An Approach for Early Prediction of Academic Procrastination in e-Learning Environment

Nisha S. Raj* and Renumol V. G.

***Abstract*—Ubiquitous learning is a new educational paradigm partly created by the affordance of digital media. This trend has continued to expand over time. The emergence of ubiquitous computing has created unique conditions for people working as education professionals and learning as students. Procrastination is one of the characteristics that has been seen in students that forces them to set back and sit back without achieving their goals. It has been estimated that almost 70% of college students or even school students engage in frequent academic procrastination and purposive delays in the beginning or completing tasks. Throughout this study, we concentrated on different predictive measures that can be used to identify procrastination behaviour among students. These measures include the usage of ensemble classification models such as Logistic Regression, Stochastic Gradient Descent, K-Nearest Neighbours, Decision Tree and Random Forest. Of these, the random forest model achieved the best predictive outcome with an accuracy of almost 85%. Moreover, earlier prediction of such procrastination behaviours would assist tutors in classifying students before completing any task or homework which is a useful path for developing sustainability in the learning process. A strength of this study is that the parameters discussed can be well defined in both virtual and traditional learning environments. However, the parameters defining students' cognitive or emotional states were not explored in this study.**

***Index Terms*—Procrastination, learning analytics, virtual learning environment, educational data mining, learning patterns.**

I. Introduction

Academic and public interest in mobile and ubiquitous learning is growing, particularly in connection with its implementation in higher-education settings. U-learning includes any environment that allows mobile devices to access learning and other teaching content via wireless networks at any location and time [1]. Adaptability, permanency, and accessibility are key characteristics of ubiquitous learning. Previous research has demonstrated that u-learning has constructive effects on the teaching/learning process [2]. However, in contrast to face-to-face interactions, this mode prevents teachers from clearly evaluating the status of learning progression [3]-[5]. As a result of the technological advancements we are witnessing, a large percentage of students have developed a sense of procrastination [6], [7]. Procrastination is trouble persuading oneself to do things that you should or would like to do. Some studies indicate that students' procrastination has a major impact on their performance in online learning environments [8], [9]. Therefore, it is important for educators and teachers to be aware of procrastination behaviour. In this study, we investigated how to predict academic procrastination at an earlier stage by exploring learner data using six different machine-learning algorithms. The earlier prediction of procrastination behaviour would assist tutors in classifying students before completing the assigned tasks or homework which is useful and builds a sustainable learning path [10]. To achieve this, we first used a logistic regression model as the baseline model in order to check the initial prediction accuracy. We have then used different prominent machine learning models such as Naive Bayes Classifier, Decision Tree, Random Forest, K-Nearest Neighbours, and Stochastic Gradient Descent, to find which one of them would fetch us the highest prediction accuracy.

Understanding the procrastination behaviours of students that affect their learning processes will help improve their self-learning behaviour [11]. Early prediction of procrastination behaviour in students helps teachers understand students before completing tasks and helps them change their learning path to more adaptable ones [8]. By taking advantage of machine learning, this study proposes accurate prediction models compared with the baseline model. The major contributions of this study are summarized as follows: 1. A machine learning model to predict academic procrastination at an early stage of the course, considering students' task completion activities. 2. The evaluation of the model is verified by using an annotated dataset.

II. Research Objective

To develop a classification model that can predict academic procrastination by identifying unique patterns in a dataset.

Research Questions

1) How accurately do the chosen algorithms would predict academic procrastination and determine students' current status?
2) Which of these classification algorithms achieved the best prediction accuracy and offered superior predictive power?

III. Literature Study

In this section, we discuss studies of academic procrastination, its prediction, impacts, and related topics. "Educational Sustainability through Big Data Assimilation to Quantify Academic Procrastination Utilizing Ensemble Classifiers" is a data-driven methodology for forecasting student procrastination in web-based online homework [12].

Manuscript received May 15, 2022; revised June 13, 2022; accepted June 24, 2022.

The authors are with the School of Engineering, Cochin University of Science and Technology, Kerala, India.

*Correspondence: nishasraj@cusat.ac.in

doi: 10.18178/ijiet.2023.13.1.1782

They pointed out that, when students use an intelligent tutoring system to complete assignment activities, they engage in various actions that may be used to quantify procrastination. The early detection of student procrastination can also aid in creating a friendly and feasible educational environment that satisfies societal responsibilities.

According to previous studies, students with lower procrastination tendencies achieve higher procrastination tendencies than do students with higher procrastination tendencies. As a result, it is critical for teachers to be aware of students' behavior, particularly their procrastinating habits. Some EDM techniques can be used to analyse data obtained through computer-supported learning environments and to estimate student behaviour according to the authors of [13].

According to an extensive amount of research, students' procrastination habits have been found to be a significant factor affecting their performance in online learning. In their research suggested a novel algorithm that uses the assignment submission behaviour of learners to forecast performance by learning problems via procrastinating behaviour [8]. Predicting student performance and deficiencies in mastering MOOC knowledge points using multitask learning allows people to offer academic help in intervening and enhancing the learning outcomes of students confronting problems [14].

Some research on procrastination, achievement goal orientations, and learning strategies [15] found that students who set the goal of learning everything was less likely to procrastinate, while those who set the goal of avoiding failing to learn everything were more likely to procrastinate. These findings highlight the necessity of separately analysing the approach and avoidance types of mastery orientation when identifying their relationship with procrastination.

The rapid rise in educational data suggests that extracting large amounts of data will require a more sophisticated collection of algorithms. As a result, the discipline of educational data mining (EDM) was created. A comprehensive analysis of the clustering algorithm's applicability and usefulness in the context of EDM [16] leads to the finding that traditional data mining algorithms, which may have a specific aim and function, cannot be directly applied to educational problems. This means that a preprocessing procedure must be implemented first, followed by the application of appropriate data-mining approaches. Clustering is one such EDM preprocessing method. The application of numerous data-mining methods to educational attributes has been the goal of many EDM studies. Based on the existing literature, this study also presents some future insights into educational data clustering and proposes other research avenues.

Prediction is an important aspect of educational data-mining. Kaur *et al.* identified an approach using classification-based techniques to identify slow learners among the pupils [17]. They collected high school student data and used WEKA, the Waikato Environment for Knowledge Analysis, to determine the parameters that affect student behavior. They then used various categorization techniques. This study highlights the importance of using classification and prediction techniques in educational data analysis.

