On Clustering and Interpreting with Rules by Means of Mathematical Optimization

Carrizosa, Emilio; Kurishchenko, Kseniia ; Marín, Alfredo ; Romero Morales, Dolores

Document Version
Final published version

Published in:
Computers & Operations Research

DOI:
10.1016/j.cor.2023.106180

Publication date:
2023

Citation for published version (APA):
Carrizosa, E., Kurishchenko, K., Marín, A., & Romero Morales, D. (2023). On Clustering and Interpreting with Rules by Means of Mathematical Optimization. *Computers & Operations Research*, *154*, Article 106180. https://doi.org/10.1016/j.cor.2023.106180

Computers & Operations Research 154 (2023) 106180

Contents lists available at ScienceDirect

Computers and Operations Research

journal homepage: www.elsevier.com/locate/cor

On clustering and interpreting with rules by means of mathematical optimization

Emilio Carrizosa [a], Kseniia Kurishchenko [b,*], Alfredo Marín [c], Dolores Romero Morales [b]

[a] Instituto de Matemáticas de la Universidad de Sevilla, Seville, Spain
[b] Department of Economics, Copenhagen Business School, Frederiksberg, Denmark
[c] Departamento de Estadística e Investigación Operativa, Universidad de Murcia, Murcia, Spain

A R T I C L E I N F O

Keywords:
Machine learning
Interpretability
Cluster analysis
Rules
Mixed-integer programming

A B S T R A C T

In this paper, we make Cluster Analysis more interpretable with a new approach that simultaneously allocates individuals to clusters and gives rule-based explanations to each cluster. The traditional homogeneity metric in clustering, namely the sum of the dissimilarities between individuals in the same cluster, is enriched by considering also, for each cluster and its associated explanation, two explainability criteria, namely, the accuracy of the explanation, i.e., how many individuals within the cluster satisfy its explanation, and the distinctiveness of the explanation, i.e., how many individuals outside the cluster satisfy its explanation. Finding the clusters and the explanations optimizing a joint measure of homogeneity, accuracy, and distinctiveness is formulated as a multi-objective Mixed Integer Linear Optimization problem, from which non-dominated solutions are generated. Our approach is tested on real-world datasets.

1. Introduction

Researchers and practitioners need to interpret the results of black-box machine learning models for model selection (Baesens et al., 2003; Bertsimas and King, 2016; Carrizosa and Romero Morales, 2013; Carrizosa et al., 2021; Hazimeh and Mazumder, 2020; Mišić, 2020), as well as to comply with legal and ethical requirements (European Commission, 2020; Goodman and Flaxman, 2017; Rader et al., 2018; Rodrigues, 2020). This explains the growing literature on Interpretable Machine Learning, such as transparent neural networks (Samek et al., 2021), interpretable random forests (Bénard et al., 2019), or sparse support vector machines (Benítez-Peña et al., 2019; Carrizosa et al., 2016; Jiménez-Cordero et al., 2021). In this paper, we contribute to the literature of Cluster Analysis (Aloise et al., 2012; Kaufmann and Rousseeuw, 1990), which is important in applications arising in, e.g., security (Corral et al., 2009), internet traffic (Morichetta et al., 2019), finance (Gibert and Conti, 2016), sales profiling (Thomassey and Fiordaliso, 2006), or astronomy (Ma et al., 2018). Our goal is to enhance the interpretability of Cluster Analysis by providing accurate and distinctive explanations for the clusters.

Two different scenarios are considered. In the first one, clusters are externally given, as is the case in Balabaeva and Kovalchuk (2020), Carrizosa et al. (2022), Davidson et al. (2018), De Koninck et al. (2017), Kauffmann et al. (2022) and Lawless et al. (2022). The goal of the problem is to find a rule-based explanation for each cluster, such that the explanation is as accurate and distinctive as possible. In the second scenario, both clusters and rule-based explanations are to be found, seeking for each cluster intra-homogeneity as well as an explanation that is as accurate and distinctive as possible.

Throughout this paper, we assume we are given a set of auxiliary features to construct the explanations of the clusters, as is done in other Data Analysis tools (Carrizosa et al., 2020; Taeb and Chandrasekaran, 2018). We explain clusters by a combination of rules defined by these features, and joined with the *AND* operator. To ensure these explanations are easily understood, we limit to a small number ℓ (in our numerical results $\ell = 2$) the number of rules to be concatenated by the AND operator.

As a running example, we will use the `housing` dataset, one the datasets used in our numerical section, where the observations correspond to houses characterized by the thirteen features found in Table 2. Records in the housing dataset are labeled, and their label identifies the cluster. In this case we are thus assuming that (two) clusters are already defined, and that we are interested in associating to them an explanation. With our methodology, a possible explanation for cluster 1 will be (RM > 5.9505) AND (LSTAT ≤ 13.33), while a possible one for cluster 2 would be (RM ≤ 6.75) AND (LSTAT > 7.765), see Table 4.

The first contribution of this paper is to design a procedure to explain existing clusters in a post-hoc fashion with our rule-based

* Corresponding author.
E-mail addresses: ecarrizosa@us.es (E. Carrizosa), kk.eco@cbs.dk (K. Kurishchenko), amarin@um.es (A. Marín), drm.eco@cbs.dk (D. Romero Morales).

https://doi.org/10.1016/j.cor.2023.106180
Received 9 February 2022; Received in revised form 21 October 2022; Accepted 31 January 2023
Available online 2 March 2023

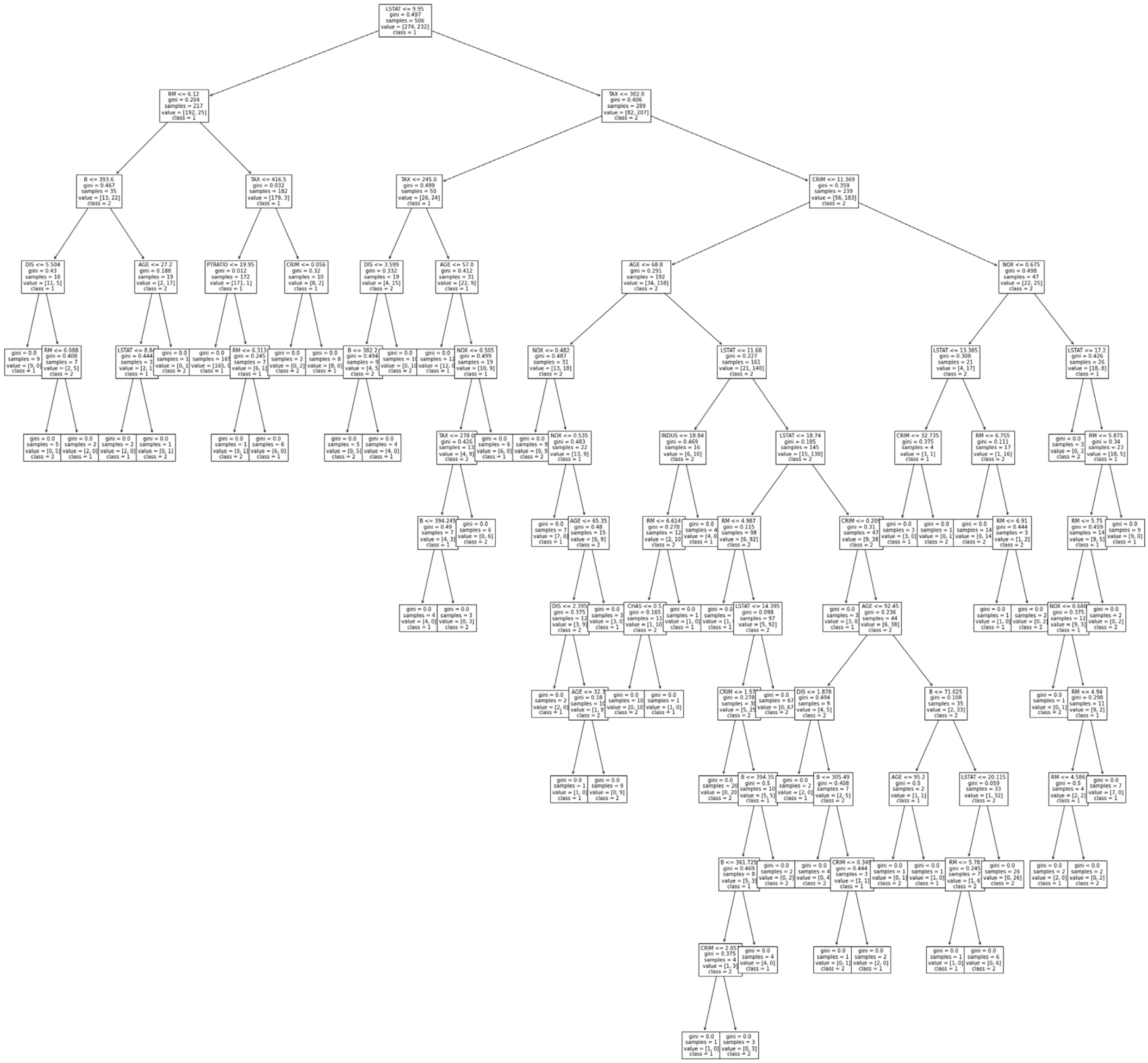


Fig. 1. The post-hoc explanations provided by CART for the `housing` dataset for clusters (classes) 1 and 2.

explanations. Since clusters are already given, we can see the problem as a supervised classification problem in which we want to link via rules the features with the clusters labels. To address this problem, any rule-based supervised classification methodology, such as Classification and Regression Trees (CART), could be used to obtain the rules explaining the clusters. This is illustrated in Fig. 1 for the `housing` dataset. CART, in general, provides explanations which are long with several rules joined with AND and OR operators, while the goal of our approach will be to derive easy to understand explanations using only a few rules joined by the AND operator that are not necessarily arranged in a tree hierarchical structure. The second contribution of this paper is a novel clustering approach to simultaneously find clusters and a rule-based explanation for each of them.

There is a stream of literature on approaches, where interpretability is sought by constructing unsupervised decision trees, see Bertsimas et al. (2021), Basak and Krishnapuram (2005) and Fraiman et al. (2013) and references therein. A set of features is used to measure the intra-homogeneity of the clusters, as well as to define explanations for the clusters. The leaf nodes of the tree define the clusters, while the splitting rules at the branch nodes are used to explain the clusters. In the simplest case, in which each cluster is assigned to a single leaf node, the explanation will correspond to the conjunction of the rules found in the path from the root node to the leaf node. If a cluster is split across different leaf nodes, the explanation will combine the path rules using the OR operator. The goal is to construct an unsupervised decision tree, as well as the K clusters and their explanations, such that a measure of their intra-homogeneity of the clusters is minimized. Alternatively, in Dasgupta et al. (2020), the authors construct an unsupervised decision tree with the goal of making as few changes as possible to the clusters obtained by K-means, measuring the intra-homogeneity of new clusters using the original K-means centers. Finally, see, e.g., Chen et al. (2016), Kim et al. (2014) and Saisubramanian et al. (2020) for rule-based explanations not necessarily arranged in a tree hierarchical structure.

The quality of the explanations is measured through their accuracy (number of true positive cases) and their distinctiveness (number of false positive cases). Indeed, we would like to ensure that the explanation of cluster k, e_k, is accurate, and thus true for most of the individuals in the cluster, but also that the explanation is distinctive to the individuals in cluster k versus the rest, and thus e_k is not true for too many of the individuals outside the cluster. We therefore first count the number of individuals in cluster k that satisfy its explanation, i.e., the true positive cases of explanation e_k. Second, we count the number of individuals outside cluster k that satisfy explanation e_k, i.e., the false positive cases of e_k. Let us illustrate these two criteria in the housing dataset, when the clusters are given by the class labels mentioned above. Let us focus on cluster 1 and assume that this is explained by the rule e_1 of length two (RM > 5.9505) AND (LSTAT ≤ 13.33). There are 214 out of the 274 individuals in cluster 1 that satisfy e_1, while 42 of the individuals outside cluster 1, i.e., in cluster 2, satisfy this explanation. Thus, in relative terms, the quality of the explanation assigned to cluster 1 is the true positive rate (TPR), $\frac{214}{274} = 0.78$ (1 being the ideal value), and the false positive rate (FPR), $\frac{42}{232} = 0.18$ (0 being the ideal value).

In this paper, we propose a mathematical optimization formulation for each of the problems described above. In the first formulation, we simultaneously split the individuals into K clusters using a dissimilarity $\boldsymbol{\delta}$ to measure the intra-homogeneity of the clusters, and choose the rule-based explanations of length at most ℓ. We consider three objectives, namely, the maximization of the intra-homogeneity of the clusters, by minimizing the sum of the dissimilarities between individuals in the same cluster, the maximization of the accuracy of explanations, by maximizing the total number of true positive cases across all clusters, and the maximization of the distinctiveness of explanations, by minimizing the total number of false positive cases across all clusters. We address this multi-objective optimization problem using a weighted approach and formulate it as a Mixed Integer Linear Programming (MILP) problem. In the second formulation, in which the clusters are given, the accuracy and the distinctiveness of the explanations are optimized.

The paper is organized as follows. In Section 2, we introduce the mathematical optimization model that clusters individuals and assigns rule-based explanations to them. In Section 3, this model is tailored to the post-hoc setting in which the clusters are given and we just seek an explanation for each of them. In Section 4, we illustrate the performance of these two models on real-world datasets. By solving the MILP formulations with different weights, different non-dominated solutions of clusters and explanations are obtained. In Section 5 we provide some conclusions and discuss future lines of research. The paper ends with an appendix, containing some of the numerical results from Section 4.

2. Building simultaneously clusters and explanations

In this section, we introduce a mathematical optimization model that finds clusters and explanations for them simultaneously. We assume that we have at hand a dissimilarity between the individuals, δ_{ij}, and that, in addition, the individuals have associated a set of auxiliary features. The dissimilarity can be a distance-based one, such as the squared Euclidean distance, but also a dissimilarity violating e.g. the triangle inequality (Kaufmann and Rousseeuw, 1990). Moreover, $\boldsymbol{\delta}$ does not need to be based on the features used to build rules and explanations.

With the features, we can build $\mathcal{N}$, a collection of N *if-then rules*. We assume that $\mathcal{N}$ is split into S groups, $\mathcal{N} = \cup_{s=1}^{S} \mathcal{N}_s$ and $\mathcal{N}_s \cap \mathcal{N}_{s'}$ if $s \neq s'$, and define the possible explanations for a cluster as the combination of at most ℓ rules joined with the AND operator, where we select at most one rule from each set $\mathcal{N}_s$. To ensure that the explanations are easy to understand, ℓ should be small, ideally $\ell \leq 2$. The group $\mathcal{N}_s$ is composed of the rules relating to one feature, but they could be associated with a group of features, such as socio-economic features or demographic ones. In our numerical section, we have 13 groups for the housing dataset, one per each feature in Table 2.

