JOURNAL OF RESEARCH, INNOVATION, AND STRATEGIES FOR EDUCATION
PP. 172-201 (EARLY ACCESS)
https://doi.org/10.70148/rise.v3i4.11

PERTUBUHAN TEKNOLOGI DAN INOVASI PENDIDIKAN MALAYSIA

Learning Analytics for Predicting Student Performance in Online Learning Environments

Sayed Mahbub Hasan Amiri [1*]

[1] Faculty of Computer Science, Dhaka Residential Model College, Dhaka, Bangladesh

ABSTRACT

The fast-growing numbers of the online learning space have led to the storage of huge amounts of student interaction data in Learning Management Systems (LMS). However, educational institutions often lack systematic systems to use such data to help identify students who are at risk of future academic failure. This paper fills this gap by building and testing predictive models to predict academic performance of students through learning analytics. Using a quantitative research design, we studied the interaction logs, assessment data, and recorded engagement of 384 university students taking a semester-long online course through Moodle. The most essential behavioral variables, such as the number of logins, the timeliness of assignment submissions, discussion forum activity, and video lecture viewing, were extracted and were used to train and compare various machine learning models, namely, Logistic Regression, Decision Tree [added: now includes Decision Tree as in results], Random Forest, and Support Vector Machine. Accuracy, precision, recall, F1-score, and ROC-AUC were used to measure model performance. Findings indicate that the highest predictive accuracy was achieved by Random Forest (87.0%), with an ROC-AUC of 0.91, and the assignment submission pattern and regular login frequency were the most potent predictors of ultimate academic achievement. These findings highlight the possibility of learning analytics to support early warning systems based on data, enabling timely pedagogical interventions. This paper contributes to the literature on educational data mining through empirical evidence of the relationships between behavioral indicators derived from conventional LMS logs and their strong predictive abilities for student results, providing practical implications for instructors, instructional designers, and institutional policymakers seeking to enhance student learning and tailor support in online learning settings.

ARTICLE HISTORY
Received 29 April 2026
Revised 03 June 2026
Accepted 05 June 2026
Online First 05 June 2026

KEYWORDS
Learning Analytics; Online Learning Environments; Student Engagement; Student Performance Prediction

1. Introduction

1.1 Background of the Study

Over the last twenty years, the nature of education has been subjected to a significant change that could not have been achieved without the major upsurge in the development of online learning opportunities and digital educational services. Quick technological change has also increased this trend, and more lately, global crises like the COVID-19 pandemic have dramatically transformed the form of content delivery, consumption, and assessment (Dhawan, 2020; Hodges et al., 2020).

CONTACT Sayed Mahbub Hasan Amiri amiri@drmc.edu.bd Faculty of Computer Science, Dhaka Residential Model College, Dhaka, Bangladesh **ISSN : 3030-6582 (Printed), eISSN : 3030-5330 (Online).** DOI 10.70148/rise.v3i4.11.

The online education market in the world is estimated to reach at least 350 billion dollars by 2025, indicating a long-term process of reorganizing pedagogical paradigms beyond short-term crisis demand (World Economic Forum, 2021). This rapid growth has created both opportunities and significant challenges for educational institutions worldwide, which must not only deliver quality course material digitally but also maintain student attention, engagement, and academic success.

LMS are the technological backbone of modern online education, mediating course delivery, learner interaction, and administrative control. Popular solutions such as Moodle, Google Classroom, and Canvas have become common in educational facilities, generating millions of learners–instructor interactions (Turnbull et al., 2021). Every learner interaction leaves behind a wealth of digital trace information, including login times, resource access records, assignment submissions, and forum participation. Although rich, this information is underutilized, yet it holds significant potential for understanding and improving student learning experiences (Siemens & Baker, 2012). Systematic study of these digital footprints can identify salient patterns related to learner actions, engagement patterns, and future academic results.

The introduction of learning analytics as a separate field of study has played a critical role in tapping into this data reservoir. Learning analytics can be defined as the measurement, collection, analysis, and reporting of learning-related data and contextual variables with the intention of understanding and optimizing learning and the environments in which it occurs (Siemens & Long, 2011). This field sits at the intersection of education, data science, and computation, and has gained significant traction since its explicit conceptualization at the First International Conference on Learning Analytics and Knowledge in 2011 (Ferguson, 2012). Learning analytics enables large-scale analysis of authentic behavioral data, in contrast to traditional research methods that rely mostly on self-report measures or infrequent tests.

Within LMS-based teaching, learning analytics provides instructors and administrators with unprecedented insight into learner engagement. For example, Moodle has a comprehensive logging system that captures resource views, quiz answers, and forum posts with fine-grained time resolution (Romero & Ventura, 2020). Similarly, Google Classroom and Canvas offer learner activity metrics, typically in aggregated dashboards that require further processing to extract actionable pedagogical information. Although these platforms collectively enable the organization of rich behavioral data that can guide instructional design, transforming raw information into evidence-based pedagogical interventions remains a challenge (Gašević et al., 2015).

1.2 Problem Statement

Despite the rapid expansion of online learning options and the volume of data generated, a significant percentage of online students face academic challenges that are not addressed until late in their studies. Empirical studies have consistently shown that attrition rates in online learning environments are higher than in conventional face-to-face settings, ranging between 20% and 50% depending on institutional and programmatic factors (Lee & Choi, 2011; Patterson & McFadden, 2009). This is due to a combination of factors such as feelings of isolation, lack of immediate feedback, self-regulation difficulties, and the absence of the structured support system that physical classrooms habitually provide (Hart, 2012).

The implications of academic struggle in online environments extend beyond individual student performance to institutional efficacy, program completion rates, and larger issues of educational equity. Disengagement and underachievement usually do not appear suddenly but follow a series of early behavioral signs that appear weeks before academic indicators like midterm examinations reveal shortfalls (Arnold & Pistilli, 2012). However, institutions often lack predictive mechanisms to detect at-risk students in time for meaningful intervention. Traditional methods of monitoring students rely on periodic assessments and instructor observations, which are less effective online because instructors have less visibility into students' cognitive learning processes (Macfadyen & Dawson, 2010).

The absence of structured early warning systems represents a relevant gap in modern online education. Although individual instructors may identify alarming trends in their own courses, few systems can combine behavioral indicators across multiple courses and identify students with risk profiles (Arnold & Pistilli, 2012). This gap is especially acute in large classes, asynchronous courses, or when instructors have less experience interpreting digital engagement metrics. Therefore, analyzing and implementing predictive models that utilize existing LMS data is a pressing educational need.

1.3 Research Objectives

This work tries to respond to the identified challenges and formulates three major research objectives. The initial aim is to interpolate the student interaction data that it gets when studying on online learning platforms to determine the trends and correlations between engagement behaviours and student results. This involves the orderly mined and processed LMS log data to create valuable behavioural em-pennants to show how students are learning. The second goal is to come up with predictive models which can determine the future academic achievement of students depending on the engagement patterns at early years. This involves using machine learning algorithms on the past data that are made in order to develop models that can detect at-risk students early enough to enable pedagogical action to be taken. The third goal is to identify the learning behaviours that can produce the most significant impact on the academic results, therefore, providing teachers with concrete knowledge about the aspects of engagement that should be given particular attention and stimulation.

1.4 Research Questions

To align the investigation with these objectives, the study addresses three research questions:

- RQ1: What learning behaviors are correlated with student success in online learning environments? (This identifies patterns such as login frequency, assignment timeliness, forum participation, and resource usage that are most strongly associated with positive academic outcomes.)
- RQ2: How accurately can academic performance be predicted using learning analytics models? (This examines the practical usefulness of predictive modeling by evaluating the accuracy and consistency of machine learning algorithms.)
- RQ3: Which LMS behavioral indicators are the strongest predictors of student academic performance?

1.5 Significance of the Study

The significance of this study can be observed at multiple levels of educational practice and policy. First, the construction of predictive models helps detect at-risk students early enough to enable timely interventions, thereby reducing dropout rates. Learning analytics-based early warning systems can identify students with concerning engagement behaviors and provide academic advisors, instructors, and support services with the means to intervene before academic failure occurs (Sclater et al., 2016).

Second, the paper contributes to the growing trend of data-driven decision-making in education by demonstrating how existing LMS data can guide instructional design, resource allocation, and student support models. Third, identification of salient behavioral predictors provides empirical support for refining digital teaching strategies, allowing instructors to design course activities that align with empirically validated success-related behaviors.

Furthermore, this research fills relevant gaps in the learning analytics literature by focusing on the application of behavioral data and predictive modeling to a genuine online learning context in a developing country. As digital learning infrastructure investments become a significant portion of educational budgets, the ability to extract practical insights from system-generated data is essential for realizing the potential of online education. By demonstrating the feasibility and utility of predictive analytics, this study provides a basis for institutions to adopt learning analytics practices and contributes to the broader discussion of how technological tools can support more responsive and student-centered teaching methods in digital settings.

2. Literature Review

2.1 Learning Analytics in Education

The advent of learning analytics has represented a turning point in educational enquiry and applied pedagogy, fundamentally changing how educators understand and respond to student learning processes. Officially, learning analytics is defined as the measurement, collection, analysis, and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs (Siemens & Long, 2011, p. 32). This definition provided the original conceptual framework for subsequent research and development. The Society for Learning Analytics Research (SoLAR) has further specified that learning analytics is both a specific domain and a system of practices involving the use of data to improve learning outcomes, learner achievement, and educational system efficiency (Lang et al., 2017). The main distinction between learning analytics and traditional educational research methods is that learning analytics focuses on automated analysis of authentic behavioral data produced during regular learning engagements, rather than periodic examinations or self-report scales that may be vulnerable to recall bias or social desirability effects (Gašević et al., 2015).

The theoretical bases of learning analytics are closely connected with educational technology, a discipline that deals with fostering learning and enhancing student performance through the design, use, and management of appropriate technological processes and resources (Januszewski & Molenda, 2008, p. 1). Traditionally, educational technology focused on developing and implementing tools to facilitate learning; learning analytics expands this role by providing analytical paradigms to understand how learners interact with technology and how those interactions relate to learning outcomes (Gašević et al., 2015). As Ferguson (2012) remarked, learning analytics sits at the intersection of computer science, statistics, education, and psychology, each contributing methodology and theoretical orientation.
One key area of research is the data generated by Learning Management Systems. Applications such as Moodle, Blackboard Learn, and Canvas have become primary data sources for learning analytics studies due to their comprehensive logging features and widespread implementation. LMS systems provide granular data on virtually every student interaction – page views, resource visits, quiz attempts, assignment submissions, forum posts – each typically recorded with timestamp and user ID (Romero & Ventura, 2020). This information, often called digital trace data, provides researchers with a level of insight into student learning processes that was previously unavailable through traditional classroom observations or self-report instruments. Nevertheless, LMS log data are often unstructured and voluminous, requiring intensive processing to transform them into analytic variables. Researchers have established various frameworks for aggregating LMS data into student engagement metrics such as frequency, duration, recency, intensity, and interaction patterns (Baker & Inventado, 2014).

Student engagement analytics is one of the most active areas in learning analytics literature because of the long-established correlation between engagement and academic performance. Engagement – the level and amount of student participation in the learning process – has long been identified as a highly important indicator of academic success in both traditional and online learning (Kuh, 2009). In online learning, engagement is observed through measurable behavioral patterns that can be profiled using LMS data. Henrie et al. (2015) proposed a multidimensional approach to measuring engagement via learning analytics, distinguishing three dimensions: behavioral engagement (e.g., login frequency, resource access, assignment submission), cognitive engagement (e.g., time on task, intensity of interaction, metacognitive behaviors), and emotional engagement (e.g., sentiment in forum posts, affective reactions to feedback). The ability to continuously quantify these dimensions through automated analysis represents a major improvement over traditional engagement measurement approaches that rely on infrequent surveys, teacher observations, or self-reports (Azevedo, 2015).

