

Received:
22 October 2018
Revised:
10 January 2019
Accepted:
11 February 2019

Cite as:
Aderibigbe Israel Adekitan,
Odunayo Salau. The impact of
engineering students'
performance in the first three
years on their graduation
result using educational data
mining.
Heliyon 5 (2019) e01250.
doi: [10.1016/j.heliyon.2019.e01250](https://doi.org/10.1016/j.heliyon.2019.e01250)



The impact of engineering students' performance in the first three years on their graduation result using educational data mining

Aderibigbe Israel Adekitan^{a,*}, Odunayo Salau^b

^a Department of Electrical and Information Engineering, Covenant University, Ota, Ogun State, Nigeria

^b Department of Business Management, Covenant University, Ota, Ogun State, Nigeria

* Corresponding author.

E-mail address: aderibigbe.adekitan@covenantuniversity.edu.ng (A.I. Adekitan).

Abstract

Research studies on educational data mining are on the increase due to the benefits obtained from the knowledge acquired from machine learning processes which help to improve decision making processes in higher institutions of learning. In this study, predictive analysis was carried out to determine the extent to which the fifth year and final Cumulative Grade Point Average (CGPA) of engineering students in a Nigerian University can be determined using the program of study, the year of entry and the Grade Point Average (GPA) for the first three years of study as inputs into a Konstanz Information Miner (KNIME) based data mining model. Six data mining algorithms were considered, and a maximum accuracy of 89.15% was achieved. The result was verified using both linear and pure quadratic regression models, and R^2 values of 0.955 and 0.957 were recorded for both cases. This creates an opportunity for identifying students that may graduate with poor results or may not graduate at all, so that early intervention may be deployed.

Keywords: Education, Information science, Computer science

1. Introduction

Higher institutions are setup to provide quality education capable of transforming the level of awareness, knowledge, and the capacity of the human mind. In Africa, the educational sector has suffered many setbacks due to under development, economic hardship, insufficient budget and corruption. Nigeria is the most populous country in Africa, and the country has the largest university system in the sub-Saharan region (Olaleye and Oyewole, 2016). Engineering education is a driver of global sustainable development and innovation (Agboola and Elinwa, 2013). In Nigeria, engineering education is regulated by the National University Commission (NUC) that acts as a proxy for the government (Mahmud et al., 2012). NUC and the Council for the Regulation of Engineering in Nigeria (COREN) together with relevant stakeholders provide regulatory oversight on the activities and operations of Nigerian universities via relevant policies aimed at ensuring quality in the educational system (Agboola and Elinwa, 2013). NUC provides the regulatory needs for university education in Nigeria and ensures along with other stakeholders (Akerlele, 2008), the implementation of all quality assurance policies and periodic accreditation of Nigerian universities (Ajayi and Ekundayo, 2008; Akerlele, 2008) toward turning out quality students (Idachaba, 2018; Obadara and Alaka, 2013).

A prospective student is required to satisfy the minimum entry requirements as stipulated by NUC to be qualified for admission consideration by any Nigerian University (Adeyemi, 2001; Odukoya et al., 2018; Oladipo et al., 2009). Engineering curriculum in Nigerian universities run for 5 years comprising ten academic semesters. In the second semester of the 4th year of an engineering program, students go on internship termed Student's Industrial Work Experience Scheme (SIWES) for the full semester and return the next session as final year students. The internship program is aimed at bridging the gap between theory and practice, as it exposes students to the realities of the engineering profession.

In the first and second year of an engineering program, students are exposed to knowledge on sciences and basic introduction to engineering as a continuum of their secondary school education, and as an introduction to general engineering. In the third year, the curriculum is more focused on the core discipline of each engineering student, that is, electrical engineering, mechanical engineering, civil engineering, and so forth. By the end of the third year, engineering students are already grounded in the basics of their profession. The academic performance of engineering students from their first year to the third year is very vital in terms of acquisition of foundational knowledge, and its impact on their final graduation Cumulative Grade Point Average (CGPA). It is often said that beyond the third year it is very challenging for a student to move from the current class of grade (first class - 1st, second class upper division - 2|1, second class lower division - 2|2, and third class - 3rd) to a

higher one due to the nature of academic courses at 400L (fourth year) and 500L (fifth year) which are more robust and touch core foundation of engineering disciplines. The Nigerian educational curriculum has also been criticized as being overloaded and not adequately meeting the societal knowledge needs of the Nigerian learner (Obadara and Alaka, 2013; Omoregie, 2008).

The students' grades are classified as follows; first class for a CGPA greater than 4.50, second class upper division for a CGPA between 3.50 to 4.49, second class lower division for the CGPA range of 2.50–3.49, and third class for a CGPA of 1.50–2.49. A student with a CGPA of less than 1.50 will be placed on academic probation in the following semester, and if already in the final semester, it is classified as a fail grade in Covenant university. A student that does not satisfy the requirements for the honours' category but with a CGPA of at least 1.0 may be awarded a pass degree in some Nigerian universities. To graduate with a good class of grade, students must work towards it early right from 100L (first year) because by 300L (third year) the CGPA would have taken form and may be had to improve. This study considers the impact of the Grade Point Average (GPA) of engineering students from 100L to 300L on the final graduation CGPA using Covenant University in Nigeria as a case study. An attempt was made in this study to establish the extent to which the final year CGPA of an engineering student can be predicted using the GPA of the first three years of study. The academic performance dataset of 1841 students that were admitted and graduated within the period of 2002–2014 across seven engineering programs in Covenant University was analysed using a predictive Konstanz Information Miner (KNIME) based data mining model and regression analysis in MATLAB.

2. Background

The quality of student-teacher interaction, student participation and engagement processes are vital indices that affect student overall performance, failure and dropout rate. For optimal learning and for an educational system to be able to improve on its practices and gaps, it is vital to have a means of collecting performance related feedback data (Daradoumis et al., 2019) and deploying systemic appraisal towards identifying weaknesses in the delivery of education to the student populace. Educational data mining is one of the methods that can be used for pattern recognition and trend identification.

