Volume 6 No. 4, April 2016 ISSN 2223-4985

International Journal of Information and Communication Technology Research

©2016 ICT Journal. All rights reserved

http://www.esjournals.org

Data Mining: A Prediction for Academic Performance Improvement of Science Students using Classification

I.A Ganiyu
Department of Computer Science, Ramon Adedoyin College of Science and Technology,
Oduduwa University, Ipetumodu, Ile-Ife, Osun State, Nigeria.

ABSTRACT

It is highly important to evaluate and predict Student's Academic performance in Academic settings because it plays an important role in guiding the Students towards becoming great leaders of tomorrow and source of manpower for the country. The amount of data stored in educational database increasing rapidly. These databases contain hidden information for improvement of students' performance. Data mining Classification Techniques like decision trees, Bayesian network etc. can be applied on the educational data for predicting the student's performance in examination. The C4.5, REP TREE, RANDOM TREE and Decision Stump are applied on science Student's data to predict their performance in first semester examination. The outcome of the decision tree predicted the likely performance score of a student taking the courses. The results will serve as a guide to the junior Student who are just coming to the system to prepare well in courses where the Students scored lower marks and aid the Education Management Board to know Lecturers that are applying the best teaching style in carrying out the objectives of learning which is to pass the knowledge across to the Student also for the Lecturers to check their teaching style for improvement. After the declaration of the results in the final examination, the marks obtained by the students were fed into the software system and the results were analyzed for the next examination. The comparative analysis of the results states that the prediction has helped the week Students to improve and concentrated more in some courses which has brought out betterment in the result.

Keywords: *Educational data mining; Decision tree; C4.5 algorithm; Random Tree algorithm; Rep Tree; Decision Stump: Prediction.*

1. INTRODUCTION

The major reason why data mining has attracted a great deal of attention in the information industry in recent years is due to the wide availability of huge amounts of data and the imminent need for turning such data into useful information and knowledge. The information and knowledge gained can be used for applications ranging from business management, production control, and market analysis, to engineering design and science exploration.

Statisticians were the first to use the term data mining. Originally, data mining or data dredging was a derogatory term referring to attempts to extract information that was not supported by the data. Today, data mining has taken on a positive meaning. Now, statisticians view data mining as the construction of a statistical model, that is, an underlying distribution from which the visible data is drawn.

Educational data mining has emerged as an independent research area in recent years, culminating in 2008 with the establishment of the annual International Conference on Educational Data Mining, and the Journal of Educational Data Mining.

Data mining in higher education is a recent research field and this area of research is gaining popularity because of its potentials to educational institutes. (Romero and Ventura, 2007), have a survey on educational data mining between 1995 and 2005. They concluded that educational data mining is a promising area of research and it has a specific requirements not presented in other domains. This, work is in the purview of educational of data mining.

Decision Trees are tree-shaped structures that represent decision sets. It uses real data-mining algorithms to help with classification. A decision-tree process will generate the rules followed in a process. Decision trees are useful for helping you choose among several courses of action and enable you to explore the possible outcomes for various options in order to assess the risk and rewards for each potential course of action. These decisions generate rules, which then are used to classify data. Decision trees are the favoured technique for building understandable models.

A.O.Osofisan, O.O.Adeyemo & S.T. Oluwasusi (2014) said a decision tree is a flow-chart-like tree structure, where each internal node is denoted by rectangles, and leaf nodes are denoted by ovals. All internal nodes have two or more child nodes. All internal nodes contain splits, which test the value of an expression of the attributes. Arcs from an internal node to its children are labelled with distinct outcomes of the test. Each leaf node has a class label associated with it. Decision trees are powerful and popular for both classification and prediction. The attractiveness of tree-based methods is due largely to the fact that decision trees represent rules. Rules can readily be expressed in English so that humans can understand them.