In a case study, Azeiteiro *et al.* evaluated the effectiveness of education on sustainable development through e-learning in higher education [18]. The expectations and experiences of students enrolled in more than one science program at Portuguese Distance Learning University were evaluated. The findings revealed that the students surveyed claimed they had achieved a high sense of motivation and satisfaction, as well as an effective learning outcome in terms of knowledge, skills, values, attitudes, and behaviour in environmental and/or sustainability sciences. They proved that education for sustainable development in an e-learning environment can contribute to the transition to more sustainable behavioral trends. Romero and Ventura reviewed recent studies in this area [19]. They began by introducing the EDM and describing various user groups, curricula, and the generated data. It then identifies some of the major issues in the learning environment that have been solved using data-mining approaches, followed by a discussion of some of the most promising future research lines. The EDM has been introduced as a new research area connected to e-learning, adaptive hypermedia, intelligent tutoring systems, web mining, data mining, and other well-established research areas. Mandalapu and Gong analyzed the influence of various attributes collected by the ASSISTments online learning platform on the performance of machine learning algorithms in predicting student career fields [20]. The ASSISTments platform captured US middle school student interaction data from 2004 to 2007. This dataset consisted of the system interaction data of 1709 students. Thus, this study explored the importance of feature selection and investigates the local linear correlation of predictors.

The literature is rich in studies related to predicting student engagement and its relationship with academic procrastination [21], [22]. These studies suggest that non-engaging students can be considered procrastinators.

Student engagement is highly correlated with statistics on how they interact with the ITS interface. Many studies have focused on student engagement on virtual learning platforms. The number of clicks is a parameter which can quantify the interaction behaviour of students [23]. However, interaction data is absent in many existing e-learning environments. Therefore, it is difficult to plug in a model with a complex parameter set to predict or detect whether the engagement quotient is tedious or not. Studies involving traditional classroom approaches are not suitable for the virtual world, because the mode of instruction is mostly learner-centric [24], [25].

Abidi *et al.*, worked with predictive measures to identify and predict student procrastination using ensemble machine learning models [12]. This study used a data-driven method to forecast students, and introduced a data-driven methodology for predicting student procrastination in web-based online homework using an intelligent tutoring system (ASSISTment). Researchers have concluded that, when students attempt homework activities on ITS, they undertake different actions that can be used to measure procrastination are closely connecting the ideas with the work discussed in this paper [12]. However, they did not use parameters related to the postponement of tasks, which is a strong indicator in any learning task, whether virtual or physical. In addition, their design was heavily auto-tuned, and the authors doubted the

reliability of the model for the same reason. The proposed work uses parameters that directly define procrastination and are chosen by educational experts with reference to the correlation analysis of each parameter with the procrastinator class. Statistical measures were used to weigh the parameters used to generate the model.

IV. Design and Flowchart

The ASSISTment dataset is enormous. The dataset should be adequately pre-processed to obtain better predictions from the models. We used feature extraction technique to take a subset of the dataset because processing a large dataset is challenging. The subsets were divided into two groups, training and testing. The training set was used to train the models and the testing set was used to test the trained models. In this case, 80 percentage of the pre-processed dataset is used for training, while the remaining 20 percentage is used for testing. The models were trained using the training samples. The fitted models were run with the test dataset, and the prediction performance of the model was evaluated. Variables such as accuracy, precision, recall, F1 score, and ROC were considered to evaluate the performance of the different models [26].

- Accuracy: correctly predicted observations from all observations.

$$\text{Accuracy} = (TP + TN) / (TP + FP + TN + FN)$$

- Precision: a measure of a classifier's accuracy

$$\text{Precision} = TP/ (FP + TP)$$

- Recall: Measure classifier completeness.

$$\text{Recall} = TP / (FP + FN)$$

- F1 score: F1 score conveys the balance between precision and recall.

$$\text{F1 score} = 2 *((\text{precision} * \text{recall}) / (\text{precision} + \text{recall}))$$

A confusion matrix was used to evaluate the performance of the model. The confusion matrix in the two-class label problem is a matrix with four entries: true positive (TP), false negative (FN), false positive (FP), and true negative (TN) (TN). TP denotes the number of procrastinating students who were correctly predicted, FP denotes the number of non-procrastinating student who have been predicted as procrastinating student, FN denotes the number of procrastinating students who have been predicted as non-procrastinating, and TN denotes the number of non-procrastinating students who were correctly predicted.

Fig 1 shows the chronological steps of the proposed methodology. The first step was to identify the data sources. The selected dataset was the ASSISTments dataset provided by the ASSISTments team for public use in 2017. The dataset consisted of clickstream log files describing students' interactions using the ASSISTments software. The dataset contained student actions from 2004 to 2006. The next step was to pre-process the data. The data can be in various formats, including structured tables, pictures, audio files, and movies. Data cleaning was performed to remove the incorrect, corrupted, and incorrectly formatted values. Missing values are checked and replaced with the mean values of the corresponding attributes.

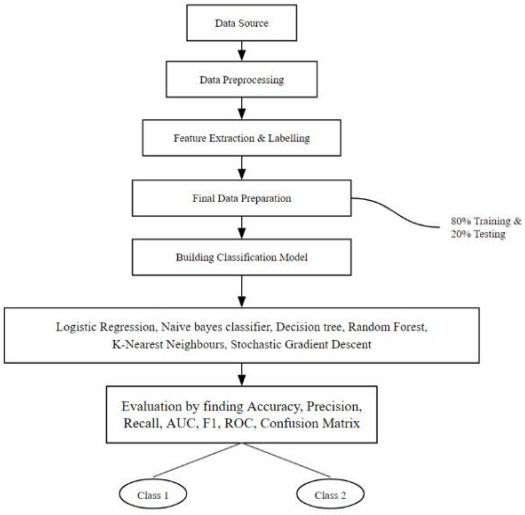


Fig. 1. Sequential steps in the proposed system.

A data preprocessing step is then performed, followed by feature extraction and labelling. Feature extraction helps minimise the length of the features in the dataset, thereby reducing the complexity of the model. Other benefits of feature extraction approaches include higher accuracy, reduced overfitting risk, faster training, and enhanced data visualization.

Data preparation includes data pre-processing, cleaning, validation, profiling, and transformation. The dataset was separated in the final phase of data preparation, with 80 percent designated as the training set and the remaining 20 percent designated as the testing dataset. The next step is to build a classification model. The classification techniques used in large transactional databases make it easier for users to obtain the information they need from large datasets. The two primary classification techniques are unsupervised and supervised. In this study, we examined six different classification algorithms to determine how they affect our goal of predicting academic procrastination. Logistic regression, naive Bayes classifier, decision tree, random forest, K-nearest neighbours, and stochastic gradient descent were the six classification algorithms chosen to build the models. Evaluation metrics, such as accuracy, precision, recall, AUC, F1, and ROC, were used to evaluate these models. The output of our model was similar to the students being classified into two classes: Class 1 will consist of those highly engaged students showing low procrastination, and Class 2 will consist of those low-engaged students exhibiting high procrastination.

