INTI JOURNAL | Vol.2023:34
eISSN:2600-7320

A Framework for Formulation of Student Dataset Using Existing and Novel Features for Analysis

Esther Samuel Alu[1*], Rashida Funke Olanrewaju[1], Afolyan A. Obiniyi[2,] Muhammad Dahiru Liman[3]

[1]Department of Computer Science, Nasarawa State University Keffi, Nigeria
[2]Department of Computer Science, Ahmadu Bello University Zaria, Nigeria
[3]Department of Computer Science, Federal University of Lafia, Nigeria

[*]Email: estheralu@nsuk.edu.ng,

Abstract

One major problem identified with most schools in Nigeria is that they lack structured educational datasets that is composed of several attributes related to each student, such as term-based grades, courses taken, student-specific details, and absences which could be easily analysed. This paper formulates a dataset with some novel features for analysing and predicting student performance. Apart from the current features like age, grade, number of failures etc. some novel features which consists of environmental factors were proposed. Students' records were collected from schools and surveys on schools' infrastructure were collected using a questionnaire. The data were analysed using NumPy and Pandas in python. Random forest was used as classifier for making prediction and detecting important features. The following features were found to influence the model decision in making decision; Average, Number of failures, students score in all the subjects, school type, portable drinking water, availability of electricity, textbook to student ratio, and availability of laboratory reagents. Four of the proposed features were among the most important features. In addition, the model was excellent in making prediction. Results of the analysis shows that there are more male than females in the dataset, this means that government, non-governmental organization and the society needs to promote and encourage girl child education.

Keywords

Student, Dataset, Feature Importance, Random Forest.

Introduction

Education is the bedrock of any society, it provides a foundation for development; therefore, this foundation must be properly built so that it can provide the desired development. The decline in performance of students in public schools raises concern considering Government spending over the last three decades in the sector (Omotor, 2004 & Onuma, 2016). This has left most government schools deserted as most parents have taken their children out of public schools to private schools

Submission: 2 June 2023; **Acceptance:** 7 August 2023

(Dixon *et al.,* 2017 & Ukpor *et al.,* 2012). Stakeholders have found themselves in the undesirable position of not being able to identify the cause of the decline in students' performance. According to West Africa Examination Council (WAEC) statistics, fewer students are registering for public schools. Private schools have a larger enrollment of students than public schools. Private schools saw a 56% increase in enrollment while public schools saw a 44% fall (Uduu, 2022). According to WAEC, private school performance increased from 54% in 2016 to 71% in 2019 among students with more than five credits. The increase in student enrollment in private schools may be responsible for this improvement.

This work focuses on the senior secondary school level of education in Nigeria, and how poor student record keeping and maintenance has hurt the quality of secondary education in Nigeria. It is this gap that this research aim to fill by formulating a dataset with relevant student features that will be used by policy makers to make decisions. It will also help school administrators to identify students who had the highest probability of failing at the end of the year. Features that influence the model decision can be the focal point of policy makers, and school administrators in tackling student failure.

Materials and Methods

Materials

The proposed dataset consists of existing and novel features.

Existing features from Literatures

The existing features used in previous related works are presented in table 1.

Table 1. Features used in previous work for predicting student performance

Features	Paper
GPA and Grades	(Huang *et al.*, 2011)
Grades	(Livieris *et al.*, 2012; Li *et al*, 2013; Arsad *et al.*, 2013; Meier *et al*, 2015; Arsad *et al.*, 2014; Buniyamin *et al*, 2016).
Grades, Backgrounds	Xu et al, 2017
Class Performance, Attendance, Assignment, Lab Work, Sessional Performance	Guleria et al, 2014
Aptitude, Personality, Motivation Learning strategies	Gray et al, 2014
student demographics, general performance, students' modules	Alharbi et al, 2016
Internal grades, sessional grades and admission score	Hamsa et al, 2016
Personal and demographics information, student satisfaction and integration	Sarker et al, 2014
Personal data, pre-university data, and university data.	Dorina 2012
Gender, marital status, admission category, family income and size, parents' qualification and	Aggarwal *et al*. (2019)

occupation, number of friends, study hours, types of school attended. and travel time to college and home	
Department satisfaction, course attendance, preferred study time, planning, and friends' contributions.	Kayri (2015)
Gender, family background, distance, GPA, entrance exam, scholarships, time, materials, internet.	Osmanbegovic and Suljic (2012)

Existing features used in this research

The features of the proposed dataset have two components namely; features from current/existing features and the novel/new features.

