International Journal of Development Mathematics Vol 1 Issue 3, Page No. 196 - 208

INTERNATIONAL JOURNAL OF DEVELOPMENT MATHEMATICS

ISSN: 3026-8656 (Print) | 3026-8699 (Online)

journal homepage: https://ijdm.org.ng/index.php/Journals

Comparative Analysis of K-Means and Naïve Bayes Algorithms for Predicting Students' Academic Performance

Mohammed M. Nasir[a*], Ahmadu A. Sandra[b] and Comfort Williams[a]

[a]*Department of Computer Science Adamawa State Polytechnic, Yola*

[b]*Department of Computer Science Modibbo Adama University, Yola*

ARTICLE INFO

Article history:
Received 17 May, 2024
Received in revised form 05 August, 2024
Accepted 23 August, 2024

Keywords:
Academic Performance, K-Means, Naïve Bayes, Prediction, Students.

MSC 2020 Subject classification:
60-02, 62-02, 62-04, 68T10, 68T45

ABSTRACT

A student's performance is a victory statistic in higher education. The university's exceptional academic record strengthens its position as one of the prerequisites for a prestigious university. Teachers need to forecast and analyze student performance to pinpoint areas of weakness and improve academic standing. In academic settings, Coordination of computational tactics to improve workforce management and academic attainment is achieved through Educational Data Mining (EDM), a theory-based approach. Classification is a broadly connected technique in forecasting student performance based on diverse criteria. Machine learning algorithms are fundamental to knowledge disclosures, permitting precise performance projection and early student-identifiable proof. This study examines how well students perform academically using the Naïve Bayes classifier (NBC) model and the K-Means clustering approach. From supervised and unsupervised machine learning, two (2) algorithms with comparable operational capacity were selected. The labeled classes in the classifier correspond to the grades in the dataset. Records were gathered from 178 students (400 levels) in Adamawa State University Mubi's computer science departments in the 2022–2023 academic sessions. The training and testing sets of the dataset are divided into two groups, each with a percentage ratio of 30% and 70%. According to the results, the Naïve Bayes model has an accuracy of 92.6%, while the K-Means model has an accuracy of 38.9%.

1. Introduction

The value of education in any country cannot be overstated. In a country with a strong economy, the quality of education is essential to promoting global economic progress (Aceredo-Dugue, 2023). A key measure of educational progress is students' academic achievement, which is impacted by various factors including age, gender, the composition of the teaching team, and learning. Predicting academic success has drawn more attention in terms of teaching and learning in education. Data mining (DM) is the process of extracting knowledge from enormous volumes of historical data. DM is a technique for getting important and useful information out of a database that may be used in the classroom and other settings. Educational data mining is used to build techniques for knowledge extraction from data in educational environments. In today's Nigeria educational system bases a large portion of students' success on test assessments, homework, attendance, assignments, practical's, and quizzes which are categorized as Continuous Assessment (CA) and their final examination. A minimum threshold for promotion is reached after these activities; as a result, it is critical to pinpoint the causes of success or failure. This will allow teachers to provide more focused counselling that addresses these factors rather than mere grades, as these factors impact the minimum criteria for promotion. Hence, a model for forecasting student's academic performance is of great importance, therefore, data mining techniques for classification and clustering are employed in this research study to anticipate the academic performance of students. A solid academic record boosts a university's reputation and increases job prospects for students because administrators may consider it to be a critical component (Alsariera *et al.,* 2022). Academic records are a means of evaluating students' accomplishments as well as the quality of the schools they attended. In the fiercely competitive world of academia, excellence in student performance is important for higher education institutions. (Ghorbani and Ghosi, 2020) defined student performance as assessed by comparing a certain student learning

* Corresponding author. Tel.: +2347061547436
E-mail address: nas9ja@gmail.com (Mohammed M. N.)
https://doi.org/10.62054/ijdm/0103.15

evaluation to Ponder instructional modules, Grade Point Average (CGPA), or the final scores, while (Nachouki and Abou, 2022) defined the students' academic success as the likelihood of gaining a long-term degree.

Weak student performance in the classroom is one of the biggest problems facing universities worldwide. The fact that a few factors affect students' achievement is one reason why it has become difficult to address this issue. Five components are noted by (Tejedor and Garcña-Valcñrcel, 2007): mental, scholastic, educational, socio-family, and identifiable proof. Different outcomes of penniless academic execution are likewise possible, even when accounting for scholastic consistent loss. According to (Vicerrectorado AcadÃmico, 2017), a major contributing factor to academic desertion is substandard performance in the classroom. According to (Viale, 2014), in similar circumstances, the proportion of students who do not pass classes at the beginning of the academic year is often high; yet, many students decide to drop out of college when they retake a lesson and fail it again. Additionally, the evaluation of these institutions and the success rates of their students are vital because they are regarded as quality indicators for instructional education.

