FUDMA Journal of Sciences (FJS)
ISSN online: 2616-1370
ISSN print: 2645 - 2944
Vol. 7 No. 6, December, 2023, pp 347 - 353
DOI: https://doi.org/10.33003/fjs-2023-0706-2103

STUDENT DROPOUT PREDICTION USING MACHINE LEARNING

***Osemwegie, E. E. and Amadin, F. I.**

[1]Department of Computer Science, Faculty of Physical Sciences, University of Benin, Nigeria.

*Corresponding authors' email: eric.osemwegie@physci.uniben.edu

ABSTRACT

In a higher education environment, we considered the likelihood of probable dropouts from a first-year undergraduate Computer Science program. In order to achieve this, data from five academic sessions were obtained from the Department of Computer Science, University of Benin, Nigeria. Out of nine hundred and forty seven (947) data obtained, only a total of nine hundred and six (906) was usable after cleaning and preprocessing. Six distinct classifiers including Naive Bayes (NB), Logistic Regression (LR), Support Vector Machine (SVM), Decision Tree (DT), K-Nearest Neighbor (KNN), and Artificial Neural Networks (ANN) were modeled for the prediction of student success and dropouts. The performance six were stated to have performed on average at 90.4%, 98.9%, 98.5%, 97.4%, 96.0% and 97.3% respectively. Although there wasn't much of a performance difference between the DT, SVM, and LR, the LR model was chosen for deployment since it performs better than the other two models in terms of F1_score and Recall.

Keywords: Student, Dropout, Prediction, Machine Learning

INTRODUCTION

Education is the sector of the economy that creates workers needed for the socioeconomic, political, and cultural advancement of any given society (Nwabueze, 2011). The systematic growth or training of the intellect, abilities, or character through instruction or study is known as education. It serves as a tool for promoting national development. It is impossible to overstate the value of education for the growth of both individuals and the country. Government measures, according to some, have raised school gross enrolment ratios over the past 60 years through a variety of educational schemes, but they haven't kept up with gross completions. School dropout is one of the significant concerns that hamper youth's progress in professional settings and it has been a subject of significant concern in most developed and developing countries including Nigeria (Udomah *et al.*,2020). Several factors could be responsible for the dropout of students from the university. Most critical factors influencing university dropout are: students' academic goals, self-evaluation capacity and academic abilities (Robbins *et al*., 2004). Also other factor includes institutional commitment, social support, social involvement, and financial support of the institutions. Irrespective of the factors that might lead to a student dropping out of school, the negative effects would affect the individual, university and the social economy (Nurmalitasari *et al*., 2023).

Finding probable dropout students is a wonderful way to improve retention strategies like help programs, training, or mentoring. Thus, a quantitative evaluation of the chances of failure can be helpful in allocating instructional, psychological, and administrative resources effectively. If captured, data produced within the academic setting can provide amazing insight that could also help identify and manage student dropouts. The study's primary goal is to design a machine-learning model for student dropout prediction. In order to achieve this, academic data of first-year undergraduate Computer Science of the University of Benin between year 2016 to 2020 were considered.

In previous study, Nurdaulet *et al*. (2021) predicted dropout and graduation from an undergraduate computer science program in a higher educational institution. The data used were sourced from the students who started the degree programme in the year 2016 and 2017. 366 participants were left after the data preprocessing and cleaning. Four different binary classifiers were considered namely: NB, SVM, LR and ANN models. The NB was the most accurate for predicting student dropout with a performance of 96%. The research by Real et *al.* (2018) predicted the likelihood of student dropout based on 17 potential predictors. To categorize students, a binary logistic regression model was utilized, and its accuracy was compared to the accuracy of three additional classification methods: One-R, KNN, and Naive Bayes. According to the findings, the binary logistic regression model provided an overall accuracy level equivalent to that of the Naive Bayes approach but superior to that of the One-R or KNN methods. Jay *et al*. (2020) study aimed to identify the underlying factors of dropout students. They predicted the student dropout by testing two classification algorithms, C4.5 and NB, on a data set containing student academic demographic details. Based on the student data gathered, the results showed that the C4.5 model had a higher accuracy rate of 98.9874% in predicting student dropout situations. There is currently a void in the use of machine learning to address dropout in developing countries (like Nigeria), despite the fact that numerous studies have been conducted in this area. This study also aims to fill that gap.

