Using Decision Tree Classifier for Analyzing Students' Activities

Snježana Milinković[1], Mirjana Maksimović[2]

[1]snjeza@etf.unssa.rs.ba, [2]mirjana@etf.unssa.rs.ba

Case study

DOI: **10.7251/JIT1302087M** UDC: **37.018.43:004.738.5**

Abstract: *In this paper students' activities data analysis in the course Introduction to programming at Faculty of Electrical Engineering in East Sarajevo is performed. Using the data that are stored in the Moodle database combined with manually collected data, the model was developed to predict students' performance in successfully passing the final exam. The goal was to identify variables that could help teachers in predicting students' performance and making specific recommendations for improving individual activities that could directly influence final exam successful passing. The model was created using decision tree classifier and experiments were performed using the WEKA data mining tool. The effect of input attributes on the model performances was analyzed and applying appropriate techniques a higher accuracy of the generated model was achieved.*

Key Words: *decision tree, moodle, students' performances, e-learning*

Introduction

The process of knowledge acquiring and transmitting was dramaticly changed by the progress in the use of information - communication technologies. Electronic learning (e-learning) has become an area where significant research efforts have been invested with aim to improve existing and find new and attractive method of knowledge dissemination. The basic tendency is to increase the motivation of e-learning courses' users and achieving the best possible outcomes. Learning Management Systems - LMS are software applications used for creation, organization and administration of e-learning courses. These softwares are specially designed for educational purposes and their applications provide user-friendly access to learning contents, easy creation and presentation of learning material, interactive communication among users, testing and polling of users, assessment activities, and so on. One of the LMSs that is widely used in academic communities around the world is Moodle (Modular Object-Oriented Dynamic Learning Environment) [5].

Moodle allows easily creation of electronic courses and adaption of traditional course to formats suitable for e-learning. In addition, it allows tracking all the activities of its users. The information about each user's activities is kept in the Moodle database of Moodle system and it is available to the system administrators at any time. This functionality option of Moodle application is very important because of vast amounts of potentially useful data that are accumulated in this way.

Applying suitable transformation and discretization techniques on data obtained from a Moodle, which can be generated from various reports on activities, it is possible to obtain a form that is suitable for the application of data mining algorithms [9]. Data mining is usually defined as the process of discovering useful patterns or knowledge from different data sources [4]. The main goal of data mining techniques is to find and describe the structural patterns in the data in order to attempt to explain connections between data and create predictive models based on them [13]. Data mining is a multidisciplinary field which includes machine learning, statistics, databases, artificial intelligence, information theory and visualization [4]. One of the most common tasks used in data mining applications is the classification. Classification is type of machine learning analogue to human learning from past experiences to gain new knowledge in order to improve our ability to perform real-world tasks [4]. Computers using machine learning learns from data which are collected in the past and represent past experiences. In most cases classification is used for learning a target function that can be used to predict the values of a discrete class attribute, e. g. classification is one type of predictions methods. The goal of prediction is to infer a target attribute, predicted variable, from some combination of other aspects of the data or another attribute. Classification here means the problem of correctly predicting the probability that an example has a predefined class from a set of attributes describing the example. In classification learning, the learning scheme is presented with a set of classified examples from which it is expected to learn a way of classifying unseen examples [13]. The process of data mining consists of three basic steps:

- Pre-processing – the raw data must be cleaned in order to become suitable for mining. Data cleaning includes removing noises and abnormalities, handling too large data, identifying and removing irrelevant attributes, and so on. Data cleaning is procedure that usually consumes a lot of time and it is very labor-intensive but it is absolutely necessary step for successful data mining.
- Application of data mining algorithms – the process of applying data mining algorithm that will produce patterns or knowledge.
- Post–processing – Among all discovered patterns or knowledge, it is necessary to discover ones that are useful for the application. For making the right decision there are many evaluation and visualization techniques that can be used.

Data mining can be applied to research and analyze the data that come from educational environments. This new developing field, known as Educational Data Mining (EDM), began to develop intensively in recent years. It is engaged in the development of methods for exploring the unique types of data that come from the educational context [10]. The main objective is to discover the implicit and useful patterns or knowledge about how students learn and the factors that affect their learning. Gained knowledge can be used to provide feedbacks to the teachers in order to improve the teaching process through more quality and easier management of students in the learning process achieving the best possible outcomes.

