Machine Learning (2024) 113:3305–3324
https://doi.org/10.1007/s10994-023-06376-z

Fair tree classifier using strong demographic parity

António Pereira Barata[1,2] · Frank W. Takes[1] · H. Jaap van den Herik[3] · Cor J. Veenman[1,4]

Received: 3 June 2022 / Revised: 13 June 2023 / Accepted: 17 July 2023 /
Published online: 24 August 2023

Abstract

When dealing with sensitive data in automated data-driven decision-making, an important concern is to learn predictors with high performance towards a class label, whilst minimising for the discrimination towards any sensitive attribute, like gender or race, induced from biased data. Hybrid tree optimisation criteria have been proposed which combine classification performance and fairness. Although the threshold-free ROC-AUC is the standard for measuring classification model performance, current fair tree classification methods mainly optimise for a fixed threshold on the fairness metric. In this paper, we propose SCAFF—splitting criterion AUC for Fairness—a compound decision tree splitting criterion which combines the threshold-free *strong* demographic parity with ROC-AUC termed, easily applicable as an ensemble. Our method simultaneously leverages multiple sensitive attributes of which the values may be multicategorical, and is tunable with respect to the unavoidable performance-fairness trade-off. In our experiments, we demonstrate how SCAFF generates effective models with competitive performance and fairness with respect to binary, multicategorical, and multiple sensitive attributes.

Keywords Fairness · Criterion · Society · Sensitive

Editors: Dana Drachsler Cohen, Javier Garcia, Mohammad Ghavamzadeh, Marek Petrik, and Philip S. Thomas.

✉ António Pereira Barata
a.p.pereira.barata@liacs.leidenuniv.nl

[1] LIACS, Leiden University, Niels Bohrweg 1, 2333 CA Leiden, The Netherlands

[2] ILT, Ministry of Infrastructure and Water Management, Graadt van Roggenweg 500, 3531 AH Utrecht, The Netherlands

[3] LCDS, Leiden University, Niels Bohrweg 1, 2333 CA Leiden, The Netherlands

[4] Netherlands Organization for Applied Scientific Research (TNO), Anna van Buerenplein 1, 2595 DA The Hague, The Netherlands

1 Introduction

The application of machine learning algorithms for classification has become ubiquitous within an abundance of domains (Brink et al., 2016; Sarker, 2021; Azar & El-Metwally, 2013; Barata et al., 2021; Dressel & Farid, 2018). Great dependency on automated decision-making, however, gives rise to concerns over model discrimination; 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, 2019).

A fair classification model has two goals: (1) to make adequate class predictions from *unseen* observations; and (2) to ensure that those class predictions are as independent of a sensitive attribute as possible (Hu et al., 2020; Cho et al., 2020). In addition, the *performance-fairness trade-off* —the inevitable situation in which the lesser the fairness of an algorithm, the greater its predictive capabilities and vice-versa (Kleinberg et al., 2016)—should be tunable to satisfy the ethical, legal, and societal needs of the end user (i.e., domain expert).

A fair classifier is most commonly learned by jointly optimising towards a classification performance measure and a fairness measure. Traditionally, fairness measures such as demographic parity (Dwork et al., 2012), equal opportunity (Corbett-Davies & Goel, 2018), or equalised odds (Hardt et al., 2016) are used. These fairness measures are all threshold-dependent. 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, the threshold-independent fairness measure termed *strong demographic parity* was proposed in Jiang et al. (2020). It extends the aforementioned demographic parity by considering fairness throughout the entire range of possible decision thresholds. However, the authors merely considered a logistic regression classifier implementation.

Meanwhile, tree-based algorithms are still regarded as a state-of-the-art solution (Zabihi et al., 2017; Dogru & Subasi, 2018). The prevalence of tree-based approaches in the literature is mostly due to (1) their tendency to not overfit when used as ensembles, (2) requiring little data pre-processing, and (3) handling mixed data types and missingness (Dogru & Subasi, 2018). Past work on tree splitting criteria has shown positive results with respect to threshold-dependent fairness (Kamiran et al., 2010).

In this work, we aim at delivering a fair splitting criterion termed *SCAFF*: Splitting Criterion AUC for Fairness, which allows for fair tree classifier learning. In particular, we propose a *fair tree classifier* learning algorithm which simultaneously (1) optimises for threshold-independent ROC-AUC classification performance and threshold-independent *strong* demographic parity, (2) handles various multicategorical sensitive attributes simultaneously, (3) is tunable with respect to the performance-fairness trade-off during learning, and (4) extends to ensemble methods.

The structure of the paper follows: Sect. 2 expresses our problem statement formally; Sect. 3 discusses related work; Sect. 4 elaborates our SCAFF method in detail; Sect. 5 describes our experiments; Sect. 6 refers to our results; and Sect. 7 concludes and recommends research directions.

2 Problem statement

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 definitions 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 tunable in this regard.

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 $X \subseteq \mathbb{R}$, $\{Y_+, Y_-\} \subseteq Y$, and $\{S_+, S_-\} \subseteq S$ be the underlying variable distributions representing the feature space, class labels, 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 fair learning algorithm is to learn the distribution for which the conditional $P(Y \mid X) \approx P(Y \mid X, S)$. In practice, this amounts to learning, from the data, a mapping function $f : x \in X \rightarrow z \in Z \subseteq \mathcal{R}$, where Z represents the model output (i.e., classification score), which aims at the strong demographic parity condition $P[(Z \mid S_+) > (Z \mid S_-)] = P[(Z \mid S_-) > (Z \mid S_+)]$, whilst maximising the *traditional* threshold-independent classification performance $P[(Z \mid Y_+) > (Z \mid Y_-)]$. The compromise between strong demographic parity and the corresponding maximal predictive performance must also be tunable.

3 Related work

In this section, we discuss the concepts from the literature related to our work: the measures of fairness (Sect. 3.1), and the fair tree splitting criteria used towards fair tree classification learning (Sect. 3.2).

3.1 Measures of fairness

Several fairness measures exist in the literature, which may be categorised as either (a) threshold-dependent or (b) threshold-independent. The three most prevalent threshold-dependent measures are: (1) demographic parity (Dwork et al., 2012); (2) equal opportunity (Corbett-Davies & Goel, 2018); and (3) equalised odds (Hardt et al., 2016).

First, *demographic parity* is the condition under which candidates of each sensitive group (e.g. male/female) should be granted positive outcomes at equal rates: $P(\hat{Y}_+ \mid S_+) = P(\hat{Y}_+ \mid S_-)$. As a fairness measure, it is defined as the absolute difference between the proportion of positive class predictions $\hat{Y}_+$ in instances with a positive sensitive attribute value S_+ and instances with a negative sensitive attribute value S_-; formally: $|P(\hat{Y}_+ \mid S_+) - P(\hat{Y}_+ \mid S_-)|$.

Second, the condition of *equal opportunity* accounts for the predictive reliability within each sensitive group: $P(\hat{Y}_+ \mid S_+, Y_+) = P(\hat{Y}_+ \mid S_-, Y_+)$. As a fairness measure, it is computed by taking the absolute difference between the true positive rate of the instance groups composed of the positive and negative sensitive values $|P(\hat{Y}_+ \mid S_+, Y_+) - P(\hat{Y}_+ \mid S_-, Y_+)|$.

Third, *equalised odds* condition extends the previous definition by also incorporating the unreliability of predictions in the sensitive groups, met when $P(\hat{Y}_+ \mid S_+, Y_-) = P(\hat{Y}_+ \mid S_-, Y_-)$. As a fairness measure, it is computed as $||P(\hat{Y}_+ \mid S_+, Y_+) - P(\hat{Y}_+ \mid S_-, Y_+)| - |P(\hat{Y}_+ \mid S_+, Y_-) - P(\hat{Y}_+ \mid S_-, Y_-)||$.

