American Journal of Engineering Research (AJER)
e-ISSN: 2320-0847 p-ISSN : 2320-0936
Volume-7, Issue-6, pp-336-343
www.ajer.org
Research Paper Open Access

Prediction Of Students' Performance In General Mathematics At Wassce Using Decision Tree

Esiefarienrhe Bukohwo Michael[1], Ochim Gold[2]
[1](Department of Mathematics, Statistics and Computer Science, College of Sciences/ Federal University of Agriculture, Makurdi, Nigeria)
[2](Department of Mathematics, Statistics and Computer Science, College of Sciences/ Federal University of Agriculture, Makurdi, Nigeria)
Corresponding Author: Esiefarienrhe Bukohwo Michael

ABSTRACT : *The most important subject in West African Examination Council (WAEC) and for admission into Nigerian universities irrespective of the course to be read is mathematics. It is the most dreaded subject by students mainly because of its logic and symbolic representations. This Research work conducted a review on education management and highlighted major factors affecting student's performances in Mathematics using the CART Algorithm incorporated in R. These factors after being trimmed were represented on a decision tree in order to develop a method for the prediction of students' performance in General Mathematics in WAEC. The most influential factor was the "Students' Previous Result in Mathematics". The results obtained from the Decision tree led to the development of various robust IF- THEN- ELSE statements which can subsequently be used to analyze future data. The result from the "prediction dataset" was represented graphically.*
KEYWORDS -: *CART,Data Mining, Decision tree, Mathematics,Performance,Prediction, WAEC.*

Date of Submission: 14-06-2018 Date of acceptance: 29-06-2018

I. INTRODUCTION

Education in recent time has taken a lot of emphasis. This is because of the obvious roles it plays in the development of a country. There is no doubt that what distinguishes the developed Nations from the developing nations of the world is the degree of science and technology prevalent in these nations and mathematics is the fulcrum of which Science and Technology rotates [1].

The West African Senior School Certificate Examination (WASSCE) is one of the most important examinations taken at the secondary school level in Nigeria senior secondary schools as it is used as an entry requirement into higher institutions.

Over the years, mathematics result in Secondary School at WASSCE has been very disheartening. Many students nearly cluster in the Pass (P) grades while majority obtain outright Fail (F) grades [2].

Many educational predictive systems have been developed in the past using several data mining techniques namely decision tree, neural networks, clustering etc.[3].

Decision tree is a popular data mining technique. It is commonly used by people because it is easy to understand and interpret, is able to handle both numerical and categorical data and is capable of mirroring human decisions more closely than other techniques.

Even though several prediction models has been developed in the past for various levels of education and for various subjects and courses, emphasis have not been laid on the prediction of performances in General Mathematics at WASSCE. There are many underlying factors influential to students' performances in General Mathematics at WASSCE. Some of the agents of influence surrounding the students' performance include not only the students themselves but the teachers, parents and peers and surroundings. Therefore, this paper is geared towards developing a decision tree to predict the performance of students in WASSCE Mathematics and visualizing the results predicted graphically. The remaining paper is discussed in sections. Sections II gives the background of the study while section III discusses the methodology used for the research. Section IV shows the results and discussions, then the section V concludes the paper and provides a way forward for future research.

II. BACKGROUND

Data mining has attracted an excellent attention within the data trade in society as an entirety in recent years, because of wide availableness of giant quantity {of knowledge and data of data} and therefore the close at hand want for turning such data into helpful information and knowledge[4].

[5]gave a clear definition on Data Mining. He said Data mining is the process of analyzing large amount of data to find out useful patterns and rules. Data mining which is the process of taking information from a data set and converting it into an understandable and meaningful structure for further use is a necessary process where various intelligent techniques are applied for extraction of useful patterns.

According to [6], most people use the word Data mining and Knowledge Discovery in Database (KDD) interchangeably. And this refers to the non-trivial extraction of implicit, previously unknown and potentially useful information from data in database. While data mining and Knowledge Discovery Database (or KDD) are frequently treated as synonyms, data mining is actually part of the Knowledge Discovery Process.

