Responsible Auditing: Privacy Constrained Fairness Estimation for Decision Trees

by

Florian van der Steen

Submitted to the Artificial Intelligence Graduate Program
in partial fulfillment of the requirements for the degree of
Master of Science

Graduate Program in Artificial Intelligence
Utrecht University
2023

Responsible Auditing: Privacy Constrained Fairness Estimation for Decision Trees

APPROVED BY:

dr. Heysem Kaya ....
(Thesis Supervisor)

Fré Vink MSc ....
(Thesis Co-supervisor)

dr. Hakim Qahtan ....

DATE OF APPROVAL: 05-07-2023

ACKNOWLEDGEMENTS

Firstly, I want to thank my supervisors, in particular Heysem Kaya, who helped tremendously and pushed me to write a thesis of this size and scope. I also want to thank Fré Vink, who co-supervised me and helped me get acquainted with the Auditdienst Rijk. My gratitude also goes to Hakim Qahtan putting time into reading my work.

Secondly, I want to thank the people close to me. My girlfriend, Floortje, for enduring me in these stress-induced months; cheering me up and motivating me when the process was rough. My friends with whom I live, for listening to my rants and for lighting my mood. And, of course, my family for supporting me during my entire academic career.

ABSTRACT

Responsible Auditing: Privacy Constrained Fairness Estimation for Decision Trees

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 are more interpretable.

NEDERLANDSE ABSTRACT

Gegevensbescherming en privacy worden steeds crucialer naarmate gegevens waardevoller en potentieel krachtiger worden. Bovendien neemt de druk van regelgeving en de samenleving toe op ontwikkelaars van *Artificial Intelligence* (AI) modellen om ervoor te zorgen dat hun modellen niet discriminerend zijn. Ten slotte is er behoefte aan interpreteerbare, transparante AI modellen voor taken met grote belangen. Over het algemeen vereist het meten van de *fairness* van ieder AI model de gevoelige kenmerken van de individuen in de dataset, waardoor privacy dus in het gedrang komt. In dit werk worden de afwegingen tussen *fairness*, privacy en interpreteerbaarheid verder onderzocht. We onderzoeken specifiek de *Statistical Parity* (SP) van beslisbomen door middel van Differentiële Privacy (DP). Dit zijn alle drie populaire methodes in hun respectievelijke onderzoeksvelden. We stellen een methode voor, genaamd *Privacy-Aware Fairness Estimation of Rules (PAFER)*, dat SP kan schatten terwijl rekening wordt gehouden met DP voor op beslisbomen. DP, waarbij gebruik wordt gemaakt van een derde partij die veilig omgaat met deze gevoelige gegevens, garandeert privacy door ruis toe te voegen aan de gevoelige gegevens. Verschillende ruis mechanismen voor DP worden empirisch vergeleken. We laten experimenteel zien dat met behulp van het Laplace-mechanisme de methode in staat is om SP met een lage fout te schatten, terwijl de privacy van de individuen in de dataset met grote zekerheid wordt gegarandeerd. We tonen ook experimenteel en theoretisch aan dat de methode beter presteert voor beslisbomen die beter interpreteerbaar zijn.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Glossary

AASPE Average Absolute Statistical Parity Error; the error measure used to estimate the performance of the method introduced in this work. 43–45, 50, 58

AI Artificial Intelligence; the scientific field concerned with developing theory and systems that perform tasks requiring intelligence. i, iv, v, 1–3, 15, 20

DP Differential Privacy; A class of methods to share information in a privacy aware manner. iv, v, ix, 20–26, 33, 34, 36, 42, 46, 58–61

DT Decision Tree; An interpretable, rule-based machine learning model that is used for classification tasks. iv, 2, 3, 15–19, 23, 24, 29–31, 33, 35–38, 42–45, 48–50, 59–62, 75, 78

EOdd Equalized Odds; a common fairness definition requiring equality of false positive and true positive rates across groups. 7, 8, 19, 25, 26, 29, 60, 61

EOpp Equality of Opportunity; a common fairness definition requiring equality of true positive rates across groups. 7, 8, 19, 26, 28, 61

ML Machine Learning; the subfield of Artificial Intelligence concerned with building systems that can learn a task without following explicit instructions. i, 1, 2, 4, 5, 10–12, 17, 19, 20, 25, 40, 62

PAFER Privacy-Aware Fairness Estimation of Rules; The proposed and studied method in this work that can estimate fairness while respecting privacy. 3, 4, 30, 31, 33–36, 38, 40, 44, 45, 47–51, 58–62, 78

PrEq Predictive Equality; a common fairness definition requiring equality of false positive rates across groups. 8, 19, 60, 61

SP Statistical Parity; a common fairness definition requiring the equality of positive outcomes across groups. iv, v, 6, 7, 10, 12, 14, 19, 25, 28, 29, 31, 33, 34, 36, 42–45, 59–62

UAR Unweighted Average Recall; A performance metric used for classification tasks that measures the recall for each class. 45, 50

XAI Explainable Artificial Intelligence; the subfield of Artificial Intelligence concerned with making systems that can be interpreted and explained. 1, 2, 15

LIST OF SYMBOLS

A	A sensitive attribute
$\mathcal{A}$	A mechanism to ensure Differential Privacy
c^*	The closest dataset-based counterfactual to an instance
C	The set of possible classes in a classification problem
$d(\cdot\,,\cdot\,)$	A distance measure
D	A dataset
$D_{A=a}$	A dataset in which for all instances $A = a$ holds
h	The height of a Decision Tree
k	The number of splits in a Decision Tree
$\mathcal{K}$	The number of sensitive groups for a sensitive attribute
L	A legitimate feature
$m(\cdot)$	The application of a model m to an instance
M	The set of all possible models
N	The total number of observations in a dataset
$p(\cdot)$	A probability
$q(D)$	A query to a dataset D
$\mathcal{R}$	The set of categorical answers a query can give
S	The score that a Machine Learning model gives to an instance
$u_D(r)$	The utility of a category for a dataset D
$U_D(m)$	The utility of a model for a dataset D
x_i	The ith instance x
x^j	The jth feature of an instance x
Y	The true outcome or label of an instance
$\hat{Y}$	The predicted outcome of an instance
$\hat{X}_{Y \leftarrow y}$	The value of variable X in the world where Y becomes y
δ	The additive term in Differential Privacy
Δ	The global sensitivity

ϵ	The privacy budget
σ	The standard deviation

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 discriminative 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 far going 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. The Dutch Tax & Customs administration used a model for fraud prediction that targeted people with multiple nationalities [5]. This later led to the resignation of the cabinet of the Dutch government [6].

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. In short, similar individuals should be treated similarly, and decisions should not lead to unjust discrimination. Non-discrimination laws for AI exist within the EU [7] and more are upcoming [8]. Partly due to the scandal, the Dutch government now has a register of all the algorithms used within it [9].

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 [10]. 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 [7]. 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 [7]. These attributes, and thus the storage practices thereof, are sensitive. Examples include the religion, marital status and gender of individuals. Another reason for the outcry regarding the fraud prediction system was the irresponsible sharing and storage within the Dutch government of the sensitive private data of thousands of individuals [5]. 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 [11, 12]. This work will investigate these three pillars of Responsible AI, investigating a novel method that is at the intersection of these three themes.

1.1. Problem Statement

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 limitedly available, 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 conflicting goals make for a difficult problem that is also fairly novel. A focus is placed on DTs, a class of interpretable models from XAI since these types of models are likely to be used in a sensitive setting. There are thus four goals in this work: fairness, privacy, interpretability and of course performance.

1.2. 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 values. 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?

(a) **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?

RSQ2.2 At what fractional `minleaf` value is the perfect estimator significantly better at estimating Statistical Parity than PAFER?

1.3. Outline

This work is divided into several Chapters, which each consist of sections and subsections. The upcoming Chapter 2 will cover the related literature and theoretical background that is relevant to this research. Chapter 3 describes the novel method that is proposed in this work. Finally, Chapter 4 describes the performed experiments, their results and thorough analysis, along with future work and a conclusion.

2. Background and Related Work

This chapter discusses work related to the research objectives and provides background to the performed research. Each section belongs to a subset in the Venn diagram that Figure 2.1 shows. It consists of the three main pillars of this work and responsible AI: fairness, interpretability and privacy, along with their intersections.

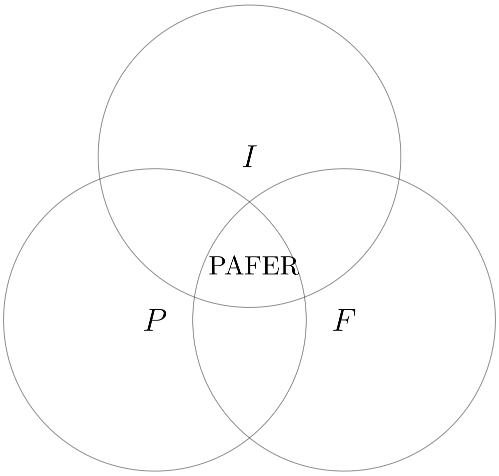

Figure 2.1. The three pillars of this proposal: Privacy (P), Interpretability (I) and Fairness (F) and their intersections $P \cap I$, $P \cap S$, $I \cap S$ and $P \cap I \cap S$ shown in a Venn diagram. Section 2.1, Section 2.2 and Section 2.4 cover the main pillars F, I and P, respectively. Section 2.3 covers $F \cap I$, Section 2.5 covers $P \cap I$, Section 2.6 covers $P \cap F$ and methods that cover all three occur in Section 2.7 and Section 2.6. PAFER, the method proposed in this work, is at the intersection of all three pillars.

2.1. Fairness Definitions

This section discusses several fairness definitions, on an individual level, as well as on a group level. 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, and this is the largest research area within algorithmic fairness.

[1] What exactly is **unjust** discrimination is a social construct and changes over time [13].

In 2016, the number of papers related to fairness surged. Partly, due to the new regulations such as the European GDPR [7] and partly due to a popular article by ProPublica which examined racial disparities in recidivism prediction software [14]. 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.

This section discusses three different types of fairness metrics: group fairness in Subsection 2.1.1, individual fairness in Subsection 2.1.2 and causal fairness in Subsection 2.1.3.

2.1.1. Group Fairness

Group fairness is the most popular type of fairness definition as it relates most closely to unjust discrimination; this subsection explains some of the most popular group fairness definitions. 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. Technically, however, in all group fairness definitions, the sensitive attribute may be any feature.

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, unless mentioned otherwise, we assume the binary classification setting in the following definitions.

2.1.1.1. 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:

$$\mathrm{SP} = p(\hat{Y} = 1|A = 1) - p(\hat{Y} = 1|A = 0), \tag{2.1}$$

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 2.1 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:

$$\text{80\%-rule } = 0.8 \leq \frac{p(\hat{Y} = 1|A = 1)}{p(\hat{Y} = 1|A = 0)} \leq 1.25, \tag{2.2}$$

where the fraction is the SP-ratio. SP is easy to compute and does not require the actual outcome labels. These advantages make it one of the most used fairness definitions.

2.1.1.2. Conditional Statistical Parity. A different version of SP is Conditional Statistical Parity [15]. This definition is similar to SP, except it allows conditioning on some legitimate features, L. Mathematically, Conditional SP is defined as:

$$\text{Conditional SP} = p(\hat{Y} = 1|A = 1, L = l) - p(\hat{Y} = 1|A = 0, L = l), \tag{2.3}$$

where l is the instantiation of the legitimate features L. An example of Conditional SP would be a comparison of the acceptance rates of male and female applicants to a university conditioned on their average final grades. If there is an imbalance, there is an even stronger indication of unjust discrimination than an imbalance in (unconditional) statistical parity. An advantage of this approach is that it is less 'naive' than regular SP. The metric, however, is only useful when every legitimate feature that is conditioned on is properly justified.

2.1.1.3. 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:

$$\text{EOdd} = p(\hat{Y} = 1|Y = y, A = 1) - p(\hat{Y} = 1|Y = y, A = 0), \;\; y \in \{0, 1\}. \tag{2.4}$$

An example of applying EOdd would be to require that both whites and people of color have equal probability to be predicted to not recidivate, under both ground truth conditions, separately. An advantage of EOdd is that, unlike SP, when the predictor is perfect, i.e. $Y = \hat{Y}$, it satisfies EOdd.

2.1.1.4. 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:

$$\text{EOpp} = p(\hat{Y} = 1|Y = 1, A = 1) - p(\hat{Y} = 1|Y = 1, A = 0). \tag{2.5}$$

An example of applying EOpp would be to just require that whites and people of color have equal probability to be predicted to not recidivate given that they did not actually end up recidivating. If there was an imbalance, one group would receive more freedom, even though the other group would equally deserve it based on their actions. An advantage of EOpp is that it is not a bi-objective, and thus is more easily optimized for compared to EOdd.

2.1.1.5. Predictive Equality. A very similar relaxation of EOdd is the fairness definition Predictive Equality (PrEq) [3]. It just requires the equality of the probabilities of wrongly 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, PrEq only requires the latter. Formally:

$$\text{PrEq} = p(\hat{Y} = 1|Y = 0, A = 1) - p(\hat{Y} = 1|Y = 0, A = 0). \tag{2.6}$$

An example of applying PrEq is to require that whites and people of color have equal probability to be predicted to not recidivate given that they did actually end up recidivating. If there is an imbalance, one of the groups would receive more freedom, even though they would abuse it. An advantage of PrEq is that it is not a bi-objective, and thus is more easily optimized for compared to EOdd.

2.1.1.6. Calibration. Calibration, also known as the sufficiency criterion, is a fairness definition based on the scores, S, of a model [18]. It requires the equality of the probabilities of the outcomes across scores and groups. Usually, the scores are binned such that S is partitioned into different score bins s. An extension of calibration is well-calibration and then these equal probabilities should be calibrated to match with the exact score. A model is calibrated if:

$$p(Y = 1|S = s, A = 1) = p(Y = 1|S = s, A = 0) = s, \ \forall s. \tag{2.7}$$

An example of applying well-calibration is requiring that for each (binned) predicted probability of receiving a loan, the fraction of people who actually pay it back is equal to the predicted score for females and males. If there is an imbalance for a certain s, one group is wrongly treated differently by the model, given the actual outcomes. A disadvantage of Calibration is that it requires the scaling of the probabilities of the outcomes, and some models may not allow for this. Calibration is useful in situations when the labels are correctly acquired, as it does not transform any of the bias in the data. All previously mentioned definitions were bias-transforming, as they required equality across predictions.

