MPRI-JAEE | INTERNATIONAL JOURNAL OF ASSESSMENT & EVALUATION IN EDU.
VOL. 04 NO. 8 – APRIL, 2024

COMPARATIVE STUDY OF MACHINE LEARNING TECHNIQUES IN PREDICTING STUDENTS DROPOUT IN COLLEGES OF EDUCATION IN NORTH-EASTERN NIGERIA

SAFIYANU AHMED
Department of Computer Science, Gombe State College of Education and Legal Studies, Nafada, P.M.B. 0191 Gombe State.
yarkada@gmail.com

Abstract

The research work aimed at predicting students' dropout from colleges of education due to poor academic performance by building four Machine learning models – Support Vector Machine (SVM), Naïve Bayes (NB), Decision Tree (DT), and K-Nearest Neighbor (KNN). The researcher sampled four public colleges of education in northeastern Nigeria, including Aminu Saleh College of Education Azare, Federal College of Education (Tech.) Gombe, Federal College of Education (Tech.) Potiskum, and Federal College of Education Yola. All first-year (N.C.E. 1) students from School of Science Education of the sampled colleges were conveniently included to participate in the study. Data were collected in two phases - **Phase I**: students' first semester results, (plus students' pre-college performance in UTME); and **Phase II**: Students second semester results, for the 2022/2023 academic session. We used WEKA application in developing our machine learning models. The predicted results (generated by the developed Machine Learning models), and the observed results (from students' second semester results) formed the basis for the study analysis. We also used Accuracy, Precision, Recall and F1-Score as the performance evaluation metrics for our developed models. It was found that Support Vector Machine was the model with highest prediction power with 89% (Accuracy), 0.92 (Precision), 0.95 (Recall), and 0.87 (F1-Score), and was thus recommended for use in our educational institutions for early detection of the students at risk of dropping out.

Keywords:
Comparative Study, Machine Learning Techniques, Students Dropout, Colleges of Education, North-Eastern.

EDITOR-IN-CHIEF
Prof. A.I. Mteiye [MALAWI]

Introduction

Machine learning is a subfield of computer science that evolved from the study of pattern recognition and computational learning theory in artificial intelligence. Machine learning explores the construction and study of algorithms that can learn from and make predictions on data. Such algorithms operate by building a model from example inputs in order to make data-driven predictions or decisions, rather than following strictly static program instructions.

There exist many reasons for students' dropout in which poor academic performance is the most influential cause. In Nigeria colleges of education, students whose Cumulative Grade Point Average (CGPA) are below the criterion point (1.00) are considered dropout. Increasingly, high dropout rates in science courses, especially in the students first year (NCE I) of study, have led to discussions for early detection of, and potential support measures for the at-risk students.

Different machine learning algorithms could be applied to help in the early detection of the at-risk students for prompt intervention and support. The effectiveness of the models in prediction might considerably vary, depending on the models used, and educational administrators might not know the best model to use to cater for their students. For such, we based our research on building and testing four machine learning models using the binary classifiers – Support Vector Machine (SVM), Naïve Bayes (NB), Decision Tree (DT), and K-Nearest Neighbor (KNN) to enable the researcher investigate their effectiveness in dropout prediction, and recommend them for use to educational administrators.

Statement of the Problem

Students' academic performances are usually classified as poor, fair, good, best, etc. Poor academic performance leads to failure and consequently dropping out of students from educational systems. High rate of students dropout in colleges of education tend to pose tremendous problems not only to the concerned students, but also to educational administrators, parents, society and other stake holders in education. It therefore became necessary for all stakeholders in education to device means by which this hypothetical problem could be checked.

Objectives of the Study

The objectives of this study are:

1. To determine science students dropout rate in colleges of education in North-Eastern Nigeria.
2. To investigate the effectiveness (prediction power) of the machine learning techniques used.
3. To identify the most accurate algorithm for students dropout prediction.
4. To recommend machine learning model(s) with highest prediction power for use in educational institutions for early detection of the students at risk so that preventive and performance improvement measures can be taken.

EDITOR-IN-CHIEF
Prof. A.I. Mteiye [MALAWI]

LITERATURE REVIEW

Exploring educational data can certainly help in achieving the desired educational goals. By applying Educational Data Mining techniques, it is possible to develop prediction models to improve student success. (Eyman A. & Dilek D., 2020). Machine learning techniques can help in predicting students' dropout in our educational institutions. Early students' dropout prediction can help academic institutions to provide a timely intervention and apply appropriate planning and training to improve students' success rate. (Meseret Y. A. & Stanislava S., 2021).

Algorithms in many research used information in a variety of ways to find the most accurate models. (Jiraporn C. et al., n.d.). Different machine learning techniques, including Decision Tree, Random Forest, Logistic Regression and Naïve Bayes were used for training and testing model in a study conducted at Hawassa University, Ehiopia. Outcome of the research benefited course instructors, institutes, and the University to decide on students' performance and apply appropriate intervention for improving students' academics in advance. The study found out that the Logistic Regression model performed better than the remaining models used in that study in predicting students' early dropouts. (Meseret Y. A. & Stanislava S., 2021).

Demonstrated in a case study was the effectiveness of several classification techniques and the cost-sensitive learning approach on the dataset from the Electrical Engineering department of Eindhoven University of Technology. The experimental results showed that rather simple classifiers gave a useful result with accuracies between 75 and 80% that was hard to beat with other more sophisticated models. (Gerben W. D. et al., 2009).

An empirical study was performed using most common Machine learning techniques on the student dataset at a Vietnamese university. The results indicated that 100% of the students were classified with a high correctly classified rate, in which Multilayer Perceptron (MLP) and Naïve Bayes gave higher TP rate and recall for most classes. It was noted that those findings could be applicable to other higher education institutions in Vietnam with regard to predicting students' final academic results, subsequently providing students with any necessary support from lecturers and university administrators. The output model could also be integrated into a management information system of a university to provide early warning function to students who are at risk. (Dinh T. H. et al., 2020).

Another study conducted in Bangladesh revealed that many students participate in board exam every year without understanding their overall performance in earlier. As a result, student dropout rates create capital wastage for all actors in the education sector and also impact the institutions' assessment processes. Through evaluating student's previous results of all regular exams using machine learning techniques, teachers can anticipate the students resulting in the board exam. Khalil(Khalil A. et al. 2021).

