Research Article

Toward an improved learning process: the relevance of ethnicity to data mining prediction of students' performance

Aderibigbe Israel Adekitan[1] · Odunayo Salau[2]

Received: 20 August 2019 / Accepted: 21 November 2019 / Published online: 2 December 2019

Abstract

The ability to predict failure is an advantageous educational tool that can be effectively used to counsel student, and this may also be used as a tool for developing, and channelling adequate academic interventions toward preventing failure and dropout tendencies. Students are generally admitted based on their evaluated academic potentials as measured using their admission criteria scores. This study seeks to identify the relationship, if any, between the admission criteria scores and the graduation grades, and to examine the influence of ethnicity using the geopolitical zone of origin of the student on the predictive accuracy of the models developed using a Nigerian University as a case study. Data mining analyses were carried out using four classifiers on the Orange Software, and the results were verified with multiple regression analysis. The maximum classification accuracy observed is 53.2% which indicates that the pre-admission scores alone are insufficient for predicting the graduation result of students but it may serve as a useful guide. By applying over-sampling technique, the accuracy increased to 79.8%. The results establish that the ethnic background of the student is statistically insignificant in predicting their graduation results. Hence, the use of ethnicity in admission processes is therefore not ideal.

Keywords Educational data mining · Nigerian university · Data mining algorithms · Performance evaluation methodologies · Knowledge discovery · Machine learning

1 Introduction

Education is generally said to be the bedrock of national growth and national integration, but studies over the years have shown that the relationship between national integration and education is not linear [1]. Higher education in Nigeria began in 1932 with the establishment of the Yaba Higher College [2], and in 1940, the University College at Ibadan was established in an effort to promote higher education within the colony. According to Okebukola [3] as quoted by Olujuwon [2], regional universities were created based on the recommendations of the Asbby commission; the University of Nigeria was established in the east in 1960, University of Ife in 1961, and in 1962, the University College at Ibadan was granted a full university status. In the North, the Ahmadu Bello University was established in 1962. This marked the beginning of region-based education in Nigeria, and as at today there are 165 Universities in Nigeria according to the National Universities Commission, 43 of which are federal universities, 47 state-owned universities and 75 private universities with only 29% of Joint Admissions and Matriculation Board (JAMB) applicants admitted into the university [4] due to the limited admission slots [5].

Nigeria, with a land area of 923,768 km^2 and an estimated population of over 190 million [7] has thirty-six (36)

✉ Aderibigbe Israel Adekitan, aderibigbe-israel.adekitan@tu-ilmenau.de | [1]Department of Electrical Engineering and Information Technology, Technische Universität Ilmenau, Ilmenau, Germany. [2]Department of Business Management, Covenant University, Ota, Ogun State, Nigeria.

SN Applied Sciences (2020) 2:8 | https://doi.org/10.1007/s42452-019-1752-1

states which are divided into six (6) geopolitical zones, namely; South South, South West, South East, North East, North West, and North Central as shown in Fig. 1. The states were aggregated based on their cultural similarities, ethnicity and common history [8, 9]. With more than two hundred and fifty ethnic groups and over 500 languages in Nigeria, the government had to develop a model for ensuring adequate allocation of political, economic, and educational resources across the regions, and this was achieved by grouping into states and geopolitical zones [10]. Overtime, it became evident that the reception and value for education varies across the geopolitical zones in Nigeria.

According to Ukiwo [1], access to education in Nigeria has been politicised due to the plural nature of the society, and this has a tendency to engender economic inequalities. Nigeria is majorly divided along ethnic and religious lines, and these factors strongly influence how government policies and efforts generally, are perceived and interpreted. The inequality and politics of higher education started in the early 90's with the intense pursuit of educational development and attainment in the south west region of the country while the Northern region was quite laid back. To put this in perspective, according to Abernethy [11] as quoted by Ukiwo [1], although the North had about 55% of the national population in 1912, 1926 and 1937, enrolment into primary schools stood at 950, 5200 and 20,250 in the North respectively while in the South, enrolment figures were 35,700, 138,250 and 218,600 and this indicates a significant disparity in regional educational status. The Eastern and Western region of the country achieved drastic growth in educational attainment as compared with the Northern region at the birth of the nation Nigeria. The old Northern region of Nigeria which was educationally laid back, now comprises three geopolitical zones and these are: North West (NW), North East (NE), and North Central (NC) while the former Eastern region and Western region in the South that were developed educationally make up the remaining three geopolitical zones.

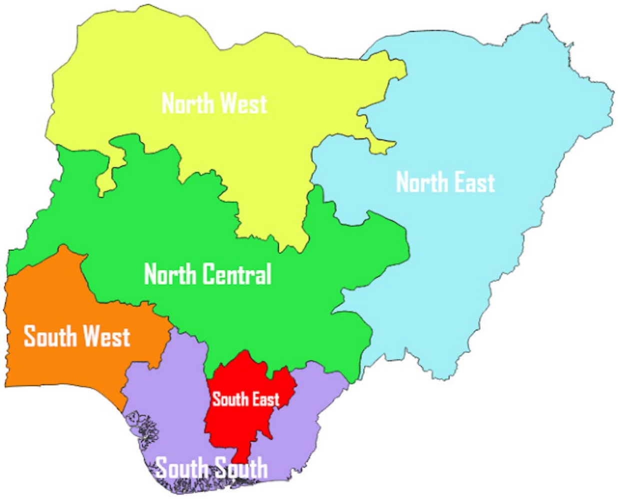


Fig. 1 The six geopolitical zones in Nigeria [6]

Primary and secondary education are vital for tackling illiteracy issues and for ensuring national development. In order to develop a total man equipped with adequate knowledge and capacity to handle today's societal problems and developmental needs, higher education is required [12]. With a focus on higher education, the National Universities Commission was setup in 1962 as an advisory agency, and it became a statutory body under Decree 1 of 1974 charged with the responsibility of ensuring adequate development and regulation of university education in Nigeria. Increasing demand for university education has created a daunting task for regulatory bodies in a bid to ensure increasing admission slots, and at the same time, enforce quality education with the reality of the ever-present national challenges such as inadequate finance, insufficient educational facilities, unavailable material resources and variances in regional educational status.

Regional differences in the quality and acceptance of education have been drastically reduced over the years due to various deliberate regulatory efforts, but evidences of it still exist in the Nigerian education sector. Admission into institutions in Nigeria is often influenced by the geopolitical zone of the applicants; some states in Nigeria are considered as educationally disadvantaged while some universities have catchment areas, and the cut-off mark (admissible score) for students from this regions are set lower than the general cut-off mark for students from other regions, even though they all sat for the same entry examination. Likewise, government related recruitment are also polarised with deliberate efforts put in place to ensure that successful applicants are selected from all the geopolitical zones. To achieve this, sometimes the pass mark for some regions is deliberately reduced as compared to others. These practices create conflict of opinions, while some Nigerians are pleased that it ensures fair sharing of opportunities among all the geopolitical zones, others think it is unfair and does not ensure success of the best candidates, and is therefore not the best option if Nigeria desires to attain its full human resource potential, and also maximize her entrepreneurial capability and governmental performance [13].

Does the geopolitical zone of origin of students affect their academic performance in higher institutions? In this study, the effect of the geopolitical zone from which a student originates on the anticipated graduation result of the student is examined using a number of selected admission criteria. The dataset analysed in this study is from a private Nigerian university with broad admission criteria that enable students from all the geopolitical zones of the country to fairly compete for the available admission slots. Predictive analyses using data mining on Orange software and regression models were carried out to identify hidden knowledge and vital statistical trends for students from the six geopolitical zones in Nigeria, in order to understand the impact, if any, of ethnicity [14] on the prediction accuracy of the graduation CGPA of University students in Nigeria using selected features.

