Comparative Study of Decision Tree-based models for Predicting Student Academic Achievement in WASSCE

Odeniyi, Latifat A. & Tunwase, A.O.
Department of Computer Science
Chrisland University
Abeokuta, Ogun State, Nigeria
E-mails: lodeniyi@chrislanduniversity.edu.ng; odeniyilateefah95@gmail.com; unwase@chrislanduniversity.edu.ng
Phone: +2348038457760; +2347061945880

ABSTRACT

Education in Nigeria has transcendent beyond an ordinary man's understanding and is of great importance to every nation which makes it attract considerable attention. Data mining is an essential phase in discovering knowledge in databases which is normally used in extracting hidden, useful patterns and knowledge from large data. Educational Data Mining (EDM) is a research area which concentrated on data mining, machine learning, and statistics in order to generate information from an educational dataset. The main objective of this paper is to compare the performance accuracy of some decision tree models from classifiers like C4.5 (also known as J48), SimpleCart, RandomTree and DecisionStump using WEKA Data Mining software for the implementation. The data collected were transformed in a form that is acceptable to the data mining software and it was then split into two sets; the training dataset and the testing dataset so that it can be acceptable to the system. The training dataset was used to train the system so as to observe the relationship between input dataset and the resulting outcome which provide a future prediction. The testing dataset contains data for testing the performance of the model. This model will be useful by decision-makers, stakeholders in the educational industry as well as parents, students and government to decide educational progress in Nigeria. RandomTree model exhibits good performance in the prediction of student success in WASSCE because of the low values generated in Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE)

Keywords: Student Academic Achievement, Decision Tree Algorithms, SimpleCart, RandomTree, Decision Stump.

iSTEAMS Proceedings Reference Format
Odeniyi, L.A. & Tunwase, A.O. (2019) Comparative Study of Decision Tree-based models for Predicting Student Academic Achievement in WASSCE Proceedings of the 15th iSTEAMS Research Nexus Conference, Chrisland University, Abeokuta, Nigeria, 16th – 18th April, 2019. Pp 247-256 www.isteams.net - DOI Affix - https://doi.org/ 10.22624/AIMS/iSTEAMS-2019/V15N1P23

1. BACKGROUND TO THE STUDY

Education is of great importance to every nation and therefore, attracts considerable attention at all levels of development and therefore need to discussed, planned and processed. Education makes both the person and the entire nation and also influences values and attitudes. The professions are similarly built through training and preparing people for different careers in life. Contemporary issues in Nigerian education have become food for thought for all meaning individuals both at home and Diaspora. It has several times been argued that Nigeria education system is at the cross-road due to recurrent student failure at the national examinations and subsequent drop out during undergraduate programmes. However, the revitalization of the lost hope rests upon the shoulders of people of wisdom and knowledge in Nigeria education industry.

Poor capacity for educational planning, administration, and management have manifested from one government to the other, which are regarded as the sine qua non of successful implementation of educational innovations. The ability to evaluate and predict student's performance is very important in educational environments because it plays an important role in producing the best quality students who will become the great leader of tomorrow and source of manpower for the country. Therefore, the performance of students in the National examination is of utmost concern. Discovering knowledge for prediction the result of the students, and prediction about student's performance is information hidden within the educational dataset. This hidden information can be extracted through data mining techniques (Asanbe et. al., 2016).

