[PREPRINT] Using Case-based Reasoning for Capturing Expert Knowledge on Explanation Methods

Jesus M. Darias, Marta Caro-Martínez,
Belén Díaz-Agudo, and Juan A. Recio-Garcia

Department of Software Engineering and Artificial Intelligence
Instituto de Tecnologías del Conocimiento
Universidad Complutense de Madrid, Spain
{jdarias,martcaro,belend,jareciog}@ucm.es

Abstract. Model-agnostic methods in (XAI) propose isolating the explanation system from the AI model architecture, typically Machine Learning or black-box models. Existing XAI libraries offer a good number of explanation methods, that are reusable for different domains and models, with different choices of parameters. However, it is not clear what would be a good explainer for a given situation, domain, AI model, and user preferences. The choice of a proper explanation method is a complex decision-making process itself. In this paper, we propose applying CBR to support this task by capturing the user preferences about explanation results into a case base. We have defined the corresponding CBR process to help retrieve a suitable explainer from a catalogue made of existing XAI libraries. CBR could help the task of learning from the explanation experiences and will help to retrieve explainers for other similar scenarios.

Keywords: XAI, Model Agnostic Models, Explanation Experiences

1 Introduction

Increasing understanding has become a requirement to trust in AI models applied to real-world tasks. Some ML models are considered intrinsically interpretable due to their simple structure, such as short decision trees, simple nearest neighbors, or sparse linear models. However, there is typically a black box nature and a lack of transparency associated with the best-performing models. This issue has triggered a new huge body of work on Explainable Artificial Intelligence (XAI), a research field that holds substantial promise for improving trust and transparency of AI-ML-based systems [11, 13, 22]. Methods for machine learning (ML) interpretability can be classified according to various criteria. Model-specific explanations are limited to specific model classes while model-agnostic methods can be used on any ML model and are applied after the model has been

trained (post hoc). Note that post hoc methods can also be applied to intrinsically interpretable models[14]. These model agnostic methods usually work by analyzing feature input and output pairs. By definition, these methods cannot have access to model internals such as weights or structural information. The main advantage of model-agnostic (post hoc) explanation methods is flexibility and re-usability, although some authors consider this type of explanation as limited *justifications* because they are not linked to the real reasoning process occurring in the ML model [2]. Another criterion is categorizing explainers as local or global. Local means that the method is applicable to explain an individual prediction, while global means that it is used for understanding the whole model learned from a certain dataset.

The background context of the research conducted in this paper is the *iSee project*[1] that aims to provide a unifying platform where personalized explanations are created by reasoning with Explanation Experiences using CBR. This is a very challenging, long-term goal as we want to capture complete user-centered explanation experiences on complex explanation strategies. We aim to be able to recommend what explanation strategy better suits an explanation situation. The contribution of this paper is the first step toward this long-term goal: we aim to capture user opinions on the preferred XAI method of a given AI model and domain. To do so, we have conducted an online experiment with several users to elicit a case base capturing their preferences regarding the explanation of several real use cases on explaining AI models. Moreover, we define the corresponding CBR system that exploits this knowledge to help users with the task of selecting an XAI method suitable for a concrete explanation scenario. The query describing the situation includes knowledge about the user expertise, the AI model and task, the data model, and the domain.

This paper runs as follows: Section 2 presents the background of this work. Section 3 describes examples of ML models and explanations with basic explainers to get a case base that captures real user preferences on explanations. Then, Section 4 describes the associated CBR system that exploits this knowledge. Section 5 presents the evaluation results and section 6 concludes the paper and opens lines of future work.

2 Background

There are several reusable model-agnostic methods that can be used on any ML model and are applied after the model has been trained (post hoc). Some relevant well-known examples are: Local Interpretable Model-Agnostic Explanations (LIME) [17], Anchors [18], Shapley Additive Explanations (SHAP) [12], Partial Dependence Plots (PDPs) [6], Accumulated Local Effects (ALE) [1] and counterfactual explanations [20]. An example of work that reviews different explanation techniques is the taxonomy proposed by Arya et al. [3]. In this work, the authors also propose Explainability 360, an open-source Python toolkit to

[1] http://isee4xai.com

implement explanation algorithms and metrics to measure them. Both resources, the taxonomy and the toolkit can help users to decide what the best implementation is to explain a specific model. In our previous work [4] we have reviewed some selected XAI libraries (Interpret, Alibi, Aix360, Dalex, and Dice) and provide examples of different model-agnostic explanations. Our work in this paper proposes using CBR to retrieve the best explainer because, even if these methods are reusable, the choice of the most suitable explanation method for a given model is a complex task where expertise is a major requirement. Moreover, one of the most original aspects of our work is that the selected explanation strategy is obtained according to users' opinions, which can enhance the performance of the CBR system since an explanation's effectiveness depends on users' opinions directly.

