American International Journ Available online at http://www.iasir.net
Research in Science, Technology,
Engineering & Mathematics
ISSN (Print): 2328-3491, ISSN (Online): 2328-3580, ISSN (CD-ROM): 2328-3629

AIJRSTEM is a refereed, indexed, peer-reviewed, multidisciplinary and open access journal published by
International Association of Scientific Innovation and Research (IASIR), USA
(An Association Unifying the Sciences, Engineering, and Applied Research)

Selection of Best Decision Tree Algorithm for Prediction and Classification of Students' Action

Alaa Khalaf Hamoud,
Department of Information Technology,
College of Computer Science and Information Technology,
Basrah University,
Iraq

Abstract: *Since the student's success rate reflects the success of educational organizations, so the trend of increasing student's success became the goal of all educational organizations. Besides that, the student's willingness of studying higher education after complete secondary school is one of the most important goals to the educational Organizations. Many reasons affect this willingness and revealing these reasons may enhance the student's will. Data mining tools (especially Decision Tree Algorithms) can be considered as the best choice to find the hidden patterns in order to achieve these goals. The experimental dataset used in this work is data set about Portuguese student on two courses (Mathematics (395 instances) and Portuguese (Portuguese language course which holds 659 instances)) which was collected and analyzed by Paulo Cortez and Alice Silva, University of Minho, Portugal. Three Decisions Tree algorithms (J48, RepTree and Hoeffding Tree (VFDT)) are applied and experimented in this work. The results showed that J48 algorithm mostly proper to classify and predict both students' willingness to complete higher education and success in courses.*
Keywords: *Educational Data Mining, Decision Tree Algorithms, J48 Algorithm, RepTree Algorithm, VFDT Algorithm.*

1. Introduction

Data mining, also called Knowledge Discovery in Databases (KDD), is the field of discovering novel and potentially useful information from large amounts of data. In recent years, there has been increasing interest in the use of data mining to investigate scientific questions within educational research, an area of inquiry termed educational data mining. Educational data mining (also referred to as "EDM"). EDM researchers study a variety of areas, including individual learning from educational software, computer supported collaborative learning, computer-adaptive testing (and testing more broadly), and the factors that are associated with student failure or non-retention in courses [12][10].

One of the key areas of applications of EDM is improvement of student models that would predict student's characteristics or academic performances in schools, colleges and other educational institutions. Prediction of student performance with high accuracy is useful in many contexts in all educational institutions for identifying slow learners and distinguishing students with low academic achievement or weak students who are likely to have low academic achievements. The end product of models would be beneficial to the teachers, parents and educational planners not only for informing the students during their study, whether their current behavior could be associated with positive and negative outcomes of the past, but also for providing advice to rectify problems [13].

Three Decision Tree algorithms (C4.5 (J48), RepTree and Hoeffding Tree) are applied. The main data set consists of two Comma Separated Values (CSV) files taken from UCI Machine Learning Repository for Students Alcohol Consumption of two courses (Portugal Language and Mathematics). The source data set files contained (1044 instances in instances) with 32 attributes. Some preprocessing operations like (cleaning data, deriving columns and removing columns) are implemented to consolidate these two source files in one data set. WEKA 3.8.0 tool is used to implement Decision Trees .

The organization of this paper is: section two viewed the related works and listed all the models of implementing data mining algorithms with education. Section three explained the concept of Educational Data Mining (EDM) briefly. Section four listed and explained the decision trees (J48, RepTree, and Hoeffding Tree) which are implemented later in the model. Section five explained the machine learning tool WEKA. Section six listed the steps and results of implementing decision trees model. The final section viewed the conclusions extracted from the whole work.

II. Related Work

In [1] the researchers compared different data mining methods and techniques for classifying students based on their Moodle usage data and the final marks obtained in their respective courses. They developed a specific mining tool for making the configuration and execution of data mining techniques easier for instructors. They used real data from seven Moodle courses with Cordoba University students. They claimed that a classifier model appropriate for educational use has to be both accurate and comprehensible for instructors in order to be of use for decision making.

In [2] different methods and techniques of data mining were compared during the prediction of students' success, applying the data collected from the surveys conducted during the summer semester at the University of Tuzla, the Faculty of Economics, academic year 2010-2011, among first year students and the data taken during the enrollment. The success was evaluated with the passing grade at the exam. The impact of students' socio-demographic variables, achieved results from high school and from the entrance exam, and attitudes towards studying which can have an effect on success, were all investigated.

