arXiv:2110.09295v1 [cs.LG] 18 Oct 2021

Fair Tree Learning

António Pereira Barata
Leiden Institute of Advanced Computer Science
Leiden University, The Netherlands
a.p.pereira.barata@liacs.leidenuniv.nl

Cor J. Veenman
Leiden Institute of Advanced Computer Science
Leiden University, The Netherlands

October 19, 2021

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 some sensitive attribute, like gender or race, induced from biased data. Various hybrid optimisation criteria exist which combine classification performance with a fairness metric. However, while the threshold-free ROC-AUC is the standard for measuring traditional classification model performance, current fair decision tree methods only optimise for a fixed threshold on both the classification task as well as the fairness metric. Moreover, current tree learning frameworks do not allow for fair treatment with respect to multiple categories or multiple sensitive attributes. Lastly, the end-users of a fair model should be able to balance fairness and classification performance according to their specific ethical, legal, and societal needs. In this paper we address these shortcomings by proposing a threshold-independent fairness metric termed *uniform demographic parity*, and a derived splitting criterion entitled *SCAFF* —Splitting Criterion AUC for Fairness— towards fair decision tree learning, which extends to bagged and boosted frameworks. Compared to the state-of-the-art, our method provides three main advantages: (1) classifier performance and fairness are defined continuously instead of relying upon an, often arbitrary, decision threshold; (2) it leverages multiple sensitive attributes simultaneously, of which the values may be multicategorical; and (3) the unavoidable performance-fairness trade-off is tunable during learning by means of a configurable hyperparameter. In our experiments with real-world datasets, we demonstrate how SCAFF attains high predictive performance towards the class label and low discrimination with respect to binary, multicategorical, and multiple sensitive attributes, further substantiating our claims. Ultimately, SCAFF allows for the learning of fairness-tunable classification models, promoting fair real-world automated data-driven decision-making.

1 Introduction

The application of machine learning algorithms for classification has become ubiquitous within an abundance of domains [1–3]. Great dependency on automated decision-making, however, gives rise to critical concerns over model discrimination; for instance, when dealing with crime prevention it was shown that the automated *Correctional Offender Management Profiling for Alternative Sanctions* (COMPAS) system was biased towards race [4]. Another phenomenon of bias was reported by Amazon's automatic recruitment tool in which, since models were trained on resumes submitted mostly by men, women unfairly scored lower [5]. As a final example, in the Netherlands, over ten thousand applications for child-care support were wrongly listed as fraudulent due to individuals having double nationality [6], incurring a severe negative socio-economic impact on the misjudged persons and their families. To prevent similar events from re-occurring, it is of the utmost importance to both develop and adequately deploy fairness-aware methods [7].

The objective of a fair classification algorithm is to learn to make decisions which are independent (or conditionally independent) of a sensitive attribute [8, 9]; e.g., by maintaining an equal proportions of positive label predictions between male and female individuals, termed *demographic parity* [10]. Note that a sensitive attribute does not need to be binary, and several sensitive attributes may be simultaneously considered. The *performance-fairness trade-off* —the phenomenon in which the lesser the fairness of an algorithm, the greater its predictive capabilities and vice-versa [11]— should be tunable to satisfy the practitioner's ethical, legal, and societal needs. Moreover, the fairness of a model should not depend on the decision threshold upon which the output of the classifier returns a positive or negative class prediction. The domain expert should be able to decide on the compromise between fairness and classification performance prior to model deployment without compromising the target statistics of the final label predictions.

Most fair classification algorithms require differentiable surrogate loss functions to approximate the target measures of performance and bias since they are not easily differentiable themselves [12, 13]. Yet, tree-based classifiers are not hindered by this requirement: a splitting criterion can be arbitrarily defined so long as it is computationally tractable. From the field of traditional classification, tree-based algorithms are, despite their simplicity, still regarded as state-of-the-art solutions in many domains [14–16]. The prevalence of tree-based approaches in the literature is mostly due to (1) model interpretability, (2) their tendency to not overfit when used as ensembles, (3) requiring little to no data pre-processing, and (4) easily handling mixed data types and missingness [15].

In this work, we aim at delivering on a fair splitting criterion which allows for fair tree learning, whilst exploiting the performance-fairness trade-off in a tunable manner. We introduce SCAFF: Splitting Criterion AUC for Fairness. By doing so, we propose the following two contributions:

- a novel decision threshold-independent fairness measure: *uniform demographic parity*;

- a *fair tree learning* algorithm which:
  - handles various multicategorical sensitive attributes simultaneously; and
  - is tunable with respect to the performance-fairness trade-off during learning.

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

2 Problem Statement

We consider the scenario in which a labelled dataset is intrinsically biased with respect to some binary sensitive attribute, such as gender. Our task is to learn a fair predictive model from the biased data, such that future predictions are independent from the sensitive attribute. We require that the definition of model fairness does not depend on a decision threshold set upon the output. In addition, multiple sensitive attributes must be handled simultaneously and each sensitive attribute may be multicategorically valued. Since there is no unique solution in the trade-off between classification performance and fairness, the fair classification model must be tunable in this regard.

