Prediction of Student Academic Performance Using a Multi-Regression and Classification-Based Model

Agagu, Modupe
Dept. of Computer Science
Olusegun Agagu University of Science and Technology,
Okitipupa, Ondo State, Nigeria
m.agagu@oaustech.edu.ng

Ikuomola, Aderonke Justina
Dept. of Computer Science
Olusegun Agagu University of Science and Technology,
Okitipupa, Ondo State, Nigeria
aj.ikuomola@oaustech.edu.ng

Obamehinti Adeolu Seun
Dept. of Computer Science
Olusegun Agagu University of Science and Technology,
Okitipupa, Ondo State, Nigeria
as.obamehinti@oaustech.edu.n

ABSTRACT

The Prediction of students' performance is a necessity because it forecasts how well a student is expected to perform during a course of study. Over the years, studies have revealed that student performance has been below average, with one of the main causes being that a thorough prediction of a student's academic potential is typically not done. To choose the best model for predicting and categorizing academic achievement, a multi-regression analysis is performed using machine learning models such as Decision Tree, K-Nearest Neighbor, Random Forest, Logistic Regression, and Support Vector Machine. Furthermore, the result shows that Random Forest is the best-performing classifier in this study, with an F1 score and accuracy of 94.9%, as well as the best-performing regression model, with a Mean Absolute Error (MAE) of 0.3711 in predicting academic success.

Keywords: Academic Performance, Classification Model, Machine Learning, Prediction, Regression Model.

1. INTRODUCTION

Machine learning algorithms have gained popularity in the education sector in recent years as a means of forecasting student success. Educational data mining is the practice of using data mining techniques to educational database data to forecast student performance. The goal of educational data mining is to enhance learning by analyzing learner behaviour through the application of machine learning algorithms and data mining techniques to data received from educational devices El et al., 2019. The ability to teach software or computers to think like humans and to learn more autonomously over time by providing them with data and knowledge in the form of real-world observations and interactions is known as machine learning (ML) Nti et al., 2022, Orji and Vassileva, 2022. Algorithms for machine learning have proven to be an effective technique for forecasting student performance. Demographic data can be used to identify students who are at risk. Predicting the learning process and analysing student performance are regarded as notable tasks in the field of educational data mining (EDM), and the application of data mining algorithms on datasets could help all participants in educational institutions Feng et al., 2022. Predictive factors included the student's gender, marital status, health status, race/ethnicity, parental education level, type of food, and amount of time spent preparing it. We developed and used five supervised machine learning (ML) models to forecast learning performance. Since these methods are frequently employed to solve regression and classification problems, they were employed in the construction of the models. These techniques include Random Forest (RF), K-Nearest Neighbours (KNN), Decision Tree (DT), Logistic Regression (LR), and Support Vector Machine (SVM). Using the test dataset, the best-performing regressor/classifier model for the academic performance of students was found. The Random Forest Regressor model is one of these algorithms that has drawn a lot of interest because of its capacity to manage big datasets and nonlinear correlations between variables. This study contributes to the literature in several ways. First, we developed and tested five predictive models using data mining classification techniques to predict student academic performance and in addition, we identified the best of five

classification techniques to predict student performance in terms of accuracy, precision, recall, and F1-Measure. The rest of this paper is organized as follows. Section 2 is the related work. Section 3 describes the methodology of the study. Section 4 explains the results and discussion. Section 5 is the conclusion. This study makes multiple contributions to the body of literature in various was. In order to forecast academic success of students, we first created and tested five predictive models using data mining classification approaches. We also determined which of the five classification strategies performed best in terms of accuracy, precision, recall, and F1-Measure in predicting student performance. The remainder of the paper is structured as follows. Section 2 contains the relevant work. Section 3 outlines the study's methodology. The discussion and results are explained in Section 4 and the conclusion is detailed in Section 5.