Technological advancements have enabled educators to employ information mining as well as explanatory methods to analyse massive databases pertaining to designs relating to the behaviour and learning of their students (Shah, 2022). Data mining is essential for sorting through an enormous amount of data to discover significant data, which supports decision-making. Data mining has numerous critical applications within the field of instruction (Delavari *et al.,* 2008). One strategy that educational institutions might employ to identify hidden patterns in educational data, broaden their understanding, as well set expectations for future student accomplishments is educational data mining (EDM) (Baashar *et al.,* 2021).

Machine learning (ML) techniques provide the tools for information extraction from data, while EDM is used to locate information within data. Generally speaking, machine learning (ML) looks at calculations that result from cases that are provided remotely (the input set) to develop common theories that predict future occurrences (Fly *et al.,* 2017). To identify patterns in data and apply those patterns to forecast outcomes, machine learning examines data. By utilizing ML in the classroom, educators will be able to identify the fundamental elements that impact a student's success. Additionally, machine learning (ML) will enable educators to identify kids who are not performing up to par, enabling them to make informed decisions.

Limitations of traditional method of learning

i. The traditional method of learning is often limited to geographical and physical constraints making it difficult to reach a large number of students it's not scalable in terms of population size which can hamper the performance of students.
ii. The human error element can lead to inaccuracy in result compilation due to population size.
iii. Manual processing can be slow during teaching and also the complying results of students

Potential Benefit of a proposed system using machine learning techniques

i. Machine Learning can handle large datasets of students and performance analysis can be carried out in other to predict their performance.
ii. Automated processing enables rapid analysis and decision-making of their performance.
iii. Machine learning can identify complex patterns and relationships in students' data leading to new insight on how to engage the students.

In this paper, we primarily focus on utilizing student data to predict academic achievement. This student dataset includes the response of Adamawa State University Mubi students to inquiries on their demographics, behavioural traits, and academic performance for the 2022–2023 academic sessions. In this paper, K-Means (KM) and Naïve Bayes (NB), machine learning methods were compared.

2. Literature Review

(Yauri *et al.,* 2023) researched predicting students' performance with artificial neural networks (ANN). By using Artificial intelligence (AI) techniques, the study produced a demonstration that forecasts student success and failure rates. In Adamawa State, Nigeria, the study looked at 720 students from three selected postsecondary institutions. From Adamawa State Polytechnic one hundred (120) students were chosen, from Adamawa State University three hundred (300) students were chosen, and from Modibbo Adama University, Yola three hundred (300) students were chosen. Descriptive measurements are used in the research to determine the elements that most likely influence students'

academic achievement. Jupyter Notebook, a Python Anaconda development environment, was used to pre-process, clean, and model the acquired data to develop a model that will forecast students' academic performance. Twelve (12) input elements in the neural network model consist of one yield layer and two hidden neuronal layers. The dataset is prepared using the back-propagation learning algorithm.

The neural network's performance is assessed through the application of cross-validation using k-fold. The neural network has an impressive 97.36% accuracy. To better prepare students for universities, (Kumar *et al.,* 2020) examined their performance on specialized exams used for university applications. The approach to the problem was to predict whether or not students would pass a specialized exam based on how well they performed in particular course subjects. Data about 200 K L University students were used in the study between 2013 and 2017. The two unique methods utilized were hierarchical clustering and K-means clustering to sort and classify the students. The authors employed the Naïve Bayes method to forecast with 72% accuracy after classifying the data. Daud *et al.* (2017) examined the causes of many students' career dropouts in their research. To determine whether a particular student would finish their education, the study took into account the individual, family, and economic factors that have the most effects on student achievement. The dataset is cleansed before testing to extract 50 students who completed their coursework and 50 who did not. The F1 score for the Naïve Bayes method was 84.8% as part of its results. A machine learning-based algorithm was presented by Ahmad and Shahzadi (2018) to determine whether or not students' academic performance was random. Making use of the students' study habits, learning aptitudes, and academic interaction traits, they were able to classify the data with 85% accuracy. The analysis concludes that the proposed model can be used to determine academically unsuccessful students.

Bernacki *et al.* (2020) conducted research to determine if the learning administration framework's log records alone could accurately predict achievement. He concluded that 75% of the people who would need to repeat a course were accurately predicted by the behaviour-based forecast. Furthermore, he expressed that students who will struggle academically wise the following semesters might be identified and assisted with the help of this approach a model utilizing an artificial neural network built by Waheed *et al.* (2020) in light of student records about their path through the LMS. It seems that student clickstream behaviours and demographics have a significant impact on academic success. Pupils who participated in course exploration outperformed others. There was no correlation between the cooperation of students in the classroom and their performance. Regardless, he concluded that a deep learning model could be a crucial instrument in forecasting students' academic performance.

