Ensemble Machine Learning Model for University Students' Risk Prediction and Assessment of Cognitive Learning Outcomes

Ananthi Claral Mary.T and Arul Leena Rose.P. J*

***Abstract*—One of the biggest challenges in higher educational institutions is to avoid students' failures. Globally student dropout is a serious issue. Risk of dropouts can be identified at an earlier stage using machine learning classifiers, as they have gained more popularity in both academia and industry. The research team suggests that early prediction facilitates educators and higher education administrators to take necessary measures to prevent dropouts. Data for the research were collected from 530 Indian students when they were engaged in online learning during pandemic crisis. This research work involves two phases. In first phase, hybrid ensemble strategy is focused that integrates two powerful machine learning algorithms namely Random Forest (RF) and eXtreme Gradient Boosting (XGBoost) for early at-risk prediction. The result is a fast procedure for classification of at-risk students which is competitive in accuracy and highly robust. Prediction models are developed using ensemble learning, furthermore ensemble models are combined into a single meta-model, which provides best outcomes to enable higher education institutions for predictive analysis. Moreover, it correctly classified students' at-risk regarding accuracy, precision, recall and F1-score with values of 93%, 91.52%, 96.42% and 93.91% respectively. In second phase, prediction model is deployed by creating a web application using. Net framework to sense students' sentiments using Azure cognitive services text analytics (Application Programming Interface) API for detecting cognitive behavioral outcomes in online learning environment.**

***Index Terms*—At-risk, cognitive services, ensemble, machine learning, online learning environment**

I. Introduction

The world is undergoing an unprecedented migration in altering of life forms, due to advancement of Information and Communication Technologies (ICT). With emerging requirement of latest tools and technologies, there is a pressing need for education sectors to create attractive Learning Management System that satisfies students' needs through e-learning, distance education, mobile learning, blended learning, online tutoring and develop algorithms which fit into changing system intelligence. In this modern era, these learning systems are becoming popular among educators and learners as it is a flexible and interactive education system, despite of time, distance, and attendance problems. However, these systems face challenges of increased dropout rates before course work completion. Students' attrition is growing in online learning environment.

Due to various potential factors that stimulate the dropouts namely demographic characteristics, previous academic achievements, device usage, self-efficacy, technical experience in online learning platforms, readiness, and effectiveness of participation.

Students' at-risk in higher education programs, can be predicted using various machine learning algorithms as they are good at processing and analyzing data from huge datasets for predicting future behavior, trends and patterns making pro-active, fast decision-making process. Machine learning is widely applied in various domains as they provide valuable outcomes. Higher education is vital for students as they learn and equip themselves with advanced technical skills for societal contribution. As, huge number of students graduate from universities every year world-wide, many students fail in their course work examinations and are forced to retake them for completing their graduation. In worst case condition, they withdraw from universities which significantly cause serious impact on valuable faculties efforts. Hence, Higher Educational Institutions (HEI) needs to construct a predictive model for identifying at-risk students and analyze their behavior which truly reflects the students' situation serving as a caution for both educators and administrators.

II. Literature Review

Karahoca and Zaripova *et al.* [1] aimed to investigate engineering university students' perceptions regarding distance learning during pandemic. They emphasized that students were able to access their courses using smart devices and satisfied with distant learning as they have not fallen behind in their studies. Badal and Sungkur [2] highlighted that pandemic outbreak created major disturbance in all areas. To confront this problem the learning procedure and content delivery mode must be altered. Most courses were offered through online learning platforms. Machine learning techniques were used to predict students' performance and to analyze effect of online learning platforms. Random Forest (RF) outperformed other machine learning classifiers namely Logistic Regression (LR), (K-Nearest Neighbors) KNN, Naive Bayes (NB), Decision Tree (DT), (Support Vector Machine) SVM, Deep Learning. 85% accuracy for student grade and 83% for engagement prediction was recorded with respect to students' demographics and learning platform interaction. This enables the educators to identify students' at-risk and take necessary corrective measures. Jawthari and Stoffa [3] compared three predictive models RF, NB, and LR to identify at-risk students who would fail the courses. Weekly predictions were made using (Virtual Learning

Manuscript received December 6, 2022; revised December 16, 2022; accepted January 28, 2023.

Ananthi Claral Mary.T and Arul Leena Rose. P. J are with the Department of Computer Science, College of Science and Humanities, SRM Institute of Science and Technology, Chengalpattu, India.

*Correspondence: leena.rose527@gmail.com (A.L.R.P.J.)

doi: 10.18178/ijiet.2023.13.6.1891

Environment) VLE data and assessment grade as first technique, and weekly assessment grade accumulation as the second approach. The findings showed that Random Forest has provided improved results compared to other algorithms.

The dominant factor that affects mental health of college students is academic stress. Few student groups may experience stress at a higher rate than others, and COVID-19 epidemic has made the stress response even more challenging. According to the survey from college students' researchers examined whether academic stress levels affected their mental health. The outcomes indicated that academic stress is significantly correlated to psychological well-being among learners stated by Barbayannis and Bandari *et al.* [4]. Student dropout is a major issue as it affects individual student also the former school in which the student studied, parents and society. Recently, a hot topic of research is students' dropout prediction. Authors proposed a novel stacking ensemble depending on a hybrid of RF, XGBoost, Gradient Boosting (GB), and Feed-forward Neural Networks (FNN) to predict student's dropout in universities. When compared with baseline models, the proposed system has shown better performance in terms of training accuracy with 93.59% and testing accuracy of 92.18% [5]. Study proposed, trained models' performance is enhanced comparing with single classifiers by using ensemble methods to predict students' performance. Higher predictive accuracy was provided by ensemble methods. Experimental results revealed that boosting gained highest accuracy for all trained features. Thus, ensemble methods performed better than single classifiers [6]. Principal Component Analysis (PCA) was used in conjunction with four machine learning classifiers namely RF, C5.0, NB, SVM to enhance the classification performance. The proposed hybrid model with 10-fold cross validation produced satisfying outcomes [7]. Authors have built a hybrid model using Generalized Linear Model that incorporates an Artificial Neural Network for predicting grades of students imposed in distance learning during pandemic. This hybrid model depends on 35 variables and is strongly correlated to students' academic performance yielding R^2 as 1 [8].

