JAIEP

Predicting Student Adaptability in Online Education: A Comparative Study of Machine Learning Models and Copula-Based Analysis

Sekar Kidambi Raju [1] *, **Marwa M. Eid**[2]
[1]School of Computing, SASTRA Deemed University, Thanjavur 613401, India
[2]Faculty of Artificial Intelligence, Delta University for Science and Technology, Mansoura 11152, Egypt

*Corresponding author: sekar1971kr@gmail.com
Emails: sekar1971kr@gmail.com, mmm@ieee.org

Abstract

The rapid shift to online education has underscored the need to understand and predict students' adaptability levels to ensure effective learning outcomes. This study aims to classify students' adaptability in online education using a range of machine learning models, including Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, K-Nearest Neighbors, Naive Bayes, Neural Network (MLP), and Gradient Boosting. The analysis is based on a dataset from Kaggle that includes features such as demographic information, educational background, and technological access. In addition to traditional machine learning approaches, the Copula method was applied to explore dependencies among features, enhancing the interpretability of the models' predictions. The models were evaluated using several performance metrics, including Accuracy, Sensitivity, Specificity, Precision, Negative Predictive Value, and F-Score. Logistic Regression emerged as the most effective model, achieving an accuracy score of 99%, demonstrating superior performance across multiple metrics. These findings offer valuable insights for educators and policymakers, highlighting the potential of machine learning models, complemented by Copula-based analysis, to enhance our understanding of student adaptability and guide the development of targeted interventions in online education.
Keywords: Student Adaptability, Online Education, Machine Learning, Copula Method, Classification Models
MSC: 68T20; 68T07; 68T20

Doi: https://doi.org/10.21608/jaiep.2024.318323.1008
Received: September 4, 2024 Revised: October 20, 2024 Accepted: November 27, 2024

1 Introduction

This paper aims to discuss the role technological advancements have played in the evolution of online education as a means of learning. Given that educational institutions around the globe are shifting to online platforms, it is pivotal to understand how students accept these changes. Flexibility, which can be described as a capacity to thrive in new conditions, is another key factor that defines the outcome of online education. The operationalization of adaptation into the low, average, and high groups shows that the combination of hypothetical constructs affects the students' adaptability differently [1–3].

Sorting information, a crucial feature in machine learning and data analysis, involves grouping information based on predetermined factors. Using classification models, it is possible to define different degrees of students' learning adaptability within the framework of online education. Potential mediators of adaptability could include students' technology skills, levels of self-control, learning styles, and backgrounds in taking online courses. Using classification algorithms, students can be classified into such types as 'high,' 'medium,' and 'low' adaptability levels regarding learning in online courses, and the correlation between these categorized groups and their performance can be analyzed [4–6].

Thus, this paper aims to review and discuss the possibilities of using classification approaches to evaluate students' adaptability to online classes. Classifications used in this study include decision trees, support vector machines, and neural network classifications. The dataset to be analyzed comprises variables that encompass the technological aptitude of students, their learning approaches or habits, and their perceptions towards online learning. The first objective is the construction of a comprehensive classification model that will correctly evaluate the students' degree of adaptability. An added goal entails determining factors concerning the differences between adaptable students and those having difficulties with the online learning environment [7–9].

In the same regard, the ideas researched in this study are crucial to educators and other institutions of learning. The present research identifies a set of predictors that can indicate students' levels of adaptation and help institutions suggest adequate support for improving online learning environments. For example, suppose it is found that students with low technological savvy are less versatile. In that case, IT-based training sessions or accessible methodologies to alter the learning platform can be incorporated to assist such learners. Additionally, such a classification model can help design individualized learning accomplishments focused on students' adaptability, with interventions designed for similar profiles [10–12].

Based on the findings on the applicability of classification techniques in identifying the student adaptability levels in online education, it is apparent that separate factors influencing student success in virtual learning environments can also be determined. In this regard, the purpose and scope of this research are to uncover these levels of adaptability and lay the ground for enhancing online education strategies and corresponding support structures. The proposed agenda focuses on improving the educational process for students from different groups to facilitate their ability to achieve educational objectives set by them in the context of the growing use of online learning environments.

2 Related Works

Identifying levels of students' adaptability within the framework of distance learning is a crucial issue and one of the directions of studies, especially with the constant advancement of technologies and approaches in distance learning. Regarding research objectives, understanding students' behavior and learning progress online enables educational institutions to optimize students' learning outcomes. Several classification methods have been used to study and forecast students' possibilities of coping with online learning environments.

