

Predictive Model for Identification and Analysis of Factors Impacting Students Academic Performance Using Machine Learning Algorithms

Ibrahim H. Ibrahim¹, Etemi J. Garba², Uwaisu A. Umar³ and Adedeji A. Adejumo⁴
^{1,2,3,4} Department of Computer Science, Modibbo Adama University, Yola, Adamawa. Nigeria
ibrahimharuna@mau.edu.ng¹, e.j.garba@mau.edu.ng², uwaisu.umar@mau.edu.ng³,
adedejijadejumo@mau.edu.ng⁴

Abstract

Background: Education is a vital component of both societal and individual growth. To create effective learning environments, it is essential to understand the factors that influence student achievement. However, the application of machine learning algorithms can drive positive change, providing higher education institutions with effective solutions to their challenges. **Aim:** This research develops a predictive model to identify and analyze the factors affecting student academic performance, aiming to provide institutional administrators and lecturers with a better understanding of the key factors impacting student success. **Method:** A comprehensive dataset was collected from undergraduate students at Modibbo Adama University (MAU), Yola, comprising student-related, home-related, lecturer-related, and institution-related factors. Model development was carried out using Python in Google Colab, a cloud-based Jupyter notebook environment. Classification algorithms such as K-Nearest Neighbors (KNN), Decision Tree (DT), Gradient Boosting Method (GBM), and Random Forest (RF) were applied to predict student academic performance. Four evaluation metrics namely; accuracy, precision, recall, and f1-score were used to analyze the models. **Results:** The Random Forest model outperformed the other machine learning models, achieving an overall accuracy of 95%. For predicting low-performing students, the model achieved a precision of 0.9677, recall of 0.9474, and f1-score of 0.9574. For high-performing students, the precision was 0.9324, recall was 0.9583, and f1-score was 0.9452. These strong performance metrics across both low and high performing student groups demonstrate the effectiveness of the Random Forest model in accurately predicting student academic performance based on the factors considered in the study. Feature importance analysis identified lecture attendance, sponsor support, quality of facilities, and lecturer clarification as the most influential factors on student performance. Other features, such as accommodation, employment status, and program preference, were found to have a low impact. These findings emphasize the importance of considering a comprehensive set of student, lecturer, institution, and home-related factors to sustain a conducive learning environment and enhance educational practices.

Keywords: Predictive Model, Student Performance, Machine Learning, Random Forest, Feature Importance

1. Introduction

Poor academic performance is a significant concern not only for universities in Nigeria but also for universities abroad. The importance of education to the individual and to the nation cannot be underestimated, because education is the foundation of national growth and development. It is unfortunate, that the quality, level and standard of education in Nigeria have drop drastically and it's still dropping by the day, because students do not have time for studies they only read during examination period to pass. To gain knowledge is out of the question. Their minds are preoccupied on how to fit in, with reference groups in order to feel accepted in the preconceived social environment (Yusuf *et al.*, 2020). Negative socialization affects students' performance, the educational institutions and the entire country as a whole.

The consequences of poor academic performance by undergraduate students can be long-term which is often exhibited as low self-esteem, fear of failure and anxiety (Nsiah, 2017). Students with poor academic grades often feel frustrated and may either resort to dropping out of learning institutions or face prolonged struggles, risking an extended stay in school. Poor academic performance also has its effects on

educational institutions and the society; for institutions, poor academic performance of students curtails the proper execution of educational operations and it reduces the amount of available manpower in different fields (Al-Zoubi & Younes, 2015).

Universities always face the challenge of providing evidence of their effectiveness to the society, government and international higher education. The top management of any university is concerned about the performance of their student to show their usefulness for the international community. In many countries in the world, the rate of graduated students, for example, has served as an indicator of the usefulness of the universities. Academic performance or achievements can be defined in terms of the performance of examinations and tests (Islam & Tasnim, 2021).

Higher education is an important factor for the development of a country. Universities take a significant role to produce highly qualified graduates who become the source of manpower for the development of a country's economy. Students are most essential asset for any educational institute. Students' academic performance plays an important role in producing the best quality graduates who will become great leader and manpower for the country thus responsible for the country's economic and social development. Students who have access to quality educational programs perform better and successful in their lives (Rahman, 2021). Students' Academic Performance is an important metric in determining the status of students in any academic institution. It allows the instructors and other education managers to get an accurate evaluation of the students in different courses in a particular semester and also serve as an indicator to the students to review their strategies for better performance in the subsequent semesters (Ahmed *et al.*, 2019).

The sacred goal of every academic institution is to entrench learning and learning cannot be said to have taken place except that the students are correctly evaluated. This evaluation which used to be in the form of class tests, assignments, practical/laboratory works, and semester examinations is considered to be holistically done to evolve the academic status of each of the students in a class in different courses (Solomon & Agrawal, 2018). Thus the performance of the student is, one of the crucial aspect of every educational institution. Unfortunately in Nigeria, there were no serious programmes or policies put in place to promote academic excellence especially in tertiary institutions (Baffa & Nasiru, 2022). Similarly, increased number of students seeking admissions into tertiary institutions put more pressure on utilities, unsanitary condition of some halls of residence and classes, breakdown of sewage disposal system, overcrowding in classes and inadequate learning facilities among others. These are always being regarded as contributory factors to declining academic standard thereby contributing to poor academic performance among students in tertiary institutions of higher learning in Nigeria (Anas *et al.*, 2023).

Academic achievement is one of the major factors in hiring employees by employers. It says especially for the fresh graduates. Thus, students put the greatest effort in their education to obtain good grades for their future career opportunities and fulfil the employer's demand. Hence, it is a need process to examine the root causes for poor and good academic performance of the students (Olufemi *et al.*, 2018). Education is one of the commanding aspects that leads to inclusive growth and improvement of the individuals, community and nation as a whole (Gonfa, 2020). Thus, good education and academic performance are an interested goal that is considered not only by educational institutions, but also by individuals, society and political decision-makers as it refers to the development of a country.

The effectiveness of AI and machine learning in advancing global education is demonstrated by their capacity to create adaptive learning platforms tailored to the individual needs of students, which greatly improves learning outcomes and supports academic achievement. AI-powered analytics and predictive modeling enable educators to identify at-risk students early and apply targeted interventions, ultimately boosting retention rates and helping more students achieve their full potential. However, in developing countries like Nigeria, the full potential of AI and machine learning for personalization, retention, and resource optimization has yet to be realized (Adenubi & Samuel, 2024).

In light of the above point, there is a need to look deeper into the underlying aspects that contribute to academic performance in order to solve the difficulties that students encounter in achieving academic success. This research intends to provide a thorough understanding of these aspects and construct a

predictive model that can precisely identify and analyze them by utilizing machine learning algorithm. By using machine learning techniques, it is possible to examine huge datasets quickly and gain insightful information. The findings of this study may help educational policy makers, institution administrators and lecturers understand the key factors that influence students' academic performance. In order to improve student result and promote academic success, customized interventions and support systems can be put in place by developing a greater understanding of these factors.

2. Related Works

The academic performance of students can be greatly influenced by the choice of feature selection algorithm used. By classifying students and making predictions about their performance, valuable insights can be provided to local ministries of education, enabling them to take proactive actions based on anticipated student outcomes. Additionally, lecturers can benefit from this approach by effectively evaluating students and focusing on areas that require improvement. Early recognition of academic performance using predictive models allows educational institutions to offer timely support and appropriate recommendations to students. Identifying underperforming students at the beginning of the academic year is particularly important as it enables the provision of additional learning resources and interventions to help them succeed (Rajendran *et al.*, 2022).

According to Ofori *et al.* (2020) the utilization of machine learning techniques to predict students' performance based on their background and in-term performance has proven to be a valuable tool for anticipating both poor and good performance across different levels of education. Early prediction of students' performance is beneficial in taking proactive measures to improve learning outcomes. Predicting academic performance is significant as it contributes to increasing graduation rates by providing appropriate guidance to students, informing changes in academic policies, guiding instructional practices, assessing the efficiency and effectiveness of learning, offering meaningful feedback to teachers and learners, and modifying learning environments. Achieving a high level of prediction accuracy in students' performance is crucial for identifying low-performing students at the beginning of the learning process. However, to accomplish these objectives, a large volume of student data needs to be analyzed and predicted using various machine learning models.

Al-Alawi *et al.* (2023) conducted a study using supervised machine learning algorithms to investigate factors that negatively affected academic performance among college students on probation. The study utilized a dataset of 6,514 college students over 11 years from a major public university in Oman, employing the Knowledge Discovery in Databases (KDD) methodology. The Information Gain (InfoGain) algorithm was used for feature selection, and ensemble methods were employed to compare the accuracy of different algorithms. The study identified study duration and previous performance in secondary school as the main factors impacting academic achievement. Gender, estimated graduation year, cohort, and academic specialization also played a significant role in determining probation status. The J48 algorithm was found to be the most effective for prediction. The findings have theoretical and practical implications for understanding and addressing underperforming students.

