Education and Information Technologies (2019) 24:1527–1543
https://doi.org/10.1007/s10639-018-9839-7

Data mining approach to predicting the performance of first year student in a university using the admission requirements

Aderibigbe Israel Adekitan[1] · Etinosa Noma-Osaghae[1]

Received: 17 September 2018 / Accepted: 13 November 2018 / Published online: 3 December 2018

Abstract

The academic performance of a student in a university is determined by a number of factors, both academic and non-academic. Student that previously excelled at the secondary school level may lose focus due to peer pressure and social lifestyle while those who previously struggled due to family distractions may be able to focus away from home, and as a result excel at the university. University admission in Nigeria is typically based on cognitive entry characteristics of a student which is mostly academic, and may not necessarily translate to excellence once in the university. In this study, the relationship between the cognitive admission entry requirements and the academic performance of students in their first year, using their CGPA and class of degree was examined using six data mining algorithms in KNIME and Orange platforms. Maximum accuracies of 50.23% and 51.9% respectively were observed, and the results were verified using regression models, with R^2 values of 0.207 and 0.232 recorded which indicate that students' performance in their first year is not fully explained by cognitive entry requirements.

Keywords Academic performance · Machine learning · Educational data mining · Data mining algorithms · Knowledge discovery · Nigerian university

1 Introduction

Nigeria, with a teeming population of over 180 million people has a very large pool of people who earnestly desire university education. The last decade has witnessed a steep rise in the demand for university education in Nigeria (Aina 2002). The demand is so

✉ Aderibigbe Israel Adekitan
aderibigbe.adekitan@covenantuniversity.edu.ng

[1] Department of Electrical and Information Engineering, Covenant University, Ota, Ogun State, Nigeria

high that only a meagre 15.3% of applicants at the most get the chance to be admitted into a Nigerian university each year (Aluede et al. 2012). Nigeria currently has one hundred and fifty two (152) universities that furnish the higher education demands of a fresh 1.7 million aspiring undergraduates each year. This is besides the five hundred thousand (at the least) young Nigerians that graduate from these higher institutions of learning every year (Saint et al. 2003). Nigerian universities have people from various culture coming together to learn. The northern region of Nigeria, which before now, were laid back about university education, is now witnessing a surge in demand for university education (Popoola et al. 2018).

With increasing growth in the national population and demand for post-graduate education, it became evident that the number of tertiary institutions in the country is grossly inadequate. The Federal Government of Nigeria has explored e-learning and distance learning to cater for the teeming number of Nigerians who desire university education. The National Open University of Nigeria (NOUN) was established chiefly for the purpose of providing e-learning and distance education services to Nigerians, and other interested parties abroad (Ajadi et al. 2008). NOUN serves as an Open and Distance Learning (ODL) platform that provides instructional materials to students, to enable self-paced learning which is very convenient for working student based on the framework by Indira Ghandi National Open University (IGNOU), India. The program needs continuous performance monitoring and improvement to ensure compliance with global best practices on ODL delivery. Implementation of online-based, peer-assisted study sessions (Nikolic and Nicholls 2018) might help in providing supplementary instruction and peer-based support for NOUN students towards enhancing student performance.

According to (Olsson and Mozelius 2016), when direct learning facilitation is not feasible, there is a need to ensure that alternative self-learning platforms are adequate, in order to prevent failure rates from increasing. The study by (Burke and Fedorek 2017; Kurt 2017) assert that self-directed learning, based on flipped classroom model enhances student learning while (Van WYK 2018) recommends the use of flipped classroom concept in open distance e-Learning programmes for enhancing students' experience, understanding and performance. In contrast, (Burke and Fedorek 2017; Ebbeler 2013) posit that flipped classroom may not necessarily improve student learning experience because students who are used to traditional classes often prefer the traditional approach while those in flipped classes may not be ready for the transition. Also, it may be challenging to teach a flipped class and likewise, flipped classes may also not be easy for students.

The deregulation of university education in Nigeria brought a new push for the creation of more universities in Nigeria (Adeogun et al. 2009). But this push came with its attendant problems of clearer disparity between the "haves" and "have nots", and declining educational quality due to profit maximization (Babalola 1998). The Federal Government of Nigeria, through the instrument of the National Universities Commission (NUC), keeps working assiduously to alleviate the myriad of problems bedevilling university education in Nigeria. The body provides the regulatory needs for university education in Nigeria and ensures along with other stakeholders, the implementation of all quality assurance policies in Nigerian universities (Ajayi and Ekundayo 2008). The National Universities Commission along with other bodies like the Joint Admissions and Matriculation Board (JAMB), the National Examination

Council (NECO) and the West African Examination Council (WAEC) work hand-in-hand to set benchmarks and standards for admission into Nigerian universities. A lot of criteria are usually used, the most prominent of them being the need to have credits in at least five (5) subjects, Mathematics and English inclusive in a recent WAEC or NECO examination (Adeyemi 2001).

Recently, a lot of universities started internal university undergraduate selection programmes based on their customized criteria. A famous example is the Covenant University Scholastic Aptitude Screening (CUSAS) programme (Popoola et al. 2018). The CUSAS of Covenant University for example, has a very stringent list of requirements for her aspiring undergraduates. These requirements cut across academic, emotional, financial, social, moral and spiritual areas (Popoola et al. 2018). The aim of selection programmes and policies for student admission is to ensure that all-round students, in terms of capacity, capability, creativity and motivation are given the few, prestigious, keenly and fairly contested admission slots available.

In this study, the relationship between cognitive entry characteristics of students at the point of admission as measured by the students' entry age, the aggregate WAEC score, the JAMB score, the university based CUSAS score and the students' first year academic performance measured by their grade class and the actual CGPA is considered using data mining. This study seeks to determine the extent of the relationship between the admission criteria used by Nigerian universities for selecting qualified prospective undergraduate applicants for 100 L admission, and the academic performance of the student after the first academic session using data mining algorithms and regression-based models.

