Machine Learning for Predicting Student Engagement and Adaptability in Online Courses

Deepa.P. S, Dr. Mukesh Kumar

[1]Research Scholar (Computer Science), School of Engineering and Technology, Shri Venkateshwara University, Gajraula, UP, India.
[2]Research Guide (Computer Science), School of Engineering and Technology, Shri Venkateshwara University, Gajraula, UP, India.

Corresponding Email: deepasurendran@gmail.com

ABSTRACT
The abstract summarizes the key findings and methodology of the study conducted on various machine learning classifiers for predictive analysis on a specific dataset. Through comprehensive exploration, classifiers like Decision Tree, Random Forest, and XGBoost emerged as high performers, achieving an average accuracy of around 90% with minimal misclassifications. Conversely, classifiers such as Logistic Regression, Gaussian Naive Bayes, and Multi-layer Perceptron demonstrated lower accuracies, indicating their limited suitability for the dataset. Additionally, hyperparameter tuning of the Decision Tree model using a grid search method led to a significant improvement in accuracy to approximately 90.87%. These findings underscore the critical role of thoughtful model selection, evaluation, and optimization in developing accurate and reliable machine learning models for real-world applications.

Keywords: Artificial Intelligence, Online Education, Machine Learning, Personalized Learning, Digital Divide, Educational Technology.

1. INTRODUCTION

Students' educational attainment and critical thinking skills are profoundly influenced by the quality of education they receive. In today's globalized and privatized world, integrating Information and Communication Technology (ICT) into teaching methodologies is imperative for educational enhancement [1]. E-Learning has emerged as a pivotal tool, particularly accentuated during the recent pandemic, facilitating the dissemination of information, concepts, and skills across diverse locations, thereby revolutionizing education [2]. This instructional paradigm shift encompasses various online delivery methods, leveraging ICT resources like computer-based learning, teleconferencing, video conferencing, emails, live chats, blogs, and social media platforms [3]. Technological advancements have broadened the scope of education beyond traditional confines, ushering in an era of Technology-Enhanced Learning (TEL) [4].

The COVID-19 pandemic acted as a catalyst for the widespread adoption of online education in India, with millions turning to digital platforms for learning [4]. This transition has not only diversified instructional approaches but also underscored the pivotal role of digitalization in education. The pandemic-induced lockdowns prompted educational institutions to pivot swiftly to online platforms, increasing parental involvement in basic education delivery [5]. With projections indicating a substantial increase in internet users by 2025 [6], the significance of digital tools in education is poised for further escalation. Online education has revitalized learning across all educational tiers, fostering creativity and engagement through the integration of recent digital tools and technologies. It offers scalability, accessibility, and cost-effectiveness while significantly enhancing learning flexibility and material reusability [7].

However, effective implementation necessitates robust infrastructure and tailored solutions to meet students' evolving digital needs. Governments and educational institutions must invest in comprehensive training programs to equip stakeholders with the requisite skills and adaptability to navigate these changes effectively.

The evolution of online education traces its roots to the 18th century's postal courses, culminating in today's sophisticated digital learning environments [9]. Technological advancements, such as high-speed internet and advanced Learning Management Systems (LMS), have facilitated seamless content delivery and collaboration, transcending geographical barriers [10]. Educators have adapted instructional methodologies to suit digital platforms, emphasizing active learning, collaboration, and social interaction [11]. Strategies like gamification and micro-learning have gained traction, enhancing engagement and outcomes in online learning environments. Furthermore, Massive Open Online Courses (MOOCs) have democratized education, offering access to high-quality content from prestigious institutions [11].

