

Predicting Student Success in MOOCs using Machine Learning: A Comparative Study of Classification Models

Jaspreet Kaur¹, Divya Gupta² and Sheenam³

¹⁻³Department of Computer Science and Engineering, Chandigarh University, Mohali - 140413, Punjab, India
Email: jassimathoda@gmail.com, divya1907gupta@gmail.com, sheenammiddhal124@gmail.com

Abstract— The proliferation of Massive Open Online Courses (MOOCs) has completely changed the educational landscape by providing students with access to high quality education content worldwide. However, issues like high dropout rates and uneven student interest levels still exist despite MOOCs' growing popularity. This paper explores the potential of artificial intelligence (AI) in evaluating and enhancing the learning process within the framework of MOOCs. We present an empirical examination of MOOC data with an emphasis on dropout trends, course completion rates, and student engagement. We forecast student achievement using machine learning models, such as Random Forest classifiers, by taking into account factors like quiz scores, platform usage, and engagement levels. Our findings show that Gradient Boosting has achieved better results in terms of accuracy and AUC. Additionally, we pinpoint important classes with high dropout rates and examine the underlying causes, providing information on how AI might be used to improve student retention. The study highlights the benefits of AI-based prediction models for individualized instruction, giving teachers the means to proactively assist and intervene on behalf of pupils who are at risk. These results highlight the important role AI can play in improving learning outcomes in the MOOC ecosystem, which will increase student retention and success.

Index Terms— MOOCs, artificial intelligence, machine learning models, student dropout rate, course.

I. INTRODUCTION

Massive Open Online Courses (MOOCs) have completely changed the educational landscape by providing students around the world with access to a wide range of quality educational content. By removing conventional hurdles like time limitations, physical location, and expensive tuition, MOOCs have democratized learning and increased access to education for a wider range of people. Millions of students participate in MOOCs each year, consuming content from prestigious universities and institutes, according to recent reports. The success of MOOCs in achieving educational outcomes hampered by a number of key issues, notwithstanding their increasing popularity. The high dropout rate, where many students do not complete the courses they begin, is one of the most urgent problems. According to research, engagement levels, time commitment, and academic background are important elements that affect students' performance in MOOCs [1].

However, it is challenging for teachers to offer individualized attention or customized learning experiences due to the scale and diversity of the student population. As a result, low retention rates arise from many students finding it difficult to stay involved. This issue is compounded by the lack of real-time feedback and personalized interventions that could potentially help students stay on track. Artificial intelligence (AI) has the potential to resolve these issues. By leveraging AI techniques such as machine learning and data analytics educational platforms can acquire a greater understanding of student behaviours, engagement patterns, and learning.

To increase retention and completion rates, these insights can be used to forecast student performance, identify students who are slow learners and suggest personalized interventions. A more effective and interesting learning environment can be created by AI-based systems by offering adaptive learning environments that customize course material to each student's needs. This study explores how AI can be used to analyse MOOC course completion rates and student engagement. In particular, we concentrate on using machine learning models to forecast student success using engagement data, including platform usage, quiz scores, and involvement in class activities. We want to show how AI-driven models may assist students have the best possible learning experience while also identifying important aspects that affect student retention and dropout rates through an empirical examination of MOOC data. By doing this, we hope to draw attention to how AI might improve learning outcomes in MOOCs and advance the area of AI in education more broadly [2].

A. Application of AI

There are several practical applications of AI that directly improve MOOC efficacy and promote student achievement. Below are some key applications showcasing how AI transforms online education:

Early Dropout Prediction and Intervention: Artificial intelligence can proactively identify early in the course, students who are at risk by examining patterns in their behaviour, including engagement levels, quiz scores, assignment submission timeliness and connection frequency. Through constant observation of these signs, AI systems are able to identify students who might be at risk of failing or dropping out. Timely interventions, including sending tailored reminders, offering focused assistance resources, or starting instructor outreach, are made possible by this early detection, increasing the likelihood that students will complete their courses and succeed overall [3].