Recently dropout prediction has received much attention. Higher dropout rate is an unsolved problem in MOOCs (Massive Open Online Courses). A novel hybrid algorithm is proposed by combining DT and Extreme Learning Machine (DT-ELM). Decision tree is applied for feature selection, then it defines enhanced weights of features selected to boost classification capability. Experimental results reveal that DT-ELM is 12.78% higher in accuracy than baseline models [9]. Any educational institution with a high dropout rate runs the risk of endangering both learners and educators. The goal of this study is to examine the problem and attempt to offer a better prediction of dropout ratio. This can assist enterprises in estimating resource usage, supplying online content, and especially allocating the appropriate number of seats to qualified students. As a result, hybrid machine learning model used in this study, gave the chosen MOOC dataset a satisfactory and enhanced accuracy ratio. The random forest model was chosen as the best model with maximum accuracy record with 90% for predicting dropout ratio when compared to the other investigations [10]. MOOC allows students to learn at their own pace. This flexibility makes students to drop out of class. Therefore, more efficient dropout prediction model is necessary for MOOC development. The goal of researchers was to predict if a learner is about to dropout in next 10 days when clickstream data for first 30 days was given. They propose a hybrid neural network model to predict student's dropout behavior. A new (Convolution Neural Network) CNN and Squeeze-and-Excitation Networks (SE-Net) and Gated Recurrent Unit (GRU) network improves prediction performance with 94% accuracy [11]. Since the advent of innovative technologies in online learning, high dropout rates have been a source of concern for MOOC providers and educators. This prompted researchers to concentrate on learner motivation studies from several angles, including the identification of demotivational signals, personalization of learning paths, and course suggestions. To choose best MOOC for each learner, authors attempted to forecast the motivation of learners for MOOCs by comparing four machine learning algorithms which shows Random Forest to be best technique with 95% accuracy [12].

Researchers proposed Stacked Generalization for Failure Prediction (SGFP) which integrates three ensemble learning classifiers, Light Gradient Boosting Machine (LGBM), XGB, and RF, using a Multilayer Perceptron (MLP). SGFP algorithm depends on heterogeneous data which indicates students' levels of interaction, performance, and training skills. The accuracy of SGFP is 97.3% [13]. The objective of researchers is to evaluate feature importance in predicting college students' grants and construct an ensemble model that balance accuracy in prediction. Results confirm that Gradient Boosting Decision Tree established on ensemble learning depicts improved classification performance with 95% accuracy than non-integrated methods such as SVM, KNN. They have highlighted that by using ensemble learning accuracy can be significantly increased [14].

At present, student dropout is one of the challenging issues of educational institutions. Various classification algorithms were used for predicting student dropout. The outcomes of Naive Bayes, Decision Tree and KNN were 64.29%, 64.84%, and 75.27% respectively. When these algorithms were combined with ensemble classifiers using (Gradient Boosting Machine) GBM as meta classifier high accuracy of 79.12% is achieved, with 10-fold cross-validation the best accuracy is 98.82% [15]. When the relevant authority identifies, when a student is absent for numerous days then it is declared that the student has dropped out of school. Even though when several efforts are taken to improve overall education status at secondary level, dropout still exists. Study used both student and school dataset to generate an ensemble model for predicting student dropout in secondary schools. Deployed model was created by integrating tuned LR and MLP. The model's purpose was to aid education stakeholders in detecting at-risk students and schools. This developed system can be used by authorities for planning and budgeting to satisfy school requirements [16]. Combining models always achieve best accuracy than individual models. To find the most appropriate model to predict a particular problem is not an easy process. Authors proposed three various models for predicting school dropout. They depend on ensemble regression models which were used in Brazilian Higher Education institutions, that detects the major factors associated with dropout. The proposed model obtained better

performance and can be used as an accurate tool for student performance prediction. This facilitates the educational administrators and policy makers for development of new student retention policies [17].

Technological advancements like Blockchain, IoT, Cloud Computing and Big Data have widened applications of Sentiment Analysis (SA) permitting to be utilized in any discipline [18]. HEI is increasingly looking for best ways to understand learning experience of their students. Sentiment analysis helps to investigate emotions and attitudes of students regarding their course experience. To analyze sentiment text was fed into Google's Cloud-based Natural Language Processing. Results presented that students' sentiment in online interaction during two online courses is more positive than in face-to-face courses [19]. During COVID-19 pandemic, to support remote and distance learning useful learning resources are lecture recordings. Students with illness, learning disabilities, and work commitments have narrated that availability of lecture recordings has shaped an inclusive education setting. Sentiment analysis was conducted using Microsoft Azure cognitive services text analytics API. With a large text dataset, machine learning was employed that were labeled for sentiments along 0 and 1. Findings depicted that lecture recordings serve as an additional resource for preparing notes or exams [20].

Researchers analyzed sentiments from tweet database using deep learning techniques. Tweets were gathered in the form of pleasant and unpleasant with high and low confidence scores. They have depicted various emotions of the people in different schedule period. The information is investigated with the assistance of API to classify outcomes as positive, negative, or neutral. For understanding the mentality of people sentiment analysis can be used [21]. In teaching-learning process VLE delivers a set of communication and interaction tools utilized by learners and educators. Researchers presented SentiEduc framework that use Multi-Agent System (MAS) to gather and analyze opinion of texts posted by learners in VLE. SenticNet tool was used to analyze sentiments automatically [22]. From an organization, probable text reviews collected on 270 training programmes by 2688 participants were analyzed. RapidMiner Text Mining package was used to track tokenization, removal of stop word, stemming, and token filtering. Authors suggested that instead of content delivery and faculty expertise, the proposed approach can further be expanded to establish sentiment expressed over several aspects like internet connection, hospitality [23].

Researchers have developed, a web-application system that uses sentiment analysis and text analytics to deliver teachers a deeper analysis of learners' response to assess the course they have taught that will enhance the students learning experience. Feedbacks were grouped into positive, negative, and neutral. The results depicted larger number of neutral sentiments. Their implementation was successful, and significantly benefits students, lecturers, and administrators [24]. Twitter is the popular free social networking channel. In Anadolu University open and distance education system, sentiment analysis for learners was performed by fetching tweets. 400 tweets were used for validation and classification outcomes were presented. The negative feelings and student complaints can be concentrated by institution managers [25]. Therefore, to solve real world problems through design and development of smart learning environment, Azure cognitive services, text analytics API is used that leverages natural language processing capabilities for deploying high-quality AI models [26].

Hence, students' performance prediction and identifying their opinions are vital to improve the quality of online education. Many applications are available to predict students at risk, but all are not useful due to their strength and weakness. Past researchers have focused students' dropout depending on certain factors namely demographics and learning platform interaction. They emphasized the use of ensemble methods and they have revealed that bagging and boosting performs well in developing ensemble models. Ensemble learning improves accuracy by fusing predictions from many learners. In this context, an ensemble meta-model is proposed that combines predictions from well-performing models to enhance students' classification performance producing a final predictive model. The surveyed literatures throws light on Random Forest algorithm, yet researchers have stated that during classification problems, traditional Random Forest may exhibit poor performance due to little attention on unstable samples. Moreover, they have shown that hybrid model reveals enhanced performance compared to traditional machine learning classifiers.

