Artificial Intelligence Research and Development
A. Cortés et al. (Eds.)

doi:10.3233/FAIA220357

Fuzzy-LORE: A Method for Extracting Local and Counterfactual Explanations Using Fuzzy Decision Trees

Najlaa MAAROOF [a,1], Antonio MORENO [a] Mohammed JABREEL [a] and Aida VALLS [a]

[a] *ITAKA-Intelligent Technologies for Advanced Knowledge Acquisition - Departament d'Enginyeria Informàtica i Matemàtiques Universitat Rovira i Virgili, Tarragona, Catalonia, Spain*

Abstract. Classification systems based Machine Learning hide the logic of their internal decision processes from the users. Hence, post-hoc explanations about their predictions are often required. This paper proposes Fuzzy-LORE, a method that generates local explanations for fuzzy-based Machine Learning systems. First, it learns a local fuzzy decision tree using a set of synthetic neighbours from the input instance. Then, it extracts from the logic of the fuzzy decision tree a meaningful explanation consisting of a set of decision rules (which explain the reasons behind the decision), a set of counterfactual rules (which inform of small changes in the instance's features that would lead to a different outcome), and finally a set of specific counterfactual examples. Our experiments on a real-world medical dataset show that Fuzzy-LORE outperforms prior approaches and methods for generating local explanations.

Keywords. Explainable AI (XAI), Machine Learning, Fuzzy Decision Tree, Diabetic Retinopathy, LORE

1. Introduction

Machine Learning (ML)-based systems have become a vital component of multiple applications in many domains, especially in healthcare. One key reason for their widespread adoption is the success in the development of accurate classification models, which help doctors in the diagnosis, treatment and prognosis of complex diseases. One example of those successful classification methods is the Fuzzy Random Forest (FRF), which can be used to solve multi-class or binary classification problems. A FRF is composed of hundreds of Fuzzy Decision Trees (FDTs) [1].

Despite their accuracy, most modern ML-based systems are considered black boxes, because it is not straightforward to understand the reasons behind their

[1] Corresponding Author: Najlaa Maaroof. E-mail: najlaamaaroofwahib.al-ziyadi@urv.cat

decisions. As a result, developing methods for explaining them has become highly demanded [2, 3].

Several kinds of explanation methods have been proposed in the literature. A popular approach is to use post-hoc explanation methods, which study the relationship between the input and the output produced by the system to extract a local explanation of a particular decision on an instance x. Most of the methods that follow this approach generate a set of inputs (neighbours of x), analyse the answers provided by the system to be explained and then create a simpler model from which a local explanation can be inferred [4, 5].

One of the most well-known post-hoc explanation methods is *Local Rule-Based Explanations* (LORE, [6]), explained briefly in section 2. LORE derives a rule-based explanation composed of the activated rule used to explain the rationale behind the system's decision, and a set of counterfactual rules which represent the minimal number of changes in the feature values of the instance that would change the conclusion of the system. Such counterfactual explanation is useful in domains like healthcare. It helps practitioners to decide what they should do to obtain a desired state instead of providing them only with important features that led to the decision.

Although LORE has shown a good performance in explaining classical ML-based systems [6], we believe that it can be improved for the particular case of fuzzy-based systems. In our previous works [7, 8] we proposed two extensions of LORE, called *Guided-LORE* and *C-LORE-F*. In the former the neighbours' generation step was formalized as a search problem and solved using Uniform Cost Search, whereas in the latter the knowledge about the definition of the fuzzy variables was used to focus the exploration of the neighbours' space. Such adaptations allowed us to make the generation process more informed and leverage more contextual information, mainly in the case in which the attributes that define the objects are fuzzy, covered in *C-LORE-F*.

Despite the promising outcome obtained with Guided-LORE and C-LORE-F, they still have some shortcomings. First, the quality of the obtained counterfactual instances should be improved [7]. Second, the basic explanation in LORE (and its variants) is limited to a single rule derived from the activated path in a decision tree, which is not very informative. Third, the method is quite rigid and the explanation can't be adapted to different applications or user types.