One way to achieve the greatest level of quality in the education system is by discovering knowledge from educational data to study the main attributes that may affect the students' performance and academic achievement. The discovered knowledge can be used as recommendations to the academic planners in education institutes to improve on their decision-making process, and also to increase students' academic achievement and reduce failure rate, to better understand students' behaviour, to assist instructors, to improve teaching and many other benefits (Kumar and Chadha, 2011). Data mining is a step in the Knowledge Discovery in Database (KDD) process that consists of applying data analysis and discovery algorithms under acceptable computational efficiency limitations, produce a particular enumeration of patterns (or models) over the data. Data mining is the extraction of hidden predictive information from large databases (Han, Kamber and Tung, 2001). The commonly used data mining technique, which develops a model that can classify the population of records at large can be referred to as classification and it frequently uses a decision tree or neural network-based classification algorithms for its classification. The data classification process involves learning and classification. In Learning, the training data are analyzed using various classification algorithm(s). In classification, test data are used to calculate the accuracy of the classification rules. If the accuracy is acceptable the rules can be applied to the new data tuples (Brijes and Saurabh, 1996). It consists of predicting the value of a categorical attribute based on the value of other attributes. Classification methods like decision trees, rule mining, Bayesian network and so on can be applied to educational data for predicting the students' behaviour, performance in an examination and so on.

Decision Tree algorithm is a well-known data mining classification tools which have the advantage of easy to use and understand for creating and displaying the results (Pornnapadol, 2004). In data mining, decision tree algorithms are very popular since they are relatively fast to train and use. A decision tree is a tree in which each branch node represents a choice between a number of alternatives, and each leaf node represents a decision (Margret, 2006). A decision tree is commonly used for gaining information for the purpose of decision making (Quinlan,1986). Decision tree always starts its classification from the root node which is for users to take actions. From this node, users split each node recursively according to the decision tree learning algorithm. The final result is a decision tree in which each branch represents a possible scenario of decisions and its outcome. The most widely used decision tree learning algorithms are ID3, C4.5, and CART.

In data mining, Decision trees are the most popular classification technique which mostly represents the group of classification rules in a tree form, and has several advantages over other techniques as stated in (H¨am¨al¨ainen and Vinni, 2010). The simplicity of its presentation makes it easy to understand and it works with different types of attributes; nominal or numerical. Considering a dataset of attributes with its classes, a decision tree produces sequences of rules which can be used to generate various classes for decision making. The Decision tree method has gained popularity due to its high accuracy of classifying the data set (Nithyassik, Nandhini, and Chandra, 2010).

1.1 Statement of Problem

The ability to evaluate and predict student's performance is very important in educational environments because it plays an important role in producing the best quality students who will become a great leader of tomorrow and source of manpower for the country. There is a wealth of data available for use in educational data mining and little has been done to use this available data to solve the challenges that face students in achieving better grades in national examination like WASSCE and NECO because there is lack of effective analysis tools to discover hidden relationships and trends in educational data.

1.1 Objective

The main objective of this paper is to compare the performance accuracy of some decision tree models from classifiers like C4.5 (also known as J48), SimpleCart, RandomTree and DecisionStump using WEKA Data Mining software for the implementation.

2. RELATED WORKS

Yadav (2011) researched on the topic data mining applications: a comparative study for predicting student's performance. Student academic performance was predicted using various data mining algorithms like ID3, C4.5 and CART. CART algorithm was classified as the best for prediction of student academic performance. Romero, Ventura, Espejo, and Hervás, (2008) in the paper titled data mining algorithms to classify students compared data mining algorithms and it was discovered that CART and C4.5 decision tree algorithms were ranked as the best for data classification purpose. Kovačić, (2010) also applied classification algorithms to predict student academic success in his paper titled early prediction of student uccess: "mining students enrolment data". He considered socio-demographic variables dataset used were collected from Open Polytechnic of New Zealand while four classification trees, namely: CHAID, exhaustive CHAID, QUEST, and CART were adopted in the study and from the result, CART had the best accuracy with 60.5%.

