# ASDF-Dashboard: Automated Subgroup Detection and Fairness Analysis

Jero Schäfer[1], Lena Wiese[1]

[1]Institute of Computer Science, Goethe University, Frankfurt am Main, Germany

### Abstract

The importance of an equal treatment of individuals by AI models drastically grows due to the demands of modern society. The potential discrimination or favoritism of specific groups of individuals is one of the common perspectives for the evaluation of model behavior. However, most of the available fairness tools require human intervention in the selection of subgroups of interest and therefore expert knowledge. In this paper we propose a new tool, the ASDF-Dashboard, which automates the process of subgroup fairness assessment. It automates the subgroup detection by applying a method based on unsupervised clustering algorithms and pattern extraction to ease the usage also for non-expert users.

### Keywords

Artificial Intelligence, Fairness, Clustering,

## 1. Introduction

The research, that was conducted in the past decades, has enabled a level of AI featuring complex systems of hundreds of possible applications and an ever growing interest in their further development. There has been a great effort in improving the developed techniques and algorithms with the goal of optimizing their performance. The more powerful and faster technologies available today facilitate the expressiveness and performance of such AI systems making them omnipresent and essential in the modern world. Machine learning methods already outperform humans in certain tasks and AI supported decision making is no longer a rarity in sensitive fields like finance or medicine. However, the consideration of the societal impact of such potentially life-changing decisions has become an increasingly important objective and, thus, it needs to be evaluated in a transparent and critical way. More precisely, AI systems must not only be designed and optimized for performance, accuracy and quality but also reckon with aspects like transparency, explainability or fairness.

Fair machine learning models have to provide an equal treatment to different individuals regardless of their sensitive characteristics, e.g., their genders, races or ethnics. It is crucial to ensure that no individual experiences discrimination or favoritism by the model choices as a consequence of their membership in a certain population. Nevertheless, it is challenging to test the behavior of a model for fairness against subgroups when considering the intersections of (sensitive) characteristics as this causes an exponentially large number of potentially

*LWDA'22: Lernen, Wissen, Daten, Analysen. October 05–07, 2022, Hildesheim, Germany*

jeschaef@cs.uni-frankfurt.de (J. Schäfer); lwiese@cs.uni-frankfurt.de (L. Wiese)
0000-0001-7727-1181 (J. Schäfer); 0000-0003-3515-9209 (L. Wiese)

discriminated or favored subgroups to test. Furthermore, it is not obvious, in general, which characteristics of a dataset or which intersections induce subgroups suffering discrimination by the model, and it usually is infeasible to test each possible intersectional subgroup for a fair treatment. Hence, an automated suggestion of subgroups for the fairness testing is desirable.

In this work we propose the ASDF-Dashboard tool for the automated subgroup fairness analysis of binary classification models. It implements the previously contributed methodology of automatic subgroup detection using an unsupervised clustering and a subsequent entropy-based pattern extraction [1] in a user-friendly, web-based interface. The results of the subgroup fairness assessment are visualized in different charts in our dashboard to give the user deep insights into the behavior of the tested binary classification model regarding the detected subgroups. In the following, we outline related work on (automated) subgroup fairness evaluation tools and frameworks in Section 2. Section 3 then presents the definitions of subgroup fairness metrics based on pattern-induced subgroups (Section 3.1) and the previously developed methodology of entropy-based pattern extraction (Section 3.2). We introduce our ASDF-Dashboard and describe its functionality in Section 4. Section 5 briefly refers to our experimental results [1] and discusses the implementation of the ASDF-Dashboard. Finally, a conclusion and potential directions for future work are given in Section 6.

## 2. Related Work

There exist quite some machine learning tools to support AI developers, data scientists and also end users to realized and understand a model's behavior when presented the data. Such tools enable the enhancement of the model in the development process, deep analyses of AI systems under various criteria and features, and transparency to end users, that can get an idea on the processing of their data out of it. In particular, the latter point is increasingly interesting as our modern society demands for more transparency in AI and an equal treatment of individuals regardless of their gender or ethnics, for example. This directly leads to the concept of AI fairness, which can be tested and visualized using diverse tools. A common drawback of such supporting tools is that they are not designed for non-expert user lacking deeper knowledge and, thus, cannot perform the required interaction with the tool appropriately.

The Boxer [2] tool provides the functionality to analyze and compare models for their behavior on the same task in an interactive fashion. It is able to identify intersectional bias in the predictions of the models for subgroups of interest as specified by the tool user. This functionality is also offered by the Fairkit-learn [3] toolkit in a similar way to monitor the performance and fairness of potentially discriminating models. Models for graph mining tasks can be investigated with a tool called FairRankVis [4], which allows to explore visualizations of the model fairness wrt. individuals and subgroups. Another approach is provided by the What-if tool [5] that performs a subgroup fairness analysis and automatically optimizes the classification threshold of the considered model based on the results of the fairness analysis. Morina et al. [6] developed a framework that delivers multiple intersectional fairness metrics and estimators. However, none of the previously mentioned tools or frameworks is able to perform a subgroup fairness analysis of a given model automatically as they all require human intervention when it comes to detecting the intersectional bias. In each of these tools, the user has to specify the subgroups of interest manually before a subgroup fairness metric is applied. Our approach, in contrast, facilitates the subgroup fairness analysis by an automated detection of subgroups for the assessment of the classifier's fairness.

The FairVis tool [7] suggests subgroups to the user, that were detected automatically by clustering the data with the k-means clustering algorithm and extracting patterns of instance prototypes. The prototypes describe the makeup of the clusters and the corresponding patterns are obtained from the dominant features matching most of the subgroup members. This means that the aggregation into a cluster made most of the individuals of this subgroup having a uniform value for the dominant attribute, which is then extracted as pattern to match data in the whole dataset. Our ASDF-Dashboard extends the approach behind FairVis by offering also different clustering algorithms for the initial subgroup detection and refining the pattern extraction by a more intuitive method to quantify the uniformity of a certain feature. Instead of ranking the features by their cluster feature entropy, we apply a configurable, global threshold to identify dominant features independent of the feature domains.

