

Gradient boosting algorithm for predicting student success

Brahim Jabir¹, Soukaina Merzouk¹, Radoine Hamzaoui², Nouredine Falih²

¹ESIM Research Team, Polydisciplinary Faculty of Sidi Bennour, Chouaib Doukkali University, Morocco

²Laboratory of Innovation in Mathematics and Applications and Information Technologies (LIMATI), Beni Mellal, Morocco

Article Info

Article history:

Received Sep 6, 2024

Revised Mar 25, 2025

Accepted May 23, 2025

Keywords:

Distance learning

E-learning

Machine learning

Performance prediction

XGBoost algorithm

ABSTRACT

The idea of using machine learning resolution techniques to predict student performance on an online learning platform such as Moodle has attracted considerable interest. Machine learning algorithms are capable of correctly interpreting the content and thus predicting the performance of our students. Algorithms namely gradient boosting machines (GBM) and eXtreme gradient boosting (XGBoost) are highly recommended by most researchers due to their high accuracy and smooth boosting time. This research was conducted to analyze the effectiveness of the XGBoost algorithm on Moodle platform to predict student performance by analyzing their online activities, practicing various types of online activities. The proposed algorithm was applied for the prediction of academic performance based on this data received from Moodle. The results demonstrate a strong correlation between many activities like the number of hours spent online and the achievement of academic goals, with a remarkable prediction rate of 0.949.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Brahim Jabir

ESIM Research Team, Polydisciplinary Faculty of Sidi Bennour, Chouaib Doukkali University

El Jadida, Morocco

Email: ibra.jabir@gmail.com

1. INTRODUCTION

Education has become a key sector of digital transformation initiatives, especially in regions that are striving to modernize their education systems to meet contemporary demands, especially in developing countries. In Morocco, we have noticed this evolution attempt which is embodied by the National Plan for Accelerating the Transformation of the Ecosystem launched in 2023, it focuses on digital learning platforms and the development of relevant technological skills [1]. This mission has led to a rapid expansion of online learning on several modules, with 70% of courses now being delivered online. These platforms generate large amounts of data, especially data on student interactions, offering unique opportunities to evaluate the educational processes of institutions and improve learning outcomes. Ethical challenges also pose major obstacles to research and this requires a rigorous approach to avoid problems of confidentiality [2]. Researchers must obtain explicit permission from administrators to access and use student data. In addition, handling this data requires meticulous attention to confidentiality and privacy protocols because the personally identifiable information must never be disclosed beyond the research team and must be used exclusively for educational purposes and remain under the supervision of the students' teachers who have the professional responsibility and understanding of the context to ensure appropriate use.

Machine learning (ML) has moved beyond its original niche in computer science to become a go-to tool on the toolbox of the practitioner in a range of professions; one of them is education if we focus here, where it helps data analysis, personalized learning, and predictive modeling [3]. ML as a subfield of artificial intelligence, is concerned with the design of systems that learn from data and can adapt their performance in the face of changing conditions rather than being explicitly programmed. This ability is certainly relevant in

educational contexts where a significant amount of underutilized data is produced through student interactions in e-learning environments. Various machine learning algorithms, including decision tree (DT), random forests (RFs), neural networks (NNs), support vector machines (SVMs), and K-means clustering, take part in mining this data and identifying patterns to enable stakeholders to take effective steps toward improving educational mechanisms [4], [5]. Each algorithm that we have mentioned above offers distinct advantages. For example, DT are said to excel in transparent decision making, while RF mitigates overfitting through ensemble methods. Neural networks are well suited for complex pattern recognition, while SVMs effectively classify learning behaviors to identify at-risk students. K-means clustering helps group students with similar characteristics, which allows for personalized interventions. So, in the field of online learning, research has shown that machine learning through its algorithms is a tool used to improve the learning experience by personalizing content, predicting academic performance, and identifying at-risk students and we analyzed many aspects that are beneficial for our future of education. For example, Alzubi *et al.* [6] demonstrated the effectiveness of neural networks in predicting student performance by analyzing demographic data, academic records, and online activities. In addition, Oller *et al.* [7] used decision trees to highlight important factors such as attendance, forum participation, and assignment completion, providing interpretable results for targeted interventions. However, as we can all see, despite these advances, there are still challenges that we can notice in our teaching life. Issues such as data scarcity, scalability, and cold start issues for new users hinder the full potential of ML applications in e-learning [8]. The use of advanced algorithms such as XGBoost in educational contexts remains underexplored especially in the Moroccan context with data from Moroccan higher education platforms. That is why we said that it is essential to fill this gap to improve the accuracy and applicability of predictive models, thus enabling more effective e-learning strategies. This paper responds to these challenges by proposing a theoretical model of student engagement capable of predicting academic performance based on the analysis of students' digital activities. Two major research questions orient this work: i) to what extent can success in the classroom be predicted from students' digital interaction practices? and ii) Which online activities are most associated with academic success?

