A Predictive Model for the Determination of Academic Performance in Private Higher Education Institutions

Francis Makombe[1], Manoj Lall[2]
Department of Computer Science
Tshwane University of Technology
Pretoria, South Africa

***Abstract*—The growth and development of predictive models in the current world has influenced considerable changes. Today, predictive modelling of academic performance has transformed more than a few institutions by improving their students' academic performance. This paper presents a computational predictive model using artificial neural networks to predict whether a student will pass or fail. The model is unique in the current literature as it is specifically designed to evaluate the effectiveness of the predictive strategies on neural networks as well as on five additional algorithms. The analysis of the experimental results shows that Artificial Neural Networks outperformed the eXtremeGBoost, Linear Regression, Support Vector Machine, Naive Bayes, and Random Forest algorithms for academic performance prediction.**

Keywords*—Classification modelling; data mining; higher education institutions; accuracy; academic performance

I. Introduction

Public higher education providers are institutions that have been established and funded by the state through the Department of Higher Education and Training (DHET). Public providers include universities, universities of technology, and comprehensive universities. Private providers are owned by private organizations or individuals. Higher education institutions (HEIs) operate in an increasingly complex and challenging environment. Competition has increased, and previously anticipated government funding has become scarce [1]. In such circumstances, HEIs must succeed in a financial sense or else they will go out of business [2]. In their quest for survival, common practices adopted by HEIs are to increase the intake of students and try to improve on their success rates. Since, many government and private funds depends on the throughput rates of institutions, being able to predict the chances of any new student's success is very important. This study aims to improve the pass rates of students' in a particular private academic institution by providing a classification model to assist in identifying student at risk of failing a program. Being able to identify such students, the educational institutes can provide a targeted support mechanisms to the needy students. The author in [3] mention that the reasons for the identification of a student at risk of dropouts or attrition early enough are to be able to provide necessary support and interventions for the student with the goal of reducing dropouts, increasing retention, performance and graduation rate.

Application of the appropriate data mining technique that suits the current scenario is important in order to identify useful patterns. In this article, factors that have an impact on the pass rates of students are identified and used in the classification model. The following algorithms are applied in the construction of the classification model-Artificial Neural Networks, Logic Regression, eXtremeGBoost, SVM, Naive Bayes, and Random Forest algorithms.

The rest of this article is structured as follows: the literature review is presented in Section II. The description of the data and the methodology used are presented in Sections III and IV. The results and its discussion are presented in Section V. In Section VI, conclusions and recommendations are presented.

II. Literature Review

In a research conducted by [4], the researchers attempted to explore the applicability of Fuzzy C-Means clustering technique for academic performance of students. They found that fuzzy C-Means clustering algorithm serves as a good benchmark to monitor the progression of students modelling in educational domain. The author in [5] also recommended a fuzzy logic-based expert system that periodically evaluates student performance and supplies students with feedback on progress within data grid environment. The system made use of the fuzzy logic theory and develop the decision making process based on fuzzy rules to assess whether a student gets very poor, poor, good, average or excellent performance.

In an attempt to identify the main attributes that may affect the performance of students in engineering, [6] applied data mining concepts such as k-Means clustering and Decision tree Techniques. They used records of 1500 students enrolled for various subjects in engineering. The author in [7] investigated the impact of classroom attendance and gender on academic performance of university students in an Organic Chemistry course. Data was collected through survey involving real time documentation of attendance for each student at each class lesson over a three month period. Their findings show that attendance had a significant impact on the performance. In another study, [8] analysed the impact of class attendance, practical work and assignments in a course on the success rate.

They found that the number of given assignment has a negative impact on the academic performance. They used C4.5 as the classification algorithm for their work. Several other studies conducted have shown that class attendance is an important predictor of academic outcomes which conclude that students who attend more classes generally earn higher final grades [9].

In a study by [10], one of the factors that influences a student's ability to succeed is the socioeconomic conditions. This fact is supported by [11] who state that Student poverty and the lack of sufficient funding have consistently been cited as key reasons for student academic failure and progression difficulties. In the study by [12], they used marks of four academic batches of Computer Science & Information Technology (CS&IT) students for predicting performance. In their study, they collected records of 347 undergraduate students have been mined with classifiers such as Decision tree, Neural Networks and Naive Bayes.