Gupta *et al.* (2023) implemented a project to analyze and categorize student performance in a university using processing techniques. The proposed Student Performance Analysis System (SPAS) aimed to overcome privacy settings that limited lecturers' access to student records at UNIMAS. The system tracked students' results and employed data mining techniques, particularly classification, to predict students likely to perform poorly in their courses. SPAS's main advantage is that it helps lecturers assess students' performance, enabling them to identify students who are likely to fail a course using this technique. The project conducted experiments using a dataset from an online learning environment to predict student performance. Three classifiers, namely Random Forest, Support Vector Machines (SVM), and Decision Tree, were used to build predictive models. The classifiers were evaluated using the 10-fold cross-validation method, dividing the dataset into 10 subsets for training and testing. The results showed promising accuracy between 81% and 85% for the implemented classifiers, with an overall accuracy of 84% for random forest, 77% for decision tree, 89% for Extra boost, 36% for AdaBoost, and 70% for KNN. These findings

demonstrate the potential of using data mining and machine learning techniques to predict student performance and improve educational outcomes.

Ekubo and Esiefarienrhe (2022) developed a model to assess the performance of different machine-learning techniques in predicting low academic performance among Nigerian tertiary students. The research utilized data from 2,348 low-performing students at Niger Delta University in Bayelsa State. The study employed the cross-industry standard process for data mining (CRISP-DM) methodology and the Waikato Environment for Knowledge Analysis (WEKA) tool for modeling. Five machine-learning classifier algorithms, including J48 decision tree, logistic regression (LR), multilayer perceptron (MLP), naïve Bayes (NB), and sequential minimal optimization (SMO), were tested. The MLP algorithm was found to be the most effective classifier for the dataset. Additionally, the study developed a predictive software application using PHP and Python, which achieved an accuracy of 98% when implementing the MLP model.

Similarly Alhazmi and Sheneamer (2023) presented an analysis of student performance using data mining methods was presented. Clustering and classification techniques were employed to explore the impact of various factors on students' GPA at an early stage. The clustering technique utilized dimensionality reduction with the T-SNE algorithm to examine the relationship between factors such as admission scores, first-level courses, academic achievement tests (AAT), and general aptitude tests (GAT) with GPA. The classification technique involved experiments with different machine learning models to predict student performance using features such as course grades and admission test scores. The results indicated that early identification of students at risk can help educational systems mitigate the risk of failure. The study also compared the performance of different classifiers, with Random Forest, SVM, and Gaussian Naive Bayes (GNB) showing superior results. The use of machine learning algorithms combined with data dimensionality reduction by T-SNE provided insights into algorithm efficiency.

Okereke *et al.* (2020) undertook a study with the aim of analyzing students' record data and utilizing it to train decision tree classifiers for predictive purposes. The authors demonstrated that the choice of classifier does not solely determine prediction accuracy; rather, it is influenced by the characteristics of the datasets. Rapid Miner was employed to preprocess, train, and test the datasets. The results showed variations in accuracy when using the decision tree algorithm to train different datasets, reflecting differences in students' aptitude for specific courses. For instance, the decision tree analysis achieved a high accuracy of 96.7% in predicting results for the COS101 course, accurately identifying 95 out of 103 students. Conversely, for the STA172 course, the same algorithm yielded a lower accuracy of 71.91%, correctly predicting only 72 out of 103 students. The decision tree was constructed using the gain ratio attribute selection method, with the attribute exhibiting the highest information gain serving as the classification node.

Ojajuni *et al.* (2021) carryout a study aiming to demonstrate how machine learning can support educators in making informed decisions and predicting student academic performance. Multiple machine learning classification models were utilized, resulting in accuracy rates ranging from 33% to 98% and cross-validation rates ranging from 30% to 37%. The XGBoost Model emerged as the most suitable machine learning model, achieving an accuracy of 97.12% and a cross-validation rate of 35.67%. The study identified several significant features that impact student academic performance, including the number of school absences, current health status, socializing with friends, free time after school, and the quality of family relationships. The research concludes that these findings can assist educators in identifying learning gaps and detecting underachieving students early, thereby enabling informed decision-making and improving student academic success and the learning process.

In a very recent study, Anusha *et al.* (2019) applied a different machine learning algorithms i.e. k nearest neighbor (k-NN), Decision Tree, Random Forest and a conventional linear regression model to predict undergraduate computer science students' performances. A set of historical data for a total of four years which constitutes theory marks, term test marks, viva as well as practical marks for each student was used to predict the students' GPA. It was demonstrated from the results that the conventional linear regression was able to provide a good prediction of the student's GPA with an overall R^2 of 94%. The

authors concluded that the developed model could help the students to be fully engaged towards improving their academic performances by virtue of their historical record of performances.

Yadav *et al.* (2018) conducted a research using data mining techniques to predict academic performance based on students' emotional skills, socio-economic factors, and previous academic performance. The study employed descriptive and predictive data mining tasks, utilizing algorithms such as J48 and Random Tree. The research aimed to predict the third-semester performance of graduation students, considering that many students drop out after the first year and require time to adapt to the academic environment. The J48 algorithm achieved an accuracy of 64.94% for cross-validation and 65% for split testing. The study found that the J48 classifier had the best performance among the tested classifiers, followed by the K-NN classifier and the rule learner (JRip). However, all test classifiers had an overall accuracy below 70%, indicating that the predictions were not highly reliable. This research emphasizes the impact of emotional skills and other factors on students' academic performance, contributing to the field of predictive modeling in education.

Anas *et al.* (2023) initiated an investigation exploring the factors and consequences associated with poor academic performance among students at Gombe State University in Nigeria. The research aimed to address the lack of research in this area, particularly in the northeast region of the country. Data was collected from 584 respondents using a questionnaire, and descriptive methods were employed for analysis. The findings revealed several common factors contributing to poor academic performance, such as family stress, poor economic conditions, lack of proper orientation programs, students' lack of commitment to academic activities, distractions from social activities, overcrowding in classrooms, lack of interest in the chosen course, inadequate facilities and teaching personnel, and the type of secondary school attended. Consequences of poor academic performance included low semester results, examination malpractice, and instances of probation and dismissal. The study recommends that parents/guardians provide support to students by relieving them of family responsibilities during school sessions, and emphasizes the need for the university management to provide more facilities to reduce overcrowding and ensure adequate teaching resources.

In their study, Hasan *et al.* (2019) addressed the issue of low student achievement to prevent anticipated poor outcomes. The objective of the research was to forecast the final marks of courses across three semesters. The primary techniques employed were the K-Nearest Neighbors (KNN) and Decision Tree (DT) algorithms. The results demonstrated that DT achieved a prediction accuracy of 94.44%, while KNN achieved 89.74% in forecasting student performance on final exams. The authors excluded certain fundamental attributes such as attendance, assignment, and presentation. They concluded that the final exam grade was dependent on the marks obtained in the midterm and class tests within a given semester.

The research performed by Oreški and Zamuda (2022) applies the CRISP-DM standard and the decision tree algorithm to develop predictive models for students' academic outcomes based on a Croatian dataset. The research focuses on predicting students' performance and grade point average using machine learning techniques. The results demonstrate that the decision tree modeling approach achieves high accuracy, reliability, and interpretability in predicting students' academic performance. The study suggests that this approach holds promise for automated student success prediction, which can be utilized to improve students' learning experiences, provide personalized recommendations, inform professors about important determinants of academic success, assist higher education institutions in adjusting strategies, and automate the adaptation of course modules and faculty programs.

To predict undergraduate students' exams grades, Yağcı (2022) proposed a prediction model. The exam grades were divided into four groups using a comparison of a variety of machine learning methods, including RF, SVM, LR, NB, ANN, and k-nearest neighbor (KNN). These groups were "32.5," "32.5-55," "55-77.5," and "77.5." The grades of 1854 students who took Turkish Language-I were taken into account in the proposed model using data that was gathered from a Turkish state university. The department, faculty, and midterm exam scores from the student's tree characteristics have been used to forecast the final exam grades. In comparison to other classification models, RF and ANN performed the best, classifying final exam grades with an area under the curve (AUC) of 86% and 74% accuracy, respectively.

Ariful and Tasnim (2021) performed a study with the objective of identifying the factors that influence the academic performance of undergraduate students at Rabindra University, Bangladesh. The study utilized primary data collected through questionnaires and opinionnaires administered to students and faculty members. The findings revealed that factors such as regular attendance, regular study, hard work, dedication, self-confidence, and support from family members and others positively impact student performance. Conversely, insufficient effort in studying, lack of interest in the subject, and a noisy and unfriendly environment in the institution were found to have a negative influence. These results can aid students, faculty, and administration in designing policies that enhance the education system at the university and contribute to the development of the nation as a whole. The study utilized a convenient and cluster random sampling method, with a sample size of 130 participants (118 students and 12 teachers) from Rabindra University, Bangladesh.

In summary most of the existing research predominantly focuses on academic factors, such as admission scores, first-level courses, test scores, and general aptitude tests (GAT), along with GPA (e.g., Alhazmi & Sheneamer, 2023), and exam grades (Yağcı, 2022). However, there is a significant gap in understanding the potential influence of non-academic factors on academic performance. Specifically, limited attention has been given to factors such as socio-economic status (e.g., sponsor type, sponsor income, and sponsor support), student motivation (e.g., classroom activeness, technology utilization, study hours), and personal well-being (e.g., health condition, stress management, and sadness frequency).

