International Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868
Foundation of Computer Science FCS, New York, USA
Volume 12 – No. 14, July 2018 – www.ijais.org

# Academic Performance Prediction for Success Rate Improvement in Higher Institutions of Learning: An Application of Data Mining Classification Algorithms

Ishaq Oyebisi Oyefolahan
Federal University of Technology, Minna, Nigeria
Information & Media Technology Department

Suleiman Idris
Federal University of Technology, Minna, Nigeria
Computer Science Department

Stella Oluyemi Etuk
Federal University of Technology, Minna, Nigeria
Information & Media Technology Department

Isiaq Oludare Alabi
Federal University of Technology, Minna, Nigeria
Information & Media Technology Department

## ABSTRACT
The abolition of pass grade for any degree course and the consequent change in cumulative grade point for any student to remain within an academic system at University level in Nigeria has led to withdrawal of many students. Thus, it becomes imperative for academic institutions managements to ensure that all necessary steps are taken to enable student graduate successfully. This study explores the usefulness of data mining in unravelling hidden knowledge in students' academic record, particularly the students' specific characteristics which managements or decision makers can leverage upon to ensure improvement in academic success rate of the students. In addition, the study provides a guide through which predicting algorithms can be used by senior academics to predict the performances of students in their respective classes. The conclusion of the study advocates for the use of data mining as decision making tool in academic institutions.

## General Terms
Data Mining, Algorithms

## Keywords
Data Mining, Students' academic performance, Classification models, Higher institution of learning, WEKA

## 1. INTRODUCTION
Data mining is an emerging powerful tool for analysis and prediction. It has been successfully applied in the area of fraud detection, advertising, marketing, loan assessment and prediction [1], but it is in nascent stage in the field of education. Considerable amount of work is being done in this direction, but still there are many untouched areas. Moreover, there is no unified approach among these researches [2].

In the area of education, data mining can be used to efficiently learn from previous data and use the knowledge to predict students' future academic performance. In addition, the knowledge gotten from applying data mining on previous students' data can also be used to advice both new and returning students in a particular institution on steps to take in other to achieve better academic performance.

Educational organizations are one of the important parts of our society and playing vital roles for the growth and development of any nation. Growth in the educational system would surely be enhanced if data mining approach is adopted as a futuristic strategic management tool. The Data Mining tool is able to facilitate better resource utilization in terms of student performance, course development and finally the development of nation's education related standards [3]. In Nigeria, many students have had to withdraw from universities in line with National University Commission (NUC) directives that universities should phase out pass certificates and as a consequence, students who are not able to make up to the required cumulative grade point average (CGPA) are advised to quit the university system. So, with the increasing numbers of students withdrawing after having invested a lot of money as tuition fees, the idea of optimizing student's performance in higher institution using data mining techniques becomes a research interest of this paper.

## 2. LITERATURE REVIEW
### 2.1 Data Mining Techniques
Data Mining is a process of extracting previously unknown, valid, potentially useful and hidden patterns from large data sets [4]. Data mining techniques includes clustering, classification, prediction, regression and outlier detection.

#### *2.1.1 Clustering*
In clustering, a large database is divided into number of subgroups called clusters. The data is divided based on the similarities that existed between the data. Objects in the same clusters are more similar to each other than object in other groups [5]. It is a collection of objects that are similar. It is used to find similarities between data according to their characteristics. Problems that clustering can be used to address includes image processing, pattern recognition, city planning [6].

#### *2.1.2 Classification*
Classification is the most commonly applied data mining technique, which employs a set of pre-classified examples to develop a model that can classify the population of records at large [7]. In a classification problem, data set is partitioned as training and testing dataset. It consists of predicting a certain

outcome based on a given input. Training set is the algorithm which consists of a set of attributes in order to predict the outcome. In order to predict the outcome, it attempts to discover the relationship between attributes and goal or prediction is its outcome. There is another algorithm known as prediction set. It consists of same set of attributes as that of training set. But in prediction set, prediction attribute is yet to be known. In order to process the prediction, it mainly analyses the input. The term which defines how "good" the algorithm is its accuracy [8] (Parvez, Saqib and Syed, 2015). Classification can be applied in the area of fraud detection and credit risk [7].

### *2.1.3 Regression*
Regression is a statistical method which investigates relationships between variables. By using Regression dependences of one variable upon others may be established. Regression is widely used in medical field for predicting the diseases or survivability of a patient [9]. Regression is another data mining technique which is based on supervised learning and is used to predict a continuous and numerical target. It predicts number, sales, profit, square footage, temperature or mortgage rates [10]. In the regression techniques target value are known. For example, you can predict the child behavior based on family history [6].

### *2.1.4 Prediction*
It is one of the data mining techniques that discover the relationship between independent variables and the relationship between dependent and independent variables. It models based on continuous or ordered value [6].