To answer these questions, this study analyzes quantitative data from Moodle interactions of 290 students, including time spent, type of interaction (e.g., helpful, collaborative, creative), response patterns, and participation quality. The research draws from three theoretical frameworks: first, digital learning analytics includes the measurement and evaluation of learner data in online environments. Second is predictive modeling in education, oriented toward the application of machine learning to predict academic success. The last framework is the Student Engagement Theory, about the connection between educational engagement and learning outcomes. This study not only recommends the application of XGBoost in academic prediction but also fills the gaps identified in the literature by exploring different forms of digital interaction. Following a quantitative research design, the current study incorporates data mining from Moodle, multiple machine learning algorithms, statistical correlation analyses, and accuracy comparisons. The rest of the paper is organized as follows: section 2 describes the data collection, preprocessing, and analytical approaches; section 3 reports the results, including the performance metrics and correlation findings; section 4 discusses the results and their implications, and section 5 concludes with key insights on future directions.

2. METHODS AND TOOLS

2.1. Data collection

The data upon which this study is based is provided by the official Moodle system that our university has established specifically for the 'university work methodology' course and for the 'digital culture' course for which this study is relevant since they are transverse modules within the framework of the reform plan mentioned in the introduction to this study. The data extraction process from Moodle was performed using the official Moodle Web Services API. Access was secured through an authorized API token, ensuring compliance with institutional data protection policies. For this study, we focused on a course containing 290 students, extracting all available information about these students' interactions with the platform. Therefore, this data set has many fields such as first and last name, total time taken to complete the course, total number of messages or publication shared by a particular student, reaction to these publications and the final outcome. These tools enable us to monitor and save students' activities, sorting out their responses to publications based on several specified parameters. The main variables considered for this study are, student information (first and last names), engagement metrics (total time spent on the course, number of messages/posts), reactions (categorized into helpful, nice, collaborative, confused, creative, bad, and amazing), and final outcome (success or failure in the course).

2.2. Data preprocessing

To prepare the data for analysis, the following pre-processing steps were performed: Data cleaning, which checked for the absence of missing data, or replicating data, and checked for wrong entries in the values. Blank cells were replaced with 0, meaning that students who never visited the platform were assigned zero access. Categorization of reactions, these collected reactions were grouped as helpful, nice, collaborative, confused, creative, poor, and amazing. The material was coded based on a coding schedule that was decided prior to coding, and the reliability between two coders was compared. The data splitting was done in an attempt to create cross-validation so that independent validation of the results can be done, as recommended by [9]. This means that $\text{test_size}=0.2$ was assigned to the test data and 80% for training data. The dataset was comprised of 5000 records of 290 students, which would allow of sufficient sample of records required for accurate model testing.

2.3. The proposed machine learning model

XGBoost or eXtreme gradient boosting is powerful and frequently used for classification and/or regression. It operates through constructing a sequence of weak models commonly known as decision trees that help improve on the previous model's mistakes in the face of the actual model. XGBoost has been chosen because of its performance and versatility in dealing with different kinds of datasets, it was found to be very efficient with most problems. XGBoost's major is written in C++ language to get a better training of gradient boosting [10]. As shown in Figure 1, XGBoost develops from existing knowledge on DTs and RFs, as well as introduces key improvements such as gradient boosting, regularization, parallelization, tree pruning, and customization [11].

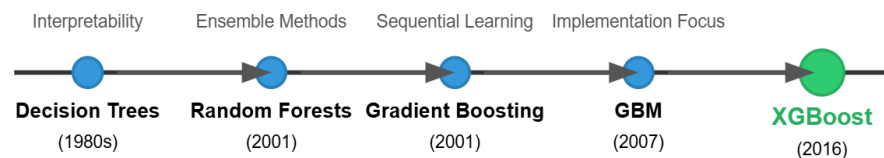


Figure 1. Evolution of XGBoost algorithm

In order to show why the XGBoost model was chosen out of other models mentioned in the literature, a comparison was made. Based on the comparison in Table 1, XGBoost is adopted because it is adaptive to missing value and quick in training and accurate in both regression and classification tasks. This gives XGBoost an edge over other models due to several reasons highlighted by literature as well as its complete features, which include the ability to predict the student success in the online course. However, the efficiency of the proposed model will be proved by numerical comparison in the results and discussion section. This will afford a definite and measurable method of indicating that indeed XGBoost is a better and suitable model for this particular use.

Table 1. Comparison of XGboost with other algorithms

| Model | Advantages | Disadvantages | References |
|---------|---|---|------------|
| DTs | Easy to interpret and visualize | Prone to overfitting | [12], [13] |
| RF | Reduces overfitting by averaging multiple trees | Computationally intensive | [14], [15] |
| SVM | Effective in high-dimensional spaces | Requires careful tuning of parameters | [16], [17] |
| NNs | Capable of capturing complex patterns | Require large datasets and extensive tuning | [18], [19] |
| XGBoost | Handles missing values well, efficient and scalable | Requires parameter tuning, complex to implement | [20], [21] |

