2024 IEEE International Conference on Teaching, Assessment and Learning for Engineering (TALE) | 979-8-3503-7623-4/24/$31.00 ©2024 IEEE | DOI: 10.1109/TALE62452.2024.10834292

Analyzing the Interpretability of Machine Learning Prediction on Student Performance Using SHapley Additive exPlanations

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 study compared several machine learning algorithms to predict student programming learning performance in an online learning environment. It used Explainable Machine Learning (EML) techniques to enhance interpretability. A range of algorithms, including Random Forest, Extra Trees, CatBoost, XGBoost, Naive Bayes, and K-Nearest Neighbors (KNN), were compared, with Extra Trees delivering the best results. Distinct from other EDM research mainly focused on predictive efficiency, we contributed by using the EML technique of SHapley Additive Explanations (SHAP), rooted in the Game Theory framework, to enhance model interpretability at both global and individual levels. At the global level, summary plots showed overall feature impacts, bar plots quantified the average effect of each feature, and dependence plots highlighted specific relationships. At the individual level, force plots identified critical features for individual predictions, decision plots traced the cumulative impact of features from the base value to the final output, and waterfall plots provided a breakdown of predictions. This study contributes to EDM by offering accurate predictive models and detailed interpretability, helping educational stakeholders make data-informed decisions to improve student outcomes.**

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

I. Introduction

Educational Data Mining (EDM) has emerged as a rapidly evolving field, focusing on extracting actionable knowledge from educational datasets to enhance learning experiences and outcomes. With the increasing complexity and volume of educational data generated through digital learning platforms, EDM faces significant challenges in accurately predicting student performance and providing personalized interventions.

Our study focused on predicting learning performance using machine learning algorithms and interpreting the results using Explainable Machine Learning (EML) techniques at both global and individual levels. By analyzing data from multiple dimensions in the Open University Learning Analytics Dataset (OULAD) [1], including demographic data, behavioral metrics, and academic achievements, we aimed to construct a holistic view of student performance, uncovering patterns and predictors of academic success or challenges. This study sought to address three research questions:

(1) Which machine learning algorithm performs best in predicting student academic outcomes?

(2) How does SHAP contribute to understanding predictive outcomes at the global level?

(3) How does SHAP assist in understanding predictive outcomes at the individual level?

To address these research questions, we undertook the following three main tasks.

First, this study employed six machine learning algorithms chosen for their demonstrated efficacy in addressing the educational data. We compared their performance to identify the most effective one for predicting student outcomes.

Second, distinct from other EDM research mainly focused on predictive efficiency, we utilized SHapley Additive Explanations (SHAP), grounded in the theoretical framework of Game Theory [2], to interpret the outcomes of these machine learning models at the global level. SHAP is an EML technique that assigns importance values to each feature, indicating their contribution to the model's predictions. We employed three types of SHAP visualizations: summary plots, bar plots, and dependence plots to elucidate feature importance and interactions.

Finally, our study used three other types of SHAP visualization, force plots, decision plots, and waterfall plots, to interpret the outcomes at the individual level, enabling a deeper understanding of the factors influencing the predictions for each student.

The remainder of the paper is structured as follows: Section II reviews the related work, Section III describes our methodology, Section IV presents the results, Section V discusses our findings, and Section VI concludes our study.

II. Related Work

EDM aims to optimize the educational process by providing practical student guidance, personalized feedback, evaluation of learning materials and curriculum structures, and prompt identification of unusual learning patterns or potential obstacles [3] [4]. However, accurately predicting student performance remains a complex challenge, influenced by factors such as the student's background, prior academic achievements, behavioral patterns [5] [6], and psychometric data [7]. Overcoming these challenges and continuing research on predicting student performance is essential for improving educational outcomes.

Several studies have utilized the OULAD dataset to showcase diverse EDM approaches. Heuer and Breiter [8] used the SVM algorithm, emphasizing demographic and behavioral features, and achieved 88% accuracy in predicting student performance. Rizvi et al. [9] applied the Decision Tree

algorithm, focusing on demographic features such as regional affiliation and socioeconomic status, resulting in an accuracy of 83.1%. Waheed et al. [10] employed the Artificial Neural Network (ANN) algorithm, tackling class imbalance and sequential data complexities, and attained an accuracy of 89%. Adnan et al. [11] implemented the Random Forest algorithm, incorporating demographic, assessment, and behavioral features, and achieved 91% accuracy. Finally, Esteban et al. [12] utilized the MLP (Multilayer Perceptron) algorithm, concentrating on assessment features, and obtained a remarkable accuracy of 93.9%. These studies demonstrated the diverse methodologies and high accuracy of machine learning in EDM.

EML has emerged as a crucial field within EDM. It aims to demystify and interpret the decision-making processes of machine learning models, especially black box models, for user comprehension. EML focuses on analyzing the underlying mechanisms of these models, making them more transparent and trustworthy. Its applications span various domains, including medical and healthcare data analysis [13], industrial data analysis [14], and smart city solutions [15].

One prominent EML method is the SHAP [16], which is based on the theoretical framework of Game Theory [2]. SHAP calculates the contribution of different features to the model's output by considering all possible combinations of features and their respective contributions. It assigns each feature a shapely value, also called SHAP value, allowing for a clear interpretation of the model's decision-making process.

