PRIMER: Jurnal Ilmiah Multidisiplin
2024, Vol. 2, No. 4, 251-262
https://doi.org/10.55681/primer.v2i4.361

Behavioral Analytics and Machine Learning for Predicting Student Course Completion in E-Learning Systems

Hafiz Khoiru Zaman[1*], Diana Saputri Sri Wahyuningtyas[2]
[1]Informatics Engineering, University Negeri Malang, Indonesia
[2]Development Economics, University Negeri Malang, Indonesia
Corresponding Author's e-mail : dianasaputrisw09@gmail.com

e-ISSN: 2985-7996

Article History:
Received: 02-08-2024
Accepted: 29-08-2024

Abstract : Low course completion rates remain a major challenge in online learning environments, affecting the effectiveness and overall quality of educational outcomes. This study aims to predict students' likelihood of completing online courses using behavioral indicators collected from their interactions with an e-learning platform. A quantitative approach was employed using the Random Forest classification algorithm. The dataset consisted of learner characteristics, course information, and video interaction behaviors, including watch time, pause count, skip count, and disengagement score. Data analysis followed the Knowledge Discovery in Databases (KDD) framework, which included data selection, preprocessing, transformation, modeling, and evaluation stages. The results demonstrated that the Random Forest model achieved excellent predictive performance, with an accuracy of 94.36% and an ROC AUC score of 0.9927. Feature importance analysis revealed that disengagement score and time watched were the most influential predictors of course completion. These findings indicate that behavioral indicators can effectively identify learners at risk of non-completion and support the development of adaptive learning systems. Therefore, the proposed model has the potential to enhance student retention and improve the overall effectiveness of online learning programs.

Kata Kunci : : Online Learning; Course Completion; Student Engagement; Behavioral Indicators; Random Forest; Learning Analytics

INTRODUCTION

The fast pace of growth of information and communication technologies has changed the delivery of education, pushing the adoption of online learning platforms across the world at a faster rate. Online learning platforms offer flexible and scalable learning environments that are trans-geography and trans-time in nature, enabling learners to consume quality learning material at their convenience. Despite all their benefits, web course platforms are plagued by recurring problems with rates of course completions and rates of enrollment. The majority of students register for courses but do not end up doing them, indicating some underlying issues such as lack of motivation, not having adequate support, and behavior of disengagement (Bedi, 2023; Werang & Leba, 2022).

Current research has shown the importance of behavior signals to forecast and explain students' learning performance and activity in web-based courses. Video view time, frequency of skipping and pausing, login behavior, and overall activity rate are increasingly employed proxies to assess cognitive and affective engagement (Luo et al., 2021; Redmond et al., 2023). Current research emphasizes the importance of behavior signals in predicting students' motivation and performance in e-learning (Nia et al., 2023). Measures like login rate, video watch time, session length, and content interaction are typically employed as proxies for cognitive and affective engagement (Rajagopal et al., 2023; Jaiswal et al., 2020). Representative is a PLOS ONE study demonstrating that the length of total study session time was strongly correlated with effort on the part of students and also predicted student performance in the course (Munir et al., 2022). This positions behavioral engagement as a primary determinant of student persistence and course completion.

Machine learning (ML) has good capabilities for the analysis of student activity at scale. In education, ML has been used in personal learning systems, predictive analytics, and intelligent tutoring systems (Yildirim & Celepcikay, 2021; Liu et al., 2023). Trends in large-volume data sets can be identified using ML algorithms, as well as trends of low-achieving students based upon patterns of interaction, and thus intervention can be earlier and more specifically targeted (Kuleto et al., 2021). These methods not only improve retention but can also lead to adaptive instruction sensitive to student need.

Further, ML-based learning technologies can also be used to increase interactivity in the forms of instant feedback and adaptive learning routes (Hilbert et al., 2021). Nevertheless, the integration of ML in learning has its side effects, i.e., concerns regarding data privacy, algorithmic fairness, and instructors having to adapt to pedagogy backed by data (Shamir & Levin, 2021; Tedre et al., 2021). This study proposes a behavior data-driven predictive model for e-learning course completion probability prediction. Based on video interaction and platform use measures, the proposed model is intended to allow educators to recognize disengaging students early enough and institute timely evidence-based interventions to promote performance in virtual learning environments.