V. Methodology

A. ASSISTments Dataset

This study adopts the ASSISTments dataset provided during the 2017 Educational Data Mining (EDM) competition [27], [28]. The ASSISTments platform captured US middle school student interaction data from 2004 to 2007.

This dataset consisted of the system interaction data of 1709 students. These students were requested to participate in a survey conducted to record their post-high school career achievement. This survey provided the career choices of 591 students. Of these, 466 students belonged to the non-STEM field and 125 students belonged to the STEM field. The data were structured in a. csv file. This dataset contains 82 attributes related to a student and details about different assignments taken by the student. Some of the attributes are hintcount, which is the number of hints taken, gender of the student, time taken for the completion of an assignment, and so on.

B. *Pre-processing of Dataset*

The first stage in data pre-processing is the data-cleaning process. Data cleaning is a process of repairing or deleting incorrect, corrupted, poorly formatted, duplicate, or incomplete data from a dataset [29]. There are numerous ways in which data can be duplicated or mislabelled when merging multiple data sources. First, we checked for missing values in the data set. We identified that two attributes, InferredGender and isSTEM, contain a large number of missing values. Two options to fill the values in these fields are either to fill with the mean or median of other values, or to fill with random values of the attributes. However, doing so did not have a positive impact on students' procrastination behavior. Hence, these two attributes were removed from the dataset and were not considered for further studies. The remaining missing values were updated using the mean value of the corresponding attributes. To make the initial dataset suitable for this study. We conducted a detailed study on every attribute present in the dataset, and from our observations, we identified the attribute that positively affects student procrastinating behavior. We then formed certain rules to calculate the total score for each sample using the identified attributes. Table1 shows the rules and weights used for the selected attributes. The weights and rules are fixed by running trials on the existing dataset to obtain greater correlation with the procrastination nature of the students.

TABLE I: THE FEATURE DESCRIPTION TABLE

Feature	Weight	Rule
AveKnow	3	<= 0.45 then procrastination , >0.45 then no procrastination
AveCarelessness	5	>= 0.36 then procrastination , <0.36 then no procrastination
AveCorrect	2	<= 0.58 then procrastination ,>0.58 then no procrastination
NumActions	4	<=1500 then procrastination , >1500 then no procrastination
AveResBored	4	>=0.25 then procrastination, < 0.25 then no procrastination
AveResOfftask	5	>= 0.55 then procrastination , < 0.55 then no procrastination
timeTaken	5	>= 25 seconds then procrastination. <25 seconds then no procrastination
hintCount	4	>=5 then procrastination
original	3	if value is 0 then procrastination
attemptCount	4	<= 50 then procrastination >50 no procrastination
totalFrAttempted	5	=0 then procrastination, no procrastination
consecutiveErrorsInRow	4	=0 then no procrastinate, otherwise procrastinate
sumTimePerSkill	4	<602 then procrastination, >= 602 no procrastination
RES_CONCENTRATING	3	<0.65 then procrastination, >=0.65 no procrastination
RES_OFFTASK	3	>=0.17 then procrastination, else no procrastination

The total score for each sample was calculated using these rules and weights. We then used two threshold values: the mean and median values of the total score. Here, the mean value was 22, and the median was 21. Subsequently, using these two threshold values and total scores, we assigned a class label for each sample. is_procrastinate will have values of 1 and 0, 1 denotes a procrastinating student, and 0 denotes a non-procrastinating student. We assign the class label based on the mean and median values of the total score; thus, we will have two datasets, one based on the mean value and the other based on the median value. We plan to conduct experiments on these two datasets in a similar manner. Fig. 2 and Fig. 3 depict parts of the mean and median datasets.

C. *Feature Extraction*

We used a correlation-based feature selection technique. We used three methods to determine the correlation values of each attribute with is_procrastinate. The three methods are the Pearson correlation, Kendall rank correlation, and Spearman correlation [30], [31]. Pearson correlation is one of the most commonly used correlations. This corresponds to the covariance of the two variables, normalized (i.e., divided) by the product of their standard deviations. The Spearman correlation between two variables is equal to the Pearson correlation between the rank scores of those two variables; however, while Pearson's correlation analyses linear correlations, Spearman's correlation assesses monotonic relationships. Because it has lower gross error sensitivity (GES) and asymptotic variance (AV), the Kendall correlation is chosen over the Spearman correlation, making it more robust and efficient.

