Trustworthy and Explainable AI for Learning Analytics

Min-Jia Li [1], Shun-Ting Li[1], Albert C. M. Yang[2], Anna Y.Q. Huang[1] and Stephen J.H. Yang [1*]

[1] *Department of Computer Science and Information Engineering, National Central University, Taiwan*
[2] *Department of Computer Science and Engineering, National Chung Hsing University, Taiwan*
**Corresponding author*

Abstract

In recent years, there has been a surge of interest in combining artificial intelligence (AI) with education to enhance learning experiences. However, one major concern is the lack of transparency in AI models, which hinders our ability to understand their decision-making processes and establish trust in their outcomes. This study aims to address these challenges by focusing on the implications of explainable and trustworthy AI in education. The primary objective of this research is to improve trust and acceptance of AI systems in education by providing comprehensive explanations for model predictions. By doing so, it seeks to equip stakeholders with a better understanding of the decision-making process and increase their confidence in the outcomes. Additionally, the study highlights the importance of evaluation metrics in assessing the quality and effectiveness of explanations generated by explanation AI models. These metrics serve as vital tools for ensuring reliable system performance and upholding the fundamental principles necessary for building trustworthy AI.

To accomplish these goals, the study utilizes the LBLS-467 dataset to predict high-risk students, employing both logistic regression and neural networks as AI models. Subsequently, explanation artificial intelligence techniques such as LIME (Local Interpretable Model-agnostic Explanations) and SHAP (Shapley Additive Explanations) are utilized to evaluate students' learning outcomes and provide explanations. Finally, six evaluation indicators are adopted to assess the accuracy and stability of these explanations. In conclusion, this study addresses the challenges associated with inconsistencies in explainable AI models within the field of education. It emphasizes the need for explainability and trust when applying AI systems in educational contexts. By providing comprehensive explanations and evaluation metrics, this research empowers education teams to make informed decisions and fosters a positive environment for the integration of AI. Ultimately, it contributes to the reliable implementation of AI technologies, enabling their full potential to be harnessed in educational settings for the benefit of learners and educators alike.

Keywords

Explainable AI, Trustworthy, Learning Analytics [1]

1. Introduction

In recent years, artificial intelligence (AI) has been widely used in various fields. AI has shown great potential in these areas due to its ability to address specific needs within specific domains. However, as artificial intelligence continues to integrate into our lives, people are increasingly applying it to the decision-making process. Whether using AI for human resource decision-making[1]or for triaging and assisting in investigations of AI-related crimes[2], these examples demonstrate the significant impact of AI on humanity. Against this backdrop, the credibility of artificial intelligence has become one of the most critical issues of our time.

Although machine learning models were able to identify high-risk students early on, the black-box nature of these models created challenges in explaining their decision-making process and predicting outcomes. As a result, education teams have found it difficult to trust the decisions made by the models, leading to unexpected limitations in the use of AI in education [3]. Therefore, in recent years, researchers have increasingly combined explainable AI with predicting student learning outcomes to enable the explainability of model prediction processes[4, 5].

As we strive to build trustworthy artificial intelligence, it is critical to follow certain fundamental principles to ensure that it functions positively and reliably in a variety of contexts.

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

minjia9916@gmail.com (M. Li); ariel055132@g.ncu.edu.tw (S. Li); albertyang1994@gmail.com (C. M. Yang); anna.yuqing@gmail.com (Y.Q. Huang); stephen.yang.ac@gmail.com (J. H. Yang)

0000-0003-2168-2816 (C. M. Yang);0000-0002-2075-2256 (Y.Q. Huang); 0000-0003-1059-620X (J. H. Yang)

Evaluation is an important aspect to ensure reliability, and evaluation models and indicators need to be established to evaluate system performance [6, 7].