## 2.2 Related Work
Many areas of human endeavor have adopted data mining approach to solve their problems; for example, in banking and finance, medical, marketing, the stock market, telecommunication, manufacturing, health care and customer relationship [1].

In finance, data mining techniques has been used for risk prediction, financial fraud detection, stock forecasting, price prediction [7]. The application of data mining techniques on financial data can contribute to the solution of classification and prediction problems and facilitate decision making processes. Typical examples of financial classification problems are corporate bankruptcy, credit risk estimation, financial distress and corporate performance prediction. The importance of data mining in finance and accounting has been recognized by many organizations [11].

The telecommunications industry was an early adopter of data mining technology and therefore many data mining applications exist. Data mining application in telecommunication industry are divided into three areas: fraud detection, marketing/customer profiling and network fault isolation [12, 13]. The telecommunications industry also implements data mining because they have huge data and have a very large customer, and rapidly changing and highly competitive environment. This industry uses data mining technique to improve their marketing efforts, detection of fraud, and better management of telecommunication networks [6].

Another area that data mining has been used to extract salient information is in the field of healthcare. Data mining have a great potential to enable healthcare systems to use data more efficiently and effectively. Hence, it improves care and reduces costs [8]. Data mining has been applied to predicting or analyzing different diseases like diabetes, heart disease, liver disease, cancer and others [5].

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 [14]. Data mining can be applied to obtain hidden knowledge that has to do with student poor academic performance and how this knowledge can be used to improve students' performance. In this area, different work has been done with different approach and motivation.

In a study by [3], different data classification approaches and comparative analysis were carried out on students' data taken from community college database. In the study they carried out a comparative analysis of all support vector machine kernel types and the Radial Kernel is identified as a best choice for Support Vector Machine. A decision tree approach was proposed which may be taken as an important basis of selection during any course program. The paper aimed to develop a faith on data mining techniques so that present education and business system may adopt this as a strategic management tool.

In their study, [1] focuses on data mining for small student data sets and aimed to find out specific attributes which can be associated with the student success rate. In addition, they also try to find out any relevant student data available to higher education institution on the basis of which they could predict the student success rate. They applied two data mining tools namely Weka and Microsoft Excel to students' data collected from higher institution of learning. The results of the study yielded that student data, available to higher education decision makers (such as professors) via export/import features, carries enough student-specific characteristics in the sense of hidden knowledge which can be successfully associated with student success rates. In addition, the Weka tool, that was used as a comparative analysis to MS Excel, showed that by using decision tree models a high prediction accuracy (especially with the REPTree model) is obtained (the accuracy was verified during the test phase).

Data mining techniques were applied by [15] to predict and analyze students' academic performance based on their academic record and forum participation. They collected students' data from two undergraduate courses. Three different data mining classification algorithms (Naïve Bayes, Neural Network and Decision Tree) were used on the dataset. The prediction performance of three classifiers were measured and compared. They observed that Naïve Bayes classifier outperformed the two classifiers by achieving overall prediction accuracy of 86%. The research will assist teachers to detect early students that are expected to fail the course. The instructor can therefore provide special attention to those students and help them to enhance their academic performance.

[16] presents the classification and prediction-based data mining algorithms to predict slow learners in education sector using Waikato Environment for Knowledge Analysis (WEKA) to classify slow learners. The authors reiterate that the academic performance of student is not always a result of one deciding factor. It heavily hinges on various factors such as personal, socio-economic, psychological and other environmental variables. The paper tried to generate data source of predictive variables, data mining methodologies to

study student performance at high school level, identification of the slow learners' performance, identification of the highly influencing predictive variables on the academic performance of high school students and to find the best classification algorithm. The limitation of the research work is that it only focused on identifying slow learning, it could not emphasize on how slow learners' academic performance can be improved. [17] looks at the analysis of student database using classification techniques as one area in which data mining can be applied. The authors explained that data analysis can be categorized into two; extracting models describing important classes and predicting future trends. The research works aims at predicting students' job position based on their academic performance. Decision Tree Algorithm also known as J48 from Waikato Environment for Knowledge Analysis (WEKA) is one of the methodologies used for the analysis of data for knowledge discovery. The limitation of the work is evident in the fact that the work did not state that job positions were advertised with respect to students' academic performance and so there was no basis for predicting the placement of students' performance to job position, and the number of students shown in the data analysis are relatively few.

