Explore the Explanation and Consistency of Explainable AI in the LBLS Data Set

Tiffany T.Y Hsu, Owen H.T. Lu

International College of Innovation, National Chengchi University, Taiwan

Abstract

Learning Analytics (LA) is a field focusing on analyzing educational data, utilizing machine learning. One of the most discussed topics is at-risk student prediction. However, the application of these methods for predicting students' academic behaviors has faced criticism due to concerns about context insensitivity, potentially leading to prejudice and discrimination against students. While some methods in explainable AI (xAI) have been proposed to address these issues, there remains uncertainty regarding the consistency of their results. In response, we incorporate two popular explainable AI (xAI) methods SHAP (Shapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations), to interpret the predicting models. These methods attribute the output of these models to individual features, providing a clearer understanding of how each features contributes to the overall prediction. This approach is exemplified in the LBLS467 dataset, which includes data on 467 students' academic performance and learning behaviors in computer programming courses, encompassing a range of metrics from programming behavior to self-regulated learning and language learning strategies. Concerning the consistency of interpretations derived from SHAP and LIME, analysis via Kendall's tau coefficients reveals a moderate alignment in their feature weight rankings. Additionally, this alignment is substantiated by a highly significant confidence level, affirming that the observed alignment is not a mere coincidence.

Keywords

Learning Analytics, Explainable AI, SHAP, LIME

1. Introduction

Learning Analytics (LA) is a research field centered on measuring, collecting, analyzing, and reporting data about learners and their contexts [1]. Within this field, predicting student academic achievement is a foundational and significant topic [2]. Risk student prediction involves identifying students at risk of academic failure using data-driven insights and has been used to enhance web-based learning environments [3]. This process is not about labelling or categorizing students, rather, it aims to foresee students' performance in classes in advance. This foresight enables educators to offer timely assistance and intervention, tailored to each student's needs, thereby enhancing their academic outcomes and experiences.

Machine learning is often criticized for being overly generalized, and overlooking the context of the individual. Reflecting on the limitations of generalizations in understanding human behavior, anthropologist Clifford Geertz suggests that theories and generalizations inevitably lack deep and contextual understanding of human thought. 'Theoretical disquisitions stand far from the immediacies of social life,' he notes. 'Any generalization or theory constructed in the absence of deep understanding, not grounded in the concrete and particular, is vacuous.' [4]. The approach of risk student prediction has also faced similar criticism of over-generalizing. The fact that machine learning models do not provide a causal effects between features and prediction is overlooking the individuality of students. In machine learning predictions, we are confronted solely with the dichotomous outcomes: students being classified as either 'at risk' or 'not at risk.' While the purpose of such predictions is not to categorize students, the absence of interpretability in these outcomes can inadvertently result in failure to recognize individuality and risk of discrimination and stereotyping [5]. Explainable Artificial Intelligence (xAI) appears to be a

LAK-WS 2024: Joint Proceedings of LAK 2024 Workshops, March 18–19, Kyoto, Japan

solution to address these concerns, helping educators understand the differences among individual students. xAI refers to methods to explain and interpret predictions made by machine learning models [6]. Recently, artificial intelligence has been integrated into many areas of society. At the same time, debates surrounding AI, particularly in the context of ethics remain active. One of the most popular topics is transparency. In discussions about transparency, besides disclosing training data and sources, another prevalent approach is the application of xAI to render the decision-making processes transparent. When a model's decision process is transparent, it becomes simpler to monitor and assess its accuracy, thereby enhancing the model's accountability. Moreover, the comprehensible predictions offered by interpretable models play a vital role in fostering people's acceptance and trust in the decisions made by the model [7].In this study, we will answer two research questions:

RQ1: What are the successful factors in the LBLS dataset explored by SHAP and LIME?
RQ2: How consistent are SHAP and LIME in interpreting a student's learning performance?

2. Literature Review

The global community has developed an extensive variety of xAI approaches, which have been applied across various domains to interpret a wide range of machine learning models, including several complex models that were previously considered too intricate to interpret [8]. These advancements in xAI have enabled a deeper understanding of machine learning outputs, enhancing transparency and trust, especially in critical sectors. In line with these developments, a systematic review of xAI applications reveals a concentrated focus in specific sectors, notably healthcare, industry, and transportation [9]. As for the field of education, despite the relatively lower number of scholarly articles compared to other domains, the application of xAI has been noted in the review. It is noteworthy that 27% of xAI applications in these articles are utilized for decision support, which is the highest proportion of application in this context. Therefore, employing xAI as a tool for decision support in predicting whether students are at risk is justified.