Data mining is the extraction of hidden useful information from a dataset through scientific analysis and methods that identify data trends, and hidden patterns within the given dataset, and as such, data mining can be referred to as knowledge discovery (Azevedo, 2018; Hussain et al., 2018). The discovery of hidden information is achieved by running data mining algorithms that combine statistics with computer

science to mine valuable information from a seemingly meaningless data jumble. Data mining can be applied in various fields such as engineering (Adekitan et al., 2019; Saini and Aggarwal, 2018), business management (Zuo et al., 2016), marketing and product design (Jin et al., 2019), computer science (Mahendra et al., 2019), education (Ibrahim et al., 2019; Porouhan, 2018), genetics (Noreña et al., 2018), biological studies (Gu et al., 2018), facility maintenance management (Miguel-Cruz et al., 2019), health and drug development studies (Keserci et al., 2017), chemistry and toxicity analysis (Saini and Srivastava, 2019), meteorology (Kovalchuk et al., 2019), transportation safety (Divya et al., 2019) and traffic management (Amiruzzaman, 2019), fraud detection (Vardhani et al., 2019), and so forth. In the educational sector, volumes of data are daily generated from various teaching and learning activities within an institution. Educational data mining is the use of data mining techniques to extract vital information from a dataset generated within the educational context. Such information provides clues on previously unknown trends that relate to student performance (Adekitan and Noma-Osaghae, 2018; Roy and Garg, 2018) and learning behaviours (Kim et al., 2018), the efficiency and quality of teaching, student potential prediction (Yang and Li, 2018), the propensity of students to dropout etc. This provides an opportunity for identifying quality gaps, and for proposing improvement and necessary policies for improving the quality, and the delivery of educational systems (Agarwal et al., 2012; Baepler and Murdoch, 2010; Osmanbegović and Suljic, 2012).

Educational data mining is a machine learning process that has been applied for studying and predicting student performance (Asif et al., 2017; Gasevic et al., 2014; Kostopoulos et al., 2018), for evaluating learning technologies integration process (Angeli et al., 2017), and for identifying learning challenges. Relevant educational dataset that has accumulated overtime and depicts the operational process under evaluation must be available, and adequately processed to support a data mining analysis toward achieving a reasonable accuracy (Almarabeh, 2017).

Research studies have been carried out to enable the prediction of the final graduation CGPA of students using the lower level performance grades of the students. In the study by Pelumi Oguntunde et al. (2018), the final CGPA of students was predicted using multiple linear regression and correlation to analyse the yearly GPA, and various inferential statistics were developed. The study determined the correlation between the first-year result and the final-year result of the student. With the aid of a regression plot, the students' GPA for the five years of study was fitted using multiple linear regression in order to explain how the GPA for each year contributed to the variations in the final CGPA of the students at graduation. In (Bucos and Drăgulescu, 2018) features such as student attendance, average scores, relevant course data, the level of student participation in class etc. were deployed in a data mining model for predicting the performance of 908 students. A decision tree model was applied by Ahmed and Elaraby (2014) to predict the probability of failure of

1,547 students such that relevant knowledge can be acquired that will enable the management team to be able to deploy adequate and early intervention. In the study, the student grades were classified into five categories, and these are: excellent, very good, good, acceptable and fail. Ten input features that include the student's department, high school grades, level of participation in class, attendance, midterm scores, lab reports, homework grades, seminar score, completion of assignments and the overall grades were applied in the decision tree model developed. Similarly, in [Yadav et al. \(2012\)](#) by using decision tree classifiers, the likelihood of a student to drop out of an institution was predicted through educational data mining.

In [Tair and El-Halees \(2012\)](#), association, classification, clustering and outlier detection data mining techniques were applied to analyse 3,314 graduate student performance records over a fifteen-year period. The dataset was analysed using Rule Induction, Naïve Bayesian classifier, K-Means clustering algorithm followed by density-based and distance-based outlier detection methods. 18 attributes of the student dataset were considered, and only 6 attributes: matriculation GPA, gender, specialty of the students, the city of the student, the grade and the type of secondary school attended were selected for the data mining analysis. The remaining 12 attributes were dropped due to their large variances and because some of the attributes are personal information that did not provide useful knowledge. The unsupervised clustering analysis performed, identified four unique clusters in the dataset using k-means algorithm. Data mining method was applied by [Al-Radaideh et al. \(2006\)](#) to evaluate student data towards identifying the key attributes that influence the academic performance of students. This provides an opportunity for improving the quality of higher education. In [\(Kabakchieva, 2013\)](#), data mining technique was applied to analyse student data at a Bulgarian university. The student dataset that was analysed, contained the personal and pre-admission attributes of each student. The Decision Tree Classifiers (J48), k-Nearest Neighbour, Bayesian, Naïve Bayes classifiers, the OneR, and the JRip Rule learners were applied to extract knowledge from the student dataset, and accuracy of 52–67% was achieved. The result showed that the number of courses failed in the first academic year and the admission score of the student are two major features among the very influential features in the classification analysis.

3. Analysis

The dataset investigated in this study contains the GPA data for the first three academic years and the final CGPA of 1,841 students from 2002 to 2014, across seven engineering departments, and these are: Information and Communication Engineering (ICE), Chemical Engineering (CHE), Computer Engineering (CE), Mechanical Engineering (MECH), Electrical and Electronics Engineering (ELECT), Civil (CVE), and Petroleum Engineering (PE).

The student performance dataset of the engineering students of Covenant University in Nigeria was obtained from the study by Popoola et al. (2018). The dataset file in spreadsheet format is named “Supplementary material” and is available from Popoola et al. (2018) via a link in appendix A. The statistical attributes of the dataset were determined and presented in Table 1 as descriptive statistics. The GPA data for the first year, the second year, the third year and the final CGPA are fitted using Gamma Distribution, Normal Distribution, Weibull Distribution and Logistic Distribution respectively. Fig. 1(a) shows the probability density function plot and Fig. 1(b) shows the cumulative probability function plot for the GPA of the engineering students in the first year. Fig. 2(a) shows the probability density function plot and Fig. 2(b) shows the cumulative probability function plot for the GPA of the engineering students in the second year. The probability density function plot and the cumulative probability function plot for the third year GPA is presented in Fig. 3(a) and (b) respectively while the probability density function plot and the cumulative probability function plot for the final year CGPA is presented in Fig. 4(a) and (b) respectively.

The key features of the dataset comprise the course of study or program, the year of entry, the first-year GPA, the second-year GPA, and the third-year GPA. The class of degree (1st, 2nd, 2nd and 3rd class) was applied as the target in the KNIME model developed while in the regression models, the class of degree was replaced with the actual CGPA as the model target.

Table 1. Descriptive statistics of 1841 students’ yearly GPA and final CGPA.