2 Background

Data related research is on the increase due to the enormous potential benefit in terms of knowledge acquisition and application [15, 16]. Educational data mining is the application of data mining methodologies in educational-data related research studies toward solving education related issues [17]. It entails the stepwise extraction of hidden and useful information from a dataset [18] generated within the education sector in a bid to further understand students, and the effectiveness of the learning process. Educational data mining converts seemingly meaningless data into useful knowledge that can greatly impact the practices and regulatory methods within the education domain. Generally, educational data mining comprises data collection, data sorting and pre-processing, data mining, and post processing of data mining results. Some of the data mining techniques deployed includes clustering, text mining, association rule mining, classification and so forth. The mode of delivery of education has transcended from the traditional classroom-based method to various web-based teaching platforms and adaptive e-learning systems [19] due to the availability of modern learning technologies [20–22], and this therefore enables diverse education related data to be generated, logged, processed and studied for a better understanding of the learning process and learners [23], and educational technology integration practices [24].

In a higher institution, various types of data can be collected with different levels of relationship and hierarchy, and as such, the type of knowledge that can be mined from an educational dataset is a function of the nature and the origin of the data. Educational data mining can be used for different objectives [17]. Educational data mining helps to understand the relationship between the educators and the students, it reveals weaknesses and gaps in the learning process, it can be used to predict potential for negative student behaviours [23], and for predicting dropout potential and students performance [25, 26], it aids the development and review of learning models, it can be used to measure the effectiveness of any intervention deployed, and may also be used to guide the learning efforts of learners. The quality and effectiveness of decision processes can be greatly enhanced using educational data mining [27], and likewise, vital feedbacks from students can also be evaluated using data mining techniques in order to identify lapses, areas of need and improvement in teaching and learning processes. Through data mining techniques, students can be classified into unique groups based on well-defined criteria to enable the deployment of purpose-specific and targeted learning interventions, and for identifying common skill set, social attitudes, learning behaviours [28, 29] and interests [30, 31]. The effectiveness of academic modules and the developments of new contents can also be evaluated using data mining techniques [32].

2.1 Related literatures

In the study by Hussain et al. [33], 24-attribute based dataset was evaluated using Bayes Network, Random Forest, J48 and PART classifiers on WEKA data mining platform to predict the semester performance of students. A predictive Multi-Layer Perceptron model was developed by Nurhayati et al. [34] for evaluating the student record of 292 students by using 5 features to predict the graduation potential of the students. In Adekitan and Salau [35] the data of 1445 students covering academic sessions from 2005 to 2009 was evaluated using data mining techniques to determine the extent of the correlation between the admission selection scores and the scholarly performance of student in their first academic year. Similarly, the performance of students in their first year was predicted by Ahmad, Ismail [36] using rule-based, decision tree (J48) and naïve bays classifiers to data mine 399 student records. Using dataset generated at a Bulgarian university, the study by Kabakchieva [37] demonstrated the use of data mining techniques for enhancing university management decision making process by extracting knowledge from 10,330 students' record with 20 features using the WEKA software.

The study by Alharbi et al. [38] carried out data mining analysis for early identification of at-risk students using the admission records and performance in their first year of study. The results of student-failure potential analysis create an opportunity for warning at-risk student of a potential failure early enough so that drastic intervention can be deployed [39]. In the study by Atta Ur et al. [40], the level of acceptance of course time table and teaching methods by student was measured using machine learning algorithms, by administering an investigative questionnaire consisting of 38 teaching and learning related questions. Learning analytics can be defined as the measurement, gathering, investigation and reporting of relevant data about student, and the learning process, and methods [41, 42]. The research by Bharara et al. [43] identified new metrics that are relevant to the learning process in order to develop a robust model for evaluating student performance. Using 4 categories of features; these are interactional, academic, the level of parent's participation in

Table 1 Descriptive statistics of the numeric data features

	Min	Mean	Median	Max	SD	Skewness	Kurtosis
JAMB score	66	219.1575	220	304	26.7225	−0.5226	1.2657
WAEC aggregate	1.2847	3.0723	3.0357	4.9306	0.5969	0.2038	−0.3589
CGPA	1.62	3.4144	3.45	4.96	0.6785	−0.135	−0.6645

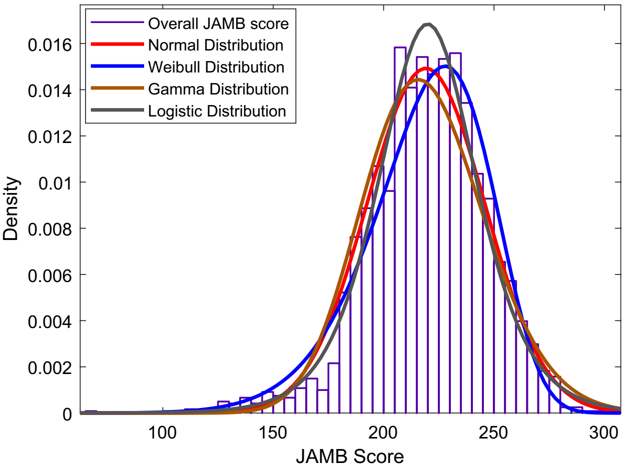


Fig. 2 Probability density function plot of JAMB score

the education of their children, and demographic features. In the study, student dataset containing 500 samples was analysed using clustering data mining methodologies.

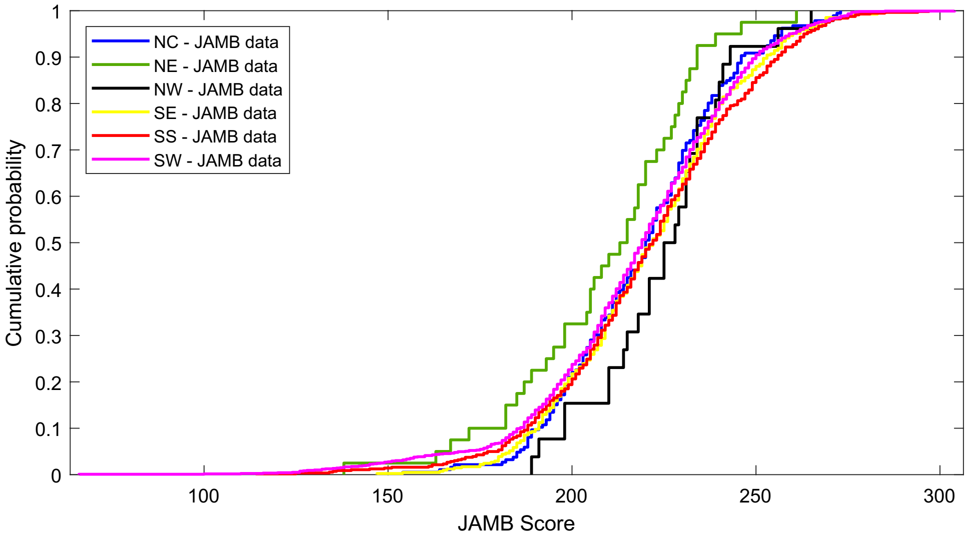
3 Data descriptive statistics

Statistical attributes of 2413 student dataset across the 4 colleges in Covenant university is presented in this section. Table 1 shows the descriptive statistics of the JAMB Score, WAEC Aggregate and the CGPA of the 2413 students. Figures 2 and 3 show the probability density function plot of the JAMB score and the cumulative probability plot of the JAMB score. Figures 4 and 5 present the probability density function plot and the cumulative probability plot of the WAEC aggregate score respectively, while Figs. 6 and 7 show the probability density function plot and the cumulative probability plot of the CGPA of the students at graduation respectively. To show the variations in the graduation CGPA of students across the six geopolitical zones, the CGPA data is presented as a box plot in Fig. 8, while in Fig. 9 the box plot shows the CGPA variations across the four colleges. As shown in Fig. 8, the North east and North west students had the lowest average CGPA of 3.369 and 3.366 respectively. From Fig. 9, it was observed that students from the college of engineering had the highest average CGPA of 3.52 while the college of science and technology had the least average CGPA of 3.34. Figure 10 presents the distribution of the number of student per grade classification for the six geopolitical zones.

