STUDENTS' PERFORMANCE PREDICTION USING CLASSSIFICATION ALGORITHMS

ABSTRACT

It is imperative to analyze educational data especially as it relates to students' performance. Educational institutions need to have a fairly accurate prior admitted students' knowledge to predict their future academic performance. This helps to identify the good students and also provides an opportunity to pay attention to and improve those who would possibly not perform too well. As a solution, this paper proposed a system which can predict the performance of students from their previous academic record using concepts of data mining techniques under Classification. The dataset containing information about students, such as gender, age, SSCE grade, UTME score, post UTME score and grade in students first year. ID3 (Iterative Dichotomiser 3) and C4.5 classification algorithms was applied on the data to predict the academic performance of students in future examinations.

***Keywords**: Classification; Decision Tree; Data Mining; Academic Performance, Machine Learning*

1. INTRODUCTION

Classification method is the most frequent technique which is used to classify data set. Presently classification is used in several fields such as education, industrial, medical and other many places [1]. Classification is basically a data mining technique in which some input pattern is applied to get desired output by using any classification algorithm. The task of developing effective academic prediction system is a critical issue for educators [2].On yearly basis, higher institutions admit students from different locations and educational background with varying scores in entrance examinations into various departments. Previous studies have revealed that various factors are responsible for students' failure which includes low socio-economic background, student's intellectual capacity, school and home environment, or the support given by parents and other family members [3]. Methodologically, analysis of the previous academic performance of students admitted can be used to better predict their future performance using the concept of machine learning. In this regard, the data of students enrolled in 2008/2009 academic session of Joseph Ayo Babalola University was obtained and used in this study. This data includes attributes such as gender, age, SSCE grade, UTME score, post UTME score and grade in student's first year, category and admission type. Two decision tree algorithms (ID3 and C4.5

algorithms) were used to predict the futureperformance of the student using the dataset. The results of the two algorithms were then compared to determine the most effective algorithm for the prediction.

2. LITERATURE REVIEW

A review of relevant literatures was carried out.Abeer and Elaraby [4] analysed previously enrolled students' data in a specific course program across 6 years (2005–2010), with multiple features collected from the university database. The work predicted the students' final grades in the particular course program. Pandey and Pal [5] presented a data mining approach to classify students' according to performers or underperformers class using Naïve Bayes algorithm, classify. Bhardwaj and Pal [6] did a comparative study to test multiple decision tree algorithms on an academic dataset in order to classify the student's academic performance. The work primarily concentrates on choosing the best decision tree algorithm from among commonly used decision tree algorithms, and then provides a standard for them individually. It was discovered that the CART decision tree technique performed reasonably better on the dataset used for testing, that was obtained based on the accuracy and precision produced at the validation stage. Livieris, *et al.* [7] developed an Artificial Neural Network (ANN) classifier to predict the performance of students in Mathematics. From their experiments they discovered that the modified spectral Perry trained artificial neural network performs better classification compared to other classifiers. Kotsiantis, *et al* [8] explored machine learning techniques for dropout prediction of students in distance learning. This study contributed in that it carved the path for educational data mining and one of the first works to implemented machine learning methods in an academic environment. Their algorithm was fed on demographic data and several project assignment rather than class performance data to make prediction of students. Moucary, *et al.* [9] applied a hybrid technique on K-Means Clustering and Artificial Neural Network for students who are pursuing higher education while adopting a new foreign language as a means of instruction and communication. Firstly, Neural Network was used to predict the student's performance and then fitting them in a particular cluster which was form using the K-Means algorithm. This clustering helped in serving a powerful tool to the instructors to identify students capabilities during their early stages of academics. Hongsuk, *et al*. [10] develop a Deep Neural Network supervised model to estimate link based flow of traffic conditions. A Traffic Performance Index was used for logistic regression to distinguish between a congested traffic condition and a non-congested traffic condition. The 3 layer model was able to estimate the congestion with a 99% of accuracy. Yadavto estimate the congestion with 99% accuracy. Yadav and Pal [11] proposed a prediction model for students' performance based on data mining methods with some few features called student's behavioral features. The model was evaluated using three different classifiers; Naïve Bayesian, Artificial Neural Network and Decision Tree. Random Forest, Bagging and Boosting were used as ensemble methods to improve the classifier's performance. The model achieved up to 22.1% more in accuracy compared when behavioral features were removed. It increased up to 25.8% accuracy after using the ensemble methods.