Decision trees are produced by algorithms that identify various ways of splitting a dataset into branch-like segments. These segments form an inverted decision tree that originates with a root node at the top of the tree. The object of analysis is reflected in this root node as a simple, one-dimensional display in the decision tree interface. The name of the field of data that is the object of analysis is usually displayed, along with the spread or distribution of the values that are contained in that field.

The most frequently used decision tree algorithms are:

- ID3
- C4.5 and
- CART

A. ID3 (Iterative Dichotomiser 3)

This is a decision tree algorithm introduced in 1986 by Quinlan Ross (Quinlan, 1986). It is based on Hunts algorithm. The tree is constructed in two phases. The two phases are tree building and pruning.

ID3 uses information gain measure to choose the splitting attribute. It only accepts categorical attributes in building a tree model. It does not give accurate result when there is noise. To remove the noise pre-processing technique has to be used.

To build decision tree, information gain is calculated for each and every attribute and select the attribute with the highest information gain to designate as a root node. The possible values of the attribute are represented as arcs. Then all possible outcome instances are tested to check whether they fall under the same class or not. If all the instances fall under the same class, the node is represented with single class name, otherwise a splitting attribute is chosen to classify the instances.

Continuous attributes can be handled using the ID3 algorithm by discretizing or directly by considering the values to find the best split point by taking a threshold on the attribute values. ID3 does not support pruning.

B. C4.5

This algorithm is a successor to ID3 developed by Quinlan Ross (Quinlan, 1986). It is also based on Hunt’s algorithm.C4.5 handles both categorical and continuous attributes to build a decision tree. In order to handle continuous attributes, C4.5 splits the attribute values into two partitions based on the selected threshold such that all the values above the threshold as one child and the remaining as another child. It also handles missing attribute values. C4.5 uses Gain Ratio as an attribute selection measure to build a decision tree. It removes the biasness of information gain when there are many outcome values of an attribute.

The first step is to calculate the gain ratio of each attribute. The root node will be the attribute whose gain ratio is maximum. C4.5 uses pessimistic pruning to remove unnecessary branches in the decision tree to improve the accuracy of classification

C. CART (Classification And Regression Trees)

CART is also based on Hunt’s algorithm (Quinlan, 2000). CART handles both categorical and continuous attributes to build a decision tree. It handles missing values. CART uses Gini Index as an attribute selection measure to build a decision tree unlike ID3 and C4.5 algorithms, CART produces binary splits. Hence, it produces binary trees. Gini Index measure does not use probabilistic assumptions like ID3, and C4.5. CART uses cost complexity pruning to remove the unreliable branches from the decision tree to improve accuracy.

2. RELATED WORKS

Surjeet K. Y and Saurabh P (2012) applied Classification methods like C4.5, ID3 and CART decision tree algorithms on engineering student’s data to predict their performance in the final exam. The outcome of the decision tree predicted the number of students who are likely to pass, fail or promoted to next year.

Khan Z. N (2005) conducted a performance study on 400 students comprising 200 boys and 200 girls selected from the senior secondary school of Aligarh Muslim University, Aligarh, India with a main objective to establish the prognostic value of different measures of cognition, personality and demographic variables for success at higher secondary level in science stream. The selection was based on cluster sampling technique in which the entire population of interest was divided into groups, or clusters, and a random sample of these clusters was selected for further analyses. It was found that girls with high socio-economic status had relatively higher academic achievement in science stream and boys with low socioeconomic status had relatively higher academic achievement in general.

Chin Chia Hsu and Tao Huang (2006) applied data classification and decision tree methods in order to improve the student performance. The data set used was obtained from M.Sc. IT department of Information Technology 2009 to 2012 batch. Extracurricular activities were also included. The information generated after the implementation of the data mining techniques will help the teachers to predict those students who have lesser performance and also to develop them with special attention.

Students’ academic performance in higher education is affected by various socioeconomic, psychological, and environmental factors (Hijazi & Naqvi, 2006). It is always in the best interest of educators to measure students’ academic performance. This allows them to evaluate not only students’ knowledge levels but also the effectiveness of their own teaching processes, and perhaps, provide a gauge of student satisfaction.