Data mining techniques could be applied in a wide range of organizations so long as they deal with collecting data. There are several data mining software being made available to the market today, to help companies tackle decision making problems and invariably overcome competition from other companies in the same business [7].

Data mining functionalities are used to specify the kind of patterns to be found in data mining tasks. Data mining tasks can be classified into two categories – descriptive and predictive.

[8]said Predictive mining tasks perform inferences on the current data in order to make predictions. The purpose of the predictive model is to allow the data miner to predict an unknown (often future) value of a specific variable; the target variable. If the target value is one of a predefined number of discrete (class) labels, the data mining task is called Classification. If the target variable is a real number, the task is regression.

There are increasing research interests in using data mining techniques in education. This emerging field is called educational data mining. [9].

Educational data mining is a process used to extract useful information and patterns from a huge educational database [10].It inherits its properties from areas like Learning Analytics, Psychometrics, Artificial Intelligence, Information Technology, Machine learning, Statistics, Database Management System, Computing and Data mining [11].

While the Analysis of educational data is not itself a new practice, recent advances in educational technology, including the increase in computing power and the ability to log fine-grained data about student's use of a computer-based learning environment, have led to an increased interest in developing techniques for analyzing the large amounts of data generated in educational settings [12].

According to [13], the useful information and patterns can be used in predicting students' performance. As a result, it would assist educationists in providing an effective learning approach and students to improve the learning activities.

The educational data mining can be applied to various data that are used in communicating with Stakeholders, students modeling, structuring the educational domain, predicting students grades etc.

In educational data mining method, the most popular task to predict students' performance is classification. There are several algorithms under classification tasks that have been applied to predict students' performance. Among the algorithms used is Decision Tree.

Decision tree is one of the popular techniques for prediction. Most researchers have used this technique because of its simplicity and comprehensibility to uncover small or large data structure and predict the value [3]

A decision tree is a flow chart likes structure where each node denotes a test on an attribute value, each branch represents an outcome of the test and tree leaves represents classes or class distribution. It is a prediction model that can be viewed as a tree where each branch of the tree is a classification question and leaves represent the partition of the data set with their classification [8].

[14] found the Decision tree as the most accurate technique in predicting grade point average.

According to [15], the key to construct decision tree is how to choose better logical judgment or attribute. There can be many choices to the same set of examples. Research shows that in general, the smaller the tree, the stronger forecasting ability. The key of constructing decision tree as small as possible is to choose the proper attribute caused branch.

[16]stated that there are various decision tree algorithms used; the ID3, C4.5 CART. CART stands for Classification And Regression Trees. It was introduced by Breiman in 1984. It builds both classifications and regression trees. The classification tree construction by CART is based on Binary Splitting of the attributes. CART can automatically perform variable selection, use any combination of continuous or discrete variables, is average in speed, is Non parametric (no probabilistic assumptions) and uses Gini Index as an attribute selection measure to build a decision tree unlike ID3 and C4.5.

One major distinguishing feature is that CART uses cost complexity pruning to remove the unreliable branches from the decision tree to improve accuracy [17].

Certain models have been built so far to predict students' performance. The accuracy of a model is likely to improve if we add the attributes that reveal the current performance (e.g. attendance, test marks etc.) and consider more instances [18].

After studying the data mining approaches for predicting student's performances, [19] suggested that experiments be extended with more distinctive attributes to get more accurate results, useful to improve the students learning outcomes. And also, experiments could be done using other data mining algorithms to get a broader approach and more valuable and accurate outputs. Different software may be utilized while at the same time various factors as well. In a nutshell, a wide range of relevant attributes, factors and techniques should be used when predicting student's academic performance.

