MEDWIN PUBLISHERS
Committed to Create Value for Researchers

Physical Science & Biophysics Journal
ISSN: 2641-9165

A Machine Learning Based Framework for Predicting Student's Academic Performance

Okereke GE*, Mamah CH, Ukekwe EC and Nwagwu HC

Department of Computer Science, University of Nigeria, Nigeria

***Corresponding author:** George Okereke E, Department of Computer Science, University of Nigeria, Nsukka, Enugu, Nigeria, Tel: +2348037784613; Email: george.okereke@unn.edu.ng

Research Article

Volume 4 Issue 2

Received Date: June 30, 2020

Published Date: July 13, 2020

Abstract

Educational Data Mining (EDM) as a new technology has become a field of research as a result of continuous improvement in numerous approaches in statistics, exploring hidden data in educational environment. An application associated with EDM is a predictive system that can be deployed in early prediction of student academic performance. The importance is to identify poor performers and provide necessary remediation to avoid school drop outs and also encourage high performers. This paper explores certain features of a population of 103 first year students majoring in Computer Science at University of Nigeria, Nsukka. Due to the high number of predicting variables determining student's performance, it is necessary to apply feature selection mechanism using rapid miner to filter these variables. Decision tree, a Machine Learning Algorithm (MLA) was used in training and testing. It was observed that the accuracy is dependent on the datasets on which the model is trained. Two dissimilar datasets achieve different accuracy on the same algorithm. This leads us to conclude that the greatest factor in achieving higher accuracy is the type of datasets not actually the type of classification algorithm.

Keywords: Machine learning; Educational data mining; Training dataset; Students' academic performance; Predictor variables

Introduction

Students' performance in the institutions have been a common concern to the education management, parents, governments and other stakeholders in the education system because of the great importance education has on the development of any nation. It is in lieu of this obvious reason that most academic institutions especially in western world, saw the urgent need to monitor the performance of both lecturers and students.

In order to achieve this, they leveraged on the technological advancement of data mining to extract information from large educational data repositories [1]. When mining is introduced in educational environment, it is regarded as Educational Data Mining (EDM), which undoubtedly has become the useful way of discovering hidden information from very large educational databases. Such information before the innovation of data mining is not utilized in decision making process. Data mining tools have brought in, its usefulness in analyzing student's trends and behaviours towards education thereby removing the use of intuition by decision makers in decision making process [1,2]. There is need for institution managers to expedite action in utilizing the vast amount of dataset of learners and other stakeholders in academic environment for multipurpose decision making. The advancement of web technology has helped in reshaping the management of data in the current educational system in many universities, increasing the actual amount of data about students, teachers and their interactions with learning and educational systems [3].

Prediction of student's academic results has become a necessity in the universities all over the world. This is because of the great need to identify poor performers early enough and offer them remediation to avoid expulsion and secondly to identify those who would perform well in

the end of semester examination. The students' academic activities in present day educational system in Nigerian universities are determined by the combination of pre-exam assessment which includes, attendance, test and end of semester examination mark. There is always a bench mark that each student is expected to get in both assessments in order to pass the course. Ironically, the vital predictor variables inherent in those assessments and exams are not utilized to predict student's performance in future semester examinations.

Several important factors do affect student's academic performance, these include, past examination records, financial status, family upbringing, performance in class test, attendance to lectures, level of understanding of the courses, the teaching capacity and general commitment to private studies etc. Using these factors we can successfully apply supervised MLA (Decision Tree) in this work which can more accurately predict student results.

In the past, researchers have used various classes of MLAs such as support vector machines (SVM), Artificial Neural Network, Bayesian network etc. in trying to predict student's academic performance [4]. Most of the previous studies used either the student's background check or one semester academic performance as the variable for predicting the student's next semester academic performance. We intend to use hybrid features from demographic data and psychometric data obtained from an online survey which all students must take before the end of semester examination. There is also need to use class attendance records, previous academic records and other student's records in educational databases. There is high possibility of this system with more predictor variables outclassing the existing ones with fewer variables in accurate predictions of the semester results. This current work applied many predicator variables in carrying out prediction of student's performance.

