Performance Evaluation of Classification Algorithms on Academic Performance of Postgraduate Students

O.A. Okunlola
Department of Computer Science
University of Ibadan, Nigeria

A.K. Ojo
Department of Computer Science
University of Ibadan, Nigeria

ABSTRACT
Educational data mininghas contributed to enhancing student academic performance by way of enabling stakeholders in academic institutions to have a pre-knowledge of the risks and dangers ahead and how to mitigate them. Prediction algorithms perform differently on dataset, and so, the need to develop models using different prediction algorithms and evaluating the result of such predictions is very important in order to be sure the best algorithm for a particular dataset is used.This work employed four classifiers: K-Nearest-Neighbour, Neural Network, Naïve Bayes and Decision Tree to model and, evaluated their models to know the performance of each on the target dataset. Their results were evaluated based on the various performance metrics. The results showed that Decision Tree had the highest accuracy on the dataset with test accuracy of 48.25% and therefore is the most suitable out of the four classifiers for performing prediction modelling on the dataset. Naïve Bayes is the second-best prediction model that can be used for predicting academic performance with an accuracy of 36.25%., followed by Neural Network with accuracy of 32.5 % and then K-Nearest Neighbour with accuracy of 32.5% but with lower precision, recall and area under Receiver Operating Characteristic curve.

Keywords
Decision Tree, Educational Data mining, K-Nearest Neighbour, Neural Network, Naïve Bayes

1. INTRODUCTION
EducationalData Mining (EDM) is the application of data mining techniques in educational framework and has attracted significant interest of many researchers in recent years spanning various fields due to its potentials for education. The quality of education needs to be improved and EDM is a tool for this improvement and hence the reason many educational institutes' need data mining for their strategy and plans. One of the educational problems that are solved with data mining is the prediction of students' academic performances and has helped in extracting several useful knowledge and insights, very crucial at all levels such that all stakeholders in academic institutions can employ strategies that can, to a large extent, guarantee better performance of students, educational process and the institution's rating in the perceived competition [1]. This students' performance prediction has become a major research area of EDM. Several research works have been carried out in this area using different computational strategies on varieties of data attributes. These computational strategies on the same data set have produced different results and hence, the suitability of one strategy over the other, based on the output recorded [2].

There are different classification algorithms such as Decision Tree, Neural Network, Naïve Baiyes, K-Nearest Neighbour, Random Forest, Support Vector Machines, AdaBoost and PART. Some of these algorithms have been used by many researchers for predictive purpose especially on educational data and in many cases - for a more refined and accurate result for informed decision-making processes.

The question of why a specific machine learning technique or statistical measure is more suitable than others is seldomly addressed but is more likely to reveal additional, yet crucial, information about the dataset under consideration and may even indicate which variables have higher impact on the predicted outcome [3].

In this work, some of these classification algorithms were used for mining the dataset of the Postgraduate Students of the Department of Computer Science, University of Ibadan. The algorithms or classifiers are Naïve Bayes, Neural Network, K-Nearest Neighbour (KNN) and Decision Tree. The data mining tool used for this work was the Waikato Environment for Knowledge Analysis (WEKA).

2. LITERATURE REVIEW
The amount of data in the world and in our lives seems ever-increasing—and there is no end in sight. As the volume of data increases, inexorably, the proportion of it that people understand decreases alarmingly. Lying hidden in all this data is information— potentially useful information—that is rarely made explicit or taken advantage of. As the flood of data swells and machines that can undertake the searching become commonplace, the opportunities for data mining increase. As the world grows in complexity, overwhelming us with the data it generates, data mining becomes our only hope for elucidating hidden patterns. Intelligently analysed data is a valuable resource. It can lead to new insights, and, in commercial settings, to competitive advantages. Data mining is about solving problems by analysing data already present in databases [4].

[5]carried out a research work using KNN and Naive Bayesian on educational dataset for secondary schools in Gaza Strip collected from the Ministry of Education, Gaza Strip. Eight (8) attributes out of 14 were selected from 500 records after data pre-processing. Accuracy, recall and precision were used as the performance metrics and were measured for the two algorithms. Naïve Bayesian had the highest accuracy of 93.17% which is better than KNN which has 63.45%. So, for this particular dataset, Naïve Bayes is the preferred classification algorithm out of the two used.