In another study, [13] applied Naïve Bayes for the classification of student evaluation. Their dataset consisted of the following parameter-age, place of birth, gender, high school status (public or private), department in high school, organization activeness, age at the start of high school level, and progress GPA score.

Discriminate analysis was done by [14] to predict the success and failure of students in a specific physics course. Discriminate analysis is a similar technique to multiple regression except that it is used for categorized data. They used this technique to provide a function that contains the variables that should be used for predicting the success of a student. They collected the data for 1622 students who enrolled in Electricity & Magnetism course, which had a high rate of failure. At first they identified many possible predictors such as, SAT grade, MATH GPA, Overall GPA. In another study [15], applied predictive modelling techniques to identity students at risk of dropping out of their registered qualification. They used Support Vector Machine, Naïve Bayes, Decision tree, K-nearest neighbors and Random Forest on 1156 students.

III. Data Description

This research followed a quantitative approach. Questionnaires were administered to private academic institutions in an anonymously manner to enhance the privacy and anonymity of the participants. The questionnaires in this study were distributed in two ways: manually handed out and also using the online survey tool survey monkey. The dataset consisted of the following attributes:

- Study hours per week.
- Bursary - whether a student has a bursary or not.
- Class Attendance.
- Student workload (number of modules registered).
- Fulltime study or attending through part-time classes.
- English language proficiency marks.
- Number of employed parents or guardians.
- Group Assignment marks.
- Test marks.
- Individual Assignment marks.

The scatterplot (Fig. 1) shows the distribution of individual test marks in relation to the individual assignment marks. In analysis of this scatterplot, most of the students perform well in both tests and individual assignments. There are a few outliers who perform very well in individual assignments but poorly in tests. According to this scatterplot, the approximate range for tests with most students' marks is 40 to 80, and that for the individual assignments is 50 to 90. This shows that students are generally performing better in individual assignments than in tests.

The scatterplot (Fig. 2) for Test and Group assignment marks shows that a greater proportion of students perform very well in group assignments, where they take part in research activities. By comparison, a lot of students fail the tests as shown by the large concentration of test marks below the mark of 50, compared to the test mark greater than 50. This could provide a basis for intervention by the private institution in efforts to assist the students prepare better for tests.

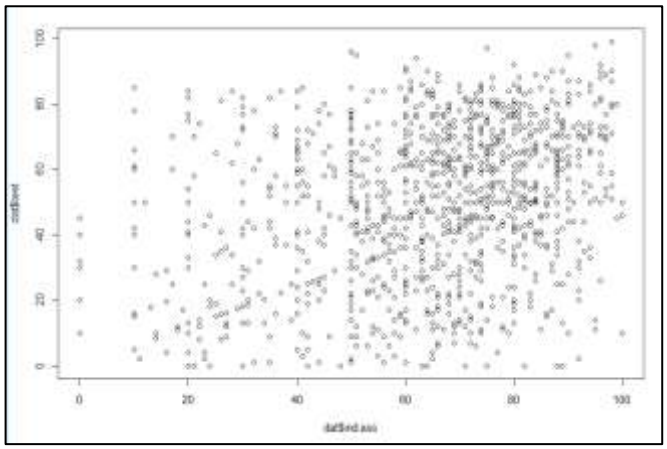

Fig. 1. Scatterplot of Test and Individual Assignment marks.

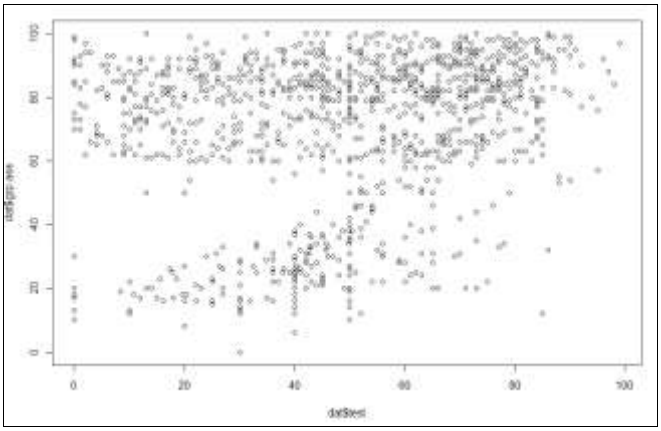

Fig. 2. Scatterplot of Test and Group Assignment Marks.