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. By being threshold-dependent, these measures of fairness are limited to being exclusively reliable for the specific threshold which produces the class prediction; i.e., there is no guarantee that fairness holds for different threshold values. In other words, a learned classifier is restricted to the decision threshold for which the fairness measure was optimised. This is an issue since, in real-world applications, the decision-threshold is volatile and dependent on the specific domain requirements.

To counter this, the notion of threshold-dependent demographic parity has been extended to the threshold-independent case, termed the *strong demographic parity* condition, introduced in Jiang et al. (2020). It considers the *continuous* model output Z, with respect to the sensitive groups $\{S_+, S_-\} \subseteq S$, and is met when $P[(Z \mid S_+) > (Z \mid S_-)] = P[(Z \mid S_-) > (Z \mid S_+)]$ (Sect. 2). Put simply, the strong demographic parity condition considers the *ordering* (or *ranking*) of the output Z, and is met when the sensitive groups are indiscernible by score. In Sect. 4.1, we describe its computation as a fairness measure.

However, the aforementioned work only considered the implementation of strong demographic parity for the logistic regression case. This impacts applicability since state-of-the-art non-linear models cannot be learned which directly optimise towards the strong demographic parity condition. We therefore focus on expanding the implementation of strong demographic parity towards non-linear models, specifically to tree-based architectures.

3.2 Fair tree splitting criteria

One clear advantage of tree learning algorithms is that they may be designed with any arbitrary splitting-selection criterion. The criterion does not have to be differentiable, as long as it is computationally tractable. A second advantage of tree frameworks over other architectures is their verified performance within different domains, making them a state-of-the-art solution to classification problems (Zabihi et al., 2017; Dogru & Subasi, 2018).

The practice of learning fairness-aware tree classifiers is directly linked to the splitting criterion used to construct the tree structure, the possibility to tune the performance-fairness trade-off, and their ensemble capability. Within the*fair tree* literature, we recommend the works by Zhang & Weiss (2022); Zhang et al. (2021), and Kamiran et al. (2010).

Although the two former works are the most recent, managing to ensure both (a) performance-fairness tunability and (b) ensemble capability, in this paper, we will *explicitly* focus on the work by Kamiran et al. (2010). The reasons for this follow. On the one hand, Zhang & Weiss (2022) focus on a *censored* setup in which a class label is not directly available; this is inherently distinct from our case in which we *assume* class label availability. On the other hand, from Zhang et al. (2021), the fair classification setup is based on online component classifiers with constant updates using streaming data (i.e., Hoeffding trees); our setup does *not* involve nor support streaming data. In contrast to the aforementioned works, Kamiran et al. (2010) focus on static non-streaming and non-censored data, which is akin to our problem.

In their work, Kamiran et al. (2010) address the fair classification problem 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:

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

and the information gain with respect to the sensitive attribute:

$$IG_S = H_S(D) - \sum_{i=1}^{k} \frac{|D_i|}{|D|} \cdot H_S(D_i), \tag{2}$$

where H_Y and H_S denote the entropy with respect to the class label and the sensitive attribute, respectively, and $D_i, i = 1, \dots, k$ denotes the partitions of D induced by the split under evaluation. Both information gains are then merged to produce three distinct compound splitting criteria, each using a different operator: (1) $IG_Y + IG_S$ (Kamiran$_{\text{Sum}}$); (2) $IG_Y - IG_S$ (Kamiran$_{\text{Sub}}$); or (3) $\frac{IG_Y}{IG_S}$ (Kamiran$_{\text{Div}}$);

In addition, the authors incorporate a leaf-relabelling step after a tree is constructed, in which a terminal leaf node assigned as a class label prediction is re-assigned the opposite class label prediction. In other words, if a node represents $\hat{Y}_+$, then it swaps to $\hat{Y}_-$, and vice-versa.

This step aims at minimising the discrimination—as a function of demographic parity—of the classifier, whilst minimising the loss in accuracy stemming from node relabelling. It is approached heuristically, as a form of the knapsack problem: in which the ranking order for each leaf node relabelling is given as a function of its impact on the change in discrimination (demographic parity) and accuracy $\frac{\Delta\text{discrimination}}{\Delta\text{accuracy}}$. Given a *discrimination allowance* $\epsilon \in [0, 1]$, then the leaf nodes are relabelled successively by rank, until ϵ is met.

In their experiments, it is shown that, for *individual* trees, the ϵ parameter is able to tune the performance-fairness trade-off well. Moreover, shape of the trade-off curve is shown to be dependent of the operator approach used when generating the compound fairness criterion.

The approaches described according to splitting criterion (Kamiran$_{\text{Sum}}$, Kamiran$_{\text{Sub}}$, and Kamiran$_{\text{Div}}$), alongside ϵ-tuning of the performance-fairness trade-off, present some limitations, three of which deserve to be named in particular: (1) the construction processes was developed with only threshold-dependent fairness in mind; (2) only single binary sensitive attributes are considered; and (3) the ensemble capability was not assessed (e.g., random forest experimentation). In the following section, we describe our proposed treed-based framework which lifts these limitations.

4 Method

In this section we describe our proposed method. It is a probabilistic tree learning framework which (1) optimises for strong demographic parity, (2) is tunable with respect to the performance-fairness trade-off, and (3) addresses multiple multicategorical sensitive attributes simultaneously.

We begin by addressing how the measure of strong demographic parity is tractable in Sect. 4.1. In Sect. 4.2, we provide our compound splitting criterion which incorporates a tunable parameter towards the trade-off between classification performance and fairness. Following, Sect. 4.3 details the assignment of Z scores to instances, which are required for our method. Section 4.4 addresses the AUC computation under a decision tree framework. Finally, in Sect. 4.5, we describe the tree construction process. A working Python implementation of our algorithm can be found in Pereira Barata (2021).

4.1 Strong demographic parity

The strong demographic parity condition is met when there is homogeneity in the rankings of scores of candidates across sensitive groups, regardless of any arbitrary decision threshold t. For simplicity, here we follow with the binary sensitive attribute case.

The strong demographic parity condition is met when, as per Sects. 2 and 3.1, $P[(Z \mid S_+) > (Z \mid S_-)] = P[(Z \mid S_-) > (Z \mid S_+)]$ which may be rewritten as $P[(Z \mid S_+) > (Z \mid S_-)]-$ $P[(Z \mid S_+) > (Z \mid S_-)] - P[(Z \mid S_-) > (Z \mid S_+)] = 0$. In practice, it is found by minimising $|P[(Z \mid S_+) > (Z \mid S_-)] - P[(Z \mid S_-) > (Z \mid S_+)]|$. Since $P[(Z \mid S_-) > (Z \mid S_+)] = 1 - P[(Z \mid S_+) \geq (Z \mid S_-)]$, then, by allowing $P[(Z \mid S_+) > (Z \mid S_-)] = P[(Z \mid S_+) \geq (Z \mid S_-)]$ under *ordinal* constraint (i.e., of any two samples, it is *always* possible to assess which has greater Z):

$$\textit{strong demographic parity} = |2 \cdot P[(Z \mid S_+) \geq (Z \mid S_-)] - 1|. \tag{3}$$

The probability term is computed as the *normalised* Mann–Whitney U statistic (Mann & Whitney, 1947), from groups $\{S_+, S_-\} \subseteq S$ and scoring Z:

$$P[(Z \mid S_+) \geq (Z \mid S_-)] = \frac{U(Z,S)}{s_+ \cdot s_-} = \frac{\sum_{i=1}^{s_+} \sum_{j=1}^{s_-} \sigma(Z_i, Z_j)}{s_+ \cdot s_-}, \tag{4}$$

$$\sigma(Z_i, Z_j) = \begin{cases} 1, & \text{if } Z_i > Z_j \\ \frac{1}{2}, & \text{if } Z_i = Z_j \\ 0, & \text{otherwise} \end{cases}. \tag{5}$$

Here, s_+ and s_- are the number of instances S_+ and S_- respectively, and Z_i and Z_j represent the Z output scores of each corresponding instance. To note, the $Z_i = Z_j$ condition accounts for ties in ranking when the ordinality of Z cannot be imposed. Furthermore, the normalised Mann–Whitney U statistic is equivalent to the ROC-AUC (hereinafter, AUC) Mason and Graham (2002):

$$\mathrm{AUC}(Z,S) = \frac{U(Z,S)}{s_+ \cdot s_-}. \tag{6}$$

Finally, we have:

$$\textit{strong demographic parity} = |2 \cdot \mathrm{AUC}(Z,S) - 1|. \tag{7}$$

As a result, the strong demographic parity is minimal when $\mathrm{AUC}(Z,S) = 0.5$.

