NIPES

Journal of Science and Technology Research

Journal homepage: www.nipesjournals.org.ng

A Multiple Linear Regression Model for Analyzing and Predicting Educational Development in Nigeria

Orogun, A.O.[1,*]; Ariyo, O.E[1]; Ajayi O.O.[1]; Onyekwelu, B.[2]

[1] Adekunle Ajasin University, Akungba-Akoko. Ondo State. Nigeria
[2] Elizade University, Ilara-Mokin, Ondo State. Nigeria
* Corresponding Author: adebola.orogun@aaua.edu.ng

Article Info

***Keywords*:**
multiple linear regression, educational development, educational organization, machine learning, performance, predicting

Received 27 October 2020
Revised 9 November 2020
Accepted 20 November 2020
Available online 01 March 2021

https://doi.org/10.37933/nipes/3.1.2021.11

https://nipesjournals.org.ng

Abstract

This paper explores the application of machine learning for predicting educational development in Nigeria, to determine its progress or regress in Ondo State. Numerous factors impact the performance of educational organizations, viz; infrastructure, learning condition, student ability, the immediacy of teachers, methods of educating and learning, research centers, grounds areas, and so forth served as the variable used for the prediction. The data for the study was obtained from Nigeria Education Digest Statistics Indicator 2005-2016, Ondo State Demographic Digest Statistics 2009. The study developed a predictive model using multiple linear regression, which considers the various factors that determines the progress of education in Nigeria. The result of the evaluation of the model shows that a somewhat rise in the educational development in the country, with regards to the feature attributes considered. The model showed that literacy rate has a higher correlation coefficient of 0.22.

1.Introduction

The main institutional component for developing human skills and knowledge is the formal education system. Most countries of the world understand that rapid quantitative extension of education opportunities is the way to national development [1][2][24]. Education at present plays a crucial role in encouraging individuals to live their lives more comfortably. According to [3], education is a procedure of acquiring skills, relevant knowledge and aptitude to survive in the technological world and become global citizens. In the present competitive world, education is an important vehicle for the socially and impeded countries to meet the global standard, it is one of the fundamental needs for human development and to escape from poverty [4][5]; it is necessary for national and worldwide improvement and for a prosperous society. It has the ability to make things change (transformation) and therefore can build up an underdeveloped country, push forward a backward or retarded state, transforms an uncivilized country, restores and empowers the disheartened or dejected person, etc. [6]. In the last decade, education in Nigeria has witnessed a critical development in terms of population expansion through an increase in enrolment and the establishment of additional institutions, [7][8]. However, it is disheartening to take note of that large number of lists that can ensure qualitative education is not taken into consideration in the country's quest to meet the quantitative target [9]. Furthermore, greater education institutions are striving hard to keep the value that is provided to the learners. Many variables or factors influence the performance of educational organizations such as inadequate infrastructure, low literacy rate, poor funding, poor monitoring/evaluation/quality control, quality teachers, laboratories, campus locations, links with industries, e.t.c [10][22]. In Nigeria, there have been complaints on rate of

education development, Nigerians are always on logger head concerning educational development, [11][12]. Parents prefer sending their children abroad to receive the so-called good education, which some Nigerians claim such education cannot be found in Nigeria education system, while others believe education in Nigeria is improving rapidly, [13][14]. This study will evaluate past historical data of the education system in Ondo State and predict the trend of an education system based on the data collected. There is an increasing trend among Nigerians towards the issue of educational development in the country, [9], due to the challenges that the education system is facing; that includes the availability of infrastructure, learning environments, quality of teachers, promptness of teachers, distance between student and school, etc. [10][23]. Having noted the challenges and issues which trends as a major problem, there is a need to analyze education data to predict the future of Nigeria education development, whether there is growth or not.