Bhardwaj and Pal (2011) conducted a performance analysis on 50 students whose records were taken from VBS Purvanchal University, Jaunpur (Uttar Pradesh) with the objective to study student's performance using 8 attributes. Decision tree method was used to classify the data. Study helped teachers to improve the result of the student.

Yadav, Brijesh and Pal (2011) conducted study to predict students' performance with 48 students dataset and 7 attributes obtained from VBS Purvachal University,Janupur (UP),India on the sampling method of computer Applications department of course MCA(master of Computer Applications) from session 2008 to 2011.Different Decision tree algorithms like ID3,C4.5,CART were used for classification. Results show that CART is the best
Algorithm for classification of data. This study will help teachers to identify those students who need special attention and also this work will help to reduce fail ratio.

Pandey and Pal (2011) conducted study on the student performance based by selecting 600 students from different colleges of Dr. R. M. L. Awadh University, Faizabad, India. By means of Bayes Classification on category, language and background qualification, it was found that whether new comer students will performer or not.

Z. J. Kovacic (2010) presented a case study on educational data mining to identify up to what extent the enrolment data can be used to predict student's success. The algorithms CHAID and CART were applied on student enrolment data of information system students of open polytechnic of New Zealand to get two decision trees classifying successful and unsuccessful students. The accuracy obtained with CHAID and CART was 59.4 and 60.5 respectively.

Al-Radaideh, et al (2006). applied a decision tree model to predict the final grade of students who studied the C++ course in Yarmouk University, Jordan in the year 2005. Three different classification methods namely ID3, C4.5, and the Naïve Bayes were used. The outcome of their results indicated that Decision Tree model had better prediction than other models.

Bharadwaj and Pa (2011). Obtained the university students data like attendance, class test, seminar and assignment marks from the students' previous database, to predict the performance at the end of the semester.

3. MATERIALS AND METHODOLOGY

This section explains the methodology adopted in this work.

3.1 Data Collection

In this study, the data set used was obtained from Oduduwa University Ile-Ife. Osun State Nigeria in Ramon Adedoyin College of Science and Technology (RACOSAT) for Three (3) First Semesters (2012/2013, 2014/2015 & 2015/2016) 100 Level.

The four classifiers (ID3, J48 an implementation of C4.5, Bayes Net and Naïve Bayes) were used in the WEKA toolkit. The test dataset has 5,860 instances of Distinct 577 records of Student that sat for an average minimum of 1 course and maximum of 12 courses in a semester with 8 attributes. Here only 8 attributes have been selected which are required for data mining. The information was selected from the student information system and stored in an excel file which then was converted into .csv file. The student related attributes have been shown in the table 1 given below with the necessary description and the domain values, these attributes with their descriptions are specified in Table 1.

