An Early Intervention Technique for At-Risk Prediction of Higher Education Students in Cloud-based Virtual Learning Environment using Classification Algorithms during COVID-19

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

Abstract—**Higher Education is considered vital for societal development. It leads to many benefits including a prosperous career and financial security. Virtual learning through cloud platforms has become fashionable as it is expediency and flexible to students. New student learning models and prediction outcomes can be developed by using these platforms. The appliance of machine learning techniques in identifying students at-risk is a challenging and concerning factor in virtual learning environment. When there are few students, it is easy for identification, but it is impractical on larger number of students. This study included 530 higher education students from various regions in India and the outcomes generated from online survey data were analyzed. The main objective of this research is to predict early identification of students at-risk in cloud virtual learning environment by analyzing their demographic characteristics, previous academic achievement, learning behavior, device type, mode of access, connectivity, self-efficacy, cloud platform usage, readiness and effectiveness in participating online sessions using four machine learning algorithms namely K Nearest Neighbor (KNN), Support Vector Machine (SVM), Linear Discriminant Analysis (LDA) and Random Forest (RF). Predictive system helps to provide solutions to low performance students. It has been implemented on real data of students from higher education who perform various courses in virtual learning environment. Deep analysis is performed to estimate the at-risk students. The experimental results exhibited that random forest achieved higher accuracy of 88.61% compared to other algorithms.**

Keywords—Prediction; at-risk; machine learning; virtual learning environment; cloud platforms; classification; COVID-19; random forest; student academic performance

I. Introduction

During COVID-19 crisis, the entire education system all over the world has shifted towards virtual learning. Online teaching is highly dependent on the successful delivery of the content. Cloud-based Virtual Learning Environment (VLE) is vital component of education in university environments. This interactive platform enables learners to achieve education objectives during pandemic outbreak. The growth of cloud computing technology has brought new opportunities in the field of education as it facilitates effective and efficient learning mechanism. Benefits of cloud computing includes collaborative learning environment, expense reduction, scalability, shareable content, usability, and global education. This induces the way online learning can be shared and distributed on diverse types of devices and platforms. Some of the biggest organizations including Amazon, Google, Microsoft, and Oracle are selecting the cloud due to its several benefits. Currently, due to the rising of educational institutions every day there exists a gap between industrial requirements and educational institutions. The technological enhancement of cloud computing might fill the gap by rendering free or paid training to the students' using systems that do not require any additional cost. For instance, educators can widen most crucial sections, upload necessary audio/video materials to support contents in cloud-based platforms, etc. However, despite of these advantages, students' learning behaviors and interaction with digital contents is still limited. Hence, this research analyzes at-risk students in cloud-based virtual learning platforms. Early prediction using classification techniques is an efficient and significant way to deliver timely intervention for dropout.

Many researchers have investigated that there are numerous ways of applying machine learning algorithms in education field. One of the vital focuses is to predict at-risk students in VLE by observing diverse student's attributes. Compared with traditional face-to-face education, dropout rates are higher as there was a sudden transformation towards virtual learning. COVID-19 pandemic has led to increase in digitalization and triggered online learning implementation in education sectors. All over the globe, learning sectors are focusing on virtual learning platforms to enhance the procedure of enlightening students. The effectiveness of online medium depends on students. The predominant usage of VLE has helped them to complete their course work examinations. However, evaluating student's results and predicting at-risk students is complex. Personalized assistance must be given to the students at-risk. In this context, this research focuses on identifying exact predictive models to investigate unidentified information regarding students' demographics characteristics, previous academic history, learning behavior, device type, mode of access and connectivity for online classes, self-efficacy, cloud platform usage, readiness, and effectiveness of participation in virtual

*Corresponding Author

classes. We have examined that these are the appropriate features that makes an impact of student learning and helps in prediction.

Recently, for schools and higher education sectors online learning has emerged as a vital source and will prolong in future. In virtual learning courses, the collaboration between learners and educators is mediated by virtual learning environment. Chatbox helps the students to interact with instructors and take part in the classroom discussions. Machine learning is widely used in education sectors for classroom management, scheduling, etc. It is a method of personalized learning that provides an individualized educational experience to students where they can manage their own learning, at their own pace, make their own decisions about what to learn. In a classroom using personalized learning, students select what they are interested in, and educators fit the curriculum and guidelines to students' interests. Scheduling helps in searching for an optimal and adaptive teaching policy that helps students learn more efficiently. Dynamic scheduling matches students requiring assistance with teachers allocating time. It is regarding designing algorithms that automatically bring out valuable information from the given data. While machine learning has many success stories, there are software available to design and train rich and flexible machine learning systems. The mathematical foundations are important to build complicated machine learning systems. These models can be created by using diverse features of student data. The machine learning algorithms can be used to develop successful classifiers.

II. Literature Review

Three determinant factors based on technological, organizational, and environmental contexts impact the acceptance of cloud computing in higher education sectors. Survey method was used to collect data from respondents and powerful statistical tool SmartPLS examined the significance of each of these influencers. Results show that factors of compatibility, security, peak management support, authoritarian policy and relative advantage have positive effect [1]. Dataset was chosen from UK Open University, which enclosed students' activity logs, assignment, and final marks, all stored in VLE logs. Feature selection was essential for designing accurate prediction models thereby facilitating students to improve their performance. A predictive model was constructed by the researchers to early forecast students at-risk of dropout. SELogisticRegression and Input-Output Hidden Markov Model (IOHMM) outperformed other baseline models. The overall accuracy of their proposed model is 84% [2]. Almajali & Masadeh ascertained how enabling circumstances, social media, comfort of utilization affect students' opinion for online learning during COVID-19. Their study proved that enabling conditions had a positive impact. Furthermore, they have discussed about the difficulties faced by the students during pandemic like anxiety, lack of device, issues in internet connectivity, etc. Their findings revealed that students who are expertise in utilization of online learning technologies have positive perception towards it [3]. In educational institutions identifying at-risk students was a problematic task. Naive Bayes classifier was selected for the progress of early warning system. Four classification colors were used to represent the various warning levels namely green for non-at-risk, yellow represents possible at-risk, red for at-risk and black color indicates dropout based on students' grades [4]. The study employs dataset from two academic writing courses in Hong Kong University. Logistic regression and classification trees can be utilized in higher education perspective, but ANN is not appropriate. The research team suggested that accuracy can be improved with other datasets [5]. The researchers have used extremely limited student attributes. Their predictions help to classify the students who are potentially at-risk. This information facilitates tutors to make timely intervention to increase student success. Three machine learning techniques DT, KNN, RF were used. Static attributes namely sexual category, age and previous educational results were excluded [6].