Another approach to automatically detect subgroups uses frequent-pattern mining on the dataset. The Divexplorer [8] tool searches possible patterns to evaluate differences in the model's behavior between subgroups and the whole population in the dataset. The search space of possible patterns is explored exhaustively while considering only patterns with a specific degree of support and dropping less supported patterns. The model fairness regarding the subgroups is then evaluated as the difference in the probability for prediction using FPR or FNR. Similarly, the DENOUNCER [9] tool generates possible patterns by traversing the pattern graph and searches for the most general patterns which have support above a given threshold and define subgroups where the model performs poorly (low accuracy). As the space of patterns grows exponentially with the number of features and highly depends on the complexity of the domains of the features, the detection of subgroups and the assessment of the model fairness wrt. to the detected subgroups can be very time consuming. Hence, the support thresholds need to be defined very carefully to prune the search space appropriately while also generating patterns inducing meaningful subgroups.

## 3. Automated Subgroup Fairness

The ASDF-Dashboard automatically assesses the subgroup fairness of a binary classifier on a given dataset. To this end, the system detects subgroups in the data by computing a clustering of the data. The found clusters are then treated as subgroups themselves while alternatively general patterns are derived from the clusters. The obtained patterns also induce subgroups for an evaluation of the classifier's fairness. This procedure facilitates the assessment as no set of protected attributes has to be predefined and the intersections of multiple protected attributes are covered implicitly.

### 3.1. Subgroup Fairness Metrics

Formally, we denote a dataset as $\mathcal{D} = \{x_1, \ldots, x_n\}$ of $n$ instances over a set of attributes $\mathcal{A} = \{A_1, \ldots, A_p\}$ with the possible values $v \in Dom(A_j)$ for $A_j \in A$. Given a dataset $\mathcal{D}$ and a subset of protected attributes $\{A_1, \ldots, A_q\} \subseteq \mathcal{A}$, we define a pattern $P = (a_1, \ldots, a_q) \in$

$Dom(A_1) \times \cdots \times Dom(A_q)$ over $\mathcal{D}$ such that an instance $x = (v_1, \ldots, v_p)$ satisfies $P$ if its attribute values match the pattern values ($v_i = a_i$ for $i \in \{1, ..., q\}$). Then, $P$ partitions $\mathcal{D}$ into a protected subgroup $\mathcal{D}_P = \{x \in \mathcal{D} \mid x \vDash P\}$ and an unprotected subgroup $\mathcal{D}_{\bar{P}} = \{x \in \mathcal{D} \mid x \nvDash P\} = \mathcal{D} \setminus \mathcal{D}_P$ [1]. With this notion of patterns, that induce subgroups, a binary classification model $\hat{M}$, that was trained on a dataset $\mathcal{D}$ to predict the class $\hat{y} = \hat{M}(x) \in \{0, 1\}$ of an input instance $x$, can be evaluated for its fairness wrt. the performance on the subgroups. The probabilities under which a model $\hat{M}$ predicts the positive/negative class label for an instance $x$ are denoted as $\mathbb{P}(\hat{y} = 1)$ and $\mathbb{P}(\hat{y} = 0)$, respectively. The classifier $\hat{M}$ predicts the class label $c \in \{0, 1\}$ for the protected subgroup with probability $\mathbb{P}(\hat{y} = c \mid x \ \in \mathcal{D}_P)$ and correct or wrong predictions given the real label $g \in \{0, 1\}$ are expressed as $\mathbb{P}(\hat{y} = c \mid y = g, x \in \mathcal{D}_P)$. The probabilities for the unprotected group $\mathcal{D}_{\bar{P}}$ are expressed analogously.

Many subgroup fairness metrics quantify the model fairness by using the values derived from confusion matrices [10, 11] such as the positive predictive value (PPV) or the true positive rate (TPR). Barocas et al. [12] further categorize subgroup fairness metrics by the three criteria "independence", "separation" and "sufficiency" which relate to most of the proposed fairness definitions. Regarding independence, a fair classifier satisfies non-discrimination if the classification is statistically independent from the membership in the protected or unprotected subgroup. The rate of acceptance (or denial), i.e., $\mathbb{P}(\hat{y} = 1 \mid x \in \mathcal{D}_P)$ (or $\hat{y} = 0$), is then equal between the two subgroups. Separation extends this category by also considering a potential correlation between the subgroup membership and the ground-truth class such that the protected and unprotected subgroup should experience equal TPRs and FPRs. Finally, sufficiency requires an independence of the probability for the ground-truth class given a positive or negative prediction. This results in the same positive/negative predictive values for the protected and unprotected subgroup. The ASDF-Dashboard computes three different subgroup fairness metrics for a broader analysis and investigation of the classification model, namely, *statistical parity*, *equal opportunity* and *equalized odds*. In the following, the formulas of these criteria are given in the context of our notion of patterns and the induced protected and unprotected subgroups as introduced in [1].

Statistical parity (Eq. 1) is satisfied if the protected subgroup $\mathcal{D}_P$ has the same chance for the prediction of a positive outcome ($\hat{y} = 1$) as the unprotected subgroup $\mathcal{D}_{\bar{P}}$ [11]:

$$\mathbb{P}(\hat{y} = 1 \mid x \in \mathcal{D}_P) = \mathbb{P}(\hat{y} = 1 \mid x \in \mathcal{D}_{\bar{P}}) \tag{1}$$

This definition requires that a fair classifier predicts the favorable label ($\hat{y} = 1$) with a probability independently of the protected attribute values. The same is also implied for the unfavorable label ($\hat{y} = 0$) due to the complementary probability. However, if instances fall into multiple of the protected groups, statistical parity tends to magnify the bias of the classifier against them [13].

