ISSN No:-2456-2165

An Evaluation of Ensemble and Non-Ensemble Data Mining Techniques in Predicting Students' Performance on E-learning Dataset

[1]Yusuf Abubakar, [2]Aminu Umar Abubakar, [2]Saidu Tasi'u, [2]Ahmad Muhammad Lawal
[1]Department of Computer Science, Kaduna Polytechnic, Kaduna, Nigeria.
[2]Department of Computer Science, Nuhu Bamalli Polytechnic, Zaria, Nigeria.

Abstract:- The excessive use of e-learning technology today has resulted in a massive growth in educational data. Students' interactions with the e-learning environment, particularly learning management systems, create huge amount of data within the shortest period of time. The data contains hidden information about students' engagement in various e-learning activities, which can be linked to their performance. Predicting student performance based on the usage of e-learning systems in educational institutions is a big concern, and it has become critical for educational administrators to better understand why so many students do badly or fail their courses. However, due to the numerous features that influence their performance, making a prognosis is challenging. This study aims to compare various ensemble techniques against their non-ensemble counterpart in predicting students' performance on data generated from learning management system. Five popular algorithms were used: Decision Tree (DT), K-Nearest Neighbor (KNN), Discriminant Analysis (DA), Nave Bayes (NB) and Support Vector Machine (SVM). To improve the performance of the classifiers, ensemble techniques such as RUSBoost, Bag and AdaBoost were employed to increase the accuracy of the students' performance prediction models. The findings revealed that after applying ensemble methods, it achieved a higher accuracy was obtained.

***Keywords**: E-Learning, Ensemble, Performance, Algorithm, Data.*

I. INTRODUCTION

Predicting students' academic achievement in educational environments could be extremely beneficial in a variety of ways. Worldwide, educational systems have rapidly evolved as a result of extensive study in the computational intelligence techniques such as Educational Data Mining (EDM) and Learning Analytics (LA) fields. Thus, educational data are analyzed using these methodologies to deduce distinct patterns used by students in order to forecast their academic performance. (Abubakar & Ahmad, 2017). Academic performance of students has consistently been a deciding factor in determining a student's career path and the Institutions' stature. EDM is a method that is used to wring valuable knowledge from educational settings, as such its applications aid in forecasting students' academic performance. As a result, researchers are delving further into alternative data mining methods in order to improve existing ones. The application of data mining approaches to forecast students' performance in light of their background and behavior has proven to be beneficial in predicting students' varying performances at various levels. Application of such data mining algorithms enables instructors to forecast pupils who are at a high risk of failing, with the purpose of providing a solution to the student. Additionally, it can help identify highly capable students in an educational institution and assist them in obtaining grants (Amrieh, Hamtini, & Aljarah, 2016).

Initially, Educational Data Mining and Learning Analytics were key fields of study in education, but prediction of students' performance become more prominent over time, as the primary goal of this field of study is to examine and forecast student performance in order for them to achieve better perfromance. The primary objective of this study is to investigate the performance of different data mining algorithms for predicting student performance by introducing a new feature category called behavioral characteristics. The educational dataset is derived from the Kalboard 360 eLearning system. The data collection was accomplished through the usage of the Experience API, a mechanism for tracking student activities. The collected features are classified into three classes: academic background, residential, and behavioral characteristics.

The research process incorporates a new feature category called behavioral characteristics, which are connected to the learners' experiences. In this paper, we used educational data to forecast students' academic achievement. As a result, this model analyzed the effect of students' learning behavioral characteristics on their academic achievement. This activity is accomplished through the use of heterogeneous data mining algorithms called classification. Five popular algorithms were used: Decision Tree (DT), K-Nearest Neighbor (KNN), Discriminant Analysis (DA), Nave Bayes (NB) and Support Vector Machine (SVM). To improve the performance of the classifiers, ensemble techniques such as RUSBoost, Bag and AdaBoost were employed to increase the accuracy of the students' performance prediction models.