There are other approaches in the CBR literature related to XAI. Some relevant early works can be found in the review by Leake and McSherry [9]. In the work by Sørmo et al. [19], authors present a framework for explanation in CBR focused on explanation goals, whereas the publication by Doyle et al.[5] develops the idea of explanation utility, a metric that may be different to the similarity metric used for nearest neighbor retrieval. Recently, there is a relevant body of work on CBR to explain other black-box models, the so-called *CBR Twins*. In the paper by Keane et al.[8], the authors propose a theoretical analysis of a post-hoc explanation-by-example approach that relies on the twinning of artificial neural networks with CBR systems. The work by Li et al. [10] combines the strength of deep learning and the interpretability of case-based reasoning to make an interpretable deep neural network. The paper by Gates et al. [7] investigates whether CBR competence can be used to predict confidence in the outputs of a black box system when the black box and CBR systems are provided with the same training data. In the publication by Weber et al. [21], the authors demonstrate how CBR can be used for an XAI approach to justify solutions produced by an opaque learning method, particularly in the context of unstructured textual data. Additionally, CBR has been proven as a suitable strategy to select the most suitable explanation method for a given model outcome. This way, our previous work has analysed its applicability to select explanation methods for image classification tasks [16] and to configure these explanation methods with an optimal setup [15].

3 Case-based elicitation

The first contribution of this paper is the elicitation of a case base capturing user preferences on the explanation of existing ML models. To acquire this knowledge we have generated several use cases reproducing real XAI scenarios where, given an ML task, several alternative explanation methods are applied. Then users are asked to select the best explanation according to their expertise and expectations. This section describes the structure of the cases and the elicitation process to collect user knowledge on selecting a proper explanation method.

3.1 Case structure

Each case is structured as a tuple $\langle D, S, R \rangle$ containing: (1) a description D of the ML model to be explained; (2) a solution S, that describes the explanation method (or explainer); and (3) a result R, which is the opinion (score) of the users about how good the solution is for this specific description. The case description D includes:

Domain: the domain is the situation where the AI model and the explanation system are applied. We have defined some domains: Medicine, Economics, Social, Security, Entertainment, and Image Recognition, although our model is extensible to others.
DataType: the type of data that the *Explainer* accepts as input. It can be text, images, or tabular data.
TrainingData: if the training data of the model is available to feed the explainer methods.
AITask: the artificial intelligence task we want to make interpretable for users. In the examples, we have only used classification and regression tasks, although it is extensible to other AI tasks such as computer vision, information retrieval, robot control, natural language processing, etc.
ModelBackend: the library or technology used to implement the AI model: python libraries $Sklearn$, $Torch$ and $TensorFlow$.
ModelType: The AI model we use to carry out the $AITask$, for example, an artificial neural network (ANN), Random Forest (RF), support vector machine (SVM), and so on.
DomainKnowledgeLevel: the level of knowledge of the target user about the $AITask$, and the $Domain$. It can be low or expert.
MLKnowledge: if the user has some knowledge of machine learning. This is a yes or no attribute.
ExplanationScope: if the target explanations are global, explaining the whole AI model, or local, explaining a single prediction.

According to all these features, we have an *explainer* as a solution S, which will fit the problem described and will be able to generate explanations. The third component in the case structure is the result R, a score that represents the users' opinions about this solution through a 7-point Likert scale. In Table 1, we show an example of a case where we can see its description, its solution, and its user score.

3.2 Case base acquisition

We have elicited a case base with real user input. We elaborated a series of use cases where an AI model was applied to solve an AI task. For each case, the user rates different explanations generated by several alternative explainers. One of the advantages of using use cases is that the users do not need to interact directly with the system and worry about aspects such as parameter configuration. In addition, we provide users with the background needed about the case and the

Description	Domain	Economics
	DataType	Tabular
	TrainingData	Yes
	AITask	Regression
	ModelBackend	Sklearn
	ModelType	RF
	DomainKnowledgeLevel	Low
	MLKnowledge	Yes
	ExplanationScope	Global
Solution		Tabular/Importance
UserScore		6

Table 1. Example of the structure of a case in our case base. This case is related to use case 3, described in section 3.

description of the explainers being applied. The purpose of each use case is to ask the user about their degree of satisfaction with the explanations proposed using a Likert scale (from 1 to 7).