## 3. METHODOLOGY

In other to build the student data model, five stages were followed which include data or attribute identification, data collection, preprocessing, classification and interpretation. Figure 1 shows the stages followed in this paper to predict the students' academic performances

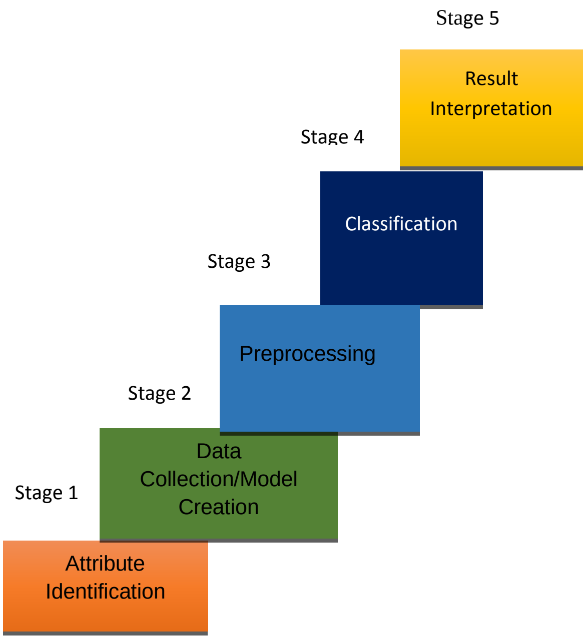


**Fig. 1 Data Mining Steps**

Data acquisition involves gathering all information that is assumed to affect students' performance. To realize this stage, we choose a course in the department of information and media technology – web applications development for the 2015/2016 and 2016/2017 academic sessions with total numbers of students 143. The student records were gotten from the lecturer in charge of the course. Attributes like matriculation number, names, CA point, Exam point, total point and final grade were acquired from the course lecturer. To acquire other attributes needed for this research, we printed research questionnaire which were distributed to the students. The research questions contain attribute like student name, matric number, state of origin, where the student stays (on or off campus), sport, gender, year of birth, registration (first or second). All these attributes were combined to make the datasets as shown in figure 2.

The next stage is the preprocessing stage where data cleaning (removal of incomplete and inconsistent data) was performed. Attribute that has no impact on the prediction were also removed after calculating the information gain (entropy) for each of the attributes. The entropy was calculated for each of the attribute using the formula below

$$Entropy = (P1, P2, \ldots Pn)$$
$$= -P1logp1 - P2logP2 \ldots - PnlogPn$$

Also, data like the class value were changed from numeric to nominal value (high, medium, low) to allow algorithms like the J48 and REPTree work effectively. Preprocessing is an important stage of the data mining and was carried out carefully since the data model gotten from the stage will be acted upon directly by the classification algorithms (J48, Reptree and M5P). Figure 3 shows the attribute gotten after the preprocessing stage.

After preprocessing, the next stage of the model is classification. Five decision tree techniques were used to classify the data in this paper namely J48, REP Tree, Random Tree, Naïve Bayes and Decision Stump. These techniques were chosen based on their good performances as found in earlier studies. The 2015/16 dataset was divided into two, the first part was used for training and the second part was used for testing. The 2016/17 dataset was used for validation.

International Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868
Foundation of Computer Science FCS, New York, USA
Volume 12 – No. 14, July 2018 – www.ijais.org

