Increasing Accuracy and Explainability in Fuzzy Regression Trees: An Experimental Analysis

Alessio Bechini, José Luis Corcuera Bárcena, Pietro Ducange, Francesco Marcelloni, Alessandro Renda,
Department of Information Engineering, University of Pisa, Largo Lucio Lazzarino 1, 56122 Pisa, Italy
Email: {alessandro.renda, alessio.bechini, pietro.ducange, francesco.marcelloni }@unipi.it,
joseluis.corcuera@phd.unipi.it

***Abstract*—Regression Trees (RTs) have been widely used in the last decades in various domains, also thanks to their inherent explainability. Fuzzy RTs (FRTs) extend RTs by using fuzzy sets and have proven to be particularly suitable for dealing with noisy and/or uncertain environments. The modelling capability of FRTs depends, among other factors, on the model used in the leaves for determining the output, and on the inference strategy. Nevertheless, the impact of such factors on FRTs accuracy and explainability has not been adequately investigated.**

In this paper, we extend a recently proposed learning scheme for FRTs by employing both linear models in the leaves and the maximum matching inference strategy. The former extension aims to increase accuracy, and the latter to improve explainability. We carried out an extensive experimental analysis by comparing the four FRT versions corresponding to any possible combination of the two extensions introduced in the paper. The results show that the best trade-off between accuracy and explainability is obtained by employing both of them.

***Index Terms*—Fuzzy Regression Trees, Fuzzy Decision Trees, Regression Models, Explainable Artificial Intelligence**

I. Introduction

In the last years, machine learning (ML) and artificial intelligence (AI) algorithms have spawned a new wave of applications employed in a wide range of industrial settings. In some fields, e.g., medicine, defence and finance, users' trust in such new AI tools depends on the ability to understand their structure and reasoning [1], [2], [3]. Explainable AI (XAI) is concerned with devising AI systems understandable to humans, possibly keeping high levels of performance. In fact, whenever the target task entails a certain complexity, and enough training data are available, the accuracy of ML/AI models and their interpretability are typically at odds: frequently the most accurate approaches (e.g., neural networks and deep learning) are the least interpretable, and the most interpretable (e.g., decision trees) are not the most accurate [4]. The inherent interpretability of AI/ML models can be characterized from a *global* and a *local* point of view: the former refers to the structural properties of the model, whereas the latter refers to the adopted inference process, focusing on how prediction is carried out for any single instance.

Several approaches have been proposed to address *regression* problems: linear and generalized linear regression, least absolute shrinkage and selection operator (LASSO) and ridge regression, least and partial least squares regression (LS and PLS), least angle regression (LARS), and multivariate adaptive regression splines (MARS), among others [5]. Furthermore, there exist methods developed in the field of ML/AI that have been demonstrated to be universal function approximators. Among them, Fuzzy Rule-Based Systems (FRBSs) [6] and Fuzzy Regression Trees (FRTs) [7] are also characterized by a high level of intrinsic interpretability. Indeed, collections of linguistic *if-then* rules can be extracted from these models and used for prediction purposes.

As regards FRBSs, several approaches have been proposed for designing Mamdani-Type and Takagi-Sugeno-Kang (TSK) models. In the former, both antecedent and consequent parts of the rules are expressed linguistically. On the contrary, in TSK models only antecedents are expressed linguistically, while a regression model is used in the consequent part, thus providing a higher modelling capability than the Mamdani-Type. In several practical applications, zero-order and first-order polynomial regression models are used.

FRTs can be considered functionally equivalent to TSK fuzzy systems, as a rule can be extracted by following the branches from the root to a leaf. As per the global interpretability of FRTs, the model complexity can be ordinarily expressed by the number of nodes or leaves, and it can be controlled by properly tuning some stopping criteria, e.g., the maximum depth of the tree. However, the system complexity also depends on the specific regression model used in the leaves: similarly to TSK fuzzy systems, also the leaves of FRTs can hold regression models. In a recent work on a decision support tool for oil spill response in the Arctic [8], authors compared different regression models for leaf nodes of an embedded FRT. Specifically, they adopted first-order and second-order polynomial regression models, and a kernel-based probabilistic regression (Gaussian process) model. The latter model, based on a squared exponential kernel, proved to be the most accurate in their case study. However, authors do not investigate the implications of the adoption of a complex non linear regression model over the system interpretability, and no comparison is given with the classical, zero-order, regression model.