Formally, consider a dataset D with n samples, m features, a binary class, and a binary sensitive attribute. Let X be a $n \times m$ matrix representing all samples and their respective feature values. Also, let Y and S both be n-length vectors which represent the binary class and the binary sensitive attribute of the samples, respectively. The goal is to learn from inputs X, Y, and S, a mapping function $f : X \mapsto Z$, where $Z \in \mathbb{R}$ and upon which a decision threshold $t \in \mathbb{R}$ induces Y. While the learning of the target function is done with input X, Y, and S, the learned function only takes input X. The classifier function f should be learned such that $f(X|Y) \approx Y$ given some threshold t (i.e., the classification objective) while also holding $f(X|S) \approx f(X)$ (i.e., the bias/fairness objective). Accordingly, let now $f_Y : (Z, Y) \mapsto \mathbb{R}$ and $f_S : (Z, S) \mapsto \mathbb{R}$ be classification performance and bias-measuring functions, respectively. Assume also that the relationship between them is positive; i.e., higher values of performance translate to higher values of model bias and the converse holds for lower values of model fairness. The output of f_Y and f_S should not change, regardless of the decision threshold t selected to infer Y. The fair classification algorithm must also allow for the tuning of the trade-off between the terms of performance f_Y and fairness f_S. The problem statement is to find a suitable threshold-free fairness measure and a corresponding compound measure for fair tree learning, which allows for the tuning of the performance-fairness trade-off.

3 Related Work

In this section, we introduce the concepts from the fair machine learning literature which relate to our problem statement. We discuss the typically-used metrics for measuring fairness (Sec. 3.1), and the known fair tree learning methods in the literature (Sec. 3.2).

3.1 Measures of Fairness

The purpose of a fairness measure is to relate, in a meaningful way, the label predictions of a trained classifier to the observed sensitive attribute values of the instances to be classified; e.g., the difference in number of job positions granted to men versus women. Several measures exist in the literature, of which the three most prevalent are: (1) demographic parity [17]; (2) equal opportunity [18]; and (3) equalised odds [19]. First, demographic parity 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_- and is formally given as $|P(\hat{Y}_+|S_+) - P(\hat{Y}_+|S_-)|$. Second, the measure of equal opportunity is computed by taking the absolute difference of either the true positive rate (TPR) or the false positive rate (FPR) between the instance groups composed of the positive and negative sensitive attribute values. For the TPR case, it follows $|P(\hat{Y}_+|S_+, Y_+) - P(\hat{Y}_+|S_-, Y_+)|$, while the FPR case is defined as $|P(\hat{Y}_+|S_+, Y_-) - P(\hat{Y}_+|S_-, Y_-)|$. Third, equalised odds is defined as the absolute difference between the equal opportunity with respect to the TPR and the equal opportunity with respect to the FPR. We note that for tree frameworks specifically, the common measures of fairness used during learning are demographic parity [20], and entropy with respect to the sensitive attribute [21]. The key limitation of all these measures is that their values are dependent on the decision threshold placed upon the trained model output Z. We advocate that model fairness should hold, regardless of the decision threshold.

3.2 Fair Tree Splitting Criteria

An important 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. Moreover, their performance provides state-of-the-art solutions to many classification problems [14–16]. Within the fair tree literature, we recommend the works by Kamiran [21] and Zhang [20], in which different splitting criteria were used to leverage classification performance and fairness.

The work by Kamiran extends the concept of information gain in traditional classification towards the sensitive attribute. Each potential feature-value split is evaluated in terms of the conventional information gain with respect to the class label IG_Y and the information gain with respect to the sensitive attribute IG_S. 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 $\text{Kamiran}_{\text{Div}}$.

In their work, Zhang et al. propose FAHT: a fairness-aware Hoeffding tree. Although the method was developed with online streaming classification as its focus, the splitting criterion developed may be generalisable. Both works use the same class label information gain IG_Y. However, Zhang et al. define the fairness gain FG of a split as a function of the observed demographic parity of the system. These approaches present, nevertheless, some limitations of which we list three: (1) only a single binary sensitive attribute may be considered; (2) there exists no performance-fairness trade-off tuning parameter built into the splitting criteria; and (3) the construction process only allows for threshold-based criteria.

4 Method

In this section we describe out proposed method. It is a probabilistic tree learning framework which uses a proposed threshold-free fairness metric and a derived novel compound threshold-independent splitting criterion. Without loss of generality, we describe our method within the fair learning problem class with a single binary sensitive attribute. We begin by addressing our fairness measure in Sec. 4.1. In Sec. 4.2, we provide our compound splitting criterion which incorporates a tunable parameter towards the trade-off between classification performance and fairness. In Sec. 4.3, we describe the tree construction process, reporting on how our method is extends towards the multivariate and multicategorically valued sensitive attribute scenario. A working Python implementation of our algorithm can be found in [22].