RESEARCH METHODOLOGY

This research employs a quantitative approach, applying statistical analysis with the help of machine learning methods. The secondary data source is an open dataset that captures the activities and engagement of students enrolled in online courses. These variables are student engagement with course content, study time, learning style, course difficulty, and course completion. In addition, secondary data are derived from past journal quotes and scientific studies that focus on predicting course completion along with learning behavior classification.

Previous research revealed inconsistent perceptions among students regarding online learning in quantitative research methods courses. For instance, studies conducted among the students of the KPI2 class at IAIN Parepare revealed that, while 60.7% of students were not in favor of online learning, 71.4% of them concurred that it made lecturing easier. Nevertheless, most still preferred face-to-face traditional learning because there was a weak connection and other issues during online learning (Ramadhani et al., 2022; Navis, 2023).

Moreover, the transformation to online learning has also influenced shifts in research methods, e.g., the integration of big data and learning analytics. Online environments are significant in rebuilding the processes of teaching and learning to facilitate analysis of students' workload and learning habits, which provides robust quantitative measures for experimental research in online learning (Toto & Limone, 2021; Velazquez et al., 2022).

The effectiveness of e-learning has been highly controversial. Meta-analyses have established that online learning tends to be less effective than conventional face-to-face learning, mostly due to problems such as the quality of course content, tools employed, and levels of student participation (Prestiadi et al., 2020; Batdı et al., 2021). However, there have been results that indicated students' and instructors' flexibility and creativity in an online setting can significantly enhance learning results (Popa et al., 2020).

Ultimately, there are organizational concerns such as leadership, policy, and management that must be addressed in online learning environments. Also, access, culture, equity, inclusion, and ethics concerns should be considered in order to increase the overall effectiveness of online learning (Martin et al., 2020). Steps in the research conducted by the authors are illustrated in Figure 1 below:

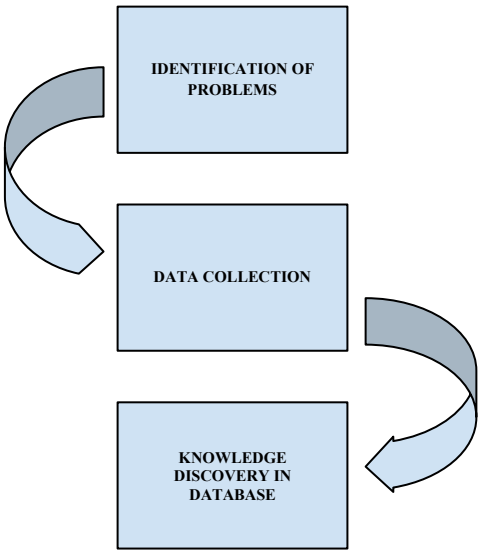

Figure 1. Research Steps

Problem Identification

The low completion rate of courses for online learners is a major challenge in today's digital learning. While flexibility and wide access are brought to learners through online learning, most of them are unable to stay consistently motivated, which negatively impacts learning achievements as well as the performance of learning initiatives (Bedi, 2023; Paulsen & McCormick, 2020). Student engagement behavioral, cognitive, and affective is a main predictor of success in online learning (Luo et al., 2021). Motivation

and self-efficacy are tied together with active engagement, where students who believe in their own abilities are likely to complete their courses successfully (Chiu, 2021).

However, restricted access to technology, inadequate course design, and reduced social interaction often hinder effective engagement (Bedi, 2023; Paulsen & McCormick, 2020). Therefore, predictive approaches grounded in machine learning are needed to identify at-risk learners early and enable timely, targeted interventions that can support course completion.

Data Collection

Analysis employed secondary data derived from an available dataset hosted by Kaggle on the activity monitoring of students undertaking online learning modules. Data consist of various features that include student profiles, learning preferences, qualities of a course, as well as user activities with learning contents based on video. All of the analyses were done using Google Colab. The data is split into three general sections depending on the type of attributes. The first section, Student Information, contains attributes such as student_id, name, age, experience_level, learning_style, and interests. The second section, Course Information, contains course_id, course_name, category, difficulty_level, ratings, num_reviews, and time_spent_on_course. The third section, Video Interaction Data, includes video_id, video_topic, video_duration, time_watched, skip_count, pause_count, and disengagement_score.

