Enhance Learning Performance Predictions With Explainable Machine Learning

Wan-Chong Choi
Faculty of Applied Sciences
Macao Polytechnic University
Macao SAR, China
& CISUC, University of Coimbra, Coimbra, Portugal
wanchong.choi@mpu.edu.mo

Chan-Tong Lam
Faculty of Applied Sciences
Macao Polytechnic University
Macao SAR, China
ctlam@mpu.edu.mo

António José Mendes
Dep. of Informatics Engineering, CISUC,
University of Coimbra
Coimbra, Portugal
toze@dei.uc.pt

***Abstract*—This Research Full Paper focuses on predicting learning performance using machine learning algorithms and interpreting the results using Explainable Machine Learning (EML) techniques.**

The study compared a comprehensive set of machine learning algorithms, including Logistic Regression, Decision Trees, AdaBoost, XGBoost, SVM, and KNN. The performance of these algorithms in predicting students' final grades in a course was accessed using various evaluation metrics.

Our study used feature selection to identify the most relevant predictors to enhance predictive accuracy, implemented the Synthetic Minority Over-sampling Technique (SMOTE) to address class imbalance, and performed hyperparameter optimization to find the most effective model settings. This comprehensive approach improved the predictive accuracy of our models over previous studies.

Additionally, the importance of early prediction in identifying at-risk students was explored, with models demonstrating promising accuracy at the first checkpoint of the course.

Departing from traditional machine learning research that often focused on model performance, our study integrated the EML technique of Shapley Additive exPlanations (SHAP), which is grounded on the theoretical framework of Game Theory, to facilitate the interpretation of the predictive outcomes. This approach offered an explanatory perspective on the key factors influencing model decisions. By contributing to the predictability and interpretability of student performance, this research enriched the field of Educational Data Mining (EDM) and enhanced the understanding of student learning trajectories.

Keywords—Learning performance prediction, Explainable machine learning, SHapley Additive exPlanations, Educational data mining, EDM

I. INTRODUCTION

Educational Data Mining (EDM), a subfield of data mining, utilizes machine learning, deep learning, or statistical methods to analyze educational data. One important use of EDM is predicting student performance. This involves analyzing data from various sources, including student demographics, engagement, academic achievements, and course quality, to predict learning outcomes. This comprehensive approach enables a nuanced understanding and prediction of student performance [1].

The advent of online platforms in virtual learning environments has significantly contributed to collecting large datasets, though it also presented unique challenges in extracting meaningful insights. Within this context, Explainable Machine Learning (EML) has emerged within EDM, allowing the interpretation and clarification of the decision-making processes of the black box models. This approach is ideally suited to tackle the complexities of data, enhance the predictive modeling of student performance, and improve the interpretability of the prediction results, thereby enriching the educational landscape.

Our study focused on predicting learning performance using machine learning algorithms and interpreting the results using EML techniques. Specifically, we assessed the performance of machine learning algorithms in predicting students' final grades in a course using the Open University Learning Analytics Dataset (OULAD) [2]. Furthermore, to gain deeper insights into the predictive outcomes, we used the EML technique of Shapley Additive exPlanations (SHAP) to explain the results based on the theoretical framework of Game Theory. The following research questions guided this study:

(1) How do different enhancement methods affect the performance of various machine learning algorithms?

(2) Which machine learning algorithm demonstrates the highest accuracy in predicting student performance?

(3) How early can student performance be reliably predicted?

(4) How does the EML technique of SHAP contribute to explaining the predictive outcomes?

We conducted a comprehensive study encompassing four main aspects to answer these questions.

First, we investigated the impact of various data enhancement techniques, such as the Synthetic Minority Over-sampling Technique (SMOTE) and hyperparameter optimization, on the performance of machine learning algorithms in predicting student learning performance.

Second, we compared a comprehensive set of machine learning algorithms, including Logistic Regression, Decision Trees, Adaptive Boosting (AdaBoost), Extreme Gradient Boosting (XGBoost), Support Vector Machine (SVM), and K-Nearest Neighbors (KNN), to identify the best-performing algorithm for predicting student performance.

Next, we evaluated the performance of various algorithms at eight checkpoints, corresponding to the timings of eight assessments during the course, to determine the earliest point at which student performance could be accurately predicted.

Lastly, we integrated the EML technique of SHAP, rooted in the theoretical framework of Game Theory, to facilitate the interpretation of predictive outcomes and provide insights into the key factors influencing model decisions.

The rest of the paper is organized as follows: Section II provides a literature review, Section III outlines the research methodology, Section IV details the study's results, Section V discusses the results, and Section VI concludes a summary.

II. Literature Review

EDM emerged as a pivotal technology in addressing complex educational challenges. It helps to enhance students' learning performance by refining the learning process, guiding learners, and providing tailored feedback and suggestions based on individual learning behaviors. Additionally, EDM has been instrumental in assessing learning materials, shaping curriculum design, and promptly identifying learning patterns and potential issues [3] [4].

A critical aspect of EDM has been incorporating diverse data sources, including student demographics, previous academic performance, course quality, and levels of engagement. Dringus [5] emphasized the necessity of basing performance prediction on extensive data collection throughout the learning process, ensuring transparency in processing and delivering results. Complementing this, Muhammad et al. [6] identified significant variables, like student demographics and course quality, as influential factors in student performance.

Various studies demonstrated diverse approaches using the OULAD dataset. Heuer and Breiter [7] utilized the SVM algorithm focusing on demographic and behavioral features, achieving an 88.0% accuracy in predicting student success. Rizvi et al. [8] employed the Decision Tree algorithm, concentrating on demographic features like regional belonging and socio-economic status, and achieved an 83.1% accuracy. Waheed et al. [9] adopted the ANN algorithm targeting behavioral features, reaching an 89.0% accuracy with their novel approach using adversarial networks. Adnan et al. [10] applied the Random Forest algorithm, blending demographic, assessment, and behavioral features to achieve a 91.0% accuracy. Esteban et al. [11] employed the MLP algorithm, focusing on assessment features to attain a 93.9% accuracy.