Mu et al. [17] utilized SHAP values to elucidate machine learning models, such as Linear Regression and XGBoost, within an educational program that provides self-paced learning software for reading, writing, and mathematics. Applying SHAP values helped identify students struggling with academic tasks, as indicated by repeated failures. Alexandra and Costin [18] used another EML technique, Local Interpretable Model-agnostic Explanations (LIME), to interpret and explain the predictions of various ensemble tree-based models for predicting students' performance.

Despite the growing interest in EML, their application in educational research remains limited. Furthermore, the literature reveals that few studies systematically analyze how to use EML to explain model interpretability at both the global and individual levels. Our study aims to fill this research gap by using SHAP to explain model interpretability comprehensively. By offering a thorough understanding of the factors that influence student achievement, our study has the potential to contribute to the development of more effective individualized educational interventions.

In conclusion, EDM has evolved significantly, leveraging advanced analytical methods to interpret educational data. In EDM, incorporating interpretability approaches, such as SHAP, is increasingly essential for informed decision-making and targeted educational interventions.

III. Methodology

A. Dataset

Our study utilized the OULAD dataset [1]. The dataset includes various educational data, such as course information, student demographics, and interactions with the online learning platform. To ensure privacy, courses are represented by anonymized codes. We randomly selected the course denoted by the code CCC as our data source.

B. Implementation Tools

Python was the primary programming language used in this study, supported by various libraries that enhanced our data processing and analysis capabilities. We employed Scikit-learn [19] for machine learning algorithms, NumPy for numerical computations, Pandas for data manipulation, SciPy for additional functionality, Matplotlib for data visualization, and Imblearn for addressing class imbalances. The EML technique of SHAP [16] was implemented using the SHAP library in Python.

C. Research Design

We conducted the following sequential steps to compare the performance of different models and interpret the predictive results by EML.

1) Step 1. Data Preprocessing and Normalization

In this initial step, we preprocessed the OULAD dataset by excluding records with missing or incomplete data, ensuring our models' quality and reliability. After preprocessing, the dataset was reduced to 4,434 student records. We utilized the Min-max scaler to normalize the data and scale the features to a certain range.

2) Step 2. Feature Selection

Feature selection is a pivotal step in EDM [20]. Initially, we selected 22 features relevant to the target variable, including demographics (e.g., education level, age), assessment scores from eight assessments, and behavioral features (e.g., clicks and attempts made across eight assessments). Using the Chi-square test [21], we excluded features with low Chi-square scores, including gender, age, total credits studied, disability status, and the number of attempts made across assessments. These features were less relevant and had a low contribution to the model's predictive power. As a result, the 17 features included in Table I were identified as the most relevant predictors of student performance [22].

Demographic: The first feature measures the highest educational attainment before course enrollment, scaled from 0 for no formal education to 1 for a postgraduate degree.

Assessment: Features 2 to 9, representing assessment scores, were initially on a 0 to 100 scale, normalized to 0 to 1.

Behavioral: Features 10 to 17 show the sum of clicks across assessments 1 to 8, initially ranging from 0 to N, normalized to 0 to 1.

Target variable: The final result in the course is used as the model's predictive target. This variable is binary, classifying students into two categories: 1 (Pass) or 0 (Fail).

TABLE I. Description of Features Data and Target Variable

ID	Type	Description	Value	Description
1	Demographic Data	Educational level at course start	[0, 0.25, 0.5, 0.75, 1]	0: None 0.25: Below A-level 0.5: A-level 0.75: Higher Education 1: Postgraduate
2–9	Assessment Data	Scores across eight assessments	[0 - 1]	Normalized from original scores of 0-100 to [0 – 1]
10–17	Behavioral Data	Clicks across eight assessments	[0 - 1]	Normalized from original count 0 to N to [0 – 1]
18	Target Variable	The final grade in the course	[0, 1]	0: Fail 1: Pass

3) Step 3. Predictive Performance Comparison

To address the imbalance in the final grade, we implemented the Synthetic Minority Over-sampling Technique (SMOTE) [23]. This study optimized hyperparameters [24] using a randomized search [25] and employed six machine learning algorithms: Random Forest, Extra Trees, CatBoost, XGBoost, Naive Bayes, and K-Nearest Neighbors (KNN). We employed 10-fold cross-validation to train and evaluate the machine learning models. The model performance was assessed using accuracy, recall, precision, and F1-Score.

4) Step 4. Interpret the Predictive Results by SHAP

We used the EML technique of SHAP to interpret the predictive results at both global and individual levels.

For global-level interpretability, we employed summary plots to provide an overview of feature impacts, bar plots to quantify the average effect of each feature, and dependence plots to illustrate the relationship between specific features and the model's predictions.

For individual-level interpretability, we utilized force plots to understand the most influencing feature for the individual student, decision plots to trace the cumulative impact of features from the base value to the final output, and waterfall plots to provide a step-by-step breakdown of the prediction process.

IV. Results

A. Results of Prediction Using Various Algorithms

Table II presents a comparative analysis of diverse machine learning algorithms for performance prediction, delineating the algorithms' efficacy in accuracy, precision, recall, and the F1-Score.