The features from existing literature used in this work are; scores of different subjects (performance of student in relevant subject like Mathematics, English, Physics, Chemistry, Biology, Agriculture, Financial Accounting, Literature, Commerce, Civic Education, Economics, and Geography), Number of times student has failed, course type, school name, gender, number of terms, student average scores, and grade.

Apart from the current features some novel/new features were proposed. These features are school infrastructures and/or school facilities namely; teacher to student ratio, availability of laboratory reagent, availability of laboratory equipment, availability of textbooks in the library, textbook to student ratio, availability of visual equipment, number of students per seat, the type of board used in school, availability of electricity, and availability of portable drinking water.

Method of Data Collection

The method of data collection used is the primary and secondary methods. The primary method is the questionnaire given to schools to obtain data about the school's facilities and infrastructures. The secondary method is the students' academic records obtained from the various schools.

Population, and Sample

Nasarawa state has 13 local government areas with 297 public or government schools and 298 private schools.

Method

Part of the contribution of this paper is to derive insights from the proposed dataset. To derive insights there is need to clean the data. One of the insights is to know the features that are important. To determine the features that are important there is a need for a model to be developed that will detect these important features.

Data Cleaning

This stage involves cleaning of the dataset, converting categorical data to numerical variables and normalizing the data. Categorical variables were converted to numerical variables, this is because the model expects numbers. The type of encoding use is one hot encoding. The AVG column is derive by getting the average scores of the nine subjects. The N_fail is derived by counting the number of subjects a student has failed.

Algorithm

Random forest classifier was used to detect features that were important. The following steps were followed to create and train the model. Load dataset, data preparation, visualize data, Model creation, train model, test model, and evaluate model. The steps are shown in figure 1.

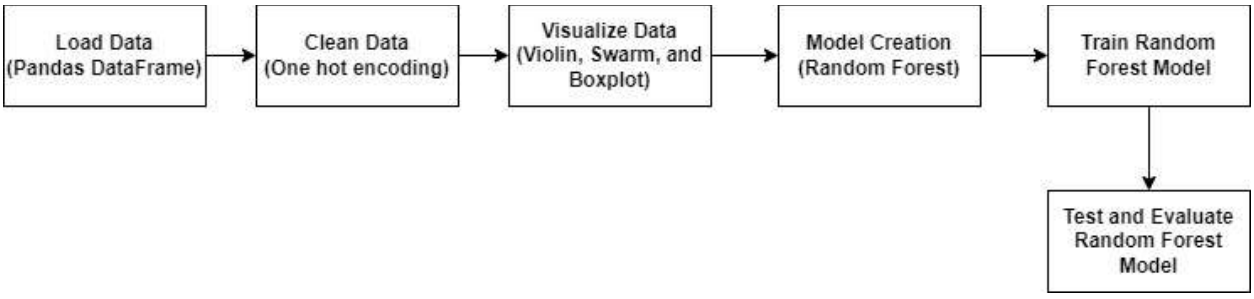

Figure 1. Model flow diagram.

The dataset was loaded into Pandas data frame for easy analysis. The data was prepared by cleaning the data as previously explained in the data cleaning stage. The data was visualized to know the distribution of the classes as shown in figure 2. The models were created using Scikit-learn. The model created is Random Forest Classifier. The model was trained using training data. The training data is 72% of the whole data. 8% of the data was used for validation. The model was tested using test data. The test data was 20% of the whole data. The model was evaluated using accuracy, precision, recall, and confusion matrix.

