Enhancing Student Performance Prediction in e-Learning Ecosystems Using Machine Learning Techniques

Fatima Ezzahraa EL Habti[1,*], Mustafa Hiri[1], Mohamed Chrayah[2], Abdelhamid Bouzidi[1], and Noura Aknin[1]

[1]TIMS Laboratory, Faculty of sciences Tetouan, Abdelmalek Essaadi University Morocco, Morocco
[2]TIMS Laboratory, ENSA Tetouan, Abdelmalek Essaadi University Morocco, Morocco
Email: fatimaezzahraaelhabti@gmail.com (F.Z.E.H.); mustafa.hiri@gmail.com (M.H.); chrayah@gmail.com (M.C.); mr.abdelhamid.bouzidi@gmail.com (A.B.); noura.aknin@uae.ac.ma (N.A.)
[*]Corresponding author

Manuscript received September 28, 2024; revised November 1, 2024; accepted November 22, 2024; published February 14, 2025

***Abstract*—Accurately predicting student performance in e-learning environments is a significant challenge that is essential for personalizing education and enhancing learning outcomes. This study examines the effectiveness of machine learning techniques in forecasting learner success within e-learning ecosystems, using the Open University Learning Analytics Dataset (OULAD). We conducted a comparative analysis of four machine learning algorithms—Random Forest, Logistic Regression, Support Vector Machine (SVM), and Linear Discriminant Analysis (LDA)—employing comprehensive data pre-processing and feature engineering methods. The Random Forest algorithm outperformed the others, achieving a 91% accuracy rate in classifying student outcomes into "Distinction," "Pass," and "Fail" categories. Despite this high accuracy, differentiating between certain performance classes, especially "Distinction" and "Fail," remained challenging, highlighting the complexity of student performance metrics in online learning contexts. These results demonstrate the potential of machine learning, particularly the Random Forest algorithm, as a valuable tool for enhancing predictive analytics in e-learning systems. The study contributes to the optimization of educational technologies by indicating that refined predictive models can lead to more effective, data-driven interventions and personalized learning pathways.**

***Keywords*—machine learning, predictive analysis, e-learning, student performance, learning analytics, random forest algorithm**

I. Introduction

In the era of digital transformation, the education sector has radically evolved, moving from traditional classrooms to sophisticated online learning ecosystems [1]. These ecosystems offer unprecedented possibilities for learners and educators, marking the beginning of a new era of learning characterized by accessibility, flexibility, and interactivity [2].

Several key elements, including content providers, learners (content consumers), consultants, infrastructure, and learning materials [3], play crucial roles in shaping this enriching learning experience [4]. In the e-learning ecosystem, these elements interact dynamically, forming a living network that promotes intellectual development and facilitates the exchange of knowledge [5, 6].

Understanding learner behavior within the e-learning ecosystem is essential for optimizing learning outcomes and improving the overall effectiveness of e-learning [7, 8]. Learner behavior, which encompasses interactions with learning materials, levels of engagement, assessment results, demographic information, and other parameters, plays a key role in shaping individual and collective learning experiences. By analyzing these behavioral patterns and their relationships with other ecosystem components, valuable insights can be gained into learner performance, allowing for tailored interventions to meet specific learning needs [9, 10].

Integrating machine learning into the e-learning ecosystem is a promising way to achieve these goals [11]. By analyzing data on learner behavior, engagement patterns, and performance metrics, machine learning algorithms can identify trends and make predictions, enabling personalized learning experiences and targeted interventions that enhance educational outcomes. This article explores the application of machine learning within the e-learning ecosystem, utilizing the OULAD dataset, a comprehensive repository of student data, to demonstrate the potential of these techniques in improving educational experiences and outcomes.

This study provides a novel approach by applying and comparing four distinct machine learning models—Random Forest, Logistic Regression, SVM, and LDA—to predict student outcomes within an e-learning ecosystem. Unlike prior studies that typically focus on a single model or less robust datasets, this study utilizes comprehensive feature engineering and preprocessing techniques to handle large-scale educational data effectively. The findings offer valuable insights into personalized learning interventions, as the models enable accurate predictions of distinct outcomes like 'Distinction,' 'Pass,' and 'Fail,' which are crucial for adaptive learning systems. By achieving high prediction accuracy with the Random Forest model, this study provides valuable insights into model selection for data-driven interventions, supporting educators in tailoring personalized learning experiences.

The main objective of this study is to develop machine learning models capable of predicting learner performance in an online learning environment. By analyzing various data such as learner demographic information, interactions with learning platforms, assessment results, and engagement in course activities, this research aims to identify factors that influence learner success and propose personalized interventions to improve their learning experience and outcomes.

Additionally, this research contributes to the wider field of educational technology by providing empirical evidence of the effectiveness of machine learning models in predicting student performance. By analyzing the OULAD, this study aims not only to improve educational outcomes but also to

enrich the literature on learning analytics and machine learning applications in education. The results of this study should enable educators, course designers, and policymakers to adapt educational content, interventions, and support services to meet the needs of learners in e-learning ecosystems.

The paper is structured as follows: following the introduction, Section II presents the background and related work, exploring the theoretical foundations and prior research on the application of machine learning in e-learning ecosystems. Section III describes the dataset and methodology, detailing the data preprocessing, analysis, and evaluation of machine learning models, including Random Forest, Logistic Regression, SVM, and LDA. Section IV presents the results of our analysis, highlighting the performance of each model, with particular emphasis on the Random Forest algorithm. Section V discusses these results, interpreting their significance for the e-learning ecosystem and suggesting directions for future research. Finally, Section VI concludes the paper by summarizing the main findings and proposing future research avenues in the field of machine learning within e-learning ecosystems.

II. Literature Review

A. Evolution of e-Learning Ecosystems

The educational landscape has undergone significant transformations over the last few decades, driven by rapid technological advancements and evolving societal needs [12]. This section explores the historical development and evolution of e-learning ecosystems, from traditional classrooms to today's sophisticated online platforms.

The shift from traditional classrooms to online learning platforms represents a major change in teaching methodologies. Initially, e-learning emerged as a solution to overcome geographical and temporal barriers, providing learners with global access to educational content [13]. Over time, technological advancements have transformed e-learning into interactive, learner-centered ecosystems [14, 15]. Modern digital platforms now integrate multimedia content, collaboration tools, and communication features such as forums, chats, and videoconferencing, catering to diverse learning preferences and styles [16].

The development of Learning Management Systems (LMS) and other educational technologies has further accelerated this transition, enabling educators to create, distribute, and manage educational content more efficiently [17, 18]. These systems provide a structured environment for online learning, supporting both synchronous and asynchronous learning activities [4, 19].

In summary, the e-learning ecosystem has rapidly evolved from a solution to geographical barriers into a platform rich with pedagogical possibilities. This evolution has made education more accessible, adaptable, and focused on individual learners' needs, profoundly transforming both teaching methods and the educational experience as a whole.

B. Overview of e-Learning Ecosystems

The e-learning ecosystem is a major evolution in digital learning, representing a comprehensive approach that includes all the elements necessary to support e-learning. This concept, viewed as the next step in digital education, combines e-learning with the ecosystem metaphor, emphasizing the importance of integrating technologies, teaching methods, and participant interactions to create an effective, holistic learning experience.