Advancements in LMS and learning analytics research lead to modular systems storing personal data in multiple locations. Anonymity is crucial for safeguarding data in integrated systems.[8]

By combining these aspects, we attempt to address the problem of inconsistent explanations produced by different explainable AI models when presented with the same dataset. This study will explore the selection of evaluation metrics to provide a comprehensive approach to this problem. Ultimately, this work will help strengthen the education team's understanding and trust in the model and promote the sustainable development of artificial intelligence in the field of education. The research questions are as follows:

RQ1: What specific evaluation metrics can be employed to assess the quality and effectiveness of explanations generated by explainable AI models?

RQ2: How to solve the problem of inconsistent explanations produced by different explainable AI models when presented with the same dataset?

2. Experiment Design

Figure 1 shows the experimental design flowchart outlining the sequential steps involved in conducting the study. The LBLS-467 dataset was obtained, the data will undergo preprocessing to handle missing values and identify high-risk and low-risk students. Following that, feature selection will be conducted to categorize questionnaire questions and learning behaviors into relevant features, and the data will be normalized. These features will be used for model training. After training, the model's effectiveness will be evaluated, and SHAP and LIME will be utilized for model explanation. Finally, six evaluation indicators will assess the quality of explanation.

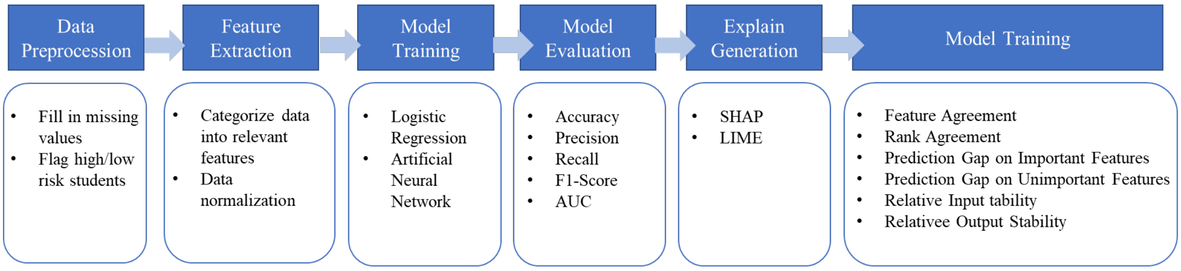


Figure 1: Experimental Design Flowchart

2.1. Dataset

The LBLS-467 (Learning Behavior Learning Strategy-467) dataset will be utilized as the data source for this experiment. LBLS-467 is an upgraded version of the LBLS-160 dataset[9]. The dataset consists of two components: Learning Behavior and Learning Strategy. Under the Learning Behavior section, the dataset captures students' learning behaviors in two online learning environments, namely BookRoll and VisCode. BookRoll is a digital material platform designed to facilitate online learning. Students can access materials, add bookmarks, and highlight important points. The system automatically records these learning behaviors for further analysis[10]. VisCode is an online Python development environment. When students practice using VisCode, the system automatically records their activities, including the time spent, any errors encountered, and the overall duration, for learning analysis purposes[11].

Regarding Learning Strategy, it collects data on students' self-regulated learning (SRL) through the use of the Motivated Strategies for Learning Questionnaire (MSLQ) and the Strategy Inventory of Language Learning (SILL) questionnaire. The MSLQ measures six dimensions of learning motivation, while the SILL assesses language learning strategies based on Oxford's categorization of language learning strategies proposed in 1990. It comprises six dimensions with a total of 50 items[12], although the dataset only includes 48 items as two questions were deemed irrelevant for programming language learning and were excluded. The modified questionnaire items were tailored to suit language learning in the context of programming languages[13].

2.2. Data Preprocessing and Feature Extraction

The LBLS-467 dataset was obtained, followed by preprocessing of the data, where missing values were filled with 0 and questionnaire data with a standard deviation of 0 was removed. Next, the data was divided into two categories: pass and fail. The fail category represents the 25% of students with lower learning status, indicating relatively backward academic performance, i.e., high-risk students[14].

