Privacy Constrained Fairness Estimation for Decision Trees

Florian van der Steen[1*], Fré Vink[2] and Heysem Kaya[1]

[1*]Department of Information and Computing Sciences, Utrecht University, Princetonplein 5, Utrecht, 3584 CC, the Netherlands.
[2]Responsible AI Team, Dutch Central Government Audit Service, Korte Voorhout 7, Den Haag, 2511 CW, the Netherlands.

*Corresponding author(s). E-mail(s): florianvandersteen@gmail.com;
Contributing authors: f.t.vink@minfin.nl; h.kaya@uu.nl;

Abstract

The protection of sensitive data becomes more vital, as data increases in value and potency. Furthermore, the pressure increases from regulators and society on model developers to make their Artificial Intelligence (AI) models non-discriminatory. To boot, there is a need for interpretable, transparent AI models for high-stakes tasks. In general, measuring the fairness of any AI model requires the sensitive attributes of the individuals in the dataset, thus raising privacy concerns. In this work, the trade-offs between fairness, privacy and interpretability are further explored. We specifically examine the Statistical Parity (SP) of Decision Trees (DTs) with Differential Privacy (DP), that are each popular methods in their respective subfield. We propose a novel method, dubbed Privacy-Aware Fairness Estimation of Rules (PAFER), that can estimate SP in a DP-aware manner for DTs. DP, making use of a third-party legal entity that securely holds this sensitive data, guarantees privacy by adding noise to the sensitive data. We experimentally compare several DP mechanisms. We show that using the Laplacian mechanism, the method is able to estimate SP with low error while guaranteeing the privacy of the individuals in the dataset with high certainty. We further show experimentally and theoretically that the method performs better for DTs that humans generally find easier to interpret.

Keywords: responsible AI, fairness, interpretability, differential privacy

1 Introduction

The methods from the scientific field of AI, and in particular Machine Learning (ML), are increasingly applied to tasks in socially sensitive domains. Due to their predictive power, ML models are used within banks for credit risk assessment [1], aid decisions within universities for new student admissions [2] and aid bail decision-making within courts [3]. Algorithmic decisions in these settings can have fargoing impacts, potentially increasing disparities within society. Numerous notorious examples exist of algorithms causing harm in this regard. In 2015, Google Photos new image recognition model classified some black individuals as gorillas [4]. This led to the removal of the category within Google Photos. A report by Amnesty International concluded that the Dutch Tax & Customs administration used a model for fraud prediction that discriminated against people with multiple nationalities [5].

The application of ML should clearly be done responsibly, giving rise to a field that considers the fairness of algorithmic decisions. Fair ML is a field within AI concerned with assessing and developing fair ML models. Fairness in this sense closely relates to equality between groups and individuals. The main notion within the field is that models should not be biased, that is, have tendencies to over/underperform for certain (groups of) individuals. This notion of bias is different from the canonical definition of bias in statistics, i.e. the difference between an estimator's expected value and the true value. Essentially, similar individuals should be treated similarly, and decisions should not lead to unjust discrimination. Non-discrimination laws for AI exist within the EU [6] and more are upcoming [7]. The Dutch government now has a register of all the algorithms used within it [8].

An additional property that responsible ML models should have, is that they are interpretable. Models of which the decision can be explained, are preferred as they aid decision-making processes affecting real people. In a loan application setting, users

have the right to know how a decision came about [9]. The field of Explainable Artificial Intelligence (XAI), is concerned with building models that are interpretable and explainable.

Inherently, ML models use data. Thus, there is also a tension between the use of these models and privacy, especially for socially sensitive tasks. Individuals have several rights when it comes to data storage, such as the right to be removed from a database [6]. It is also beneficial for entities to guarantee privacy so that more individuals trust the entity with their data. Some data storage practices are discouraged such as the collection of several protected attributes [6]. These attributes, and thus the storage practices thereof, are sensitive. Examples include the religion, marital status, and gender of individuals. In industrial settings, numerous data leaks have occurred. Social media platforms are especially notorious for privacy violations, with Facebook even incurring data breaches on multiple occasions [10, 11]. The report by Amnesty International also concluded that the Dutch Tax & Customs Administration in the Dutch childcare benefits scandal failed to safely handle the sensitive private data of thousands of individuals, while they used a biased model [5]. This work will investigate these three pillars of Responsible AI, investigating a novel method that is at the intersection of these three themes.

To assess and improve fairness precisely, one needs the sensitive attributes of the individuals that a ML model was trained on. But these are often absent or have limited availability, due to privacy considerations. Exactly here lies the focal point of this work, the assessment of the fairness of ML models, while respecting the privacy of the individuals in the dataset. These antagonistic goals make for a novel, highly constrained, and hence difficult problem. A focus is placed on DTs, a class of interpretable models from XAI since these types of models are likely to be used in critical tasks involving humans due to the GDPR (in particular Art. 22) [6] and its national

3

implementations. There are thus four goals we try to optimize in this work: fairness, privacy, interpretability, and predictive performance.

1.1 Research Questions

The main goal of this work is to develop a method that can estimate the fairness of an interpretable model with a high accuracy while respecting privacy. A method, named Privacy-Aware Fairness Estimation of Rules (PAFER), is proposed that can estimate the fairness of a class of interpretable models, DTs, while respecting privacy. The method is thus at the intersection of these three responsible AI pillars. The research questions (RQs), along with their research subquestions, (RSQs) are:

RQ1 What is the optimal privacy mechanism that preserves privacy and minimizes average Statistical Parity error?

RSQ1.1 Is there a statistically significant mean difference in Absolute Statistical Parity error between the Laplacian mechanism and the Exponential mechanism?

RQ2 Is there a statistically significant difference between the Statistical Parity errors of PAFER compared to other benchmarks for varying Decision Tree hyperparameter values?

RSQ2.1 At what fractional `minleaf` value is PAFER significantly better at estimating Statistical Parity than a random baseline?

1.2 Outline

The remainder of the paper is organized as follows. The upcoming section 2 will provide the theoretical background, which is followed by section 3 that covers the related literature. Section 4 describes the novel method that is proposed in this work. Subsequently, section 5 describes the performed experiments, their results, and thorough analysis. Finally, section 6 concludes with limitations and future directions.

2 Preliminaries

This section discusses work related to the research objectives and provides background to the performed research. Subsection 2.1 describes fairness theory, subsection 2.2 provides background on interpretable models and subsection 2.3 explains notions of privacy.

2.1 Fairness Definitions

Fairness in an algorithmic setting relates to the way an algorithm handles different (groups of) individuals. Unjust discrimination[1] is often the subject when examining the behavior of algorithms with respect to groups of individuals. For this work, only fairness definitions relating to supervised ML were studied, as this is the largest research area within algorithmic fairness.

In 2016, the number of papers related to fairness surged. Partly, due to the new regulations such as the European GDPR [6] and partly due to a popular article by ProPublica which examined racial disparities in recidivism prediction software [13]. Because of the young age of the field and the sudden rise in activity, numerous definitions of fairness have been proposed since. Most of the definitions also simultaneously hold multiple names; this section aims to include as many of the names for each definition.

The performance-oriented nature of the ML research field accelerated the development of fairness metrics, quantifying the fairness for a particular model. The majority of the definitions can therefore also be seen, or rewritten, as a measuring stick for the fairness of a supervised ML model. This measurement may be on a scale, which is the case for most group fairness definitions, or binary, which is the case for some causal fairness definitions.

---

[1]What exactly is **unjust** discrimination is a social construct and changes over time [12].

5

The fairness definitions, namely the mathematical measures of fairness, can be categorized into group fairness, individual fairness and causal fairness [14]. Considering the space limitations and the relevance to our work, in this section, we will focus on group fairness and provide the definitions of the most prominent measures used in the literature. Group fairness is the most popular type of fairness definition as it relates most closely to unjust discrimination. Individuals are grouped based on a sensitive, or protected attribute, A, which partitions the population. This partition is often binary, for instance when A denotes a privileged and unprivileged group. In this subsection, we assume a binary partition for ease of notation, but all mentioned definitions can be applied to $\mathcal{K}$-order partitions. Some attributes are protected by law, for example, gender, ethnicity and age.

The setting for these definitions is often the binary classification setting where $Y \in \{0, 1\}$, with Y as the outcome. This is partly due to ease of notation, but more importantly, the binary classification setting is common in impactful prediction tasks. Examples of impactful prediction tasks are granting or not granting a loan [1], accepting or not accepting students to a university [2] and predicting recidivism after a certain period [3]. In each setting, a clear favorable (1) and unfavorable (0) outcome can be identified. Thus, we assume the binary classification setting in the following definitions.

2.1.1 Statistical Parity

Statistical Parity (SP) is a decision-based definition, which compares the different positive prediction rates for each group [15]. SP, also known as demographic parity, equal acceptance rate, total variation or the independence criterion, is by far the most popular fairness definition. The mathematical definition is:

$$\text{SP} = p(\hat{Y} = 1 | A = 1) - p(\hat{Y} = 1 | A = 0), \tag{1}$$

where $\hat{Y}$ is the decision of the classifier. An example of SP would be the comparison of the acceptance rates of males and females to a university.

Note that Equation 1 is the SP-difference but the SP-ratio also exists. US law adopts this definition of SP as the 80%-rule [16]. The 80%-rule states that the ratio of the acceptance rates must not be smaller than 0.8, i.e. 80%. Formally:

$$80\%\text{-rule } = 0.8 \leq \frac{p(\hat{Y} = 1|A = 1)}{p(\hat{Y} = 1|A = 0)} \leq 1.25, \tag{2}$$

where the fraction is the SP-ratio. SP is easy to compute and merely uses the model's predictions. SP therefore does not require labelled data. These advantages make it one of the most used fairness definitions.

2.1.2 Equalized Odds

Another, also very common, fairness definition is the Equalized Odds (EOdd) metric [17]. It is also known as disparate mistreatment or the separation criterion. EOdd requires that the probabilities of being correctly positively classified and the probabilities of being incorrectly positively classified are equal across groups. Thus, the definition is twofold; both false positive classification probability and true positive classification probability should be equal across groups. Formally:

$$\text{EOdd} = p(\hat{Y} = 1|Y = y, A = 1) - p(\hat{Y} = 1|Y = y, A = 0), \ \ y \in \{0, 1\}. \tag{3}$$

An advantage of EOdd is that, unlike SP, when the predictor is perfect, i.e. $Y = \hat{Y}$, it satisfies EOdd.

2.1.3 Equality of Opportunity

A relaxation of EOdd is the fairness definition Equality of Opportunity (EOpp) [17]. It just requires the equality of the probabilities of correctly predicting the positive class

across groups. In other words, where EOdd requires that both true positive and false positive classification rates are equal across groups, EOpp only requires the former. Formally:

$$\text{EOpp} = p(\hat{Y} = 1 | Y = 1, A = 1) - p(\hat{Y} = 1 | Y = 1, A = 0). \tag{4}$$

An advantage of EOpp is that it is not a bi-objective, and thus is more easily optimized for compared to EOdd.

