Reliable and fair machine learning for risk assessment

Pereira Barata, A.P.

Chapter 5

Fair Tree Classifier

When learning classification models from biased data, the resulting classifiers tend to exacerbate the biases present [Richardson, 2022]. With respect to the Inspectorate, a case in point is *confirmation* bias.

Consider the following example. In international cargo ship risk assessment, a prevailing trait towards selecting a ship for inspection is the colour of the flag of the ship. A reputable country is assigned a white flag. However, the flag may be either white, grey, or black, and reflects the detention rate of ships for that country. Indeed, inspectors may be disproportionately influenced by the colour of a flag, causing more frequent and stringent inspections of ships with non-white flags, leading to confirmation bias in data.

To learn a classifier from such biased data, the standard classification problem becomes three-fold: (1) it is necessary to learn a model with high classification performance; (2) the impact of the biases on the model must be suppressed (i.e., model fairness); and (3) the performance-fairness trade-off must be tunable such that the requirements by the relevant stakeholders can be easily met.

In this chapter, we propose SCAFF: a solution to the problem at hand in the form of a compound splitting criterion which combines (a) AUC, (b) strong demographic parity, and (c) a performance-fairness trade-off tunability parameter. In our experimental results, we show via performance-fairness trade-off curves how SCAFF generates effective models with competitive performance and high fairness. This result answers RQ3: how can we, from biased data, learn a model tunable with respect to the performance-fairness trade-off such that the selection of the trade-off point is made intuitive for the relevant stakeholders?

The current chapter corresponds to the following publication:
Pereira Barata, A., Takes, F. W., van den Herik, H. J., and Veenman, C. J. (2022). Fair tree classifier using strong demographic parity. *Machine Learning (under review)*

5.1 Algorithmic Fairness

The application of machine learning algorithms for classification has become ubiquitous within an abundance of domains [Brink et al., 2016, Sarker, 2021, Azar and El-Metwally, 2013, Pereira Barata et al., 2021, Dressel and Farid, 2018]. Great dependency on automated decision-making, however, gives rise to concerns over model bias; e.g., bias was reported by Amazon's automatic recruitment tool in which women unfairly scored lower. It turns out that models were trained on resumes submitted mostly by men, thus disadvantaging women a priori [Dastian, 2018]. To prevent the modelling of historical biases, it is of the utmost importance to develop fairness-aware methods [European Commission, 2019c].

A fair classification model has three goals: (1) to make adequate class predictions from *unseen* observations; (2) to ensure that the bias in data is suppressed from those predictions [Cho et al., 2020]; and (3) to allow for the tunability of the inherent trade-off between the aforementioned two goals —the performance-fairness trade-off [Kleinberg et al., 2016]— such that the ethical, legal, and societal needs of the end user (i.e., domain expert) are met. Here we remark that the third goal is of greatest importance, as achieving it provides a manner by which trade-off points can be made selectable by the relevant stakeholders.

To quantify model fairness (i.e., the extent to which the biases in data have been suppressed) different fairness measures have been proposed (see Definitions 1.5, 1.6, and 1.7). Traditionally, fairness measures such as demographic parity [Dwork et al., 2012], equal opportunity [Corbett-Davies and Goel, 2018], or equalised odds [Hardt et al., 2016] are used. These fairness measures are all *threshold-dependent*. A threshold-dependent fairness measure is defined as follows.

Definition 5.1 – ***Threshold-dependent fairness measure***

A threshold-dependent fairness measure is a quantification of algorithmic bias with respect to some sensitive group, measured as a function of the class predictions induced by applying a threshold to the (continuous) model output.

Considering a classification model with continuous output, a decision threshold must be set to produce class predictions, upon which those measures are reliant. In other words, fairness would only be ensured with respect to that particular threshold. To counter this limitation, a *threshold-independent* fairness measure can be used instead. A threshold-independent fairness measure is defined as follows.

Definition 5.2 – ***Threshold-independent fairness measure***

A threshold-independent fairness measure is a quantification of algorithmic bias with respect to some sensitive group, measured as a function of the (continuous) model output, rather than the class predictions.

One such measure is the *strong demographic parity*. The strong demographic parity extends the aforementioned demographic parity by considering fairness throughout the entire range of possible decision thresholds. Although having been proposed in [Jiang et al., 2020], the authors provided an implementation of strong demographic parity merely towards the linear classifier case.

Tree-based algorithms are regarded as a state-of-the-art solution for the classification problem [Zabihi et al., 2017, Dogru and Subasi, 2018, Angenent et al., 2020]. Their prevalence in the literature is mostly due to (1) model interpretability, (2) their tendency to not overfit when used as ensembles, (3) requiring little data pre-processing, and (4) handling mixed data types and missingness [Dogru and Subasi, 2018]. Past work on tree splitting criteria has shown positive results with respect to threshold-dependent fairness [Kamiran et al., 2010]. There is a desire to extend it towards the threshold-independent case.

In this work, we propose SCAFF: the Splitting Criterion AUC For Fairness. SCAFF allows for fair tree classifier learning by directly optimising for the threshold-independent fairness measure of strong demographic parity. In particular, we propose a *fair tree classifier* learning algorithm which simultaneously (1) optimises for threshold-independent classification performance (i.e., AUC); (2) suppresses the impact of bias directly in terms of strong demographic parity; and (3) is tunable with respect to the performance-fairness trade-off during learning. In addition, our method handles various multicategorical sensitive attributes simultaneously, and easily extends to bagging (i.e., random forest) and (gradient) boosting frameworks.

The structure of the chapter follows: Section 5.2 expresses our problem description formally; Section 5.3 discusses related work; Section 5.4 elaborates our SCAFF method; Section 5.5 describes our experiments; Section 5.6 refers to our results; and Section 5.7 concludes and recommends research directions.

5.2 Problem Description

We consider the scenario in which a labelled dataset is intrinsically biased with respect to one or more sensitive attributes of which the values may be either binary or multicategorical. Our task is to learn a fair predictive model from the biased data, such that future predictions are independent from the sensitive attribute(s).

We require that the measures of model performance and fairness do not depend on a decision threshold set upon the output. Since there is no unique solution in the trade-off between performance and fairness, the fair model must also be readily tunable in this regard, as to meet the requirements of the application domain.