Logistic regression and decision tree classifiers, which are conventional machine learning algorithms, have been applied to classify students according to their ease with online learning. The significant advantage of logistic regression models is their simplicity and interpretability; thus, the impact of different factors on adaptability levels can be investigated [13]. For instance, in studying the students' success or failure rates within online classes, logistic regression can estimate relevant factors such as past academic records, IT literacy, and activity levels.

Decision trees, in turn, are more detailed as they illustrate the hierarchy of the factors determining the students' flexibility [14]. Scholars have used algorithms such as decision trees to ascertain the factors that have a major influence on adaptability. These factors include self-regulation skills, time management, and previous experience with technology. The outcomes of decision trees assist educators in understanding which factors are most influential for efficient online learning.

In this regard, new and more sophisticated classification techniques have been adopted to address the aforementioned disadvantages. Random forests, an extension of decision trees that integrate results from a large

Doi: https://doi.org/10.21608/jaiep.2024.318323.1008
Received: September 4, 2024 Revised: October 20, 2024 Accepted: November 27, 2024

number of trees, are effective in increasing classification performance and model stability even when the relationships between predictors and the outcome variable cannot be accurately captured by a linear model [15]. For instance, decision trees in random forests have been applied to sort students into different levels of adaptability based on parameters such as preferred learning mode, activity level, and demographics.
Support Vector Machines (SVMs) have also been used to classify students' adaptability levels, mainly when classifications are carried out in high dimensions where linearity is not achievable [16]. SVMs can model higher-order non-linear polynomial interactions between variables, allowing them to classify students successfully based on academic history, behavioral data, and self-reports of adaptability.
Artificial neural networks, especially deep neural networks, have increasingly attracted interest due to their capacity to automatically learn high-level representations from extensive data [17]. These models can process raw data to identify pertinent features and interactions, such as online learning interactions and engagement levels, to forecast students' adaptability. For instance, clickstream data and forum contributions have been used as input features in deep learning models to categorize learners according to their flexibility and activity.
Ensemble learning, which involves combining several models' forecasts, has been researched to improve classification accuracy and reliability. Techniques such as boosting and bagging combine individual classifiers' decisions to enhance performance. For example, boosting algorithms like AdaBoost have been employed to improve base classifiers' distinctions in predicting students' adaptability levels by focusing on complex cases [18].
Combining machine learning techniques has also been explored to enhance traditional classification methods. For instance, hybrid models that combine logistic regression with decision trees or SVMs leverage the advantages of various techniques for better classifications [19]. Such models are preferable for modeling students' adaptability because they incorporate different data types and analytical methods.
Factor analysis and feature selection are critical for enhancing classification models by identifying key indicators of adaptability [20]. Techniques such as Recursive Feature Elimination (RFE) and Lasso regression are used to select the most relevant features and remove less significant ones [21]. Additionally, incorporating contextual factors like socio-economic status, resource availability, and institutional support improves the models' accuracy by accounting for external influences [22].
By integrating these approaches, classification models can more effectively diagnose adaptability challenges and suggest targeted interventions to support students in virtual learning environments.

3 Dataset Description

The dataset applied in this study is sourced from Kaggle and is titled *"Students Adaptability Level in Online Education."* This dataset contains various characteristics that can help explain the factors affecting students' adaptability to online learning.
The given dataset includes several features that provide information about demographic, educational, and technological aspects influencing students' adaptability. These features include:

- **Gender:** This feature indicates the gender of the student; this may influence adaptability, as male and female students may learn differently or encounter specific challenges.
- **Age:** This feature categorizes students by their age, allowing for the examination of flexibility across different stages in a student's life.
- **Education Level:** This feature identifies the current course that the student is undertaking, such as undergraduate or postgraduate, which may affect their flexibility in online classes.
- **Institution Type:** This feature aims to identify whether the student is in a public or private educational institution, as the extent to which online education is supported may vary between institutions.
- **IT Student:** This binary feature indicates whether the student is studying an IT-related course, which may influence their ease of use of online resources.
- **Location:** This feature identifies whether the student lives in a town, which can influence their access to educational materials and internet availability.
- **Load-shedding:** This feature captures the intensity of load-shedding regimes that students may experience, affecting their ability to attend online classes regularly.
- **Financial Condition:** This feature relates to the student's family financial status, which could impact the availability of the required technology and a stable internet connection needed for online learning.

- **Internet Type:** This feature specifies the type of internet connection the student typically uses, which is crucial for the stability and quality of virtual classes.
- **Network Type:** This feature refers to the type of network connectivity available to the student, such as Broadband or Mobile Data, thus directly affecting the level or extent of their use of online education platforms.