In [3] data mining techniques intend to approach students' achievement of secondary school using real-world data. The two core classes (Mathematics and Portuguese) were modeled under binary/five-level classification and regression tasks. Four DM models (i.e. Decision Trees, Random Forest, Neural Networks and Support Vector Machines) and three input selections (e.g. with and without previous grades) were tested. The results show that a good predictive accuracy can be achieved, provided that the first and/or second school period grades are available. Although student achievement is highly influenced by past evaluations, an explanatory analysis has shown that there are also other relevant.

In [4] the researcher presented the initial results from a data mining research project implemented at a Bulgarian university, aimed at revealing the high potential of data mining applications for university management.

In [6] C4.5 decision tree algorithm is applied on student's internal assessment data to predict their performance in the final exam. The outcome of the decision tree predicted the number of students who are likely to fail or pass. The result is given to the tutor and steps were taken to improve the performance of the students who were predicted to fail. After the declaration of the results in the final examination the marks obtained by the students are fed into the system and the results were analyzed. The accuracy of the algorithm is compared with ID3 algorithm and found to be more efficient in terms of the accurately predicting the outcome of the student and time taken to derive the tree.

III. Educational Data Mining (EDM)

EDM methods often differ from methods from the broader data mining literature, in explicitly exploiting the multiple levels of meaningful hierarchy in educational data.Methods from the psychometrics literature are often integrated with methods from the machine learning and data mining literatures to achieve this goal. For example, in mining data about how students choose to use educational software, it may be worthwhile to simultaneously consider data at the keystroke level, answer level, session level, student level, classroom level, and school level. Issues of time, sequence, and context also play important roles in the study of educational data [10].

Once a construct of educational interest (such as off-task behavior, or whether or not a skill is known) has been empirically defined in data, it can be transferred to new data sets. The transfer of constructs is not trivial – often, the same construct can be subtly different at the data level, within data from a different context or system – but transfer learning and rapid labeling methods have been successful in speeding up the process of developing or validating a model for a new context. This has led to many educational data mining analyses being replicated across data from several learning systems or contexts [9].

IV. Decision Tree Algorithms

Trees are directed graphs beginning with one node and branching to many. They are fundamental to computer science (data structures), biology (classification, psychology (decision theory), and many other fields. Classification and regression trees are used for prediction. In the last two decades, they have become popular as alternatives to regression, discriminant analysis, and other procedures based on algebraic models. Tree-fitting methods have become so popular that several commercial programs now compete for the attention of market researchers and others looking for software [9].

4.1 J48

J48graft is an extended version of J48 that considers grafting additional branches onto the tree in a post processing phase (Webb, 1999). The grafting process attempts to achieve some of the power of ensemble methods such as bagged and boosted trees while maintaining a single interpretable structure. It identifies regions of the instance space that are either empty or contain only misclassified examples and explores alternative classifications by considering different tests that could have been selected at nodes above the leaf containing the region in question [7].

4.2 RepTree

RepTree builds a decision or regression tree using information gain/variance reduction and prunes it using reduced-error pruning. Optimized for speed, it only sorts values for numeric attributes once. It deals with missing values by splitting instances into pieces, as C4.5 does. You can set the minimum number of instances per leaf,

maximum tree depth (useful when boosting trees), minimum proportion of training set variance for a split (numeric classes only), and number of folds for pruning [7] [8].

4.3 Hoeffding Tree

A Hoeffding tree (VFDT) is an incremental, anytime decision tree induction algorithm that is capable of learning from massive data streams, assuming that the distribution generating examples does not change over time. Hoeffding trees exploit the fact that a small sample can often be enough to choose an optimal splitting attribute. This idea is supported mathematically by the Hoeffding bound, which quantifies the number of observations (in our case, examples) needed to estimate some statistics within a prescribed precision (in our case, the goodness of an attribute). A theoretically appealing feature of Hoeffding Trees not shared by other incremental decision tree learners is that it has sound guarantees of performance. Using the Hoeffding bound one can show that its output is asymptotically nearly identical to that of a non-incremental learner using infinitely many examples [14].

V. WEKA tool

Weka is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes [5].