Below we introduce the notation used in this section relating to the individuals, the dissimilarity between them, the rules based on features characterizing the individuals, and whether the individuals satisfy the rules or not. In addition, we also present the notation for the decision variables in our mathematical optimization formulation of the problem, namely, decisions on the cluster membership for each individual, the choice of the rules composing the explanation of maximum length ℓ for each cluster, and decision variables about the true positive cases and the false positive cases of the explanation assigned to each cluster.

Indices and sets	
k	$\in \{1,\dots,\mathrm{K}\}$ for clusters,
i, j	$\in \{1,\dots,\mathrm{I}\} = \mathcal{I}$ for individuals,
s	$\in \{1,\dots,\mathrm{S}\}$ for groups of rules,
n	$\in \{1,\dots,\mathrm{N}\} = \mathcal{N} = \cup_{s=1}^{\mathrm{S}} \mathcal{N}_s : \mathcal{N}_s \cap \mathcal{N}_{s'} = \emptyset$ for rules,
Data	
$\boldsymbol{\delta}$	Matrix of dissimilarities δ_{ij} between each pair of individuals i and j,
b_{isn}	$= \begin{cases} 1, & \text{if individual } i \text{ is explained by rule } n \in \mathcal{N}_s \\ 0, & \text{otherwise} \end{cases}$,
Decision variables	
x_{ki}	$= \begin{cases} 1, & \text{if individual } i \text{ belongs to cluster } k \\ 0, & \text{otherwise} \end{cases}$,
z_{ksn}	$= \begin{cases} 1, & \text{if rule } n \in \mathcal{N}_s \text{ is chosen for cluster } k \\ 0, & \text{otherwise} \end{cases}$,
α_i	$= \begin{cases} 1, & \text{if individual } i \text{ is a true positive case to the explanation assigned to its cluster} \\ 0, & \text{otherwise} \end{cases}$,
β_{ki}	$= \begin{cases} 1, & \text{if individual } i \text{ is outside cluster } k \text{ and is a false positive case to the explanation assigned to cluster } k \\ 0, & \text{otherwise} \end{cases}$,
Parameters	
$\theta_1 \geq 0$	Weight for true positive cases across the K clusters,
$\theta_2 \geq 0$	Weight for false positive cases across the K clusters,
ℓ	Maximum length of the clusters' explanations.

In the following, we provide a mathematical optimization formulation to cluster the individuals in $\mathcal{I}$ using the dissimilarity $\boldsymbol{\delta}$ while selecting for each cluster a rule-based explanation of maximum length ℓ combining the rules of $\mathcal{N}_s$, $s = 1,\dots,S$:

$$\min_{\mathbf{x},\mathbf{z},\boldsymbol{\alpha},\boldsymbol{\beta}} \quad \sum_{k=1}^{\mathrm{K}} \sum_{i=1}^{\mathrm{I}-1} \sum_{j=i+1}^{\mathrm{I}} \delta_{ij} x_{ki} x_{kj} - \theta_1 \sum_{i=1}^{\mathrm{I}} \alpha_i + \theta_2 \sum_{k=1}^{\mathrm{K}} \sum_{i=1}^{\mathrm{I}} \beta_{ki} \tag{1}$$

$$\text{s.t.} \quad \sum_{k=1}^{\mathrm{K}} x_{ki} = 1, \qquad i = 1\dots\mathrm{I} \tag{2}$$

$$\sum_{n\in\mathcal{N}_s} z_{ksn} \leq 1, \qquad k = 1\dots\mathrm{K},\ s = 1\dots\mathrm{S} \tag{3}$$

$$1 \leq \sum_{s=1}^{\mathrm{S}} \sum_{n\in\mathcal{N}_s} z_{ksn} \leq \ell, \qquad k = 1\dots\mathrm{K} \tag{4}$$

$$\alpha_i + x_{ki} + \sum_{n\in\mathcal{N}_s} (1 - b_{isn}) z_{ksn} \leq 2, \qquad i = 1\dots\mathrm{I},\ k = 1\dots\mathrm{K},\ s = 1\dots\mathrm{S} \tag{5}$$

$$\beta_{ki} + x_{ki} + \sum_{s=1}^{\mathrm{S}} \sum_{n\in\mathcal{N}_s} (1 - b_{isn}) z_{ksn} \geq 1, \qquad i = 1\dots\mathrm{I},\ k = 1\dots\mathrm{K} \tag{6}$$

$$x_{ki} \in \{0,1\}, \qquad i = 1\dots\mathrm{I},\ k = 1\dots\mathrm{K} \tag{7}$$

$$z_{ksn} \in \{0,1\}, \qquad s = 1\dots\mathrm{S},\ n \in \mathcal{N}_s,\ k = 1\dots\mathrm{K} \tag{8}$$

$$\alpha_i \in [0,1], \qquad i = 1\dots\mathrm{I} \tag{9}$$

$$\beta_{ki} \in [0,1], \qquad i = 1\dots\mathrm{I},\ k = 1\dots\mathrm{K}. \tag{10}$$

The objective function (1) consists of three terms: the minimization of intra-homogeneity of clusters, the maximization of the total true

Table 1
Description of the datasets used to illustrate the quality of the explanations provided by (CinterP) and (InterP).

Name of dataset	#Individuals (I)	#Classes (C)	#Features (d)
housing	506	2	13
breast cancer	683	2	10
PIMA	768	2	8
abalone	835	2	8
wine	178	3	13
glass	214	6	9

Table 2
Description of the features in the housing dataset and the $C = 2$ classes.

Feature	Description
CRIM	Per capita crime rate by town
ZN	Proportion of residential land zoned for lots over 25,000 sq.ft.
INDUS	Proportion of non-retail business acres per town
CHAS	Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
NOX	Nitric oxides concentration (parts per 10 million)
RM	Average number of rooms per dwelling
AGE	Proportion of owner-occupied units built prior to 1940
DIS	Weighted distances to five Boston employment centres
RAD	Index of accessibility to radial highways
TAX	Full-value property-tax rate per $10,000
PTRATIO	Pupil–teacher ratio by town
B	$1000(\mathrm{Bk}-0.63)^2$ where Bk is the proportion of blacks by town
LSTAT	% lower status of the population
Class	Higher (class 1) or lower (class 2) than the median value of owner-occupied homes in $1000's

positive cases weighted by the parameter θ_1, and minimization of the total false positive cases by weighted by the parameter θ_2. The intra-homogeneity can take different forms (Rao, 1971; Basak and Krishnapuram, 2005), and we have considered here the sum of the dissimilarities within each cluster. We now discuss the constraints, and note that the correctness of the formulation is driven by the direction of the optimization, as we will see below. Constraints (2) ensure that each individual is assigned to exactly one cluster. For each cluster, constraints (3) ensure that at most one rule of group s is chosen, while constraints (4) impose that at least one rule is chosen for each cluster but no more than ℓ. Constraints (5) and (6) ensure that α_i and β_{ki} are well-defined. Because of the direction of the objective function, we only need to ensure that $\alpha_i = 0$ and $\beta_{ki} = 1$ are well-defined. Let us start with $\alpha_i = 0$ and note that $\sum_{n\in\mathcal{N}_s}(1-\mathcal{b}_{isn})z_{ksn} \le 1$. Thanks to this inequality, constraints (5) are redundant if individual i does not belong to cluster k, $x_{ki} = 0$. If individual i belongs to cluster k, $x_{ki} = 1$, and it is not explained by the explanation assigned to this cluster, then for each $s, n \in \mathcal{N}_s$ such that $z_{ksn} = 1$, we have that $\mathcal{b}_{isn} = 0$. This means that $\sum_{n\in\mathcal{N}_s}(1-\mathcal{b}_{isn})z_{ksn} = 0$, yielding $\alpha_i \le 0$. This, together with the fact that α_i cannot be negative, ensures that $\alpha_i = 0$. We now analyze the case of $\beta_{ik} = 1$. If individual i does not belong to cluster k, $x_{ki} = 0$, but satisfies the chosen explanation for that cluster, then $\forall s, n \in \mathcal{N}_s$ such that $z_{ksn} = 1$ we have $\mathcal{b}_{isn} = 1$. With this $\sum_{s=1}^{S}\sum_{n\in\mathcal{N}_s}(1-\mathcal{b}_{isn})z_{ksn} = 0$, and thus $\beta_{ki} \ge 1$, which together with the upper bound on β_{ki}, ensures that $\beta_{ki} = 1$. The integrality of the decision variables $\mathbf{x}$ and $\mathbf{z}$ is enforced by constraints (7) and (8). Decision variables α_i and β_{ki} were defined as integer variables, but as seen above we can assume them to be continuous without of loss of optimality, see constraints (9)–(10).

The intra-homogeneity term contains the product of binary decision variables $\mathbf{x}$. We linearize them by adding new decision variables $y_{kij} = x_{ki}x_{kj}$ and new constraints. With this the clustering and interpreting problem can be written as the following MILP formulation:

$$\min_{\mathbf{x},\mathbf{z},\boldsymbol{\alpha},\boldsymbol{\beta},\mathbf{y}} \quad \sum_{k=1}^{K}\sum_{i=1}^{I-1}\sum_{j=i+1}^{I}\delta_{ij}y_{kij} - \theta_1\sum_{i=1}^{I}\alpha_i + \theta_2\sum_{k=1}^{K}\sum_{i=1}^{I}\beta_{ki},$$

$$\text{s.t.} \quad (2)\text{–}(10)$$

$$x_{ki} + x_{kj} - y_{kij} \le 1, \qquad i = 1\ldots I-1,\ j = i+1\ldots I,\ k = 1\ldots K$$

$$y_{kij} \in [0,1], \qquad i = 1\ldots I-1,\ j = i+1\ldots I,\ k = 1\ldots K.$$

We will refer to this MILP formulation as (CinterP), which has $I + K(2 + S + SI + I + \frac{I(I-1)}{2})$ linear constraints, $(I + N)K$ binary decision variables, and $I(1 + K + \frac{K(I-1)}{2})$ continuous decision variables between 0 and 1.

The formulation (CinterP) can be enriched with desirable properties on the explanations associated with the clusters. In the pursue of distinctiveness, we discuss below two possibilities. For instance, one could impose that a feature (or one group of them) is used to explain at most one cluster. Alternatively, one could wish that a rule is associated with a cluster and that its complement is associated with another cluster. For instance, we could have (TAX > 398) associated with one cluster and (TAX ≤ 398) with another one. These constraints can be easily incorporated into (CinterP), while still being an MILP formulation.

3. Constructing explanations when clusters are given

Our proposed methodology can be used in a post-hoc step, where the goal is to explain the clusters that have been built previously with a Cluster Analysis approach, or that are simply available to the user in the form of cluster membership labels of the individuals. This means that we are given the set of individuals already split into K clusters, i.e., $\mathcal{I} = \cup_{k=1}^{K}\mathcal{I}_k$ with $\mathcal{I}_k \cap \mathcal{I}_{k'}$ with $k \ne k'$. In the following, we present the mathematical optimization formulation that selects rule-based explanations for the clusters, that are accurate and distinctive, of maximum length ℓ combining the rules of $\mathcal{N}_s$, $s = 1,\ldots,S$.

The decision variables z_{ksn} are defined as above, but we use slightly different decision variables to measure the quality of the explanations, i.e., the total number of true positive cases across all the clusters, as well as the false positive ones. Let γ_{ki} be a binary decision variable. Let us assume that i is in cluster k. The decision variable γ_{ki} is equal to 1 if individual i satisfies the explanation assigned to cluster k, and otherwise zero. For $k' \ne k$, $\gamma_{k'i}$ is equal to 1 if i satisfies the explanation chosen for cluster k' and 0 otherwise. The model for interpreting clusters $\mathcal{I}_k$, for $k = 1,\ldots,K$, reads as follows:

$$\min_{\mathbf{z},\boldsymbol{\gamma}} \quad -\sum_{k=1}^{K}\sum_{i\in\mathcal{I}_k}\gamma_{ki} + \theta\sum_{k=1}^{K}\sum_{\substack{k'=1\\k\ne k'}}^{K}\sum_{i\in\mathcal{I}_{k'}}\gamma_{ki} \tag{11}$$

$$\text{s.t.} \quad \sum_{n\in\mathcal{N}_s} z_{ksn} \le 1, \qquad k = 1\ldots K,\ s = 1\ldots S \tag{12}$$

$$1 \le \sum_{s=1}^{S}\sum_{n\in\mathcal{N}_s} z_{ksn} \le \ell, \qquad k = 1\ldots K \tag{13}$$

$$\gamma_{ki} + \sum_{n\in\mathcal{N}_s}(1-\mathcal{b}_{isn})z_{ksn} \le 1, \qquad i\in\mathcal{I}_k,\ k = 1\ldots K,\ s = 1\ldots S \tag{14}$$

$$\gamma_{ki} + \sum_{s=1}^{S}\sum_{n\in\mathcal{N}_s}(1-\mathcal{b}_{isn})z_{ksn} \ge 1, \qquad i\in\mathcal{I}_{k'},\ k,k' = 1\ldots K, k\ne k' \tag{15}$$

$$z_{ksn} \in \{0,1\}, \qquad s = 1\ldots S,\ n\in\mathcal{N}_s,\ k = 1\ldots K \tag{16}$$

$$\gamma_{ki} \in [0,1], \qquad i = 1\ldots I,\ k = 1\ldots K. \tag{17}$$

The objective function (11) maximizes total true positive cases and minimizes total false positive cases weighted by the parameter $\theta \ge 0$. Constraints (12)–(13) are exactly the same as constraints (3)–(4). Constraints (14)–(15) resemble constraints (5)–(6), but they are slightly different since the cluster membership is known, and ensure that γ_{ki} is well defined. The nature of decision variables is specified in constraints (16)–(17), where, as before, we can assume that γ_{ki} is a continuous

Table 3
The clusters and the explanations provided by (CinterP), $\theta_1 \in \{2^p\}_{p=-1,0,1}$ and $\theta_2 \in \{2^p\}_{p=-1,0,1}$, for the `housing` dataset, with K = 2 clusters, explanations of a maximum length of $\ell = 2$ constructed with N = 187 rules using the deciles of the continuous features and all attributes of the categorical features.