According to Hasan *et al.* (2019), using artificial intelligence to predict student performance can help students avoid poor grades and prepare them for upcoming tests. Teachers can provide students with appropriate guidance by differentiating between circumstances and course prerequisites. Teachers can use the framework to screen students and provide individualized assistance to help reduce students' laziness. With a 94.88% accuracy rate, the research benefits teachers and students alike. According to Alamri *et al.* (2020), a variety of factors could influence the performance of students in the most recent exam. The study predicts final grades in scientific and Portuguese linguistic courses using the Support Vector Machine (SVM) and Irregular Woodland (RF) algorithms. It appears that a 93% accuracy rate is achieved by twofold classification, whereas In Random Forest (RF), regression has the lowest RMSE (1.13). Early preparation can assist educational establishments in creating strategies for students who perform well, by further improving their academic outcomes. The study recommends that informative organizations operate more effectively.

There are many prediction models available today with different approaches to students' performance; the reviews of the literature reveal that most prediction models are based mainly on continuous assessment and final examinations and a few on behavioural factors. The Data Mining Techniques that are chosen in most of the reviews are from supervised and unsupervised learning with no similar operational properties to obtain an accuracy level. This study will design a predictive model that factors the relationships between students' behavioural factors, social factors, family factors, and academic factors, the model will be developed using the k-mean clustering method and Naïve Bayes classifier in Python Jupyter notebook as Data Mining Techniques to analyse, this is to show that behavioural factors as independent variables while the grades as dependent variables.

The Data Mining Techniques chosen for this study have similar operational properties to see if based on their similar operational properties there will be less bias between both algorithms in terms of performance and the accuracy level will increase by close merge. The result obtained from the study is validated using a cross-validation technique to show which algorithm is more accurate and efficient.

3. Materials and Method

The purpose of this research is to forecast students' academic success using K-means and Naïve Bayes. The result of this work will help the educators/faculty to improve the teaching approach constructively. In addition, the teachers could observe students' achievements and also have an idea of how to help the weak ones. The data Normalized are data consisting of rows and columns. There are thirty-one (31) columns of attributes and one hundred and seventy-eight (178) rows of instances in the dataset used. The 31 columns of attributes are gender, Age, Address, State, Occupation, Parent marital status, Mother education level, Mother job, Father education, Father job, Sponsor, Family size, family relationship, Reason for choosing school, Time taken to class, Study time, Past failure, Believe, School support, Family support, Extracurricular activities, Internet usage, Romantic relation, Free time, Class absence, Outing, Course understanding, Course Interest, Lecturer relation, Grade 1 and Grade 2. Emphases are given to attributes that are classroom-related during analysis. K-mean clustering method and Naïve Bayes classifier are used on all 178 instances grouping them into classes.

All necessary libraries in Python were used to solve the Naïve Bayes and K-mean algorithms based on the instances in the dataset and the results are used to compare the performance of each in WEKA.

K-mean mode of operation

i. Choose k of cluster
ii. Select at random k points that will be the centroid (not necessary from the data points).
iii. Assign each data point to the closest centroid that forms k clusters.
iv. Compute and place a new centroid of each cluster.
v. Reassign each data point to the new closest centroid. If any reassignment takes place, go to step (iv) otherwise go to finish.

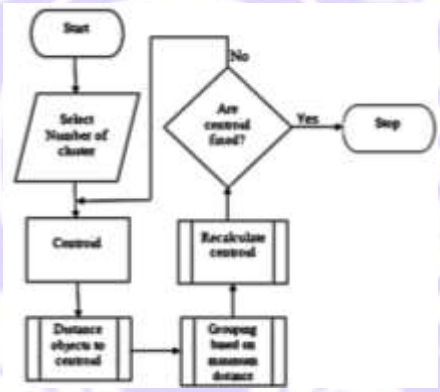

Figure 1: Flowchart of k-mean Clustering Operation.

Naïve Bayes mode of operation

Naïve Bayes is a simple and powerful algorithm for predictive analysis; it is a classification technique that uses Bayes theorem with an assumption of independence among predictors. In simple terms Naïve Bayes classifier assumes that the presence of a particular feature in a class is unrelated to any other feature even if these feature depends on each other or upon the existence of the other feature, all of these properties independently contribute to the probability. The Naïve Bayes model is easy to build and very useful to large datasets. Naïve Bayes classifier works on the principles of conditional probability as given by Bayes theorem.

Given a hypothesis c and evidence, Bayes theorem states that the relationship between the probability of the hypothesis before getting the evidence P(c) and the probability of the hypothesis after getting the evidence P(c/) is

$$P(c/x)=\frac{P(x/c)P(c)}{P(x)},\ P(x)\neq 0 \quad (1)$$

where,

$P(c/x)=$ Conditional probability of c given x

$P(x/c)=$ Conditional probability of x given c

$P(c)=$ Probability of c

$P(x)=$ Probability of x

3.1 Study area

This study is conducted at Adamawa State University, Mubi, Nigeria. The dataset was gathered at Adamawa State University, Mubi, and attributes include student behavior demographic, family demographic, social demographic, and academic demographic. All the data are obtained from student interviews and questionnaires. During data collection, a dataset with one hundred and seventy-eight (178) instances was generated

3.2 Method of data collection

In this study, a structured questionnaire is used to collect data, copies of the questionnaire are distributed (online) directly to respondents of the Computer Science department and enough time was given to the respondents to fill the questionnaires administered to them online. Primary as well as secondary sources are used, and the primary data is sourced through structured questionnaires. The questionnaire is broken down into several sections that are pertinent to the following: grades of the student, family support system of a student, engaged time by the student, interest of the student, and study habits of the student. The questionnaire comprises four (4) sections.

