Journal of Applied Sciences, Information and Computing

Volume 4, Issue 2, November 2023

ISSN: 1813-3509 https://doi.org/10.59568/JASIC-2023-4-2-04

ENHANCED STUDENT RETENTION IN OPEN AND DISTANCE EDUCATION THROUGH EFFECTIVE ACADEMIC PERFORMANCE MODEL USING NAÏVE BAYES AND K-NEAREST NEIGHBOR MACHINE LEARNING ALGORITHMS

[1]Ezeanya C. U., [2]Onyeji E. M., [3]Ejimofor I. A.

[1]Directorate of Information and Technology, National Open University of Nigeria; cezeanya@noun.edu.ng
[2]emmanuelonyeji@gmail.com
[3]Department of Computer Engineering, Madona University, Akpugo; iaejims2@yahoo.com

Abstract
Improving student performance in an academic pursuit is one of the key concerns of institutions especially open and distance learning institutions where learners are separated from the institution by geographical region. The current observation of low-quality graduates from colleges and universities, particularly in open and distance learning, can be attributed to the lack of mechanisms that could help administrators at universities to forecast the academic achievement of the concerned students in the coming years. The goal of data mining in education is to create models, algorithms, and techniques for analyzing information gathered from learning environments to comprehend and enhance the learning process. The goal of this research is to identify patterns in the measures of academic achievement and how they relate to admission, high school, and personal information about the students. These findings can serve as a solid basis for customizing and enhancing the curriculum for open and distance learning to better suit the needs of individual students. Also, the research work identified factors that had a crucial influence on overall students' performance. Hybridizing Naïve Bayes and K-Nearest Neighbor were used as Classifiers to develop a model for predicting the performance of students. The new model which is the hybridized model (combined Naïve Bayes and K-Nearest Neighbor) predicts better results than individual Naïve Bayes and K-Nearest Neighbor algorithms which shown itself as the best prediction and classification model.
Keywords: Naïve Bayes and K-Nearest Neighbor algorithms

I Introduction

The key indicator of whether a student or educational institution has met its short- or long-term educational objectives is its student academic performance, sometimes

known as "academic achievement of learners." There is no consensus on how it should be evaluated or which components are most crucial, although it is frequently measured by examination or ongoing assessment. Academic achievement is crucial in establishing the value of graduates who will be in charge of the nation's social and economic development (Nuankaew & Nuankaew, 2022).

The field known as EDM (Educational Data Mining) is a young one that focuses on understanding how to use educational data to create models, analyze algorithms, and gain a deeper understanding of students' performance. Academic institutions now urgently need educational data mining to raise the standard of instruction. Modeling a variety of research important to student learning in online and distance education systems has been achieved with the use of educational data mining techniques (Mgala & Mgbogho, 2022). Every year, models improve in accuracy and are validated to become more generalizable. Numerous innovative techniques, theories of facilitation, and enhancements have emerged as a result of education research. The way we learn and live has been completely changed by information technology. Today, a second wave of transformation in all areas of learning and accomplishment is supported by the utilization of data gathered through these technologies.

The idea behind predictive modeling is to build a model that is capable of making predictions (Manika & Madhusudhan, 2022). A machine learning algorithm is typically part of such a model, which learns specific attributes from a training dataset before making predictions. Regression and pattern categorization are two other subfields of predictive modeling. Regression models are used to forecast continuous variables, such as the maximum temperature for the upcoming days in the weather forecast, by examining relationships between variables and trends. The goal of pattern classification, in contrast to regression models, is to designate discrete class labels to particular data as results of a prediction.

Providing a good education to students and improving the quality of management decisions is the main goal of any academic institution. Support services especially to open and distance learners play a major role because it helps to alleviate student dropout rate (Onu *et al*, 2023).To help academic planners in remote learning institutions make better decisions and boost student academic performance, the knowledge obtained can be used to make constructive and helpful recommendations and reduce dropout rate (Aman *et al*, 2019), better understand student behavior (Karalar *et al*, 2021) and support moderators, improve moderation and many others

There is always difficulty in choosing the best machine learning classification model to classify student academic performance with a significant accuracy rate, Identifying the most important key indicators that could be helpful in creating the classification model for predicting the grades of students' dissertation projects, Choosing the right variables/attributes for correct prediction and using the right prediction technique and tools to help uncover hidden features for early identification of at-risk students. Due to several problems and other circumstances, it is still difficult to forecast student performance with any degree of accuracy. Therefore, this research deals with the possibilities of data mining in education to improve the quality of the decision-making process in distance learning institutions by proposing the student achievement predictor model.

The evaluation of students' academic achievement is influenced by a variety of elements, including a student's psychological, socioeconomic, and personal characteristics. The traditional

metric for evaluating a student's academic success is their prior cumulative grade point average (CGPA), but there are numerous other crucial factors that influence the final result. All of these factors must be included in predictive models in order to accurately forecast student success. It is advantageous to identify students who perform poorly academically by accurately predicting student performance. The school administration can provide individualized support to the identified children so that their performance can advance in the future.