Moreover, EML emerged as a crucial field within EDM for user comprehension, especially in black box models. EML interprets and clarifies the decision-making processes of these models in various domains [12]. In education, researchers explored the application of EML techniques. For instance, a deep learning-based knowledge tracing model was developed and interpreted using the layer-wise relevance propagation method [13], enabling more personalized learning experiences. In another example, Local Interpretable Model-agnostic Explanations (LIME) were utilized to develop an explainable model predicting at-risk students using demographic and clickstream data [14].

However, while LIME approximates local surrogate models for individual predictions, it does not provide consistent feature importance scores across the entire dataset. To address this limitation, Lundberg and Lee introduced SHAP in 2017 [15], with its theoretical framework rooted in Game Theory's Shapley values, introduced by Lloyd Shapley [16] as a fair way to allocate gains in cooperative games.

In SHAP's theoretical framework, features are viewed as players in a game, and the model's predictions represent the game's outcomes. This analogy helps to understand how Shapley's values explain the model's decision-making process.

In a cooperative game, players work together to achieve a common goal, and the Shapley value is a way to distribute the total gains or losses among the players fairly. Similarly, in a machine learning model, features work together to make predictions, and the Shapley value helps to determine each feature's contribution to the final prediction.

To calculate the Shapley value for a feature, SHAP considers all possible combinations of features and computes the marginal contribution of the feature in each combination. This is done by comparing the model's prediction with and without the feature in question while keeping all other features constant. The Shapley value is then the average marginal contribution of the feature across all possible combinations.

By using Shapley values, SHAP ensures that the feature contributions are additive, meaning that the sum of all feature contributions equals the difference between the model's prediction and the average prediction for the dataset. This property makes Shapley values a consistent and comprehensive way to explain the model's decision-making process. The application of SHAP in education showed promising results. Research [17] used a neural network for university dropout prediction, leveraging SHAP to interpret the importance of various features, including academic and exam performance data, thus enabling customized intervention strategies for at-risk students.

However, despite its growing popularity in interpreting machine learning outcomes, the role of EML in education remained relatively underexplored. Our study predicts learning performance using machine learning algorithms and analyzes the results using EML techniques, specifically SHAP. By applying SHAP to our models, we aim to understand better the factors influencing student learning outcomes.

In summary, EDM emerged as a crucial interdisciplinary field, leveraging advanced techniques to enhance the understanding and improvement of educational processes. Despite its growing popularity, the role of EML in education remained relatively underexplored, presenting opportunities for further research to understand better factors influencing student learning outcomes and develop effective intervention strategies.

III. Methodology

A. Dataset

In this study, we utilized the OULAD [2], provided by the Open University, UK. This public dataset contains detailed educational data on courses, students, and their interactions with the virtual learning environment. The dataset includes various courses, represented by codes such as AAA, BBB, CCC, etc., to protect privacy. In this study, we have randomly selected the CCC course from the OULAD dataset as our data source.

B. Implementation tools

Python was the primary tool for building prediction models. The Scikit-learn library [18] implemented the machine learning algorithms. NumPy and Pandas were used for data manipulation and preprocessing, while SciPy was used for scientific computing. Matplotlib was used for data visualization, and Imblearn handled a class imbalance in the dataset.

The EML technique of SHAP [15] was implemented using the SHAP library in Python. This library generated visual force plots, analyzed why machine learning algorithms made specific predictions, and offered an understanding of the factors significantly affecting student performance.

C. Selected Algorithms

This study utilized six machine learning algorithms for their demonstrated efficacy in educational settings.

1) Logistic Regression

Logistic Regression, traditionally used for modeling the likelihood of binary outcomes based on predictor variables [19], can be adapted to a multi-class setting. This method demonstrated its flexibility by categorizing educational data into multiple distinct classes. Its application extended beyond the conventional binary framework, showcasing its effectiveness in handling complex, multi-class educational datasets.

2) Decision Tree

Decision Tree provides an intuitive classification method, segmenting data based on feature nodes and decision branches [20]. Their straightforward nature and minimal computational demands make them suitable for preliminary investigations of datasets. However, their effectiveness may be limited when handling continuous features or large, diverse datasets.

3) AdaBoost

Adaptive Boosting, commonly known as AdaBoost, focuses on sequentially improving classification by adjusting to incorrectly classified samples [21]. Its structure, which combines several decision trees, helps prevent overfitting and effectively refining predictions, making it suitable for adaptive learning scenarios in educational data.

4) XGBoost

Extreme Gradient Boosting, also known as XGBoost, is a machine learning technique that enhances prediction accuracy by sequentially building decision trees and fine-tuning the significance of variables [22]. Known for efficiently correcting its mistakes through iterative refinement, XGBoost excels in processing and interpreting complex datasets, making it well-suited for the educational domain.

5) SVM

Support Vector Machine, also known as SVM, was rooted in the statistical learning theory [23] and excelled in data classification by maximizing the separation between categories in feature space. Notable for its effectiveness in linear and non-linear classification and its high accuracy attributed to the large margin property, SVM was particularly apt for educational datasets, which often encompass complex, multi-dimensional relationships.

6) KNN

K-Nearest Neighbors, also known as KNN, distinguish themselves through their simplicity and efficacy in classifying data based on proximity to the nearest neighbors [24]. This method was valuable in educational datasets where patterns might not be linearly discernable, and the relationship between features and performance was not explicitly defined.

D. Data Preprocessing

Initially, we excluded records containing missing values. After processing the data, the dataset comprised 4,434 student records.