Formally, consider a dataset D with n samples, m features, and two classes. Without loss of generality, assume the case in which a single binary sensitive attribute exists. Let X, Y, and S be the underlying variable distributions representing the feature space, classes, and sensitive attribute, respectively, from which the n samples were drawn. Accordingly, each sample may be represented as (x_i, y_i, s_i), for $i = 1, 2, \ldots, n$.

The goal of the learning algorithm is to learn the distribution for which the conditional $P(Y|X) \approx P(Y|X, S)$. In practice, this amounts to learning from the data a mapping function $f : x \in X \rightarrow z \in Z$ where Z represents the model output (i.e., classification score) upon which a threshold t induces a class prediction, and under which the condition of strong demographic parity must be met, $\forall t \in Z : P(Z \geq t|S_+) = P(Z \geq t|S_-)$, while maximising for the threshold-independent classification performance $P[(Z|Y_+) \geq (Z|Y_-)]$. The compromise between strong demographic parity and the corresponding maximal predictive performance must also be tunable.

5.3 Related Work

In this section, we discuss the concepts from the literature related to our work: the measures of fairness (Section 5.3.1), and the fair tree splitting criteria used towards fair tree classification learning (Section 5.3.2).

5.3.1 Measures of Fairness

Fairness measures in the literature may be categorised as being either (a) threshold-dependent or (b) threshold-independent. With respect to threshold-dependent measures, the three most prevalent are: (1) demographic parity [Dwork et al., 2012]; (2) equal opportunity [Corbett-Davies and Goel, 2018]; and (3) equalised odds [Hardt et al., 2016].

First, *demographic parity* (see Definition 1.6) is the condition under which each sensitive group (e.g. male/female) should be granted a positive outcome, at equal rates. It is the absolute difference between the proportion of positive class predictions $\hat{Y}_+$ in samples with a positive sensitive attribute value S_+ and samples with a negative sensitive attribute value S_-, and is computed as $|P(\hat{Y}_+|S_+) - P(\hat{Y}_+|S_-)|$.

Second, the measure of *equal opportunity* is defined as follows.

Definition 5.3 – ***Equal opportunity***

Equal opportunity is the fairness measure which considers the absolute difference between the conditional TPR of each sensitive group

Equal opportunity is the fairness measure which accounts for the predictive reliability within each sensitive group and is computed as the absolute difference $|P(\hat{Y}_+|S_+, Y_+) - P(\hat{Y}_+|S_-, Y_+)|$.

Third, the definition of *equalised odds* follows.

Definition 5.4 – ***Equalised odds***

Equalised odds is the fairness measure which considers the absolute difference between the conditional TPR of each sensitive group, as well as the difference between the conditional FPR of each sensitive group.

Equalised odds extends from the measure of equal opportunity by also incorporating the unreliability of predictions in the sensitive groups. It is computed as $||P(\hat{Y}_+|S_+, Y_+) - P(\hat{Y}_+|S_-, Y_+)| - |P(\hat{Y}_+|S_+, Y_-) - P(\hat{Y}_+|S_-, Y_-)||$.

Albeit computationally different, the three measures share at least one common aspect: the output of the classification model must be binary; i.e., a decision threshold must be placed upon the continuous output which induces the class prediction. As a result, a problem arises when applying these measures towards learning a fair classifier. These measures of fairness are limited to being exclusively reliable for the specific threshold which produces the class prediction: there is no guarantee that fairness holds for different threshold values.

In practice, when learning several fair classifiers for real-world applications, (i.e., hyperparameter optimisation), the final classification model should not be dependent on any arbitrary threshold, as fairness should be maintained throughout. Rather, the decision threshold should only be placed a posteriori, according to the performance requirements of the end user (e.g., precision vs recall) whilst incurring minimal impact over fairness.

With respect to threshold-independent fairness measures, the notion of demographic parity has been extended into strong demographic parity (see Definition 1.7). Strong demographic parity takes into account the continuous output of the model, such that the ordering of the output should be independent of the sensitive groups. It is computed as the absolute difference between the probabilities $|P[(Z|S_+) \geq (Z|S_-)] - P[(Z|S_+) < (Z|S_-)]|$.

Although strong demographic parity was proposed with a working fair learning framework in [Jiang et al., 2020], their implementation only considers the linear classifier case. We focus on extending the implementation towards non-linear models, specifically towards tree-based architectures.

5.3.2 Fair Tree Splitting Criteria

The practice of learning a tree classifier from biased data is directly linked to the splitting criterion used to construct the tree structure. Within the fairness literature with respect to tree-based algorithms, we recommend the works by [Kamiran et al., 2010] and [Zhang and Ntoutsi, 2019], in which different approaches are used to measure classification performance and fairness. The measures are then jointly used as splitting criteria during training to select the best split.

In the work by [Kamiran et al., 2010], the authors propose to address the fair splitting criterion problem, by accounting for the impact of bias in the model during learning. They do so by extending the concept of information gain in traditional classification towards the sensitive attribute. Given data D, a split is evaluated as the information gain with respect to the class label:

$$IG_Y = H_Y(D) - \sum_{i=1}^{k} \frac{|D_i|}{|D|} \cdot H_Y(D_i), \tag{5.1}$$

and the information gain with respect to the sensitive attribute:

$$IG_S = H_S(D) - \sum_{i=1}^{k} \frac{|D_i|}{|D|} \cdot H_S(D_i), \tag{5.2}$$

where H_Y and H_S denote the entropy with respect to the class label and the sensitive attribute, respectively, and $D_i, i = 1, \ldots, k$ denotes the partitions of D induced by the split under evaluation.

Both information gains are then merged to produce two distinct compound splitting criteria by either: (1) subtracting IG_Y by IG_S, hereinafter termed $\text{Kamiran}_{\text{Sub}}$, or (2) dividing IG_Y by IG_S, hereinafter denoted as $\text{Kamiran}_{\text{Div}}$. Although this work was fundamental in establishing fair tree-learning frameworks, it is limited in scope since fairness is only considered as the threshold-dependent demographic parity.