Machine Learning

Machine learning has been used to describe a system that can automatically gather and synthesize knowledge. Because of its capacity for learning through analytical observations, examples, and experiences, the system can continually advance and deepen its understanding while delivering better outcomes. Machine learning can be supervised learning, Unsupervised learning, Semi supervised learning or Reinforcement learning.

Machine Learning Algorithm: This refers to a piece of program code (mathematics or program logic) that enables professionals to study, analyze, understand, and explore large complex data sets. Each algorithm follows a series of instructions to achieve the goal, make predictions, or categorize information by learning, creating, and discovering patterns embedded in the data. Some machine learning algorithms that are commonly used to achieve actual results are: Decision Trees, Artificial Neural Networks (ANNs), and Random Forest Algorithm, K-Nearest Neighbors (ANN) Algorithm, Support Vector Machines (SVMs) and Naive Bayes algorithm. Each of these algorithms belongs to the supervised machine learning family. According to Amin et al (2017), he suggested the use of machine learning techniques in building a predictive GPA model where students GPA was predicted in order to help counsel the student where necessary and 53% accuracy was achieved. Also Predictive model for the forecast of student academic performance was also developed by Yağcı (2022) and in the process of analyzing the data, it was discovered that students need to be helped in their learning process in order to reduce academic risk and dropout. A study by Kotsiantis et al (2018), Sivasakthi (2017) also suggested the use of five classification algorithms to forecast the performance of distance learners.

From the literature review, it was discovered that the most commonly used machine learning algorithms for predicting student academic performance are Decision Tree (DT), Naive Bayes (NB), Artificial Neural Networks (ANN), Rule-based (RB), K- Nearest Neighbor (KNN), and Support Vector Machine (SVM). Academic performance is influenced by various factors including psychometric factors, demographic factors, work-related factors, social factors, etc. From the reviewed literatures, the knowledge gaps include researchers' inability to identify a helpful indication and the criteria for attaching the advice to evaluate analytic results in practice. Providing better education requires many parameters to understand the process at the level of student understanding.

II Methodology

The selection of the appropriate component and pertinent features with the appropriate prediction approach presents another hurdle in forecasting student academic achievement. Most researchers used a mixed method approach to choose an appropriate method, integrating the best prediction methodology to boost the model's robustness. However, the availability of student data inputs for the model to do the calculation for an accurate forecast also plays a role in selecting an effective approach. The goal of analysis is

to get a realistic and accurate insight into a system and its problem areas in order to design an improved system. This study has used the technique of knowledge discovery in databases (KDD) and object-oriented analysis and design methodology (OOADM). Two instruments were examined: Naive Bayes Classifier and K-Nearest Neighbor (KNN)

Data Sources and Data Collection Techniques

Both primary and secondary information were gathered from various sources to conduct a thorough analysis of the current system.. Primary data was collected from the institution and secondary data was collected from literature review, which includes understanding and observation of the available Academic Performance Prediction System. Secondary data was also collected from a range of sources to conduct an insightful study of the systems in place, how they work, and how they operate. Internet resources, publications, articles from newspapers, and the guide evaluations of educational achievement forecasting are among the sources.
The dataset used in this study was obtained from the Department of Computer Science at the National Open University of Nigeria (NOUN) Study Center in Enugu State. First, a sample of about 1000 postgraduates from different centers was collected. During this phase, the data collection process was examined in order to select an appropriate dataset to work with. At that time, the Information Technology course in the Computer Science department was selected because of its large number of postgraduates and also because of the high dropout statistics of the course. The rules and procedures for collecting data on examination results were also reviewed. Records of 499 master's students were extracted from the records. The results for the first semester of the academic year 2017 to 2020 were also used. A total of nine courses are selected and recorded for students. Student demographics, behavioral and attitudinal data, parental and school factors were collected using questionnaires given to the students. These questionnaires were uploaded to the Google form and the links were sent to the students so that they could fill in the questionnaire. Printed copies were also distributed to the students. The results of the first semesters were collected by the Exams Unit of the Postgraduate School. Courses include both core and elective courses. The result showed the overall performance of the candidates in each course (100 in total). The cumulative first semester grade point average (CGPA) for each student was calculated.

Transforming the collected data into a suitable format to be used to perform structured analyses was also done. The researcher also used a discretization mechanism to convert the student-related factors and student performance grade from numerical values to nominal values that represent the class labels of the classification problem. To achieve this step, the researcher divided the dataset into three nominal intervals (high level, medium level, and low level) based on the students' CGPA.