While some studies (e.g., Yadav *et al.* 2018) have considered a few non-academic factors like emotional skills, socio-economic factors, and previous academic performance, there is still a need for comprehensive analysis. Therefore, this study aims to bridge this gap by examining the relationship between academic performance and a wide range of input parameters, including student related factors (e.g., age, gender, marital status, accommodation, health condition, lecture attendance, CGPA, etc.), home related factors (e.g., sponsor type, sponsor support, family size, home responsibilities etc.), institutional related factors (e.g., preference program enrollment, quality of facilities, class size, etc.), and lecturer related factors (e.g., lecturer clarification, lecture topic density, lecturer absence frequency, lecturer punctuality etc.). By incorporating these non-academic factors into our predictive model, we aim to gain a more comprehensive understanding of student academic performance. Furthermore, this approach can provide more meaningful insights from available data and improve the reliability of predictions.

3. Methodology

3.1 Tools Used

The model development was carried out using Python programming language in Google colab, a cloud-based Jupyter notebook environment. Scikit-learn library was used to import machine learning models such as K-Nearest Neighbors, Random Forest, Gradient Boosting Method and Decision Tree. Similarly Pandas and Numpy libraries were used for exploratory data analysis and data cleaning while Matplotlib and Seaborn were used for data visualization and creation of predictive model graphics.

3.2 Conceptual Framework

A conceptual framework is a visual representation in the form of a diagram that illustrates the interrelationships between key concepts and variables. It provides a structured framework for researchers to organize their thoughts and guide their analysis. The framework and methodology of the proposed approach is presented in Figure 1.

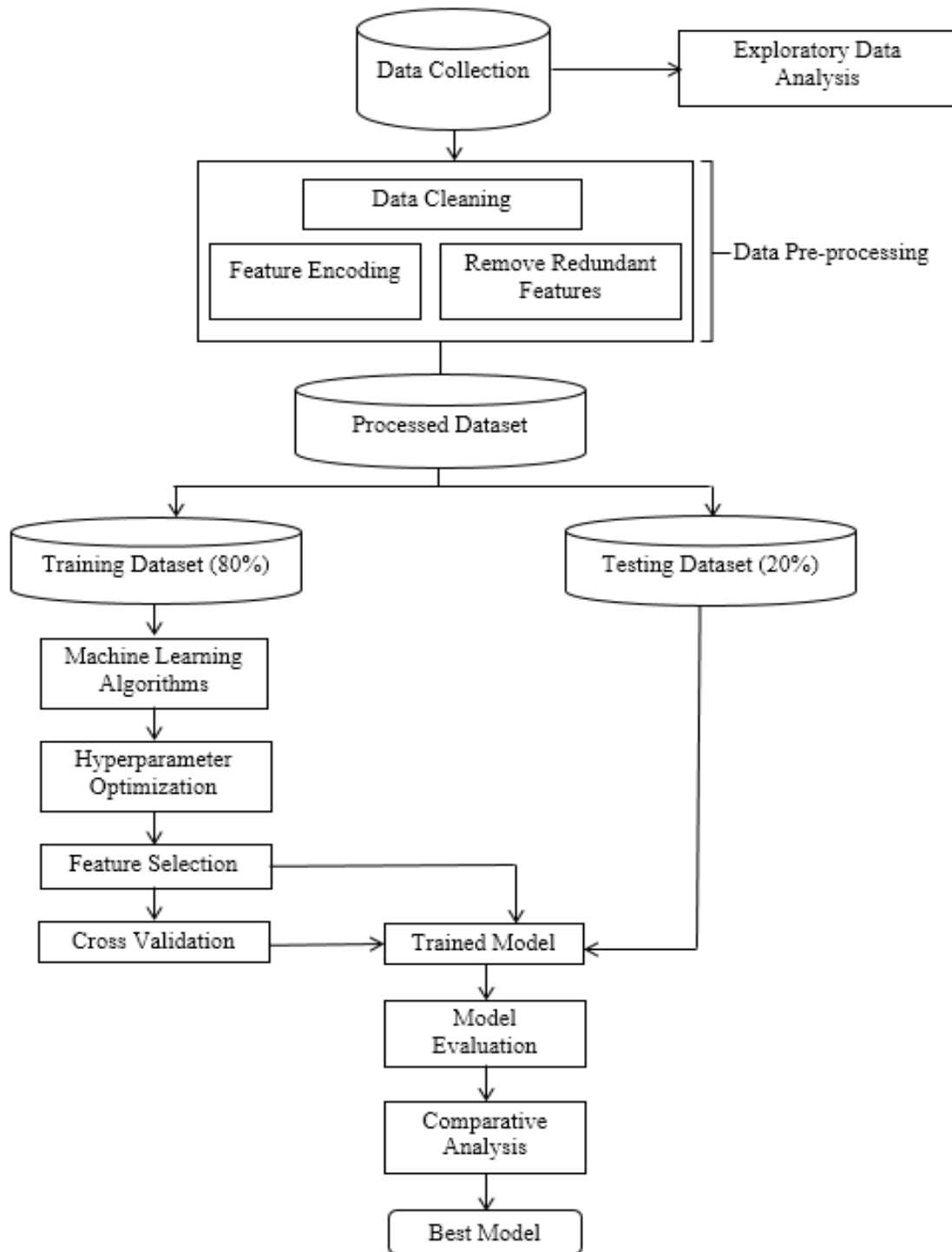


Figure 1: Conceptual framework for the model

3.2.1 Data collection and description

The data collection process was crucial for obtaining relevant information to train and evaluate the machine learning models contributing to the advancement of research in the field of predicting factors impacting academic performance. A comprehensive dataset was collected from students at MAU Yola from various levels and faculties within the university, encompassing factors such as student-related, home-related, lecturer-related, and institutional-related attributes. The data was collected using secure online questionnaires, ensuring utmost privacy and confidentiality for all respondents. The careful design and implementation of the survey-based data collection process ensured the acquisition of robust and reliable data, fostering a comprehensive understanding of the various factors influencing academic performance. The following Table 1 presents the attributes collected during the data collection process.

Table 1: Description of attributes and their respective type and values

Attribute Category	Name of the Attributes	Questions	Attribute Type and Values
	Age	Select appropriate range that corresponds to your age	Categorical (16-20, 21-25, 26-30, 31-35, 36 or above)
	Gender	What is your gender?	Categorical(Male, Female)
	Marital status	What is your marital status?	Categorical (Single, Married, Divorced, Widowed)
	Accommodation	What was your accommodation during the previous sessions?	Categorical (On-campus, Off-campus, Living with family/relatives)
	Faculty	Which faculty do you belong to?	Categorical(Physical Sciences, Life Sciences, Agriculture, Education, Engineering, Environmental, Social and Management Sciences)
	Health condition	Do you have any health issue?	Categorical (Yes, No)
	Lecture attendance	I am sometimes absent from lectures	Categorical (Strongly agree, Agree, Disagree, Strongly disagree)
	Secondary school type	Which secondary school did you attend?	Categorical (Public, Private)
	Level of study	Indicate your current level of study	Numerical (200, 300, 400, 500)
	Classroom activeness	How often do you actively participate in classroom discussions and ask Questions?	Categorical (Never, Rarely, Occasionally, Frequently, Very frequently)
	Lateness frequency	Average number of days late to school each month	Categorical (Never, Rarely, Once a week, Twice a week, Almost every day)
	Study hours	How many hours do you plan to spend each week on studying or preparing assignment?	Categorical (Less than 5 hours, 5-10 hours, 10-15 hours, 15-20 hours, more than 20 hours)

Student related factors	CGPA	Select the appropriate range that corresponds to your CGPA	Ordinal(0.00 - 1.49, 1.50 - 2.39, 2.40 - 3.49, 3.50 - 4.49, 4.50 - 5.00)
	Technology utilization	How often do you utilize technology (e.g., smartphone, computer) or online resources for studying purposes?	Categorical (Never, Rarely, Occasionally, Frequently, Very frequently)
	Employment status	Are you currently employed while studying?	Categorical (Yes, No)
	Social media usage	How many hours do you spend on social media in a day?	Categorical (Less than 1 hour, 1-2 hours, 2-3 hours, 3-4 hours, More than 4 hours)
	Stress management	How would you rate your ability to manage stress and deal with academic pressure?	Categorical (Very low, Low, Moderate, High, Very high)
	Sadness frequency	How often have you been feeling sad or experiencing lack of motivation?	Categorical (Very often, Often, Sometimes, Rarely, Never)
	Student association engagement	Are you an active member of your departmental association or a member of your student union government?	Categorical (Yes, No)
	Extracurricular activities	How often do you engage in extracurricular activities (e.g., clubs, sports, cultural events)?	Categorical (Never, Rarely, Occasionally, Frequently, Very frequently)
	Years before admission	How long did it take you to gain admission into the university?	Categorical (Less than 1 year, 1-2 years, 2-3 years, More than 3 years)
Home related factors	Sponsor type	What is the type of your sponsor?	Categorical (Parents, Guardian, Scholarship, Self-sponsored)
	Sponsor income	What is the estimated average income of your sponsor?	Categorical (Below ₦80,000, ₦80,000-₦150,000, Above ₦150,000)
	Sponsor support	How would you rate the level of support provided by your sponsor?	Categorical (Low, Average, High)
	Family size	What is the size of your family?	Categorical (1-4 members, 5-9 members, Above 9 members)
	Home responsibilities	To what extent do you have extra responsibilities at home?	Categorical (Not at all, Occasionally, Frequently, Very frequently)
	Lecturer clarification	How would you rate your lecturer's ability to explain complex concepts clearly?	Categorical (Very poor, Poor, Average, Good, Excellent)
	Lecturer supportiveness	How would you rate your lecturer's accessibility and	Categorical (Very poor, Poor, Average, Good, Excellent)