Equal opportunity (Eq. 2) judges a classifier based on the probability of giving instances $x$ of the favorable class ($y = 1$) a correct prediction, i.e., $x$ is assigned the favorable class label by classifier $\hat{M}$. Formally, it is fulfilled if

$$\mathbb{P}(\hat{y} = 1 \mid y = 1,\ x \in \mathcal{D}_P) = \mathbb{P}(\hat{y} = 1 \mid y = 1,\ x \in \mathcal{D}_{\bar{P}}). \tag{2}$$

Assuming equal opportunity, the TPRs for instances regardless of their subgroup membership have to coincide. From Equation 2 also follows that the probability of a false prediction of the unfavorable class given $x$ actually is a member of the favorable class has to be equal between the subgroups (FNR) as $\mathbb{P}(\hat{y} = 0 \mid y = 1,\ x \in \mathcal{D}_P) = 1 - \mathbb{P}(\hat{y} = 1 \mid y = 1,\ x \in \mathcal{D}_P)$.

The equalized odds subgroup fairness metric extends the equal opportunity definition by additionally forcing the equality of the subgroup's FPRs:

$$\mathbb{P}(\hat{y} = 1 \mid y = 0,\ x \in \mathcal{D}_P) = \mathbb{P}(\hat{y} = 1 \mid y = 0,\ x \in \mathcal{D}_{\bar{P}}) \tag{3}$$

Thus, equalized odds is satisfied if the probabilities of correct positive predictions (Eq. 2) and incorrect positive predictions (Eq. 3) are the same for the protected and unprotected subgroup.

The previously defined fairness criteria can be used to derive metrics that quantify the subgroup fairness of the binary classifier instead of enforcing the strict equality only. Hence, the model can then considered fair if the probabilities for an equal treatment are similar and unfair if they are not close. The ASDF-Dashboard relaxes the three fairness criteria as shown in Table 1. For $F_{spd}$ and $F_{eod}$ the probability for an instance of the protected subgroup $\mathcal{D}_P$ is subtracted from the probability for an instance of the unprotected subgroup $\mathcal{D}_{\bar{P}}$. The equalized odds metric $F_{aod}$ is computed as the average of the equal opportunity metric and the difference between the probability for an incorrect positive prediction by the classifier on the unprotected and protected subgroup. These relaxations of the three fairness criteria are implemented as fairness metrics in the "AI Fairness 360" toolkit [14]. Alternatively, ratios of the subgroup probabilities can be computed, e.g., as applied in the $\epsilon$-differential fairness definitions [6, 15] of statistical parity, equal opportunity or equalized odds.

**Table 1**
Fairness metrics for pattern-induced subgroups

| Definition | Fairness Metric |
|---|---|
| Statistical parity | $F_{spd}(P) = \mathbb{P}(\hat{y} = 1 \mid x \in \mathcal{D}_{\bar{P}}) - \mathbb{P}(\hat{y} = 1 \mid x \in \mathcal{D}_P)$ |
| Eq. opportunity | $F_{eod}(P) = \mathbb{P}(\hat{y} = 1 \mid y = 1, x \in \mathcal{D}_{\bar{P}}) - \mathbb{P}(\hat{y} = 1 \mid y = 1, x \in \mathcal{D}_P)$ |
| Equalized odds | $F_{aod}(P) = \frac{1}{2}\big[\mathbb{P}(\hat{y} = 1 \mid y = 0, x \in \mathcal{D}_{\bar{P}}) - \mathbb{P}(\hat{y} = 1 \mid y = 0, x \in \mathcal{D}_P)$ |
|  | $+ \quad \mathbb{P}(\hat{y} = 1 \mid y = 1, x \in \mathcal{D}_{\bar{P}}) - \mathbb{P}(\hat{y} = 1 \mid y = 1, x \in \mathcal{D}_P)\big]$ |

Whenever one of these fairness metrics given a pattern $P$ over dataset $\mathcal{D}$ is close to zero, it means that the classifier produces fair results on individuals from the subgroup $\mathcal{D}_P$ as they are treated similar to the rest of the population. Fairness metric values less than zero indicate a favoritism of the individuals from $\mathcal{D}_P$ over the rest of the population due to a higher probability for a positive prediction. If the fairness metrics, in contrast, yield a value greater than zero, the classifier discriminates against individuals from the protected subgroup $\mathcal{D}_P$ according to the underlying fairness definition.

## 3.2. Pattern Extraction for Subgroup Detection

The unsupervised task of detecting meaningful groups in a dataset $\mathcal{D}$ can be performed by computing a clustering $\mathcal{C} = \{C_1, \ldots, C_k\}$ that divides $\mathcal{D}$ into such groups of similar instances. The groups are so-called clusters and a pair of instances $x_1$ and $x_2$, that belong to the same cluster, shares some similarity. The cluster structure and degree of similarity between the individuals in the same group depend on the clustering type, distance measure and parameter selection. Our ASDF-Dashboard computes such a clustering either in an automated fashion or controlled by the parameters the user specified. After the clusters are found, we employ out notion of the previously defined patterns and the induced protected and unprotected subgroups to assess the classifiers fairness.