1.1 Related Literature

Analyzing and predicting educational progress has been an active area of research in recent times. [15] predicted student academic performance in engineering courses using a machine learning technique to predict student academic performance in engineering courses. In their research, the input features included course grades from all semesters and the output variable was exam scores. The researchers observed that SVMs are suitable for predicting an individual student's performance and that multi-linear regression is suitable for forecasting the performance of all students in a course. However, the author was unable to justify why SVM was optimal as other machine learning techniques were not tested with the dataset. [16], forecasted students' performances using machine learning techniques like C4.5, sequential minimal optimization (SMO), Naïve Bayes, 1-NN (1-Nearest Neighborhood), and MLP (multi-layer perceptron) with input features (e.g., gender, income, board marks, and attendance). Correlation-based feature selection (CBFS) techniques was applied to improve the model performances and determined that SMO achieves a higher effective average testing accuracy (66%) in comparison with other methods. [17] predicted student performance using data mining techniques like Regression and decision trees to know the academic failure of students. Their research clustered students using their academic performance as metric. They identified that apart from the challenges and costs involved in educational data mining, EDM implementation requires the privacy and ethics of all the stakeholders involved in the EDM process. However, this study was unable to assert a good predictive accuracy. Furthermore, [18], examined student difficulties in a course on mathematics, system analysis, and design using data mining techniques. The paper examined student difficulties in a course on mathematics, system analysis, and design using data mining techniques. Test grades was used as input features and determined that AdaBoost was the best classifier for predicting the difficulties that students would experience in subjects. Also, [19] carried out a predictive analysis on Students Academic Performance utilizing Naive Bayes Algorithm. The limitation of this study was that the research was conducted on few factors affecting academic performance. In [20], an analysis was carried to asset the relationship between grades and students' learning processes using DEEDS data. Process Mining (PM) and Complexity Matrix (CM) methods were used to analyze the relationship between grades and students' learning processes using DEEDS data. It was then concluded that the average student grades are positively correlated with the CM and that difficulty is negatively correlated with the CM. In addition, they determined that process discovery using PM and CM models provides valuable information regarding student learning processes. [21] at-risk students in advance of the next course were identified**.** In this study, logistic regression, support vector machines (SVMs), decision trees (DTs), ANNs and a Naïve Bayes classifier (NBC) were used to identify at-risk students in advance of the next course. This study used input features, such as grades, attendance, quizzes, weekly homework, team participation, project milestones, mathematical modeling activity tasks, and exams from an offline course. Analysis of the results found that the NBC algorithm provided satisfactory accuracy (85%).

2.0 Methodology

In this research, multiple linear regression was used. It was concluded that the outcomes are continuous variables. This technique is useful when the variable depends on more than one independent variable. Another consideration is that multiple linear regression models are computationally cheap and extremely fast and achieve better accuracy in a statistical problem. For achieving the above-mentioned research methodology, the following concepts were used: data collection, data preparation which comprises data transformation, data cleaning, feature selection, data visualization, split data into train and test set. Figure 1 shows the systemic diagram of the model.

