Comparing Random Forest and Logistic Regression for Predicting Student Completion in Online University Courses Using Behavioral Data

Muhamad Irfan[1,*], Abdul Sattar[2], Ahmad Sher[3], Muhamad Ijaz[4]

[1]Institute of Banking and Finance, Bahauddin Zakariya University Bosan Road Multan, Multan, Pakistan

[2,3,4]College of Agriculture, Bahauddin Zakariya University Bahadur Sub-Campus, Layyah, Pakistan

ABSTRACT

This paper compares the performance of two machine learning algorithms, Random Forest and Logistic Regression, in predicting student course completion in online university courses using behavioral data. Behavioral data, including interaction logs and submission records, has proven to be crucial in identifying students at risk of non-completion. The study evaluates the models using standard classification metrics such as accuracy, precision, recall, and F1-score, based on real-world data from online courses. Both models demonstrate exceptionally high predictive accuracy, with Logistic Regression achieving perfect classification and Random Forest closely following. While Logistic Regression is favored for its simplicity and interpretability, Random Forest excels in handling complex, non-linear relationships within the data. The analysis of feature importance reveals that student engagement, particularly through viewing and passing course materials, is a strong predictor of course completion. These findings offer significant practical implications for online education, supporting early interventions to enhance student retention. However, limitations such as the absence of certain behavioral data and the linear assumption in Logistic Regression suggest areas for future research. Expanding the dataset to include discussion forums, peer interactions, or additional machine learning models may provide deeper insights into improving student success in online courses.

Keywords Predictive analytics, Random Forest, Logistic Regression, Online education, Student retention

Submitted 8 January 2025
Accepted 17 February 2025
Published 3 March 2025

Corresponding author
Muhamad Irfan,
dr.mirfan@bzu.edu.pk

Additional Information and Declarations can be found on page 16

Introduction

Online education has transformed the learning landscape, making education more accessible to a diverse range of students. This shift has been particularly pronounced during the COVID-19 pandemic, which necessitated a rapid transition to online learning environments. The increased reliance on digital platforms has highlighted the potential of online education to reach learners who may have previously faced barriers to traditional education, such as geographical constraints or scheduling conflicts [1]. The flexibility and convenience of online courses have made them an attractive option for many, contributing to a significant increase in enrollment across various educational institutions.

However, despite the growing importance and accessibility of online education, challenges persist, particularly in student retention and success. Research indicates that retention rates for online courses are significantly lower than those for traditional face-to-face courses, with dropout rates estimated to be 15-20%

How to cite this article: M. Irfan, et. al., "Comparing Random Forest and Logistic Regression for Predicting Student Completion in Online University Courses Using Behavioral Data," *Artif. Intell. Learn.*, vol. 1, no. 1, pp. 1-19, 2025.

higher in online settings [2]. Factors contributing to this phenomenon include a lack of social interaction, insufficient instructor presence, and inadequate student support systems. The absence of a structured learning environment can lead to feelings of isolation among students, which negatively impacts their motivation and engagement [3].

To address these challenges, educational institutions must implement strategies that enhance student satisfaction and retention. Studies have shown that student satisfaction is closely linked to their overall retention in online programs [4]. Factors such as instructor accessibility, timely feedback, and the use of engaging instructional materials play a vital role in fostering a positive online learning experience. Additionally, leveraging learning analytics to identify at-risk students and provide targeted support can significantly improve retention rates. Institutions must also consider students' diverse needs and preferences when designing online courses to ensure that the learning experience is effective and engaging.

Predictive modelling in education has emerged as a crucial tool for identifying students at risk of not completing their courses. By leveraging data analytics, educators and institutions can analyze various factors that contribute to student performance and retention, allowing for timely interventions that can significantly enhance student success rates. This approach is particularly vital in online learning environments, where the lack of physical presence can increase feelings of isolation and disengagement among students [5].

In recent years, the application of machine learning in predictive analytics has gained prominence, particularly in understanding and forecasting complex patterns, such as those observed in digital marketing and e-commerce, where Random Forest and Logistic Regression have been extensively evaluated for accuracy and robustness [6], [7]. This methodological rigor extends into the educational field, where sustainable data mining studies explore key predictors of student success, offering insights into factors that influence academic performance and retention [8], [9]. In educational research, understanding user behaviors through clustering and anomaly detection has proven effective in identifying at-risk students, as clustering techniques help reveal latent patterns that correlate with performance outcomes [10], [11]. Furthermore, advancements in predictive modeling for sequential data, as demonstrated in studies of time series forecasting, underscore the potential for using these models to monitor and predict student engagement over time [12], [13]. Collectively, these studies highlight the growing importance of applying machine learning techniques to educational datasets to derive actionable insights for enhancing student success and retention.

Research indicates that predictive modelling can effectively identify at-risk students by analyzing historical data, including academic performance, engagement metrics, and demographic information. For instance, studies have shown that prior academic performance, such as GPA and previous course outcomes, are strong predictors of future success in online courses. By utilizing these indicators, institutions can develop targeted support strategies that address the specific needs of students who may be struggling, thereby improving retention rates [14].

Moreover, predictive modelling can facilitate early intervention strategies. For example, by monitoring student engagement through learning management systems, educators can identify patterns that may signal a decline in performance or motivation. This allows for timely outreach to students, offering additional resources or support before they reach a critical point of disengagement [15]. Such proactive measures have been shown to positively impact student retention, as they provide the necessary support to help students navigate challenges they may encounter in their coursework.

Additionally, integrating predictive analytics into educational practices can foster a more personalized learning experience. By understanding individual students' unique challenges, educators can tailor their approaches to meet diverse learning needs, thereby enhancing student satisfaction and engagement [16]. This personalized approach not only aids in retention but also improves academic outcomes, as students feel more supported and connected to their learning environment [17].

Predicting student completion in online university courses increasingly relies on analyzing student behavioral data, such as interaction logs and submission records. This data provides valuable insights into student engagement and performance, which are critical for identifying those at risk of not completing their courses. Research has shown that students' engagement characteristics, including their interaction with course materials and participation in discussions, are strong predictors of their success and completion rates in online courses [18]. By analyzing these behavioral patterns, educators can tailor their interventions to support at-risk students better.

