ANN-based Model For predicting Academic Success In A Private Educational Institution

Francis Makombe, Manoj Lall

Abstract: The pressure to improve on success rates is greater on private HEIs than on public funded HEIs, as their main source of funding is from fees collected from students. Poor success rates will undoubtedly affect private HEIs' reputation and funding. To minimise the impact of a poor success rate amongst students, it is important to be able to identify students at risk of failing at an early stage, so that a more targeted remedial action could be taken. Private institutions apply various strategies such as making provision for extra tuition, extended laboratory access and establishing learning communities. From the discussion presented here, it is apparent that the timely identification of students at risk of failing a particular programme is of significant importance to both the students and the institutions they are registered with. In this article, artificial neural networks, extreme gradient boost, logistic regression, support vector machine, naive Bayes, and random forest are used for the classification of students. A dataset of 3 000 students were collected from a private higher education provider. It was observed that artificial neural networks produced the best performing model with an accuracy of 88.07%.

Index Terms: Accuracy; Academic performance; Artificial Neural Networks; Classification modelling; Data mining; Higher education institutions; Predictive model.

1 INTRODUCTION

OVER the past decades, there have been several developments in the higher education sector, notably an increase in the number of institutions which prospective students can choose from when considering furthering their education. Along with the increase in public universities, there has also been substantial growth in the private higher education sector. Public higher education providers are institutions that have been established and funded by the state through the Department of Higher Education and Training (DHET). Whereas, private education providers are owned by private organizations or individuals. Whilst both public and private institutions may offer the same qualifications, a major difference lies in their funding models - public institutions are subsidised by the government while private institutions are not and must generate their own funding [1]. Students' academic performance plays a vital role in educational institutions, as it is often used as a metric for the institution's performance [2]. The impact of poor success rate on institutes of higher learning, whether these institutions are government or privately funded, can be serious as most are 'tuition dependent' [3]. Additionally, a poor success rate is perceived by some as a measure of quality of education offered by a particular educational institution [4]. According to the American Council on Education (ACE) success rates have become a key component of discussions about accountability in higher education.

- *Francis Makombe is currently pursuing master's degree program in the Department of Computer Science, Tshwane University of Technology, Pretoria, South Africa. E-mail: francismakombe@gmail.com*
- *Manoj Lall, Department of Computer Science, Tshwane University of Technology, Pretoria, South Africa E-mail: LallM@tut.ac.za*

To address the problem of poor success rates amongst its students, many private institutions apply some forms of progress monitoring techniques to identify academically weak students and then apply a more focused remedial actions such as assigning tutors to needy students and providing labs access beyond the normal allocated time. Early detection of students at risk, along with the institution of preventive measures, can drastically improve their success [5]. Amongst the various approaches adopted to address the problem of identifying students at risk of failing, applying educational data mining (EDM) techniques continue to be in the forefront. The main objective of EDM is to analyse educational data in an attempt to provide answers to questions emanating in the domain of education [6]. This research attempts to enhance students' pass rates at a private academic institution by developing a machine learning based classification model that can help identify students at risk of failing a program. In order to build the classification model, the following algorithms are used, Logic Regression, eXtremeGBoost, Artificial Neural Networks, SVM, Naive Bayes, and Random Forests. This paper is organised as follows: literature review is presented in Section 2 while the methodology, experimental results and its discussions are presented in Sections 3 and 4 respectively. Finally, in Section 5, conclusion and future works are presented.