Based on the clustering $\mathcal{C}$, a pattern can be extracted that partitions the dataset $\mathcal{D}$ into protected and unprotected subgroups according to the clusters. To this end, a clustering-based pattern [1] $P_i^{\mathcal{C}} = (i)$ is defined over the artificial cluster label attribute $A_{\mathcal{C}}$ for each cluster $C_i \in \mathcal{C}$. These patterns map the instances $x \in C_i$ to protected subgroups $\mathcal{D}_{P_i^{\mathcal{C}}}$ and the subgroup fairness of $\hat{M}$ is then calculated for a fairness metric $F$ by averaging over all clusters:

$$\bar{F}(P^{\mathcal{C}}) = \frac{1}{k} \cdot \sum_{i=1}^{k} |F(P_i^{\mathcal{C}})| \text{ for } F \in \{F_{spd}, F_{eod}, F_{aod}\} \tag{4}$$

As an alternative method we refined the clustering-based subgroup detection by deriving more sophisticated patterns from the clusters [1]. These patterns describe the makeup of the found clusters and map instances from $\mathcal{D}$ to the protected subgroups. The patterns are therefore derived from the most meaningful attributes that dominate a cluster $C_i \in \mathcal{C}$, i.e., the majority of instances $x \in C_i$ have the same value $v_j \in Dom(A_j)$ for the dominant attribute $A_j \in \mathcal{A}$. The dominant features are determined by calculating the normalized cluster feature entropy [1]

$$H_{i,j} = -\frac{1}{\log_2 |Dom(A_j)|} \cdot \sum_{v \in Dom(A_j)} \frac{N_{i,j,v}}{N_i} \cdot \log_2 \left( \frac{N_{i,j,v}}{N_i} \right) \tag{5}$$

where $N_i$ is the size of $C_i$ and $N_{i,j,v}$ denotes the number of instances $x \in C_i$ that have value $v$ for attribute $A_j$. The closer $H_{i,j}$ is to zero, the more instances with the same value for feature $A_j$ are contained in the cluster $C_i$ and a single value for $A_j$ is found at all instances if the $H_{i,j} = 0$. If $H_{i,j}$ is close to 1, this indicates more variation in the feature values across the instances in the cluster $C_i$. The set of dominant features of a cluster $C_i$ is determined as $A^i = \{A_j \in \mathcal{A} \mid H_{i,j} \leq t\}$ for some threshold $0 \leq t \leq 1$. An entropy-based pattern

$$P_i^t = (a_1, \dots, a_q) \in Dom(A_1) \times \dots \times Dom(A_q)$$

is then obtained for each cluster $C_i \in \mathcal{C}$ by extracting the most frequent values of each of the dominant features $A_j \in A^i$. These patterns $P_i^t$ map all instances $x \in \mathcal{D}$ to the protected subgroup $\mathcal{D}_{P_i^t}$ that exactly match the most frequent values of the dominant features of cluster $C_i$. However, if all candidate features exceed the threshold $t$, i.e., $A^i = \emptyset$, no pattern can be extracted. In contrast to the clustering-based patterns, the protected subgroup does not exclusively contain individuals from $C_i$ but also other individuals from other clusters that match the dominant attributes' values. Here, the normalization of the feature entropy ensures that an appropriate global threshold can be set ignoring differing sizes of the active domains of the attributes throughout the dataset [1]. In the following, we also refer to the subgroups induced by clustering-based patterns as clusters or clustering-based subgroups and to the subgroups induced by entropy-based patterns as entropy-based subgroups.

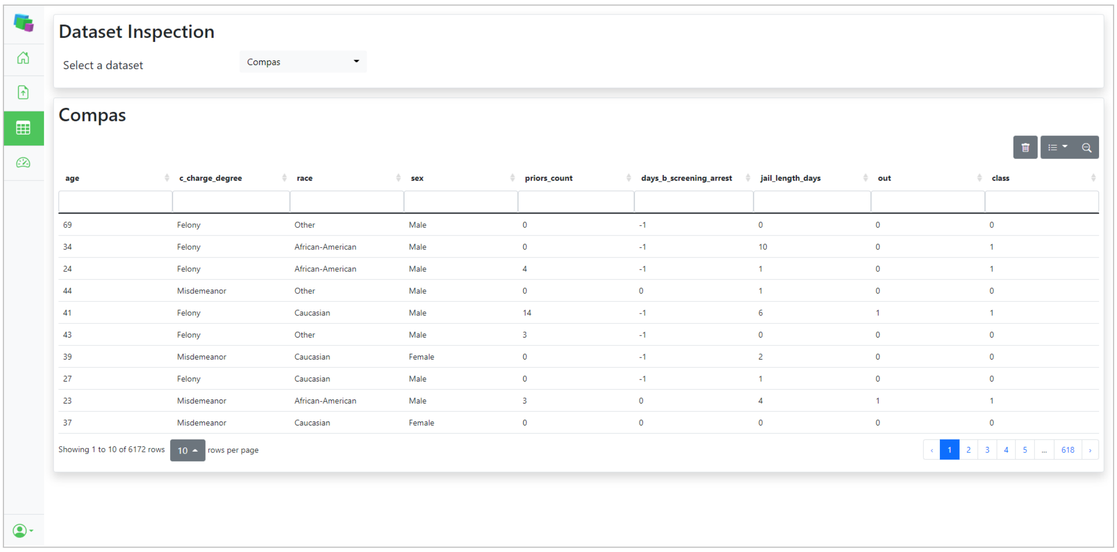
Consider a clustering-based pattern $P_i^{\mathcal{C}}$ and an entropy-based pattern $P_i^t$ extracted from the same cluster $C_i \in \mathcal{C}$ of dataset $\mathcal{D}$. The two patterns induce the different protected subgroups $\mathcal{D}_{P_i^{\mathcal{C}}}$ and $\mathcal{D}_{P_i^t}$, respectively. Generally, there might be instances $x \in C_i$ with $x \nvDash P_i^t$ such that $x \in \mathcal{D}_{P_i^{\mathcal{C}}}$ but $x \notin \mathcal{D}_{P_i^t}$. On the contrary side, there might be also individuals $x \in \mathcal{D}_{P_i^t}$ from other clusters $C_j, j \neq i$ that satisfy $x \vDash P_i^t$ and, thus, are member of the protected subgroup $\mathcal{D}_{P_i^t}$ but not of $\mathcal{D}_{P_i^{\mathcal{C}}}$. However, the protected subgroups of two entropy-based patterns $P_i^t$ and $P_j^t$ might share some individuals or even coincide due to the same dominant features and most frequent values in both $C_i$ and $C_j$. This is not possible for the protected subgroups of two clustering-based patterns $P_i^{\mathcal{C}}$ and $P_j^{\mathcal{C}}$ as we assume a hard partitional clustering with disjoint clusters, i.e., $\mathcal{D}_{P_i^{\mathcal{C}}} \cap \mathcal{D}_{P_j^{\mathcal{C}}} = \emptyset$. Two entropy-based patterns, in contrast, might be identical ($P_i^t = P_j^t$) which causes an induction of the same subgroup $\mathcal{D}_{P_i^t} = \mathcal{D}_{P_j^t}$.