For instance, interaction logs can reveal how frequently students log into the course, the time spent on various activities, and their participation in collaborative tasks. These metrics are essential for understanding engagement levels. Studies indicate that higher levels of engagement correlate with improved academic performance and course completion. Conversely, a decline in interaction frequency or a lack of assignment submission can signal potential dropout risks. By monitoring these indicators, educators can intervene early, providing additional support or resources to help students stay on track [19].

Moreover, the use of predictive analytics allows institutions to create models that forecast student outcomes based on historical data. For example, predictive models can analyze submission records to identify patterns that precede course failures, such as late submissions or incomplete assignments. This proactive approach enables educators to reach out to students who exhibit these behaviors, offering personalized support to significantly enhance their chances of successfully completing the course [20].

Integrating learning analytics into educational practices also facilitates a more nuanced understanding of student behavior. By employing data mining techniques, institutions can uncover trends and correlations within large datasets, allowing for more informed decision-making regarding course design and student support services. This data-driven approach not only aids in identifying at-risk students but also enhances the overall learning experience by ensuring that educational resources are allocated effectively.

The existing literature on predicting student completion in online university courses has made significant strides, but there are notable gaps in identifying the most effective algorithms for this task. Much of the research to date has focused on traditional machine learning models such as logistic regression, decision trees, and support vector machines, which have shown success in some contexts. However, there has been limited comparative analysis of more advanced ensemble-based algorithms like Random Forest, which may offer greater accuracy and robustness in modeling the complex interactions present in student behavioral data. This gap is particularly relevant as online learning environments continue to expand, requiring models that can better handle the non-linear relationships and high-dimensional data often seen in these settings.

Additionally, a substantial portion of current studies emphasizes only basic engagement metrics—such as login frequency or time spent on course materials—while overlooking the potential for combining these with richer data sources like submission records, assignment completion rates, or interaction patterns. The lack of integrated behavioral data in many predictive models limits their ability to capture the nuances of student engagement and performance fully. Studies incorporating multiple types of student behavior, including submission logs, could offer more precise predictions of student outcomes. However, this area remains underexplored, allowing further research into holistic models that merge various behavioral datasets to enhance prediction accuracy.

Furthermore, while some research has explored the use of predictive analytics in educational settings, direct algorithm comparisons are scarce across different learning environments, particularly in large-scale online courses. Most studies either examine small datasets or focus on traditional classroom settings, failing to address the unique challenges posed by online education [21]. This gap highlights the need for more comprehensive evaluations of algorithm performance in diverse online education contexts, particularly with large, heterogeneous datasets typical of university-level online courses. The absence of such evaluations makes it difficult to determine which models generalize best to these environments, underscoring the need for more research in this area to ensure robust, scalable solutions for predicting student completion.

The primary objective of this study is to compare the effectiveness of two widely used machine learning algorithms—Random Forest and Logistic Regression—in predicting student course completion in online university environments. Both algorithms have been employed in educational data mining, but their comparative performance remains underexplored when applied to behavioral data, such as interaction logs and submission records. Understanding how these models predict course completion can provide valuable insights for educational institutions seeking to improve student retention through data-driven interventions.

Additionally, the findings of this study contribute to the broader educational technology landscape by providing a scalable solution that can be adapted across various learning management systems and educational platforms. The insights gained from comparing Random Forest and Logistic Regression can inform the development of predictive tools integrated into these platforms, making it easier for educators to monitor student progress and intervene when necessary. This research thus holds the potential to significantly enhance the personalization and effectiveness of online education, ensuring that data-driven methodologies are better utilized to meet the diverse needs of students in the digital learning environment.

Literature Review

Overview of Predictive Analytics in Education

Predictive analytics in education, especially in online learning environments, has become a central focus for researchers aiming to enhance student retention and academic success. This field involves using machine learning (ML) algorithms and statistical models to analyze student data, which helps institutions forecast academic outcomes and intervene before students disengage or fail to complete their courses. The growing availability of student interaction data, such as participation in learning management systems (LMS) and submission patterns, has further fueled the development of predictive models designed to identify at-risk students. These models enable educators to provide timely support, improving both retention and student outcomes.

Various machine learning models have been applied to predictive analytics in education. Neural networks, for instance, are frequently praised for their ability to process large datasets and capture non-linear relationships within student behavior, leading to high prediction accuracy in student performance. Other models, such as logistic regression and decision trees, remain popular due to their interpretability and ease of implementation. However, their predictive power may be less robust when dealing with complex datasets [22]. Moreover, combining multiple data sources—ranging from demographic data to behavioral metrics—has enhanced the predictive capabilities of these models, allowing them to account for a wider range of factors influencing student success [23].

The application of predictive analytics in online learning has yielded valuable insights into student engagement and dropout rates. For example, studies have demonstrated that clickstream data, which tracks how students interact with course materials, can be used to predict dropout rates with considerable accuracy, particularly in Massive Open Online Courses (MOOCs). Research [24] also identified indicators such as discussion forum participation and peer evaluation quality as significant predictors of academic success. Early warning systems (EWS) that utilize predictive models have been developed to alert educators when students risk failing, allowing for timely interventions. These systems often display data through dashboards, which make predictions and student engagement metrics easily accessible for teachers and administrators.

Despite the advancements in predictive analytics, challenges remain regarding these technologies' ethical use and actual impact. The study [25] noted that while there is a theoretical basis for predictive analytics, empirical evidence showing substantial improvements in student outcomes is still limited. Furthermore, concerns about data privacy and potential biases in the algorithms require careful consideration, as they can disproportionately affect underrepresented student populations. Addressing these challenges is crucial for ensuring that predictive models are accurate, equitable, and beneficial to all students. As the field progresses, the ethical implications and transparency in predictive modeling need to be central to developing and implementing these systems in educational settings.

Behavioral Data in Predicting Student Success

Student interaction data, such as clickstream data, submission logs, and event-based data, has gained increasing prominence in educational data mining. This type of behavioral data is crucial for understanding how students engage with online learning materials and predicting their academic success. Researchers have focused on leveraging these digital footprints to build models that can accurately forecast outcomes such as course completion, academic performance, and retention. Clickstream data, which tracks student navigation through online course platforms, is particularly valuable for identifying engagement patterns that correlate with success or risk of failure. Submission logs detailing the frequency and timing of task completions and event-based data, such as attempts to engage with specific activities, provide a comprehensive view of student behavior, making these data points integral to predicting success metrics.

Several studies have explored the predictive power of this behavioral data.

