Novel Mathematical Optimization Models for Explainable and Fair Machine Learning

Kurishchenko, Kseniia

Document Version
Final published version

DOI:
10.22439/phd.21.2024

Publication date:
2024

License
Unspecified

Citation for published version (APA):
Kurishchenko, K. (2024). *Novel Mathematical Optimization Models for Explainable and Fair Machine Learning*. Copenhagen Business School [Phd]. PhD Series No. 21.2024 https://doi.org/10.22439/phd.21.2024

COPENHAGEN BUSINESS SCHOOL

Solbjerg Plads 3
DK-2000 Frederiksberg
Danmark

www.cbs.dk

ISSN 0906-6934

Print ISBN: 978-87-7568-273-7
Online ISBN: 978-87-7568-274-4

NOVEL MATHEMATICAL OPTIMIZATION MODELS FOR EXPLAINABLE AND FAIR MACHINE LEARNING

PhD Series 21-2024

CBS PhD School
Department of Economics

PhD Series 21.2024

KSENIIA KURISHCHENKO

NOVEL MATHEMATICAL OPTIMIZATION MODELS FOR EXPLAINABLE AND FAIR MACHINE LEARNING

CBS

Copenhagen Business School
Department of Economics

Doctor of Philosophy
Doctoral Thesis in Economics

Thesis:

Novel Mathematical Optimization Models for Explainable and Fair Machine Learning

Kseniia Kurishchenko

Primary supervisor:
Prof Dolores Romero Morales
Secondary supervisors:
Prof Emilio Carrizosa
Prof Ralf Andreas Wilke

Kseniia Kurishchenko
Novel Mathematical Optimization Models for Explainable and Fair Machine Learning

First edition 2024
Ph.D. Series 21.2024

ISSN 0906-6934

Print ISBN: 978-87-7568-273-7
Online ISBN: 978-87-7568-274-4

DOI: https://doi.org/10.22439/phd.21.2024

Acknowledgement

I am honored to present my thesis, which is the result of my Ph.D. studies at the Department of Economics, Copenhagen Business School. During my studies, the Department provided me with unwavering support and encouragement, for which I am immensely grateful.

However, I must acknowledge that these times have been difficult. At the onset of my Ph.D., a pandemic emerged, and we were forced to adjust to a new normal that kept us away from our friends and colleagues. This was a trying time for us all, and I am grateful for the support and guidance provided by my main supervisor, Dolores Romero Morales, and my secondary supervisor, Emilio Carrizosa. Their invaluable advice, continuous support, patience, and kindness have been a lifeline throughout my Ph.D. studies. Their wealth of experience and expertise has encouraged me throughout my research and daily life. I have learned much from scientific discussions with them. They have always been thoughtful and have put much time into my research-related and personal issues. I am also grateful to my secondary supervisor, Ralf Andreas Wilke, for his insightful comments and feedback on my work.

Lastly, I would like to take this opportunity to express my heartfelt gratitude to my family including my husband and my daughter. They have always been my backbone, inspiring me to pursue my curiosity and passion for science. Their unwavering support and encouragement have been essential in helping me to complete my studies.

Abstract

This thesis consists of six chapters including the introduction and the conclusions. The chapters are dedicated to enhancing the transparency of key models in Machine Learning. In this dissertation, I propose novel Mathematical Optimization models to trade off accuracy and transparency in Cluster Analysis, Supervised Classification, and Treatment Allocation.

In Chapter II, co-authored with Emilio Carrizosa, Alfredo Marín, and Dolores Romero Morales, we tackle the problem of enhancing the interpretability/explainability of the results of Cluster Analysis, which is one of the transparency criteria pursued in this dissertation. Our goal is to find an explanation for each cluster, such that clusters are characterized as precisely and distinctively as possible, i.e., the explanation is fulfilled by as many as possible individuals of the corresponding cluster, true positive cases, and by as few as possible individuals in the remaining clusters, false positive cases. We assume that a dissimilarity between the individuals is given, and propose distance-based explanations, namely those defined by individuals that are close to its so-called prototype. To find the set of prototypes, we address the bi-objective optimization problem that maximizes the total number of true positive cases across all clusters and minimizes the total number of false positive cases, while controlling the true positive rate as well as the false positive rate in each cluster. We develop two Mixed Integer Linear Programming (MILP) models, inspired by classic Location Analysis problems, that differ in the way individuals are allocated to prototypes. We illustrate the explanations provided by these models and their accuracy in both real-world data as well as simulated data.

In Chapter III, co-authored with Emilio Carrizosa, Alfredo Marín, and Dolores Romero Morales, 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 individu-

als 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 MILP problem, from which non-dominated solutions are generated. We illustrate the clusters and the accuracy of the corresponding explanations in real-world data.

In Chapter IV, co-authored with Emilio Carrizosa and Dolores Romero Morales, we investigate how to make tree ensembles in Supervised Classification more transparent, incorporating by design explainability and fairness criteria. While explainability helps the user understand the key features that play a role in the classification task, with fairness we ensure that the ensemble does not discriminate against a group of observations that share a sensitive attribute. We propose an MILP formulation to train an ensemble of trees that apart from minimizing the misclassification error, controls for sparsity as well as the accuracy in the sensitive group. Our formulation is scalable in the number of observations since its number of binary decision variables is independent of the number of observations. In our numerical results, we show that for standard datasets used in the fairness literature, we can dramatically enhance the fairness of the benchmark, namely the popular Random Forest, while using only a few features, all without damaging the misclassification error.

In Chapter V, I investigate the Treatment Allocation problem, where one has to decide which individuals will receive treatment and which not. If not carefully trained, the algorithm may provide unfair results, unequally allocating treatment to individuals in the sensitive (e.g., females) and non-sensitive (e.g., males) groups. To deal with it I propose to measure unfairness as the difference between the average treatment effects in the sensitive group and the non-sensitive group. I introduce a Mathematical Optimization model to have accurate heterogeneous treatment effect predictions and a good level of fairness, which will be the basis for the treatment allocation in forthcoming individuals. I present results on simulated datasets, illustrating that my model provides fairer predictions of the treatment effect than the benchmark.

Danish abstract

Denne afhandling består af seks kapitler inklusiv introduktionen og konklusionen. Kapitlerne er dedikeret til at gøre en række modeller inden for Machine Learning mere transparente. I afhandlingen foreslår jeg nye matematiske optimeringsmodeller til at afveje præcision og transparens i klyngeanalyse, supervised klassifikation, og allokering af behandlinger.

I Kapitel II, som er skrevet i samarbejde med Emilio Carrizosa, Alfredo Marín og Dolores Romero Morales, håndterer vi problemet om at øge fortolkeligheden af resultater fra klyngeanalyse, som er en af de transparenskriterier der forfølges i afhandlingen. Vores mål er at finde en forklaring for hver klynge, således at klynger er karakteriseret så præcist og distinkt som muligt. Med andre ord at forklaringen er opfyldt for flest mulige individer i den tilsvarende klynge, de sandt positive tilfælde, og af færrest mulige individer i de resterende klynger, de falsk positive tilfælde. Vi antager at der er givet et forskellighedsmål mellem individerne, og foreslår afstandsbaserede forklaringer, som er dem der er defineret af individer tæt på den såkaldte prototype. For at finde den mængde af prototyper adresserer vi det bi-objektive optimeringsproblem der maksimerer det totale antal sandt positive tilfælde på tværs af alle klynger og minimerer det totale antal falsk positive tilfælde, samtidig med at vi kontrollerer sandt positiv raten og falsk positiv raten i hver klynge. Vi konstruerer to lineære blandede heltalsprogrammeringsmodeller, som er inspireret af klassiske lokationsanalyse problemer, der adskiller sig i måden individer bliver allokeret prototyper. Vi illusterer forklaringerne givet af disse modeller og deres præcision på både virkelig data og simuleret data.

I Kapitel III som er skrevet i samarbejde med Emilio Carrizosa, Alfredo Marín og Dolores Romero Morales, gør vi klyngeanalyse mere fortolkelig med en ny fremgangsmåde der samtidig allokerer individer til klynger og giver regel baseret forklaringer til hver klynge. Det traditionelle homogenitetsmål i klyngeanalyse, som er summen af forskelle mellem individer i samme klynge, udvides ved også at tage højde for to forklaringenskriterier for hver klynge og dets tilhørende forklaring, nemlig præcisionen af forklaringerne, med andre ord hvor mange individer i klyngen opfylder forklaringen, og hvor distinkt forklaringen er, med andre ord hvor mange individer uden for

klyngen opfylder dens forklaring. At finde klyngerne og forklaringener der optimerer homogenitet, præcision, og distinktionen er formuleret som et multi-objektiv lineært blandet heltalsprogram, hvorfra ikke domineret løsninger genereres. Vi illustrerer klyngerne og præcisionen af de tilhørende forklaringer på virkelig data.

I Kapitel IV, som er skrevet i samarbejde med Emilio Carrizosa og Dolores Romero Morales, undersøger vi hvordan man kan lave "tree ensembles" i supervised klassifikation mere transparente ved at inkorporere forklarligheds og "fairness" kriterier. Mens forkarlighed hjælper brugeren med at forstå vigtige karakteristika der spiller en rolle i klassifikationen, sikrer vi os med "fairness" at vores "ensemble" ikke diskriminerer mod bestemte grupper af individer der kan karakteriseres af følsom information. Vi foreslår et lineært blandet heltalsprogram til at træne en "ensemble" af træer der udover at minimere fejlklassificering, også kontrollerer for "sparsity" og præcision i den beskyttede gruppe. Vores program er skalerbart i forhold til antallet af observationer, da antallet af binære variable er uafhængig af antallet af observationer. I vores numeriske resultater viser vi at på standard datasæt brugt i "fairness" literaturen kan vi drastisk øge "fairness" på vores benchmark, nemlig "random forest", imens vi kun bruger få kovariater og uden at øge fejlklassificering.

I Kapitel V undersøger jeg "Treatment Allocation" problemet, hvor man skal beslutte hvilke individer der får behandling. Hvis algoritmen ikke trænes forsigtigt kan den give unfair resultater som en ulige fordeling af behandlinger mellem den beskyttede (f.eks. kvinder) og ikke beskyttede (f.eks. mænd) gruppe. For at håndtere dette, foreslår jeg at måle "unfairness" som forskellen i behandlingseffekten mellem den beskyttede og ikke beskyttede gruppe. Jeg introducerer en matematisk optimeringsmodel til at have præcise prædiktioner af heterogene behandslingseffekter og et godt niveau af "fairness", som er basis for allokeringen af behandlinger til kommende individer. Jeg viser resultaterne på simulerede datasæt, der illusterer at min model giver mere fair prædiktioner af behandlingseffekter end benchmarken.

Contents

List of Figures

List of Tables

Chapter I

Introduction

The use of Artificial Intelligence (AI) and Machine Learning (ML) to aid Data Driven Decision Making is increasing dramatically. The wide availability of AI/ML algorithms brings important advantages, such as the improved accuracy of decisions and the reduction in the resources required to make them [Athey, 2017, Bertsimas et al., 2022, Jordan and Mitchell, 2015]. Despite excellent accuracy, state-of-the-art ML tools such as Deep Learning [Goodfellow et al., 2016], Random Forests [Breiman, 2001], and Support Vector Machines [Cortes and Vapnik, 1995], are effectively black boxes that complicate model trustworthiness and may provide unfair outcomes. The use of these methodologies requires caution when deploying them in high-stakes decision making due to social, ethical, and legal concerns [Gunning et al., 2019, Shin, 2021]. The need for transparent Machine Learning models is huge in many areas, e.g., credit scoring, medical diagnosis and regulatory benchmarking [Baesens et al., 2003, Benítez-Peña et al., 2020, Carrizosa and Romero Morales, 2013, Di Teodoro et al., 2024, Freitas, 2014, Kleinberg et al., 2018, Lepri et al., 2017]. Public authorities are also demanding transparency in algorithmic decision making [Goodman and Flaxman, 2017]. This Ph.D. dissertation contributes to modeling the trade-off between accuracy and transparency for important tasks in Unsupervised Learning (namely, Cluster Analysis), Supervised Learning (namely, classification and regression), and treatment allocation.

Transparency in AI can take various forms such as accountability, explainability or fairness [Hutchinson et al., 2021, Panigutti et al., 2023], and in this dissertation, I focus on the last two.

Explainability is the concept that an ML model and its output can be easily explained to a human being. There is a growing literature on Interpretable ML, such as transparent neural networks [Samek et al., 2021, Wu et al., 2021], interpretable random forests [Bénard et al., 2021, Vidal and Schiffer, 2020], or sparse support vector machines [Benítez-Peña et al., 2019, Carrizosa et al., 2016, Jiménez-Cordero et al., 2021]. In this dissertation, several forms of explainability are used and described in what follows. The ML models can be explainable by construction. One of the main exponents are decision trees [Carrizosa et al., 2021b, Gordon et al., 1984]. This type of model builds a tree where the obtained path from the root node to a given leaf node can be considered as an if-then rule explanation for any observation falling in that leaf. Another type of explanation is prototype-based [Carrizosa et al., 2007, Cover and Hart, 1967]. This explanation consists of a prototype or representative individual that speaks for a group of individuals. They are used in, e.g., Cluster Analysis when one needs to understand the "average" individual of the group. Another approach to improve the explainability of a model is to enhance its sparsity, i.e., to use fewer explanatory variables, where one of the main exponents is LASSO [Hastie et al., 2019, Tibshirani, 1996]. Alternatively, one can explain the decisions made by the ML model after it

has been trained. This is a rapidly growing research direction referred as Explainable Artificial Intelligence (XAI) [Barredo Arrieta et al., 2020, Heaton and Fung, 2023, Lakkaraju et al., 2017]. The most popular XAI techniques are LIME [Ribeiro et al., 2016] and SHAP [Lundberg and Lee, 2017].

Explainability is not the only concern in ML models. Indeed, nowadays there are many examples of algorithmic bias against sensitive groups, that are, for example, exposed to structural discrimination, sexism, racism, or the like [Corbett-Davies and Goel, 2018, Miron et al., 2020, Romei and Ruggieri, 2014]. If not carefully trained, the ML model can provide unfair results [Mehrabi et al., 2022]. To minimize this effect many fairness measures have been introduced in the literature [Hort et al., 2022, Zafar et al., 2017]. Fairness measures control the training process so that algorithmic bias can be mitigated, as will be done in this thesis for classification and regression as well as treatment allocation tasks.

In this dissertation, I make more transparent key methodologies in Cluster Analysis, classification and regression, and treatment allocation. The main goal of Cluster Analysis is to split similar objects into groups or clusters. It can be useful in applications such as recommendation systems where a recommendation is given based on similar individuals (clusters), market and customer segmentation, biological data segmentation, etc. Another type of problem I consider is classification and regression via Tree Ensembles such as Random Forests [Biau and Scornet, 2016, Breiman, 2001] and XGBoost [Chen and Guestrin, 2016]. In Tree Ensembles, a collection of trees is built and the final prediction is made by combining the predictions obtained from each of the trees. The last type of problem I consider is treatment allocation. This is a task where a decision maker needs to decide whom to allocate to treatment (surgery, loan, etc). In order to satisfy the budget constraints the decision maker can only allocate treatment to those who get the highest benefit from the intervention. Thus, to define those most beneficial individuals the prediction of the treatment effect is needed.

Mathematical Optimization is the core methodology to address these key tasks [Carrizosa and Romero Morales, 2013, Carrizosa et al., 2021b]. In Chapters II–V, I propose Mathematical Optimization problems to build ML models that balance accuracy, explainability and fairness. By imposing objectives and constraints, these Mathematical Optimization problems allow us to achieve high accuracy while enhancing explainability and fairness. In particular, Chapters II and III relate to Cluster Analysis, Chapter IV to classification and regression via Tree Ensembles, while Chapter V to treatment allocation via Tree Ensembles.

In Chapter II, based on the work in Carrizosa et al. [2022b], we associate to each cluster a

prototype. As the measure of quality of such an allocation, we consider the number of individuals of a cluster that are closer to their cluster prototype than to the other prototypes. We develop two Mixed Integer Linear Programming (MILP) models that select the prototypes requiring only a dissimilarity between the individuals. We illustrate on real-world as well as simulated datasets that we manage to find accurate explanations for the clusters.

In Chapter III, based on the work in Carrizosa et al. [2023a], we associate with each cluster a small set of clauses joined by the AND operator. As the measure of quality of such an allocation, we consider the number of individuals in a cluster that satisfy all the clauses from their cluster but not all the clauses from the other clusters. We develop two MILP models that select the clauses for each cluster. We illustrate on several real-world datasets that accurate explanations for the clusters can be found.

In Chapter IV, based on the work in Carrizosa et al. [2023b], we consider the problem of making Classification and Regression Tree Ensembles tasks more explainable and fairer to sensitive groups. We propose a tree weighting approach via an MILP problem that allocates higher weight to "better" trees and lower weight or even zero weight to "bad" trees. With this, we are able to control for sparsity (a proxy of explainability) and the accuracy of the Tree Ensemble in the sensitive group (our measure of fairness). We illustrate on several real-world datasets that we manage to increase the sparsity and fairness of the original Tree Ensemble without harming its accuracy.

In Chapter V, based on the work in Kurishchenko [2023], I extend the idea of Chapter IV to consider the treatment allocation problem. I use Tree Ensembles to predict treatment effects, which will be the basis for the treatment allocation in forthcoming individuals. I introduce a fairness measure to ensure that the predicted treatment effects in the sensitive and non-sensitive groups on average do not differ much. I reweight the trees in the ensemble via a Convex Quadratic Programming problem to have accurate outcome predictions and a good level of fairness of the treatment effect predictions. I illustrate on simulated datasets that my model provides fairer predictions of the treatment effect than the benchmark.

Finally, some general conclusions and possible lines of future research are provided in Chapter VI.

Chapter II

Interpreting clusters via prototype optimization

II.1 Introduction

This chapter is devoted to the interpretability of one of the most popular Unsupervised Learning tasks, namely, Cluster Analysis [Aloise et al., 2012, Gan et al., 2007, Kaufmann and Rousseeuw, 1990]. The need for interpretability in Cluster Analysis arises in many applications, such as security [Corral et al., 2009], internet traffic [Morichetta et al., 2019], finance [Gibert and Conti, 2016], sales profiling [Thomassey and Fiordaliso, 2006], and astronomy [Ma et al., 2018].

There are two ways of enhancing interpretability in Cluster Analysis: intrinsic models and post-hoc models. Intrinsic models build simultaneously clusters and their explanations [Bertsimas et al., 2021, Chen et al., 2016], while post-hoc approaches are needed to interpret existing clusters, that have been built in the past, and for which we only have a label for each individual. There are some works in the literature on post-hoc approaches. In Davidson et al. [2018], the authors assume that the individuals have been evaluated on a set of features and propose rule-based explanations. There are also ad-hoc approaches as those in, e.g., Balabaeva and Kovalchuk [2020], De Koninck et al. [2017], Kauffmann et al. [2022], for specific types of data. In this chapter, we propose a post-hoc approach for interpreting clusters via means of prototypes.

Throughout this section, we will use a running example with clusters given, namely the real-world dataset containing twelve countries about the opinions of political science students, see Table II.1. In Rousseeuw [1987], three clusters are given for this dataset, cluster 1 composed by Belgium, Egypt, France, Israel, and USA; cluster 2 with Brasil, India, and Zaire; and cluster 3 with China, Cuba, USSR, and Yugoslavia.

Table II.1: Dissimilarities on opinions of political science students between the twelve countries in our running example, Rousseeuw [1987].

Country	Dissimilarities to other countries										
	Belgium	Brasil	China	Cuba	Egypt	France	India	Israel	USA	USSR	Yugoslavia
Brasil	5.58										
China	7.00	6.50									
Cuba	7.08	7.00	3.83								
Egypt	4.83	5.08	8.17	5.83							
France	2.17	5.75	6.67	6.92	4.92						
India	6.42	5.00	5.58	6.00	4.67	6.42					
Israel	3.42	5.50	6.42	6.42	5.00	3.92	6.17				
USA	2.50	4.92	6.25	7.33	4.50	2.25	6.33	2.75			
USSR	6.08	6.67	4.25	2.67	6.00	6.17	6.17	6.92	6.17		
Yugoslavia	5.25	6.83	4.50	3.75	5.75	5.42	6.08	5.83	6.67	3.67	
Zaire	4.75	3.00	6.08	6.67	5.00	5.58	4.83	6.17	5.67	6.50	6.92

Our starting point is the predefined clusters in $\mathcal{C}$, which have been obtained applying a clustering procedure to the set of individuals $\mathcal{N}$ [Aloise et al., 2012, 2009, Grötschel and Wakabayashi, 1989, Jain, 2010, Maldonado et al., 2015, Rao, 1971, Seref et al., 2014]. We propose a methodology to improve the interpretability of the results of Cluster Analysis, by giving an explanation to each cluster $c \in \mathcal{C}$ that characterizes as precisely and distinctively as possible c. In other words, the explanation is to be fulfilled by as many as possible individuals of c (and these will be referred to as *true positive* cases) and by as few as possible individuals in the remaining clusters (which will be referred to as *false positive* cases).

Our explanations are distance-based, as in clustering procedures attempting to partition the set of individuals such that individuals that are close to each other are allocated to the same cluster, whereas individuals that are far from each other are expected to be in different clusters. It is then natural to *explain* cluster c following a distance-based explanation such as

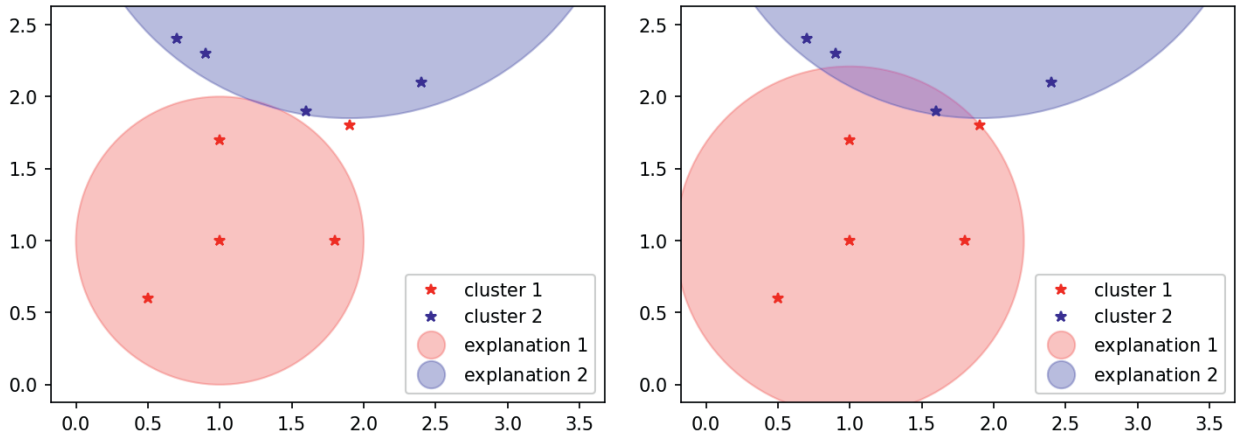
c is the set of individuals of $\mathcal{N}$ that are close to a given individual i.

To define distance-based explanations, we assume we are given a dissimilarity $\boldsymbol{\delta}$ to measure the closeness between individuals [Kaufmann and Rousseeuw, 1990]. The dissimilarity between the twelve countries in our running example is given in Table II.1. Note that, in general, $\boldsymbol{\delta}$ does not need to be the dissimilarity used to construct the clusters in $\mathcal{C}$. Actually, that dissimilarity may not be available to us.

How well this explains cluster c depends on the choice of individual i to which we will refer as the prototype of cluster c [Carrizosa et al., 2007, Cover and Hart, 1967], in other words, the "face" chosen for the cluster. Our aim is to select the set of prototypes that maximizes the total number of true positive cases across all clusters and minimizes the total number of false positive cases while controlling in each cluster the true positive rate, i.e., the number of true positive cases divided by the size of the cluster as well as the false positive rate, i.e., the number of false positive cases divided by the size of the cluster. With the methodology proposed in this chapter, the chosen prototypes for our example are: France for cluster 1, Brasil for cluster 2, and Yugoslavia for cluster 3. For cluster 1, all 5 countries are true positive cases, while none of the 7 countries in the other two clusters are false positive cases, yielding to the ideal quality of the explanation, namely 100% true positive rate and 0% false positive rate. The same holds for the other two clusters.

In general, one cannot expect to find perfect explanations. In Figure II.1, we can see that by trying to improve the number of true positive cases of an explanation we may harm the number of false positive cases. There we have two clusters, cluster 1 with 5 individuals represented by a red star and cluster 2 with 4 individuals represented by a blue star. If we look at the explanation in

Figure II.1a for cluster 1, the circle in red containing 4 of the individuals from cluster 1 and none from cluster 2, we see that there are 4 true positive cases (or, equivalently, an 80% true positive rate) and 0 false positive cases (or, equivalently, a 0% false positive rate), while for the alternative explanation for cluster 1 in Figure II.1b, the number of true positive cases has increased to 5 (achieving a 100% true positive rate) but the number of false positive cases has gone up to 1 (25% false positive rate).



(a) An explanation for cluster 1

(b) An alternative explanation for cluster 1

Figure II.1: Illustration of the trade-off between true positive and false positive cases when interpreting clusters via means of prototypes.

To find the set of prototypes, we propose two mathematical optimization models, the covering and the partitioning ones, inspired by classic Location Analysis problems, namely the covering [García and Marín, 2019] and the p-median problems [García et al., 2011, Marín and Pelegrín, 2019]. In the covering model, a cluster is explained as the individuals whose distance to its prototype is below a threshold value, i.e., the explanation of cluster c can be visualized as the ball in the distance $\boldsymbol{\delta}$ centered at its prototype and radius equal to the corresponding threshold value. Instead, in the set-partitioning model, cluster c is explained as the individuals that are the closest to the prototype of c than to the prototypes of the other clusters. In this case, the explanations can be visualized as Voronoi diagrams. For both models, we provide a Mixed Integer Linear Programming (MILP) formulation, where in the covering one, in addition to the prototypes, we need to decide the size of the radii.

The remainder of the chapter is organized as follows. Section II.2 presents the covering model, while Section II.3 the partitioning model. Section II.4 provides numerical results for real-world data as well as simulated data. Section II.5 summarizes the chapter and proposes future lines of

research.

II.2 The covering model

In this model, given a cluster c, a prototype i, an individual will be considered covered by cluster c if it is close enough to i. By close enough we mean that their dissimilarity is below a threshold value r_c, which is the coverage radius. Our aim is thus to find the prototypes and the cluster radii. Observe that, with this approach, an individual could be covered by more than one cluster if some of the radii are large, while some individuals may not be covered by any cluster when the radii are small. We obtain an MILP formulation for this problem, which is separable on the clusters. We show how the radii can only take on a discrete amount of values, and give an alternative Integer Programming (IP) formulation for a fixed radius. We focus on the most interpretable case in which only one prototype per cluster is to be selected. The extension to more than one prototype is straightforward.

Let us introduce the problem more formally. We are given a clustering $\mathcal{C}$ obtained from splitting the individuals in $\mathcal{N}$, $\mathcal{N} = \bigcup_{c\in\mathcal{C}} \mathcal{N}_c$. The prototype of cluster c is chosen from set $\mathcal{I}_c \subseteq \mathcal{N}_c$, with $\mathcal{I} = \bigcup_{c\in\mathcal{C}} \mathcal{I}_c$. We are also given the dissimilarity between prototype i and individual n, δ_{in}, for every $i \in \mathcal{I}$ and $n \in \mathcal{N}$. This dissimilarity does not need to be the one that was used to construct the clusters. As pointed out in the introduction, we may have been given only clusters, and neither the method nor the dissimilarity used to build them.

Let r_c be the radius of the explanation chosen for cluster c. For $i \in \mathcal{I}_c$, let π_{in} be the binary decision variable which takes on the value 1 if $n \in \mathcal{N}$ lies in the ball of radius r_c centered at prototype $i \in \mathcal{I}$, and 0 otherwise. Moreover, let z_i be the binary decision variable which takes on the value 1 if i is chosen as prototype and 0 otherwise. Throughout this chapter, we use bold typesetting to denote the vectors, e.g., $\mathbf{r} = (r_c)_{c\in\mathcal{C}}$.

With these variables, the number of *true positive* cases in cluster c is equal to $\sum_{i\in\mathcal{I}_c} \sum_{n\in\mathcal{N}_c} \pi_{in} z_i$ and the corresponding True Positive Rate (TPR_c) is

$$\text{TPR}_c = \frac{\sum_{i\in\mathcal{I}_c} \sum_{n\in\mathcal{N}_c} \pi_{in} z_i}{|\mathcal{N}_c|}, \tag{II.2.1}$$

while the number of *false positive* cases in cluster c is equal to $\sum_{i\in\mathcal{I}_c} \sum_{n\in\mathcal{N}\setminus\mathcal{N}_c} \pi_{in} z_i$ and the corresponding False Positive Rate (FPR_c) is

$$\text{FPR}_c = \frac{\sum_{i\in\mathcal{I}_c} \sum_{n\in\mathcal{N}\setminus\mathcal{N}_c} \pi_{in} z_i}{|\mathcal{N} \setminus \mathcal{N}_c|}. \tag{II.2.2}$$

The covering model reads as follows:

$$\max_{\mathbf{z},\boldsymbol{\pi},\mathbf{r}} \quad \sum_{c\in\mathcal{C}}\sum_{i\in\mathcal{I}_c}\sum_{n\in\mathcal{N}_c}\pi_{in}z_i - \theta\sum_{c\in\mathcal{C}}\sum_{i\in\mathcal{I}_c}\sum_{n\in\mathcal{N}\setminus\mathcal{N}_c}\pi_{in}z_i \tag{II.2.3}$$

$$\text{s.t.} \quad \sum_{i\in\mathcal{I}_c} z_i = 1, \qquad \forall c\in\mathcal{C} \tag{II.2.4}$$

$$r_c \geq \delta_{in}\pi_{in}, \qquad \forall (i,n)\in\mathcal{I}_c\times\mathcal{N}_c, \forall c\in\mathcal{C} \tag{II.2.5}$$

$$r_c \leq \delta_{in} + (r_c^{\max} - \delta_{in})\pi_{in}, \qquad \forall (i,n)\in\mathcal{I}_c\times(\mathcal{N}\setminus\mathcal{N}_c), \forall c\in\mathcal{C} \tag{II.2.6}$$

$$\sum_{i\in\mathcal{I}_c}\sum_{n\in\mathcal{N}_c}\pi_{in}z_i \geq \lceil\lambda_c|\mathcal{N}_c|\rceil, \qquad \forall c\in\mathcal{C} \tag{II.2.7}$$

$$\sum_{i\in\mathcal{I}_c}\sum_{n\in\mathcal{N}\setminus\mathcal{N}_c}\pi_{in}z_i \leq \lfloor\mu_c|\mathcal{N}\setminus\mathcal{N}_c|\rfloor, \qquad \forall c\in\mathcal{C} \tag{II.2.8}$$

$$r_c^{\min} \leq r_c \leq r_c^{\max}, \qquad \forall c\in\mathcal{C} \tag{II.2.9}$$

$$z_i \in \{0,1\}, \qquad \forall i\in\mathcal{I}_c, \forall c\in\mathcal{C} \tag{II.2.10}$$

$$\pi_{in} \in \{0,1\}, \qquad \forall (i,n)\in\mathcal{I}_c\times\mathcal{N}, \forall c\in\mathcal{C}. \tag{II.2.11}$$

The objective function is equal to the total number of true positive cases across all clusters minus the total number of false positive cases weighted by the trade-off parameter $\theta \geq 0$. Constraints (II.2.4) ensure that one single prototype is chosen for each cluster. Constraints (II.2.5) and (II.2.6) ensure that the decision variables π_{in} are well defined. Note that because of the shape of the objective function, for $n\in\mathcal{N}_c$, we only need to ensure that if $r_c < \delta_{in}$ then $\pi_{in} = 0$, which is done by constraint (II.2.5). For $n\in\mathcal{N}\setminus\mathcal{N}_c$, we only need to ensure that if $r_c > \delta_{in}$ then $\pi_{in} = 1$, which is done by constraints (II.2.6). Note that if $r_c = \delta_{in}$ then $\pi_{in} = 1$ for individuals inside the cluster c and $\pi_{in} = 0$ for individuals outside the cluster c. It is easy to see that constraints (II.2.7) control the true positive rate in cluster c, TPR_c, via the parameter $\lambda_c \in [0,1]$. Similarly, constraints (II.2.8) control the false positive rate in cluster c, FPR_c, via the parameter $\mu_c \in [0,1]$. Finally, constraints (II.2.9)–(II.2.11) define the nature of the decision variables. The radius of cluster c is bounded from below and above by $r_c^{\min}$ and $r_c^{\max}$, respectively. Straightforward values for these parameters are $r_c^{\min} = \min_{(i,n)\in\mathcal{I}_c\times\mathcal{N}_c, i\neq n}\delta_{in}$ and $r_c^{\max} = \max_{(i,n)\in\mathcal{I}_c\times\mathcal{N}_c}\delta_{in}$.

Note that the objective function contains the total number of true and false positive cases across all clusters, while constraints (II.2.7)–(II.2.8) allow us to control these two criteria in each cluster. These constraints can be useful when we want to prioritize how well we explain certain clusters, or when the clusters are of very different size and we want to ensure a good performance independently of their size, as we do in the numerical section for the real-world dataset.

In formulation (II.2.3)–(II.2.11), we have the product of two decision variables, i.e., π_{in} and z_i,

which makes the problem bilinear. We can obtain an equivalent MILP formulation, by applying the Fortet transformation [Fortet, 1960]. Let us introduce the new decision variable $y_{in} = \pi_{in} z_i$ and the corresponding constraints to ensure y_{in} is well-defined. The covering model can be reformulated as the following MILP

$$\max_{\mathbf{z},\boldsymbol{\pi},\mathbf{r},\mathbf{y}} \quad \sum_{c\in\mathcal{C}}\sum_{i\in\mathcal{I}_c}\sum_{n\in\mathcal{N}_c} y_{in} - \theta \sum_{c\in\mathcal{C}}\sum_{i\in\mathcal{I}_c}\sum_{n\in\mathcal{N}\setminus\mathcal{N}_c} y_{in} \tag{II.2.12}$$

$$\text{(II.2.4)} - \text{(II.2.11)}$$

$$y_{in} \leq \pi_{in}, \qquad \forall (i,n) \in \mathcal{I}_c \times \mathcal{N}, \forall c \in \mathcal{C} \tag{II.2.13}$$

$$y_{in} \leq z_i, \qquad \forall (i,n) \in \mathcal{I}_c \times \mathcal{N}, \forall c \in \mathcal{C} \tag{II.2.14}$$

$$y_{in} \geq \pi_{in} + z_i - 1, \qquad \forall (i,n) \in \mathcal{I}_c \times \mathcal{N}, \forall c \in \mathcal{C} \tag{II.2.15}$$

$$y_{in} \geq 0, \qquad \forall (i,n) \in \mathcal{I}_c \times \mathcal{N}, \forall c \in \mathcal{C}, \tag{II.2.16}$$

with $|\mathcal{I}| \times |\mathcal{N}| + |\mathcal{I}|$ binary and $|\mathcal{C}| + |\mathcal{I}| \times |\mathcal{N}|$ continuous decision variables, and $4|\mathcal{I}| \times |\mathcal{N}| + 5|\mathcal{C}|$ linear constraints.

The following result allows us to decompose the covering model into smaller subproblems.

Proposition II.2.1. *The covering model formulation is separable on the clusters.*

Proof. The objective function of the covering model consists of a summation across the clusters of the number of true positive cases minus the number of false positive cases weighted by θ. In addition, the constraints relevant to c only involve decision variables relating to c. With this, the desired result easily follows. □

We have modeled the radius of cluster c, r_c, as a continuous variable. However, it is easy to show that we only need to consider a discrete amount of values, namely, $r_c \in \{\delta_{in} : \ (i,n) \in \mathcal{I}_c \times \mathcal{N}_c\}$. Suppose that we solve the covering model for one of these values. Since the radius is fixed, the values of π_{in} are known and can be calculated in a preprocessing step, as well as the true positive cases and false positive cases associated with i if i is chosen as a prototype.

Let us denote by π^r_{in} the value of π_{in} when the radius of cluster c, r_c, is fixed to r. Let us define

$$\phi^r_{ic} = \sum_{n\in\mathcal{N}_c} \pi^r_{in},$$

$$\psi^r_{ic} = \sum_{n\in\mathcal{N}\setminus\mathcal{N}_c} \pi^r_{in}.$$

With this, the covering model for cluster c and radius $r_c = r$ can be formulated as follows:

$$\max_{\mathbf{z}} \quad \sum_{i\in\mathcal{I}_c} \phi^r_{ic} z_i - \theta \sum_{i\in\mathcal{I}_c} \psi^r_{ic} z_i \tag{II.2.17}$$

$$\text{s.t.} \quad \sum_{i \in \mathcal{I}_c} z_i = 1 \tag{II.2.18}$$

$$\sum_{i \in \mathcal{I}_c} \phi^r_{ic} z_i \geq \lceil \lambda_c |\mathcal{N}_c| \rceil, \tag{II.2.19}$$

$$\sum_{i \in \mathcal{I}_c} \psi^r_{ic} z_i \leq \lfloor \mu_c |\mathcal{N} \setminus \mathcal{N}_c| \rfloor, \tag{II.2.20}$$

$$z_i \in \{0, 1\}, \qquad \forall i \in \mathcal{I}_c. \tag{II.2.21}$$

Note that the set of candidates to prototype for cluster c, $\mathcal{I}_c$, can be reduced to $\mathcal{I}'_c \subset \mathcal{I}_c$. Some candidates can be removed because $\phi^r_{ic} < \lceil \lambda_c |\mathcal{N}_c| \rceil$ and others because $\psi^r_{ic} > \lfloor \mu_c |\mathcal{N} \setminus \mathcal{N}_c| \rfloor$. After reducing the set of candidates from $\mathcal{I}_c$ to $\mathcal{I}'_c$, we can eliminate constraints (II.2.19) and (II.2.20), and the problem is equivalent to choosing the prototype from $\mathcal{I}'_c$ with the largest $\phi^r_{ic} - \theta \psi^r_{ic}$.

To tackle large instances of the problem, i.e., with many individuals, we propose a heuristic approach based on combining our covering model with a sampling procedure from the set of individuals and/or the set of candidates to prototype. Indeed, we can sample from the set of candidates to prototype for cluster c, yielding $\tilde{\mathcal{I}}_c \subset \mathcal{I}_c$, for all c, and/or sample from the set of individuals from cluster c, yielding $\tilde{\mathcal{N}}_c \subset \mathcal{N}_c$, and solve the reduced covering model. Let z^{R}_i and r^{R}_c, $i \in \tilde{\mathcal{I}}_c$ and $c \in \mathcal{C}$, be the chosen prototypes and the chosen radii of the reduced problem if this is feasible. We can use this partial solution to find a feasible solution to the original problem, $(\mathbf{z}^{\mathrm{O}}, \boldsymbol{\pi}^{\mathrm{O}}, \mathbf{r}^{\mathrm{O}})$ with $\mathbf{z}^{\mathrm{O}}_i = \mathbf{z}^{\mathrm{R}}_i$, for all $i \in \tilde{\mathcal{I}}_c$ and $c \in \mathcal{C}$, and $\mathbf{r}^{\mathrm{O}} = \mathbf{r}^{\mathrm{R}}$, satisfying constraints (II.2.7), imposing a lower bound on TPR_c, and constraints (II.2.8), imposing an upper bound on FPR_c. Needless to say that this approach may not yield a feasible solution to the original problem, and we may need to sample more or make the values of λ_c and μ_c less restrictive.

II.3 The partitioning model

An alternative way of explaining clusters by means of prototypes is the partitioning model. In this case, each individual is assigned to exactly one prototype, namely the closest one. To do this, in addition to the z_i variables defined as before, we also need the binary variables ρ_{in} that allocate individuals to prototypes. Let ρ_{in} take on the value 1 if prototype i is the closest one to individual n from the chosen ones, and 0 otherwise. With these variables, the number of *true positive* cases in cluster c is equal to $\sum_{i \in \mathcal{I}_c} \sum_{n \in \mathcal{N}_c} \rho_{in}$ and the corresponding true positive rate

$$\mathrm{TPR}_c = \frac{\sum_{i \in \mathcal{I}_c} \sum_{n \in \mathcal{N}_c} \rho_{in}}{|\mathcal{N}_c|}, \tag{II.3.1}$$

while the number of *false positive* cases in cluster c is equal to $\sum_{i\in\mathcal{I}_c}\sum_{n\in\mathcal{N}\setminus\mathcal{N}_c}\rho_{in}$ and the corresponding false positive rate

$$\mathrm{FPR}_c = \frac{\sum_{i\in\mathcal{I}_c}\sum_{n\in\mathcal{N}\setminus\mathcal{N}_c}\rho_{in}}{|\mathcal{N}\setminus\mathcal{N}_c|}. \tag{II.3.2}$$

The partitioning model reads as follows:

$$\max_{\mathbf{z},\boldsymbol{\rho}} \quad \sum_{c\in\mathcal{C}}\sum_{i\in\mathcal{I}_c}\sum_{n\in\mathcal{N}_c}\rho_{in} - \theta\sum_{c\in\mathcal{C}}\sum_{i\in\mathcal{I}_c}\sum_{n\in\mathcal{N}\setminus\mathcal{N}_c}\rho_{in} \tag{II.3.3}$$

$$\text{s.t.} \quad \sum_{i\in\mathcal{I}_c} z_i = 1, \qquad \forall c\in\mathcal{C} \tag{II.3.4}$$

$$\sum_{j\in\mathcal{I}_c:\,\delta_{jn}\le\delta_{in}} z_j + \sum_{j\in\mathcal{I}:\,\delta_{jn}>\delta_{in}} \rho_{jn} \le 1 \qquad \forall (i,n)\in\mathcal{I}_c\times\mathcal{N}, \forall c\in\mathcal{C} \tag{II.3.5}$$

$$\rho_{in} \le z_i, \qquad \forall (i,n)\in\mathcal{I}\times\mathcal{N} \tag{II.3.6}$$

$$\sum_{i\in\mathcal{I}}\rho_{in} = 1, \qquad \forall n\in\mathcal{N} \tag{II.3.7}$$

$$\sum_{i\in\mathcal{I}_c}\sum_{n\in\mathcal{N}_c}\rho_{in} \ge \lceil\lambda_c|\mathcal{N}_c|\rceil, \qquad \forall c\in\mathcal{C} \tag{II.3.8}$$

$$\sum_{i\in\mathcal{I}_c}\sum_{n\in\mathcal{N}\setminus\mathcal{N}_c}\rho_{in} \le \lfloor\mu_c|\mathcal{N}\setminus\mathcal{N}_c|\rfloor, \qquad \forall c\in\mathcal{C} \tag{II.3.9}$$

$$z_i\in\{0,1\}, \qquad \forall i\in\mathcal{I} \tag{II.3.10}$$

$$\rho_{in}\in\{0,1\}, \qquad \forall (i,n)\in\mathcal{I}\times\mathcal{N}. \tag{II.3.11}$$

The objective function (II.3.3) is as in the covering model, as well as constraints (II.3.4) ensuring that we choose exactly one prototype for cluster c and constraints (II.3.8)-(II.3.9) controlling TPR_c and FPR_c for all $c\in\mathcal{C}$. Constraints (II.3.5) are the closest assignment constraints and reinforce Wagner and Falkson [1975] using the fact that, for each cluster, only one prototype is chosen. These constraints make sure that if individual n is assigned to a prototype, then there cannot be another prototype closer to n. Constraints (II.3.6) ensure that individuals are assigned to prototypes that have been selected. Constraints (II.3.7) impose that the model assigns each individual to a single prototype. Constraints (II.3.10)–(II.3.11) define the nature of the decision variables.

Two observations are noted on the partitioning formulation (II.3.3)-(II.3.11). First, there is a clear difference between the partitioning model and the covering model introduced in the previous section. To define the explanations in the partitioning model, we need to know the prototypes for all clusters, while with the covering model, due to its separability on the clusters, see Proposition II.2.1, we can obtain explanations for one single cluster without knowing the prototypes from other clusters. Second, in the partitioning formulation above we have chosen one prototype per

cluster. If we were to choose more than one, we will obviously need to change the right-hand side of constraints (II.3.4), as well as replace (II.3.5) by the original Wagner and Falkson [1975] constraints

$$z_i + \sum_{j \in \mathcal{I}: \delta_{in} < \delta_{jn}} \rho_{jn} \leq 1, \qquad \forall (i,n) \in \mathcal{I}_c \times \mathcal{N}, \forall c \in \mathcal{C}.$$

The following result on the objective function (II.3.3) easily follows.

Lemma II.3.1. *The objective function of the partitioning problem is equivalent to*

$$\max_{\mathbf{z},\boldsymbol{\rho}} \quad \sum_{c \in \mathcal{C}} \sum_{i \in \mathcal{I}_c} \sum_{n \in \mathcal{N}_c} \rho_{in}. \tag{II.3.12}$$

Proof. The result follows thanks to constraints (II.3.7). Indeed, the objective function in (II.3.3) can be rewritten as

$$\begin{aligned}
& \sum_{c \in \mathcal{C}} \sum_{i \in \mathcal{I}_c} \sum_{n \in \mathcal{N}_c} \rho_{in} - \theta \sum_{c \in \mathcal{C}} \sum_{i \in \mathcal{I}_c} \sum_{n \in \mathcal{N} \setminus \mathcal{N}_c} \rho_{in} = \\
= \; & \sum_{c \in \mathcal{C}} \sum_{i \in \mathcal{I}_c} \sum_{n \in \mathcal{N}_c} \rho_{in} - \theta \sum_{c \in \mathcal{C}} \sum_{i \in \mathcal{I}_c} \Big(\sum_{n \in \mathcal{N} \setminus \mathcal{N}_c} \rho_{in} + \sum_{n \in \mathcal{N}_c} \rho_{in} - \sum_{n \in \mathcal{N}_c} \rho_{in} \Big) \\
= \; & (1+\theta) \sum_{c \in \mathcal{C}} \sum_{i \in \mathcal{I}_c} \sum_{n \in \mathcal{N}_c} \rho_{in} - \theta \sum_{c \in \mathcal{C}} \sum_{i \in \mathcal{I}_c} \sum_{n \in \mathcal{N}} \rho_{in} \\
= \; & (1+\theta) \sum_{c \in \mathcal{C}} \sum_{i \in \mathcal{I}_c} \sum_{n \in \mathcal{N}_c} \rho_{in} - \theta \sum_{n \in \mathcal{N}} \sum_{i \in \mathcal{I}} \rho_{in} \\
= \; & (1+\theta) \sum_{c \in \mathcal{C}} \sum_{i \in \mathcal{I}_c} \sum_{n \in \mathcal{N}_c} \rho_{in} - \theta |\mathcal{N}|,
\end{aligned}$$

where constraints (II.3.7) have been used in the last step. Now the desired result follows removing the constant term $\theta|\mathcal{N}|$ and noting that $1 + \theta > 0$. □

The following result on the integrality of variables ρ_{in} easily follows.

