Students' success prediction using Weka tool

Milos Ilic, Petar Spalevic
Electrical and Computing Engineering
University of Pristina, Faculty of Technical Science
Kosovska Mitrovica, Serbia
milos.ilic@pr.ac.rs, petar.spalevic@pr.ac.rs

Mladen Veinovic, Wejdan Saed Alatresh
Singidunum University
Belgrade, Serbia
mveinovic@singidunm.ac.rs, she11hab@yahoo.com

Abstract— **One of the biggest challenges for higher education today is to predict the paths of students through the educational process. Institutions would like to know, which students will need assistance in order to finish course successfully. Successful students' result prediction in early course stage depends on many factors. Data mining techniques could be used for this kind of job. Based on collected students' information, different data mining techniques need to be used. For the purpose of this research WEKA data mining software was used for the prediction of final student mark based on parameters in two different datasets. Each dataset contains information about different students from one college course in the past fourth semesters. Student data from the last semester are used for test dataset.**

Keywords-Classification; data mining; J48; prediction; SMO; weka; ZeroR

I. INTRODUCTION

The main objective of higher education institutions is to provide quality education to its students. One way to achieve highest level of quality in higher education system is by discovering knowledge for prediction regarding enrolment of students in a particular course, detection of abnormal values in the result sheets of the students, prediction about students' performance and so on. For the purpose of students' data processing, data mining techniques could be used. Data Mining or knowledge discovery has become the area of growing significance because it helps in analyzing data from different perspectives and summarizing it into useful information [1].

The main data mining functions are applying various methods and algorithms in order to discover and extract patterns of stored data. Data mining or knowledge discovery applications have got a rich focus due to its significance in decision making and it has become an essential component in various organizations. Data mining enables organizations to use their current reporting capabilities to uncover and understand hidden patterns in vast databases. These patterns are then built into data mining models and used to predict individual behavior with high accuracy. As a result of this insight, institutions are able to allocate resources and staff more effectively. Educational data mining is a new emerging technique of data mining that can be applied on the data related to the field of education. There are increasing research interests in using data mining in education. This new emerging field, concerns with developing methods that discover knowledge from data originating from educational environments. Data mining may, for example, give an institution the information necessary to take action before a student drops out, or to efficiently allocate resources with an accurate estimate of how many students will take a particular course Educational data mining uses many techniques such as decision trees, neural networks, k-nearest neighbor, naive bayes, support vector machines and many others [2]. All data mining techniques are implemented as part of different software applications. Some of them have specific purpose, and some applications can be used for different problems, and on concrete datasets different techniques and algorithms can be applied.

One of the open source software designed for data analysis and knowledge discovering is WEKA [3]. WEKA or Waikato Environment for Knowledge Analysis software is product of the University of Waikato (New Zealand) and was first implemented in its modern form in 1997. It uses the GNU General Public License (GPL). The software is written in the Java™ language and contains a GUI for interacting with data files and producing visual results. It also has a general API, so WEKA can be embedded in other applications like any other library. WEKA has several standard data mining tasks, data preprocessing, clustering, classification, association visualization, and feature selection.

In this research we use WEKA data mining tool to predict students' results in the early stage of particular course. More precisely, prediction is based on two different training datasets. First training data set contains information about number of students' visits to the lectures and laboratory exercises. Second training dataset contains more students' data beside lecture visits. Data mining classification algorithms was applied on both datasets separately and our goal was to predict student final score and mark based on those two dataset. Research includes students' data collected in the period of four semesters. These data was collected from one particular course. Data from the three of the mentioned four semesters are used for model training, and the data from the last fourth semester are used for testing and prediction.

The paper is organized as follows. Second part represents literature review of similar researches. Third part presents short explanation about WEKA software possibilities, and data mining algorithms implemented in it. The fourth section represents analysis and discussion of obtained classification and prediction results in our research. Fifth section presents main conclusions and ideas for future research, and the last section presents used references.