In conclusion, prediction of individual Students Performance in the May/June WASSCE is a need for Nigerians as the poor performance so far in Mathematics over the years has been a source of concern. The use of reliable data mining tools to effectively predict the future performance promises to make the prediction reliable and help stakeholders make good decisions about the future.

Although there are a lot of researches on Educational Data Mining, there still remains a vast area of researches to cover in Educational Data Mining and a wider area to cover in WAEC Performances in various subjects.

III. METHODOLOGY

This section basically describes the method or approaches used by the researcher in order to arrive at a productive conclusive research work.

The Data mining technique used to analyze the data is The Decision Tree. The Classification and Regression Tree (CART) Algorithm was used for the formation of the Decision Tree. This is chosen because of the ability of the Algorithm to properly eliminate those factors that do not influence performance.

3.1 The Proposed Model

The methodology used in this work is the Knowledge Discovery in Databases (KDD) which can be seen as an automatic non-trivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data. Data Mining is a step in the KDD process. It consists of the application of particular data mining algorithms to extract the information and patterns derived by the KDD process. The specific method to be utilized is the data mining techniques called Decision Tree.

The Diagram below shows the Knowledge Discovery in Database Process Model.

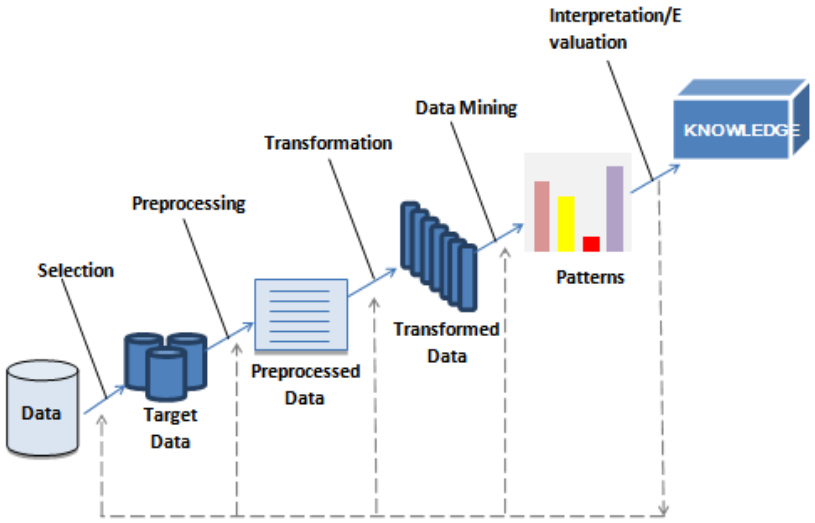

Figure 1: The KDD Process Model

The Decision Tree Prediction Model was formed using the CART algorithm in R. The data collected was for the subject General mathematics in the West African Senior School Certificate Exam. The data set collected from the admission records of the first to third year students of the Federal University of Agriculture, Makurdi. The dataset mined was 200 in number.

3.2 Extraction of Data

The factors that influenced past performance in Mathematics, stemming not from the Exam body but from the Students immediate surrounding environments are recorded in a tabular form as shown in table 1.