A recent algorithm for learning FRTs for large scale and

This work has been partly funded by the European Commission through the H2020 project Hexa-X (Grant Agreement no. 101015956), by the Italian Ministry of Education and Research (MIUR) in the framework of the CrossLab project (Departments of Excellence) and by the University of Pisa Project PRA2020 "mOSAIc: Open Source as key enabling approach for Artificial Intelligence in healthcare" (PRA_2020_38).

high dimensional regression problems has been proposed in [9]. In this case, the improvement in interpretability due to the adoption of zero-order regression models comes at the cost of a reduction of global transparency, as a consequence of using a binary version of the tree: here, the tests at decision nodes, and thus the rules that can be extracted from the tree, are not expressed in terms of linguistic values. Furthermore, the inference stage is carried out by combining the estimations in each leaf, thus reducing the local interpretability of the trees as well.

In this paper, with the objective of improving both the global and the local explainability of FRTs, while maintaining a good level of accuracy, we extend the state of the art learning scheme proposed in [9]: first, we adopt a multi-way rather than a binary FRT, aiming to obtain highly interpretable rules expressed in terms of linguistic values; second, in an attempt to increase prediction accuracy, in each leaf we use a first-order rather than a zero-order polynomial model; third, as inference strategy, we just consider the most highly activated path (maximum matching policy) rather than aggregating the contributions of multiple rules. In the presented experimental analysis, we compare four variants of the multi-way FRT, and assess the impact of the proposed modifications: specifically, we explore all the possible combinations between the order of the polynomial in the leaves (zero-order and first-order) and the inference strategy (maximum matching and weighted rule aggregation). Results show that, as expected, the FRTs with first-order polynomial models in the leaves outperform the FRTs with zero-order polynomial models. Moreover, the adoption of the maximum matching strategy only marginally impacts the accuracy of FRTs, although yielding a higher explainability level.

The paper is organized as follows: Section II provides some background on fuzzy partitions, FRTs, and their interpretability. Section III describes the proposed modifications on the FRT in [9], to enhance accuracy and interpretability. The setup and results of the experimental analysis are reported and discussed in Section IV. Finally, Section V draws proper concluding remarks.

II. Background

In this section, we introduce some preliminaries about FRTs and their interpretability.

A. Fuzzy Regression Trees

A decision tree (DT) is a rooted tree where the topmost node and each internal (non-leaf) node represents a test on an input variable. Each path from the root node to a leaf node is a sequence of tests on input variables. In the case of numerical input variables, tests are defined by intervals in the form of $X_f > x_{f,s}$ and $X_f \leq x_{f,s}$, where X_f is an input variable and $x_{f,s} \in \mathbb{R}$. In the case of categorical input variables, tests evaluate whether $X_f \subseteq L_{f,s}$, where $L_{f,s}$ is a subset of possible categorical values for X_f. Each test results in a number of branches originating from the node where the test is applied. When a DT allows only two branches in each non-leaf node, the tree is called a *binary* or two-way tree; otherwise, the tree is called a *multi-way* tree.

In classification problems, the sequence of tests from the root to the leaves partitions the input space into subspaces the contain subsets of the training set as "pure" as possible, i.e. containing training instances that belong to the same class. Each leaf node is characterised by a class label, which corresponds to the class of the majority of the instances in the subset of the training set isolated by the sequence [10].

In regression problems, the sequence of tests aims to partition the input space into subspaces that contain subsets of training set with output values very close to each other. Since the goal is the prediction of a real value, leaf nodes are characterized by a regression model defined on the input variables [11], [12]. In this case, DTs are named as *regression trees* (RTs).

The choice of the input variable to be used in the decision node is performed by exploiting appropriate indexes during the learning phase. In DTs, very popular indexes are Gini Index or Information Gain. In RT, variance of the output values is generally used.

Fuzzy set theory has been integrated with DTs and RTs, leading to Fuzzy DTs (FDT) [13], [14] and Fuzzy RT (FRTs) [9].

Let us consider a regression dataset defined by a set of input variables $X = \{X_1, X_2, \ldots, X_F\}$ and the output variable $Y \in \mathbb{R}$. Let U_f, $f = 1, 2, \ldots, F$, be the universe of discourse of variable X_f. Then, fuzzy partitions over U_f with T_f fuzzy sets are defined as $P_f = \left\{A_{f,1}, A_{f,2}, \ldots, A_{f,T_f}\right\}$. In our experiments, we adopt triangular fuzzy sets $A_{f,i}$ described by the tuple ($a_{f,i}$, $b_{f,i}$, $c_{f,i}$), where $b_{f,i}$ is the core and $a_{f,i}$ and $c_{f,i}$ are the left and right extremes of the support of $A_{f,i}$. In order to maintain a high level of explainability of the partitions, we adopt strong uniform fuzzy partitions, where $a_{f,1} = b_{f,1}, b_{f,T_f} = c_{f,Tf}$ and, for $j = 2, ..., T_f - 1$, $b_{f,j} = c_{f,j-1}$ and $b_{f,j} = a_{f,j+1}$. Indeed, since strong fuzzy partitions ensure a high level of coverage, completeness and complementarity, they are widely recognized to be highly interpretable [15]. Unlike classical DTs/RTs, in FDTs/FRTs the tests on internal nodes use fuzzy sets and are expressed in the form of X_f is $A_{f,i}$, where $A_{f,i}$ is a fuzzy set defined over the fuzzy partition of the input variable X_f. One instance may activate different branches, thus reaching multiple leaves with different activation degrees.