The global accessibility of online education, coupled with immersive learning experiences and learning analytics, has further propelled its prominence [12]. Insights gleaned from online learning platforms empower educators to personalize interventions and curriculum adjustments, ensuring enhanced learning outcomes for diverse student populations. As online education continues to evolve, its potential to revolutionize educational paradigms and cater to the needs of learners worldwide remains unparalleled. Machine learning can play a significant role in testing and evaluating educational software. By harnessing the power of machine learning algorithms, educators and developers can enhance the efficiency, accuracy, and effectiveness of software testing processes. Educational applications are highly precise because of their potential influence on students' education. They are also used for grading and guiding pupils in their studies. So, they need to be put through rigorous testing before being used in the real world. The purpose of software testing is to ensure that the system's qualities and the software's capabilities are in agreement and that the software can achieve its goals as expected. The software testing procedure becomes more laborious as its complexity grows [3].

2. LITERATURE REVIEW

Intelligent learning environments leverage both supervised and unsupervised learning to model students' knowledge, meta-cognitive skills, and behaviors, enabling the anticipation of future actions [13]. By statistically recognizing patterns, data is collected, processed, learned from, and tested. Manual labeling of data facilitates the use of supervised learning techniques to determine activity nature, while unsupervised learning techniques like k-means clustering identify common learning behavior patterns [13]. Integrating both supervised and unsupervised learning minimizes implementation time, offering a powerful and adaptive learning experience [13].

Supervised learning provides personalized recommendations and skill assessments by analyzing learner data. Unsupervised learning aids in clustering learners, identifying knowledge gaps, and organizing learning materials efficiently [13]. Machine learning plays a crucial role in career prediction and planning, utilizing multi-modal, multi-task learning for career forecasting [14]. It leverages various techniques like graph-based learning and multi-view learning to analyze users' social media engagement and interests, aiding in career planning [14].

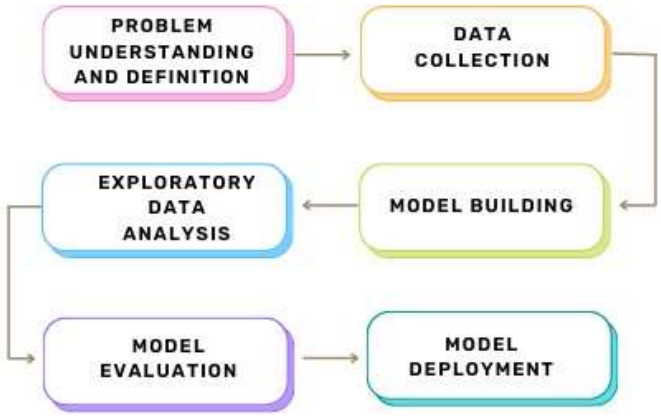

Figure 1. Predictive Analytics Methodology

Automated assessment benefits from machine learning techniques like neural networks, long short-term memory networks, convolutional neural networks, deep reinforcement learning, and natural language processing [15]. These methods enable objective, computer-based evaluation and question generation, enhancing assessment accuracy [16]. Machine learning facilitates predicting learning outcomes, reducing dropout rates by intervening early [17]. By automatically collecting exam scores and related information, machine learning models predict students' performance, considering factors like prior knowledge and aptitude [40].
Agent-based educational applications capitalize on machine learning to offer individualized instruction, enhancing the effectiveness of educational software [18]. Dynamic adaptive learning systems, incorporating agents and learning objects, cater to each student's unique needs, thereby improving learning outcomes [18]. These applications demonstrate the versatility and efficacy of machine learning in enhancing various aspects of education, from personalized learning experiences to automated assessment and career planning [19].

3. MACHINE LEARNING MODEL FOR PREDICTING STUDENT ADAPTABILITY ON ONLINE EDUCATION

The methodology outlined in Figure 1 offers a systematic approach to identifying factors influencing students' adaptability to online learning. Leveraging advanced statistical techniques and machine learning algorithms, this study aims to provide actionable insights for enhancing online learning strategies and supporting students effectively. The dataset used for analysis includes key characteristics such as gender, age, education level, institution type, IT student status, location, and more (Table I)

Dataset Selection Preparation

The dataset taken from Kaggle [20] contains comprises 1013 records with no missing fields, but approximately 30% of rows are duplicates. These duplicates cannot be removed due to potential distinct entries. With 13 independent variables and 1 dependent variable (Adaptivity Level), the dataset is prepared for predictive analysis by converting categorical features into numerical values. This is achieved using encoding techniques like OneHotEncoder and LabelEncoder to ensure the data is suitable for machine learning models.

