-11-

DEVELOPMENT OF STUDENTS' ACADEMIC PERFORMANCE PREDICTION MODEL USING DECISION TREE

[1]Lawal O.O; [2]Adepoju A.O; [3]Makinde, O.E. (PhD) and [4]Osunade. O.(PhD)
[1,2] Department of Computer Science, D.S Adegbenro I.C.T Polytechnic, Itori, Ogun State
[3]Department of Computer Science, Ajayi Crowther University, Oyo, Oyo State
[4]Department of Computer Science, University of Ibadan, Ibadan, Nigeria.
Corresponding Author: opepolawal@gmail.com, 08062160129

ABSTRACT

Students' academic performance prediction model is a model developed to reduce the students' failure rate in higher institutions of learning. The model is developed to ascertain the students' academic performance by looking at the extent of socio-economic background on the student's performance and the effect of the student's previous academic status (GPA) on his/her future performance. This paper proposed the J48 Decision tree model for predicting the academic performance of third-year Bachelor of Science (B. Sc) students in Computer Science at Ajayi Crowther University, Oyo as they proceed to the final year. The data collected from the department contain the students' Name, Matriculation number, Gender, State of Origin, Current G.P.A. The parameter used as the Key Performance Indicator (KPI) is the Current G.P. The model is created using a Training data set, and the performance is confirmed by the Supplied test set and applied to the student's data to generate the Prediction model. The extracted knowledge from the prediction model is used to identify and profile the student to determine the student's level of success in the final year. The outcome of the work showed that the students have a high tendency to have the same GPA in the final year.

Keywords: Data mining, Educational Data Mining, Key Performance Indicator, Prediction, Decision tree, WEKA, Classification.

1. INTRODUCTION

Education is an important factor that contributes immensely to the growth of society at large. Many educational organizations and school administrations today, leave no stone unturned to improve their student's academic performance, in which the marks obtained by the student in the examination decide his/her future. They want to increase the number of students getting passed yearly to develop the best quality of the education process in their institution, to maintain the brand name of the organization, and to educate students in a better way.

To achieve this target, Data Mining (DM) is considered one of the most suitable techniques for giving additional insights to the Institutions of Higher Learning community to help them make better decisions in educational activities. Data mining is the process used to transform data into useful information by extracting and making use of relatively unknown patterns, trends, and datasets from large amounts of stored raw data in the repository. Though many factors determine a good institution, the academic performance of the student plays a vital role in it (Jai & David, 2014).

Educational Data Mining (EDM) is an emerging trend in data mining, which is concerned with developing techniques for exploring and mining useful educational patterns/datasets from the database of an educational institution and using the extracted data to make decisions and predictions for the enhancement of the educational system as it relates to students' performance and systemic improvement. Researchers in this field focus on discovering useful knowledge either to help educational institutes manage their students better or to help students to manage their education and deliverables better and enhance their performance (Amjad, 2016).

2. LITERATURE REVIEW

Data Mining

According to Mehmed (2011), *Data mining is a process of discovering various models, summaries, and derived values from a given collection of data.* It has been widely used in recent years due to the availability of huge amounts of data in electronic form, and there is a need for turning such data into useful information and knowledge for large applications. These applications are found in fields such as Artificial Intelligence, Machine Learning, Market Analysis, Statistics and Database Systems, Business Management, and Decision Support Systems.

In practice, the two primary goals of data mining tend to be *prediction*and *description*(Mehmed, 2011). *Prediction*involves using some variables or fields in the data set to predict unknown or future values of other variables of interest. *Description*, on the other hand, focuses on finding patterns describing the data that can be interpreted by humans.

Therefore, it is possible to put data-mining activities into one of two categories:

1. Predictive data mining, which *produces the model*of the system described by the given data set, or
2. Descriptive data mining, which *produces new, nontrivial information*based on the available data set.

On the predictive end, the goal of data mining is to produce a model, expressed as an executable code, which can be used to perform classification, prediction, estimation, or other similar tasks. On the descriptive end of the spectrum, the goal is to gain an understanding of the analyzed system by uncovering patterns and relationships in large data sets.

Data mining techniques

Data mining techniques are used to build a model to identify new knowledge information (Osmanbegovic and Mirza, 2012). Several major data mining techniques have been developed and used including association, classification, clustering, prediction and sequential patterns.

Mehmed (2011) postulated that the goals of prediction and description are achieved by using data-mining techniques for the following primary data-mining tasks:

1. *Classification* -Discovery of a predictive learning function that classifies a data item into one of several predefined classes.
2. *Regression* -Discovery of a predictive learning function that maps a data item to a real - value prediction variable.
3. *Clustering* -A common descriptive task in which one seeks to identify a finite set of categories or clusters to describe the data.
4. *Summarization* -An additional descriptive task that involves methods for finding a compact description for a set (or subset) of data.
5. *Dependency Modelling* -Finding a local model that describes significant dependencies between variables or between the values of a feature in a data set or in a part of a data set.
6. *Change and Deviation Detection* -Discovering the most significant changes in the data set.