	Min	Max	Mean	Std. deviation	Variance	Skewness	Kurtosis
First Year GPA	1.6000	4.9600	3.7977	0.6591	0.4344	-0.6254	0.0265
Second Year GPA	1.1900	4.9600	3.3070	0.7435	0.5528	-0.0407	-0.5667
Third Year GPA	0.9700	5.0000	3.3935	0.8535	0.7285	-0.3226	-0.6562
Final CGPA	1.8000	4.9300	3.5605	0.6599	0.4355	-0.2190	-0.5777

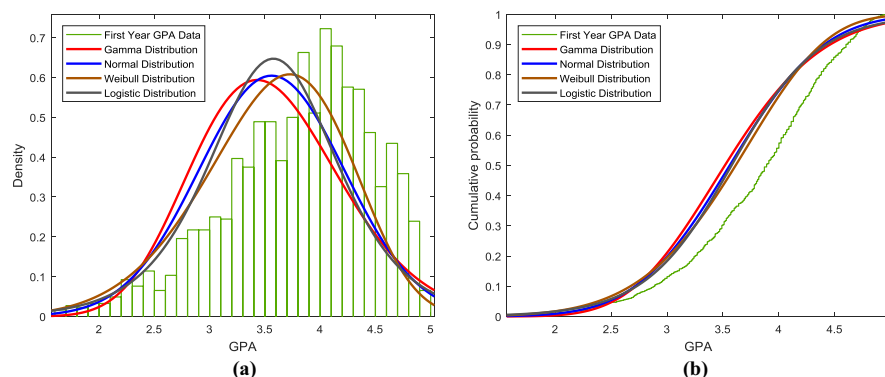


Fig. 1. First year GPA plots (a) Probability density function (b) Cumulative probability function.

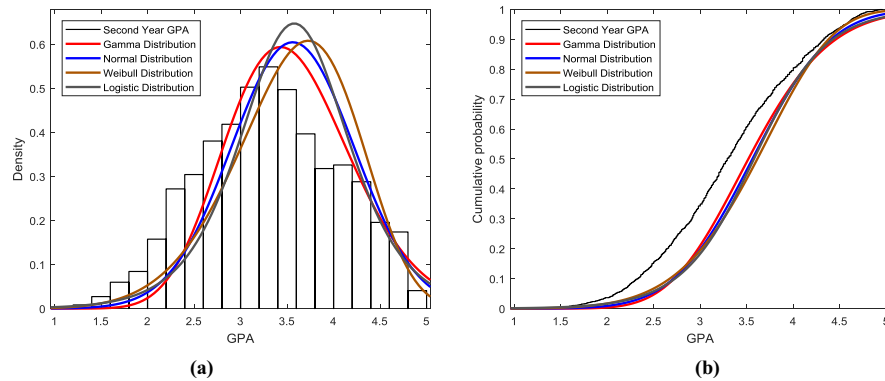


Fig. 2. Second year GPA plots (a) Probability density function (b) Cumulative probability function.

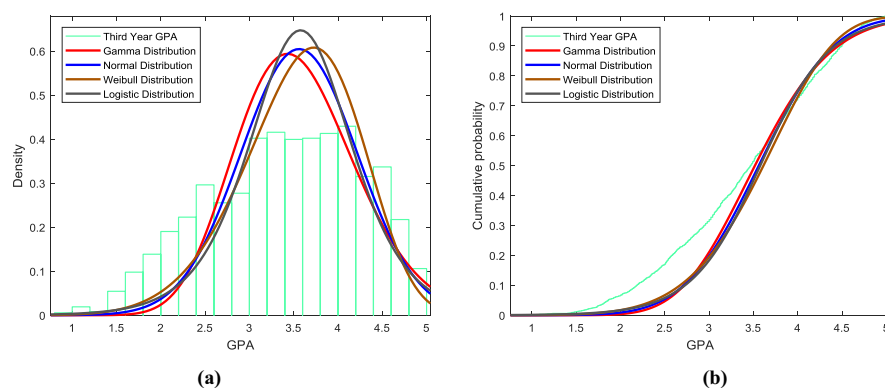


Fig. 3. Third year GPA plots (a) Probability density function (b) Cumulative probability function.

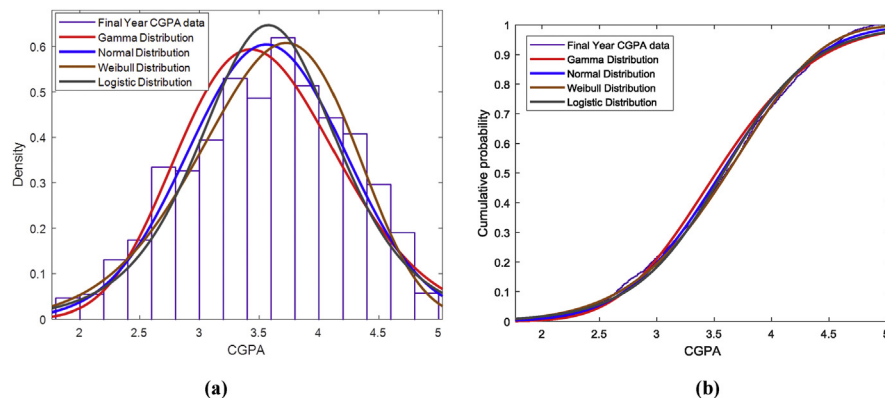


Fig. 4. Final year CGPA plots (a) Probability density function (b) Cumulative probability function.

4. Methodology

KNIME (Konstanz Information Miner) is a well-known, Java based, modular data mining application which facilitates interactive, visual, easy assembling, testing and running of data mining pipelines. The graphical workflow in KNIME is made possible by means of an Eclipse plug-in (Berthold et al., 2009) and the KNIME

application is available under Open Source GPLv3 license ([KNIME, 2018b](#)). The application has been applied for many data mining analyses and projects ([Wahbeh et al., 2011](#); [Yu et al., 2016](#)) by setting up workflows on the KNIME platform. For the classification analysis on the KNIME data mining platform, the following inputs were applied; the engineering departments or program of study, the year of entry, the students' GPA in the first year, the GPA in the second year, and the GPA in the third year. The seven engineering programs are ICE, CHE, CE, MECH, ELECT, CVE and PE while the year or session of admission stretched from the 2002/2003 session to 2009/2010 session.