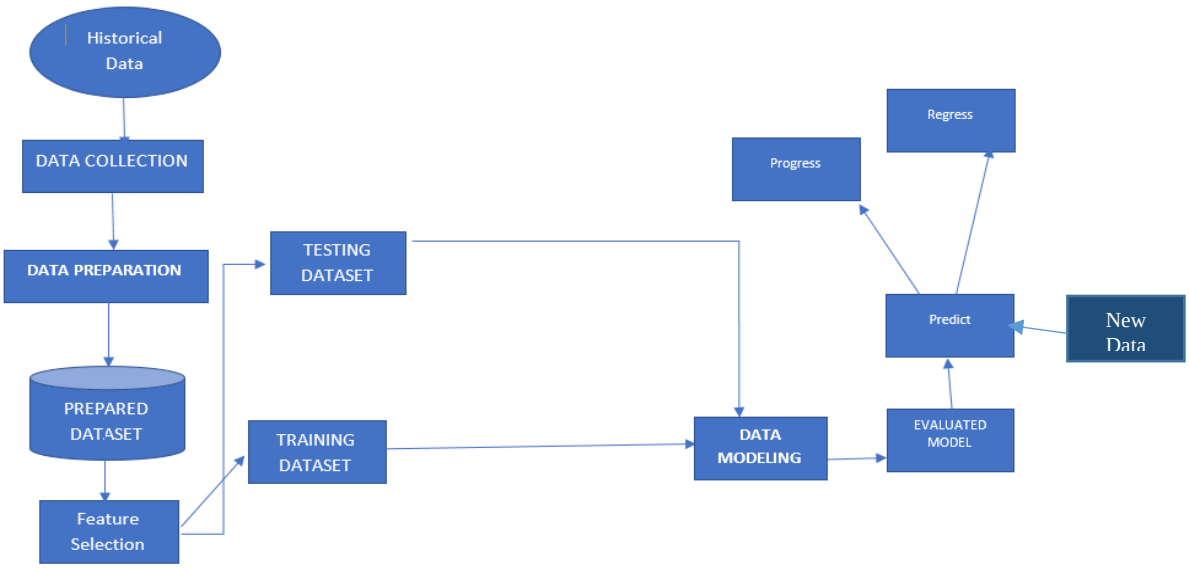

Figure 1. Flowchart of the study methodology

2.1 Data Collection

Database, data warehouse, site visitation, interview, internet, questionnaire, observation, and other documents were the actual sources of data, with data collection ranging between the years 2006 through 2016. A large volume of historical data is needed for a machine learning model to be successful. This project was achieved by gathering materials from different sources. There is no ready-to-use data set, data was gathered or extracted from the historical dataset (Digest Education Statistics 2006-2010, Nigeria Digest of Education Statistics 2011-2016, Nigeria Education Indicator 2016 and Digest of Demographic and Vital Statistics of Ondo State 2009).

2.1.1 Data Set Description

The dataset is composed of historical records of education statistics in Ondo state between the years 2006 – 2016. The dataset was categorized into two perspectives (Primary institution dataset and Secondary Institution dataset. The data set was organized in columns and rows. The columns represent a variable and the row represents a single record of different attributes of education statistics. In general, there were 9 variables (columns) and 102 rows in the dataset. The data were extracted and stored in MS excel format and saved as CSV (comma separated values). Tables 1, 2, and 3 presents variables used and their different representations, with samples of datasets used.

Table 1. Variables, definitions and representation

Variables	Definitions	Representation
Enrollment	The number of people enrolled at a school	IDV1
Population (Age group)	Distribution of the population according to age	IDV2
Gross Enrollment Rate	Percentage of total enrollment at a specific level of education, regardless of age. Enrolment/Population (Age group) * 100, in a given school year	IDV3
Literacy Rate	It is the percentage of people in a certain sample of the population	IDV4
Qualified Teacher Rate	Percentage of teachers by level of education taught who have at least the minimum academic qualifications required to teach. (Qualified Teacher/All Teachers) * 100	IDV5
School Funding Rate	Allocation of the percentage of government revenue, from sales and income tax, to use towards education	IDV6
Learning Environment Rate	It is the percentage of educational approaches, cultures, and physical settings for all types of learning activities	IDV7
Pupil/Student-Teacher Ratio	The average number of pupils/student per teacher at a specific level of education in a given school year	IDV8
Health Service Rate	It is the percentage of improvement or maintenance of health in schools.	IDV9

Table 2. Sample of the dataset (Primary Institution)

Year	Category	IDV1	IDV2	IDV3	IDV4	IDV5	IDV6	IDV7	IDV8	IDV9
2006	Primary	744120	877028	84.84	76.6	75.38	10.61	40.14	51	10.61
2007	Primary	827632	930118	88.98	83.12	73.38	9.51	34.51	66	8.80
2008	Primary	865387	909896	95.11	89.32	79.34	10.04	43.05	66	15.87