E. Data Normalization

In the subsequent step, we employed the Min-max scaler for data normalization, a technique frequently adopted in EDM studies [25] [26]. This method was chosen due to the considerable range diversity in the raw data of the educational features, which was critical for ensuring the reliability of data predictions. It scaled the data to fall within a [0, 1] range, using the equation (1), where R_{max} and R_{min} denoted the maximum and minimum values of the dataset, respectively.

$$R' = \frac{R - R_{min}}{R_{max} - R_{min}} \in [0, 1] \quad (1)$$

F. Feature Selection

In our study, feature selection was vital in improving our predictive models in EDM [27]. The target variable, which is the variable that the model aims to predict, is the final grade of the students in the course. Initially, we selected 22 input features based on their relevance to the target variable. These features included demographic data (gender, highest level of education, age, number of studied credits, and disability status), assessment data (scores from eight assessments), and behavioral data (sum of clicks for each assessment and the number of attempts across all assessments).

Next, we used the Chi-square test [28] to identify the most significant features by evaluating the strength of the association between each feature and the target variable. We also used the Pearson correlation test, which yielded results that closely resemble the Chi-square test results.

Our findings in Table I show that higher Chi-square scores indicated a stronger relationship with the student's final grade in the course. This statistical approach allowed us to determine that demographic features were less impactful on final grades than assessment scores and behavioral features, helping us decide which features to include or exclude.

We excluded features (ranked 18 to 22) with a Chi-Square score less than 12 from the analysis. These features had weaker relevance and might not contribute significantly to the model's predictive power, potentially even reducing the predictive performance. We tested the model and found that excluding these features improved the model's predictive performance.

Figure 1 graphically emphasizes the hierarchical significance of the features, illustrating that assessment scores are significantly more predictive than demographic and behavioral data for determining the final grade.

TABLE I. CHI-SQUARE TEST RESULTS

Rank	Type	Feature	Chi-Square Score	Result
1	Assessment	assessment8_score	1529.9523	Included
2	Assessment	assessment5_score	1335.6470	Included
3	Assessment	assessment7_score	1215.1540	Included
4	Assessment	assessment6_score	1158.4380	Included
5	Assessment	assessment4_score	928.1105	Included
6	Assessment	assessment3_score	890.4325	Included
7	Assessment	assessment2_score	478.2245	Included
8	Assessment	assessment1_score	427.9368	Included
9	Behavioral	assessment7_click	113.7401	Included
10	Behavioral	assessment5_click	108.1367	Included
11	Behavioral	assessment6_click	106.5908	Included
12	Behavioral	assessment4_click	94.2063	Included
13	Behavioral	assessment8_click	90.1063	Included
14	Behavioral	assessment3_click	73.6920	Included
15	Behavioral	assessment1_click	38.8124	Included
16	Behavioral	assessment2_click	38.6289	Included
17	Demographic	highest_education	12.5810	Included
18	Behavioral	num_of_prev_attempts	10.3623	Excluded
19	Demographic	age_band	6.8003	Excluded
20	Demographic	gender	5.3832	Excluded
21	Demographic	studied_credits	4.0641	Excluded
22	Demographic	disability	3.9571	Excluded

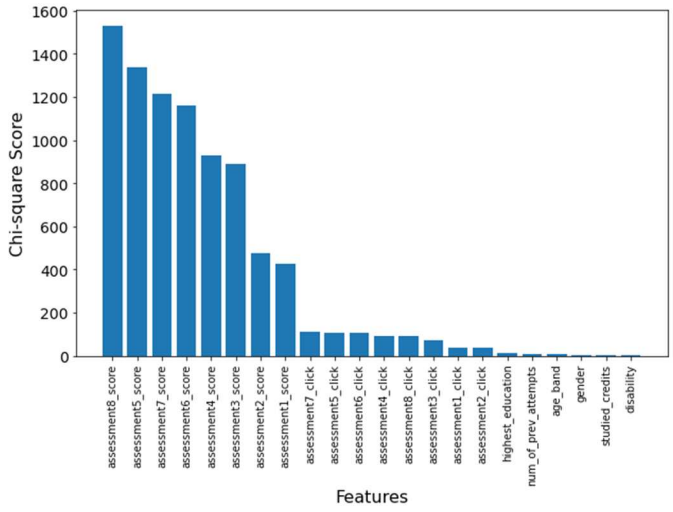

Fig. 1. Visualization of Feature Importance based on Chi-square Score

Table II shows the final 17 features and one target variable after excluding the less influential features.

TABLE II. FEATURES AND TARGET VARIABLE

ID	Type	Description	Value	Normalization
1	Demographic Feature	Highest education upon module enrollment	[0, 0.25, 0.5, 0.75, 1]	0: No formal 0.25: Below A level 0.5: A level or equivalent 0.75: HE qualification 1: Postgraduate
2–9	Assessment Feature	Scores from assessment 1 to 8	[0 - 1]	0 – 100 scaled to [0 – 1]
10–17	Behavioral Feature	Sum of clicks from assessment 1 to 8	[0 - 1]	0 – N scaled to [0 – 1]
18	Target Variable	Final grade of a course	[0,1]	0: Fail 1: Pass

Demographic data: The first feature represents students' highest level of education, scaled from 0 (no formal education) to 1 (postgraduate degree), with intermediate values denoting various educational stages.

Assessment data: Features 2 to 9 are assessment scores initially on a scale from 0 to 100 and normalized to a 0 to 1 range.

Behavioral data: Features 10 to 17 capture the sum of clicks from assessments 1 to 8, with the original click counts ranging from 0 to N, normalized to a 0 to 1 range.

Target variable: The final grade is the predictive target of the study, utilizing a binary classification that categorizes students into two groups: 1 denotes Pass, and 0 denotes Fail.

G. *Cross-Validation*

Our study utilized 10-fold cross-validation, a widely used method in EDM [29] [30], to effectively train and evaluate machine learning models. This method divides the dataset into ten equal parts, using nine folds for training and one for testing in each iteration [31].

By employing this strategy, we reduced biases and the risk of overfitting, enhancing the reliability and generalizability of our models. The results from all iterations were averaged to provide a comprehensive measure of the model's effectiveness.