Additionally, for each use case, two questions were included to do basic profiling of the user where we represent their knowledge in the specific domain and their expertise in machine learning. With this information, each user's answer to a specific explanation is represented as a case. This way, the description of the use case is associated with the explanation method (as its solution) and the scores given by the users are the result of the case. Next, we describe the use cases presented to the participants according to the case structure described in the previous section.

Use case 1: cervical cancer prediction[2]. The *Domain* of this use case is Medicine, one of the most critical domains to apply AI prediction tasks and where explanation systems are required. In questionnaire 1, we have random forest and neural networks as the *ModelType* to classify (*AITask*) the high risk of having cervical cancer. These models consider features, represented as tabular data (*DataType*), like the age of the individuals, sexual partners, and the number of pregnancies, among others. We have some explanation methods to justify the model behaviors (i.e., different *Solutions*): some models with global scope (*ExplanationScope*), like Variable Importance and Accumulated Local Effects, and some local models (*ExplanationScope*) like LIME, SHAP, Anchors, and DiCE.

Use case 2: depression screening[3]. The second use case is a problem of the psychology field (Medicine *Domain*). In this problem, we try to explain why a

[2] https://forms.gle/ctJZx53wRhTb7hMf8
[3] https://forms.gle/2jYBkWgNcWjNKRLs6

machine learning model (*ModelType*) predicts depression in students (classification *AItask*). The models use tabular data (*DataType*) collected through a questionnaire. We have also some explanation methods (*Solutions*) to understand this task: Variable Importance, Accumulated Local Effects (global *ExplanationScope*), LIME, SHAP, and Anchors (local *ExplanationScope*).

Use case 3: cost prediction[4]. It predicts the price per square meter of apartments in Poland. We can consider the *Domain* of this case as Economics. The main goal is the prediction of product prices. Both employers and consumers need to know how artificial intelligence works to avoid mistakes in prediction. In the cases related to this questionnaire, we use random forest (*AI Task*) to solve this regression problem (*AITask*) that considers tabular data (*DataType*) describing attributes about the apartment: surface, floor, location, and others. We have proposed some explanation methods (*Solutions*) to try to understand the random forest working: Variable Importance, Accumulated Local Effects (global *ExplanationScope*), LIME, and SHAP (local *ExplanationScope*).

Use case 4: income prediction[5]. The *AITask* to solve is a classification to predict if a person earns more than 50K dollars a year. We use ML models (*ModelType*) that use *TrainingData*, using variable importance, Accumulated Local Effects (global *ExplanationScope*), LIME, SHAP or public DiCE (local *ExplanationScope*), as the proposed explanation methods (*Solutions*). If the machine learning model does not use *TrainingData*, we have proposed to use private DiCE, which is also local. The *Domain* of this problem was labeled as Economics.

Use case 5: fraud detection[6]. We show the results of some explanation methods when applying a ML model (*ModelType*) to classify (*AITask*) if a transaction was fraudulent. The model is trained using tabular data (*DataType*). We have different *Solutions* to explain these machine learning models: variable importance, Accumulated Local Effects (global *ExplanationScope*), LIME, SHAP, and Anchors (local *ExplanationScope*). The *Domain* of these cases is set to Security. The main goal of this domain is fighting against vulnerabilities in critical systems.

Use case 6: social problems identification[7]. This questionnaire is related to Social *Domain*. Artificial intelligence can be applied to solve and detect very important social problems, for instance, alcohol consumption in young people, or discrimination based on race or sex. To make the artificial intelligence model transparent is necessary to understand how to fight against all these problems. Variable importance, Accumulated Local Effects (global *ExplanationScope*), LIME and SHAP (local *ExplanationScope*) are the explanation methods (*Solutions*) proposed in this questionnaire. They try to explain the behavior of machine learning models (*ModelType*) applied to a

[4] https://forms.gle/Kc91FWF9gKgg5yfS6
[5] https://forms.gle/KHXTGbJydXHAHH2p6
[6] https://forms.gle/mFe9ccVhZiLEjk4u6
[7] https://forms.gle/mFe9ccVhZiLEjk4u6

regression problem that tries to predict the final grades of Portuguese students (*AITask*). Tabular *DataType* as the student's school, her age, study time, etc are used by the machine learning methods.