Proposition II.3.1. *Suppose that $\delta_{in} \neq \delta_{jn}$ for all $i, j \in \mathcal{I}, i \neq j$, and $n \in \mathcal{N}$. Without loss of optimality, the integrality constraint* (II.3.11) *on variables ρ_{in} can be relaxed to*

$$\rho_{in} \geq 0, \quad \forall (i,n) \in \mathcal{I} \times \mathcal{N}. \tag{II.3.13}$$

Proof. Let $\mathbf{z}^F$ be a partial solution for the partitioning model satisfying (II.3.4) and (II.3.10). For each $n \in \mathcal{N}$, define $i(n) := \arg\min\{\delta_{in} \,:\, z_i^F = 1\}$ and $\rho_{jn}^F = 1$ if $j = i(n)$ and 0 otherwise. It is easy to see that $\boldsymbol{\rho}^F$ is the only feasible solution to the following constraints

$$\sum_{j \in \mathcal{I}:\, \delta_{jn} > \delta_{in}} \rho_{jn} \leq 1 - \sum_{j \in \mathcal{I}_c:\, \delta_{jn} \leq \delta_{in}} z_j^F, \qquad \forall (i,n) \in \mathcal{I}_c \times \mathcal{N}, \forall c \in \mathcal{C} \tag{II.3.14}$$

$$\rho_{in} \leq z_i^F, \qquad \forall (i,n) \in \mathcal{I} \times \mathcal{N} \tag{II.3.15}$$

$$\sum_{i \in \mathcal{I}} \rho_{in} = 1, \qquad \forall n \in \mathcal{N} \tag{II.3.16}$$

$$\rho_{in} \geq 0, \qquad \forall (i,n) \in \mathcal{I} \times \mathcal{N}, \tag{II.3.17}$$

For a given n, consider constraints (II.3.14) for $i = i(n)$. It is easy to see that the rhs is equal to 0, since $j = i(n)$ is one of the indices in the summation and by definition $z^F_{i(n)} = 1$. This means that the lhs of (II.3.14) for $i = i(n)$ should be zero, and thus $\rho^F_{jn} = 0$ for all j such that $\delta_{jn} > \delta_{ii(n)}$, which is equivalent to $j \neq i(n)$ since there are no ties in the dissimilarities. Now from (II.3.16), we have that $\rho^F_{i(n)n} = 1$. Note that this solution clearly satisfies (II.3.15) and (II.3.17), and that by construction $\rho^F_{in} \in \{0,1\}$. Consider now $(\mathbf{z}^F, \boldsymbol{\rho}^F)$. If this solution satisfies constraints (II.3.8)–(II.3.9), then $(\mathbf{z}^F, \boldsymbol{\rho}^F)$ is feasible for the partitioning model. Otherwise, there is no feasible solution for the partitioning model with $\mathbf{z} = \mathbf{z}^F$. □

With Lemma II.3.1 and Proposition II.3.1, the partitioning problem has been written as the MILP (II.3.12), (II.3.4)-(II.3.10) and (II.3.13) model, with $|\mathcal{I}|$ binary and $|\mathcal{I}| \times |\mathcal{N}|$ continuous decision variables and $2|\mathcal{I}| \times |\mathcal{N}| + 3|\mathcal{C}| + |\mathcal{N}|$ linear constraints.

To tackle large instances of the partitioning problem with many individuals, we can use a similar heuristic approach as in Section II.2. Recall that this consists on solving a reduced partitioning model by sampling in the set of individuals and/or the set of candidates to prototype. For each c, recall that $\tilde{\mathcal{I}}_c \subset \mathcal{I}_c$ is the subsample of candidates to prototype for cluster c in the reduced problem and $\tilde{\mathcal{N}_c} \subset \mathcal{N}_c$ the subsample of individuals. Let z_i^{R}, $i \in \tilde{\mathcal{I}}_c$ and $c \in \mathcal{C}$, be the chosen prototypes if the reduced problem is feasible. We can use this partial solution to find a feasible solution to the original problem, $(\mathbf{z}^{\mathrm{O}}, \boldsymbol{\rho}^{\mathrm{O}})$ with $\mathbf{z}_i^{\mathrm{O}} = \mathbf{z}_i^{\mathrm{R}}$ for all $i \in \tilde{\mathcal{I}}_c$ and $c \in \mathcal{C}$, satisfying constraints (II.3.8), imposing a lower bound on TPR_c, and constraints (II.3.9), imposing an upper bound on FPR_c. As pointed out in the previous section, this heuristic approach may not yield a feasible solution to the original problem, and we may need to sample more or make the values of λ_c and μ_c less restrictive.

II.4 Numerical results

In this section, we illustrate the quality of the cluster explanations provided by the covering and the partitioning models using both real-world data and simulated data. We measure the goodness of cluster explanations by the true positive ratio TPR_c and the false positive ratio FPR_c in each of the clusters, defined in (II.2.1) and (II.2.2) for the covering problem and in (II.3.1) and (II.3.2)

for the partitioning problem. The explanations are obtained assuming that $\lambda = \lambda_1 = \ldots = \lambda_{|\mathcal{C}|}$ and $\mu = \mu_1 = \ldots = \mu_{|\mathcal{C}|}$. This means that throughout this section, and with loss of generality, we impose the same requirements on TPR_c to all clusters, as well as on FPR_c.

We have set the parameter in the objective function of the covering model, θ, which weighs between the total number of true positive cases and false positive ones, equal to 1. This parameter does not play a role in the partitioning model as shown in Lemma II.3.1, where we have proved that this model maximizes the total number of true positive cases subject to the performance constraints on TPR_c and FPR_c. To illustrate the trade-off between TPR_c and FPR_c, we vary the parameters λ and μ on a grid in $[0, 1] \times [0, 1]$.

As real-world data, we use functional data relating to Canadian weather data, see Figure II.2 and Section II.4.1, publicly available in the *R* package *fda* [Febrero-Bande and Oviedo de la Fuente, 2012]. With this data we illustrate that our approach can generate good explanations,

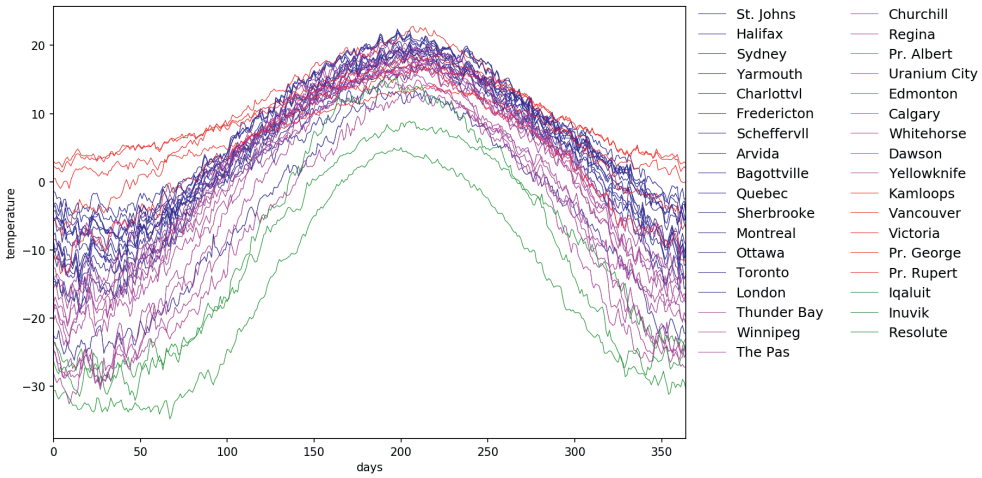


Figure II.2: The Canadian weather data to test the covering model and the partitioning one. The data is grouped into four clusters by climate's type: Atlantic - blue, Continental - pink, Pacific - red, Arctic - green. Days are along the horizontal axis, temperatures are along the vertical axis.

i.e., with high TPR_c and with low FPR_c, and that for some of the clusters we even obtain perfect explanations, i.e., with $\text{TPR}_c = 1$ and $\text{FPR}_c = 0$. Our grid results illustrate how by increasing the requirements on TPR_c through the parameter λ, we have to compromise the FPR_c of some clusters. In terms of simulated data, we use synthetic clusters in $\mathbb{R}^2$, see Figure II.3 and Section II.4.2, and illustrate how our approach achieves good explanations in terms of TPR_c and FPR_c, even for large number of individuals $|\mathcal{N}|$.

To solve the mathematical optimization models arising we use *Gurobi* [Gurobi Optimization, 2020] with *Python* [Python Core Team, 2015] on a PC Intel®Core TM i7-8665U, 16GB of RAM.

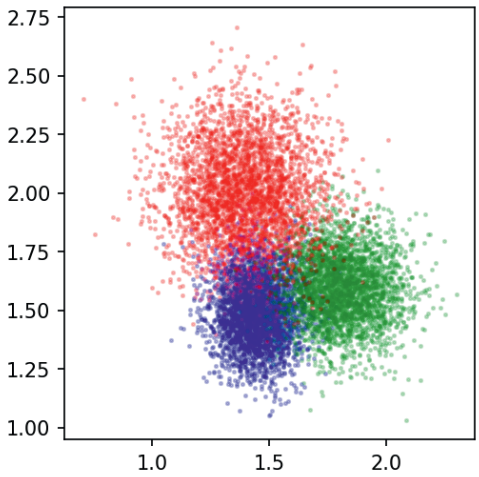


Figure II.3: Simulated data in $\mathbb{R}^2$ with three clusters to test the covering model and the partitioning one.

We have imposed a time limit of 300 seconds to each optimization model. Within this time limit, in our numerical results below, we have been able to prove optimality or to show that the problem is infeasible.

II.4.1 Results for real-world data

The Canadian weather data contains 365 days of temperature observations for $|\mathcal{N}| = 35$ cities grouped into $|\mathcal{C}| = 4$ types of climates: Atlantic ($|\mathcal{N}_{\text{Atlantic}}| = 15$), Continental ($|\mathcal{N}_{\text{Continental}}| = 12$), Pacific ($|\mathcal{N}_{\text{Pacific}}| = 5$), and Arctic ($|\mathcal{N}_{\text{Artic}}| = 3$). The data are depicted in Figure II.2, where the clusters are identified by a color, namely, blue for Atlantic, pink for Continental, red for Pacific, and green for Arctic. To build the dissimilarity measure, we use a vectorial representation of each observation with the 365 daily temperatures. We measure the dissimilarity between n and i as the Euclidean distance between the corresponding vectors of temperatures. In both the covering and the partitioning models, we consider $\mathcal{I} = \mathcal{N}$, i.e., all individuals are candidates to prototype.

To illustrate the trade-off between TPR_c and FPR_c for each cluster, we vary λ and μ on a grid in $[0, 1] \times [0, 1]$, namely, $\lambda, \mu \in \{0.0, 0.1, 0.2, \ldots, 1.0\}$. Recall that we impose the same requirements on TPR_c as well as on FPR_c to all clusters independently of their size, avoiding thus that our approach is significantly biased towards those clusters with most individuals. The results for the covering model can be found in Figure II.4, where we report the TPR_c and the FPR_c for each cluster, separately. We use a white background to denote a combination of (λ, μ) for which the corresponding model is infeasible, i.e., no explanation can be found ensuring a TPR_c of at least λ and a FPR_c of at most μ, for each of the clusters. In general, the covering model finds good

explanations, i.e., explanations that have an attractive trade-off between TPR_c and FPR_c for all the clusters. This is the case for $(\lambda, \mu) = (0.80, 0.20)$, for which $\text{TPR}_{\text{Atlantic}} = 0.80$, $\text{TPR}_{\text{Continental}} = 0.92$, $\text{TPR}_{\text{Pacific}} = 0.80$ and $\text{FPR}_{\text{Artic}} = 1.00$, while $\text{FPR}_{\text{Atlantic}} = 0.00$, $\text{FPR}_{\text{Continental}} = 0.13$, $\text{FPR}_{\text{Pacific}} = 0.03$ and $\text{FPR}_{\text{Artic}} = 0.00$.

The explanations of the covering model for $(\lambda, \mu) = (0.80, 0.20)$ are depicted in Figure II.5. In Figure II.5a we highlight in boldface the selected prototypes for each of the clusters. Figures II.5b-II.5e zoom in on each of the prototypes and the individuals explained by them (true positive and false positive), as well as the ones that should have been explained but were not (false negative). To visualize this, we use lines of the same color as the prototype to denote true positive cases; the lines with a color different from the one of the prototype denote false positive cases; while the dashed lines of the same color as the prototype denote false negative cases. For instance, in Figure II.5c, we can see that the prototype of the Continental climate cluster is Uranium City (in boldface pink), Dawson is a true positive (pink line), Inuvik is a false positive (green line), while Calgary is a false negative (dashed line in pink). We can see that the covering model can find more than one explanation for an individual, e.g., Inuvik is explained by the prototypes from the Continental and the Arctic clusters, or not explained at all, e.g., Calgary.

To end with the covering model we briefly discuss the range of values of TPR_c and FPR_c in Figure II.4. By definition, the higher the value of λ, i.e., the stricter we are on the minimum requirement on TPR_c for all clusters, the worse the FPR_c. For instance, for $\mu = 0.10$, $\text{FPR}_{\text{Continental}}$ worsens from 0.04 to 0.09 when increasing λ. Similarly, the lower the value of μ, i.e., the stricter we are on the maximum requirement on FPR_c for all clusters, the worse the TPR_c. For instance, for $\lambda = 0.70$, $\text{TPR}_{\text{Continental}}$ worsens from 0.92 to 0.75 when decreasing μ.

We now briefly discuss the results of the partitioning model for the Canadian weather data in Figure II.6. Note that in this case, the partitioning model gives for each cluster the same TPR_c and the same FPR_c for all combinations of (λ, μ) in the chosen grid for which there is a feasible solution, i.e., for $\lambda \leq 0.80$ and $\mu \geq 0.10$. More detailed information on this solution can be found in Figure II.7. There we can see that, as expected, the partitioning model gives a unique explanation for each individual.

II.4.2 Results for simulated data

In this section we consider simulated data in $\mathbb{R}^2$. The simulated data consist of three clusters, see Figure II.3 where cluster 1 is depicted in blue, cluster 2 in green, and cluster 3 in red. The coordinates of the individuals in cluster c are randomly drawn from a multivariate normal distribution,

$\mathbb{N}(\boldsymbol{\beta}^c, \Sigma^c)$, with

$$\boldsymbol{\beta}^1 = (1.45, 1.5) \qquad \boldsymbol{\beta}^2 = (1.8, 1.6) \qquad \boldsymbol{\beta}^3 = (1.4, 2.0)$$

$$\Sigma^1 = \begin{pmatrix} 0.01 & 0.00 \\ 0.00 & 0.02 \end{pmatrix} \quad \Sigma^2 = \begin{pmatrix} 0.02 & 0.00 \\ 0.00 & 0.02 \end{pmatrix} \quad \Sigma^3 = \begin{pmatrix} 0.03 & 0.00 \\ 0.00 & 0.04 \end{pmatrix}.$$

We split the individuals in $\mathcal{N}$ roughly equally across the three clusters.

The goal of this experiment is to show that our methodology is scalable, i.e., it can handle datasets with large number of individuals and it can obtain good explanations in terms of TPR_c and FPR_c for all the clusters with both the covering and the partitioning models. For this we consider instances with $|\mathcal{N}| \in \{10^4, 10^5, 10^6\}$, and we vary λ and μ on a grid in $[0,1] \times [0,1]$, namely, $\lambda \in \{0.85, 0.86, 0.87, 0.88, 0.89, 0.90\}$ and $\mu \in \{0.05, 0.06, 0.07, 0.08, 0.09, 0.10\}$.

To obtain the explanations, we apply the reduction technique described in Sections II.2 and II.3 for the covering and the partitioning models, respectively. This consists of three steps, namely, (i) defining the data for the reduced model, (ii) finding the explanations with this new model, and (iii) evaluating the quality of the explanations in the original data. When performing (i), we select $\tilde{\mathcal{N}}_c \subset \mathcal{N}_c$ using hierarchical clustering with the Euclidean distance as the dissimilarity between the individuals in $\mathcal{N}_c$. We then choose the threshold that yields $|\tilde{\mathcal{N}}_c|$ groups of individuals. From each of these groups, we choose a representative randomly, which becomes an individual of $\tilde{\mathcal{N}}_c$. The selected individuals, with weights $\tilde{w}_n$ equal to the size of their group, across the three clusters compose $\tilde{\mathcal{N}}$. We apply a similar approach to select the individuals in $\tilde{\mathcal{I}}_c \subset \mathcal{I}_c$, for each c, by using as starting point $\tilde{\mathcal{I}}_c$ and then partition it into $|\tilde{\mathcal{I}}_c|$ groups, and select a representative randomly that becomes a member of $\tilde{\mathcal{I}}_c$. In (ii), we solve the covering and the partitioning models with individuals in $\tilde{\mathcal{N}}_c$ weighted by $\tilde{w}_n$ and candidates to prototype in $\tilde{\mathcal{I}}_c$. Third, for the obtained explanations, we calculate TPR_c and FPR_c on the original dataset $\mathcal{N}$, with $|\mathcal{N}| \in \{10^4, 10^5, 10^6\}$. In the numerical results below, we take $|\tilde{\mathcal{N}}_c| = 125$ and $|\tilde{\mathcal{I}}_c| = 25$, $c = 1, 2, 3$.

We now discuss the results for the covering model, see Figures II.8 and II.9. We can see that the explanations obtained with the reduced problem show a good performance on the original dataset even when the number of individuals is very large, namely $|\mathcal{N}| = 10^6$. To illustrate this, let us start with $(\lambda, \mu) = (0.90, 0.10)$. In terms of true positive cases, for $|\mathcal{N}| \in \{10^4, 10^5, 10^6\}$, we have TPR_c equal to 0.91, 0.90, 0.90, for $c = 1, 2, 3$. In terms of false positive cases, for $|\mathcal{N}| = 10^4$, we have FPR_c equal to 0.08, 0.05, 0.05, for $c = 1, 2, 3$, while for $|\mathcal{N}| = 10^5$ and 10^6, FPR_1 worsens to 0.09. This means that with the optimal solution of the reduced problem, we have been able to find explanations to the clusters that satisfy constraints (II.2.7) for $\lambda = 0.90$ and (II.2.8) for $\mu = 0.10$. For other combinations of λ and μ, the quality of the explanations provided by the

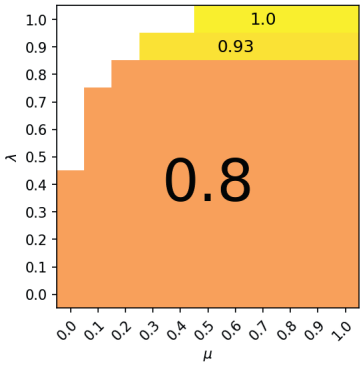
reduced problem is also good, with possible minor violations of constraints (II.2.7) or (II.2.8).

For the partitioning model, we use a similar procedure and the results can be found in Figures II.10 and II.11. We can see from those figures that the conclusions are similar.

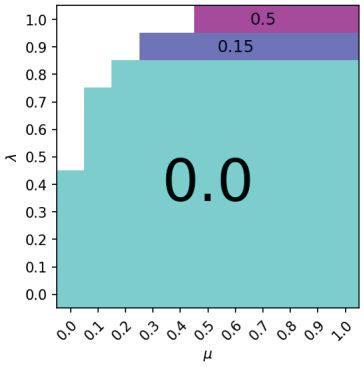
II.5 Conclusions

In this chapter, we have proposed a methodology to derive explanations for the clusters obtained from a Cluster Analysis procedure. The explanations are distance-based and defined as the set of individuals that are close to the so-called prototypes. To find explanations that are as accurate as possible, we select the prototypes that maximize the total number of true positive cases across all clusters and minimize the total number of false positive cases, while controlling the true positive rate as well as the false positive rate in each cluster. We have introduced two prototype optimization models, namely, the covering and the partitioning models. Both models can be formulated as MILPs. We have illustrated the good performance of the explanations provided by these models in terms of true positive and false positive rates using both real-world data and simulated data.

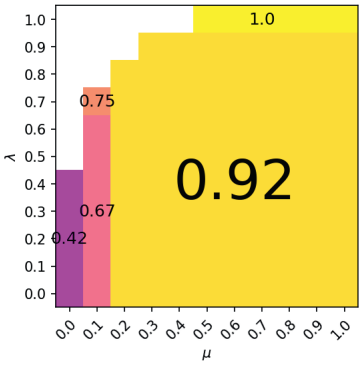
There are two interesting lines of future research. The first one is to strengthen the mathematical optimization formulations provided in this chapter. The second one is to study the problem of building the clusters and find distance-based explanations simultaneously.



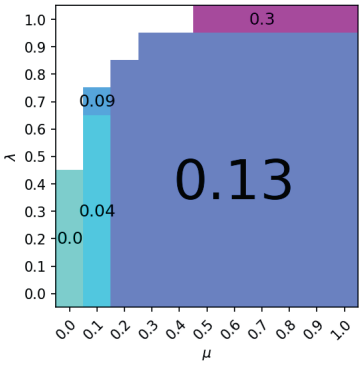
(a) $\mathrm{TPR}_{\mathrm{Atlantic}}$



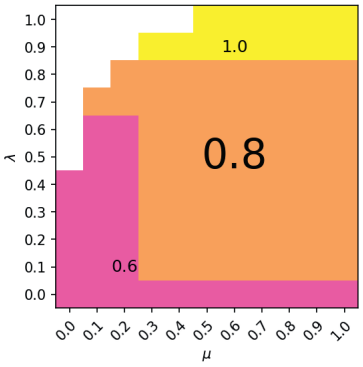
(b) $\mathrm{FPR}_{\mathrm{Atlantic}}$



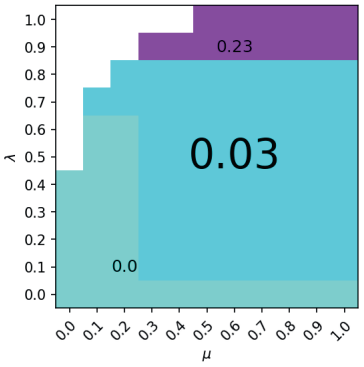
(c) $\mathrm{TPR}_{\mathrm{Continental}}$



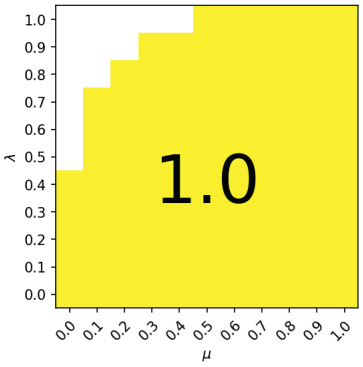
(d) $\mathrm{FPR}_{\mathrm{Continental}}$



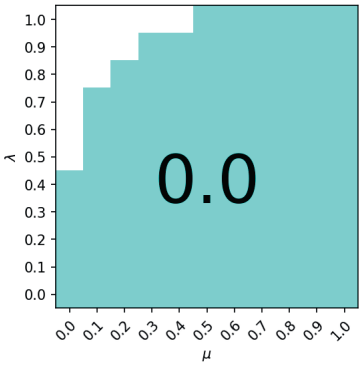
(e) $\mathrm{TPR}_{\mathrm{Pacific}}$



(f) $\mathrm{FPR}_{\mathrm{Pacific}}$

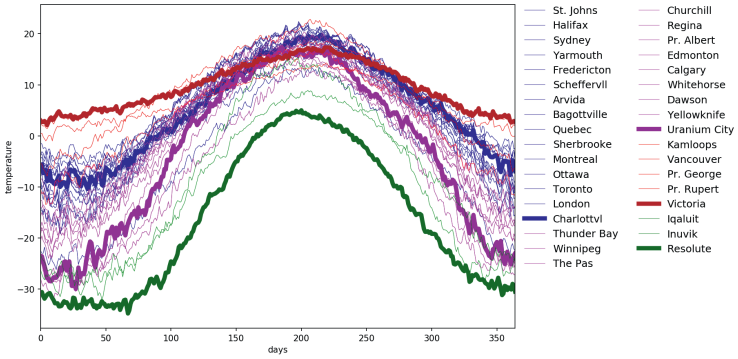


(g) $\mathrm{TPR}_{\mathrm{Arctic}}$

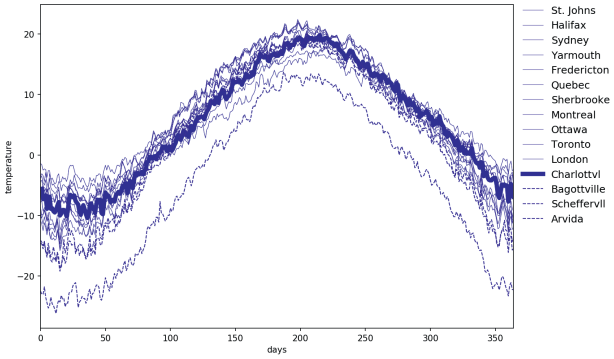


(h) $\mathrm{FPR}_{\mathrm{Arctic}}$

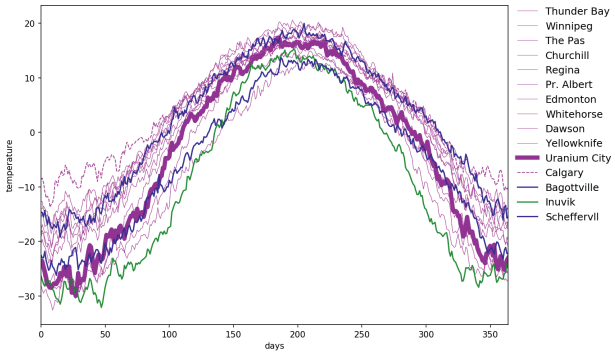
Figure II.4: For each cluster of the Canadian weather data, the true positive ratio and false positive ratio given by the covering model when λ and μ vary on a grid in $[0,1] \times [0,1]$.



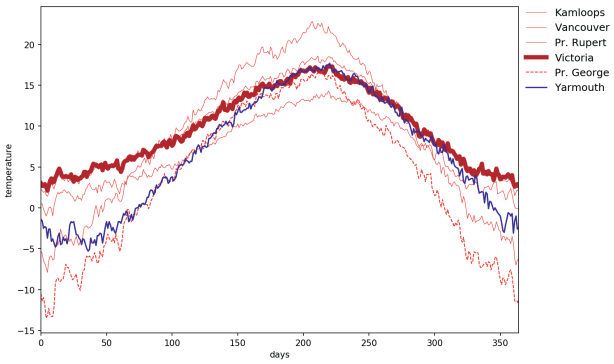
(a) The prototypes of the covering model for $\lambda = 0.80$ and $\mu = 0.20$



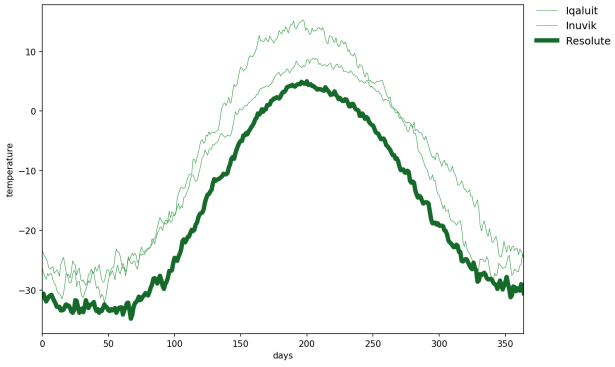
(b) $\mathrm{TPR}_{\mathrm{Atlantic}} = 0.80$, $\mathrm{FPR}_{\mathrm{Atlantic}} = 0.00$



(c) $\mathrm{TPR}_{\mathrm{Continental}} = 0.92$, $\mathrm{FPR}_{\mathrm{Continental}} = 0.13$

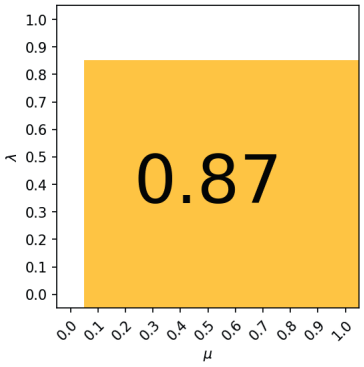


(d) $\mathrm{TPR}_{\mathrm{Pacific}} = 0.80$, $\mathrm{FPR}_{\mathrm{Pacific}} = 0.03$

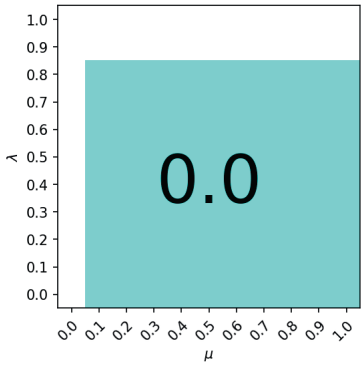


(e) $\mathrm{TPR}_{\mathrm{Artic}} = 1.00$, $\mathrm{FPR}_{\mathrm{Artic}} = 0.00$

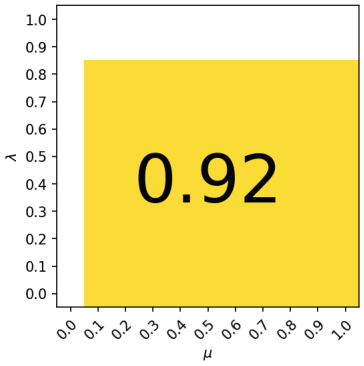
Figure II.5: The chosen prototypes for the Canadian weather dataset highlighted in boldface, with $\lambda = 0.80$ and $\mu = 0.20$, for the covering model. The lines of the same color as the cluster denote true positive cases; the lines of color different from the one of the cluster denote false positive cases; the dashed lines of the same color as the cluster denote false negative cases.



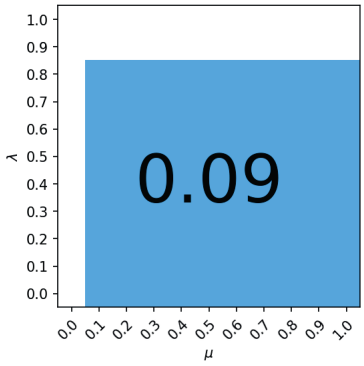
(a) $\mathrm{TPR}_{\mathrm{Atlantic}}$



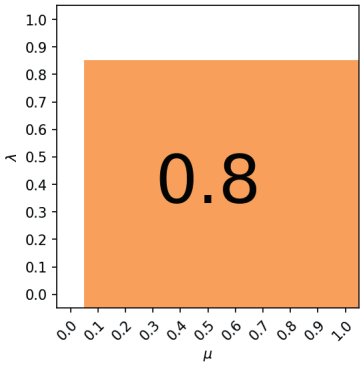
(b) $\mathrm{FPR}_{\mathrm{Atlantic}}$



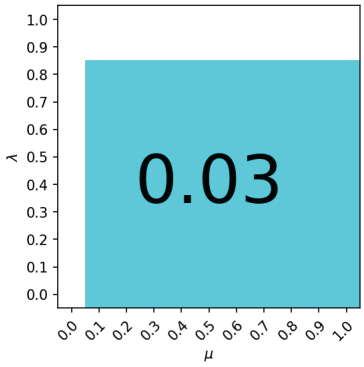
(c) $\mathrm{TPR}_{\mathrm{Continental}}$



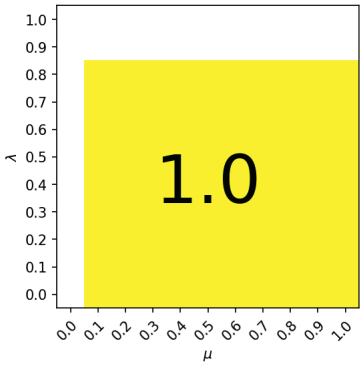
(d) $\mathrm{FPR}_{\mathrm{Continental}}$



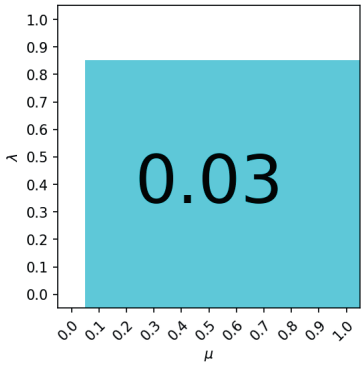
(e) $\mathrm{TPR}_{\mathrm{Pacific}}$



(f) $\mathrm{FPR}_{\mathrm{Pacific}}$

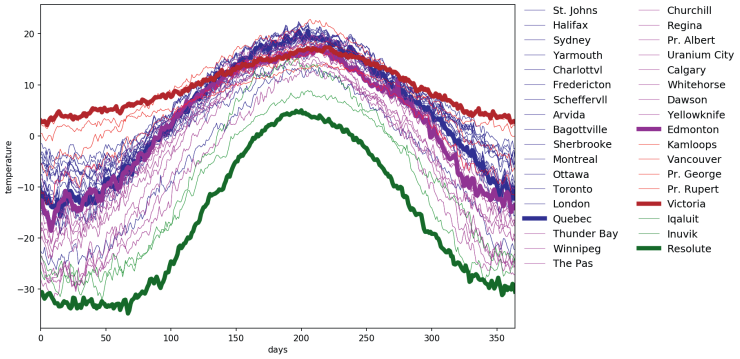


(g) $\mathrm{TPR}_{\mathrm{Arctic}}$

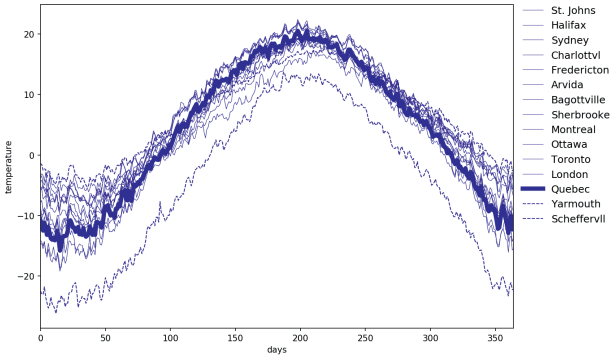


(h) $\mathrm{FPR}_{\mathrm{Arctic}}$

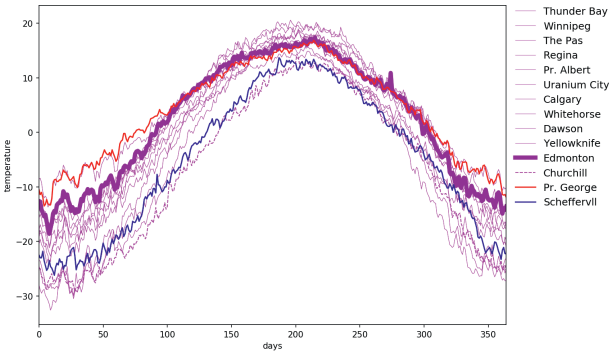
Figure II.6: For each cluster of the Canadian weather data, the true positive ratio and false positive ratio given by the partitioning model when λ and μ vary on a grid in $[0,1] \times [0,1]$.



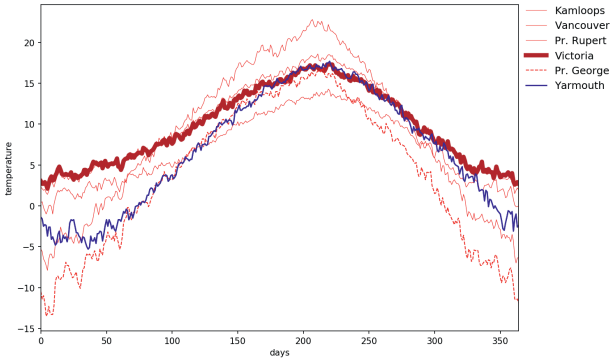
(a) The prototypes of the partitioning model for $\lambda = 0.80$ and $\mu = 0.10$



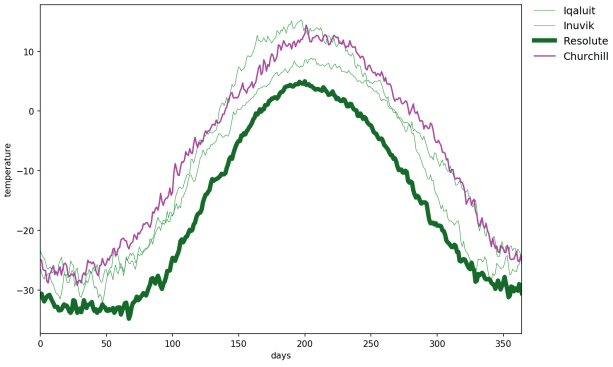
(b) $\mathrm{TPR}_{\mathrm{Atlantic}} = 0.87$, $\mathrm{FPR}_{\mathrm{Atlantic}} = 0.00$



(c) $\mathrm{TPR}_{\mathrm{Continental}} = 0.92$, $\mathrm{FPR}_{\mathrm{Continental}} = 0.09$

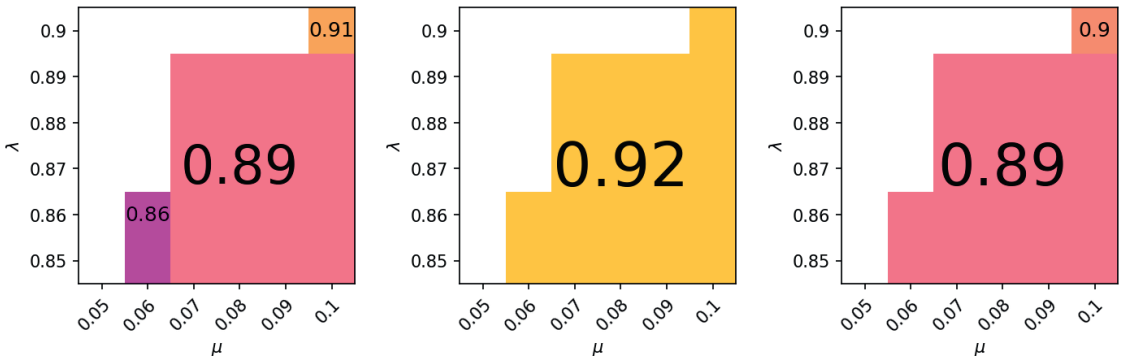


(d) $\mathrm{TPR}_{\mathrm{Pacific}} = 0.80$, $\mathrm{FPR}_{\mathrm{Pacific}} = 0.03$

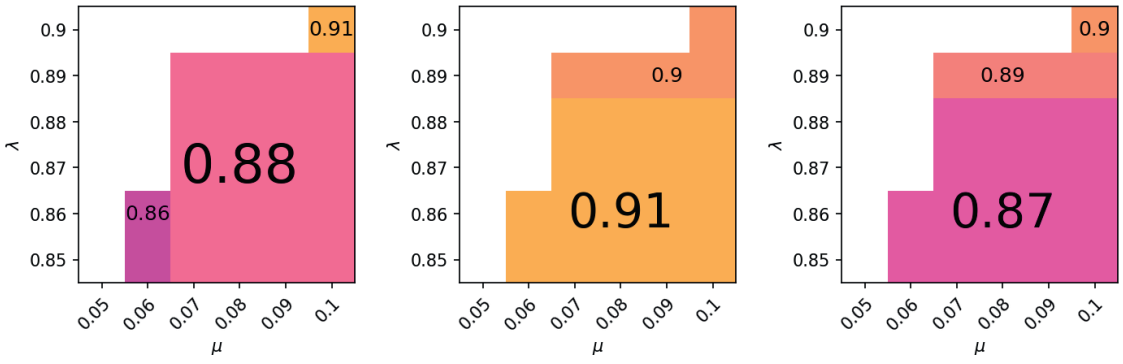


(e) $\mathrm{TPR}_{\mathrm{Artic}} = 1.00$, $\mathrm{FPR}_{\mathrm{Artic}} = 0.03$

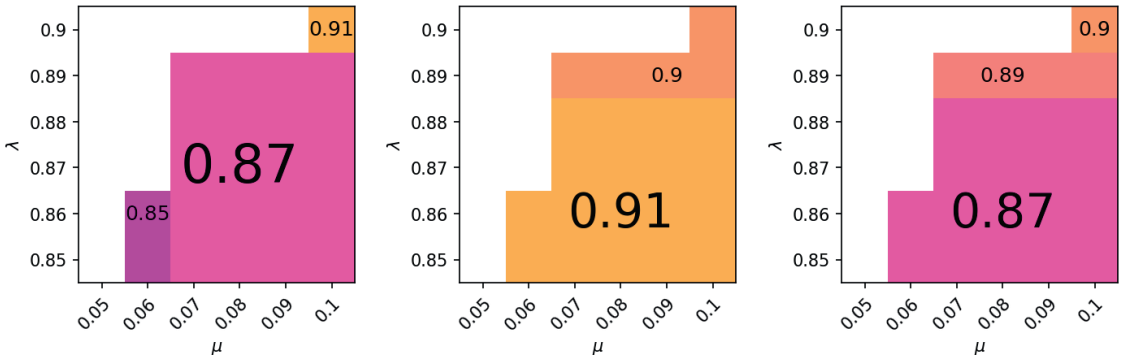
Figure II.7: The chosen prototypes for the Canadian weather dataset highlighted in boldface, with $\lambda = 0.80$ and $\mu = 0.10$, for the partitioning model. The lines of the same color as the cluster denote true positive cases; the lines of color different from the one of the cluster denote false positive cases; the dashed lines of the same color as the cluster denote false negative cases.



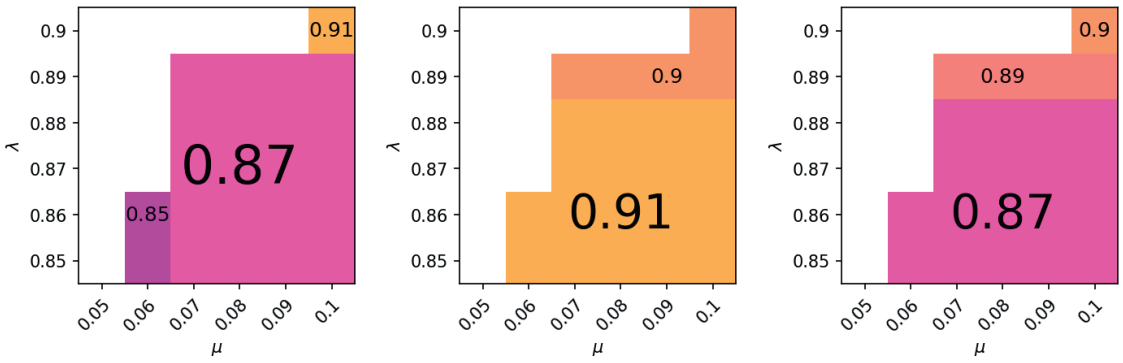
(a) $|\tilde{\mathcal{N}}| = 375$, $\mathrm{TPR}_c, c = 1, 2, 3$.



(b) $|\mathcal{N}| = 10^4$, $\mathrm{TPR}_c, c = 1, 2, 3$.

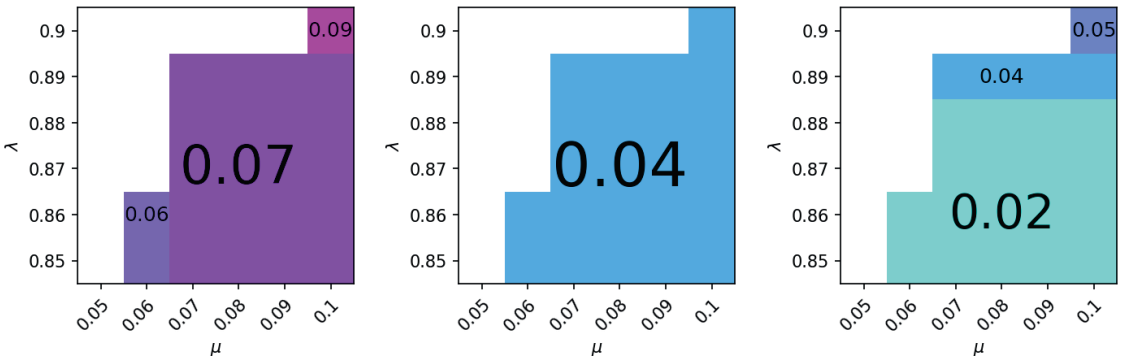


(c) $|\mathcal{N}| = 10^5$, $\mathrm{TPR}_c, c = 1, 2, 3$.

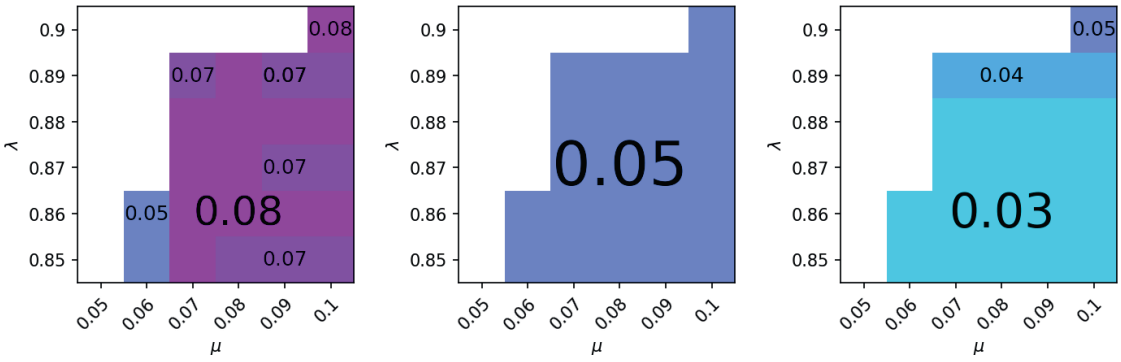


(d) $|\mathcal{N}| = 10^6$, $\mathrm{TPR}_c, c = 1, 2, 3$.

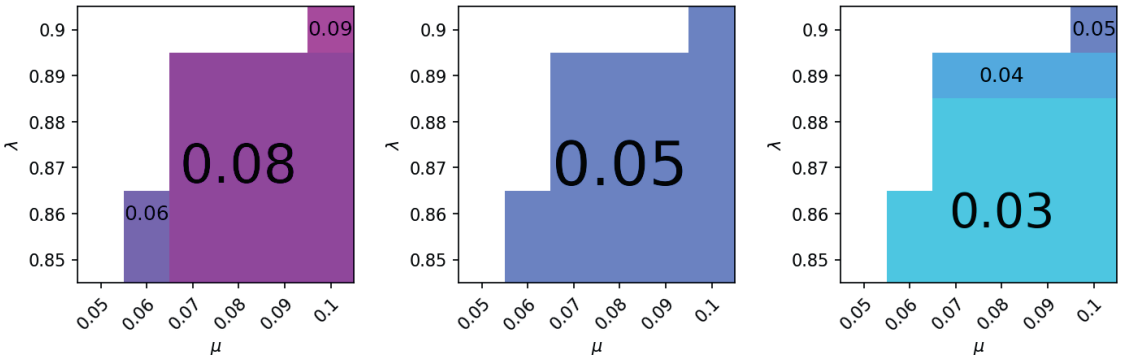
Figure II.8: For each cluster of the simulated data, the true positive ratio given by the covering model when λ and μ vary on a grid in $[0.85, 0.90] \times [0.05, 0.10]$, for the reduced problem as well as the original problem with $|\mathcal{N}| \in \{10^4, 10^5, 10^6\}$.



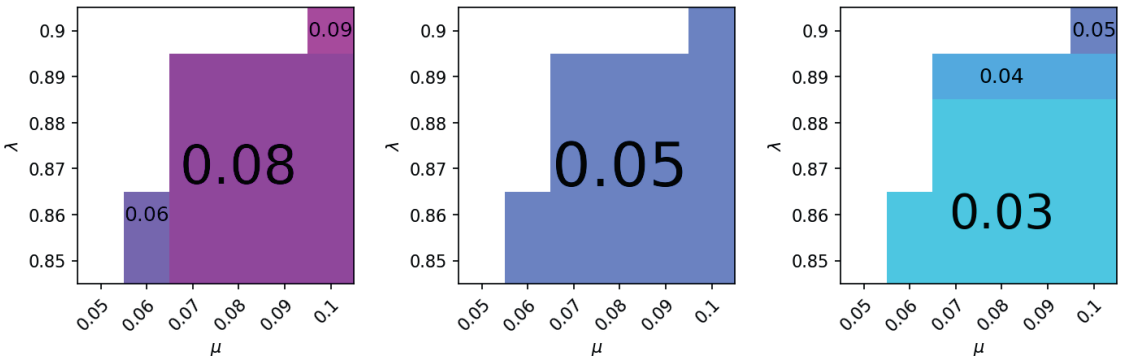
(a) $|\tilde{\mathcal{N}}| = 375$, $\mathrm{FPR}_c, c = 1, 2, 3$.