Research [26] demonstrated the efficacy of integrating clickstream data with content-based resources in predicting student performance. In their study, the combination of interaction data from a MOOC and traditional content-related data significantly improved the accuracy of predictions. The research highlighted the importance of understanding how students engage with the content and how much they learn. Similarly, research [27] a systematic literature review was conducted that underscored the increasing need for data-driven, evidence-based decision-making in education. Their work pointed to the opportunities presented by large-scale behavioral data in crafting more personalized and effective interventions to improve student outcomes.

Random Forest and Logistic Regression in Educational Data Mining

In educational data mining, both Random Forest and Logistic Regression have been extensively applied for classification tasks, each offering distinct advantages and limitations. These models have been pivotal in predicting student outcomes, such as course completion, performance, and dropout risk, by leveraging various behavioral and academic datasets. This section reviews key studies that have applied these models in educational contexts, emphasizing their strengths and challenges in handling classification tasks.

Random Forest, an ensemble learning method that constructs multiple decision trees and outputs the mode of their predictions, has demonstrated effectiveness in educational settings. The research [28] applied the Random Forest algorithm to predict student dropout rates, highlighting its capability to manage imbalanced datasets and deliver accurate predictions. One of the main advantages noted in their study is Random Forest's ability to handle many input variables without discarding irrelevant factors—a significant benefit in education, where diverse, interconnected variables often influence student success. Research [29] also applied Random Forest to analyze the performance of engineering students, concluding that the model was particularly useful in capturing complex interactions between predictors, such as early academic achievements and graduation outcomes.

While Random Forest is known for its high accuracy and the ability to measure variable importance, it has certain drawbacks. Its computational intensity and the need for careful tuning of hyperparameters to avoid overfitting are commonly cited limitations, especially with smaller datasets. Additionally, the model's complexity can make it challenging to provide clear interpretations of how individual predictors relate to student outcomes, which can be a critical factor in educational research where stakeholders value interpretability.

Logistic Regression, a statistical model used for binary classification, has also been widely employed in educational data mining. This model is valued for its simplicity and interpretability, making it suitable for understanding the factors influencing student success.

Despite its advantages, Logistic Regression assumes a linear relationship between the log odds of the outcome and the predictor variables, which can limit its performance when dealing with non-linear relationships. This constraint makes the model less flexible than Random Forest when working with complex educational datasets. Furthermore, Logistic Regression may face difficulties in handling high-dimensional data where the number of predictors exceeds the number of observations, potentially leading to overfitting.

Both Random Forest and Logistic Regression have demonstrated their utility in educational data mining, yet their applications depend heavily on the specific context and research goals. Random Forest is highly effective for large-scale datasets with numerous predictors and complex interactions, making it ideal for projects where accuracy and robustness are prioritized. In contrast, Logistic Regression's simplicity and interpretability make it a preferred choice in educational studies where understanding the relationship between variables is essential. Given these trade-offs, the choice between the two models should be guided by the nature of the dataset and the need for either interpretability or handling complexity.

Summary and Research Gap

The literature review highlights the effectiveness of both Random Forest and Logistic Regression in educational data mining, particularly for classification tasks involving student outcomes. Random Forest is valued for its ability to handle complex interactions within large datasets, providing robust predictions while mitigating overfitting through its ensemble nature. On the other hand, Logistic Regression offers a simpler, more interpretable model, often favored for its transparency in understanding the relationships between predictor variables and student success. Both models have demonstrated utility in educational contexts, yet their comparative effectiveness in predicting student completion in online university courses, specifically using behavioral data, remains underexplored.

The need to compare these two algorithms arises from the unique challenges online education poses. Behavioral data, including clickstream data, submission logs, and other interaction-based metrics, provides a wealth of information about how students engage with their coursework. However, accurately predicting course completion from this data requires models that can effectively capture and interpret these patterns. Random Forest’s ability to model non-linear relationships and feature interactions contrasts with Logistic Regression’s straightforward, interpretable approach, compelling the need to determine which model performs better under the specific conditions of online learning environments.

While existing studies have applied these algorithms individually to educational datasets, there is a significant gap in research comparing their performance specifically in the context of student behavioral data in online education. As online learning becomes increasingly prevalent, understanding which model offers superior predictive power is essential for educators and administrators seeking to enhance student retention through data-driven interventions. This study addresses this gap by directly comparing random forest and logistic regression, contributing to the growing field of educational predictive analytics.

Method

The research method for this study consists of several steps to ensure a comprehensive and accurate analysis. The flowchart in Figure 1 outlines the detailed steps of the research method.

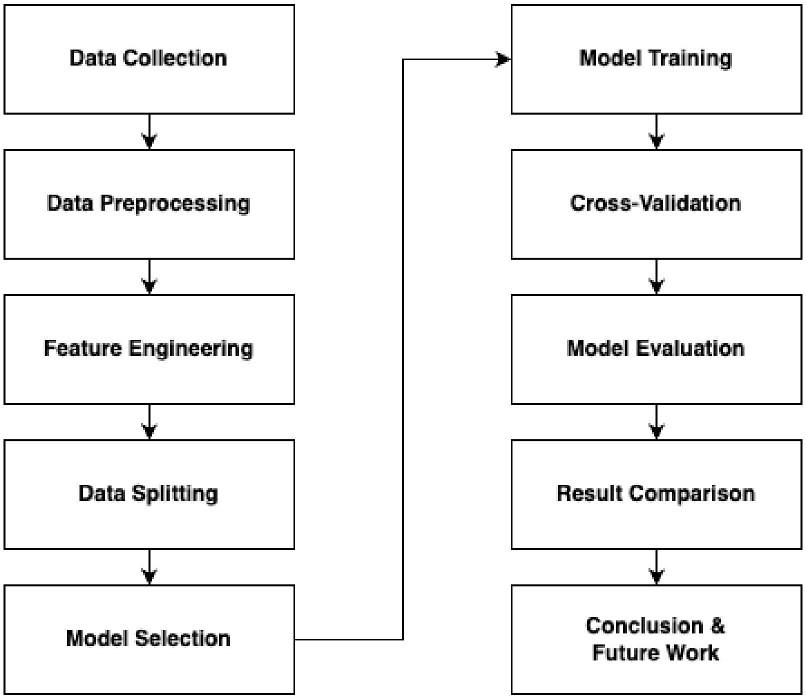

Figure 1 Research Method Flowchart

Dataset Description

This study utilizes two primary datasets, event_data_train.csv and submissions_data_train.csv, to analyze student behavior and predict course completion in an online university setting. These datasets provide a detailed view of student interactions and submission patterns, which serve as key indicators for predicting whether students successfully complete their courses. Both datasets offer granular information about how students engage with the online learning platform and their progress through practical tasks, forming the basis for feature engineering and model training.