2.1.1.7. Accurate Coverage. Accurate Coverage is a fairness definition that requires that the prediction probabilities are equal to the outcome rates in the data [19]. Formally:

$$p(\hat{Y} = 1|A = a) = p(Y = 1|A = a), \quad a \in \{0, 1\}. \tag{2.8}$$

For instance, in a loan application setting, when the payback rate of males is 60% and for females is 50% in the data, the classifier should also predict 60% of males and 50% of females to pay back their loan. Accurate Coverage is also a definition that maintains the status quo, as it assumes the correctness of the labels in the dataset.

2.1.1.8. Rawlsian Min-Max. Rawlsian Min-Max fairness is a popular definition in the research area of fairness where sensitive attributes are (partially) unavailable [20, 21]. The definition originates from the work by philosopher John Rawls on distributive justice and his difference principle [22, p.155]. It is based on the idea that the group with the worst utility should be maximal (minimal group–maximal utility). The utility can be any sort of function for a group but is often an evaluation metric of a ML model specifically for that group. Formally:

$$\text{Rawlsian Min-Max} = \arg\max_{m \in M} \min_{a \in \{0,1\}} U_{D_{A=a}}(m), \tag{2.9}$$

where $U_{D_{A=a}}(m)$ is the utility of model m out of all possible models M, applied only on instances from a dataset D that have protected attribute a. This notation was borrowed from [21]. The definition is similar to requiring a minimal performance for each group.

2.1.1.9. Burden. A very different kind of definition but nonetheless group related, is Burden [23]. The definition relates to model-based counterfactuals; a perturbation of an instance such that the classification, $m(x)$, of the instance changes. Given a method that can generate these counterfactuals, the distance can be calculated for each individual between their instance, x, and their nearest counterfactual, c^*. Formally, the definition of Burden then is:

$$\text{Burden} = \frac{1}{|D_{A=1}|} \sum_{x \in D_{A=1}} d(x, c^*) - \frac{1}{|D_{A=0}|} \sum_{x \in D_{A=0}} d(x, c^*) \;\; \forall x \to m(x) = 0, \tag{2.10}$$

where $|D_{A=a}|$ is the size of (un)privileged group a in dataset D and $d(x, c^*)$ is the distance between individual x and its closest counterfactual c^*. The distance is an indication of recourse; the amount of attributes an individual has to change to be classified favorably [24]. An imbalance in Burden implies that, on average, one group must change more of their observed behavior to be correctly classified. Burden can be seen as an extension of SP but counterfactual generation is often a time-heavy process, making the extension costly [25]. Burden can also be used to compare the recourse of two similar individuals, i.e. be used as an individual fairness metric. The next subsection details some of these metrics.

2.1.2. Individual Fairness

Whereas the previous definitions were all based on comparisons between groups, the following individual fairness definitions look at the fairness of a ML model given two individuals. Individual fairness may be preferred as the overall goal of fairness is to obtain equality across individuals and not groups. Although the upcoming definitions often still involve sensitive attributes, they do not partition the population based on them.

2.1.2.1. Causal Discrimination. An intuitive definition of individual fairness is that of Causal Discrimination [26]. A classifier is fair based on Causal Discrimination if the model classifies instances which have the same attribute values to the same class. If individuals only differ on sensitive attributes, they are also considered equal and must also belong to the same class according to the model. Formally:

$$\text{Causal Discrimination} = (x_i = x_j) \rightarrow (m(x_i) = m(x_j)), \quad i \neq j, \tag{2.11}$$

where x_i is the ith instance x. Duplicating instances and changing sensitive attributes is an often-used method to test Causal Discrimination. Although an intuitive definition, ironically, Causal Discrimination is often insufficient as a metric because of the underlying causal influences of the sensitive attributes.

2.1.2.2. Fairness through awareness. Fairness through awareness is based on the idea that similar individuals, who are similar with respect to the classification task, should be treated similarly by the ML model [15]. It is the most well-known individual fairness definition and is sometimes even called individual fairness. Fairness through awareness requires that if two individuals x_1, x_2, are within distance $d_1(x_1, x_2)$ of each other, then their prediction distributions must be no greater than $d_2(x_1, x_2)$ of each other. These distance and distribution difference functions are task-specific and the Dwork et al. give some pointers towards sensible functions [15]. For example, the distance metric could be the normalized difference in age and the distribution difference function could be the difference in predicted probabilities for the positive outcome. An advantage of this metric is that it is general, in fact, causal discrimination can be expressed in terms of fairness through awareness. Moreover, the definition is more fine-grained than statistical parity, as adhering to it also reduces SP in sub-groups [15].

2.1.2.3. Fairness through unawareness. Fairness through unawareness is the idea that blinding a ML model from sensitive features ensures that individuals are all treated fairly. It is thus a procedural fairness definition. This definition is merely an idea, as it does not apply in practice. More often than not, task-related features are correlated with sensitive features, which the model still picks up on [27]. Still, recent legislation requires omitting the sensitive features from training data [8].

2.1.3. Causal Fairness

The final type of fairness definition that is discussed in this section is causal fairness. Causal fairness requires a causal model that represents the causal relations between the features and the outcome. Figure 2.2 gives two examples of such models. Using this causal model one can infer some fairness properties of the classification task. The field often assumes that these causal models were built using expert knowledge, as opposed to using the dataset.

2.1.3.1. Unresolved Discrimination. Unresolved Discrimination directly relates to the structure of the causal model [28]. When each path from the sensitive attribute to the outcome goes only via resolving variables, the causal model satisfies Unresolved Discrimination. Resolving variables are justified causes of the outcome (similar to conditional statistical parity). The right causal model of Figure 2.2 displays no unresolved discrimination, as the only path from vertex A to Y is via R. For this definition, it is necessary that the causal model is correct to not falsely assure or alarm.

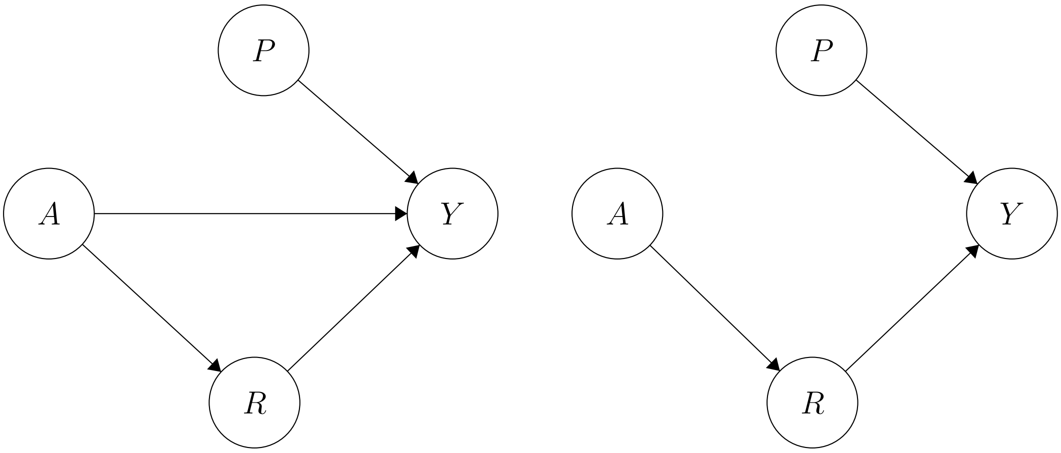

Figure 2.2. Two examples of Causal Graphs. A = Sensitive Attribute, Y = Outcome, R = Resolving Feature and P = Proxy Feature.

2.1.3.2. Proxy Discrimination. Proxy Discrimination also directly relates to the structure of the causal model [28]. Proxy features are those that are highly indicative of the sensitive attribute. The causal model satisfies Proxy Discrimination, if there is no path from the sensitive attribute to the outcome via a proxy feature. The left causal model of Figure 2.2 displays no proxy discrimination, as there is no path from vertex A to Y via P. For this definition, it is necessary that the causal model is correct to not falsely assure or alarm.

2.1.3.3. Counterfactual Fairness. A well-known causal fairness definition is Counterfactual Fairness [29]. Counterfactual fairness requires that if we generate a counterfactual, using a fully specified causal model, in which we change a sensitive attribute for an individual, x, the distribution of the prediction for that individual does not change. It is thus a causal individual fairness definition. Formally, the following equation must hold in Counterfactual Fairness:

$$p(\hat{Y}_{A\leftarrow 1} = y|X = x, A = 1) = p(\hat{Y}_{A\leftarrow 0} = y|X = x, A = 1) \;\; y \in \{0, 1\}, \tag{2.12}$$

where $\hat{Y}_{A\leftarrow a}$ is the prediction for an individual in the circumstance where $A = a$.

For example, if we want to predict the success of a single student at a university and observe and use their final grades, but these are influenced by their race, the model might be counterfactually unfair. If we change the race of an individual, the final grades change and the distribution of the prediction might also change. An advantage of this approach is that it is less naive than other instance perturbation methods such as Causal Discrimination, because we take into account the change of the other variables. Creating a fully specified causal model, however, is time-consuming and complex, especially for a large number of features.

2.1.3.4. Individual Direct Discrimination. Individual Direct Discrimination compares the outcome rates of two groups that are similar to an individual [30]. It is also a causal individual fairness definition. The first group is the one in which the sensitive attribute is the same, and the other when it is different. A causal model is used to determine the similar individuals from both groups. The similarity is based on what degree features are causes of the outcome; features that differ greatly but have the same causal effect on the outcome are considered similar. If the difference in outcome rates of the two groups exceeds some threshold, the model is deemed causally unfair, according to Individual Direct Discrimination. An advantage of this metric is that it evades the use of counterfactuals.

2.1.3.5. Total Effect. Total Effect compares the acceptance rates across groups, by changing the sensitive attribute and measuring the difference in outcome along all causal paths connecting the two [31]. If the difference in acceptance rates exceeds some threshold, the model is deemed causally unfair, according to Total Effect. This calculation also uses counterfactuals, like counterfactual fairness, except it is now applied on a group level. Total Effect, also known as Average Causal Effect and Average Treatment Effect, is the fairness definition that coincides the most with SP. Because Total effect uses a causal model, Total Effect measures the change in acceptance rates for the entire population, whereas SP measures it only for a sample of the population, i.e. those in the dataset.

2.1.3.6. Equality of Effort. Equality of Effort is similar to the previously discussed Burden fairness definition. It also focuses on how much an individual has to change in order to change their outcome. Like Burden, a model satisfies Equality of Effort if individuals in both groups have to change a certain (indirect) causal influence equally as much to change the outcome. The causal model is used in a similar fashion as Individual Direct Discrimination; to represent each group. Then, the minimal change to a certain causal influence such that it changes the outcome is calculated and averaged over the group. If the differences in minimal change exceed some threshold, the model is deemed causally unfair, according to Equality of Effort. While an intuitive definition, only assessing the change in one causal influence is often insufficient. Equality of Effort concludes the discussion of fairness metrics; the next section highlights another pillar of responsible AI: interpretability.

2.2. Interpretable Models

This section 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. Figure 2.3 gives an example of a DT. 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

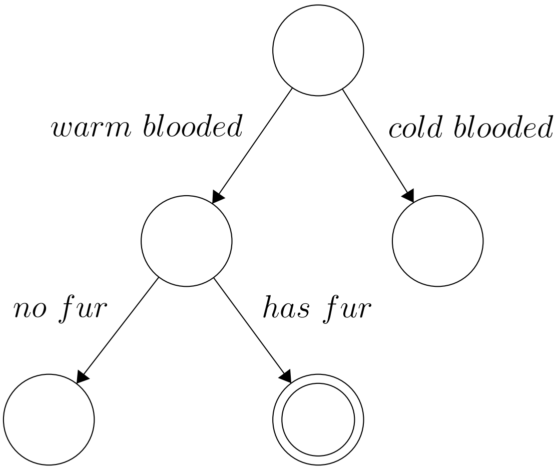

Figure 2.3. A simple DT with two branching nodes and three leaf nodes, that determines whether an animal is a mammal. A leaf node with an inner circle denotes a positive classification, i.e. the majority of the instances in that node are mammals.

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, e.g. $x^1 < 7$, 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 [32]. 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. The DT in Figure 2.3 has a height of 2. Very shallow Decision Trees are sometimes also called decision stumps [33]. 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 [34]. Major disadvantages of DTs are that they are incapable of capturing linear relations and that their construction is very sensitive to changes in the data. Still, their performance, especially the ensembles of DTs, are state-of-the-art for prediction tasks on tabular data [35].

2.3. Fair Interpretable Models

Where the previous sections were on improving the responsibility of ML models in one dimension, this section highlights models that focus on both fairness and interpretability. Several attempts have been made to improve the fairness of interpretable models[1]. In general, fairness-enhancing methods can fall under three categories: pre-processing, in-processing and post-processing [36]. Pre-processing methods focus on the data aspect of the ML pipeline, aiming to eliminate bias before the model training phase. In-processing methods work in tandem with the ML model during training, often adding a fairness objective to the optimization function of a model. Post-processing methods work on the outputs of a ML model, changing the classifications to ensure the

[1] Note that estimating the fairness of interpretable models is trivial if privacy is not a factor.

satisfaction of some fairness metric. Multiple types of methods may be combined. The focus of this section is on the interpretable models that were previously introduced. For each model, we name the type of processing method.

2.3.1. Fair Decision Trees

As mentioned, two well-performing approaches exist for the construction of DTs. Fairness-enhancing methods exist for both approaches which are highlighted in the same order as they appeared in the previous subsection.

2.3.1.1. Heuristic-Based Decision Trees. Kamiran & Calders introduced the most popular method for enhancing the fairness of DTs [37]. The method, now known as Discrimination Aware Decision Trees (DADT), has two phases. The first phase of DADT constructs a tree where the homogeneity of the sensitive attribute is incorporated into the splitting heuristic function. DADT can thus be considered an in-processing method. The second phase of DADT relabels the decision nodes in such a way that fairness is maximally improved and accuracy minimally worsened. The relabeling phase is phrased as a KNAPSACK problem [38], and is also solved greedily. Due to this second phase, DADT can also be considered a post-processing method. The heuristic that creates decision nodes that are both homogeneous w.r.t. group membership and class membership, pairs the best with the relabeling phase. The paper verifies this both experimentally and explains it by the fact that fewer leaf nodes have to be relabeled to prevent discrimination.

