A User Study on the Utility of Context-aware Explanations for Assisting Data Scientists in Error Analysis of Fuzzy Decision Trees

Guillermo Fernández[1], Jose A. Gámez[1], Jose M. Puerta[1], Juan A. Aledo[1], Jose M. Alonso-Moral[2], Alberto Bugarin[2]
[1]*Departamento de Sistemas Informáticos, Universidad de Castilla-La Mancha*, Albacete, Spain
[2]*Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS), Universidade de Santiago de Compostela*, Santiago de Compostela, Spain
{Guillermo.Fernandez, Jose.Gamez, Jose.Puerta, JuanAngel.Aledo}@uclm.es,
{josemaria.alonso.moral, alberto.bugarin.diz}@usc.es

2024 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE) | 979-8-3503-1954-5/24/$31.00 ©2024 IEEE | DOI: 10.1109/FUZZ-IEEE60900.2024.10611812

***Abstract*—The current growth of interest in the field of eXplainable Artificial Intelligence has led to the rise of multiple explanation generation techniques to bridge the gap between the algorithmic complexity of the most powerful algorithms and their end users, who are expected to take advantage of them.**

In addition, explanations can assist with other tasks, such as explaining the algorithm to a model designer, who can use that information to fine-tune the system that implements such an algorithm. This gives the model designer some insight into the inner workings of the system to guide design decisions as well as identify and prevent potential errors, resulting in a myriad of improvements regarding model properties such as accuracy, explainability, trustworthiness, coherence with existing knowledge, etc. In this paper, we introduce a method to enrich local explanations of fuzzy decision tree classifications with context information. The goodness of the proposed method was validated with a user study in which 26 participants had to detect classification errors made by a fuzzy decision tree. Reported results show that enriched explanations allowed the participants to be more accurate in detecting errors in the classification system, albeit taking longer to solve the task with respect to participants who were shown regular local explanations.

***Index Terms*—Explainable AI, Trustworthy AI, Interpretable Fuzzy Systems, Fuzzy Decision Trees, Factual Explanations, Error Analysis**

I. Introduction

In recent times, we have experienced a massive rise in the amount of available data, which has led to new, more complex models to be developed to tackle a wide range of tasks [1]–[3]. The problem arises when these new models are accompanied by a decrease in interpretability [4]. In certain critical fields, e.g., medicine, aviation, law, etc., users may need to trust the models before deploying them, given the consequences of the decisions such models can promote.

The omnipresence of this increase in the use of models has even caused legislation to appear to regulate it. The General Data Protection Regulation (GDPR) [5] is a European regulation that, among other things, contemplates the *right to explanation*, which is not only applicable to humans but also to machines and Artificial Intelligence (AI) techniques. The European AI Act[1] goes a step forward, paying special attention to high-risk applications when regulating the development of human-centric Trustworthy AI, which is demanded to be well aligned with European values.

All of this motivates the rise of a new research field, *eXplainable AI* (XAI) [6], [7], intending to push interpretability and explainability into AI models so as to gain an understanding of their inner working, which can enable them to be used in sensitive applications. XAI research into the generation of new explanation techniques is gaining a lot of popularity, to the point where there are multiple taxonomies [6], [8] that divide these techniques into groups, depending if they are *model agnostic* or *model specific*, if they are *local explanations* or *global explanations*, if they are *ad-hoc* or *post-hoc*, etc. These different techniques are also tested through human evaluation [9], [10] to determine the best way of explaining the model to end users.

We can push the usefulness of explanations even further. There is a possibility of using explanations for error analysis [11], and XAI techniques can help to detect false predictions [12]. Indeed, the so-called *Error Analysis* can help data scientists get a deeper understanding of machine learning model errors. Accordingly, this paper focuses on using context-aware local explanations to detect errors made by a fuzzy white box classifier, in this case, a Fuzzy Decision Tree. The main contributions of the paper are:

This work has been funded by the following projects: SBPLY/21/180225-/000062 (Government of Castilla-La Mancha and ERDF funds); PID2022–106758GB–C33, and FPU19/02930 (MCIN/AEI/10.13039/501100011033 and ERDF Next Generation EU); and 2022-GRIN-34437 (Universidad de Castilla-La Mancha and ERDF funds). This work is also supported under Grants PID2021-123152OB-C21 and PID2020-112623GB-I00, funded by MCIN/AEI/10.13039/501100011033 and by "ESF Investing in your future", under Grant TED2021-130295B-C33 funded by MCIN/AEI/10.13039/501100011033 and by the "European Union NextGenerationEU/PRTR", and the Galician Ministry of Culture, Education, Professional Training and University (Grants ED431G2019/04 and ED431C2022/19 cofunded by the European Regional Development Fund, ERDF/FEDER program).

[1]https://artificialintelligenceact.eu/

- First, we propose a novel way of extracting context information automatically from a Fuzzy Decision Tree.
- Second, we propose a novel method for generating a context-aware local explanation for each data instance classified by a Fuzzy Decision Tree.
- Third, we carry out a user study in which we survey a group of experts in fuzzy systems with the aim of validating the utility of context-aware local explanations for assisting in the error analysis task. Half of the participants interact with context-aware local explanations. The rest of the participants interact with regular local explanations.

The rest of the paper is structured as follows. Section II describes the current state of the art. Section III introduces some necessary concepts and notation. Section IV illustrates which are the research questions we are trying to answer and how the survey has been designed. Section V explains the material employed, how the context information and explanations are generated, and includes an illustrative example. Section VI describes and discusses the survey results. Finally, Section VII provides readers with concluding remarks and points out some future lines of work.

