Erasmus University Rotterdam
Erasmus School of Economics
Bachelor Thesis Econometrics and Operations Research

Interpretable Cluster Analysis with Imputed Data

Eva van Dijk (570467)

Supervisor:	R.S.H. Willemsen
Second assessor:	W. van den Heuvel
Date final version:	1st July 2023

Abstract

The use of machine learning programs is increasing but these programs often do not explain how the results are obtained. To add to the research of interpretable machine learning, this research paper validates the workings of the interpretable Clustering Analysis MILP formulations introduced by Carrizosa et al. (2023). The two MILP models form rule-based explanations for obtained clusters. The quality of the explanations is measured in accuracy and distinctiveness. One of the MILP models simultaneously provides explanations and clusters, while the other attains explanations for predetermined clusters. To further enhance the interpretability of Cluster Analysis, an additional MILP model is introduced that simultaneously forms clusters and their corresponding explanations while accounting for imputed values. The MILP model penalizes the use of rules based on features containing imputed values attempting to find better explanations based solely on observed data. In the paper, we find that the additional MILP model finds similar and in some cases better results than the MILP of Carrizosa et al. (2023).

The views stated in this thesis are those of the author and not necessarily those of the supervisor, second assessor, Erasmus School of Economics or Erasmus University Rotterdam.

1 Introduction

The use of machine learning (ML) is an up-and-coming practice in our daily life. ML has become a crucial aspect of the digital world in the last few years because of its accuracy and simplicity of use (Aggarwal et al., 2022). While the ML systems do not provide an explanation of their results, the demand for the use of the systems seems to rise and broaden towards different fields, such as banking (Doumpos et al., 2022), criminal justice (Saunders, Hunt & Hollywood, 2016) and medicine (Patterson et al., 2019). However, it is not always ethical to use ML in these fields. For example, in determining if a client is credit-worthy the decision must be supported by logical explanations that the client can understand. ML programs often do not provide these explanations and are therefore not ethical to use for creditworthiness determination. To make use of ML systems ethically in these fields, users should be able to understand the decisions and functionality of ML machines (Bacelar, 2021).

Together with ML, Explainable Artificial Intelligence (XAI) has gained attention in recent years (Kukkonen, Lindroos & Brauer, 2022). XAI methods supply the ML user with further explanation about the decisions the system makes and what it learns from the provided input data. The predominance of XAI research focuses on supervised ML methods while research on interpretable unsupervised methods is more scarce. Unsupervised methods run on unlabeled data, where the correct answers are unknown. The performance of these methods is therefore difficult to measure. Additionally, the methods are often applied to search for hidden patterns within the data. Explaining the results provides insight into the uncovered patterns that had previously gone unnoticed (Montavon et al., 2022). So not only can supervised methods benefit from explanations, but also unsupervised methods. Especially one of the most common classes of unsupervised learning algorithms, namely Clustering. A clustering algorithm assigns observations to clusters based on their features such that intra-homogeneous and inter-heterogeneous clusters are formed.

Current research focuses on interpretable Cluster Analysis by applying Integer Programming techniques. Carrizosa et al. (2023) constructed a Mixed Integer Linear Programming (MILP) problem that simultaneously forms intra-homogeneous clusters while creating rule-based explanations for each cluster. They also provide a MILP to extract distinct explanations for already predetermined clusters.

This paper utilizes the two models of Carrizosa et al. (2023) to validate their results and to further extend the models. Carrizosa et al. (2023) assume excess to a complete dataset, while in most real-life cases a complete dataset is hard to obtain (Altman & Bland, 2007). The chances are high that a dataset contains missing values. The missing values can be imputed, altering the distribution, the correlation between variables and the sample variance. As a consequence, imputation can provide biased results (White, Daniel & Royston, 2010). In the case of the models created by Carrizosa et al. (2023), imputation can lead to biased cluster explanations.

An extended version of the clustering and explanation providing MILP that penalizes the use of features with imputed data to explain the clusters is investigated in this paper. The goal is to find better explanations for a cluster taking into account which features contain imputed values to prevent the explanations from being biased. When considering an incomplete dataset the extended MILP seems to provide similar and in some cases better results for the real data than for the original MILP.

The remainder of the paper is structured as follows. The next section provides a literature overview on interpretable Cluster Analysis and dealing with missing data while clustering. Section 3 gives a short description of dealing with the missing data problem. In Section 4, the MILP formulations are explained in detail. Next, the results are presented. Lastly, Section 6 concludes the paper.

2 Literature

There is a need for interpretable unsupervised machine learning methods, particularly interpretable Cluster Analysis (CA). We first discuss the already available interpretable CA algorithms. Next, a shortcoming that all of the mentioned methods have in common and how it is dealt with while performing CA is discussed, namely the assumption of excess to a complete dataset.

2.1 Interpretable Cluster Analysis

Recent research focuses on developing Integer Programming (IP) formulations that provide distinctive rule-based explanations to previously determined clusters, a so-called post-hoc approach. In 2022, Lawless and Gunluk (2022) framed an IP technique regarding a polyhedral description problem which explains the grouping of data points within a cluster by forming a polyhedron surrounding them.

Carrizosa et al. (2022) proposed two MILP formulations, a covering and partitioning model based on classic Location Analysis problems, the covering and the p-median problem. The covering problem maximizes demand satisfied among customers while having excess to p facilities that can be opened or closed (García & Marín, 2015). The p-median problem allocates demand to customers with a fixed number of p open facilities while minimizing the total travel distance (Daskin & Maass, 2015). At the end of the paper, the authors suggest the next step in their research, namely presenting a MILP technique that simultaneously divides the data into intra-homogeneous inter-heterogeneous clusters and provides corresponding explanations. Such a MILP formulation is also referred to as an intrinsic model. The following year the writers published a paper covering a MILP formulation that concurrently builds and explains clusters (Carrizosa et al., 2023). In addition, another post-hoc model based on this intrinsic model is formulated in the same paper.

Lawless et al. (2022) also constructed an intrinsic model. The model is a Mixed Integer Non-Linear Programming (MINLP) problem which surrounds each cluster with a polytope and searches for separating planes, also known as hyperplanes, to aid the interpretability. Polytopes are geometric shapes with a flat side that surround the observations. The authors claim that their model has a higher expressive power than decision-tree approaches since polytopes can be plotted into a decision tree together with hyperplanes that are aligned with the axes.

2.2 Handling Missing Data

An important observation is that all IP techiniques denoted in the previously mentioned papers assume the availability of a complete dataset. In practice, the gathered data has almost always some proportion that is incomplete. This missing data can arise in only one attribute or more, caused completely at random, at random or not at random (Lin & Tsai, 2020).

There are two ways to deal with missing data: deletion or imputation. Once the percentage of missing values exceeds 15% of the data, the observations with missing features can no longer be deleted without it having a significant effect on the final clustering result (Acuna & Rodriguez, 2004). When this occurs the missing data is often imputed. There are various imputation methods. Popular methods in comparative research are K-Nearest-Neighbour (KNN), Singular Value Decomposition, Bayesian Principal Component Analysis (BPCA) and Median imputation (Celton, Malpertuy, Lelandais & De Brevern, 2010).

For decades, the statistical technique of replacing missing values with the mean has been well-researched and validated as a reliable approach to complete a dataset (Little & Rubin, 2019). De Souto, Jaskowiak and Costa (2015) concluded in their comparative analysis on the impact of different imputation methods that the mean imputation does not significantly impact the quality of the clusters compared to the results from the complete dataset. The mean method performed as well as more complex strategies, such as KNN and BPCA.

Researchers also found multiple ways to perform clustering on the incomplete dataset without the need for deletion or imputation. Chi, Chi and Baraniuk (2016) proposed a version of the k-means clustering algorithm that is applicable even when the dataset is incomplete. The method determines the differences between the observations only using the observed values, avoiding wasted data as a consequence of deleting observations and avoiding erroneous differences caused by wrongly imputed values. Wagstaff (2004) constructed a clustering algorithm that contains a set of constraints based solely on known values. Again providing a clustering method that can handle missing data.

The next section introduces the goal of this paper, why it is important to account for missing data and the composition of the rules used to form explanations.

3 Problem description

To contribute to the research of interpretable Machine Learning, we have set a goal to heighten the interpretability of Cluster Analysis by providing accurate and distinctive explanations for the clusters. We attempt to achieve our goal by considering three different models that are applicable in different scenarios.

In the case that clusters are provided, the clusters are explained by means of the post-hoc model. Are the clusters yet to be formed, then one of the intrinsic models is considered. When the dataset is complete the original intrinsic model of Carrizosa et al. (2023) is employed to search for cluster explanations. Cases with missing values are clustered and explained by virtue of an extended version of the original intrinsic model, referred to as the imputed data intrinsic model.

Performing CA while only having excess to an incomplete dataset is a well-researched topic.

On the other hand, there has been no research on missing data while working with interpretable CA. Imputation seems to be a great solution to deal with missing values while forming clusters. However, the explanations that result from the interpretable CA based on the features with imputed values are less reliable than the explanations based on features that do not contain missing values (Rodwell, 2014). We can not state for certain that the imputed values portray the missing values well, resulting in explanations formed by features from which we can not guarantee they portray reality. To penalize the use of imputed-value-based explanations and stimulate the selection of explanations based on observed data, the imputed data intrinsic model is introduced in this paper.

