Applying machine learning approach to predict students' performance in higher educational institutions

Article *in* Kybernetes · June 2021

DOI: 10.1108/K-12-2020-0865

CITATIONS
8

READS
1,416

2 authors:

Nas Yakubu
Arden University
11 PUBLICATIONS **327** CITATIONS
SEE PROFILE

A. Mohammed Abubakar
Antalya Bilim University
69 PUBLICATIONS **2,476** CITATIONS
SEE PROFILE

Applying machine learning approach to predict students' performance in higher educational institutions

Mohammed Nasiru Yakubu
American University of Nigeria, Yola, Nigeria, and

A. Mohammed Abubakar
Antalya Bilim Universitesi, Antalya, Turkey

Received 13 December 2020
Revised 4 February 2021
Accepted 23 March 2021

Abstract

Purpose – Academic success and failure are relevant lifelines for economic success in the knowledge-based economy. The purpose of this paper is to predict the propensity of students' academic performance using early detection indicators (i.e. age, gender, high school exam scores, region, CGPA) to allow for timely and efficient remediation.

Design/methodology/approach – A machine learning approach was used to develop a model based on secondary data obtained from students' information system in a Nigerian university.

Findings – Results revealed that age is not a predictor for academic success (high CGPA); female students are 1.2 times more likely to have high CGPA compared to their male counterparts; students with high JAMB scores are more likely to achieve academic success, high CGPA and vice versa; students from affluent and developed regions are more likely to achieve academic success, high CGPA and vice versa; and students in Years 3 and 4 are more likely to achieve academic success, high CGPA.

Originality/value – This predictive model serves as a classifier and useful strategy to mitigate failure, promote success and better manage resources in tertiary institutions.

Keywords Information systems, Education, ICT, Artificial intelligence, Academic success, Machine learning, Logistic regression, Enrollment data, Higher education, Nigeria

Paper type Research paper

1. Introduction

Developments in educational technologies have resulted in the accumulation of large amounts of data on students and their learning activities. The exploration of these data has empowered researchers and educators with knowledge and intelligence to support teaching and learning activities (Sedrakyan *et al.*, 2020) and to determine academic outcomes (Al-Sudani and Palaniappan, 2019). A common indicator of academic outcome in higher education institutions is the cumulative grade point average (CGPA). Higher education institutions require their students to maintain a minimum CGPA of 2.0 on a 4.0 scale, to enable them progress in their studies. High CGPA (i.e. above 2.0) are considered as academic success that is accompanied with rewards, and low CGPA (i.e. less than 2.0) results in a series of penalties ranging from a warning, probation, restriction, suspension or dismissal.

This research did not receive any specific grant from funding agencies in the public, commercial or not-for-profit sectors. The authors have no conflict of interest.

Kybernetes

0368-492X
DOI 10.1108/K-12-2020-0865

There are two streams of research on the factors that contribute to the academic success of students in higher educational institutions. The former, learning analytics are based on data mainly derived from the use of learning management systems (LMS) applications (Huang *et al.*, 2020). Variables used in learning analytics include clickstream data, forum posts, login frequency, course duration, time spent, exercises and peer interaction (Mwalumbwe and Mtebe, 2017; Moreno-Marcos *et al.*, 2020). The latter focuses on socio-demographic variables such as enrollment data (Rizvi *et al.*, 2019; Aulck *et al.*, 2019). Students' enrollment information is mostly obtained at registration time such as age, gender, ethnicity, prior education, poverty levels and parents' qualification (Aulck *et al.*, 2019; Hoffait and Schyns, 2017; Kamal and Ahuja, 2019).

Research found that self-reported survey data designed to predict academic success lacks objectivity in comparison to LMS data (Tempelaar *et al.*, 2020). Thus, academic advising and counseling staff in higher educational institutions rely on the CGPA scores and LMS data to predict students' future grades and detect those needing guidance or at risk of falling (Gašević *et al.*, 2016; Shum and Ferguson, 2012). This approach is vital as it allows for the implementation of intervention strategies (Olaya *et al.*, 2020). However, this process captures failing students a little too late. Similar arguments were echoed in Tempelaar *et al.* (2015) influential work, which found that entry test data and demographic variables are more robust compared to LMS data in determining student academic success. Cooper *et al.* (2001) *argued that qualitative and quantitative intervention strategies are needed to improve professional practice in education.*

The deficiencies of LMS data in predicting academic success has been acknowledged in past studies (Conijn *et al.*, 2016). For instance, LMS data cannot address course differences and complexity. Consequently, Francis *et al.* (2020) stated that structural disadvantage does impact student outcomes, particularly for disadvantaged students and this impact cannot be identified in LMS data. Scholars recommend the use of alternative data sources (Conijn *et al.*, 2016; Francis *et al.*, 2020; Tempelaar *et al.*, 2015) and machine learning approach in determining students' academic success (Rincón-Flores *et al.*, 2020). In response to these calls, this paper opts for socio-demographic data and machine learning approach to determine students' academic success.

Because understanding the factors that contribute towards academic failure via an early detection system would be more beneficial as timely intervention, in the form of guidance, it can be made prior to students selecting courses at the beginning of a new session. In addition, universities in the Sub-Saharan region are under resourced coupled with lack of technologies to capture data (Bawack and Kamdjoug, 2020). This has contributed to limited understanding of and research from the region. Few universities in the region are now adopting technology to facilitate learning and improve their processes.

This paper's aim is to fill the gap in the literature and also provide insights for administrators in the region. The rest of this study is arranged as follows: the next section presents the literature review. This is followed by the methodology section. The discussion and results section summarize the findings, the take home lesson and ends with limitations, and future research directions.

2. Literature review

There are numerous benefits that accompany the prediction of university students' academic success. The most significant of these benefits is to improve the success rates of their students. According to Alyahyan and Düştegör (2020), university administrators rely on the early detection of students who are at-risk, as an institution's students' success is an important metric used to measure the institution's performance. Another identified benefit

of predicting students' success is the reduction of student attrition. Student attrition is the decrease in student numbers (dropouts)with the passage of time. Student attrition represents a misuse of resources (Berens *et al.*, 2018). These include public and private resources in the form of money and time.

In addition, Berens *et al.* (2018) identified the "feelings of inadequacy", which leads to social stigmatization of the individual. In countries where there is a high competition for spaces in university programs, there is the possibility that students who dropout would have wasted the space that could have been allocated to a student who could have fared better. This is another result of student attrition, which could be avoided if the detection of students' risk is done prior to acceptance into the university. The early prediction of academic success is also useful in allocating support and intervention procedures to at-risk students, this helps university administrators in the planning and allocation of resources required for the guidance and support of such students.

Data on the students is usually required to predict the success of students, and there are several statistical techniques used in the prediction analysis. Universities collect large amounts of data from students before they are admitted (enrollment data) and also while they are enrolled. The former data type, enrollment data, is usually collected to gather information about the students and is often used to process admission. In Nigerian universities, the admission office uses this type of data to decide the programs to be offered to applicants. For example, criteria such as the region where the student is from and their previous academic records are used to offer programs to students. A typical academic record used to offer admission is the Joint Admission and Matriculation Board (JAMB) examination score.

JAMB is responsible for conducting matriculation examinations for entry into all Nigerian universities, polytechnics and colleges of education. All university programs have a cut-off mark, for instance, the minimum cut off mark for entry into federal universities is between 180 and 200 points, out of a possible 400 points, depending on the university. These cut-off marks do not only depend on the admitting university. Other variables are also considered such as the program of study and the catchment area (region) for the university (Kanyip, 2013). Programs such as Medicine (MBBS) and Law (LLB) require significantly higher JAMB scores to be admitted and students in the catchment areas are favored for admission as their score requirements are lower than students outside of the catchment area.