II. Related Work

Fuzzy classifiers [13] are a specific kind of fuzzy model which can predict a discrete (crisp) target variable, i.e., the class. Fuzzy rule-based classifier systems (FRBCS) are a specific type of fuzzy classifier defined by a set of rules with the class variable as consequent and a subset of the predictive fuzzy variables as antecedent. Fuzzy Decision Trees (FDT) [14] are a knowledge representation for fuzzy rule-based classifiers. We focus on FDT as the fuzzy rule-based model to be explained, since it is a widely employed algorithm in cases where the knowledge needs to be induced from data. However, the proposed method can be generalized for other types of FRBCS. The implementation used in this paper is the Fuzzy Multiway Decision Tree described in [15], using the *aggregated vote* process to compute the class value.

Regarding explanations, one important distinction within XAI taxonomy is whether the method generates a *local* or a *global* explanation. A *local explanation* is a particular type of explanation that focuses on a single instance and the neighborhood around it, while *global explanations* aim to explain the entire behavior of the model. *Factual explanations* [16], [17] are a common type of local explanations, which can be defined as an answer to the question about why a particular instance is classified into a certain class value.

A well-established method of extracting factual explanations for decision trees is by using the path from the tree's root to the leaf that classifies the instance as a decision rule. This method is simple enough to extract an explanation from a crisp tree, where only a single leaf node is activated. In addition, there are different ways of extracting a local explanation from a fuzzy tree where multiple leaf nodes can be activated. One such option is to use the path with the maximum activation degree as the factual explanation, which works best together with trees that use maximum matching to determine the output of the class value [18].

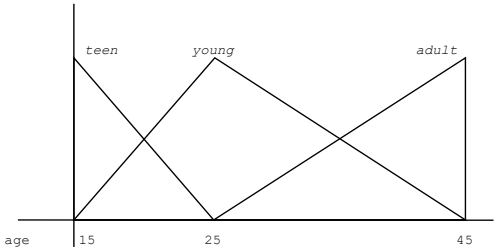

Fig. 1. Illustrative example of Ruspini Partition for age.

The FDT chosen in this paper uses *aggregated vote*, which considers all the fired leaf nodes. This is why the factual explanations for this tree are extracted using the $mr\text{-}factual$ method described in [16]. This method considers all the rules that have been used for the classification and selects one or multiple rules to explain an instance. The added activation degree of all the rules with the classified class value with respect to the added activation degree of the rest of the class values is the metric by which the rule or rules are then selected.

Finally, it is worth noting that the idea of enhancing local explanations by using *global knowledge* has already been discussed in the literature. In [19], the authors turn local Shapley values into global explanations by means of functional decomposition. Other publications propose merging local and global explanations, for example, through feature importance [20], concept relevance [21], saliency maps [21] or strategy summaries [22].

III. Preliminary Concepts

In this section, let us begin with the revision of some related concepts. In a classification problem, an instance $x = (x_1, \ldots, x_n) \in \mathcal{X}_1 \times \cdots \times \mathcal{X}_n$, where $\mathcal{X}_1, \ldots, \mathcal{X}_n$ are n sets of *input variables*, is mapped to a decision $y \in \mathcal{Y} = \{y_1, \ldots, y_m\}$ by a function (classifier) $f : \mathcal{X}_1 \times \cdots \times \mathcal{X}_n \to \mathcal{Y}$. We write $f(x) = y$ to denote the classification y given to x. Let us assume that, associated with each continuous input variable $\mathcal{X}_i$, there is a fuzzy (linguistic) variable $\mathcal{F}_i = \{v_{i,1} \ldots, v_{i,k_i}\}$ defined through a Ruspini partition [23] (Strong Fuzzy Partition, SFP) of k_i ordered fuzzy sets (see Figure 1), where $\mathcal{F}$ is the set of all fuzzy (linguistic) variables. We use v_{i,z_i} to denote both the fuzzy set and its corresponding associated linguistic label, indistinctly. Notice that a triangular fuzzy set is defined by a triple of real-valued points (start, modal point, end). For example, in Figure 1, we have *teen* = (15, 15, 25) and *young* = (15, 25, 45).

Given a value $\delta \in dom(\mathcal{X}_i)$, let $\mu_i(\delta) = (\mu_{i,1}(\delta), \ldots, \mu_{i,k_i}(\delta))$ be the vector of membership degrees of δ to the k_i fuzzy sets of $\mathcal{F}_i$. In other words, $\mu_{i,z_i}(\delta)$ is the membership degree of δ to the set v_{i,z_i}.

Let $e = \{r_1, \ldots, r_e\}$ be an explanation formed by one (or more) fuzzy decision rules. Each rule $r = P(r) \to y(r)$ consists of a set of premises in conjunctive form $P(r) =$

$p_{s_1} \wedge \cdots \wedge p_{s_r}$ and an outcome $y(r) \in \mathcal{Y}$. Each premise $p_i = \langle \mathcal{F}_i, v_{i,z_i} \rangle$ is an attribute-value pair where $\mathcal{F}_i$ is a fuzzy variable and v_{i,z_i} is one of its corresponding fuzzy sets.