4 Methodology

The dataset sourced from Badejo et al. [14] contains 2413 student records as generated by the students' records department of Covenant University with the support of Covenant University Centre for Systems and Information

Fig. 3 Cumulative probability plot of the JAMB score per geopolitical zone



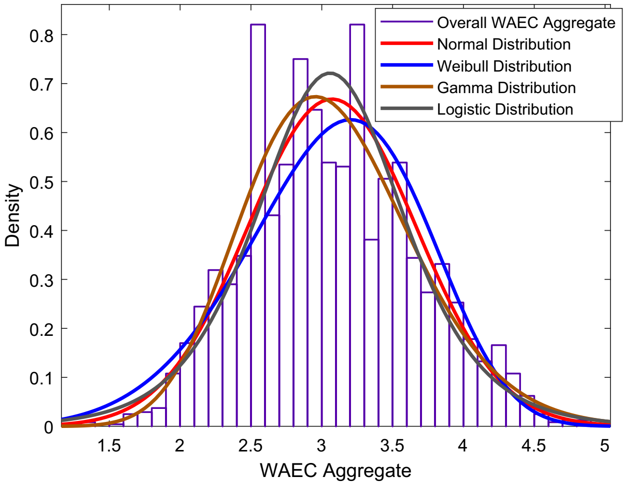


Fig. 4 Probability density function plot of WAEC aggregate

Services. The data contains records of students across the 6 geopolitical zones in Nigeria i.e. South East (SE), South West (SW), South South (SS), North West (NW), North East (NE), and North Central (NC). The data span over a 5-year period from 2008 to 2013 graduation set, across the four colleges of the University i.e. College of Leadership and Development Studies (CLDS), College of Business and Social Sciences (CBSS), College of Science and Technology (CST), and College of Engineering (COE). The dataset contains the year of graduation, the geopolitical zone, the college, pre-admission scores for the West African Examination Council (WAEC) examination and the Joint Admission Matriculation Board (JAMB), and also the Cumulative Grade Point Average (CGPA) at the point of graduation. The class of grade at the point of graduation was added to the features, and the colleges were coded numerically as 1, 2, 3, and 4 respectively, while the geopolitical zones were coded as 1, 2, 3, 4, 5 and 6 respectively. WAEC conducts the West African Senior School Certificate Examination for west African countries. The examination board was established in 1952 for conducting the final examination for graduating high school students.

In this study, the student record was analysed using Orange data mining software as discussed in detail in Sect. 5.1, and regression analysis as discussed in Sect. 5.3. For the two models, the predictive analysis was done in two folds: in the first mode, the geopolitical zone of each student was considered and in the second mode the geopolitical zone was skipped in order to identify if there is any significant variation in model performance as a result of the geopolitical zone feature. Also, statistical analysis and various plots were also developed to show data trends, and variations among the geopolitical zones. For the data mining analysis, the class of grade was used as the target while the CGPA was ignored, whereas for the