Table 1. Datasets features

Feature	Description	Example	Mode	Freq-uency
Gender	Gender of the student	Boy, Girl	Boy	663
Age	Age group of the student	11-15, 16-20, 21-25	21-25	374

Education Level	Current educational level of the student	School, College, University	School	530
Institution Type	Type of educational institution	Government, Non-Government	Non-Government	823
IT Student	Whether the student is an IT student	Yes, No	No	901
Location	Whether the student resides in an area with load-shedding	Yes, No	Yes	935
Load-shedding	Frequency of power outages in the student's area	Low, High	Low	1004
Financial Condition	Financial condition of the student's household	Poor, Mid, High	Mid	878
Internet Type	Type of internet connection used by the student	Wifi, Mobile Data	Mobile Data	695
Network Type	Type of network connectivity	3G, 4G, 5G	4G	775
Class Duration	Duration of the student's online classes	0, 1-3, 3-6	1-3	840
Self LMS	Whether the student uses a self-learning management system	Yes, No	No	995
Device	Primary device used for online learning	Mobile, Tab, Laptop	Mobile	1013
Adaptivity Level	Level of adaptability to online learning environments	Low, Moderate, High	Moderate	625

With 1013 records and no missing fields, approximately 30% of rows are duplicates, which cannot be removed due to potential distinct entries. Featuring 13 independent variables and 1 dependent variable (Adaptivity Level), the dataset is prepared for predictive analysis by converting categorical features into numerical values using encoding techniques like OneHotEncoder and LabelEncoder, ensuring suitability for machine learning models.

Exploratory Data Analysis

The exploratory data analysis (EDA) of the dataset uncovers several notable correlations. Age demonstrates a strong positive correlation with Education Level (0.852), indicating older students tend to be in higher educational tiers. Additionally, moderate positive correlations exist between Age and IT Student status (0.505), Education Level and IT Student status (0.523), suggesting older students and those in advanced education levels are more likely to be IT students. Age also moderately correlates with Self LMS use (0.439) and Class Duration (0.320), implying older students favor self-learning systems and longer classes. Notably, Financial Condition correlates with Adaptivity Level (0.311) and Internet Type (0.282), indicating better financial conditions lead to higher adaptability and better internet connectivity. Adaptivity Level and Institution Type (0.295) also have a moderate correlation, suggesting the institution type might influence adaptability. Moreover, Gender exhibits low

correlations with most features, Load-shedding displays weak correlations, and Network Type correlates with Internet Type (0.349) and Device (0.112), indicating relationships with network quality.

Model Selection

The model selection process involves considering various classifiers for implementation in predictive analysis. Logistic Regression (LRC) is chosen for its simplicity and interpretability, utilizing a linear model for binary classification. K-Nearest Neighbors (KNN) is selected for its intuitive nature and lack of distribution assumptions. Gaussian Naive Bayes (GNB) is preferred for its simplicity and speed, despite assuming independence among features. Support Vector Classifier (SVC) is chosen for its effectiveness in high-dimensional spaces and robustness to overfitting. Decision Tree Classifier is opted for its interpretability and minimal data preprocessing requirements, despite susceptibility to overfitting. Random Forest Classifier is selected to reduce overfitting by combining predictions from multiple decision trees. Multi-layer Perceptron (MLP) is chosen for its ability to model complex relationships, despite requiring significant computational power and being prone to overfitting. XGBoost Classifier (XGB) is preferred for its efficiency and high performance, albeit requiring careful tuning to avoid overfitting and being more complex to interpret. Each classifier offers distinct advantages and disadvantages, allowing for informed selection based on the specific requirements of the predictive analysis task.