2. RELATED WORK

All material on each page should fit within a rectangle on the page, beginning 2.54 cm (1") from the top of the page and ending with 2.54 cm (1") from the bottom. The right and left margins should be 1.9 cm (.75”). The text should be in single column various data mining approaches, such as classification and regression, have been used to develop a model for predicting student performance. When the outcome variables are definite (or discrete), the classification technique is used, but the regression technique is used when the outcome variables are numerical (or continuous) according to the author in El et al., 2019.

The authors in Obsie and Adem, 2018 used Neural Network (NN), Logistic Regression (LR), and Support Vector Regression (SVR) to predict student academic achievement. The dataset for this study was obtained from Hawassa University Student Information System (SIS) for the School of Computer Science. The dataset included 134 undergraduate degree students who graduated from the university in 2015, 2016 and 2017, with 52 (38.81%), 39 (29.10%), and 43 (32.09%) students respectively. The information gathered was organized in a Microsoft Excel spreadsheet. The prediction accuracy for NN was 0.9763, SVR was 0.9805 and LR was 0.9805. Overall, the NN technique produced the least accurate prediction result for all cases. The study demonstrated that data mining techniques can be utilized to forecast students' academic achievement at higher education institutions. The researcher in Hasan et al., 2019 studied the issues concerning low performing student success to avoid a predicted negative outcome. The study attempts to forecast the final grade for the course over three semesters. KNN and DT algorithms were the main strategies used. The results presented that the prediction correctness using DT was 94.44% and the prediction accuracy using KNN was 89.74% in forecasting student final exam performance. Attendance, assignment, and presentation were all avoided by the authors. The author discovered that the final exam is influenced by the midterm and class test grades during the semester. With respect to forecasting the student's final exam outcome, this paper is similar to our model. However, we added more courses and took the student's demography and marital status into account. Using a data mining technique, the authors of Ikuomola and Nwanze, 2020 created a model that forecasts the academic success of students. The university student database, data pre-processing, and the data mining method are the three primary components of their concept and they worked with a dataset from a university. To categorize the data pieces into distinct groups, classification techniques like J.48, Naïve Bayes Bayesian Network, JRip, OneR, and PART were employed. When it comes to forecasting patterns and informing crucial courses that can impact each final student's CGPA, Naïve Bayes, Bayesian Network, and PART perform better than average, according to the evaluation findings of the classification models created using the chosen data mining. The authors' study in Mengash, 2020 concentrated on how to employ data mining techniques to forecast applicants' academic achievement at the university to assist universities with their admissions decision-making. The suggested methodology was validated using a data set of 2,039 students enrolled at a Saudi public university's Computer Science and Information College between 2016 and

2019. The findings show that, depending on a few pre-admission parameters, applicants' early university performance can be anticipated before admission (High School Grade Average, Scholastic Achievement Admission Test Score, and General Aptitude Test Score). The outcomes also demonstrate that the pre-admission criterion that most closely forecasts future student success is the score on the Scholastic Achievement Admission Test. Therefore, this score should be assigned more weight in admissions systems. Also, it was found that the Artificial Neural Network technique has an accuracy rate above 79%, making it superior to other classification techniques considered such as Decision Trees, Support Vector Machines, and Naïve Bayes. A method to increase the student's final grade prediction model's accuracy for a given course was provided by the authors in Jishan et al., 2025. Comparing their study's outcomes, they found that only two of the models they used had the best accuracy, which is almost 75%. Neural Network and Naive Bayes are the models. In the study published in Al-Alawi et al., 2023, supervised machine learning techniques were used to investigate the variables that adversely affected college probationary students' academic performance (underperforming students). A public institution in Oman offered a sample of N = 6514 college students across 11 years (2009 to 2019) for which they employed the Knowledge Discovery in Databases (KDD) methodology. The best features were chosen using the information gain (InfoGain) algorithm, and ensemble approaches were then used to compare the accuracy with more reliable algorithms including logit boost, vote, and bagging. The algorithms underwent validation by 10-fold cross-validation after being assessed using performance assessment measures like accuracy, precision, recall, F-measure, and ROC curve. The study found that length of study at the university and prior success in secondary school are the primary factors influencing students' academic progress. The study also showed that whether a student was placed on probation depended greatly on their academic specialization, cohort, gender, and predicted graduation year.