2.2 Interpretable Models

This subsection outlines a class of models with inherently high interpretability, DTs, that are central to this work. The interpretability of a model is the degree to which the classifications and the decision-making mechanism can be interpreted. The field of XAI is concerned with building systems that can be interpreted and explained. Complex systems might need an explanation function that generates explanations for the outputs of the system. Some methods may inherently be highly interpretable, requiring no explanation method, such as DTs. Interpretability may be desired to ensure safety, gain insight, enable auditing or manage expectations.

2.2.1 Decision Trees (DTs)

A DT is a type of rule-based system that can be used for classification problems. The structure of the tree is learned from a labelled dataset. DTs consist of nodes, namely branching nodes and leaf nodes. The upper branching node is the root node. To classify an instance, one starts at the root node and follows the rules which apply to the instance from branching node to branching node until no more rules can be applied. Then, one reaches a decision node, also called a leaf node. Every node holds the

instances that could reach that node. Thus, the root node holds every instance. Decision nodes classify instances based on the class that represents the most individuals within that node.

There are two effective ways to determine the structure of a DT, given a labelled dataset. The most common way is to have a function that indicates what should be the splitting criterion in each branching node. These heuristic functions look at splitting criteria to partition the data in the node such that each partition is as homogeneous as possible w.r.t. class. An example of such a heuristic is entropy, intuitively defined as the degree to which the class distribution is random in a partition. A greedy process then constructs the tree, picking the best split in each individual node. Optimal DTs are a newer set of approaches, that utilize methods from dynamic programming and constrained optimization [18]. Their performance is generally better as they approach the true DT more closely than greedily constructed DTs. However, their construction is computationally heavy.

The interpretability of a DT is determined by several factors. The main factor is its height, the number of times the DT partitions the data. Very shallow Decision Trees are sometimes called Decision Stumps [19]. The `minleaf` DT hyperparameter also influences the interpretability of a DT. The `minleaf` value constrains how many instances should minimally hold in a leaf node. The smaller the value, the more splits are required to reach the set `minleaf` value. Optimal DTs cannot have a tall height due to their high computational cost. Greedy DTs can be terminated early in the construction process to maintain interpretability. Closely related to height is the number of decision nodes in the tree. This also influences the interpretability of DTs, as the more decision nodes a DT has, the more complex the DT is. Finally, DTs built with numeric features might become uninterpretable because they use the same numeric feature over and over, leading to unintuitive decision boundaries.

In general, DTs are interpretable because they offer visualizations and use rules, which are both easy to understand for humans [20]. Major disadvantages of DTs include their incapability of efficiently modeling linear relationships and their sensitivity to changes in the data. Still, their performance, especially ensembles of DTs, are state-of-the-art for prediction tasks on tabular data [21].

2.3 Privacy Definitions

The final main pillar of responsible AI that this work discusses is privacy. Privacy, in general, is a term that can be used in multiple contexts. In its literal sense, privacy relates to one's ability to make personal and intimate decisions with nothing interfering. In this work, however, privacy refers to the degree of control one has over others accessing personal data about themselves. This is also known as informational privacy. The less personal data others access about an individual, the more privacy the individual has. This subsection discusses several techniques to increase informational privacy.

2.3.1 Differential Privacy (DP)

Differential Privacy (DP) [22] is a notion that gives mathematical guarantees on the membership of individuals in a dataset. In principle, it is a promise to any individual in a dataset, namely: 'You will not be affected, adversely or otherwise, by allowing your data to be used in any analysis of the data, no matter what other analyses, datasets, or information sources are available' [23]. More specifically, an adversary cannot infer if an individual is in the dataset. DP can be applied when sharing data, or an analysis of the data. ML models are ways of analysing data and therefore can also promise to adhere to DP. Another guarantee that DP makes is that it is immune to post-processing, i.e. DP cannot be undone [23].

Definition

The promise of DP can be mathematically guaranteed up to a parameter ε. A higher ε guarantees more privacy. This parameter ε is the privacy budget. The main means of guaranteeing the promise of DP is by perturbing the data, i.e. adding noise to the data. In the context of building ML models, this noise may be added to the parameters of the ML model or to its training data. At any rate, there is a query, $q(\cdot)$, for data[2], to which DP adds noise. Because DP is based on membership inference, the formal definition compares two neighboring datasets, D and D', in which only one instance differs. For these datasets, (ε, δ)-DP formally is:

$$p(\mathcal{A}(q(D)) \subseteq range(\mathcal{A})) \leq \exp(\varepsilon) \cdot p(\mathcal{A}(q(D')) + \delta \subseteq range(\mathcal{A})), \qquad (5)$$

where $\mathcal{A}$ is a randomized mechanism around a query $q(\cdot)$ and $range(\mathcal{A})$ is the range of all outcomes the mechanism can have. If $\delta = 0$, ε-DP is satisfied. DP-mechanisms thus randomize query answers in some way.

Global Sensitivity

How much noise ought to be added, depends on the difference the inclusion of one worst-case individual in the dataset makes for the query answer. This is known as the sensitivity, Δq, how sensitive a query answer is to a change in the data [22]. Formally:

$$\Delta q = \max_{D,D'} ||q(D) - q(D')||_1, \qquad (6)$$

which is also know as the ℓ_1-sensitivity or the global sensitivity.

---

[2]This query may come from a user of a ML model or from a developer that requires training data.

Laplace Mechanism

Several techniques exist to randomize query answers, of which the most common one is the Laplacian mechanism [22], for queries requesting real numbers[3]. The mechanism involves adding noise to a query answer, sampled from the Laplace distribution, centered at 0 and with a scale equal to $\frac{\Delta q}{\varepsilon}$. The Laplace mechanism can be formalised as:

$$\mathcal{A}(D, q(\cdot), \varepsilon) = q(D) + Lap(\frac{\Delta q}{\varepsilon}),\tag{7}$$

where $Lap(\frac{\Delta q}{\varepsilon})$ is the added Laplacian noise.

Exponential Mechanism

A different noise schema is the Exponential mechanism [24], used for categorical, utility-related queries[4]. For these sorts of queries, a small amount of noise may completely destroy the utility of the query answer. A utility function, $u_D(r)$, is defined over the categories, $r \in \mathcal{R}$, for a certain dataset D. The exponential mechanism is sensitive w.r.t. the utility function, Δu, not with respect to changes in r. The exponential mechanism can be formally defined as:

$$p(\mathcal{A}(D, u, \mathcal{R}, \varepsilon) = r) \propto \exp(\frac{\varepsilon u_D(r)}{2\Delta u}).\tag{8}$$

In other words, the probability of the best category being chosen is proportional to $e^{\frac{\varepsilon u_D(r)}{2\Delta u}}$.

Gaussian Mechanism

The Gaussian mechanism adds noise based on the Gaussian distribution, with $\mathcal{N}(0, \sigma)$. The mechanism is similar to the Laplacian mechanism in this sense. DP holds if $\sigma \geq \sqrt{2\ln(\frac{1.25}{\delta})}\frac{\Delta_2}{\varepsilon}$ [23]. The term Δ_2 is the global ℓ_2-sensitivity; instead of using the

---

[3]An example of such a query might be: 'What is the average age of females in the dataset?'.
[4]An example of such a query might be: 'What is the optimal attribute to partition the dataset in terms of class?' Such a query can be found in the next subsection.

ℓ_1-norm in Equation 6, Δ_2 uses the ℓ_2-norm. The Gaussian mechanism can be deemed a more 'natural' type of noise, as it adds noise that is often assumed to be present in measurements. A disadvantage is that both δ and ε must be in $(0, 1)$, so ε-DP can never be met.

3 Related Work

This section discusses work related to the research objectives. Whereas the previous section discussed background related to only one pillar of Responsible AI, this section will highlight methods at the intersection of these fields. It concludes by relating the proposed method, PAFER, to the current landscape of methods.

3.1 Fair Decision Trees

Some of the earliest work regarding fair DTs was performed by Kamiran & Calders and is now known as Discrimination Aware Decision Trees (DADT). They proposed a Heuristic-Based DT that incorporates the homogeneity of the sensitive attribute into the splitting criterion [25]. DADT also performs some post-processing s.t. certain decision nodes change their decision. This step is phrased as a KNAPSACK problem [26], and is also solved greedily.

In terms of optimal DTs, Linden et al. achieve excellent results with a method named DPFair [27]. Their work significantly improves the speed of the work of Jo et al., who formulate the optimal DT problem with an additional fairness objective [28].

3.2 Privacy-aware Decision Trees

DTs with privacy guarantees are best represented by the work of Mohammed et al.. The method, named Private Decision tree Algorithm (PDA), uses the Exponential mechanism and queries the required quantities for greedily building op the DT [29]. For an in-depth overview of DTs with privacy guarantees, the reader is referred to [30].

3.3 Fair Privacy-aware models

There is an upcoming field within responsible AI that is aimed at improving fairness, without accessing sensitive data. Prominent examples include Adversarially Reweighted Learning (ARL) and Fair Related Features (FairRF) [31, 32], respectively. While we highly value this line of work, it does not allow for the evaluation or estimation of fairness, as the field assumes sensitive attributes are entirely unavailable. Therefore, we consider these methods to be insufficient for our purpose, as we aim to provide guarantees on the degree of fairness a model exhibits, e.g. adherence to the 80%-rule.

The method most closely related to ours is named AttributeConceal and was introduced by Hamman et al.. They explore the idea of querying the group fairness metrics [33]. The scenario they assume is that ML developers have some dataset without sensitive attributes for which they build models, and therefore query SP and EOdd from a data curator. They establish that if the developers have bad intentions, they can identify a sensitive attribute of an individual using one unrealistic query, or two realistic ones. The main idea is that the models, for which they query fairness metrics, differ only on one individual, giving away their sensitive attribute via the answer. This result is then extended using any number of individuals. When the sizes of the groups differ greatly, i.e. $|D_{A=0}| \ll |D_{A=1}|$, using compressed sensing [34], the number of queries is in $O(|D_{A=0}| \log(\frac{N}{|D_{A=1}|}))$, with $N = |D_{A=1} + D_{A=0}|$, the total number of instances. The authors propose a mitigation strategy named AttributeConceal, using smooth sensitivity. This is a sensitivity notion that is based on the worst-case individual in the dataset. DP is ensured for any number of queries by adding noise to each query answer. It is experimentally verified that using AttributeConceal, an adversary can predict sensitive attributes merely as well as a random estimator.

Table 1 Overview of methods that are similar to PAFER. Fairness-Aware methods are methods that aim to improve or estimate fairness.