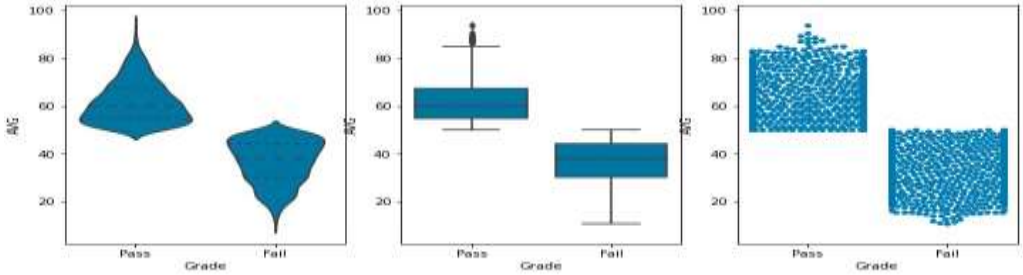

Figure 2. Violin plot, Boxplot, and swarm plot showing the distribution of the two classes.

Results and Discussion

The formulated dataset and the insights derive from the dataset are presented here.

Formulation of Student Dataset

The formulated dataset consists of 28 features from current literatures and proposed novel features. The students' dataset has a total of 3543 records from both private and public schools. The dataset had two categories (Fail and Pass). Figure 3 shows a sample of the formulated dataset.

	NAMES	School_Name	MTH	ENG	BIO	GEO	AGRIC	ECO	CIVIC	PHY_GOVT_COMM	...	Laboratory_equipments
0	02. CHUE DESMOND	Anty Dele College	41.666667	56.000000	46.666667	74.000000	67.000000	64.666667	60.000000	37.333333	...	Good
1	03. JOSEPH SUNDAY	Anty Dele College	35.666667	42.000000	32.000000	68.666667	59.000000	44.666667	58.666667	42.666667	...	Good
2	04. FRANCIS EPHRAIM	Anty Dele College	74.333333	74.666667	79.000000	91.666667	83.666667	81.000000	74.000000	88.000000	...	Good
3	05. ISHAYA EMMANUEL	Anty Dele College	40.000000	67.000000	55.000000	73.333333	69.666667	73.000000	71.666667	59.000000	...	Good
4	06. AKOR EMMANUEL	Anty Dele College	56.333333	73.000000	58.000000	81.666667	74.000000	69.333333	69.666667	71.333333	...	Good

5 rows × 28 columns

Figure 3. Sample of the formulated student dataset.

The essence of creating this dataset is to make it public for researchers to carry out research and derive insights from the dataset. Some of the insights derived from the dataset are discussed in succession.

Insights 1: Is the dataset balance?

This dataset was not balanced, because it has 3543 students, with 2016 that Failed and 1527 Passed.

Insights 2: What is the gender count?

The gender count was determined by distributing the classes on a pie chart with result as follows; 1792 males and 1751 females, which means males are more in the dataset. This result shows that there is need for government to encourage girl child education for gender equality.

Insights 3: Are the proposed features important?

Features importance shows if the proposed novel features are important, whether they can be useful for making decision. Figure 2 shows a plot of the importance of each feature.

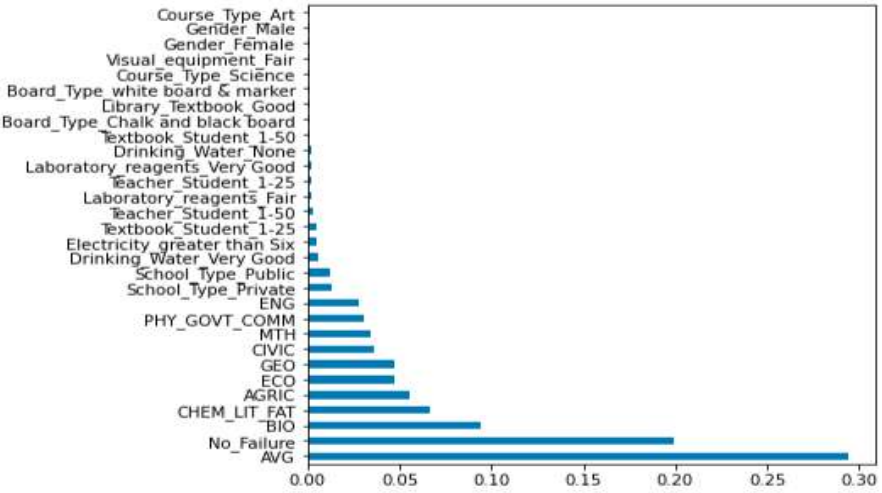

Figure 4. Feature importance.

