FIPER: a Visual-based Explanation Combining Rules and Feature Importance

Eleonora Cappuccio[1,2,3][0000-0002-6105-2512], Daniele Fadda[2][0000-0002-0051-0604], Rosa Lanzilotti[3][0000-0002-2039-8162], and Salvatore Rinzivillo[2][0000-0003-4404-4147]

[1] Università di Pisa, Italy
{name.surname}@phd.unipi.it
[2] Università degli Studi di Bari Aldo Moro
{name.surname}@uniba.it
[3] KDDLab - ISTI - CNR, Pisa, Italy
{name.surname}@isti.cnr.it

Abstract. Artificial Intelligence algorithms have now become pervasive in multiple high-stakes domains. However, their internal logic can be obscure to humans. Explainable Artificial Intelligence aims to design tools and techniques to illustrate the predictions of the so-called black-box algorithms. The Human-Computer Interaction community has long stressed the need for a more user-centered approach to Explainable AI. This approach can benefit from research in user interface, user experience, and visual analytics. This paper proposes a visual-based method to illustrate rules paired with feature importance. A user study with 15 participants was conducted comparing our visual method with the original output of the algorithm and textual representation to test its effectiveness with users.

Keywords: User-centric Explainable AI · Visual Analytics · User Interfaces for Explainable AI

1 Introduction

Artificial Intelligence (AI) and Machine Learning (ML) decision-making systems are widely used in high-stakes domains such as healthcare, justice, and finance. Their usefulness in solving increasingly complex tasks comes at a cost: the internal logic behind the model is often unintelligible to humans. Following the General Data Protection Regulations (GDPR) by the European Union, which establishes a right to an explanation for a user affected by an automated decision-making system [34], there has been an emergence of explainable artificial intelligence (XAI) techniques in recent years [1, 16, 10]. These techniques seek to make AI and ML models interpretable by humans. However, several studies [22, 3, 19, 1, 17] have pointed out that most of the work in XAI is built upon researchers' belief of what a "good" explanation is [3, 22], framing XAI mainly as an algorithmic problem and not focusing enough on the user's point of view [23].

The Human-Computer Interaction (HCI) community has recently directed its attention toward the intersection of Artificial Intelligence (AI) and Explainable AI (XAI), incorporating theories and concepts from the rich research field of HCI and reframing XAI as a design problem[17]. In the present study, we introduce a novel and innovative approach named FIPER (Feature Importance Plot for Explanatory Rules) that leverages the visualization of explanations through the fusion of rules and feature importance. While rules play a central role in our investigation, we augment their significance by incorporating feature importance, which serves to provide valuable context regarding the role and relevance of a specific feature upon which a rule's predicate has been generated. By integrating these two components, we anticipate a significant enhancement in the user experience, facilitating a more comprehensive and intuitive understanding of the underlying prediction algorithm.

The central role in this study is played by rules; however, they are supported by Feature Importance(FI). FI provides context for the significance and relevance of a feature for which a rule predicate has been formulated. This integration is expected to enhance both the user experience and the user's conceptual understanding of the prediction algorithm. *FIPER* exploits interactivity by design to manage datasets with a high number of features. The user can filter the attributes to focus only on those that are predicates of the rule. In this way, the cognitive workload is lowered. *FIPER* is designed around the data scientist/developer enabling the user to verify hypotheses and to assess that the AI model works accordingly to the expected behavior [28]. A user study was performed to investigate if *FIPER* effectively supports the users in their work. [1, 25].

2 Related Work

This work contributes to the field of human-centered explainable AI, which seeks to bring techniques and methodologies from HCI into the design of explanations [17]. Our primary objective is to leverage a visual representation of the explanation to facilitate and enhance human interpretation. In 2018 the Defense Advanced Research Projects Agency (DARPA) "explainable AI initiative" framed the explainable AI process as a three-stage approach, distinguishing between the explainable model, the explanation user interface, and the psychological requirements crucial for their design. By differentiating between the model responsible for generating explanations for machine learning algorithms and the means employed to effectively communicate these explanations to the user, this framework provides a comprehensive understanding of the multifaceted nature of explainable AI.

By aligning our research with these established frameworks and principles, we aim to contribute to advancing the field by proposing a novel visual-based approach that caters to the psychological requirements of users. By focusing on the design of an intuitive and visually appealing explanation user interface, we aspire to bridge the gap between complex machine learning models and human comprehension, thereby enabling users to gain meaningful insights and a deeper understanding of the underlying AI algorithms [12, 4].