There is an abundance of studies on the prediction of academic success, which is based on enrollment data in developed countries, for instance, Kovacic (2010) identified ethnicity, course program and course block as significant variables that differentiate a successful from an unsuccessful student. More recently, Saa *et al.* (2019) identified gender, nationality, high school attended and student performance before joining the university as predictors of student academic performance using random forest algorithm. Similar studies that have employed demographic or enrollment data for prediction analysis, predictor variables such as high school scores (Garg, 2018; Mohamed and Waguih, 2017), gender (Garg, 2018; Putpuek *et al.*, 2018), age (Hamoud *et al.*, 2018) and ethnicity (Ahmad *et al.*, 2015) have been observed to determine the academic success of students. According to Francis *et al.* (2020), social-economic and cultural features such as age, gender, race or ethnicity and other personal characteristics can exacerbate differential student outcomes.

The other type of data collected from students pertains to their learning behavior while enrolled in school is referred to as learning analytics. Learning analytics is the analysis of data that captures a student's learning behavior for the purpose of understanding and improving the learning process, experience and environment. This is mainly captured via surveys admitted to students or derived LMS. Factors such as login frequency (Hamoud

et al., 2018; Mwalumbwe and Mtebe, 2017), discussion posts (Hamoud *et al.*, 2018; Mwalumbwe and Mtebe, 2017), number of downloads (Mwalumbwe and Mtebe, 2017) and peer interaction (Mwalumbwe and Mtebe, 2017) have all been observed to be predictors of academic success.

Traditional techniques are not suitable for analyzing large data sets (big data) obtained from student LMS (Akgül, 2018, 2019; Baleanu *et al.*, 2020). Sin and Muthu (2015) identified four suitable machine learning-based techniques for educational data, namely, logistic regression, nearest neighbor, clustering and classification. Regression is a statistical process that involves estimating the relationship between independent variables (predictors) and dependent variables (outcome), regression is a suitable technique for prediction and forecasting (Kumari and Yadav, 2018). Nearest neighbour analysis is a measure of the spread of a variable over a geographical point and groups similar observation based on the distance to the point (Lee *et al.*, 2019). Clustering categorizes similar objects into groups called clusters where objects in a cluster have similar attributes and are different to objects in other clusters (Musumeci *et al.*, 2018). Classification identifies and assigns a certain value to a group based on previously categorized values (Soofi and Awan, 2017).

Alyahyan and Düştegör (2020) suggest that regression and classification are typical examples of prediction models while clustering and association are used for descriptive models. Prediction models use supervised learning algorithms to estimate the values expected from the dependent variables based on the independent variables (Bramer, 2016). Several authors have tried to identify the best educational data mining technique (regression and classification) with regards to student prediction and the results have been inconclusive. For instance, Rusli *et al.* (2008) observed that the Neuro-fuzzy model was better than artificial neural network (ANN) and logistic regression, while Shahiri and Wahidah (2015) found ANN to be more accurate than SVM, decision tree and Naive Bayes methods. ANN has superior performance to regression analysis for prediction problem; however, ANN performance for classification problem is the same with those of logistic regression and discriminant analysis based on student data (Paliwal and Kumar, 2009).

Yaacob *et al.* (2019) compared five models in trying to predict student success and found the Naives-Bayes performed better, closely followed by the logistic regression model. Regression has also been observed to perform better than neural networks (Oyedeji *et al.*, 2020). In reviewing the "best practices in predicting academic success in higher education", Alyahyan and Düştegör, (2020) mentioned prior-academic achievement, and student demographics as the two most used variables in 69% of the papers they reviewed. The authors also identified the "CGPA as the most common factors used to predict student performance" (Alyahyan and Düştegör, 2020). Building on this line of argument, this study uses socio-demographic data obtained from active students (i.e. currently enrolled) as the input variables and CGPA as the target variable. Based on the extant literature and discussions, the following research question is proposed:

RQ. What variables (i.e. entry age, gender, state of origin, JAMB score and level of study) are relevant for predicting Sub-Saharan students' academic success?

3. Methodology

3.1 Data preprocessing stage

The proposed associations in this study were modelled using machine learning approach, specifically, logistic regression. The input and target variables are enrollment data captured by the student information system of American University of Nigeria (AUN). The historical data comprised *978* undergraduate students from the AUN. The students

were termed as active students, i.e. students enrolled in the current semester. Table 1 shows that there is a decrease in students as the level of study increases which might be associated with number of admissions and/or dropouts. An overwhelming number of the students (91.7%) are between 16 and 20 years of age and the rest are above 20 years. Pertaining to gender, 51.9% of the students are males and the rest are females. AUN is located in the North Eastern region of Nigeria, majority of the students (295, 39.5%) are from this region followed closely by the North central region (150, 20.1%), which is the closest region to the northeast in Nigeria. Consequently, 34.4% of the students are from the Southern regions of Nigeria with the South-south region contributing to almost half (115, 15.4%) of the students (Figure 1).

A total of 13 attributes were identified in the data set, namely, home address, telephone number, email address, gender, admission date, date of birth, state of origin, country of origin, JAMB score, level of study, marital status, name of previous school and CGPA. Six irrelevant attributes for student performance prediction were deleted, namely, home address, country of origin, telephone number, email address, marital status and name of previous school. Country of origin was removed as 99.99% of the students were from Nigeria and marital status was removed as only 0.02% of the students were married. In total, 28 records were discarded due to erroneous entry and/or missing value. First year students' (n = 202) information were excluded due to lack of CGPA score, as they have not earned any credit hours. Thus, seven relevant attributes were retained for analysis following the recommendation of past studies (Ahmad *et al.*, 2015; Garg, 2018; Hamoud *et al.*, 2018;

	Frequency	(%)
Year		
1	228	30.5
2	207	27.7
3	178	23.8
4	134	17.9
Entry age		
15	20	2.7
16	157	21.0
17	262	35.1
18	164	22.0
19	74	9.9
20	28	3.7
21	24	3.2
22	9	1.2
23	3	0.4
26	2	0.3
27	2	0.3
28	1	0.1
29	1	0.1
Gender		
Male	388	51.9
Female	359	48.1
CGPA		
Fail	128	17.1
Pass	619	82.9

Table 1. Demographic breakdown of the students

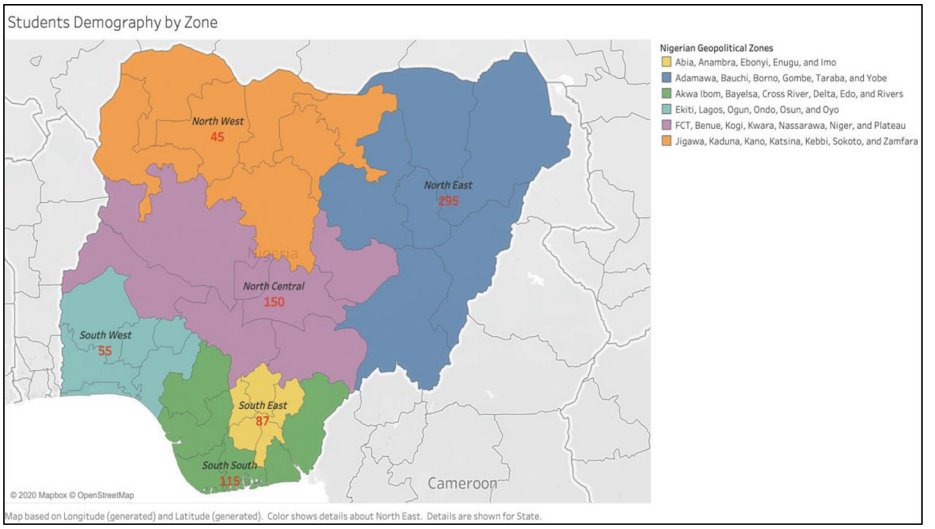

Figure 1.
Student count by geopolitical zones

Mohamed and Waguih, 2017; Putpuek *et al.*, 2018; Saa *et al.*, 2019). Preprocessing of the data involved the following steps for the listed variables:

