MSc Artificial Intelligence
Master Thesis

Profile-based Subgroup Discovery for Fairness Analysis

by
Dionne Gantzert
12058866

July 14, 2024

48EC
November 2023 - June 2024

Supervisor:
Isabel Barberá

Examiner:
Dr Giovanni Sileno

Second reader:
Dr Fernando Santos

Universiteit van Amsterdam

Contents

1 Introduction **1**

2 Related Work **4**
2.1 Subgroup Discovery 4
2.2 Clustering for Subgroup Discovery 5
2.3 Subgroup Quality 6
2.4 Subgroup Fairness 7

3 Theoretic Background **8**
3.1 Subgroup Discovery Tasks 8
3.1.1 DFS 8
3.1.2 VLSD 8
3.2 Variability Controlled Hierarchical K-medoids (VHK) 9
3.2.1 Variability 10
3.2.2 VHK 11
3.3 Subgroup Quality Evaluation 11
3.3.1 Quality of subgroups 12
3.4 Subgroup Fairness Evaluation 13
3.4.1 Individual Fairness Notions for Subgroup Fairness 13
3.4.2 Group Fairness Notions for Subgroup Fairness 14

4 Methodology **15**
4.1 Subgroup Discovery Techniques 15
4.2 Profile-based Subgroup Discovery 15
4.2.1 Clustering Pipeline 16
4.2.2 Profile Descriptions based on Clusters 16
4.3 Evaluation 17
4.3.1 Description Quality 18
4.3.2 Subgroup Quality 18
4.3.3 Subgroup Fairness 18

5 Experiments **21**
5.1 Data 21
5.1.1 Data Description 21
5.1.2 Data Preprocessing 21
5.2 Preliminary Experiment: Numerical Descriptions 22
5.3 Experiment 1: Description Quality Clustering 23
5.4 Experiment 2: Description Comparison for SD 23
5.4.1 VHK 23
5.4.2 Depth-First Search Algorithm (DFS) 24
5.4.3 Vertical List Subgroup Discovery (VLSD) 24

5.5 Experiment 3: Subgroup Quality Comparison for SD 24
5.6 Experiment 4: Subgroup Fairness: Individual Notions 24
5.7 Experiment 5: Subgroup Fairness: Group Notions 25

6 Results **26**
6.1 Experiment 1: Description Quality 26
6.2 Experiment 2: Description Comparison for SD 27
6.3 Experiment 3: Subgroup Quality Comparison for SD 28
6.4 Experiment 4: Subgroup Fairness: Individual Notions 29
6.4.1 Logistic Regression (LR) 29
6.4.2 XGBoost Classifier 30
6.5 Experiment 5: Subgroup Fairness: Group Notions 32
6.5.1 Logistic Regression (LR) 32
6.5.2 XGBoost Classifier 34

7 Discussion **36**

8 Conclusion **39**

Abstract

Machine learning models are increasingly employed in all types of applications, as for instance financial applications like credit scoring, where they play a crucial role in loan approvals and financial inclusion. However, these models can reproduce biases present in the data, leading to unfair outcomes, as for instance reinforcing unacceptable gender bias. Existing fairness approaches primarily focus on group fairness, often overlooking disparities within subgroups, a phenomenon known as fairness gerrymandering. To address this issue, this research emphasizes the importance of subgroup fairness and proposes a clustering-based method for subgroup discovery, called Profile-based Subgroup Discovery (PSD). By means of experiments on the well-known German credit dataset, we show that our method generates simple and interpretable clusters, using discriminative decision rules to ensure practical utility and fairness. This study investigates two key areas: the impact of our PSD approach on subgroup identification and on subgroup fairness, particularly within the context of gender bias in credit scoring. The results indicate that while PSD does not surpass our competitor subgroup discovery methods in identifying high-quality subgroups, PSD demonstrated strengths in pinpointing distinct subpopulations of likely non-creditworthy applicants. Regarding subgroup fairness, PSD identifies more subgroups exhibiting gender bias, yet a deeper qualitative analysis of subgroup descriptions is needed to understand these biases fully. PSD maintains fairness in group notions and data alignment but requires further refinement to effectively detect biases while preserving fairness and data fidelity.

Chapter 1

Introduction

Machine learning models are increasingly being utilized in all types of applications, as for instance financial applications like credit scoring, where they play a crucial role in loan approvals and financial inclusion. However, these models can reproduce biases present in the data, leading to unfair outcomes, as for instance reinforcing unacceptable gender bias. In credit scoring, gender bias can result in qualified women being denied loans, hindering their access to financial resources. Although fairness in machine learning has gained significant attention in recent years, most existing approaches focus on group fairness metrics, which are limited in their reliance on group-level averages, which can conceal outcome disparities for subgroups within a disadvantaged group. For instance, even if the average loan approval rates for men and women are similar, the model might still disadvantage women in specific subgroups, such as single mothers. This phenomenon is known as *fairness gerrymandering* [23], where subgroup fairness is sacrificed to achieve parity across entire groups [6]. From a societal point of view, this type of treatment is clearly unfair, and thus unacceptable.

From a technical point of view, it is crucial to incorporate subgroup fairness in the evaluation of AI systems to address fairness gerrymandering. An essential aspect of subgroup fairness is the identification of subgroups. While recent studies have primarily focused on intersectional bias, which involves the combination of multiple sensitive attributes [27], defining subgroups should not be limited to this method. Predefining subgroups based solely on sensitive attributes may overlook other important biased relationships. In this context, it seems sound to utilize subgroup discovery (SD). SD aims to describe relationships between independent variables and specific target variable values, extracting significant rules through data mining techniques [8]. This approach avoids predefined subgroups, identifying those that are most relevant to the decision-making process. Clustering is a method that can be utilized as a subgroup discovery technique by dividing unlabeled data into subgroups based on similarity [11, 12]. Clustering for subgroup discovery (CSD) produces clusters that can be easily distinguished using simple decision rules, resulting in interpretable subgroups. While clustering can be used to identify subgroups, it differs from traditional subgroup discovery, which focuses on finding significant rules related to a target variable. In contrast, CSD aims to find significant rules that describe similar individuals without considering the target variable. This raises questions about how CSD would perform if the target variable were taken into account during the subgroup identification process. Additionally, since subgroup identification is an essential part of subgroup fairness, it also prompts inquiry into how CSD can identify subgroups that are treated unfairly by a classifier compared to conventional SD methods. To address these questions, we propose a novel clustering method designed to generate simple and interpretable clusters for subgroup discovery, called Profile-based Subgroup Discovery (PSD), based on previous semi-hierarchical methods for profile extraction[41, 5]. Our methodology involves two steps: first, partitioning

the data space based on the target variable and then applying iterative clustering to obtain profiles; second, extracting descriptive rules from these profiles to identify subgroups. Like other CSD and SD techniques, PSD relies on discriminative decision rules that can be applied in real-world applications. Our method stands out by integrating the target variable into the clustering process, aligning it closely with subgroup discovery techniques. We will refer to the obtained descriptive rules as 'profile descriptions' for clusters and as 'subgroup descriptions' for those derived from traditional subgroup discovery techniques. This distinction will be consistently maintained throughout the study to ensure clarity.

While significant strides have been made in subgroup discovery, studies regarding clustering for subgroup discovery have often overlooked the importance of integrating target variables directly into the discovery process. This research introduces PSD as a novel technique that enhances the field of subgroup discovery by addressing this gap. PSD improves upon existing CSD methods by incorporating the target variable into the clustering process, allowing for the identification of subgroups that are not only similar but also relevant to specific outcomes. This approach ensures that the discovered subgroups are directly tied to the target variable, providing more meaningful insights. Additionally, this research contributes to the evolving landscape of AI fairness by emphasizing the importance of subgroup fairness in the following manner. Firstly, by introducing PSD as a subgroup discovery technique, this research addresses the limitation of predefined subgroups in subgroup fairness research, providing a more flexible method for identifying and understanding biased relationship within data. Secondly, by focusing on the interpretability of these profiles, the proposed methodology ensures that the identified subgroups can be effectively utilized to evaluate fairness in AI systems. Consequently, this research contributes to the broader goal of creating more equitable and transparent AI systems, aligning with and expanding upon existing studies in the field.

Our study evaluates the performance of our PSD method across two critical dimensions: subgroup identification and subgroup fairness. Regarding the identification of subgroups, we investigate how well instances within a cluster align with their corresponding profile descriptions. We expect profile descriptions generated by PSD to highly align with the instances in the cluster, since the description method will be based on the cluster instances. We expect that the profile descriptions will be more specific for PSD than traditional clustering, since incorporating the target variable is likely to limit the variability in descriptions. Additionally, we will compare the subgroup descriptions generated through PSD with those produced by other subgroup discovery techniques. We hypothesize that PSD will produce lower-quality subgroups than traditional SD techniques, as PSD does not optimize for subgroup quality. Another critical aspect of this study is to examine the subgroup fairness of the subgroups identified by PSD. We will investigate how fairness metrics vary for individuals within the same subgroup when identified by different subgroup discovery methods, with a particular focus on gender bias within the credit scoring context. Furthermore, the study will analyze how fairness metrics differ across subgroups identified by various subgroup discovery methods, as compared to the overall dataset. Because PSD's descriptions are based on similarity within subgroups, we anticipate discovering more diverse subgroups, thus identifying more instances of unfair treatment. Finally, we hypothesize that PSD will be more fair in terms of aligning with the data, as its descriptions are obtained directly from the data rather than by combining features until a satisfactory quality threshold is reached, as in traditional SD methods. These investigations aim to enhance our understanding of PSD's capabilities in identifying unfair subgroups, thereby contributing to the broader discourse on AI fairness.

This thesis is structured as follows: Chapter 2 provides an overview of related work, cov-

ering subgroup discovery, clustering, subgroup quality, and subgroup fairness. This chapter establishes the current state of research, outlining existing challenges and approaches. Chapter 3 discusses the theoretical background, encompassing relevant frameworks and metrics essential for understanding the algorithms used in this study. Chapter 4 presents the methodology and its rationale. In Chapter 5, we detail the experimental design and hyperparameter settings of the algorithms. Chapter 6 presents our research results, emphasizing key findings and insights. Chapter 7 offers a detailed discussion of these findings, examining our methodological choices, study limitations, and suggesting future research directions. Finally, Chapter 8 concludes by summarizing the main contributions of the research and its significance in advancing AI fairness.

Chapter 2

Related Work

In this chapter, we review existing research relevant to our study. We begin by introducing the latest research on subgroup discovery and clustering for subgroup discovery. Then, we examine subgroup discovery techniques and clustering in terms of subgroup quality measures. Finally, we investigate how subgroup fairness has been addressed in prior studies, focusing in particular on subgroup discovery and clustering.

2.1 Subgroup Discovery

Subgroup discovery methods vary based on several criteria, including the number of subgroups returned, the employed data structures, and the search strategies utilized. In terms of output, SD algorithms can either return all subgroups, those above a specified quality threshold, or the top-k subgroups, with the latter being preferred in many studies due to its efficiency in memory usage [2, 34]. Data structures play a crucial role in SD methods; examples include frequent pattern trees (FP-Tree) and Bitset, each tailored to efficiently manage and retrieve patterns from datasets [3, 32]. SD-map [3] employs an FP-tree data structure to efficiently store and retrieve frequent patterns from a dataset. It builds upon the concept of a prefix-tree but includes additional features to manage and represent frequent patterns more effectively. BSD [32] utilizes a Bitset representation to refine subgroups using logical AND operations, optimizing time and memory efficiency across different programming languages. Another distinguishing factor among SD algorithms is their search strategy, where "search" refers to the systematic exploration of the possible subgroup space within a dataset. Among the most prevalent strategies are exhaustive search and heuristic search algorithms [25]. Exhaustive subgroup discovery (SD) algorithms such as depth-first search and breadth-first search traverse the entire search space to ensure optimal subgroup identification. Some commonly used exhaustive search based SD algorithms are APRIORI-SD, EXPLORA and MIDOS. APRIORI-SD [22] employs *unusualness* as a postprocessing measure to evaluate the quality of induced rules and probabilistically classify examples. The evaluation of rule sets includes metrics such as the area under the ROC curve, support, significance of individual rules, and the overall size and accuracy of the rule set [20]. EXPLORA [26] utilizes decision trees to generate subgroups, assessing rule interestingness through statistical metrics like generality and simplicity. MIDOS [42], designed for multirelational databases, evaluates subgroups using a function combining unusualness and subgroup size, applying both minimum support and optimistic estimate pruning. Due to the expensive search space of exhaustive search strategies, these algorithms are often impractical in real-world applications. A common approach to mitigate this issue involves searching for all potential subgroup candidates and subsequently pruning irrelevant ones based on predefined constraints, thereby reducing the hypothesis space.
We identify two types of pruning strategies: anti-monotone constraints and optimistic esti-

mates [25]. The first strategy involves specifying anti-monotone constraints, such as setting a minimum subgroup size N_{min} or a maximum search depth d. These constraints are termed anti-monotone because if a subgroup fails to meet the constraint, none of its more specific subgroups will meet it either, allowing an entire branch of subgroups to be pruned. This approach is highly effective in limiting the otherwise exponentially expanding search space [25]. The second strategy involves calculating optimistic estimates or upper bounds for an interestingness measure of subgroups that have not yet been explored. The second strategy is used in two recent studies, which introduce H-DivExplorer and VLSD. H-DivExplorer [37] combines both of these strategies, using a minimum support threshold as anti-monotone constraint and polarity pruning to improve the performance of subgroup exploration. Polarity pruning focuses on identifying item sets with high divergence, which can be either positive or negative. Polarity pruning works by assigning a polarity to each item based on its impact on divergence when considered alone. The heuristic then combines only items with the same polarity to form item sets, ensuring that item sets consist solely of items that collectively maximize divergence in the same direction. This selective combination effectively narrows down the search space, making the exploration process more efficient. By focusing on high-divergence item sets and combining items with consistent polarity, polarity pruning maintains the quality of the subgroups identified while optimizing the search process. The VLSD algorithm [34] introduces another heuristic exploration approach with an optimistic estimate for Vertical List Subgroup Discovery (VLSD). VLSD uses an optimistic estimate based on the WRAcc score during the exploration process. If the optimistic estimate for a potential subgroup's WRAcc is below a certain threshold, it can be pruned from the search space early, as it is unlikely to lead to a high-quality subgroup. This allows the algorithm to focus on more promising areas of the search space, thus improving efficiency. In contrast to exhaustive search strategies, heuristic SD algorithms such as H-DivExplorer and VLSD use a heuristic function to explore the search space of the problem, which is more efficient and possibly reduces the number of subgroups that must be explored. However, heuristic SD algorithms do not guarantee that the best subgroups will be found [2, 20].

2.2 Clustering for Subgroup Discovery

Another essential focus of this research is the application of clustering algorithms for subgroup discovery, an area that has not been extensively studied. While there is limited literature on the use of clustering algorithms in subgroup discovery, a key consideration is how to effectively describe clusters using combinations of features, akin to traditional subgroup discovery methods. This section reviews recent studies that have explored this theme.