2.3.1. Features

This XGBClassifier is a class of the library XGBoost that contains the algorithm of classification based on the gradient boosting. XGBClassifier, built particularly for classification tasks, has several benefits at its disposal. It handles missing data effectively without extra pre-processing steps and can handle parallel computing [22] this optimizes learning large datasets [23]. It emerges that constructing multiple decision trees sequentially leads to a better predictive performance through gradient boosting [24]. Moreover, parameter tuning in XGBClassifier is integrated with intelligent prevention of overfitting by the penalty on complex trees [25], and the identification of the useful features are also separate [26]. In addition, it can support the distributed computing and can be used on Apache Hadoop, Apache Spark, and Dask and others

[27]. The model's hyperparameters were tuned using grid search to optimize performance [28]. Table 2 defines the configuration of the hyperparameter grid for tuning our XGBoost model, including their descriptions and possible values. Hyperparameters are crucial in optimizing the performance of machine learning models, and this grid includes a variety of values for each hyperparameter to explore the best combination.

Table 2. Hyperparameters

| Hyperparameter | Description | Values |
|------------------|--|----------------------------------|
| base_score | Initial prediction score for all instances | [0.25, 0.5, 0.75, 0.1] |
| n_estimators | Number of boosting rounds | [100, 200, 500, 900, 1500, 2000] |
| max_depth | Maximum depth of each tree | [2, 3, 5, 10, 12, 15] |
| booster | Type of booster to use | 'gbtree', 'gblinear' |
| learning_rate | Step size shrinkage used to prevent overfitting | [0.01, 0.05, 0.1, 0.2, 0.5] |
| min_child_weight | Minimum sum of instance weight (hessian) needed in a child | [1, 2, 3, 4, 5] |

Combining these hyperparameters into a single configuration, named `hyperparameter_grid` can be useful in hyperparameter tuning procedures as grid search in our case, is a solution to find the best fitting configuration of the model XGBoost. It also means that in this approach the model is being trained to deliver its best by covering all the aspects of the problem. The pseudocode in Algorithm 1 provides the acceptable values for each of the hyperparameters (as mentioned in Table 2) of the XGBClassifier and establishes a duct for the hyperparameter grid configurations. The model attained a high accuracy on the test set, along with precision, recall, and F1 measure to determine its efficiency. For feature selection we utilized XGBoost's built-in feature importance mechanism based on gain, which measures each feature's contribution to model improvement when used in trees. These features were ranked according to their importance scores, and we retained those that cumulatively contributed to 95% of the total importance. This technic allowed us to focus on the most influential factors affecting student performance while maintaining model interpretability.

Algorithm 1. Hyperparameter_grid

```

DEFINE hyperparameter_grid AS DICTIONARY
  SET hyperparameter_grid['n_estimators'] TO n_estimators_values
  SET hyperparameter_grid['max_depth'] TO max_depth_values
  SET hyperparameter_grid['learning_rate'] TO learning_rate_values
  SET hyperparameter_grid['min_child_weight'] TO min_child_weight_values
  SET hyperparameter_grid['booster'] TO booster_values
  SET hyperparameter_grid['base_score'] TO base_score_values

```

3. RESULTS

3.1. Research question 1: from online interactions to a predicted success rate

Our first research question was to determine whether student success could be predicted solely from their online interactions, or whether other factors directly influence success were needed. For that, the analysis involved extracting and processing interaction data from our university's Moodle platform, including time spent, message frequency, and different types of reactions, to establish a predictive framework. The analysis gives us these key findings:

3.1.1. The model performance

To investigate various behaviors and the correlation to student's achievement, interactions divided into types Helpful, Well-written, collaborative, confused, creative, Inappropriate and Remarkable were examined. This was the view that we sought to understand if type of interactions influenced the success rate of the human resource functions. This data was used for training our model and the analysis of these interactions for predicting the levels of success of the student groups. Given the high performance of the XGBoost algorithm in classification tasks, we then used it to explore more into student success. As presented in Figure 2, the model performance was assessed based on precision, recall, F1 and overall accuracy. The accuracy of positive prediction rates was 94% for class 0 (students not succeeding), while the success rating was 93% for class 1 (students succeeding). The recall was 0.91 for class 0 and was 0.96 for class 1, which shows that the model is good at identifying most of the positive samples. The F1-scores which are measures of model accuracy in terms of both precision and recall were 0.91 for class 0 and 0.96 for class 1. Lastly, the accuracy of the model was 0.949 suggesting little omission and commission errors were made when using the data set.

| | precision | recall | f1-score |
|--------------------|-----------|--------|----------|
| 0 | 0.95 | 0.83 | 0.91 |
| 1 | 0.93 | 0.96 | 0.96 |
| accuracy | | | 0.95 |
| macro avg | 0.97 | 0.92 | 0.94 |
| weighted avg | 0.95 | 0.95 | 0.95 |
| 0.9491525423728814 | | | |

Figure 2. Performance of the proposed model

Figure 3 presents the receiver operating characteristic (ROC) curve illustrating the performance of our proposed model, and then the curve visualizes the relationship between the true positive rate and the false positive rate at different classification thresholds. The area under the curve (AUC) is 0.93 which shows that the model has a discriminatory ability that is significantly superior to random classification (represented by the dashed diagonal line). This high AUC value confirms the effectiveness of our algorithmic approach in identifying correlations between student engagement patterns and their academic performance. Our analysis shows that the model can effectively distinguish successful students from those at risk based solely on their online learning activity, which supports our research hypothesis.