Quadri and Kalyankar, (2010) carried out a research on applied decision trees to analyze the problem of dropouts in any higher educational institution. Decision trees are used to make important design decisions and explain the interdependencies among the properties of drop out students; providing an instance machine learning technique that can be used to improve the effectiveness and efficiency of a modeling process. The study addresses the capabilities and strengths of the decision tree algorithm in identifying drop out students to guide the teachers in concentrating on appropriate features associated with counseling students or arranging financial aid to them. Paris *et al*, (2010) presented a paper titled improving academic performance prediction using voting technique in data mining. The paper evaluated the performance of different prediction techniques for prediction of students' CGPA. Decision trees and Bayesian methods were employed and proposed voting technique accuracy was compared with C4.5, NBTree, BayesNet, naïve Bayes and hidden naïve Bayes (HNB). Voting technique also performed on three (3) weak classifiers (naïve Bayes, OneR and Decision stump). From the result, it was realised that HNB performed well while decision stump classifier was next to HNB in term of performance.

Affendey *et al* (2010) in his paper titled ranking of influencing factors in predicting students' academic performance used ranked students registered courses as attributes to predict student grade at graduation to determine students' academic performance. It was discovered from the results that Naïve Bayes, AODE and RBF Network had an accuracy of 95.29% when CFS was used as attribute selection, while, AODE had an accuracy of 95.29% when CoE was used as attribute selection. Al-Radaideh, Al-Shawakfa, and Al-Najjar (2006) conducted a research with the topic mining student data using decision trees.

From the study, ID3 algorithm, C4.5, and Naïve Bayes algorithm were used for classification and the result showed that C4.5 decision tree had better results for student performance prediction. Osofisan et al., (2014) researched on an educational data mining with the title empirical study of decision tree and artificial neural network algorithm for mining educational database. From their study, they compared decision tree algorithm and neural network with the aim of determining the best algorithm for classification and prediction. The result showed that Neural Network out-performed decision tree in both classification and prediction.

3. METHODOLOGY

The research makes use of educational data mining to predict students' academic achievement. The Knowledge Discovery in Database (KDD) methodology was adopted which consists of data collection and pre-processing, building of model, model evaluation and finally, prediction of student academic achievement.

3.1 Research Design

Dataset used for this research was collected from WASSCE result of students from five private and five public senior secondary schools in Ibadan South East Local Government of Oyo State. Initially, more than 20 attributes were collected and some of the attributes were eliminated since they were considered as irrelevant (noise) to the study. Finally, eight (8) out of 20 conditional attributes were used and one was considered as class attribute for the model classification. The attributes were described in the table 1 below.

Table 1: Students attributes and their description

Attributes used	Description of the attributes	Output values
School-Environment	This explains how conducive is the school environment in term of infrastructures like the seat, tables to write, good classroom structure, ventilation, writing board, and necessary amenities.	Conducive, Nonconducive and Partially-Conducive
Nature of School	The previous school the students attended their junior secondary education	Private, Public
Type-of-School	The type of school where the student sat for their examinations	Private, Public
Class type	This explains the type of class of the students	Science, Commercial, Arts
Sex	The sex of the student either male or female	Male, Female
Mode	This variable indicates student mode of study	Day, Boarder
Student-Age	Age of student on entry to senior secondary school	13,14,15,16 and 17
Result	the Result of students from WAEC	Pass, fail

4. DATA COLLECTION AND PRE-PROCESSING

Collection of data for this study was done through direct collection from the schools repository of data for five years. Some missing and incomplete data were removed from the dataset during pre-processing in order to improve quality of the data to be used since this will determine the classification performance of the model. The data were later integrated (both from private and public school) into a comprehensive dataset using Microsoft Excel 2007. The dataset was entered in WEKA acceptable format which is Comma Separated Value (CSV).