In a study carried out on 72,598 pupils, a random forest achieved an accuracy of 93.5% and an AUC of 0.965. Thus, the predictor was accurate enough to be used as a useful support

tool for teachers allowing them to take early countermeasures for preventing dropout. (Nicolae-Bogdan S. et al., 2015).

Bayesian Networks have a high potential to predict students at risk of not passing their core first-year courses in Computer Science. In particular, failing at least one of the first-year mathematics and computer science courses can be predicted with a 90.64% accuracy (on average). This finding justifies the method of identifying At-Risk-students automatically. Once these students are identified, they can be enrolled in the EDP in order to improve students' academic success and graduation throughput. (Zachary B. N. et al., 2019).

Students' dropout rate is one of the primary performance indicators of e-learning. It is crucial in the case of the virtual learning environment at Constantine, the Philosopher University in Nitra, Slovakia, as well as in the case of MOOCs (Massive Open Online Course). The reliable prediction can encourage students to study harder when they know that they are at risk and plan their semester workload carefully. The school management can take this information into account when deciding whether to confirm or reject students' requests who appeal for repeating failed online courses. Predicting students at-risk of dropout also gives educators a signal to intervene when students are at risk and improve students' involvement in interactive activities in time. It is essential to reveal what courses have in common and how this set of features can influence students' behavior and performance. The results discovered that the classification models could be considered trustworthy enough to make a proper prediction of completing or early dropout before the end of a semester. In fact, good results of prediction were obtained, regardless of the employed model. The features selected in that study proved to be influential in the prediction of completer and non-completers. The prediction accuracy varied between 77 and 93%. (Janka K. & Martin D., 2021).

Another research provided a comprehensive review of the most recent and relevant research endeavors on machine learning application toward predicting, explaining and solving the problem of student dropout in MOOCs. It highlighted both student-related factors and MOOC related factors that lead to a high number of dropouts. The paper also identified some of the critical challenges associated with student dropout prediction and provided recommendations and proposal to assist researchers employing various machine learning techniques in solving it timely and efficiently. Additionally, the paper floated the idea of unification of a clickstream data and student-provided data as a standard, similar to various learning objects metadata standards. (Fisnik D. et al., 2018).

Also, it drew on the detailed student logs, extracted from the Learning Management System that hosted the e-learning courses, to dynamically make its estimations and adapt them to student progress throughout the course. The method used three machine learning techniques, namely feed-forward neural networks, support vector machines and probabilistic ensemble simplified fuzzy ARTMAP. To overcome individual technique inaccuracies in identifying dropout students, the method combined their estimations

using three different decision schemes. The three individual machine learning techniques, as well as the three decision schemes were tested using data from two e-learning courses, to determine their overall accuracy, sensitivity and precision. The experimental results showed that the most successful technique in promptly and accurately predicting dropout-prone students was the decision scheme which identified a student as being a dropout if that student had been determined as such by at least one of the three machine learning techniques. Using these schemes, the proposed method achieved a 75–85% overall student classification rate from the first section of the two courses, to reach a 97–100% rate in the final sections. The results concerning the sensitivity and precision criteria were also high, indicating that the scheme was accurate in both correctly identifying dropouts and avoiding completer misclassifications. (Ionna et al., 2009).

In another study, the researchers aimed at building a dropout early warning system that maximizes their chance of supporting the successful learning of all students, and minimizes the cost for intervention by addressing the class imbalance issue using the SMOTE and ensemble methods in machine learning, and also by evaluating the trained classifiers with both ROC and PR curves. ROC curves were not very informative, whereas PR curves were informative. According to their PR curves, BDT showed the best performance. (Sunbok L. & Jae Y. C., 2019).

An ML-based system was also designed and developed to evaluate the performance of high school students. The system evaluated students from different aspects and classified each student into four classes: very well, good, medium and bad. Due to the lack of appropriate educational databases, the best information needed was collected individually with the help of online questionnaires and the collaboration of teachers. The most influential features were determined and analyzed, and the most efficient and optimal ML models were identified as well. Moreover, the models were trained again with the decreased dataset that contained data with the most important features. Therefore, more efficient models were made, and students were evaluated without human intervention. (Mostafa et al., 2021).

In another scenario, it was proved that the learning algorithms predict dropout of new students with satisfying accuracy and thus become a useful tool in an attempt to prevent and therefore reduce dropouts. The accuracy reached 63% in the initial predictions based only on demographic data of the students and exceeded 83% before the middle of the academic period. The comparison of the six algorithms showed that the Naive Bayes algorithm is the most appropriate. A prototype web based support tool, which could recognize students with high probability of dropout, was constructed for implementing that algorithm. (Sotiris K. et al., 2003).

Furthermore, an AI based expert system was proposed for predicting the dropout rate of students in a college using data mining classification techniques. The system might give benefit to the students, parents, teachers, academicians, society and owners to analyze the performance of the students and to impart the quality of education in the educational

institutions and make decision regarding the improvements. The classification results illustrated that the EMLA was the best algorithm with detection rate of 78.31% as compared to NDTREE and Decision Stump algorithms. (Behal S. & Sallan G., 2020).