During pandemic contentment of higher education students aspired for learning in virtual modality was forecasted. Students responded that they obtain their classes, seminars, videoconferences using Google Meet and Zoom compared to WebEx and Blackboard. Researchers have concluded that students have better opinion towards online learning, the difficulty do not stretch out in technology usage, but it lies in the teachers' teaching strategy [7]. A tool that allows estimating the hazard of quitting an academic course was proposed. Python programming language was used to implement several machine learning algorithms namely LDA, SVM and RF. LDA and SVM was proved to have utmost performance with a slight superior variance for SVM results. When additional learning requirements feature were introduced, in random forest, the final performance was improved compared to LDA and SVM results. The suggestion for the future development is to have more data regarding student performance by considering the outcome of their activities completed in virtual education environments namely Moodle, Google Classroom and Edmodo [8]. Francis Ofori et al. reviewed the literature and summarized the various machine learning models with their corresponding prediction accuracy. It helps to improve the graduation rates by providing feedbacks to educators and students thereby modifying learning environments. They concluded that most machine learning models dwelled in studying students' performance prediction but failed to identify the best model [9]. Identifying students' at-risk by using eBook interaction logs have employed 13 prediction algorithms. Results revealed that random forest performed better than other algorithms with accuracy 82.3% and kappa 64.7% for raw data, J48 algorithm with accuracy 83.3% and kappa 66.5% performed better with transformed data. Naive Bayes accuracy 81.1% and kappa 62.1% outperformed other algorithms for categorical data [10]. Automated machine learning usage was proposed to improve correctness of prediction percentage depending on the data before commencing of their academic year. This aids students to moderate their risk failure and has achieved the overall accuracy of 75.9% [11]. Perceptions of post graduate students towards responding online learning process have been discussed. Majority of the students used Zoom cloud platform for learning from home activities. Others preferred Google Classroom, Whatsapp and other applications based on the agreement provided by each platform. Few students declared that limitations in technology and usage of applications hampered the online learning process [12]. The data of 2097 students of higher education were investigated, and the system was trained with Logistic Regression and ANN with four attributes related to student socio-economic and academic details. Results have shown that highest risk of dropping out of students has lowest grade. Comparing to Logistic Regression, ANN has better classification accuracy. However, the accuracy of ANN did not exceed 80% [13].

The researchers developed an early identification system using student performance and administrative data from private and state university. AdaBoost algorithm was used to predict the student dropout. The results revealed that prediction accuracy improves at fourth semester when compared to first semester. The demographic data available at the time of enrollment does not improve prediction accuracy when performance data is available in increasing semesters; this issue must be solved [14]. Khadija Alhumaid focused on the fear of technology usage by students and educators during Covid-19 pandemic. The various fear factors are uncertainty, anxiety, and fear of losing loved ones. Hence m-learning was adopted by the educational institutions, the results of studying and teaching was promising. She has concluded that with the assistance of mobile learning the fear factors can be reduced. It has a high perceived usefulness and perceived ease of use that can decrease the fear and enable the respondents to achieve their classes on time [15]. Reduced Training Vector-SVM was proposed to predict marginal and at-risk students. Analysis revealed that this algorithm can diminish number of training vectors and training time of classifier by at least 60% thereby retaining accurateness. Results represented an overall accuracy of 93.5% [16]. Two open-source datasets namely mathematics and Portuguese was selected for predicting student educational performance. It was identified that prior grade has most impact on finishing grade. Random Forest has gained higher accuracy in mathematics dataset. SVM attained better accuracy than Random Forest in Portuguese dataset [17]. The four main difficulties of students in virtual learning classes were meager internet connection, lack of experience to virtual education applications and tools namely Teams, Padlet, Socrative and Miro, students' restricted English proficiency and difficulty in concentration. Besides, results have proved that online learning is cost effective, as it's not necessary for students to arrive to campus. They had better time management and have utilized digital devices to access online classes. The students get encouraged when they had quizzes established through interactive applications namely Kahoot, Padlet, Micro, Socrative and many other. Lecturers gave more assignments during online learning period than traditional face-face learning [18]. The troubles faced by at-risk students were analyzed in VLE. These predictive models can be used for avoiding student dropouts. Feature engineering was used to enhance performance of predictive models. Experimental results revealed that random forest provides excellent results when compared to other baseline models [19].

Several machine learning algorithms were employed to the dataset to predict low-engagement students in web-centered learning systems from log data of VLE. Kappa and accuracy values were compared for the models. Outcomes proved that J48, decision tree, JRIP, gradient-boosted classifiers revealed improved results [20]. Predicting students' difficulties in digital design course session were investigated. SVM has achieved 80% performance accuracy compared to other classifiers namely ANN, LR, NBC, DT for predicting student obscurity [21]. Deep long short-term memory model was deployed for students' performance prediction. This model tends to monitor the week-wise pattern of students' interaction and their engagement activities to learn their behavior and generate better outcomes. It outperformed other baseline models namely logistic regression, ANN. It predicted with 90% accuracy of student interaction in VLE [22]. An application was implemented that utilizes academic information of university students and generated classification models by using ANN, ID3 and C4.5. Decision tree built by C4.5 has higher performance measure. Suggestions were given to increase the number of variables and including the institutional and socio-economic variables for further research [23]. The ways to measure fairness in VLE was examined. CGPA was referred as main attribute for student performance prediction. A great underrepresentation of disabled students was determined. This leads to misclassification that disable students were predicted to fail the course. These guidelines must be considered by the researchers [24]. Zulherman analyzed the strength, weakness, opportunities, and threats to Zoom Cloud Meeting, a reliable video platform. The results focused behavioral intention drivers of ZCM usage during crisis are hedonic inspiration and perceived self-efficacy. Influence among these attributes was stronger [25]. Moreover, researchers focused on predicting student's dropout at course level in e-learning course using various machine learning techniques. Five attributes reflecting course activities namely accesses, assignments, tests, exam, and project were considered. Pearson correlation was conducted to identify correlations between independent variables and results. The most appropriate attribute for prediction with high correlation besides project and tests is access that detects students who do not obtain time for accesses that expose them to higher risk. Most adopted algorithms are Logistic Regression, Decision Tree, Naive Bayes Classifier, Support Vector Machine, Random Forest, Neural Network. Results proved that Random Forest classifier obtained best accuracy with 93%, precision reached 86%, F1 score was 91% compared to other classifiers [26]. Researchers have collected real students' data with various information namely personal, economic, and academic records and evaluated by statistics values to find most effective one. They used three prediction algorithms Decision tree, SVM and KNN to detect student dropout. Decision tree reaches better performance for identifying students at high risk. In decision tree they have used J48 technique that represents real dropout groups. Authors have stated that it is vital to evaluate student behavior through multi objective algorithms that assess their skills and emotions [27]. On the other hand, dropout rates are higher in online learning than offline as students must control their own study time without the help of educators. It is vital for professors to assist students in a timely intervention to avoid dropout. Lowering dropout rate is an important challenge for universities. The experiment collected actual log and historical records from Cyber University Learning Management System (LMS) to predict drop-out risk during learning period. They have used four machine learning algorithms Decision Tree, Random Forest, SVM and Deep Neural Network. Random Forest shows the best performance with 96% accuracy [28].