Relation: Student

| No. | 1: Student year Numeric | 2: Student Numeric | 3: Gender Nominal | 4: YOB Numeric | 5: Region Nominal | 6: PL Nominal | 7: AM Nominal | 8: LofIPC Nominal | 9: Sport Nominal | 10: Registration Nominal | 11: Campus Nominal | 12: CA Numeric | 13: Exam Point Numeric | 14: Final Point Numeric | 15: Type Nominal |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 201516.0 | 1.0 | Male | 199... | North | Yes | Yes | High | Yes | First | Yes | 34.0 | 40.0 | 74.0 | High |
| 2 | 201516.0 | 2.0 | Female | 199... | North | Yes | Yes | Average | No | First | No | 24.0 | 34.0 | 58.0 | Low |
| 3 | 201516.0 | 3.0 | Female | 199... | North | Yes | No | High | No | First | No | 28.0 | 37.0 | 65.0 | Mediu... |
| 4 | 201516.0 | 4.0 | Male | 199... | North | Yes | No | Average | Yes | First | No | 24.0 | 19.0 | 43.0 | Low |
| 5 | 201516.0 | 6.0 | Male | 198... | North | Yes | Yes | High | Yes | First | No | 18.0 | 13.0 | 31.0 | Low |
| 6 | 201516.0 | 7.0 | Female | 199... | North | Yes | Yes | Average | Yes | First | Yes | 20.0 | 27.0 | 47.0 | Low |
| 7 | 201516.0 | 8.0 | Female | 199... | North | Yes | Yes | Average | No | First | No | 22.0 | 23.0 | 45.0 | Low |
| 8 | 201516.0 | 9.0 | Male | 199... | North | Yes | Yes | Average | No | First | No | 24.0 | 17.0 | 41.0 | Low |
| 9 | 201516.0 | 10.0 | Male | 199... | North | Yes | No | High | No | First | No | 29.0 | 27.0 | 56.0 | Low |
| 10 | 201516.0 | 12.0 | Male | 199... | North | Yes | Yes | Average | Yes | First | No | 28.0 | 36.0 | 64.0 | Mediu... |
| 11 | 201516.0 | 13.0 | Male | 199... | South | Yes | Yes | Average | No | First | No | 29.0 | 39.0 | 68.0 | Mediu... |
| 12 | 201516.0 | 14.0 | Male | 199... | North | Yes | Yes | Average | Yes | First | No | 24.0 | 24.0 | 48.0 | Low |
| 13 | 201516.0 | 15.0 | Male | 199... | North | Yes | No | Average | No | First | No | 23.0 | 18.0 | 41.0 | Low |
| 14 | 201516.0 | 16.0 | Male | 199... | North | Yes | Yes | Average | Yes | First | Yes | 20.0 | 27.0 | 47.0 | Low |
| 15 | 201516.0 | 17.0 | Male | 199... | North | Yes | Yes | High | Yes | First | No | 22.0 | 23.0 | 45.0 | Low |
| 16 | 201516.0 | 18.0 | Male | 199... | North | Yes | No | Average | Yes | First | No | 16.0 | 17.0 | 33.0 | Low |
| 17 | 201516.0 | 19.0 | Female | 199... | South | Yes | No | High | No | First | Yes | 33.0 | 27.0 | 60.0 | Mediu... |
| 18 | 201516.0 | 20.0 | Female | 199... | South | Yes | Yes | Average | Yes | First | No | 24.0 | 31.0 | 55.0 | Low |
| 19 | 201516.0 | 21.0 | Male | 199... | South | Yes | No | Average | No | First | No | 24.0 | 33.0 | 57.0 | Low |
| 20 | 201516.0 | 22.0 | Male | 199... | North | No | No | Average | Yes | First | No | 22.0 | 26.0 | 48.0 | Low |
| 21 | 201516.0 | 24.0 | Male | 199... | North | Yes | Yes | High | No | First | No | 24.0 | 29.0 | 53.0 | Low |
| 22 | 201516.0 | 25.0 | Female | 199... | South | Yes | No | Low | No | First | No | 19.0 | 24.0 | 43.0 | Low |
| 23 | 201516.0 | 27.0 | Female | 199... | North | Yes | No | Low | No | First | No | 17.0 | 15.0 | 32.0 | Low |
| 24 | 201516.0 | 29.0 | Male | 199... | North | Yes | Yes | High | Yes | First | No | 19.0 | 11.0 | 30.0 | Low |
| 25 | 201516.0 | 30.0 | Male | 199... | North | Yes | No | Low | No | First | No | 25.0 | 16.0 | 41.0 | Low |
| 26 | 201516.0 | 31.0 | Male | 199... | South | No | Yes | High | No | First | No | 19.0 | 16.0 | 35.0 | Low |
| 27 | 201516.0 | 32.0 | Male | 199... | North | Yes | Yes | Average | No | First | No | 24.0 | 23.0 | 47.0 | Low |
| 28 | 201516.0 | 33.0 | Male | 199... | North | Yes | Yes | High | No | First | No | 31.0 | 30.0 | 61.0 | Mediu... |
| 29 | 201516.0 | 34.0 | Female | 199... | North | Yes | Yes | Average | No | First | No | 25.0 | 16.0 | 41.0 | Low |
| 30 | 201516.0 | 36.0 | Male | 199... | South | Yes | Yes | Average | Yes | First | No | 22.0 | 25.0 | 47.0 | Low |
| 31 | 201516.0 | 37.0 | Female | 199... | North | Yes | Yes | High | No | First | Yes | 27.0 | 20.0 | 47.0 | Low |
| 32 | 201516.0 | 38.0 | Male | 199... | North | Yes | No | Average | No | First | No | 20.0 | 12.0 | 32.0 | Low |
| 33 | 201516.0 | 39.0 | Male | 198... | North | Yes | No | High | No | First | No | 25.0 | 37.0 | 62.0 | Mediu... |
| 34 | 201516.0 | 40.0 | Male | 199... | North | Yes | Yes | High | Yes | First | No | 21.0 | 22.0 | 43.0 | Low |
| 35 | 201516.0 | 42.0 | Male | 199... | North | Yes | No | Average | No | First | No | 25.0 | 23.0 | 48.0 | Low |
| 36 | 201516.0 | 43.0 | Female | 199... | North | Yes | Yes | High | No | First | No | 27.0 | 28.0 | 55.0 | Low |
| 37 | 201516.0 | 44.0 | Female | 199... | North | No | Yes | Average | No | First | No | 23.0 | 21.0 | 44.0 | Low |

