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 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]. 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.
  \[
  \text{Accuracy} = \frac{(\text{TP} + \text{TN})}{(\text{TP} + \text{FP} + \text{TN} + \text{FN})}
  \]

- Precision: a measure of a classifier’s accuracy
  \[
  \text{Precision} = \frac{\text{TP}}{(\text{FP} + \text{TP})}
  \]

- Recall: Measure classifier completeness.
  \[
  \text{Recall} = \frac{\text{TP}}{(\text{FP} + \text{FN})}
  \]

- F1 score: F1 score conveys the balance between precision and recall.
  \[
  \text{F1 score} = 2 \ast \frac{(\text{precision} \ast \text{recall})}{(\text{precision} + \text{recall})}
  \]

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). 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.](image)

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. Table I 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>
<thead>
<tr>
<th>Feature</th>
<th>Weight</th>
<th>Rule</th>
</tr>
</thead>
<tbody>
<tr>
<td>AveKnow</td>
<td>3</td>
<td>&lt;= 0.45 then procrastination, &gt;0.45 then no procrastination</td>
</tr>
<tr>
<td>AveCarelessness</td>
<td>5</td>
<td>&gt;= 0.36 then procrastination, &lt;0.36 then no procrastination</td>
</tr>
<tr>
<td>AveCorrect</td>
<td>2</td>
<td>&lt;= 0.58 then procrastination, &gt;0.58 then no procrastination</td>
</tr>
<tr>
<td>NumActions</td>
<td>4</td>
<td>&lt;=1500 then procrastination, &gt;1500 then no procrastination</td>
</tr>
<tr>
<td>AveResBored</td>
<td>4</td>
<td>&gt;=0.25 then procrastination, &lt;0.25 then no procrastination</td>
</tr>
</tbody>
</table>

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_procrastinate value.

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.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Weight</th>
<th>Rule</th>
</tr>
</thead>
<tbody>
<tr>
<td>AveResOfftask</td>
<td>5</td>
<td>&gt;=0.55 then procrastination, &lt;0.55 then no procrastination</td>
</tr>
<tr>
<td>timeTaken</td>
<td>5</td>
<td>&gt;= 25 seconds then procrastination, &lt;25 seconds then no procrastination</td>
</tr>
<tr>
<td>hintCount</td>
<td>4</td>
<td>&gt;5 then procrastination</td>
</tr>
<tr>
<td>original</td>
<td>3</td>
<td>if value is 0 then procrastination</td>
</tr>
<tr>
<td>attemptCount</td>
<td>4</td>
<td>&lt;= 50 then procrastination &gt;50 no procrastination</td>
</tr>
<tr>
<td>totalFrAttempted</td>
<td>5</td>
<td>=0 then procrastination, no procrastination</td>
</tr>
<tr>
<td>consecutiveErrorsInRow</td>
<td>4</td>
<td>=0 then no procrastinate, otherwise procrastinate</td>
</tr>
<tr>
<td>sumTimePerSkill</td>
<td>4</td>
<td>&lt;602 then procrastination, &gt;= 602 no procrastination</td>
</tr>
<tr>
<td>RES_CONCENTRATING</td>
<td>3</td>
<td>&lt;0.65 then procrastination, &gt;=0.65 no procrastination</td>
</tr>
<tr>
<td>RES_OFFTASK</td>
<td>3</td>
<td>&gt;=0.17 then procrastination, else no procrastination</td>
</tr>
</tbody>
</table>
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, naïve 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.

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

<table>
<thead>
<tr>
<th>Features</th>
<th>Correlation Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>AveKnow</td>
<td>0.070230374</td>
</tr>
<tr>
<td>AveCarelessness</td>
<td>0.065049633</td>
</tr>
<tr>
<td>AveCorrect</td>
<td>0.136910239</td>
</tr>
<tr>
<td>AveResBored</td>
<td>0.1578585</td>
</tr>
<tr>
<td>AveResConf</td>
<td>0.203335259</td>
</tr>
<tr>
<td>AveResOffTask</td>
<td>0.3249286</td>
</tr>
<tr>
<td>action_num</td>
<td>0.116029646</td>
</tr>
<tr>
<td>timeTaken</td>
<td>0.354265338</td>
</tr>
<tr>
<td>consecutiveCorrectRow</td>
<td>0.138577777</td>
</tr>
<tr>
<td>consecutiveResBored</td>
<td>0.35412744</td>
</tr>
<tr>
<td>timeOver80</td>
<td>0.27831743</td>
</tr>
<tr>
<td>confidence(BORED)</td>
<td>0.3582190</td>
</tr>
<tr>
<td>confidence(CONFUSED)</td>
<td>0.16222356</td>
</tr>
<tr>
<td>confidence(OPT TASK)</td>
<td>0.34521834</td>
</tr>
<tr>
<td>RES_BORED</td>
<td>0.358275926</td>
</tr>
<tr>
<td>RES_CONFUSED</td>
<td>0.1622138</td>
</tr>
</tbody>
</table>
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.

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.

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.
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].

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.
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.

CONFICT 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


Procrastination: Associations with...