Table 1: Factors that influenced performance in Mathematics

S/N	NAME	POSSIBLE VALUES
1	Gender	Male / Female
2	Age	<=20 / >20
3	Marital Status (Masta)	Single / Married
4	Living with Parents/Guardian (Livpg)	Parents(P) / Guardian (G)
5	Parents cohabitation status (pcs)	Living Together(Tog) / Living Apart (Apa)
6	Family size (Fs)	<=4 / >4
7	Exam phobia	Yes / No
8	Private Home Lessons (phl)	Yes / No
9	School Type (st)	Boarding (Bor) / Day
10	Scholarship Beneficiary (sb)	Yes / No
11	Interest in Higher Education (Ihe)	Yes / No
12	Alcohol Consumption (ac)	Yes / No
13	Have a Boyfriend/Girlfriend (hbg)	Yes / No
14	Hangouts with Friends (hwf)	Yes / No
15	Repeated an exam/class (rc)	Yes / No
16	Use of Phone/laptops to solve math problems (upl)	Yes / No
17	Social Media Status (sms)	Always Active(Ac) / Sometimes Active (So) / Rarely Active (Ra)
18	Math Teacher's educational Qualification (meq)	WASSCE (W) / First Degree(F) / Postgraduate Degree (P)
19	Father's Highest Educational Qualification (fea)	FSLC (F) / WASSCE (W) / Tertiary (T)
20	Mother's Highest Educational Qualification (mea)	FSLC (F) / WASSCE (W) / Tertiary (T)
21	Parent's total incomes (pti)	<=N70,000 / N70,000-N150,000 / N150,000 and above
22	Family House Location (fhl)	Highly populated (hp) / Low populated (lp) / Averagely populated(ap)
23	Performance in other subjects (pios)	Very bad (vb) / Fair (Fa) / Good (Go)
24	School Previous Result in Mathematics (sprimp)	Good (Go) / Bad (Bd) / Average (Av)
25	Attendance in School (ais)	<=40% / 41-75% / 76-100%
26	Study Hours (sh)	<=2hours / 2-5hours / 5-10hours
27	Class Type (ct)	Science (Sci) / Arts / Social Science (SocSci)
28	Interactions with classmates (iwc)	Good (Go) / Poor (Pr) / Average (Av)
29	Health	Good (Gd)/ Poor(Pr) / Average (Av)

A data set "**tablegrade"** was inputted and created in the R console. The data was split into two; the training datasets (to enable the machine learn) and the testing dataset (to discover the accuracy of the test datasets). Out of the 200 dataset, 150 datasets were used for training and 50 datasets were used for testing. The predication model was then developed after the training and the testing and the model created was then represented graphically.

3.3 Assumptions of the Model:

It is assumed that the data collected and inputted which forms the derivative of the model is accurate and without errors. This is assumed because change in the collected or inputted data will cause a change in the model.

3.4 Advantage of the Model

- The model uses the basic influential attributes filtered by the Algorithm as the independent variables hence leading to a stable model.

IV. RESULTS AND DISCUSSION

The diagram below shows the training data set represented on a decision tree:

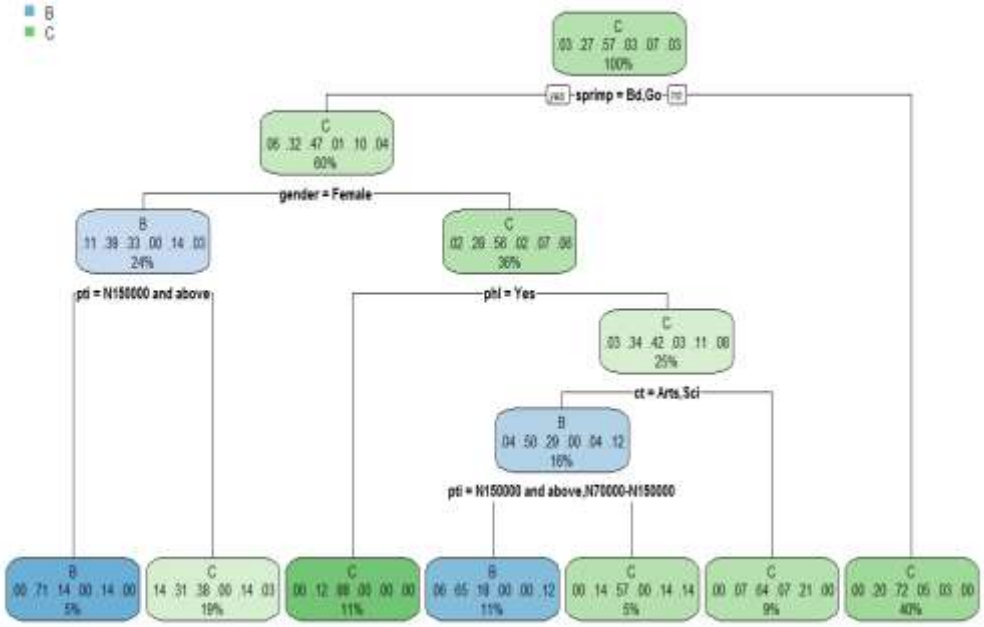

Figure 2: decision tree for the Training Datasets

From the data extracted, the decision tree above shows that the most popular grades in the WASSCE mathematics in the past has been the C. The data has grades and corresponding probability thus: A(0.03), B(0.27), C(0.57), D(0.03), E(0.07), F(0.03).

The most important attribute as classified by the Decision Tree Algorithm (Classification and Regression Tree Algorithm) is the School Previous Result in Mathematics (Sprimp). A greater percentage (60%) has School Previous Result in Mathematics as "Good" and "Bad". And a smaller percentage (40%) has the School Previous Result in Mathematics as "Average". From this, a student's performance can be predicted from the knowledge of the student's previous result. Hence if a student has a trend of performing well in Mathematics in the school's internal examination, the student is supposed to get a good grade at the WASSCE General Mathematics.

The next important attribute is the Gender. Out of the 60% whose School performance in Math Previous Result was "Good" and "Bad", 24% of the data are Female and 36% male. The male have majority of the grade clustered around C and then followed by B, E, F, A and D in the mentioned order. The Female have their grades clustered around B firstly, then C, E, A and F (no D). This shows that in as much as the male pass mathematics more with a C, the females when they pass Mathematics in WASSCE pass with better grades and when the female fail, their grades cluster between E and F unlike the Male whose failures are spread amongst D, E and F. Hence the Male pass Mathematics in WASSCE with an A-C grade of 86% and the female with an A-C grade of 83%.

From the decision tree, we have 5% of the Female whose "Parents Total Income" is greater than N150,000 and 19% for those whose "Parent's Total Income" is below N150,000. For the female whose Parents Total Income is greater N150,000, majority score a B in General Mathematics WASSCE followed by a C and then an E. For those Female students whose Parents Total Income is less than N150,000, their grades cluster around C and then B, A, E and an F. This shows that Parent's Total Income encourages very good performance in Mathematics. For the Male whose School Previous Result in Mathematics was Good/Bad, 11% had Private Home Lessons and had their grades basically at C after which it had grades clustered at B and none from the data got a D,E,F and A. 25% of those who don't have Private Home Lessons had a cluster of C as well and their grades spread over B, E, F, A and D with frequency of occurrence as in the mentioned Sequence. This shows that Private Home Lessons help to Prevent Failure in Mathematics Performance in WASSCE even though Excellence is not promised.

For the male population sample who didn't receive Private Home Lessons, 16% of the data are Science and Arts students and 9% are Social Science students. The Arts and Science students have a higher number of B as grades then C, F and then E. The Social Science Students have their grades centered on C, then E, B, and D. So from the tree, we can see that the Science and Arts Students perform better in General Mathematics in WASSCE.

Finally, for the male students who don't have lesson Teachers and are in the Arts and Science Class, 11% their Parents earn N70, 000 and above and 5% earn below N70, 000. Their grades show that those students whose parents earn above N70, 000 perform better in General Mathematics in WASSCE than those whose parents earn less than N70, 000

In summary, the major attributes affecting the previous performance in General Mathematics in WASSCE include School Performance in Previous Math Result, Gender, Parents Total Income and Class Type. This means that the other attributes such as Alcohol Consumption, Social Media Status have no strong influence on the performance in General Mathematics in WASSCE.