The primary target of classification here is the completion_status attribute, indicating whether a student has passed a course or not. All the other attributes are assumed to be predictor variables to model and analyze the learning behavior of the students throughout the whole process of online learning. The dataset consists of 21 attributes and contains 14,101 data instances. For a better description of the structure of the dataset, Table 1 lists each attribute along with its corresponding data type.

Table 1. Dataset Attributes and Their Data Types

Attribute	Data Type	Description
student_id	object	Unique identifier for each student.
name	object	Name of the student.
age	float64	Age of the student.
experience_level	object	Student's online learning experience level.
learning_style	object	Dominant learning style of the student.
interests	object	Topics the student is interested in.
course_id	object	Unique identifier for each course.
course_name	object	Name of the course.
category	object	Category of the course.
difficulty_level	object	Difficulty level of the course.
ratings	float64	Average rating of the course.
num_reviews	float64	Number of reviews for the course.
time_spent_on_course	float64	Total time spent by the student on the course.
completion_status	object	Course completion status (target variable).

Attribute	Data Type	Description
video_id	object	Unique identifier for each video.
video_topic	object	Topic covered in the video.
video_duration	float64	Duration of the video.
time_watched	float64	Time the student spent watching the video.
skip_count	float64	Number of times the student skipped parts of the video.
pause_count	float64	Number of times the student paused the video.
disengagement_score	float64	Score indicating the level of disengagement.

Knowledge Discovery in Database (KDD) Process

Data classification and processing in the study follow the Knowledge Discovery in Database (KDD) process. KDD procedures are necessary for data quality and consistency assurance in machine learning models. A step-by-step overview of the KDD process is presented in Figure 2 and includes the following:

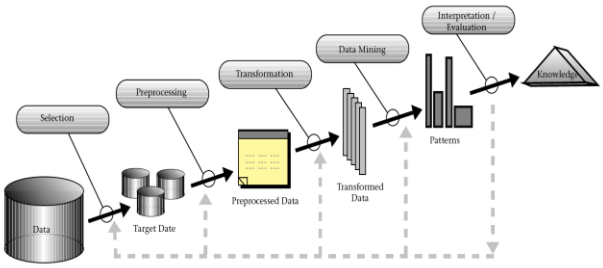

Figure 2. Steps in the KDD Process

1. Data Selection

In this step, the meaningful and relevant factors which influence course completion are selected. Irrelevant factors such as user ID, name, course name, and topic of the video are removed because they do not contribute to the prediction.

2. Preprocessing

The data is preprocessed to eliminate missing values and inconsistencies. Numerical missing values are imputed with the median, and categorical data is imputed with the mode. Column names are normalized as well to facilitate smoother modeling. Categorical columns such as learning_style, experience_level, and difficulty_level are also encoded through label encoding.

3. Data Transformation

The data is converted to a numerical format appropriate for machine learning algorithms. The data set is split into two subsets: training set as 80% and testing set as 20%. The data is also normalized by using StandardScaler in order to enhance the accuracy of the classification model.

4. Data Mining

In this phase, the Random Forest classification technique is used for modeling. The technique was picked because it consistently performs well on data sets with a large

number of characteristics and can handle both numerical and category features. To predict whether or not students finish courses, the model is trained on the training set and evaluated on the testing set.

5. Evaluation

Model performance is done through metrics such as accuracy, classification report, and confusion matrix to determine how well the model classifies the data. Feature importance is also analyzed to determine which features most significantly play a role in predicting course completion. Through the processes described, this study aims to provide a data-driven approach towards solving low online course completion rates through the use of an accurate and effective classification model.

RESULT AND DISCUSSION

Initial Data Exploration

The data of the study reflect the behavior of the students in the online courses. The following attributes are analyzed: time video watched (time_watched), overall time spent on the course (time_spent_on_course), course difficulty level (difficulty_level), number of skips and pauses, and disengagement score. The data were preprocessed prior to modeling. Mode was used to impute missing values for categorical features, while median was used to impute missing values for numerical features. Non-informative features don't contribute a lot to the prediction task and hence were eliminated, like student_id, name, course_id, course_name, video_id, video_topic and video name. The distribution of each feature is illustrated in Figure 3