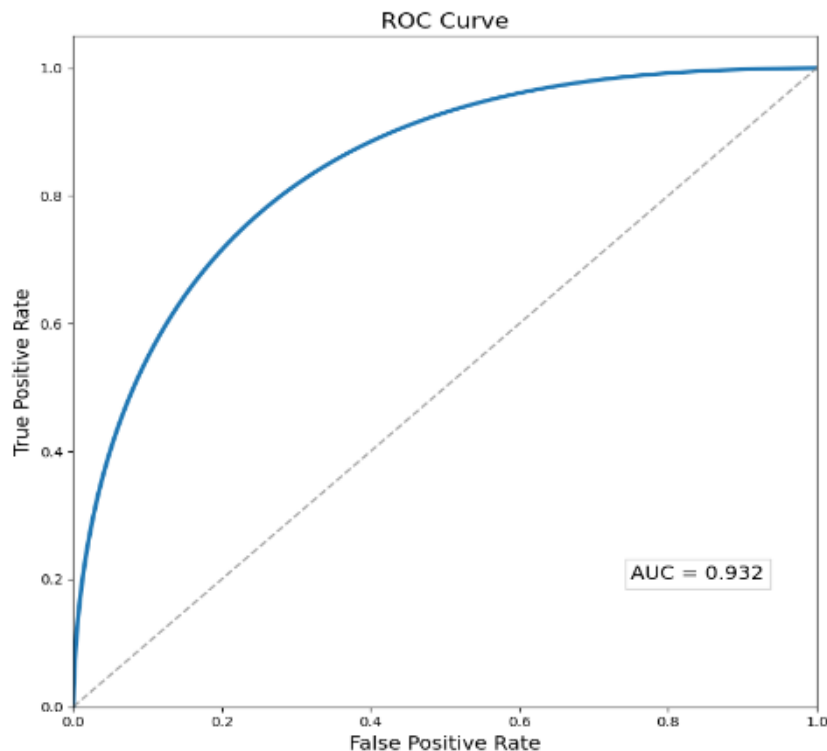


Figure 3. ROC curve for the proposed model

3.1.2. The confusion matrix

The confusion matrix [29] presented in Figure 4 shows the result of a prediction on 290 cases in total. The matrix can be interpreted as follows: True positives $TP = 228$ mean that the model has correctly classified 228 instances as positive class. For this case, specificity is $1 - (FP / (FP + TP))$, therefore, false negatives (FN) are 0, and it also shows that no example from the positive class were classified as negatives. Specifically, number 3 on FP is the number of instances that were falsely classified as the positive case. False negatives (FN) are 15 hence there are 15 examples the model incorrectly classified as being from the positive class. Thus, the model returns generally high true positive and true negative values and low values of false positive ones.

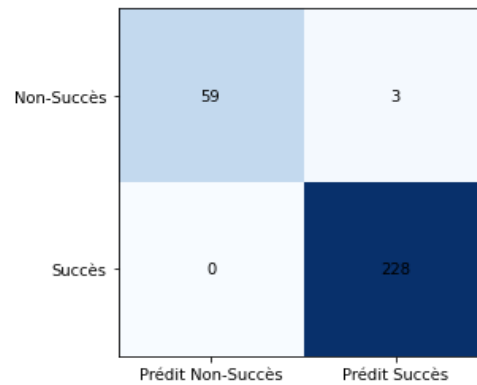


Figure 4. The confusion matrix

3.2. Research question 2: correlation between the internet activities and achievement results

The second research question is based on identifying distinct associations between different forms of platform use and academic performance, for answering this question; we conducted several correlation analyses between specific online behaviors and exam results, using quantitative and qualitative methods to establish reliable trends. Our analytical approach incorporated Pearson correlation coefficients to measure the strength and direction of relationships between variables complemented by statistical significance tests to validate the reliability of these associations. The analysis revealed several significant patterns:

3.2.1. Time performance correlation

In order to answer this research question regarding the possible predictability of the students' success based on their interaction, we analyzed the correlation between the online platform interaction of the students, which is measured in hours spent per day spent on the platform and the grade they received at the end of the course. This implicit association was tested using the Pearson correlation coefficient as an index of this linear relationship. Positive coefficients equal to 1 or close to 1 are stronger while negative ones equal to -1 or close to this are stronger; coefficients equal to 0 means that there is no relationship at all in terms of the linearity of the data [30]. Table 3 offers analysis of the Pearson correlation coefficient $r=0.807$, with a statistically significant difference at $p=0.001$. These results suggest that there exists a high positive relationship whereby more interaction per hour on the online platform leads to better final grade. This is an implication that the frequency of practicing the activities associated with the internet brings out better performance among students.