II. LETERATURE REVIEW

According to [4] predicting students' profiling indicate that data mining allows building customer models each describing the specific habits, need and behavior of group of customers. It classifies new customers and predicts their special need. Consequently, data mining can help management to identify the demographic, geographic and psychographic characteristics of students based on information provided by the students at the time of admission. Profiles are often based on demographic and geographic variables. Furthermore, surveys are one common method of building customer profiles. Neural networking technique can be used to identify different types of students. In addition, discriminant analysis can also be used to identify patterns. Regression analysis, decision tree and Bayesian classification can be applied. Consequently, cluster analysis can be done to students' profiling and separate marketing strategies can be prepared to target segmented students.

Authors in [5] studied students' performance in the course using data mining techniques, particularly classification techniques such as Naive Bayes and Decision tree based on students ID and marks scored in course. Furthermore, they suggest that data mining process can be done to the teachers for classifying performance which helps in improving higher education system. Data mining methods helps students and teachers to improve students' performance.

In [6] authors uses the data mining prediction technique to identify the most effective factor to determine a student's test score, and then adjusting these factors to improve the student's test score performance in the following year. Author in [7] present the various techniques of data mining which is used to analysis the student records in order to categorize the students into grade order in all their education studies and it helps in interview situation. It examines that which factors helps to categorize students in rank order to arrange for the recruitment process. Due to this, we can easily discover the eligible student and it also reduces the short listings. For this job data mining techniques are efficiently used to manage the performance level of students. Classification is one of the data mining techniques which is used to accurately classifies the data for categorizing student based on the levels. Clustering is one important function of data mining to analysis discovering data sources distribution of information and the cluster analysis is an important research topic.

Results presented in [8] describe the application of k- mean clustering algorithm to provide the result of student academic performance. The main aim is to analysis the student's performance by using k mean implementation in clustering. In this paper authors combined the k mean model with the deterministic model to analyze the students' results of a private Institution in Nigeria which is a good benchmark to monitor the students' progression of academic performance in higher institution for the purpose of making an effective decision by the academic planners. Authors simply compare the predictive power of clustering algorithm and the Euclidean distance as a measure of similarity distance. They provide better result compare the earliest model of k-mean.

Authors in [9] have studied how data mining can be applied to educational systems. They show how useful data mining can be in higher education, particularly to improve students' performance. In the research they used students' data from the database of final year students' for Information Technology UG course, and available data including their performance at university examination in various subjects. They applied classification and clustering algorithms ZeroR and DBSCAN respectively. Based on DBSCAN algorithm noisy data was detected. Their conclusion is that each of this knowledge can be used to improve students' performance.

III. WEKA DATA MINING TOOL

Weka is portable and platform independent software because it is fully implemented in the Java programming language and thus runs on almost any modern computing platform. This software has several standard data mining tasks, data preprocessing, clustering, classification, association visualization, and feature selection. The weka GUI chooser launches the weka's graphical environment which has four buttons: Explorer, Experimenter, Knowledge Flow and Simple CLI. Data mining techniques and algorithms used in this research are placed in Explorer interface, which has several panels that give access to the main components of the workbench [10]. The start point in weka explorer is preprocessing panel. From this panel user can load datasets, browse the characteristics of attributes and apply any combination of weka's unsupervised filters to the data. When data are loaded user can apply data mining techniques divided in two main groups: classifier and cluster group. From classifier panel user could configure and execute any of the weka classifiers on the current dataset. User can choose to perform a cross validation or test on a separate dataset. Classification errors can be visualized in a pop-up data visualization tool. If the classifier produces a decision tree it can be displayed graphically in a pop-up tree visualizer. Another group of techniques are clustering techniques. From the cluster panel user can configure and execute any of the clusters on the current dataset. Clusters can be visualized in a pop-up data visualization tool. The next three panels provide different possibilities for data association and visualization to the user. From the associate panel user can mine the current dataset for association rules using the weka associators [11]. User through the select attributes panel can configure and apply any combination of attribute evaluator and search method to select the most pertinent attributes in the dataset. Visualize panel is the last panel in the explorer weka card. This panel displays a scatter plot matrix for the current dataset. The number of cells in the matrix can be changed by pressing the *Select Attributes* button and then choosing those attributes to displayed. This panel allows user to visualize the current dataset in one and two dimensions [11]. When the coloring attribute is discrete, each value is displayed as a different color; when the coloring attribute is continuous, a spectrum is used to indicate the value. When the class is discrete, misclassified points are shown by a box in the color corresponding to the class predicted by the classifier; when the class is continuous, the size of each plotted point varies in proportion to the magnitude of the error made by the classifier.