Related Works

Using Kernel K-Means (unsupervised ML) and Smooth Support Vector Machine (supervised ML) in Predicting Student's Academic Performance.

This study applied two distinct ML techniques in predicting students results using the same predictor variables extracted from some psychometric factors like, study habit, interest, family support etc. Both models were based on kernel methods which have grown in popularity recently in the field of data mining application. It has the capability of processing huge amount of data, especially data that are highly dimensional, nonlinear and non-separable. The data were collected from undergraduate student's academic databases in University of Malaysia; the researcher was able to generate some psychometric factors already enumerated as training sets for the models. The support vector machine is a flexible algorithm and so its flexibility aided the researcher to successfully implement both algorithms in kernel which can transform non-separable data to separable by addition of more dimension to it. The result obtained from the study proved that there is positive correlation between predictor variables applied and the target class. Predictor variables were very significant in determining the percentage of the outcome which was 52.2% (R2=0.522). In this work, it was observed that the cluster model of students was entirely based on their performances, with every student in cluster labeled with their performance index, in order to show their present situation with the models. The prediction accuracy of the models showed lowest accuracy of 61% (R2 = 0.61) in predicting "Good" performance index and the highest accuracy 93.67% (R2 = 0.9367) in predicting "Poor" performance index [5].

The study claimed that kernel method does what data mining technique can do, on educational databases and that the knowledge gained in the study can be used in monitoring students' academic progression every semester. In contrast, this work applied machine learning technique (Decision tree) for predicting student's performance using various sensitive predictor variables extracted from student's records and questionnaires. Also the above study based their most important predictor variables on psychometric factors in predicting student's next semester academic performance. Student's previous exam scores or assignments and other factors revolving around academics were not considered in the models. The exclusion of essential predictor variables in academics in the work ex rayed has created a gap, which this work is set to address in order to achieve a higher accuracy. The present work applies extensive data from academics, demographic and psychometric data. Great emphasis was on student's disposition in a particular course ie the level of understanding in a particular course. A statistical study of student's interest and understanding in academic courses proved that not all students who were interested in a particular course passed. Interest can be there but understanding of a particular course is a pivot variable that determined a pass or fail.

Predicting Student Academic Achievement by Using the Decision Tree and Neural Network Techniques

This study analysed factors affecting academic achievement which have impact on the prediction of student's

Copyright© Okereke GE, et al.

academic performance. The researcher used WEKA software, an open source data mining tool to analyse attributes for predicting undergraduate academic performance in an international program. The datasets which included data from researcher's questionnaires were obtained from 1,600 student records registered between year 2001 and 2011 in the university in Thailand. The attributes extracted from the datasets were 22 and the data were pre-processed with attribute importance analysis. The dataset was further divided into training set, validation set and testing set. The researcher training sets which were applied to two classifiers allowed the system to observe the relationships existing between the input variables and the output labels. To understand the performance of the trained models with the data, validation sets and testing sets were applied. The main aim of the validation set was to detect over fitting and estimate prediction errors for the models while the test sets determine the overall performance, what the models have achieved. The researcher in evaluating the prediction accuracy of both classifiers did an experimental comparison of both models where he found that decision tree classifier achieved higher accuracy of 85.188% than that of neural network classifier with 71.313%. They further analyzed the important factors for grouping students in this manner:

Firstly, it was observed that the students with a high risk of low academic performance were those who never studied English language courses while those who had a good grasp of English language before entering university had high performance. The study found out that the high performers were single, work few hours per semester, and registered for 12-15 credits per semester.

In the second consideration, the students having risk of low academic achievements are the students who have need for additional study of English language courses. It was observed that many of them were either married or divorced, work moderate or high number of hours per semester and registered for either less than 12 credits per semester (students are not allowed to register more than 12 credits if the CGPA is low) or more than 15 credits per semester (students are allowed to register more than 15 credits if the CGPA is high) [6].