II. RELATED WORKS

A number of studies in the area of Computational Intelligence and EDM have been conducted in the last few years, resulting in the publication of numerous papers. As a result, numerous studies have used a variety of data mining algorithms for the prediction of students' performance in order to uncover the hidden information contained in their data. This has benefited students and instructors by enabling the e-Learning system to be enhanced. (Bithari, Thapa, & K.C., 2020) used classification techniques such as Decision Tree, Artificial Neural Network, and Naive Bayes to forecast students' academic performance and estimate the impact of students' behavioral characteristics; thus, the Artificial Neural Network performs better.

(Satyanarayana & Nuckowski, 2016) used a number of popular ensemble techniques, such as Bagging, Random Forest and Adaboosting to improve the accuracy of predicting students' performance. The study used Adaboosting on Artificial Neural Networks (ANN) which produced the overall accuracy of 79.1%. (Costa, et al., 2017) also used on bagging technique which is an ensemble of different algorithms predict students' performance, this resulted to achieving high level accuracy of 95%. (Latham, Crockett & Mclean, 2013) proposed a Deep Neural Network model for predicting student performance. In their experiment, they demonstrated that a DNN can perform significantly better with less data by developing a deep understanding of the dataset and fine-tuning the model's quality. As a result, their proposed model was 84.3 percent accurate.

According to (Romero & Ventura, 2010) various types of attributes can affect students' performance, including demographic, social, and school-related attributes. These features were classified using three different algorithms namely; Naïve Bayes, J48 and Multi-Layer Perceptron (MLP). MLP had a 51.2 percent accuracy in their experiment, while Naive Bayes had a 68.6 percent accuracy and J48 decision tree had a 73.9 percent accuracy. As a result, J48 had the highest accuracy rating of the three classification algorithms used. (Romero, Ventura and Garcia, 2008) used the CRISP-DM model to forecast students' academic performance. The J48 Decision tree, Bayesian and the Nearest Neighbor algorithms were used. The Bayesian Classifiers achieved the lowest accuracy, while the J48 achieved the highest.

Usman, et al., (2019) used a variety of classification methods to forecast students' performance at the semester's end. They used four decision tree algorithms namely; J48, CART, C4.5 and ID3 as well as the Naive Bayes Classifier. CART had the highest accuracy of 40%, followed by ID3 with 33.3 percent, C4.5 with 35.2%, and CHAID with 34.2 percent. Naive Bayes algorithm had a precision of 36.4%. (Amrieh, Thair & Ibrahim, 2016) employed four distinct classification algorithms: Naïve Bayes, Decision Tree, Random Forest, and Rule Induction. Decision Tree achieved highest accuracy of 90 percent, followed by the Random Forest at 85 percent, Nave Bayes at 84 percent, and Rule induction at 82 percent. (Kamal & Ahuja, 2019) incorporated categorization techniques to predict and analyze students' academic performance which used clustering techniques to group students according to their cognitive style in using Learning Management System (LMS). An interface for instructors was proposed that would assist them in differentiating students based on their academic strength, so that the weaker students would receive more attention.

(Kakasevski et al., 2008) proposed a predictive model that utilizes data from female students and the ID3 algorithm to identify courses that contribute to poor academic performance at King Saud University's information technology department in Riyadh, Saudi Arabia. Numerous models were constructed using ID3 algorithm, the dataset was partitioned into three groups to allow for the creation of unique models for each group. According to the findings, the model based on second-year student outcomes had the highest accuracy. IT 221 and the different programming courses, CSC111 and CSC113, are excellent indicators of students' achievement. (Guo et al., 2015) forecasted students' final outcomes using the ID3 decision tree. The algorithm informed us of the information gain conditions prior to splitting. 90% was obtained through the application of the model.