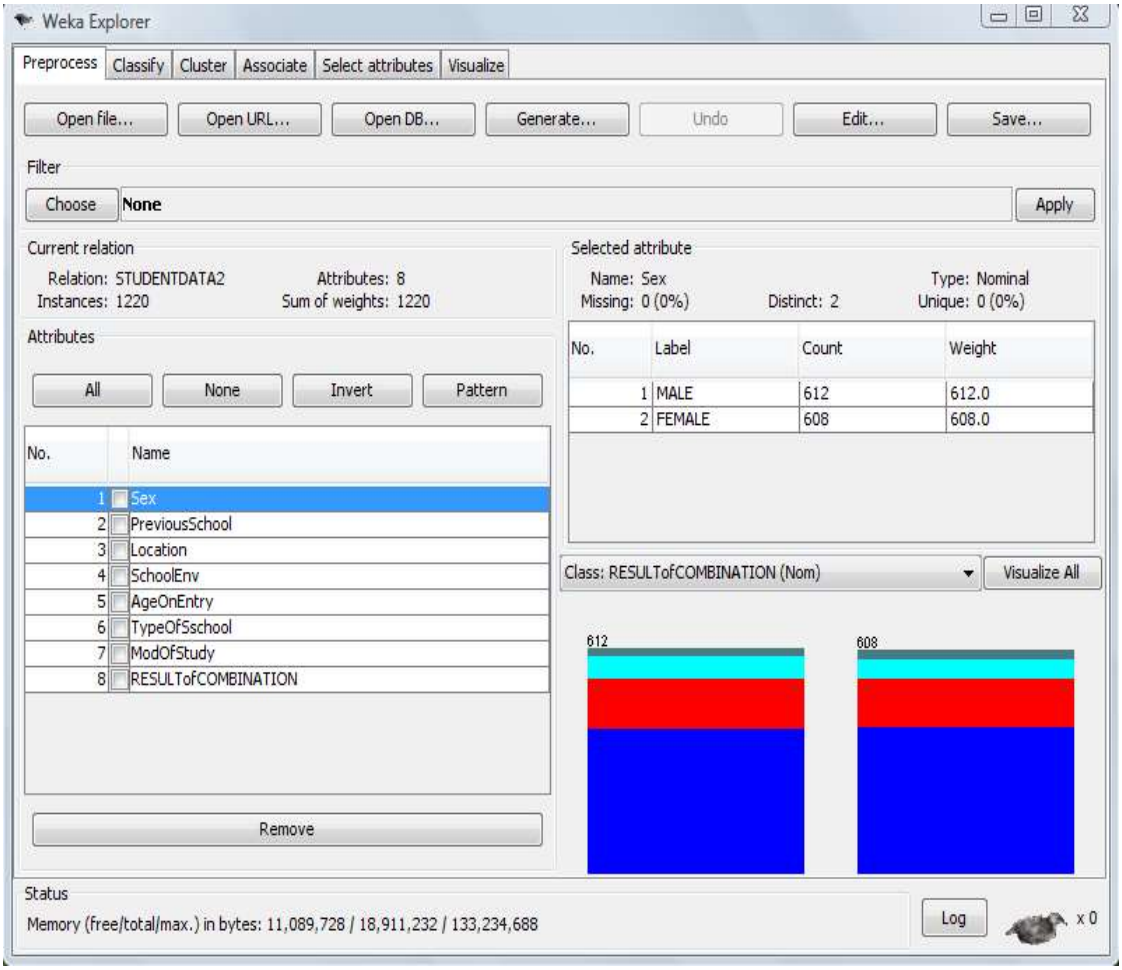

Figure 1 displays the Pre-process stage of data

4.1 Building of model

In this stage, classification model was built using decision tree because decision tree was is a good and practical method since it is relatively fast, and can easy to use and understand and can as well be easily converted to simple classification rules. Information gain metric which determines the attribute that is most useful and gain ratio was used to rank attributes and to build the decision tree where each attribute is located according to its gain ratio. The attribute ranked first was considered as the root-node of the decision tree for each model.