This study has some limitations which further researches like this intend to address. There seem to be concentration on international students registered for undergraduate course. It is believed that further research should look into other programs of the university.

Secondly, the researcher concentrated on international study and made English proficiency the major factor in predicting results of possible poor performers. It did not put into consideration student's previous exam scores even though he applied many other demographic and psychometric factors. The current system will not be looking at English proficiency of the student and would not concentrate on international study in the university but will consider many variables ranging from, student's financial background, gender, assignment scores with attendance (CA), and some other factors obtained from compulsory questionnaires.

Methodology

The datasets used were obtained from first year computer science students of University of Nigeria, Nsukka on two courses Cos101 and STA172. The choice of these courses came from the researchers understanding of the impact of the courses in computer science field. The gender, martial-status and CA (which is a combination of pre-exam test and mark allotted for attendance) were extracted from the results of the courses. Also other factors which played important role in determining the prediction accuracy of student's results were obtained through a compulsory online survey which students take before semester exam. From the student's answers, we obtain demographic factors like income- level, religious- involvement, and also psychometric factors like, level of understanding of courses, study time or habit, social life in school. The dynamic nature in student's disposition to studies makes the researchers to propose that the survey should be taken every semester after the CA has been conducted and results known. It has been observed that a student can improve in their level of understanding every semester or the factors that helped a student to pass may be absent in the next semester. The data from survey and the CA were entered in Microsoft Excel and using Rapid Miner, it was preprocessed and used for both training and testing a machine learning algorithm called Decision Tree. The rapid miner helps in cleaning the datasets in order to remove irrelevant features that cannot impact on the prediction. These two courses created two different models for prediction, while the COS101 model will be used in predicting all COS courses in computer science department, the STA model would also be used in predicting all STA related courses in the department.

ID3 decision tree uses the information gained to determine the root node of the tree and other paths of the tree. These are usually done behind the scene by the rapid miner. The rapid miner does not take numeric values, therefore data containing numeric values are classified as low, average, and high, for instance; in CA a student score>=20 is high, while the score>=10 is average, else it is low .The models which can predict all COS. courses in the department and all STA courses, generates two different trees. Through the logical part of the tree our system can be designed and implemented.

Copyright© Okereke GE, et al.

Result and Discussion

Accuracy Calculation of the Decision Tree Model

There are various performance metrics used in assessing the results obtained from Machine learning algorithm. Accuracy report is one important metric in determining the efficiency of the algorithm used. Some of the terms used in accuracy calculation are as below:

1. True positives (tpos): this represents the number of positive tuples that were labeled correctly with a positive class.
2. False positives (fpos): this represents the negative tuples labeled incorrectly with a positive class
3. True negatives (tneg): this represents the negative tuples labeled correctly with a negative class.
4. False negatives (fneg): this represents the positive tuples labeled incorrectly, with a negative class.

Accuracy also involves using terms such as *sensitivity* which is known as "true positive rate"; *specificity* "true negative rate" and *precision*, percentage of the positive tuples which were labeled correctly.

Accuracy (A) is the probability of choosing true positives and negatives from all positive and negative tuples or the probability of correct prediction for data stated as;

$$A = \frac{T_{pos} + T_{neg}}{pos + neg}$$

Evaluation of Performance Metric

Tables 1 & 2 shows accuracy table and precision table for the performance metric using the trained system to predict results of new students in first year for COS 101. A total of 103 students were tested and the accuracy was depicted as shown by rapid miner. Figure 1 shows the lift chart for COS 101.

Accuracy: 92.27%+/-9.70% (micro average: 92.23%)

	True Pass	True Fail	Class Precision
Pred.Pass	88	3	96.70%
Pred.Fail	5	7	58.33%
Class recall	94.62%	70.00%	

Table 1: Accuracy table of decision tree with COS 101.