4.1 Uniform Demographic Parity

We design our fairness measure with respect to classification threshold-independence. The demographic parity measure aims to minimise the difference in candidates from the sensitive groups among the selected candidates. We propose to impose this criterion without a decision threshold on Z towards $Y \in \{+, -\}$. In other words, our aim is to use the classifier scores $Z \in \mathbb{R}$—as defined in Sec. 2— to quantify the amount of fairness of that classifier with respect to the sensitive attribute S. Regardless of the threshold on Z, it always requires equal proportions of the demographic groups. We call this the *uniform demographic parity* condition. A method satisfies the uniform demographic parity condition if and only if for all thresholds on Z (resulting in $Y+$ and $Y-$) it satisfies the demographic parity condition. Formally, for decision threshold t, it is given as:

$$\forall t \in \mathbb{R} : P(Z \leq t|S_+) = P(Z \leq t|S_-). \tag{1}$$

By retroceding the classifier scores Z and the imposed threshold t into the function f with input X, the uniform demographic parity condition can be rewritten as:

$$P[f(X|S_+) > f(X|S_-)] = P[f(X|S_+) < f(X|S_-)]. \tag{2}$$

In practice, this translates to finding the function f which minimises the difference between the two terms in the previous expression; i,.e., a function which is not able to discriminate samples S_+ and S_- given their respective X. Formally:

$$\arg\min_f |P[f(X|S_+) > f(X|S_-)] - P[f(X|S_+) < f(X|S_-)]|. \quad (3)$$

This amounts to finding the function f which randomly orders the samples from the different sensitive groups. In machine learning, the ROC-AUC (hereinafter, AUC) is a measure which expresses the quality of a sample ordering with respect to a binary label, where a random order results in AUC$= 0.5$. Accordingly, we could find the fair classifier f by optimizing for an AUC value of 0.5 on the sensitive attribute. In order to solve the optimisation problem, we aim to minimise the AUC with S_+ as positive class, which we call AUC_{S_+}. Since $\mathrm{AUC}_{S_+} = 0$ is also maximally unfair, we define *sensitive AUC* ($\mathrm{AUC_S}$) — the bias-measuring function f_S from Sec. 2— as $\max(\mathrm{AUC}_{S_+}, \mathrm{AUC}_{S_-})$, where $0.5 \leq \mathrm{AUC_S} \leq 1$. A classifier f which satisfies the uniform demographic parity condition has $\mathrm{AUC_S} = 0.5$ and a completely unfair classifier has $\mathrm{AUC_S} = 1$.

4.2 Splitting Criterion AUC for Fairness

Now that a fairness measure has been defined, a classification performance measure is needed to build a compound optimisation criterion. In addition to also requiring threshold-independence, the target performance measure should be in the same range, such that tuning the performance-fairness trade-off becomes intuitive. In other words, the range of possible values of each measure and their interpretation should be similar, making for a sensible compound splitting criterion. For that reason we introduce as the performance measure, the standard classification AUC metric towards the positive class [23], subsequently termed $\mathrm{AUC_Y}$ —the classification performance-measuring functionf_Y in Sec. 2— which is computationally tractable with complexity $O(n \cdot \log(n))$. With both measures defined, we can integrate them into a single expression. The objective becomes finding a split which maximises $\mathrm{AUC_Y}$ (towards $\mathrm{AUC_Y} = 1$), while minimising $\mathrm{AUC_S}$ (towards $\mathrm{AUC_S} = 0.5$). To allow for the tunability of the trade-off between these terms, we propose an *orthogonality* parameter $\Theta \in [0, 1]$ which we incorporate into our splitting criterion. The Θ parameter regulates the splitting criterion score towards either classification performance ($\Theta = 0$) or fairness ($\Theta = 1$). 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:

$$\mathrm{SCAFF}(Z, Y, S, \Theta) = (1 - \Theta) \cdot \mathrm{AUC_Y}(Z, Y) - \Theta \cdot \mathrm{AUC_S}(Z, S). \quad (4)$$

4.3 Tree Construction

As with any typical tree architecture, learning is done by finding, 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 children nodes and is evaluated according to the Z scores of the parent node and the new Z' scores of the children 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 parent node scores Z and children scores Z' induced by a split, the SCAFF Gain (SG) associated with that split is defined as:

$$SG = \mathrm{SCAFF}(Z', Y, S, \Theta) - \mathrm{SCAFF}(Z, Y, S, \Theta). \tag{5}$$

The split with maximal SG across all evaluated splits is selected if and only if its corresponding $SG > 0$. If the condition is not met, no splitting occurs and the parent node becomes a leaf node. The process of assigning instance scores Z and computing the AUC values necessary to evaluate a split according to the SCAFF method can be viewed in the mock example provided in Fig. 1. For simplicity, assume the classifier score Z of an instance is defined as the probability $P(Y_+)$ within its current node. Trivially, a node has only one unique value of Z assigned to all its instances, which results in $\mathrm{AUC_Y}$ and $\mathrm{AUC_S}$ both equal to 0.5.