MATERIALS AND METHODS

In this study, a student dropout prediction model based on binary classification leveraging several machine learning classifiers was developed. The fundamental concept is to employ different algorithms for machine learning to a particular data collection for more dropout modeling and prediction with regard to first-year computer science student result. The logistic regression classifier, which was chosen as the best model based on the data set and the models that were compared, was then used to predict whether or not a student would pass. The system chart is shown in figure 1 below. It shows the system chart and procedures utilized in this investigation are shown below:

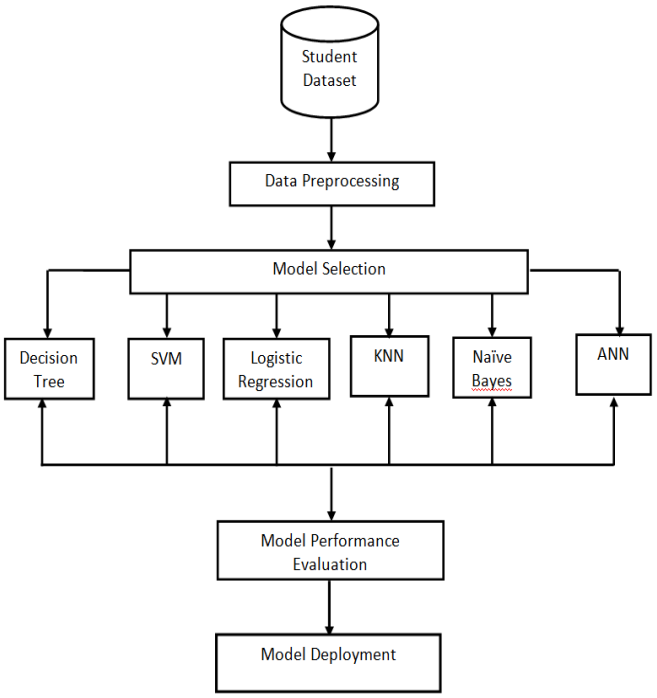

Figure 1: System chart

Dataset Description

This study used real-world data from the University of Benin, Department of Computer Science. The dataset includes first-year students data of five academic sessions from 2016/2017, 2017/2018, 2018/2019, 2019/2020, 2020/2021 sessions. To improve the outcome as well as its applicability, seventeen courses comprising core, electives, and mandatory required to make up a total credit load of 49 credits were chosen. The selected courses all satisfied the criteria for importance as well as a negligible quantity of data was missing. A detailed overview of the dataset is shown in Table 1. The names of the students were not included and their matriculation numbers were not included to ensure the anonymity of students' grades and data confidentiality.

Table 1: Dataset Description

Features	Values
Mat. No.	Char
Names/Gender	String
Year	Date
CHM111	Num/Char
CSC111	Num/Char
…	…
GST	Num/Char
Total Credit Passed	Number
Total Credit Failed	Number
Total Credit Registered	Number

Data Preprocessing

The data collected included extraneous information that were unnecessary for the forecasts and modelling such as multiple column attributes, multiple values on every cell, and unregistered students. Additionally, since unstructured data cannot be utilized as input for the classification model, it must be converted. The following techniques were employed in the investigation to preprocess the data:

Data Cleaning

In other to make the dataset more workable, identifying and fixing data flaws or anomalies such as missing numbers, deviations, and repetitions were taken into account. Unregistered students' data with no values were removed, the student's gender was separated from the student's Full Name column, grade and score columns were reduced to just the grade column. Microsoft Excel was used in carrying out the data cleaning process.

Data Integration

The result of the various sessions was merged into a single large dataset. Combining the data is required for the model classification. Microsoft Excel was used to integrate the datasets of the various academic session into a single dataset.

Data Transformation

Data transformation involves putting the data into an analysis-ready format. The data acquired was unfit for the model and hence required to be transformed to ensure its validity for the modelling. One-hot encoding was used to convert the gender features to take on values 0 for female and 1 for male. Also, to ensure a high precision for the KNN models, the input dataset was standardized. The StandardScaler function was used to implement the standardization of the input data. JupyterLab was used to achieve the data transformation. The dataset was saved as CSV file. Table 2 below shows the sample dataset after the transformation was done.