The event_data_train.csv dataset contains records of various student interactions with course steps, which are represented as events. Key features in this dataset include \`step_id\`, \`timestamp\`, \`user_id\`, and \`action\`. The \`action\` column logs different types of student activities, such as \`viewed\` (viewing a course step), \`started_attempt\` (beginning an attempt to solve a problem), \`passed\` (successfully completing a practical task), and \`discovered\` (transitioning to a new course step). Each event provides valuable insight into student engagement, allowing for the measurement of activity frequency, engagement duration, and interaction patterns. These behavioral metrics are essential for identifying trends in student participation, which can be linked to course completion rates.

The submissions_data_train.csv dataset captures detailed submission records for practical tasks, providing complementary information to the interaction data. The dataset includes \`step_id\`, \`timestamp\`, \`user_id\`, and \`submission_status\`. The \`submission_status\` field indicates whether a student's submission was \`correct\` or \`wrong\`, offering a clear measure of student performance on assignments. By tracking the timing and outcomes of these submissions, this dataset allows for an analysis of how student efforts in practical assignments impact their likelihood of completing the course. For instance, patterns of frequent incorrect submissions or delayed task completions may indicate students who are at risk of not finishing the course.

Together, these datasets form a comprehensive picture of student engagement and performance. The interaction logs from event_data_train.csv provide insights into students' behavioral patterns, while the submission records from submissions_data_train.csv offer a performance-based perspective on student progress. By integrating these two datasets, the study aims to create predictive models that leverage both engagement and performance features to accurately predict course completion. These datasets enable the extraction of key behavioral features, such as the total number of events per student, the ratio of correct to incorrect submissions, and the frequency of engagement with specific course steps, all of which are critical to the analysis.

Exploratory Data Analysis (EDA)

The initial step in the exploratory data analysis involved cleaning the datasets to ensure consistency and usability. The event_data_train.csv and submissions_data_train.csv datasets contained no missing values in their primary columns, such as `step_id`, `timestamp`, `user_id`, and `action`. However, the `timestamp` fields were stored in Unix format, which was converted into readable date-time formats for easier analysis of time-based patterns in student activity. The `action` and `submission_status` columns, which contained categorical variables representing different types of student actions and submission outcomes, were encoded for modeling purposes. Specifically, actions such as `viewed`, `passed`, `discovered`, and `started_attempt` were encoded into numerical labels, and the submission statuses `correct` and `wrong` were similarly transformed. To gain insight into student behavior, several visualizations were generated.

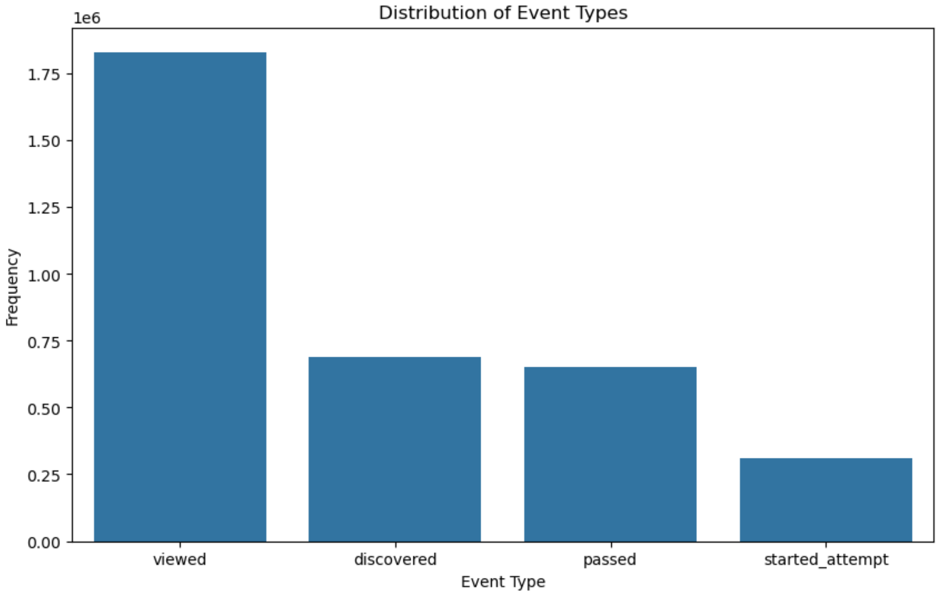

Figure 2 Distribution of Event Types

Figure 2 illustrates the distribution of different event types in an online learning dataset, with the x-axis representing various student actions and the y-axis displaying their frequencies. The most common event type is "viewed," with approximately 1.8 million occurrences, indicating that a large portion of student interactions involve passively viewing course content. This suggests that students are primarily consuming material without necessarily engaging in more active learning processes. The "discovered" event type, which represents students transitioning to new steps or topics, is the second most frequent event, with fewer than 1 million instances. This reflects moderate engagement as students explore different sections of the course. The "passed" event, which represents students successfully completing tasks, shows a lower frequency than both viewing and discovering events. This implies that fewer students are completing practical tasks, highlighting a potential drop-off in engagement when it comes to active participation. Finally, "started_attempt" has the lowest frequency among the event types, indicating that only a small subset of students are attempting tasks in the first place. The disparity between viewing content and attempting or completing tasks suggests that while many students passively engage with the course, fewer take the necessary steps to actively participate and succeed. This highlights a potential challenge in moving students from passive engagement to active learning, which could impact overall course completion rates.