## 4. ASDF-Dashboard

Our ASDF-Dashboard[1] implements the subgroup fairness analysis based on the proposed methodology of clustering- and entropy-based subgroups. It is hosted as a publicly accessible web application that supports an automatic, user-friendly subgroup fairness analysis and provides a broad visualization of the subgroup fairness results. Registered users can upload their datasets, that contain also the ground-truth labels as well as the labels obtained as the predictions of their binary classifier, to the system. The ASDF-Dashboard further provides a tabular view of each of the uploaded dataset that can be used to interactively browse the uploaded data using sorting and column filters to get a better insight into the structure of the data. Figure 1 shows an exemplary table for the COMPAS [16] dataset, which is provided in the FairVis [7] repository incl. the predicted labels but without the clustering labels. Each of the table rows corresponds to one defendant whose recidivism for a period of 2 years was predicted by a binary classifier.

To perform the subgroup fairness analysis, at least a dataset $\mathcal{D}$, the positive (favorable) class label (0 or 1) and the entropy threshold ($0 \leq t \leq 1$) for the pattern extraction have to specified in the control tile, which is depicted at the top of Figure 2). Then, the ASDF-Dashboard can already compute the subgroup fairness of the given classfier automatically. Optionally, the user can also specify the categorical columns by selection, which need to be one-hot encoded for the computation of the clustering as the distances measures most commonly require vectors of numeric data as input. If no categorical attributes are selected, the system automatically detects them to apply the one-hot encoding. Furthermore, the fully numeric features of the selected dataset are then also scaled using the min-max normalization before computing the clustering. However, clustering a mixture of numeric and categorical attributes is very sensitive to the choice of algorithm and distance metric as there is no consensus on the optimal technique [17]. Therefore, we decided for the usual processing involving encoding and scaling. In addition to the automatic subgroup fairness calculation, the classifier's fairness can also be evaluated manually by choosing a clustering algorithm and specifying its parameters. In case of the automatic fairness assessment, the SLINK clustering algorithm is applied to dataset $\mathcal{D} = \{x_1, \ldots, x_n\}$

[1] https://github.com/jeschaef/ASDF-Dashboard



Dataset Inspection

Select a dataset Compas

Compas

| age | c_charge_degree | race | sex | priors_count | days_b_screening_arrest | jail_length_days | out | class |
|---|---|---|---|---|---|---|---|---|
| 69 | Felony | Other | Male | 0 | -1 | 0 | 0 | 0 |
| 34 | Felony | African-American | Male | 0 | -1 | 10 | 0 | 1 |
| 24 | Felony | African-American | Male | 4 | -1 | 1 | 0 | 1 |
| 44 | Misdemeanor | Other | Male | 0 | 0 | 1 | 0 | 0 |
| 41 | Felony | Caucasian | Male | 14 | -1 | 6 | 1 | 1 |
| 43 | Felony | Other | Male | 3 | -1 | 0 | 0 | 0 |
| 39 | Misdemeanor | Caucasian | Female | 0 | -1 | 2 | 0 | 0 |
| 27 | Felony | Caucasian | Male | 0 | -1 | 1 | 0 | 0 |
| 23 | Misdemeanor | African-American | Male | 3 | 0 | 4 | 1 | 1 |
| 37 | Misdemeanor | Caucasian | Female | 0 | 0 | 0 | 0 | 0 |

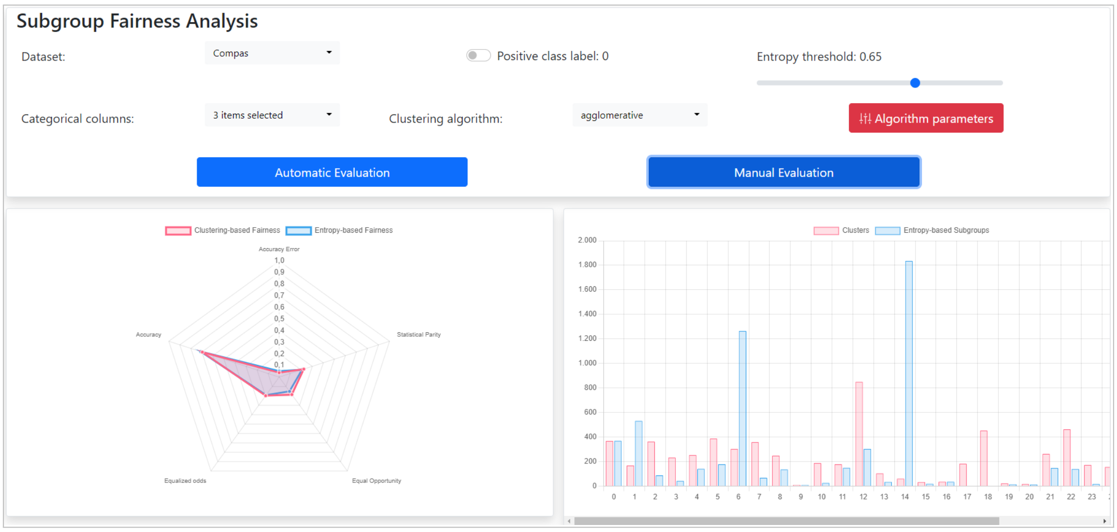
Showing 1 to 10 of 6172 rows 10 rows per page