We calculated the correlation value of each attribute with the is_procrastinate value for both the mean and median datasets. Then the attribute which shows positive and high correlation value with is_procrastinate value in all the three methods and in both the datasets is taken into consideration. These attributes were considered in the creation of the classification model. The other attribute was eliminated from the dataset. Table II shows the selected features, and their correlation values with is the procastinate value.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	AveKnow	AveCarele	AveCorrec	AveResBo	AveResCo	AveResOff	action_nu	timeTaker	consecutiv	sumTime3	timeOver8	confidenc	confidenc	confidenc	RES_BORE	RES_CONF	RES_OFFT.	is_procrastinate	
2	0.352416	0.183276	0.483902	0.208389	0.115905	0.156503	9950	49	0	0	0	0.597865	0	0.83871	0.376427	0	0.785585	1	
3	0.352416	0.183276	0.483902	0.208389	0.115905	0.156503	9951	4	0	0	0	0.355694	0.9375	0.6	0.156027	0.887452	0.468252	1	
4	0.352416	0.183276	0.483902	0.208389	0.115905	0.156503	9952	6	0	0	0	0.355694	0.9375	0.6	0.156027	0.887452	0.468252	1	
5	0.352416	0.183276	0.483902	0.208389	0.115905	0.156503	9953	18	1	0	0	0.355694	0	0.204082	0.156027	0	0.108417	1	
6	0.352416	0.183276	0.483902	0.208389	0.115905	0.156503	9954	2	0	0	0	0.355694	0	0.204082	0.156027	0	0.108417	0	
7	0.352416	0.183276	0.483902	0.208389	0.115905	0.156503	9955	59	0	0	0	0.597865	0	0.83871	0.376427	0	0.785585	1	
8	0.352416	0.183276	0.483902	0.208389	0.115905	0.156503	9956	15	0	0	0	0.355694	0	0.204082	0.156027	0	0.108417	0	
9	0.352416	0.183276	0.483902	0.208389	0.115905	0.156503	9957	61	0	0	0	0.597865	0.378151	0.83871	0.376427	0.060808	0.785585	1	
10	0.352416	0.183276	0.483902	0.208389	0.115905	0.156503	9958	18	0	0	0	0.355694	0.378151	0.9375	0.156027	0.060808	0.916914	1	
11	0.352416	0.183276	0.483902	0.208389	0.115905	0.156503	9959	0	0	0	0	0.355694	0.378151	0.9375	0.156027	0.060808	0.916914	1	
12	0.352416	0.183276	0.483902	0.208389	0.115905	0.156503	9960	14	0	0	0	0.355694	0	0.166667	0.156027	0	0.088541	0	
13	0.352416	0.183276	0.483902	0.208389	0.115905	0.156503	9961	8	0	0	0	0.355694	0	0.6	0.156027	0	0.468252	1	
14	0.352416	0.183276	0.483902	0.208389	0.115905	0.156503	9962	28	0	0	0	0.597865	0.79646	0.83871	0.376427	0.633474	0.785585	1	
15	0.352416	0.183276	0.483902	0.208389	0.115905	0.156503	9963	51	0	0	0	0.355694	0	0.204082	0.156027	0	0.108417	1	
16	0.352416	0.183276	0.483902	0.208389	0.115905	0.156503	9964	29	0	0	0	0.597865	0.79646	0.83871	0.376427	0.633474	0.785585	1	
17	0.352416	0.183276	0.483902	0.208389	0.115905	0.156503	9965	4	0	0	0	0.355694	0	0.204082	0.156027	0	0.108417	0	
18	0.352416	0.183276	0.483902	0.208389	0.115905	0.156503	9966	83	0	0	1	0.355694	0	0.280702	0.156027	0	0.149121	1	
19	0.352416	0.183276	0.483902	0.208389	0.115905	0.156503	9967	51	0	0	0	0.597865	0.378151	0.222222	0.376427	0.060808	0.118054	1	
20	0.352416	0.183276	0.483902	0.208389	0.115905	0.156503	9968	8	0	0	0	0.355694	0	0.204082	0.156027	0	0.108417	0	
21	0.352416	0.183276	0.483902	0.208389	0.115905	0.156503	9969	9	0	0	0	0.355694	0	0.204082	0.156027	0	0.108417	0	
22	0.352416	0.183276	0.483902	0.208389	0.115905	0.156503	9970	21	0	0	0	0.355694	0	0.6	0.156027	0	0.468252	1	
23	0.352416	0.183276	0.483902	0.208389	0.115905	0.156503	9971	15	0	-0.0585	0	0.355694	0	0.6	0.156027	0	0.468252	1	
24	0.352416	0.183276	0.483902	0.208389	0.115905	0.156503	9972	36	0	0	0	0.597865	0.79646	0.222222	0.376427	0.633474	0.118054	1	
25	0.352416	0.183276	0.483902	0.208389	0.115905	0.156503	9973	18	0	0	0	0.355694	0	0.204082	0.156027	0	0.108417	0	
26	0.352416	0.183276	0.483902	0.208389	0.115905	0.156503	9974	12	0	0	0	0.355694	0	0.6	0.156027	0	0.468252	1	

newmeandataset

Fig. 2. Mean dataset after data preprocessing stage.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	AveKnow	AveCarele	AveCorrec	AveResBo	AveResCo	AveResOff	action_nu	timeTaker	consecutiv	sumTime3	timeOver8	confidenc	confidenc	confidenc	RES_BORE	RES_CONF	RES_OFFT.	is_procrastinate	
2	0.352416	0.183276	0.483902	0.208389	0.115905	0.156503	9950	49	0	0	0	0.597865	0	0.83871	0.376427	0	0.785585	1	
3	0.352416	0.183276	0.483902	0.208389	0.115905	0.156503	9951	4	0	0	0	0.355694	0.9375	0.6	0.156027	0.887452	0.468252	1	
4	0.352416	0.183276	0.483902	0.208389	0.115905	0.156503	9952	6	0	0	0	0.355694	0.9375	0.6	0.156027	0.887452	0.468252	1	
5	0.352416	0.183276	0.483902	0.208389	0.115905	0.156503	9953	18	1	0	0	0.355694	0	0.204082	0.156027	0	0.108417	1	
6	0.352416	0.183276	0.483902	0.208389	0.115905	0.156503	9954	2	0	0	0	0.355694	0	0.204082	0.156027	0	0.108417	0	
7	0.352416	0.183276	0.483902	0.208389	0.115905	0.156503	9955	59	0	0	0	0.597865	0	0.83871	0.376427	0	0.785585	1	
8	0.352416	0.183276	0.483902	0.208389	0.115905	0.156503	9956	15	0	0	0	0.355694	0	0.204082	0.156027	0	0.108417	0	
9	0.352416	0.183276	0.483902	0.208389	0.115905	0.156503	9957	61	0	0	0	0.597865	0.378151	0.83871	0.376427	0.060808	0.785585	1	
10	0.352416	0.183276	0.483902	0.208389	0.115905	0.156503	9958	18	0	0	0	0.355694	0.378151	0.9375	0.156027	0.060808	0.916914	1	
11	0.352416	0.183276	0.483902	0.208389	0.115905	0.156503	9959	0	0	0	0	0.355694	0.378151	0.9375	0.156027	0.060808	0.916914	1	
12	0.352416	0.183276	0.483902	0.208389	0.115905	0.156503	9960	14	0	0	0	0.355694	0	0.166667	0.156027	0	0.088541	0	
13	0.352416	0.183276	0.483902	0.208389	0.115905	0.156503	9961	8	0	0	0	0.355694	0	0.6	0.156027	0	0.468252	1	
14	0.352416	0.183276	0.483902	0.208389	0.115905	0.156503	9962	28	0	0	0	0.597865	0.79646	0.83871	0.376427	0.633474	0.785585	1	
15	0.352416	0.183276	0.483902	0.208389	0.115905	0.156503	9963	51	0	0	0	0.355694	0	0.204082	0.156027	0	0.108417	1	
16	0.352416	0.183276	0.483902	0.208389	0.115905	0.156503	9964	29	0	0	0	0.597865	0.79646	0.83871	0.376427	0.633474	0.785585	1	
17	0.352416	0.183276	0.483902	0.208389	0.115905	0.156503	9965	4	0	0	0	0.355694	0	0.204082	0.156027	0	0.108417	0	
18	0.352416	0.183276	0.483902	0.208389	0.115905	0.156503	9966	83	0	0	1	0.355694	0	0.280702	0.156027	0	0.149121	1	
19	0.352416	0.183276	0.483902	0.208389	0.115905	0.156503	9967	51	0	0	0	0.597865	0.378151	0.222222	0.376427	0.060808	0.118054	1	
20	0.352416	0.183276	0.483902	0.208389	0.115905	0.156503	9968	8	0	0	0	0.355694	0	0.204082	0.156027	0	0.108417	0	
21	0.352416	0.183276	0.483902	0.208389	0.115905	0.156503	9969	9	0	0	0	0.355694	0	0.204082	0.156027	0	0.108417	0	
22	0.352416	0.183276	0.483902	0.208389	0.115905	0.156503	9970	21	0	0	0	0.355694	0	0.6	0.156027	0	0.468252	1	
23	0.352416	0.183276	0.483902	0.208389	0.115905	0.156503	9971	15	0	-0.0585	0	0.355694	0	0.6	0.156027	0	0.468252	1	
24	0.352416	0.183276	0.483902	0.208389	0.115905	0.156503	9972	36	0	0	0	0.597865	0.79646	0.222222	0.376427	0.633474	0.118054	1	
25	0.352416	0.183276	0.483902	0.208389	0.115905	0.156503	9973	18	0	0	0	0.355694	0	0.204082	0.156027	0	0.108417	0	
26	0.352416	0.183276	0.483902	0.208389	0.115905	0.156503	9974	12	0	0	0	0.355694	0	0.6	0.156027	0	0.468252	1	