Table 3. The correlation results between the total time spent and the result

| Total_duration_spent | Pearson Correlation | Total_duration_spent | Result |
|----------------------|---------------------|----------------------|--------|
| | | 1 | ,891** |
| | Sig. (2-tailed) | | <,001 |

** Correlation is significant at the 0.01 level (2-tailed).

3.2.2. Interaction type analysis

Explorations of student engagement with learning management systems (LMSs) extend beyond mere contact duration and encompass quantitative features that suggest great roles. This study therefore identified quantifiable relationships between different types of interactions and academic outcomes, worthy of understanding the effects of social interactions on the Moodle platform. Amiable interactions especially useful ($r=0.724$), synergy ($r=0.683$) and innovative ($r=0.651$) interactions show relatively high correlation with success. These behaviors say that students are actively working together to achieve common learning goals. The quality of contributions is also important with well-structured messages (correlation of 0.612) and outstanding contributions (correlation of 0.589) having a positive impact on academic achievement. On the other hand, confusing messages are negatively associated with performance ($r=-0.342$) and so are the messages that are perceived as being inappropriate ($r=-0.289$) suggesting orientation problems or lack of active involvement. These outcomes prove the claim that variety and the specifics of the interactions are most influential for achieving academic performance. Such observations suggest the development of rich and diverse engagements along with appropriate instruction of need academic intervention, and assistance for

struggling students. If this data could be incorporated into early warning systems, one would be able to come up with suggested interventions for every student making the student engagement and their performance better. The following Table 4 shows the correlation between different sorts of interactions and academic success, obviously indicating the necessity of positive qualitative interaction for the prediction of performance.

Table 4. Correlation between interaction types and success

| Interaction type | Correlation (r) | p-value | Effect size |
|---------------------|-----------------|---------|-------------|
| Helpful posts | 0.724 | < 0.001 | Large |
| Collaborative posts | 0.683 | < 0.001 | Medium |
| Creative posts | 0.651 | < 0.001 | Medium |
| Well-Written posts | 0.612 | < 0.001 | Medium |
| Remarkable posts | 0.589 | < 0.001 | Medium |
| Confused posts | -0.342 | < 0.001 | Small |
| Inappropriate posts | -0.289 | < 0.001 | Small |

Effect size interpretation: Small: $r < 0.3$, Medium: $0.3 \leq r < 0.7$, Large: $r \geq 0.7$

3.3. Model comparison

Some of these findings accord with some previous research findings. In Ashima's study, a system based on RF was used with a prediction accuracy of 96% while a model we have developed here based on XGBoost has slightly lesser but comparable accuracy to that work. In the research where the authors enhanced the classification accuracy from 77% using ensemble techniques with basic data mining techniques [31], it seems higher than they are. For the study done by Theophilus team [32], the accuracy achieved was 90% by applying SVM in order to predict learners' engagement levels in online learning. Comparing this with our study, it is clear that our model yields infinitesimal higher accuracy. In Padmalaya's study [33] the accuracy for likelihood ratio obtained from a NNs while predicting the students' academic performance ranged between 87.14% to 90.74% and this is in comparison to our model, which seems to yield similar accuracy. In their experiment to predict physical education student academic performance, [34] obtained an overall average accuracy of 85.74% with the DT classifier utilized tenfold cross-validation. However, the results derived from our proposed approach produce substantially higher accuracy than those from theirs. When we compare our results with literature review, we observe that our model is highly proficient and can accurately predict whether a student will succeed or fail based on their engagement. Therefore, this study reveals that communications students engage in on the e-learning platform could serve as predictive markers of their performance. From this finding, it becomes clear that teachers and online course developers should ensure that students engage in the platform to enhance their success. Table 1 indicates the advantages of the used algorithm over other algorithms for choosing the required algorithm. However, we need to work out the original comparative results of each performance metric after the training of our model, which is illustrated in Table 5 illustrating the performance of the proposed model. This comparison reveals that XGBoost model has higher accuracy than most of the traditional machine learning models. It is also evident from this all-round comparison that this XGBoost model is quite effective in predicting student success, which makes it very useful for e-Learning activities.

Table 5. Performance comparison of algorithms

| Model | Accuracy [31]–[34] |
|---------------------------------|--------------------|
| DTs | 85,74 |
| RF | 96% |
| SVM | 90% |
| NNs | 90,74% |
| Traditional data mining methods | 77% |
| Proposed XGBoost | 95% |

4. DISCUSSION

4.1. Performance and theoretical contributions

The Accuracy achieved by XGBoost model was 0.949 and AUC=0,93, which proves that what we built can be used to predict the students at risk based on their online behaviors. This result shows once again that interaction patterns are highly predictive of future academic success, as reflected by the digital footprints left behind. Within the given dataset, the number of false positives (3) and more importantly, the false negatives are non-existent which points to the ability of distinguishing at-risk students. Second, this capacity

can be used at an early intervention level for students in at-risk situations to help them before academic difficulties reach this level of severity and before the final exam. The observed significance which is quite high and positive ($r=0.807$) actually underlines the importance of consistent use by relating it with the academic performance of the students. Furthermore, interaction types such as helpful, collaboration, and creation useful, show that the degree and intensity of interaction are important as well. As shown in Table 6, our research findings contribute to two well-established theories.