θ_1	θ_2	Intra-homogeneity	Cluster	TPR	FPR	Explanations
2^{-1}	2^{-1}	$0.6 \cdot 10^5$	1	1.00	0.04	TAX > 398 AND INDUS > 12.83
			2	0.97	0.00	NOX ≤ 0.605 AND RAD ≤ 8
2^{-1}	2^0	$0.6 \cdot 10^5$	1	0.90	0.00	INDUS > 12.83 AND PTRATIO > 19.7
			2	0.97	0.00	NOX ≤ 0.605 AND RAD ≤ 8
2^{-1}	2^1	$0.6 \cdot 10^5$	1	0.90	0.00	INDUS > 12.83 AND PTRATIO > 19.7
			2	0.97	0.00	NOX ≤ 0.605 AND RAD ≤ 8
2^0	2^{-1}	$0.6 \cdot 10^5$	1	1.00	0.04	TAX > 398 AND INDUS > 12.83
			2	1.00	0.09	TAX ≤ 437 AND NOX ≤ 0.668
2^0	2^0	$0.6 \cdot 10^5$	1	1.00	0.04	TAX > 398 AND INDUS > 12.83
			2	0.97	0.00	NOX ≤ 0.605 AND RAD ≤ 8
2^0	2^1	$0.6 \cdot 10^5$	1	0.90	0.00	INDUS > 12.83 AND PTRATIO > 19.7
			2	0.97	0.00	NOX ≤ 0.605 AND RAD ≤ 8
2^1	2^{-1}	$0.6 \cdot 10^5$	1	1.00	0.04	TAX > 398 AND INDUS > 12.83
			2	1.00	0.09	TAX ≤ 437 AND NOX ≤ 0.668
2^1	2^0	$0.6 \cdot 10^5$	1	1.00	0.04	TAX > 398 AND INDUS > 12.83
			2	1.00	0.09	TAX ≤ 437 AND NOX ≤ 0.668
2^1	2^1	$0.6 \cdot 10^5$	1	1.00	0.04	TAX > 398 AND INDUS > 12.83
			2	0.97	0.00	NOX ≤ 0.605 AND RAD ≤ 8

variable. Model (11)–(17), hereafter (InterP), is an MILP problem with $\mathrm{K}(\mathrm{S}+2)+\mathrm{I}(\mathrm{S}+1)$ constraints, K N integer decision variables and K I continuous decision variables between 0 and 1. Please note that (InterP) is separable yielding an MILP for each cluster. Nevertheless, when incorporating the two desirable properties on the explanations to enhance their distinctiveness, namely, a feature can be used by at most one cluster or the complementarity of the explanations of two clusters, the problem is not separable anymore.

The sizes of (CinterP) and (InterP) depend on the number of rules available to construct the explanations of the clusters, i.e., N. For continuous features, the number of rules can be controlled by choosing the level of granularity of the thresholds defining these rules. First, in the most granular case, one can use all possible thresholds corresponding to all distinct values of the features in the dataset. This may lead to a redundancy since many values may be very close to each other, and thus yielding the same accuracy and distinctiveness of the explanation. Second, in a less granular case, we could use as thresholds some percentiles of the features, say, the deciles. This dramatically reduces the number of rules we start with, but it also enhances the interpretation of the rule, by saying that this is the value of the feature that leaves 10% of the observations in the dataset above (respectively, below), if the ninth decile is chosen. These different sources of if-then rules will be tested in the numerical section. For (InterP), where the clusters are given, there is another alternative to generate the rules. They can be extracted from an additive tree model based on stumps, such as an XGBoost of depth 1, which uses the cluster labels as the class labels. In this way, we expect more granularity in some features than in others because they are more relevant to explain the clusters.

4. Numerical section

In this section, we illustrate our methodology on well-known real-world datasets from the UCI Repository (Dua and Graff, 2017). In Section 4.1, we present the benchmark datasets and the rules used to build the explanations. In Section 4.2, we focus on our novel clustering and interpreting model in which we perform these two tasks simultaneously, namely (CinterP). We discuss the intra-homogeneity of the clusters, the accuracy and the distinctiveness of our explanations. In Section 4.3, we focus on our post-hoc model in which the clusters are given and we aim to explain them, namely (InterP). We discuss the accuracy and the distinctiveness of our explanations and compare them to those obtained with CART. In Section 4.4, the impact of the source of the rules used to construct the explanations on (CinterP) and (InterP) is analyzed. To enhance the clarity of the presentation, some of the tables and figures have been placed in the Appendix.

For interpretability purposes, we limit the maximum length of explanations to $\ell = 2$ for both (CinterP) and (InterP). In (CinterP), we take as dissimilarity δ_{ij} the squared Euclidean distance between the (normalized) feature vectors of individuals i and j. To solve the optimization models we use *Gurobi* (Gurobi Optimization, 2020) with *Python* (Python Core Team, 2015) on a PC Intel®Core TM i7-8665U, 16 GB of RAM. For each instance of (CinterP), we impose a time limit of 10 min, which allows us to get solutions in which the clusters and explanations show a good tradeoff in the three criteria optimized, namely intra-homogeneity, accuracy and distinctiveness of the explanations. For (InterP), all the instances were solved in less than 10 s.

4.1. The datasets and the set of rules

The benchmark datasets are from Supervised Classification, with $C = 2, 3$ and 6 classes. We use these C classes as the clusters to be explained in the post-hoc approach (InterP), while our clustering and interpreting model (CinterP) ignores this information and constructs the C clusters and their corresponding explanations. The description of the datasets can be found in Tables 1, 2 and Tables A.7–A.11. Table 1 contains information on the name of the dataset, the number of individuals, the number of classes and the number of features used to construct the rules, while Tables 2 and A.7–A.11 contains a brief description of each of these features and the classes.

We make two observations on these datasets. First, all features are continuous except for the `housing` dataset that has one binary feature and `abalone` that has one categorical variable with three categories, for which we have constructed a binary feature for each category. Second, the dataset `abalone` has been obtained by drawing a random sample from the original dataset, which has more than 4000 observations.

The rules we consider in Sections 4.2 and 4.3 are of the following form. We have a group of rules for each feature, i.e., S = d. If feature s is continuous, we consider the rules: $feature_s \leq threshold$, $feature_s > threshold$, where $threshold$ takes on the deciles of $feature_s$. For binary features, the two rules are defined as $feature_s = 1$, $feature_s = 0$. This choice of rules is further analyzed in Section 4.4.

Table 4
The clusters and the explanations provided by (InterP), $\theta \in \{2^p\}_{p=-5,\dots,5}$, for the `housing` dataset, with K = 2 clusters, explanations of a maximum length of ℓ = 2 constructed with N = 187 rules using the deciles of the continuous features and all attributes of the categorical features.

θ	Cluster	TPR	FPR	Explanations
2^5	1	0.45	0.00	RM > 6.376 AND LSTAT ≤ 7.765
	2	0.14	0.00	PTRATIO > 20.9 AND LSTAT > 11.36
2^4	1	0.59	0.01	RM > 6.2085 AND LSTAT ≤ 9.53
	2	0.14	0.00	PTRATIO > 20.9 AND LSTAT > 11.36
2^3	1	0.59	0.01	RM > 6.2085 AND LSTAT ≤ 9.53
	2	0.14	0.00	PTRATIO > 20.9 AND LSTAT > 11.36
2^2	1	0.59	0.01	RM > 6.2085 AND LSTAT ≤ 9.53
	2	0.41	0.05	CRIM ≤ 10.753 AND LSTAT > 15.62
2^1	1	0.70	0.06	RM > 6.086 AND LSTAT ≤ 11.36
	2	0.70	0.15	CRIM ≤ 10.753 AND LSTAT > 11.36
2^0	1	0.70	0.06	RM > 6.086 AND LSTAT ≤ 11.36
	2	0.81	0.23	AGE > 26.95 AND LSTAT > 11.36
2^{-1}	1	0.78	0.18	RM > 5.9505 AND LSTAT ≤ 13.33
	2	0.97	0.40	RM ≤ 6.75 AND LSTAT > 7.765
2^{-2}	1	0.98	0.83	PTRATIO ≤ 20.9
	2	0.99	0.46	LSTAT > 7.765
2^{-3}	1	0.98	0.83	PTRATIO ≤ 20.9
	2	0.99	0.46	LSTAT > 7.765
2^{-4}	1	1.00	1.00	All in
	2	0.99	0.46	LSTAT > 7.765
2^{-5}	1	1.00	1.00	All in
	2	1.00	0.63	LSTAT > 6.29
CART	1	0.75	0.12	LSTAT ≤ 9.95 AND RM > 6.12 OR LSTAT > 9.95 AND TAX ≤ 302
	2	0.88	0.25	LSTAT ≤ 9.95 AND RM ≤ 6.12 OR LSTAT > 9.95 AND TAX > 302

4.2. Illustrating the clustering and interpreting model (CinterP)

The results of (CinterP) can be found in Tables 3 and B.12–B.16, where a table is devoted to each benchmark dataset. For each dataset, the corresponding table shows the value of the three objectives in (CinterP) and the explanations obtained for each cluster. For the first objective, we report the total intra-homogeneity, while for the other two objectives, namely the accuracy and the distinctiveness, we report those in relative terms, i.e., the true and false positive rates for each cluster.

Model (CinterP) has two parameters, θ_1 and θ_2, which are weights of the accuracy and the distinctiveness of the explanations, respectively. To have both objectives in roughly the same scale, we divide the intra-homogeneity by the constant $\mathrm{I}^2 \max_{ij} \delta_{ij}^2$, while the other two objectives are divided by I. Once this is done, we consider a grid of parameters, namely, $(\theta_1, \theta_2) \in \{2^p\}_{p=-1,0,1} \times \{2^p\}_{p=-1,0,1}$. We first solve (CinterP) for the smallest value of θ_1 and each value of θ_2, the latter taken in increasing order. We continue in a similar fashion with the values of θ_1 taken in increasing order. For each problem, we start with an initial solution: clusters and explanations. We consider two options and give to the solver the one with the best objective function. Initial clusters can be constructed using K-means clustering or can be simply the ones obtained when solving (CinterP) with the previous combination of θ_1 and θ_2 in our grid. We use these clusters in (InterP) to obtain the corresponding initial explanations, with $\theta = \theta_2/\theta_1$.

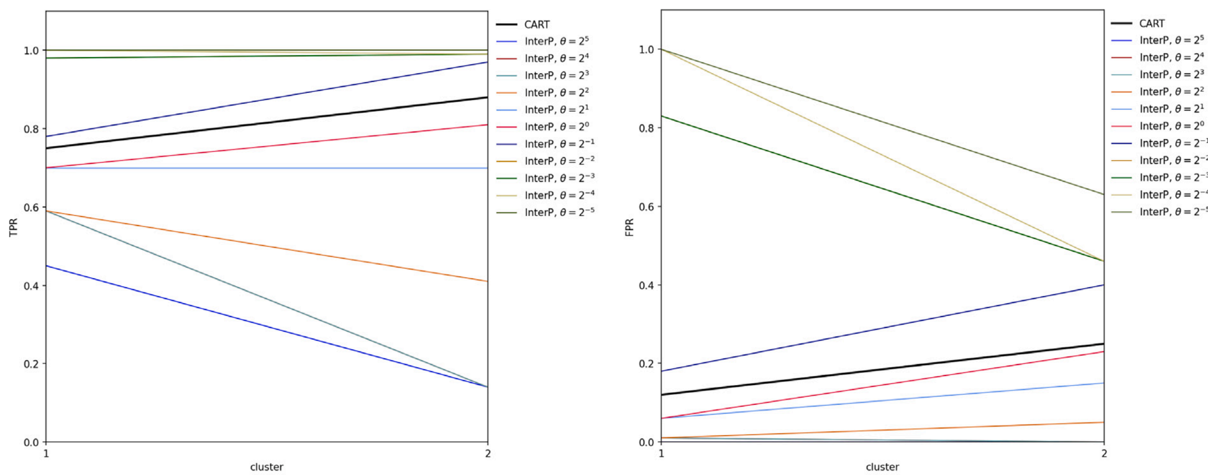
Let us start discussing the results for the `housing` dataset found in Table 3. The intra-homogeneity stays the same for all the combinations of the parameters in the grid, namely, $0.6 \cdot 10^5$. After inspecting the clusters, we note that those are the ones from the initial solution, namely the K-means solution. As we will see below, when we enlarge the number of rules, problem (CinterP) will yield different partitions. The explanations obtained for these clusters are very good in terms of the accuracy and distinctiveness of the explanations. Indeed, the true positive rate of the first cluster ranges from 90% to 100% and the false positive rate from 0% to 4%, while for the second cluster, the true positive rate ranges from 97% to 100% and the false positive rate from 0% to 9%. As we will see below, (CinterP) will slightly improve these metrics when we enlarge the number of rules.

Similar conclusions can be drawn for the other datasets. For `breast cancer`, for the best value of the intra-homogeneity, the explanations have a true positive rate of 97% and 90%, respectively, and a false positive rate of 2% in both clusters. For `PIMA`, for the second best value of the intra-homogeneity, the explanations have a true positive rate of 80% and 100%, respectively, and the false positive rate is perfect, i.e., 0% in both clusters. For `abalone`, for the second best value of the intra-homogeneity, the explanations have a true positive rate of 82% and 100%, respectively, and a false positive rate of 16% and 0%, respectively. For `wine`, we obtain perfect explanations for all three clusters. To end, for `glass`, for the best value of the intra-homogeneity, the explanations have a true positive rate of 80%, 100%, 95%, 100%, 100% and 50%, respectively, and a false positive rate of 3%, 0%, 9%, 1%, 2% and 0%, respectively.

To end, we note that we have not been able to obtain a proof of optimality for the solutions above within the time limit of 10 min. Indeed, for `housing`, the MIPGAP ranges from 3.05% to 11.77%, for `breast cancer` from 1.60% to 9.76%, for `PIMA` from 3.65% to 25.76%, for `abalone` from 8.93% to 62.30%, for `wine` from 1.89% to 10.06%, for `glass` from 8.84% to 41.42%. This is not surprising since it is known that clustering is already a difficult problem, and (CinterP) here needs to cluster approximately hundreds of individuals, and, in addition, explain the clusters, all within the same mathematical optimization model.

4.3. Illustrating the interpreting model (InterP)

To illustrate (InterP) and its natural benchmark, namely CART, we assume that the clusters are given by classes reported in Tables 2 and A.7–A.11. To make the comparison fair, we train a CART of depth 2



(a) True Positive Rate   (b) False Positive Rate

Fig. 2. The `housing` data:the interpretability results obtained by (InterP).