In this paper we focus on the zero-order polynomial FRT proposed in [9]. Here, the value assigned to each leaf node is computed as a weighted average (zero-order polynomial) of the output values for all the training set instances that activate such leaf node, using the activation degree as instance weight. Unlike the version described in [9], however, we will adopt a multi-way FRT (instead of a binary one) and uniform partitions instead of the partitions generated according to the approach in [16]. Section III provides more details.

B. Interpretability of Fuzzy Regression Trees

Transparency of ML models, i.e., the property whereby an observer can understand the structure of the model itself, is a key enabler towards XAI [4]. RTs are generally considered among the most transparent models for regression tasks, as the inference process, equivalent to the application of simple *if-then* rules, is very much akin to human reasoning. Furthermore, in the context of FRTs, the adoption of linguistic representation of numerical variables allows a direct human interaction, further enhancing the model interpretability. However, even within the scope of tree-based models, the degree of interpretability may vary depending on several factors.

First, the structural properties, namely number of nodes and leaves, impact on the so-called *global* interpretability [17]: clearly, the more compact the trees, the easier their interpretation is.

Second, the properties of the input variables partition affect the semantic interpretability of FRTs: requirements of coverage, completeness, distinguishability and complementarity, fulfilled e.g. by strong uniform triangular fuzzy partitions, are essential to achieve high interpretability [15].

Third, the inference strategy also plays a crucial role: typically FRTs adopt a weighting approach, in which the output is determined by combining the contributions of all the leaves activated by an input pattern. Thus, the output depends on different paths, making it hard to explain how the result has been obtained. Instead, the maximum matching approach guarantees that only one path is considered for generating the output, leading to a very intuitive and easily comprehensible explanation of how this output has been obtained, according to the linguistic rule relative to the activation path.

Finally, the order of the polynomial model used in the leaf nodes impacts not only global, but also local interpretability, which is associated with the inference process and focuses on how each instance is processed. The interpretability of traditional FRTs with zero-order polynomial regression models in the leaves can be easily ascertained by analysing the formulation of the extracted rules, expressed as:

$$\begin{aligned} R_k : \textbf{IF}\ \ X_1\ is\ A_{1,j_{k,1}}\ \ \textbf{AND}\ \ \ldots\ \ \textbf{AND}\ \ X_F\ is\ A_{F,j_{k,F}} \\ \textbf{THEN}\quad y_k = c_k \end{aligned} \tag{1}$$

However, also in the case of first-order polynomial regression models, when the consequent part is in the form $y_k = \gamma_{k,0} + \sum_{f=1}^{F} \gamma_{k,f} \cdot X_f$, FRTs can be considered interpretable. The local linear model can be interpreted reporting the effect of each input variable on the output value, as expressed by the related coefficient. Furthermore, if all the input variables are defined in the same range of values, we can adopt linguistic labels to characterize the *impact* of an attribute on the output. Formally,

$$Impact_f = \begin{cases} Low & \text{if } |\gamma_f| \le \tan\frac{\pi}{6} \\ Medium & \text{if } \tan\frac{\pi}{6} < |\gamma_f| \le \tan\frac{\pi}{3} \\ High & |\gamma_f| > \tan\frac{\pi}{3} \end{cases} \tag{2}$$

Note that the requirement on the range of the values can be easily met through normalization of the input variables. Figure 1 shows the resulting labelling. Trivially, the impact is positive or negative (i.e., the output increases or decreases) depending on the sign of the coefficient.

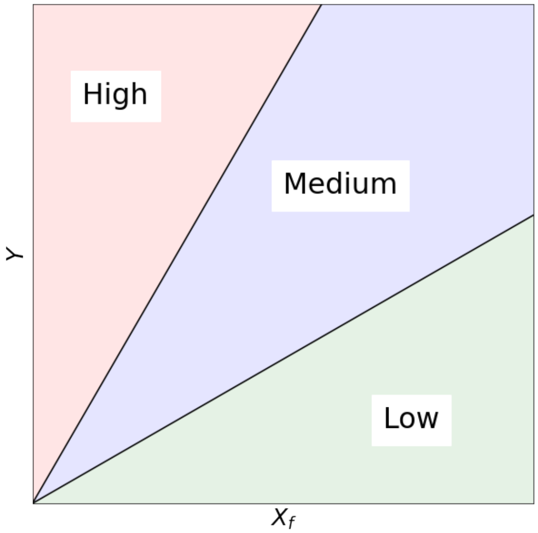

Fig. 1. Linguistic labels used to define the impact of a generic input variable X_f on the output Y.

