The Scientific Temper (2024) Vol. 15 (2): 2199-2206
Doi: 10.58414/SCIENTIFICTEMPER.2024.15.2.32
E-ISSN: 2231-6396, ISSN: 0976-8653
https://scientifictemper.com/

RESEARCH ARTICLE

Predictive modeling of dropout in MOOCs using machine learning techniques

Kinjal K. Patel*, **Kiran Amin**

Abstract

The advent of massive open online courses (MOOCs) has revolutionized education, offering unprecedented access to high-quality learning materials globally. However, high dropout rates pose significant challenges to realizing the full potential of MOOCs. This study explores machine learning techniques for predicting student dropout in MOOCs, utilizing the open university learning analytics dataset (OULAD). Seven algorithms, including decision tree, random forest, Gaussian naïve Bayes, AdaBoost classifier, extra tree classifier, XGBoost classifier, and multilayer perceptron (MLP), are employed to predict student outcomes and dropout probabilities. The XGBoost classifier emerges as the top performer, achieving 87% accuracy in pass/fail prediction and 86% accuracy in dropout prediction. Additionally, the study proposes personalized interventions based on individual dropout probabilities to enhance student retention. The findings underscore the potential of machine learning in addressing dropout challenges in MOOCs and offer insights for instructors and educational institutions to proactively support at-risk students.

Keywords: Machine learning, Predictive modeling, Dropout prediction, MOOCs, Learning analytics.

Introduction

The continuous evolution of technology has ushered in a transformative era for educational institutions, prompting a paradigm shift toward the adoption of scalable e-learning solutions (Xing, Chen, Stein, & Marcinkowski, 2016; Xing & Du, 2018). In response to this digital revolution, educational providers are increasingly embracing innovative methods to deliver content, creating an environment where learners can seamlessly access educational materials on various devices at any time (Lee, Hong, & Hwang, 2018; Lemay & Doleck, 2020). One of the pivotal advantages of this technological progression is the flexibility it affords learners. The traditional constraints of time and location are gradually fading away as e-learning solutions empower students to engage with educational content at their own pace (Joshi, 2024; Thulasi, 2020). This adaptability accommodates diverse learning styles, allowing individuals to absorb information in a manner that suits their preferences and capabilities, ultimately enhancing the overall learning experience (Jeon, Park, & Bang, 2020).

The advent of massive open online courses (MOOCs) marks a pivotal moment in the evolution of education, transforming traditional paradigms and making learning accessible on a global scale (Feng, Tang, & Liu, 2019). Introduced in 2008 by Georges Siemens, MOOCs have rapidly gained prominence due to their openness, simplicity, quality, and the unprecedented reach they offer (Xing & Du, 2018). The year 2012, often dubbed "The year of the MOOC," witnessed a significant surge in their popularity, marked by the launch of prominent platforms like Coursera, edX, and Udacity (Laura, 2012). These platforms revolutionized distance education by providing free access to a diverse range of subjects from top universities worldwide (Ahmed Ali Mubarak, Cao, & Ahmed, 2021). This democratization of knowledge enables learners, regardless of their geographical location or socio-economic status, to engage with high-quality educational content (S. Gupta & Sabitha, 2019; Sanchez-Gordon & Lujan-Mora, 2016).

MOOC courses are designed for large-scale participation by leveraging the power of the internet, making education more accessible to a global audience with internet connectivity (R. Gupta & Sambyal, 2013; Xing *et al.*, 2016). They cover a diverse range of subjects and are often available

Faculty of Engineering and Technology, Ganpat University, Mehsana, Gujarat, India.

***Corresponding Author:** Kinjal K. Patel, Faculty of Engineering and Technology, Ganpat University, Mehsana, Gujarat, India., E-Mail: cekinjalvp@gmail.com

How to cite this article: Patel, K. K., Amin, K. (2024). Predictive modeling of dropout in MOOCs using machine learning techniques. The Scientific Temper, **15**(2):2199-2206.

Doi: 10.58414/SCIENTIFICTEMPER.2024.15.2.32

Source of support: Nil

Conflict of interest: None.