Data mining process

The general experimental procedure adapted to data–mining problems involves the following steps:

1. State the problem and formulate the hypothesis
2. Collect the data
3. Pre-process the data
4. Estimate the model
5. Interpret the model and draw conclusions

Educational Data Mining

According to Amjad (2016), Educational Data Mining (EDM) is a new trend in the data mining and Knowledge Discovery in Databases (KDD) field which focuses on mining useful patterns and discovering useful knowledge from the educational information systems, such as admissions systems, registration systems, course management systems (module, blackboard, etc…), and any

other systems dealing with students at different levels of education, from schools to colleges and universities.
Researchers in this field focus on discovering useful knowledge either to help educational institutes manage their students better or to help students to manage their education and deliverables better and enhance their performance. It is an emerging discipline, concerned with developing methods for exploring the unique types of data that come from educational settings and using those methods to better understand students and the settings in which they learn (Kumar and Radhika, 2014).
In this paper, the Traditional Class Room Educational System is considered. The performance of the students depends on many factors such as their family background. Also important is the academic background i.e. Secondary School Certificate Examination (SSCE), all previous semesters' results and Cumulative Grade Point Average (CGPA) has to be considered. Those students who are at risk can be found out and some remedial action can be taken to improve their performance. Due to these results, the institution can improve the chance(s) of the students who will pass. Educational data mining can be used to get feedback for the teachers so the teacher can improve the teaching method. Also, the information can be useful for those who are designing the course content.

Key Performance Indicator

Osmanbegovic and Mirza (2012) presented a model to predict student performance. They evaluate student success by passing grades on the exam. Parameters addressed for prediction include students' socio-demographic variables, achieved results from high school, the entrance exam, and attitudes toward studying which can affect success.
Ramesh et al. (2013) presented a valuable study to figure out factors that influenced student success. They focused on parents' occupations and school types. Their obtained results from hypothesis testing reveal that type of school does not influence student performance and parents' occupation plays a major role in predicting grades. The ability to predict students' marks could be useful in a great number of different ways associated with university-level learning.
Even though there are many prediction models available with different approaches to student performance, Sajadin (2011) posited that there is no certainty if there are any predictors that accurately determine whether a student will be an academic genius, a dropout, or an average performer.
Graded Point Average (GPA) is a commonly used indicator of academic performance. Many universities set a minimum GPA that should be maintained. Therefore, GPA remains the most common factor used by academic planners to evaluate progression in an academic environment. Many factors could act as barriers to students attaining and maintaining a high GPA that reflects their overall academic performance, during their tenure in university. These factors could be targeted by the department in developing strategies to improve student learning and improve their academic performance by way of monitoring the progression of their performance (Oyelade et al., 2010).
In this work, the key Performance Indicator used is the Graded Point Average (GPA) of the student, bearing in mind that the student's family background and SSCE result cannot be fully relied upon as key performance indicators to predict his/her future performance.

The Waikato Environment for Knowledge Analysis (WEKA)

According to Frank et al. (2005), WEKA is a data mining system developed by the University of Waikato in New Zealand that implements data mining algorithms. WEKA is a state-of-the-art facility for developing machine learning (ML) techniques and their application to real-world data mining problems. It is a collection of machine learning algorithms for data mining tasks. The algorithms are applied directly to a dataset. WEKA implements algorithms for data pre-processing, classification, regression, clustering, and association rules; it also includes visualization tools.
WEKA would not only afford a toolbox of learning algorithms but also a framework inside which researchers could implement new algorithms without having to be concerned with supporting

infrastructure for data manipulation and scheme evaluation. WEKA is open-source software issued under General Public License (Pallavi, 2015).

The data file normally used by WEKA is in Attribute Relation File Format (.arff), which consists of special tags to indicate different things in the data file foremost: attribute names, attribute types, attribute values, and the data. For working with WEKA we do not need deep knowledge of data mining which is the reason why it is a very popular data mining tool. It also provides the Graphical User Interface of the user and provides many facilities. The GUI Chooser consists of four buttons—one for each of the four major WEKA applications. The buttons can be used to start the following applications:

Explorer: It is the main interface in WEKA. It has a set of panels, each of which can be used to perform a certain task. Once a dataset has been loaded, one of the other panels in the Explorer can be used to perform further analysis.

Experimenter: An environment for performing experiments and conducting statistical tests between learning schemes.

Knowledge Flow: This environment supports essentially the same functions as the Explorer but with a drag – and drop interface. One advantage is that it supports incremental learning.

Simple CLI: Provides a simple Command-Line Interface that allows direct execution of WEKA commands for operating systems that do not provide their command-line interface.

A trial version of the WEKA package can be downloaded from the University of Waikato website at http://www.cs.waikato.ac.nz/~ml/weka/index.html.



Figure 1: WEKA tool front view (Source: Purva and Kamal, 2015)

Execution in WEKA

Execution in WEKA is a step-by-step process (Purva and Kamal, 2015). First is data loading, Data can be loaded from various sources, including files, URLs, and databases. WEKA can read in ".csv" format (Comma Separated Value). Firstly we take Excel datasheet from the real world, the first row contains the attribute names (separated by commas) followed by each data row with attribute values listed in the same order (also separated by commas), and convert it into .csv file format. Then go to the explore button in WEKA and save this .csv file. Once data is loaded into WEKA, the data set is automatically saved into ARFF format (Purva and Kamal, 2015).