Hybrid method (combination of Nave Bayes and KNN algorithm) was applied to provide an accurate assessment of the characteristics that may affect student performance/grades and to improve the performance of the students' predictive model. These methods resample the original data into samples of the data set, and then each sample is trained by a classifier. The classifiers used in the student prediction model were K-nearest neighbor (KNN) and Nave Bayesian (NB). To select a tool and the best algorithm to serve as the basis for developing the new multi-agent student academic performance prediction model, WEKA, JADE and Netbeans IDE were selected. The study results of the previous semester, based on 48 attributes

such as matriculation number, semester results, etc., which make up a large part of the intra-semester grades, were used to predict the final exam results.

KNN Classifier

Ihsan and Ashraf (2017) predicted student performance using the KNN classifier. The basis for the nearest neighbor is the categorization of an unknown data point whose class is already known. Here it is called nonparametric because it makes no assumptions about the underlying data distribution, i.e. H. the model structure is determined by the data. The K value establishes the class of the sample information point by determining how many nearest neighbors to take into account when calculating the nearest neighbor Bayesian (NB). It works by finding the distances between a query and all samples (k) closest to the query, and then voting for the most common label (in the case of classification). K Nearest Neighbour Easily Explained with Implementation in Figure 1.

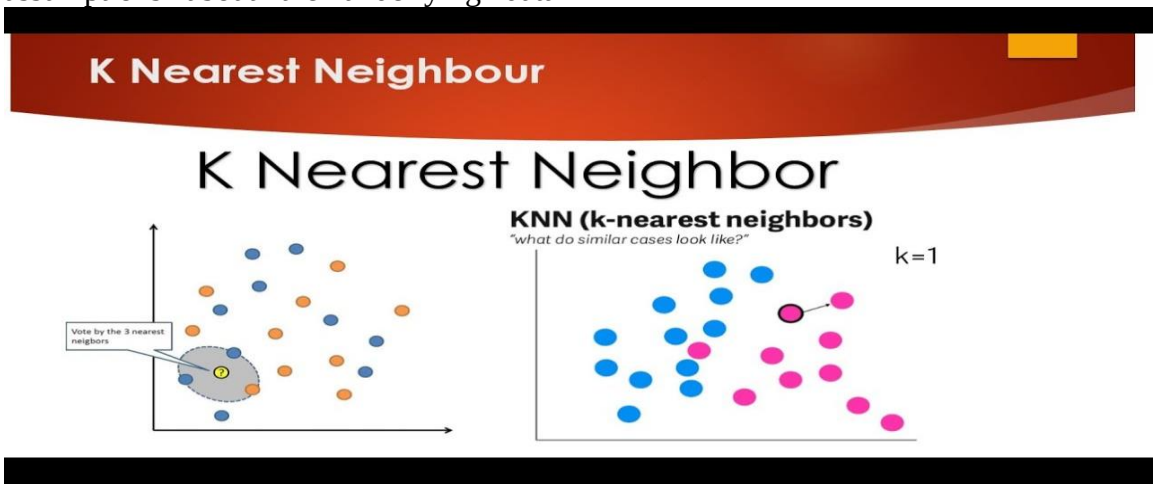

Figure 1. K Nearest Neighbour (KNN)

The **K nearest neighbor algorithm** is shown below. Depending on the function to be performed, the classification uses the K-label mode while the regression calculates the mean of the K-labels as shown in Figure 2 and Figure 3

The KNN Algorithm

1. Load thedata
2. Initialize K to your chosen number of neighbors
3. For each example in the data
   Calculate thedistancebetweenthequeryexampleandthecurrentexamplefromthedata. Add the distance and the index of the example to an ordered collection
4. Sort the ordered collection of distances and indices from smallest to largest (inascending order) by the distances
5. Pick the first Kentries from the sorted collection
6. Get the labels of the selected Kentries
7. If regression, return the mean of the K labels

Figure 2 Algorithm for KNN Classifier (https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm)

k←the number of nearest neighbor
for each object Z do
Calculate the distance between every object x and x in the training set d(x,z) Neighborhood← the

k neighbors, closest to Z in the training set

Z Class←Select Class (according to neighbourhood)
End for

Figure 3: Pseudocode for KNN Classification Algorithm (Ihusan and Ashraf 2020)

Naïve Bayes Algorithm

Dake and Gyimah (2017) used the Nave-Bayes classifier to examine student grades. Based on the Bayes theorem, it works best with high data dimensionality. Based on the input, the Bayesian classifier can determine the maximum output that can be produced. A more accurate probabilistic classifier is produced at runtime by adding new raw data. In this case, the existence of one feature in a class does not imply the existence of other features.. In Nave Bayes, the probability of each class in the training set is calculated and the value with the highest probability becomes the predicted value, as shown in the algorithm in Figure 4

Input:
Training dataset T,
F= (f1, f2, f3,..,fn) // value of the predictor variable
in testing dataset. Output:
A class of testing dataset. Step:

1. Read the training taset T;
2. Calculate the mean and standard deviation of the predictor variables in each class;
3. Repeat

Calculate the probability of fi using the gauss density equation in each class; Until the probability of all predictor variables (f1, f2,f3,.., fn) has been calculated.