[6]used Decision Tree and Neural Network algorithms to mine the data of 800 students collected from the National Centre for Education Statistics, a principle federal organization that analyses the insights and compares the educational data throughout the Colleges and Universities in the United States of America. The features were extracted and minedusing the R programming language with its attendant packages. The data processing phase was implemented under the condition that all the students' records had to be randomized first so the model could avoid gathering students with similar background. The dataset was then split into training set and testing set with a ratio

of 7:3. Decision Tree gave an accuracy of 92.8% while Neural Network produced an accuracy of 97.1 using one hidden layer. Therefore, Neural network was the preferred algorithm for this dataset.

Similarly, [7] used J48, Bayesnet, PART and Random Forest classification algorithms to predict the final semester percentage of students from three different colleges in Assam, India. The total number of records were 300 with 24 attributes. Two attributes were dropped at the data cleaning stage and 12 influential attributes were selected using feature selection. Random Forest has more correctly classified instances than other classification methods. It produced an accuracy percentage of 99% while PART produced 74.33%, J48 produced 73% and BayesNet with 65.33%. Random Forest also outperformed others with minimum errors in terms of Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Relative Absolute Error (RAE) and Root Relative Squared Error (RRSE) when compared with other classifiers.

Kappa statistic value is 0.9859 for Random Forest which shows that the model is statistically significant and therefore the preferred algorithm for the data set based on the overall performance Criteria.

While some Classification Algorithms achieve a superior result with a particular dataset, the performance of such classifiers might be discouraging when applied on other dataset as seen in [8],[9],[10] and [11].

3. METHODOLOGY

For this study, the dataset used was extracted from the Computer Science department, University of Ibadan Postgraduate College within the period of 2008 and 2014 sessions; the data was carefully reviewed, pre-processed and organized in a new excel flat file before converting to WEKA tool file format (ARFF). Attributes such as Surname, Other names, University Attended, Course studied, Class of Degree, Date of Birth, matric number,State of Origin, Local Government Area, Session resumed, other sessions registered, Sponsor, Gender, Master of Science (MSc) Grade, Ordinary Level (O/L) Mathematics, O/L English, O/L Physics, O/L Chemistry, Sponsor, Marital Status, Hall of Residence were captured in the raw data.

Data cleaning was done to remove attributes with no impact on the predicting process such as Surname, Other Names, Matric number, State of origin, Local Government Area, Other sessions registered and Hall of Residence. 1093 records were retrieved and cleaned, remaining 274 which was then used for the work.

The Attributes were evaluated using the Attribute Evaluators in WEKA. Two Evaluators used were ClassifierAttributeEval with Ranker Search Method and GainRatioAttributeEval with Ranker Search methods. The two Attribute Evaluators selected the 12 independent Attributes listed in Table 1.

3.1 Model Evaluation and Error Measurement

These were the evaluation criteria for the model. Performance measures were derived from confusion matrix. These are metrics used to compute model correctness and predictive suitability.

3.1.1 Sensitivity (Recall or True Positive rate)

Recall is the number of correct classifications divided by the total number of positive classifications.

$$Recall = \frac{TruePositive\ \ (TP)}{(TruePositive\ +FalseNegative\ )} \qquad (1)$$

divided by total number of positive classifications.

3.1.2 Precision

The number of correct positive classifications

$$Precision = \frac{TruePositive\ \ (TP)}{(TruePositive\ +FalsePositive\ )} \qquad (2)$$

3.1.3 Receiver Operating Characteristic (ROC) curve

This measures model accuracy in a weighted sort way.

The greater the area under the curve the better.

3.1.4 Mean Absolute Error (MAE)

This estimates how far the predictions or forecasts differ from the actual values.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|x_i - x| \qquad (3)$$

where n = the number of errors, $|x_i - x|$= the absolute errors.

3.1.5 Root Mean Square Error (RMSE)

This is an evaluator of the differences between the predictor values and the actual observed values.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(X_{obsi} - X_{modeli}\ )^2}{n}} \qquad (4)$$

where X_{obs} are observed values and X_{model} are modeled values at time/place i

4. RESULTS AND FINDINGS

In this section, the results from the study were discussed. The following metrics were employed to evaluate the performance of each of the Algorithms on the dataset; Accuracy, Precision, Recall, F-Measure, MAE, RMSE, Receiver Operation Curve (ROC) and Kappa Statistic. Tables 2 to 8 and Figures 1 to 7depict the performance of the four algorithms.