Use case 7: text classification[8]. This model classifies a newsgroup post in a topic (*AITask*). The machine learning model (*ModelType*) only uses the text (*DataType*) from the post to classify it in religion, autos, baseball, among other topics. Therefore, the *Domain* of this problem is Entertainment. Although this is a domain not as critical as some of the previous ones, explanations in the entertainment industry have many advantages, for example, increasing the user's acceptance and satisfaction, or even persuading users to consume new products. The *Solution* we have for this use case is LIME (local *ExplanationScope*).

Use cases 8, 9, 10: image recognition[9] [10] [11]. These three questionnaires are included in the *Domain* Image Recognition. This domain is specific for understanding the prediction of the objects that appear in images. In these questionnaires we have toy examples related to the classification (*AITask*) of images (*DataType*) using machine learning (*ModelType*). In the questionnaire 8, we present several *Solutions* for image classifications of animals: LIME and Anchors (local *ExplanationScope*). In the questionnaire 9, we have Anchors and Counterfactuals (local *ExplanationScope*) to classify black and white images of clothes. Finally, in the last questionnaire, we have Counterfactuals as the *Solution* to explain black and white images of handwritten digits classification.

Figure 1 shows a screenshot of one of the use cases created to collect user preferences regarding the explanation methods. Next, we detail the CBR process that exploits this knowledge to propose the most suitable XAI method for a given explanation scenario.

4 CBR process

One of the most important aspects of a CBR system is how similar cases are retrieved. In our system, the proposed retrieval function can be decomposed into two steps: filtering and sorting. Given a case description D, the filtering step takes into account certain attributes that allow identifying the explanation methods that are compatible with that case. Namely, these attributes are *DataType, TrainingData, AITask, ModelBackend, ModelType* and *ExplanationScope*. This filter guarantees that all the retrieved explainers are valid solutions. For example, suppose we have a random forest regressor that works with tabular data, and we want an explanation for an instance. Just by using the *DataType*

[8] https://forms.gle/KitNg2FnkTbuL3KR6
[9] https://forms.gle/MCtagTCMB9jiFdgk6
[10] https://forms.gle/YHYga6d9eqLVFvsh7
[11] https://forms.gle/tZxzH8ZyY3VejhVv7

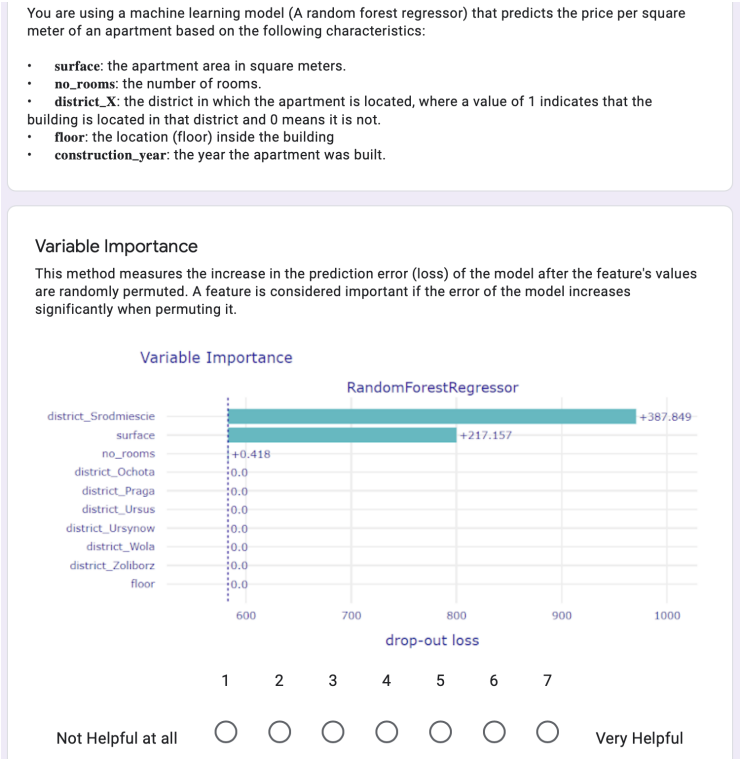

Fig. 1. Screenshot of use case 3. It describes to the user the AI task and proposes several alternative explanation methods to understand the corresponding model.

attribute, only the tabular explainers are retrieved. Since the *AITask* is regression, explainers that only work with classification, such as counterfactuals, will be discarded. Finally, since we want explanations for an instance, the *Explanation-Scope* will be local, and thus the final retrieved explainers will be *Tabular/LIME* and *Tabular/SHAP*.