We extend the notion of strong demographic parity via $\mathrm{AUC}(Z, S)$ to the multicategorical case by following a one-versus-rest (OvR) approach:

$$\textit{strong demographic parity} = \max_{S_k \in S}[|2 \cdot \mathrm{AUC}(Z, S_k) - 1|], \tag{8}$$

$$\mathrm{AUC}(Z, S_k) = P[(Z \mid S_k) \geq (Z \mid \neg S_k)], k \in \{1, 2, \dots m\}. \tag{9}$$

Here, m represents the number of *unique* sensitive values for a specific sensitive attribute, and S_k represents the sensitive value itself. For multiple sensitive attributes, the same approach is applied, in which the max is taken across all sensitive attributes, after considering the max across all values per attribute.

4.2 Splitting criterion AUC for fairness

In order to reach our goal of a fair tree learner (Sect. 2), our proposed compound splitting criterion must join the strong demographic measure with the *traditional* threshold-independent classification performance measure given as $P[(Z \mid Y_+) \geq (Z \mid Y_-)]$, or simply the $\mathrm{AUC}(Z, Y)$ (Mason & Graham, 2002).

Yet, during tree construction (i.e., node split evaluation), the straightforward measure of $\mathrm{AUC}(Z, Y)$ is lacking. It lacks because it considers *only* $P[(Z \mid Y_+) \geq (Z \mid Y_-)]$: a split which maximises $P[(Z \mid Y_-) \geq (Z \mid Y_+)]$ $(= 1 - P[(Z \mid Y_+) \geq (Z \mid Y_-)])$ is equally as good at separability (Lee, 2019). Therefore, we define the *classification performance component* of our criterion:

$$\textit{classification performance component} = |2 \cdot \mathrm{AUC}(Z, Y) - 1|. \tag{10}$$

The objective becomes finding a split which maximises the classification performance component (Eq. 10) and minimises the strong demographic parity component (Eq. 7). Moreover, we propose an *orthogonality* parameter $\Theta \in [0, 1]$ which we incorporate into our splitting criterion to tune the trade-off between classification performance and fairness: $\Theta = 0$ only promotes classification performance; and $\Theta = 1$ only promotes fairness. Accordingly, for the simplest fair classification problem given instance scores Z, class label Y, and sensitive attribute S, we define SCAFF —Splitting Criterion AUC for Fairness—as:

$$\mathrm{SCAFF}(Z, Y, S, \Theta) = (1 - \Theta) \cdot |2 \cdot \mathrm{AUC}(Z, Y) - 1| - \Theta \cdot |2 \cdot \mathrm{AUC}(Z, S) - 1|. \tag{11}$$

4.3 Assigning *Z* scores

For tree building, a straightforward solution is to consider the *proportion of positive class* label instances in that node $Z = P(Y_+ \mid \textit{node})$. To be more explicit, this is the *ratio* between (a) the number of positive class samples Y_+ in the node and (b) the total number of samples in that node. Conversely, $Z = P(Y_- \mid \textit{node})$ is also a viable option since, as per Sect. 4.2, our classification performance component is *invariant* to which class label is being considered.

Yet, other solutions are possible. For example, under a *boosting* ensemble, the Z scores of the instances in a node are iteratively updated such that, at each iteration, instances in the same *node* may have different scores. In other words, to apply *SCAFF* within a boosting framework, it is only necessary to append its computation to an existing algorithm (such as the standard gradient boosting algorithm (Friedman, 2001)) which already determines the Z scores of the instances. At each split evaluation and instance score update, *SCAFF* would be computed and serve as the splitting criterion.

To make predictions, it is sensible that *only* $Z = P(Y_+ \mid node)$ is considered as traditionally high scores are associated with the positive class label. This approach trivially enables *bagging* ensembles (i.e., random forest), by computing the average Z across all weak learners. We further remark that this is the setup we will use in our upcoming experimental setup.

4.4 AUC computation

Traditionally, the computation of AUC has a time complexity $O(n \cdot \log(n))$ (Eq. 5). Yet, from Lee (2019), the AUC can re-formulated as a function of the true positive rate and the false positive rate, if there exist at most two unique Z values across the instances ($|\{z \in Z\}| \leq 2$). This is the case for the example $Z = P(Y_+ \mid node)$ or $Z = P(Y_- \mid node)$. The AUC function then becomes:

$$\mathrm{AUC}(Z, Y) = \frac{1 + \mathrm{TPR}(\hat{Y}, Y) - \mathrm{FPR}(\hat{Y}, Y)}{2}, \tag{12}$$

$$\hat{Y} = \begin{cases} \hat{Y}_+ \text{ if } Z \geq \max\{z \in Z\}, \\ \hat{Y}_- \text{ otherwise} \end{cases}, \tag{13}$$

$$\mathrm{TPR}(\hat{Y}, Y) = P(\hat{Y}_+ \mid Y_+) = \frac{P(\hat{Y}_+ \cap Y_+)}{P(Y_+)}, \text{ and} \tag{14}$$

$$\mathrm{FPR}(\hat{Y}, Y) = P(\hat{Y}_+ \mid Y_-) = \frac{P(\hat{Y}_+ \cap Y_-)}{P(Y_-)}. \tag{15}$$

These computations hold for $\mathrm{AUC}(Z, S)$ by replacing the class Y with sensitive attribute S, and defining $\hat{S}$ according to Z (Eq. 13).

4.5 Tree construction

During the iterative tree learning process, (*feature*, *value*) pairs are considered, which induce *child* nodes $node_L$ and $node_R$ from a *parent* node $node_P$. Given $node_P$ with instance scores Z, and $node_L$ and $node_R$ with instance scores Z', the SCAFF Gain (SG) associated with that split is defined as:

$$SG = \mathrm{SCAFF}(Z', Y, S, \Theta) - \mathrm{SCAFF}(Z, Y, S, \Theta). \tag{16}$$

In addition, as per Lee (2019), we incorporate into SG a normalisation component which is simply the information entropy of the frequency of instances in the child nodes with respect to the parent node. Let $|node|$ indicate the number of samples a node represents; naturally: $|node_P| = |node_L| + |node_R|$. The normalisation component *Split Info* (SI) of a candidate split is defined as:

$$SI = -\frac{|node_L|}{|node_P|} \cdot \log_2\left(\frac{|node_L|}{|node_P|}\right) - \frac{|node_R|}{|node_P|} \cdot \log_2\left(\frac{|node_R|}{|node_P|}\right). \tag{17}$$

Finally, we define the *SCAFF Gain Ratio* (*SGR*) of a split as:

$$SGR = \frac{SG}{SI} \tag{18}$$

The split with maximal *SGR* across all splits is selected if its corresponding $SGR \geq 0$. An example of SCAFF evaluation can be viewed in Fig. 1.