IV. EXPERIMENT DISCUSSION

Datasets used in this research are created and saved as .arff (Attribute - Relation File Format) file [12]. An .arff file is an ASCII text file that describes a list of instances sharing a set of attributes. This file has specific structure. If data are saved in other file, that file must be converted in .arff, because weka works with .arff files. ARFF files have two distinct sections. The first section is the Header information, which is followed the Data information. The Header of the ARFF file contains the name of the relation, a list of the attributes (the columns in the data), and their types. Different attribute types can be used for different types of information. The last attribute is class attribute.

In the case of this research the class attribute contains information about students' grade. In Data information section data are put in the same order as attributes in the header (columns in data row), and all data are comma separated. One such row of data is named instance. From database about students two initial training datasets are created. For the training datasets "old" students' data from the past three semesters were used. Those three semesters represent school years in the period from 2013 to the 2015. As we mentioned above research is carried out based on two separated experiments. The both experiments have the same goal, to predict students' final grade bases on information collected in the early stage of the course.

The first prediction experiment is based on the students' presence at lectures. Information about students' presents at the lectures is numeric attribute, calculated as the difference between the total number of lectures and the number of students' arrivals to the lectures. In the same way another attribute about students' presents on the laboratory exercises is calculated, because these teaching activities were held at different times. The second prediction experiment besides the attributes about students' presence on the lectures depends on two more parameters. These two parameters are students' results on two tests performed during the semester.

Experiments in this paper are based on prediction functionalities provided by classification techniques. Classification is a data mining task that maps the data into predefined groups and classes. First step after the dataset creation and loading in weka preprocess panel is model construction. Model construction consists of set of predetermined classes. Each sample is assumed to belong to a predefined class. The set of sample used for model construction is mentioned training set. For model creation the most important is to choose best classification algorithm. On both training datasets same techniques were applied, and technique with best performances was selected for model creation. Classification results are presented in Table 1 and Table 2 for the first and second training dataset respectively. The model can be represented as classification rules, decision trees, or mathematical formulae. Created model is used for classifying future or unknown objects. For all classifiers presented in the tables we perform 10-fold cross-validation, without percentage split. This means that we use whole dataset for training, another datasets are used for testing and prediction.

TABLE I. CLASSIFICATION RESULTS AND STATISTICS

First training dataset				
Parameters	***ZeroR***	***IBk***	***J48***	***Part***
Correctly Classified Instances [%]	49.57	74.50	71.10	67.70
Mean absolute error	0.23	0.11	0.13	0.14
Root mean squared error	0.34	0.23	0.26	0.27
Relative absolute error [%]	100	46.82	75.97	62.26

TABLE II. CLASSIFICATION RESULTS AND STATISTICS

Second training dataset				
Parameters	***ZeroR***	***IBk***	***J48***	***Part***
Correctly Classified Instances [%]	50.71	98.58	86.40	84.14
Mean absolute error	0.23	0.01	0.06	0.08
Root mean squared error	0.34	0.05	0.18	0.19
Relative absolute error [%]	100	3.67	28.87	57.97

From the above presented classification results we can conclude that in the both cases (first training and second training dataset) best performances provides *IBk* classification algorithm. That is implementation of K-nearest neighbor classifier. Beside this classification technique *J48* classification algorithm provides good performances too. *J48* is decision tree classification algorithm and provides possibility for decision rules creation. Because of that fact both algorithms are used for model creation and future prediction.