(b) $|\mathcal{N}| = 10^4$, $\mathrm{FPR}_c, c = 1, 2, 3$.

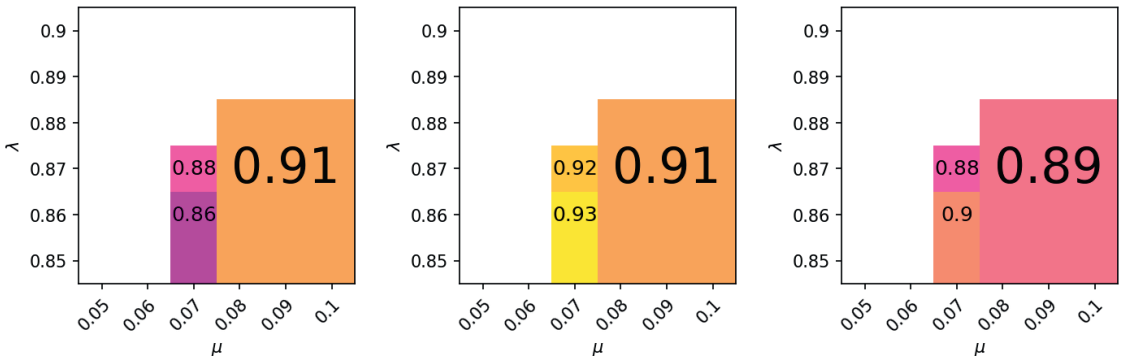


(c) $|\mathcal{N}| = 10^5$, $\mathrm{FPR}_c, c = 1, 2, 3$.

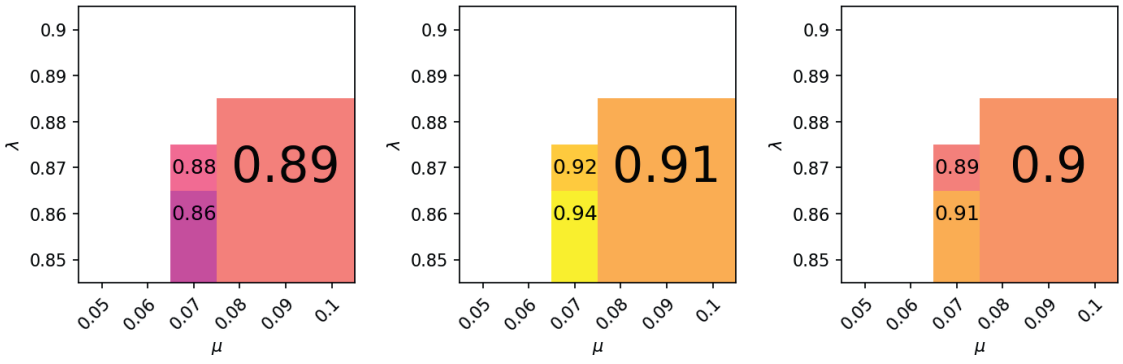


(d) $|\mathcal{N}| = 10^6$, $\mathrm{FPR}_c, c = 1, 2, 3$.

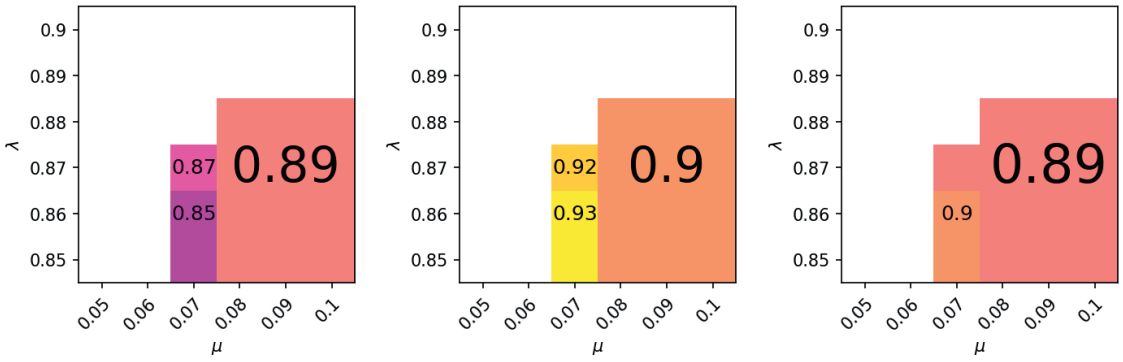
Figure II.9: For each cluster of the simulated data, the false positive ratio given by the covering model when λ and μ vary on a grid in $[0.85, 0.90] \times [0.05, 0.10]$, for the reduced problem as well as the original problem with $|\mathcal{N}| \in \{10^4, 10^5, 10^6\}$.



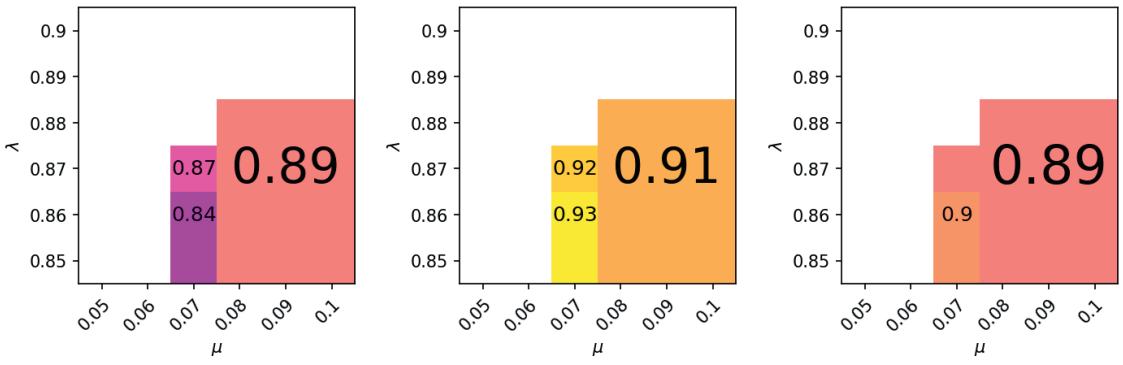
(a) $|\tilde{\mathcal{N}}| = 375$, $\mathrm{TPR}_c, c = 1, 2, 3$.



(b) $|\mathcal{N}| = 10^4$, $\mathrm{TPR}_c, c = 1, 2, 3$.

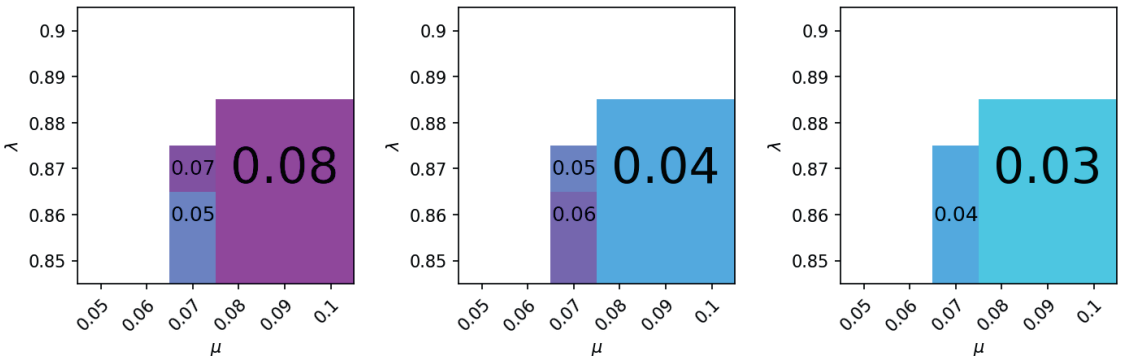


(c) $|\mathcal{N}| = 10^5$, $\mathrm{TPR}_c, c = 1, 2, 3$.

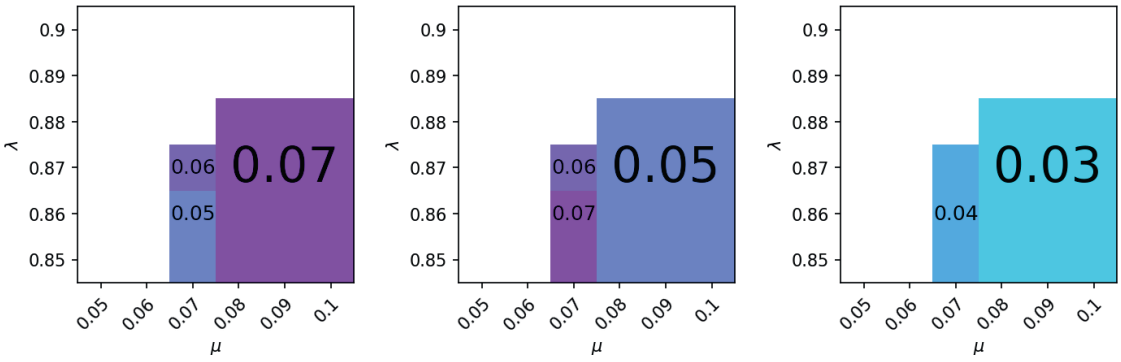


(d) $|\mathcal{N}| = 10^6$, $\mathrm{TPR}_c, c = 1, 2, 3$.

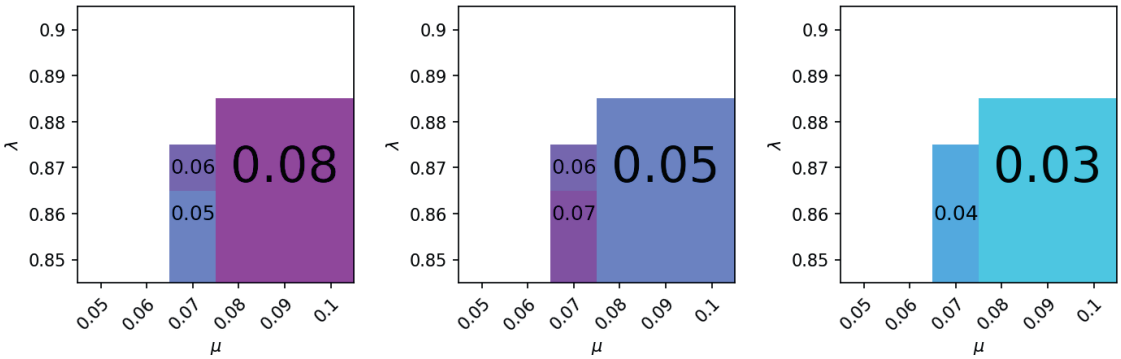
Figure II.10: For each cluster of the simulated data, the true positive ratio given by the partitioning model when λ and μ vary on a grid in $[0.85, 0.90] \times [0.05, 0.10]$, for the reduced problem as well as the original problem with $|\mathcal{N}| \in \{10^4, 10^5, 10^6\}$.



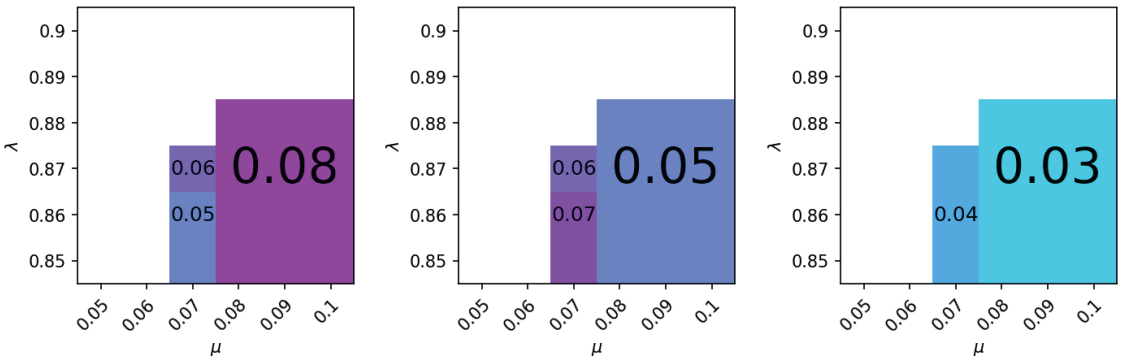
(a) $|\tilde{\mathcal{N}}| = 375$, $\mathrm{FPR}_c, c = 1, 2, 3$.



(b) $|\mathcal{N}| = 10^4$, $\mathrm{FPR}_c, c = 1, 2, 3$.



(c) $|\mathcal{N}| = 10^5$, $\mathrm{FPR}_c, c = 1, 2, 3$.



(d) $|\mathcal{N}| = 10^6$, $\mathrm{FPR}_c, c = 1, 2, 3$.

Figure II.11: For each cluster of the simulated data, the false positive ratio given by the partitioning model when λ and μ vary on a grid in $[0.85, 0.90] \times [0.05, 0.10]$, for the reduced problem as well as the original problem with $|\mathcal{N}| \in \{10^4, 10^5, 10^6\}$.

Chapter III

On clustering and interpreting with rules by means of mathematical optimization

III.1 Introduction

In this chapter, 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 Chapter II [Carrizosa et al., 2022b] and in Balabaeva and Kovalchuk [2020], Davidson et al. [2018], De Koninck et al. [2017], Kauffmann et al. [2022], Lawless et al. [2022]. Our goal 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 chapter, 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. To ensure these explanations are easily understood, we join them with the *AND* operator and 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 III.3. Records in the `housing` dataset are labelled, 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 $\leq$ 13.33), while a possible one for cluster 2 would be (RM $\leq$ 6.75) AND (LSTAT $>$ 7.765), see Table III.15.

The first contribution of this chapter is to design a procedure to explain existing clusters in a post-hoc fashion with our rule-based 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 Figure III.1 for the `housing` dataset. CART, in general, provides explanations which are long with several rules joined with AND and OR operators. Linking rules by the OR operator is more difficult to understand since no conjunctive explanation is found out to explain the whole cluster. Instead, 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 chapter is a novel clustering approach

to simultaneously find clusters and a rule-based explanation for each of them.

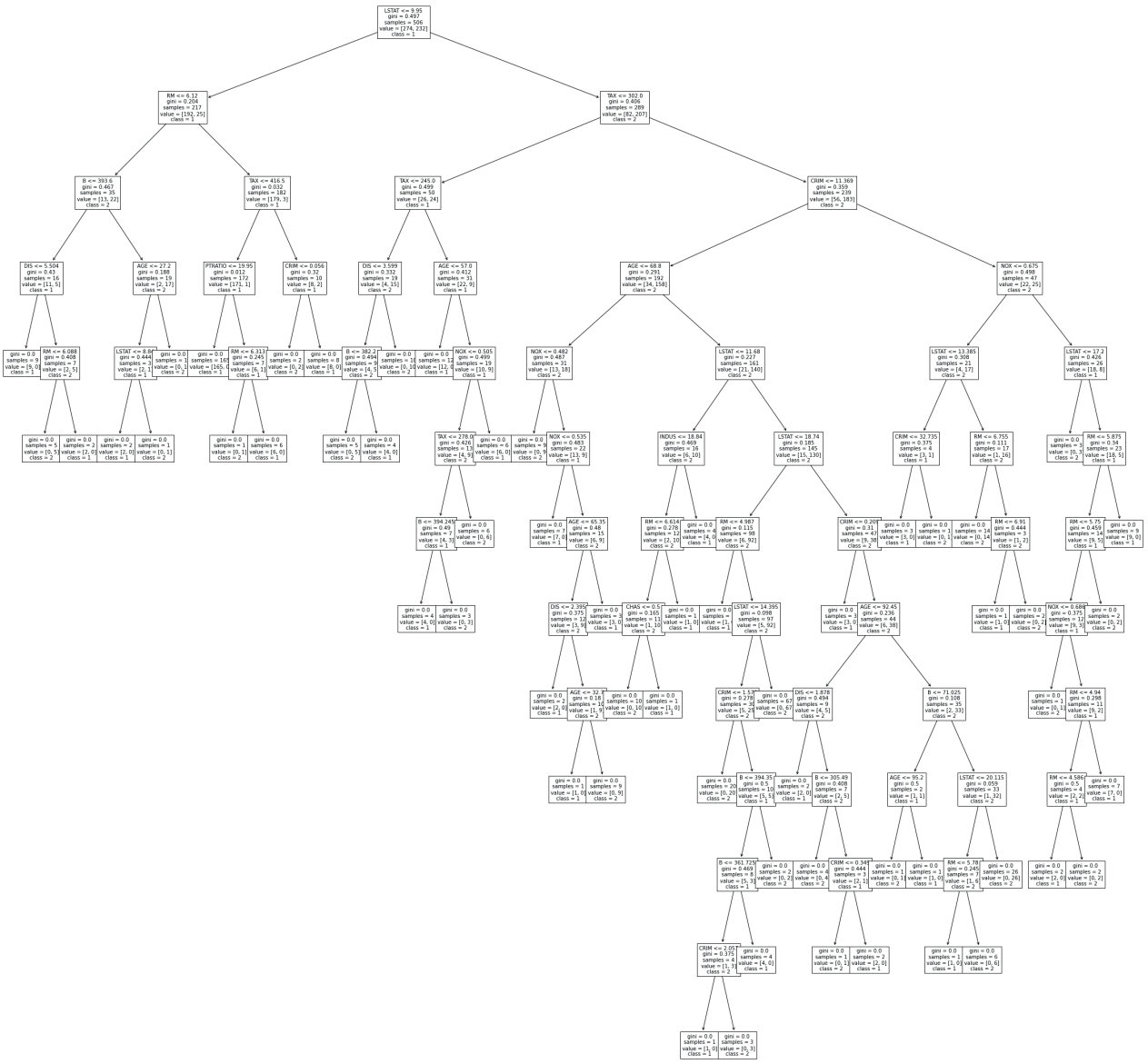


Figure III.1: The post-hoc rule-based explanations provided by CART for the `housing` dataset for clusters (classes) 1 and 2.

There is a stream of literature on approaches, where interpretability is sought by constructing unsupervised decision trees, see Basak and Krishnapuram [2005], Bertsimas et al. [2021], 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 C 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], 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 c, e_c, 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 c versus the rest, and thus e_c is not true for too many of the individuals outside the cluster. We therefore first count the number of individuals in cluster c that satisfy its explanation, i.e., the true positive cases of explanation e_c. Second, we count the number of individuals outside cluster c that satisfy explanation e_c, i.e., the false positive cases of e_c. 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 $\leq$ 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 chapter, we propose a mathematical optimization formulation for each of the problems described above. In the first formulation, we simultaneously split the individuals into C 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 chapter is organized as follows. In Section III.2, we introduce the mathematical optimization model that clusters individuals and assigns rule-based explanations to them. In Section III.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 III.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 III.5 we provide some conclusions and discuss future lines of research.

III.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, $\mathcal{N} = \cup_{s=1}^{\mathrm{S}} \mathcal{N}_s$ 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 III.3.

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

c	$\in \{1, \ldots, \mathrm{C}\}$ for clusters,

i, j	$\in \{1, \ldots, \mathrm{I}\} = \mathcal{I}$ for individuals,
s	$\in \{1, \ldots, \mathrm{S}\}$ for groups of rules,
n	$\in \{1, \ldots, \mathrm{N}\} = \mathcal{N} = \cup_{s=1}^{\mathrm{S}} \mathcal{N}_s : \mathcal{N}_s \cap \mathcal{N}_{s'} = \emptyset$ 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_{ci}	$= \begin{cases} 1, & \text{if individual } i \text{ belongs to cluster } c \\ 0, & \text{otherwise} \end{cases}$,
z_{csn}	$= \begin{cases} 1, & \text{if rule } n \in \mathcal{N}_s \text{ is chosen for cluster } c \\ 0, & \text{otherwise} \end{cases}$,
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_{ci}	$= \begin{cases} 1, & \text{if individual } i \text{ is outside cluster } c \text{ and is a false positive case to the explanation assigned to cluster } c \\ 0, & \text{otherwise} \end{cases}$,
Parameters	
$\theta_1 \geq 0$	weight for true positive cases across the C clusters,
$\theta_2 \geq 0$	weight for false positive cases across the C 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 $\mathcal{N}_s$, $s = 1, \ldots, S$:

$$\min_{\mathbf{x},\mathbf{z},\boldsymbol{\alpha},\boldsymbol{\beta}} \quad \sum_{c=1}^{\mathrm{C}} \sum_{i=1}^{\mathrm{I}-1} \sum_{j=i+1}^{\mathrm{I}} \delta_{ij} x_{ci} x_{cj} - \theta_1 \sum_{i=1}^{\mathrm{I}} \alpha_i + \theta_2 \sum_{c=1}^{\mathrm{C}} \sum_{i=1}^{\mathrm{I}} \beta_{ci} \tag{III.2.1}$$

$$\text{s.t.} \quad \sum_{c=1}^{\mathrm{C}} x_{ci} = 1, \quad i = 1 \ldots \mathrm{I} \tag{III.2.2}$$

$$\sum_{n \in \mathcal{N}_s} z_{csn} \leq 1, \quad c = 1 \ldots \mathrm{C},\ s = 1 \ldots \mathrm{S} \tag{III.2.3}$$

$$1 \leq \sum_{s=1}^{\mathrm{S}} \sum_{n \in \mathcal{N}_s} z_{csn} \leq \ell, \quad c = 1 \ldots \mathrm{C} \tag{III.2.4}$$

$$\alpha_i + x_{ci} + \sum_{n \in \mathcal{N}_s} (1 - b_{isn}) z_{csn} \leq 2, \quad i = 1 \dots \mathrm{I},\ c = 1 \dots \mathrm{C},\ s = 1 \dots \mathrm{S} \tag{III.2.5}$$

$$\beta_{ci} + x_{ci} + \sum_{s=1}^{\mathrm{S}} \sum_{n \in \mathcal{N}_s} (1 - b_{isn}) z_{csn} \geq 1, \quad i = 1 \dots \mathrm{I},\ c = 1 \dots \mathrm{C} \tag{III.2.6}$$

$$x_{ci} \in \{0, 1\}, \quad i = 1 \dots \mathrm{I},\ c = 1 \dots \mathrm{C} \tag{III.2.7}$$

$$z_{csn} \in \{0, 1\}, \quad s = 1 \dots \mathrm{S},\ n \in \mathcal{N}_s,\ c = 1 \dots \mathrm{C} \tag{III.2.8}$$

$$\alpha_i \in \{0, 1\}, \quad i = 1 \dots \mathrm{I} \tag{III.2.9}$$

$$\beta_{ci} \in \{0, 1\}, \quad i = 1 \dots \mathrm{I},\ c = 1 \dots \mathrm{C}. \tag{III.2.10}$$

The objective function (III.2.1) consists of three terms: the minimization of intra-homogeneity of clusters, the maximization of the total true 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 [Basak and Krishnapuram, 2005, Rao, 1971], 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 (III.2.2) ensure that each individual is assigned to exactly one cluster. For each cluster, constraints (III.2.3) ensure that at most one rule of group s is chosen, while constraints (III.2.4) impose that at least one rule is chosen for each cluster but no more than ℓ. Constraints (III.2.5) and (III.2.6) ensure that α_i and β_{ci} are well-defined. Because of the direction of the objective function, we only need to ensure that $\alpha_i = 0$ and $\beta_{ci} = 1$ are well-defined. Let us start with $\alpha_i = 0$ and note that $\sum_{n \in \mathcal{N}_s} (1 - b_{isn}) z_{csn} \leq 1$. Thanks to this inequality, constraints (III.2.5) are redundant if individual i does not belong to cluster c, $x_{ci} = 0$. If individual i belongs to cluster c, $x_{ci} = 1$, and it is not explained by the explanation assigned to this cluster, then for each $s, n \in \mathcal{N}_s$ such that $z_{csn} = 1$, we have that $b_{isn} = 0$. This means that $\sum_{n \in \mathcal{N}_s} (1 - b_{isn}) z_{csn} = 0$, yielding $\alpha_i \leq 0$. This, together with the fact that α_i cannot be negative, ensures that $\alpha_i = 0$. We now analyze the case of $\beta_{ic} = 1$. If individual i does not belong to cluster c, $x_{ci} = 0$, but satisfies the chosen explanation for that cluster, then $\forall s, n \in \mathcal{N}_s$ such that $z_{csn} = 1$ we have $b_{isn} = 1$. With this $\sum_{s=1}^{\mathrm{S}} \sum_{n \in \mathcal{N}_s} (1 - b_{isn}) z_{csn} = 0$, and thus $\beta_{ci} \geq 1$, which together with the upper bound on β_i, ensures that $\beta_i = 1$. Constraints (III.2.7)–(III.2.10) define the nature of the decision variables.

The following result on the integrality of variables α_i and β_{ci} easily follows.

Proposition III.2.1. *Without loss of optimality, the integrality constraints* (III.2.9)-(III.2.10) *on variables α_i and β_{ci} can be relaxed to*

$$\alpha_i \in [0, 1], \quad i = 1 \dots \mathrm{I} \tag{III.2.11}$$

$$\beta_{ci} \in [0,1], \quad i = 1 \dots \mathrm{I},\ c = 1 \dots \mathrm{C}. \tag{III.2.12}$$

Proof. This result easily follows from the discussion on constraints (III.2.5)-(III.2.6) and the direction of the optimization. Let $(\mathbf{x}^*, \mathbf{z}^*, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$ be an optimal solution to (III.2.1)-(III.2.8) and (III.2.11)-(III.2.12). Suppose that there exists $\hat{\imath}$ such that $\alpha^*_{\hat{\imath}} \in (0,1)$. It is easy to show that we can improve the objective function, which is in contradiction with the fact that $(\mathbf{x}^*, \mathbf{z}^*, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$ is an optimal solution. Since $\mathbf{x}^*$ and $\mathbf{z}^*$ are vectors of binary decision variables, $b_{isn} \in \{0,1\}$ for all i, s, n, and $\alpha^*_{\hat{\imath}}$ is fractional, we have that constraint (III.2.5) for $i = \hat{\imath}$ is not binding and thus we can increase $\alpha_{\hat{\imath}}$ to $\alpha^*_{\hat{\imath}} + \epsilon$, which is an improvement on the objective function. Similar, if there exist $\hat{c}$ and $\hat{\imath}$ such that $\beta^*_{\hat{c}\hat{\imath}} \in (0,1)$, using a similar argument, we can decrease $\beta_{\hat{c}\hat{\imath}}$ to $\beta^*_{\hat{c}\hat{\imath}} - \epsilon$, and thus improve again the objective function. □

The intra-homogeneity term contains the product of binary decision variables x_{ci} and x_{cj}, for all i, j, c. Note that these bilinear terms are different to the ones in Chapter II, as they relate to the clustering decisions, which are not present there. We linearize them by applying the Fortet transformation [Fortet, 1960]. Let $y_{cij} = x_{ci}x_{cj}$. With this the clustering and interpreting problem can be written as the following MILP formulation:

$$\begin{aligned}
\min_{\mathbf{x},\mathbf{z},\boldsymbol{\alpha},\boldsymbol{\beta},\mathbf{y}} \quad & \sum_{c=1}^{\mathrm{C}} \sum_{i=1}^{\mathrm{I}-1} \sum_{j=i+1}^{\mathrm{I}} \delta_{ij} y_{cij} - \theta_1 \sum_{i=1}^{\mathrm{I}} \alpha_i + \theta_2 \sum_{c=1}^{\mathrm{C}} \sum_{i=1}^{\mathrm{I}} \beta_{ci}, \\
\text{s.t.} \quad & (\text{III.2.2}) - (\text{III.2.8}); (\text{III.2.11}) - (\text{III.2.12}) \\
& y_{cij} \leq x_{ci}, \quad i = 1 \dots \mathrm{I}-1,\ j = i+1 \dots \mathrm{I},\ c = 1 \dots \mathrm{C} \\
& y_{cij} \leq x_{cj}, \quad i = 1 \dots \mathrm{I}-1,\ j = i+1 \dots \mathrm{I},\ c = 1 \dots \mathrm{C} \\
& y_{cij} \geq x_{ci} + x_{cj} - 1, \quad i = 1 \dots \mathrm{I}-1,\ j = i+1 \dots \mathrm{I},\ c = 1 \dots \mathrm{C} \\
& y_{cij} \geq 0, \quad i = 1 \dots \mathrm{I}-1,\ j = i+1 \dots \mathrm{I},\ c = 1 \dots \mathrm{C},
\end{aligned}$$

with $\mathrm{I} + \mathrm{C}(2 + \mathrm{S} + \mathrm{SI} + \mathrm{I} + 3\,\frac{\mathrm{I}(\mathrm{I}-1)}{2})$ linear constraints, $(\mathrm{I} + \mathrm{N})\mathrm{C}$ binary decision variables, and $\mathrm{I}(1 + \mathrm{C} + \frac{\mathrm{C}(\mathrm{I}-1)}{2})$ continuous decision variables. We will refer to this MILP formulation as (CinterP).

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.

III.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 C clusters, i.e., $\mathcal{I} = \cup_{c=1}^{\mathrm{C}} \mathcal{I}_c$ with $\mathcal{I}_c \cap \mathcal{I}_{c'}$ with $c \neq c'$. 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 $\mathcal{N}_s$, $s = 1, \ldots, \mathrm{S}$.

The decision variables z_{csn} 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_{ci} be a binary decision variable. Let us assume that i is in cluster c. The decision variable γ_{ci} is equal to 1 if individual i satisfies the explanation assigned to cluster c, and otherwise zero. For $c' \neq c$, $\gamma_{c'i}$ is equal to 1 if i satisfies the explanation chosen for cluster c' and 0 otherwise. The model for interpreting clusters $\mathcal{I}_c$, for $c = 1, \ldots, \mathrm{C}$, reads as follows:

$$\min_{\mathbf{z},\boldsymbol{\gamma}} \quad -\sum_{c=1}^{\mathrm{C}} \sum_{i \in \mathcal{I}_c} \gamma_{ci} + \theta \sum_{c=1}^{\mathrm{C}} \sum_{\substack{c'=1 \\ c \neq c'}}^{\mathrm{C}} \sum_{i \in \mathcal{I}_{c'}} \gamma_{ci} \tag{III.3.1}$$

$$\text{s.t.} \quad \sum_{n \in \mathcal{N}_s} z_{csn} \leq 1, \qquad c = 1 \ldots \mathrm{C},\ s = 1 \ldots \mathrm{S} \tag{III.3.2}$$

$$1 \leq \sum_{s=1}^{\mathrm{S}} \sum_{n \in \mathcal{N}_s} z_{csn} \leq \ell, \qquad c = 1 \ldots \mathrm{C} \tag{III.3.3}$$

$$\gamma_{ci} + \sum_{n \in \mathcal{N}_s} (1 - b_{isn}) z_{csn} \leq 1, \qquad i \in \mathcal{I}_c,\ c = 1 \ldots \mathrm{C},\ s = 1 \ldots \mathrm{S} \tag{III.3.4}$$

$$\gamma_{ci} + \sum_{s=1}^{\mathrm{S}} \sum_{n \in \mathcal{N}_s} (1 - b_{isn}) z_{csn} \geq 1, \qquad i \in \mathcal{I}_{c'},\ c, c' = 1 \ldots \mathrm{C}, c \neq c' \tag{III.3.5}$$

$$z_{csn} \in \{0, 1\}, \qquad s = 1 \ldots \mathrm{S},\ n \in \mathcal{N}_s,\ c = 1 \ldots \mathrm{C} \tag{III.3.6}$$

$$\gamma_{ci} \in \{0, 1\}, \qquad i = 1 \ldots \mathrm{I},\ c = 1 \ldots \mathrm{C}. \tag{III.3.7}$$

The objective function (III.3.1) maximizes total true positive cases and minimizes total false positive cases weighted by the parameter $\theta \geq 0$. Constraints (III.3.2)–(III.3.3) are exactly the same as constraints (III.2.3)–(III.2.4). Constraints (III.3.4)-(III.3.5) resemble constraints (III.2.5)-(III.2.6), but they are slightly different since the cluster membership is known, and ensure that γ_{ci} is well-defined. The nature of decision variables is specified in constraints (III.3.6)–(III.3.7).

In the same vein as Proposition III.2.1, the following result on the integrality of variables γ_{ci}

easily follows.

Proposition III.3.1. *Without loss of optimality, the integrality constraints* (III.3.7) *on variables* γ_{ci} *can be relaxed to*

$$\gamma_{ci} \in [0,1], \quad i = 1 \dots \mathrm{I},\ c = 1 \dots \mathrm{C} \tag{III.3.8}$$

Proof. The proof is similar to the one for Proposition III.2.1. □

With this, the problem to interpret clusters has been formulated as (III.3.1)–(III.3.6) and (III.3.8), which is an MILP model with $\mathrm{C}(\mathrm{S}+2) + \mathrm{I}(\mathrm{S}+1)$ constraints, $\mathrm{C}\,\mathrm{N}$ integer decision variables and $\mathrm{C}\,\mathrm{I}$ continuous decision variables. Hereafter, we will refer to this MILP as (InterP). In addition, the following result allows us to decompose (InterP) into smaller subproblems.

Proposition III.3.2. *(InterP) is separable on the clusters.*

Proof. The objective function of (InterP) is separable on the clusters, while the constraints relevant to c only involve decision variables relating to c. With this, the desired result easily follows. □

As mentioned in the previous section, we can incorporate 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. However, in this case, Proposition III.3.2 does not hold.

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.

III.4 Numerical results

In this section, we illustrate our methodology on well-known real-world datasets from the UCI Repository [Dua and Graff, 2017]. In Section III.4.1, we present the benchmark datasets and the rules used to build the explanations. In Section III.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 III.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 III.4.4, the impact of the source of the rules used to construct the explanations on (CinterP) and (InterP) is analyzed.

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, 16GB of RAM. For each instance of (CinterP), we impose a time limit of 600 seconds, which allows us to get solutions in which the clusters and explanations show a good trade-off 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 seconds, which is explained in part by the absence of clustering decisions in this model.

III.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 III.2–III.8. Table III.2 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 III.3–III.8 contain 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 4,000 observations.

Table III.2: Description of the datasets used to illustrate the quality of the rule-based 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 III.3: 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	$1000(\text{Bk} - 0.63)^2$ 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

Table III.4: 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)

The rules we consider in Sections III.4.2 and III.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: $\mathit{feature}_s \leq \mathit{threshold}$, $\mathit{feature}_s > \mathit{threshold}$, where *threshold* takes on the deciles of $\mathit{feature}_s$. For binary features, the two rules are defined as $\mathit{feature}_s = 1$, $\mathit{feature}_s = 0$. This choice of rules is further analyzed in Section III.4.4.

Table III.5: 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 hours 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 III.6: 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

Table III.7: 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)

III.4.2 Illustrating the clustering and interpreting model (CinterP)

The results of (CinterP) can be found in Tables III.9–III.14, 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

Table III.8: 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)

cluster.

Model (CinterP) has two parameters, θ_1 and θ_2, which are weights of 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 C-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 III.9. 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 600 seconds. 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.

Table III.9: The clusters and the rule-based 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 C = 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 $\leq$ 0.605 AND RAD $\leq$ 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 $\leq$ 0.605 AND RAD $\leq$ 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 $\leq$ 0.605 AND RAD $\leq$ 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 $\leq$ 437 AND NOX $\leq$ 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 $\leq$ 0.605 AND RAD $\leq$ 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 $\leq$ 0.605 AND RAD $\leq$ 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 $\leq$ 437 AND NOX $\leq$ 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 $\leq$ 437 AND NOX $\leq$ 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 $\leq$ 0.605 AND RAD $\leq$ 8

Table III.10: The clusters and the rule-based 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 C = 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_1	θ_2	intra-homogeneity	cluster	TPR	FPR	explanations
2^{-1}	2^{-1}	$1.73 \cdot 10^5$	1	1.00	0.00	Thickness ≤ 3
			2	1.00	0.00	Thickness > 3
2^{-1}	2^{0}	$0.67 \cdot 10^5$	1	0.97	0.02	Size ≤ 4 AND Nuclei ≤ 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 ≤ 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 ≤ 1
			2	1.00	0.00	Shape > 1
2^{0}	2^{0}	$1.24 \cdot 10^5$	1	1.00	0.00	Shape ≤ 1
			2	1.00	0.00	Shape > 1
2^{0}	2^{1}	$1.24 \cdot 10^5$	1	1.00	0.00	Shape ≤ 1
			2	1.00	0.00	Shape > 1
2^{1}	2^{-1}	$1.24 \cdot 10^5$	1	1.00	0.00	Shape ≤ 1
			2	1.00	0.00	Shape > 1
2^{1}	2^{0}	$1.24 \cdot 10^5$	1	1.00	0.00	Shape ≤ 1
			2	1.00	0.00	Shape > 1
2^{1}	2^{1}	$1.24 \cdot 10^5$	1	1.00	0.00	Shape ≤ 1
			2	1.00	0.00	Shape > 1

III.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 III.3–III.8. To make the comparison fair, we train a CART of depth 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 III.15–III.21. These two criteria are depicted in Figures III.2–III.7 for both methodologies. The CART trees can be found in Figures III.8–III.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 III.15 and Figure III.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

Table III.11: The clusters and the rule-based 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 C = 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	0.75	0.03	Pregnancies $>$ 3 AND Age $>$ 33
			2	1.00	0.05	Pregnancies $\leq$ 5 AND Age $\leq$ 42.6
2^{-1}	2^{0}	$0.48 \cdot 10^5$	1	0.72	0.01	Pregnancies $>$ 4 AND Age $>$ 33
			2	1.00	0.04	Pregnancies $\leq$ 5 AND Age $\leq$ 42.6
2^{-1}	2^{1}	$0.57 \cdot 10^5$	1	0.80	0.00	BMI $>$ 33.7
			2	1.00	0.00	BMI $\leq$ 32
2^{0}	2^{-1}	$1.22 \cdot 10^5$	1	1.00	0.00	all in
			2	–	–	–
2^{0}	2^{0}	$1.22 \cdot 10^5$	1	1.00	0.00	all in
			2	–	–	–
2^{0}	2^{1}	$1.22 \cdot 10^5$	1	1.00	0.00	all in
			2	–	–	–
2^{1}	2^{-1}	$1.22 \cdot 10^5$	1	1.00	0.00	all in
			2	–	–	–
2^{1}	2^{0}	$1.22 \cdot 10^5$	1	1.00	0.00	all in
			2	–	–	–
2^{1}	2^{1}	$1.22 \cdot 10^5$	1	1.00	0.00	all in
			2	–	–	–

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 trade-off 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 CART, with more complex explanations, namely ((LSTAT $\leq$ 9.95) AND (RM $>$ 6.12)) OR ((LSTAT $>$ 9.95) AND (TAX $\leq$ 302)) for cluster 1, with a true positive rate of 75% and false positive rate of 12%, and ((LSTAT $\leq$ 9.95) AND (RM $\leq$ 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 for 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

Table III.12: The clusters and the rule-based 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 C = 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	0.82	0.16	Length > 0.415 AND Viscera weight > 0.1435
			2	1.00	0.00	Sex = I
2^{-1}	2^{0}	$2.17 \cdot 10^5$	1	0.82	0.16	Length > 0.415 AND Viscera weight > 0.1435
			2	1.00	0.00	Sex = I
2^{-1}	2^{1}	$2.16 \cdot 10^5$	1	0.56	0.00	Sex = M
			2	0.93	0.00	Sex = I
2^{0}	2^{-1}	$2.52 \cdot 10^5$	1	0.95	0.40	Whole weight > 0.3625 AND Shell weight > 0.103
			2	1.00	0.00	Sex = I AND Length $\leq$ 0.54
2^{0}	2^{0}	$2.52 \cdot 10^5$	1	0.90	0.22	Length > 0.415 AND Viscera weight > 0.10775
			2	1.00	0.00	Sex = I AND Length $\leq$ 0.54
2^{0}	2^{1}	$2.52 \cdot 10^5$	1	0.82	0.07	Length > 0.415 AND Viscera weight > 0.1435
			2	1.00	0.00	Sex = I AND Length $\leq$ 0.54
2^{1}	2^{-1}	$2.52 \cdot 10^5$	1	0.98	0.65	Whole weight > 0.1955 AND Viscera weight > 0.04
			2	1.00	0.00	Sex = I AND Length $\leq$ 0.54
2^{1}	2^{0}	$2.52 \cdot 10^5$	1	0.95	0.40	Whole weight > 0.3625 AND Shell weight > 0.103
			2	1.00	0.00	Sex = I AND Length $\leq$ 0.54
2^{1}	2^{1}	$2.52 \cdot 10^5$	1	0.90	0.22	Length > 0.415 AND Viscera weight > 0.10775
			2	1.00	0.00	Sex = I AND Length $\leq$ 0.54

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.

III.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 III.22. For (InterP), small improvements are also reported for the most granular option, see Table III.23.

Table III.13: The clusters and the rule-based 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 C = 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.

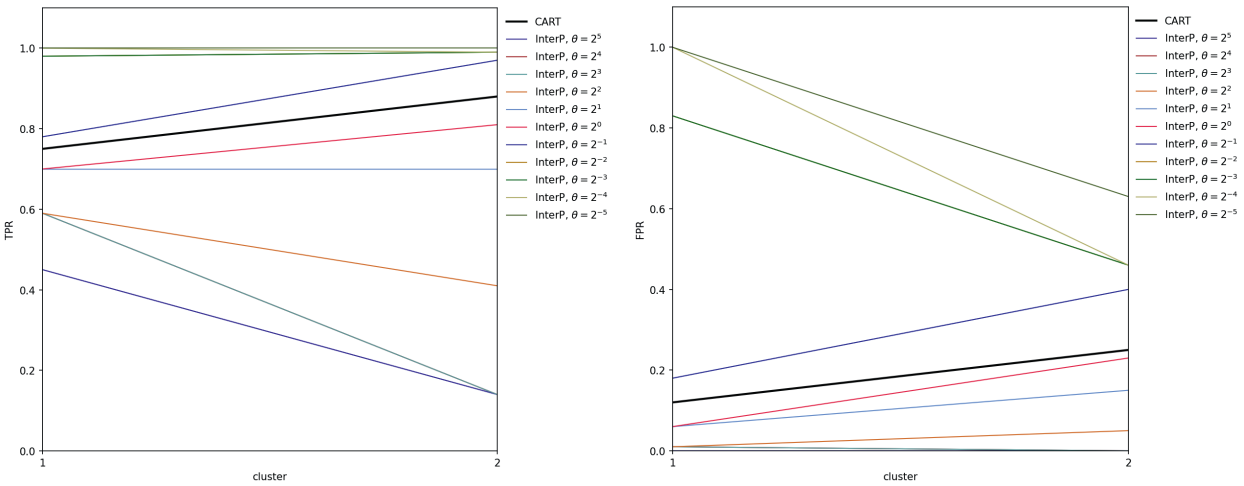
θ_1	θ_2	intra-homogeneity	cluster	TPR	FPR	explanations
2^{-1}	2^{-1}	$4.99 \cdot 10^3$	1	1.00	0.00	Ash $>$ 2.3 AND Totalphenols $>$ 1.881
			2	1.00	0.00	Ash $\leq$ 2.3 AND Totalphenols $>$ 1.881
			3	1.00	0.00	Totalphenols $\leq$ 1.881
2^{-1}	2^{0}	$5.22 \cdot 10^3$	1	1.00	0.00	Ash $\leq$ 2.61 AND Totalphenols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

Table III.14: The clusters and the rule-based 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 C = 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 $\leq$ 1.36 AND Si $\leq$ 72.132
			2	1.00	0.00	Mg $\leq$ 2.805 AND Ca $>$ 10.443
			3	0.95	0.08	K $>$ 0.492 AND Fe $\leq$ 0.128
			4	1.00	0.01	Ca $\leq$ 10.443 AND Fe $>$ 0.128
			5	0.96	0.01	Mg $\leq$ 0.6 AND Ba $>$ 0
			6	0.44	0.00	Si $\leq$ 71.773 AND Ca $\leq$ 8.6
2^{-1}	2^{0}	$9.17 \cdot 10^2$	1	0.53	0.02	Al $\leq$ 1.146 AND Si $\leq$ 72.132
			2	1.00	0.00	Mg $\leq$ 2.805 AND Ca $>$ 10.443
			3	1.00	0.04	K $>$ 0.492 AND Fe $\leq$ 0.128
			4	1.00	0.01	Ca $\leq$ 10.443 AND Fe $>$ 0.128
			5	0.91	0.00	K $\leq$ 0.08 AND Ba $>$ 0
			6	0.44	0.00	RI $\leq$ 1.51869 AND Si $\leq$ 71.773
2^{-1}	2^{1}	$8.59 \cdot 10^2$	1	0.24	0.00	Mg $>$ 3.757 AND K $\leq$ 0.19
			2	1.00	0.00	Mg $\leq$ 2.805 AND Ca $>$ 10.443
			3	1.00	0.04	K $>$ 0.492 AND Fe $\leq$ 0.07
			4	0.90	0.00	Ca $\leq$ 10.443 AND Fe $>$ 0.128
			5	0.91	0.00	K $\leq$ 0.08 AND Ba $>$ 0
			6	0.40	0.00	Si $\leq$ 71.773 AND Ca $\leq$ 8.6
2^{0}	2^{-1}	$7.79 \cdot 10^2$	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	1.00	0.15	K $>$ 0.19 AND Fe $\leq$ 0.128
			4	1.00	0.01	Ca $\leq$ 10.443 AND Fe $>$ 0.128
			5	0.92	0.01	Mg $\leq$ 0.6 AND Ba $>$ 0
			6	0.67	0.00	Si $\leq$ 72.132 AND Ca $\leq$ 7.97
2^{0}	2^{0}	$9.07 \cdot 10^2$	1	0.75	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	0.97	0.07	K $>$ 0.492 AND Fe $\leq$ 0.128
			4	1.00	0.01	Ca $\leq$ 10.443 AND Fe $>$ 0.128
			5	0.91	0.00	K $\leq$ 0.08 AND Ba $>$ 0
			6	0.60	0.00	Si $\leq$ 72.132 AND Ca $\leq$ 7.97
2^{0}	2^{1}	$9.07 \cdot 10^2$	1	0.75	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	0.97	0.07	K $>$ 0.492 AND Fe $\leq$ 0.128
			4	1.00	0.01	Ca $\leq$ 10.443 AND Fe $>$ 0.128
			5	0.91	0.00	K $\leq$ 0.08 AND Ba $>$ 0
			6	0.60	0.00	Si $\leq$ 72.132 AND Ca $\leq$ 7.97
2^{1}	2^{-1}	$7.75 \cdot 10^2$	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	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	0.96	0.02	Al $>$ 1.748 AND Ba $>$ 0
			6	0.63	0.00	RI $\leq$ 1.51735 AND Si $\leq$ 72.132
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 III.15: The clusters and the rule-based explanations provided by (InterP), $\theta \in \{2^p\}_{p=-5,\ldots,5}$, for the `housing` dataset, with C = 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	cluster	TPR	FPR	explanations
2^5	1	0.45	0.00	RM $>$ 6.376 AND LSTAT $\leq$ 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 $\leq$ 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 $\leq$ 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 $\leq$ 9.53
	2	0.41	0.05	CRIM $\leq$ 10.753 AND LSTAT $>$ 15.62
2^1	1	0.70	0.06	RM $>$ 6.086 AND LSTAT $\leq$ 11.36
	2	0.70	0.15	CRIM $\leq$ 10.753 AND LSTAT $>$ 11.36
2^0	1	0.70	0.06	RM $>$ 6.086 AND LSTAT $\leq$ 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 $\leq$ 13.33
	2	0.97	0.40	RM $\leq$ 6.75 AND LSTAT $>$ 7.765
2^{-2}	1	0.98	0.83	PTRATIO $\leq$ 20.9
	2	0.99	0.46	LSTAT $>$ 7.765
2^{-3}	1	0.98	0.83	PTRATIO $\leq$ 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 $\leq$ 9.95 AND RM $>$ 6.12 OR LSTAT $>$ 9.95 AND TAX $\leq$ 302
	2	0.88	0.25	LSTAT $\leq$ 9.95 AND RM $\leq$ 6.12 OR LSTAT $>$ 9.95 AND TAX $>$ 302



(a) True Positive Rate

(b) False Positive Rate

Figure III.2: The `housing` data: the interpretability results obtained with rule-based explanations given by (InterP).