In recent years, a lot of research in the field of educational data mining was performed. An overview of the current state and the progress made in the development and implementation of educational data mining is given in [10]. Prediction of the achieved success and the final grade in the exam can be performed applying data mining algorithms. In [1], the ranking of factors that influence the prediction of academic performance in order to identify students who will need to study harder to pass the exam was performed by the application of data mining methods. An experiment with pattern classification for student performance prediction is performed in [2]. The obtained results illustrate that recognition for a certain class on a large data set can be obtained by a classifier built from a small size data set. The scope of [7] was to identify the factors influencing the performance of students in final examinations and find out a suitable data mining algorithm to predict the grade of students. The obtained results reveals that type of school does not have influence on students' performance while parents' occupation plays a major role in predicting grades. The focus of the research can be put on usage of data mining methods for analyzing the quality and methods that e-learning courses content is presented to the students [6]. The impact of the certain e-learning tools on the achievement of students' objectives is discussed in [3]. In [8] a survey about the application of data mining to web-based electronic courses and learning content management systems was performed. As a result, a general model that represents the whole process of application of data mining techniques in educational systems was created (Figure 1). A specific Moodle mining tool oriented for the use of not only experts in data mining is described in [11]. Also, the performances of different data mining techniques for classifying students are compared. Performed experiments show that in general there is not one single algorithm that obtains the best classification accuracy in all cases while some pre-processing task like filtering, discretization or rebalancing can be very important to obtain better or worse results.

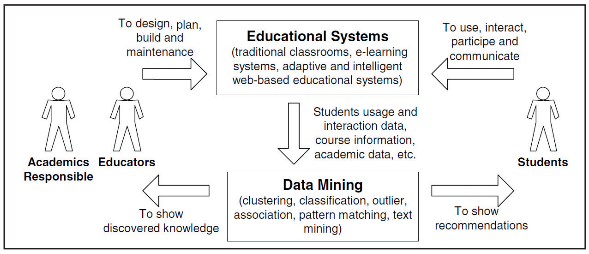

To design, plan, build and maintenance
Educational Systems
(traditional classrooms, e-learning systems, adaptive and intelligent web-based educational systems)
To use, interact, participe and communicate
Students usage and interaction data, course information, academic data, etc.
Academics Responsible
Educators
Students
Data Mining
(clustering, classification, outlier, association, pattern matching, text mining)
To show discovered knowledge
To show recommendations

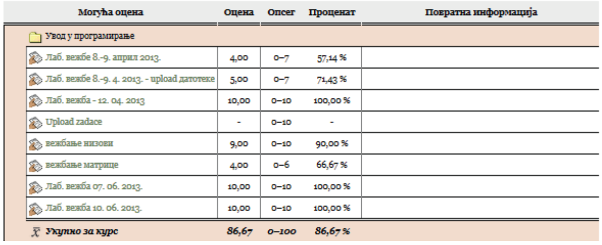
Figure 1 Application of data mining in educational systems

This paper analyzes the impact of specific pre-exam activities on actual student performance in the course Introduction to programming that is performed in Faculty of Electrical Engineering in East Sarajevo. A model for predicting students' performance in the final exam was developed by analyzing course activities. Most of the data about these activities were stored in the Moodle database while some of the data were manually collected (students' attendance on lectures). From the Moodle database, a randomly selected data for one generation of students were chosen. The model was created using a decision tree classifier. Presented experiments were performed using WEKA data mining tool [12]. The influence of input attributes on the performance of the model was analyzed and higher accuracy of the generated model was achieved by application of appropriate techniques.

The rest of this paper is organized as follows. The course organization, data collection and preprocessing are described in second section. Third section presents J48 Decision tree algorithm while in fourth section simulation results of four proposed experiments are shown. Finally, fifth section provides conclusion remarks and outlines directions for future work.