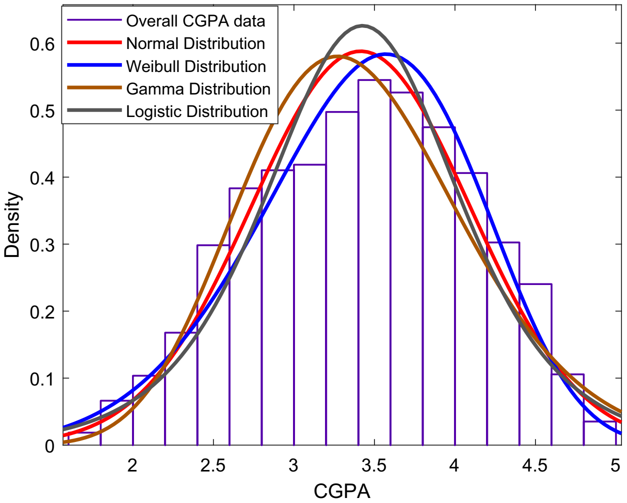


Fig. 6 Probability density function plot of CGPA

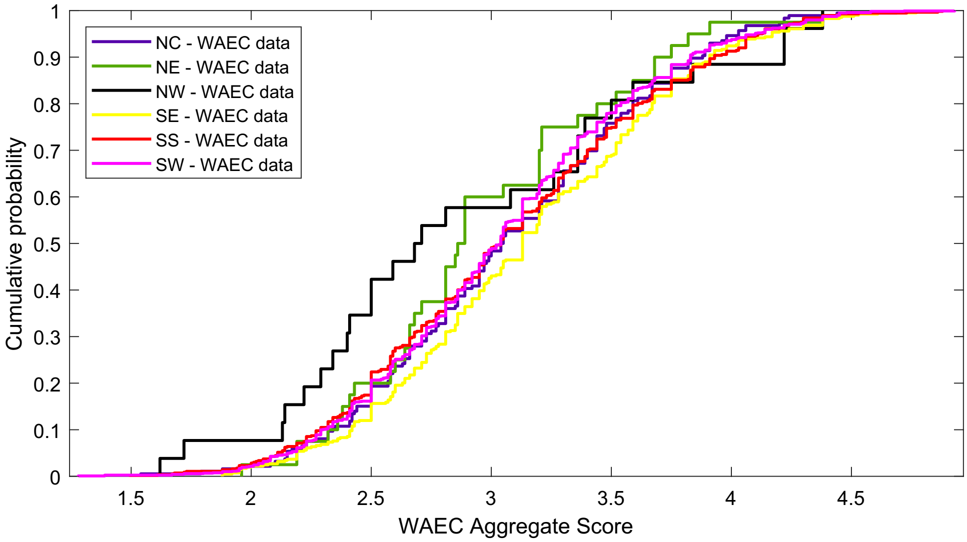


Fig. 5 Cumulative probability plot of the WAEC score per geopolitical zone

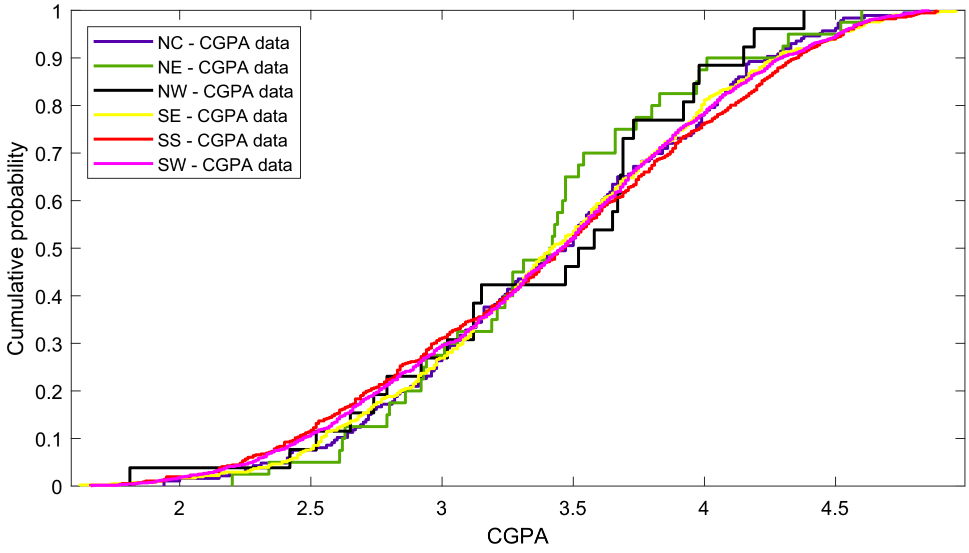


Fig. 7 Cumulative probability plot of the CGPA per geopolitical zone

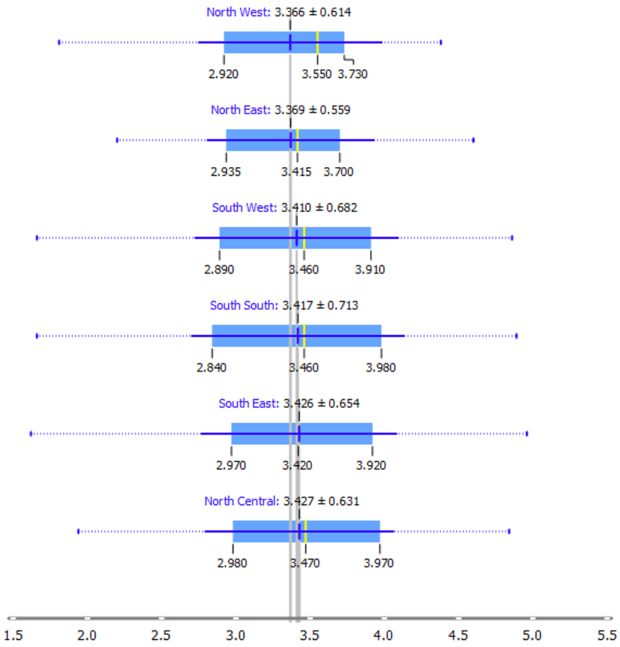


Fig. 8 CGPA variations across the six geopolitical zones

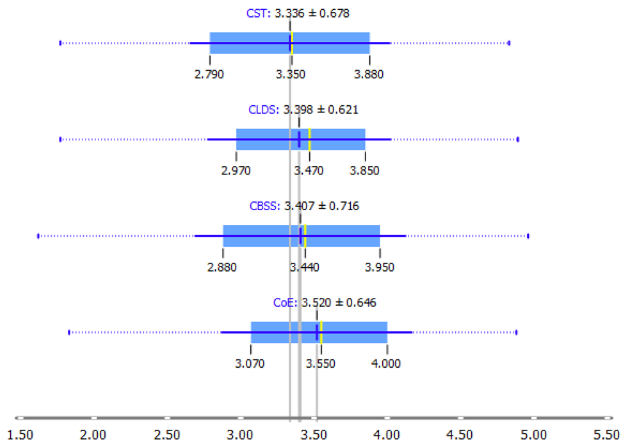
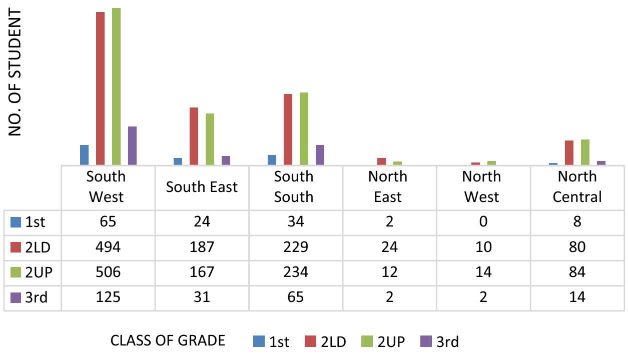


Fig. 9 CGPA variations across the four colleges



	South West	South East	South South	North East	North West	North Central
■ 1st	65	24	34	2	0	8
■ 2LD	494	187	229	24	10	80
■ 2UP	506	167	234	12	14	84
■ 3rd	125	31	65	2	2	14

Fig. 10 Distribution of the class of grades across the six geopolitical zones

regression-based analysis, the class of grade was skipped while the CGPA was used as the dependent variable. The class of grade is coded as follows; 1st, for first class, 2UP for second class upper, 2LD for second class lower division, and 3rd for third class. To acquire hidden knowledge from the dataset, the study was carried out in stages which comprise data collection, data cleaning and pre-processing, determination of the descriptive statistics of the dataset, predictive modelling using data mining and regression, result evaluation and interpretation. The impact of data balancing techniques; over-sampling and under-sampling on the predictive accuracy of the algorithms, on the imbalanced dataset was also investigated and presented in Sect. 5.2.

5 Results

5.1 Educational data mining using Orange application

The Orange software is open-source software that provides a visual approach to machine learning for an interactive data analysis which enables easy construction and configuration of workflows for various machine learning studies. In this study, an Orange workflow was developed as shown in Fig. 11. Four data mining algorithms; the Classification Tree, the Neural Network [44], the Naïve Bayes and the Random Forest algorithms were applied to evaluate the predictive capabilities of the student features considered. 70% of the data population was randomly selected using stratified sampling for training the model, while the remaining 30% was deployed for performance evaluation, and the training-test sequence was repeated ten times. The result of the analysis is presented in two folds: considering the geopolitical zone as a feature and excluding the geopolitical zone in the analysis. The lift curve shows comparatively the performance of the data mining algorithms using the average over class as the target class. The confusion matrix of the sub-samples is presented in Tables 2 and 3 for the Classification Tree algorithm, while Tables 4 and 5 show the confusion matrix for the Neural Network algorithm as a percentage of the actual. Tables 6 and 7 show the confusion matrix for the Naïve Bayes, while Tables 8 and 9 present the confusion matrix for the Random Forest data mining algorithm. A comparative evaluation of the performance of the four data mining algorithms is presented in Tables 10 and 11 respectively for the two cases i.e. when the geopolitical zone was considered as a feature, and when it was excluded in the analysis. The data mining classifiers are rated according to their Classification Accuracy (CA), the Precision rate, the Area under ROC Curve (AUC), the F1 score, and the Recall.

5.1.1 The tree algorithm

Table 2 Confusion matrix for the tree algorithm considering geopolitical zone

	Predicted			
	1st (%)	2LD (%)	2UP (%)	3rd (%)
Actual				
1st	17.00	31.80	49.80	1.50
2LD	4.80	52.00	35.00	8.20
2UP	9.60	42.70	44.20	3.50
3rd	3.10	45.80	28.90	22.20

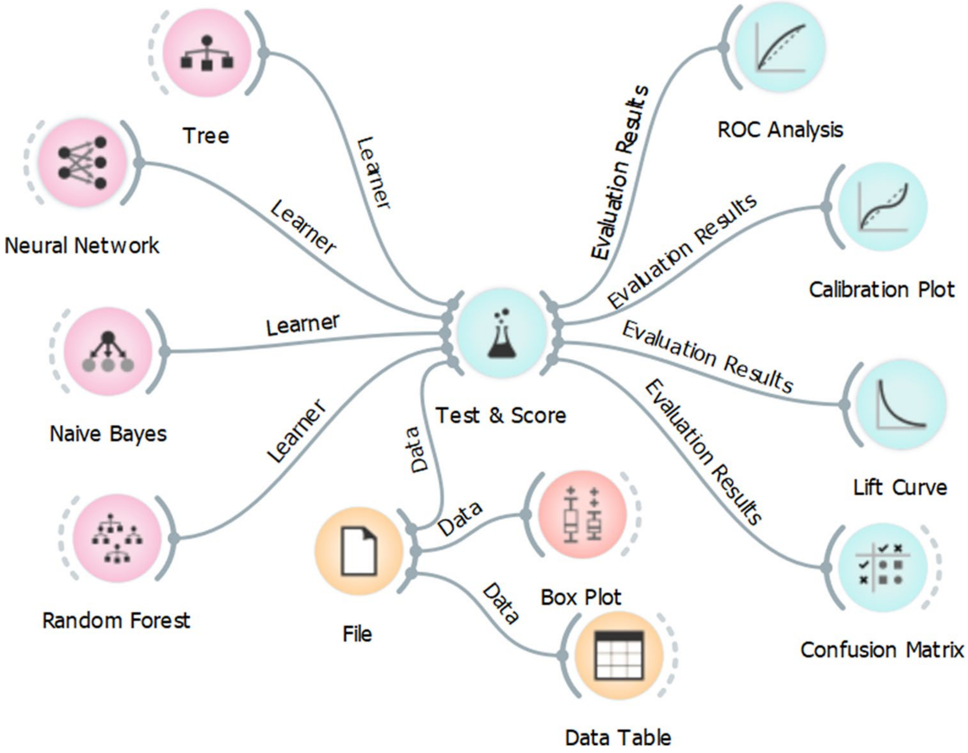


Fig. 11 The Orange data mining workflow

Table 3 Confusion matrix for the tree algorithm excluding geopolitical zone

	Predicted			
	1st (%)	2LD (%)	2UP (%)	3rd (%)
Actual				
1st	17.20	31.80	50.00	1.00
2LD	4.40	52.20	35.00	8.40
2UP	9.10	42.10	45.20	3.60
3rd	3.50	52.20	23.80	20.60

5.1.2 The neural network algorithm

Table 4 Confusion matrix for the neural network considering geopolitical zone

	Predicted			
	1st (%)	2LD (%)	2UP (%)	3rd (%)
Actual				
1st	8.20	19.00	72.80	0.00
2LD	0.10	61.40	34.50	4.00
2UP	1.20	39.60	57.80	1.30
3rd	0.00	69.20	14.70	16.10