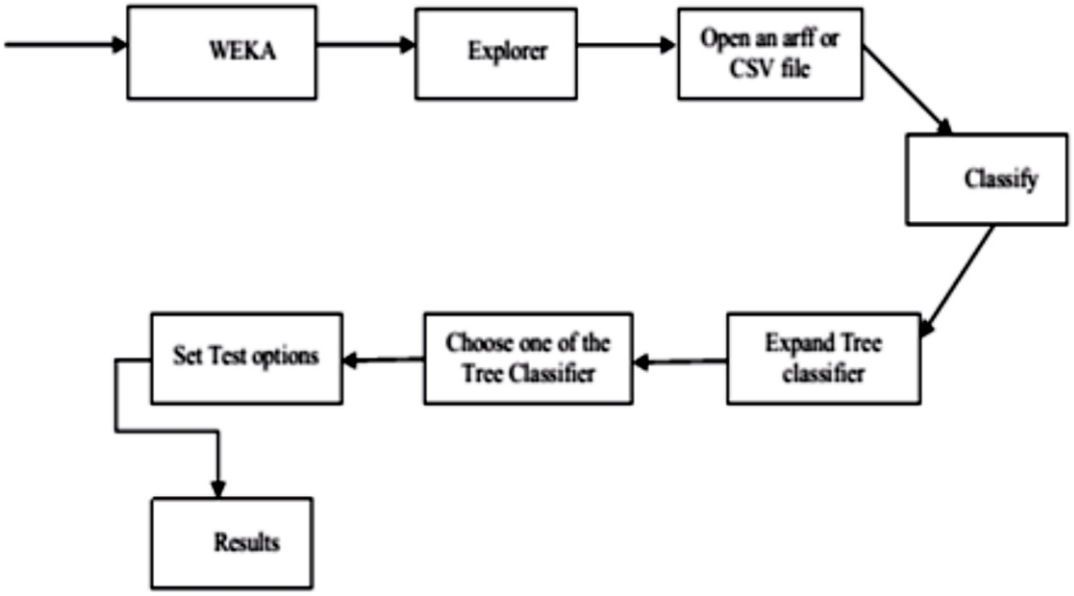


Figure 2: Execution in WEKA tool (Source: Purva and Kamal, 2015)

Decision Tree Classification in WEKA

Classification is a process of finding a model that describes and distinguishes data classes and concepts to predict the class of objects whose class label is unknown. The derived model should be represented in a decision tree or neural network.

The Classification process involves the following steps:

1. Create a training dataset.
2. Identify class attributes and classes.
3. Identify useful attributes for classification (Relevance analysis).
4. Learn a model using training examples in the Training set.
5. Use the model to classify the unknown data samples.

Classification is possibly the most frequently used data mining technique. It is the process of finding a set of models that describe and differentiate data classes and concepts, to be able to use the model to predict the class whose label is unknown. Many algorithms can be used for classification, such as decision trees, neural networks, logistic regression, etc., and various decision tree algorithms are used in classification like ID3, AD Tree, REP, J48, FT Tree, LAD Tree, decision stamp, LMT, random forest, random tree, etc. In this work we are using J48 decision tree algorithm for classification.

1. METHODOLOGY

In this study, the students' Grade Point Average (GPA) is selected as a dependent parameter. The dataset collected of forty-nine (49) instances (students) was used to develop a training model whose precision level and other parameters were considered to ascertain the model's accuracy. The model's status was further confirmed by the User-supplied Test set which was supplied to the system, and the statistic of the result was compared to the Training model to confirm the accuracy of the model, this was done by supplying the model result of the first nine (9) students and comparing the result generated.

To generate the prediction model, a class of unknown results (Predicted GPA) was created and supplied to the model. The prediction level of the result and other performance metrics signifies the accuracy of the prediction for each student as revealed.

Algorithm of the model

The training dataset (I1…In) which is the total number of instances was loaded to WEKA. The GPA (Key Performance Indicator) for each student (GPA1…GPAn) from the dataset with other

information about each student is used to build the model, and data of unknown GPA (nG1…nGn) for each student was fed into the model for prediction. The result generated displays the Previous GPA and the Predicted GPA.

Step 1 - Start
Step 2 - Take input which is given by User
In= {I1,……...,In}
Step 3 - Dataset preparation
Dn = {{I1,…In}D}
Step 4 - Dataset elaboration
DI = {GPA1,…GPAn, I1,…..In, nG1,…..nGn}
Step 5 - Processing
While (Dn!=0)
{
If (nGn==In)
Check GPAn;
}
Step 6 - Result Generation
R = {GPA, nG};
Step 7 - Stop

Where,

In = Input given by users
Dn = Dataset
D = Database
DI = Dataset contents
GPA1….GPAn = Previous GPA for each student
nG1…..nGn = New GPA (Predicted) for each student
R = Result generated

Flowchart of the Model

The system flowchart for training the data is shown in the Figure below:

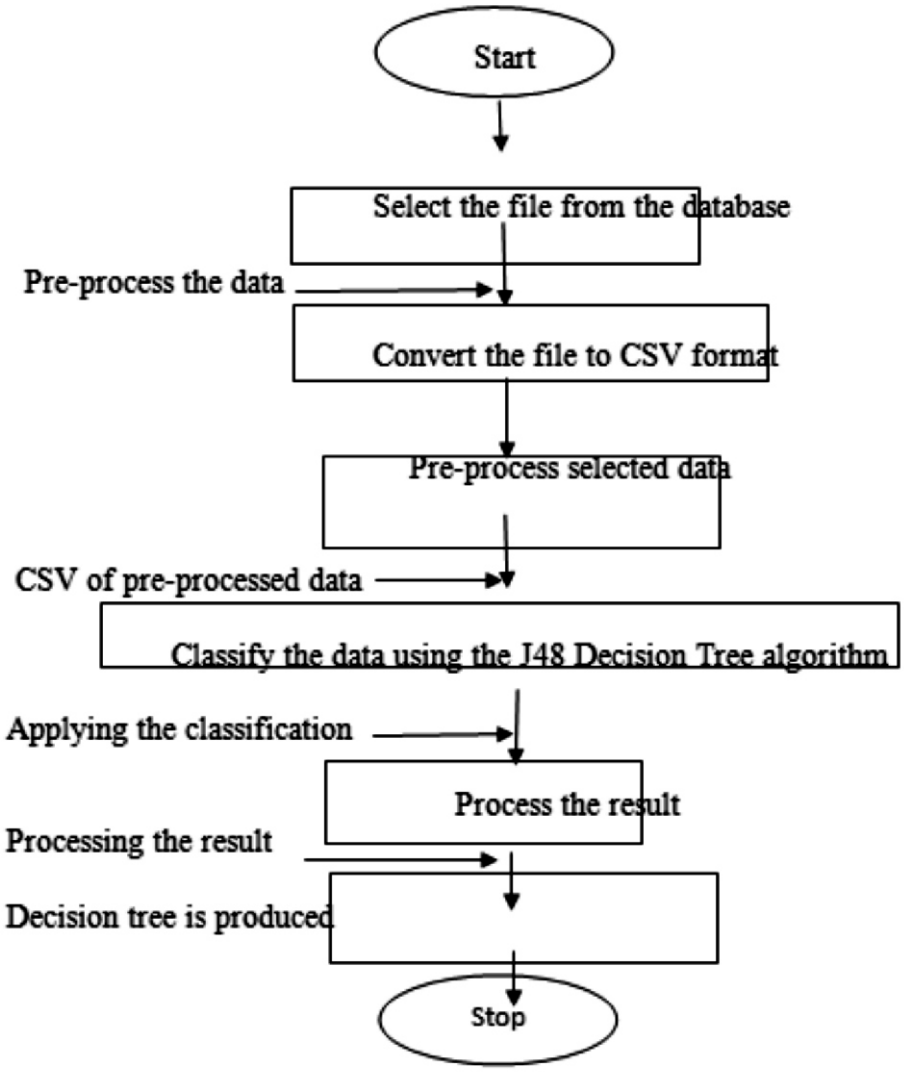
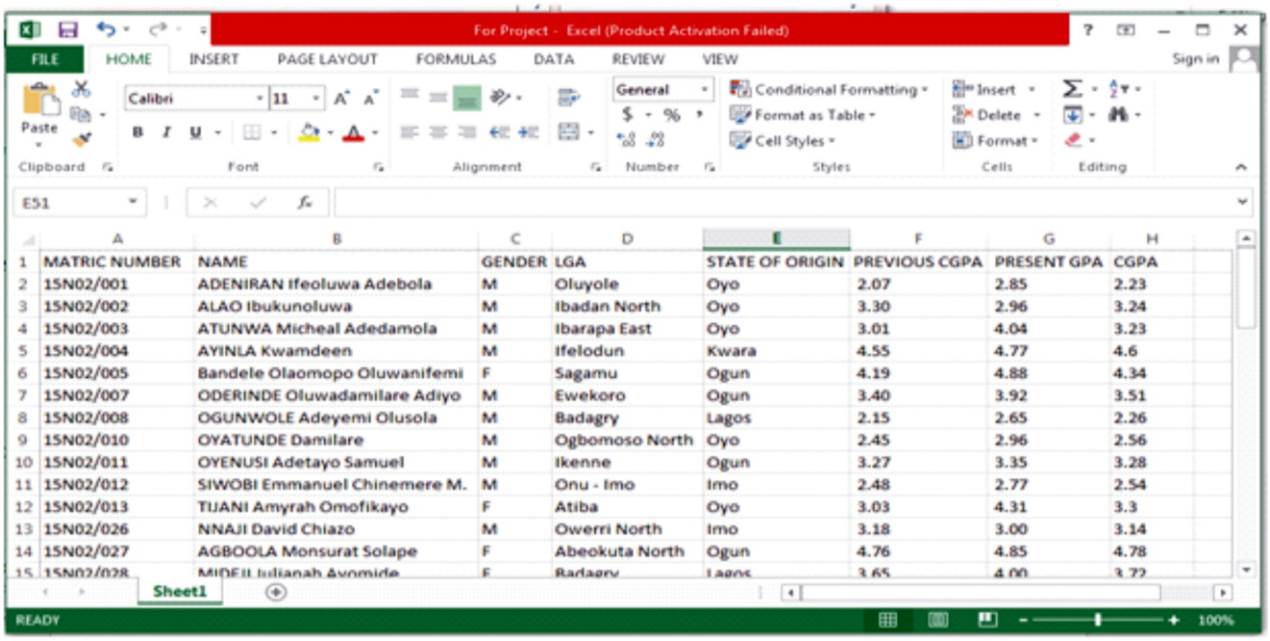