3.3 Method of data analysis

This study develops a model in Python to help analyze data in the Naïve Bayes classifier and K-mean algorithm in other to develop patterns for a predictive model.

3.4 Data preprocessing

The sample data, which has numerous columns, numerous rows, and some missing values, is presented as an Excel or CSV file. The most important step in developing a data culture and leveraging data to generate precise projections is cleaning the data. Information is preprocessed and sanitized. To prepare data for use, data cleaning involves deleting or altering information that is inaccurate, missing, unnecessary, duplicated, or formatted wrongly. Another critical stage in the creation of deep machine learning or machine learning algorithms is preprocessing data. Both the quality and quantity of data are increased and the noise decreased.

Procedure for the proposed model

Data Identification: Finding the relevant data required for the model is the first stage in data identification. This entails comprehending the issue that needs to be resolved, locating pertinent data sources, and gathering or gaining access to the required data. Ensuring the data collected is both large enough and of high enough quality for the modeling assignment is essential.

Data Preprocessing: To clean up and substitute meaningful values for missing values, data preprocessing is necessary. The next step after obtaining the data is preprocessing. The data is cleansed and ready for modeling in this place. This includes encoding categorical data, adjusting or scaling features, addressing missing values, and handling outliers. Making sure the data is in a format that is appropriate for training the machine learning model is the aim.

Training Set: The model is trained using a subset of the dataset, known as the training set. It is a portion of the data that the model is trained on. The model uses the training set, which typically comprises of target values (dependent variables) and input features (independent variables), to identify patterns and correlations in the data. 70% of the dataset was utilized in this instance.

Train Model: The training procedure begins after the model and training set are in place. The model learns from the input data during training by modifying its parameters to minimize a selected loss function. By identifying underlying patterns in the training data, the objective is to improve the model's performance.

Testing Set: A different subset of the dataset apart from the training data is the testing set. It is employed to assess the trained model's ability to generalize to fresh, untested data. The testing set aids in evaluating the model's efficacy and its capacity to forecast correctly on data that was not used in training. Here, thirty percent of the data were utilized.

Prediction: After training and testing, the model is used for new, untested data to draw conclusions or make predictions. When fresh data is added to the trained model, it will make predictions based on the lessons it has learned during training.

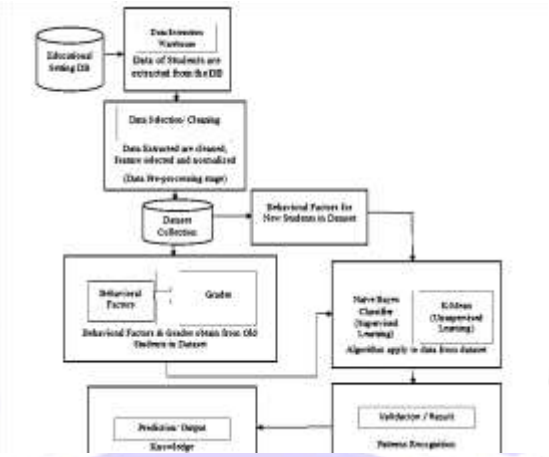

Figure 2: Model of the Proposed System

Step 1: Extraction of Students Data from the institution Database (DB)
Step 2: Data cleaning, feature selection, and normalization are carried out on the data extracted.
Step 3: Create a Dataset based on students' behavioural and grade evaluation for existing students and new intake students.
Step 4: Evaluate new student's based on just behavioural factors while existing students on both behavioural factors and grades in Step 3.
Step 5: Apply Algorithms to Step 4.
Step 6: Validate results
Step7: Predict output

3.4 Dataset

A dataset is a collection of data organized into a table, with each column representing a different variable. It is primarily used for analysis.