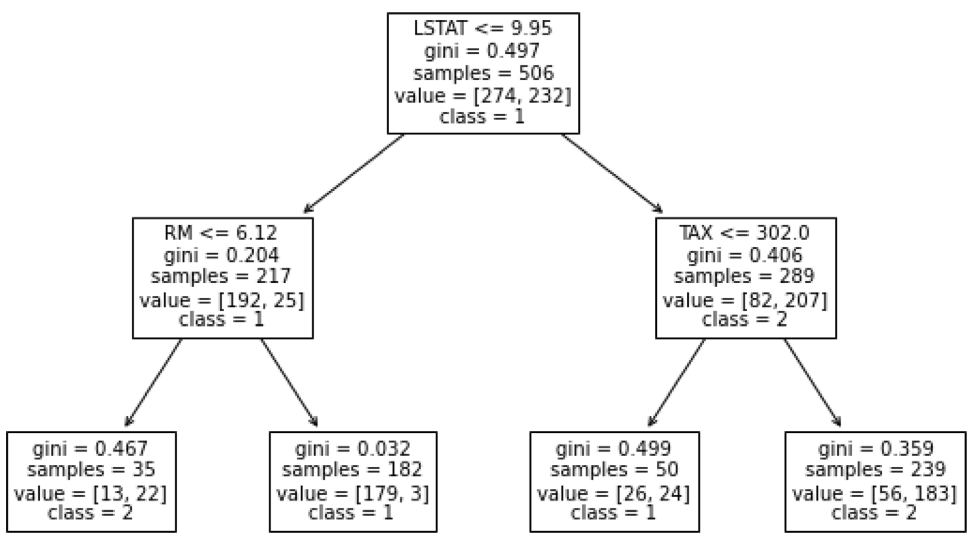


Fig. 3. The post-hoc explanations provided by a CART of depth 2 for the `housing` dataset for clusters (classes) 1 and 2.

for these benchmark datasets with $C = 2$ classes, while for `wine` and `glass`, the chosen depth is 2 and 4, which is the minimum one to ensure that all classes are represented in the leaf nodes.

The explanations provided by (InterP) and CART for these clusters, as well as the accuracy and distinctiveness can be found in Tables 4 and Tables C.17–C.22. These two criteria are depicted in Figs. 2 and C.4–C.8 for both methodologies. The CART trees can be found in Figs. 3 and C.9–C.13.

For the only parameter in (InterP), namely θ, we consider the grid of values $\theta \in \{2^p\}_{p=-5,\ldots,5}$. We solve the problem instances of (InterP) in increasing order of θ. For each value of the parameter, we give to the solver as the initial solution the one obtained with the previous value of θ.

We focus on the `housing` dataset, as the results for the rest datasets are similar. From Table 4 and Fig. 2, we can see that the true positive rate of the first cluster ranges from 45% to 100% and the false positive rate from 0% to 100%. For the second cluster, the true positive rate ranges from 14% to 100% and the false positive rate 0% to 63%. The low (respectively the high) values of the grid are not very interesting, since they correspond to extreme solutions with a very low true positive rate (respectively very high false positive rate). Indeed, they provide explanations that are hardly satisfied by any member of the cluster (respectively explanations that are satisfied by all clusters marked as "all in"). Therefore, we focus on the central values of the chosen grid. There, we find a good tradeoff between the accuracy and the distinctiveness for both clusters. Indeed, we see that for cluster 1 the explanation (RM > 6.086) AND (LSTAT $\leq$ 11.36) has a true positive rate of 70% and a false positive rate of 6%, while for cluster 2 (AGE > 26.95) AND (LSTAT > 11.36) has a true positive rate of 81% and a false positive rate of 23%. This is a similar performance to that of

Table 5
The clusters and the explanations provided by (CinterP), $\theta_1 \in \{2^p\}_{p=-1,0,1}$ and $\theta_2 \in \{2^p\}_{p=-1,0,1}$, for the `housing` dataset, with K = 2 clusters, explanations of a maximum length of $\ell = 2$ constructed with N = 5646 rules using the unique values of the continuous features and all attributes of the categorical features.

θ_1	θ_2	Intra-homogeneity	Cluster	TPR	FPR	Explanations
2^{-1}	2^{-1}	$6.03 \cdot 10^4$	1	1.00	0.04	INDUS > 15.04 AND RAD > 3
			2	1.00	0.00	TAX ≤ 432 AND NOX ≤ 0.647
2^{-1}	2^0	$6.04 \cdot 10^4$	1	0.91	0.00	TAX > 432
			2	1.00	0.00	TAX ≤ 432 AND NOX ≤ 0.647
2^{-1}	2^1	$6.04 \cdot 10^4$	1	0.91	0.00	TAX > 432
			2	1.00	0.00	TAX ≤ 432 AND NOX ≤ 0.647
2^0	2^{-1}	$6.03 \cdot 10^4$	1	1.00	0.04	TAX > 402 AND INDUS > 15.04
			2	1.00	0.00	TAX ≤ 432 AND NOX ≤ 0.647
2^0	2^0	$6.03 \cdot 10^4$	1	1.00	0.04	TAX > 402 AND INDUS > 15.04
			2	1.00	0.00	TAX ≤ 432 AND NOX ≤ 0.647
2^0	2^1	$6.04 \cdot 10^4$	1	0.91	0.00	TAX > 432
			2	1.00	0.00	TAX ≤ 432 AND NOX ≤ 0.647
2^1	2^{-1}	$6.03 \cdot 10^4$	1	1.00	0.04	INDUS > 15.04 AND RAD > 3
			2	1.00	0.00	TAX ≤ 432 AND NOX ≤ 0.647
2^1	2^0	$6.03 \cdot 10^4$	1	1.00	0.04	TAX > 402 AND INDUS > 15.04
			2	1.00	0.00	TAX ≤ 432 AND NOX ≤ 0.647
2^1	2^1	$6.03 \cdot 10^4$	1	1.00	0.04	INDUS > 15.04 AND RAD > 3
			2	1.00	0.00	TAX ≤ 432 AND NOX ≤ 0.647

CART, with more complex explanations, namely ((LSTAT ≤ 9.95) AND (RM > 6.12)) OR ((LSTAT > 9.95) AND (TAX ≤ 302)) for cluster 1, with a true positive rate of 75% and false positive rate of 12%, and ((LSTAT ≤ 9.95) AND (RM ≤ 6.12)) OR ((LSTAT > 9.95) AND (TAX > 302)) for cluster 2, with a true positive rate of 88% and false positive rate of 25%. These explanations, linking rules by an OR operator, seem to imply that the given clusters are not the natural clusters, since no conjunctive explanation is found out to explain the whole cluster. This unpleasant fact observed in CARTs is, by construction, impossible in our approach. In addition, our explanations above use as thresholds the deciles, as opposed to CART that may use any possible value of the features in the dataset. This lower granularity we have chosen may affect the two metrics measuring the quality of the explanations, i.e., it may lower the accuracy and/or the distinctiveness, but it will enhance the interpretability of these thresholds.

4.4. Source of rules

In this section we present the results of (CinterP) and (InterP) with alternative sources of explanations for the `housing` dataset. We would like to understand the impact of increasing the granularity of the rules used to construct the explanations. We test (CinterP) and (InterP) when all distinct values of the features in the dataset are considered as thresholds. This increases the total number of rules from N = 187 to N = 5646.

With the increase of granularity, (CinterP) now improves the true positive rate of the first cluster, yielding explanations that are almost perfect for a 4% false positive rate of the second cluster, see Table 5. For (InterP), small improvements are also reported for the most granular option, see Table 6.

Table 6
The clusters and the explanations provided by (InterP), $\theta \in \{2^p\}_{p=-5,\dots,5}$, for the `housing` dataset, with K = 2 clusters, explanations of a maximum length of $\ell = 2$ constructed with N = 5646 rules using the unique values of the continuous features and all attributes of the categorical features.

θ	Cluster	TPR	FPR	Explanations
2^5	1	0.51	0.00	RM > 6.31 AND LSTAT ≤ 8.61
	2	0.14	0.00	LSTAT > 11.25 AND PTRATIO > 20.9
2^4	1	0.58	0.00	Al ≤ 1.146 AND Si ≤ 72.132
	2	0.14	0.00	Mg ≤ 2.805 AND Ca > 10.443
2^3	1	0.64	0.01	RM > 6.144 AND LSTAT ≤ 9.93
	2	0.14	0.00	LSTAT > 11.25 AND PTRATIO > 20.9
2^2	1	0.64	0.01	RM > 6.144 AND LSTAT ≤ 9.93
	2	0.45	0.05	LSTAT > 14.81 AND CRIM ≤ 10.6718
2^1	1	0.70	0.04	RM > 6.12 AND LSTAT ≤ 11.66
	2	0.70	0.14	LSTAT > 11.66 AND CRIM ≤ 11.1604
2^0	1	0.73	0.06	RM > 6.059 AND LSTAT ≤ 11.66
	2	0.80	0.20	LSTAT > 11.66 AND CRIM ≤ 37.6619
2^{-1}	1	0.78	0.19	LSTAT ≤ 11.66 AND B > 172.91
	2	0.99	0.44	LSTAT > 7.67 AND PTRATIO > 14.4
2^{-2}	1	0.98	0.80	PTRATIO ≤ 20.9 AND B > 6.68
	2	0.99	0.44	LSTAT > 7.67 AND PTRATIO > 14.4
2^{-3}	1	1.00	0.90	PTRATIO ≤ 21 AND B > 6.68
	2	0.99	0.44	LSTAT > 7.67 AND PTRATIO > 14.4
2^{-4}	1	1.00	0.97	PTRATIO ≤ 21.2 AND B > 6.68
	2	0.99	0.44	LSTAT > 7.67 AND PTRATIO > 14.4
2^{-5}	1	1.00	0.97	PTRATIO ≤ 21.2 AND B > 6.68
	2	1.00	0.53	LSTAT > 6.73 AND PTRATIO > 14.4

5. Conclusions

In this paper, we have introduced an MILP model to simultaneously cluster individuals and provide rule-based explanations for the clusters. We have assumed that we have at hand a dissimilarity between the individuals. We have also assumed that we have rules based on features characterizing the individuals, which are to be combined with the AND operator to obtain explanations for the clusters. We have measured the quality of the clustering by minimizing the total dissimilarity between individuals in the same cluster, while the goodness of the explanations has been pursued by maximizing the number of true positive cases across all clusters and minimizing the number of false positive cases. Our approach can be applied in a post-hoc fashion to interpret the clusters of any Cluster Analysis approach or the clusters available to the user in the form of cluster membership labels.

To end, it would be interesting to sharpen the corresponding mathematical optimization formulation for (CinterP), as well as to model alternative forms of intra-homogeneity of the clusters. Another line of future research that is worth considering is the modeling of fairness constraints (Abraham et al., 2020).

Table A.7
Description of the features in the `breast cancer` dataset and the $C = 2$ classes.

Feature	Description
Thickness	Clump Thickness
Size	Uniformity of Cell Size
Shape	Uniformity of Cell Shape
Adhesion	Marginal Adhesion
Epithelial Size	Single Epithelial Cell Size
Nuclei	Bare Nuclei
Nuclei	Bland Chromatin
Normal Nucleoli	Normal Nucleoli
Mitoses	Mitoses
Class	Benign (class 1) or malignant (class 2)

Table A.8
Description of the features in the `PIMA` dataset and the $C = 2$ classes.

Feature	Description
Pregnancies	Number of times pregnant
Glucose	Plasma glucose concentration a 2 h in an oral glucose tolerance test
BloodPressure	Diastolic blood pressure (mm Hg)
SkinThickness	Triceps skin fold thickness (mm)
Insulin	2-Hour serum insulin (mu U/ml)
BMI	Body mass index (weight in kg/(height in m)^2)
DiabetesPedigree	Diabetes pedigree function
Age	Age (years)
Class	Diabetes (class 2) or not (class 1)

Table A.9
Description of the features in the `abalone` dataset and the $C = 2$ classes.

Feature	Description
Sex	Sex
Length	Length
Diameter	Diameter
Height	Height
Whole weight	Whole weight
Shucked weight	Shucked weight
Viscera weight	Viscera weight
Shell weight	Shell weight
Class	Higher (class 2) or lower (class 1) than the median value of the number of the rings

CRediT authorship contribution statement

Emilio Carrizosa: Conceptualization, Methodology, Writing – original draft, Writing – review & editing, Supervision. **Kseniia Kurishchenko:** Conceptualization, Methodology, Writing – original draft, Writing – review & editing. **Alfredo Marín:** Conceptualization, Methodology, Writing – original draft, Writing – review & editing, Supervision. **Dolores Romero Morales:** Conceptualization, Methodology, Writing – original draft, Writing – review & editing, Supervision.

Data availability

The data is available on the UCI repository

Acknowledgments

This research has been financed in part by research projects EC H2020 MSCA RISE NeEDS (Grant agreement ID: 822214); FQM-329, P18-FR-2369 and US-1381178 (Junta de Andalucía, Spain); PID2019-110886RB-I00 (Ministerio de Ciencia e Innovación, Spain). This support is gratefully acknowledged. Part of this research was conducted while the third author, A. Marín, was on sabbatical at Instituto de Matemáticas de la Universidad de Sevilla, Seville, Spain.

Appendix A. Description of the features and classes in the datasets

See Tables A.7–A.11.

Table A.10
Description of the features in the `wine` dataset and the $C = 3$ classes.

Feature	Description
Alcohol	Alcohol
Malic acid	Malic acid
Ash	Ash
Alcalinity of ash	Alcalinity of ash
Magnesium	Magnesium
Total phenols	Total phenols
Flavanoids	Flavanoids
Nonflavanoid phenols	Nonflavanoid phenols
Proanthocyanins	Proanthocyanins
Color intensity	Color intensity
Hue	Hue
OD280andOD31ofdilutedwines	OD280/OD315 of diluted wines
Proline	Proline
Class	Type of wine ($C = 3$)

Table A.11
Description of the features in the `glass` dataset and the $C = 6$ classes.

Feature	Description
RI	Refractive index
Na	Sodium
Mg	Magnesium
Al	Aluminum
Si	Silicon
K	Potassium
Ca	Calcium
Ba	Barium
Fe	Iron
Class	Type of glass ($C = 6$)

Table B.12
The clusters and the explanations provided by (CinterP), $\theta_1 \in \{2^p\}_{p=-1,0,1}$ and $\theta_2 \in \{2^p\}_{p=-1,0,1}$, for the `breast cancer` dataset, with K = 2 clusters, explanations of a maximum length of ℓ = 2 constructed with N = 83 rules using the deciles of the continuous features and all attributes of the categorical features.