The application of xAI in education manifests primarily in two aspects: data usage and stakeholder engagement [10]. Application in data usage enables the explanatory models to improve prediction models after identifying the characteristics of student success in the classroom. In terms of stakeholder engagement, it allows teachers to adjust their teaching methods based on the results provided by the explanations.

Reflecting on previous studies, there was research focused on the automatic generation of explanations in virtual learning environments. In [11], a tool was developed to generate multi-modal explanations regarding predictions of whether a student will pass or fail. The study compared the accuracy of various classifiers. Under the conditions of most models demonstrated high accuracy, it opts for simpler models including J48, Rep-Tree, and RandomTree over complex ones like SVM to achieve a balance between accuracy and interpretability. [12] also indicates that when models achieve high predictive accuracy, simpler models may yield higher quality explanations. Therefore, this study follows this direction by comparing the predictive accuracy of multiple models and selecting a simpler model for explanation under the premise of high accuracy.

According to [6], the most prominent repositories on GitHub in 2022 for xAI, as measured by the number of stars, were slundberg/shap (Shapley Additive exPlanations) and marcotcr/lime (Local Interpretable Model-agnostic Explanations). SHAP operates on game theory principles, attributing a machine learning model's output to the contributions of individual features [13]. Conversely, LIME elucidates the predictions of classifiers or regressors faithfully by locally approximating them with an interpretable model [8]. Both methods are adept at explaining machine learning models, regardless of their complexity. Given the active community engagement on GitHub, the high level of attention these methods have garnered, and their open-source status, this study will incorporate both SHAP and LIME. Utilizing these approaches, we aim

to pinpoint key features that determine the classification of individual students as at-risk. We will compare the outcomes from each method and assess the consistency between the two.

3. Methods

3.1. LBLS Dataset

LBLS467 is a dataset that collects data on 467 students' academic performance and learning behaviors in computer programming courses. It encompasses students' programming editing behaviors, questionnaire survey results on Self-regulated Learning (SRL) and the Strategy Inventory for Language Learning (SILL). This dataset includes a total of 208 features, covering a wide range of learning behaviors and performance indicators. The dataset is utilized to propose a series of challenging suggestions for the LBLS dataset and was used in a data challenge workshop organized by the Society for Learning Analytics Research (SoLAR) [14] [15]. 'At-risk' has diverse definition, in this study, we defined 'At-risk' students are those who fail or are on the verge of failing the course in this study. Specifically, risk students are those whose performance is comparatively worse than at least 75% of the students in their class."

3.2. Feature Extraction and Classification

In this study, we employ Principal Component Analysis (PCA) as our primary tool for feature extraction. PCA, a common preprocessing step for machine learning algorithms [16], is followed by the application of three different models, ranging from the most explainable to the least: Decision Tree, Logistic Regression, and Support Vector Machine (SVM). We create a graph to demonstrate how model accuracy relates to the number of PCA components, aiming to find the most accurate model for a given component count. The most accurate model is then further analyzed using SHAP and LIME.

$$\text{Accuracy} = \frac{(\text{TP} + \text{TN})}{(\text{TP} + \text{FP} + \text{TN} + \text{FN})} \tag{1}$$

- TP (True Positives): The number of correct predictions that an instance is positive.
- TN (True Negatives): The number of correct predictions that an instance is negative.
- FP (False Positives): The number of incorrect predictions that an instance is positive (actually negative).
- FN (False Negatives): The number of incorrect predictions that an instance is negative (actually positive).

3.3. SHAP and LIME

SHAP is an xAI method grounded in game theory, designed to interpret the predictions of complex machine learning models. It employs the Shapley value to calculate the contribution of each feature to the model's output. This approach facilitates a comprehensive understanding of how different features influence the model's predictions. The weights of the features are derived from the following [13]:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F|-|S|-1)!}{|F|!} [f_{S \cup i}(x_{S \cup i}) - f_S(x_S)] \tag{2}$$

- F : All features.
- $|F|$: The total number of features.
- φ_i : The SHAP value of feature i .
- S : A subset of all features set F excluding the feature i.
- $|S|$: The size of subset S.

- $f_{S\cup i}(x_{S\cup i})$: The prediction of model f when the feature set S includes the feature i.
- $f_S(x_S)$: The prediction of the model f with only the feature set S.