Machine learning developers used feature importance to select features that will help their model in making decision. From the figure 4, the features that are important are ranked accordingly; starting with Average, number of failures, scores of all the subjects, school type, drinking water, electricity, textbook to student ratio, and laboratory reagent. Based on figure 4, the

following proposed features were proven to be important namely; availability of drinking water, availability of electricity, textbook to student ratio, and laboratory reagent.

Models Evaluation

The results of the model performance are presented in figures 5 and 6 based on the metrics. The random forest classifier achieved an accuracy of 1.0.

Random forest Classification Report

	precision	recall	f1-score	support
0	1.00	1.00	1.00	421
1	1.00	1.00	1.00	288
accuracy			1.00	709
macro avg	1.00	1.00	1.00	709
weighted avg	1.00	1.00	1.00	709

Figure 5. Classification report for Random Forest.

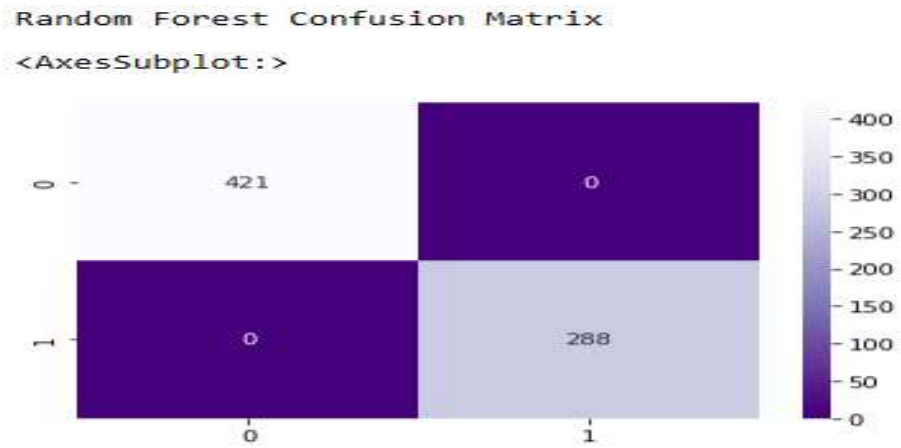

Figure 6. Confusion matrix for random forest.

Figure 6 shows there were 421 True positives, 0 False Negatives, 0 False Positives, and 288 True Negatives. The major focus is the False Negatives, because it means the model predicted these instances that Failed as passed, which would have been very unfortunate because the students will miss the necessary intervention intended for them.

Conclusion

The results show that the novel features proposed were really important and these features were not considered by the previous authors. From the preceding literature review, it is clear that fundamental data attributes which have significant impact on the performance of students were not considered by the authors. This is understandably so because the studies were conducted in different climes where each has its own peculiarities.

In other climes, electricity is available for 24 hours, there is portal drinking water, laboratories are well equipped, there are no congestion in classes, there are textbook in libraries for students to study, and there are enough teachers to meet the required number of teachers to student ratio.

In Nigeria, there is problem of electricity, some places are not even connected to the national grid, most places don't have portal drinking water, laboratories are not well equipped due to underfunding of education sector, the textbooks in the libraries are not enough and some are outdated, some schools don't have enough teachers, and due to lack of enough schools in some places the classes are congested. These are the reasons why these attributes were considered. Important features presented by the model can be used by stakeholders to make informed decisions. The model used here is a classification model, the problem can also be addressed using a regression model.

Acknowledgements

Authors are thankful to Nasarawa State University Keffi.

References

Aggarwal, D., Mittal, S., & Bali, V. (2019). Prediction model for classifying students based on performance using machine learning techniques. *International Journal of Recent Technology and Engineering*, *8*(257), 496–503. https://doi.org/10.35940/ijrte.B1093.0782S719

Alharbi, Z., Cornford, J., Dolder, L., & De La Iglesia, B. (2016). Using data mining techniques to predict students at risk of poor performance. In *Proceedings of the 2016 SAI Computing Conference (SAI 2016)* (pp. 523–531).

Arsad, P. M., Buniyamin, N., & Manan, J. L. A. (2013). A neural network students' performance prediction model (NNSPPM). In *2013 IEEE International Conference on Smart Instrumentation, Measurement and Applications (ICSIMA 2013)* (pp. 26–27).