θ_1	θ_2	Intra-homogeneity	Cluster	TPR	FPR	Explanations
2^{-1}	2^{-1}	$1.73 \cdot 10^5$	1	1.00	0.00	Thickness $\leq$ 3
			2	1.00	0.00	Thickness > 3
2^{-1}	2^0	$0.67 \cdot 10^5$	1	0.97	0.02	Size $\leq$ 4 AND Nuclei $\leq$ 4
			2	0.90	0.02	Size > 2 AND Nuclei > 2
2^{-1}	2^1	$1.1 \cdot 10^5$	1	0.97	0.00	Nuclei $\leq$ 4
			2	1.00	0.00	Size > 2 AND Nuclei > 4
2^0	2^{-1}	$1.24 \cdot 10^5$	1	1.00	0.00	Shape $\leq$ 1
			2	1.00	0.00	Shape > 1
2^0	2^0	$1.24 \cdot 10^5$	1	1.00	0.00	Shape $\leq$ 1
			2	1.00	0.00	Shape > 1
2^0	2^1	$1.24 \cdot 10^5$	1	1.00	0.00	Shape $\leq$ 1
			2	1.00	0.00	Shape > 1
2^1	2^{-1}	$1.24 \cdot 10^5$	1	1.00	0.00	Shape $\leq$ 1
			2	1.00	0.00	Shape > 1
2^1	2^0	$1.24 \cdot 10^5$	1	1.00	0.00	Shape $\leq$ 1
			2	1.00	0.00	Shape > 1
2^1	2^1	$1.24 \cdot 10^5$	1	1.00	0.00	Shape $\leq$ 1
			2	1.00	0.00	Shape > 1

Appendix B. The results for (CinterP) from Section 4.2

See Tables B.12–B.16.

Appendix C. The results for (InterP) and CART from Section 4.3

See Tables C.17–C.22 and Figs. C.4–C.13.

Table B.13
The clusters and the explanations provided by (CinterP), $\theta_1 \in \{2^p\}_{p=-1,0,1}$ and $\theta_2 \in \{2^p\}_{p=-1,0,1}$, for the PIMA dataset, with K = 2 clusters, explanations of a maximum length of $\ell = 2$ constructed with N = 135 rules using the deciles of the continuous features and all attributes of the categorical features.

θ_1	θ_2	Intra-homogeneity	Cluster	TPR	FPR	Explanations
2^{-1}	2^{-1}	$0.48 \cdot 10^5$	1 2	0.75 1.00	0.03 0.05	Pregnancies > 3 AND Age > 33 Pregnancies ≤ 5 AND Age ≤ 42.6
2^{-1}	2^0	$0.48 \cdot 10^5$	1 2	0.72 1.00	0.01 0.04	Pregnancies > 4 AND Age > 33 Pregnancies ≤ 5 AND Age ≤ 42.6
2^{-1}	2^1	$0.57 \cdot 10^5$	1 2	0.80 1.00	0.00 0.00	BMI > 33.7 BMI ≤ 32
2^0	2^{-1}	$1.22 \cdot 10^5$	1 2	1.00 –	0.00 –	All in –
2^0	2^0	$1.22 \cdot 10^5$	1 2	1.00 –	0.00 –	All in –
2^0	2^1	$1.22 \cdot 10^5$	1 2	1.00 –	0.00 –	All in –
2^1	2^{-1}	$1.22 \cdot 10^5$	1 2	1.00 –	0.00 –	All in –
2^1	2^0	$1.22 \cdot 10^5$	1 2	1.00 –	0.00 –	All in –
2^1	2^1	$1.22 \cdot 10^5$	1 2	1.00 –	0.00 –	All in –

Table B.14
The clusters and the explanations provided by (CinterP), $\theta_1 \in \{2^p\}_{p=-1,0,1}$ and $\theta_2 \in \{2^p\}_{p=-1,0,1}$, for the abalone dataset, with K = 2 clusters, explanations of a maximum length of $\ell = 2$ constructed with N = 130 rules using the deciles of the continuous features and all attributes of the categorical features.

θ_1	θ_2	Intra-homogeneity	Cluster	TPR	FPR	Explanations
2^{-1}	2^{-1}	$2.17 \cdot 10^5$	1 2	0.82 1.00	0.16 0.00	Length > 0.415 AND Viscera weight > 0.1435 Sex = I
2^{-1}	2^0	$2.17 \cdot 10^5$	1 2	0.82 1.00	0.16 0.00	Length > 0.415 AND Viscera weight > 0.1435 Sex = I
2^{-1}	2^1	$2.16 \cdot 10^5$	1 2	0.56 0.93	0.00 0.00	Sex = M Sex = I
2^0	2^{-1}	$2.52 \cdot 10^5$	1 2	0.95 1.00	0.40 0.00	Whole weight > 0.3625 AND Shell weight > 0.103 Sex = I AND Length ≤ 0.54
2^0	2^0	$2.52 \cdot 10^5$	1 2	0.90 1.00	0.22 0.00	Length > 0.415 AND Viscera weight > 0.10775 Sex = I AND Length ≤ 0.54
2^0	2^1	$2.52 \cdot 10^5$	1 2	0.82 1.00	0.07 0.00	Length > 0.415 AND Viscera weight > 0.1435 Sex = I AND Length ≤ 0.54
2^1	2^{-1}	$2.52 \cdot 10^5$	1 2	0.98 1.00	0.65 0.00	Whole weight > 0.1955 AND Viscera weight > 0.04 Sex = I AND Length ≤ 0.54
2^1	2^0	$2.52 \cdot 10^5$	1 2	0.95 1.00	0.40 0.00	Whole weight > 0.3625 AND Shell weight > 0.103 Sex = I AND Length ≤ 0.54
2^1	2^1	$2.52 \cdot 10^5$	1 2	0.90 1.00	0.22 0.00	Length > 0.415 AND Viscera weight > 0.10775 Sex = I AND Length ≤ 0.54

Table B.15
The clusters and the explanations provided by (CinterP), $\theta_1 \in \{2^p\}_{p=-1,0,1}$ and $\theta_2 \in \{2^p\}_{p=-1,0,1}$, for the wine dataset, with K = 3 clusters, explanations of a maximum length of $\ell = 2$ constructed with N = 235 rules using the deciles of the continuous features and all attributes of the categorical features.

θ_1	θ_2	Intra-homogeneity	Cluster	TPR	FPR	Explanations
2^{-1}	2^{-1}	$4.99 \cdot 10^3$	1 2 3	1.00 1.00 1.00	0.00 0.00 0.00	Ash > 2.3 AND Totalphenols > 1.881 Ash ≤ 2.3 AND Totalphenols > 1.881 Totalphenols ≤ 1.881
2^{-1}	2^0	$5.22 \cdot 10^3$	1 2 3	1.00 1.00 1.00	0.00 0.00 0.00	Ash ≤ 2.61 AND Totalphenols > 2.05 Ash ≤ 2.61 AND Totalphenols ≤ 2.05 Ash > 2.61
2^{-1}	2^1	$6.15 \cdot 10^3$	1 2 3	1.00 1.00 1.00	0.00 0.00 0.00	Malicacid > 1.247 AND Proline ≤ 742 Malicacid ≤ 1.247 Malicacid > 1.247 AND Proline > 742
2^0	2^{-1}	$4.99 \cdot 10^3$	1 2 3	1.00 1.00 1.00	0.00 0.00 0.00	Ash > 2.3 AND Totalphenols > 1.881 Ash ≤ 2.3 AND Totalphenols > 1.881 Totalphenols ≤ 1.881

(continued on next page)

Table B.15 (*continued*).

θ_1	θ_2	Intra-homogeneity	Cluster	TPR	FPR	Explanations
2^0	2^0	$4.99 \cdot 10^3$	1	1.00	0.00	Ash > 2.3 AND Totalphenols > 1.881
			2	1.00	0.00	Ash ≤ 2.3 AND Totalphenols > 1.881
			3	1.00	0.00	Totalphenols ≤ 1.881
2^0	2^1	$4.99 \cdot 10^3$	1	1.00	0.00	Ash > 2.3 AND Totalphenols > 1.881
			2	1.00	0.00	Ash ≤ 2.3 AND Totalphenols > 1.881
			3	1.00	0.00	Totalphenols ≤ 1.881
2^1	2^{-1}	$4.99 \cdot 10^3$	1	1.00	0.00	Ash > 2.3 AND Totalphenols > 1.881
			2	1.00	0.00	Ash ≤ 2.3 AND Totalphenols > 1.881
			3	1.00	0.00	Totalphenols ≤ 1.881
2^1	2^0	$4.99 \cdot 10^3$	1	1.00	0.00	Ash > 2.3 AND Totalphenols > 1.881
			2	1.00	0.00	Ash ≤ 2.3 AND Totalphenols > 1.881
			3	1.00	0.00	Totalphenols ≤ 1.881
2^1	2^1	$4.99 \cdot 10^3$	1	1.00	0.00	Ash > 2.3 AND Totalphenols > 1.881
			2	1.00	0.00	Ash ≤ 2.3 AND Totalphenols > 1.881
			3	1.00	0.00	Totalphenols ≤ 1.881

Table B.16
The clusters and the explanations provided by (CinterP), $\theta_1 \in \{2^p\}_{p=-1,0,1}$ and $\theta_2 \in \{2^p\}_{p=-1,0,1}$, for the `glass` dataset, with K = 6 clusters, explanations of a maximum length of $\ell = 2$ constructed with N = 139 rules using the deciles of the continuous features and all attributes of the categorical features.

θ_1	θ_2	Intra-homogeneity	Cluster	TPR	FPR	Explanations
2^{-1}	2^{-1}	$7.79 \cdot 10^2$	1	0.77	0.03	Al ≤ 1.36 AND Si ≤ 72.132
			2	1.00	0.00	Mg ≤ 2.805 AND Ca > 10.443
			3	0.95	0.08	K > 0.492 AND Fe ≤ 0.128
			4	1.00	0.01	Ca ≤ 10.443 AND Fe > 0.128
			5	0.96	0.01	Mg ≤ 0.6 AND Ba > 0
			6	0.44	0.00	Si ≤ 71.773 AND Ca ≤ 8.6
2^{-1}	2^0	$9.17 \cdot 10^2$	1	0.53	0.02	Al ≤ 1.146 AND Si ≤ 72.132
			2	1.00	0.00	Mg ≤ 2.805 AND Ca > 10.443
			3	1.00	0.04	K > 0.492 AND Fe ≤ 0.128
			4	1.00	0.01	Ca ≤ 10.443 AND Fe > 0.128
			5	0.91	0.00	K ≤ 0.08 AND Ba > 0
			6	0.44	0.00	RI ≤ 1.51869 AND Si ≤ 71.773
2^{-1}	2^1	$8.59 \cdot 10^2$	1	0.24	0.00	Mg > 3.757 AND K ≤ 0.19
			2	1.00	0.00	Mg ≤ 2.805 AND Ca > 10.443
			3	1.00	0.04	K > 0.492 AND Fe ≤ 0.07
			4	0.90	0.00	Ca ≤ 10.443 AND Fe > 0.128
			5	0.91	0.00	K ≤ 0.08 AND Ba > 0
			6	0.40	0.00	Si ≤ 71.773 AND Ca ≤ 8.6
2^0	2^{-1}	$7.79 \cdot 10^2$	1	0.80	0.03	Al ≤ 1.36 AND Si ≤ 72.132
			2	1.00	0.00	Mg ≤ 2.805 AND Ca > 10.443
			3	1.00	0.15	K > 0.19 AND Fe ≤ 0.128
			4	1.00	0.01	Ca ≤ 10.443 AND Fe > 0.128
			5	0.92	0.01	Mg ≤ 0.6 AND Ba > 0
			6	0.67	0.00	Si ≤ 72.132 AND Ca ≤ 7.97
2^0	2^0	$9.07 \cdot 10^2$	1	0.75	0.03	Al ≤ 1.36 AND Si ≤ 72.132
			2	1.00	0.00	Mg ≤ 2.805 AND Ca > 10.443
			3	0.97	0.07	K > 0.492 AND Fe ≤ 0.128
			4	1.00	0.01	Ca ≤ 10.443 AND Fe > 0.128
			5	0.91	0.00	K ≤ 0.08 AND Ba > 0
			6	0.60	0.00	Si ≤ 72.132 AND Ca ≤ 7.97
2^0	2^1	$9.07 \cdot 10^2$	1	0.75	0.03	Al ≤ 1.36 AND Si ≤ 72.132
			2	1.00	0.00	Mg ≤ 2.805 AND Ca > 10.443
			3	0.97	0.07	K > 0.492 AND Fe ≤ 0.128
			4	1.00	0.01	Ca ≤ 10.443 AND Fe > 0.128
			5	0.91	0.00	K ≤ 0.08 AND Ba > 0
			6	0.60	0.00	Si ≤ 72.132 AND Ca ≤ 7.97
2^1	2^{-1}	$7.75 \cdot 10^2$	1	0.80	0.03	Al ≤ 1.36 AND Si ≤ 72.132
			2	1.00	0.00	Mg ≤ 2.805 AND Ca > 10.443
			3	1.00	0.16	K > 0.19 AND Fe ≤ 0.128
			4	1.00	0.01	Ca ≤ 10.443 AND Fe > 0.128
			5	0.96	0.02	Al > 1.748 AND Ba > 0
			6	0.63	0.00	RI ≤ 1.51735 AND Si ≤ 72.132

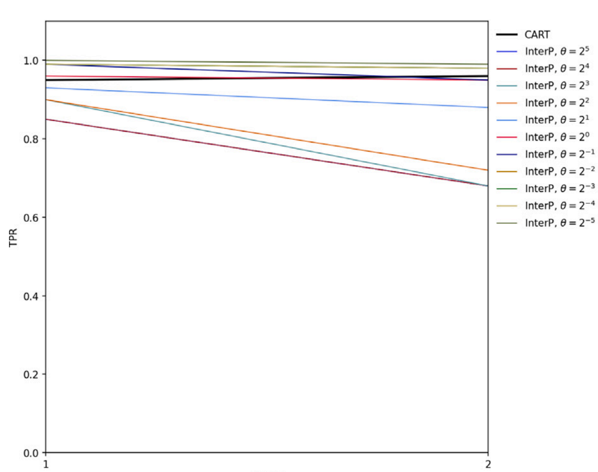
(*continued on next page*)

Table B.16 (*continued*).