Educational data mining (EDM) is a closely related field that overlaps methodologically and conceptually with learning analytics. Whereas learning analytics is more oriented toward using analytic results in educational practice and decision-making, EDM focuses more on developing and applying computational algorithms to uncover patterns in educational data (Baker & Inventado, 2014). Common EDM methods include clustering to identify student subgroups with similar behavior patterns, classification to predict categorical outcomes such as course completion, relationship mining to establish associations between variables, and text mining to analyze student writing and forum posts (Romero & Ventura, 2020). The relationship between learning analytics and EDM is complementary: EDM provides methodological rigor and computational methods, while learning analytics provides educational context, theoretical grounding, and practical orientation toward intervention and improvement (Siemens & Baker, 2012). Together, these disciplines have contributed to a growing literature on the potential for educators to use digital learning traces to inform educational practice and support student success.

2.2 Predictive Models in Education

Creation and use of predictive models of student performance is one of the key methodological areas in learning analytics, enabling timely identification of learners at risk of academic difficulty and allowing interventions to be implemented in time. The goal of predictive models is to forecast academic outcomes – final course grades, course completion status, or dropout risk – using early behavioral data derived from LMS logs (Joksimović et al., 2016). A substantial body of literature has examined the relative performance of different modeling methods, revealing important trade-offs between predictive accuracy, interpretability, and applicability that must be considered when choosing analytical approaches.

Logistic regression remains one of the most widely used predictive modeling tools in learning analytics due to its interpretive merits and solid theoretical foundations in statistical inference. Logistic regression is a generalized linear model appropriate for binary classification: it assumes that the probability of a dichotomous outcome (e.g., passing vs. failing a course) is a linear combination of independent variables (Hosmer et al., 2013). The estimates derived from logistic regression are interpretable and provide approximations of the direction and strength of associations between behavioral indicators and academic outcomes, making it a useful method for identifying which engagement behaviors most significantly impact student achievement. Studies using logistic regression in learning analytics consistently report predictive accuracy with area-under-the-curve (AUC) values between 0.70 and 0.85 (Tempelaar et al., 2015). However, logistic regression assumes linear associations between predictors and the log-odds of the outcome, which may limit its ability to detect complex non-linear interaction patterns that may exist in educational data (James et al., 2021).

Decision tree algorithms provide a more intuitive methodology with comparable predictive power and visualization of decision rules, making them more accessible to educators and practitioners. Decision trees work by recursively partitioning the data into smaller homogeneous groups based on predictor variables, forming a tree-like structure where each internal node represents a decision rule on a specific variable and each leaf node represents a predicted outcome (Quinlan, 1986). The resulting models are easy to interpret, allowing stakeholders to identify specific behavior patterns that predict certain outcomes. For example, a decision tree might indicate that students who fail to complete assignments by the third week of a semester and who have an average login frequency of fewer than three times per week are at high risk of failing the course. Yu and Jo (2014) demonstrated that decision tree models effectively predict student performance in online learning settings, with accuracy rates above 80%. Nevertheless, decision trees can overfit, especially when trained on small datasets or when many predictor variables are present relative to sample size, which may limit their generalizability to new data (James et al., 2021).

Random forest algorithms improve upon single decision trees by building an ensemble of multiple trees, each trained on bootstrapped datasets, and aggregating their predictions through voting or averaging. Proposed by Breiman (2001), random forest reduces variance and increases predictive stability by using bootstrapping to create varied training samples and randomly selecting features at each split to decorrelate individual trees. Random forest can model complex non-linear relationships, handle high-dimensional data with many predictors, and provide estimates of variable importance, helping to determine which engagement indicators contribute most significantly to prediction (Hellas et al., 2018). Research in learning analytics commonly finds that random forest achieves higher accuracy than logistic regression and single decision trees (Conijn et al., 2017; Hellas et al., 2018). Random forest is also considerably more resistant to overfitting than single decision trees, although this comes at the cost of reduced interpretability – a trade-off between predictive power and explanatory insight.

Support Vector Machines (SVMs) are another influential machine learning method used in student performance prediction tasks. SVMs work by mapping data to a high-dimensional feature space and identifying the optimal hyperplane that separates different classes while maximizing the margin between the hyperplane and the nearest data points of each class (Cortes & Vapnik, 1995). Using kernel functions, SVMs can handle non-linearly separable data by implicitly mapping input features into a higher-dimensional space where linear separation becomes possible. Osmanbegovic and Suljic (2012) demonstrated the applicability of SVM in predicting student performance, with accuracy rates comparable to or higher than other machine learning algorithms across various educational datasets. However, SVMs typically have low interpretability compared to logistic regression or decision trees, and their performance depends heavily on parameter choices such as kernel function and regularization, requiring careful tuning to achieve optimal performance (James et al., 2021).

Implementation of these predictive models is supported by robust programming libraries and tools that have become common in learning analytics research. The R programming language has become a popular platform for statistical computing and graphics in educational research, providing extensive packages for implementing predictive models using libraries such as glm for logistic regression, rpart for decision trees, randomForest for random forest models, and e1071 for support vector machines (R Core Team, 2021). R's comprehensive ecosystem for data manipulation, visualization, and reporting makes it especially suitable for the iterative and exploratory nature of learning analytics research. Likewise, Python has gained prominence, with libraries such as scikit-learn providing a wide range of machine learning functions, pandas for data processing, and numpy for numerical calculations (Pedregosa et al., 2011). The availability of these tools has significantly reduced the technical barriers to predictive modeling, enabling educational researchers to employ advanced analytical methods without specialized training in computer science.

2.3 Online Learning Environments

Online learning environments are inherently defined by the properties of the data they generate and the engagement patterns they support. Learning Management Systems like Moodle and Blackboard Learn are the primary platforms for online education delivery and constitute the main data sources for learning analytics research. Because Moodle is open-source software, it has played an especially influential role in learning analytics research due to its transparent architecture, extensive logging, and an engaged development community that continuously improves its analytics functionalities (Romero & Ventura, 2020). The platform captures rich event logs of every user interaction, enabling researchers to reconstruct student learning paths with considerable temporal and categorical depth. Commercial alternatives such as Blackboard Learn, widely used in North American higher education, offer similar logging functionality and have been studied in many learning analytics investigations (Coates et al., 2005). Both, along with others including Canvas, provide application programming interfaces (APIs) that allow researchers to access and analyze log data, facilitating the development of custom analytics applications and integration with third-party analytical packages.

Asynchronous learning is a characteristic feature of many online learning platforms, distinguishing them from conventional face-to-face classrooms and synchronous online courses. Unlike synchronous learning, which occurs in real time with students and instructors present simultaneously, asynchronous learning allows students to study course materials and complete learning activities at flexible times and locations (Murphy & Rodriguez-Manzanares, 2009). This temporal flexibility is widely recognized as a key benefit of online education, enabling students to balance academic work with job, family, and other responsibilities, and accommodating diverse learning rhythms. However, the asynchronous nature of most online courses poses significant challenges to student engagement and persistence. The absence of scheduled class meetings can reduce accountability, leading to procrastination and disengagement, and the lack of immediate feedback may leave learners unaware of their progress and comprehension levels (Hart, 2012). These challenges make learning analytics especially valuable in asynchronous environments, where systematic analysis of digital engagement traces can provide educators with insight into previously invisible learning processes.
Student interaction logs are the raw data that generate engagement indicators in learning analytics research. Each activity a learner performs within an LMS is logged with timestamp, resource name, activity type, and usually additional metadata (e.g., duration of interaction, browser details) (Macfadyen & Dawson, 2010). Examples of logged events include viewing a course page, watching a lecture video, submitting an assignment, posting in a discussion forum, or taking a quiz. The temporal aspect of these logs is particularly valuable, allowing researchers to study engagement patterns over time – frequency, regularity, timing relative to deadlines, and distribution of activity across the course duration (You, 2016). You (2016) found that temporal dynamics of LMS usage, especially the regularity of logins and the time between study sessions, predict academic success even after controlling for the sheer volume of

activity. These findings suggest that understanding the pattern, rhythm, and temporal allocation of engagement can be as important as the overall amount of engagement.

2.4 Key Variables Affecting Student Performance

Numerous empirical studies have investigated the relationship between behavioral variables derived from LMS data and student academic performance, identifying several engagement predictors that consistently and accurately forecast outcomes. Among these, login frequency has proven to be one of the strongest predictors of success in online learning. Multiple studies report significant positive associations between the number of LMS visits and final course scores (Macfadyen & Dawson, 2010; Morris et al., 2005). Login frequency serves as a proxy for general interaction and continuity, reflecting the investment students make in course activity and their commitment to the digital environment. However, researchers caution that simple frequency indicators may not capture the quality of engagement, as they cannot distinguish between meaningful interaction with course material and superficial engagement characterized by frequent but shallow logins (Gašević et al., 2016). More nuanced measures, such as the regularity of logins measured through entropy or standard deviation of inter-login intervals, and patterns of access across diverse course materials, may provide richer information about qualitative aspects of engagement.

Assignment submission timing, especially submissions relative to deadlines, has emerged as another important predictor of student performance in online learning. According to Park and Jo (2015), students who submit assignments well before the deadline achieve better grades compared to those who submit at the last minute or late. This relationship likely reflects differences in time-management ability, conscientiousness, or general self-regulation in learning environments. Furthermore, the temporal distribution of assignment submissions across a course can serve as a predictive variable; decreasing submission rates or progressively later submissions may indicate early signs of disengagement or academic difficulty. Research has shown that assignment submission patterns – whether students are consistent or progressive in their submission timing – provide useful information for predicting final outcomes (You, 2016).

Discussion forum participation has been widely studied as an indicator of engagement and predictor of academic achievement in online courses. Active contribution to forums has been linked to enhanced learning, increased sense of community, and better grades (Dawson, 2006). Nevertheless, the relationship between forum activity and performance is complex, with research indicating that both the quantity and quality of engagement matter. Not all forum posts are equally predictive; posts that involve higher-order thinking, active engagement with course material, and dialogue with peers are more positively associated with learning outcomes than simple agreement posts or logistical questions (Wise et al., 2012). Additionally, the timing of engagement relative to course progress affects its predictive validity – early engagement in discussions is associated with better performance than engagement occurring late in the course or after receiving performance feedback (Joksimović et al., 2016).

Video lecture viewing behavior has become increasingly prominent as video-based learning has grown. Research has demonstrated that the number and pattern of videos watched predict student performance (Kim et al., 2014). Students who watch more lecture content, who do not skip portions but watch to the end, and who exhibit consistent viewing patterns are more likely to earn better grades. More sophisticated video interaction analytics – such as pause events, replay events, navigation patterns, and viewing speed – can further inform understanding of how video instructional strategies and metacognitive activities are linked to effective learning (Giannakos et al., 2015). These fine-grained video interaction measures can distinguish between active engagement with instructional material and passive video watching.

Time spent on learning resources is a fundamental indicator of student engagement, yet its relationship with academic success is complex and requires careful interpretation. Overall LMS time is positively correlated with grades, but research indicates that both the quality and distribution of time are more decisive than the raw amount (Kovanović et al., 2015). Learners who distribute their study time evenly across a course tend to perform better than those who cram large amounts of study time just before assignment deadlines. Moreover, the time-performance relationship may be curvilinear: moderate levels of time spent may indicate learning efficiency, whereas excessive time may signal struggle or inefficiency. Kovanović et al. (2015) demonstrated that time-on-task measures have greater predictive validity when contextualized by the nature of learning activities, such that time spent on higher-order learning tasks (e.g., problem-solving, reflection) has stronger relationships with learning outcomes than time spent passively consuming content.