Table 1: Data Fields

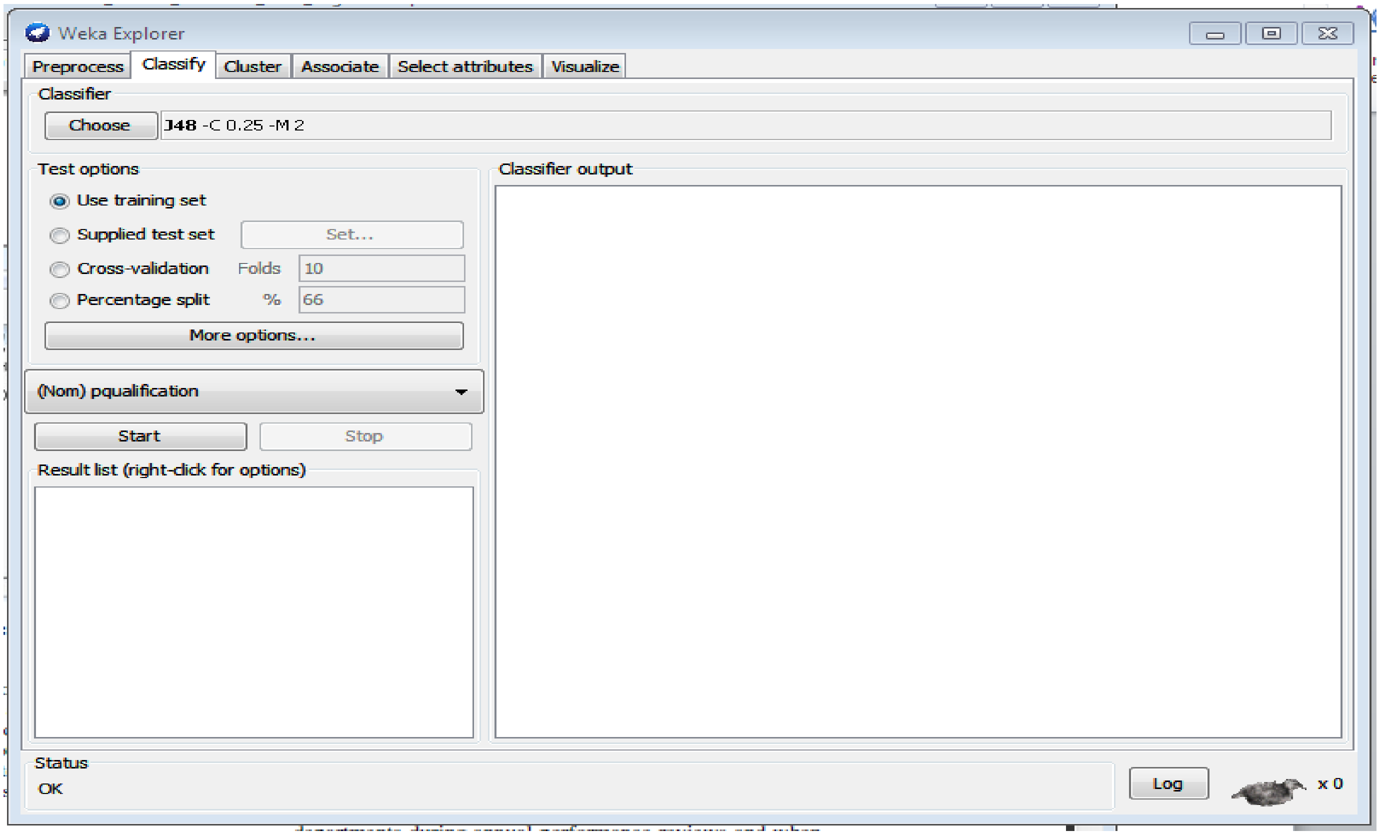
Field	Description
Name	Name of the Student
Matric Number	Unique Key that identify a student.
Department	Department where the Student belongs
Session	Session of the examination
Level	Level of the student.
Semester	semester of the examination
Course	Name of the course i.e. (CHM 103, PHY 107, PHY 101, CHM 101 ,BIO 107, MTH 103, CHM 107, GST 101, GST 105, BIO 101, MTH 101, STA 111, GST 107, CSC 101)
Score	Score of a Student in a course

This work was conducted using WEKA (Waikato Environment for Knowledge Analysis) version 3.7.4 - an open source software developed by a team of researchers of Waikato University and certified by IEEE. WEKA is a popular suite of machine learning software written in Java, developed at the University of Waikato, New Zealand. The WEKA workbench contains a collection of visualization tools and algorithms for data analysis and predictive modeling, together with graphical user interfaces for easy access to this functionality. It is written in Java and runs cross platform. WEKA supports several standard data mining tasks, more specifically, data preprocessing, clustering, classification, regression, visualization, and feature selection.