Table III.16: The clusters and the rule-based explanations provided by (InterP), $\theta \in \{2^p\}_{p=-5,\dots,5}$, for the `breast cancer` dataset, with C = 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

Table III.17: The clusters and the rule-based explanations provided by (InterP), $\theta \in \{2^p\}_{p=-5,\dots,5}$, for the `PIMA` dataset, with C = 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 $\leq$ 102 AND BMI $\leq$ 25.9
	2	0.04	0.00	Glucose $>$ 167 AND SkinThickness $>$ 40
2^4	1	0.19	0.00	Glucose $\leq$ 102 AND BMI $\leq$ 28.2
	2	0.04	0.00	Glucose $>$ 167 AND SkinThickness $>$ 40
2^3	1	0.30	0.02	BMI $\leq$ 30.1 AND Age $\leq$ 27
	2	0.10	0.00	Glucose $>$ 167 AND SkinThickness $>$ 31
2^2	1	0.45	0.07	Glucose $\leq$ 117 AND Age $\leq$ 29
	2	0.23	0.02	Glucose $>$ 167 AND BMI $>$ 28.2
2^1	1	0.68	0.25	Pregnancies $\leq$ 7 AND Glucose $\leq$ 125
	2	0.23	0.02	Glucose $>$ 167 AND BMI $>$ 28.2
2^0	1	0.88	0.49	Glucose $\leq$ 147 AND BMI $\leq$ 41.5
	2	0.55	0.13	Glucose $>$ 125 AND BMI $>$ 30.1
2^{-1}	1	0.98	0.76	Glucose $\leq$ 167
	2	0.68	0.23	Glucose $>$ 117 AND BMI $>$ 28.2
2^{-2}	1	0.98	0.76	Glucose $\leq$ 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 $\leq$ 127.5 OR Glucose $>$ 127.5 AND BMI $\leq$ 29.95
	2	0.56	0.23	Glucose $\leq$ 127.5 AND BMI $>$ 29.95

Table III.18: The clusters and the rule-based explanations provided by (InterP), $\theta \in \{2^p\}_{p=-5,\ldots,5}$, for the `abalone` dataset, with C = 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 $\leq$ 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 $\leq$ 0.085
	2	0.20	0.00	Shell weight $>$ 0.41125
2^3	1	0.34	0.01	Sex = I AND Height $\leq$ 0.105
	2	0.20	0.00	Shell weight $>$ 0.41125
2^2	1	0.50	0.04	Sex = I AND Height $\leq$ 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 $\leq$ 0.135
	2	0.65	0.14	Diameter $>$ 0.4 AND Shell weight $>$ 0.268
2^0	1	0.71	0.18	Height $\leq$ 0.14 AND Shell weight $\leq$ 0.23475
	2	0.76	0.23	Diameter $>$ 0.365 AND Shell weight $>$ 0.23475
2^{-1}	1	0.88	0.41	Height $\leq$ 0.16 AND Shell weight $\leq$ 0.3065
	2	0.86	0.34	Whole weight $>$ 0.521 AND Shell weight $>$ 0.19
2^{-2}	1	1.00	0.74	Height $\leq$ 0.185 AND Shell weight $\leq$ 0.41125
	2	0.97	0.63	Whole weight $>$ 0.1955 AND Shell weight $>$ 0.103
2^{-3}	1	1.00	0.74	Height $\leq$ 0.185 AND Shell weight $\leq$ 0.41125
	2	1.00	0.78	Whole weight $>$ 0.1955 AND Viscera weight $>$ 0.04
2^{-4}	1	1.00	0.74	Height $\leq$ 0.185 AND Shell weight $\leq$ 0.41125
	2	1.00	0.80	Whole weight $>$ 0.1955 AND all in
2^{-5}	1	1.00	0.74	Height $\leq$ 0.185 AND Shell weight $\leq$ 0.41125
	2	1.00	0.80	Whole weight $>$ 0.1955
CART	1	0.73	0.27	Shell weight $\leq$ 0.217
	2	0.8	0.2	Shell weight $>$ 0.217

Table III.19: The clusters and the rule-based explanations provided by (InterP), $\theta \in \{2^p\}_{p=-5,\ldots,5}$, for the `wine` dataset, with C = 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.

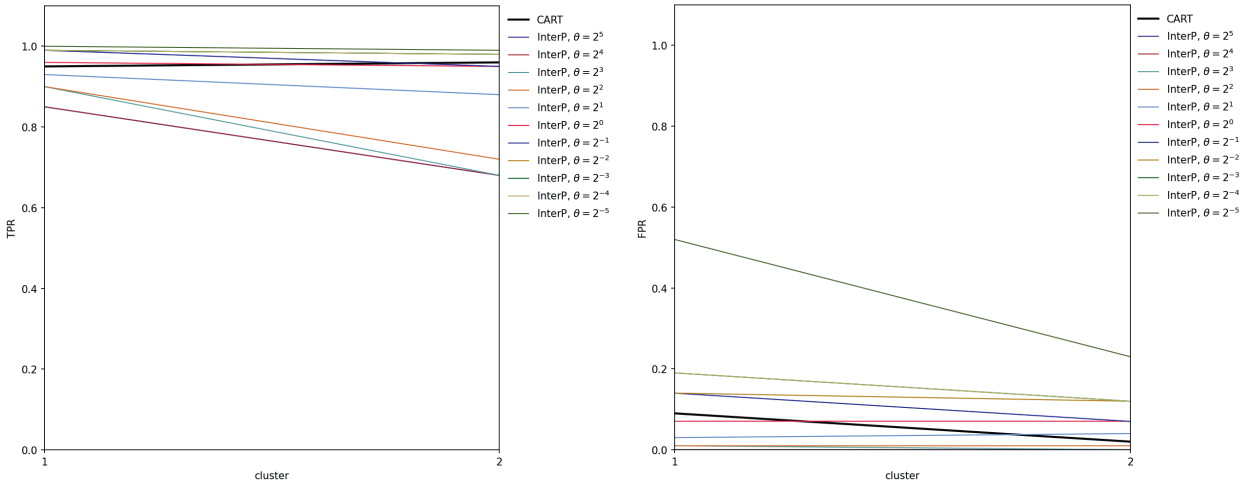
θ	cluster	TPR	FPR	explanations
2^5	1	0.78	0.00	Alcohol > 13.05 AND Proline > 879
	2	0.77	0.00	Colorintensity ≤ 3.4
	3	0.90	0.00	Flavanoids ≤ 1.324 AND Colorintensity > 4.08
2^4	1	0.78	0.00	Alcohol > 13.05 AND Proline > 879
	2	0.77	0.00	Colorintensity ≤ 3.4
	3	0.90	0.00	Flavanoids ≤ 1.324 AND Colorintensity > 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
				Proline > 755.0 AND Flavanoids ≤ 2.165
	3	0.96	0.02	OR Proline ≤ 755.0 AND OD280andOD31ofdilutedwines ≤ 2.115

Table III.20: The clusters and the rule-based explanations provided by (InterP), $\theta \in \{2^p\}_{p=-5,\ldots,5}$, for the `glass` dataset, with C = 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 $\leq$ 1.5163 AND Fe $>$ 0.22
	2	0.14	0.00	Mg $>$ 3.757 AND Ca $\leq$ 8.6
	3	0.06	0.00	Na $>$ 14.018 AND Fe $>$ 0.22
	4	0.23	0.00	RI $\leq$ 1.51591 AND Si $\leq$ 71.773
	5	0.22	0.00	K $\leq$ 0 AND Ca $\leq$ 7.97
	6	0.79	0.00	Na $>$ 14.018 AND Ba $>$ 0
2^4	1	0.06	0.00	RI $\leq$ 1.5163 AND Fe $>$ 0.22
	2	0.14	0.00	Mg $>$ 3.757 AND Ca $\leq$ 8.6
	3	0.06	0.00	Na $>$ 14.018 AND Fe $>$ 0.22
	4	0.23	0.00	RI $\leq$ 1.51591 AND Si $\leq$ 71.773
	5	0.22	0.00	K $\leq$ 0 AND Ca $\leq$ 7.97
	6	0.79	0.00	Na $>$ 14.018 AND Ba $>$ 0
2^3	1	0.06	0.00	RI $\leq$ 1.5163 AND Fe $>$ 0.22
	2	0.14	0.00	Mg $>$ 3.757 AND Ca $\leq$ 8.6
	3	0.06	0.00	Na $>$ 14.018 AND Fe $>$ 0.22
	4	0.23	0.00	RI $\leq$ 1.51591 AND Si $\leq$ 71.773
	5	0.22	0.00	K $\leq$ 0 AND Ca $\leq$ 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 $\leq$ 8.6
	3	0.06	0.00	Na $>$ 14.018 AND Fe $>$ 0.22
	4	0.23	0.00	RI $\leq$ 1.51591 AND Si $\leq$ 71.773
	5	0.22	0.00	K $\leq$ 0 AND Ca $\leq$ 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 $\leq$ 8.12
	3	0.06	0.00	Na $>$ 14.018 AND Fe $>$ 0.22
	4	0.23	0.00	RI $\leq$ 1.51591 AND Si $\leq$ 71.773
	5	0.22	0.00	K $\leq$ 0 AND Ca $\leq$ 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 $\leq$ 8.339
	3	0.06	0.00	Na $>$ 14.018 AND Fe $>$ 0.22
	4	0.23	0.00	RI $\leq$ 1.51591 AND Si $\leq$ 71.773
	5	0.67	0.01	Si $\leq$ 72.79 AND K $\leq$ 0
	6	0.79	0.00	Na $>$ 14.018 AND Ba $>$ 0

Table III.21: The clusters and the rule-based explanations provided by (InterP), $\theta \in \{2^p\}_{p=-5,\ldots,5}$, for the `glass` dataset, with C = 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.)

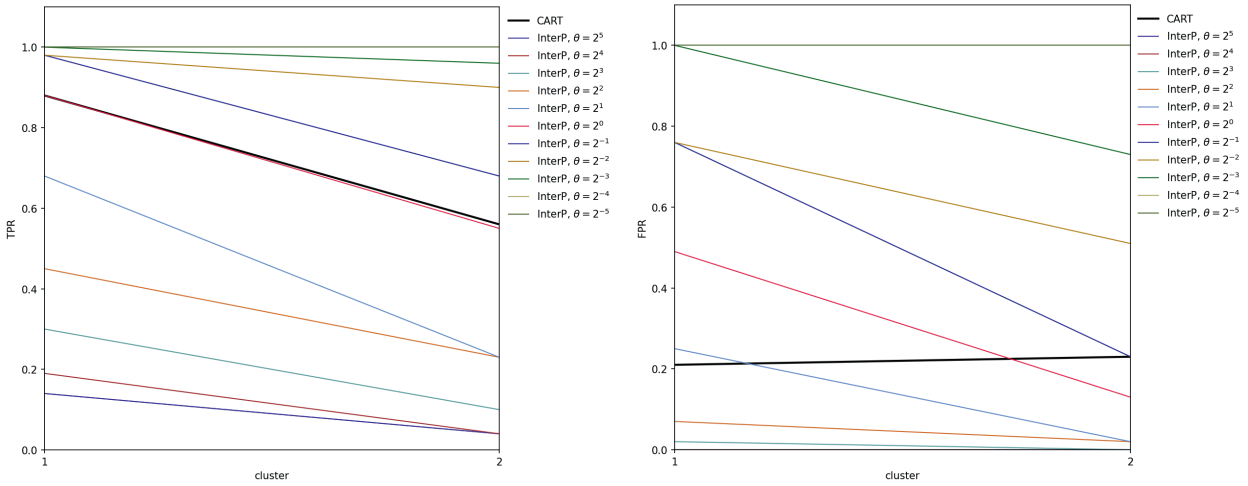
θ	cluster	TPR	FPR	explanations
2^{-1}	1	0.86	0.23	RI > 1.51735 AND Mg > 2.805
	2	0.62	0.20	Mg > 2.805 AND Ca ≤ 8.482
	3	0.12	0.01	Na > 13.3 AND Fe > 0.22
	4	0.92	0.05	Na ≤ 13.44 AND Mg ≤ 2.805
	5	1.00	0.02	K ≤ 0 AND Ba ≤ 0
	6	0.90	0.02	Na > 13.3 AND Al > 1.748
2^{-2}	1	0.93	0.35	Al ≤ 1.488 AND Ca ≤ 10.443
	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



(a) True Positive Rate

(b) False Positive Rate

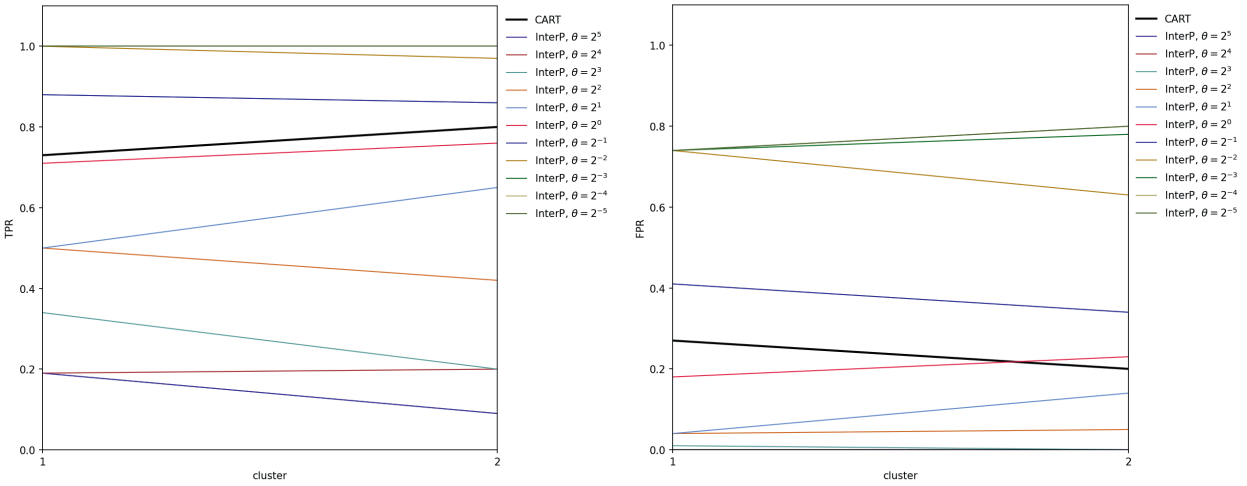
Figure III.3: The `breast cancer` data: the post-hoc interpretability results obtained with rule-based explanations given by (InterP) and CART.



(a) True Positive Rate

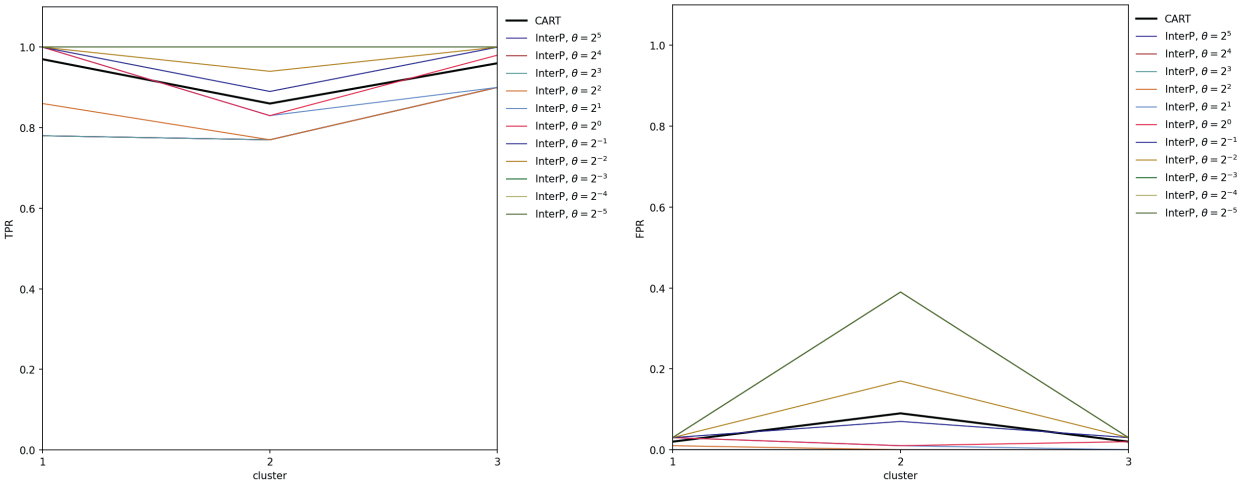
(b) False Positive Rate

Figure III.4: The `PIMA` data: the post-hoc interpretability results obtained with rule-based explanations given by (InterP) and CART.



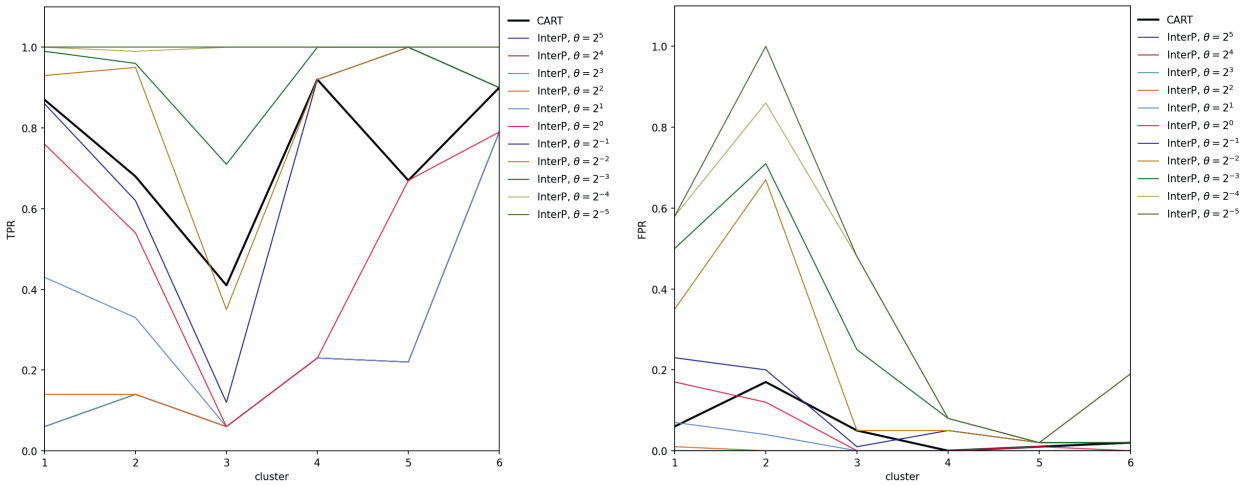
(a) True Positive Rate (b) False Positive Rate

Figure III.5: The `abalone` data: the post-hoc interpretability results obtained with rule-based explanations given by (InterP) and CART.



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

Figure III.6: The `wine` data: the post-hoc interpretability results obtained with rule-based explanations given by (InterP) and CART.



(a) True Positive Rate

(b) False Positive Rate

Figure III.7: The `glass` data: the post-hoc interpretability results obtained with rule-based explanations given by (InterP) and CART.

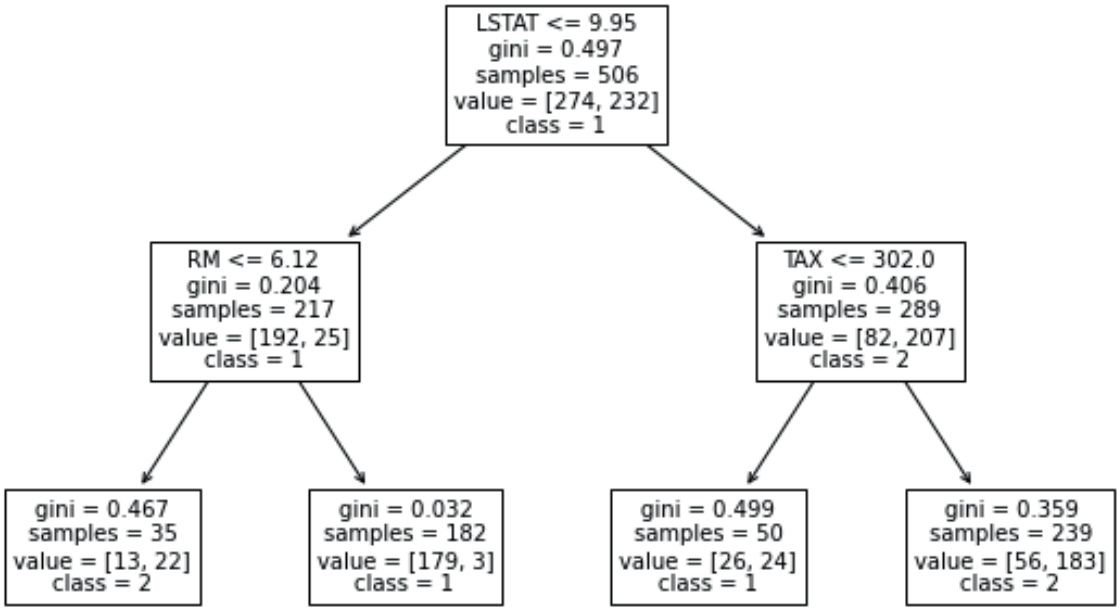


Figure III.8: The post-hoc rule-based explanations provided by a CART of depth 2 for the `housing` dataset for clusters (classes) 1 and 2.

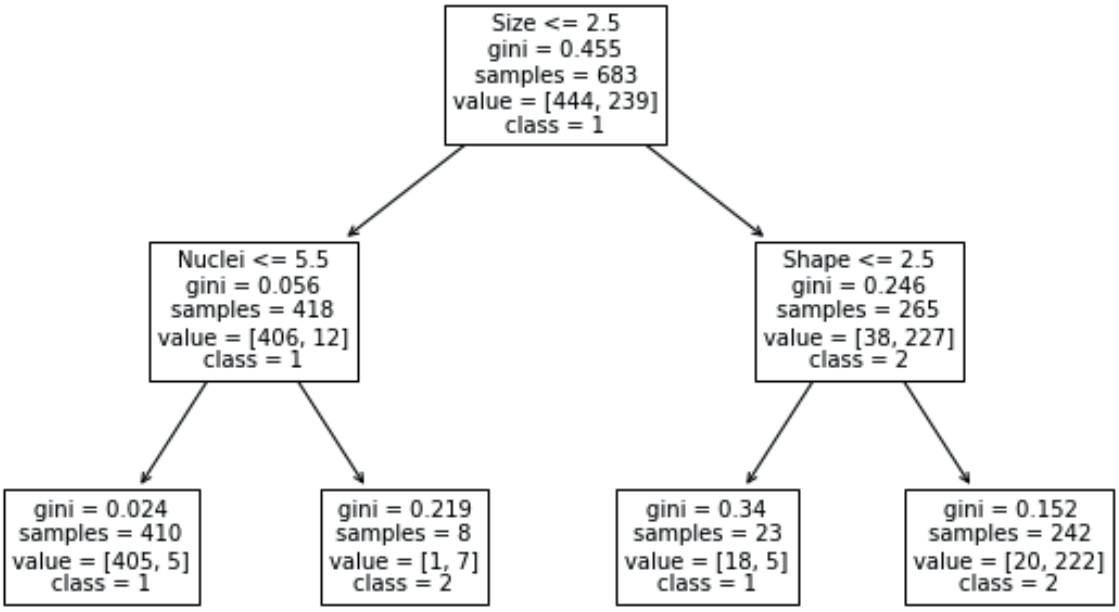


Figure III.9: The post-hoc rule-based explanations provided by a CART of depth 2 for the `breast cancer` dataset for clusters (classes) 1 and 2.

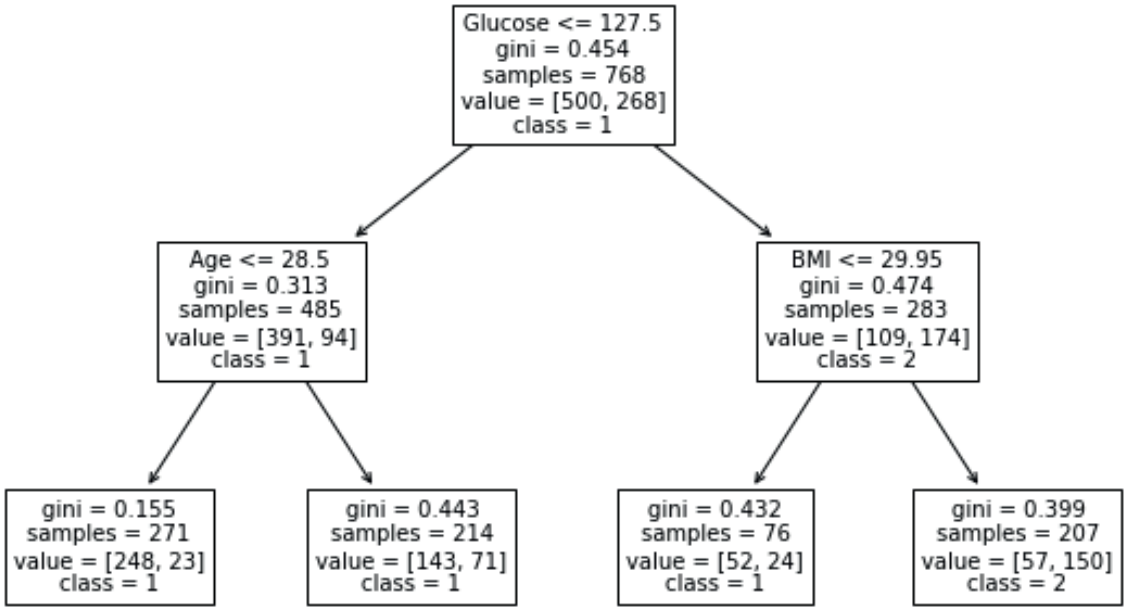


Figure III.10: The post-hoc rule-based explanations provided by a CART of depth 2 for the `PIMA` dataset for clusters (classes) 1 and 2.

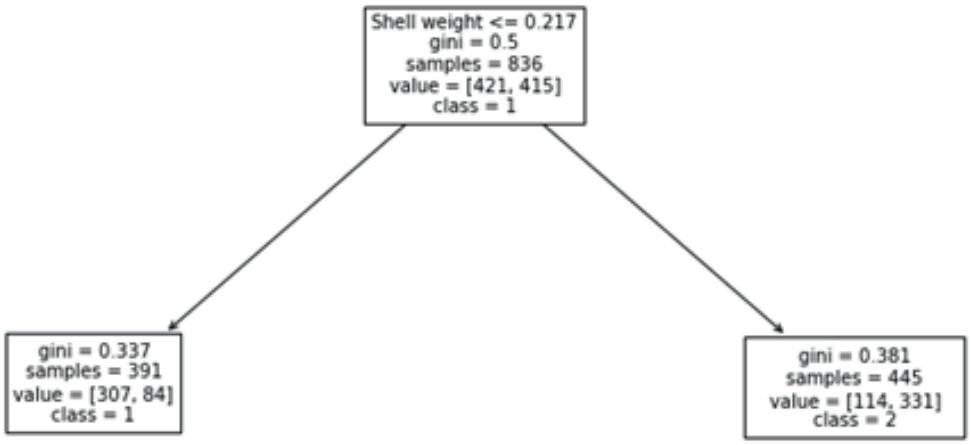


Figure III.11: The post-hoc rule-based explanations provided by a CART of depth 1 for the `abalone` dataset for clusters (classes) 1 and 2.

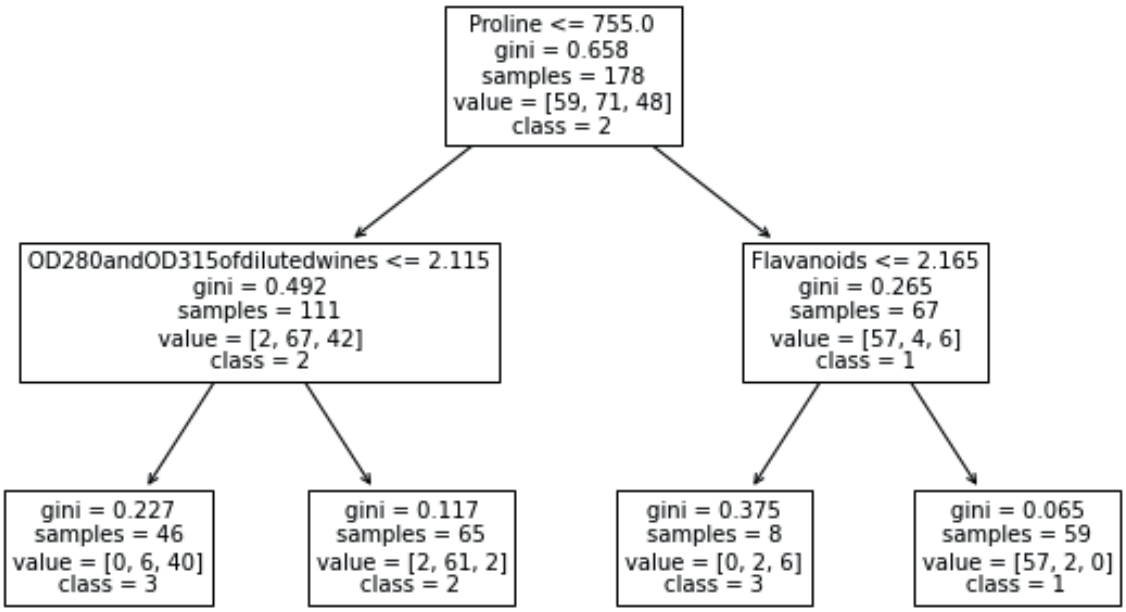


Figure III.12: The post-hoc rule-based explanations provided by a CART of depth 2 for the `wine` dataset for clusters (classes) 1, 2 and 3.

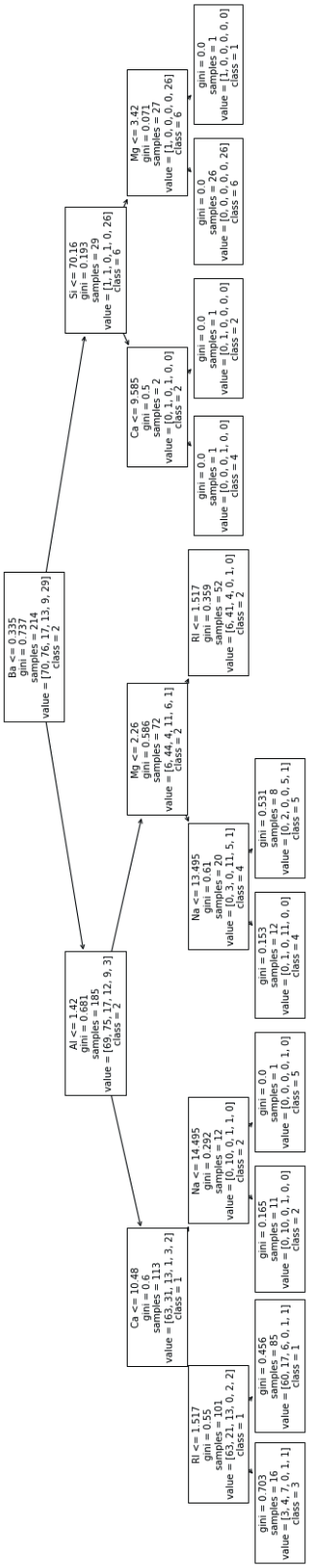


Figure III.13: The post-hoc rule-based explanations provided by a CART of depth 4 for the `glass` dataset for clusters (classes) 1, 2, 3, 4, 5 and 6.

Table III.22: The clusters and the rule-based 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 C = 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 $\leq$ 432 AND NOX $\leq$ 0.647
2^{-1}	2^{0}	$6.04 \cdot 10^4$	1	0.91	0.00	TAX $>$ 432
			2	1.00	0.00	TAX $\leq$ 432 AND NOX $\leq$ 0.647
2^{-1}	2^{1}	$6.04 \cdot 10^4$	1	0.91	0.00	TAX $>$ 432
			2	1.00	0.00	TAX $\leq$ 432 AND NOX $\leq$ 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 $\leq$ 432 AND NOX $\leq$ 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 $\leq$ 432 AND NOX $\leq$ 0.647
2^{0}	2^{1}	$6.04 \cdot 10^4$	1	0.91	0.00	TAX $>$ 432
			2	1.00	0.00	TAX $\leq$ 432 AND NOX $\leq$ 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 $\leq$ 432 AND NOX $\leq$ 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 $\leq$ 432 AND NOX $\leq$ 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 $\leq$ 432 AND NOX $\leq$ 0.647

Table III.23: The clusters and the rule-based explanations provided by (InterP), $\theta \in \{2^p\}_{p=-5,\ldots,5}$, for the `housing` dataset, with C = 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 $\leq$ 8.61
	2	0.14	0.00	LSTAT $>$ 11.25 AND PTRATIO $>$ 20.9
2^4	1	0.58	0.00	Al $\leq$ 1.146 AND Si $\leq$ 72.132
	2	0.14	0.00	Mg $\leq$ 2.805 AND Ca $>$ 10.443
2^3	1	0.64	0.01	RM $>$ 6.144 AND LSTAT $\leq$ 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 $\leq$ 9.93
	2	0.45	0.05	LSTAT $>$ 14.81 AND CRIM $\leq$ 10.6718
2^1	1	0.70	0.04	RM $>$ 6.12 AND LSTAT $\leq$ 11.66
	2	0.70	0.14	LSTAT $>$ 11.66 AND CRIM $\leq$ 11.1604
2^0	1	0.73	0.06	RM $>$ 6.059 AND LSTAT $\leq$ 11.66
	2	0.80	0.20	LSTAT $>$ 11.66 AND CRIM $\leq$ 37.6619
2^{-1}	1	0.78	0.19	LSTAT $\leq$ 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 $\leq$ 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 $\leq$ 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 $\leq$ 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 $\leq$ 21.2 AND B $>$ 6.68
	2	1.00	0.53	LSTAT $>$ 6.73 AND PTRATIO $>$ 14.4

III.5 Conclusions

In this chapter, 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. We have illustrated, in real-world datasets, the good performance of these explanations already for length $\ell = 2$, i.e., for very concise ones.

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].

Chapter IV

On enhancing the explainability and fairness of tree ensembles

IV.1 Introduction

The literature has reported some controversial/unfair decisions made with Artificial Intelligence / Machine Learning algorithms when, e.g., assessing the risk of potential recidivism or making social benefit allocations [Rudin, 2019]. This, together with the need of users (e.g., physicians, judges, civil servants, citizens) to understand why the model made a decision, calls for enhancing the transparency of Supervised Learning algorithms [Blanquero et al., 2020, European Commission, 2020, Goodman and Flaxman, 2017, Panigutti et al., 2023, Rudin et al., 2022]. In this chapter, we contribute to this stream of literature enhancing the explainability and fairness of tree ensembles for classification and regression tasks.

Decision trees are seen as the benchmark methodology in transparent classification [Carrizosa et al., 2021b]. A decision tree is defined by a series of if-then queries, which are easy to explain/interpret, in which features are compared against cutoff values. However, decision trees may not be that accurate and they may also suffer from instability, i.e., negligible changes in one feature may yield a rather different accuracy. To overcome these shortcomings, tree ensembles, in which a collection of decision trees are combined [Gambella et al., 2021, Mišić, 2020], have been proposed. The most common strategies to train tree ensembles are bagging or boosting [Friedman, 2001]. In the former one, bootstrapping defines the training sample for each tree, while random sampling on the set of features is used to reduce the number of if-then rules checked in each of the branch nodes. A classic example of this is the Random Forest [Biau and Scornet, 2016, Breiman, 2001]. In boosting, a sequential approach is used in which a new decision tree is added in each iteration with the aim to improve the error made by the tree ensemble at hand. A classic example of this is the XGBoost [Chen and Guestrin, 2016]. By construction, tree ensembles are far less explainable than decision trees, since, in general, almost all features are used in prediction with many cutoffs [Vidal and Schiffer, 2020]. Also by construction, tree ensembles do not allow a proper modeling of fairness.

In this chapter, we assume that we have a tree ensemble at hand and propose the Explainable and Fair Tree Ensemble (EFTE) methodology. This consists in modifying the original tree ensemble to, possibly at the expense of a decrease in accuracy, improve its explainability and fairness. As is customary in Supervised Classification, we have observations split into K classes and characterized by p features, either numerical or categorical. Some of these observations share a sensitive attribute, such as race or low income, and we need to ensure that the classifier does not discriminate against them and/or amplify the biases that may be present in the dataset [Romei and Ruggieri, 2014, Zafar et al., 2017].

The goal of the EFTE is to enhance the transparency of tree ensembles *by design*, i.e., by imposing it in the training process, while aiming for a competitive accuracy. The starting point of our approach is a collection of classification trees $\mathcal{T}$, where for each tree we have the if-then rule associated with each branch node and the class assigned to each leaf node. The source of these trees can vary. They may have been obtained by training a Random Forest or an XGBoost, or they can simply be a collection of weak learners each of them with a tree structure. The EFTE associates a weight to each tree in $\mathcal{T}$. The weights are optimized to ensure a good trade-off between classification accuracy, explainability, and fairness. Our measure of explainability is sparsity, the typical surrogate [Carrizosa et al., 2021b], and thus we impose an upper bound on the number of features used by EFTE. Our fairness measure is the classification accuracy for the sensitive observations, which we aim to have as high as possible. See Besse et al. [2022], Miron et al. [2020], Carrizosa et al. [2022a], Mehrabi et al. [2022] and references therein for other fairness metrics. Inspired by the models to train Support Vector Machines [Carrizosa and Romero Morales, 2013, Vapnik, 1995, 1998], we model the EFTE using a Mixed Integer Linear Programming (MILP) formulation, where there are only binary decision variables to model the sparsity. In our numerical results, we show that for standard datasets used in the fairness literature, we can dramatically enhance the fairness of the benchmark, namely the popular Random Forest, while using only a few features, all without damaging the misclassification error.

The chapter is organized as follows. In Section IV.2 we introduce the EFTE and the MILP formulation. In Section IV.3 we illustrate the performance of the EFTE in terms of misclassification error, fairness, and explainability in real-world datasets. In Section IV.4 we conclude the chapter and propose a number of lines of future research.

IV.2 The EFTE model

In this section, we introduce the Explainable and Fair Tree Ensemble (EFTE) classifier and an MILP formulation that is scalable in the number of observations. We start by presenting the information available from the collection of trees at hand that will be combined to yield an EFTE.

We have a classification problem with K classes indexed by the set $\mathcal{K} = \{1, \ldots, \mathrm{K}\}$, defined in a feature space $\mathcal{X} \subset \mathbb{R}^p$. Note that we can handle both numerical and categorical features, where for the latter ones we transform them into 0-1 features using the one-hot encoding.

Our methodology requires as a starting point a set of classification trees $\mathcal{T}$ of cardinality T. Since the trees are in place, we know what features are used in the branch nodes. We use for this the notation f_j^t, $j = 1, \ldots, p$ and $t = 1, \ldots, \mathrm{T}$, such that f_j^t is equal to 1 if feature j is used at least

once in tree t and 0 otherwise. We also know the class assignment rule used by each of the trees. Note that one of the most popular ways to make the class assignment is using the majority rule, where any individual in a given leaf node of the tree is associated with the most frequent class in such a leaf node. We use for this the notation $\psi_t : \mathcal{X} \to \mathcal{K}$, $t = 1, \ldots, \mathrm{T}$, where $\psi_t(\boldsymbol{x})$ denotes the class assigned by tree t to datapoint $\boldsymbol{x} \in \mathcal{X}$. We would like to stress that both types of data f_j^t and $\psi_t(\boldsymbol{x})$ are obtained prior to the training of the EFTE.

To build the EFTE we have a training sample $\mathcal{I}$ of $n = |\mathcal{I}|$ observations, namely, $\{(\boldsymbol{x}_i, k_i)\}_{i \in \mathcal{I}}$, where $\boldsymbol{x}_i \in \mathbb{R}^p$ is the feature vector characterizing observation i and $k_i \in \mathcal{K}$ is its class membership. Recall that we have the so-called sensitive individuals that we want to protect against an unfair treatment in terms of misclassification error in the training process. Therefore, we introduce notation $\mathcal{I}_1 \subset \mathcal{I}$ for the individuals in $\mathcal{I}$ that belong to the sensitive group, with $n_1 = |\mathcal{I}_1|$. As abovementioned, we know the class assigned by each tree to each observation in the training sample, namely, $\psi_t(\boldsymbol{x}_i)$ for $i = 1, \ldots, n$ and $t = 1, \ldots, \mathrm{T}$. For each tree t, we define the parameter $y_k^t(\boldsymbol{x}_i)$ that is equal to 1 if observation $i = 1, \ldots, n$ is predicted class k by tree t and 0 otherwise, $i = 1, \ldots, n$, $k = 1, \ldots, \mathrm{K}$ and $t = 1, \ldots, \mathrm{T}$. This notation will be convenient when defining the class score that EFTE uses to make predictions.

The EFTE has two goals, namely, to achieve a good and fair classification accuracy as well as a good sparsity, a surrogate of explainability. To this aim, we propose a mathematical optimization model to select only a few features in the classifier, to eliminate the trees using features outside the set of selected ones, and to weigh the remaining trees to achieve a good and fair classification accuracy. Therefore, we define the following decision variables and parameters. Let $\omega^t \in [0, 1]$ be a continuous decision variable that models the weight that the EFTE allocates to tree t, $t = 1, \ldots, \mathrm{T}$. Let ϕ_j be a binary decision variable equal to 1 if feature j is used in the model and 0 otherwise. Let $\eta \in (0, 1]$ be an upper bound on the maximum weight allocated to each of the trees and $\nu \in \{1, 2, \ldots, p\}$ an upper bound on the number of features used by the EFTE.

Usually, one requires a good classification accuracy by minimizing the misclassification error in the training sample, hereafter $\mathrm{misclas}(\boldsymbol{\omega}, \mathcal{T}, \mathcal{I})$. In this chapter, we also aim to have a fair misclassification error. Therefore, in addition, we propose to minimize the misclassification error in the subsample of sensitive individuals of the training sample, i.e., $\mathcal{I}_1$, hereafter $\mathrm{misclas}(\boldsymbol{\omega}, \mathcal{T}, \mathcal{I}_1)$. We follow a weighted approach and combine these two terms using the parameter $\alpha \geq 0$, defining the fair misclassification error as:

$$\mathrm{fairmisclas}(\boldsymbol{\omega}, \mathcal{T}, \mathcal{I}; \alpha) := \mathrm{misclas}(\boldsymbol{\omega}, \mathcal{T}, \mathcal{I}) + \alpha\, \mathrm{misclas}(\boldsymbol{\omega}, \mathcal{T}, \mathcal{I}_1). \tag{IV.2.1}$$

With this, our performance measure *fairmisclas* gives weight $1 + \alpha$ to the misclassification error

incurred on an individual of the sensitive group and 1 on an individual outside the protected group. The higher α the more stress we put in ensuring the correct classification in the individuals from the sensitive group.

Once we know the weights ω^t for each tree in the ensemble, the EFTE makes class assignment using $\sum_{t=1}^{\mathrm{T}} \omega^t y_k^t(\boldsymbol{x}_i)$, i.e., the score associated to class k for individual $i = 1, \ldots, n$. We assign class $\tilde{k}_i \in \{1, \ldots, \mathrm{K}\}$ to individual $i = 1, \ldots, n$ if

$$\tilde{k}_i \in \arg\max_k \sum_{t=1}^{\mathrm{T}} \omega^t y_k^t(\boldsymbol{x}_i).$$

With this, we consider the prediction is correct if

$$\sum_{t=1}^{\mathrm{T}} \omega^t y_{k_i}^t(\boldsymbol{x}_i) \geq \sum_{t=1}^{\mathrm{T}} \omega^t y_k^t(\boldsymbol{x}_i) + \varepsilon \quad \forall k \neq k_i, \tag{IV.2.2}$$

with $\varepsilon > 0$. Note that this parameter ε is used to model a conservative approach, such that for a record where we have a tie, and thus the difference between the best score and the second best score is below ε, we count this as a misclassification error.

The most straightforward way to count the number of individuals in which we incur a misclassification error would be to introduce binary decision variables for each individual and each class [Carrizosa et al., 2021b], to check whether the inequalities in (IV.2.2) are satisfied. However, this hard way of modeling errors is not scalable for large training samples since it includes as many binary decision variables as observations in the data set, and it can overfit the training data. Instead, we propose a soft approach using deviation decision variable which is a continuous decision variable $\xi_i \geq 0$, to measure the violation of the inequalities in (IV.2.2). This strategy is similar to the one used to train Support Vector Machines [Carrizosa and Romero Morales, 2013, Vapnik, 1995, 1998].

The second goal of the EFTE is to ensure that only a few features are used, i.e., the classifier is sparse. This is achieved by imposing an upper bound on the number of features used by the EFTE.

The MILP formulation of the EFTE that we will use in the numerical section reads as follows:

$$\min_{\boldsymbol{\omega}, \boldsymbol{\phi}, \boldsymbol{\xi}} \quad \frac{1}{n} \sum_{i=1}^{n} \xi_i + \alpha \frac{1}{n_1} \sum_{i=1}^{n_1} \xi_i \tag{IV.2.3}$$

$$\text{s.t.} \quad \sum_{t=1}^{\mathrm{T}} \omega^t y_{k_i}^t(\boldsymbol{x}_i) \geq \sum_{t=1}^{\mathrm{T}} \omega^t y_k^t(\boldsymbol{x}_i) - \xi_i + \varepsilon, \quad i = 1, \ldots, n,\ k = 1, \ldots, \mathrm{K} : k \neq k_i, \tag{IV.2.4}$$

$$\sum_{t=1}^{\mathrm{T}} \omega^t = 1, \tag{IV.2.5}$$

$$\sum_{j=1}^{p} \phi_j \leq \nu, \tag{IV.2.6}$$

$$\omega^t \leq \eta\, \phi_j, \quad j = 1, \ldots, p,\, t = 1, \ldots, \mathrm{T} : f_j^t = 1, \tag{IV.2.7}$$

$$\omega^t \geq 0, \quad t = 1, \ldots, \mathrm{T} \tag{IV.2.8}$$

$$\phi_j \in \{0, 1\}, \quad j = 1, \ldots, p, \tag{IV.2.9}$$

$$\xi_i \geq 0, \quad i = 1, \ldots, n. \tag{IV.2.10}$$

Let us discuss the objective function (IV.2.3) and constraints (IV.2.4) together. Constraints (IV.2.4) ensure that ξ_i is well-defined. If $\sum_{t=1}^{\mathrm{T}} \omega^t y_{k_i}^t(\boldsymbol{x}_i) \geq \sum_{t=1}^{\mathrm{T}} \omega^t y_k^t(\boldsymbol{x}_i) + \varepsilon$ for all $k \neq k_i$, then without loss of optimality we can choose $\xi_i = 0$, i.e., there is no misclassification error. However, if this is not the case, then $\xi_i > 0$. Now, it is clear that the objective function (IV.2.3) minimizes a proxy for the fair misclassification error with the help of deviation variables ξ_i. Note that the deviations of the protected observations in $\mathcal{I}_1$ are weighted with $1 + \alpha$ and the rest with 1. Constraint (IV.2.5) ensures that the weights ω^t sum up to 1 across the T trees. Therefore ω^t is the fraction of the total weight, and thus the importance, allocated to tree t. Constraint (IV.2.6) ensures that the EFTE can use at most ν features. Constraints (IV.2.7) are twofold. First, they ensure that ϕ_j is well-defined, i.e., if feature j cannot be used in the EFTE, namely $\phi_j = 0$, then $\omega^t = 0$ for each tree using that feature. Second, they impose the upper bound η on the weight ω^t, for each t. Constraints (IV.2.8)–(IV.2.10) specify the nature of the decision variables $\boldsymbol{\omega}, \boldsymbol{\phi}$ and $\boldsymbol{\xi}$. In sum, EFTE has been formulated as an MILP problem with at most $\mathrm{T}\, p + 2 + n\,(\mathrm{K} - 1)$ linear constraints, $\mathrm{T} + n$ non-negative decision variables, and p binary decision variables.