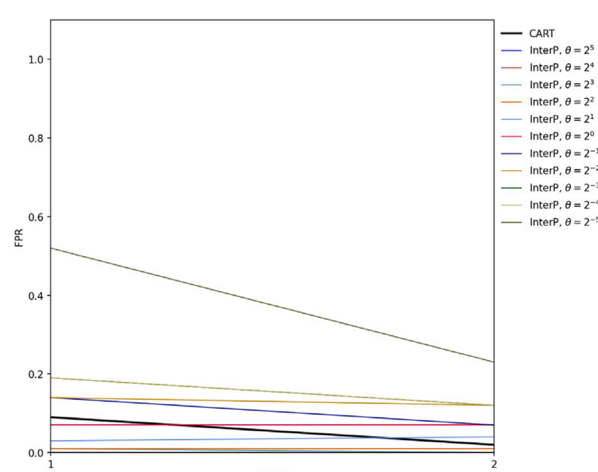
θ_1	θ_2	Intra-homogeneity	Cluster	TPR	FPR	Explanations
2^1	2^0	$7.73 \cdot 10^2$	1	0.83	0.03	Al $\leq$ 1.36 AND Si $\leq$ 72.132
			2	1.00	0.00	Mg $\leq$ 2.805 AND Ca > 10.443
			3	1.00	0.16	K > 0.19 AND Fe $\leq$ 0.128
			4	1.00	0.01	Ca $\leq$ 10.443 AND Fe > 0.128
			5	1.00	0.02	Al > 1.748 AND Ba > 0
			6	0.50	0.00	RI $\leq$ 1.51735 AND Si $\leq$ 72.132
2^1	2^1	$7.71 \cdot 10^2$	1	0.80	0.03	Al $\leq$ 1.36 AND Si $\leq$ 72.132
			2	1.00	0.00	Mg $\leq$ 2.805 AND Ca > 10.443
			3	0.95	0.09	K > 0.492 AND Fe $\leq$ 0.128
			4	1.00	0.01	Ca $\leq$ 10.443 AND Fe > 0.128
			5	1.00	0.02	Al > 1.748 AND Ba > 0
			6	0.50	0.00	RI $\leq$ 1.51735 AND Si $\leq$ 72.132

Table C.17
The clusters and the explanations provided by (InterP), $\theta \in \{2^p\}_{p=-5,\dots,5}$, for the `breast cancer` dataset, with K = 2 clusters, explanations of a maximum length of $\ell = 2$ constructed with N = 83 rules using the deciles of the continuous features and all attributes of the categorical features.

θ	Cluster	TPR	FPR	Explanations
2^5	1	0.85	0.00	Epithelial Size $\leq$ 3 AND Nuclei $\leq$ 1
	2	0.68	0.00	Size > 4 AND Adhesion > 1
2^4	1	0.85	0.00	Epithelial Size $\leq$ 3 AND Nuclei $\leq$ 1
	2	0.68	0.00	Size > 4 AND Adhesion > 1
2^3	1	0.90	0.01	Epithelial Size $\leq$ 3 AND Nuclei $\leq$ 2
	2	0.68	0.00	Size > 4 AND Adhesion > 1
2^2	1	0.90	0.01	Epithelial Size $\leq$ 3 AND Nuclei $\leq$ 2
	2	0.72	0.01	Size > 4
2^1	1	0.93	0.03	Shape $\leq$ 3 AND Chromatin $\leq$ 3
	2	0.88	0.04	Size > 1 AND Nuclei > 2
2^0	1	0.96	0.07	Size $\leq$ 4 AND Nuclei $\leq$ 4
	2	0.95	0.07	Size > 2 AND Shape > 1
2^{-1}	1	0.99	0.14	Size $\leq$ 4 AND Nuclei $\leq$ 9
	2	0.95	0.07	Size > 2 AND Shape > 1
2^{-2}	1	0.99	0.14	Size $\leq$ 4 AND Nuclei $\leq$ 9
	2	0.98	0.12	Size > 1 AND Shape > 1
2^{-3}	1	0.99	0.19	Thickness $\leq$ 9.8 AND Size $\leq$ 4
	2	0.98	0.12	Size > 1 AND Shape > 1
2^{-4}	1	0.99	0.19	Thickness $\leq$ 9.8 AND Size $\leq$ 4
	2	0.98	0.12	Size > 1 AND Shape > 1
2^{-5}	1	1.00	0.52	Thickness $\leq$ 9.8 AND Normal Nucleoli $\leq$ 9
	2	0.99	0.23	Shape > 1
CART	1	0.95	0.09	Size > 2.5 AND Shape $\leq$ 2.5 OR Size $\leq$ 2.5 AND Nuclei $\leq$ 5.5
	2	0.96	0.02	Size > 2.5 AND Shape > 2.5 OR Size $\leq$ 2.5 AND Nuclei > 5.5

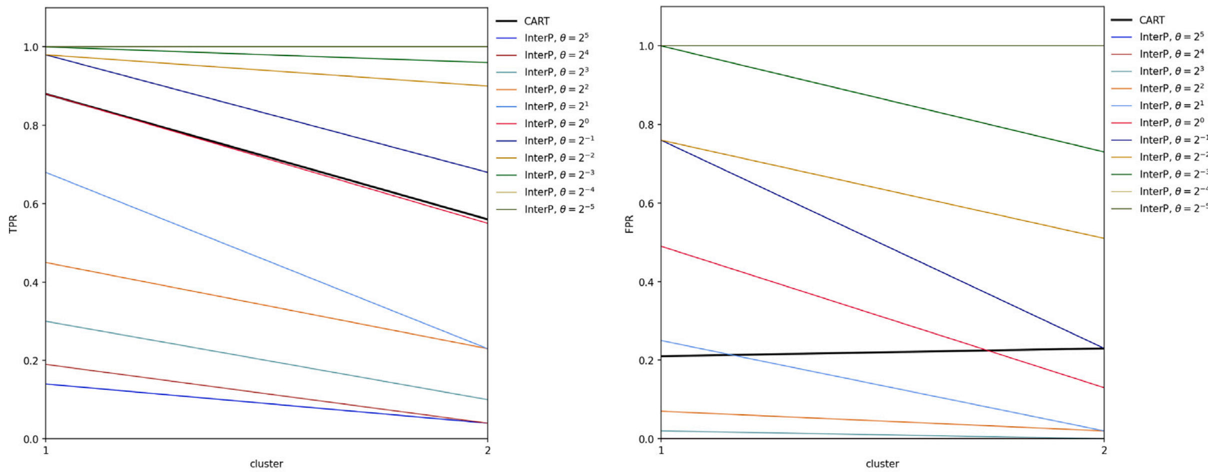


(a) True Positive Rate



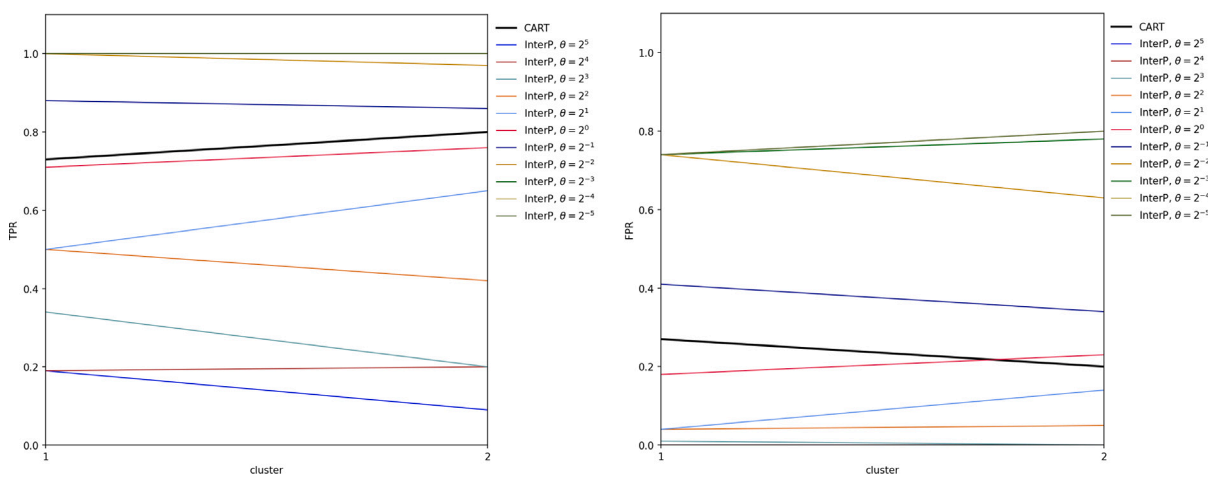
(b) False Positive Rate

Fig. C.4. The `breast cancer` data: the post-hoc interpretability results obtained by (InterP) and CART.



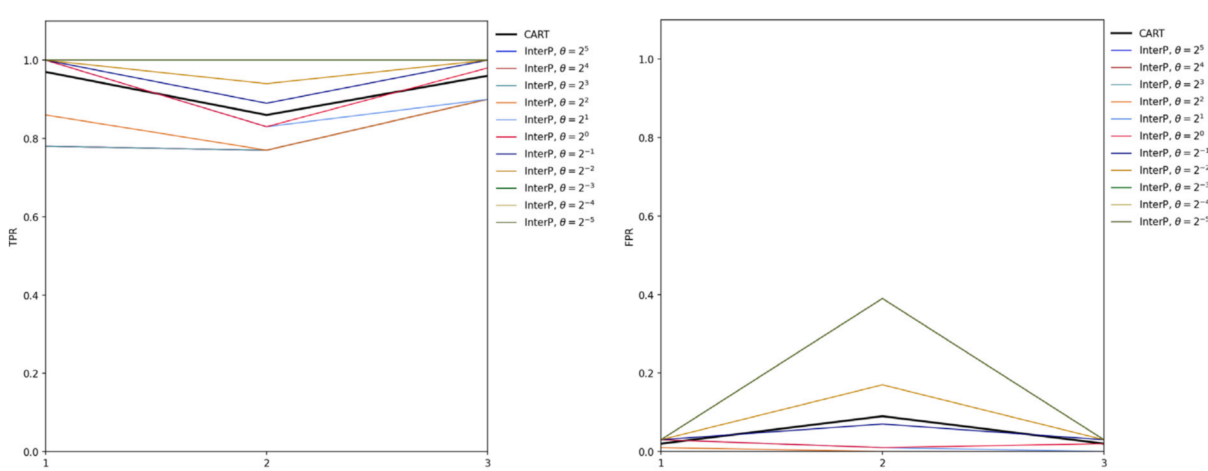
(a) True Positive Rate (b) False Positive Rate

Fig. C.5. The PIMA data: the post-hoc interpretability results obtained by (InterP) and CART.



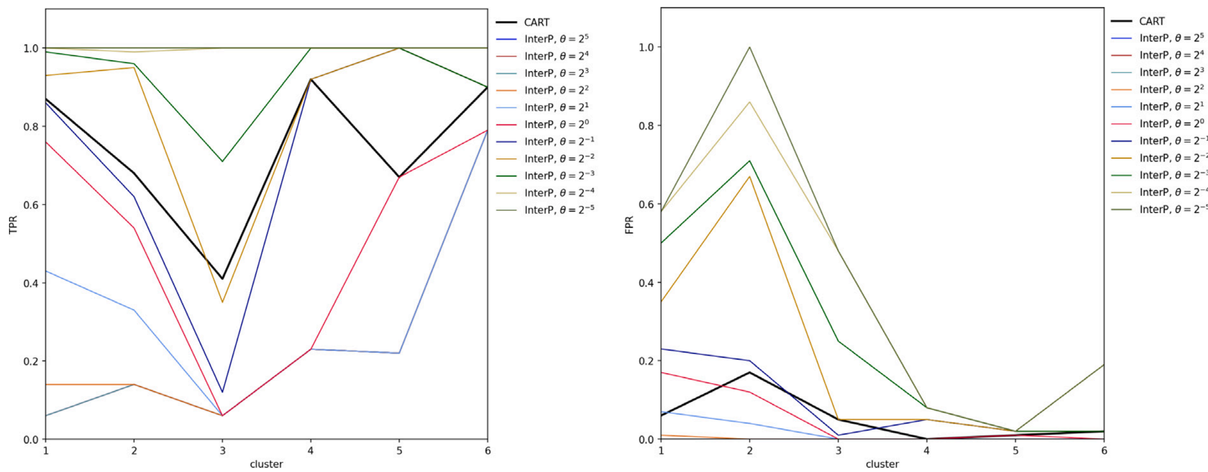
(a) True Positive Rate (b) False Positive Rate

Fig. C.6. The abalone data: the post-hoc interpretability results obtained by (InterP) and CART.



(a) True Positive Rate (b) False Positive Rate

Fig. C.7. The wine data: the post-hoc interpretability results obtained by (InterP) and CART.



(a) True Positive Rate (b) False Positive Rate

Fig. C.8. The `glass` data: the post-hoc interpretability results obtained by (InterP) and CART.

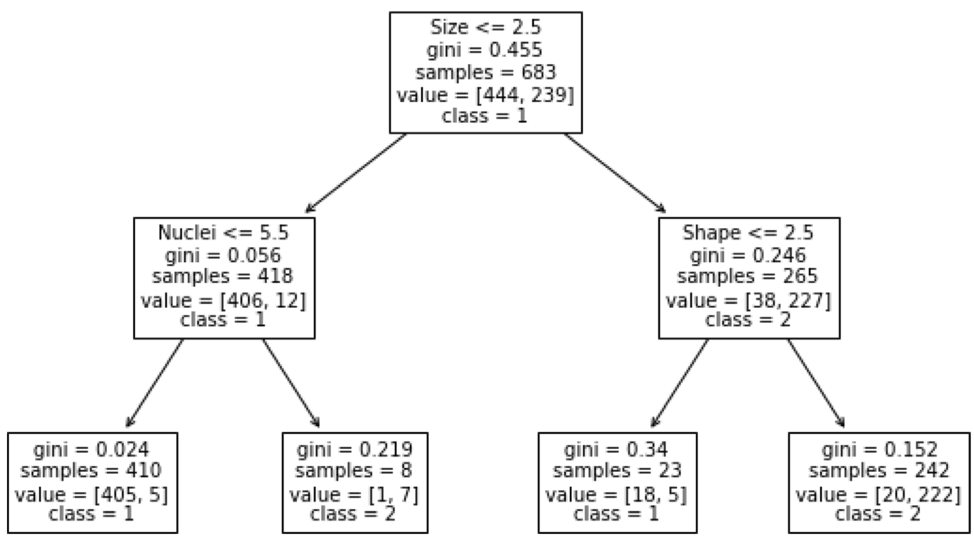


Fig. C.9. The post-hoc explanations provided by a CART of depth 2 for the `breast cancer` dataset for clusters (classes) 1 and 2.

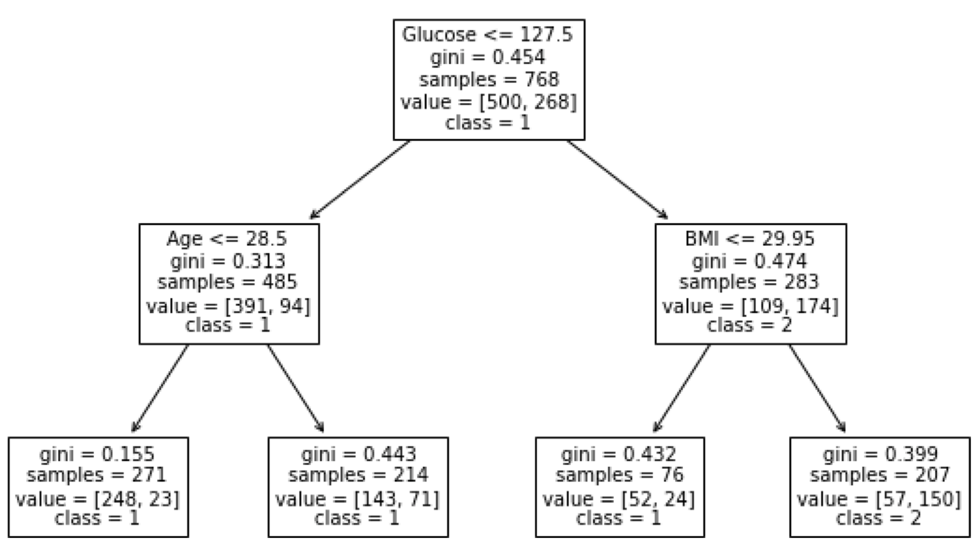


Fig. C.10. The post-hoc explanations provided by a CART of depth 2 for the `PIMA` dataset for clusters (classes) 1 and 2.