Buniyamin, N., Bin Mat, U., & Arshad, P. M. (2016). Educational data mining for prediction and classification of engineering students' achievement. In *2015 IEEE 7th International Conference on Engineering Education (ICEED 2015)* (pp. 49–53).

Dorina, K. (2012). Student performance prediction by using data mining classification algorithms. *International Journal of Computer Science and Management Research, 1*(4).

Dixon, P., Humble, S., & Tooley, J. (2017). How school choice is framed by parental preferences and family characteristics: A study in poor areas of Lagos State, Nigeria. *Institute of Economic Affairs*, 53–65.

Gray, G., McGuinness, C., & Owende, P. (2014). An application of classification models to predict learner progression in tertiary education. In *Proceedings of the 2014 IEEE International Advance Computing Conference (IACC)*.

Guleria, P., Thakur, N., & Sood, M. (2015). Predicting student performance using decision tree classifiers and information gain. In *Proceedings of the 2014 3rd International Conference on Parallel, Distributed and Grid Computing (PDGC 2014)* (pp. 126–129).

Hamsa, H., Indiradevi, S., & Kizhakkethottam, J. J. (2016). Student academic performance prediction model using decision tree and fuzzy genetic algorithm. *Procedia Technology*.

Huang, S., & Fang, N. (2012). Work in progress: Early prediction of students' academic performance in an introductory engineering course through different mathematical modeling techniques. In *Proceedings of the Frontiers in Education Conference (FIE)* (Vol. 1, pp. 3–4).

Kayri, M. (2015). An intelligent approach to educational data: Performance comparison of the multilayer perceptron and the radial basis function artificial neural network. *Educational Sciences: Theory & Practice, 15*(5), 1247–1255. https://doi.org/10.12738/estp.2015.5.0238

Li, K. F., Rusk, D., & Song, F. (2013). Predicting student academic performance. In *Proceedings of the 2013 7th International Conference on Complex, Intelligent, and Software Intensive Systems (CISIS 2013)* (pp. 27–33).

Livieris, I. E., Drakopoulou, K., & Pintelas, P. (2012). Predicting students' performance using artificial neural networks. In *Proceedings of the 8th Pan-Hellenic Conference on Information and Communication Technologies in Education* (pp. 28–30).

Livieris, I. E., Drakopoulou, K., Tampakas, V., Mikropoulos, T., & Pintelas, P. (2019). Predicting secondary school students' performance using a semi-supervised learning approach. *Journal of Educational Computing Research, 57*(2), 448–470.

Meier, Y., Xu, J., Atan, O., & Van Der Schaar, M. (2016). Predicting grades. *IEEE Transactions on Signal Processing, 64*(4), 959–972.

Mohd Arsad, P., Buniyamin, N., & Ab Manan, J. L. (2014). Neural network and linear regression methods for prediction of students' academic achievement. In *IEEE Global Engineering Education Conference (EDUCON)* (pp. 916–921).

Omotor, D. G. (2004). An analysis of federal government expenditure in the education sector of Nigeria: Implications for national development. *Journal of Social Sciences, 9*(2), 105–110.

Onuma, N. (2016). Financial allocation to secondary education in Nigeria: Implication for students' performance. *IOSR Journal of Research & Method in Education, 6*(3), 42–47.

Sarker, F., Tiropanis, T., & Davis, H. C. (2014). Linked data, data mining and external open data for better prediction of at-risk students. In *Proceedings of the International Conference on Control, Decision and Information Technologies (CoDIT)*.

Uduu, O. (2022). Public schools record a 73.81% success rate in 2021 WAEC, highest in 6 years. *Dataphyte*. https://www.dataphyte.com/latest-reports/educationdevelopment/public-schools-record-a-73-81-success-rate-in-2021-waec-highest-in-6-years/

Ukpor, C. O., Ubi, I. O., & Okon, A. E. (2012). Assessment of factors determining parents' preference for private secondary schools in rural communities of Cross River State. *Global Journal of Educational Research, 11*(2), 99–106.

Xu, J., Moon, K. H., & Van Der Schaar, M. (2017). A machine learning approach for tracking and predicting student performance in degree programs. *IEEE Journal of Selected Topics in Signal Processing, 11*(5), 742–753.