Among the various algorithms compared, the Extra Trees algorithm demonstrated the best performance, with the highest accuracy of 95.37%, precision of 93.13%, recall of 98.01%, and F1-Score of 95.50%.

TABLE II. Performance Comparison of Various Algorithms

Algorithm	Accuracy	Precision	Recall	F1-Score
Random Forest	0.9531	**0.9327***	0.9769	0.9541
Extra Trees	**0.9537***	0.9313	**0.9801***	**0.9550***
CatBoost	0.9515	0.9295	0.9774	0.9527
XGBoost	0.9522	0.9291	0.9792	0.9534
Naive Bayes	0.9218	0.9117	0.9347	0.9229
KNN	0.9456	0.9256	0.9697	0.9469

* Asterisks (*) indicate the best values in each evaluation metric.

B. Analyzing the Global Level Interpretability with SHAP

We conducted the SHAP analysis on the best-performing Extra Trees algorithm model to interpret the predictive results. Various SHAP plots, including summary, bar, and dependence plots, were utilized to explain the model's behavior and the importance of different features. As an EML technique, SHAP provides valuable insights into the model's decision-making process. These plots allowed us to explore the intricate relationships between features and their impact on the global level of the model's predictions.

1) Summary Plots

The summary plot in Figure 1 shows a consolidated view of feature impacts on the model output. Each point on the plot represents a SHAP value for a feature and an instance in the dataset. Based on Game Theory [2], SHAP values quantify each feature's contribution to the model's prediction.

Notably, assessment scores impacted the model significantly, especially from assessments 8, 5, and 7. They were positioned higher on the y-axis, indicating a positive correlation with student success. The color coding provided additional information, where features with low values were shown in blue, and those with high values were in red.

Specifically, in analyzing the scores for assessments 8, 5, and 7, we observed a distinct pattern: higher scores, represented by red dots, were primarily situated on the positive side of the x-axis. This placement correlated with higher SHAP values, implying a beneficial influence on the model's prediction of student success. Conversely, blue dots depicting lower scores were predominantly clustered on the negative side of the x-axis. This location corresponded to reduced SHAP values, signifying a detrimental impact on the model's forecast, often suggesting a likelihood of student failure.

The trend was similar in the click data for assessments 7 and 8, where more clicks (red dots) positively influenced the model's passing prediction. In contrast, fewer clicks (blue dots) were associated with failure predictions.

Interestingly, not all click data exhibited the same impact. The distribution of red dots for clicks in assessments 1-3, on the left side of the x-axis, indicated an inverse relationship: more clicks correlated with lower SHAP values. As SHAP values reflect the direction and strength of a feature's impact on the model's prediction, this pattern implies that increased clicks on early assessments could link to forecasts of underperformance or potential failure in students.

This differentiation in click rates and SHAP values highlighted the need for early identification of such patterns to inform interventions. This can allow tailoring teaching strategies and support for at-risk students, underlining the importance of early detection systems to identify students requiring additional academic support.

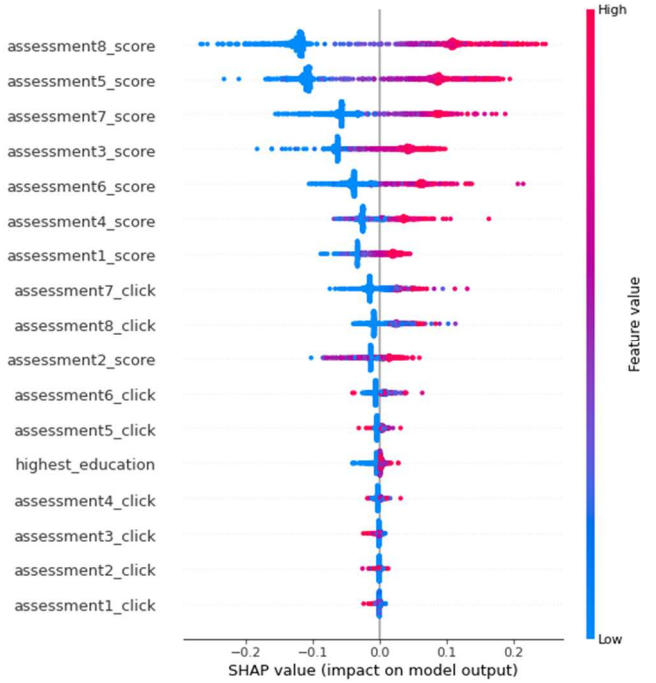

Fig. 1. SHAP Summary Plot for Overall Prediction

Overall, the summary plot was a valuable EML tool for understanding the key factors influencing model predictions. It provided insight into how assessment scores and click data correlate with student performance predictions.

2) *Bar Plots*

SHAP values measure each feature's contribution to the model's predictions by comparing the expected change in output when a feature is included versus when it is absent. This approach highlights the impact of each feature on individual predictions relative to the model's overall output.

In Figure 2, the bar plot visualizes the mean SHAP values, succinctly demonstrating the significance of each feature in predicting student success.