Method	Interpretable	Privacy-aware	Fairness-Aware
DADT [25]	✓	✗	✓
DPFair [27]	✓	✗	✓
PDA [29]	✓	✓	✗
ARL [31]	✗	✓	✓
FairRF [32]	✗	✓	✓
AttributeConceal [33]	✗	✓	✓
PAFER	✓	✓	✓

3.4 PAFER & Related Work

Table 1 shows methods from the domain of responsible AI that have similar goals to PAFER. In general, we see a lack of fair, privacy-preserving methods for rule-based methods, specifically DTs. Hamman et al. investigate the fairness of models in general without giving in on privacy [33], but the method lacks validity. The developers, in their setting, do not gain intuition on what should be changed about their model to improve fairness. One class of models that lends itself well to this would be DTs, as these are modular and can be pruned, i.e. rules can be removed. DTs are the state-of-the-art for tabular data [21] and sensitive tasks are often prediction tasks for tabular data[5]. A method that can identify unfairness in a privacy-aware manner for DTs would be interpretable, fair and differentially private, respecting some of the most important pillars of responsible AI. PAFER aims to fill this gap, querying the individual rules in a DT. The next section will introduce the method.

4 Proposed Method

In this section, we introduce PAFER, a novel method to estimate the fairness of DTs in privacy constrained manner. The following subsections dissect the proposed method, starting with subsection 4.1, on the assumptions and specific scenarios for which the

---

[5]Examples are university acceptance [2], bail decision making [3] and credit risk assessment [1].

method is built. Successively, subsection 4.2 provides a detailed description of the procedure, outlining the pseudocode and some theoretical properties.

4.1 Scenario

PAFER requires a specific, albeit common, scenario for its use. This subsection describes that scenario and discusses how common the scenario actually is.

4.1.1 Assumptions

PAFER is a method that requires a certain setting, which comes with several assumptions. Firstly, PAFER is made for an auditing setting, in the sense that it is a method that is assumed to be used at the end of a development cycle. PAFER does not mitigate bias, it merely estimates the fairness of the rules in a DT. Secondly, we assume that a developer has constructed a DT that makes binary decisions on a critical task (e.g., about people). The developer may have had access to a dataset containing individuals and some task-specific features, but this dataset does not contain a full specification of sensitive attributes on an instance level. The developer (or the algorithm auditor) wants to assess the fairness of their model using SP. We lastly assume that a legal, trusted third party exists that knows these sensitive attributes on an instance or aggregate level,[6] and is willing to share them using some safe private protocol. Based on these assumptions, the fairness of the DT can be assessed, using the third party and PAFER.

4.1.2 Prevalance of Scenario

The scenario that was described in the previous subsection can occur in the real world under varying circumstances. This subsection enumerates some assumptions and their prevalence in the real world. Firstly, it is common to see a rule-based method built

---

[6]These sensitive data can be kept at the aggregate level at the legal party to minimize sensitive data leakage and to conform to privacy laws.

for a sensitive task [35, 36]. Rules are able to explain the decision process, allowing individuals that are affected by the system to receive explanations about the decision affecting them. Secondly, binary decision-making is also quite common for sensitive tasks. Prominent examples include university acceptance decision making [2], recidivism prediction [13] and loan application evaluations [1]. Moreover, multiclass decision-making problems can be rewritten as binary decision problems, as shown in Corollary 1. Thirdly, it is often the case that model developers do not have access to sensitive attributes. Simply because of regulations [6], or because they were not deemed necessary when gathering the data. Lastly, it is quite common that a developer worries about fairness after the construction of their model. This may be due to newly imposed regulations [7], due to a compliance check by an auditing body or due to newly created awareness of machine bias [13]. Furthermore, when sensitive data is absent, the development of a fair rule-based system becomes difficult. There are currently no fair, interpretable, sensitive attribute agnostic classifiers, as is apparent from section 3.

What is uncommon, however, is a third party that has the sensitive attribute data of the individuals in the dataset, albeit at an aggregate level, and is also willing to share them. As data is the new oil fueling modern machines [37], sharing data becomes more and more difficult. Since, however, fair and interpretable sensitive attribute agnostic classifiers are currently lacking, this assumption becomes necessary. This work can thus be seen as an exploration of this cooperation between the developer and the data holder, to determine the privacy risks and utility of such an exchange.

4.2 Privacy-Aware Fairness Estimation of Rules: PAFER

We propose Privacy-Aware Fairness Estimation of Rules (PAFER), a method based on DP [23], that enables the calculation of SP for DTs while guaranteeing privacy. PAFER sends specifically designed queries to a third party to estimate SP. PAFER sends one

17

query for each decision-making rule and one query for the overall composition of the sensitive attributes. The size of each (un)privileged group, along with the total number of accepted individuals from each (un)privileged group, allows us to calculate the SP. Let $\mathcal{X}$ be the data used to train a DT, with x_i^j the jth feature of the ith individual. Let a rule be of the form $x^1 < 5 \wedge x^2 = True$. The query then asks for the distribution of the sensitive attributes for all individuals that have properties $x^1 < 5$ and $x^2 = True$. In PAFER, each query is a histogram query as a person cannot be both privileged and unprivileged. The query to determine the general sensitive attribute composition of all individuals can be seen as a query for an 'empty' rule; a rule that applies to everyone[7]. It can also be seen as querying the root node of a DT.

4.2.1 PAFER and the privacy budget

A property of DTs is that only one rule applies to a person. Therefore, PAFER queries each decision-making rule without having to share the privacy budget between these queries. Although we calculate a global statistic in SP, we query each decision-making rule. This is possible due to some noise cancelling out on aggregate, and, for DTs, because we can share the privacy budget over all decision-making rules. This intuition was also noted in [30].

Because PAFER queries every individual at least once, half of the privacy budget is spent on the query to determine the general sensitive attribute composition of all individuals, and the other half is spent on the remaining queries. Still, reducing the number of queries reduces the total amount of noise. PAFER therefore prunes non-distinguishing rules. A redundant rule can be formed when the splitting criterion of the DT improves but the split does not create a node with a different majority class.

---

[7]In logic this rule would be a tautology, a statement that is always true, e.g. $x^1 < 5 \vee x^1 \geq 5$.

4.2.2 PAFER and Statistical Parity

The definition of SP that PAFER calculates differs slightly from the most common, original definition [15], to support intersectional fairness analyses and to ensure the SP value is in $[0, 1]$. When A is a $\mathcal{K}$-ary sensitive attribute, the metric that PAFER calculates is:

$$\text{SP} = \min \left(\frac{p(\hat{Y} = 1 | A = a)}{p(\hat{Y} = 1 | A = b)} \right), \ a, b \in \{0, 1, 2, \ldots, k - 1\}, a \neq b. \tag{9}$$

The SP value is always in $[0, 1]$, as we arrange the fraction such that the smallest 'acceptance rate' is in the numerator and the largest is in the denominator.

4.2.3 DP mechanisms for PAFER

Three commonly used DP mechanisms are apt for PAFER, namely the Laplacian mechanism, the Exponential mechanism and the Gaussian mechanism. The Laplacian mechanism is used to perform a histogram query and thus has a sensitivity of 1 [23]. The Exponential mechanism uses a utility function such that $u_D(r) = q(D) - |q(D) - r|$ where r ranges from zero to the number of individuals that the rule applies to, and $q(D)$ is the true query answer. The sensitivity is 1 as it is based on its database argument, and this count can differ by only 1 [23]. The Gaussian mechanism is also used to perform a histogram query and has a sensitivity of 2, as it uses the Δ_2-sensitivity.

4.2.4 Invalid Answer Policies

The Laplacian mechanism and Gaussian mechanism add noise in such a way that invalid query answers may occur. A query answer is invalid if it is negative, or if it exceeds the total number of instances in the dataset[8]. A policy for handling these

---

[8]Note that is common for a histogram query answer to exceed the number of individuals in a decision node by a certain amount. We, therefore, do not deem it as an invalid query answer.

invalid query answers must be chosen. In practice, these are mappings from invalid values to valid values. We provide several options in this subsection.

Table 2 The proposed policy options for each type of invalid query answer.

Negative	Too Large
0	uniform
1	total - valid
uniform	
total - valid	

Table 2 shows the available options for handling invalid query answers. A policy consists of a mapping chosen from the first column and a mapping chosen from the second in this table. The first column shows policies for negative query answers and the second column shows policies for query answers that exceed the number of individuals in the dataset. The 'uniform' policy replaces an invalid answer with the answer if the rule would apply to the same number of individuals from each un(privileged) group. The 'total - valid' policy requires that all other values in the histogram were correct and thus together allow for a calculation of the missing value by subtracting it from the total.

4.2.5 PAFER Pseudocode

Algorithm 1 shows the pseudocode for PAFER.

4.2.6 Theoretical Properties of PAFER

We theoretically determine a lower and upper bound of the number of queries that PAFER requires for a k-ary DT in Theorem 1. The lower bound is equal to two, and the upper bound is $2^{h-1} + 1$, dependent on the height of the DT, h. Note that PAFER removes redundant rules to reduce the number of rules. The larger the number of rules, the more noise is added on aggregate.

Algorithm 1 PAFER

```
 1: procedure PAFER(𝒜, D, ε, DT, π, 𝒦)
 2:      ▷ 𝒜 is a DP mechanism that introduces noise
 3:      ▷ D is a database with N instances
 4:      ▷ ε is the privacy budget
 5:      ▷ DT is a binary Decision Tree composed of rules
 6:      ▷ π is a policy that transforms invalid query answers to valid query answers
 7:      ▷ 𝒦 is the number of sensitive groups for the sensitive attribute
 8:      accept_rates ← zeros(1, 𝒦) ▷ accept_rates is a row vector of dimension 𝒦,
    initialized at 0
 9:      total ← 𝒜(True, D, ½ε)
10:     for q ∈ DT do
11:         if q is favorable then
12:             accept_rates += π(𝒜(q,D,½ε))/total
13:         end if
14:     end for
15:     ŜP = min(accept_rates)/max(accept_rates)
16:     return ŜP
17: end procedure
```

Corollary 1. *Any DT that uses non-binary splits and that classifies for a binary decision problem, can be converted to a DT that solely uses binary splits.*

Proof. Assume a DT has nodes with an arbitrary number of splits k, with clauses $A, B, C, \ldots, K$. Converting this to a binary decision process can be achieved by chaining each clause, i.e. for each clause a split is created of the form A or $\neg A$. The latter of the two branches is then chained to B or $\neg B$, and so forth. This process is schematically shown in Figure A1 in Appendix A. Since we have proven this property for an arbitrary number of splits in a node, the property holds for any k-ary DT. ☐

Theorem 1. *The number of queries required by PAFER to estimate SP for a binary DT is lower bounded by 2 and upper bounded by $2^{h-1} + 1$.*

Proof. Assume that we have constructed a DT for a binary classification task. By Corollary 1, the DT can be converted to a binary tree, since it classifies for a binary classification problem. Further, let the height that this (converted) binary DT has be h. To estimate SP, for each sensitive attribute the total size is required, $|D_{A=a}|$, as well

as the number of individuals from each (un)privileged group that is classified favorably by the DT. By definition, the first quantity requires 1 histogram query. The latter quantity requires a query for each favorable decision rule in the tree. A branching node that creates one leaf node and one other branching node, adds either an unfavourable or a favourable classification rule to its DT. The most shallow binary tree is schematically shown in Figure A2 in Appendix A. Only 1 histogram query is required for this tree, thus the lower bound for the number of required queries for PAFER is $1 + 1 = 2$. A perfectly balanced binary tree is shown in Figure A3 in Appendix A. In this case, the number of favourable decision rules in the tree is $\frac{1}{2}2^h = 2^{-1}2^h = 2^{h-1}$. As, by the properties of PAFER, each split that creates two leaf nodes adds both a favourable and an unfavourable classification rule to the DT. In a perfectly balanced tree (amongst others), all nodes at $h - 1$ are such nodes. Half of the nodes at h (i.e., leaf nodes) are thus favourable and half are unfavourable. This amounts to 2^{h-1} histogram queries. The upper bound for the number of required queries for PAFER is thus $2^{h-1} + 1$. □

5 Evaluation

This section evaluates the proposed method in the previous section, PAFER. Firstly, subsection 5.1 describes the experimental setup, detailing the used datasets and the two experiments. Secondly, subsection 5.2 displays and discusses the results of the experiments.