The Weka workbench is a collection of state-of-the-art machine learning algorithms and data preprocessing tools. It includes virtually all the algorithms described in this book. It is designed so that you can quickly try out existing methods on new datasets in flexible ways. It provides extensive support for the whole process of experimental data mining, including preparing the input data, evaluating learning schemes statistically, and visualizing the input data and the result of learning. As well as a wide variety of learning algorithms, it includes a wide range of preprocessing tools. This diverse and comprehensive toolkit is accessed through a common interface so that its users can compare different methods and identify those that are most appropriate for the problem at hand [8].

VI. Decision Trees Models

This section describes the stages of building decision trees. The first step is data preprocessing in which the data is visualized, cleaned and unified in order to be prepared to the second step. The second step, many decision trees algorithms are applied and the run information of each tree compared to find the best decision tree.

A. Data Preprocessing

The data set (Student Alcohol Consumption Data Set [11]) is depended on in this model. The data consists of two data sets student-mat.csv (Math course which holds 395 instances) and student-por.csv (Portuguese language course which holds 659 instances). Both of these data sets are consisting of 32 attributes shown in the table (1).

Table (1): Students Data Set Attributes

Attribute	Description	Values
School	student's school	nominal: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira
Sex	student's sex	nominal: 'F' - female or 'M' - male
Age	student's age	numeric: from 15 to 22
Address	student's home address type	nominal: 'U' - urban or 'R' - rural
Famsize	family size	nominal: 'LE3' - less or equal to 3 or 'GT3' - greater than 3
Pstatus	parent's cohabitation status	nominal: 'T' - living together or 'A' - apart
Medu	mother's education	numeric: 0= none, 1 =primary education (4th grade), 2 =from 5th to 9th grade, 3=secondary education or 4 = higher education
Fedu	Father's education	numeric: 0= none, 1 =primary education (4th grade), 2 =from 5th to 9th grade, 3=secondary education or 4 = higher education
Mjob	mother's job	nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other'
Fjob	Father's job	nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other'
Reason	reason to choose this school	nominal: close to 'home', school 'reputation', 'course' preference or 'other'
Guardian	student's guardian	nominal: 'mother', 'father' or 'other'
Traveltime	home to school travel time	numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour
Studytime	weekly study time	numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours
Failures	number of past class failures	numeric: n if 1<=n<3, else 4
Schoolsup	extra educational support	nominal: yes or no
Famsup	family educational support	nominal: yes or no
Paid	extra paid classes within the course subject (Math or Portuguese)	nominal: yes or no
Activities	extra-curricular activities	nominal: yes or no
Nursery	attended nursery school	nominal: yes or no
Higher	wants to take higher education	nominal: yes or no
Internet	Internet access at home	nominal: yes or no

Romantic	with a romantic relationship	nominal: yes or no
Famrel	quality of family relationships	numeric: from 1 - very bad to 5 - excellent
Freetime	free time after school	numeric: from 1 - very low to 5 - very high
Goout	going out with friends	numeric: from 1 - very low to 5 - very high
Dalc	workday alcohol consumption	numeric: from 1 - very low to 5 - very high
Walc	weekend alcohol consumption	numeric: from 1 - very low to 5 - very high
Health	current health status	numeric: from 1 - very bad to 5 - very good
Absences	number of school absences	numeric: from 0 to 93
G1	first period grade	numeric: from 0 to 20
G2	second period grade	numeric: from 0 to 20
G3	final grade	numeric: from 0 to 20

From the first observation to data sets using WEKA tool, no missing values found. These two data sets are downloaded with file type csv (Comma Separated File) and all text values contained “” with their context. Some cleaning and unifying operations are done such as (removing “”” from “”M”” to produce M and removing “”yes”” to produce yes) . The second step is consolidating these two data sets which results one data set with 1044 instances. In order to perform consolidating stage, attribute (Course) is added to describe the course (Portugal Language course or Math Course) and took values (P or M). Many of derived attributes obtained such as G1Grade, G2Grade, G3Grade and AbsRate as shown in the table(2).

Table(2): Derived Attributes

Derived Attribute	Source Attribute	Description	Value
G1Grade	G1	Grade of first period	Nominal :P =Pass or F= Fail
G2Grade	G2	Grade of second period	Nominal :P =Pass or F= Fail
G3Grade	G3	Grade of final period	Nominal :P =Pass or F= Fail
AbsRate	Absence	Absence rate	Numeric: 1=0, 2= >=1<=5, 3=other

B. Decision Trees Algorithm’s Tests

In the first test to construct decision trees group, all attributes are selected unless (G1, G2, G3, G1Grade, G2Grade) and G3Grade remained in order to use in tree building. Absence attribute removed and AbsRate used instead. The test mode is 10 cross validation with higher as the goal class. Table () shows the results of run information after applying three algorithms (J48, RepTree and Hoeffding Tree).

