Comparison of Predictive Machine Learning Models to Predict the Level of Adaptability of Students in Online Education

Orlando Iparraguirre-Villanueva[1], Carmen Torres-Ceclén[2], Andrés Epifanía-Huerta[3], Gloria Castro-Leon[4], Melquiades Melgarejo-Graciano[5], Joselyn Zapata-Paulini[6], Michael Cabanillas-Carbonell[7]
Facultad de Ingeniería y Negocios, Universidad Privada Norbert Wiener, Lima, Perú[1]
Facultad de Ingeniería, Universidad Católica los Ángeles de Chimbote, Perú[2]
Facultad de Ingeniería, Universidad Tecnológica del Perú, Chimbote, Perú[3]
Facultad de Ingeniería y Gestión, Universidad Nacional Tecnológica de Lima Sur, Lima, Perú[4]
Facultad de Ciencias Empresariales, Universidad Científica del Sur, Lima, Perú[5]
Escuela de Posgrado, Universidad Continental, Lima, Perú[6]
Facultad de Ingeniería, Universidad Privada del Norte, Lima, Perú[7]

***Abstract*—With the onset of the COVID-19 pandemic, online education has become one of the most important options available to students around the world. Although online education has been widely accepted in recent years, the sudden shift from face-to-face education has resulted in several obstacles for students. This paper, aims to predict the level of adaptability that students have towards online education by using predictive machine learning (ML) models such as Random Forest (RF), K-Nearest-Neighbor (KNN), Support vector machine (SVM), Logistic Regression (LR) and XGBClassifier (XGB).The dataset used in this paper was obtained from Kaggle, which is composed of a population of 1205 high school to college students. Various stages in data analysis have been performed, including data understanding and cleaning, exploratory analysis, training, testing, and validation. Multiple parameters, such as accuracy, specificity, sensitivity, F1 count and precision, have been used to evaluate the performance of each model. The results have shown that all five models can provide optimal results in terms of prediction. For example, the RF and XGB models presented the best performance with an accuracy rate of 92%, outperforming the other models. In consequence, it is suggested to use these two models RF and XGB for prediction of students' adaptability level in online education due to their higher prediction efficiency. Also, KNN, SVM and LR models, achieved a performance of 85%, 76%, 67%, respectively. In conclusion, the results show that the RF and XGB models have a clear advantage in achieving higher prediction accuracy. These results are in line with other similar works that used ML techniques to predict adaptability levels.**

Keywords*—*Machine learning; adaptability; students; online education; prediction; models

I. Introduction

In recent years, online education has experienced an unprecedented boom. The COVID-19 pandemic further accelerated this process, forcing millions of students and educators to adapt to a fully digital educational environment [1]. In response to the educational crisis caused by the COVID-19 pandemic, UNESCO has established partnerships with various international organizations, such as the International Labour Organization (OIT), the United Nations High Commissioner for Refugees (UNHCR), the United Nations Children's Fund (UNICEF) and the World Health Organization (WHO) [2]. Companies such as Microsoft, Google, Facebook, and Coursera [3] also participate. With the aim of addressing the problems of connectivity, content through digital tools and seeking distance learning solutions for children, youth, and adults through innovation. Today, online education is not only a viable alternative to face-to-face education, but has become an indispensable tool for many students, professionals, and individuals around the world. This type of education is offered through online platforms and tools, such as websites, mobile applications, discussion forums, videoconferencing, among others [4], [5]. Since the advent of the COVID-19 pandemic, online education has been an indispensable alternative for students around the world [6]. However, the sudden transition from face-to-face to online education has posed many challenges to students, especially in terms of their ability to adapt to this new educational environment [7]. Adaptability is a fundamental skill that allows us to face new and unfamiliar situations with flexibility and creativity [8]. In the case of online education, adaptability refers to the ability of students to adjust to a new educational model [9], which requires different skills and tools than those used in face-to-face education [10]. However, not all students have the same ability to adapt to online education and some may have difficulty staying motivated and engaged [11]. Early identification of these students can be key to providing them with appropriate support and improving their online academic success [12]. For example, in Latin America and the Caribbean, Brazil has the highest rate of online students with 49%, Mexico with 15%, Colombia with 11%, Argentina with 8%, Chile with 6% and Peru with 5% [13] as shown in Fig. 1.

Fig. 1. Representation of online education in Latin America and the Caribbean.

Online education has revolutionized the world of teaching and learning, providing new opportunities to access education from anywhere in the world [14], [15]. In recent years, ML has started to play a very important role in online education, enhancing the learning experience and personalizing education for each student [16], [17]. ML is a discipline within the field of artificial intelligence (AI) that provides the ability for computers to learn and improve their performance on a specific task without being explicitly programmed. In the field of online education, ML is used to analyze student data and create personalized learning models [18], [19]. This work aims to predict the level of adaptability of students in online education for which we use the following models: RF, KNN, SVM, LR and XGB of ML. Furthermore, this work aims to provide a more in-depth understanding of how ML is transforming online education and how it can be used to improve education in the future. In addition, it will examine the challenges and limitations associated with the use of ML in online education. The predictive capacity of the models is also sought, for which the results of the trained models are compared.

This document is structured as follows: Section II presents the related works. Section III explains the materials and methods used, as well as the process of data collection and understanding. Section IV presents the results obtained after model processing and training. Section V discusses the analysis and discussion of the results, as well as the performance of the models. Finally, Section VI presents the conclusions drawn from the study and proposes future work.