One property of multi-rule explanations is that, given an explanation e that explains the instance x, then $y(r) = b(x)$ for all $r \in e$. Fuzzy rules differ from crisp rules in that, while a crisp rule has a binary (0 or 1) match with an instance x, a fuzzy rule r has a *matching degree* with the instance, $md(r, x)$, defined as:

$$md(r, x) = \min_{i \in \{s_1, \ldots, s_r\}} \{\mu_{i,z_i}(x_i)\} \in [0, 1].$$

Let $t()$ be a Fuzzy Decision Tree (FDT) whose decision-making process needs to be explained, learned from a training dataset $TR = \{(x_1^t, \ldots, x_n^t, y^t)\}_{t=1}^T$. Each leaf node h of the FDT is formed by multiple class values with a weight each, meaning each branch can be written as the set of rules with the premise formed by each decision node and one of the classes in h. Let R_t be the ruleset composed by all the rules extracted from t. Each rule $r \in R_t$ has a weight $w(r)$ that is used to compute the *activation degree* of the rule and an instance x, $AD(r, x)$, defined as:

$$AD(r, x) = md(r, x) \cdot w(r)$$

For a given set of instances X and a ruleset R_t, let us denote $X(r) = \{x \in X \mid AD(r, x) > 0\}$, $R_t^y = \{r \in R_t \mid y(r) = y\}$ and $R_t(v_{i,z_i}) = \{r \in R_t \mid v_{i,z_i} \in P(r)\}$.

We define the mean of the activation degrees of a rule r for a dataset X as the sum of the activation degrees of that rule and each instance of the dataset, divided by the number of instances that r fires, i.e.:

$$mAD(r, X) = \frac{\sum_{x \in X(r)} AD(r, x)}{|X(r)|}.$$

In addition, the mean of activation degrees of a fuzzy set v_{i,z_i} for X in a ruleset R_t is defined as the sum of the mean of the activation degrees of each rule that has v_{i,z_i} in any premise, divided by the number of rules with v_{i,z_i} in any premise, i.e.:

$$mAD(v_{i,z_i}, X, R_t) = \frac{\sum_{r \in R_t(v_{i,z_i})} mAD(r, X)}{|R_t(v_{i,z_i})|}.$$

We also define the weighted mean of activation degrees of a fuzzy set v_{i,z_i} for X in a ruleset R_t as the product of $mAD(v_{i,z_i}, X, R_t)$ and the number of times each rule in R_t is fired, i.e.:

$$wAD(v_{i,z_i}, X, R_t) = mAD(v_{i,z_i}, X, R_t) \cdot \sum_{r \in R_t(v_{i,z_i})} |X(r)|.$$

The intuition behind $wAD(v_{i,z_i}, X, R_t)$ is as follows. It represents the average activation degree related to a certain fuzzy set v_{i,z_i} weighted by the number of times it is relevant in X. Therefore, the more times it appears (as we count any rule with v_{i,z_i} in any premise), the higher the relevance degree associated with that fuzzy set.

Let us now introduce the *context information* which includes the *most important* fuzzy sets needed to infer a class from a dataset. Moreover, we define the *importance* $I(v_{i,z_i}, X, R_t^y)$ for the fuzzy set v_{i,z_i}, the dataset X, the ruleset R_t, and the class value y as:

$$I(v_{i,z_i}, X, R_t^y) = \frac{wAD(v_{i,z_i}, X, R_t^y)}{\sum_{\mathcal{F}_j \in \mathcal{F}} \sum_{v_{j,z_j} \in \mathcal{F}_j} wAD(v_{j,z_j}, X, R_t^y)}.$$

where the importance $I \in [0, 1]$ and it comes out as result of dividing the weighted mean of activation degrees of a fuzzy set by the sum of all weighted means of all fuzzy sets.

Accordingly, we define the *important sets* of a fuzzy variable $\mathcal{F}_i$ and a class y, computed for a dataset X and a ruleset R_t, $IS(\mathcal{F}_i, X, y, R_t)$, as the set of fuzzy sets which have importance I greater than a threshold th:

$$IS(\mathcal{F}_i, X, y, R_t) = \{v_{i,z_i} | v_{i,z_i} \in \mathcal{F}_i, I(v_{i,z_i}, X, R_t^y) > th\}.$$

Finally, the *context information* of a class y, computed for a dataset X and a ruleset R_t, $Ctx(X, y, R_t)$, is defined as:

$$Ctx(X, y, R_t) = \{\langle \mathcal{F}_i, IS(\mathcal{F}_i, X, y, R_t) \rangle \ \forall \ \mathcal{F}_i \in \mathcal{F}\}$$

IV. Evaluation Methodology

We follow the evaluation methodology introduced in [24], which consists of two stages: Planning and Execution, in agreement with the best practices described in [10]. The first stage determines *what* will be evaluated, while the second stage focuses on *how* the evaluation is carried out.

A. Planning Stage

During the planning stage, we carry out the following steps: (i) we determine the goal of the evaluation, (ii) we establish the type of research, and (iii) we set the sample of participants.

The goal of the evaluation is answering to the following research question:

- **RQ1:** Does presenting context-aware local explanations help detect errors of the classifiers better than presenting regular local explanations?

In order to answer RQ1, we first define the following hypotheses:

- **H1:** Users presented with context-aware local explanations will be more accurate in determining if the model is right or wrong.
- **H2:** Users presented with context-aware local explanations will be faster in determining if the model is right or wrong.

There are two types of evaluations we can perform for testing H1 and H2:

- **Qualitative Evaluation** focuses on improving the system, usually gathering non-numerical data [25], indicating how the users perceive the system. These evaluations often employ open-ended questionnaires and free-text comments.
- **Quantitative Evaluation** focuses on aggregating numerical data into categories, ordering or measuring data into appropriate units [26]. These evaluations often employ close-ended questionnaires.

Given the nature of our two hypotheses, a **Quantitative Evaluation** where we ask the participants to solve several close-ended tasks and then aggregate the collected data to validate the hypothesis is the most suitable type of evaluation. However, we also include open-ended questions so the participants can express their opinions on each task.