In the example, the *Z* scores are defined as $P(Y_+ \mid node)$, which (as previously mentioned) trivially enables *bagging* methods. For orthogonality $\Theta = 0.5$, then $\text{SCAFF}(Z', Y, S, \Theta) = (1 - 0.5) \cdot |2 \cdot 0.8 - 1| - 0.5 \cdot \max(|2 \cdot 0.6 - 1|, |2 \cdot 0.917 - 1|)$, $\text{SCAFF}(Z, Y, S, \Theta) = (1 - 0.5) \cdot |2 \cdot 0.5 - 1| - 0.5 \cdot \max(|2 \cdot 0.5 - 1|, |2 \cdot 0.5 - 1|)$, resulting in a $SG = -0.117 \log_2$. The evaluation of the regularisation component is $SI = -\frac{5}{10} \cdot \log_2\left(\frac{5}{10}\right) - \frac{5}{10} \cdot \log_2\left(\frac{5}{10}\right) = 1$. Finally, $SGR = \frac{-0.117}{1} = -0.117$. The reason for the poor *SGR* is clear: the split heavily segregated individuals based on *race*. Since the value of *SGR* is negative, the split is not selected.

5 Experiments

For the description of our experiments, we begin by mentioning the datasets and how we used them (Sect. 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 (Sect. 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 category-valued sensitive attributes, we focus on the single binary sensitive attribute case first. We additionally experimented on a single dataset

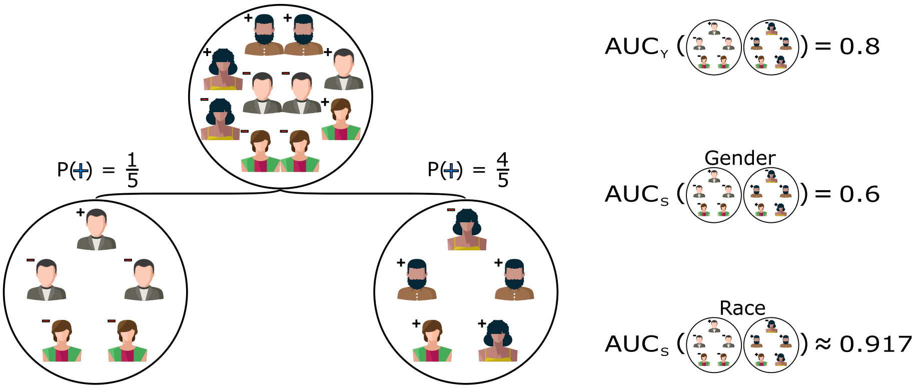


Fig. 1 Computing necessary AUC values for split evaluation, with 2 sensitive attributes (*race* and *gender*). Instances pertaining to a node are assigned its respective positive class probability as their *Z* scores, denoted above each child node

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).

5.1 Datasets

Two binary classification datasets were used, typically used in the fairness literature (Le Quy et al., 2022). Specifically, we employed the following: (a) *Adult* (30, 913 instances, 10 features), where the sensitive attribute may be either (1) *gender* $\in \{\text{male}, \text{female}\}$, (2) *race* $\in \{\text{black}, \text{white}\}$, or (3) the discretised *age* $\in \{\leq 24, 25\text{–}49, \geq 50\}$; and (b) *Bank* (45, 203 instances, 14 features) in which the sensitive attribute is the discretised *age* $\in \{\leq 21, 22\text{–}64, \geq 65\}$.

For the *binary* sensitive attribute case, we considered the following dataset-attribute pairs individually: *Adult* (*Gender*), and *Adult* (*Race*). With respect to the scenario of a *multicategorical* sensitive attribute, we considered *Bank (Age)*, and *Adult* (*Age*). Finally, for the scenario under which *multiple* sensitive attributes exist—binary or multicategorical—we considered simultaneously all sensitive attributes of *Adult* (*All*).

5.2 Experimental setup

To provide an adequate evaluation of our method, we considered (a) *traditional* –baseline– non-fairness-aware methods, and (b) previous work in fair splitting criteria by Kamiran et al. (2010). The purpose of the former is to establish control measurements for classification performance and fairness, while the latter serves as the benchmark upon which we wish to improve. To note, the fair splitting criteria methods from the literature could only be deployed upon the datasets of which the sensitive attributes were binary, as per their design.

Two baseline methods were used: (1) $\text{Baseline}_{\text{IG}}$, the standard tree learning algorithm based on *information gain* (Eq. 1); and (2) $\text{Baseline}_{\text{AUC}}$, the tree learning algorithm based on *AUC Gain Ratio* Lee (2019). To note, the second baseline method is equivalent to our SCAFF method when $\Theta = 0$. With respect to fairness-aware tree learners, three methods were implemented: (1) $\text{Kamiran}_{\text{Sum}}$; (2) $\text{Kamiran}_{\text{Sub}}$; and (3) $\text{Kamiran}_{\text{Div}}$. Each of these was deployed with the inclusion of the post-processing leaf relabelling step (Sect. 3.2).

To assess the tunability of the our method, a range of 9 values were used for orthogonality $\Theta \in [0.1, 0.9]$. Towards comparing this tunability with the fairness-aware method, 9 values for $\epsilon \in [0.1, 0.9]$ were used in the post-processing leaf relabelling (Sect. 3.2). Here, we note that $\Theta = 1 - \epsilon$ and that, for simplicity, we hereinafter consider tunability in terms of Θ.

We measured classification performance and fairness in both (1) threshold-independent and (2) threshold-dependent fashions. For the first, the measures of *AUC classification performance* and *strong demographic parity* were used. For the second, classification performance was measured via *macro* F_1-*score*—the unweighted average of the harmonic mean of the precision and recall across class labels—and fairness was measured via *demographic* *parity*. For threshold-dependent measures, 9 decision thresholds were selected based on a range of quantiles of Z.

We computed also the difference between performance and fairness as a function of each of the aforementioned measures. Specifically, prior to evaluating the difference, the performance and fairness measures were normalised to be within the same range of values [0, 1] by using the maximum and minimum values of the respective measures across all methods. This measure serves as a *normalised score* for the performance-fairness trade-off: large values of this score translate to *large fairness gains* at the cost of *small performance losses*.

To relate the threshold-dependent and threshold-independent demographic parities, we measured at each decision threshold—along Θ values—the Pearson correlation coefficient Pearson's correlation coefficient (2008), and the respective null hypothesis p-values, between *strong demographic parity* and *demographic parity*. The purpose is to check whether the behaviour of strong demographic parity across Θ transfers to that of the demographic parity.

For each dataset configuration, the same 10-fold cross validation was applied across all methods, and the means and standard deviations were recorded for performance and fairness. The classification scores Z of samples were computed as the proportion of positive samples of the terminal leaf node—$P(Y_+ \mid node)$, Sect. 4.3—of a single tree, as illustrated in Fig. 1.

To achieve state-of-the-art performance, all methods were deployed as a bagging ensemble (i.e., random forest) Breiman (2001). Accordingly, the final classification score of a sample is the average Z across all different trees. Bootstrapping, random feature selection, and continuous-feature discretisation were also applied, given their prevalence in real-world implementations of tree-based algorithms, such as XGBoost Chen and Guestrin (2016). For the full implementation details, see Pereira Barata (2021).

6 Results

In this section, we present the results of our experiments with benchmark fairness datasets. We invite the reader to access our results online Pereira Barata (2021), where the three-dimensional plots are interactive. We begin by reporting on the classification performance and fairness obtained across our method and the competing approaches for the binary sensitive attribute configurations (Sect. 6.1). We follow with the performance and fairness produced by our SCAFF approach for the multicategorical case (Sect. 6.2). In Sect. 6.3, we show the results of our method applied simultaneously to binary and multicategorical sensitive attributes. Lastly, we show how strong demographic parity relates to demographic parity across different decision thresholds (Sect. 6.4).