WEKA provides access to SQL databases using Java Database Connectivity and can process the result returned by a database query. WEKA allows many algorithms giving room for comparison to determine the better classifier among those used for the study. For the course of this research work, J48 which is an open source Java implementation of the C4.5 algorithm in the Weka data mining tool was considered because C4.5 made a number of improvements to other algorithms.

4.0 DISCUSSION OF RESULTS

Clicking on Classify tab on the panel open a menu which allows the user to choose the corresponding J48 algorithm.



J48 -C 0.25 -M 2

● Use training set

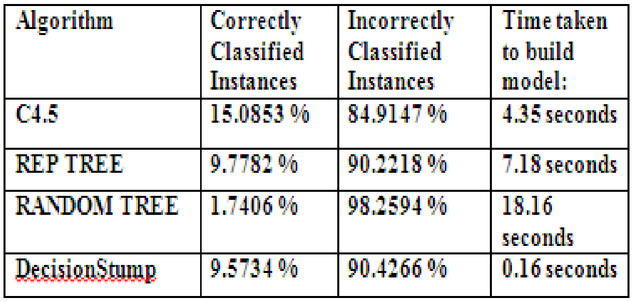
Folds 10

% 66

(Nom) pqualification

Status OK

Figure 1: Classify Tab



Algorithm	Correctly Classified Instances	Incorrectly Classified Instances	Time taken to build model:
C4.5	15.0853 %	84.9147 %	4.35 seconds
REP TREE	9.7782 %	90.2218 %	7.18 seconds
RANDOM TREE	1.7406 %	98.2594 %	18.16 seconds
DecisionStump	9.5734 %	90.4266 %	0.16 seconds

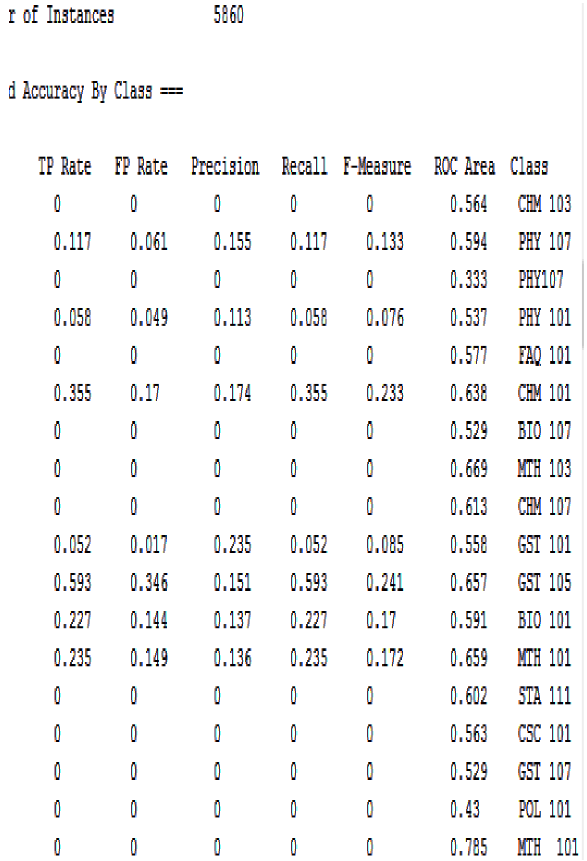
Figure 2 : Classifiers Accuracy

As shown in Figure 2 above, J48 algorithm which actually implements a later and slightly improved version called C4.5 revision 8, which was the last public version of this family of algorithms before the commercial implementation of C5.0 was released was chosen because of its best result analysis as compared to other decision tree algorithm present in Weka, The performance metrics used in assessing the performance of the classifier models are: Correctly Classified Instances, Incorrectly Classified Instances, Time taken to build model, True Positive (TP) Rate and False Positive (FP) rate.
True Positive Rate is the proportion of cases which were classified as the actual class, indicating how much part of the class was correctly captured. It is equivalent to Recall. False

Positive Rate is the proportion of cases which were classified as one class but belongs to a different class.