3. METHODOLOGY

To determine the impact of some learning attributes on the academic performance of students, we used well-known machine learning methods, which are described in this section. We employed questionnaires and student data from the university database to collect study data, employing a quantitative research approach. The ideal method for gathering primary data for research projects based on surveys, experiments, and observation is to use questionnaires. Thus, data on participants' views, attitudes, feelings, and expected behaviour are provided via the survey method to researchers. The demographic information and qualities utilized in the questionnaire design are what make up the model. We divided our dataset into training and test sets, preprocessed the data prior the analysis, and then created five supervised machine learning classifiers and regressors to forecast students' academic success. We trained and evaluated our regression models for academic performance prediction and classification models prediction using the metrics stated in section 4. The test sets was used to determine the performance of the models and compared the performance of the models built to determine the best-performing regressor classifier. Preprocessing of the dataset and prediction experiment in this study were performed using Python and the sci-kit-learn library. The models were implemented and compared including the Random Forest (RF), Logistic Regression (LR), Support Vector Machine (SVM), Decision Tree (DT), and K-Nearest Neighbors (KNN), for both the regression and classification problems of student academics. The dataset used in the model building was split into training and test sets in the ratio of 80%:20% for the regression experiment and 80%:20% for the classification.

3.1 Model Development Dataflow

The ML model for predicting students' academic performance includes several steps: data collection, data pre-processing, splitting dataset, and regression. The flowchart of the prediction model is shown in Figure 1 below.

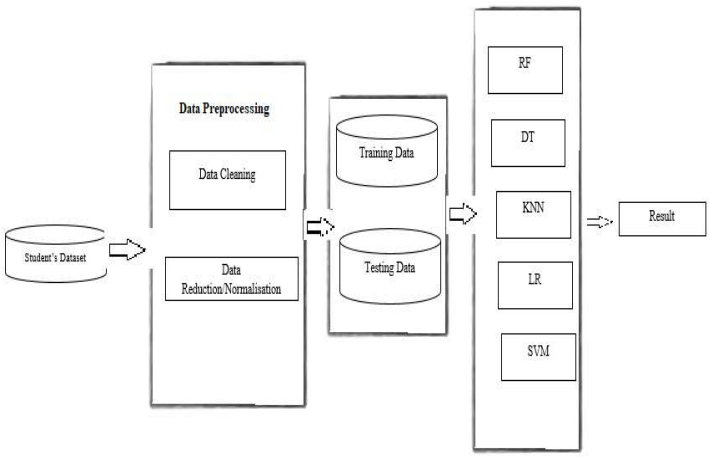

Figure 1: Flow Diagram of Model Development

3.1.1 Data Collection

The data of final year students of the Department of Computer Science, Olusegun Agagu University of Science and Technology, Okitipupa which contains the academic record of the students were collected. A total number of 124 final-year students were used. The students' information was collected from the school's database management system. The dataset used contains academic and demographic data about the students. The data were collected through a questionnaire, including:

i. Marital status: depicts the students' marital engagement. It states whether the student is maritally engaged or not and if they have offspring or not.
ii. Health status: depicts the health-related matters of the student, in the case of a student suffering from any illnesses or not.
iii. Home Address (City): states the town where the student resides.
iv. Food type: give information about the type of diet students take.
v. Parent level of education: this talks about the parent's education either literate or illiterate
vi. Ethnicity
vii. Gender: this includes whether the student is male or female.

3.1.2 Data Preprocessing

One crucial phase in the data mining process is pre-processing. The information collected from the university database is unreliable due to noise and missing numbers. Transforming the data into a format that the algorithms can use is the goal of data pre-processing. The dataset has undergone three primary preprocessing steps: feature encoding, feature scaling, and data cleaning. Microsoft Excel and the Python programming language were used to carry out the pre-processing.

