Predictive Insights into Course Completion Rates in MOOCs

Syed Arham Akheel

Solutions Architect
Redmond, WA
arhamakheel@yahoo.com

Abstract

Massive Open Online Courses (MOOCs) have revolutionized higher education by providing scalable and flexible learning opportunities. However, completion rates remain consistently low, prompting researchers to investigate key predictors of student success. This study combines a literature review with an empirical experiment using machine learning models to predict course completion based on behavioural and demographic data. Decision Tree, Random Forest, and regression models are employed to identify features that significantly impact course outcomes. The results highlight the crucial role of engagement, prior education, and social interactions in improving completion rates.

Keywords: MOOCs, Machine Learning, Dropout Prediction, Course Completion, Engagement

I. INTRODUCTION

MOOCs have gained popularity in recent years as an affordable and accessible alternative to traditional higher education [1]. Despite their potential to democratize learning, the dropout rates in MOOCs are high, with completion rates often below 10% [8]. These low completion rates present significant challenges for both students and course creators. For students, the lack of interaction, motivation, and support often leads to disengagement and eventual dropout [3], [4]. Pursel et al. [3] found that nearly 70% of learners cited lack of motivation as the primary reason for not completing a course, while Sunar et al. [4] emphasized the importance of social interactions in sustaining learner engagement.

For course creators, maintaining student engagement and ensuring course quality is particularly challenging due to the scale of MOOCs and the diversity of learners [2]. Dalipi et al. [2] highlighted that course creators struggle to design content that caters to different learning styles, as MOOCs often attract learners from varying educational backgrounds and with diverse levels of prior knowledge. Furthermore, Hughes and Dobbins [6] demonstrated that course interactions, such as quizzes and video lectures, need to be optimized to improve engagement, but achieving this at scale remains a complex task.

Another critical challenge is the generalizability of predictive models across different MOOCs. Although machine learning models, such as Random Forest and Decision Tree, have shown promise in predicting at-risk students, their effectiveness can vary depending on course-specific features [6], [2]. Borrella et al. [7] found that using predictive analytics to trigger interventions reduced dropout rates by 15%, but the effectiveness of these interventions often depends on the course structure and the type of content being delivered. Moreover, Moreno-Marcos et al. [8] emphasized that temporal patterns of engagement play a crucial role in understanding dropout behaviour, but incorporating these patterns into predictive models is challenging due to the variability in learner behaviour over time.

Understanding and predicting student success has become a central focus in research, with machine learning (ML) models emerging as promising tools to identify students at risk of dropout [2], [10]. Nagrecha

et al. [10] pointed out that interpretability of ML models is crucial for their adoption by educators, as black-box models may not provide actionable insights for improving course design. This study explores the use of ML techniques to predict student success and proposes potential interventions to enhance completion rates. By leveraging both demographic and engagement data, this study aims to provide insights into the key drivers of student success and offer recommendations for improving the learning experience in MOOCs.

II. LITERATURE REVIEW

A. Engagement and Motivation in MOOCs

Engagement is a key factor that significantly impacts course completion rates in MOOCs. Studies have shown that learner behaviour, such as watching video lectures and participating in discussions, is positively correlated with course completion [1]. Behavioural engagement, which refers to participation in course activities, is critical for academic success, as learners who interact with course content frequently are more likely to complete the course [1], [4]. According to Al-Shabandar et al. [1], students who watched over 75% of video lectures had a 60% higher completion rate compared to those who watched less than 25%. Motivation also plays an essential role in MOOCs, with highly motivated students exhibiting greater persistence and success [3]. Pursel et al. [3] found that 80% of students with strong intrinsic motivation completed their courses, whereas only 30% of students with low motivation did so. However, lack of motivation is a leading cause of dropout, and interventions aimed at increasing motivation can effectively enhance retention [2]. Dalipi et al. [2] reported that dropout rates can be reduced by up to 20% with targeted motivational interventions, such as personalized reminders and feedback.

B. Predictive Modeling for Dropout Prediction

Previous research has extensively utilized ML models to predict student dropout in MOOCs. Dalipi et al. [2] highlighted that student-related factors such as motivation and time availability, along with MOOC-specific factors like course design, contribute to high dropout rates. Models like Random Forest, Decision Tree, and regression have been applied to identify students at risk of dropping out by analyzing features such as course interactions and assessment scores [2], [6]. For instance, Hughes and Dobbins [6] demonstrated that Random Forest models could achieve an accuracy of up to 75% in predicting student dropout based on interaction data. Predictive models not only help in identifying students at risk but also allow timely interventions to reduce attrition [7]. Borrella et al. [7] found that using predictive models to trigger interventions reduced dropout rates by 15% in a MOOC-based program. However, challenges remain in ensuring the generalizability of these models across different courses, as course-specific features can significantly affect model performance.