In the work of [Zhang and Ntoutsi, 2019], a fairness-aware Hoeffding tree (FAHT) is introduced. Although the method was developed with online streaming classification as its focus, the splitting criterion developed may be generally applicable to the fair learning problem. The FAHT approach relies, as with the previous work, on a compound criterion composed of a class label part and a sensitive attribute part and addresses demographic parity. Both works use the same class label information gain IG_Y. However, the fairness component is computed differently between them. For FAHT, the fairness gain FG of a split is given as a function of $Disc(D)$ of a set of data:

$$FG = Disc(D) - \sum_{i=1}^{k} \frac{|D_i|}{|D|} \cdot Disc(D_i). \tag{5.3}$$

The bias term is defined as the observed demographic parity of the system $|P(Y_+|S_+) - P(Y_+|S_-)|$. The splitting criterion of FAHT evaluates as follows:

$$\begin{cases} IG_Y, & \text{if } FG = 0 \\ IG_Y \cdot FG, & \text{otherwise} \end{cases}. \tag{5.4}$$

The two proposed splitting criteria present some limitations, three of which deserve to be named in particular: (1) the construction processes were developed with only threshold-dependent fairness in mind; (2) both implementations only address a single binary sensitive attribute; and (3) there exists no performance-fairness trade-off tuning parameter built into the splitting criteria. In the following section, we propose our method which lifts these limitations.

5.4 The SCAFF Method

In this section we propose our SCAFF method. It is a probabilistic learning framework which (1) optimises for threshold-independent classification performance (i.e., AUC); (2) addresses fairness in terms of strong demographic parity; and (3) is tunable with respect to the performance-fairness trade-off. In addition, SCAFF leverages multiple sensitive attributes simultaneously and easily extends to bagging and boosting frameworks.

We begin by addressing the implementation of the classification performance in Section 5.4.1, followed by the implementation of the fairness measure of strong demographic parity in Section 5.4.2. In Section 5.4.3, we provide our compound splitting criterion which incorporates a tunable parameter towards the trade-off between classification performance and fairness. In Section 5.4.4, we describe the tree construction process, reporting on how our method leverages multiple sensitive attributes simultaneously and extends to bagging and boosting frameworks. A working Python implementation of our algorithm can be found in [Pereira Barata, 2021].

5.4.1 AUC Computation

In machine learning, the AUC is a measure which expresses the quality of a sample ordering with respect to a binary label $\{Y_-, Y_+\}$. It computes the probability $P[(Z|Y_+) \geq (Z|Y_-)]$. Here, a random order results in AUC $= 0.5$ and a perfect order results in AUC $= 1$; conversely AUC $= 0$ if all labels are flipped and still perfectly ordered.

Traditionally, computing the AUC has a time complexity $O(n \cdot \log(n))$:

$$\text{AUC}(Z, Y) = \frac{\sum_{i=1}^{y_+} \sum_{j=1}^{y_-} \sigma(Z_i, Z_j)}{y_+ \cdot y_-}, \tag{5.5}$$

where

$$\sigma(Z_i, Z_j) = \begin{cases} 1, & \text{if } Z_i > Z_j \\ \frac{1}{2}, & \text{if } Z_i = Z_j \\ 0, & \text{otherwise} \end{cases}. \tag{5.6}$$

Here, y_+ and y_- are the number of all instances Y_+ and Y_- respectively, and Z_i and Z_j represent the Z output scores of each corresponding instance.

Yet, for the scenario in which a parent node is split into two child nodes —towards candidate split evaluation—the time complexity of computing the AUC may be reduced. From [Lee, 2019], the AUC of a split may be re-written as a function of the TPR and the FPR induced by the split. The AUC then becomes:

$$\text{AUC} = \frac{1 + \text{TPR} - \text{FPR}}{2}. \tag{5.7}$$

For each candidate split, the child node with highest $P(Y_+)$ is assigned as the positive prediction node such that all samples contained in it are labelled $\hat{Y}_+$. The other child node induces $\hat{Y}_-$. This strategy is equivalent to computing the AUC traditionally; i.e., assigning samples in each node with Z scores equal to the proportion of ground truth positive labels $P(Y_+)$ of their corresponding node. Hereinafter, we denote AUC_Y as the AUC with respect to the class label.

5.4.2 Strong Demographic Parity

The strong demographic parity condition aims to minimise the difference in candidates from the sensitive groups among the selected candidates, regardless of any arbitrary decision threshold t. The goal is to minimise the expression $|P[(Z|S_+) \geq (Z|S_-)] - P[(Z|S_+) < (Z|S_-)]|$ from Section 5.3.1. The condition is reached by learning the target function f which randomly orders the samples towards the sensitive groups; i.e., the AUC towards the sensitive attribute.

We find the fair classifier f by optimising for an AUC value of 0.5 on the sensitive attribute. In order to solve the optimisation problem, we aim at minimising the AUC with S_+ as the positive class, which we denote as AUC_{S_+}.

Since $\text{AUC}_{\text{S}_+} = 0$ is as maximally unfair as $\text{AUC}_{\text{S}_-} = 1$, we define *sensitive AUC* (AUC_S) —f_S from Section 5.2— as the following:

$$\text{AUC}_\text{S} = \max[1 - \text{AUC}(Z, S), \text{AUC}(Z, S)], \tag{5.8}$$

such that the $\max$ operator bounds the range of possible AUC_S values to $[0.5, 1]$.

Definition 5.5 – ***Sensitive AUC***

Sensitive AUC is the AUC towards a sensitive attribute, bounded to values $[0.5, 1]$ and is proportional to the strong demographic parity. AUC_S can be computed as a function of the strong demographic parity:

$$\text{AUC}_\text{S} = \frac{\text{strong demographic parity} + 1}{2}$$

AUC_S of 1 indicates that the model is completely biased, while 0.5 indicates that the model is complete fair.

Now that both classification performance AUC_Y and fairness measure AUC_S have been described, the splitting criterion may be constructed.

5.4.3 Splitting Criterion AUC For Fairness

Towards tunability of the performance-fairness trade-off, we define the *orthogonality* parameter Θ as follows.