II. Related Works

Predicting the level of student adoptability in online education has become an important issue. In this context, researchers and academics have carried out work related to ML models for the analysis of distance education and student adaptability, these methods involve the use of algorithms that allow the analysis of large volumes of data and the extraction of significant patterns. For example, in [20] they developed a study on the role of student adaptability in virtual education during the COVID-19 pandemic, in this work they used a population of 1548 students. The findings showed that adaptability is a crucial personal resource that can help students in their online learning process. Similarly, in [21] The authors conducted an investigation to estimate the level of student adaptability in virtual education using ML techniques, where they explored the gap between virtual and face-to-face education from a Bangladeshi context. They found that 8.3% of the students adapted with high ease to online education, they also identified factors that make it difficult for them to adapt, such as Internet connectivity, lack of knowledge in the use of digital tools, problems of access to electricity, among others. Also, in [22] they presented the results of a survey conducted in June 202 to beginning teachers, analyzing factors such as technology, Internet accessibility, teachers' competencies, pedagogical knowledge and learning opportunities. The results showed that for teachers to adapt easily to online education, they must have certain digital competencies, such as the adoption of information and communication technologies. Similarly, in the following work [23] they proposed an adaptive system to predict an educational path by applying different data mining algorithms to extract relevant features and build a personalized model. The results showed that ML algorithms are efficient in their performance and accuracy. Similarly, [24] made use of ML algorithms to predict student performance in online education. For which they used deep neural network algorithms and SVM and decision tree (DT) classifiers. As a result, they concluded that the SVM classifier is the one that obtains the best results in its accuracy and performance. Furthermore, in [25] examined the prospects, advantages, and problems of artificial intelligence (AI) in online education. They used ML algorithms to classify the level of adaptability of the students. As a result, an accuracy level of 93% was obtained, RF models and neural networks proved to be the most efficient. Similarly, the authors in [26] developed a study to determine the influence of IA on students' social adaptability. They used a population of 1328 students. They used ML algorithm to determine the influence on student adaptability, using variables such as: teacher-student, parents and society, age, connectivity, technological competencies, among others. The results showed that IA has a positive influence on adaptability in online education. Neural network classifiers are contributing significantly to learning prediction. For example, in the work [27] they developed a paper that seeks to associate learning styles with distance education student behavior, in which they applied linear regression models. The results of the study indicated that no significant relationship was found between learning styles and student performance in distance education. Furthermore, in the article [28], a study was conducted on the performance of students in online education using the deep learning short- and long-term memory (LSTM) technique, in which a total of 22437 students participated. As a result, an accuracy of 0.8457, precision of 0.8224 and F1-Score of 0.7943 were obtained using the LSTM technique.

III. METHODOLOGY

The purpose of this section is to present the methodology used in this work, which is divided into two parts: A) description of the ML models (RF, KNN, SVM, LR and XGB); B) the development of the case study. Within the case development process, the following stages are followed: 1) data set collection; 2) data set processing and exploratory data analysis; 3) training and validation of the ML models.

A. Description of the ML Models

ML is composed of a set of algorithms for regression, classification, reduction, clustering, among other types of algorithms. However, for this work that aims to predict the adaptability of online students, five classification algorithms have been chosen, such as: RF, KNN, SVM, LR and XGBC.

1) Random forest: The supervised learning (SL) algorithm known as RF model, according to [29], is used for classification and regression. This model is made up of a series of decision trees, where each is trained with a random subset of data and cast its vote for the outcome, as described in [30], as depicted in Fig. 2. This reduces overfitting and increases the accuracy of the model. It also allows the identification of the most important features for prediction.

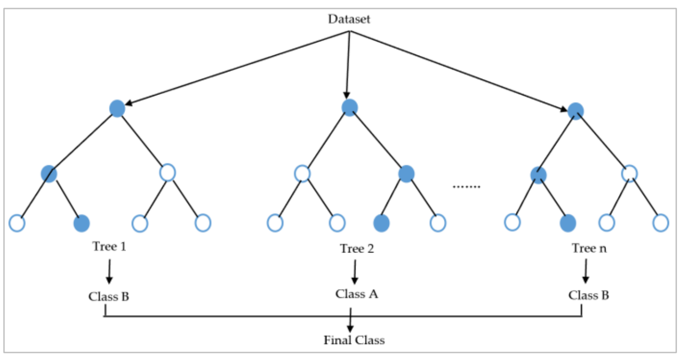

Fig. 2. RF model classification process.

2) K-Nearest neighbor: The K-NN model is an SL algorithm used for classification and regression [31]. In the context of regression, the k nearest neighbor's method consists of estimating the value of a data point from the values of the k nearest neighbors to that point. K-NN is simple, efficient, and easy to understand, but it can be sensitive to outliers and does not work well with high dimensionality data [32]. The formulation for calculating the vector distance is represented in equation (1).

$$d(q,p) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2} \qquad (1)$$

where q_n represents the results of one observation and p_n also represents the results of another observation.

3) Support vector machine: The SVM model is an SL algorithm used for classification and regression[33]. This hyperplane is found by maximizing the distance between the closest points of each class, called support vectors [34]. SVM is effective in high dimensionality spaces and can handle nonlinear data using kernel functions [35]. However, it can be sensitive to hyperparameter selection and requires a large amount of data processing resources.

4) Logistic regression: The LR model is an SL algorithm used for binary classification. It models the probability that a data point belongs to the positive or negative class using the logistic function [36]. The model is trained using the likelihood function and fit using optimization methods such as gradient descent [37]. The LR is simple and easy to understand but can have problems with nonlinear data and can be sensitive to outliers [38]. The model is represented in equation (2) and (3).

$$PY = 1|X = \frac{1}{1+exp\, w_{o+} \sum_{i=1}^{n} w_i x_i} \qquad (2)$$

and

$$PY = 0|X = \frac{exp\, w_{o+} \sum_{i=1}^{n} w_i x_i}{1+exp\, w_{o+} \sum_{i=1}^{n} w_i x_i} \qquad (3)$$

Where, PY|X represents the parametric distribution, Y is the described value, and X= x1, ... xn is a vector with continuous values.

5) eXtreme Gradient Boosting: XGB is a decision tree-based ML algorithm used for regression and classification [39]. XGB improves the model at each iteration by adding sequential trees that focus on the most difficult to predict data points [40]. It uses a combination of gradient boosting and regularization to avoid overfitting and improve model accuracy [41]. XGB is scalable and can be used on large and complex datasets.