3.1 Explanation assessment

The quality of the explanations is assessed based on their accuracy and their distinctiveness. The accuracy is measured as the number of true positive cases in a cluster divided by the number of individuals assigned to the cluster, also called the true positive rate (TPR). When an individual in cluster k meets the requirements of the explanation of cluster k, e_k, it is counted as a true positive case. A false positive case is added when an individual outside cluster k is explained by e_k. The distinctiveness is equal to the number of false positive cases for a cluster divided by the number of individuals outside of the cluster, referred to as the false positive rate (FPR). We desire a TPR close to 1 and a FPR close to 0.

We denote the number of clusters by K. For each cluster k holds that $k \in \{1, ..., K\}$. The number of clusters within a dataset is set equal to the number of predetermined clusters. The post-hoc model explains these clusters. The individuals are thus divided into K different groups, such that $J = \cup_{k=1}^{K} J_k$ for which holds that $J_k \cap J_{k'} = \emptyset$ where $k \neq k'$ and J represents the set of individuals.

3.2 Rule formulation

The number of rules by which the explanations of clusters are constructed can be determined in multiple ways. For continuous features, the level of granularity of the threshold determines the number of rules. Each threshold has two corresponding rules. For a continuous feature s, the rules are $feature_s \leq threshold$ and $feature_s > threshold$. A binary feature s has rules $feature_s = 0$ and $feature_s = 1$. We refer to these rules as if-then rules, meaning that if the rule n is selected for a cluster k and individual i satisfies n then we want individual i to belong to cluster k.

For the highest level of granularity, all possible values of the threshold are considered based on the values the features hold within the datasets. Because of the abundance of rules, similar explanations lead to the same level of accuracy and distinctiveness. A less granular case is to consider only the deciles. With the deciles, the rules for each feature can be ascertained for which the dataset is divided into 10 ranked groups of equal size, i.e. if the first decile is selected then 10% of the dataset is below the corresponding threshold of a feature.

Both proposed ways to form rules will be applied to the post-hoc model and to the intrinsic model, the results of which can be found in Section 5. For the imputed data intrinsic model, only the decile thresholds are utilized.

The if-then rules that can be formed by means of the features are stored in a collection A. Collection A consists of N rules that can be split into S number of subgroups, A_s, for which $A = \bigcup_{s=1}^{S} A_s$ and $s \in \{1, ..., S\}$. Each subgroup contains the if-then rules related to one specific observed feature and has the property $A_s \bigcap A_{s'} = \emptyset$ where $s \neq s'$. From A, a maximum of l rules may be selected and joined by an AND operator to form an explanation e_k for a cluster k. When more than one rule is selected, each rule must originate from different subgroups of A. In the models, we denote a rule by n for which $n \in \{1, ..., N\}$. An individual is either denoted by i or j. The total amount of individuals is denoted by I, this indicates that $i, j \in \{1, ..., I\}$.

4 Methodology

In this section, we present the model that provides an explanation for predetermined clusters. The original model by Carrizosa et al. (2023) that builds and explains clusters concurrently is discussed next. Lastly, we formulate how this model can be expanded such that it accounts for imputed values while selecting the explanations.

4.1 Post-hoc model

This section introduces the notation needed for the post-hoc model. The post-hoc model explains the clusters that have been formed beforehand. For each individual in the clusters, it can be extracted from the data if the individual is explained by a rule n of feature s. We denote this using b_{isn} for an individual i, see the definition below.

$$b_{isn} = \begin{cases} 1, & \text{if rule } n \in A_s \text{ explains individual } i \\ 0, & \text{otherwise.} \end{cases}$$

For the model, the binary decision variable γ_{ki} is defined. When an individual i belongs to cluster k, γ_{ki} is one if i complies with the explanation of k, otherwise it equals zero. This definition makes counting the true positive cases possible. In the case of $\gamma_{k'i}$ where $k \neq k'$, the decision variable equals one if i complies with the explanation of cluster k', otherwise zero. This interpretation can provide the number of false positive cases. In addition, the decision variables z_{ksn} are introduced to track the rule selection.

$$z_{ksn} = \begin{cases} 1, & \text{if rule } n \in A_s \text{ is selected to explain cluster } k \\ 0, & \text{otherwise} \end{cases}$$

The model below is the post-hoc MILP model that interprets the clusters J_k for $k = 1, ..., K$. It selects rule-based explanations by combining maximum l rules of the groups A_s, $s = 1, ..., S$. The model includes NK binary decision variables, IK continuous decision variables and $S(K + I) + 2K + I$ constraints.

$$min_{z,\gamma} - \sum_{k=1}^{K}\sum_{i\in J_k}\gamma_{ki} + \theta\sum_{k=1}^{K}\sum_{k'=1,k\neq k'}^{K}\sum_{i\in J_{k'}}\gamma_{ki} \tag{1}$$

$$\text{s.t.} \quad \sum_{n\in A_s} z_{ksn} \leq 1, \qquad k = 1,...,K; s = 1,...,S \tag{2}$$

$$1 \leq \sum_{s=1}^{S}\sum_{n\in A_s} z_{ksn} \leq l, \qquad k = 1,...,K \tag{3}$$

$$\gamma_{ki} + \sum_{n\in A_s}(1-b_{isn})z_{ksn} \leq 1, \qquad i \in J_k; k = 1,...,K; s = 1,...,S \tag{4}$$

$$\gamma_{ki} + \sum_{s=1}^{S}\sum_{n\in A_s}(1-b_{isn})z_{ksn} \geq 1, \qquad i \in J_{k'}; k,k' = 1,...,K; k \neq k' \tag{5}$$

$$z_{ksn} \in \{0,1\}, \qquad s = 1,...,S; n \in A_s; k = 1,...,K \tag{6}$$

$$\gamma_{ki} \in [0,1], \qquad i = 1,...,I; k = 1,...,K. \tag{7}$$

The objective 1 of the above model maximizes the accuracy and distinctiveness of the explanations selected for each cluster by maximizing the number of true positive cases and minimizing the weighted number of false positive cases respectively. The false positive cases are weighted by a parameter $\theta \geq 0$ which can be altered. A higher value of θ increases the importance of a low FPR.

The first Constraints 2 ensure that at most one rule is selected from each A_s, $s = 1,...,S$, for the explanation of a cluster. The restriction that at most l rules are allowed to be combined into an explanation is secured by Constraints 3. These constraints concurrently guarantee that the explanation of each cluster contains at least one rule.

Constraints 4 and 5 define the lower and upper bound of γ_{ki}. First, we look at the case where an individual i that is assigned to cluster k is not explained by the rules selected for k, γ_{ki} should be equal to zero. For a rule n selected from A_s it holds that z_{ksn} =1 but $b_{isn} = 0$. Recognise that $\sum_{n\in A_s}(1-b_{isn})z_{ksn} \leq 1$, since at maximum only one rule n of A_s is selected to explain cluster k. In this case $\sum_{n\in A_s}(1-b_{isn})z_{ksn} = 1$, therefore $\gamma_{ki} \leq 0$. Since γ_{ki} must be non-negative, γ_{ki} equals zero.

In the case that individual i is not a member of cluster k but does fit the explanation of k, we want γ_{ki} to equal one. For each rule $n \in A_s$ for all groups $s = 1,...,S$ that is assigned to the explanation of cluster k, it holds that $z_{ksn} = 1$ and $b_{isn} = 1$. As a consequence, γ_{ki} must be larger or equal to one. Following from the fact that γ_{ki} is smaller or equal to one, γ_{ki} is set to one.

Lastly, the binary characteristic of z_{ksn} is ensured by Constraints 6 and the assumption that γ_{ki} is a continuous variable between the values zero and one is established by Constraints 7. The relaxation of γ_{ki} from a binary variable to a continuous variable does not lead to any loss of optimization because of the formulation of the objective function and the constraints.

4.2 Intrinsic model

In this section, we introduce the notation for the intrinsic model related to the clusters, the individuals and the dissimilarities between these individuals as defined by Carrizosa et al. (2023).

Before the intrinsic model can be applied, the dissimilarities between data points must be determined. Carrizosa et al. (2023) used the Euclidean distance as a measure of dissimilarity. The Euclidean distance formula is $\delta_{ij} = \sqrt{\sum_{s=1}^{S}(i_s - j_s)^2}$, where i and j denote the indices of two different observations that contain the values of S features. The matrix Δ is a matrix of each dissimilarity between all possible combinations of individuals.

The use of the Euclidean distance may lead to undesired outcomes. Namely, the largest-scaled feature can dominate the other features creating a skewed result. Therefore, it is important that the attributes are normalized before the distances are calculated. In this paper, we apply the min-max normalization which transforms every feature value to a value between zero and one.

In Table 1, we introduce the definitions of the decision variables and the parameters in the intrinsic model related to the rules that explain each cluster, the individuals that need to be assigned to a cluster, and the true positive and false positive cases.