2.3.1.2. Optimal Decision Trees. For group fairness, there is one line of work regarding optimal fair DTs that was initiated by Aghei et al. [39]. This work was extended by Jo et al. who gave a formulation to construct optimally fair DTs based on Mixed Integer Optimization [40]. The method is obviously an in-processing approach as it adds a fairness objective to the optimal DT problem. Aside from this formulation, the authors propose an approach that solves the problem for a certain DT height, and then optimally branches out from that tree. Both implementations, however, are very slow and practically, heights of more than three are out of reach. The framework can be used to optimize for SP, conditional SP, PrEq, EOpp and EOdd. For each definition, a slack parameter can be specified, like in the 80%-rule, such that an array of DTs can be generated. The paper shows an improvement in terms of fairness and accuracy over DADT.

Recent work hugely improves the speed of the former method, but can only improve statistical parity [41]. Linden et al. do note that the method "can easily be extended to support other notions of group fairness" [41, p.3]. The method is named DPFair. The speed improvements mainly stem from more smartly pruning the search space and writing the algorithm in a faster programming language. One major downside of these approaches is that they require the binarization of the features in the dataset. And even then, for large datasets with a large number of features, construction of optimally fair DTs can take more than an hour.

For individual fairness, there is one method that was proposed by Ranzato & Zanella [42]. The paper uses the equivalence between individual fairness and stability, a notion found in robust ML. The authors build upon their own work [43], to propose a method that constructs an individually fair set of DTs, using a genetic algorithm. The authors name the method Fairness Aware Tree Training (FATT). The paper experimentally verifies that the method is far more individually fair than other fairness unaware DT methods while giving in little on accuracy. The authors also propose to let the parameters of the trees generated by FATT be the hyperparameters for fairness unaware Decision Trees, e.g. setting the maximum height of the 'naive' DTs to the height of the trees generated by FATT. These types of DTs give in far less in terms of accuracy and, using the 'fair' parameters, still provide a substantial individual fairness improvement. FATT is an interesting approach but as it is the only work considering individual fairness, it cannot yet be compared to other methods and definitions of individual fairness.

2.4. 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 relates to the degree of control one has over others accessing personal data about them. This is also known as informational privacy. The less personal data others access about an individual, the more privacy the individual has. This section discusses several techniques to increase informational privacy.

2.4.1. Differential Privacy (DP)

Differential Privacy (DP) [44] 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' [45]. 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 [45].

2.4.1.1. Definition. The promise of DP can be mathematically guaranteed up to a probability 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[3], to which DP adds noise. Because DP is based on membership inference, the formal definition compares two neighboring datasets, D and D', in which

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

only one instance differs. For these datasets, ε, δ-DP formally is:

$$p(\mathcal{A}(q(D)) \subseteq range(\mathcal{A})) \leq \exp(\varepsilon) \cdot p(\mathcal{A}(q(D')) + \delta \subseteq range(\mathcal{A})), \tag{2.13}$$

where $\mathcal{A}$ is a randomized mechanism around a query $q(\cdot)$, $range(\mathcal{A})$ 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.

2.4.1.2. 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 [44]. Formally:

$$\Delta q = \max_{D,D'} ||q(D) - q(D')||_1, \tag{2.14}$$

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

2.4.1.3. Laplace Mechanism. Several techniques exist to randomize query answers, of which the most common one is the Laplacian mechanism [44], for queries requesting real numbers[4]. 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:

$$\mathcal{A}(D, q(\cdot), \varepsilon) = q(D) + Lap(\frac{\Delta q}{\varepsilon}), \tag{2.15}$$

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

2.4.1.4. Randomized Response. For answers with a binary answer, Randomized Response may be used [46]. This procedure is ln(3)-differentially private [45]. The pro-

[4] An example of such a query might be: ‘What is the average age of females in the dataset?’.

cedure is as follows:

(i) Flip a coin.
(ii) If it is heads, respond truthfully.
(iii) Else, flip another coin.
(iv) If it is heads, respond 0, else 1.

The responses 0 and 1 are placeholders for actual answers and should be mapped to the query appropriately. The procedure originates in social sciences where respondents might be not so inclined to answer truthfully with regard to criminal activities. This procedure ensures that the respondents cannot be charged for their answers.

2.4.1.5. Exponential Mechanism. A different noise schema is the Exponential mechanism [47], used for categorical, utility-related queries[5]. 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:

$$p(\mathcal{A}(D, u, \mathcal{R}, \varepsilon) = r) \propto \exp(\frac{\varepsilon u_D(r)}{2\Delta u}). \tag{2.16}$$

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

2.4.1.6. 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 $\sigma \geq \sqrt{2\ln(\frac{1.25}{\delta})\frac{\Delta_2}{\varepsilon}}$ [45]. The term Δ_2 is the global ℓ_2-sensitivity; instead of using the ℓ_1-norm in Equation 2.14, Δ_2 uses the ℓ_2-norm. The Gaussian mechanism can be deemed a more 'natural' type of noise, as it adds noise

[5] 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 section.

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.

2.5. Privacy Aware Interpretable Models

This section discusses the construction of DTs in a DP-aware manner. At the intersection of privacy and interpretability, several works exist that prevent data leakage via interpretable models or via developers of interpretable models.

2.5.1. Privacy Aware Decision Trees

There are three main works on the construction of DTs with DP guarantees, the rest of the field is more concerned with creating decision forests which have better performance. This holds in general, not only in a privacy-constrained setting. This subsection discusses the three works in chronological order. The setting that this body of work assumes is that a DT developer has limited access to the data via a curator that they can send queries to. The answers to these queries should be perturbed via DP-mechanisms.

Blum et al. first introduced DTs with DP [48]. It was more of a proof-of-concept; the authors rewrote the information gain splitting criterion to make it differentially private. Querying the necessary quantities for each node and adding Laplacian noise to the answers ensures DP. For the leaf nodes, the class counts are queried, as is the case for all other approaches mentioned. The method, however, requires a large privacy budget which in turn makes the answers to the queries noisy. It also can not handle continuous features but does allow for trees with a height equal to the total number of features.

The improvement on this method came from offloading the bulk of the computation to the data curator [49]. The method that is proposed in [49] simply queries for the quantities in each node and the best attribute to split on. The latter is used to construct the tree and the former to cleverly determine the termination of the tree construction. The improvement also stems from the fact that the method in [48] used overlapping queries which hurts the privacy budget. This problem is not present in [49], where the queries for nodes for each height are non-overlapping. Friedman & Schuster used the exponential mechanism, which relies on the sensitivity of the utility function, in this case, the splitting criterion. It is experimentally verified that when the criterion is the error rate, the accuracy is the highest. This method can handle continuous variables in theory but the inclusion of them in the training set severely hurts the predictive performance. Moreover, the height of the DT can at maximum be five. The method still improved performance significantly, however, due to the more clever queries and noise addition.

The final improvement in this line of work was found by Mohammed et al.. They disregard the first query concerned with the node quantities and instead focused solely on the queries for the best splitting criterion, allowing more privacy budget to be spent on each query [50]. This approach comes at the cost of a more robust termination criterion, that has less flexibility than the one in [49]. Through experimental evaluation, a very robust termination criterion is determined, which is: stop at a height of four. Using this termination procedure, the performance of the method is experimentally shown to outperform the previous method. However, this method excludes the possibility of using continuous features, but this is not a large downside as it is discouraged for the approach in [49] that this method builds upon.

2.6. Privacy Aware Fair Models

This section discusses methods that simultaneously have an eye for fairness and DP. These objectives may be competing not only in the sense that fairness sometimes requires sensitive attributes but also in the sense that fairness-enhancing models might leak more information from certain groups [51]. Note that this section discusses works from different fields and with different settings, unlike previous sections where goals were more uniform.

2.6.1. Querying Fairness

Hamman et al. explore the idea of querying the actual group fairness metrics in a recent paper [52]. 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. It is established in [52] 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 [53], 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 Attribute Conceal, 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 Attribute Conceal, an adversary can predict sensitive attributes merely as well as a random estimator.

2.6.2. Post-processing Method

Jagielski et al., in [54], transform a fairness-enhancing post-processing [17] and in-processing approach [55]. They also consider the setting where only the protected attribute remains to be private. They adapt both fairness enhancing algorithms, optimizing for EOdd, to also adhere to DP. The former is based on the fractions in the data adhering to different combinations of $\hat{Y} = \hat{y}, A = a$ and $Y = y$. These are fed into a linear program to determine the optimal decision thresholds for each individual group, a. The privacy extension comes from perturbing these fractions using the Laplacian distribution. Albeit intuitive, the method performs quite badly and requires access to the sensitive attributes at test time. Therefore, the latter approach is introduced, which is based on a zero-sum game between a hypothesis selector that finds the best-performing model and a regulator that points out EOdd violations to them based on gradient descent. The equilibrium that is arrived at in the game, is the best trade-off between EOdd and accuracy. The hypothesis selector is considered to adhere to DP if sensitive attributes are absent from its input. The fairness regulator is made differentially private by adding Laplacian noise to the gradients of the gradient descent solver. The results of this approach are only satisfactory for large privacy budgets.

2.7. Fairness Without Sensitive Data

This section highlights a number of methods that aim to enhance fairness without accessing sensitive data. Research towards these methods is a reasonably new subarea within fairness research. Research has addressed a range of availabilities, from one sensitive attribute missing to no information on the sensitive groups at all. These methods are often developed using datasets with sensitive features because evaluation would otherwise be impossible.

2.7.1. Proxy Fairness

Gupta et al. were one of the first to measure fairness when a sensitive feature is missing [19]. They propose to measure and improve the fairness of another sensitive group instead, e.g. mitigating gender bias as a proxy for racial bias. The bias mitigation is most effective if the proxy group and the true sensitive group are semantically related, but this need not be the case. The authors then experimentally show that using the post-processing method from [17] on a proxy group, increases fairness in terms of Accurate Coverage and EOpp for the proxied group. The effectiveness, however, depends on the fairness metric. The experiments show that EOpp is more difficult to improve using a proxy group. An advantage of Proxy Fairness is that it can be used with any class of models, and thus also interpretable models.

2.7.2. Fairness Using Distributionally Robust Optimization (DRO)

Hashimoto et al. were one of the very first methods to investigate bias in a setting where sensitive data is entirely unavailable [20]. The authors even presume that the number of sensitive groups is unknown. The proposed method can be applied to the class of stochastic gradient descent models. The method is applied in a setting where users from a group give queries to a model, e.g. a speech recognition model, and are assumed to use the model less if the performance is worse. This gives rise to the Rawlsian Min-Max fairness definition, as this aims to have a maximal minimum performance for each group. By sampling around the data generation process, and minimizing the maximum loss for all possible groups within that sample, Rawlsian Min-Max fairness is ensured. By careful sampling, each group is represented. Distributionally Robust Optimization (DRO), upweights the samples in the minority groups. The method performs well, i.e. the maximum loss is low, if the instances with a low loss are all from minority groups. The paper empirically verifies the results on an auto-complete service, observing retention when DRO is used or not. Applying DRO causes higher retention rates for both groups.

2.7.3. Adversarially Reweighted Learning (ARL)

Lahoti et al. build upon DRO and also investigate Rawlsian Min-Max fairness [21]. The authors apply Rawlsian Min-Max fairness to computationally identifiable groups, instead of all possible groups. An adversarial neural network identifies these groups, by identifying regions with high expected error rates. Another model is responsible for classification. The paper introduces Adversarially Reweighted Learning (ARL), upweighting samples in identified regions, such that the classification model performs better for these regions, thus optimizing Rawlsian Min-Max fairness. The paper details numerous experiments including comparisons with DRO, Inverse Probability Weighting [56] and Min-Diff [57], as well as an analysis of the learned weights, the identified groups and the robustness of ARL to label and representation bias. It is concluded that ARL achieves comparable or better performance than the aforementioned methods. The learned weights show to be higher for minority groups and become lower when the size of the minority group increases. As with DRO, ARL is shown to be prone to label bias, because then some misclassified instances are not from minority groups. The authors conclude that representation bias is not an issue for both methods. Finally, the authors observe that ARL performs worse when the groups are less identifiable. ARL is thus preferred over DRO when one suspects that the groups can be easily identified, perhaps if some features are historically proxy features for sensitive attributes.

2.7.4. Fair Related Features (FairRF)

Zhao et al. also consider a loss-based model and the unavailability of any sensitive information [58]. However, they assume that some features are heavily related to sensitive features and that these are known or estimated. By regularizing the covariance between the predictions of the model and the related features, fairness is enhanced. The method, named Fair Related Features (FairRF), learns a specific penalty term for each related feature, based on its covariance with the predictions. This leads to a constrained optimization problem that is solved by alternatingly updating the model parameters and the penalty terms. An advantage of FairRF is that it incorporates a regularization importance hyper-parameter, enabling developers to find the right balance between fairness and accuracy. Through empirical verification, it is shown that the selected set of related features may be noisy. FairRF is also experimentally compared with other methods, including ARL. FairRF gives up a bit of accuracy for fairer results in terms of SP and EOpp.

2.7.5. Fairness Using Knowledge Distillation

Chai et al. introduce the latest advancement in fairness without sensitive data in [59]. It utilizes knowledge distillation, the concept that an overfit complex teacher model can provide the labels for a simpler student model. The predictions of the teacher model and the actual labels are combined to form a smooth continuous labelling. The hypothesis in [59] is that discrimination is most easily prevented close to the decision boundary[1]. The student model is trained on the smooth labelling, which causes a focus on correctly classified instances for which there is a large difference between the actual label and the smoothed label. This improves the performance in regions with more instances from minority groups. Via empirical evaluation, the method is shown to outperform DRO, ARL and FairRF, giving up less accuracy while achieving more fair predictions in terms of SP and EOdd. When the smoothing parameter is increased, i.e. the amount of influence the teacher has on the new labels, EOdd is shown to improve.

2.7.6. Proxy Models

Several approaches exist that replace the missing sensitive data with a prediction of these variables. The models predicting sensitive variables are often called proxy models. Proxy models get their training labels from other (publicly available) datasets. A notorious example is Bayesian Improved Surname Geocoding, which used the Naive Bayes approach to predict ethnicity based on surname and address [61]. In general, proxy models underestimate fairness [62]. Nevertheless, a bank that applied Bayesian Improved Surname Geocoding received a fine of 98 million US dollars [63]. The main disadvantage of proxy models is that they deteriorate in performance the fewer proxy features are present in the data. Due to their dual-use[2], also noted by Awasthi et al. they are not reviewed in this work [64]. The approach, however, is quite popular and a decently sized body of research aims towards it.

2.8. Related Work & Background Conclusion

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 [52], 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

[1] This idea is also explored in the post-processing method named Reject Option Classification [60].