Table 2: Transformed dataset

	Gender	Year	1-TCE	1-CE	1-GPA	2-TCE	2-CE	2-GPA	TCP	TCF	TCR	GPA	Status
0	M	2016	24	9	2.88	22	8	2.95	43	3	46	2.826	1
1	F	2016	24	9	2.50	22	8	1.41	30	16	46	2.022	0
2	F	2016	24	9	3.38	22	8	2.41	46	0	46	2.761	1
3	M	2016	24	9	4.00	22	8	3.00	46	0	46	3.326	1
4	M	2016	24	9	2.63	22	8	2.64	41	5	46	2.565	1

After the transformation of the dataset, the following variables reflected on the new datasets and were used for the modelling: Gender, Year of admission, 1-TCE : Total Credit enrolled in First Semester, 1-CE: Number of courses enrolled in the first semester, 1-GPA: First semester grade point average, 2-TCE : Total Credit enrolled in Second Semester, 2-CE: Number of courses enrolled in the second semester, 2-GPA: second semester grade point average, TCP: Total Courses Passed, TCF: Total Courses Failed, TCR: Total Courses Registered, GPA: Grade Point Average, and Status (1 means Successful, 0 means Failed).

Exploratory Data Analysis

This phase focused on data exploration in order to spot evident flaws and better comprehend patterns within the data, detect unusual events or outlier, and discover interesting relationships between variables. Plots, charts and graphs were used to explore data and were combined to provide further insights.

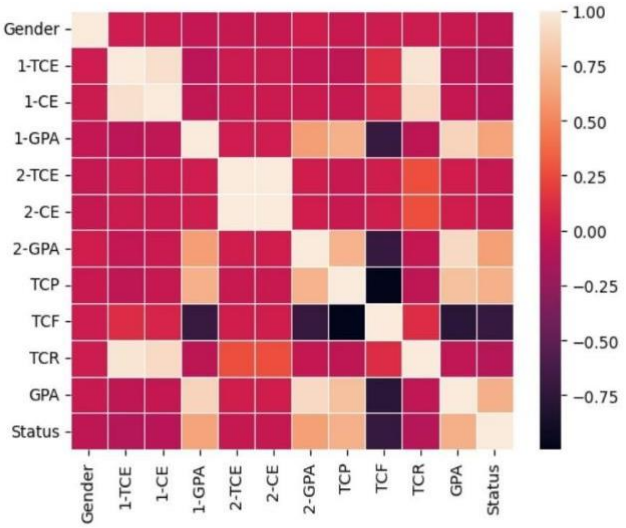

Figure 2: Correlation among the variables

The correlation heatmap shown in Figure 2 depicts the correlation between the variables. Each square represents the correlation of the variables on each axis. We can confirm the correlation between the selected features and the features to remove that are not linked to the dropout analysis by assessing the correlation among the selected features. Figure 3 shows the histogram for each numerical attribute.

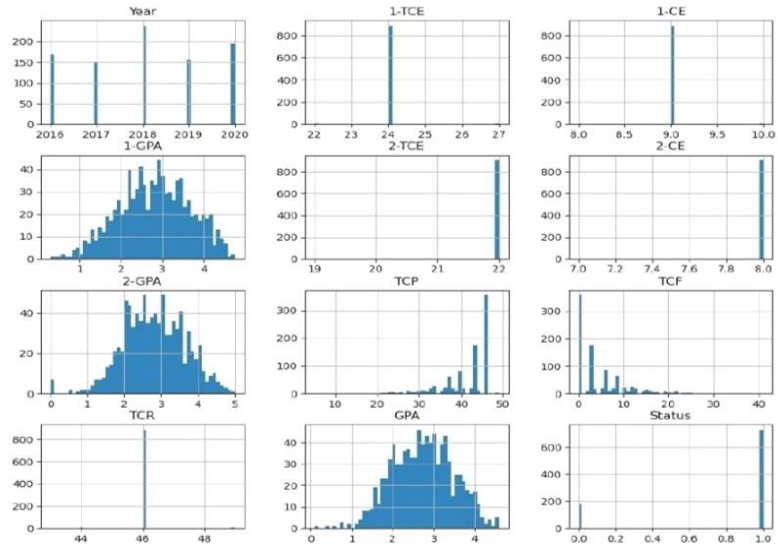

Figure 3: A histogram for each numerical attribute

Model Selection

Different machine-learning models are applied to the dataset in this work. In order to determine whether or not a student would drop out of school, a binary classifier is used to solve the classification problem of student dropout prediction. LR, SVM, DT, KNN, NB, ANN are the machine learning models chosen for this project.

