COMPARATIVE AND PREDICTIVE ANALYTICS FOR STUDENTS' ACADEMIC PERFORMANCE USING SUPERVISED CLASSIFIERS

BASHIRU, A.S.[1]*, BAOKU, I.G.[2], BASHIR, A.J.[3] AND ILYASU, U.[4]

[1] Department of Computer Science, Abdu Gusau Polytechnic Talata Mafara, Zamfara State, Nigeria, [2]Department of Mathematical Sciences, Faculty of Physical Sciences, [3]Department of Cyber Security, [4]Department of Computer Science, Faculty of Computer Science and Artificial Intelligence, Federal University Dutsin-Ma, Katsina State, Nigeria

ABSTRACT

The greatest aim of every educational setup is giving the best educational experience and knowledge to the students. Discovering the students who need extra support and guidance so as to carry out the necessary actions to enhance their performance plays an important role in achieving that aim. In this research work, five machine learning algorithms have been used to build a classifier that can predict the performance of students in higher institutions considering two institutions, which are Abdu Gusau Polytechnic Talata Mafara and College of Education Maru, Zamfara State are considered. The machine learning algorithms include Random Forest, Support Vector Machine, Logistic Regression, K-Nearest Neighbour and Naive Bayes. The models have been compared using the Precision, Recall, F1-Score and Support classification accuracy. The dataset used to build the models is collected from the MIS centre of each of the institutions. The Random Forest model is found to achieve the best performance.

Keywords: Feature Selection, Logistic Regression, Random Forest, Support Vector Machine, Students' Performance Prediction,

***Correspondence:** baliyusani@gmail.com**,** +2348062477796

INTRODUCTION

With the wide usage of computers and internet, there has recently been a huge increase in publicly available data that can be analyzed; from online sales information, website traffic or user habits, data are generated every day. Such a large amount of data presents both a problem and an opportunity. The problem is that it is difficult for humans to analyze such large data. The opportunity is that this type of data is ideal for computers to process, because it is stored digitally in a well-formatted way, and computers can process data much faster than humans.

Although machine learning applications vary, their general functions are similar throughout their applications. The computer analyzes a large amount of data, and finds patterns and rules hidden in the data. These patterns and rules are mathematical in nature; they can be easily defined and processed by a computer. The computer can then use those rules to meaningfully characterize new data according to [1].

Image recognition technologies also use machine learning to identify particular objects in an image, such as faces. The machine learning algorithm analyzes images that contain a certain object. If given enough images to process, the algorithm is able to determine whether an image contains that object or not [3]. In addition, machine learning can be used to understand the kind of products a customer might be interested in [2]. The research focuses on supervised learning, more specifically predictive analytics, which is the process of using machine learning to predict future outcomes [5]. Predictive analytics has a wide range of applications, such as fraud detection, analyzing population trends, understanding user behaviour, etc. [6].

This technique produced a higher accuracy rate in classification and prediction in comparison with the other algorithms such as Naive Bayes, Bagging, Boosting and Random Forest. In this research, Gender, Family Size, Parents Status, Mother and Father Education, Mother and Father Job are some of the influential factors that can adversely impact the achievement of students' academic history, pre-enrollment status, socioeconomic status, psychological factors are considered and analyzed to understand the progression from one level to another. These served as the dataset features. Naïve Bayes, Random Forest and ensemble methods have classified the attributes and shown the highly influential attributes or the factors that determine the performance of students. Random Forest algorithm with increased iteration and bag size has performed better than the other ensemble methods and the Naïve Bayes algorithm [7]. Their research work dealt with the predicting students' academic performance in a technical institution in India. A dataset was obtained using a questionnaire-based survey and the academic section of the chosen institution. Data-pre-processing and factor analysis have been performed on the obtained dataset to remove the anomalies in the data; reduce the dimensionality of data and to obtain the most correlated feature [8].