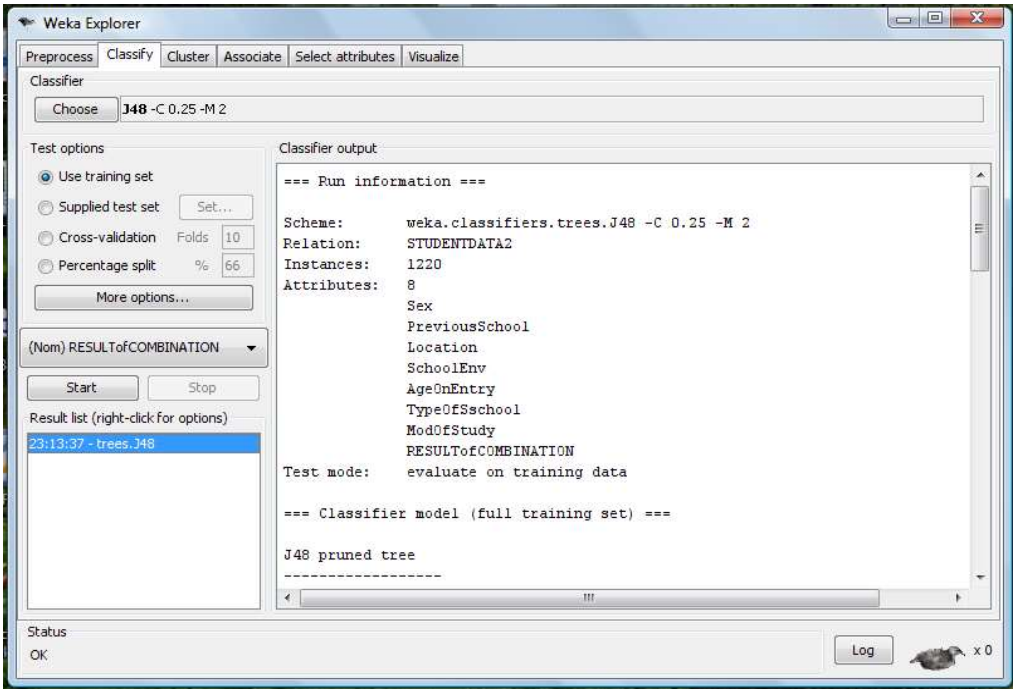

Figure 2 displays the C4.5 (J48) classifier model building

4.2 Description of (Waikato Environment for Knowledge Analysis) WEKA Software Tool

(Waikato Environment for Knowledge Analysis) WEKA was used for analysis in this paper and here are the steps involved in the process of its use:

1. The user launches WEKA
2. Use one of the 4(four) applications present on the WEKA GUI Chooser
3. Assuming Explorer was chosen among the application, then there is a need to start with preprocessing in order to load data saved in comma separated value(CSV). The class attribute should be the last column.
4. All data could be visualized in the preprocessing section of the application.
5. Now to classification, there is a test option for data depending on the chosen option. The options are: Use training set, Supplied test set, Cross-validation, and Percentage split.
6. Select the classifier that you prefer among the pool of classifiers present in WEKA. Though, different classifiers can be tested to choose the most accurate among them.
7. Among the icon in Explorer is a select attribute which can be used to determine the most relevant attribute to the study.
8. The User can as well choose to generate rules from classifier icon, to choose the most relevant rules from the rules generated.
9. The Tree can be generated by right-clicking on a particular classifier after training with such classifier.
10. Then the user can now develop its own tree from the result gathered from the training with WEKA applications as well as predict from the gathered information.