IV. Methodology

In order to assess the effect of data quality, data attribute significance and class number in the academic performance prediction in this study, six classification algorithms have been selected and implemented in R programming language. These algorithms were chosen because they cover the different approaches used by classifiers for learning and they are state of the art algorithms that are often used in data mining applications [16].

A. Random Forests

Instead of building a single tree for classification, the Random Forests constructs a set of trees, and uses them all to classify or to predict. Random Forests where developed by [17], and they create a (forest) collection of decision trees by the method of bagging. Random Forests are sets of learning models where the unknown input is listed according to the majority vote of decision-making bodies. This means that the class predicted by most of the trees would be the last class in the set. Random Forests, increase the classification performance, avoiding overfitting and are robust to outliers and noise [17].

B. Neural Network- Multilayer Perceptron (MLP)

This refers to an artificial neural feed network class in which at least three layers of nodes are present: one input layer, one hidden layer and one output layer. Every node is a non-linear activation neuron except for the input nodes. MLP uses a supervised learning method called training backpropagation [18]. Every node layer is fully connected to the next layer, which generates a finite acyclic graph (DAG).Except the input node, each node is a processing node that is used to calculate the output based on an input using a non-linear activation function. Each link of two nodes has a change in weight depending on the training data set. The weight adjustments are based on the error of the measured output difference and the predicted output. The weights are adjusted to reduce the error by using a gradient descent.

C. Support Vector Machines (SVMs)

SVMs for binary classification where developed by [19]. This is an approach that is used to solve classification problems using linear methods for both datasets having linearly and nonlinearly separable classes [20].

D. Linear Regression

Linear regression helps to predict the value of the Y outcomes variable on the basis of one or more X variables (Equation 1). The objective is to create a linear relationship (a mathematical formula) between a predictor variable(s) as well as the response variable, such that the value of the Y answer is determined by using this formula only when the values (Xs) of the predictors are known. In general, the formula for linear regression is provided as follows:

$$Y = \beta 1 + \beta 2X + \epsilon \quad (1)$$

where, $\beta 1$ is the intercept and $\beta 2$ is the slope. These are called regression coefficients, and ϵ is the error term, which refers to the area of Y, that the regression model cannot be able to explain.

E. Naïve Bayes Classifiers

These refer to a collection of "probabilistic classifiers" which are based on the application of Bayes 'theorem with strict (Naïve) independence assumptions amongst features. Naïve Bayes classifiers are very scalable.

F. eXtreme Gradient Boosting

eXtreme Gradient Boosting (XGBoost) is a versatile and enhanced gradient algorithm booster variant designed for efficiency, machine speed and performance of the model. It is an ensemble learning technique that combines multiple machine learning algorithms to lessen errors and increase prediction accuracies.

V. Results and Discussion

The following chart demonstrates the different accuracy, sensitivity, and F-measure values obtained (Fig. 3). Inaccuracies are also shown for each of the six algorithms used in this research. Fig. 1 shows that neural networks algorithm had the best accuracies which also had the least inaccuracies. It also had high precision, and F-measure values where a good classifier has an F-measure value of close to 1, whilst the worst classifier has an F-measure close to 0.

A. Receiver Operating Characteristic (ROC) Curve

The purpose of the Receiver Operating Characteristic (ROC) curve is to primarily assess the accuracy of a continuous measurement that is performing a binary outcome prediction. The best classifier has an area under the curve (AUC) value close to 1. Fig. 4 shown below are the AUC values for three classifiers, two with the best performance and one with a poor performance.

The following values where obtained for the AUC. This was done for three classifiers, which where the two best classifiers, and the worst classifier Table I.

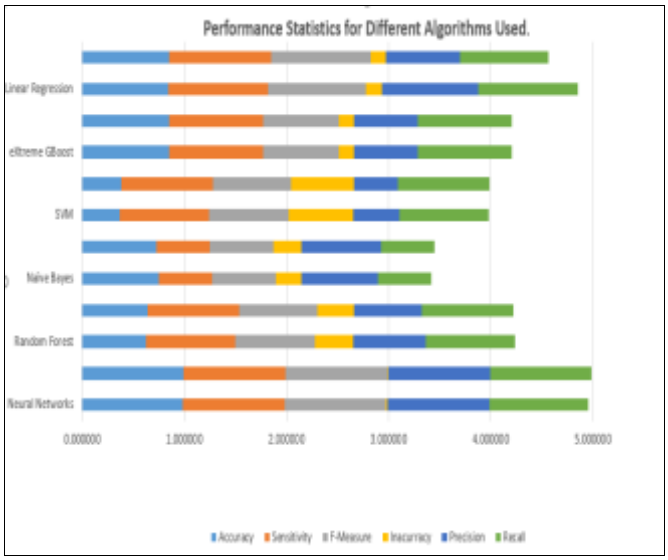

Fig. 3. Performance Statistics for different Algorithms used.