Stacking fusion model based on RF-CART–XGBoost–LightGBM is proposed, which plays a dominant role in students' performance prediction depending on their learning behavior. Prediction accuracy of RF is 71%, (Classification and Regression Tree) CART 68%, XGBoost is 73%, and LightGBM 76%. Stacking fusion model performs better than single models with 84% accuracy [27]. Compared to previous existing studies proposed by researchers, this research aims to merge Random Forest (88.61%) and eXtreme Gradient Boosting (92%) to produce an ensemble meta-model with higher accuracy of 93% for students' at-risk prediction in online learning environment using real-time dataset.

III. Materials and Methodology

A. Problem Statement

The primary aim of this research is to detect at-risk students and analyze their sentiments in online learning environment. Within this framework, various multideterminant characteristics were investigated in this study. Successful online communities can be achieved with existence of these influential factors, as they can determine students' at-risk and their emotional state. They are the main predictors of learners' performance in cloud-based learning system. This paper addresses the subsequent research questions:

1) To develop an ensemble model for risk prediction of students in online learning environment.
2) To evaluate the proposed models' performance using indicators namely accuracy, precision, recall, F1-score, and Area under the Curve.
3) To identify students' sentiment for smart device possession characteristics, self-efficacy, technical experience in cloud platforms, level of readiness, and

effectiveness of online classes compared to traditional settings.

B. Data Collection from Participants

Real data was collected from students of various higher education institutions throughout India. In this cross-sectional study, participants were selected through convenience sampling. The research was carried out through an online survey by using free cloud service Google Forms. Questionnaire was designed and circulated to target respondents belonging to diverse academic divisions and levels. Few sections of questionnaire were developed exclusively to address the research objectives. Link of the questionnaire was circulated to students through online, when they were engaged in online classes via cloud meeting platforms namely Google Classroom, Zoom, Microsoft Teams, Webex to identify students' at-risk and analyze their sentiments based on their interactions with online learning platforms. Approximately, 700 students submitted the completed questionnaire. As the dataset contains redundant data and outliers, after careful investigation 530 responses were chosen.

1) Students' personal information/demographic data

Among the total sample of participants, 66.60% were males, while that of 33.40% were females. 29.81% of students were aged between 17 to 19 years old, 69.81% were between 20 to 29 and 0.38% were 30 to 45. Considering academic qualification of students, they were from various colleges pursuing different engineering courses like mechanical, computer science, electronics & communication, electronics & electrical, biomedical, and various disciplines of science & humanities namely, computer science, physics, chemistry, mathematics, commerce, business administration. 68.68% of respondents held bachelor's degrees, 31.32% held master's degrees.

2) Survey instrument and structure

Online survey questionnaire developed for this research was framed with five major sections: 1) Demographic information of students; 2) Device possession characteristics to access online classes; 3) Self-efficacy of participants in online learning, 10 questions were framed with 5-point Likert scale; 4) Familiarity of technological experience with cloud platforms, 10 questions were designed which was estimated with 4-point Likert items; 5) Readiness and effectiveness presents 10 questions about readiness of respondents in participating online learning with 4-point Likert scale, and effectiveness represents five questions with 5-point Likert scale to measure effectiveness of online learning compared to conventional classroom settings.

IV. Machine Learning Model Implementation

Many machine learning techniques are effectively used in educational settings. In this research, real data of students in engineering and science & humanities based on their interactions with online learning platforms were collected. Dataset consists of 56 features with duplicate data and outliers. To maintain quality of prediction, redundant records and outliers were eliminated by preprocessing the data. After preprocessing, two consistent measures of RF algorithm namely %IncMSE and IncNodePurity were used to extract optimal features. 46 features were selected for each student with target variable as (Cumulative Grade Point Average) CGPA. Features extracted from the model provide outstanding directions that instructors can apply to provide early suggestions to students prior to course work completion. For improving model accuracy and to adjust the feature scale, data is normalized using Min-Max normalization. Hybrid machine learning algorithms: RF and XGBoost are proposed. These algorithms were chosen as they provide best results for prediction. The ensemble meta-model was employed for prediction of students' at-risk in online learning. The main idea behind this is to boost the prediction performance of classifiers using ensemble methods that depend on decision tree algorithms. In next phase, the prediction model is deployed using Azure cognitive services to identify students' sentiments. The text analytics API automatically classifies text responses as positive, negative, or neutral depending on students' comments. This enables the instructors to identify their cognitive behavioral outcomes. Flow of the research model is shown in Fig. 1.

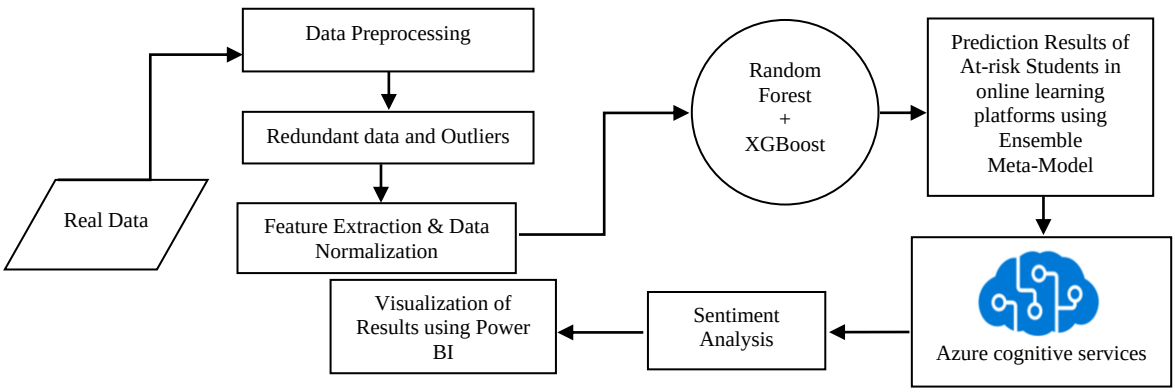

Fig. 1. Proposed diagram of the research model.

A. Random Forest

This is one of the ensemble learning methods which is superior to other methods owing to its simple structure, understandability, higher efficiency [28]. This is a useful tool for tasks involving predictive analytics in institutional research. Random forest is flexible to apply, adaptable, and cost-effective to compute, with the decision-tree infrastructure offering a comprehensible alternative to conventional regression techniques. To rank feature

importance successfully random forest algorithm was used. Random Forest outperformed other baseline models with highest predictive accuracy [29]. In terms of accuracy and computational speed through its combination of bagging and random subspace methods, RF persists to be top among ensemble classifiers [30]. This is a process of integrating various classifiers for solving complex problems by increasing model performance. RF builds several decision trees on randomly selected data samples and takes majority vote for classification and average for regression. Large number of trees in forest leads to greater accuracy and avoids overfitting. Random Forest algorithm can be represented as:

1) Select random k data points from training set
2) With selected data point subsets build decision trees
3) Choose the number N for building decision trees
4) Repeat steps 1 & 2
5) When a new datapoint enters, find prediction of each decision tree, and allocate new datapoint to category which wins majority of votes.