Definition 5.6 – ***Orthogonality parameter*** Θ

The orthogonality parameter $\Theta \in [0, 1]$ is the parameter of SCAFF which regulates the performance-fairness trade-off of the learned model: $\Theta = 0$ results in a completely biased but most performing model, whereas $\Theta = 1$ results in a completely fair but nonperforming model.

The objective is then to find a split which, for a given Θ, maximises AUC_Y (towards $\text{AUC}_\text{Y} = 1$), while minimising AUC_S (towards $\text{AUC}_\text{S} = 0.5$). Accordingly, for the fair classification problem given instance scores Z, class label Y, and sensitive attribute S, we define SCAFF:

$$\text{SCAFF}(Z, Y, S, \Theta) = (1 - \Theta) \cdot \text{AUC}_\text{Y} - \Theta \cdot \text{AUC}_\text{S}. \tag{5.9}$$

The purpose of Θ is to change the direction of the splitting criterion score towards either classification or fairness. To illustrate this effect, consider Fig. 5.1.

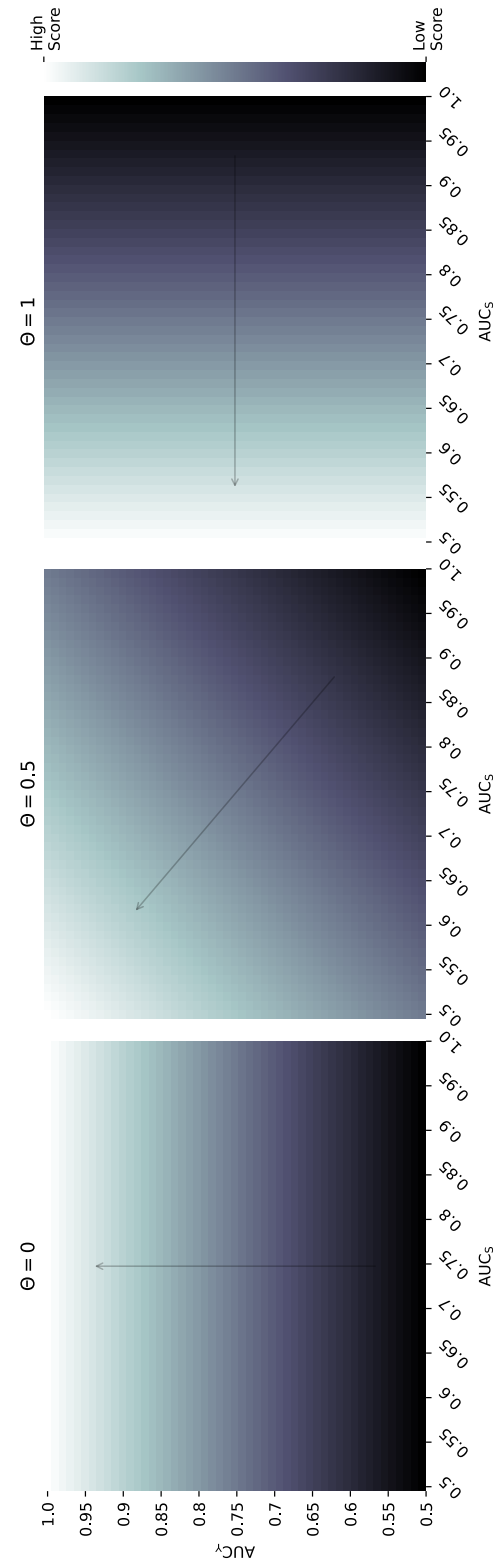

Figure 5.1: **Effect of the orthogonality Θ over SCAFF scores.** Each heatmap depicts the SCAFF scores at three different values of Θ. Each heatmap cell corresponds to a specific combination of AUC_Y (vertical axis) and AUC_S (horizontal axis) from which the score is computed.

Each heatmap represents, for varying values of Θ, the split evaluation scores for all possible values of AUC_Y (vertical axis) and AUC_S (horizontal axis), according to Eq. 5.9. The direction of the optimal score, from darkest to brightest tones, is additionally represented as an arrow. From left to right, the optimal score direction rotates along Θ. We call it the orthogonality parameter since it rotates the direction of the optimal scores, making $\Theta = 0$ and $\Theta = 1$ orthogonal score directions.

5.4.4 Tree Construction

As with any typical tree architecture, learning is done by selecting, at each step (i.e., depth), the split which optimises the splitting criterion score. A split at some feature value partitions a node into two child nodes and is evaluated according to the Z scores of the parent node and the new Z' scores of the child nodes induced by that split. The optimal split is the one which, across all possible feature value split points, maximises the splitting criterion score.

Given (a) the parent node scores Z and (b) the child scores Z' induced by a split, the SCAFF gain SG associated with that split is defined as:

$$SG = \text{SCAFF}(Z', Y, S, \Theta) - \text{SCAFF}(Z, Y, S, \Theta). \tag{5.10}$$

The split with maximal SG across all evaluated splits is selected if and only if its corresponding $SG \geq 0$. Otherwise, no splitting occurs.

SCAFF is not only able to handle binary sensitive attributes but also extends to the multivariate and multicategorical scenarios, including intersectional factors (i.e., the combination of sensitive attributes) [Buolamwini and Gebru, 2018] via a one-versus-rest (OvR) approach [Tax and Duin, 2002]. The AUC_S used in SCAFF is the maximum OvR, since no sensitive attribute should have priority over fairness.

An example of SCAFF evaluation can be viewed in Fig. 5.2, in which the OvR $\text{AUC}_\text{S} = \max(0.6, 0.917) = 0.917$. In the aforementioned example, we mention that Z scores are given as $P(Y_+)$ in a node. We remark that our methods extends trivially to the bagging (i.e., random forest) case by considering the final score of a sample as the average score across all trees. Yet, other Z score definitions are viable; e.g., (gradient) boosting techniques compute Z by iteratively updating existing sample scores [Hastie et al., 2009]. In that sense, samples within the same child node may have distinct Z scores.

Our method extends to such boosting cases since SG relies on Z, regardless of its computation. In contrast, traditional tree learning algorithms do not extend to boosting, since no Z scores are incorporated into the splitting criteria. We remark that, for samples in the same node which have distinct Z scores, the computation of the AUC must follow the traditional approach (Eq. 5.5).