Lecturer related factors		willingness to address students' questions or concerns?	
	Illustrative teaching approach	How often does your lecturers provide real-life examples or case studies to illustrate theoretical concepts?	Categorical (Never, Rarely, Occasionally, Frequently, Always)
	Use of relevant teaching resources	How often does your lecturers use relevant and up-to-date resources or materials for teaching?	Categorical (Never, Rarely, Occasionally, Frequently, Always)
	Lecture topic density	My lecturers discuss many topics in a short period	Categorical (Strongly disagree, Disagree, Agree, Strongly agree)
	Lecturer absence frequency	My lecturers are sometimes absent from class	Categorical (Strongly disagree, Disagree, Agree, Strongly Agree)
	Lecturer punctuality	My lecturers are sometimes late to class	Categorical (Strongly disagree, Disagree, Agree, Strongly Agree)
	Teaching experience	How would you rate the teaching experience or method of your lecturers?	Categorical (Very poor, Poor, Average, Good, Excellent)
Institution related factors	Preference program enrollment	Are you enrolled in your preferred program?	Categorical (Yes, No)
	Online resources	How would you rate the availability and quality of online resources (e.g., online libraries, Google class room) provided by your institution?	Categorical (Very poor, Poor, Average, Good, Excellent)
	Teaching staff competency	How would you rate the quality and effectiveness (e.g., knowledge) of the teaching staff at your institution?	Categorical (Very poor, Poor, Average, Good, Excellent)
	Quality of facilities	How would you rate the availability and reliability of lab facilities at your institution?	Categorical (Very poor, Poor, Average, Good, Excellent)
	Class size	Do you find the classroom size and the number of students in your class satisfactory?	Categorical (Yes, No)
	Internet accessibility	Internet facility is good and is always available at the institution	Categorical (Strongly disagree, Disagree, Agree, Strongly agree)

3.2.2 Data preprocessing

The collected data was preprocessed, which included handling inconsistencies, errors, missing values, encoding categorical variables, removing redundant features, and normalizing numerical features to ensure the data is in a suitable format for machine learning algorithms.

3.3 Classification Algorithms

The Classification algorithms is a supervised learning technique that is used to identify the category of new observations on the basis of training data. In classification, a program learns from the given dataset or observations and then classifies new observation into a number of classes or groups. Such as, Yes or No, 0 or 1, low or high, etc. Classes can be called as targets/labels or categories. Since the classification algorithm is a supervised learning technique, hence it takes labeled input data, which means it contains input with the corresponding output. The main goal of the Classification algorithm is to identify the category of a given dataset, and these algorithms are mainly used to predict the output for the categorical data. In this research, different classification algorithms were used for the prediction of the student academic performance, these include K-Nearest Neighbors (KNN), Decision Tree (DT) Random Forest (RF) and Gradient Boosting Method (GBM).

3.3.1 K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) is a simple yet effective machine learning algorithm that can be used for both classification and regression tasks. It is commonly known for its classification capabilities where the target variable is categorical e.g., low or high. In KNN classification, the algorithm makes predictions for new data points by considering the k nearest neighbors in the training dataset. The value of k represents the number of neighboring data points to consider. These neighbors are determined based on the distance of their feature values to the feature values of the new data point. For CGPA prediction task, the algorithm would find the k nearest students in the training data based on features like age, gender, study hours, family size, etc. It would then look at the CGPA categories (High or Low) of these k nearest neighbors. The class (High or Low) that appears most frequently through majority vote among the k neighbors would be assigned to the new student.

3.3.2 Decision Tree (DT)

Decision Tree is a supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome. In a decision tree, there are two nodes, which are the Decision node and leaf node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches. A decision tree simply asks a question, and based on the answer (Yes/No), it further split the tree into sub trees for CGPA classification, a Decision Tree model would learn a set of rules (e.g., if absent from lecture= strongly agree and study hours < 5 hours and social media usage is >4 hours, then predict "high" CGPA, else predict "low" CGPA) based on the input features. The model would then use these learned rules to make predictions on new instances.

3.3.3 Random Forest (RF)

Random Forest (RF) is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the

concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model. As the name suggests, Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset. Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output. The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting. For CGPA classification, Random Forest would train multiple decision trees on different subsets of the training data and features (e.g., study hours, age, gender, etc.). Each Decision Tree would make a prediction, and the final prediction would be the majority vote (or average) of all the individual tree predictions.

3.3.4 Gradient Boosting Method (GBM)

Gradient Boosting Method (GBM) is another ensemble learning method that iteratively trains weak classification models (e.g., Decision Trees) and combines them to create a strong predictive model. The "boosting" process works by training multiple decision trees in a sequential manner. After the first tree is trained, the data points that were misclassified are given higher weights and vice versa. This iterative process of training trees and adjusting weights continues, with each subsequent tree in the ensemble aiming to correct the mistakes of the previous ones. The final model would be a weighted combination of all the individual Decision Trees, designed to optimally predict the "low" or "high" class.

3.4 Hyperparameter Optimization

Hyperparameter optimization involves finding the optimal combination of hyperparameters for a machine learning model to maximize its performance. Hyperparameters, set prior to model training, influence the model's behavior and characteristics, such as learning rate, regularization strength, or the number of hidden layers. These parameters cannot be directly learned from the data because they are manually engineered. When handling different prediction modeling problems, it is usually necessary to adjust the model hyperparameters. In our research, we employed a grid search approach to identify the best hyperparameter values for the machine learning algorithms utilized in our experiments.

3.5 Feature Selection

After the initial step of removing redundant features, feature selection involves identifying and retaining the most relevant and impactful features from the remaining set, while eliminating those that do not contribute significantly to the analysis or modeling tasks. By leveraging techniques like feature importance ranking, machine learning algorithms can identify the key factors that contribute to academic success or failure, thereby focusing the analysis on the most influential variables.

3.6 Cross Validation

This technique involves dividing the training set into k equal-sized subsets or folds. The model is then trained and evaluated k times, with each iteration using one fold as the validation set and the remaining $k-1$ folds for training. This process yields k models and their classification accuracy on the validation set is recorded. By averaging the k classification accuracy rates, a performance indicator that reflects the model's ability to generalize to unseen data is obtained. In this study, ten-fold cross-validation was specifically employed to mitigate the risk of overfitting the training data. This approach divides the training set into ten equal-sized folds, allowing the model's performance to be evaluated on multiple validation sets.

3.7 Model Evaluation

Model evaluation refers to the process of assessing the performance and quality of a trained machine learning model. It involves measuring how well the model performs on a specific task or problem such as classification. It emphasizes the importance of determining the model's suitability for the intended application. Evaluation metrics such as accuracy, precision, recall, and f1-score can be used to assess the model's effectiveness in predicting factors that influence student academic performance. Accuracy measure the overall performance of the model, precision measures the accuracy of the predictions for each class low and high, recall measures how well the model identifies all instances of each class, and f1-score provides a balanced measure of precision and recall for each class. By evaluating these metrics, we can gain valuable insights into the models' performance and their ability to predict factors that impact student academic success. The confusion matrix is shown in Figure 2.

		Predicted Label	
		Predicted Low	Predicted High
Actual Label	Actual Low	True Low	False High
	Actual High	False Low	True High

Figure 2: Confusion matrix

$$accuracy = \frac{True\ Low + True\ High}{True\ low + True\ High + False\ Low + False\ High} \tag{1}$$

$$precision\ Low = \frac{True\ Low}{True\ low + False\ low} \tag{2}$$

$$precision\ High = \frac{True\ High}{True\ High + False\ High} \tag{3}$$

$$recall\ Low = \frac{True\ Low}{True\ Low + False\ High} \tag{4}$$

$$recall\ High = \frac{True\ High}{True\ High + False\ Low} \tag{5}$$

$$f1 - score\ Low = 2 \times \frac{precision\ Low \times recall\ Low}{precision\ Low + recall\ Low} \tag{6}$$

$$f1 - score\ High = 2 \times \frac{precision\ High \times recall\ High}{precision\ High + recall\ High} \tag{7}$$

4. Results and Discussions

4.1 Data Analysis

The online questionnaire was sent to selected student participants of MAU Yola across seven different faculties in the university and was accessible to them for a specified period, during which they were encouraged to complete and submit their responses within that timeframe. Thereafter, a total of 835 student respondents' data was collected. It can be observed that most of the respondents from the university are male consisting of 79.3% (662) while female is 20.7% (173) as shown in Figure 3. The age of the respondent used for this study ranges from 16-20 years and above however the response indicate that most of the respondent fall in the age bracket 21-25 years with 51.1% (427) of the sample as shown in Figure 4.