III. Multi-way Fuzzy Regression Trees as explainable and accurate regression models

In this section, first we describe the basic scheme of the multi-way FRT learning algorithm that we adopted in our experimental analysis. The algorithm is based on the FRT learning scheme, discussed in [9], for generating binary FRTs. Then, we discuss about two possible options for creating the local regression models in the leaves of the trees. Finally, we argue on the inference strategies that can be adopted to estimate the output values, namely the *maximum matching approach* and the *weighted average strategy*.

A. Model Learning Scheme

Algorithm 1 shows the pseudo-code of the scheme for learning multi-way FRTs, where **TR** = $\{\ (x_1, y_1), ..., (x_N, y_N)\ \}$ is the training set, **X** is the set of input variables, *SelectBestInput* is the method for selecting the most relevant input variable and *StopMet* is the method that checks if a stopping condition is verified.

The algorithm starts from creating a root node. Then, it builds a multi-way FRT model by calling the recursive function *TreeGrowing*. This function first checks if the node, passed as a parameter, is a leaf node through the *StopMet* function. In case this check is not satisfied, the *SelectBestInput* function is called. This function selects the best input variable, which partitions the input space. The *SelectBestInput* requires the definition of a criterion for the choice of the input variable to be used in the node. As in [9], we adopt the Partition Fuzzy Gain (PFGain), which exploits the concept of fuzzy variance [18], as a metric for selecting the most relevant input variable

```
Data: training set TR, set X of input variables, input
      variable selection method $SelectBestInput$,
      stopping method $StopMet$
Result: FRT
$root$ ←create a new node;
$tree$ ←TreeGrowing(root, TR, X, SplitMet, StopMet);
return $tree$;
Function TreeGrowing(node, S, X, SplitMet,
 StopMet)
    if $StopMet(node)$ then
        $node$ ←mark node as leaf;
    else
        $X_{best}$ ←SelectBestInput( X, S, SplitMet);
        $X \leftarrow X - X_{best}$;
        foreach $A_{X_{best},i}$ in $P_{X_{best}}$ do
            $S_{A_{X_{best},i}}$ ←get instances from $S$ falling in
             $A_{X_{best},i}$;
            $child_{A_{X_{best},i}}$ ←create a node by using
             $A_{X_{best},i}$ and $S_{A_{X_{best},i}}$;
            $node$ ←connect the node with
             TreeGrowing($child_{A_{X_{best},i}}$, $S_{A_{X_{best},i}}$, $X$
             $SplitMet$, $StopMet$) ;
        end
    end
    return $node$;
end
```

Algorithm 1: Pseudo-code of the Multi-way FRT learning process

at each recursive stage of the FRT learning algorithm. Once the most relevant input variable has been selected, a splitting node is created for each fuzzy set of its fuzzy partition and the instances falling in its support are assigned to the new node. After the creation of the new nodes, the *TreeGrowing* is recursively called from each of them.

As regards the original learning scheme discussed in [9], there are two main differences. First, with the aim of obtaining the highest integrity level of the input fuzzy partitions, we adopted strong triangular fuzzy partitions with only three fuzzy sets, rather than running a fuzzy discretization algorithm. Indeed, the fuzzy discretization algorithm can automatically generate a strong fuzzy partition in which the fuzzy sets in the partitions are optimized in terms of both granularity (i.e. number of fuzzy sets) and the position of the cores. We realized that running the discretization algorithm we can take the risk of generating partitions with a large number of fuzzy sets even very irregularly distributed. Second, it is worth to notice that, in the original scheme for generating binary FRTs, the method for selecting the most relevant input variable chooses a splitting point from a subset of possible splitting points for generating two branches. In the scheme for learning multi-way decision trees, the decision branches are represented by the fuzzy sets which compose the partition of the selected input variable.

In our experimental analysis, as in [9], we use the following criteria to stop the tree growth (*StopMet* function):

- when the number of instances in a node is lower than a fraction λ of instances in the training set;
- when PFGain computed for the best-selected attribute is lower than a fixed threshold ϵ;
- when the maximum depth allowed in the FRT, defined by the parameter β, is achieved.

B. Local Regression Model Estimation

After learning the tree structure, each leaf node estimates a regression model from the instances of the training set that belong to the leaf node with a membership degree larger than zero. In our experimental analysis, we adopted both zero-order and first-order polynomial models.