This plot revealed that assessment 8's score had the highest mean SHAP value, signifying its paramount average contribution to the model's predictions and marking it as a critical factor in determining student success. Conversely, assessment 1's number of clicks showed the smallest mean SHAP value, highlighting its minimal impact on the model's predictive capability.

Upon examining the bar plot, it became evident that assessment scores substantially impacted the prediction outcome more than click behaviors. This plot was instrumental in deepening the understanding of the model's predictive dynamics and identifying critical factors for further analysis, potentially guiding targeted educational strategies to enhance student outcomes.

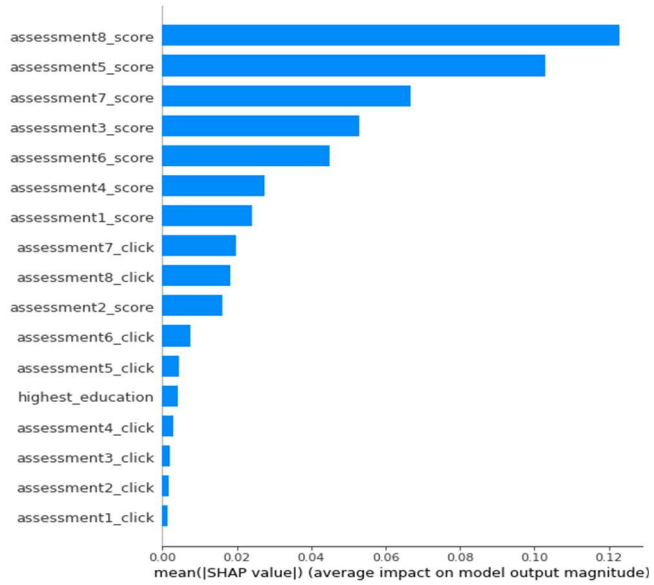

Fig. 2. SHAP Bar Plot for Overall Prediction

3) *Dependence Plots*

Figure 3 displays a SHAP dependence plot, illustrating the relationship between assessment 5's scores and clicks with the model's predictions.

The X-axis shows the scores, which ranged from 0 to 1. A clear upward trend in SHAP values was observed on the Y-axis, suggesting that higher scores were often associated with a better chance of passing.

The color gradient in the plot is essential for interpretation. Blue indicates lower-frequency clicks, while red represents higher engagement levels. We noticed that the red dots were mainly concentrated in the upper right corner of the plot, signifying higher click counts in assessment 5. These corresponded to higher assessment 5 scores and SHAP values, indicating a prediction of passing by the model.

Conversely, some blue dots were distinctly clustered in the lower left corner, demonstrating that lower click counts were more often associated with lower scores and SHAP values, leading the model to predict failure.

From this dependence plot, we can discern a clear correlation between the number of clicks and scores, which shows that more clicks are associated with higher scores and SHAP values in assessment 5.

From an educational perspective, early monitoring of student engagement with specific assessments becomes a viable method for identifying students potentially at risk of low performance.

These findings are invaluable in understanding how particular combinations of student performance metrics can indicate better academic outcomes.

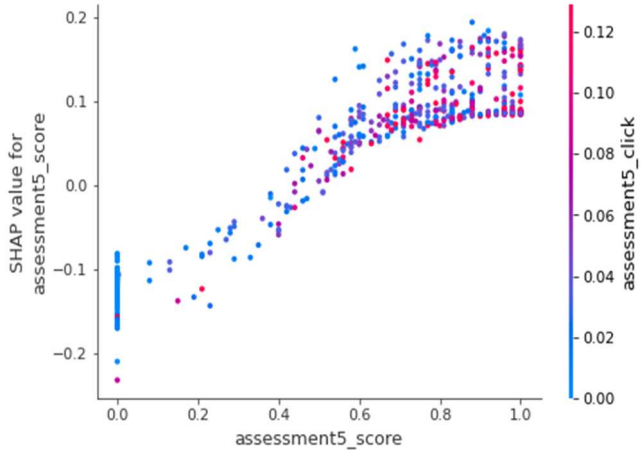

Fig. 3. SHAP Dependence Plot for Scores and Clicks of Assessment 5

Fig. 4. SHAP Force Plot of Student A with a Failed Grade Prediction

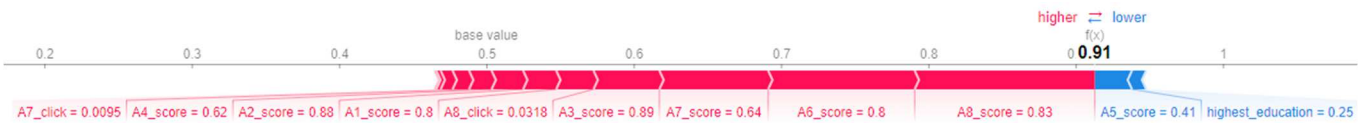

Fig. 5. SHAP Force Plot of Student B with a Passed Grade Prediction

C. Analyzing the Indivial Level Interpretability with SHAP

As an EML technique, SHAP provided different plots for understanding the factors contributing to each student's predicted outcome. Our study employed a range of SHAP plots to delve into individual-level interpretability within our predictive model. These included force plots, decision plots, and waterfall plots, offering unique insights into how various features influenced the model's predictions for individual students.