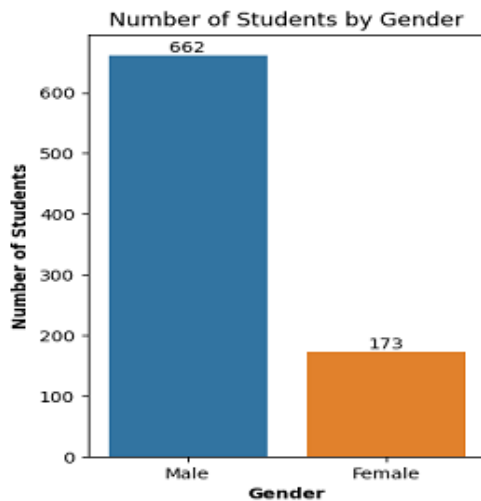


Figure 3: Students Gender Bar chart

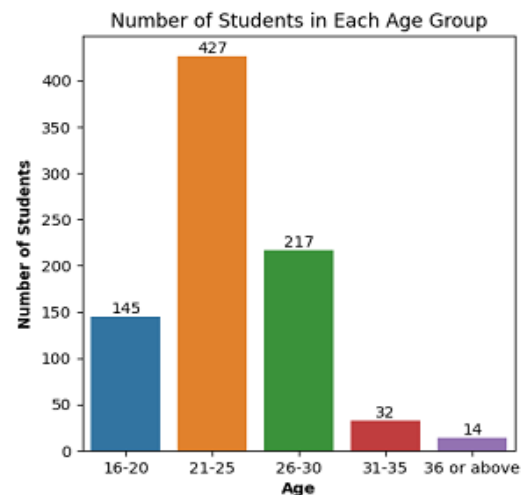


Figure 4: Students Age group Bar chart

The faculty with the highest representation was Physical Sciences with 40.1% (335) respondents, followed by Life Sciences with 17.7% (148) and Environmental with 12.7% (106), as illustrated in Figure 5. The academic level of the respondent used for this study ranges from 200 - 500 level however the highest proportion of the respondents are 500-level students with 39.2% (327), followed by 300-level with 29.5% (246) as shown in Figure 6.

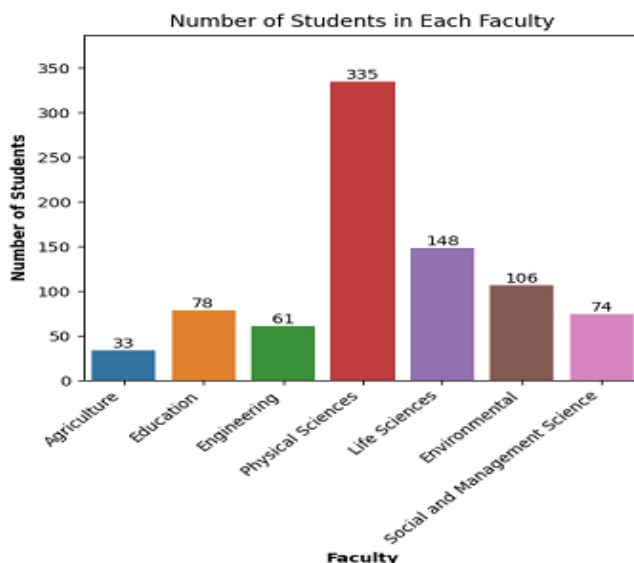


Figure 5: Students Faculty Bar chart

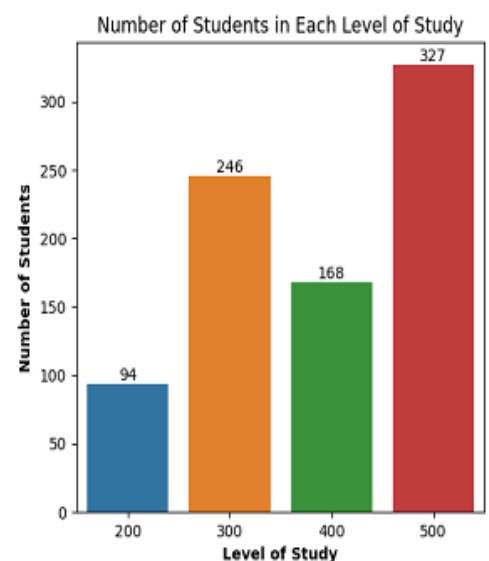


Figure 6: Students Level Bar chart

4.2 Dataset Overview

Before delving into the results, it is important to provide a brief overview of the dataset used in this study. Out of the total of 835 student respondents, each sample represents a student's related factors, lecturer-related factors, institution-related factors and home-related factors which includes 40 features. These features can be categorized into two main types - namely independent variables and dependent variables. The independent variables are the attributes that were used to predict the dependent variable, which is the CGPA range that has been categorized into Low and High.

Table 2: Sample of the original dataset

	Gender	MaritalStatus	Age	Accommodation	Faculty	SecSchType	HealthCondition	LevelOfStudy	LectureAttend	ClassActive	...	InternetAccess
0	Male	Single	31-35	Living with family/relatives	Physical Sciences	Public	No	500	Strongly disagree	Occasionally	...	Strongly disagree
1	Male	Single	31-35	Off-campus	Physical Sciences	Public	No	500	Disagree	Frequently	...	Strongly disagree
2	Male	Single	21-25	Off-campus	Physical Sciences	Private	No	300	Agree	Never	...	Strongly disagree
3	Female	Single	21-25	On-campus	Life Sciences	Public	No	500	Disagree	Never	...	Agree
4	Male	Single	16-20	Off-campus	Physical Sciences	Private	No	200	Agree	Rarely	...	Disagree

In order to begin the development of the machine learning model for predicting CGPA category, the collected data underwent a thorough cleaning process to remove any inconsistencies, missing values, or outliers. Additionally, preprocessing techniques such as feature encoding, which involve handling categorical variables, were applied to ensure the data was suitable for model training and evaluation. Techniques related to removing redundant features were also applied.

Table 3: Student dataset after preprocessing

	Gender	MaritalStatus	Age	Accommodation	Faculty	SecSchType	HealthCondition	LevelOfStudy	LectureAttend	ClassActive	...	InternetAccess
0	1	2	3	0	5	1	0	500	0	2	...	0
1	1	2	3	1	5	1	0	500	1	3	...	0
2	1	2	1	1	5	0	0	300	2	0	...	0
3	0	2	1	2	4	1	0	500	1	0	...	2
4	1	2	0	1	5	0	0	200	2	1	...	1

4.3 Data Splitting

The preprocessed data were randomly split into training and testing sets, with 80% of the data allocated for training the machine learning model as shown in Table 4. This provides a substantial amount of data for the model to learn patterns and relationships within the dataset. The remaining 20% of the data that the model has not seen during training was allocated for testing the model's performance as shown in Table 5. This allows for a reliable evaluation of the model's performance on unseen data.

Table 4: Split training dataset

Training Set:

	Gender	MaritalStatus	Age	Accommodation	Faculty	SecSchType	HealthCondition	LevelOfStudy	LectureAttend	ClassActive	...	InternetAccess
186	1	2	0	1	3	0	0	200	2	0	...	1
576	0	2	0	2	4	1	0	200	1	1	...	1
135	0	2	1	0	4	0	0	400	0	1	...	1
546	0	2	2	1	5	1	0	300	2	2	...	2
613	1	2	0	2	1	1	0	200	0	1	...	0

Table 5: Split testing dataset

Testing Set:

	Gender	MaritalStatus	Age	Accommodation	Faculty	SecSchType	HealthCondition	LevelOfStudy	LectureAttend	ClassActive	...	InternetAccess
780	1	2	1	1	3	1	0	500	2	1	...	2
767	1	2	1	2	0	1	0	300	1	1	...	1
61	1	2	1	2	5	0	0	500	2	3	...	2
347	1	2	2	2	2	1	0	500	2	2	...	0
273	1	2	1	1	5	0	0	300	1	1	...	0

Additionally, stratified sampling techniques were utilized during the random split of data to ensure that the distribution of the CGPA categories was preserved in both the training and testing sets. For the Low CGPA category, 80% (288) of the total 360 instances were used for training, while 20% (72) were allocated for testing. Similarly, for the High CGPA category, 80% (380) of the total 475 instances were used in the training set, with the remaining 20% (95) reserved for the testing set. This approach maintains the original class distribution in both subsets, ensuring a representative sample for model development and evaluation.

4.4 Comparative Model Performance Evaluation

This section evaluates the performance of four different machine learning models developed and tested for the given research. The performance evaluation for these models is displayed in the appendix. Table 6 presents a side-by-side comparison of the key performance metrics for each model, including accuracy, precision, recall, and f1-score to assess the models' classification abilities. This tabular format allows for an easy visualization and assessment of the relative strengths and weaknesses of the different models and also enable the process of selecting the best model. Similarly Table 7 present comparison of the classifier models performance based on correctly and incorrectly classified instances.