LR:derives a probability result for the dependent variable by estimating the probability of an event occurring given a set of independent factors.

SVM: discriminative machine learning model that model that maximizes the prediction accuracy of a model without overfitting the training data.

DT: non-parametric supervised learning algorithm mostly utilized to solve classification issues, but it may also be applied as a regression model to forecast numerical results

KNN: makes assumptions or classifications about how to group a single data point based on its proximity.

ANN: modeled after the neural network of the human brain and imitates the way that biological neurons communicate with one another.

NB: generative learning algorithm that aims to simulate the distribution of inputs for a certain class or category.

Model Evaluation

The dataset used for the training and testing of the proposed models were split into 70% for training and 30% of the dataset for testing the model. Also, the confusion matrix was used as the performance evaluation tool evaluates the performance of the classification model. The model was evaluated using accuracy, precision, recall, F1_Score.

RESULTS AND DISCUSSION

Table 3: Dataset table in the study

Data Set	Dropout (0)	Successful (1)	Total
Train	127	507	637
Test	50	222	272
Sum	177	729	906

Nine hundred and six of the 947 data points that were collected between 2016 and 2021 are used in the experiments that were conducted. Learning on Decision Tree, KNN, NB, LR, ANN, and SVM were achieved by dividing collected and pretreated data into learning and test datasets in a 7:3 ratio. We apply a tester to the trained model in order to evaluate the accuracy of the prediction. The following are the evaluation findings for every model:

Table 4: Decision Tree Model

	precision	recall	f1-score	support
0	0.915	0.935	0.925	46
1	0.987	0.982	0.984	226
accuracy			0.974	272
macro avg	0.951	0.959	0.955	272
weighted avg	0.975	0.974	0.974	272

The prediction verification using the DT model yielded a considerably high performance with 97.4% accuracy, 91.5% precision, and 93.5% f1-score. The DT model also did quite well in terms of learn rate, with an execution time of 0.0127secs recorded. The execution time for each of the models is shown in Table 10 below.

Table 5: KNN Model

	precision	recall	f1-score	support
0	0.831	0.980	0.899	50
1	0.995	0.955	0.975	222
accuracy			0.975	272
macro avg	0.913	0.967	0.937	272
weighted avg	0.965	0.960	0.961	272

Although not as good as the DT, the prediction verification using the KNN model also produced a high performance. The KNN model had an 89.9% F1-score, 96.0% accuracy, 83.1% precision, and 98.0% recall. With an execution time of 0.0797 seconds, KNN also tends to do better in terms of learn rate.

Table 6: NB Model

	precision	recall	f1-score	support
0	0.681	0.942	0.790	52
1	0.985	0.895	0.938	220
accuracy			0.904	272
macro avg	0.833	0.919	0.864	272
weighted avg	0.927	0.904	0.910	272

The NB model performed well, with a prediction accuracy of 90.4%, however it has the lowest precision of all models. Table 9 compares the performance of all models used in this study. In terms of learn rate, NB surpassed all other models except the DT, with an execution time of 0.0163 seconds.

Table 7: LR Model

	precision	**recall**	**f1-score**	**support**
0	1.000	0.940	0.969	50
1	0.87	1.000	0.993	222
accuracy			0.989	272
macro avg	0.993	0.970	0.981	272
weighted avg	0.989	0.989	0.989	272

With 98.9% accuracy and 100% precision, the prediction verification results employing the LR model demonstrate noticeably better performance. Though it is better than ANN and SVM, the learning rate is not as good as DT, KNN, and NB. In terms of accuracy, precision, and F1-score, LR fared better than any of the models taken into consideration in this study.

Table 8: ANN Model

	precision	**recall**	**f1-score**	**support**
0	0.969	0.886	0.925	35
1	0.973	0.993	0.983	147
accuracy			0.973	182
macro avg	0.971	0.939	0.954	182
weighted avg	0.972	0.973	0.972	182

The ANN model performs with 97.3% accuracy and 96.9% precision, however, it requires a lot of time to learn, making it unsuitable for the prediction system. Although the SVM demonstrates a strong performance of 98.5% accuracy, it is also not suitable for the prediction system because it requires a significant amount of time to learn.