Table 1: Definitions of decision variables and parameters denoted in the intrinsic model

Decision variables	
$x_{ki} = \begin{cases} 1, & \text{if individual } i \text{ is assigned to cluster } k \\ 0, & \text{otherwise} \end{cases}$	
$\alpha_i = \begin{cases} 1, & \text{if the explanation selected for the cluster of individual } i \text{ is a true positive case for } i \\ 0, & \text{otherwise} \end{cases}$	
$\beta_{ki} = \begin{cases} 1, & \text{if the explanation selected for cluster } k \text{ is a false positive case} \\ & \text{for individual } i \text{ and } i \text{ is not assigned to } k \\ 0, & \text{otherwise} \end{cases}$	
Parameters	
$\theta_1 \geq 0$	Weight assigned to the true positive cases across all clusters
$\theta_2 \geq 0$	Weight assigned to the false positive cases across all clusters
l	Maximum number of rules that can form a cluster's explanation

Below, we present the IP formulation to assign the individuals to the K clusters utilizing the dissimilarity matrix Δ while simultaneously determining a rule-based explanation consisting of maximum l if-then rules from A_s, $s = 1, \ldots S$, for each cluster. The model can be a MILP problem by introducing a decision variable y_{kij}. We define y_{kij} as the product of two x variables, $y_{kij} = x_{ki}x_{kj}$, which equals one when both individuals i and j are assigned to cluster k, otherwise zero.

$$min_{x,z,\alpha,\beta,y} \sum_{k=1}^{K} \sum_{i=1}^{I-1} \sum_{j=i+1}^{I} \delta_{ij} y_{kij} - \theta_1 \sum_{i=1}^{I} \alpha_i + \theta_2 \sum_{k=1}^{K} \sum_{i=1}^{I} \beta_{ki} \tag{8}$$

$$\text{s.t.} \quad \sum_{k=1}^{K} x_{ki} = 1, \qquad i = 1, ..., I \tag{9}$$

$$\sum_{n \in A_s} z_{ksn} \leq 1, \qquad k = 1, ..., K; s = 1, ..., S \tag{10}$$

$$1 \leq \sum_{s=1}^{S} \sum_{n \in A_s} z_{ksn} \leq l, \qquad k = 1, ..., K \tag{11}$$

$$\alpha_i + x_{ki} + \sum_{n \in A_s} (1 - b_{isn}) z_{ksn} \leq 2, \qquad i = 1, ..., I; k = 1, ..., K; s = 1, ..., S \tag{12}$$

$$\beta_{ki} + x_{ki} + \sum_{s=1}^{S} \sum_{n \in A_s} (1 - b_{isn}) z_{ksn} \geq 1, \qquad i = 1, ..., I; k = 1, ..., K \tag{13}$$

$$x_{ki} + x_{kj} - y_{kij} \leq 1, \qquad i = 1, ..., I-1; j = i+1, ..., I; k = 1, ..., K \tag{14}$$

$$x_{ki} \in \{0, 1\}, \qquad i = 1, ..., I; k = 1, ..., K \tag{15}$$

$$z_{ksn} \in \{0, 1\}, \qquad s = 1, ..., S; n \in A_s; k = 1, ..., K \tag{16}$$

$$y_{kij} \in [0, 1], \qquad i = 1, ..., I-1; j = i+1, ..., I; k = 1, ..., K \tag{17}$$

$$\alpha_i \in [0, 1], \qquad i = 1, ..., I \tag{18}$$

$$\beta_{ki} \in [0, 1], \qquad i = 1, ..., I; k = 1, ..., K. \tag{19}$$

The objective 8 is composed of three terms. The first term minimizes the intra-homogeneity (IH) of clusters in the form of summating the dissimilarities within clusters. The second maximizes the total number of true positive cases with an assigned weight θ_1. The third term minimizes the total number of false positive cases with an assigned weight θ_2. The solution must comply with the following constraints. Firstly, Constraints 9 verify that each individual is allocated to precisely one cluster. Constraints 10 and 11 are identical to Constraints 2 and 3. Because of the way that the objective is formulated, we only need to include the definition of $\alpha_i = 0$ and $\beta_{ki} = 1$, as explained by Carrizosa et al. (2023). To ensure that these values of α_i and β_{ki} are well-defined, the Constraints 12 and 13 are included in the formulation. They define α_i and β_{ki} is a similar way as Constraints 4 and 5 define γ_{ki}. The decision variables x and z are binary variables, which are imposed by Constraints 15 and 16. Decision variables α_i and β_{ki} are defined as binary variables as well, however, because of the formulation of the model we can presume the variables as continuous without loss of optimality (Constraints 18 and 19).

The intrinsic model has $K(I+N)$ binary and $I(\frac{K(I-1)}{2}+K+1)$ continuous decision variables between the values zero and one. The total number of linear constraints is $I+K(\frac{I(I-1)}{2}+IS+I+S+2)$.

4.3 Imputed Data Intrinsic model

The intrinsic model does not consider the possibility that there are missing values. If there are missing values these values could be replaced by imputed values, for example taking the form of the mean of the values of other observations. However, explaining the clusters based on the feature for which there is a high percentage of imputed values would not be reliable since the explanation would not be based on observed data. The use of rules about features with imputed values should be penalized.

To penalize the use of rules about features with imputed values, we introduce a new matrix M with length I and width S for which each m_{is} equals one if individual i had a missing value for feature s which is replaced by an imputed value and zero otherwise. We multiply m_{is} with z_{ksn} and x_{ki} for every possible combination of clusters, individuals and rules and then add up the multiplications. This summation is multiplied with a parameter λ and added to the objective. The objective takes the following form.

$$min_{x,z,\alpha,\beta} \sum_{k=1}^{K} \sum_{i=1}^{I-1} \sum_{j=i+1}^{I} \delta_{ij} y_{kij} - \theta_1 \sum_{i=1}^{I} \alpha_i + \theta_2 \sum_{k=1}^{K} \sum_{i=1}^{I} \beta_{ki} + \lambda \sum_{k=1}^{K} \sum_{i=1}^{I} \sum_{s}^{S} \sum_{n \in A_s} m_{is} z_{ksn} x_{ki} \tag{20}$$

The added term is defined as the number of individuals that are appointed to a cluster whose explanation contains a rule based on a feature for which the individual has an imputed value. We want to minimize this term such that the number of individuals explained solely based on their non-imputed values is maximized. Because of the formulation of the added term, the selection of specific feature rules is penalized more strictly when the percentage of individuals in the cluster with imputed values for the corresponding feature is higher.

To make the model a linear problem again, we introduce decision variable $w_{ksni} = z_{ksn} x_{ki}$. The definition of w_{ksni} makes it a binary decision variable which can be relaxed to a continuous decision variable between 0 and 1 without loss of optimality (Constraints 22). With the new decision variable also comes new constraints, Constraints 21 such that $w_{ksni} = 1$ is well-defined. The imputed data intrinsic model is formed by Objective 20 and the Constraints (9) - (19), (21) and (22).

$$z_{ksn} + x_{ki} - w_{ksni} \leq 1, \quad k = 1, ..., K; s = 1, ..., S; n \in A_s; i = 1, ..., I \tag{21}$$

$$w_{ksni} \in [0, 1], \quad k = 1, ..., K; s = 1, ..., S; n \in A_s; i = 1, ..., I \tag{22}$$

When λ equals 1, the model prefers to select explanations that are based on observed data unless the rules including imputed features provide a significantly better explanation. In the case of λ set to 0, the model reduces to the original intrinsic model.

This extended intrinsic model has $K(I+N)$ binary and $I(\frac{K(I-1)}{2} + K + KN + 1)$ continuous decision variables between the values zero and one. The total number of linear constraints is $I + K(\frac{I(I-1)}{2} + IS + NI + I + S + 2)$.

5 Results

For the results, we first discuss the data used to apply the different models. Next, the results of the post-hoc model are shown and are followed by the results of the intrinsic model. Next, we discuss how the missing values are generated and imputed. Lastly, the outcome of the imputed data intrinsic model is presented.

5.1 The data

We apply the original intrinsic and the post-hoc model to two datasets (Carrizosa et al., 2023). All utilized datasets contain classes formed based on the class definitions which can be found in Tables 9 to 12 in Appendix B. These tables also provide the definitions of the features. For the intrinsic model, we ignore the predetermined classes to create our own while the post-hoc model explains the predetermined classes. The two datasets, the housing and breast cancer dataset, each contain 2 classes.

For the imputed data intrinsic model, two other datasets are utilized. Since the imputed data intrinsic model has even more decision variables and constraints than the other two models, this model is more difficult to solve. The extended intrinsic model is tested using more compact datasets with fewer observations and features than the datasets used for the models used by Carrizosa et al. (2023). We refer to these datasets as the cryotherapy and forest fires datasets. Both comprise two classes.

Further description of the datasets is denoted in Table 2. Table 2 contains information on each dataset on the number of observations, the number of predetermined classes and the number of features that are observed. All datasets are available on the UCI repository.

The housing dataset will be used throughout this paper as an example. The data points correspond to owner-occupied homes in the area of Boston Mass collected by the US Census Service. The homes are divided into two classes, namely whether or not the home is worth more than the median value of owner-occupied homes.

Table 2: Information on the datasets

Name of dataset	# Observations	# Classes	# Features
housing	506	2	13
breast cancer	683	2	10
cryotherapy	90	2	5
forest fires	122	2	10

5.2 Post-hoc model results

The post-hoc model is solved for multiple values of θ. We alter the parameter $\theta \geq 0$ between the values 2^p with $p \in \{-5, -4, ..., 4, 5\}$. The post-hoc model is first solved for the case where θ equals 2^{-5} without an initial solution. The model is then applied in increasing order of θ where the initial solution is set equal to the solution of the model where θ equals the previous value.