Table 6: Models performance metrics comparison

Models	Metrics	Low	High	accuracy (%)
KNN	precision	0.8317	0.8333	83
	recall	0.8842	0.7639	
	f1-score	0.8571	0.7971	
DT	precision	0.8812	0.9091	89
	recall	0.9368	0.8333	
	f1-score	0.9082	0.8696	
GBM	precision	0.8515	0.8636	86

	recall	0.9053	0.7917	
	f1-score	0.8776	0.8261	
RF	precision	0.9677	0.9324	95
	recall	0.9474	0.9583	
	f1-score	0.9574	0.9452	

Table 7: Comparison of correctly and incorrectly classified instances

Student Classification	KNN	DT	GBM	RF
Correctly classified student	139	149	143	159
Incorrectly classified student	28	18	24	8
Correctly classified low student	84	89	86	90
Incorrectly classified low student	17	12	15	3
Correctly classified high student	55	60	57	69
Incorrectly classified high student	11	6	9	5

The comparative evaluation of the four machine learning models reveals that the Random Forest (RF) model outperformed the others with the highest overall accuracy of 95%. This high accuracy indicates that the RF model is highly effective in correctly classifying both low and high CGPA values. The RF model correctly classifies a large majority of instances, 159 out of the 167 total instances, while incorrectly classifying only 8 instances. This demonstrates the model's excellent precision, recall, and f1-score (all above 0.93) for both the Low and High CGPA classes. In contrast, the Decision Tree (DT) model also show promising result which correctly classifies a larger number of instances (149 out of 167) and incorrectly classified 18, achieving the second-highest overall accuracy of 89% but with the performance on the High class slightly weaker than on the Low class, with a recall of 0.83 compared to 0.94 for the Low class. The K-Nearest Neighbors (KNN) model had the lowest performance which correctly classified 139 instances and incorrectly classified a larger number of instances (28 out of 167), achieving an accuracy of 83%, with a balanced performance between the Low and High classes but a weaker recall score of 0.76 for the High class compared to 0.88 for the Low class. The Gradient Boosting Method (GBM) model correctly classified 143 instances and incorrectly classified a significant number of instances (24 out of 167) compared to Decision Tree and Random Forest, achieving an accuracy of 86% and maintaining a relatively balanced performance between the Low and High classes, with f1-scores around 0.88 and 0.83, respectively, but with a lower recall score of 0.79 for the High class. The detailed performance metrics for each model, including classification reports, precision, recall, f1-score, and confusion matrices, further highlight the superiority of the Random Forest model. Based on this comprehensive comparative analysis, the Random Forest model can be identified as the most effective model for this classification task.

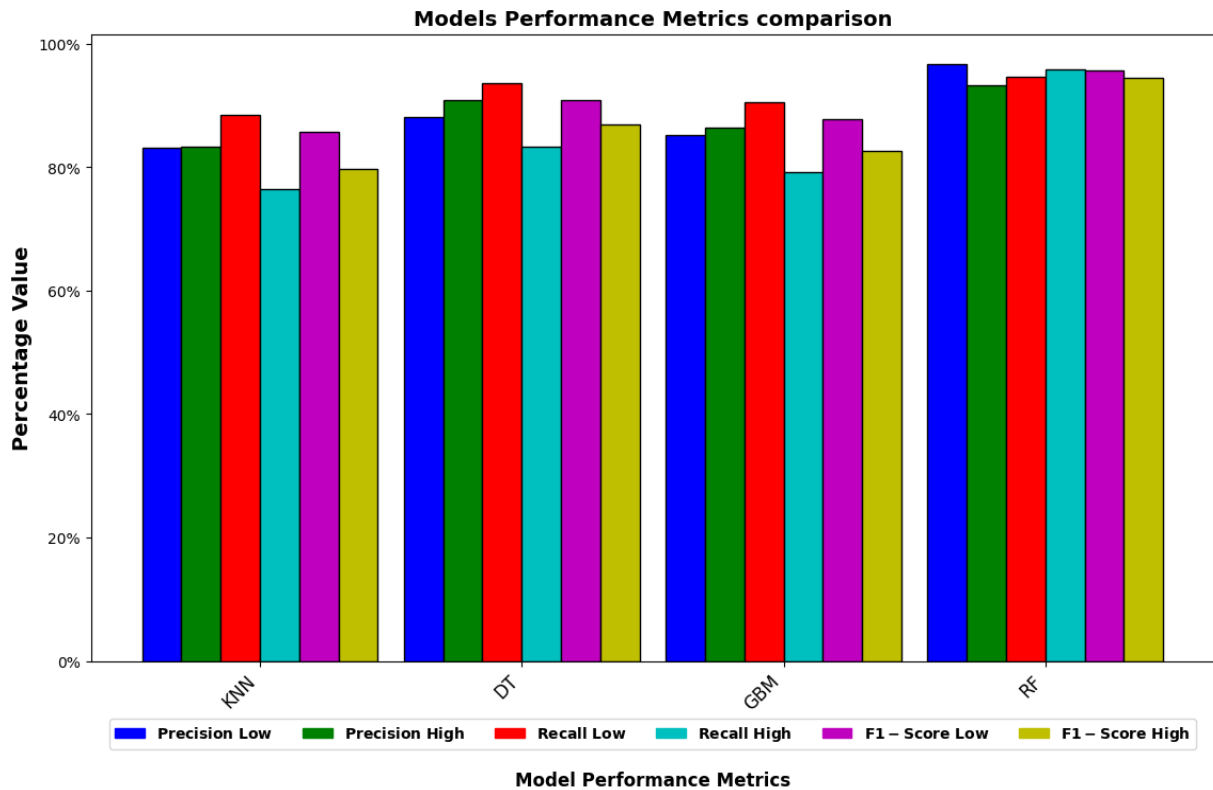


Figure 7: Bar chart showing the performance metrics

Figure 7 shows a grouped bar chart that compare the algorithms across the performance metrics. The bars for each algorithm are grouped together, and each group represents the values of the different metrics (precision low, precision high, recall low, recall high, f1-score low and f1-score high) for that algorithm. The height of each bar represents the value of the corresponding metric. Since Random Forest outperformed the other algorithms, Random Forest can be identified as the most effective model for predicting CGPA category.

4.5 Feature Importance Analysis

The dataset may contain many features, but some may not affect students’ academic performance. Hence, it is recommended to know the most important features. The purpose of this is to identify the most influential features or variables that have the greatest impact on predicting the student’s CGPA. This analysis utilizes model-based techniques to quantify the relative importance of each feature in the dataset. Specifically, this analysis applies three different machine learning algorithms, namely; Decision Tree, Random Forest, and Gradient Boosting Method to assess feature importance. Each of these algorithms has built-in methods for evaluating the contribution of individual features to the overall predictive performance of the model. By comparing the feature importance rankings across these three different models, the features that are consistently deemed highly important by multiple algorithms are likely to be the most critical input variables for accurately predicting a student’s academic performance.

4.5.1 Feature importance for Decision Tree (DT)

The diagram in Figure 8 shows the Decision Tree in Google colab ranking the features in order of importance from highest to lowest. From the diagram it can be observed that the top four most important

attributes are *Years before admission, Lecture attendance, Lecturer clarification and Sponsor support* while the least three features for this algorithm are *Accommodation, Employment status, and program preference*.

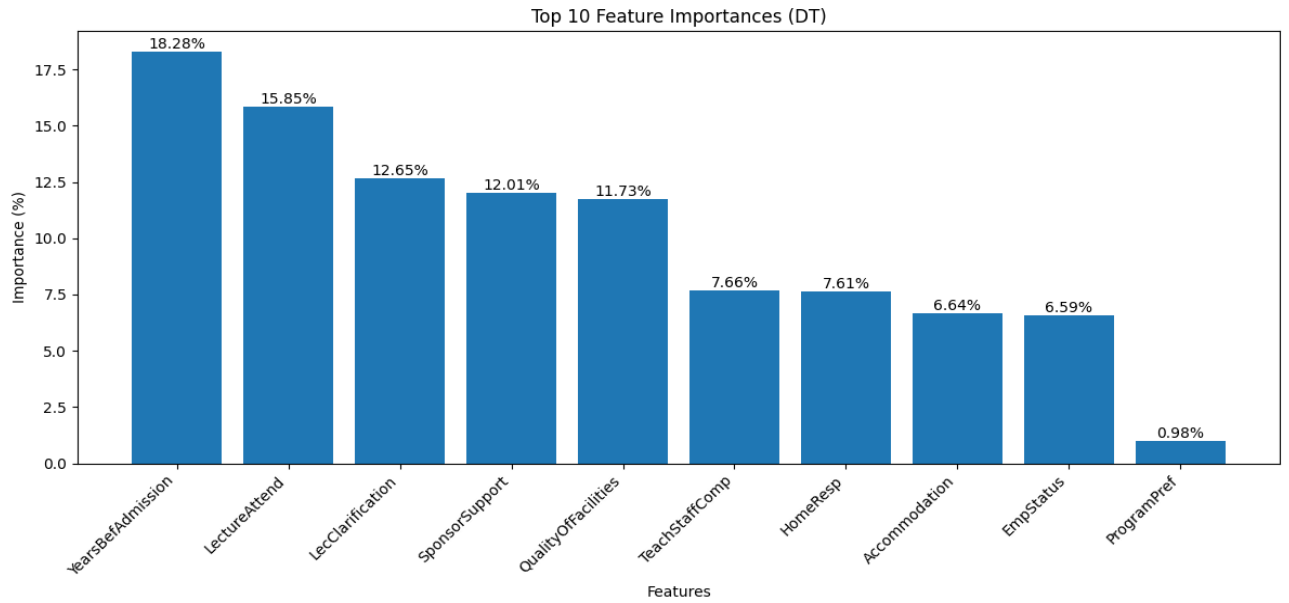


Figure 8: Decision Tree ranked feature importance

4.5.2 Feature importance for Gradient Boosting Method (GBM)

The feature ranked with Gradient Boosting Method is presented in Figure 9. From the diagram it can be observed that the best four attributes are *Lecture attendance, Sponsor support, Quality of facility and Lecturer clarification* while the least three features for this algorithm are *Accommodation, Employment status, and Program preference*. The Gradient Boosting Methods shares three best attributes (*Lecture attendance, Sponsor support, Lecturer clarification*) and three least attribute (*Accommodation, Employment status, and Program preference*) with the Decision Tree.