4. RESULTS AND DISCUSSION

Model Training

For model training, the performance of various machine learning classifiers on a specific dataset is summarized in Table 27. Each classifier, including Logistic Regression (LRC), K-Nearest Neighbors (KNN), Gaussian Naive Bayes (GNB), Support Vector Classifier (SVC), Decision Tree, Random Forest, XGBoost, and Multi-layer Perceptron (MLP), is evaluated based on cross-validation scores, average accuracy, and the number of mislabeled data points. Key observations from the table reveal high-performing models such as Decision Tree, Random Forest, and XGBoost, which achieve around 90% average accuracy with a low number of misclassified points. These models demonstrate effectiveness in learning the data patterns for the given task. Mid-range performers like KNN and SVC exhibit decent average accuracies with a moderate number of misclassifications, making them suitable choices for interpretable or efficient models. However, lower performers like LRC, GNB, and MLP show lower average accuracies and higher misclassification rates, suggesting they may not be the best options for this dataset. Next steps involve further analysis, including visualizing cross-validation scores, considering feature importance, and hyperparameter tuning, with a focus on Decision Tree due to its promising performance and interpretability.

Model Evaluation

The model evaluation process involves analyzing the performance of various machine learning classifiers on a specific dataset, as summarized in Table 2. Each classifier is assessed based on its cross-validation score, average accuracy, and the number of mislabeled data points. The results reveal notable differences in performance among the classifiers. High-performing models like Decision Tree, Random Forest, and XGBoost achieve a very high average accuracy (around 90%) with a low number of misclassified points, indicating their effectiveness in learning the data patterns. Mid-range performers like KNN and SVC demonstrate decent average accuracies (around 77% and 73%) with a moderate number of misclassifications, making them suitable choices for scenarios where interpretability or efficiency is important. However, lower performers such as LRC, GNB, and MLP exhibit lower average accuracies

(around 65% to 78%) with a higher number of misclassifications, suggesting they might not be the best options for this dataset. To further refine the model selection process, potential next steps include visualizing the data to identify outliers or models with high variance, considering feature importance to understand influential features, and conducting hyperparameter tuning, especially for promising models like Decision Tree, to potentially enhance their performance further. The classification report for the Gini model, likely a decision tree using the Gini impurity criterion, provides insights into the model's precision, recall, and F1-score for each class, indicating its overall effectiveness in distinguishing between different classes. Overall, these evaluations offer valuable insights for optimizing model performance and guiding future steps in the analysis process.

Table 2. Results obtained from different classifiers on the dataset

Classifier	**Cross Validation Score**	**Average Accuracy**	**Mislabeled Data Points**
LRC	[0.71502591, 0.66321244, 0.66839378, 0.66321244, 0.69270833]	0.68(+/- 0.04)	84
KNN	[0.77202073, 0.77202073, 0.75129534, 0.74611399, 0.80208333]	0.77 (+/- 0.04)	44
GNB	[0.65284974, 0.57512953, 0.63212435, 0.70466321, 0.68229167]	0.65 (+/- 0.09)	87
SVC	[0.77202073, 0.73056995, 0.68393782, 0.7357513, 0.75]	0.73 (+/- 0.06)	67
Decision Tree	[0.86010363, 0.90673575, 0.92746114, 0.89119171, 0.890625]	0.90 (+/- 0.04)	22
Random Forest	[0.88082902, 0.90673575, 0.92746114, 0.88601036, 0.90104167]	0.90 (+/- 0.03)	22
XGB	[0.89119171, 0.89119171, 0.92746114, 0.9015544, 0.890625]	0.90 (+/- 0.03)	22
MLP	[0.78756477, 0.76165803, 0.78756477, 0.76683938, 0.80208333]	0.78 (+/- 0.03)	46

Here are some key observations from this table 2 and figure 2:

High Performers: Decision Tree, Random Forest, and XGBoost all achieve a very high average accuracy (around 90%) with a low number of misclassified points. These models seem to be effective at learning the patterns in the data for this particular task.

