The Fifth International Conference on e-Learning (eLearning-2014), 22-23 September 2014, Belgrade, Serbia

BOOSTING - A METHOD FOR IMPROVING THE ACCURACY OF PREDICTIVE MODEL

SNJEŽANA MILINKOVIĆ
University of East Sarajevo, Faculty of Electrical Engineering, snjeza@etf.unssa.rs.ba

Abstract: *In this paper, some models for predicting students' performance in the final exam have been shown. Applying special machine learning techniques and algorithms more accurate classification/predictive models can be obtained. Analyzing generated models the specific administrative and demographic data that most influence students' performance in the course Introduction to programming that is performed in Faculty of Electrical Engineering in East Sarajevo were identified. The models were created using WEKA data mining tool.*

Keywords: *Educational data mining, Classification, Predictive model*

1. INTRODUCTION

The highest quality in performing the teaching process should be the priority task of any education institution. One of the key factors for quality assurance of teaching process is its organizational strategy and educational institutions should pay special attention on it. To create a good organizational strategy it is necessary to have available as many as possible data about students for whom the teaching process will be organized. Efficient analysis of those data would provide information for teachers and management of educational institution in order to perform maximal adoption of teaching process to the needs of students who will attend it. In this way students' motivation and learning outcomes could be significantly improved. Information systems of educational institutions store large amounts of data about students. Some of those data are more or less important but as a whole they contain potentially useful information and knowledge about the students. In order to gain that knowledge it is necessary to perform the efficient processing of those data. One efficient way for performing that process is applying data mining techniques.

Data mining involves techniques for discovering implicit patterns in the data that could provide new knowledge. Input data for applying data mining techniques are presented in the form of a set of examples, and the output can be expressed in the predictive or descriptive form of the analyzed data structure. Data mining is a multi-disciplinary field involving machine learning, statistics, databases, artificial intelligence, information retrieval, and visualization [1]. There are four the most common tasks used in data mining applications: supervised learning (or classification), unsupervised learning (or clustering), association rule mining, and sequential pattern mining. Each of them is characterized by different styles of learning but all of them provide necessary guidance for better understanding of analyzed data and some useful knowledge about connections between input and output data. One of the most commonly used data mining task is creation of classification or predictive model. It is desirable for generated models to have as higher as possible the accuracy of classification and prediction. In addition, it is also desirable for the model to be present in the form of some comprehensible formalism and to be easily interpretable by those users who are not data mining experts.

Data mining applied for analyzing the data that come from different types of educational environments present special research field known as Educational Data Mining (EDM) [2]. EDM analyzes the unique types of data generated by any kind of information system that is used for supporting learning or education. These data can be generated through interactions of individual students with an educational system but they might also include administrative data (e.g. school, school district), demographic data (e.g. gender, age, school grades), data about student affectivity (e.g. motivation, emotional states), etc. [2]. The main objective of educational data mining is to extract implicit and useful patterns or to obtain useful knowledge about the ways students learn and factors that affect their learning. Different data mining models can be implemented to evaluate students' performance. Analyzing those models it is possible to identify some connections between data and the factors that have key influence on students' achievements. That knowledge can help teachers to get proper understanding of student's learning capabilities and provide useful guidance for improvement of teaching process.

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 [2]. In [3], 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. Applying different data mining classification techniques for predicting the marks in the final exam of the students that use Moodle courses has been shown in [4]. Using clustering analysis comparing of two algorithms for measuring the potential of students' academic skills has been done in [5]. The impact of the certain e-learning tools on the achievement of students' objectives is discussed in [6]. A survey about the application of data mining to web-based electronic courses and learning content management systems was performed in [7].

In this paper, some models for predicting students' performance in the final exam has been shown. Applying special machine learning techniques and algorithms more accurate classification/predictive models can be obtained. Analyzing generated models the specific administrative and demographic data that most influence students' performance in the course Introduction to programming that is performed at the Faculty of Electrical Engineering in East Sarajevo were identified. The models were created using WEKA data mining tool [8].

The rest of this paper is organized as follows. The main characteristics of applied data mining methods and techniques are described in second section. The third section describes input data for creating classification/predictive model. Performed experiments are described in fourth section, and fifth section provides conclusion remarks and outlines directions for future work.