Table 1 representing learning behaviors, focused on the analysis of student interactions with the BookRoll platform. We assessed the frequency of e-book openings, page turns, and page skips as indicators of active learning. Additionally, in the VisCode platform, we collected data to assess students' programming behaviors. This included recording the total usage time, frequency of opening VisCode, instances of code copying and pasting, lines of code written, and the overall count of error codes executed. These indicators provided valuable insights into students' programming engagement and proficiency.

Table 1
Features of Learning Behavior Extracted from the Learning Environment

System	Feature	Description
BookRoll	Marker_Operation	Number of markers added and deleted
	Memo_Operation	Number of times to add, delete, and modify note
	Bookmark_Operation	Number of times to add, delete, and skip to a bookmark
	Prev_Next_Operation	Number of times to turn page
	Jump_Operation	Number of times to jump to a bookmark, notes or key point
	Open_Num	Number of open the e-book
	Marker_Num	Number of markers (highlight key point)
	Memo_Num	Number of memos
	Bookmark_Num	Number of bookmarks
VisCode	Error_Num	Number of errors that occurred when VisCode ran the code
	Used_time	Total usage time
	Code_Length	Total number of lines of all programs
	Execute_Times	Total number of code executions
	Notebook_Open	Total number of times to open VisCode
	Code_Copy	Total number of code copies
	Code_Paste	Total number of times code was pasted

Table 2 presented a comprehensive questionnaire that evaluated various dimensions of learning strategies. By categorizing the questionnaire items according to the authors' descriptions, we gained a deep understanding of how students approach their learning. One example is the SRL (Self-Regulated Learning) questionnaire, which focused on rehearsal strategies involving repetitive review for better retention and comprehension.

The table encompassed different features and descriptions, including SRL Learning Motivation (intrinsic and extrinsic motivation, task value, control beliefs, self-efficacy, and test anxiety), SRL Learning Strategy (rehearsal, elaboration, organization, critical thinking, metacognitive self-regulation, time and study environment management, effort regulation, peer learning, and help-seeking), and SILL (Strategy Inventory for Language Learning) strategies (memory, cognitive, compensation, metacognitive, affective, and social strategies).

Data normalization is an effective data preprocessing strategy for data mining and machine learning[15-17]. In this study, Min-Max Normalization is employed to scale the feature data to a range of 0 to 1 while preserving the original data distribution.

Table 2
Learning Strategies Features Extracted from Learning Questionnaires

Questionnaire	Feature	Description
SRL Learning Motivation	intrinsic	The internal drive and enjoyment individuals experience when engaging in learning activities.
	extrinsic	External factors, such as rewards or recognition, that influence individuals' engagement in learning activities.
	task_value	Assess students' perceptions of the interest, importance and usefulness of course content
	control_beliefs	Assesses whether students believe their hard work will lead to positive outcomes
	self_efficacy	Assesses the judgment and confidence that the student can complete the task independently
	test_anxiety	Student anxiety levels about tests in the course
SRL Learning Strategy	rehearsal	Repetitive or repeated review of study material to enhance retention and understanding.
	elaboration	Enhancing understanding by making connections and creating meaningful associations with prior knowledge.
	organization	The act of structuring and arranging information in a systematic and logical manner to facilitate comprehension and retrieval.
	critical_thinking	The process of objectively analyzing and evaluating information to make informed judgments and decisions.
	metacognitive_self_regulation	The skill of monitoring and controlling one's own learning process for better outcomes.
	time_and_study_environment	Encompass managing study time effectively and creating an optimal setting for focused learning.
	effort_regulation	The skill of consciously managing and adjusting one's level of effort to maximize learning outcomes.
	peer_learning	The process of students learning from and with their peers, through collaborative activities and discussions.
	help_seeking	Students seek support from others when they encounter difficulties in their studies.
SILL	memory	The ability and strategies used to effectively remember and recall the materials learned.
	cognitive	The mental processes and abilities involved in learning, such as attention, memory, thinking, and problem-solving.
	compensation	The use of alternative strategies or resources to overcome difficulties or limitations in language skills or knowledge.
	metacognitive	Students will plan, organize, evaluate and monitor their own language learning
	affective	Students regulate their emotions, motivations and attitudes when learning a language
	social	Student interacts with others while learning