In previous studies it is evident that predicting student dropout is challenging task. When different machine learning algorithms are employed, it reveals varying prediction accuracy of students' at-risk. It is inexplicit that which algorithm is most excellent for predicting at-risk students in virtual learning environments, and what are the most appropriate features to be considered for various machine learning classifiers, as the determinants of attrition depend on multi-dimensional character. Also, during time of registering for online classes, the student data collected at the university are not sufficient for dropout prediction. In this scenario, our research focuses on several factors namely prior academic achievement, demographic characteristics, learning behavior, device type, access, connectivity, self-efficacy, cloud platform usage, readiness, and effectiveness of participation in online classes through Google Classroom and Zoom cloud platforms. We have not focused on a specific predictive model; instead, we have considered four machine learning algorithms to identify best model.

In the light of the reviewed literature above, cloud computing adoption in higher education institutions have a positive effect. Rest part of paper was organized as follows. Section 3 explains research method that provides a description on real time student dataset, preprocessing, feature extraction and normalization. Section 4 describes machine learning algorithms; Section 5 reveals experimental results. Section 6 draws conclusions and the further research guidelines.

III. Research Methodology

A. Research Questions of Inquiry

The intention of this study is to extend an early detection model for finding at-risk students. Machine learning algorithms are implemented on the student's real time dataset collected from various educational institutions, and its accuracy has been analyzed. Within an online learning environment, the system can be globalized to any type of course. Proposed model of research is depicted in Fig. 1. This research aims to respond subsequent questions:

1) To build a predictive system to classify the at-risk students of cloud virtual learning platform.

2) To estimate model accuracy by means of various machine learning algorithms namely KNN, SVM, LDA and Random Forest.

3) To investigate most suited machine learning algorithm for predicting student difficulties based on demographics, device type, access, connectivity, self-efficacy, cloud platform usage, readiness, and effectiveness of participation in online learning sessions.

B. Methodology

A sequence of steps has been adopted prior to the model is organized for assessment and final reporting. Methodological data flow of the complete process is visualized in Fig. 2. Initial step is collecting real data from students, and the data is preprocessed to eliminate missing value, duplications, and outliers. Feature extraction is employed to extract features. Data is normalized and is passed to the four classifiers KNN, SVM, LDA and Random Forest. Classification models have discrete outcome, we need a metric that compares discrete classes in some form. A model's performance can be evaluated using classification metrics and it determines how well or bad the classification is, but each of them evaluates it in a different way. Each classification result is tested with the performance metrics. Finally, the results are visualized.

C. Dataset Description

To meet the objective of the study, an electronic survey was conducted to gather real time data from higher education students belonging to various academic institutions throughout India. The instrument used for this research is E-questionnaire. It was chosen as it was considered as an efficient and effective approach. The target population for this study consisted of 530 students enrolled in online courses. The survey was broadly classified into five factors. (1) Demographic characteristics (2) Device type, mode of access, connectivity (3) Self-Efficacy (4) Familiarity with cloud platform usage (5) Readiness and Effectiveness of online learning compared to regular classroom setting. Demographic information of the respondents is presented in Table I, which shows that 66.60% are male respondents, and 33.40% are females between age group of 17 and 45 years. They are from diverse departments namely B.Sc Computer Science (20.38%), B.Com (14.34%), M.C.A (11.32%), B.C.A (11.32%), B.Tech (11.32%), M.sc (15.09%), M.com (4.91%) and BBA (11.32%) degree programmes. Online classes were managed for the students through cloud platforms. We have used free cloud services Google Forms and Google Sheets for collecting and analyzing the data. Conceptually, Self-Efficacy was estimated with 5-point Likert-scale ranging from 1= "Strongly Disagree" to 5= "Strongly Agree". Cloud platform usages are measured using 4-point range from "Not Familiar" to "Very Much Familiar". Readiness is given 4-point range from 1= "Not Ready" to 4= "Very Much Ready" and Effectiveness from 1="Much Less Effective" to 5="Much More Effective".

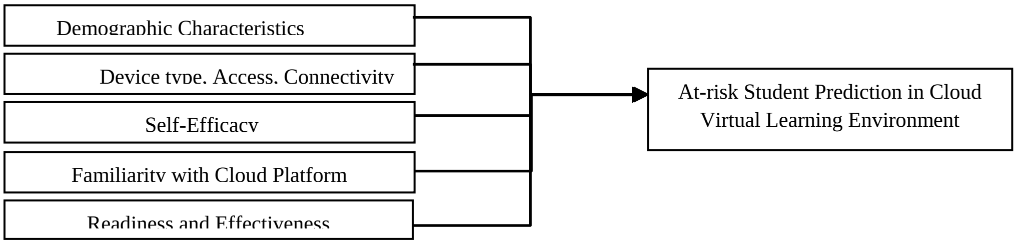
Demographic Characteristics
Device type, Access, Connectivity
Self-Efficacy
Familiarity with Cloud Platform
Readiness and Effectiveness
At-risk Student Prediction in Cloud Virtual Learning Environment

Fig. 1. Proposed Model.