Fig 4: Algorithms for Naïve Bayes Classifier

In this research, two existing academic performance prediction systems were analyzed using Naive Bayes and KNN Classifier.

Students Grades Predictor using Naïve Bayes Classifier

Dake and Gyimah (2017) proposed a Nave Bayes approach to predict students' final grades. The classifier model is based on data from previous students who have taken the same course. Attributes/functions used included attendance, assignment, test score, class attendance, and proximity to hostel, gender, and academic rank. Attributes and their values have been chosen at discretion, which may affect a student's ability to pass or fail an exam.

50 cases in total were collected for analysis to test the classifier. The comma-separated values (CSV) were transformed into the Weka Attribute-Relation File Format (ARFF) using the ARFF-View. The training data set was then subjected to Nave Bayes classification. The result obtained gave an accuracy of 88% for correctly classified instances. The model was evaluated with 10-fold validation. Of the 50 cases, 44 were classified correctly and 6 incorrectly. The student's academic status was then predicted. The chart in Figure 4 highlights the levels for classifying student achievement. First, data is collected from various sources and relevant attributes or characteristics are selected. Next, pre-processing is performed. In this phase, the data is transformed into the format in which the classification can be done. It includes feature extraction, normalization and discretization. The Nave Bays classifier is then applied to analyze the patterns and discover important features in the data. Finally, the results are evaluated.

Predicating Academic Performance of Students in Higher Institutions with KNN Classifier

Omisore and Azeez (2018) developed a predictive model using the ANN classification to predict the academic performance of students in higher institutions. The educational dataset of 310

students of all levels in the 2013/2014 session was collected and used by the University of Lagos, Akoka. Further relevant information was collected via questionnaires. Various selected characteristics were used for the prediction. Variables/student attributes used include student demographics, current and previous academic status, department structure, and family background. Then the data collected in different tables was merged and lower entropy attributes were removed as shown in Figure 6. The academic standing of the students was stratified into five different groups - excellent, good, average, poor and poor. 10-fold cross-validation was used to assess performance. The result obtained gave a prediction accuracy of 58.3%.

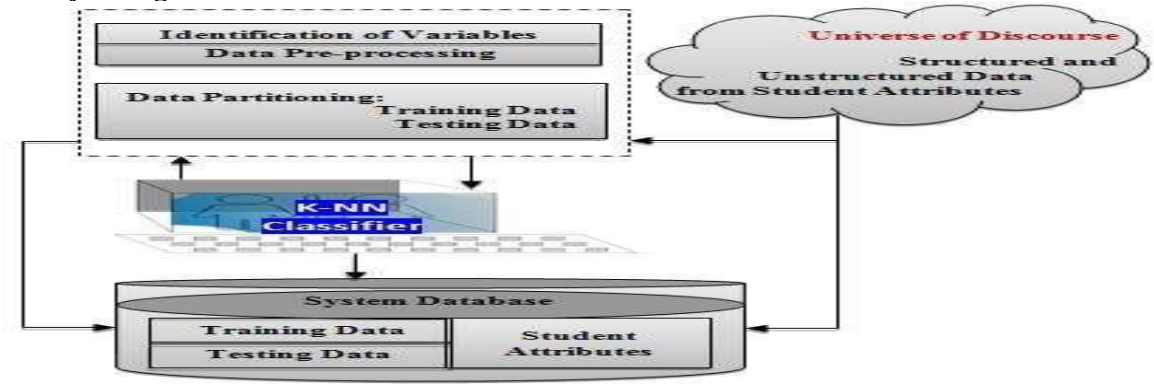

Figure 6 Conceptualized models for predicting student academic performance (Omisore and Azeez 2018)

Analysis of the new Algorithm

Today, distance learning and e-learning are heavily influenced by information technology. Everything is shifting from manual to automated methods. In this research, a technique is developed to predict the academic performance of postgraduate students by analyzing the students' performance in their first semester courses using the K-Bay classification method; an integration of K-nearest neighbor classifier and Nave Bayes. Attributes of the student such as demographic factors, social and academic factors, unit test grades, final semester exams, and the student's aggregate cumulative grade point average (CGPA) over previous semesters are used in the prediction.Combining classifiers to improve accuracy is a common phenomenon nowadays, as both Nave Bayes and KNN are simpler but powerful algorithms that are ideal candidates for combining to achieve higher accuracy. The hybrid system is used to compare the two individual classifiers, Nave Bayes and KNN, to find the more efficient data mining classifier. This would lead to finding a more efficient and time-saving algorithm to predict a student's performance. The new system will be cost and time efficient. This will have simple operations. Using the model, students, teachers, and curriculum reviewers can easily access an up-to-date curriculum from their various departments. We can describe the new algorithm (see also Figure 7) as follows:

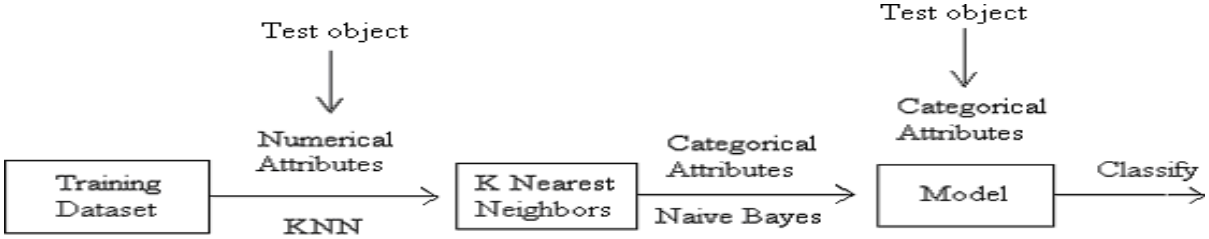

Figure 7. Graphical representation of the proposed algorithm

Step:1. Obtain the K-Nearest Neighbor of a new observation based on the numerical attributes.

Step:2. Use the set of K observations, found in step 1 as training data and use it to build a model exploiting the Naïve Bayes algorithm based only on the categorical attributes.

Step:3. Use the model built in step 2 to classify the new observation.

The idea of the new algorithm is very simple. KNN and Nave Bayes are used to the training and testing phase. The two classifiers are then combined to give a powerful classifier that increases the accuracy of the prediction at a lower time. The new system was also able to predict students' academic performance and offer academic advice based on academic standing. The new system was able to query the factors that affect each student's performance, as well as the important factors that affect students' academic performance in general.

The process model of the new algorithm

The process model of the new algorithm is shown in Figure 8 below. Starting with the collection of data from exam units and questionnaires, followed by pre-processing, where the data is converted into the format that the algorithm can understand. The classification process is complete and the result for the analysis is evaluated.

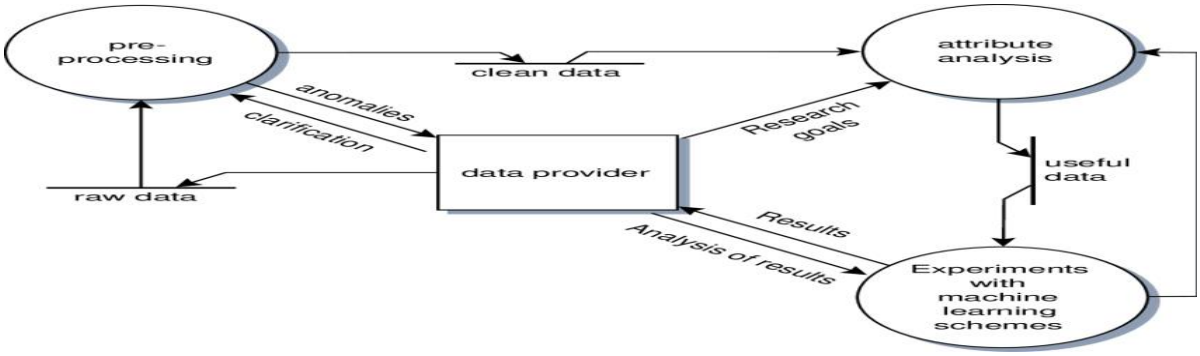

Fig 8: Process Model for a Machine Learning Application (Garner et al 2020)

III Training Set/Test Set

The training set is used for model building and for testing the model evaluation. The training dataset is implemented to build the model, while the testing (or validation) set is used to validate the model. All data were split using the 90% and 10% division. 90% of the data was used as a training dataset and 10% as a test classification task, K-NN and Nave Bayes were used to build the performance model and optimize performance.

Results and Discussions

Table 4 to Table 10 shows the performance metrics using all the attributes and the exam scores.

Table 4: Contingency Table for Naïve Bayes and K-NN Evaluation (CrossValidation)

		A	B	C	<-- classified as	
Naïve Bayes	A	24	0	6	a =HIGH b =LOW c =MEDIUM	Predicted Class by Naïve Bayes classifier
	B	0	250	18		
	C	2	19	180		
K-NN	AB	4	5	21	a =HIGH b =LOW c =MEDIUM	Class by K-NN classifier
	C	6	180	82		

		18	66	117		

Table 5 shows the analysis results from 10 validations. The dataset used for cross-validation/90.98% was correctly classified by Naïve bayes while 60.32% was correctly classified using the KNN algorithm.

Table 5: Filtered results for KNN and Naïve Bayes (Cross-validation using all attributes)

Classifier	**Train (499 instances) Cross Validation using all attributes**				
	Correctly Classified Instances Student	Incorrectly Classified Instances Student	Accuracy	Kappa statistic	Mean absolute error
Naïve Bayes	90.982%	9.018%	81.76%	0.8338	0.0834
K-NN	60.3206 %	39.6794 %	53.91%	0.2771	0.2657

Table 6 shows the metrics result for the two classifiers using cross validation with their various accuracy measures. Performance metrics like Precision, Recall, F-Measure, Mathews Correlation Coefficient (MCC), and Receiver Operator Characteristics (ROC) were used for the analysis.