2.3. Model Training, Evaluation and Explanation

Commonly used machine learning models in educational scenarios include random forest (RF), support vector machine (SVM), decision tree (DT), logistic regression (LR), K-nearest neighbor algorithm (KNN), and artificial neural network (ANN). Logistic regression and neural networks, however, have been found to provide more accurate predictions compared to other methods[18, 19]. Hence, this experiment utilizes logistic regression and neural networks for prediction.

To assess the predictive performance of various machine learning algorithms, this study employs five indicators: Accuracy, Precision, Recall, F1-Measure, and Area Under Curve (AUC). These indicators are widely used to evaluate the classification performance of models[20, 21]

This study utilizes LIME[22] and SHAP[23] as explanation generators for model predictions. Six evaluation indicators are employed to assess the quality of explanations, measuring authenticity and stability[24]: Feature Agreement (FA), Rank Agreement (RA), Prediction Gap on Important Features (PGI), Prediction Gap on Unimportant Features (PGU), Relative Input Stability (RIS) and Relative Output Stability (ROS).

The first four evaluation indicators (FA, RA, PGI, PGU) assess the accuracy of the explanations, while RIS and ROS evaluate their stability. FA and RA are specific to linear models (e.g., linear regression, logistic regression)[24], whereas PGI, PGU, RIS, and ROS are applicable to all models.

Feature Agreement (FA) quantifies the proportion of the top K features that exhibit consistent rankings between explanations generated by AI and predictions made by the model. Prediction Gap on Important Features (PGI) measures the difference in prediction probabilities when influential features, as identified by explainable AI-generated explanations, are perturbed. Higher PGI values indicate a stronger correspondence between the explanation and the prediction. Conversely, Prediction Gap on Unimportant Features (PGU) measures the change in prediction probabilities when non-influential features, as identified by explainable AI explanations, are perturbed. PGU serves as an indicator of the explanation's accuracy in capturing non-influential factors.

Relative Input Stability (RIS) and Relative Output Stability (ROS) quantify the maximum change in the explanation generated by explainable AI in relation to the predicted input and output probabilities, respectively. These metrics evaluate the stability of the explanation.

3. Result

3.1. Evaluation of Model Efficacy

Table 3 displays the training performance results using LR and ANN, indicating that logistic regression outperforms artificial neural networks in predicting whether a student is high-risk or low-risk.

Table 3
Predictive Performance Results of Different Models

Method	Accuracy	Precision	Recall	F1-Score	AUC
LR	85.8%	78.9%	85.8%	81.0%	60.8%
ANN	63.9%	65.3%	63.9%	61.3%	58.8%

3.2. Explanation of Discrepancy Results

Both SHAP and LIME can provide explanations for model predictions on individual student data. SHAP offers explanations through waterfall plots, while LIME utilizes its own graphical representation.

The waterfall plot in SHAP is designed for analyzing the most important features contributing to a high-risk prediction for a single data point. The X-axis represents the SHAP value, indicating the impact (positive or negative) of the corresponding feature on the prediction. The Y-axis represents the data features and their values for that particular data point (e.g., 337 = Error_Num, which represents the total number of errors the student encountered while compiling code). The function f(x) represents the prediction result given by SHAP, considering all features. If f(x) equals 1, it indicates a high-risk student, while f(x) equals 0 represents a low-risk student. E[f(x)] represents the average prediction value of the model across the dataset.

Figure 2 illustrates how SHAP explains the predictions of a logistic regression model through a waterfall plot for a high-risk student. The following information can be observed: The majority of students are predicted as low-risk, as the value of E[f(x)] is 0.111, approaching 0. In the plot, this student is predicted as high-risk, with a value of 1 for f(x).