2 Educational data mining

Data mining is a knowledge discovery process which entails the extraction of intelligent information from a given dataset using scientific methodologies (Azevedo 2018; Hussain et al. 2018). Dataset accumulated overtime from a process or system contains hidden historical information which can be data mined for enhancing the quality of decision-making processes. Data mining entails the use of algorithms for identifying patterns and trends within a dataset. Educational data mining is the extraction of useful information from dataset generated in the educational domain (Tair and El-Halees 2012; Ryan and Baker 2010; Senthil and Lin 2018; Bharara et al. 2018). Educational databases are rich sources of information for evaluating student performance and for various predictive analyses (Ahmed and Elaraby 2014; Khedr and El Seddawy 2015; Bharara et al. 2018). It also helps in identifying any hidden relationship between students' performance and their learning characteristics (Hussain et al. 2019; Ahuja et al. 2019) and behaviours (Kim et al. 2018). Useful information obtained from data mining can be used to improve the quality and the mode of delivery of higher education systems (Daradoumis et al. 2019) in order to improve teaching efficiency, and ultimately student performance (Baepler and Murdoch 2010; Osmanbegović and Suljic 2012; Agarwal et al. 2012). Data mined information can guide in identifying areas of intervention such as course redesign, modification of communication strategies, improved assessment methods, and so forth towards improving the quality of education and aptitude (Baepler and Murdoch 2010).

Educational data mining is a data driven process for identifying student learning issues and performance trends in institutions of learning (Bucos and Drăgulescu 2018). Machine learning has found application in studying the academic behaviour and performance prediction of students (Kostopoulos et al. 2018). The various areas of application of machine learning in education includes prediction of drop out and graduation potential (Ahuja and Kankane 2018; Nurhayati et al. 2018), prediction of academic performance (Roy and Garg 2018; Fernandes et al. 2019), assessment of the learning process (Khan and Ghosh 2018) and identification of learning risks, evaluation of administrators, evaluation of students' textual feedback (Ibrahim et al. 2019; Atta UR et al. 2018), assessment of the interactions among the educational stakeholders, as a pilot for guiding the implementation and integration of institution-based educational technologies (Angeli et al. 2017), and so forth (Rodrigues et al. 2018). Data-driven learning is the order of the day, and more research works are being dedicated toward improving knowledge in this field. In educational data mining, the accuracy of the algorithm and the model implemented depend on the nature, and the size of the data under study (Almarabeh 2017). Educational data mining and data mining generally, are strongly data-driven, and as such, issues of privacy and consent may arise with respect to the potential commercial value of data (Lynch 2017). Hence, adequate modalities for managing such must be ensured.

3 Related studies

The study by (Baepler and Murdoch 2010) identified a distinction between academic analytics and data mining. Academic analytics is a hypothesis driven process which involves the use of dataset for solving an academic problem while data mining is a speculative process for identifying wealth of useful information from seemingly insensible information without any predetermined hypothesis. According to (Bharara et al. 2018), learning analytics entails the application of analytic tools for studying students, the collection and analysis of student's virtual learning platform interaction-related data, while educational data mining is a computer-based learning approach for detecting unique patterns in student related dataset for identifying new findings and for testing theories.

A predictive analysis was performed by (Ahmed and Elaraby 2014) using 1548 student sample records from 2005 to 2010. In the study, features such as lab test grade, midterm grade, scored level of student participation etc. were applied using a decision tree to identify students that are likely to fail. K-means clustering algorithm was applied by (Bhise et al. 2013) to predict student result using student database records, while in the study by (Yadav et al. 2012), the potential of a student to drop out was predicted using decision tree classifiers. Using selected features from students' database, the performance of students was classified by (Kaur and Kaur 2018) using six different data mining algorithms; Naïve Bayes, K-NN, 1-NN and Decision Tree (C5.0, C4.5, CART). By applying knowledge discovery techniques using data mining, the study by (Burgos et al. 2018) evaluated student records from e-learning platforms for students taking distance learning courses using predictive models. After model implementation, a teaching plan was developed and deployed, and the teaching plan was able to reduce student dropout rate by 14% as compared with previous academic sessions.

In the study by (Bucos and Drăgulescu 2018), educational data mining was carried out using 908 pre-processed student data samples and five classification algorithms. The features considered include the course information, student attendance, average scores, level of student activity and the number of credits passed in the preceding session. The use of correlation and multiple linear regression analysis for predicting the graduation CGPA of students using their GPA for a period of five years was demonstrated by (Pelumi Oguntunde et al. 2018). In (Sivakumar and Selvaraj 2018), the performance of student was predicted using multiple supervised classifiers for performance comparison. In the study, four categories of performance ranging from excellent to poor were identified. Using 300 student sample records from 3 colleges, a predictive analysis of student performance was carried out by (Hussain et al. 2018) on WEKA platform by analysing 12 significant features using 4 classification algorithms.

4 Methodology

In this study, an analysis was carried out to show the extent of the relationship between the academic record of students of Covenant University in Nigeria at the point of admission, based on the university admission entry requirements, and the academic performance of the admitted students in the first year as measured by their 100 L CGPA and the class of grade using predictive data mining and regression models. Methods such as regression analysis and artificial neural networks have been deployed in this area (Arsad and Buniyamin 2014) but in this study, a data mining model was deployed on KNIME and Orange platform, and the veracity of the prediction was checked using regression analysis for performance comparison. Based on the dataset by (Odukoya et al. 2018), 1445 student records from 2005 to 2009 were analysed. The following features in the dataset were examined: the student's entry age, the aggregate WAEC score, the JAMB score, the university based CUSAS score, the first-year grade classification while the actual CGPA was considered for the regression analysis. These features represent the key requirements by the Nigerian University Commission (NUC) for admission into engineering programmes in Nigeria. The student grades fall into the following categories, first class (1st), second class lower (2|1), second class upper (2|2) and third class (3rd) respectively. The KNIME and Orange Analytic applications were run on Core ™ i3-7100 U CPU 2.4GHz, 6 GB Ram computer running Windows 10 operating system.