3. METHODOLOGY

The methodology of this study is composed of: identification of the required variables for students performance prediction, the collection and preparation of data, formulation of the predictive models using the supervised machine learning algorithm (Decision Tree), simulation of the predictive models using the WEKA simulation environment and the performance evaluation metrics applied during model validation for the predictive models performance evaluation.

a. Data Collection and Preparation

The dataset used in this study was obtained from the academic record office, Joseph Ayo Babalola University. The data was anonymously obtained without any bias. Personal and academic record of students admitted in 2008/2009 into the university from Six (6) major departments namely: Computer Science (CSC), Accounting (ACC), Political Science (POL), Microbiology (MCB), Economics (ECO), Business Administration (BUS) was used. The size of the dataset is 100 records.

b. Model Formulation

In this study, decision tree algorithm was used in formulating the model for prediction because the pattern explaining the link between the attributes identified (input attributes) and the student's performance (the target attribute) was needed. The pattern identified was then converted into a set of rules that can assist in making informed decisions regarding the performance of students. In formulating a predictive model using supervised machine learning algorithm, a mapping function is used to easily state the general expression. The dataset S which consists of the records of students containing fields representing the set of classification factors (i number of input variables for j students), X_{ij} alongside the respective target variable (student's performance) denoted by the variable Y_j – the student's performance for the j^{th} individual in the j records of data for the study. The mapping function that defines the link between the classification features and the target attribute – classification of student's performance is given in equation (1).

$$\varphi: X \quad \rightarrow \quad Y \qquad (1)$$
$$defined\ as: \varphi(X) = Y$$

The equation shows the relationship between the set of classification factors represented by a vector, X consisting of the values of i variables and the label Y which defines the student's performance – First, Two-1, Two-2 and Third of each student as expressed in equation (2). Assuming the values of the set of variables for a student is represented as $X = \{X_1, X_2, X_3, \ldots\ldots, X_i\}$ where X_i is the value of each variablei = 1 to j; then the mapping φ which represents the predictive model for student's performance maps the variables of each one to their corresponding student's performance according to equation (2).

$$\varphi(X) = \begin{cases} First \\ Two - 1 \\ Two - 2 \\ Third \end{cases} \qquad (2)$$

The decision trees developed for the performance of students was used to propose a set of rules that can be used to determine the student's performance directly just by observing the value of the variables identified by the model and the succession of events. Also, the set of attributes identified in the final decision trees model for student's performance are the variables which have the most relevant importance to the determination of each student's performance. It was proposed to be given much consideration during performance assessment of students.

For the training dataset, S is a set containing $S_1, S_2, \ldots., S_j$ of samples that have been classified already of the students' records which consist the values of their variables, $X = \{X_1, X_2, \ldots., X_i\}$ together with the classification of student's performance, $Y = \{First, Two - 1, Two - 2, Third\}$ such that, $S = (X, Y)$ for all students from 1 to j.

In this work, C4.5 and ID3 classification algorithms were used for the predictive model formulation. The two conditions used by the C4.5 decision trees in developing its decision trees are stated in equations (3) and (4) defined as the information gain and the split criteria respectively. Equation (3) is used in determining which attribute is used to split the dataset at every iteration while equation (4) is used to determine which of the selected attribute split is most effective in splitting the dataset after attribute selection by equation (3).

$$IG(X_i) = H(X_i) - \sum_{t \epsilon T} \frac{|t|}{|X_{ij}|} \cdot H(X_i) \ (3)$$

where:

$$H(X_i) = -\sum_{t \epsilon T} \frac{|t, X_i|}{|X_{ij}|} \cdot \log_2 \frac{|t, X_i|}{|X_{ij}|}$$

$$Split(T) = -\sum_{t \epsilon T} \frac{|t|}{|X_{ij}|} \cdot \log_2 \frac{|t|}{|X_{ij}|} \qquad (4)$$

T is the set of values for a given attribute X_i.