1) Force Plots

SHAP force plots revealed how various features contributed to the model's predictions, distinguishing factors that led to either a Fail (Student A) or a Pass (Student B).

A SHAP force plot visually represented each feature's impact on the prediction. Features that pushed the prediction towards failure were shown in blue, while those indicating a pass appeared in red. The plot arranged these features along a spectrum, with the position and width signifying their impact. A broader feature on the plot indicated a more substantial influence on the prediction.

In Figure 4, the force plot analyzed Student A's performance, where the predictive model showed a high probability of failure with an output value of fx(0.27). This value, much lower than the base value of 0.5, indicated a strong tendency towards failing.

The primary factors influencing this prediction were the blue features, mainly the low assessment scores ranging from 0 to 0.48 across six assessments. These scores highlighted a consistent pattern of underperformance in the course, suggesting difficulties in understanding or effectively engaging with the course material.

On the other hand, Figure 4 also shows some positive influences, marked as red features. Notably, a score of 0.51 in assessment 5 was slightly higher than the others, indicating some areas of relative proficiency or better understanding. Additionally, the student's engagement levels, as shown by their click data during assessments 5 and 8, were 0.1325 and 0.0266, respectively. This indicated a certain level of participation. Although these engagements were low, they suggested that the student tried to interact with the course material. Despite these positive aspects, more was needed to override the impact of the low assessment scores, which primarily drove the model's prediction toward failure. This case highlighted the importance of consistent performance in assessments for academic success.

Conversely, in Figure 5, the force plot provided a detailed illustration of Student B, where the predictive model indicated a high likelihood of success, as evidenced by an output value of fx(0.91). This value, significantly higher than the base value of 0.5, suggested a strong prediction in favor of passing.

Central to this prediction were the red features, predominantly high assessment scores. These scores were notably superior, with 0.83 in assessment 8, 0.8 in assessment 6, and consistently high marks in other assessments. Such scores indicated a good understanding of the subject matter and consistent academic performance. Additionally, the model considered the student's active engagement with the course material. This was reflected in the click data, such as 0.0318 clicks in assessment 8 and 0.0095 in assessment 7, which suggested a high level of interaction and involvement with the course content.

However, some blue features, like a lower score of 0.41 in assessment 5, indicated areas where Student B could improve. Furthermore, the student's initial educational background was noteworthy. The highest education feature score was 0.25, showing that Student B had an education level below A level when this student enrolled in this course. 'A level' is an advanced qualification typically required for university entry in the UK, so being below this level implies that Student B had less preparation for higher education. The model considered this a disadvantage. This factor negatively impacted the model's predictions, indicating that students with lower education levels at the start might encounter more significant challenges in the course.

Despite these factors, the overall prediction for Student B remained optimistic. The combination of high assessment scores and active engagement outweighed the potential negative impact of the lower score in assessment 5 and the initial educational background. This case exemplified how a student could overcome initial educational setbacks through consistent performance and engagement in academic activities.

Overall, the force plot visualizations for Students A and B provided insight into the factors influencing their academic performance predictions. Student A's lower assessment scores and moderate engagement levels pointed towards a failure prediction, underscoring the importance of consistent understanding and assessment performance. For Student B, high assessment scores and active engagement supported a successful prediction, illustrating the positive impact of consistent performance and engagement in academic outcomes. These visualizations allowed the identification of areas where educational support could be optimized, enhancing the potential for student success.

2) Decision Plots

The decision plots, analyzed from a bottom-up perspective, provided an in-depth view of the model's decision-making process for individual predictions. By tracing the path from the base value to the final output, we could see the cumulative impact of each feature. These plots showed how the interaction of different features led to the model's final prediction.

Regarding Student A, as depicted in Figure 6, the model leaned towards a failure prediction. This inference began with the clicks on assessment 1 at the bottom, which exerted a slight negative influence at a value of 0.022, slightly reducing the model's output value. Progressing upwards, we encountered the feature highest education, which failed to shift the prediction towards a pass despite its high value of 0.75 (Higher Education). This suggested that the highest education level was not a decisive factor in determining the likelihood of passing or failing Student A.

Further analysis showed that the clicks on assessments 7 and 3, denoted by values of 0.004 and 0.029, respectively, had a minor effect on the output value. This trend continued with scores and clicks on other assessments like the score of assessment 7 and the click count of assessments 2, 4, and 6. These features, with low values, contributed to a slight negative impact on the output value, resulting in the prediction of failure.