**Fig 2: Attribute with entropy greater than zero (gender, region, parent literacy, academic mentor, year of birth, Academic mentor, Level of interest in present course, sport, campus, CA, Exam point, Final point)**

| No. | 1: Gender Nominal | 2: Region Nominal | 3: PL Nominal | 4: AM Nominal | 5: LofIPC Nominal | 6: Sport Nominal | 7: Campus Nominal | 8: CA Numeric | 9: Exam Point Numeric | 10: Final Point Numeric | 11: Type Nominal |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Male | North | Yes | Yes | High | Yes | Yes | 34.0 | 40.0 | 74.0 | High |
| 2 | Female | North | Yes | Yes | Average | No | No | 24.0 | 34.0 | 58.0 | Low |
| 3 | Female | North | Yes | No | High | No | No | 28.0 | 37.0 | 65.0 | Mediu... |
| 4 | Male | North | Yes | No | Average | Yes | No | 24.0 | 19.0 | 43.0 | Low |
| 5 | Male | North | Yes | Yes | High | Yes | No | 18.0 | 13.0 | 31.0 | Low |
| 6 | Female | North | Yes | Yes | Average | Yes | Yes | 20.0 | 27.0 | 47.0 | Low |
| 7 | Female | North | Yes | Yes | Average | No | No | 22.0 | 23.0 | 45.0 | Low |
| 8 | Male | North | Yes | Yes | Average | No | No | 24.0 | 17.0 | 41.0 | Low |
| 9 | Male | North | Yes | No | High | No | No | 29.0 | 27.0 | 56.0 | Low |
| 10 | Male | North | Yes | Yes | Average | Yes | No | 28.0 | 36.0 | 64.0 | Mediu... |
| 11 | Male | South | Yes | Yes | Average | No | No | 29.0 | 39.0 | 68.0 | Mediu... |
| 12 | Male | North | Yes | Yes | Average | Yes | No | 24.0 | 24.0 | 48.0 | Low |
| 13 | Male | North | Yes | No | Average | No | No | 23.0 | 18.0 | 41.0 | Low |
| 14 | Male | North | Yes | Yes | Average | Yes | Yes | 20.0 | 27.0 | 47.0 | Low |
| 15 | Male | North | Yes | Yes | High | Yes | No | 22.0 | 23.0 | 45.0 | Low |
| 16 | Male | North | Yes | No | Average | Yes | No | 16.0 | 17.0 | 33.0 | Low |
| 17 | Female | South | Yes | No | High | No | Yes | 33.0 | 27.0 | 60.0 | Mediu... |
| 18 | Female | South | Yes | Yes | Average | Yes | No | 24.0 | 31.0 | 55.0 | Low |
| 19 | Male | South | Yes | No | Average | No | No | 24.0 | 33.0 | 57.0 | Low |
| 20 | Male | North | No | No | Average | Yes | No | 22.0 | 26.0 | 48.0 | Low |
| 21 | Male | North | Yes | Yes | High | No | No | 24.0 | 29.0 | 53.0 | Low |
| 22 | Female | South | Yes | No | Low | No | No | 19.0 | 24.0 | 43.0 | Low |
| 23 | Female | North | Yes | No | Low | No | No | 17.0 | 15.0 | 32.0 | Low |
| 24 | Male | North | Yes | Yes | High | Yes | No | 19.0 | 11.0 | 30.0 | Low |
| 25 | Male | North | Yes | No | Low | No | No | 25.0 | 16.0 | 41.0 | Low |
| 26 | Male | South | No | Yes | High | No | No | 19.0 | 16.0 | 35.0 | Low |
| 27 | Male | North | Yes | Yes | Average | No | No | 24.0 | 23.0 | 47.0 | Low |
| 28 | Male | North | Yes | Yes | High | No | No | 31.0 | 30.0 | 61.0 | Mediu... |
| 29 | Female | North | Yes | Yes | Average | No | No | 25.0 | 16.0 | 41.0 | Low |

**Fig 3: Attribute with entropy greater than zero (gender, region, parent literacy, academic mentor, year of birth, Academic mentor, Level of interest in present course, sport, campus, CA, Exam point, Final point)**

## 3.1 Performance Evaluation

The performance of the classifiers used were evaluated using the following metrics

**Precision:** proportion of correct positive observation

$$\frac{TP}{TP + FP}$$

**Accuracy:** Proportion of total number of correct prediction

$$\frac{TP + TN}{P + N}$$

**Recall:** Proportion of positives correctly predicted as positive

$$\frac{TP}{P}$$

**F-Measure:** This is derived from precision and recall values. The F-Measure produces a high result when Precision and Recall are both balanced, thus this is very significant.

$$\frac{2 * Recall * Precision}{Recall + Precision}$$

# 4. RESULTS

The study which aims at exploring attributes that contribute to the prediction of students' academic performances achieved it aim by identifying 11 attributes out of 15 as predicting factors. In addition, as there are different classification algorithms that have been put to use in literature, the study uses the data gathered to evaluate the performances of these different classification algorithms (J48, REPTree, RandomTree, NaiveBayes, and DecisionTree) with a view of ascertaining the goodness of the classification outputs of the algorithms.