Table 6: Naïve Bayes and K-NN Performance metrics results using all attributes

Algorithm	**TP Rate**	**FP Rate**	**Precision**	**Recall**	**F-Measure**	**MCC**	**ROC Area**	**PRC Area**	**Class**
Naïve Bayes	0.800	0.004	0.923	0.800	0.857	0.851	0.993	0.937	HIGH
	0.933	0.082	0.929	0.933	0.931	0.851	0.968	0.973	LOW
	0.896	0.081	0.882	0.896	0.889	0.813	0.954	0.930	MEDIUM
K-NN	0.133	0.051	0.143	0.133	0.138	0.085	0.488	0.067	HIGH
	0.672	0.307	0.717	0.672	0.694	0.363	0.683	0.661	LOW
	0.582	0.346	0.532	0.582	0.556	0.234	0.614	0.477	MEDIUM

Using a split of 90% training data and 10% percent test data, the system was trained and tested using Naïve Bayes and KNN algorithms. The analysis results as shown in Figure 7 show the contingency table for the percentage distribution results for the learning test and the set of tests using the classifier, while Table 8 shows the accuracy of the prediction.

Table 7: Contingency Table of the result (Model Building and Evaluation using all attributes)

	A	**B**	**C**	<-- classified as	
Naïve Bayes	**22**	0	2	**a** =HIGH **b** =LOW **c** =MEDIUM	Predicted Class by Naïve Bayes classifier
	0	**235**	12		
	5	19	**154**		
K-NN	14	1	9	**a** =HIGH **b** =LOW **c** =MEDIUM	Predicted Class by K-NN classifier
	0	**216**	31		
	8	33	**137**		

Table 8: Prediction Accuracy and Computational Results using all attributes

	Train set (449 instances)	**Test set(49 instances)**

Classifier	Correctly Classified Instances Student	Incorrectly Classified Instances Student	Accuracy	Kappa statistic	Mean absolute error	Correctly Classified Instances Student	Incorrectly Classified Instances Student	Accuracy	Kappa statistic	Mean absolute error
Naïve Bayes	91.5367 %	8.4633 %	91.54	0.8425	0.0718	71.4286%	28.5714%	71.43%	0.4503	0.3965
K-NN	81.7372 %	18.2628 %	81.74	0.6585	0.1236	69.3878%	30.6122%	69.39%	0.4258	0.4381

Table 9 shows the results of the measures for the two classifiers using percentage distributions with their different measures of precision. Measures such as accuracy, recall, measure F, Mathews correlation coefficient (MCC) and receiver operator characteristics (ROC) were used for the analysis.

Table 9: Naïve Bayes and Performance metrics results using all attributes

Algorithm	**TP Rate**	**FP Rate**	**Precision**	**Recall**	**F-Measure**	**MCC**	**ROC Area**	**PRC Area**	**Class**
Naïve Bayes	0.917	0.012	0.815	0.917	0.863	0.856	0.996	0.929	HIGH
	0.951	0.094	0.925	0.951	0.938	0.861	0.973	0.979	LOW
	0.865	0.052	0.917	0.865	0.890	0.822	0.963	0.946	MEDIUM
K-NN	0.583	0.019	0.636	0.583	0.609	0.588	0.773	0.415	HIGH
	0.874	0.168	0.864	0.874	0.869	0.707	0.847	0.813	LOW
	0.770	0.148	0.774	0.770	0.772	0.623	0.807	0.717	MEDIUM

Table 10 describes the performance evaluation of the hybrid proposal between Naïve Bayes and our k-nearest neighbor. Calculation results show that the accuracy level in each data mining classifier with Naïve Bayes has an accuracy of 71.43%, K-NN is 69.39% and K-Bay is 95.92, respectively. %. Model testing time for K-Bay is lower than for KNN and Naïve Bayes. This result shows that the hybrid model gives better performance than the simple model.

Table 10 K-Bay performance Evaluation using all attributes

Classifier	**Test set (49 instances)**					
	Correctly Classified Instances Student	Incorrectly Classified Instances Student	Accuracy	Kappa statistic	Mean absolute error	Model Testing Time (Sec.)
Naïve Bayes	71.4286%	28.5714%	71.43%	0.4503	0.3965	0.357
K-NN	69.3878%	30.6122%	69.39%	0.4258	0.4381	0.18
K-Bay	95.9184 %	4.0816 %	95.92%	0.9213	0.1395	0.134

Performance Evaluation using highly Influencing Factors/attributes

Performance results using only high-impact attributes using only high-impact factors/attributes and evaluation scores are shown in Tables 11 to Table 17