Precision is the proportion of the cases which truly have the actual class among all the instances which were classified as the class. F-Measure: is a combined measure for Precision and Recall and is simply calculated as: 2*Precision*Recall/ (Precision + Recall). Receiving Operating Characteristic (ROC) is the graphical display of TPR against FPR while AUC represents the area under ROC curve.

The result of its performance can be seen below



```
r of Instances          5860


d Accuracy By Class ===


   TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
     0         0         0          0        0          0.564    CHM 103
     0.117     0.061     0.155      0.117    0.133      0.594    PHY 107
     0         0         0          0        0          0.333    PHY107
     0.058     0.049     0.113      0.058    0.076      0.537    PHY 101
     0         0         0          0        0          0.577    FAQ 101
     0.355     0.17      0.174      0.355    0.233      0.638    CHM 101
     0         0         0          0        0          0.529    BIO 107
     0         0         0          0        0          0.669    MTH 103
     0         0         0          0        0          0.613    CHM 107
     0.052     0.017     0.235      0.052    0.085      0.558    GST 101
     0.593     0.346     0.151      0.593    0.241      0.657    GST 105
     0.227     0.144     0.137      0.227    0.17       0.591    BIO 101
     0.235     0.149     0.136      0.235    0.172      0.659    MTH 101
     0         0         0          0        0          0.602    STA 111
     0         0         0          0        0          0.563    CSC 101
     0         0         0          0        0          0.529    GST 107
     0         0         0          0        0          0.43     POL 101
     0         0         0          0        0          0.785    MTH  101
```

Figure 3: Results showing the performance of C4.5 (J48) classifier with Confusion Matrix.

The confusion matrix is commonly named contingency table. The number of correctly classified instances is the sum of the diagonals in the matrix; all others are incorrectly classified.

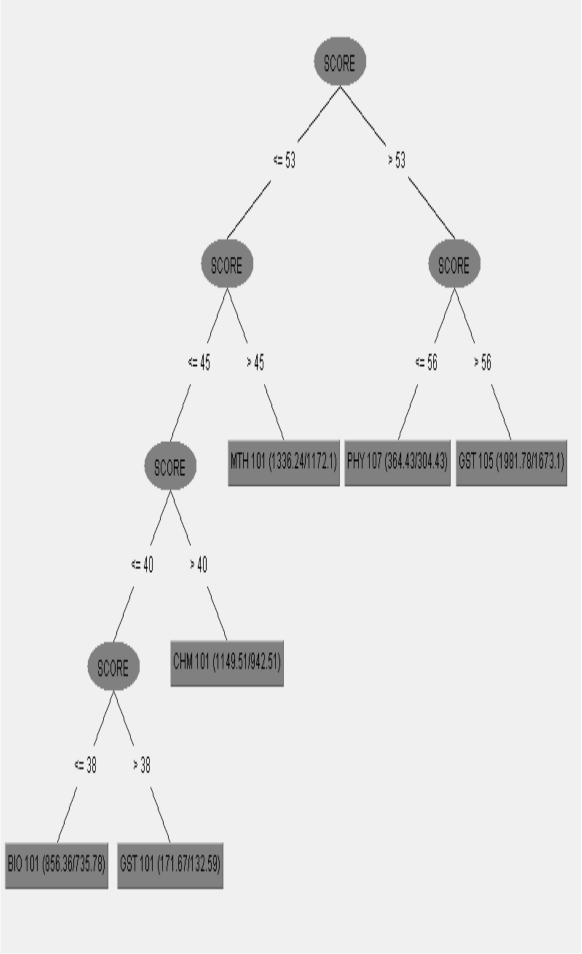


Figure 4 : Decision tree rules

5. RESULT ANALYSIS

Figure 4 is the decision tree constructed by the J48 classifier. This indicates how the classifier uses the attributes to make a decision. The leaf nodes indicate the outcome of a test, and each leaf (terminal) node holds a class label and the topmost node is the root node (Experience).