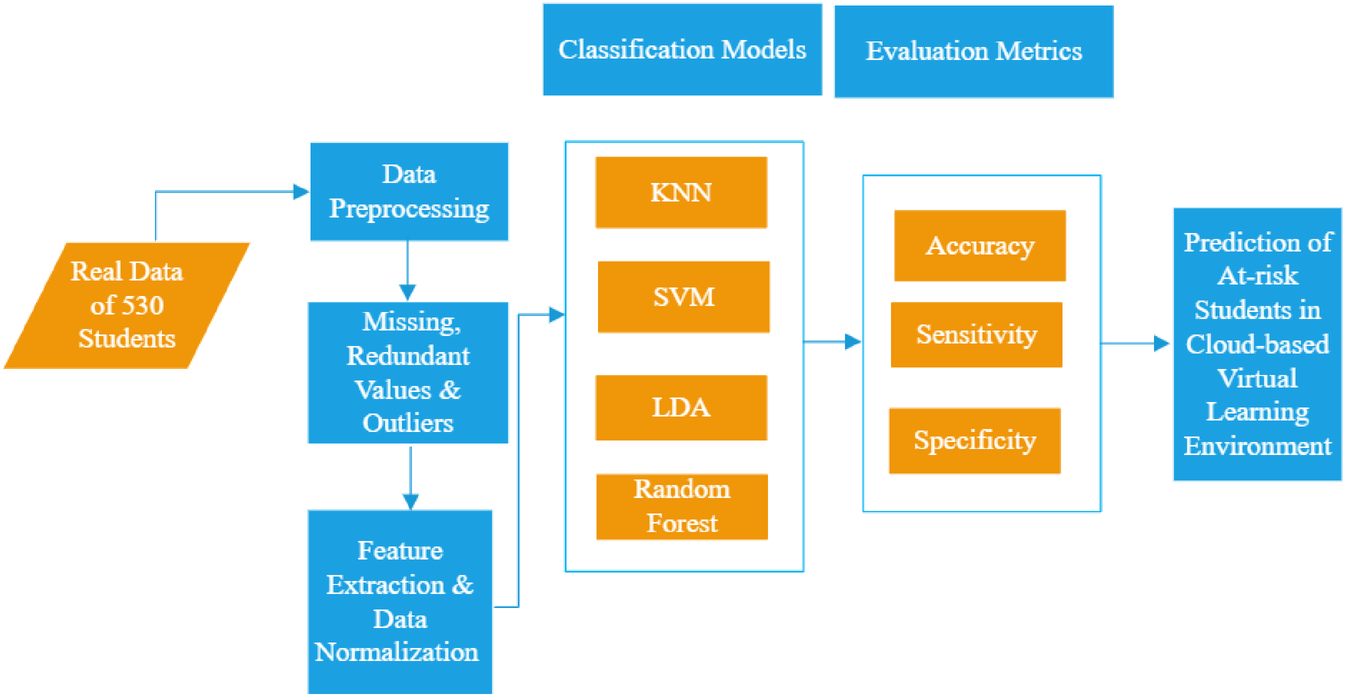
Classification Models
Evaluation Metrics
Real Data of 530 Students
Data Preprocessing
Missing, Redundant Values & Outliers
Feature Extraction & Data Normalization
KNN
SVM
LDA
Random Forest
Accuracy
Sensitivity
Specificity
Prediction of At-risk Students in Cloud-based Virtual Learning Environment

Fig. 2. Conceptual Framework of the Research Model.

TABLE I. SUMMARY OF DEMOGRAPHIC CHARACTERISTICS OF STUDENTS

Variable	Label	n=530	%
Gender	Male	353	66.60
	Female	177	33.40
Age	17-19	158	29.81
	20-29	370	69.81
	30-45	2	0.38
Education qualification	B.Sc Computer Science	108	20.38
	B.Com	76	14.34
	M.C.A	60	11.32
	B.C.A	60	11.32
	B.Tech	60	11.32
	M.Sc	80	15.09
	M.Com	26	4.91
	B.B.A	60	11.32

D. *Data Preprocessing*

Before evaluation phase of the classification model, the dataset was passed through a preprocessing stage. Initially, we observed that the real time dataset contains fifty-six features with missing data, redundant data, and outliers. Few students have submitted the questionnaire redundantly and missed to fill some data. When a specific value could not be estimated for a data sample, it leads to missing data. They have not provided the correct data that tends to form outliers, as the data plunge outside the range of typical distribution. By preprocessing the data, they were eliminated to maintain the quality of prediction. R4.1.1 statistical software was used as an experimental tool. Machine learning was conducted using caret package that is a comprehensive framework for constructing machine learning models in R. The model was fed with more students' engagement activities; self-efficacy in online learning tends to learn about the behavioral patterns. The final dataset was exported into a .csv file and it is ready to be trained with different machine learning algorithms and evaluated to select most accurate one. We have used Microsoft Excel for visualizing the outcomes.

E. *Feature Selection*

It is the procedure of choosing optimal number of attributes from a larger dataset, which is the most difficult and challenging task. From this procedure, we come to know the utmost useful set of features for predicting target variable. To diminish computational cost and to increase model performance it is desirable to reduce number of features. To find the top variables, we are attaining improved cross-validated accuracy and data can be used to identify most likely elements to predict at-risk students. Two reliable measures of random forest algorithm namely %IncMSE and IncNodePurity were utilized for feature selection to generate an optimal subset of features. Finally, we have extracted forty-six features for each student with CGPA as the target variable. The selected features by our model render excellent instructions that educators can utilize to offer early assessments to learners prior to closure of course work.

1) %IncMSE: The first measure is %IncMSE which is a highly informative measure of importance. It defines the mean decrease in accuracy or how the prediction gets poorer when the variable alters its value. The variation among original mean error and randomly permuted mean error is computed and this forms the fundamental idea for measuring important score of variables. For each tree prediction error on test is recorded (Mean Squared Error- MSE). Same process is carried after permuting each predictor variable. Higher the difference, then variable is more important. The general equation is,

MSE = mean ((actual_y – predicted_y) ^2)

2) Mean decrease gini (IncNodePurity): Depending on the gini impurity index this is a variable importance measure for computing splits in trees. Higher value of mean decrease gini score, higher importance of variable.