During the initial filtering, we use the case description attributes so only the compatible explainers are returned. The solutions of these compatible cases contain different potential explainers to solve (explain) the query. We denote the set of cases sharing the same explainer as a solution with $\mathcal{C}^S$. At this point, some solutions (explanation methods) may be more suitable than others for a particular query. For this reason, the sorting phase arranges the cases in ($\mathcal{C}^S$) according to the following similarity metric:

$$sim(q, c) = \frac{1}{W} \sum_{a \in SimAttr} w_a \cdot equal(q(a), c(a))$$

where *SimAttr* represents the following attributes of the case description: *Domain*, *DomainKnowledgeLevel*, *MLKnowledge*, *AItask*, *ModelBackend*, and *Mod-*

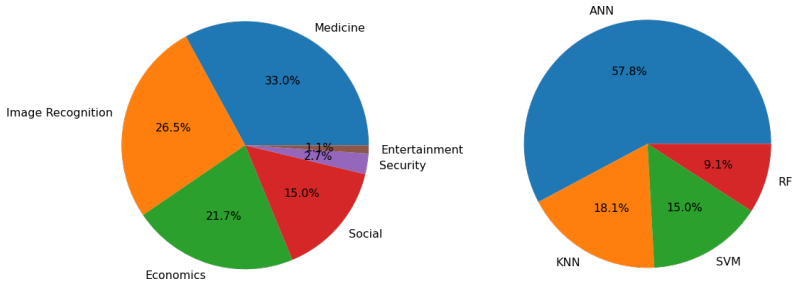

Fig. 2. Stratified analysis of cases per domain (left) and ML model (right).

elType. The values $w_a \in [0..1]$ are weights that have been computed to obtain the minimum error using a greedy optimization method, and $W = \sum w_a$.

Once the similarity values are calculated for all the cases in $\mathcal{C}^S$, the score (R) of the k most similar cases to the query are averaged to obtain the *Mean Estimated Explanation Utility Score*. This score represents the expected user satisfaction for that explainer. The same process applies for the rest of the compatible solutions: all cases with that explainer as *Solution* are retrieved creating the $\mathcal{C}^S$ set, their similarity values are calculated, and then the scores of the most similar cases are aggregated to determine the estimated utility value. We do not apply a filter considering a minimum number of agreeing with reviews to include a case because doing the mean we are taking into account all the opinions, and not only the majority opinion. Lastly, the explainers are ranked according to the expected score to be proposed to the user.

5 Evaluation and Discussion

The case base includes a total of 746 cases, where users evaluated the explanations for the domain models using 11 different explanation methods from some libraries. 30 different users participated by filling in the questionnaires with different skills in machine learning. Particularly, only 10.7% of cases referred to a user with no previous knowledge of machine learning. Although with a greater amount of responses we could get better results, we think this amount is enough to consider the system trustworthy.

The stratified analysis of the cases regarding the application domain and ML model is shown in Figure 2. The domains with the most cases are Medicine and Image recognition. There is a greater number of cases for these domains because there were more use cases associated with them. Regarding the distribution of models, there is a majority of cases applied to artificial neural networks (ANN).

The analysis of cases for each explainer and datatype is presented in Figure 3. Regarding the explainers, all the solutions are guaranteed to be valid methods for their case descriptions thanks to the filtering step of the retrieval phase that was applied when elaborating the questionnaires to evaluate the use cases. As expected, global methods, that can be applied in almost any case involving

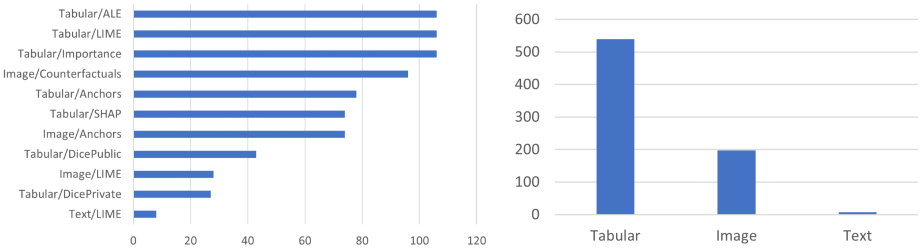

Fig. 3. Stratified analysis of the number of cases per explainer (left) and data type (right).