newmediandataset

Fig. 3. Median dataset after the data preprocessing stage.

D. Model Validation

We used the Anaconda data science platform to train and test the predictive models. In this study, we evaluated six machine learning models. We selected logistic regression, naive Bayes classifier, decision tree, random forest, K-nearest neighbours, and stochastic gradient descent [32]-[34]. From this logistic regression, was considered as the baseline model. Logistic regression models the probabilities of possible outcomes of a single trial. The naïve Bayes classifier uses Bayes' theorem, assuming a relationship between the feature pairs. A decision tree is used to develop a hierarchical model of decisions and probable outcomes. The inner nodes represent the events, and the leaf nodes represent the outcomes. The edges represent the costs and probabilities incurred. The random forest classifier tries to fit a few decision trees in the subsamples of the dataset and uses the mean value to improve the performance of the model and avoid overfitting. The sub sample is created by replacement and are of same size as the original dataset.

The KNN considers a group of labelled points and uses them to learn how to label other points. The new data are labelled by considering the similarity between them and the already labelled points. The stochastic gradient descent (SGD) is an effective approach for fitting linear models. SGD is useful when the sample size is large, as it supports different loss functions and penalties for classification.

TABLE II: SELECTED FEATURES AND THEIR CORRELATION VALUES WITH ARE PROCASTINATED VALUES

Features	Correlation value
AveKnow	0.070230374
AveCarelessness	0.065049633
AveCorrect	0.136910239
AveResBored	0.354712744
AveResConf	0.203335259
AveResOfftask	0.322842984
action_num	0.116029646
timeTaken	0.354265338
consecutiveErrorsInRow	0.185657777
sumTime3SDWhen3RowRight	0.044596677
timeOver80	0.278831743
confidence(BORED)	0.357612908
confidence(CONFUSED)	0.166222236
confidence(OFF TASK)	0.345427163
RES_BORED	0.358275926
RES_CONFUSED	0.16622138
RES_OFFTASK	0.345216486

The experimental setup used the standard settings for these models. Numpy, Pandas, Matplotlib, and Sklearn were the library modules used in the implementation. We have also attempted to implement the support vector, but it takes more than seven hours to train the model because the number of samples is large. So, we ignored it. We trained these models with the mean dataset and evaluated the performance of each model by calculating its accuracy, precision, recall, and F1-score. The same procedure was performed using the median dataset. The models were trained using the median dataset, and the performance of each model was evaluated by calculating its accuracy, precision, recall, and F1 score.

VI. Results and Discussion

Finally, the two datasets that were considered, namely, the mean and median datasets, were fed to each of the six machine learning models separately, including the base model. The optimal combination of relevant features was chosen based on a correlation study, considering the impact that each feature can have on predicting academic procrastination. Of the six machine learning models that were trained and tested, the random forest model outperformed all other models, achieving an overall accuracy of approximately 87%. The results are in correlation with studies in the field of learning analytics using machine learning approaches [12], [20], [23].

Several methods are available for evaluating the classification models. Accuracy is the most intuitive performance measure and is simply the ratio of correctly predicted observations to total observations. Accuracy is a significant measure, but only when we have symmetric datasets, where the values of false positives and false negatives are almost the same. Therefore, we had to look at other parameters to evaluate the performance of our model. Precision is the ratio of the correctly predicted positive observations to the total number of predicted positive observations. High precision is associated with a low false positive rate. The recall is the ratio of correctly predicted positive observations to all the observations in the actual class. The F1 score is the weighted average of precision and recall. Accuracy works best if false positives and false negatives have similar costs. If the costs of false positives and false negatives are very different, it is better to consider both precision and recall. AUC or area under the curve is also widely chosen a performance metric. The evaluation used two datasets extracted from the ASSISTments dataset. One dataset, the mean dataset, is prepared by taking the mean value of the selected parameters, as discussed in the Data Pre-processing section. Similarly, the median dataset was prepared using the median values of the selected parameters from the large ASSISTment dataset. Fig. 4 and Fig. 5 show the performance analysis of the models with the mean and median datasets, respectively.

The X-axes in Fig. 4 and Fig. 5 represent the different models studied. The Y-axis represents the performance quotient on a scale of 0-1. We plotted the accuracy, precision, recall, AUC, and F1 score of each model. From Fig. 4 and Fig. 5, we can observe that the model with RF as the classification algorithm outperformed all other models. The accuracy of RF was 83% for the mean dataset and 87% for the median dataset. The basic nature of the RF algorithm is to consider a few subsamples from the input dataset that are of equal length and try to fit the model with these subsamples. Here, it is more powerful in training a large dataset, such as the ASSISTments dataset, when compared with other common machine learning algorithms.

Fig. 4. Performance analysis of different models when evaluated with the mean dataset.

The random forest-based model showed a steady performance with the median dataset compared to the mean dataset. This was because of the inherent variability in the mean dataset, as we considered the average of the parameter values. The median dataset always had the highest value recorded for any parameter, and the RF fit comfortably with those values. Thus, the RF algorithm shows better results in all evaluation metrics than the other models because it can work with a stronger median dataset. Fig. 6 shows the performance of the RF model with mean and median datasets.