In this work, we propose a novel method called Fuzzy-LORE to address the shortcomings of standard LORE-based methods (i.e., LORE, Guided-LORE and C-LORE-F) and provide better explanations in the case of fuzzy-based ML systems. Fuzzy-LORE adapts our previous LORE-based methods by using fuzzy decision trees as an alternative to the classical decision trees. First, it learns a local fuzzy decision tree predictor on a synthetic neighbourhood of the instance x to be explained. Then, it extracts from the logic of the fuzzy decision tree a meaningful explanation consisting of a set of decision rules, a set of counterfactual rules, and a set of counterfactual examples. We will focus only on binary classification.

We evaluated the proposed method on a private dataset, used to train a FRF-based binary classifier that assesses the risk of developing diabetic retinopathy in diabetic patients. The experimental results show that, according to several metrics, Fuzzy-LORE outperforms the prior classical LORE-based methods, mainly in the generation of counterfactual examples.

The rest of this article is structured as follows. Section 2 provides an overview of the classical LORE-based methods. Section 3 explains the proposed method. In Section 4, we describe the experimental setup and discuss the obtained results. Finally, in section 5, we conclude the paper and list some points for future work.

2. Preliminaries

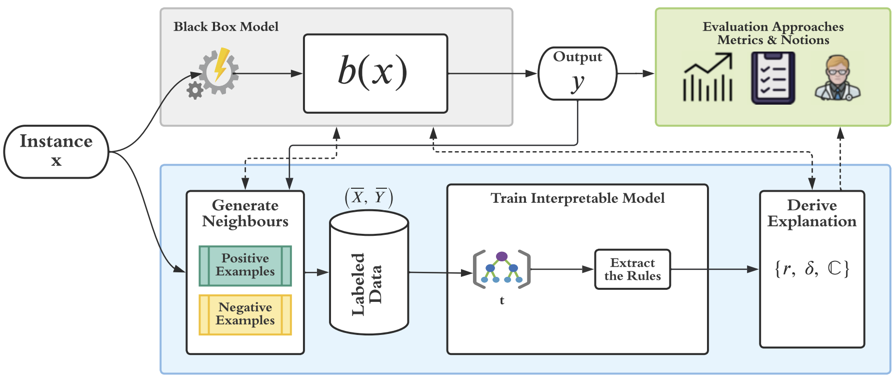

Figure 1. Architecture of the LORE-based explanation methods.

This section provides a brief background of the classical LORE-based methods, which use decision trees to provide a post-hoc explanation for the decision assigned to a specific instance. The inputs of the LORE-based method are a trained ML model, b, and an example x. Figure 1 shows the main steps and the general architecture of LORE. First, b is applied to x to get a decision y. Then, we obtain a set of neighbours of x, $\mathcal{D}$, and a rule-based model t (a decision tree) is built by considering the output of b in these points. From this model t it is possible to derive an explanation that contains the rule r used to classify x, a set of counterfactual rules δ and a set of counterfactual instances $\mathbb{C}$.

The set $\mathcal{D}$ is obtained by merging two subsets, $\mathcal{D}^+$ and $\mathcal{D}^-$. The first one is called the *positive set*, and it contains a set of instances that belong to the same class than x. The second one, the *negative set*, contains examples with a different class. We obtain $\mathcal{D}^-$ by looking at an auxiliary set T and finding the closest example to x , i.e., x^-, that has a different label than y. T can be the training set used to train the black-box model b, if accessible, or any other data set from the same distribution.

The main difference between LORE and its extensions Guided-LORE and C-LORE-F, lies in the neighbours' generation step. LORE uses a genetic algorithm to do it, whereas our methods define the neighbourhood generation as a search problem in which we explore the neighbourhood space of a point x by applying a Uniform Cost Search.