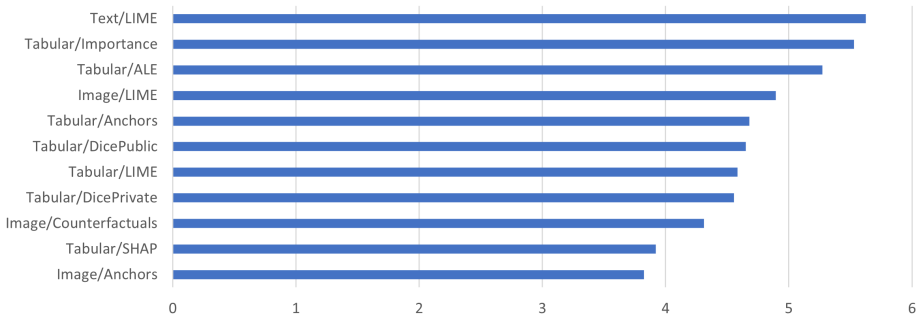

Fig. 4. Average user score by explainer.

tabular data (ALE and Feature Importance), have more cases than most of the local explainers. Although logical, it is important to note that the number of cases per explainer is directly proportional to the number of cases of the data type such explainer uses. It is also worth noting that there are few cases representative of the text data type, mainly because we only used one method to generate explanations (LIME), but also because the Entertainment domain, where this type of explainer was used, did not count with enough cases.

In Figure 4, we present the average user score of each explainer method. It is worth emphasizing that the scores assigned by the user go from 1 to 7. Although LIME for text data has the higher score, this result is not reliable since the number of cases where this explainer was used is too little. However, the tabular global methods, ALE, and Variable Importance, were the next best-rated explainers with an average score of 5.53 and 5.27, respectively. This does not mean that users disliked local explainers. In fact, most of the local explainers have an average score above the neutral mark of 4. However, explainers such as SHAP for tabular data and Anchors for images did not receive good ratings in general.

The same pattern is identified upon analyzing the mean scores per explainer grouping by the different domains in Table 2. Again, the global methods ALE and Variable Importance are the preferred ones by the users in most of the do-

Domain	Solution	UserScore
Social	Tabular/ALE	6.07
	Tabular/Importance	5.89
	Tabular/LIME	5.25
	Tabular/SHAP	4.50
Security	Tabular/Importance	4.80
	Tabular/Anchors	3.60
	Tabular/LIME	3.40
	Tabular/ALE	3.20
Medicine	Tabular/Importance	5.34
	Tabular/LIME	4.89
	Tabular/ALE	4.82
	Tabular/Anchors	4.56
	Tabular/DicePublic	3.93
	Tabular/SHAP	3.56
Image Recognition	Image/LIME	4.89
	Image/Counterfactuals	4.31
	Image/Anchors	3.82
Entertainment	Text/LIME	5.62
Economics	Tabular/ALE	5.59
	Tabular/Importance	5.59
	Tabular/Anchors	5.07
	Tabular/DicePublic	5.07
	Tabular/DicePrivate	4.55
	Tabular/LIME	3.59

Table 2. Mean score per explainer by domain.

mains. However, there is an exception where ALE is the worst-rated explainer in the Security domain. One possible explanation for this is that in the only model explained in this domain the meaning of the features was not provided because of data privacy reasons. The low score given by the users may imply that this method is not particularly helpful when the intrinsic meaning of the attributes is unknown. Nevertheless, this is one of the domains with the lowest number of cases in the case base, so the standard deviation is considerably higher. Regarding the local tabular methods, LIME, DiCE (counterfactuals), and Anchors were preferred over SHAP in all the domains. As for the image explainers, LIME was the better-rated explainer, followed by image counterfactuals. Anchors for images did not prove to be helpful for the users and its average score fell below the neutral mark of 4.

A similar outcome is obtained when grouping by the AI Task, as shown in Table 3. However, it is worth noting that the same explainers obtained a considerably higher score when used for regression tasks than for classification. One reason for this may be that the regression models proposed in the questionnaires are easier to interpret than the classification ones since the value of a feature is

AITask	Solution	UserScore
Regression	Tabular/ALE	6.07
	Tabular/Importance	5.89
	Tabular/LIME	5.25
	Tabular/SHAP	4.50
Classification	Text/LIME	5.62
	Tabular/Importance	5.39
	Tabular/ALE	4.98
	Image/LIME	4.89
	Tabular/Anchors	4.67
	Tabular/DicePublic	4.65
	Tabular/DicePrivate	4.55
	Tabular/LIME	4.34
	Image/Counterfactuals	4.31
	Image/Anchors	3.82
	Tabular/SHAP	3.56

Table 3. Mean user score per explainer by AI task.

proportional to the predicted value, while classification models work with probabilities. However, it is worth pointing out that the sample size was not large enough as only two of the models presented in the questionnaires were regressors.