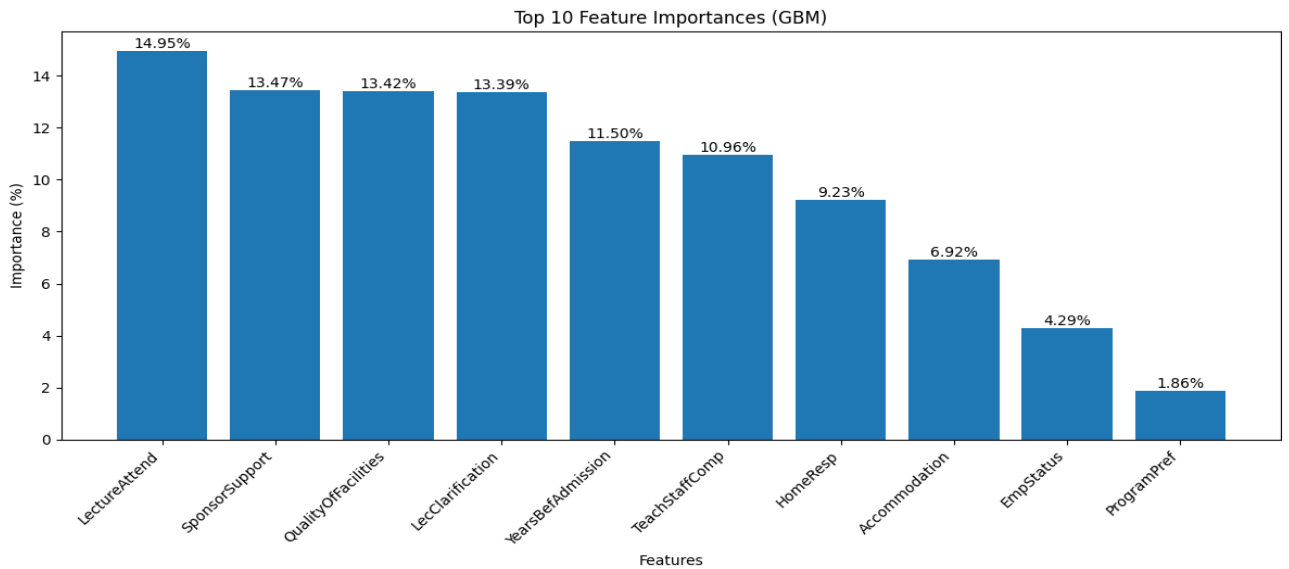


Figure 9: Gradient Boosting Method ranked feature importance

4.5.3 Feature importance for Random Forest (RF)

The diagram (Figure 10) present the feature ranked by Random Forest. The attributes considered as the best four with this algorithm are *Lecture attendance*, *Quality of facility*, *Lecturer clarification* and *Home responsibilities* while the least three features for this algorithm are *Accommodation*, *Employment status*, and *Program preference*. Random Forest shares part of the top four attributes (*Lecture attendance*, *Lecturer clarification*) and the least three features (*Accommodation*, *Employment status*, *Program preference*) with the Decision Tree and Gradient Boosting Methods. Additionally, the Random Forest aligned with Gradient Boosting Method results in identifying *Quality of facility* as one of the top four ranked attributes.

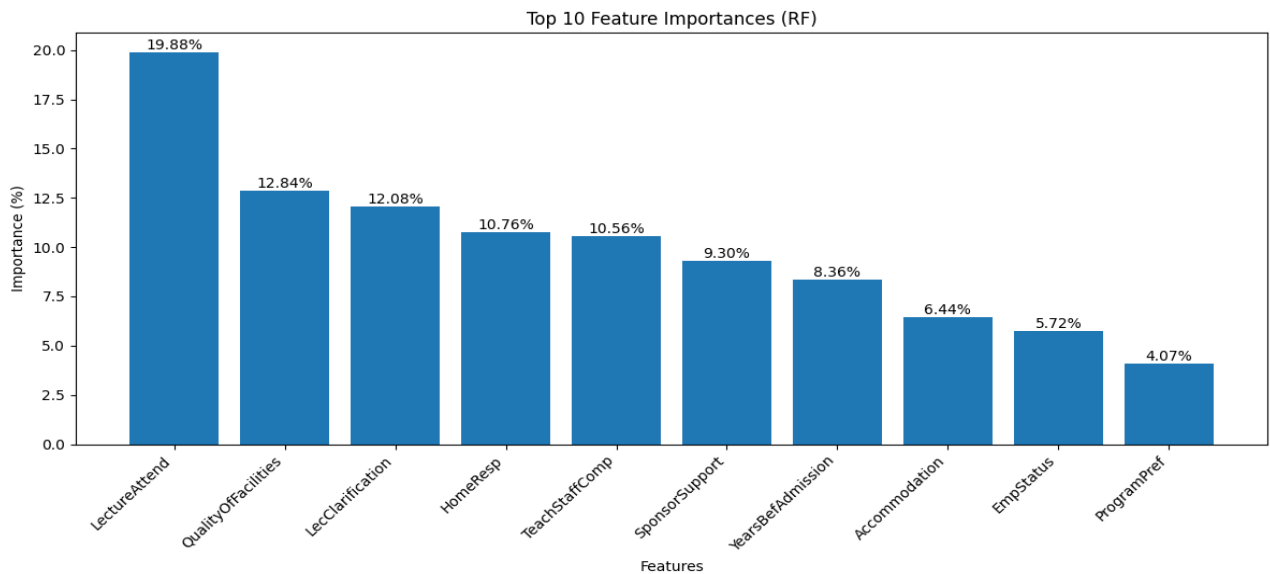


Figure 10: Random Forest ranked feature importance

In summary, of all the feature importance ranking algorithms used in this study, the attributes shared as part of the top four attributes by two or more of the algorithms are *Lecture attendance*, *Sponsor support*, *Lecturer clarification* and *Quality of facility*. This indicates that these attributes could contribute highly to how students are ultimately classified into performance groups. The attributes shared as part of the least three attributes by two or more of the algorithms are *Accommodation*, *Employment status* and *Program preference*, suggesting and strongly imply that these features may contribute least to students' classified group. The insights gained from this comprehensive feature importance analysis are that all of the four factors considered in this research, namely; Student related factors (lecture attendance), Lecturer related factors (lecturer clarification), Institution related factors (quality of facility) and Home related factors (sponsor support) had a great impact on the academic performance of students.

5. Conclusion

This research aimed to develop a robust machine learning model for predicting and analyzing the key factors impacting student academic performance. By leveraging a comprehensive dataset collected using secure online questionnaires from undergraduate students at MAU Yola, the study explored the influence of student-related, home-related, lecturer-related, and institution-related factors on academic achievement. Four machine learning models, including Random Forest (RF), Decision Tree (DT), K-Nearest Neighbors (KNN), and Gradient Boosting (GBM) were comparatively evaluated based on performance metrics such

as accuracy, precision, recall and f1-score. Based on the results of the evaluation, it is evident that the Random Forest model outperformed the other machine learning models with an overall accuracy of 95%. For predicting low performing students, the model had a precision of 0.9677, recall of 0.9474, and f1-score of 0.9574. For high performing students, the precision was 0.9324, recall was 0.9583, and f1-score was 0.9452. These strong performance metrics across both low and high performing student groups demonstrate the effectiveness of the Random Forest model in accurately predicting student academic performance based on the factors considered in the study. The feature importance analysis identified lecture attendance, sponsor support, quality of facilities, and lecturer clarification as the most influential factors that have the greatest impact on student academic performance, however for other features such as accommodation, employment status, and program preference, they have a low impact on student performance.

Early identification of academic performance allows educational institutions to offer timely assistance and suitable recommendations to students. Recognizing underperforming students at the beginning of the academic year can be essential for enhancing their learning skills. This study emphasizes the importance of data-driven machine learning methods in addressing various challenges faced by students, thus helping educational institutions formulate effective strategies for improvement. The study contributes to the existing body of knowledge by showcasing the potential of machine learning techniques in education research. The developed model and the identified influential factors can serve as a valuable resource for educational stakeholders in developing more effective strategies and interventions for improving student academic performance. Future research can explore different machine learning algorithms to enhance predictive accuracy. Additionally, collecting data from diverse institutions could expand the understanding of factors influencing student success.