Niemann et al. (2017) [35] propose a combination of subgroup discovery and clustering that aims to identify diverse subgroups while maintaining interpretability. Their method hierarchically reorganizes descriptions obtained from SD by clustering similar descriptions together, ensuring that the resulting clusters capture distinct patterns within the description rules. For each cluster of descriptions, a representative description rule is chosen based on a trade-off between rule confidence and coverage towards the target variable. This allows for easier interpretation by domain experts, who can then delve deeper into individual cluster members. Niemann et al. (2017) propose an approach for creating subgroups based on hierarchically clustering similar descriptions together. However, a key limitation lies in their reliance on pre-defined thresholds for both confidence and coverage. Confidence is the proportion of instances that satisfy both the conditions and the result of the description rule among those that satisfy the conditions, while coverage denotes the fraction of all instances in the dataset that satisfy the conditions of the rule. These pre-defined thresholds can restrict the approach's adaptability to various

datasets. Datasets with unique characteristics or imbalanced class distributions might require adjustments to the thresholds for successful subgroup identification. Additionally, setting these thresholds presents a challenge in balancing certainty with comprehensiveness. Overly high confidence thresholds might lead to overlooking potentially valuable, but less certain, patterns that could hold important insights. Conversely, prioritizing high coverage can mask the presence of smaller, yet important, subgroups. This is particularly problematic in fairness analysis, where understanding the behavior of minority subgroups is crucial.

In contrast, Cooper et al. (2021) introduce a multistep clustering approach aimed at generating interpretable clusters within a two-dimensional space for subgroup discovery [11]. They define these clusters using simple but highly discriminative decision rules that are human-readable and suitable for use in realistic, real-world scenarios. Their process involves calculating SHAP values for a trained XGBoost model to identify how each variable contributes to the model's predictions for each individual instance. The resulting SHAP values are then clustered based on the similarity of the factors predicting a positive outcome. This approach ensures that the clustering process prioritizes the most relevant variables, instead of treating all variables equally and potentially grouping participants based on similarities in irrelevant factors. A limitation of this approach is the reliance on SHAP values, which is inherently dependent on the quality and accuracy of the initial XGBoost model, which may not always capture the underlying data patterns accurately.

Another research, Wilms et al. (2022) [41], introduced a Profile-based Evaluation method for Bias Assessment on Mixed datasets (PEBAM). PEBAM is designed to assess biases in profiles represented within a dataset, which might remain undetected when using only individual or group-bias metrics. Their method involves extracting profiles from the medoids of clusters generated through a variability controlled hierarchical K-medoids (VHK) clustering approach. Buijs (2023)[5] builds upon VHK by introducing a descriptive algorithm to describe the profiles, based on feature relevance. The feature relevance is dependent on the difference in variability, as the smaller the variability is compared to the whole dataset, the more specified a feature has become. This research builds upon these two studies by utilizing the VHK clustering approach in the subgroup discovery field and refining the descriptive algorithm to improve subgroup identification methods.

2.3 Subgroup Quality

A third important aspect of this research is the assessment of the subgroups obtained from both the CSD and the SD algorithms. Rizkallah et al. (2019)[38] conducted an analysis of SD quality measures, offering researchers an overview of various approaches to identifying meaningful subgroups. The term 'meaningful' is subjective and is often represented in this context by the concept of 'interestingness'. One widely accepted measure of interestingness is *unusualness*, also called the Weighted Relative Accuracy (WRAcc)[36], which is defined in section 3.3.1. WRAcc is frequently used in algorithms like CN2-SD[29], Apriori-SD[22], and VLSD[34]. The size of a subgroup is also an important factor, evaluated through metrics like coverage and support. Coverage denotes the fraction of all instances in the dataset that satisfy the conditions of the rule[35, 37] and support denotes the fraction of all positive instances in the dataset that satisfy the conditions of the rule[38]. These metrics are employed in algorithms such as CN2-SD, H-DivExplorer[37], and Cluster Grouping (CG)[44]. Additionally, the confidence score, or precision, is used by researchers like Niemann et al. (2017)[35] and Carmona et al.(2011)[8] as an internal quality criterion for generating candidate subgroups. Carmona et al. (2011) evaluate their method based on coverage, significance, WRAcc, support, and confidence.

Herrera et al. (2011)[20] provided a comprehensive overview of quality measures for subgroup discovery, further defining additional measures like false alarm, novelty, interest, and precision. These various measures help ensure that the subgroups identified are not only statistically significant but also practically meaningful and relevant to the research context.

2.4 Subgroup Fairness

A notable limitation of group fairness metrics is their dependence on group-level averages, which can mask disparities in outcomes for subgroups within a disadvantaged group. This issue, known as fairness gerrymandering, is highlighted by Kearns et al. (2018) [23]. Research on subgroup fairness often addresses intersectional bias, focusing on the intersection of sensitive attributes, such as 'black females.' Expanding on the Max-Min fairness definition, Ghosh et al. (2021)[16] introduced a method to address intersectionality in fairness assessments. Their approach assesses the fairness of any combination of subgroups for any given fairness measure, and then calculates a ratio based on the most and least fair outcomes across these subgroup combinations. While this method is flexible and can work with various fairness measures, it faces challenges with data sparsity. As the number of factors considered for intersectionality increases, the data available for each specific subgroup combination can become scarce, potentially impacting the reliability of the fairness assessment [17]. Additionally, Foulds et al. (2020) [15] introduced differential fairness, a multi-attribute fairness definition informed by intersectionality, positing that the probabilities of outcomes should be similar regardless of the combination of protected attributes. However, this approach only considers the intersectionality of protected attributes, overlooking potential biased relationships between protected and non-protected attributes. Auditing for intersectional bias is straightforward when the analysis involves a limited number of protected attributes and a single fairness metric. However, the complexity increases exponentially as the dimensionality of intersectionality expands and multiple fairness metrics are considered. This combinatorial explosion in the number of subgroups to be examined renders an exhaustive analysis computationally intractable [7]. To assist data scientists in auditing models for intersectional bias, various visualization tools such as FairVis [7] and FairCompass [33] have been developed in recent years. Regarding fairness metrics, Chang et al. (2023)[9] investigated the relationship between the average feature importance disparity of a subgroup and their fairness, using true positive rate, false positive rate and expected calibration error as detailed by Barocas et al. (2023)[4]. Zhang et al. (2021) [43] proposed an explorative model building system, FairRover, for responsible fair model building. FairRover can evaluate models for different performance metrics, including demographic parity, false discovery rate, false omission rate, and error rate. Their approach provides a comprehensive framework for assessing bias across multiple dimensions, highlighting the complex interplay between different fairness criteria in real-world applications. Kearns et al. (2018) [23] propose a new set of fairness definitions that bridge the gap between statistical and individual fairness, combining the strengths of both approaches. Specifically, they introduced variants of fairness metrics such as statistical parity subgroup fairness and false positive subgroup fairness. By redefining these metrics to apply to subgroups instead of individuals, they established new subgroup fairness metrics. Since our research focuses on evaluating subgroup fairness, these measures form the foundational steps of this study.

Chapter 3

Theoretic Background

This chapter outlines the key concepts fundamental to this research. We begin with an introduction to subgroup discovery, exploring the DFS and VLSD algorithms in detail, considering DFS as a baseline method and VLSD as a significant competitor for evaluation due to its advancements in subgroup identification. Next, we explain the variability-controlled hierarchical K-medoids (VHK) clustering and its role in identifying subgroups. Finally, we present the metrics used for evaluating both the quality and fairness of the profiles.

3.1 Subgroup Discovery Tasks

Subgroup discovery aims to describe relationships between independent variables and specific target variable values, extracting significant rules through data mining techniques [8]. This research compares our proposed method with two established subgroup discovery approaches: depth-first-search (DFS) and vertical list subgroup discovery (VLSD).

3.1.1 DFS

One widely used and straightforward approach in subgroup discovery is depth-first search (DFS). This algorithm begins with a subgroup defined by a single selector and gradually extends the search by adding more selectors to the subgroup. If it reaches a maximum depth or another specified pruning criterion, the algorithm backtracks to explore other potential subgroups similarly within the search tree [30, 25]. While DFS is efficient in terms of memory usage, its ability to incorporate pruning through optimistic estimates is limited, as such estimates typically rely on data that may only become available after visiting a significant portion of the subgroup space [25]. The primary advantage of DFS lies in its ability to exhaustively identify all possible subgroups.

3.1.2 VLSD

The Vertical List Subgroup Discovery (VLSD) algorithm, as proposed by Lopez-Martinez-Carrasco et al. (2023) [34], is a subgroup discovery algorithm based on a vertical list data structure. VLSD combines an equivalence class exploration strategy and a pruning strategy based on optimistic estimates. The equivalence class exploration strategy begins by dividing the dataset into equivalence classes based on the values of one or more attributes. This initial clustering of similar data points facilitates subsequent processing. The algorithm then hierarchically explores these classes. An illustration regarding the exploration strategy is displayed in Figure 3.1. The algorithm begins by creating an empty list to store the discovered subgroups and calculating the true and false populations of the dataset. This essentially determines how

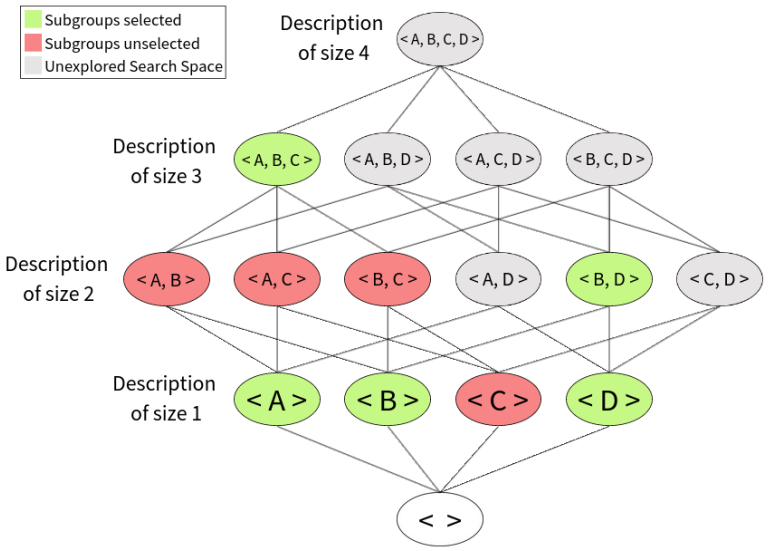


Figure 3.1: Examples of VLSD search spaces with optimistic estimate [34]

many instances meet the criteria defined by the subgroup being considered ('true' population) and how many do not ('false' population). It then generates all possible subgroups with a single attribute, evaluates their quality, and adds them to the list, sorting them based on a given criterion (e.g., WRAcc). Next, the algorithm initializes a triangular matrix to store size-2 subgroups, which involves pairing each attribute with every other attribute and evaluating the resulting subgroups. The matrix helps in efficiently pruning the search space by eliminating subgroups that do not meet the quality thresholds. For each attribute, the algorithm evaluates the size-2 subgroups, adds them to the list, and recursively explores further subgroups by refining these initial pairs. By leveraging the matrix, VLSD avoids unnecessary computations, making it more efficient. This process of evaluation, storage, and recursive exploration continues until all promising subgroups have been identified and evaluated. The result is a comprehensive list of subgroup descriptions that are significant according to the quality measures, providing insights into the structure and patterns within the data. To improve efficiency, VLSD employs an optimistic estimate to prune unpromising paths early in the search process. This estimate predicts the maximum potential quality of subgroups that could be found along a particular path, and if it falls below a threshold, the path is discarded and no further subgroups along that path will be explored [34]. The VLSD algorithm leverages vertical lists to manage equivalence classes, allowing for quick access and manipulation of data points during the search. This approach not only reduces computational complexity but also facilitates parallel execution, significantly speeding up the subgroup discovery process. The combination of hierarchical exploration, early pruning, and efficient data management ensures that the algorithm can handle large datasets effectively and identify high-quality subgroups.

3.2 Variability Controlled Hierarchical K-medoids (VHK)

The clustering algorithm used for this research is the variability controlled hierarchical K-medoids (VHK) clustering algorithm from Wilms et al. (2022) [41]. They have implemented this clustering algorithm by repeatedly partitioning the dataset into clusters using a semi-hierarchical approach. Initially, all data points are grouped together and progressively divided into smaller clusters across multiple layers, using k-medoids clustering with Gower distance to handle mixed data types. At each layer, the variability of subsets is assessed to determine whether further splitting is necessary, ensuring clusters maintain sufficient variability. Since

our approach is based on this clustering algorithm and variability, variability is being expanded upon below.

3.2.1 Variability

Whether clusters have sufficient variability is determined by the variability score of the features. This score indicates the extent to which a feature f varies within a dataset D [41]. Variability scores range from 0 to 1, with 0 indicating no variation and 1 indicating high variation. Since a dataset consists of numerical and categorical features, the variability score is defined for each. The numerical variability score can be calculated as follows:

Definition 3.2.1 (Numerical Variability Score)**.** Given a numerical non-constant feature X, its variability score is calculated based on the sample standard deviation s. Normalize the feature X between 0 and 1. Then, compute the sample standard deviation s from the normalized samples. Finally, to measure the variability, standardize the sample standard deviation ss as follows:

$$\text{ss} = \frac{s}{0.29}$$

where 0.29 is the approximate standard deviation of a random variable uniformly distributed between 0 and 1.

The standardized standard deviation provides an indication of how much the data points deviate from the mean. A low value of ss (close to zero) indicates that most sample points are close to the mean, suggesting less variability and thus X is a potentially relevant feature. In contrast, a high value of ss suggests greater variability in the data points, suggesting X is an irrelevant feature.

The categorical variability score used is defined by Allaj (2018) [1] as follows:

Definition 3.2.2 (Categorical Variability Score)**.** The variability of an outcome resulting from categorical data with n elements falling in k categories 0, 1, 2, ..., k-1 with vector of the relative frequencies equal to

$$f = (f_0, f_1, ..., f_{k-1})$$

is measured by the following variability measures

$$v_k = 1 - \parallel f \parallel_k = 1 - \sqrt{f_0^2 + f_1^2 + ... + f_{k-1}^2}$$

$$v_{k,s} = \frac{v_k}{1 - \frac{1}{\sqrt{k}}}$$

where $f_i = n_i/n$, for any i = 0, 1, 2, ..., k - 1, where n_i gives the number of elements falling in category i with n = $\Sigma_{i=0}^{k-1} n_i \leqslant 1$, is the relative frequency of the ith category and $\| \cdot \|_k$ is the Euclidean norm defined on the real space R_k.

The variability score is bounded by 0 and 1 - 1/k, where values close to 1 - 1/k show greater variability in the data points, suggesting an irrelevant feature. In contrast, values close to 0 show low variability, suggesting a relevant feature.

3.2.2 VHK

VHK is implemented by repeatedly partitioning the dataset into clusters using a semi-hierarchical approach. Initially, all data points are grouped together and progressively divided into smaller clusters across multiple layers, using k-medoids clustering with Gower distance to handle mixed data types. At each layer, the variability of subsets is assessed to determine whether further splitting is necessary, ensuring clusters maintain sufficient variability. Clusters with less than a pre-defined percentage of the total data or insufficient variability were not further divided. Whether a cluster had insufficient variability was decided as follows. For each feature in the dataset, it was checked whether the variability within the current subset falls within the specified range (between 0.2 and 0.8). If any feature meets this criterion, the subset is considered to have sufficient variability for further clustering, and only those subsets with meaningful diversity are split further. To clarify the reasoning of setting this range between 0.2 and 0.8, a variability below 0.2 suggest that there is high homogeneity between the cluster instances. This means that the instances within the cluster are very similar to each other, which indicates that the cluster is already cohesive and well-defined, making further splitting unnecessary. In contrast, a variability above 0.8 suggests that the instances of that cluster are very diverse, meaning that there may not be meaningful, homogeneous subgroups to be found. Splitting such a cluster thus may not be useful or result in interpretable clusters. Furthermore, when the variability is very high, any further splitting might result in overfitting to noise rather than capturing meaningful patterns. For each cluster that reaches its final state, where no further splitting occurs, the algorithm stores the instances and medoids.