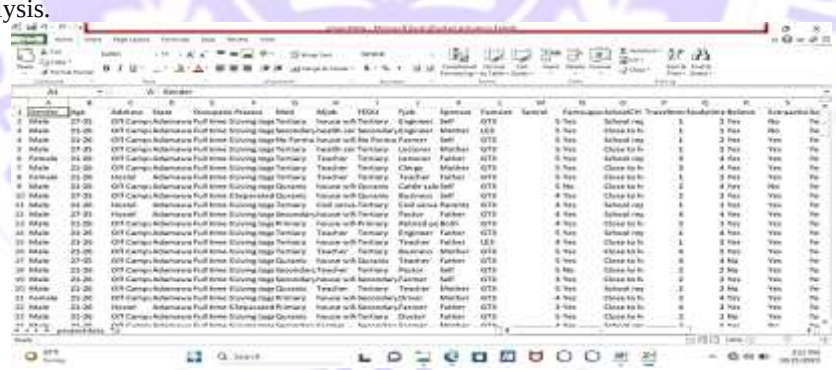

Figure 3: Screenshot of the Dataset for Students

The figure above displays a screenshot of the student's dataset utilized for this study. The data are stored in an Excel sheet with a. CSV file (C|omma-Separated Values) format. A total of one hundred and seventy-eight (178) respondents with thirty-one (31) attributes helped determine students' academic performance.

4. Results

Table 1 Pre-processing dataset

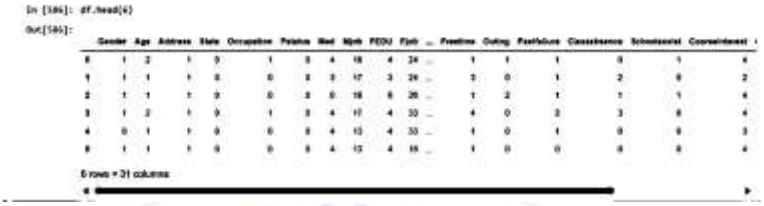

Table 1 show all the data in the dataset has been converted to a usable format, which is an integer, to obtain an accurate output. Some columns that have little or no effect on student academic performance are dropped using the syntax “df=df.drop (["column_name], axis=1). Those attributes that were not classroom-related and would not affect the activities of the classroom, like relationship, state, address, belief, school choice, and extracurricular activities, were dropped because they were insignificant to the outcome of a student's performance.

4.1 Training and testing set

The training and testing sets of the dataset are divided into percentage ratios of 70% and 30%, respectively. The data with a large percentage of 70% is used as a training set (124 respondent’s data), while a lower percentage ratio of 30% is used as a testing set (54 respondent’s data), and a holdout/cross-validation of k-fold is employed. "x_train, x_test, y_train, y_test=train_test_split(x, y, test_size=0.3, random_state=10)" is the syntax for the training and testing sets.

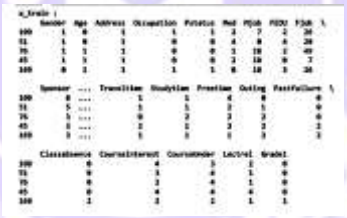

Figure 4: Screenshot of x_train

Figure 5: Screenshot of x_test

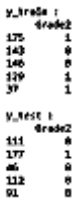

Figure 6: Screenshot of y_train and y_test (class)

4.2 Naïve Bayes and K-means analysis on dataset

On the dataset, the following is the result of the accuracy of the Naïve Bayes classifier:

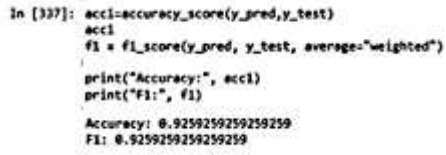

Figure 7: Screenshot of Naïve Bayes accuracy and F-measure of the dataset

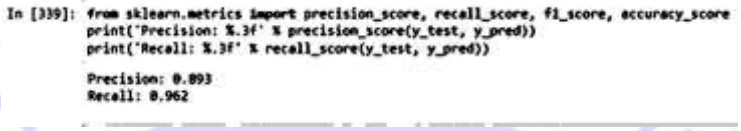

Figure 8: Screenshot of Naïve Bayes Precision and Recall of the prediction

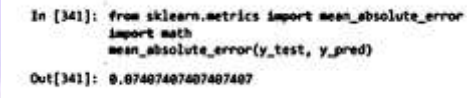

Figure 9: Screenshot the Mean Absolute Error of the prediction

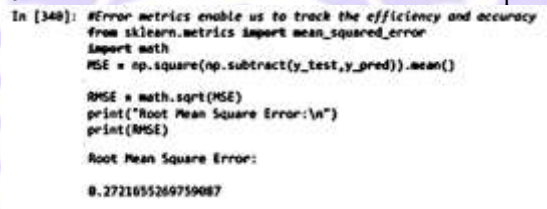

Figure 10: Screenshot RootMean Square Error of the prediction

Table 2: Bayes Result

Algorithm	Accuracy	F1	Precision	Recall
Naïve Bayes	0.926	0.926	0.893	0.962
Percentage (%)	92.6%	92.6%	89.3%	96.2%

Table 2 above shows that the Naïve Bayes algorithm has an accuracy of 92.6%, an F1 of 92.6%, a precision of 89.3%, and a recall of 96.2%. The Naïve Bayes algorithm model will make approximately 93% correct instances prediction and 7% incorrect instances prediction, which means it predicts 165 respondents on the dataset have a CGPA greater than 2, while 13 respondents will have a CGPA less or equal to 2.