[2] The prediction of sensitive attributes also enables unjust discrimination.

are modular and can be pruned, i.e. rules can be removed. DTs are the state-of-the-art for tabular data [35] and sensitive tasks are often prediction tasks for tabular data[1]. A method that can identify unfairness in a privacy-aware manner for DTs would be interpretable, fair and differentially private, respecting some of the pillars of responsible AI. PAFER aims to fill this gap, querying the individual rules in a DT. The next chapter will introduce the method.

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

3. Proposed Method

This chapter describes Privacy-Aware Fairness Estimation of Rules (PAFER), the core of this work. PAFER is a novel method to estimate the fairness of DTs. The following sections dissect the proposed method, starting with a section on the assumptions and specific scenarios for which the method is built (Section 3.1). The successive section provides a detailed description of the procedure (Section 3.2), followed by the pseudocode (Subsection 3.2.5) and a section that details some theoretical properties (Subsection 3.2.6).

3.1. Scenario

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

3.1.1. 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 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 now wants to assess the fairness of their model using SP. We lastly assume that a third party exists that does know these sensitive attributes on an instance level, 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.

3.1.2. 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 for a sensitive task [65, 66]. 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 [14] and loan application evaluations [1]. Moreover, multiclass decision-making problems can be rewritten as binary decision problems, as shown in Corollary 3.1. Thirdly, it is often the case that model developers do not have access to sensitive attributes. Simply because of regulations [7], 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 [8], due to a compliance check by an auditing body or due to newly created awareness of machine bias [14]. 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 Chapter 2.

What is uncommon, however, is a third party that has all the sensitive attributes of the individuals in the dataset, and is also willing to share them. As data is the new oil [67], sharing data becomes more and more difficult. Since, however, fair and interpretable sensitive attribute agnostic classifiers are currently lacking (Chapter 2), this assumption becomes necessary. This work can thus be seen as an exploration of this cooperation between developer and data holder, to determine the privacy risks and utility of such an exchange.

3.2. Privacy-Aware Fairness Estimation of Rules: PAFER

We propose Privacy-Aware Fairness Estimation of Rules (PAFER), a method based on DP [45], 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 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 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[1]. It can also be seen as querying the root node of a DT.

3.2.1. 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 [68].

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.

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

3.2.2. 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:

$$\text{SP} = \min\left(\frac{p(\hat{Y} = 1|A = a)}{p(\hat{Y} = 1|A = b)}\right), \; a, b \in \{0, 1, 2, \ldots, k-1\}, a \neq b. \tag{3.1}$$

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.

3.2.3. 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 [45]. 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 [45]. The Gaussian mechanism is also used to perform a histogram query and has a sensitivity of 2, as it uses the Δ_2-sensitivity.

3.2.4. 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[1]. A policy for handling these invalid query answers must be chosen. In practice, these are mappings from invalid

[1] 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.

values to valid values. We provide several options in this subsection.

Table 3.1. The proposed policy options for each type of invalid query answer. A policy consists of a mapping chosen from the first column and a mapping chosen from the second.

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

Table 3.1 shows the available options for handling invalid query answers. 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.

3.2.5. PAFER Pseudocode

Algorithm 1 shows the pseudocode for PAFER.

3.2.6. 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 3.1. 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.

Corollary 3.1. *Any DT that classifies for a binary decision problem that uses non-binary splits, can be converted to a DT that solely uses binary splits.*

Algorithm 1 PAFER($\mathcal{A}, D, \varepsilon, DT, policy, \mathcal{K}$), outputs estimated SP

```
{$\mathcal{A}$ is a DP mechanism that introduces noise}
{$D$ is a database with $N$ instances}
{$\varepsilon$ is the privacy budget}
{$DT$ is a binary Decision Tree composed of rules}
{$policy$ is a mapping that transforms invalid query answers to valid query answers}
{$\mathcal{K}$ is the number of sensitive groups for the sensitive attribute}
$accept\_rates \leftarrow zeros(1, \mathcal{K})$ {$accept\_rates$ is a row vector of dimension $\mathcal{K}$, initialized at 0}
$total \leftarrow \mathcal{A}(True, D, \frac{1}{2}\varepsilon)$
for $q \in DT$ do
    if $q$ is favorable then
        $accept\_rates \mathrel{+}= \frac{policy(\mathcal{A}(q, D, \frac{1}{2}\varepsilon))}{total}$
    end if
end for
$\widehat{SP} = \frac{\min(accept\_rates)}{\max(accept\_rates)}$
return $\widehat{SP}$
```

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 3.1. Since we have proven this property for an arbitrary number of splits in a node, the property holds for any k-ary DT. □

Theorem 3.1. *The number of queries required to estimate SP for PAFER 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 3.1, the DT can be converted to a binary tree, since it classifies for a binary classification problem. Further assume that this (converted) binary DT, has height h. To estimate SP, for each sensitive attribute the total size is required, $|D_{A=a}|$, as well as

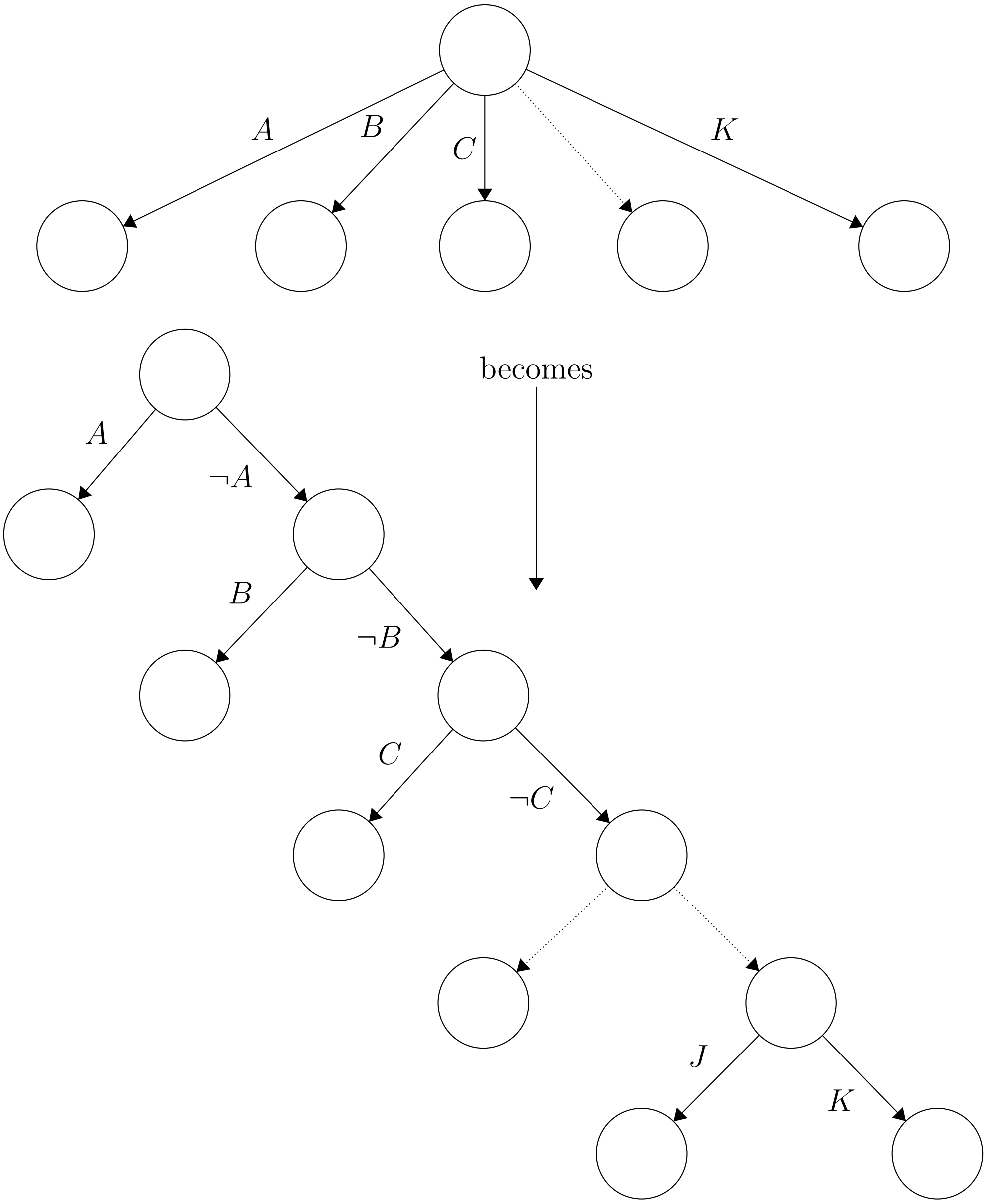

Figure 3.1. A schematic display of the process by which a binary tree that has non-binary splits can be converted into a binary tree for a binary decision process. The dotted lines $\cdot^{\cdot^{\cdot}}$, denote that the pattern of the DT can be repeated an arbitrary number of times.

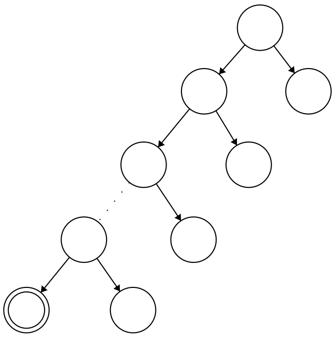

Figure 3.2. The smallest number of favorable decision rules in a decision tree for a binary classification problem. The leaf node with an inner circle denotes a leaf node in which the majority of the individuals are classified favorably in the training set. The dotted line, $\iddots$, denotes that the pattern can go on indefinitely.

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 3.2. 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 3.3. 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 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 $2^{h-1}+1$. □

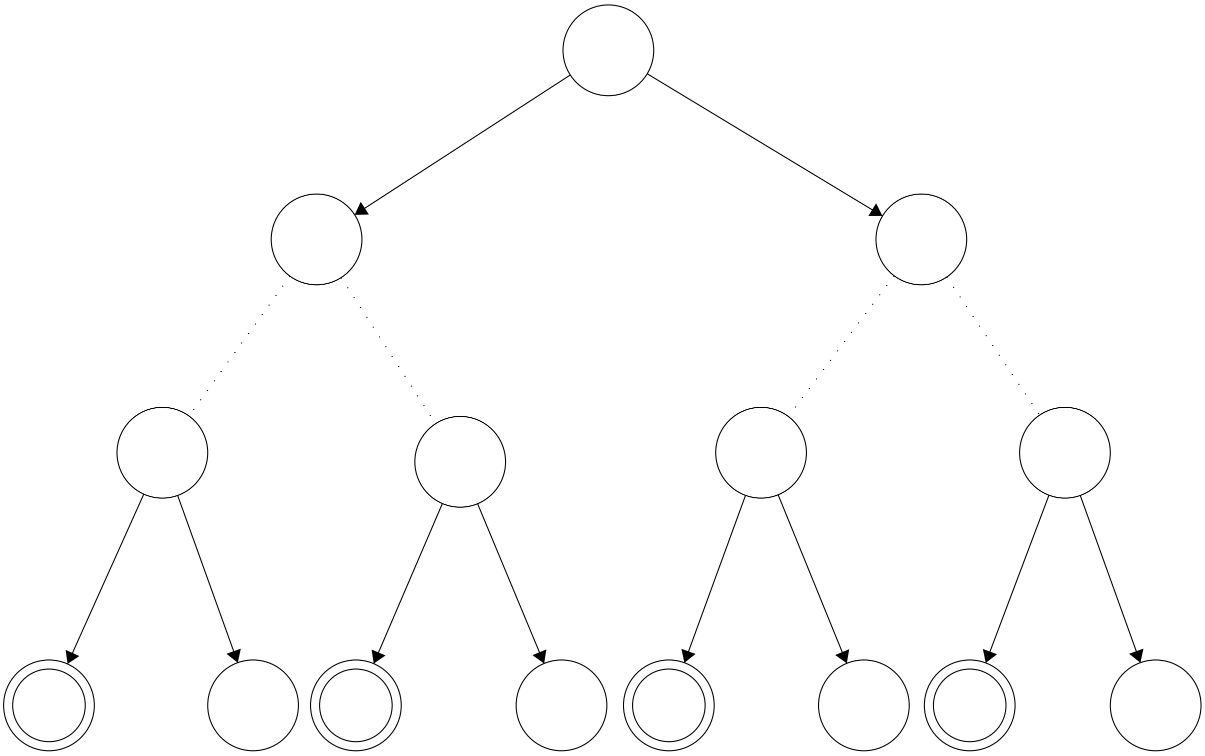

Figure 3.3. The largest number of favorable decision rules in a decision tree for a binary classification problem. The leaf nodes with an inner circle denote a leaf node in which the majority of the individuals are classified favorably in the training set. The dotted lines, $\cdot^{\cdot^{\cdot}}$, denote that the pattern can go on indefinitely.

4. Evaluation

This chapter evaluates the proposed method in the previous chapter, PAFER. Firstly, Section 4.1 describes the experimental setup, detailing the used datasets in Subsection 4.1.1, and the two experiments in Subsection 4.1.2 through Subsubsection 4.1.3.1. Secondly, Section 4.2 displays and discusses the results of the experiments.

4.1. 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.

4.1.1. 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 [69], COMPAS [14] and German [70]. They are all well known in the domain of fairness for ML, and can be considered benchmark datasets. The datasets are publicly available and pseudonymized; every privacy concern is thus merely for the sake of argument.

4.1.1.1. Properties. These three datasets were chosen because they possess some important properties. Importantly, they vary in size and are very popular in fairness for ML research. Each dataset models a binary classification problem, enabling the calculation of various fairness metrics. Table 4.1 shows some other important characteristics of each dataset.

Table 4.1. Properties of the three chosen publicly available datasets.

Dataset	# Rows	# Features	Sens. attrib.	Task
Adult	48842	14	race, sex, age, country of origin	Income > $50000
COMPAS	7214	53	race, sex, age	Recidivism after 2 years
German	1000	24	race, sex, age, country of origin	Loan default

4.1.1.2. Pre-processing. This subsubsection 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 kept separate. Every sensitive attribute except age was binarized, distinguishing between privileged and unprivileged groups. The privileged individuals were White men living in their original country of birth, and the unprivileged individuals were those who were not male, not White and not living in their original country of birth. 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 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 [14]. 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[1], this number reduced to eleven, of which three are sensitive attributes. The other pre-processing step in the article is to remove cases in

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

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 dataset does not come with a preset train set and test set. The dataset was manually split according to the same proportions as the Adult dataset (roughly $\frac{1}{3}$). 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 dataset does not come with a preset train set and test set. The dataset was, therefore, manually split according to the same proportions as the Adult dataset (roughly $\frac{1}{3}$). The final number of rows is 667 for the train set and 333 for the test set, totalling 1000 rows.