5 Descriptive statistics of the dataset

This section presents the descriptive statistics of the student dataset analysed in this study. Table 1 shows the descriptive statistics of the numeric variables in the dataset. In Figs. 1, 2, 3 and 4 an attempt was made to fit the data using 5 different distributions. Figures 1 (a) & 1 (b) present the cumulative probability and the probability density function plots for the CGPA data while Figs 2 (a) & (b) show the cumulative probability and the probability density function plots for the CUSAS score data for all the 1445 students. In Figs. 3 (a) & (b), the cumulative probability and the probability density function plots for the JAMB score are

Table 1 Descriptive statistics of the entry requirements

	Min	Max	Mean	Std. deviation	Variance	Skewness	Kurtosis
Entry Age	15	24	17.9232	1.2067	1.4560	0.9648	1.5256
JAMB Score	133	298	222.7723	23.2200	539.1690	−0.0798	−0.1973
CUSAS Score	41	110	69.4999	8.9183	79.5363	−0.0232	0.0136
CGPA	1.8	4.93	3.5853	0.6394	0.4088	−0.2462	−0.4884
WAEC	1.67	4.88	3.2566	0.6068	0.3682	0.0508	−0.6116

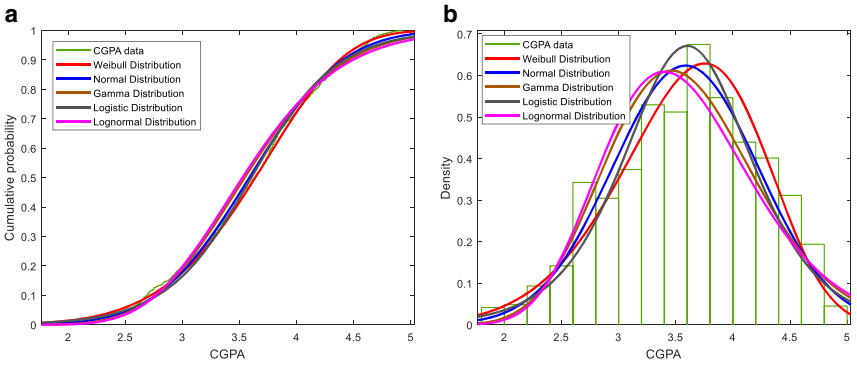

Fig. 1 CGPA data plot showing the (**a**) Cumulative probability (**b**) Probability density function

presented while Figs. 4 (a) & 4 (b) show the cumulative probability and the probability density plot of the Aggregate WAEC score. The data features analysed are presented as boxplots which categorised the student scores based on the class

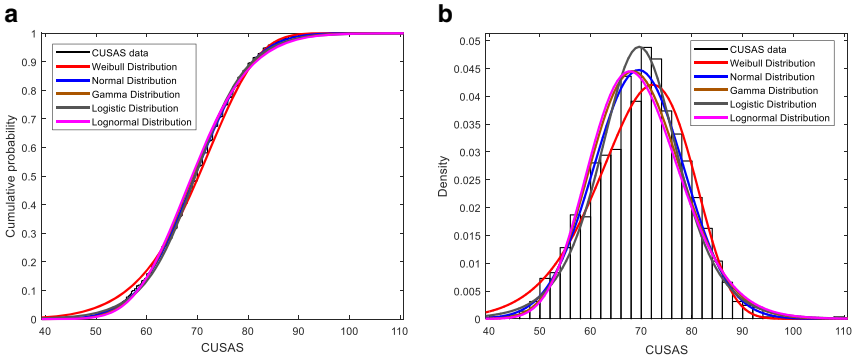

Fig. 2 CUSAS data plot showing the (**a**) Cumulative probability (**b**) Probability density function

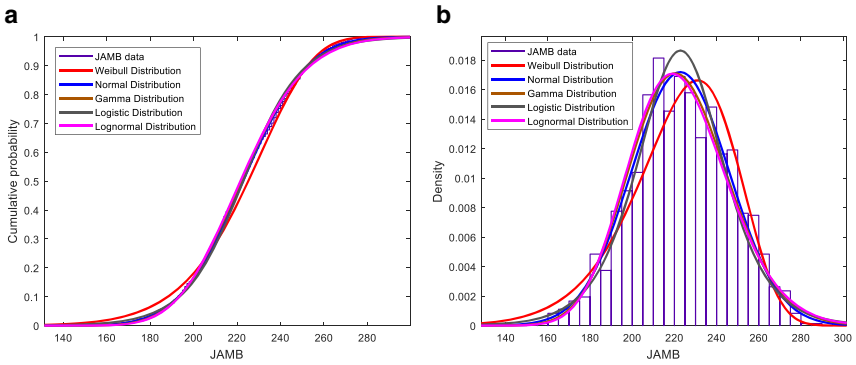

Fig. 3 JAMB data plot showing the (**a**) Cumulative probability (**b**) Probability density function

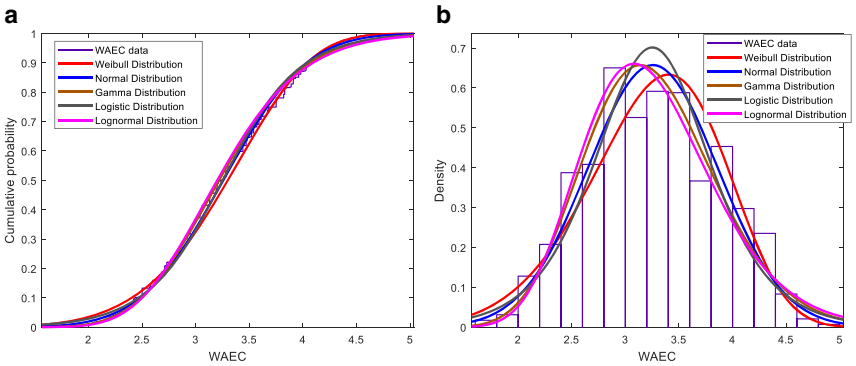

Fig. 4 WAEC data plot showing the (**a**) Cumulative probability (**b**) Probability density function

of grade (1st, 2|1, 2|2 and 3rd class) of the student for the first year. Figure 5 (a) presents the box plot of the CUSAS score, Fig. 5 (b) presents the JAMB score,