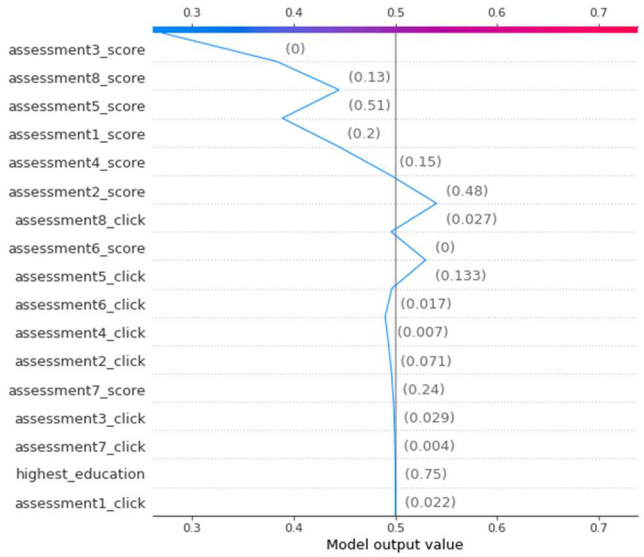

Fig. 6. SHAP Decision Plot of Student A with a Failed Grade Prediction

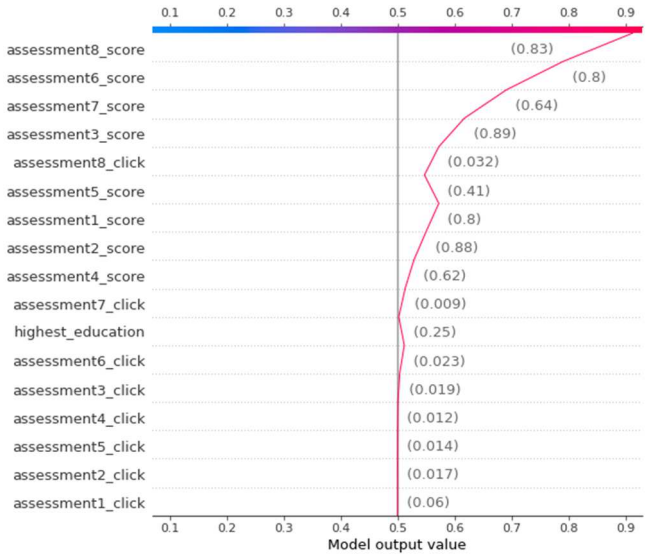

Fig. 7. SHAP Decision Plot of Student B with a Passed Grade Prediction

The click counts of assessment 5 at 0.133 and assessment 8 at 0.027 positively influenced the model's prediction, causing a slight rightward shift in the curve. The zero score for assessment 6 was crucial in predicting failure, notably lowering the output value. This effect was compounded by lower assessment scores 2, 4, and 1.

Although assessment 5 scored 0.51, impacting the model positively, it wasn't enough to counteract the trend towards failure. Finally, the value of 0.13 for the score of assessment 8 further strengthened the model's inclination towards predicting failure. Notably, the score of assessment 3, being zero, indicated poor performance and solidly drove the model to predict a fail outcome for Student A.

In the SHAP decision plot for Student B, shown in Figure 7, the model's prediction for a passing grade was initially influenced by the click counts from assessments 1, 2, 5, 4, 3, and 6 with values of 0.06, 0.017, 0.014, 0.012, 0.019, and 0.023 respectively, consistently exerted a slight positive impact on the prediction. This pattern suggests that while these clicks contributed to the prediction, they were not the most decisive factors in determining the model's outcome.

As we moved upward, the feature of the highest education emerged, valued at 0.25. This subtly shifted the model's prediction towards a negative aspect, revealing that Student B had an education level below A level at the start of the course, a factor considered disadvantageous by the model.

The influence became more pronounced with the click of assessment 7, valued at 0.009, and the scores for assessments 4, 2, and 1, valued at 0.62, 0.88, and 0.8, respectively. These values indicated robust performances, significantly propelling the model's prediction toward a pass. In contrast, the score of assessment 5, at a value of 0.41, represented a lower score, pushing the model's prediction towards the negative side, as indicated by a leftward shift in the curve.

Approaching the upper section of the plot, the click count of assessment 8, valued at 0.032, and the scores of assessments 3, 7, and 6, with values of 0.89, 0.64, and 0.8, respectively, emerged as critical factors. Notably, the score of assessment 8, at 0.83, strongly affirmed the model's pass prediction, showcasing a solid performance that solidified the positive outcome for Student B.

Overall, the decision plots clearly illustrated the model's decision-making process. Starting from a base value, they showed the impact of each feature, mirroring sequential moves in a cooperative game, aligning with Game Theory principles [2]. We explored two cases through these plots, uncovering how various features affected the model's output. These insights revealed that academic performance and student engagement significantly shaped predictions, influencing success or failure likelihood. This approach enhanced our understanding of the factors driving educational outcomes and refined our model interpretation.

3) Waterfall Plots

SHAP waterfall plots offered a detailed insight into how each feature impacted the final prediction from a baseline value, providing a more detailed step-by-step breakdown of the prediction process. Similar to decision plots, waterfall plots showed additional details, such as the actual feature values on the Y-axis, and used colored arrows to indicate SHAP values. This visual representation made the influence of each feature more apparent. Furthermore, waterfall plots hide the features with minimal impact on the model's prediction, enabling a focus on the most significant features.

Figure 8 illustrates the waterfall plot for Student A, predicting a failed grade. The plot commenced from the model's baseline value, with each row indicating a feature's positive (red) or negative (blue) contribution. This representation effectively demonstrated how each feature moved the value from the model's expected output on the dataset to the predicted output value, resembling the additive nature of SHAP values in cooperative games.