The last step of the planning stage is to determine the sample of participants. In other words, we have to define the *target audience*. Given the constraints of our hypotheses, the target audience is made up of *researchers with a background in fuzzy sets and systems*, who can evaluate the result of the classification of an FDT. Accordingly, we look for a small but well-focused sample of participants. Therefore, to gather participants, we submitted the survey to the email distribution lists of EUSFLAT[2], NAFIPS[3], and the IEEE-CIS Task Force on Explainable Fuzzy Systems.

B. Execution Stage

During the execution stage, we carry out the following steps: (i) we develop the consent form and debriefing statement, (ii) we develop the survey questionnaire, (iii) we distribute the link to the questionnaire by email distribution lists, and (iv) we analyze the collected data. In the next section, we will go into depth about how the execution stage is implemented.

V. Implementation Details

To address the research question, we need to implement the specific details of the methodology defined in the previous Section. For that, the following elements must be specified: (i) the material used, (ii) the generation of the explanations, and (iii) the generation of the survey.

A. Material

As a use case, we have selected a beer-style classification task using the dataset[4] first introduced in [27] and later used in [23] for the design and validation of an explainable fuzzy beer style classifier. The task consists of identifying one out of the eight beer styles (Blanche, Pilsner, Lager, IPA, Barleywine, Belgian Strong Ale, Porter, and Stout) in terms of three attributes (Color, Bitterness, and Strength). Color is translated into a number in terms of the Standard Reference Method (SRM[5]). Bitterness is measured on the International Bittering Units (IBU[6]) scale. Strength is measured in terms of Alcohol content by Volume (ABV[7]).

Regarding the model, we make two assumptions: (i) the model to be explained is already learned, and (ii) the data for learning the model, extracting the context information, and the instances to be explained are independent. For that, the dataset is divided into a training set (TR) to learn the model, a validation set (VAL) to extract the context information, and a test set ($TEST$) to explain the instances. The context information used in the survey ($Ctx(VAL, y, R_t)$) is described in Table I.

TABLE I
Context information ($th = 0.2$)

Class value	Context information
$Blanche$	$\langle Color, \{Pale, Straw\}\rangle$ $\langle Bitterness, \{Low, Low\text{-}medium\}\rangle$
$Pilsner$	$\langle Color, \{Straw\}\rangle$ $\langle Bitterness, \{Medium\text{-}high\}\rangle$
$Lager$	$\langle Color, \{Straw, Amber\}\rangle$ $\langle Bitterness, \{Low\text{-}medium\}\rangle$
IPA	$\langle Color, \{Straw, Amber\}\rangle$ $\langle Bitterness, \{Medium\text{-}high\}\rangle$
$Barleywine$	$\langle Color, \{Amber\}\rangle$
$Belgian\text{-}Strong\text{-}Ale$	$\langle Color, \{Brown\}\rangle$ $\langle Strength, \{High, Very\ high\}\rangle$
$Porter$	$\langle Color, \{Brown\}\rangle$ $\langle Strength, \{Session, Standard\}\rangle$
$Stout$	$\langle Color, \{Brown, Black\}\rangle$ $\langle Strength, \{Standard\}\rangle$

As fuzzy classifier, we use the Fuzzy Multiway Decision Tree described in [15]. We apply the aggregated vote process to compute the class value. The variables are partitioned according to fuzzy sets previously defined by experts, which are described as triangular SFP in Table II.

For the evaluation platform, we have considered Qualtrics[8]. It is an online questionnaire generation tool with a user-friendly interface. Qualtrics manages multiple types of questions, a timer for each question, an enriched text editor to format the questions properly, and a powerful database filter tool from which to extract the relevant data to analyze results. After generating the questionnaire, Qualtrics produces a general link that has been shared via email distribution lists.

TABLE II
Fuzzy Set Definitions

Variable	Fuzzy Sets	Definition
$Color$	$Pale$	$(0, 0, 5.25)$
	$Straw$	$(0, 5.25, 13.25)$
	$Amber$	$(5.25, 13.25, 24)$
	$Brown$	$(13.25, 24, 45)$
	$Black$	$(24, 45, 45)$
$Bitterness$	Low	$(7, 7, 26.75)$
	$Low\text{-}medium$	$(7, 26.75, 40)$
	$Medium\text{-}high$	$(26.75, 40, 250)$
	$High$	$(40, 250, 250)$
$Strength$	$Session$	$(0.035, 0.035, 0.06)$
	$Standard$	$(0.035, 0.06, 0.07875)$
	$High$	$(0.06, 0.07875, 0.136)$
	$Very\ High$	$(0.07875, 0.07875, 0.136)$

[2]European Society for Fuzzy Logic and Technology (https://www.eusflat.org/)

[3]North American Fuzzy Information Processing Society.

[4]We have taken the dataset in ARFF format from https://gitlab.citius.usc.es/jose.alonso/xai

[5]https://beerandbrewing.com/dictionary/xizZHSE4me/

[6]https://www.thebrewenthusiast.com/ibus

[7]https://alcohol.org/statistics-information/abv/

[8]https://www.qualtrics.com/

Finally, it is worth noting that all the required code has been developed using Python 3.10. The implementation of the multi-rule local explanations is supported by open-source code, which has been taken from the GitHub[9] associated with [16]. Moreover, all the code used for the preparation of the survey, including the extraction of the context information, as well as the (anonymized) results and processing of the survey, are open-source and available in GitHub[10].