The adaptivity of IT students to the online learning environment is described in a pie chart presented in Figure 1. This picture divides students according to the adaptability level; hence, it explains which group of students specializing in Information Technology is doing well to meet the demands and challenges posed by online schooling. The chart shows the distribution of adaptability in this particular group, which is essential owing to the technical subject of their study and the level of difficulty, or otherwise, they are likely to encounter in a digital learning environment.

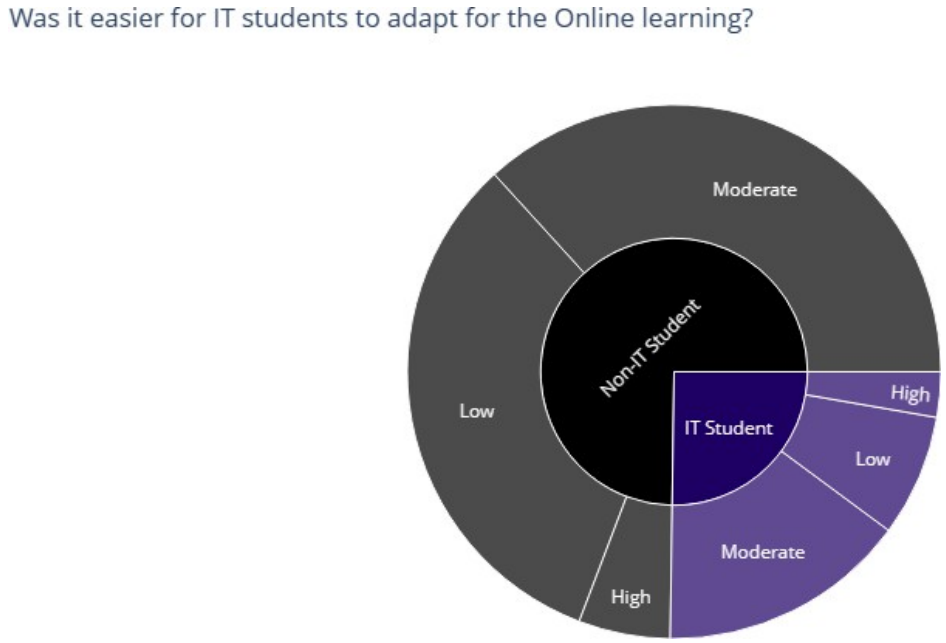

Figure 1: Pie Chart for IT Student Adaptivity to Online Learning

Figure 2 illustrates a histogram of the tested student's adaptability percentage in a given data set. This statistic shows the distribution of students' adaptability levels; it provides practical information about the distribution of the student populace in accordance with their comfort and efficiency when dealing with online classes. This is because the histogram is useful in narrowing down the possible adaptability trends and any possible skewness in the data that would require intervention.
Combined, these features offer a wide range of information concerning various elements that could potentially impact a student's readiness for an online learning environment. Therefore, they are appropriate for use in classification and prediction models as predictors and inputs.

4 Methodology

4.1 Data Preprocessing

Data preprocessing is one of the initial steps in machine learning that involves preparing raw data for use in machine learning models. The following steps were taken to preprocess the data in this study:

Handling Missing Data Missing data can pose a significant threat to the accuracy and reliability of machine learning models if left unaddressed. In our dataset, missing values were identified, and various strategies were employed to address them:

- **Mean/Median Imputation:** For numerical features with missing observations, mean or median imputation was applied. This method helps maintain the integrity of the dataset while preserving the overall distribution of the data.

Doi: https://doi.org/10.21608/jaiep.2024.318323.1008
Received: September 4, 2024 Revised: October 20, 2024 Accepted: November 27, 2024