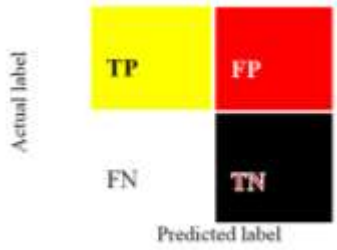

Figure 11: Different representation of a confusion matrix outcome

The F1 offers a balance between recall and precision since it is the harmonic mean of both.

$$F1 = \frac{2 \times Precision * Recall}{Precision + Recall} \quad (2)$$

where,
F1 = F measure score

Precision tells how much of the correctly predicted outcome turned out positive, and it's obtained by dividing the proportion of instances that are true of a class by the total instances of the class.

$$Precision = \frac{True\ positive\ (Tp)}{true\ positive(tp) + false\ positive(fp)} \quad (3)$$

where,
Tp = True positive of the prediction outcome
Fp = False positive of the prediction outcome

Recall tells how much the actual positive cases were able to be predicted correctly by the model, and it is obtained by dividing the proportion of instances that are true of a class by the actual total and false negative of the class.

$$Recall = \frac{True\ positive(Tp)}{true\ positive(Tp) + false\ negative(Fn)} \quad (4)$$

where,
Tp= True positive of the predicted outcome
Fn = False negative of the predicted outcome

$$Accuracy = \frac{True\ positive(Tp) + True\ negative(Tn)}{True\ positive\ (Tp) + false\ positive(Fp) + True\ negative(Tn) + false\ Negative(fn)} \quad (5)$$

where,
Tp = True positive of the predicted outcome
Tn= True negative of the predicted outcome
Fp= False positive of the predicted outcome
Fn = False negative of the predicted outcome

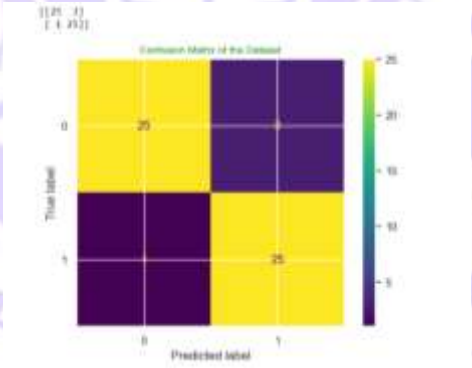

Figure 12: Screenshot of the confusion matrix

The information for the confusion matrix predicted label is given as follows:

True positive (Tp)= 25
False positive (Fp)= 3
False negative (Fn)= 1
True negative (TN)= 25

The above means 25 true positives means 25 cases in the model correctly predicted that the students had GT2

(> 2) grades in their academic performance, 3 false positives means 3 cases where the model incorrectly predicted the students had GT2 (>2) grades in their academic performance; 25 true negatives means 25 cases where the model correctly predicted that the students had LE2 (<= 2) grades in their academic performance; and 1 false negative means 1 case where the model incorrectly predicted that the students had LE2 (<=2) grades in their academic performance.

The outcome of the accuracy of K-means clustering on the dataset is observed as follows:

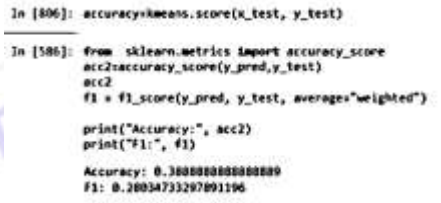

Figure 13: Screenshot of K-means accuracy and F-measure of the dataset

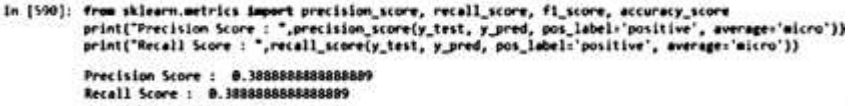

Figure 14: Screenshot of K-means Precision and Recall of the prediction

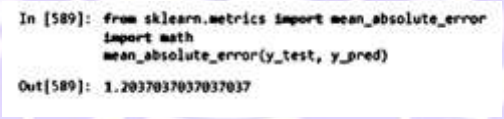

Figure 15: Screenshot the Mean Absolute Error of the prediction

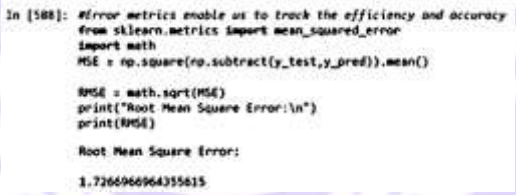

Figure 16: Screenshot RootMean Square Error of the Prediction.

Table 3: K-means Result

Algorithm	Accuracy	F1	Precision	Recall
K-means	0.389	0.389	0.389	0.389

Table: 3. above shows that the K-means algorithm has an accuracy of 38.9%, an F1 of 38.9%, a precision of 38.9%, and a recall of 38.9%.