3.3 Subgroup Quality Evaluation

The subgroup quality evaluation metrics used in this study are based on the definitions by Lopez-Martinez-Carrasco et al. (2023) [34]: Sensitivity, Specificity, WRAcc score, and WRAcc optimistic. While these metrics are commonly used for quality evaluation, Lopez-Martinez-Carrasco et al. (2023) have adapted them specifically for subgroup quality analysis. Our study adopts these same evaluation metrics but renames them to avoid confusion regarding ground truths. Instead of using traditional terms like tp (true positive), fp (false positive), fn (false negative), and tn (true negative), we use different names that better reflect the concepts in our context. This change is necessary because there is no classifier making predictions here; we are only comparing subgroup descriptions with their corresponding labels. Sensitivity, Specificity, WRAcc score and WRAcc optimistic will be renamed as follows: SD Sensitivity, SD Specificity, SD WRAcc score and SD Optimistic WRAcc, respectively. By placing 'SD' before each scores, we hope to remind you that these metrics are not applied in their traditional way, but in the context of assessing the quality of subgroups without looking at predictions. To provide clarity regarding the values of the metrics, we have defined the metrics below, based on the confusion matrix from table 3.1

Table 3.1: Confusion matrix of a subgroup [34]

		Instance in subgroup description		
		True	**False**	**Total**
Instance has label	**Positive**	SG^+	$NotSG^+ = N_{positives} - SG^+$	$N_{positives}$
	Negative	SG^-	$NotSG^- = N_{negatives} - SG^-$	$N_{negatives}$
	Total	SG_{size}	$NotSG_{size} = N - SG_{size}$	N

Table 3.1 shows the amount of positive instances within a subgroup description as SG^+, the amount of negative instances within a subgroup description SG^-, the amount of positive instances that fall outside the subgroup description $NotSG^+$ and the amount of negative instances that fall outside the subgroup description $NotSG^-$. Then, SG_{size} refers to the subgroup size, $NotSG_{size}$ refers to the size of the data minus the subgroup size and N is the total amount of instances. Similarly, $N_{positives}$ refers to the amount of positive instances in the dataset and $N_{negatives}$ to the amount of negative instances in the dataset.

3.3.1 Quality of subgroups

Now that we have explained the values used in the evaluation metrics, we explore the evaluation metrics further. The quality of the subgroups will be evaluated based on SD Sensitivity, SD Specificity, SD WRAcc and SD Optimistic WRAcc, as these metrics provide a comprehensive and balanced assessment of subgroup performance.

First, the SD Sensitivity measures the proportion of positive instances within the subgroup. It gives a measure of how prevalent the positive instances are in the identified subgroup:

$$\text{SD Sensitivity} = \frac{SG^+}{SG_{size}} \tag{3.1}$$

Second, the SD Specificity evaluates the proportion of actual negatives in the rest of the data. This metric provides complementary insight to the SD Sensitivity metric by focusing on the instances outside the subgroup. Together, these metrics give a complete picture of the distribution of positive and negative instances both inside and outside the subgroup.

$$\text{SD Specificity} = \frac{NotSG^-}{NotSG_{size}} \tag{3.2}$$

Third, the SD WRAcc addresses the trade-off between discovering large subgroups and ensuring they are highly relevant to the target property. By taking the subgroup size into account and the relative concentration of positive instances within a subgroup, the SD WRAcc is defined as:

$$\text{Positive Rate} = \frac{N_{positives}}{N} \tag{3.3}$$

$$\text{Relative Concentration } SG^+ = \frac{SG^+}{N_{positives}} - \frac{SG_{size}}{N} \tag{3.4}$$

$$\begin{aligned} \text{SD WRAcc} &= \text{Positive Rate} \cdot \text{Relative Concentration } SG^+ \\ &= \frac{N_{positives}}{N} \cdot \left(\frac{SG^+}{N_{positives}} - \frac{SG_{size}}{N}\right) \end{aligned} \tag{3.5}$$

Additionally, the SD Optimistic WRAcc score [34, 18] provides an upper bound estimate of the SD WRAcc, which helps in assessing the potential maximum quality of the subgroups. The SD Optimistic WRAcc score highlights subgroups that could be highly valuable if the optimistic conditions hold true:

$$\text{SD Optimistic WRAcc} = \frac{(SG^+)^2}{N_{positives}} \cdot \left(1 - \frac{SG_{size}}{N}\right) \tag{3.6}$$

3.4 Subgroup Fairness Evaluation

The subgroup fairness evaluation metrics used to investigate the fairness of the subgroups within the subgroup are the individual fairness notions named wrong disadvantage, demographic parity difference and equalized odds difference. The subgroup fairness evaluation metrics used to investigate the fairness between subgroups and the data are the group fairness notions named demographic parity subgroup fairness and equalized odds subgroup fairness.

3.4.1 Individual Fairness Notions for Subgroup Fairness

Wrong Disadvantage (WD), as defined by Wilms et al.[41], measures the percentage of instances incorrectly labeled as negative by a classifier. This metric is particularly relevant as it highlights the potential for unjust exclusion from financial opportunities based on erroneous predictions.

$$\text{Wrong Disadvantage (WD)} = \frac{fn}{size} \tag{3.7}$$

In addition to WD, we also examine demographic parity difference and equalized odds difference with respect to protected attributes (in our case: 'Sex'). Demographic parity, as defined by Rosenblatt et al. (2023)[39], measures the extent to which the probability of receiving a positive outcome is independent of membership in a protected class:

Definition 3.4.1 (Demographic parity). Given a predictor $\hat{Y} : X \times A \to Y$, we say $\hat{Y}$ satisfies demographic parity if, for all instances of protected attributes a and a',

$$\Pr(\hat{Y}(x, a) = y \mid A = a) = \Pr(\hat{Y}(x, a') = y \mid A = a') \tag{3.8}$$

where the probability is taken over the conditional distribution of X and the possible randomness of $\hat{Y}$.

Demographic parity (DP) is satisfied when the outcomes of a model's classification are independent of a specific sensitive attribute, such as 'Sex' in our case. This means that the positive rate between different subgroups should be equal. Disparity metrics, like demographic parity difference, assess the extent to which a particular predictor deviates from meeting a parity requirement. Demographic parity difference (DPD) is defined as the difference in positive rate between the largest and smallest two demographic groups [13, 28]:

$$\text{DPD} = \max_a \Pr(\hat{Y}(x, a) = y \mid A = a) - \min_a \Pr(\hat{Y}(x, a') = y \mid A = a') \tag{3.9}$$

The advantage of DPD in comparison to DP is that it is more measurable. A DPD value of 0 indicates perfect demographic parity, all groups have equal prediction probabilities. The higher the difference, the greater the disparity in model outcomes between groups. In contrast to demographic parity, equalized odds allows the prediction $\hat{Y}$ to depend on the sensitive attribute A, but only through the target variable Y. Equalized odds, as defined by Hardt et al. (2016)[19], ensures that the model's predictions are equally accurate across subgroups by requiring that the true positive rate and the false positive rate be the same for all subgroups.

Definition 3.4.2 (Equalized Odds). We say that a predictor $\hat{Y}$ satisfies equalized odds with respect to protected attribute A and outcome Y , if $\hat{Y}$ and A are independent conditional on Y.

This definition promotes the use of features that accurately predict Y while preventing A

from being used as a proxy for Y. Since our problem is a binary scenario, equalized odds requires that:

$$\Pr(\hat{Y} = 1 \mid A = l, Y = y) = \Pr(\hat{Y} = 1 \mid A = a', Y = y), \quad y \in \{0, 1\} \tag{3.10}$$

This constraint requires that for the outcome $y = 1$ the true positive rates across the two demographics $A = 0$ and $A = 1$ need to be equal. Similarly, for $y = 0$, the false positive rates need to be equal. The equalized odds difference (EOD) is defined as the larger of the true positive rate difference (TPD) and the false positive rate difference (FPD) [28]. This can be expressed as:

$$TPD = \max_a(\Pr(\hat{Y} = 1 \mid A = a, Y = 1)) - \min_a(\Pr(\hat{Y} = 1 \mid A = a', Y = 1)) \tag{3.11}$$

$$FPD = \max_a(\Pr(\hat{Y} = 1 \mid A = a, Y = 0)) - \min_a(\Pr(\hat{Y} = 1 \mid A = a', Y = 0)) \tag{3.12}$$

$$EOD = \max(TPD, FPD) \tag{3.13}$$

where $A = l$ stands for the largest value of that attribute in the subgroup and $A = s$ stands for the smallest value of that attribute in the subgroup. An EOD value of 0 indicates that all demographic groups have the same true positive, false positive, true negative and false negative rates. The higher the difference, the greater the disparity in model outcomes between groups.

3.4.2 Group Fairness Notions for Subgroup Fairness

The DPD Subgroup Fairness and EOD Subgroup Fairness are defined by Kearns et al. (2018) [24] as follows:

Definition 3.4.1 (Demographic Parity Subgroup Fairness)**.** Fix any classifier D, distribution P, collection of group indicators G, and parameter $\gamma \in [0, 1]$. For each $g \in G$, define:

$$\alpha_{DP}(g, P) = \Pr_P[g(x) = 1]$$
$$\beta_{DP}(g, D, P) = |DP(D) - DP(D, g)|$$

where $DP(D) = \Pr -P, D[D(X) = 1]$ and $DP(D, g) = \Pr_{P,D}[D(X) = 1|g(x) = 1]$ denote the overall acceptance rate of D and the acceptance rate of D on group g respectively. We say that D satisfied γ-demographic parity Fairness with respect to P and G if for every $g \in G$, $\alpha_{DP}(g, P)\beta - DP(g, D, P) \leq \gamma$. We will sometimes refer to $DP(D)$ as the SP base rate.

It is important to understand that the definitions Kearns et al. (2018) give refer to two approximation parameters. The parameter α determines the maximum fraction of the population that can be disregarded. Likewise, we permit deviations up to β from the base rate in the probability of a positive classification for each subgroup, rather than requiring an exact match.

Definition 3.4.2 (Equalized Odds Subgroup Fairness)**.** Fix any classifier D, distribution P, collection of group indicators G, and parameter $\gamma \in [0, 1]$. For each $g \in G$, define:

$$\alpha_{EO}(g, P) = \Pr_P[g(x) = 1, y = 0]$$
$$\beta_{EO}(g, D, P) = |EO(D) - EO(D, g)|$$

where $EO(D) = \Pr_{D,P}[D(X) = 1|y = 0]$ and $EO(D, g) = \Pr_{D,P}[D(X) = 1|g(x) = 1, y = 0]$ denote the overall false-positive rate of D and the false-positive rate of D on group g respectively. We say that D satisfies γ-False Positive (FP) Fairness with respect to P and G if for every $g \in G$, $\alpha_{EO}(g, P)\beta_{EO}(g, D, P) \leq \gamma$. We will sometimes refer to $FP(D)$ as the FP-base rate.

If the classifier D does not meet the γ-fairness condition for either DP or EO fairness, we consider D to be γ-unfair with respect to P and G. Any subgroup g that demonstrates this lack of fairness is referred to as a γ-unfair certificate for (D, P).

Chapter 4

Methodology

This chapter outlines the methodology employed in this research. We begin by explaining the selection of subgroup discovery techniques. Following this, we describe the overall pipeline of our method, which is VHK as a subgroup discovery technique, explaining the choice for VHK and how we describe the subgroups. Additionally, we discuss our evaluation approach, focusing on the descritpion quality, subgroup quality and subgroup fairness.

4.1 Subgroup Discovery Techniques

This research employs two subgroup discovery algorithms; the Depth First Search (DFS) algorithm and the Vertical List Subgroup Discovery (VLSD). The DFS algorithm is chosen for its simplicity, providing a systematic approach to exploring potential subgroups and facilitating a thorough examination of the dataset. Since the DFS algorithm is an exhaustive search algorithm, it guarantees to find all relevant subgroups. For these reasons, the DFS algorithm will serve as the baseline for this research. The VLSD algorithm, introduced by Lopez-Martinez-Carrasco et al. (2023) [34], is a new and efficient SD algorithm that has demonstrated superior performance compared to other SD algorithms such as SD-Map [3] and BSD [32], making it a valuable choice for comparison. Moreover, its widespread availability enhances its suitability for this research. For both subgroup discovery techniques, the same methodology is applied. The algorithms are executed on the dataset, excluding the target feature and any sensitive attributes, focusing solely on deriving descriptive rules. These rules are then standardized into a unified format to facilitate subsequent evaluation processes. The evaluation of the SD algorithms are described in section 4.3.

4.2 Profile-based Subgroup Discovery

This research investigates the use of clustering algorithms for subgroup discovery. Unlike traditional SD techniques like DFS and VLSD that identify subgroups based on a target variable, clustering algorithms group instances based on their inherent similarities. To leverage clustering for SD, we propose a two-step approach: (1) adapting the clustering pipeline to incorporate the target variable and (2) developing a method to interpret and describe the resulting clusters. Listing 4.1 provides pseudocode to illustrate the main idea of our approach. The following sections detail the modifications to the clustering pipeline and the subsequent method for obtaining profile descriptions based on clusters.

```
def obtain_profiles(data, clusters):
    profile_descriptions = []
    for depth in [1, 2, 3]:
        for cluster in clusters:
            # Get variability difference scores for each feature
            # (Buijs (2023))
            variability_difference = variability_difference(data, cluster)

            # Get top-d descriptions based on variability difference score
            description = get_variability_difference_descriptions(cluster,
                variability_difference, depth)
            profile_descriptions.append(description)

    return profile_descriptions

def PSD(data):
    # (1) Clustering pipeline
    positive_data, negative_data = split_data_by_label(data)

    # Obtain clusters from VHK (Wilms et al. (2022))
    positive_clusters = VHK(positive_data), len(positive_data), k=2, l=25)
    negative_clusters = VHK(negative_data), len(negative_data), k=2, l=25)


    # (2) Profile Descriptions based on clusters
    # Obtain profile descriptions based on Buijs (2023)
    positive_profiles = obtain_profiles(positive_data, positive_clusters)
    negative_profiles = obtain_profiles(negative_data, negative_clusters)

    return positive_profiles, negative_profiles
```

Listing 4.1: Pseudocode PSD method

4.2.1 Clustering Pipeline

To incorporate the target variable as in subgroup discovery, we employ a targeted clustering approach. This involves splitting the data into two subsets based on the target variable's values (i.e., positive and negative labels). Subsequently, we apply the chosen clustering algorithm independently to each subset. Figure 4.1 illustrates the difference between applying clustering directly to the data and using our approach. This strategy offers several advantages. Firstly, the resulting clusters will only contain instances with the same target label (positive or negative), facilitating a clear comparison between positive and negative profiles. For clarity, we will refer to the positive profiles as PSD+ and the negative profiles as PSD-. This allows us to investigate potential similarities or significant differences in their characteristics. Additionally, this approach enables a more focused analysis of subgroups within each target class, potentially leading to more actionable insights in machine learning fairness. By concentrating on one specific target variable at a time, such as clustering positive instances, this approach aligns with subgroup discovery techniques that also focus on describing rules for a specific target variable. The clustering algorithm used in this research is the variability controlled hierarchical K-medoids (VHK), as proposed by Wilms et al. (2022)[41]. The primary advantage of K-medoids clustering compared to K-means is its increased resilience to noise and outliers [41].