Chromik *et al.* define an explanation user interface (XUI) *"as the sum of outputs of an XAI system that the user can directly interact with. An XUI may tap into the ML model or may use one or more explanation-generating algorithms to provide relevant insights for a particular audience"* [4]. The separation between the explainable algorithm and how the explanation is presented to the users has also been pointed out by [6]: the authors differentiate between explanation techniques and explanation visualizations. The first involves the generation of *rough explanations*, usually propounded by AI researchers, while the latter concerns how these rough explanations are presented to users. Text can be used to convey a simple form of explanation, while the conjunctive use of text and visual cues can enhance how explanations are delivered to the users [8]. However, visualization is better suited to communicate complex concepts [2]. [35] cite basic charts for raw data, as well as tornado diagrams for list attribution, and saliency heatmaps for image-based models. It has been proven by [25] that the way an explanation is displayed has an effect on how the user makes decisions. In [36], the authors investigate how different visual displays of example-based explanation affect the user appropriate trust of the ML classification. However, as reported by [25], the design and testing of different visualizations for Explainable AI are still under-studied. The pure text has been used to show rules [24, 9]; however, some visual examples can be found in [25, 24, 2]. Another key point for explanations is the implementation of interactivity: although advocated in several studies, its integration within explanations is still limited [1, 5, 21]

3 Visual Explanation for Rules and Feature Importance

3.1 XAI methods

We address the problem of representing explanations based on rules in a visual format to enable the user to investigate the relationship of the input with the outcome of the decision system. Accordingly to [11], a rule can be formally defined as a statement like $p \rightarrow y$, where the *consequence* y is the output of the black-box and the *premise* p is a conjunction of split conditions on the observed features, where each condition can be represented as a predicate of the form $a_i \in [v_{i,l}, v_{i,h}]$, where a_i is one of the features of the data and $v_{i,l}$ and $v_{i,h}$ are respectively the lower and upper bounds for the domain of a_i where the predicate is valid. For categorical data types the predicate has the form $a_i \in \{v_l, v_j, \ldots, v_k\}$, where each v_i is a value of a_i. An instance x is covered by a rule r if all the predicates of the premise of r are satisfied by x.

Automatic scripts and programs can efficiently manage rules to enable support for reasoning and exploration. However, this formal representation may present a high cognitive load for the user. Moreover, rule predicates do not provide an explicit ranking of the features of the data. Explanation methods based on Feature Importance provide a ranking of each feature based on the relevance of the feature in the final decision. Given an instance x, a FI method returns an ordered sequence of pairs $[(a_j, w_1), (a_l, w_2), \dots, (a_k, w_m)]$, where each feature a_i is associated with a weight w_i that represents the relevance of a_i in the decision. For local explanation, the reference to a feature a_i intentionally means the actual value observed for a_i in the current instance x. The explanations based on FI are lightweight from a cognitive point of view, although they provide less information than those based on rules.

We propose a visual interface that combines the strong points of both groups of explanation strategies. In particular, we exploit FI to enforce a ranking on the visualized features to guarantee that the most relevant are the firsts shown to the user. We introduce a visual encoding to represent the intervals yielded by the rule predicates to easily catch the relationship of each interval with the global distribution of the data. Without loss in generality, we identified two methods for both families of explanation strategies: LORE [11] and SHAP [20]. We opted for LORE due to prior experience of its adoption in previous case studies. Nonetheless, the visualization is versatile enough to accommodate other rule-generating algorithms like Anchor [31], by adopting a translation interface to match the input schema of our tool. For calculating Feature Importance, we employed SHAP, a widely recognized standard in the field. Alternately, other algorithms such as LIME [30], which also generate Feature Importance, can be employed.

3.2 Visualization

Our proposal organizes the explanation's visual space to combine the information yielded by the FI and the rule-based methods. Figure 1 shows a visualization of an instance extracted from the *German Credit Risk* dataset from UCI [7], a dataset widely used for educational purposes. The visualization comprises two panels: on the left, the weights of the FI methods are reported and sorted accordingly to their absolute values; on the right, the rule predicates are visualized following the order of the first diagram. The FI panel represents the weights by color coding the corresponding feature's positive (blue) or negative (magenta) contribution. The visualization is designed using a color-blind-friendly palette. The rule predicates panel visualizes all the features with a specific chart aligned with the elements in the FI panel. We use two different types of charts based on the type of each feature. For *categorical data types*, we adopt a stacked bar chart to show the part-of-the-whole relationship of each possible value. With this representation, the user can catch the internal distribution of the values. A diamond point is located in the center of the value observed for the attribute in x. For *numerical data types*, we use a box plot chart that shows a compact visualization of the data distribution: min, max, first quartile, third quartile, and median. The observed value for x is represented by a diamond point located within the scale of the box plot. For those attributes for which exists a predicate p in the rule r we add a second layer to highlight the intervals of the rule. The interval visualization changes accordingly to the data type. For categorical data, the intervals contained in the rule premise are highlighted in yellow. For numerical data, a yellow bar represents the extent of the predicate values.