F. Normalization of Data

3) For machine learning this is a procedure employed as a part of data preparation. Extracted features were originally at various scales. To adjust scales of features to have a standard scale of measure the data was normalized to improve the model accuracy. The formula is given by,

Min-Max Normalization: (x - min(x)) / (max(x) - min(x))

IV. Machine Learning Models

Machine learning is regarding designing algorithms that automatically bring out valuable information from the given data. It is not possible for an “AI” to be trained without data. In every project, classifying and labeling datasets takes most of the time, when it reflects the real time data. The techniques applied to predict the students at risk occur as training and testing phases. To train an algorithm to know how to pertain the concepts, to identify and produce outcomes the training dataset is utilized. When we train a model on dataset, measuring its performance on the same tells how good it is at making predictions on data it has already seen. Training a model on a subset of our data, we can then use the data the model was not trained on to calculate how this would perform on unseen data. These purposefully hold out part of our dataset from training and then use the performance on this held-out dataset as a proxy of our model’s performance in production. The model is constructed on training set and examined on held-out testing set. This allows us to test that our model can generalize to unseen data. 530 students’ academic performance data were randomly divided into two datasets. Training set makes up most total data, around 70% (372 student records). The test data represents 30% (158 student records) that is used to estimate how well our algorithm was qualified with training data. Each technique is presented with data that have not been used during training to observe the classification performance during the testing phase. Before applying classifiers, feature selection methods were implemented using random forest algorithm. CGPA (Cumulative Grade Point Average) indicates the percentage of marks scored by the students. This attribute was used as response or target variable, which logically is the best predictor for course grade. This dependent variable depends on various independent variables namely demographics, previous academic outcomes, learning behaviors, device usage related factors, familiarity of cloud platform usage, self-efficacy, and readiness & effectiveness in participation of online classes. Based on the CGPA we have calculated the risk category of students. The criteria if CGPA<=55, classifies the students as “At-Risk”. Else they are considered as “Non-AtRisk” students. We have used the variable Academic Performance as Boolean attribute depending upon CGPA. This attribute indicates class for supervised binary classification task and result of prediction. For our model, a negative outcome means student was not at-risk. A positive outcome defines student was at-risk. Thus, the students are classified into two categories.

For our research we have used supervised machine learning algorithms, as our dataset have labels. We considered K Nearest Neighbor (KNN), Support Vector Machine (SVM), Linear Discriminant Analysis (LDA) and Random Forest (RF) as they are the most preferable algorithms by researchers to resolve related issues. We validated our methodology by supplying appropriate set of estimation measures. We have assessed the performance of different classifiers with model parameters.

A. KNN Algorithm

KNN categorizes a new data point into target class, based on similarity of its neighboring data points. We input a dataset of atrisk and non-atrisk students. We train our model to identify students’ performance depending on extracted features.

Choose the number k of neighbors

Compute Euclidian Distance between the data points. It is given by,

Euclidian Distance=sqrt$(X_2\text{-}X_1)^2$+ $(Y_2\text{-}Y_1)^2$

Select K nearest neighbor.

Count number of data points in each group, among these K neighbors.

Allot new data point to that group for which number of neighbor is maximum.

B. SVM

A Support Vector Machine is utilized for classification and regression problems. In our dataset, we have used kernel SVM to conduct non-linear partitioning. The main idea behind this algorithm is,

Find the lines that separate the classes optimally. This dividing line is called a hyperplane.

Find the optimal hyperplane that helps in maximizing the margin between two classes.

Transform it into a higher dimension by employing a kernel function to dataset.

Clearly separate the two groups with a plane with the end goal of maximizing the margin.

Once the data points are separated into dimensions, SVM classifies the two groups.

C. LDA

Linear Discriminant Analysis (LDA) considers a data set of observations as input. We require having a categorical variable to define class and several predictor variables for each observation. Steps involved in this algorithm,

Compute mean vectors of each class of dependent variable
Compute with-in class and between-class scatter matrices
Calculate eigenvalues and eigenvector for scatter matrix within class and between class
Sort eigenvalues in descending order and select top k.
Create a new matrix containing eigenvectors that maps to k eigenvalues.
Obtain linear discriminants by taking the dot product of data and matrix.

D. *Random Forest*

This model extends and integrates multiple decision trees to create a forest. It allows for more correct and constant outcomes as it relies on multitude of trees. Steps involved in this algorithm,

From the dataset having k number of records, n records are taken randomly.

For each sample individual decision trees are constructed.

Each decision tree will generate a result.

Outcome is measured depending on majority of the votes for classification or averaging the output of all trees for regression respectively.

V. Result and Discussion

In this part of the study, we predicted the at-risk students from their multidimensional characteristics. To answer the research questions, we performed several experiments. The educators can utilize the predictive model to determine students having difficulties. They can deliver relevant materials, increase the student engagement activities, and improve the marks of such candidates in cloud-based learning platforms. They can acquire corrective actions at former stage, to offer supplementary assist to the students at-risk. This is vital to exactly rank the classifiers depending on their prediction potential of at-risk and subsequent decision making.

A. *Binary Classifier Evaluation Metrics*

Model performance in classification problem is assessed through confusion matrix. It represents four numbers in a two-by-two matrix. Each element displays class-wise accuracy. Reason for depicting this is to obtain benefit of our outstanding visual abilities to process more information. The elements of the confusion matrix are used to find three important parameters namely accuracy, sensitivity, and specificity. We implemented our experiment with 10-fold cross validation; each model we constructed has ensued in Kappa and Accuracy. These assessment metrics can be utilized to assess the value of classifiers for ranking various models. This generates True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN). Accuracy, specificity, sensitivity is used to enhance experimental results in case of binary classification.

- Accuracy is the ratio between accurate predictions over entire number of instances. It is used to determine number of times classifier is correct. This is given by,

Accuracy= (TP+TN)/ Total

- Sensitivity (True Positive Rate) refers to proportion of correct positives that are accurately detected as positives by classifier. Formula is,

Sensitivity=TP/ (TP+FN)

- Specificity (True Negative Rate) relates to classifier's ability to find negative results. The equation is,

Specificity=TN/ (TN+FP)

Table II presents results of KNN, SVM, LDA and RF for real-time dataset. The similar data was passed for all these techniques. As the research aims to classify students' at-risk, it is vital to attain better predictive accurateness for unsuccessful learners. Accuracy from Random Forest is 88.61% it is representing good performance based on accuracy, sensitivity and specificity compared to KNN, SVM and LDA. Fig. 3 visualizes the classification accuracy, sensitivity and specificity obtained using various classifiers. The output shows our model accuracy for test set. The figure clearly depicts that Random Forest algorithm outperformed the other machine learning techniques KNN, SVM and LDA for identifying at-risk students.