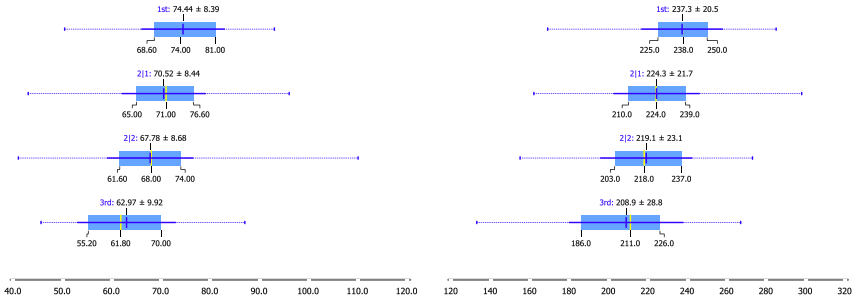

Fig. 5 Box plot showing data classification by grade class for (**a**) CUSAS score (**b**) JAMB score

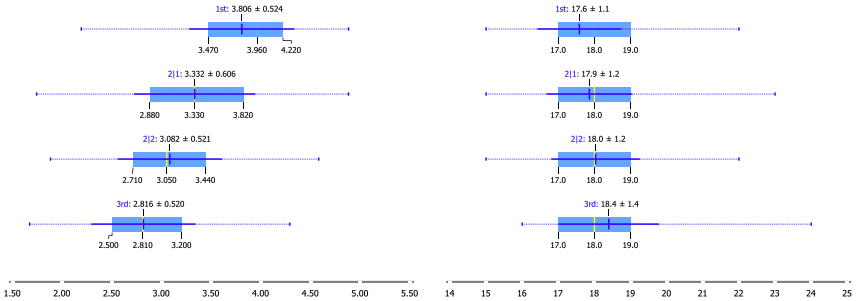

Fig. 6 Box plot showing data classification by grade class for (**a**) WAEC score (**b**) AGE score

Fig. 6 (a) shows the box plot of the aggregate WAEC score and Fig. 6 (b) shows the box plot of the AGE data feature.

6 The results of the predictive analysis using the KNIME platform

The predictive KNIME based model deployed in this study is shown in Fig. 7. The samples were selected using stratified sampling, and 70% of the data sample was used to train the model using 6 data mining algorithms; the Random Forest, the Tree Ensemble, the Decision Tree, the Naive Bayes, the Logistic Regression, and the Resilient back propagation (Rprop) Multi-Layer Perceptron algorithms. The remaining 30% of the samples were deployed for evaluating the performance of the model. Dimension reduction using principal component analysis was carried out to improve the model accuracy.

The following admission entry requirement features were applied in the data mining model; the student's entry age, the aggregate WAEC score, the JAMB score and the university based CUSAS score to predict the class of the student's first year grade. A comparative analysis of the performance of the six models is presented in Table 2. The Logistic Regression algorithm had the highest prediction accuracy of 50.23%, followed

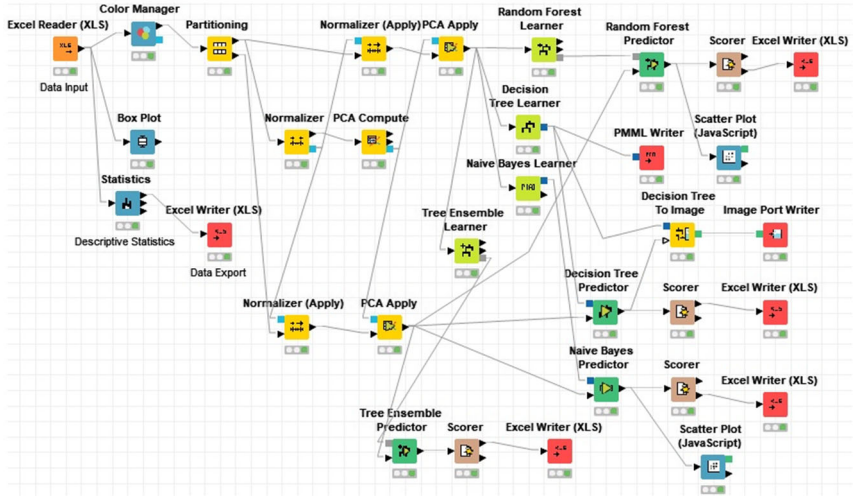

Fig. 7 The KNIME workflow

Table 2 Model performance comparison

	Random Forest	Tree Ensemble	Decision Tree	Naive Bayes	Logistic Regression	Rprop. MLP
Correct Classified	210	205	172	186	218	213
Accuracy	48.38%	47.24%	39.631%	42.857%	50.23%	49.078%
Cohen's Kappa (k)	0.09	0.071	0.0041	0.106	0.081	0.084
Wrong Classified	224	229	262	248	216	221
Error	51.613%	52.765%	60.368%	57.143%	49.77%	50.922%

by the Rprop MLP algorithm with an accuracy of 49.078% and this is followed by the Random Forest algorithm with 48.38% accuracy. Next in hierarchy is the Tree Ensemble algorithm with an accuracy of 47.24%, and this is followed by the Naive Bayes algorithm with an accuracy of 42.857%. The Decision Tree algorithm had the least accuracy of 39.631%. Table 2 shows comparatively the variations in the number of samples correctly and wrongly classified by the six algorithms.

The accuracy of the models is far lower than expected, and this expectation is because of the common assumption that the academic performance of a student based on entry requirements is a strong indicator of the performance of the student once in the university system. Although, this belief may not be totally wrong but the extent of its validity needs to be determined. For example, in Figs. 5 (a), 5 (b) and 6 (a) using the class of the first year grades (1st, 2|1, 2|2 and 3rd class) it can be seen from the box plots that the order of the highest scores in each group follows the order of the class of grades for the CUSAS, the JAMB and the WAEC scores respectively which implies that indeed there is a relationship between academic excellence based on admission requirements and the performance afterwards. Figure 6 (b) shows that there is no significant variation in the ages of students for all the class of grades. Table 3 compares the True Positive and False Positive predictions for the 6 data mining algorithms considered in the KNIME model.