The results of the post-hoc model for the housing dataset can be found in Tables 3 and 4. For the case where the rules are formed using the decile values, the results differ from the results found by Carrizosa et al. (2023). The most significant cause is the difference in predetermined classes. Carrizosa et al. (2023) have found clusters that split the observations into a cluster containing 274 houses and one formed by 232 houses. The split we find has 255 observations in the first cluster and 251 in the second. The different clusters lead to different explanations and therefore divergent TPRs and FPRs. The paper of Carrizosa et al. (2023) will also be called the reference paper.

Table 3: The clusters and explanations derived from the post-hoc model for the housing dataset constructed with K = 2 clusters, a maximum length of $l = 2$, utilizing N = 189 rules generated from the deciles of the continuous features and considering all attributes of the categorical features, along with the TPR and FPR found in the reference paper.

$\boldsymbol{\theta}$	**C**	**TPR**	**FPR**	**Ref TPR**	**Ref FPR**	**Explanations**
2^5	1	0.27	0.00	0.14	0.00	CRIM>5.581 AND NOX>0.668
	2	0.48	0.00	0.45	0.00	RM>6.376 AND LSTAT≤7.765
2^4	1	0.45	0.01	0.14	0.00	DIS≤2.6403 AND LSTAT>15.62
	2	0.64	0.01	0.59	0.01	RM>6.2085 AND LSTAT≤9.53
2^3	1	0.55	0.02	0.14	0.00	PTRATIO>19.7 AND LSTAT>13.33
	2	0.64	0.01	0.59	0.01	RM>6.2085 AND LSTAT≤9.53
2^2	1	0.66	0.04	0.41	0.05	TAX>289 AND LSTAT>13.33
	2	0.64	0.01	0.59	0.01	RM>6.2085 AND LSTAT≤9.53
2^1	1	0.66	0.04	0.70	0.15	TAX>289 AND LSTAT>13.33
	2	0.76	0.05	0.70	0.06	RM>5.9505 AND LSTAT≤11.36
2^0	1	0.83	0.16	0.70	0.06	INDUS>2.91 AND LSTAT>11.36
	2	0.86	0.15	0.81	0.23	RM>5.9505 AND LSTAT≤13.33
2^{-1}	1	0.99	0.42	0.78	0.18	LSTAT>7.765
	2	0.86	0.15	0.97	0.40	RM>5.9505 AND LSTAT≤13.33
2^{-2}	1	0.99	0.42	0.99	0.46	LSTAT>7.765
	2	0.96	0.43	0.98	0.83	TAX≤66 AND LSTAT≤15.62
2^{-3}	1	0.99	0.42	0.99	0.46	LSTAT>7.765
	2	0.98	0.61	0.98	0.83	TAX≤666 AND LSTAT≤18.06
2^{-4}	1	0.99	0.42	0.99	0.46	LSTAT>7.765
	2	1.00	0.79	1.00	1.00	TAX≤666 AND LSTAT≤23.035
2^{-5}	1	1.00	0.60	1.00	0.63	LSTAT>6.29
	2	1.00	0.79	1.00	1.00	TAX≤666 AND LSTAT≤23.035

From Table 3, we denote that the TPR ranges from 0.27 to 1 and the FPR takes forms between values 0 and 0.79. The table depicts the same trend in the true and false positive ratios as found by Carrizosa et al. (2023). For the lowest value of θ the TPR and the FPR are the highest. When the value of θ increases, the TPR and FPR decline. This indicates that the best combination of TPR and FPR, and therefore the best explanations, can be found for θ values around 0.5, 1 and 2.

The instance with rules formed by all unique values of the features portrays the same correlation between the value of θ and the true and false positive ratios. The result most optimal is that of the central values for θ. Modest improvement in the TPR and FPR can be noticed for the more granular case compared to the model with decile rules. The lowest TPR is now 0.51 instead of 0.27 and the highest FPR equals 0.69 rather than 0.79. In addition, the central θ values results improve when the granularity increases.

The result for the breast cancer dataset is denoted in Table 13 in Appendix C. The breast cancer dataset is used only in the case of decile thresholds. The TPR ranges from 0.68 to 1, while the FPR equals values between 0 and 0.52. The obtained output deviates far less from the Carrizosa et al. (2023) results than that of the housing dataset. The few differences that occur are dissimilarities in explanations, meaning that the explanations observed in this paper

Table 4: The clusters and explanations derived from the post-hoc model for the housing dataset constructed with K = 2 clusters, a maximum length of $l = 2$, utilizing N = 5646 rules generated from the unique values of the continuous features and considering all attributes of the categorical features, along with the TPR and FPR found in the reference paper.

$\boldsymbol{\theta}$	C	TPR	FPR	Ref TPR	Ref FPR	Explanations
2^5	1	0.51	0.01	0.51	0.00	TAX>402 AND LSTAT>14.37
	2	0.52	0.00	0.14	0.00	NOX≤0.51 AND RM>6.279
2^4	1	0.51	0.01	0.14	0.00	TAX>402 AND LSTAT>14.36
	2	0.69	0.01	0.58	0.00	RM>6.144 AND LSTAT≤
2^3	1	0.52	0.01	0.14	0.00	TAX>AND LSTAT>14.1
	2	0.69	0.01	0.64	0.01	RM>6.144 AND LSTAT≤9.93
2^2	1	0.66	0.04	0.45	0.05	TAX>300 AND LSTAT>13.33
	2	0.70	0.01	0.64	0.01	RM>6.12 AND LSTAT≤9.93
2^1	1	0.65	0.04	0.70	0.04	TAX>305 AND LSTAT>13.33
	2	0.79	0.05	0.70	0.14	RM>6.059 AND LSTAT≤11.66
2^0	1	0.81	0.13	0.80	0.20	INDUS>4.86 AND LSTAT>11.66
	2	0.79	0.05	0.73	0.06	RM>6.059 AND LSTAT≤11.66
2^{-1}	1	1.00	0.40	0.99	0.44	PTRATIO>14.4 AND LSTAT>7.67
	2	0.86	0.15	0.78	0.19	RM>5.957 AND LSTAT≤13.27
2^{-2}	1	0.99	0.40	0.99	0.44	PTRATIO>13 AND LSTAT>7.67
	2	0.96	0.38	0.98	0.80	B>127.36 AND LSTAT≤14.81
2^{-3}	1	0.99	0.40	0.99	0.44	PTRATIO>13 AND LSTAT>7.67
	2	0.99	0.63	1.00	0.90	B>127.36 AND LSTAT≤19.78
2^{-4}	1	0.99	0.40	0.99	0.44	PTRATIO>13 AND LSTAT>7.67
	2	1.00	0.69	1.00	0.97	B>127.36 AND LSTAT≤21.52
2^{-5}	1	1.00	0.49	1.00	0.53	PTRATIO>13 AND LSTAT>6.73
	2	1.00	0.78	1.00	0.97	CRIM≤14.4383 AND B>127.36

and that of Carrizosa et al. (2023) have identical true and false positive rates. Since the IH is identical, the results are equally as good.

5.3 Intrinsic model results

We apply the intrinsic model with different values for the two parameters θ_1 and θ_2. The values of θ_1 and θ_2 are set to combinations of $\{0.5, 1, 2\}$ to investigate which combination gives the highest number of true positives and the lowest number of false positives while maintaining an acceptable dissimilarity level. The first combination is θ_1 and θ_2 equal 0.5. After that, θ_2 is adjusted until it has taken all forms of $\{0.5, 1, 2\}$. The next step is to vary θ_1.

The model starts with an initial solution and has a time limit of 10 minutes. For the first combination of the θ parameters, the initial solution of the cluster allocation equals the solution of k-means clustering. The k-means clustering results are used to solve the post-hoc model, providing an initial solution for the rule selection. Further combinations have an initial clustering solution set to the solution of the previous θ combination and the initial explanation selection is equal to the outcome of the post-hoc model with this clustering solution as input.

While replicating the intrinsic model of Carrizosa et al. (2023), the exact same explanations are found for the housing dataset when using deciles as threshold (Table 5). One small distinction is that the TPR ranges from 0.91 to 1 instead of 0.90 to 1, presumably because of a difference in rounding. The FPR has an equally small range between 0 and 0.09.

The results for a higher granularity, denoted in Table 6, have true and false positive rates equal to that of the original authors. The explanations do differ slightly. Since the TPR, FPR and the IH are identical to the original results, it indicates that the explanations found in this paper are equivalent in interpretability.

Important to note is that these results are not proven to be optimal because of the 10-minute time limit. We do see while comparing the cases with decile values and the unique values as rule thresholds that the range of the FPR becomes smaller when the number of rules rises. The TPR takes a value between 1 and 0.91 in both cases, but the FPR range reduces to values between 0 and 0.04 in the case of higher granularity.

The results for the breast cancer dataset in this paper vary greatly from the Carrizosa et al. (2023) results. In the original results, multiple intra-homogeneity values are found. This paper only reports one value for the intra-homogeneity. The magnitude and complexity of the problem of the breast cancer dataset are higher than that of the housing dataset problem. Therefore, another grouping of observations could not be found within the 10-minute time limit. We do observe that the peak values of TPR are again obtained for high values of θ_1 and θ_2.