B. Generation of explanations

1) Context information: To write the context information for the validation set, a class value y and the rules extracted from the Fuzzy Multiway Decision tree R_t, $Ctx(VAL, y, R_t)$ (in short, Ctx) as natural text, the following template is followed:

(T1) $[y]$ *beer usually has* $[Ctx_1]$ *and* $\cdots$ *and* $[Ctx_n]$.

where each element of the context information $Ctx_i = \langle \mathcal{F}_i, \{v_{i,1}, \ldots, v_{i,z_i}\}\rangle$ is written as:

(T2) *a* $[v_{i,1}]$ *or* $\cdots$ *or* $[v_{i,z_i}]$ $[\mathcal{F}_i]$

2) Regular local explanations: Given an explanation $e = \{r_1, \ldots, r_e\}$, we group the fuzzy sets by their corresponding fuzzy variables:

$$G(e) = \{\langle \mathcal{F}_i, \{v_{i,z_i}\}\rangle, \forall \mathcal{F}_i \in \mathcal{F}, v_{i,z_i} \in P(r_1) \vee \cdots \vee v_{i,z_i} \in P(r_e)\}$$

Then, we write the local explanation as:

(T3) *The instance belongs to the family of* $[y]$ *beer because it has* $[G_1]$ *and* $\cdots$ *and* $[G_n]$.

Where each $G_i = \langle \mathcal{F}_i, \{v_{i,1}, \ldots, v_{i,z_i}\}\rangle$ is written as:

(T4) *a* $[v_{i,1}]/\cdots/[v_{i,z_i}]$ $[\mathcal{F}_i]$

3) Context-aware local explanations: To enrich explanations with context information, we structure a sentence that summarizes the expected behavior for the class value. For that, we add a first sentence with the expected behavior of the fuzzy sets and a second sentence with the unexpected behavior of the fuzzy sets.

For the first sentence, we define the expected behavior EB as the set of fuzzy sets that appear in the factual explanation and the related context information. Then, we write the sentence using the following template:

(T5) *As usual for this family, the beer has* $[EB_1]$ *and* $\cdots$ *and* $[EB_n]$.

where each EB_i is written using Template **(T2)**.

For the second sentence, we define the unexpected behavior EB' as the group of fuzzy sets that appear in the factual explanation but not in the related context information. Then, we write the sentence using the following template:

(T6) *However, it is not usual for this beer to have* $[EB'_1]$ *and* $\cdots$ *and* $[EB'_n]$.

4) Illustrative example: Let us suppose a class value $y = Stout$ with context information:

$$Ctx = \{\langle Color, \{Brown, Black\}\rangle, \langle Strength, \{Standard\}\rangle\}$$

According to Templates **T1** and **T2**, the context information for that class is written as:

(E1) *Stout beer usually has a Brown or Black color and a Standard strength.*

Let us suppose a factual explanation $e = \{r_1, r_2\}$ with rules:

$$r_1 = \langle Color, Brown\rangle \wedge \langle Strength, Standard\rangle \rightarrow Stout$$
$$r_2 = \langle Color, Black\rangle \wedge \langle Strength, Session\rangle \rightarrow Stout$$

We group the sets and variables as follows:

$$G(e) = \{\langle Color, \{Brown, Black\}\rangle, \langle Strength, \{Standard, Session\}\rangle\}$$

According to Templates **T3** and **T4**, we write the explanation as follows:

(E2) *The instance belongs to the family of Stout beer because it has a Brown/Black color and a Session/Standard Strength.*

Finally, we compute the expected behavior EB as:

$$EB = \{\langle Color, \{Brown, Black\}\rangle, \langle Strength, \{Standard\}\rangle\}$$

and the unexpected behavior EB' as:

$$EB' = \{\langle Strength, \{Session\}\rangle\}$$

According to Templates **T5** and **T6**, we write the context-aware local explanation as follows:

(E3) *The instance belongs to the family of Stout beer because it has a Brown/Black color and a Session/Standard Strength. As usual for this family, the beer has a Brown/Black color and a Standard strength. However, it is not usual for this beer to have a Session strength.*

C. Generation of the survey

The survey participants are randomly divided into two groups (with a 50% chance of belonging to each group) in terms of the type of explanation they will interact with: *regular local explanations* or *context-aware local explanations*. Participants do not know which group they are assigned to. Moreover, they do not know that there are different groups.

The structure of the questionnaire developed with Qualtrics is as follows:

[9]https://github.com/Kaysera/teacher

[10]https://github.com/Kaysera/knowledge-extraction

1) **Welcome**: A text that explains the purpose of the survey, shows information about the research groups involved in the development of the survey and provides contact information to the lead researcher for any questions and/or issues.
2) **Consent**: A consent form that consists of several items that confirm that the participant has reached the age of majority, knows the participation in the survey is voluntary, has enough information to proceed, and knows that the data collected is anonymous and properly treated according to the GDPR.
3) **Instructions**: A text explaining the use case under study. It also includes some information regarding the dataset, the model, and the test instances. In addition, the format of the tasks presented during the rest of the survey is described together with some important notes.
4) **Comprehension Check**: The participant is shown an illustrative task (with the same structure as the tasks that she/he will have to solve later) to verify if she/he understood the previous instructions properly.
5) **Tasks**: There are four groups of tasks, and one task from each group will be presented to each participant:
    a) *High Confidence Hit*: The tree correctly classifies the test instance with high confidence (i.e., there is a big difference in activation degree between the correct class and the second-best possible class).
    b) *Low Confidence Hit*: The tree correctly classifies the test instance with low confidence.
    c) *High Confidence Miss*: The tree incorrectly classifies the test instance with high confidence.
    d) *Low Confidence Miss*: The tree incorrectly classifies the test instance with low confidence.