C. Social Interactions and Peer Engagement

Social engagement has been identified as a critical component in sustaining student interest in MOOCs. Sunar et al. [4] found that students who actively engage in discussions and follow peers are more likely to complete courses. Their study reported that students who participated in at least five discussion forums had a 50% higher completion rate than those who did not participate. Social network analysis has shown that repeated interactions with peers can significantly improve course completion rates, as learners feel a sense of community and support [4]. Encouraging social interactions through group discussions and forums has been proposed as an effective intervention strategy to increase learner retention [4]. However, fostering meaningful interactions remains challenging due to the scale of MOOCs and the varying levels of student engagement. Providing structured opportunities for collaboration, such as group projects or peer reviews, has been suggested as a potential solution to enhance social engagement.

III. METHODOLOGY

A. Data Collection and Preprocessing

The dataset used in this study was derived from the Open University Learning Analytics Dataset (OULAD) [5]. The OULAD is a publicly available dataset containing information on 32,593 students enrolled in seven different courses offered by the Open University. The dataset includes a variety of features, such as:

- Demographic Data: Age, gender, highest education level, region, and disability status.
- Engagement Metrics: Number of clicks on course materials, forum participation, assessment submissions, and time spent on the Virtual Learning Environment (VLE).
- Performance Metrics: Scores on assignments, quizzes, and final exam results.
- Intervention Indicators: Records of whether students received additional support, such as reminder emails or feedback.

The dataset was collected by the Open University and provides detailed records of student interactions with the learning platform. The data is structured to facilitate learning analytics research, making it suitable for evaluating factors that contribute to student success and dropout rates.

Feature selection was conducted using both statistical methods and domain knowledge. Correlation analysis was used to identify features with a significant relationship to the target variable (course completion). Features with a high correlation to course success, such as video engagement, assessment scores, and discussion participation, were retained, while features with little or no correlation were removed to reduce model complexity and avoid overfitting. Additionally, domain knowledge was leveraged to ensure that features important for understanding student behaviour, such as the number of course interactions and prior academic background, were included in the analysis.

Handling missing values was also an essential part of the preprocessing pipeline. Missing demographic data, such as age or educational background, was imputed using median values, while missing engagement metrics were filled with zeros, assuming that lack of recorded activity indicated nonparticipation. Categorical features, such as education level and gender, were encoded using one-hot encoding to ensure that machine learning models could process these features effectively. Numerical features, such as assessment scores and the number of interactions, were normalized to bring all features onto a similar scale, which helps in improving the convergence of gradient-based optimization algorithms used in certain machine learning models.

The importance of effective data preprocessing and feature selection cannot be overstated, as these steps directly impact the quality of the machine learning models. By reducing noise and focusing on the most relevant features, the models were better able to identify patterns that predict student success. This approach also improved the interpretability of the models, allowing educators to gain actionable insights into which factors most significantly influence course completion.

B. Experiment Design

To predict student success, we employed several machine learning models, including Decision Tree, Random Forest, Linear Regression, Ridge Regression, and Lasso Regression. Each model was chosen for its unique strengths in capturing different aspects of the data's underlying patterns, providing a holistic view of the features influencing course completion rates.

1) Model Selection and Justification: The Decision Tree and Random Forest models were chosen due to their interpretability and ability to capture non-linear relationships between features and outcomes. Decision Trees are particularly useful for understanding feature importance, which makes them valuable tools for educational researchers who require

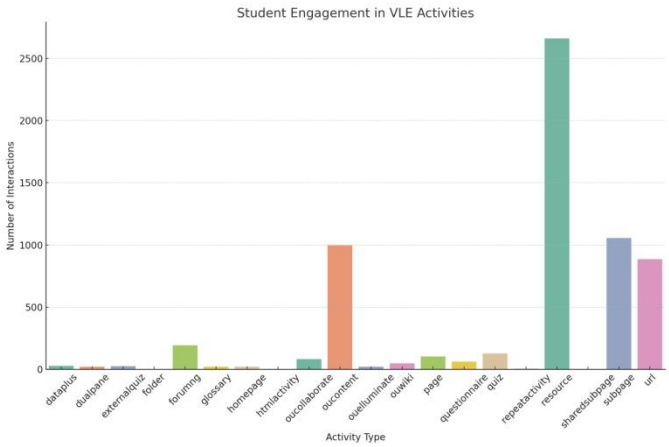

Fig. 1. Student Engagement in VLE Activities

actionable insights into student behaviours [6]. Random Forest, which is an ensemble of Decision Trees, has the advantage of reducing overfitting and increasing accuracy through averaging multiple decision paths [6]. Studies have shown that Random Forest can achieve an accuracy of up to 75% in predicting student dropout based on interaction data [6].