Table 5: The clusters and explanations derived from the intrinsic model for the housing dataset constructed with K = 2 clusters, a maximum length of $l = 2$, utilizing N = 189 rules generated from the deciles of the continuous features and considering all attributes of the categorical features, along with the TPR, FPR and IH found in the reference paper.

θ_1	θ_2	IH	Ref IH	C	TPR	FPR	Ref TPR	Ref FPR	Explanations
0.5	0.5	$0.6*10^5$	$0.6*10^5$	1	1.00	0.00	1.00	0.00	INDUS>12.83 AND TAX>398
				2	0.97	0.04	0.97	0.04	NOX≤6.5025 AND RAD≤8
0.5	1	$0.6*10^5$	$0.6*10^5$	1	0.91	0.00	0.90	0.00	INDUS>12.83 AND PTRATIO>19.7
				2	0.97	0.00	0.97	0.00	NOX≤0.605 AND RAD≤8
0.5	2	$0.6*10^5$	$0.6*10^5$	1	0.91	0.00	0.90	0.00	INDUS>12.83 AND PTRATIO>19.7
				2	0.97	0.00	0.97	0.00	NOX≤0.605 AND RAD≤8
1	0.5	$0.6*10^5$	$0.6*10^5$	1	1.00	0.04	1.00	0.04	INDUS>12.83 AND TAX>398
				2	1.00	0.09	1.00	0.09	NOX≤0.668 AND TAX≤437
1	1	$0.6*10^5$	$0.6*10^5$	1	1.00	0.04	1.00	0.04	INDUS>12.83 AND PTRATIO>19.7
				2	0.97	0.00	0.97	0.00	NOX≤0.605 AND RAD≤8
1	2	$0.6*10^5$	$0.6*10^5$	1	0.91	0.00	0.90	0.00	INDUS>12.83 AND PTRATIO>19.7
				2	0.97	0.00	0.97	0.00	NOX≤0.605 AND RAD≤8
2	0.5	$0.6*10^5$	$0.6*10^5$	1	1.00	0.04	1.00	0.04	INDUS>12.83 AND TAX>398
				2	1.00	0.09	1.00	0.09	NOX≤0.668 AND TAX≤437
2	1	$0.6*10^5$	$0.6*10^5$	1	1.00	0.04	1.00	0.04	INDUS>12.83 AND TAX>398
				2	1.00	0.09	1.00	0.09	NOX≤0.668 AND TAX≤437
2	2	$0.6*10^5$	$0.6*10^5$	1	1.00	0.04	1.00	0.04	INDUS>12.83 AND TAX>398
				2	0.97	0.00	0.97	0.00	NOX≤0.605 AND RAD≤8

Table 6: The clusters and explanations derived from the intrinsic model for the housing dataset constructed with K = 2 clusters, a maximum length of $l = 2$, utilizing N = 5646 rules generated from the unique values of the continuous features and considering all attributes of the categorical features, along with the TPR, FPR and IH found in the reference paper.

θ_1	θ_2	IH	Ref IH	C	TPR	FPR	Ref TPR	Ref FPR	Explanations
0.5	0.5	0.6*10^5	0.6*10^5	1	1.00	0.04	1.00	0.04	INDUS>15.04 AND RAD>3
				2	1.00	0.00	1.00	0.00	NOX≤0.647 AND TAX≤432
0.5	1	0.6*10^5	0.6*10^5	1	0.91	0.00	0.91	0.00	INDUS>15.04 AND TAX>432
				2	1.00	0.00	1.00	0.00	NOX≤0.647 AND TAX≤432
0.5	2	0.6*10^5	0.6*10^5	1	0.91	0.00	0.91	0.00	INDUS>15.04 AND TAX>422
				2	1.00	0.00	1.00	0.00	NOX≤0.647 AND TAX≤432
1	0.5	0.6*10^5	0.6*10^5	1	1.00	0.04	1.00	0.04	INDUS>15.04 AND TAX>351
				2	1.00	0.00	1.00	0.00	NOX≤0.647 AND TAX≤432
1	1	0.6*10^5	0.6*10^5	1	1.00	0.04	1.00	0.04	INDUS>15.04 AND RAD>3
				2	1.00	0.00	1.00	0.00	NOX≤0.647 AND TAX≤432
1	2	0.6*10^5	0.6*10^5	1	0.91	0.00	0.91	0.00	INDUS>15.04 AND TAX>432
				2	1.00	0.00	1.00	0.00	NOX≤0.647 AND TAX≤432
2	0.5	0.6*10^5	0.6*10^5	1	1.00	0.04	1.00	0.04	INDUS>15.04 AND TAX>198
				2	1.00	0.00	1.00	0.00	NOX≤0.647 AND TAX≤432
2	1	0.6*10^5	0.6*10^5	1	1.00	0.04	1.00	0.04	INDUS>15.04 AND TAX>351
				2	1.00	0.00	1.00	0.00	NOX≤0.647 AND TAX≤432
2	2	0.6*10^5	0.6*10^5	1	1.00	0.04	1.00	0.04	INDUS>15.04 AND RAD>3
				2	1.00	0.00	1.00	0.00	NOX≤0.647 AND TAX≤432

5.4 Generating missing and imputed values

Before we can move on to the imputed data intrinsic model, we need an incomplete dataset. To test the effect of imputed values on the imputed data intrinsic model, 20% of the cryotherapy and forest fires dataset is deleted. We form incomplete datasets by deleting data completely at random in specific features. In both cases, around half of the features are selected. These features are denoted by * in the Tables 11 and 12 which can be found in Appendix B. The newly formed incomplete datasets mimic real-life datasets for which, for example, the measuring machinery malfunctioned.

The missing values are then filled by the mean of the remaining values of the corresponding feature. We use mean imputation because it is easy and fast but most importantly reliable. The datasets made up of imputed values are now ready to function as the input data of the extended intrinsic model.

5.5 Imputed data intrinsic model

In this section, we compare the results for the extended intrinsic model of the instance where λ equals one and the case in which λ is set to zero. The value of λ is equal to one when we want to penalize the use of features with imputed values for the explanations. When discussing the results we refer to this case as the imputed data intrinsic model. We set λ to zero when we want to reduce the problem to the original intrinsic model. The values of parameters θ_1 and θ_2 are varied in the same fashion as for the standard intrinsic model.

We report the best feasible solution found within the time limits including the IH, TPR and FPR of the imputed dataset and the selected explanations. In addition, the IH, TPR and FPR for the complete dataset are determined based on the clusters and explanations found for the imputed dataset.

The initial solutions of the imputed intrinsic model are set similarly to the initial solutions of the original intrinsic model. The initial solution is set to the original intrinsic model's clustering solution of the corresponding θ combination. The initial explanations are then the explanations for the clustering solution selected by the post-hoc model. The first initial clustering solution is determined by the k-means clustering outcome.

5.5.1 Cryotherapy results

The time limit for the cryotherapy dataset is set to 10 minutes and the results are portrayed in Table 7. The features that contain imputed values are Time, NumbWarts and Area (Table 11) of which Area is used multiple times to explain clusters when the original intrinsic model is applied. This is just over half of the number of features in the dataset. When λ is equal to 1, the IH values remain unchanged. However, the explanations that contain Area in the case of the original intrinsic model are altered when accounting for imputed values. The explanations exclude the Area rules and maintain the remaining rules. Despite these changes, the TPR and FPR of these new explanations retain the same value, indicating that the explanations are equally as accurate and distinctive as the explanations in the case of λ is zero.

One combination of θ_1 and θ_2 reports different explanations for both λ values that are not related to the imputed data. For the θ combination where both parameters are assigned value 0.5, the clustering solution and therefore the explanations differ in the two cases of λ. This is likely caused by the increase in the model's complexity when the λ parameter is fixed at 1. Despite the complexity increase, the time limit remains of equal length in both cases of λ causing the solutions to differ for this particular θ combination. Because of this divergent clustering solution, the range of the TPR varies. For λ equal to zero, the TPR ranges from 0.97 to 1 and the FPR has a value of 0 or 0.03. Changing λ to one results in TPR values between 0.91 and 1. The FPR range remains the same. The TPR and FPR values in both cases are identical for the imputed and complete datasets.

5.5.2 Forest fires results

The forest fires dataset contains twice the amount of features as that of the cryotherapy dataset, increasing the complexity of the problem. The time limit is therefore set to 20 minutes. Again half of the features have generated missing values that are imputed by the mean, namely the features FFMC, DMC, DC, ISI and BUI (Table 12). These features are determined by a Fire Weather Index (FWI) system. This means the imputed dataset illustrates a situation where the FWI system malfunctions.

Without taking the imputed values into account while selecting the explanations, the explanations contain rules concerning features with imputed values in more than half of the results for

Table 7: The clusters and explanations derived from the imputed data intrinsic model for the cryotherapy dataset constructed with K = 2 clusters, a maximum length of l = 2, utilizing N = 62 rules generated from the deciles of the continuous features and considering all attributes of the categorical features, along with the intra-homogeneity, TPR and FPR values for the complete dataset.