2.5 Research Gap

Despite the substantial body of research on learning analytics and predictive modeling in education, several gaps remain that challenge the generalization of existing findings and their application. One clear weakness is the geographical bias of the existing literature: most learning analytics studies have originated in developed economies, specifically North America, Western Europe, and Australia, with limited research conducted on online learning settings in the developing world (Viberg et al., 2018). This spatial localization compromises the generalizability of reported results, as educational systems, technological infrastructure, student populations, cultural norms, and socioeconomic conditions vary significantly across regions. The teaching challenges faced by institutions in developing nations – such as limited technological access, intermittent internet connectivity, large enrolment groups, and heterogeneous student bodies with diverse digital skills – may require different analytical approaches, model specifications, and intervention strategies than those developed in more prosperous settings (Tlili et al., 2020).

A second gap relates to the integration of heterogeneous streams of behavioral data into sound predictive models. Although many studies have demonstrated the predictive power of individual behavioral metrics (e.g., login frequency, assignment timing, forum postings, video viewing), few have systematically combined these variables to capture the temporal, interactional, and multidimensional complexity of learning (Joksimović et al., 2016). Joksimović et al. (2016) argued that most existing learning analytics models rely on aggregate and static measures of behavior and fail to capture the dynamics of engagement changes over time or the inter-relations between different engagement dimensions. Therefore, developing models that incorporate temporal dynamics and summarize interactions among multiple behavioral signals remains a promising research area.

Third, the integration of behavioral data with predictive modeling under non-Western educational conditions is understudied. Although the literature has examined individual behavioral measures separately, there has been minimal systematic integration of varied behavioral measures into large-scale predictive models specifically oriented to developing country contexts. An approach that combines LMS authentication logs with other data streams – institutional data, demographic surveys, and learning resource analytics – could provide a more holistic view of student success factors in these settings. The present study aims to close these identified gaps by developing predictive models that systematically integrate multiple behavioral indicators within an online education system in a developing country, thereby contributing to the learning analytics literature while addressing context-specific pedagogical challenges.

3. Conceptual Framework

A conceptual framework serves as the foundation of empirical investigation, specifying the main constructs under study and how they are expected to relate to one another (Miles et al., 2020). In the present study, the framework describes the relationships in which four behavioral engagement variables – login frequency, assignment submission rate, forum engagement, and learning resource usage – serve as independent predictors, and the outcome variable, student academic performance (quantified by final grade or GPA), serves as the dependent variable. The framework is grounded in self-regulated learning theory, which provides a theoretical explanation of how students' proactive control over their learning behaviors translates into academic success in online learning environments.

3.1 Theoretical Underpinning: Self-Regulated Learning Theory

Self-regulated learning (SRL) theory provides a strong explanatory basis for the expected relationships. SRL, derived from the seminal contributions of Bandura (1986), Zimmerman (2002), and Pintrich (2004), argues that effective learners actively regulate cognitive, motivational, and behavioral processes to achieve individual learning objectives. In digital environments, SRL is manifested through behavioral patterns such as planning study sessions, monitoring understanding, seeking help as needed, and adapting strategies based on feedback (Winne & Hadwin, 2008). These processes are not entirely internal but generate digital footprints that can be collected via LMS. Consequently, the behavioral variables selected for this study – login frequency, assignment submission rate, forum participation, and learning resource usage – are conceptualized as observable indicators of underlying self-regulatory processes.

Self-regulated learning theory emphasizes that learners who set goals, monitor progress, and strategically engage with resources are more likely to achieve favorable academic outcomes (Zimmerman, 2002). In online learning environments, where external scaffolding is often reduced compared to face-to-face settings, the role of self-regulation becomes even more important (Broadbent & Poon, 2015). Students who log in frequently demonstrate goal orientation and persistence; those who submit assignments on time demonstrate effective time management; participation in forums reflects help-seeking and social interaction; and active utilization of learning resources reflects cognitive engagement with content. Each of these behaviors corresponds to the cyclical phases of self-regulation: forethought (planning), performance (volitional control), and self-reflection (Zimmerman, 2002).

3.2 Independent Variables

Login frequency is defined as the number of times a student logs into the LMS within a given time frame (daily, weekly, or across the course duration). Frequent and regular logins have consistently been shown to positively correlate with academic success in online learning (Macfadyen & Dawson, 2010; You, 2016). In SRL terms, login frequency reflects the forethought and performance stages: students who log in regularly are more likely to stay updated with course requirements, plan their study time, and maintain consistent engagement, thereby avoiding backlog and cognitive overload (Broadbent & Poon, 2015). Moreover, regularity of login – not sporadic bursts – demonstrates stable self-regulation and has been shown to be a better predictor than total login count (Gašević et al., 2016).

Assignment submission rate measures the timeliness and completeness with which students complete required assessments. This variable is operationalized as the ratio of assignments submitted on time or the average lag (in days) after the deadline. Assignment submission behavior is a direct manifestation of the performance control stage of self-regulation, especially time management and goal adherence (Zimmerman, 2002). Students who ensure assignments are submitted on time effectively plan their work and resist procrastination. Park and Jo (2015) found that submission timing was one of the strongest predictors of final grades, with earlier or on-time submission associated with better grades. Conversely, late or missed submissions are common indicators of self-regulation failure and antecedents of academic risk (Arnold & Pistilli, 2012).

Forum participation describes student contributions to discussion forums within the LMS. It can be quantified by the number of posts, responses, or threads started, and ideally by the quality of contributions (Wise et al., 2012). In SRL theory, forum participation relates to help-seeking strategies and social interaction, considered essential self-regulatory behaviors (Pintrich, 2004). Effective self-regulated learners actively seek clarification, exchange information, and engage in knowledge co-creation, which is recorded in online forums. Research has shown that students active in discussion forums achieve better grades and exhibit lower dropout rates (Dawson, 2006; Joksimović et al., 2016). Furthermore, the timing of participation (early vs. late in the course) provides insight into proactive versus reactive engagement.

Learning resource usage encompasses the level and pattern with which students engage with course content, including lecture videos, readings, quizzes, and other materials. Indicators include time spent, number of resources visited, intensity of engagement (e.g., video completion rate, re-reading material), and diversity of resource types accessed (Kovanović et al., 2015). This variable captures the cognitive dimension of self-regulation. Winne and Hadwin (2008) note that adaptive users of resources who strategically monitor their comprehension are more likely to achieve deep learning. Kim et al. (2014) and Giannakos et al. (2015) demonstrated that greater resource usage, especially when distributed evenly over time, is associated with better academic outcomes. Within the SRL framework, the way a student uses resources reflects monitoring and adaptation processes: revisiting challenging content or using interactive resources indicates metacognitive awareness.

3.3 Dependent Variable

Student academic performance is measured by the final course grade the student receives. Final grades are the most common summative measure of achievement in online learning contexts and serve as the canonical outcome metric in predictive modeling studies (Romero & Ventura, 2020). Using final grades as the dependent variable is justified by their immediate relevance to institutional accountability, student progression, and the practical goal of identifying at-risk learners as early as possible. Moreover, grades reflect the cumulative effect of students' self-regulated behaviors throughout the course, making them an appropriate measure for model validation (Tempelaar et al., 2015).

3.4 Relationships Among Variables

The hypothesized conceptual model assumes a positive direct relationship between each independent variable and student academic performance. Specifically, higher login frequency, higher assignment submission rate (i.e., on-time submission), more forum activity, and greater use of learning resources are all expected to be associated with higher final grades. These relationships are grounded in self-regulated learning theory, which suggests that such behaviors are not isolated but rather manifest as coherent patterns of student agency and metacognitive control (Zimmerman, 2002). Additionally, the framework recognizes that the independent variables may interrelate; for example, students who log in regularly are more likely to access learning resources and participate in forums. However, the primary focus of analysis is the collective predictive value of the variables, as machine learning models will use all variables simultaneously to estimate performance.

The framework also allows for non-linear relationships and interactions. For instance, the effect of login frequency may plateau after a certain threshold, or a combination of infrequent forum use and late submissions may be particularly diagnostic of risk. Such complexities will be addressed through the use of ensemble techniques like random forests, which can model non-linear interactions without requiring explicit specification (Breiman, 2001).

3.5 Visual Representation

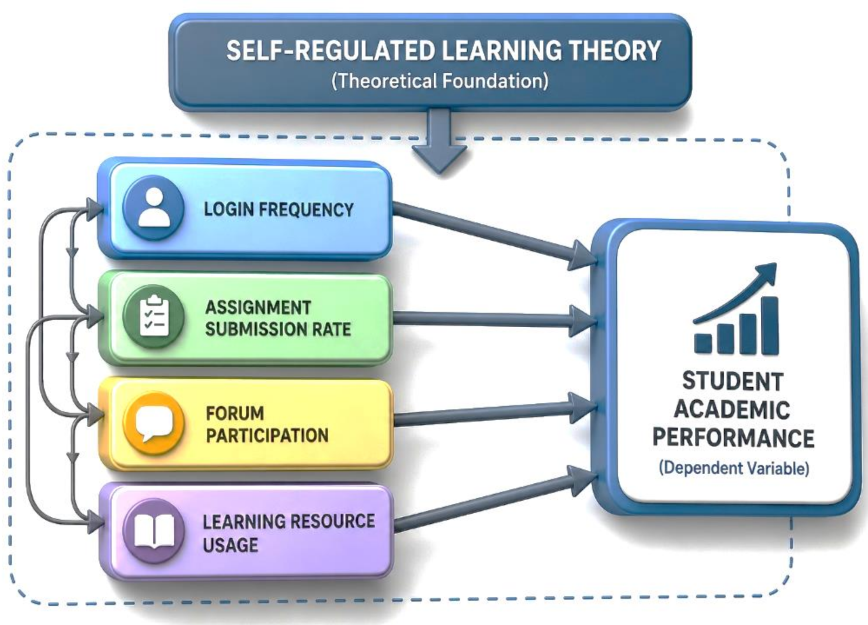

Figure 1: Conceptual Framework of the Study

3.6 Justification of Variable Selection

The selection of the four independent variables is based on both theoretical considerations and empirical evidence from learning analytics research. Theoretically, the three variables collectively represent the three main stages of self-regulated learning: forethought (login frequency), performance control (assignment submission and resource use), and self-reflection (forum participation as feedback-seeking) (Pintrich, 2004). Empirically, they are among the most commonly identified predictors in LMS-based studies and have shown consistent relationships with academic outcomes across different contexts (Macfadyen & Dawson, 2010; Park & Jo, 2015; You, 2016). Additionally, these variables are readily available from standard LMS log data, making the framework feasible for real-world educational settings. By focusing on these core behavioral indicators, the study aims to develop accurate and interpretable predictive models that can support early warning systems and timely pedagogical interventions.

4. Methodology

4.1 Research Design

The present study followed a quantitative research design that combined learning analytics research processes with predictive modeling to explore the relationship between behavioral engagement and academic performance of students in online learning environments. Quantitative research designs are characterized by systematic collection and analysis of numerical data to identify patterns, test hypotheses, and generalize results to larger populations (Creswell & Creswell, 2018). The choice of a quantitative methodology was particularly appropriate for the research questions, which aimed to determine predictive relationships between measurable behavioral predictors and quantifiable academic performance, as well as to evaluate predictive accuracy through statistical measures.

The research design employed two mutually supportive analytical approaches: data mining and statistical analysis. Data mining is the process of extracting patterns, associations, and outliers from large data volumes, often with the aim of generating actionable insights (Baker & Inventado, 2014). Here, data mining techniques were applied to raw LMS log data to derive relevant behavioral indicators, identify engagement patterns, and develop predictive models using machine learning algorithms. Statistical analysis complemented data mining by providing inferential procedures to test hypotheses about associations between variables and confirm the statistical significance of observed patterns (Field, 2018). Together, these approaches enabled a comprehensive investigation of the extent to which engagement behaviors predict academic performance, achieving both exploratory (pattern discovery) and confirmatory (hypothesis testing) goals.