H. *Evaluation Metrics*

Our study employed several vital metrics to comprehensively analyze our model's cross-validation performance. These metrics, which are essential for understanding the effectiveness of a classification model, include accuracy, recall, precision, and F1-Score.

I. *Application of SMOTE*

The used dataset showed a notable imbalance in the target variable: 2756 students failed while only 1678 passed. This uneven class distribution can hinder the performance of machine learning models, often causing a bias towards the majority class (Fail class).

To counteract the data imbalance, we employed the SMOTE [32]. This method creates synthetic examples of the minority class (Pass class) to balance the dataset. By using SMOTE, we aimed to create a more evenly distributed dataset for training our models, enhancing their accuracy and generalizability.

J. *Hyperparameter Optimization*

In our pursuit of optimal model performance, we employed a randomized search [33] strategy for hyperparameter optimization [34], following the application of SMOTE. Randomized search, known for its efficiency in both time and computational resources, was pivotal in exploring the extensive hyperparameter space. The randomized approach injected an element of stochasticity, enabling a more diverse and unexpected exploration of potential parameter configurations. This method proved invaluable in uncovering the most effective settings for our models, thereby substantially elevating their predictive accuracy. The significance of this step lies in its ability to tailor the models precisely to our data characteristics through a strategic and efficient discovery of optimal parameters, leading to enhanced prediction precision.

IV. RESULTS

A. Comparison of Different Enhancement Methods

This section compares the efficacy of various machine learning models augmented with different enhancement methods.

The algorithms' performance, as delineated in Table III, was quantified by their F1 scores under four scenarios: the basic model (no use of enhancement methods), the application of SMOTE, the combination of SMOTE and parameter optimization with all features, and the combination of SMOTE and parameter optimization with feature selection (excluding the features ranked 18 to 22 with weaker relevance).

TABLE III. PERFORMANCE COMPARISON OF ENHANCEMENT METHODS

Algorithm	Basic Model	SMOTE	SMOTE+ Parameter Optimization+ All Features	SMOTE+ Parameter Optimization+ Feature Selection	Increase
Logistic Regression	0.8936	0.9263	0.9288	0.9297	4.0%
Decision Tree	0.9078	0.9336	0.9339	0.9424	3.8%
AdaBoost	0.9035	0.9402	0.9428	0.9442	4.5%
XGBoost	0.9066	0.9505	0.9492	0.9534	5.2%
SVM	0.9106	0.9391	0.9465	0.9502	4.4%
KNN	0.8998	0.9427	0.9457	0.9496	5.5%

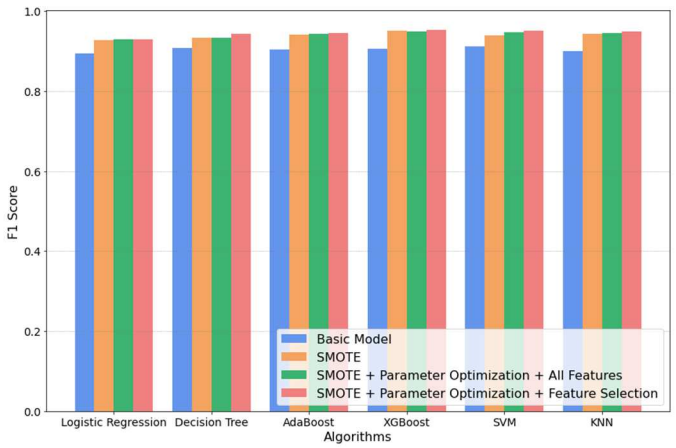

Fig. 2. Comparison of F1 across Different Algorithms and Methods

Figure 2 illustrates the F1 scores for various machine learning algorithms, each undergoing progressive enhancements. Implementing SMOTE notably improved the models' performance, exemplified by the Logistic Regression's F1 score rising from 89.36% to 92.63% and the Decision Tree's from 90.78% to 93.36%. This enhancement trend was consistently observed across all models after applying SMOTE.

Moreover, the improvement through parameter optimization resulted in increments in model performance. We tested the model with all demographic factors (SMOTE combined with parameter optimization and all features) and compared it to the feature selection model (SMOTE combined with parameter optimization and feature selection, excluding the features with weaker relevance). We found that the latter approach performed better. This finding underscores the importance of feature selection in enhancing model performance.

Using SMOTE combined with parameter optimization and feature selection yielded the best performance. For instance, the F1 score for XGBoost achieved the highest performance in this study at 95.34%. The SVM model also improved, reaching 95.02%. Although the improvement from parameter optimization was relatively minor, it proved that it can further enhance the model's performance.

In conclusion, using SMOTE combined with parameter optimization, the F1-score metric improved by 3.8% to 5.5% compared to the basic models. It confirmed the effectiveness of these enhancement methods in improving our models' performance, resulting in more accurate and reliable predictive models for EDM.

B. Prediction Result of Different Algorithms

Our analysis of the performance of various machine learning algorithms in outcome prediction, as detailed in Table IV, reflects results achieved by applying SMOTE and hyperparameter optimization across all algorithms.

TABLE IV. COMPARATIVE PERFORMANCE OF DIFFERENT ALGORITHMS

Algorithm	Accuracy	Precision	Recall	F1
Logistic Regression	0.9286	0.9164	0.9438	0.9297
Decision Tree	0.9415	0.9275	0.9583	0.9424
AdaBoost	0.9435	**0.9319***	0.9574	0.9442
XGBoost	**0.9522***	0.9291	**0.9792***	**0.9534***
SVM	0.9490	0.9290	0.9728	0.9502
KNN	0.9481	0.9244	0.9764	0.9496