SHAP provides a method to quantify the contribution to the change in prediction when feature i is added to the model for every possible feature set S. The idea of LIME, on the other hand, is to approximate the behavior of a complex model near the prediction of a specific instance using a simpler model. The formula of LIME is as follows [8]:

$$\xi(x) = \underset{g\in G}{\operatorname{argmin}}\, L(f, g, \pi_x) + \Omega(g) \tag{3}$$

- g : A simple model used to approximate the behavior of the complex model f near the instance x.
- G : The set of all possible simple models.
- π_x : A weighting function that assigns higher weights to points closer to the instance x.
- $\Omega(g)$: A complexity measure that penalizes the model g.

LIME begins by selecting a specific instance x (in this case, an individual student), already predicted by a complex model. It generates a series of perturbed samples around this instance to explore the model's behavior locally. To approximate the behavior of the complex model in this localized region, a simpler model, such as linear regression, is employed. The key objective is to assess the alignment between the outputs of this simpler model, denoted as g, and the original complex model, denoted as f , within the local context. This process is represented in the formulation by minimizing the loss function between g and f, complemented by the minimization of g's complexity measure.

In this study, the Logistic Regression model was trained using data transformed through PCA. Consequently, to maintain consistency with the training data, the data samples generated by LIME need to be transformed into the same dimensional space. A wrapper function is implemented to facilitate this process, transforming the LIME-generated data via PCA to ensure that the data is in the appropriate form for the trained model to process effectively. And the same wrapper function has been applied on SHAP.

3.4 Consistency Evaluation

SHAP and LIME operate on distinct principles to determine the contribution of each feature to the outcome. The critical question lies in the extent of the differences in the explanations derived from these two methods. To evaluate their consistency, our approach involved identifying features that show statistical correlation with the predicted results, as assessed by Spearman's correlation with a significance threshold set at $\alpha = 0.05$. This threshold was chosen to discern features significantly correlated with the outcomes. The next step is to compare the ranks of contributions as provided by SHAP and LIME, utilizing the Kendall's tau for this comparative analysis.

The Kendall correlation coefficient measures the degree of similarity between two sets of rankings assigned to the same group of objects [18]. Firstly, we rank the selected features based on their influence on the prediction outcome. This process results in two sets of ranking data, each ordering the same set of features. For each pair of features, we examine their respective positions in both ranking sets and calculate their relative positions. Consequently, if a feature ranks higher than another in both sets, the pair is deemed 'consistent'; the opposite scenario indicates inconsistency. Once considers all pairs of features, calculating the difference between the number of consistent pairs and inconsistent pairs, divided by the total number of pairs. Following is the formula of Kendall correlation coefficient:

$$\tau = \frac{n_c - n_d}{\frac{1}{2}\times n\times(n-1)} \tag{4}$$

- n_c : The number of concordant pairs

- n_d : The number of discordant pairs.
- n : The sample sizes.

The value of this coefficient ranges from -1 and 1. A value approaching 1 indicates a high level of consistency in the rankings, while a value approaching -1 signifies a substantial degree of inconsistency [19].

4. Results and Discussion

4.1 Reply RQ1

As illustrated in Figure 1, the accuracy assessments demonstrate that all models achieved accuracy rates around 80%. Notably, both Logistic Regression and Decision Tree models showed remarkable performance. Logistic Regression achieved an 84.6% accuracy rate with 16 components, while the Decision Tree reached a same level of accuracy with 58 components.

The final decision to focus on Logistic Regression for in-depth analysis stems from a crucial observation. Under the premise of using PCA as a method for feature extraction, the Decision Tree model becomes less interpretable. Initially, the Decision Tree was a preferred choice due to its well-known ease of interpretability. However, it was crucial to assess whether its performance was sufficiently superior to warrant detailed explanation. Upon further analysis, it was found that its accuracy was comparable to that of the Logistic Regression model. Therefore, we decided to apply SHAP and LIME to the Logistic Regression model.