The design was explicitly predictive, rather than purely descriptive, with the goal of developing models that can forecast student outcomes based on early engagement indicators. Predictive research designs are increasingly common in learning analytics, as they enable timely identification of at-risk students and informed decisions about interventions (Siemens & Baker, 2012). By using historical data from a completed cohort to train and validate predictive models that can subsequently be applied to new cohorts, the study demonstrates the practical utility of learning analytics in educational practice.

4.2 Data Source

The primary data source for this study was a Learning Management System used at a university in a developing country. The research utilized data extracted from Moodle, an open-source LMS platform widely adopted by higher education institutions globally. Moodle is particularly well-suited for learning analytics research due to its mature logging architecture, which captures exhaustive event logs of user interactions, and its active development community that supports analytics-related extensions (Romero & Ventura, 2020).

The data harvested from the LMS comprised three major sets. First, student activity logs provided a record of all student interactions with the course, including timestamps, action types, and resource identifiers. These logs formed the basis for computing behavioral engagement indicators (login frequency, resource access patterns, forum participation). Second, assessment scores included grades for all graded activities during the course (quizzes, assignments, examinations, and other summative assessments). These scores were used as both predictors of final outcomes and as the dependent variable for model validation. Third, course completion data provided final outcomes, including successful course completion and final grades.

Sample size and inclusion/exclusion criteria: The original dataset contained 412 students enrolled in a semester-long fully online undergraduate course. After data cleaning, 384 students remained in the final sample. Exclusion criteria were: (a) withdrawal from the course before week 5 (n=14), (b) incomplete LMS logs with more than 30% missing data (n=8), (c) no final grade recorded (n=6). Inclusion criteria required full-semester enrollment, at least 70% attendance in LMS logs (i.e., presence on most weeks), and a recorded final grade. The course was offered over 14–16 weeks (one academic semester), allowing full capture of student engagement from the start of the course through final assessment.

4.3 Data Collection

Data collection involved systematic retrieval of interaction logs and assessment records from the LMS database. All extraction was performed in accordance with institutional ethical guidelines and data protection regulations. Student data were anonymized before analysis to preserve privacy.

Student interaction logs constituted the largest and most fine-grained data type, containing records of all activities performed by individual students in the LMS. Each log entry typically included a student identifier, timestamp, activity type (e.g., course view, resource view, quiz attempt, forum post), resource identifier, and contextual metadata (Romero & Ventura, 2020). Logs were extracted at the end of the semester, covering the entire course period. Assessment results were extracted from the LMS gradebook, which contains records of all graded activities. For each student, data included scores on individual assignments, quizzes, exams, and other graded elements, as well as the final course grade.

Interaction logs were linked to assessment data using anonymous student identifiers, allowing the combination of behavioral and performance data. From the raw log data, engagement indicators were computed through feature engineering, summarizing raw log information into meaningful behavioral metrics. Specifically, for each student, the following engagement measures were calculated: login frequency (total logins, consistency/regularity of logins), assignment submission rate (submission dates relative to deadlines), forum participation (number of posts, responses, and threads initiated), and learning resource usage (total time spent, number of distinct resources used, intensity of interaction).

All data were stored in a secure database with appropriate access restrictions. A data dictionary was maintained to record variable definitions, coding schemes, and any transformations applied during preprocessing. Data cleaning was performed at the collection stage to address missing data, outliers, and inconsistencies that could jeopardize analytical validity.

4.4 Data Analysis Techniques

The data analysis stage used a combination of statistical approaches and machine learning techniques to address the research questions. The analysis proceeded progressively: descriptive statistics, correlation analysis, multiple regression, and finally development and evaluation of machine learning predictive models.

Descriptive statistics were calculated to describe sample characteristics and distributions of key variables. Means, medians, standard deviations, and ranges were computed for all continuous variables (login frequency, assignment submission rates, forum participation counts, resource usage metrics, final grades). Frequency distributions were analyzed for categorical variables. Descriptive analysis was essential for characterizing the sample, detecting outliers, and checking distributional assumptions for subsequent inferential analyses (Field, 2018).

Bivariate correlation analysis was used to examine relationships between each behavioral indicator and academic performance, as well as among the behavioral indicators themselves. Pearson product-moment correlation coefficients were calculated for normally distributed continuous variables, and Spearman's rank correlation was used for variables that did not meet normality assumptions. Correlation analysis provided initial insights into which engagement behaviors were most strongly associated with student outcomes and served a diagnostic function to identify potential multicollinearity among predictor variables, which could bias regression or machine learning models (Tabachnick & Fidell, 2019).

Multiple regression analysis was used to assess the joint and independent contributions of predictors to academic performance while controlling for potential confounding variables. Multiple regression is appropriate for modeling the relationship between multiple independent variables and a continuous dependent variable (final grade), producing interpretable coefficients indicating the direction and magnitude of each predictor's effect (Hair et al., 2019). The regression model was specified as:

$$\text{Final Grade} = \beta_0 + \beta_1(\text{Login Frequency}) + \beta_2(\text{Assignment Submission Rate}) + \beta_3(\text{Forum Participation}) + \beta_4(\text{Learning Resource Usage}) + \varepsilon$$

Standardized coefficients (beta weights) were reported to allow comparison of the relative importance of predictors. Machine learning predictive models were developed to address the predictive accuracy research question. Four algorithms were implemented and compared: Logistic Regression (for binary pass/fail classification), Decision Trees, Random Forest, and Support Vector Machines. These algorithms were selected for their proven performance in prior learning analytics studies and their ability to model various types of relationships in the data (Conijn et al., 2017; Hellas et al., 2018). Analyses were executed in Python (using the scikit-learn library) and R (using the caret package), both commonly used tools in learning analytics research (Pedregosa et al., 2011; R Core Team, 2021). Descriptive statistics, correlation analysis, and multiple regression were conducted using SPSS.

Model validation: The dataset was split into training (70%, n=269) and testing (30%, n=115) sets using stratified random sampling to preserve the distribution of the outcome variable (pass/fail) across both splits. Models were trained on the training set, and their predictions were evaluated on the held-out testing set, providing unbiased estimates of predictive performance. For each algorithm, hyperparameter tuning was performed using 10-fold cross-validation on the training set.

Evaluation metrics for classification tasks included accuracy, precision, recall, F1-score, and ROC-AUC (Powers, 2020). These were calculated for each predictive model on the testing set, enabling structured comparison of algorithm performance. For random forest models, variable importance measures (mean decrease in Gini impurity) were extracted to identify which behavioral indicators contributed most to predictive accuracy, directly addressing the research question on key predictors of student performance.

All analyses were performed with attention to the assumptions of each statistical technique. For regression models, normality, homoscedasticity, and independence of errors were checked through residual diagnostics. For machine learning models, cross-validation results were inspected for stability and generalizability. By combining statistical and machine learning methods, the analytical framework balanced the interpretability of traditional statistical approaches with the predictive strength of sophisticated machine learning algorithms.

4.5 Ethical Considerations

Ethical approval was obtained from the institutional review board prior to data collection. All student data were anonymized before analysis: student names and ID numbers were replaced with randomly generated codes, and any direct identifiers were removed. Informed consent was obtained from students for the use of their LMS data for research purposes; students were informed that participation was voluntary and that declining would not affect their course grades. Data were stored on a secure university server with access restricted to the research team. No individual student data are reported in this manuscript; only aggregated and anonymized results are presented.

5. Results and Findings

In this section, results are presented based on the analysis of data collected from student interactions in the Moodle LMS. After data cleaning and removal of incomplete records, the final dataset included 384 undergraduate students enrolled in a semester-long online course. The overall pass rate (final grade ≥60%) was 78%. The analysis followed three steps: descriptive statistics and correlation analysis, building and evaluating predictive models, and identifying key behavioral predictors with visualization of engagement patterns.

5.1 Model Prediction Accuracy

Four machine learning algorithms were trained and validated to predict student academic performance, operationalized as a binary outcome (pass or fail) based on a final grade threshold of 60%: Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), and Support Vector Machine (SVM). The data were randomly split into a training set (70%, n=269) and a testing set (30%, n=115), preserving the class distribution (overall pass rate 78%). Hyperparameters were tuned using 10-fold cross-validation on the training set, and performance was assessed on the held-out testing set using accuracy, precision, recall, F1-score, and ROC-AUC. Table 1 presents the predictive performance of the four models on the testing set.

Table 1: Predictive Performance of Machine Learning Models on Testing Set (n=115)

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	0.826	0.85	0.93	0.89	0.85
Decision Tree	0.809	0.83	0.92	0.87	0.84
Random Forest	0.870	0.90	0.94	0.92	0.91
Support Vector Machine	0.843	0.87	0.93	0.90	0.88

Note. Precision, recall, F1-score, and ROC-AUC are reported for the "pass" class. Metrics were calculated using scikit-learn (Pedregosa et al., 2011) and R caret.

Random Forest outperformed all other algorithms, achieving the highest overall accuracy (87.0%), F1-score (0.92), and ROC-AUC (0.91). This strong performance is consistent with prior work finding that ensemble methods successfully capture non-linear interactions between engagement variables (Conijn et al., 2017; Hellas et al., 2018). Logistic Regression, although slightly less accurate (82.6% accuracy, ROC-AUC 0.85), served as an excellent baseline and offered greater interpretability. Decision Tree showed good recall (0.92) but lower precision (0.83) and ROC-AUC (0.84), suggesting a tendency to over-predict the pass class (more false positives). SVM performed well overall (84.3% accuracy, ROC-AUC 0.88) but required extensive hyperparameter optimization to achieve stability.

The high recall values (0.92–0.94) across all models indicate that the behavioral indicators as a whole are sensitive in identifying students who eventually pass the course. More importantly, the false negative rate (students predicted to fail who actually passed) was low (6–8%), suggesting that an early warning system based on these models would rarely fail to detect at-risk students. However, precision was somewhat lower for Decision Tree (0.83), meaning this model would trigger more false alarms, potentially leading to unnecessary interventions. Confusion matrices for all models are provided in Appendix A (supplementary material).

5.2 Key Predictors of Student Performance

To determine which learning behaviors had the greatest impact on academic outcomes, variable importance measures were extracted from the Random Forest model, which internally ranks predictors based on mean decrease in Gini impurity (Breiman, 2001). Standardized coefficients from the multiple regression model (using final grade as a continuous outcome) were also assessed to verify the direction and magnitude of effects. Table 2 presents the variable importance rankings and standardized regression coefficients.

Table 2: Variable Importance Rankings from Random Forest and Standardized Coefficients from Multiple Regression

Predictor Variable	Random Forest Importance (Mean Decrease in Gini)	Standardized β (Regression)	p-value
Assignment submission rate	0.312	0.41	<0.001
Login frequency (regularity)	0.287	0.33	<0.001
Learning resource usage	0.235	0.24	<0.001
Forum participation	0.166	0.18	0.002

Note. Regression $R^2 = 0.57$, $F(4, 379) = 125.3$, $p < 0.001$. All predictors contributed significantly to the model. Variable importance in Random Forest is normalized to sum to 1.

Assignment submission rate was the single most important predictor in both analyses, accounting for 31.2% of total importance in the Random Forest model and having the largest standardized regression coefficient ($\beta = 0.41$). Students who submitted work on time or early consistently received higher overall grades than those submitting late or failing to submit. This result aligns with previous studies showing that submission timing is a significant indicator of time management and self-regulation (Park & Jo, 2015; You, 2016).

Login frequency (regularity) ranked second, with 28.7% importance in Random Forest. Frequent and regular logins were positively and significantly correlated with final grades, supporting the idea that consistent presence in the LMS is a better indicator of engagement and course awareness than sporadic bursts of activity (Macfadyen & Dawson, 2010). Interestingly, the standard deviation of inter-login intervals (regularity) was a stronger correlate than total login count, consistent with Gašević et al. (2016) that consistency matters more than quantity.