The K-means algorithm model will make approximately 39% correct instances prediction and 61% incorrect instances prediction, which means it predicts that 69 respondents on the dataset have a CGPA greater than 2, while 109 respondents will have a CGPA less or equal to 2. Based on this, the K-mean model will be a weak algorithm for predicting students' academic performance.

4.3 Comparison between Naïve Bayes and K-means

The results of both algorithms were compared to determine which was more accurate and efficient in predicting students' academic performance.

Table 4: Comparison of Naïve Bayes and K-means Result

Algorithm	Accuracy correct instances	Incorrect instances	Percentage (%) correct instances	Percentage (%) incorrect instances	Total No, of Respondents with correct instance	Total No, of Respondents with incorrect instance
Naïve Bayes	0.926	0.074	92.6%	7.4%	165	13
K-means	0.389	0.611	38.9%	61.1%	69	109

Table 4 above shows that the Naïve Bayes algorithm is more accurate and efficient compared to the K-means algorithm because, for every prediction made by the K-means model, more than half of it will be incorrect.

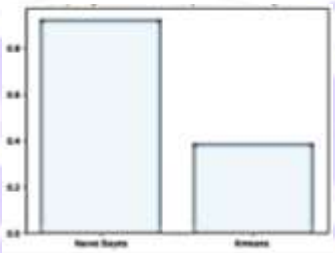

Figure 17: Screenshot Graphical Representation of the Comparison between Naïve Bayes and K-means.

5. Discussion

This paper proposes a machine learning algorithm-based application to forecast the academic success of students in the computer science department at Adamawa State University Mubi's. To predict students' academic achievements, the algorithms K-Means and Naive Bayes were computed and compared in display form. This investigation focused on two factors. The anticipation of academic success based on the data gathered on the students was the main parameter. The second involved comparing the machine learning calculations' execution markers.

In data mining, performance analysis of results based on learning is a system that aims for excellence at several levels and in various dimensions. This study explores students' academic performance using both the Naïve Bayes classifier and the K-means cluster. The grades in the dataset are the labelled classes in the classifier. The dataset is split into training and testing sets; the two parts have a percentage ratio of 70% and 30%, respectively. Python Jupyter notebook and Weka were used to evaluate both Naïve Bayes and K-means, and in both validations, the Naïve Bayes showed more accuracy and efficiency in predicting students' academic performance. The results shows that Naïve Bayes has an accuracy of 92.6%, an incorrect instances of 7.4%, a total number of respondents with correct instances of 165 (predicted students with CGPA greater than 2) and a total number of respondents with incorrect instances of 13 (predicted students with CGPA Less or equal to 2), with a Recall of 96.2%, F1 value of 92.6% and a Precision 89.3% while for K-means the model showed an accuracy of 38.9%, an incorrect instances of 61.1%, a total number of respondents with correct instances of 69 (predicted students with CGPA greater than 2) and a total number of respondents with incorrect instances of 109 (predicted students with CGPA Less or equal to 2), with a Recall of 38.9%, F1 value of 38.9% and a Precision 38.9% while Weka showed that Naïve Bayes has an accuracy of 83.1%, an incorrect instances of 16.9%, a total number of respondents with correct instances of 148 (predicted students with CGPA greater than 2) and a total number of respondents with incorrect instances of 30 (predicted students with CGPA Less or equal to 2), with a Recall of 92.8%, F1 value of 89.5% and a Precision 86.5% while for K-means Weka showed an accuracy of 48%, an incorrect instances of 52%, a total number of respondents with correct instances of 85 (predicted students with CGPA greater than 2) and a total number of respondents with incorrect instances of 93 (predicted students with CGPA Less or equal to 2), with a Recall of 48%, F1 value of 48% and a Precision 48%

6. Conclusion

The study investigates two (2) algorithms of different categories, one a classifier algorithm (Naïve Bayes) and the other a clustering algorithm (K-means). The idea is to help all stakeholders in the educational sector improve the outcome of students' academic performance and also help lecturers improve the way they teach by adapting new methods of tutoring their students to improve the outcome of students' performance. Weka and Python Jupyter notebook was used to evaluate both Naïve Bayes and K-means on the dataset, and in both validations, the Naïve Bayes showed more accuracy and efficiency in predicting the students' academic performance. It was found that Naïve Bayes gives a better accuracy level for the training and testing set of the dataset than K-means. It is found that the Naïve Bayes algorithm is best suited for the model based on the lower mean absolute error and root mean square error than k-means. The management of the institution can benefit from analysing and evaluating the results of the model for the decision-making process.

References:

Acevedo-Duque, Á., Jiménez-Bucarey, C., Prado-Sabido, T., Fernández Mantilla, M. M., Merino-Flores, I., Izquierdo-Marín, S. S. & Valle-Palomino N. (2023). Education for Sustainable Developments: Challenges for Postgraduate Programmes. *Int. J. Environ. Res. Public Health*. 20:1759. doi: 10.3390/ijerph20031759.