    Figure 2 shows one example of a task. The picture includes five main elements:

    a) *Context information*: It includes a piece of text related to each class, and it is generated as described in Section V-B.
    b) *Tree activation*: An image with the Fuzzy Decision Tree represents graphically the most relevant information that is required to understand the inference process (see further details in Fig. 3 where the tree is zoomed in). The image includes (in the top-right part) the numerical values of the test instance along with the range of possible values for each feature. In addition, all nodes and leaves in the tree (with the weight of each class in each leaf) are printed. Moreover, the nodes activated by the test instance are highlighted in green, with their corresponding membership degrees and final match printed below.
    c) *Explanation*: Depending on the group the participant has been assigned at the beginning of the survey, she/he will see a *regular local* or *context-aware* explanation in natural language below the image of the tree.
    d) *Close-answer question*: The participant is asked if (according to the context information, the image with the tree activation details, and the related explanation) she/he believes the classification is correct, incorrect, or doubtful.
    e) *Open-answer question*: Participants can justify their answer in a free text if they believe it necessary.
6) **Global Satisfaction**: Participants can indicate the helpfulness of the given explanations and context information on a 5-point Likert scale.
7) **Demographics**: Two general questions to determine the region and education level of the participant.

To sum up, each participant must complete a comprehension check, do four tasks, answer a global satisfaction question, and answer a couple of demographic questions.

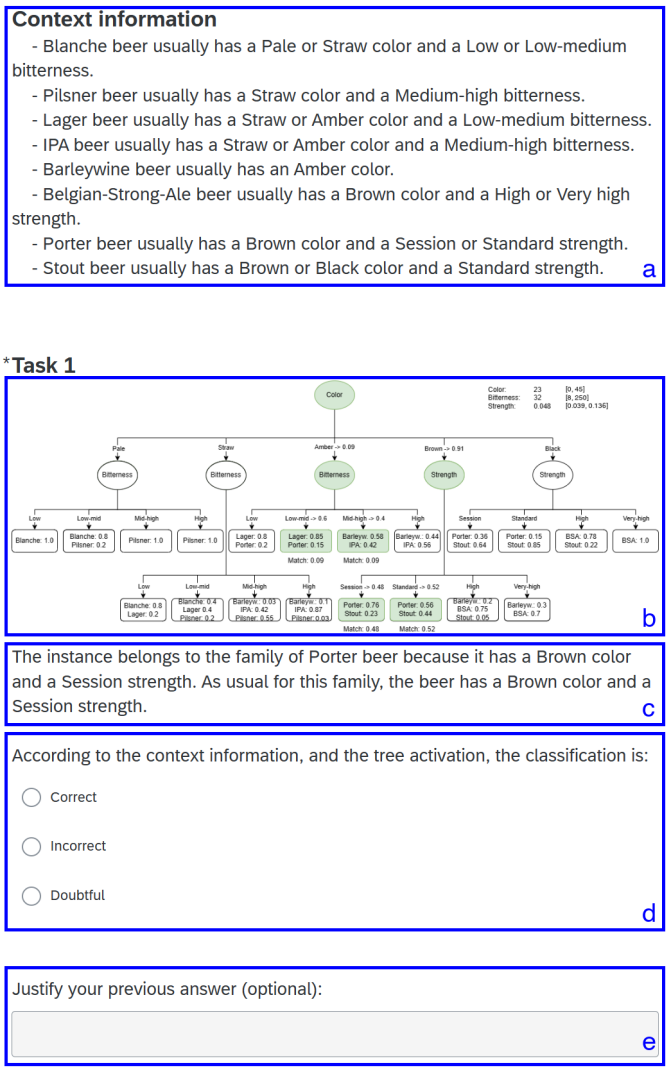

Fig. 2. Example of a task from the survey (*Context-aware* group).

VI. Results of the User Study

The survey was available for 15 days, during which a total of 26 participants took part. Most of the participants were

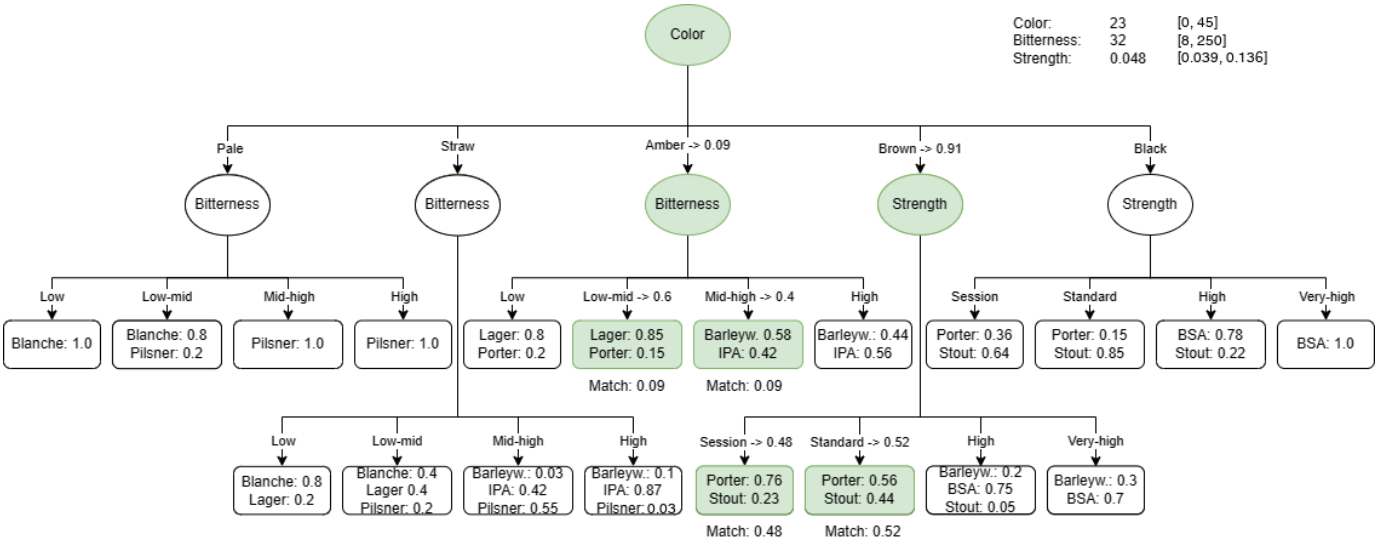

Fig. 3. Representation of a Fuzzy Decision Tree.