Baker, Gowda, & Corbett, (2011) forecasted students' performance using educational data. The model developed in this study was used to examine the effect of attitudinal characteristics on academic performance of students. Four algorithms were used: Support Vector Machines (SVM), ID3, Naive Bayes, and K-Nearest Neighbor (KNN). To increase the accuracy of the classifiers' student performance model, ensemble methods such as Boosting, Voting, and Bagging were used.

III. METHODOLOGY

In this section, a technique for students' performance is presented which is based on a number of heterogeneous classification algorithms in order to investigate the features that can affect students' academic performance. Figure 1 illustrates the proposed methodology. The methodology began with the collection of data from the Kalboard 360 e-Learning platform via the XAPI. Following this is the data preprocessing phase, which involves converting the collected data into an appropriate form that will be suitable for the experiments.

ISSN No:-2456-2165

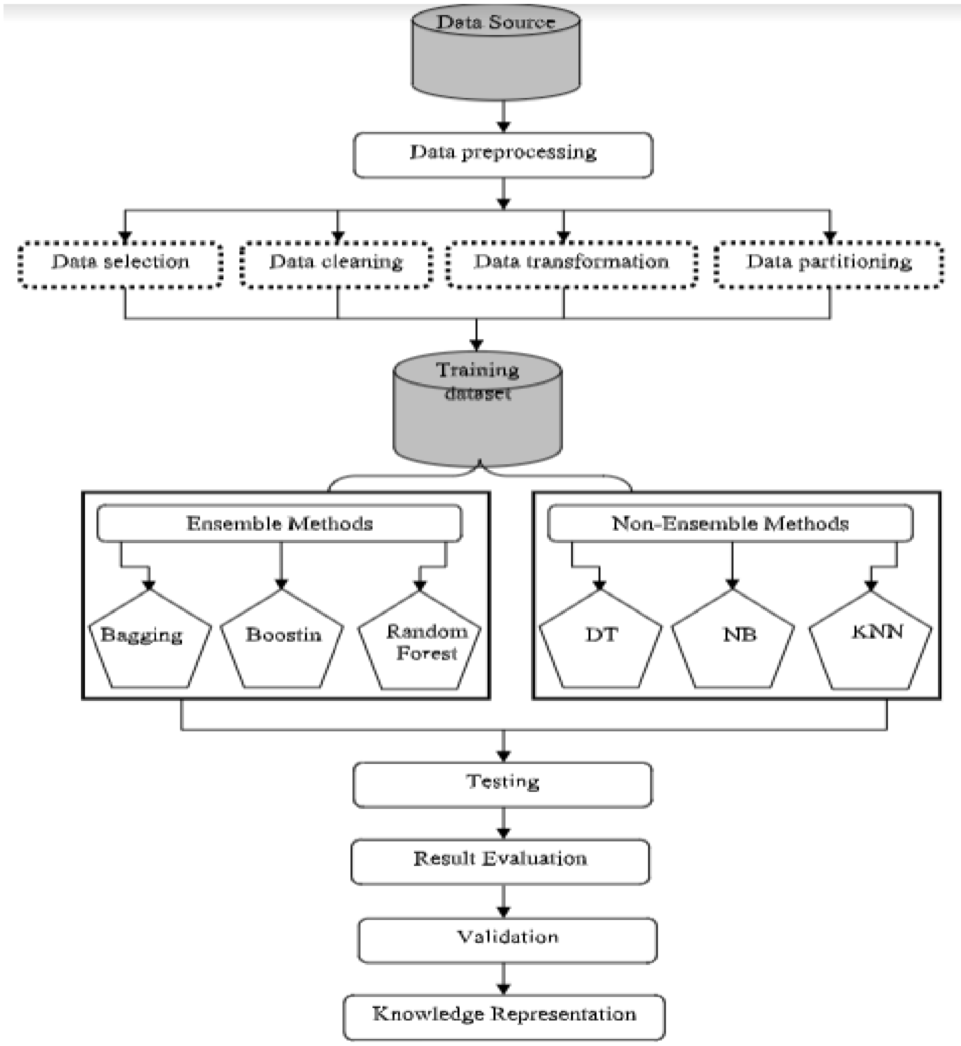

Figure 1: System Architecture

Following that, the discretization technique was applied to transform the data from numerical to nominal values, which corresponds to classification class label. To accomplish this step, the dataset was divided into three classes based on the final grades of students: low-level, middle-level, and high-level. Low-level scores ranges from 0–59, middle-level scores ranges from 60–79, and high-level scores ranges from 80–100. After this division, the best feature set with the highest rank was chosen using feature selection technique. The information gain method is used to select features, and the ensemble model is then constructed using five distinct classification algorithms.

A. Data Set

The usage of the internet in education has increased dramatically, resulting in the development of a new concept called as a Learning Management System (LMS). The LMS is a digital framework that organizes and facilitates online learning. The LMS's objective is to manage all learners, to monitor their interactions, and record their progress throughout the period of their engagemennt. The dataset used for this paper was compiled using Kalboard 360 which is a multiagent learning management system (LMS) that was created to promote learning via the use of cutting-edge technologies. This type of system enables users have access to educational resources synchronously from a variety of devices. The data was gathered using the Experience API (XAPI) learner activity tracking tool (Amrieh, Thair & Ibrahim, 2016). The XAPI is a component of the Training and Learning