4.1 Derived Rules from the Decision Tree

The Corresponding decision Rules from the tree is given below:

1. If School Performance in Previous Math Result is Equal to Good/Bad and
i. Gender is Equal to Female and
a) Parents Total Income is Greater than N150, 000 then the Grade is likely to be B.
b) Parents Total Income is less than N150, 000 then the Grade is likely to be C.
ii. Gender is equal to Male and
a) Private home lesson is Equal to Yes then the Grade is likely to be C.
b) Private home lesson is Equal to No and Class type is Equal to Science or Art and Parents Total Income is Greater than or Equal to N70,000 then the Grade is likely to be B.
c) Private home lesson is Equal to No and Class type is Equal to Science or Art and Parents Total Income is less than N70, 000 then the Grade is likely to be C.
d) Private home lesson is Equal to No and Class type is Equal to Social Science then the Grade is likely to be C.
2. If School Previous Result in Mathematics is Equal to Average, then the Grade is likely to be C.

4.2 Prediction Analysis

The mined data after being computed and analyzed using the Classification and Regression Tree Algorithm (CART), has the following result for its prediction using Test Values.

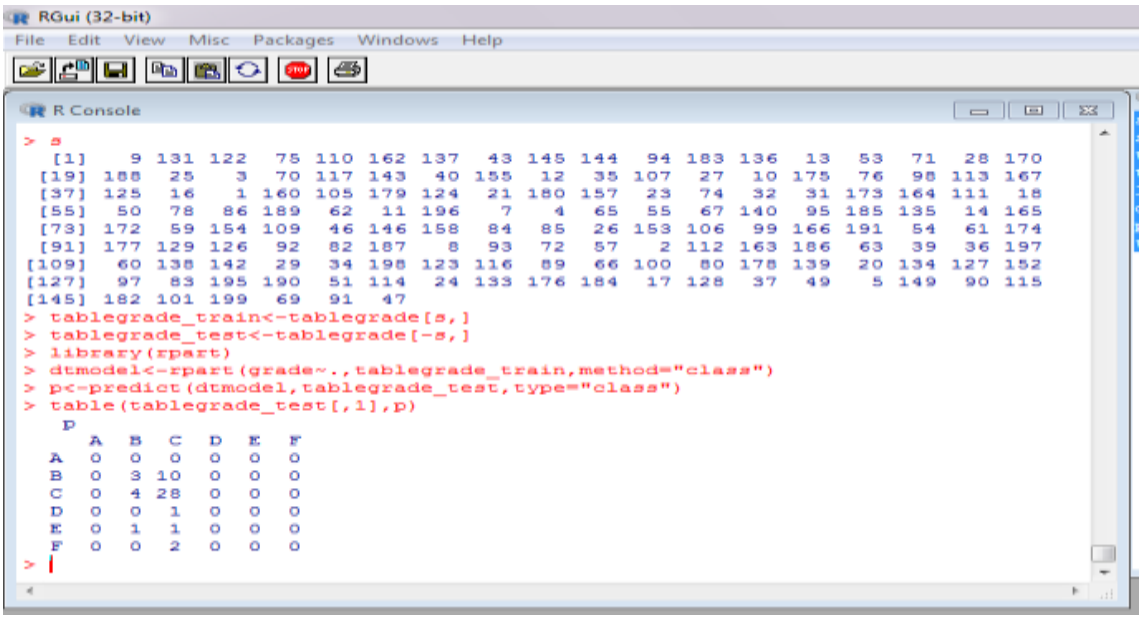

Figure 3: prediction results from the Test Dataset

The above shows that the grade A was never predicted wrongly. The table goes further to show that the grade B was wrongly predicted as C, 10 times in all with a 20% error rate. C was predicted wrongly as B 4 times with an error rate of 6%. D was wrongly predicted as C once and E wrongly predicted as B once with both error rates as 2%. The grade E was also wrongly predicted as C once with an error rate of 2% also. F was wrongly predicted as C 2 twice with error rate as 4%. Generally the error was highest at predicting B as C though the error rate is minimal.

From the table also, 31 out of 50 grades were correctly predicted. Hence the accuracy for the prediction using the Classification and Regression Tree (Decision Tree) technique is 62%. This prediction is a lot higher

than average and hence can be regarded as averagely accurate. Especially in regards to the small size of Test dataset used. The predictor centered on predicting correctly the most popular grades gotten in WAEC; the grade C.