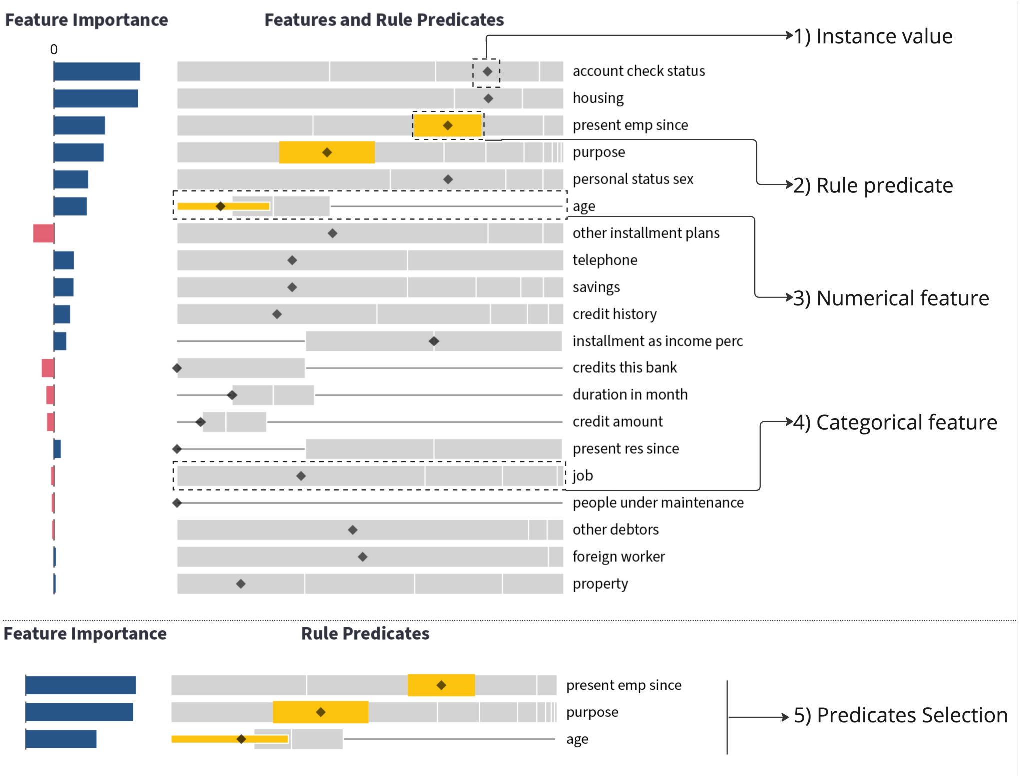
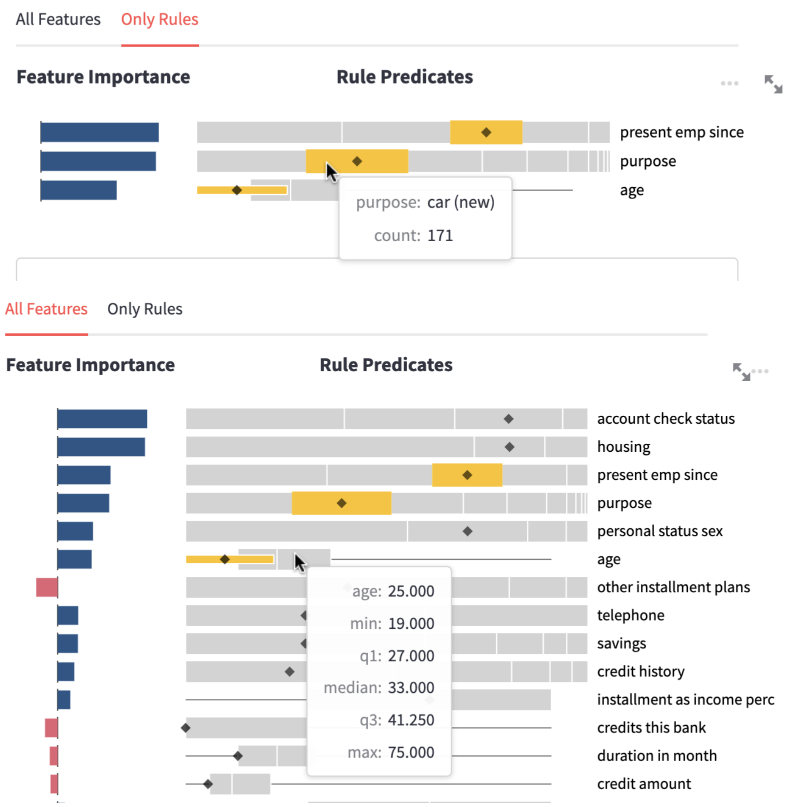


Fig. 1: *FIPER* visualization of one instance of the *German Credit Risk* dataset. *(Top)* Attributes are sorted by the absolute value of FI. Categorical attributes are represented as stacked absolute bar charts. Numerical values are represented as box plots. The interval contained in the predicates of the rule are highlighted in yellow. *(Bottom)* Filtered view of the visualization, showing only the attributes referred in the rule premise