Received: 28/04/2024 **Accepted:** 23/05/2024 **Published** : 15/06/2024

at little to no cost, democratizing access to education on a global scale (Itani, Brisson, & Garlatti, 2018). The emergence of MOOCs has ushered in a new era of learning, extending the boundaries of education beyond traditional confines. This development has been particularly impactful in making education accessible to individuals worldwide, irrespective of geographical location or socio-economic status (Xing *et al*., 2016). The affordability and flexibility associated with MOOCs have played a crucial role in breaking down barriers to education and fostering a more inclusive and equitable learning environment (Ali & Arts, 2023; Lambert, 2020). Learners have the freedom to engage with course materials at their own pace and from any location, providing a level of convenience that was previously unimaginable. This adaptability not only caters to diverse learning styles but also accommodates the busy schedules of individuals who may not have had the opportunity to pursue education through traditional means (R. Gupta & Sambyal, 2013).

Moreover, the impact of MOOCs extends beyond individual learners. Recognizing the potential of these platforms, colleges and universities are increasingly integrating MOOCs into their offerings (Guerrero, Heaton, & Urbano, 2021). By leveraging the expansive reach and interactive features of MOOC platforms, educational institutions can augment their programs, reaching a broader audience and diversifying their educational delivery methods (Biswas & Sarkar, 2022). MOOCs have created a dynamic learning ecosystem, enabling global interaction among learners, professors, and peers. The ability to connect with individuals from around the world fosters a collaborative and enriching educational experience (Guerrero *et al*., 2021). Whether learners seek free access to knowledge or opt for paid certifications, MOOCs have become a valuable resource for both those seeking to expand their skills and trainers aiming to offer accessible and effective online education (Lambert, 2020).

Despite the widespread popularity of MOOCs, it is essential to acknowledge and address the existing limitations that hinder their transformative potential. One of the most significant challenges is the persistently high dropout rates, with less than 10% of enrollees successfully completing the course and obtaining a certificate (Mourdi, Sadgal, Berrada Fathi, & El Kabtane, 2020). This issue is underscored by concrete examples, such as a software engineering course offered by MIT and Berkeley, which experienced a pass rate as low as 7% among its 50,000 registrations (R. Gupta & Sambyal, 2013). Similarly, Duke University's Bioelectricity MOOC saw only 2.6% of the initially registered 12,175 participants completing the course (Onah, Sinclair, & Boyatt, 2014).

This prevalent pattern of high incompletion rates poses a substantial obstacle to the realization of the full benefits of MOOCs. While the low completion rate is often attributed to a scale-efficacy tradeoff (Onah *et al*., 2014), it remains a significant challenge that needs to be addressed to harness the potential of these online learning platforms fully.

One promising avenue for improvement involves the development of efficient student success prediction models within MOOCs (Gardner & Brooks, 2018). These predictive models can analyze various factors and patterns to anticipate student dropout, completion, and overall learning outcomes (Mourdi *et al*., 2020). By identifying early indicators of potential disengagement or struggles, instructors and educational platforms can intervene proactively to support learners in overcoming challenges and staying engaged throughout the course (Bergdahl, 2022). The implementation of effective predictive models offers a strategic approach to enhance enrollment, completion rates, and the overall learner experience in MOOCs. Once these models are refined and proven effective, personalized interventions can be tailored to the specific needs of individual learners. This targeted support can take various forms, including additional resources, personalized feedback, or adaptive learning strategies, all designed to improve learner outcomes and foster more meaningful interaction between learners and instructors (Gardner & Brooks, 2018; Mourdi, Sadgal, El Kabtane, & El Abdallaoui, 2021).

The growing concern over the high attrition rates in MOOCs has prompted researchers to explore innovative solutions, leading to the integration of learning analytics methods for early prediction of learners who may be at risk of dropping out. This paper details a thorough examination of dropout rate prediction through the application of various machine-learning techniques. Therefore, this paper aims to achieve two primary research goals: firstly, to explore the diverse machine learning techniques employed for dropout prediction, and secondly, to investigate personalized interventions based on individual dropout probabilities.

The predictive models discussed herein have the potential to aid educational institutions and instructors in promptly identifying students at risk of academic struggles. This early detection enables timely interventions, allowing educators to employ suitable persuasive techniques to motivate struggling students, thereby improving their performance and encouraging them to stay on course. This research not only underscores the interdisciplinary nature of learning analytics but also highlights the potential impact of data-driven insights on shaping the future of MOOCs. As machine learning continues to advance, the findings from this study may pave the way for more effective strategies in addressing the challenges associated with high attrition rates and optimizing the learning experience for a diverse range of MOOC participants.