In Figure 5, we analyze the mean score given to the explainers depending on the previous domain knowledge of the users. One of the main aspects is that expert users in the proposed domains tend to evaluate the explainers more positively. Although there seems to be a greater disparity for image explainers, it is important to highlight that only one user claimed not to have knowledge in the Image Recognition domain, so it would be incorrect to make interpretations about the suitability of this explainer solely for expert users. However, in domains involving tabular data, the number of users with little knowledge about the domain is more similar to the number of expert users. Thus, the results obtained are more reliable and although the score distance between these types of users is lower, users with low domain knowledge give lower scores to the proposed explanations.

Lastly, we have used cross-validation to evaluate the performance of the CBR system. From the original case base, 15% of the cases were used as the test set, and the rest represented the case base used by the CBR system. Since each case from the test set was composed of the case description, solution, and user score (from 1 to 7), we calculated the predicted score of that explainer by feeding the case description to the retrieval function. This process was repeated using all the cases to obtain the mean error. In Figure 6, the absolute error distribution is displayed. The mean absolute error was 1.03 with a standard deviation of 0.83.

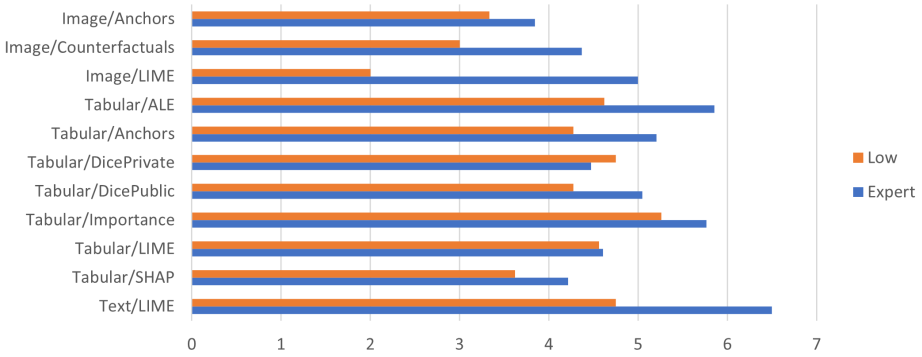

Fig. 5. Mean user score per explainer according to the users' knowledge of the domain.

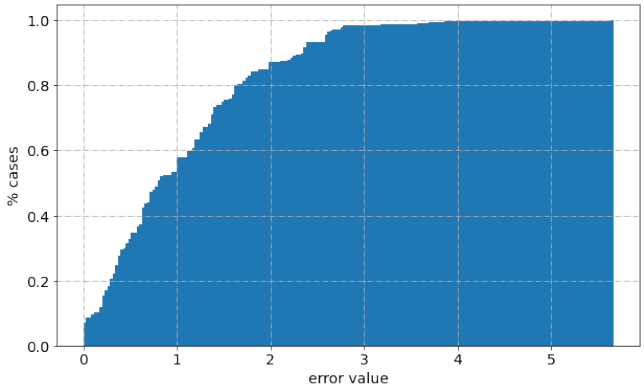

Fig. 6. Cumulative histogram displaying the absolute error distribution of the CBR system. The y-axis represents the percentage of cases (value 1.0 is equal to 100%) having the error value represented by the x-axis.

6 Conclusions

When it comes to interpretability, one of the limitations of an AI engineer is selecting a method to explain the behavior of a particular model, identifying which of those tools produce the most meaningful explanations for users. This is a clear example of a user-centered task, where we propose applying CBR to capture and reuse the expert's knowledge.

In this paper, we have built a simple CBR system that aims to recommend the most suitable explanation method for ML models from different domains and users. By specifying the intrinsic characteristics of the model such as the data type it works with, the task it achieves, and the ML architecture that was used to build it, the compatible explainers are easily filtered in the retrieval phase of our CBR system. Nonetheless, the fact that an explainer method is compatible with a certain model does not mean it will yield good results.