Table 5 Confusion matrix for the neural network excluding geopolitical zone

	Predicted			
	1st (%)	2LD (%)	2UP (%)	3rd (%)
Actual				
1st	0.00	15.80	84.20	0.00
2LD	0.10	58.60	38.60	2.70
2UP	0.20	36.90	62.40	0.50
3rd	0.00	74.30	17.50	8.20

5.1.3 The Naïve Bayes algorithm

Table 6 Confusion matrix for the Naïve Bayes considering geopolitical zone

	Predicted			
	1st (%)	2LD (%)	2UP (%)	3rd (%)
Actual				
1st	0.00	15.20	84.80	0.00
2LD	0.20	58.00	38.50	3.40
2UP	0.20	37.50	61.70	0.60
3rd	0.10	71.00	18.20	10.70

Table 7 Confusion matrix for the Naïve Bayes excluding geopolitical zone

	Predicted			
	1st (%)	2LD (%)	2UP (%)	3rd (%)
Actual				
1st	0.00	15.80	84.20	0.00
2LD	0.10	58.60	38.60	2.70
2UP	0.20	36.90	62.40	0.50
3rd	0.00	74.30	17.50	8.20

5.1.4 The random forest algorithm

Table 8 Confusion Matrix for the Random Forest considering geopolitical zone

	Predicted			
	1st (%)	2LD (%)	2UP (%)	3rd (%)
Actual				
1st	11.20	26.20	62.00	0.50
2LD	0.60	54.30	39.20	5.90
2UP	2.90	41.00	54.30	1.80
3rd	0.10	54.70	23.50	21.70

Table 9 Confusion matrix for the random forest excluding geopolitical zone

	Predicted			
	1st (%)	2LD (%)	2UP (%)	3rd (%)
Actual				
1st	8.20	27.00	64.50	0.20
2LD	0.70	54.80	38.00	6.60
2UP	3.40	39.90	54.40	2.40
3rd	0.10	54.20	23.80	21.90

The quality of a data mining prediction can be examined visually using a lift curve. The Lift curve is obtained by plotting the true positive predictions (TP Rate) against the actual total number of positive instances (P rate) as a measure of the proficiency of the classifier algorithm. The lift curves for the data mining analyses are presented for the four target classes in Figs. 12 to 19 showing the variations in the predictive capabilities of the four data mining algorithms. Figures 12 and 13 show the lift curve for students that graduated with 1st class grade as the target class. Figures 14 and 15 present the lift curve for the second-class upper grade (2UP), both for the inclusion and exclusion of the geopolitical zone

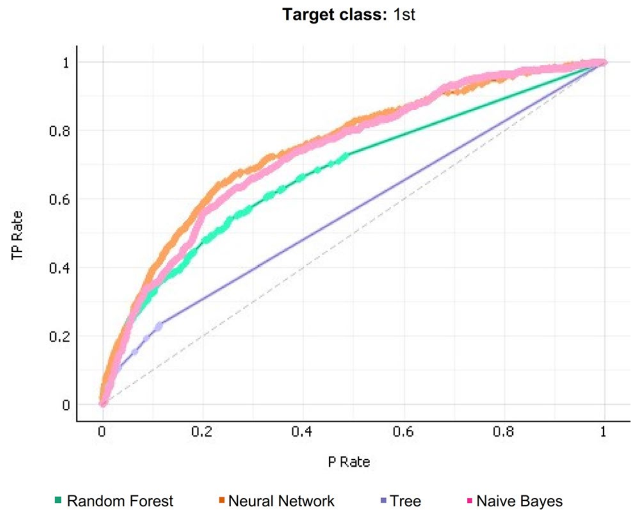


Fig. 12 The lift curve for the 1st class grade considering geopolitical zone

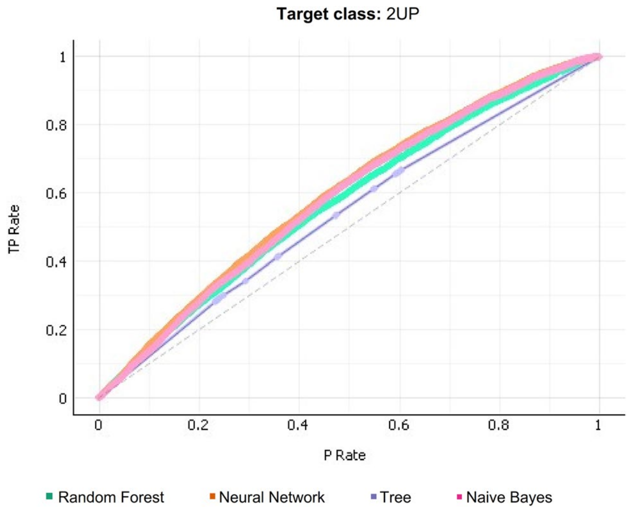


Fig. 14 The lift curve for the 2UP grade considering geopolitical zone

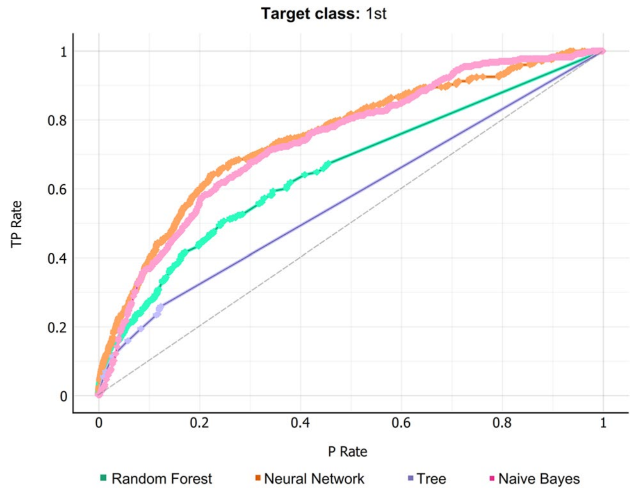


Fig. 13 The lift curve for the 1st class grade excluding geopolitical zone

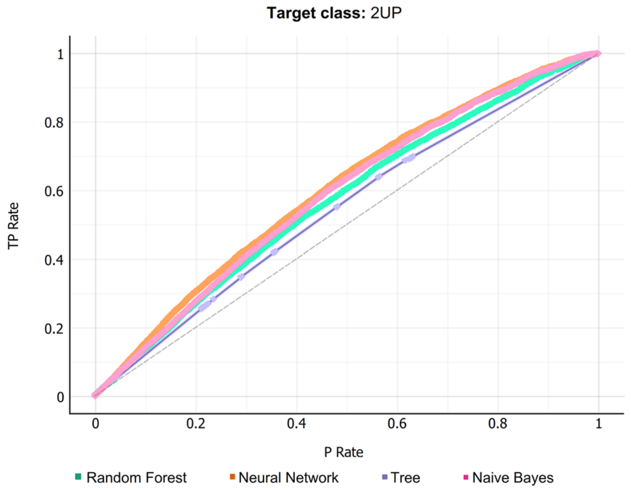


Fig. 15 The lift curve for the 2UP grade excluding geopolitical zone

as a feature respectively. The lift curve for the prediction of students that graduated with second class lower grade (2LD) is displayed in Figs. 16 and 17 respectively, while Figs. 18 and 19 show the lift curve expressing the predictive capabilities of the four data mining algorithms for the third-class graduation grade (3rd). From the lift curves, it is observed that the algorithms were better able to predict the 1st class grade and the 3rd class grade than the second class upper and the second-class lower division graduation CGPA grades.