The simulation of the predictive model was performed in WEKA environment.

c. 10-fold Cross Validation (Model Validation)

Cross-validation procedure was used in this work. This entails splitting the entire datasets into some folds (or partitions). Each fold was selected for testing, with the remaining k – 1 fold; the subsequent fold was used for testing with the remaining fold (together with the first fold used) used for training, pending when all k partitions had been selected for testing. The error rate recorded from each process was added up with the mean the mean error-rate recorded

d. Performance Evaluation of Model Validation Process

In the course of evaluating the predictive model, the models' performance was quantified using some metrics. Basically, four (4) parameters must be known from the model testing of predictions made by the classifier during model testing. These parameters are: true positive (TP), true negative (TN), false positive (FP) and false negative (FP). TPs refers to the accurate prediction of positive cases, TNs refers to the accurate prediction of negative cases, and FPs indicates the negative cases predicted as positives while FNs indicates the positive cases predicted as negatives. The results were then obtainable on confusion matrix which is a 4 x 4 matrix table owing to the four (4) labels of the output class (see Figure 1). Correct classifications were plotted along the diagonal from the north-west position for first predicted as first (A), 2-1 predicted as 2-1 (F), followed by 2-2 predicted as 2-2 (K) and third predicted as third (P) on the south-east corner (also called true positives and negatives). The incorrect classifications were plotted in the remaining cells of the confusion matrix (also called false positives). Also, the actual first cases are A+B+C+D, actual 2-1 cases are E+F+G+H, actual 2-2 cases are I+J+K+L while actual third are M+N+O+P and the predicted first are A+E+I+M, 2-1 are B+F+J+N, predicted 2-2 are C+G+K+O and predicted third are D+H+L+P.

The developed model was validated with a number of performance metrics based on the values of *A – P* in the confusion matrix for each predictive model. They are presented as follows.

a. Accuracy: the total number of correct classification.

$$Accuracy = \frac{A + F + K + P}{total\_cases} \tag{5}$$

b. TP rate (recall/sensitivity): the amount of actual cases accurately classified.

$$TP_{first} = \frac{A}{A + B + C + D} \tag{6}$$

FIRST	2-1	2-2	THIRD	
A	**B**	**C**	**D**	FIRST
E	**F**	**G**	**H**	2-1
I	**J**	**K**	**L**	2-2
M	**N**	**O**	**P**	THIRD

Figure 1: Confusion Matrix

$$TP_{2-1} = \frac{F}{E + F + G + H} \tag{7}$$

$$TP_{2-2} = \frac{K}{I + J + K + L} \tag{8}$$

$$TP_{third} = \frac{P}{M + N + O + P} \tag{9}$$

c. FP (false alarm/1-specificity): the amount of negative cases inaccurately classified as positive.

$$FP_{first} = \frac{E + I + M}{actual_{2-1} + actual_{2-2} + actual_{third}} \tag{10}$$

$$FP_{2-1} = \frac{B + J + N}{actual_{first} + actual_{2-2} + actual_{third}} \tag{11}$$

$$FP_{2-2} = \frac{C + G + O}{actual_{first} + actual_{2-1} + actual_{third}} \tag{12}$$

$$FP_{third} = \frac{D + H + L}{actual_{first} + actual_{2-1} + actual_{2-2}} \tag{13}$$

d. Precision: the proportion of predictions that are correct.

$$Precision_{first} = \frac{A}{A + E + I + M} \tag{14}$$

$$Precision_{2-1} = \frac{F}{B + F + J + N} \tag{15}$$

$$Precision_{2-2} = \frac{K}{C + G + K + O} \tag{16}$$

$$Precision_{third} = \frac{P}{D + H + L + P} \tag{17}$$

Using the aforementioned performance metrics, the performance of the predictive model for the classification of student's performance was evaluated by validation, using a dataset. The TP rate and precision lie within the interval [0, 1], accuracy within the interval of [0, 100] % while the FP rate lies within an interval of [0, 1]. The closer the accuracy is to 100% the better the model, the closer the value of the TP rate and precision is to 1 the better. While the closer the value of FP rate is to 0, the better. Therefore, the evaluation of an effective model has a high TP/Precision rates and a low FP rates.

4. RESULTS AND DISCUSSION

Table 1 shows a description of the variables that were discretized and the nominal variables to which they were converted to for clarity of model complexity. Afterwards, the pre-processed dataset was saved in the acceptable format (attribute relation file format (.arff)) for the machine learning simulation environment.