For the future prediction which is based on these models the same classifiers need to be used. If we observe results presented in the above tables from the aspect of students' data used for classification we can see that in both cases results are similar. Classification results show that beside more parameters in the second training dataset number of correctly classified instances is similar. The quest is in which measure that can or can't affect the prediction results.

Test datasets (new data) contains students' data from the last (fourth) semester which is covered by this research. In the moment when the prediction was performed those students was not finish final exam yet. Based on that we were able to compare predicted final grade and final student' grade after the exam. Test datasets (two of them) are created in the same structure like training datasets. One and only difference between training and testing dataset is in class attribute. In the case of test dataset class attribute can be question mark, or some value predicted by user. In both cases after prediction weka inputs predict value on the place for class attribute in each instance row. In the prediction phase the known label of test sample is compared with the classified result from the model. Accuracy rate is the percentage of test set samples that are correctly classified by the model. Test set is independent of training set, otherwise over-fitting will occur. The decision tree is used to represent logical rules of student final grade based on presented parameters. Some of decision tree combinations are presented in continuation for the first and second dataset on the Fig. 1 and Fig. 2 respectively.

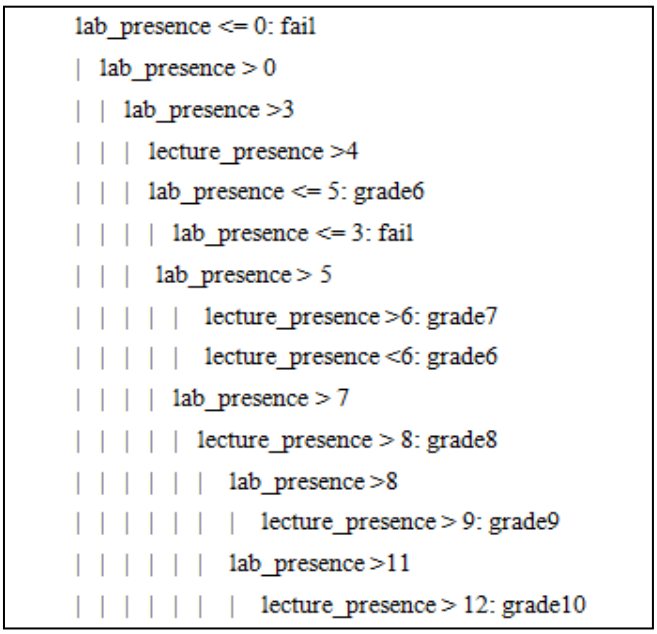

Figure 1. Decision tree for the first dataset

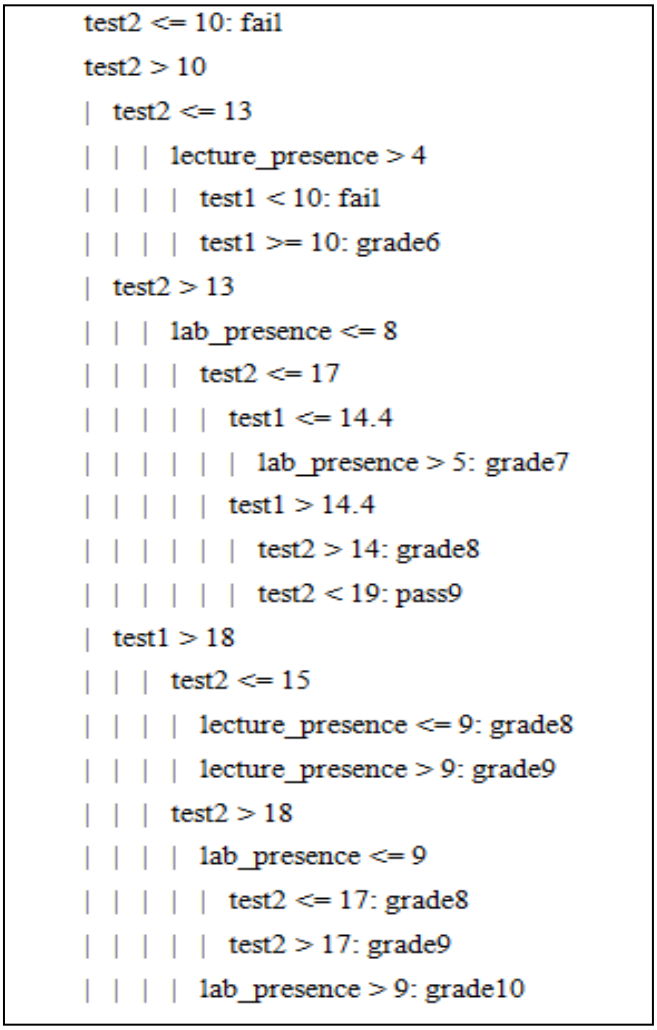

Figure 2. Decision tree for the second dataset

TABLE III. COMPARISON OF PREDICTED VALUES AND FINAL GRADES

First test dataset			**Second test dataset**	
Grade	***IBk***	***J48***	***IBk***	***J48***
Fail exam	89%	87.01%	90.8%	88.05%
Six (pass)	89.50%	87%	91.02%	88.80%
Seven	88.03%	86.5%	90%	90.01%
Eight	87.20%	85%	93%	95.20%
Nine	87%	85.02%	95.89%	97%
Ten	88%	86%	97.20%	97.06%

Presented decision trees provide basic rules for creating final students' grades. Decision tree as can be seen from Fig. 1 provides rules for final students' grades based on students' presents on lectures and laboratory exercises.

As we can see, student must be present on one third of total number of lecture and laboratory classes to pass particular exam. When the total number of students' presents on the lecture and laboratory classes is greater than two thirds of the total number of classes, we can expect that the students will get a high grades. Generally number of lecture and laboratory classes on which student need to be present to get grades in the range from six to ten varies. Different tree paths end up with the same grades based on different values for both parameters. In some cases value for one parameter can be higher than value for another, but such values provide the same grade as some other combination.