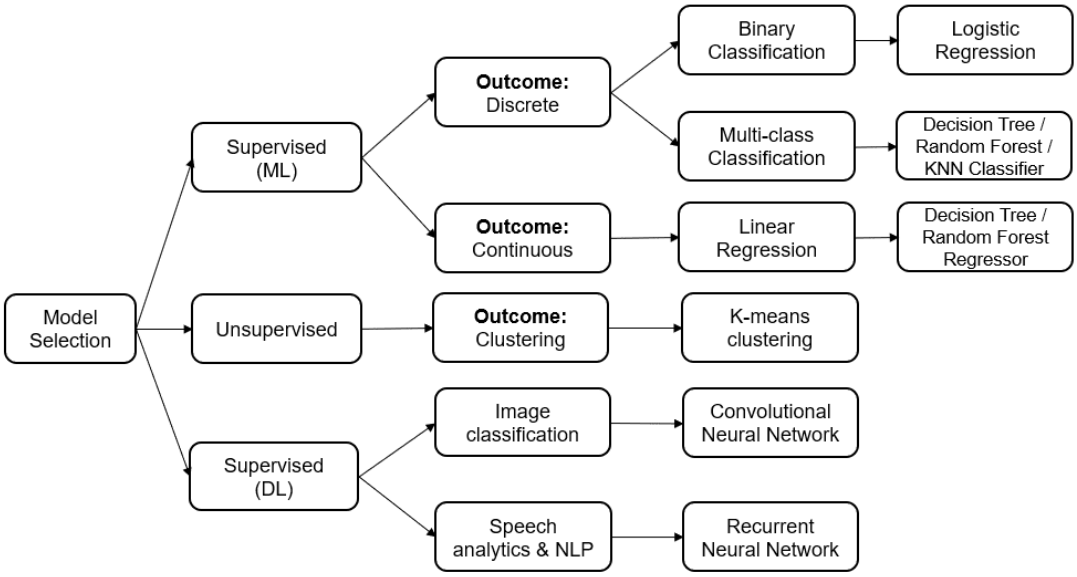
A survey of machine learning techniques on addressing student dropout problem was presented in research. The survey drew several conclusions; First, while several techniques were proposed for addressing student dropout in developed countries, there was lack of research on the use of machine learning for addressing that problem in developing countries. Second, despite the major efforts on using machine learning in education, data imbalance problem were ignored by many researchers. That facilitated using improper evaluation metrics on analyzing performance of the algorithms. Third, many researches focus on providing early prediction rather than including ranking and forecasting mechanisms on addressing the problem of student dropout. Lastly, school level datasets must be considered when addressing that problem, in order to come up with the proposed solutions to facilitate the authorities on identifying at risk schools for early intervention. (Mduma N., etal., 2019).

METHODOLOGY

Machine Learning Model Development

The machine learning model development involves data acquisition from multiple trusted sources, data processing to make suitable for building the model, choosing algorithm to build the model, building the model, computing performance metrics and choosing best performing model.

Building an ML model requires splitting of data into two sets, such as 'training set' and 'testing set' in the ratio of 80:20 or 70:30. A set of supervised (for labelled data) and unsupervised (for unlabeled data) algorithms are available to choose from depending on the nature of input data and business outcome to predict. In case of labelled data, it is recommended to choose logistic regression algorithm if the outcome to predict is of binary (0/1) in nature; choose decision tree classifier (or) Random Forest classifier (or) KNN if the outcome to predict is of multi-class (1/2/3/...) in nature; and choose linear regression algorithm (e.g., decision tree regressor, random forest regressor) if the outcome to predict is continuous in nature. The clustering (Unsupervised) algorithm is preferred to analyze unlabeled / unstructured (text) data (e.g., k-means clustering). Artificial Neural Network (ANN) algorithms are suggested to analyze other unstructured (image/voice) data types, such as Convolutional Neural Networks (CNN) for image recognition and Recurrent Neural Networks (RNN) for voice recognition and Natural Language Processing (NLP). The model is built using training dataset and make prediction using test dataset. Use of deep learning (neural networks) models is preferred over regression models (ML models) for better performance as these models introduce extra layer of non-linearity with the introduction of Activation Function (AF). The algorithm selection under various scenarios has been represented in figure below.



Algorithm Selection Criteria

Computation of Model Performance is next logical step to choose the right model. The model performance metrics will decide on final model selection, that include computation of accuracy, precision, recall, F1 score (weighted average of precision and recall), along with confusion matrix for classification models and co-efficient of determination for regression models. It is not recommended to use accuracy as a measure to determine the performance of classification models that are trained with imbalanced/skewed datasets, rather precision and recall are recommended to be computed to choose right classification model.

Study Structure

The study covered northeast geopolitical zone of Nigeria. The zone consisted of six states including Adamawa, Bauchi, Borno, Gombe, Taraba and Yobe. Four public colleges of education were strategically sampled, including Aminu Saleh College of Education Azare (in Bauchi State), Federal College of Education (Tech.) Gombe (in Gombe State), Federal College of Education (Tech.) Potiskum (in Yobe State), and Federal College of Education Yola (in Adamawa State). The researcher used convenience sampling to include all first-year (N.C.E. 1) students from School of Science Education of the sampled colleges to participate in the study. Outcome of this study formed the basis for generalization to the entire colleges of education within the study area. Therefore, all first-year (N.C.E. 1) science education students from colleges of education within the study area served as the population of this study.

Method of Data Collection

Students' first semester results for the 2022/2023 academic session (plus students' pre-college performance in Unified Tertiary Matriculation Examination (UTME)) were

obtained from Students Information System (SIS) from each of the four sampled colleges to form a dataset for that college. Each dataset was split into two – training (70%) and testing (30%) sub-datasets. The training sub-datasets were used to train machine learning models, while the testing sub-datasets were used to test (validate) the accuracy of the models. The models with best performances for the applied machine learning techniques were used in predicting students' dropout beforehand. We later obtained the students second semester results (for the same academic session), and compared the outcomes of the results (**observed results**) with the predictions (**predicted results**) made by our preferred models.

Method of Data Analysis

The researcher used an all-inclusive WEKA application (version 3.8.6) in developing machine learning models using classification algorithms. We started by loading the WEKA application, and opened the Explorer window by clicking on the "Explorer" tab. The Explorer window consisted of the "Preprocess", "Classify", "Cluster", "Associate", "Select Attributes", "Visualize" and "Auto WEKA" tabs. By default, only the "Preprocess" tab becomes active, to initially allow user to import data file (in a .csv – Comma Separated Values file format) into the WEKA Application by clicking on the "Open file…" button. We then selected "Reg. Num", "UTME", "GPA" and "Status", under the "Attributes" list. The values of the "Status" attribute were jointly derived based on the values of the "UTME" and "GPA" using the following Excel formula:

C2 = IF(OR((AND(B5>=140, C5>=0.75)), (AND(B5>=120, C5>=1)), (AND(B5>=100, C5>=1.25))), "Pass", "Fail")

where C2 is the hosting cell.