Course Organization, Data Collection and Preprocessing

For the purposes of this study, data about students who have attended the Introduction to Programming course, which is performed during the summer semester of the first year of study in Electrical Engineering in East Sarajevo, were collected and analyzed. Randomly sampling, the data of the students from all three study programs that are running at the faculty were collected. Electronic course was implemented as a complement to the traditional way of teaching what means that concept of blended learning is applied. The main objective of the electronic course creation is to improve the efficiency of traditional ways of teaching. The course was created using the Moodle platform and has been used to provide various learning resources and facilitate communication among its participants. Traditional course content, and thus supporting electronic course, was organized through three parts: lecture, problem solving exercises and laboratory exercises. Through the pre-exam activities, students are required to attend and successfully complete 3 cycles of laboratory exercises. Other activities in the course (homework assignments, successfully done tasks on preparatory cycle of laboratory exercises, lessons access, forums, and access to other resources of course) are not part of the mandatory pre-exam activities but some of them are scored. The number of points gained through these non-mandatory activities represents bonus in adition to the maximum required number of points which student can earn. A percentage values of the successful completion of certain activities are stored in the Moodle database for each student, as it shown in Figure 2.



Могућа оцена	Оцена	Опсег	Проценат	Повратна информација
Увод у програмирање				
Лаб. вежбе 8.-9. април 2013.	4,00	0–7	57,14 %	
Лаб. вежбе 8.-9. 4. 2013. - upload датотеке	5,00	0–7	71,43 %	
Лаб. вежба - 12. 04. 2013	10,00	0–10	100,00 %	
Upload zadace	-	0–10	-	
вежбање низови	9,00	0–10	90,00 %	
вежбање матрице	4,00	0–6	66,67 %	
Лаб. вежба 07. 06. 2013.	10,00	0–10	100,00 %	
Лаб. вежба 10. 06. 2013.	10,00	0–10	100,00 %	
Укупно за курс	86,67	0–100	86,67 %	

Figure 2 A percentage values of the successful completion of certain activities for one student

This information is extracted from the Moodle database for randomly chosen students. Manually collected data about points that students earned by attendance on lectures during semester are added to this information and together they present input data for data mining process. To be able to apply data mining techniques, it is necessary to preprocess input data. In the initial stage of preprocessing, data of students who have not obtained the minimum points required for the successful defense of mandatory laboratory exercises are discarded. Next step is identifing and discarding the attributes that have no predictive value (the index number, name, and so on). After that, all percentage values extracted from Moodle database are recalculated in the number of points for particular activities. By manually discretization process [11] a numerical values which represent the final grade of class attribute ‚results' were transformed into nominal values in accordance with the specific needs of the individual experiments performing. After that, using the filter method Infogain the values of input attributes in relation to the class attribute have been evaluated. In this manner, in data set used in the study, attributes that have no impact on values of the class attribute are identified and discarded. All discarded attributes in this preprocessing step belonged to the set of attributes that describe the non-mandatory course activities (graded and ungraded).

Data preprocessing is a procedure that usually consumes a bulk of time and requires a lot of work, but it is an absolutely necessary step for the successful application of data mining techniques and algorithms.

J48 Decision Tree Algorithm

The decision tree is a very popular method for classification and decision making. It is a decision making technique based on the relationship between strategy and conditions, and it is used to solve many problems. It predicts outcomes using a series of questions and rules for data classification. The decision tree branching occurs as a result of meeting the requirements of classification issues. Each question will divide data into subsets that are more homogeneous than the senior set. If the question has two answers, then the response to the question arise two subsets (binary tree). Subsets arise according to number of questions answers. Therefore the classification of certain data are carried out. Predicting the behavior of a particular client can be made on the basis of its belonging to a particular event (which is classified based on a number of issues and conditions), for which we know how it acts. During the construction of decision trees is important to know the right questions. The main advantage of decision tree classifier is its classification speed. The models which are based on the decision tree algorithms differ in certain data characteristics which are required and in which basis issues are created [13]. In this paper, J48 decision tree, which is an implementation of C4.5 algorithm in WEKA data mining tool [12], is used.

Simulation Results

In order to obtain as much useful information of the individual attributes impact on students' performance in the course Introduction to Programming with aim of obtaining a large percentage of correctly classified instances, the work presented in this paper is carried out through several experiments:

- 1st experiment:
- Used attributes are: laboratories (total), student attendance on lectures and results (passed and failed).
- 2nd experiment:
- Used attributes are: laboratory exercises of first, second and third cycle (L1, L2 and L3, respectively), student attendance on lectures and results (passed and failed).
- 3rd experiment:
- Used attributes are: laboratories (total), student attendance on lectures and results (passed in June-July period, passed in other periods and failed).
- 4th experiment:
- Used attributes are: laboratory exercises of first, second and third cycle (L1, L2 and L3, respectively), student attendance on lectures and results (passed in June-July period, passed in other periods and failed).