In the work of Ravina *et al*. [8], Python 3 tool was used for the comparison of machine learning

algorithms. The support vector regression linear algorithm provided superior prediction of 83.44% after applying multiple linear regression, support vector regression_rbf, support vector regression_poly and support vector regression linear for determining the academic performance of the final year undergraduate students of the chosen academic institution. Adaptive dynamic tests were used for assessing student academic performance. There are four main stages in this method: Data Collection, Classification, Creation of a Predictive Model and Evaluation. Among other conclusions, it was shown that the prediction accuracy of the framework, when the Adadelta and Adagrad optimizers were used, was found to be 76.73% and 82.39%, respectively, while its overall respective learning performance was 80.76% and 86.57%; meaning that it was capable of predicting teams' performance adequately and accurately.

The specific focus of this research work is education. The aim is to predict student performance. First, the trained dataset is taken as input. There are different datasets, containing different types of information like certain information about a student, such as age, gender, family background or medical information. Secondly, the algorithms were trained in the model, which is outputs success or failure of the student, using other variable. This research compares five models after thoroughly preprocessing best features selection and then made use of Random Forest, Support Vector Machine, Logistic Regression, K-Nearest Neighbor and Naive Bayes as the classifiers. To determine which classifier provides the best results, Precision, Recall, F1-Score and Support were used and the implementation was done on Python.

MATERIALS AND METHODS

This study focuses on a predictive model using the features of low, average and high performing higher institution students in two different schools, Abdu Gusau Polytechnic, Talata Mafara and College of Education, Maru, Zamfara State. Through this model, new students with similar characteristics can be identified early and the institutions can set up the essential interventions to cater for these students' needs. To achieve this aim, the research looked at five machine-learning classifiers and used the data collected from each of the schools to build models using the Python modeling tool and libraries. Figure 1 depicts the architectural framework of the study. In the following sections, a description of each step of the research process is presented. To address the common issues of above review literatures such as class imbalance in the feature selected, and classification errors, this research has used a model which has following the phases.

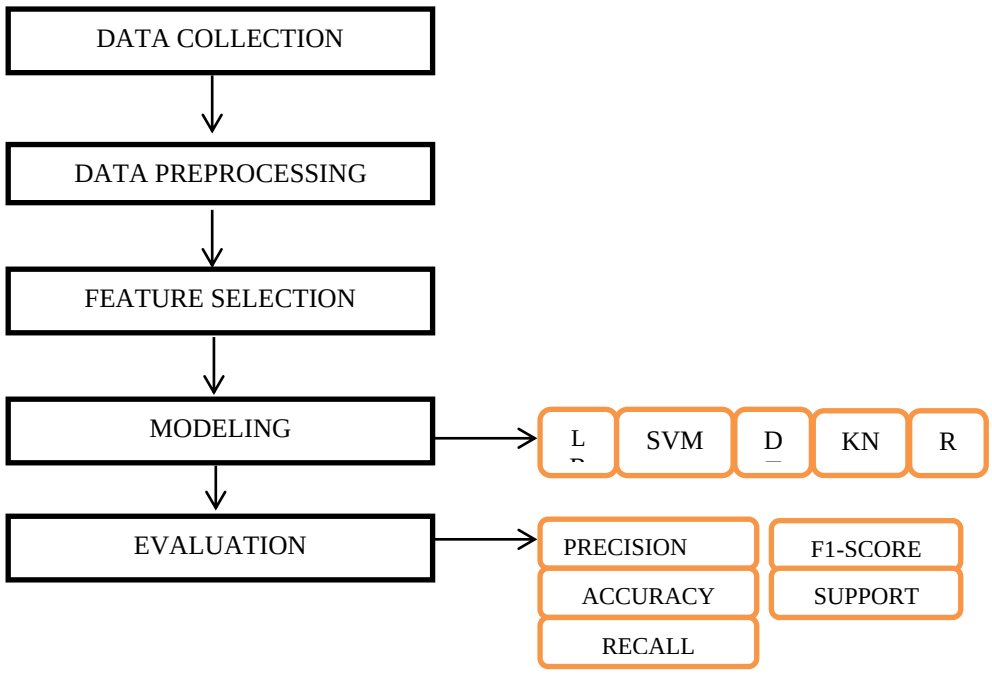

Figure 1: Architectural frame involving methodology

Step 1: Data collection

The dataset used in this research was collected from Abdu Gusau Polytechnic, Talata Mafara and College of Education, Maru, Zamfara State. We had access to the necessary data for this study. Students' data were obtained from the Management Information System (MIS) centres for Abdu Gusau Polytechnic, Talata Mafara and College of Education, Maru, Zamfara State only and about 1044 dataset of students were used. The data collected consist of the following features:

Table 1: Initial feature selected

PARAMETER	DESCRIPTION	VALUE
School	School of student	Binary (GP or MS)
Sex	Student gender	Binary (Male or Female)
Age	Student age	Numeric (15 to 22)
Address	Student home address type	Binary (Urban or Rural)
Famsize	Family size	Binary (<=3 or >3)
Pstatus	Parent status	Binary (living together or apart)
Medu	Mother education	0-4
Fedu	Father education	0-4
Mjob	Mother job	Nominal
Fjob	Father job	Nominal
Romantic	With a romantic relationship	Yes or No
Famrel	Family relationship	Numeric: from 1-very bad to 5- Excellent
Free time	Free time after school	Numeric: from 1-very bad to 5- very high
Go out	Going out with friends	Numeric: from 1-very bad to 5- very high
Dalc	Daily alcohol consumption	Numeric: from 1-very bad to 5- very high
Walc	Weekend alcohol consumption	Numeric: from 1-very bad to 5- very high
Health	Current health state	Numeric: from 1-very bad to 5- very good
Absences	Number of student absences	Numeric (0 to 100)
Travel time	Home to school travel time	Numeric
Reason	Reason for choosing the school	Nominal
Studytime	Weekly study time	Numeric
Failures	Number of past class failures	Numeric
Schoolsup	Extra educational school support	Binary (yes or no)
Famsup	Family educational support	Binary (yes or no)
Paid	Tuition paid	Binary (yes or no)
Internet	Internet access to study	Binary (yes or no)

Higher	Wants to take higher education	Binary (yes or no)
Activities	Extra curriculum activities	Binary (yes or no)
Nursery	Attended nursery school	Binary (yes or no)
Guardian	Student guardian	Nominal
Total grades	Total student grade	Numeric (0 to 20)

Step 2: Data pre-processing

Pre-processing plays an important role in data mining. Its purpose is to convert raw data into a suitable form which can be used by mining algorithms. Following tasks are performed in this phase. The parameters used are School, Sex, Age, Address, Family size, Parent status, Mother education, Father education, Mother job, Father job, Romantic, Family relation, Free time, Go out, Dalc, Walc, Health, Absences, reason, guardian, travel time, study time, failures, school support, family support, paid, activities, nursery, higher, internet, Total grades and Grades.The following were used for the pre-processing:

A. Data cleaning

Once the data are collected, what is required is to clean them of inconsistencies, errors and noises before the data is ready for use. Data cleaning is the number one step that needed to be implemented in data pre-processing.

Data integration

This refers to gathering of the data from the multiple sources into single repository. As we have gathered the data from two different schools. Redundancy is one of the common problems that arose when integrating data. The dataset consists of two commas separated values (CSV) files which were taken from MIS repository of AGP T/Mand COE Maru, Zamfara State. These files contained the performance data of various courses which were studied by Students and were integrated into a single file.

B. Data transformation

In this phase, a thorough examination of attributes and their corresponding values was performed to reduce any irregularities in the data, handle missing values in the data, and enhance the reliability of the data. Additionally, the pre-processing phase involved transforming the input data into a form that is preferred by the Machine Learning algorithms. One of the most common methods for transforming input data is z-score standardization which scales different features' values such that they follow a standard normal distribution. Consequently, the transformed features will have a comparable range or scale of measurement, such that none will have more influence than the others on the learning algorithm. As shown in equation (3.1), standardization of an element in feature X involves subtracting the mean value from feature X, and then dividing the outcome by the standard deviation of feature X and this was used in this research.

$$X_{transformed} = \frac{X-\mu}{\sigma} = \frac{X-Mean(X)}{SD(X)} \quad (3.1)$$

C. Data reduction

Some attributes will not deliver sufficient input required for the prediction. Therefore, they need to be eradicated. The eradication is done by selecting the attributes related to data prediction. Many attributes have been derived for the analysis like time spent on social network age, free time, mother education, father education, mother job, father job etc. However, the increase in feature space leads to difficulties for supervised learning. As a result, a high number of features can prompt a decrease in classification accuracy. For this, to select the best attributes for prediction purposes, this research work used correlation which is an important pre-processing step for feature reduction.