TABLE II. Experimental Results of Real-Time Dataset for Predicting Student's At-Risk using Various Classifiers

Classifier	**Accuracy**	**Sensitivity**	**Specificity**
KNN	79.11	97.66	31.32
SVM	84.18	96.09	33.33
LDA	82.28	92.97	36.67
RF	88.61	96.09	56.67

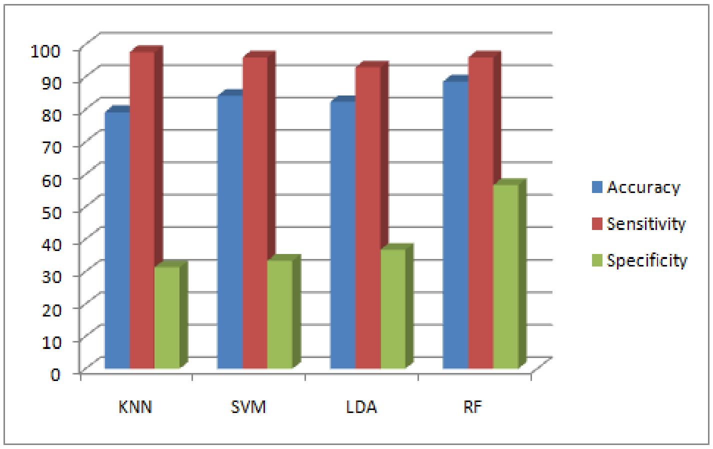

Fig. 3. Visualization Results of Different Classifier for Real-Time Student Dataset.

B. ROC Curves

When we are evaluating classifier performance it is the dominant visualization tool. Performance metrics helps generate an aggregate perspective of a model's performance. For binary classification problems, receiver operating characteristic (ROC) curves can be very informative. It illustrates true positive rate as a function of false positive rate, hence prompting sensitivity of classifier. We partition predictions into positive and negative classes for purpose of obtaining ROC measurements namely specificity and sensitivity normally used on ROC curve axis. For our real-world dataset, if we are classifying whether a student is at-risk or non-atrisk, it is significantly better to categorize a small number of additional non-atrisk students as at-risk and avoid classifying any at-risk as non-at-risk students. We would choose threshold that reduces false-negative rate, increase true-positive rate, and rest us at the top of the ROC plot. This gives an observation of overall performance of the classifier. Fig. 4 visualizes column-wise area under ROC curve (AUC) for KNN, SVM, LDA and RF. For classification models, another performance metric is AUC. This is measure of capability of classifier to differentiate among classes and is utilized as a summary of ROC curve. The classification performance is enhanced when this area is larger. For evaluating and comparing models this is an extensively used option.

An ROC curve gives us a more nuanced view of how a model's performance changes as we make predictions conservative. Table III illustrates the results of ROC curve for various machine learning algorithms. After observing ROC curves of each classifier, it's clear that Random Forest algorithm has the highest statistic of ROC 76.38%. Classification models often use the area under the curve (AUC) to represent performance. It delivers an agreement measure of performance across all possible classification thresholds. Fig. 5 illustrates the highest Area Under Curve that corresponds to random forest algorithm. Higher AUC depicts classifier has outstanding performance to differentiate among positive and negative classes. From these series of experiments, it is apparent that in overall random forest algorithm can be used to detect the students regarded at-risk in cloud-based virtual learning environment based on multidimensional variants.

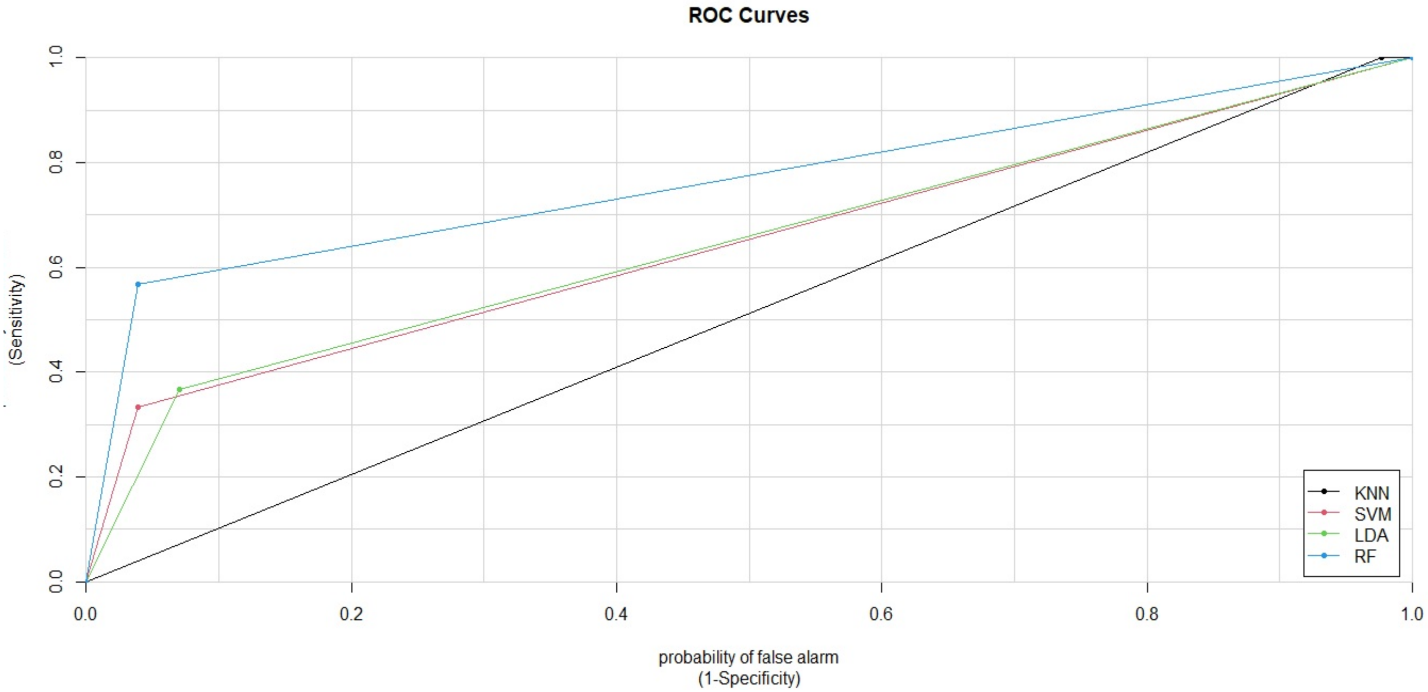

Fig. 4. Column-Wise Area Under ROC Curve (AUC) of Classifiers for Students' At-Risk Prediction.