Random Forest Classifier is imported for creating the prediction model with train and test sets to forecast the students' at-risk in online learning environment by setting n_estimators=200, max_depth=6. Random Forest achieved 88.61% accuracy, 56.67% precision, 33.33% recall and 41.86% of F1 score.

B. Extreme Gradient Boosting Algorithm

XGBoost method achieves best results when compared to classical machine learning models for predicting student performance [31]. XGBoost is a tree-based supervised machine learning algorithm which is highly scalable and exact execution of gradient boosting that increases computing power, being designed largely for improving model performance and computational speed. When compared to other algorithms XGBoost is a faster algorithm as of its parallel and distributed computing. In terms of systems optimization and machine learning principles, this is built with deep considerations. In XGBoost, decision trees are formed in sequential form and weights perform a significant role. Weights are allocated to all independent variables that are then fed into DT which predicts outcomes. Weight of wrongly predicted variables by the tree are boosted and variables are then fed to second DT. These individual classifiers then ensemble to provide a robust and more accurate model. Like gradient boosting, this generates an additive expansion of objective function by minimizing loss function. To reduce step size in additive expansion, shrinkage is an additional regularization parameter in XGBoost. Using other approaches like depth of trees, tree complexity can be reduced. Another advantage of tree complexity reduction is models are trained faster and requires less memory space. To avoid overfitting and to increase training speed randomization techniques are implemented. Mathematically the model is denoted in the form represented by Eq. (1).

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), f_k \in F \quad (1)$$

where K is number of trees, f is functional space of F; F is set of possible CARTS. Objective function for model is represented in Eq. (2).

$$obj(\phi) = \sum_i^n l\left(y_i, \hat{y}_i\right) + \sum_{k=1}^{K} \Omega(f_k) \quad (2)$$

where, first term is loss function and second is regularization parameter. Optimization becomes harder when all trees are learned at once. Thus, an additive strategy is applied, which minimizes loss function and add a new tree that is summarized as given in Eq. (3).

$$\hat{y}_i^{(t)} = \sum_{k=1}^{t} f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i) \quad (3)$$

Taylor series expansion up to second order. Refer to Eq. (4).

$$obj^{(t)} = \sum_{i=1}^{n} \left[l\left(y_i, \hat{y}_i^{(t-1)}\right) + g_i f_t(x_i) + {}^1/_2\, h_i f_t^2(x_i) \right] + \Omega(f_t) + constant \quad (4)$$

The regularization term of the model is defined by Eq. (5).

$$\Omega(f) = YT + {}^1/_2\, \lambda \sum_{j=1}^{T} w_j^2 \quad (5)$$

The objective function becomes as presented in Eq. (6).

$$\sum_{j=1}^{T} \left[\left(\sum_{i \in I_j} g_i \right) w_j + {}^1/_2 \left(\sum_{i \in I_j} h_i + \lambda \right) w_j^2 \right] + YT \quad (6)$$

The simplified formula for XGBoost algorithm is given by Eq. (7).

$$obj^{(t)} = \sum_{j=1}^{T} \left[G_j w_j + {}^1/_2 \left(H_j + \lambda \right) w_j^2 \right] + YT \quad (7)$$

XGBoost uses hyperparameters which is a type of parameter, set before the learning process commences. It's tunable and directly affects the model performance. The most common parameters are:

- max_depth: Maximum depth per tree. Input value of max_depth is 6.
- learning_rate: This identifies the step size at each iteration while the model optimizes towards its objective. Value is 0.3
- n_estimators: Number of trees in the ensemble meta-model. This is equivalent to number of boosting rounds. Value is 100.
- colsample_bytree: This represents fraction of columns to be sampled randomly for each tree. Value is 1.
- subsample: Represents fraction of observations to be sampled for each tree. Value is 1.

To achieve maximum performance, hyperparameters are optimized by using random search that allows more precise discovery of best values by selecting random combinations to train the model. Thus, the output obtained from XGBoost algorithm recognized students' at-risk with accuracy of 92%, with a precision of 70%, recall of 38.88% and had the F1-score of 50%.

V. Experimental Results

This meticulous meta-modeling technique is crucial for decision makers and education administrators, as it identifies students' at-risk. The global implication of this research is its capability to provide an early warning to educators and this novel model facilitates the HEI to develop innovative strategies and teaching principles relevant to student retention. Proposed model is assessed on real-time data

gathered from the Indian students. They used various cloud platforms namely Google Classroom, ZOOM, Teams, etc to complete their course work examinations. HEI utilized these platforms for best teaching and content delivery. Importance of online learning data is highlighted in this research as it served as a best platform during the pandemic crisis. A hybrid model is created which is a combination of random forest and XGBoost that had highest accuracy percentage compared to other baseline models. When more data is used for training, the classification results are better. To avoid the problem of over fitting ten-fold cross-validation is applied. The various parameters used to evaluate are Accuracy, Precision, Recall, and F1-score. These indicators perform a dominant role by providing better foresight for prediction outcomes.

A. Model Evaluation Metrics

The primary research objective is to correctly predict the risk students in online learning environment, as it is dependent on multiple factors including demographics, device characteristics, self-efficacy, technical experience, level of readiness & effectiveness of online sessions. For evaluating the model performance, four evaluation metrics are considered. Equation uses True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN). To weight each instance equally micro-average is used, whereas a macro-average is used when all classes are considered equally to estimate overall classifier's performance about most frequent class labels.

1) Accuracy

Accuracy is equal to ratio of total correct predictions by classifier to total number of data points. For assessing performance of classification model, this is a vital measure, and is the frequently applied metric to evaluate classifier's quality. It is calculated as shown in Eq. (8).

$$Accuracy = \frac{TP+TN}{Total\ Number\ of\ Samples} \tag{8}$$

2) Precision

This is a vital measure to detect number of correctly classified students in the dataset. Precision is equal to ratio of True Positive samples to sum of True Positive and False Positive samples. Refer to Eq. (9).

$$Precision = \frac{TP}{TP+FP} \tag{9}$$

3) Recall

This is important measure to detect number of correctly classified students in dataset out of all students that have been accurately predicted. Refer to Eq. (10).

$$Recall = \frac{TP}{TP+FN} \tag{10}$$

4) F-Measure (F-Score)

This provides an accurate balance between Precision and Recall by giving an accurate evaluation of model's performance in classifying students' at-risk. It is equal to harmonic mean of precision and recall value. This is an important metric to estimate the model. Eq. (11) is calculated, and ensemble models' classification report is depicted in Fig. 2.

$$F1\ Score = \frac{2*Precision*Recall}{Precision+Recall} \tag{11}$$

```
print(classification_report(y_test, y_pred))

              precision    recall  f1-score

           0       0.93      0.97      0.95
           1       0.96      0.92      0.94

    accuracy                           0.94
   macro avg       0.94      0.94      0.94
weighted avg       0.94      0.94      0.94
```

Fig. 2. Classification report of ensemble model.