Personalized Learning Pathways: Artificial intelligence is able to provide personalized learning experiences by suggesting content, adjusting the pace of instructions, and tailoring difficulty level of exercises based on each student's unique learning preferences, patterns, and performance information. AI systems are able to develop adaptive paths that can be tailored to the needs of individual learners by evaluating data such as quiz scores, time spent on learning modules, and interaction history. Better educational outcomes are eventually the result of this personalisation, which also helps to close knowledge gaps and increase learning efficiency while maintaining student engagement [2], [3].

Intelligent Tutoring Systems: When a student is facing problem with a concept, task, or test, artificial intelligence can provide them with immediate feedback, cues, or thorough explanations. Artificial intelligence (AI)-driven tutoring systems can quickly identify misconceptions or knowledge gaps by tracking student replies and inter action patterns. They can then offer tailored advice, recommend different learning materials, or give detailed explanations to help students grasp the material. In large scale online learning settings like MOOCs, this instant support is very helpful since it enables students to make corrections early, gain confidence, and continue making steady progress without waiting for teacher intervention [4].

Real-Time Engagement Monitoring: AI can track student activity patterns to identify early indicators of dis engagement, like prolonged periods of inactivity, little to no participation in discussion boards, and a tendency to skip assignments or tests. Continuous analysis of these behavioural characteristics enables AI systems to promptly detect students who are in danger of falling behind or quitting school. This makes it possible to implement prompt interventions, such automatic reminders, customised cheerleading messages, or suggestions for re engagement tactics, which keep students interested in the material and raise retention rates overall [5].

II. RELATED WORK

MOOCs are becoming increasingly popular as a flexible way to learn at one's own time and place, and as a result, many students are enrolling in these courses. MOOC providers keep logs to monitor student activities [6]. This widespread use has attracted the attention of many researchers who are interested in analysing these logs to uncover hidden insights. The author's strategy [6], which uses decision trees on MOOC data and clickstream data, is able to predict dropout rates much more accurately. Although clickstream data are the main emphasis of this paper, the methodology may theoretically also be integrated with forum data (for example, by employing

multiple kernel learning), a topic that can be addressed in future research. Xing et al. [7] developed time-aware predictive models that predicted student dropout using logistic regression and decision trees. During various course weeks, they examined characteristics including video views, assignment submissions, and forum participation. To identify at-risk students early in the course lifecycle, their method placed more emphasis on temporal behaviour than static traits. In this study author [8] used comprehensive clickstream data to record specific student behaviours such as timing of interactions, video pausing, and page navigation. According to their research, these small-scale interactions are powerful markers of student involvement and, when paired with more conventional activity elements, can greatly increase the accuracy of dropout predictions. Early-week behaviour was found to be especially indicative of long-term results.

Clickstream data offers high-resolution insights, but it necessitates a lot of preprocessing and storage. In this research, the authors [9] provide a process for determining which students are most likely to discontinue a course, and they create and assess a strategy to mitigate this risk. author employ machine learning techniques to forecast likely dropouts using clickstream data from MOOCs provided through the MITx Micro Master's program in Supply Chain Management.

80% of actual dropouts are accurately predicted by their model. they used these findings to inform their focused intervention, which involves sending individualised emails of encouragement to learners who have been identified as at risk in order to increase their motivation and engagement. Nevertheless, the intervention had little effect on dropout rates, even with a high email open rate. This study [10] offers a thorough analysis of several MOOC dropout prediction techniques, emphasising distinct approaches and areas in need of further research. It draws attention to the expanding MOOC landscape and the possibility of technologically driven improvements in outcome accuracy. The use of sophisticated models based on machine learning, deep learning, and meta-heuristics techniques to enhance learning outcomes, increase course completion rates, and offer individualised educational experiences is also covered in this review.

III. METHODOLOGY

The objective of this study is to explore the potential applications of Artificial Intelligence (AI) approaches that can be used to improve student engagement and course completion rates in Massive Open Online Courses (MOOCs). we empirically analyse MOOC data by employing machine learning models to predict student performance based on a variety of engagement indicators. The following subsections describes the dataset, exploratory data analysis (EDA), machine learning models, data preprocessing, and assessment measures used in this work.

A. Dataset description

In this we have used publicly available dataset of MOOCs platform. The dataset includes student information, engagement metric, course information, completion status. The dataset also having information regarding records of student interactions with course content, as well as their final outcomes (e.g., completion or dropout).