Architecture (TLA) that enables learners' learning experiences and actions, such as reading an article or viewing an educational video, to be tracked. The Experience API enables suppliers of learning activities to define the learner, the activity, and the items that make up the learning experience (Amrieh, Thair & Ibrahim, 2016). The goal of using XAPI in this research is track students' behavior throughout the period of their interaction with the system in order to estimate the characteristics that may have an effect on students' academic achievement. The educational dataset used in (Usman, et al., 2019) has just 150 records of students with 11 characteristics. The dataset is extended in this study to 500 students with 16 features. These characteristics are grouped into three broad categories: (1) Demographic characteristics such as gender and country origin (2)

Characteristics of the academic background, such as educational stage, grade level, and section (3) Behavioral characteristics such as raised hand in class, resource visits, parent survey responses, and parent school satisfaction. These characteristics include the student's and instructor's progress. The attributes, their description and data type for the dataset are listed in Table 1.

Table 1. Prediction attributes, description and data type.

Category	Attributes	Description	Type
Demographical Attributes	Gender	Student's gender	Nominal
	Nationality	Country of origin	Nominal
	Place of birth	Where the students was given birth	Nominal
	Sponsorship	Father, Mother or guardian of the student	Nominal
Academic background features	Stage ID	Category Student belongs to (low-level, Middle-level, High-level)	Nominal
	Grade ID	Students grade	Nominal
	Class	Students' Classroom	Nominal
	Semester	Semester	Nominal
	Topic	Topics offered	Nominal
	Days Absent	Number of days students was absent	Nominal
Parents participation on learning process	Parent answering survey	Whether parents responds to the survey sent from the school	Nominal
	Parent school satisfaction	Degree of parent satisfaction from school as follow (Good, Bad)	Nominal
Behavioral features	Discussion groups	Student behavior during interaction with Kalboard 360 e-learning System.	Numeric
	Visited resources		Numeric
	Raised hand on class		Numeric
	Viewing announcements		Numeric

B. Data Preprocessing

➢ *Data cleaning*

This is used to deal with missing values or noisy data in a dataset. The dataset used in this study contains 500 records; therefore, after deleting the missing values (20 in number), the dataset contains 480 records. The remaining dataset is being subjected to discretization and resampling, followed by Feature Selection.