4.2.2 Profile Descriptions based on Clusters

The second step in our approach is developing a method to interpret and describe the resulting clusters. After applying the clustering algorithm to both the positive instances and the nega-

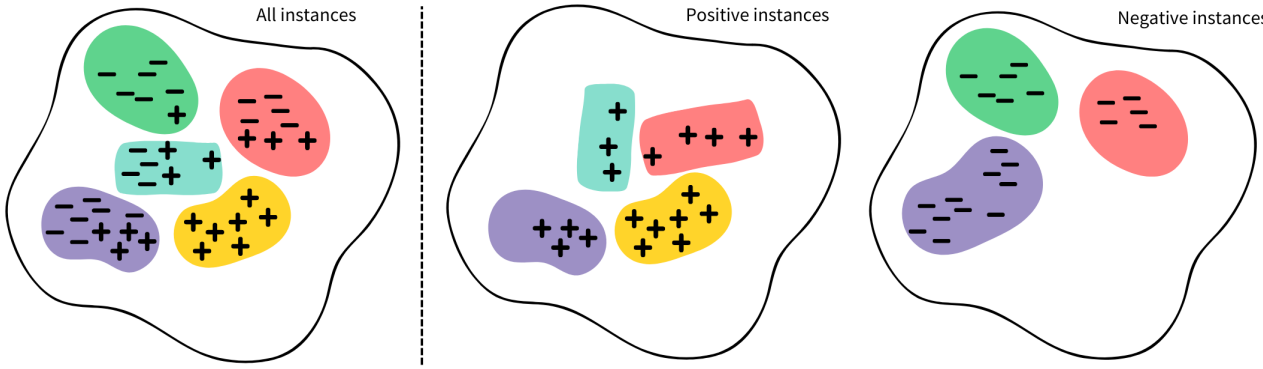


Figure 4.1: Illustration of traditional clustering (left) in comparison to our clustering method (right)

tive instances, we obtain two sets of clusters: positive clusters and negative clusters. For each cluster set, we will describe each cluster based on our description method. After applying our description method on a cluster set, all duplicate profile descriptions are removed. This way, only the unique profile descriptions are left for evaluation.

Our description method is defined as follows. The profile descriptions are obtained based on the d most relevant features to describe the cluster. One way to obtain the most relevant features is the variability score, as employed by Wilms et al. (2022) [41] and described in section 3.2.1. However, one of the disadvantages of the variability score is that it doesn't take the whole dataset into account. The smaller the variability is compared to the whole dataset, the more specified this feature has become. Therefore, in our approach, the relevance of the features is determined through the variability difference score. The variability difference is defined by Buijs (2023) [5] as follows:

$$\mathrm{R} = -\Delta\mathrm{var} = -(\mathrm{var}(\mathrm{Cluster}) - \mathrm{var}(\mathrm{Data})) \tag{4.1}$$

where var(Cluster) is the variability score of the cluster of interest and var(Data) is the variability score over the whole dataset. The variability difference ranges from -1 to 1, where -1 indicates the least relevant features and 1 the most relevant features. In contrast to Buijs (2023) [5], however, instead of classifying all features into relevant, neutral and irrelevant, we only consider d most relevant features to describe the cluster. Therefore, we sort the relevant features from highest to lowest variability difference and pick the top-d relevant features.

4.3 Evaluation

There are three main aspects on which our method needs to be evaluated. First, the quality of the profile descriptions need to be analyzed to assess our description method and gain an understanding in how well instances within a cluster align with their corresponding profile descriptions. Second, the quality of the profiles need to be analyzed and compared to the subgroups derived from subgroup discovery. Third, the fairness of the profiles need to be analyzed and compared to the subgroups obtained from subgroup discovery. We will investigate how fairness metrics vary for individuals within the same subgroup when identified by different subgroup discovery methods, with a particular focus on gender bias within the credit scoring context. Furthermore, we will analyze how fairness metrics differ across subgroups identified by various subgroup discovery methods, as compared to the overall dataset.

4.3.1 Description Quality

To assess our description method and gain an understanding in how well instances within a cluster align with their corresponding profile descriptions, we employ coverage and purity metrics. Coverage measures the proportion of instances in the profile description that are actually present in the cluster. Purity measures the proportion of instances in the cluster that actually belong to the profile description. In essence, coverage tells us if the profile description captures most of the relevant instances, and purity tells us if the profile description captures the instances in the cluster [38]. To evaluate the description quality, we will look at the difference in coverage and purity for our clustering method for SD in comparison to traditional clustering, as illustrated in Figure 4.1.

4.3.2 Subgroup Quality

Motivated by their popularity in the literature, the quality of the profiles and subgroups will be evaluated based on the SD Sensitivity, the SD Specificity, the SD Weighted Relative Accuracy (WRAcc) score and the SD Optimistic WRAcc, as these metrics provide a comprehensive and balanced assessment of subgroup quality. The SD Sensitivity measures the proportion of actual positive instances within a subgroup. This metric helps to understand the internal composition of the subgroup. A higher value indicates that the subgroup predominantly consists of positive instances. In the context of fairness, it can reveal whether certain subgroups have a disproportionate number of positive instances compared to others, which might indicate bias. Second, the SD Specificity evaluates the proportion of actual negative instances outside the subgroup. When the SD Specificity is lower than the proportion of negative instances in the whole dataset, this might indicate a negative bias in the subgroup. This metric serves as a starting point for investigating potential bias in the subgroup. It helps identify subgroups that might warrant further analysis to understand if they are enriched with the negative class in a fair and meaningful way. Third, the SD WRAcc is a metric that combines the positive rate and the relative concentration of positive instances in the subgroup, prioritizing subgroups that are both significant in size and have a high positive rate. This score addresses the trade-off between discovering large subgroups and ensuring they are highly relevant to the target property. Additionally, the SD Optimistic WRAcc score [34, 18], provides an upper bound estimate of the SD WRAcc, which helps in assessing the potential maximum quality of the subgroups.

4.3.3 Subgroup Fairness

Subgroup fairness depends significantly on the classifier applied to the data, as each classifier may treat certain subgroups differently. To evaluate subgroup fairness, we will employ two distinct classifiers in our analysis. Additionally, we will introduce and discuss the metrics used to assess both internal and external subgroup fairness, providing a robust framework for understanding how each classifier impacts subgroup equity.

Classifiers

The classifiers employed in this research are logistic regression and XGBoost. These choices were made due to their relevance, interpretability, and prevalence in credit scoring. Logistic regression is widely used for its simplicity and transparency. It outputs coefficients that directly translate to the impact of each feature on the probability of default. This allows banks and financial institutions to understand how factors such as checking account influence the probability of a client defaulting. XGBoost, on the other hand, is a popular gradient boosting

algorithm known for its high predictive accuracy and efficiency, often outperforming other models in credit scoring tasks [10]. Financial institutions frequently adopt XGBoost and similar gradient boosting algorithms because they achieve a balance between performance and interpretability. While XGBoost does not provide the same level of individual feature interpretation as logistic regression, it offers feature importance scores, indicating which features contribute most to the model's predictions. By utilizing both logistic regression and XGBoost, we aim to gain a comprehensive understanding of subgroup fairness in credit scoring.

Fairness Metrics

To comprehensively assess subgroup fairness, we differentiate between fairness within the subgroup and fairness between a subgroup and the data, internal and external fairness respectively. Internal subgroup fairness examines the gender bias within the subgroups, while external fairness compares the treatment of entire subgroups relative to the whole dataset. We will explore metrics for both aspects, where internal fairness is evaluated based on individual notions of subgroup fairness, and external fairness is evaluated based on group notions of subgroup fairness.

Internal Fairness
To evaluate the fairness within a subgroup we employ three distinct fairness metrics; wrong disadvantage, demographic parity difference and equalized odds difference, as described in section 3.4. For each of these metrics, we will evaluate each identified subgroup for the logistic regression classifier and the XGBoost algorithm.

External Fairness
To evaluate the subgroup fairness in comparison to the whole dataset, we employ four subgroup fairness metrics; weighted demographic parity subgroup fairness, weighted true positive difference subgroup fairness, weighted false positive difference subgroup fairness and weighted equalized odds subgroup fairness. Each identified subgroup is evaluated using these metrics for both the logistic regression classifier and the XGBoost algorithm. These subgroup fairness metrics are variations of those introduced by Kearns et al. (2018)[23]. Instead of introducing new subgroup fairness metrics, we adapt existing group fairness metrics to fit our context of external fairness. For instance, while Rosenblatt et al. (2023)[40] defined demographic parity as the probability of receiving a positive outcome being independent of membership in a protected class, we redefine demographic parity for subgroups as the probability of receiving a positive outcome being independent of subgroup membership. We aim to obtain measurable results using the demographic parity difference (DPD) subgroup fairness metric, which is defined as:

$$\text{DPD Subgroup Fairness} = \Pr(\hat{Y}(x) = 1) - \Pr(\hat{Y}(x) = 1 \mid g(x) = 1) \tag{4.2}$$

which can be written in simple terms:

$$\text{DPD Subgroup Fairness} = \frac{tp + fp}{tp + fp + tn + fn} - \frac{tp_g + fp_g}{tp_g + fp_g + tn_g + fn_g} \tag{4.3}$$

where the right side of the subtraction sign are the ground truths of the subgroup specifically.

The other metric we introduce based on Kearns et al. (2018) and the equalized odds difference is the equalized odds subgroup fairness. This can be defined as:

$$\text{EOD Subgroup Fairness} = max(TPD, FPD) \tag{4.4}$$

with:

$$TPD = \frac{tp}{tp + fn} - \frac{tp_g}{tp_g + fn_g} \tag{4.5}$$

$$FPD = \frac{fp}{fp + tn} - \frac{fp_g}{fp_g + tn_g} \tag{4.6}$$

These metrics will be reweighted by the subgroup size and used to evaluate the profile descriptions and to assess the subgroup descriptions from DFS and VLSD, facilitating a comparison among these three subgroup discovery techniques.

Chapter 5

Experiments

This chapter describes the experiments done in this research. First, we will describe the dataset used in detail and explain the preprocessing steps. Then, we describe a preliminary experiment regarding the numerical features. Consequently to this result, we added a preprocessing step, meant to transform the mixed dataset to a nominal dataset. Additionally, we describe a first experiment we perform to evaluate the quality of the profile descriptions. Furthermore, the experiments for the subgroup quality assessment are described, focusing on the descriptions and quality measures. Lastly, we describe the experiments for subgroup fairness, focusing on the individual notions and group notions separately.

5.1 Data

The data used for this research is data concerning the task of credit scoring. The German Credit Risk dataset is chosen based on a variety of reasons. First, the German Credit Risk dataset concerns a binary classification task, which simplifies the evaluation of fairness metrics. The binary classification task allows for clear and interpretable assessments of ground truths across different subgroups, which are essential for evaluating fairness criteria such as demographic parity and equalized odds. Second, the German Credit Risk dataset is a readily available dataset of manageable size, allowing for detailed analysis without requiring extensive computational resources. Finally, since the German Credit Risk dataset is a real-world dataset, the findings of this research are more applicable to practical scenarios, are more credible and generalizable as they are based on real-world data, reflecting true complexities and biases present in financial decision-making.

5.1.1 Data Description

The German Credit Risk dataset[1] is a variation of the original German Credit Data [21] and contains 1000 entries with 3 numerical features and 7 categorical features. Each entry corresponds to an individual applying for credit from a bank. Each individual is categorized as either a 'good' or 'bad' credit risk based on a set of attributes. The attributes in this dataset are Age, Sex, Job, Housing, Saving accounts, Checking account, Credit amount, Duration, Purpose and the Label. The attributes Age, Credit amount and Duration are the numerical features.

5.1.2 Data Preprocessing

The data preprocessing involved numerous steps. Initially, the categorical labels 'good' and 'bad' were encoded as 1 and 0, respectively, to enable binary classification. Missing values

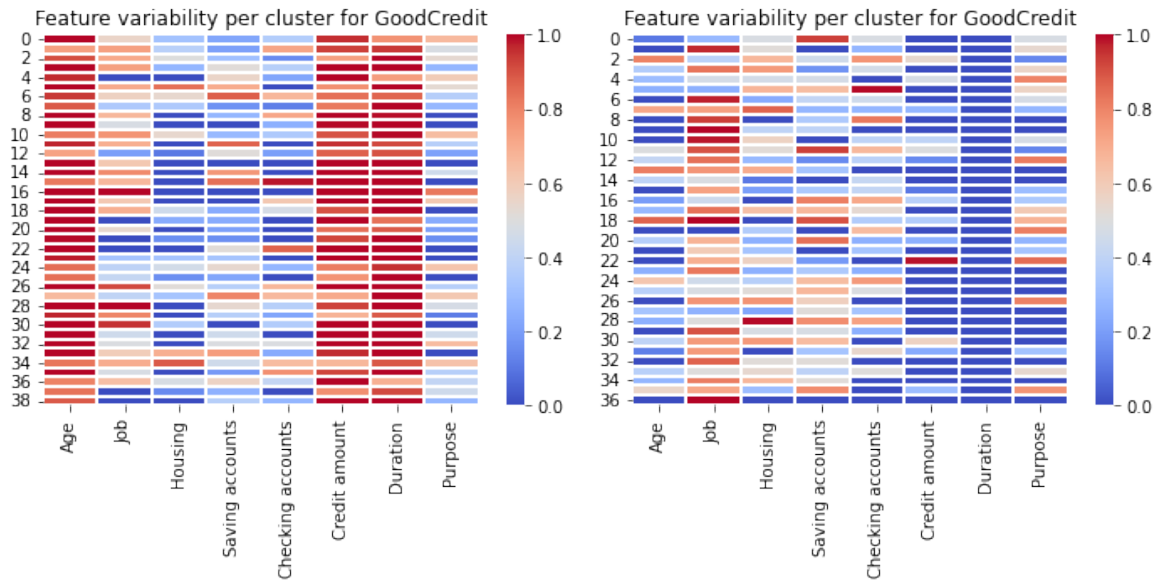
[1] https://www.kaggle.com/datasets/uciml/german-credit

were addressed by replacing all NaN entries with the label 'unknown', thereby preserving the completeness of the data. The dataset was then divided into three distinct subsets: BadCredit, including all instances labeled as 0; GoodCredit, including all instances labeled as 1 and AllCredit, containing all instances regardless of their label. Subsequently, target variable BadLabels, GoodLabels, and AllLabels were assigned to BadCredit, GoodCredit and AllCredit respectively, to ensure accurate labeling for analysis. BadLabels is a list of zeros with the same length as BadCredit, GoodLabels is a list of ones with the same length as GoodCredit, and AllLabels contains a list of zeros and ones, with a zero corresponding to an instance that also belongs in BadCredit and a one corresponding to an instance that also belongs in GoodCredit. This preprocessing approach ensures that the data is prepared for clustering algorithms and subgroup discovery techniques.

Additionally, based on the outcome of the preliminary experiment described below, we adjusted the data preprocessing steps as follows. Similar to the process described above, the categorical labels 'good' and 'bad' were encoded as 1 and 0, respectively, to enable binary classification. Missing values were addressed by replacing all NaN entries with the label 'unknown', thereby preserving the completeness of the data. The discretization of numerical values was done by using the Entropy based method[14]. This aligns with the approach of Lopez-Martinez-Carrasco et al. (2023) [34].