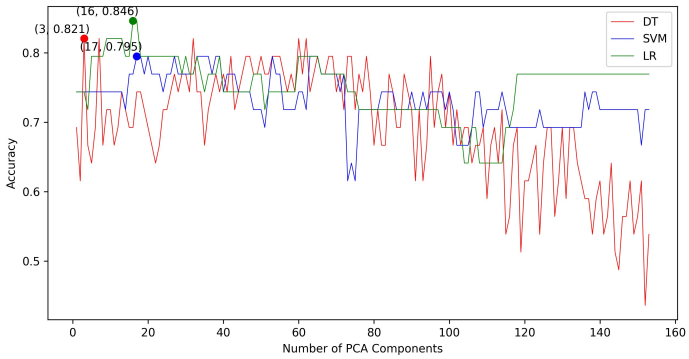

Figure 1: Accuracy vs. Number of PCA Components

In the results, we present the explanation of SHAP's prediction for individual instance in the form of a waterfall plot. This mode of presentation is very similar to the way data is represented in LIME results, which aids in our comparison of each instance.

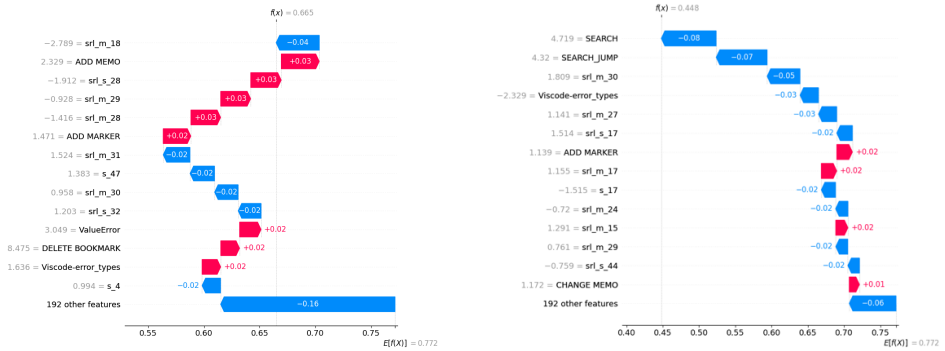

Figure 2: SHAP interpretation of student A, B (left to right)

This SHAP waterfall plot illustrates how feature contributions (red and blue bars) move the model prediction from a baseline value (the average output of the model) $E[f(x)]$ to the final prediction $f(x)$. Blue bars represent features that decrease the prediction probability, while red bars indicate those that increase it. The gray texts in front of the feature names are the value to each features.

In Figure 2, the model predicts Student A as at-risk with a probability value of 0.665, surpassing the threshold for risk. Key features like 'ADD_MEMO', 'srl_s_28', and 'srl_s_29'' positively influence this outcome. In contrast, 'srl_m_18' and 192 other features collectively decrease the prediction probability by about 0.16. Figure 2 shows Student B as not at-risk with a predictive value of 0.448, influenced by features like 'SEARCH', 'SEARCH_JUMP', and 'srl_m_30' which lower the risk probability.

LIME's plot in Figure 3 indicates Student A as at-risk with a 0.66 predictive probability, consistent with the number displayed on SHAP analysis. Influential features include 'LINK_CLICK', 'SEARCH_JUMP', and 'srl_s_3. Conversely, features like 'CLEAR_HW_MEMO, and 'OPEN_RECOMMENDATION' contribute to a lower risk prediction. Figure 3 predicts Student B as not at-risk at 0.55 probability, significantly influenced by 'RuntimeError' and 'HTTPError'.

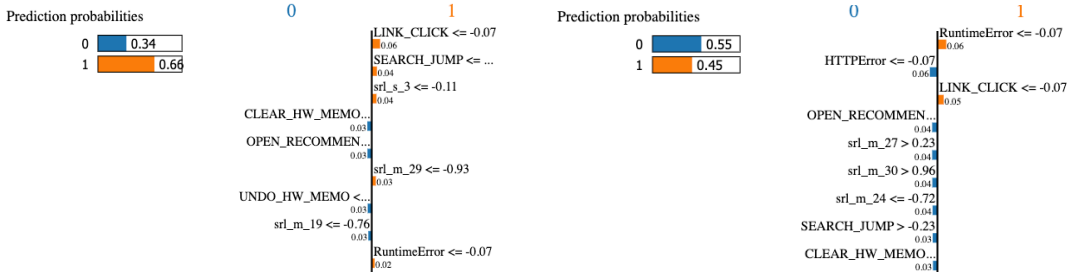

Figure 3: LIME interpretation of student A, B (left to right)

To identify the key features contributing to succeed in the LBLS dataset, we assess the contribution of features to the prediction. For SHAP analysis, we employed global explanation to find out the top five features with the highest contribution values. Since LIME lacks a global explanation mechanism, we aggregated the top five features with the most significant impact from each predictions. The five most influential features in the global explanation of SHAP were ADD_RECOMMENDATION, ADD_HW_MEMO, s_41, s_26, and TabError; whereas, for LIME, they were LINK_CLICK, HTTPError, ZeroDivisionError, RecursionError, and s_32.