B. Performance Comparison of Machine Learning Models

After the implementation of machine learning model, confusion matrix is verified for each model to determine performance of the classification models for a given test data. Ensemble meta-model obtained an accuracy of 93%, recall results on test set shows that model can predict nearly 96% of at-risk students. Precision value of 91.52% means predicting both students' at-risk and non-atrisk were accurate after the delivery of course content. The outcomes depict that this model has capability of processing effectively every class and ultimately, this is more optimum for predicting students' at-risk in higher education when compared to other classifiers (Table I). Regarding F1-score, the method reached overall F1-Score of 93.91% which reveals efficient results. Fig. 3 shows better prediction performance of ensemble method. To yield better performance on machine learning problems ensemble learning techniques were used. The final risk prediction model is derived by combining results from random forest and extreme gradient boosting to produce a strong learner. Outcomes of the proposed technique will be beneficial for HEI within online learning framework to reduce failure rates and enhance the learning performance.

TABLE I: PERFORMANCE METRICS OF ENSEMBLE META-MODEL VERSUS BASE CLASSIFIERS

Model	Accuracy %	Precision %	Recall %	F1-Score %
Random Forest	88.61	56.67	33.33	41.86
Extreme Gradient Boosting	92	70.0	38.88	50.0
Ensemble Meta-Model	93	91.52	96.42	93.91

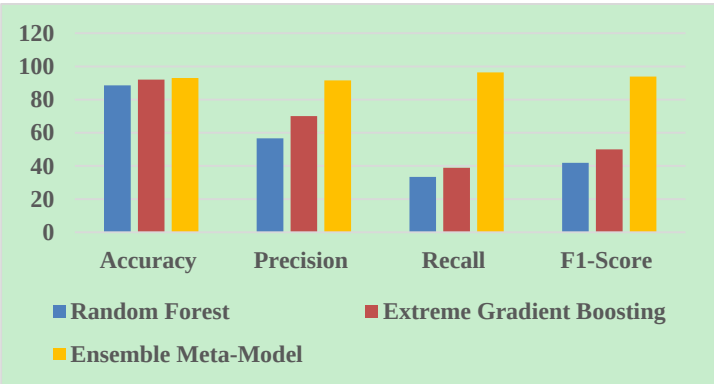

Fig. 3. Overall prediction results for students' at-risk.

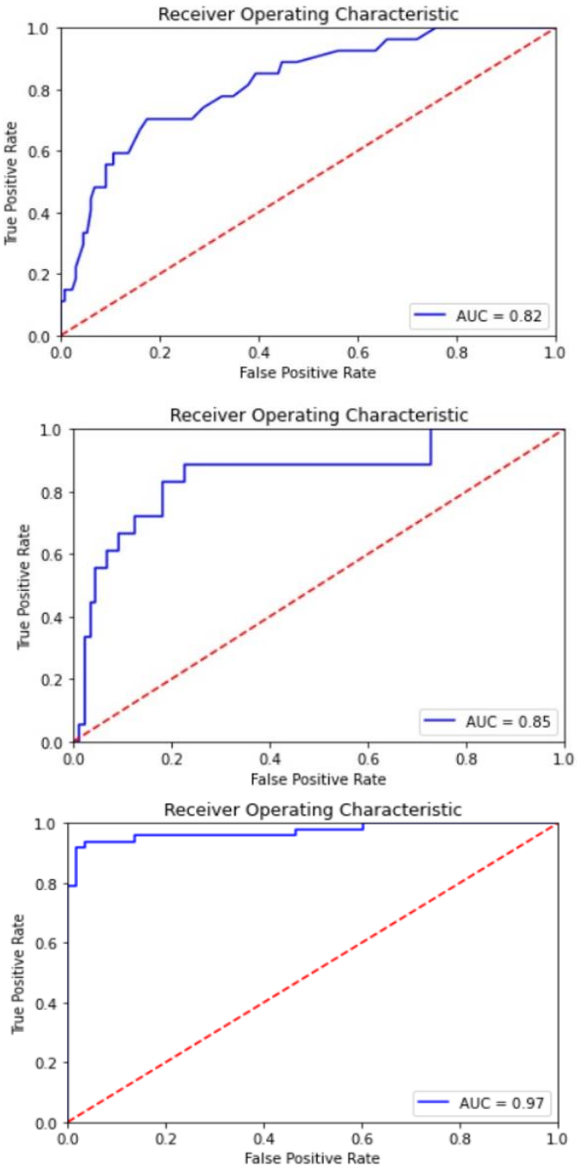

Fig. 4. ROC curve and AUC results of random forest, XGBoost, ensemble meta-model.

C. ROC and Area under the Curve (AUC)

The performance of classification models namely random forest, XGBoost and ensemble model at all classification thresholds can be depicted using Receiver Operating Curve (ROC). This curve plots two parameters namely True Positive Rate (TPR) and False Positive Rate (FPR) at various threshold settings. ROC curve is thus sensitivity or recall as a function of fall-out. Fig. 4 depicts ROC and Area Under Curve (AUC) results of random forest, eXtreme Gradient Boosting, ensemble meta-model.

VI. Detecting Students' Cognitive Learning Outcomes Using Sentiment Analysis

Azure cognitive services text analytics API is used in this research to identify sentiment scores of learners. This applies machine learning classification algorithm for sentiment analysis. The machine learning approach employed was trained with a huge text of students' real dataset of records. Sentiment analysis is a part of text analytics that identifies level of positive, negative, or neutral sentiment of input text using a confidence score found by the service at a sentence and document-level across a variety of languages. API returns sentiment score of a raw text between 0 and 1. Scores closer to 1 indicate positive, scores nearer to 0.5 represents neutral while scores closer to 0 represent negative sentiment. This provides natural language-based processing APIs to process raw text. To extract quality strings for a higher level of document or textual content summarization key phrase extraction API is used which extracts key phrases to identify key points in input text. Language detection API takes text documents as input and returns language identifiers. It enables the application to detect language in which user is submitting the text. The named entity recognition API takes a JSON document and returns a list of entities with links to more information on the web. Text Analytics API can be invoked as depicted in Fig. 5.

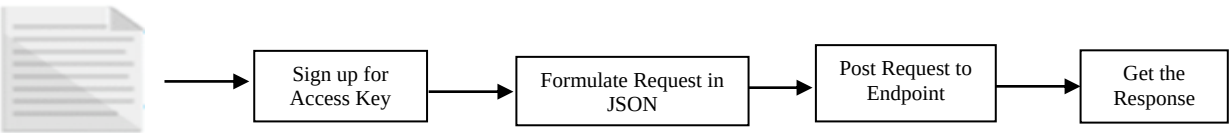

Fig. 5. Text analytics API invocation.