2 LITERATURE REVIEW

Being able to predict student performance is essential in order to help at-risk students and assure their retention, and at the same time improve the university's ranking and reputation. In essence, strong emphasis is being placed on knowing which students are candidates of poor academic performance and what factors contributed to this, so that early action could be taken to support the students. In a study conducted by [7], the researchers used 'O' level grades (terminal high school results) and their first 3 semesters marks to predict the performance of the students. They applied SVM, RF, KNN, decision trees, and linear regression algorithms in their models. They reported that SVM and the RF performed better than other algorithms. In an attempt to predict the dropout rate of students at an institution of higher learning, [8] conducted a study which consisted of variables such as accommodation, age, credits collected, disability, financial aid, gender, and

years spent in the system. Their study made use of Random Forests, Support Vector Machines, Decision Trees, Naïve Bayes, K-Nearest Neighbor, and Logistic Regression for classification purposes. They obtained an accuracy of 96.2% using Random Forest. In another study conducted in Portugal by [9] to determine the students likely to drop out of the university, a decision tree model was constructed using data of 2,970 first year university students. Results obtained from their study showed that the academic performance variable was confirmed as a significant determining factor. A multiple regression technique was applied by [10] to predict the success of students by analysing students' profiles data. It was reported that mother's education level plays a positive role in a student's success. In another study by [11], J48, random forest, multilayer perceptron, IB1 and decision trees algorithms were used for predicting student performance on attributes such as attendance, students' result, economic conditions, parent education, locality, and gender. Their study concluded that the performance of Random Forest was better than that of other classifiers. In another study conducted by [12], variables such as midterm marks, lab test grade, seminar performance, assignment, and attendance were used in creating a decision trees model to predict students' academic performance. They observed that economic status, family and relation support had high probabilities of affecting student's performance in an examination. In their research to create a predictive model [13], using a dataset consisting of variables such as students' presence in the laboratory exercises, and students' results in their first two tests, they concluded that the J48 classification algorithm provided good predictive performances. In another study by [14], variables such as semester hours, overall GPA for the student, and percentage of attendance in a particular class where utilized to predict performance in a subject. By applying a regression model, they concluded that the overall student's GPA and the attendance percentage are the most significant factors in determining the grade attained in a specific subject. The authors in [15] undertook a study to predict the final marks of students. Their dataset consisted of the following variables: student's first year result, class attendance, parent income, and the daily distance a student travels to college. They made us of the ID3 algorithm for this purpose. In another study, [16], C4.5, ID3 and CART decision tree algorithms were applied on engineering students' data to predict their performance in the final exam. They collected data on variables such as branch, sex, students' grade in high school, father's education level and family size. The accuracy values obtained in their experiments showed decision trees to be able to successfully identify students likely to fail. In a study to predict the performance of Information Technology (IT) students at the end of the first year, [17] utilised C4.5 and Naive Bayes algorithms and obtained an accuracy of 98.64%. In order to predict the dropout rate of students at an institution of higher learning, [8] used a dataset consisting of variables such as accommodation, age, credits collected, disability, financial aid, gender and years spent in the system. The study made use of the following algorithms for classification purposes: random forests, support vector machines, decision trees, naïve Bayes, K-nearest neighbor and logistic regression. They observed that random forest, with an accuracy rate of 96.2% was the best performing algorithm. In the current study, the following six algorithms are used for model creation: Random forests, Artificial neural network, Support vector machine, Logistic regression, Naive Bayes, and eXtreme gradient boosting.

2.1 Random Forests

The Random Forest (RF) algorithm is a supervised learning model. It uses labelled data to 'learn' how to classify unlabelled data. Instead of building a single tree for classification, random forest constructs a set of trees and uses them all to classify or to predict. Random forests are sets of learning models where an unknown input is listed according to the majority vote of decision-making bodies. Random forests provide advantages such as increased classification performance, avoid overfitting and are robust to outliers and noise [18]. Besides high prediction accuracy, a random forest is efficient, interpretable and non-parametric for various types of datasets [19]. In order to decide how the nodes on a decision tree branch into several possible outcomes, Gini index and the generalised entropy measures are used (see Equations 1 and 2) [20].

$$Gini = \sum_{i=1}^{c} (p_i)^2 \quad (1)$$

$$Entropy = \sum_{i=1}^{c} - p_i * \log_2(p_i) \quad (2)$$

Where pi represents the relative frequency of the class observed in the dataset and c represents the number of classes.