In our previous work [4] we concluded that one of the greatest disadvantages of the available XAI libraries was the lack of personalization of the explanations. However, by identifying the explanation methods that users consider helpful, it is easier to identify the factors that come into play to make a certain explanation better than others in a specific situation. For this reason, we collected a case base gathering user feedback on numerous explanation methods applied to ML models from different domains. The collected cases let us conclude valuable insight regarding preferred explainers given the domain or AI task. However, we are aware that CBR systems are live systems and that XAI is a research field in continuous change, therefore our case base has to be updated with new explainers adapted to new problems and solutions as they arise. This paper is the first step in a very challenging, long-term goal as we want to capture complete user-centered explanation experiences on complex and combined explanation strategies. We are defining an ontology to help with the knowledge-intensive representation of previous experiences, different types of users and explanation needs, characterization of the data, the black-box model, and the contextual properties of the application domain and task. In future work, we will use this ontology to improve user modelling and to provide personalized explanations that suit the needs of the person receiving them, by modelling user intentions and needs. In this way, it will be possible to merge the already existing explainability methods with a user-oriented approach.

Acknowledgements This research is a result of the Horizon 2020 Future and Emerging Technologies (FET) programme of the European Union through the iSee project (CHIST-ERA-19-XAI-008, PCI2020-120720-2) funded by MCIN/AEI and European Union "NextGenerationEU"/PRTR".

References

1. Apley, D.W., Zhu, J.: Visualizing the effects of predictor variables in black box supervised learning models. Journal of the Royal Statistical Society **82**(4), 1059–1086 (2020)
2. Arrieta, A.B., et al.: Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information fusion **58**, 82–115 (2020)
3. Arya, V., et al.: Ai explainability 360: An extensible toolkit for understanding data and machine learning models. J. Mach. Learn. Res. **21**(130), 1–6 (2020)
4. Darias, J.M., et al.: A systematic review on model-agnostic xai libraries **3017**, 28–39 (2021)
5. Doyle, D., et al.: Explanation oriented retrieval. In: ECCBR. pp. 157–168 (2004)
6. Friedman, J.H.: Greedy function approximation: a gradient boosting machine. Annals of statistics pp. 1189–1232 (2001)
7. Gates, L., et al.: Cbr confidence as a basis for confidence in black box systems. In: ICCBR. pp. 95–109 (2019)
8. Keane, M.T., et al.: How case-based reasoning explains neural networks: A theoretical analysis of xai using post-hoc explanation-by-example from a survey of ann-cbr twin-systems. In: ICCBR. pp. 155–171 (2019)

9. Leake, D., Mcsherry, D.: Introduction to the special issue on explanation in case-based reasoning. The Artificial Intelligence Review **24**(2), 103 (2005)
10. Li, O., et al.: Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions. In: AAAI Conference on AI. vol. 32 (2018)
11. Lipton, Z.C.: The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. Queue **16**(3), 31–57 (2018)
12. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. Advances in neural information processing systems **30** (2017)
13. Miller, T.: Explanation in artificial intelligence: Insights from the social sciences. Artificial intelligence **267**, 1–38 (2019)
14. Molnar, C.: Interpretable Machine Learning. 2 edn. (2022), https://christophm.github.io/interpretable-ml-book
15. Recio-García, J.A., et al.: Cbr-lime: a case-based reasoning approach to provide specific local interpretable model-agnostic explanations. In: ICCBR. pp. 179–194 (2020)
16. Recio-García, J.A., et al.: A case-based approach for the selection of explanation algorithms in image classification. In: ICCBR. pp. 186–200 (2021)
17. Ribeiro, M.T., et al.: "Why should i trust you?" Explaining the predictions of any classifier. In: ACM SIGKDD. pp. 1135–1144 (2016)
18. Ribeiro, M.T., et al.: Anchors: High-precision model-agnostic explanations. In: AAAI conference on AI. vol. 32 (2018)
19. Sørmo, F., et al.: Explanation in case-based reasoning–perspectives and goals. Artificial Intelligence Review **24**(2), 109–143 (2005)
20. Verma, S., et al.: Counterfactual explanations for machine learning: A review. arXiv preprint arXiv:2010.10596 (2020)
21. Weber, R.O., et al.: Investigating textual case-based xai. In: ICCBR. pp. 431–447 (2018)
22. Weld, D.S., Bansal, G.: The challenge of crafting intelligible intelligence. Communications of the ACM **62**(6), 70–79 (2019)