5.2 Preliminary Experiment: Numerical Descriptions

The VHK algorithm can handle both categorical and numerical features to find similar instances. Similarly, our description method can also use numerical features in the profile description. However, we are interested in whether the cluster instances are indeed similar with respect to the numerical values, or that our clustering algorithm finds categorical variables more important. If that is the case for our research, there won't be any descriptions containing numerical features, and therefore those features would not be interesting at all. If this happens, we can choose to discretize the numerical values in order to still keep those features in the data. To test this, we did a preliminary experiment. The experiment is based the VHK algorithm and our description method. We applied VHK clustering to our datasets BadCredit and GoodCredit separately, and described the clusters as discussed in section 4.2.2. When inspecting the profiles, we look for numerical features in the profile descriptions. Our conclusion was that there were none, for both the BadCredit data and the GoodCredit data. This meant that the numerical features were less relevant for the VHK clustering algorithm. In order to keep the numerical features in the data, we discretized them as described in the data preprocessing section above. This conclusion is also supported by the feature variability, as displayed in Figure 5.1a. Here you can see that the numerical features Age, Credit amount and Duration have high variability, causing the algorithm to not choose those features for similarity and description. In contrast, Figure 5.1b shows that the variability of the discretized numerical features show low variability, causing them to be picked to describe the profiles. Based on the results of this preliminary experiment, we have concluded that the current definition of numerical variability is not effective, and therefore will not be used.



(a) Mixed dataset

(b) Categorical dataset

Figure 5.1: Feature variability for different datasets, where (a) concerns the mixed dataset and (b) concerns the categorical dataset

5.3 Experiment 1: Description Quality Clustering

The first experiment is regarding the description quality to assess our description method and gain an understanding in how well instances within a cluster align with their corresponding profile descriptions. For this experiment, the VHK clustering algorithm is applied to the BadCredit and GoodCredit datasets to represent our PSD approach, and applied to the AllCredit dataset to represent a traditional clustering approach. For all datasets, the VHK algorithm is initialized with $k = 2$ clusters and $l = 25$ layers. We test the stability of the VHK clustering towards the initialization of the clusters by running the algorithm 100 times for different (random) initializations and compare each cluster. All unique cluster combinations are chosen for description. For each cluster combination, all clusters are described based on the variability difference score with a depth $d \in [1, 2, 3]$. For each cluster, the descriptions of all lengths d are obtained and the coverage and purity are calculated. We will compare these scores for BadCredit, GoodCredit and AllCredit.

5.4 Experiment 2: Description Comparison for SD

The second experiment is regarding the descriptions to assess the difference in descriptions between VHK, DFS and VLSD. This section describes how the descriptions are obtained from these three algorithms, specifying the parameter setup.

5.4.1 VHK

Subgroup discovery focuses on finding interested subgroups regarding a target variable, in this case the GoodCredit data instances. Therefore, the profile descriptions are based on the GoodCredit data. We take the same GoodCredit profile descriptions as obtained in Experiment 1.

5.4.2 Depth-First Search Algorithm (DFS)

The depth-first (DFS) search algorithm implementation is from PySubgroup[2][31]. The binary target variable is the risk of default, which is the positive label 'good', or 1 as previously indicated. Given our intent to investigate gender bias at a later stage, we opt to disregard 'Sex' as an attribute for this experiment. The quality function used to evaluate the discovered subgroups is the WRAcc score, which is defined as the StandardQF with an α of 1.0 in the python package from PySubgroup. The search strategy used is of course depth-first search and the DFS algorithm will output maximum 3871 subgroups, as this is the number of all possible subgroup combinations for our dataset. For evaluation, we are only interested in the top-30 subgroups based on the WRAcc score.

5.4.3 Vertical List Subgroup Discovery (VLSD)

The Vertical List Subgroup Discovery (VSLD) algorithm implementation is from the subgroups[3] python library from Lopez-Martinez-Carrasco et al. (2023)[34]. Similar to the DFS algorithm, the quality measure used is WRAcc score. The attribute 'Sex' is excluded to allow for an investigation of gender bias at a later stage. The search strategy employed is the VLSD algorithm with an optimistic estimate. VLSD also considers a quality threshold and an optimistic quality threshold. Since we are interested in the positive instances, we want the quality threshold to be zero. The optimistic quality estimate is set to 0.25, in aligning with Lopez-Martinez-Carrasco et al. (2023). The VLSD algorithm will output numerous subgroups, but for evaluation we are only interested in the top-30 subgroups based on the WRAcc score.

5.5 Experiment 3: Subgroup Quality Comparison for SD

Experiment 2 provides us with the descriptions from VHK, DFS and VLSD, which form the basis of this experiment. For each profile or subgroup description, we count the number of instances belonging to that subgroup description, and the number of positive and negative instances belonging to that subgroup description. Additionally, we count the number of instances in the whole dataset, including the number of positive and negative instances. These values are needed for the evaluation metrics described in section 3.3. Based on these values, the sensitivity, specificity, WRAcc score and Optimistic WRAcc can be calculated. For each description, these scores are stored in a dataframe. The dataframe is then sorted by WRAcc score, in order to find the top-30 most relevant subgroup descriptions.

5.6 Experiment 4: Subgroup Fairness: Individual Notions

To assess the subgroup fairness regarding the individual notions of fairness, we design the following experiment. Based on the profile and subgroup descriptions obtained from Experiment 2, we assess the individual fairness notions scores of each SD method for both the Logistic Regression (LR) and XGBoost classifier. We used K-Fold splitting with n_splits = 5, shuffle = True and random_state = 42. This means that the classifier is trained 5 times with a random split of 80% training data and 20% test data. The LR classifier is initialized with random_state = 42 and a maximum iteration of 10000. The XGBoost classifier is initialized with a max_depth of 5, gamma=1, and eval_metric is 'error'. For each of these 5 runs, we calculate the fairness

[2] https://github.com/flemmerich/pysubgroup
[3] https://github.com/antoniolopezmc/subgroups

metrics as described in section 3.4. At last, we take the average of these results to obtain more reliable results. This experiment is repeated for the profile descriptions of VHK, and the subgroup descriptions from DFS and VLSD.

5.7 Experiment 5: Subgroup Fairness: Group Notions

To assess the subgroup fairness regarding the group notions of fairness, we design the following experiment. Based on the profile and subgroup descriptions obtained from Experiment 2, we assess the group fairness notions scores of each SD method for both the Logistic Regression (LR) and XGBoost classifier. We used K-Fold splitting with n_splits = 5, shuffle = True and random_state = 42. This means that the classifier is trained 5 times with a random split of 80% training data and 20% test data. The LR classifier is initialized with random_state = 42 and a maximum iteration of 10000. The XGBoost classifier is initialized with a max_depth of 5, gamma=1, and eval_metric is 'error'. For each of these 5 runs, we calculate the fairness metrics as described in section 3.4. At last, we take the average of these results to obtain more reliable results. This experiment is repeated for the profile descriptions of VHK, and the subgroup descriptions from DFS and VLSD.

Chapter 6

Results

This chapter presents the findings of our research. We begin by comparing the quality of descriptions produced by our method versus traditional clustering techniques. Following this, we examine the differences in descriptions generated by VHK, DFS, and VLSD. We then assess the subgroup quality across these three methods, with a focus on their overall effectiveness. Additionally, we evaluate the individual fairness notions for each method, analyzing results separately for logistic regression and XGBoost. Finally, we explore the group fairness notions for all three methods, once again dividing the analysis between logistic regression and XGBoost outcomes.

6.1 Experiment 1: Description Quality

This section describes the results regarding the description quality in terms of coverage and purity. Table 6.1 shows the average coverage and purity for the profile descriptions, alongside the standard deviation. The results are displayed for our method (PSD) and for traditional clustering (VHK), where traditional clustering refers to not partitioning the data based on the target variable before applying our method, as shown in figure 4.1. Regarding the coverage, the average coverage of the descriptions increase with description depth d, for both methods. For a description depth of 1, our method has a lower coverage on average than traditional clustering. In contrast, for descriptions with depth two or three, our method has a higher coverage on average than traditional clustering. Independently of the depth of the description, the average coverage of our method ranges between 5.5% and 8.5%, and the average coverage of traditional clustering ranges from 6.0% and 7.4%. Regarding the purity metric, as the depth increases, the purity decreases for all methods. For a depth of one, the descriptions include almost all instances of the cluster, with 99.9% and 99.8% for BadCredit and GoodCredit, respectively. For a depth of three, the descriptions cover around 97% - 98% of the cluster instances on average. These values are higher for our method than for the traditional clustering method, with traditional clustering achieving 99.6% purity on average for a depth of 1, and a 97.3% on average for a depth of 3. Based on these results, we can conclude that there is no significant difference in performance by splitting the dataset based on the target variable.

	Coverage			Purity		
	PSD		VHK	PSD		VHK
depth	BadCredit	GoodCredit	AllCredit	BadCredit	GoodCredit	AllCredit
1	0.0561 ±0.05	0.0553 ±0.03	**0.0601 ±0.03**	0.9995 ±0.01	**0.9985 ±0.01**	0.9965 ±0.01
2	0.0678 ±0.06	**0.0727 ±0.05**	0.0644 ±0.04	**0.9981 ±0.01**	0.9947 ±0.03	0.9915 ±0.03
3	0.0788 ±0.07	**0.0848 ±0.06**	0.0739 ±0.05	**0.9873 ±0.04**	0.9725 ±0.07	0.9727 ±0.06

Table 6.1: Description Quality for Clustering as SD for the average coverage ±std and average purity ±std

6.2 Experiment 2: Description Comparison for SD

This section analyzes the descriptions of the subgroups identified by the different subgroup discovery (SD) techniques. We focus on descriptions associated with individuals in the GoodCredit data (PSD+), representing our target population of interest with good creditworthiness. We employ the WRAcc score to evaluate the ranking of the descriptions.

We compared the top-30 subgroup descriptions (based on WRAcc score) generated by PSD+, DFS, and VLSD algorithms. Figure 6.1 illustrates the overlap between these methods. PSD+ and DFS had 8 out of 30 descriptions that overlapped (∼ 27%). PSD+ and VLSD exhibited the least overlap, with only 7 out of 30 descriptions (around 23%) shared. VLSD had the most overlap with DFS (24 out of 30 descriptions, or 80%).

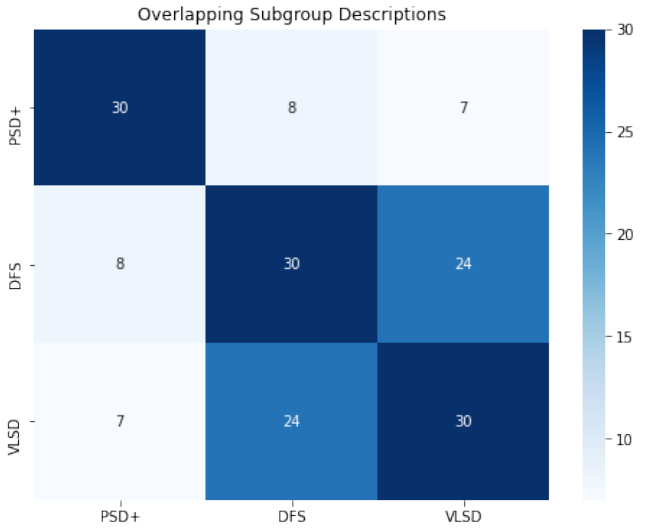


Figure 6.1: The number of overlapping subgroup descriptions between PSD+, DFS and VLSD

Regarding the ranking of descriptions by WRAcc score (Table 6.2), the top description was identical for all three algorithms. Notably, all the top-15 highest scoring descriptions were identified by VLSD and DFS, while PSD+ only found three within this group. The second and third best descriptions for PSD+ were ranked 14th and 15th by VLSD, and 13th and 14th by DFS, respectively. Additionally, the 15th ranking description for PSD+, is equal to the 54th ranking description of DFS, and 79th ranking description for VLSD. The results of this experiment highlight a key distinction between PSD and SD: PSD does not rely on a quality measure for finding subgroups, resulting in less similarity with the subgroups identified by DFS and VLSD. Furthermore, this also leads to less high-quality profile descriptions, since we are not optimizing for this.

no.	description	PSD+ rank	DFS rank	VLSD rank	WRAcc score
0	{'Checking account': 'unknown'}	1	1	1	0.0722
1	{'Age': '>=25.5', 'Checking account': 'unknown'}	-	2	2	0.0672
2	{'Checking account': 'unknown', 'Credit amount': '<3913.5'}	-	3	3	0.0653
3	{'Age': '>=25.5', 'Checking account': 'unknown', 'Credit amount': '<3913.5'}	-	4	4	0.0603
4	{'Housing': 'own', 'Checking account': 'unknown'}	-	5	5	0.0592
5	{'Housing': 'own', 'Age': '>=25.5', 'Checking account': 'unknown'}	-	6	6	0.0554
6	{'Housing': 'own', 'Checking account': 'unknown', 'Credit amount': '<3913.5'}	-	7	7	0.0531
7	{'Job': 2, 'Checking account': 'unknown'}	-	8	8	0.0528
8	{'Housing': 'own', 'Age': '>=25.5', 'Credit amount': '<3913.5'}	-	9	9	0.0489
9	{'Job': 2, 'Age': '>=25.5', 'Checking account': 'unknown'}	-	10	10	0.0482
10	{'Job': 2, 'Checking account': 'unknown', 'Credit amount': '<3913.5'}	-	11	11	0.048
11	{'Age': '>=25.5', 'Credit amount': '<3913.5'}	-	12	12	0.046
12	{'Duration': '<15.5', 'Age': '>=25.5', 'Credit amount': '<3913.5'}	2	13	13	0.0453
13	{'Housing': 'own', 'Credit amount': '<3913.5'}	-	14	14	0.0442
14	{'Duration': '<15.5', 'Credit amount': '<3913.5'}	3	15	15	0.0442
15	{'Duration': '<15.5', 'Checking account': 'unknown'}	4	17	18	0.042
16	{'Duration': '<15.5', 'Checking account': 'unknown', 'Credit amount': '<3913.5'}	5	18	19	0.0413
17	{'Duration': '<15.5'}	6	20	22	0.0403
18	{'Duration': '<15.5', 'Housing': 'own', 'Credit amount': '<3913.5'}	7	22	26	0.0391
19	{'Duration': '<15.5', 'Housing': 'own'}	8	25	32	0.0362
20	{'Duration': '<15.5', 'Housing': 'own', 'Checking account': 'unknown'}	9	33	41	0.031
21	{'Duration': '>=15.5', 'Checking account': 'unknown'}	10	34	42	0.0302
22	{'Duration': '<15.5', 'Job': 2, 'Checking account': 'unknown'}	11	35	44	0.0298
23	{'Duration': '>=15.5', 'Age': '>=25.5', 'Checking account': 'unknown'}	12	36	45	0.0297
24	{'Duration': '<15.5', 'Job': 2, 'Credit amount': '<3913.5'}	13	46	60	0.0265
25	{'Duration': '<15.5', 'Job': 2}	14	48	64	0.0261
26	{'Duration': '>=15.5', 'Checking account': 'unknown', 'Credit amount': '<3913.5'}	15	54	79	0.024

Table 6.2: Ranking of descriptions by different algorithms

6.3 Experiment 3: Subgroup Quality Comparison for SD

Our evaluation focused on four key subgroup discovery (SD) metrics: SD Sensitivity, SD Specificity, SD Weighted Relative Accuracy (WRAcc), and Optimistic WRAcc. These metrics assess the effectiveness of the discovered subgroups in capturing positive and negative instances (creditworthy applicants). Again, the top-30 descriptions are used, with respect to the WRAcc score. The results can be found in Figure 6.2.