6.1 Binary sensitive attribute

We present the results pertaining to the *Adult* (*Race*) setup in Fig. 2, and the *Adult* (*Gender*) setup in Fig. 3. The left side of the figures correspond to the threshold-independent

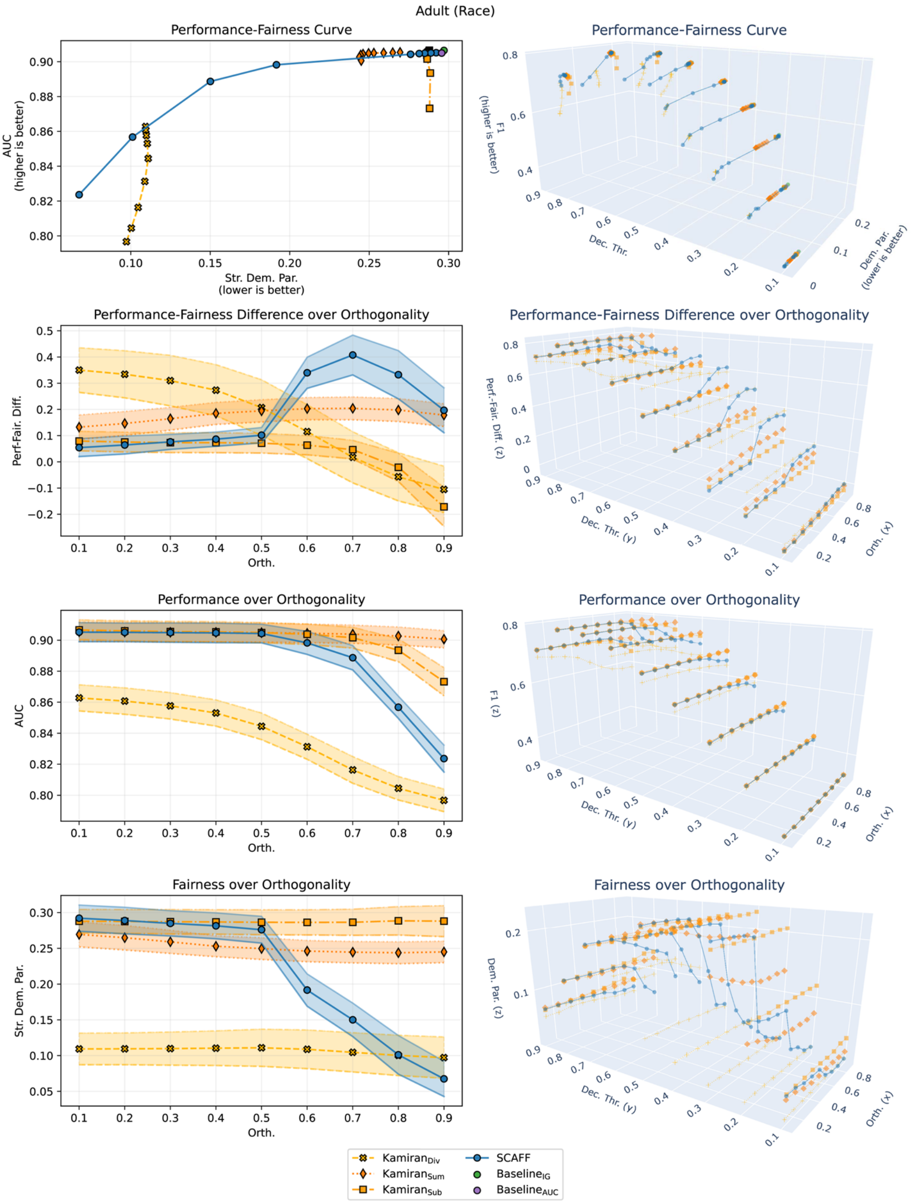


Fig. 2 Performance-fairness in the binary sensitive attribute setup *Adult* (*Race*)

measures (AUC and strong demographic parity), while the right side corresponds to the threshold-dependent measures (F_1 and demographic parity).

Overall, our method consistently performs better in the *combination* of classification performance and fairness, allowing for a suitable target point of orthogonality Θ. Explicitly, our method provides a larger coverage of the performance-fairness trade-off

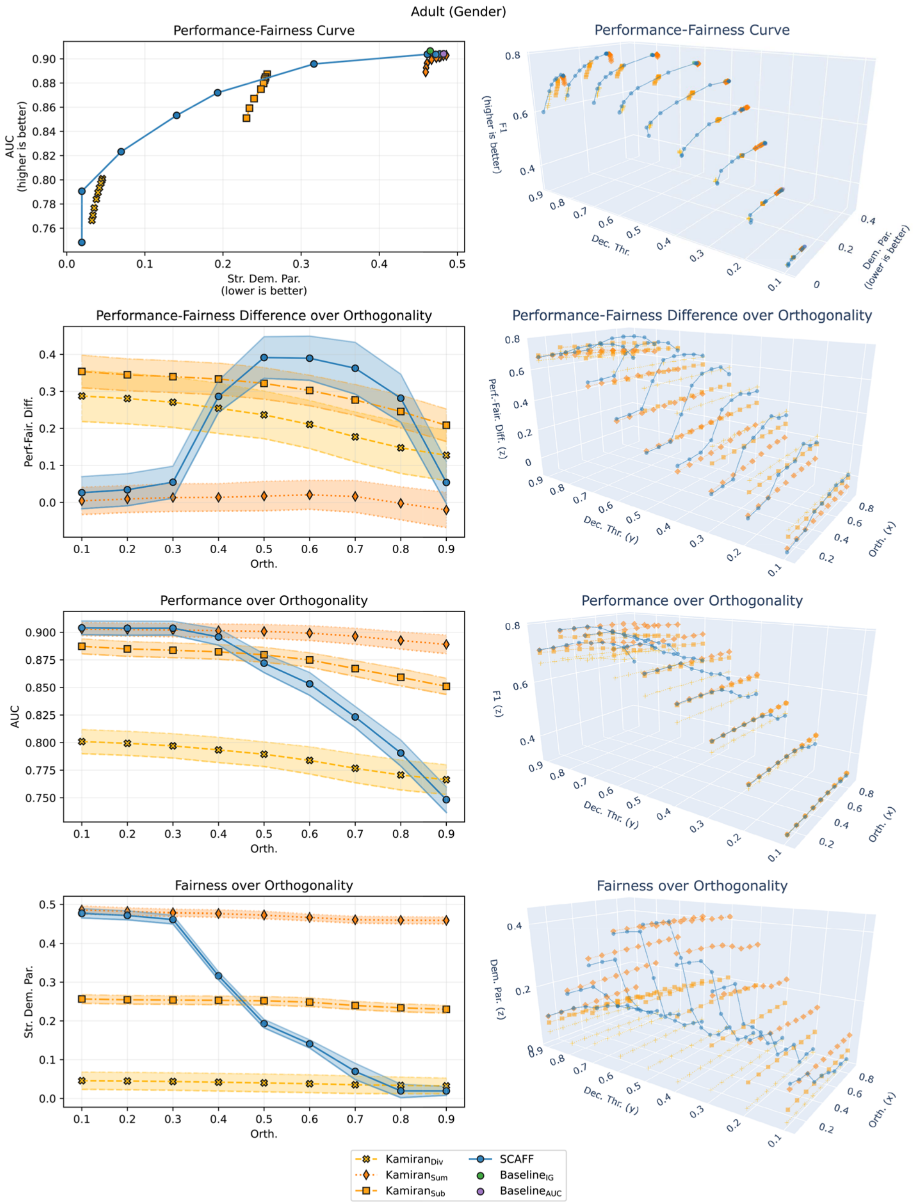


Fig. 3 Performance-fairness in the binary sensitive attribute setup *Adult* (*Gender*)

(see *performance-fairness curve*). Moreover, our method is able to achieve higher values of performance-fairness difference.