The example in Figure 1 shows a rule for an instance of the dataset that is classified as *Bad Credit Risk*. The attributes are sorted by the absolute values of weights of FI. For instance, the attributes *account check status* and *housing* (both categorical) are the most relevant for FI, even if they are not mentioned by the rule associated with the prediction. The predicates of the rule refer to three attributes: *present employed since*, *purpose*, and *age*. The interval for the predicate for age is relevant since it covers the lower part of the distribution. The diamond shows that the associated value is below the first quartile. As suggested by [1, 22, 4], two forms of interactivity are implemented:

- To focus the user's attention only on the predicates, it is possible to dynamically restrict the view only to those attributes mentioned within the rule. This interaction follows the principles of giving users easy access to relevant and important information [32, 18]. The lower part of Figure 1 *(Bottom)* shows the restricted version.

– For each attribute, the user may get access to a finer level of details of the corresponding distribution by hovering the pointer over the visualization. Figure 2 shows two different styles of tooltips for two different data types. For the categorical data type (Top), we show the selected value and its cardinality. For numerical data type, we show a set of representative values (min, max, Q1, Q3, median) and the value of the corresponding feature.



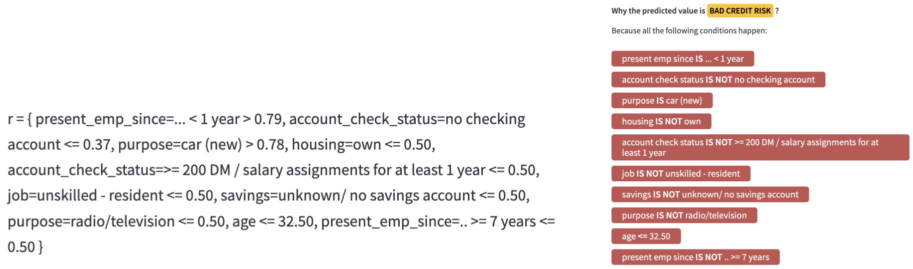
purpose: car (new)
count: 171

age: 25.000
min: 19.000
q1: 27.000
median: 33.000
q3: 41.250
max: 75.000

Fig. 2: Finer details of a specific feature, selected by hovering the mouse on the corresponding row. (Top) Tooltip for a categorical data type, where the feature's actual value is shown with its class's cardinality. (Bottom) Tooltip for a numerical data type, where statistical central values are shown: min, max, median, Q1, and Q3.

3.3 Other explanation modalities.

We compared *FIPER* with two other interfaces, presented in Figure 3. LORE output is the raw output of the algorithm as text. The XAI library visualization is the implementation already available within the XAI Library[4]. The latter enhances the rule's content by improving each predicate's readability with a sequence of graphical blocks.



r = { present_emp_since=... < 1 year > 0.79, account_check_status=no checking account <= 0.37, purpose=car (new) > 0.78, housing=own <= 0.50, account_check_status=>= 200 DM / salary assignments for at least 1 year <= 0.50, job=unskilled - resident <= 0.50, savings=unknown/ no savings account <= 0.50, purpose=radio/television <= 0.50, age <= 32.50, present_emp_since=.. >= 7 years <= 0.50 }

Why the predicted value is BAD CREDIT RISK ?

Because all the following conditions happen:

- present emp since **IS** ... < 1 year
- account check status **IS NOT** no checking account
- purpose **IS** car (new)
- housing **IS NOT** own
- account check status **IS NOT** >= 200 DM / salary assignments for at least 1 year
- job **IS NOT** unskilled - resident
- savings **IS NOT** unknown/ no savings account
- purpose **IS NOT** radio/television
- age <= 32.50
- present emp since **IS NOT** .. >= 7 years

(a) LORE Output (b) XAI Library

Fig. 3: The same instance of Figure1 visualized as LORE output and XAI library visualization

4 User Study

A user study was performed to better understand the value of *FIPER* in providing clear and understandable explanations. More specifically, three different explanation modalities were compared (i.e., Lore simple output, XAI library output, and *FIPER*) to answer the following research questions:

- **RQ1**. Can the explanation modalities support data scientists in understanding the AI model?
- **RQ2**. What is the difference among the explanation modalities regarding data scientist satisfaction?

4.1 Participants and design

A total of 15 students in the second year of the Master's degree in Data Science (7 females) participated in the study. Their mean age was 25 years (SD = 2.7, min = 22, max = 31). Out of 15 students, 12 declared their knowledge of the dataset as "very low" and "low", while 3 stated that they had good knowledge of it. Considering the number of participants, a within-subject design was performed

[4] `https://pypi.org/project/XAI-Library/`

[29], with the explanation modality as an independent variable and three within-subject factors, i.e., Lore simple output, XAI library output, and *FIPER*. The participants worked individually with the three modalities and provided their opinion.