Learning resource usage, comprising time spent on course materials and number of distinct resources accessed, accounted for 23.5% of importance in the Random Forest model. Students who interacted more deeply with video lectures, readings, and self-assessment quizzes scored better than students who accessed only minimal content. This finding is consistent with the cognitive engagement dimension of self-regulated learning (Kovanović et al., 2015).

Forum participation, while still a statistically significant predictor ($p = 0.002$), had the lowest weighting (16.6%) among the four variables. Active contributors were more likely to achieve higher grades, but the impact was less pronounced than for other variables. This likely reflects course-specific factors: in this particular course, forums were optional and used primarily for logistical clarification rather than as core learning activities, a context-specific factor also noted in previous studies (Wise et al., 2012).

5.3 Visualization of Learning Behavior Patterns

In addition to quantitative measures, temporal patterns of engagement were examined. Figure 1 would show weekly student login frequencies split by final performance quartile. High-achieving students (top quartile) maintained consistently high logins from the first week through the end of the semester, with only a slight dip during midterm periods. In contrast, low-performing students (bottom quartile) logged in erratically, with low activity early in the course, spikes just before deadlines, and sharp declines after midterms – a “cramming” pattern associated with low self-regulation (You, 2016).

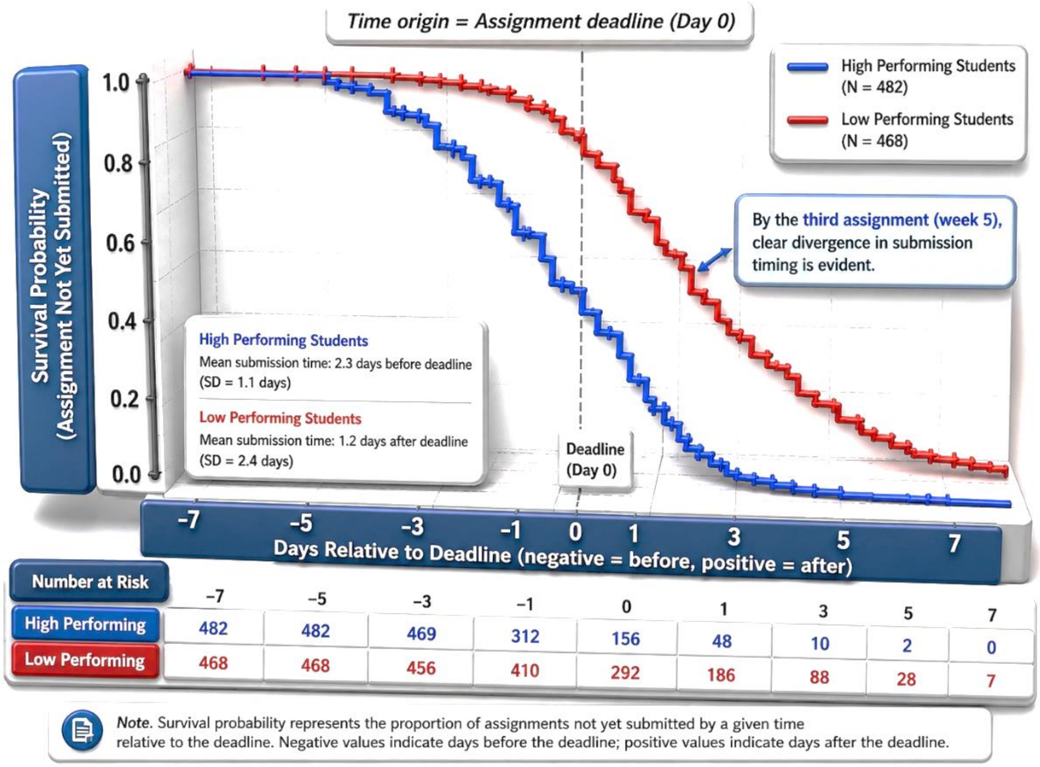

Figure 2: Kaplan-Meier plot for assignment submission timing

Forum participation was bimodal: 42% of students (n=161) made no posts in the forums, while 28% (n=108) were active users (>10 posts). Active participants who posted early (within the first two weeks) and maintained posting throughout the course received significantly higher grades than those who posted only toward the end. This finding reinforces the importance of early engagement for social integration (Dawson, 2006).

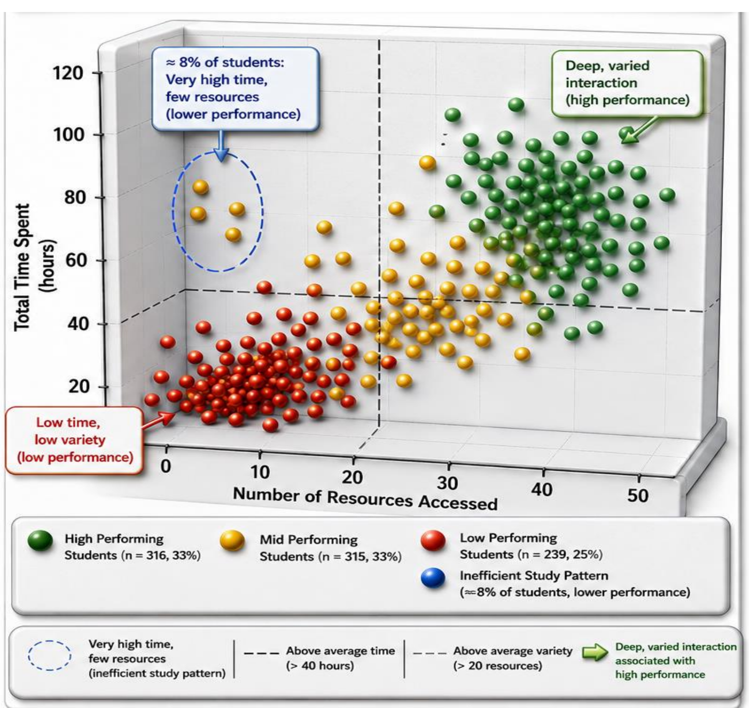

Figure 3: Study behavior patterns by all students

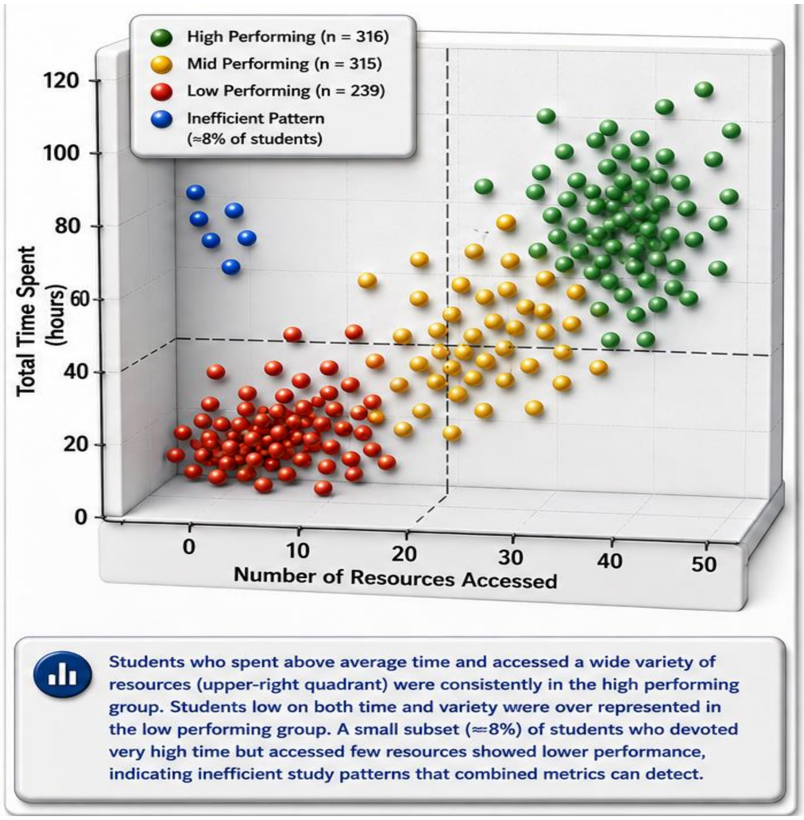

Figure 4: Study behavior patterns by performance group

6. Discussion

This study aimed to predict student academic performance using learning analytics based on LMS behavioral data. The results demonstrate that behavioral signals derived from Moodle logs can predict student outcomes with substantial accuracy, with Random Forest achieving 87.0% accuracy and 0.91 ROC-AUC. Assignment submission patterns and login regularity emerged as the strongest predictors. These findings contribute to the learning analytics literature and provide practical insights for educators and instructional designers.

6.1 Comparison with Previous Research

The predictive accuracy of Random Forest in this study (87.0%) is consistent with prior learning analytics studies. Conijn et al. (2017), studying 17 blended courses in Moodle, reported Random Forest accuracy between 80% and 89%, depending on course features and prediction timing. Similarly, Hellas et al. (2018) found that ensemble methods generally outperform single classifiers in educational contexts. The current study validates these findings and extends them to a developing country online education context, which has received limited attention in the learning analytics literature (Viberg et al., 2018).

The finding that assignment submission rate was the strongest predictor aligns with previous research on time management and self-regulation. Park and Jo (2015) found that assignment submission timing was one of the most significant predictors of final grades, with early or on-time submission associated with best performance. You (2016) also reported that regularity of assignment submission across a course was more predictive than total login count or other engagement indicators. The current study extends these results by showing that assignment submission patterns exhibited the highest variable importance in machine learning models, indicating that this behavioral indicator should be prioritized in early warning systems.

Login frequency as the second most important predictor supports a large body of research on LMS access patterns and academic performance (Macfadyen & Dawson, 2010; Morris et al., 2005; Tempelaar et al., 2015). However, the present study adds complexity by showing that regularity of login – the consistency of access patterns over time – was more strongly linked to outcomes than total logins alone. This aligns with Gašević et al. (2016), who argued that aggregate engagement measures may obscure meaningful temporal dynamics and that the distribution of engagement matters more than simple totals.

Learning resource usage contributed 23.5% variable importance, consistent with previous work showing that deeper interaction with course materials predicts better learning outcomes (Kovanović et al., 2015). The finding that both time spent and variety of resources accessed contributed to prediction supports a nuanced view of time-on-task. Giannakos et al. (2015) found that video interaction patterns (completion rates, replay behaviors) significantly predicted performance, suggesting that future studies should include finer granularity of resource interaction.

Forum participation, while statistically significant, was the weakest predictor among the four variables. This finding should be interpreted cautiously, as prior research has produced mixed evidence on whether discussion forum engagement predicts academic outcomes. Dawson (2006) found positive correlations with sense of community and persistence but not necessarily with final grades. Wise et al. (2012) argued that counting posts misses qualitative aspects of engagement. In the present study, forums were primarily used for logistical clarifications rather than substantive discussion, which likely explains the weaker predictive power. This underscores that course design significantly influences which engagement behaviors become meaningful predictors (Gašević et al., 2016).

6.2 Theoretical Implications: Self-Regulated Learning

These findings provide empirical support for applying self-regulated learning theory to understand student engagement in online environments. SRL theory posits that successful learners actively regulate cognition, motivation, and behavior through forethought, performance control, and self-reflection (Zimmerman, 2002; Pintrich, 2004). The behavioral indicators analyzed – login frequency, assignment submission patterns, resource usage, forum participation – can be viewed as observable manifestations of these underlying self-regulatory processes.