4.1 K-Nearest Neighbour (K-NN)

Table 2. Precision, Recall and F-Measure for K-NN

MSc Grade	Precision	Recall	F-measure
PhD	0.452	0.4	0.424
Mphil/PhD	0.25	0.24	0.245
Mphil	0.176	0.3	0.222
Terminal	1	0.333	0.5
Non-Graduating (NG)	0	0	0

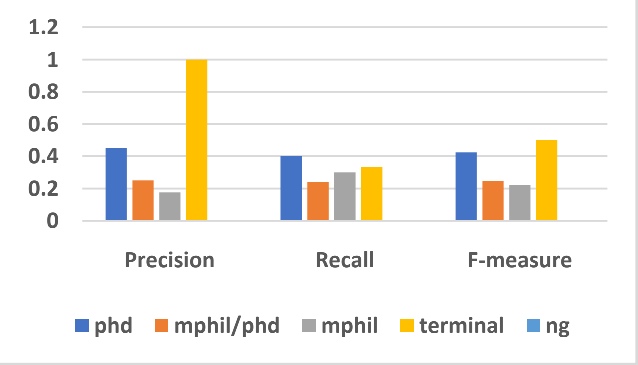

Figure 1. Chart Visualization of Precision, Recall and F-measure for K-NN

4.2 Neural Network (NN)

Table 3Precision, Recall and F-Measure for NN

MSc Grade	Precision	Recall	F-measure
PhD	0.528	0.543	0.535
Mphil/PhD	0.29	0.36	0.321
Mphil	0.125	0.1	0.111
Terminal	0	0	0
NG	0	0	0

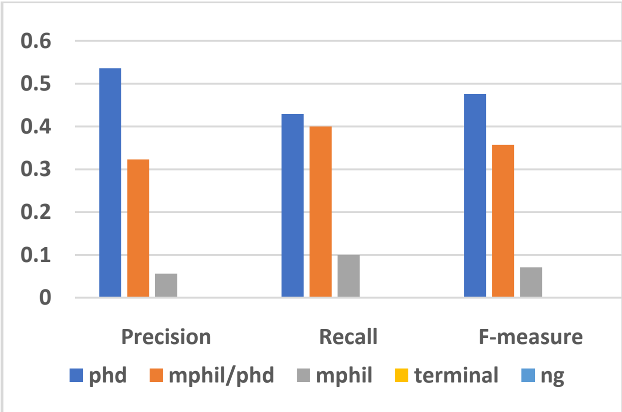

Figure2. Bar Chart Visualization of Precision, Recall and F-measure for Neural Network.

4.3 Naïve Bayes

Table 4.Precision, Recall and F-Measure for Naïve Bayes

MSc rade	Precision	Recall	F-measure
PhD	0.528	0.543	0.535
Mphil/PhD	0.29	0.36	0.321
Mphil	0.125	0.1	0.111
Terminal	0	0	0
NG	0	0	0

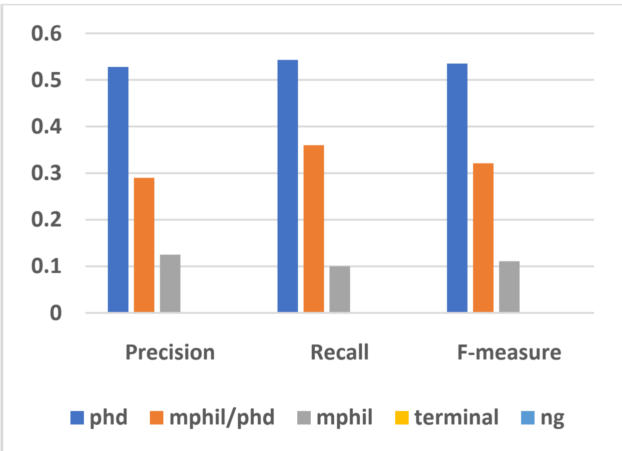

Figure3. Chart Visualization of Precision, Recall and F-measure for Naïve Bayes

4.4 Decision Tree

Table 5. Precision, Recall and F-Measure for Decision Tree

MSc rade	Precision	Recall	F-easure
PhD	0.61	0.714	0.658
Mphil/PhD	0.406	0.52	0.456
Mphil	0.143	0.1	0.118
Terminal		0	
NG		0.488	