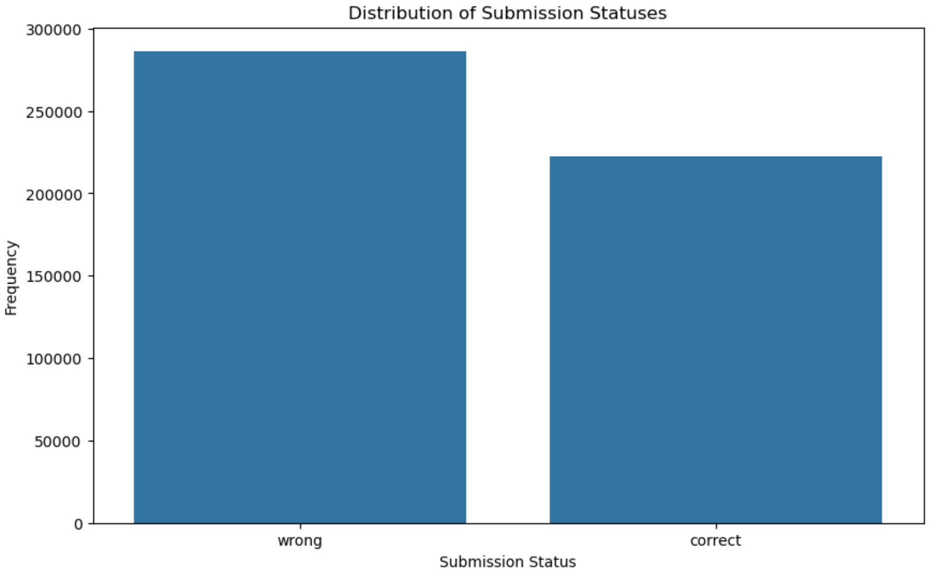

Figure 3 Distribution of Submission Statuses

Figure 3 shows the distribution of submission statuses in the online learning dataset, with the x-axis representing the different submission outcomes, and the y-axis showing the frequency of each outcome. The chart highlights two submission statuses: "wrong" and "correct." The "wrong" submission status is more frequent, with nearly 290,000 occurrences, while the "correct" submission status has fewer occurrences, totaling just under 250,000. This indicates that students, on average, submit more incorrect solutions before achieving a correct answer. The difference in frequency between wrong and correct submissions suggests that many students face challenges when attempting tasks, leading to a higher proportion of incorrect submissions. This pattern may reflect difficulties in understanding course materials or the complexity of tasks, requiring students to make multiple attempts before succeeding. The higher frequency of incorrect submissions could be a useful indicator for identifying areas where students struggle, highlighting the potential need for additional instructional support or feedback to help students overcome obstacles and improve task completion rates.

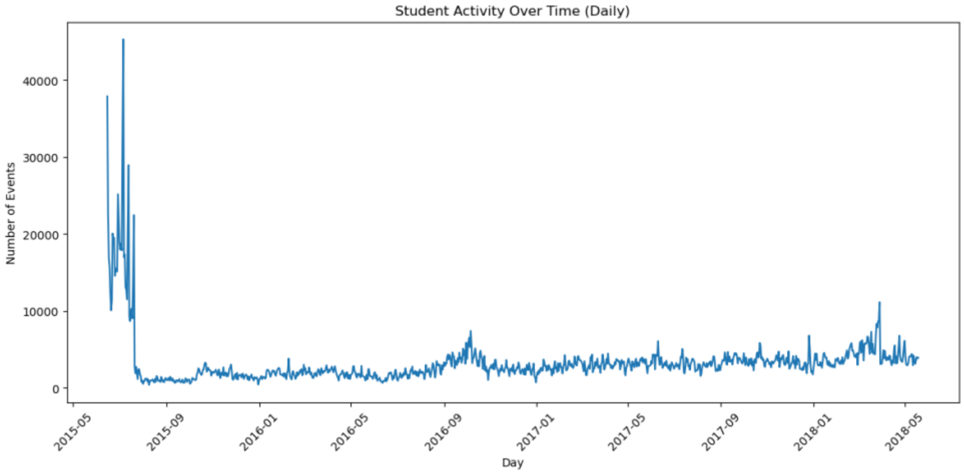

Figure 4 Student Activity Over Time (Daily)

Figure 4 represents student activity over time, where the x-axis shows the timeline (from mid-2015 to mid-2018), and the y-axis shows the number of events (student interactions) occurring daily. At the beginning of the time series, there is a significant spike in student activity, with the number of events reaching over 40,000 daily. This likely reflects a strong initial engagement at the start of the course or program, possibly due to onboarding or the release of critical course materials. However, after this initial surge, the activity quickly drops off and stabilizes at a lower baseline. Between 2016 and 2018, the activity remains relatively stable, with smaller peaks and troughs indicating periodic increases in engagement, possibly corresponding to specific milestones, assignments, or exams. Toward the end of the timeline, there is a noticeable increase in activity once again, possibly due to final assessments or project deadlines. This pattern of high initial activity followed by a period of relative inactivity is typical in online learning environments, where many students engage early and fall off in participation over time.

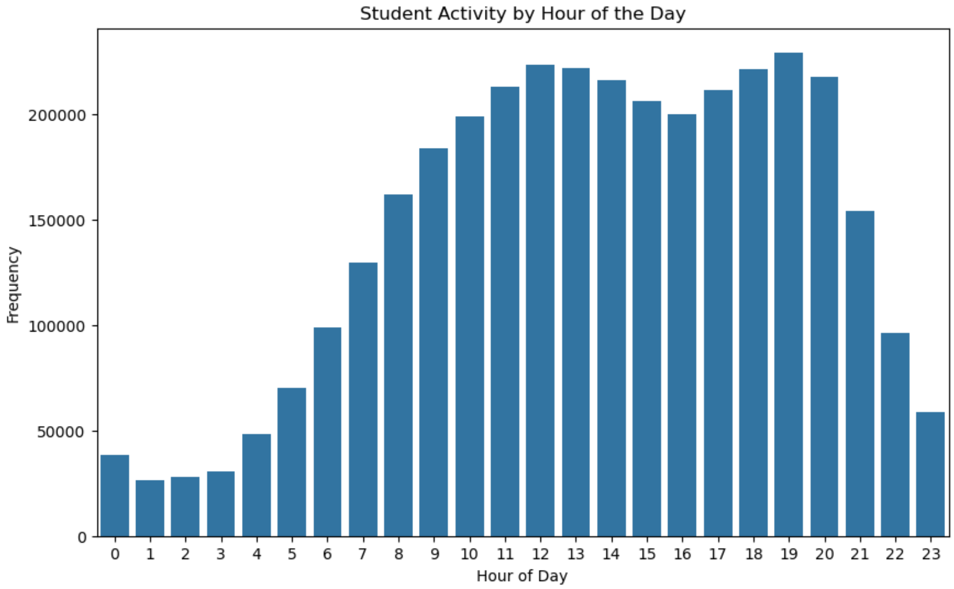

Figure 5 Student Activity by Hour of the Day

Figure 5 shows student activity by the hour of the day, where the x-axis represents the 24-hour day, and the y-axis represents the frequency of student interactions. The chart reveals that student activity increases steadily throughout the day, starting around 5:00 a.m. Activity peaks between 10:00 a.m. and 7:00 p.m., with the highest frequency of interactions occurring around noon and early evening. After 8:00 p.m., activity starts to decline sharply, with the lowest levels of engagement occurring between 1:00 a.m. and 5:00 a.m. This distribution suggests that students tend to engage with their courses during typical working or study hours, with a preference for late mornings and afternoons. There is significantly less engagement during nighttime, particularly in the early morning hours, which is consistent with typical online learning patterns where students manage their coursework around their daily schedules.