In the process of developing our models, we clicked on the "Classify" tab in the Explorer window and selected the intended classifier from a drop down menu, by clicking on the "Choose" button under the "Classifier" section. We then selected and set the value of the "Percentage Split" option under the "Test options" section to 70%, which split the dataset into training and testing sub-datasets in the ratio 70:30. We then clicked on the "More options…" button and set other options in the "Classifier evaluation options" dialog box. We also selected either "Plain Text" (or "CSV") from the drop-down menu by clicking the "Choose" button right to the "Output predictions". We further set the class attribute to Status (a user-defined attribute from the imported file) from a drop-down menu by clicking an arrow with the tip "select the attribute to use as the class". We finally triggered action (executed programs) by clicking on the "Start" button which resulted in some output in the "Classifier output" environment. A unique file name is assigned to the resulted output which automatically appeared in the "Result list" environment. You can right-click on the file name to perform other actions as "Save result buffer", "Delete result buffer(s)", "Load model", Save model", etc.

With the above developed model settings, we proceeded and selected "Use training set" option under "Test options" section, and clicked the "Start" button again, which generated the predicted results in the "Classifier output" environment using the entire dataset (100%). We renamed the predicted result by right-clicking on the output file name and selected "Save result buffer" from the drop-down menu, we then saved the file with the .csv file extension.

Furthermore, we opened our saved file (e.g. fceGombe_NV_predictedResult.csv) using Microsoft Excel application, and moved to the "=== Predictions on training set ===" section, where we then added two fields – "observed" and "remark" next to the five existing fields – "inst#", "actual", "predicted", "error" and "prediction". We then manually populated (filled in) the "observed" column with the current status of the students based on the outcome of their second semester results as either "Pass" (succeeded) or "Fail" (dropout). Finally, we inserted the following Excel formula in the immediate cell below the "remark" field:

G2 =IF(AND(C2="Pass", F2="Pass"), "TP", IF(AND(C2 = "Fail", F2="Fail"), "TN", IF(AND(C2 = "Pass", F2="Fail"), "FP", IF(AND(C2 = "Fail", F2="Pass"), "FN", "Error"))))

where G2 was the hosting cell, which aided in populating the "remark" column with TP, FP, TN, and FN parameters (for confusion matrix), by comparing the predicted results with the observed results, with the following interpretation:

1. If "predicted" = "Pass" and "observed" = "Pass", then "remark" = "TP" (True Positive)
2. If "predicted" = "Fail" and "observed" = "Fail", then "remark" = "TN" (True Negative)
3. If "predicted" = "Fail", while "observed" = "Pass", then "remark" = "FP" (False Positive)
4. If "predicted" = "Pass", while "observed" = "Fail", then "remark" = "FN" (False Negative)

Confusion Matrix

We used Confusion Matrix in evaluating the performance of our machine learning models. The table below gathered four important parameters related to a given success prediction model:

Class	Predicted Class	
	P	**N**
P	True Positive (TP)	False Negative (FN)
N	False Positive (FP)	True Negative (TN)

Confusion Matrix

Where:

- True Positive (TP): number of successful students classified correctly as "successful".

EDITOR-IN-CHIEF
Prof. A.I. Mteiye [MALAWI]

- True Negative (TN): number of not successful students classified correctly as "non-successful".
- False Positive (FP): number of successful students incorrectly classified as "non-successful".
- False Negative (FN): number of not successful students classified incorrectly as "successful".

Performance Evaluation Metrics

We evaluated the performances of our developed machine learning models using Accuracy, Precision, Recall, and F1-Score, as provided by the following formulae.

- **Accuracy**

  The accuracy is calculated by dividing the number of correct classes predicted by the total number of types (classes). Thus, it measures the overall rate or classification correctness.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{FP} + \text{FN} + \text{TP} + \text{TN}}$$

- **Recall**

  It is the proportion of real positives predicted to be positive. Recall ensures that the predictive model is not overlooking a few students who are Fail or Pass. The Recall is used to evaluate the actual success rate/ students who passed successfully.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

- **Precision**

  Precision is the proportion of real negatives expected to be negative. It determines the fraction of true positives among true positives and false positives predicted. The precision is used to assess the dropout of students, and it is determined as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

- **F1score**

  The F1score determines the harmonic mean of recall and precision of a predictive model. Therefore, the F1score is suitable for classification problems where the target labels are imbalanced.

The F1score shows the balance between two measures of classification. It represents a measure widely used trade-off calculation for imbalanced datasets. It is determined as follows:

$F1 - Score = \sqrt{Recall * Precision}$ Or $\frac{\text{TP}}{\text{TP} + \frac{1}{2}(\text{FP} + \text{FN})}$

Results and Discussion

The outcomes of the study were presented in both tabular and bar chart forms, discussed and analyzed based on the findings from each of the four sampled institutions in relations

to the four machine learning techniques used. The performance evaluation metrics – Accuracy, Precision, Recall, and F1-Score obtained were also compared to facilitate in the realization of the stated objectives of the study.

Table 1: Aminu Saleh College of Education Azare

S/N	Machine Learning Technique	Performance Evaluation Metrics			
		Accuracy	Precision	Recall	F1-Score
1.	Support Vector Machine (SVM)	0.92	0.96	0.94	0.88
2.	Naïve Bayes (NB)	0.88	0.92	0.93	0.88
3.	Decision Tree (DT)	0.85	0.90	0.91	0.88
4.	K-Nearest Neighbor (KNN)	0.82	0.89	0.88	0.87