Key findings focused on students' perceptions and learning activities in an online learning environment. Their opinions had an impact of their professional development. Motivated by the need for more concrete and accurate classification model, five important factors were identified that may be used to better understand students' cognitive behavior and attitude towards online learning. Overall, the conceptual model was proposed and validated to evaluate students' sentiments from various degree programmes on device usage, self-efficacy, technological experience, readiness, and effectiveness in cloud-based online learning platforms. This investigation utilized 2650 records and sentiment results obtained depicts that 53.40% indicates positive, 38.11% being neutral and 8.49% being negative (see Table II).

TABLE II: Overall Percentage of Sentiment Against Five Factors for 530 Students

Sentiment Label	Frequency Count	Percentage (%)
Positive	1415	53.40%
Neutral	1010	38.11%
Negative	225	8.49%
Total	2650	100.00%

A. Smart Device Usage

Utilization of smart devices emphasizes the learners to study anywhere and anytime. They use smart devices namely smartphone, laptop, tablet, desktop PC for accessing digital resources through broadband, WIFI, mobile data and other modes. Majority of the responses show that 0.75% being

positive, 91.51% with neutral and 7.74% negative sentiments. Findings from open-ended responses revealed several themes explaining conditions in which students depict larger percentage of neutral sentiments while engaging with online classes through smart devices (see Table III). With the growth of pervasiveness towards smart devices, this depends on various technical factors like slow internet connection, network bandwidth breakdown, and poor response time. Moreover, Prakasha and Sangeetha *et al.* acknowledged that there is a necessity to examine the existing digital infrastructure of learners of Indian families. Government, non-governmental organizations, educational institutions should consider necessary measures to assist students' requirements during pandemic rather than recklessly insisting on remote learning. This information will provide valuable insights and ensures equity in access to higher education and best usage of ICT with purpose of nurturing their talents and life-long learning [32].

TABLE III: DISTRIBUTION OF SENTIMENTS FOR DEVICE USAGE

Sentiment Label	Frequency Count	Percentage (%)
Positive	4	0.75
Neutral	485	91.51
Negative	41	7.74
Total	530	100.00

B. Self-efficacy

Psychologist Albert Bandura has stated self-efficacy as a person's belief in his or her capacity for executing behaviors required to produce performance achievements. This is significant variable of success in participating online classes. The students' behavior and engagement activities facilitate them to improve their academic performance. The questionnaire framed represents self-efficacy of respondents in managing to solve difficult problems, finding means & ways to acquire what they desire, stick to aims & achieve their goals, handle unforeseen situations, invest necessary effort, finding several solutions and handle whatever comes their way while participating in online classes. Sentiment analysis results exhibited that 72.07% of the sentiments were positive, 1.51% neutral and 26.42% negative (see Table IV). These results indicated that respondents reveal greater percentage of positive sentiments towards online learning self-efficacy.

TABLE IV: DISTRIBUTION OF SENTIMENTS FOR SELF-EFFICACY

Sentiment Label	Frequency Count	Percentage (%)
Positive	382	72.07
Neutral	8	1.51
Negative	140	26.42
Total	530	100.00

C. Familiarity with Cloud Technologies

This highlights on technology students are familiar with accessing online teaching materials, as technology-based setting enables them to learn better. Technology utilization contributes a lot in pedagogical and practical aspects that leads to effective learning with the assistance and support from cloud components. Outside the classroom, all these technologies have provided many useful added advantages to educators and learners thereby, allowing for professional development, co-curricular learning, conference support, greater information sharing, collaboration, and ease of accessibility. These cloud platforms prepare students to make decisions independently, by promoting self-directed learning, creativity, and innovation. Results illustrates that most of the sentences labeled as neutral sentiments with 91.70% and 8.30% were negative (see Table V). However, this shows that most of the students did not have sufficient experience with these technologies in accessing online course contents. Hence participants require more technical expertise and training in online learning platforms. Therefore, HEI need to focus on identifying the appropriate platforms and curriculum design for online courses.

TABLE V: DISTRIBUTION OF SENTIMENTS FOR CLOUD PLATFORM USAGE

Sentiment Label	Frequency Count	Percentage (%)
Positive	0	0
Neutral	486	91.70
Negative	44	8.30
Total	530	100.00

D. Readiness of Participation in Online Learning

In addition, students' readiness of participation in online learning plays a significant role to encourage learners to be involved in learning activities. Readiness was one of the strongest predictors of satisfaction for students in online courses. This can be estimated by evaluating learners' familiarity with online learning, sufficient requirement of knowledge and skills, readiness with PCs and stable internet connection for attending online classes, efficiency of device they use for accessing online classes, motivation to utilize features of an online learning environment, time management, submission of assignments and complete courses through online program, readiness to foster a positive online learning environment with peers and teachers. Therefore, readiness can be perceived as a vital factor to be considered in any progress of online learning environments. Outcomes of sentiment analysis indicates that 94.15% were positive and 5.85% were neutral sentiments (see Table VI). The students reveal greater percentage of positive sentiments for readiness of participation.

TABLE VI: DISTRIBUTION OF SENTIMENTS FOR READINESS

Sentiment Label	Frequency Count	Percentage (%)
Positive	499	94.15
Neutral	31	5.85
Negative	0	0
Total	530	100.00

E. Effectiveness of Online Learning Compared to Traditional Classroom Environment

This is important to determine efficiency of online learning compared with traditional educational settings. Effectiveness helps to understand value and efficacy of a particular education system. It can be estimated by assessing students' expediency, satisfying their individual learning

requirements, contributing to efficient communication, strengthening the sense of community with educators and peers, supporting greater student participation and interaction. All these effectiveness factors shape the online learning systems more efficiently. According to students' perceptions, effectiveness of online learning reveals 100% positive sentiments (see Table VII). This shows online learning is more effective than traditional learning, when considering the above stated factors. Sentiment scores for an individual student, is depicted in Table VIII. Similarly, sentiments scores are calculated for entire dataset (*n*=530).

TABLE VII: DISTRIBUTION OF SENTIMENTS FOR EFFECTIVENESS

Sentiment Label	Frequency Count	Percentage (%)
Positive	530	100
Neutral	0	0
Negative	0	0
Total	530	100

TABLE VIII: EXAMPLES OF TEXT AND SENTIMENT SCORES FOR AN INDIVIDUAL STUDENT

Objectives	Text	Score	Sentiment
Device Usage	Type of smart device used, mode of device availability, type of device connectivity, number of hours device is connected to access online classes?	0.5	Neutral
Self-Efficacy	Can you manage to solve difficult problems, can you find means and ways, stick to your aims, and accomplish your goals, do you know to handle unforeseen situations, invest necessary effort, can you find several solutions, and handle whatever situations in online learning?	0.999	Positive
Familiarity with Cloud Technologies	Technology you are familiar with in using online teaching materials namely Videos from YouTube, Facebook Classroom, Google Classroom, ZOOM, Gmail & yahoo, Twitter, Whatsapp & Webchat, Google Meet, Flipgrid	0.233	Negative
Readiness in Participation	Already are you familiar with online learning, your knowledge and skills are sufficient, ready with PCs and stable internet connection, does your device works efficiently, ready to motivate yourself, ready to manage the time, ready to submit your assignments and take exams through online program, ready to foster a positive online learning environment with peers and teachers?	0.997	Positive
Effectiveness of Online Learning	Does online learning offer convenience, meet your individual needs, contribute to effective communication, increase sense of community, and promotes greater student participation and interaction?	0.999	Positive

Power BI is used in conjunction with features of cognitive services text analytics to figure out sentiments. Fig. 6 shows collection of visualization charts that appear together within power BI dashboard. This clearly illustrates effectiveness and readiness has highest positive sentiment scores.