$\boldsymbol{\theta_1}$	$\boldsymbol{\theta_2}$	$\boldsymbol{\lambda}$	**IH**	**Actual IH**	**C**	**TPR**	**FPR**	**Actual TPR**	**Actual FPR**	**Explanation**
0.5	0.5	0	$1.12*10^3$	$1.39*10^3$	1	0.97	0.00	0.97	0.00	Type>1.4
					2	1.00	0.03	1.00	0.03	Type≤1.4
		1	$1.16*10^3$	$1.43*10^3$	1	0.91	0.00	0.91	0.00	Age>21.6 AND Type>1.4
					2	0.95	0.00	0.95	0.00	Type≤1.4
0.5	1	0	$1.12*10^3$	$1.39*10^3$	1	0.97	0.00	0.97	0.00	Type>1.4
					2	0.98	0.00	0.98	0.00	Age≤41.9 AND Type≤1.4
		1	$1.12*10^3$	$1.39*10^3$	1	0.97	0.00	0.97	0.00	Type>1.4
					2	0.98	0.00	0.98	0.00	Age≤41.9 AND Type≤1.4
0.5	2	0	$1.12*10^3$	$1.39*10^3$	1	0.97	0.00	0.97	0.00	Type>1.4
					2	0.98	0.00	0.98	0.00	Age≤41.9 AND Type≤1.4
		1	$1.12*10^3$	$1.39*10^3$	1	0.97	0.00	0.97	0.00	Type>1.4
					2	0.98	0.00	0.98	0.00	Age≤41.9 AND Type≤1.4
1	0.5	0	$1.12*10^3$	$1.39*10^3$	1	0.97	0.00	0.97	0.00	Area>9.8 AND Type>1.4
					2	1.00	0.03	1.00	0.03	Type≤1.4
		1	$1.12*10^3$	$1.39*10^3$	1	0.97	0.00	0.97	0.00	Type>1.4
					2	1.00	0.03	1.00	0.03	Type≤1.4
1	1	0	$1.12*10^3$	$1.39*10^3$	1	0.97	0.00	0.97	0.00	Type>1.4
					2	1.00	0.03	1.00	0.03	Type≤1.4
		1	$1.12*10^3$	$1.39*10^3$	1	0.97	0.00	0.97	0.00	Type>1.4
					2	1.00	0.03	1.00	0.03	Type≤1.4
1	2	0	$1.12*10^3$	$1.39*10^3$	1	0.97	0.00	0.97	0.00	Type>1.4
					2	0.98	0.00	0.98	0.00	Age≤41.9 AND Type≤1.4
		1	$1.12*10^3$	$1.39*10^3$	1	0.97	0.00	0.97	0.00	Type>1.4
					2	0.98	0.00	0.98	0.00	Age≤41.9 AND Type≤1.4
2	0.5	0	$1.12*10^3$	$1.39*10^3$	1	0.97	0.00	0.97	0.00	Area>9.8 AND Type>1.4
					2	1.00	0.03	1.00	0.03	Type≤1.4
		1	$1.12*10^3$	$1.39*10^3$	1	0.97	0.00	0.97	0.00	Type>1.4
					2	1.00	0.03	1.00	0.03	Type≤1.4
2	1	0	$1.12*10^3$	$1.39*10^3$	1	0.97	0.00	0.97	0.00	Area>9.8 AND Type>1.4
					2	1.00	0.03	1.00	0.03	Type≤1.4
		1	$1.12*10^3$	$1.39*10^3$	1	0.97	0.00	0.97	0.00	Type>1.4
					2	1.00	0.03	1.00	0.03	Type≤1.4
2	2	0	$1.12*10^3$	$1.39*10^3$	1	0.97	0.00	0.97	0.00	Type>1.4
					2	1.00	0.03	1.00	0.03	Type≤1.4
		1	$1.12*10^3$	$1.39*10^3$	1	0.97	0.00	0.97	0.00	Type>1.4
					2	1.00	0.03	1.00	0.03	Type≤1.4

all θ combinations. It was less than half of the cases for the cryotherapy dataset. Therefore, changing the λ value should have a bigger effect on the explanations while using the forest fires dataset. The results are represented in Table 15 in Appendix D.

For the original intrinsic model, the intra-homogeneity remains constant for all combinations of the θ_1 and θ_2 parameter values, indicating that the clustering performance remains constant. The true positive rates take on values between 0.16 and 0.98. The FPR range from 0.03 to 0.31. The TPR and FPR for the complete dataset differ for the forest fires dataset from the TPR and FPR for the imputed dataset. Nonetheless, the ranges remain consistent across both datasets.

The TPR ranges from 0.82 to 1 in the case of λ equal to 1. For the false positive rates, the range is 0.03 to 0.52. In the imputed data intrinsic model case, the ranges of the TPR and FPR for both the imputed dataset and the complete dataset also remain identical. Nevertheless, the specific values of TPR and FPR differ from the actual TPR and FPR for certain combinations of the θ parameters. In addition, the imputed data intrinsic model provides better TPR scores, but also worse FPR scores compared to the original intrinsic model.

The IH values do change in some instances for the imputed data intrinsic model. For these instances, the model that accounts for features with imputed values finds lower IH values than the model that does not account for these features. Unfortunately, a few of the corresponding false positive rates rise and true positive rate values decline. It is difficult to conclude from these results which model finds the better solution. To make the comparison easier, the objective of the original intrinsic model (Equation 8) is calculated for each solution. The values are denoted in Table 8 and referred to as the 'Obj value'. The solution that has the lowest value for this summation is the better option since we want a high IH, high TPR and low FPR.

In the case that θ_1 and θ_2 both equal 0.5, changing the value of λ results in dissimilar solutions. In Table 8, we see that for the imputed data intrinsic model, the solution has a lower objective value for both the imputed data and the complete data compared to the original intrinsic model. The imputed data intrinsic model again finds a better solution for θ_1 equal to 1 and θ_2 equal to 0.5. Both θ combinations mentioned result in explanations including features with imputed values when λ is set to 0. If λ is equal to 1, the explanations do not contain these features.

For a few other results of θ combinations varying λ does not impact the solution for these combinations. This indicates that, for these cases, the imputed data intrinsic model can not find a solution that does not contain rules based on features with imputed values rules without providing a significantly higher objective value.

Table 8: The original objective functions' values for the solutions obtained from the extended intrinsic model using imputed data examined across all combinations of λ, θ_1, and θ_2 parameters in relation to the forest fires dataset calculated for the imputed and complete datasets.

$\boldsymbol{\theta_1}$	$\boldsymbol{\theta_2}$	$\lambda = 0$		$\lambda = 1$	
		Obj value	**Actual Obj value**	**Obj value**	**Actual Obj value**
0.5	0.5	$2.60*10^3$	$3.71*10^3$	$2.59*10^3$	$3.64*10^3$
0.5	1	$2.58*10^3$	$3.68*10^3$	$2.58*10^3$	$3.68*10^3$
0.5	2	$2.58*10^3$	$3.68*10^3$	$2.58*10^3$	$3.68*10^3$
1	0.5	$2.52*10^3$	$3.60*10^3$	$2.51*10^3$	$3.60*10^3$
1	1	$2.52*10^3$	$3.62*10^3$	$2.52*10^3$	$3.62*10^3$
1	2	$2.53*10^3$	$3.63*10^3$	$2.53*10^3$	$3.63*10^3$
2	0.5	$2.40*10^3$	$3.50*10^3$	$2.40*10^3$	$3.50*10^3$
2	1	$2.41*10^3$	$3.54*10^3$	$2.41*10^3$	$3.54*10^3$
2	2	$2.42*10^3$	$3.52*10^3$	$2.42*10^3$	$3.52*10^3$

6 Conclusion

In this paper, we replicate the MILP model created by Carrizosa et al. (2023) that provides explanations for predetermined clusters based on the features of the observations. In addition, their MILP formulation that simultaneously assigns observations to clusters and selects corresponding explanations is validated. The Euclidean distance portrays the dissimilarity between observations. Furthermore, we form rules based on the features that characterize the individuals within a dataset. The rules are joined by an AND operator to form explanations for the clusters. Lastly, the availability of a complete dataset is assumed.

The goal of both models is to maximize the accuracy of the explanations by maximizing the number of true positive cases and to maximize the distinctiveness of the explanations by minimizing the number of false positive cases. The intrinsic model has an additional goal to minimize the dissimilarity between individuals that belong to the same cluster. We validate in this paper that the two MILP models do indeed work and can provide accurate and distinctive explanations.

On top of the MILP models formulated by Carrizosa et al. (2023), a new MILP model is introduced. The new model is an extension of the intrinsic model. The extended model, referred to as the imputed data intrinsic model, penalizes the selection of rules based on features with imputed values for cluster explanations. When the assumption of availability to a complete dataset is violated, researchers will resort to utilizing an incomplete dataset. The missing values contained in the incomplete dataset are imputed or the observations containing missing values are deleted. Explanations formed by rules based on features that contain imputed values will be less reliable than explanations consisting of rules based on features that only contain observed data since it can not be stated for certain that the rules of features with imputed values represent reality. Our goal is to find explanations that portray the real world the best by searching for more or as accurate and distinctive explanations using only features consisting of observed data as the explanations found that include all features.