5.1 Experimental Setup

This section describes the experiments that answer the research questions. The first subsection describes these datasets and details their properties. The subsections thereafter describe the experiments in order, corresponding to the research question they aim to answer.

5.1.1 Datasets

This subsection describes the datasets that are used to answer the research questions. The datasets form the test bed on which the experiments can be performed. We chose three datasets, namely Adult [38], COMPAS [13] and German [39]. They are all well known in the domain of fairness for ML, and can be considered benchmark datasets. Importantly, they vary in size and all model a binary classification problem, enabling the calculation of various fairness metrics. The datasets are publicly available and pseudonymized; every privacy concern is thus merely for the sake of argument. Table 3 shows some other important characteristics of each dataset.

Table 3 Properties of the three chosen publicly available datasets.

Dataset	# Rows	# Features	Sens. attrib.	Task
Adult	48842	14	race, sex, age, country of origin	Income $> \$50\,000$
COMPAS	7214	53	race, sex, age	Recidivism after 2 years
German	1000	24	race, sex, age, country of origin	Loan default

Pre-processing

This paragraph describes each pre-processing step for every chosen dataset. Some pre-processing steps were taken for all datasets. In every dataset, the sensitive attributes were separated from the training set. Every sensitive attribute except age was binarized, distinguishing between privileged and unprivileged groups. The privileged individuals were White men who lived in their original country of birth, and the unprivileged individuals were those who were not male, not White or lived abroad. We now detail the pre-processing steps that are dataset-specific.

Adult. The Adult dataset comes with a predetermined train and test set. The same pre-processing steps were performed on each one. Rows that contained missing values were removed. The "fnlwgt" column, which stands for "final weight" was removed

23

as it is a relic from a previously trained model and unrelated features might cause overfitting. The final number of rows was 30162 for the train set and 15060 for the test set.

COMPAS. The COMPAS article analyzes two datasets, one for general recidivism and one for violent recidivism [13]. Only the dataset for general recidivism was used. This is a dataset with a large number of features (53), but by following the feature selection steps from the article[9], this number reduced to eleven, of which three are sensitive attributes. The other pre-processing step in the article is to remove cases in which the arrest date and COMPAS screening date are more than thirty days apart. The features that contain dates are then converted to just the year, rounded down. Missing values are imputed with the median value for that feature. Replacing missing values with the median value ensures that no out-of-the-ordinary values are added to the dataset. The final number of rows was 4115 for the train set and 2057 for the test set, totalling 6172 rows.

German. The German dataset is a nearly perfect dataset for our purposes; it contains no missing values. The gender attribute is encoded in the marital status attribute, which required separation. The final number of rows is 667 for the train set and 333 for the test set, totalling 1000 rows.

5.1.2 Experiment 1: Comparison of DP mechanisms for PAFER

Experiment 1 was constructed such that it answered **RQ1**; what DP mechanism is optimal for what privacy budget? The best performing shallow DT was constructed for each dataset, using grid search and cross-validation, optimizing for balanced accuracy. The height of the DT, the number of leaf nodes and the number of selected features were varied. The parameter space can be described as $\{2, 3, 4\} \times \{3, 4, 5, 6, 7, 8, 9, 10, 11, 12\} \times \{\text{sqrt, all}, \log_2\}$, constituting tuples of (height, # leaf nodes, # selected features). The out-of-sample SP of each DT is also provided in Table 4. The experiment

---

[9]https://github.com/propublica/compas-analysis/blob/master/Compas%20Analysis.ipynb

was repeated fifty times with this same DT, such that the random noise, introduced by the DP mechanisms, could be averaged. Initially, we considered the Laplacian, Exponential and Gaussian mechanisms for the comparison. However, after exploratory testing, we deemed the Gaussian mechanism to perform too poorly to be included. Table 5 shows some of these preliminary results. The performance of each mechanism was measured using the Average Absolute Statistical Parity Error (AASPE), defined as follows:

$$\text{AASPE} = \sum_{i}^{\#\,\text{runs}} \frac{1}{\#\,\text{runs}} |SP_i - \widehat{SP_i}|, \tag{10}$$

where # runs is the number of times the experiment was repeated, SP_i and $\widehat{SP_i}$ are the true and estimated SP of the ith run, respectively. The metric was calculated out of sample, i.e., on the test set. The differences in performance were compared using an independent t-test. The privacy budget was varied such that forty equally spaced values were tested with $\varepsilon \in (0, \frac{1}{2}]$. Initial results showed that privacy budgets larger than $\frac{1}{2}$ offered very marginal improvements. Table 5 shows a summary of the preliminary results for Experiment 1. Experiment 1 was performed for both ethnicity, sex and the two combined. The former two sensitive features were encoded as a binary feature, distinguishing between a privileged (white, male) and an unprivileged (non-white, non-male) group. The latter sensitive feature was encoded as a quaternary feature, distinguishing between a privileged (white-male) and an unprivileged (non-white or non-male) group. Whenever a query answer is invalid, as described in subsubsection 4.2.4, a policy must be chosen for calculation of the SP metric. In Experiment 1, the uniform answer approach was chosen, i.e., the size of the group was made to be proportional to the number of sensitive features and the total size. The proportion of invalid query answers, i.e., $\frac{\#\,\text{invalid answers}}{\#\,\text{total answers}}$, was also tracked during this experiment. This invalid value ratio provides some indication of how much noise is added to the query answers.

25

5.1.3 Experiment 2: Comparison of different DTs for PAFER

Experiment 2 was constructed in such a way that it answered **RQ2**; what is the effect of DT hyperparameters on the performance of PAFER? The `minleaf` value was varied such that eighty equally spaced values were tested with `minleaf` $\in (0, \frac{1}{5}]$. In the initial results, shown in Table 6, when the `minleaf` value exceeded $\frac{1}{5}$, the same split was repeatedly chosen for each dataset. Even though `minleaf` $< \frac{1}{2}$, a risk still occurs that one numerical feature is split over and over, which hinders interpretability. Therefore, each numerical feature is categorized by binning it. The bins were established by generating five different DTs, that used all the numerical features. An average splitting value was determined for each height across DTs, that was kept at a maximum of seven[10]. Averages were rounded to the nearest natural number. The privacy budget was defined such that $\varepsilon \in \{\frac{1}{20}, \frac{2}{20}, \frac{3}{20}, \frac{4}{20}, \frac{5}{20}\}$. The performance was again measured in

---

[10]Based on the "Magic Number 7", as humans can generally hold seven $\pm$ two pieces of information in memory, and thus, also, seven rule clauses in memory [40].

Table 4 The out-of-sample Statistical Parity of each constructed DT in Experiment 1. Note that the Sex-Ethnicity attribute is encoded using four (un)privileged groups, and the others are encoded using two.

Dataset A	Adult	COMPAS	German
Ethnicity	0.65	0.78	0.90
Sex	0.30	0.84	0.90
Sex-Ethnicity	0.23	0.72	0.78

Table 5 Preliminary results for Experiment 1 with larger privacy budgets. The Gaussian mechanism was tested with $\delta = \frac{1}{1000}$. The performance was measured using the AASPE on the Adult dataset.

ε	Laplacian	Exponential	Gaussian	Gauss. Invalid Ratio
0.50	0.02320	0.34350	-	-
0.55	0.02065	0.30289	0.32484	0.330
0.60	0.01872	0.25780	0.28916	0.305
0.65	0.01329	0.27566	0.26961	0.230
0.70	0.01026	0.30831	0.27676	0.250
0.75	0.01353	0.32444	0.26572	0.260

Table 6 Preliminary results for Experiment 2. The performance was measured using AASPE on the Adult dataset. The results were averaged over 25 runs.

minleaf \ ε	$\frac{1}{20}$	$\frac{1}{10}$	$\frac{3}{20}$	$\frac{1}{5}$	$\frac{1}{4}$
$\frac{1}{5}$	.0828	.0532	.0407	.0323	.0194
$\frac{1}{4}$	.0711	.0325	.0235	.0187	.0119
$\frac{3}{10}$	.0486	.0282	.0188	.0149	.0128

AASPE, as shown in Equation 10. The metric was measured out of sample, i.e., on the test set. The performance for each `minleaf` value was averaged over fifty potentially different DTs. The same invalid query answer policy was chosen as in Experiment 1, replacing each invalid query answer with the uniformly distributed answer. The performance of PAFER was compared with a baseline that uniformly randomly guesses an SP value in the interval $[0, 1)$. A one-sided t-test determined whether PAFER significantly outperformed the random baseline.

Experiment 2.1: Interaction between ε and `minleaf` hyperparameters

The SP metric is also popular due to its legal use in the United States, where it is used to determine compliance with the 80%-rule [16]. Thus, the UAR (Unweighted Average Recall) of PAFER was calculated for each `minleaf` value, to obtain an indication of whether PAFER was able to effectively measure this compliance. UAR is the average of class-wise recall scores. This was done by rounding each estimation down to its decimal value, thus creating 'classes' that the UAR could be calculated for. To gain more intuition about the interaction between ε and `minleaf` value, the following metric was calculated for each combination:

$$\text{UAR} - \text{AASPE} = \sum_{c \in C} \frac{1}{|C|} \times \frac{\#\,\text{true}\,c}{\#c} - \sum_{i}^{\#\,\text{runs}} \frac{1}{\#\,\text{runs}} |SP_i - \widehat{SP_i}| \qquad (11)$$

Ideally, AASPE is minimized and UAR is maximized, thus maximizing the metric shown in Equation 11. Besides the metric, the experimental setup was identical to

27

Fig. 1 A comparison of the Laplacian and Exponential DP mechanism for different privacy budgets ε. When indicated, from the critical ε value to $\varepsilon = \frac{1}{2}$, the Laplacian mechanism performs significantly better ($p < .05$) than the Exponential mechanism. The uncertainty is pictured in a lighter color around the average.

Experiment 2. Therefore, the same DTs were used for this experiment, only the metrics differed.