➢ *Data Discretization*

Discretization is used to convert numerical features to nominal ones depending on the class data. Discretization tool was used to convert students' performance indicator from numerical to nominal values, which correspond to the classification problem's class labels. Discretization is accomplished by dividing the data set into three nominal intervals (low-level, medium-level and high-level) depending on the final grade of students. The low-level interval includes scores ranging from 0 to 59, the Middle-Level interval includes scores ranging from 60 to 89, and the High-Level interval includes numbers ranging from 80–100. After discretization, the dataset contains 125 records with a low level of proficiency, 209 records with a middle level of proficiency, and 146 record with a high level of proficiency.

➢ *Data Resampling*

This is a technique for condensing large datasets. It generates a stratified subsample from the original dataset. Resample filter that generates an arbitrary subsample of a dataset with or without substitution.

➢ *Feature Selection:*

This process selects a subset of attributes that most accurately describes the data and eliminates repeated and uninteresting ones. The process will enhance the data's quality, hence improving the algorithm's performance. Filter-based feature selection is a method for rating features in which the most highly ranked ones are chosen and the other ones are ignored and put to the algorithm. Numerous feature ranking metrics, such as the Information Gain Ratio, have been presented. Thus, a filter approach based on information gain is employed to determine the best attributes/features for creating the ensemble model in order to study the feature rank.

ISSN No:-2456-2165

IV. EXPERIMENTS AND RESULTS

A. Environment

RapidMiner, a data mining tool was used to assess the proposed ensemble technique. Additionally, we employed 10-fold cross validation to split the dataset into training and testing partitions. Thus, we divided the dataset into ten equal-sized subsets; one subset is used for testing, while the remaining nine are used for training. After ten iterations, the final result is approximated as the average error rate across all test examples.

B. Evaluation Metrics

Four of the most frequently used measures for evaluating the classification results were used in this research. Accuracy, Precision, Recall, and F-Measure are the four components. Calculate the measures by referring to Table 2, which contains a classification confusion matrix based on four equations.

Table 2: Confusion Matrix

		Predicted label	
		Positive	**Negative**
True label	Positive	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

- *Accuracy*

This is the proportion of total number of predictions that were correct. The accuracy of a predictive model is calculated as follows:

$$A(\%) = \frac{\text{TP + TN}}{\text{TP+FN+FP+TN}} X\ 100 \qquad (1)$$

- *Precision*

This is the percentage of correctly classified samples compared to the total number of misclassified and correctly classified samples.

$$P(\%) = \frac{TP}{TP+FP} X\ 100 \qquad (2)$$

- *Recall or Sensitivity*

The proportion of actual positive cases which are correctly identified

$$R(\%) = \frac{TP}{TP+FN} X\ 100 \qquad (3)$$

F-Measure: This is the accuracy of harmonic mean of precision and recall that is the weighted average of the class

$$F(\%) = 2X \frac{Precision\ X\ Recall}{Precision+Recall} X\ 100 \qquad (4)$$

C. Results Evaluation

- *Evaluating Results Using Base Classifiers*

Numerous factors can influence the predicted model when predicting student performance. We considered behavioral characteristics as a significant characteristics that can affect student's academic performance in this work. In Table 3, we demonstrate the impact of behavioral features by utilizing KNN, NB, Disc, DT and SVM). With each classifier, we divided the classification output into two sections: Classification Results with Behavioral Attributes (BA) and Classification Results without Behavioral Attributes (WBA).

Table 3: Classification Algorithms results with Behavioral Attributes (BA) and without Behavioral Attributes (WBA).

Metric	**NB**		**DT**		**KNN**		**DISC**		**SVM**	
Behavioral Attribute	**BA (%)**	**WBA (%)**	**BA (%)**	**WBA (%)**	**BA (%)**	**WBA (%)**	**BA (%)**	**WBA (%)**	**BA (%)**	**WBA (%)**
Accuracy	93	77	94	82	89	69	93	78	92	75
Precision	85	60	85	67	79	47	85	60	84	59
Recall	86	62	87	68	81	48	86	60	83	52
F-Measure	84	50	85	66	79	47	85	59	82	50

As shown in Table 3, the DT model outperforms techniques. The DT model obtained an accuracy of 94 percent when using BA and 82 percent when not using BF. Precision was 85 percent with BA and 67 percent without BA. The recall rate is 87 percent with BA and 68 percent without BA. The results for F-Measure are 85 percent with BF and 66 percent without BA. As a result of the foregoing analysis, it is clear that learner behavior has a significant effect on students' academic performance.