Linear Regression, Ridge Regression, and Lasso Regression were used to predict the likelihood of course completion as a continuous variable. These regression models were selected to capture linear relationships between features, allowing us to estimate the impact of each feature on the probability of completion. Ridge and Lasso are regularized forms of Linear Regression that help mitigate multicollinearity and overfitting by adding penalties to the regression coefficients. In MOOCs, where engagement metrics often correlate with each other (e.g., video views and discussion participation), regularization helps in producing more stable and generalizable models [2].

2) Experimental Setup: The dataset was split into training and testing sets with an 80:20 ratio, ensuring that sufficient data was available for model training while keeping a separate set for evaluating generalization performance. The target variable was defined as course completion (pass/fail), and the features included demographic information, engagement metrics (e.g., video views, discussion participation), and assessment scores.

The features were carefully selected based on correlation analysis and domain knowledge, as previously described. The most relevant features included:

- Video Engagement: Number of video lectures watched.
- Discussion Participation: Frequency of participation in forums.
- Assessment Scores: Scores achieved in quizzes and assessments.
- Demographic Information: Age, gender, and educational background.

These features have been identified in the literature as significant predictors of course success [1], [3]. For example, Al-Shabandar et al. [1] reported that students who watched more than 75% of the video content were 60% more likely to complete the course.

3) Model Evaluation: The models were evaluated using several metrics to assess their predictive performance:

- Accuracy: Used to evaluate the classification models (Decision Tree, Random Forest) by measuring the percentage of correct predictions.

- Mean Squared Error (MSE) and Mean Absolute Error (MAE): These metrics were used for regression models to evaluate the difference between predicted and actual values. Lower values of MSE and MAE indicate better predictive performance.
- R-squared (R^2): Measures the proportion of variance in the target variable that can be explained by the model. Higher R^2 scores indicate better model fit.

Additionally, confusion matrices were used to assess the performance of the classification models in terms of correctly and incorrectly predicted outcomes. The confusion matrices highlighted specific challenges in differentiating between students who withdrew and those who failed, indicating the need for more nuanced features or additional data sources.

4) Scientific Reasoning Behind the Design: The combination of classification and regression models provided a comprehensive approach to understanding student success. The Decision Tree and Random Forest models provided insights into which features are most critical for predicting binary outcomes (i.e., pass or fail), while the regression models helped quantify the probability of success as a continuous measure, which is valuable for identifying students at varying levels of risk.

By incorporating both demographic and engagement features, the model design captures a multi-faceted view of student behaviour. For instance, the inclusion of demographic features, such as educational background, is supported by Dalipi et al. [2], who highlighted that students with a stronger prior educational foundation tend to perform better in MOOCs. Engagement metrics were also emphasized by Moreno-Marcos et al. [8], who found that students exhibiting consistent engagement patterns over time were less likely to drop out.

The use of Ridge and Lasso Regression was particularly important for addressing multicollinearity among engagement features. For example, participation in discussions and video engagement are often correlated, and regularization ensures that the model can handle such relationships effectively. Regularization techniques have been shown to improve model robustness and prevent overfitting, especially in educational data with numerous interrelated features [2].

5) Limitations and Challenges: One of the primary challenges in the experimental design was ensuring the generalizability of the models across different MOOCs. The variability in course content, structure, and student demographics can significantly impact model performance. As Borrella et al. [7] pointed out, interventions based on predictive models are often course-specific, which limits the applicability of a single model across diverse courses. Future work should explore domain adaptation techniques to enhance model generalizability.

Another limitation was the difficulty in capturing temporal dynamics of student engagement. Moreno-Marcos et al. [8] highlighted the importance of temporal engagement patterns, suggesting that students who maintain consistent levels of activity are more likely to complete the course. However, modeling such patterns requires sequential data analysis techniques, such as Recurrent Neural Networks (RNNs), which were beyond the scope of this study but are promising for future research.