D. Data discretization

The discretization mechanism is a technique used to transform the needed data from numerical values into nominal values. Some classifiers are not applicable on continuous data. That is why target attribute which is Total Grade has been converted into nominal.

E. Class balancing

This phase deals with data balancing; data balancing technique is applied after data pre-processing for solving the class imbalance problem. The class imbalanced problem arises when the number of instances in one class is much smaller than the number of instances in another class or other classes. Traditional classification algorithms provide high accuracy for majority classes when data is imbalanced because during classification, they have much intension towards majority class instances and have less

intension for minority class instances. The adjustment of the ratio of two class samples can improve the machine's learning performance. Therefore, we employed a class balancing method known as boxplot in this phase.

Feature selection

Feature selection is the process of selecting a suitable subset of relevant and informative features that can be used to construct a model that can achieve an equal or better accuracy models constructed with the full feature set. While the performance of Machine Learning algorithms is significantly dependent on the quality of the selected features, eliminating redundant or irrelevant features results in a reduction in the training time and computational cost of the Machine Learning algorithms. This is a reduction in the model complexity; in addition to the improvement in the model performance. The feature selection methods can be classified into two main categories:

- Filter method
- Wrapper method.

The datasets were carefully analyzed to identify features that have a greater impact on the output variables. This study applied filter method using information gain-based selection algorithm to evaluate the feature ranks. It is checking which features are most important to build students' performance model. During feature selection, a rank value is assigned to each feature according to their influence on data classification. Filter feature selection methods were utilized to perform this analysis since they are faster and computationally more efficient than wrapper-based methods. The SBF (Selection By Filter) function in Python provides a simple interface that can be used to screen the predictors and selects the optimal feature subset based on univariate statistical methods.

Step 3: Data splitting

After completing the pre-processing task, the datasets were split into two datasets, training and test sets. The training sets were used to construct the models, while the test set were used to evaluate the performance of the models. In this phase, 80% of the data was allocated to the training set, and the remaining 20% was allocated to the test set and this was done using the library called SKLEARN.

Step 4: Models training and tuning

In this phase, multiple models including single classifiers (i.e., Random Forest, Support Vector Machine, Logistic Regression, K-Nearest Neighbour and Naive Bayes) were fit to the training set. For each model, a set of hyperparameters were optimized to identify the best model fit. The algorithms used are discussed below:

Random Forests

Random Forests, as the name suggests, is a group of Decision Trees. Moreover, it is a meta estimator that actually fits a number of various decision tree classifiers that are based on various sub-samples of the dataset and then uses average evaluations to improve the predictive accuracy and the control of over-fitting. In addition to classification, it can also be used for regression. It can successfully create a model despite missing values and also be used for feature engineering.

Logistic Regression

Logistic Regression is a useful algorithm when the output required is categorical in nature. It is based on the logistic or sigmoid function from statistics. The Logistic Regression class from the linear models' package in the scikit-learn library was used to build the model in python.

K-Nearest Neighbors

The K-Nearest Neighbors (KNN) classifier is a type of "lazy" learning algorithm. The algorithm uses the data points to create the model structure. It uses all the data points in the testing phase to determine groups and clusters in the data set. It is highly efficient when the dataset does not follow mathematical theoretical assumptions.

Naïve Bayes

The Naïve Bayes (NB) classifiers are a family of easy to train classifiers, which are powerful in determining the probability of the outcome based on a given set of conditions to the Bayes theorem. In this approach, the conditional probabilities are inverted to represent the data as a function of measurable quantities.

a. The Gaussian model is a Naïve Bayes classifier, which is a continuous distribution characterized by mean and variance.

b. The Bernoulli model is a Naïve Bayes classifier that generates Binary/Boolean indicators, in contrast to the multinomial NB model. The BernoulliNB class from the scikit-learn library was used to build the model in python.

RESULTS AND DISCUSSION

This section discusses the result gotten from the methodology used. This made up of Precision, Recall, F1-score and Support and lastly states the prediction scores.