The main reason for labeling this student as high-risk is the high value of Error_Num (with a SHAP value of 0.69), which far exceeds the SHAP values of Marker_Num and Marker_Operation (0.07 and 0.06, respectively). This indicates that the student's excessive errors during code compilation are the primary factor contributing to their high-risk prediction.

Throughout the semester, this student made a total of 337 errors while compiling code, with the third quartile of Error_Num being 152. This confirms the student's tendency to make a relatively high number of errors during code compilation.

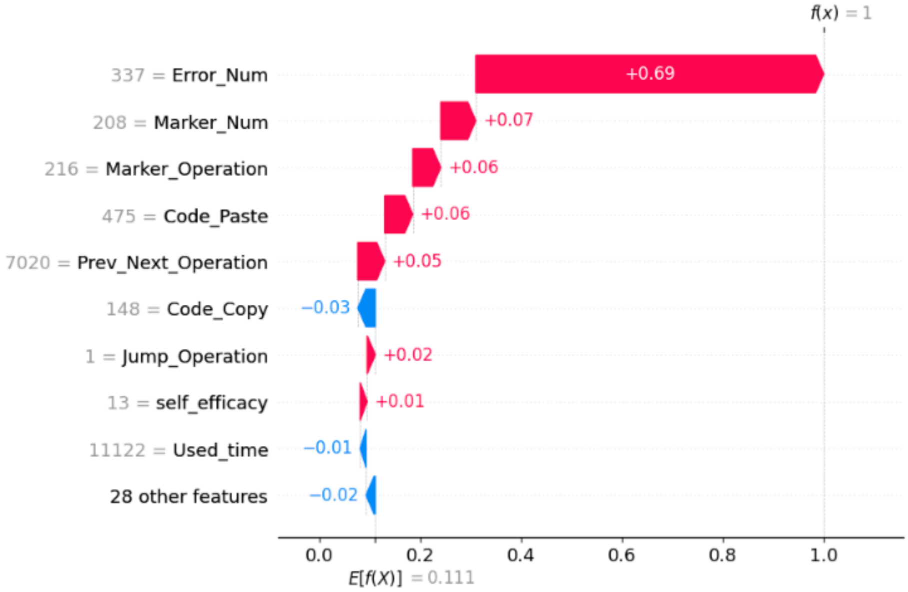
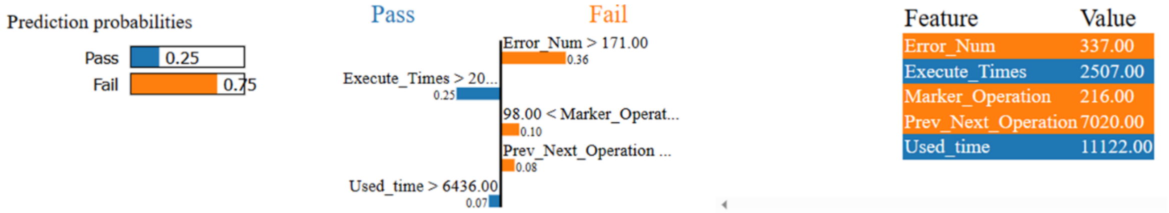


Figure 2 Waterfall plot: Using SHAP to Explain High-Risk Students

Figure 3 demonstrates how LIME explains the predictions of a logistic regression model for the same student as a high-risk student. From the Prediction Probabilities on the left, it can be seen that LIME predicts the probability of this student being a low-risk student as 0.25, while the probability of being a high-risk student is 0.75. This indicates that LIME leans towards considering this student as a high-risk student.

The middle chart indicates the five most important features and their contributions to the prediction, as well as the prediction rules. For example, the top five features are Error_Num (total number of errors during code compilation), Execute_Times (total number of program executions), Marker_Operation (frequency of using key functions), Prev_Next_Operation (total number of page flips), and Used_Time (total duration of using VisCode). It is also stated that if Error_Num exceeds 171, LIME considers this student as a high-risk student, with Error_Num contributing 0.36 to this prediction.