Table 3. Sample of the dataset (Secondary Institution)

Year	Category	IDV1	IDV2	IDV3	IDV4	IDV5	IDV6	IDV7	IDV8	IDV9
2006	Secondary	188320	478435	39.36	69.23	75.38	10.61	39.55	19	12.87
2007	Secondary	238676	490489	48.66	68.66	73.38	9.51	37.56	37	9.56
2008	Secondary	222251	516421	53.30	73.39	79.34	10.04	31.45	21	13.03

2.2 Data Preprocessing

This involved the combination of data from multiple sources into a unified data store as the data collected were from various data stores. The data collected were from the historical dataset (Digest Education Statistics 2006-2010, Nigeria Digest of Education Statistics 2011-2016, Nigeria Education Indicator 2016 and Digest of Demographic and Vital Statistics of Ondo State 2009) and so they were merged and collated together in Microsoft excel 2016. In this study, the Pearson correlation coefficient a type of filter feature selection method was used in selecting the best feature. The features are ranked by the features having a high correlation and closely related and two (2) variable "enrollment" and "population" were removed from the dataset due to their low correlation. For this research, multiple linear regression will be used. Let y denotes the dependent variable that is linearly related to k independent variable $X_1, X_2, \ldots, X_k$ through the parameters $\beta_1, \beta_2, \ldots, \beta_k$ and we write:

$$\boldsymbol{y} = \boldsymbol{X}_1\beta_1 + X_2\beta_2 + \ldots + X_k\beta_k + \varepsilon. \qquad (1)$$

Equation (1) is called the multiple linear regression model. The parameters $\beta_1, \beta_2, \ldots \beta_k$ are the regression coefficients associated with $X_1, X_2, \ldots, X_k$ respectively and ε is the random error component reflecting the difference between the observed and fitted linear relationship. There can be various reasons for such difference, e.g. joint effect of those variables not included in the model, random factors that cannot be accounted for in the model etc.

The j^{tk} regression coefficient β_j represents the expected change in y per unit change in j^{tk} independent variable X_j. Assuming $E(\varepsilon) = 0$,

$$B_j = \frac{\delta E(y)}{\delta X j} \quad (2)$$

2.2.1 M**odel set up:**
Let an experiment be conducted n times and the data is obtained in Table 4

Table 4: Explanatory variables for n observations

Observation number	Response y	Explanatory variables X_1 X_2 … X_k
1	y1	X_{11} $X1_2$ … X_{1k}
2	y2	X_{21} X_{22} … X_{2k}
.	.	. . .
.	.	. . .
.	.	. . .
N	yn	X_{n1} X_{n2} … X_{nk}
		.

Assuming that the model is

$y = \beta_0 + \beta_1 X_1 + \beta_2 X_{2 + \ldots +} \beta_k X_{k + \varepsilon}$ (3)

The n-tuples of observations are also assumed to follow the same model.
Thus they satisfy

$y = \beta_0 + \beta_1 X_{11} + \beta_2 X_{12 + \ldots +} \beta_k X_{1k + \varepsilon}$ (4)

$y = \beta_0 + \beta_1 X_{21} + \beta_2 X_{22 + \ldots +} \beta_k X_{2k + \varepsilon}$ (5)

.

..$y = \beta_0 + \beta_1 X_{n1} + \beta_2 X_{n2 + \ldots +} \beta_k X_{nk + \varepsilon}$ (6)

Assumptions in the multiple linear regression model
Some assumptions are needed in the model $y = X\beta + \varepsilon$ for the drawing statistical inferences. The following assumption are made:

(i) $E(\varepsilon) = 0$
(ii) $E(\varepsilon \varepsilon') = \sigma^2 I_n$
(iii) $Rank(X) = k$
(iv) X is a non-stochastic matrix
(v) $E \sim N(0, \sigma^2 I_n)$

These assumptions are used to study the statistical properties of estimator of regression coefficient. The following assumption is required to study particularly the large sample properties of the estimators

(vi) $\lim_{n\to\infty} \left(\frac{X'X}{n}\right)$ exists and is a non-stochastic and nonsingular matrix (with finite elements).