*Asterisks indicate the highest values in each metric

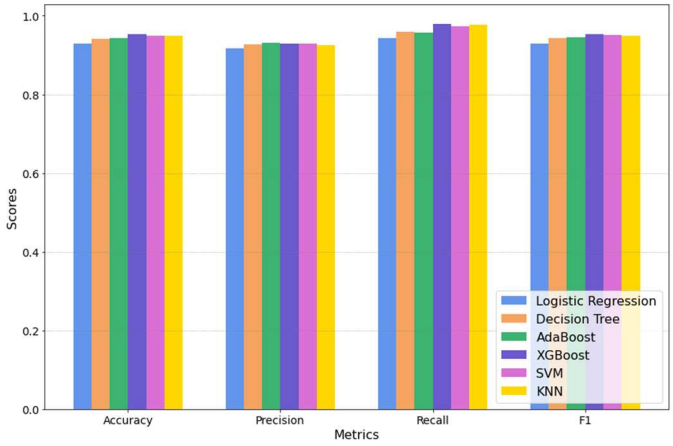

Fig. 3. Comparison of Different Algorithms Across Various Metrics

The Logistic Regression algorithm exhibited the least favorable performance among all the algorithms, with an accuracy of 92.86%, precision of 91.64%, recall of 94.38%, and F1 score of 92.97%.

The Decision Tree algorithm demonstrated moderate performance, with an accuracy of 94.15%, precision of 92.75%, recall of 95.83%, and F1 score of 94.24%.

AdaBoost showcased good results, with an accuracy of 94.35%, the highest precision of 93.19%, a recall of 95.74%, and an F1 score of 94.42%.

Compared to other algorithms, XGBoost yielded the best predictive performance, with the highest accuracy of 95.22%, precision of 92.91%, best recall of 97.92%, and highest F1 score of 95.34%.

The SVM algorithm provided robust results, with an accuracy of 94.90%, precision of 92.90%, recall of 97.28%, and F1 score of 95.02%. Lastly, the KNN algorithm performed well with an accuracy of 94.81%, precision of 92.44%, recall of 97.64%, and F1 score of 94.96%.

Figure 3 facilitates a visual comparison of the algorithms. The bar chart distinctly illustrates the dominance of XGBoost across most metrics, as it achieves the highest accuracy, recall, and F1 scores. While the other algorithms performed commendably, it is noteworthy that AdaBoost outperformed XGBoost in precision.

In conclusion, the XGBoost algorithm was the standout performer, excelling in accuracy. recall, and F1 score. This analysis highlighted its superiority over other algorithms, with AdaBoost also demonstrating notable performance in terms of precision.

C. Predicting Student Performance at an Early Stage

The ability to anticipate student performance, particularly identifying those who may require additional support early in their academic journey, is essential for enhancing educational outcomes. We evaluated the performance of different algorithms in predicting student success across eight checkpoints (Corresponding to the time of the eight assessments), with their F1 scores detailed in Table V and presented in Figure 4.

TABLE V. F1 SCORES OF ALGORITHMS ACROSS ASSESSMENTS

Checkpoint	Logistic Regression	Decision Tree	AdaBoost	XGBoost	SVM	KNN
Assessment1	0.8059	0.7927	0.8000	0.7944	0.8069	0.7821
Assessment2	0.8008	0.8305	0.8307	0.8350	0.8384	0.8489
Assessment3	0.8722	0.8710	0.8745	0.8879	0.8762	0.8835
Assessment4	0.8751	0.8828	0.9018	0.9115	0.8968	0.9074
Assessment5	0.9146	0.9136	0.9163	0.9308	0.9235	0.9288
Assessment6	0.9210	0.9248	0.9238	0.9362	0.9330	0.9347
Assessment7	0.9209	0.9338	0.9377	0.9491	0.9418	0.9429
Assessment8	0.9297	0.9424	0.9442	0.9534	0.9502	0.9496

At assessment 1, most algorithms demonstrated promising predictive accuracy, with the F1 score reaching 78.21% to 80.69%. For instance, SVM exhibited an F1 score of 80.69%. This early indication was crucial as it can allow educators to identify potentially struggling students right from the beginning of the course.

As the course progressed to assessment 2, predictive accuracy was markedly improved across most algorithms. Notably, KNN peaked with an F1 score of 84.89%, suggesting increasing reliability of the predictions as more data became available.

The trend of increasing accuracy continued into assessment 3, where XGBoost achieved an F1 score of 88.79%. This benchmark indicated the model's high performance in detecting students who might encounter academic difficulties early.

By assessment 4, the accuracy of predictions surpassed 90% for several models. This highlighted a significant point in the course where predictions became highly reliable, allowing for even more precise identification and support for at-risk students. For example, XGBoost reached an F1 score of 91.15%.

Moving on to assessments 5 and 6, the performance of the algorithms remained consistently high. The XGBoost model achieved an F1 score of 93.08% at assessment 5, and 93.62% at assessment 6.

Notably, as the course approached the final assessments, the incremental improvements in prediction accuracy became less pronounced. For instance, between assessment six and assessment 8, the increase in accuracy was relatively minor, albeit maintaining a high level. This plateau suggested that the most critical predictive data was captured in the earlier stages of the course.

These findings underscored the potential of machine learning in educational settings, especially in the early stages of a course. By leveraging predictive models, educators can identify students who are at risk of failing early. This enables them to tailor their teaching methods and interventions, providing the necessary support for those students. This proactive approach could be pivotal in enhancing overall learning outcomes and reducing failure rates.

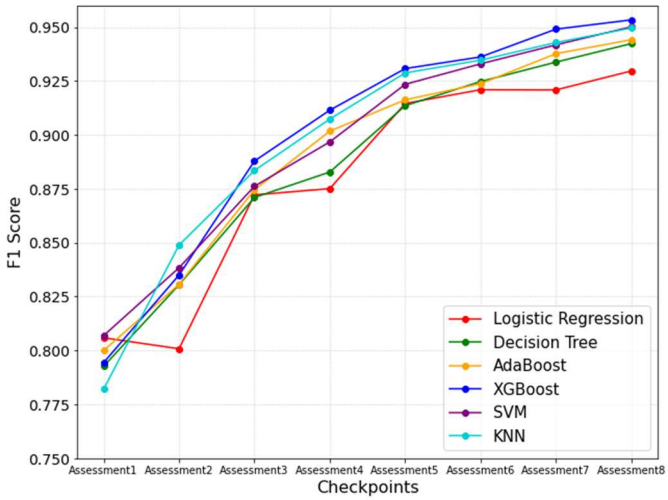

Fig. 4. Prediction Results of Different Algorithms Over Time

D. Interpret the Predictive Outcome by SHAP

To further explain the predictive outcome, we integrated the EML technique of SHAP to facilitate the interpretation of the top-performing XGBoost model. This approach could help the understanding of the inner workings of black-boxed machine learning algorithms by leveraging the theoretical framework of Game Theory.