B. Case Study

1) Understanding and collecting the data set: In this section, the main task is to analyze and understand the dataset on online student adaptability to extract relevant information that can help improve the online learning experience. The first step is to understand the context of online education and how students' adaptability affects their experience. The dataset used in this work was obtained from Kaggle, this dataset is composed of a population of 1205 students between high school and college. The dataset is composed of 14 attributes such as: gender, age, educational level, type of educational institution, technology student (yes/no), location, internet quality, economic condition, type of internet, type of device to connect to the internet, duration of classes, knowledge of LMS, class schedule and level of adaptability, as shown in Fig. 3.

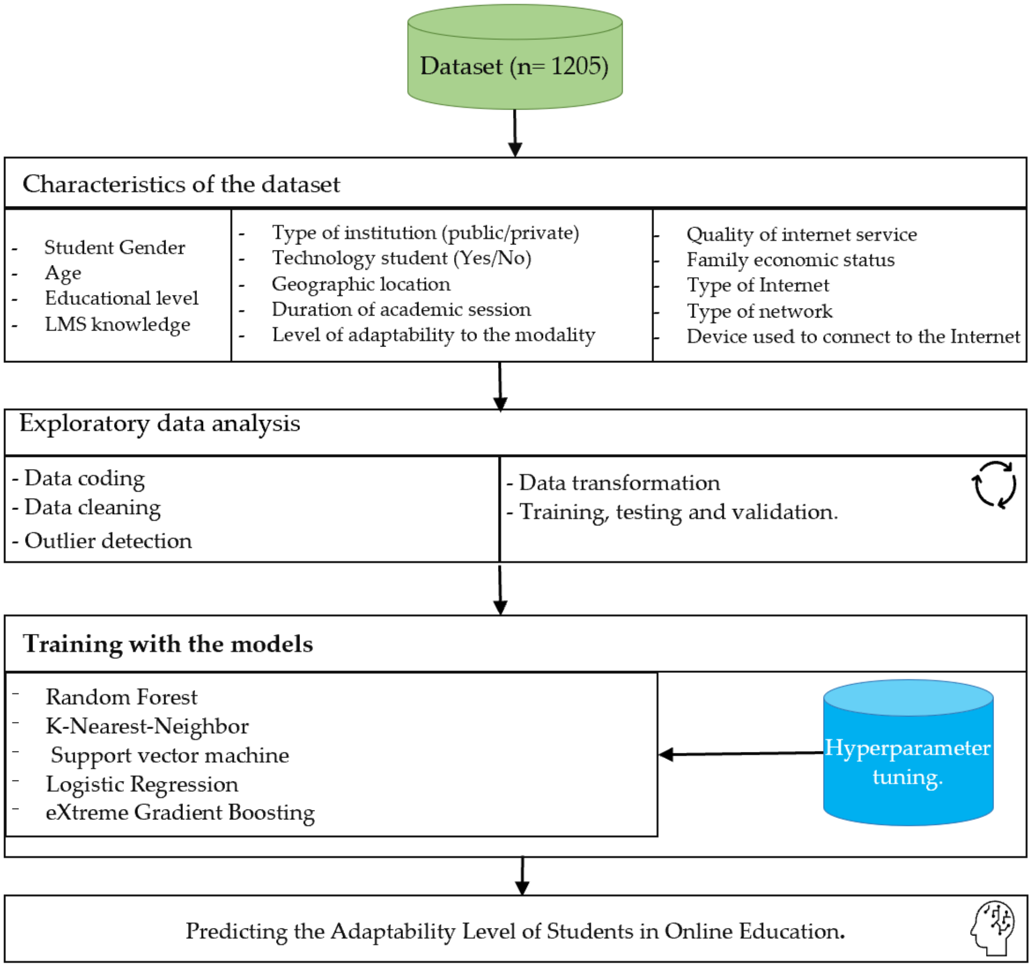

Fig. 3. The process shows the extraction, selection of attributes, analysis, training with ML models and training results.

Generally, when the data set is small, as is the case, and to solve this limitation, Synthetic Minority Over-sampling Technique (SMOTE) is used in this paper to deal with class imbalance in the data. SMOTE solves the problem by creating synthetic instances by interpolating the minority class in the dataset in order to increase the size of the minority class in the dataset. In this context, predicting the adaptability of online learners directly contributes to educators and educational program managers by anticipating students' needs and difficulties and providing immediate answers and supports to improve their online learning experience. In addition, ML methods directly help designers of online educational programs to improve the quality of their educational offerings and make them more accessible and effective for students.

2) Data set processing: The processing of the dataset involves performing a set of tasks, such as: data cleaning, transforming, and analyzing the collected data. As a first step we have to: import the libraries for loading the dataset, then we generally explore the data. In it, we can find the data types for each attribute, the content and description of the attributes, as shown in Table I. As a second step we proceed to eliminate any duplicate, missing or erroneous information. Also, outliers and inconsistencies in the data are checked to ensure that they are accurate and reliable.

Next, the statistical values of the data set are checked for example, the number of students, unique values, frequency, and highest values as presented in Table II.

Performing a simplified analysis of the data set by constructing a general distribution of the characteristics, the following interpretation is reached. The gender variable is balanced, since the same number of males and females is applied to obtain a relevant result. Most of the students correspond to the age of 11-25 years; this is the age at which one speaks with confidence of sustainable adaptation to online learning. The values provided by the level of education indicate that most of the students have only a school education. This data correlates with a given timetable, in which many students are between the ages of 7 and 20. Existing educational services are provided by private institutions. This is due to the development of the educational market. It should also be noted that most of the educational programs are sold by telephone through the 4G network, which shows that people receive education at any place convenient for them.

Next, we analyze the number of students according to their level of adaptation in online education.