Table 9: Summary of the Performance Evaluation

Prediction Model	**Performance Measure**			
	Accuracy	**Precision**	**Recall**	**F1_Score**
Decision Tree	0.974	0.915	0.935	0.925
K Nearest Neighbor	0.960	0.831	0.980	0.899
Naïve Bayes	0.904	0.681	0.942	0.790
Logistic Regression	0.989	1.000	0.940	0.969
Artificial Neural Network	0.973	0.969	0.886	0.925
SVM	0.985	0.960	0.960	0.960

Table 9 shows the experimental results from the predictive models produced by the DT, KNN, NB, LR, ANN, and SVM. With relative success rates of 98.9%, 98.5%, 97.4% and 97.3% on the assessment dataset, the LR, SVM, DT and ANN classifiers produced the highest accuracy results.

The LR model, however, ran significantly more quickly than the SVM. Prior to tweaking the hyperparameters, the SVM's execution time was 0.0328 seconds. Although the recall value, F1_score and accuracy of the LR were greater than that of the DT, the DT performed better than the LR model in terms of the execution time as observed in this research. The LR model, which performed better than all other models according to the accuracy, recall and f1_score as examined in this study effort, was used for the deployment.

The execution times of the various models are displayed in seconds in Table 10 below.

Table 10: Execution time (in seconds)

Model	**Execution Time**
DT	0.0127
KNN	0.0797
NB	0.0163
LR	0.2310
ANN	2.8650
SVM	2.6700

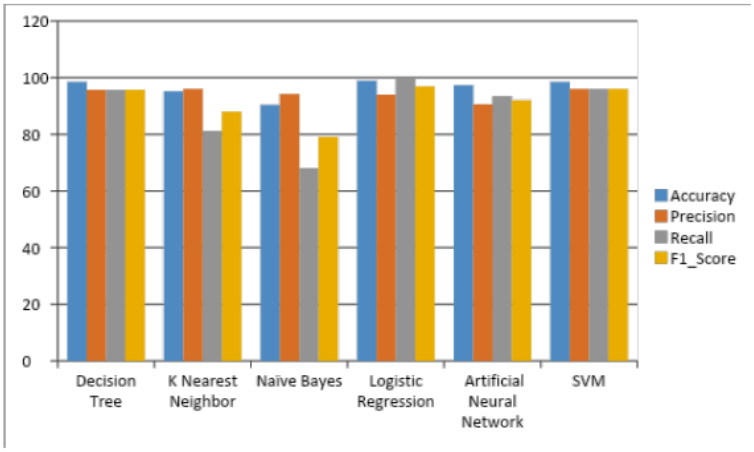

Figure 4: Chart of the various performance evaluation

The test dataset for DT revealed that 222 of the 226 instances of successful students given as 1 and 43 of the 46 instances of non-successful (dropout) students given as 0 were correctly identified. Using the test dataset for KNN, 212 of the 222 cases of successful students were correctly identified, as were 49 of 50 non-successful students. Following the NB test dataset, 49 out of 52 unsuccessful students and 197 out of 220 successful students were accurately classified. The LR test dataset also revealed that 222 of 222 successful students and 49 of 52 unsuccessful students are correctly classified. While the test dataset for SVM demonstrates that 220 of 222 successful students were correctly identified, 48 of 50 non-successful students were correctly classified. The test dataset properly identified 146 of 147 successful students and 31 of 35 unsuccessful students for ANN. Details about the comparison of the accurate prediction are provided in the table 11 below.

Table 11: Shows the comparison of the accurate prediction of the 6 different models

Model	Correct Prediction		Incorrect Prediction	
	Non Successful 0	Successful 1	Non Successful 0	Successful 1
DT	3	222	43	4
KNN	49	212	1	10
NB	49	197	3	23
LR	47	222	3	0
ANN	31	146	4	1
SVM	48	220	2	2

When compared to a related study by Haarika *et al.* (2022) on the use of ML techniques to predict student dropout, the LR model demonstrates a high degree of prediction efficiency. Furthermore, the LR model demonstrated good prediction accuracy with a percentage of 90% and an F1-Score of 0.85 in the study by Ujkani *et al.* (2022), suggesting that the model is excellent in forecasting student dropout.