B. Machine learning models

Several machine learning methods are used to forecast dropout rates and student progress. The models were chosen based on their ability to handle the complexity and size of the data. The following models are considered. Random forest: Random Forest is a strong ensemble-based model and is especially well-suited for managing large datasets and identify complex as well as non-linear relationship among features. These models avoid overfitting and increase accuracy by combining the prediction capabilities of several decision trees. These models are used in MOOCs to categorise students into groups like "completed" or "dropped out" by examining a range of engagement-related characteristics, including as quiz scores, time spent on course content, frequency of logins, and discussion involvement. In online learning environments, they are useful tools for comprehending and forecasting student achievement due to their capacity to process high-dimensional data and produce feature importance scores.

Logistic Regression: Logistic regression is a popular statistical model for binary classification tasks, such predicting whether a student will complete a course or drop out. Unlike more complex models, Logistic Regression provides clear insights into how individual input features—such as quiz scores, time spent on the platform, or number of discussion posts—contribute to the final prediction. This model helps in determining which characteristics have the most impact on student outcomes by evaluating the likelihood of course completion based on these qualities. It is especially useful in situations where explainability is just as crucial as prediction accuracy because of its transparency and simplicity of use.

Support Vector Machine: SVM is a powerful and adapt able model, that is frequently employed for binary classification tasks, particularly when the aim is to determine a distinct margin of separation between two groups. SVM is utilised in MOOCs to identify the appropriate decision border between students who are likely to complete the course and those who are at risk of dropping out. If necessary, SVM can handle both linear and non-linear correlations in the data by mapping input features—like engagement metrics, quiz scores, and interaction frequencies—into a higher-dimensional space. Its strength is its ability to optimise the margin between classes, which frequently results in better generalisation performance, particularly for datasets of modest or smaller sizes.

Gradient Boosting: Gradient Boosting is an advanced ensemble learning method that builds extremely accurate predictive models by combining the outputs of several weak learners, usually decision trees, one after the other. The model's performance can be improved gradually by training each new tree to fix the mistakes caused by the ones that came before it. Gradient boosting in the contexts of MOOCs, is particularly effective for capturing intricate, non-linear correlations between engagement features like quiz performance, frequency of logins, and forum participation. This method is more robust than standard models in handling noisy or unbalanced data, and it also improves forecast accuracy. For predicting student success and dropout chances in extensive online education datasets, it is a highly effective option due to its capacity to rank and evaluate the most informative aspects.

IV. EVALUATION METRICS

A variety of indicators are used to assess the performance of machine learning models. The accuracy, precision, recall, F1 score and Area Under the Curve (AUC) are the metrics used for evaluating performance.

Accuracy: The most widely used evaluation metric is accuracy, where TP (True Positive) is the number of positive samples that the model predicts as positive classes and TN (True Negative) is the number of negative samples that the model predicts as negative classes. $P + N$ is the number of all samples. It is given by

$$Accuracy = \frac{TP+TN}{P+N} \quad (1)$$

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

where FP (False Positive) stands for negative samples predicted as positive by the model.

$$Recall = \frac{TP}{TP} \quad (3)$$

where FN (False Negative) is the number of negative samples predicted as positive by the model.

$$F1_score = 2 \cdot \frac{Precision \cdot Recall}{Precision + r} \quad (4)$$

AUC (Area Under the Curve) is a popular statistic for evaluating tasks involving binary classification. A classifier's capacity to discriminate between positive and negative classes is gauged by its AUC score.

$$AUC = \frac{1}{N_+ \cdot N_-} \sum_{i=1}^{N_+} \sum_{j=1}^{N_-} I(S_i > S_j) \quad (5)$$

Where N_+ is number of positive samples, N_- is number of negative samples and S_i is score of the i th positive sample and S_j is score of the j th negative sample. $I(S_i > S_j)$ is 1 if $S_i > S_j$ and it is 0.5 if $(S_i = S_j)$ and it is zero if $(S_i < S_j)$.