- *Date of birth and admission date*: These two attributes were used to obtain the university entry age of the students and were subsequently standardized.
- *Gender:* The gender attribute was converted into male and female dummy variables where 1 indicated the presence and 0 the absence of the male and female variables.
- *State of origin*: There are six geo-political zones in Nigeria which are divided according to the ethnical preferences within the country. The Hausa ethnic groups make up the north-west (NW) and north-east (NE), while the Yoruba ethnic groups make up the south-west (SW), and south-east (SE) is made up of the Igbo ethnic groups. The north-central (NC) and the south-south (SS) *consist of ethnic minorities in neighboring* states with similar cultures. As every state can only be part of one geopolitical zone, six new attributes (regions) were created (NW, NE, NC, SS, SW and SE) and the values of 1 or 0 were assigned to indicate if a student was from the geopolitical region or not respectively, i.e. dummy variables.
- *JAMB score*: JAMB score values are whole numbers between 0 and 400. Similar to the entry age attribute, the JAMB scores were standardized.
- *Level of study*: Majority of the programs in the university are four-year programs with the exception of the engineering and law programs which are usually five years. As these programs are new to the university, the oldest students are in their fourth year. Similar to the regions attribute, the values of 1 or 0 were assigned to indicate if a student was in a particular year or not respectively (dummy variables).
- *CGPA (Target variable):* The CGPA, as mentioned earlier is on a scale of 4, thus a CGPA value could be a number between 0 and 4 and rounded to 2 significant figures. As a passing score is a CGPA of over 2.0, the value 1 was assigned to any CGPA over 2.0 and 0 was assigned to values less than 2.0. Therefore, the target variable, for the purpose of this study, is a discrete variable (1, 0), which is a mandatory requirement for logistic regression methods.

3.2 Logistic regression

The data preparation and processing resulted in six main attributes and 747 observations, which are further processed into 14 attributes, namely, Entry Age, NC, NE, NW, SE, SS, SW, Level 1, Level 2, Level 3, Level 4, Gender, JAMB and CGPA. The purpose of this study is to create a machine-learning algorithm that will predict if a student will achieve a passing CGPA, i.e. over 2.0 on a 4.0 scale. Logistics regression, being a predictive analysis model, is the preferred statistical method for this study as the dependent variable (CGPA) is dichotomous. After preprocessing the data, 13 attributes were identified as the input variables (Entry Age, NC, NE, NW, SE, SS, SW, Level 1, Level 2, Level 3, Level 4, Gender and JAMB score) and one attribute as the target variable (CGPA). Scikit-learn (Sklearn)an open-source machine-learning library with various algorithms such as clustering, regression and classification was used in this study.

After loading the data set, the data set was randomly split in to 2: training data with 522 observations (70 %) and test data with 225 observations (30 %). From the 747 observations, 619 students had passing CGPAs while the remaining 128 had CGPAs below 2.0. Sklearn was used to build and fit the logistic regression model using the training data. The observed accuracy of the model based on the training data was 84.7% (0.8467432950191571). This means that the model was 84.7% accurate in matching the outputs and the targets. Next, the intercept, the coefficients of the variables, as well as their corresponding odds ratio (exponential of the coefficients) were obtained as shown in Table 2.

From Table 2, the entry age variable seems to be the only variable that does not really contribute to the model, as its odd ratio is approximately 1.0. Sklearn's predict probability (predict_proba) method was used to extract the probability of earning a CGPA (>2.0) for each observation in the test data. Then a comma-delimited file (CSV) of the test cases was extracted for a visual analysis of inputs versus their probabilities. Tableau software (Tableau Desktop Professional 2020.1.2 version) was used to analyze the observations in the test dataset along with their corresponding predictions obtained from running the model with the test data set as mentioned above. Tableau is a sophisticated and interactive data visualization application used to easily transform raw data into comprehensible formats. Figures 2–6 show the output of the analysis between the variables and the predicted targets performed using Tableau.

Input variables	Coefficient	Odds ratio
Year 4	1.713867	5.550385
Year 3	1.285781	3.617493
South West	0.952796	2.592948
JAMB Score	0.935355	2.548117
North Central	0.334934	1.397848
North West	0.292712	1.340057
Female	0.21119	1.235147
Entry Age	0.06319	1.065229
Male	−0.211244	0.809576
North East	−0.423244	0.654919
South East	−0.543134	0.580925
South South	−0.614118	0.541118
Year 2	−1.243248	0.288446
Year 1	−1.756455	0.172656
Intercept (Bias)	*2.983912*	*19.764993*