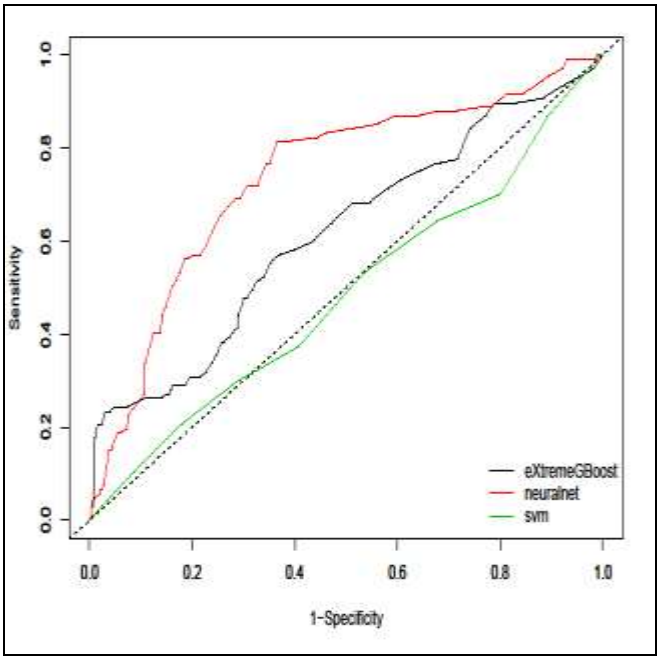

Fig. 4. ROC for Two Most Accurate and the Least Accurate Classifier.

TABLE I. AUC VALUES FOR TWO MOST ACCURATE CLASSIFIERS AND THE LEAST ACCURATE CLASSIFIER

Algorithm	AUC
eXtremeGBoost	0.62
Neural networks	0.86
SVM	0.43

By making use of the AUC and accuracy values obtained in this experiment the neural networks algorithm was selected to be the most suitable algorithm for the prediction of academic performance for this study. The performance of ANN was followed by eXtremeGBoost and then SVM. It can therefore be concluded that the neural net algorithm outperformed the other five algorithms for academic performance classification.

B. Confusion Matrix Results

Table II below summarizes the experimental results obtained for both the training and testing dataset, and it also demonstrates the accuracies and misclassification errors obtained using a neural network defined with the simple learning rate algorithm.

TABLE II. NEURAL NETWORK ALGORITHM WITH SIMPLE LEARNING RATE CONFUSION MATRIX RESULTS

CONFUSION MATRIX					
Dataset	**True Positive**	**False Positive**	**False Negative**	**True Negative**	**Misclassification Error**
Training data	5478	88	86	3636	0.019
Test data	1223	17	23	1059	0.017

Fig. 5 shows the predictions of a sample of six students using the neural networks. These are the computed values which show the predicted value of whether a student will pass or fail a module. The simple learning rate algorithm was used for these predictions. The value of 0.4566725 for the first student in the dataset means that the student is more likely to fail this module. Similarly, the value of 0.6010540 (which is greater than 0.5) for the second student would mean that this student is more likely to pass this module.

```
> head(output$net.result)
          [,1]
[1,] 0.4566725
[2,] 0.6010540
[3,] 0.5961648
[4,] 0.4566725
[5,] 0.4566725
[6,] 0.5130902
>
```

Fig. 5. Output of Neural Networks.

VI. CONCLUSIONS AND RECOMMENDATIONS

In the study the researcher shows the degree of accuracy of the six algorithms used in the study, and their related misclassification errors. It was observed that ANN performed better than Logic Regression, eXtremeGBoost, SVM, Naive Bayes, and Random Forest algorithms. It was observed that bursary and group assignments had a positive correlation with the pass rate. The recommendations, based on the results obtained, are: (1) Group assignments have a positive correlation concerning whether a student will pass or fail as they have a direct effect. Hence it is recommended that students should be encouraged to take a more active role in group assignments. (2) Bursaries have a positive correlation with academic performance; therefore, it is recommended for the private institute to provide bursaries to successful applicants. (3) There should be provision made for booster or support classes meant for students predicted to fail. To have a more accurate assessment of a student's academic performance, data from other domains of higher education value chain such as psychosocial domain, cognitive domain, institutional domain, personality domain, and demographic domain should be considered as future work.