1) E-Learning

E-learning is at the forefront of modern education [20], embodying the fusion of pedagogy and digital innovation. It refers to a teaching method that uses digital technologies to deliver courses and training remotely [21]. Its primary objective is to offer flexible access to education, enabling learners to take courses at their own pace and according to their schedules [20, 22]. E-learning transcends geographical barriers, offering learners access to educational resources and experiences via digital platforms and tools [23]. It allows learning paths to be tailored to individual needs, reducing the time and distance constraints associated with traditional face-to-face education [24]. E-learning also offers a diverse range of educational resources (text, video, audio, articles, podcasts, quizzes) and methods (virtual classrooms, MOOCs, social learning, mobile learning, discussion forums, online conferences) [25], making it possible to learn anytime, anywhere.

E-learning can occur synchronously or asynchronously [26]. In synchronous learning, learners participate in real-time online sessions with instructors and other learners [27]. In asynchronous learning, they can access resources and complete activities at their own pace, without real-time constraints [27]. Both formats offer flexibility to accommodate the diverse needs of learners.

2) Ecosystem

The term "ecosystem" originates from biology and was first coined by British botanist Arthur George Tansley in 1935 [28]. It refers to a community of living and non-living entities interacting with each other and their environment [29]. Ecosystems are characterized by features such as self-organization, self-regulation, dynamics, flexibility, evolution, and collaboration [30], all of which facilitate the flow and exchange of energy and matter necessary for the maintenance and sustainability of life [30]. This concept has been metaphorically extended to fields such as business, technology, industry, and education [31].

In the context of e-learning, an ecosystem consists of various elements—learners, educators, content providers, platforms, tools, and policies—interconnected and interdependent, working together to support learning activities and outcomes [32]. Like a natural ecosystem, the e-learning ecosystem is characterized by complexity, adaptability, and the symbiotic relationships among its components.

3) E-Learning ecosystem

The term "e-learning ecosystem" encapsulates the holistic nature of digital learning environments, highlighting the intricate web of interactions between learners, educators, content providers, platforms, and infrastructures [33]. It emphasizes that successful e-learning experiences depend not only on the provision of content but also on the creation of

immersive and engaging environments that cater to diverse learning styles and preferences. By conceptualizing e-learning as an ecosystem, stakeholders can foster environments that promote the creation, exchange, and application of knowledge [33].

The e-learning ecosystem comprises multiple interdependent components that collectively enhance the learning experience [33–35]. The key players are:

Content providers: These include educators and institutions that create, curate, and deliver the learning materials essential for learner engagement and success. Their contributions range from traditional course materials to innovative digital resources, covering a wide array of subjects and learning styles.

Content consumers (learners): At the center of the ecosystem, learners interact with the provided content to acquire knowledge and skills. This group includes students from diverse backgrounds and educational levels, each with unique learning objectives and needs, making personalized learning a critical feature of the e-learning ecosystem.

Consultants: These experts provide technical and pedagogical support, ensuring that the ecosystem's offerings are both accessible and effective. Their advice ranges from technical assistance to curriculum development.

Infrastructure: The backbone of the e-learning ecosystem, infrastructure includes LMS platforms, digital tools, and network services that support content delivery, communication, collaboration, and learner progress tracking.

Learning Materials: A diverse range of educational resources, from textbooks to interactive simulations and multimedia elements, designed to engage learners and accommodate different learning preferences, enhancing the overall educational experience.

Together, these elements form a dynamic and integrated network that supports personalized learning pathways, highlighting the e-learning ecosystem's adaptability to individual learner needs and goals.

C. Literature Review on the Application of Machine Learning in e-Learning

Machine learning applications in e-learning have significantly evolved, driven by the increasing availability of educational data and advancements in analytical techniques. Previous research has laid a solid foundation for using machine learning to improve educational outcomes, presenting various approaches and methodologies. One key area of study involves the development of predictive models for student performance and engagement [36]. Researchers have investigated factors such as demographic data, prior academic performance, learning behaviors, and socio-economic background to predict academic success [37]. These models have proven useful in identifying at-risk students, enabling timely interventions and support mechanisms [38].

Additionally, research has focused on personalized learning systems that adjust content and activities based on individual learner needs and preferences [39]. Machine learning algorithms analyze learner interactions, preferences, and performance data to offer tailored learning experiences [40]. Adaptive learning platforms powered by artificial intelligence dynamically adjust content difficulty, pace, and presentation based on real-time feedback and assessments [41].

Machine learning techniques have also been applied to automated assessment and feedback generation [42]. Natural language processing algorithms assess written responses, essays, and discussion forum posts, providing instant feedback to learners and educators. Automated scoring systems, powered by machine learning, allow scalable and efficient processing of large assessment volumes while maintaining reliability and validity [42].

Overall, research on the application of machine learning in e-learning shows great potential for improving educational experiences and learner outcomes. However, challenges related to data privacy, algorithmic bias, interpretability, and scalability persist, requiring ongoing interdisciplinary collaboration and ethical considerations in future projects [43].

III. Dataset and Methods

The main objective of this study is to develop machine learning models capable of predicting learner performance in an e-learning environment. By analyzing a variety of data, including learner demographic information, interactions with learning platforms, assessment results, and engagement in course activities, this research aims to identify factors that influence learner success and propose personalized interventions to improve their learning experiences and outcomes.

This study employs a quantitative research approach to examine the impact of machine learning techniques on improving the e-learning ecosystem. Using the OULAD dataset, the goal is to develop predictive models that can accurately forecast learner performance and engagement. Specifically, the research explores four distinct machine learning algorithms to determine the most effective model for predicting outcomes in our e-learning context. This approach aims to contribute to more personalized and effective learning experiences by identifying the algorithm that best captures the complexities of learner data and interactions within the e-learning ecosystem.

A. Dataset Description

The dataset used in this study is the OULAD [44], which provides a comprehensive repository of student data collected from various courses offered by the Open University. The dataset includes demographics, course interactions (aggregated clickstream data of student interactions in the Virtual Learning Environment (VLE)), assessment results, engagement metrics, and course materials. It covers 22 courses, 32,593 students, their assessment results, and logs of their interactions with the VLE, represented by daily summaries of student clicks (10,655,280 entries). This data is utilized to train and test predictive models aimed at improving educational outcomes and personalizing learning experiences.

The dataset is structured into several tables, each capturing specific aspects of student behavior and performance. For simplicity, an integrated view of the key variables is presented in Table 1. To provide a glimpse of the dataset, a sample of

actual data entries is shown in Table 2, illustrating the structure and content used for training and testing predictive models.