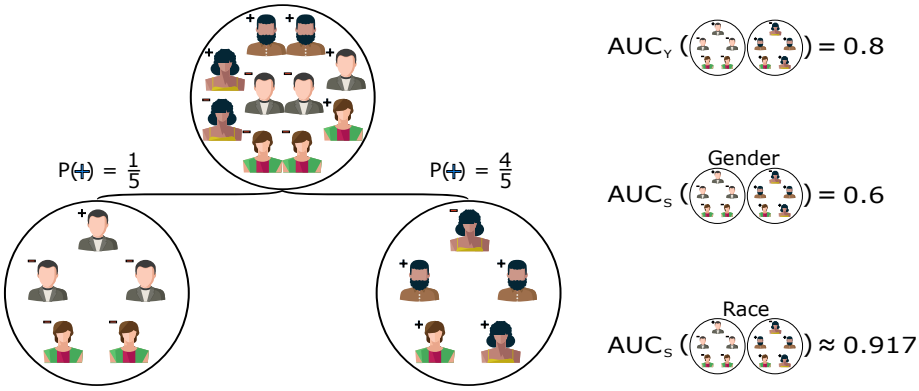

Figure 5.2: **Computing AUC values for SCAFF.** AUC_Y and AUC_S in a system with 10 samples, a class label, and two sensitive attributes (gender and race).

5.5 Experiments

For the description of our experiments, we begin by mentioning the datasets and how we used them (Section 5.5.1); we then characterise the experimental setup deployed to (1) gather the performance and fairness values and (2) report on the relationship between the threshold-independent and threshold-dependent demographic parities (Section 5.5.2). We compared SCAFF against other fair splitting criteria by using benchmark fairness datasets. Since the methods against which we compare our approach are neither suited for multivariate nor multicategorical sensitive attributes, we focus on the single binary sensitive attribute case first. We additionally experimented on a single dataset to explore how SCAFF handles multiple sensitive attributes simultaneously as well as multicategorical values. Lastly, we tested the quantitative relationship of the strong demographic parity yielded by our method with the corresponding demographic parity at different decision-thresholds. For reproducibility, our experiments are made available in [Pereira Barata, 2021].

5.5.1 Datasets

Three binary classification datasets were used. These are benchmark datasets used for fairness methods [Quy et al., 2021]. Each of them has at least one sensitive attribute. Specifically, we employed the following: (a) *Bank* ($45,211$ instances, 50 features) in which the sensitive attribute is the binary condition of age ≥ 65 (b) *Adult* ($45,222$ instances, 97 features), where the sensitive attribute may be either (i) race $\in \{\text{white}, \text{non-white}\}$ or (ii) gender $\in \{\text{male}, \text{female}\}$; and (c) *Recidivism* ($6150$ instances, 8 features) of which the sensitive attributes may be either (i) race $\in \{\text{white}, \text{non-white}\}$ or (ii) gender $\in \{\text{male}, \text{female}\}$.

For the binary sensitive attribute case, we considered each dataset-sensitive attribute configuration, making for a total of five different dataset configurations. Two scenarios were further set in which the *Adult* dataset was considered: (i) the multiple sensitive attribute scenario such that both sensitive attributes (race and gender) were handled simultaneously; and (ii) the multicategorical sensitive attribute scenario in which the intersectional attributes {non-white female (NWF), non-white male (NWM), white female (WF), white male (WM)} were concurrently considered.

5.5.2 Experimental Setup

To provide an adequate comparison between our splitting criterion and the state-of-the-art, we considered previous works in fair splitting criteria; specifically, the works proposed by [Kamiran et al., 2010] and [Zhang and Ntoutsi, 2019]. For each dataset configuration, and for all methods, the same 10-fold CV was applied.

To measure classification performance and algorithm fairness, AUC_Y (the accepted standard measure for classifier performance) and AUC_S were used, respectively. The performance and fairness measures across test folds were averaged to produce a single value pair for each dataset, per method, and in our case for each value of orthogonality Θ. For all methods, the classification output scores Z of samples were computed as the $P(Y_+)$ of the terminal leaf node of a single tree, as previously shown in Fig. 5.2.

To be able to achieve state-of-the-art performance, each method was deployed as a random forest (i.e., bagging) [Breiman, 2001]. As such, the final classification score of a sample is the average Z model output of all terminal nodes across the different trees generated. Throughout all methods, the same set of hyperparameters was used, such as the number of trees (500), the maximum depth of each tree (4), and the random seed initialisation.

Bootstrapping, random feature selection, and continuous-feature discretisation were also applied, given their prevalence in real-world implementations of tree-based algorithms, such as [Chen and Guestrin, 2016]. For our method, a range of 11 values for Θ was used between 0 and 1. For details of the implementation, see [Pereira Barata, 2021].

To relate the threshold-dependent and threshold-independent demographic parities, decision thresholds were applied to the classifier outputs of our method across different values of Θ for the different datasets. The thresholds were considered as 9 quantiles values between 0.1 and 0.9 of each test set output and, consequently, demographic parity —defined in Section 5.3.1— was averaged over folds.

We measured, at each decision threshold —along Θ values— the Pearson correlation coefficient [Kirch, 2008], and the respective null hypothesis p-values, between strong demographic parity (as AUC_S) and demographic parity. The purpose is to check whether the behaviour of strong demographic parity across Θ transfers to that of the induced demographic parity.

5.6 Results

In this section, we present the results of our experiments. We report on the classification performance, fairness, and tunability of the performance-fairness trade-off achieved by our method via orthogonality Θ. We do so for the aforementioned sensitive attribute configurations: binary (Section 5.6.1), and non-binary (Section 5.6.2). Specifically for the binary configuration, we compare our method to the competing approaches. Finally, we show how the strong demographic parity (measured in AUC_S) yielded by our method translates to the induced demographic parity across different (a) decision thresholds and (b) values of orthogonality Θ (Section 5.6.3).