2.2 Artificial Neural Network

An artificial neural network (ANN) is a well-documented artificial intelligence (AI) model inspired by the framework of biological human neurons. ANN has been successfully applied to numerous problems in different fields [21]. In essence, it consists of three layers, namely an input layer, a hidden layer and an output layer (see Fig 1).

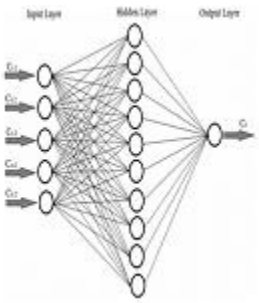

Fig. 1. *ANN model with hidden layers [22]*

Except for the input node, each node is a processing node that is used to calculate the output based on an input using an activation function. Commonly used activation functions are the linear function, sigmoid function and the tanh function [22]. The goal of the ANN learning algorithm is to determine a set of weights that minimises the total sum of square errors.

2.3 Support Vector Machines

Support vector machines (SVMs) are machine learning algorithms widely used for classification and in prediction

problems [23]. SVMs are described as sets of related supervised learning techniques used for classification and regression tasks [24]. An SVM utilizes a kernel function to perform both linear and non-linear classifications. For a simple binary high dimensional linear classification problem, the SVM algorithm builds a hyperplane with the intention of maximizing the distance between the hyperplane and the nearest data points on each side, that is the margins (See Fig 2) [25], [26].

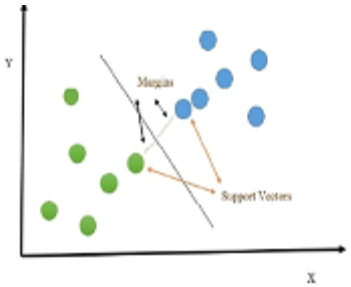

Fig 2. *Support Vector Machine [25]*

2.4 Logistic Regression

Logistic regression is well-known to the data mining research community as a tool for classification and modelling [27]. In modern years, the usage of logistic regression modelling has received a great deal of attention in the literature. Unlike linear regression, logistic regression can directly predict probabilities (values that are restricted to the 0-1 interval). Logistic regression preserves the marginal probabilities of the training data. The coefficients of the model also provide some hint of the relative importance of each input variable.

2.5 Naïve Bayes classifiers

Naïve Bayes classifiers refer to a collection of 'probabilistic classifiers' which are based on the application of Bayes' theorem with strict (naïve) independence assumptions amongst its features [28]. Several comparative studies of classification methods found that the Naïve Bayes classifier algorithm is known to outperform even highly sophisticated classification methods. It provides a way of calculating posterior probability P(c|x) from P(c), P(x) and P(x|c). Some of the advantages of the Naïve Bayes algorithm are that it is easy to implement, possesses improved time efficiency, can handle missing data [29].

2.6 eXtreme Gradient Boosting

The extreme gradient boosting (XGBoost) algorithm is a scalable machine learning system for tree boosting. Boosting is a machine learning technique that can be used for regression and classification problems. The main difference between random forest (RF) and gradient boosted machines (GBMs) is that while in RF, trees are built independently to each other, GBM adds a new tree to complement already built ones. XGBoost is an implementation of gradient boosted decision trees designed for speed and performance.

3 RESEARCH METHODOLOGY

The knowledge discovery in databases (KDD) process is applied for the creation of the proposed predictive models. Fig 3 depicts the overview of the steps involved in the KDD process. Data is subjected to various processes till knowledge is extracted, and the steps followed are selection, preprocessing, transformation, data mining, and evaluation.