Table 1: Student's performance data showing the discretized numeric variables

Name of Variable	Raw Label	Interval	Discretized Value
UTME Score	Numeric (0 – 400)	1 to 100 101 to 200 201 to 300 301 to 400	1 – 100 101 – 200 201 – 300 301 – 400
Age	Numeric (in years)	Less than 18 years 18 years and above	Below-18 18-above
SSCE Score	Numeric (0 – 30)	1 to 10 1 to 20 21 to 30	1 – 10 11 – 20 21 – 30
100 Level Grade	Numeric (0.0 – 5.0)	Below 2.00 2.00 - 2.50	Pass Third

		2.50 – 2.49 3.50 – 4.49 4.50 – 5.00	Two-2 Two-1 First

Table 2 gives the narrative of the number of students with their individual classification of student's performance from the record of 47 student chosen for formulating and validating the model which were saved in the Student-Train-Data.arff file. The table shows that of the 100 students used; 2% had first class, 22% had second class upper, 61% had second class lower while 15% had third class degree by the time of graduation. The results showed that majority of the students had second class lower degrees amounting for about 70% of the student population selected for this study.

Table 2: Distribution of student performance among historical dataset

Student's Performance	Frequency	Percentage (%)
First	2	2.0
Two – 1	22	22.0
Two – 2	61	61.0
Third	15	15.0
Total	100	100.0

Results of Model Formulation and Simulation

Following theidentification of the factors that are associated with student performance, the next phase is model formulation using the aforementioned decision trees algorithms available in the WEKA environment. The 10-fold cross validation technique was used in evaluating the performance of the developed predictive model for student performance using the historical dataset used for training the model. This process was performed for both decision trees algorithm used with their respective performance compared for the most effective.

a. Results of model formulation and simulation using the ID3 algorithm

The results of the formulation of the predictive model using the ID3 decision trees algorithm showed that a limited number of variables were the most important classification factors. Identified variables in the order of their significance are:

- 100 level grade;
- Subject grades in core subjects such as physics, mathematics, and English;
- UTME score;
- Age at admission; and
- Student’s gender.

The predictive model was formulated based on ID3 identified variables, using the results of the simulation with the C4.5 algorithm in WEKA simulation environment. The ID3 was used to formulate a tree that was adopted in deducing the set of rules used for the classification of student’s performance. Following the simulation of the predictive using the ID3 and C4.5 decision trees algorithm, after 10-fold cross validation, the result of the performance evaluation of the model was recorded. The confusion matrix used to interpret TP and TN alongside the FP and FN of the validation result is shown in Figure 2 and Figure 3 respectively. The results showed that out of the figure 2 actual first classes, all were correctly classified, out of the 22 actual two-1, all were correct classified, out of the 61 two-2, all were correctly classified and out of the 15 third class cases, all were correctly classified. Hence, all 100 instances in the dataset were correctly classifier by the ID3 decision trees classifier meaning 100% accuracy.

FIRST	2-1	2-2	THIRD	
2	0	0	0	FIRST
0	2	0	0	2-1
0	0	61	0	2-2
0	0	0	15	THIRD

Figure 2: Confusion matrix of performance evaluation using ID3

b. Model Formulation and Simulation in C4.5 algorithm

The C4.5 algorithm was also used to implement predictive model in the simulation environment. From the result, the algorithm could not identify the variables that were the most important factors of student’s performance.

The confusion matrix in figure 3 was used to evaluate the performance of the predictive model for classification of student’s performance. The results further showed that using the C4.5 decision trees algorithm to formulate the model for the classification of student's performance, all 61 two-2 cases were correctly classified while all 2 first class cases, 22 two-1 cases and 15 third class cases were misclassified as two-2 cases. Therefore, 61 out of the 100 students' instances were correctly classified by the C4.5 decision trees classifier for the model development owing for an accuracy of 61%.

FIRST	2-1	2-2	THIRD	
0	0	2	0	FIRST
0	0	22	0	2-1
0	0	61	0	2-2
0	0	15	0	THIRD

Figure 3: Confusion matrix of performance evaluation using C4.5

c. Discussion of results

The result of the performance evaluation of the machine learning algorithms are presented in Table 3 which presents the average values of each performance evaluation metrics considered for this study. For the ID3 decision trees algorithm based on the results presented in the confusion matrix presented in figure 3. The results showed that the TP rate which gave a description of the proportion of actual cases that was correctly predicted was 1 which implied that 100% of the actual cases were correctly predicted; the FP rate was 0 which implied that 0% of actual cases were not accurately classified while the precision was 1 which implied that 100% of the predictions made by the classifier were correct.