TABLE I. GENERAL ANALYSIS OF THE DATA SET

	Student Gender	Age	Educational level	Institution Type	Financial Condition	Internet Type	Device
0	Male	21	University	Non-public	Middle class	Wi-Fi	Tab
1	woman	20	University	Non-public	Middle class	Mobile Data	Mobile
2	woman	16	College	Non-public	Middle class	Wi-Fi	Mobile
3	woman	12	School	Non-public	Middle class	Mobile Data	Mobile
4	woman	15	School	Non-public	Lower class	Wi-Fi	Mobile
…	…	…	….	…	…	…	…
1201	woman	16	College	Non-public	Middle class	Wi-Fi	Mobile
1202	woman	16	College	Non-public	Middle class	Wi-Fi	Mobile
1203	Male	12	School	Non-public	Middle class	Mobile Data	Mobile
1204	woman	17	College	Non-public	Middle class	Wi-Fi	Mobile
1205	woman	11	College	Non-public	Lower class	Mobile Data	Mobile

TABLE II. DATA SET STATISTICS

	Count	unique	top	freq
Student's gender	1205	2	Male	663
Age range	1205	6	21-25	374
Student's educational level	1205	3	School	530
Type of educational institution	1205	2	private	823
Student's IT knowledge	1205	2	No	901
Student's geographic location	1205	2	Yes	935
Power system interruption	1205	2	Low	1004
Student's economic status	1205	3	Mid	878
Type of mobile / fixed Internet	1205	2	Mobile Data	695
Type of connectivity	1205	3	4G	775
Session duration	1205	3	1-3	840
LMS knowledge	1205	2	No	995
Device used	1205	3	Mobile	1013
Degree of student adaptability	1205	3	Moderate	625

According to the graphical representation presented in Fig. 4, it is observed that men tend to adjust more easily to new knowledge, while the rate of maladaptation is similar for both genders. In addition, it is evident that the optimal age for successful adaptation is between 21 and 25 years, while the worst adaptation occurs in people older than 26 years and in the age range of 16 to 20 years. This information suggests that age may play an important role in the ability to adapt to online education. The lower capacity to adapt to new knowledge can be attributed to various factors, both social and physiological. Importantly, those who belong to the middle class tend to take better advantage of online educational materials. In addition, a greater adaptive capacity has been observed among those students who reside in organized communities or environments, which may be related to economic and social factors. It is also important to consider that the level of adaptation to new knowledge is also influenced by the quality of the available Internet service.

Once the exploratory analysis of the data set has been completed, the next step is to perform the processing of the data set. As a starting point, it consists of dividing the data set in two, 80% will be used for training and 20% for testing. With Sklearn.Model_selection.train_test_split(), it is possible to perform the splitting process by simply specifying the size of the test part for training. As in the exploratory analysis of the data set, nulls or missing values have been discovered. Therefore, it is necessary to remove them from the dataset in order to handle them, estimate those values and then add a mean value to the dataset, for which we use libraries such as: Simple-Imputer () and impute.fit_transform(x_train, x_test). We then proceed to normalize the data. There is no doubt that data normalization is a very important factor, since it consists of reorganizing and scaling each of the features according to the unit variance, so that all features can be compared equally. The following libraries are used for feature normalization: StandardScaler(), tandardScaler.fit_transform(x_train_Impute, x_test_Impute) and data_train_normalized(). Since the data set contains a relatively small number of observations, the SMOTE technique will be used to avoid over-fitting the data set. The objective of this technique is to compare the accuracy of the original data set with the accuracy of the sampled data set. This comparison involves comparing their accuracy. By using the SMOTE technique, class imbalance can be balanced, since class imbalance can lead to asymmetry and overfitting in training.

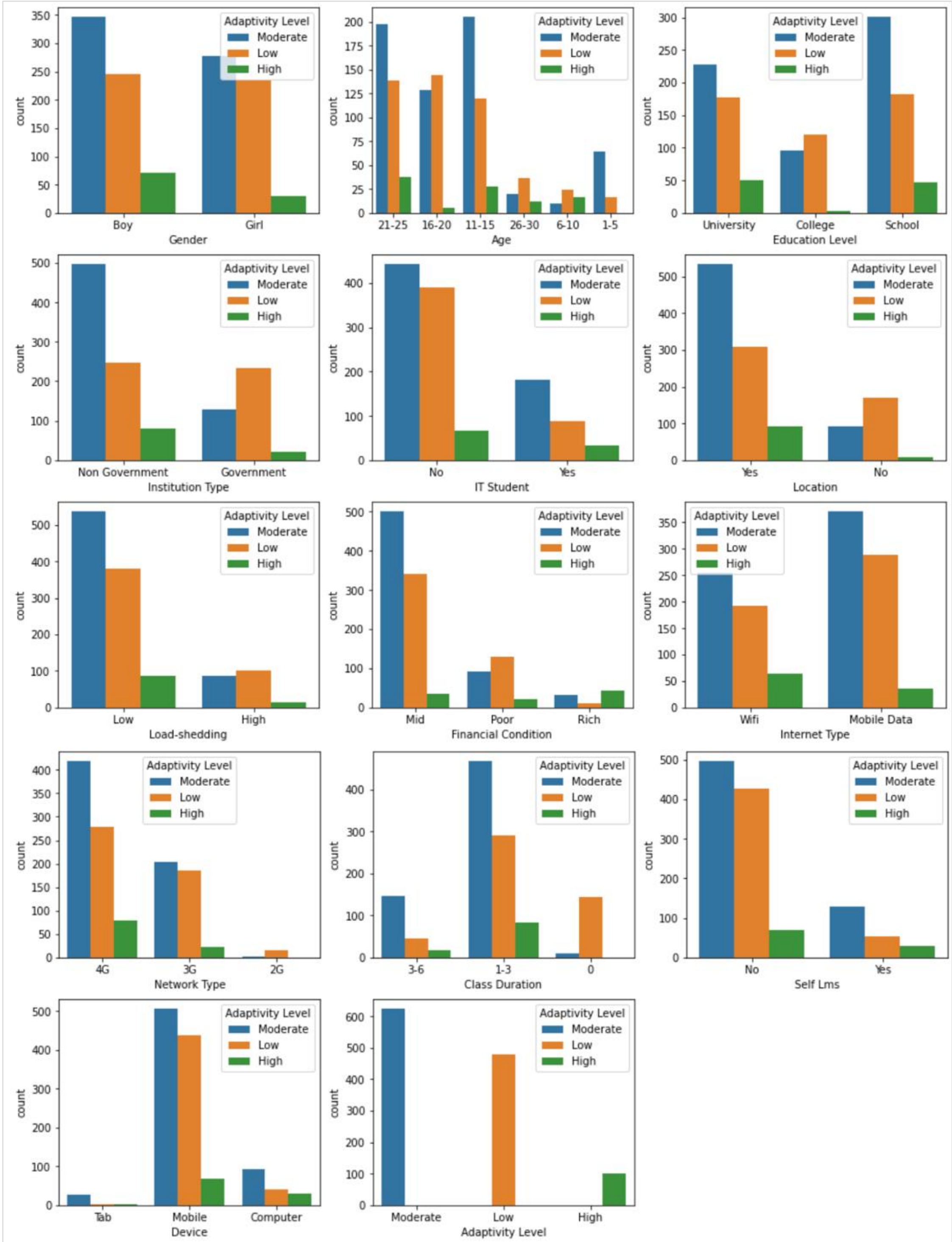

Fig. 4. Analysis of the level of adaptability.