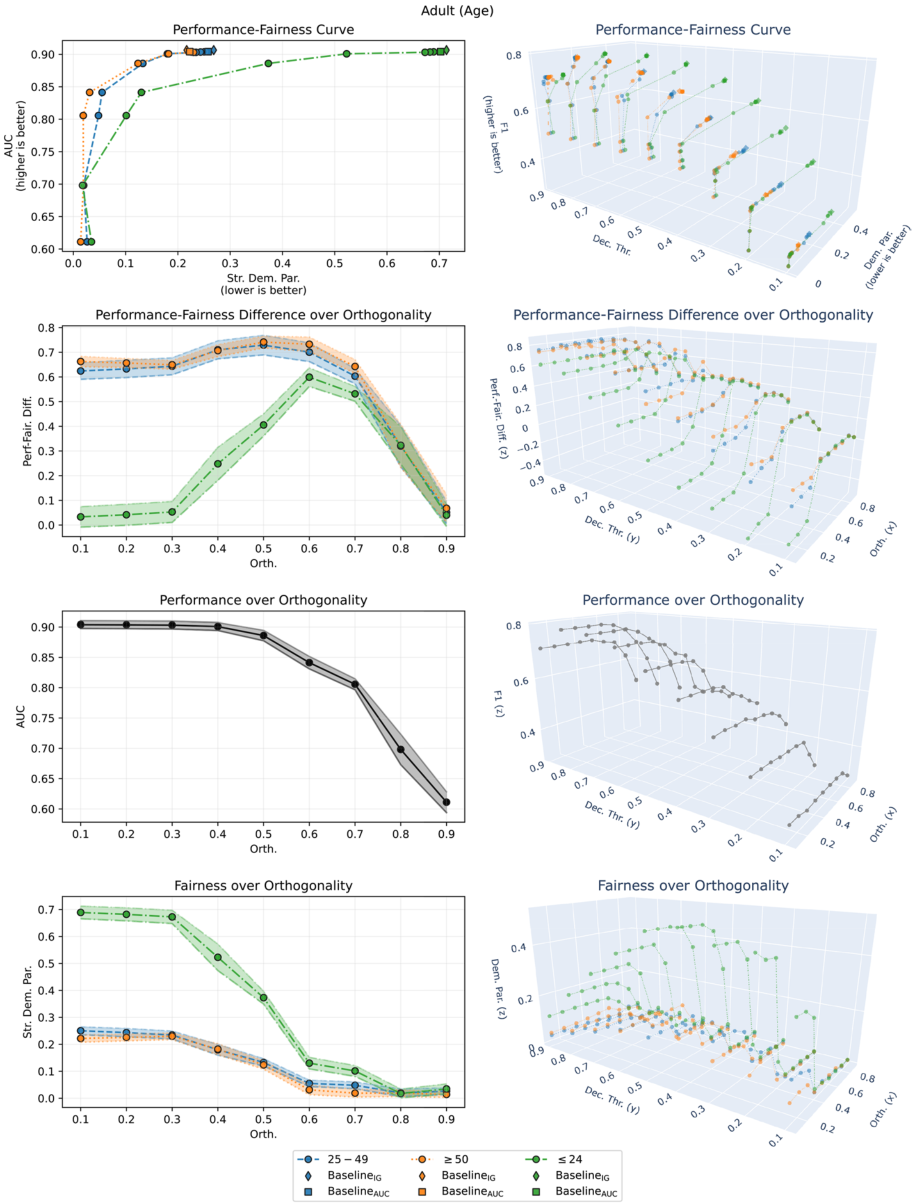


Fig. 4 Performance-fairness in the multicategorical sensitive attribute setup *Adult* (*Age*)

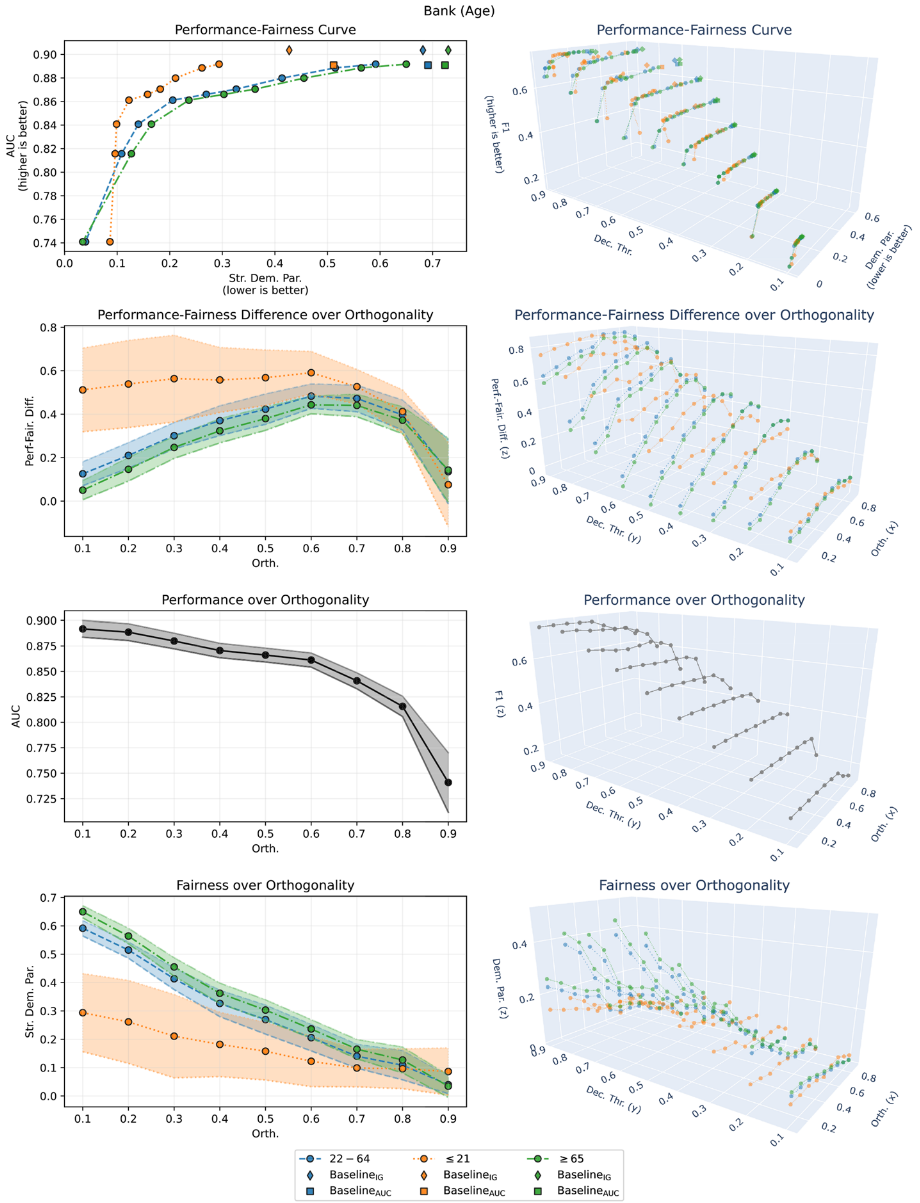


Fig. 5 Performance-fairness in the multicategorical sensitive attribute setup *Bank* (*Age*)