Precision: 58.33% (positive class: FAIL)

	True Pass	True Fail	Class Precision
Pred.Pass	88	3	96.70%
Pred.Fail	5	7	58.33%
Class recall	94.62%	70.00%	

Table 2: Precision Table of COS 101.

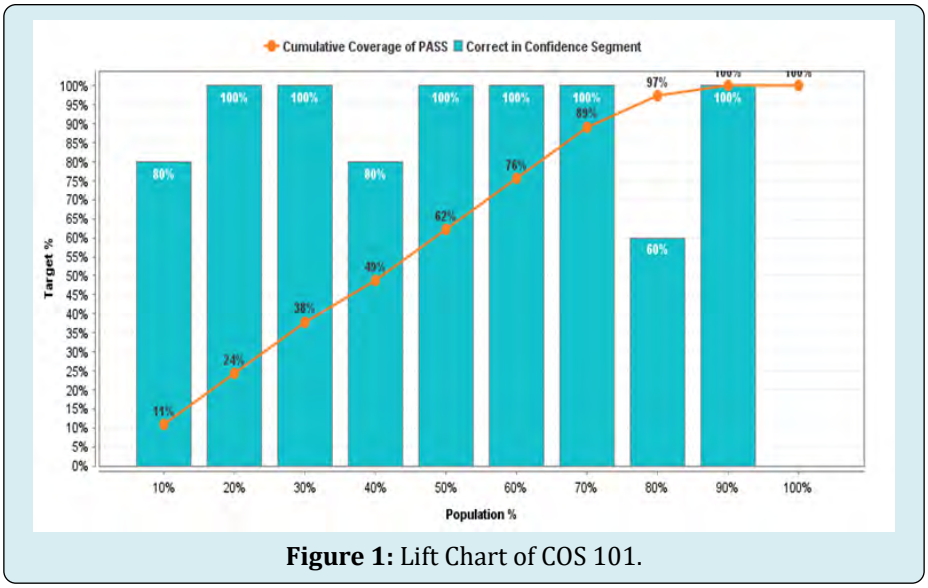

Figure 1: Lift Chart of COS 101.

Tables 3 & 4 shows accuracy table and precision table for the performance metric using the trained system to predict results of new students in first year for STA 172. A total of 103 students were tested and the accuracy was depicted as shown by rapid miner. Figure 2 shows the lift chart for STA 172.

Copyright© Okereke GE, et al.

Accuracy: 70.00%+/-7.84% (micro average:69.90%)

	True Pass	True Fail	Class Precision
Pred.Pass	64	25	71.91%
Pred.Fail	6	8	57.14%
Class recall	91.43%	24.24%	

Table 3: Accuracy Prediction of STA 172.

Precision: 57.14% (positive class: Fail)

	True Pass	True Fail	Class Precision
Pred.Pass	64	25	71.91%
Pred.Fail	6	8	57.14%
Class recall	91.43%	24.24%	

Table 4: Class Precision of STA 172.

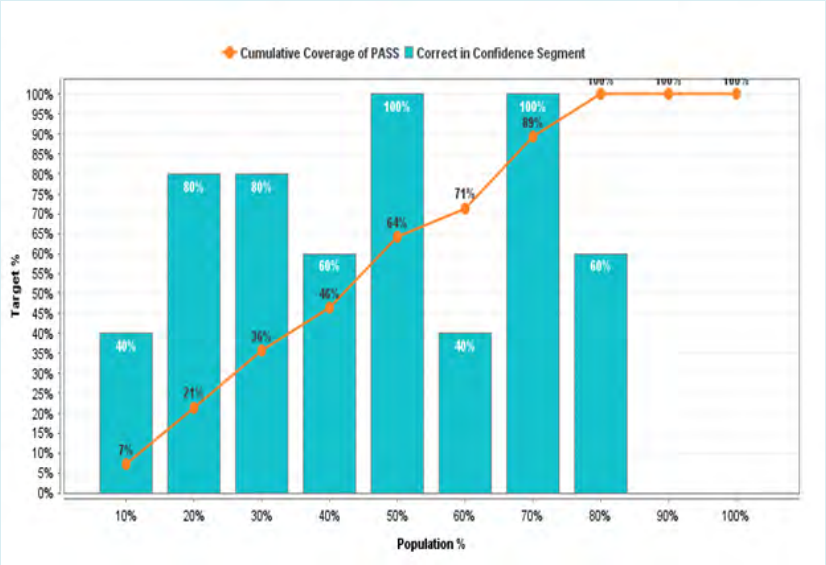

Figure: 2 Lift chart of STA 172.