3) Training and data validation: Dataset training is a process, in which ML models are fit to a dataset. During this process, the models are fitted to minimize the margin of error between the model predictions and the actual values of the data used in training the model. The purpose of this process is to achieve a model that can generalize well to unseen data, i.e., that can make accurate predictions on new data. For each model is fitted with SMOTE technique, it means that, one uses the original data set, and the other uses the SMOTE balanced data set. For the parameter selection process, the cross-validation technique was used, which consists of training in two parts, first selecting one part of the data (training) and evaluating it with another part (test set), this process is repeated several times for different combinations. The result is obtained by promising the results of each experiment, which gives us a more accurate assessment of the predictive ability of the model. Cross-validation is a very useful technique because it allows using all available data to train and evaluate the model, which can increase accuracy and reduce the risk of over-fitting. In addition, cross-validation allows us to compare different ML models and select the best model for our data set. To evaluate the performance of ML models, several measures are employed, which are defined as: a) True Positives (TP), in this work are classified as positive samples which are those students who have moderate level of adaptability to online education; b) True Negatives, in this work are classified as negative samples those students who do not have a level adaptability to online education; c) False Positives, are the samples that have been misclassified as positive and d) False Negatives, are the samples that have been misclassified as negative, as seen in Fig. 5.

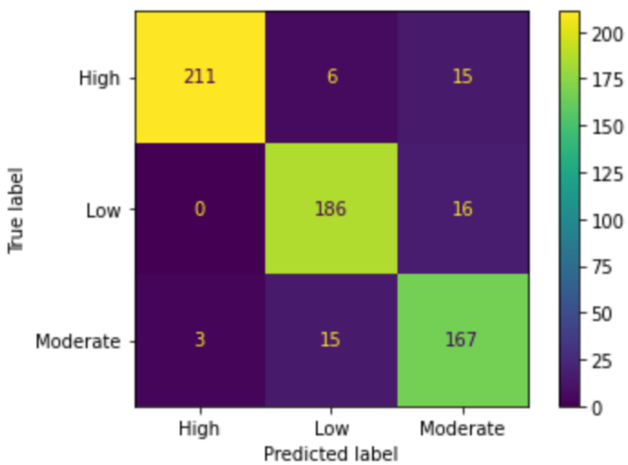

Fig. 5. Variable confusion matrix.

Fig. 5 shows the confusion matrix in general. For example, it can be seen that 211 samples are classified as positive (students with high level of adaptability), 186 samples are classified as positive (have high internet connectivity) and 167 samples are also classified as positive (students with moderate level of adaptability). After training the models, it is important to evaluate their performance to make sure it is accurate. This process follows the following steps: testing the models, in this work 20% of the dataset is used; evaluation metrics are used to assess the performance of the models, accuracy, recall, F1-Score and support, as shown in the following section in Table III.

IV. Results

Training of various ML models, including RF, K-NN, SVM, LR and XGB, was carried out using a specific Kaggle dataset. Subsequently, a learning algorithm was developed to train the models. Subsequently, the performance of each model was evaluated using unobserved data. To carry out this evaluation, various metrics such as accuracy, precision, recall, ROC curve and F1 score were used. The results of these evaluations are presented in Table III.

TABLE III. Results of Trained Models

Random Forest				
	Precision (%)	**Recall (%)**	**f1-score (%)**	**support**
0	91	98	94	214
1	93	91	92	207
2	91	85	88	198
accuracy			92	619
macro avg	92	91	91	619
weighted avg	92	92	92	619
KNN classifier				
	Precision (%)	**Recall (%)**	**f1-score (%)**	**support**
0	86	100	93	214
1	86	82	84	207
2	86	73	78	198
accuracy			85	619
macro avg	85	85	85	619
weighted avg	85	85	85	619
Support vector machine				
	Precision (%)	**Recall (%)**	**f1-score (%)**	**support**
0	83	86	85	214
1	83	67	74	207
2	65	76	70	198
accuracy			76	619
macro avg	77	76	76	619
weighted avg	77	76	76	619
Logistic Regression				
	Precision (%)	**Recall (%)**	**f1-score (%)**	**support**
0	82	70	76	214
1	71	63	66	207
2	54	69	61	198
accuracy			67	619
macro avg	69	67	68	619
weighted avg	69	67	68	619
xGBClassifier				
	Precision (%)	**Recall (%)**	**f1-score (%)**	**support**
0	91	100	95	214
1	92	91	92	207
2	93	84	88	198
accuracy			92	619
macro avg	92	92	92	619
weighted avg	92	92	92	619

In this study, a standardized procedure has been carried out to calculate the true positive (TP) and false positive (FP) rate of all trained models. To illustrate this process, in the case of the RF model, the TP rate is obtained by dividing the number of true positives by the total number of positive cases. The FP rate, on the other hand, is obtained by dividing the number of false positives by the total number of negative cases. The same formula is applied for the other models (K-NN, SVM, LR, XGB). These rates are important metrics to evaluate the performance of a classification model and is applied to adjust the classification threshold of a model. For example, if you want to increase the rate of true positives, you can decrease the classification threshold of the model so that it predicts more samples as positive. However, this could also increase the FP rate. Therefore, it is important to find an appropriate balance between PT and FP. In this case, the RF, K-NN, SVM, LR and XGB models were trained, and the following performances were obtained: 91%, 85%, 76%, 67% and 92%, respectively.