Literature Review

The surge in MOOC popularity has led to a plethora of enrolled individuals generating extensive log files capturing their activities, and researchers are delving into these datasets to extract meaningful insights (Boyer, Gelman, Schreck, & Veeramachaneni, 2015). Several studies have focused on predicting dropout rates among MOOC learners, employing diverse approaches and models.

Taylor, Veeramachaneni, and O'Reilly (2014) utilized clickstream and forum submission data to train a logistic regression classifier, predicting the likelihood of students discontinuing their learning journey. Similarly, He, Bailey, Rubinstein, and Zhang (2015) employed logistic regression, considering factors such as course completion, assignment completion, and scoring to predict dropout instances. Time series classification methods, including hidden Markov chains, nonlinear state space models, and recurrent neural networks (RNNs), have also been explored. Kizilcec, C. Piech, and Schneider (2013) categorized learner engagement into four classes based on video lectures and assignment grades, using clustering techniques to describe engagement activity. Ahmed A. Mubarak, Cao, and Zhang (2020) introduced a predictive model combining logistic regression with an input-output hidden Markov model. Santana, Costa, Fonseca, Rêgo, and Araújo (2017) implemented four machine learning (ML) algorithms, with the support vector machine (SVM) achieving the highest accuracy in identifying students at risk of failure. Additionally, Li, Baker, and Warschauer (2020) analyzed clickstream data to uncover the relationship between online engagement and academic performance.

Dataset

In this study, the researchers aim to contribute to the understanding of dropout rate prediction using machine learning techniques by utilizing The open university learning analytics dataset (OULAD). The OULAD serves as the foundational source for training and testing models in the investigation of dropout rate prediction using a variety of machine learning techniques in this study. Published by the Open University, a prominent publicly funded British university, the dataset is a valuable resource in the realm of educational research.

The OULAD comprises an extensive collection of over 300,000 records of student activity, providing a rich and comprehensive dataset for analysis. This dataset encompasses various dimensions of student engagement, including log data from virtual learning environments (VLE), details about seven distinct courses offered, student demographic information, and course-related data such as grades and assessments. The inclusion of diverse data points allows researchers to explore a multifaceted view of student behavior and academic performance within the context of online learning environments. The dataset was meticulously collected as part of the open university learning analytics project, reflecting the university's commitment to leveraging analytics and data-driven interventions to enhance student success and retention. The overarching goal of this project aligns with the broader

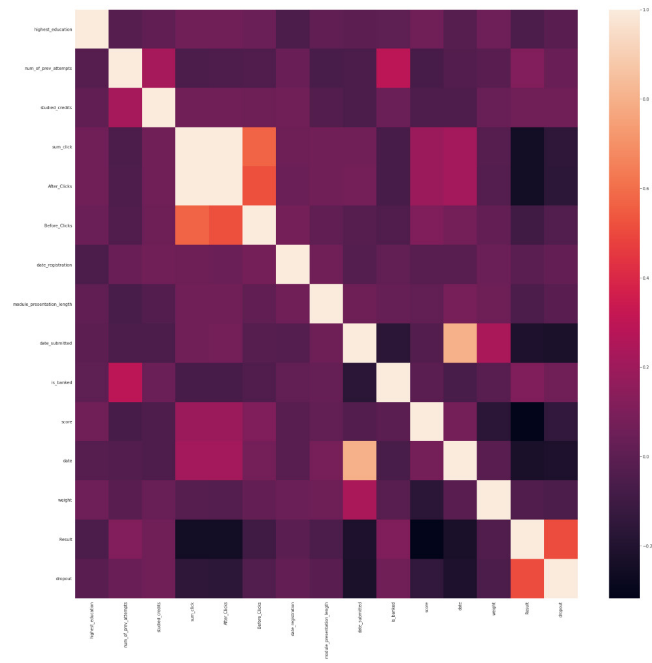

Figure 1: Heatmap for different features

trends in the educational landscape, where institutions are increasingly turning to learning analytics to gain insights into student behavior and implement proactive measures to support their academic journey.

Data Preprocessing

Prior to analysis, the data underwent meticulous processing to guarantee accuracy, relevance, and suitability for analytical progress. This preparatory phase involved employing various techniques, including cleaning, normalization, scaling, feature extraction, and selection. Incomplete records were addressed by either eliminating them from the dataset or substituting missing values with appropriate replacements, such as mean or mode values, chosen based on the specific context of the data being analyzed. This rigorous data processing approach lays the groundwork for robust, insightful analyses, ensuring the reliability and integrity of our research findings.