Once the synthetic neighbours, i.e., $\mathcal{D}$ have been obtained, a decision tree predictor, t, is trained on $\mathcal{D}$. Finally, the explanation is derived from t by finding the activated path and converting it to a rule format. Then, the counterfactual rules and their corresponding counterfactual instances are extracted from the analysis of t.

As stated in the introduction, the replacement of the decision tree by a fuzzy decision tree requires changes in the tasks of extracting the decision rules, the counterfactual rules and the counterfactual instances, which are detailed in the next section.

3. Fuzzy-LORE

Fuzzy-LORE is an adaptation of the LORE explanation methods that employs a FDT local predictor to obtain richer linguistic explanations of binary classification systems based on fuzzy variables. Following the LORE scheme, Fuzzy-LORE first generates synthetic neighbours of the instance of interest, using the method proposed in C-LORE-F [8]. This mechanism utilizes contextual information from the definitions of the fuzzy sets of the attributes. Fuzzy-LORE uses fuzzy tools in all of its steps, from generating the neighbours to constructing the FDT local predictor model and obtaining the explanation components. The next subsections explain in detail all the steps of Fuzzy-LORE.

3.1. FDT Construction

The first novelty in Fuzzy-LORE is the construction of a local interpretable model consisting of a small fuzzy decision tree. The fact that the tree uses the same linguistic variables than the fuzzy black-box model will help to interpret the explanations. Fuzzy-LORE uses the induction algorithm proposed in [9] to construct a local fuzzy decision tree with conjunctive rules. This algorithm is an adaptation of ID3 for fuzzy datasets with linguistic variables. It uses two parameters during the construction process. The first one is called the significance level, α, which filters out the evidence that is not relevant enough. The second parameter is the truth level threshold, β, which controls the tree's growth as it defines the minimum level for ending a branch. In the experimental section, the values $\alpha = 0.1$ and $\beta = 0.9$ have been empirically obtained.

The main steps to construct the FDT are the following: (1) Select the best attribute as the root of the tree, based on the ambiguity function [1]. (2) For each linguistic term of this attribute, create a branch with the examples with support of at least α, and compute the truth level of classification for each class in the set of classes. (3) If the truth level of classification is above β for at least one class, terminate the branch and set the label as the class with the highest truth level. (4) Otherwise, check if an additional attribute will further reduce the classification ambiguity. If that is the case, select the best one as a new decision node of the branch and repeat step 2 until no further growth is possible. (5) Otherwise, terminate the branch as a leaf with a label corresponding to the class with the highest truth level. After constructing the tree, each branch can be considered as a classification rule with a degree of support equal to the truth level of its conclusion.

3.2. Inference in FDT

The Mamdani inference procedure is applied to find the decision class for an input, x, as follows: (1) Calculate the satisfaction level of the premises of each rule, using the t-norm minimum. (2) Calculate the membership of x to the conclusion class as the product between the satisfaction level of the premises and the degree of support of the rule. (3) Aggregate all the memberships for the same class, given by different rules, using the t-conorm maximum. The result is the confidence on the class. (4) For binary problems, compare the confidences of class 0 and class 1 and choose the one that has the highest value as the final decision class.

3.3. Explanation Extraction from FDT

The second change in Fuzzy-LORE is the explanation extraction process. Fuzzy-LORE derives an explanation from the constructed FDT, slightly different from the one derived by the LORE-based methods. Having in mind that we are dealing with binary classification problems, and given that $b(x) = y$, the explanation of this classification has the form of a triplet $(\mathbb{R}, \Delta, \mathbb{C})$, where:

- $\mathbb{R}$ is the set of decision rules that cover the instance x and have y as output. Each rule $r \in \mathbb{R}$ tells which conditions are satisfied by the object x for being classified as y. Thus, they indicate several minimal sets of conjunctive conditions necessary for belonging to that class.
- Δ is the set of counterfactual rules that lead to the opposite class.
- $\mathbb{C}$ is a set of counterfactual instances, that represent examples of objects that do not belong to class y and have the minimum changes with respect to the original input object x.