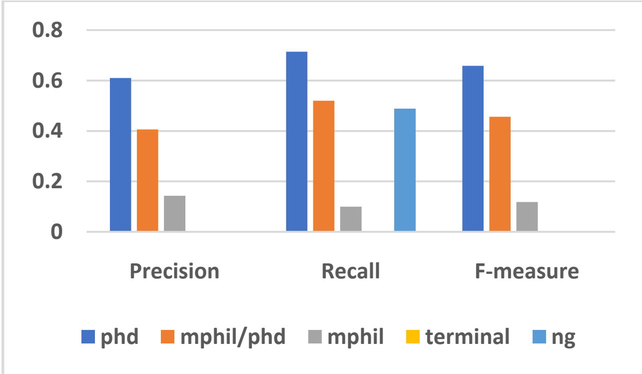

Figure 4. Chart Visualization of Precision, Recall and F-measure for Decision Tree

Table 6. ROC for K-NN, NN, Naïve Bayes and Decision Tree

MSc rade	KNN	NN	Naïve Bayes	Decision Tree
PhD	0.466	0.551	0.665	0.706
Mphil/PhD	0.489	0.508	0.515	0.543
Mphil	0.521	0.43	0.613	0.609
Terminal	0.714	0.651	0.753	0.84
NG	0.557	0.027	0.937	0.38

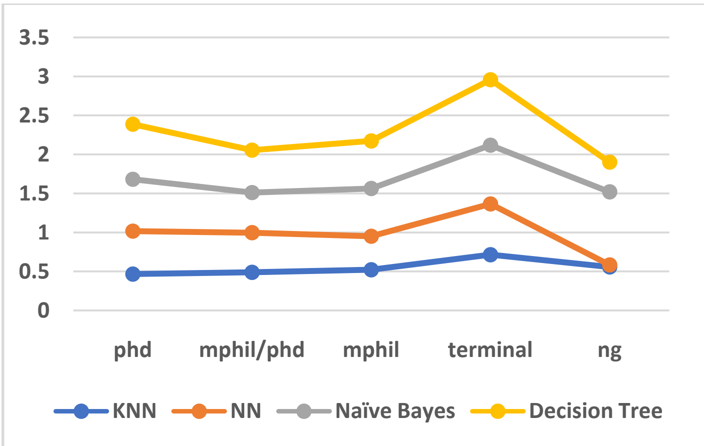

Figure5. ROC for K-NN, NN, Naïve Bayes and Decision Tree.

Table 7. MAE, RMSE and Kappa Statistic for K-NN, NN, Naïve Bayes and Decision Tree

Performance Measures	KNN	Neural Network	Naïve Bayes	Decision Tree
Kappa Statistic	0.0428	0.0318	0.0449	0.199
MAE	0.278	0.2742	0.2546	0.259
RMSE	0.5084	0.4687	0.3971	0.3694

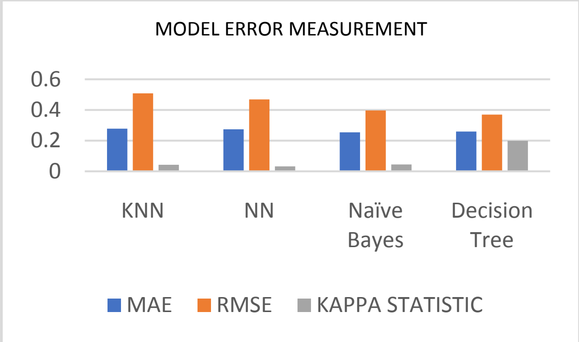

Figure 6. Chart Visualization of MAE, RMSE and Kappa Statistic

4.5 Model Accuracy

Table 8. Prediction Accuracy of the four classifiers/Algorithms

Prediction Model	***Test Accuracy***
K-Nearest Neighbour (K-NN)	32.5
Neural Network	32.5
Naïve Bayes	36.25
Decision Tree (Random Forest)	48.75