According to [18], classification is one of commonly used data mining method; which utilizes pre-classified data to develop a predicting model that can be used to classify records of dataset. In building the predictive model; the arff format of the selected dataset was loaded into WEKA, and for each selected algorithm, two experimental procedures were performed. After applying the various data mining algorithms on the students' dataset, the summary of results obtained are presented in Table 1.

## 4.1 J48 Model

Using the J48 algorithm, training phase gave 93.3333% where 56 instances were correctly classified and 6.6667% 2 instances were incorrectly classified. The testing phase showed 100% where 51 instances were correctly classified and 0% instances incorrectly classified. The decision tree for the testing phase is presented in Fig. 4.

## 4.2 REPTree Model

Training set gave 90% where 54 instances were correctly classified and 10% where 6 instances were classified incorrectly. The testing data showed 94.1176% where 48 instances were correctly classified and 5.8824% where 3 instances were incorrectly classified. (see Fig. 5)

## 4.3 Random Tree Model

The RandomTree model gave a classification accuracy of 86.6667% where 52 instances were correctly classified and 13.3333% where 8 instances were incorrectly classified. The testing data gave 93.1712% accuracy where 43 instances were correctly classified and 15.6863% where 8 instances were incorrectly classified. (see Fig. 6)

## 4.4 Naïve Bayes Model

The Naïve Bayes model gave 90% accuracy where 54 instances where correctly classified and 10% where 6 instances were incorrectly classified on the training data set. The testing data showed 98.0392% accuracy where 50 instances were correctly classified and 1.9608% where one instance was incorrectly classified

## 4.5 Decision Stump Model

This model gave 91.6667% accuracy on training data where 55 instances were correctly classified and 8.3333% where 5 instances were incorrectly classified. The testing data showed 94.1176% where 48 instances were correctly classified and 5.8824 were 3 instances were incorrectly classified.