Many Rules were generated from the decision tree and it can be expressed in English so that we humans can understand them.

Decision rules

Since the decision tree can be linearized into decision rules, where the outcome is the contents of the leaf node, and the conditions along the path form a conjunction in the if clause. In general, the rules are of the form:

if condition1 and condition2 and condition3 then outcome.

The rules show that:

1. If Score is Greater than 56 then it is likely to be GST 105
2. If Score is Less than or Equal to 56 then it is likely to be PHY 107
3. If Score is Greater 45 then it is likely to be MTH 101
4. If Score is Less than or Equal to 45 and

i. Score Greater than 40 then it is likely to be CHM 101

ii. Score Less than or Equal to 40 And
   (a) Score Greater than 38 then it is likely to be GST 101
   (b) Score Less than or Equal to 38 then it is likely to be BIO 101

6. CONCLUSION

We applied data mining techniques to discover knowledge in Education domain and the result of the analysis has shown that Student passed the following courses GST 105, PHY 107, MTH 101, CHM 101 and score lesser mark in GST 101 and BIO 101. From the analysis above it can be concluded that Students should concentrate more on GST 101 and BIO101 or Lecturers taking these courses should check for the effectiveness of their own teaching processes for betterment of the Result. The future work is to deal with more records of students in order to obtain better generalization and accurate instances.

REFERENCES

[1]. Romero, C. and Ventura, S. (2007) ‘Educational data Mining: A Survey from 1995 to 2005’, Expert Systems with Applications (33), pp. 135-146

[2]. U . K. Pandey, and S. Pal, “Data Mining: A prediction of performer or underperformer using classification”, (IJCSIT) International Journal of Computer Science and Information Technology, Vol. 2(2), pp.686-690, ISSN: 0975-9646, 2011.

[3]. B.K. Bharadwaj and S. Pal. “Mining Educational Data to Analyze Students Performance”, International Journal of Advance Computer Science and Applications (IJACSA), Vol. 2, No. 6, pp.63-69, 2011.

[4]. AI-Radaideh,Q. A., AI-Shawakfa, E.M., and AI-Najjar, M. I., “Mining Student Data using Decision Trees”, International Arab Conference on Information Technology(ACIT'2006), Yarmouk University, Jordan, 2006.

[5]. Z. J. Kovacic, “Early prediction of student success: Mining student enrollment data”, Proceedings of Informing Science & IT Education Conference 2010.

[6]. P.K.Srimani and Annapurna S Kamath. ”Data Mining Techniques for the Performance Analysis of a Learning Model-A case study”International Journal of Computer Applications (0975-8887), volume 53-No 5 September 2012.

[7]. J. R. Quinlan, “Introduction of decision tree”, Journal of Machine learning”, pp. 81-106, 1986.

[8]. Surjeet K. Y and Saurabh P (2012)’ Data Mining: A Prediction for Performance
Improvement of Engineering Students using Classification” World of Computer Science and Information Technology Journal (WCSIT) ISSN: 2221-0741 Vol. 2, No. 2, 51-56, 2012

[9]. A.O. Osofisan, O.O. Adeyemo & S.T. Oluwasusi “Empirical Study of Decision Tree and Artificial Neural Network Algorithm for Mining Educational Database” African Journal of Computing & ICT Reserved - ISSN 2006-1781

[10]. Z. N. Khan “Scholastic achievement of higher secondary students in science stream”, Journal of Social Sciences, Vol. 1, No. 2, pp. 84-87, 2005.

[11]. Hijazi, T. S., & Naqvi, R. S. M. M. (2006). Factors affecting students’ performance: A case of private colleges. Bangladesh e-Journal of Sociology, 3(1).

[12]. Chin Chia Hsu and Tao Huang, The use of Data Mining Technology to evaluate student’s academic achievement via multiple channels of enrolment: An empirical analysis of St. John’s University of Technology. The IABPAD Conference Proceedings Orlando, Florida, January 3-6, 2006.