Upon thorough examination of the 'date_submitted' and 'date' columns, three new attributes were created: 'click_timing', 'before_click', and 'after_click.' These supplementary attributes played a pivotal role in assessing the timeliness of student assignment submissions. To operationalize this, the submission date was systematically compared to the assignment deadline, facilitating the classification of submissions into 'on-time' (assigned a value of 1) or 'late' (assigned a value of 0).

To explore the relationships between different features and student performance, we constructed a correlation matrix and generated a heatmap using the Pearson correlation coefficient (Figure 1). This enabled us to measure the intensity and direction of the linear association between

	highest_education	num_of_prev_attempts	studied_credits	sum_click	After_Clicks	Before_Clicks	date_registration	module_presentation_length	date_submitted	is_banked	score	date	weight	Result	dropout
highest_education	1	-0.0239	0.01096	0.06038	0.05956	0.04338	-0.06	0.00819	-0.0009	0.00471	0.05905	-0.021	0.05536	-0.0541	-0.0132
num_of_prev_attempts	-0.0239	1	0.22497	-0.0505	-0.0495	-0.04	0.04082	-0.0629	-0.0599	0.29079	-0.0664	-0.0342	-0.0106	0.10973	0.03783
studied_credits	0.01096	0.22497	1	0.06328	0.06201	0.0507	0.06335	-0.0335	-0.0569	0.04331	-0.0476	-0.0481	0.03268	0.06005	0.05945
sum_click	0.06038	-0.0505	0.06328	1	0.99792	0.57193	0.04798	0.06608	0.06664	-0.067	0.18846	0.21265	-0.0231	-0.2517	-0.156
After_Clicks	0.05956	-0.0495	0.06201	0.99792	1	0.51779	0.04391	0.06812	0.07105	-0.0667	0.18825	0.21623	-0.0258	-0.255	-0.1599
Before_Clicks	0.04338	-0.04	0.0507	0.57193	0.51779	1	0.0779	0.01012	-0.0197	-0.0407	0.10552	0.07051	0.021	-0.0953	-0.0351
date_registration	-0.06	0.04082	0.06335	0.04798	0.04391	0.0779	1	0.06195	-0.0282	0.01244	-0.0184	-0.0174	0.04392	-0.0085	0.02071
module_presentation_length	0.00819	-0.0629	-0.0335	0.06608	0.06812	0.01012	0.06195	1	0.04782	0.02846	0.01433	0.08333	0.05247	-0.053	-0.0134
date_submitted	-0.0009	-0.0599	-0.0569	0.06664	0.07105	-0.0197	-0.0282	0.04782	1	-0.1725	-0.0339	0.79715	0.23809	-0.2153	-0.2219
is_banked	0.00471	0.29079	0.04331	-0.067	-0.0667	-0.0407	0.01244	0.02846	-0.1725	1	-0.0081	-0.0612	-0.0146	0.10487	0.0604
score	0.05905	-0.0664	-0.0476	0.18846	0.18825	0.10552	-0.0184	0.01433	-0.0339	-0.0081	1	0.07606	-0.1664	-0.3177	-0.1471
date	-0.021	-0.0342	-0.0481	0.21265	0.21623	0.07051	-0.0174	0.08333	0.79715	-0.0612	0.07606	1	-0.0127	-0.2277	-0.2158
weight	0.05536	-0.0106	0.03268	-0.0231	-0.0258	0.021	0.04392	0.05247	0.23809	-0.0146	-0.1664	-0.0127	1	-0.0436	-0.0508
Result	-0.0541	0.10973	0.06005	-0.2517	-0.255	-0.0953	-0.0085	-0.053	-0.2153	0.10487	-0.3177	-0.2277	-0.0436	1	0.50877
dropout	-0.0132	0.03783	0.05945	-0.156	-0.1599	-0.0351	0.02071	-0.0134	-0.2219	0.0604	-0.1471	-0.2158	-0.0508	0.50877	1

Figure 2: Correlation matrix

each pair of variables (Figure 2). The output suggests a potentially strong correlation between students' level of engagement in the course, measured by the number of clicks, and their final performance. Students who clicked more frequently may have been more actively engaged with the material, more proactive in seeking out resources and support, and consequently more likely to perform well on assessments. Similarly, there appears to be a robust relationship between assessment scores and final performance. Students with higher assessment scores are more likely to excel in the course and continue without dropping out.