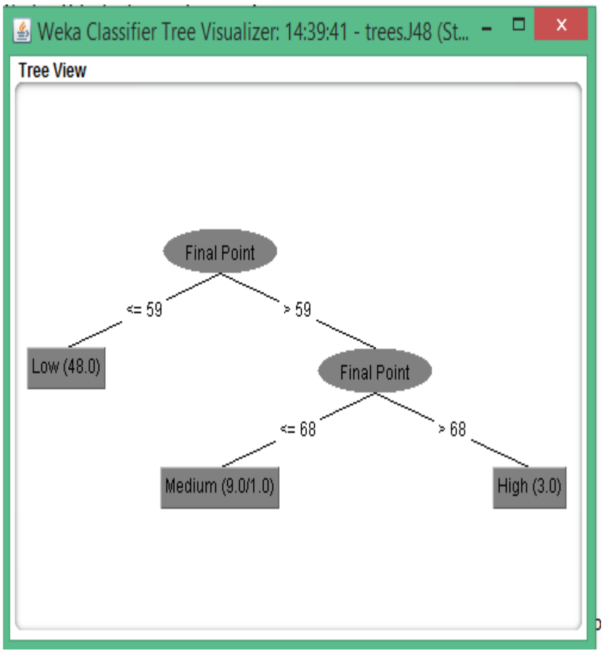


**Fig 4: Decision Tree for the testing dataset using J48**

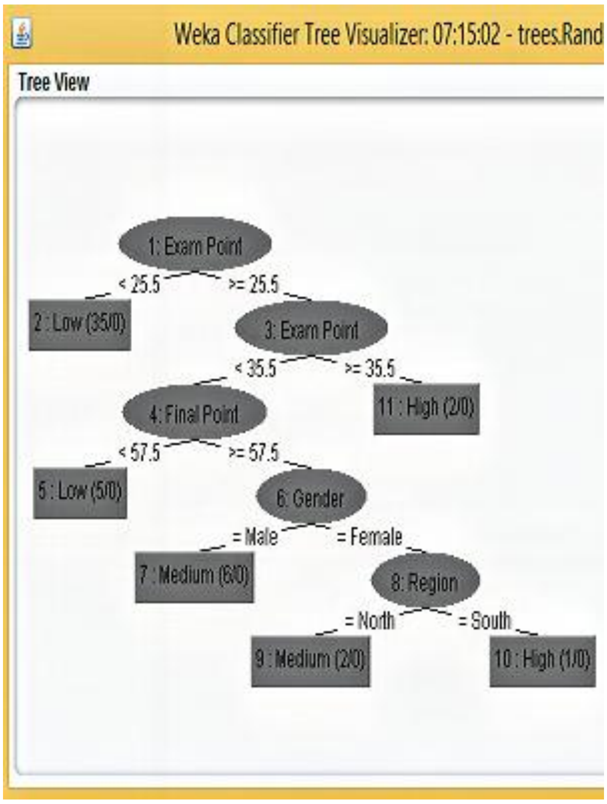


**Figure 5: Decision Tree for Testing data set using Random Tree**

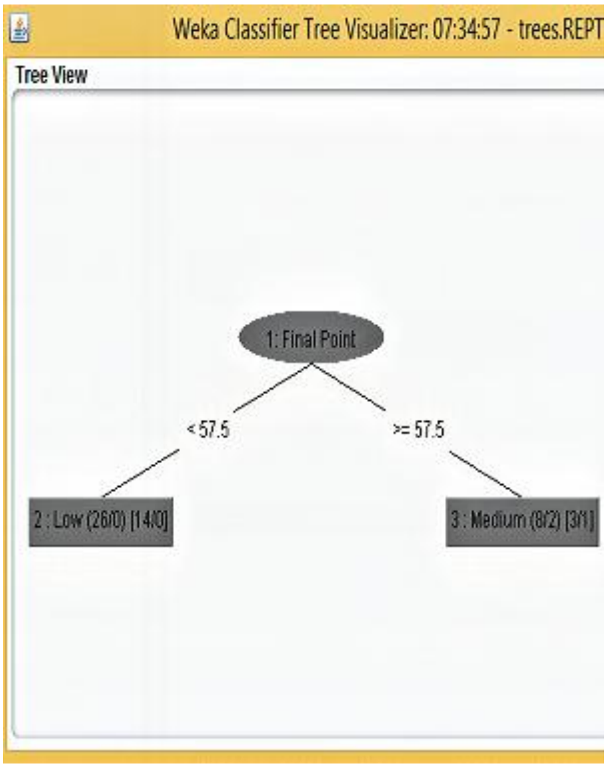


**Figure 6: Decision Tree for Testing data set using RandomTree**

**Table 1. Result Outputs of Classifier Algorithms**

| Experiment | Algorithm | Precision% | | Recall% | | F-measure% | | Accuracy% | |
|---|---|---|---|---|---|---|---|---|---|
| | | 10 fold CV | 20% Test Set | 10 fold CV | 20% Test Set | 10 fold CV | 20% Test Set | 10 fold CV | 20% Test Set |
| 1 | J48 | 94.3 | 88.1 | 93.3 | 89.6 | 93.7 | 87.9 | 93.333 | 89.583 |
| 2 | REPTree | 87.7 | 66.0 | 90.0 | 81.3 | 88.5 | 72.8 | 90.000 | 81.250 |
| 3 | RandomTree | 83.4 | 88.8 | 86.7 | 91.7 | 84.9 | 89.6 | 86.667 | 91.667 |
| 4 | NaïveBayes | 87.7 | 79.8 | 90.0 | 85.4 | 88.5 | 82.5 | 90.000 | 85.417 |
| 5 | DecisionStump | 88.5 | 88.9 | 91.7 | 91.7 | 89.8 | 89.6 | 91.667 | 91.667 |

## 5. CONCLUSION

In this study, effort has been made to predict students academic performances based on attributes associated with the students. The use of WEKA which is a free desktop tool for data mining has shown that data mining today can be carried out without the challenge of big investment in analytical tools. In addition, the usual notion that only very large data can be used to unravel hidden knowledge is dymistified as the tool has proven that knowledge for decision making can also be deduced from not so large data. The identification of attributes with entropy higher than zero (contributing attributes) provides further insight into key influencers of students academic outputs. Five classsification techniques were used to predict the students academic performances based on the data utilized .

The analyses above show that J48 (DecisionTree) with accuracy of 93.333% using 11 attributes and 10 fold cross validation is more appropriate in building the predicting model for the students academic perfomances based on the dataset. Compared to four other algorithm used in the study, J48 shows better prediction accuracy and followed by DecisionStump which have prediction accuracy of 91.667% for both 10 fold cross validation and 20% supplied test set. A closer look shows that though J48 has the highest prediction accuracy, DecisionStump has higher prediction accuracy over J48 when 20% supplied test set is used. It therefore means that

DecisionStump provide better generalization output compared to other classification tools. A likelihood reason for the performance of J48 on the test set may be the higher sensitivity of J48 to missing values.