Table 1. Overview of key variables in the OULAD

VARIABLE NAME	DESCRIPTION
Code_module	course module identifier
Code_presentation	course session identifier
Id_student	Unique student identifier
Gender	gender of student
Region	geographic region of student's residence
High_education	Student's highest level of education at entry
Imd_band	multiple deprivation index band for place of residence
Age_band	student's age group
Num_of_prev_attempts	number of previous attempts at the module by the student
Students_credits	total number of credits for the modules studied by the student
Handicap	indicates whether the student has declared a disability
Final_result	final results of the student in the module presentation
Date_registration	Number of days between the student's registration on the module presentation and the course start date
Date_unregistration	Number of days between unregistration and the course start date, if the student withdrew before completion
id_assessment	Unique ID for assessment
assessment_type	Type of assessment (TMA, CMA, Exam)
Date_submit	The date of student submission, measured as the number of days since the start of the module presentation
Score	The student's score in this assessment. The range is from 0 to 100. The score <40 is interpreted as Fail
Sum_click	Number of times a student interacts with the material during the day

Table 2. Sample entries from OULAD dataset illustrating student data structure

Code_module	Code_presentation	Id_student	Gender	Region	Age_band	Final_result	Sum_clik	Date_registration	Score
AAA	2013J	123456	M	North West	30-39	Pass	120	50	78
BBB	2014J	789012	F	South East	40-49	Distinction	340	10	92
CCC	2013B	345678	M	London	20-29	Fail	90	15	35

B. Methods

Our study follows the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology [45] to apply machine learning techniques aimed at improving outcomes within the e-learning ecosystem. This structured approach is illustrated in Fig. 1 below, which visually outlines the essential phases of our methodology. These phases guide the entire process, from initial data understanding to the deployment of the final model.

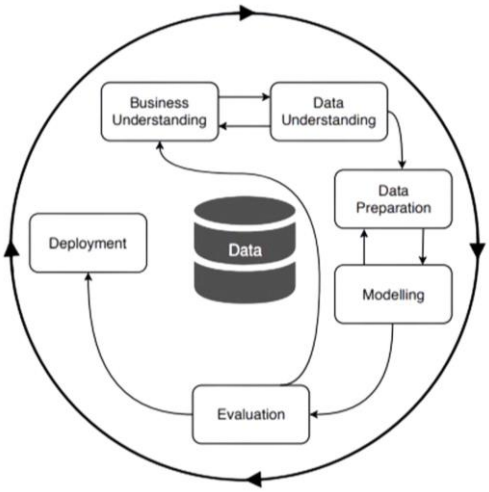

Fig. 1. CRISP-DM process model applied in machine learning for e-learning ecosystem [46].

1) Business understanding

The Business Understanding phase is crucial in defining the scope and objectives of our study within the context of the e-learning ecosystem. This initial step involves identifying the specific educational goals that our machine learning project aims to support, such as enhancing student engagement, improving assessment outcomes, and personalizing learning experiences. We collaborate with educational stakeholders to gather requirements and define Key Performance Indicators (KPIs) for student success. By understanding these elements, we ensure that our data mining efforts are aligned with educational objectives and designed to address real-world educational challenges effectively. This phase establishes a clear framework for what the machine learning models need to achieve, guiding the selection and application of analytical methods and tools in the subsequent phases.

2) Data understanding

This phase begins with the initial loading of data from the OULAD. We analyze the structure of the dataset, identifying key variables such as student age, assessment scores, and interaction data. Understanding these variables allows us to formulate hypotheses about the factors influencing student success and engagement.

3) Data preparation

Data preparation is a critical step in the machine learning process, essential for transforming raw data into a structured and usable format suitable for modeling. This phase involves data loading, transformation, cleaning, and organization to ensure compatibility with machine learning algorithms. Each step is meticulously designed to improve the quality and reliability of the data, facilitating robust and accurate predictive modeling.

The initial stage involves extracting information from structured sources, typically formatted into tables where each row represents an observation (e.g., a student or interaction) and each column corresponds to a specific variable (e.g., student age, assessment score). The extracted data undergoes preliminary processing, including categorizing assessments by type due to their distinct properties and impacts on student performance. For instance, exams—often key indicators of subject mastery—are treated separately from other types of assessments.

Assessments are tallied for each course and session, providing insights into the required level of academic engagement and helping weigh overall student performance. Success on these assessments is determined using a threshold function, where a score exceeding a predefined pass mark (typically 40%) is considered successful, in line with institutional academic standards. Furthermore, student scores are adjusted by weight, acknowledging that not all assessments contribute equally to the final grade.

Engagement metrics, such as the average number of clicks per resource, serve as key indicators of student interaction with online learning materials. These metrics are crucial for understanding student engagement levels. Students who withdraw before course completion are excluded from these calculations to ensure that performance metrics accurately reflect the outcomes of those who complete their learning journey.

The next phase involves merging all relevant information into a single DataFrame for comprehensive analysis. This integrated DataFrame combines data from multiple sources and tables, covering various aspects of student behavior and performance. Data cleansing is performed to remove inconsistencies or missing values, ensuring the reliability of the modeling process. Additionally, unnecessary or redundant columns are removed to streamline the dataset, enhancing its usability for machine learning applications.

Fig. 2 outlines the essential steps in the data preparation process, visually representing the sequence from data extraction to the final preparation of data for modeling.

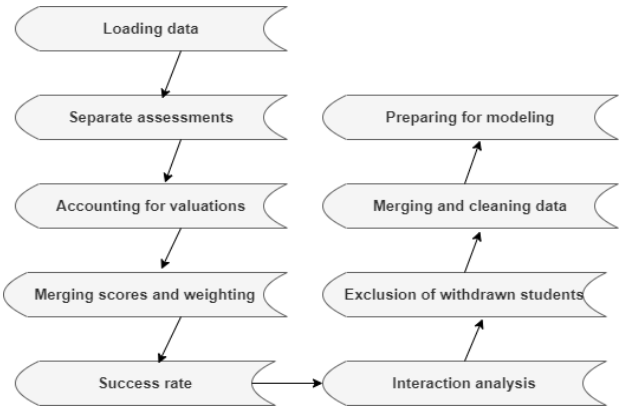

Fig. 2. Steps in data preparation for machine learning.

After cleansing, the final DataFrame is meticulously prepared for predictive modeling. This preparation involves transforming and structuring the data to facilitate the application of machine learning techniques. The final DataFrame, which contains 4950 rows, includes variables that capture the complex relationships between student characteristics, engagement, and performance. Table 3 provides an overview of the key variables included in the final DataFrame:

Careful data preparation is essential to ensure the accuracy of predictive models and their capacity to generate meaningful insights. This process reflects the rigor necessary for data manipulation and analysis in online education, aiming to enhance learning experiences and outcomes for students through evidence-based interventions.

Table 3. Variables of the final DataFrame used for predictive modeling

Variables	Description
Num_of_prev_attempts	Number of previous attempts of the module by the student.
Weighted_grade	Student's weighted average grade
Pass_rate	Student success rate, perhaps as a percentage or normalized score.
Score_exam	Student's final exam score.
Date	Date associated with the assessment or interaction, expressed in number of days.
Sum_click	Total number of student interactions with VLE material on a given day.
Final_result	The model's target variable, which indicates the student's final result in the module, such as "Pass", "Fail" or "Distinction".