2. CLASSIFICATION AND BOOSTING

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 [1]. 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.

A lot of different classification algorithms have been developed, but the most popular are so called "white-box" classification algorithms. They provide an explanation for the classification result and their results are directly suitable for decision making. Among those algorithms one of the most popular is C4.5 decision tree algorithm. Decision tree based algorithms 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 classification results obtained by applying C4.5 algorithm is usually very comprehensible, but drawback can be pretty low accuracy of predictive model - those classifiers are usully pretty weak.

Special kind of classification learning is so called ensemble learning which includes techniques based on combining different models learned from the data [9]. Applying these techniques several different training sets are derived from original training set and for each of them a classification model is learned. The ensemble classifier combines these models and produces one ensemble of learned models. In that way, relatively weak classifier can be transformed into very powerful ensemble classifier. These techniques are particularly suitable for applications with so-called unstable learning algorithms like Decision tree and Neural networks [9]. Unstable learning algorithms usually produce quite different classification results even if only small changes in the input data happened. From the perspective of ensemble learning classifier these instabilities are desirable: combining multiple models makes sense only if these models are different from one another.

One of the most frequently used and very powerful ensemble machine learning scheme is boosting. Boosting can be applied for creating classification model and predictive accuracy of generated model is very often significantly higher than the one obtained using a single model. For creating single classification/predictive model boosting uses voting: it combines classification results obtained performing classification algorithms of the same type over different subsets of training dataset. Boosting is iterative process in which each new model is influenced by the performance of models that have been built previously. It creates single prediction combining the outputs of individual models using voting together with weighting. It gives greater weights to those instances that haven't been handled correctly performing previous models. In that way, boosting forced every new model to try to obtain correct classification result for those instances. Combining voting and weighting on each test instance more reliable prediction can be obtained in most cases.

There are many variants on the idea of boosting. Two the most commonly used are AdaBoost.M1 developed by Freund and Schapire (1996) and LogitBoost algorithm developed by Friedman et al. (2000) [9], [10].

One of the drawbacks of ensemble learning techniques is loss of interpretability of the obtained classification/predictive model. In recent years some methods that combine the performance benefits with comprehensible models have been developed. Some of them produce standard decision tree models while others introduce new variants of trees that provide optional paths [9]. All of them are part of so called Interpretable ensembles. One approach for creating a single tree structure that can represent an ensemble of classifiers compactly can be done if the ensemble consists of decision trees. The result of this approach is called an option tree. Option trees differ from decision trees in that they contain two types of node: decision nodes and option nodes. For classifying an instance it is necessary to filter

it down through the tree. At a decision node just one of the branches has to be taken but at an option node take all of the branches have to be taken. In such way, the instance ends up in more than one leaf, and the classifications obtained from those leaves must somehow be combined into an overall classification. This can be done simply by voting, taking the majority vote at an option node to be the prediction of the node [9].

Option trees can be generated by incrementally adding nodes to it. This is commonly done using a boosting algorithm and this is one of the approaches implemented in Weka data mining tool. The resulting trees are usually called alternating decision trees instead of option trees. In that case, the decision nodes are called splitter nodes and the option nodes are called prediction nodes [9]. The standard alternating decision tree applies to two-class problems. A positive or negative numeric value is associated with each prediction node. To obtain a prediction for an instance it has to be filtered down all applicable branches and sum up the values from any prediction nodes that are encountered. Depending on whether the obtained sum is positive or negative the predicting class is generated. Alternating decision trees always have a prediction node at the root. The alternating tree can be grown using a boosting algorithm that employs a base learner for numeric prediction, such as the LogitBoost method, and can be extending for solving the multiclass problems by splitting the problem into several two-class problems [11].