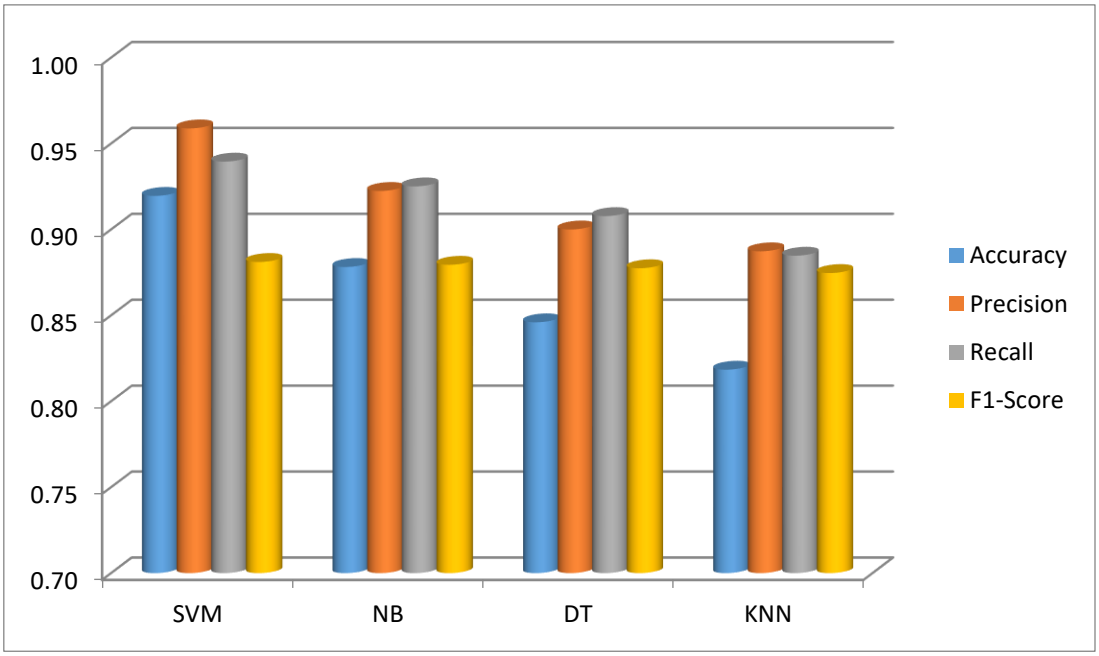


Fig. 1: Aminu Saleh College of Education Azare

Table 1 presented the results obtained from Aminu Saleh College of Education Azare. The experiment was carried out on 435 participants. In ***Accuracy***, the applied models showed that Support Vector Machine had 0.92 (92%), Naïve Bayes had 0.88 (88%), Decision Tree had 0.85 (85%), and K-Nearest Neighbor had 0.82 (82%). While in ***Precision***, the applied models showed that Support Vector Machine had 0.96, Naïve Bayes had 0.92, Decision Tree had 0.90, and the K-Nearest Neighbor had 0.89. While in ***Recall***, the applied models showed that Support Vector Machine had 0.94, Naïve Bayes had 0.93, Decision Tree had 0.91, and the K-Nearest Neighbor had 0.88. While in the ***F1-Score***, the applied models showed that Support Vector Machine, Naïve Bayes, and Decision Tree, all had 0.88 each, while the K-Nearest Neighbor had 0.87.

We can therefore infer, from the performance evaluation metrics in Table 1 that Support Vector Machine, with 92% Accuracy, 0.96 Precision, 0.94 Recall, and 0.88 F1-Score, outperformed excelled other machine learning models used.

Table 2: Federal College of Education (Tech.) Gombe

S/N	Machine Learning Technique	Performance Evaluation Metrics			
		Accuracy	Precision	Recall	F1-Score
1.	Support Vector Machine (SVM)	0.90	0.91	0.97	0.86
2.	Naïve Bayes (NB)	0.88	0.89	0.96	0.85
3.	Decision Tree (DT)	0.83	0.88	0.89	0.85
4.	K-Nearest Neighbor (KNN)	0.76	0.81	0.90	0.85

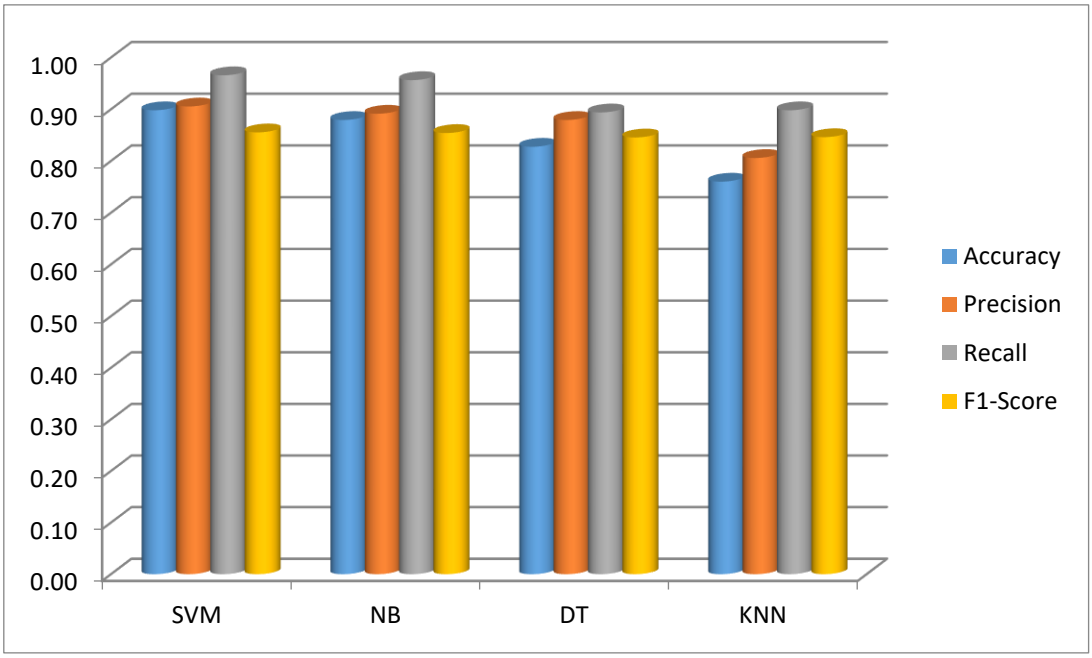


Fig. 2: Federal College of Education (Tech.) Gombe

Table 2 presented the results obtained from Federal College of Education (Tech.) Gombe. The experiment was carried out on 1,275 participants. In ***Accuracy***, the applied models showed that Support Vector Machine had 0.90(90%), Naïve Bayes had 0.88 (88%), Decision Tree had 0.83 (83%), and K-Nearest Neighbor had 0.76 (76%). While in ***Precision***, the applied models showed that Support Vector Machine had 0.91, Naïve Bayes had 0.89, Decision Tree had 0.88, and the K-Nearest Neighbor had 0.81. While in ***Recall***, the applied models showed that Support Vector Machine had 0.97, Naïve Bayes had 0.96, Decision Tree had 0.89, and the K-Nearest Neighbor had 0.90. While in the ***F1-Score***, the applied models showed that Support Vector Machine had 0.86, while Naïve Bayes, Decision Tree and the K-Nearest Neighbor all had 0.85 each.