Table 2. Bias, variables coefficient and odds ratio

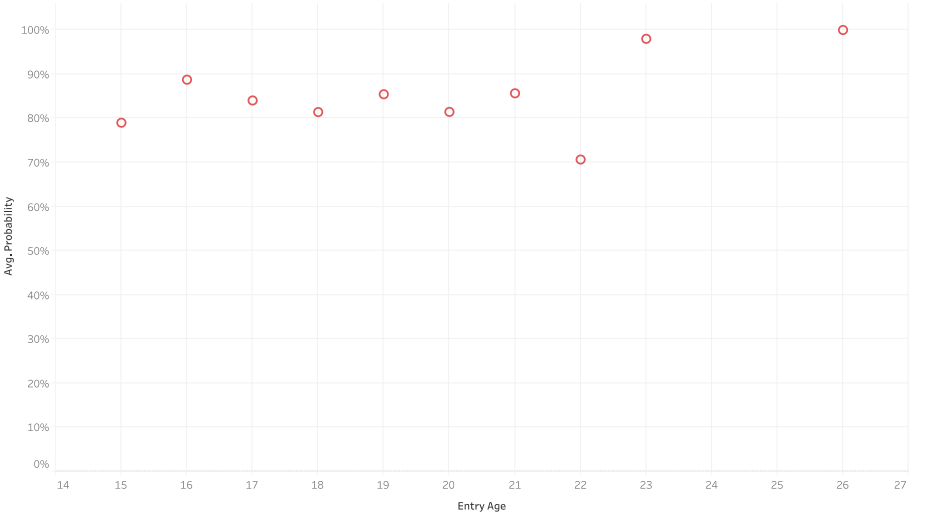

Figure 2.
Probability of academic success versus entry age

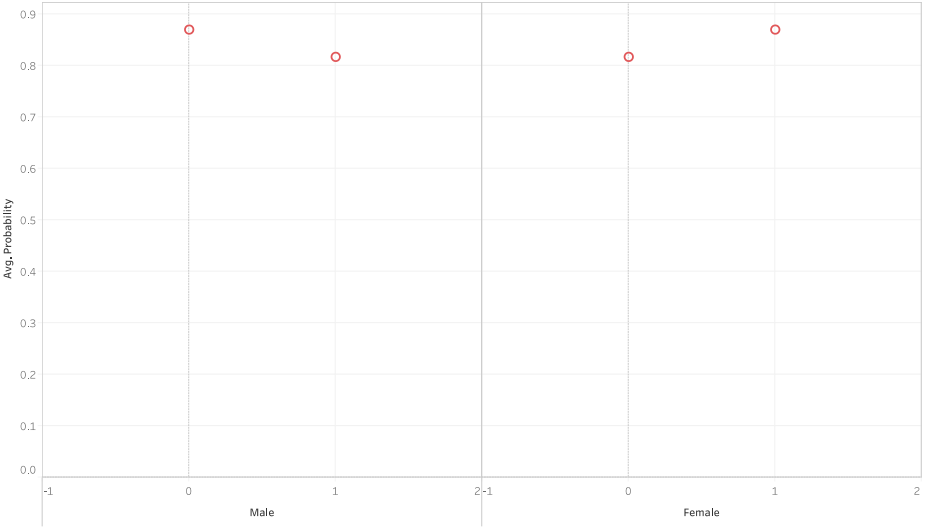

Figure 3.
Probability of academic success versus gender

4. Results and discussion

This study sets out to use enrollment data to identify significant factors that can be used to predict if students in a private university in Nigeria will achieve a passing CGPA. A supervised machine-learning approach was adopted, and the data set was divided into training data (70%) and a test data set (30%). Running a logistics regression model on the training data, the model was trained and able to predict the students' academic success/

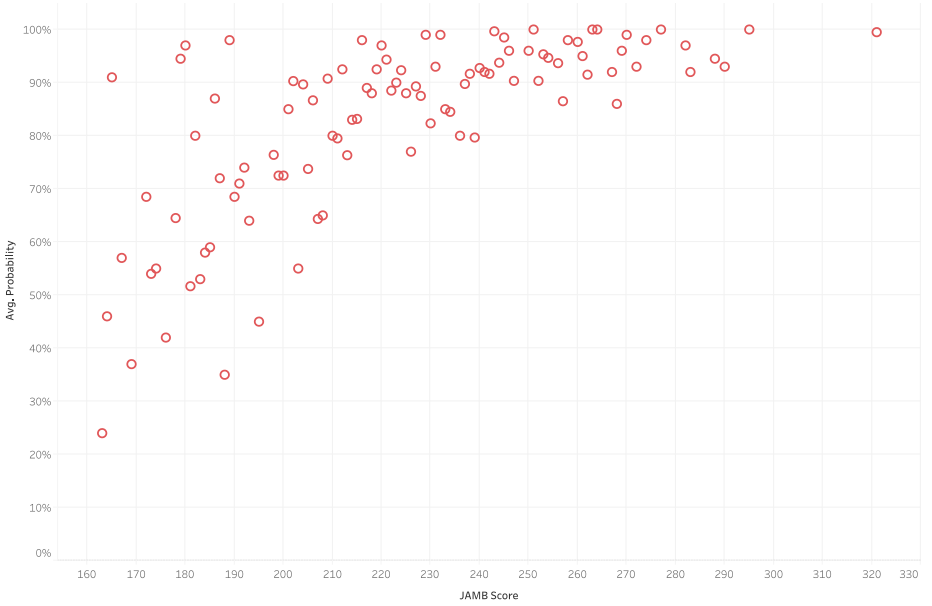

Figure 4.
Probability of academic success versus JAMB score

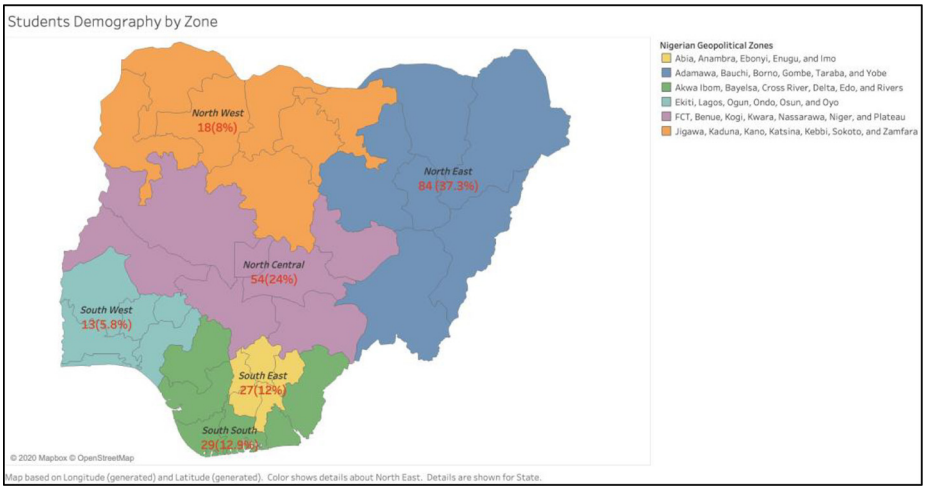

Figure 5.
Student count by geopolitical zones (test data)

failure with an accuracy of 84.7%. The model was then used to the test data set and the accuracy was observed to be 83.5%. Table 2 shows the bias (intercept) as well as the coefficients of the input variables obtained from the trained model. The probability of earning a passing CGPA for each observation in the test data set was extracted along with the corresponding observations and analyzed with the Tableau visualization software. The subsections below will discuss the analysis of the input variables and their associated probabilities based on the test data set.

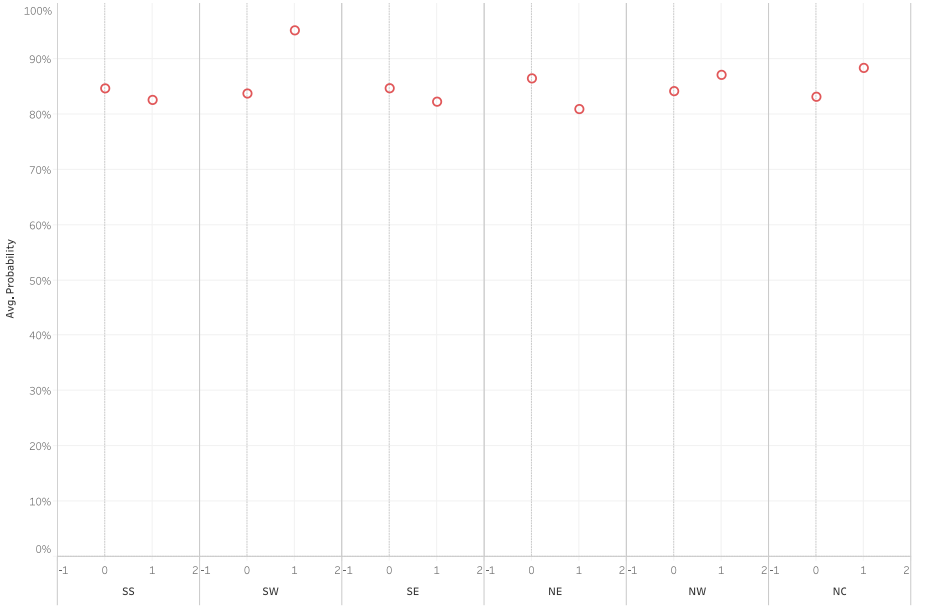

Figure 6.
Probability of academic success versus regions

4.1 Entry age

In this study, entry age signifies the age the student was accepted into university. The influence of the variable entry age in the logistic regression model is negligible with an odds ratio of 1.06 ($\beta = 0.06$). Previous studies have examined the impact of age on performance of students. Öner *et al.* (2018) observed that younger ages in children resulted in lower grades in mathematics as well as inattention and higher hyperactivity. While Öner's study was based on children at a lower education level, the principle applies to university students where academic performance has been observed to depend on demographic data such as age (Fernandes *et al.*, 2019; Helal *et al.*, 2019). Figure 2 plots the average probability of academic success versus the age of students based on the test data set. The entry ages of the students captured in the test data set was between 15 and 26 years of age.