3.1.2.1 Data Cleaning

Real-world data are usually unstructured and noisy. The data cleaning procedure seeks to address irregularities in the data by filling in missing values and reducing noise. A number of students failed to complete the questionnaire's demographic sections, which resulted in missing values in the demographic data columns. After deleting the rows with missing values, the dataset—which originally contained 637 instances—now has 621 instances and 124 students. Considering that only 2.5 percent of the observations were removed, there are most likely no significant distortions.

3.1.2.2 Features Encoding

All inputs and outputs in machine learning models must be numerical variables. As a result, feature encoding—the process of encoding categorical data—must be done before the data can be used in the model. Label Encoder was used in our dataset to convert features into numerical features.

3.1.2.3 Features Scaling

With this technique, a group of independent variables or data features is normalized by scaling the data to lie between a narrower range, like 0.0 and 1.0. This could hasten the training process and lower the error rate of algorithms. We applied the Standard Scaler method to our dataset.

3.1.3 Splitting Dataset

The dataset is split into training and testing datasets. The training dataset is used to develop the model and the testing dataset is used to evaluate the model. It was 80% for training data and 20% for testing data.

4. EVALUATION METRICS

The algorithms; SVM, RF, KNN, DT, and LR algorithms were used to create ML models. The performance metrics are mean absolute error (MAE), accuracy, recall, precision and F1- score.

4.1 Mean absolute error (MAE)

The MAE is common performance metrics often commonly used for evaluation. It is used to estimate the prediction error of the model. The MAE measures the average magnitude of the errors in a set of predictions. MAE equation can be seen in Equation (1).

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \quad (1)$$

Where y_i represents the actual value, $\hat{y}_i$ represents the predicted value of y_i, and N represents the number of instances.

4.2 Accuracy

Accuracy score is a common evaluation metric used in classification tasks to measure the proportion of correctly predicted instances out of the total instances in a dataset. It is calculated as

$$\text{Accuracy} = \frac{Number\ of\ Correct\ Predictions}{Total\ Number\ of\ Predictions} \quad (2)$$

4.3 Recall

This is defined as the percentage ratio of correctly classified instances of a given class.

$$(TPR) = \frac{TP}{(TP+FN)} \quad (3)$$

4.4 Precision

This is defined as a proportion of instances that are true of a class divided by the total instances classified as that class.

$$\text{Precision} = TP/(TP+FP) \quad (4)$$

4.5 F1-Score

The F1-score is a widely used evaluation metric in classification tasks that provides a balance between precision and recall. It is a way to assess a model's accuracy while considering both false positives and false negatives.

$$\text{F1-Score} = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (5)$$

5. RESULT AND DISCUSSION

To analyze the performance of regression models, many evaluation metrics have been applied. The performance measures how close the expected results are to the actual values. In research studies, measures such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and others are widely utilized. The accuracy of the regression models in this study was compared by estimating the individual MAE. The average models' prediction errors can be readily analyzed using this measure (as the average difference between actual and anticipated values). When the MAE is low, the model's accuracy improves. We also employed four commonly used evaluation metrics for the classification problem: accuracy, F1-score, precision, and recall. Using the test dataset, the performance of the regression and classification models was compared using the above evaluation measures. Tables 1 show the results of regression performance for students' academic performance prediction using the five different machine learning models and the mean absolute error for each regression model. The RF model performed better than the other models, while the KNN model gave the least accurate result. The accuracy of classifier models is shown in Table 2. RF has the highest overall score in terms of accuracy, precision, recall, and F1 among the classifiers, followed by DT. LR produced the least accurate performance result. These findings indicate that the attributes utilized in this study are adequate for predicting students' academic performance and academic success.