4.2 Reply RQ2

We select 93 features that are statistically correlated to the result using Spearman's correlation coefficient. In the next step of the analysis, we will employ SHAP and LIME to evaluate the prediction concerning student A. Our focus will be on capturing the rankings of all 93 features. Following this, using Kendall's tau coefficient to assess the similarity between these rankings.

Table 1
Features that are statistically correlated to predicting result

Features	Description	ρ
CLOSE	Closed the book.	-0.11*
OPEN	Opened the book.	-0.12**
PAGE_JUMP	Jumped to a particular page.	0.17***
CODE_COPY	Number of times a student copy codes.	0.33***
CODE_EXECUTION	Number of times a student execute codes.	0.28***
Other 88 features		

*p < .05 **p < .01 ***p < .001

Each blue dot on the graph represents a unique feature. When a dot aligns with the diagonal, it signifies that SHAP and LIME assign the same ranking to that feature's weight. For the interpretative analysis of Student A and Student B, the Kendall's tau are 0.66 and 0.64, respectively, suggesting a moderate but noticeable positive correlation between the two datasets. This implies that an increase in one dataset's values is generally mirrored by an increase in the other, although the relationship is not exceptionally strong. The analysis yields remarkably low P-values for the weight rankings, all of which are below 0.001, reinforce the significance of this correlation.

The graph reveals a tendency for the features' weight rankings, as determined by both interpretation methods, to cluster near the diagonal, particularly those with higher (towards the start) and lower weights (towards the end). This pattern suggests a greater consistency in how both methods evaluate these features. Conversely, the rankings of features in the central region of the graph tend to be more dispersed.

In the two prediction points, SHAP and LIME show a moderate level of consistency in assessing feature importance, with a tendency for feature rankings to cluster near the diagonal line indicating higher consistency in evaluating the most and least important features. The dispersion of feature rankings in the central area of the graph suggests greater variability in interpreting features of medium importance. The low P-values enhance the credibility of the results, suggesting that the observed correlations are not random but reflect the underlying patterns in the data.

Table 2
Features that are statistically correlated to predicting result

Rank	Feature Weight Ranking by SHAP	Feature Weight Ranking by LIME
1	srl_s_9	srl_m_17
2	FileNotFoundError	srl_m_15
3	srl_m_3	PREV
4	srl_m_9	srl_s_23
5	ConversionError	s_10
And other 88 features		

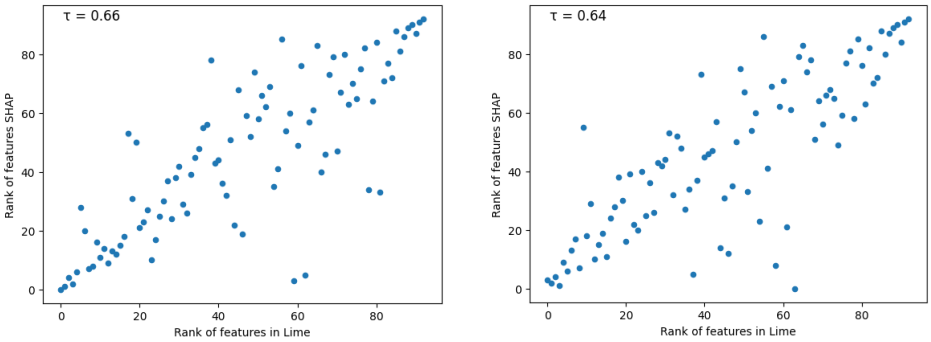

Figure 4: Kendall's Tau Rank Correlation of student A and B

Finally, we compared the feature weight rankings explained by SHAP and LIME for each prediction point pairwise, calculating the average of Kendall's tau and p-value. We obtained an average Kendall's tau of 0.623 and an average p-value of 0.000979. This suggests that there is also a moderate to strong correlation in the feature importance rankings between the two methods for each prediction point. In other words, the rankings of feature importance are relatively consistent between the two methods, and the p-value being far below 0.05 shows that the correlation in rankings between SHAP and LIME is statistically significant.