Attribute 'results' in all 4 experiments is referred to as a class variable.

1st experiment

Attributes evaluation can be performed using Info-Gain Attribute Evaluation and Gain-Ratio Attribute Evaluation. Info-Gain evaluates attributes by measuring their information gain with respect to the class. This method can treat missing as a separate value or distribute the counts among other values in proportion to their frequency. Gain-Ratio Attribute Evaluation evaluates attributes by measuring their gain ratio with respect to the class. Attributes with estimates of less than 0.01 should be excluded from the analyzed data set. For attributes proposed in 1st experiment, their evaluation is given in Table 1.

Table 1. Attributes evaluation – 1st experiment

Atribute	*InfoGain AttributeEval*	*GainRatio AttributeEval*
Lab.	0.203	0.208
Attendance	0.125	0.126

Table 1 show that laboratory excercies have the major impact to final results. The attribute with the maximum gain ratio is selected as the splitting attribute what can be seen from Figure 3 a.

In the first experiment, after applying the J48 classifier an accuracy of 71.8 % is achieved and created tree is shown in Figure 3 a. The numbers given in parentheses are the number of instances assigned to that node number followed incorrectly classified instances. The minimum number of instances per node (minNumObj) was kept at 2 and during the experiment 10-fold cross-validation is applied, which is a standard method for predicting the error rate learning techniques of a given fixed sample of data. The data were divided into 10 subsets where classes are represented in approximately the same proportion as in the full data set. Each part is done in order and learning scheme is trained on the remaining nine-tenths, and the error rate is calculated on the set of the test sample. Thus, the learning is performed a total of 10 times on different training sets (each set has much in common with the other). Finally, there is an average value of 10 estimated errors to obtain an estimate of the total error [13]. If the minimum number of instances per node (minNumObj) is increased to 3 a simplier tree shown in Figure 3 b is obtained, but the accuracy of correctly classified instances is less than the previous case -70.4%.

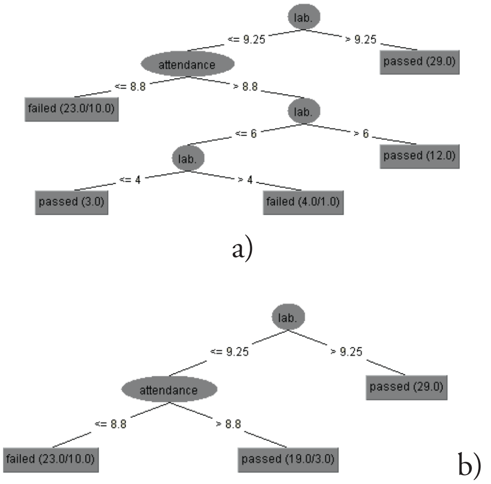


Figure 3 Decision tree (1st experiment): a) the initial model, b) the model with increased minimum number of instances per node

2nd experiment

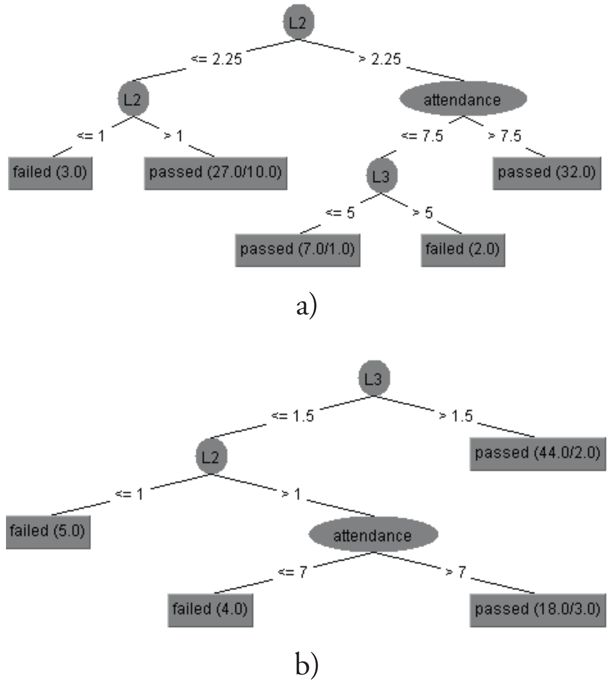
Results of attributes evaluation in second experiment are shown in Table 2.