We can therefore infer, from the performance evaluation metrics in Table 2 that Support Vector Machine, with 90% Accuracy, 91% Precision, 97% Recall, and 86% F1-Score, also excelled other machine learning models used.

Table 3: Federal College of Education (Tech.) Potiskum

S/N	Machine Learning Technique	Performance Evaluation Metrics			
		Accuracy	Precision	Recall	F1-Score
1.	Support Vector Machine (SVM)	0.93	0.97	0.94	0.88
2.	Naïve Bayes (NB)	0.90	0.92	0.95	0.88
3.	Decision Tree (DT)	0.85	0.93	0.88	0.87
4.	K-Nearest Neighbor (KNN)	0.73	0.82	0.84	0.86

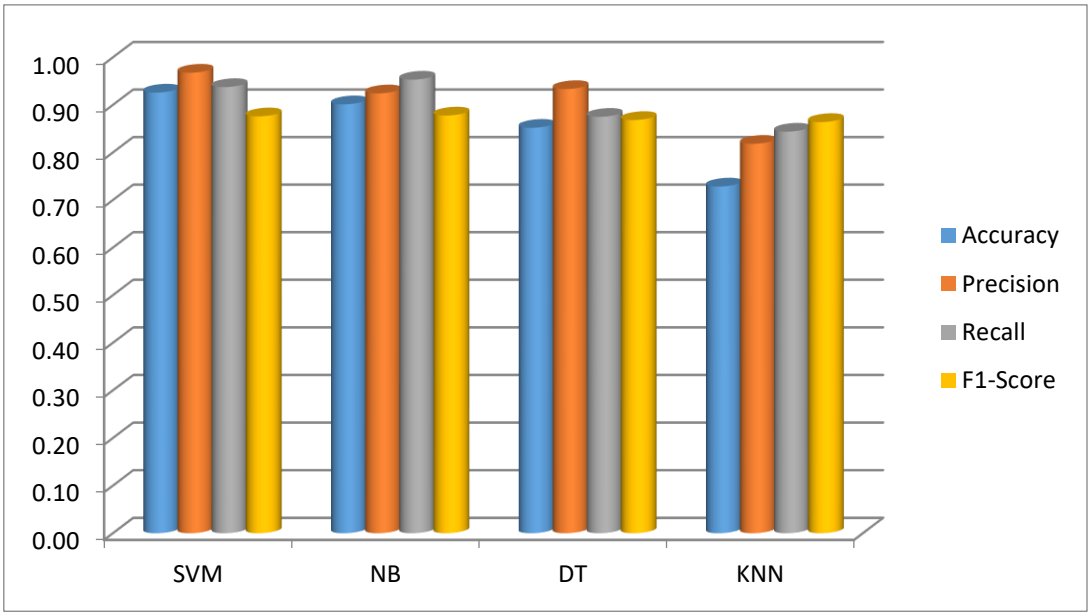


Fig. 3: Federal College of Education (Tech.) Potiskum

Table 3 presented the results obtained from Federal College of Education (Tech.) Potiskum. The experiment was carried out on only 81 participants, which had the least number of participants from the sampled institutions. In ***Accuracy***, the applied models showed that Support Vector Machine had 0.93 (93%), Naïve Bayes had 0.90 (90%), Decision Tree had 0.85 (85%), and K-Nearest Neighbor had 0.73 (73%). While in ***Precision***, the applied models showed that Support Vector Machine had 0.97, Naïve Bayes had 0.92, Decision Tree had 0.93, and the K-Nearest Neighbor had 0.82. While in ***Recall***, the applied models showed that Support Vector Machine had 0.94, Naïve Bayes had 0.95, Decision Tree had 0.88, and the K-Nearest Neighbor had 0.84. While in ***F1-Score***, the applied models showed that both Support Vector Machine and Naïve Bayes had 0.88 each, Decision Tree had 0.87, while K-Nearest Neighbor had 0.86.

We can therefore infer, from the performance evaluation metrics in Table 3 that, Support Vector Machine, with 93% Accuracy, 0.97 Precision, 0.94 Recall, and 0.88 F1-Score, also excelled other machine learning models used.

Table 4: Federal College of Education Yola

S/N	Machine Learning Technique	Performance Evaluation Metrics			
		Accuracy	Precision	Recall	F1-Score
1.	Support Vector Machine (SVM)	0.87	0.90	0.94	0.89
2.	Naïve Bayes (NB)	0.90	0.96	0.91	0.88
3.	Decision Tree (DT)	0.85	0.91	0.90	0.88
4.	K-Nearest Neighbor (KNN)	0.79	0.89	0.84	0.88