V. RESULT

This section presents the result of our machine learning models that forecast MOOC dropout rates and student achievement. We use a number of metrics, including accuracy, precision, recall, F1-score, and AUC, to assess the performance of many models, including Random Forest, Logistic Regression, Support Vector Machines (SVM), and Gradient Boosting Machines (GBM). We also examine the connections between course completion rates and important engagement characteristics. The Performance of each machine learning model is assessed using cross validation techniques and results has been shown in table1. Gradient Boosting outperformed all other models of machine learning with the highest accuracy of 86.7%, an f1-score of 0.86 and AUC of 0.92 demonstrating higher predictive power in distinguishing between students who complete the course and those who drop out. Random Forest came next, showing a good trade-off between recall and precision with an AUC of

0.91 and an F1-score of 0.85.SVM and Logistic Regression also produced good results with accuracies of 82.3% and 80.5% respectively, though they were slightly less effective in capturing the complexity of the learning patterns.

TABLE I: COMPARISON OF MODEL PERFORMANCE METRICS

Model	Accuracy (%)	Precision	Recall	F1-score	AUC
Random Forest	85.2	0.82	0.88	0.85	0.91
Logistic Regression	80.5	0.78	0.84	0.81	0.87
Support Vector Machine	82.3	0.80	0.86	0.83	0.88
Gradient Boosting	86.7	0.83	0.89	0.86	0.92

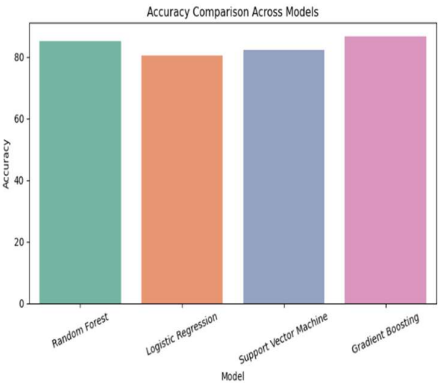


Figure 1. Accuracy Comparison of different techniques

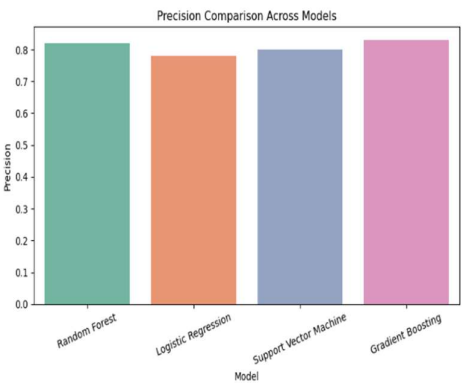


Figure 2. Precision Comparison of different techniques

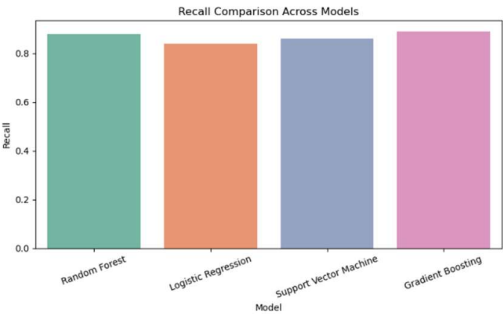


Figure 3. Recall Comparison of different techniques

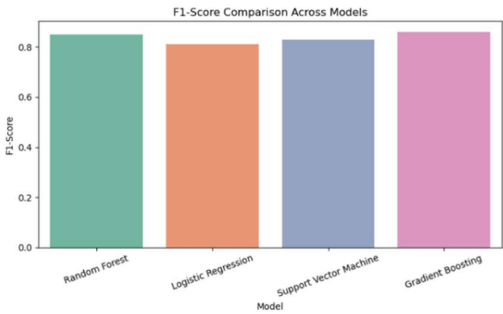


Figure 4. F1-score Comparison of different techniques

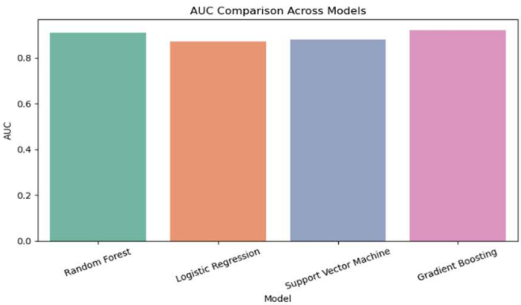


Figure 5. AUC Comparison of Different Techniques

The Gradient Boosting model outperformed the other models in terms of total prediction performance, achieved the best accuracy (86.7%) as shown in figure 1. In terms of predicting student success, the figure 2 shows that Random Forest and Gradient Boosting both have the highest precision scores, indicating that they are more effective at reducing false positives.