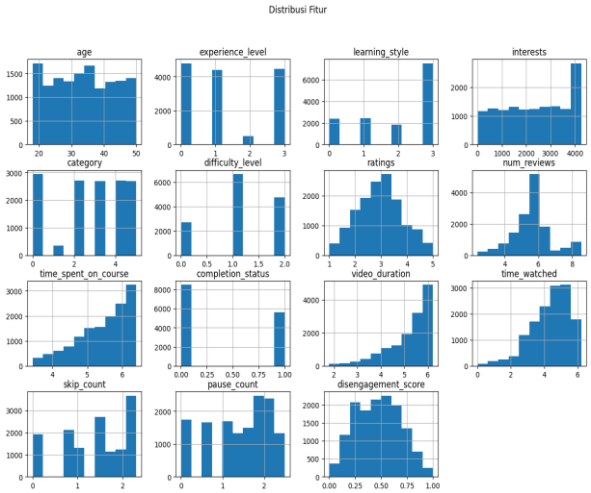

Figure 3. Feature Distribution in the Dataset

To detect potential outliers, a boxplot visualization was used as shown in Figure 4.

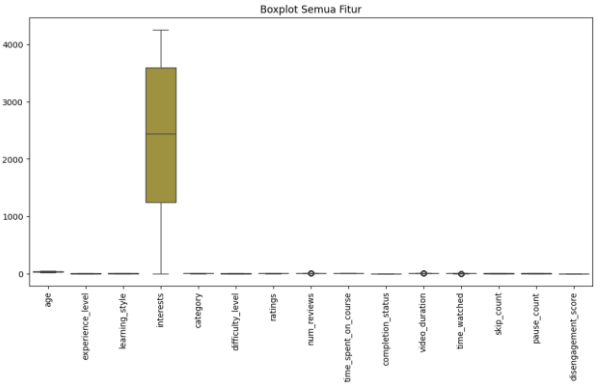

Figure 4. Boxplot of Dataset Features

Feature Correlation

The feature correlations are depicted in a heatmap in Figure 5. Exploration into relationships showed features such as time_watched and disengagement_score have significantly correlated relationships with the target variable (completion_status). A disengagement_score showed a negative correlation (-0.67), whereas time_watched appeared to be a strong positive correlation (0.73), where the more time members spent watching videos, the greater the probability of course completion.

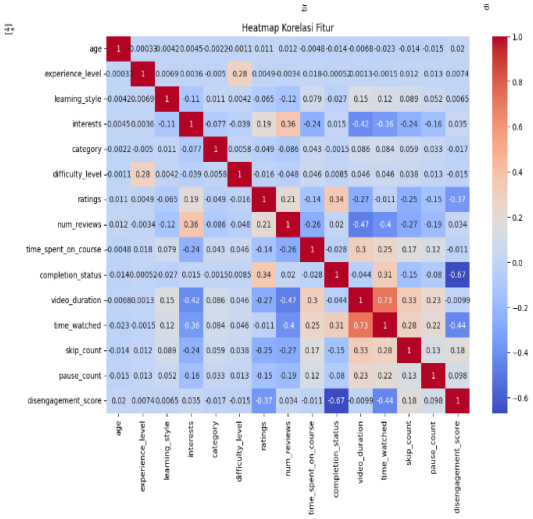

Figure 5. Feature Correlation Heatmap

Random Forest Model Training

The data was split into test sets and settings in a ratio of 80:20. For course imbalance, the Engineered Minority Over-sampling Technique (Destroyed) was implemented. Hyperparameter tuning was also done using GridSearchCV to achieve optimal performance of the Irregular Woodland demonstration. The presentation was done by setting using the optimized parameters with the number of choice trees being 100.

Model Evaluation

The model was evaluated on precision, recall, F1-score, accuracy, and ROC AUC Score classification metrics. All performance results are provided in Table 2 and graphically illustrated using the confusion matrix (Figure 6) and the ROC curve (Figure 7).