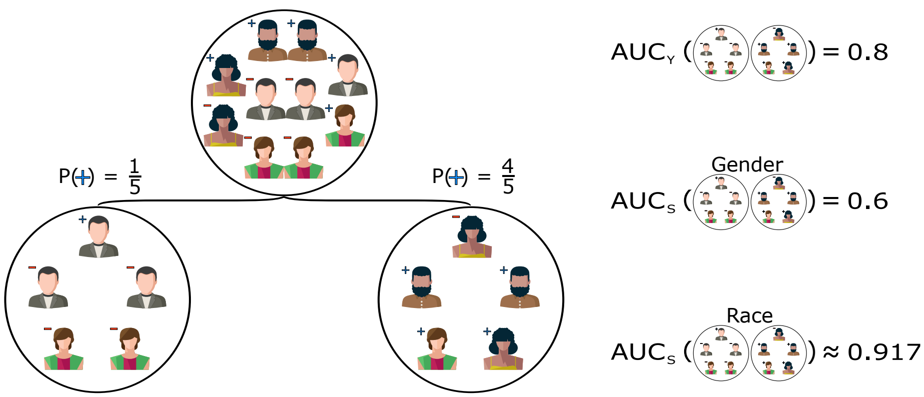

Figure 1: Computing necessary AUC values for split evaluation. Instances pertaining to a node are assigned its respective probability as their Z scores. Considering orthogonality $\Theta = 0.5$, then $SG = (0.5 \cdot 0.8 - 0.5 \cdot 0.6) - (0.5 \cdot 0.5 - 0.5 \cdot 0.5) = 0.1$ towards gender and, for race it follows $SG = (0.5 \cdot 0.8 - 0.5 \cdot 0.917) - (0.5 \cdot 0.5 - 0.5 \cdot 0.5) \approx -0.059$.

While we mention that Z scores are defined as $P(Y_+)$ in a node, enabling bagging, other definitions are also viable. For example, (gradient) boosted techniques compute Z by iteratively updating existing sample scores [24]. Our method is extendable to this a scenario since SG relies on Z, regardless of its computation, whereas traditional fair tree learning algorithms cannot: their splitting criteria cannot handle Z scores. In addition, SCAFF extends to multivariate and multicategorical sensitive attributes, including intersectional factors (i.e., the combination of different sensitive attributes) [25]. We propose a one-versus-rest (OvR) approach [26] to compute the $\mathrm{AUC_S}$ of all sensitive attribute values. The $\mathrm{AUC_S}$ to be used in the splitting criterion score is the maximum $\mathrm{AUC_S}$ across all OvR: since no sensitive attribute should have priority over fairness, the maximum bias should be considered. Following from the previous example in Fig. 1, the OvR $\mathrm{AUC_S} = \max(0.6, 0.917) = 0.917$.

5 Experiments

In this section we describe the experiments performed to evaluate the introduced fairness measure and compound splitting criterion for fair tree learning. We begin by describing the datasets and how we used them in our experiments (Sec. 5.1); we then characterise the experimental setup deployed to gather the performance and fairness values and report on the relationship between our threshold-independent performance and fairness measures and the traditionally-used threshold-dependent ones (Sec. 5.2). We compared SCAFF against other fair splitting criteria in a controlled scenario by using real-world benchmark fairness datasets. Under this scenario, 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. 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 relationship between the threshold-independent measures upon which our method relies, and the typically used decision threshold-based performance and fairness metrics. For reproducibility, our experiments are made available in [22].

5.1 Datasets

Three binary classification datasets were used which have at least one sensitive attribute. Specifically, we employed the following: (a) *Bank* ($45,211$ instances, 50 features), in which the classification goal is to predict if a client will subscribe a term deposit, and the sensitive attribute is the binary condition of age ≥ 65; (b) *Adult* ($45,222$ instances, 97 features), with the goal of predicting if the income of an individual $\geq 50,000$ per year, and the sensitive attribute may be either (a) race $\in \{\text{white}, \text{non-white}\}$ or (b) gender $\in \{\text{male}, \text{female}\}$; and (c) *Recidivism* (6150 instances, 8 features), in which the classification task is to predict whether a convicted criminal will reoffend based on their criminal history, and the sensitive attributes may be either (a) race $\in \{\text{white}, \text{non-white}\}$ or (b) gender $\in \{\text{male}, \text{female}\}$.

For the single binary sensitive attribute case, we considered each dataset-sensitive attribute configuration separately, making for a total of five different dataset configurations: *Bank (Age)*, *Adult (Race)*, *Adult (Gender)*, *Recidivism (Race)*, and *Recidivism (Gender)*. Two scenarios were further set in which the *Adult* dataset was considered: (1) the multiple sensitive attribute scenario —hereinafter *Adult (Multiple)*— such that both sensitive attributes (race and gender) were handled simultaneously; and (2) the multicategorical sensitive attribute scenario —hereinafter *Adult (Intersectional)*— such that the sensitive attribute $S \in \{\text{non-white female}, \text{non-white male}, \text{white female}, \text{white male}\}$ was concurrently handled. In total, seven different dataset configurations were generated.

5.2 Experimental Setup

To provide an adequate comparison between our splitting criterion and the state-of-the-art, we look at previous works in fair tree learning. Specifically, we consider the splitting critera proposed by Kamiran et al. [21] and Zhang et al. [20], on which we elaborated in Sec. 3.2. For each dataset configuration, 10-fold cross validation was applied. All methods were provided the same training sets for learning, and the resulting classification outputs were gathered from the respective test sets. To measure classification performance and algorithm fairness, $\mathrm{AUC_Y}$ (the accepted standard measure for classifier performance) and $\mathrm{AUC_S}$ were used. In line with our argumentation for using $\mathrm{AUC_S}$ as a fairness measure in our splitting criterion, we use it to measure the (un)fairness of the learned classifier. 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. Each tree was constructed following the splitting criterion of each corresponding method. The Z scores for our method were computed as the $P(Y_+)$ in each node (see mock example, Fig. 1 in Sec. 4.3). For all methods, the classification scores of samples were computed as the $P(Y_+)$ of the terminal leaf node. To simulate a state-of-the-art fair classification scenario, each method was deployed as a random forest [27]. Across 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. For our method in particular, a range of 11 values for orthogonality Θ was used between 0 to 1. For the experimental implementation, see [22].