Once the EFTE has been trained, we make class predictions in new individuals in the following manner. For an individual with feature vector $\boldsymbol{x}$, recall that $y_k^t(\boldsymbol{x})$ is equal to 1 if tree t assigns class k to it, and otherwise 0, $t = 1, \ldots, \mathrm{T}$ and $k = 1, \ldots, \mathrm{K}$. Then, the EFTE predicts class

$$\tilde{k} \in \arg\max_k \sum_{t=1}^{\mathrm{T}} \omega^t y_k^t(\boldsymbol{x}),$$

where, in case of ties, we can break them randomly.

A few remarks can be made about the EFTE formulation (IV.2.3)–(IV.2.10). The first one is on the feasibility of the formulation. For small values ν, the problem may be infeasible. This is probably the case for trees coming from training a random forest in which no pruning has been applied, and therefore there may not be trees using only a few features. This is less of an issue in XGBoost where many trees of very small depth are combined. The same holds for η, namely, for small values of this parameter the problem is infeasible. This will certainly be the case if $\eta < \frac{1}{\mathrm{T}}$, as even by taking the maximum possible value of the weights would violate constraint (IV.2.5). The

second remark is on the nature of the parameter η. This can be seen as a regularization parameter on the tree ensemble [Hastie et al., 2009]. If $\eta = 1$ only a few trees may be chosen, with the risk of overfitting to the training sample. However, as we decrease the value of this parameter we force more trees to be part of EFTE.

To end this section, we briefly discuss two important extensions of the EFTE. First, our methodology can easily incorporate more sophisticated forms of sparsity. In the current formulation, we control the total number of features used in the EFTE. We can also control the number of features used from a given group. This is meaningful for categorical features, where for each j categorical we have a group of 0-1 features (one per category) associated with j coming from its one-hot encoding. We may want to impose sparsity within the group, and thus using as few categories as possible from j. The grouping of features may also appear when we have features of different natures such as socioeconomic ones or demographic ones. Again, one may want to impose group sparsity, i.e., controlling the sparsity within each of these groups [Benítez-Peña et al., 2021, Friedman et al., 2010]. Second, the EFTE can also deal with regression tasks, where the response variable is a continuous amount. In this case, for each tree in $\mathcal{T}$, we would know the predicted response for each individual, as well as the features used at least once. The EFTE would combine these predictions with the weights of the trees. The goal of the EFTE in regression would be to make these predictions as accurate as possible, as fair as possible, while using as few as possible features. To train this model we need to solve a Mixed Integer Convex Quadratic Problem with linear constraints, where again we only have binary decision variables associated with the selection of features. Indeed, the decision variables are still the weights associated to the trees ω^t and the 0-1 variables ϕ_j to decide which features can be used by the EFTE. The objective function minimizes the mean squared error in $\mathcal{I}$, as opposed to the misclassification error, and similarly for the sensitive individuals in $\mathcal{I}_1$. As for the feasible region, we only need constraints (IV.2.5)–(IV.2.9), as the remaining ones we saw above relate to the definition of the misclassification error and are not needed in the regression task.

IV.3 Numerical results

In this section we illustrate the performance of the EFTE on two publicly available datasets in terms of misclassification error, fairness, and explainability, benchmarking our approach against a very well-known class of tree ensembles, namely, Random Forests [Breiman, 2001].

We illustrate our methodology on two binary classification datasets, i.e., with K = 2 classes, often used in the fairness literature [Le Quy et al., 2022], namely the `PIMA` diabetes dataset [Dua

Table IV.1: Description of the features in the COMPAS dataset and the K = 2 classes.

Features	Description
Sex	Sex
Age	Age in years
Age_cat	Age category
Race	Race
Days_b_screening_arrest	The number of days between COMPAS screening and arrest
Decile_score	A continuous variable, the decile of the COMPAS score
Priors_count	The prior offenses count
C_charge_degree	Charge degree of original crime
Score_text	ProPublica-defined category of decile score
Class	Defendant is rearrested within 2 years (class 1) or not (class 2)

and Graff, 2017] and the COMPAS dataset [Angwin et al., 2016]. From Chapter III, we may recall that the PIMA dataset contains patients records and is used to predict whether a patient has diabetes. The COMPAS dataset contains records of crime defendants and is used to assess potential recidivism risk. The description of the features, the two classes and the sensitive group can be found in Table III.5 for the PIMA dataset and in Table IV.1 for the COMPAS dataset. For the PIMA dataset, the sensitive group is the set of individuals with diabetes, i.e., those in class $k = 2$. For the COMPAS dataset, the sensitive group is the set of African-Americans not being rearrested within 2 years, i.e., African-Americans in class $k = 2$. Note that in the PIMA dataset we are performing a classical cost-sensitivity analysis [Carrizosa et al., 2021b, Turney, 1995], where EFTE focuses on ensuring that the misclassification error is small for individuals at risk of diabetes, class $k = 1$, while ensuring that the overall misclassification error is also small. For the COMPAS dataset, EFTE focuses on some of the individuals of class $k = 2$ at risk of racial discrimination when predicting recidivism, namely, the ones showing the attribute African-American for the categorical feature Race in Table IV.1.

The dimension of the datasets is provided in Table IV.2, including the number of observations, the percentage of observations in the sensitive group, the number of features (p), the number of classes (K), and the class split. Note that we have both numerical and categorical features and therefore p refers to the number of features after the categorical ones have been coded through binary features, having one for each category of each categorical feature. The last column of Table IV.2 refers to the accuracy of a standard Random Forest (RF) with 500 trees of unlimited depth, trained using the *scikit-learn* library [Pedregosa et al., 2011].

Variable importance metrics have been developed to enhance the transparency of Random Forests and other tree ensemble models [Altmann et al., 2010]. To give a first impression on the importance of the features for the classification task, we report the variable importance metric that is given by the *scikit-learn* library when training the random forest in Table IV.2, namely,

Table IV.2: The dimension of the benchmark datasets to test EFTE and the out-of-sample misclassification error of the standard RF.

dataset	# observations	% sensitive observations	p	K	class split	RF error	RF error in sensitive group
PIMA	768	35%	8	2	35%/65%	23.78%	42.94%
COMPAS	6172	25%	20	2	46%/54%	36.27%	40.11%

the Mean Decrease in Impurity (MDI). This can be found in Figure IV.1 for the PIMA dataset and Figure IV.2 for the COMPAS dataset. For the PIMA dataset, Glucose is by far the feature with the highest importance, BMI, Age, DiabetesPedigreeFunction follow, while the remaining four features have a much lower value of the importance. For the COMPAS dataset, Age is the feature with the highest importance, closely followed by Priors_count, Decile_score and Days_b_screening_arrest. The remaining sixteen features, many of them associated with categorical features such as Race, have a much lower value of the importance metric.

The design of the experiments is as follows. We split the dataset into 3 samples: training (67%), validation (16.5%), and testing (16.5%). The training sample is used to build and reweight the initial trees. To build the EFTE, we consider stump trees. For each continuous feature, we construct (at most) 100 trees based on the percentiles of the corresponding feature where we split the observations below the percentile from the rest. Note that we may have fewer than 100 trees, as there may be repeated trees if the percentiles coincide. For each categorical feature, we build one tree per category where we split the observations showing that category from the rest. The validation sample is used to choose the best values of the EFTE parameters ε and η. We consider $\varepsilon \in \{2^{-3}, 2^{-2}, 2^{-1}\}$ and $\eta \in \{2^{-5}, 2^{-4}, 2^{-3}, 2^{-2}, 2^{-1}, 2^{0}\}$. We use the testing sample to report the performance of the EFTE in terms of misclassification error, fairness, and explainability. To end, we run five Monte Carlo simulations and report the average performance for each criterion across all runs. The number of stump trees in the EFTE, i.e., T, for the first run is 508 for the PIMA dataset and 199 for the COMPAS one, but other runs have similar values.

To illustrate the trade-off between the different criteria, we show results for a set of values of the parameters α and ν. Recall that $1+\alpha$ is the weight we give to the misclassification error incurred in individuals from the sensitive group while this weight is 1 for the remaining individuals, and ν is the maximum number of features that the EFTE can use. We consider $\alpha \in \{0\} \cup \{2^{-3}, 2^{-2}, 2^{-1}, 2^{0}\}$, where the higher the value of α the fairer we are towards individuals in the sensitive group. For ν, we use all possible values, namely $\nu \in \{1, 2, \ldots, p\}$.

To solve the MILP formulation (IV.2.3)–(IV.2.10), that builds the EFTE, we use *Gurobi* [Gurobi Optimization, 2020] with *Python* [Python Core Team, 2015] on a PC Intel®Core TM i7-8665U, 16GB of RAM. Each of the MILP instances, for different training samples and different values of the parameters, was solved to optimality in less than 1 second for the `PIMA` dataset and less than 6 seconds for the `COMPAS` one.

We benchmark our methodology against a Random Forest (RF) with 500 trees of the unlimited depth. As for the EFTE, we consider the RF with a limited number of features, namely the ν features with the highest value of the variable importance of the standard RF with all features, for $\nu \in \{1, 2, \ldots, p\}$. For instance, when $\nu = 1$, the RF is trained only with Glucose for the `PIMA` dataset and with Age for the `COMPAS` dataset, as seen in Figures IV.1 and IV.2 respectively. Note that when $\nu = p$ the RF is trained using all the features and thus it coincides with the standard RF, for which the out-of-sample accuracies were reported in Table IV.2.

The results on misclassification error, fairness and explainability for the `PIMA` dataset can be found in Figures IV.3, IV.5 and IV.7 (left panel), while for the `COMPAS` dataset can be found in Figures IV.4, IV.6 and IV.7 (right panel).

We start discussing the misclassification error and the fairness of the EFTE in the `PIMA` dataset. Figure IV.3a plots the average out-of-sample misclassification error, while Figure IV.3b refers to the average out-of-sample misclassification error for the sensitive observations in the different testing samples, our measure of fairness. Note that the misclassification error of the standard RF, i.e., when all features can be used and given in Table IV.2, corresponds to the point at the far right of the RF line. As one can see in Figure IV.3a, the EFTE gives similar results to the RF in terms of out-of-sample misclassification error, or even better, for small values of the parameter α tested, namely, $\alpha \in \{0\} \cup \{2^{-3}, 2^{-2}, 2^{-1}\}$. It is natural to see that the larger the value of α the higher the value of the misclassification error of EFTE. In Figure IV.3b we can see that the EFTE shows better results in terms of out-of-sample misclassification error in the sensitive observations compared to the RF for higher values of α, namely, $\alpha \in \{2^{-1}, 2^{0}\}$, for smaller values of α the EFTE and RF are comparable, and for $\alpha = 0$ EFTE is slightly worse. In sum, this means that there are values of α for which the EFTE gives similar overall misclassification error to RF and is much more fair towards the sensitive group than its benchmark.

We now discuss the explainability of the EFTE in the `PIMA` dataset, using two metrics, namely, the number of features and the number of stump trees used. Recall that in the MILP formulation (IV.2.3)–(IV.2.10), we have the parameter ν which is an upper bound on the number of features used in the EFTE. However, the fair and explainable tree ensemble may use even fewer features.

In Figure IV.5a, we count the actual number of features used by the EFTE and display the average number across the five folds, for each value of ν and α tested. Combining Figures IV.3 and IV.5a, we can see that the misclassification error and the fairness are very similar for all values of $\nu \geq 4$, meaning that with half of the features, we can find a good trade-off between these two criteria. For $\nu \geq 4$, the number of features used by the EFTE is between 3.6 and 5, and thus, our methodology is selective and not using the whole budget ν.

To complement Figure IV.5a, Figure IV.7 (left panel) displays the average number of folds in which a feature is used. We have a plot for each value of α, where the color coding in each plot is the same as in Figure IV.3. To ease the visualization of these averages, we use a heatmap with a cell for each value of ν (horizontal) and each feature (vertical), where the features are in decreasing order of the variable importance metric in Figure IV.1. The darker the cell the more folds are using that feature for the corresponding value of ν. In Figure IV.7 we can see that Glucose is always used by the EFTE, and most of the times BMI is, which we recall are the two features with the largest values of the variable importance metric in Figure IV.1. BloodPressure is never used, except for one value of α and ν, while Pregnancies, with a lower value of the variable importance is often used by the EFTE. Table IV.3 displays the EFTE with $\alpha = 0.5$, $\nu = 4$, $\epsilon = 0.125$ and $\eta = 0.5$, for one of the five Monte Carlo simulations. We can see that the variables used are Glucose, with the stump with the highest weight, and then Pregnancies, BMI and DiabetesPedigreeFunction.

Table IV.3: The EFTE (trees and weights) for the `PIMA` dataset, with $\alpha = 0.5$, $\nu = 4$, $\epsilon = 0.125$ and $\eta = 0.5$, for one of the five Monte Carlo simulations.

Left Node of Stump Tree t	ω^t
Pregnancies ≤ 0.352941	0.1250
Glucose ≤ 0.572864	0.4375
Glucose ≤ 0.723618	0.1250
BMI ≤ 0.582593	0.1250
DiabetesPedigreeFunction ≤ 0.242955	0.1250
DiabetesPedigreeFunction ≤ 0.248565	0.0625

We continue discussing another explainability metric of the EFTE, namely, the number of stump trees actually used in the tree ensemble, i.e., those for which the continuous decision variable $\omega^t \neq 0$. Figure IV.5b displays the results for the `PIMA` dataset, where we count the number of trees active in the ensemble and plot the average results across the five folds, for each value of α and ν tested. We can see that from the roughly 500 trees we start with, only a few of them are actively used by EFTE. We use no more than 60 trees, in general, while for $\nu \geq 4$ we use no more than 20 trees. Since we are using stumps this means that only a few thresholds of the (continuous) features play a role in the classification task.

To end with the PIMA dataset, we note that the EFTE with $\alpha = 0$ and $\nu = 1$ obtains a comparable misclassification error to the standard RF i.e. when all RF can make use of all the features, and better to the one obtained by RF when trained only on Glucose, the feature with the highest variable importance in Figure IV.1. As we can see in Figure IV.7, the top plot of the left panel, Glucose is also the feature selected by the EFTE, while from Figure IV.5b we can see that the EFTE uses close to 60 stump trees all defining thresholds of the feature Glucose. The outperformance of the EFTE for $\alpha = 0$ against RF may be explained by the fact that the weights assigned to the stump trees are optimized.

The conclusions that can be drawn for the COMPAS dataset are similar, as Figures IV.4, IV.6 and IV.7 (right panel) depict. Except for $\alpha = 1$, the EFTE gives similar results to the RF in terms of out-of-sample misclassification error or slightly better. For all the values of α, the EFTE shows better results in terms of out-of-sample misclassification error for the sensitive observations. In terms of explainability, we can see that the misclassification error and the fairness are very similar for all values of $\nu \geq 8$, while for $\nu \geq 8$ the number of features used by the EFTE is between 4 and 7. As before, our methodology selects only a few features even for larger values of the budget ν.

In Figure IV.7 we can see that Age, Priors_count, and Decile_score are always used by the EFTE, for $\nu \geq 4$, which we recall are the three features with the largest values of the variable importance metric in Figure IV.2. As for the remaining chosen features by the EFTE, the ranking of the variable importance is not necessarily followed for larger values of ν. Table IV.4 displays the EFTE with $\alpha = 0.5$, $\nu = 4$, $\epsilon = 0.125$ and $\eta = 0.5$, for one of the five Monte Carlo simulations. We can see that in addition to Age, Priors_count, and Decile_score, we also use Days_b_screening_arrest.

Table IV.4: The EFTE (trees and weights) for the COMPAS dataset, with $\alpha = 0.5$, $\nu = 4$, $\epsilon = 0.125$ and $\eta = 0.5$, for one of the five Monte Carlo simulations.

Left Node of Stump Tree t	ω^t
Age $\leq$ 0.012821	0.041667
Age $\leq$ 0.025641	0.083333
Age $\leq$ 0.192308	0.041667
Age $\leq$ 0.269231	0.041667
Age $\leq$ 0.833333	0.104167
Priors_count $\leq$ 0.052632	0.041667
Priors_count $\leq$ 0.131579	0.041667
Priors_count $\leq$ 0.552632	0.041667
Priors_count $\leq$ 0.578947	0.041667
Priors_count $\leq$ 0.657895	0.041667
Days_b_screening_arrest $\leq$ 0.5	0.020833
Days_b_screening_arrest $\leq$ 0.516667	0.020833
Days_b_screening_arrest $\leq$ 0.716667	0.020833
Days_b_screening_arrest $\leq$ 0.733333	0.020833
Decile_score $\leq$ 0	0.354167
Decile_score $\leq$ 0.666667	0.041667

From Figure IV.6b, we can see that we use less than 35 trees, in general, although we have around 200 to start with. Some of those correspond to thresholds on the continuous variables, while others to critical categories of the categorical features, such as Race.

We note that for the `COMPAS` dataset, our methodology outperforms RF in terms of fairness for $\alpha = 0$, while is slightly better in terms of misclassification error. This outperformance may be explained by the fact that we only use a few thresholds of the continuous features as well as a few attributes of the categorical features, and thus we are less prone to overfit [Carrizosa et al., 2021a, 2022c].

To end this section, we show similar plots to those in Figure IV.3 for the `PIMA` dataset, when the stump trees used as starting point for the EFTE methodology come from building an XGBoost in the training sample, while the remaining of the design of experiments stays the same. The results can be found in Figure IV.8 for XGBoosts with 1000 stump trees. For this dataset, both figures show a similar tradeoff between the misclassification error and the fairness of the EFTE, with slightly better fairness in Figure IV.8 at the cost of higher misclassification error.

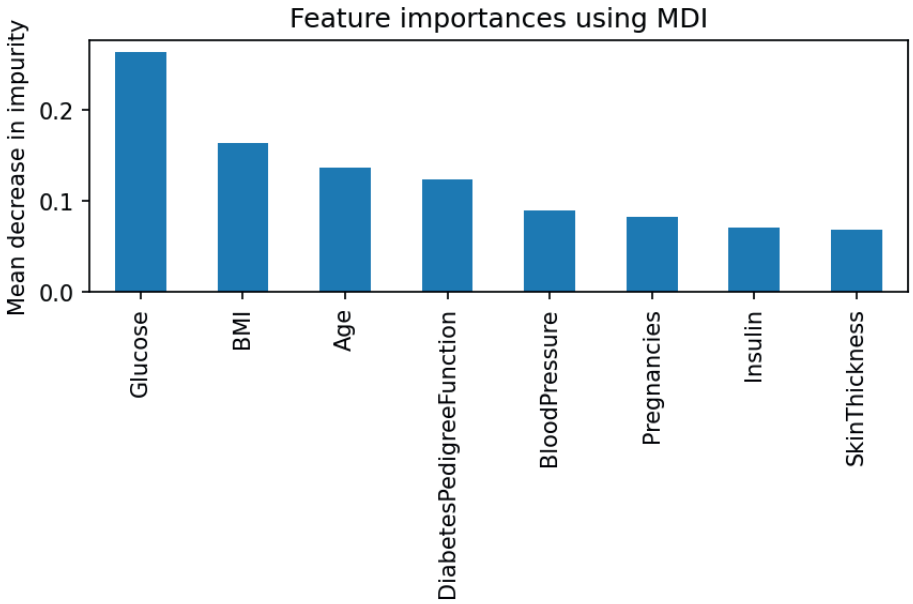


Figure IV.1: Variable importance plot for a standard RF trained on the `PIMA` dataset.

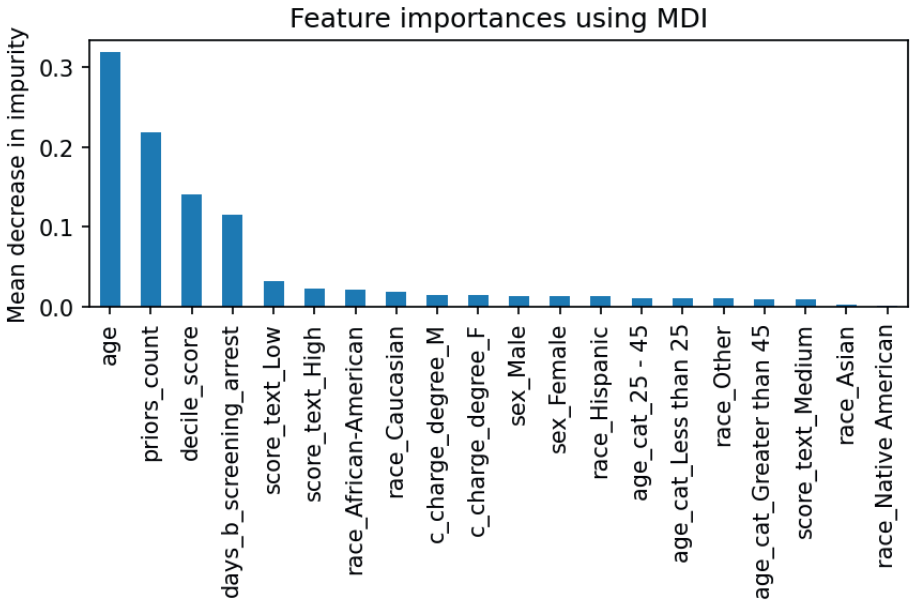


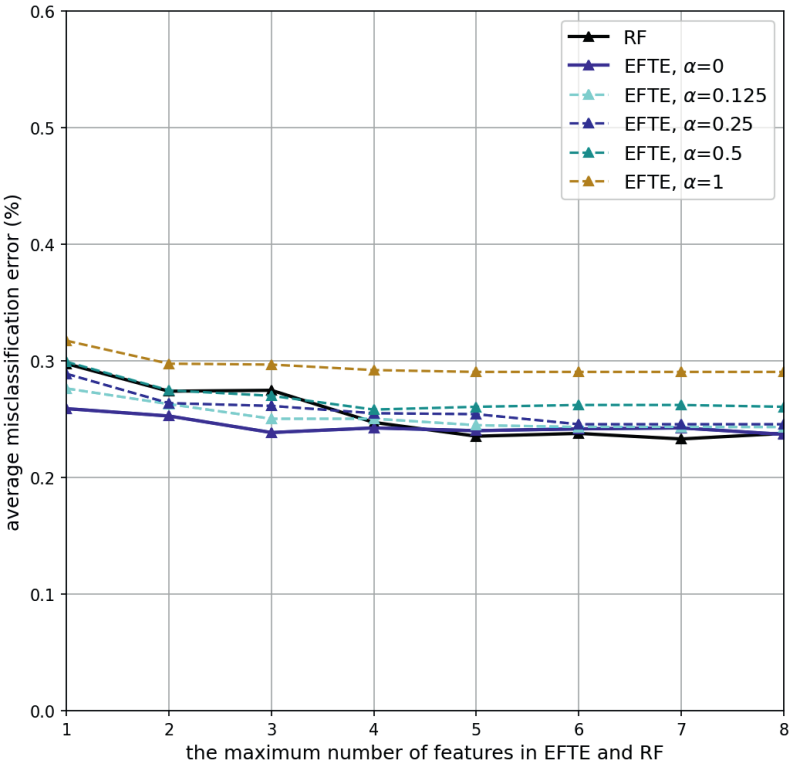
Figure IV.2: Variable importance plot for a standard RF trained on the `COMPAS` dataset.

IV.4 Conclusions

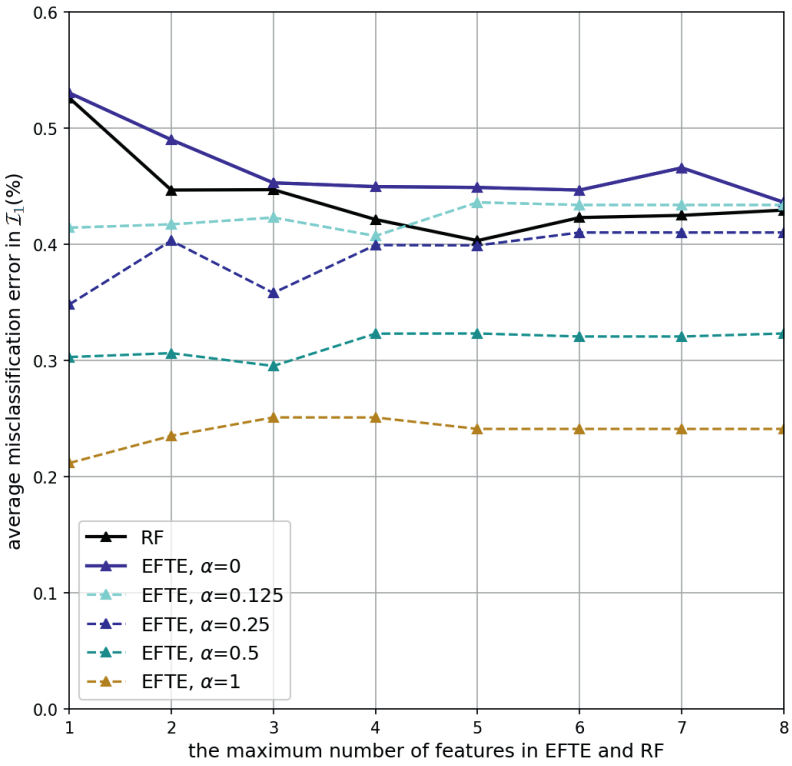
In this chapter, we trade off some of the accuracy of the tree ensemble to enhance its sparsity, ensuring that we use fewer features, and its fairness towards a group sharing a sensitive attribute, ensuring that the accuracy in this group is as high as possible. This means that the feature selection

is not only guided by the overall misclassification error but also the misclassification error in the sensitive group. We propose an MILP formulation to train the Explainable and Fair Tree Ensemble (EFTE), where the classification error is modeled through continuous decision variables as opposed to binary ones. Therefore, our formulation has the advantage of being scalable in the number of observations. Our numerical results illustrate the EFTE built from a pool of stump trees. For two datasets often used in the fairness literature, we can show that the EFTE dramatically improves the fairness of the ensemble without harming the overall misclassification error, and that this is true even if we use less than half of the features.

As for future research, there are two interesting directions. The first one is about the collection of classification trees $\mathcal{T}$ at hand to train the EFTE. To warrant a good sparsity of the EFTE, it is important that each individual tree in $\mathcal{T}$ is using only a few features. This is the case if the trees are shallow, such as those coming from an XGBoost. If the trees are deep, as is normally the case for trees coming from a Random Forest, we can prune them in a preprocessing step [Liu and Mazumder, 2023], before adding them to $\mathcal{T}$. Instead, one can simultaneously prune the trees and train the EFTE. Hence, in addition to the decision variables associated with the weights of the trees and the selection of the features, we need new ones to model the pruning of the trees. The second one is about the fairness measure optimized by the EFTE. We have modeled the misclassification error in the sensitive group, but there are other criteria we could have considered such as the disparate mistreatment [Miron et al., 2020]. The study of efficient mathematical optimization formulations for these two problems is left as an open question.

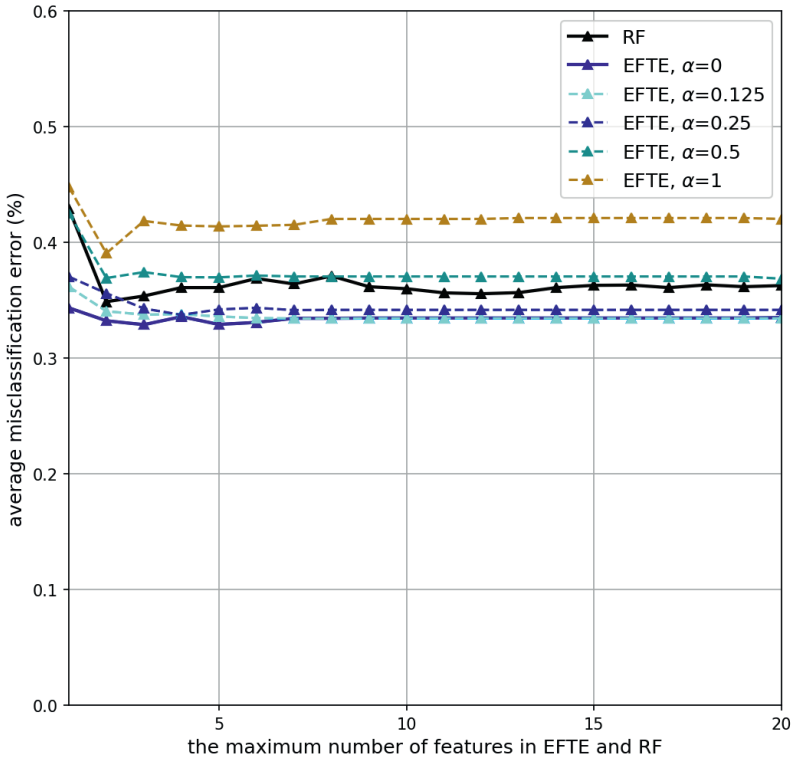


(a) Out-of-sample misclassification error

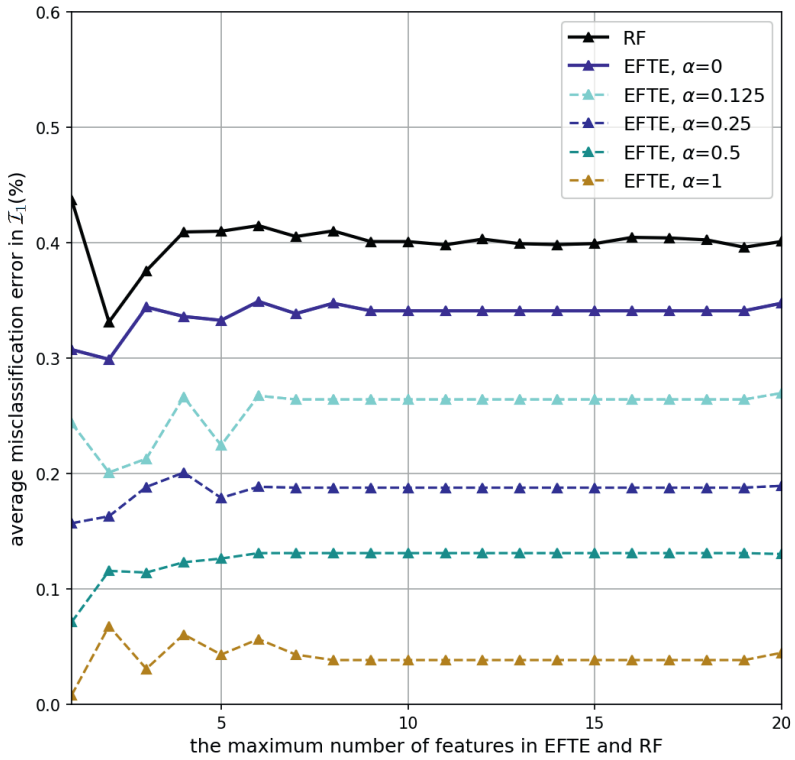


(b) Out-of-sample misclassification error in sensitive observations

Figure IV.3: Out-of-sample misclassification error and fairness in the `PIMA` dataset of the EFTE and RF.

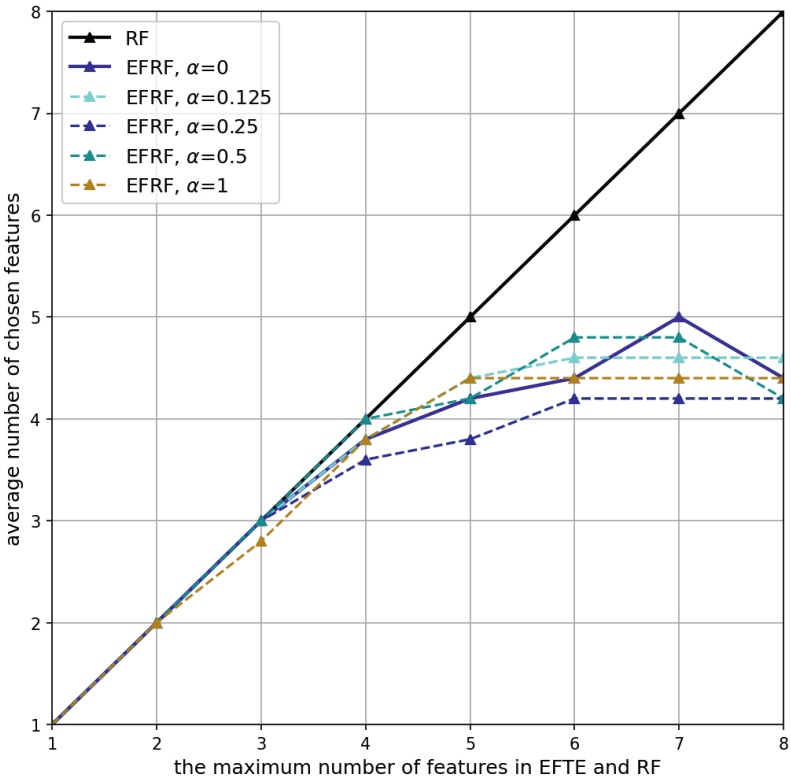


(a) Out-of-sample misclassification error

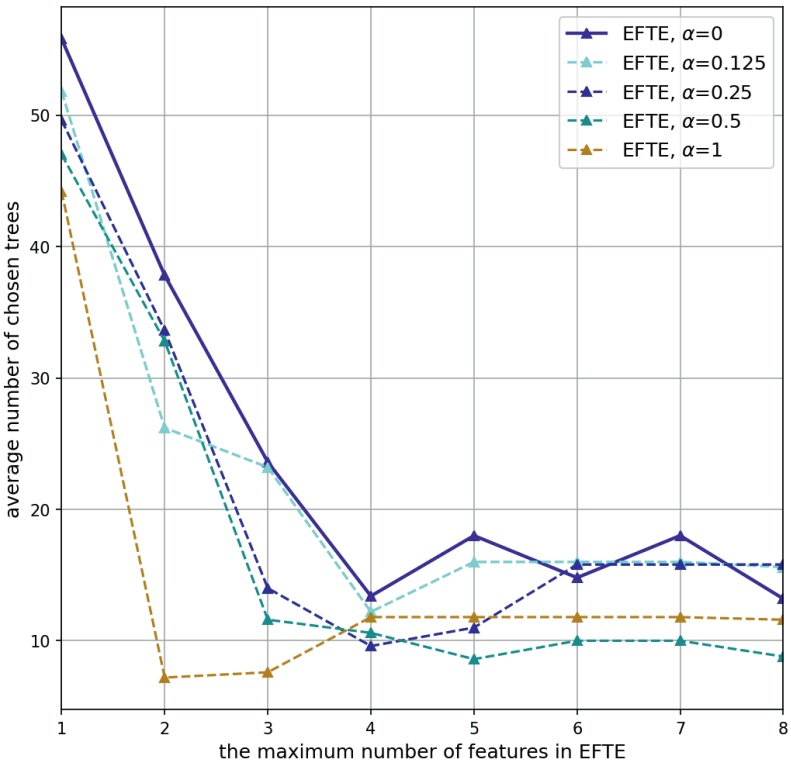


(b) Out-of-sample misclassification error in sensitive observations

Figure IV.4: Out-of-sample misclassification error and fairness in the `COMPAS` dataset of the EFTE and RF.

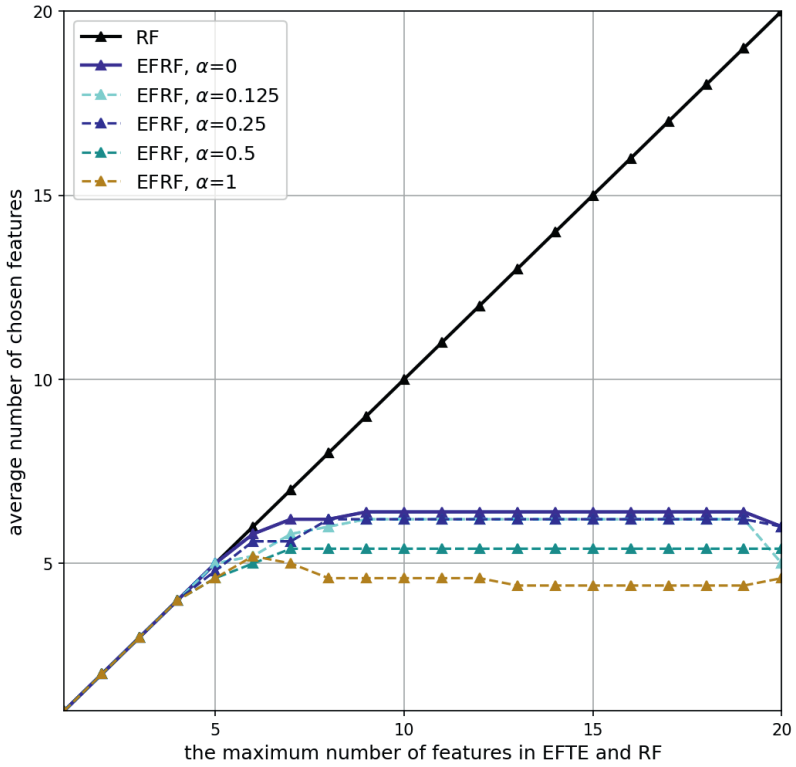


(a) Number of used features

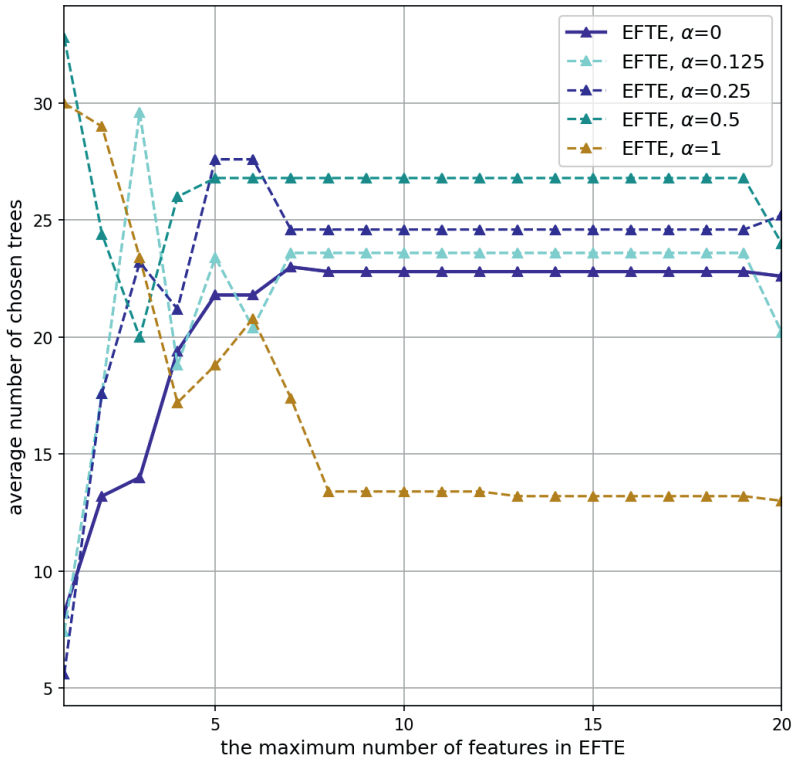


(b) Number of used trees

Figure IV.5: Average number of features (above) and average number of trees (below) used by the EFTE in the `PIMA` dataset.



(a) Number of used features



(b) Number of used trees

Figure IV.6: Average number of features (above) and average number of trees (below) used by EFTE in the `COMPAS` dataset.

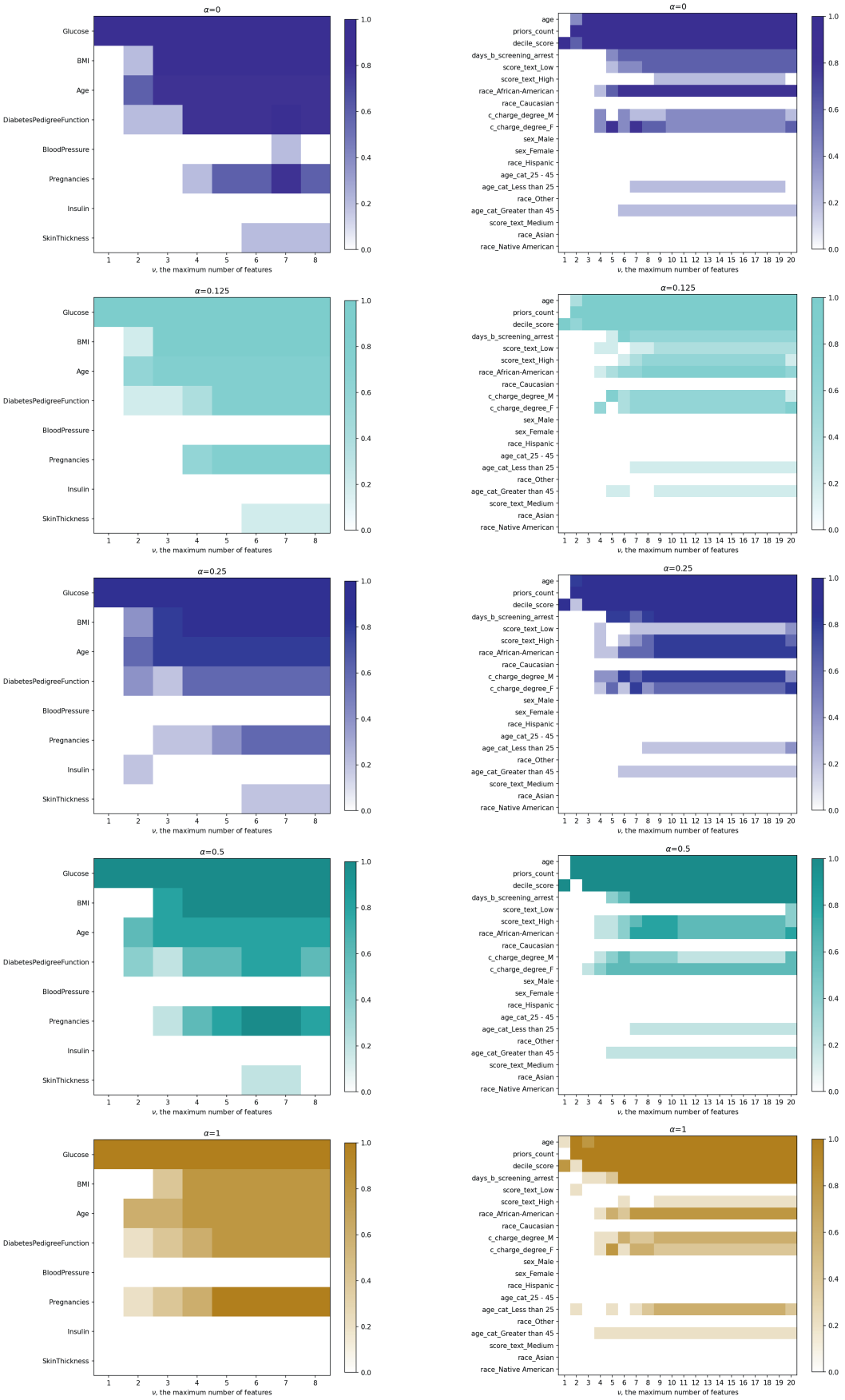
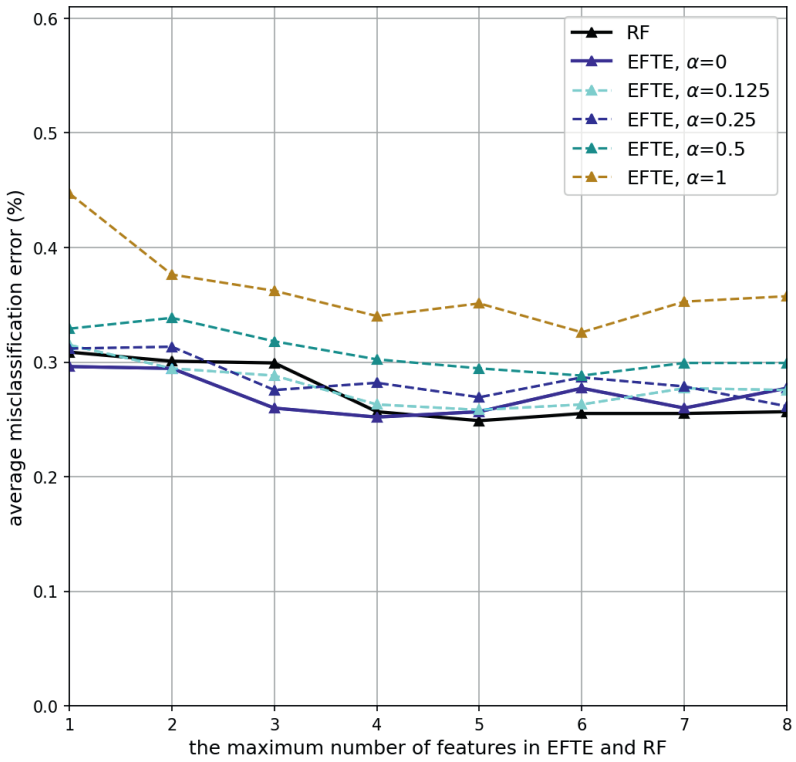
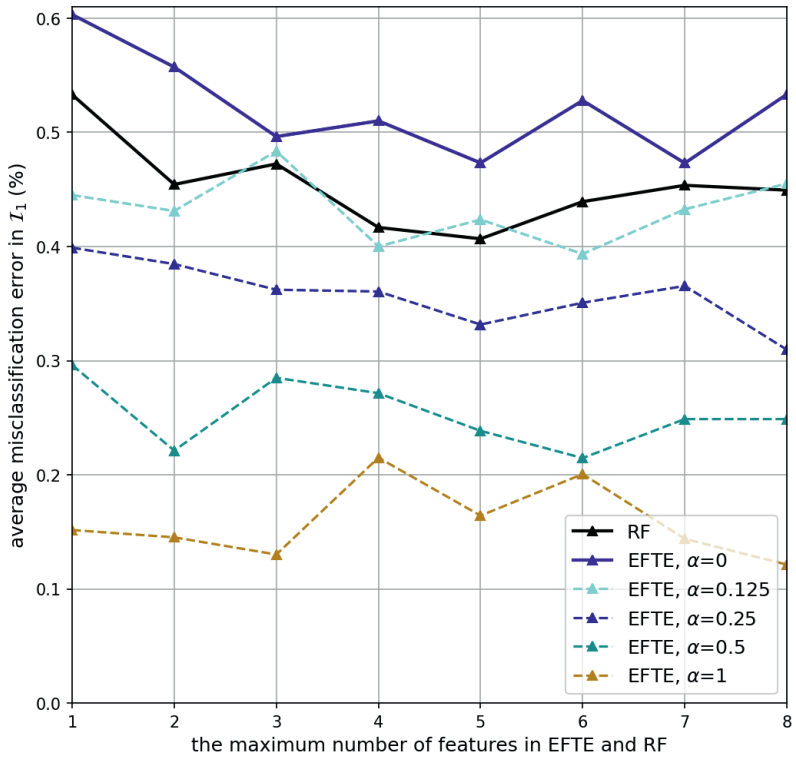


Figure IV.7: Heatmap of the average number of folds in which a feature is used by the EFTE in the `PIMA` dataset (left) and the `COMPAS` dataset (right).



(a) Out-of-sample misclassification error



(b) Out-of-sample misclassification error in sensitive observations

Figure IV.8: Out-of-sample misclassification error and fairness in the `PIMA` dataset of the EFTE and RF, when XGBoost is used to generate the stump trees.