Descriptive statistics for the key features in both datasets helped summarize the overall trends in student behavior and performance. In the event_data_train.csv dataset, the number of unique `step_id` values was 198, and the `user_id` column indicated that 19,234 unique students participated in the online courses. The mean and median timestamps reflected a broad range of activity, with students interacting with course steps from mid-2015 to mid-2018. The number of `viewed` events dominated, while `passed` and `started_attempt` actions were less frequent, highlighting the passive nature of many students' engagement. Similarly, the submissions_data_train.csv dataset showed that the `wrong` submission status occurred more frequently than the `correct` status, with 286,399 incorrect submissions compared to 222,705 correct ones. The mean submission time, measured by the `timestamp` column, aligned with the events dataset, indicating that both submission and event activities followed similar temporal patterns. Additionally, the distribution of submission attempts across the 9,940 unique students further emphasized the challenges many students faced when attempting to complete practical tasks, as reflected in the higher frequency of incorrect submissions.

Feature Engineering

To enable the prediction of student course completion, key features were derived from the available event and submission data. For each student, we calculated the total counts of event types such as "viewed," "discovered," "started_attempt," and "passed." This allowed us to quantify the level of interaction students had with course materials. In addition to these counts, we computed the passed/viewed ratio, a critical feature that reveals the proportion of viewed steps that were successfully completed by the student. This ratio is an important indicator of student progress. Furthermore, submission data was used to create features capturing the count of correct and wrong submissions per student. These features help differentiate between students who persist and improve through attempts and those who struggle to submit correct solutions. The combination of interaction-based features and submission-based features provided a holistic view of student engagement and task completion efforts, laying the groundwork for training predictive models.

Model Training

Two models, Random Forest and Logistic Regression, were chosen for comparison in this study. Random Forest was selected due to its robustness in handling high-dimensional data and its ability to model complex interactions between variables. It has been frequently used in educational data mining for its high accuracy in classification tasks. Logistic Regression, on the other hand, was chosen for its simplicity and interpretability, making it ideal for situations where the relationship between predictors and the outcome needs to be easily understood. The comparison between these two models aimed to determine which method is more effective for predicting course completion using behavioral data. The dataset was divided into training and testing sets using an 80-20 split to ensure that the model could generalize well on unseen data. The training set was used to build the models, while the testing set was reserved for evaluating their performance. Cross-validation with five folds was also applied to further assess the consistency of the models across different subsets of data, ensuring that the results were not dependent on any particular random split. The performance of both models was evaluated using several standard classification metrics: accuracy, precision, recall, and the F1-score. Accuracy measures the overall correctness of the model’s predictions. Precision is the proportion of positive predictions that are correct, while recall indicates how well the model identifies true positives. The F1-score provides a balance between precision and recall, making it useful for datasets where class distribution may be imbalanced. The results showed that both models performed well, with near-perfect scores across all metrics, particularly in terms of recall and F1-score, suggesting that both Random Forest and Logistic Regression are highly effective in predicting student course completion based on behavioral data.

Result and Discussion

Model Performance

The performance of the Random Forest and Logistic Regression models was evaluated using several classification metrics, including accuracy, precision, recall, and the F1-score. Both models demonstrated exceptionally high performance across all metrics when predicting student completion in online university courses based on behavioral data, shown in Table 1. The Random

Forest model achieved an accuracy of 0.9995, with a precision of 1.0 and a recall of 0.9991. The F1-score for this model was calculated as 0.9995, reflecting its balanced performance in identifying both students who completed their courses and those who did not. Similarly, the Logistic Regression model performed with an accuracy of 1.0, yielding perfect scores for precision, recall, and F1-score. These results suggest that Logistic Regression accurately classified all instances in the test set without error, which might be attributed to the strong linear relationships present in the dataset. In terms of cross-validation, Random Forest yielded a slightly lower variance in its metrics, with an average accuracy of 0.9999 (± 0.0001), while Logistic Regression also performed consistently, showing similar variance in its accuracy (0.9999 ± 0.0001).

Table 1. Model Performance Results

Model	Accuracy	Precision	Recall	F1-Score
Random Forest	0.9995	1	0.9991	0.9995
Logistic Regression	1	1	1	1

Both models performed remarkably well in this classification task, but the slight differences in recall and F1-score for Random Forest highlight its ability to handle more nuanced cases, particularly in large datasets with complex relationships. While Logistic Regression offers simplicity and interpretability, Random Forest may provide additional robustness, particularly in cases with non-linear patterns.

Discussion

These results suggest that both Random Forest and Logistic Regression are highly effective for predicting student course completion based on behavioral data, with minimal differences in their overall performance. Logistic Regression's perfect results, however, should be interpreted with caution, as its linear assumption may not always capture underlying complexities in more varied datasets. On the other hand, Random Forest, through its ensemble nature, is better suited to detect intricate patterns and interactions between the features, making it a more versatile choice in diverse educational settings.

Given the near-identical performance of both models, the choice between Random Forest and Logistic Regression in educational data mining applications might depend on specific project requirements. For example, if interpretability and simplicity are key concerns, Logistic Regression is an excellent choice. However, if the dataset exhibits complex relationships and interactions that need to be captured, Random Forest may offer superior flexibility without compromising accuracy.

The comparison between Random Forest and Logistic Regression models for predicting student course completion revealed that both algorithms performed with extremely high accuracy. Logistic Regression achieved a perfect accuracy score, classifying all instances correctly. Random Forest, while slightly lower in terms of recall (0.9991), demonstrated robust performance across all metrics, including an F1-score of 0.9995. Despite the near-identical results in predictive accuracy, Logistic Regression’s simplicity and linearity contrasted with Random Forest’s more complex decision tree ensemble approach. The latter’s strength lies in its ability to handle large datasets with intricate interactions, which suggests that Random Forest may be better suited to handle more complicated educational datasets in future research.