4.1.2. 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 $\{2, 3, 4\} \times \{3, 4, 5, 6, 7, 8, 9, 10, 11, 12\} \times \{\text{sqrt}, \text{all}, \log_2\}$, constituting tuples of (height, # leaf nodes, # selected features). Section A.1, in the Appendix, shows the final trained pruned DTs. The out-of-sample SP of each DT is also provided in Table 4.2. The experiment 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 4.3 shows some of these preliminary results. The performance of each mechanism was measured using the Average Absolute Statistical Parity Error (AASPE), defined as follows:

$$\text{AASPE} = \sum_{i}^{\#\,\text{runs}} \frac{1}{\#\,\text{runs}} |SP_i - \widehat{SP_i}|, \tag{4.1}$$

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 4.3 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 Subsection 3.2.4, 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.

Table 4.2. 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 4.3. Preliminary results for Experiment 1 with larger privacy budgets. Results were averaged over 25 runs, except for $\varepsilon = \frac{1}{2}$. The Gaussian mechanism was tested with $\delta = \frac{1}{1000}$. The performance was measured using the AASPE.

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

4.1.3. Experiment 2: Comparison of different DTs for PAFER

Experiment 2 was constructed in such a way that it answered **RQ1**; 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 4.4, 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. 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[1]. 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 AASPE, as shown in Equation 4.1. 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

[1] Based on the "Magic Number 7", as humans can generally hold seven ± two pieces of information in memory, and thus, also, seven rule clauses in memory [71].

Table 4.4. Preliminary results for Experiment 2. The performance was measured using AASPE. 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

value in the interval $[0, 1)$. A one-sided t-test determined whether PAFER significantly outperformed the random baseline.

4.1.3.1. 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:

$$\mathrm{UAR} - \mathrm{AASPE} = \sum_{c \in C} \frac{1}{|C|} \times \frac{\#\,\mathrm{true}\,c}{\#c} - \sum_{i}^{\#\,\mathrm{runs}} \frac{1}{\#\,\mathrm{runs}} |SP_i - \widehat{SP_i}| \tag{4.2}$$

Ideally, AASPE is minimized and UAR is maximized, thus maximizing the metric shown in Equation 4.2. Besides the metric, the experimental setup was identical to Experiment 2. Therefore, the same DTs were used for this experiment, only the metrics differed.

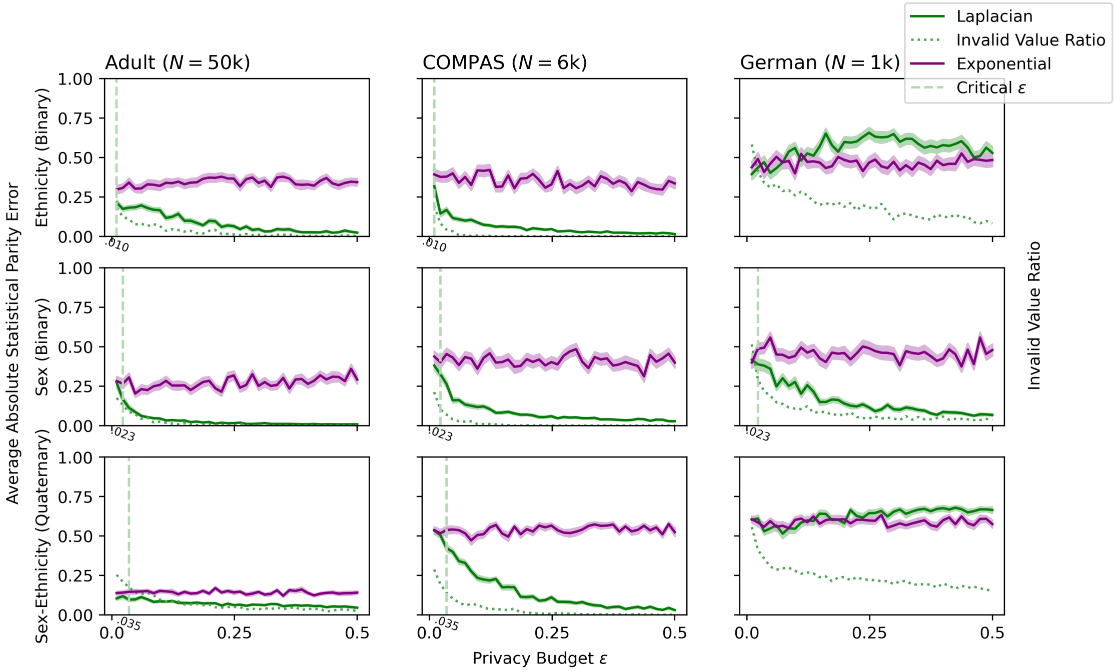

Figure 4.1. A comparison of the Laplacian and Exponential DP mechanism for different privacy budgets ε. When indicated, from the critical ε value to $\varepsilon = \frac{1}{2}$, the Laplacian mechanism performs significantly better ($p < .05$) than the Exponential mechanism. The uncertainty is pictured in a lighter color around the average.

4.2. 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.

4.2.1. Results for Experiment 1

Figure 4.1 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 an 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 Subsection 3.2.4. 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 4.1 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.

4.2.2. Results for Experiment 2

Table 4.5 through Table 4.10 show the results for Experiment 2. The tables clearly show that PAFER generally significantly outperforms the random baseline. For

Table 4.5. 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.

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

Table 4.6. 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 $\Diamond$ indicates that the random baseline performed significantly better than PAFER.

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

Table 4.7. 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.

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

small privacy budgets ($\varepsilon \leq \frac{1}{10}$) and small `minleaf` values (`minleaf` $= \frac{1}{1000}$), PAFER does not strictly perform better, for instance in Table 4.9. PAFER is even significantly outperformed by the random baseline in some cases, such as in Table 4.6 and Table 4.10, 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

Table 4.8. 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.

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

Table 4.9. 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.

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

Table 4.10. 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.

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

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 [72].

Another factor negatively impacting the performance of PAFER is the size of the dataset and the number of (un)privileged groups. The baseline significantly outperforms PAFER in Table 4.10 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.

The results for PAFER compared to a perfect estimator are very lop-sided; the perfect estimator significantly outperforms PAFER across all datasets, sensitive attribute analyses, `minleaf` values and privacy budgets with $p \lll .001$.

4.2.2.1. Results for Experiment 2.1. Figure 4.2 through Figure 4.7 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 4.6 and Figure 4.2. 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 4.3.

Nonetheless, a pattern emerges from Figure 4.2 through Figure 4.7 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 $\frac{7}{100}$.

This final experiment also replicates some of the results of Experiment 1 and Experiment 2. The middle plot in Figure 4.2 through Figure 4.7 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. Experiment 2.1 thus aptly acted as a replicating sanity check.

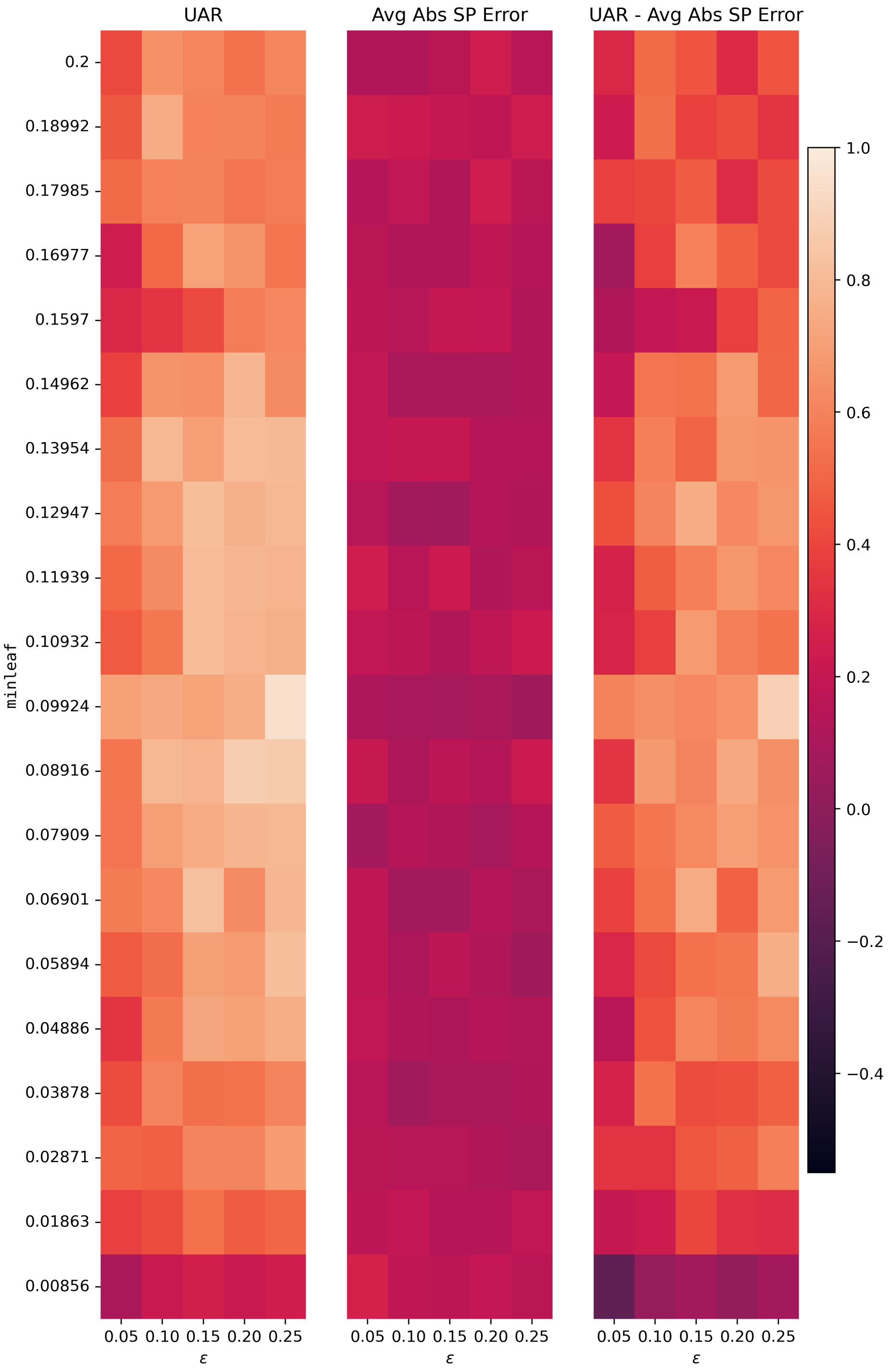

Figure 4.2. The hyperparameter space for the Adult dataset and the binary ethnicity attribute.

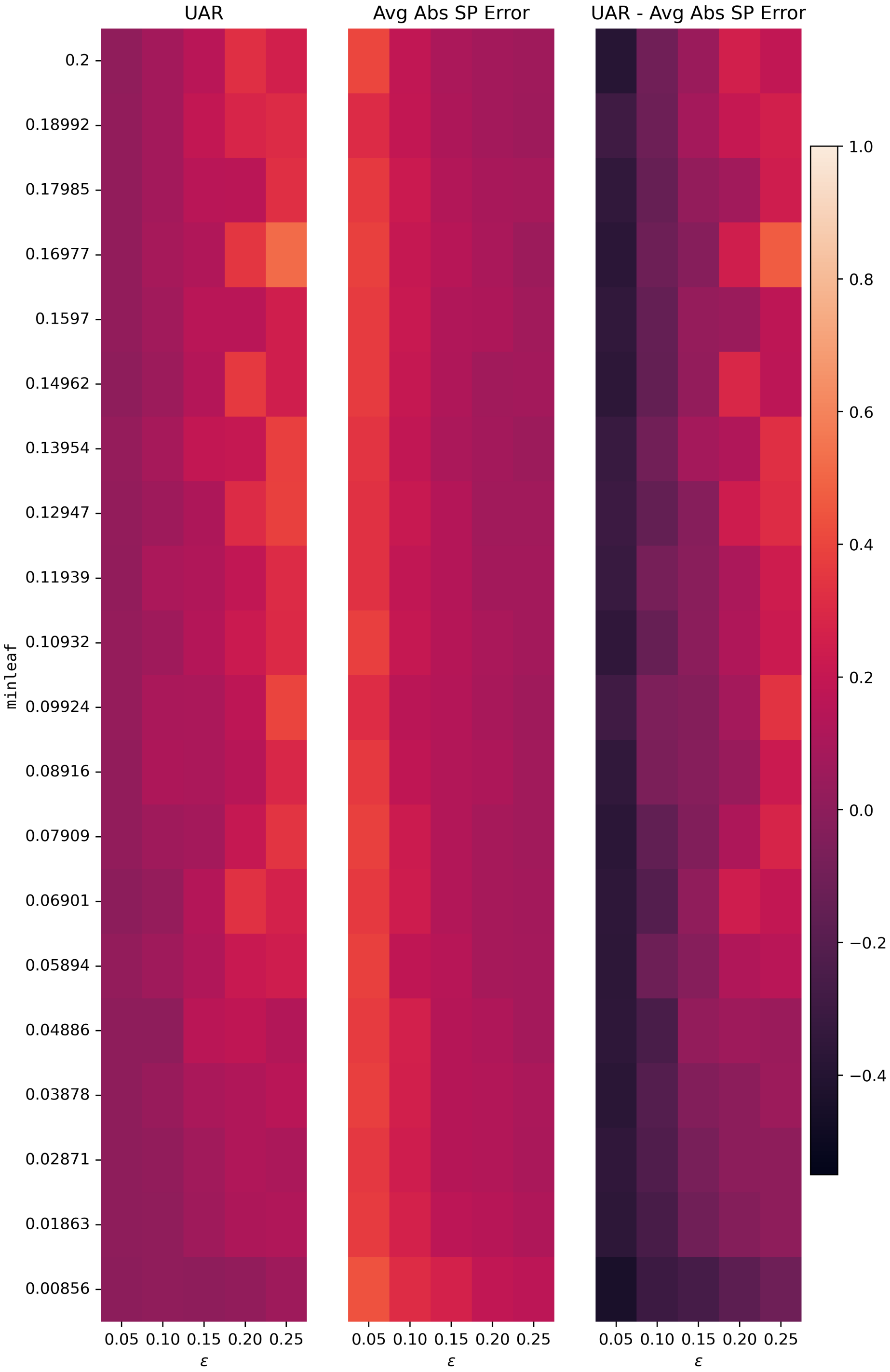

Figure 4.3. The hyperparameter space for the Adult dataset and the binary sex attribute.