4) Modeling

Once the data has been prepared, we select and apply appropriate machine learning models to predict student performance. This phase involves an iterative process of model selection, parameter tuning and training. We chose models including Random Forest, SVM, Logistic Regression and LDA, is based on their ability to handle the specific data types and patterns identified during the data understanding phase. To assess model performance, we use accuracy, precision, recall, and F1-score, enabling a multi-dimensional evaluation of classification accuracy for student outcomes. In the following paragraphs, we detail the modeling process and results for each of these algorithms, discussing their individual performances and the insights gained from their application.

5) Evaluation

For this study, we selected four machine learning algorithms—Random Forest, Logistic Regression, SVM, and LDA—based on their unique strengths in handling educational data and their suitability for classification tasks. Random Forest is known for its robustness and ability to reduce overfitting by aggregating multiple decision trees [47]. Logistic Regression is well-regarded for providing probabilistic outputs, which are beneficial for multiclass classification tasks [48]. SVM is particularly effective in high-dimensional spaces and seeks to maximize the margin between classes, making it suitable for distinguishing between performance levels [49]. Finally, LDA enhances class separability, facilitating classification into distinct categories [50] such as 'Distinction,' 'Pass,' and 'Fail.' These models offer a diverse basis for comparison, allowing us to evaluate their effectiveness in predicting student performance outcomes.

To assess the effectiveness of each model, we utilized four key performance metrics: accuracy, precision, recall, and F1-score. Accuracy provides an overall measure of correct predictions, precision assesses the accuracy of positive predictions within each category, recall reflects the model's sensitivity in identifying actual instances, and F1-score balances precision and recall, providing a robust measure when class distributions vary [51]. These metrics offer a comprehensive basis for evaluating the classification performance of each model.

6) Deployment

The deployment phase consists of implementing the

best-performing model in a real educational environment, such as within a Learning Management System (LMS) or as a standalone application accessible to educators and administrators. We continuously monitor the model's performance using relevant educational metrics, such as student engagement, assessment scores, and completion rates, to evaluate whether the model is effectively enhancing learning outcomes. Feedback from educators and students is collected periodically to assess user satisfaction and identify any necessary adjustments. Based on this feedback and ongoing performance data, we make optimizations and recalibrate the model as needed, ensuring its continued relevance and effectiveness over time. This deployment and improvement cycle are the focus of our future work, where we aim to refine these processes further to adapt to the dynamic needs of e-learning environments.

IV. Results

In our exploration of machine learning methods for enhancing the e-learning ecosystem, we have entered a critical phase: predictive modeling. This step has enabled us to develop models capable of predicting student performance with notable accuracy. To achieve this, we divided our dataset into two parts: 80% for training and 20% for testing.

This strategic split ensures that our models are trained on a representative portion of the data while being evaluated on a separate set, thereby verifying their ability to generalize and perform under realistic conditions. The test set, distinct from the training data, acts as an unbiased measure of the model's effectiveness when faced with new data.

The models selected for this comparative analysis include: Random Forest, Logistic Regression, SVM, and LDA. These models were chosen due to their widespread recognition and proven performance across many areas of machine learning, making them reliable choices for this study. They have been tested and validated in various contexts, consistently yielding strong results.

A. Random Forest

Random Forest is an ensemble model consisting of multiple individual decision trees, each trained on a subset of the data [47]. For predictions, Random Forest aggregates the outcomes from all decision trees and either follows the majority vote for classification or takes the average prediction for regression tasks [47]. This model is well-known for its robustness and ability to mitigate overfitting, due in part to the diversity of its trees and the bagging (Bootstrap Aggregating) method it employs.

1) Confusion matrix

The performance of the Random Forest model is summarized visually in the confusion matrix shown in Fig. 3. This matrix illustrates the number of true and false predictions for each class, providing an immediate assessment of the model's performance. Since our objective is classification rather than regression or forecasting, the confusion matrix is appropriate for evaluating how well the model differentiates between the discrete classes. For predictive tasks involving continuous outcomes, metrics such as MAE, MSE, RMSE, and R-squared are commonly used. However, for this classification task, we focus on classification-specific metrics.

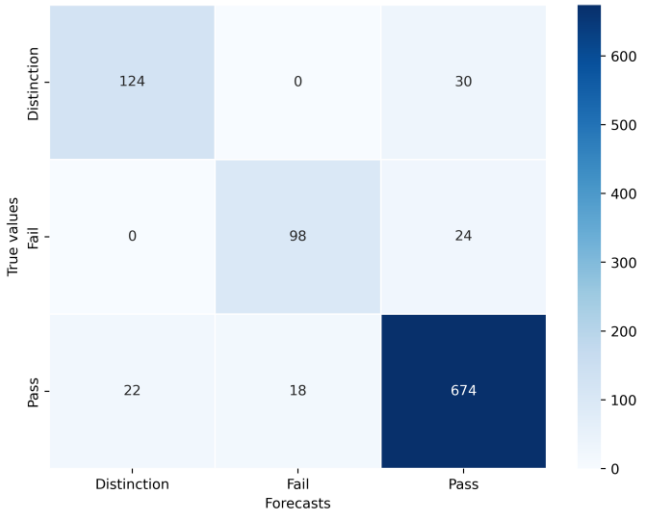

Fig. 3. Confusion matrix for random forest model.

The confusion matrix details:

- Distinction: The model correctly predicted 124 students as achieving Distinction (True Positives), while 22 students who should have been classified as Distinction were missed (False Negatives).
- Fail: The model correctly identified 98 students as Fail (True Positives), while 18 students who failed were incorrectly classified as passing (False Negatives).
- Pass: The model accurately predicted 674 students as Pass (True Positives), though 40 students were mistakenly classified as Pass (False Positives).

2) Classification report

As shown in Table 4, the Random Forest model achieves an overall accuracy of 91%. Its strongest performance is in predicting students who will pass, with a precision of 0.93 and a recall of 0.94, indicating a high number of correct predictions and comprehensive coverage of actual pass cases. Predictions for Distinction are also fairly accurate, although with a slightly lower recall, indicating that some high-performing students were not identified. The model performs well in identifying Failures, with balanced precision and recall values around 0.80. These results demonstrate the model's robustness in distinguishing between different levels of student performance, particularly excelling at recognizing students who will pass.

Table 4. Classification report for random forest model

	Precision	Recall	F1-score	Support
Distinction	0.85	0.81	0.83	154
Fail	0.84	0.80	0.82	122
Pass	0.93	0.94	0.93	714
Accuracy			0.91	990
Macro avg	0.87	0.85	0.86	990
Weighted avg	0.90	0.091	0.90	990

B. Logistic Regression

Despite its name, logistic regression is primarily used for binary classification but can also be extended to handle multiclass classification [48]. It estimates the probability that a given input belongs to a particular class, which can then be converted into a binary prediction by applying a threshold

[48]. Logistic regression is particularly useful in cases where explicit probabilities are required, and it offers the advantage of being easily updated with new data.