4.2 The experimental tasks

The participants were asked to carry out a sequence of 3 instances of increasing difficulty, defined by the number of predicates in the rule. For each instance, the participant had to answer three experimental questions. The first one asked them to indicate which features are present within the rule. The second question requested them to identify (if any) rule predicates that are insignificant to the prediction. Finally, they were asked to identify the most relevant predicate of the rule to determine the prediction. Because each of the 15 participants answered the above 3 questions for each of the 3 instances on 3 modalities, the total number of answers collected was 405 ($15 \times 3 \times 3 \times 3$). To avoid possible unfair effects of learning from the first task (i.e. order effect) [33], the questions and the explanation modality order were counterbalanced across the participants, according to a Latin Square design.

4.3 Procedure

The study occurred in a quiet university room where the students attended their classes. Three researchers were involved, who intervened just if technical problems emerged. The study lasted one hour and a half, starting from the presentation of the study goal to the participants, including the interaction with the three visualization modalities to answer the experimental questions, until the completion of questionnaires administered before and after the interaction to collect data about the participants and their opinion on visualizations. All participants followed the same procedure. First, they were introduced to the study purpose and what they had to do. Participants were asked to sign an informed consent as our university's ethics committee requires for the user study. All participants provided consent. Then, the participants were invited to the study via a link to a web platform that allowed them to answer the experimental questions using the three modalities. Once the participant clicked on the link, a page providing an overview of the study and its goals appeared. At this stage, the platform requested participants to fill in a questionnaire to collect their demographic data, and their familiarity with the domain on a scale from 1 to 7 (1 - being not familiar at all, 7 - being very familiar). The other data were collected anonymously, with no means of identifying individual participants. The platform randomly assigned participants to one of the three explanation modalities. Thus, a training session where they saw the instructions on how to read each explanation was followed by one practice trial. The actual study session then started. The participant interacted with the first visualization modality to answer the three experimental questions for each instance. Then, the participant completed an online questionnaire including NASA-TLX [13]. This procedure was the same for all the 3 conditions. However, before repeating the same procedure with the next modality, the participants were invited to relax for 5 minutes. Finally, the platform asked participants to fill in a final questionnaire to express their satisfaction with the modalities they had just used and to vote for the best explanation modality and explain why; the questionnaire included the User Engagement Scale (UES) [26] in its short form related to the preferred visualization. At the end of the study, participants were thanked for their participation. A pilot study involving three participants was conducted to check the overall research methodology.

4.4 Data collection and analysis.

Quantitative and qualitative data were collected to answer the two research questions. To analyze the support (RQ1) provided by the explanation modalities to data scientists, metrics such as the error rate and task execution time were considered.

4.5 Error rate.

Participants could make two different types of errors while performing the tasks: when asked to list a set of features, they could either enter a feature that was not present (E1) or not enter a feature that was present (E2). To calculate the error rate, we created nine vectors containing the correct answers, three for each instance. The elements of the vectors were equal to the number of features in the train set of the dataset, plus the "I' don't know" option. The 9 vectors were compared with those of the responses given by the users to compute the error rate; in Figure **??**, the error rate of Lore Output is compared to the other two visualizations. During the task for each instance and visualization, we noted the completion time each participant took to analyze the visualizations and complete the task. We use these time measurements as a proxy for the cognitive workload of each condition.

4.6 User satisfaction.

The online questionnaire to investigate satisfaction (RQ2) with the explanation modality was composed of two sections. The first section proposed the NASA-TLX questionnaire, used as "Raw TLX" [13]. It is a 6-item survey that rates perceived workload using a system through 6 subjective dimensions, i.e., Mental Demand, Physical Demand, Temporal Demand, Performance, Effort and Frustration, rated within a 100-point range with 5-point steps (lower is better). These ratings were combined to calculate the overall NASA-TLX workload index [14]. Specifically, the NASA-TLX was used to assess the workload caused by each modality because the user's workload when using a software tool influences user satisfaction. The second section presented the new UES (User Engagement Scale) short form, derived from the UES long form. It is a 12-item survey used to measure user engagement, a quality characterized by the depth of a user's investment when interacting with a digital system [27]. It typically results in positive outcomes [26]. This tool measures user engagement by summarizing an index that ranges from 0 to 5. It also provides detailed information about four dimensions of user engagement, i.e., Focused Attention (FA), Perceived Usability (PU), Aesthetic Appeal (AE), and Reward (RW). The last questionnaire was administered when the participants used all three explanation modalities. It evaluated the participant's satisfaction by asking them to rank the three modalities based on their Utility, Completeness, Understandability, and Helpfulness (from 1 to 3, 1 is the best) and to vote for the best visualization explaining why they preferred a modality over the others.