5.2 Results

This section describes the results of the experiments and also provides an analysis of the results. Results are ordered to match the order of the experiments.

5.2.1 Results for Experiment 1

Figure 1 answers **RQ1**; the Laplacian mechanism outperforms the Exponential mechanism on seven out of the nine analyses. The Laplacian mechanism is significantly better even at very low privacy budgets ($\varepsilon < 0.1$). The error of the mechanism generally decreases steadily, as the privacy budget increases. This is expected behavior.

As the privacy budget increases, the amount of noise decreases. The Laplacian mechanism performs the best on the Adult and COMPAS datasets, because their invalid value ratio is small, especially for $\varepsilon > \frac{1}{10}$.

The Exponential mechanism performs relatively stable across analyses, however, its performance is generally bad, with errors even reaching the maximum possible error for the German dataset. This is probably due to the design of the utility function, $u_D(r)$, which does not differentiate enough between good and bad answers. Moreover, the Exponential mechanism consistently adds even more noise because it guarantees valid query answers. The Laplacian mechanism does not give these guarantees, and thus relies less on the chosen policy, as described in subsubsection 4.2.4. The mechanism performs somewhat decently on the intersectional analysis for the Adult dataset. This is due to it being an easy prediction task, the Laplacian mechanism starts at a similarly low error.

Figure 1 shows that the invalid value ratio consistently decreases with the privacy budget. This behavior is expected, given that the amount of noise decreases as the privacy budget increases. The invalid value ratio is the largest in the intersectional analyses because then the sensitive attributes are quaternary. The difference between the invalid value ratio progression for the Adult and COMPAS datasets is small, whereas the difference between COMPAS and German is large. Thus, smaller datasets only become problematic for PAFER between 6000 and 1000 rows. Experiment 2 sheds further light on this question.

For the two cases where the Exponential mechanism is competitive with the Laplacian mechanism, the invalid value ratio is also large. When the dataset is small, the sensitivity is relatively larger, and the chances of invalid query answers are larger. Note that the error is measured out-of-sample, so, for the German dataset, the histogram queries are performed on a dataset of size 333. This effect is also visible in the next experiment.

Table 7 Results for Experiment 2 on the Adult dataset and the binary ethnicity sensitive attribute. A * indicates that PAFER performed significantly better than the random baseline.

minleaf \ ε	$\frac{1}{20}$	$\frac{1}{10}$	$\frac{3}{20}$	$\frac{1}{5}$	$\frac{1}{4}$
$\frac{1}{1000}$	$p < .001^*$	$p < .001^*$	$p < .001^*$	$p = .001^*$	$p = .039^*$
$\frac{1}{100}$	$p < .001^*$	$p < .001^*$	$p < .001^*$	$p < .001^*$	$p < .001^*$
$\frac{1}{5}$	$p < .001^*$	$p < .001^*$	$p < .001^*$	$p < .001^*$	$p < .001^*$

Table 8 Results for Experiment 2 on the Adult dataset and the binary sex sensitive attribute. A * indicates that PAFER performed significantly better than the random baseline. A $\lozenge$ indicates that the random baseline performed significantly better than PAFER.

minleaf \ ε	$\frac{1}{20}$	$\frac{1}{10}$	$\frac{3}{20}$	$\frac{1}{5}$	$\frac{1}{4}$
$\frac{1}{1000}$	$p = .999\lozenge$	$p = .87$	$p = .57$	$p = .02^*$	$p = .02^*$
$\frac{1}{100}$	$p < .001^*$	$p < .001^*$	$p < .001^*$	$p < .001^*$	$p < .001^*$
$\frac{1}{5}$	$p < .001^*$	$p < .001^*$	$p = .02^*$	$p = .02^*$	$p < .001^*$

Table 9 Results for Experiment 2 on the Adult dataset and the quaternary sex-ethnicity sensitive attribute. A * indicates that PAFER performed significantly better than the random baseline.

minleaf \ ε	$\frac{1}{20}$	$\frac{1}{10}$	$\frac{3}{20}$	$\frac{1}{5}$	$\frac{1}{4}$
$\frac{1}{1000}$	$p < .001^*$	$p < .001^*$	$p < .001^*$	$p < .001^*$	$p < .001^*$
$\frac{1}{100}$	$p < .001^*$	$p < .001^*$	$p < .001^*$	$p < .001^*$	$p < .001^*$
$\frac{1}{5}$	$p < .001^*$	$p < .001^*$	$p < .001^*$	$p < .001^*$	$p < .001^*$

Table 10 Results for Experiment 2 on the COMPAS dataset and the binary ethnicity sensitive attribute. A * indicates that PAFER performed significantly better than the random baseline.

minleaf \ ε	$\frac{1}{20}$	$\frac{1}{10}$	$\frac{3}{20}$	$\frac{1}{5}$	$\frac{1}{4}$
$\frac{1}{1000}$	$p < .001^*$	$p < .001^*$	$p < .001^*$	$p < .001^*$	$p < .001^*$
$\frac{1}{100}$	$p < .001^*$	$p < .001^*$	$p < .001^*$	$p < .001^*$	$p < .001^*$
$\frac{1}{5}$	$p < .001^*$	$p < .001^*$	$p < .001^*$	$p < .001^*$	$p < .001^*$

5.2.2 Results for Experiment 2

Table 7 through Table 12 show the results for Experiment 2. The tables clearly show that PAFER generally significantly outperforms the random baseline. For small privacy budgets ($\varepsilon \leq \frac{1}{10}$) and small minleaf values (minleaf $= \frac{1}{1000}$), PAFER does not

Table 11 Results for Experiment 2 on the COMPAS dataset and the binary sex sensitive attribute. A * indicates that PAFER performed significantly better than the random baseline.

minleaf \ ε	$\frac{1}{20}$	$\frac{1}{10}$	$\frac{3}{20}$	$\frac{1}{5}$	$\frac{1}{4}$
$\frac{1}{1000}$	$p = .94$	$p = .46$	$p = .27$	$p = .015^*$	$p < .001^*$
$\frac{1}{100}$	$p < .001^*$	$p < .001^*$	$p < .001^*$	$p < .001^*$	$p < .001^*$
$\frac{1}{5}$	$p < .001^*$	$p < .001^*$	$p < .001^*$	$p < .001^*$	$p < .001^*$

Table 12 Results for Experiment 2 on the COMPAS dataset and the quaternary sex-ethnicity sensitive attribute. A * indicates that PAFER performed significantly better than the random baseline. A $\Diamond$ indicates that the random baseline performed significantly better than PAFER.

minleaf \ ε	$\frac{1}{20}$	$\frac{1}{10}$	$\frac{3}{20}$	$\frac{1}{5}$	$\frac{1}{4}$
$\frac{1}{1000}$	$p = 1\Diamond$	$p = 1\Diamond$	$p = 1\Diamond$	$p = 1\Diamond$	$p = .98\Diamond$
$\frac{1}{100}$	$p = .38$	$p < .001^*$	$p < .001^*$	$p < .001^*$	$p < .001^*$
$\frac{1}{5}$	$p = 0.99\Diamond$	$p < .001^*$	$p < .001^*$	$p < .001^*$	$p < .001^*$

strictly perform better, for instance in Table 11. PAFER is even significantly outperformed by the random baseline in some cases, such as in Table 8 and Table 12, for similarly small values of ε and minleaf. PAFER thus performs poorly with a small privacy budget, but also on less interpretable DTs. When the minleaf value of a DT is small, it generally has more branches and branches are longer, as it takes more splits to reach the desired minleaf size. Both of these factors worsen the interpretability of a DT [41].

Other factors negatively impacting the performance of PAFER are a small dataset size and the number of (un)privileged groups. Therefore, the results for the German dataset are omitted, as PAFER is entirely outperformed by the random baseline. This also occurs in Table 12, for all ε and minleaf $= \frac{1}{1000}$. This is due to the smaller leaf nodes, but also due to the smaller dataset (N = 6000), and the quaternary sex-ethnicity sensitive attribute. This reduces the queried quantities even further, resulting in worse performance for PAFER. Then, the (un)privileged group sizes are closer to zero per rule, which increases the probability of invalid query answers. PAFER's worse

performance on smaller datasets, and less interpretable DTs is a clear limitation of the method.

For the sake of succinctness, the results and respective plots for Experiment 2.1 are given in Appendix B. This final experiment also replicates some of the results of Experiment 1 and Experiment 2. The middle plot in Figure A4 through Figure A9 shows that PAFER with the Laplacian mechanism performs better for larger privacy budgets. These plots also show the previously mentioned trade-off between interpretability and performance of PAFER; the method performs worse for smaller `minleaf` values. Lastly, the performance is generally lower for the COMPAS dataset, which holds fewer instances.

6 Conclusion & Future Work

This section concludes the work with answers to the research questions in subsection 6.1, summarizes the entire work in subsection 6.2, and provides suggestions for future work in subsection 6.3.

6.1 Answers to the Research Questions

This section will answer the research questions (RQs) and research subquestions (RSQs), as posed in subsection 1.1.

RQ1 What is the optimal privacy mechanism that preserves privacy and minimizes average Statistical Parity error?

The optimal DP mechanism in Experiment 1 was the Laplacian mechanism, as shown in Figure 1. It performed optimally, in the sense that it achieved a low AASPE at small privacy budgets. This varied from 0.05 error at $\varepsilon = 0.1$, to an error of 0.1 at $\varepsilon = 0.25$. The preliminary results showed that the Gaussian mechanism was also far from optimal, even for large privacy budgets (Table 5).

RSQ1.1 Is there a statistically significant mean difference in Absolute Statistical Parity error between the Laplacian mechanism and the Exponential mechanism?

Yes, the Laplacian mechanism significantly outperformed the Exponential mechanism at very low privacy budgets, on seven out of the nine performed analyses (Figure 1). The Gaussian mechanism proved also to be of no match for the Laplacian mechanism, even at large privacy budgets (Table 5).

RQ2 Is there a statistically significant difference between the Statistical Parity errors of PAFER compared to other benchmarks for varying Decision Tree hyperparameter values?

Yes, for nearly all trials in Experiment 2, there was a significant difference in error between PAFER and the random baseline.

RSQ2.1 At what fractional `minleaf` value is PAFER significantly better at estimating Statistical Parity than a random baseline?

The answer depends on the sensitive attribute that is analyzed and the dataset. In Experiment 2, for the Adult dataset, a `fractional minleaf` value of $\frac{1}{100}$ ensured that PAFER significantly outperformed the random baseline, (Table 9). For the COMPAS dataset and intersectional analysis, a privacy budget of $\varepsilon = \frac{1}{20}$ was not enough to statistically prove that PAFER outperformed the random baseline (Table 12).

6.2 Summary

This work has shed light on the trade-offs between fairness, privacy and interpretability, by introducing a novel, privacy-aware fairness estimation method called PAFER. There is a natural tension between the estimation of fairness and privacy, given that sensitive attributes are required to calculate fairness. This applies also to interpretable, rule-based methods. The proposed method, PAFER, alleviates some of this tension. PAFER should be applied on a DT in a binary classification setting, at the end of a development cycle. PAFER guarantees privacy using mechanisms from DP, allowing it to measure SP for DTs.