3.3.1. Decision rules

Let $R_x = R_x^+ \cup R_x^-$ be the set of all fuzzy rules of the constructed FDT given the instance x. R_x^+ refers to the set of rules that have the conclusion y, and R_x^- is the set of fuzzy rules that have the opposite conclusion. Each rule $r \in R_x^+$ has the following format:

$$IF\ (f_i\ IS\ t_{i,a})\ AND\ (f_j\ IS\ t_{j,b})\ ...AND\ (f_z\ IS\ t_{z,c})\ THEN\ class\ IS\ y$$

Each attribute f_i is a linguistic variable with a set of terms $t_{i,1}, t_{i,2}, \ldots$. Each term has an associated fuzzy set, $\mu_{t_{i,a}}$, and they define a fuzzy partition.

R_x^+ contains the FDT rules activated by x that lead to the conclusion y, so they constitute the base of the decision rules of the explanation. In order to present a comprehensible explanation, Fuzzy-LORE is more flexible than LORE, that has one single activated crisp rule. In this new version, $\mathbb{R}$ can be defined as a subset of the rules in R_x^+, taking advantage of the fuzzy activation of several rules. Depending on the application and on the user type, $\mathbb{R}$ can be either all the rules in R_x^+, the top k rules with highest confidence scores, or the set of rules with a confidence above a certain threshold. In the experiments presented in this paper, the top 3 rules with highest confidence were included in $\mathbb{R}$.

3.3.2. Counterfactual rules

The aim of this step is to find rules similar to those of $\mathbb{R}$ which lead to the opposite conclusion. After finding these counterfactual rules, in the next step it will be possible to compute counterfactual instances (individuals close to x that belong to a different class).

To extract the counterfactual rules Δ, we consider each rule $r_c \in R_x^-$. For each condition $c_i = (f_i \text{ IS } t_{i,a})$ in r_c, we check if f_i appears in any condition of the rules in $\mathbb{R}$ with a different term, i.e. $t_{i,b}$. We change the membership of $t_{i,b}$ by its negation $1 - \mu_{t_{i,b}}(x)$. Then, we re-calculate the final confidence score of the rule r_c with the new membership values as mentioned in subsection 3.2. With the negated membership functions we are analysing what would happen if we changed the original value of x in f_i to a value that activated that condition.

After that, we filter out those rules in R_x^- that have confidence smaller than the maximum confidence score in $\mathbb{R}$ and these are the final counterfactual rules of the explanation component Δ.

3.3.3. Counterfactual instances

Finally, for each rule r_c in Δ we create a counterfactual instance $x_c \in \mathbb{C}$ by making a copy of x and changing only the values of the features that appear in r_c. Concretely, each term t_i of r_c is defuzzified using the Center-of-Maximum method (using the membership values calculated in the previous step). The rationale for substituting the original value by the center of maximum is to put a value in the attribute that maximally activates the condition of the counterfactual rule.

Let us illustrate the process of generating a counterfactual instance with an example from the problem of diagnosis of diabetic retinopathy, used in the experimental section. The instance that is classified is $x = \{$Age=61, Sex=1, EVOL=19, TTM=2, HbA1c=9, CDKEPI=106.21, MA=0, BMI=44.21, HTAR=1$\}$. For simplification we only consider one decision rule r in R_x^+.