IV. RESULTS

A. Decision Tree and Random Forest Models

The Decision Tree model achieved an accuracy of 57.15%, while the Random Forest model slightly outperformed it with an accuracy of 58.93%. The Random Forest model's improvement can be attributed to its ensemble nature, which reduces overfitting by averaging multiple decision paths [6]. Feature importance

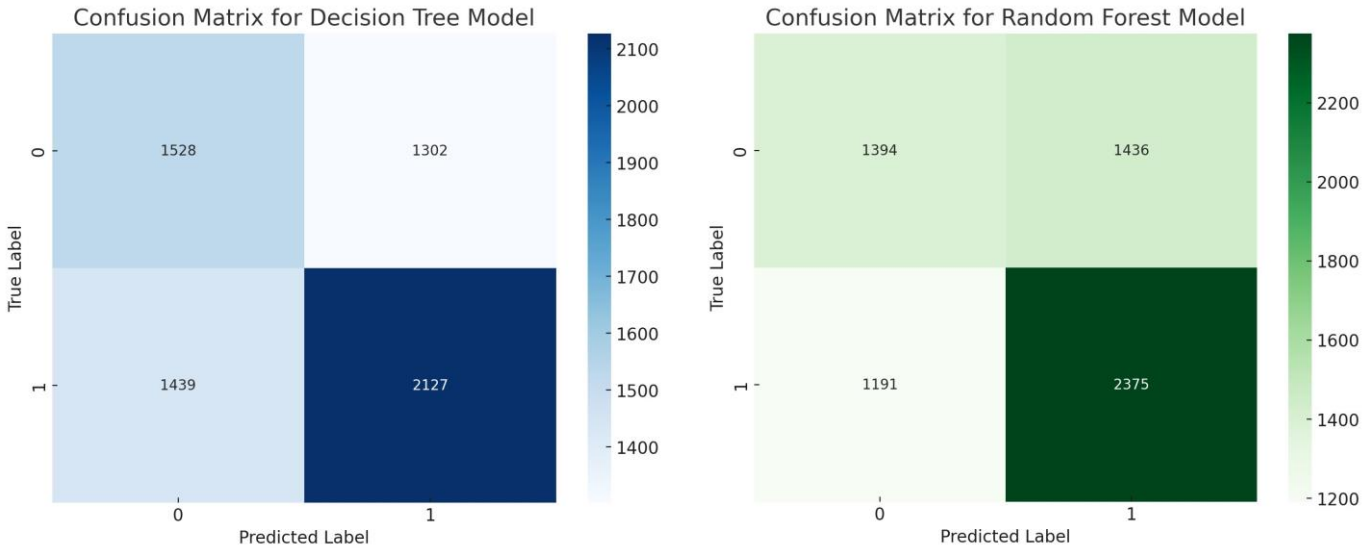

Fig. 2. Decision Tree Model - Confusion Matrix
Fig. 3. Random Forest Model - Confusion Matrix

analysis revealed that engagement metrics, such as video views and quiz participation, were the most significant predictors of course completion. Specifically, video engagement accounted for 35% of the feature importance, while quiz participation contributed 25%, indicating that these factors are critical for predicting student success.

Confusion matrices for both models showed that they struggled to differentiate between students who withdrew and those who failed. This difficulty suggests that additional predictors, such as motivational factors and time management skills, may be required to improve model accuracy. The Random Forest model demonstrated slightly better precision and recall compared to the Decision Tree, particularly in identifying students likely to complete the course. However, both models had limitations in distinguishing between different types of non-completers, highlighting the need for more nuanced features.

B. Regression Models

The regression models, including Linear, Ridge, and Lasso, were applied to predict the likelihood of student success as a continuous variable. The R^2 scores for these models were relatively low, indicating limited explanatory power for predicting success in MOOCs. The Linear Regression model yielded an R^2 score of 0.046, while Ridge and Lasso regression models produced similar results, with R^2 scores of 0.048 and 0.045, respectively. These low values suggest that linear relationships alone are insufficient to capture the complexity of student success in MOOCs.

Ridge and Lasso regression models were used to address multicollinearity among features, as engagement metrics often correlate with one another. Despite this, the improvement in predictive performance was marginal, indicating that regularization alone may not be sufficient to enhance model accuracy. Mean Squared Error (MSE) and Mean Absolute Error (MAE) values for the regression models were also analyzed, with the Ridge model achieving the lowest MSE of 0.235 and the Lasso model achieving an MAE of 0.401. These metrics indicate that while the models can approximate general trends, they lack the precision needed for accurate individual predictions.