Chapter V

Fair treatment allocation via tree ensembles

V.1 Introduction

Nowadays, it is common to use data from observational studies in treatment allocation problems [Athey et al., 2019, Bertsimas et al., 2019, Jo et al., 2021], where one has to decide which individuals will receive treatment and which not. Examples abound in, e.g., personalized medicine, lending services, and online advertising. It has been shown that individuals may react differently to the treatment of interest [Xie et al., 2012]. With the ability to collect large datasets, researchers can now design personalized treatment allocation policies [Fernández-Loría and Provost, 2022, Kleinberg et al., 2015] through newly developed non-parametric methods predicting heterogeneous treatment effects (HTE), i.e., the expected value of the difference of outcomes between being treated or not, conditional to the values of the covariates linked with each individual, see Tran et al. [2023] for a review.

Once the HTE has been predicted for the individuals, the treatment is allocated to those with the highest values of HTE using a threshold of interest. However, this plausible allocation strategy may lead to unfair decisions [Athey, 2017, Baumgaertner, 2022, Bénard et al., 2021, Constantaras et al., 2023, Kayser-Bril, 2020, Kim and Zubizarreta, 2023, Nabi et al., 2019]. Indeed, historical data may suffer from biases linked to sensitive attributes, such as gender or age. If the prediction algorithm is not carefully designed, the predictions may inherit these biases, yielding unequal treatment allocations to individuals in the sensitive (e.g., women or elderly patients) and the non-sensitive (e.g., men or non-elderly patients) groups. This may cause discrimination in socially impactful settings such as in criminal justice [Gelman et al., 2007], healthcare [Obermeyer et al., 2019], or credit scoring [Das et al., 2023].

Let me illustrate the need for fairness considerations with an example of access to healthcare resources. For patients with advanced osteoarthritis, surgical intervention is one of the effective treatments. The replacement of the affected joints can improve quality of life. It has been found that some racial minorities show worse outcomes [Usiskin and Misra, 2022]. If a decision maker needs to target surgery to the most cost-effective patients and uses for it an algorithm without any fairness awareness, then, due to the existing biases in the data, the prediction of the treatment effect might be lower for individuals from these racial minorities (sensitive group). This leads to a lower chance of being selected for surgery in the sensitive group, which may be socially undesirable and needed to be mitigated. In Kim and Zubizarreta [2023] and references therein, it is advocated that the allocation of healthcare resources should be based not only on cost-effective but also on ethical values.

To mitigate algorithmic bias many fairness metrics have been proposed in the literature [Baro-

cas et al., 2023, Khademi et al., 2019, Jo et al., 2021, Zafar et al., 2017], mainly for Supervised Learning, as in Chapter IV. In this chapter, I introduce a fairness measure to ensure that the predicted treatment effects in the two groups (sensitive and non-sensitive) on average do not differ much. Designing a prediction model for the HTE that takes into account, on top of accuracy, my fairness measure, yields corrected HTE predictions both in the sensitive and the non-sensitive groups. These corrected predictions can enhance the availability of the treatment to the sensitive group, hopefully with a small impact on the population.

In this chapter, I propose the Fair Heterogeneous Treatment Effect Forest (FhteF) methodology, which aims to predict the HTE for treatment allocation while ensuring that the predictions in the sensitive group do not differ significantly from those in the non-sensitive group. The main idea of the FhteF is to use a given ensemble of treatment effect predictors and assign weights to each of them such that good predictors in terms of accuracy and fairness contribute more. As the predictors of the treatment effect I leverage the Tree Ensemble with linear models in the leaves.

Before describing the FhteF, I introduce some notation. Individuals have associated the following random variables:

- $\boldsymbol{X}$ is a vector of p explanatory variables.
- W is a binary variable indicating whether the individual has been treated ($W = 1$) or not ($W = 0$).
- $Y(W)$ is the outcome for W given, i.e., $Y(W) = W\,Y(1) + (1 - W)\,Y(0)$ [Imbens and Rubin, 2015]. With this, the treatment effect is expressed as $Y(1) - Y(0)$. Note that for a given individual, $Y(W)$ is the observed outcome, given as either $Y(1)$ or $Y(0)$, while $Y(1 - W)$, referred in the literature as the potential outcome, is not observed, and therefore the treatment effect $Y(1) - Y(0)$ cannot be observed either.
- A novel aspect of this chapter is considering in addition the binary variable Z indicating whether the individual belongs to the sensitive group ($Z = 1$) or not ($Z = 0$).

I adapt the definition of heterogeneous treatment effect [Künzel et al., 2019] to this fairness setting. The HTE for an individual with $\boldsymbol{X} = \boldsymbol{x}$ as the vector of explanatory variables and $Z = z$ as the sensitive group membership value is thus defined as $\tau(\boldsymbol{X}, z) = \mathbb{E}\left[Y(1) - Y(0) | (\boldsymbol{X}, Z) = (\boldsymbol{x}, z)\right]$. The treatment allocation then can be done via a policy function $g(\boldsymbol{X}, z) = 1\{\tau(\boldsymbol{X}, z) > \bar{\tau}\}$, where $\bar{\tau}$ a threshold of interest.

To train the prediction model for the HTE, the sample $\{(\boldsymbol{X}_i, Z_i, W_i, Y_i(W_i))\}_{i \in \mathcal{I}}$ of size $|\mathcal{I}| = n$ is at hand. Hereafter, I denote by $\mathcal{I}_1$ the set of training sample observations in the sensitive group,

i.e., $\mathcal{I}_1 = \{i \in \mathcal{I} : Z_i = 1\}$, of size $|\mathcal{I}_1| = n_1$, and by $\mathcal{I}_0$ those in the non-sensitive group, namely, $\mathcal{I}_0 = \{i \in \mathcal{I} : Z_i = 0\}$, of size $|\mathcal{I}_0| = n_0$. With this, $\mathcal{I} = \mathcal{I}_0 \cup \mathcal{I}_1$ and $n = n_0 + n_1$.

The HTE is the expected value of the treatment effect conditional to given values of $(\boldsymbol{X}, Z)$. Therefore, I need to predict the potential outcome for each of the individuals in the training sample, which automatically gives their corresponding treatment effect. For a given individual, with characteristics $(\boldsymbol{X}_i, Z_i, W_i, Y_i(W_i))$, there are mainly two plausible approaches to predicting the potential outcome $Y_i(1 - W_i)$. The first one considers similar individuals in the covariates, in my case $(\boldsymbol{X}, Z)$, but with treatment value $1 - W_i$ [Frölich, 2004] to make the predictions. The second one adjusts the well-known machine learning methodology Random Forests (RFs) [Breiman, 2001]. In this case, obviously, the treatment variable W cannot be used in the splitting rules, while the goal of the splitting rule is to maximize heterogeneity in the covariates while balancing the individuals from the treated and not treated groups. The Generalized Random Forest [Athey et al., 2019, Wager and Athey, 2018] is one of the most well-known of these methodologies. None of these approaches allows direct control on the differences in the predictions made for the sensitive and the non-sensitive groups.

In contrast with the aforementioned methods, the FhteF predicts HTE with fairness considerations. The procedure has two steps. In the first step, a random forest consisting of T decision trees is built. In each tree, branching is performed taking into account only the features in $(\boldsymbol{X}, Z)$. At each leaf node, a linear regression model is built to predict both $Y(1)$ and $Y(0)$ from the observed values of $(\boldsymbol{X}, Z, W, Y(W))$. In the second step, predictions for $Y(1)$ and $Y(0)$ are obtained for any individual. This is done by defining weights for the different trees, and predicting $Y(1)$ and $Y(0)$ as the weighted averages of the predictions obtained in the first step at the different leaf nodes of the forest. Similarly to Chapter IV, these weights are obtained by optimizing a convex combination of the mean squared error of the predictions and my fairness measure.

By design, all trees in the RF yield accurate predictions (I optimize mean squared errors), and, thanks to the reweighting, FhteF gives more importance to fairier trees. I model the FhteF using a Convex Quadratic Programming formulation with a linear constraint, which can be efficiently solved with existing commercial solvers for small and medium sizes of the problem.

The remainder of the chapter is organized as follows. In Section V.2, I introduce a mathematical optimization formulation for the FhteF, that reweights the trees in the RF to predict fair HTEs. In Section V.3, I illustrate the performance of the FhteF on simulated datasets. The results of the study show that the FhteF significantly improves fairness compared to the benchmark, namely the Generalized Random Forest. In Section V.4, I provide concluding remarks and lines of future

research.

V.2 The FhteF model

In this section, I provide a Mathematical Optimization formulation for the Fair Heterogeneous Treatment Effect Forest (FhteF).

Recall that in the first step of the FhteF I build an RF, with a linear model at each leaf node. The model for leaf node l in tree t has the form

$$Y(W) = \beta^0_{tl} + W\beta^w_{tl} + \boldsymbol{X}^\top \boldsymbol{\beta}^x_{tl} + Z\beta^z_{tl} + \varepsilon_{tt}, \tag{V.2.1}$$

where β^0_{tl} is the independent term, β^w_{tl} the coefficient of the treatment variable, β^z_{tl} the one of the sensitive attribute, $\boldsymbol{\beta}^x_{tl}$ the vector of coefficients of the p explanatory variables, and ε the error term. Clearly, this linear model gives me the prediction for the two possible outcomes, for all observations i falling in this leaf node, namely, $\hat{Y}_{it}(W_i)$ and $\hat{Y}_{it}(1-W_i)$, while the HTE prediction is equal to the estimated coefficient of the treatment variable, i.e., $\hat{\tau}_{tl} = \hat{\beta}^w_{tl}$.

In the second step of the FhteF, the trees in the RF are reweighted to enhance the fairness of the HTE predictions. Let $\omega_t \geq 0$ be the continuous decision variable representing the weight of tree t. The predictions returned by the FhteF are a combination of the leaf nodes' predictions with the tree weights ω_t. Let me illustrate it for observation i. For tree t, i falls in a single leaf node denoted as $l(i)$. Thus, the HTE prediction for i is equal to $\hat{\tau}_i = \sum_{t=1}^{\mathrm{T}} \omega_t \hat{\beta}^w_{t\,l(i)}$, while the outcome predictions are $\hat{Y}_i(W_i) := \sum_{t=1}^{\mathrm{T}} \omega_t \hat{Y}_{it}(W_i)$ and $\hat{Y}_i(1-W_i) := \sum_{t=1}^{\mathrm{T}} \omega_t \hat{Y}_{it}(1-W_i)$. I define my measure of fairness as the absolute value of the difference between the average predicted HTE in $\mathcal{I}_1$ and that in $\mathcal{I}_0$, i.e.,

$$\left| \frac{1}{n_1} \sum_{i \in \mathcal{I}_1} \sum_{t=1}^{\mathrm{T}} \omega_t \hat{\beta}^w_{t\,l(i)} - \frac{1}{n_0} \sum_{i \in \mathcal{I}_0} \sum_{t=1}^{\mathrm{T}} \omega_t \hat{\beta}^w_{t\,l(i)} \right|. \tag{V.2.2}$$

With this, the mathematical formulation of the FhteF reads as follows:

$$\min_{\boldsymbol{\omega}} \quad \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \sum_{t=1}^{\mathrm{T}} \omega_t \hat{Y}_{it}(W_i) \right)^2 + \alpha \left| \frac{1}{n_1} \sum_{i \in \mathcal{I}_1} \sum_{t=1}^{\mathrm{T}} \omega_t \hat{\beta}^w_{t\,l(i)} - \frac{1}{n_0} \sum_{i \in \mathcal{I}_0} \sum_{t=1}^{\mathrm{T}} \omega_t \hat{\beta}^w_{t\,l(i)} \right| \tag{V.2.3}$$

$$\text{s.t.} \quad \sum_{t=1}^{\mathrm{T}} \omega_t = 1, \tag{V.2.4}$$

$$\omega_t \geq 0, \quad \forall t. \tag{V.2.5}$$

The objective function (V.2.3) is the weighted sum of two terms. The first term minimizes the mean squared error (MSE) of the predicted observed outcome. The second term, with a weight

of $\alpha \geq 0$, measures the fairness of the HTE predictions. The higher the weight α the fairer the model, while when $\alpha = 0$, the model ignores fairness and the only goal is accuracy. Constraint (V.2.4) ensures that the weights $\boldsymbol{\omega}$ are normalized, which is a form of regularization [Hastie et al., 2009]. Constraint (V.2.5) specifies the nature of the decision variable $\boldsymbol{\omega}$.

The objective function (V.2.3) is nonsmooth as it includes an absolute value. With standard techniques, an equivalent smooth formulation is obtained. With this, model (V.2.3)–(V.2.5) can be reformulated as a Convex Quadratic Programming problem with a linear constraint, with $\mathrm{T}+1$ continuous variables, where T variables relate to the weights and the last variable relates to the linearization of the absolute value.

The FhteF predicts the HTE for the observations in the training sample. In addition, the FhteF can also be used to make predictions in new observations. Once the FhteF has been built, the predicted treatment effect of a new individual, say s, is equal to $\sum_{t=1}^{\mathrm{T}} \hat{\omega}_t \hat{\beta}^w_{t\,l(s)}$.

It is worth mentioning that the FhteF can handle other methods instead of the RF, such as an XGBoost [Chen and Guestrin, 2016]. We would expect in this case more trees but shallower, and therefore a larger size of the FhteF. With respect to the linear model in the leaf nodes, in the numerical section I have used an OLS regression, but I could have used instead, e.g., a LASSO regression [Tibshirani, 1996].

V.3 Numerical results

In this section, I present the obtained results for the FhteF for different values of the parameter α. In Section V.3.1, I first consider the simulated data generating model as in Athey and Imbens [2016]. Please note that this paper is not devoted to fairness, and therefore this simulated data does not take into account any sensitive attribute. In Section V.3.2, I modify this simulated data generating model to incorporate a sensitive attribute and different degrees of unfairness. I benchmark the FhteF against the Generalized Random Forest (GRF) [Athey et al., 2019, Wager and Athey, 2018]. To build the GRF I used the *grf* package [Tibshirani et al., 2023] for *R* [R Core Team, 2023].

The design of the experiments is as follows. I draw a training sample $\{(\boldsymbol{X}_i, Z_i, W_i, Y_i(W_i))\}_{i\in\mathcal{I}}$ with $n = 50,000$ individuals. I train an RF using the *scikit-learn* library [Pedregosa et al., 2011] with $\mathrm{T} \in \{1,000,\ 2,000\}$ regression trees. The trees are of unlimited depth but with a limit of having at least 50 observations in each leaf node to be able to estimate the treatment effect via an OLS regression. Also, all leaf nodes have at least 10 observations of the treated and non-treated groups respectively.

To solve the smooth reformulation of (V.2.3)–(V.2.5) *Gurobi* [Gurobi Optimization, 2020] and *Python* [Python Core Team, 2015] on a PC Intel®Core TM i7-8665U, 16GB of RAM are used. All the instances are solved to optimality within 5 seconds.

V.3.1 Results for the case without sensitive attributes

I consider the simulated data generating model introduced in Athey and Imbens [2016], which does not contain any sensitive attribute. The generating model has the form

$$Y(W) = \eta\left(\boldsymbol{X}\right) + \frac{1}{2}(2W-1)\kappa\left(\boldsymbol{X}\right) + \epsilon, \tag{V.3.1}$$

where $\boldsymbol{X} \sim \mathbb{N}(\mathbf{0}, \mathrm{I})$ is a multivariate normal distribution with mean vector $\mathbf{0} = (0, \ldots, 0)$ and covariance matrix equal to the identity matrix I, $\epsilon \sim \mathbb{N}(0, 0.01)$ is normally distributed, and W follows a Bernoulli distribution, $W \sim \mathbb{B}(0.5)$. I draw i.i.d. realizations from $\boldsymbol{X}$, ϵ and W, indexed by i, yielding $(\boldsymbol{X}_i, W_i, Y_i(W_i))$. The first $50,000$ vectors will define the training sample, and the remaining $50,000$ the testing sample.

I use the three types of simulated data in Athey and Imbens [2016], which I denote by $\mathbb{D}1$, $\mathbb{D}2$ and $\mathbb{D}3$, see Table V.1. They depend on the number of features p and the functional forms of $\eta(\cdot)$ and $\kappa(\cdot)$. Note that datasets $\mathbb{D}2$ and $\mathbb{D}3$ include noisy features that are used to build the model but not involved in the ground truth function. For each dataset in Table V.1, the HTE for an individual with $\boldsymbol{X} = \boldsymbol{x}$ as the vector of explanatory variables is known and given by the corresponding $\kappa(\boldsymbol{x})$.

Table V.1: Datasets without fairness considerations to test FhteF. The number of features p and the functional forms of $\eta(\boldsymbol{x})$ and $\kappa(\boldsymbol{x})$ for the simulated data generating model in (V.3.1) are displayed.

Dataset	p	$\eta(\boldsymbol{x})$	$\kappa(\boldsymbol{x})$
$\mathbb{D}1$	2	$\frac{1}{2}x_1 + x_2$	$\frac{1}{2}x_1$
$\mathbb{D}2$	10	$\frac{1}{2}\sum_{k=1}^{2} x_k + \sum_{k=3}^{6} x_k$	$\sum_{k=1}^{2} 1\{x_k > 0\} x_k$
$\mathbb{D}3$	20	$\frac{1}{2}\sum_{k=1}^{4} x_k + \sum_{k=5}^{8} x_k$	$\sum_{k=1}^{4} 1\{x_k > 0\} x_k$

The obtained results on the test sample can be seen in Table V.2. Since I have the data generating model, I can, for each individual, observe both $Y_i(1)$ and $Y_i(0)$, and thus the true HTE are observed: $\tau(\boldsymbol{x}) = \kappa(\boldsymbol{x})$. Hence, I report the average Euclidean distance between the true value and predicted value of the HTE to show how far the obtained predictions of the treatment effects are from the true values and refer to it as *error*. As one can see, the FhteF has the same result as the GRF for dataset $\mathbb{D}1$, while with more features the FhteF is less accurate but overall the error

is still low. Also, as expected, the FhteF is more accurate with $\mathrm{T} = 2,000$ trees than with $1,000$. In what follows, I proceed with $\mathrm{T} = 2,000$ trees.

Table V.2: Results on datasets without fairness considerations. The FhteF consists of $\mathrm{T} \in \{1,000,\ 2,000\}$ trees. The GRF consists of the default number of trees, $\mathrm{T} = 2,000$.

Dataset	T	$\text{FhteF}_{\text{error}}$	$\text{GRF}_{\text{error}}$
$\mathbb{D}1$	$1,000$ $2,000$	0.00006 0.00005	0.00005
$\mathbb{D}2$	$1,000$ $2,000$	0.0016 0.0015	0.0003
$\mathbb{D}3$	$1,000$ $2,000$	0.0029 0.0028	0.0009

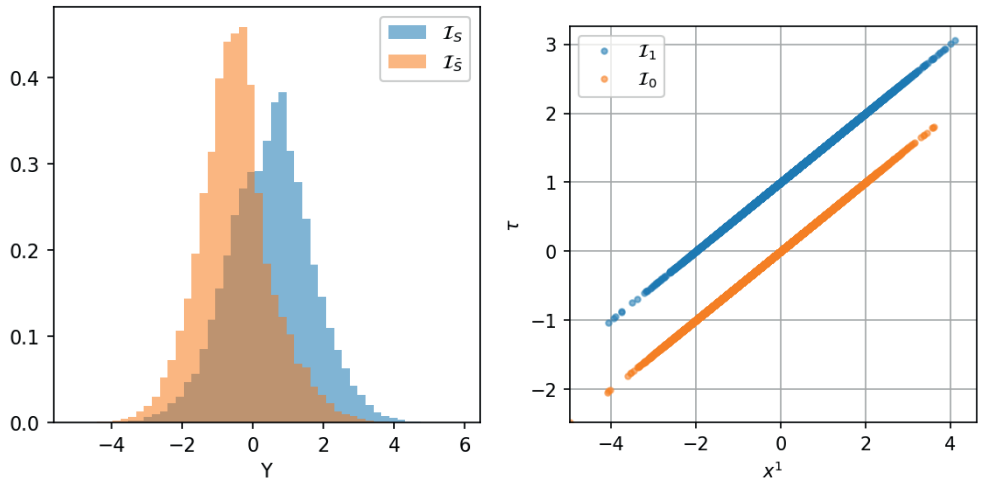
V.3.2 Results for the case with a sensitive attribute

In the fairness analysis, I still consider the simulated data generating model in (V.3.1) but with an adjustment that leads to unfairness in the data, those datasets I denote by $\mathbb{D}1\text{f}$, $\mathbb{D}2\text{f}$ and $\mathbb{D}3\text{f}$. The explanatory variables $\boldsymbol{X}$, the treatment variable W and the error term ϵ are the same but now there is an additional sensitive feature Z defined in Table V.3. Note, in datasets $\mathbb{D}2\text{f}$ and $\mathbb{D}3\text{f}$ there are noisy features as in datasets $\mathbb{D}2$ and $\mathbb{D}3$. As before, for each dataset in Table V.3, the true HTE for an individual with $\boldsymbol{X} = \boldsymbol{x}$ as the vector of explanatory variables and $Z = z$ as the sensitive group membership value is known and given by the corresponding $\kappa(\boldsymbol{x}, z)$.

Table V.3: Datasets with unfairness to test FhteF. The number of features p, the functional forms of $\eta(\boldsymbol{x})$ and $\kappa(\boldsymbol{x}, z)$ for the simulated data generating model in (V.3.1), and the probability distribution of the sensitive attribute are displayed. Note that a Bernoulli draw decides the membership the sensitive group.

Dataset	p	$\eta(\boldsymbol{x})$	$\kappa(\boldsymbol{x}, z)$	Z
$\mathbb{D}1\text{f}$	3	$\frac{1}{2}x_1 + x_2$	$\frac{1}{2}x_1 + z$	$\mathbb{B}\left(0.2 + 1\left\{\sum_{k=1}^{2} x_k \geq 0\right\} 0.6\right)$
$\mathbb{D}2\text{f}$	11	$\frac{1}{2}\sum_{k=1}^{2} x_k + \sum_{k=3}^{6} x_k$	$\sum_{k=1}^{2} 1\{x_k > 0\} x_k + z$	$\mathbb{B}\left(0.2 + 1\left\{\sum_{k=1}^{4} x_k \geq 0\right\} 0.6\right)$
$\mathbb{D}3\text{f}$	21	$\frac{1}{2}\sum_{k=1}^{4} x_k + \sum_{k=5}^{8} x_k$	$\sum_{k=1}^{4} 1\{x_k > 0\} x_k + z$	$\mathbb{B}\left(0.2 + 1\left\{\sum_{k=1}^{7} x_k \geq 0\right\} 0.6\right)$

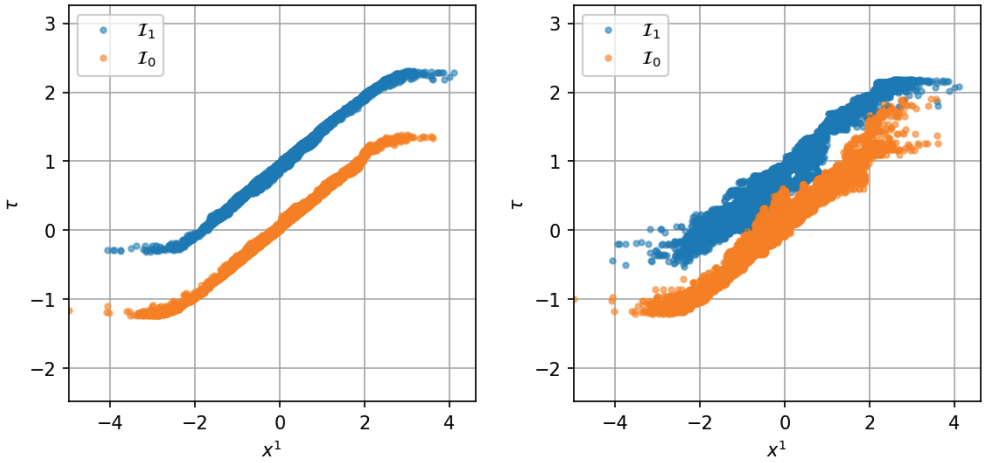
For simplicity, let me visualize the first unfair dataset, $\mathbb{D}1\text{f}$, which has three covariates, $p = 3$: x_1, x_2, z. Figure V.1 shows the output distribution and the true treatment effects $\kappa(\boldsymbol{x}, z) = \frac{1}{2}x_1 + z$. As one can see, there is a clear difference between the sensitive, $\mathcal{I}_1$, and non-sensitive, $\mathcal{I}_0$, groups in both functions. The FhteF is able to reduce this gap with different degree depending on parameter α, see Figure V.2.



(a) Y distribution

(b) $\tau(\boldsymbol{x}, z) = \frac{1}{2}x_1 + z$

Figure V.1: The distribution of the outcome Y and the true treatment effects for the sensitive and non-sensitive groups for dataset $\mathbb{D}$1f.



(a) predicted treatment effects, $\alpha = 2^{-6}$

(b) predicted treatment effects, $\alpha = 2^{-1}$

Figure V.2: The comparison of predicted treatment effects for the sensitive and non-sensitive groups for dataset $\mathbb{D}$1f.

In the following, I discuss the results for the FhteF with $\alpha \in \{0\} \cup \{2^n\}_{n=-6}^{4}$ and the GRF with $\mathrm{T} = 2,000$ trees, see Figure V.3. The upper plot of each subplot depicts the error for the FhteF (black line) and GRF (red line) showing how far the predicted treatment effects are from the true treatment effects. Despite one point ($\mathbb{D}$1f, $\alpha = 0$), the GRF error is lower than the FhteF error. The bottom plot of each subplot refers to the unfairness. Please note that the actual unfairness present in the dataset is equal to $|\frac{1}{n_1}\sum_{i\in\mathcal{I}_1}\kappa(\boldsymbol{X}_i, Z_i = 1) - \frac{1}{n_0}\sum_{i\in\mathcal{I}_0}\kappa(\boldsymbol{X}_i, Z_i = 0)|$. I then plot

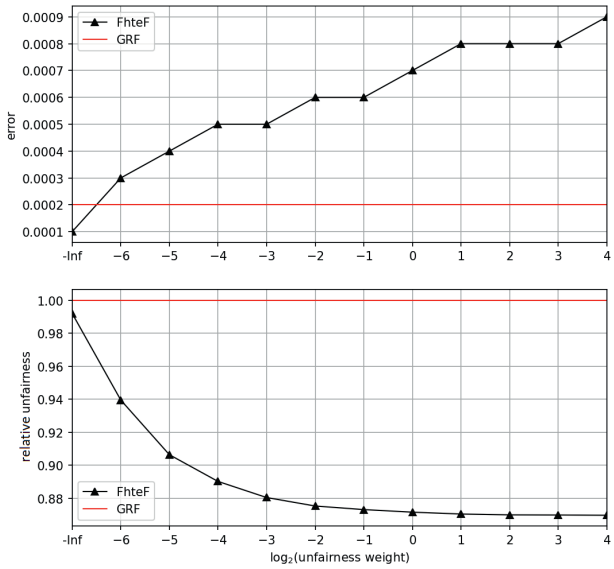
the unfairness of the predictions of the HTE relative to the actual unfairness for the FhteF (black line) and GRF (red line). With any value of α, the relative unfairness for the FhteF is lower than that of the GRF, which is very close to the actual unfairness. The behavior of the error and relative unfairness shows the clear trade-off between these two objectives in the FhteF: the higher the error, the lower the relative unfairness. Let me discuss the results for two values of α, namely, $\alpha = 0$, i.e., when we do not care about fairness, and $\alpha = 2^{-6}$, i.e., meaning that we put a small weight towards having fairer outcomes. With these two values of the parameter α the error does not differ much, but we observe a significant reduction in unfairness especially for the $\mathbb{D}1$f dataset. To end, it is worth mentioning that for the three datasets, the decrease in the relative unfairness slows down when $\alpha \geq 2^1$.

V.4 Conclusions

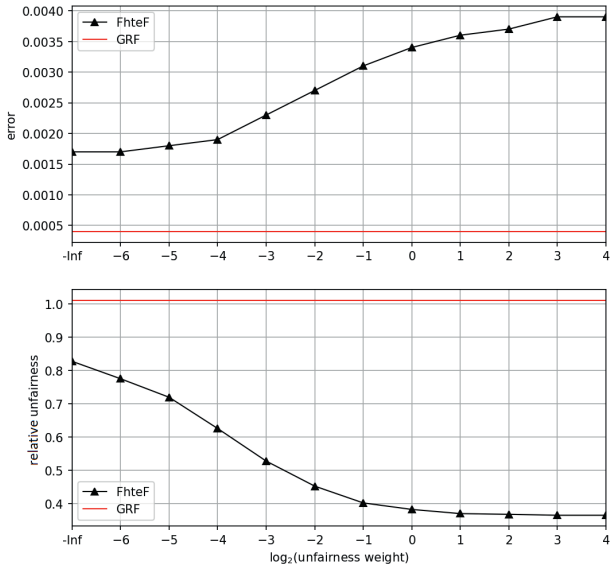
In this chapter, I propose the Fair Heterogeneous Treatment Effect Forest (FhteF) to predict treatment effects, where the goal is to minimize a weighted combination of the mean squared error of the outcome predictions as well as my fairness measure, namely, the absolute value of the difference between the in-sample average predicted treatment effect in the sensitive and in the non-sensitive observations. The main idea of the developed model is to reweight the trees of the initial RF predicting the treatment effects with linear models in the leaf nodes, such that the weight associated with fairer trees is higher. I model this as a Convex Quadratic Programming problem and present numerical results on simulated data, illustrating that I can reduce unfairness without sacrificing much accuracy. With the FhteF, researchers are able to properly address the existing unfairness in the data, to improve the availability of the treatment in the sensitive group, hopefully with a small decrease in the overall impact on the population.

There are some lines of interest for future research. The first one relates to handling more general cases of the treatment variable and of the sensitive group membership variable. In Chapter V, I have assumed two values for the treatment variable, namely, being treated or not. As for the sensitive attribute, I have assumed that observations are sensitive or not. Dealing with multiple values for the treatment variables and/or the existence of multiple sensitive groups requires extending the definition of the fairness measure in (V.2.2).

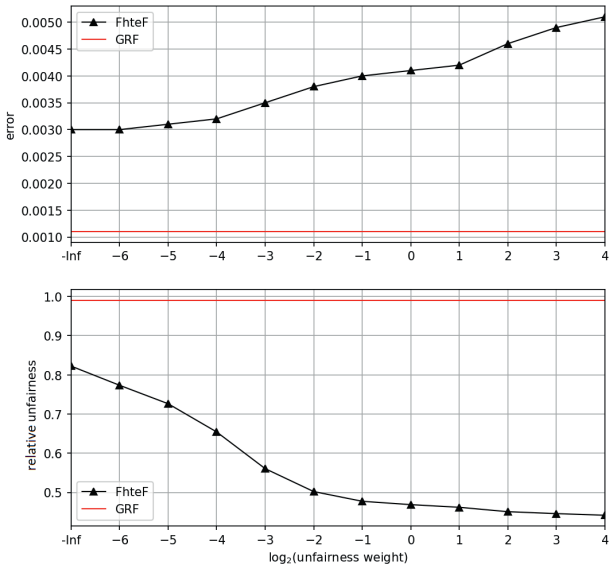
The second one refers to the prescription of the best treatment among the collection of available treatments. One option to make the prescription is to use the Empirical Welfare Maximization (EWM) approach [Kitagawa and Tetenov, 2018] which learns a treatment assignment policy by maximizing the data-driven average social welfare. This requires extending FhteF by adding the



(a) Dataset $\mathbb{D}1f$



(b) Dataset $\mathbb{D}2f$



(c) Dataset $\mathbb{D}3f$

Figure V.3: Results on datasets with unfairness. The FhteF and GRF consist of $T = 2,000$ trees.

EWM term to the objective function. Other natural extensions are incorporating the cost of treatment and/or constraints on the budget available.

Chapter VI

General conclusions and future work

In this Ph.D. dissertation, I enhance the transparency of well-known Machine Learning methodologies. More precisely, I use Mathematical Optimization to take into account explainability and fairness, the forms of transparency we consider, while preserving accuracy.

Chapters II and III are dedicated to enhancing explainability in Cluster Analysis. In both chapters, we assume that a dissimilarity between the individuals is given. We consider different types of explanations, namely prototype-based and rule-based ones. A prototype-based explanation is a distance-based explanation, where a prototype individual, close to the cluster, is chosen to represent it. A rule-based explanation is a feature-based explanation, where clauses defined by the features and joined by the AND operator, are assigned to the clusters.

In Chapter II, based on the work in Carrizosa et al. [2022b], we have proposed two models to find prototypes for each cluster, such that the selected prototype is close (similar) to its cluster and far (dissimilar) from other clusters. We develop two MILP models, inspired by classic Location Analysis problems, that differ in the way individuals are allocated to prototypes. In Chapter III, based on the work in Carrizosa et al. [2023a], we introduced an MILP model to simultaneously cluster individuals and provide rule-based explanations for the clusters. This approach can be applied either in a post-hoc manner to interpret existing clusters as done in Chapter II, or when clusters are sought along with explanations.

A possible extension concerning the results of Chapters II and III would be to consider other clustering objectives. In this dissertation, we examine the partition clustering type of models, but we could investigate, e.g., density-based clustering [Kriegel et al., 2011]. On the other hand, an alternative perspective can be to model fairness measures that are suitable for Cluster Analysis [Chhabra et al., 2021].

In Chapter IV, based on the work in Carrizosa et al. [2023b], we consider another Machine Learning task, namely classification and regression via Tree Ensembles, where we need to protect a group of observations sharing a sensitive attribute. In order to enhance explainability, we model sparsity such that we use fewer features. In order to enhance fairness, we aim for the accuracy in the sensitive group to be as high as possible. Thus, we optimize three objectives, namely, accuracy, sparsity, and fairness, via reweighting the trees in the ensemble. We propose an MILP formulation to train the Explainable and Fair Tree Ensemble (EFTE), where the misclassification error is modeled through continuous decision variables as opposed to binary ones.

In this chapter, we consider global sparsity as the proxy for explainability. As for future research, we could implement other types of sparsity. For instance, we could model forms of sparsity suitable for complex data such as hierarchical data [Carrizosa et al., 2022c].

In Chapter V, based on the work in Kurishchenko [2023], which is my solo-authored paper, I consider a treatment allocation problem. I propose the Fair Heterogeneous Treatment Effect Forest (FhteF) to predict treatment effects, where the goal is to have high accuracy and fairness. I use Tree Ensembles, where in the leaves linear models are used to predict the treatment effect. I reweight the obtained trees such that the total accuracy is maximized as well as the fairness. I model this as a Convex Quadratic Programming problem.

Finally, the results of Chapter V can be extended as follows. Firstly, I could generalized the model to the multivariate or continuous cases of the treatment variable and the sensitive group membership variable. Secondly, one can reformulate the model by adding the Empirical Welfare Maximization term. With this, the model would give direct prescriptions for the best treatment.

To end, there exist interesting extensions of the works presented in this Ph.D. dissertation, e.g., to model Differential Privacy or other desirable properties of algorithmic decisions.

Bibliography

S. S. Abraham, D. Padmanabhan, and S. S. Sundaram. Fairness in clustering with multiple sensitive attributes. In *EDBT/ICDT 2020 Joint Conference*, pages 287–298, 2020.

D. Aloise, A. Deshpande, P. Hansen, and P. Popat. NP-hardness of Euclidean sum-of-squares clustering. *Machine Learning*, 75(2):245–248, 2009.

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

A. Altmann, L. Toloşi, O. Sander, and T. Lengauer. Permutation importance: a corrected feature importance measure. *Bioinformatics*, 26(10):1340–1347, 2010.

J. Angwin, J. Larson, S. Mattu, and L. Kirchner. Machine bias. In *Ethics of Data and Analytics*, pages 254–264. Auerbach Publications, 2016.

S. Athey. Beyond prediction: Using big data for policy problems. *Science*, 355(6324):483–485, 2017.

S. Athey and G. Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016.

S. Athey, J. Tibshirani, and S. Wager. Generalized random forests. *The Annals of Statistics*, 47 (2):1148–1178, 2019.

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

K. Balabaeva and S. Kovalchuk. Post-hoc interpretation of clinical pathways clustering using bayesian inference. *Procedia Computer Science*, 178:264–273, 2020.

S. Barocas, M. Hardt, and A. Narayanan. *Fairness and machine learning: Limitations and opportunities*. MIT Press, 2023.

A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, R. Chatila, and F. Herrera. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58:82–115, 2020.

J. Basak and R. Krishnapuram. Interpretable hierarchical clustering by constructing an unsupervised decision tree. *IEEE Transactions on Knowledge and Data Engineering*, 17(1):121–132, 2005.

E. Baumgaertner. Medication treatment for addiction is shorter for black and hispanic patients, study finds. *The New York Times*, 2022. URL `https://www.nytimes.com/2022/11/09/health/opioid-addiction-treatment-racial-disparities.html?smid=url-share`.

C. Bénard, G. Biau, S. Da Veiga, and E. Scornet. Sirus: Stable and interpretable rule set for classification. *Electronic Journal of Statistics*, 15(1):427–505, 2021.

S. Benítez-Peña, R. Blanquero, E. Carrizosa, and P. Ramírez-Cobo. Cost-sensitive feature selection for Support Vector Machines. *Computers & Operations Research*, 106:169–178, 2019.

S. Benítez-Peña, P. Bogetoft, and D. Romero Morales. Feature selection in data envelopment analysis: A mathematical optimization approach. *Omega*, 96:102068, 2020.

S. Benítez-Peña, E. Carrizosa, V. Guerrero, M. D. Jiménez-Gamero, B. Martín-Barragán, C. Molero-Río, P. Ramírez-Cobo, D. Romero Morales, and M. R. Sillero-Denamiel. On sparse ensemble methods: An application to short-term predictions of the evolution of COVID-19. *European Journal of Operational Research*, 295(2):648–663, 2021.

D. Bertsimas, J. Dunn, and N. Mundru. Optimal prescriptive trees. *INFORMS Journal on Optimization*, 1(2):164–183, 2019.

D. Bertsimas, A. Orfanoudaki, and H. Wiberg. Interpretable clustering: an optimization approach. *Machine Learning*, 110(1):89–138, 2021.

D. Bertsimas, J. Pauphilet, J. Stevens, and M. Tandon. Predicting inpatient flow at a major hospital using interpretable analytics. *Manufacturing & Service Operations Management*, 24(6): 2809–2824, 2022.

P. Besse, E. del Barrio, P. Gordaliza, J.-M. Loubes, and L. Risser. A survey of bias in machine learning through the prism of statistical parity. *The American Statistician*, 76(2):188–198, 2022.

G. Biau and E. Scornet. A random forest guided tour. *TEST*, 25(2):197–227, 2016.

R. Blanquero, E. Carrizosa, C. Molero-Río, and D. Romero Morales. Sparsity in optimal randomized classification trees. *European Journal of Operational Research*, 284(1):255–272, 2020.

L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

E. Carrizosa and D. Romero Morales. Supervised classification and mathematical optimization. *Computers & Operations Research*, 40(1):150–165, 2013.

E. Carrizosa, B. Martín-Barragán, D. Romero Morales, and F. Plastria. On the selection of the globally optimal prototype subset for nearest-neighbor classification. *INFORMS Journal on Computing*, 19(3):470–479, 2007.

E. Carrizosa, A. Nogales-Gómez, and D. Romero Morales. Strongly agree or strongly disagree?: Rating features in Support Vector Machines. *Information Sciences*, 329:256–273, 2016.

E. Carrizosa, V. Guerrero, D. Romero Morales, and A. Satorra. Enhancing interpretability in factor analysis by means of mathematical optimization. *Multivariate Behavioral Research*, 55 (5):748–762, 2020.

E. Carrizosa, M. Galvis Restrepo, and D. Romero Morales. On clustering categories of categorical predictors in generalized linear models. *Expert Systems with Applications*, 182:115245, 2021a.

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

E. Carrizosa, M. Galvis Restrepo, and D. Romero Morales. Improving fairness of generalized linear models by feature shrinkage. Technical report, Copenhagen Business School, Denmark, `https://www.researchgate.net/publication/358614960_Improving_fairness_of_Generalized_Linear_Models_by_feature_shrinkage`, 2022a.

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

E. Carrizosa, L. H. Mortensen, D. Romero Morales, and M. R. Sillero-Denamiel. The tree based linear regression model for hierarchical categorical variables. *Expert Systems With Applications*, 203(7):117423, 2022c.

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

E. Carrizosa, K. Kurishchenko, and D. Romero Morales. On enhancing the explainability and fairness of tree ensembles. Technical report, Copenhagen Business School, Denmark, 2023b. URL `https://www.researchgate.net/publication/374013681_On_enhancing_the_explainability_and_fairness_of_tree_ensembles`.

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

T. Chen and C. Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, 2016.

A. Chhabra, K. Masalkovaitė, and P. Mohapatra. An overview of fairness in clustering. *IEEE Access*, 9:130698–130720, 2021.

E. Constantaras, G. Geiger, J.-C. Braun, D. Mehrotra, and H. Aung. Inside the suspicion machin. *Wired*, 2023. URL `https://www.wired.com/`.

S. Corbett-Davies and S. Goel. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023*, 2018.

G. Corral, E. Armengol, A. Fornells, and E. Golobardes. Explanations of unsupervised learning clustering applied to data security analysis. *Neurocomputing*, 72(13):2754–2762, 2009.

C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20:273–297, 1995.

T. Cover and P. Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13:21–27, 1967.

S. Das, R. Stanton, and N. Wallace. Algorithmic fairness. *Annual Review of Financial Economics*, 15:565–593, 2023.

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

I. Davidson, A. Gourru, and S. Ravi. The cluster description problem - complexity results, formulations and approximations. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

P. De Koninck, J. De Weerdt, and S. L. vanden Broucke. Explaining clusterings of process instances. *Data Mining and Knowledge Discovery*, 31(3):774–808, 2017.

G. Di Teodoro, M. Monaci, and L. Palagi. Unboxing tree ensembles for interpretability: a hierarchical visualization tool and a multivariate optimal re-built tree. *EURO Journal on Computational Optimization*, 12:100084, 2024.

D. Dua and C. Graff. UCI Machine Learning Repository, 2017. URL `http://archive.ics.uci.edu/ml`.

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

M. Febrero-Bande and M. Oviedo de la Fuente. Statistical computing in functional data analysis: The R package fda.usc. *Journal of Statistical Software*, 51(4):1–28, 2012.

C. Fernández-Loría and F. Provost. Causal decision making and causal effect estimation are not the same... and why it matters. *INFORMS Journal on Data Science*, 1(1):4–16, 2022.

R. Fortet. Applications de l'algebre de boole en recherche opérationelle. *Revue Française de Recherche Opérationelle*, 4(14):17–26, 1960.

R. Fraiman, B. Ghattas, and M. Svarc. Interpretable clustering using unsupervised binary trees. *Advances in Data Analysis and Classification*, 7(2):125–145, 2013.

A. A. Freitas. Comprehensible classification models: a position paper. *ACM SIGKDD Explorations Newsletter*, 15(1):1–10, 2014.

J. Friedman. Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, 29(5):1189–1232, 2001.

J. Friedman, T. Hastie, and R. Tibshirani. A note on the group lasso and a sparse group lasso. *arXiv preprint arXiv:1001.0736*, 2010.

M. Frölich. Finite-sample properties of propensity-score matching and weighting estimators. *Review of Economics and Statistics*, 86(1):77–90, 2004.

C. Gambella, B. Ghaddar, and J. Naoum-Sawaya. Optimization models for machine learning: A survey. *European Journal of Operational Research*, 290(3):807–828, 2021.

G. Gan, C. Ma, and J. Wu. *Data clustering: theory, algorithms, and applications*. SIAM, 2007.

S. García and A. Marín. Covering location problems. *Location Science*, pages 99–119, 2019.

S. García, M. Labbé, and A. Marín. Solving large p-median problems with a radius formulation. *INFORMS Journal on Computing*, 23(4):546–556, 2011.

A. Gelman, J. Fagan, and A. Kiss. An analysis of the new york city police department's "stop-and-frisk" policy in the context of claims of racial bias. *Journal of the American Statistical Association*, 102(479):813–823, 2007.

K. Gibert and D. Conti. On the understanding of profiles by means of post-processing techniques: an application to financial assets. *International Journal of Computer Mathematics*, 93(5):807–820, 2016.

I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016.

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

A. Gordon, L. Breiman, J. Friedman, R. Olshen, and C. J. Stone. Classification and regression trees. *Biometrics*, 40(3):874, 1984.

M. Grötschel and Y. Wakabayashi. A cutting plane algorithm for a clustering problem. *Mathematical Programming*, 45(1):59–96, 1989.

D. Gunning, M. Stefik, J. Choi, T. Miller, S. Stumpf, and G.-Z. Yang. XAI—explainable artificial intelligence. *Science Robotics*, 4(37), 2019.

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

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, New York, 2nd edition, 2009.

T. Hastie, R. Tibshirani, and M. Wainwright. *Statistical learning with sparsity: the lasso and generalizations*. Chapman and Hall/CRC, 2019.

H. Heaton and S. W. Fung. Explainable AI via learning to optimize. *Scientific Reports*, 13(1): 10103, 2023.

M. Hort, Z. Chen, J. M. Zhang, F. Sarro, and M. Harman. Bias mitigation for machine learning classifiers: A comprehensive survey. *arXiv preprint arXiv:2207.07068*, 2022.

B. Hutchinson, A. Smart, A. Hanna, E. Denton, C. Greer, O. Kjartansson, P. Barnes, and M. Mitchell. Towards accountability for machine learning datasets: Practices from software engineering and infrastructure. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 560–575, 2021.

G. W. Imbens and D. B. Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

A. Jain. Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, 31(8):651–666, 2010.

A. Jiménez-Cordero, J. M. Morales, and S. Pineda. A novel embedded min-max approach for feature selection in nonlinear support vector machine classification. *European Journal of Operational Research*, 293(1):24–35, 2021.

N. Jo, S. Aghaei, A. Gómez, and P. Vayanos. Learning optimal prescriptive trees from observational data. *arXiv preprint arXiv:2108.13628*, 2021.

M. Jordan and T. Mitchell. Machine learning: Trends, perspectives, and prospects. *Science*, 349 (6245):255–260, 2015.

J. Kauffmann, M. Esders, L. Ruff, G. Montavon, W. Samek, and K.-R. Müller. From clustering to cluster explanations via neural networks. Forthcoming in *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

L. Kaufmann and P. J. Rousseeuw. Finding groups in data: an introduction to cluster analysis. *Wiley, New York*, 1990.

N. Kayser-Bril. Austria's employment agency rolls out discriminatory algorithm, sees no problem. *Algorithm Watch*, 2020. URL `https://algorithmwatch.org/en/austrias-employment-agency-ams-rolls-out-discriminatory-algorithm`.

A. Khademi, S. Lee, D. Foley, and V. Honavar. Fairness in algorithmic decision making: An excursion through the lens of causality. In *The World Wide Web Conference*, pages 2907–2914, 2019.

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

K. Kim and J. R. Zubizarreta. Fair and robust estimation of heterogeneous treatment effects for policy learning. *arXiv preprint arXiv:2306.03625*, 2023.

T. Kitagawa and A. Tetenov. Who should be treated? Empirical welfare maximization methods for treatment choice. *Econometrica*, 86(2):591–616, 2018.

J. Kleinberg, J. Ludwig, S. Mullainathan, and Z. Obermeyer. Prediction policy problems. *American Economic Review*, 105(5):491–495, 2015.

J. Kleinberg, H. Lakkaraju, J. Leskovec, J. Ludwig, and S. Mullainathan. Human decisions and machine predictions. *The Quarterly Journal of Economics*, 133(1):237–293, 2018.

H.-P. Kriegel, P. Kröger, J. Sander, and A. Zimek. Density-based clustering. *WIREs Data Mining and Knowledge Discovery*, 1(3):231–240, 2011.