The graph in Figure 2 shows that for the age range of 15 to 21, the probability of achieving a passing CGPA is between 80 and 90%. At the age of 22 years however, there is a noticeable drop. The graph also shows that students above the age of 23 have the highest chances of academic success; however, this could be misleading considering that there are only two students in the test sample over the age of 23. Agreeing with previous research (Mills *et al.*, 2009; Amuda *et al.*, 2016), the findings of this study suggest that age is not a predictor of university students' academic success. A possibility for this observation could be that the entry age for majority (96.88%) of the students fall within the same age group (i.e. youth between 15 and 24 years of age), Table 3, which has been defined as the active and receptive age (Adelakun, 2017). Therefore, it is expected that the students in this age group will have similar social and behavioral traits.

This finding supports the fact that as the level of study for the student increases, the probability of success also increases. This is as a result of maturity that is associated with the increase in age as well as the motivation to complete their studies as they approach graduation.

4.2 Gender: male and female

Table 4 shows that the gender of the students in the test data set is almost balanced, 113 are male (50.2%), while there are 112 female students (49.80%). With an odds ratio of 1.23 (β = 0.21), the female gender attribute is a significant variable in the logistic regression model. The male gender also contributes to the logistic regression model with an odds ratio of 0.81 (β = −0.21). These findings indicate that the odds of academic success are 1.2 times more likely if the student is a female. This supports previous research that shows that gender is a predictor of academic success (Anderton and Chivers, 2016).

Figure 3 also shows that female students have a higher probability of achieving a passing CGPA in comparison to male students. This finding is supported by previous research (Reddy *et al.*, 2017; Adelakun, 2017). Fischer *et al.* (2013) attributes this advantage over male students to females showing a higher accomplishment motivation, stating that female students show "more compensatory effort as well as self-control and taking more pride in their own productivity which helps the female students to outperform their male counterparts" (Fischer *et al.*, 2013). Similarly, Spinath *et al.* (2014) believe that female students are better suited for educational environments as they are verbally more intelligent, stronger in terms of self-discipline and have a higher motivation (Spinath *et al.*, 2014).

4.3 JAMB score

In the test data set, 208 out of the 226 students scored over the 180 JAMB entry score requirement and in comparison, 213 students achieved a passing CGPA. With an odds ratio of 2.55 (β = 0.94), the JAMB score is an important feature in predicting academic success, especially considering that the JAMB score variable was standardized. Admission into Nigerian universities typically depends on cognitive entry characteristics of the students that are mostly academic in nature such as the JAMB and the Senior School Certificate Examination (SSCE) score. Adekitan and Noma-Osaghae (2019) suggest that cognitive entry characteristics of students may not explain academic success, though in their data mining study to predict the performance of first year university students, they observed that

Age	Frequency	(%)
15	7	3.11
16	47	20.89
17	81	36.00
18	44	19.56
19	29	12.89
20	9	4.00
21	3	1.33
22	3	1.33
23	1	0.44
26	1	0.44

Table 3. Age distribution

Gender		
Male	113	50.20
Female	112	49.80

Table 4. Gender distribution

students with a higher JAMB score performed better (Adekitan and Noma-Osaghae, 2019). Bamgboye *et al.* (2001) also observed that students performed better in preclinical MBBS examination than their counterparts with lower JAMB scores though their findings indicate that WAEC/SSCE scores are a better predictor of the students' performance. The trend shown in Figure 4 suggests that the students' probability of academic success increases with an increasing JAMB score and corroborates the findings of Bamgboye *et al.* (2001) and Adekitan and Noma-Osaghae (2019).

4.4 Region (ethnicity)

Nigeria has six geo-political zones based on ethnicity. Figure 5 shows the demographic representation, with regards to the region/zones, of the students in the test dataset. Majority of the students are from the northern part of Nigeria, specifically the northeast and north central due to proximity to the university, which is in the northeast region. Our findings show that the region is a significant predictor for the outcome variable. Students from the southwest were having the highest probability of achieving a passing CGPA with a coefficient of 0.95, followed by students from the north west ($\beta = 0.33$) and the north central ($\beta = 0.30$) regions.

The northeast, southeast and south-south regions had negative coefficients with odds ratios of 0.65, 0.58 and 0.54 respectively. This indicates that for students in either of the three regions, the odds of achieving a CGPA of 2.0 or more are 35, 43 and 46% lower than the base model (Figure 6). Contrariwise to our findings, past research found that ethnic background is not a predictor of students' final graduation results (Adekitan and Salau, 2020). Their study uses the following predictive algorithms: classification tree, random forest, Naïve Bayes and neural networks. Similarly, Ojetunde *et al.* (2019) did not find evidence to support that the geopolitical/ethnic background of students determines the degree outcome of undergraduate students using a logistic regression model.

4.5 Level of study

In this study, the duration of all the programs available is 4 years with the exception of the law and engineering programs, which last for 5 years. These two programs are relatively new, and the most senior classes are year 4 and year 1, respectively. Table 5 shows the number of students in each year within the test dataset. As expected, the frequency decreases as the level increases due to students dropping out or not meeting the requirements to enter the next level. Eddy (2016), suggests that as students' progress in their academics, the courses they take tend to be less general and more difficult in nature. This would be one of the reasons why there is a drop in the number of students as their education level increases.

The Year 4 ($\beta = 1.71$) and Year 3 ($\beta = 1.21$) variables have the largest coefficients in the logistic regression model. This is an expected result as the students at this stage will be more dedicated and put in more effort not only to graduate but also to improve their CGPA

Table 5. Study year distribution

	Frequency	(%)
Year		
Year 1	73	32.40
Year 2	60	26.70
Year 3	52	23.10
Year 4	40	17.80

to obtain a higher classed degree. As expected, the Year 1 variable ($\beta = -1.24$) has the lowest coefficient followed by Year 2 ($\beta = -1.76$) among the student years as this is where students are adjusting to the demands of a new academic and social environment (Junco, 2015). While it is expected that the courses students take become more difficult as they progress in their education, the probability of them achieving a passing CGPA increases as they are more determined to excel and those who are less likely to succeed would have dropped out of the program (Figure 7).

4.6 Implications for theory and practice

In this study, a supervised machine-learning approach was used to predict students' academic success using enrolment data. The enrolment data was trained to accurately predict 84.7% of the students CGPA using the logistics regression algorithm. The algorithm was applied to a test data resulting in 83.5% accuracy. The most significant contributors in the model, of decreasing magnitude, include the following variables: Year 4, Year 3, SW, JAMB Score, Gender, NC and NW. The present work unveils catchy theoretical implications. A burgeoning stream of research highlighted several limitations of self-reported survey data and LMS data in predicting students' academic success (Conijn *et al.*, 2016; Rincón-Flores *et al.*, 2020; Tempelaar *et al.*, 2020; Tempelaar *et al.*, 2015).