We showed that the minimum number of required queries for PAFER is 2. We also showed that the maximum number of queries depends on the height of the DT via $2^{h-1} + 1$, where h is the height.

In our experimental comparison of several DP mechanisms, PAFER showed to be capable of accurately estimating SP for low privacy budgets ($\varepsilon = \frac{1}{10}$) when used with the Laplacian mechanism. This confirms that the calculation of SP for DTs while respecting privacy is possible using PAFER.

Experiment 2 showed that the smaller the leaf nodes of the DT are, the worse the performance is. PAFER thus performs better for more interpretable DTs; as the smaller the `minleaf` value is, the less interpretable a DT is.

Future work can look into other types of DP mechanisms to use with PAFER, and other types of fairness metrics, e.g. EOdd.

6.3 Limitations & Future Work

This section describes some avenues that could be further explored regarding PAFER, with an eye on the limitations that became apparent from the experimental results. We suggest an extension of PAFER that can adopt two other new fairness metrics in subsubsection 6.3.1 and suggest examining the different parameters of the PAFER algorithm in subsubsection 6.3.2.

6.3.1 Other fairness metrics

The most obvious research avenue for PAFER is the extension to support other fairness metrics. SP is a popular, but simple metric that is not correct in every scenario. We thus propose two other group fairness metrics that are suitable for PAFER. However, with the abundance of fairness metrics, multiple other suitable metrics are bound to exist.

The EOdd metric compares the acceptance rates across (un)privileged groups and dataset labels. In our scenario (subsection 4.1), we assume to know the dataset labels,

as this is required for the construction of a DT. Therefore, by querying the sensitive attribute distributions for favorably classifying rules, only for those individuals for which $Y = y$, PAFER can calculate EOdd. Since these groups are mutually exclusive, ε does not have to be shared. Since EOpp is a variant of EOdd, this can naturally also be measured using this approach. A downside is that the number of queries is multiplied by a factor of two, which hinders performance. However, this is not much of an overhead because it is only a constant factor.

6.3.2 Other input parameters

Examining the input parameters of the PAFER estimation algorithm in Algorithm 1, two clear candidates for further research become visible. These are the DP mechanism, $\mathcal{A}$ and the model that is audited, DT. Paragraph 6.3.2 and paragraph 6.3.2 discuss these options.

The Differential Privacy mechanism

The performance of other DP mechanisms can be experimentally compared to the currently examined mechanisms, using the experimental setup of Experiment 1. Experiment 2 shows that there is still room for improvement, as a random guessing baseline significantly outperforms the Laplacian mechanism on multiple occasions.

The work of Hamman et al. in [33] shows promising results for a simple SP query. They use a DP mechanism based on smooth sensitivity [42]; a sensitivity that adds data-specific noise to guarantee DP. If this DP mechanism could be adopted for histogram queries, PAFER might improve in accuracy. Currently, PAFER improves poorly on less interpretable DTs. An improvement in accuracy might also enable PAFER to audit less interpretable DTs.

The audited model

PAFER, as the name suggests, is currently only suited for rule-based systems, and in particular DTs. Further research could look into the applicability of PAFER for other rule-based systems, such as fuzzy-logic rule systems [43], rule lists [44] and association rule data mining [45]. The main point of attention is the distribution of the privacy budget. For DTs, only one rule applies to each person, so PAFER can query all rules. For other rule-based methods, this might not be the case.

Aytekin made the connection between Neural Networks and DTs explicit, showing that for any activation function, a Neural Network can be written as a DT [46]. Applying PAFER to extracted DTs from Neural Networks could also be a future research direction. However, the Neural Network must have a low number of parameters, or else the associated DT would be very tall. DTs with a tall height work worse with PAFER, so the applicability is limited.

Declarations

- Funding: The research leading to this article was conducted during an internship at the Dutch Central Government Audit Service (ADR) as part of the Utrecht University MSc thesis study of the first author.
- Conflict of interest/Competing interests: Authors declare no conflict of interest.
- Ethics approval: Not applicable.
- Consent to participate: Not applicable.
- Consent for publication: Authors received written consent from the ADR for the publication of this work.
- Availability of data and materials: The datasets, Adult [38], COMPAS [13] and German [39], used in the study are publicly available.
- Code availability: Codes of PAFER will be made publicly available upon acceptance of the paper.

- Authors' contributions: The research is conducted by FvdS and supervised by FV and HK. The proposed method is conceptualized by FvdS and matured in consultation with the other authors. Experiments are conducted and discussed by all three authors. The manuscript is written by FvdS, then reviewed and revised by FV and HK. All authors have read and approved the final version.

References

[1] Zhu, L., Qiu, D., Ergu, D., Ying, C., Liu, K.: A study on predicting loan default based on the random forest algorithm. Procedia Computer Science **162**, 503–513 (2019) https://doi.org/10.1016/j.procs.2019.12.017

[2] Bickel, P.J., Hammel, E.A., O'Connell, J.W.: Sex bias in graduate admissions: Data from berkeley: Measuring bias is harder than is usually assumed, and the evidence is sometimes contrary to expectation. Science **187**(4175), 398–404 (1975) https://doi.org/10.1126/science.187.4175.398

[3] Chouldechova, A.: Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. Big Data **5**(2), 153–163 (2017) https://doi.org/10.1089/big.2016.0047 . Accessed 2023-01-13

[4] Barr, A.: Google Mistakenly Tags Black People as 'Gorillas,' Showing Limits of Algorithms. Section: Digits (2015). https://www.wsj.com/articles/BL-DGB-42522 Accessed 2023-02-04

[5] Xenophobic machines: Discrimination through unregulated use of algorithms in the Dutch childcare benefits scandal (2021). https://www.amnesty.org/en/documents/eur35/4686/2021/en/ Accessed 2023-02-04

[6] Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation). Official Journal of the European Union **L 119**, 1–88 (14-04-2016)

[7] Proposal for a regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts. Official Journal of the European Union **COM/2021/206 final**, 1–107 (21-04-2021)

[8] Het Algoritmeregister van de Nederlandse overheid (2022). https://algoritmes.overheid.nl/ Accessed 2023-02-04

[9] Directive 2014/17/eu of the european parliament and of the council of 4 february 2014 on credit agreements for consumers relating to residential immovable property and amending directives 2008/48/ec and 2013/36/eu and regulation (eu). Official Journal of the European Union **L 60/34**, 34–85 (04-02-2014)

[10] Cadwalladr, C., Graham-Harrison, E.: Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach. The Guardian (2018). Chap. News. Accessed 2023-02-04

[11] Losing Face: Two More Cases of Third-Party Facebook App Data Exposure | UpGuard (2019). https://www.upguard.com/breaches/facebook-user-data-leak Accessed 2023-02-04

[12] Berendt, B., Preibusch, S.: Better decision support through exploratory discrimination-aware data mining: foundations and empirical evidence. Artificial Intelligence and Law **22**(2), 175–209 (2014)

https://doi.org/10.1007/s10506-013-9152-0

[13] Mattu, S., Larson, J., Kirchner, L., Angwin, J.: Machine Bias (2016). https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing Accessed 2023-01-06

[14] Verma, S., Rubin, J.: Fairness definitions explained. In: Proceedings of the International Workshop on Software Fairness, pp. 1–7. ACM, Gothenburg Sweden (2018). https://doi.org/10.1145/3194770.3194776

[15] Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R.: Fairness through awareness. In: Proceedings of the 3rd Innovations in Theoretical Computer Science Conference On - ITCS '12, pp. 214–226. ACM Press, Cambridge, Massachusetts (2012). https://doi.org/10.1145/2090236.2090255

[16] Questions and answers to clarify and provide a common interpretation of the uniform guidelines on employee selection procedures. FEDERAL REGISTER **44**(43) (01-03-1979)

[17] Hardt, M., Price, E., Price, E., Srebro, N.: Equality of Opportunity in Supervised Learning. In: Advances in Neural Information Processing Systems, vol. 29, pp. 3315–3323. Curran Associates, Inc., Red Hook, New York (2016). https://proceedings.neurips.cc/paper/2016/hash/9d2682367c3935defcb1f9e247a97c0d-Abstract.html Accessed 2023-01-10

[18] Bertsimas, D., Dunn, J.: Optimal classification trees. Machine Learning **106**, 1039–1082 (2017) https://doi.org/10.1007/s10994-017-5633-9

[19] Oliver, J.J., Hand, D.: Averaging over decision stumps. In: Machine Learning: ECML-94: European Conference on Machine Learning Catania, Italy, April 6–8, 1994 Proceedings 7, pp. 231–241 (1994). https://doi.org/10.1007/3-540-57868-4

61 . Springer

[20] Molnar, C.: Interpretable Machine Learning, 2nd edn. (2022). https://christophm.github.io/interpretable-ml-book

[21] Borisov, V., Leemann, T., Seßler, K., Haug, J., Pawelczyk, M., Kasneci, G.: Deep neural networks and tabular data: A survey. IEEE Transactions on Neural Networks and Learning Systems (2022) https://doi.org/10.1109/TNNLS.2022.3229161

[22] Dwork, C.: Differential privacy. In: 33rd International Colloquium, ICALP 2006 on Automata, Languages and Programming, pp. 1–12. Springer, Berlin, Heidelberg (2006). https://doi.org/10.1007/11787006_1

[23] Dwork, C., Roth, A.: The Algorithmic Foundations of Differential Privacy. Foundations and Trends in Theoretical Computer Science **9**(3-4), 211–407 (2013) https://doi.org/10.1561/0400000042 . Accessed 2023-01-27

[24] McSherry, F., Talwar, K.: Mechanism design via differential privacy. In: 48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07), pp. 94–103 (2007). https://doi.org/10.1109/FOCS.2007.66 . IEEE

[25] Kamiran, F., Calders, T., Pechenizkiy, M.: Discrimination Aware Decision Tree Learning. In: 2010 IEEE International Conference on Data Mining, pp. 869–874. IEEE, Sydney, Australia (2010). https://doi.org/10.1109/ICDM.2010.50

[26] Ausiello, G., Crescenzi, P., Gambosi, G., Kann, V., Marchetti-Spaccamela, A., Protasi, M.: Complexity and Approximation: Combinatorial Optimization Problems and Their Approximability Properties. Springer, Berlin, Heidelberg (1999). https://doi.org/10.1007/978-3-642-58412-1
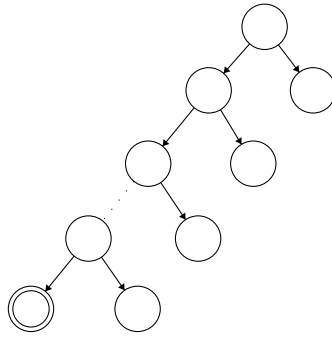
[27] Linden, J., Weerdt, M., Demirović, E.: Fair and optimal decision trees: A dynamic programming approach. In: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A. (eds.) Advances in Neural Information Processing Systems, vol. 35, pp. 38899–38911. Curran Associates, Inc., Red Hook, New York (2022). https://proceedings.neurips.cc/paper_files/paper/2022/file/fe248e22b241ae5a9adf11493c8c12bc-Paper-Conference.pdf