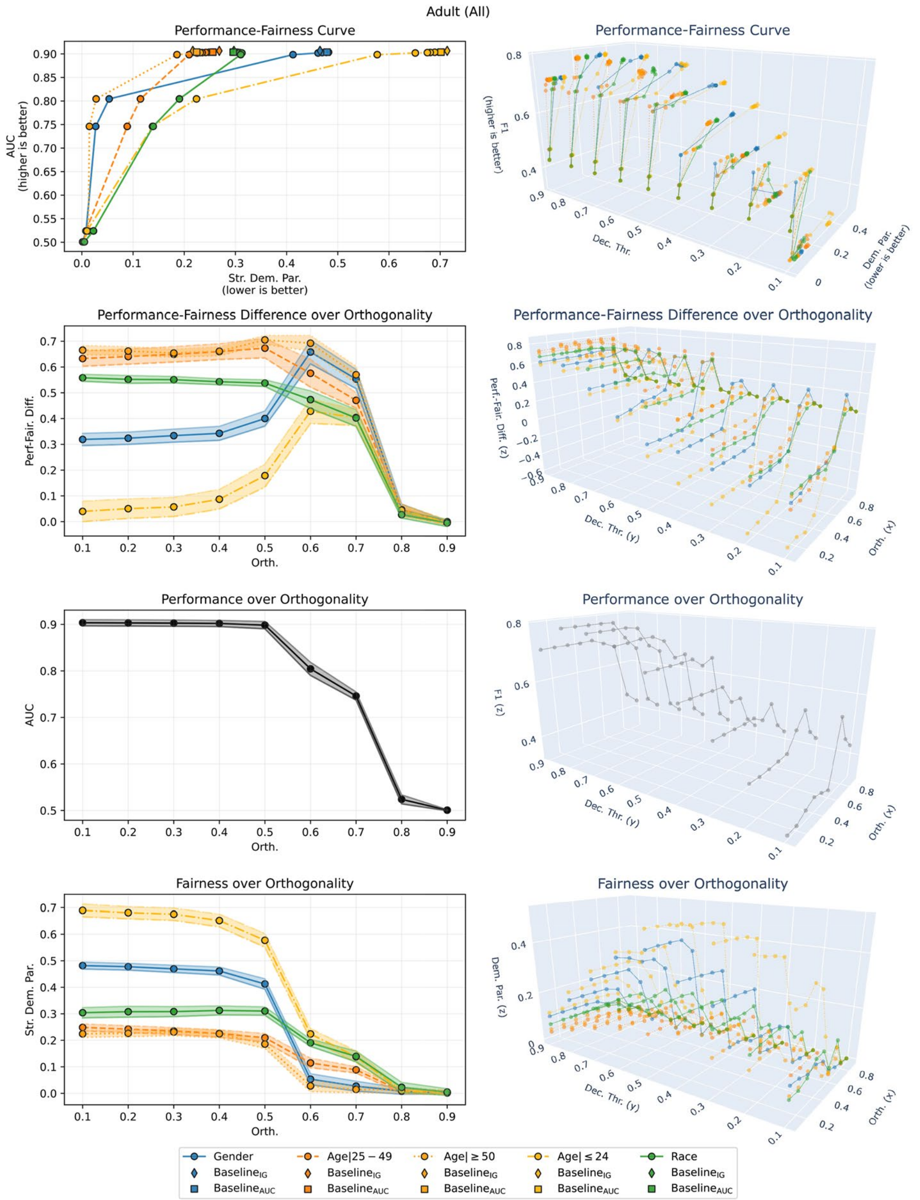
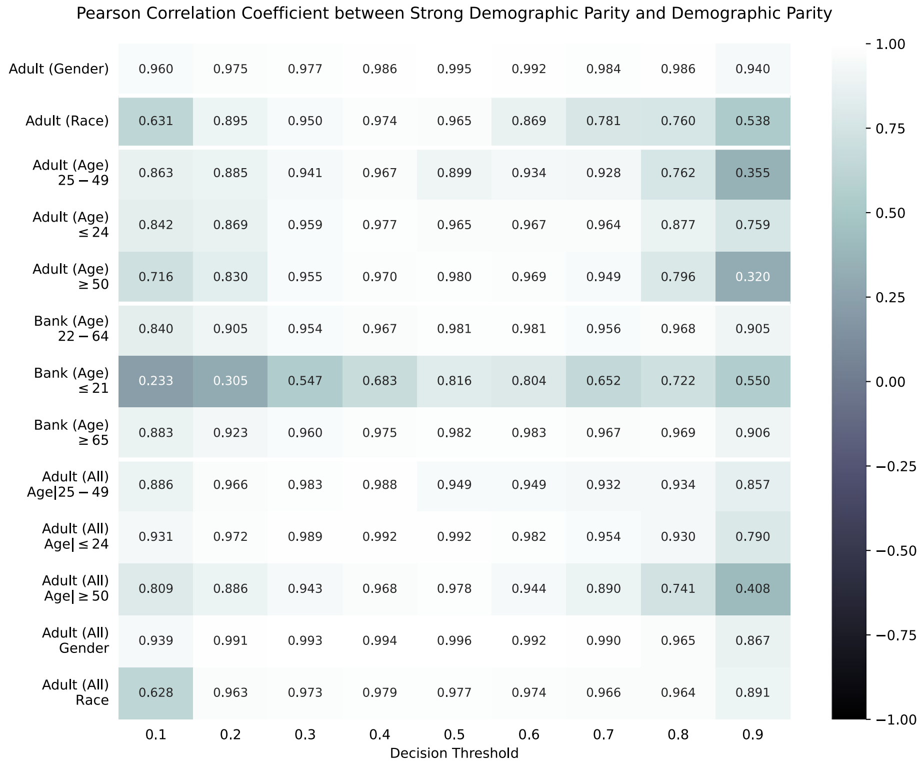


Fig. 6 Performance-fairness in the multiple sensitive attribute setup *Bank* (*All*)



Pearson Correlation Coefficient between Strong Demographic Parity and Demographic Parity

	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Adult (Gender)	0.960	0.975	0.977	0.986	0.995	0.992	0.984	0.986	0.940
Adult (Race)	0.631	0.895	0.950	0.974	0.965	0.869	0.781	0.760	0.538
Adult (Age) $25-49$	0.863	0.885	0.941	0.967	0.899	0.934	0.928	0.762	0.355
Adult (Age) ≤ 24	0.842	0.869	0.959	0.977	0.965	0.967	0.964	0.877	0.759
Adult (Age) ≥ 50	0.716	0.830	0.955	0.970	0.980	0.969	0.949	0.796	0.320
Bank (Age) $22-64$	0.840	0.905	0.954	0.967	0.981	0.981	0.956	0.968	0.905
Bank (Age) ≤ 21	0.233	0.305	0.547	0.683	0.816	0.804	0.652	0.722	0.550
Bank (Age) ≥ 65	0.883	0.923	0.960	0.975	0.982	0.983	0.967	0.969	0.906
Adult (All) Age$\|25-49$	0.886	0.966	0.983	0.988	0.949	0.949	0.932	0.934	0.857
Adult (All) Age$\| \leq 24$	0.931	0.972	0.989	0.992	0.992	0.982	0.954	0.930	0.790
Adult (All) Age$\| \geq 50$	0.809	0.886	0.943	0.968	0.978	0.944	0.890	0.741	0.408
Adult (All) Gender	0.939	0.991	0.993	0.994	0.996	0.992	0.990	0.965	0.867
Adult (All) Race	0.628	0.963	0.973	0.979	0.977	0.974	0.966	0.964	0.891

Fig. 7 Pearson correlation coefficient between strong demographic parity and demographic parity at different decision thresholds. Each cell corresponds to the correlation between the fairness measures over orthogonality Θ across all folds

6.2 Multicategorical sensitive attribute

We present the results pertaining to the *Adult* (*Race*) setup in Fig. 4, and the *Bank* (*Age*) setup in Fig. 5. The left side of the figures correspond to the threshold-independent measures (AUC and strong demographic parity), while the right side corresponds to the threshold-dependent measures (F_1 and demographic parity). Different colours indicate different *age* categories.

For both experimental setups, our method was able to simultaneously increase model fairness towards all categories, while maintaining adequate classification performance. Remarkably, SCAFF was able to achieve large values of performance-fairness difference (more pronounced for *Adult (Age)*, which translates to the method being able to generate models which are able to immensely increase fairness at a minimal cost of classification performance.