5.6.1 Binary Sensitive Attribute

To regard the performance and fairness of all methods per dataset configuration, see Fig. 5.3. For our method, each point corresponds to a value of $\Theta \in [0, 1]$. An orthogonality value $\Theta = 0$ is equivalent to a traditional classifier and corresponds to the right-most point. Conversely, $\Theta = 1$ corresponds to the left-most point. In the horizontal axis, AUC_S represents (un)fairness. The vertical axis depicts AUC_Y as classification performance.

Unlike the other methods which output a single performance-fairness value (represented as a point), our SCAFF method produces a performance-fairness trade-off curve along Θ. This is advantageous as it provides a way for practitioners to make informed decisions. The impact of Θ on the tunability of the performance-fairness trade-off for each dataset-sensitive attribute pair is consistent: as increasingly greater values of Θ are used, the greater the fairness and lesser the classification performance.

Noticeably, in *Bank (Age)*, SCAFF was able to reduce AUC_S by 0.2 at a loss in performance of only 0.02. Overall, our method consistently performs better in the combination of classification performance and fairness, allowing for a suitable target point. It is a convincing result of (1) the use of AUC in the splitting criterion and (2) the flexibility of the orthogonality parameter Θ.

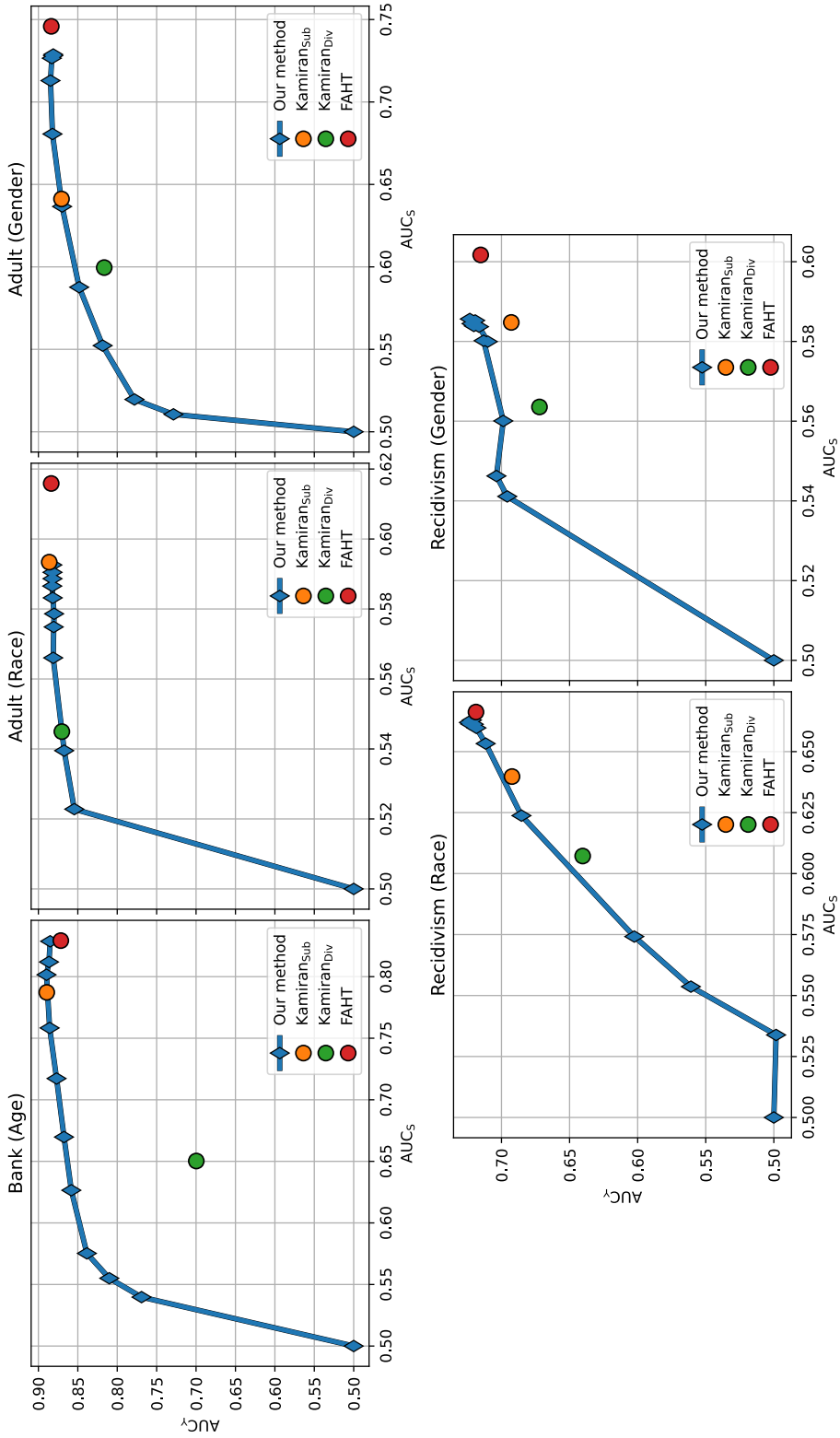

Figure 5.3: **Performance-fairness across methods per dataset.** Measures of AUC_S (horizontal axis) and AUC_Y (vertical axis) are shown. For our method, each point from right to left indicates a value of $\Theta \in [0, 1]$ in increasing order.

5.6.2 Multiple and Multicategorical Cases

We present in Fig. 5.4 the outcomes of the dataset configurations for multiple sensitive attributes —*Adult (Multiple)* in the left panel — and multicategorical sensitive attribute values, considered as the intersectional values: *Adult (Intersectional)* in the right panel.

For both panels, across different values of orthogonality Θ (horizontal axis), the classification performance AUC_Y is shown in blue and the different AUC_S are provided (vertical axis). To the left, the AUC_S for race and gender can be regarded; to the right, the AUC_S for each of the different intersectional sensitive attribute values are displayed: NWF, NWM, WF, and WM.

Focusing on the *Adult (Multiple)* configuration, it is witnessable that the behaviour of the fairness measures along Θ match those of the *Adult (Race)* and *Adult (Gender)* previously shown in Fig. 5.3: greater values of orthogonality translate to greater values of fairness (decreasing AUC_S) and lesser classification performance AUC_Y. This is expected, since the performance-fairness trade-off phenomenon is known.