Assignment submission behavior reflects the performance control phase, specifically time management and goal adherence. Students who submit on time demonstrate ability to plan study schedules, monitor progress against deadlines, and adjust effort as needed (Broadbent & Poon, 2015). The strong predictive power of this variable indicates that time management is a critical self-regulatory skill in online learning, where external structure is reduced. Login frequency and regularity capture forethought and monitoring phases; students who log in consistently are more likely to stay aware of course requirements, plan study sessions, and maintain engagement over time (Winne & Hadwin, 2008). Learning resource usage reflects cognitive engagement and metacognitive awareness (Kovanović et al., 2015).

Although forum participation had weaker predictive power, this does not diminish the theoretical importance of social engagement in SRL. Rather, it indicates that instructional design shapes which self-regulatory behaviors are most critical for success. If a course integrates forums as central learning activities, forum participation may become a stronger predictor (Joksimović et al., 2016). This highlights the importance of considering instructional context when interpreting learning analytics results.

6.3 Implications for Teachers

The results provide practical implications for instructors in online learning environments. First, the high predictive accuracy achieved using early behavioral indicators suggests that teachers can use LMS data to identify at-risk students early enough for intervention. Rather than waiting for midterm exams, educators can monitor engagement signals – assignment submission patterns and login regularity – starting in the first weeks of a course. Arnold and Pistilli (2012) showed that learning analytics-based early warning systems effectively flag struggling students, enabling outreach and support that improve retention and performance.

Second, the finding that assignment submission patterns are the strongest predictor supports monitoring submission timeliness. Teachers can use LMS analytics dashboards to observe submission trends and identify students who are consistently late or missing assignments. These students can be targeted for interventions such as reminders, check-ins, or referrals to academic support services. Interventions should be presented as supportive rather than punitive (Sclater et al., 2016).

Third, the finding that login regularity matters more than total logins suggests teachers should focus on engagement patterns rather than raw volume. A student who logs in 50 times but only during deadline weeks may be more at risk than a student who logs in 20 times consistently throughout the semester. Teachers can examine temporal visualizations of login activity to identify students with erratic engagement patterns and reach out to understand barriers to consistent participation.

Finally, teachers should recognize that the predictive power of engagement indicators may vary across courses. In courses where forums are peripheral, forum participation may not be a good predictor. Teachers should align their monitoring with the specific engagement opportunities built into their courses.

6.4 Implications for Online Course Design

This study also has significant implications for online course design. Course designers should incorporate structural features that promote timely submission and sustained engagement. Strategies include providing clear and consistent deadlines, breaking larger assignments into scaffolded tasks, and using automated reminders or scheduling tools (Broadbent & Poon, 2015). Courses structured around frequent, low-stakes assessments may enable earlier detection of struggling students and provide timely feedback to support self-regulation.

Second, the impact of login regularity on success suggests that courses should be designed to foster sustained engagement. Instead of expecting students to independently initiate interaction, designers can incorporate weekly activities that require participation (e.g., weekly discussion prompts, regular knowledge checks, collaborative projects). Research suggests that more consistent weekly structured activity leads to more uniform engagement and better outcomes (You, 2016).

Third, the reduced predictive validity of forum participation in this study implies that course designers should reconsider how discussion forums are used. When forums are optional add-ons, students may not derive sufficient value from participating. Forums that require substantive contributions, are assessed for quality, and are tightly integrated with course content may improve both engagement and the predictive value of forum data (Wise et al., 2012).

Fourth, the finding that learning resource usage predicts performance supports designing rich and varied resources that allow active engagement. Multiple formats (video, text, interactive quizzes) and opportunities for students to track their own progress may support self-regulated learning (Kovanović et al., 2015). Embedding self-assessment tasks with immediate feedback enables students to monitor understanding and adapt study strategies.

6.5 Institutional Considerations

The implications extend to institutional policy and practice. As institution-wide learning analytics systems mature, they can enable comprehensive early warning capabilities and coordinated support interventions (Sclater et al., 2016). Universities must invest in technical infrastructure and human capacity to implement learning analytics ethically and effectively, including data governance frameworks, faculty training, and support services for flagged students.

In summary, these findings corroborate and extend previous work showing that LMS behavioral metrics significantly predict student performance, with assignment submission patterns and login regularity as highly influential predictors. The results contribute to self-regulated learning theory by providing empirical evidence linking observable behaviors to underlying self-regulatory processes, and they offer practical guidance for educators, course designers, and institutions seeking to use learning analytics to improve student outcomes in online courses.

7. Implications

7.1 Practical Implications

7.1.1 Early Warning Systems for Struggling Students

The most direct practical implication is the development and implementation of early warning systems using learning analytics to detect students at risk of academic failure. The Random Forest model achieved high predictive accuracy (87.0% and ROC-AUC 0.91), demonstrating that machine learning algorithms can successfully differentiate students likely to complete a course from those at risk of dropping out, using data available from the first few weeks. This capability fills a major gap identified in the problem statement.

Learning analytics-based early warning systems offer advantages over traditional monitoring methods, which depend on assessments like midterm exams that may occur weeks into a course (Arnold & Pistilli, 2012). By then, struggling students may have already fallen significantly behind. Learning analytics enables real-time monitoring of engagement markers that can foreshadow risk before formal assessments show problems. For example, students who miss early assignments or have login frequency below a threshold can be flagged within the first two to three weeks, when the opportunity for recovery is greatest.

The finding that assignment submission patterns are most predictive has specific implications for early warning system design. Submission timeliness is both real-time observable and actionable: when a student misses a deadline or submits late, the system can trigger automated notifications to both student and instructor for immediate intervention. Park and Jo (2015) found that near-real-time interventions can improve student outcomes, especially when combined with tailored feedback and support resources.

However, implementation must address ethical and logistical challenges. Institutions should develop clear policies on data collection, privacy, and use of predictive analytics to ensure students understand how their data are used and that interventions are supportive rather than punitive (Sclater et al., 2016). Early warning systems should complement human judgment, not replace it. The models are designed to aid educators by flagging students who may need attention; decisions about interventions should remain with instructors and advisors who can contextualize analytics against qualitative observations.

Additionally, appropriate response mechanisms must be in place when students are flagged as at risk. Support services (academic advising, tutoring, counseling, accessibility resources) must be aware of identified students and resourced to assist them (Siemens & Long, 2011). An early warning system that identifies risk but does not connect students to support will likely not improve outcomes and may cause frustration or anxiety. Therefore, early warning system development should be accompanied by investments in student support infrastructure and training for staff responding to alerts.

7.1.2 Improved Online Teaching Strategies

The results guide improved online teaching practices by clarifying which engagement behaviors have the highest impact on student success. The finding that assignment submission rate is the strongest predictor indicates that instructors should focus on encouraging assignment completion and time management strategies. This support may include breaking larger assignments into scaffolded tasks with interim deadlines, providing clear rubrics, or using automated reminders (Broadbent & Poon, 2015). Instructors can use LMS analytics to track submission patterns and flag students struggling with time management before they fall too far behind.

The finding that login regularity (consistency of access patterns) is more important than total login count informs how instructors think about and encourage engagement. Instead of simply motivating students to log in multiple times per week, instructors should emphasize the need for regular, consistent engagement across the entire semester. Incorporating weekly activities, consistent feedback loops, and predictable interaction rhythms can encourage steadier engagement patterns (You, 2016). Instructors can model desired behaviors by showing up consistently, responding promptly to student questions, and discussing why regular attendance matters.

Given that learning resource usage is a crucial predictor, instructors should design courses with rich, diversified learning materials that allow active interaction. Offering multiple formats (video lectures, interactive modules, readings, self-assessment quizzes) can accommodate diverse learning preferences and enhance engagement (Kovanović et al., 2015). Instructors can use analytics to see how students engage with resources and identify materials that may be particularly difficult or engaging, leading to iterative course improvement.

The weaker predictive power of forum participation in this study highlights the importance of intentional design for social learning activities. When forums are optional or marginal, students may not benefit enough from participating. To make discussion forums impactful, instructors can design forums that require meaningful contributions, assess forum participation, and set expectations for quality and volume (Wise et al., 2012). Instructors can also model effective participation and provide timely feedback.

Finally, grounded in self-regulated learning theory, instructors can promote effective performance through explicit instruction in self-regulation strategies, including goal setting, time management, metacognitive monitoring, and help-seeking behaviors (Zimmerman, 2002). Embedding self-regulation prompts into course activities (e.g., pre-assignment planning exercises, post-activity reflection questions) may mitigate the risk of students failing to succeed online.

7.2 Policy Implications

These findings have important implications for institutional policy on learning analytics adoption. As digital educational infrastructure grows, extracting actionable insights from generated data is a fundamental step in realizing online education's potential (Siemens & Long, 2011). The predictive accuracy demonstrated by LMS behavioral data provides a strong rationale for moving beyond isolated pilot projects toward institution-wide frameworks.

Broad adoption requires robust data governance regarding ownership, privacy, security, and ethical use. Students should be informed about what data are collected, how they are used, and with whom (Sclater et al., 2016). Institutions need clear protocols for informed consent, anonymization for research, and ensuring predictive models support rather than penalize students. Issues of algorithmic fairness and bias must be addressed so that models do not disproportionately flag students from particular demographic groups or perpetuate existing inequities (Ferguson, 2012).

A second policy consideration is integration of learning analytics with existing institutional systems and processes. Effective early warning systems require coordination between academic departments, student support services, IT, and institutional research. Institutions need clear pathways for how analytics-generated alerts will be acted upon to ensure timely and effective assistance (Arnold & Pistilli, 2012). This may involve rethinking traditional roles, creating new communication channels, and investing in staff training for data-informed practice.

Institutional policy should address sustainability of learning analytics initiatives. Ongoing investment in technical infrastructure, data management, and analytical expertise is required to develop and maintain predictive models. Institutions should consider resource allocation across the entire learning analytics lifecycle: data collection, model building, implementation, evaluation, and iteration (Lang et al., 2017). This may involve building specialized learning analytics teams, partnering with vendors, or developing academic expertise through training.

Finally, institutions need policies that support research and evaluation to advance the evidence base. Although this study shows predictive power in one course context, additional research is needed to test generalizability across courses, disciplines, and institutions. Institutions should develop processes for ethical access to learning analytics data, encourage faculty engagement with learning analytics research, and disseminate findings to contribute to the broader educational community (Viberg et al., 2018).

7.3 Summary

This study has implications across multiple levels of educational practice and policy. Practically, it supports the development of early warning systems for identifying at-risk students in time for meaningful intervention, and more strategic teaching practices focused on assignment submission, login regularity, and resource usage. At the policy level, it makes a case for institutional adoption of learning analytics frameworks with attention to data governance, integration with student support, sustainability, and research capacity. Collectively, these implications point toward a future where learning analytics are systematically used to support student success in online learning.

8. Limitations

Although this study provides important insights into the use of learning analytics for predicting student performance in online learning contexts, several limitations affect the generalizability and scope of these findings. Understanding these limitations is important for appropriate interpretation and for guiding future research.

8.1 Data Limited to One Institution

A major limitation is that data were collected from only one institution (a university in a developing country). Although this focus allowed detailed analysis and control over institutional variables, it severely limits generalizability to other educational contexts. Learning analytics research has shown that predictive models developed in one context may not transfer directly to another due to differences in student populations, institutional policies, course designs, and technology infrastructures (Gašević et al., 2016).

The single-institution limitation is especially pronounced given that most learning analytics research has been conducted in developed countries (Viberg et al., 2018). Although the current study provides needed evidence from a developing country, findings from one institution cannot capture the diversity of online learning environments across regions with different cultural and educational systems. Technological infrastructure, internet reliability, and digital literacy levels vary across contexts and likely affect engagement rates and the predictive utility of behavioral indicators. Additionally, the study focused on only one course at the institution. Although this controlled for instructional design and assessment variables, it limits generalizability to other disciplines, course levels, and instructional formats. A predictive model built for a specific course may not generalize to another course with different content, teaching style, or student population (Conijn et al., 2017). The specific course examined may have had unique features (e.g., optional discussion forums, assessment frequency) that influenced which behaviors emerged as strong predictors.