A predictive model was developed on the KNIME analytics platform for identifying the hidden relationship among the features of the dataset which may enable a reasonable prediction of the class of grade of the final year CGPA of the engineering students, using their GPA for the first three academic years out of the total five years of study for engineering programs in Nigeria. In the KNIME workflow, the data was imported into the platform using the excel reader, and the statistical properties of the dataset was extracted with the statistics node. The data was pre-processed by splitting the data samples into two using stratified sampling; 70% for training and 30% for testing. The class of grade is colour coded to enable visual evaluation using scatter plots. The samples were normalized, and principal component analysis was carried out to reduce the dimensionality of the variables. The learners Meta node contains six data mining learner algorithms while the predictors Meta node contains their corresponding predictors. The performance of each algorithm was evaluated using scorers, and the result was exported through the excel writers. The arrows indicate the direction of data flow from one stage to another in the KNIME workflow. The forward feature selection Meta node enabled the classification of the influence of each input variable.

Six (6) main data mining algorithms: the Probabilistic Neural Network (PNN) based on the DDA (Dynamic Decay Adjustment), the Random Forest Predictor, the Decision Tree Predictor, the Naïve Bayes Predictor, the Tree Ensemble Predictor, and the Logistic Regression Predictor were applied in the KNIME model for predicting the class of grade of the final CGPA of the students at graduation. The data mining algorithms were applied as pre-configured by the developers with the class of grade selected as the target column, further information on the algorithms and configuration can be sourced from KNIME and NodePit online documentations ([KNIME, 2018a](#); [NodePit, 2018](#)). The data was partitioned into two in the ratio 70:30 using stratified sampling, and 70% of the students' performance data samples was applied in training the model while the remaining 30% was deployed for testing the performance of the predictive algorithms. The dataset was normalized, and principal component analysis was carried out to transform the possibly correlated features into uncorrelated principal components in order to improve the predictive accuracy of the model. For a comparative analysis, and to further validate the performance of

the data mining model; linear and pure quadratic regression models were also developed to analyse the students' performance dataset.

5. Results & discussion

This section presents the results of the predictive models using both KNIME and regression-based models.

5.1. Results using KNIME-based model

The predictive capability of each of the five input variables was evaluated using the forward feature selection Meta node on the KNIME server, and the result shows that in terms of accuracy, the third year GPA is the most influential variable followed by the second year GPA, the first year GPA, the program and the year of entry is the least influential variable in the classification experiment. The performance of the algorithms in terms of the prediction confusion is presented as confusion matrices in Table 2 for the PNN predictor, in Table 3 for the Random Forest predictor, in Table 4 for the Decision Tree predictor, in Table 5 for the Naïve Bayes predictor, in Table 6 for the Tree Ensemble predictor and in Table 7 for the Logistic Regression predictor. The confusion matrix shows the predictive performance of the algorithms for each class of the classification analysis for identifying the True Positive, False Positive, True Negative and False Negative predictions (Davis and Goadrich, 2006; Parker, 2001; Visa et al., 2011) for the 1st, 2|1, 2|2, and 3rd class grade classifications. A comparative performance analysis for the six algorithms is presented in Table 8. In terms of prediction accuracy, the Logistic Regression predictor had the highest accuracy of 89.15% followed by the

Table 2. Confusion matrix for the Probabilistic Neural Network (PNN) Predictor.

	2 2	3rd	2 1	1st
2 2	181	7	23	0
3rd	16	16	0	0
2 1	18	0	242	1
1st	0	0	13	36

Table 3. Confusion matrix for the Random Forest predictor.

	2 2	3rd	2 1	1st
2 2	179	7	25	0
3rd	14	18	0	0
2 1	13	0	244	4
1st	0	0	5	44

Table 4. Confusion matrix for the Decision Tree predictor.

	2 2	3rd	2 1	1st
2 2	176	9	21	0
3rd	12	20	0	0
2 1	16	0	237	3
1st	0	0	5	44

Table 5. Confusion matrix for the Naïve Bayes predictor.

	2 2	3rd	2 1	1st
2 2	177	16	18	0
3rd	7	25	0	0
2 1	23	0	236	2
1st	0	0	9	40

Table 6. Confusion matrix for the Tree Ensemble predictor.

	2 2	3rd	2 1	1st
2 2	178	7	26	0
3rd	12	20	0	0
2 1	13	0	244	4
1st	0	0	5	44

Table 7. Confusion matrix for the Logistic Regression predictor.

	2 2	3rd	2 1	1st
2 2	184	7	20	0
3rd	11	21	0	0
2 1	13	0	246	2
1st	0	0	7	42

Table 8. Model performance comparison.

	PNN	Random Forest	Decision Tree	Naive Bayes	Tree Ensemble	Logistic Regression
Correct Classified	475	485	477	478	486	493
Accuracy	85.895%	87.70%	87.85%	86.438%	87.884%	89.15%
Cohen's Kappa (k)	0.767	0.799	0.803	0.782	0.803	0.823
Wrong Classified	78	68	66	75	67	60
Error	14.105%	12.297%	12.155%	13.562%	12.116%	10.85%

Tree Ensemble with an accuracy of 87.884%. The Decision Tree predictor had the third best accuracy of 87.85%, and the Random Forest predictor had the fourth best accuracy with an accuracy of 87.70%. The Naive Bayes predictor had an accuracy of 86.438% while the PNN predictor had the least accuracy of 85.89%. The performance of the algorithms in terms of the number of True Positive and False Positive predictions is presented comparatively in Table 9.

5.2. Results using regression-based models

For a comparative analysis, and to further validate the performance of the data mining model which had a least accuracy of 85.89%; linear and pure quadratic regression models were developed to analyse the students' performance dataset using Matrix Laboratory (MATLAB). The regression analysis was not performed on KNIME in order to develop a regression model that incorporates all the data samples, rather than splitting the data into two for training and testing as required by KNIME data mining nodes. Also, Analysis of Variance (ANOVA), F statistics, and the response plot of the dependent variable to any of the independent variables can be easily developed on MATLAB. The following independent variables were considered for the regression analysis; the program of study coded numerically from 1 to 7, the year of entry coded numerically from 1 to 8 for the 2002/2003 session up to 2009/2010 academic sessions, the GPA of the first year, second year and third year as independent variables X_1 , X_2 , X_3 , X_4 and X_5 respectively. The relationship between the independent variables and the Final CGPA as the dependent variable is measured using the coefficient of determination (R^2) value of the regression model. Using the standardized coefficient (β) in Table 10, the third year GPA has the highest effect size or influence on the predicted CGPA followed by the second year GPA, the first year GPA, and the program of study while the year of entry has the least influence on the dependent variable.

Table 9. Prediction confusion of the six data mining predictors.