Data collection and analysis

The data are already collected from two institutions which are AGP Talata Mafara and COE, Maru, Zamfara State and the overall total number of students datasets collected are 1044; 394 data from AGP Talata Mafara represented as (GP) and 650 datasets from College of Education, Maru, Zamfara State indicated as (MS) and consist of 31 features that were later reduced into 25 features after the feature selection.

The number of students was categorized into three which are low (students that performed below average), average (students that performed averagely) and high (students that performed above average). And it was discovered as depicted in Figure 2 that students that performed below average were 90, those that performed averagely were 750 and those that performed above average were 204 making the total of 1044 students. And these were saved in csv file for processing.

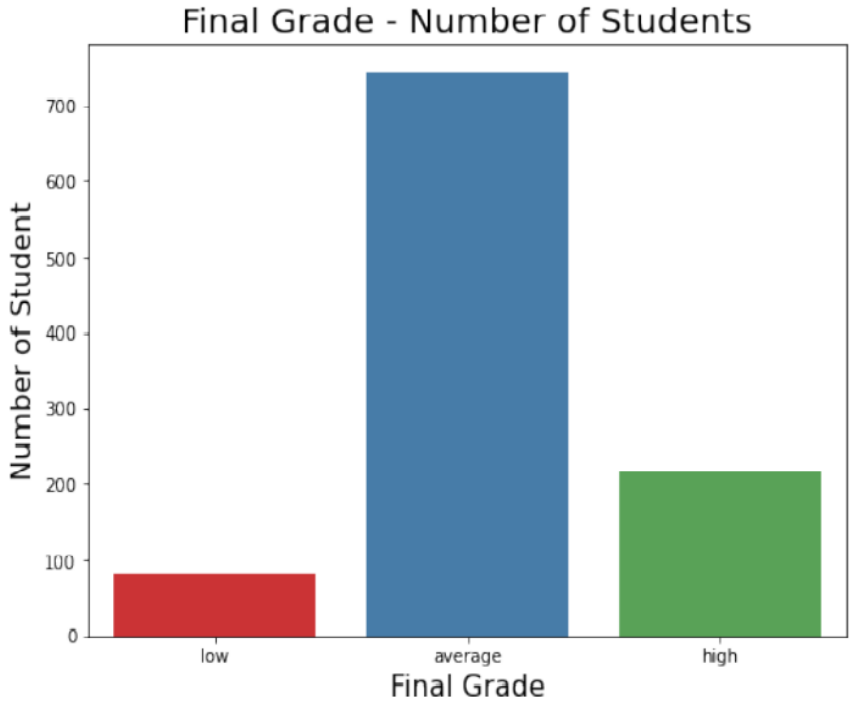

Figure 2: Graph of number of students against final grade

Classification of report

To measure how good a prediction is, we will count how many of the predicted values are equal to the actual values, some of them are positive and some are negative. For binary classification problems, the four important quantities are True Positives, False Positives, True Negatives, and False Negatives. They are defined as follows and use the actual and predicted values:

1. **True Positive**: This is the case where the actual and predicted values were both positive.
2. **False Positive**: This is the case where the actual value was negative but the predicted value was positive.
3. **True Negative**: This is the case where the actual and predicted values were both negative
4. **False Negative**: This is the case where the actual value was positive but the predicted value was negative.

Based on these values, we can generate four main classification metrics called Precision, Recall, F1-score, and Support. The definition of these follow:

1. **Precision**: It measures the proficiency of the classifier to not label negative instances as positive. It indicates how well the classifier labels the positive predictions. The formula for Precision is as follows:

$Precision = TruePositive\ (TruePositive + FalsePositive)$ *(4.1)*

2. **Recall**: It measures the proficiency of the classifier to predict all the positive instances. It indicates how many correct positive labels are assigned by the classifier. The formula for Recall is as follows:

$Recall = TruePositive\ (TruePositive + FalseNegative)$ *(4.2)*

3. **F1-score**: It is an accuracy measure that utilizes a combination of Precision and Recall. It is a harmonic or weighted average of Precision and Recall where the $F1$ score is between 0 and 1. It is denoted by the following formula:

$F1score = 2 * (Precision * Recall)\ (Precision + Recall)$ *(4.3)*

4. **Support**: It is the total number of occurrences of each label in the actual values. It is the number of samples of true responses that lie in that particular class and is to measure imbalances in the dataset.