1) Confusion matrix

The performance of the logistic regression model, as shown in the confusion matrix (Fig. 4), is summarized as follows:

- Distinction: The model correctly predicted 118 students as achieving Distinction (True Positives), with zero false positives. However, 36 students who should have been classified as Distinction were incorrectly predicted as Pass (False Negatives).
- Fail: The model correctly identified 99 students as Fail (True Positives), with no instances wrongly classified as Fail. However, 23 students who should have been classified as Fail were misclassified as Pass (False Negatives).
- Pass: The model correctly predicted 658 students as Pass (True Positives), but 32 Distinction and 24 Fail students were misclassified as Pass, totaling 56 false positives.

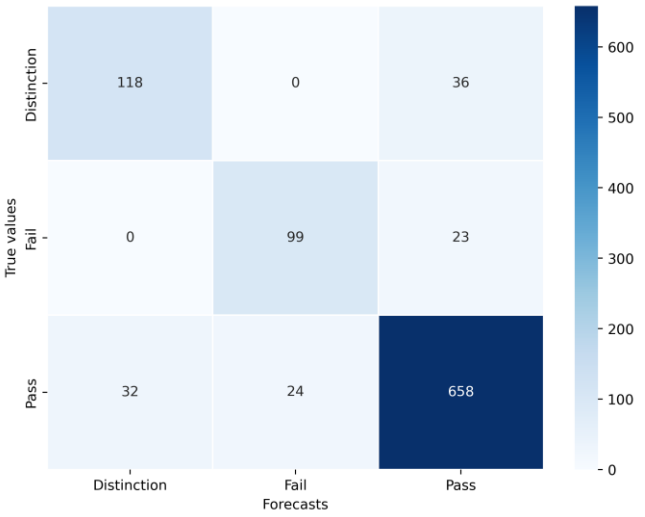

Fig. 4. Confusion matrix for logistic regression model.

The confusion matrix highlights the logistic regression model's strong ability to classify 'Pass' instances with high accuracy. However, it also reveals challenges in distinguishing 'Distinction' and 'Fail' from 'Pass,' as evidenced by the respective false negatives and false positives.

2) Classification report

As shown in Table 5, the logistic regression model achieves an overall accuracy of 88%, reflecting a high level of correct predictions across all classes. The model performs best in predicting 'Pass' instances, with a precision and recall of 92%, indicating both accuracy and a high rate of capturing true 'Pass' cases.

For 'Distinction' and 'Fail,' the model demonstrates reasonable performance, with precision and recall rates near 80%. The F1-scores align with these precisions and recall values, suggesting balanced classification ability. True positives are highest for 'Pass' predictions, showcasing the model's strong predictive power for this category. However, for 'Distinction' and 'Fail,' the model exhibits some limitations due to the higher number of false negatives and false positives.

Overall, these metrics underscore the model's reliability, particularly in identifying students who will pass, while also highlighting areas for improvement in distinguishing between 'Distinction' and 'Fail.'

Table 5. Classification report for Logistic Regression model

	Precision	Recall	F1-score	Support
Distinction	0.79	0.77	0.78	154
Fail	0.80	0.81	0.81	122
Pass	0.92	0.92	0.92	714
Accuracy			0.88	990
Macro avg	0.84	0.83	0.83	990
Weighted avg	0.88	0.88	0.88	990

C. SVM

SVM is a powerful and versatile supervised learning model used for classification, regression, and anomaly detection [49]. SVM is particularly effective in high-dimensional spaces and is characterized by its use of kernels, which allow it to handle complex data, as well as its focus on minimizing classification errors and maximizing the margin between data classes [52]. For classification tasks, SVM creates a hyperplane, or set of hyperplanes, in a high-dimensional space. The best hyperplane is one that maximizes the distance to the nearest data points from all classes, ensuring the best possible separation of the classes [53].

1) Confusion matrix

The performance of the SVM model, as shown in the confusion matrix (Fig. 5), is summarized as follows:

- Distinction (0): The model correctly predicted 115 instances of Distinction (True Positives). However, it misclassified 39 instances as Pass (False Negatives) and 29 instances as Fail (False Negatives). There were no instances of Fail or Pass misclassified as Distinction.
- Fail (1): The model accurately predicted 95 instances of Fail (True Positives), but 27 instances were misclassified as Pass, and 21 instances were incorrectly predicted as Distinction.
- Pass (2): The model showed high effectiveness in identifying Pass cases, with 664 correct predictions (True Positives). However, it misclassified 29 Distinction and 21 Fail instances as Pass (False Positives), totaling 48 incorrect classifications.

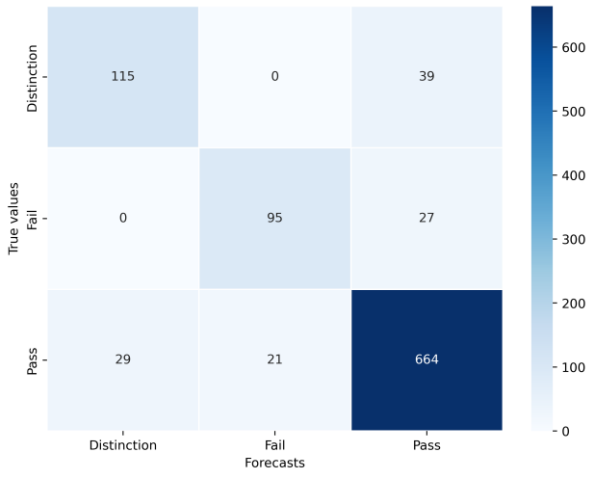

Fig. 5. Confusion matrix for SVM model.

The confusion matrix highlights the SVM model's strong predictive ability for 'Pass' outcomes, while also pointing out areas for improvement in distinguishing between 'Distinction' and 'Fail' classes.

2) Classification report

As shown in Table 6, the SVM model achieves an overall accuracy of 88%, indicating solid performance in classifying students into 'Distinction,' 'Fail,' and 'Pass' categories. The model is particularly effective in predicting 'Pass' outcomes, with a precision of 91% and a recall of 93%, reflecting its ability to correctly identify students who will pass while capturing the majority of actual Pass cases.

The model's performance for 'Distinction' and 'Fail' categories is slightly less accurate, with precision and recall values around 80%. The F1-scores are consistent with these precisions and recall figures, indicating balanced accuracy and coverage across all categories. The SVM model's strengths are most evident in the 'Pass' category, though there is room for improvement in accurately classifying 'Distinction' and 'Fail' cases.

Table 6. Classification report for SVM model

	Precision	Recall	F1-score	Support
Distinction	0.80	0.75	0.77	154
Fail	0.82	0.78	0.80	122
Pass	0.91	0.93	0.92	714
Accuracy			0.88	990
Macro avg	0.84	0.82	0.83	990
Weighted avg	0.88	0.88	0.88	990

D. LDA

LDA is a technique used for both classification and dimensionality reduction. It works by finding a linear combination of features that best separates two or more classes [50]. The goal is to project the data onto a lower-dimensional space that maximizes class separability while retaining as much relevant information as possible. LDA is also known for being resistant to overfitting, particularly when classes are well-separated and the number of features is relatively small.