Let K be the number of leaf nodes in an FRT. The path from the root to the generic k^{th} leaf node can be described by the following rule:

$$R_k : \textbf{IF} \ X_1 \ is \ A_{1,j_{k,1}} \ \textbf{AND} \ \dots \ \textbf{AND} \ X_{F_k} \ is \ A_{F_k,j_{k,F}} \quad \textbf{THEN} \ \ Y = f_k(\mathbf{X}) \tag{3}$$

where $j_{k,f} \in [1, T_f]$ identifies the index of the fuzzy set of partition P_f of input variable X_f used in the rule R_k. In the case of the zero-order polynomial regression model, which has been adopted in [9], $f_k(\mathbf{X}) = c_k$, where c_k is calculated as the weighted average of the output values y_z of each training instance. The weight is the strength of activation of the rule R_k, which, given the input pattern $\mathbf{x}_z$ corresponding to the output value y_z, is computed as:

$$w_k(\mathbf{x}_z) = \prod_{f=1}^{F_k} \mu^k_{f,j_{k,f}}(x_{z,f}) \tag{4}$$

where $\mu^k_{f,j_{k,f}}(x_f)$ is the membership degree of $x_{z,f}$ to the fuzzy set $A^k_{f,j_{k,f}}$ of the partition of each input variable chosen in the path from the root to the k^{th} leaf node. We observe that only the instances $\mathbf{x}_n$ with $w_k(\mathbf{x}_n) > 0$ are considered in the computation of c_k.

In the case of first-order polynomial regression model, $f_k(\mathbf{X}) = \gamma_{k,0} + \sum_{f=1}^{F} \gamma_{k,f} \cdot X_f$. The coefficients $\boldsymbol{\gamma_k} = \{\gamma_{k,0}, \gamma_{k,1}, \dots \gamma_{k,F}\}$ of the consequent part of each rule can be estimated by applying a local weighted least-squared method. Specifically, in the estimation of the parameters, each training sample ($\mathbf{x}_z$, y_z) with a membership value different from 0 to the specific leaf is weighted by its strength of activation of the rule.

C. Inference Stage

In the inference stage, an input pattern is fed to the FRT for estimating the corresponding output value. Due to the fuzzy partitioning of the input variables, the input pattern may activate more than one path from the root to the leaves. Thus, more than one rule is fired with different strengths of activation. Two main strategies have been adopted in the literature for generating an output value: *voting method* and *maximum matching* [19]. Traditional fuzzy models based on rules for regression problems adopt the *weighted average* as voting method: the output values estimated by each rule are weighted by the strength of activation. This approach has been also adopted in the binary FRT proposed in [9].

When the maximum matching approach is adopted, only the rule with the highest strength of activation is used for estimating the output value. Using a single rule may reduce the modelling capability of the FRT, and consequently decrease its accuracy. However, the local interpretability of an FRT which takes its decision based on a single rule is much higher than the one of an FRT which adopts the weighted average for taking decisions [17].

In our experimental analysis we compare the two inference strategies, assessing their impact on the modelling capability of FRT models. In the experiments, we consider the two cases in which zero-order and first-order polynomial regression models are used in the leaves.

IV. Experimental Analysis

In this section, we first describe our experimental setup, including details regarding the datasets employed and the FRTs configuration. Then, we show and discuss the results obtained by four variants of multi-way FRTs. Such variants pursue different trade-offs between interpretability and accuracy, and stem from the state-of-the-art proposal for building FRTs [9], as widely discussed in Section III.

A. Experimental Setup

The four different variants of FRTs used in the experiments are generated by combining two dimensions, namely the inference strategy (maximum matching or weighted average) and the model used in the leaves (zero-order or first-order polynomial regression models). The resulting variants are described in the following:

- **FRT-MM-0**: FRT with zero-order polynomial regression model in the leaves and maximum matching strategy for inference.
- **FRT-WA-0**: FRT with zero-order polynomial regression model in the leaves and weighted average strategy for inference.
- **FRT-MM-1**: FRT with first-order polynomial regression model in the leaves and maximum matching strategy for inference.
- **FRT-WA-1**: FRT with first-order polynomial regression model in the leaves and weighted average strategy for inference.

All the variants use the following parameter configuration:

- $T_f = 3, \quad \forall f \in \{1, \ldots, F\}$, to ensure high interpretability: the three fuzzy sets in the partitions are associated with the linguistic terms *Low*, *Medium* and *High*, respectively, for linguistically describing the rules;
- $\lambda = 0.05$, as the minimum fraction of instances of the training set that activate a node to continue building the tree; it is used as a stopping criterion (see Section III-A);
- $\epsilon = 0.0001$, as the minimum PFGain threshold for a split during the FRT induction; it is used as a stopping criterion (see Section III-A)

The configurations of FRTs with zero-order and first-order polynomial regression models in the leaves only differ in the value of the parameter β, which defines the maximum depth of the FRT. For a fair comparison, we set the parameter β so as to achieve a comparable complexity between the two types of FRT and specifically $\beta = 8$ for FRT-MM-0 and FRT-WA-0, and $\beta = 4$ for FRT-MM-1 and FRT-WA-1.