Gradient Boosting and Random Forest lead in recall, demonstrating their superiority in accurately recognizing students who actually completed the course as shown in figure 3. Figure 4 shows the F1-score of different machine learning techniques and demonstrates that the Gradient Boosting has the best balance between precision and recall. Gradient Boosting has the maximum area under the ROC curve (0.92), according to the figure 5, demonstrating its ability to differentiate between students who completed and those who dropped out.

V. CONCLUSION

The study emphasizes how machine learning may enhance MOOC learning by detecting at-risk individuals early and facilitating personalized interventions. The findings of this research show the potential of machine learning models in forecasting MOOC dropout rates and student performance, offering important new information about the variables affecting course completion. Predictive performance was assessed using different machine learning models including Random Forest, Logistic Regression, Support Vector Machine (SVM), and Gradient Boosting. Each model was evaluated using accuracy as the main evaluation criterion. Among the models tested Gradient boosting achieved highest accuracy followed by random Forest. The results emphasize the significance of engagement metrics and suggest that MOOC providers can use AI-based models to customize learning routes and offer support to improve retention and completion rates. Future research should focus on further refining these models, integrating real-time data, and exploring more advanced techniques, such as deep learning, to enhance prediction accuracy and create more dynamic learning environments.

REFERENCES

- [1] Yu, J., Luo, G., Xiao, T., Zhong, Q., Wang, Y., Feng, W., ... & Tang, J. (2020, July). MOOCCube: A large-scale data repository for NLP applications in MOOCs. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 3135-3142).
- [2] Bai, X., Zhang, F., Li, J., Guo, T., Aziz, A., Jin, A., & Xia, F. (2021). Educational big data: Predictions, applications and challenges. *Big Data Research*, 26, 100270.
- [3] Nguyen, A. K., Nguyen, D. K., Vo, K. T., Nguyen, T., Nguyen-Hoang, T. A., Dinh, N. T., & Nguyen, H. T. (2024, October). Proposing big data architecture for addressing dropout problem in mooc platforms. In *2024 International Conference on Advanced Technologies for Communications (ATC)* (pp. 522-527). IEEE.
- [4] Li, B., Li, G., Xu, J., Li, X., Liu, X., Wang, M., & Lv, J. (2023). A personalized recommendation framework based on MOOC system integrating deep learning and big data. *Computers and Electrical Engineering*, 106, 108571.
- [5] Basnet, R. B., Johnson, C., & Doleck, T. (2022). Dropout prediction in Moocs using deep learning and machine learning. *Education and Information Technologies*, 27(8), 11499-11513.
- [6] Kloft, M., Stiehler, F., Zheng, Z., & Pinkwart, N. (2014, October). Predicting MOOC dropout over weeks using machine learning methods. In *Proceedings of the EMNLP 2014 workshop on analysis of large scale social interaction in MOOCs* (pp. 60-65).
- [7] 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.
- [8] Whitehill, J., Mohan, K., Seaton, D., Rosen, Y., & Tingley, D. (2017). Delving deeper into MOOC student dropout prediction. *arXiv preprint arXiv:1702.06404*.
- [9] Borrella, I., Caballero-Caballero, S., & Ponce-Cueto, E. (2019, June). Predict and intervene: Addressing the dropout problem in a MOOC- based program. In *Proceedings of the sixth (2019) ACM conference on learning@ scale* (pp. 1-9).
- [10] Alghamdi, S., Soh, B., & Li, A. (2025). A Comprehensive Review of Dropout Prediction Methods Based on Multivariate Analysed Features of MOOC Platforms. *Multimodal Technologies and Interaction*, 9(1), 3.
- [11] Chi, Z., Zhang, S., & Shi, L. (2023). Analysis and prediction of MOOC learners' dropout behavior. *Applied Sciences*, 13(2), 1068.