Decision tree shown in Fig. 2 on the other hand provides rules for students' final grade prediction based on more parameters. Because of more parameters, second decision tree has more possible paths and concrete tree is much bigger. In accordance with this fact, here is shown only one part of the decision tree. Two more parameters represent scores from two partial tests during the semester. In accordance with decision tree, we can expect that student will pass the exam, if he/her has scores on both tests equals to ten or greater than ten points. In the same time another two parameters that represent student presence on classes must take the value at the range of at least one third of total number of classes.

If we carefully observe both trees we can see that parameters for lecture and laboratory presence are in the same range for the same grades. With more parameters for decision rules in second tree prediction model provides a more accurate calculation and prediction. Student' presence on lectures may not be a reliable parameter that student will successfully pass the exam, and because of that test results are desirable to secure better prediction.

After prediction based on created models and completion of the final exam in the faculty, confirmation of prediction was calculated. Confirmation is calculated and presented in the percentages, as the number of predicted concrete students' grade and the total number of students who get that particular grade. As we can see from Table 3, confirmation of prediction success is calculated for the case of both test datasets and both created classification models.

Also calculation was performed for all grades, including students who have not passed the exam. Based on Table 3 we can see that in the case of second test dataset (dataset with more than two parameters) the accuracy of prediction is higher for higher grades. It is because students who are not attending the classes usually are not even passed the exam. In those cases information about students' presents on the lectures is enough for successful prediction.

In fact the percentage of matches to predict the failure of the examination and the real failure is the most important, because these students are actually target group for early success prediction. The benefit of early prediction of students who are not able to pass the exam is possibility that for these students professors and faculty could organize additional classes, or to provide additional attention when they working with them. In such cases, with additional work and appropriate help, students can achieve better results. Another students' grade prediction in the early phase of the course provides for faculty useful information about all students who enroll to the next year of study.