Analyzing Table III, the RF and XGBClassifier models achieved the best rates in the f1-score, recall and accuracy metrics, in this order are analyzed. RF, 92%, 92%, 92%; XGBClassifier, 92%, 92% and 92%, although it is true that in this case, we analyze it at the average level. Therefore, it can be affirmed that the two models present optimal performances to predict the level of adaptability in online education; the second model to obtain better metrics is KNN with the following rates: 85% in F1-score, 85% in recall and 85% in precision; it is followed by the SVM model with 76% of F1-Score, 76% in recall and 77% of precision. Finally, the LR model achieved 68% in F1-Score, 67% in recall and 69% in precision. To reach a better interpretation of the results, it was decided to analyze the importance of each characteristic of the data set. For this purpose, the feature_importances technique was used to determine which features are the most important for the model. The calculation is based on how much each feature contributes to the model in terms of impurity reduction at the node of the tree where it is located. Therefore, the feature that obtains the highest score is considered the most important for the model. This is shown in Table IV.

TABLE IV. IMPORTANT CHARACTERISTICS OF THE DATA SET

#	Features	Feature importances
7	Financial Condition	16.35 %
1	Age	13.72 %
10	Class Duration	12.69 %
0	Gender	8.56 %
9	Network Type	7.54 %
3	Institution Type	6.08 %
2	Education Level	6.08 %
5	Location	5.59 %
8	Internet Type	5.54 %
12	Device	4.96 %
11	Self Lms	4.82 %
4	IT Student	4.37 %
6	Load-shedding	3.64 %

V. DISCUSSION

In the last three years, online education has experienced tremendous growth and has become a popular form of learning due to the COVID-19 pandemic. However, not all students have the same ability to adapt to this type of education. Some students may have more difficulty adapting to online education due to the lack of face-to-face interaction with the professor and their classmates, the need to be more autonomous and organized in their own study time. Predicting the level of adaptability of students in online education is an important factor for online educators, as it allows them to identify and support students who may have difficulties adapting to this type of education. ML methods are a useful tool for predicting the level of student adaptability in online education. Table IV shows that the variable economic condition, which represents the most important factor for student adaptability in online education, is followed in importance by the following variables: age, duration of sessions, type of network used, type of institution, level of education, geographic location, type of Internet, among other variables. ML models such as RF, KNN, LR, SVM, and XGB were used to analyze the data and rank which of the models is optimal for predicting the level of student adaptability in online education. These ML models provided valuable information about which variables are the most important for students' adaptability in online education and how they are related to each other, as shown in Table IV. For example, Economic condition is the most important variable for adaptability, which represents 16.35% in importance, which is related to the results obtained in the study [20], where they concluded that student's financial condition plays a very important role in student's adaptability in online education. Similarly, the results of this work are related to [21], where they explored the gaps that exist in online and face-to-face education, and concluded that the variable of economic condition, age and Internet connection are determining factors for adaptability to online education. The results of the training with the predictive models showed that the RF (92%) and XGB (92%) model achieved the best rates for predicting the level of student adaptability in online education, which is related to the results obtained in the work [24], where they used ML classifiers and deep neural networks, and concluded that the SVM classifiers and the Decision tree, achieved the best results in accuracy and performance. AI has contributed significantly in the academic field, so much so that, unlike this work, in [23] and [25] examined perspectives, advantages and problems of AI in online education, for which they used different ML algorithms, and concluded that RF is the something-rhythm that achieved the best metrics, with an accuracy of 93%, therefore, it is the best suited for this type of tasks, this result is superior to that achieved in this work. The results will depend on the nature of the research and the volume of the data set. In the same line, in [26] they used IA to determine the influence on the adaptability of students in online education, using variables such as: teacher-student, parents and society, age, connectivity, among others. The results showed that AI has a positive influence on the adaptability of students in online education. Unlike the ML models used in this work to predict the level of adaptability of students in online education, in [28], they used the deep learning LSTM technique to analyze the performance of students in online education, where they obtained that deep learning techniques are a very good option to predict the level of adaptability, reaching an accuracy of 84. Predicting the adaptability level of students in online education using ML methods is a valuable tool for online educators. It can provide important information about which variables are most important for student adaptability in online education and how they relate to each other. In addition, it can be used to develop personalized support strategies for students who may have difficulty adapting to this type of education.

VI. CONCLUSIONS

After performing training and comparison of ML (RF, KNN, VSM, LR and XGB) predictive models for predicting the Adaptability Level of students in online education, using the Kaggle dataset consisting of 14 attributes, it reaches the following conclusions.

It was determined that the RF and XGBClassifier models obtained the best results in accuracy and performance. Therefore, for predicting the level of adaptability of students in online education, they are the best predictors, which suggests that these models can be useful for improving the effectiveness of online education. Without detracting from the other models, they also obtained excellent performance results. The models used in this study demonstrated high prediction accuracy.

It was also found that the variables used in the set of variables are important for the accuracy of the predictions. Table IV shows the relevance of each of the variables for predicting the level of adaptability of students in online education, the most important being economic condition, age, duration of sessions, gender, type of network, type of institution, directly influencing the level of adaptability of students.

Finally, predictive ML models can be a valuable tool for predicting the level of student adaptability in online education. The use of ML models, such as RF, KNN, VSM, LR, and XGB, and the careful selection of variables can significantly improve the accuracy of predictions. These findings contribute in a significant way for online education and can help educators improve the effectiveness of their online educational programs. In the future, a possible development that would complement the use of the models would be the development of work to assess the academic performance of students in online education.

REFERENCES

[1] R. E. Baticulon et al., "Barriers to Online Learning in the Time of COVID-19: A National Survey of Medical Students in the Philippines," Med Sci Educ, vol. 31, no. 2, pp. 615–626, Apr. 2021, doi: 10.1007/S40670-021-01231-Z.