REFERENCES

[1] E. J. Dumond and T. W. Johnson, "Managing university business educational quality: ISO or AACSB?," Quality Assurance in Education, 2013.

[2] H. J. Juhl and M. Christensen, "Quality management in a Danish business school–A head of department perspective," Total Quality Management, vol. 19, no. 7-8, pp. 719-732, 2008.

[3] O. W. Adejo and T. Connolly, "Predicting student academic performance using multi-model heterogeneous ensemble approach," Journal of Applied Research in Higher Education, 2018.

[4] R. S. Yadav and V. P. Singh, "Modeling academic performance evaluation using fuzzy c-means clustering techniques," International Journal of Computer Applications, vol. 60, no. 8, 2012.

[5] S. Patel, P. Sajja, and A. Patel, "Fuzzy logic based expert system for students performance evaluation in data grid environment," International Journal of Scientific & Engineering Research, vol. 5, no. 1, 2014.

[6] V. Sreenivasarao and C. G. Yohannes, "Improving academic performance of students of defence university based on data warehousing and data mining," Global Journal of computer science and technology, 2012.

[7] O. D. Ayodele, "Class attendance and academic performance of second year university students in an organic chemistry course," African Journal of Chemical Education, vol. 7, no. 1, pp. 63-75, 2017.

[8] N. A. Yassein, R. G. M. Helali, and S. B. Mohomad, "Predicting student academic performance in KSA using data mining techniques," Journal of Information Technology and Software Engineering, vol. 7, no. 5, pp. 1-5, 2017.

[9] A. Kirby and B. McElroy, "The effect of attendance on grade for first year economics students in University College Cork," Vol. XX, No. XX, Issue, Year, 2003.

[10] D. E. Roby, "Research on school attendance and student achievement: A study of Ohio schools," Educational Research Quarterly, vol. 28, no. 1, pp. 3-16, 2004.

[11] S. Mngomezulu, R. Dhunpath, and N. Munro, "Does financial assistance undermine academic success? Experiences of'at risk'students in a South African university," Journal of Education (University of KwaZulu-Natal), no. 68, pp. 131-148, 2017.

[12] R. Asif, A. Merceron, and M. K. Pathan, "Predicting student academic performance at degree level: a case study," International Journal of Intelligent Systems and Applications, vol. 7, no. 1, p. 49, 2014.

[13] N. Dengen, E. Budiman, M. Wati, and U. Hairah, "Student Academic Evaluation using Naïve Bayes Classifier Algorithm," in 2018 2nd East Indonesia Conference on Computer and Information Technology (EIConCIT), 2018: IEEE, pp. 104-107.

[14] E. W. Thomas, M. J. Marr, A. Thomas, R. M. Hume, and N. Walker, "Using discriminant analysis to identify students at risk," in Technology-Based Re-Engineering Engineering Education Proceedings of Frontiers in Education FIE'96 26th Annual Conference, 1996, vol. 1: IEEE, pp. 185-188.

[15] R. Lottering, R. Hans, and M. Lall, "A model for the identification of students at risk of dropout at a university of technology," in 2020 International Conference on Artificial Intelligence, Big Data, Computing and Data Communication Systems (icABCD), 2020: IEEE, pp. 1-8.

[16] I. H. Witten and E. Frank, "Data mining: practical machine learning tools and techniques with Java implementations," Acm Sigmod Record, vol. 31, no. 1, pp. 76-77, 2002.

[17] L. Breiman, "Random forests," Machine learning, vol. 45, no. 1, pp. 5-32, 2001.

[18] J. Gao, X. He, and L. Deng, "Deep learning for web search and natural language processing," 2015.

[19] C. Cortes and V. Vapnik, "Support-vector networks," Machine learning, vol. 20, no. 3, pp. 273-297, 1995.

[20] D.-M. Tsai and C.-C. Lin, "Fuzzy C-means based clustering for linearly and nonlinearly separable data," Pattern recognition, vol. 44, no. 8, pp. 1750-1760, 2011.