D. Results of the Ensemble Technique

We used ensemble techniques in this subsection to improve the accuracy of the evaluation results from traditional Data Mining techniques. Table 4 illustrates the enhanced results obtained by combining five traditional classifiers using ensemble techniques (SVM, KNN, NB, DISC, DT). Each ensemble trains five classifiers separately and then uses a majority voting to combine the results in order to achieve optimal prediction. For DT and NB algorithms, boosting technique outperform other ensemble techniques, with accuracy increased from 93% to 95%, precision increased from 85% to

ISSN No:-2456-2165

87%, and recall increased from 86% to 87%. Then, for DT, the accuracy is increased from 94% to 96%, the precision from 85% to 88%, and the recall from 87% to 88%.

After training the classification model using 10-fold cross validation, the validation process begins. Validation is a critical phase in the development of predictive models; it determines the predictive models' accuracy. The model is trained on a total of 500 students and validated on 25 newcomer students in this study. The results of evaluation using classification techniques (SVM, KNN, NB, DISC, DT) during the testing and validation stages are summarized in Table 5.

Table 4. Classification results using ensemble techniques.

Evaluation metric	**Classification Algorithm**					**AdaBoost**					**Bagging**					**Random under sampling**
	NB	DT	KNN	DISC	SVM	NB	DT	KNN	DISC	SVM	NB	DT	KNN	DISC	SVM	DT
Accuracy	0.93	0.94	0.89	0.93	0.92	0.96	0.95	0.89	0.92	0.90	0.91	0.94	0.88	0.92	0.90	0.98
Precision	0.85	0.85	0.79	0.85	0.84	0.87	0.87	0.79	0.83	0.82	0.82	0.84	0.78	0.82	0.85	0.90
Recall	0.86	0.87	0.81	0.86	0.83	0.87	0.89	0.81	0.86	0.82	0.82	0.87	0.81	0.84	0.82	0.90
F-Measure	0.84	0.85	0.79	0.85	0.82	0.86	0.86	0.79	0.84	0.82	0.83	0.84	0.78	0.84	0.82	0.89

Table 5. Classification Algorithms results through testing and validation

Evaluation measure	**Testing results**					**Validation results**				
Classifier	**NB**	**DT**	**KNN**	**DISC**	**SVM**	**NB**	**DT**	**KNN**	**DISC**	**SVM**
Accuracy	0.93	0.94	0.89	0.93	0.92	0.96	0.98	0.92	0.96	0.95
Precision	0.85	0.85	0.79	0.85	0.84	0.96	0.99	0.86	0.96	0.94
Recall	0.86	0.87	0.81	0.86	0.83	0.97	0.99	0.90	0.96	0.94
F-Measure	0.84	0.85	0.79	0.85	0.82	0.96	0.99	0.87	0.96	0.94

V. CONCLUSION AND FUTURE WORK

Predicting students' academic performance has long been a source of concern for higher education institutions worldwide. Due to the enormous usability of LMSs, a massive amount of data on communication and interaction between instructors and learners has been generated. The data gathered contains some hidden knowledge that is used to help students improve their academic performance. The purpose of this research is to propose a new students' performance prediction model that is based on a variety of data mining techniques and incorporates new attributes known as behavioral attributes. These attributes are associated with learners' interaction with the LMS. Predictive models are evaluated using a variety of classifiers, namely: SVM, KNN, NB, DISC and DT. Additionally, Bagging, Ada-Boost, and Random Under sampling ensemble techniques were used to improve the classifiers' performance. The findings indicated that there is a significant and robust relationship between learners' behavior and academic performance. The predictive model's accuracy with behavioral features is 94 percent, while without behavioral features it is 82 percent. After applying ensemble methods, it achieved a precision of 98 percent. In future works, we will analyze student data in order to identify additional characteristics that will help us identify students with lower achievement and performance levels. Other ensemble methods could also be used to help students improve their academic performance.