To show-case how threshold-independent measures compare to the traditionally-used threshold-dependent ones, different decision thresholds were applied to the classifier outputs of our method across different values of Θ for the *Recidivism* dataset with respect to gender. The decision thresholds were considered as quantiles of each test set output. The quantiles were considered as a range of 9 values between 0.1 and 0.9. At each threshold, the predictive accuracy and the fairness metrics mentioned in Sec. 3.1 —demographic parity, equal opportunity (TPR and FPR), and equalised odds— were computed. We then measured at each decision threshold —along Θ values— the Pearson correlation coefficient [28], and the respective null hypothesis p-values, between $\mathrm{AUC_Y}$ and accuracy, and $\mathrm{AUC_S}$ and each of the fairness measures.

6 Results

In this section, we present the results of our experiments. We begin by reporting on the classification performance and fairness obtained across our method and the competing state-of-the-art approaches towards fair tree learning for the binary sensitive attribute configurations (Sec. 6.1). We follow with the performance and fairness for the non-binary case (Sec. 6.2). Finally, we show how the $\mathrm{AUC_Y}$ classification performance and novel uniform demographic parity (expressed as $\mathrm{AUC_S}$) correlate to the commonly-used and threshold-dependent accuracy and fairness metrics (Sec. 6.3).

6.1 Binary Sensitive Attribute

To analyse the trade-off between performance and fairness of our model across different values of orthogonality Θ, regard the left panels of Fig. 2. Along increasing values of orthogonality Θ (horizontal axis), the average classification performance tends do decrease ($\mathrm{AUC_Y}$ marked by upward-facing triangles in the vertical axis), as expected given the performance-fairness trade-off. For completeness, we also include the standard deviations for each measure, denoted as the upper and lower fills of the graph. Similarly, the average uniform demographic parity ($\mathrm{AUC_S}$ marked as downward-facing triangles in the vertical axis) also tends to decrease as the orthogonality increases. Naturally, a value of $\Theta = 0$ corresponds to the traditional classification problem (with no regard for the sensitive attribute), while $\Theta = 1$ translates to a completely fair classifier.

Albeit differently-valued, the performance-fairness trade-off for each dataset-sensitive attribute pair (denoted at the top left of each graph) is consistent: the greater the fairness (smaller values for $\mathrm{AUC_S}$), resulting from increasingly greater values of Θ, the lesser its classification performance. To compare our method to the state-of-the-art approaches, we focus on the right panels of the figure. In the horizontal axis, bias is represented as $\mathrm{AUC_S}$, while the vertical axis depicts the $\mathrm{AUC_Y}$ classification performance.

With respect to our method, each point is depicting a specific orthogonality value Θ, corresponding to the left-hand side. The competing methods are shown as idividual circles Note that, unlike the other methods which output a single performance-fairness value (represented as a point), our SCAFF method is able to produce a performance-fairness curve. This is advantageous as it provides a way for domain experts to make an informed decision which suits their requirements. The optimal fair classification solution should be to the top-left: *top* indicating high predictive performance, and *left* indicating low bias towards the sensitive attribute (in which a value of 0.5 indicates a perfectly un-biased (or conversely, completely fair) classifier. It is, therefore, noticeable how our method outperforms the competing approaches. For example, the *Bank (Age)* dataset configuration shows our method outperforming in both classification performance and bias, against all competing methods. It is a convincing result of (1) the use of AUC in the splitting criterion and (2) the flexibility of the Θ parameter. We also note the following for the *Bank (Age)* dataset; for the trade-off between performance and fairness, our algorithm was able to achieve a drop in bias of circa 0.3 $\mathrm{AUC_S}$ at the expense of only 0.1 $\mathrm{AUC_Y}$ at the value of $\Theta = 0.9$. Overall, our method consistently performs better in the combination of classification performance and fairness, allowing for a suitable target point.