Figure 3 WEKA software tool interface

5. DATA PRESENTATION

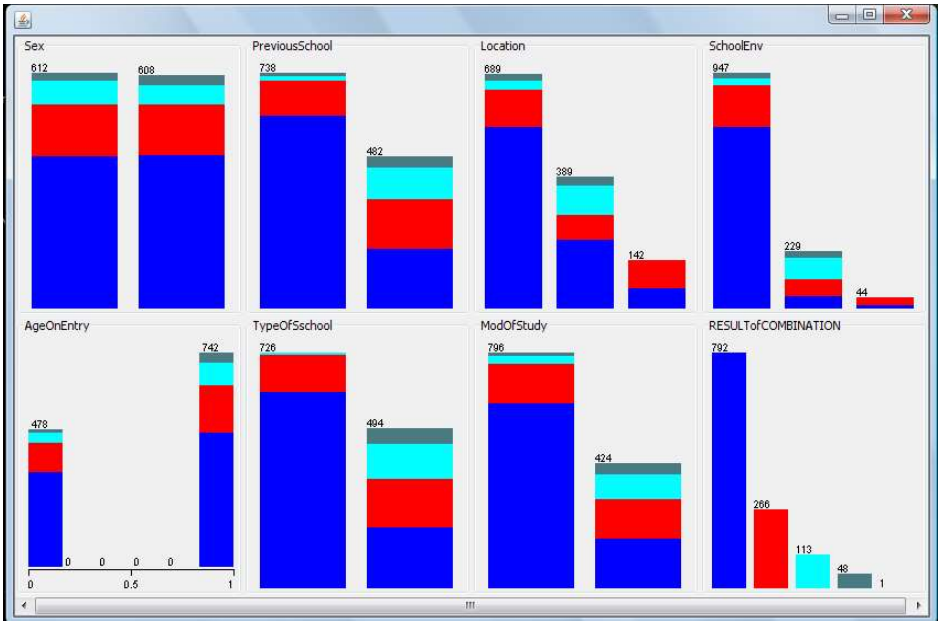

Figure 4 displays the visualize all feature of some of the attributes used in the research

6. DISCUSSION OF FINDINGS

The following metrics were adopted in determining the performance of the models to determine which of them perform better in predicting students' academic achievement: Time taken to build the model,correctlt classified instances of each model, incorrectly classified instances of each model and performance accuracy of each model. The findings from the instances was shown and discussed in Table 2 below.

Table 2: performance of the classification models

Criteria considered	C4.5 (J48)	SimpleCart	RandomTree	Decisionstump
Time taken to build each model	0.05 seconds	0.03 seconds	0.05 seconds	0.02 seconds
Correctly-classified instances of each model	893	892	899	792
Incorrectly-classified instances of each model	327	328	321	428
Generated accuracy for each model	73.1967%	73.1148%	73.6885%	64.918%

Among the tested classifiers, Random tree was classified as the most accurate classifiers considering the number of data that was correctly classified as well as the root mean square error of the algorithm.. So RandomTree classifier has more accuracy compared to C4.5, SimpleCart, and decisionstump classifiers but from the table, decisionStump takes the shortest time to build the model compared to others. Moreover, results from WEKA classifiers, and Attribute Ranking indicates that among the factors that contribute to the difference in the two results are school environment, type of school and previous school. It was also deduced that student from private school with conducive atmosphere for learning pass than their counterpart from public school.

7. CONCLUDING REMARKS

In conclusion, four classifiers such as SimpleCart, C4.5, RandomTree, and DecisionStump were used for predicting student success in the national examination and it was observation that RandomTree has more accuracy than the remaining classifiers. Randomtree classifier implemented on the attributes had classification accuracy of 73.6885 % but in terms of speed, DecisionStump classifier was found to be the best as it took only 0.02 second to train the data while RAndomTree used seconds. The best model selected for predicting student academic achievement in WAEC national examination as discovered from this study could not exceed a classification accuracy of 73.6885 % and still much remains to fill the gap of 26.3115 % misclassified cases. Therefore, data mining techniques can be used efficiently to model and predict student academic data cases. The outcome of this study can be used by government at all levels to decide education progress in Nigeria.

8. CONTRIBUTIONS TO KNOWLEDGE

This study has indicated that data mining techniques can be applied in the prediction of typhoid fever and the resulting models of this study are worthy of educational use. To improve the classification accuracy of the models' further researches should be conducted using different classification algorithms and other data mining techniques such as Naïve Bayes classifier, genetic algorithm, etc. can be used for prediction. Finally, expanded data set with more distinctive attributes to get more accurate results can also be used to carry out predictions to improve the classification accuracy.