Table 2. Attributes evaluation – 2nd experiment

Attribute	*InfoGain AttributeEval*	*GainRatio AttributeEval*
L1	0	0
L2	0.226	0.258
L3	0.193	0.219
Attendance	0.126	0.125

Table 2 shows that the attribute with the maximum gain ratio is L2 and it is selected as the splitting attribute while first laboratory exercise L1 evaluation is less than 0.01 so it should be excluded from the dataset.

Observing the effects of individual laboratory excercies (L1, L2 and L3) and the student attendance on lectures on achieved results at the exam in the second experiment, using the J48 decision tree and the 10-fold cross-validation, obtained accuracy is 74.6 %. Created tree is shown in Figure 4 a.



a)

b)

Figure 4 Decision tree (2nd experiment): a) the initial model, b) the model achieved over a balanced data

In multiclass prediction, the result on a test set is often displayed as a two-dimensional confusion matrix with a row and column for each class. Each matrix element shows the number of test examples for which the actual class is the row and the predicted class is the column. Good results correspond to large numbers down the main diagonal and small, ideally zero, off-diagonal elements. The results are shown in Table 3.

Table 3. Confusion matrix

Predictied class		
a	b	**Real class**
50	5	*a=pass*
13	**3**	*b=failed*

Originally generated model have shown unbalanced distribution of examples per class variables, what indicated that the data were not well prepared. In the case of unbalanced data sets, examples of small classes are more difficult to train. The problem with unbalanced data arises because learning algorithms tend to overlook less frequent classes (minority classes), paying attention just to the most frequent ones (majority classes). As a result, the classifier obtained is not able to correctly classify data instances corresponding to poorly represented classes. One of the most frequent methods used to learn from unbalanced data consists of re-sampling the data. To solve this problem, in this paper resampling was performed using Resample Weka filter for supervised learning with or without instance replacement. Created decision tree on re-sampled data is shown in Figure 4 b. Achieved accuracy in this case is 87.3%. It can be concluded that accuracy is significantly increased on re-sampled data and the decision tree clasificator created more precise model.

3rd experiment

In the third experiment, the emphasis is on the impact of attributes laboratory excercises (total) and student attendance on lectures on passing the exam in the first, June-July, final exam term.

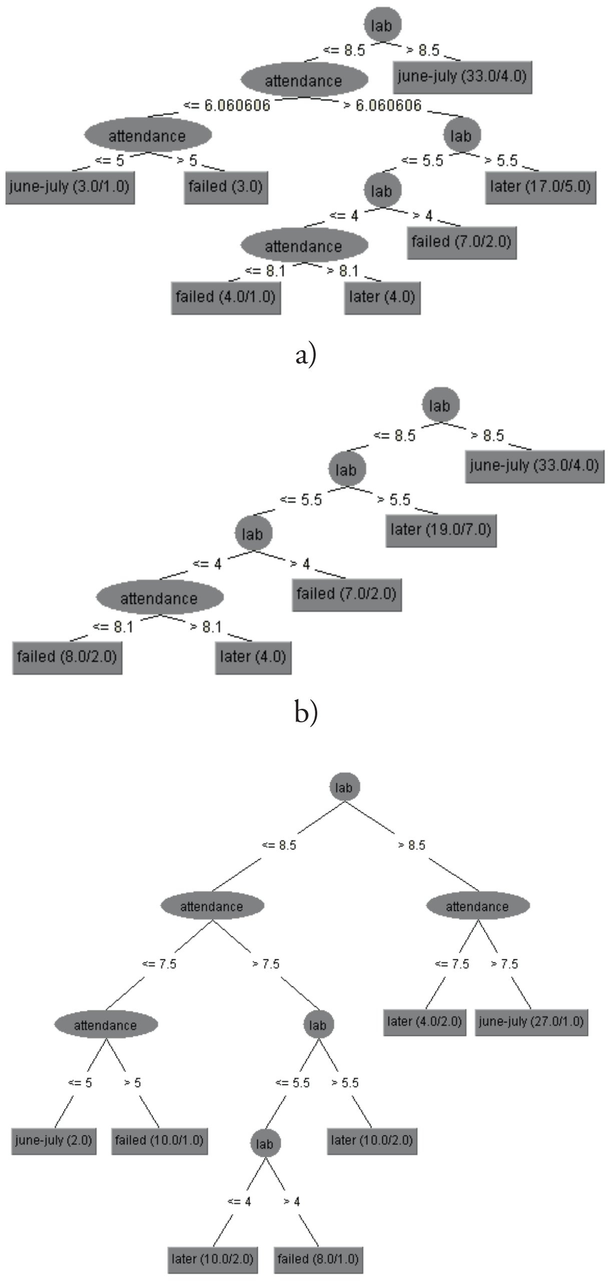
For those proposed attributes their evaluation is given in Table 4.