Reference

- Adenubi, A. O., & Samuel, N. (2024). Revolutionizing education with artificial intelligence and machine learning: personalization, retention, and resource optimization. *Kasu Journal of Computer Science*, 1(2), 378-391. <https://doi.org/10.47514/kjcs/2024.1.2.0015>
- Ahmed, O. A., Moshood A. A., & Kayode, S. A. (2019). Students' academic performance and dropout predictions: A Review *Malaysian Journal of Computing*, 4 (2), 278-303.
- 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*, 1-26. doi: <https://doi.org/10.1007/s10639-023-11700-0>
- Alhazmi, E., & Sheneamer, A. (2023). Early predicting of students performance in higher education. *IEEE Access*, 11, 27579- 27589.
- Al-Zoubi, S. M., & Younes, M. B. (2015). Low academic achievement: causes and results. *Theory and Practice in Language Studies*, 5(11), 2262-2268.
- Anas, N., Usman, A., Abdullahi, I., & Jalo, A., I. (2023). An empirical study of the common factors affecting students' academic performance in Gombe state university, Gombe-Nigeria. *International Journal of Social Science and Human Research*, 6(3), 1410-1415. <https://doi.org/10.47191/ijsshr/v6-i3-09>
- Anusha, M., Karthik, K., Rani, P., & Srikanth, V. (2019). Prediction of student performance using machine learning. *International Journal of Engineering and Advanced Technology*, 8(6), 247-255. Retrieved from <https://doi.org/10.35940/ijeat.E7520.088619>.

- Ariful I., & Tasnim, S. (2021). “An analysis of factors influencing academic performance of undergraduate students: A case study of Rabindra university, Bangladesh (RUB).” *Shanlax International Journal of Education*, 9(3), 127-135. doi: <https://doi.org/10.34293/education.v9i3.3732>
- Baffa, A. U., & Nasiru, A. (2022). Conceptual and empirical review of internal conflict management strategies in Nigerian universities. *Gusau Journal of Sociology*, 3 (1), 54-74.
- Ekubo, E. A., & Esiefarienrhe, B. M. (2022). Using machine learning to predict low academic performance at a Nigerian university. *The African Journal of Information and Communication (AJIC)*, 30, 1-33. <https://doi.org/10.23962/ajic.i30.14839>
- Gonfa, R. (2020), “Reviewing the Factors that hinder academic achievement of the students”, *Global Scientific Journal*, 8(5), 47-56.
- Gupta, S. K., Kumar, I., Manish, K., Kumari, S., & Devi, A. S. (2023). Student performance analysis using machine learning. *International Journal of Progressive Research in Engineering Management and Science (IJPREMS)*, 3(5), 480-484. doi: <https://www.doi.org/10.58257/IJPREMS31181>
- Hasan, H. R., Rabby, A. S. A., Islam, M. T., & 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)* (pp. 1-7). Kanpur, India.
- Islam, A., & Tasnim, S. (2021). An analysis of factors influencing academic performance of undergraduate students: A case study of Rabindra university, Bangladesh (RUB). *Shanlax International Journal of Education*, 9(3), 127-135.
- Nsiah, H. (2017). Fear of failure and the academic performance of students from low-income families. *International Journal of Education and Social Science*, 4(10), 19-26.
- Ofori, F., Maina, E. & Gitonga, R. (2020). Using machine learning algorithms to predict students’ performance and improve learning outcome: A literature based review, *Journal of Information and Technology*, 4(1), 33-55.
- Ojajuni, O., Ayeni, F., Akodu, O., Ekanoye, F., Adewole, S., Ayo, T., Misra, S., & Mbarika, V. (2021). Predicting student academic performance using machine learning. *ICCSA*, 481-491.
- Okereke, G.E, Mamah C.H, Ukekwe E.C & Nwagwu H.C. (2020). A machine learning based framework for predicting student’s academic performance. *Physical Science & Biophysics Journal*, 4(2).
- Olufemi, O. T., Adediran, A. A., & Oyediran, W. O. (2018). “Factors affecting students’ academic performance in colleges of education in southwest, Nigeria”, *British Journal of Education*, 6(10), 43-56.
- Oreški, D., & Zamuda, D. (2022). Machine learning based model for predicting student outcomes. *Proceedings of the International Conference on Industrial Engineering and Operations Management Istanbul, Turkey*.
- Rahman, M. (2021). Impact of socio-economic factors on undergraduate students’ academic performance in Bangladesh: A case study at social science faculty, university of Dhaka. *Asian Journal of Social Sciences and Legal Studies*, 3(4), 147-157.
- Rajendran, S., Sinhaa, A. A., & Chamundeswari S. (2022). Predicting the academic performance of middle- and high-school students using machine learning algorithms. *Social Sciences & Humanities Open*, 1-15. <https://doi.org/10.1016/j.ssaho.2022.100357>

Solomon, D., Patil, S., & Agrawal, P. (2018). Predicting performance and potential difficulties of university student using classification : survey paper. *International Journal of Pure and Applied Mathematics*, 118(18), 2703–2707.

Yadav, K. R., Singh, O. K., Choudhary, A. U., & Singh, M. (2018). Tracking and predicting student performance in degree programs using machine learning approach. *International Journal of Research and Analytical Reviews (IJRAR)*, 5(4), 1809-1811.

Yağcı, M. (2022). Educational data mining: Prediction of students’ academic performance using machine learning algorithms. *Smart Learning Environments*, 9(1), 1–9. <https://doi.org/10.1186/s40561-022-00192-z>

Yusuf, E. I., Audu, G. T., & Yusuf, F. I. (2020). A theoretical analysis of the divergence factors that affect students academic performance in university of Jos Nigeria. *International Journal of Innovative Research and Development*, 9(8), 9-18.

Appendix: Models Performance Evaluation for Different Machine Learning Algorithms

```

--- Evaluation for KNN ---

Correctly classified instances: 139
Incorrectly classified instances: 28
Total number of instances: 167

Classification Report:

```

	precision	recall	f1-score	support
Low	0.83	0.88	0.86	95
High	0.83	0.76	0.80	72
accuracy			0.83	167
macro avg	0.83	0.82	0.83	167
weighted avg	0.83	0.83	0.83	167

```

--- Metrics Performance Details: ---

Precision (Low): 0.8317
Precision (High): 0.8333
Recall (Low): 0.8842
Recall (High): 0.7639
F1 Score (Low): 0.8571
F1 Score (High): 0.7971

Confusion Matrix:

```

	Predicted Low	Predicted High
Actual Low	84	11
Actual High	17	55

(a) K-Nearest Neighbors (KNN)

--- Evaluation for GBM ---

Correctly classified instances: 143
 Incorrectly classified instances: 24
 Total number of instances: 167

Classification Report:

	precision	recall	f1-score	support
Low	0.85	0.91	0.88	95
High	0.86	0.79	0.83	72
accuracy			0.86	167
macro avg	0.86	0.85	0.85	167
weighted avg	0.86	0.86	0.86	167

--- Metrics Performance Details: ---

Precision (Low): 0.8515
 Precision (High): 0.8636
 Recall (Low): 0.9053
 Recall (High): 0.7917
 F1 Score (Low): 0.8776
 F1 Score (High): 0.8261

Confusion Matrix:

	Predicted Low	Predicted High
Actual Low	86	9
Actual High	15	57

(b) Gradient Boosting Method (GBM)

--- Evaluation for DT ---

Correctly classified instances: 149
 Incorrectly classified instances: 18
 Total number of instances: 167

Classification Report:

	precision	recall	f1-score	support
Low	0.88	0.94	0.91	95
High	0.91	0.83	0.87	72
accuracy			0.89	167
macro avg	0.90	0.89	0.89	167
weighted avg	0.89	0.89	0.89	167

--- Metrics Performance Details: ---

Precision (Low): 0.8812
 Precision (High): 0.9091
 Recall (Low): 0.9368
 Recall (High): 0.8333
 F1 Score (Low): 0.9082
 F1 Score (High): 0.8696

Confusion Matrix:

	Predicted Low	Predicted High
Actual Low	89	6
Actual High	12	60

(c) Decision Tree (DT)

--- Evaluation for RF ---

Correctly classified instances: 159
 Incorrectly classified instances: 8
 Total number of instances: 167

Classification Report:

	precision	recall	f1-score	support
Low	0.97	0.95	0.96	95
High	0.93	0.96	0.95	72
accuracy			0.95	167
macro avg	0.95	0.95	0.95	167
weighted avg	0.95	0.95	0.95	167

--- Metrics Performance Details: ---

Precision (Low): 0.9677
 Precision (High): 0.9324
 Recall (Low): 0.9474
 Recall (High): 0.9583
 F1 Score (Low): 0.9574
 F1 Score (High): 0.9452

Confusion Matrix:

	Predicted Low	Predicted High
Actual Low	90	5
Actual High	3	69

(d) Random Forest (RF)