The right chart indicates the actual values of these features. For example, this student made 337 errors while compiling code and executed a total of 2507 programs.



Feature	Value
Error_Num	337.00
Execute_Times	2507.00
Marker_Operation	216.00
Prev_Next_Operation	7020.00
Used_time	11122.00

Figure 3 Using LIME to Explain High-Risk Students

This study found inconsistencies in the features used by LIME and SHAP, indicating a discrepancy in the explanations generated by these two explainers. The detailed features are shown in Table 4.

Table 4
The Five Most Important Features Explained for the Same High-Risk Student

LIME	SHAP
Error_Num	Error_Num
Execute_Times	Marker_Num
Marker_Operation	Marker_Operation
Prev_Next_Operation	Code_Paste
Used_Time	Prev_Next_Operation

3.3. The quality of Explanation

Table 5 presents the explanation performance results obtained by using logistic regression and artificial neural networks with two different explainers, LIME and SHAP. It is observed that when using logistic regression, the explanation quality generated by the LIME explainer is superior to that of SHAP. On the other hand, when using artificial neural networks, the explanation quality generated by the SHAP explainer is better than that of LIME.

Table 5
Explanation Quality of Various Explainable AI Algorithms with Different models

Method	FA	RA	PGI	PGU	RIS	ROS
LR+SHAP	0.70	0.62	0.0072	0.0036	3.23	4.69
LR+LIME	0.97	0.72	0.0081	0.0043	0.22	3.68
ANN+SHAP	N/A	N/A	0.0133	0.0023	0.21	2.43
ANN+LIME	N/A	N/A	0.0135	0.0026	0.30	3.69

4. Discussion

We noticed in Table 4 that the five most important features are highlighted by LIME and SHAP. There are three common features, meanwhile, two different features came from each explanation model. This raises an intriguing perspective that the shared emphasis on these features may indicate higher importance and reliability.

This viewpoint sparks our interest in delving deeper into the correlation between model predictive performance and explanatory performance. We plan to further explore this in the upcoming discussion. This not only enriches our discourse but also contributes to providing a more comprehensive perspective to address various viewpoints and concerns.

4.1. Methods for Assessing Explanations

To address the first research question regarding the evaluation of explanations from explainable AI models, we utilized six metrics, as detailed in Section 2.3. The first two metrics, Feature Agreement (FA) and Rank Agreement (RA), assess fidelity and consistency in explanations generated by post-hoc models like SHAP or LIME. FA measures the shared top-K features between the post-hoc explanation and the model's feature-based importance ranking, while RA evaluates feature ordering consistency. These metrics contribute to understanding the explainability and trustworthiness of AI systems.

The third and fourth metrics, Prediction Gap on Important Features (PGI) and Prediction Gap on Unimportant Features (PGU), quantify the impact of perturbations on identified influential and unimportant features. PGI reflects the alignment between influential features and the model's prediction, while PGU assesses the model's disregard for unimportant features. These metrics enhance our understanding of explainability and trustworthiness.

Two additional metrics, Relative Input Stability (RIS) and Relative Output Stability (ROS), were incorporated. RIS measures changes in the explanation due to slight input modifications, indicating explanation stability. A smaller RIS value signifies higher stability. Similarly, ROS quantifies changes in the explanation relative to variations in output probabilities, assessing explanation robustness. A lower ROS value indicates higher stability. These metrics aim to comprehensively evaluate the fidelity, alignment, and stability of post-hoc explanations in diverse AI domains.

4.2. Exploring the Quality of Explanations

Although the accuracy of ANN is lower, we are still interested in seeing how LIME and SHAP perform on LR and ANN. It is worth noting that six evaluation metrics, including FA, RA, PGI, PGU, RIS, and ROS, are specifically designed to evaluate explanatory artificial intelligence (XAI) and are not affected by model accuracy. Therefore, although the accuracy of ANN is lower, it does not affect our evaluation of LIME and SHAP in terms of interpretation.