3. INPUT DATA FOR CLASSIFICATION

For the purposes of this study, administrative and demographic data of students who have attended the Introduction to Programming course were collected and their impact on students' performance was analyzed. This course is performing during the summer semester of the first year of study at the Faculty of Electrical Engineering in East Sarajevo. Randomly sampling, the data of the 2013/14 generation of students from all three study programs that are running at the Faculty have been taken into account. Open source data mining tool WEKA [8] was used to apply the learning methods to a dataset and analyze their output to extract useful information about the data and their impact on students' performance. The data collected, which represent the attributes for data mining process, include:

- city from where students came (*city*),
- high school they graduated (*school*),
- obtained mark of subject mathematics in all four high school years (*m1*, *m2*, *m3*, *m4*),
- obtained mark of subject informatics in all four high school years (*i1*, *i2*, *i3*, *i4*),
- average mark of subject mathematics in high school (*matav*),
- average mark of subject informatics in high school (*infav*),
- graduated average mark in high school (*hsav*),
- points obtained on the faculty qualification exam (*test*),
- total number of points collected for enrolment to faculty (*total*),
- enrolment period (*enroll*),
- way of financing the study period (status) ,
- department (*depar*),
- average mark obtained on passed exams at the beginning of second semester (*exam_av*),
- number of passed exams at the beginning of second semester (*exam_num*),
- obtained mark on 6 subject performed during the first semester (*eng1*, *math1*, *fct*, *phy*, *fee1*, *man*),
- obtained mark in the course Introduction to Programming (*mark*).

The last attribute was used as a class attribute.

To be able to apply data mining techniques, it was necessary to pre-process input data. In the initial stage of pre-processing step the attributes that have no predictive value are identified and discarded (the index number, student name, and so on). By manually discretization process [2] a numerical values which represented the marks obtained on six subject performed in the first semester are transformed into two nominal values, passed or failed, and the final grade of class attribute '*mark*' were transformed into the same nominal values (grade 5 – failed, grades 6, 7, 8, 9, 10 - passed). Excel .csv file is formed of these data and exported to WEKA data mining tool.

Three experiments were conducted and performance and results of obtained classification models were analyzed.

4. EXPERIMENTAL RESULTS

The first experiment was performed applying J48 decision tree classification algorithm which is Weka implementation of above described C4.5 algorithm. The obtained results are shown in Image 1. The classification accuracy which is the number of correctly classified instances in the test set divided by the total number of instances in the test set was used as a measure for estimation the strength and the accuracy of a classification/predictive model.

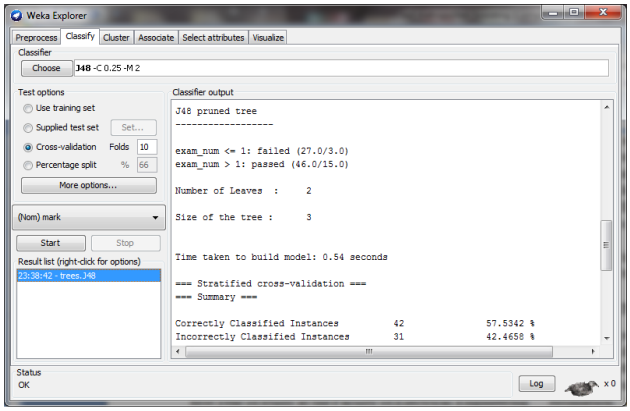

Image 1: J48 classification results

From Image 1 it can be seen that relatively low accuracy is obtained, 57,53% correctly classified instances. Generated tree is shown in Image 2.

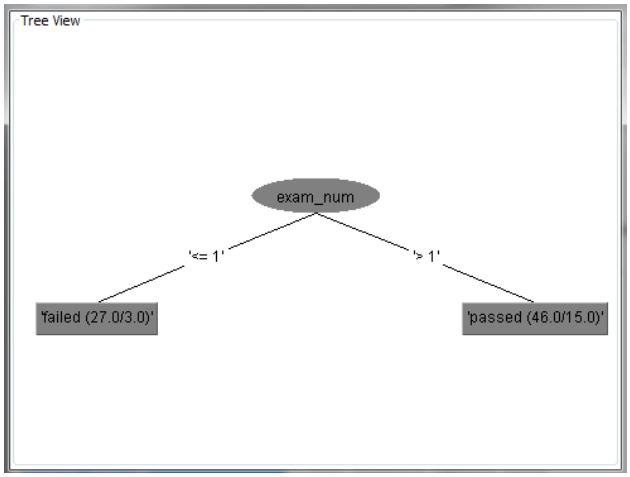

Image 2: J48 decision tree

From Image 2 it can be seen that attribute *exam_num*, the number of the exam passed until the beginning of the second semester is identified as a key tree splitting attribute. The numbers in the brackets present the total number of instances of that class/the number of misclassified instances of that class. It can be seen that the number of misclassified instances of passed class is pretty high. Even though obtained tree is very comprehensible the obtained result cannot be suitable for any further use.