The analysis done by the data mining Technique "Decision Tree" has been used to first of all discover the most influential factors affecting the grades in Mathematics in WASSCE. And then discover the Hierarchy of these discovered influential factors. It can be observed that out of the 29 factors that were thought about by the Researcher, only 5 factors were discovered to really influence the grades gotten in General Mathematics in WASSCE. The model derived from this reduced factors can be used for further activities as were described in

Fig 1. The resulting graph for the predicted dataset is shown below:

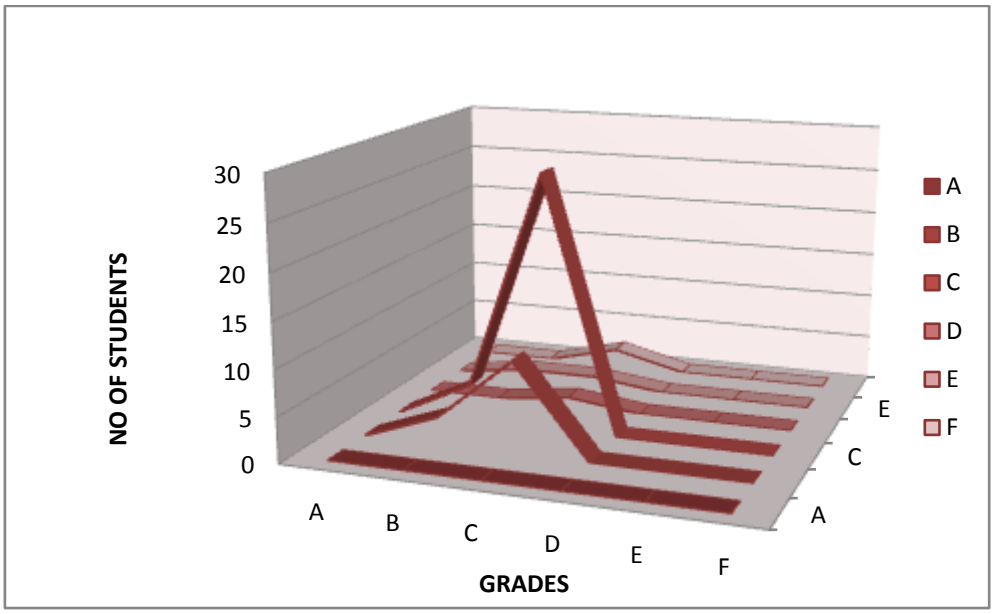

Figure 4: graph for the predicted Test Dataset

V. CONCLUSION

Prediction models in Educational Data Mining has caught a lot of attention in Recent time and the performance in General Mathematics in the WASSCE over the years has given cause to students, teachers, the Government and Parents to worry. Placing the blame on the WAEC body as a general problem has done little or no good in improving Performances and hence the need to predict performance using a wide range of relevant factors that surrounds the students' immediate environment cannot be over-emphasized. To this end, the author has used the Decision Tree Technique (CART) to analyze these data that greatly affects performances in Mathematics. The objectives of this work has been achieved by the use of a very easy to understand mining Technique; Decision tree (CART). The use of this technique has helped to limit the wide range of attributes originally used for the study. These attributes were trimmed based on how influential they were to students WAEC results in General Mathematics. Hence the use of wide range of attributes (factors) of data should not be regarded as important as the strength of influence of attributes on performance. For future purposes, the researchers recommend the use of very large data to develop a more stable model and discover more knowledge in order to assist in understanding and correcting the problems of poor performances in the WASSCE General Mathematics.

REFERENCES

[1]. M. Musa and E.S.Dauda,Trends Analysis of Students' Mathematics Performance in West African Senior School Certificate Examination from 2004 to 2013: Implication for Nigeria's Vision 20:2020, *British Journal of Education, 2(7)*, 2014, 50-60.