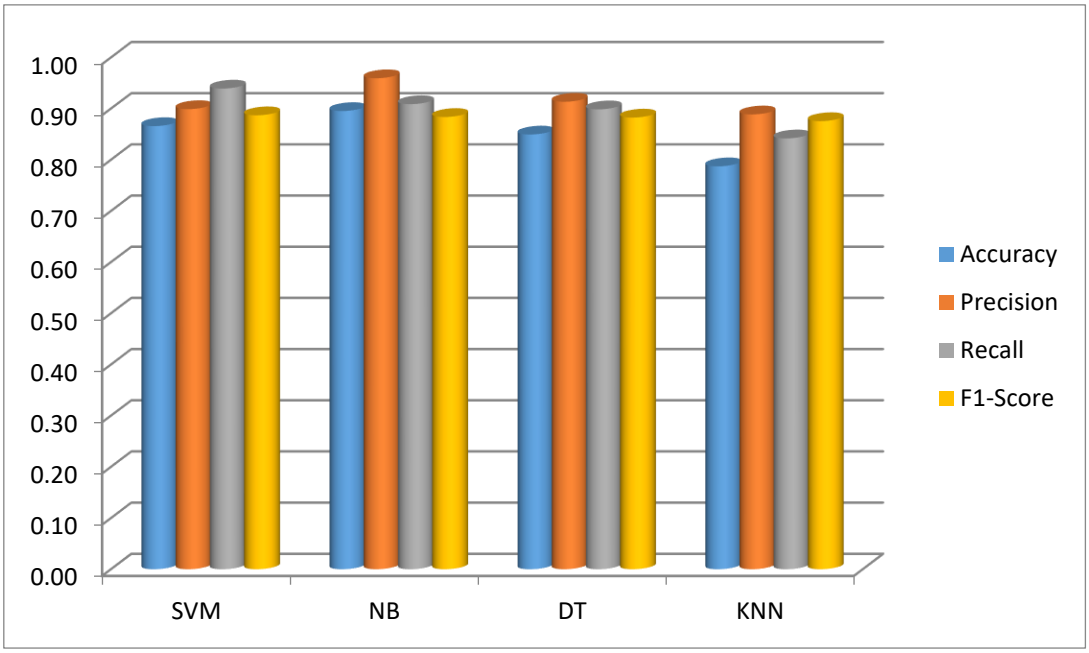


Fig. 4: Federal College of Education Yola

Table 4 presented the results obtained from Federal College of Education Yola. The experiment was carried out on 611 participants. In ***Accuracy***, the applied models showed that Support Vector Machine had 0.87 (87%), Naïve Bayes had 0.90 (90%), Decision Tree had 0.85 (85%), and K-Nearest Neighbor had 0.79 (79%). While in ***Precision***, the applied models showed that Support Vector Machine had 0.90, Naïve Bayes had 0.96, Decision Tree had 0.91, and the K-Nearest Neighbor had 0.89. While in ***Recall***, the applied models showed that Support Vector Machine had 0.94, Naïve Bayes had 0.91, Decision Tree had 0.90, and the K-Nearest Neighbor had 0.84. While in the ***F1-Score***, the applied models showed that Support Vector Machine had 0.89, while Naïve Bayes, Decision Tree and K-Nearest Neighbor all had 0.88 each.

We can therefore infer, from the performance evaluation metrics in Table 4 that Naïve Bayes, with 90% Accuracy, 0.96 Precision, 0.91 Recall, and 0.88 F1-Score, excelled other machine learning models used.

Table 5: Merged Result for the Four Sampled Colleges

S/N	Machine Learning Technique	Performance Evaluation Metrics			
		Accuracy	Precision	Recall	F1-Score
1.	Support Vector Machine (SVM)	0.89	0.92	0.95	0.87
2.	Naïve Bayes (NB)	0.88	0.91	0.94	0.87
3.	Decision Tree (DT)	0.84	0.89	0.90	0.86
4.	K-Nearest Neighbor (KNN)	0.78	0.84	0.88	0.86

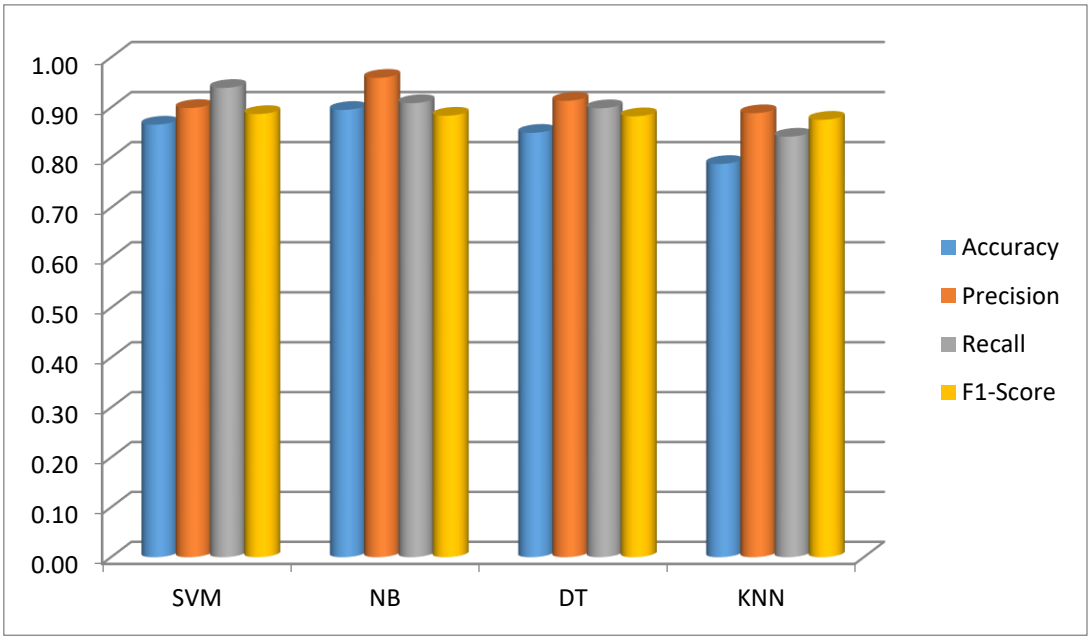


Fig. 5: Merged Result for the Four Sampled Colleges

Table 5 presented the merged results for the four sampled colleges of education in the study area, including 435 participants from Aminu Saleh College of Education Azare, 1,275 participants from Federal College of Education (Tech.) Gombe, 81 participants from Federal College of Education (Tech.) Potiskum, and 611 participants from Federal College of Education Yola. Therefore, the experiment was carried out on 2,402 participants altogether.

In ***Accuracy***, the applied models showed that Support Vector Machine had 0.89 (89%), Naïve Bayes had 0.88 (88%), Decision Tree had 0.84 (84%), and K-Nearest Neighbor had 0.78 (78%). While in ***Precision***, the applied models showed that Support Vector Machine had 0.92, Naïve Bayes had 0.91, Decision Tree had 0.89, and the K-Nearest Neighbor had

0.84. While in ***Recall***, the applied models showed that Support Vector Machine had 0.95, Naïve Bayes had 0.94, Decision Tree had 0.90, and the K-Nearest Neighbor had 0.88. While in the ***F1-Score***, the applied models showed that both Support Vector Machine and Naïve Bayes had 0.87 each, while both Decision Tree and K-Nearest Neighbor had 0.86 each.
We can therefore infer, from the performance evaluation metrics in Table 5, that Support Vector Machine, with 89% Accuracy, 0.92 Precision, 0.95 Recall, and 0.87 F1-Score, excelled other machine learning models used.