For this research work, the result obtained for the precision, recall, F1-score and Support for each of the classifiers i.e. Random forest, Support Vector Machine, Linear Regression, K-Nearest Neighbour and Naïve Bayes is shown in the table .5

Table1 shows that for Random forest classifier, precision for Average performance students, High performance students and Low performance students are 0.79, 0.67 and 0.40 respectively. For Recall, it shows that High performance students and Low performance students are 0.93, 0.42 and 0.11 respectively. For F1-score, High performance students and Low performance students are 0.85, 0.52 and 0.17 respectively. For Support, High performance students and Low performance students are 153 , 38 and 18 respectively.

Table 2: Result for random forest classifier

Random Forest Classifier Report:

	precision	recall	f1-score	support
average	0.79	0.93	0.85	153
high	0.67	0.42	0.52	38
low	0.40	0.11	0.17	18
accuracy			0.77	209
macro avg	0.62	0.49	0.51	209
weighted avg	0.73	0.77	0.73	209

Table 2 shows that for Support Vector Machine classifier, precision for Average performance students, High performance students and Low performance students are 0.73, 0.00 and 0.00 respectively. For Recall, it shows that High performance students and Low performance students are 1.00, 0.00 and 0.00 respectively. For F1-score, High performance students and Low performance students are 0.85, 0.00 and 0.00 respectively. For Support, High performance students and Low performance students are 153 , 38 and 18 respectively.

Table 3: Result for SVM

Support Vector Classifier Report:

	precision	recall	f1-score	support
average	0.73	1.00	0.85	153
high	0.00	0.00	0.00	38
low	0.00	0.00	0.00	18
accuracy			0.73	209
macro avg	0.24	0.33	0.28	209
weighted avg	0.54	0.73	0.62	209

Table 3: shows that for Linear Regression classifier, precision for Average performance students, High performance students and Low performance students are 0.77, 0.69 and 0.40 respectively. For Recall, it shows that High performance students and Low performance students are 0.95, 0.29 and 0.11 respectively. For F1-score, High performance students and Low performance students are 0.85, 0.41 and 0.17 respectively. For Support, High performance students and Low performance students are 153 , 38 and 18 respectively.

Table 4: Result for logistic regression classifier

Logistic Regression Classfier Report:

	precision	recall	f1-score	support
average	0.77	0.95	0.85	153
high	0.69	0.29	0.41	38
low	0.40	0.11	0.17	18
accuracy			0.76	209
macro avg	0.62	0.45	0.48	209
weighted avg	0.72	0.76	0.71	209

Table 4 shows that for K-Nearest Neighbour classifier, precision for Average performance students, High performance students and Low performance students are 0.76, 0.41 and 0.00 respectively. For Recall, it shows that High performance students and Low performance students are 0.90, 0.29 and 0.00 respectively. For F1-score, High performance students and Low performance students are 0.83, 0.34 and 0.00 respectively. For Support, High performance students and Low performance students are 153 , 38 and 18 respectively.

Table 5: Result for KNN classifier

K-Nearest Classifier Report:

	precision	recall	f1-score	support
average	0.76	0.90	0.83	153
high	0.41	0.29	0.34	38
low	0.00	0.00	0.00	18
accuracy			0.71	209
macro avg	0.39	0.40	0.39	209
weighted avg	0.63	0.71	0.67	209

Table 5 shows that for Random forest classifier, precision for Average performance students, High performance students and Low performance students are 0.85, 0.24 and 0.22 respectively. For Recall, it shows that High performance students and Low performance students are 0.18, 0.97 and 0.28 respectively. For F1-score, High performance students and Low performance students are 0.30, 0.39 and 0.24 respectively. For Support, High performance students and Low performance students are 153 , 38 and 18 respectively.

Table 6: Result for Naïve Bayes classifier

Naive Bayes Classifier Report:

	precision	recall	f1-score	support
average	0.85	0.18	0.30	153
high	0.24	0.97	0.39	38
low	0.22	0.28	0.24	18
accuracy			0.33	209
macro avg	0.44	0.48	0.31	209
weighted avg	0.68	0.33	0.31	209

Final prediction of result

Table 7: The prediction performance score, shows for each of the classifier and random forest prediction score shows the highest value which is 76.56%.

