. However, much greater contribution to final analysis is expected of data that could be gathered for specific course tracking. The metadata about subject matter taught should be constructed for that purpose. Concept maps are informal modelling technique suitable for that purpose and already developed, tested and used in academic community for teaching and learning purposes. They are actually knowledge visualisation tools which can provide both a 'course map' view of constructional topics and methodology to support personal knowledge acquisition (as described in [8]).

Each covered concept should be interrelated with other concepts – those which are prerequisites for specific concept understanding and those which are build upon it. Some concepts are more basic and some are specific. Every test question is, on the other hand, related to specific concept or a few of them. Linking specific questions to concepts on one side and students and their answers on the other, could reveal non-comprehension of some concepts. In that way student could be directed to some specific teaching materials (elaborated in detail in [8]) and not just warned or directed to some general literature or class. In that way student could face his/her weaknesses and have better chance to acquire requested knowledge.

Some additional attributes could be connected to the available materials for some concepts (number of examples used to introduce a concept, exercise materials, related quizzes, complexity of related mathematical background, number of prerequisite concepts, additional materials and literature given to students etc.). There were some investigation about that matter and some results are presented in [9].

Visualisation tools that demand student engagement are proven to be very successful methods for knowledge transfer. On that trace attributes that would describe involvement of such tools in concepts taught should be also used (investigated and elaborated in [10]).

Some attributes could be created dynamically – based on success or evaluations of previous generations of students.

## VI. CONCLUSION

Knowledge creation about existing teaching and learning processes is always welcome as it enables various improvements.

Even though available data for presented analyses is rather modest, presented analysis enable concrete steps to be undertaken to improve the way of teaching and testing the student population.

First, test objectivity was analyzed. Some abnormalities were noticed and suggestions for future improvements were given. Data mining was used for timely prediction of possible dropouts. Presented regression model and decision trees demonstrate much better results than simple score border setting. However, there is space for further improvements. Future directions are also described and explained with references on supplemental literature given.

## ACKNOWLEDGMENT

We'd like to thank the dean of Faculty of Electrical Engineering and Computing, University of Zagreb - prof. Vedran Mornar, Ph.D, as well as the vice-dean prof. Mario Cifrek, Ph.D for their assistance in data gathering phase. We also thank the university office for students and prof. Mirta Baranović Ph.D.

It must also be mentioned that Ivan Felja, M.Sc. has designed and made the system for tracking success of students for this course (Electrical engineering fundamentals). We also thank him and Jakov Pavlek, B.Sc. for enabling us to get access to the required data.

## LITERATURE

[1] I. Nonaka: 'A Dynamic Theory of Organizational Knowledge Creation', Organization Science, Vol.5, No.1, pp. 14-37, Feb. 1994.

[2] M.K. Brohman: 'Knowledge Creation Opportunities in the Data Mining Process', Proceedings of 39th Hawaii International Conference on System Sciences, 2006.

[3] M. K. Brohman, M. Parent, M. R. Pearce, M. Wade: 'The Business Intelligence Value Chain', Proceedings from the 33rd Hawaii International Conference on System Sciences, 2000.

[4] G. Knezović: "School management as business challenge", In special enclosure for 'Business' magazine: "Knowledge and business", p.3, 31.01.2008

[5] Z. Perić: Interview with Ivan Vidaković – Member of National Competitiveness Council, also director of IBM Croatia, in special enclosure for 'Business' magazine: "Knowledge and business", p.12-13, 31.01.2008.

[6] M. Vranić, D. Pintar, Z. Skočir: 'The use of data mining in education environment', Proceedings of Contel, Zagreb, 2007.

[7] Y. Ma, B. Liu, C. K. Wong, P.S. Yu, S. M. Lee, "Targeting the Right Students Using Data Mining", Proceedings of the 6th ACM SIGKDD international conference on knowledge discovery and data mining, Boston 2000.

[8] B. Marshall, H. Chen, R. Shen, E.A. Fox: 'Moving digital libraries into the student learning space: The GetSmart experience', ACM Journal on Educational Resources in Computing (JERIC), Vol. 6 , No. 1, Article 2, March 2006.

[9] R. Peress: 'Real-World Projects Can Make a Difference', IEEE The Institute, March 2007, Vol. 31, No. 1, pp.14

[10] Group of authors: ' Exploring the role of visualization and engagement in computer science education', ITiCSE 2002 working group report, ACM SIGCSE Bulletin archive, Vol. 35 , No 2, June 2003., pp. 131-152