SCAFF was able to reduce the bias towards both sensitive attributes simultaneously whilst maintaining adequate classification performance; in particular at $\Theta = 0.7$, both race and gender $\text{AUC}_\text{S} = 0.55$ (a remarkably low bias value), and AUC_Y is above 0.8 indicating model adequacy. Similarly for *Adult (Intersectional)* at the same value of the orthogonality parameter $\Theta = 0.7$, our method was able to converge the bias of all sensitive attribute values to sensible values concurrently whilst maintaining proper classification performance.

These results show our proposed method is able to produce adequate classification models with regards to multiple and multicategorical sensitive attributes which maximise performance with the least decrease in fairness. To put it differently, our method is able to exploit the performance-fairness trade-off even for multiple and multicategorical sensitive attributes.

We remark, however, one limitation of our OvR approach. Since the OvR AUC_S along multiple attributes or values is evaluated as its maximum (as described in Section 5.4.3), there is no guarantee that all attributes will have their biases decreased along Θ: regard the *slight* increase in NWM bias.

Yet, this characteristic of our approach inherently bounds the highest possible value of bias. In other words, along Θ, the maximum value of AUC_S is strictly monotonically decreasing. The remark is further corroborated by the NWF, WF, and WM intersectional sensitive attributes, of which the curves behave in a nearly-identical manner along the different values of Θ. Under the assumption that none of the sensitive attributes is of greater importance than any other, the maximally-valued sensitive attribute should always be considered as the attribute by which fairness is measured.

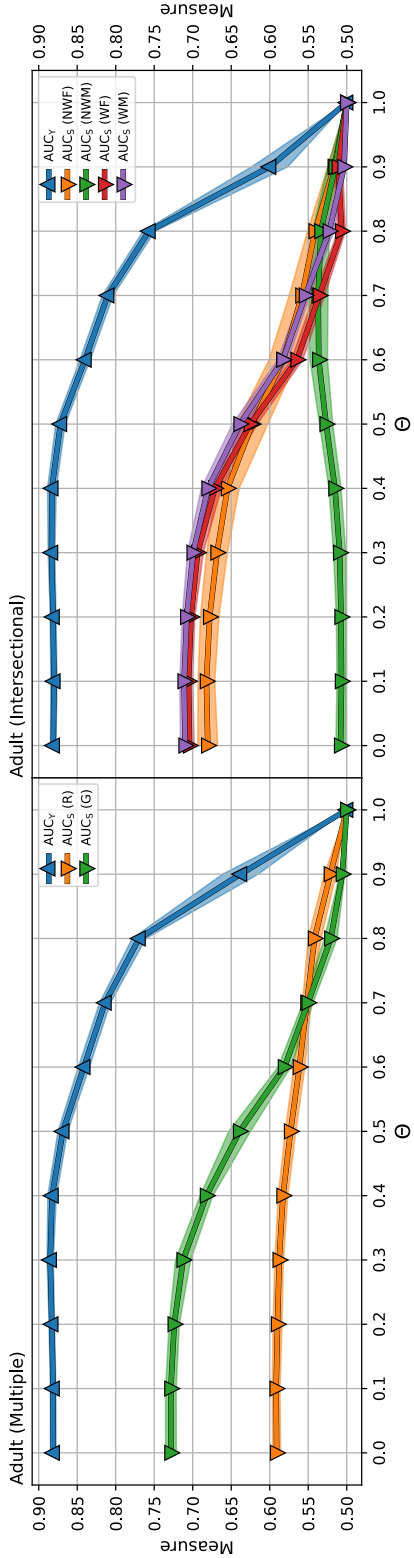

Figure 5.4: **Performance-fairness for the multiple and intersectional cases.** Across different values of orthogonality Θ (horizontal axis), average and standard deviation of AUC_Y and AUC_S values (vertical axis) are shown for the *Adult* dataset.

5.6.3 Relationship with Demographic Parity

Below, we describe the results of applying our method to the five dataset configurations for different values of Θ, and measuring the corresponding (threshold-dependent) demographic parity at different decision thresholds. The purpose is to determine if (1) threshold-independence extends across arbitrary decision thresholds, and (2) if changes in Θ induce an equivalent behaviour between demographic parity and strong demographic parity.

In Fig. 5.5, it is shown how for different decision thresholds (horizontal axis), the mean demographic parity (vertical axis) —across all test folds— behaves with different values of Θ (differently-coloured lines), for the five binary sensitive attribute dataset configurations. An additional panel is provided (bottom-right), where for each value of Θ (horizontal axis), the variation of demographic parity across decision thresholds for each dataset is present.

Across all dataset configurations, and particularly noticeable in those with high demographic parity —concretely *Bank (Age) and Adult (Gender)*— the effect of the orthogonality parameter Θ is generally the same: as orthogonality values increase, values for demographic parity decrease, regardless of the decision threshold selected.

The spread of demographic parity (measured as standard deviation) also decreases along Θ, for different decision thresholds. To put it differently, higher values of Θ translate to greater threshold-independence. This is sensical since, by definition, SCAFF directly optimises for threshold-independent measures.

To grasp the relationship between strong and threshold-dependent demographic parities, regard Table 5.1. Each row depicts a decision threshold at which demographic parity was computed; a column indicates a dataset configuration. A cell depicts the Pearson correlation coefficient between the two measures of fairness along the parameter Θ, for the decision threshold. The coefficients represent how similar the behaviour between threshold-dependent and - independent demographic parities is, induced by shifts in Θ.

Noteworthily, bolded entries indicate a statistical significance of $\alpha = 0.05$ towards the null hypothesis of no correlation. Safe for a single outlying entry —threshold 0.9 in the *Adult (Race)* configuration, in which the value of demographic parity is negligible— all table entries are consistently high and of statistical significance. This shows that the effect of shifting the orthogonality parameter Θ is, in practice, identical for both types of demographic parity regardless of the decision threshold selected, validating our method with respect to threshold independence.