The explanatory variables can also be stochastic in some cases. We assume that X is non-stochastic unless stated separately.

$$\lim_{n\to\infty} \left(\frac{X'X}{n}\right)$$

2.4 Model Evaluation

In this research, several metrics are used for assessing or evaluating the performance of the machine learning model which includes, regression accuracy, mean absolute error, mean squared error, and R^2 score.

- Mean absolute error is the average of the absolute differences between predictions and actual values. It gives an idea of how wrong the predictions were.

- Accuracy is the number of correct predictions as a ratio of all predictions made.

$$\frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

$$\text{Accuracy} = \frac{\text{number c}}{\text{total number of prediction}}$$

- Mean squared error is the average squared distance of a data point from the fitted line.

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

- R^2 (R Squared) metric provides an indication of the goodness of fit of a set of predictions to the actual values. This is a value between 0 and 1 for no-fit and perfect fit respectively.

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

Where SS_{res} is the residual sum of squares and SS_{tot} is the total sum of squares
There are two major problems of prediction error for a model: bias and variance. A model with a high bias will produce similar errors for an input regardless of the training set it was trained with; the model biases its own assumptions about the real relationship over the relationship demonstrated in the training data. A model with high variance, conversely, will produce different errors for an input depending on the training set that it was trained with. That is, a model with high variance over-fits the training data, while a model with high bias under fits the training data. Ideally, a model will have both low bias and variance, but efforts to decrease one will frequently increase the other. This is known as the bias-variance trade-off.

3. Implementation

The model was implemented on Window 10 pro OS, Intel Core i5 processor, 2.4Ghz, 4GB RAM using Python programming language on Jupyter Notebook IDE (Integrated Development Environment). The model was tested at least 5 times to ensure high prediction accuracy. Analysis of the data is done by visualizing the data through a series of plots to gain knowledge of the data. According to the data collected, primary education development has 60% highest rate in 2014 and 49% lowest rate in 2005, which show a significant improvement in primary education growth, secondary school development highest rate was at 51% in 2014 and recorded the lowest rate of 35% in 2011. Figure 2a and b shows the rate of development. Figures 3 and 4 shows the education rate for both primary and secondary education, with their corresponding independent variables. The regression correlation for all independent variables is presented in Figure 5.