Methodology

In the dataset, a systematic categorization process was implemented to classify student outcomes into distinct classes for analysis. Initially, the pass and distinction results were grouped to constitute a unified PASS class, emphasizing successful outcomes. Simultaneously, the failed and withdrawn results were combined to create a consolidated FAIL class, encompassing instances of academic challenges or discontinuation. Moreover, a broader classification was introduced where PASS, FAIL, and DISTINCTION results were grouped together, collectively forming a NON-DROPOUT class. Conversely, instances marked as WITHDRAWN were specifically categorized as indicative of a DROPOUT. This classification schema allowed for a comprehensive examination of student outcomes, facilitating distinct analyses of academic success, failure, and dropout patterns.

The dataset exhibits a notable imbalance in the distribution of PASS and FAIL categories, as well as in the proportion of DROPOUT and NON-DROPOUT students. To address this issue and enhance the robustness of the analysis, this study adopts the Synthetic minority over-sampling technique (SMOTE) to balance the data by generating synthetic instances of the minority class, thereby mitigating the imbalance and ensuring a more equitable representation of both outcomes. This technique contributes to a more balanced training set, enabling predictive models to better capture patterns and relationships within both classes.

The literature lacks definitive guidelines for the optimal division of a dataset into training (analysis) and test (holdout) groups. Divergent recommendations exist, with some researchers endorsing an 80 to 20 split between the analysis and holdout samples, while others advocate for a 75 to 25 division. In alignment with the decision-making process, an 80:20 train-test split was implemented to partition the dataset for training and testing purposes. The 80:20 split aligns with common practices in the field, facilitating a robust evaluation of the model's predictive capabilities on unseen data.

Experimental Result

This paper utilizes a range of predictive models, including decision tree, random forest, Gaussian naïve Bayes, AdaBoost classifier, extra tree classifier, XGBoost Classifier, and multilayer perceptron (MLP), to anticipate dropouts within the OLAUD dataset. The objective of the study was to predict whether a student would pass or fail the course and to determine whether they would drop out or continue with the course. To achieve this, the performance of each model is evaluated using diverse metrics such as accuracy, precision, recall, and F1-score. Accuracy provides an overall measure of how often the model correctly predicts both dropout and non-dropout instances. Precision focuses on the proportion of correctly predicted dropout instances out of all instances predicted as dropouts, aiming to minimize false positive predictions. Recall, also known as sensitivity, measures the proportion of actual dropout instances that are correctly identified by the model, thereby mitigating false negative predictions. The F1-score, which harmonizes precision and recall, offers a balanced assessment of a model's predictive capability by considering both false positives and false negatives. The following Tables 1 and 2 display the outcomes achieved by all algorithms using the balanced dataset.

According to the findings presented in Tables 1 and 2, the XGBoost classifier demonstrates superior performance, achieving an accuracy of 87% in predicting student pass/fail outcomes. Moreover, it exhibits a remarkable accuracy of 86% in forecasting whether a student will persist or drop out of the course. These results underscore the efficacy of the XGBoost algorithm in predictive modeling within the context of student academic performance and retention.

Dropout Probability for Personalization

In the next study, this study proposes to predict the dropout probabilities of each student. As shown in Tables 3 and 4, each of these algorithms produced the fail probability and dropout probability of each student.

Table 1: PASS/FAIL prediction using balanced dataset

Algorithm	*Precision*		*Recall*		*F1-score*		*Accuracy (%)*
	PASS	*FAIL*	*PASS*	*FAIL*	*PASS*	*FAIL*	
Decision tree classifier	0.82	1.00	1.00	0.31	0.90	0.48	84
Gaussian naïve bayes	0.88	0.53	0.82	0.65	0.85	0.58	78
Random forest classifier	0.89	0.47	0.75	0.70	0.81	0.56	74
Extra tree classifier	0.87	0.63	0.89	0.58	0.88	0.60	82
XGB classifier	0.94	0.68	0.88	0.82	0.91	0.74	87
Adaboost classifier	0.91	0.55	0.82	0.73	0.86	0.63	80
MLP	0.87	0.82	0.96	0.53	0.91	0.64	86