Figure 3: System flowchart for the training of the data

Method of Data Collection

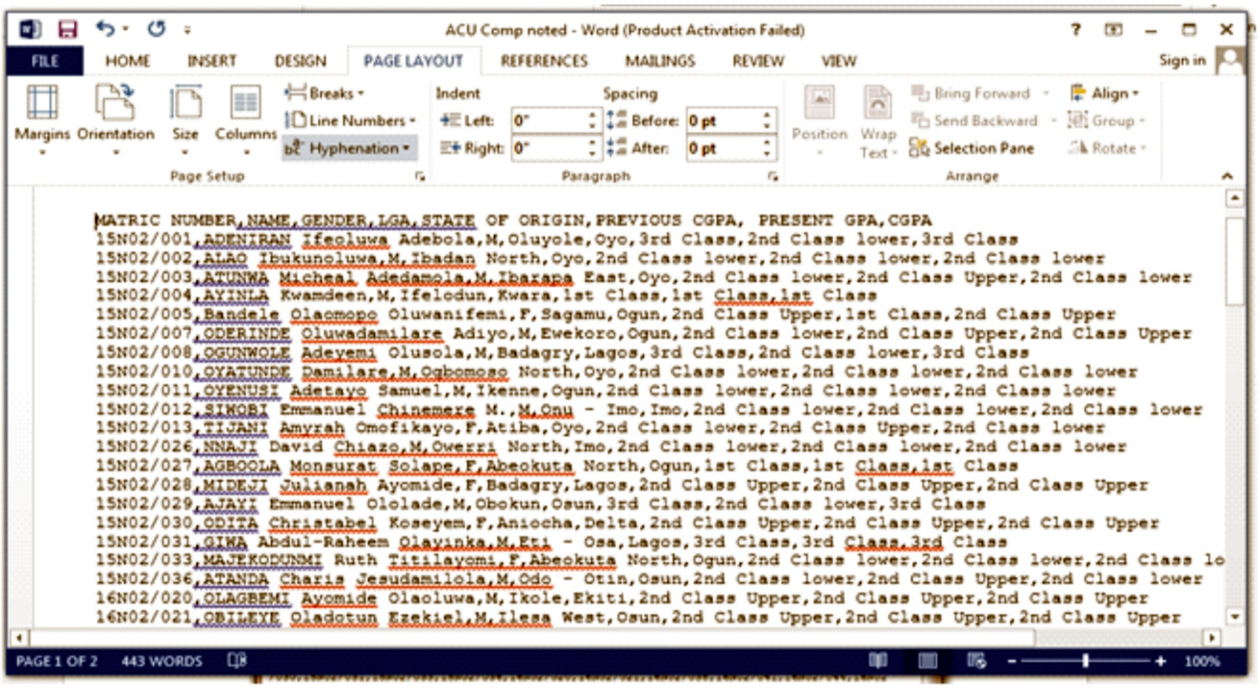
The method used for the collection of data was by getting the student's basic information such as the Matriculation number, Name of the student, Sex of the student, State of origin, and previous academic status of the student. The data collected from the department are in a Word format. The data was copied to the Microsoft Excel (spreadsheet) format:



	A	B	C	D	E	F	G	H
1	MATRIC NUMBER	NAME	GENDER	LGA	STATE OF ORIGIN	PREVIOUS CGPA	PRESENT GPA	CGPA
2	15N02/001	ADENIRAN Ifeoluwa Adebola	M	Oluyole	Oyo	2.07	2.85	2.23
3	15N02/002	ALAO Ibukunoluwa	M	Ibadan North	Oyo	3.30	2.96	3.24
4	15N02/003	ATUNWA Micheal Adedamola	M	Ibarapa East	Oyo	3.01	4.04	3.23
5	15N02/004	AYINLA Kwamdeen	M	Ifelodun	Kwara	4.55	4.77	4.6
6	15N02/005	Bandele Olaomopo Oluwanifemi	F	Sagamu	Ogun	4.19	4.88	4.34
7	15N02/007	ODERINDE Oluwadamilare Adiyo	M	Ewekoro	Ogun	3.40	3.92	3.51
8	15N02/008	OGUNWOLE Adeyemi Olusola	M	Badagry	Lagos	2.15	2.65	2.26
9	15N02/010	OYATUNDE Damilare	M	Ogbomoso North	Oyo	2.45	2.96	2.56
10	15N02/011	OYENUSI Adetayo Samuel	M	Ikenne	Ogun	3.27	3.35	3.28
11	15N02/012	SIWOBI Emmanuel Chinemere M.	M	Onu - Imo	Imo	2.48	2.77	2.54
12	15N02/013	TIJANI Amyrah Omofikayo	F	Atiba	Oyo	3.03	4.31	3.3
13	15N02/026	NNAJI David Chiazo	M	Owerri North	Imo	3.18	3.00	3.14
14	15N02/027	AGBOOLA Monsurat Solape	F	Abeokuta North	Ogun	4.76	4.85	4.78
15	15N02/028	MIDEJI Julianah Ayomide	F	Badagry	Lagos	3.65	4.00	3.72

Figure 4: Microsoft Excel format of the collected data

And saved as “comma separated value” format i.e.csv.



```
MATRIC NUMBER,NAME,GENDER,LGA,STATE OF ORIGIN,PREVIOUS CGPA, PRESENT GPA,CGPA
15N02/001,ADENIRAN Ifeoluwa Adebola,M,Oluyole,Oyo,3rd Class,2nd Class lower,3rd Class
15N02/002,ALAO Ibukunoluwa,M,Ibadan North,Oyo,2nd Class lower,2nd Class lower,2nd Class lower
15N02/003,ATUNWA Micheal Adedamola,M,Ibarapa East,Oyo,2nd Class lower,2nd Class Upper,2nd Class lower
15N02/004,AYINLA Kwamdeen,M,Ifelodun,Kwara,1st Class,1st Class,1st Class
15N02/005,Bandele Olaomopo Oluwanifemi,F,Sagamu,Ogun,2nd Class Upper,1st Class,2nd Class Upper
15N02/007,ODERINDE Oluwadamilare Adiyo,M,Ewekoro,Ogun,2nd Class lower,2nd Class Upper,2nd Class Upper
15N02/008,OGUNWOLE Adeyemi Olusola,M,Badagry,Lagos,3rd Class,2nd Class lower,3rd Class
15N02/010,OYATUNDE Damilare,M,Ogbomoso North,Oyo,2nd Class lower,2nd Class lower,2nd Class lower
15N02/011,OYENUSI Adetayo Samuel,M,Ikenne,Ogun,2nd Class lower,2nd Class lower,2nd Class lower
15N02/012,SIWOBI Emmanuel Chinemere M.,M,Onu - Imo,Imo,2nd Class lower,2nd Class lower,2nd Class lower
15N02/013,TIJANI Amyrah Omofikayo,F,Atiba,Oyo,2nd Class lower,2nd Class Upper,2nd Class lower
15N02/026,NNAJI David Chiazo,M,Owerri North,Imo,2nd Class lower,2nd Class lower,2nd Class lower
15N02/027,AGBOOLA Monsurat Solape,F,Abeokuta North,Ogun,1st Class,1st Class,1st Class
15N02/028,MIDEJI Julianah Ayomide,F,Badagry,Lagos,2nd Class Upper,2nd Class Upper,2nd Class Upper
15N02/029,AJAYI Emmanuel Ololade,M,Obokun,Osun,3rd Class,2nd Class lower,3rd Class
15N02/030,ODITA Christabel Koseyem,F,Aniocha,Delta,2nd Class Upper,2nd Class Upper,2nd Class Upper
15N02/031,GIWA Abdul-Raheem Olayinka,M,Eti - Osa,Lagos,3rd Class,3rd Class,3rd Class
15N02/033,MAJEKODUNMI Ruth Titilayomi,F,Abeokuta North,Ogun,2nd Class lower,2nd Class lower,2nd Class lo
15N02/036,ATANDA Charis Jesudamilola,M,Odo - Otin,Osun,2nd Class lower,2nd Class Upper,2nd Class lower
16N02/020,OLAGBEMI Ayomide Olaoluwa,M,Ikole,Ekiti,2nd Class Upper,2nd Class Upper,2nd Class Upper
16N02/021,OBILEYE Oladotun Ezekiel,M,Ilesa West,Osun,2nd Class Upper,2nd Class Upper,2nd Class Upper
```