Figure 2a, b. Primary and secondary Education Development Rate

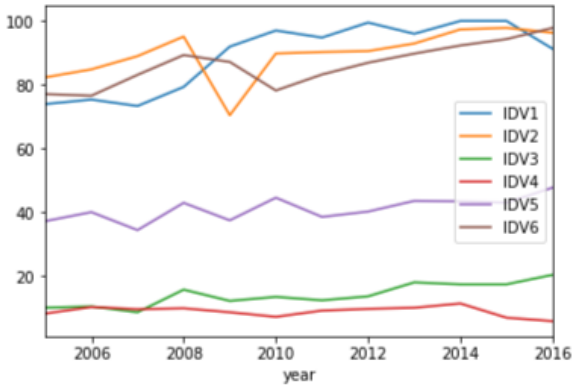

Figure 3. Primary Education Rate for each independent variable

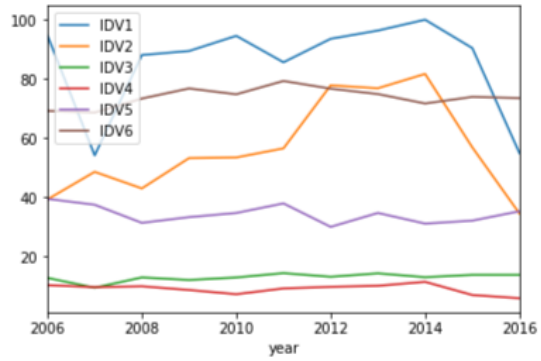

Figure 4. Secondary Education Rate for each independent variable

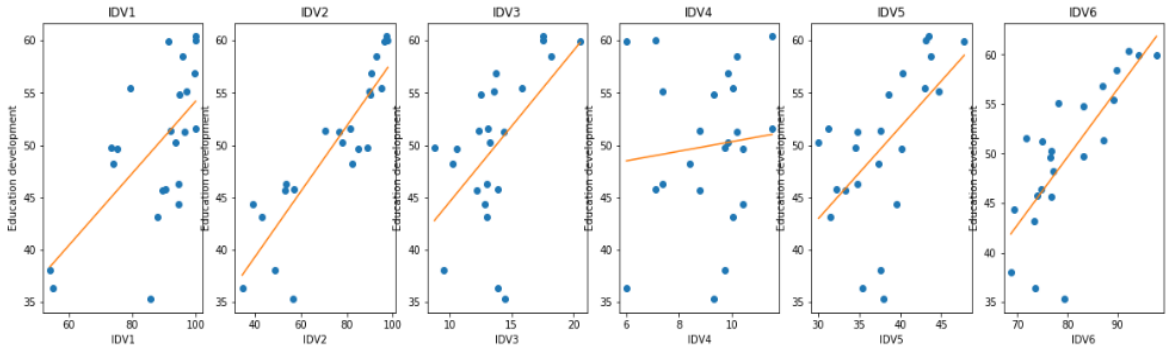

Figure 5. Regression correlation line for each independent variable

.

3.1 Results and Discussion

This experiment was performed to predict educational development in Nigeria using multiple linear regression. The prediction results are reported in this section. To measure the quality of the multiple linear regression equation, the predicted development is compared with actual development. Table 4 shows the details of the predictor's correlation for prediction. The correlation results show that IDV2 (Literacy Rate) has a higher correlation coefficient (0.220567). The summary statistics are shown in Table 5.

Table 4. Correlation of predictor with the education development rate

S/N	Predictor	Correlation Coefficient
1	IDV1 (Gross Enrollment rate)	0.167220
2	IDV2 (Literacy Rate)	0.220567
3	IDV3 (Qualified Teacher rate)	0.113300
4	IDV4 (School Funding Rate)	-0.090887
5	IDV5 (Learning Environment rate)	0.124007
6	IDV6 (Health Service Rate)	0.096333

Table 5: Regression analysis and descriptive statistics

	category	IDV1	IDV2	IDV3	IDV4	IDV5	IDV6	mean
count	23.000000	23.000000	23.000000	23.000000	23.000000	23.000000	23.000000	23.000000
mean	0.478261	87.584957	73.880174	13.705000	9.097391	37.945217	80.392174	49.915036
std	0.510754	13.243169	20.787032	2.796334	1.588845	4.840276	8.269675	7.341164
min	0.000000	54.200000	34.580000	8.795000	6.010000	30.090000	68.660000	35.285000
25%	0.000000	82.482000	55.052000	12.445000	7.885000	34.645000	74.390000	45.697500
50%	0.000000	91.960000	81.710000	13.250000	9.750000	37.570000	77.100000	50.223333
75%	1.000000	96.160000	90.420000	14.445000	10.125000	41.685000	87.050000	55.285000
max	1.000000	100.000000	97.840000	20.510000	11.540000	47.810000	97.820000	60.381667

Here, multiple regression approach on the data set. From this approach, we can predict educational development in any one of the future's years by using the selected factors. When the multiple linear regression equation is used with test data for testing the accuracy of the multiple linear regression equation, we obtain the educational development rate which is close to the actual educational development rate data, the graphical representation between the actual and predicted value of educational development is represented in the graph given below.