1) Confusion matrix

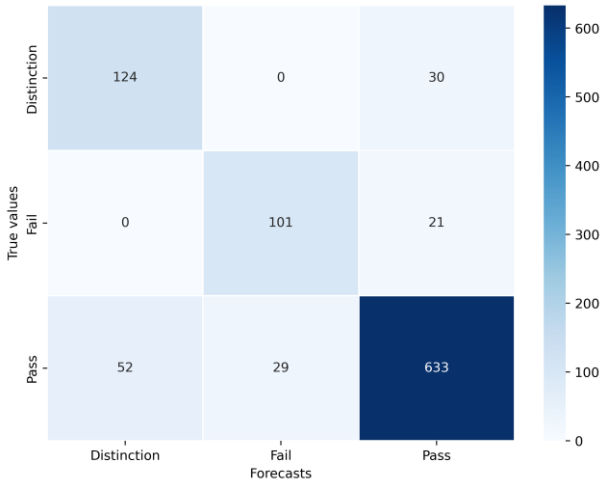

Fig. 6. Confusion matrix for LDA model.

The performance of the LDA model, as shown in the confusion matrix (Fig. 6), is summarized as follows:

- Distinction: The model successfully predicted 124 instances as Distinction (True Positives), but misclassified 30 Distinction students as Pass (False Negatives) and 52 as Fail (False Negatives). There were no instances of students from Fail or Pass categories being misclassified as Distinction.
- Fail: The model correctly predicted 101 Fail instances (True Positives). However, 21 students who should have been classified as Fail were misclassified as Pass, and 29 as Distinction (False Negatives).
- Pass: The model accurately predicted 633 instances as Pass (True Positives), but incorrectly classified 52 Distinction and 29 Fail students as Pass (False Positives), leading to a total of 81 misclassifications.

The confusion matrix for the LDA model reveals commendable accuracy in identifying 'Pass' instances, while also indicating areas for improvement in correctly classifying 'Distinction' and 'Fail' outcomes.

2) Classification report

As summarized in Table 7, the LDA model achieves an overall accuracy of 87%. It is highly precise in predicting students who will pass, with a precision of 93%, though there is a notable number of false positives in this category. The model shows good sensitivity in identifying students who will achieve Distinction and those who will fail, with recall rates of 81% and 83%, respectively.

However, precision for the 'Distinction' and 'Fail' categories is lower, at 70% for Distinction and 78% for Fail, indicating a cautious approach in predicting both higher and lower achievements. The F1-scores reflect a balanced accuracy between precision and recall across all categories. Overall, the model effectively predicts student performance, with a strong emphasis on correctly identifying students who will pass.

Table 7. Classification report for LDA model

	Precision	Recall	F1-score	Support
Distinction	0.70	0.81	0.75	154
Fail	0.78	0.83	0.80	122
Pass	0.93	0.89	0.91	714
Accuracy			0.87	990
Macro avg	0.80	0.84	0.82	990
Weighted avg	0.87	0.87	0.87	990

V. Discussion

The bar chart (Fig. 7) illustrates the accuracy of four classification models—Random Forest, Logistic Regression, SVM, and LDA—on both the training and test sets. The blue bars represent the accuracy on the test set, while the cyan bars indicate the accuracy on the training set.

Among the models, the Random Forest achieved the highest test set accuracy at 91%, demonstrating its superior performance in classifying student outcomes. This suggests that Random Forest is the most effective model for predicting student performance, particularly when accuracy is the primary consideration.

This research explored machine learning techniques for improving the predictability and understanding of learner performance within the e-learning ecosystem. Utilizing the

Open University Learning Analytics Dataset (OULAD), our aim was to develop models that provide accurate, personalized insights into student success, supporting data-driven decision-making in e-learning strategies.

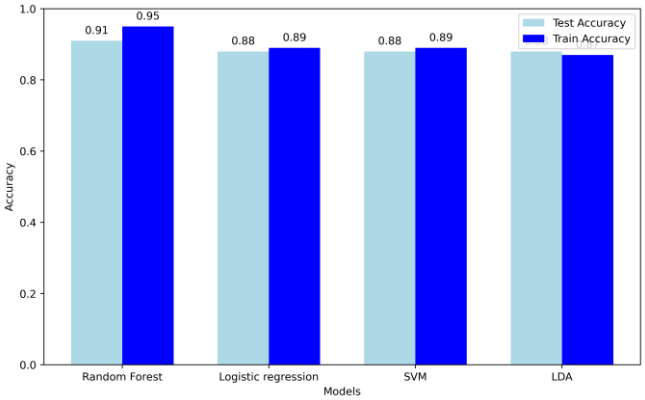

Fig. 7. Accuracy of classification models on training and test sets.

We conducted a comparative analysis of machine learning models to evaluate their effectiveness in predicting learner performance. This approach allowed us to thoroughly examine the strengths and limitations of each model while identifying key factors that influence learner outcomes. The objective was to create a predictive model that accurately classifies students into "Distinction," "Pass," and "Fail" categories using the rich feature set provided by OULAD.

Our methodology included rigorous data preprocessing, covering data loading, separation, assessment categorization, success thresholds, score merging and weighting, interaction analysis, and the exclusion of withdrawn students. This pre-treatment ensured that the dataset was optimized for modeling, accounting for student engagement, academic performance, and learning behaviors.

The selection of machine learning models—Random Forest, Logistic Regression, SVM, and LDA—was strategic, aimed at evaluating diverse algorithms with varying complexities. Dividing the dataset into 80% training and 20% testing ensured that the models were trained on representative data and evaluated under realistic conditions, thereby validating their generalizability and performance.

The results highlighted Random Forest's superior accuracy and its effectiveness in generalizing to unseen data. It was particularly adept at predicting "Pass" outcomes, indicating its potential as a reliable tool in the e-learning context for anticipating student needs. While Logistic Regression and SVM performed well in identifying "Pass" instances, they struggled with differentiating between "Distinction" and "Fail" cases. Similarly, LDA showed promise but required refinement for classifying higher success levels like "Distinction."

However, across all models, distinguishing between "Distinction" and "Fail" categories presented challenges. This limitation reflects the complexity of student performance metrics and the need for models capable of better navigating these subtleties.

The study's findings offer insights for future research and application. Incorporating machine learning models, such as Random Forest, into e-learning platforms could lead to more adaptive, responsive educational experiences. The classification challenges highlighted by this study suggest that further refinement of predictive models is necessary, potentially through the use of more nuanced features or advanced modeling techniques.

In conclusion, this research demonstrates the effectiveness of machine learning in e-learning ecosystems. While our models successfully predicted student performance, continuous improvements are needed to address the evolving dynamics of e-learning environments. Future research could focus on integrating additional datasets, applying more complex algorithms, and implementing real-time predictive systems that dynamically respond to student activities. This approach not only underscores the challenges and opportunities in predictive analytics for online learning but also contributes to the ongoing discussion on optimizing educational technologies for personalized learning and informed academic decision-making.

VI. Conclusion

This study has successfully demonstrated the potential of machine learning techniques to analyze and predict learner performance in e-learning ecosystems. By utilizing the Open University Learning Analytics dataset, we developed and compared various predictive models, with the Random Forest algorithm exhibiting superior accuracy and generalizability. These findings reinforce the hypothesis that machine learning can significantly enhance predictive analytics within e-learning systems, paving the way for more personalized and effective learning interventions.