The SD Sensitivity indicates the proportion of positive instances covered by the subgroup descriptions. VLSD achieved the highest average SD Sensitivity, suggesting its descriptions generally encompass a larger portion of creditworthy applicants. While PSD+ had a maximum sensitivity of 49.71% for the "unknown checking account" subgroup, its overall average was lower than VLSD and DFS. Furthermore, this description is also included in the DFS and VLSD descriptions. Notably, DFS exhibited the highest average sensitivity (0.4029). The SD Specificity reflects the proportion of negative instances (non-creditworthy applicants) that fall outside the identified subgroups. Here, the average scores were closer for all algorithms. However, DFS displayed the greatest variation and the lowest average specificity (0.8154). Conversely, PSD+ achieved the highest average SD Specificity, with an average score of 0.8930, implying its subgroups tend to exclude most negative instances in general. The SD WRAcc score addresses the trade-off between discovering large subgroups and ensuring they are highly relevant to the target property. The average SD WRAcc score is the highest for VLSD (0.0475), followed by DFS and PSD+ with a SD WRAcc score of 0.0471 and 0.0459, respectively. The PSD+ algorithm varies between 0.0142 and 0.0722, with half of the WRAcc scores between 0.0235 and 0.0349 and an average of 0.0283. This is significantly lower than the DFS and VLSD algorithms. Lastly, the Optimistic WRAcc score. DFS achieved the highest average Optimistic WRAcc (71.1119), followed by VLSD (65.8182) and then PSD+ (42.8713). The results demonstrate that PSD+ shows lower performance in SD sensitivity and overall SD (optimistic) WRAcc compared to VLSD and DFS. However, PSD+ excels in achieving high SD specificity,

effectively excluding non-creditworthy applicants from identified subgroups.

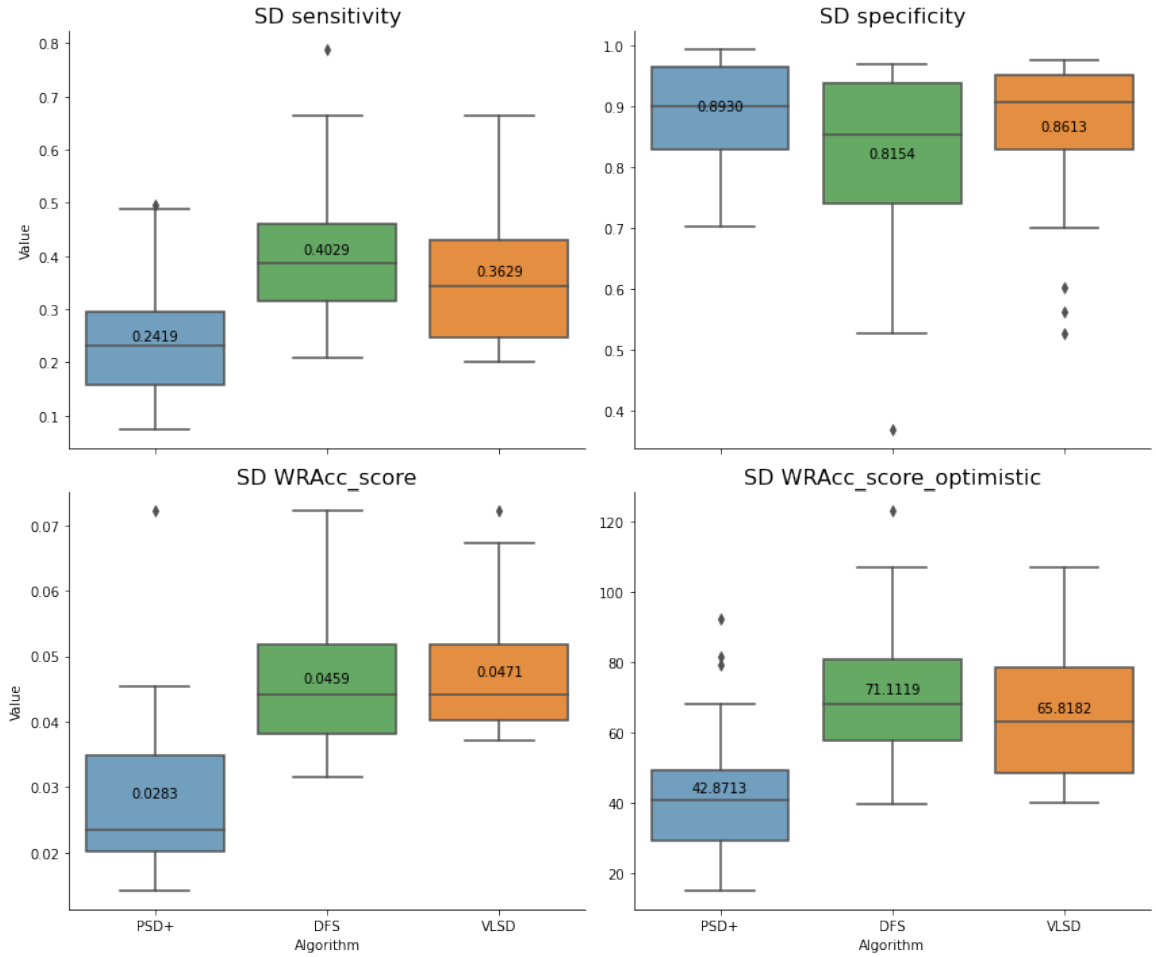


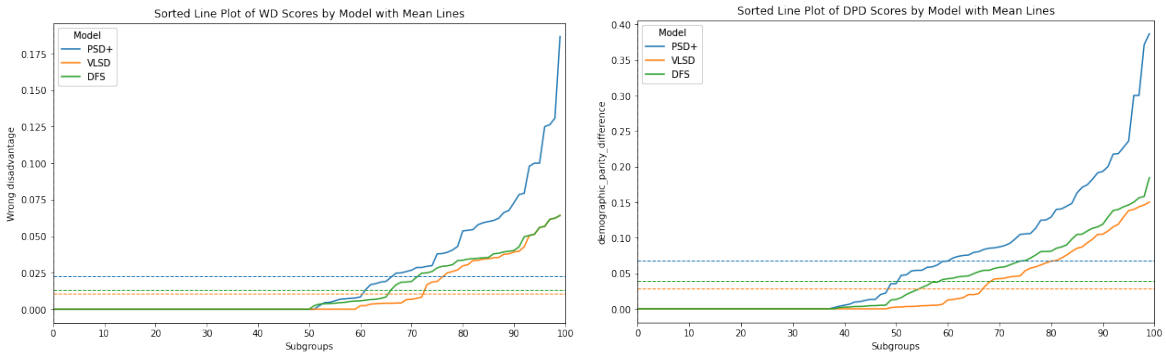
Figure 6.2: Subgroup quality metrics comparison for PSD+, DFS and VLSD

6.4 Experiment 4: Subgroup Fairness: Individual Notions

This experiment investigates the capability of subgroup discovery techniques to identify unfair subgroups, in terms of gender bias. For the internal subgroup fairness we use demographic parity difference (DPD), equalized odds difference (EOD) and wrong disadvantage (WD). This experiment is applied to the logistic regression and XGBoost classifiers, we will make a distinction between those when discussing the results.

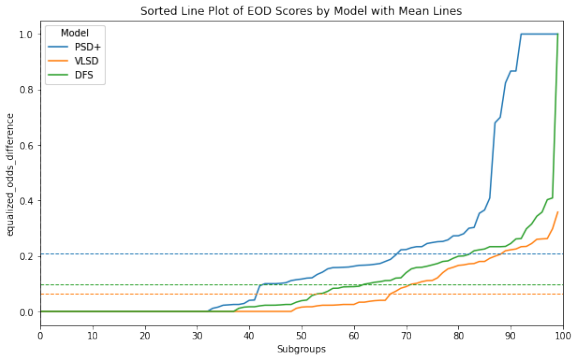
6.4.1 Logistic Regression (LR)

Figure 6.3 presents the sorted line plots of the Wrong Disadvantage (WD), Demographic Parity Difference (DPD), and Equalized Odds Difference (EOD) metrics across the three models: VLSD, DFS, and PSD+ for the LR. A consistent trend emerges across all three metrics, with VLSD showing the least amount of subgroups with a score above zero, indicating better internal fairness, while PSD+ exhibits the most amount of subgroups above zero and the widest range of internal fairness scores. Table 6.3 complements these findings by detailing the minimum, maximum, and mean values of each metric, along with the frequency of occurrences for the minimum and maximum values. These metrics are crucial for understanding how many



(a) WD

(b) DPD



(c) EOD

Figure 6.3: Internal Subgroup Fairness measures for LR

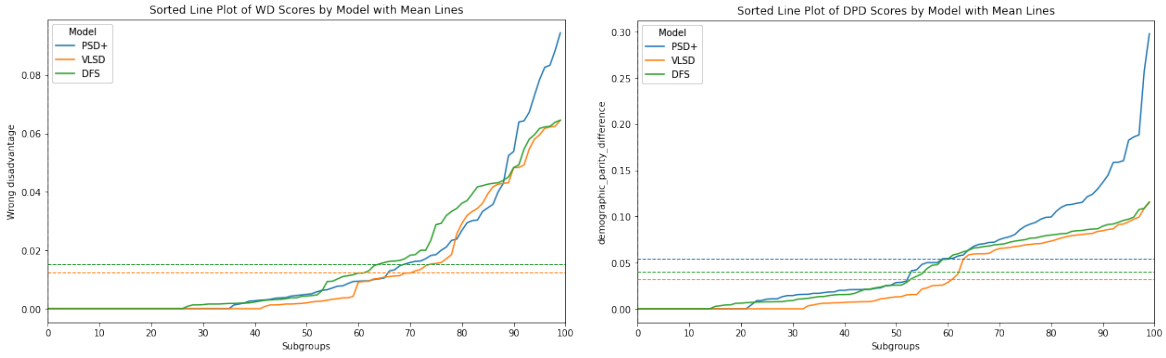
identified subgroups achieve demographic parity, equalized odds, and a zero score for WD. For WD scores, VLSD has the majority of values close to zero, with a maximum WD of 0.0642. PSD+, on the other hand, has fewer values near zero and a maximum WD of 0.1867. DFS lies between these two, with a maximum WD of 0.0642 and 51% of its descriptions achieving a WD of 0. In comparison, PSD+ achieves a WD of 0 in 52% of cases, while VLSD reaches this in 60% of cases. Regarding DPD scores, 38% of PSD+ profiles have a DPD of 0.0, compared to 39% for DFS and 49% for VLSD. The maximum DPD scores for PSD+, DFS, and VLSD are 0.3867, 0.1844, and 0.15, respectively. The average DPD scores are 0.0674 for PSD+, 0.0395 for DFS, and 0.0291 for VLSD. In terms of EOD scores, both PSD+ and DFS identified subgroups with EOD values ranging from 0 to 1, indicating significant variability. In contrast, VLSD's EOD values range from 0.0 to 0.3582, indicating better performance in achieving equalized odds. Nearly half of the VLSD subgroups (49%) have an EOD of 0.0, whereas PSD+ and DFS achieve this in 33% and 38% of cases, respectively. PSD+ finds 8 subgroup descriptions with an EOD of 1.0, whereas DFS finds 1 and VLSD 0. Although VLSD finds the most amount of subgroups that are treated fairly according to our fairness metrics, PSD+ identifies most subgroup descriptions that exhibit gender bias.

6.4.2 XGBoost Classifier

Figure 6.4 presents the sorted line plots of the Wrong Disadvantage (WD), Demographic Parity Difference (DPD), and Equalized Odds Difference (EOD) metrics across the three models: VLSD, DFS, and PSD+ for the LR. A consistent trend emerges across all three metrics, with VLSD showing the least amount of subgroups with a score above zero, indicating better internal fairness, while PSD+ exhibits the most amount of subgroups above zero and the widest range of internal fairness scores.

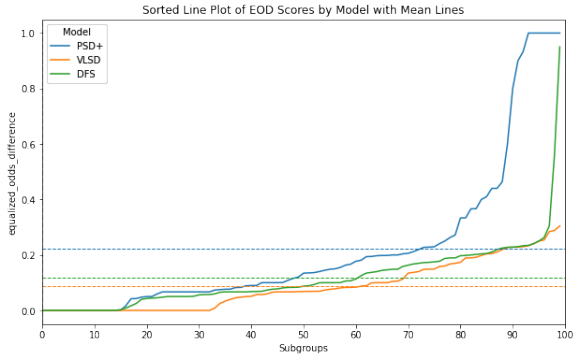
		Models		
Metric	Value	PSD+	DFS	VLSD
Wrong disadvantage	min (occurence)	0.0 (52)	0.0 (51)	**0.0 (60)**
	mean	0.0227	0.0131	**0.0108**
	max (occurence)	0.1876 (1)	**0.0642 (1)**	**0.0642 (1)**
Demographic parity difference	min (occurence)	0.0 (38)	0.0 (39)	**0.0 (49)**
	mean	0.0674	0.0395	**0.0291**
	max (occurence)	0.3867 (1)	0.1844 (1)	**0.15 (1)**
Equalized odds difference	min (occurence)	0.0 (33)	0.0 (38)	**0.0 (49)**
	mean	0.2094	0.0975	**0.0633**
	max (occurence)	1.0 (8)	1.0 (1)	**0.3582 (1)**

Table 6.3: Range of internal subgroup fairness metrics for three different models (LR)



(a) WD

(b) DPD



(c) EOD

Figure 6.4: Internal Subgroup Fairness measures for XGBoost

Table 6.4 complements these findings by detailing the minimum, maximum, and mean values of each metric, along with the frequency of occurrences for the minimum and maximum values. These metrics are crucial for understanding how many identified subgroups achieve demographic parity, equalized odds, and a zero score for WD. For WD scores, VLSD has the majority of values close to zero, with a maximum WD of 0.0645. PSD+, on the other hand, has fewer values near zero and a maximum WD of 0.0944. DFS lies between these two, with a maximum WD of 0.0645 and 27% of its descriptions achieving a WD of 0. In comparison, PSD+ achieves a WD of 0 in 36% of cases, while VLSD reaches this in 42% of cases. Regarding DPD scores, 22% of PSD+ profiles have a DPD of 0.0, compared to 15% for DFS and 33% for VLSD. The maximum DPD scores for PSD+, DFS, and VLSD are 0.2977, 0.1155, and 0.1155, respectively. The average DPD scores are 0.0541 for PSD+, 0.0400 for DFS, and 0.0325 for VLSD. In terms of EOD scores, the PSD+ algorithm identified subgroups with EOD values

ranging from 0 to 1, indicating significant variability. Similar results are found for DFS, with EOD values ranging from 0 to 0.95. In contrast, VLSD's EOD values range from 0.0 to 0.3049, indicating better performance in achieving equalized odds. The VLSD finds around twice as many subgroups (33%) with an EOD of 0.0 in comparison to PSD+ and DFS, whereas PSD+ and DFS achieve this in 16% and 15% of cases, respectively. PSD+ finds 1 subgroup description with an EOD of 1.0, whereas DFS and VLSD find none. Although VLSD again finds the most amount of subgroups that are treated fairly according to our fairness metrics, PSD+ identifies most subgroup descriptions that exhibit gender bias.