	PNN		Random Forest		Decision Tree		Naive Bayes		Tree Ensemble		Logistic Regression	
	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP
2 2	181	34	179	27	176	28	177	30	178	25	184	24
3rd	16	7	18	7	20	9	25	16	20	7	21	7
2 1	242	36	244	30	237	26	236	27	244	31	246	27
1st	36	1	44	4	44	3	40	2	44	4	42	2
Overall	475	78	485	68	477	66	478	75	486	67	493	60

TP — True Positive.

FP — False Positive.

Table 10. Linear regression model results.

	Estimate	Standard Error (SE)	Beta (β)	tStat	pValue
(Intercept)	0.4865	0.0197	-	24.6470	4.18E-116
Program (X_1)	-0.0057	0.0017	-0.0164	-3.2952	0.0010
Year of Entry (X_2)	-0.0016	0.0018	-0.0053	-0.8765	0.3809
First Year GPA (X_3)	0.1811	0.0090	0.1809	20.1100	1.90E-81
Second Year GPA (X_4)	0.2788	0.0089	0.3141	31.3090	9.00E-173
Third Year GPA (X_5)	0.4404	0.0064	0.5696	68.8750	0

Number of observations: 1841, Error degrees of freedom (EDoF): 1835.

Root mean square (RMS) Error: 0.140.

R^2 : 0.955, Adjusted R^2 : 0.955.

F-statistic vs. constant model: 7.78e+03, p-value = 0.

5.2.1. Linear regression model

The equation for the linear regression model obtained from the regression analysis is shown in Eq. (1).

$$y = 0.4865 - 0.0057x_1 - 0.0016x_2 + 0.1811x_3 + 0.2788x_4 + 0.4404x_5 \quad (1)$$

The summary of the linear regression model estimates is presented in Table 10. The model has an R^2 value of 0.955 which indicates that the final CGPA of engineering students in Covenant University can be reasonably determined by their performance (GPA) in the first three years of the five-year study. The F-statistic value for the components is shown in Table 11 while Table 12 displays the ANOVA for the linear regression model. The overall model prediction is presented as an adjusted variable plot in Fig. 5 (a) while Figs. 5 (b), 6 (a) and (b) present the adjusted response of the model to the first year GPA, the second year GPA and the third year GPA respectively.

5.2.2. Pure quadratic regression model

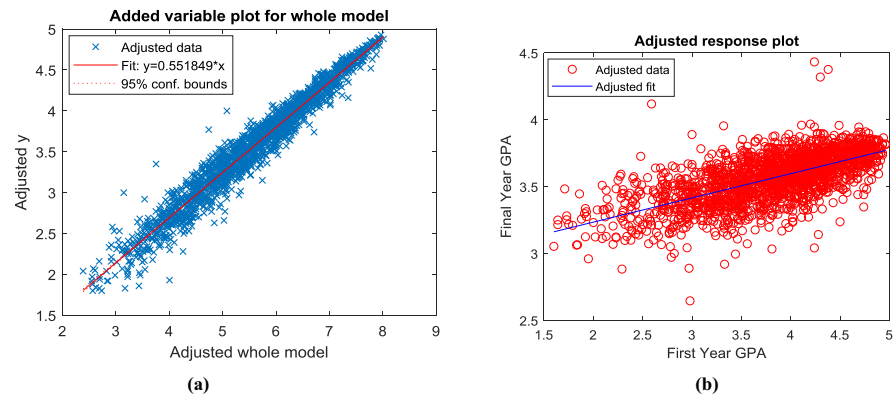
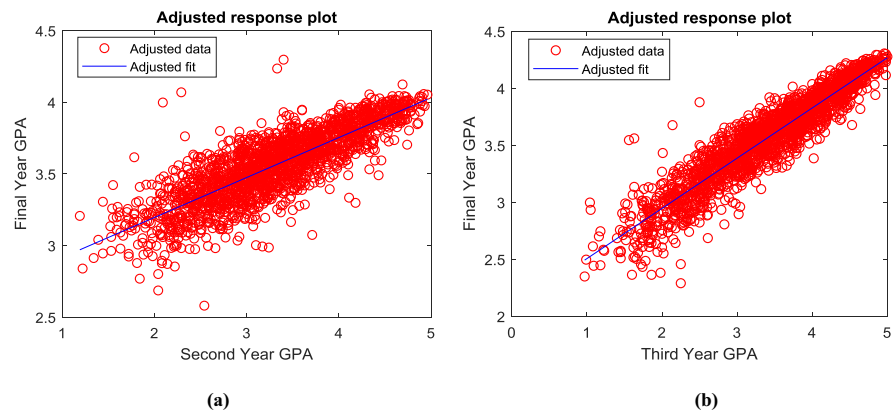
The equation for the pure quadratic regression model obtained from the regression analysis is shown in Eq. (2).

Table 11. F-statistic values for the components, except for the constant term.

	Sum of Square (Sum Sq.)	Degree of Freedom (DF)	Mean Square (Sq.)	F	pValue
Program	0.2135	1	0.2135	10.8580	0.0010
Year of Entry	0.0151	1	0.0151	0.7683	0.3809
First Year GPA	7.9512	1	7.9512	404.41	1.90E-81
Second Year GPA	19.2730	1	19.273	980.24	9.00E-173
Third Year GPA	93.2670	1	93.267	4743.8	0
Error	36.0780	1835	0.0197		

Table 12. ANOVA for the linear regression model.

	Sum Sq.	DF	Mean Sq.	F	pValue
Total	801.38	1840	0.4355		
Model	765.3	5	153.06	7784.9	0
Residual	36.078	1835	0.0197		

**Fig. 5.** (a) Added variable plot for the whole model (b) Adjusted response plot using the first year GPA.**Fig. 6.** (a) Adjusted response plot using the second year GPA (b) Adjusted response plot using the third year GPA.

$$y = 0.7380 - 0.0570x_1 - 0.0439x_2 + 0.1043x_3 + 0.2522x_4 + 0.4806x_5 + 0.0069x_1^2 + 0.0044x_2^2 + 0.0121x_3^2 + 0.0053x_4^2 - 0.008x_5^2 \quad (2)$$

The summary of the quadratic regression model estimates is presented in Table 13; the model has an R^2 value of 0.957. The F-statistic value for the components is shown in Table 14 while Table 15 displays the ANOVA for the quadratic regression model. The overall model prediction is presented as an adjusted variable plot in Fig. 7 (a) while Figs. 7 (b), 8 (a) and (b) present the adjusted response of the model to the first year GPA, the second year GPA and the third year GPA respectively.