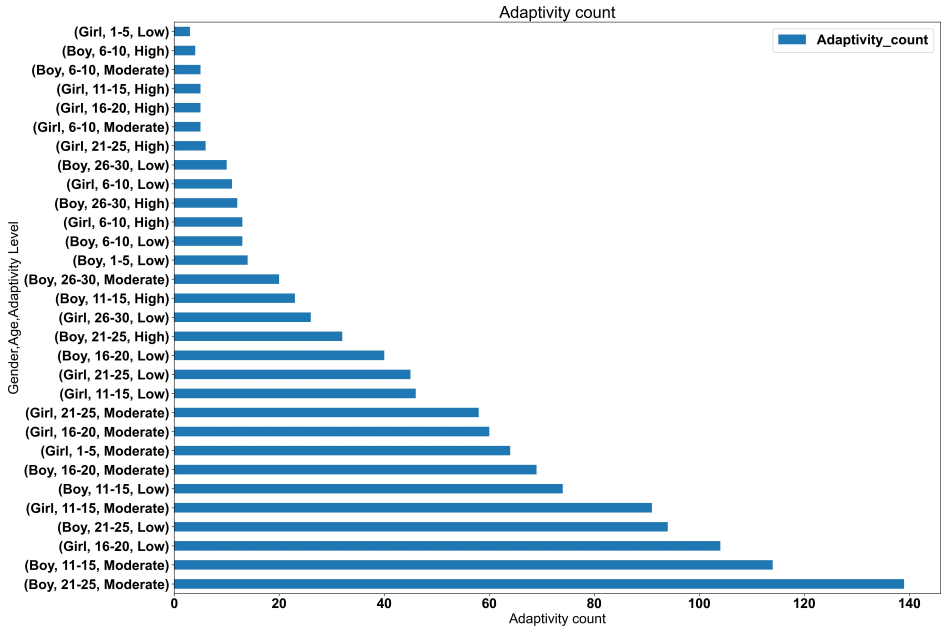

Figure 2: Histogram of the Student Adaptivity

- **Mode Imputation for Categorical Data:** For categorical features, missing values were imputed using the mode, which is the most frequent category. This ensures that the categorical variables retain their sample proportions within the dataset.

- **Advanced Techniques:** For features with a large percentage of missing values or complex missing data patterns, more advanced imputation techniques, such as k-Nearest Neighbors (k-NN) imputation, were employed to provide a more accurate fill for the missing values.

By carefully addressing missing data, we ensured that no biases were introduced into the dataset, allowing the machine learning models to perform effectively with complete data.

Encoding Categorical Variables Many machine learning algorithms require numerical data, so categorical variables were converted into numerical representations:

- **Label Encoding:** For ordinal categorical variables, such as Education Level, label encoding was used. This method maps categories to integer values while preserving the order of the categories.

- **One-Hot Encoding:** For nominal categorical variables, such as Gender and Institution Type, one-hot encoding was applied. This technique creates binary columns for each category, ensuring no hierarchy is implied within the data.

Proper encoding of categorical features ensures that machine learning models can effectively interpret and utilize these variables, leading to improved model performance and predictive accuracy.

Scaling Numerical Features Numerical features in the dataset often have different scales or ranges, which can impact the performance of machine learning models:

- **Min-Max Scaling:** This method scales features to a range of [0, 1], preserving the differences in values while reducing the scale of the numbers.

- **Standardization:** For features with varying means and standard deviations, standardization was applied. This method transforms features into a mean of 0 and a standard deviation of 1, which is particularly useful for algorithms that assume normally distributed data.

Normalizing numerical features ensures that no single feature dominates the learning process, leading to more balanced model training and improved temporal reliability by reducing inconsistencies and errors.

4.2 Copula Approach

To complement the analysis, this study adopted the Copula method to present the degrees of interdependence among the features in the dataset. This method is particularly beneficial for capturing relationships that cannot be easily explained by traditional correlation measures, especially when dealing with non-linear structures.
The Copula approach assesses relationships between multiple random variables without assuming a specific probability distribution. While most correlation measures rely on linear correlation coefficients, Copulas offer greater flexibility by capturing all forms of correlation, including tail correlation, which refers to the relationship between two variables' extreme values (upper or lower tails).

Role in Understanding Dependencies The Copula approach enabled a detailed examination and comparison of various dependencies, such as the influence of financial condition on internet type or the impact of load-shedding on online learning about the student's location. It was possible to model these interactions accurately by resolving such dependencies during the analysis.
Incorporating Copulas significantly enhanced the modeling of dependence patterns among the variables, leading to more comprehensible machine learning models. This approach facilitated a deeper understanding of factors crucial for classification outcomes, allowing the identification of issues related to students' adaptability that are central to developing more accurate classification models.
In Figure 3 we can find the correlation heatmap resulting from the Copula analysis, based on the data set described herein. This kind of heatmap focuses on the pairwise combinations of variables in a dataset, which enables analysts to decipher the strength of connections between these features. Specifically, being used in generating these correlations, Copula is useful in capturing these non-linear, or hidden, dependencies that might not be captured by more standard correlation approaches, providing a deeper understanding of the factors that determine students' adaptability.