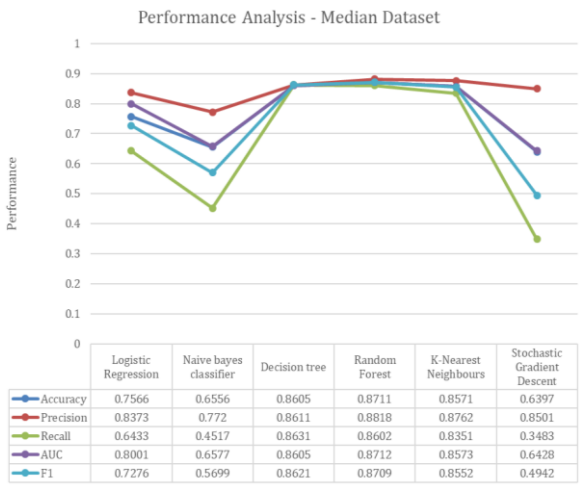
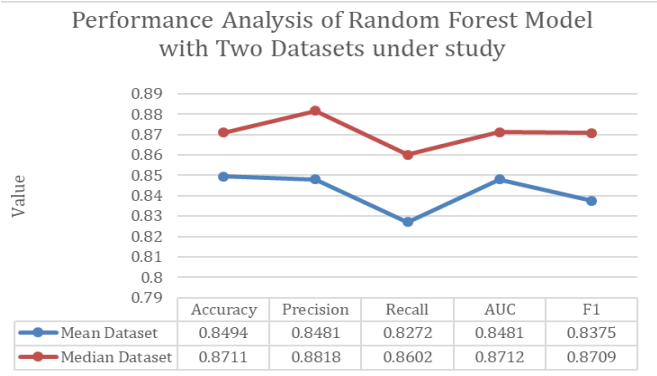


Fig 5. Performance Analysis of different models when evaluated with the mean dataset.

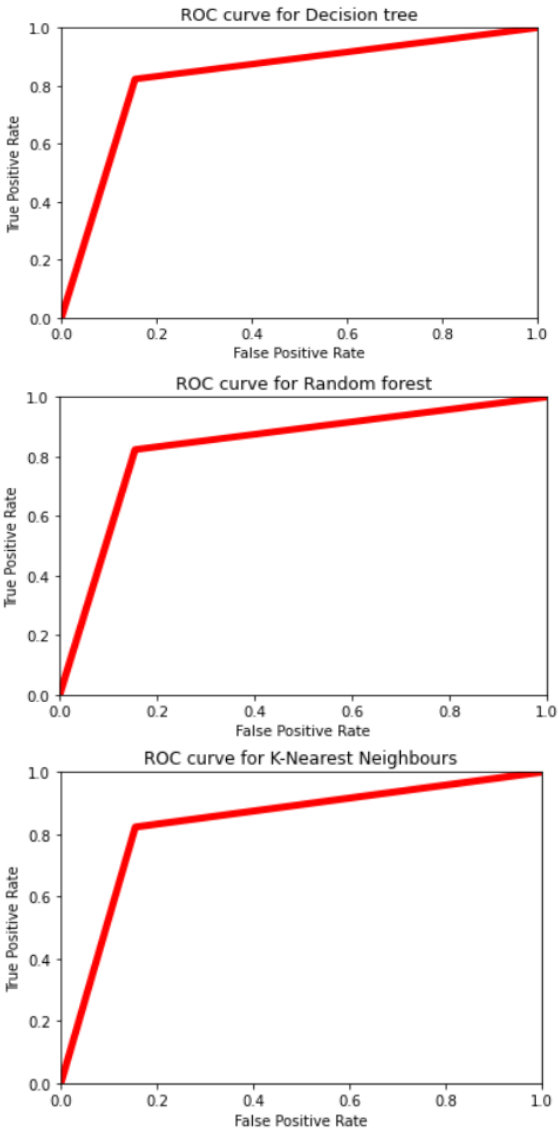
In their paper Abidi *et al.*, discussed the analysing of procrastinating behaviour and found that the tree-ensemble models are better in identifying student procrastination [12]. Autotuning was used for the parameter extraction. In our study, we analysed the correlation of each variable with the procrastinating behaviour of the student. The dataset is annotated with procrastination or non-procrastination

categories through statistical analysis and adjusting the weight of the parameters involved in obtaining a better prediction performance in the training phase. Thus, data pre-processing and feature extraction are performed in a more controlled manner. The student engagement studies, which are closely related to the procrastination studies also concludes that, out of the basic machine learning prediction logics, the random forest works well for a longer feature set [23], [34].



	Accuracy	Precision	Recall	AUC	F1
Mean Dataset	0.8494	0.8481	0.8272	0.8481	0.8375
Median Dataset	0.8711	0.8818	0.8602	0.8712	0.8709

Fig. 6. Performance analysis of random forest model with two datasets under study.



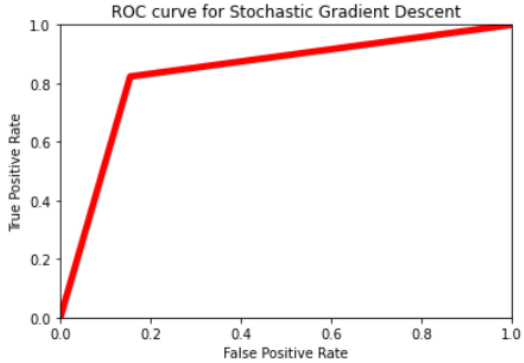
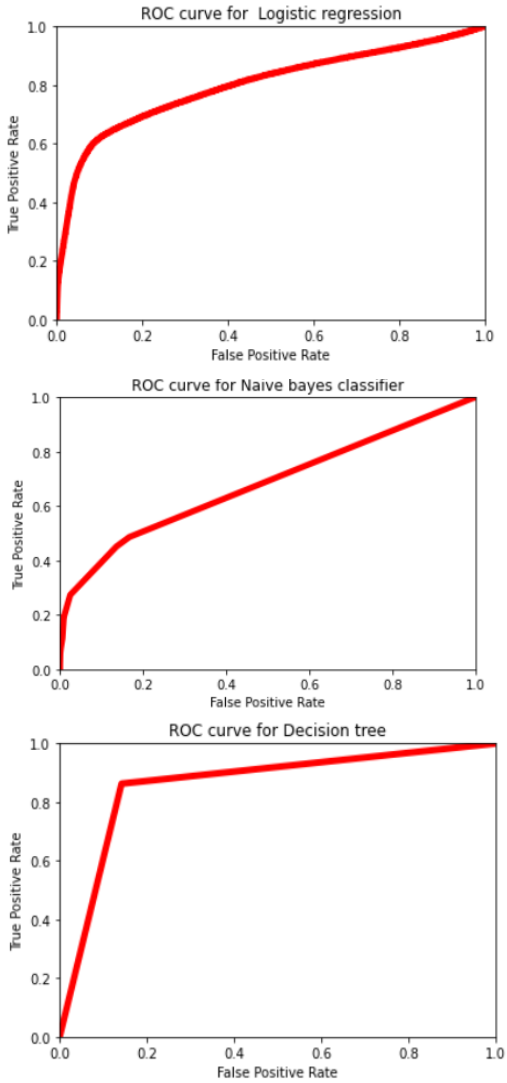


Fig. 7. ROC curves for the mean dataset.