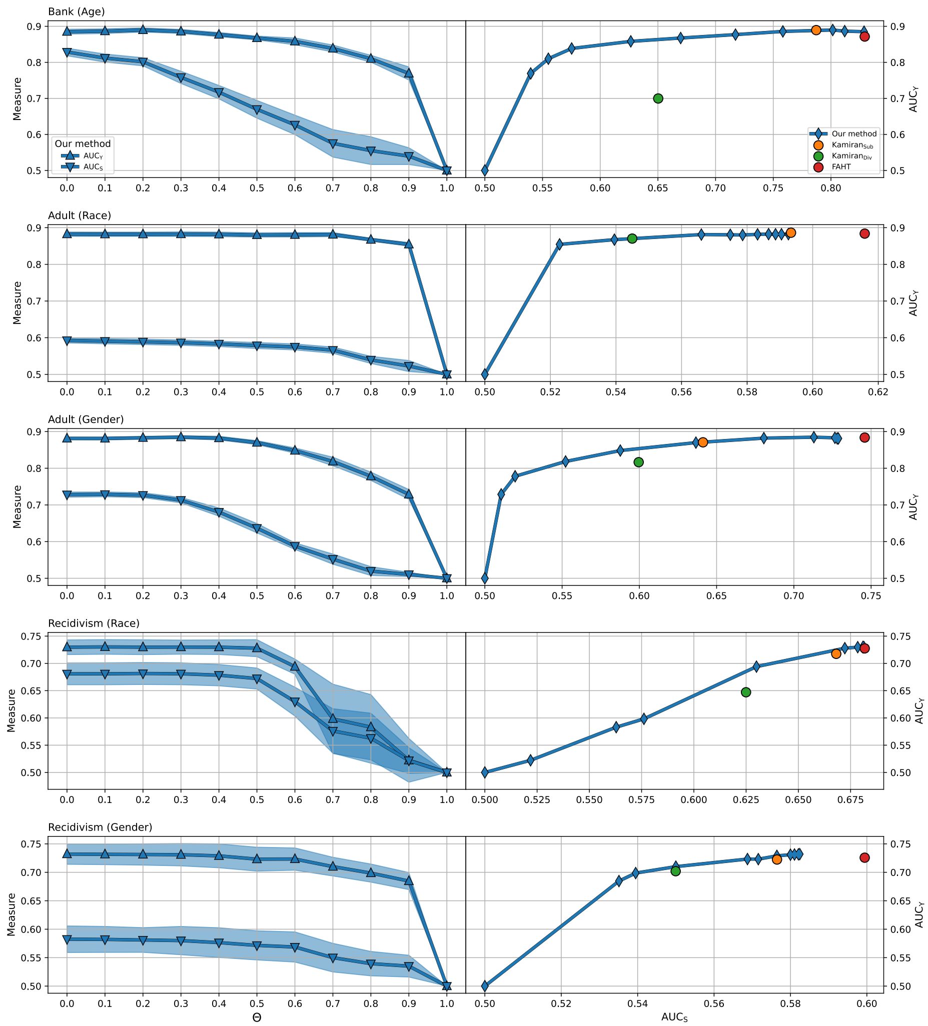

Figure 2: Model performance and fairness across methods, per dataset-sensitive attribute configuration. Each dataset and sensitive attribute is framed at the top left of each graph. Left: average and standard deviation values of $\mathrm{AUC_Y}$ and $\mathrm{AUC_S}$ values (vertical axis) across different values of orthogonality Θ (horizontal axis) for our method. Right: measures of $\mathrm{AUC_Y}$ (vertical axis) and $\mathrm{AUC_S}$ (horizontal axis) for all methods, including the different values of Θ corresponding to the left panel.

6.2 Multivariate and Multicategorical Sensitive Attributes

We present in Fig. 3 the outcomes of the dataset configurations *Adult (Multiple)* in the left panel —multiple sensitive attributes— and *Adult (Intersectional)* in the right panel —multicategorical sensitive attribute values. For both panels, the classification performance $\mathrm{AUC_Y}$ is shown in blue, and the different $\mathrm{AUC_S}$ are provided, across different values of orthogonality Θ. To the left, the $\mathrm{AUC_S}$ for race (R) and gender (G) can be regarded; to the right, the $\mathrm{AUC_S}$ for each of the different intersectional sensitive attribute values are displayed: non-white female (NWF), non-white male (NWM), white female (WF), and white male (WM). Focusing on the *Adult (Multiple)* configuration, it is clearly witnessable that the behaviour of the fairness measures along Θ match those of the *Adult (Race)* and *Adult (Gender)* previously shown in Fig. 2: as with the single binary sensitive attribute case, larger values of orthogonality translate to greater fairness and, conversely, lesser classification performance, which is expected (i.e., the performance-fairness trade-off). Noteworthily, 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 $\mathrm{AUC_S} = 0.55$ (a remarkably low bias value), and $\mathrm{AUC_Y}$ is above 0.8 indicating model prediction adequacy. Similarly for *Adult (Intersectional)* at the same orthogonality $\Theta = 0.7$, our method was able to converge the bias of all sensitive attribute values to prudent values concurrently and, at the same time, provide proper classification performance.

One limitation of our OvR approach to non-binary sensitive attributes is, however, regardable. Since the OvR $\mathrm{AUC_S}$ along multiple attributes or values is evaluated as its maximum (as described in Sec. 4.2), there is no guarantee that all but the most biased attribute will have its fairness increased: regard the slight increase in bias for non-white males. Yet, this characteristic of our approach

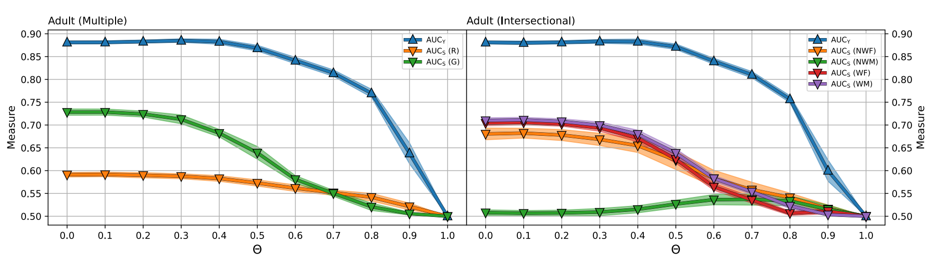

Figure 3: SCAFF model performance and fairness for multiple (left) and intersectional (right) sensitive attributes of the *Adult* dataset. Average and standard deviation values of $\mathrm{AUC_Y}$ and $\mathrm{AUC_S}$ values (vertical axis) across different values of orthogonality Θ (horizontal axis) are shown. Left: race (R) and gender (G). Right: non-white female (NWF), non-white male (NWM), white female (WF), and white male (WM).