For the C4.5 decision trees algorithm based on the results presented in the confusion matrix presented in figure 3. The results showed that the TP rate was 1 for two-2 but 0 for first/two-1/third which implied that 100% and 0% of the actual two-2 cases and first/two-1/third cases respectively were correctly predicted; the FP rate was 1 for two-2 but 0 for first/two-1/third which implied that 100% and 0% of actual cases were misclassified while the precision which gave a description of the proportion of predictions that were correctly classified was 0.61 for two-2 but 0 for first/two-1/third which implied that 61% and 0% of the predictions made by the classifier were correct.

From the study, it was discovered that ID3 decision trees algorithm was able to classify the performance of students by graduation better than the C4.5 decision trees algorithm. The ID3 decision trees algorithm was able to accurately classify all cases of students with a value of 100% showing that it had the capacity to identify the complex patterns that existed within the dataset than the C4.5 decision trees algorithm. The variables identified by the ID3 decision trees algorithm can also be given very close attention and observed in order to better understand the students' performance and proper monitoring.

Table 3: Performance Evaluation Result Summary for the machine learning algorithms selected

Algorithm Used	Correct Classification	Accuracy (%)	TP Rate	FP Rate	Precision
ID3 Decision Tree Algorithm	61	61.0	1.000	0.000	1.000
C4.5Decisio Tree Algorithm	100	100.0	0.333	0.333	0.203

5. CONCLUSION

The study proposed a predictive model for student performance using relevant classification factors selected from a predefined set of factors of student performance. The ID3 decision trees algorithms identified few factors which were more related in determining the performance of students. The predictive model was formulated using the variables identified by ID3 decision trees for this study and the performance evaluation of both models showed that the model developed using the ID3 decision trees algorithm was a better model. Unlike the C4.5 decision trees algorithm which could not clearly state the relevant attributes, ID3 was able to identify the important variables and used them in developing the decision trees for students' performance classification. The results of the study revealed the variables that were identified by the ID3 decision trees algorithm as relevant for identifying the classification of student's performance. The ID3 algorithm was observed to show a better accuracy compared to that of the C4.5 algorithm using the training dataset presented in the study.

6. REFERENCES

[1] B. Baradwaj and S. Pal, 2011. Mining Educational Data to Analyze Students' Performance. (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 2, No. 6, 2011.

[2] K. Adhatrao, A. Gaykar A, A. Dhawan, R. Jha. and V. Honrao. Predicting Students' Performance Using ID3 and C4.5 Classification Algorithms. International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.3, No.5, September 2013

[3] Z. Xiaoliang,., W. Jian, Y. Hongcan, and W. Shangzhuo, Research and Application of the improved Algorithm C4.5 on Decision Tree. International Conference on Test and

Measurement (ICTM), Vol. 2, 2009.

[4] A. Ahmed,. and I. Elaraby, 2014. Data Mining: A prediction for Student's Performance Using Classification Method. World Journal of Computer Application and Technology, 2(2), pp.43-47.

[5] U. Pandeyand S. Pal. Data Mining: A Prediction of performer or underperformer using Classification. (IJCSIT) International Journal of Computer Science and Information Technology, Vol. 2(2), 2011

[6.] B. Bhardwaj and S. Pal. Data Mining: A prediction for performance improvement using classification. (IJCSIS) International Journal of Computer Science and Information Security, Vol. 9, No. 4, 2012.

[7] I. Livieris and P. Pintela. An improved spectral conjugate gradient neural network training algorithm. *International Journal on Artificial Intelligence Tools*, 21(1), 2012.

[8] S. Kotsiantis, C. Pierrakeas and P. Pintelas. Predicting students' performance in distance learning using machine learning techniques, *Journal of Applied Artificial Intelligence*, 18(5), p.p. 411-426, 2004.

[9] C. Moucary, M. Khair and W. Zakhem, 2011. Improving student's performance using data clustering and neural networks in foreign-language based higher education. The Research Bulletin of Jordan ACM, 2(3), pp 27-34

[10] Y. Hongsuk, H. Jung. and S. Bae, 2017. Deep Neural Networks for traffic flow prediction. In Big Data and Smart aComputing (BigComp), IEEE International Conference on (pp. 328-331). IEEE.

[11] S. Yadav, and S. Pal. Data mining: A prediction for performance improvement of engineering students using classification. World of Computer Science and Information Technology Journal (WCSIT). (ISSN: 2221-0741), Vol. 2, No. 2, 51-56, 2012