Model complexity (C_{FRT}) depends on the structural properties of the FRT and is defined as follows. For FRTs with zero-order polynomial regression model in the leaves, complexity is defined as the total number of nodes N_{FRT} in the FRT. Formally:

$$C_{FRT-0} = N_{FRT} \tag{5}$$

For FRTs with a first-order polynomial regression model in the leaves, complexity is defined as the sum of the total number of internal nodes plus the total number of coefficients defined in the first-order polynomial regression model for all leaf nodes. Formally:

$$C_{FRT-1} = IN_{FRT} + \sum_{LN \in FRT} N_{Coeff}(LN) \tag{6}$$

where IN is the total number of internal nodes, LN is a generic leaf node in the FRT, and $N_{Coeff}(LN)$ is the number of the coefficients of the linear model used in LN. As the first-order polynomial regression model is estimated considering all the attributes of the input patterns, the number of coefficients is $F+1$ for any leaf.

We employ ten well-known regression datasets available within the KEEL [20] and Torgo's[1] dataset repositories: Weather Izmir, Treasury, Mortgage, Computer Activity, California Housing, Analyzing Categorical Data, Elevators, House_16H, MV Artificial Domain and Pumadyn. Details about the datasets are reported in Table I.

TABLE I
Dataset description

Dataset	Dimensionality (F)	Samples (N)
Weather Izmir (WI)	9	1461
Treasury (TR)	15	1049
Mortgage (MO)	15	1049
Computer Activity (CA)	21	8192
California Housing (CH)	8	20460
Analyzing Categorical Data (AN)	8	4052
Elevators (EL)	19	16599
House_16H (HO)	17	22784
MV Artificial Domain (MV)	11	40768
Pumadyn (PM)	9	8192

The quality of prediction of the FRTs is evaluated through the *Mean Squared Error* (MSE):

$$MSE = \frac{1}{N_{test}} \sum_{i=1}^{N_{test}} (y_i - \hat{y_i})^2 \tag{7}$$

[1] https://www.dcc.fc.up.pt/ ltorgo/Regression/DataSets.html

where N_{test} is the number of samples considered for the evaluation, y_i and $\hat{y}_i$ are the ground truth value and the predicted value associated with the i-th instance of the test set, respectively. Results are evaluated in terms of average values over 5-fold cross-validation: for a fair comparison, at each iteration of the cross-validation, the same split is used for the four types of FRTs.

B. Experimental Results

Table II shows the results obtained by the four FRT variants described in subsection IV-A. As 5-fold cross-validation is adopted, for each FRT configuration and each dataset, we report the average MSE over 5 values measured on both the test and training sets. Best values are in bold.

Results point out that FRT-MM-1 and FRT-WA-1 always outperform FRT-MM-0 and FRT-WA-0. Further, none of the FRTs is particularly affected by overtraining. Also, results suggest that the impact of the inference strategy is more evident in FRTs with zero-order rather than first-order polynomial models in the leaves: the generalization capability (measured as MSE on test set) of FRT-MM-0 is almost halved compared to FRT-WA-0 in four out of ten datasets (WeatherIzmir, Treasury, Mortgage, MV Artificial Domain); on the other hand, the increased modelling capability of first-order polynomial models makes the FRTs more robust with respect to the choice of the inference strategy: FRT-MM-1 is comparable or just slightly outperformed by FRT-WA-1, which in turn is less interpretable than FRT-MM-1.

In summary, when the interpretability of the model is an essential requirement, FRT-MM-0 has to be used as it provides the highest degree of interpretability; however, it has a restricted modelling power. Instead, the proposed FRT-MM-1 represents a suitable trade-off in applications for which both accuracy and interpretability are deemed crucial.

Table III reports the complexity of the models, as discussed in Section IV-A. The lowest values of number of rules and complexity are highlighted in bold.

As the depth parameter is discrete, it does not allow for an exact match between the complexity of zero-order and first-order polynomial models; indeed, first-order polynomial models are in general slightly more complex than the zero-order counterparts, based on the definition of complexity provided in Eq. 5 and Eq. 6. However first-order polynomial model (FRT-MM-1 and FRT-WA-1) features a lower number of leaves (and consequently of rules) compared to the zero-order model (FRT-MM-0 and FRT-WA-0) and the antecedent of each rule has a lower maximum number of conditions.

Finally, in the following, we report an example of rule generated from an FRT learned from the *California* dataset.

TABLE II
AVERAGE MSE AND STANDARD DEVIATION OVER CROSS-VALIDATION FOR EACH DATASET AND FOR EACH FRT VARIANT.