As an EML technique, SHAP calculates each feature's SHAP values by analyzing every possible subset of features and assessing the feature's effect on the model's predictions. This method provides a detailed understanding of each feature's influence on the model's decision-making process.

Fig. 5. SHAP Force Plot of Student A with a Failed Grade Prediction

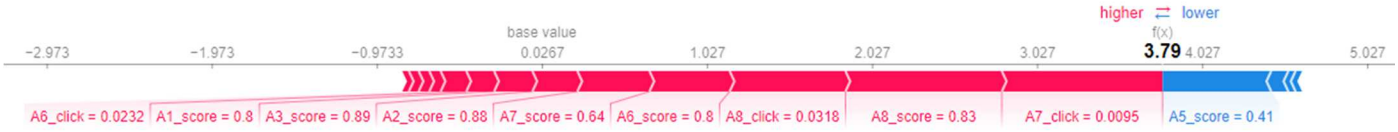

Fig. 6. SHAP Force Plot of Student B with a Passed Grade Prediction

SHAP force plots are typically used to visualize the feature contributions for individual instances. In our study, the individual instances are students. Therefore, SHAP force plots were used to measure the contribution of each feature in the model toward the prediction results of the particular students, providing essential insights into the model's decision-making process at the individual student level.

These plots visualize each feature's contribution along a spectrum, where the position and width of each feature indicate its relative impact on the model's prediction. Blue features represent factors influencing the model toward a failure prediction, while red features indicate factors pushing the prediction toward passing. The extent of each feature in the graph corresponds to its impact, with broader features showing a more substantial influence on the prediction's direction.

In the game theory theoretical framework, the SHAP force plots for Students A and B can be interpreted as a cooperative game in which each feature (player) contributes to the model's prediction (outcome). The SHAP values represent the fair allocation of each feature's contribution to the prediction.

1) SHAP Analysis for Fail Prediction

In Figure 5, dedicated to student A, the model predicted failure, aligning with the student's actual outcome. The visualization showcased a model output value of fx(-2.49), a significant deviation from the base value of 0.0267. This strongly indicated that the model predicted the student would fail.

The blue features, predominantly assessment scores, illuminated the student's challenges. The scores ranged from 0 to 0.51 in eight relatively low assessments. These lower scores indicated struggles in understanding or performing well in these assessments, significantly swaying the model towards a failure prediction. To address these challenges, teachers can offer personalized tutoring or extra help sessions focusing on the topics where the student scored low.

In contrast, the red features revealed a different perspective on Student A. The behavioral data showed the student's engagement with course assessments, including click counts: 0.0266 for assessment 8, 0.0039 for assessment 7, 0.0225 for assessment 1, and 0.1325 for assessment 5. Although these features indicated some level of participation, they were not enough to shift the model's prediction to a pass.

2) SHAP Analysis for Pass Prediction

Turning to Student B in Figure 6, the model presented an optimistic prediction of success with an output value of fx(3.79). This value, when compared to the base value of 0.0267, strongly suggested a pass, aligning with the student's successful academic performance.

The red features were pivotal, pushing the model's prediction toward the pass. The most influential factors included the scores in assessment 8 at 0.83, assessment 6 at 0.8, assessment 7 at 0.64, assessment 2 at 0.88, assessment 3 at 0.89, and assessment 1 at 0.8. Additionally, the clicks of engagement at assessment 7 at 0.0095, assessment 8 at 0.0318, and assessment 6 at 0.0232. To further cultivate the potential, Student B demonstrated excellence and could be provided with advanced learning materials and opportunities for deeper engagement, such as participation in higher-level courses or academic competitions.

However, not all factors were positively aligned for Student B. The blue features, such as a lower score of 0.41 in assessment 5, suggested areas for improvement. This score, albeit lower, did not significantly impact the model's overall optimistic passing prediction.

3) Summary of SHAP Analysis

Overall, the EML technique of SHAP allowed us to interpret the predictive outcomes of Students A and B. It provided clear insights into the factors influencing their academic performance predictions, leveraging the theoretical framework of Game Theory.

Student A's lower assessment scores and moderate engagement levels led to a failure prediction. This highlighted the importance of consistent and thorough understanding and assessment performance for academic success.

High assessment scores and active engagement supported the model's successful prediction for Student B. This case illustrated the positive impact of consistent high performance and engagement on academic outcomes.

These visualizations are valuable in pinpointing areas where educational support can be optimized, thus enhancing student success potential.

V. DISCUSSION

A. RQ1) How do different enhancement methods affect the performance of various machine learning algorithms?

Our study demonstrated that by applying SMOTE combined with parameter optimization, the F1-score metric of the machine learning models improved by 3.8% to 5.5% compared to the basic models without these enhancements.

These results highlighted the effectiveness of SMOTE in balancing the class distribution, which was crucial in educational datasets due to the inherent imbalance between pass and fail outcomes. This technique ensured that the models did not disproportionately favor the majority class, thereby improving the recall for minority classes. Furthermore, hyperparameter optimization complemented SMOTE by fine-tuning the models' parameters to align with the specific characteristics of educational data, further enhancing their predictive performance.