Notably, the plot marked the score of assessment 5 and clicks of assessments 8 and 5 in red, with SHAP values of +0.06, +0.04, and +0.03, respectively. This clarity allowed for a clear interpretation of how these three features influenced the model's prediction positively, while other features in blue pushed it towards a negative outcome. Such insights have the potential to guide educators in focusing on the weaknesses indicated by the blue features, especially those with the most comprehensive spans and highest negative SHAP values, for targeted interventions to enhance students' learning performance.

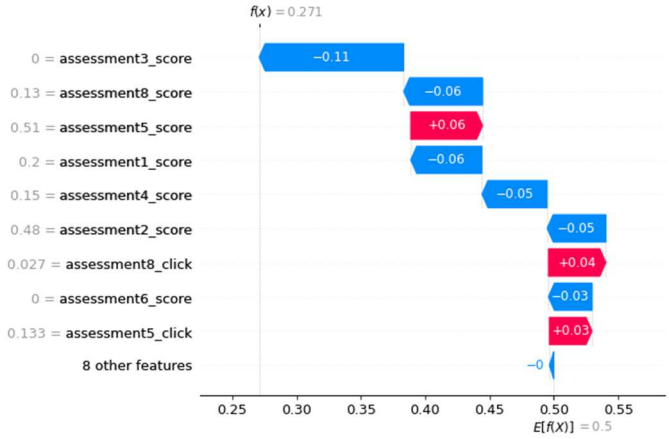

Fig. 8. SHAP Waterfall Plot of Student A with a Failed Grade Prediction

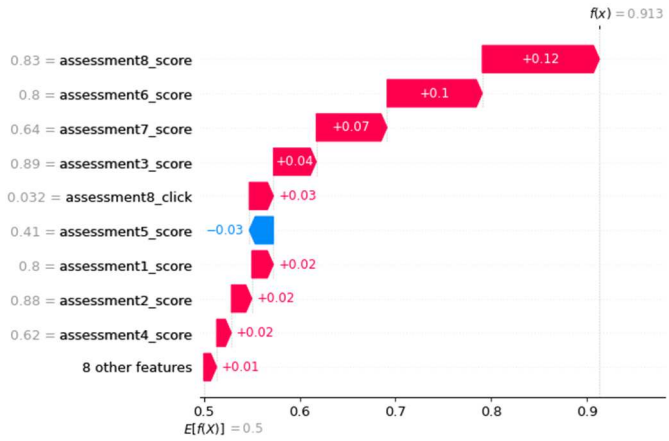

Fig. 9. SHAP Waterfall Plot of Student B with a Passed Grade Prediction

Similarly, Figure 9 presents the waterfall plot for Student B, predicting a passing grade. It followed the same structure, marking the features pushing the prediction positively in red and those influencing it negatively in blue. Educators could use the insights from the blue sections to offer personalized interventions, helping students enhance their performance.

Overall, the advantage of waterfall plots lies in their precise representation of how each feature influenced a model's final prediction. They excelled in displaying both the actual values of features and their SHAP values, offering a step-by-step breakdown of the prediction process. This clarity and specificity make waterfall plots useful for educators to understand and act on the predictive models' insights.

V. Discussion

A. *RQ1) Which machine learning algorithm performs best in predicting student academic outcomes?*

This study analyzed different machine learning algorithms and found that the Extra Trees algorithm performed best. To benchmark against prior research, we evaluated our optimal algorithm (Extra Trees) against results from earlier studies using the OULAD dataset with accuracy as the evaluation metric. Adnan et al. [11] achieved 91% accuracy with the Random Forest, incorporating demographic, assessment, and behavioral features. Esteban et al. [12] employed MLP and focused on assessment features, attaining 93.9% accuracy.

Our findings indicated that the Extra Trees algorithm, with 95.37% accuracy, outperformed previous studies, highlighting the effectiveness of our approach and the impact of SMOTE and hyperparameter optimization on improving accuracy.

B. *RQ2) How does SHAP contribute to understanding predictive outcomes at the global level?*

SHAP made markable contributions to understanding predictive outcomes at the global level.

Summary plots provided an overview of feature impacts on the model output. They clearly showed which features had the most significant positive or negative influence on predictions. In our study, assessment scores, especially from assessments 8, 5, and 7, greatly impacted the model's student success or failure predictions. The color coding in summary plots added an extra dimension, revealing how feature values correlated with their impact.

Bar plots quantified the average impact of each feature on the model's output using mean SHAP values. They allowed easy comparison of features' relative importance, similar to ranking players in a cooperative game based on their SHAP values. In our case, assessment 8 scores had the highest mean SHAP value, identifying it as the most critical factor in determining student outcomes. Bar plots focused on the most influential features for further analysis and interventions.

Dependence plots illustrated the relationship between specific features and the model's predictions. They provided insights into how feature values and their interactions affected outcomes, drawing parallels to the interaction of players in a cooperative game. Our dependence plot showed a clear correlation between the number of clicks and scores in assessment 5, with higher engagement linked to better performance. Such plots could guide early monitoring of student engagement to identify those at risk.