S. R. Künzel, J. S. Sekhon, P. J. Bickel, and B. Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, 116 (10):4156–4165, 2019.

K. Kurishchenko. On enhancing fairness in heterogeneous treatment effects via ensembles. Technical report, Copenhagen Business School, Denmark, 2023. URL `https://www.researchgate.net/publication/377306392_On_Enhancing_Fairness_in_Heterogeneous_Treatment_Effects_via_Ensembles`.

H. Lakkaraju, E. Kamar, R. Caruana, and J. Leskovec. Interpretable & explorable approximations of black box models. *arXiv preprint arXiv:1707.01154*, 2017.

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

T. Le Quy, A. Roy, V. Iosifidis, W. Zhang, and E. Ntoutsi. A survey on datasets for fairness-aware machine learning. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 12 (3):e1452, 2022.

B. Lepri, N. Oliver, E. Letouzé, A. Pentland, and P. Vinck. Fair, transparent, and accountable algorithmic decision-making processes. *Philosophy & Technology*, 31(4):611–627, 2017.

B. Liu and R. Mazumder. ForestPrune: Compact Depth-Pruned Tree Ensembles. In F. Ruiz, J. Dy, and J.-W. van de Meent, editors, *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 9417–9428. PMLR, 2023.

S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30:4765–4774, 2017.

R. Ma, R. A. Angryk, P. Riley, and S. F. Boubrahimi. Coronal mass ejection data clustering and visualization of decision trees. *The Astrophysical Journal Supplement Series*, 236(1):14, 2018.

S. Maldonado, E. Carrizosa, and R. Weber. Kernel penalized k-means: A feature selection method based on kernel k-means. *Information Sciences*, 322:150–160, 2015.

A. Marín and M. Pelegrín. p-median problems. In G. Laporte, S. Nickel, and F. Saldanha da Gama, editors, *Location Science*, pages 25–50. Springer International Publishing, Cham, 2019.

N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54:1–35, 2022.

M. Miron, S. Tolan, E. Gómez, and C. Castillo. Addressing multiple metrics of group fairness in data-driven decision making. *arXiv preprint arXiv:2003.04794*, 2020.

V. V. Mišić. Optimization of tree ensembles. *Operations Research*, 68(5):1605–1624, 2020.

A. Morichetta, P. Casas, and M. Mellia. EXPLAIN-IT: Towards Explainable AI for Unsupervised Network Traffic Analysis. *Proceedings of the 3rd ACM CoNEXT Workshop on Big DAta, Machine Learning and Artificial Intelligence for Data Communication Networks - Big-DAMA '19*, pages 22–28, 2019.

R. Nabi, D. Malinsky, and I. Shpitser. Learning optimal fair policies. In *International Conference on Machine Learning*, pages 4674–4682. PMLR, 2019.

Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019.

C. Panigutti, R. Hamon, I. Hupont, D. Fernandez Llorca, D. Fano Yela, H. Junklewitz, S. Scalzo, G. Mazzini, I. Sanchez, J. Soler Garrido, and E. Gomez. The Role of Explainable AI in the

Context of the AI Act. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, pages 1139–1150, New York, NY, USA, 2023.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Python Core Team. Python: A dynamic, open source programming language. Python Software Foundation, 2015. URL `https://www.python.org`.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2023. URL `https://www.R-project.org/`.

M. Rao. Cluster analysis and mathematical programming. *Journal of the American Statistical Association*, 66(335):622–626, 1971.

M. T. Ribeiro, S. Singh, and C. Guestrin. "Why should I trust you?" Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144, 2016.

A. Romei and S. Ruggieri. A multidisciplinary survey on discrimination analysis. *The Knowledge Engineering Review*, 29(5):582–638, 2014.

P. J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987.

C. Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

C. Rudin, C. Chen, Z. Chen, H. Huang, L. Semenova, and C. Zhong. Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistics Surveys*, 16:1–85, 2022.

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

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

O. Seref, Y.-J. Fan, and W. A. Chaovalitwongse. Mathematical programming formulations and algorithms for discrete k-median clustering of time-series data. *INFORMS Journal on Computing*, 26(1):160–172, 2014.

D. Shin. The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. *International Journal of Human-Computer Studies*, 146:102551, 2021.

A. Taeb and V. Chandrasekaran. Interpreting latent variables in factor models via convex optimization. *Mathematical Programming*, 167(1):129–154, 2018.

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

J. Tibshirani, S. Athey, E. Sverdrup, and S. Wager. *grf: Generalized Random Forests*, 2023. URL `https://CRAN.R-project.org/package=grf`. R package version 2.3.0.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

C. Tran, K. Burghardt, K. Lerman, and E. Zheleva. Data-driven estimation of heterogeneous treatment effects. *arXiv preprint arXiv:2301.06615*, 2023.

P. Turney. Cost-sensitive classification: Empirical evaluation of a hybrid genetic decision tree induction algorithm. *Journal of Artificial Intelligence Research*, 2:369–409, 1995.

I. Usiskin and D. Misra. Racial disparities in elective total joint arthroplasty for osteoarthritis. *ACR Open Rheumatology*, 4(4):306–311, 2022.

V. Vapnik. *The Nature of Statistical Learning Theory*. Springer Verlag, 1995.

V. Vapnik. *Statistical Learning Theory*. Wiley, 1998.

T. Vidal and M. Schiffer. Born-again tree ensembles. In *International Conference on Machine Learning*, pages 9743–9753. PMLR, 2020.

S. Wager and S. Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

J. Wagner and L. Falkson. The optimal nodal location of public facilities with price-sensitive demand. *Geographical Analysis*, 7(1):69–83, 1975.

M. Wu, S. Parbhoo, M. C. Hughes, V. Roth, and F. Doshi-Velez. Optimizing for interpretability in deep neural networks with tree regularization. *Journal of Artificial Intelligence Research*, 72: 1–37, 2021.

Y. Xie, J. E. Brand, and B. Jann. Estimating heterogeneous treatment effects with observational data. *Sociological Methodology*, 42(1):314–347, 2012.

M. B. Zafar, I. Valera, M. Gomez Rodriguez, and K. P. Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web*, pages 1171–1180. International World Wide Web Conferences Steering Committee, 2017.

TITLER I PH.D.SERIEN:

2004

1. Martin Grieger
Internet-based Electronic Marketplaces and Supply Chain Management

2. Thomas Basbøll
LIKENESS
A Philosophical Investigation

3. Morten Knudsen
Beslutningens vaklen
En systemteoretisk analyse of moderniseringen af et amtskommunalt sundhedsvæsen 1980-2000

4. Lars Bo Jeppesen
Organizing Consumer Innovation
A product development strategy that is based on online communities and allows some firms to benefit from a distributed process of innovation by consumers

5. Barbara Dragsted
SEGMENTATION IN TRANSLATION AND TRANSLATION MEMORY SYSTEMS
An empirical investigation of cognitive segmentation and effects of integrating a TM system into the translation process

6. Jeanet Hardis
Sociale partnerskaber
Et socialkonstruktivistisk casestudie af partnerskabsaktørers virkelighedsopfattelse mellem identitet og legitimitet

7. Henriette Hallberg Thygesen
System Dynamics in Action

8. Carsten Mejer Plath
Strategisk Økonomistyring

9. Annemette Kjærgaard
Knowledge Management as Internal Corporate Venturing
– a Field Study of the Rise and Fall of a Bottom-Up Process

10. Knut Arne Hovdal
De profesjonelle i endring
Norsk ph.d., ej til salg gennem Samfundslitteratur

11. Søren Jeppesen
Environmental Practices and Greening Strategies in Small Manufacturing Enterprises in South Africa
– A Critical Realist Approach

12. Lars Frode Frederiksen
Industriel forskningsledelse
– på sporet af mønstre og samarbejde i danske forskningsintensive virksomheder

13. Martin Jes Iversen
The Governance of GN Great Nordic
– in an age of strategic and structural transitions 1939-1988

14. Lars Pynt Andersen
The Rhetorical Strategies of Danish TV Advertising
A study of the first fifteen years with special emphasis on genre and irony

15. Jakob Rasmussen
Business Perspectives on E-learning

16. Sof Thrane
The Social and Economic Dynamics of Networks
– a Weberian Analysis of Three Formalised Horizontal Networks

17. Lene Nielsen
Engaging Personas and Narrative Scenarios – a study on how a user-centered approach influenced the perception of the design process in the e-business group at AstraZeneca

18. S.J Valstad
Organisationsidentitet
Norsk ph.d., ej til salg gennem Samfundslitteratur

19. Thomas Lyse Hansen
Six Essays on Pricing and Weather risk in Energy Markets

20. Sabine Madsen
Emerging Methods – An Interpretive Study of ISD Methods in Practice

21. Evis Sinani
The Impact of Foreign Direct Investment on Efficiency, Productivity Growth and Trade: An Empirical Investigation

22. Bent Meier Sørensen
Making Events Work Or, How to Multiply Your Crisis

23. Pernille Schnoor
Brand Ethos
Om troværdige brand- og virksomhedsidentiteter i et retorisk og diskursteoretisk perspektiv

24. Sidsel Fabech
Von welchem Österreich ist hier die Rede?
Diskursive forhandlinger og magtkampe mellem rivaliserende nationale identitetskonstruktioner i østrigske pressediskurser

25. Klavs Odgaard Christensen
Sprogpolitik og identitetsdannelse i flersprogede forbundsstater
Et komparativt studie af Schweiz og Canada

26. Dana B. Minbaeva
Human Resource Practices and Knowledge Transfer in Multinational Corporations

27. Holger Højlund
Markedets politiske fornuft
Et studie af velfærdens organisering i perioden 1990-2003

28. Christine Mølgaard Frandsen
A.s erfaring
Om mellemværendets praktik i en transformation af mennesket og subjektiviteten

29. Sine Nørholm Just
The Constitution of Meaning – A Meaningful Constitution?
Legitimacy, identity, and public opinion in the debate on the future of Europe

2005

1. Claus J. Varnes
Managing product innovation through rules – The role of formal and structured methods in product development

2. Helle Hedegaard Hein
Mellem konflikt og konsensus – Dialogudvikling på hospitalsklinikker

3. Axel Rosenø
Customer Value Driven Product Innovation – A Study of Market Learning in New Product Development

4. Søren Buhl Pedersen
Making space
An outline of place branding

5. Camilla Funck Ellehave
Differences that Matter
An analysis of practices of gender and organizing in contemporary workplaces

6. Rigmor Madeleine Lond
Styring af kommunale forvaltninger

7. Mette Aagaard Andreassen
Supply Chain versus Supply Chain Benchmarking as a Means to Managing Supply Chains

8. Caroline Aggestam-Pontoppidan
From an idea to a standard
The UN and the global governance of accountants' competence

9. Norsk ph.d.

10. Vivienne Heng Ker-ni
An Experimental Field Study on the

Effectiveness of Grocer Media Advertising
Measuring Ad Recall and Recognition, Purchase Intentions and Short-Term Sales

11. Allan Mortensen
Essays on the Pricing of Corporate Bonds and Credit Derivatives

12. Remo Stefano Chiari
Figure che fanno conoscere
Itinerario sull'idea del valore cognitivo e espressivo della metafora e di altri tropi da Aristotele e da Vico fino al cognitivismo contemporaneo

13. Anders McIlquham-Schmidt
Strategic Planning and Corporate Performance
An integrative research review and a meta-analysis of the strategic planning and corporate performance literature from 1956 to 2003

14. Jens Geersbro
The TDF – PMI Case
Making Sense of the Dynamics of Business Relationships and Networks

15 Mette Andersen
Corporate Social Responsibility in Global Supply Chains
Understanding the uniqueness of firm behaviour

16. Eva Boxenbaum
Institutional Genesis: Micro – Dynamic Foundations of Institutional Change

17. Peter Lund-Thomsen
Capacity Development, Environmental Justice NGOs, and Governance: The Case of South Africa

18. Signe Jarlov
Konstruktioner af offentlig ledelse

19. Lars Stæhr Jensen
Vocabulary Knowledge and Listening Comprehension in English as a Foreign Language
An empirical study employing data elicited from Danish EFL learners

20. Christian Nielsen
Essays on Business Reporting
Production and consumption of strategic information in the market for information

21. Marianne Thejls Fischer
Egos and Ethics of Management Consultants

22. Annie Bekke Kjær
Performance management i Proces-innovation
– belyst i et social-konstruktivistisk perspektiv

23. Suzanne Dee Pedersen
GENTAGELSENS METAMORFOSE
Om organisering af den kreative gøren i den kunstneriske arbejdspraksis

24. Benedikte Dorte Rosenbrink
Revenue Management
Økonomiske, konkurrencemæssige & organisatoriske konsekvenser

25. Thomas Riise Johansen
Written Accounts and Verbal Accounts
The Danish Case of Accounting and Accountability to Employees

26. Ann Fogelgren-Pedersen
The Mobile Internet: Pioneering Users' Adoption Decisions

27. Birgitte Rasmussen
Ledelse i fællesskab – de tillidsvalgtes fornyende rolle

28. Gitte Thit Nielsen
Remerger
– skabende ledelseskræfter i fusion og opkøb

29. Carmine Gioia
A MICROECONOMETRIC ANALYSIS OF MERGERS AND ACQUISITIONS

30. Ole Hinz
Den effektive forandringsleder: pilot, pædagog eller politiker?
Et studie i arbejdslederes meningstilskrivninger i forbindelse med vellykket gennemførelse af ledelsesinitierede forandringsprojekter

31. Kjell-Åge Gotvassli
Et praksisbasert perspektiv på dynamiske
læringsnettverk i toppidretten
Norsk ph.d., ej til salg gennem Samfundslitteratur

32. Henriette Langstrup Nielsen
Linking Healthcare
An inquiry into the changing performances of web-based technology for asthma monitoring

33. Karin Tweddell Levinsen
Virtuel Uddannelsespraksis
Master i IKT og Læring – et casestudie i hvordan proaktiv proceshåndtering kan forbedre praksis i virtuelle læringsmiljøer

34. Anika Liversage
Finding a Path
Labour Market Life Stories of Immigrant Professionals

35. Kasper Elmquist Jørgensen
Studier i samspillet mellem stat og erhvervsliv i Danmark under
1. verdenskrig

36. Finn Janning
A DIFFERENT STORY
Seduction, Conquest and Discovery

37. Patricia Ann Plackett
Strategic Management of the Radical Innovation Process
Leveraging Social Capital for Market Uncertainty Management

2006

1. Christian Vintergaard
Early Phases of Corporate Venturing

2. Niels Rom-Poulsen
Essays in Computational Finance

3. Tina Brandt Husman
Organisational Capabilities, Competitive Advantage & Project-Based Organisations
The Case of Advertising and Creative Good Production

4. Mette Rosenkrands Johansen
Practice at the top
– how top managers mobilise and use non-financial performance measures

5. Eva Parum
Corporate governance som strategisk kommunikations- og ledelsesværktøj

6. Susan Aagaard Petersen
Culture's Influence on Performance Management: The Case of a Danish Company in China

7. Thomas Nicolai Pedersen
The Discursive Constitution of Organizational Governance – Between unity and differentiation
The Case of the governance of environmental risks by World Bank environmental staff

8. Cynthia Selin
Volatile Visions: Transactons in Anticipatory Knowledge

9. Jesper Banghøj
Financial Accounting Information and Compensation in Danish Companies

10. Mikkel Lucas Overby
Strategic Alliances in Emerging High-Tech Markets: What's the Difference and does it Matter?

11. Tine Aage
External Information Acquisition of Industrial Districts and the Impact of Different Knowledge Creation Dimensions

A case study of the Fashion and Design Branch of the Industrial District of Montebelluna, NE Italy

12. Mikkel Flyverbom
Making the Global Information Society Governable
On the Governmentality of Multi-Stakeholder Networks

13. Anette Grønning
Personen bag
Tilstedevær i e-mail som interaktionsform mellem kunde og medarbejder i dansk forsikringskontekst

14. Jørn Helder
One Company – One Language?
The NN-case

15. Lars Bjerregaard Mikkelsen
Differing perceptions of customer value
Development and application of a tool for mapping perceptions of customer value at both ends of customer-supplier dyads in industrial markets

16. Lise Granerud
Exploring Learning
Technological learning within small manufacturers in South Africa

17. Esben Rahbek Pedersen
Between Hopes and Realities: Reflections on the Promises and Practices of Corporate Social Responsibility (CSR)

18. Ramona Samson
The Cultural Integration Model and European Transformation.
The Case of Romania

2007

1. Jakob Vestergaard
Discipline in The Global Economy
Panopticism and the Post-Washington Consensus

2. Heidi Lund Hansen
Spaces for learning and working
A qualitative study of change of work, management, vehicles of power and social practices in open offices

3. Sudhanshu Rai
Exploring the internal dynamics of software development teams during user analysis
A tension enabled Institutionalization Model; "Where process becomes the objective"

4. Norsk ph.d.
Ej til salg gennem Samfundslitteratur

5. Serden Ozcan
EXPLORING HETEROGENEITY IN ORGANIZATIONAL ACTIONS AND OUTCOMES
A Behavioural Perspective

6. Kim Sundtoft Hald
Inter-organizational Performance Measurement and Management in Action
– An Ethnography on the Construction of Management, Identity and Relationships

7. Tobias Lindeberg
Evaluative Technologies
Quality and the Multiplicity of Performance

8. Merete Wedell-Wedellsborg
Den globale soldat
Identitetsdannelse og identitetsledelse i multinationale militære organisationer

9. Lars Frederiksen
Open Innovation Business Models
Innovation in firm-hosted online user communities and inter-firm project ventures in the music industry
– A collection of essays

10. Jonas Gabrielsen
Retorisk toposlære – fra statisk 'sted' til persuasiv aktivitet

11. Christian Moldt-Jørgensen
Fra meningsløs til meningsfuld evaluering.
Anvendelsen af studentertilfredshedsmålinger på de korte og mellemlange videregående uddannelser set fra et psykodynamisk systemperspektiv

12. Ping Gao
Extending the application of actor-network theory
Cases of innovation in the telecommunications industry

13. Peter Mejlby
Frihed og fængsel, en del af den samme drøm?
Et phronetisk baseret casestudie af frigørelsens og kontrollens sameksistens i værdibaseret ledelse!

14. Kristina Birch
Statistical Modelling in Marketing

15. Signe Poulsen
Sense and sensibility:
The language of emotional appeals in insurance marketing

16. Anders Bjerre Trolle
Essays on derivatives pricing and dynamic asset allocation

17. Peter Feldhütter
Empirical Studies of Bond and Credit Markets

18. Jens Henrik Eggert Christensen
Default and Recovery Risk Modeling and Estimation

19. Maria Theresa Larsen
Academic Enterprise: A New Mission for Universities or a Contradiction in Terms?
Four papers on the long-term implications of increasing industry involvement and commercialization in academia

20. Morten Wellendorf
Postimplementering af teknologi i den offentlige forvaltning
Analyser af en organisations kontinuerlige arbejde med informationsteknologi

21. Ekaterina Mhaanna
Concept Relations for Terminological Process Analysis

22. Stefan Ring Thorbjørnsen
Forsvaret i forandring
Et studie i officerers kapabiliteter under påvirkning af omverdenens forandringspres mod øget styring og læring

23. Christa Breum Amhøj
Det selvskabte medlemskab om managementstaten, dens styringsteknologier og indbyggere

24. Karoline Bromose
Between Technological Turbulence and Operational Stability
– An empirical case study of corporate venturing in TDC

25. Susanne Justesen
Navigating the Paradoxes of Diversity in Innovation Practice
– A Longitudinal study of six very different innovation processes – in practice

26. Luise Noring Henler
Conceptualising successful supply chain partnerships
– Viewing supply chain partnerships from an organisational culture perspective

27. Mark Mau
Kampen om telefonen
Det danske telefonvæsen under den tyske besættelse 1940-45

28. Jakob Halskov
The semiautomatic expansion of existing terminological ontologies using knowledge patterns discovered

on the WWW – an implementation and evaluation

29. Gergana Koleva
European Policy Instruments Beyond Networks and Structure: The Innovative Medicines Initiative

30. Christian Geisler Asmussen
Global Strategy and International Diversity: A Double-Edged Sword?

31. Christina Holm-Petersen
Stolthed og fordom
Kultur- og identitetsarbejde ved skabelsen af en ny sengeafdeling gennem fusion

32. Hans Peter Olsen
Hybrid Governance of Standardized States
Causes and Contours of the Global Regulation of Government Auditing

33. Lars Bøge Sørensen
Risk Management in the Supply Chain

34. Peter Aagaard
Det unikkes dynamikker
De institutionelle mulighedsbetingelser bag den individuelle udforskning i professionelt og frivilligt arbejde

35. Yun Mi Antorini
Brand Community Innovation
An Intrinsic Case Study of the Adult Fans of LEGO Community

36. Joachim Lynggaard Boll
Labor Related Corporate Social Performance in Denmark
Organizational and Institutional Perspectives

2008

1. Frederik Christian Vinten
Essays on Private Equity

2. Jesper Clement
Visual Influence of Packaging Design on In-Store Buying Decisions

3. Marius Brostrøm Kousgaard
Tid til kvalitetsmåling?
– Studier af indrulleringsprocesser i forbindelse med introduktionen af kliniske kvalitetsdatabaser i speciallægepraksissektoren

4. Irene Skovgaard Smith
Management Consulting in Action
Value creation and ambiguity in client-consultant relations

5. Anders Rom
Management accounting and integrated information systems
How to exploit the potential for management accounting of information technology

6. Marina Candi
Aesthetic Design as an Element of Service Innovation in New Technology-based Firms

7. Morten Schnack
Teknologi og tværfaglighed
– en analyse af diskussionen omkring indførelse af EPJ på en hospitalsafdeling

8. Helene Balslev Clausen
Juntos pero no revueltos – un estudio sobre emigrantes norteamericanos en un pueblo mexicano

9. Lise Justesen
Kunsten at skrive revisionsrapporter. En beretning om forvaltningsrevisionens beretninger

10. Michael E. Hansen
The politics of corporate responsibility: CSR and the governance of child labor and core labor rights in the 1990s

11. Anne Roepstorff
Holdning for handling – en etnologisk undersøgelse af Virksomheders Sociale Ansvar/CSR

12. Claus Bajlum
Essays on Credit Risk and Credit Derivatives

13. Anders Bojesen
The Performative Power of Competence – an Inquiry into Subjectivity and Social Technologies at Work

14. Satu Reijonen
Green and Fragile
A Study on Markets and the Natural Environment

15. Ilduara Busta
Corporate Governance in Banking
A European Study

16. Kristian Anders Hvass
A Boolean Analysis Predicting Industry Change: Innovation, Imitation & Business Models
The Winning Hybrid: A case study of isomorphism in the airline industry

17. Trine Paludan
De uvidende og de udviklingsparate
Identitet som mulighed og restriktion blandt fabriksarbejdere på det aftayloriserede fabriksgulv

18. Kristian Jakobsen
Foreign market entry in transition economies: Entry timing and mode choice

19. Jakob Elming
Syntactic reordering in statistical machine translation

20. Lars Brømsøe Termansen
Regional Computable General Equilibrium Models for Denmark
Three papers laying the foundation for regional CGE models with agglomeration characteristics

21. Mia Reinholt
The Motivational Foundations of Knowledge Sharing

22. Frederikke Krogh-Meibom
The Co-Evolution of Institutions and Technology
– A Neo-Institutional Understanding of Change Processes within the Business Press – the Case Study of Financial Times

23. Peter D. Ørberg Jensen
OFFSHORING OF ADVANCED AND HIGH-VALUE TECHNICAL SERVICES: ANTECEDENTS, PROCESS DYNAMICS AND FIRMLEVEL IMPACTS

24. Pham Thi Song Hanh
Functional Upgrading, Relational Capability and Export Performance of Vietnamese Wood Furniture Producers

25. Mads Vangkilde
Why wait?
An Exploration of first-mover advantages among Danish e-grocers through a resource perspective

26. Hubert Buch-Hansen
Rethinking the History of European Level Merger Control
A Critical Political Economy Perspective

2009

1. Vivian Lindhardsen
From Independent Ratings to Communal Ratings: A Study of CWA Raters' Decision-Making Behaviours

2. Guðrið Weihe
Public-Private Partnerships: Meaning and Practice

3. Chris Nøkkentved
Enabling Supply Networks with Collaborative Information Infrastructures
An Empirical Investigation of Business Model Innovation in Supplier Relationship Management

4. Sara Louise Muhr
Wound, Interrupted – On the Vulnerability of Diversity Management

5. Christine Sestoft
Forbrugeradfærd i et Stats- og Livsformsteoretisk perspektiv

6. Michael Pedersen
Tune in, Breakdown, and Reboot: On the production of the stress-fit self-managing employee

7. Salla Lutz
Position and Reposition in Networks – Exemplified by the Transformation of the Danish Pine Furniture Manufacturers

8. Jens Forssbæck
Essays on market discipline in commercial and central banking

9. Tine Murphy
Sense from Silence – A Basis for Organised Action
How do Sensemaking Processes with Minimal Sharing Relate to the Reproduction of Organised Action?

10. Sara Malou Strandvad
Inspirations for a new sociology of art: A sociomaterial study of development processes in the Danish film industry

11. Nicolaas Mouton
On the evolution of social scientific metaphors:
A cognitive-historical enquiry into the divergent trajectories of the idea that collective entities – states and societies, cities and corporations – are biological organisms.

12. Lars Andreas Knutsen
Mobile Data Services:
Shaping of user engagements

13. Nikolaos Theodoros Korfiatis
Information Exchange and Behavior
A Multi-method Inquiry on Online Communities

14. Jens Albæk
Forestillinger om kvalitet og tværfaglighed på sygehuse
– skabelse af forestillinger i læge- og plejegrupperne angående relevans af nye idéer om kvalitetsudvikling gennem tolkningsprocesser

15. Maja Lotz
The Business of Co-Creation – and the Co-Creation of Business

16. Gitte P. Jakobsen
Narrative Construction of Leader Identity in a Leader Development Program Context

17. Dorte Hermansen
"Living the brand" som en brandorienteret dialogisk praxis:
Om udvikling af medarbejdernes brandorienterede dømmekraft

18. Aseem Kinra
Supply Chain (logistics) Environmental Complexity

19. Michael Nørager
How to manage SMEs through the transformation from non innovative to innovative?

20. Kristin Wallevik
Corporate Governance in Family Firms
The Norwegian Maritime Sector

21. Bo Hansen Hansen
Beyond the Process
Enriching Software Process Improvement with Knowledge Management

22. Annemette Skot-Hansen
Franske adjektivisk afledte adverbier, der tager præpositionssyntagmer indledt med præpositionen à som argumenter
En valensgrammatisk undersøgelse

23. Line Gry Knudsen
Collaborative R&D Capabilities
In Search of Micro-Foundations

24. Christian Scheuer
Employers meet employees
Essays on sorting and globalization

25. Rasmus Johnsen
The Great Health of Melancholy
A Study of the Pathologies of Performativity

26. Ha Thi Van Pham
Internationalization, Competitiveness Enhancement and Export Performance of Emerging Market Firms:
Evidence from Vietnam

27. Henriette Balieu
Kontrolbegrebets betydning for kausativalternationen i spansk
En kognitiv-typologisk analyse

2010

1. Yen Tran
Organizing Innovationin Turbulent Fashion Market
Four papers on how fashion firms create and appropriate innovation value

2. Anders Raastrup Kristensen
Metaphysical Labour
Flexibility, Performance and Commitment in Work-Life Management

3. Margrét Sigrún Sigurdardottir
Dependently independent
Co-existence of institutional logics in the recorded music industry

4. Ásta Dis Óladóttir
Internationalization from a small domestic base:
An empirical analysis of Economics and Management

5. Christine Secher
E-deltagelse i praksis – politikernes og forvaltningens medkonstruktion og konsekvenserne heraf

6. Marianne Stang Våland
What we talk about when we talk about space:
End User Participation between Processes of Organizational and Architectural Design

7. Rex Degnegaard
Strategic Change Management
Change Management Challenges in the Danish Police Reform

8. Ulrik Schultz Brix
Værdi i rekruttering – den sikre beslutning
En pragmatisk analyse af perception og synliggørelse af værdi i rekrutterings- og udvælgelsesarbejdet

9. Jan Ole Similä
Kontraktsledelse
Relasjonen mellom virksomhetsledelse og kontraktshåndtering, belyst via fire norske virksomheter

10. Susanne Boch Waldorff
Emerging Organizations: In between local translation, institutional logics and discourse

11. Brian Kane
Performance Talk
Next Generation Management of Organizational Performance

12. Lars Ohnemus
Brand Thrust: Strategic Branding and Shareholder Value
An Empirical Reconciliation of two Critical Concepts

13. Jesper Schlamovitz
Håndtering af usikkerhed i film- og byggeprojekter

14. Tommy Moesby-Jensen
Det faktiske livs forbindtlighed
Førsokratisk informeret, ny-aristotelisk ἦθος-tænkning hos Martin Heidegger

15. Christian Fich
Two Nations Divided by Common Values
French National Habitus and the Rejection of American Power

16. Peter Beyer
Processer, sammenhængskraft og fleksibilitet
Et empirisk casestudie af omstillingsforløb i fire virksomheder

17. Adam Buchhorn
Markets of Good Intentions
Constructing and Organizing Biogas Markets Amid Fragility and Controversy

18. Cecilie K. Moesby-Jensen
Social læring og fælles praksis
Et mixed method studie, der belyser læringskonsekvenser af et lederkursus for et praksisfællesskab af offentlige mellemledere

19. Heidi Boye
Fødevarer og sundhed i senmodernismen
– En indsigt i hyggefænomenet og de relaterede fødevarepraksisser

20. Kristine Munkgård Pedersen
Flygtige forbindelser og midlertidige mobiliseringer
Om kulturel produktion på Roskilde Festival

21. Oliver Jacob Weber
Causes of Intercompany Harmony in Business Markets – An Empirical Investigation from a Dyad Perspective

22. Susanne Ekman
Authority and Autonomy
Paradoxes of Modern Knowledge Work

23. Anette Frey Larsen
Kvalitetsledelse på danske hospitaler
– Ledelsernes indflydelse på introduktion og vedligeholdelse af kvalitetsstrategier i det danske sundhedsvæsen

24. Toyoko Sato
Performativity and Discourse: Japanese Advertisements on the Aesthetic Education of Desire

25. Kenneth Brinch Jensen
Identifying the Last Planner System
Lean management in the construction industry

26. Javier Busquets
Orchestrating Network Behavior for Innovation

27. Luke Patey
The Power of Resistance: India's National Oil Company and International Activism in Sudan

28. Mette Vedel
Value Creation in Triadic Business Relationships. Interaction, Interconnection and Position

29. Kristian Tørning
Knowledge Management Systems in Practice – A Work Place Study

30. Qingxin Shi
An Empirical Study of Thinking Aloud Usability Testing from a Cultural Perspective

31. Tanja Juul Christiansen
Corporate blogging: Medarbejderes kommunikative handlekraft

32. Malgorzata Ciesielska
Hybrid Organisations.
A study of the Open Source – business setting

33. Jens Dick-Nielsen
Three Essays on Corporate Bond Market Liquidity

34. Sabrina Speiermann
Modstandens Politik
Kampagnestyring i Velfærdsstaten.
En diskussion af trafikkampagners styringspotentiale

35. Julie Uldam
Fickle Commitment. Fostering political engagement in 'the flighty world of online activism'

36. Annegrete Juul Nielsen
Traveling technologies and transformations in health care

37. Athur Mühlen-Schulte
Organising Development
Power and Organisational Reform in the United Nations Development Programme

38. Louise Rygaard Jonas
Branding på butiksgulvet
Et case-studie af kultur- og identitetsarbejdet i Kvickly

2011

1. Stefan Fraenkel
Key Success Factors for Sales Force Readiness during New Product Launch
A Study of Product Launches in the Swedish Pharmaceutical Industry

2. Christian Plesner Rossing
International Transfer Pricing in Theory and Practice

3. Tobias Dam Hede
Samtalekunst og ledelsesdisciplin
– en analyse af coachingsdiskursens genealogi og governmentality

4. Kim Pettersson
Essays on Audit Quality, Auditor Choice, and Equity Valuation

5. Henrik Merkelsen
The expert-lay controversy in risk research and management. Effects of institutional distances. Studies of risk definitions, perceptions, management and communication

6. Simon S. Torp
Employee Stock Ownership:
Effect on Strategic Management and Performance

7. Mie Harder
Internal Antecedents of Management Innovation

8. Ole Helby Petersen
Public-Private Partnerships: Policy and Regulation – With Comparative and Multi-level Case Studies from Denmark and Ireland

9. Morten Krogh Petersen
'Good' Outcomes. Handling Multiplicity in Government Communication

10. Kristian Tangsgaard Hvelplund
Allocation of cognitive resources in translation - an eye-tracking and key-logging study

11. Moshe Yonatany
The Internationalization Process of Digital Service Providers

12. Anne Vestergaard
Distance and Suffering
Humanitarian Discourse in the age of Mediatization

13. Thorsten Mikkelsen
Personligsheds indflydelse på forretningsrelationer

14. Jane Thostrup Jagd
Hvorfor fortsætter fusionsbølgen udover "the tipping point"?
– en empirisk analyse af information og kognitioner om fusioner

15. Gregory Gimpel
Value-driven Adoption and Consumption of Technology: Understanding Technology Decision Making

16. Thomas Stengade Sønderskov
Den nye mulighed
Social innovation i en forretningsmæssig kontekst

17. Jeppe Christoffersen
Donor supported strategic alliances in developing countries

18. Vibeke Vad Baunsgaard
Dominant Ideological Modes of Rationality: Cross functional

integration in the process of product innovation

19. Throstur Olaf Sigurjonsson
Governance Failure and Icelands's Financial Collapse

20. Allan Sall Tang Andersen
Essays on the modeling of risks in interest-rate and inflation markets

21. Heidi Tscherning
Mobile Devices in Social Contexts

22. Birgitte Gorm Hansen
Adapting in the Knowledge Economy Lateral Strategies for Scientists and Those Who Study Them

23. Kristina Vaarst Andersen
Optimal Levels of Embeddedness The Contingent Value of Networked Collaboration

24. Justine Grønbæk Pors
Noisy Management A History of Danish School Governing from 1970-2010

25. Stefan Linder
Micro-foundations of Strategic Entrepreneurship Essays on Autonomous Strategic Action

26. Xin Li
Toward an Integrative Framework of National Competitiveness An application to China

27. Rune Thorbjørn Clausen
Værdifuld arkitektur Et eksplorativt studie af bygningers rolle i virksomheders værdiskabelse

28. Monica Viken
Markedsundersøkelser som bevis i varemerke- og markedsføringsrett

29. Christian Wymann
Tattooing The Economic and Artistic Constitution of a Social Phenomenon

30. Sanne Frandsen
Productive Incoherence A Case Study of Branding and Identity Struggles in a Low-Prestige Organization

31. Mads Stenbo Nielsen
Essays on Correlation Modelling

32. Ivan Häuser
Følelse og sprog Etablering af en ekspressiv kategori, eksemplificeret på russisk

33. Sebastian Schwenen
Security of Supply in Electricity Markets

2012

1. Peter Holm Andreasen
The Dynamics of Procurement Management - A Complexity Approach

2. Martin Haulrich
Data-Driven Bitext Dependency Parsing and Alignment

3. Line Kirkegaard
Konsulenten i den anden nat En undersøgelse af det intense arbejdsliv

4. Tonny Stenheim
Decision usefulness of goodwill under IFRS

5. Morten Lind Larsen
Produktivitet, vækst og velfærd Industrirådet og efterkrigstidens Danmark 1945 - 1958

6. Petter Berg
Cartel Damages and Cost Asymmetries

7. Lynn Kahle
Experiential Discourse in Marketing A methodical inquiry into practice and theory

8. Anne Roelsgaard Obling
Management of Emotions in Accelerated Medical Relationships

9. Thomas Frandsen
Managing Modularity of Service Processes Architecture

10. Carina Christine Skovmøller
CSR som noget særligt
Et casestudie om styring og meningsskabelse i relation til CSR ud fra en intern optik

11. Michael Tell
Fradragsbeskæring af selskabers finansieringsudgifter
En skatteretlig analyse af SEL §§ 11, 11B og 11C

12. Morten Holm
Customer Profitability Measurement Models
Their Merits and Sophistication across Contexts

13. Katja Joo Dyppel
Beskatning af derivater
En analyse af dansk skatteret

14. Esben Anton Schultz
Essays in Labor Economics
Evidence from Danish Micro Data

15. Carina Risvig Hansen
"Contracts not covered, or not fully covered, by the Public Sector Directive"

16. Anja Svejgaard Pors
Iværksættelse af kommunikation
- patientfigurer i hospitalets strategiske kommunikation

17. Frans Bévort
Making sense of management with logics
An ethnographic study of accountants who become managers

18. René Kallestrup
The Dynamics of Bank and Sovereign Credit Risk

19. Brett Crawford
Revisiting the Phenomenon of Interests in Organizational Institutionalism
The Case of U.S. Chambers of Commerce

20. Mario Daniele Amore
Essays on Empirical Corporate Finance

21. Arne Stjernholm Madsen
The evolution of innovation strategy
Studied in the context of medical device activities at the pharmaceutical company Novo Nordisk A/S in the period 1980-2008

22. Jacob Holm Hansen
Is Social Integration Necessary for Corporate Branding?
A study of corporate branding strategies at Novo Nordisk

23. Stuart Webber
Corporate Profit Shifting and the Multinational Enterprise

24. Helene Ratner
Promises of Reflexivity
Managing and Researching Inclusive Schools

25. Therese Strand
The Owners and the Power: Insights from Annual General Meetings

26. Robert Gavin Strand
In Praise of Corporate Social Responsibility Bureaucracy

27. Nina Sormunen
Auditor's going-concern reporting
Reporting decision and content of the report

28. John Bang Mathiasen
Learning within a product development working practice:
- an understanding anchored in pragmatism

29. Philip Holst Riis
Understanding Role-Oriented Enterprise Systems: From Vendors to Customers

30. Marie Lisa Dacanay
Social Enterprises and the Poor
Enhancing Social Entrepreneurship and Stakeholder Theory

31. Fumiko Kano Glückstad
Bridging Remote Cultures: Cross-lingual concept mapping based on the information receiver's prior-knowledge

32. Henrik Barslund Fosse
Empirical Essays in International Trade

33. Peter Alexander Albrecht
Foundational hybridity and its reproduction
Security sector reform in Sierra Leone

34. Maja Rosenstock
CSR - hvor svært kan det være?
Kulturanalytisk casestudie om udfordringer og dilemmaer med at forankre Coops CSR-strategi

35. Jeanette Rasmussen
Tweens, medier og forbrug
Et studie af 10-12 årige danske børns brug af internettet, opfattelse og forståelse af markedsføring og forbrug

36. Ib Tunby Gulbrandsen
'This page is not intended for a US Audience'
A five-act spectacle on online communication, collaboration & organization.

37. Kasper Aalling Teilmann
Interactive Approaches to Rural Development

38. Mette Mogensen
The Organization(s) of Well-being and Productivity
(Re)assembling work in the Danish Post

39. Søren Friis Møller
From Disinterestedness to Engagement
Towards Relational Leadership In the Cultural Sector

40. Nico Peter Berhausen
Management Control, Innovation and Strategic Objectives – Interactions and Convergence in Product Development Networks

41. Balder Onarheim
Creativity under Constraints
Creativity as Balancing 'Constrainedness'

42. Haoyong Zhou
Essays on Family Firms

43. Elisabeth Naima Mikkelsen
Making sense of organisational conflict
An empirical study of enacted sense-making in everyday conflict at work

2013

1. Jacob Lyngsie
Entrepreneurship in an Organizational Context

2. Signe Groth-Brodersen
Fra ledelse til selvet
En socialpsykologisk analyse af forholdet imellem selvledelse, ledelse og stress i det moderne arbejdsliv

3. Nis Høyrup Christensen
Shaping Markets: A Neoinstitutional Analysis of the Emerging Organizational Field of Renewable Energy in China

4. Christian Edelvold Berg
As a matter of size
THE IMPORTANCE OF CRITICAL MASS AND THE CONSEQUENCES OF SCARCITY FOR TELEVISION MARKETS

5. Christine D. Isakson
Coworker Influence and Labor Mobility
Essays on Turnover, Entrepreneurship and Location Choice in the Danish Maritime Industry

6. Niels Joseph Jerne Lennon
Accounting Qualities in Practice
Rhizomatic stories of representational faithfulness, decision making and control

7. Shannon O'Donnell
Making Ensemble Possible
How special groups organize for collaborative creativity in conditions of spatial variability and distance

8. Robert W. D. Veitch
Access Decisions in a Partly-Digital World
Comparing Digital Piracy and Legal Modes for Film and Music

9. Marie Mathiesen
Making Strategy Work
An Organizational Ethnography

10. Arisa Shollo
The role of business intelligence in organizational decision-making

11. Mia Kaspersen
The construction of social and environmental reporting

12. Marcus Møller Larsen
The organizational design of offshoring

13. Mette Ohm Rørdam
EU Law on Food Naming
The prohibition against misleading names in an internal market context

14. Hans Peter Rasmussen
GIV EN GED!
Kan giver-idealtyper forklare støtte til velgørenhed og understøtte relationsopbygning?

15. Ruben Schachtenhaufen
Fonetisk reduktion i dansk

16. Peter Koerver Schmidt
Dansk CFC-beskatning
I et internationalt og komparativt perspektiv

17. Morten Froholdt
Strategi i den offentlige sektor
En kortlægning af styringsmæssig kontekst, strategisk tilgang, samt anvendte redskaber og teknologier for udvalgte danske statslige styrelser

18. Annette Camilla Sjørup
Cognitive effort in metaphor translation
An eye-tracking and key-logging study

19. Tamara Stucchi
The Internationalization of Emerging Market Firms:
A Context-Specific Study

20. Thomas Lopdrup-Hjorth
"Let's Go Outside":
The Value of Co-Creation

21. Ana Alačovska
Genre and Autonomy in Cultural Production
The case of travel guidebook production

22. Marius Gudmand-Høyer
Stemningssindssygdommenes historie i det 19. århundrede
Omtydningen af melankolien og manien som bipolære stemningslidelser i dansk sammenhæng under hensyn til dannelsen af det moderne følelseslivs relative autonomi.
En problematiserings- og erfarings-analytisk undersøgelse

23. Lichen Alex Yu
Fabricating an S&OP Process
Circulating References and Matters of Concern

24. Esben Alfort
The Expression of a Need
Understanding search

25. Trine Pallesen
Assembling Markets for Wind Power
An Inquiry into the Making of Market Devices

26. Anders Koed Madsen
Web-Visions
Repurposing digital traces to organize social attention

27. Lærke Højgaard Christiansen
BREWING ORGANIZATIONAL RESPONSES TO INSTITUTIONAL LOGICS

28. Tommy Kjær Lassen
EGENTLIG SELVLEDELSE
En ledelsesfilosofisk afhandling om selvledelsens paradoksale dynamik og eksistentielle engagement

29. Morten Rossing
Local Adaption and Meaning Creation in Performance Appraisal

30. Søren Obed Madsen
Lederen som oversætter
Et oversættelsesteoretisk perspektiv på strategisk arbejde

31. Thomas Høgenhaven
Open Government Communities
Does Design Affect Participation?

32. Kirstine Zinck Pedersen
Failsafe Organizing?
A Pragmatic Stance on Patient Safety

33. Anne Petersen
Hverdagslogikker i psykiatrisk arbejde
En institutionsetnografisk undersøgelse af hverdagen i psykiatriske organisationer

34. Didde Maria Humle
Fortællinger om arbejde

35. Mark Holst-Mikkelsen
Strategieksekvering i praksis
– barrierer og muligheder!

36. Malek Maalouf
Sustaining lean
Strategies for dealing with organizational paradoxes

37. Nicolaj Tofte Brenneche
Systemic Innovation In The Making
The Social Productivity of Cartographic Crisis and Transitions in the Case of SEEIT

38. Morten Gylling
The Structure of Discourse
A Corpus-Based Cross-Linguistic Study

39. Binzhang YANG
Urban Green Spaces for Quality Life
- Case Study: the landscape architecture for people in Copenhagen

40. Michael Friis Pedersen
Finance and Organization:
The Implications for Whole Farm Risk Management

41. Even Fallan
Issues on supply and demand for environmental accounting information

42. Ather Nawaz
Website user experience
A cross-cultural study of the relation between users´ cognitive style, context of use, and information architecture of local websites

43. Karin Beukel
The Determinants for Creating Valuable Inventions

44. Arjan Markus
External Knowledge Sourcing and Firm Innovation
Essays on the Micro-Foundations of Firms' Search for Innovation

2014

1. Solon Moreira
Four Essays on Technology Licensing and Firm Innovation

2. Karin Strzeletz Ivertsen
Partnership Drift in Innovation Processes
A study of the Think City electric car development

3. Kathrine Hoffmann Pii
Responsibility Flows in Patient-centred Prevention

4. Jane Bjørn Vedel
Managing Strategic Research
An empirical analysis of science-industry collaboration in a pharmaceutical company

5. Martin Gylling
Processuel strategi i organisationer
Monografi om dobbeltheden i tænkning af strategi, dels som vidensfelt i organisationsteori, dels som kunstnerisk tilgang til at skabe i erhvervsmæssig innovation

6. Linne Marie Lauesen
Corporate Social Responsibility in the Water Sector: How Material Practices and their Symbolic and Physical Meanings Form a Colonising Logic

7. Maggie Qiuzhu Mei
LEARNING TO INNOVATE: The role of ambidexterity, standard, and decision process

8. Inger Høedt-Rasmussen
Developing Identity for Lawyers Towards Sustainable Lawyering

9. Sebastian Fux
Essays on Return Predictability and Term Structure Modelling

10. Thorbjørn N. M. Lund-Poulsen
Essays on Value Based Management

11. Oana Brindusa Albu
Transparency in Organizing: A Performative Approach

12. Lena Olaison
Entrepreneurship at the limits

13. Hanne Sørum
DRESSED FOR WEB SUCCESS? An Empirical Study of Website Quality in the Public Sector

14. Lasse Folke Henriksen
Knowing networks How experts shape transnational governance

15. Maria Halbinger
Entrepreneurial Individuals Empirical Investigations into Entrepreneurial Activities of Hackers and Makers

16. Robert Spliid
Kapitalfondenes metoder og kompetencer

17. Christiane Stelling
Public-private partnerships & the need, development and management of trusting A processual and embedded exploration

18. Marta Gasparin
Management of design as a translation process

19. Kåre Moberg
Assessing the Impact of Entrepreneurship Education From ABC to PhD

20. Alexander Cole
Distant neighbors Collective learning beyond the cluster

21. Martin Møller Boje Rasmussen
Is Competitiveness a Question of Being Alike? How the United Kingdom, Germany and Denmark Came to Compete through their Knowledge Regimes from 1993 to 2007

22. Anders Ravn Sørensen
Studies in central bank legitimacy, currency and national identity Four cases from Danish monetary history

23. Nina Bellak
Can Language be Managed in International Business? Insights into Language Choice from a Case Study of Danish and Austrian Multinational Corporations (MNCs)

24. Rikke Kristine Nielsen
Global Mindset as Managerial Meta-competence and Organizational Capability: Boundary-crossing Leadership Cooperation in the MNC The Case of 'Group Mindset' in Solar A/S.