also bounds the highest possible value of bias: along Θ, the maximum value of $\mathrm{AUC_S}$ is strictly monotonically decreasing. This is further corroborated by the behaviour of the similarly-valued biases NWF, WF, and MF, which have near identical curves along the different values of orthogonality Θ. These results show that our proposed SCAFF method is able to produce adequate fair classification models (low $\mathrm{AUC_S}$ and high $\mathrm{AUC_Y}$) with regards to a multitude of sensitive attributes and values.

6.3 Traditional Performance and Fairness Across Decision Thresholds

Below, we describe the results of applying our method to the *Recidivism (Gender)* dataset for different values of Θ and different decision thresholds. To quantify the relationship between measures, we present Table 1. Each cell depicts the Pearson correlation coefficient between threshold-independent and threshold-dependent measures of performance and bias along the parameter Θ, across different decision thresholds in the *Recidivism* dataset with respect to gender as the sensitive attribute. In other words, the coefficients represent the correlation between the values shown in the left-hand side of the two-dimensional Fig. 2 for either $\mathrm{AUC_Y}$ and accuracy, or $\mathrm{AUC_S}$ and the remaining fairness measures. In addition, bolded entries indicate a statistical significance of $\alpha = 0.95$ towards the null hypothesis of no correlation. As shown, the behaviour of the threshold-independent measure heavily correlates with that of the threshold-dependent ones, which is intended.

Table 1: Pearson correlation coefficients between threshold-independent and threshold-dependent measures of performance and bias along the parameter Θ, across different decision thresholds in the *Recidivism (Gender)* dataset. Bolded entries indicate a null hypothesis p-value ≤ 0.05

	Threshold								
Measure	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Accuracy	**0.990**	**0.991**	**0.994**	**0.993**	**0.997**	**0.999**	**0.998**	**0.998**	**0.997**
Dem. Parity	**0.962**	**0.843**	**0.914**	**0.984**	**0.975**	**0.985**	**0.988**	**0.991**	**0.966**
Equalised Odds	0.176	−.067	**0.608**	0.602	**0.755**	**0.917**	**0.821**	**0.978**	**0.974**
Eq. Opp. (TPR)	**0.772**	0.572	**0.773**	**0.759**	**0.841**	**0.953**	**0.953**	**0.972**	**0.975**
Eq. Opp. (FPR)	**0.900**	**0.761**	**0.708**	**0.682**	**0.813**	**0.937**	**0.961**	**0.754**	**0.776**

As a final remark, even though not experimentally shown, our method is also trivially applicable to (gradient) boosted techniques. The Z and Z' scores are computed according to the framework in question, but the selection of the split is made according to SG defined by equation (5) in Sec. 4.3.

7 Conclusion

In the present work we introduced SCAFF: the Splitting Criterion AUC for Fairness towards tree learning, which extends to bagged and boosted frameworks.

By doing so, we introduced (1) a decision threshold-independent fairness measure which we term *uniform demographic parity* which is computable as the *sensitive AUC*; and (2) a fair tree learning algorithm making use of this novel fairness measure which: (a) handles multicategorically valued sensitive attributes as well as multiple sensitive attributes simultaneously; and (b) is tunable with respect to the performance-fairness trade-off via an orthogonality parameter Θ.

Finally, we empirically validated our method through extensive experimentation, improving upon the current state-of-the-art in fairness literature. Within our experiments with real datasets, our method outperformed the state-of-the-art not only in terms of predictive performance and model fairness, but also by its capability of handling multiple sensitive attributes, of which the values may be valued multicategorically. Moreover, we showed how our novel threshold-independent fairness measure correlates with the traditionally-used threshold-dependent fairness measures. As future work, it would be advantageous to exploit different threshold-independent performance measures used in machine learning towards the fairness classification scenario. Ultimately, the deployment of fair machine learning approaches within domains where automated decision-making is required should be the ulterior goal in this field of research.

References

[1] Ahmad Taher Azar and Shereen M El-Metwally. "Decision tree classifiers for automated medical diagnosis". In: *Neural Computing and Applications* 23.7 (2013), pp. 2387–2403.

[2] António Pereira Barata et al. "The eXPose approach to crosslier detection". In: *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE. 2021, pp. 2312–2319.

[3] Julia Dressel and Hany Farid. "The accuracy, fairness, and limits of predicting recidivism". In: *Science Advances* 4.1 (2018), eaao5580.

[4] Sam Corbett-Davies et al. "A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear". In: *The Washington Post* (Oct. 17, 2016). URL: `https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas/` (visited on 04/21/2021).