7 The results of the predictive analysis using the Orange platform

To verify the veracity of the performance of the KNIME model a similar analysis was performed using the Orange data mining platform as shown in Fig. 8. Six data mining

Table 3 Prediction confusion of the three data mining predictors

	Random Forest		Tree Ensemble		Decision Tree		Naive Bayes		Logistic Regression		Rprop. MLP	
	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP
3rd	0	5	0	6	2	17	8	61	0	0	0	1
2|1	134	128	127	125	101	122	125	108	164	152	140	129
2|2	73	80	76	90	59	92	41	54	53	63	72	88
1st	3	11	2	8	10	31	12	25	1	1	1	3
Overall	210	224	205	229	172	262	186	248	218	216	213	221

TP True Positive, *FP* False Positive

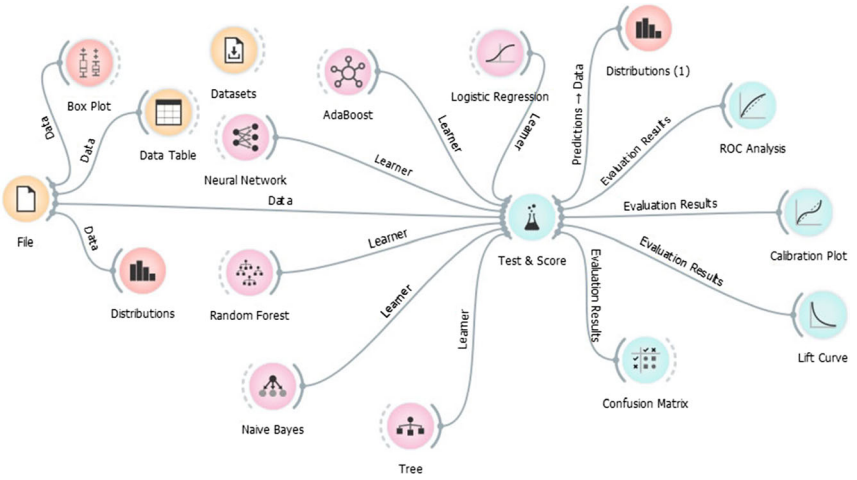

Fig. 8 The Orange model workflow

algorithms were considered for the supervised learning model, and these are: the Tree, the Random Forest, the Neural Network, the Naïve Bayes, the Logistic Regression and the AdaBoost algorithm respectively. The performance of each of the algorithm is presented comparatively in Table 4. The Neural Network had the highest classification

Table 4 Performance comparison for the data mining algorithms on the Orange platform

Method	AUC	(CA)	F1	Precision	Recall
Tree	0.552	0.436	0.436	0.439	0.436
Random Forest	0.607	0.503	0.484	0.489	0.503
Neural Network	0.635	0.519	0.494	0.486	0.519
Naïve Bayes	0.642	0.500	0.489	0.486	0.500
Logistic Regression	0.644	0.511	0.463	0.52	0.511
AdaBoost	0.531	0.428	0.429	`0.430	0.428

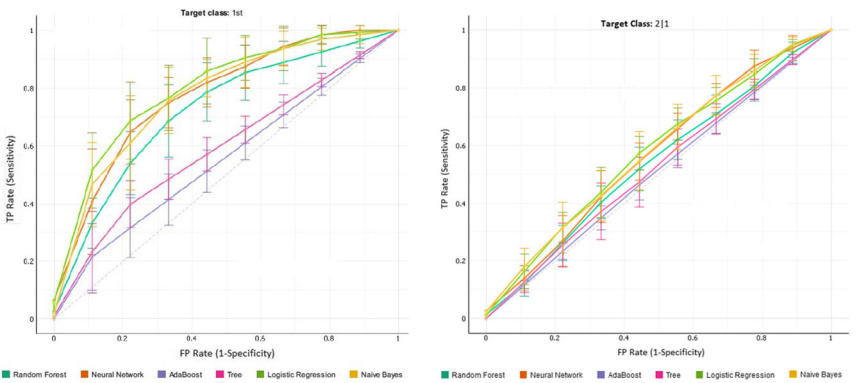

Fig. 9 The model sensitivity for (**a**) 1st Class grade (**b**) 2|1 grade

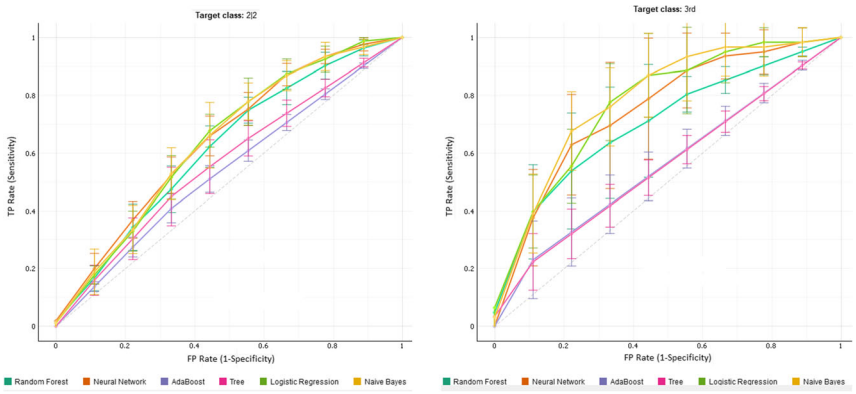

Fig. 10 The model sensitivity for (**a**) 2|2 class grade (**b**) 3rd class grade

accuracy of 51.9%, followed by the Logistic Regression with an accuracy of 51.1%, and this is followed by the Random Forest with an accuracy of 50.3%. The Naïve Bayes had the 4th best performance with an accuracy of 50% and next in hierarchy is the Tree algorithm with an accuracy of 43.6%. The AdaBoost algorithm had the least accuracy of 42.8%. Table 4 was obtained using the average over classes as the target class and a stratified 10-fold cross validation sampling type. Table 4 compares the Orange algorithms in terms of the Area under ROC Curve (AUC), the Classification Accuracy (CA), the F1 score, the Precision rate, and the Recall.

The sensitivity of each of the algorithm in the Orange model is shown in Figs. 9 (a), (b), 10 (a) and 10 (b) using the TP rate for the four class of grades i.e. first class, second class lower, second class upper and third class respectively.