Generally, different studies have found mixed outputs in terms of classification algorithm performances. For example REPTree performs better that J48 and M5P Decision Tree

[1] while J48 has better acuracy compared to Naïve Bayes and Random Forest [19].

In conclusion, the research achieved its objectives by identify the influencers of students academic performance and exploring the level of predicting accuracy of different algorithms in the WEKA environment. The study would assist decision maker in the academic industry particularly in Nigeria on important factors during planning and evaluation stages of academic delivery. In addition, it will enable lecturers the basic understanding on how to counsel students for good academic performance.

Future research can look into more attributes such semestarial work load of students, the prerequisite knowledge prior to taking some course and the use of large dataset for analysis.

## 6. REFERENCES

[1] Natek, S. and Zwilling, M. (2014) Student data mining solution-knowledge management system related to higher education institutions. *Expert Systems with Applications* 41, 6400 – 6407

[2] Thakar, P. Mehta, A. and Manisha, (2015). Performance Analysis and Prediction in Educational Data Mining: A Research Travelogue. *International Journal of Computer Applications* 110(15), 60 – 68

[3] Agarwal, S. Pandey, G. N., and Tiwari, M. D. (2012). Data Mining in Education, Data Classification and Decision Tree Approach. *International Journal of e-Education, e-Business, e-Management and e-Learning* 2(2), 140-144

[4] Tarekegn, G. B. and Sreenivasarao V. (2016). Application of Data Mining Technique on Pre ART Data: The Case Study of Felege Hiwot Referral Hospital. International Journal of Research Studies in Computer Science and Engineering. 3(2), 1- 9

[5] Patel, S. and Patel, H. (2016). Survey of Data Mining Technique used in Health Care Domain. *International Journal of Information Science and Techniques.* 6(1/2), 53 – 60

[6] Smita and Priti, M. (2014). Use of Data Mining in Various Field: A Survey Paper. *Journal of Computer Engineering.* 16(3). 18 – 21.

[7] Nagarathna, M., and Moorthi, M. (2015). A Review of Data Mining and its Techniques in Finance Sector. *International Journal of Advance Research in Computer Science and Management Studies. 3(3).*

[8] Ahmed, P., Qamar, S. and Rizvi, S. Q. A. (2015). Techniques of Data Mining in HealthCare: A Review. *International Journal of Computer Application.* 120(15). 38 – 50.

[9] Tomar, D. and Agarwal, S. (2013). A Survey on Data Mining Approaches for HealthCare. *International journal of Bio-Science and Bio-Technology.* Volume 5 issue (5).

[10] Gera, M. and Goel S. (2015). Data Mining – Techniques, Methods and Algorithms: A Review on Tools and their validity. *International Journal of Computer Applications.* 113(18), 22 – 29

[11] Soumya, S. B. and Deepika, N. (2016). Data Mining with Predictive Analytics for Financial Applications. *International Journal of Scientific Engineering and Applied Science.* 2(1). 310 – 317. Available at: www.ijseas.com

[12] Weiss, G M. . (n.a). Data Mining in Telecommunication. Department of Computer and Information Science. Fordham University.

[13] Umayaparvath, V and Iyakutti, K., (2012). Application of Data Mining Technique in Telecom Churn Prediction. *International Journal of Computer Applications*. 42(20) 5-9.

[14] Kumar, V and Chadha A. (2011). An Emperical Study of the Application of Data Mining Technique in Higher Education. *International Journal of Advanced Computer Science and Applications.* 2(3). 80 – 84.

[15] Mueen, A. Zafar, B. and Manzoor, U. (2016). Modelling and Predicting Students Academic Performance Using Data Mining Techniques. *International Journal of Modern Education and Computer Science.* 11. 36 – 42.

[16] Kaur, P., M. Singh, and G.S. Josan, (2015). Classification and Prediction Based Data Mining Algorithms to Predict Slow Learners in Education Sector. *Procedia Computer Science*. 57: p. 500-508.

[17] Sumathi, K., Kannan, S., and Nagarajan, K., (2016). Data Mining: Analysis of Student Database using Classification Techniques. International Journal of Computer Applications. 141(8). 22 – 27.

[18] Singh, S. and Kumar, V. (2012). Classification of Student's data Using Data Mining Techniques for Training & Placement Department in Technical Education. International Journal of Computer Science and Network (IJCSN) Volume 1, Issue 4. Available at: http://ijcsn.org/IJCSN-2012/1-4/IJCSN-2012-1-4-63.pdf.

[19] Tarekegun, G. B. and Sreenivasarao, V. (2016) Application of Data Mining to Predict Students Placement in to Departments. *International Journal of Research Studies in Computer Science and Engineering.* 3(2) 10-14.