In light of these reasonings, this paper strives to predict students' academic success using machine learning approach and enrolment data from SIS. In doing so, this paper responds to research calls and adds insights to this line of research. Theoretically, a complementary perspective is needed to move beyond the approach of emphasis on one or two predictors of student academic performance toward complementary considerations. Thus, this paper adds to the literature by showing the value and complementary nature of several predictors in Sub-Saharan region, where limited empirical findings are available. This work also paves the way for future work towards exploration of potential moderators

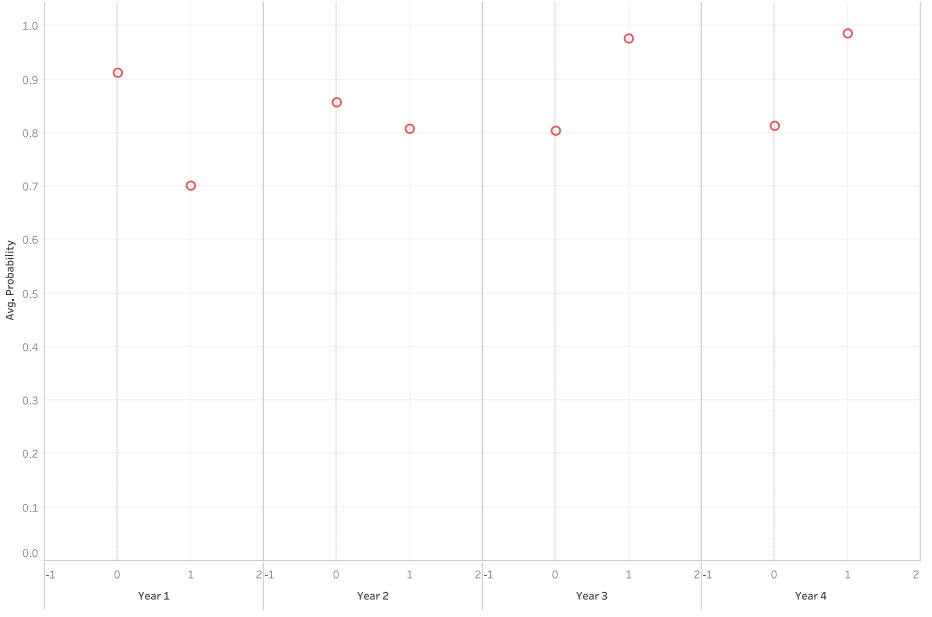

Figure 7.
Probability of academic success versus level of study

and mediators. We envisage research opportunities that leverage other machine learning algorithms.

This present study divulged several complementary practical implications. Structural disadvantage appears to have adverse effects on student outcomes. Past work in the UK echoed similar arguments, as BAME (Black, Asian and Minority Ethnic) students from lower socio-economic backgrounds were less likely to attain academic success (Crawford, 2014; Soria and Stebleton, 2012). The results show that students from the northeast, southeast and south-south do not perform as well as students from the northwest, south west and north central. This could be attributed to the quality of the universities and teachers, which have been affected by conflicts, such as Boko Haram in the North east (Bertoni *et al.*, 2019; Abiodun *et al.*, 2020) and Niger Delta militants in the south-south region (Pepple and Ogologo, 2017).

Technically, the three regions can be characterized as having structural disadvantage, that are inflated by other factors such as community and engagement (influence student's intellect, behavior and desire for studies) and finances in the south east. Similar arguments were echoed by (Crawford, 2014; Morrow and Ackermann, 2012). Federal, state and local governments need to do more to support students in these regions. *One*, interventions programs to develop the quality of tutors, tutoring and learning activities should be implemented in the region. *Two*, government needs to increase educational investments and staff posting, for instance, experience and high performing professors and lecturers should be posted to these regions to revive the educational system in the region starting from primary to tertiary level. Alternatively, pay increase can be used in the affected regions to attract teachers, lecturers and professors to work in the regions.

Three, the ability to identify students who are at risk of achieving a passing CGPA is important for educators as they can apply timely intervention measures to aid the students which can lead to a reduction of attrition rates. For instance, based on the findings of this study, universities are advised to focus on students at risk, e.g. Years 1 and 2 students. The academic advising unit as well as the school psychologist can arrange sessions with the students that aims at helping them adjust to the demands of a new academic and social environment. Societies, social and cultural students' clubs are facilitators for greater integration among new students, as integration increases, information flows within the students which could serve as a helper in academic activities. At faculty and departmental level, peer-to-peer tutoring can be offered where high performing students can assist new, struggling and/or students at risk of failing.

Four, apart from encouraging peer-to-peer tutoring, implementing this model could help universities manage intellectual talent, in that academically gifted and motivated students are to be identified, attracted, groomed and guided to pursue their career of interests, who might be attracted to other universities. In particular, their accomplishments can not only create an academically vigorous climate for competition and knowledge sharing but also boost the universities' prestige and reputation as highlighted in past work (Miguéis *et al.*, 2018).

Fifth, the results from this study can also help in recruitment of students. The ability to identify students who are at risk of achieving a passing CGPA is important for educators as they can apply timely intervention measures to aid the students, which can lead to a reduction of attrition rates. Most universities used to accept a minimum JAMB score of 180, this study shows that majority of the students that scored between 160 and 180 falls into the group that has more than a 50% chance of achieving a passing CGPA. Therefore, admission teams can use the findings of this study to improve their decision-making process for admitting students into the university. Past works made similar suggestions about the

opportunities surrounding knowledge management and organizational intelligence (Danõa *et al.*, 2020; Holley, 2009; Metcalfe, 2005).

4.7 Strengths, limitations and future research direction

Applicable to most research, this study has several limitations and strengths that should be noted. *In terms of limitations*, limited number of variables were available in the student information system at the time of the study. For instance, West African Examinations Council (WAEC) results were not considered, and past work shows WAEC results as important predictors (Bamgboye *et al.*, 2001). *Second*, other important variables such as personality traits, stress, anxiety and motivation are not captured which limits our ability to draw sound conclusions. Future studies are encouraged to consider pre-enrollment data (i.e. WAEC results) and other variables, for instance, a mixture of secondary that captures demographic data and primary data that captures perceptions, can be used to model and predict students' academic performance.

Third, the present work uses one machine learning algorithm (i.e. logistics regression). Future studies can compare the logistics regression model with other machine learning models, specifically, classification methods such as neural networks, decision trees, random forests, for additional insights. Past work showed that a combination of machine learning techniques can help to compare the accuracy and performance and identify the suitable techniques for any given situation (Rusli *et al.*, 2008; Amirhajlou *et al.*, 2019). *In terms of strengths*, self-reported and cross-sectional data are vulnerable to common method variance (Podsakoff *et al.*, 2012). This study uses secondary and multisource data to evade the potential threat of common method variance and to increase our ability to draw concrete causal inference. Finally, the use of machine learning technique increases the reliability and validity of our findings, which is almost impossible using conventional methods.

References

Abiodun, T.F., Omolayo, O.O., Tomisin, A.D. and Chinedu, O.C. (2020), "Assessment of boko haram insurgents' threats to educational development in the northeast Nigeria: the way forward", *Assessment*, Vol. 3 No. 1, pp. 31-43.

Adekitan, A.I. and Noma-Osaghae, E. (2019), "Data mining approach to predicting the performance of first year student in a university using the admission requirements", *Education and Information Technologies*, Vol. 24 No. 2, pp. 1527-1543.

Adekitan, A.I. and Salau, O. (2020), "Toward an improved learning process: the relevance of ethnicity to data mining prediction of students' performance", *SN Applied Sciences*, Vol. 2 No. 1, p. 8.

Adelakun, O.E. (2017), "Gender disparity in academic performance of students in the faculty of agriculture and forestry, university of Ibadan, Oyo state", *International Journal of Agricultural Economics and Rural Development*, Vol. 9 No. 1.

Ahmad, F., Ismail, N.H. and Aziz, A.A. (2015), "The prediction of students' academic performance using classification data mining techniques", *Applied Mathematical Sciences*, Vol. 9 No. 129, pp. 6415-6426.