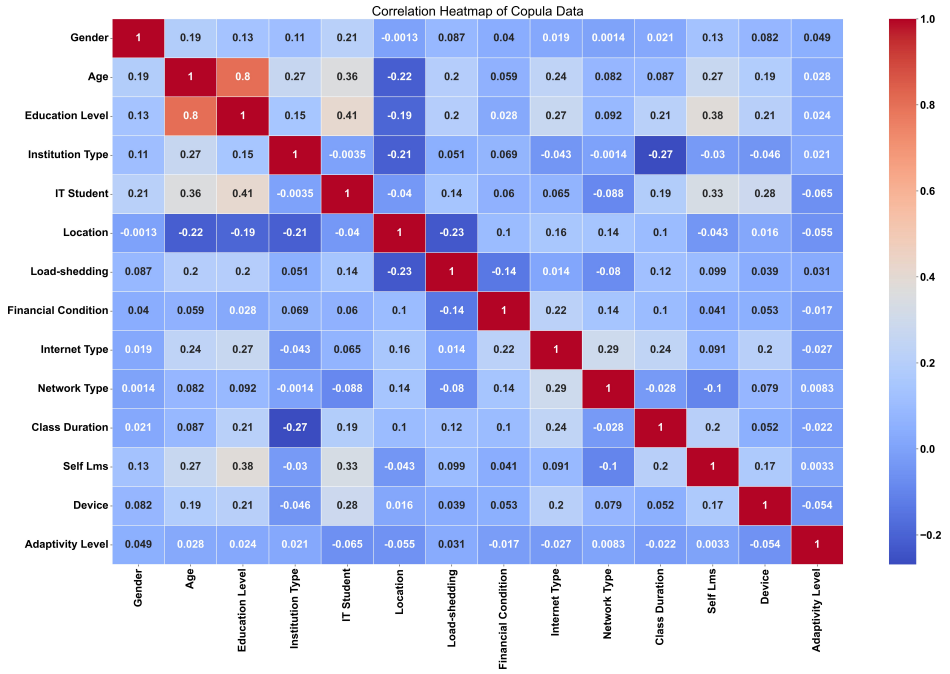

Figure 3: Correlation Heatmap of the Copula Result

Figure 4 shows the pair plot of this study based on Copula analysis that describes the relationship between various features in the large dataset of students. This plot gives a three-dimensional look of the data, demonstrating the strength of the relationship of variables in terms of pairs. The pair plot is highly revealing when it comes to viewing graphical patterns, trends, as well as outliers in the data, which in turn allows for developing even clearer perception of feature dependencies that define the flexibility of students in the conditions of online learning and teaching.
Thus, we established a solid and stable foundation for developing accurate, reliable, and interpretable NLP machine-learning models by performing thorough data preprocessing and incorporating the Copula method in our analysis. These steps were instrumental in building models capable of identifying students' adaptability levels in online education, providing valuable insights to practitioners and policymakers in the field of

Doi: https://doi.org/10.21608/jaiep.2024.318323.1008
Received: September 4, 2024 Revised: October 20, 2024 Accepted: November 27, 2024

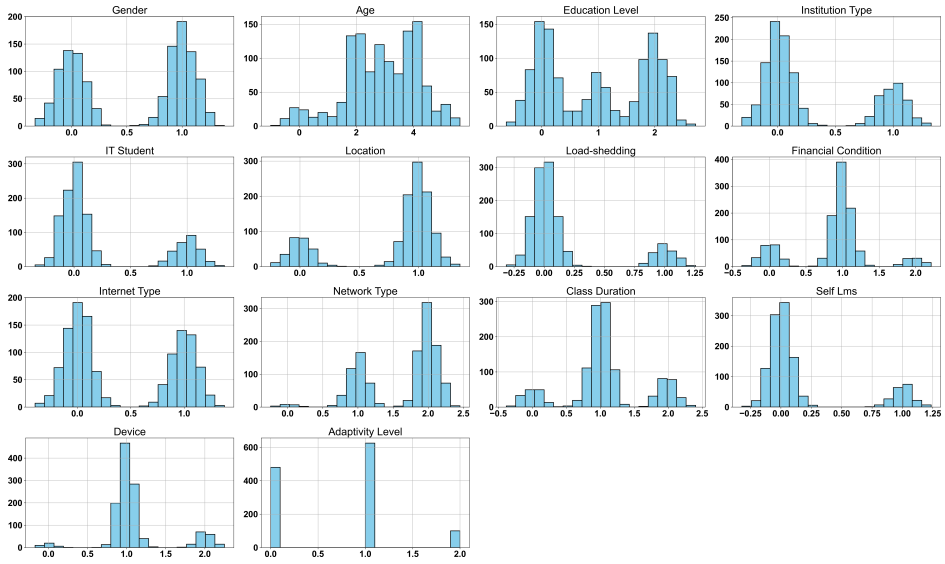

Figure 4: Pair Plot of the Copula Result

education.