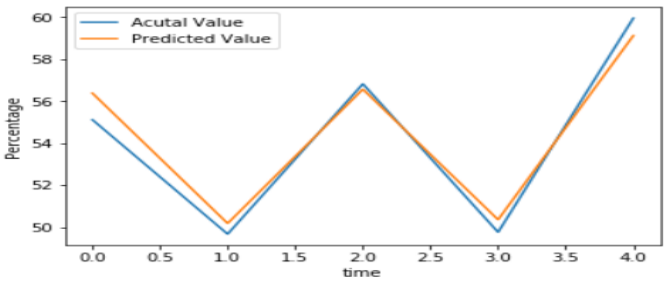

Figure 6. Comparison between Actual Value and Predicted Value

3.2 Model Evaluation

Figure 7 shows the mean absolute error (average of the absolute differences between predictions and actual values) with the value of 0.6914458946945743. The model accuracy (number of correct predictions as a ratio of all predictions made) was 0.712343 i.e. 71.24%.

```
In [201]: ##calculate MAE using scikit-learn
          from sklearn import metrics
          metrics.mean_absolute_error(y_test,y_pred)

Out[201]: 0.6914458946945743
```

Figure 7. Model Mean absolute error.

4.5 Model Summary

Tables 6 and Table 7 show the model summary generated using the fitted model summary function.

Table 6 Model Summary

Dep. Variable	Mean	R-squared	0.895
Model:	OLS	Adj. R-squared	0.855
Method	Least Squares	F-statistic	22.65
Date:	Mon, 06 Jan 2019	Prob(F-statistic)	5.60e-07
Time	05:27:12	Log-Likelihood	-52.093
No. Observation	23	AIC	118.2
Df Residuals	16	BIC	126.1
Df Model	6		
Covariance Type	nonroboust		

Table 7 Model Summary

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	5.7970	10.475	0.553	0.588	-16.409	28.003
IDV1	0.1672	0.060	2.800	0.013	0.041	0.294
IDV2	0.2206	0.052	4.237	0.001	0.110	0.331
IDV3	0.1133	0.366	0.309	0.761	-0.664	0.890
IDV4	-0.0909	0.429	-0.212	0.835	-1.000	0.818
IDV5	0.1240	0.189	0.658	0.520	-0.276	0.524
IDV6	0.0963	0.156	0.619	0.544	-0.233	0.426

4.0 Conclusion

This study investigated the ability to predict educational development in Nigeria using machine learning technique. The dataset was extracted from Nigeria Education Digest Statistics Indicator 2005-2016, Ondo State Demographic Digest Statistics 2009. The model was trained using training data and tested the models on testing data and modal evaluation was used to evaluate the performance of the models, the model had a high predicting accuracy of 71 percent.

References

[1] Gbenu, J. P. (2012). State of Nigerian Secondary Education & The Need for Quality Sustenance: Green Journal of Educational Research, Vol. 2(1), pp.007-012.
[2] Lawrence N. N.,(2016). Nigeria Education Development and Quality Sustenance. Universitatis Islandis Sigillum, Haskoli Island, 2016
[3] Oyewole B.K., Osalusi F.M., (2016). Towards Actualizing Sustainable Education Standards in Nigeria. International Journal of Learning, Teaching and educational Research. Volume 15(2), pp 44-54.
[4] Sivakumar, M., Sarvalingam, A. (2016). Human deprivation index: A measure of multidimensional poverty.