Table 2. Random Forest Model Evaluation Results (SMOTE + GridSearchCV)

Class	Precision	Recall	F1-Score	Support
0 (Not Completed)	0.95	0.95	0.95	1710
1 (Completed)	0.93	0.93	0.93	1111
Accuracy			0.9436	2821
Macro Avg	0.94	0.94	0.94	2821
Weighted Avg	0.94	0.94	0.94	2821

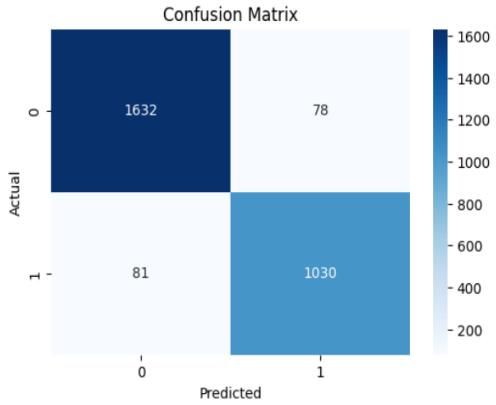

Figure 6. Confusion Matrix of the Random Forest Model

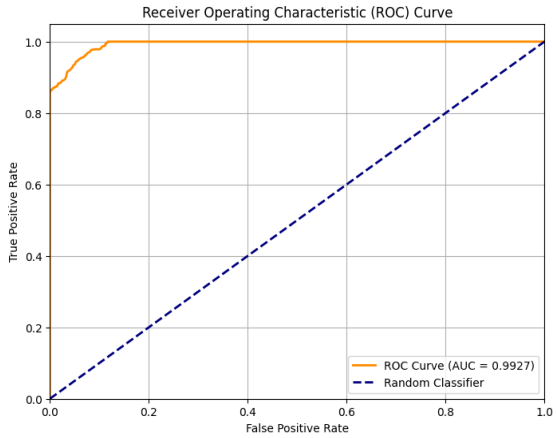

Figure 7. ROC Curve of the Random Forest Model

From the confusion matrix, the model correctly predicted 1,632 students not finishing the course (True Negatives) and 1,030 students finishing the course (True Positives). However, the model also produced 78 False Positives (students predicted to finish the course but did not) and 81 False Negatives (students finishing the course but were predicted otherwise).

Overall, the model performed extremely well in predicting, correctly classifying the majority of the students with 94.36% accuracy. Further, the ROC AUC Score of 0.9927 indicates the model's excellent capability to differentiate between the two classes. The ROC curve (Figure 7) also supports this by exhibiting a high true positive rate at various thresholds.

Feature Importance Analysis

Figure 7 below shows the contribution of each feature to the prediction of course completion status according to the Random Forest model:

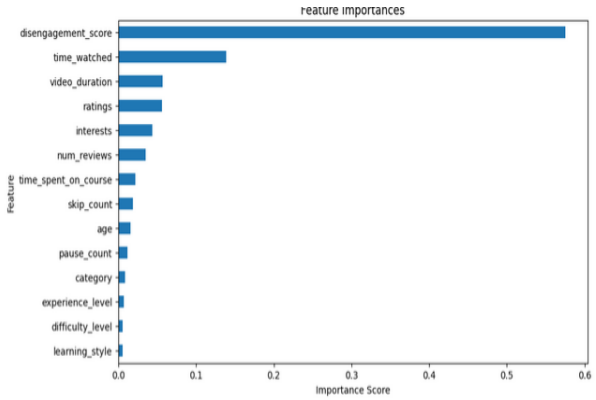

Figure 8. Feature Importance Visualization

It can be observed from the graph that the most influential feature on the prediction is disengagement_score, with an importance score of nearly 0.6. This shows that the level of disengagement is the most significant indicator for predicting whether a participant will complete the course. The time_watched feature is also highly important, indicating that the duration of watching videos is an indicator of participants' engagement in the learning process.

Likewise, features ratings and video_duration are also important, referring to the quality and popularity of the course content. Additionally, interests and num_reviews relate to course topics of interest to learners and popularity of the course among other learners. These elements influence learners' motivation levels to complete the course. Other behavioral markers such as time_spent_on_course, skip_count, and pause_count also have an impact but to a lesser extent. These markers have a direct bearing on learning behavior and how learners engage with the learning content. On the other hand, markers such as learning_style, difficulty_level, and experience_level appear to have a relatively small impact and suggest that course completion is less likely to be predicted by the level of learning style or the level of experience in this data set.

Discussion

The Random Forest classifier applied here was very effective in predicting course completion status. The accuracy was 94.36% and ROC AUC was 0.9927, and it was very effective in discriminating between the course completers and non-completers. The metrics for classification show precision and recall for both classes are very evenly balanced, at 93% to 95%. This suggests the model not only predicts correctly overall but also both classes equally well, which is extremely important in online learning in order not to favor one class of students.