6.3 Multiple sensitive attributes

We present the results pertaining to the *Bank* (*All*) setup in Fig. 6. The left side of the figures correspond to the threshold-independent measures (AUC and strong demographic parity), while the right side corresponds to the threshold-dependent measures (F_1 and demographic parity).

For multiple binary and multicategorical sensitive attributes, SCAFF was able to generate adequately-fair classification models with respect to all sensitive attributes. We note here that, the greater the number of constraints the more difficult it is to minimise the performance detriment which accompanies the gain in fairness. Yet, the results are overall positive even within this difficult scenario.

6.4 Relationship with demographic parity

In Fig 7, the relationship between the threshold-dependent and threshold-independent fairness measures is made explicit. Each row depicts a decision threshold upon which demographic parity was computed, whereas a column indicates a dataset configuration. Accordingly, a cell depicts the Pearson correlation coefficient between the two measures of fairness along the parameter Θ, for a given decision threshold. The coefficients consider (a) the strong demographic parity and (b) its threshold-induced counterpart, as seen in the left and right *Fairness over Orthogonality* visualisations, respectively (Sects. 6.1–6.3).

The coefficients represent how similar the behaviour between threshold-dependent and - independent demographic parities is, induced by shifts in Θ. It is advantageous to maintain the behaviours similar, regardless of the selected threshold. Noteworthily, all statistical results tested significantly ($\alpha = 0.05$) against the null hypothesis of no correlation; i.e., there is statistical evidence indicative of positive correlation. This shows that the effect of shifting the orthogonality parameter Θ in our method, which optimises for the threshold-independent strong demographic parity, mostly carries over to the threshold-dependent demographic parity.

We remark, however, that the distribution of the correlation coefficients is not homogeneous. Namely, for extreme decision thresholds (e.g., 0.1 and 0.9), the correlation drops albeit *still positive* and of statistical significance. On a similar note, for *Bank* (*Age*) ≤ 21, the correlation coefficients were lower than those of other configurations. This is congruent with the (high) standard deviations in fairness shown in Fig. 5 which, in turn, are most probably caused by the low frequency of instances in which *age* ≤ 21; per test fold, only 17 instances have that sensitive attribute value. In other words, we attribute (a) the lower correlation values and (b) higher standard deviation of fairness to (c) the low frequency of samples exhibiting that specific attribute value.

7 Conclusion

In the present work, we introduced SCAFF: the Splitting Criterion AUC for Fairness. By doing so, we proposed a learning algorithm which simultaneously (1) optimises for threshold-independent ROC-AUC classification performance and threshold-independent *strong* demographic parity, (2) handles various multicategorical sensitive attributes simultaneously, (3) is tunable with respect to the performance-fairness trade-off, and (4) extends to ensemble methods.

We empirically validated our method through experimentation on benchmark datasets traditionally used in the fairness literature. Within our experiments with real datasets, we showed that our approach outperformed the competing state-of-the-art criteria methods, not only in terms of predictive performance and model fairness, but also 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 *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 should be the ulterior goal in this field of research.

Author contributions APBand CJV contributed to the study conception and design. Material preparation, data collection and analysis were performed by APB. The first draft of the manuscript was written by APB. All authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding Funding for this study was provided by the Ministry of Infrastructure and Water Management of the Netherlands.

Code availability The current version of the code is available at https://github.com/pereirabarataap/fair_tree_classifier.

Declarations

Conflict of interest The authors declare no Conflict of interest.

References

Azar, A. T., & El-Metwally, S. M. (2013). Decision tree classifiers for automated medical diagnosis. *Neural Computing and Applications, 23*(7), 2387–2403.

Barata, A. P., Takes, F. W., van den Herik, H. J., Veenman, C. J., & (2021). The eXPose approach to crosslier detection. In: *2020 25th international conference on pattern recognition (ICPR* (pp. 2312–2319).

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

Brink, H., Richards, J., & Fetherolf, M. (2016). *Real-world machine learning*. Simon and Schuster.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785–794). ACM. https://doi.org/10.1145/2939672.2939785

Cho, J., Hwang, G., & Suh, C. (2020). A fair classifier using kernel density estimation. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, & H. Lin (Eds.), *Advances in neural information processing systems* (Vol. 33, pp. 15088–15099). Curran Associates, Inc.

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

Dastian, J. (2018). Amazon scraps secret ai recruiting tool that showed bias against women. Reuters. https://www.reuters.com/article/ us-amazon-com-jobs-automation-in...-ai-recruiting-tool-that-showed -bias-against-women-idUSKCN1MK08G. Retrieved 21 April 2021 from https://www.reuters.com/article/us-amazon-com-jobs-automation-in...- ai-recruiting-tool-that-showed-bias-against-women-idUSKCN-1MK08G

Dogru, N., & Subasi, A. (2018). Traffic accident detection using random forest classifier. *2018 15th learning and technology conference (l &t)* (pp. 40–45).

Dressel, J., & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science Advances, 4*(1), eaao5580.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. *Proceedings of the 3rd innovations in theoretical computer science conference* (pp. 214–226).

European Commission (2019). *Proposal for a regulation on a European approach for artificial intelligence*. https://digital -strategy.ec.europa.eu/en/library/proposal-regulation-european -approach-artificial-intelligence. Retrieved 21, April 2021 from https://digital-strategy.ec.europa.eu/en/library/proposal-regulationeuropean- approach-artificial-intelligence

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of statistics, 29*, 1189–1232.

Hardt, M., Price, E., & Srebro, N. (2016). *Equality of opportunity in supervised learning*. arXiv preprint arXiv:1610.02413 .

Hu, Y., Wu, Y., Zhang, L., & Wu, X. (2020). Fair multiple decision making through soft interventions. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, & H. Lin (Eds.), *Advances in neural information processing systems* (Vol. 33, pp. 17965–17975). Curran Associates Inc.

Jiang, R., Pacchiano, A., Stepleton, T., Jiang, H., & Chiappa, S. (2020). Wasserstein fair classification. In L. N. Kanal & J. F. Lemmer (Eds.), *Uncertainty in artificial intelligence* (pp. 862–872). Elsevier.

Kamiran, F., Calders, T., & Pechenizkiy, M. (2010). Discrimination aware decision tree learning. In *2010 IEEE international conference on data mining* (pp. 869–874).

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). *Inherent trade-offs in the fair determination of risk scores*. arXiv preprint arXiv:1609.05807

Lee, J.-S. (2019). AUC4.5: AUC-based C4.5 decision tree algorithm for imbalanced data classification. *IEEE Access, 7*, 106034–106042.

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

Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics, 18*, 50–60.

Mason, S. J., & Graham, N. E. (2002). Areas beneath the relative operating characteristics (roc) and relative operating levels (rol) curves: Statistical significance and interpretation. *Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography, 128*(584), 2145–2166.

Pearson's correlation coefficient. (2008). In Kirch, W (Ed.), *Encyclopedia of public health* (pp. 1090–1091). Springer. https://doi.org/10.1007/978-1-4020-5614-7_2569

Pereira Barata, A. (2021). *Fair tree classifier*. https://github.com/ pereirabarataap/fair tree classifier. GitHub.

Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. *SN Computer Science, 2*(3), 1–21.

Zabihi, M., Rad, A.B., Katsaggelos, A.K., Kiranyaz, S., Narkilahti, S., & Gabbouj, M. (2017). Detection of atrial fibrillation in ecg hand-held devices using a random forest classifier. In *2017 computing in cardiology (cinc)* (pp. 1–4).

Zhang, W., Bifet, A., Zhang, X., Weiss, J.C., & Nejdl, W. (2021). Farf: A fair and adaptive random forests classifier. In *Advances in knowledge discovery and data mining: 25th pacific-asia conference, pakdd 2021, virtual event, may 11–14, 2021, proceedings, part ii* (pp. 245–256).

Zhang, W., & Weiss, J.C. (2022). Longitudinal fairness with censorship. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 36, pp. 12235–12243).

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.