Table (3): Decision Trees run Information

Algorithm	CCI	ICI	Precision	Recall	F-Measure	ROC Area	Time
J48	**91.954 %**	**8.046 %**	**0.904**	**0.920**	**0.908**	**0.665**	**0.16 seconds**
RepTree	**91.4751 %**	**8.5249 %**	**0.885**	**0.915**	**0.887**	**0.623**	**0.09 seconds**
Hoeffding Tree	**91.4751 %**	**8.5249 %**	**0.837**	**0.915**	**0.874**	**0.494**	**0.11 seconds**

The table consists of nine attributes used for comparing the results of applying the data mining algorithms after selecting specific attributes from source dataset. The table attributes are:

Algorithm represents the name of used algorithm during the test .

1. CCI (Correctly Classified Instances) represents the number of correctly classified instances divided by the total instances and multiplied by 100.
2. ICI (Incorrectly Classified Instances) represents the number of incorrectly classified instances divided by the total instances and multiplied by 100.
3. Precision: of algorithm represents the percentage of accurate classified instances from all truly classified instances.
4. Recall reflects the division number of correctly classified instances by the total number of all instances (almost recall value be same as CCI).
5. F-Measure: measured from recall and precision values (double value of precision multiplied by recall divided by the value of summation of recall and precision).
6. ROC Area stands for Receiver Operation Characteristic Area which depicts the performance of classifier without regard to class distribution or error costs.
7. Time: second parts were taken to build the tree.

Table (3) shows that J48 algorithm took the high percentage value ICI (Incorrectly Classified Instances) and less percentage value with J48. J48 takes the high precision and recall values compared with (RepTree and Hoeffding

Tree) algorithms. ROC and F-Measure values are highest with J48 algorithm while J48 takes the second score time to build the tree.

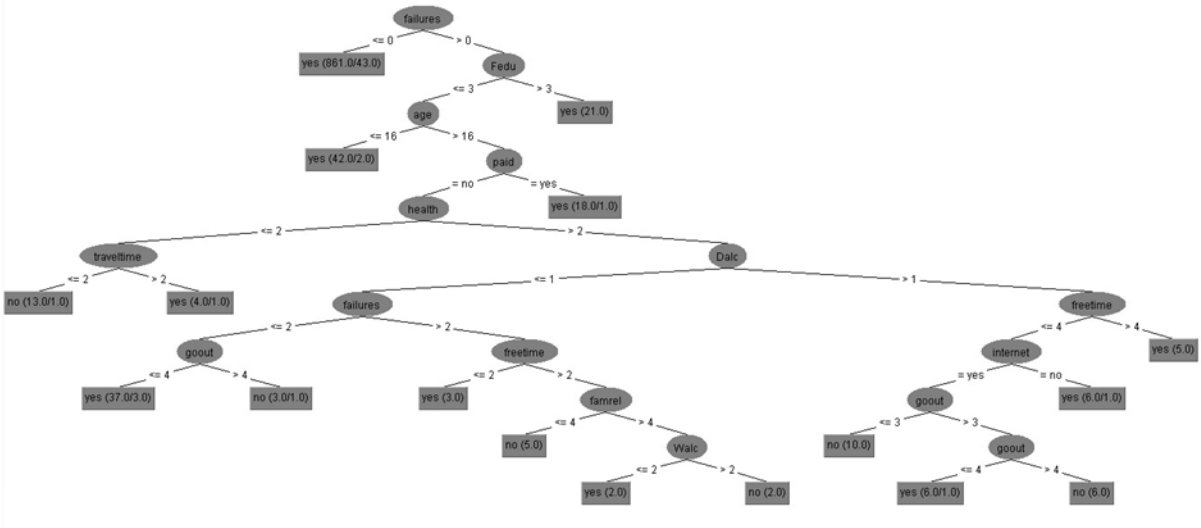

Figure (1): J48 decision Tree of Higher Education Willingness

The second group of decision trees is constructed based on (Course, school, sex, age, address, famsize, Pstatus, Medu, Fedu, Mjob Fjob, Reason, guardian, traveltim, studytime, failures, schoolsup, famsup, paid, activities,nursery, internet, romantic, famrel, freetime, gout, health, absence, AbsRate and G3Grade). The Test mode is 10-fold cross-validation to build tree with high rate accuracy and the final leaf (class is set to higher (student wants to take higher education or not) in order to build a decision tree which classifying and predicting the students based on input Attributes.