Using several similarity measures show that this study attains a predictability rate, which pioneers many studies in the past. The relationship of engagement with success complements similar literature, while a successful case in our university in Morocco, extends applicability of the model to cross-national e-learning initiatives. Secondly, the use of different kinds of interactions reveals new findings, underlining the quantity and the kind of interactions in online environments.

Table 6. Implications of the research

| Contributions | Implications |
|-------------------------|--|
| Pedagogical Design | Based on the research outcomes, course designers should ensure that aspects that may enhance participation and interactions are formalized for frequent and purposeful use, and some of these include: The different engagement types as should include collaborative and creative based type of engagements while the assessment types should reflect engagement types. The use of interactive characteristics together with the distribution of such elements throughout the timeline of the course can also increase students' involvement. |
| Student Support Systems | Based on the model's output, decision making for early warning system for student vulnerable can be made. Specific approaches can be launched, targeted to certain engagement profiles, meanwhile, feedback systems can foster the drift towards productive behaviors, and they are useful in the allocation of institutional resources for students based on anticipated needs of the student. |
| Institutional Policy | In light of this, one perhaps would wish to encourage or recommend features during the development of the platform that are associated with success, at a policy level these are characteristics of educational institutions. Developing issues for specific populations may require the enhancement of digital engagement for teaching faculty and development programs to focus on the utilization of online engagement metrics in assessment policies. |

4.2. Limitations and recommendations

This study has methodological limitations. The first limitation related to the data; the data were collected from a single Moroccan university and relating to specific courses, the results might not apply to other institutions, educational systems or cultural contexts where the behaviors and learning styles of students might differ. In addition, our approach focuses on the quantitative aspect favoring measurable engagement indicators like: time spent online, number of publications and interactions, while neglecting somewhat crucial qualitative factors such as students' motivation and satisfaction, their learning preferences and their emotional states, which surely influence their academic results.

As recommendations to policymakers, educators, administrators, and platform developers, the following insights aim to enhance the design and implementation of effective online learning environments. For educators: It is suggested that educators would be influential in mapping the temporal dynamics of the different types of interactions at work, as well as in designing activities that support varying levels of interactions. The more often feedback of the quality and quantity of engagement is provided to students, the more effective they can be in inspiring students and in getting the results that are desired. For administrators: There is the need for incorporating high-quality learning analytics systems and social media skills in faculty members. Any policy should include meaningful communication and balance institutional resources with the likely needs of the students. For platform developers: It is recommended that engagement metrics should be well incorporated, and there should be an integration of auto-responsive features. The creation of personalized intervention management tools and the incorporation of technology which will promote different usages are advised.

5. CONCLUSION

The application of AI for predicting the students' success in the framework of e-learning platform considering their activity on the learning environment is effective. The results obtained indicate that participation of students in online activities improves their performance hence a need for students to embrace the online activities. It also shows how applicable the XGBoost algorithm is to this specific task with an average accuracy of 0.949 prediction. This study is going to help in improving strategies used by online teachers and support offered to learners. Regarding future research avenues, several key areas emerge. The first is the exploration of deep learning architecture such as neural networks to provide valuable comparative information with our XGBoost model. Second, expanding our database to a wider range of students from

multiple disciplines, courses, and academic years would significantly improve the generalizability and robustness of the model. Finally, optimizing the model's performance through rigorous hyperparameter tuning and the integration of additional predictor variables would refine the accuracy of our system.

FUNDING INFORMATION

Authors state no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

| Name of Author | C | M | So | Va | Fo | I | R | D | O | E | Vi | Su | P | Fu |
|------------------|---|---|----|----|----|---|---|---|---|---|----|----|---|----|
| Brahim Jabir | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | |
| Soukaina Merzouk | | ✓ | | | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | | |
| Radoine Hamzaoui | ✓ | | ✓ | ✓ | | | ✓ | | | | | | ✓ | ✓ |
| Noureddine Falih | | ✓ | | ✓ | | | | | | ✓ | ✓ | | | |

C : Conceptualization

M : Methodology

So : Software

Va : Validation

Fo : Formal analysis

I : Investigation

R : Resources

D : Data Curation

O : Writing - Original Draft

E : Writing - Review & Editing

Vi : Visualization

Su : Supervision

P : Project administration

Fu : Funding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

INFORMED CONSENT

We have obtained informed consent from all individuals included in this study.

DATA AVAILABILITY

The data that support the findings of this study are available on request from the corresponding author, B.J.