Figure 5: CSV format of the collected data

The data are loaded in CSV format to the WEKA tool. During the scan of the data, WEKA computes some basic statistics on each attribute. The following statistics are shown in the 'Selected attribute' box on the right panel of the 'Pre-process' window:

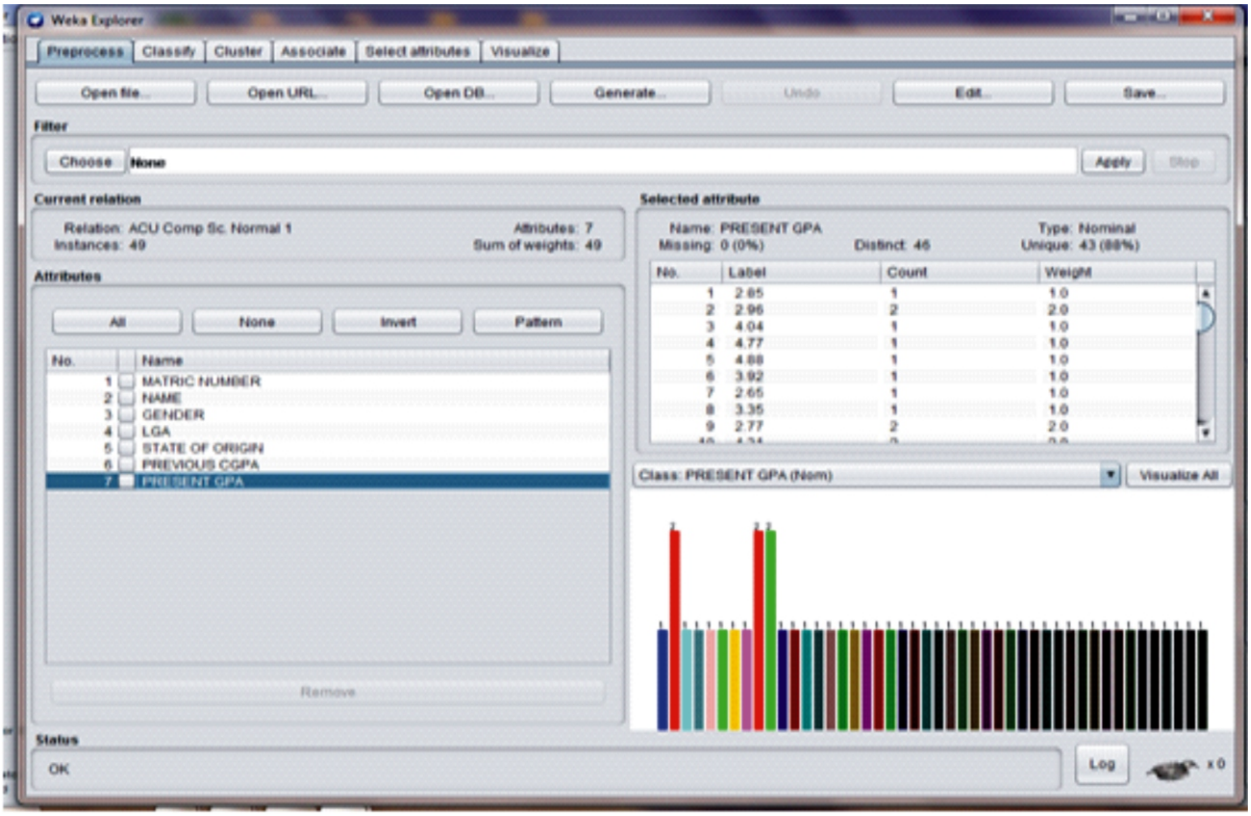


Figure 6: The Pre-process panel showing the attributes and basic statistics

Techniques for Data Analysis

The technique used for data analysis consists of the Training phase, Testing phase, Prediction phase, and Knowledge representation. The data set is divided into two (**2**) sets, a training set and a testing set. The training set is used to build the model from the J48 decision tree classifier, it evaluates the classifier on how well it predicts the class of the instances it was trained on as the student's Current G.P.A was used as the Key Performance Indicator, and various statistics in the output such as Precision, F-measure, TP, FP are analyzed. A total of forty-nine instances (49 students) were trained. The testing set is used to validate the model by checking the accuracy of the model on how well it predicts the class of a set of instances loaded from a file as a user-supplied test set. The prediction phase consists of the Prediction model created by leaving a blank column in which the prediction level is inserted by the classifier from which the students' performance is predicted based on the level (ranging between 0.5 to 1). The result generated by the model can further be represented in visualized (chart etc.) or tree form for further clarification and easy understanding.