**Figure 1: Dataset inspection.** View for the inspection of an uploaded dataset in the ASDF-Dashboard. As an example, the COMPAS [16] dataset including the predicted class (column "out") and ground-truth class (column "class") is shown here.

with the desired number of clusters $2 \leq k \leq \lfloor \sqrt{\frac{n}{2}} \rfloor$, which we estimate before by using the x-means clustering algorithm.

Figure 2 shows the subgroup fairness analysis on the COMPAS dataset with entropy threshold $t = 0.65$, favorable class label $0$ (not recidivism in two years) and the three categorical columns "c_charge_degree" (felony or misdemeanor), "race" (african-american, caucasian, asian, hispanic or other) and "sex" (female or male) using the SLINK clustering algorithm (agglomerative clustering with single linkage) with $k = 30$. The average absolute values of statistical parity, equal opportunity, equalized odds, accuracy and difference between the subgroup and global accuracy (accuracy error) over both the clustering- (red) and entropy-based subgroups (blue) are shown by the radar chart in the left tile of the dashboard (Figure 2). The average absolute values give a good insight over violations of any of the fairness definitions by discrimination or favoritism throughout all the detected subgroups. The tile next to it displays the sizes of the clusters and entropy-based subgroups by bars. Here, it can be clearly seen that the both types of subgroups do not coincide in general as the sizes differ. Especially, the cluster $C_{14}$ is much smaller than the entropy-based subgroup $\mathcal{D}_{P^t_{14}}$, for instance, as $P^t_{14} =$ ("Felony", "African-American", "Male") is a common attribute pattern in the dataset matching 1836 individuals.

To get an overview over the subgroups, the extracted entropy-based patterns are displayed in a table (Figure 3). Each row in the table corresponds to one entropy-based pattern $P^t_i$ extracted from cluster $C_i \in \mathcal{C}$. The columns of the table represent all the features $A_j \in A^i$ of the dataset $\mathcal{D}$ that were found to be dominant in at least one of the clusters $C_i \in \mathcal{C}$, i.e., $H_{i,j} \leq 0.65$ for some $i \in \{1, ..., k\}$.The remaining features are not relevant for the entropy-based patterns and, thus, not listed in the table. The column "id" identifies cluster $C_i$ and the corresponding entropy-based pattern $P^t_i$. For example, the entropy-based subgroup 8 in Figure 3 is defined by $P^t_8 =$ ("Felony", "Caucasian", "Female", $-1$) for the set of dominant attributes



Subgroup Fairness Analysis

Dataset: Compas

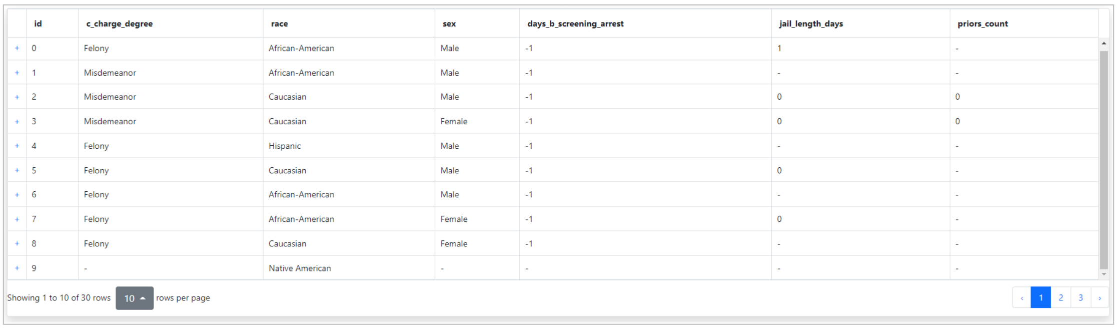
Positive class label: 0

Entropy threshold: 0.65

Categorical columns: 3 items selected

Clustering algorithm: agglomerative

**Figure 2: Subgroup fairness analysis.** The upper tile initiates the fairness analysis of a classifier by selecting a dataset (incl. predictions) and setting the positive (favorable) class label, the entropy threshold $t$, the categorical features to one-hot encode and optionally the clustering algorithm with parameters. The radar chart (bottom left) displays global fairness metrics for both clustering- and entropy-based subgroups and the bar chart (bottom right) visualizes the detected subgroup sizes.



| | id | c_charge_degree | race | sex | days_b_screening_arrest | jail_length_days | priors_count |
|---|---|---|---|---|---|---|---|
| + | 0 | Felony | African-American | Male | -1 | 1 | - |
| + | 1 | Misdemeanor | African-American | Male | -1 | - | - |
| + | 2 | Misdemeanor | Caucasian | Male | -1 | 0 | 0 |
| + | 3 | Misdemeanor | Caucasian | Female | -1 | 0 | 0 |
| + | 4 | Felony | Hispanic | Male | -1 | - | - |
| + | 5 | Felony | Caucasian | Male | -1 | 0 | - |
| + | 6 | Felony | African-American | Male | -1 | - | - |
| + | 7 | Felony | African-American | Female | -1 | 0 | - |
| + | 8 | Felony | Caucasian | Female | -1 | - | - |
| + | 9 | - | Native American | - | - | - | - |

Showing 1 to 10 of 30 rows 10 rows per page

**Figure 3: Pattern information.** The extracted entropy-based patterns are listed as rows in a table. The table columns represent the features found to be dominant in at least one cluster. The minus symbol ("-") indicates that a feature does not occur in a pattern due to its high normalized entropy.