Table 2: Dropout/Non-dropout prediction using balanced dataset

Algorithm	*Precision*		*Recall*		*F1-score*		*Accuracy (%)*
	Non dropout	*Dropout*	*Non dropout*	*Dropout*	*Non dropout*	*Dropout*	
Decision tree classifier	0.98	0.13	0.52	0.87	0.68	0.22	54
Gaussian naïve bayes	0.98	0.14	0.57	0.87	0.72	0.24	59
Random forest classifier	0.97	0.19	0.73	0.76	0.83	0.30	73
Extra tree classifier	0.97	0.16	0.66	0.79	0.79	0.27	67
XGB classifier	0.99	0.34	0.86	0.92	0.92	0.50	86
Adaboost classifier	0.98	0.20	0.73	0.82	0.83	0.32	73
MLP	0.98	0.34	0.84	0.90	0.90	0.54	83

Table 3: Probability of failing the student

	DT	*GNB*	*RF*	*ET*	*XGB*	*ADABOOST*	*MLP*
student1	0.40674	0.57247	0.57206	0.47274	0.52068	0.52068	0.68413
student2	0.40674	0.12910	0.41047	0.44724	0.44826	0.45826	0.02503
student3	0.40674	0.00688	0.51350	0.45333	0.61844	0.51844	0.79951
student4	0.40674	0.00025	0.39746	0.36079	0.44428	0.44728	0.00052
student5	0.40674	0.00043	0.39792	0.34749	0.36845	0.46845	0.00532
student6	0.40674	0.04686	0.40050	0.41324	0.76845	0.46845	0.35875
student7	0.40674	0.10443	0.42147	0.42117	0.36694	0.48694	0.11783
student8	0.40674	0.22558	0.53615	0.47328	0.42239	0.46870	0.08656
student9	0.40674	0.00879	0.40537	0.38021	0.35726	0.45826	0.00216
student10	0.40674	0.19525	0.43047	0.41247	0.34565	0.49505	0.28688

Table 4: Probability of dropout of students

	DT	*GNB*	*RF*	*ET*	*XGB*	*ADABOOST*	*MLP*
student1	0.64177	0.44072	0.49961	0.50845	0.48818	0.48818	0.15261
student2	0.64177	0.56097	0.39880	0.42488	0.42488	0.49613	0.21755
student3	0.64177	0.99950	0.57104	0.61565	0.57104	0.52331	0.74082
student4	0.20145	0.00207	0.38005	0.46064	0.00207	0.41531	0.00855
student5	0.20145	0.08722	0.42841	0.43359	0.50129	0.50129	0.28374
student6	0.20145	0.00000	0.34150	0.26617	0.26617	0.42488	0.00010
student7	0.20145	0.10739	0.44669	0.38238	0.20145	0.42056	0.03234
student8	0.64177	0.99885	0.56697	0.58025	0.52331	0.52331	0.38604
student9	0.64177	0.00000	0.44435	0.29889	0.29889	0.50330	0.08346
student10	0.64177	0.94268	0.57267	0.46523	0.36416	0.49152	0.36416

Discussion and Conclusion

MOOCs are becoming more population nowadays due to their numerous benefits, including flexibility, accessibility, and the democratization of education. These platforms allow learners worldwide to access high-quality educational content from prestigious institutions, often at little to no cost. This accessibility breaks down traditional barriers related to geographical location and socio-economic status, making education more inclusive. Additionally, MOOCs provide learners with the freedom to engage with course materials at their own pace and convenience, accommodating diverse learning styles and schedules.

Despite these advantages, MOOCs face significant challenges, notably high dropout rates. High attrition rates undermine the potential of MOOCs, suggesting a need for more effective strategies to maintain learner engagement and course completion. This paper explores diverse machine learning approaches to predict student dropouts, utilizing the OULAD. The study focuses on building a prediction model for the early identification of at-risk students using seven well-known machine learning algorithms: Decision tree, random forest, Gaussian naïve Bayes, AdaBoost classifier, extra tree classifier, XGBoost classifier, and multilayer perceptron (MLP).

The experimental results demonstrate that among these predictive models, the XGBoost classifier performs the best, achieving 87% accuracy in predicting pass/fail outcomes and 86% in dropout predictions. These results underline the effectiveness of the XGBoost algorithm in educational contexts, suggesting its potential for broader application in predicting and mitigating dropout rates. Moreover, personalized interventions based on predicted dropout probabilities were proposed, aiming to enhance learner retention and success. The predictive models' ability to generate dropout probabilities for each student enables tailored support, which is critical for improving learner outcomes and fostering meaningful interactions between learners and instructors.