Table 11: Naïve Bayes and K-NN Performance results using highly influencing attributes

Naïve Bayes	A	B	C	<-- classified as	
	0	6	**24**	**a** = HIGH	Predicted Class by Naïve Bayes classifier
	265	3	0	**b** = LOW	
	0	**201**	0	**c**= MEDIUM	
K-NN	**0**	23	7	**a** = HIGH	Predicted Class by K- NN classifier
	218	**49**	1	**b** = LOW	
	50	137	**14**	**c**= MEDIUM	

Table 12: Naïve Bayes and K-NN Performance results using highly influencing attributes

Algorithm	**TP Rate**	**FP Rate**	**Precision**	**Recall**	**F-Measure**	**MCC**	**ROC Area**	**PRC Area**	**Class**
Naïve Bayes	0.871	0.007	0.900	0.871	0.885	0.877	0.991	0.954	HIGH
	0.983	0.014	0.987	0.983	0.985	0.969	0.998	0.998	LOW
	0.968	0.031	0.958	0.968	0.963	0.936	0.991	0.981	MEDIUM
K-NN	0.677	0.029	0.636	0.677	0.656	0.630	0.886	0.543	HIGH
	0.857	0.247	0.785	0.857	0.819	0.614	0.825	0.782	LOW
	0.676	0.146	0.770	0.676	0.720	0.542	0.792	0.710	MEDIUM

Table 13: Contingency Table of the result (Cross Validation using highly influencing attributes)

Classifier	**Train (499 instances) using Ranking (5 attributes)**				
	Correctly Classified Instances Student	Incorrectly Classified Instances Student	Accuracy	Kappa statistic	Mean absolute error
Naïve Bayes	98.1964	1.8036 %	98.40%	0.9667	0.0558
K-NN	72.5451 %	27.4549 %	72.55%	0.4918	0.1848

Table 14: Prediction Accuracy and Computational Results using highly influencing attributes

Classifier	**Train set (449 instances)**					**Test set(49 instances)**				
	Correctly Classified Instances Student	Incorrectly Classified Instances Student	Accuracy	Kappa statistic	Mean absolute error	Correctly Classified Instances Student	Incorrectly Classified Instances Student	Accuracy	Kappa statistic	Mean absolute error
Naïve Bayes	96.882 %	3.118 %	96.88	0.944	0.0615	75.5102 %	24.4898 %	75.51%	0.5859	0.1771
K-NN	76.8374 %	23.1626 %	76.84	0.5824	0.156	59.1837%	40.8163 %	59.18%	0.3436	0.2826

Table 15: Naïve Bayes and Performance metrics results using highly influencing attributes

Classifier	**Train set (449 instances)**					**Test set(49 instances)**				
	Correctly Classified Instances Student	Incorrectly Classified Instances Student	Accuracy	Kappa statistic	Mean absolute error	Correctly Classified Instances Student	Incorrectly Classified Instances Student	Accuracy	Kappa statistic	Mean absolute error
Naïve Bayes	96.882 %	3.118 %	96.88	0.944	0.0615	75.5102 %	24.4898 %	75.51%	0.5859	0.1771
K-NN	76.8374 %	23.1626 %	76.84	0.5824	0.156	59.1837%	40.8163 %	59.18%	0.3436	0.2826

Table 16: Contingency Table (Model Building and Evaluation using highly influencing attributes)

	A	**B**	**C**	<-- classified as	
Naïve Bayes	**0**	4	27	**a** =HIGH **b** =LOW **c** =MEDIUM	Predicted Class by Naïve Bayes classifier
	226	**4**	0		
	3	182	**3**		
K-NN	**3**	7	21	**a** =HIGH **b** =LOW **c** =Medium MEDIUM	dicted Class by K-NN classifier
	197	**31**	2		
	51	127	**10**		

Table 17 describes the performance evaluation of our naive Bayesian and k-nearest neighbor hybrid proposal using only highly influential factors/attributes. Calculation results show that the accuracy level in each mining classifier with Naïve Bayes has an accuracy of 75.51%, K-NN is 59.38% and K-Bay is 99%, respectively. Model testing time for K-Bay is lower than for KNN and Naïve Bayes. According to the results, KNN performed better when all attributes were used, while Naïve Bayes performed better when only the most influential attributes were used. The hybrid model also gives better performance when using only the powerful attributes than using all the attributes.

Table 17: K-Bay Performance Ranking Using All Attributes High Impact Attribute Usage

Classifier	**Test set (49 instances)**				
	Correctly Classified Instances Student	Incorrectly Classified Instances Student	Accuracy	Kappa statistic	Mean absolute error
Naïve Bayes	75.5102 %	24.4898 %	75.51%	0.5859	0.1771
K-NN	59.1837%	40.8163 %	59.18%	0.3436	0.2826
K-Bay	99%	1%	99%	99%	0.0019

The Naïve Bayes and K-Nearest Neighbor (KNN) algorithms were evaluated in terms of their performance using the free software tool WEKA on the data. Feature selection techniques are used to rank certain attributes/features in order to determine the most important features and their ranking order. Use correlation-based feature selection; the five most important ranked attributes; These include employment status, overly busy supervisors with high commitment, poor library facilities, standard laboratory equipment and facilities, use of stimulants, learning-enhancing drugs, use of Internet regularly to surf and socialize. Network connections.