5. Conclusion

In this study, we emphasize the importance of xAI in preventing over-generalization of machine learning algorithms, especially in fields of learning analytics. We use PCA for feature extraction, comparing accuracies of multiple models, and selected one that is both simple to use and highly accurate. We then combine various statistical methods to check if SHAP and LIME explanations of feature weight rankings are consistent. The results show moderate consistency in SHAP and LIME rankings among 93 selected features related to prediction outcomes, with high confidence. In learning analytics, divergent results from xAI in predicting at-risk students can complicate strategy formulation for stakeholders. Our study has analyzed explanations for two students predicted with different labels. Future research could explore which explanation is more trustable when there is a lack of consistency, whether to sacrifice model accuracy for higher consistency, or to involve more human intuition in assessing the reasonableness of explanations. As for key feature identification for student learning performance and strategy formulation for adaptive development, it undoubtedly requires involvement from school teachers, educators, and psychologists.

Acknowledgments

This study is supported in part by the National Science and Technology Council of Taiwan under contract numbers NSTC 112-2410-H-004 -063 -.

References

[1] Siemens, G., & Baker, R. S. d. (2012). Learning analytics and educational data mining: towards communication and collaboration. In Proceedings of the 2nd international conference on learning analytics and knowledge (pp. 252–254).

[2] Ak¸capınar, G., Altun, A., & A¸skar, P. (2019). Using learning analytics to develop early-warning system for at-risk students. International Journal of Educational Technology in Higher Education, 16 (1), 1–20.

[3] Romero, C., & Ventura, S. (2010). Educational data mining: a review of the state of the art. IEEE Transactions on Systems, Man, and Cybernetics, Part C (applications and reviews), 40 (6), 601–618.

[4] Birhane, A. (2021). Algorithmic injustice: a relational ethics approach. Patterns, 2 (2).

[5] Scholes, V. (2016). The ethics of using learning analytics to categorize students on risk. Educational Technology Research and Development, 64 (5), 939–955.

[6] Holzinger, A., Saranti, A., Molnar, C., Biecek, P., & Samek, W. (2020). Explainable ai methods-a brief overview. In International workshop on extending explainable ai beyond deep models and classifiers (pp. 13–38).

[7] Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature machine intelligence, 1 (5), 206–215.

[8] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). " why should i trust you?" explaining the predictions of any classifier. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 1135–1144).

[9] Islam, M. R., Ahmed, M. U., Barua, S., & Begum, S. (2022). A systematic review of explainable artificial intelligence in terms of different application domains and tasks. Applied Sciences, 12 (3), 1353.

[10] De Laet, T., Millecamp, M., Broos, T., De Croon, R., Verbert, K., & Duorado, R. (2020). Explainablelearning analytics: challenges and opportunities. In Companion proceedings of the 10th internationalconference on learning analytics & knowledge lak20 society for learning analytics research (solar)(pp. 500–510).

[11] Alonso, J. M., & Bugar´ın, A. (2019). Expliclas: automatic generation of explanations in natural language for weka classifiers. In 2019 ieee international conference on fuzzy systems (fuzz-ieee) (pp.1–6).

[12] Liu, B., & Udell, M. (2020). Impact of accuracy on model interpretations. arXiv preprint arXiv:2011.09903 .
[13] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in neural information processing systems, 30
[14] Lu, O. H., Huang, A. Y., Flang, B., Ogata, H., & Yang, S. J. (2022). A quality data set for data challenge: Featuring 160 students' learning behaviors and learning strategies in a programming course. In 2022 30th International Conference on Computers in Education. ICCE.
[15] Flanagan, B., Ogata, H. (2018). Learning Analytics Platform in Higher Education in Japan, Knowledge Management & E-Learning (KM&EL), Vol.10, No.4, pp.469-484.
[16] Wold, S., Esbensen, K., & Geladi, P. (1987). Principal component analysis. Chemometrics and intelligent laboratory systems, 2 (1-3), 37–52.
[17] DW, G. D. A. (2019). Darpa's explainable artificial intelligence program. AI Mag, 40 (2), 44
[18] Kendall, M. G. (1938). A new measure of rank correlation. Biometrika, 30 (1/2), 81 93.
[19] Abdi, H. (2007). The kendall rank correlation coefficient. Encyclopedia of Measurement and Statistics. Sage, Thousand Oaks, CA, 508–510.