The findings from this study offer crucial insights for managers and decision-makers within MOOC platforms and educational institutions. By leveraging advanced machine learning algorithms, several key actions can be implemented to improve student outcomes and resource allocation. First, the developed prediction model allows instructors to identify students at higher risk of dropping out early in the course. This proactive identification enables the deployment of timely interventions, such as personalized feedback, additional tutoring, or motivational communications, which can significantly reduce dropout rates and enhance overall student retention. Second, understanding the specific factors contributing to dropout probabilities—such as low engagement or poor assessment scores—enables instructors to design and implement targeted support strategies. For instance, students demonstrating low engagement, as indicated by the number of clicks, can be offered engagement-boosting activities or resources, while those with poor assessment scores might benefit from supplementary instructional materials or personalized coaching. Lastly, the predictive insights from the machine learning algorithms enable MOOC providers to allocate their resources more efficiently, focusing on students who are most likely to benefit from additional support and interventions. The predictive models can inform broader strategic planning and policy development. By understanding patterns and trends in student engagement and dropout rates, instructors can develop long-term strategies aimed at continuous improvement of MOOC offerings, ensuring they remain competitive and effective in meeting the educational needs of diverse learner populations. This strategic approach ensures that MOOCs can continue to deliver high-quality education to a broad audience, maximizing their potential impact on global education.

This study has certain limitations that are indeed critical for understanding the scope and applicability of its findings. By focusing solely on data from a single MOOC course offered by Open University, the study may not capture the full spectrum of factors influencing student dropout across various MOOC platforms and course types. This limitation could impact the generalizability of the findings to broader contexts. To address this limitation, future research could replicate the study using data from multiple MOOC courses offered by different providers. Additionally, adopting different feature selection techniques can help ensure that the features retained in the analysis are the most relevant and informative for predicting student dropout. Future research endeavors should aim to address these limitations to advance our understanding of student attrition in online learning environments.

References

Ali, A. J. I. J. o. M. S., & Arts. (2023). Exploring the Transformative Potential of Technology in Overcoming Educational Disparities. *2*(1).

Bergdahl, N. (2022). Engagement and disengagement in online learning. *Computers & Education, 188*, 104561. doi:https://doi.org/10.1016/j.compedu.2022.104561

Biswas, S., & Sarkar, M. (2022). MOOC : Challenges & Prospects in Indian Higher Education. *Journal of Information and Computational Science, 10*(2), 1015-1024.

Boyer, S., Gelman, B. U., Schreck, B., & Veeramachaneni, K. (2015, 19-21 Oct. 2015). *Data science foundry for MOOCs*. Paper presented at the 2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA).

Feng, W., Tang, J., & Liu, T. X. (2019). Understanding Dropouts in MOOCs. *Proceedings of the AAAI Conference on Artificial Intelligence, 33*(01), 517-524. doi:10.1609/aaai.v33i01.3301517

Gardner, J., & Brooks, C. (2018). Student success prediction in MOOCs. *User Modeling and User-Adapted Interaction, 28*(2), 127-203. doi:10.1007/s11257-018-9203-z

Guerrero, M., Heaton, S., & Urbano, D. (2021). Building universities' intrapreneurial capabilities in the digital era: The role and impacts of Massive Open Online Courses (MOOCs). *Technovation, 99*, 102139. doi:https://doi.org/10.1016/j.technovation.2020.102139

Gupta, R., & Sambyal, N. (2013). An understanding Approach towards MOOCs. *International Journal of Emerging Technology and Advanced Engineering, 3*(6), 312-315.

Gupta, S., & Sabitha, A. S. (2019). Deciphering the attributes of student retention in massive open online courses using data mining techniques. *Education and Information Technologies, 24*(3), 1973-1994. doi:10.1007/s10639-018-9829-9

He, J., Bailey, J., Rubinstein, B., & Zhang, R. (2015). Identifying At-Risk Students in Massive Open Online Courses. *Proceedings of the AAAI Conference on Artificial Intelligence, 29*(1).

Itani, A., Brisson, L., & Garlatti, S. (2018, 2018//). *Understanding Learner's Dropout in MOOCs*. Paper presented at the Intelligent Data Engineering and Automated Learning – IDEAL 2018, Cham.