$r: IF\ (HbA1c\ IS\ More9)\ THEN\ class\ IS\ Class1$ (confidence = 0.757)

One of the candidate counterfactual rules is

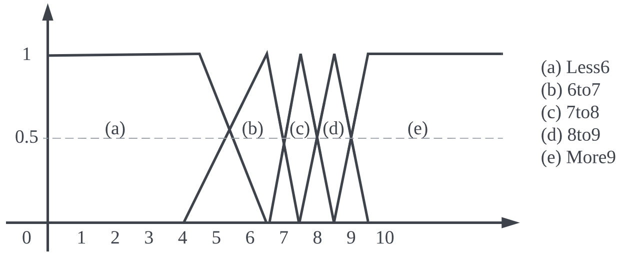

Figure 2. The membership functions of the HBA1c variable.

r_c : $IF\ (HbA1c\ IS\ 6to7)\ AND\ (MA\ IS\ Correct)\ THEN\ class\ IS\ Class0$ with a confidence of 0 based on x.

So, to construct a counterfactual instance only the value of attribute HbA1c must be changed as it appears in both r and r_c. Given the fuzzy set definitions for HbA1c shown in Figure 2, the fuzzification of the value HbA1c=9 of x gives the following membership scores for each term: $\{Less6 = 0, 6to7 = 0, 7to8 = 0, 8to9 = 0.2, More9 = 0.8\}$.

To activate the condition $(HbA1c\ IS\ 6to7)$ in r_c, we take the negated membership of this term, $\mu_{6to7} = 1$. Based on that, the confidence of r_c will be 0.78 instead of 0 and it will be included in Δ as its confidence is larger than 0.757.

Finally, the following counterfactual example is obtained: {Age=61, Sex=1, EVOL=19, TTM=2, **HbA1c=6.5**, CDKEPI=106.21, MA=0, BMI=44.21, HTAR=1}

4. Experiments and Results

This section describes the experimental setup, and discusses the obtained results by comparing the performance of different methods and evaluating the generated counterfactual examples.

4.1. Experimental Setup

We evaluated Fuzzy-LORE on a private data set that shows if a diabetic patient has (or not) a high risk of developing diabetic retinopathy. It is composed of 2323 examples of binary classification. The Diabetic-Retinopathy data set was used to develop a fuzzy random forest-based classifier, called RETIPROGRAM, which is currently being used in the Hospital de Sant Joan in Reus (Tarragona). Each instance in the data set is defined by nine attributes: current age, sex, years since diabetes detection, type of diabetes treatment, good or bad control of arterial hypertension, HbA1c level, glomerular filtrate rate estimated by the CKD-EPI value, microalbuminuria, and body mass index. The data was split into a training set of 1212 examples and a test set of 1111 examples. The classification model used in RETIPROGRAM achieves an accuracy of 80%, with a sensitivity of 81.3% and a specificity of 79.7% [10]. We used the test set in all our experiments to evaluate the effectiveness of Fuzzy-LORE.

4.2. Evaluation of the Explanation Results

As described in the previous section, a Fuzzy-LORE explanation contains the explanation decision rules $\mathbb{R}$ and a set of counterfactual rules Δ, from which the counterfactual examples, $\mathbb{C}$, are derived. These components are obtained from a fuzzy decision tree (a set of fuzzy decision rules), that we call the explanation model. In this section we evaluate the quality of the rules generated by the proposed method and compare it to the LORE-based methods using the following evaluation metrics:

- **Hit**: this metric computes the similarity between the output of the explanation model and the black-box, b, for all the testing instances. It returns 1 if they are equal and 0 otherwise.
- **Fidelity**: this metric measures to which extent the explanation model can accurately reproduce the black-box predictor for the particular case of instance x. It answers the question of how good is the explanation model at mimicking the behaviour of the black-box by comparing its predictions and the ones of the black-box on the instances that are neighbours of x, which are in $\mathcal{D}$.
- **l-Fidelity**: it is similar to the *fidelity*; however, it is computed on the subset of instances from $\mathcal{D}$ covered by the explanation rules, $\mathbb{R}$. It is used to measure to what extent these rules are good at mimicking the black-box model on similar data of the same class.
- **c-Hit**: this metric compares the predictions of the explanation model and the black-box model on all the counterfactual instances of x, $\mathbb{C}$.