Higher accuracy of classification model can be obtained using decision trees implemented using powerful boosting technique. One of Weka implementation of alternating decision tree is so called ADTree interpretable ensemble. The classification results obtained using ADTree is shown in Image 3.

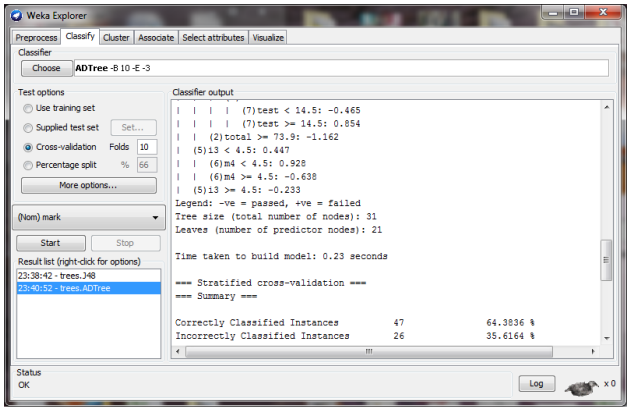

Image 3: ADTree classification results

From Image 3 it can be seen that better accuracy is obtained, 64,38% correctly classified instances. Generated tree is shown in Image 4.

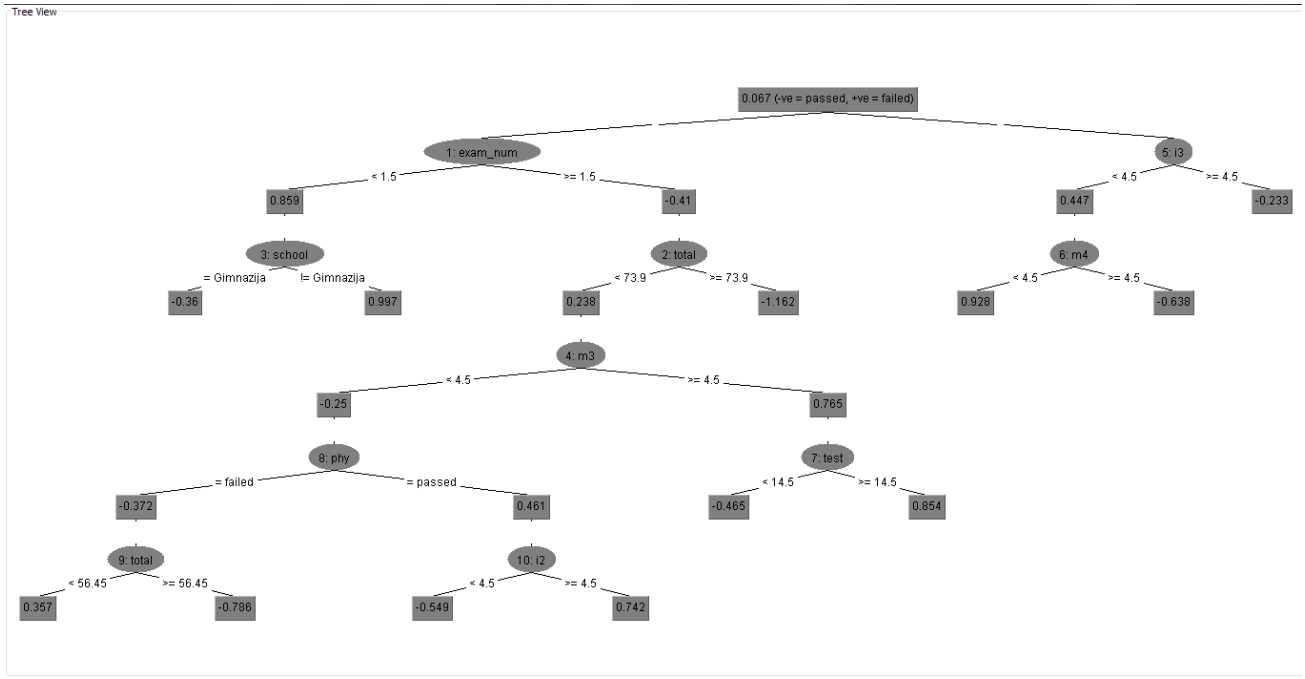

Image 4: ADTree alternating decision tree

From Image 4 it can be seen that alternating decision tree with splitter nodes and prediction nodes is created. This tree is not easily interpretable and to obtained classification result for every instance of data set it has to be filtered down the tree as it was mentioned above in this paper. To classify an instance we have to go down the tree according to the values of its attributes and sum up the numerical values from any prediction nodes that are encountered. The predicted class depends on the obtained sum value: if that value is positive the class is failed, and the class is passed if obtained sum is negative, as it was explained in the first, base prediction node shown in image 4. From Image 4 it can also be seen that the same splitting attribute *exam_num* is again chosen, but in this case, more attributes are includes in tree decision making process and that is why better classification results is obtained. Loss of interpretability is price for that.

The third experiment is performed using another Weka interpretable ensemble classifier: even more powerful alternating decision tree performed using LogitBoost algorithm. The obtained results are now much better, Image 5.