From Table 5, it is evident that LIME produces explanations with higher accuracy and consistency compared to SHAP for the LR model. While LIME exhibits a higher PGU compared to SHAP, indicating a greater variance in predictions concerning unimportant features, it is essential to consider other factors in evaluating the overall performance. It is crucial to recognize that PGU primarily emphasizes the explanation of unimportant features. In this particular experiment, we emphasize the significance of paying more attention to explanations related to important features. Additionally, the model's explanatory performance is not solely determined by a single indicator; instead, a comprehensive assessment considering multiple indicators is necessary.

LIME performance better than SHAP for the LR model can be attributed to the similarity in functionality between LIME and logistic regression. LIME operates by providing a locally interpretable model (using linear regression) for complex and opaque models, aiming to find a simple and understandable model for a specific instance to address the question of "why the model classifies an instance into a specific category"[22].

Both logistic regression and linear regression employ similar formulas, with the distinction that logistic regression applies a sigmoid function to transform the regression results into predicted probabilities, while linear regression does not involve this sigmoid transformation. This resemblance in approach between logistic regression and LIME explains why LIME performs better when explaining logistic regression models[18, 22].

On the other hand, SHAP generates explanations with higher fidelity and stability in the ANN model compared to LIME. This may be attributed to the fact that SHAP's functioning is more similar to artificial neural networks. SHAP operates by analyzing the explainability of a model's predictions in terms of the contribution of each feature, calculating the Shapley values for each feature to measure its impact on the predictions. Higher contribution indicates higher importance of that feature. However, the training methodology of artificial neural networks involves transformations through the states of neurons in hidden layers, representing nonlinear classification[25]. Therefore, linear regression, which is used by LIME, may not provide explanations of higher quality, leading to SHAP's feature contributions aligning better with the training methodology of artificial neural networks.

4.3. Resolving Differences in Explanations

Firstly, user feedback can be gathered through methods such as questionnaires or interviews to assess the quality of explanations. Secondly, in the absence of user or expert input, explanations can be selected based on their accuracy and stability. If the explanations generated by a specific explainer demonstrate better stability compared to those produced by other explainers, that specific explainer can be chosen. For example, when dealing with predictions made by artificial neural network models, SHAP's explanations outperform LIME in three out of four evaluation criteria (PGI, PGU, RIS, and ROS). Therefore, SHAP explanations can be employed in such scenarios. Conversely, LIME's explanations surpass SHAP in five out of six evaluation criteria (FA, RA, PGI, RIS, and ROS) for logistic regression models. Hence, LIME explanations can be utilized when working with logistic regression predictions.

5. Conclusion

In conclusion, this study emphasizes the significance of explainable and trustworthy AI in the field of education. By employing two machine learning methods (logistic regression and neural networks) and two explainable AI packages (LIME, SHAP), the research evaluates and generates explanations for students' learning outcomes. The use of six evaluation metrics for explainability ensures the accuracy and stability of these explanations.

The findings of this study contribute to the development of explainable AI models that are transparent and can be trusted by education teams. By providing comprehensive explanations for model predictions, the study enhances the understanding and confidence of stakeholders in the decision-making process of AI systems. This promotes the responsible and sustainable integration of artificial intelligence in educational settings. Moreover, the research highlights the importance of evaluation metrics in assessing the quality and effectiveness of explanations generated by explainable AI models. Establishing such metrics not only ensures reliable system performance but also supports the establishment of fundamental principles for building trustworthy AI in various contexts.

In our future work, we plan to extend our analysis to include diverse model architectures, which will allow us to highlight differences in particular features and explore the aspects of importance that transcend specific model structures. This expansion could contribute to a more robust evaluation of model interpretability and feature importance across various modeling paradigms.

Acknowledgements

This work is supported by Ministry of Science and Technology, Taiwan under grants MOST-112-2410-H-008 -019 -MY3, MOST-111-2410-H-008 -010 -MY3, and Ministry of Education, Taiwan. Lastly, I acknowledge the assistance of ChatGPT in refining the language of this paper.