Table 1 shows the means and standard deviations of the metrics for Fuzzy-LORE and the previous LORE-based methods on the test set. It may be seen that Fuzzy-LORE and C-LORE-F show almost the same performance in the Hit and Fidelity measures. C-LORE-F is slightly better than Fuzzy-LORE in terms of l-Fidelity. However, Fuzzy-LORE outperforms clearly all the other methods in terms of c-Hit. We can attribute such improvement in the c-Hit measure to the quality of the generated counterfactual examples (which are evaluated in more depth in Section 4.3).

Table 1. Evaluation of the explanation results for Fuzzy-LORE vs other LORE-based methods.

Methods	Hit	Fidelity	l-Fidelity	c-Hit
LORE	0.95±0.13	0.96±0.05	0.95±0.09	0.79±0.32
Guided-LORE	0.99±0.02	0.98±0.06	0.99±0.03	0.83±0.28
C-LORE-F	1.00±0.00	0.99 ± 0.002	0.99 ± 0.002	0.89 ± 0.29
Fuzzy-LORE	1.00±0.00	0.99±0.03	0.98±0.04	0.96 ± 0.17

4.3. Evaluation of the counterfactual examples

Counterfactual examples help to understand what changes may be applied to an object to obtain a different outcome. This is particularly interesting in healthcare applications. Hence, it is important to have counterfactual examples that balance a wide range of suggested modifications (diversity) and the relative facility of adopting those modifications (proximity to the actual input). Moreover, counterfactual examples must be actionable, e.g., people can not reduce their age or change their race.

In this subsection, we evaluate the generated counterfactual examples for C-LORE-F and Fuzzy-LORE (as they showed almost the same performance) using the following evaluation metrics [11]:

- **Validity**: is the number of counterfactual examples with a different outcome than the original input, i.e., x, divided by the total number of counterfactual examples.

$$\text{Validity} = \frac{|\hat{x} \in \mathbb{C} \ s.t. b(x) \neq b(\hat{x})|}{|\mathbb{C}|} \tag{1}$$

Here $\mathbb{C}$ refers to the set of returned counterfactual examples and b is the black-box model.

- **Proximity**: is the mean of feature-wise normalised distances between a counterfactual example c and the original input x.

$$\text{Proximity} = 1 - \frac{1}{|\mathbb{C}|} \sum_{c \in \mathbb{C}} dist(c, x) \tag{2}$$

- **Sparsity**: it measures the average of attribute value changes between a counterfactual example and the original input.

$$\text{Sparsity} = 1 - \frac{1}{|\mathbb{C}| * |\mathbb{F}|} \sum_{c \in \mathbb{C}} \sum_{f \in \mathbb{F}} \mathbb{1}[c_f \neq x_f] \tag{3}$$

Here, $\mathbb{F}$ is the set of features, and $\mathbb{1}$ is the indicator function.

- **Diversity**: it is similar to proximity. However, instead of computing the feature-wise distance between the counterfactual example and the original input, we compute it between each pair of counterfactual examples.

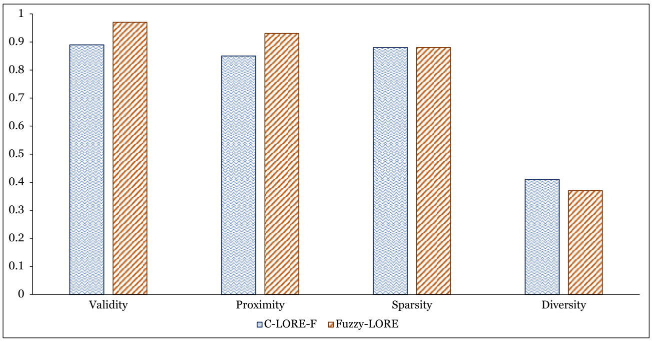

Figure 3. Evaluation of the counterfactual examples for C-LORE-F and Fuzzy-LORE.