4.7 Results

Figure 4*(Top)* shows all tasks' errors and completion times for all the users. The results are organized by instance (rows) and visualizations (columns). For each cell of this grid, the chart shows each user's actual time to complete the three tasks (bar chart on the top, with a line showing the median completion time) and the number of errors (heatmap on the bottom). We use two distinct colors scales for the heatmaps. The LORE Output visualization (first column) uses grayscale to represent each task's absolute number of errors. The other two visualizations (columns 2 and 3) show the difference from the LORE Output errors using a divergent color scale: purples maps to better performance and oranges to higher errors. The tasks with no errors are denoted with a thicker black stroke. Figure 4*(Bottom)* reports the absolute errors for each task and each condition. In this case, each circle has a color proportional to the number of errors. Each circle reports the actual number for further details. Gray circles denote tasks where there are no errors. Although considered the easiest, the first instance shown to the user required a higher completion time. This might be because the user had to get acquainted with the different visualization strategies and gain the correct way to read the outputs. Among the three visualizations, *FIPER* appears to be the most time-consuming but with a significant improvement in the error rate. We can conclude from this observation that although *FIPER* is slightly more time-demanding than the other two visualizations, it performs better on all the tasks, even with the most difficult instances. From the user satisfaction questionnaire, *FIPER* visualization was chosen as the top preferred by 10 participants, 4 preferred the XAI Library visualization, and only one chose Lore simple output. *FIPER* was considered the most valuable visualization by 13 participants (the other two chose the XAI library). Overall the *FIPER* was considered more understandable by 10 participants.

The XAI library Visualization was appreciated by some of the participants for its conciseness, and it was pointed out that it may be more suited for datasets with a low number of features. A participant commented that *FIPER is the most easily readable visualization even though XAI Library Viz might be more immediate for certain questions.* This follows what was stated by [15] about

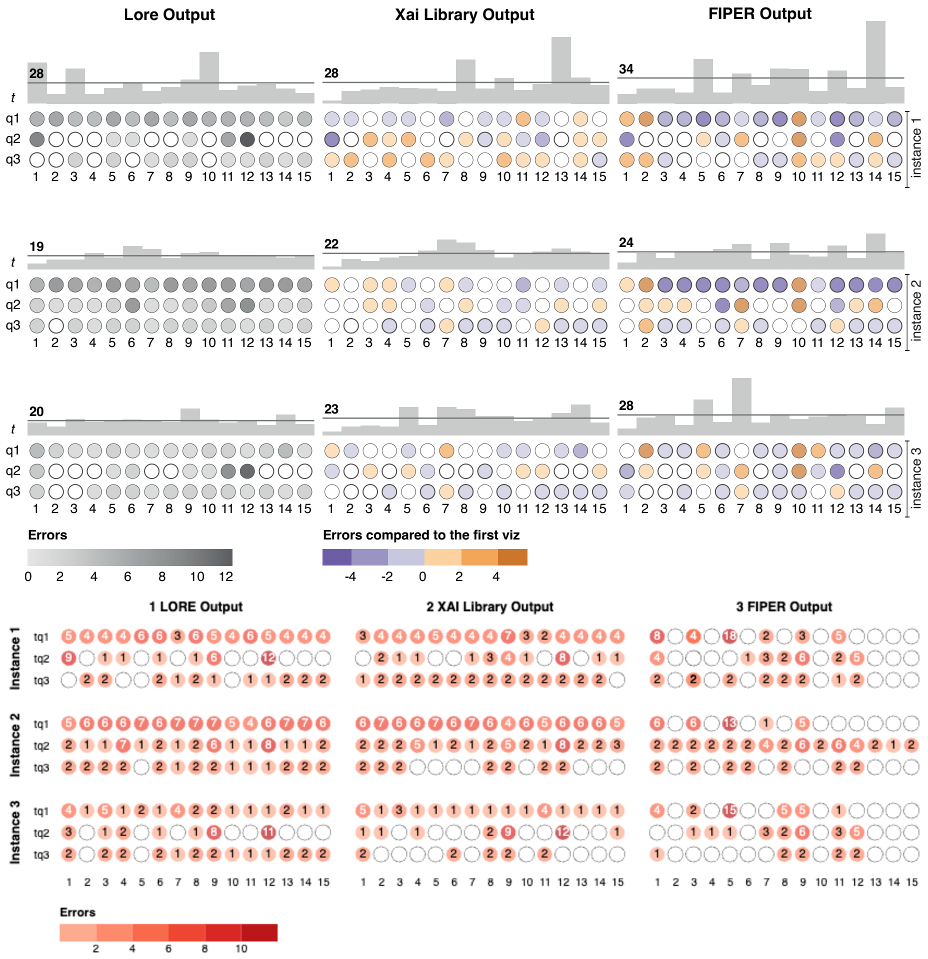


Fig. 4: Errors rate and completion times for tasks in each condition.*(Top)* The first column shows the absolute errors of the LORE output. The other two columns show the difference w.r.t. the first column, with divergent color scale to highlight increment or decrement in errors. *(Bottom)* Absolute number of errors for each output and each task

the completeness of explanations which is positively correlated with improved mental models and does not impair user experience or task.