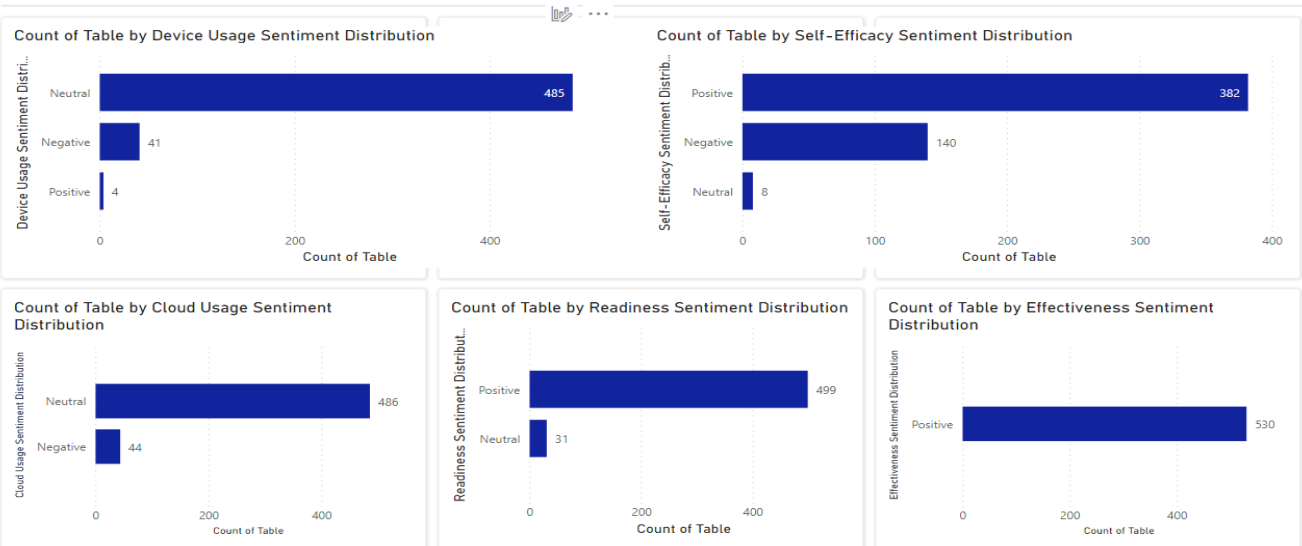

Fig. 6. Power BI detailed results for each fold.

VII. CONCLUSIONS AND FUTURE WORK

Detecting students' at-risk and assessing their sentiments is more beneficial for higher education institutions. Over recent years, researchers have shown ensemble models in machine learning is a proven method for improving model accuracy, robustness, generalizability, prediction performance and has become one of the hottest trends in data science. Depending on the literature review of risk prediction in higher education institutions, the most appropriate machine learning algorithm is selected for learner classification. When boosting is accomplished appropriately by choosing tuning parameters, then it can generalize well and convert a weak model to robust model. Ensemble methods outperform a single learner that is prone to overfitting or underfitting and integrate them to generate a stronger model. In most scenarios, ensembling makes the model more robust and stable by confirming decent

performance on test cases. They were adopted as optimal solutions to obtain satisfied accuracy of prediction. The integrated ensemble framework simultaneously tunes the parameters and increases stability of the model. The novelty of this research is demonstrated through improving Random Forest boosted by XGBoost algorithm that has greater influence of final performance. The results revealed that this novel method gives better accuracy of 93%.

Another vision of this research emphasizes the role of sentiment analysis and shows how to identify cognitive behavior of students in online learning, providing better insights on usage of sentiment analysis in higher education sector. Students' effectiveness regarding online learning mode was highly positive. Contrary to traditional classrooms, virtual classrooms offer limitless opportunities for integrating innovative teaching techniques. Students believe that online learning is more effective as each student receives the lecture, irrespective of number of students present in the virtual classroom. Moreover, by posting their queries in chatbox or discussion forums students can clarify their doubts from instructors and peers. These multifaceted factors permit learners to improve their credibility by introducing them to digitalized curriculum thus integrating global immersions to stay relevant with the changing educational trends and demands by enhancing their behavioral characteristics and making them better citizens. As a future work, chatbots can be developed that serve as virtual advisers which can communicate with students through messaging apps. This way, students can follow self-paced learning, at their own relaxed environment.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Ananthi Claral Mary.T conducted the research, analyzed the data, and wrote the article; Dr. Arul Leena Rose.P.J evaluated, supervised, and approved the article; all authors had approved the final version.

References

[1] D. Karahoca, Z. F. Zaripova, A. R. Bayanova, L. S. Chikileva, S. V. Lyalyaev, and X. Baoyun, "During the Covid-19 pandemic, students' opinions on distance education in department of engineering," *Int. J. Eng. Pedagog.*, vol. 12, no. 2, pp. 4–19, 2022, doi: 10.3991/IJEP.V12I2.29321.

[2] Y. T. Badal and R. K. Sungkur, "Predictive modelling and analytics of students' grades using machine learning algorithms," *Education and Information Technologies*, 0123456789, Springer US, 2022.

[3] M. Jawthari and V. Stoffa, "Predicting at-risk students using weekly activities and assessments," *Int. J. Emerg. Technol. Learn.*, vol. 17, no. 19, pp. 59–73, 2022, doi: 10.3991/ijet.v17i19.31349.

[4] G. Barbayannis, M. Bandari, X. Zheng, H. Baquerizo, K. W. Pecor, and X. Ming, "Academic stress and mental well-being in college students: Correlations, affected groups, and COVID-19," *Front. Psychol.*, vol. 13, pp. 1–10, 2022, doi: 10.3389/fpsyg.2022.886344.

[5] J. Niyogisubizo, L. Liao, E. Nziyumva, E. Murwanashyaka, and P. C. Nshimyumukiza, "Predicting student's dropout in university classes using two-layer ensemble machine learning approach: A novel stacked generalization," *Comput. Educ. Artif. Intell.*, vol. 3, 2022, doi: 10.1016/j.caeai.2022.100066.

[6] E. D. Evangelista, "A hybrid machine learning framework for predicting students' performance in virtual learning environment," *Int. J. Emerg. Technol. Learn.*, vol. 16, no. 24, pp. 255–272, 2021, doi: 10.3991/ijet.v16i24.26151.