[2]. H.S. Tsok, S.S. Kpanja, andS.M. Hwere, A Comparative of Student's Performance in S.S.C.E Mathematics and Pre-National Diploma (PRE-ND) Programs Mathematics: A case study of Nassarawa State Polytechnic Lafia,*Journal of Education and Practice, 4(27)*, 2013, 69-70.

[3]. S. Natek andM. Zwilling, Student data mining solution–knowledge management system related to higher education institutions, *Expert systems with applications, 41(14)*, 2014, 6400–6407.

[4]. R. Kalpana, N. Shanthi and S. Arumugam, A survey on Data Mining Techniques in Agriculture,*International Journal of Advances in Computer Science and Technology, 3(8)*, 2014, 426.

[5]. H. Kaur, A Review of Application of Data Mining in the Field of Education,*International Journal of Advanced Research in Computer and Communication Engineering, 4(4)*, 2015, 409.

[6]. J.C Aruna andP.K. Butey, Importance of Data Mining with Different Types of Data Application and Challenging Areas,*Journal of Engineering, Research and Applications,4(5)*, 2014, 38.

[7]. K. Amarendra, A Survey on Data Mining and its applications, *International Journal of Emerging Trends of Technology in Computer Science (IJETTCS),3(3)*, 2014, 163.

[8]. N. Jain and V. Srivastava, Data Mining Techniques: A Survey Paper, *International Journal of Research in Engineering and Technology, 02(11)*, 2013, 116.

[9]. A. AL-Malaise, A. Malibari, and M. Alkhozae, Students Performance Prediction System using Multi Agent Data Mining Technique,*International Journal of Data Mining and Knowledge Management Process,4(5)*, 2014, 1.

[10]. D. M. D. Angeline, Association rule generation for student performance analysis using Apriori algorithm,*The SIJ Transactions on Computer Science Engineering & its Applications (CSEA),1(1)*, 2013, 12–16.

[11]. R. Jindal and M.D. Borah, A study on Educational Data mining and Research Trends,*International Journal of Database Management Systems (IJDMS), 5(3)*, 2013, 53.

[12]. P. Nithya, A. Umamaheswari and A. Umadevi, A survey on Educational Data mining in Field of Education, *Journal of Computer Science and Software Development,1(1)*, 2016, 1.

[13]. M.A. Shahiri, W. Husain, and A.N. Rashid, A Review on Predicting Students Performance using Data Mining Techniques, *Proc. Computer Science 3r International Conference on Information System*, Penang Malaysia, 2015, 417-420.

[14]. N.S. Shah, Predicting Factors that Affect Student's Academic Performance by using Data Mining Techniques. *Pakistan Business Review, 13(4)*, 2012, 631-638.

[15]. Q. Dai, C. Zhang, andH. Wu, Research of Decision Tree Classification Algorithm in Data Mining,*International Journal of Database Theory and Application,9(5)*, 2016, 2.

[16]. H. Sharma andS. Kumar, A survey on Decision Tree Algorithm of Classification in Data Mining,*International Journal of Science and Research (IJSR), 5(4)*, 2016, 2095.

[17]. I.A. Ganiyu, Data Mining: A Prediction for Academic Performance Improvement of Science Students using Classification,*International Journal of Information and Communication Technology Research, 6(4)*, 2016, 2224.

[18]. R.R. Kabra and Bichkar R.S. Performance Prediction of Engineering Students Using Decision Trees,*International Journal of Computer Applications, 36(11)*, 2011, 12.

[19]. E. Osmanbegovic and M. Suljic.Data Mining Approach for Predicting Students Performance,*Journal of Economic Review, 10(1)*, 2012, 11.

Esiefarienrhe Bukohwo Michael "Prediction of Students' Performance in General Mathematics at Wasco Using Decision Tree." American Journal of Engineering Research (AJER), vol. 7, no. 06, 2018, pp. 336-343.