		Models		
Metric	Value	PSD+	DFS	VLSD
Wrong disadvantage	min (occurence)	0.0 (36)	0.0 (27)	**0.0 (42)**
	mean	0.01523	0.01540	**0.0125**
	max (occurence)	0.0944 (1)	**0.0645 (1)**	**0.0645 (1)**
Demographic parity difference	min (occurence)	0.0 (22)	0.0 (15)	**0.0 (33)**
	mean	0.0541	0.0400	**0.0325**
	max (occurence)	0.2977 (1)	**0.1155 (1)**	**0.1155 (1)**
Equalized odds difference	min (occurence)	0.0 (16)	0.0 (15)	**0.0 (33)**
	mean	0.2210	0.1169	**0.0866**
	max (occurence)	1.0 (1)	0.95 (1)	**0.3049 (1)**

Table 6.4: Range of internal subgroup fairness metrics for three different models (XGBoost)

6.5 Experiment 5: Subgroup Fairness: Group Notions

This experiment investigates the capability of subgroup discovery techniques to identify unfair treatment of subgroups. For the external subgroup fairness we use the weighted DPD Subgroup Fairness, weighted EOD Subgroup Fairness, weighted TPD Subgroup Fairness and weighted FPD Subgroup Fairness. This experiment is applied to the logistic regression and XGBoost classifiers, we will make a distinction between those when discussing the results.

6.5.1 Logistic Regression (LR)

Figure 6.5 presents sorted line plots for the weighted DPD, EOD, TPD, and FPD Subgroup Fairness metrics across three models: VLSD, DFS, and PSD+. A consistent trend is observed in both the weighted DPD and EOD Subgroup Fairness metrics, where the PSD+ model generally shows lower values compared to DFS and VLSD, indicating better external fairness. Similarly, for the weighted TPD and FPD Subgroup Fairness metrics, the PSD+ model tends to achieve scores closer to zero, further indicating superior external fairness.

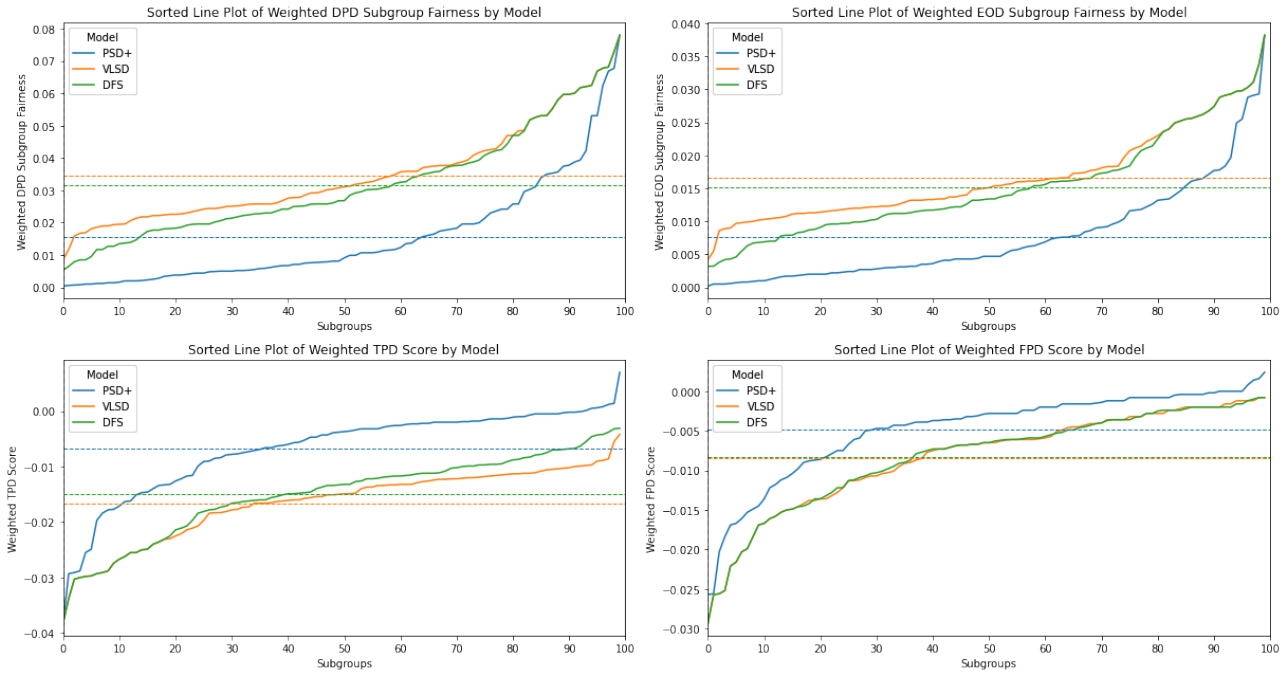


Figure 6.5: External Subgroup Fairness measures for LR

Table 6.5 complements these findings by detailing the minimum, maximum, and mean values of each metric, along with the frequency of occurrences for the minimum and maximum values. Notably, PSD+ consistently exhibits lower values for weighted Demographic Parity Difference (DPD) and weighted Equalized Odds Difference (EOD) compared to DFS and VLSD. This trend is evident in Figure 6.5 and corroborated by Table 6.5, where PSD+ frequently achieves the best metric values. In terms of weighted DPD subgroup fairness, PSD+ shows significantly lower average scores than DFS and VLSD. The worst recorded weighted DPD score is 0.078, identified as an overlapping subgroup by all three algorithms, with VLSD exhibiting the highest scores overall. For weighted EOD subgroup fairness scores, PSD+ again demonstrates the lowest average scores at 0.0076, followed by DFS at 0.0151 and VLSD at 0.0166. An overlapping subgroup among all three algorithms achieves the maximum weighted EOD score of 0.0382. Considering weighted True Positive Difference (TPD) and weighted False Positive Difference (FPD), nearly all values are negative, indicating a higher positive rate within the subgroups compared to the entire dataset. The minimum weighted TPD subgroup fairness score observed is -0.0382, also identified as an overlapping subgroup. Similar to weighted DPD, PSD+ shows average scores closer to zero for weighted TPD subgroup fairness compared to DFS and VLSD. Regarding weighted FPD subgroup fairness, PSD+ achieves the closest average score to zero at -0.0048, indicating minimal difference on average between the false positive rate within the subgroup and the dataset. The maximum value that is closest to zero is -0.0008, which is shared by subgroups identified by both DFS and VLSD algorithms. These results indicate that PSD+ finds subgroups most similar to the data. For identifying subgroups that are treated unfairly, VLSD and DFS seem to be a better fit, as these score worse on the fairness metrics.

		Models		
Metric	Value	PSD+	DFS	VLSD
weighted DPD subgroup fairness	min (occurence)	**0.0004 (1)**	0.0053 (1)	0.0085 (1)
	mean	**0.0154**	0.0317	0.0346
	max (occurence)	**0.078 (1)**	**0.078 (1)**	**0.078 (1)**
weighted EOD subgroup fairness	min (occurence)	**0.0002 (1)**	0.0032 (1)	0.0042 (1)
	mean	**0.0076**	0.0151	0.0166
	max (occurence)	**0.0382 (1)**	**0.0382 (1)**	**0.0382 (1)**
weighted TPD subgroup fairness	min (occurence)	**-0.0382 (1)**	**-0.0382 (1)**	**-0.0382 (1)**
	mean	**-0.0068**	-0.0150	-0.0165
	max (occurence)	**0.007 (1)**	-0.0031 (1)	-0.0042 (1)
weighted FPD subgroup fairness	min (occurence)	**-0.0257 (1)**	-0.0293 (1)	-0.0293 (1)
	mean	**-0.0048**	-0.0083	-0.0083
	max (occurence)	0.0024	**-0.0008 (2)**	**-0.0008 (2)**

Table 6.5: Range of external subgroup fairness metrics for three different models (LR)

6.5.2 XGBoost Classifier

Figure 6.6 presents sorted line plots for the weighted DPD, EOD, TPD, and FPD Subgroup Fairness metrics across three models: VLSD, DFS, and PSD+. A consistent trend is observed in both the weighted DPD and EOD Subgroup Fairness metrics, where the PSD+ model generally shows lower values compared to DFS and VLSD, indicating better external fairness. Similarly, for the weighted TPD and FPD Subgroup Fairness metrics, the PSD+ model tends to achieve scores closer to zero, further indicating superior external fairness.

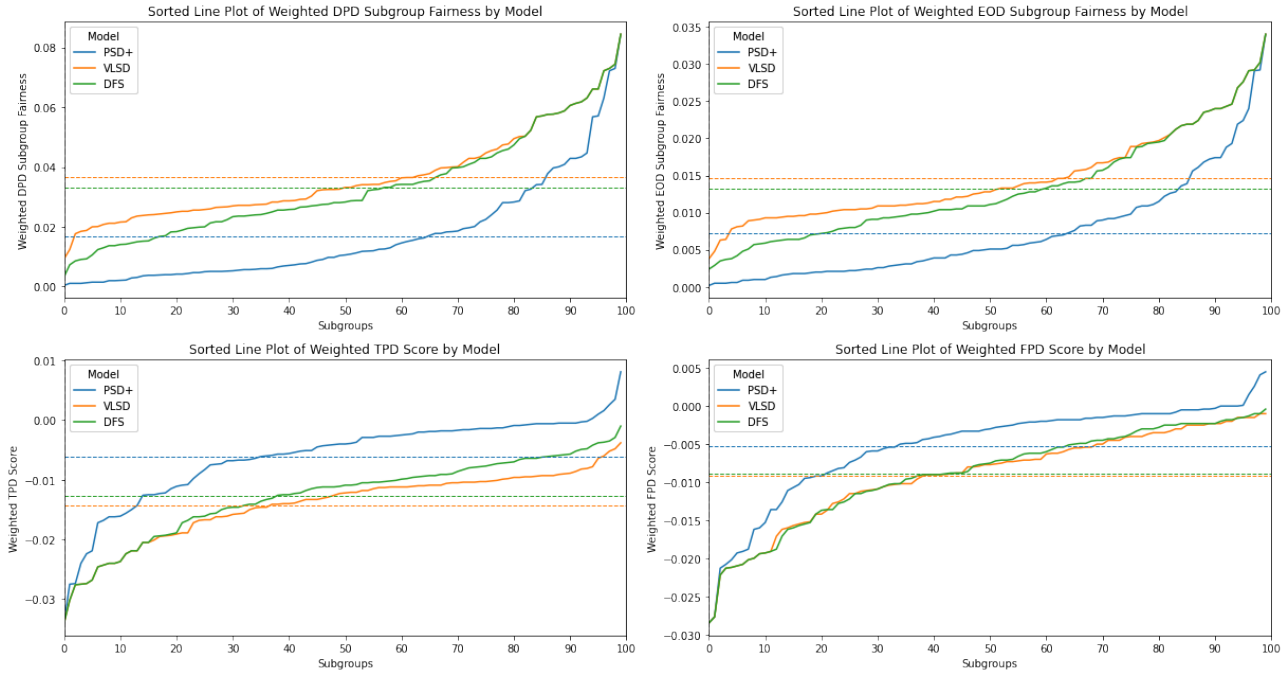


Figure 6.6: External Subgroup Fairness measures for XGBoost

Table 6.6 complements these findings by detailing the minimum, maximum, and mean values of each metric, along with the frequency of occurrences for the minimum and maximum values. Notably, PSD+ consistently exhibits lower values for weighted Demographic Parity Difference (DPD) and weighted Equalized Odds Difference (EOD) compared to DFS and VLSD. This

trend is evident in Figure 6.6 and corroborated by Table 6.6, where PSD+ frequently achieves the best metric values. In terms of weighted DPD subgroup fairness, PSD+ shows significantly lower average scores than DFS and VLSD. The worst recorded weighted DPD score is 0.0844, identified as an overlapping subgroup by all three algorithms, with VLSD exhibiting the highest scores overall. For weighted EOD subgroup fairness scores, PSD+ again demonstrates the lowest average scores at 0.0072, followed by DFS at 0.0131 and VLSD at 0.0146. An overlapping subgroup among all three algorithms achieves the maximum weighted EOD score of 0.034. Considering weighted True Positive Difference (TPD) and weighted False Positive Difference (FPD), nearly all values are negative, indicating a higher positive rate within the subgroups compared to the entire dataset. The minimum weighted TPD subgroup fairness score observed is -0.034, also identified as an overlapping subgroup. Similar to weighted DPD, PSD+ shows average scores closer to zero for weighted TPD subgroup fairness compared to DFS and VLSD. Regarding weighted FPD subgroup fairness, PSD+ achieves the closest average score to zero at -0.0053, indicating minimal difference on average between the false positive rate within the subgroup and the dataset. The maximum value that is closest to zero is -0.0004, which is a subgroup identified by DFS. These results indicate that PSD+ finds subgroups most similar to the data. For identifying subgroups that are treated unfairly, VLSD and DFS seem to be a better fit, as these score worse on the fairness metrics.

		Models		
Metric	Value	PSD+	DFS	VLSD
weighted DPD subgroup fairness	min (occurence)	**0.0004 (1)**	0.0033 (1)	0.0093 (1)
	mean	**0.0168**	0.0330	0.0365
	max (occurence)	**0.0844 (1)**	**0.0844 (1)**	**0.0844 (1)**
weighted EOD subgroup fairness	min (occurence)	**0.0002 (1)**	0.0024 (1)	0.0038 (1)
	mean	**0.0072**	0.0131	0.0146
	max (occurence)	**0.034 (1)**	**0.034 (1)**	**0.034 (1)**
weighted TPD subgroup fairness	min (occurence)	**-0.034 (1)**	**-0.034 (1)**	**-0.034 (1)**
	mean	**-0.0061**	-0.0127	-0.0165
	max (occurence)	0.0081 (1)	**-0.001 (1)**	-0.0038 (1)
weighted FPD subgroup fairness	min (occurence)	**-0.0285 (1)**	**-0.0285 (1)**	**-0.0285 (1)**
	mean	**-0.0053**	-0.0089	-0.0091
	max (occurence)	0.0045	**-0.0004 (1)**	-0.001 (2)

Table 6.6: Range of external subgroup fairness metrics for three different models (XGBoost)

Chapter 7

Discussion

This chapter provides a comprehensive overview of the key findings from this research, offering an in-depth interpretation of the results, exploring broader implications, and critically examining limitations. Recommendations for future research are also presented to guide subsequent investigations in this field.

This study introduces a novel clustering method called Profile-based Subgroup Discovery (PSD), designed to generate simple and interpretable clusters for subgroup discovery. Building on previous semi-hierarchical methods for profile extraction, PSD aims to enhance the clarity and utility of subgroup analysis. We evaluated the performance of PSD across three key dimensions: profile description alignment, subgroup description quality, and subgroup fairness. First, we assessed how well instances within a cluster align with their corresponding profile descriptions, ensuring the clarity and coherence of the generated profiles. Second, we compared the ability of PSD to identify high-quality profiles against other subgroup discovery techniques, highlighting its effectiveness and areas for improvement. Third, we examined PSD's capability to identify subgroups that are treated unfairly, with a specific focus on gender bias in the context of credit scoring. Additionally, the study analyzed the differences in fairness metrics across subgroups identified by PSD and traditional subgroup discovery methods, in comparison to the overall dataset.

Before the main experiments, a preliminary experiment was conducted to investigate the numerical variability metric and its potential influence on the profile descriptions. The results revealed that the numerical variability metric used in the VHK algorithm from Wilms et al. (2022)[41] predominantly identified high variability scores for numerical features, making categorical features appear more relevant. Consequently, it was concluded that the current definition of numerical variability is ineffective, leading us to exclude it from our study. This decision restricted our research to categorical data instead of mixed datasets. Although numerical features were discretized, this process resulted in information loss. In the credit scoring domain, specific numerical values, such as the amount of credit or savings, are critical. Therefore, future research should explore alternative numerical variability metrics or different approaches for handling numerical data.