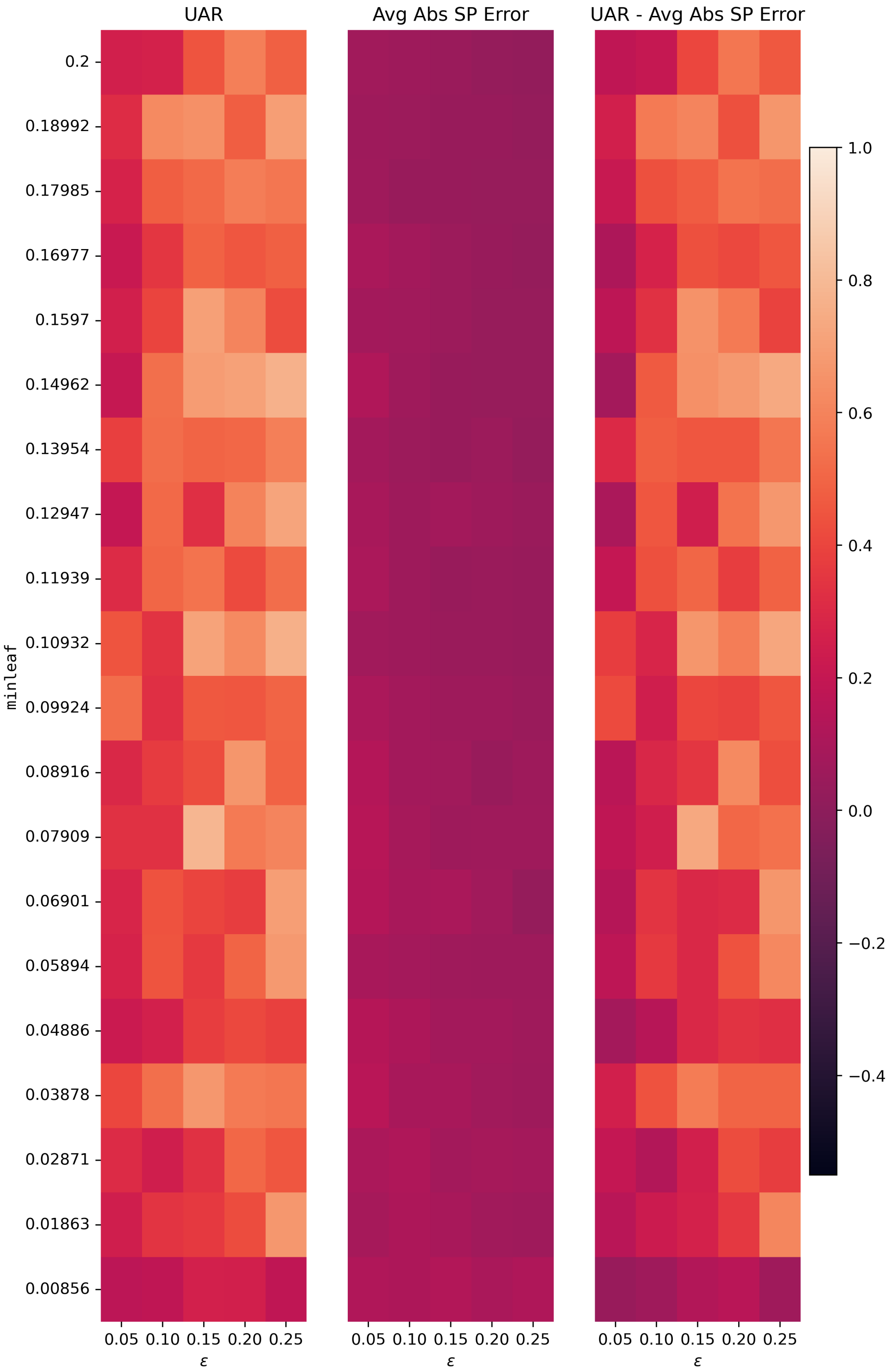

Figure 4.4. The hyperparameter space for the Adult dataset and the quaternary sex-ethnicity attribute.

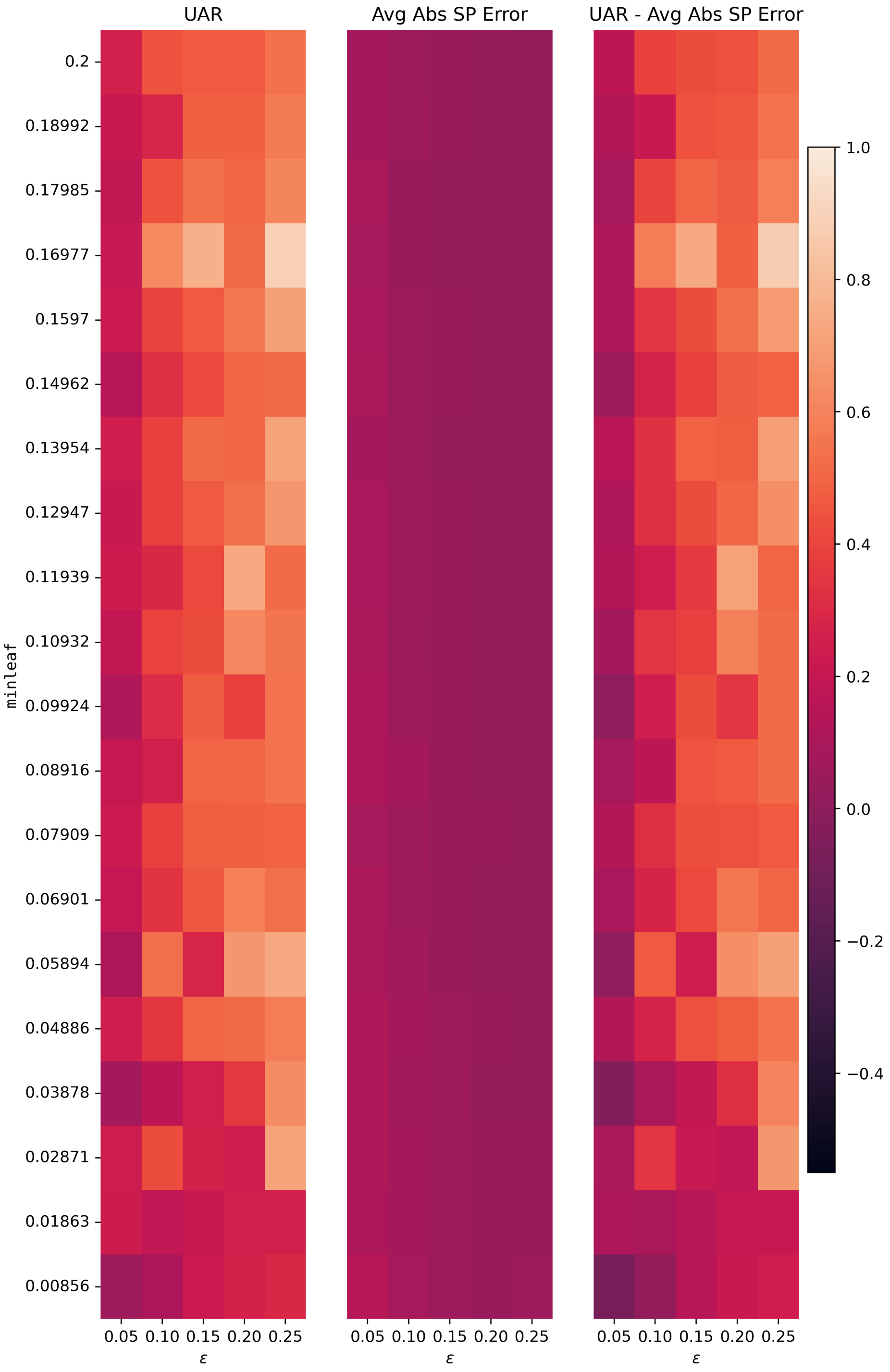

Figure 4.5. The hyperparameter space for the COMPAS dataset and the binary ethnicity attribute.

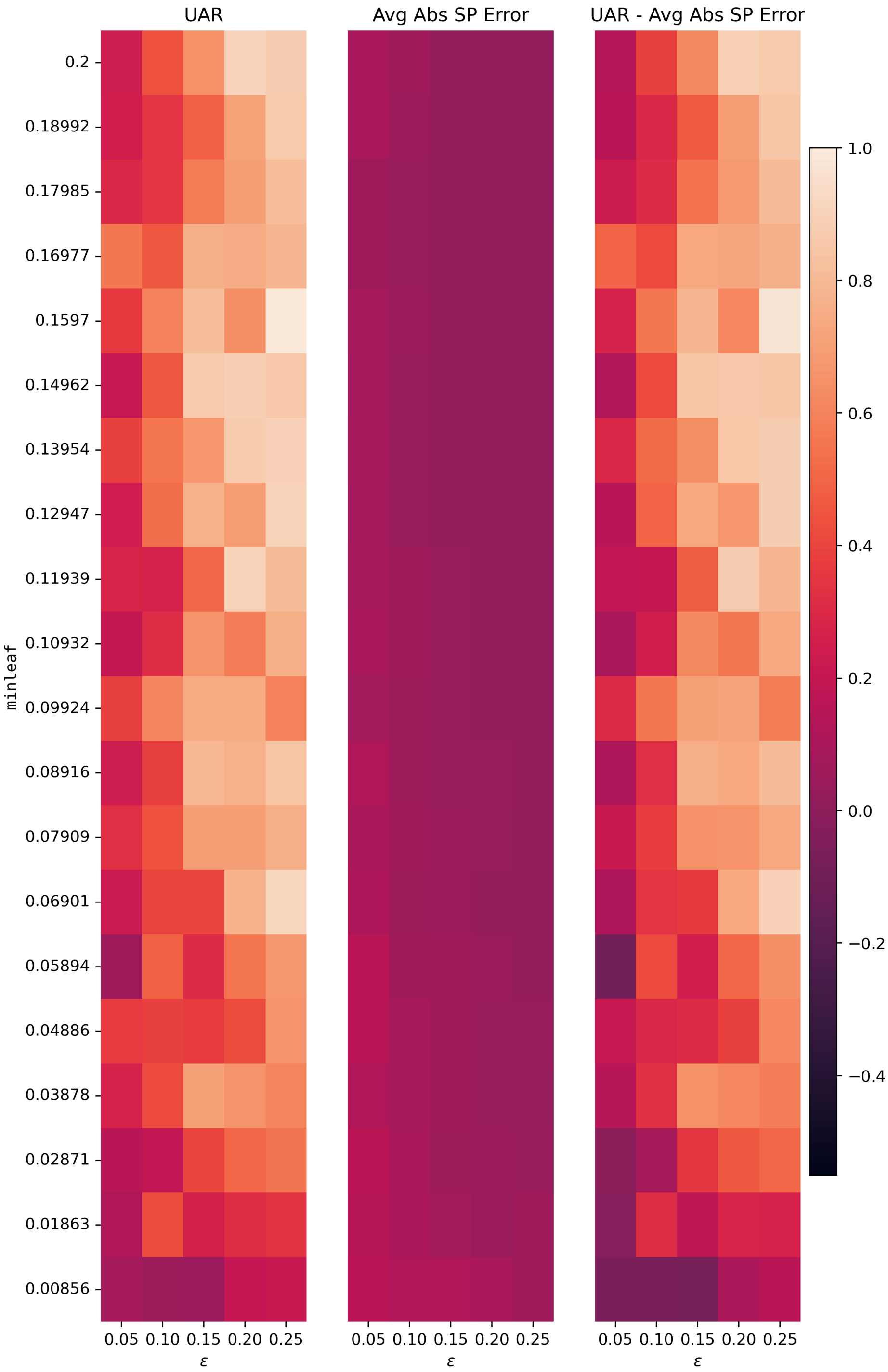

Figure 4.6. The hyperparameter space for the COMPAS dataset and the binary sex attribute.

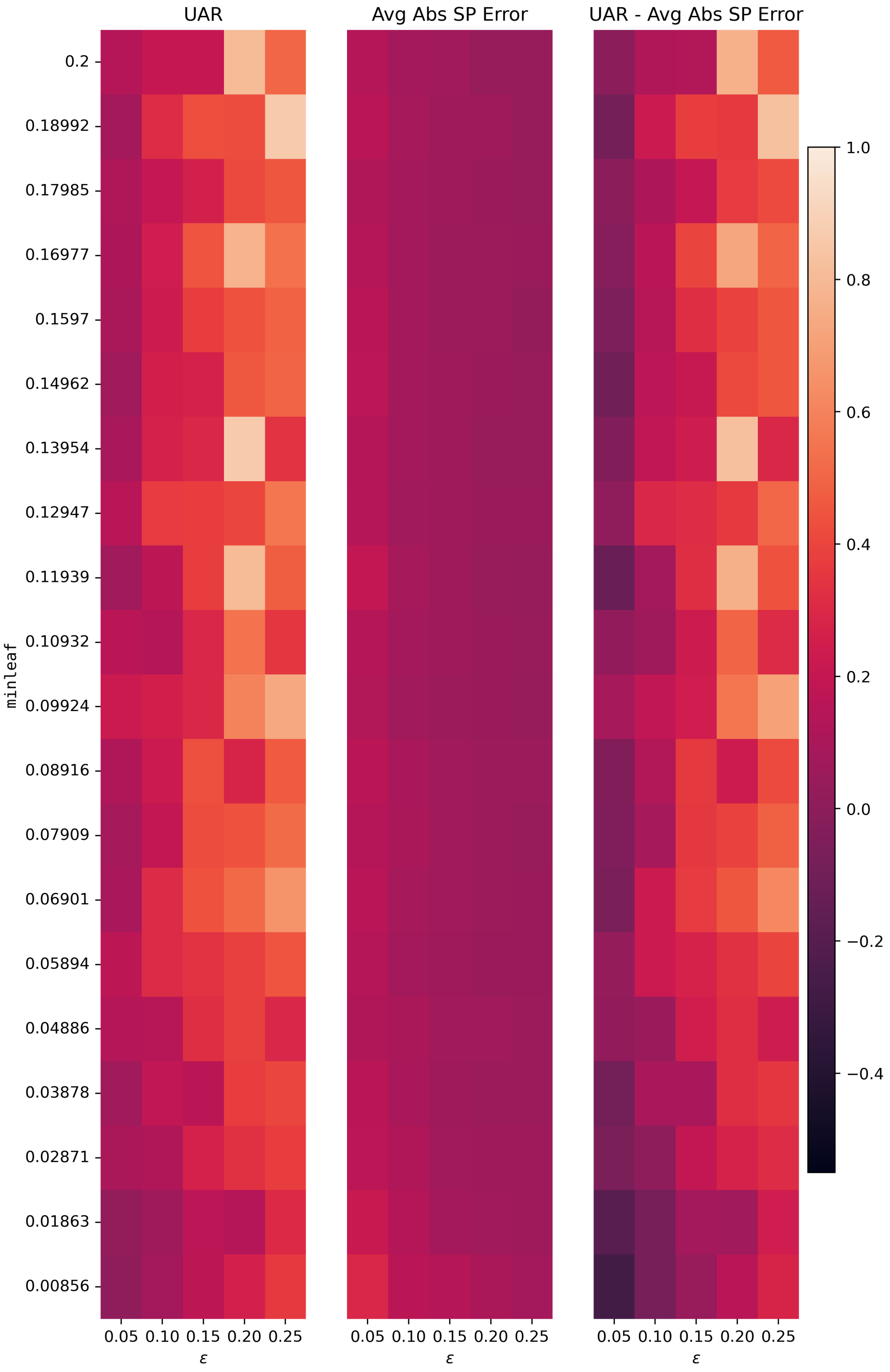

Figure 4.7. The hyperparameter space for the COMPAS dataset and the quaternary sex-ethnicity attribute.