Table(4): Decision Trees run information

Algorithm	CCI	ICI	Precision	Recall	F-Measure	ROC Area	Time
J48	**91.4751 %**	**8.5249 %**	**0.897**	**0.915**	**0.902**	**0.615**	**0.04 seconds**
RepTree	**91.3793 %**	**8.6207 %**	**0.879**	**0.914**	**0.882**	**0.663**	**0.01 seconds**
Hoeffding Tree	**91.4751 %**	**8.5249 %**	**0.837**	**0.915**	**0.874**	**0.494**	**0.03 seconds**

The results in the table (4) can be clearly showed the optimal algorithm since it has the best values in almost all attributes. CCI and ICI percentage with J48 are same as with Hoeffding Tree algorithm. Precision, recall, and F-Measure values are high in J48 compared with the other algorithms. RepTree algorithm took the best values with ROC Area and time. Totally, J48 algorithm is the best since it covered all the attributes and took the high accuracy.

The third test is done after selecting all attributes and the goal class is G3Grade in order to find the best algorithm to classify the students based on passing the final period.

Table(5): Decision Trees run information

Algorithm	CCI	ICI	Precision	Recall	F-Measure	ROC Area	Time
J48	**90.3257 %**	**9.6743 %**	**0.904**	**0.903**	**0.904**	**0.857**	**0.02 seconds**
RepTree	**91.4751 %**	**8.5249 %**	**0.918**	**0.915**	**0.916**	**0.905**	**0.01 seconds**
Hoeffding Tree	**89.3678 %**	**10.6322 %**	**0.902**	**0.894**	**0.897**	**0.888**	**0.06 seconds**

The table shows that RepTree is the best algorithm from observing the high rate CCI, low rate ICI, and high precision score with minimum time to build the tree. Recall, F-Measure and ROC area also took the best score with RepTree compared with J48 and Hoeffding Tree. Figure (2) shows the result RepTree graph, and it is obvious that RepTree did not cover all the attributes. In figure (3), J48 took the second score rate and covered nearly all the attributes. The result tree of J48 and RepTree are very simple and easy to understand by the specialists.