Akgül, A. (2018), "A novel method for a fractional derivative with non-local and non-singular kernel", *Chaos, Solitons and Fractals*, Vol. 114, pp. 478-482.

Akgül, E.K. (2019), "Solutions of the linear and nonlinear differential equations within the generalized fractional derivatives", *Chaos: An Interdisciplinary Journal of Nonlinear Science*, Vol. 29 No. 2, p. 023108.

Al-Sudani, S. and Palaniappan, R. (2019), "Predicting students' final degree classification using an extended profile", *Education and Information Technologies*, Vol. 24 No. 4, pp. 2357-2369.

Alyahyan, E. and Düştegör, D. (2020), "Predicting academic success in higher education: literature review and best practices", *International Journal of Educational Technology in Higher Education*, Vol. 17 No. 1.

Amirhajlou, L., Sohrabi, Z., Alebouyeh, M.R., Tavakoli, N., Haghighi, R.Z., Hashemi, A. and Asoodeh, A. (2019), "Application of data mining techniques for predicting residents' performance on pre-board examinations: a case study", *Journal of Education and Health Promotion*, Vol. 8.

Amuda, B.G., Bulus, A.K. and Joseph, H.P. (2016), "Marital status and age as predictors of academic performance of students of colleges of education in the North-Eastern Nigeria", *American Journal of Educational Research*, Vol. 4 No. 12, pp. 896-902.

Anderton, R. and Chivers, P. (2016), "Predicting academic success of health science students for first year anatomy and physiology", *International Journal of Higher Education*, Vol. 5 No. 1.

Aulck, L., Nambi, D., Velagapudi, N., Blumenstock, J. and West, J. (2019), "Mining university registrar records to predict First-Year undergraduate attrition", *International Educational Data Mining Society*.

Baleanu, D., Fernandez, A. and Akgül, A. (2020), "On a fractional operator combining proportional and classical differ integrals", *Mathematics*, Vol. 8 No. 3, p. 360.

Bamgboye, E.A., Ogunnowo, B.E., Badru, O.B. and Adewoye, E.O. (2001), "Students admission grades and their performance at Ibadan university pre-clinical MBBS examinations", *African Journal of Medicine and Medical Sciences*, Vol. 30 No. 3, pp. 207-211.

Bawack, R.E. and Kamdjoug, J.R.K. (2020), "The role of digital information use on student performance and collaboration in marginal universities", *International Journal of Information Management*, Vol. 54, p. 102179.

Berens, J., Schneider, K., Görtz, S., Oster, S. and Burghoff, J. (2018), "Early detection of students at risk–predicting student dropouts using administrative student data and machine learning methods", *Journal of Educational Data Mining*, Vol. 11 No. 3, pp. 1-41.

Bertoni, E., Di Maio, M., Molini, V. and Nistico, R. (2019), "Education is forbidden: the effect of the boko haram conflict on education in North-East Nigeria", *Journal of Development Economics*, Vol. 141, p. 102249.

Bramer, M. (2016), *Principles of Data Mining*, Springer, London.

Conijn, R., Snijders, C., Kleingeld, A. and Matzat, U. (2016), "Predicting student performance from LMS data: a comparison of 17 blended courses using moodle LMS", *IEEE Transactions on Learning Technologies*, Vol. 10 No. 1, pp. 17-29.

Cooper, H., Carlisle, C., Gibbs, T. and Watkins, C. (2001), "Developing an evidence base for interdisciplinary learning: a systematic review", *Journal of Advanced Nursing*, Vol. 35 No. 2, pp. 228-237.

Crawford, C. (2014), "Socio-economic differences in university outcomes in the UK: drop-out, degree completion and degree class (no. W14/31)", IFS Working Papers.

Danõa, J., Caputo, F. and Ráček, J. (2020), "Complex network analysis for knowledge management and organizational intelligence", *Journal of the Knowledge Economy*, Vol. 11 No. 2, pp. 405-424.

Eddy, T.M. (2016), *Reflecting Back: Do Senior Students Believe They Experienced a Sophomore Slump*?.

Fernandes, E., Holanda, M., Victorino, M., Borges, C., Carvalho, R. and Van Erven, G. (2019), "Educational data mining: predictive analysis of academic performance of public-school students in the capital of Brazil", *Journal of Business Research*, Vol. 94, pp. 335-343.

Fischer, F., Schult, J. and Hell, B. (2013), "Sex differences in secondary school success: why female students perform better", *European Journal of Psychology of Education*, Vol. 28 No. 2, pp. 529-543.

Francis, P., Broughan, C., Foster, C. and Wilson, C. (2020), "Thinking critically about learning analytics, student outcomes, and equity of attainment", *Assessment and Evaluation in Higher Education*, Vol. 45 No. 6, pp. 811-821, doi: 10.1080/02602938.2019.1691975.

Garg, R. (2018), "Predict student performance in different regions of Punjab", *International Journal of Advanced Research in Computer Science*, Vol. 9 No. 1, pp. 236-241.

Gašević, D., Dawson, S., Rogers, T. and Gasevic, D. (2016), "Learning analytics should not promote one size fits all: the effects of instructional conditions in predicting academic success", *The Internet and Higher Education*, Vol. 28, pp. 68-84.

Hamoud, A.K., Hashim, A.S. and Awadh, W.A. (2018), "Predicting student performance in higher education institutions using decision tree analysis", *International Journal of Interactive Multimedia and Artificial Intelligence*, Vol. 5 No. 2.

Helal, S., Li, J., Liu, L., Ebrahimie, E., Dawson, S. and Murray, D.J. (2019), "Identifying key factors of student academic performance by subgroup discovery", *International Journal of Data Science and Analytics*, Vol. 7 No. 3, pp. 227-245.

Hoffait, A.S. and Schyns, M. (2017), "Early detection of university students with potential difficulties", *Decision Support Systems*, Vol. 101, pp. 1-11.

Holley, K.A. (2009), "Understanding interdisciplinary challenges and opportunities in higher education", *ASHE Higher Education Report*, Vol. 35 No. 2, pp. 1-131.

Huang, A.Y., Lu, O.H., Huang, J.C., Yin, C.J. and Yang, S.J. (2020), "Predicting students' academic performance by using educational big data and learning analytics: evaluation of classification methods and learning logs", *Interactive Learning Environments*, Vol. 28 No. 2, pp. 206-230.

Junco, R. (2015), "Student class standing, Facebook use, and academic performance", *Journal of Applied Developmental Psychology*, Vol. 36, pp. 18-29.

Kamal, P. and Ahuja, S. (2019), "Academic performance prediction using data mining techniques: identification of influential factors effecting the academic performance in undergrad professional course", *Harmony Search and Nature Inspired Optimization Algorithms*, Springer, Singapore, pp. 835-843.

Kanyip, B.P. (2013), *Admission Crisis in Nigerian Universities: The Challenges Youth and Parents Face in Seeking Admission*.

Kovacic, Z. (2010), *Early Prediction of Student Success: Mining Students' Enrolment Data*.

Kumari, K. and Yadav, S. (2018), "Linear regression analysis study", *Journal of the Practice of Cardiovascular Sciences*, Vol. 4 No. 1, p. 33.

Lee, T.R., Wood, W.T. and Phrampus, B.J. (2019), "A machine learning (kNN) approach to predicting global seafloor total organic carbon", *Global Biogeochemical Cycles*, Vol. 33 No. 1, pp. 37-46.

Metcalfe, A. (Ed.), (2005), *Knowledge Management and Higher Education: A Critical Analysis: A Critical Analysis*, IGI Global.