Table 4. Attributes evaluation – 3rd experiment

Atribut	*InfoGain AttributeEval*	*GainRatio AttributeEval*
Lab.	0.489	0.548
Attendance	0.441	0.289

Table 4 shows that laboratory excercies have the maximum gain ratio and it is selected as the splitting attribute what can be seen from Figure 4.

In this case, the achieved accuracy is 61.9% and created decision tree is shown in Figure 5 a. Increasing the minimum number of instances per node (minNumObj) from 2 to 4, the accuracy drops to 57.5% creating simplier tree shown in Figure 4 b. Applying decision tree classifier on balanced data the achieved accuracy is 73.2%. Tree created on balanced data set is shown in Figure 5 c.



a)

b)

c)

Figure 5 Decision tree (3rd experiment): a) the initial model, b) the model achieved by increasing the minimum number of instances per node, c) model achieved over a balanced data

4th experiment

In the fourth experiment the influence of individual laboratory excercises (L1, L2 and L3) and the student attendance on lectures to passing the exam in the first, June-July, final exam term is analyzed.

Results of attributes evaluation in this experiment are shown in Table 5.

Table 5. Attributes evaluation – 4th experiment

Attribute	*InfoGain AttributeEval*	*GainRatio AttributeEval*
L1	0.212	0.214
L2	0.384	0.335
L3	0.363	0.321
Attendance	0	0

Table 5 shows that the attribute with the maximum gain ratio is L2 and it is selected as the splitting attribute while evaluation of attribute attendance is 0.

In this case the J48 decision tree classifier achieves an accuracy of 47.8% and created decision tree is shown in Figure 6 a. From confusion matrix (Table 6) it can be seen that there is an imbalance in the distribution of the value of output classes and the accuracy of small classes is less than the accuracy of the higher class.

Tabela 6. Confusion matrix – 4th experiment

Predicted class			
a	b	c	**Real class**
25	4	6	*a=june-july*
7	**4**	5	*b=failed*
10	5	**5**	*c=later*

Applying the function Resample data distribution balance is improved, which affects the result. In this case the accuracy of 80.2% is achieved. Thus, the predictive accuracy of a balanced data is significantly increased. If in the balanced data the number of instances per node is increased from 2 to 3, the accuracy of model predictions fails to 77.4% and the decision tree classifier creates a simplier decision tree Figure 6 b.

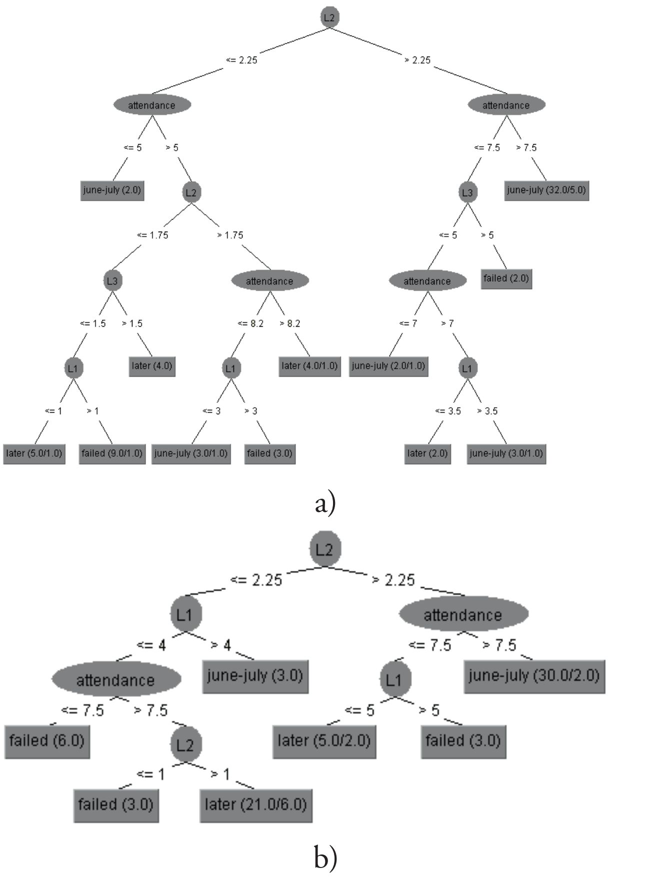


Figure 6 Decision tree (4th experiment): a) the initial model, b) the model achieved by increasing the minimum number of instances per node over a balanced data