5 Conclusions and future work

This research paper introduces a novel visual-based approach for rule representation. It evaluates its effectiveness compared to two existing textual-based approaches within the context of data science education. By integrating the rules with feature importance, we aim to enhance the user experience and promote a better understanding of the underlying prediction algorithm.

The preliminary findings of our study indicate that the visual-based approach demonstrates superior suitability for datasets containing a high number of features. Conversely, qualitative feedback suggests that the XAI library visualization may be more suitable for datasets with fewer attributes. It is worth noting that our initial testing of *FIPER* was conducted with data scientists; however, we intend to extend its application and customization to various scenarios, specifically targeting experts in different domains. Furthermore, *FIPER* offers users a certain degree of interactivity, allowing them to engage with the explanations provided.

This work is not free of limitations. Feature importance is used to sort predicates, even if it can contrast with the rule-based approach. However, it's worth noting that rule-based methods don't consistently provide a logical order when displaying rules and the order of presentation of the predicates may not be aligned with their relevance. Thus, our approach makes a design choice to exploit FI ranking in the visualization. The user retains the option to decide whether to apply this sorting, for example prioritizing the visualization of those features that are present in one of the predicates of the rule. A future development comprises the possibility for the user to select the FI method among a given set, allowing the users to confront different FI algorithms.

FIPER enables us to assess an instance's ranking within the distributions of individual features. We are enhancing the interface to permit instance editing, granting users the capacity to modify specific feature values and delve into the black box's response. We are also actively developing a *FIPER* version that visualizes counter rules, which are logical predicates based on features that result in an alternate classification of the chosen instance.

Eventually, for specific explainers, like LORE in our instance, a synthetic neighborhood is established around the instance to build the explanation. We are working to add a layer of presentation of the distribution of feature values within this neighborhood.

6 Acknowledgements

This work has been supported by the European Community Horizon 2020 programme under the funding scheme ERC-2018-ADG G.A. 834756 *XAI: Science and technology for the eXplanation of AI decision making*, by the European Union's Horizon Europe Programme under the CREXDATA project, grant agreement no. 101092749, by the Next Generation EU: NRRP Initiative, Mission 4, Component 2, Investment 1.3, PE0000013 – "Future Artificial Intelligence Research – FAIR" - CUP: H97G22000210007 and "SoBigData.it - Strengthening the Italian RI for Social Mining and Big Data Analytics" - Prot. IR0000013.

References

1. Abdul, A. M., Vermeulen, J., Wang, D., Lim, B. Y. & Kankanhalli, M. S. *Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda* (eds Mandryk, R. L., Hancock, M., Perry, M. & Cox, A. L.) 2018.
2. Andrienko, N. V., Andrienko, G. L., Adilova, L., Wrobel, S. & Rhyne, T. Visual Analytics for Human-Centered Machine Learning. *IEEE Computer Graphics and Applications* **42,** 123–133 (2022).
3. Cheng, H. F. *et al. Explaining Decision-Making Algorithms through UI: Strategies to Help Non-Expert Stakeholders* (eds Brewster, S. A., Fitzpatrick, G., Cox, A. L. & Kostakos, V.) 2019.
4. Chromik, M. & Butz, A. *Human-XAI Interaction: A Review and Design Principles for Explanation User Interfaces* (eds Ardito, C. *et al.*) 2021.
5. Chromik, M. & Schuessler, M. *A Taxonomy for Human Subject Evaluation of Black-Box Explanations in XAI* (eds Smith-Renner, A. *et al.*) 2020.
6. Danilevsky, M. *et al. A Survey of the State of Explainable AI for Natural Language Processing* (eds Wong, K., Knight, K. & Wu, H.) 2020.
7. Dua, D. & Graff, C. *UCI Machine Learning Repository* 2017.
8. Ehsan, U., Tambwekar, P., Chan, L., Harrison, B. & Riedl, M. O. *Automated rationale generation: a technique for explainable AI and its effects on human perceptions* (eds Fu, W., Pan, S., Brdiczka, O., Chau, P. & Calvary, G.) 2019.
9. Freitas, A. A. Comprehensible classification models: a position paper. *SIGKDD Explor.* **15,** 1–10 (2013).
10. Guidotti, R. *et al.* A Survey of Methods for Explaining Black Box Models. *ACM Comput. Surv.* **51,** 93:1–93:42 (2019).
11. Guidotti, R. *et al.* Factual and Counterfactual Explanations for Black Box Decision Making. *IEEE Intell. Syst.* **34,** 14–23 (2019).
12. Gunning, D. & Aha, D. W. DARPA's Explainable Artificial Intelligence (XAI) Program. *AI Mag.* **40,** 44–58 (2019).
13. Hart, S. G. *NASA-task load index (NASA-TLX); 20 years later* Sage publications Sage CA: Los Angeles, CA, 2006.
14. Hart, S. G. & Staveland, L. E. in *Advances in psychology* 139–183 (Elsevier, 1988).
15. Kulesza, T. *et al. Too much, too little, or just right? Ways explanations impact end users' mental models* (eds Kelleher, C., Burnett, M. M. & Sauer, S.) 2013.
16. Liao, Q. V., Gruen, D. M. & Miller, S. *Questioning the AI: Informing Design Practices for Explainable AI User Experiences* (eds Bernhaupt, R. *et al.*) 2020.
17. Liao, Q. V. & Varshney, K. R. Human-Centered Explainable AI (XAI): From Algorithms to User Experiences. *CoRR* **abs/2110.10790.** arXiv: `2110.10790` (2021).