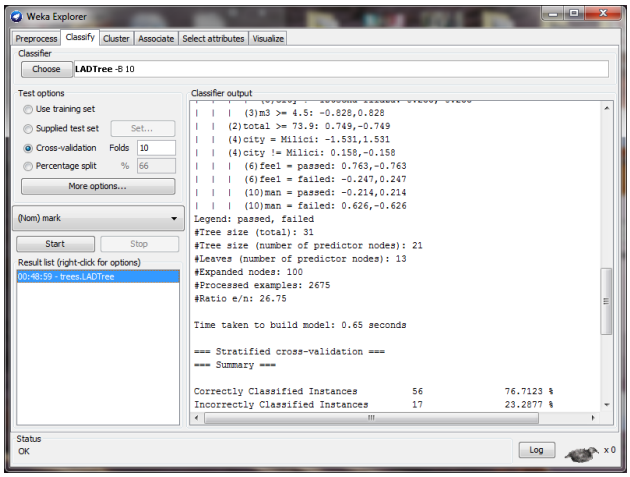

Image 5: LADTree classification results

High accuracy of 76,71% is now obtained, which present very good classification result, but method for computing the class of particular instances using this tree is even more complicated than for the standard alternating decision tree [9]. Generated tree is shown in Image 6.

From Image 6 it can be seen that again the same splitting attribute *exam_num* is chosen as the most important one for decision making. Comparing trees obtained performing ADTree and LADTree algorithms a few attributes are identified as very important for decision making in both methods (*exam_num*, *total*, *test* and *m3*). These attributes are important for early identification of different groups of students, especially the ones who have low possibility to pass the exam (*exam_num* < 1.5). This information can be very useful to the teacher in order to try to pay more attention to those students and try to adapt the teaching material to motivate them to study more. In addition, from generated trees it is obviuos that there are some attributes that never show up in decision making process. The futher experiments have to be performed in order to determine whether these attributes present unnecessary noise and to make a conclusion whether they need to be a part of input data set or not.

It is obvious that so called ensemble learning classifier are very powerful data minig tools. The lack of easy interpretation of their classification results is a serious obstacle to their massive use and focus of future research should be a way to translate their result in the form that will be easily understood by non-expert data mining users.