Jeon, B., Park, Namyong , & Bang, S. (2020). Dropout Prediction over Weeks in MOOCs via Interpretable Multilayer Representation Learning. *arXiv:2002.01598 [cs.LG]*.

Joshi, M. (2024). Adaptive Learning through Artificial Intelligence. *International Journal on Integrated Education, 7*, 41-43. doi:10.2139/ssrn.4514887

Kizilcec, R. F., C. Piech, & Schneider., E. (2013). Deconstructing disengagement: analyzing learner subpopulations in massive open online courses. *in Proceedings of the third international conference on learning analytics and knowledge*.

Lambert, S. R. (2020). Do MOOCs contribute to student equity and social inclusion? A systematic review 2014–18. *Computers & Education, 145*, 103693. doi:https://doi.org/10.1016/j.compedu.2019.103693

Laura, P. (2012). The year of the MOOC. The New York Times. *The New York Time*

Lee, J., Hong, A., & Hwang, J. (2018). *A Review of Massive Open Online Courses: MOOC's Approach to Bridge the Digital Divide*.

Lemay, D. J., & Doleck, T. (2020). Predicting completion of massive open online course (MOOC) assignments from video viewing behavior. *Interactive Learning Environments*, 1-12. doi:10.1080/10494820.2020.1746673

Li, Q., Baker, R., & Warschauer, M. (2020). Using clickstream data to measure, understand, and support self-regulated learning in online courses. *The Internet and Higher Education, 45*, 100727. doi:https://doi.org/10.1016/j.iheduc.2020.100727

Mourdi, Y., Sadgal, M., Berrada Fathi, W., & El Kabtane, H. (2020). A Machine Learning Based Approach to Enhance MOOC Users' Classification. *Turkish Online Journal of Distance Education, 21*(2), 47-68.

Mourdi, Y., Sadgal, M., El Kabtane, H., & El Abdallaoui, H. E. A. (2021). A Multilayers Perceptron for predicting weekly learner commitment in MOOCs. *Journal of Physics: Conference Series, 1743*, 012027. doi:10.1088/1742-6596/1743/1/012027

Mubarak, A. A., Cao, H., & Ahmed, S. A. M. (2021). Predictive learning analytics using deep learning model in MOOCs' courses videos. *Education and Information Technologies, 26*(1), 371-392. doi:10.1007/s10639-020-10273-6

Mubarak, A. A., Cao, H., & Zhang, W. (2020). Prediction of students' early dropout based on their interaction logs in online learning environment. *Interactive Learning Environments*, 1-20. doi:10.1080/10494820.2020.1727529

Onah, D. F., Sinclair, J., & Boyatt. (2014). *Dropout rates of massive open online courses: Behavioral pattens of MOOC droput and completion: Existing evalutaion*. Paper presented at the 6th International Conference on Education and New Learning Technologies (EDULEARN14).

Sanchez-Gordon, S., & Lujan-Mora, S. (2016). How Could MOOCs Become Accessible? The Case of edX and the Future of Inclusive Online Learning. *Journal of Universal Computer Science, 22(1), 55–81., 22*(1), 55-81.

Santana, M., Costa, E., Fonseca, B., Rêgo, J., & Araújo, F. (2017). Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses. *Computers in Human Behavior, 73*. doi:10.1016/j.chb.2017.01.047

Taylor, C., Veeramachaneni, K., & O'Reilly, U. M. (2014). Likely to stop? Predicting stopout in massive open online courses. *arXiv:1408.3382 [cs.CY]*.

Thulasi, B. S. (2020). Analysis on Massive Open Online Course (MOOC) Dropout Prediction Using Machine Learning Techniques The State of the Art. *The International journal of analytical and experimental modal analysis, XII*(III), 1287-1290.

Xing, W., Chen, X., Stein, J., & Marcinkowski, M. (2016). Temporal predication of dropouts in MOOCs: Reaching the low hanging fruit through stacking generalization. *Computers* *in Human Behavior, 58*, 119-129. doi:https://doi.org/10.1016/j.chb.2015.12.007

Xing, W., & Du, D. (2018). Dropout Prediction in MOOCs: Using Deep Learning for Personalized Intervention. *Journal of Educational Computing Research, 57*(3), 547-570. doi:10.1177/0735633118757015