Out of the confusion matrix, 1,632 non-passing students and 1,030 passing students were correctly classified by the model. The misclassifications were modestly low at 78 false positives and 81 false negatives, demonstrating the validity of the model in measuring the course completion status. In feature importance, learner engagement is the most dominant in learning outcomes by far. Disengagement_score was the most prominent feature, pointing out that disengaged or passive behavior is strongly predictive of course non-completion. Time_watched, the total actual video watching time, was also among the most dominant features, suggesting that sustained engagement is key for successful learning.

Features such as video_duration and ratings provided us with information about the quality of content. Higher rated and more duration courses are more interesting and can retain learners. Participant's interests and reviews for each course also proved to be important, as the popularity of the course and relevance of the subject have a positive effect on motivation and completion. Other behavioral traits like time_spent_on_course, skip_count, and pause_count were measuring the learners' interaction behaviors, which also have effects on course completion. Although features like learning style, difficulty level, and experience level were relatively less effective, they may potentially hold contextual value depending on course design or learner background. In general, the results indicate that not only is the model accurate in prediction, but the predictions are also useful for online course designers. One can intervene among the participants who are showing symptoms of disengagement or low engagement, allowing educational institutions to implement preventive strategies and substantially increase course completion rates in the long run.

CONCLUSION AND RECOMMENDATIONS

This study successfully developed a machine learning-based predictive model to estimate the likelihood of course completion among learners in an online learning environment. By employing the Random Forest algorithm combined with SMOTE for class balancing and hyperparameter tuning through GridSearchCV, the model achieved a classification accuracy of 94.36% and an ROC AUC score of 0.9927, demonstrating its strong capability in distinguishing between learners who complete courses and those who do not. Feature importance analysis revealed that behavioral indicators, particularly disengagement_score and time_watched, are the most dominant predictors of course completion status, affirming that active and sustained engagement with learning materials is a critical determinant of success in online education. Overall, this study validates the effectiveness of educational data mining approaches in producing accurate predictive models while generating actionable insights for educators and course developers to improve online course completion rates.

Future research and practice should consider several directions based on the findings of this study. Online learning platforms are encouraged to integrate real-time machine learning-based monitoring systems to detect at-risk learners at the earliest possible stage, enabling timely and targeted interventions. The development of new features reflecting temporal engagement patterns, such as consistency of study time and daily access frequency, should also be explored in subsequent studies to further enhance model predictive power. Deep learning architectures, such as LSTM or Transformer-based models, present promising avenues for capturing more complex sequential engagement patterns that traditional classifiers may overlook. Moreover, ethical considerations surrounding data privacy and algorithmic fairness must remain a central concern to ensure responsible and inclusive deployment of predictive systems in educational settings. Finally, validation of the proposed model across datasets from diverse platforms and cultural contexts is necessary to establish the generalizability of the findings presented in this study.

ACKNOWLEDGMENTS

The authors would like to express their sincere gratitude to all parties who have provided support, both moral and material, in the completion of this research. Special thanks are extended to Universitas Negeri Malang for providing academic facilities and a conducive research environment. The authors also wish to thank fellow researchers for

their constructive discussions, as well as the open-source community for providing the dataset and analytical tools used in this study.

REFERENCES

Bedi, A. (2023). Keep learning: Student engagement in an online environment. *Online Learning*, 27(2). https://doi.org/10.24059/olj.v27i2.3287

Werang, B., & Leba, S. (2022). Factors affecting student engagement in online teaching and learning: A qualitative case study. *The Qualitative Report*, 27(4). https://doi.org/10.46743/2160-3715/2022.5165

Luo, N., Li, H., Zhao, L., Wu, Z., & Zhang, J. (2021). Promoting student engagement in online learning through harmonious classroom environment. *Asia-Pacific Education Researcher*, 31. https://doi.org/10.1007/s40299-021-00606-5

Redmond, P., Alexsen, M., Maloney, S., *et al.*(2023). Student perceptions of online engagement. *Online Learning*, 27(1). https://doi.org/10.24059/olj.v27i1.3320

Nia, H., Marôco, J., She, L., *et al.*(2023). Student satisfaction and academic efficacy during online learning with the mediating effect of student engagement: A multi-country study. *PLoS One*, 18. https://doi.org/10.1371/journal.pone.0285315