REFERENCES

1. Affendey, L.S, Paris I.H, Mustapha N, Suleiman M.N., and Muda Z., (2010) , Ranking of influencing factors in predicting students' academic performance, *Journal of Information Technology*.; Vol. 9, No. 4, pp832–837
2. Al-Radaideh, Q. A., Al-Shawakfa, E. W and Al-Najjar, M. I.(2006), "Mining student data using decision trees", International Conference on Information Technology (ACIT'2006), Yarmouk University, Jordan.
3. Asanbe M.O., Osofisan A.O. and William W.F. (2016), Teachers Performance Evaluation in Higher Educational Institution using data mining technique, *International Journal of Applied Information Systems (IJAIS),* New York, USA, Vol. 10, No. 7, pp 2249-0868.
4. Ashish Kumar, Pranav Bhatia, Anshul Goel, Silica Kole (2015), Implementation and Comparison of Decision Tree Based Algorithms, *International Journal of Innovations & Advancement in Computer Science (IJIACS)*
5. Brijesh Kumar Baradwaj and Saurabh Pal(2011),"Mining Educational Data to Analyze Students' Performance**",** *International Journal of Advanced Computer Science and Applications*, Vol. 2, No. 6.pp 63-69.
6. Han, J., M. Kamber, and Tung, A. K. H. (2001), "Spatial Clustering Methods in Data Mining: A Survey", H. Miller and J. Han (eds.), Geographic Data Mining and Knowledge Discovery, Taylor and Francis.
7. Kumar, V. and Chadha, A. (2011), 'An Empirical Study of the applications of Data Mining Techniques in Higher Education', *International Journal of Advanced Computer Science and Applications*, vol. 2, no. 3, pp. 80-84
8. Margret H. Dunham, "Data Mining: Introductory and advance topic", Pearson Education, India, 2006.
9. Nithyassik, B., Nandhini, and Chandra, E. (2010),"Classification Techniques in Education Domain", International Journal on Computer Science and Engineering, Vol. 2, No.5, pp.1647-1684.
10. Osofisan A.O., Adeyemo Omowunmi.O. & Oluwasusi, S.T.,(2014). Empirical Study of Decision Tree and Artificial Neural Network Algorithm for Mining Educational Database, *African Journal of Computing & ICT Vol 7. No. 2*
11. Paris I.H., Affendey L.S., and Mustapha N., (2010) Improving academic performance prediction using voting technique in data mining, *Journal of World Academy of Science, Engineering and Technology*, Vol. 62, pp 820–3.
12. Pornnapadol, "Children who have learning disabilities", Child and Adolescent Psychiatric Bulletin Club of Thailand, October-December, 2004, pp.47-48..

13. Quadri, M. N. and Kalyankar, N.V. (2010), Drop Out Feature of Student Data for Academic Performance Using Decision Tree Techniques, *Global Journal of Computer Science and Technology*, Vol. 10 Issue 2 (Ver 1.0).
14. Quinlan, J.R. (1986),"Induction of Decision Tree", *Journal of Machine learning*, Morgan Kaufmann Vol.1, pp.81-106.
15. Romero, C. and Ventura, S. (2007), 'Educational data mining A Survey from 1995 to
16. 2005', Expert Systems with Applications (33), pp. 135-146.
17. Romero, C., Ventura, S., Espejo, P. G., and Hervas, C. (2008), "Data mining algorithms to classify students", International Conference on Educational Data Mining, No. 1, pp 8-17.
18. Yadav, S. K. Bharadwaj, B.K. and Pal, S. (2011), Data Mining Applications: A comparative study for Predicting Student's Performance*, International Journal of Innovative Technology and Creative Engineering (IJITCE)*, Vol. 1, No. 12, pp. 13-19.