The sample size (n=384) was sufficient for the analyses but remains a limitation. Larger, more diverse samples would provide greater statistical power, enable more complex modeling, and increase confidence in the stability and generalizability of findings. The sample came from a single semester and may not reflect between-term or between-cohort differences. Replication across multiple semesters and courses is needed.

8.2 Limited Variables

A second important limitation relates to the variables considered. The study included four behavioral indicators from LMS logs: login frequency, assignment submission rate, forum participation, and learning resource usage. Although these are widely studied predictors (Macfadyen & Dawson, 2010; Park & Jo, 2015), they account for only a fraction of the drivers of academic outcomes. Omitted variables could provide additional explanatory power, potentially leading to omitted variable bias.

Importantly, the analysis did not control for demographic factors such as age, gender, prior academic performance, socioeconomic status, or educational background. Research has consistently shown that these factors relate to student outcomes and may interact with engagement behaviors (Tempelaar et al., 2015). The study cannot determine whether the predictive power of behavioral indicators is independent of demographic characteristics or whether certain demographic groups are systematically flagged by models. This raises questions of algorithmic fairness and whether predictive models could perpetuate existing inequities (Ferguson, 2012).

The study also omitted measures of student motivation, self-efficacy, and other psychological constructs that are theoretically relevant to self-regulated learning. Although behavioral markers are viewed as observable manifestations of self-regulation, the absence of direct assessments of motivation or metacognition prevents confirmation of the theoretical framework (Zimmerman, 2002). Future research should combine surveys with behavioral data to provide a richer understanding.

The study also lacked data on the quality of engagement, focusing only on quantitative measures. Forum participation was recorded as number of posts, not the depth or quality of contributions. Learning resource usage was quantified by time and number of resources accessed, not the nature of interaction (e.g., replay patterns, note-taking). Research by Wise et al. (2012) and Giannakos et al. (2015) suggests that qualitative dimensions of engagement are often more predictive than naive quantitative measures, so the current models may underestimate the predictive utility of engagement data.

Finally, the study did not aggregate institutional data such as academic advising records, library usage, or participation in support services. These variables could provide context for understanding student success and may moderate the engagement-outcomes relationship. Their absence limits the ability to make operational recommendations for institutional support mechanisms.

8.3 Additional Methodological Limitations

Several methodological limitations also deserve acknowledgment. The study used a retrospective design, analyzing data from an already completed course. Although this is typical in learning analytics literature, it precludes real-time intervention and dynamic adaptation. Experimental designs that evaluate predictive models in real time with actual interventions are needed to confirm causal efficacy (Sclater et al., 2016).

Binary classification of academic performance (pass/fail) is also a limitation, as it diminishes the richness of continuous grade data and may hide variation among students who pass. Future research using regression-based approaches could capture more nuanced relationships.

The study did not account for changes in engagement behaviors due to the COVID-19 pandemic; data were collected during a period when students faced disruptions to their learning environments. It is unclear how much these findings reflect normal online learning circumstances versus pandemic-induced conditions. Replication under more stable conditions would be valuable.

8.4 Final Statement

In conclusion, this study has limitations related to a single institution, a single course, a limited set of behavioral variables, and a retrospective design. These limitations affect generalizability and suggest the need for future research encompassing multiple institutional contexts, broader variable ranges, and prospective designs. Recognizing these limitations does not diminish the study’s contributions but rather places its findings in context and highlights fruitful avenues for future research.

9. Future Research Directions

Although this study has validated the usefulness of learning analytics for predicting student performance in online spaces, the limitations and findings suggest several future research directions.

9.1 Cross-Institution Datasets

A key avenue is collecting and analyzing cross-institution datasets to explore the diversity of online learning contexts across institutional types, geographic regions, and student populations. Most existing learning analytics literature, including the current study, is constrained to one institution, raising questions about generalizability (Viberg et al., 2018). Future research should establish large multi-institutional datasets to investigate whether behavioral markers have universal predictive power or vary across contexts.

Cross-institution datasets would enable several important lines of inquiry. First, researchers could examine how well predictive models transfer from one institution to another. This has practical implications because institutions adopting learning analytics often lack historical data to build models (Gašević et al., 2016). If models show generalizability, it may reduce duplicative work; if models are context-specific, each institution may need to build its own models or use domain adaptation procedures.

Second, cross-institution datasets would allow investigation of how institutional characteristics (size, selectivity, public vs. private status, resource availability) moderate the relationship between engagement behaviors and academic outcomes. Such analyses could determine which engagement indicators are more or less predictive in different contexts and help target interventions accordingly. Comparative studies across countries and cultural contexts would address the research gap that learning analytics is under-studied in developing countries. Future research should proactively include institutions from Africa, Asia, Latin America, and the Middle East to understand how cultural context, educational traditions, and technological infrastructure influence engagement and predictive validity (Tlili et al., 2020).

Cross-institution datasets require coordination and collaboration, attention to data governance, privacy, and ethical considerations. Initiatives like the Society for Learning Analytics Research (SoLAR) (Lang et al., 2017) can support shared data repositories and standards. Funding agencies can play a catalytic role.

9.2 AI-Based Adaptive Learning Systems

A second key direction is integrating predictive analytics with AI-based adaptive learning systems capable of delivering personalized, real-time interventions based on student engagement patterns. This study shows that predictive models can identify at-risk students, but it does not address how such predictions might inform practice. Future research should investigate the design and evaluation of adaptive learning systems that use engagement and performance data to make real-time adjustments (Siemens & Baker, 2012).

AI-based adaptive systems move from "one-size-fits-all" instruction toward personalized learning experiences that respond to individual student needs. These systems may incorporate machine learning for student modeling, natural language processing for analyzing student responses, and recommender systems to suggest learning resources and activities (Koedinger et al., 2015). Combined with predictive analytics, they could automatically detect learners with at-risk behavior patterns and deliver specific support (personalized feedback, recommended study strategies, guided learning paths, or links to support resources).

Specific research questions include: Which types of AI-driven interventions are most effective for at-risk students? How do adaptive systems balance automation and human oversight? What qualitative dimensions of engagement (beyond simple behavioral metrics) can adaptive systems leverage? Natural language processing to analyze forum contributions, machine learning for classifying resource interaction patterns, and computer vision for video viewing behavior could provide richer engagement measures (Giannakos et al., 2015). Long-term follow-up studies are needed to track lasting effects of adaptive systems on student outcomes across multiple courses and academic terms.

9.3 Additional Research Directions

Other directions include incorporating more variables (demographic, psychological, qualitative engagement metrics) and using mixed methods that combine quantitative analytics with qualitative insights from student interviews or instructor observations. Prospective, longitudinal designs with randomized controlled trials would provide causal evidence of effectiveness. Finally, systematic exploration of ethical dimensions – data privacy, algorithmic transparency, student agency, potential bias – is critical for responsible learning analytics practice (Ferguson, 2012).

10. Conclusion

The rapid expansion of online learning has generated vast amounts of student interaction data within Learning Management Systems. This study has demonstrated that learning analytics, grounded in self-regulated learning theory, can effectively predict student academic performance using behavioral indicators readily extracted from LMS log data. By developing and validating predictive models that leverage login frequency, assignment submission timing, forum activity, and learning resource usage, this research advances the application of predictive modeling in learning analytics.

Learning analytics is critically important in today's educational landscape. As digital learning infrastructure grows and the volume of data increases, extracting actionable insights is essential for realizing the potential of online education (Siemens & Long, 2011). Learning analytics represents a maturation of educational technology practice – moving from deploying digital tools to using data generated by those tools to understand and improve learning (Gašević et al., 2015). The discipline fills a major gap in online education: the ability to deliver early alerts that identify at-risk students before academic failure becomes inevitable.

This study identified key predictors of student performance. Assignment submission rate was the strongest predictor, highlighting that time management and goal adherence are critical aspects of self-regulated learning (Zimmerman, 2002). Students who submitted assignments on time or early consistently outperformed those who submitted late. Login frequency ranked second, with regularity of access more predictive than raw counts, indicating that sustained, consistent engagement matters more than sporadic bursts. Learning resource usage was a substantial contributor, supporting the role of active and strategic interaction with course materials as reflecting cognitive and metacognitive aspects of SRL (Kovanović et al., 2015). Forum participation was the weakest statistically significant predictor, pointing to course design as a key factor in which engagement behaviors serve as success indicators.

The practical implications are substantial. The high predictive accuracy achieved (87.0% with Random Forest, ROC-AUC 0.91) provides a foundation for early warning systems that identify at-risk students in time for adequate intervention. Such systems can flag students with concerning engagement patterns – missed assignments, infrequent logins, limited resource access – prompting timely outreach by instructors, advisors, or support services (Arnold & Pistilli, 2012). Early intervention is crucial in online learning, where students can disengage unnoticed until they are irrecoverably behind. The finding that assignment submission patterns are the strongest predictor suggests that monitoring submission timeliness from the first weeks of a course provides actionable intelligence for intervention.

Beyond early warning systems, learning analytics provides insights for improved online pedagogies and course designs. Instructors can use analytics to understand which engagement behaviors are most predictive of success in their courses and adjust teaching accordingly. Strategies that promote regular login and on-time submission, or activities structured to encourage steady study rhythms, can foster self-regulation (Broadbent & Poon, 2015). Course designers can use analytics to assess the effectiveness of different pedagogical strategies and make evidence-based decisions about course structure, assessment design, and activity pacing.

At the institutional level, learning analytics frameworks encourage data-informed decision-making and strategic improvements in student outcomes. Institutions that invest in learning analytics infrastructure can coordinate support services across campus and evaluate intervention effectiveness at scale (Sclater et al., 2016). Development of such systems requires attention to data governance, privacy, ethics, and integration with existing processes. When implemented appropriately, learning analytics can improve institutional capacity to support student success while maintaining trust and transparency.

This study also emphasizes the need for context in learning analytics research. The weaker predictive power of forum participation appears partly due to course-specific factors (optional discussion forums). This underscores the importance of institutions developing risk models based on their specific populations. The limitations of this study – particularly the single-institution focus and narrow variable set – point to important future research directions, including cross-institution datasets, integration of AI-based adaptive learning systems, and evaluation of algorithmic fairness and ethical considerations.

Ultimately, learning analytics offers a fundamental way to understand and support student success in online learning. By leveraging behavioral data produced in routine LMS interactions to build predictive models that can identify at-risk students, tailor teaching strategies, and promote evidence-based decision-making, learning analytics contributes to the realization of more responsive and student-centered digital education. The findings show that assignment submission patterns and regular login frequency are the strongest behavioral predictors of academic performance, providing practical guidance for educators, instructional designers, and institutions. As online learning continues to expand globally, the appropriate use of learning analytics will be essential to help every student reach their potential. The potential benefits – early intervention for struggling students, improved instructional quality, and enhanced institutional capacity to support student success – position learning analytics as a prominent research and practice priority for digital education.

Declarations

Competing Interests

The author declares no competing interests. No financial, personal, or professional relationships have influenced the design, execution, or reporting of this research. The study was conducted independently without funding from any organization that might have a stake in the findings.

Author's Contribution

The author is the sole contributor to this work and is responsible for all aspects of the research. This includes conceptualization, study design, data collection, data analysis, interpretation of results, drafting of the manuscript, and critical revision. The author has reviewed and approved the final version of the manuscript and agrees to be accountable for all aspects of the work.

Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request. The data are not publicly available due to privacy and ethical restrictions, as they contain sensitive student interaction logs and assessment records that could compromise participant confidentiality. Access to anonymized data may be granted to qualified researchers for academic purposes, subject to approval by the institutional ethics committee and execution of a data sharing agreement.