[2] UNESCO, "List of Geoparks & Regional Networks," 2023. https://en.unesco.org/global-geoparks/list (accessed Mar. 08, 2023).

[3] U. Nations, "Human Development Report 2021-22," Human Development Reports, 2022, Accessed: Mar. 08, 2023. [Online]. Available: https://hdr.undp.org/content/human-development-report-2021-22

[4] M. M. H. Suzan, N. A. Samrin, A. A. Biswas, and M. A. Pramanik, "Students' Adaptability Level Prediction in Online Education using Machine Learning Approaches," 2021 12th International Conference on Computing Communication and Networking Technologies, ICCCNT 2021, 2021, doi: 10.1109/ICCCNT51525.2021.9579741.

[5] A. Jennifer G., M. George Thomas, and R. Vijay Solomon, "Does Virtual Titration Experiment Meet Students' Expectation? Inside Out from Indian Context," J Chem Educ, vol. 99, no. 3, pp. 1280–1286, Mar. 2022, doi: 10.1021/ACS.JCHEMED.1C01034.

[6] A. Besser, G. L. Flett, and V. Zeigler-Hill, "Adaptability to a Sudden Transition to Online Learning During the COVID-19 Pandemic: Understanding the Challenges for Students," Scholarsh Teach Learn Psychol, vol. 8, no. 2, pp. 85–105, 2022, doi: 10.1037/STL0000198.

[7] L. Yazel, C. Bishop, and H. Britt, "Adapting During the Ever-Changing Pandemic Environment: A One-Year Examination of How Health Education Specialists Remain Adaptable," Am J Health Educ, vol. 53, no. 3, pp. 142–148, 2022, doi: 10.1080/19325037.2022.2048746.

[8] Y. Zhang, G. Zhao, and B. Zhou, "Does learning longer improve student achievement? Evidence from online education of graduating students in a high school during COVID-19 period," China Economic Review, vol. 70, p. 101691, Dec. 2021, doi: 10.1016/J.CHIECO.2021.101691.

[9] C. Guo and B. Wan, "The digital divide in online learning in China during the COVID-19 pandemic," Technol Soc, vol. 71, p. 102122, Nov. 2022, doi: 10.1016/J.TECHSOC.2022.102122.

[10] E. M. Azila-Gbettor, M. K. Abiemo, and S. N. Glate, "University support and online learning engagement during the Covid-19 period: The role of student vitality," Heliyon, vol. 9, no. 1, p. e12832, Jan. 2023, doi: 10.1016/J.HELIYON.2023.E12832.

[11] F. Li et al., "Perceptions towards online learning among medical students during the COVID-19 pandemic," Heliyon, vol. 9, no. 2, p. e13119, Feb. 2023, doi: 10.1016/J.HELIYON.2023.E13119.

[12] C. Tarchi, E. W. Brante, M. Jokar, and E. Manzari, "Pre-service teachers' conceptions of online learning in emergency distance education: How is it defined and what self-regulated learning skills are associated with it?," Teach Teach Educ, vol. 113, p. 103669, May 2022, doi: 10.1016/J.TATE.2022.103669.

[13] M. Bower, "eLearning Statistics: 2023 Data, Trends & Predictions," Design of Technology-Enhanced Learning, pp. 261–304, Aug. 2023, doi: 10.1108/978-1-78714-182-720171011.

[14] Y. Ma, R. Zhang, and Y. Zhang, "Visualization analysis of junior school students' pubertal timing and social adaptability using data mining approaches," Heliyon, vol. 8, no. 8, p. e10443, Aug. 2022, doi: 10.1016/J.HELIYON.2022.E10443.

[15] M. Gamboa-Ramos, R. Gómez-Noa, O. Iparraguirre-Villanueva, M. Cabanillas-Carbonell, and J. L. H. Salazar, "Mobile Application with Augmented Reality to Improve Learning in Science and Technology," International Journal of Advanced Computer Science and Applications, vol. 12, no. 10, pp. 487–492, 2021, doi: 10.14569/IJACSA.2021.0121055.

[16] A. A. Biswas, A. Majumder, M. J. Mia, I. Nowrin, and N. A. Ritu, "Predicting the enrollment and dropout of students in the post-graduation degree using machine learning classifier," International Journal of Innovative Technology and Exploring Engineering, vol. 8, no. 11, pp. 3083–3088, Sep. 2019, doi: 10.35940/IJITEE.K2435.0981119.

[17] A. Qazi, N. Hasan, C. M. Owusu-Ansah, G. Hardaker, S. K. Dey, and K. Haruna, "SentiTAM: Sentiments centered integrated framework for mobile learning adaptability in higher education," Heliyon, vol. 9, no. 1, p. e12705, Jan. 2023, doi: 10.1016/J.HELIYON.2022.E12705.

[18] A. Morgan, R. Sibson, and D. Jackson, "Digital demand and digital deficit: conceptualising digital literacy and gauging proficiency among higher education students," Journal of Higher Education Policy and Management, vol. 44, no. 3, pp. 258–275, 2022, doi: 10.1080/1360080X.2022.2030275.

[19] J. C. Bricout, "Making computer-mediated education responsive to the accommodation needs of students with disabilities," J Soc Work Educ, vol. 37, no. 2, pp. 267–281, 2001, doi: 10.1080/10437797.2001.10779053.

[20] A. J. Martin, R. J. Collie, and R. P. Nagy, "Adaptability and High School Students' Online Learning During COVID-19: A Job Demands-Resources Perspective," Front Psychol, vol. 12, p. 3181, Aug. 2021, doi: 10.3389/FPSYG.2021.702163/BIBTEX.

[21] M. M. H. Suzan, N. A. Samrin, A. A. Biswas, and M. A. Pramanik, "Students' Adaptability Level Prediction in Online Education using Machine Learning Approaches," 2021 12th International Conference on Computing Communication and Networking Technologies, ICCCNT 2021, 2021, doi: 10.1109/ICCCNT51525.2021.9579741.

[22] J. König, D. J. Jäger-Biela, and N. Glutsch, "Adapting to online teaching during COVID-19 school closure: teacher education and teacher competence effects among early career teachers in Germany," https://doi.org/10.1080/02619768.2020.1809650, vol. 43, no. 4, pp. 608–622, Aug. 2020, doi: 10.1080/02619768.2020.1809650.