Figure 3 shows the comparative results of Fuzzy-LORE vs C-LORE-F with respect to these evaluation metrics. In general, Fuzzy-LORE showed better performance than C-LORE-F, mainly in terms of validity and proximity. Both Fuzzy-LORE and C-LORE-F have similar performance in terms of sparsity. Looking at the diversity results, we can find that C-LORE-F generates slightly more diverse counterfactual examples than the proposed method. However, both of them showed a low performance. This issue will be studied in future work.

5. Conclusion

Fuzzy-LORE is a new post-hoc explanation method for fuzzy binary classifiers. It learns a local fuzzy decision tree on a synthetic neighbourhood of an instance. Then, it extracts from it a meaningful explanation consisting of : (1) A set of decision rules that explain the reasons behind the classification decision. (2) A set of counterfactual rules that suggest a minimal number of changes in the instance features to get a different outcome. (3) A set of counterfactual examples. The method has been evaluated on a dataset to assess the risk of developing diabetic retinopathy. The evaluation results revealed that using the fuzzy decision tree as an explanation model gives better explanations than the decision tree, mainly in the counterfactual rules and instances. However, Fuzzy-LORE failed to generate diverse counterfactual examples. Hence, in our future work, we plan to improve the diversity of the generated counterfactual examples. We also plan to extend the current work to provide explanations for multi-class fuzzy-based classifiers.

Acknowledgements

Research projects PI21/00064 and PI18/00169 from ISCIII & FEDER funds and URV grants 2020PFR-B2-61 & 2019PFR-B2-61. First author has a URV Martí Franquès predoctoral grant.

References

[1] Saleh E, Valls A, Moreno A, Romero-Aroca P, Torra V, Bustince H. Learning fuzzy measures for aggregation in fuzzy rule-based models. In: International Conference on Modeling Decisions for Artificial Intelligence. Springer; 2018. p. 114-27.

[2] Adadi A, Berrada M. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE access. 2018;6:52138-60.

[3] Carvalho DV, Pereira EM, Cardoso JS. Machine learning interpretability: A survey on methods and metrics. Electronics. 2019;8(8):832.

[4] Strumbelj E, Kononenko I. An efficient explanation of individual classifications using game theory. The Journal of Machine Learning Research. 2010;11:1-18.

[5] Ribeiro MT, Singh S, Guestrin C. 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. ACM; 2016. p. 1135-44.

[6] Guidotti R, Monreale A, Ruggieri S, Pedreschi D, Turini F, Giannotti F. Local rule-based explanations of black box decision systems. arXiv preprint arXiv:180510820. 2018.

[7] Maaroof N, Moreno A, Valls A, Jabreel M. Guided-LORE: Improving LORE with a Focused Search of Neighbours. In: Heintz F, Milano M, O'Sullivan B, editors. Trustworthy AI - Integrating Learning, Optimization and Reasoning. Springer; 2021. p. 49-62.

[8] Maaroof N, Moreno A, Jabreel M, Valls A. Contextualized LORE for Fuzzy Attributes. Artificial Intelligence Research and Development. 2021:435.

[9] Yuan Y, Shaw MJ. Induction of fuzzy decision trees. Fuzzy Sets and systems. 1995;69(2):125-39.

[10] Blanco MES, Romero-Aroca P, Pujol RV, Valls A, SaLeh E, Moreno A, et al. A Clinical Decision Support System (CDSS) for diabetic retinopathy screening. Creating a clinical support application. Investigative Ophthalmology & Visual Science. 2020;61(7):3308-8.

[11] Kommiya Mothilal R, Sharma A, Tan C. Explaining Machine Learning Classifiers through Diverse Counterfactual Explanations. arXiv e-prints. 2019:arXiv-1905.