Random Forest’s inherent ability to measure feature importance allowed for deeper insights into the variables that most significantly impacted student completion. Among the most influential features were "viewed" and "passed" events, indicating that students who consistently interacted with course materials and successfully completed steps were more likely to finish the course. Additionally, the ratio of "passed" to "viewed" events played a key role in prediction, demonstrating that students who frequently engaged with materials but did not pass them were at greater risk of not completing the course. Logistic Regression, although not providing feature importance directly, showed similar tendencies in its coefficient analysis, with higher engagement and successful task completion correlating with higher probabilities of course completion.

The findings from both models suggest critical insights for educators seeking to improve student retention in online courses. High engagement with learning materials (viewed events) and successful progression through the course (passed events) are crucial indicators of student success. Identifying students who engage with content but struggle to pass can help educators intervene early, providing additional support to those at risk of not completing their courses. The ability to leverage behavioral data to predict student outcomes offers significant potential for improving personalized learning experiences, allowing institutions to tailor interventions and resources to individual student needs based on their interaction patterns.

From a practical perspective, these results can assist educational institutions in developing more effective retention strategies. By using predictive models like Random Forest or Logistic Regression, administrators can monitor student activity in real-time and intervene when patterns suggest a risk of dropout. This approach not only enhances course retention rates but also promotes a more proactive and supportive learning environment. For example, targeted outreach campaigns could be designed for students with low passed/viewed ratios, or additional resources might be directed toward those consistently viewing but not progressing in course steps. In conclusion, applying these predictive models can provide actionable insights that ultimately help improve student retention in online university courses.

Conclusion

The comparison between Random Forest and Logistic Regression for predicting student course completion in online university courses revealed that both models achieved exceptional performance. Logistic Regression slightly outperformed Random Forest, achieving a perfect accuracy, precision, recall, and F1-score across all evaluation metrics. Random Forest followed closely with similarly high scores but showed a minor drop in recall compared to Logistic Regression. Both models demonstrated their ability to effectively predict course completion based on student behavioral data, although the simplicity and interpretability of Logistic Regression make it an ideal model in many educational contexts. The practical implications of these findings suggest that predictive analytics can play a transformative role in online education. By leveraging student interaction data, such as event logs and submission records, educators can proactively identify at-risk students who might struggle to complete their courses. Early identification through predictive modeling allows institutions to tailor interventions, such as offering additional support, personalized feedback, or providing supplementary resources. This proactive approach enhances student success and retention, helping to minimize dropout rates in online learning environments. Moreover, applying such models contributes to data-driven decision-making in education, which ultimately improves the overall quality of online learning experiences.

Despite the strong results, this study faced several limitations. First, the analysis was based on a specific set of behavioral data from events and submission logs, which may not capture the full complexity of student learning experiences. The models, while highly accurate, were also limited by the nature of the features available, such as lacking data on student participation in forums, peer interactions, or instructor feedback. Additionally, while Random Forest is known for managing complex datasets, it can be computationally intensive, and Logistic Regression, despite its simplicity, assumes a linear relationship between features and the outcome, which may not hold true in all contexts. Potential biases may have also emerged due to the focus on a single dataset from online courses, which could affect the generalizability of the findings. Future research could address these limitations by incorporating additional data sources, such as discussion forum participation, peer evaluations, or interaction data with instructors. Testing other machine learning algorithms, such as Support Vector Machines or deep learning models, might also yield new insights into the predictive power of different approaches. Expanding the study to other types of online courses, such as MOOCs or blended learning environments, would further validate the findings and increase their applicability. Additionally, research focusing on developing hybrid models that combine the strengths of both Random Forest and Logistic Regression could offer more robust solutions for predicting student success in online education.

Declarations

Author Contributions

Conceptualization: M.I.; Methodology: A.S.; Software: M.I.; Validation: M.I.; Formal Analysis: A.S.; Investigation: M.I.; Resources: M.I.; Data Curation: A.S.; Writing Original Draft Preparation: A.S.; Writing Review and Editing: A.S.; Visualization: M.I.; All authors have read and agreed to the published version of the manuscript.

Data Availability Statement

The data presented in this study are available on request from the corresponding author.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] W. Ali, “Online and Remote Learning in Higher Education Institutes: A Necessity in Light of COVID-19 Pandemic,” *High. Educ. Stud.*, vol. 10, no. 3, p. 16, 2020, doi: 10.5539/hes.v10n3p16.

[2] E. M. Leeds, S. M. Campbell, H. Baker, R. Ali, D. Brawley, and J. D. C. Crisp, “The Impact of Student Retention Strategies: An Empirical Study,” *Int. J. Manag. Educ.*, vol. 7, no. 1/2, p. 22, 2013, doi: 10.1504/ijmie.2013.050812.

[3] P. S. Muljana and T. Luo, “Factors Contributing to Student Retention in Online Learning and Recommended Strategies for Improvement: A Systematic Literature Review,” *J. Inf. Technol. Educ. Res.*, vol. 18, pp. 019–057, 2019, doi: 10.28945/4182.

[4] M. T. Cole, D. J. Shelley, and L. B. Swartz, “Online Instruction, E-Learning, and Student Satisfaction: A Three Year Study,” *Int. Rev. Res. Open Distrib. Learn.*, vol. 15, no. 6, 2014, doi: 10.19173/irrodl.v15i6.1748.

[5] S. James, K. Swan, and C. Daston, “Retention, Progression and the Taking of Online Courses,” *Online Learn.*, vol. 20, no. 2, 2015, doi: 10.24059/olj.v20i2.780.