$A^8 = \{$"c_charge_degree", "race","sex","days_b_screening_arest"$\}$. Each table row can also be expanded to show an embedded table listing the individual fairness metrics.

These fairness metrics can also be investigated for each cluster and entropy-based subgroup in the chart displayed in the left tile of Figure 4 individually. The individual fairness metrics are visualized on click on one of the subtables in the pattern table or on one of the cluster-/entropy-based subgroup size bars. Here, the selected cluster and entropy-based subgroup are $\mathcal{D}_{P_0^C}$ and $\mathcal{D}_{P_0^t}$, respectively. The bars reveal that they share a similar accuracy score of $\approx 65\%$. However, the tested classifier slightly discriminates the protected instances $x \in \mathcal{D}_{P_0^t}$ as compared to the unprotected instances according to the three subgroup fairness metrics whereas it treats

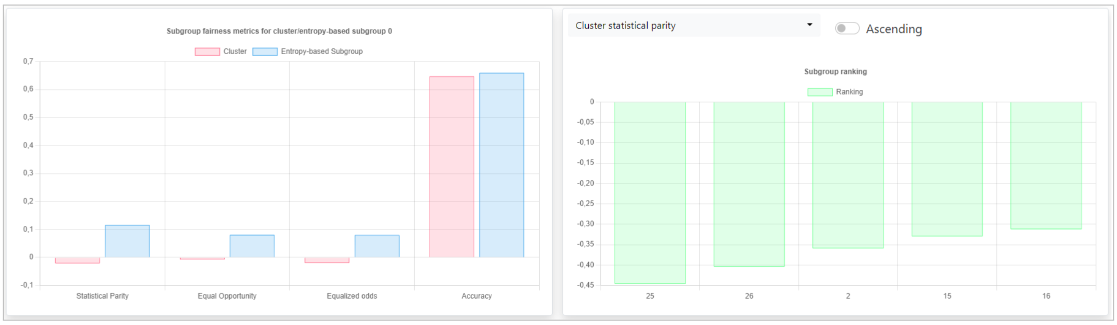


**Figure 4: Individual subgroup metrics.** The subgroup fairness metric values are shown by the bar plots (left) for an individual cluster and subgroup induced by the extracted entropy-based pattern (here: cluster/subgroup 0). The dropdown menu enables the selection of statistical parity, equal opportunity, equalized odds or prediction accuracy to visualized the top five clusters or entropy-based subgroups regarding the selected criterion, e.g., the five clusters with the lowest statistical parity values are shown.

the protected instances wrt. the clustering-based subgroup $\mathcal{D}_{P_0^{\mathcal{C}}}$ equally to the unprotected individuals. The five most discriminated or favored subgroups by subgroup fairness metric can be quickly found by the ranking chart in the right tile (Figure 4). Sorting the individual subgroup fairness values in ascending order yields the top five favored subgroups and the descending sorting order yields the most discriminated ones. Next to the three subgroup fairness metrics also the cluster or entropy-based subgroup accuracy values can be ranked in the same manner.

## 5. Evaluation

In our experiments [1] we tested our system using the COMPAS [2] dataset version from FairVis [7] ($n = 6172, p = 7$), an updated version of the Statlog German Credit [3] dataset ($n = 1000, p = 20$) and the Medical Expenditure Panel Survey (MEPS) [4] dataset from panel 19 of 2015 ($n = 15830$, $p = 40$). For each of the datasets we compared multiple clustering algorithms for the automated subgroup detection, namely, k-Means, DBSCAN, OPTICS, Spectral Clustering, SLINK, Ward, BIRCH, SSC-BP, SSC-OMP, and EnSC. Based on the dataset, small sets of individual parameter values (usually the number of clusters and the main parameters (e.g., $\epsilon$ at DBSCAN)) were tested in a grid search fashion for the detection of subgroups. We chose to report the subgroup fairness results for each parameter setting only for the run that had maximal clustering performance and showed the highest fairness violation. To this end, we measured the silhouette score $S_{\mathcal{C}}$ of clustering $\mathcal{C}$ and the mean absolute error between the prediction accuracy of classifier $\hat{M}$ on the clustering-based subgroups in comparison to the global accuracy (Eq. 7 [1]).

As our previous experiments [1] have shown, our proposed subgroup detection methods are applicable for the automated subgroup fairness analysis of a binary classifier. The applied clustering algorithms showed a varying performance as measured by mean absolute error in prediction accuracy on the clusters and sometimes multiple algorithms provided an equally

[2] https://github.com/poloclub/FairVis/blob/master/models/processed/compas_out.csv
[3] https://archive.ics.uci.edu/ml/datasets/South+German+Credit+%28UPDATE%29
[4] https://meps.ahrq.gov/mepsweb/data_stats/download_data_files_detail.jsp?cboPufNumber=HC-183

good performance in different settings. The SLINK clustering algorithm yielded a strong overall performance at detecting unfairly treated subgroups. In fact, it outperformed the other clustering algorithms in many of the experimental settings including different datasets and subgroup fairness metrics. Due to the outstanding results, we implemented it for the fully automated subgroup fairness analysis in our tool. Additionally, the ASDF-Dashboard offers users the opportunity to select and configure any of the clustering algorithms for the subgroup detection.