However, the challenge of accurately distinguishing between different performance categories, particularly "Distinction" and "Fail," highlights the complexity of learning analysis and the need for more nuanced modeling approaches. Insights gained from this research underscore the importance of ongoing model refinement and the potential benefits of integrating multi-faceted data sources to capture the full range of factors influencing student outcomes. Further exploration of advanced machine learning techniques and a more granular approach to data analysis could improve the accuracy and effectiveness of predictive models.

Looking ahead, the integration of machine learning models into e-learning platforms has the potential to revolutionize education by enabling more adaptive, personalized, and responsive teaching and learning experiences. Future research should focus on operationalizing these models within real e-learning environments, assessing their impact on educational outcomes, and ensuring their ethical and fair application. The advancement of e-learning through machine learning is underway, and this study represents a critical step toward realizing the full potential of this dynamic field.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Fatima Ezzahraa EL HABTI designed and implemented the machine learning system, and wrote the manuscript. Mustafa Hiri assisted in the development of the machine learning system and contributed to the revision and review of

the manuscript. Mohamed Chrayah proposed the main idea for the topic and contributed to the conceptualization of the research. Abdelhamid Bouzidi contributed to the revision and editing of the manuscript. Noura Aknin also contributed to the revision and editing of the manuscript. All authors had approved the final version.

REFERENCES

[1] A. Collins and R. Halverson, *Rethinking Education in the Age of Technology: The Digital Revolution and Schooling in America*, Teachers College Press, 2018.

[2] C. Timbi-Sisalima, M. Sánchez-Gordón, J. R. Hilera-Gonzalez, and S. Otón-Tortosa, "Quality assurance in e-learning: a proposal from accessibility to sustainability," *Sustainability*, vol. 14, no. 5, 2022, pp. 1–27.

[3] W. Luna-Encalada, J. Guaiña-Yungan, and F. Molina-Granja, "E-learning ecosystem's to implement virtual computer labs," in *Proc. Learning Technology for Education Challenges: 9th International Workshop*, vol. 1428, 2021, pp. 77–89.

[4] M. S. Contreras-Ortiz, P. P. Marrugo, and J. C. R. Ribón, "E-learning ecosystems for people with Autism spectrum disorder: A systematic review," *IEEE Access*, 2023, vol. 11, pp. 49819–49832.

[5] L. T. Nguyen and K. Tuamsuk, "Digital learning ecosystem at educational institutions: A content analysis of scholarly discourse," *Cogent Education*, vol. 9, no. 1, 2022.

[6] F. E. Habti, M. Chrayah, A. Bouzidi, H. A. Ali, and N. Aknin, "Building a sustainable e-learning ecosystem: Strategies for long-term success," in *Proc. 2023 XIII International Conference on Virtual Campus (JICV)*, 2023, pp. 1–5.

[7] K. R. Premlatha, B. Dharani, and T. V. Geetha, "Dynamic learner profiling and automatic learner classification for adaptive e-learning environment," *Interactive Learning Environments*, vol. 24, no. 6, 2016, pp. 1054–1075.

[8] R. Panigrahi, P. R. Srivastava, and P. K. Panigrahi, "Effectiveness of e-learning: the mediating role of student engagement on perceived learning effectiveness," *Information Technology & People*, vol. 34, no. 7, 2021, pp. 1840–1862.

[9] K. Mangaroska, B. Vesin, V. Kostakos, P. Brusilovsky, and M. N. Giannakos, "Architecting analytics across multiple e-learning systems to enhance learning design," vol. 14, no. 2, 2021, pp. 173–188.

[10] S. N. Kew and Z. Tasir, "Developing a learning analytics intervention in e-learning to enhance students' learning performance: A case study," *Education and Information Technologies*, vol. 27, no. 5, 2022, pp. 7099–7134.

[11] M. Tan and P. Shao, "Prediction of student dropout in e-Learning program through the use of machine learning method," *International Journal of Emerging Technologies in Learning*, vol. 10, no. 1, 2015.

[12] A. Oke and F. A. P. Fernandes, "Innovations in teaching and learning: Exploring the perceptions of the education sector on the 4th industrial revolution (4IR)," *Journal of Open Innovation: Technology, Market, and Complexity*, vol. 6, no. 2, p. 3, 2020.

[13] Y. Wang, X. Liu, and Z. Zhang, "An overview of e-learning in China: History, challenges and opportunities," *Research in Comparative and International Education*, vol. 13, no. 1, 2018, pp. 195–210.

[14] W. Fasso, C. Knight, and B. A. Knight, "A learner-centered design framework for e-learning," *International Journal of Online Pedagogy and Course Design (IJOPCD)*, vol. 4, no. 4, 2014, pp. 44–59.

[15] S. Ahmad, A. S. M. M. Noor, A. A. Alwan, Y. Gulzar, W. Z. Khan, and F. A. Reegu, "eLearning Acceptance and Adoption Challenges in Higher Education," *Sustainability*, vol. 15, no. 7, 2023.

[16] M. Liu and C. Liu, "The adoption of e-learning beyond MOOCs for higher education," *International Journal of Accounting & Information Management*, 2020.

[17] P. Veluvali and J. Surisetti, "Learning management system for greater learner engagement in higher education—A review," *Higher Education for the Future*, vol. 9, no. 1, 2022, pp. 107–121.

[18] A. Al-Hunaiyyan, S. Al-Sharhan, and R. AlHajri, "Prospects and challenges of learning management systems in higher education," *(IJACSA) International Journal of Advanced Computer Science and Applications*, vol. 11, no. 12, 2020, pp. 73–79.

[19] V. M. Bradley, "Learning Management System (LMS) use with online instruction," *International Journal of Technology in Education*, vol. 4, no. 1, 2021, pp. 68–92.

[20] Q. N. Naveed, A. Muhammad, S. Sanober, M. R. N. Qureshi, and A. Shah, "A mixed method study for investigating critical success factors (CSFs) of e-learning in Saudi Arabian universities," *(IJACSA) International Journal of Advanced Computer Science and Applications*, vol. 8, no. 5, 2017, pp. 171–178.

[21] S. R. Sobral, "Two decades of research in e-learning: A deep bibliometric analysis," *International Journal of Information and Education Technology*, vol. 11, no. 9, 2021, pp. 398–404

[22] M. Shkoukani, "Explore the major characteristics of learning management systems and their impact on e-learning success," *(IJACSA) International Journal of Advanced Computer Science and Applications*, vol. 10, no. 1, 2019, pp. 296–301.

[23] S. K. Basak, M. Wotto, and P. Belanger, "E-learning, m-learning and d-learning: Conceptual definition and comparative analysis," *E-Learning and Digital Media*, vol. 15, no. 4, 2018, pp. 191–216.

[24] M. Liu and D. Yu, "Towards intelligent e-learning systems," *Education and Information Technologies*, vol. 28, no. 7, 2023, pp. 7845–7876.