Rajagopal, M., Ali, B., Priya, S., *et al.*(2023). Machine learning methods for online education case. In *2023 8th Int. Conf. on Science Technology Engineering and Mathematics (ICONSTEM)*. https://doi.org/10.1109/ICONSTEM56934.2023.10142626

Jaiswal, G., Sharma, A., & Sarup, R. (2020). Machine learning in higher education. In *Handbook of Research on Emerging Trends and Applications of Machine Learning*, 2020. https://doi.org/10.4018/978-1-5225-9643-1.ch002

Munir, H., Vogel, B., & Jacobsson, A. (2022). Artificial intelligence and machine learning approaches in digital education: A systematic revision. *Information*, 13(4). https://doi.org/10.3390/info13040203

Yildirim, Y., & Celepcikay, A. (2021). Artificial intelligence and machine learning applications in education. *Eurasian J. of Higher Education*, 4. https://doi.org/10.31039/ejohe.2021.4.49

Liu, L., Wang, S., Britton, T., & Abebe, R. (2023). Reimagining the machine learning life cycle to improve educational outcomes of students. *Proc. Natl. Acad. Sci. U.S.A.*, 120. https://doi.org/10.1073/pnas.2204781120

Kuleto, V., Ilić, M., Dumangiu, M., *et al.*(2021). Exploring opportunities and challenges of artificial intelligence and machine learning in higher education institutions. *Sustainability*, 13(18). https://doi.org/10.3390/su131810424

Hilbert, S., Coors, S., Eb, K., *et al.*(2021). Machine learning for the educational sciences. *Review of Education*, 2021. https://doi.org/10.31234/OSF.IO/3HNR6

Shamir, G., & Levin, I. (2021). Teaching machine learning in elementary school. *Int. J. Child-Computer Interaction*, 31. https://doi.org/10.1016/j.ijcci.2021.100415

Tedre, M., Toivonen, T., Kahila, J., *et al.*, (2021). Teaching machine learning in K–12 classroom: Pedagogical and technological trajectories for artificial intelligence education. *IEEE Access*, 9. https://doi.org/10.1109/ACCESS.2021.3097962

Ramadhani, R., Dayanti, I., Ramadhani, A., *et al.*, (2022). The influence of online learning on students' level of understanding in courses of quantitative research methods KPI2 class. *Palakka: Media and Islamic Communication*, 2022. https://doi.org/10.30863/palakka.v3i1.2531

Navis, W. (2023). Persepsi pembelajaran daring. *PSIKOSAINS (J. Penelitian dan Pemikiran Psikologi)*, 2023. https://doi.org/10.30587/psikosains.v18i2.5749

Toto, G., & Limone, P. (2021). Online quantitative research methods: A scoping review on education, psychological factors in e-learning, and working from home.

Velazquez, L., Atenas, B., & Castro-Palacio, J. (2022). Quantitative methods to determine the student workload: Empirical study based on digital platforms. *Chaos*, 32(10). https://doi.org/10.1063/5.0103719

Prestiadi, D., Arifin, I., & Bhayangkara, A. (2020). Meta-analysis of online learning implementation in learning effectiveness. In *2020 6th Int. Conf. on Education and Technology (ICET)*. https://doi.org/10.1109/ICET51153.2020.9276557

Batdı, V., Doğan, Y., & Talan, T. (2021). Effectiveness of online learning: A multi-complementary approach research with responses from the COVID-19 pandemic period. *Interactive Learning Environments*, 31. https://doi.org/10.1080/10494820.2021.1954035

Popa, D., Repanovici, A., Lupu, D., *et al.*(2020). Using mixed methods to understand teaching and learning in COVID-19 times. *Sustainability*, 12. https://doi.org/10.3390/su12208726

Martin, F., Sun, T., & Westine, C. (2020). A systematic review of research on online teaching and learning from 2009 to 2018. *Computers & Education*, 159. https://doi.org/10.1016/j.compedu.2020.104009

Paulsen, J., & McCormick, A. (2020). Reassessing disparities in online learner student engagement in higher education. *Educ. Res.*, 49. https://doi.org/10.3102/0013189X19898690

Chiu, T. (2021). Student engagement in K–12 online learning amid COVID-19: A qualitative approach from a self-determination theory perspective. *Interact. Learn. Environ.*, 31. https://doi.org/10.1080/10494820.2021.1926289