B. RQ2) Which machine learning algorithm demonstrates the highest accuracy in predicting student performance?

Among the machine learning algorithms evaluated in our study, the XGBoost algorithm exhibited the best performance, with an accuracy of 95.22%, precision of 92.91%, recall of 97.92%, and F1 score of 95.34%. XGBoost's superior performance could be attributed to its ability to effectively handle diverse and preprocessed educational data, identify underlying patterns, and manage complex feature interactions and dependencies.

To assess the performance of our XGBoost model with previous research, we compared its accuracy with other studies that utilized the OULAD dataset, as shown in Table VI.

TABLE VI. ACCURACY COMPARISON WITH PREVIOUS OULAD STUDIES

Ref	Algorithm	Features	Accuracy
[7]	SVN	Demographic, Behavioral	88.0%
[8]	Decision Tree	Demographic	83.1%
[9]	ANN	Behavioral	89.0%
[10]	Random Forest	Demographic, Assessment, Behavioral	91.0%
[11]	MLP	Assessment	93.9%
Our study	XGBoost	Demographic, Assessment, Behavioral	95.22%

In the accuracy comparison with previous OULAD studies, Heuer and Breiter [7] implemented the SVN algorithm with demographic and behavioral features, achieving 88% accuracy. Rizvi et al. [8] utilized the Decision Tree algorithm focusing on demographic features and attained an 83.1% accuracy. Waheed et al. [9] employed the ANN algorithm, targeting behavioral features, and reached an 89.0% accuracy. Adnan et al. [10] applied the Random Forest algorithm, incorporating demographic, assessment, and behavioral features, achieving a 91% accuracy. Esteban et al. [11] also employed the Multilayer Perceptron (MLP) algorithm, focusing on assessment features, achieving 93.9% accuracy.

The comparison revealed that our XGBoost model, with an accuracy of 95.22%, outperformed the previous studies, demonstrating an advancement in predictive accuracy.

C. RQ3) How early can student performance be reliably predicted?

Our study emphasized the significance of early prediction in identifying at-risk students and demonstrated the promising accuracy of the models in making predictions at an early stage. We could measure the models' predictive effectiveness at different course stages by assessing the performance of various algorithms at eight checkpoints, corresponding to the timings of eight assessments throughout the course.

The findings revealed that our models achieved a promising predictive accuracy, ranging from 78.21% to 80.69%, at the first checkpoint. This early indication of student performance underscored the potential of machine learning in educational contexts, particularly in enabling educators to provide timely interventions and support to students who might be struggling from the outset of the course.

D. RQ4) How does the EML technique of SHAP contribute to explaining the predictive outcomes?

The SHAP EML technique, grounded in the theoretical framework of Game Theory, provided valuable insights into the predictive outcomes of machine learning models. In our study, SHAP visualizations, particularly force plots, illustrated how different features contributed to the predictions made by the XGBoost model. These visualizations offered a clear and intuitive explanation of why Student A was predicted to fail, and Student B was expected to pass based on the contributions of their respective features.

For Student A, lower assessment scores and moderate engagement were the primary factors leading to the failure prediction, highlighting areas where personalized interventions could be beneficial. In contrast, Student B's consistently high scores and active participation drove the passing prediction, demonstrating the positive impact of academic diligence and engagement on success.

By leveraging the theoretical framework of Game Theory, SHAP simplified complex predictive models into easily interpretable visual explanations that educators could use to tailor their support strategies effectively. This approach positioned SHAP as a valuable tool in translating predictive outcomes into actionable insights, facilitating the development of personalized intervention plans that address students' weaknesses and build upon their strengths.

VI. CONCLUSION

Our findings revealed that the XGBoost algorithm performed best, improving upon previous research. This underscored the effectiveness of our approach, which included enhancement methods like SMOTE and hyperparameter optimization. These methodological refinements played a pivotal role in improving the predictive accuracy of the machine learning models.

The study also highlighted the critical role of early prediction in identifying at-risk students. It indicated that at the first checkpoint of the course, most algorithms demonstrated promising predictive accuracy, which is a significant advantage for educational institutions. This early detection enables timely

interventions, potentially improving student outcomes and reducing failure rates.

Departing from traditional machine learning research that often focused on model performance, our study integrated the EML technique of SHAP to facilitate interpreting the predictive outcomes. Using the theoretical framework of Game Theory, SHAP translates complex machine-learning models into intuitive visual explanations. This empowers educators to design tailored support strategies that address student weaknesses and build on their strengths, transforming predictive outcomes into actionable insights.

In summary, this study contributed to the field of EDM by demonstrating the effectiveness of machine learning algorithms in predicting student performance. The insights enhanced our understanding of student learning patterns and offered practical implications for educational institutions seeking to leverage data-driven strategies to improve student outcomes. Moreover, the EML technique of SHAP provided an explanatory perspective on the key factors influencing model decisions, offering a deeper understanding of student learning trajectories. Thus, our study contributed to both the predictability and interpretability of student performance in EDM.

REFERENCES

[1] A. Abu Saa, M. Al-Emran, and K. Shaalan, 'Factors affecting students' performance in higher education: a systematic review of predictive data mining techniques', *Technology, Knowledge and Learning*, vol. 24, pp. 567–598, 2019.

[2] J. Kuzilek, M. Hlosta, and Z. Zdrahal, 'Open university learning analytics dataset', *Scientific data*, vol. 4, no. 1, pp. 1–8, 2017.

[3] R. S. Baker and K. Yacef, 'The state of educational data mining in 2009: A review and future visions', *Journal of educational data mining*, vol. 1, no. 1, pp. 3–17, 2009.

[4] C. Romero, S. Ventura, and E. García, 'Data mining in course management systems: Moodle case study and tutorial', *Computers & education*, vol. 51, no. 1, pp. 368–384, 2008.

[5] L. P. Dringus, 'Learning analytics considered harmful.', *Journal of Asynchronous Learning Networks*, vol. 16, no. 3, pp. 87–100, 2012.

[6] M. H. bin Roslan and C. J. Chen, 'Educational Data Mining for Student Performance Prediction: A Systematic Literature Review (2015-2021).', *International Journal of Emerging Technologies in Learning*, vol. 17, no. 5, 2022.