Table 13. Quadratic regression model results.

	Estimate	SE	Beta	tStat	pValue
(Intercept)	0.7380	0.0788	-	9.3629	2.20E-20
Program (X ₁)	-0.0570	0.0069	-0.012941	-8.209	4.16E-16
Year of Entry (X ₂)	-0.0439	0.0079	0.012463	-5.5939	2.56E-08
First Year GPA (X ₃)	0.1043	0.0497	0.19571	2.0989	0.035958
Second Year GPA (X ₄)	0.2522	0.0432	0.32351	5.8433	6.04E-09
Third Year GPA (X ₅)	0.4806	0.0319	0.55104	15.066	1.93E-48
Program ² (X ₁) ²	0.0069	0.0009	0.037883	7.7051	2.13E-14
Year of Entry ² (X ₂) ²	0.0044	0.0008	0.031763	5.4661	5.23E-08
First Year GPA ² (X ₃) ²	0.0121	0.0070	0.0079419	1.7162	0.0863
Second Year GPA ² (X ₄) ²	0.0053	0.0067	0.0044	0.7869	0.4315
Third Year GPA ² (X ₅) ²	-0.0080	0.0050	-0.0088762	-1.6088	0.1078

Number of observations: 1841, EDoF: 1830.

RMS Error: 0.137.

R²: 0.957, Adjusted R²: 0.957.

F-statistic vs. constant model: 4.11e+03, p-value = 0.

Table 14. F-statistic values for the components, except for the constant term.

	Sum Sq.	DF	Mean Sq.	F	pValue
Program	0.1711	1	0.1711	9.1635	0.0025
Year of Entry	0.0267	1	0.0267	1.4279	0.2323
First Year GPA	8.4678	1	8.4678	453.42	4.54E-90
Second Year GPA	18.556	1	18.556	993.61	1.42E-174
Third Year GPA	81.8	1	81.8	4380.1	0
Program2	1.1087	1	1.1087	59.369	2.13E-14
Year of Entry2	0.5580	1	0.5580	29.879	5.23E-08
First Year GPA2	0.0550	1	0.0550	2.9455	0.0863
Second Year GPA2	0.0116	1	0.0116	0.6192	0.4315
Third Year GPA2	0.0483	1	0.0483	2.5882	0.1078
Error	34.176	1830	0.0187		

Table 15. ANOVA for the quadratic regression model.

	Sum Sq.	DF	Mean Sq.	F	pValue
Total	801.38	1840	0.4355		
Model	767.2	10	76.72	4108.1	0
Linear	765.3	5	153.06	8195.9	0
Nonlinear	1.9023	5	0.38046	20.372	7.76E-20
Residual	34.176	1830	0.0187		

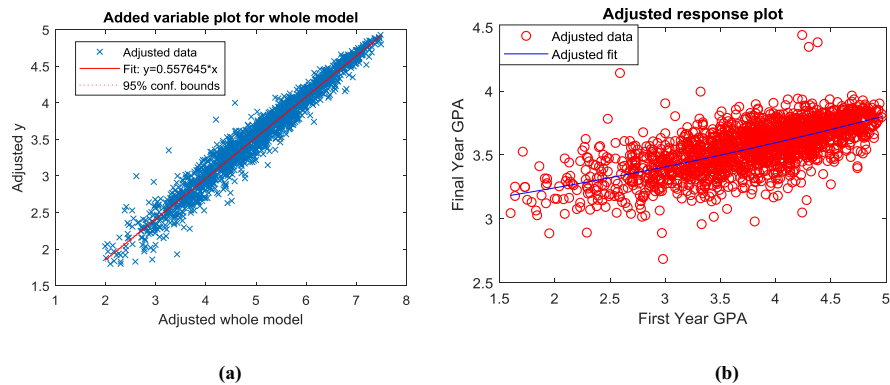


Fig. 7. (a) Added variable plot for the whole model (b) Adjusted response plot using the first year GPA.

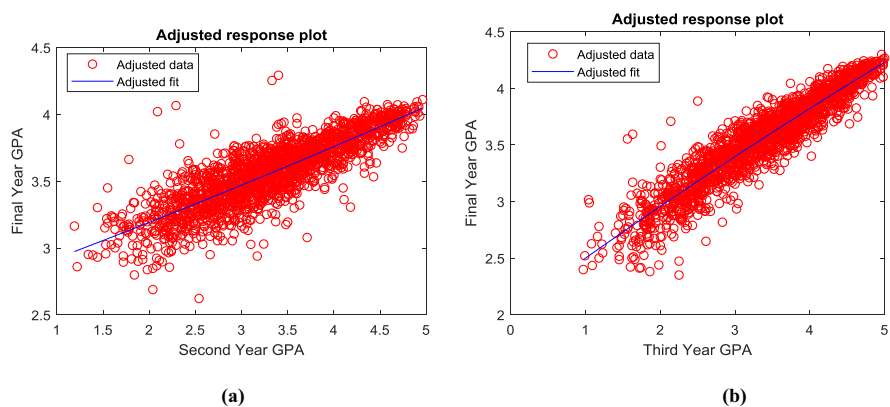


Fig. 8. (a) Adjusted response plot using the second year GPA (b) Adjusted response plot using the third year GPA.

The results of the KNIME based predictive model show that the class of grade of a student's final, fifth-year graduation result can be reasonably predicted using the student's GPA for the first three years of study. A maximum accuracy of 89.15% was achieved using the Logistic Regression algorithm while the PNN algorithm had the least accuracy of 85.895%. Likewise, using regression models R^2 value of 0.955 was achieved for the linear, and 0.957 for the pure quadratic regression models which indicate that the final CGPA of engineering students in Covenant University can be reasonably determined by their performance (GPA) in the first three years of the five-year study.

6. Conclusion

The management of higher educational system has transformed over the years from being reactive to being proactive in decision making and system performance analysis. Ensuring adequate quality in the educational system is vital to students' performance, and the overall value of the knowledge being imparted. There are many

benefits for detecting student issues and learning difficulties early, because this presents a unique opportunity to address the causal factors on time in order to prevent student failure and drop out tendencies. The performance of engineering student within the first three years of study is often said to be the most important in determining the final CGPA and class of grades of students in Nigeria due to the difficulty of improving the grades significantly at higher levels as the courses become more robust and intensive.