FRT	**MSE train**	**STD train**	**MSE test**	**STD test**
Weather Izimir				
FRT-MM-0	**2**7.56	6.83	27.98	6.31
FRT-WA-0	14.55	0.29	14.89	1.14
FRT-MM-1	1.31	0.07	1.40	0.30
FRT-WA-1	**1.22**	0.07	**1.32**	0.28
Treasury (x 10^{-3})				
FRT-MM-0	789.32	66.74	844.54	91.02
FRT-WA-0	398.72	6.31	407.18	73.62
FRT-MM-1	31.49	4.29	43.03	19.13
FRT-WA-1	**30.45**	4.14	**39.05**	19.67
Mortgage (x 10^{-3})				
FRT-MM-0	635.03	71.88	620.32	125.12
FRT-WA-0	303.06	1.05	310.16	24.13
FRT-MM-1	5.90	0.32	8.31	1.84
FRT-WA-1	**5.12**	0.20	**6.90**	1.36
Computer Activity				
FRT-MM-0	66.89	8.96	68.14	9.92
FRT-WA-0	61.36	8.89	62.18	9.90
FRT-MM-1	**5.94**	0.09	**6.30**	0.29
FRT-WA-1	5.95	0.04	8.48	4.39
California Housing (x 10^{-9})				
FRT-MM-0	8.56	0.08	8.57	0.16
FRT-WA-0	9.54	0.03	9.54	0.14
FRT-MM-1	4.25	0.06	4.28	0.19
FRT-WA-1	**4.11**	0.03	**4.15**	0.14
Analyzing Categorial Data				
FRT-MM-0	0.14	0.00	0.14	0.02
FRT-WA-0	0.14	0.00	0.15	0.02
FRT-MM-1	0.04	0.00	0.04	0.00
FRT-WA-1	**0.03**	0.00	**0.03**	0.01
Elevators (x 10^{-5})				
FRT-MM-0	3.03	0.03	3.05	0.18
FRT-WA-0	2.67	0.03	2.67	0.16
FRT-MM-1	0.62	0.01	0.62	0.03
FRT-WA-1	**0.60**	0.00	**0.60**	0.02
House_16H (x 10^{-9})				
FRT-MM-0	2.43	0.03	2.44	0.10
FRT-WA-0	2.43	0.03	2.44	0.10
FRT-MM-1	**1.63**	0.03	**1.68**	0.14
FRT-WA-1	**1.63**	0.02	**1.68**	0.10
MV Artificial Domain				
FRT-MM-0	27.51	0.25	27.49	0.74
FRT-WA-0	14.77	0.06	14.78	0.25
FRT-MM-1	**0.05**	0.00	**0.05**	0.00
FRT-WA-1	**0.05**	0.00	**0.05**	0.00
Pumadyn				
FRT-MM-0	16.22	0.81	16.56	0.64
FRT-WA-0	16.90	0.19	17.24	0.62
FRT-MM-1	11.82	0.28	12.33	0.55
FRT-WA-1	**11.25**	0.13	**11.62**	0.58

TABLE III
COMPLEXITY ANALYSIS: NUMBER OF RULES AND OVERALL COMPLEXITY OF FRTS (SEE EQ.5 AND EQ. 6)

Dataset	Num. rules	Complexity
FRT-MM-0 / FRT-WA-0		
Weather Izimir	241	**361**
Treasury	335	**553**
Mortgage	453	**752**
Computer Activity	285	**428**
California Housing	229	**366**
Analyzing Categorial Data	17	**28**
Elevators	202	**305**
House_16H	248	**372**
MV Artificial Domain	63	**99**
Pumadyn	733	1098
FRT-MM-1 / FRT-WA-1		
Weather Izimir	**64**	674
Treasury	**46**	761
Mortgage	**53**	872
Computer Activity	**42**	952
California Housing	**51**	486
Analyzing Categorial Data	**11**	88
Elevators	**36**	618
House_16H	**24**	448
MV Artificial Domain	**37**	426
Pumadyn	**47**	**442**

$$\begin{aligned}
R_k :\ & \textbf{IF}\ MedianIncome\ is\ Low \\
& \textbf{AND}\ Latitude\ is\ Low \\
& \textbf{AND}\ Longitude\ is\ Medium \\
& \textbf{THEN} : MedianHouseValue = 0.89+ \\
& - 1.10 \cdot Longitude - 1.03 \cdot Latitude+ \\
& + 0.10 \cdot HousingMedianAge - 1.56 \cdot TotalRooms+ \\
& + 2.08 \cdot TotalBedrooms - 2.33 \cdot Population+ \\
& + 0.41 \cdot Households + 1.27 \cdot MedianIncome
\end{aligned} \tag{8}$$

As discussed in Section II-B, we can characterize the impact of each attribute on the output value produced by the local linear model. For instance, the *Population* has a *high* negative impact on the *MedianHouseValue* (coefficient -2.33), whereas the *TotalBedrooms* has a *high* positive impact on it (coefficient $+2.08$).