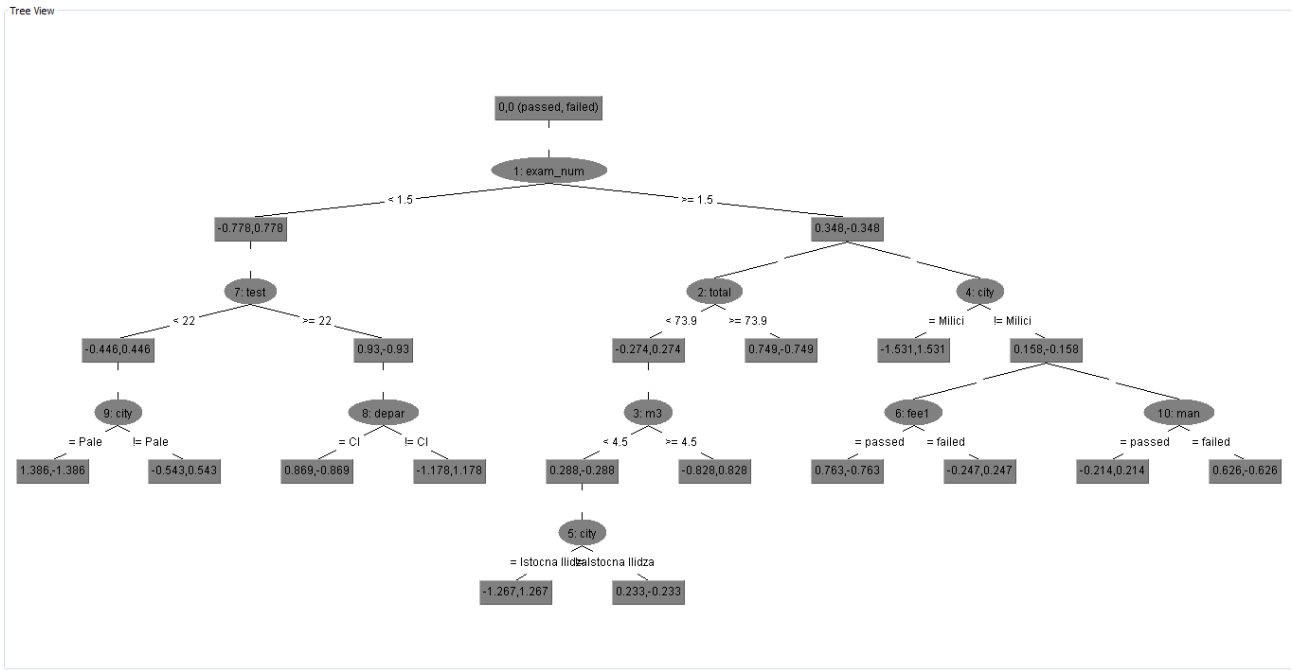

Image 6: LADTree alternating decision tree

5. CONCLUSION

The main goal of this paper was to investigate the possibilities for creation as much as possible more accurate predictive model for one educational data set. One of the advanced machine learning techniques, boosting and its algorithms has been analyzed. Performed experiments have shown that quite satisfactory predictive result can be obtained. The main drawback of analyzed algorithms is pretty hard interpretation of the obtained classification results. The main goal of future research should be seeking the ways to translate obtained results on some easily understandable formalism.

One of the goals is also to investigate what impact on created models could have their combining with some pre-processing techniques applied on input data like filtering or select attributes in Weka implementation.

All the performed experiments were conducted using default values of boosting algorithms' parameters. Changing the number of boosting iteration and their impact on accuracy of predictive model can also be the subject of future work.

LADTree is multiclass algorithm so that it can be applied on this input data set with more than two classes, and some guidance for more precise grouping of students could be obtained.

LITERATURE

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

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

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

[4] C. Romero, E.G. Pedro, A. Zafra, J.R. Romero, S. Ventura, "Web Usage Mining for Predicting Final Marks of Students Thas Use Moodle Courses", © 2010 Wiley Periodicals, Inc.

[5] Dewi, A. O. P. Utomo, W. H. Sri Yulianto J. P., "Identification of Potential Student Academic Ability using Comparison Algorithm K-Means and Farthest First", International Journal of Computer Applications (0975 – 8887) Volume 63– No.17, 2013

[6] 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

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

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

[9] I. H. Witten, E. Frank, M.A. Hall, "Data mining: practical machine learning tools and techniques", 3rd edition, Elsevier, 2011

[10] J. Friedman, T. Hastie, R. Tibshirani, Additive logistic regression: a statistical view of boosting, The Annals of Statistics, Vol. 28, o. 2, 337-407, 2000

[11] G. Holmes at al. Multiclass Alternating Decision Trees, University of Waikato, Hamilton, New Zealand