In this study, data mining approach was applied to evaluate the validity of this assumption by performing a predictive analysis to determine the final graduation CGPA and the class of grades of students in their final year using their GPA for the first three years of study. The program and the year of entry were applied as predictive inputs into a KNIME workflow using six independent data mining algorithms that were executed separately for a comparative performance analysis of the result of each of the six different algorithms. A maximum accuracy of 89.15% was achieved, and using regression models for performance validation, R^2 values of 0.955 and 0.957 were achieved using both linear and pure-quadratic based regression models. This indicates that indeed the graduating results of engineering students in Nigeria, in the fifth and final year of study can be reasonably predicted using their performance in the first three academic sessions. For future studies, as an alternative to running six data mining algorithms separately as implemented in this study, the KNIME workflow could be modified to incorporate the six data mining algorithms together in a model using a voting system such that the benefits of each algorithm can be combined.

Declarations

Author contribution statement

Aderibigbe Israel Adekitan: Conceived and designed the experiments. Performed the experiments; Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data; Wrote the paper.

Odunayo Salau: Performed the experiments; Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data; Wrote the paper.

Funding statement

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Competing interest statement

The authors declare no conflict of interest.

Additional information

There is no additional information for this paper.

Acknowledgements

The Authors appreciate the Department of Student Records and Academic Affairs, SmartCU and Covenant University Data Analytics Centre Research Clusters for the dataset generated, and Covenant University Centre for Research, Innovation and Discovery for providing an enabling research environment.

References

- Adekitan, A.I., 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.*
- Adekitan, Aderibigbe Israel, Adewale, Adeyinka, Olaitan, Alashiri, 2019. Determining the operational status of a Three Phase Induction Motor using a predictive data mining model. *Int. J. Power Electron. Drive Syst.* 10 (1).
- Adeyemi, K., 2001. Equality of access and catchment area factor in university admissions in Nigeria. *High. Educ.* 42 (3), 307–332.
- Agarwal, S., Pandey, G., Tiwari, M., 2012. Data mining in education: data classification and decision tree approach. *Int. J. e-Educ. e-Bus. e-Manag. e-Learn.* 2 (2), 140.
- Agboola, O.P., Elinwa, U.K., 2013. Accreditation of engineering and architectural education in Nigeria: the way forward. *Proc. Soc. Behav. Sci.* 83, 836–840.
- Ahmed, A.B.E.D., Elaraby, I.S., 2014. Data mining: a prediction for student's performance using classification method. *World J. Comp. Appl. Technol.* 2 (2), 43–47.
- Ajayi, I., Ekundayo, H.T., 2008. The deregulation of University education in Nigeria: implications for quality assurance. *Nebula* 5 (4), 212–224.
- Akerele, W., 2008. Quality assurance in Nigeria's university system: the imperatives for the 21 st century. In: *Towards Quality in African Higher Education*. Walacrown venture, Ibadan, pp. 19–141.
- Al-Radaideh, Q., Al-Shawakfa, E., Al-Najjar, M., 2006. Mining student data using decision trees. In: *Paper Presented at the the 2006 International Arab Conference on Information Technology. ACIT'2006*.

- Almarabeh, H., 2017. Analysis of students' performance by using different data mining classifiers. *Int. J. Mod. Educ. Comput. Sci.* 9 (8), 9.
- Amiruzzaman, M., 2019. Prediction of traffic-violation using data mining techniques. *Adv. Intell. Syst. Comp.* 880, 283–297.
- Angeli, C., Howard, S.K., Ma, J., Yang, J., Kirschner, P.A., 2017. Data mining in educational technology classroom research: can it make a contribution? [Article]. *Comput. Educ.* 113, 226–242.
- Asif, R., Merceron, A., Ali, S.A., Haider, N.G., 2017. Analyzing undergraduate students' performance using educational data mining. [Article]. *Comput. Educ.* 113, 177–194.
- Azevedo, A., 2018. Data mining and knowledge discovery in databases. In: *Encyclopedia of Information Science and Technology*, fourth ed. IGI Global, pp. 1907–1918.
- Baepler, P., Murdoch, C.J., 2010. Academic analytics and data mining in higher education. *Int. J. Scholarsh. Teach. Learn.* 4 (2), 17.
- Berthold, M.R., Cebon, N., Dill, F., Gabriel, T.R., Kötter, T., Meinl, T., 2009. KNIME-the Konstanz information miner: version 2.0 and beyond. *AcM SIGKDD Explor. Newsl.* 11 (1), 26–31.
- Bucos, M., Drăgulescu, B., 2018. Predicting student success using data generated in traditional educational environments. [Article]. *TEM J.* 7 (3), 617–625.
- Daradoumis, T., Marquès Puig, J.M., Arguedas, M., Calvet Liñan, L., 2019. Analyzing students' perceptions to improve the design of an automated assessment tool in online distributed programming. [Article]. *Comput. Educ.* 128, 159–170.
- Davis, J., Goadrich, M., 2006. The relationship between Precision-Recall and ROC curves. In: *Paper Presented at the Proceedings of the 23rd International Conference on Machine Learning*.
- Divya, N., Preetam, R., Deepthishree, A.M., Lingamaiah, V.B., 2019. Analysis of road accidents through data mining. *Lect. Notes Electr. Eng.* 500, 287–294.
- Gasevic, D., Rose, C., Siemens, G., Wolff, A., Zdrahal, Z., 2014. Learning analytics and machine learning. In: *Paper Presented at the Proceedings of the Fourth International Conference on Learning Analytics and Knowledge*.
- Gu, J., Gao, B., Zhu, S., 2018. Characterization of bi-domain drosomycin-type anti-fungal peptides in nematodes: an example of convergent evolution. [Article]. *Dev. Comp. Immunol.* 87, 90–97.