Miguéis, V.L., Freitas, A., Garcia, P.J. and Silva, A. (2018), "Early segmentation of students according to their academic performance: a predictive modelling approach", *Decision Support Systems*, Vol. 115, pp. 36-51.

Mills, C., Heyworth, J., Rosenwax, L., Carr, S. and Rosenberg, M. (2009), "Factors associated with the academic success of first year health science students", *Advances in Health Sciences Education*, Vol. 14 No. 2, pp. 205-217.

Mohamed, M.H. and Waguih, H.M. (2017), "Early prediction of student success using a data mining classification technique", *International Journal of Science and Research*, Vol. 610, pp. 126-131.

Moreno-Marcos, P.M., Pong, T.C., Muñoz-Merino, P.J. and Kloos, C.D. (2020), "Analysis of the factors influencing learners' performance prediction with learning analytics", *IEEE Access*, Vol. 8, pp. 5264-5282.

Morrow, J. and Ackermann, M. (2012), "Intention to persist and retention of first-year students: the importance of motivation and sense of belonging", *College Student Journal*, Vol. 46 No. 3, pp. 483-491.

Musumeci, F., Rottondi, C., Nag, A., Macaluso, I., Zibar, D., Ruffini, M. and Tornatore, M. (2018), "An overview on application of machine learning techniques in optical networks", *IEEE Communications Surveys and Tutorials*, Vol. 21 No. 2, pp. 1383-1408.

Mwalumbwe, I. and Mtebe, J.S. (2017), "Using learning analytics to predict students' performance in moodle learning management system: a case of Mbeya university of science and technology", *The Electronic Journal of Information Systems in Developing Countries*, Vol. 79 No. 1, pp. 1-13.

Ojetunde, I., Sule, A.I., Kemiki, O.A. and Olatunji, I.A. (2019), "Factors affecting the academic performance of real estate students in a specialized federal university of technology in Nigeria", *Property Management*, Vol. 38 No. 2.

Olaya, D., Vásquez, J., Maldonado, S., Miranda, J. and Verbeke, W. (2020), "Uplift modeling for preventing student dropout in higher education", *Decision Support Systems*, Vol. 134, p. 113320.

Öner, Ö., Çalışır, Ö., Ayyıldız, N., Çelikağ, I., Uran, P., Olkun, S. and Çiçek, M. (2018), "Effects of changed school entry rules: age effects within third grade students", *EURASIA Journal of Mathematics, Science and Technology Education*, Vol. 14 No. 6, pp. 2555-2562.

Oyedeji, A.O., Salami, A.M., Folorunsho, O. and Abolade, O.R. (2020), "Analysis and prediction of student academic performance using machine learning", *JITCE (Journal of Information Technology and Computer Engineering)*, Vol. 4 No. 1, pp. 10-15.

Paliwal, M. and Kumar, U.A. (2009), "A study of academic performance of business school graduates using neural network and statistical techniques", *Expert Systems with Applications*, Vol. 36 No. 4, pp. 7865-7872.

Pepple, T.F. and Ogologo, G.A. (2017), "The effects of the Niger Delta crisis on educational resources, attitude to schooling, and academic achievement of basic science students in Rivers state, Nigeria", *Journal of the International Society for Teacher Education*, Vol. 21 No. 1, pp. 67-76.

Podsakoff, P.M., MacKenzie, S.B. and Podsakoff, N.P. (2012), "Sources of method bias in social science research and recommendations on how to control it", *Annual Review of Psychology*, Vol. 63 No. 1, pp. 539-569.

Putpuek, N., Rojanaprasert, N., Atchariyachanvanich, K. and Thamrongthanyawong, T. (2018), "Comparative study of prediction models for final GPA score: a case study of rajabhat rajanagarindra university", *2018 IEEE/ACIS 17th International Conference on Computer and Information Science*, pp. 92-97.

Reddy, B.V., Gupta, A. and Singh, A.K. (2017), "A study to assess factors affecting the performance of undergraduate medical students in academic examination in community medicine", *Int J Med. Public Health*, Vol. 4 No. 4, pp. 1066-1070.

Rincón-Flores, E.G., López-Camacho, E., Mena, J. and López, O.O. (2020), April). "Predicting academic performance with artificial intelligence (AI), a new tool for teachers and students", *2020 IEEE Global Engineering Education Conference (EDUCON)*, *IEEE*, pp. 1049-1054.

Rizvi, S., Rienties, B. and Khoja, S.A. (2019), "The role of demographics in online learning; a decision tree based approach", *Computers and Education*, Vol. 137, pp. 32-47.

Rusli, N.M., Ibrahim, Z. and Janor, R.M. (2008), "Predicting students' academic achievement: comparison between logistic regression, artificial neural network, and neuro-fuzzy", *2008 International Symposium on Information Technology*, IEEE, Vol. 1, pp. 1-6.

Saa, A.A., Al-Emran, M. and Shaalan, K. (2019), "Mining student information system records to predict students' academic performance", *International Conference on Advanced Machine Learning Technologies and Applications*, Springer, Cham, pp. 229-239.

Sedrakyan, G., Malmberg, J., Verbert, K., Järvelä, S. and Kirschner, P.A. (2020), "Linking learning behavior analytics and learning science concepts: designing a learning analytics dashboard for feedback to support learning regulation", *Computers in Human Behavior*, Vol. 107, p. 105512.

Shahiri, A.M. and Wahidah, H. (2015), "A review on predicting student's performance using data mining techniques", *Procedia Computer Science*, Vol. 72, pp. 414-422.

Shum, S.B. and Ferguson, R. (2012), "Social learning analytics", *Journal of Educational Technology &Society*, Vol. 15 No. 3, pp. 3-26.

Sin, K. and Muthu, L. (2015), "Application of big data in education data minning and learning analytics–a literature review", *ICTACT Journal on Soft Computing*, Vol. 5 No. 4.

Soofi, A.A. and Awan, A. (2017), "Classification techniques in machine learning: applications and issues", *Journal of Basic and Applied Sciences*, Vol. 13, pp. 459-465.

Soria, K.M. and Stebleton, M.J. (2012), "First-generation students' academic engagement and retention", *Teaching in Higher Education*, Vol. 17 No. 6, pp. 673-685.

Spinath, B., Eckert, C. and Steinmayr, R. (2014), "Gender differences in school success: what are the roles of students' intelligence, personality and motivation?", *Educational Research*, Vol. 56 No. 2, pp. 230-243.

Tempelaar, D.T., Rienties, B. and Giesbers, B. (2015), "In search for the most informative data for feedback generation: Learning analytics in a data-rich context", *Computers in Human Behavior*, Vol. 47, pp. 157-167.

Tempelaar, D., Rienties, B. and Nguyen, Q. (2020), "Subjective data, objective data and the role of bias in predictive modelling: lessons from a dispositional learning analytics application", *Plos One*, Vol. 15 No. 6, p. e0233977.

Yaacob, W.F., Nasir, S.A., Yaacob, W.F. and Sobri, N.M. (2019), "Supervised data mining approach for predicting student performance", *Indonesian Journal of Electrical Engineering and Computer Science*, Vol. 16 No. 3, pp. 1584-1592.

Further reading

Akçapınar, G., Altun, A. and Aşkar, P. (2019), "Using learning analytics to develop early-warning system for at-risk students", *International Journal of Educational Technology in Higher Education*, Vol. 16 No. 1, p. 40.

Corresponding author

Mohammed Nasiru Yakubu can be contacted at: yakubu.m@aun.edu.ng