5. Conclusion & Future Work

This Chapter concludes this work. It provides answers to the research questions in Section 5.1, summarizes the entire work in Section 5.2 and provides suggestions for future work in Section 5.3.

5.1. Answers to the Research Questions

This section will answer the research questions (RQs) and research subquestions (RSQs), as posed in Section 1.2.

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 4.1. 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 4.3.

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. The Gaussian mechanism proved also to be of no match for the Laplacian mechanism, even at large privacy budgets Table 4.3.

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. In fact, for all trials in Experiment 2, there was a significant difference in error between PAFER and the perfect estimator, in favor of the perfect estimator.

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 `f`ractional minleaf value of $\frac{1}{100}$ ensured that PAFER significantly outperformed the random baseline, as shown in Table 4.7. 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, as shown in Table 4.10.

RSQ2.2 At what fractional `minleaf` value is the perfect estimator significantly better at estimating Statistical Parity than PAFER?
A `minleaf` value of $\frac{1}{5}$ is not large enough to make PAFER competitive with the perfect estimator. The perfect estimator outperformed PAFER across all datasets, sensitive attribute analyses, `minleaf` values, and privacy budgets with $p \lll .001$.

5.2. 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. PAFER further showed to perform worse when the audited DT is less interpretable.

Experiment 2 showed that the smaller the leaf nodes of the DT are, the worse the performance is. PAFER thus trades off privacy and accuracy of estimation with interpretability; 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 and PrEq.

5.3. 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 three other new fairness metrics in Subsection 5.3.1 and suggest examining the different parameters of the PAFER Algorithm in Subsection 5.3.2.

5.3.1. 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 three 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 (Section 3.1), 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. PrEq would also fit this approach. 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.

5.3.2. Other input parameters

Examining the input parameters of the PAFER estimation algorithm in Algorithm 1, three clear candidates for further research become visible. These are the DP mechanism, $\mathcal{A}$, the model that is audited, DT, and the *policy* for handling invalid query answers. Subsubsection 5.3.2.1 through Subsubsection 5.3.2.3 discuss each input parameter, in order.

5.3.2.1. The DP 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 [52] shows promising results for a simple SP query. They use a DP mechanism based on smooth sensitivity [73]; 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.

Other mechanisms may be available if the query merely asks whether the DT adheres to the 80%-rule, i.e. if the SP of the DT is larger than 0.8. In this case, the Sparse Vector Technique is available [45]. This technique would allow for the auditing of multiple models without sharing the privacy budget. This would also make the method suitable for different types of ML models.

5.3.2.2. 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 [74], rule lists [75] and association rule data mining [76]. 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 [77]. 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.

5.3.2.3. The invalid value policy. Every experiment in this work used the same invalid value policy, namely the uniform approach. Further research can test other policies, to potentially improve PAFER's performance on smaller datasets and less interpretable DTs. Especially the 0 and 1 negative policies are intuitive, given that a negative query answer has a relatively high probability of being close to 0 or 1. The Gaussian mechanism currently relies a lot on the chosen invalid value policy, so investigation of better policies might make the Gaussian mechanism more competitive with the Laplacian. In total, Table 3.1 gives rise to $2 \times 4 = 8$ policies, providing enough experimental avenues.

REFERENCES

1. Zhu, L., D. Qiu, D. Ergu, C. Ying and K. Liu, “A study on predicting loan default based on the random forest algorithm”, *Procedia Computer Science*, Vol. 162, pp. 503–513, 2019, https://www.sciencedirect.com/science/article/pii/S1877050919320277.

2. Bickel, P. J., E. A. Hammel and J. W. O’Connell, “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*, Vol. 187, No. 4175, pp. 398–404, 1975, https://www.science.org/doi/abs/10.1126/science.187.4175.398.

3. Chouldechova, A., “Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments”, *Big Data*, Vol. 5, No. 2, pp. 153–163, Jun. 2017, http://www.liebertpub.com/doi/10.1089/big.2016.0047.

4. Barr, A., “Google Mistakenly Tags Black People as ‘Gorillas,’ Showing Limits of Algorithms”, , jul 2015, https://www.wsj.com/articles/BL-DGB-42522, section: Digits.

5. “Xenophobic machines: Discrimination through unregulated use of algorithms in the Dutch childcare benefits scandal”, , Oct. 2021, https://www.amnesty.org/en/documents/eur35/4686/2021/en/.

6. Adriaanse, M. L., “Kabinet-Rutte III gevallen om Toeslagenaffaire - NRC”, , Jan. 2021, https://web.archive.org/web/20210117020719/https://www.nrc.nl/nieuws/2021/01/15/kabinet-rutte-iii-gevallen-om-toeslagenaffaire-a4027684.

7. “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*, Vol. L 119, pp. 1–88, 14-04-2016.

8. "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*, Vol. COM/2021/206 final, pp. 1–107, 21-04-2021.

9. "Het Algoritmeregister van de Nederlandse overheid", , 2022, `https://algoritmes.overheid.nl/`.

10. "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*, Vol. L 60/34, pp. 34–85, 04-02-2014.

11. Cadwalladr, C. and E. Graham-Harrison, "Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach", *The Guardian*, mar 2018, `https://www.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-influence-us-election`.

12. "Losing Face: Two More Cases of Third-Party Facebook App Data Exposure | UpGuard", , 2019, `https://www.upguard.com/breaches/facebook-user-data-leak`.

13. Berendt, B. and S. Preibusch, "Better decision support through exploratory discrimination-aware data mining: foundations and empirical evidence", *Artificial Intelligence and Law*, Vol. 22, No. 2, pp. 175–209, 2014, `https://link.springer.`

com/article/10.1007/s10506-013-9152-0.

14. Mattu, S., J. Larson, L. Kirchner and J. Angwin, "Machine Bias", , 2016, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

15. Dwork, C., M. Hardt, T. Pitassi, O. Reingold and R. Zemel, "Fairness through awareness", *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference on - ITCS '12*, pp. 214–226, ACM Press, Cambridge, Massachusetts, 2012, http://dl.acm.org/citation.cfm?doid=2090236.2090255.

16. "Questions and Answers to Clarify and Provide a Common Interpretation of the Uniform Guidelines on Employee Selection Procedures", *FEDERAL REGISTER*, Vol. 44, No. 43, 01-03-1979.

17. Hardt, M., E. Price, E. Price and N. Srebro, "Equality of Opportunity in Supervised Learning", *Advances in Neural Information Processing Systems*, Vol. 29, Curran Associates, Inc., 2016, https://proceedings.neurips.cc/paper/2016/hash/9d2682367c3935defcb1f9e247a97c0d-Abstract.html.

18. Kleinberg, J., S. Mullainathan and M. Raghavan, "Inherent Trade-Offs in the Fair Determination of Risk Scores", C. H. Papadimitriou (Editor), *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*, Vol. 67 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pp. 43:1–43:23, Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, Dagstuhl, Germany, 2017, http://drops.dagstuhl.de/opus/volltexte/2017/8156.

19. Gupta, M., A. Cotter, M. M. Fard and S. Wang, "Proxy Fairness", , Jun. 2018, http://arxiv.org/abs/1806.11212, arXiv:1806.11212 [cs, stat].

20. Hashimoto, T., M. Srivastava, H. Namkoong and P. Liang, "Fairness Without Demographics in Repeated Loss Minimization", *Proceedings of the 35th Inter-*

national Conference on Machine Learning, pp. 1929–1938, PMLR, Jul. 2018, `https://proceedings.mlr.press/v80/hashimoto18a.html`, iSSN: 2640-3498.

21. Lahoti, P., A. Beutel, J. Chen, K. Lee, F. Prost, N. Thain, X. Wang and E. Chi, “Fairness without Demographics through Adversarially Reweighted Learning”, *Advances in Neural Information Processing Systems*, Vol. 33, pp. 728–740, Curran Associates, Inc., 2020, `https://proceedings.neurips.cc/paper/2020/hash/07fc15c9d169ee48573edd749d25945d-Abstract.html`.

22. Rawls, J., *Justice as fairness: A restatement*, Harvard University Press, 2001.

23. Sharma, S., J. Henderson and J. Ghosh, “CERTIFAI: A Common Framework to Provide Explanations and Analyse the Fairness and Robustness of Black-Box Models”, *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, pp. 166—-172, Association for Computing Machinery, New York, NY, 2020, `https://dl.acm.org/doi/abs/10.1145/3375627.3375812`.

24. Upadhyay, S., S. Joshi and H. Lakkaraju, “Towards Robust and Reliable Algorithmic Recourse”, *Advances in Neural Information Processing Systems*, Vol. 34, pp. 16926–16937, Curran Associates, Inc., 2021, `https://proceedings.neurips.cc/paper/2021/hash/8ccfb1140664a5fa63177fb6e07352f0-Abstract.html`.

25. van Rosmalen, Y., F. van der Steen, S. Jans and D. van der Weijden, “Bursting the Burden Bubble? An Assessment of Sharma et al.’s Counterfactual-based Fairness Metric”, , Nov. 2022, `https://bnaic2022.uantwerpen.be/wp-content/uploads/BNAICBeNeLearn_2022_submission_4430.pdf`.

26. Galhotra, S., Y. Brun and A. Meliou, “Fairness testing: testing software for discrimination”, *Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering*, pp. 498–510, ACM, Paderborn Germany, Aug. 2017, `https://dl.acm.org/doi/10.1145/3106237.3106277`.

27. Pedreschi, D., S. Ruggieri and F. Turini, "Discrimination-aware Data Mining", *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 560–568, 2008, https://dl.acm.org/doi/abs/10.1145/1401890.1401959.

28. Kilbertus, N., M. Rojas Carulla, G. Parascandolo, M. Hardt, D. Janzing and B. Schölkopf, "Avoiding discrimination through causal reasoning", *Advances in Neural Information Processing Systems*, Vol. 30, pp. 656–666, 2017, https://proceedings.neurips.cc/paper_files/paper/2017/file/f5f8590cd58a54e94377e6ae2eded4d9-Paper.pdf.

29. Kusner, M. J., J. Loftus, C. Russell and R. Silva, "Counterfactual Fairness", *Advances in Neural Information Processing Systems*, Vol. 30, Curran Associates, Inc., 2017, https://proceedings.neurips.cc/paper/2017/hash/a486cd07e4ac3d270571622f4f316ec5-Abstract.html.

30. Zhang, L., Y. Wu and X. Wu, "Situation Testing-Based Discrimination Discovery: A Causal Inference Approach.", *IJCAI*, Vol. 16, pp. 2718–2724, 2016, https://www.ijcai.org/Proceedings/16/Papers/386.pdf.

31. Pearl, J., *Causality*, Cambridge university press, 2009.

32. Bertsimas, D. and J. Dunn, "Optimal classification trees", *Machine Learning*, Vol. 106, pp. 1039–1082, 2017, https://link.springer.com/article/10.1007/s10994-017-5633-9.

33. Oliver, J. J. and D. Hand, "Averaging over decision stumps", *Machine Learning: ECML-94: European Conference on Machine Learning Catania, Italy, April 6–8, 1994 Proceedings 7*, pp. 231–241, Springer, 1994, https://link.springer.com/chapter/10.1007/3-540-57868-4_61.

34. Molnar, C., *Interpretable Machine Learning*, 2 edn., 2022, https://christophm.

github.io/interpretable-ml-book.

35. Borisov, V., T. Leemann, K. Seßler, J. Haug, M. Pawelczyk and G. Kasneci, "Deep neural networks and tabular data: A survey", *IEEE Transactions on Neural Networks and Learning Systems*, 2022, https://ieeexplore.ieee.org/abstract/document/9998482.

36. Mehrabi, N., F. Morstatter, N. Saxena, K. Lerman and A. Galstyan, "A Survey on Bias and Fairness in Machine Learning", *ACM Computing Surveys*, Vol. 54, No. 6, pp. 1–35, Jul. 2022, https://dl.acm.org/doi/10.1145/3457607.

37. Kamiran, F., T. Calders and M. Pechenizkiy, "Discrimination Aware Decision Tree Learning", *2010 IEEE International Conference on Data Mining*, pp. 869–874, IEEE, Sydney, Australia, Dec. 2010, http://ieeexplore.ieee.org/document/5694053/.

38. Ausiello, G., P. Crescenzi, G. Gambosi, V. Kann, A. Marchetti-Spaccamela and M. Protasi, *Complexity and approximation: Combinatorial optimization problems and their approximability properties*, Springer Science & Business Media, 2012.

39. Aghaei, S., M. J. Azizi and P. Vayanos, "Learning Optimal and Fair Decision Trees for Non-Discriminative Decision-Making", , Mar. 2019, http://arxiv.org/abs/1903.10598, arXiv:1903.10598 [cs, stat].

40. Jo, N., S. Aghaei, J. Benson, A. Gómez and P. Vayanos, "Learning Optimal Fair Classification Trees", , Jun. 2022, http://arxiv.org/abs/2201.09932, arXiv:2201.09932 [cs, math].