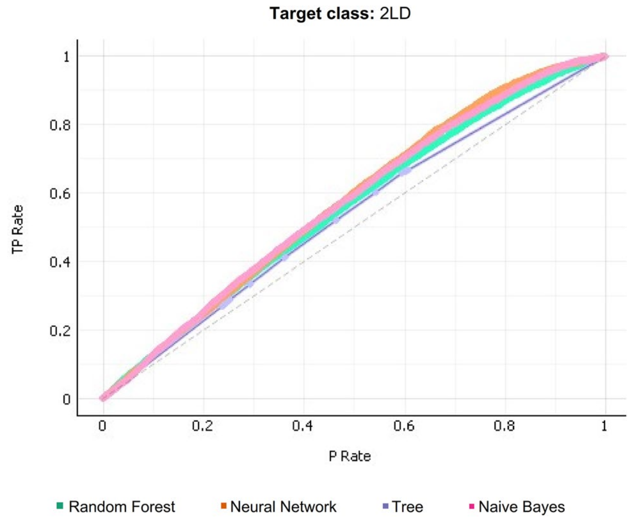


Fig. 16 The lift curve for the 2LD grade considering geopolitical zone

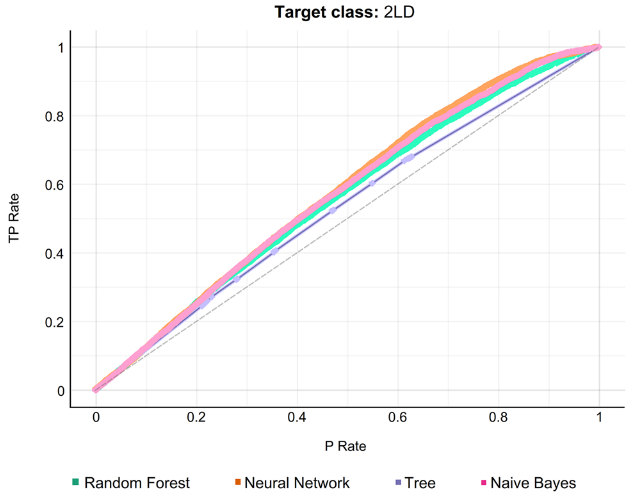


Fig. 17 The lift curve for the 2LD grade excluding geopolitical zone

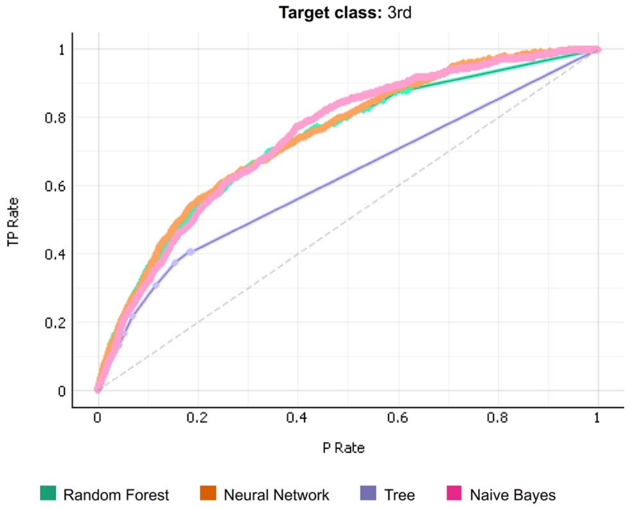


Fig. 18 The lift curve for the 3rd class grade considering geopolitical zone

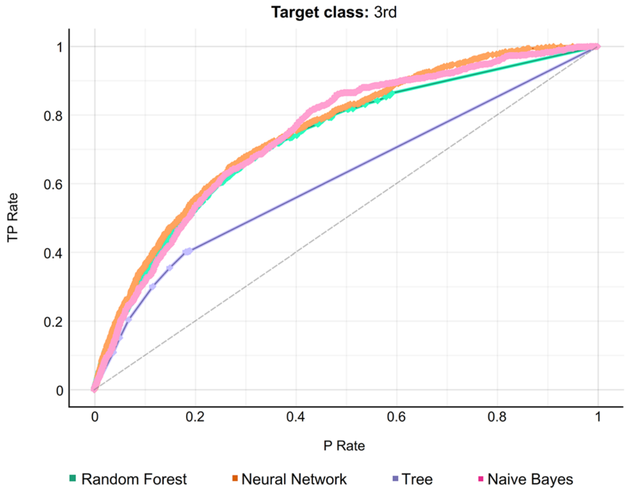


Fig. 19 The lift curve for the 3rd class grade excluding geopolitical zone

5.1.5 Performance comparison of all the algorithms

Table 10 Performance comparison for the data mining algorithms considering geopolitical zone

Method	AUC	CA	F1	Precision	Recall
Tree	0.573	0.438	0.437	0.440	0.438
Naive Bayes	0.665	0.517	0.488	0.479	0.517
Random forest	0.639	0.496	0.484	0.483	0.496
Neural network	0.674	0.528	0.508	0.520	0.528

Table 11 Performance comparison for the data mining algorithms excluding geopolitical zone

Method	AUC	CA	F1	Precision	Recall
Tree	0.576	0.442	0.440	0.442	0.442
Naive Bayes	0.666	0.520	0.488	0.480	0.520
Random forest	0.628	0.487	0.476	0.475	0.487
Neural network	0.681	0.532	0.510	0.528	0.532

5.2 The impact of under-sampling and over-sampling on model performance

As a result of the typical variation in student performance, the dataset analysed in this study is imbalanced in terms of the number of student in each class of grade (first class, second class upper, second class lower, and third class). To evaluate the effect of this imbalance, a balanced dataset was generated using data-based techniques. Two data mining experiments were performed using under-sampling and over-sampling while excluding the geopolitical zone feature. For the under-sampling analysis, the number of total samples was reduced to 532 i.e. 133 samples per class, and for the over-sampling analysis, the number of samples was increased to 5096 i.e. 1024 samples per class by sampling with replacement.

The result of the under-sampling is presented in Table 12, and it does not show a significant difference from the performance of the model using the actual dataset without under-sampling. By under-sampling, a lot of information is lost. For example, the 2LD samples is reduced from 1024 to 133 which does not depict reality. The result of the over-sampling which is often performed to improve classification accuracy is presented in Table 13, and it reveals an improvement in performance for the Tree, and Random Forest algorithms. Although, over-sampling improved the classification accuracy, but it has created a bias towards the minority class. For example, the total number of first class grade is 133, but this was scaled up to 1024 to create a balanced dataset, and as such, this does not depict reality. In most educational institutions, there will always be more average student, than exceptionally good or poor student. This is a typical challenge with data mining, and as such, the predictive accuracy is further investigated using traditional regression analysis.

Table 12 Performance comparison using under-sampling while excluding geopolitical zone

	AUC	CA	F1	Precision	Recall
Tree	0.6200	0.3820	0.3849	0.3937	0.3820
Naive Bayes	0.7226	0.4644	0.4479	0.4489	0.4644
Random forest	0.7103	0.4607	0.4571	0.4552	0.4607
Neural network	0.7447	0.4888	0.4837	0.4831	0.4888

Table 13 Performance comparison using over-sampling while excluding geopolitical zone

Method	AUC	CA	F1	Precision	Recall
Tree	0.874	0.750	0.740	0.736	0.750
Naive Bayes	0.743	0.491	0.483	0.479	0.491
Random forest	0.938	0.798	0.790	0.788	0.798
Neural network	0.794	0.539	0.535	0.533	0.539

5.3 Analysis using multiple linear regression

Multiple linear regression is a statistical model that represents the relationship between a dependent variable, and more than one independent variables. It attempts to explain the extent to which variations in the dependent variable is explained by variations in each of the independent variables. To further evaluate the accuracy and results of the data mining analysis; multiple linear regression study of the dataset was also carried out. Just like the case of the data mining analysis, the regression analysis is also in two folds. The first case considers the coded geopolitical zone of the students as an independent variable, while the second case excludes the insignificant predictors identified in the first case. The variables applied in the analysis are the year of graduation, the aggregate WAEC score, the JAMB score, the coded college of the student, the coded geopolitical zone as independent variables, and the CGPA as the dependent variable.