European (88.4% vs. 11.6% of American participants), and all of them were Computer Science majors with a post-graduate level (76.9% of the participants had a Ph.D. while 23.1% had a master's degree). They were equally distributed into the two groups (13 participants in each group). Participants in the first group interacted only with *regular local* explanations, while participants in the second group interacted only with *context-aware* explanations. For short, the two groups will be referred to as *local* and *Ctx-A*, respectively.

Table III contains the survey results. As we can see, most people answered the comprehension check question correctly, but some of them expressed doubts regarding the result of the classification. We considered all answers valid since no participant declared the classifier's result incorrect. It can also be observed that the comprehension check took longer than the subsequent tasks. This was expected, as participants must understand what the task to solve is about and familiarize themselves with the survey format.

TABLE III
SURVEY RESULTS: REGULAR LOCAL EXPLANATIONS (LOCAL) VERSUS CONTEXT-AWARE EXPLANATIONS (CTX-A).

		Correct	**Doubtful**	**Incorrect**	**Time**
Comp. Check	Ctx-A	0.77	**0.23**	0	**03:16**
	Local	**0.85**	0.15	0	02:45
Task 1	Ctx-A	0.92	**0.08**	0	**01:24**
	Local	**1.00**	0	0	00:59
Task 2	Ctx-A	**0.62**	0.23	**0.15**	**01:06**
	Local	0.54	**0.31**	**0.15**	00:50
Task 3	Ctx-A	**0.62**	0.23	**0.15**	**01:11**
	Local	**0.62**	**0.31**	0.07	01:06
Task 4	Ctx-A	**0.62**	0.15	**0.23**	**01:22**
	Local	**0.62**	**0.31**	0.07	00:29

Regarding Group 1 (*High Confidence Hit*), we can see that all *local* participants (and most *Ctx-A* participants) were able to identify the answer correctly. There is a minimal difference between groups, in which it seems that when the classifier has a very clear hit, showing less information that can distract the participant is beneficial.

Regarding Group 2 (*Low Confidence Hit*), we can see that more *Ctx-A* participants can identify the solution correctly. Nevertheless, *Local* participants were more prone to doubt the classifier's validity in this context. Notice that rule activation degrees are closer in Group 2, which is likely to make the task harder.

Regarding Group 3 (*High Confidence Miss*), the same number of participants in both groups identified the answer incorrectly. Still, more *Ctx-A* participants could correctly identify the answer as Incorrect in this case, while *local* participants were less confident.

Regarding Group 4 (*Low Confidence Miss*), we can see a behavior that is very similar to the observed for Group 3. However, in this case, a greater percentage of *Ctx-A* participants can identify the answer correctly.

We can see that overall, a context-aware local explanation favors the error detection (or lack thereof) in the classifier in all but the clearest hits. In contrast, a local explanation is more prone to cause doubts in less clear-cut cases. This supports **H1** as expected.

Regarding the time taken to solve the different tasks, we can see that the comprehension check took around 3 minutes. At the same time, once the general procedure was understood, participants needed only about 1 minute to do each task. It is noticeable that *local* participants are faster than *Ctx-A* participants, possibly due to being shown less information to process before answering the question. This rejects **H2**, indicating that less information does not mean more time to think about the answer (although it does result in somehow smaller accuracy).

Finally, regarding satisfaction, most participants believe that the information shown is useful for solving the task (3.9% Strongly Agree, 61.5% Agree, 23.1% Neutral, and 11.5% Disagree).

VII. Conclusions and Future Work

In this paper, we have developed a new algorithm for extracting context information from a Fuzzy Decision Tree, and we have carried out a user study to validate the goodness of local explanations enriched with such information, what we call context-aware local explanations. The new algorithm exploits the underlying structure of a Fuzzy Decision Tree in which each node represents a fuzzy set that has a *matching degree* for each classified instance. We use *matching degree* as a starting point from which to determine the *importance* of each fuzzy set with respect to a dataset, which we then use to extract *context information*.

In addition, we have shown in a use case related to beer style classification how context-aware local explanations can assist data scientists in detecting errors made by a Fuzzy Decision Tree. Namely, we have empirically validated the goodness of our proposal with a user study that is implemented in the form of an online survey. We distinguished two groups of participants: a first group where participants only interacted with regular local explanations and a second group where participants only interacted with the new context-aware local explanations. A total of 26 subjects took part in the survey (13 participants assigned randomly to each group). In light of the reported results, we can conclude that the research question proposed is correct, given that an improvement in error detection for those participants in the second group is observed, albeit taking more time to fulfill the tasks due to the greater amount of information that they had to process.

Future lines of research can take advantage of the algorithm proposed in this paper to enhance other explanation generation approaches. We use multi-rule factual explanations in this work, but the proposal can be easily applied to single-rule factual explanations that are extracted with different methods. Moreover, we use Fuzzy Decision Trees as the classifier to be explained. Still, the algorithm can also be expanded to other Fuzzy Models with similar characteristics, such as Fuzzy Rule-Based Classifiers. Finally, we test the error analysis detection with these explanations, but they can have more classic XAI uses, such as helping a final user better understand the decision taken by a classifier.