REFRENCES

[1]. Abubakar, Y., & Ahmad, N. B. H. (2017). Prediction of Students Performance in E-Learning Environment Using Random Forest. *International Journal of Innovative Computing*, *7*(2).

[2]. Amrieh, E.A., Hamtini, T., Aljarah, I.: Mining educational data to predict student’s academic performance using ensemble methods. Int. J. Database Theory Appl. **9**, 119–136 (2016)

[3]. Bithari, T. B., Thapa, S., & K.C., H. (2020). Predicting Academic Performance of Engineering Students Using Ensemble Method. *Technical Journal*, *2*(1), 89–98. *https://doi.org/10.3126/tj.v2i1.32845*

[4]. C. R. Romero and S. Ventura, "Educational Data Mining: A Review of the State of the Art", *IEEE Transactions on Systems Man and Cybernetics Part C: Applications and Reviews*, vol. 40, no. 6, pp. 601-618, 2010.

[5]. C. Romero, S. Ventura and E. Garcia, "Data mining in course management systems: Moodle case", *Comput Educ*, vol. 51, pp. 368-384, 2008.

[6]. Costa, E.B., Fonseca, B., Santana, M.A., De Araújo, F.F., Rego, J.: Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses. Comput. Hum. Behav. **73**, 247–256 (2017)

[7]. E.A. Amrieh, H. Thair and A. Ibrahim, "Preprocessing and analyzing educational data set using X-API for improving student's performance", *2015 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT)*, December, 2015.

[8]. G. Kakasevski, M. Mihajlov, S. Arsenovski and S. Chungurski, "Evaluating usability in learning management system Moodle", *ITI 2008-30th International Conference on Information Technology Interfaces*, pp. 613-618, 2008.

[9]. Guo, B., Zhang, R., Xu, G., Shi, C., Yang, L.: Predicting students' performance in educational data mining. In: 2015 International Symposium on Educational Technology (ISET), pp. 125–128. IEEE (2015)

[10]. Kamal, P., & Ahuja, S. (2019). An ensemble-based model for prediction of academic performance of students in undergrad professional course. *Journal of Engineering, Design and Technology*.

[11]. Latham, A., Crockett, K., Mclean, D. Profiling student learning styles with multilayer perceptron neural networks. In: 2013 IEEE International Conference on Systems, Man, and Cybernetics, pp. 2510–2515 (2013)

[12]. Pe, M., Barrag, M.: Feature Engineering based on ANOVA, cluster validity assessment and KNN for fault diagnosis in bearings. J. Intell. Fuzzy Syst. **34**, 3451–3462 (2018)

[13]. R.S. Baker, S. Gowda and A. Corbett, "Automatically Detecting a Student's Preparation for Future Learning: Help Use Is Key", *Proceedings of the 4th International Conference on Educational Data Mining*, pp. 179-188, 2011.

[14]. Satyanarayana, A., & Nuckowski, M. (2016). Data mining using ensemble classifiers for improved prediction of student academic performance.

[15]. Usman, U. I., Salisu, A., Barroon, A. I. E., & Yusuf, A. (2019). A COMPARATIVE STUDY OF BASE CLASSIFIERS IN PREDICTING STUDENTS'PERFORMANCE BASED ON INTERACTION WITH LMS PLATFORM. *FUDMA JOURNAL OF SCIENCES-ISSN: 2616-1370*, *3*(1), 231-239.