8 Predictive analysis using regression

To further validate the accuracies of the data mining models, the dataset was further assessed using regression. Two regression models were considered, the linear regression and the quadratic regression models as presented in the following sections. In the regression model, the relationship existing between the cognitive entry requirements (the student's entry age (x1), the aggregate WAEC score (x2), the JAMB score (x3) and the university based CUSAS score (x4) and the first year CGPA (y) of the students was evaluated.

8.1 Linear regression model

The linear regression model which represents the relationship of the data analysed is shown in eq. 1. The results of the model are displayed in Table 5, and it can be observed that for the 1445 student sample data, the model R-squared value is 0.207 which indicates a weak relationship. The relationship of the linear regression model parameters is presented in eq. 1.

$$y = 1 + x_1 + x_2 + x_3 + x_4 \tag{1}$$

Table 5 Linear regression model results

	Estimate	SE	tStat	pValue
(Intercept)	1.6187	0.30209	5.3584	9.76E-08
×1	−0.03928	0.012715	−3.089	0.002047
×2	0.004029	0.000678	5.9463	3.44E-09
×3	0.010895	0.001817	5.996	2.55E-09
×4	0.3119	0.026534	11.755	1.58E-30

No. of observations: 1445, Error degrees of freedom: 1440

RMS Error: 0.57

R^2 : 0.207, Adjusted R^2 : 0.205

F-statistic vs. constant model: 94.2, p value = 3.01E-71

8.2 Quadratic regression model

The quadratic regression model which represents the relationship among the data features analysed is shown in eq. 2. The results of the model are displayed in Table 6, and it can be observed that for the 1445 student sample data analysed, the R-squared value is 0.232 which also indicates a weak relationship

Table 6 Quadratic regression model results

	Estimate	SE	tStat	pValue
(Intercept)	4.1111	3.9768	1.0338	0.30141
x_1	−0.23418	0.29097	−0.80484	0.42105
x_2	0.012516	0.014585	0.85812	0.39097
x_3	0.007849	0.041166	0.19067	0.84881
x_4	−0.61633	0.54023	−1.1409	0.25412
x_1:x_2	0.000208	0.000564	0.36847	0.71258
x_1:x_3	0.000888	0.00159	0.5588	0.57639
x_1:x_4	−0.0099	0.022109	−0.44799	0.65423
x_2:x_3	0.000236	7.82E-05	3.0195	0.002577
x_2:x_4	0.002636	0.00122	2.1617	0.030809
x_3:x_4	−0.00136	0.00329	−0.4119	0.68047
x_1^2	0.003301	0.006568	0.50266	0.61528
x_2^2	−8.20E-05	2.32E-05	−3.533	0.000424
x_3^2	−0.00046	0.000154	−2.9928	0.002811
x_4^2	0.092437	0.039376	2.3476	0.019033

No. of observations: 1445, Error degrees of freedom: 1430

RMS Error: 0.563

R^2 : 0.232, Adjusted R^2 : 0.224

F-statistic vs. constant model: 30.8, p value = 4.71E-72

considering the standard R-squared value range of 0 to 1. The relationship of the quadratic regression model parameters is presented in eq. 2.

$$y = 1 + x_1 \cdot x_2 + x_1 x_3 + x_1 x_4 + x_2 x_3 + x_2 x_4 + x_3 x_4 + {x_1}^2 + {x_2}^2 + {x_3}^2 + {x_4}^2 \quad (2)$$

9 Summary of results

In the KNIME model, the maximum accuracy is 50.23% and the lowest accuracy is 39.631% whereas for the Orange model the accuracy varied from 51.9% to 42.8%. The performance of both models is quite close, which implies that the data mining algorithms on both platforms were able to identify the extent of the hidden relationship between the cognitive admission entry requirements and the class of grade of the students in their first year. With a maximum 51.9% accuracy for both data mining models; it implies that the students' entry age, the aggregate WAEC score, the JAMB score and the university based CUSAS score which are the key admission entry criteria into 100 L programmes, only partially explain the performance (class of grade) of the admitted student in their first year at the university. This means that other non-academic factors and attributes of each admitted student which may not be easy to measure at the point of admission, coupled with the personal lifestyle and struggles while in the university are potential determinants of the academic performance of students in their first year. The linear regression model had a F-statistic of 94.2 which is significant at a p value of 3.01E-71 but with a R^2 value (coefficient of determination) of 0.207; it implies that the admission entry criteria only explains 20.7% of the 100 L CGPA variation. A similar result was observed for the quadratic regression model with a R^2 value of 0.232 which indicates a weak predictive relationship.

The results of the KNIME and Orange data mining models, and the regression analysis carried out in this study show that there is a relationship between the admission entry requirements, and the academic performance of the admitted student in their first year but this relationship is not very strong as revealed by the accuracies of the predictive algorithms. Selection criteria for admission into the university in Nigeria are mostly based on the academic profile of applicants. The result of this study therefore emphasizes the need to re-evaluate the admission criteria and the need to consider other non-academic factors that may be indicators of student dedication, motivation and passion for success. Excellence in sports, leadership experiences, participation in student projects at secondary school level etc. are likely factors that should also be considered in the admission process. Apart from the natural effects of an individual's academic prowess and learning skills on their academic excellence in a university, other factors such as the level of attendance at lectures, number of sick days, family issues at home that may affect a student's psychological state, financial challenges, average scores in the first set of tests, social lifestyle, availability of support programmes within the university e.g. counselling and so forth also play key roles in the overall performance of a student.

For the most part, a well-executed selection process usually yields a good percentage of academically sound scholars (Idachaba 2018a), but other factors like teaching and

research, the quality of facilities and the general institutional management (Idachaba 2018b), can greatly affect academic performance in universities. Other factors that could affect academic performance in universities include administrative bottlenecks and the flaunting of laid down policies and ground rules. Ensuring academic excellence for students in their first year of study requires a holistic approach that encompasses the selection of academically promising applicants, provision of adequate facilities, staff and faculty; availability of purpose-fit student orientation, counselling and mentorship programmes, availability of a proactive performance evaluation system, and so forth.