[25] M. Meryem, R. Najat, A. Jaafar, and B. Salmane, "Impact of e-learning on the environment and the optimization of the use of natural resources," *E3S Web of Conferences. EDP Sciences*, vol. 412, 2023, p. 01098.

[26] Y. Gao, S. L. Wong, M. N. M. Khambari, and N. Noordin, "A bibliometric analysis of the scientific production of e-learning in higher education (1998-2020)," *International Journal of Information and Education Technology*, vol. 12, no. 5, 2022, pp. 390–399.

[27] C. C. Ogbonna, N. E. Ibezim, and C. A. Obi, "Synchronous versus asynchronous e-learning in teaching word processing: An experimental approach," *South African Journal of Education*, vol. 39, no. 2, 2019, pp. 1–15.

[28] F. B. Golley, "Historical origins of the ecosystem concept in biology," *The Ecosystem Concept in Anthropology*, Routledge, 2019, p. 33–49.

[29] G. Upreti, "Understanding ecosystem evolution and behavior," *Ecosociocentrism: The Earth First Paradigm for Sustainable Living*, Cham: Springer Nature Switzerland, 2023, pp. 65–89.

[30] M. G. Russell and N. V. Smorodinskaya, "Leveraging complexity for ecosystemic innovation," *Technological Forecasting and Social Change*, vol. 136, 2018, p. 114–131.

[31] E. Jeladze, K. Pata, and J. S. Quaicoe, "Factors determining digital learning ecosystem smartness in schools," *Interaction Design and Architecture(s) Journal*, vol. 35, 2017, pp. 32–55.

[32] L. T. Nguyen, I. Kanjug, G. Lowatcharin, T. Manakul, K. Poonpon, W. Sarakorn, A. Somabut, N. Srisawasdi, S. Traiyarach, and K. Tuamsuk, "Digital learning ecosystem for classroom teaching in Thailand high schools," *SAGE Open*, vol. 13, no. 1, 2023.

[33] F. E. E. Habti, M. Chrayah, and A. Bouzidi, "Towards e-learning ecosystem model based on cloud computing," in *Proc. 2020 X International Conference on Virtual Campus (JICV)*, IEEE, 2020, pp. 1–4.

[34] F. E. E. Habti, M. Chrayah, A. Bouzidi, and H. A. Ali, "Blended learning platform model," in *Proc. 2022 XII International Conference on Virtual Campus (JICV)*, 2022, pp. 1–4.

[35] V. Heyde and A. Siebrits, "The ecosystem of e-learning model for higher education," *South African Journal of Science*, vol. 115, no. 5–6, 2019, pp. 1–6.

[36] F. Qiu, G. Zhang, X. Sheng, L. Jiang, L. Zhu, Q. Xiang, B. Jiang, and P. K. Chen, "Predicting students' performance in e-learning using learning process and behaviour data," Scientific Reports, vol. 12, no. 1, 2022, p. 453.

[37] S. Batool, J. Rashid, M. W. Nisar, J. Kim, H. Y. Kwon, and A. Hussain, "Educational data mining to predict students' academic performance: A survey study," *Education and Information Technologies*, vol. 28, no. 1, 2023, ïpp. 905–971.

[38] K. S. Na and Z. Tasir, "Identifying at-risk students in online learning by analysing learning behaviour: A systematic review," in *Proc. 2017 IEEE Conference on Big Data and Analytics (ICBDA)*, IEEE, 2017, pp. 118–123.

[39] O. Aissaoui, Y. Alami, L. Oughdir, and Y. Allioui, "A hybrid machine learning approach to predict learning styles in adaptive e-learning system," in *Proc. International Conference on Advanced Intelligent Systems for Sustainable Development*, 2018, pp. 772–786.

[40] I. Gligorea, M. Cioca, R. Oancea, A. T. Gorski, H. Gorski, and P. Tudorache, "Adaptive learning using artificial intelligence in e-learning: A literature review," *Education Sciences*, vol. 13, no. 12, 2023, pp. 1–27.

[41] K. Colchester, H. Hagras, D. Alghazzawi, and G. Aldabbagh, "A survey of artificial intelligence techniques employed for adaptive educational systems within e-learning platforms," *Journal of Artificial Intelligence and Soft Computing Research*, vol. 7, no. 1, 2017, pp. 47–64.

[42] Z. R. Alhalalmeh, Y. M. Fouda, M. A. Rushdi, and M. El-Mikkawy, "Automating assessment and providing personalized feedback in e-learning: The power of template matching," *Sustainability*, vol. 15, no. 19, 2023, pp. 1–22.
[43] D. Tzimas and S. Demetriadis, "Ethical issues in learning analytics: A review of the field," *Educational Technology Research and Development*, vol. 69, 2021, pp. 1101–1133.
[44] J. Kuzilek, M. Hlosta, and Z. Zdrahal, "Open university learning analytics dataset," *Scientific Data*, vol. 4, no. 1, 2017, pp. 1–8.
[45] S. Huber, H. Wiemer, D. Schneider, and S. Ihlenfeldt, "DMME: Data mining methodology for engineering applications–a holistic extension to the CRISP-DM model," *Procedia Cirp*, vol. 79, 2019, pp. 403–408.
[46] F. Martinez-Plumed, L. Contreras-Ochando, C. Ferri, J. Hernández-Orallo, M. Kull, N. Lachiche, M. Jose Ramirez-Quintana, and P. Flach, "CRISP-DM twenty years later: From data mining processes to data science trajectories," *IEEE Transactions on Knowledge and Data Engineering*, vol. 33, no. 8, 2019, pp. 3048–3061.
[47] M. Hiri, M. Chrayah, N. Ourdani, and N. Aknin, "Machine learning techniques for diabetes classification: A comparative study," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 9, 2023, pp. 785–790.
[48] A. H. Osman and H. M. Aljahdali, "Feature weight optimization mechanism for email spam detection based on two-step clustering algorithm and logistic regression method," *International Journal of Advanced Computer Science and Applications*, vol. 8, no. 10, 2017, pp. 420–429.
[49] M. Zbakh, N. Aknin, M. Chrayah, and A. Bouzidi, "Enhancing employee performance management: A data-driven decision support model using machine learning algorithms," *International Journal of Advanced Computer Science and Applications*, vol. 15, no. 3, 2024.
[50] S. A. P. Raj and Vidyaathulasiraman, "Prioritization of e-learners activities using principal component analysis method," *International Journal of Information Technology*, vol. 13, no. 6, 2021, pp. 2439–245.
[51] N. Chandrasekhar and S. Peddakrishna, "Enhancing heart disease prediction accuracy through machine learning techniques and optimization," *Processes*, vol. 11, no. 4, 2023, p. 1210.
[52] A.U. Khasanah and H. Harwati, "Educational data mining techniques approach to predict student's performance," *International Journal of Information and Education Technology*, vol. 9, no. 2, 2019, pp. 115–118.
[53] A .S. Paramita and L. M. Tjahjono, "Implementing machine learning techniques for predicting student performance in an e-learning environment," *International Journal of Informatics and Information Systems*, vol. 4, no. 2, 2021, pp. 149–156.