Mid-Range Performers: KNN and SVC have decent average accuracies (around 77% and 73%) with a moderate number of misclassifications. They might be suitable choices if interpretability or efficiency is a concern, as these models can be easier to understand than Decision Tree-based methods.

Lower Performers: LRC, GNB, and MLP show lower average accuracies (around 65% to 78%) with a higher number of misclassifications. These might not be the best options for this specific dataset.

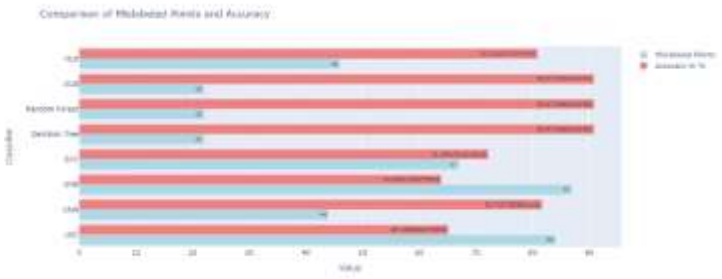

Figure 2. Performance evaluation

Hyperparameter Tunning

A grid search method is employed to optimize the hyperparameters of a decision tree classifier, streamlining the process into several key steps. Initially, the classifier is imported and hyperparameters are defined, specifying parameters such as criterion, splitter, max_depth, min_samples_split, min_samples_leaf, max_features, and max_leaf_nodes. Subsequently, a GridSearchCV object is created to execute the grid search, utilizing arguments like the classifier instance, the defined parameter grid, the number of cross-validation folds, and the scoring metric (accuracy).

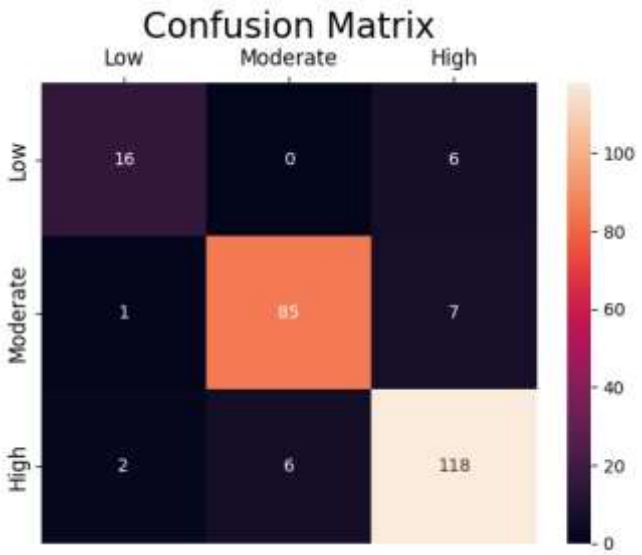

Figure 3. Confusion matrix obtained for decision tree classifier

The grid search is then trained on the provided data, evaluating each model variant's performance via cross-validation. Results are obtained by retrieving the best hyperparameter combination and the best decision tree model, showcasing the ideal settings and the achieved accuracy (approximately 90.87% in this case) (Figure 3). The output highlights the significance of tuning hyperparameters to enhance model accuracy, with the identified settings offering insights into optimizing decision tree performance for the specific dataset.

The classification results in the table provide a performance summary of a model across three classes (Table 3). For class 0, the precision is 0.84, meaning 84% of predicted instances for class 0 are correct, while the recall is 0.73, indicating 73% of actual class 0 instances are correctly identified. For class 1, the model performs strongly, with a precision of 0.93 and recall of 0.91, yielding an F1-score of 0.92, reflecting a good balance between precision and recall. Class 2 has the highest recall of 0.94 and a precision of 0.90, indicating strong performance in identifying this class. The overall accuracy of the model is 91%, with a macro average F1-score

of 0.87, which takes into account the performance for each class equally, and a weighted average F1-score of 0.91, emphasizing the model's robustness given the distribution of classes. This evaluation highlights the model's high effectiveness across all classes, particularly in classes 1 and 2.