An ROC curve was generated by plotting the false-positive rate of a model against its true-positive rate for each possible cut-off value. The area under the curve (AUC) was then calculated and used as a metric to show how well the model can classify data points. The following tables show the results obtained after the successful implementation and testing of our models using the mean and median datasets. The evaluation metrics chosen were the accuracy, precision, recall, F1, and AUC. For the median dataset, Random Forest model achieved the best prediction accuracy of approximately 87%. Fig. 7 and Fig. 8 shows the resulting ROC curves.



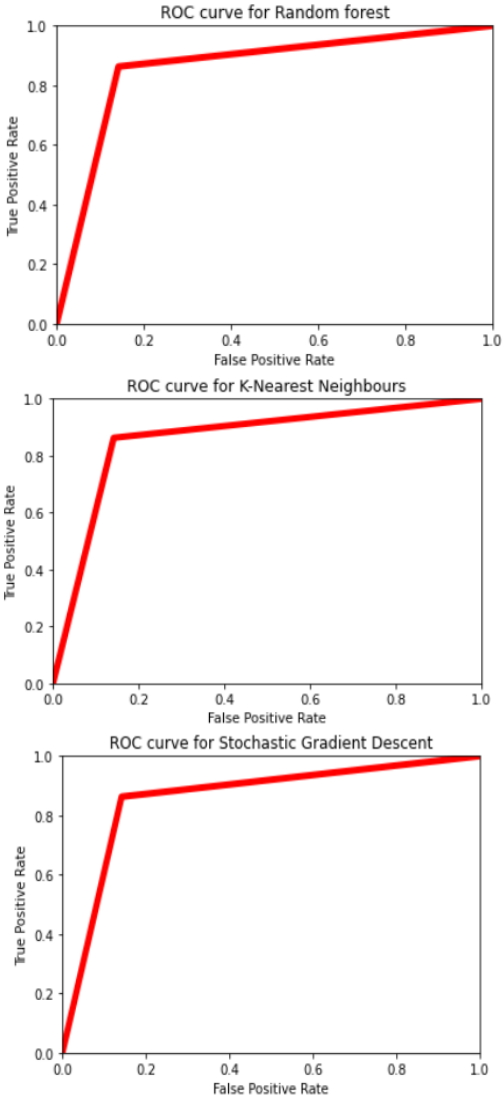


Fig. 8. ROC curves for median dataset.

VII. Conclusion and Future Scope

Several researchers have identified students' procrastination as an essential element that negatively influences their online learning performance, making its prediction a very helpful task for both educators and students. In this study, we used six machine learning algorithms to predict academic procrastination among students by analysing student data. For this purpose, we used the ASSISTments dataset, which was made public and provided during the educational mining competition held in 2017.It consisted of clickstream log files describing students' interaction with ASSISTment's tutoring system. Because there were no features or attributes in the dataset with values related to academic procrastination, our first task was to formulate the output column based on a predefined function using weights assigned to the relevant features. The next stage was to select relevant features for model implementation. The final set of features was chosen based on a correlation study and taking into consideration the influence each of these features can have on predicting academic procrastination. Six machine-learning classification models were chosen and explored: logistic regression, naive Bayes classifier, decision tree, random forest, K-nearest neighbours, and stochastic gradient descent. The random forest model outperformed all other models, achieving an overall accuracy of 87% when using the median dataset. These earlier projections have the potential to increase tutor effectiveness and assist students in need, thereby enhancing educational sustainability.

A future perspective in this area can identify more prominent features that could help predict procrastination by performing more experiments using real-time interactive environments. This model can also be implemented in a real-world scenario by taking data from a real-world case, and should identify how the model works and how the best performance can be obtained. More technologies can also be incorporated, such as different classification algorithms or deep learning methods, to enhance the performance of these models.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Both the authors discussed and expanded the methodology of the work. The first author of the paper, Nisha S Raj conducted the experiments and wrote the manuscript. The second author Dr Renumol V G revised the experimentation and discussion sections of the paper and both agreed on the final draft of the paper.

References

[1] M. Abbasi, S. Dargahi, Z. Pirani, and F. Bonyadi, "Role of procrastination and motivational self-regulation in predicting students' academic engagement," *Iranian Journal of Medical Education*, vol. 15, pp. 160-169, Apr. 2015.

[2] S. S. M. R Abidi, W. Zhang, S. A. Haidery, S. S. Rizvi, R. Riaz, H. Ding, and S. J. Kwon, "Educational sustainability through big data assimilation was used to quantify academic procrastination using ensemble classifiers," *Sustainability*, vol. 12, no. 15, p. 6074, Jul. 2020.

[3] T. Akinci. "Determination of predictive relationships between problematic smartphone use, self-regulation, academic procrastination and academic stress through modelling," *International Journal of Progressive Education*, vol. 17, no. 1, pp. 35-53, 2020.

[4] H. Akoglu. "User's guide to correlation coefficients," *Turkish Journal of Emergency Medicine*, vol. 18, no. 3, pp. 91-93, Sep 2018.

[5] A. Akram, C. Fu, Y. Li, M. Y. Javed, R. Lin, Y. Jiang, and Y. Tang. "Predicting students' academic procrastination in blended learning course using homework submission data," *IEEE Access*, vol. 7, pp. 102487-102498, Jul. 2019.

[6] D. D Atsa'am and R. Wario. "Classifier selection for the prediction of dominant transmission mode of coronavirus within localities: Predicting COVID-19 transmission mode," *International Journal of e-Health and Medical Communications (IJEHMC)*, vol. 12, no. 6, pp. 1-12, Nov. 2021.

[7] U. M. Azeiteiro, P. Bacelar-Nicolau, F. J. Caetano, and S. Caeiro. "Education for sustainable development through e-learning in higher education: Experiences from Portugal," *Journal of Cleaner Production*, vol. 106, pp. 308-319, 2015.

[8] E. Buraimoh, R. Ajoodha, and K. Padayachee, "Prediction of student success using student engagement with learning management system," *Interdisciplinary Research in Technology and Management CRC Press*, pp. 577-583, Sep. 2021.