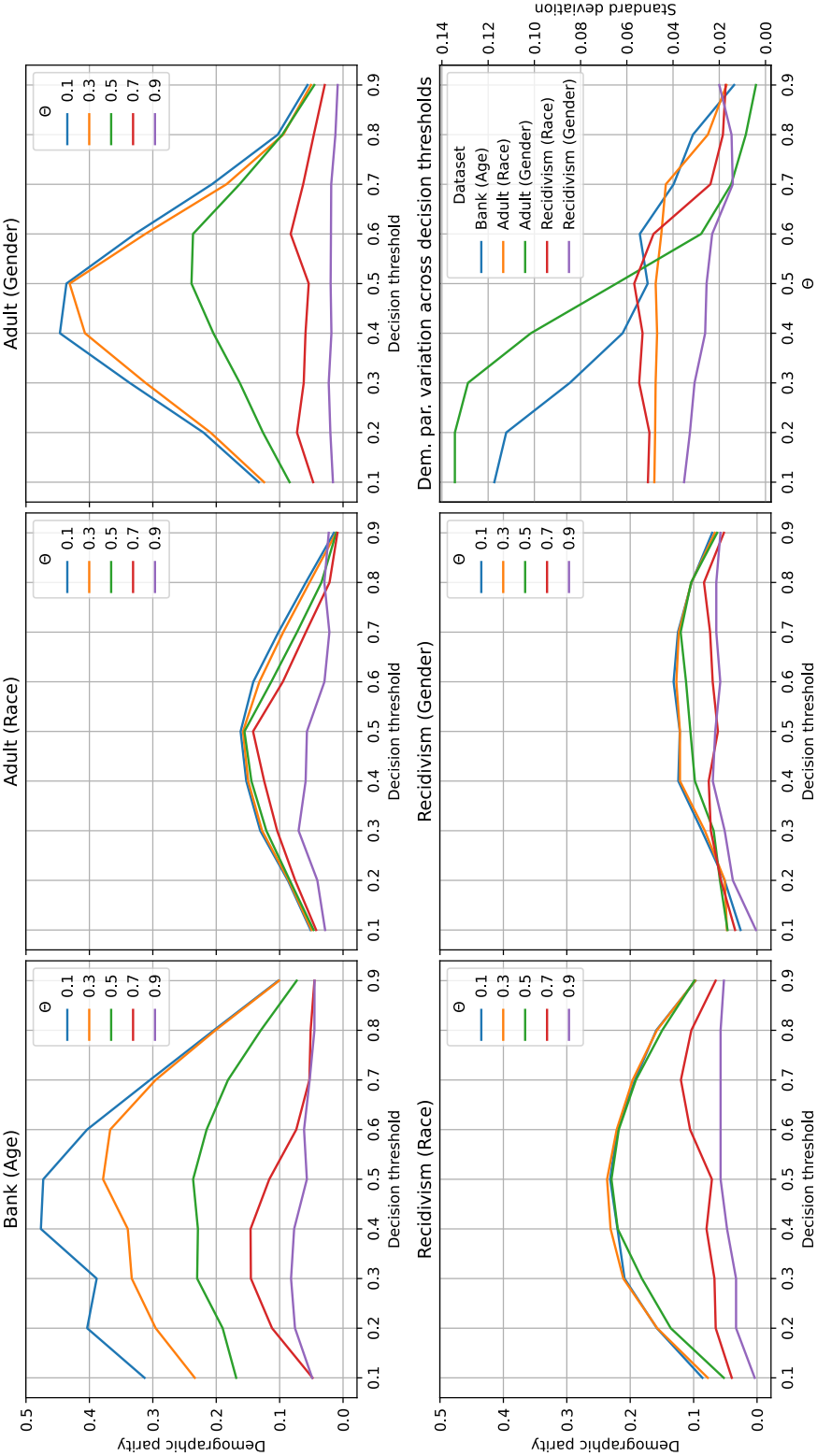

Figure 5.5: **Effect of orthogonality over demographic parity.** Decision thresholds are shown in the horizontal axis. Values of demographic parity are represented in the vertical axis. Values of Θ are highlighted with different colours. Bottom-right panel: standard deviation (vertical axis) of demographic parity across thresholds and Θ (horizontal axis).

Table 5.1: Pearson correlation coefficients between strong demographic parity (measured as AUC_S) and demographic parity, along Θ, for different decision thresholds in the five dataset configurations; bolded entries indicate a null hypothesis p-value ≤ 0.05.

	Dataset				
Th	Bank (A)	Adult (R)	Adult (G)	Recid. (R)	Recid. (G)
0.1	**0.983**	**0.963**	**0.994**	**0.937**	**0.839**
0.2	**0.984**	**0.965**	**0.997**	**0.995**	**0.895**
0.3	**0.993**	**0.971**	**0.994**	**0.987**	**0.968**
0.4	**0.988**	**0.992**	**0.991**	**0.995**	**0.949**
0.5	**0.997**	**0.988**	**0.995**	**0.990**	**0.973**
0.6	**0.993**	**0.994**	**0.995**	**0.998**	**0.975**
0.7	**0.984**	**0.979**	**0.984**	**0.991**	**0.992**
0.8	**0.975**	**0.871**	**0.919**	**0.983**	**0.984**
0.9	**0.941**	0.267	**0.947**	**0.944**	**0.922**

5.7 Chapter Conclusion

In the present work, we introduced SCAFF. By doing so, we proposed a learning algorithm which simultaneously (1) optimises for threshold-independent performance —AUC— and fairness —strong demographic parity— (2) is able to handle various multicategorical sensitive attributes simultaneously, (3) is tunable with respect to the performance-fairness trade-off via an orthogonality parameter Θ, and (4) easily extends to bagging and (gradient) boosting.

Moreover, we empirically validated our method through experimentation on benchmark datasets traditionally used in the fairness literature. Then we validated our experiments with real datasets. Here, we showed that our approach outperformed the competing state-of-the-art criteria methods, by its predictive performance and model fairness, as well as by its capability of handling multiple sensitive attributes simultaneously, of which the values may be valued multicategorically. Moreover, we demonstrated how the behaviour of strong demographic parity induced by our method extends to the threshold-dependent demographic parity.

As future work, we recommend to extend the current framework from learning classification problems towards other learning paradigms. Ultimately, the development and deployment of fair machine learning approaches within sensitive domains is the goal in this field of research.