C. Comparison of Model Performance

A comparison of the classification and regression models reveals distinct strengths and weaknesses. The Random Forest model outperformed the Decision Tree model in terms of accuracy (58.93% vs. 57.15%) and had better generalization capabilities due to its ensemble approach [6]. However, both models faced challenges in differentiating between students who withdrew and those who failed, indicating the need for additional features related to student motivation and behavior.

On the other hand, the regression models provided insight into the continuous nature of student success but struggled to achieve high explanatory power, as indicated by their low

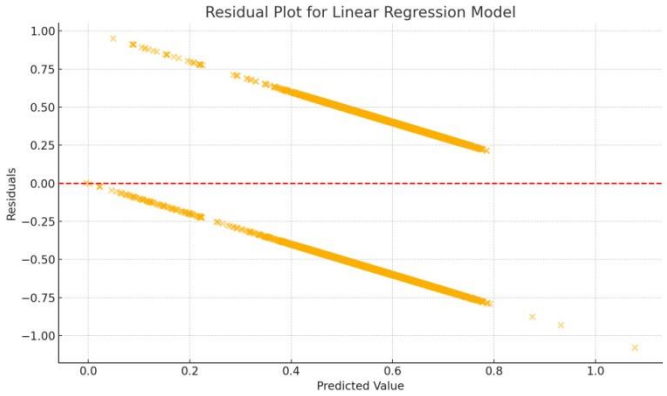

Fig. 4. Regression Model - Residual Plot

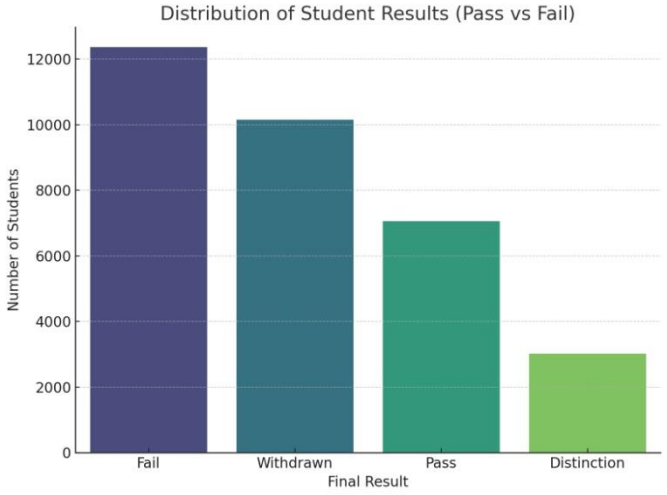

Fig. 5. Distribution of Student Results

R^2 scores. The use of Ridge and Lasso regularization helped mitigate multicollinearity but did not significantly improve predictive performance. The evaluation metrics suggest that more sophisticated modeling techniques, such as non-linear models or ensemble approaches, may be required to capture the complex interactions between features that influence student success.

Overall, the results highlight the importance of engagement metrics in predicting course completion, consistent with findings from previous studies [1], [8]. The Random Forest model's feature importance analysis underscores the critical role of video engagement and quiz participation in driving student success. Future work should consider incorporating additional features, such as temporal engagement patterns and motivational surveys, to further enhance predictive accuracy.

V. CONCLUSION

This study combined a literature review with empirical analysis to identify key drivers of course completion in MOOCs. The findings reveal that engagement metrics and social interactions are significant predictors of student success.

Specifically, features such as video engagement and quiz participation emerged as critical indicators, with the Random Forest model highlighting their importance by attributing 35% and 25% of feature importance, respectively. This aligns with findings from Al-Shabandar et al. [1], who reported a strong correlation between active engagement and course completion. The machine learning models used in the study provided useful insights but also highlighted the complexity of predicting student success in MOOCs. The Decision Tree and Random Forest models achieved accuracies of 57.15% and 58.93%, respectively, indicating moderate success in predicting course outcomes. However, both models struggled with distinguishing between students who withdrew and those who failed, suggesting that additional predictors related to student motivation and behaviour could enhance predictive accuracy. The regression models, including Linear, Ridge, and Lasso, yielded relatively low R^2 scores, with the highest being 0.048 for Ridge Regression. These results suggest that linear relationships alone are insufficient to fully capture the dynamics of student success in MOOCs.