Table 3. Performce evaluation of classification algorithms

Class	**Precision**	**Recall**	**F1-Score**	**Support**
0	0.84	0.73	0.78	22
1	0.93	0.91	0.92	93
2	0.90	0.94	0.92	126
Accuracy			0.91	241
Macro Avg	0.89	0.86	0.87	241
Weighted Avg	0.91	0.91	0.91	241

This is a classification report for a Gini model (likely a decision tree using the Gini impurity criterion) that has three classes (labeled 0, 1, and 2). Let's break down the information presented:

Metrics

- ***Precision:*** This metric tells you the proportion of positive predictions that were actually correct. For example, for class 1, a precision of 0.93 means that out of all the data points the model predicted as class 1, 93% were truly class 1.
- ***Recall:*** This metric tells you the proportion of actual positive cases that were correctly identified by the model. For example, for class 1 again, a recall of 0.91 means that out of all the actual class 1 data points, the model identified 91% of them correctly.
- ***F1-Score:*** This metric is a harmonic mean of precision and recall, combining their information into a single score. A value close to 1 indicates good performance on both precision and recall.

Results by Class

- The table shows these metrics (precision, recall, F1-score) for each of the three classes (0, 1, and 2).
- We can see that the model performs well on all three classes, with precision and recall values above 0.84 for each class. The F1-score also reflects this good performance, staying above 0.78 for all classes.

Overall Performance

- ***Accuracy:*** This is the overall percentage of predictions that the model classified correctly. Here, the accuracy is 0.91, indicating that the model classified 91% of the 241 data points correctly.
- ***Macro Average:*** This is the average of the individual class metrics (precision, recall, F1-score) across all classes. Here, the macro averages are around 0.89, which is slightly lower than the overall accuracy but suggests a balanced performance across classes.
- ***Weighted Average*****:** This takes into account the class distribution when calculating the average. Here, the weighted averages are the same as the overall accuracy (0.91) because the report likely assumes a balanced class distribution (or close to it).

Overall Interpretation

This classification report suggests that the Gini model performs well on this 3-class classification task. It achieves high accuracy (91%) and shows good precision, recall, and F1-score for each individual class, indicating the model can effectively distinguish between the different classes.

5. CONCLUSION

The exploration of various machine learning classifiers on a specific dataset, coupled with model evaluation and hyperparameter tuning, has provided valuable insights into the effectiveness of different algorithms for predictive analysis. The evaluation of classifier performance revealed notable variations in accuracy, with Decision Tree, Random Forest, and XGBoost emerging as high performers, achieving an average accuracy of around 90% with a low number of misclassified points. Conversely, classifiers like Logistic Regression, Gaussian Naive Bayes, and Multi-layer Perceptron exhibited lower accuracies, suggesting their limited suitability for the dataset. Further, the hyperparameter tuning of a Decision Tree model using a grid search method demonstrated the importance of parameter optimization in enhancing model performance. The process identified the optimal combination of hyperparameters, resulting in a significant improvement in accuracy to approximately 90.87%, highlighting the effectiveness of automated tuning techniques in refining model robustness and predictive capabilities. These findings underscore the importance of thoughtful model selection, evaluation, and optimization in ensuring the development of accurate and reliable machine learning models for real-world applications.