5 Results

The current capability of each machine learning model was assessed based on the metrics discussed earlier. Table 1 below presents the Accuracy, Sensitivity (True Positive Rate), Specificity (True Negative Rate), Precision (Positive Predictive Value), Negative Predictive Value, and F-score for each model.

Table 1: Model Performance Metrics

Model	Accuracy	Sensitivity (TRP)	Specificity (TNP)	PPV (P-value)	NPV (P-value)	F-Score
Logistic Regression	0.99	0.9968	0.9667	0.9904	0.9886	0.9936
Decision Tree	0.9875	0.9935	0.9667	0.9904	0.9775	0.9919
Random Forest	0.9675	0.9903	0.8889	0.9685	0.9639	0.9793
Support Vector Machine	0.9575	0.9774	0.8889	0.9681	0.9195	0.9727
K-Nearest Neighbors	0.955	0.9871	0.8444	0.9563	0.9500	0.9714
Naive Bayes	0.9525	0.9903	0.8222	0.9505	0.9610	0.9700
Neural Network (MLP)	0.9275	0.9710	0.7778	0.9377	0.8861	0.9540
Gradient Boosting	0.845	0.9871	0.3556	0.8407	0.8889	0.9080

Figure 5 depicts a scatter plot that compares different machine learning models in terms of classification accuracy obtained in this research. This plot shows a comparison of all accuracies that have been obtained from the different models, providing a visual perception of the models' ability to give accurate results in assessing students' adaptability.

The classification accuracy results are also presented in Figure 6 in the form of a box plot to give an overall view of the dispersion of the accuracy scores for all the explored machine learning models. The box plot shows the median accuracy and the distribution range, along with information about quartiles and potential outliers, providing an overview of performance variation.

5.1 Performance Analysis

The logistic regression model achieved the highest accuracy score of 99%, demonstrating consistently high performance across all measured aspects. Therefore, it is considered the most efficient model in this study. The Decision Tree and Random Forest models also produced strong results across all parameters but with slightly lower accuracy than Logistic Regression.

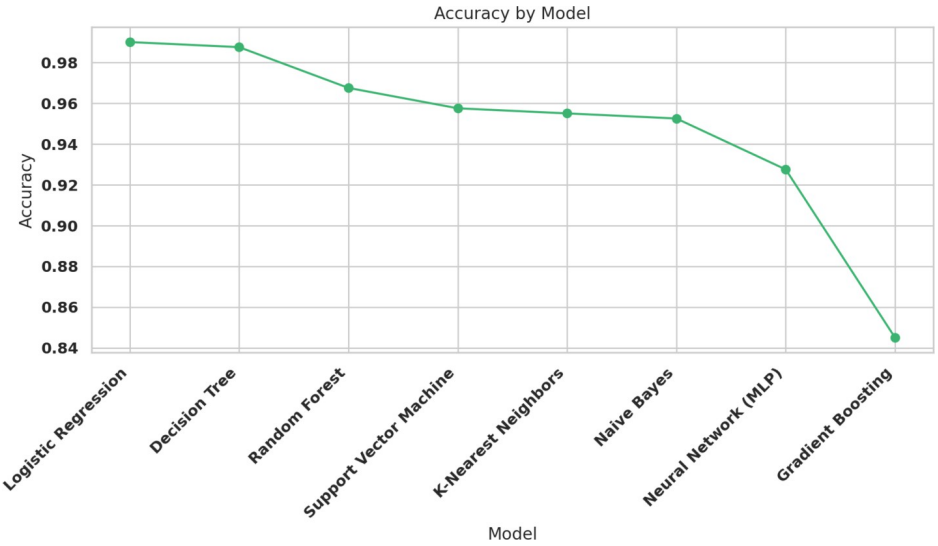

Figure 5: Scatter Plot of the Classification Accuracy Results

5.2 Comparison and Analysis

Best Performing Models: Logistic Regression delivered the best performance, with an accuracy of 99%, a sensitivity of 0.9968, and a specificity of 0.9667. The combination of high sensitivity and specificity, along with uniformly strong results in other metrics, demonstrates this model's stability and predictive power in estimating students' success in online learning environments. The Decision Tree and Random Forest models also showed moderate-to-high accuracy and good performance, particularly in sensitivity (True Positive Rate), making them useful depending on the application requirements.