TABLE III. COMPARATIVE RESULTS OF RECEIVER OPERATOR CHARACTERISTIC (ROC) CURVES FOR KNN, SVM, LDA AND RF

Classifier	ROC Curve Statistic
KNN	51.17
SVM	64.71
LDA	64.82
RF	76.38

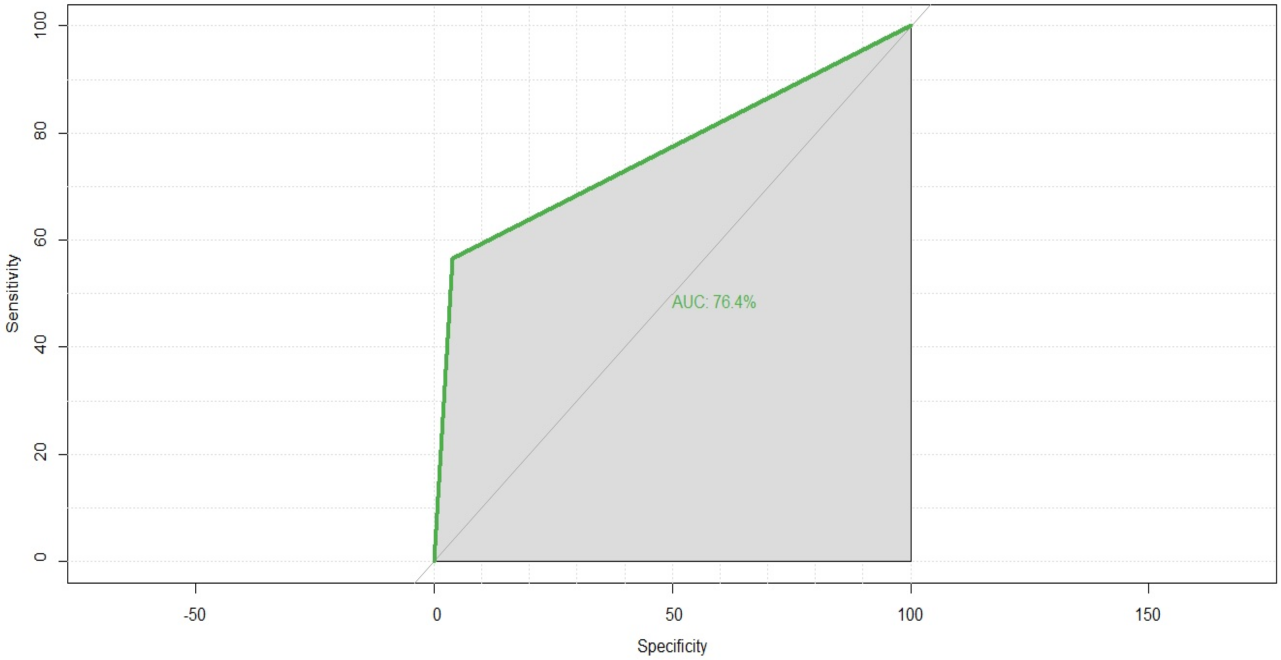

Fig. 5. ROC Curve Illustrating the Highest Area Under Curve corresponding to Random Forest Algorithm.

VI. Conclusion and Future Work

As there is a sudden shift away from the traditional classroom in many parts of the globe, the adoption of virtual learning is more effective during the COVID-19 crisis. The students are in diverse geographical locations, the cloud platforms have empowered them to learn in their own comfort zone. They had an opportunity to participate in an interactive and collaborative learning environment. The instructors can upload and share the course material in cloud platforms. The students can access the materials from their own individual portal that enables them as independent learners. By applying machine learning algorithms in cloud platforms progress of the students can be tracked hence the learning gaps can be identified. This research work focuses on implementing students' outcome based early prediction model. We employed different machine learning approaches namely KNN, SVM, LDA and RF to detect at-risk students. The lower academic results decrease self-confidence of the learners and depletion of valuable educating efforts. These students always have an intention of dropping out from college. To solve the problem of drop out, this system helps the higher education institutions to classify the risky students at earlier stage. The amount of data generations depends on the students' multivariate characteristics for the courses enrolled in online classes. This was used to create the machine learning model that resourcefully utilizes this data, hence forth bring outcomes that can be used additional in students' wellbeing in terms of their performance and personal growth. The versatility of these systems also helps the teachers to take potential efforts towards the risky students. A sequence of experiments has been managed to identify the most excellent model.

Comparing to the existing research and results, current research revealed that the most promising random forest algorithm achieved high accuracy with 88.61% and outperformed other binary classification models. These algorithms were used to classify the students' at-risk in VLE by considering their multideterminant characteristics. The outcomes from this model can profoundly help educators, to upgrade their existing teaching methodologies and the implementation of new techniques facilitates the students to pay additional attention in their studies. This data-driven study can support VLE administrators, instructors, and course co-coordinators in the articulation of effective virtual learning structure that can bestow to process of decision-making. This early intervention technique that was implemented in virtual learning environment motivates the students to have high academic scores.

Depending on our results machine learning is highly recommended to be integrated with cloud platforms. It gives an insight into real-time situations that allows the higher educational institutions to forecast future outcomes. In further research, we plan to deploy predictive model in cloud computing platform by means of helping the educators to progress the performance of unprepared students and for automating the decision-making process.

References

[1] Z. M. A. Abdullah Alghushami, Nur Haryani Zakaria, "The determinants impacting the adoption of cloud computing in Yemen institutions," in *AIP Conference Proceedings*, 2018, pp. 1–7, Available: https://doi.org/10.1063/1.5055424.

[2] H. C. & W. Z. Ahmed A. Mubarak, "Prediction of students' early dropout based on their interaction logs in online learning environment," *Interact. Learn. Environ.*, 2020.

[3] D. A. A. and R. Masa'deh, "Antecedents of students' perceptions of online learning through covid-19 pandemic in Jordan," *Int. J. Data Netw. Sci.*, vol. 5, no. 4, pp. 587–592, 2021.

[4] M. S. David Baneres, M.Elena Rodriguez-Gonzalez, "An Early Feedback Prediction System for Learners At-Risk Within a First-Year

Higher Education Course," *IEEE Trans. Learn. Technol.*, vol. 12, no. 2, pp. 1–14, 2019.

[5] D. Foung, "Redesigning Prediction Algorithms for At-Risk Students in Higher Education: The Opportunities and Challenges of Using Classification Techniques in a University Academic Writing", Book: "Redesigning Higher Education Initiatives for Industry 4.0," IGI Global, pp. 232-250, 2019.

[6] Y. S. Edward Wakelam, Amanda Jefferies, Neil Davey, "The potential for student performance prediction in small cohorts with minimal available attributes," *Br. J. Educ. Technol.*, vol. 51, no. 2, pp. 347–370, 2020.