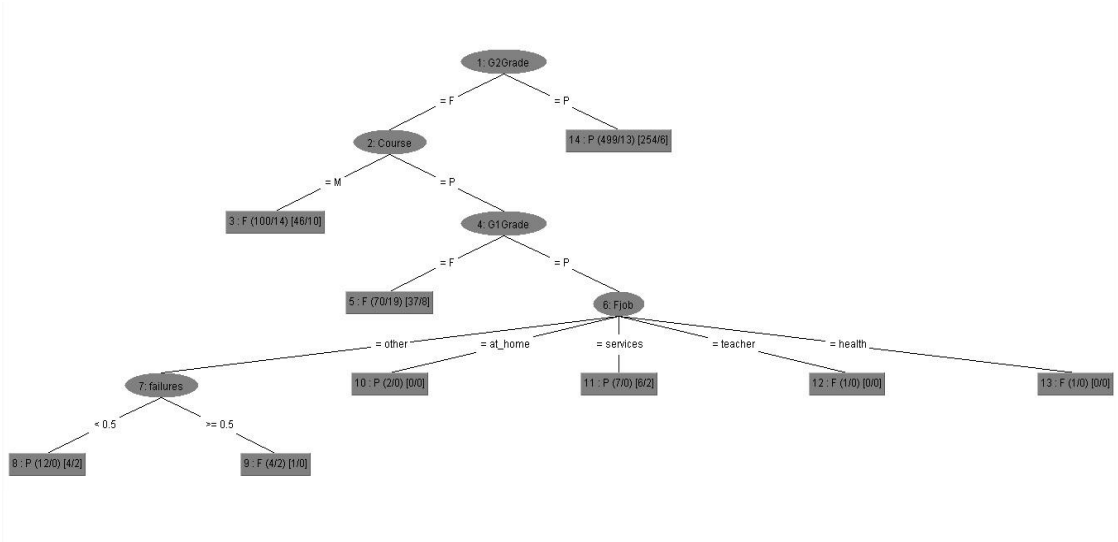

Figure (2): RepTree Algorithm

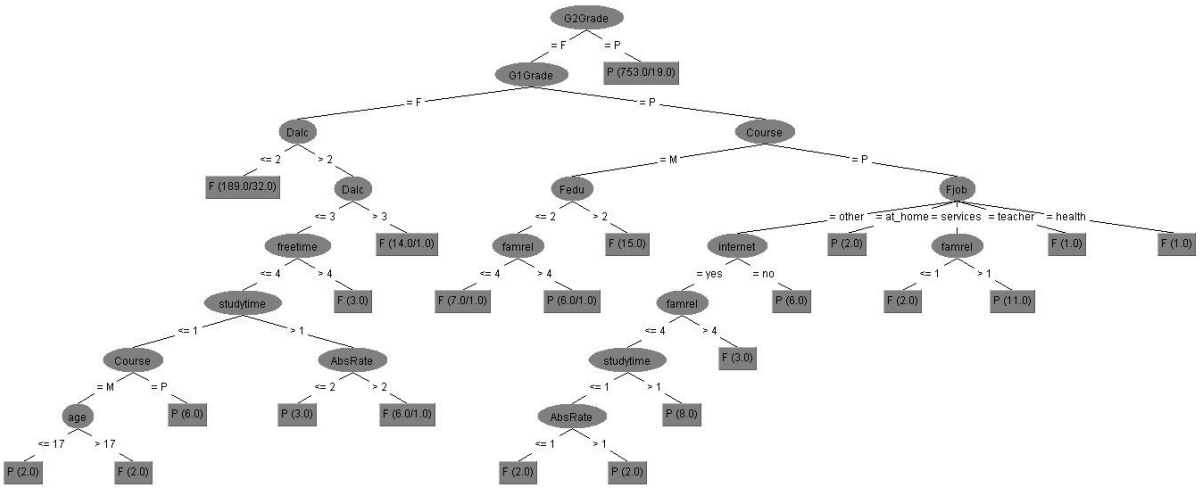

Figure (3): J48 Algorithm

VII. Conclusion

This paper listed and compared the results of implementing three different decision trees algorithms. Decision tree graphs affected by the number of input attributes and the end class attribute. Two main classes (Success of the student (G3Grade) and Willingness of studying higher education (higher)) are chosen to build the tree graph. The results showed J48 is the best decision tree algorithm which can be used as a prediction and classification road map of student's action. The selection of J48 came from the compared results beside the number of nodes in the graph which affected on the visibility of the tree.

References

[1] Romero, Cristóbal, et al. "Data mining algorithms to classify students." Educational Data Mining 2008. 2008.
[2] Osmanbegović, Edin, and Mirza Suljić. "Data mining approach for predicting student performance." Economic Review 10.1 (2012).
[3] Cortez, Paulo, and Alice Maria Gonçalves Silva. "Using data mining to predict secondary school student performance." (2008).
[4] Kabakchieva, Dorina. "Predicting student performance by using data mining methods for classification." Cybernetics and information technologies 13.1 (2013): 61-72.
[5] http://www.cs.waikato.ac.nz/~ml/weka/
[6] Kumar, S. Anupama, and M. N. Vijayalakshmi. "Efficiency of decision trees in predicting student's academic performance." First International Conference on Computer Science, Engineering and Applications, CS and IT. Vol. 2. 2011.
[7] Witten, Ian H., and Eibe Frank. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, Second Edition, 2005.
[8] Witten, Ian H., and Eibe Frank. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, Third Edition, 2011.
[9] Wilkinson, Leland. "Classification and regression trees." Systat 11 (2004): 35-56.
[10] Baker, R. S. J. D. "Data mining for education." International encyclopedia of education 7 (2010): 112-118.

[11] Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[12] Baker, Ryan SJD, and Kalina Yacef. "The state of educational data mining in 2009: A review and future visions." JEDM-Journal of Educational Data Mining 1.1 (2009): 3-17.

[13] Ramaswami, M., and R. Bhaskaran. "A study on feature selection techniques in educational data mining." arXiv preprint arXiv:0912.3924 (2009).

[14] Hulten, Geoff, Laurie Spencer, and Pedro Domingos. "Mining time-changing data streams." Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2001.