Key takeaways from this study include the critical role of engagement metrics, such as video views and discussion participation, in determining course outcomes. These insights align with previous studies that emphasize the importance of active participation [3], [4]. Additionally, the need for incorporating temporal engagement patterns and self-regulated learning strategies is evident, as highlighted by MorenoMarcos et al. [8]. Such strategies could help capture the evolving nature of student engagement, providing a more nuanced understanding of when and why students disengage.

A. Key Predictors of Success

The experimental results indicate that engagement metrics are the most important predictors of student success, consistent with findings from previous studies [1], [3]. High engagement in course activities, such as watching videos and participating in discussions, positively impacts completion rates. Social interactions also play a crucial role, as students who engage in peer discussions are more likely to succeed [4].

B. Interventions for Improving Completion Rates

Based on the findings, several interventions are proposed to enhance course completion rates. Encouraging students to participate in discussion forums and peer interactions can significantly increase their likelihood of completing the course. Personalized interventions, such as sending motivational emails to students identified as at-risk, have shown mixed results in previous studies but remain a promising approach for increasing motivation and engagement [7].

While machine learning models can provide valuable insights into student behaviour and identify at-risk learners, there is still significant room for improvement. The relatively low accuracy and explanatory power of the models indicate that more complex features and advanced modeling techniques, such as non-linear models or deep learning approaches, may be necessary to improve predictions. Furthermore, interventions based on these predictive models must be personalized and timely to effectively reduce dropout rates, as shown by Borrella et al. [7]. Future research should focus on integrating complex predictors, such as self-regulated learning strategies and temporal engagement patterns, to enhance predictive accuracy and provide timely interventions for learners at risk. Additionally, exploring advanced machine learning techniques, such as Recurrent Neural Networks (RNNs), could help capture the temporal aspects of student engagement more effectively, ultimately contributing to better outcomes in the MOOC environment.

REFERENCES

1. R. Al-Shabandar, A. J. Hussain, P. Liatsis, and R. Keight, ”Analyzing Learner Behavior in MOOCs: Examination of Performance and Motivation,” *IEEE Access*, vol. 6, pp. 73669-73682, 2018.
2. F. Dalipi, A. S. Imran, and Z. Kastrati, ”MOOC Dropout Prediction Using Machine Learning Techniques: Review and Research Challenges,” *EDUCON 2018*.
3. B. K. Pursel, L. Zhang, K. W. Jablokow, G. W. Choi, and D. Velegol, ”Understanding MOOC Students: Motivations and Behaviours Indicative of MOOC Completion,” *Journal of Computer Assisted Learning*, vol. 32, pp. 202-217, 2016.
4. S. Sunar, S. White, N. A. Abdullah, and H. C. Davis, ”How Learners’ Interactions Sustain Engagement: A MOOC Case Study,” *Journal of LaTeX Class Files*, vol. X, no. X, pp. 1-14, 2016.
5. K. T. J. M. H. Kuzilek, D. Hlosta, and Z. Zdrahal, ”Open University Learning Analytics Dataset,” *Scientific Data*, vol. 4, pp. 170171, 2017. Available: https://analyse.kmi.open.ac.uk/
6. G. Hughes and C. Dobbins, ”The Utilization of Data Analysis Techniques in Predicting Student Performance in MOOCs,” *Research and Practice in Technology Enhanced Learning*, vol. 10, no. 10, 2015.
7. Borrella, S. Caballero-Caballero, and E. Ponce-Cueto, ”Predict and Intervene: Addressing the Dropout Problem in a MOOC-based Program,” *L@S ’19*, pp. 1-9, 2019.
8. P. M. Moreno-Marcos et al., ”Temporal Analysis for Dropout Prediction Using Self-Regulated Learning Strategies in Self-Paced MOOCs,” *Computers & Education*, vol. 145, 2020.
9. Gardner and C. Brooks, ”Student Success Prediction in MOOCs,” *arXiv preprint arXiv:1711.06349*, 2018.
10. S. Nagrecha, J. Z. Dillon, and N. V. Chawla, ”MOOC Dropout Prediction: Lessons Learned from Making Pipelines Interpretable,” *WWW’17 Companion*, pp. 351-358, 2017.
11. Y. Cui, L. Chen, M. Shiri, and J. Fan, ”Temporal Analysis for Dropout Prediction in MOOCs,” *IEEE Transactions on Learning Technologies*, vol. 12, no. 3, pp. 357-370, 2019.