[7] E. García-Salirrosas, "Satisfaction of university students in virtual education in a COVID-19 scenario," 3rd International Conference on Education Technology Management (ICETM), 2020, pp. 41-47.

[8] S. P. Z. Francesca Del Bonifro, Maurizio Gabbrielli, Giuseppe Lisanti, "Student Dropout Prediction," *21st Int. Conf. Artif. Intell. Educ. AIED*, pp. 129–140, 2020.

[9] D. R. G. Francis Ofori, Dr. Elizaphan Maina, "Using Machine Learning Algorithms to Predict Students' Performance and Improve Learning Outcome: A Literature Based Review," *J. Inf. Technol.*, vol. 4, no. 1, pp. 33–55, 2020.

[10] H. O. GokhanAkcapinar, Mohammad Nehal Hasnine, Rwitajit Majumdar, Brendan Flanagan, "Developing an early-warning system for spotting at-risk students by using eBook interaction logs," *Smart Learn. Environ.*, vol. 6, no. 4, pp. 1–15, 2019.

[11] A. F. Hassan Zeineddine, Udo Braendle, "Enhancing prediction of student success: Automated machine learning approach," *Comput. Electr. Eng.*, vol. 89, no. 4, 2020, [Online]. Available: https://doi.org/10.1016/j.compeleceng.2020.106903.

[12] A. F. Ihsana El Khuluqo, Abdul Rahman A. Ghani, "Postgraduate students' perspective on supporting 'learning from home' to solve the COVID-19 pandemic," *Int. J. Eval. Res. Educ.*, vol. 10, no. 2, pp. 615–623, 2021.

[13] R. G.-C. Ivan Sandoval-Palis, David Naranjo, Jack Vidal, "Early Dropout Prediction Model: A Case Study of University Leveling Course Students," *Sustainability*, vol. 12, no. 9314, pp. 1–17, 2020.

[14] J. B. Johannes Berens, Kerstin Schneider, Simon Gortz, Simon Oster, "Early Detection of Students at Risk- Predicting Student Dropouts Using Administrative Student Data from German Universities and Machine Learning Models," *J. Educ. Data Min.*, vol. 11, no. 3, pp. 1–41, 2019.

[15] K. Alhumaid, "Developing an educational framework for using mobile learning during the era of COVID-19," *Int. J. Data Netw. Sci.*, vol. 5, no. 3, pp. 215–230, 2021.

[16] T. M. L. Kwok Tai Chui, Dennis Chun Lok Fung, Miltiadis D. Lytras, "Predicting At-risk University Students in a Virtual Learning Environment via a Machine Learning Algorithm," *Comput. Human Behav.*, vol. 107, 2020, Available: https://doi.org/10.1016/j.chb.2018.06.032.

[17] N. A. Leena H. Alamri, Ranim S. Almuslim, Mona S. Alotibi, Dana K. Alkadi, Irfan Ullah Khan, "Predicting Student Academic Performance using Support Vector Machine and Random Forest," in *3rd International Conference on Education Technology Management (ICETM)*, 2020, pp. 1–8.

[18] M. Revani Putri, K. Oktriono, C. Sidupa, M. Willyarto "Portraying Students' Challenges and Expectations toward Online Learning in Embracing Industrial Revolution 4.0 Era: A case in ELT in the COVID-19 Outbreak,"3rd International Conference on Education Technology Management (ICETM), 2020, pp. 36-40.

[19] M. B. and S. U. K. Muhammad Adnan, Asad Habib, Jawad Ashraf, ShafaqMussadiq, Arsalan Ali Raza, Muhammad Abid, "Predicting at-Risk Students at Different Percentages of Course Length for Early Intervention Using Machine Learning Models," *IEEE Access*, vol. 9, pp. 7519–7539, 2021.

[20] S. M. R. A. Mushtaq Hussain, Wenhao Zhu, Wu Zhang, "Student Engagement Predictions in an e-Learning System and Their Impact on Student Course Assessment Scores," *Comput. Intell. Neurosci.*, vol. 2018, no. 6347186, pp. 1–21, 2018.

[21] S. A. Mushtaq Hussain, Wenhao Zhu, Wu Zhang, Syed Muhammad Raza Abidi, "Using machine learning to predict student difficulties from learning session data," *Artif. Intell. Rev.*, vol. 52, pp. 381–407, 2019, Available: https://doi.org/10.1155/2018/6347186.

[22] S.-U. H. Naif RadiAljohani, Ayman Fayoumi, "Predicting At-Risk Students Using Clickstream Data in the Virtual Learning Environment," *Sustainability*, vol. 11, no. 24, pp. 1–12, 2019.

[23] P. NorkaBedregal-Alpaca, Víctor Cornejo-Aparicio, Joshua Zárate-Valderrama, "Classification Models for Determining Types of Academic Risk and Predicting Dropout in University Students," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 11, no. 1, pp. 266-272, 2020.

[24] K. S. and V. S. Shirin Riazy, "Fairness in Learning Analytics: Student At-risk Prediction in Virtual Learning Environments, *"In Proceedings of the 12th International Conference on Computer Supported Education (CSEDU)*, 2020, pp. 15–25.

[25] F. M. Z. Zulherman, ZalikNuryana, AstadiPangarso, "Factor of Zoom cloud meetings: Technology adoption in the pandemic of COVID-19," *Int. J. Eval. Res. Educ.*, vol. 10, no. 3, pp. 816–825, 2021.

[26] M. D. Janka Kabathova, "Towards Predicting Student's Dropout in University Courses Using Different Machine Learning Techniques," *Appl. Sci.*, vol. 11, no. 7, pp. 1–19, 2021, Available: https://doi.org/10.3390/app11073130.

[27] J. A. C. R. Luis Earving Lee, Salvador Ibarra Martinez, M. G. T. B. Jesus David Teran Villanueva, Julio Laria Menchaca, and E. C. Rocha, "Evaluation of Prediction Algorithms in the Student Dropout Problem," *J. Comput. Commun.*, vol. 8, pp. 20–27, 2020, Available: https://doi.org/10.4236/jcc.2020.83002.

[28] S. J. Y. Hee Sun Park, "Early Dropout Prediction in Online Learning of University using Machine Learning," *Int. J. Informatics Vis.*, vol. 5, no. 4, pp. 347–353, 2021, Available: http://dx.doi.org/10.30630/joiv.5.4.732.