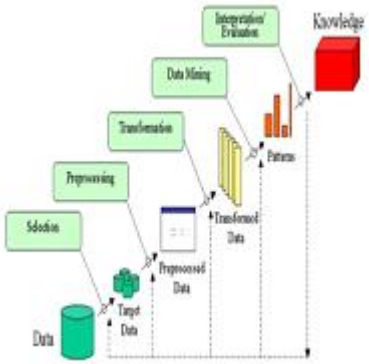

Fig 3: *Knowledge Discovery in Databases [30]*

A dataset consisting of 3000 student records was obtained from the National Office of a Private Tertiary Education Provider (PTEP). Additional data were collected from students by means of questionnaires using an online survey tool (survey monkey) and merged together. The data attributes and their descriptions are listed in Table 1.

TABLE 1: *DESCRIPTION OF THE DATASET ATTRIBUTES*

	Attribute	Description	Values
1.	STUDYHRS	Indicates the number of study hours per week done by the student.	Min 0, Max 56
2.	STDWRKLD	Refers to the number of modules a student is registered for in an academic year.	Min 1, Max 8
3.	STDPVTY	Indicates the number of employed parents or guardians.	0,1,2
4.	ENGP	Indicates the English language proficiency marks obtained by the student.	Min 0. Max 100
5.	CLASSA	Refers to a student's class attendance statistics per semester.	Min 0, Max 80
6.	BURSARY	Indicates if a student has a bursary or not.	Yes/NO
7.	FINAL MARK	Refers to the aggregate mark for all the subject taken.	Min 0, Max 100
8.	FTPT	Refers to whether a student is registered either as a fulltime student, or as a part-time student.	FT/PT
9	AGE	Indicates the age of the student.	Min10, Max 90
10	GENDER	Indicates the sex of the student.	Female/ Male
11	PROVINCE	Refers to the province the student comes from.	1 -9 (SA provinces) 0 – for international
12	HOME_LANGUAGE	Indicates the home language of the student.	AFRI, ENGL, ISIN, ISIX,ISIZ,

			OTHR, SESO, SESS, SETS, SISW, TSHI, XITS
13	PASSORFAIL (CLASS ATTRIBUTE)	Indicates if the student passes the qualification or not during a registered year. NOTE a student is considered to pass if he/she obtains a maark of 50% or more., Otherwise the student is considered to have failed the qualification.	Pass / Fail

Before being used for developing a predictive model, the collected raw data were scaling and normalization. In addition, the raw data are checked for missing values and outliers. In this research, as the number of outliers were few, they were removed from the dataset. Missing values were imputed using the mean values. In addition, data cleaning was performed to eliminate typographical errors and remove duplicate records. Feature selection outlines the process of reducing the number of input variables through the removal of noise from the dataset [31]. This research made use of random forest for feature selection. Table 2 shows the features that were confirmed to have predictive powers.

TABLE 2: *RELEVANT FEATURES*

Variable	Relevance Decision
1. STUDYHRS	Confirmed
2. STDWRKLD	Confirmed
3. STDPVTY	Rejected
4. ENGP	Confirmed
5. CLASSA	Confirmed
6. BURSARY	Confirmed
7. FINAL MARK	Confirmed
8. FTPT	Confirmed
9. AGE	Rejected
10. GENDER	Rejected
11. PROVINCE	Rejected
12. HOME_LANGUAGE	Rejected

4. RESULTS AND DISCUSSION

The following table outlines the performance measures obtained for the six models.

TABLE 3: *SUMMARY OF PERFORMANCE MEASURES OF THE SIX MODELS*

	ANN	RF	NB	SVM	XG_ Boost	LR
Accuracy	0.831	0.626	0.748	0.605	0.809	0.802
Sensitivity	0.706	0.586	0.526	0.425	0.696	0.679
F-Measure	0.739	0.681	0.619	0.711	0.702	0.696
Precision	0.775	0.705	0.750	0.453	0.7516	0.7227

The ROC-AUC curve of the six models are presented in Fig 4. Since, ANN has a better performance than the other models, it was selected for further fine-tuning to improve its predictive performance.