18. Lim, B. Y. & Dey, A. K. *Toolkit to support intelligibility in context-aware applications* (eds Bardram, J. E., Langheinrich, M., Truong, K. N. & Nixon, P.) 2010.
19. Lipton, Z. C. The mythos of model interpretability. *Commun. ACM* **61,** 36–43 (2018).
20. Lundberg, S. M. & Lee, S. *A Unified Approach to Interpreting Model Predictions* (eds Guyon, I. *et al.*) 2017.
21. Madumal, P., Miller, T., Sonenberg, L. & Vetere, F. *A Grounded Interaction Protocol for Explainable Artificial Intelligence* (eds Elkind, E., Veloso, M., Agmon, N. & Taylor, M. E.) 2019.
22. Miller, T. Explanation in artificial intelligence: Insights from the social sciences. *Artif. Intell.* **267,** 1–38 (2019).
23. Miller, T., Howe, P. & Sonenberg, L. Explainable AI: Beware of Inmates Running the Asylum Or: How I Learnt to Stop Worrying and Love the Social and Behavioural Sciences. *CoRR* **abs/1712.00547.** arXiv: `1712.00547` (2017).
24. Ming, Y., Qu, H. & Bertini, E. RuleMatrix: Visualizing and Understanding Classifiers with Rules. *IEEE Trans. Vis. Comput. Graph.* **25,** 342–352 (2019).
25. Mucha, H., Robert, S., Breitschwerdt, R. & Fellmann, M. *Interfaces for Explanations in Human-AI Interaction: Proposing a Design Evaluation Approach* (eds Kitamura, Y., Quigley, A., Isbister, K. & Igarashi, T.) 2021.
26. O'Brien, H. L., Cairns, P. A. & Hall, M. A practical approach to measuring user engagement with the refined user engagement scale (UES) and new UES short form. *Int. J. Hum. Comput. Stud.* **112,** 28–39 (2018).
27. O'Brien, H. Theoretical perspectives on user engagement. *Why engagement matters: Cross-disciplinary perspectives of user engagement in digital media,* 1–26 (2016).
28. Preece, A. D., Harborne, D., Braines, D., Tomsett, R. & Chakraborty, S. Stakeholders in Explainable AI. *CoRR* **abs/1810.00184.** arXiv: `1810.00184` (2018).
29. Preece, J., Sharp, H. & Rogers, Y. *Interaction design: beyond human-computer interaction* (John Wiley & Sons, 2015).
30. Ribeiro, M. T., Singh, S. & Guestrin, C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier (eds Krishnapuram, B. *et al.*) 1135–1144 (2016).
31. Ribeiro, M. T., Singh, S. & Guestrin, C. *Anchors: High-Precision Model-Agnostic Explanations* in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018* (eds McIlraith, S. A. & Weinberger, K. Q.) (AAAI Press, 2018), 1527–1535.

32. Schaffer, J. *et al. Getting the Message?: A Study of Explanation Interfaces for Microblog Data Analysis* (eds Brdiczka, O., Chau, P., Carenini, G., Pan, S. & Kristensson, P. O.) 2015.
33. Sharp, H., Preece, J. & Rogers, Y. *Interaction Design: Beyond Human-Computer Interaction* (Wiley, 2019).
34. Sovrano, F., Vitali, F. & Palmirani, M. Making Things Explainable vs Explaining: Requirements and Challenges under the GDPR. *CoRR* **abs/2110.00758.** arXiv: `2110.00758` (2021).
35. Wang, D., Yang, Q., Abdul, A. M. & Lim, B. Y. *Designing Theory-Driven User-Centric Explainable AI* (eds Brewster, S. A., Fitzpatrick, G., Cox, A. L. & Kostakos, V.) 2019.
36. Yang, F., Huang, Z., Scholtz, J. & Arendt, D. L. *How do visual explanations foster end users' appropriate trust in machine learning?* (eds Paternò, F., Oliver, N., Conati, C., Spano, L. D. & Tintarev, N.) 2020.