[6] Q. Siddique, “Comparative Analysis of Sentiment Classification Techniques on Flipkart Product Reviews: A Study Using Logistic Regression, SVC, Random Forest, and Gradient …,” *J. Digit. Mark. Digit. Curr.*, no. Query date: 2024-10-12 10:42:49, 2024, [Online]. Available: http://jdmdc.com/index.php/JDMDC/article/view/4

[7] B. Hayadi and I. E. Emary, “Predicting Campaign ROI Using Decision Trees and Random Forests in Digital Marketing,” *… Digit. Mark. Digit. Curr.*, no. Query date: 2024-10-12 10:42:49, 2024, [Online]. Available: http://jdmdc.com/index.php/JDMDC/article/view/5

[8] M. Murnawan, S. Lestari, R. Samihardjo, and ..., “Sustainable Educational Data Mining Studies: Identifying Key Factors and Techniques for Predicting Student Academic Performance,” *J. Appl. Data …*, no. Query date: 2024-10-12 10:59:38, 2024, [Online]. Available: http://www.bright-journal.org/Journal/index.php/JADS/article/view/347

[9] H. Sukmana, Y. Durachman, A. Amri, and ..., “Comparative Analysis of SVM and RF Algorithms for Tsunami Prediction: A Performance Evaluation Study,” *J. Appl. Data …*, no. Query date: 2024-10-12 10:59:38, 2024, [Online]. Available: http://bright-journal.org/Journal/index.php/JADS/article/view/159

[10] T. Hariguna and A. Al-Rawahna, “Unsupervised Anomaly Detection in Digital Currency Trading: A Clustering and Density-Based Approach Using Bitcoin Data,” *J. Curr. Res. Blockchain*, no. Query date: 2024-10-12 10:44:40, 2024, [Online]. Available: http://jcrb.net/index.php/Journal/article/view/12

[11] D. Sugianto and A. R. Hananto, “Geospatial Analysis of Virtual Property Prices Distributions and Clustering,” *Int. J. Res. Metaverese*, vol. 1, no. 2, Art. no. 2, Sep. 2024, doi: 10.47738/ijrm.v1i2.10.

[12] S. A. Ghaffar and W. C. Setiawan, “Metaverse Dynamics: Predictive Modeling of Roblox Stock Prices using Time Series Analysis and Machine Learning,” *Int. J. Res. Metaverese*, vol. 1, no. 1, Art. no. 1, Jun. 2024, doi: 10.47738/ijrm.v1i1.6.

[13] A. Hananto, “Assessing the Efficacy of Convolutional Neural Networks in Recognizing Handwritten Digits,” *J. Curr. Res. Blockchain*, no. Query date: 2024-10-12 10:44:40, 2024.

[14] B. Arnbjörnsdóttir, “Determining Factors in Student Retention in Online Courses,” pp. 116–121, 2017, doi: 10.14705/rpnet.2017.eurocall2017.699.

[15] P. Shea and T. Bidjerano, “A National Study of Differences Between Distance and Non-Distance Community College Students in Time to First Associate Degree

Attainment, Transfer, and Dropout," *Online Learn.*, vol. 20, no. 3, 2016, doi: 10.24059/olj.v20i3.984.

[16] M. Mattila and A. Mattila, "Dimensions of Likelihood to Recommend an Online Course," 2016, doi: 10.18638/hassacc.2016.4.1.210.

[17] J. C. Drew *et al.*, "Development of a Distance Education Program by a Land-Grant University Augments the 2-Year to 4-Year STEM Pipeline and Increases Diversity in STEM," *Plos One*, vol. 10, no. 4, p. e0119548, 2015, doi: 10.1371/journal.pone.0119548.

[18] T. Soffer and A. Cohen, "Students' Engagement Characteristics Predict Success and Completion of Online Courses," *J. Comput. Assist. Learn.*, vol. 35, no. 3, pp. 378–389, 2019, doi: 10.1111/jcal.12340.

[19] C. Wladis, K. M. Conway, and A. C. Hachey, "Assessing Readiness for Online Education — Research Models for Identifying Students at Risk," *Online Learn.*, vol. 20, no. 3, 2016, doi: 10.24059/olj.v20i3.980.

[20] Y. Ma, S. Lee, J. Lu, Y. Pan, and J. Sun, "Construction of Data-Driven Performance Digital Twin for a Real-World Gas Turbine Anomaly Detection Considering Uncertainty," *Sensors*, vol. 23, no. 15, p. 6660, 2023, doi: 10.3390/s23156660.

[21] K. A. Doorn, V. Békés, and R. A. Zweig, "Clinical Psychology Graduate Students: Lessons Learned From a Sudden Transition to Online Education.," *Scholarsh. Teach. Learn. Psychol.*, vol. 8, no. 4, pp. 279–294, 2022, doi: 10.1037/stl0000317.

[22] F. Chen and Y. Cui, "Utilizing Student Time Series Behaviour in Learning Management Systems for Early Prediction of Course Performance," *J. Learn. Anal.*, vol. 7, no. 2, pp. 1–17, 2020, doi: 10.18608/jla.2020.72.1.

[23] K. Fahd and S. J. Miah, "Designing and Evaluating a Big Data Analytics Approach for Predicting Students' Success Factors," 2022, doi: 10.21203/rs.3.rs-2075479/v1.

[24] G. Akçapınar, A. Altun, and P. Aşkar, "Using Learning Analytics to Develop Early-Warning System for at-Risk Students," *Int. J. Educ. Technol. High. Educ.*, vol. 16, no. 1, 2019, doi: 10.1186/s41239-019-0172-z.

[25] K. Bird, B. Castleman, Z. Mabel, and Y. Song, "Bringing Transparency to Predictive Analytics: A Systematic Comparison of Predictive Modeling Methods in Higher Education," *Aera Open*, vol. 7, 2021, doi: 10.1177/23328584211037630.

[26] A. Hernández-Blanco, B. Herrera-Flores, D. Tomás, and B. Navarro-Colorado, "A Systematic Review of Deep Learning Approaches to Educational Data Mining," *Complexity*, vol. 2019, no. 1, 2019, doi: 10.1155/2019/1306039.

[27] S. K. Banihashem, K. Aliabadi, S. P. Ardakani, A. Delaver, and M. N. Ahmadabadi, "Learning Analytics: A Systematic Literature Review," *Interdiscip. J. Virtual Learn. Med. Sci.*, vol. 9, no. 2, 2018, doi: 10.5812/ijvlms.63024.

[28] M. Utari, B. Warsito, and R. Kusumaningrum, "Implementation of Data Mining for Drop-Out Prediction Using Random Forest Method," 2020, doi: 10.1109/icoict49345.2020.9166276.

[29] A. I. Adekitan and O. P. Salau, "The Impact of Engineering Students' Performance in the First Three Years on Their Graduation Result Using Educational Data Mining," *Heliyon*, vol. 5, no. 2, p. e01250, 2019, doi: 10.1016/j.heliyon.2019.e01250.