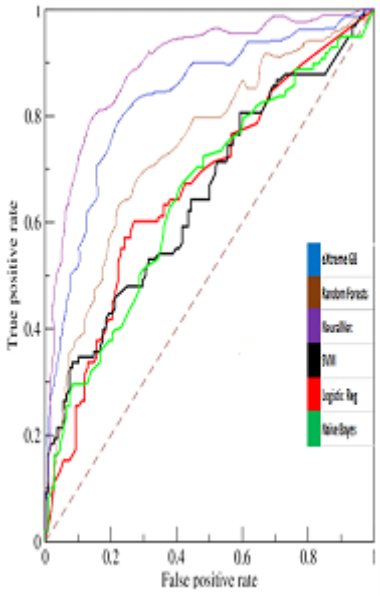

Fig 4: *ROC Curves of the six classification models.*

The fine-tuning was conducted by varying the number of hidden layers between one and three and applying either the rectified linear unit Relu or Sigmoid activation functions. An activation function may be regarded as a mapping of summed weighted input to the output of the neuron. Table 4 shows the Accuracy and loss when using different activation functions and the number of hidden layer.

TABLE 4: *MODEL ACCURACIES FOR DIFFERENT ACTIVATION FUNCTIONS AND HIDDEN LAYERS.*

	Activation Function	Number of Hidden Layers	Accuracy (%)	Loss score
1	relu	1	83.07	0.5212
2	relu	2	87.14	0.3987
3	relu	3	88.07	0.3642
4	sigmoid	1	78.94	0.4983
5	sigmoid	2	82.57	0.4357
6	sigmoid	3	85.60	0.3763

5 CONCLUSIONS AND RECOMMENDATIONS

In this research, it was established that using a predictive model based on machine learning could be an approach that a private tertiary education institution can apply to identify students at risk of poor performance. For this purpose, this research made created six models based on popular machine learning algorithms. It was observed that the ANN using 3 hidden layer and relu as the activation function was the best performing model for the given dataset. 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] DHET. 2013. White paper for post-school education and training. building an expanded, effective and integrated post-school system. November 2013.

[2] M.N. Quadri and N. V. Kalyankar, "Drop out feature of student data for academic performance using decision tree techniques." Global Journal of Computer Science and Technology, 2010.

[3] F. Makombe and M. Lall. "A predictive model for the determination of academic performance in private higher education institutions." International Journal of Advanced Computer Science and Applications (IJACSA), 2020.

[4] J. Liang, C. Li, and L. Zheng, "Machine learning application in MOOCs: Dropout prediction", 2016 11th International Conference on Computer Science & Education (ICCSE).52-57, 2016.

[5] B. O. Barefoot, "Higher education's revolving door: confronting the problem of student drop out in US colleges and universities", The Journal of Open, Distance and e-Learning, 19(1):9-18, 2004.

[6] T. Barnes, M. Desmarais, C. Romero, and S. Ventura, "Educational data mining", Proceedings of 2nd International Conference on Educational Data Mining., 2009.

[7] J. Akinode, and S, Oloruntoba, "Student academic performance prediction using support vector machine", International Journal of Engineering Sciences & Research Technology, 2017.

[8] R. Lottering, R. Hans and M. Lall, "A model for the identification of students at risk of dropout at a university of technology", International Conference on Artificial Intelligence, Big Data, Computing and Data Communication Systems (icABCD) 6-7 Aug. 2020. DOI:10.1109/icABCD49160.2020.9183874.

[9] J. Casanova, A. Cervero, C. Núñez, S. Almeida, and A. Bernardo, "Factors that determine the persistence and dropout of university students", Psicothema, 30(4), 408-414, 2018.

[10] M. NAQVI, and S. HIJAZI, "Factors affecting Students' Performance A Case Of Private Colleges", Bangladesh e-Journal of Sociology., 3 (1), 65-99, 2006.

[11] Shanavas and Mythili., "An Analysis of students' performance using classification algorithms", Journal of Computer Engineering (IOSR-JCE), 16(1):63-69, 2014.