References

[1] T. Chen, T. He, M. Benesty, V. Khotilovich, Y. Tang, H. Cho, K. Chen, R. Mitchell, I. Cano, T. Zhou, *et al.*, "XGBoost: extreme gradient boosting," *R package version 0.4-2*, vol. 1, no. 4, pp. 1–4, 2015.

[2] Z. Dai, H. Liu, Q. V. Le, and M. Tan, "CoAtNet: Marrying convolution and attention for all data sizes," *Advances in neural information processing systems*, vol. 34, pp. 3965–3977, 2021.

[3] S. Zhang, M. Liu, and J. Yan, "The diversified ensemble neural network," *Advances in Neural Information Processing Systems*, vol. 33, pp. 16001–16011, 2020.

[4] P. P. Angelov, E. A. Soares, R. Jiang, N. I. Arnold, and P. M. Atkinson, "Explainable artificial intelligence: an analytical review," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 11, no. 5, p. e1424, 2021.

[5] European Parliament, "General data protection regulation (GDPR)," *Intersoft Consulting, Accessed in October 24*, 2018.

[6] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, *et al.*, "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Information fusion*, vol. 58, pp. 82–115, 2020.

[7] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," *ACM Comput. Surv.*, vol. 51, no. 5, pp. 93:1–93:42, 2019.

[8] T. Speith, "A review of taxonomies of explainable artificial intelligence (XAI) methods," in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 2239–2250, 2022.

[9] R. R. Hoffman, S. T. Mueller, G. Klein, and J. Litman, "Metrics for explainable AI: Challenges and prospects," *arXiv preprint arXiv:1812.04608*, 2018.

[10] C. van der Lee, A. Gatt, E. van Miltenburg, and E. Krahmer, "Human evaluation of automatically generated text: Current trends and best practice guidelines," *Computer Speech & Language*, vol. 67, p. 101151, 2021.

[11] V. Hassija, V. Chamola, A. Mahapatra, A. Singal, D. Goel, K. Huang, S. Scardapane, I. Spinelli, M. Mahmud, and A. Hussain, "Interpreting black-box models: a review on explainable artificial intelligence," *Cognitive Computation*, pp. 1–30, 2023.

[12] E. Lee, Y. Lee, and T. Lee, "Automatic false alarm detection based on XAI and reliability analysis," *Applied Sciences*, vol. 12, p. 6761, 2022.

[13] L. I. Kuncheva, *Fuzzy Classifier Design*. Physica-Verlag GmbH, 2010.

[14] Y.-L. Chen, T. Wang, B. sheng Wang, and Z. jun Li, "A survey of fuzzy decision tree classifier," *Fuzzy Information and Engineering*, vol. 1, no. 2, pp. 149–159, 2009.

[15] A. Segatori, F. Marcelloni, and W. Pedrycz, "On distributed fuzzy decision trees for big data," *IEEE Transactions on Fuzzy Systems*, vol. 26, no. 1, pp. 174–192, 2017.

[16] G. Fernandez, J. A. Aledo, J. A. Gamez, and J. M. Puerta, "Factual and counterfactual explanations in fuzzy classification trees," *IEEE Transactions on Fuzzy Systems*, 2022.

[17] R. Guidotti, A. Monreale, F. Giannotti, D. Pedreschi, S. Ruggieri, and F. Turini, "Factual and counterfactual explanations for black box decision making," *IEEE Intelligent Systems*, vol. 34, pp. 14–23, 2019.

[18] I. Stepin, J. M. Alonso, A. Catala, and M. Pereira-Fariña, "Generation and evaluation of factual and counterfactual explanations for decision trees and fuzzy rule-based classifiers," in *2020 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pp. 1–8, IEEE, 2020.

[19] M. Hiabu, J. T. Meyer, and M. N. Wright, "Unifying local and global model explanations by functional decomposition of low dimensional structures," in *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics* (F. Ruiz, J. Dy, and J.-W. van de Meent, eds.), vol. 206 of *Proceedings of Machine Learning Research*, pp. 7040–7060, PMLR, 2023.

[20] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee, "From local explanations to global understanding with explainable AI for trees," *Nature machine intelligence*, vol. 2, no. 1, pp. 56–67, 2020.

[21] J. Schrouff, S. Baur, S. Hou, D. Mincu, E. Loreaux, R. Blanes, J. Wexler, A. Karthikesalingam, and B. Kim, "Best of both worlds: local and global explanations with human-understandable concepts," *CoRR*, vol. abs/2106.08641, 2021.

[22] T. Huber, K. Weitz, E. André, and O. Amir, "Local and global explanations of agent behavior: Integrating strategy summaries with saliency maps," *Artificial Intelligence*, vol. 301, p. 103571, 2021.

[23] J. M. Alonso *et al.*, "Explainable fuzzy systems: Paving the way from interpretable fuzzy systems to explainable AI systems," *Studies in Computational Intelligence; Springer Nature: Cham, Switzerland*, 2021.

[24] R. Confalonieri and J. M. Alonso-Moral, "An operational framework for guiding human evaluation in explainable and trustworthy AI," *IEEE Intelligent Systems*, vol. 39, pp. 18–28, 2024.

[25] K. F. Punch, *Introduction to social research: Quantitative and qualitative approaches*. sage, 2013.

[26] S. Mcleod, "Qualitative vs quantitative research methods & data analysis." https://www.simplypsychology.org/qualitative-quantitative.html. Accessed: 2024-01-10.

[27] G. Castellano, C. Castiello, and A. M. Fanelli, "The FISDeT software: Application to beer style classification," in *2017 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pp. 1–6, 2017.