6. REFERENCES

[1] Jena, D. P. K. (2020). Impact of pandemic covid. International journal of current research, 12(7). https://doi.org/10.24941/ijcr.39209.07.2020

[2] Dutta, D. A. (2020). Impact of Digital social media on Indian Higher Education: Alternative Approaches of Online Learning during COVID-19 Pandemic Crisis. International Journal of Scientific and Research Publications (IJSRP), 10(05), 604–611.https://doi.org/10.29322/ijsrp.10.05.2020.p10169

[3] Mahapatra, a., & Sharma, p. (2021). Education in times of covid-19 pandemic: academic stress and its psychosocial impact on children and adolescents in India. In international journal of social psychiatry (vol. 67, issue 4, pp. 397–399). Sage publications ltd. Https://doi.org/10.1177/0020764020961801

[4] Gali, Y., & Schechter, C. (2020). NGO involvement in education policy: Principals' voices. International Journal of Educational Management, 34(10), 1509-1525.

[5] Subroto, W. (2021). Prevention acts towards bullying in Indonesian schools: A systematic review. AL-ISHLAH: Jurnal Pendidikan, 13(3), 2889-2897.

[6] Singh, S., Singh, U. S., & Nermend, M. (2022). Decision analysis of e-learning in bridging digital divide for education dissemination. Procedia Computer Science, 207, 1970-1980.

[7] Prajapati, B. (2019). Entrepreneurial intention among business students: The effect of entrepreneurship education. Westcliff International Journal of Applied Research, 3(1), 54-67.

[8] Singh, Swati, & Ranjith, M. (2021). Lockdown and its Impact on Education, Environment and Economy. Journal of Advances in Education and Philosophy, 5(4), 116–119. https://doi.org/10.36348/jaep.2021.v05i04.005

[9] Khatri, S. K. (2024). Higher Education in the Digital Era: Exploring the Transformative Effects of Technological Advancements on Higher Education. Dr. Seema Saxena, 19(5).

[10] Sunita, M. L. Education: The 2021 (Post-COVID) Edition.

[11] Priyadarshini, V., Ahmed, M. S., Sathya, R., Koteeswari, D., & Ragothaman, S. (2023). Direct Test Effect of Disruptive Technology Acceptance Model (DTAM) on Massive Online Open Courses (MOOCS) Learners' Satisfaction. Indian Journal of Science and Technology, 16(8), 590-597.

[12] Subha, S., & Priya, S. B. (2021, November). Comparative analysis of supervised machine learning algorithms for evaluating the performance level of students. In 2021 Fifth

International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC) (pp. 1-10). IEEE.

[13] Conway, C. M. (2020). How does the brain learn environmental structure? Ten core principles for understanding the neurocognitive mechanisms of statistical learning. Neuroscience & Biobehavioral Reviews, 112, 279-299.

[14] Yoon, I., & Kim, M. (2023). Determinants of teachers' positive perception on their professional development experience: an application of LASSO-based machine learning approach. Professional Development in Education, 1-15.

[15] Wan, Q., & Ye, L. (2022). [Retracted] Career Recommendation for College Students Based on Deep Learning and Machine Learning. Scientific Programming, 2022(1), 3437139.

[16] Blšták, M., & Rozinajová, V. (2022). Automatic question generation based on sentence structure analysis using machine learning approach. Natural Language Engineering, 28(4), 487-517.

[17] He, S., Epp, C. D., Chen, F., & Cui, Y. (2024). Examining change in students' self-regulated learning patterns after a formative assessment using process mining techniques. Computers in Human Behavior, 152, 108061.

[18] Ionescu, Ș., Delcea, C., Chiriță, N., & Nica, I. (2024). Exploring the Use of Artificial Intelligence in Agent-Based Modeling Applications: A Bibliometric Study. Algorithms, 17(1), 21.

[19] Luan, H., & Tsai, C. C. (2021). A review of using machine learning approaches for precision education. Educational Technology & Society, 24(1), 250-266.

[20] Onyema, E. M., Almuzaini, K. K., Onu, F. U., Verma, D., Gregory, U. S., Puttaramaiah, M., & Afriyie, R. K. (2022). Prospects and challenges of using machine learning for academic forecasting. Computational Intelligence and Neuroscience, 2022(1), 5624475.