- Hussain, S., Dahan, N.A., Ba-Alwib, F.M., Ribata, N., 2018. Educational data mining and analysis of students' academic performance using WEKA. [Article]. *Indone. J. Electr. Eng. Comp. Sci.* 9 (2), 447–459.
- Ibrahim, Z.M., Bader-El-Den, M., Cocea, M., 2019. Mining unit feedback to explore students' learning experiences. *Adv. Intell. Syst. Comp.* 840, 339–350.
- Idachaba, F., 2018. Development of a rapid mentoring scheme for managing large classes in engineering departments. In: Paper Presented at the INTED2018 Conference, Valencia, Spain, 5th–7th March 2018.
- Jin, J., Liu, Y., Ji, P., Kwong, C.K., 2019. Review on recent advances in information mining from big consumer opinion data for product design. [Article]. *J. Comput. Inf. Sci. Eng.* 19 (1).
- Kabakchieva, Dorina, 2013. Predicting student performance by using data mining methods for classification. *Cybern. Inf. Technol.* 13 (1).
- Keserci, S., Livingston, E., Wan, L., Pico, A.R., Chacko, G., 2017. Research synergy and drug development: bright stars in neighboring constellations. *Heliyon* 3 (11), e00442.
- Kim, D., Yoon, M., Jo, I.H., Branch, R.M., 2018. Learning analytics to support self-regulated learning in asynchronous online courses: a case study at a women's university in South Korea. [Article]. *Comput. Educ.* 127, 233–251.
- KNIME, 2018a. Classification and Predictive Modelling. Retrieved 21-Dec-2018, from. <https://www.knime.com/nodeguide/analytics/classification-and-predictive-modelling>.
- KNIME, 2018b. KNIME Open Source Story, 22-Dec-2018, from. <https://www.knime.com/knime-open-source-story>.
- Kostopoulos, G., Kotsiantis, S., Pierrakeas, C., Koutsonikos, G., Gravvanis, G.A., 2018. Forecasting students' success in an open university. [Article]. *Int. J. Learn. Technol.* 13 (1), 26–43.
- Kovalchuk, V., Demchuk, O., Demchuk, D., Voitovich, O., 2019. Data mining for a model of irrigation control using weather web-services. *Adv. Intell. Syst. Comp.* 754, 133–143.
- Mahendra, M., Kishore, C., Prathima, C., 2019. Data Mining Efficiency and Scalability for Smarter Internet of Things. *SpringerBriefs Appl. Sci. Technol.* 119–125.
- Mahmud, J.O., Ismail, M.S.M., Taib, J.M., 2012. Engineering education and product design: Nigeria's challenge. *Proc. Soc. Behav. Sci.* 56, 679–684.

Miguel-Cruz, A., Aya-Parra, P.A., Rodríguez-Dueñas, W.R., Camelo-Ocampo, A.F., Plata-Guao, V.S., Correal, O.H.H., 2019. Using data mining techniques to determine whether to outsource medical equipment maintenance tasks in real contexts. In: Paper Presented at the IFMBE Proceedings.

NodePit, 2018. Mining from. <https://nodepit.com/category/analytics/mining>.

Noreña, A.P., González Muñoz, A., Mosquera-Rendón, J., Botero, K., Cristancho, M.A., 2018. Colombia, an unknown genetic diversity in the era of Big Data. [Article]. BMC Genomics 19.

Obadara, O.E., Alaka, A., 2013. Accreditation and quality assurance in Nigerian universities. J. Educ. Pract. 4 (8).

Odukoya, J.A., Bowale, E.I., Okunlola, S., 2018. Formulation and implementation of educational policies in Nigeria. Afr. Educ. Res. J. 6 (1).

Oguntunde, Pelumi, Okagbue, Hilary, Oguntunde, Omoleye Adedoyin, Opanuga, A., 2018. Analysis of the inter-relationship between students' first year results and their final graduating grades. Int. J. Adv. Appl. Sci. 5 (10), 1–6.

Oladipo, A., Adeosun, O., Oni, A., 2009. Quality Assurance and Sustainable University Education in Nigeria.

Olaleye, F.O., Oyewole, B.K., 2016. Quality assurance in Nigerian university education: the role of the National universities commission (NUC) as a regulatory body. Int. J. Acad. Res. Bus. Soc. Sci. 6 (12), 2222–6990.

Omoregie, N., 2008. Quality assurance in Nigerian university education and credentialing. Education 129 (2).

Osmanbegović, E., Suljic, M., 2012. Data Mining Approach for Predicting Student Performance, vol. 10.

Parker, J., 2001. Rank and response combination from confusion matrix data. Inf. Fusion 2 (2), 113–120.

Popoola, S.I., Atayero, A.A., Badejo, J.A., John, T.M., Odukoya, J.A., Omole, D.O., 2018. Learning analytics for smart campus: data on academic performances of engineering undergraduates in Nigerian private university. Data Brief 17, 76–94.

Porouhan, P., 2018. Process mining and interaction data analytics in a web-based multi- tabletop collaborative learning and teaching environment. [Article]. Int. J. Web Based Learn. Teach. Technol. 13 (4), 34–61.

Roy, S., Garg, A., 2018. Predicting academic performance of student using classification techniques. In: Paper Presented at the 2017 4th IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics. UPCON 2017.

- Saini, M.K., Aggarwal, A., 2018. Detection and diagnosis of induction motor bearing faults using multiwavelet transform and naive Bayes classifier. [Article]. *Int. Trans. Electr. Energy Syst.* 28 (8).
- Saini, B., Srivastava, S., 2019. Nanoinformatics: Predicting Toxicity Using Computational Modeling. *SpringerBriefs Appl. Sci. Technol.* 65–73.
- Tair, M.M.A., El-Halees, A.M., 2012. Mining educational data to improve students' performance: a case study. *Int. J. Inf. Commun. Technol. Res.* 2 (2), 140–146.
- Vardhani, P.R., Priyadarshini, Y.I., Narasimhulu, Y., 2019. CNN data mining algorithm for detecting credit card fraud. *SpringerBriefs Appl. Sci. Technol.* 85–93.
- Visa, S., Ramsay, B., Ralescu, A., van der Knaap, E., 2011. Confusion matrix-based feature selection. In: Paper Presented at the Midwest Artificial Intelligence and Cognitive Science Conference.
- Wahbeh, A.H., Al-Radaideh, Q.A., Al-Kabi, M.N., Al-Shawakfa, E.M., 2011. A comparison study between data mining tools over some classification methods. *Int. J. Adv. Comput. Sci. Appl.* 8 (2), 18–26.
- Yadav, S.K., Bharadwaj, B., Pal, S., 2012. Mining Education Data to Predict Student's Retention: A Comparative Study *arXiv preprint arXiv:1203.2987*.
- Yang, F., Li, F.W.B., 2018. Study on student performance estimation, student progress analysis, and student potential prediction based on data mining. [Article]. *Comput. Educ.* 123, 97–108.
- Yu, S., Zhao, D., Chen, W., Hou, H., 2016. Oil-immersed power transformer internal fault diagnosis research based on probabilistic neural network. *Proc. Comp. Sci.* 83, 1327–1331.
- Zuo, Y., Kajikawa, Y., Mori, J., 2016. Extraction of business relationships in supply networks using statistical learning theory. *Heliyon* 2 (6), e00123.