10 Conclusion and future scope

The use of various selection criteria by regulatory bodies in Nigeria and Nigerian universities has led to the admission of a higher percentage of academically sound and aspiring undergraduates, and a concentration of great talents in Nigerian universities. It has also helped to sift out a higher percentage of unqualified candidates and reduce the rate of underhand admission practices in Nigerian universities, although there are existing opportunities for further improvement in admission procedures and practices.

Admission entry requirements are vital indices for selecting students for admission in higher institutions. The analysis in this study reveals the extent to which the admission entry requirements can be used as performance predictors once in the university system, using Covenant University in Nigeria as a case study. Data trends presented in box plots show that the WAEC, CUSAS and JAMB entry scores of students are higher for those who had 1st class grade, followed by those on 2|1, and then 2|2 and finally students on 3rd class had the least average scores. The educational data mining carried out using KNIME and Orange platform had maximum accuracies of 50.23% and 51.9% respectively, and this shows that the cognitive entry characteristics of a student as defined by the admission entry requirements, does not fully explain the eventual first year class of grade of that student. To validate the results of the data mining models, linear and quadratic regression analytical models were developed and R-squared values of 0.207 and 0.232 were observed respectively, indicating a weak relationship between the first year CGPA and the admission entry requirements. This therefore makes it vital to ensure that admission selection criteria are robust enough to allow for a well-mixed breed of students with varying unique strength, not necessarily academic that may eventually translate to success once in the university.

At present, the analysis so far has focused generally on the relationship of the admission entry criteria and the 100 L performance of the student. It will be interesting to consider classifying the student records based on their demographic attributes such as gender, family income level, religion, average family size, state of origin etc. in a bit to identify new trends and hidden knowledge. The effect of flipped classroom approach to student learning can also be measured in a particular academic session by classifying students after admission for a case-control study based on the admission criteria, and then selecting study and control samples from each group to be differently subjected to the traditional and the flipped classroom teaching models. This will allow for a comparative analysis of the performance of the study and control students' CGPA and class of grade at the end of the academic

year, so as to identify if their will be a significant difference between the performance of students exposed to the flipped classroom model, and those subjected to the traditional teaching model.

Acknowledgements The Authors appreciate Covenant University Centre for Research, Innovation and Development for the commitment to innovative research, and for providing an enabling research environment.

References

Adeogun, A., Subair, S., & Osifila, G. (2009). Deregulation of university education in Nigeria: Problems and prospects. *Florida Journal of Educational Administration & Policy, 3*, 1–8.

Adeyemi, K. (2001). Equality of access and catchment area factor in university admissions in Nigeria. *Higher Education, 42*, 307–332.

Agarwal, S., Pandey, G., & Tiwari, M. (2012). Data mining in education: Data classification and decision tree approach. *International Journal of e-Education, e-Business, e-Management and e-Learning, 2*, 140.

Ahmed, A. B. E. D., & Elaraby, I. S. (2014). Data mining: A prediction for Student's performance using classification method. *World Journal of Computer Application and Technology, 2*, 43–47.

Ahuja, R. & Kankane, Y. (2018). Predicting the probability of student's degree completion by using different data mining techniques. 474–477.

Ahuja, R., Jha, A., Maurya, R. & Srivastava, R. (2019). Analysis of educational data mining. 4th International Conference on Harmony Search, Soft Computing and Applications, ICHSA 2018, Gurgaon; India. Advances in Intelligent Systems and Computing, 897–907.

Aina, O. I. 2002. Alternative modes of financing higher education in Nigeria and the implications for university governance. *Africa Development/Afrique et Développement*, 236–262.

Ajadi, T. O., Salawu, I. O. & Adeoye, F. A. (2008). E-learning and distance education in Nigeria. Online Submission, 7.

Ajayi, I., & Ekundayo, H. T. (2008). The deregulation of university education in Nigeria: Implications for quality assurance. *Nebula, 5*, 212–224.

Almarabeh, H. (2017). Analysis of Students' performance by using different data mining classifiers. *International Journal of Modern Education and Computer Science, 9*, 9.

Aluede, O., Idogho, P. O. & Imonikhe, J. S. (2012). Increasing access to university education in Nigeria: Present challenges and suggestions for the future. The African Symposium. 3–13.

Angeli, C., Howard, S. K., Ma, J., Yang, J., & Kirschner, P. A. (2017). Data mining in educational technology classroom research: Can it make a contribution? *Computers and Education, 113*, 226–242.

Arsad, P. M. & Buniyamin, N. (2014). Neural Network and Linear Regression methods for prediction of students' academic achievement. Global Engineering Education Conference (EDUCON), IEEE, 2014. IEEE, 916–921.

Atta UR, R., Sultan, K., Aldhafferi, N., & Alqahtani, A. (2018). Educational data mining for enhanced teaching and learning. *Journal of Theoretical and Applied Information Technology, 96*, 4417–4427.

Azevedo, A. (2018). Data mining and knowledge discovery in databases. In *Encyclopedia of information science and technology, fourth edition*. IGI Global.

Babalola, J. B. (1998). Cost and financing of university education in Nigeria. *Higher Education, 36*, 43–66.

Baepler, P., & Murdoch, C. J. (2010). Academic analytics and data mining in higher education. *International Journal for the Scholarship of Teaching and Learning, 4*, 17.

Bharara, S., Sabitha, S., & Bansal, A. (2018). Application of learning analytics using clustering data Mining for Students' disposition analysis. *Education and Information Technologies, 23*, 957–984.

Bhise, R., Thorat, S., & Supekar, A. (2013). Importance of data mining in higher education system. *IOSR Journal Of Humanities And Social Science (IOSR-JHSS), 6*, 18.

Bucos, M., & Drăgulescu, B. (2018). Predicting student success using data generated in traditional educational environments. *TEM Journal, 7*, 617–625.

Burgos, C., Campanario, M. L., Peña, D. D. L., Lara, J. A., Lizcano, D., & Martínez, M. A. (2018). Data mining for modeling students' performance: A tutoring action plan to prevent academic dropout. *Computers and Electrical Engineering, 66*, 541–556.