The LR Model outperformed the other models in terms of accuracy, recall, and F1-score, as Figure 5 illustrates. The model that performed the best after it was the SVM Model, however it required a long time to learn and was not suitable for the prediction system.

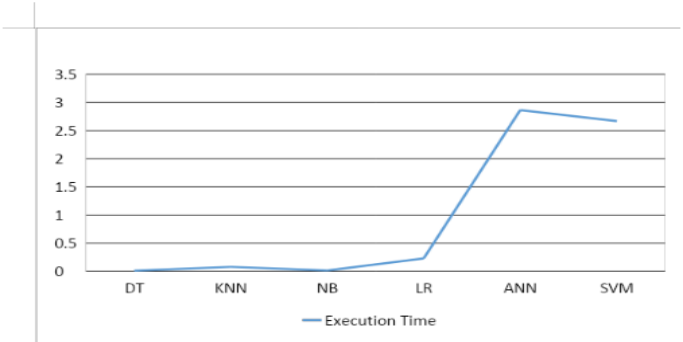

Figure 5: Line graph showing the execution time of the various ML models

Figure 5 depicts the execution times of the various machine learning models, with the DT executing the fastest and the SVM executing the slowest.

CONCLUSION

Early prediction of student dropout can assist academic institutions in providing timely intervention as well as suitable planning and training to improve students' success rate. This

study used a variety of machine learning techniques to predict academic dropout among students. The model was trained and tested using DT, LR, NB, SVM, KNN, and ANN. With the use of the suggested prediction approach, course advisors, organizations, and the university will be able to assess students' performance and put effective interventions in place to raise their academic performance in advance. This study discovered that the Logistic Regression Model outperformed the other models employed in this investigation in predicting student dropouts. To increase accuracy, the proposed model may need to be re-evaluated using additional datasets, perhaps drawn from academic Big Datasets

REFERENCES

Haarika S., Srinivas K. (2022). "Student Dropout Prediction Using Machine Learning Techniques". International Journal of Intelligent Systems and Applications in Engineering ISSN: 2147-6799

Jay S. Gil, Allemar J. P., Ramcis N. V. (2020). "Predicting Students'Dropout Indicators in Public School using Data Mining Approaches". International Journal of Advanced Trends in Computer Science and Engineering ISSN 2278-3091. Available online: http://www.warse.org/IJATCSE/static/pdf/file/ijatcse110912020.pdf
https://doi.org/10.30534/ijatcse/2020/110912020

Nurdaulet S., Alibek O. Yershat Sapazhanov, Shirali K., (2021). "Prediction of Student's Dropout from a University Program". International Conference on Electronics Computer and Computation (ICECCO). DOI:10.1109/ICECCO53203.2021.9663763

Nurmalitasari N., Zalizah A. L., Faizuddin M. N., (2023). "Factors Influencing Dropout Students in Higher Education". Education Research International. 2023.1-13.10.1155/2023/7704142.

Real A. C., Oliveira C. B., Borges J. L., (2018). "Using Academic Performance to Predict College Students Dropout: A Case Study". https://doi.org/10.1590/S1678-4634201844180590

Nwabueze A.I. (2011). "Achieving MDGs through ICTs Usage in Secondary Schools in Nigeria: Developing Global Partnership with Secondary Schools." Germany: Lambert Academic Publishing.

Robbins S. B, Lauver K, Le H, Davis D, Langley R, Carlstrom A. (2004). "Do psychosocial and study skill factors predict college outcomes? A meta-analysis". Psychological Bulletin, *130*(2), 261–288. https://doi.org/10.1037/0033-2909.130.2.261

Udomah N. G., and Dr. Archbong U. I. (2020). "Psychological Factors and Drop out Tendency of Year One Students in Schools of Nursing, Akwa Ibom State, Nigeria". International Journal of Research in Education and Management Science Vol 3 NO 2, United Kingdom

Ujkani B., Minkovska D., Lyudmila Y. S. (2022). "Application of Logistic Regression Technique for Predicting Student Dropout. International Scientific Conference Electronic