[28] Jo, N., Aghaei, S., Benson, J., Gómez, A., Vayanos, P.: Learning Optimal Fair Classification Trees. arXiv. arXiv:2201.09932 [cs, math] (2022). http://arxiv.org/abs/2201.09932 Accessed 2022-12-08

[29] Mohammed, N., Barouti, S., Alhadidi, D., Chen, R.: Secure and Private Management of Healthcare Databases for Data Mining. In: 2015 IEEE 28th International Symposium on Computer-Based Medical Systems, pp. 191–196. IEEE, Sao Carlos, Brazil (2015). https://doi.org/10.1109/CBMS.2015.54

[30] Fletcher, S., Islam, M.Z.: Decision Tree Classification with Differential Privacy: A Survey. ACM Computing Surveys **52**(4), 1–33 (2020) https://doi.org/10.1145/3337064

[31] Lahoti, P., Beutel, A., Chen, J., Lee, K., Prost, F., Thain, N., Wang, X., Chi, E.: Fairness without Demographics through Adversarially Reweighted Learning. In: Advances in Neural Information Processing Systems, vol. 33, pp. 728–740. Curran Associates, Inc., Red Hook, New York (2020). https://proceedings.neurips.cc/paper/2020/hash/07fc15c9d169ee48573edd749d25945d-Abstract.html Accessed 2022-11-21

[32] Zhao, T., Dai, E., Shu, K., Wang, S.: Towards Fair Classifiers Without Sensitive Attributes: Exploring Biases in Related Features. In: Proceedings of the Fifteenth ACM International Conference on Web Search And Data Mining, pp. 1433–1442.

ACM, Virtual Event AZ USA (2022). https://doi.org/10.1145/3488560.3498493

[33] Hamman, F., Chen, J., Dutta, S.: Can querying for bias leak protected attributes? achieving privacy with smooth sensitivity. In: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, pp. 1358–1368 (2023). https://doi.org/10.1145/3593013.3594086

[34] Candès, E.J., Wakin, M.B.: An introduction to compressive sampling. IEEE signal processing magazine **25**(2), 21–30 (2008) https://doi.org/10.1109/MSP.2007.914731

[35] Navada, A., Ansari, A.N., Patil, S., Sonkamble, B.A.: Overview of use of decision tree algorithms in machine learning. In: 2011 IEEE Control and System Graduate Research Colloquium, pp. 37–42 (2011). https://doi.org/10.1109/ICSGRC.2011.5991826

[36] Wang, T., Rudin, C., Doshi-Velez, F., Liu, Y., Klampfl, E., MacNeille, P.: A Bayesian Framework for Learning Rule Sets for Interpretable Classification. Journal of Machine Learning Research **18**(70), 1–37 (2017)

[37] Szczepański, M.: Is data the new oil? competition issues in the digital economy. EPRS in-depth analysis, 1–8 (2020)

[38] Kohavi, R., Becker, B.: UCI Machine Learning Repository: Adult Data Set (2016). https://archive.ics.uci.edu/ml/datasets/adult Accessed 2023-02-05

[39] Hofmann, H.: Statlog (German Credit Data) Data Set (2013). https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data) Accessed 2023-02-05

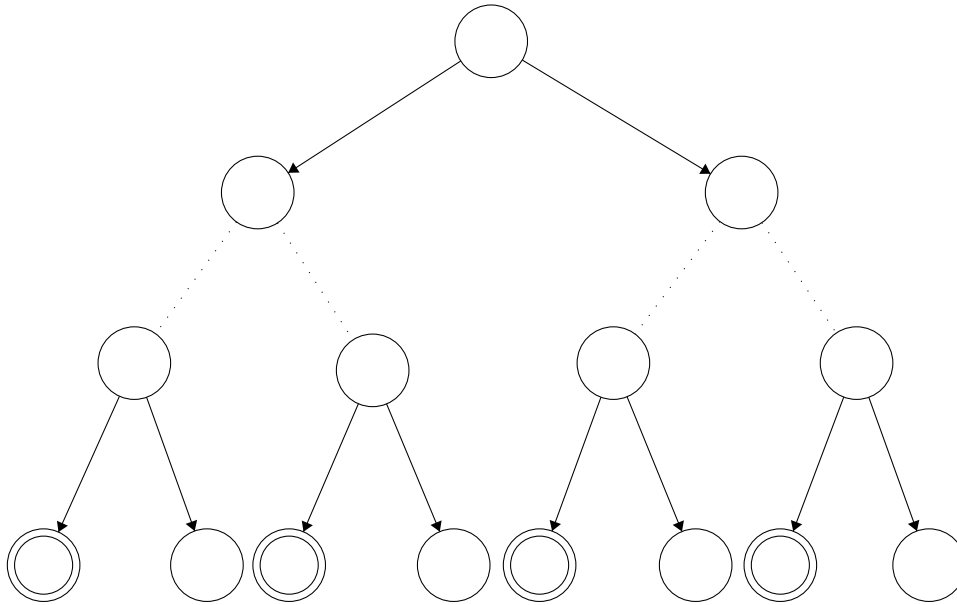[40] Miller, G.A.: The magical number seven, plus or minus two: Some limits on our

capacity for processing information. Psychological Review **63**(2), 81–97 (1956) https://doi.org/10.1037/h0043158 . Accessed 2023-07-11

[41] Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., Herrera, F.: Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion **58**, 82–115 (2020) https://doi.org/10.1016/j.inffus.2019.12.012

[42] Nissim, K., Raskhodnikova, S., Smith, A.: Smooth sensitivity and sampling in private data analysis. In: Proceedings of the Thirty-ninth Annual ACM Symposium on Theory of Computing, pp. 75–84. ACM, San Diego California USA (2007). https://doi.org/10.1145/1250790.1250803

[43] Mendel, J.M.: Uncertain Rule-Based Fuzzy Systems: Introduction and New Directions, 2nd Edition. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-51370-6

[44] Angelino, E., Larus-Stone, N., Alabi, D., Seltzer, M., Rudin, C.: Learning certifiably optimal rule lists for categorical data. Journal of Machine Learning Research **18**(234), 1–78 (2018)

[45] Yazgana, P., Kusakci, A.O.: A literature survey on association rule mining algorithms. Southeast Europe Journal of soft computing **5**(1) (2016) https://doi.org/10.21533/scjournal.v5i1.102

[46] Aytekin, C.: Neural Networks are Decision Trees. arXiv. arXiv:2210.05189 [cs] (2022). http://arxiv.org/abs/2210.05189 Accessed 2023-06-08

Fig. A1 A schematic display of the process by which a binary tree that has non-binary splits can be converted into a binary tree for a binary decision process. The dotted lines ⋅⋅⋅, denote that the pattern of the DT can be repeated an arbitrary number of times.

Fig. A2 The smallest number of favorable decision rules in a decision tree for a binary classification problem. The leaf node with an inner circle denotes a leaf node in which the majority of the individuals are classified favorably in the training set. The dotted line, ⋰, denotes that the pattern can go on indefinitely.

Fig. A3 The largest number of favorable decision rules in a decision tree for a binary classification problem. The leaf nodes with an inner circle denote a leaf node in which the majority of the individuals are classified favorably in the training set. The dotted lines, ⋰, denote that the pattern can go on indefinitely.

Appendix A   Figures Illustrating PAFER's Theoretical Properties

Appendix B   Results for Experiment 2.1

Figure A4 through Figure A9 show the results for Experiment 2.1. Experiment 2.1 shows that PAFER is unreliable in its ability to predict adherence to the 80%-rule.

For some datasets and sensitive attributes, PAFER performs quite well, e.g. reaching around 90% UAR, as shown in Figure A8 and Figure A4. For other datasets and sensitive attributes, PAFER performs rather poorly, reaching no higher than 50% UAR on the Adult dataset with the binary sex attribute, as shown in Figure A5.