V. CONCLUSION

Higher education institutions are nucleus of research and future development of scientific and technical personnel. Higher education institutions acting in a competitive environment, with the prerequisite mission to generate, accumulate and share knowledge. The chain of generating knowledge inside and among external organizations is considered essential to reduce the limitations of internal resources and could be plainly improved with the use of data mining technologies. In this paper authors presented one of open source data mining tool that can be used to improve educational process. Quality of students' lectures and at the end quality of students' knowledge is very important for all academic stuff employed in the higher educational institutions.

Presented experiment which provides students' grade prediction present that data mining techniques can be used to provide better quality of student knowledge. If professors are able to predict students' final grade in early phase of course, they can spot potentially shortcomings and students who need more attention. If they have such information in right moment they could help most of them to finish course successfully. In that way those student's will be encouraged by the progress, and that filing will help them to overcome the upcoming course materials. Actually through the additional attention and learning students would learn more, and based on that knowledge they will able easier to learn new material.

The future research on this topic will be to extend dataset with new data from the other courses and other students' knowledge testing methods. We want to find the balance between the required number of attributes in the datasets and the most successful prediction. In the same way weka data mining tool can be used for other educational data processing like course enrolment, timetable optimization and prediction of number of students which will enroll some concrete course beads on information from the past. We can conclude that weka is powerful tool, but like any other tool requires appropriate dataset, and if it is possible big dataset with accurate information from the past. Each prediction depends on the accuracy of information on which is the creation of training models based.

ACKNOWLEDGMENT

This work has been supported by the Ministry of Education, Science and Technological Development of Republic of Serbia within the projects TR 32023 and TR 35026.

REFERENCES

[1] C. Romero, S. Ventura, E. Garcia, "Data mining in course management systems: Moodle case study and tutorial", Computers & Education, vol. 51, no. 1, 2008, pp. 368-384.

[2] S. Ayesha, T. Mustafa, A. Sattar, M. Khan, "Data mining model for higher education system", Europen Journal of Scientific Research, vol.43, no.1, 2010. pp.24-29.

[3] Weka 3: Data Mining Software in Java, University of Waikato, [Online]. Available: http://www.cs.waikato.ac.nz/ml/index.html.

[4] L. Romdhae, N. Fadhel, B. Ayeb, "An efficient approach for building customer profiles from business data", Expert System with Applications, vol. 37, 2010, pp. 1573-1585.

[5] A. Kumar, G. Uma, "Improving academic performance of students by applying data mining techniques", European Journal of Scientific Research, no. 4, 2009, pp. 526-534.

[6] S. Gabrilson, D. Fabro, P. Valduriez, Towards the efficient development of model transformations using model weaving and matching transformations, Software and Systems Modeling 2003. Data Mining with CRCT Scores, Office of information technology, Geogia Department of Education.

[7] K. Umamaheswari, S. Niraimathi "A study on student data analysis using data mining techniques", International Journal of Advanced Research in Computer Science and Software Engineering, vol. 3, Issue 8, August 2013, pp. 117-120.

[8] O. Oyelade, O. Oladipupo, I. Obagbuwa, "Application of k-Means clustering algorithm for prediction of students' academic performance" (IJCSIS) International Journal of Computer Science and Information Security, Vol. 7, num. 1, 2010, pp. 292-295.

[9] S. Aher, L.M.R.J. Lobo, "Data mining in educational system using weka", International Conference on Emerging Technology Trends, 2011, pp. 20-25.

[10] R. Kirkby, E. Frank, Weka explorer user guide for version 3-4-3, The University of Waikato, 2004, pp. 1-13.

[11] Weka knowledge explorer, [Online]. Available: http://www.cs.waikato.ac.nz/~ml/weka/gui_explorer.html.

[12] Attribute-Relation File Format (ARFF), University of Waikato, [Online]. Available: http://www.cs.waikato.ac.nz/ml/weka/arff.html.