Ahmad, Z., & Shahzadi, E. (2018). Prediction of students' academic performances using artificial neural networks. *Bulletin of Education and Research, 40*(3), 157–164.

Alamri, L. H., Almuslim, R, S., Alotibi, M, S., Alkadi, D. K. I., Ullah Khan, I. & N. Aslam, N. (2020) Predicting students' academic performance using support vector machine and random forest in Proceeding of the 2020 3rd International Conference on Education Technology Management, pp. 100–107, London, UK, June 2020.

Alsariera, Y, A Y., Baashar, G, Alkawsi, G., Mustafa, A. Alkahtani, A. A. & Ali, N. (2022). Assessments and evaluations of different machine learning algorithms for predicting student performance, Computational Intelligence and Neuroscience 1–11, 2022.

Baashar, Y., Alkawsi, G., Ali, N. Alhussian, H. & Bahbouh, H. T. (2021). Predicting student performance using machine learning method: a systematic literature review, in Proceedings of the 2021 International Conference on Computer and Information Sciences (ICCOINS), pp. 357–362, Kuching, Malaysia, June 2021.

Bernacki, M. L., Chavez, M. M., and Uesbeck, P. M. (2020). Predicting achievement and providing support before STEM major begin to fail. *Computer and Education*, 158(August), 103999. https://doi.org/10.1016/j.compedu.2020.103999.

Daud, A., Aljohani, N. R., Abbasi, R. A., Lytras, M. D., Abbas, F. & Alowibdi, J. S. (2017). "Predicting students' performance using advanced learning analytics." in Proceeding of the 26th International Conference on World Wide Web Companion—WWW '17 Companion. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee. 415–421. doi: 10.1145/3041021.3054164.

Fly, O., Jet, A., Awodele, O., Hinmikaiye, J., O., Olakanmi, O. & J. Akinjobi, J. (2008). Supervised machine learning algorithms: classification and comparison," *International Journal of Computers Trends and Technology*, vol. 48, no. 3, pp. 128–138, 2017.

Delavari, N., Phon-Amnuaisuk, S., & Beikzadeh, M. R., (2008). Data mining applications in higher learning institutions, *Informatics in Education,* 7(1),31–54.

Ghorbani, R., & Ghousi, R. (2020). Comparing different resampling method in predicting student performance using

machine learning techniques, *IEEE Access*, 8, 67899–67911.

Kumar, V. U., Krishna, A., Neelakanteswara, P., and Basha, C. Z. (2020). "Advanced prediction of performance of a students in an university using machine learning techniques." in 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC). Coimbatore, India. 121–126. doi: 10.1109/ICESC48915.2020.9155557

Hasan, H, M., R., Rabby, A. K, M. A., Islam, M. T. & Hossain, S.A. (2019). Machine learning algorithms for student performance prediction in Proceedings of the 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 1–7, Kanpur, India, July 2019.

Nachouki, M., & Abou Naaj, M. (2022). Predicting student performance to improve academic advising using the random forest algorithms, *International Journal of Distance Educ. Technol.,* 20(1)*,*1–17. Jan.2022.

Shah, T. H. (2022). Research Anthology on Big Data Analytics, Architectures, and Application. Information Resource Management Association; Hershey, PA, USA: 2022. Big data analytics in higher education. 1275–1293.

Tejedor, F. & García-Valcárcel, A. (2007). Causas del bajo rendimiento del estudiante universitario (en opinión de los profesores y alumnos). Propuestas de mejora en el marco del EEES. Revista de Educación 342, 443–473. Available at: https://dialnet.unirioja.es/servlet/articulo?codigo=2254218.

Viale, H. (2014). Una aproximación teórica a la deserción estudiantil. Revista Digital de Investigación en Docencia Universitaria. 8, 59–76. doi: 10.19083/ridu.8.366.

Vicerrectorado Académico, (2017). Cuando lo que se sabe nos dice cuánto no se sabe—Vicerrectorado Académico. Available at:https://vicerrectorado.pucp.edu.pe/academico/noticias/cuando-lo-que-se-sabe-nos-dice-cuanto-no-se-sabe/.

Waheed, H., Hassan, S. U., Aljohani, N. R., Hardman, J., Alelyani, S. & Nawaz, R. (2020). Predicting academic performances of student from VLE big data using deep learning models. *Computer in Human Behavior,* 104(October 2019), 106189. https://doi.org/10.1016/j.chb.2019.106189.

Yauri, R. A., Suru, H. U., Afrifa, J. & Moses, H. G. (2023). A Machine Leaning Approach in Predicting Student Academic Performance Using Artificial Neural Networks. Journal of Computational and Cognitive Engineering https:// doi.org/10.47852/bonviewJCCE3202470.