We prove that the imputed data intrinsic model provides as accurate and distinctive explanations as the model that does not account for imputed values in almost all cases. In some cases, the imputed data intrinsic model resulted in an even better solution for the clustering and for the selected explanations.

It would be interesting to investigate the imputed data intrinsic model further. This paper tests the imputed data intrinsic model for datasets that contain missing completely at random values. In reality, values are more commonly missing at random or missing not at random. Further research could consider the different forms of missing data to make the imputed data intrinsic model even more applicable in real-life situations. Perhaps new constraints can be introduced to increase the performance of the models. Lastly, the dissimilarity between individuals can be determined in various different ways. A comparative analysis of the intrinsic models utilizing diverse dissimilarity measures would be a valuable addition to the research field of interpretable Machine Learning.

References

Acuna, E. & Rodriguez, C. (2004). The treatment of missing values and its effect on classifier accuracy. In *Classification, clustering, and data mining applications: Proceedings of the meeting of the international federation of classification societies (ifcs), illinois institute of technology, chicago, 15–18 july 2004* (pp. 639–647).

Aggarwal, K., Mijwil, M. M., Al-Mistarehi, A.-H., Alomari, S., Gök, M., Alaabdin, A. M. Z. & Abdulrhman, S. H. (2022). Has the future started? the current growth of artificial intelligence, machine learning, and deep learning. *Iraqi Journal for Computer Science and Mathematics*, *3*(1), 115–123.

Altman, D. G. & Bland, J. M. (2007). Missing data. *Bmj*, *334*(7590), 424–424.

Bacelar, M. (2021). Possible ethics on machine learning biases and their impacts in future prospects. *ScienceOpen Preprints*.

Carrizosa, E., Kurishchenko, K., Marín, A. & Morales, D. R. (2022). Interpreting clusters via prototype optimization. *Omega*, *107*, 102543.

Carrizosa, E., Kurishchenko, K., Marín, A. & Morales, D. R. (2023). On clustering and interpreting with rules by means of mathematical optimization. *Computers & Operations Research*, *154*, 106180.

Celton, M., Malpertuy, A., Lelandais, G. & De Brevern, A. G. (2010). Comparative analysis of missing value imputation methods to improve clustering and interpretation of microarray experiments. *BMC genomics*, *11*, 1–16.

Chi, J. T., Chi, E. C. & Baraniuk, R. G. (2016). k-pod: A method for k-means clustering of missing data. *The American Statistician*, *70*(1), 91–99.

Daskin, M. S. & Maass, K. L. (2015). The p-median problem. In *Location science* (pp. 21–45). Springer.

De Souto, M. C., Jaskowiak, P. A. & Costa, I. G. (2015). Impact of missing data imputation methods on gene expression clustering and classification. *BMC bioinformatics*, *16*(1), 1–9.

Doumpos, M., Zopounidis, C., Gounopoulos, D., Platanakis, E. & Zhang, W. (2022). Operational research and artificial intelligence methods in banking. *European Journal of Operational Research*.

García, S. & Marín, A. (2015). Covering location problems. *Location science*, 93–114.

Kukkonen, J., Lindroos, A. & Brauer, A. (2022). Explainable ai for the natural sciences.

Lawless, C. & Gunluk, O. (2022). Cluster explanation via polyhedral descriptions. *arXiv preprint arXiv:2210.08798*.

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).

Lin, W.-C. & Tsai, C.-F. (2020). Missing value imputation: a review and analysis of the literature (2006–2017). *Artificial Intelligence Review*, *53*, 1487–1509.

Little, R. J. & Rubin, D. B. (2019). *Statistical analysis with missing data* (Vol. 793). John Wiley & Sons.

Montavon, G., Kauffmann, J., Samek, W. & Müller, K.-R. (2022). Explaining the predictions of unsupervised learning models. In *xxai-beyond explainable ai: International workshop, held in conjunction with icml 2020, july 18, 2020, vienna, austria, revised and extended papers* (pp. 117–138).

Patterson, B. W., Engstrom, C. J., Sah, V., Smith, M. A., Mendonça, E. A., Pulia, M. S., ... Shah, M. N. (2019). Training and interpreting machine learning algorithms to evaluate fall risk after emergency department visits. *Medical care*, *57*(7), 560.

Rodwell, L. (2014). Comparison of methods for imputing limited-range variables: a simulation study. *BMC Medical Research Methodology*.

Saunders, J., Hunt, P. & Hollywood, J. S. (2016). Predictions put into practice: a quasi-experimental evaluation of chicago's predictive policing pilot. *Journal of experimental criminology*, *12*, 347–371.

Wagstaff, K. (2004). Clustering with missing values: no impuation required.

White, I. R., Daniel, R. & Royston, P. (2010). Avoiding bias due to perfect prediction in multiple imputation of incomplete categorical variables. *Computational statistics & data analysis*, *54*(10), 2267–2275.

A Programming code description

In total 3 Python and 6 Java classes were used to obtain the results denoted in this paper.

The following classes are the Python classes that have been used.

- **MCAR&Impute**: Generates missing data points, imputes missing data points using the mean of non-missing data points belonging to the corresponding feature and saves a matrix for which a value equals 1 if the corresponding index in the imputed dataset is an imputed value, otherwise the value is zero.

- **Deciles**: Determines the decile values of the data features.

- **Normalization**: Normalizes the data utilizing the min-max normalization. It also contains the code to perform k-means clustering providing the k-means cluster assignments.

Now we will shortly describe the Java classes.

- **InputData**: Consists of methods to determine the input for the models, namely the dissimilarity matrix (*dissimilarityMatrix* method) and the rule matrix (*ruleMatrix* method). It also reads the other input such as the initial X and Z in the form of a CSV file and transforms them into Java matrices using the *kmeans* method. This class is extended by all of the following classes.
- **PostHocModel**: Solves the post-hoc model for instances with binary variables.
- **PostHocModel_BreastCancer**: Solves the post-hoc model for instances without binary variables.
- **IntrinsicModel_Housing**: Tries to find a feasible solution for the intrinsic model for instances with binary variables.
- **IntrinsicModel_BreastCancer**: Tries to find a feasible solution for the intrinsic model for instances without binary variables.
- **IntrinsicModel_CryoExt**: Tries to find a feasible solution for the imputed data intrinsic model.

To obtain the results in Table 3, run the class PostHocModel with the housing data and the decile values as the rule thresholds. Run the same class with the housing data only with all the unique values of the features to get the results from Table 4.

When you run the PostHocModel_BreastCancer class with the breast cancer dataset and the decile values as the rule thresholds, the results from Table 13 can be obtained.

The IntrinsicModel_Housing can be run on the housing dataset with the decile values and all unique values as the rule thresholds to acquire the values stated in Tables 5 and 6 correspondingly.

To attain the results of Table 14, run the IntrinsicModel_BreastCancer on the breast cancer dataset with the decile values as the rule thresholds.

Running the IntrinsicModel_CryoExt with the cryotherapy will provide the results of Table 7. When you use the forest fires data the results of Table 15 are obtained and the values of Table 8 can be calculated by means of the method's output.

B Data feature description

Table 9: Feature descriptions of the housing dataset

Feature	Description
CRIM	Crime rate by town per capita
ZN	Proportion of residential land zoned for lots over 25,00 squared feet
INDUS	Proportion of non-retail business acres per town
CHAS	Dummy variable for Charles River (equal to 1 if river is bounded by tract; o otherwise)
NOX	Nitric oxides concentration in parts per 10 million
RM	Average number of rooms per residence
AGE	Proportion of owner-occupied units built before 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(Bk - 0.63)ˆ2 where Bk denotes the proportion of black people by town
LSTAT	Percentage of lower status of the population
Classes	Value above (class 1) or below (class 2) the median value of owner-occupied homes in \$1000's

Table 10: Feature descriptions of the breast cancer dataset

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
Chromatin	Bland Chromatin
Normal Nucleoli	Normal Nucleoli
Mitoses	Mitoses
Classes	Benign (class 1) or Malignant (class 2)

Table 11: Feature descriptions of the cryotherapy dataset where * indicates the features for which missing values are generated

Feature	Description
Age	Age of the patient
Time*	Time elapsed before treatment in months
NumbWarts*	Number of warts
Type	Type of warts
Area*	Surface are that warts cover in mm^2
Classes	Malignant (class 1) or Benign (class 2)

Table 12: Feature descriptions of the forest fires dataset where * indicates the features for which missing values are generated

Feature	Description
Temp	Temperature at noon in Degrees Celsius
RH	Relative humidity in percentages
WS	Wind speed in km/h
Rain	Total rainfall in mm
FFMC*	Fine Fuel Moisture Code index from the FWI system
DMC*	Duff Moisture Code index from the FWI system
DC*	Drought Code index from the FWI system
ISI*	Initial Spread Index from the FWI system
BUI*	Buildup Index from the FWI system
FWI	Fire Weather Index
Classes	Forest fires did not occur (class 1), Forest fires did occur (class 2)

C Results of the breast cancer dataset

Table 13: The clusters and explanations derived from the post-hoc model for the breast cancer dataset constructed with K = 2 clusters, a maximum length of l = 2, utilizing N = 84 rules generated from the deciles of the continuous features and considering all attributes of the categorical features, along with the TPR and FPR found in the reference paper.