References

[1] Park, H., et al. Human-AI interaction in human resource management: Understanding why employees resist algorithmic evaluation at workplaces and how to mitigate burdens. in Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. 2021.

[2] Sibai, F.N. AI crimes: A classification. in 2020 International Conference on Cyber Security and Protection of Digital Services (Cyber Security). 2020. IEEE.

[3] Baker, R.S., Challenges for the future of educational data mining: The Baker learning analytics prizes. Journal of Educational Data Mining, 2019. **11**(1): p. 1-17.

[4] Alwarthan, S., N. Aslam, and I.U. Khan, An explainable model for identifying at-risk student at higher education. IEEE Access, 2022. **10**: p. 107649-107668.

[5] Jang, Y., et al., Practical early prediction of students' performance using machine learning and eXplainable AI. Education and Information Technologies, 2022: p. 1-35.

[6] Díaz-Rodríguez, N. and G. Pisoni. Accessible Cultural Heritage through Explainable Artificial Intelligence. 2020.

[7] Cheng, L., K.R. Varshney, and H. Liu, Socially responsible AI algorithms: Issues, purposes, and challenges. Journal of Artificial Intelligence Research, 2021. **71**: p. 1137-1181.

[8] Flanagan, B. and H. Ogata, Learning analytics platform in higher education in Japan. Knowledge Management & E-Learning: An International Journal, 2018. **10**(4): p. 469-484.

[9] LUa, O.H., et al., A Quality Data Set for Data Challenge: Featuring 160 Students' Learning Behaviors and Learning Strategies in a Programming Course.

[10] Ogata, H., et al. E-Book-based learning analytics in university education. in International conference on computer in education (ICCE 2015). 2015.

[11] Lu, O.H., et al. Early-Stage Engagement: Applying Big Data Analytics on Collaborative Learning Environment for Measuring Learners' Engagement Rate. in 2016 International Conference on Educational Innovation through Technology (EITT). 2016. IEEE.

[12] Pintrich, P.R., A manual for the use of the Motivated Strategies for Learning Questionnaire (MSLQ). 1991.

[13] Oxford, R., Language learning strategiesWhat every teacher should know. 1990: Heinle & heinle Publishers.;.

[14] Osmanbegovic, E. and M. Suljic, Data mining approach for predicting student performance. Economic Review: Journal of Economics and Business, 2012. **10**(1): p. 3-12.

[15] Gao, J., Machine learning applications for data center optimization. 2014.

[16] O'shea, T.J. and N. West. Radio machine learning dataset generation with gnu radio. in Proceedings of the GNU Radio Conference. 2016.

[17] Zitnik, M., et al., Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities. Information Fusion, 2019. **50**: p. 71-91.

[18] Oqaidi, K., S. Aouhassi, and K. Mansouri, Towards a Students' Dropout Prediction Model in Higher Education Institutions Using Machine Learning Algorithms. International Journal of Emerging Technologies in Learning (Online), 2022. **17**(18): p. 103.

[19] Tomasevic, N., N. Gvozdenovic, and S. Vranes, An overview and comparison of supervised data mining techniques for student exam performance prediction. Computers & education, 2020. **143**: p. 103676.

[20] Ferri, C., J. Hernández-Orallo, and R. Modroiu, An experimental comparison of performance measures for classification. Pattern recognition letters, 2009. **30**(1): p. 27-38.

[21] Hossin, M. and M.N. Sulaiman, A review on evaluation metrics for data classification evaluations. International journal of data mining & knowledge management process, 2015. **5**(2): p. 1.

[22] Ribeiro, M.T., S. Singh, and C. Guestrin. " 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. 2016.

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

[24] Agarwal, C., et al., Rethinking stability for attribution-based explanations. arXiv preprint arXiv:2203.06877, 2022.

[25] Jain, A.K., J. Mao, and K.M. Mohiuddin, Artificial neural networks: A tutorial. Computer, 1996. **29**(3): p. 31-44.