Regarding profile descriptions, we expected them to highly align with the instances in the cluster, as they were obtained through variability difference. More specifically, we anticipated the profile descriptions from PSD to be more specific and align better than traditional clustering with VHK, given that PSD takes the target variable into account, limiting the variability in descriptions. However, our results indicate no significant difference in performance by splitting the dataset based on the target variable. Although PSD obtained higher coverage and

purity scores than VHK, these results were not significantly better. One possible explanation for this finding could be that the metrics used to evaluate the quality of profile descriptions (coverage and purity) might not fully capture the benefits of using the target variable in clustering. Alternative metrics such as support, as in Herrera (2011)[20], or a more comprehensive evaluation framework might be needed to detect differences in future research. Another explanation could be that the dataset used may not have strong inherent relationships between the target variable and the features used for clustering. If the target variable does not significantly influence the variability in the data, partitioning based on the target variable might not yield better-aligned profile descriptions. Further investigation regarding the relationships in the data is needed to confirm this explanation. Another important limitation concerns the definition of variability difference used in describing the clusters. We described the clusters using variability difference based on the most relevant features, but this approach may not always capture the most significant features, as the definition of relevance is subjective. Our approach focuses on the variability within a cluster compared to the entire dataset, potentially overlooking the importance of features for distinguishing between positive and negative classes. This limitation is reflected in our results, where our method performed significantly worse in terms of the SD WRAcc score. To address this, future research could consider incorporating SHAP values, as utilized by Cooper et al. (2021)[11], which provide a more nuanced understanding of feature relevance and do not treat all variables equally.

In terms of subgroup quality, we expected PSD to produce lower quality subgroups than SD techniques, as we do not optimize for this while SD techniques do. We also expected less overlapping subgroup descriptions, since high-quality subgroups are more likely to be found when specifically searching for them. Our results confirm this hypothesis, with PSD having the least overlapping top-30 descriptions compared to DFS and VLSD. The evaluation of the ranking using the SD WRAcc score revealed that PSD indeed identified fewer high-quality subgroups compared to DFS and VLSD. Furthermore, PSD exhibited lower SD sensitivity in identifying creditworthy applicants compared to the other algorithms, indicating a stricter definition of subgroup profiles. This strictness may lead to the exclusion of potentially creditworthy applicants whose characteristics do not align closely with PSD's narrow profile descriptions. Conversely, PSD excelled in achieving the highest average SD specificity, indicating its strength in identifying distinct subpopulations of likely non-creditworthy applicants, which is crucial in credit scoring. This trade-off suggests that the inherent design of PSD, which includes partitioning based on the target variable and focusing on variability differences, might lead to a natural trade-off where the method excels in excluding negatives but at the cost of missing some positives. Future research should investigate other quality metrics and these trade-offs, as different contexts may require more or less specified subgroups.

Regarding subgroup fairness, specifically the individual notions, we expected PSD to identify more subgroups that are treated unfairly. Since the descriptions are based on similarity within subgroups, we anticipated finding more diverse subgroups and therefore more subgroups that are treated unfairly. Our results found that PSD indeed identified more subgroups containing gender bias, with a higher demographic parity difference and equalized odds difference. These results were consistent for both the logistic regression classifier and the XGBoost classifier. Although we hypothesize that this is due to more diverse subgroups, we cannot confirm this without a qualitative analysis of the subgroup descriptions. This limitation means we found more gender bias in the subgroups from PSD but cannot explain why. Additionally, the fairness metrics used are common but do not provide a complete picture of fairness. For example, Zhang et al. (2021)[43] also used the false discovery rate, false omission rate, and error rate to investigate gender bias. Future research should focus on a more comprehensive

qualitative analysis of fairness metrics and the qualitative differences in subgroup descriptions, paying particular attention to the relationships between 'Sex' and other features in the data, which might influence fairness scores.
Regarding the group fairness metrics used to evaluate the subgroups, we expected PSD to be more fair in terms of aligning with the data, as the descriptions are obtained directly from the data rather than by combining features until a satisfactory quality threshold is reached, as in subgroup discovery. The results confirm this hypothesis, with PSD fitting the data best across all four subgroup fairness metrics employed. This implies that PSD is less effective at identifying subgroups that are treated unfairly compared to VLSD and DFS. Despite its alignment with data and fairness, PSD's strength in fairness metrics implies a limitation in identifying unfairly treated subgroups. Traditional SD methods like VLSD and DFS, which are optimized to find subgroups with high-quality thresholds, may be better suited for detecting instances of unfair treatment. For applications where fairness and alignment with the data distribution are critical, PSD might be a more suitable method. However, for tasks focused on uncovering and addressing unfair treatment or biases, traditional SD methods like VLSD and DFS might be more effective. The findings suggest that while PSD is advantageous for maintaining fairness and data alignment, there may be a need to refine the method to better identify unfairly treated subgroups. Future research could explore ways to enhance PSD's ability to detect bias while preserving its strengths in fairness and data representation.

Chapter 8

Conclusion

This research has explored the critical intersection of subgroup discovery, clustering methodologies, and fairness considerations in machine learning, particularly within the domain of credit scoring. The overarching goal was to introduce and evaluate Profile-based Subgroup Discovery (PSD) as a novel methodological approach that applies target variable-dependent clustering for subgroup discovery, with a specific focus on analyzing subgroup fairness in AI.

We proposed a two-step approach: (1) adapting the clustering pipeline to incorporate the target variable, and (2) developing a method to interpret and describe the resulting clusters, referred to as profile descriptions. The quality of these profile descriptions generated by PSD was compared with those produced by traditional clustering techniques. Additionally, we evaluated the meaningfulness of subgroup descriptions generated through PSD versus other Subgroup Discovery (SD) techniques. We examined differences in subgroup fairness metrics between PSD and other SD methods, with a particular focus on gender bias in credit scoring. Furthermore, we compared the subgroup fairness metrics for subgroups identified by PSD with those identified by alternative techniques.

The main takeaways of this research are as follows: The numerical variability metric used by the underlying VHK algorithm predominantly identified high variability scores for numerical features, making categorical features appear more relevant. Consequently, the current definition of numerical variability was deemed ineffective, leading the study to focus on categorical data. While PSD aimed for highly aligned profile descriptions, the chosen metrics (coverage and purity) did not show significant improvement over traditional clustering methods, which already had highly aligned profile descriptions. In terms of profile quality, PSD did not surpass traditional subgroup discovery methods like DFS and VLSD in identifying high-quality subgroups, potentially overlooking some creditworthy individuals due to stricter subgroup definitions. However, PSD demonstrated strengths in pinpointing distinct subpopulations of likely non-creditworthy applicants, achieving the highest average SD specificity. This indicates its effectiveness in excluding negative instances and reducing the inclusion of non-creditworthy applicants. Regarding subgroup fairness, PSD identified more subgroups that exhibited gender bias. A deeper understanding of these biases would require qualitative analysis of subgroup descriptions. PSD excels in maintaining group notions of subgroup fairness and aligning with the data distribution. However, further research is needed to enhance its capability to detect bias effectively while maintaining its strengths in fairness and data representation.

By integrating the target variable into the clustering process, our method closely aligns with subgroup discovery techniques. By analyzing the identified subgroups for bias and fairness, this research advances our understanding of identifying bias and achieving fairness within subgroups

discovered through clustering and subgroup discovery. Future subgroup discovery algorithms could be designed to consider both aspects during the clustering process, potentially incorporating fairness metrics alongside traditional quality measures. While this study focused on categorical data and specific fairness metrics in credit scoring, future research could focus on incorporating numerical data into the variability-controlled K-medoids (VHK) algorithm.

Overall, while PSD shows promise in fairness and data alignment, ongoing research is necessary to further refine its capabilities, identifying more high-quality subgroups and ensuring it can effectively detect bias while leveraging its strengths in fairness and data representation. This work contributes to the broader goal of creating more equitable and transparent AI systems.

Bibliography

[1] Erindi Allaj. Two simple measures of variability for categorical data. *Journal of applied statistics*, 45(8):1509, 2018.

[2] Martin Atzmueller. Subgroup discovery. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 5(1):35–49, 2015.

[3] Martin Atzmueller and Frank Puppe. Sd-map–a fast algorithm for exhaustive subgroup discovery. In *Knowledge Discovery in Databases: PKDD 2006: 10th European Conference on Principles and Practice of Knowledge Discovery in Databases Berlin, Germany, September 18-22, 2006 Proceedings 10*, pages 6–17. Springer, 2006.

[4] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and machine learning: Limitations and opportunities*. MIT Press, 2023.

[5] Sacha E. Buijs. Clustering algorithms and concept descriptors in constructing conceptual spaces. Bachelor's thesis, University of Amsterdam, Faculty of Science, Science Park 900, 1098 XH Amsterdam, 2023. Supervisor: Dr. G. Sileno.

[6] Maarten Buyl and Tijl De Bie. Inherent limitations of ai fairness. *Communications of the ACM*, 67(2):48–55, 2024.

[7] Ángel Alexander Cabrera, Will Epperson, Fred Hohman, Minsuk Kahng, Jamie Morgenstern, and Duen Horng Chau. Fairvis: Visual analytics for discovering intersectional bias in machine learning. In *2019 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 46–56, 2019.

[8] CJ Carmona and David Elizondo. Subgroup discovery: Real-world applications. Technical report, Techincal Report, 2011.

[9] Peter W Chang, Leor Fishman, and Seth Neel. Feature importance disparities for data bias investigations. *arXiv preprint arXiv:2303.01704*, 2023.

[10] Tianqi Chen and Carlos 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.

[11] Aidan Cooper, Orla Doyle, and Alison Bourke. Supervised clustering for subgroup discovery: An application to covid-19 symptomatology. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 408–422. Springer, 2021.

[12] Edwin S Dalmaijer, Camilla L Nord, and Duncan E Astle. Statistical power for cluster analysis. *BMC bioinformatics*, 23(1):205, 2022.

[13] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012.

[14] Usama M Fayyad and Keki B Irani. Multi-interval discretization of continuous-valued attributes for classification learning. In *Ijcai*, volume 93, pages 1022–1029. Citeseer, 1993.

[15] James R Foulds, Rashidul Islam, Kamrun Naher Keya, and Shimei Pan. An intersectional definition of fairness. In *2020 IEEE 36th International Conference on Data Engineering (ICDE)*, pages 1918–1921. IEEE, 2020.

[16] Avijit Ghosh, Lea Genuit, and Mary Reagan. Characterizing intersectional group fairness with worst-case comparisons. In *Artificial Intelligence Diversity, Belonging, Equity, and Inclusion*, pages 22–34. PMLR, 2021.

[17] Usman Gohar and Lu Cheng. A survey on intersectional fairness in machine learning: Notions, mitigation, and challenges. *arXiv preprint arXiv:2305.06969*, 2023.

[18] Henrik Grosskreutz, Stefan Rüping, and Stefan Wrobel. Tight optimistic estimates for fast subgroup discovery. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 440–456. Springer, 2008.

[19] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. *CoRR*, abs/1610.02413:3, 2016.

[20] Franciso Herrera, Cristóbal José Carmona, Pedro González, and María José Del Jesus. An overview on subgroup discovery: foundations and applications. *Knowledge and information systems*, 29:495–525, 2011.

[21] Hans Hofmann. Statlog (German Credit Data). UCI Machine Learning Repository, 1994. DOI: https://doi.org/10.24432/C5NC77.

[22] Branko Kavšek and Nada Lavrač. Apriori-sd: Adapting association rule learning to subgroup discovery. *Applied Artificial Intelligence*, 20(7):543–583, 2006.

[23] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International conference on machine learning*, pages 2564–2572. PMLR, 2018.

[24] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International conference on machine learning*, page 2566. PMLR, 2018.

[25] Christoph Kiefer, Florian Lemmerich, Benedikt Langenberg, and Axel Mayer. Subgroup discovery in structural equation models. *Psychological Methods*, 2022.

[26] Willi Klösgen. Explora: A multipattern and multistrategy discovery assistant. In *Advances in knowledge discovery and data mining*, pages 249–271. 1996.

[27] Kenji Kobayashi and Yuri Nakao. One-vs.-one mitigation of intersectional bias: A general method to extend fairness-aware binary classification. *arXiv preprint arXiv:2010.13494*, 2020.

[28] Kenneth Lai, Vlad Shmerko, and Svetlana Yanushkevich. Fairness on synthetic visual and thermal mask images. *arXiv preprint arXiv:2209.08762*, 2022.

[29] Nada Lavrac, Branko Kavsek, Peter Flach, and Ljupco Todorovski. Subgroup discovery with cn2-sd. *J. Mach. Learn. Res.*, 5(2):153–188, 2004.

[30] Florian Lemmerich. *Novel techniques for efficient and effective subgroup discovery*. Bayerische Julius-Maximilians-Universitaet Wuerzburg (Germany), 2014.

[31] Florian Lemmerich and Martin Becker. pysubgroup: Easy-to-use subgroup discovery in python. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2018, Dublin, Ireland, September 10–14, 2018, Proceedings, Part III 18*, pages 658–662. Springer, 2019.

[32] Florian Lemmerich, Mathias Rohlfs, and Martin Atzmueller. Fast discovery of relevant subgroup patterns. In *Twenty-Third International FLAIRS Conference*, 2010.

[33] Jessica Liu, Huaming Chen, Jun Shen, and Kim-Kwang Raymond Choo. Faircompass: Operationalising fairness in machine learning. *IEEE Transactions on Artificial Intelligence*, pages 1–10, 2024.

[34] Antonio Lopez-Martinez-Carrasco, Jose M Juarez, Manuel Campos, and Bernardo Canovas-Segura. Vlsd—an efficient subgroup discovery algorithm based on equivalence classes and optimistic estimate. *Algorithms*, 16(6):274, 2023.

[35] Uli Niemann, Myra Spiliopoulou, Bernhard Preim, Till Ittermann, and Henry Völzke. Combining subgroup discovery and clustering to identify diverse subpopulations in cohort study data. In *2017 IEEE 30th International Symposium on Computer-Based Medical Systems (CBMS)*, pages 582–587. IEEE, 2017.

[36] Petra Kralj Novak, Nada Lavrač, and Geoffrey I Webb. Supervised descriptive rule discovery: A unifying survey of contrast set, emerging pattern and subgroup mining. *Journal of Machine Learning Research*, 10(2), 2009.

[37] Eliana Pastor, Elena Baralis, and Luca de Alfaro. A hierarchical approach to anomalous subgroup discovery. In *2023 IEEE 39th International Conference on Data Engineering (ICDE)*, pages 2647–2659. IEEE, 2023.

[38] LW Rizkallah and NM Darwish. An analysis of subgroup discovery quality measures. *Journal of engineering and applied science*, 66:109–131, 2019.

[39] Lucas Rosenblatt and R Teal Witter. Counterfactual fairness is basically demographic parity. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, page 14464, 2023.

[40] Lucas Rosenblatt and R Teal Witter. Counterfactual fairness is basically demographic parity. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 14461–14469, 2023.

[41] Mieke Wilms, Giovanni Sileno, and Hinda Haned. Pebam: A profile-based evaluation method for bias assessment on mixed datasets. In *German Conference on Artificial Intelligence (Künstliche Intelligenz)*, pages 209–223. Springer, 2022.

[42] Stefan Wrobel. An algorithm for multi-relational discovery of subgroups. In *European symposium on principles of data mining and knowledge discovery*, pages 78–87. Springer, 1997.

[43] Hantian Zhang, Nima Shahbazi, Xu Chu, and Abolfazl Asudeh. Fairrover: explorative model building for fair and responsible machine learning. In *Proceedings of the Fifth Workshop on Data Management for End-To-End Machine Learning*, pages 1–10, 2021.

[44] Albrecht Zimmermann and Luc De Raedt. Cluster-grouping: from subgroup discovery to clustering. *Machine Learning*, 77:125–159, 2009.