The visualizations of the fairness analysis results support the comprehension of the classification model's behavior when presented individuals of different subgroups in the data. The ASDF-Dashboard presents various charts with the fairness metric values for a broad coverage of diverse aspects. The users can investigate the characteristics of the found subgroups, i.e., the sizes of the clustering- and entropy-based subgroups and the extracted patterns for each cluster, as well as the subgroup fairness metrics as measured for each subgroup individually. The rankings of the clusters or entropy-based subgroups allow for a direct access to the most discriminated or favored subgroups assuming a certain subgroup fairness metric. Additionally, the global fairness values are displayed to the user as an overall judgement of the classifier's fairness. However, our tool is limited to fairness assessment for the task of binary classification and can not be applied to multi-class settings which require different subgroup fairness definitions and metrics. Another limitation is that datasets and models are not uploaded separately for modular compositions of dataset and model but just the dataset containing the predictions.

## 6. Conclusion

In this work we presented the ASDF-Dashboard for carrying out a subgroup fairness analysis of a binary classifier. Our tool is able to detect meaningful subgroups being treated unfairly by the classification model as measured by three common subgroup fairness metrics. The detection of the discriminated or favored subgroup uses an unsupervised clustering and an entropy-based pattern approach to automatically identify subgroups of similar instances with as little user interaction as possible. After the subgroup fairness assessment, users can explore the visualizations of the analysis results in various ways including global and local fairness measurements. In future research one could revise and further improve the subgroup detection methods by testing more clustering algorithms and datasets to get more insights into the performance and robustness of the proposed methods in various scenarios. Another future direction could be a qualitative comparison between the clustering-based and other approaches like frequent pattern mining approaches. In particular, an investigation on the properties of the extracted patterns could yield valuable information. Furthermore, it might also be beneficial to derive a cluster validation index based on some subgroup fairness criterion that allows to select the best out of multiple clustering models for the subgroup detection.

## References

[1] J. Schäfer, L. Wiese, Clustering-Based Subgroup Detection for Automated Fairness Analysis, in: S. Chiusano, T. Cerquitelli, R. Wrembel, K. Nørvåg, B. Catania, G. Vargas-Solar,

E. Zumpano (Eds.), New Trends in Database and Information Systems, Springer International Publishing, Cham, 2022, pp. 45–55.

[2] M. Gleicher, A. Barve, X. Yu, F. Heimerl, Boxer: Interactive comparison of classifier results, in: Computer Graphics Forum, volume 39, Wiley Online Library, 2020, pp. 181–193.

[3] B. Johnson, Y. Brun, Fairkit-learn: A Fairness Evaluation and Comparison Toolkit, 44th International Conference on Software Engineering Companion (ICSE '22 Companion) (2022).

[4] T. Xie, Y. Ma, J. Kang, H. Tong, R. Maciejewski, FairRankVis: A Visual Analytics Framework for Exploring Algorithmic Fairness in Graph Mining Models, IEEE Transactions on Visualization and Computer Graphics 28 (2021) 368–377.

[5] J. Wexler, M. Pushkarna, T. Bolukbasi, M. Wattenberg, F. Viégas, J. Wilson, The what-if tool: Interactive probing of machine learning models, IEEE transactions on visualization and computer graphics 26 (2019) 56–65.

[6] G. Morina, V. Oliinyk, J. Waton, I. Marusic, K. Georgatzis, Auditing and Achieving Intersectional Fairness in Classification Problems, arXiv preprint arXiv:1911.01468 (2019).

[7] A. A. Cabrera, W. Epperson, F. Hohman, M. Kahng, J. Morgenstern, D. H. Chau, FAIRVIS: Visual Analytics for Discovering Intersectional Bias in Machine Learning, 2019 IEEE Conference on Visual Analytics Science and Technology (VAST) (2019).

[8] E. Pastor, L. de Alfaro, E. Baralis, Looking for Trouble: Analyzing Classifier Behavior via Pattern Divergence, in: Proceedings of the 2021 International Conference on Management of Data, 2021, pp. 1400–1412.

[9] J. Li, Y. Moskovitch, H. Jagadish, DENOUNCER: Detection of Unfairness in Classifiers, Proceedings of the VLDB Endowment 14 (2021) 2719–2722.

[10] C. Hertweck, C. Heitz, A Systematic Approach to Group Fairness in Automated Decision Making, in: 2021 8th Swiss Conference on Data Science (SDS), IEEE, 2021, pp. 1–6.

[11] S. Verma, J. Rubin, Fairness Definitions Explained, in: 2018 ieee/acm international workshop on software fairness (fairware), IEEE, 2018, pp. 1–7.

[12] S. Barocas, M. Hardt, A. Narayanan, Fairness and Machine Learning, fairmlbook.org, 2019. http://www.fairmlbook.org.

[13] M. H. Teodorescu, L. Morse, Y. Awwad, G. C. Kane, Failures of Fairness in Automation Require a Deeper Understanding of Human-ML Augmentation, MIS Quarterly 45 (2021).

[14] R. K. E. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilovic, S. Nagar, K. N. Ramamurthy, J. Richards, D. Saha, P. Sattigeri, M. Singh, K. R. Varshney, Y. Zhang, AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias, 2018.

[15] J. R. Foulds, R. Islam, K. N. Keya, S. Pan, An Intersectional Definition of Fairness, in: 2020 IEEE 36th International Conference on Data Engineering (ICDE), IEEE, 2020, pp. 1918–1921.

[16] J. Angwin, J. Larson, S. Mattu, L. Kirchner, Machine Bias, 2016. URL: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing, Accessed 27 June 2022.

[17] M. van de Velden, A. Iodice D'Enza, A. Markos, Distance-based clustering of mixed data, Wiley Interdisciplinary Reviews: Computational Statistics 11 (2019) e1456.