Upon comparing and analyzing the model's performance and accuracy, it is evident that while Logistic Regression is the best among the models, Decision Tree and Random Forest models also perform well and have their own unique strengths. The specificity and sensitivity metrics vary across the models, underscoring that different applications may require models that prioritize balance or the trade-off between these metrics. For instance, when aiming to accurately identify adaptable students or minimize the rate of false positives, the choice of model may vary depending on the specific classification problem.

6 Conclusion

The classification of students' adaptability levels in the online education system was addressed using the machine learning models of Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, K-Nearest Neighbor, Naive Bayes, Neural Network (MLP), and Gradient Boosting. Regarding the accuracy of the classification, Logistic Regression produced the highest accuracy of 99% and satisfying sensitivity and specificity values. The authors of the study also emphasized the necessity of the Copula method for revealing the dependencies between features and boosting the interpretability of the model.

Based on all the research, it was discovered that machine learning models, specifically Logistic Regression, can be useful in identifying students at risk and, therefore, can be used to provide early interventions for students in online classes depending on issues like financial status and internet connectivity. Policymakers and educators are urged to embrace these approaches to enhance learners' achievements.

Future directions for the study include collecting data for a more significant number of students to detect the influence of different features on adaptability, examining other aspects that affect this factor, and utilizing new algorithms to enhance prediction and contribute to formulating strategies concerning online learning.

References

[1] A. F. Agudo-Peregrina, S. Iglesias-Pradas, M. A. Conde-González, and A. Hernández-García. Can we predict success from log data in vles? classification of interactions for learning analytics and their relation with performance in vle-supported f2f and online learning. *Computers in Human Behavior*, 31:542–550, 2014.

Doi: https://doi.org/10.21608/jaiep.2024.318323.1008
Received: September 4, 2024 Revised: October 20, 2024 Accepted: November 27, 2024

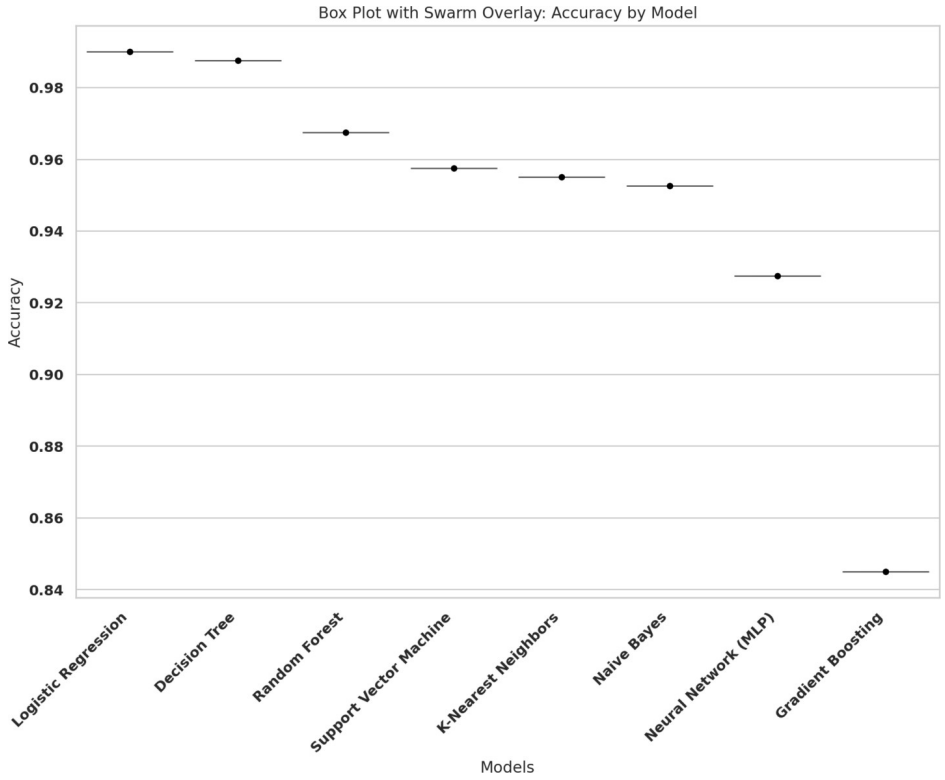

Figure 6: Box Plot of the Classification Accuracy Results

[2] Ş. Aydoğdu. Predicting student final performance using artificial neural networks in online learning environments. *Education and Information Technologies*, 25(3):1913–1927, 2020.

[3] I. Azzi, A. Jeghal, A. Radouane, A. Yahyaouy, and H. Tairi. A robust classification to predict learning styles in adaptive e-learning systems. *Education and Information Technologies*, 25(1):437–448, 2020.