Third, an associative model was developed using a combination of Naïve Bayes and K-Nearest Neighbor to predict student

learning outcomes. First Semester Results of 499 graduate students were used in the prediction.

Fourth, the new model was tested in two parts; use all the attributes and use the influential ones as shown in table 3 and table 4 below.

Using all attributes, the system achieved an accuracy of 95.92% against single classifiers; Naïve Bayes and KNN have an accuracy of 69.39% and 71.43/% respectively. The execution time of the new model is 0.134 seconds while KNN and Naive Bay are 0.357 and 0.18 seconds, respectively. Using only high-impact attributes, the system achieves 99% accuracy against single classifiers; Naïve Bayes and KNN have an accuracy of 75.51% and 59.18%, respectively. In summary, this research can motivate and help universities regularly perform data mining tasks on student data to uncover interesting results and patterns that could help the whole university and students as well.

IV Conclusion

Research has revealed that data mining has the potential to become an important part of the decision-making and knowledge management processes of educational institutions. Research shows that student data available at higher education institutions can be used to predict student learning outcomes. Enroll in various courses that use educational machine learning algorithms. Student performance is not the result of a single determinant. It depends on many factors such as personal, socioeconomic, psychological and other environmental factors.

Predicting student learning outcomes with a high accuracy rate will improve the educational services of higher education institutions. The reliability of the model can help the educational institution to know the student's learning status in advance and identify the students with a high probability of failing to take appropriate measures such as advising students and providing solutions timely. It also allows the institution to identify bright students and foster their future growth by encouraging them. In the long run, this can help students improve their learning and ultimately lead to better academic performance, thereby reducing dropout rates and depression.

V References

[1]Aman F., Rauf A., Ali R., Iqbal F. and Khattak A.M.(2019) "A Predictive Model for Predicting Students Academic Performance," *2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA)*, Patras, Greece, 2019, pp. 1-4, doi: 10.1109/IISA.2019.8900760.

[2]Amin. Z., Refik, C., Yau, H., & Hernandez-Torrano, D., (2017). Predicting Students‟ GPA and Developing Intervention Strategies Based on Self-Regulatory Learning Behaviors. 2017, IEEE

[3]Dake, D.K., & Gyimah, E. (2017). Students Grades Predictor using Naïve Bayes Classifier – A Case Study of University of Education, Winneba.

[4]Ihsan A. & Ashraf M. (2017). Students performance prediction using KNN and Naïve Bayesian. 909-913. 10.1109/ICITECH.2017.8079967.

[5] Karalar, H., Kapucu, C. & Gürüler, H (2021). Predicting students at risk of academic failure using ensemble model during pandemic in a distance learning system. *Int J Educ Technol High Educ* **18**, 63 (2021). https://doi.org/10.1186/s41239-021-00300-y

[6] Kotsiantis, S., Zaharakis, I., & Pintelas, P. (2018). Supervised machine learning: A review of classification techniques. Ed, 2018.

[7]Manika L. & Madhusudhan, M. (2022). Predictive Modeling. 10.1007/978-3-030-85085-2_8.

[8]Mgala, M., & Mbogho, A. (2022). Data-driven intervention-level prediction modeling for academic performance. 15 Proceedings of the Seventh International Conference on Information and Communication Technologies and Development. May 2022. DOI: 10.1145/2737856.2738012

[9]Nuankaew P, Nuankaew WS(2022). Student performance prediction model for predicting academic achievement of high school students. *European J Ed Res*. 2022;11(2):949-963. doi: 10.12973/eu-jer.11.2.949

[10]Omisore, O., & Azeez, N. (2018). Predicting Academic Performance of Students with KNN Classifier.Conference:ACMInternationalConferenceonComputerScienceResearchand Innovations (CoSRI2018)

[11]Onu F.U., Ezeanya C.U., Ezea I. and Obabueki O. (2023). Enhanced Student Support System in Open and Distance Education Using Long Short-Term Memory Recurrent Neural Network, FUOYE Journal of Engineering and Technology(FUOYEJET), 8(1), 10-16. http://doi.org/10.46792/fuoyejet.v8i1.955

[12]Sivasakthi, M.. (2017). Classification and prediction based data mining algorithms to predict students' introductory programming performance. 346-350. 10.1109/ICICI.2017.8365371.

[13]Yağcı, M. Educational data mining: prediction of students' academic performance using machine learning algorithms. *Smart Learn. Environ.* **9**, 11 (2022). https://doi.org/10.1186/s40561-022-00192-z