[5] Jeffrey Dastian. "Amazon scraps secret AI recruiting tool that showed bias against women". In: *Reuters* (Oct. 11, 2018). URL: `https://www.reuters.com/article/us-amazon-com-jobs-automation-in...-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G` (visited on 04/21/2021).

[6] Pieter Klein. "Belastingdienst geeft toe: toch sprake van etnisch profileren". In: *RTL Nieuws* (May 11, 2020). URL: `https://www.rtlnieuws.nl/nieuws/artikel/5117616/belastingdienst-toeslagen-profileren-nationaliteit` (visited on 04/21/2021).

[7] European Comission. “Proposal for a regulation on a European approach for artificial intelligence”. In: (Apr. 21, 2021). URL: https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-european-approach-artificial-intelligence (visited on 04/21/2021).

[8] Yaowei Hu et al. “Fair Multiple Decision Making Through Soft Interventions”. In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., 2020, pp. 17965–17975. URL: https://proceedings.neurips.cc/paper/2020/file/d0921d442ee91b896ad95059d13df618-Paper.pdf.

[9] Jaewoong Cho, Gyeongjo Hwang, and Changho Suh. “A Fair Classifier Using Kernel Density Estimation”. In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., 2020, pp. 15088–15099. URL: https://proceedings.neurips.cc/paper/2020/file/ac3870fcad1cfc367825cda0101eee62-Paper.pdf.

[10] Toon Calders and Sicco Verwer. “Three naive bayes approaches for discrimination-free classification”. In: *Data Mining and Knowledge Discovery* 21.2 (2010), pp. 277–292.

[11] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. “Inherent trade-offs in the fair determination of risk scores”. In: *arXiv preprint arXiv:1609.05807* (2016).

[12] Tongxin Hu et al. “Fairnn-conjoint learning of fair representations for fair decisions”. In: *International Conference on Discovery Science*. Springer. 2020, pp. 581–595.

[13] Muhammad Bilal Zafar et al. “Fairness constraints: Mechanisms for fair classification”. In: *Artificial Intelligence and Statistics*. PMLR. 2017, pp. 962–970.

[14] Morteza Zabihi et al. “Detection of atrial fibrillation in ECG hand-held devices using a random forest classifier”. In: *2017 Computing in Cardiology (CinC)*. IEEE. 2017, pp. 1–4.

[15] Nejdet Dogru and Abdulhamit Subasi. “Traffic accident detection using random forest classifier”. In: *2018 15th Learning and Technology Conference (L&T)*. IEEE. 2018, pp. 40–45.

[16] Mitch N Angenent, António Pereira Barata, and Frank W Takes. “Large-scale machine learning for business sector prediction”. In: *Proceedings of the 35th Annual ACM Symposium on Applied Computing*. 2020, pp. 1143–1146.

[17] Cynthia Dwork et al. “Fairness through awareness”. In: *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*. 2012, pp. 214–226.

[18] Sam Corbett-Davies and Sharad Goel. “The measure and mismeasure of fairness: A critical review of fair machine learning”. In: *arXiv preprint arXiv:1808.00023* (2018).

[19] Moritz Hardt, Eric Price, and Nathan Srebro. “Equality of opportunity in supervised learning”. In: *arXiv preprint arXiv:1610.02413* (2016).

[20] Wenbin Zhang and Eirini Ntoutsi. “FAHT: An adaptive dairness-aware decision tree classifier”. In: *Proceedings of the 28th International Joint Conference on Artificial Intelligence, IJCAI-19*. International Joint Conferences on Artificial Intelligence Organization, July 2019, pp. 1480–1486. DOI: 10.24963/ijcai.2019/205. URL: https://doi.org/10.24963/ijcai.2019/205.

[21] Faisal Kamiran, Toon Calders, and Mykola Pechenizkiy. “Discrimination aware decision tree learning”. In: *2010 IEEE International Conference on Data Mining*. IEEE. 2010, pp. 869–874.

[22] Anonymous. *Fair tree learning*. https://anonymous.4open.science/r/fair_tree_learning-29BA. 2021.

[23] Tom Fawcett. “An introduction to ROC analysis”. In: *Pattern Recognition Letters* 27.8 (2006), pp. 861–874.

[24] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. “Boosting and additive trees”. In: *The elements of statistical learning*. Springer, 2009, pp. 337–387.

[25] European Union. “Gender & intersectional bias in artificial intelligence”. In: (Sept. 28, 2020). URL: https://op.europa.eu/en/publication-detail/-/publication/286e1432-021a-11eb-836a-01aa75ed71a1 (visited on 04/21/2021).

[26] David MJ Tax and Robert PW Duin. “Using two-class classifiers for multiclass classification”. In: *Object Recognition Supported by User Interaction for Service Robots*. Vol. 2. IEEE. 2002, pp. 124–127.

[27] Leo Breiman. “Random forests”. In: *Machine learning* 45.1 (2001), pp. 5–32.

[28] “Pearson’s Correlation Coefficient”. In: *Encyclopedia of Public Health*. Ed. by Wilhelm Kirch. Dordrecht: Springer Netherlands, 2008, pp. 1090–1091. ISBN: 978-1-4020-5614-7. DOI: 10.1007/978-1-4020-5614-7_2569. URL: https://doi.org/10.1007/978-1-4020-5614-7_2569.