Results Analysis

This paper uses machine learning algorithm to predict students' performance in exams in order to provide earlier warning to the students to focus more attention on those courses so as to improve their academic performance. It established the number of students likely to fail COS101 and STA172 - Pass: positive class and Fail: negative class -. The data sets used were obtained from records of previous first year students and were used in training and testing the decision tree algorithm.

In COS101 accuracy table, (table 1), those predicted to pass and truly passed were 88 students. However, there were 3 students classified incorrectly, they actually failed but were predicted as passed (Fpos). In the second row of table 1, there are 5 instances predicted as fail but actually passed (Fneg). Secondly, 7 instances were predicted to fail but actually failed (Tneg).

In table 3, the first row of STA172, 64 students were predicted to pass and truly passed (actually passed), but 25 students were predicted to pass but failed (Fneg). In the second row, 6 students were also predicted incorrectly as fail but actually pass (Tneg). The paper showed that although past examination records, financial status, family upbringing, performance in class test, attendance to lectures, the teaching capacity and general commitment to private studies affect students' performance in a course, however student's level of understanding of a courses affects the performance more.

Conclusion

The focus of this paper is to analyze data from students' record and use it in training the classifiers generated using decision tree algorithm to enable them make predictions. We also demonstrated in this paper that the choice of a classifier does not determine the accuracy of prediction but the nature of the datasets. Rapid miner aided us not only in preprocessing but also in training and testing our datasets. From the results obtained it was observed that one classifier, a decision tree was used in training two different datasets and the accuracy of the results obtained were different. This was as a result of different dispositions of students to a particular course. A student may have high understanding of COS101 but poor understanding of STA172 etc.

Copyright© Okereke GE, et al.

A comparison of the two results on the two courses under review, the Decision tree analysis for COS101 yielded highest accuracy of 96.7%, predicting the correct results of 95 students while only 8 students out of 103 students were predicted incorrectly. In the case of STA172, it was different, the same algorithm achieved 71.91% in accuracy prediction by predicting correctly only 72 students while 31 students were incorrectly predicted. We used the gain ratio attribute selection method in building the decision tree for classification with the attribute having the highest information gain being the node. From the rules or knowledge patterns based on the classifiers and the tree paths generated through entropy calculations, they were mapped into PHP code. This was done in form of class methods in PHP. The trees were translated into if-else code blocks and switch statements. The PHP class is what the web application used to make predictions for each course group.

References

1. Nwagwu HC, Okereke GE, Nwobodo C (2017) Mining and Visualizing Contradictory Data. Journal of Big Data 4.

2. Latha R (2017) Data Mining Techniques: Educational System. International journal of Advanced Research Trends in Engineering and Technology 4(Special Issue).

3. Ali D, Rabeeh NR, Abbasi RA, Lytras MD, Farhat A, et al. (2017) Predicting students' performance using Advanced learning Analytics. Proceedings of the 26th international Conference on worldwide web companion.

4. Romero C, Ventura S (2007) Educational data mining: A survey from 1995 to 2005. Expert systems with applications 33(1): 135-146.

5. Sajadin S (2012) An Application of predicting student performance using Kernel K-means and Smooth Support Vector Machine. University Malaysia Pahang, Malaysia.

6. Affendey L, Paris H, Mustapha N, Nasir SM, Muda Z (2018) Ranking of Influencing Factors in Predicting Students academic Performance. International Technology Journal 9(6): 832-837.

Copyright© Okereke GE, et al.