Overall, SHAP visualizations at the global level offered a powerful tool for interpreting black box machine learning models in educational contexts. They provided a transparent, visual way to understand which features drove predictions and how they interacted. This understanding was crucial for designing targeted interventions and support strategies to improve student outcomes. The insights from summary, bar, and dependence plots could inform data-driven decision-making and help optimize educational practices.

C. *RQ3) How does SHAP assist in understanding predictive outcomes at the individual level?*

It is important to note that while global interpretations are valuable, they may only capture some of the nuances of individual student experiences. Therefore, combining global and individual-level interpretations is essential for comprehensively understanding predictive models. In our study, SHAP played a crucial role in understanding predictive outcomes at the individual level through various plots.

Force plots provided a detailed view of how different features contributed to the model's prediction for an individual student. They visually represented each feature's influence, distinguishing factors that led to a fail or pass prediction. This information could help identify areas where students need additional support. In our study, the force plots for Students A and B revealed the significant impact of specific assessment scores on individual predictions. For instance, Student A's low score in assessment 3 was the most important negative feature, leading to a fail prediction. This insight allows teachers to

target interventions specifically for assessment 3, potentially enhancing Student A's understanding and performance.

Decision plots provided an in-depth view of the model's decision-making for individual predictions, showing the cumulative impact of each feature from base value to final output. Our analysis of Students A and B highlighted how academic performance and engagement shaped predictions, helping to guide targeted interventions by identifying the most influential factors for each student.

Waterfall plots provided a detailed breakdown of the prediction process, displaying actual feature values and SHAP values. They showed how each feature moved the prediction from the baseline to the final output. In our study, the waterfall plots for Students A and B allowed a clear interpretation of how specific features influenced the model's predictions, guiding educators to focus on areas needing improvement.

These SHAP visualizations at the individual level offered valuable insights into the unique factors affecting each student's predicted outcomes. They provided a granular understanding of how the model arrived at its predictions, enabling personalized interventions and support.

VI. Conclusion

This study explored the application of machine learning algorithms to predict student performance and interpret the predictive results using the EML technique of SHAP in EDM. The Extra Trees algorithm outperformed other algorithms. It underscored the effectiveness of our approach, which included feature selection, enhancement methods like SMOTE, and hyperparameter optimization. These approaches played a pivotal role in improving predictive accuracy.

SHAP visualizations, including summary plots, bar plots, and dependence plots, provided valuable insights into the predictive model's global interpretability. These plots revealed the impact of features such as assessment scores and engagement metrics on the model's predictions.

Furthermore, force, decision, and waterfall plots showed SHAP's ability to interpret individual-level predictions. These visualizations offered a granular perspective on the factors influencing a student's predicted success or failure, enabling personalized interventions and support.

Our findings highlighted the potential of machine learning and SHAP in EDM, providing a foundation for data-informed decision-making and personalized educational strategies. This study encourages further exploration of SHAP in diverse EDM scenarios to understand better the factors influencing student success and develop targeted intervention strategies.

However, our study has limitations. First, the OULAD dataset is specific to the online learning environment, which may limit its generalizability to other contexts. Second, while SHAP was used, other EML methods like LIME could be explored in future research to compare their effectiveness in explaining predictions.

References

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

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

[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] W. Hämäläinen and M. Vinni, 'Classifiers for educational data mining', *Handbook of Educational Data Mining, Chapman & Hall/CRC Data Mining and Knowledge Discovery Series*, pp. 57–71, 2011.

[6] F. Araque, C. Roldán, and A. Salguero, 'Factors influencing university drop out rates', *Computers & Education*, vol. 53, no. 3, pp. 563–574, 2009.

[7] W.-C. Choi, C.-T. Lam, and A. J. Mendes, 'How Various Educational Features Influence Programming Performance in Primary School Education', in *2024 IEEE Global Engineering Education Conference (EDUCON)*, Kos Island, Greece: IEEE, May 2024, pp. 1–8. doi: 10.1109/EDUCON60312.2024.10578608.

[8] 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.

[9] 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.

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

[11] 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.

[12] 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.

[13] M. S. Kamal, A. Northcote, L. Chowdhury, N. Dey, R. G. Crespo, and E. Herrera-Viedma, 'Alzheimer's patient analysis using image and gene expression data and explainable-AI to present associated genes', *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–7, 2021.

[14] 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.

[15] O. Embarak, 'Explainable artificial intelligence for services exchange in smart cities', *Explainable Artificial Intelligence for Smart Cities*, pp. 13–30, 2021.

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

[17] T. Mu, A. Jetten, and E. Brunskill, 'Towards Suggesting Actionable Interventions for Wheel-Spinning Students.', *International Educational Data Mining Society*, 2020.

[18] A. Vultureanu-Albişi and C. Bădică, 'Improving students' performance by interpretable explanations using ensemble tree-based approaches', in *2021 IEEE 15th International Symposium on Applied Computational Intelligence and Informatics (SACI)*, IEEE, 2021, pp. 215–220.

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

[20] 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.

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

[22] W.-C. Choi, C.-T. Lam, and A. J. Mendes, 'Predicting Learning Performance with Explainable Machine Learning Algorithms', in *2024 IEEE Frontiers in Education Conference (FIE)*, 2024.

[23] 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.

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

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