Table C.18
The clusters and the explanations provided by (InterP), $\theta \in \{2^p\}_{p=-5,\ldots,5}$, for the PIMA dataset, with K = 2 clusters, explanations of a maximum length of $\ell = 2$ constructed with N = 135 rules using the deciles of the continuous features and all attributes of the categorical features.

θ	Cluster	TPR	FPR	Explanations
2^5	1	0.14	0.00	Glucose ≤ 102 AND BMI ≤ 25.9
	2	0.04	0.00	Glucose > 167 AND SkinThickness > 40
2^4	1	0.19	0.00	Glucose ≤ 102 AND BMI ≤ 28.2
	2	0.04	0.00	Glucose > 167 AND SkinThickness > 40
2^3	1	0.30	0.02	BMI ≤ 30.1 AND Age ≤ 27
	2	0.10	0.00	Glucose > 167 AND SkinThickness > 31
2^2	1	0.45	0.07	Glucose ≤ 117 AND Age ≤ 29
	2	0.23	0.02	Glucose > 167 AND BMI > 28.2
2^1	1	0.68	0.25	Pregnancies ≤ 7 AND Glucose ≤ 125
	2	0.23	0.02	Glucose > 167 AND BMI > 28.2
2^0	1	0.88	0.49	Glucose ≤ 147 AND BMI ≤ 41.5
	2	0.55	0.13	Glucose > 125 AND BMI > 30.1
2^{-1}	1	0.98	0.76	Glucose ≤ 167
	2	0.68	0.23	Glucose > 117 AND BMI > 28.2
2^{-2}	1	0.98	0.76	Glucose ≤ 167
	2	0.90	0.51	Glucose > 95 AND BMI > 25.9
2^{-3}	1	1.00	1.00	All in
	2	0.96	0.73	Glucose > 85 AND BMI > 23.6
2^{-4}	1	1.00	1.00	All in
	2	1.00	1.00	All in
2^{-5}	1	1.00	1.00	All in
	2	1.00	1.00	All in
CART	1	0.88	0.21	Glucose ≤ 127.5 OR Glucose > 127.5 AND BMI ≤ 29.95
	2	0.56	0.23	Glucose ≤ 127.5 AND BMI > 29.95

Table C.19
The clusters and the explanations provided by (InterP), $\theta \in \{2^p\}_{p=-5,\ldots,5}$, for the abalone dataset, with K = 2 clusters, explanations of a maximum length of $\ell = 2$ constructed with N = 130 rules using the deciles of the continuous features and all attributes of the categorical features.

θ	Cluster	TPR	FPR	Explanations
2^5	1	0.19	0.00	Sex = I AND Height ≤ 0.085
	2	0.09	0.00	Sex = M AND Shell weight > 0.41125
2^4	1	0.19	0.00	Sex = I AND Height ≤ 0.085
	2	0.20	0.00	Shell weight > 0.41125
2^3	1	0.34	0.01	Sex = I AND Height ≤ 0.105
	2	0.20	0.00	Shell weight > 0.41125
2^2	1	0.50	0.04	Sex = I AND Height ≤ 0.135
	2	0.42	0.05	Height > 0.16 AND Shell weight > 0.3065
2^1	1	0.50	0.04	Sex = I AND Height ≤ 0.135
	2	0.65	0.14	Diameter > 0.4 AND Shell weight > 0.268
2^0	1	0.71	0.18	Height ≤ 0.14 AND Shell weight ≤ 0.23475
	2	0.76	0.23	Diameter > 0.365 AND Shell weight > 0.23475
2^{-1}	1	0.88	0.41	Height ≤ 0.16 AND Shell weight ≤ 0.3065
	2	0.86	0.34	Whole weight > 0.521 AND Shell weight > 0.19
2^{-2}	1	1.00	0.74	Height ≤ 0.185 AND Shell weight ≤ 0.41125
	2	0.97	0.63	Whole weight > 0.1955 AND Shell weight > 0.103
2^{-3}	1	1.00	0.74	Height ≤ 0.185 AND Shell weight ≤ 0.41125
	2	1.00	0.78	Whole weight > 0.1955 AND Viscera weight > 0.04
2^{-4}	1	1.00	0.74	Height ≤ 0.185 AND Shell weight ≤ 0.41125
	2	1.00	0.80	Whole weight > 0.1955 AND all in
2^{-5}	1	1.00	0.74	Height ≤ 0.185 AND Shell weight ≤ 0.41125
	2	1.00	0.80	Whole weight > 0.1955
CART	1	0.73	0.27	Shell weight ≤ 0.217
	2	0.8	0.2	Shell weight > 0.217

Table C.20
The clusters and the explanations provided by (InterP), $\theta \in \{2^p\}_{p=-5,\dots,5}$, for the `wine` dataset, with K = 3 clusters, explanations of a maximum length of $\ell = 2$ constructed with N = 235 rules using the deciles of the continuous features and all attributes of the categorical features.

θ	Cluster	TPR	FPR	Explanations
2^5	1	0.78	0.00	Alcohol > 13.05 AND Proline > 879
	2	0.77	0.00	Colorintensity ≤ 3.4
	3	0.90	0.00	Flavanoids ≤ 1.324 AND Colorintensity > 4.08
2^4	1	0.78	0.00	Alcohol > 13.05 AND Proline > 879
	2	0.77	0.00	Colorintensity ≤ 3.4
	3	0.90	0.00	Flavanoids ≤ 1.324 AND Colorintensity > 4.08
2^3	1	0.78	0.00	Alcohol > 13.05 AND Proline > 879
	2	0.77	0.00	Colorintensity ≤ 3.4
	3	0.90	0.00	Flavanoids ≤ 1.324 AND Colorintensity > 4.08
2^2	1	0.86	0.01	Flavanoids > 2.46 AND Proline > 742
	2	0.77	0.00	Colorintensity ≤ 3.4
	3	0.90	0.00	Flavanoids ≤ 1.324 AND Colorintensity > 4.08
2^1	1	1.00	0.03	Flavanoids > 2.135 AND Alcohol > 12.76
	2	0.83	0.01	Alcohol ≤ 12.76 AND Colorintensity ≤ 4.69
	3	0.90	0.00	Flavanoids ≤ 1.324 AND Colorintensity > 4.08
2^0	1	1.00	0.03	Flavanoids > 2.135 AND Alcohol > 12.76
	2	0.83	0.01	Alcohol ≤ 12.76 AND Colorintensity ≤ 4.69
	3	0.98	0.02	Flavanoids ≤ 1.738 AND Hue ≤ 0.91
2^{-1}	1	1.00	0.03	Flavanoids > 2.135 AND Alcohol > 12.76
	2	0.89	0.07	Alcohol ≤ 13.05 AND Colorintensity ≤ 4.69
	3	1.00	0.03	Flavanoids ≤ 1.738 AND Colorintensity > 3.4
2^{-2}	1	1.00	0.03	Flavanoids > 2.135 AND Alcohol > 12.76
	2	0.94	0.17	Proline ≤ 1048 AND Colorintensity ≤ 4.69
	3	1.00	0.03	Flavanoids ≤ 1.738 AND Colorintensity > 3.4
2^{-3}	1	1.00	0.03	Flavanoids > 2.135 AND Alcohol > 12.76
	2	1.00	0.39	Proline ≤ 1048 AND Colorintensity ≤ 6.99
	3	1.00	0.03	Flavanoids ≤ 1.738 AND Colorintensity > 3.4
2^{-4}	1	1.00	0.03	Flavanoids > 2.135 AND Alcohol > 12.76
	2	1.00	0.39	Proline ≤ 1048 AND Colorintensity ≤ 6.99
	3	1.00	0.03	Flavanoids ≤ 1.738 AND Colorintensity > 3.4
2^{-5}	1	1.00	0.03	Flavanoids > 2.135 AND Alcohol > 12.76
	2	1.00	0.39	Proline ≤ 1048 AND Colorintensity ≤ 6.99
	3	1.00	0.03	Flavanoids ≤ 1.738 AND Colorintensity > 3.4
CART	1	0.97	0.02	Proline > 755.0 AND Flavanoids > 2.165
	2	0.86	0.09	Proline ≤ 755.0 AND OD280andOD31ofdilutedwines > 2.115
	3	0.96	0.02	Proline > 755.0 AND Flavanoids ≤ 2.165 OR Proline ≤ 755.0 AND OD280andOD31ofdilutedwines ≤ 2.115

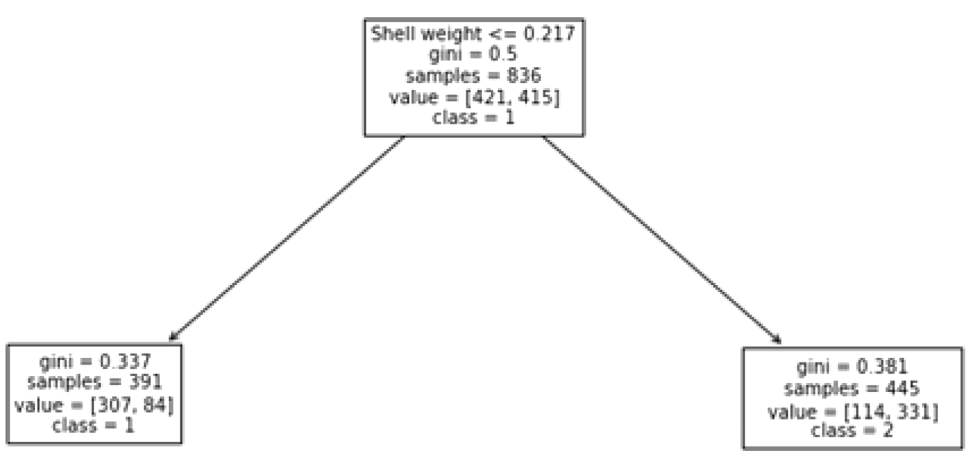


Fig. C.11. The post-hoc explanations provided by a CART of depth 1 for the `abalone` dataset for clusters (classes) 1 and 2.

Table C.21
The clusters and the explanations provided by (InterP), $\theta \in \{2^p\}_{p=-5,\ldots,5}$, for the `glass` dataset, with K = 6 clusters, explanations of a maximum length of $\ell = 2$ constructed with N = 139 rules using the deciles of the continuous features and all attributes of the categorical features.

θ	Cluster	TPR	FPR	Explanations
2^5	1	0.06	0.00	RI $\le$ 1.5163 AND Fe > 0.22
	2	0.14	0.00	Mg > 3.757 AND Ca $\le$ 8.6
	3	0.06	0.00	Na > 14.018 AND Fe > 0.22
	4	0.23	0.00	RI $\le$ 1.51591 AND Si $\le$ 71.773
	5	0.22	0.00	K $\le$ 0 AND Ca $\le$ 7.97
	6	0.79	0.00	Na > 14.018 AND Ba > 0
2^4	1	0.06	0.00	RI $\le$ 1.5163 AND Fe > 0.22
	2	0.14	0.00	Mg > 3.757 AND Ca $\le$ 8.6
	3	0.06	0.00	Na > 14.018 AND Fe > 0.22
	4	0.23	0.00	RI $\le$ 1.51591 AND Si $\le$ 71.773
	5	0.22	0.00	K $\le$ 0 AND Ca $\le$ 7.97
	6	0.79	0.00	Na > 14.018 AND Ba > 0
2^3	1	0.06	0.00	RI $\le$ 1.5163 AND Fe > 0.22
	2	0.14	0.00	Mg > 3.757 AND Ca $\le$ 8.6
	3	0.06	0.00	Na > 14.018 AND Fe > 0.22
	4	0.23	0.00	RI $\le$ 1.51591 AND Si $\le$ 71.773
	5	0.22	0.00	K $\le$ 0 AND Ca $\le$ 7.97
	6	0.79	0.00	Na > 14.018 AND Ba > 0
2^2	1	0.14	0.01	Mg > 3.39 AND Ca > 9.57
	2	0.14	0.00	Mg > 3.757 AND Ca $\le$ 8.6
	3	0.06	0.00	Na > 14.018 AND Fe > 0.22
	4	0.23	0.00	RI $\le$ 1.51591 AND Si $\le$ 71.773
	5	0.22	0.00	K $\le$ 0 AND Ca $\le$ 7.97
	6	0.79	0.00	Na > 14.018 AND Ba > 0
2^1	1	0.43	0.07	Mg > 3.39 AND Ca > 8.6
	2	0.33	0.04	Mg > 3.48 AND Ca $\le$ 8.12
	3	0.06	0.00	Na > 14.018 AND Fe > 0.22
	4	0.23	0.00	RI $\le$ 1.51591 AND Si $\le$ 71.773
	5	0.22	0.00	K $\le$ 0 AND Ca $\le$ 7.97
	6	0.79	0.00	Na > 14.018 AND Ba > 0
2^0	1	0.76	0.17	RI > 1.51735 AND Mg > 3.39
	2	0.54	0.12	Mg > 2.805 AND Ca $\le$ 8.339
	3	0.06	0.00	Na > 14.018 AND Fe > 0.22
	4	0.23	0.00	RI $\le$ 1.51591 AND Si $\le$ 71.773
	5	0.67	0.01	Si $\le$ 72.79 AND K $\le$ 0
	6	0.79	0.00	Na > 14.018 AND Ba > 0

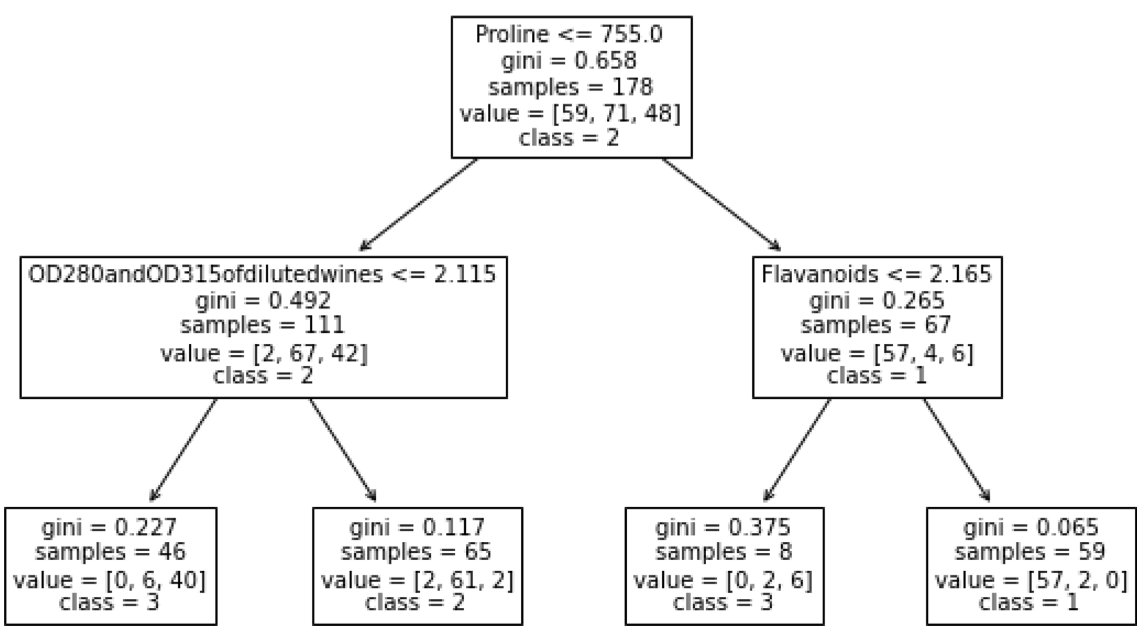


Fig. C.12. The post-hoc explanations provided by a CART of depth 2 for the `wine` dataset for clusters (classes) 1, 2 and 3.