[23] M. Ezz and A. Elshenawy, "Adaptive recommendation system using machine learning algorithms for predicting student's best academic program," Educ Inf Technol (Dordr), vol. 25, no. 4, pp. 2733–2746, Jul. 2020, doi: 10.1007/S10639-019-10049-7/METRICS.

[24] X. Ma, Y. Yang, and Z. Zhou, "Using machine learning algorithm to predict student pass rates in online education," ACM International Conference Proceeding Series, pp. 156–161, Apr. 2018, doi: 10.1145/3220162.3220188.

[25] R. G. Tiwari, A. Misra, V. Kukreja, A. K. Jain, and N. Ujjwal, "Education 4.0: Classification of Student Adaptability Level in E-Education," 2022 10th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions), ICRITO 2022, 2022, doi: 10.1109/ICRITO56286.2022.9964851.

[26] C. Xie et al., "Influence of Artificial Intelligence in Education on Adolescents’ Social Adaptability: A Machine Learning Study," International Journal of Environmental Research and Public Health

2022, Vol. 19, Page 7890, vol. 19, no. 13, p. 7890, Jun. 2022, doi: 10.3390/IJERPH19137890.

[27] R. D. Costa, G. F. Souza, R. A. M. Valentim, and T. B. Castro, "The theory of learning styles applied to distance learning," Cogn Syst Res, vol. 64, pp. 134–145, Dec. 2020, doi: 10.1016/J.COGSYS.2020.08.004.

[28] H. Waheed, S. U. Hassan, R. Nawaz, N. R. Aljohani, G. Chen, and D. Gasevic, "Early prediction of learners at risk in self-paced education: A neural network approach," Expert Syst Appl, vol. 213, p. 118868, Mar. 2023, doi: 10.1016/J.ESWA.2022.118868.

[29] P. Palimkar, R. N. Shaw, and A. Ghosh, "Machine Learning Technique to Prognosis Diabetes Disease: Random Forest Classifier Approach," Lecture Notes in Networks and Systems, vol. 218, pp. 219–244, 2022, doi: 10.1007/978-981-16-2164-2_19/COVER.

[30] O. Iparraguirre-Villanueva, A. Epifanía-Huerta, C. Torres-Ceclén, J. Ruiz-Alvarado, M. Cabanillas-Carbonell, and N. Wiener, "Breast Cancer Prediction using Machine Learning Models," IJACSA) International Journal of Advanced Computer Science and Applications, vol. 14, no. 2, p. 2023, Accessed: Mar. 14, 2023. [Online]. Available: www.ijacsa.thesai.org

[31] X. Zhang, H. Xiao, R. Gao, H. Zhang, and Y. Wang, "K-nearest neighbors rule combining prototype selection and local feature weighting for classification," Knowl Based Syst, vol. 243, p. 108451, May 2022, doi: 10.1016/J.KNOSYS.2022.108451.

[32] L. Sun, J. Zhang, W. Ding, and J. Xu, "Feature reduction for imbalanced data classification using similarity-based feature clustering with adaptive weighted K-nearest neighbors," Inf Sci (N Y), vol. 593, pp. 591–613, May 2022, doi: 10.1016/J.INS.2022.02.004.

[33] M. Tanveer, T. Rajani, R. Rastogi, Y. H. Shao, and M. A. Ganaie, "Comprehensive review on twin support vector machines," Ann Oper Res, pp. 1–46, Mar. 2022, doi: 10.1007/S10479-022-04575-W/METRICS.

[34] A. B. YILMAZ, Y. S. TASPINAR, and M. Koklu, "Classification of Malicious Android Applications Using Naive Bayes and Support Vector Machine Algorithms," International Journal of Intelligent Systems and Applications in Engineering, vol. 10, no. 2, pp. 269–274, May 2022, Accessed: Mar. 14, 2023. [Online]. Available: https://www.ijisae.org/index.php/IJISAE/article/view/2010

[35] Joseph Redshaw, D. S. J. Ting, Alex Brown, J. D. Hirst, and Thomas Gärtner, "Krein support vector machine classification of antimicrobial peptides," Digital Discovery, 2023, doi: 10.1039/D3DD00004D.

[36] Y. Hu, Y. Fan, Y. Song, and M. Li, "A general robust low–rank multinomial logistic regression for corrupted matrix data classification," Applied Intelligence, pp. 1–17, Feb. 2023, doi: 10.1007/S10489-022-04424-0/METRICS.

[37] M. Zivkovic et al., "Training Logistic Regression Model by Hybridized Multi-verse Optimizer for Spam Email Classification," pp. 507–520, 2023, doi: 10.1007/978-981-19-6634-7_35.

[38] S. R. Darawsheh, A. S. Al-Shaar, F. A. Haziemeh, and M. T. Alshurideh, "Classification Thyroid Disease Using Multinomial Logistic Regressions (LR)," pp. 645–659, 2023, doi: 10.1007/978-3-031-12382-5_34.

[39] H. Nguyen and N. D. Hoang, "Computer vision-based classification of concrete spall severity using metaheuristic-optimized Extreme Gradient Boosting Machine and Deep Convolutional Neural Network," Autom Constr, vol. 140, p. 104371, Aug. 2022, doi: 10.1016/J.AUTCON.2022.104371.

[40] T. Thenmozhi and R. Helen, "Feature Selection Using Extreme Gradient Boosting Bayesian Optimization to upgrade the Classification Performance of Motor Imagery signals for BCI," J Neurosci Methods, vol. 366, p. 109425, Jan. 2022, doi: 10.1016/J.JNEUMETH.2021.109425.

[41] A. Ramón, A. M. Torres, J. Milara, J. Cascón, P. Blasco, and J. Mateo, "eXtreme Gradient Boosting-based method to classify patients with COVID-19," Journal of Investigative Medicine, vol. 70, no. 7, pp. 1472–1480, Oct. 2022, doi: 10.1136/JIM-2021-002278.