Table 1: Regressor Performance for Students' Academic Performance Prediction

Regressors	Mean absolute Error
i. Logistic Regression	0.3912
ii. Random Forest	0.3711
iii. Support Vector Machine (SVM)	0.4120
iv. Decision Tree (DT)	0.3770
v. K-Nearest Neighbors (KNN)	0.432

Table 2: Classifiers Performance Evaluation in percentages

Metrics	RF	LR	SVM	DT	KNN
Accuracy	94.9	58.9	59.9	88.0	68.5
Precision	94.9	56.9	63.2	89.2	68.6
Recall	95.0	57.0	61.1	88.5	68.7
	94.9	56.8	58.7	88.0	68.5

6. CONCLUSION

The capacity of certain features to predict students' academic performance and study habits was investigated using the supervised machine learning (ML) approach. To ascertain the overall influence of the variables on research method and success, five ML regression and classification models were specifically applied, and their performances were compared. This study produced useful models that may be used in a variety of higher education courses to predict students' academic achievement based on general traits. Our models' outcomes showed that the characteristics had good accuracy values for predicting academic performance. In this study, the best-performing classifier has an F1 score of 94.9%, accuracy, and precision, while the best-performing regressor has an MAE of 0.3711 in predicting academic success. The findings suggest that meeting the needs of students through the creation of models that forecast their performance can enhance both the learning characteristics and the performance outcomes of students throughout their course of study. If implemented into educational systems, the model examined in this study will aid in accurately projecting and predicting students' learning development. Educational administrators can use the models developed in this study to assess students' performance and those who are at danger of dropping out of a course or pursuing further education, and then provide the appropriate support and interventions.

7. REFERENCES

[1] El Aissaoui, O., El Alami El Madani, Y., Oughdir, L., Dakkak, A., and El Allioui, Y. (2019). A multiple linear regression-based approach to predict student performance. In *International Conference on Advanced Intelligent Systems for Sustainable Development*, 9-23. Cham: Springer International Publishing.

[2] Nti, I. K., Akyeramfo-Sam, S., Bediako-Kyeremeh, B., and Agyemang, S. (2022). Prediction of social media effects on students' academic performance using Machine Learning Algorithms (MLAs). *Journal of Computers in Education*, *9*(2), 195-223.

[3] Orji, F. A., and Vassileva, J. (2022). Machine Learning Approach for Predicting Students Academic Performance and Study Strategies based on their Motivation. *arXiv preprint arXiv:2210.08186*.

[4] Feng, G., Fan, M., & Chen, Y. (2022). Analysis and prediction of students' academic performance based on educational data mining. *IEEE Access*, *10*, 19558-19571.

[5] Obsie, E., Y. and Adem , S. A. (2018). Prediction of Student Academic Performance using Neural Network, Linear Regression and Support Vector Regression: A Case Study, *International Journal of Computer Applications,* 180 (40), 39-47.

[6] Hasan, H.R.; Rabby, A.S.A.; Islam, M.T., and Hossain, S.A. (2019). Machine Learning Algorithm for Student's Performance Prediction. In Proceedings of the 2019 10th *International Conference on Computing, Communication and Networking Technologies (ICCCNT), Kanpur, India*, pp. 1–7.

[7] Ikuomola A.J and Nwanze M.N. (2020). Predicting Student Academic Performance using Data Mining Techniques, *Journal of Behavioral Informatics, Digital and Development Research*, 6(2), 45-56.

[8] Mengash, H. A. (2020). Using data mining techniques to predict student performance to support decision making in university admission systems. *Ieee Access*, *8*, 55462-55470.

[9] Jishan, S. T., Rashu, R. I., Haque, N., & Rahman, R. M. (2015). Improving accuracy of students' final grade prediction model using optimal equal width binning and synthetic minority over-sampling technique. *Decision Analytics*, *2*, 1-25.

[10] Al-Alawi, L., Al Shaqsi, J., Tarhini, A., & Al-Busaidi, A. S. (2023). Using machine learning to predict factors affecting academic performance: the case of college students on academic probation. *Education and Information Technologies*, *28*(10), 12407-12432.