1. RESULTS AND DISCUSSION

This work was developed using the J48 decision tree classifier in WEKA to predict the academic performance of forty-nine (49) students of the Department of Computer Science under seven (7) attributes.

From the experimental results, J48 algorithm predicted the unknown category of CGPA to the accuracy of 93.8776% which is fair enough for the system to rely upon for the prediction of students' academic performance and it also takes relatively little time to execute.

Setting Test Options

Two categories of data set were used in the course of the work, they are:

1. **Training data set -** evaluates the classifier on how well it predicts the class of the instances it was trained on. To produce the model, a training data was used, a data set with known output values was used to build the model. This type of model used the entire dataset as the training set which was used

to create the model and use the same dataset to test the accuracy of the model having identified the unknown attribute.
The summary of the model is as shown below:

=== Summary ===

Correctly Classified Instances	46	93.8776 %
Incorrectly Classified Instances	3	6.1224 %
Kappa statistic	0.9373	
Mean absolute error	0.0027	
Root mean squared error	0.0365	
Relative absolute error	6.2624 %	
Root relative squared error	25.0284 %	
Total Number of Instances	49	

The correctly classified instances are 46 at 93.8776% which signifies that the model is efficient enough to be relied upon in contrast to the incorrectly classified instances. The Kappa statistics is 0.9373 or 93.73% which indicates that the classifier is doing better than mere chance.
2. **Supplied test set -** evaluates the classifier on how well it predicts the class of a set of instances loaded from a file. Clicking on the '**Supplied test** Set' button brings up a dialog allowing you to choose the file to test on. The test data was created to control over-fitting after the model is created, it is tested to ensure that the accuracy of the model does not decrease with the test set. This ensures that the model will accurately predict future unknown values. Only nine (**9**) instances (the first nine students) were selected and used to evaluate the model.

=== Summary ===

Correctly Classified Instances	9	100 %
Incorrectly Classified Instances	0	0 %
Kappa statistic	1	
Mean absolute error	0	
Root mean squared error	0	
Relative absolute error	0 %	
Root relative squared error	0 %	
Total Number of Instances	9	
Ignored Class Unknown Instances	40	

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	2.85
1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	2.96
1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	4.04
1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	4.77
1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	4.88
1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	3.92
1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	2.65
1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	3.35

A total of nine (**9**) instances (students) were used to evaluate/test the model. The model has the same information as the Training data set model, thus showing that it is okay and efficient for prediction.
The correctly classified instances are 9 at 100%, the incorrectly classified instances 0 at 0%, Kappa Statistics is 1, and Ignored instances 40 (this is because only nine (**9**) students were used to test the model out of the forty-nine (**49**) students of ACU), True positive (TP) is 1, False positive (FP) is 0, Precision Rate, Recall, and F-Measure are 1. It thus reflects that the model is strong and better for prediction.

Prediction Model

After training and testing the model, data of unknown academic performance was then fed into the system for prediction. The prediction output of a given student is either "Pass" or "Fail". Out of forty-nine (49) datasets supplied into the system forty-six (46) were correctly predicted and only three (3) were incorrectly predicted. The percentage of correctly predicted datasets is 93.8776% which is fair enough to be entirely dependent on.

=== Summary ===		
Correctly Classified Instances	46	93.8776 %
Incorrectly Classified Instances	3	6.1224 %
Kappa statistic	0.9373	
Mean absolute error	0.0027	
Root mean squared error	0.0365	
Relative absolute error	6.2624 %	
Root relative squared error	25.0284 %	
Total Number of Instances	49	

The model having correctly classified instances of 46, Kappa statistics of 0.9373, True positive rate of 1, False positive rate of 0, Precision rate and Recall of 1, and F-measure of 0.667-1, shows that most of the students have a high tendency to have the same GP in the coming Semester.

The table below shows and interprets the prediction of the first nine (9) students as predicted by the model:

Student Name	Matric. Number	Current G.P	Predicted G.P	Prediction Level
ADENIRAN IfeoluwaAdebola	15N02/001	2.85	2.85	1
ALAO Ibukunoluwa	15N02/002	2.96	2.96	1
ATUNWA MichealAdedamola	15N02/003	4.04	4.04	1
AYINLA Kwamdeen	15N02/004	4.77	4.77	1
BandeleOlaomopoOluwanifemi	15N02/005	4.88	4.88	1
ODERINDE OluwadamilareAdisa	15N02/007	3.92	3.92	1
OGUNWOLE AdeyemiOlusola	15N02/008	2.65	2.65	1
OYATUNDE Damilare	15N02/010	2.96	2.96	1
OYENUSI Adetayo Samuel	15N02/011	3.35	3.35	1