S/N	**CLASSIFIER**	**PREDICTION SCORE**
1.	Random Forest	76.56
2.	Support Vector Machine	73.21
3.	Logistic Regression	75.60
4.	K- Nearest Neighbour	71.29
5.	Naïve Bayes	33.49

CONCLUSION

To predict student's performance in advance is a very important issue. We conclude from findings of the study that various datasets of student provide different results with different attributes. This is the reason why the results are varied with different evaluation measures like precision, F1- score, Recall and support. It is concluded after this study that every algorithm result is varied according to the dataset and variable attribute used for prediction. However, if we use the random forest, support vector machine, Linear Regression, K-nearest neighbour and Naïve Bayes, according to our requirements these algorithms provide extra ordinary accurate results for future prediction and help in the betterment of education system. However, Random Forest algorithm performed better which resulted into 76.56% while the least was Naïve Bayes which resulted to 33.49%.

ACKNOWLEDGEMENTS

My profound gratitude goes to Almighty Allah Who in his infinite mercy and grace grants me the opportunity to enrol my M.Sc. program. I sincerely to express my sincere appreciation to my humble and hardworking major supervisor, Prof. Ismail Gboyega Baoku for his time, suggestions, contributions, guidance and corrections to ensure the success of this research work. Permit me to also thank my minor supervisor, Mr. Bashir Ahmed Jamilu for his assistance throughout the research work. I am also indebted to the HOD, Dr. Umar Iliyasu for his guidance and support and also like to express my best regard to my able Lecturers, Dr G.N. Obunadike, Dr. Oloruwaju, M.O., Mrs Faith O. Echubo, Mr Abubakar Ahmad, Mr Muktar Abubakar Gache, Hunu Suleman, Mr Ili Jiya, Mr Aminu Bashir and Mr Yusuf Surajo for their tremendous experience and knowledge giving to me. Finally, my special thanks go to my lovely Parents (Alh. Abubakar Aliyu and Haj. Baraka) who cares, prayers and support me during my program and equally my colleagues, for their great advice and contributions toward my MSc program completion. Thanks to you all.

REFERENCE

1. HUSSAIN, S., MUHSIN, Z.F., SALAL, Y.K., THEODOROU, P., KURTOĞLU, F. & HAZARIKA, G.C. (2019). Prediction Model on Student Performance based on Internal Assessment using Deep Learning.*International Journal of Emerging Technologies in Learning*, **14**(8): 4-22.
2. WITTEN, H.I., EIBE, F., MARK A. HALL & CHRISTOPHER J. PAL. (2016). Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann.
3. HUSSAIN, S., MUHSIN, Z.F., SALAL, Y.K., THEODOROU, P., KURTOĞLU, F. & HAZARIKA, G.C. (2019). Prediction Model on Student Performance based on Internal Assessment using Deep Learning.*International Journal of Emerging Technologies in Learning*, **14**(8): 4-22.
4. BUENAÑO-FERNÁNDEZ, D., GIL, D. & LUJÁN-MORA, S. (2019). Application of Machine Learning in Predicting Performance for Computer Engineering Students: A Case Study. *Sustainability*, **11**(10): 2833-2851
5. NYCE, C. (2007). A Predictive analytics white paper. American Institute for CPCU. Insurance Institute of America, 9-10.
6. SAS, S. (2017). Predictive Analytics: What it is and why it matters. https://www.sas.com/en_us/insights/analytics/predictive-analytics.html. Retrieved April 24, 2017.
7. SUJITH, J., SANGEETHA, K. & JAIGANESH, V. (2020). Predicting Students Academic Performance using an Improved Random Forest Classifier, International Conference on Emerging Smart Computing and Informatics (ESCI) AISSMS Institute of Information Technology, Pune, India.
8. RAVINA, A., PRANAV D., ALAMEEN K., FATHIMA, R. & SRIDHARAN, G. (2021). Educational data mining for predicting students' academic performance using machine learning algorithms, International Conference on Sustainable materials, Manufacturing and Renewable Technologies 2021.Elsevier Ltd. All rights reserved. https://doi.org/10.1016/j.matpr.2021.05.646 2214-7853/ 2021