[4] K. Coussement, M. Phan, A. De Caigny, D. F. Benoit, and A. Raes. Predicting student dropout in subscription-based online learning environments: The beneficial impact of the logit leaf model. *Decision Support Systems*, 135:113325, 2020.

[5] G. Deeva, D. Bogdanova, E. Serral, M. Snoeck, and J. De Weerdt. A review of automated feedback systems for learners: Classification framework, challenges and opportunities. *Computers & Education*, 162:104094, 2021.

[6] L. E. C. Delnoij, K. J. H. Dirkx, J. P. W. Janssen, and R. L. Martens. Predicting and resolving non-completion in higher (online) education – a literature review. *Educational Research Review*, 29:100313, 2020.

[7] M. E. Dogan, T. Goru Dogan, and A. Bozkurt. The use of artificial intelligence (ai) in online learning and distance education processes: A systematic review of empirical studies. *Applied Sciences*, 13(5), 2023.

[8] O. El Aissaoui, Y. El Alami El Madani, L. Oughdir, and Y. El Allioui. A fuzzy classification approach for learning style prediction based on web mining technique in e-learning environments. *Education and Information Technologies*, 24(3):1943–1959, 2019.

[9] P. Gao, J. Li, and S. Liu. An introduction to key technology in artificial intelligence and big data-driven e-learning and e-education. *Mobile Networks and Applications*, 26(5):2123–2126, 2021.

[10] H. Hayati, M. Khalidi Idrissi, and S. Bennani. Automatic classification for cognitive engagement in online discussion forums: Text mining and machine learning approach. In I. I. Bittencourt, M. Cukurova, K. Muldner, R. Luckin, and E. Millán, editors, *Artificial Intelligence in Education*, pages 114–118. Springer International Publishing, 2020.

[11] M.-A. Kaufhold, M. Bayer, and C. Reuter. Rapid relevance classification of social media posts in disasters and emergencies: A system and evaluation featuring active, incremental and online learning. *Information Processing & Management*, 57(1):102132, 2020.

[12] Ö. N. Kenger and E. Ozceylan. A hybrid approach based on mathematical modelling and improved online learning algorithm for data classification. *Expert Systems with Applications*, 218:119607, 2023.

[13] A. Khamparia and B. Pandey. Association of learning styles with different e-learning problems: A systematic review and classification. *Education and Information Technologies*, 25(2):1303–1331, 2020.

[14] A. Khan and S. K. Ghosh. Student performance analysis and prediction in classroom learning: A review of educational data mining studies. *Education and Information Technologies*, 26(1):205–240, 2021.

[15] Y. Kirsal Ever and K. Dimililer. The effectiveness of a new classification system in higher education as a new e-learning tool. *Quality & Quantity*, 52(1):573–582, 2018.

[16] R. Lamb, K. Neumann, and K. A. Linder. Real-time prediction of science student learning outcomes using machine learning classification of hemodynamics during virtual reality and online learning sessions. *Computers and Education: Artificial Intelligence*, 3:100078, 2022.

[17] N. Mustafee and K. Katsaliaki. Classification of the existing knowledge base of or/ms research and practice (1990–2019) using a proposed classification scheme. *Computers & Operations Research*, 118:104920, 2020.

[18] F. Ouyang, L. Zheng, and P. Jiao. Artificial intelligence in online higher education: A systematic review of empirical research from 2011 to 2020. *Education and Information Technologies*, 27(6):7893–7925, 2022.

[19] Z. Quan and L. Pu. An improved accurate classification method for online education resources based on support vector machine (svm): Algorithm and experiment. *Education and Information Technologies*, 28(7):8097–8111, 2023.

[20] A. V. Savchenko, L. V. Savchenko, and I. Makarov. Classifying emotions and engagement in online learning based on a single facial expression recognition neural network. *IEEE Transactions on Affective Computing*, 13(4):2132–2143, 2022.

[21] L. Yu, X. Wu, and Y. Yang. An online education data classification model based on tr madaboost algorithm. *Chinese Journal of Electronics*, 28(1):21–28, 2019.

[22] H. Zhang, T. Huang, S. Liu, H. Yin, J. Li, H. Yang, and Y. Xia. A learning style classification approach based on deep belief network for large-scale online education. *Journal of Cloud Computing*, 9(1):26, 2020.

Doi: https://doi.org/10.21608/jaiep.2024.318323.1008
Received: September 4, 2024 Revised: October 20, 2024 Accepted: November 27, 2024