References

Arnold, K. E., & Pistilli, M. D. (2012). Course signals at Purdue: Using learning analytics to increase student success. *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge*, 267–270. https://doi.org/10.1145/2330601.2330666

Azevedo, R. (2015). Defining and measuring engagement and learning in science: Conceptual, theoretical, methodological, and analytical issues. *Educational Psychologist*, 50(1), 84–94. https://doi.org/10.1080/00461520.2015.1004069

Baker, R. S., & Inventado, P. S. (2014). Educational data mining and learning analytics. In J. A. Larusson & B. White (Eds.), *Learning analytics: From research to practice* (pp. 61–75). Springer. https://doi.org/10.1007/978-1-4614-3305-7_4

Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory*. Prentice-Hall.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

Broadbent, J., & Poon, W. L. (2015). Self-regulated learning strategies & academic achievement in online higher education learning environments: A systematic review. *The Internet and Higher Education*, 27, 1–13. https://doi.org/10.1016/j.iheduc.2015.04.007

Coates, H., James, R., & Baldwin, G. (2005). A critical examination of the effects of learning management systems on university teaching and learning. *Tertiary Education and Management*, 11(1), 19–36. https://doi.org/10.1080/13583883.2005.9967137

Conijn, R., Snijders, C., Kleingeld, A., & Matzat, U. (2017). Predicting student performance from LMS data: A comparison of 17 blended courses using Moodle LMS. *IEEE Transactions on Learning Technologies*, 10(1), 17–29. https://doi.org/10.1109/TLT.2016.2616312

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273–297. https://doi.org/10.1007/BF00994018

Creswell, J. W., & Creswell, J. D. (2018). *Research design: Qualitative, quantitative, and mixed methods approaches* (5th ed.). SAGE Publications.

Dawson, S. (2006). A study of the relationship between student communication interaction and sense of community. *The Internet and Higher Education*, 9(3), 153–162. https://doi.org/10.1016/j.iheduc.2006.06.007

Dhawan, S. (2020). Online learning: A panacea in the time of COVID-19 crisis. *Journal of Educational Technology Systems*, 49(1), 5–22. https://doi.org/10.1177/0047239520934018

Ferguson, R. (2012). Learning analytics: Drivers, developments and challenges. *International Journal of Technology Enhanced Learning*, 4(5-6), 304–317. https://doi.org/10.1504/IJTEL.2012.051816

Field, A. (2018). *Discovering statistics using IBM SPSS statistics* (5th ed.). SAGE Publications.

Gašević, D., Dawson, S., & Siemens, G. (2015). Let's not forget: Learning analytics are about learning. *TechTrends*, 59(1), 64–71. https://doi.org/10.1007/s11528-014-0822-x

Gašević, D., Dawson, S., Rogers, T., & Gasevic, D. (2016). Learning analytics should not promote one size fits all: The effects of instructional conditions in predicting academic success. *The Internet and Higher Education*, 28, 68–84. https://doi.org/10.1016/j.iheduc.2015.10.002

Giannakos, M. N., Chorianopoulos, K., & Chrisochoides, N. (2015). Making sense of video analytics: Lessons learned from clickstream interactions, attitudes, and learning outcome in a video-assisted course. *International Review of Research in Open and Distributed Learning*, 16(1), 260–283. https://doi.org/10.19173/irrodl.v16i1.1976

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2019). *Multivariate data analysis* (8th ed.). Cengage Learning.

Hart, C. (2012). Factors associated with student persistence in an online program of study: A review of the literature. *Journal of Interactive Online Learning*, 11(1), 19–42. https://www.ncolr.org/jiol/issues/view/issue/11/1

Hellas, A., Ihantola, P., Petersen, A., Ajanovski, V. V., Gutica, M., Hynninen, T., Knutas, A., Leinonen, J., Messom, C., & Liao, S. N. (2018). Predicting academic performance: A systematic literature review. *Proceedings Companion of the 23rd Annual ACM Conference on Innovation and Technology in Computer Science Education*, 175–199. https://doi.org/10.1145/3293881.3295783

Henrie, C. R., Halverson, L. R., & Graham, C. R. (2015). Measuring student engagement in technology-mediated learning: A review. *Computers & Education*, 90, 36–53. https://doi.org/10.1016/j.compedu.2015.09.005

Hodges, C., Moore, S., Lockee, B., Trust, T., & Bond, A. (2020). The difference between emergency remote teaching and online learning. *Educause Review*, 27, 1–12. https://er.educause.edu/articles/2020/3/the-difference-between-emergency-remote-teaching-and-online-learning

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (3rd ed.). John Wiley & Sons. https://doi.org/10.1002/9781118548387

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning: With applications in R* (2nd ed.). Springer. https://doi.org/10.1007/978-1-0716-1418-1

Januszewski, A., & Molenda, M. (Eds.). (2008). *Educational technology: A definition with commentary*. Routledge.

Joksimović, S., Gašević, D., Loughin, T. M., Kovanović, V., & Hatala, M. (2016). Learning at distance: Effects of interaction traces on academic achievement. *Computers & Education*, 102, 154–167. https://doi.org/10.1016/j.compedu.2016.09.002

Kim, J., Li, S., & Bonk, C. J. (2014). Factors affecting online learner success: A cross-institutional study. *Journal of Interactive Learning Research*, 25(4), 549–573. https://www.learntechlib.org/primary/p/147280/

Koedinger, K. R., D'Mello, S., McLaughlin, E. A., Pardos, Z. A., & Rosé, C. P. (2015). Data mining and education. *Wiley Interdisciplinary Reviews: Cognitive Science*, 6(4), 333–353. https://doi.org/10.1002/wcs.1350

Kovanović, V., Gašević, D., Dawson, S., Joksimović, S., & Baker, R. S. (2015). Does time-on-task estimation matter? Implications for the validity of learning analytics findings. *Journal of Learning Analytics*, 2(3), 81–110. https://doi.org/10.18608/jla.2015.23.6

Kuh, G. D. (2009). The national survey of student engagement: Conceptual and empirical foundations. *New Directions for Institutional Research*, 2009(141), 5–20. https://doi.org/10.1002/ir.283

Lang, C., Siemens, G., Wise, A., & Gašević, D. (Eds.). (2017). *Handbook of learning analytics*. Society for Learning Analytics Research. https://doi.org/10.18608/hla17

Lee, Y., & Choi, J. (2011). A review of online course dropout research: Implications for practice and future research. *Educational Technology Research and Development*, 59(5), 593–618. https://doi.org/10.1007/s11423-010-9177-y

Macfadyen, L. P., & Dawson, S. (2010). Mining LMS data to develop an "early warning system" for educators: A proof of concept. *Computers & Education*, 54(2), 588–599. https://doi.org/10.1016/j.compedu.2009.09.008

Miles, M. B., Huberman, A. M., & Saldaña, J. (2020). *Qualitative data analysis: A methods sourcebook* (4th ed.). SAGE Publications.

Morris, L. V., Finnegan, C., & Wu, S. S. (2005). Tracking student behavior, persistence, and achievement in online courses. *The Internet and Higher Education*, 8(3), 221–231. https://doi.org/10.1016/j.iheduc.2005.06.009

Murphy, E., & Rodríguez-Manzanares, M. A. (2009). Teachers' perspectives on motivation in high-school distance education. *Journal of Distance Education*, 23(3), 1–24. https://www.ijede.ca/index.php/jde/article/view/543

Osmanbegović, E., & Suljić, M. (2012). Data mining approach for predicting student performance. *Economic Review*, 10(1), 3–12. https://www.ef.untz.ba/wp-content/uploads/2017/06/ER_2012_10_1_osmanbegovic.pdf

Park, Y., & Jo, I. H. (2015). Development of the learning analytics dashboard to support students' learning performance. *Journal of Universal Computer Science*, 21(1), 110–133. https://doi.org/10.3217/jucs-021-01-0110

Patterson, B., & McFadden, C. (2009). Attrition in online and campus degree programs. *Online Journal of Distance Learning Administration*, 12(2), 1–8. https://ojdla.com/archive/spring82/patterson82.pdf

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830. https://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html

Pintrich, P. R. (2004). A conceptual framework for assessing motivation and self-regulated learning in college students. *Educational Psychology Review*, 16(4), 385–407. https://doi.org/10.1007/s10648-004-0006-x

Powers, D. M. W. (2020). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *Journal of Machine Learning Technologies*, 2(1), 37–63. https://arxiv.org/abs/2010.16061

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, 1(1), 81–106. https://doi.org/10.1007/BF00116251

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(3), e1355. https://doi.org/10.1002/widm.1355

Sclater, N., Peasgood, A., & Mullan, J. (2016). *Learning analytics in higher education: A review of UK and international practice*. Jisc. https://www.jisc.ac.uk/reports/learning-analytics-in-higher-education

Siemens, G., & Baker, R. S. J. d. (2012). Learning analytics and educational data mining: Towards communication and collaboration. *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge*, 252–254. https://doi.org/10.1145/2330601.2330661

Siemens, G., & Long, P. (2011). Penetrating the fog: Analytics in learning and education. *Educause Review*, 46(5), 30–32. https://er.educause.edu/articles/2011/9/penetrating-the-fog-analytics-in-learning-and-education

Tabachnick, B. G., & Fidell, L. S. (2019). *Using multivariate statistics* (7th ed.). Pearson Education. [no DOI]

Tempelaar, D. T., Rienties, B., & Giesbers, B. (2015). In search for the most informative data for feedback generation: Learning analytics in a data-rich context. *Computers in Human Behavior*, 47, 157–167. https://doi.org/10.1016/j.chb.2014.05.038

Tlili, A., Zhang, J., Papamitsiou, Z., Manske, S., Huang, R., Kinshuk, & Hoppe, H. U. (2020). Towards utilising emerging technologies to address the challenges of using Open Educational Resources: A vision of the future.

Educational Technology Research and Development, 68(2), 789–809. https://doi.org/10.1007/s11423-019-09732-4

Turnbull, D., Chugh, R., & Luck, J. (2021). Learning management systems: A review of the research literature. *Journal of Information Technology Education: Research*, 20, 99–121. https://doi.org/10.28945/4689

Viberg, O., Hatakka, M., Bälter, O., & Mavroudi, A. (2018). The current landscape of learning analytics in higher education. *Computers in Human Behavior*, 89, 98–110. https://doi.org/10.1016/j.chb.2018.07.027

Winne, P. H., & Hadwin, A. F. (2008). The weave of motivation and self-regulated learning. In D. H. Schunk & B. J. Zimmerman (Eds.), *Motivation and self-regulated learning: Theory, research, and applications* (pp. 297–314). Lawrence Erlbaum Associates.

Wise, A. F., Speer, J., Marbouti, F., & Hsiao, Y. T. (2012). Broadening the notion of participation in online discussions: Examining patterns in learners' online listening behaviors. *Instructional Science*, 41(2), 323–343. https://doi.org/10.1007/s11251-012-9230-9

World Economic Forum. (2021). *The future of online learning: How education technology is reshaping global education*. World Economic Forum. https://www.weforum.org/reports/the-future-of-online-learning/

You, J. W. (2016). Identifying significant indicators using LMS data to predict course achievement in online learning. *The Internet and Higher Education*, 29, 23–30. https://doi.org/10.1016/j.iheduc.2015.11.003

Yu, T., & Jo, I. H. (2014). Educational technology approach toward learning analytics: Relationship between student online learning behavior and academic performance. *Proceedings of the 4th International Conference on Learning Analytics and Knowledge*, 269–270. https://doi.org/10.1145/2567574.2567594

Zimmerman, B. J. (2002). Becoming a self-regulated learner: An overview. *Theory Into Practice*, 41(2), 64–70. https://doi.org/10.1207/s15430421tip4102_2