Analyzing the summarized results of performed experiments, presented in Table 7, it can be seen that the highest accuracy of the initial data is achieved in the second experiment, while the lowest accuracy is achieved in the fourth.

Table 7 Summarized results of performed experiments

experiment/achieved accuracy	initial model	balanced data
First experiment	71.8%	76.5%
Second experiment	74.6%	87.3%
Third experiment	61.9%	73.2%
Fourth experiment	47.8%	80.2%

Also, it is evident that with more class attributes accuracy decreases. By increasing the minimum number of instances per node a decision tree is simplier but at the same time accuracy decreases.

After balancing the data, accuracy in all four experiments is significantly increased, with the highest accuracy achieved again in the second experiment, while the largest percentage improvement over the initial model is discernible in the fourth experiment. Experiments have shown that the greatest influence on the outcome of students success in the final exam has the laboratory exercise L2 while L1 has the smallest influence.

Conclusion

After all discovered patterns or gained knowledge by applying data mining algorithms it is necessary to discover those that are useful for the particular application and to identify variables that can help teachers in predicting student performance. Experiments performed in this work using the J48 decision tree classifier showed that the laboratory excercise of the second cycle have the greatest impact on the success of passing the exam which leads to a conclusion that this teaching unit need an extra attention. Also, improving its content should lead to better overcome of those laboratory exercises and thus directly influence increase of the final exam passing rate.

From filter method and the obtained experimental results it can be concluded that the impact on the learning process have only those activities in the course that are mandatory. This imposes a recommendation that a greater number of activities should be classified into this category in order to ensure better and continuous work of the students during the semester, what will be the subject of our future research.

Authorship statement

Author(s) confirms that the above named article is an original work, did not previously published or is currently under consideration for any other publication.

Conflicts of interest

We declare that we have no conflicts of interest.

Literature

[1] Affendey, L.S., et al. (2010). Ranking of Influencing Factors in Predicting Students' Academic Performance, *Information Technology Journal* 9 (4): 832-837.

[2] Ai, J., and Laffey, J. (2007). Web Mining as a Tool for Understanding Online Learning, *MERLOT Journal of Online Learning and Teaching* 3(2): 160-169.

[3] Kickul, J., and Kickul, G. (2002). New pathways in e-learning: The role of student proactivity and technology utilization, 45rd Annual Meeting of the Midwest Academy of Management Conference, Indiana, USA.

[4] Liu, B. (2007). Web DataMining - Exploring Hyperlinks, Contents, and Usage Data, © Springer-Verlag Berlin Heidelberg

[5] Moodle, Available: https://moodle.org/

[6] Prema, M., and Prakasam, S. (2013). Effectiveness of Data Mining - based E-learning system (DMBELS), *International Journal of Computer Applications* 66(19): 31-36.

[7] Ramesh, V., et al. (2013). Predicting Student Performance: A Statistical and Data Mining Approach, *International Journal of Computer Applications* 63(8): 35-39.

[8] Romero, C., and Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005, *Expert Systems with Applications* 33, 135–146.

[9] Romero, C., et al. (2008). Data mining in course management systems: Moodle case study and tutorial, *Computers&Education*, Elsevier 55(1):368–384.

[10] Romero, C., and Ventura, S. (2013). Data mining in education, WIREs *Data Mining Knowl Discov*, 3(1): 12–27.

[11] Romero, C., et al. (2013). Web usage mining for predicting final marks of students that use Moodle courses, *Comput. Appl. Eng. Educ.*, 21: 135–146.

[12] Weka software tool, Available: http://www.cs.waikato.ac.nz/ml/weka/

[13] Witten IH et al. (2011) Data mining: practical machine learning tools and techniques, Morgan Kaufmann, Amsterdam

Submitted: November 18, 2013.
Accepted: December 20, 2013.