Table 6: Dropout Rates for the Four Sampled Colleges

S/N	Institution	Participants	Dropout	Dropout Rate
1.	Aminu Saleh College of Education Azare	435	88	20.23
2.	Federal College of Education (Tech.) Gombe	1275	314	24.63
3.	Federal College of Education (Tech.) Potiskum	81	17	20.99
4.	Federal College of Education Yola	611	118	19.31
		2402	**537**	**22.36**

Table 5 showed that Aminu Saleh College of Education Azare, Federal College of Education (Tech.) Gombe, Federal College of Education (Tech.) Potiskum and Federal College of Education Yola had dropout rate of 20.23% (88 out of 435), 24.63% (314 out of 1275), 20.99% (17 out of 81) and 19.31% (118 out of 611) respectively, with the merged dropout rate of 22.36% (537 out of 2402) for the four sampled colleges.

Conclusion

The use of machine learning models has nowadays become a handy tool for predicting success or failure of an organization in which education is not exceptional. It has become imperative for educational administrators to embark on the use of machine learning techniques for early detection of students at risk of dropout, and for the prevention of the menace the dropout poses to education system. The number of successful candidates promoted to next level has always been diminishing due to persistent increase in dropout rates from one education level to another, especially at the initial stage (first year i.e. NCE 1) of the students' endeavor, due to the fact that the students might have been new in the system and/or lack requisite education background that will enable them succeed in their study.
This study showed that, Support Vector Machine (SVM) had highest prediction power over other machine learning models used in predicting science students' dropout from colleges of education in northeastern Nigeria. We found that Support Vector Machine (SVM) outpaced Naïve Bayes (NB), Decision Tree (DT), and K-Nearest Neighbor (KNN) in

EDITOR-IN-CHIEF
Prof. A.I. Mteiye [MALAWI]

Aminu Saleh College of Education Azare, Federal College of Education (Tech.) Gombe and Federal College of Education (Tech.) Potiskum as shown in Table 1, Table 2 and Table 3 respectively. On the other hand, Naïve Bayes (NB) took ride in Federal College of Education Yola, as shown in Table 4. The study further revealed, from the performance evaluation metrics in Table 5, that both Support Vector Machine and Naïve Bayes had 0.87 F1-Score, but differ in Accuracy, Precision and Recall by 1% only. We can therefore infer that Support Vector Machine (SVM), with 89% Accuracy, 0.92 Precision, 0.95 Recall, and 0.87 F1-Score, is the model with highest prediction power, followed by Naïve Bayes (NB), with 88% Accuracy, 0.91 Precision, 0.94 Recall, and 0.87 F1-Score.

Recommendation

For the effective and efficient use of machine learning models in dropout prediction, we recommend for the use of Support Vector Machine (SVM) and Naïve Bayes (NB) for early detection of the at-risk students so that proper mechanism and preventive measures can be put in place to check the menace of students' dropout from educational institutions, especially colleges of education.

References

Behal S. & Sallan G. (2020). Prediction of student dropout using enhanced machine learning algorithm. Advances in Mathematics: *Scientific Journal* 9 (2020), no.6, 3821–3826. ISSN: 1857-8365 (printed); 1857-8438 (electronic) https://doi.org/10.37418/amsj.9.6.61

Dinh T. H., Pham T. T. L., Nguyen C. G., Nguyen T. L. H. & Loan L. (2020). An empirical study for student academic performance prediction. *International Journal of Computer Science and Information Security (IJCSIS)*, Vol. 18, No. 3, March 2020. https://sites.google.com/site/ijcsis/ ISSN 1947-5500

Eyman A. & Dilek D. (2020). Predicting academic success in higher education: literature review and best practices. *International Journal of Educational Technology in Higher*. doi.org/10.1186/s41239-020-0177-7

Fisnik D., Ali S. I. & Zenun K. (2018). MOOC dropout prediction using machine learning techniques: review and research challenges. http://10.1109/EDUCON.2018.8363340

Gerben W. D., Mykola P. & Jan M. V. (2009). Predicting students drop out: A Case Study. Educational Data Mining 2009.

Janka K. & Martin D. (2021). Towards predicting student's dropout in university courses using different machine learning techniques. Appl. Sci. 2021, 11, 3130. https://doi.org/10.3390/app11073130

Jiraporn C., Busayamas P., Somkid A., Wannapong T., & Narin N. (n.d.). A comparison of machine learning algorithms and their applications. https://doi10.5013/IJSSST.a.20.04.08. ISSN: 1473-804x online, 1473-8031 print

Khalil A., Partha C., Evana A., Umme H. F. & Saifur R. (2021). A comparative study of different machine learning techniques to predict the result of an individual student using previous performances. *International Journal of Computer Science and Information Security* (IJCSIS), Vol. 19, No. 1, January 2021. https://sites.google.com/site/ijcsis/ISSN 1947-5500

Mduma N., Khamisi K. & Dina M. (2019). A survey of machine learning approaches and techniques for student dropout prediction. Data Science Journal, 18: 14, pp. 1–10. DOI: https://doi.org/10.5334/dsj-2019-014

Meseret Y. A. & Stanislava S. (2021). Global challenges of students dropout: A prediction model development using machine learning algorithms on higher education datasets. https://doi.org/10.1051/shsconf/202112909001

Mostafa Z., Abolghasem S., Soo-Mi C. & Ali E. (2021). A practical model for the evaluation of high school student performance based on machine learning. Appl. Sci. 2021, 11, 11534. https://doi.org/10.3390/app112311534

Nicolae-Bogdan S., Rasmus H., Christian I. & Stephen A. (2015). High-school dropout prediction using machine learning. ESANN 2015 proceedings, European Symposium on Artificial Neural Networks, Computational

Intelligence and Machine Learning. Bruges (Belgium), 22-24 April 2015, i6doc.com publ., ISBN 978-287587014-8. http://www.i6doc.com/en/.

Sotiris K., Pintelas P. E. & Christos P. (2003). Preventing student dropout in distance learning using machine learning techniques. https://www.researchgate.net/publication/221018112

Sunbok L. & Jae Y. C. (2019). The machine learning-based dropout early warning system for improving the performance of dropout prediction. Appl. Sci. 2019, 9, 3093; https://doi:10.3390/app9153093

Zachary B. N., Deshendran M. & Sonia B. (2019). Using Bayesian networks and machine learning to predict computer science success. https://doi:10.1007/978-3-030-05813-5_14

EDITOR-IN-CHIEF
Prof. A.I. Mteiye [MALAWI]