[7] H. Heuer and A. Breiter, 'Student success prediction and the trade-off between big data and data minimization', *DeLFI 2018-Die 16. E-Learning Fachtagung Informatik*, 2018.

[8] S. Rizvi, B. Rienties, and S. A. Khoja, 'The role of demographics in online learning; A decision tree based approach', *Computers & Education*, vol. 137, pp. 32–47, 2019.

[9] H. Waheed *et al.*, 'Balancing sequential data to predict students at-risk using adversarial networks', *Computers & Electrical Engineering*, vol. 93, p. 107274, 2021.

[10] M. Adnan *et al.*, 'Predicting at-risk students at different percentages of course length for early intervention using machine learning models', *Ieee Access*, vol. 9, pp. 7519–7539, 2021.

[11] A. Esteban, C. Romero, and A. Zafra, 'Assignments as influential factor to improve the prediction of student performance in online courses', *Applied Sciences*, vol. 11, no. 21, p. 10145, 2021.

[12] I. Ahmed, G. Jeon, and F. Piccialli, 'From artificial intelligence to explainable artificial intelligence in industry 4.0: a survey on what, how, and where', *IEEE Transactions on Industrial Informatics*, vol. 18, no. 8, pp. 5031–5042, 2022.

[13] Y. Lu, D. Wang, Q. Meng, and P. Chen, 'Towards interpretable deep learning models for knowledge tracing', in *Artificial Intelligence in Education: 21st International Conference, AIED 2020, Ifrane, Morocco, July 6–10, 2020, Proceedings, Part II 21*, Springer, 2020, pp. 185–190.

[14] B. Pei and W. Xing, 'An interpretable pipeline for identifying at-risk students', *Journal of Educational Computing Research*, vol. 60, no. 2, pp. 380–405, 2022.

[15] S. M. Lundberg and S.-I. Lee, 'A unified approach to interpreting model predictions', *Advances in neural information processing systems*, vol. 30, 2017.

[16] L. S. Shapley, 'A value for n-person games', *Princeton University Press*, 1953.

[17] M. Baranyi, M. Nagy, and R. Molontay, 'Interpretable deep learning for university dropout prediction', in *Proceedings of the 21st annual conference on information technology education*, 2020, pp. 13–19.

[18] F. Pedregosa *et al.*, 'Scikit-learn: Machine learning in Python', *the Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011.

[19] S. Sperandei, 'Understanding logistic regression analysis', *Biochemia medica*, vol. 24, no. 1, pp. 12–18, 2014.

[20] Y.-Y. Song and L. U. Ying, 'Decision tree methods: applications for classification and prediction', *Shanghai archives of psychiatry*, vol. 27, no. 2, p. 130, 2015.

[21] Y. Freund and R. E. Schapire, 'A decision-theoretic generalization of on-line learning and an application to boosting', *Journal of computer and system sciences*, vol. 55, no. 1, pp. 119–139, 1997.

[22] 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*, 2016, pp. 785–794.

[23] D. Nasien, S. S. Yuhaniz, and H. Haron, 'Statistical learning theory and support vector machines', in *2010 Second International Conference on Computer Research and Development*, IEEE, 2010, pp. 760–764.

[24] K. Taunk, S. De, S. Verma, and A. Swetapadma, 'A Brief Review of Nearest Neighbor Algorithm for Learning and Classification', in *2019 International Conference on Intelligent Computing and Control Systems (ICCS)*, Madurai, India: IEEE, May 2019, pp. 1255–1260. doi: 10.1109/ICCS45141.2019.9065747.

[25] D. Bertovic, M. Mravak, K. Nikolov, and N. Vidovic, 'Using Moodle Test Scores to Predict Success in an Online Course', in *2022 International Conference on Software, Telecommunications and Computer Networks (SoftCOM)*, Split, Croatia: IEEE, Sep. 2022, pp. 1–7.

[26] J. Leinonen, F. E. V. Castro, and A. Hellas, 'Time-on-task metrics for predicting performance', *ACM Inroads*, vol. 13, no. 2, pp. 42–49, Jun. 2022.

[27] W.-C. Choi, C.-T. Lam, and A. J. Mendes, 'A Systematic Literature Review on Performance Prediction in Learning Programming Using Educational Data Mining', in *2023 IEEE Frontiers in Education Conference (FIE)*, IEEE, 2023, pp. 1–9.

[28] A. Chugh, *ML: chi-square test for feature selection*. GeeksforGeeks. https://www.geeksforgeeks.org/ml-chi-square-test-for-feature-selection/, 2018.

[29] P. Nuankaew, W. Nuankaew, D. Teeraputon, K. Phanniphong, and S. Bussaman, 'Prediction model of student achievement in business computer disciplines', *International Journal of Emerging Technologies in Learning (iJET)*, vol. 15, no. 20, pp. 160–181, 2020.

[30] D. Delen, 'Predicting student attrition with data mining methods', *Journal of College Student Retention: Research, Theory & Practice*, vol. 13, no. 1, pp. 17–35, 2011.

[31] C. F. De Oliveira, S. R. Sobral, M. J. Ferreira, and F. Moreira, 'How Does Learning Analytics Contribute to Prevent Students' Dropout in Higher Education: A Systematic Literature Review', *BDCC*, vol. 5, no. 4, p. 64, Nov. 2021, doi: 10.3390/bdcc5040064.

[32] A. Fernández, S. Garcia, F. Herrera, and N. V. Chawla, 'SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary', *Journal of artificial intelligence research*, vol. 61, pp. 863–905, 2018.

[33] J. Bergstra and Y. Bengio, 'Random search for hyper-parameter optimization.', *Journal of machine learning research*, vol. 13, no. 2, 2012.

[34] L. Yang and A. Shami, 'On hyperparameter optimization of machine learning algorithms: Theory and practice', *Neurocomputing*, vol. 415, pp. 295–316, 2020.