5.3.1 Regression analysis considering geopolitical zone

The multiple regression analysis has five independent variables, coded geopolitical zone (GeoZone), coded college, the year of graduation (YoG), jamb score, and the WAEC Aggregate score while the dependent variable is the graduation CGPA. The result of the analysis is presented in Table 14 as the regression model summary,

Table 14 Regression model summary

	Coefficients	Standard Error	t Stat	*P* value	Lower 95%	Upper 95%
Intercept	38.0340	20.4582	1.8591	0.0631	− 2.0836	78.1516
GeoZone	− 0.0005	0.0085	− 0.0623	0.9503	− 0.0172	0.0161
College	0.0153	0.0108	1.4246	0.1544	− 0.0058	0.0364
YoG	− 0.0185	0.0102	− 1.8172	0.0693	− 0.0385	0.0015
JAMB score	0.0066	0.0005	12.7043	0.0000	0.0056	0.0076
WAEC aggregate	0.3690	0.0219	16.8776	0.0000	0.3261	0.4118

Table 15 Regression statistics

Parameter	Value
Multiple R	0.4602
R square	0.2118
Adjusted R square	0.2101
Standard error	0.6030
Observations	2413

Table 16 Regression anova

	df	SS	MS	F	Significance F
Regression	5	235.1634	47.0327	129.3312	1.3E−121
Residual	2407	875.3311	0.3637		
Total	2412	1110.4945			

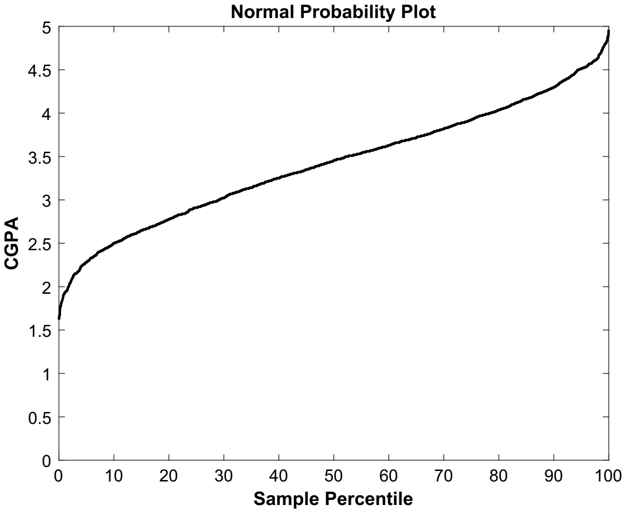


Fig. 20 The normal probability plot of the CGPA

while key regression statistics are available in Table 15. Table 16 shows the regression Anova, and the normal probability plot of the CGPA is displayed in Fig. 20. It was observed from the *P* value column in Table 14 that, at 5% significance level ($P = 0.05$), GeoZone ($0.9503 > 0.05$), College ($0.1544 > 0.05$), YoG ($0.0693 > 0.05$) and the intercept ($0.0631 > 0.05$) are not statistically significant. The Multiple R which explains the correlation between the actual values of the CGPA and the predicted CGPA is 0.4602 while the adjusted R square is 0.2101 which implies that the independent variables only explain 21.01% of the variability of the graduation CGPA of the students. To evaluate the overall model, we consider Table 16 which shows that although some of the independent variables are not strong predictors of the graduation CGPA of the students, but the *P* value of the regression model for the F test statistic (Significance F) is 1.3E−121 which is far lower than 0.05 and as such it implies that the regression model is actually significant for predicting the graduation CGPA of student.

5.3.2 Regression analysis after excluding insignificant predictors

Based on the previous multiple regression analysis in Sect. 5.3.1 above, the insignificant independent variables i.e. the coded geopolitical zone, the coded college variable, and the year of graduation were removed from the model and the regression analysis was repeated. The result of the analysis is presented in Table 17 as the regression model summary, while key regression statistics are available in Table 18, and Table19 shows the regression Anova. From Table 17, it can be observed that all the predictors including the intercept are significant with *P* values less than 0.05 which confirms that the predictors on which the model is based are statistically significant. The F statistic has also increased from 129.3312 in Table 16 to 320.2196 in Table 19, and the overall *P* value of the model is $4.7194\text{E}{-}124 < 0.05$ confirming that the regression model is significant for the prediction of the graduation CGPA of students from different geopolitical zones in Nigeria using the pre-admission features of the student. With an adjusted R square value of 0.2093, it implies the that model only explains 20.93% of the variations in the CGPA. Although the R^2 is not better than the previous case in Sect. 5.3.1 but the model is more reliable because it only contains statistically significant predictors.

Using the coefficients of the multiple linear regression model, the predictive regression equation can be written as follows:

Table 17 Regression model summary

	Coefficients	Standard Error	t Stat	*P* value	Lower 95%	Upper 95%
Intercept	0.9051	0.1089	8.3101	0.0000	0.6916	1.1187
JAMB score	0.0061	0.0005	12.8867	0.0000	0.0052	0.0070
WAEC aggregate	0.3805	0.0212	17.9096	0.0000	0.3388	0.4222

Table 18 Regression statistics

Parameter	Value
Multiple R	0.4582
R square	0.2099
Adjusted R square	0.2093
Standard error	0.6034
Observations	2413

Table 19 Regression anova

	df	SS	MS	F	Significance F
Regression	2	233.1481	116.5741	320.2196	4.7194E−124
Residual	2410	877.3464	0.3640		
Total	2412	1110.4945			

$$\text{CGPA} = 0.9051 + (0.0061 \times \text{JAMB score}) + (0.3805 \times \text{WAEC Aggregate})$$

6 Discussion

In this study, the dataset of 2413 graduated students of Covenant University in Nigeria has been evaluated, using the Orange data mining software and regression analysis. Considering the results in Tables 10 and 11 for the data mining analysis using four classifiers i.e. the Tree, Random Forest, Neural Network, and the Naive Bayes Classifiers for the two cases: using the Geopolitical zone and excluding Geopolitical zone as a feature, it can be seen that the AUC for the Tree increased slightly from 0.573 to 0.576, while that of the Random Forest reduced from 0.639 to 0.628. Likewise, for the Neural Network, the AUC increased from 0.674 to 0.681, while for the Naïve Bayes, the AUC increased slightly from 0.655 to 0.666. A similar trend can be observed for the other measure of model fitness parameters in Tables 10 and 11. This implies that the consideration of the geopolitical zone of origin of the students only increased the performance of the Random Forest algorithm, while for the other three algorithms their AUC and CA actually reduced slightly when the geopolitical zone was considered as a feature. The result of the data mining analysis was verified by multiple regression analysis, and the model revealed that the coded geopolitical zone is actually statistically insignificant in the prediction of the graduation CGPA of Covenant University using the admission scores of each student. The above-average accuracy observed, is an indication that the pre-admission academic performance of students is not a complete predictor of their performance once in the university. A number of factors; academic and non-academic such as social lifestyle, financial buoyancy, class attendance, game and internet addictions etc. will shape the performance of a student once admitted into the university. Over-sampling technique to create a balanced dataset towards improving the prediction of the minority class will create a bias towards the minority class. Except if the goal is to predict the minority class, oversampling technique for predicting student performance should be avoided as this distorts reality.

7 Conclusion

In recent times, the application of data mining is gaining ground even in new fields of study. Educational data mining is a great tool that reveals useful information by scientifically mining dataset generated within the education domain. The quality, acceptance and value for education in Nigeria is greatly influenced by ethnicity. There are cities in Nigeria where a significant number of their elites are professors and Ph.D. holders, while only few of such exists in others. In this study, the significance of the geopolitical zone of students on the prediction accuracy of their graduation CGPA in a University in Nigeria was examined. Various statistical analyses were carried out to show trends among the six geopolitical zones. The average CGPA across geopolitical zones varied from 3.366 to 3.427 which is quite close, and implies that the average performance of students across the geopolitical zones do not vary much, as it would have been in the early days of the country Nigeria. The data mining and regression analyses further revealed that the geopolitical zone of the students is not a statistically significant variable for predicting the graduation CGPA of students based on their admission criteria scores, using Covenant University; a private institution in Nigeria as a case study. Although, the multiple regression model developed is significant for predictive analysis but the R^2 observed is 0.2099.

The accuracy of the model may be improved by considering additional features, especially the academic scores of the students in their first year which is a reflection of how well the students have transformed from their secondary school perception of education, and settled into the academic rigours and demands of a university. Since the case study university is in the south west region of the country, it will be interesting to find out what the results of a similar study would be using universities in other regions as case study, in order to see if the variations in the average academic performance of students from the six geopolitical zone will have the same trend. Likewise, the analysis in this study is based on regression and data mining using the Orange software. It will be interesting to see what the results and trends would look like using alternative tools and analytical techniques.