Nonetheless, a pattern emerges from Figure A4 through Figure A9 regarding the UAR - AASPE. Of course, PAFER performs better for privacy budgets larger than $\frac{3}{20}$. However, PAFER also performs better for certain minleaf values. The 'hotspot' differs between the Adult and COMPAS dataset, minleaf $= \frac{1}{10}$ and minleaf $= \frac{3}{20}$, respectively, but the range seems to be from $\frac{7}{100}$ to $\frac{1}{5}$. The ideal scenario for PAFER thus seems to be when a privacy budget of at least $\varepsilon = \frac{3}{20}$ is available, and the examined DT has leaf nodes with a fractional minleaf value of at least $\frac{7}{100}$.
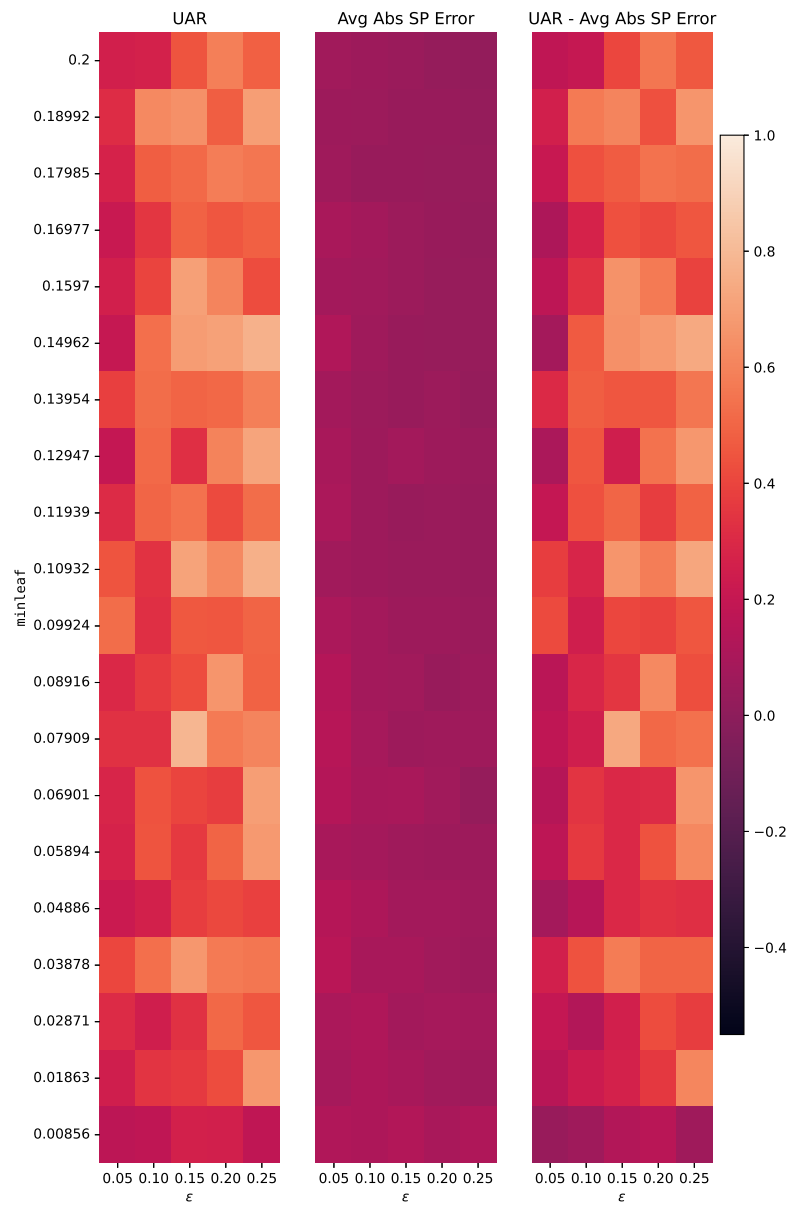
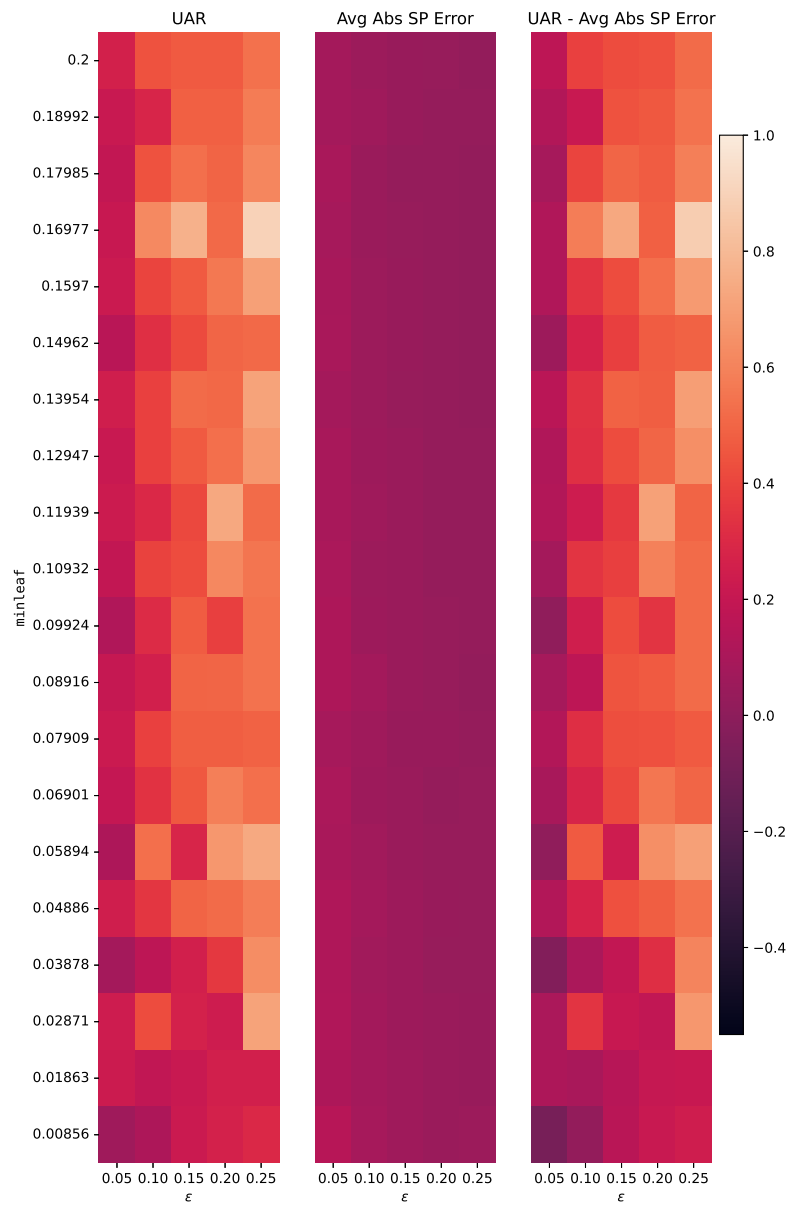**Fig. A4** The hyperparameter space for the Adult dataset and the binary ethnicity attribute.

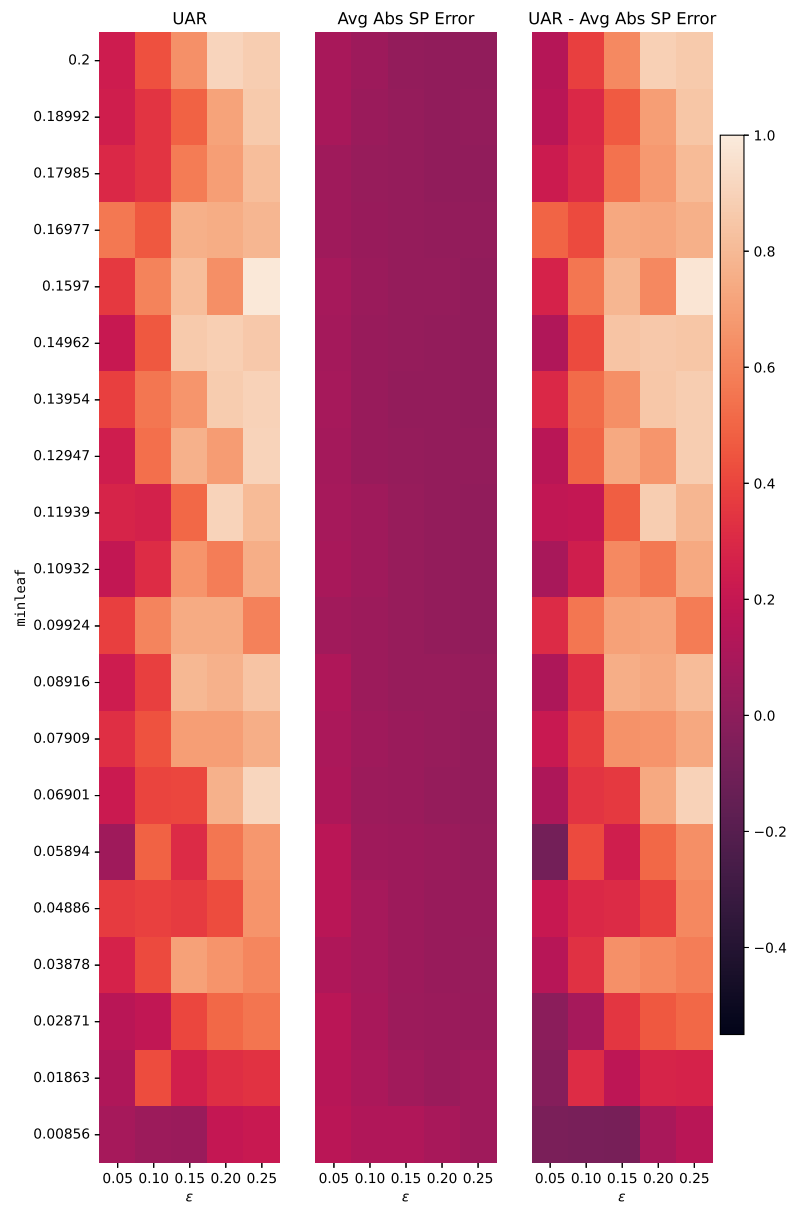**Fig. A5** The hyperparameter space for the Adult dataset and the binary sex attribute.
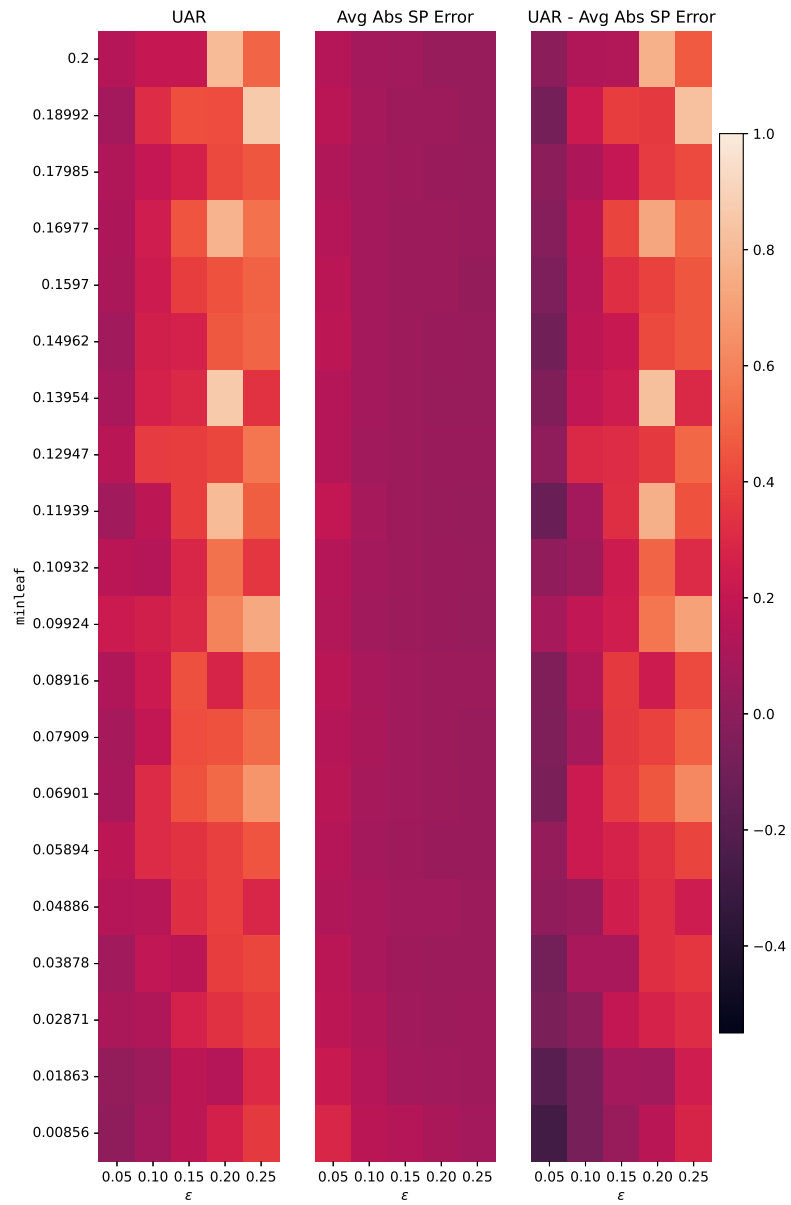
Fig. A6 The hyperparameter space for the Adult dataset and the quaternary sex-ethnicity attribute.

49

Fig. A7 The hyperparameter space for the COMPAS dataset and the binary ethnicity attribute.

Fig. A8 The hyperparameter space for the COMPAS dataset and the binary sex attribute.

Fig. A9 The hyperparameter space for the COMPAS dataset and the quaternary sex-ethnicity attribute.