[12] L. D. JAGADESH, and S. ARUNDATHI S. "Data Mining_A prediction for Student's Performance Using Decision Tree ID3 Method", International Journal of Scientific & Engineering Research Volume 5(7):13-29, 2014.

[13] I. Milos, V. Mladen, and S. Alatresh, "Students' success prediction using Weka tool", Infoteh-Jahorina . 15 (15), 2016.

[14] S. Obeidat, A. B.ashir, W. A. Jadayil, "The Importance of Class Attendance and GPA for academic success in Industrial Engineering Classes", World Academy of Science, Engineering and Technology, 61, 1192 – 1195, 2012.

[15] U. LANJEWAR and A. CHAWARE, "ID3 Derived Fuzzy Rules for Predicting the Students Academic Performance", IOSR Journal of Computer Engineering, 16(6):53-60, 2014.

[16] 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), 2(2):51-56, 2012.

[17] J. Delima, R. Vilchez and J. Alejandrino, "IT Students Selection and Admission Analysis using Naïve Bayes and C4.5 Algorithm". International Journal of Advanced Trends in Computer Science and Engineering, 9(1), .2020.

[18] L. Brieman , " Random Forests.", Machine Learning, pp 5 – 32, 2001.

[19] N. Donges, "A Complete Guide to the Random Forest Algorithm", https://builtin.com/data-science/random-forest-algorithm, 2021.

[20] M. Embrechts, L. Han, B. Szymanski, K. Sternickel, and A. Ross, "Random Forests Feature Selection with Kernel Partial Least Squares: Detecting Ischemia from Magneto Cardiograms:", European Symposium on Artificial Neural Networks, Burges, Belgium:, 221-225. 2006.

[21] R. Crespo, F. López-Martínez, E. Núñez-Valdez, and V. García-Díaz, "An artificial neural network approach for predicting hypertension using NHANES data", Scientific Reports, 10(10620), 2020. DOI: 10.1038/s41598-020-67640-z.

[22] R. NIAZKAR and M. NIAZKAR, "Application of artificial neural networks to predict the COVID-19 outbreak", Global Health Research and Policy, 5(1):50, 2020.

[23] Y. Pang, N. Juddb, J. O'brienb and M. Ben-Avieb. "Predicting Students' Graduation Outcomes through Support Vector Machines", Conference: 2017 IEEE Frontiers in Education Conference (FIE), 2017.

[24] J. Akinode, and S. Oloruntoba, " Student academic performance prediction using support vector machine.", International Journal of Engineering Sciences & Research Technology, 2017.

[25] D. Mohan and M. G. Gopal., "Quality Analysis Of Rice Grains Using ANN And SVM", Journal of Critical Reviews, 7(1), 2020.

[26] K. Baradwaj and S. Pal, "Mining Educational Data to Analyze Students Performance", International Journal of Advanced Computer Science and Applications, 2(6), 2011.

[27] PV Anusha, C Anuradha, PSRC Murthy, CS Kiran, "Logistic Regression Approach for Outlier Mining in High Dimensional Dataset", International Journal of Advanced Research in Engineering and Technology (IJARET), 12(1), 2021.

[28] M. Boulle, "Compression-Based Averaging of Selective Naive Bayes Classifiers.", Journal of Machine Learning Research , 1659-1685, 2007.

[29] A.P. Wibawa, A. Kurniawan, D. Murti, R.P. Adiperkasa, S. Putra, S. A. Kurniawan and Y, Nugraha, "Naïve Bayes Classifier for Journal Quartile Classification", *iJES,* 7(2), 91-99, 2019.

[30] N. Kumar, S. Jain, and K. Chauhan, "Knowledge Discovery from Data Mining Techniques", International Journal of Engineering Research & Technology (IJERT), 7(12), pp.1-3. 2019.

[31] J. Brownlee. , "How to choose a feature selection method for machine learning." Machine Learning Mastery, 10 , 2019.