V. CONCLUSION

In this paper we presented some variants of a state-of-the-art Fuzzy Regression Tree (FRT), and analysing the possible different trade-offs between accuracy and interpretability provided by such variants. In particular, we compared the impact of adopting a first-order versus a zero-order polynomial regression model at each leaf node; moreover, we analysed two different inference strategies, namely maximum matching and weighted average. The FRTs rely on strong uniform triangular fuzzy partitions of the input variables. The empirical comparison has been carried out on ten publicly available datasets, and results are evaluated in terms of Mean Squared Error. It has been observed that, given a comparable complexity measured in terms of number of parameters in the FRT, the adoption of a first-order polynomial model in the leaves leads to better results than the classical approach based on zero-order polynomial models. Furthermore, the adoption of a maximum matching approach does not significantly degrade the modelling power of FRTs compared to the weighted average strategy, yet ensuring a higher level of interpretability. The proposed FRT variant with first-order polynomial model in the leaves and maximum matching as inference strategy represents thus an effective solution for applications that demand for both high accuracy and high explainability.

REFERENCES

[1] W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, and K.-R. Müller, *Explainable AI: interpreting, explaining and visualizing deep learning*. Springer Nature, 2019, vol. 11700.

[2] H. J. Escalante, S. Escalera, I. Guyon, X. Baró, Y. Güçlütürk, U. Güçlü, M. van Gerven, and R. van Lier, *Explainable and interpretable models in computer vision and machine learning*. Springer, 2018.

[3] O. Biran and C. Cotton, "Explanation and justification in machine learning: A survey," in *IJCAI-17 workshop on explainable AI (XAI)*, vol. 8, no. 1, 2017, pp. 8–13.

[4] 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.

[5] D. A. Freedman, *Statistical models: theory and practice*. Cambridge University Press, 2009.

[6] B. Kosko, "Fuzzy systems as universal approximators," *IEEE transactions on computers*, vol. 43, no. 11, pp. 1329–1333, 1994.

[7] A. Suárez and J. F. Lutsko, "Globally optimal fuzzy decision trees for classification and regression," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, no. 12, pp. 1297–1311, 1999.

[8] S. Mohammadiun, G. Hu, A. A. Gharahbagh, R. Mirshahi, J. Li, K. Hewage, and R. Sadiq, "Optimization of integrated fuzzy decision tree and regression models for selection of oil spill response method in the arctic," *Knowledge-Based Systems*, vol. 213, p. 106676, 2021.

[9] J. Cózar, F. Marcelloni, J. Gámez, and L. de la Ossa, "Building efficient fuzzy regression trees for large scale and high dimensional problems," *Journal of Big Data*, vol. 5, 12 2018.

[10] L. Rokach and O. Maimon, "Decision trees," in *Data mining and knowledge discovery handbook*. Springer, 2005, pp. 165–192.

[11] J. R. Quinlan *et al.*, "Learning with continuous classes," in *5th Australian joint conference on artificial intelligence*, vol. 92. World Scientific, 1992, pp. 343–348.

[12] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*. Monterey, CA: Wadsworth and Brooks, 1984.

[13] Y.-l. Chen, T. Wang, B.-s. Wang, and Z.-j. Li, "A survey of fuzzy decision tree classifier," *Fuzzy Inf. Eng.*, vol. 1, no. 2, pp. 149–159, 2009.

[14] A. Renda, P. Ducange, G. Gallo, and F. Marcelloni, "Xai models for quality of experience prediction in wireless networks," in *2021 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. IEEE, 2021, pp. 1–6.

[15] M. J. Gacto, R. Alcalá, and F. Herrera, "Interpretability of linguistic fuzzy rule-based systems: An overview of interpretability measures," *Inf. Sci.*, vol. 181, no. 20, pp. 4340–4360, 2011.

[16] 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.

[17] J. M. Alonso, P. Ducange, R. Pecori, and R. Vilas, "Building explanations for fuzzy decision trees with the expliclas software," in *2020 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. IEEE, 2020, pp. 1–8.

[18] J. Cózar, L. d. l. Ossa, and J. A. Gámez, “Tsk-0 fuzzy rule-based systems for high-dimensional problems using the apriori principle for rule generation,” in *International Conference on Rough Sets and Current Trends in Computing*. Springer, 2014, pp. 270–279.

[19] L. Magdalena, “Fuzzy rule-based systems,” in *Springer Handbook of Computational Intelligence*. Springer, 2015, pp. 203–218.

[20] J. Alcalá-Fdez, A. Fernández, J. Luengo, J. Derrac, S. García, L. Sánchez, and F. Herrera, “Keel data-mining software tool: data set repository, integration of algorithms and experimental analysis framework.” *Journal of Multiple-Valued Logic & Soft Computing*, vol. 17, 2011.