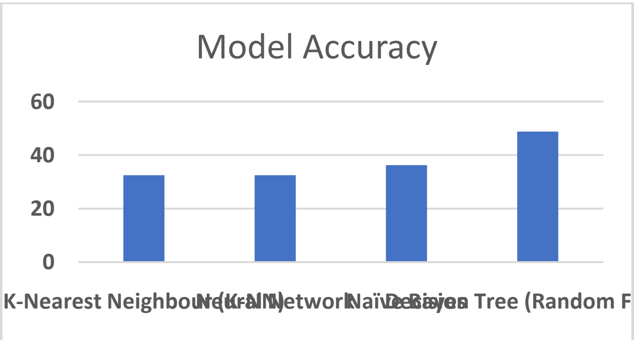

Figure 7. Chart Visualization of Accuracy for the four classifiers/Algorithms

5. CONCLUSION AND RECOMMENDATION

From the results in the section above, Decision Tree (Random Forest) has the best accuracy being the highest of all the four models that can be used for predicting the academic performance of the Masters'students in the Department of Computer Science, University of Ibadan. It has the lowest MAE, RMSE compared to other prediction models. Furthermore, the ROC curve for Random Forest has the largest area of coverage, an indication that its performance is preferred to other models.

K-NN has the least performance considering all the results in the result presentation section based on the performance parameters for the evaluation. In summary, Decision tree has the best performance followed by Naïve Bayes, then Neural Network and lastly K-NN - based on this work. It is least expected to use K-NN for this dataset. The dataset is not balanced and that accounted for the low accuracies of all the classifiers.

Due to the imbalanced nature of the dataset, appropriate sampling techniques can be applied to produce a modified dataset which in turn will guarantee a better prediction accuracy of the classifiers on the dataset. Furthermore, increasing the size of the dataset can also bring improvement to the performance of the classifiers on the dataset.

6. ACKNOWLEDGMENTS

The authors would like to thank the Postgraduate School, University of Ibadan, Nigeria for their assistance in getting the dataset used for this research work.

7. REFERENCES

[1] M. A. Umar, "Student academic performance prediction using artificial neural networks: a case study.," International Journal of Computer Application, vol. 178, no. 48, pp. 24-29, September 2019.

[2] Cortez, P. and Silva, A., "Using data mining to predict secondary school student performance.," in Proceedings of 5th Annual Future Business Technology Conference, Porto, 2008.

[3] M. M. Ashenafi, "A comparative analysis of selected studies in student performance prediction.," International Journal of Data Mining & Knowledge Management Process, vol. 7, no. 4, pp. 17-32, 2017.

[4] Han, J.W., Kamber, M. and Pei, J. , "Classification: Basic Concepts.," in Data Mining: Concepts and Techniques, Burlington, , 2012.

[5] Amra, I. A. A. and Maghari, A. Y. A. , "Students' performance prediction using KNN and naïve Bayesian," in 8th International Conference on Information Technology, 2017.

[6] J. Feng, "Predicting students' academic performance with decision tree and neural network.," 2019.

[7] Hussain, S., Dahan, N. A. and Najoua, A. , " Educational data mining and analysis of students' academic performance using WEKA.," Indonesian Journal of Electrical Engineering and Computer Science., vol. 9, no. 2, pp. 447-459, 2018.

[8] Sharma, R., Hussung, R., Keil, A., Friederich, F., Fromenteze, T., Khalily, M., Deka, B., Fusco, V., & Yurduseven, O. (2022). Performance Analysis of Classification Algorithms for Millimeter-Wave Imaging. Paper presented at European Conference on Antennas and Propagation, Madrid, Spain.

[9] E. Venkatesan "Performance Analysis of classification Algorithms using Clinical Dataset". Journal of Information and Computational Science, Vol.9 (9), PP.395-401, 2019.

[10] N.N Khanom, F. Nihar, S.S Hassan, L Islam (2020). "Performance Analysis of Algorithms on Different Types of Health Related Datasets". Journal of Physics: Conference Series 1577 012051.

[11] Bayu Adhi Tama and Sunghoon Lim (2020). " A Comparative Performance Evaluation of Classification Algorithms for Clinical Decision Support Systems". Journal of Mathematics, Vol. 8(10), 2020.