25. Rasmus Koss Hartmann
User Innovation inside government Towards a critically performative foundation for inquiry

26. Kristian Gylling Olesen
Flertydig og emergerende ledelse i folkeskolen
Et aktør-netværksteoretisk ledelses-studie af politiske evalueringsreformers betydning for ledelse i den danske folkeskole

27. Troels Riis Larsen
Kampen om Danmarks omdømme 1945-2010
Omdømmearbejde og omdømmepolitik

28. Klaus Majgaard
Jagten på autenticitet i offentlig styring

29. Ming Hua Li
Institutional Transition and Organizational Diversity: Differentiated internationalization strategies of emerging market state-owned enterprises

30. Sofie Blinkenberg Federspiel
IT, organisation og digitalisering: Institutionelt arbejde i den kommunale digitaliseringsproces

31. Elvi Weinreich
Hvilke offentlige ledere er der brug for når velfærdstænkningen flytter sig – er Diplomuddannelsens lederprofil svaret?

32. Ellen Mølgaard Korsager
Self-conception and image of context in the growth of the firm – A Penrosian History of Fiberline Composites

33. Else Skjold
The Daily Selection

34. Marie Louise Conradsen
The Cancer Centre That Never Was The Organisation of Danish Cancer Research 1949-1992

35. Virgilio Failla
Three Essays on the Dynamics of Entrepreneurs in the Labor Market

36. Nicky Nedergaard
Brand-Based Innovation Relational Perspectives on Brand Logics and Design Innovation Strategies and Implementation

37. Mads Gjedsted Nielsen
Essays in Real Estate Finance

38. Kristin Martina Brandl
Process Perspectives on Service Offshoring

39. Mia Rosa Koss Hartmann
In the gray zone With police in making space for creativity

40. Karen Ingerslev
Healthcare Innovation under The Microscope Framing Boundaries of Wicked Problems

41. Tim Neerup Themsen
Risk Management in large Danish public capital investment programmes

2015

1. Jakob Ion Wille
Film som design
Design af levende billeder i film og tv-serier

2. Christiane Mossin
Interzones of Law and Metaphysics Hierarchies, Logics and Foundations of Social Order seen through the Prism of EU Social Rights

3. Thomas Tøth
TRUSTWORTHINESS: ENABLING GLOBAL COLLABORATION
An Ethnographic Study of Trust, Distance, Control, Culture and Boundary Spanning within Offshore Outsourcing of IT Services

4. Steven Højlund
Evaluation Use in Evaluation Systems – The Case of the European Commission

5. Julia Kirch Kirkegaard
AMBIGUOUS WINDS OF CHANGE – OR FIGHTING AGAINST WINDMILLS IN CHINESE WIND POWER
A CONSTRUCTIVIST INQUIRY INTO CHINA'S PRAGMATICS OF GREEN MARKETISATION MAPPING CONTROVERSIES OVER A POTENTIAL TURN TO QUALITY IN CHINESE WIND POWER

6. Michelle Carol Antero
A Multi-case Analysis of the Development of Enterprise Resource Planning Systems (ERP) Business Practices

Morten Friis-Olivarius
The Associative Nature of Creativity

7. Mathew Abraham
New Cooperativism:
A study of emerging producer organisations in India

8. Stine Hedegaard
Sustainability-Focused Identity: Identity work performed to manage, negotiate and resolve barriers and tensions that arise in the process of constructing or ganizational identity in a sustainability context

9. Cecilie Glerup
Organizing Science in Society – the conduct and justification of resposible research

10. Allan Salling Pedersen
Implementering af ITIL® IT-governance - når best practice konflikter med kulturen Løsning af implementerings-problemer gennem anvendelse af kendte CSF i et aktionsforskningsforløb.

11. Nihat Misir
A Real Options Approach to Determining Power Prices

12. Mamdouh Medhat
MEASURING AND PRICING THE RISK OF CORPORATE FAILURES

13. Rina Hansen
Toward a Digital Strategy for Omnichannel Retailing

14. Eva Pallesen
In the rhythm of welfare creation
A relational processual investigation moving beyond the conceptual horizon of welfare management

15. Gouya Harirchi
In Search of Opportunities: Three Essays on Global Linkages for Innovation

16. Lotte Holck
Embedded Diversity: A critical ethnographic study of the structural tensions of organizing diversity

17. Jose Daniel Balarezo
Learning through Scenario Planning

18. Louise Pram Nielsen
Knowledge dissemination based on terminological ontologies. Using eye tracking to further user interface design.

19. Sofie Dam
PUBLIC-PRIVATE PARTNERSHIPS FOR INNOVATION AND SUSTAINABILITY TRANSFORMATION
An embedded, comparative case study of municipal waste management in England and Denmark

20. Ulrik Hartmyer Christiansen
Follwoing the Content of Reported Risk Across the Organization

21. Guro Refsum Sanden
Language strategies in multinational corporations. A cross-sector study of financial service companies and manufacturing companies.

22. Linn Gevoll
Designing performance management for operational level
- A closer look on the role of design choices in framing coordination and motivation

23. Frederik Larsen
Objects and Social Actions – on Second-hand Valuation Practices

24. Thorhildur Hansdottir Jetzek
The Sustainable Value of Open Government Data
Uncovering the Generative Mechanisms of Open Data through a Mixed Methods Approach

25. Gustav Toppenberg
Innovation-based M&A – Technological-Integration Challenges – The Case of Digital-Technology Companies

26. Mie Plotnikof
Challenges of Collaborative Governance
An Organizational Discourse Study of Public Managers' Struggles with Collaboration across the Daycare Area

27. Christian Garmann Johnsen
Who Are the Post-Bureaucrats? A Philosophical Examination of the Creative Manager, the Authentic Leader and the Entrepreneur

28. Jacob Brogaard-Kay
Constituting Performance Management
A field study of a pharmaceutical company

29. Rasmus Ploug Jenle
Engineering Markets for Control: Integrating Wind Power into the Danish Electricity System

30. Morten Lindholst
Complex Business Negotiation: Understanding Preparation and Planning

31. Morten Grynings
TRUST AND TRANSPARENCY FROM AN ALIGNMENT PERSPECTIVE

32. Peter Andreas Norn
Byregimer og styringsevne: Politisk lederskab af store byudviklingsprojekter

33. Milan Miric
Essays on Competition, Innovation and Firm Strategy in Digital Markets

34. Sanne K. Hjordrup
The Value of Talent Management
Rethinking practice, problems and possibilities

35. Johanna Sax
Strategic Risk Management – Analyzing Antecedents and Contingencies for Value Creation

36. Pernille Rydén
Strategic Cognition of Social Media

37. Mimmi Sjöklint
The Measurable Me
- The Influence of Self-tracking on the User Experience

38. Juan Ignacio Staricco
Towards a Fair Global Economic Regime? A critical assessment of Fair Trade through the examination of the Argentinean wine industry

39. Marie Henriette Madsen
Emerging and temporary connections in Quality work

40. Yangfeng CAO
Toward a Process Framework of Business Model Innovation in the Global Context
Entrepreneurship-Enabled Dynamic Capability of Medium-Sized Multinational Enterprises

41. Carsten Scheibye
Enactment of the Organizational Cost Structure in Value Chain Configuration
A Contribution to Strategic Cost Management

2016

1. Signe Sofie Dyrby
Enterprise Social Media at Work

2. Dorte Boesby Dahl
The making of the public parking attendant
Dirt, aesthetics and inclusion in public service work

3. Verena Girschik
Realizing Corporate Responsibility Positioning and Framing in Nascent Institutional Change

4. Anders Ørding Olsen
IN SEARCH OF SOLUTIONS
Inertia, Knowledge Sources and Diversity in Collaborative Problem-solving

5. Pernille Steen Pedersen
Udkast til et nyt copingbegreb
En kvalifikation af ledelsesmuligheder for at forebygge sygefravær ved psykiske problemer.

6. Kerli Kant Hvass
Weaving a Path from Waste to Value: Exploring fashion industry business models and the circular economy

7. Kasper Lindskow
Exploring Digital News Publishing Business Models – a production network approach

8. Mikkel Mouritz Marfelt
The chameleon workforce: Assembling and negotiating the content of a workforce

9. Marianne Bertelsen
Aesthetic encounters
Rethinking autonomy, space & time in today's world of art

10. Louise Hauberg Wilhelmsen
EU PERSPECTIVES ON INTERNATIONAL COMMERCIAL ARBITRATION

11. Abid Hussain
On the Design, Development and Use of the Social Data Analytics Tool (SODATO): Design Propositions, Patterns, and Principles for Big Social Data Analytics

12. Mark Bruun
Essays on Earnings Predictability

13. Tor Bøe-Lillegraven
BUSINESS PARADOXES, BLACK BOXES, AND BIG DATA: BEYOND ORGANIZATIONAL AMBIDEXTERITY

14. Hadis Khonsary-Atighi
ECONOMIC DETERMINANTS OF DOMESTIC INVESTMENT IN AN OIL-BASED ECONOMY: THE CASE OF IRAN (1965-2010)

15. Maj Lervad Grasten
Rule of Law or Rule by Lawyers?
On the Politics of Translation in Global Governance

16. Lene Granzau Juel-Jacobsen
SUPERMARKEDETS MODUS OPERANDI – en hverdagssociologisk undersøgelse af forholdet mellem rum og handlen og understøtte relationsopbygning?

17. Christine Thalsgård Henriques
In search of entrepreneurial learning – Towards a relational perspective on incubating practices?

18. Patrick Bennett
Essays in Education, Crime, and Job Displacement

19. Søren Korsgaard
Payments and Central Bank Policy

20. Marie Kruse Skibsted
Empirical Essays in Economics of Education and Labor

21. Elizabeth Benedict Christensen
The Constantly Contingent Sense of Belonging of the 1.5 Generation Undocumented Youth
An Everyday Perspective

22. Lasse J. Jessen
Essays on Discounting Behavior and Gambling Behavior

23. Kalle Johannes Rose
Når stifterviljen dør…
Et retsøkonomisk bidrag til 200 års juridisk konflikt om ejendomsretten

24. Andreas Søeborg Kirkedal
Danish Stød and Automatic Speech Recognition

25. Ida Lunde Jørgensen
Institutions and Legitimations in Finance for the Arts

26. Olga Rykov Ibsen
An empirical cross-linguistic study of directives: A semiotic approach to the sentence forms chosen by British, Danish and Russian speakers in native and ELF contexts

27. Desi Volker
Understanding Interest Rate Volatility

28. Angeli Elizabeth Weller
Practice at the Boundaries of Business Ethics & Corporate Social Responsibility

29. Ida Danneskiold-Samsøe
Levende læring i kunstneriske organisationer
En undersøgelse af læringsprocesser mellem projekt og organisation på Aarhus Teater

30. Leif Christensen
Quality of information – The role of internal controls and materiality

31. Olga Zarzecka
Tie Content in Professional Networks

32. Henrik Mahncke
De store gaver
- Filantropiens gensidighedsrelationer i teori og praksis

33. Carsten Lund Pedersen
Using the Collective Wisdom of Frontline Employees in Strategic Issue Management

34. Yun Liu
Essays on Market Design

35. Denitsa Hazarbassanova Blagoeva
The Internationalisation of Service Firms

36. Manya Jaura Lind
Capability development in an off-shoring context: How, why and by whom

37. Luis R. Boscán F.
Essays on the Design of Contracts and Markets for Power System Flexibility

38. Andreas Philipp Distel
Capabilities for Strategic Adaptation: Micro-Foundations, Organizational Conditions, and Performance Implications

39. Lavinia Bleoca
The Usefulness of Innovation and Intellectual Capital in Business Performance: The Financial Effects of Knowledge Management vs. Disclosure

40. Henrik Jensen
Economic Organization and Imperfect Managerial Knowledge: A Study of the Role of Managerial Meta-Knowledge in the Management of Distributed Knowledge

41. Stine Mosekjær
The Understanding of English Emotion Words by Chinese and Japanese Speakers of English as a Lingua Franca An Empirical Study

42. Hallur Tor Sigurdarson
The Ministry of Desire - Anxiety and entrepreneurship in a bureaucracy

43. Kätlin Pulk
Making Time While Being in Time
A study of the temporality of organizational processes

44. Valeria Giacomin
Contextualizing the cluster Palm oil in Southeast Asia in global perspective (1880s–1970s)

45. Jeanette Willert
Managers' use of multiple Management Control Systems: The role and interplay of management control systems and company performance

46. Mads Vestergaard Jensen
Financial Frictions: Implications for Early Option Exercise and Realized Volatility

47. Mikael Reimer Jensen
Interbank Markets and Frictions

48. Benjamin Faigen
Essays on Employee Ownership

49. Adela Michea
Enacting Business Models
An Ethnographic Study of an Emerging Business Model Innovation within the Frame of a Manufacturing Company.

50. Iben Sandal Stjerne
Transcending organization in temporary systems
Aesthetics' organizing work and employment in Creative Industries

51. Simon Krogh
Anticipating Organizational Change

52. Sarah Netter
Exploring the Sharing Economy

53. Lene Tolstrup Christensen
State-owned enterprises as institutional market actors in the marketization of public service provision:
A comparative case study of Danish and Swedish passenger rail 1990–2015

54. Kyoung(Kay) Sun Park
Three Essays on Financial Economics

2017

1. Mari Bjerck
Apparel at work. Work uniforms and women in male-dominated manual occupations.

2. Christoph H. Flöthmann
Who Manages Our Supply Chains? Backgrounds, Competencies and Contributions of Human Resources in Supply Chain Management

3. Aleksandra Anna Rzeźnik
Essays in Empirical Asset Pricing

4. Claes Bäckman
Essays on Housing Markets

5. Kirsti Reitan Andersen
Stabilizing Sustainability in the Textile and Fashion Industry

6. Kira Hoffmann
Cost Behavior: An Empirical Analysis of Determinants and Consequences of Asymmetries

7. Tobin Hanspal
Essays in Household Finance

8. Nina Lange
Correlation in Energy Markets

9. Anjum Fayyaz
Donor Interventions and SME Networking in Industrial Clusters in Punjab Province, Pakistan

10. Magnus Paulsen Hansen
Trying the unemployed. Justification and critique, emancipation and coercion towards the 'active society'. A study of contemporary reforms in France and Denmark

11. Sameer Azizi
Corporate Social Responsibility in Afghanistan
– a critical case study of the mobile telecommunications industry

12. Malene Myhre
The internationalization of small and medium-sized enterprises: A qualitative study

13. Thomas Presskorn-Thygesen
The Significance of Normativity – Studies in Post-Kantian Philosophy and Social Theory

14. Federico Clementi
Essays on multinational production and international trade

15. Lara Anne Hale
Experimental Standards in Sustainability Transitions: Insights from the Building Sector

16. Richard Pucci
Accounting for Financial Instruments in an Uncertain World Controversies in IFRS in the Aftermath of the 2008 Financial Crisis

17. Sarah Maria Denta
Kommunale offentlige private partnerskaber Regulering I skyggen af Farumsagen

18. Christian Östlund
Design for e-training

19. Amalie Martinus Hauge
Organizing Valuations – a pragmatic inquiry

20. Tim Holst Celik
Tension-filled Governance? Exploring the Emergence, Consolidation and Reconfiguration of Legitimatory and Fiscal State-crafting

21. Christian Bason
Leading Public Design: How managers engage with design to transform public governance

22. Davide Tomio
Essays on Arbitrage and Market Liquidity

23. Simone Stæhr
Financial Analysts' Forecasts Behavioral Aspects and the Impact of Personal Characteristics

24. Mikkel Godt Gregersen
Management Control, Intrinsic Motivation and Creativity – How Can They Coexist

25. Kristjan Johannes Suse Jespersen
Advancing the Payments for Ecosystem Service Discourse Through Institutional Theory

26. Kristian Bondo Hansen
Crowds and Speculation: A study of crowd phenomena in the U.S. financial markets 1890 to 1940

27. Lars Balslev
Actors and practices – An institutional study on management accounting change in Air Greenland

28. Sven Klingler
Essays on Asset Pricing with Financial Frictions

29. Klement Ahrensbach Rasmussen
Business Model Innovation The Role of Organizational Design

30. Giulio Zichella
Entrepreneurial Cognition. Three essays on entrepreneurial behavior and cognition under risk and uncertainty

31. Richard Ledborg Hansen
En forkærlighed til det eksisterende – mellemlederens oplevelse af forandringsmodstand i organisatoriske forandringer

32. Vilhelm Stefan Holsting
Militært chefvirke: Kritik og retfærdiggørelse mellem politik og profession

33. Thomas Jensen
Shipping Information Pipeline: An information infrastructure to improve international containerized shipping

34. Dzmitry Bartalevich
Do economic theories inform policy? Analysis of the influence of the Chicago School on European Union competition policy

35. Kristian Roed Nielsen
Crowdfunding for Sustainability: A study on the potential of reward-based crowdfunding in supporting sustainable entrepreneurship

36. Emil Husted
There is always an alternative: A study of control and commitment in political organization

37. Anders Ludvig Sevelsted
Interpreting Bonds and Boundaries of Obligation. A genealogy of the emergence and development of Protestant voluntary social work in Denmark as shown through the cases of the Copenhagen Home Mission and the Blue Cross (1850 – 1950)

38. Niklas Kohl
Essays on Stock Issuance

39. Maya Christiane Flensborg Jensen
BOUNDARIES OF PROFESSIONALIZATION AT WORK An ethnography-inspired study of care workers' dilemmas at the margin

40. Andreas Kamstrup
Crowdsourcing and the Architectural Competition as Organisational Technologies

41. Louise Lyngfeldt Gorm Hansen
Triggering Earthquakes in Science, Politics and Chinese Hydropower - A Controversy Study

2018

1. Vishv Priya Kohli
Combatting Falsifi cation and Counterfeiting of Medicinal Products in the E uropean Union – A Legal Analysis

2. Helle Haurum
Customer Engagement Behavior in the context of Continuous Service Relationships

3. Nis Grünberg
The Party -state order: Essays on China's political organization and political economic institutions

4. Jesper Christensen
A Behavioral Theory of Human Capital Integration

5. Poula Marie Helth
Learning in practice

6. Rasmus Vendler Toft-Kehler
Entrepreneurship as a career? An investigation of the relationship between entrepreneurial experience and entrepreneurial outcome

7. Szymon Furtak
Sensing the Future: Designing sensor-based predictive information systems for forecasting spare part demand for diesel engines

8. Mette Brehm Johansen
Organizing patient involvement. An ethnographic study

9. Iwona Sulinska
Complexities of Social Capital in Boards of Directors

10. Cecilie Fanøe Petersen
Award of public contracts as a means to conferring State aid: A legal analysis of the interface between public procurement law and State aid law

11. Ahmad Ahmad Barirani
Three Experimental Studies on Entrepreneurship

12. Carsten Allerslev Olsen
Financial Reporting Enforcement: Impact and Consequences

13. Irene Christensen
New product fumbles – Organizing for the Ramp-up process

14. Jacob Taarup-Esbensen
Managing communities – Mining MNEs' community risk management practices

15. Lester Allan Lasrado
Set-Theoretic approach to maturity models

16. Mia B. Münster
Intention vs. Perception of Designed Atmospheres in Fashion Stores

17. Anne Sluhan
Non-Financial Dimensions of Family Firm Ownership: How Socioemotional Wealth and Familiness Influence Internationalization

18. Henrik Yde Andersen
Essays on Debt and Pensions

19. Fabian Heinrich Müller
Valuation Reversed – When Valuators are Valuated. An Analysis of the Perception of and Reaction to Reviewers in Fine-Dining

20. Martin Jarmatz
Organizing for Pricing

21. Niels Joachim Christfort Gormsen
Essays on Empirical Asset Pricing

22. Diego Zunino
Socio-Cognitive Perspectives in Business Venturing

23. Benjamin Asmussen
Networks and Faces between Copenhagen and Canton, 1730-1840

24. Dalia Bagdziunaite
Brains at Brand Touchpoints A Consumer Neuroscience Study of Information Processing of Brand Advertisements and the Store Environment in Compulsive Buying

25. Erol Kazan
Towards a Disruptive Digital Platform Model

26. Andreas Bang Nielsen
Essays on Foreign Exchange and Credit Risk

27. Anne Krebs
Accountable, Operable Knowledge Toward Value Representations of Individual Knowledge in Accounting

28. Matilde Fogh Kirkegaard
A firm- and demand-side perspective on behavioral strategy for value creation: Insights from the hearing aid industry

29. Agnieszka Nowinska
SHIPS AND RELATION-SHIPS Tie formation in the sector of shipping intermediaries in shipping

30. Stine Evald Bentsen
The Comprehension of English Texts by Native Speakers of English and Japanese, Chinese and Russian Speakers of English as a Lingua Franca. An Empirical Study.

31. Stine Louise Daetz
Essays on Financial Frictions in Lending Markets

32. *Christian Skov Jensen*
Essays on Asset Pricing

33. Anders Kryger
Aligning future employee action and corporate strategy in a resource-scarce environment

34. Maitane Elorriaga-Rubio
The behavioral foundations of strategic decision-making: A contextual perspective

35. Roddy Walker
Leadership Development as Organisational Rehabilitation: Shaping Middle-Managers as Double Agents

36. Jinsun Bae
Producing Garments for Global Markets Corporate social responsibility (CSR) in Myanmar's export garment industry 2011–2015

37. Queralt Prat-i-Pubill
Axiological knowledge in a knowledge driven world. Considerations for organizations.

38. Pia Mølgaard
Essays on Corporate Loans and Credit Risk

39. Marzia Aricò
Service Design as a Transformative Force: Introduction and Adoption in an Organizational Context

40. Christian Dyrlund Wåhlin-Jacobsen
Constructing change initiatives in workplace voice activities Studies from a social interaction perspective

41. Peter Kalum Schou
Institutional Logics in Entrepreneurial Ventures: How Competing Logics arise and shape organizational processes and outcomes during scale-up

42. Per Henriksen
Enterprise Risk Management Rationaler og paradokser i en moderne ledelsesteknologi

43. Maximilian Schellmann
The Politics of Organizing Refugee Camps

44. Jacob Halvas Bjerre
Excluding the Jews: The Aryanization of Danish-German Trade and German Anti-Jewish Policy in Denmark 1937-1943

45. Ida Schrøder
Hybridising accounting and caring: A symmetrical study of how costs and needs are connected in Danish child protection work

46. Katrine Kunst
Electronic Word of Behavior: Transforming digital traces of consumer behaviors into communicative content in product design

47. Viktor Avlonitis
Essays on the role of modularity in management: Towards a unified perspective of modular and integral design

48. Anne Sofie Fischer
Negotiating Spaces of Everyday Politics: -An ethnographic study of organizing for social transformation for women in urban poverty, Delhi, India

2019

1. Shihan Du
   ESSAYS IN EMPIRICAL STUDIES BASED ON ADMINISTRATIVE LABOUR MARKET DATA

2. Mart Laatsit
   Policy learning in innovation policy: A comparative analysis of European Union member states

3. Peter J. Wynne
   Proactively Building Capabilities for the Post-Acquisition Integration of Information Systems

4. Kalina S. Staykova
   Generative Mechanisms for Digital Platform Ecosystem Evolution

5. Ieva Linkeviciute
   Essays on the Demand-Side Management in Electricity Markets

6. Jonatan Echebarria Fernández
   Jurisdiction and Arbitration Agreements in Contracts for the Carriage of Goods by Sea – Limitations on Party Autonomy

7. Louise Thorn Bøttkjær
   Votes for sale. Essays on clientelism in new democracies.

8. Ditte Vilstrup Holm
   The Poetics of Participation: the organizing of participation in contemporary art

9. Philip Rosenbaum
   Essays in Labor Markets – Gender, Fertility and Education

10. Mia Olsen
    Mobile Betalinger - Succesfaktorer og Adfærdsmæssige Konsekvenser

11. Adrián Luis Mérida Gutiérrez
    Entrepreneurial Careers: Determinants, Trajectories, and Outcomes

12. Frederik Regli
    Essays on Crude Oil Tanker Markets

13. Cancan Wang
    Becoming Adaptive through Social Media: Transforming Governance and Organizational Form in Collaborative E-government

14. Lena Lindbjerg Sperling
    Economic and Cultural Development: Empirical Studies of Micro-level Data

15. Xia Zhang
    Obligation, face and facework: An empirical study of the communicative act of cancellation of an obligation by Chinese, Danish and British business professionals in both L1 and ELF contexts

16. Stefan Kirkegaard Sløk-Madsen
    Entrepreneurial Judgment and Commercialization

17. Erin Leitheiser
    The Comparative Dynamics of Private Governance
    The case of the Bangladesh Ready-Made Garment Industry

18. Lone Christensen
    STRATEGIIMPLEMENTERING: STYRINGSBESTRÆBELSER, IDENTITET OG AFFEKT

19. Thomas Kjær Poulsen
    Essays on Asset Pricing with Financial Frictions

20. Maria Lundberg
    Trust and self-trust in leadership identity constructions: A qualitative exploration of narrative ecology in the discursive aftermath of heroic discourse

21. Tina Joanes
Sufficiency for sustainability Determinants and strategies for reducing clothing consumption

22. Benjamin Johannes Flesch
Social Set Visualizer (SoSeVi): Design, Development and Evaluation of a Visual Analytics Tool for Computational Set Analysis of Big Social Data

23. Henriette Sophia Groskopff Tvede Schleimann
Creating innovation through collaboration – Partnering in the maritime sector

24. Kristian Steensen Nielsen
The Role of Self-Regulation in Environmental Behavior Change

25. Lydia L. Jørgensen
Moving Organizational Atmospheres

26. Theodor Lucian Vladasel
Embracing Heterogeneity: Essays in Entrepreneurship and Human Capital

27. Seidi Suurmets
Contextual Effects in Consumer Research: An Investigation of Consumer Information Processing and Behavior via the Applicati on of Eye-tracking Methodology

28. Marie Sundby Palle Nickelsen
Reformer mellem integritet og innovation: Reform af reformens form i den danske centraladministration fra 1920 til 2019

29. Vibeke Kristine Scheller
The temporal organizing of same-day discharge: A tempography of a Cardiac Day Unit

30. Qian Sun
Adopting Artificial Intelligence in Healthcare in the Digital Age: Perceived Challenges, Frame Incongruence, and Social Power

31. Dorthe Thorning Mejlhede
Artful change agency and organizing for innovation – the case of a Nordic fintech cooperative

32. Benjamin Christoffersen
Corporate Default Models: Empirical Evidence and Methodical Contributions

33. Filipe Antonio Bonito Vieira
Essays on Pensions and Fiscal Sustainability

34. Morten Nicklas Bigler Jensen
Earnings Management in Private Firms: An Empirical Analysis of Determinants and Consequences of Earnings Management in Private Firms

2020

1. Christian Hendriksen
Inside the Blue Box: Explaining industry influence in the International Maritime Organization

2. Vasileios Kosmas
Environmental and social issues in global supply chains: Emission reduction in the maritime transport industry and maritime search and rescue operational response to migration

3. Thorben Peter Simonsen
The spatial organization of psychiatric practice: A situated inquiry into 'healing architecture'

4. Signe Bruskin
The infinite storm: An ethnographic study of organizational change in a bank

5. Rasmus Corlin Christensen
Politics and Professionals: Transnational Struggles to Change International Taxation

6. Robert Lorenz Törmer
The Architectural Enablement of a Digital Platform Strategy

7. Anna Kirkebæk Johansson Gosovic
Ethics as Practice: An ethnographic study of business ethics in a multinational biopharmaceutical company

8. Frank Meier
Making up leaders in leadership development

9. Kai Basner
Servitization at work: On proliferation and containment

10. Anestis Keremis
Anti-corruption in action: How is anti-corruption practiced in multinational companies?

11. Marie Larsen Ryberg
Governing Interdisciolinarity: Stakes and translations of interdisciplinarity in Danish high school education.

12. Jannick Friis Christensen
Queering organisation(s): Norm-critical orientations to organising and researching diversity

13. Thorsteinn Sigurdur Sveinsson
Essays on Macroeconomic Implications of Demographic Change

14. Catherine Casler
Reconstruction in strategy and organization: For a pragmatic stance

15. Luisa Murphy
Revisiting the standard organization of multi-stakeholder initiatives (MSIs): The case of a meta-MSI in Southeast Asia

16. Friedrich Bergmann
Essays on International Trade

17. Nicholas Haagensen
European Legal Networks in Crisis: The Legal Construction of Economic Policy

18. Charlotte Biil
Samskabelse med en sommerfugle-model: Hybrid ret i forbindelse med et partnerskabsprojekt mellem 100 selvejende daginstitutioner, deres paraplyorganisation, tre kommuner og CBS

19. Andreas Dimmelmeier
The Role of Economic Ideas in Sustainable Finance: From Paradigms to Policy

20. Maibrith Kempka Jensen
Ledelse og autoritet i interaktion - En interaktionsbaseret undersøgelse af autoritet i ledelse i praksis

21. Thomas Burø
LAND OF LIGHT: Assembling the Ecology of Culture in Odsherred 2000-2018

22. Prins Marcus Valiant Lantz
Timely Emotion: The Rhetorical Framing of Strategic Decision Making

23. Thorbjørn Vittenhof Fejerskov
Fra værdi til invitationer - offentlig værdiskabelse gennem affekt, potentialitet og begivenhed

24. Lea Acre Foverskov
Demographic Change and Employment: Path dependencies and institutional logics in the European Commission

25. Anirudh Agrawal
A Doctoral Dissertation

26. Julie Marx
Households in the housing market

27. Hadar Gafni
Alternative Digital Methods of Providing Entrepreneurial Finance

28. Mathilde Hjerrild Carlsen
Ledelse af engagementer: En undersøgelse af samarbejde mellem folkeskoler og virksomheder i Danmark

29. Suen Wang
Essays on the Gendered Origins and Implications of Social Policies in the Developing World

30. Stine Hald Larsen
The Story of the Relative: A Systems-Theoretical Analysis of the Role of the Relative in Danish Eldercare Policy from 1930 to 2020

31. Christian Casper Hofma
Immersive technologies and organizational routines: When head-mounted displays meet organizational routines

32. Jonathan Feddersen
The temporal emergence of social relations: An event-based perspective of organising

33. Nageswaran Vaidyanathan
ENRICHING RETAIL CUSTOMER EXPERIENCE USING AUGMENTED REALITY

2021

1. Vanya Rusinova
The Determinants of Firms' Engagement in Corporate Social Responsibility: Evidence from Natural Experiments

2. Lívia Lopes Barakat
Knowledge management mechanisms at MNCs:
The enhancing effect of absorptive capacity and its effects on performance and innovation

3. Søren Bundgaard Brøgger
Essays on Modern Derivatives Markets

4. Martin Friis Nielsen
Consuming Memory: Towards a conceptualization of social media platforms as organizational technologies of consumption

05. Fei Liu
Emergent Technology Use in Consumer Decision Journeys: A Process-as-Propensity Approach

06. Jakob Rømer Barfod
Ledelse i militære højrisikoteams

07. Elham Shafiei Gol
Creative Crowdwork Arrangements

08. Árni Jóhan Petersen
Collective Imaginary as (Residual) Fantasy: A Case Study of the Faroese Oil Bonanza

09. Søren Bering
"Manufacturing, Forward Integration and Governance Strategy"

10. Lars Oehler
Technological Change and the Decomposition of Innovation:
Choices and Consequences for Latecomer Firm Upgrading: The Case of China's Wind Energy Sector

11. Lise Dahl Arvedsen
Leadership in interaction in a virtual context:
A study of the role of leadership processes in a complex context, and how such processes are accomplished in practice

12. Jacob Emil Jeppesen
Essays on Knowledge networks, scientific impact and new knowledge adoption

13. Kasper Ingeman Beck
Essays on Chinese State-Owned Enterprises:
Reform, Corporate Governance and Subnational Diversity

14. Sönnich Dahl Sönnichsen
Exploring the interface between public demand and private supply for implementation of circular economy principles

15. Benjamin Knox
Essays on Financial Markets and Monetary Policy

16. Anita Eskesen
Essays on Utility Regulation: Evaluating Negotiation-Based Approaches inthe Context of Danish Utility Regulation

17. Agnes Guenther
Essays on Firm Strategy and Human Capital

18. Sophie Marie Cappelen
Walking on Eggshells: The balancing act of temporal work in a setting of culinary change

19. Manar Saleh Alnamlah
About Gender Gaps in Entrepreneurial Finance

20. Kirsten Tangaa Nielsen
Essays on the Value of CEOs and Directors

21. Renée Ridgway
Re:search - the Personalised Subject vs. the Anonymous User

22. Codrina Ana Maria Lauth
IMPACT Industrial Hackathons: Findings from a longitudinal case study on short-term vs long-term IMPACT implementations from industrial hackathons within Grundfos

23. Wolf-Hendrik Uhlbach
Scientist Mobility: Essays on knowledge production and innovation

24. Tomaz Sedej
Blockchain technology and inter-organizational relationships

25. Lasse Bundgaard
Public Private Innovation Partnerships: Creating Public Value & Scaling Up Sustainable City Solutions

26. Dimitra Makri Andersen
Walking through Temporal Walls: Rethinking NGO Organizing for Sustainability through a Temporal Lens on NGO-Business Partnerships

27. Louise Fjord Kjærsgaard
Allocation of the Right to Tax Income from Digital Products and Services: A legal analysis of international tax treaty law

28. Sara Dahlman
Marginal alternativity: Organizing for sustainable investing

29. Henrik Gundelach
Performance determinants: An Investigation of the Relationship between Resources, Experience and Performance in Challenging Business Environments

30. Tom Wraight
Confronting the Developmental State: American Trade Policy in the Neoliberal Era

31. Mathias Fjællegaard Jensen
Essays on Gender and Skills in the Labour Market

32. Daniel Lundgaard
Using Social Media to Discuss Global Challenges: Case Studies of the Climate Change Debate on Twitter

33. Jonas Sveistrup Søgaard
Designs for Accounting Information Systems using Distributed Ledger Technology

34. Sarosh Asad
CEO narcissism and board composition: Implications for firm strategy and performance

35. Johann Ole Willers
Experts and Markets in Cybersecurity On Definitional Power and the Organization of Cyber Risks

36. Alexander Kronies
Opportunities and Risks in Alternative Investments

37. Niels Fuglsang
The Politics of Economic Models: An inquiry into the possibilities and limits concerning the rise of macroeconomic forecasting models and what this means for policymaking

38. David Howoldt
Policy Instruments and Policy Mixes for Innovation: Analysing Their Relation to Grand Challenges, Entrepreneurship and Innovation Capability with Natural Language Processing and Latent Variable Methods

2022

01. Ditte Thøgersen
Managing Public Innovation on the Frontline

02. Rasmus Jørgensen
Essays on Empirical Asset Pricing and Private Equity

03. Nicola Giommetti
Essays on Private Equity

04. Laila Starr
When Is Health Innovation Worth It? Essays On New Approaches To Value Creation In Health

05. Maria Krysfeldt Rasmussen
Den transformative ledelsesbyrde – etnografisk studie af en religionsinspireret ledelsesfilosofi i en dansk modevirksomhed

06. Rikke Sejer Nielsen
Mortgage Decisions of Households: Consequences for Consumption and Savings

07. Myriam Noémy Marending
Essays on development challenges of low income countries: Evidence from conflict, pest and credit

08. Selorm Agbleze
A BEHAVIORAL THEORY OF FIRM FORMALIZATION

09. Rasmus Arler Bogetoft
Rettighedshavers faktisk lidte tab i immaterialretssager: Studier af dansk ret med støtte i økonomisk teori og metode

10. Franz Maximilian Buchmann
Driving the Green Transition of the Maritime Industry through Clean Technology Adoption and Environmental Policies

11. Ivan Olav Vulchanov
The role of English as an organisational language in international workplaces

12. Anne Agerbak Bilde
TRANSFORMATIONER AF SKOLELEDELSE - en systemteoretisk analyse af hvordan betingelser for skoleledelse forandres med læring som genstand i perioden 1958-2020

13. JUAN JOSE PRICE ELTON
EFFICIENCY AND PRODUCTIVITY ANALYSIS: TWO EMPIRICAL APPLICATIONS AND A METHODOLOGICAL CONTRIBUTION

14. Catarina Pessanha Gomes
The Art of Occupying: Romanticism as Political Culture in French Prefigurative politics

15. Mark Ørberg
Fondsretten og den levende vedtægt

16. Majbritt Greve
Maersk's Role in Economic Development: A Study of Shipping and Logistics Foreign Direct Investment in Global Trade

17. Sille Julie J. Abildgaard
Doing-Being Creative: Empirical Studies of Interaction in Design Work

18. Jette Sandager
Glitter, Glamour, and the Future of (More) Girls in STEM: Gendered Formations of STEM Aspirations

19. Casper Hein Winther
Inside the innovation lab - How paradoxical tensions persist in ambidextrous organizations over time

20. Nikola Kostić
Collaborative governance of inter-organizational relationships: The effects of management controls, blockchain technology, and industry standards

21. Saila Naomi Stausholm
Maximum capital, minimum tax: Enablers and facilitators of corporate tax minimization

22. Robin Porsfelt
Seeing through Signs: On Economic Imagination and Semiotic Speculation

23. Michael Herburger
Supply chain resilience – a concept for coping with cyber risks

24. Katharina Christiane Nielsen Jeschke
Balancing safety in everyday work - A case study of construction managers' dynamic safety practices

25. Jakob Ahm Sørensen
Financial Markets with Frictions and Belief Distortions

26. Jakob Laage-Thomsen
Nudging Leviathan, Protecting Demos - A Comparative Sociology of Public Administration and Expertise in the Nordics

27. Kathrine Søs Jacobsen Cesko
Collaboration between Economic Operators in the Competition for Public Contracts: A Legal and Economic Analysis of Grey Zones between EU Public Procurement Law and EU Competition Law

28. Mette Nelund
Den nye jord – Et feltstudie af et bæredygtigt virke på Farendløse Mosteri

29. Benjamin Cedric Larsen
Governing Artificial Intelligence – Lessons from the United States and China

30. Anders Brøndum Klein
Kollektiv meningsdannelse iblandt heterogene aktører i eksperimentelle samskabelsesprocesser

31. Stefano Tripodi
Essays on Development Economicis

32. Katrine Maria Lumbye
Internationalization of European Electricity Multinationals in Times of Transition

33. Xiaochun Guo
Dynamic Roles of Digital Currency – An Exploration from Interactive Processes: Difference, Time, and Perspective

34. Louise Lindbjerg
Three Essays on Firm Innovation

35. Marcela Galvis Restrepo
Feature reduction for classification with mixed data: an algorithmic approach

36. Hanna Nyborg Storm
Cultural institutions and attractiveness – How cultural institutions contribute to the development of regions and local communities

37. Anna-Bertha Heeris Christensen
Conflicts and Challenges in Practices of Commercializing Humans – An Ethnographic Study of Influencer Marketing Work

38. Casper Berg Lavmand Larsen
A Worker-Centered Inquiry into the Contingencies and Consequences of Worker Representation

39. Niels le Duc
The Resource Commitment of Multinational Enterprise R&D Activities

40. Esben Langager Olsen
Change management tools and change managers – Examining the simulacra of change

41. Anne Sophie Lassen
Gender in the Labor Market

42. Alison E. Holm
Corrective corporate responses to accusations of misconduct on societal issues

43. Chenyan Lyu
Carbon Pricing, Renewable Energy, and Clean Growth – A Market Perspective

44. Alina Grecu
UNPACKING MULTI-LEVEL OFFSHORING CONSEQUENCES: Hiring Wages, Onshore Performance, and Public Sentiment

45. Alexandra Lüth
Offshore Energy Hubs as an Emerging Concept – Sector Integration at Sea

2023

01. Cheryl Basil Sequeira
Port Business Development – Digitalisation of Port Authroity and Hybrid Governance Model

02. Mette Suder Franck
Empirical Essays on Technology Supported Learning – Studies of Danish Higher Education

03. Søren Lund Frandsen
States and Experts – Assembling Expertise for Climate Change and Pandemics

04. Guowei Dong
Innovation and Internationalization – Evidence from Chinese Manufacturing Enterprises

05. Eileen Murphy
In Service to Security – Constructing the Authority to Manage European Border Data Infrastructures

06. Bontu Lucie Guschke
THE PERSISTENCE OF SEXISM AND RACISM AT UNIVERSITIES – Exploring the imperceptibility and unspeakability of workplace harassment and discrimination in academia

07. Christoph Viebig
Learning Entrepreneurship – How capabilities shape learning from experience, reflection, and action

08. Kasper Regenburg
Financial Risks of Private Firms

09. Kathrine Møller Solgaard
Who to hire? – A situated study of employee selection as routine, practice, and process

10. Jack Kværnø-Jones
Intersections between FinTech Imaginaries and Traditional Banking – A study of disciplinary, implementary, and parasitic work in the Danish financial sector

11. Stine Quorning
Managing Climate Change Like a Central Banker – The Political Economy of Greening the Monetary Technocracy

12. Amanda Bille
No business without politics – Investigating the political nature of supply chain management

13. Theis Ingerslev Jensen
Essays on Empirical Asset Pricing

14. Ann Fugl-Meyer
The Agile Imperative – A Qualitative Study of a Translation Process in the Danish Tax Administration

15. Nicolai Søgaard Laursen
Longevity risk in reinsurance and equity markets

16. Shelter Selorm Kwesi Teyi
STRATEGIC ENTREPRENEURSHIP IN THE INFORMAL ECONOMY

17. Luisa Hedler
Time, Law and Tech – The introduction of algorithms to courts of law

18. Tróndur Møller Sandoy
Essays on the Economics of Education

19. Nathan Rietzler
Crowdsourcing Processes and Performance Outcomes

20. Sigrid Alexandra Koob
Essays on Democracy, Redistribution, and Inequality

21. David Pinkus
Pension Fund Investment: Implications for the Real Economy

22. Sina Smid
Inequality and Redistribution – Essays on Local Elections, Gender and Corruption in Developing Countries

23. Andreas Brøgger
Financial Economics with Preferences and Frictions

24. Timothy Charlton-Czaplicki
Arendt in the platformised world – Labour, work and action on digital platforms

25. Letícia Vedolin Sebastião
Mindfulness and Consumption: Routes Toward Consumer Self-Control

26. Lotte List
Crisis Sovereignty – The Philosophy of History of the Exception

27. Jeanette Walldorf
Essays on the Economics of Education and Labour Market

28. Juan Camilo Giraldo-Mora
It is Along Ways – Global Payment Infrastructure in Movement

29. Niels Buus Lassen
THE PREDICTIVE POWER OF SOCIAL MEDIA DATA

30. Frederik Bjørn Christensen
Essays on the Intergenerational Welfare State

31. Shama Patel
The Summer of 2020: Situating Digital Media in Scaling Affective Contagion: A Case of the George Floyd Video

32. Federico Jensen
Who rules the waves in the 21st Century?
The international political economy of global shipping

33. Tobias Berggren Jensen
Selvledende organisationer i den offentlige sektor – modsætninger og konflikter i radikal decentralisering

34. Jonathan Harmat
The Affects By Which We Are Torn
Four Essays on Government and Affect

35. Jørgen Valther Hansen
The Big 4 Audit Firms and the Public Interest
Public oversight & Audit Firm Governance

36. Stig Strandbæk Nyman
The Birth of Algorithmic Aspirational Control

37. Morten Tinning
Steaming Ahead
Experiences and the Transition from Sail to Steam

38. Oguzhan Cepni
Essays in Applied Financial Economics

39. Tim Dominik Maurer
Essays on Pension Policy

40. Aixa Y. Alemán-Díaz
Exploring Global Ideas in National Policy for Science, Technology and Innovation an Isomorphic Difference Approach

41. Michael Güldenpfennig
Managing the interrelationships between manufacturing system elements for productivity improvement in the factory

42. Jun Yuan (Julian) Seng
Essays on the political economy of innovative startups

43. Jacek Piosik
Essays on Entrepreneurial Finance

44. Elizabeth Cooper
Tourists on the Edge
Understanding and Encouraging Sustainable Tourist Behaviour in Greenland

2024

01. Marija Sarafinovska
Patients as Innovators: An Empirical Study of Patients' Role in Innovation in the Healthcare Industry

02. Niina Hakala
Corporate Reporting in the Governance of Climate Transition – Framing agency in a financialized world

03. Kasper Merling Arendt
Unleashing Entrepreneurial Education
Developing Entrepreneurial Mindsets, Competencies, and Long-Term Behavior

04. Kerstin Martel
Creating and dissolving 'identity' in global mobility studies -
a multi-scalar inquiry of belongingness and becoming on-the-move

05. Sofie Elbæk Henriksen
Big Tech to the Rescue?
An Ethnographic Study of Corporate Humanitarianism in the Refugee Crisis

06. Christina Kjær
Corporate scandals
- in the age of 'responsible business'

07. Anna Stöber
Embedded Self-Managing Modes of Organizing
Empirical Inquiries into Boundaries, Momentum, and Collectivity

08. Lucas Sören Göbeler
Shifting and Shaping
Physicality in Digital Innovation

09. Felix Schilling
Department of International Economics, Government and Business

10. Mathias Lund Larsen
China and the Political Economy of the Green State

11. Michael Bennedsen Hansen
At få sjælen med
En narrativ analyse af danske container-søfolks erindringer, fortidsbrug og identitets-konstruktioner

12. Justyna Agata Bekier
More than a numbers game
Accounting for circular economy performance in collaborative initiatives in cities

13. Frederik Schade
The Question of Digital Responsibility
An Ethnography of Emergent Institutional Formations in the Contemporary Governance of Technology

14. Alexandrina Schmidt
The Mundane in the Digital: A qualitative study of social work and vulnerable clients in Danish job centres

15. Julian Fernandez Mejia
Essays on International Finance

16. Leonie Decrinis
Nudging in the Workplace: Exploring a Micro-level Approach Towards Corporate Sustainability

17. Nina Frausing Pedersen
A Framing Contest between Institutional Actors on Crypto-Asset Policymaking in the EU

18. Amalie Toft Bentsen
The Internal Market & the EU Climate Regime
Interactions and frictions in the legal norm systems

19. Sippo Rossi
Bots on Social Media
The Past, Present and Future

20. Sumair Hussain
Essays on Disclosures

21. Kseniia Kurishchenko
Novel Mathematical Optimization Models for Explainable and Fair Machine Learning

TITLER I ATV PH.D.-SERIEN

1992

1. Niels Kornum
Servicesamkørsel – organisation, økonomi og planlægningsmetode

1995

2. Verner Worm
Nordiske virksomheder i Kina
Kulturspecifikke interaktionsrelationer ved nordiske virksomhedsetableringer i Kina

1999

3. Mogens Bjerre
Key Account Management of Complex Strategic Relationships
An Empirical Study of the Fast Moving Consumer Goods Industry

2000

4. Lotte Darsø
Innovation in the Making
Interaction Research with heterogeneous Groups of Knowledge Workers creating new Knowledge and new Leads

2001

5. Peter Hobolt Jensen
Managing Strategic Design Identities
The case of the Lego Developer Network

2002

6. Peter Lohmann
The Deleuzian Other of Organizational Change – Moving Perspectives of the Human

7. Anne Marie Jess Hansen
To lead from a distance: The dynamic interplay between strategy and strategizing – A case study of the strategic management process

2003

8. Lotte Henriksen
Videndeling
– om organisatoriske og ledelsesmæssige udfordringer ved videndeling i praksis

9. Niels Christian Nickelsen
Arrangements of Knowing: Coordinating Procedures Tools and Bodies in Industrial Production – a case study of the collective making of new products

2005

10. Carsten Ørts Hansen
Konstruktion af ledelsesteknologier og effektivitet

TITLER I DBA PH.D.-SERIEN

2007

1. Peter Kastrup-Misir
Endeavoring to Understand Market Orientation – and the concomitant co-mutation of the researched, the re searcher, the research itself and the truth

2009

1. Torkild Leo Thellefsen
Fundamental Signs and Significance effects
A Semeiotic outline of Fundamental Signs, Significance-effects, Knowledge Profiling and their use in Knowledge Organization and Branding

2. Daniel Ronzani
When Bits Learn to Walk Don't Make Them Trip. Technological Innovation and the Role of Regulation by Law in Information Systems Research: the Case of Radio Frequency Identification (RFID)

2010

1. Alexander Carnera
Magten over livet og livet som magt
Studier i den biopolitiske ambivalens