[7] P. Sokkhey and T. Okazaki, "Hybrid machine learning algorithms for predicting academic performance," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 1, pp. 32–41, 2020, doi: 10.14569/ijacsa.2020.0110104.

[8] Z. Kanetaki, C. Stergiou, G. Bekas, C. Troussas, and C. Sgouropoulou, "A hybrid machine learning model for grade prediction in online engineering education," *Int. J. Eng. Pedagog.*, vol. 12, no. 3, pp. 4–23, 2022, doi: 10.3991/IJEP.V12I3.23873.

[9] J. Chen, J. Feng, X. Sun, N. Wu, Z. Yang, and S. Chen, "MOOC dropout prediction using a hybrid algorithm based on decision tree and extreme learning machine," *Mathematical Problems in Engineering*, 2019, doi: 10.1155/2019/8404653.

[10] F. J. Alsolami, "A hybrid approach for dropout prediction of MOOC students using machine learning," *International Journal of Computer Science and Network Security*, vol. 20, no. 5, pp. 54–63, 2020.

[11] Y. Zhang, L. Chang, and T. Liu, "MOOCs dropout prediction based on hybrid deep neural network," in *Proc. 2020 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC)*, 2020, pp. 197–203, doi: 10.1109/CyberC49757.2020.00039.

[12] S. Assami, N. Daoudi, and R. Ajhoun, "Implementation of a Machine learning-based MOOC recommender system using learner motivation prediction," *Int. J. Eng. Pedagog.*, vol. 12, no. 5, pp. 68–85, 2022, doi: 10.3991/ijep.v12i5.30523.

[13] L. K. Smirani, H. A. Yamani, L. J. Menzli, and J. A. Boulahia, "Using ensemble learning algorithms to predict student failure and enabling customized educational paths," *Sci. Program.*, vol. 2022, doi: 10.1155/2022/3805235.

[14] Y. Sun, Z. Li, X. Li, and J. Zhang, "Classifier selection and ensemble model for multi-class imbalance learning in education grants prediction," *Appl. Artif. Intell.*, vol. 35, no. 4, pp. 290–303, 2021, doi: 10.1080/08839514.2021.1877481.

[15] N. Hutagaol and Suharjito, "Predictive modelling of student dropout using ensemble classifier method in higher education," *Adv. Sci. Technol. Eng. Syst.*, vol. 4, no. 4, pp. 206–211, 2019, doi: 10.25046/aj040425.

[16] N. Mduma, K. Kalegele, and D. Machuve, "An ensemble predictive model based prototype for student drop-out in secondary schools," *J. Inf. Syst. Eng. Manag.*, vol. 4, no. 3, 2019, doi: 10.29333/jisem/5893.

[17] P. M. Da Silva, M. N. C. A. Lima, W. L. Soares, I. R. R. Silva, R. A. De Fagundes, and F. F. De Souza, "Ensemble *regression* models applied to dropout in higher education," in *Proc. 2019 Brazilian Conference on Intelligent Systems, BRACIS 2019*, pp. 120–125, doi: 10.1109/BRACIS.2019.00030.

[18] M. Wankhade, A. C. S. Rao, and C. Kulkarni, "A survey on *sentiment* analysis methods, applications, and challenges," *Artificial Intelligence Review*, vol. 55, no. 7, pp. 5731-5780, 2022, doi: 10.1007/s10462-022-10144-1.

[19] T. D. Pham *et al.*, "Natural language processing for analysis of student online sentiment in a postgraduate program," *Pacific J. Technol. Enhanc. Learn.*, vol. 2, no. 2, pp. 15–30, 2020, doi: 10.24135/pjtel.v2i2.4.

[20] L. M. Nkomo and B. K. Daniel, "Sentiment analysis of student engagement with lecture recording," *TechTrends*, vol. 65, no. 2, pp. 213–224, 2021, doi: 10.1007/s11528-020-00563-8.

[21] A. M. Gattan, "Deep learning technique of sentiment analysis for twitter database," *Int. J. Interact. Mob. Technol.*, vol. 16, no. 1, pp. 184–193, 2022, doi: 10.3991/IJIM.V16I01.27575.

[22] M. A. Alencar, J. F. M. Netto, and F. Morais, "A sentiment analysis framework for virtual learning environment," *Appl. Artif. Intell.*, vol. 35, no. 7, pp. 520–536, 2021, doi: 10.1080/08839514.2021.1904594

[23] K. Ravi, V. Siddeshwar, V. Ravi, and L. Mohan, "Sentiment analysis applied to educational sector," *2015 IEEE International Conference on Computational Intelligence and Computing Research, ICCIC 2015*, doi: 10.1109/ICCIC.2015.7435667

[24] O. A. Ayeni *et al.*, "Web-based Student opinion mining system using sentiment analysis," *Int. J. Inf. Eng. Electron. Bus.*, vol. 12, no. 5, pp. 33–46, 2020, doi: 10.5815/ijieeb.2020.05.04.

[25] Z. K. Ozturk, Z. I. E. Cicek, and Z. Ergul, "Sentiment analysis: An application to Anadolu University," *Acta Phys. Pol. A*, vol. 132, no. 3, pp. 753–755, 2017, doi: 10.12693/APhysPolA.132.753.

[26] S. Pande, "An overview of sentiment analysis using azure cognitive services," *International Journal of Management, IT & Engineering*, vol. 9, no. 10, pp. 123-127, 2019.

[27] F. Yu and X. Liu, "Research on student performance prediction based on stacking fusion model," *Electron.*, vol. 11, no. 19, 2022, doi: 10.3390/electronics11193166

[28] M. Savargiv, B. Masoumi, and M. R. Keyvanpour, "A new random forest algorithm based on learning automata," *Comput. Intell. Neurosci.*, vol. 2021, doi: 10.1155/2021/5572781.

[29] L. He, R. A. Levine, J. Fan, J. Beemer, and J. Stronach, "Random Forest as a predictive analytics alternative to regression in institutional research," *Pract. Assessment, Res. Eval.*, vol. 23, no. 1, pp. 1–16, 2018.

[30] D. Yates, and M. Z. Islam, "FastForest: Increasing random *forest* processing speed while maintaining accuracy," *Information Sciences*, vol. 557, pp. 130–152, 2021, doi: 10.1016/j.ins.2020.12.067

[31] K. Yan, "Student performance prediction using XGBoost method from a macro perspective," in *Proc. 2021 2nd International Conference on Computing and Data Science, CDS 2021*, pp. 453–459.

[32] G. S. Prakasha, R. Sangeetha, S. M. Almeida, and A. Chellasamy, "Examining university students' attitude towards e-Learning and their academic achievement during COVID-19," *Int. J. Inf. Educ. Technol.*, vol. 12, no. 10, pp. 1056–1064, 2022, doi: 10.18178/ijiet.2022.12.10.1720.

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).