[9] L. M. Closson, and R.R Boutilier, "Perfectionism, academic engagement, and procrastination among undergraduates: The moderating role of honors student status," *Learning and Individual Differences*, vol. 57, pp. 157-162, Jul .2017.

[10] A. Dutt, M.A. Ismail, and T. Herawan, "A systematic review on educational data mining," *IEEE Access*, vol. 5, pp. 15991-16005, Jan. 2017.

[11] C. Goutte and E. Gaussier. "A probabilistic interpretation of precision, recall and F-score, with implication for evaluation," in *Proc. European Conference on Information Retrieval*, Springer, Berlin pp. 345-359, Mar. 2005.

[12] M. A. Hernández and S. J. Stolfo. "Real-world data is dirty: Data cleansing and the merge/purge problem," *Data Mining and Knowledge Discovery*, vol. 2, no. 1, pp. 9-37, Jan. 1998.

[13] D. Hooshyar, M. Pedaste, and Y. Yang. "Mining educational data to predict students' performance through procrastination behavior," *Entropy*, vol. 22, no. 1, p. 12, Dec 2019.

[14] A. J. Howell, and D. C. Watson, "Procrastination: Associations with achievement goal orientation and learning strategies," *Personality and Individual Differences*, vol. 43, no. 1, pp. 167-178, Jul 2007.

[15] SITE GOOGLE. [Online]. Available: https://sites.google.com/site/assistmentsdata/home/2009-2010-assistment-data#h.qrz90yfezr4h

[16] G. J. Hwang, "Definition, framework and research issues of smart learning environments-a context-aware ubiquitous learning perspective," *Smart Learning Environments*, vol. 1, no. 1, pp. 1-14, Dec. 2014.

[17] I. Jivet, J. Wong, M. Scheffel, M. Valle Torre, M. Specht, and H. Drachsler. "Quantum of Choice: How learners' feedback monitoring decisions, goals and self-regulated learning skills are related," in *Proc. LAK21: 11th International Learning Analytics and Knowledge Conference*, pp. 416-427, April 2021.

[18] X. Kang and W. Zhang. "An experimental case study on forum-based online teaching to improve student engagement and motivation in higher education," *Interactive Learning Environments*, pp. 1-12, Sep. 2020.

[19] N. Kant, K. D. Prasad, and K. Anjali. "Selecting an appropriate learning management system in open and distance learning: A strategic approach," *Asian Association of Open Universities Journal*, Mar. 2021.

[20] I. Katz, K. Eilot, and N. Nevo. "I'll do it later: Type of motivation, self-efficacy and homework procrastination," *Motivation and Emotion*, vol. 38, no. 1, pp. 111-119, Feb 2014.

[21] P. Kaur, M. Singh, and G. S. Josan. "Classification and prediction-based data mining algorithms to predict slow learners in education sector," *Procedia Computer Science*, vol. 57, pp. 500-508, Jan. 2015.

[22] R. Liu and A. Tan. "Towards interpretable automated machine learning for STEM career prediction," *Journal of Educational Data Mining*, vol. 12, no. 2, pp. 19–32, Aug. 2020.

[23] V. Mandalapu. "Profiling and modeling student learning behaviors and outcomes from digital learning environments," Doctoral dissertation, University of Maryland, Baltimore County, 2021.

[24] V. Mandalapu and J. Gong. Towards better affect detectors: Detecting changes rather than states," in *Proc. International Conference on Artificial Intelligence in Education*, Springer, Cham, pp. 199-203, Jun. 2018.

[25] M. Margaretha, S. Saragih, A. Mariana, and K. M. Simatupang. "Academic procrastination and cyberloafing behavior: A case study of students in Indonesia," *Cypriot Journal of Educational Sciences*, vol. 17, no. 3, pp. 752-764, Mar. 2022.

[26] T. Patikorn, R. S. Baker, and N. T. Heffernan. "ASSISTments longitudinal data mining competition special issue: A preface," *Journal of Educational Data Mining*, vol. 12, no. 2, pp. i-xi. Aug. 2020.

[27] P. Redmond. "From face-to-face teaching to online teaching: Pedagogical transitions," in *Proc. ASCILITE 2011: 28th Annual CONFERENCE of the Australasian Society for Computers in Learning in Tertiary Education: Changing Demands, Changing Directions*, Australasian Society for Computers in Learning in Tertiary Education (ASCILITE, 2011), pp. 1050-1060.

[28] N. S. Raj and V G. Renumol. "Early prediction of student engagement in virtual learning environments using machine learning techniques," *E-Learning and Digital Media.*, Jun. 2022.

[29] C. 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, Jul. 2010.

[30] M. T. Rupinski and W. P. Dunlap. "Approximating pearson product-moment correlations from Kendall's tau and Spearman's rho," *Educational and psychological measurement*, vol. 56, no. 3, pp. 419-429, Jun. 1996.

[31] S. F. Sabbeh. "Machine-learning techniques for customer retention: A comparative study," *International Journal of Advanced Computer Science and Applications*, vol. 9, no. 2, 2018.

[32] E. Slover and J. Mandernach. "Beyond online versus face-to-face comparisons: The interaction of student age and mode of instruction on academic achievement," *Journal of Educators Online*, vol. 15, no. 1, Jan. 2018.

[33] A. Tella, A. L. Balogun, N. Adebisi, and S. Abdullah. "Spatial assessment of PM10 hotspots using random forest, K-nearest neighbour and naïve bayes," *Atmospheric Pollution Research*, vol. 12, no. 10, p. 101202, Oct .2021.

[34] T. Qin, P. Poovendran, and S. BalaMurugan. "Student-centered learning environments based on multimedia big data analytics," *Arabian Journal for Science and Engineering*, pp. 1-11, Aug. 2021.

Copyright © 2023 by the authors. This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited (CC BY 4.0).

Nisha S. Raj is currently a research scholar in the Division of Information Technology, School of Engineering, Cochin University of Science and Technology (CUSAT), India. She secured her bachelor of technology (computer science and engineering) degree from CUSAT and a masters in computer science and engineering from Anna University, India.

Her research interests are educational data mining, learning analytics, personalized learning environments, AI in education, and technology in education. She has several publications in these areas. She has delivered talks on innovative teaching practices in various workshops.

Renumol V. G. is a professor in information technology at Cochin University of Science and Technology (CUSAT), India. She received the B.Tech. degree in computer engineering and the M.Tech. degree in software engineering from CUSAT, and a PhD degree in computing education from the Indian Institute of Technology (IIT), Madras, India. She has received a postdoctoral fellowship (PDF) in educational technology from the Indian Institute of Technology, Mumbai, India.

Her research interests include computing education, cognitive psychology, educational technology, and ICT in special education. She has published several studies in these areas.