Table C.22
The clusters and the explanations provided by (InterP), $\theta \in \{2^p\}_{p=-5,\dots,5}$, for the `glass` dataset, with K = 6 clusters, explanations of a maximum length of $\ell = 2$ constructed with N = 139 rules using the deciles of the continuous features and all attributes of the categorical features (cont.).

θ	Cluster	TPR	FPR	Explanations
2^{-1}	1	0.86	0.23	RI > 1.51735 AND Mg > 2.805
	2	0.62	0.20	Mg > 2.805 AND Ca ≤ 8.482
	3	0.12	0.01	Na > 13.3 AND Fe > 0.22
	4	0.92	0.05	Na ≤ 13.44 AND Mg ≤ 2.805
	5	1.00	0.02	K ≤ 0 AND Ba ≤ 0
	6	0.90	0.02	Na > 13.3 AND Al > 1.748
2^{-2}	1	0.93	0.35	Al ≤ 1.488 AND Ca ≤ 10.443
	2	0.95	0.67	Na ≤ 14.018 AND Ba ≤ 0.64
	3	0.35	0.05	RI ≤ 1.51735 AND Al ≤ 1.36
	4	0.92	0.05	Na ≤ 13.44 AND Mg ≤ 2.805
	5	1.00	0.02	K ≤ 0 AND Ba ≤ 0
	6	0.90	0.02	Na > 13.3 AND Al > 1.748
2^{-3}	1	0.99	0.50	Mg > 2.805 AND Al ≤ 1.748
	2	0.96	0.71	Na ≤ 14.018
	3	0.71	0.25	Na > 13.3 AND Mg > 2.805
	4	1.00	0.08	Mg ≤ 2.805 AND K > 0.08
	5	1.00	0.02	K ≤ 0 AND Ba ≤ 0
	6	0.90	0.02	Na > 13.3 AND Al > 1.748
2^{-4}	1	1.00	0.58	Al ≤ 1.748 AND Ca ≤ 10.443
	2	0.99	0.86	Ba ≤ 0.64
	3	1.00	0.48	Mg > 2.805 AND Ca > 8.12
	4	1.00	0.08	Mg ≤ 2.805 AND K > 0.08
	5	1.00	0.02	K ≤ 0 AND Ba ≤ 0
	6	1.00	0.19	Mg ≤ 3.39 AND Ca ≤ 10.443
2^{-5}	1	1.00	0.58	Al ≤ 1.748 AND Ca ≤ 10.443
	2	1.00	1.00	All in
	3	1.00	0.48	Mg > 2.805 AND Ca > 8.12
	4	1.00	0.08	Mg ≤ 2.805 AND K > 0.08
	5	1.00	0.02	K ≤ 0 AND Ba ≤ 0
	6	1.00	0.19	Mg ≤ 3.39 AND Ca ≤ 10.443
CART	1	0.87	0.06	Ba ≤ 0.335 AND Al ≤ 1.42 AND Ca ≤ 10.48 AND Rl > 1.517 OR Ba > 0.335 AND Si > 70.16 AND Mg > 3.42
	2	0.68	0.17	Ba ≤ 0.335 AND Al ≤ 1.42 AND Ca > 10.48 AND Na ≤14.495 OR Ba ≤ 0.335 AND Al > 1.42 AND Mg > 2.26 OR Ba > 0.335 AND Si ≤ 70.16 AND Ca > 9.585
	3	0.41	0.05	Ba ≤ 0.335 AND Al ≤ 1.42 AND Ca ≤ 10.48 AND Rl ≤ 1.517
	4	0.92	0.00	Ba ≤ 0.335 AND Al > 1.42 AND Mg ≤ 2.26 AND Na ≤ 13.495 OR Ba > 0.335 AND Si ≤ 70.16 AND Ca ≤ 9.585
	5	0.67	0.01	Ba ≤ 0.335 AND Al ≤ 1.42 AND Ca > 10.48 AND Na > 14.495 OR Ba ≤ 0.335 AND Al > 1.42 AND Mg ≤ 2.26 AND Na > 13.495
	6	0.90	0.02	Ba > 0.335 AND Si > 70.16 AND Mg ≤ 3.42

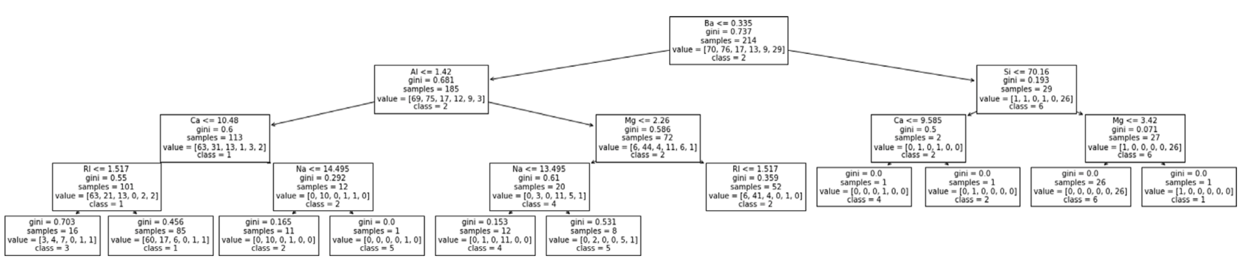


Fig. C.13. The post-hoc explanations provided by a CART of depth 4 for the `glass` dataset for clusters (classes) 1, 2, 3, 4, 5 and 6.

References

Abraham, S.S., P, D., Sundaram, S.S., 2020. Fairness in clustering with multiple sensitive attributes. In: EDBT/ICDT 2020 Joint Conference. pp. 287–298.

Aloise, D., Hansen, P., Liberti, L., 2012. An improved column generation algorithm for minimum sum-of-squares clustering. Math. Program. 131 (1–2), 195–220.

Baesens, B., Setiono, R., Mues, C., Vanthienen, J., 2003. Using neural network rule extraction and decision tables for credit-risk evaluation. Manage. Sci. 49 (3), 312–329.

Balabaeva, K., Kovalchuk, S., 2020. Post-hoc interpretation of clinical pathways clustering using Bayesian inference. Procedia Comput. Sci. 178, 264–273, 9th International Young Scientists Conference in Computational Science, YSC2020, 05-12 September 2020.

Basak, J., Krishnapuram, R., 2005. Interpretable hierarchical clustering by constructing an unsupervised decision tree. IEEE Trans. Knowl. Data Eng. 17 (1), 121–132.

Bénard, C., Biau, G., Da Veiga, S., Scornet, E., 2019. SIRUS: Making random forests interpretable. arXiv preprint arXiv:1908.06852.

Benítez-Peña, S., Blanquero, R., Carrizosa, E., Ramírez-Cobo, P., 2019. Cost-sensitive feature selection for support vector machines. Comput. Oper. Res. 106, 169–178.

Bertsimas, D., King, A., 2016. OR forum – An algorithmic approach to linear regression. Oper. Res. 64 (1), 2–16.

Bertsimas, D., Orfanoudaki, A., Wiberg, H., 2021. Interpretable clustering: An optimization approach. Mach. Learn. 110 (1), 89–138.

Carrizosa, E., Guerrero, V., Romero Morales, D., Satorra, A., 2020. Enhancing interpretability in Factor Analysis by means of Mathematical Optimization. Multivar. Behav. Res. 55 (5), 748–762.

Carrizosa, E., Kurishchenko, K., Marín, A., Romero Morales, D., 2022. Interpreting clusters via prototype optimization. Omega 107, 102543.

Carrizosa, E., Molero-Río, C., Romero Morales, D., 2021. Mathematical optimization in classification and regression trees. TOP 29 (1), 5–33.

Carrizosa, E., Nogales-Gómez, A., Romero Morales, D., 2016. Strongly agree or strongly disagree?: Rating features in support vector machines. Inform. Sci. 329, 256–273.

Carrizosa, E., Romero Morales, D., 2013. Supervised classification and mathematical optimization. Comput. Oper. Res. 40 (1), 150–165.

Chen, J., Chang, Y., Hobbs, B., Castaldi, P., Cho, M., Silverman, E., Dy, J., 2016. Interpretable clustering via discriminative rectangle mixture model. In: 2016 IEEE 16th International Conference on Data Mining. ICDM, pp. 823–828.

Corral, G., Armengol, E., Fornells, A., Golobardes, E., 2009. Explanations of unsupervised learning clustering applied to data security analysis. Neurocomputing 72 (13), 2754–2762, Hybrid Learning Machines (HAIS 2007) / Recent Developments in Natural Computation (ICNC 2007).

Dasgupta, S., Frost, N., Moshkovitz, M., Rashtchian, C., 2020. Explainable k-means and k-medians clustering. In: Proceedings of the 37th International Conference on Machine Learning. pp. 7055–7065.

Davidson, I., Gourru, A., Ravi, S., 2018. The cluster description problem - complexity results, formulations and approximations. In: Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (Eds.), Advances in Neural Information Processing Systems, Vol. 31. Curran Associates, Inc..

De Koninck, P., De Weerdt, J., vanden Broucke, S.L., 2017. Explaining clusterings of process instances. Data Min. Knowl. Discov. 31 (3), 774–808.

Dua, D., Graff, C., 2017. UCI machine learning repository. URL http://archive.ics.uci.edu/ml.

European Commission, 2020. White Paper on Artificial Intelligence: a European approach to excellence and trust. URL https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf.

Fraiman, R., Ghattas, B., Svarc, M., 2013. Interpretable clustering using unsupervised binary trees. Adv. Data Anal. Classif. 7 (2), 125–145.

Gibert, K., Conti, D., 2016. On the understanding of profiles by means of post-processing techniques: An application to financial assets. Int. J. Comput. Math. 93 (5), 807–820.

Goodman, B., Flaxman, S., 2017. European Union regulations on algorithmic decision-making and a "right to explanation". AI Mag. 38 (3), 50–57.

Gurobi Optimization, 2020. Gurobi optimizer reference manual. URL http://www.gurobi.com.

Hazimeh, H., Mazumder, R., 2020. Fast best subset selection: Coordinate descent and local combinatorial optimization algorithms. Oper. Res. 68 (5), 1517–1537.

Jiménez-Cordero, A., Morales, J.M., Pineda, S., 2021. A novel embedded min-max approach for feature selection in nonlinear support vector machine classification. European J. Oper. Res. 293 (1), 24–35.

Kauffmann, J., Esders, M., Ruff, L., Montavon, G., Samek, W., Müller, K.-R., 2022. From clustering to cluster explanations via neural networks. IEEE Trans. Neural Netw. Learn. Syst..

Kaufmann, L., Rousseeuw, P.J., 1990. Finding Groups in Data: An Introduction to Cluster Analysis. New York.

Kim, B., Rudin, C., Shah, J.A., 2014. The Bayesian case model: A generative approach for case-based reasoning and prototype classification. In: Advances in Neural Information Processing Systems. pp. 1952–1960.

Lawless, C., Kalagnanam, J., Nguyen, L.M., Phan, D., Reddy, C., 2022. Interpretable clustering via multi-polytope machines. In: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 36. pp. 7309–7316.

Ma, R., Angryk, R.A., Riley, P., Boubrahimi, S.F., 2018. Coronal mass ejection data clustering and visualization of decision trees. Astrophys. J. Suppl. Ser. 236 (1), 14.

Mišić, V.V., 2020. Optimization of tree ensembles. Oper. Res. 68 (5), 1605–1624.

Morichetta, A., Casas, P., Mellia, M., 2019. EXPLAIN-IT: Towards explainable AI for unsupervised network traffic analysis. In: Proceedings of the 3rd ACM CoNEXT Workshop on Big DAta, Machine Learning and Artificial Intelligence for Data Communication Networks - Big-DAMA '19. ACM Press, pp. 22–28.

Python Core Team, 2015. Python: A Dynamic, Open Source Programming Language. URL https://www.python.org.

Rader, E., Cotter, K., Cho, J., 2018. Explanations as mechanisms for supporting algorithmic transparency. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, New York, NY, USA, pp. 1–13.

Rao, M., 1971. Cluster analysis and mathematical programming. J. Amer. Statist. Assoc. 66 (335), 622–626.

Rodrigues, R., 2020. Legal and human rights issues of AI: Gaps, challenges and vulnerabilities. J. Responsible Technol. 4, 100005.

Saisubramanian, S., Galhotra, S., Zilberstein, S., 2020. Balancing the tradeoff between clustering value and interpretability. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. pp. 351–357.

Samek, W., Montavon, G., Lapuschkin, S., Anders, C.J., Müller, K.-R., 2021. Explaining deep neural networks and beyond: A review of methods and applications. Proc. IEEE 109 (3), 247–278.

Taeb, A., Chandrasekaran, V., 2018. Interpreting latent variables in factor models via convex optimization. Math. Program. 167 (1), 129–154.

Thomassey, S., Fiordaliso, A., 2006. A hybrid sales forecasting system based on clustering and decision trees. Decis. Support Syst. 42 (1), 408–421.