REFERENCES

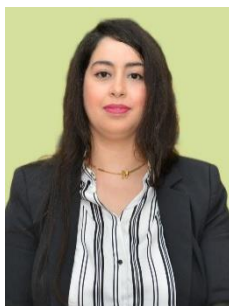
- [1] "The National Plan for Accelerating the Transformation of the Ecosystem," *Ministry of higher education scientific research and innovation*, 2022. Accessed: May 17, 2024. https://www.onousc.ma/actualites/plan-national-d-acceleration-de-la-transformation-de-l-ecosysteme-de-l-enseignement-superieur-de-la-recherche-scientifique-et-de-l-innovation?utm_source
- [2] C. A. Eden, O. N. Chisom, and I. S. Adeniyi, "Integrating AI in education: Opportunities, challenges, and ethical considerations," *Magna Scientia Advanced Research and Reviews*, vol. 10, no. 2, pp. 006–013, 2024, doi: 10.30574/msarr.2024.10.2.0039.
- [3] R. Farhat, Y. Mourali, M. Jemni, and H. Ezzedine, "An overview of machine learning technologies and their use in E-learning," in *2020 International Multi-Conference on: "Organization of Knowledge and Advanced Technologies" (OCTA)*, Feb. 2020, pp. 1–4, doi: 10.1109/OCTA49274.2020.9151758.
- [4] C. A. E. Piter, S. Hadi, and I. N. Yulita, "Multi-label classification for scientific conference activities information text using extreme gradient boost (XGBoost) method," in *2021 International Conference on Artificial Intelligence and Big Data Analytics*, Oct. 2021, pp. 1–5, doi: 10.1109/ICAIBDA53487.2021.9689699.
- [5] M. Mohammed, M. B. Khan, and E. B. M. Bashier, *Machine learning*. CRC Press, 2016.
- [6] J. Alzubi, A. Nayyar, and A. Kumar, "Machine learning from theory to algorithms: An overview," *Journal of Physics: Conference Series*, vol. 1142, p. 012012, Nov. 2018, doi: 10.1088/1742-6596/1142/1/012012.
- [7] J. Oller, A. Engel, and M. J. Rochera, "Personalizing learning through connecting students' learning experiences: an exploratory study," *The Journal of Educational Research*, vol. 114, no. 4, pp. 404–417, Jul. 2021, doi: 10.1080/00220671.2021.1960255.
- [8] A. W. Fazil, M. Hakimi, A. K. Shahidzay, and A. Hasas, "Exploring the broad impact of AI technologies on student engagement and academic performance in university settings in Afghanistan," *RIGGS: Journal of Artificial Intelligence and Digital Business*, vol. 2, no. 2, pp. 56–63, Jan. 2024, doi: 10.31004/riggs.v2i2.268.
- [9] J. J. Faraway, "Does data splitting improve prediction?," *Statistics and Computing*, vol. 26, no. 1–2, pp. 49–60, Jan. 2016, doi: 10.1007/s11222-014-9522-9.
- [10] W. Li, Y. Yin, X. Quan, and H. Zhang, "Gene expression value prediction based on XGBoost algorithm," *Frontiers in Genetics*, vol. 10, Nov. 2019, doi: 10.3389/fgene.2019.01077.
- [11] M. A. A. da Cruz, L. R. Abbade, P. Lorenz, S. B. Mafra, and J. J. P. C. Rodrigues, "Detecting compromised IoT devices through XGBoost," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 12, pp. 15392–15399, Dec. 2023, doi: 10.1109/TITS.2022.3187252.