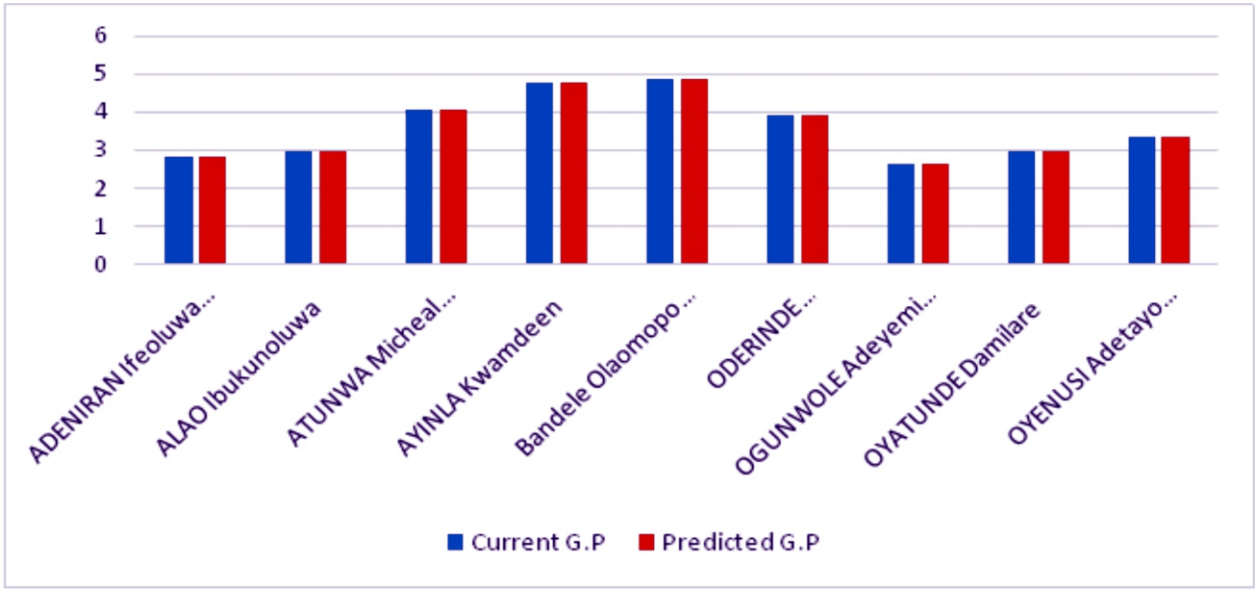


Figure 7: Prediction graph for the first nine students

From the graph, the Current G.P. is represented by the blue bar (left) while the Predicted G.P. is represented by the orange bar (right). The graph shows that the Predicted G.P. and the Current G.P. are at the same level for all the students, thus signifying that the student is liable to have the same G.P. in the coming Semester. For example, Adeniran Ifeoluwa had a GPA of 2.85 is likely to have the same GPA in the coming Semester because the prediction level is 1, the same goes for other students.

Summary

In this paper, the J48 method of classification was used to build a model to predict the student's academic performance based on its accuracy measure and prediction level for small datasets, and in terms of its merits above other decision tree algorithms. As the model was evaluated using the student's record of the Computer Science Department, Ajayi Crowther University, Oyo.

The simplicity of the J48 result output and its easy interpretation and prediction makes it a more convenient tool to predict students' academic performance.

Conclusion

The J48 decision tree presents and achieves a high rate of accuracy. It classifies the data into the correct and incorrect instances as we cover seven (**7**) attributes under forty-nine (**49**) instances and the model successfully identifies the students who are likely to fail. These students can be considered for proper counselling to improve their results in the coming Semester.

The system will generally help students to benchmark their grades from their entry point into the final year, thereby helping them to work harder to achieve this. Finally, the developed system would help to significantly reduce the overall failure rate in most academic institutions as students can be well guided and counselled.

Recommendation

From the results and findings of the experiments in this work, we recommend the adoption of the student performance prediction model as Education Data Mining is an emerging data Mining discipline. More similar studies on a different data set for the machine learning approach are needed to confirm the above finding.

Future work could also include applying data mining techniques on an expanded data set with more unique attributes to get more accurate results. Also, a comparative analysis of these results would be carried out based on other experiments results gotten from using other types of decision tree algorithms such as C4.5, CHAID, and CART.

REFERENCES

A. Dinesh Kumar, Dr. V. Radhika, A Survey on Predicting Student Performance, International Journal of Computer Science and Information Technologies (IJCSIT), Vol. 5 (5), 2014, 6147-6149.

Amjad Abu Saa (*2016*), Educational Data Mining & Students' Performance Prediction, *International Journal of Advanced Computer Science and Applications (IJACSA), Vol. 7, No. 5.*

E. Frank, M. Hall, G. Holmes, R. Kirkby, B. P Fahringer, I. H. Witten, and L. Trigg (2005), WEKA in Data Mining and Knowledge Discovery Handbook, Springer, Pp. 1305 –1314.

Jai Ruby & K. David (2014), "A study model on the impact of various indicators in the performance of students in Higher Education", IJRET International Journal of Research in Engineering and Technology, Vol. 3, Issue 5, May 2014, Pp.750-755.

Mehmed Kantardzic (2011), *Data Mining: Concepts, Models, Methods, and Algorithms*, Second Edition, Institute of Electrical and Electronics Engineers. Published 2011 by John Wiley& Sons, Inc.

Osmanbegovic E, Mirza S (2012), Data mining approach for predicting student performance. JEcon Bus 10: 3-12.

Oyelade, Oladipupo & Obagbuwa (2010), "Application of K-means clustering algorithm for prediction of student's academic performance". IJCSIS, Vol.7, No.1, Pp-292.

Pallavi, Sunila Godara, (2015) A Comparative Performance Analysis of Clustering Algorithms, International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622www.ijera.com Vol. 1, Issue 3, Pp. 441-445.

Purva Sewaiwar, Kamal Kant Verma (*2015*), Comparative Study of Various Decision Tree Classification Algorithm Using WEKA, ***International Journal of Emerging Research in Management & Technology ISSN: 2278-9359, October 2015 (Volume-4, Issue-10)***

Ramesh V, Parkavi P, Ramar K (2013), Predicting student performance: A statistical and data mining approach, IJCA 63: 35-39.

Sajadin et al (2011), **Prediction of student academic performance by an application of data mining techniques,** *International Conference on Management and Artificial Intelligence IPEDR Vol.6 IACSIT Press, Bali, Indonesia.*