[5] Romina I.A., (2015). Improving Higher Education in Nigeria: The Roles of Stakeholders. International Journal of Higher Education. Volume 4(1), pp 61-70.
[6] Uriah, O.A. Ololube, N.P., Olulube, D.O.O. (2016). Education as a viable tool for placing Nigeria in its rightful place among the comity of nations: Obstacle and solutions. Journal of Global Research in education and social sciences vol. 8, pp. 92-101.
[7] Babalola, J. B. (2007). Revitalizing Quality Higher Education in Nigeria: options and strategies. In J. B. Babalola, publication 2007.
[8] Auwal M, Umar M. K., (2017). Appraising the Quality of Education in Nigeria: A Philosophical View. Sokoto Educational Review. Volume 17(2), pp 9
[9] Famade, O., Omiyade, G., Adebola, Y. (2015): Towards improved funding of tertiary institutions in Nigeria. Asian Journal of Humanities and Social Sciences, vol. 3, pp. 83-90.
[10] Uriah, O.A., Joseph A. I (2017). "Education and Social Change" in G. Ekwueme Elechi and Ogbondah, L. (eds.). Sociological Foundations of Education. Port Harcourt: High Calling Outreach Publishers. Vol. 3 pp. 11-24.
[11] Emediegwu, L. E., Clement I., (2016). "The connection between Education and Sustainable Economic Growth in Nigeria," Zambia social science journal: vol. 6: No. 1, Article 6.
[12] Grace M. B., (2016). Perspective of Science Education in Nigeria. International Journal of Sustainable Development. Volume 11 (2), pp 01-04.
[13] Alexander T., (2017). Islamic Modernism and Colonial Education in Northern Nigeria: Naibi Sulaiman Wali (1927 – 2013). Journal Religion and Education. Volume 44. http://doi.org/10.1080/15507394.2016.1267540
[14] Onifade, A., (2006). Science Education for All Nigerians: challenges of the 21st Century. Eko Journal of Educational Research 2(2),pp 61.
[15] Huang S., Fang N (2013) Predicting student academic performance in an engineering dynamics course: a comparison of four types of predictive mathematical models. Computer Education 61: pp 133–145. https://doi.org/10.1016/j.compedu.2012.08.015.
[16] Acharya A., Sinha D (2014). Early prediction of students' performance using machine learning techniques. International Journal Computer Application 107(1): pp37–43. https://doi.org/10.5120/18717-9939
[17] Amirah M.S, Wahidah H, Nur'Aini A, (2015). A review on predicting student's performance using data mining techniques. Procedia computer science, Elsevier 2015, pp 414-422
[18] Kaur K., (2015) Analyzing the effect of difficulty level of a course on students' performance prediction using data mining. In: 1st international conference on next generation computing technologies 2015, pp. 756–761.
[19] Sujith J., Balamurugan E., Vibin C., (2015). "Students' Academic Performance using Naive Bayes Algorithm. International Journal of Computer Applications. Vol. 6 pp. 22- 35.
[20] Vahdat M., Oneto L., Anguita D., Funk M., Rauterberg M., (2015) A Learning analytics approach to correlate the academic achievements of students with interaction data from an educational simulator. In 10th International European conference on technology enhanced learning (EC-TEL) 2015. pp. 352–366. https://doi.org/10.1007/978-3-319-24258-326
[21] Marbouti F., Diefes-Dux H.A., Madhavan K., (2016) Models for early prediction of at-risk students in a course using standards-based grading. Comput .Educ. 103:1–15.
[22] Bengio K., Adepoju S. A., Kolo A. J., (2016). A Decision Tree Approach for Predicting Students Academic Performance," Int. J. Educ. Manag. Eng., vol. 5, no. 5, pp. 12–19.
[23] Blessing E.N., Thom O., Dorothy C. I., (2016). Quality Education for National Development: The Nigerian experience. African Educational Research Journal. Volume 4 (3), pp 101-108. https://doi.org/10.1109/NGCT.2015.7375222
[24] Kenneth B., Brian C., (2017). A Sociology of Education for Africa. Routeledge Library Editions: Sociology of Education. Volume 8.
[25] Lantz, B., Machine learning with R. Packt Publishing Ltd, 2013.