Compliance with ethical standards

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

1. Ukiwo U (2007) Education, horizontal inequalities and ethnic relations in Nigeria. Int J Educ Dev 27(3):266–281. https://doi.org/10.1016/j.ijedudev.2006.10.016
2. Olujuwon T (2002) Education in Nigeria: a futuristic perspective. In conference proceedings of 4th world conference of the internationalconsortium for educational development 3–6 July, Perth, Western Australia, pp 1–8
3. Okebukola P (2000) Trends in tertiary education in Nigeria. The State of Education in Nigeria, p 84–102
4. JAMB (2015) Statistics 2010–2016. 25 Nov 2018. https://www.jamb.gov.ng/Statistics.aspx. Accessed 25 Nov 2018
5. Aluede O, Idogho PO, Imonikhe JS (2012) Increasing access to university education in Nigeria: present challenges and suggestions for the future. In: The African symposium
6. Yusuf U (2018) Map of the six geopolitical zone in Nigeria showing total number of registered voters. Geospatial Solutions Expert. 07 Dec 2018; https://umar-yusuf.blogspot.com/2018/09/map-of-six-geopolitical-zone-in-nigeria.html. Accessed 7 Dec 2018
7. NBS (2017) National Population Estimates. National Population Commission and National Bureau of Statistics Estimates. https://nigerianstat.gov.ng/resource/POPULATION%20PROJECTION%20Nigeria%20sgfn.xls. Accessed 7 March 2019
8. Eze T, Sunday C, Ogbodo JC (2014) Patterns of inequality in human development across Nigeria's six geopolitical zones. Dev Ctry Stud 4(8):97–101
9. Wikipedia (2018) Geopolitical zones of Nigeria. 18 11 2018. https://en.wikipedia.org/wiki/Geopolitical_zones_of_Nigeria. Accessed 18 Nov 2018
10. Wikipedia (2018) Nigeria. 15 Nov 2018. https://en.wikipedia.org/wiki/Nigeria. Accessed 15 Nov 2018
11. Abernethy DB (1964) Nigeria creates a new region. Africa Rep 9(3):8
12. Agboola B, Ofoegbu F (2010) Access to University education in Nigeria: a review. Online Submission
13. Antoniades N, Haan P (2019) Government capabilities as drivers of performance: path to prosperity. Heliyon 5(2):e01180
14. Badejo JA et al (2018) Data sets linking ethnic perceptions to undergraduate students learning outcomes in a Nigerian Tertiary Institution. Data Brief 18:760–764
15. Jin X et al (2015) Significance and challenges of big data research. Big Data Res 2(2):59–64. https://doi.org/10.1016/j.bdr.2015.01.006
16. Alharbi FR, Khan MB (2019) Identifying comparative opinions in Arabic text in social media using machine learning techniques. SN Appl Sci 1(3):213. https://doi.org/10.1007/s42452-019-0183-3
17. Romero C, Ventura S (2010) Educational data mining: a review of the state of the art. IEEE Trans Syst Man Cybern Part C (Appl Rev) 40(6):601–618. https://doi.org/10.1109/tsmcc.2010.2053532
18. Lakshmipadmaja D, Vishnuvardhan B (2018) Classification performance improvement using random subset feature selection algorithm for data mining. Big Data Res 12:1–12. https://doi.org/10.1016/j.bdr.2018.02.007
19. Bradac V, Walek B (2017) A comprehensive adaptive system for e-learning of foreign languages. Expert Syst Appl 90:414–426
20. Hegazi MO, Abugroon MA (2016) The state of the art on educational data mining in higher education. Int J Comput Trends Technol 31(1):46–56
21. Sahin A, Top N, Delen E (2016) Teachers' first-year experience with chromebook laptops and their attitudes towards technology integration. Technol Knowl Learn 21(3):361–378. https://doi.org/10.1007/s10758-016-9277-9
22. Haraty RA, Bitar G (2019) Associating learning technology to sustain the environment through green mobile applications. Heliyon 5(1):e01141
23. Pei Z-J (2017) Educational data mining for teaching and learning. In: 2nd international conference on education and development (ICED 2017)
24. El Alfy S, Gómez JM, Ivanov D (2017) Exploring instructors' technology readiness, attitudes and behavioral intentions towards e-learning technologies in Egypt and United Arab Emirates. Educ Inf Technol 22(5):2605–2627. https://doi.org/10.1007/s10639-016-9562-1
25. Fernandes E et al (2019) Educational data mining: predictive analysis of academic performance of public school students in the capital of Brazil. J Bus Res 94:335–343. https://doi.org/10.1016/j.jbusres.2018.02.012
26. Adekitan AI, Salau O (2019) The impact of engineering students' performance in the first three years on their graduation result using educational data mining. Heliyon 5(2):e01250. https://doi.org/10.1016/j.heliyon.2019.e01250
27. Rumbold JMM, Pierscionek BK (2018) What are data? A categorization of the data sensitivity spectrum. Big Data Res 12:49–59. https://doi.org/10.1016/j.bdr.2017.11.001
28. Kim D et al (2018) Learning analytics to support self-regulated learning in asynchronous online courses: a case study at a women's university in South Korea. Comput Educ 127:233–251. https://doi.org/10.1016/j.compedu.2018.08.023

29. Ahuja R, Kankane Y (2017) Predicting the probability of student's degree completion by using different data mining techniques. In: 4th international conference on image information processing, ICIIP 2017. Shimla, India
30. Ayers E, Nugent R, Dean N (2009) A comparison of student skill knowledge estimates. International Working Group on Educational Data Mining
31. Zakrzewska D (2008) Cluster analysis for users' modeling in intelligent e-learning systems. In: International conference on industrial, engineering and other applications of applied intelligent systems. Springer
32. Tang C et al (2000) Personalized courseware construction based on web data mining. In: Proceedings of the first international conference on Web information systems engineering, 2000. IEEE
33. Hussain S et al (2018) Educational data mining and analysis of students' academic performance using WEKA. Indones J Electr Eng Comput Sci 9(2):447–459. https://doi.org/10.11591/ijeecs.v9.i2.pp447-459
34. Nurhayati OD et al (2018) Graduation prediction system using artificial neural network. Int J Mech Eng Technol 9(7):1051–1057
35. Adekitan AI, Noma-Osaghae E (2018) Data mining approach to predicting the performance of first year student in a university using the admission requirements. Educ Inf Technol. https://doi.org/10.1007/s10639-018-9839-7
36. Ahmad F, Ismail N, Aziz AA (2015) The prediction of students' academic performance using classification data mining techniques. Appl Math Sci 9(129):6415–6426
37. Kabakchieva D (2013) Predicting student performance by using data mining methods for classification. Cybern Inf Technol 13(1):61–72
38. Alharbi Z et al (2016) Using data mining techniques to predict students at risk of poor performance. In: 2016 SAI computing conference (SAI)
39. Arnold KE (2010) Signals: applying academic analytics. Educ Q 33(1):n1
40. Atta Ur R et al (2018) Educational data mining for enhanced teaching and learning. J Theor Appl Inf Technol 96(14):4417–4427
41. Gibson D, de Freitas S (2016) Exploratory analysis in learning analytics. Technol Knowl Learn 21(1):5–19. https://doi.org/10.1007/s10758-015-9249-5
42. Ifenthaler D, Erlandson BE (2016) Learning with data: visualization to support teaching, learning, and assessment. Technol Knowl Learn 21(1):1–3. https://doi.org/10.1007/s10758-015-9273-5
43. Bharara S, Sabitha S, Bansal A (2018) Application of learning analytics using clustering data Mining for Students' disposition analysis. Educ Inf Technol 23(2):957–984. https://doi.org/10.1007/s10639-017-9645-7
44. Lau ET, Sun L, Yang Q (2019) Modelling, prediction and classification of student academic performance using artificial neural networks. SN Appl Sci 1(9):982. https://doi.org/10.1007/s42452-019-0884-7