Burke, A. S., & Fedorek, B. (2017). Does "flipping" promote engagement?: A comparison of a traditional, online, and flipped class. *Active Learning in Higher Education, 18*, 11–24.

Daradoumis, T., Marquès Puig, J. M., Arguedas, M., & Calvet Liñan, L. (2019). Analyzing students' perceptions to improve the design of an automated assessment tool in online distributed programming. *Computers and Education, 128*, 159–170.

Ebbeler, J. (2013). Introduction to ancient Rome,' the flipped version. *The Chronicle of Higher Education* [online], 59. Available: https://www.chronicle.com/article/Introduction-to-Ancient/140475/.

Fernandes, E., Holanda, M., Victorino, M., Borges, V., Carvalho, R., & Erven, G. V. (2019). Educational data mining: Predictive analysis of academic performance of public school students in the capital of Brazil. *Journal of Business Research, 94*, 335–343.

Hussain, S., Dahan, N. A., Ba-Alwib, F. M., & Ribata, N. (2018). Educational data mining and analysis of students' academic performance using WEKA. *Indonesian Journal of Electrical Engineering and Computer Science, 9*, 447–459.

Hussain, S., Atallah, R., Kamsin, A. & Hazarika, J. (2019). Classification, clustering and association rule mining in educational datasets using data mining tools: A case study.

Ibrahim, Z. M., Bader-El-Den, M. & Cocea, M. (2019). Mining unit feedback to explore students' learning experiences. *Advances in Intelligent Systems and Computing.*

Idachaba, F. (2018a). Development of a rapid mentoring scheme for managing large classes in engineering departments. INTED2018 Conference, 5th–7th March 2018 2018a Valencia, Spain. 5725–5731.

Idachaba, F. (2018b). Outcome based engineering curriculum design: a system for curriculum streamlining and graduate quality improvement in engineering. INTED2018 Conference, 5th–7th March 2018 2018b Valencia, Spain. 5888–5893.

Kaur, N., & Kaur, J. (2018). Performance evaluation of data mining classification in educational system using genetic algorithm. *International Journal of Advanced Science and Technology, 114*, 127–138.

Khan, A., & Ghosh, S. K. (2018). Data mining based analysis to explore the effect of teaching on student performance. *Education and Information Technologies, 23*, 1677–1697.

Khedr, A. E., & El Seddawy, A. I. (2015). A proposed data mining framework for higher education system. *International Journal of Computer Applications, 113*, 24–31.

Kim, D., Yoon, M., Jo, I. H., & Branch, R. M. (2018). Learning analytics to support self-regulated learning in asynchronous online courses: A case study at a women's university in South Korea. *Computers and Education, 127*, 233–251.

Kostopoulos, G., Kotsiantis, S., Pierrakeas, C., Koutsonikos, G., & Gravvanis, G. A. (2018). Forecasting students' success in an open university. *International Journal of Learning Technology, 13*, 26–43.

Kurt, G. (2017). Implementing the flipped classroom in teacher education: Evidence from Turkey. *Educational Technology and Society, 20*, 211–221.

Lynch, C. F. (2017). Who prophets from big data in education? New insights and new challenges. *Theory and Research in Education, 15*, 249–271.

Nikolic, S. & Nicholls, B. (2018). Exploring student interest of online peer assisted learning using mixed-reality technology.

Nurhayati, O. D., Bachri, O. S., Supriyanto, A., & Hasbullah, M. (2018). Graduation prediction system using artificial neural network. *International Journal of Mechanical Engineering and Technology, 9*, 1051–1057.

Odukoya, J. A., Popoola, S. I., Atayero, A. A., Omole, D. O., Badejo, J. A., John, T. M., & Olowo, O. O. (2018). Learning analytics: Dataset for empirical evaluation of entry requirements into engineering undergraduate programs in a Nigerian university. *Data in Brief, 17*, 998–1014.

Oguntunde, P., Okagbue, H., Oguntunde, O. A., & Opanuga, A. (2018). Analysis of the inter-relationship between students' first year results and their final graduating grades. *International Journal of Advanced and Applied Sciences, 5*, 1–6.

Olsson, M. & Mozelius, P. (2016). On design of online learning environments for programming education. Proceedings of the European Conference on e-Learning, ECEL. 533–539.

Osmanbegović, E. & Suljic, M. (2012). Data mining approach for predicting student performance.

Popoola, S. I., Atayero, A. A., Badejo, J. A., Odukoya, J. A., Omole, D. O., & Ajayi, P. (2018). Datasets on demographic trends in enrollment into undergraduate engineering programs at Covenant University, Nigeria. *Data in Brief, 18*, 47–59.

Rodrigues, M. W., Isotani, S., & Zárate, L. E. (2018). Educational data mining: A review of evaluation process in the e-learning. *Telematics and Informatics, 35*, 1701–1717.

Roy, S. & Garg, A. (2018). Predicting academic performance of student using classification techniques. 2017 4th IEEE Uttar Pradesh section international conference on electrical, computer and electronics, UPCON 2017. 568–572.

Ryan, S. J. D., & Baker. (2010). Data mining for education. *International encyclopedia of education, 7*, 112–118.

Saint, W., Hartnett, T. A., & Strassner, E. (2003). Higher education in Nigeria: A status report. *Higher Education Policy, 16*, 259–281.

Senthil, S. & Lin, W. M. (2018). Applying classification techniques to predict students' academic results. 2017 IEEE International Conference on Current Trends in Advanced Computing, ICCTAC 2017. 1–6.

Sivakumar, S. & Selvaraj, R. 2018. Predictive modeling of students performance through the enhanced decision tree. *Lecture Notes in Electrical Engineering.*

Tair, M. M. A., & El-Halees, A. M. (2012). Mining educational data to improve students' performance: A case study. *International Journal of Information and Communication Technology Research, 2*, 140–146.

Van WYK, M. M. (2018). Economics student teachers' views on the usefulness of a flipped classroom pedagogical approach for an open distance eLearning environment. *International Journal of Information and Learning Technology, 35*, 255–265.

Yadav, S. K., Bharadwaj, B. & PAL, S. (2012). Mining education data to predict student's retention: A comparative study. *arXiv preprint arXiv:1203.2987*.