- [12] J. R. Quinlan, "Induction of decision trees," *Machine Learning*, vol. 1, no. 1, pp. 81–106, Mar. 1986, doi: 10.1007/BF00116251.
- [13] V. G. Costa and C. E. Pedreira, "Recent advances in decision trees: an updated survey," *Artificial Intelligence Review*, vol. 56, no. 5, pp. 4765–4800, May 2023, doi: 10.1007/s10462-022-10275-5.
- [14] B. Thomas and J. Chandra, "Random forest application on cognitive level classification of E-learning content," *International Journal of Electrical and Computer Engineering*, vol. 10, no. 4, pp. 4372–4380, 2020, doi: 10.11591/ijece.v10i4.pp4372-4380.
- [15] Y. Mansour and M. Schain, "Random forests," *Machine Learning*, vol. 45, no. 2, pp. 123–145, 2001, doi: 10.1023/A:1010950718922.
- [16] P. Darveau and P. Eng, "Support vector machines: modeling the dual cognitive processes of an SVM," 2023.
- [17] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, doi: 10.1007/BF00994018.
- [18] N. Qasrina Ann, D. Pebrianti, M. F. Abas, and L. Bayuaji, "Automated-tuned hyper-parameter deep neural network by using arithmetic optimization algorithm for Lorenz chaotic system," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 13, no. 2, pp. 2167–2176, Apr. 2023, doi: 10.11591/ijece.v13i2.pp2167-2176.
- [19] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015, doi: 10.1038/nature14539.
- [20] K. Budholiya, S. K. Shrivastava, and V. Sharma, "An optimized XGBoost based diagnostic system for effective prediction of heart disease," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 7, pp. 4514–4523, Jul. 2022, doi: 10.1016/j.jksuci.2020.10.013.
- [21] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Aug. 2016, pp. 785–794, doi: 10.1145/2939672.2939785.
- [22] D. A. Rusdah and H. Murfi, "XGBoost in handling missing values for life insurance risk prediction," *SN Applied Sciences*, vol. 2, no. 8, p. 1336, Aug. 2020, doi: 10.1007/s42452-020-3128-y.
- [23] X. Deng *et al.*, "Bagging–XGBoost algorithm based extreme weather identification and short-term load forecasting model," *Energy Reports*, vol. 8, pp. 8661–8674, Nov. 2022, doi: 10.1016/j.egy.2022.06.072.
- [24] X. Chen and B. Zhang, "A XGBoost algorithm-based fatigue recognition model using face detection," pp. 1–7, 2023.
- [25] Y. Guang, "Generalized XGBoost method," *arXiv preprint arXiv:2109.07473*, 2021.
- [26] A. Prabha, J. Yadav, A. Rani, and V. Singh, "Design of intelligent diabetes mellitus detection system using hybrid feature selection based XGBoost classifier," *Computers in Biology and Medicine*, vol. 136, p. 104664, Sep. 2021, doi: 10.1016/j.combiomed.2021.104664.
- [27] A. El Mezouari, A. El Fazziki, and M. Sadgal, "Hadoop–spark framework for machine learning-based smart irrigation planning," *SN Computer Science*, vol. 3, no. 1, p. 10, Jan. 2022, doi: 10.1007/s42979-021-00856-6.
- [28] D. M. Belete and M. D. Huchaiah, "Grid search in hyperparameter optimization of machine learning models for prediction of HIV/AIDS test results," *International Journal of Computers and Applications*, vol. 44, no. 9, pp. 875–886, Sep. 2022, doi: 10.1080/1206212X.2021.1974663.
- [29] P. Sedgwick, "Pearson's correlation coefficient," *BMJ*, vol. 345, no. jul04 1, pp. e4483–e4483, Jul. 2012, doi: 10.1136/bmj.e4483.
- [30] R. Susmaga, "Confusion matrix visualization," in *Intelligent Information Processing and Web Mining*, 2004, pp. 107–116.
- [31] A. A. Alsulami, A. S. A. M. AL-Ghamdi, and M. Ragab, "Enhancement of E-learning student's performance based on ensemble techniques," *Electronics (Switzerland)*, vol. 12, no. 6, 2023, doi: 10.3390/electronics12061508.
- [32] G. Theophilus and C. I. Eke, "Machine learning-based e-learners' engagement level prediction using benchmark datasets," *International Journal of Applied Information Systems*, vol. 12, no. 41, pp. 23–32, Sep. 2023, doi: 10.5120/ijais2023451951.
- [33] P. Nayak, S. Vaheed, S. Gupta, and N. Mohan, "Predicting students' academic performance by mining the educational data through machine learning-based classification model," *Education and Information Technologies*, vol. 28, no. 11, pp. 14611–14637, Nov. 2023, doi: 10.1007/s10639-023-11706-8.
- [34] S. Hussain and M. Q. Khan, "Student-performulator: predicting students' academic performance at secondary and intermediate level using machine learning," *Annals of Data Science*, vol. 10, no. 3, pp. 637–655, Jun. 2023, doi: 10.1007/s40745-021-00341-0.

BIOGRAPHIES OF AUTHORS






Brahim Jabir    Obtained his degree in computer engineering from the Faculty of Sciences and Techniques at the University Sultan Moulay Slimane in Beni Mellal Morocco. He is a professor in computer science at Chouaib Doukkali University. Previously, he taught at the Training College in the region of Beni Mellal Khenifra. Br. JABIR has worked as the Head of the ICT Department at the Teacher Training College and was the Leader of the LANDitic Research Group. Besides, he is working as an Associate Editor of International Journal of Information Technologies Systems (IGI Global). His interest in research area is how artificial intelligence affects different subject areas. His email: ibra.jabir@gmail.com.






Soukaina Merzouk    received the master's degree in information science and engineering from Hassan II University - Casablanca Faculty of Science Ben M'sik, Casablanca, Morocco, in 2016 and the Ph.D. degree in Computer Science from Hassan II University - Casablanca Faculty of Science Ben M'sik, Casablanca, Morocco, in 2022. She is currently professor of Computer Science at Polydisciplinary Faculty of Sidi Bennour, Chouaib Doukkali University El Jadida, Morocco. Her research interests include agile software development, metamodeling, internet of things, artificial intelligence applications, data mining techniques, big data, and machine learning. She can be contacted at email: merzouk.soukaina@gmail.com.



Hamzaoui Radoine    is PhD student in computer science, Innovation Laboratory in Mathematics, Applications and Information Technologies Polydisciplinary Faculty Sultan Moulay Sliman University Beni Mellal, Morocco, professor trainer of ICTE at CRMEF of Beni mellal-Khenifra, obtained a master specialized in mathematics and technology education, field of interest: ICT and e-learning, research topic: evaluation of distance learning in the bachelor cycle -case of Beni mellal. His email id is: info.hamzaoui@gmail.com.



Noureddine Falih    is PhD in computer science from Faculty of Sciences and Technologies of Mohammedia, Morocco in 2013. He is an associate professor in Polydisciplinary Faculty of Sultan Moulay Slimane University at Beni Mellal, Morocco since 2014. He has 18 years of professional experience in several prestigious companies. His research topics are related to information system governance, business intelligence, big data analytics, and digital agriculture. He can be contacted at email: nourfald@yahoo.fr.