$\boldsymbol{\theta}$	C	TPR	FPR	Ref TPR	Ref FPR	Explanations
2^5	1	0.68	0.00	0.68	0.00	Size>4 AND Adhesion>1
	2	0.86	0.00	0.85	0.00	Size≤2 AND Nuclei≤2
2^4	1	0.68	0.00	0.68	0.00	Size>4 AND Adhesion>1
	2	0.90	0.01	0.85	0.00	Epithelial Size≤3 AND Nuclei≤2
2^3	1	0.68	0.00	0.68	0.00	Size>4 AND Adhesion>1
	2	0.90	0.01	0.90	0.01	Epithelial Size≤3 AND Nuclei≤2
2^2	1	0.72	0.01	0.72	0.01	Size>4
	2	0.90	0.01	0.90	0.01	Epithelial Size≤3 AND Nuclei≤2
2^1	1	0.89	0.04	0.88	0.04	Size>2 and Nuclei>1
	2	0.93	0.03	0.93	0.03	Shape≤3 AND Chromatin≤3
2^0	1	0.95	0.07	0.95	0.07	Size>2 AND Shape>1
	2	0.97	0.07	0.96	0.07	Size≤4 AND Nuclei≤4
2^{-1}	1	0.95	0.07	0.95	0.07	Size>2 AND Shape>1
	2	0.98	0.14	0.99	0.14	Size≤4 AND Nuclei≤9
2^{-2}	1	0.98	0.12	0.98	0.12	Size>1 AND Shape>1
	2	0.99	0.14	0.99	0.14	Size≤4 AND Nuclei≤9
2^{-3}	1	0.98	0.12	0.98	0.12	Size>1 AND Shape>1
	2	0.99	0.19	0.99	0.19	Thickness≤9.8 AND Size≤4
2^{-4}	1	0.98	0.12	0.98	0.12	Size>1 AND Shape>1
	2	0.99	0.19	0.99	0.19	Thickness≤9.8 AND Size≤4
2^{-5}	1	0.99	0.23	0.99	0.23	Shape>1 AND Nuclei≤10
	2	1.00	0.52	1.00	0.52	Thickness≤9.8 AND Size≤9

Table 14: The clusters and explanations derived from the intrinsic model for the breast cancer dataset constructed with K = 2 clusters, a maximum length of l = 2, utilizing N = 84 rules generated from the deciles of the continuous features and considering all attributes of the categorical features, along with the TPR, FPR and IH found in the reference paper.

$\boldsymbol{\theta_1}$	$\boldsymbol{\theta_2}$	**IH**	**Ref IH**	**C**	**TPR**	**FPR**	**Ref TPR**	**Ref FPR**	**Explanations**
0.5	0.5	0.67*10^5	1.73*10^5	1	0.90	0.02	1.00	0.00	Size>3 AND Nuclei>2
				2	0.97	0.02	1.00	0.00	Size≤4 AND Nuclei≤4
0.5	1	0.67*10^5	0.67*10^5	1	0.90	0.02	0.90	0.02	Size>3 AND Nuclei>2
				2	0.97	0.02	0.97	0.02	Size≤4 AND Nuclei≤4
0.5	2	0.67*10^5	1.10*10^5	1	0.80	0.02	1.00	0.00	Size>3 AND Nuclei>4
				2	0.97	0.00	0.97	0.00	Size≤4 AND Nuclei≤4
1	0.5	0.67*10^5	1.24*10^5	1	0.97	0.09	1.00	0.00	Size>3 AND Shape>1
				2	0.99	0.07	1.00	0.00	Size≤4 AND Nuclei≤9
1	1	0.67*10^5	1.24*10^5	1	0.90	0.02	1.00	0.00	Size>3 AND Nuclei>2
				2	0.97	0.02	1.00	0.00	Size≤4 AND Nuclei≤4
1	2	0.67*10^5	1.24*10^5	1	0.90	0.02	1.00	0.00	Size>3 AND Nuclei>2
				2	0.97	0.02	1.00	0.00	Size≤4 AND Nuclei≤4
2	0.5	0.67*10^5	1.24*10^5	1	0.97	0.09	1.00	0.00	Size>3 AND Shape>1
				2	0.99	0.07	1.00	0.00	Size≤4 AND Nuclei≤9
2	1	0.67*10^5	1.24*10^5	1	0.97	0.09	1.00	0.00	Size>3 AND Shape>1
				2	0.99	0.07	1.00	0.00	Size≤4 AND Nuclei≤9
2	2	0.67*10^5	1.24*10^5	1	0.90	0.02	1.00	0.00	Size>3 AND Nuclei>2
				2	0.97	0.02	1.00	0.00	Size≤4 AND Nuclei≤4

D Forest fires imputed data intrinsic model results

Table 15: The clusters and explanations derived from the imputed data intrinsic model for the forest fires dataset constructed with K = 2 clusters, a maximum length of $l = 2$, utilizing N = 132 rules generated from the deciles of the continuous features and considering all attributes of the categorical features, along with the IH, TPR and FPR values for the complete dataset.

θ_1	θ_2	λ	IH	Actual IH	C	TPR	FPR	Actual TPR	Actual FPR	Explanation
0.5	0.5	0	$2.63*10^3$	$3.73*10^3$	1	0.89	0.05	0.85	0.05	Temp≤31 AND BUI≤17.86
					2	0.16	0.26	0.16	0.26	Temp>26.1 AND Rain>0.37
		1	$2.59*10^3$	$3.68*10^3$	1	1.00	0.52	1.00	0.52	Temp≤26.1 AND FWI>0.4
					2	1.00	0.21	1.00	0.21	Temp≤33 AND FWI≤7.44
0.5	1	0	$2.63*10^3$	$3.73*10^3$	1	0.90	0.03	0.90	0.03	Temp≤31 AND FWI≤7.44
					2	0.87	0.05	0.87	0.05	Temp>31 AND RH≤76
		1	$2.63*10^3$	$3.73*10^3$	1	0.90	0.03	0.90	0.03	Temp≤31 AND FWI≤7.44
					2	0.87	0.05	0.87	0.05	Temp>31 AND RH≤76
0.5	2	0	$2.63*10^3$	$3.73*10^3$	1	0.90	0.05	0.85	0.02	Temp≤31 AND FFMC≤85.32
					2	0.82	0.03	0.82	0.03	Temp>31 AND RH≤73
		1	$2.63*10^3$	$3.73*10^3$	1	0.90	0.05	0.85	0.02	Temp≤31 AND FFMC≤85.32
					2	0.82	0.03	0.82	0.03	Temp>31 AND RH≤73
1	0.5	0	$2.63*10^3$	$3.73*10^3$	1	0.90	0.03	0.90	0.03	Temp≤31 AND FWI≤7.44
					2	0.98	0.26	0.98	0.26	Temp>30 AND DC>8.4
		1	$2.62*10^3$	$3.71*10^3$	1	0.92	0.05	0.92	0.05	Temp≤31 AND FWI≤7.44
					2	0.94	0.20	0.94	0.20	Temp>30 AND RH≤76
1	1	0	$2.63*10^3$	$3.73*10^3$	1	0.90	0.03	0.90	0.03	Temp≤31 AND FWI≤7.44
					2	0.87	0.05	0.87	0.05	Temp>31 AND RH≤76
		1	$2.63*10^3$	$3.73*10^3$	1	0.90	0.03	0.90	0.03	Temp≤31 AND FWI≤7.44
					2	0.87	0.05	0.87	0.05	Temp>31 AND RH≤76
1	2	0	$2.63*10^3$	$3.73*10^3$	1	0.90	0.03	0.90	0.03	Temp≤31 AND FWI≤7.44
					2	0.87	0.05	0.87	0.05	Temp>31 AND RH≤76
		1	$2.63*10^3$	$3.73*10^3$	1	0.90	0.03	0.90	0.03	Temp≤31 AND FWI≤7.44
					2	0.87	0.05	0.87	0.05	Temp>31 AND RH≤76
2	0.5	0	$2.63*10^3$	$3.73*10^3$	1	0.98	0.31	0.98	0.31	RH>53 AND FWI≤7.44
					2	0.98	0.26	0.98	0.26	Temp>30 AND DC>8.4
		1	$2.63*10^3$	$3.73*10^3$	1	0.98	0.31	0.98	0.31	RH>53 AND FWI≤7.44
					2	0.98	0.26	0.98	0.26	Temp>30 AND DC>2.6
2	1	0	$2.63*10^3$	$3.73*10^3$	1	0.90	0.03	0.90	0.03	Temp≤31 AND FWI≤7.44
					2	0.98	0.26	0.98	0.26	Temp>30 AND DC>8.4
		1	$2.63*10^3$	$3.73*10^3$	1	0.90	0.03	0.90	0.03	Temp≤31 AND FWI≤7.44
					2	0.98	0.26	0.98	0.26	Temp>30 AND DC>2.6
2	2	0	$2.63*10^3$	$3.73*10^3$	1	0.90	0.03	0.90	0.03	Temp≤31 AND FWI≤7.44
					2	0.87	0.05	0.87	0.05	Temp>31 AND RH≤76
		1	$2.63*10^3$	$3.73*10^3$	1	0.90	0.03	0.90	0.03	Temp≤31 AND FWI≤7.44
					2	0.87	0.05	0.87	0.05	Temp>31 AND RH≤76