41. Linden, J. G. M. v. d., M. Weerdt and E. Demirović, "Fair and Optimal Decision Trees: A Dynamic Programming Approach", *Advances in Neural Information Processing Systems*, Oct. 2022, https://proceedings.neurips.cc/paper_files/paper/2022/file/fe248e22b241ae5a9adf11493c8c12bc-Paper-Conference.

pdf.

42. Ranzato, F., C. Urban and M. Zanella, "Fair Training of Decision Tree Classifiers", , Jan. 2021, http://arxiv.org/abs/2101.00909, arXiv:2101.00909 [cs].

43. Ranzato, F. and M. Zanella, "Genetic adversarial training of decision trees", *Proceedings of the Genetic and Evolutionary Computation Conference*, pp. 358–367, 2021, https://dl.acm.org/doi/abs/10.1145/3449639.3459286.

44. Dwork, C., "Differential Privacy", *33rd International Colloquium, ICALP 2006 on Automata, Languages and Programming*, pp. 1–12, Springer, 2006, https://link.springer.com/chapter/10.1007/11787006_1.

45. Dwork, C. and A. Roth, "The Algorithmic Foundations of Differential Privacy", *Foundations and Trends® in Theoretical Computer Science*, Vol. 9, No. 3-4, pp. 211–407, 2013, http://www.nowpublishers.com/articles/foundations-and-trends-in-theoretical-computer-science/TCS-042.

46. Warner, S. L., "Randomized response: A survey technique for eliminating evasive answer bias", *Journal of the American Statistical Association*, Vol. 60, No. 309, pp. 63–69, 1965, https://www.tandfonline.com/doi/abs/10.1080/01621459.1965.10480775.

47. McSherry, F. and K. Talwar, "Mechanism design via differential privacy", *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*, pp. 94–103, IEEE, 2007, https://ieeexplore.ieee.org/abstract/document/4389483/.

48. Blum, A., C. Dwork, F. McSherry and K. Nissim, "Practical privacy: the SuLQ framework", *Proceedings of the twenty-fourth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pp. 128–138, 2005, https://dl.acm.org/doi/abs/10.1145/1065167.1065184.

49. Friedman, A. and A. Schuster, "Data mining with differential privacy", *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 493–502, 2010, `https://dl.acm.org/doi/abs/10.1145/1835804.1835868`.

50. Mohammed, N., S. Barouti, D. Alhadidi and R. Chen, "Secure and private management of healthcare databases for data mining", *2015 IEEE 28th International Symposium on Computer-Based Medical Systems*, pp. 191–196, IEEE, 2015, `https://ieeexplore.ieee.org/abstract/document/7167484/`.

51. Fioretto, F., C. Tran, P. Van Hentenryck and K. Zhu, "Differential Privacy and Fairness in Decisions and Learning Tasks: A Survey", *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, pp. 5470–5477, International Joint Conferences on Artificial Intelligence Organization, Vienna, Austria, Jul. 2022, `https://www.ijcai.org/proceedings/2022/766`.

52. Hamman, F., J. Chen and S. Dutta, "Can Querying for Bias Leak Protected Attributes? Achieving Privacy With Smooth Sensitivity", , Nov. 2022, `http://arxiv.org/abs/2211.02139`, arXiv:2211.02139 [cs].

53. Candès, E. J. and M. B. Wakin, "An introduction to compressive sampling", *IEEE signal processing magazine*, Vol. 25, No. 2, pp. 21–30, 2008, `https://ieeexplore.ieee.org/abstract/document/4472240/`.

54. Jagielski, M., M. Kearns, J. Mao, A. Oprea, A. Roth, S. Sharifi Malvajerdi and J. Ullman, "Differentially Private Fair Learning", *Proceedings of the 36th International Conference on Machine Learning*, pp. 3000–3008, PMLR, May 2019, `https://proceedings.mlr.press/v97/jagielski19a.html`, iSSN: 2640-3498.

55. Agarwal, A., A. Beygelzimer, M. Dudík, J. Langford and H. Wallach, "A reductions approach to fair classification", *International Conference on Machine Learning*, pp. 60–69, PMLR, 2018, `https://proceedings.mlr.press/v80/agarwal18a.html`.

56. Höfler, M., H. Pfister, R. Lieb and H.-U. Wittchen, “The use of weights to account for non-response and drop-out”, *Social psychiatry and psychiatric epidemiology*, Vol. 40, pp. 291–299, 2005, https://link.springer.com/article/10.1007/s00127-005-0882-5.

57. Beutel, A., J. Chen, T. Doshi, H. Qian, A. Woodruff, C. Luu, P. Kreitmann, J. Bischof and E. H. Chi, “Putting fairness principles into practice: Challenges, metrics, and improvements”, *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 453–459, 2019, https://dl.acm.org/doi/abs/10.1145/3306618.3314234.

58. Zhao, T., E. Dai, K. Shu and S. Wang, “Towards Fair Classifiers Without Sensitive Attributes: Exploring Biases in Related Features”, *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, pp. 1433–1442, ACM, Virtual Event AZ USA, Feb. 2022, https://dl.acm.org/doi/10.1145/3488560.3498493.

59. Chai, J., T. Jang and X. Wang, “Fairness without Demographics through Knowledge Distillation”, *Advances in Neural Information Processing Systems*, 2022, https://proceedings.neurips.cc/paper_files/paper/2022/file/79dc391a2c1067e9ac2b764e31a60377-Paper-Conference.pdf.

60. Kamiran, F., A. Karim and X. Zhang, “Decision theory for discrimination-aware classification”, *2012 IEEE 12th International Conference on Data Mining*, pp. 924–929, IEEE, 2012, https://ieeexplore.ieee.org/abstract/document/6413831/.

61. Elliott, M. N., P. A. Morrison, A. Fremont, D. F. McCaffrey, P. Pantoja and N. Lurie, “Using the Census Bureau’s surname list to improve estimates of race/ethnicity and associated disparities”, *Health Services and Outcomes Research Methodology*, Vol. 9, pp. 69–83, 2009, https://link.springer.com/article/10.1007/s10742-009-0047-1.

62. Chen, J., N. Kallus, X. Mao, G. Svacha and M. Udell, “Fairness Under Unawareness: Assessing Disparity When Protected Class Is Unobserved”, *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 339–348, ACM, Atlanta GA USA, Jan. 2019, `https://dl.acm.org/doi/10.1145/3287560.3287594`.

63. “CFPB and DOJ Order Ally to Pay $80 Million to Consumers Harmed by Discriminatory Auto Loan Pricing”, , Dec. 2013, `https://www.consumerfinance.gov/about-us/newsroom/cfpb-and-doj-order-ally-to-pay-80-million-to-consumers-harmed-by-discriminato`

64. Awasthi, P., A. Beutel, M. Kleindessner, J. Morgenstern and X. Wang, “Evaluating fairness of machine learning models under uncertain and incomplete information”, *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 206–214, 2021, `https://dl.acm.org/doi/abs/10.1145/3442188.3445884`.

65. Navada, A., A. N. Ansari, S. Patil and B. A. Sonkamble, “Overview of use of decision tree algorithms in machine learning”, *2011 IEEE Control and System Graduate Research Colloquium*, pp. 37–42, Jun. 2011, `https://ieeexplore.ieee.org/abstract/document/5991826/`.

66. Wang, T., C. Rudin, F. Doshi-Velez, Y. Liu, E. Klampfl and P. MacNeille, “A Bayesian Framework for Learning Rule Sets for Interpretable Classification”, *Journal of Machine Learning Research*, Vol. 18, No. 70, pp. 1–37, 2017, `https://www.jmlr.org/papers/volume18/16-003/16-003.pdf`.

67. Szczepański, M., “Is data the new oil? Competition issues in the digital economy”, *EPRS in-depth analysis*, pp. 1–8, 2020, `https://www.europarl.europa.eu/RegData/etudes/BRIE/2020/646117/EPRS_BRI(2020)646117_EN.pdf`.

68. Fletcher, S. and M. Z. Islam, “Decision Tree Classification with Differential Pri-

vacy: A Survey", *ACM Computing Surveys*, Vol. 52, No. 4, pp. 1–33, Jul. 2020, https://dl.acm.org/doi/10.1145/3337064.

69. Kohavi, R. and B. Becker, "UCI Machine Learning Repository: Adult Data Set", , 2016, https://archive.ics.uci.edu/ml/datasets/adult.

70. Hofmann, H., "Statlog (German Credit Data) Data Set", , 2013, https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data).

71. Miller, G. A., "The magical number seven, plus or minus two: Some limits on our capacity for processing information.", *Psychological Review*, Vol. 63, No. 2, pp. 81–97, Mar. 1956, http://doi.apa.org/getdoi.cfm?doi=10.1037/h0043158.

72. Barredo Arrieta, A., N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila and F. Herrera, "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI", *Information Fusion*, Vol. 58, pp. 82–115, Jun. 2020, https://www.sciencedirect.com/science/article/pii/S1566253519308103.

73. Nissim, K., S. Raskhodnikova and A. Smith, "Smooth sensitivity and sampling in private data analysis", *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*, pp. 75–84, ACM, San Diego California USA, Jun. 2007, https://dl.acm.org/doi/10.1145/1250790.1250803.

74. Mendel, J. M., *Uncertain Rule-Based Fuzzy Systems: Introduction and New Directions, 2nd Edition*, Springer International Publishing, Cham, 2017, http://link.springer.com/10.1007/978-3-319-51370-6.

75. Angelino, E., N. Larus-Stone, D. Alabi, M. Seltzer and C. Rudin, "Learning Certifiably Optimal Rule Lists for Categorical Data", *Journal of Machine Learning Research*, Vol. 18, No. 234, pp. 1–78, 2018, http://jmlr.org/papers/v18/17-716.

html.

76. Yazgana, P. and A. O. Kusakci, “A literature survey on association rule mining algorithms”, *Southeast Europe Journal of soft computing*, Vol. 5, No. 1, 2016, http://scjournal.ius.edu.ba/index.php/scjournal/article/view/102.

77. Aytekin, C., “Neural Networks are Decision Trees”, , Oct. 2022, http://arxiv.org/abs/2210.05189, arXiv:2210.05189 [cs].

A. Additional experimental results

A.1. Pruned decision trees of Experiment 1

This section includes the DTs that were trained for Experiment 1. They are pruned, in the sense that non-distinguishing rules were removed.

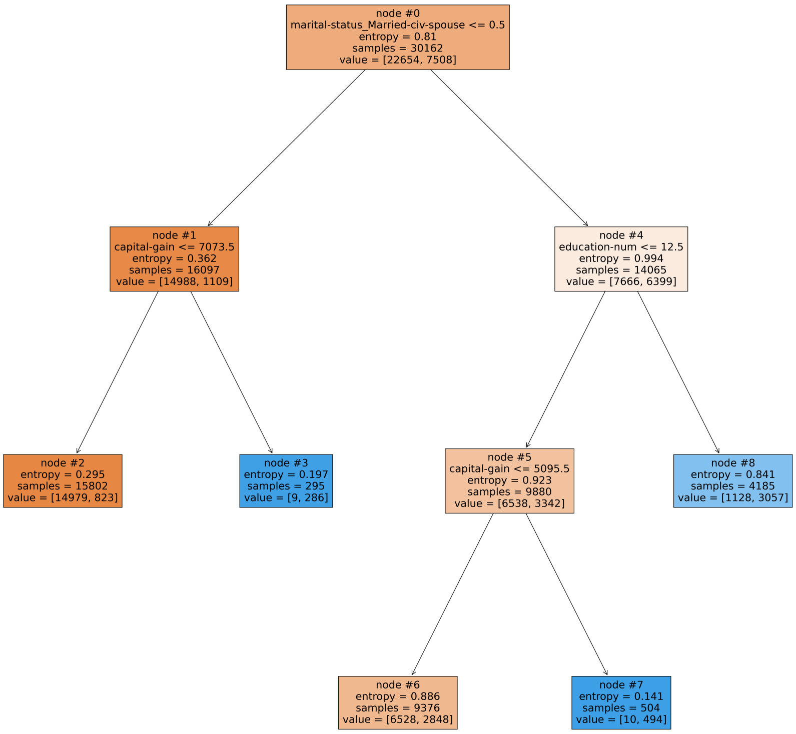

Figure A.1. The pruned DT that was used in Experiment 1 for the Adult dataset.

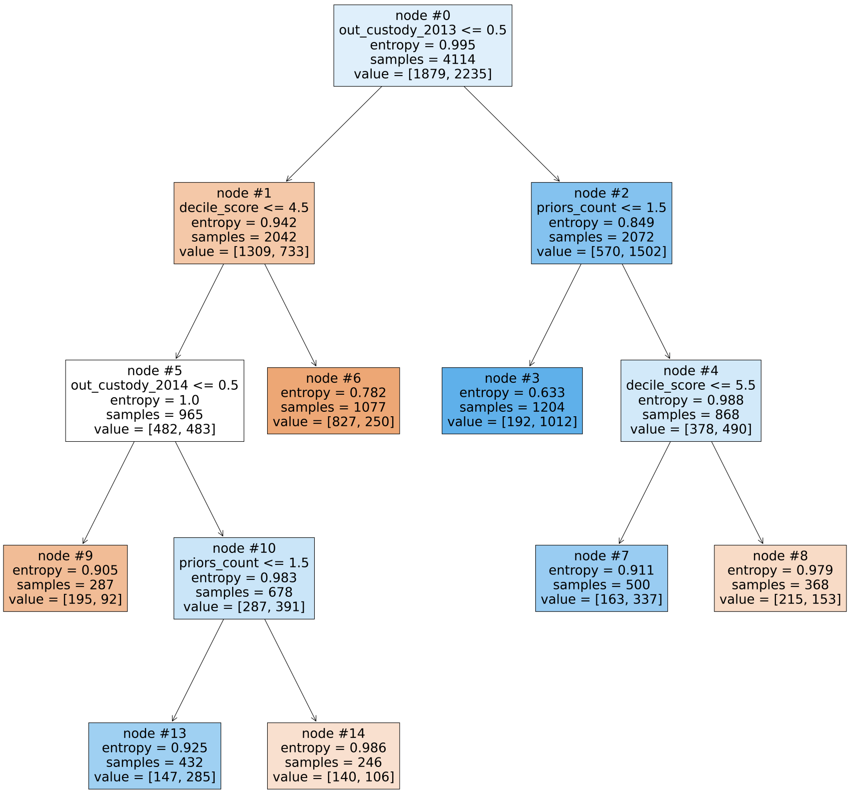

Figure A.2. The pruned DT that was used in Experiment 1 for the COMPAS dataset.

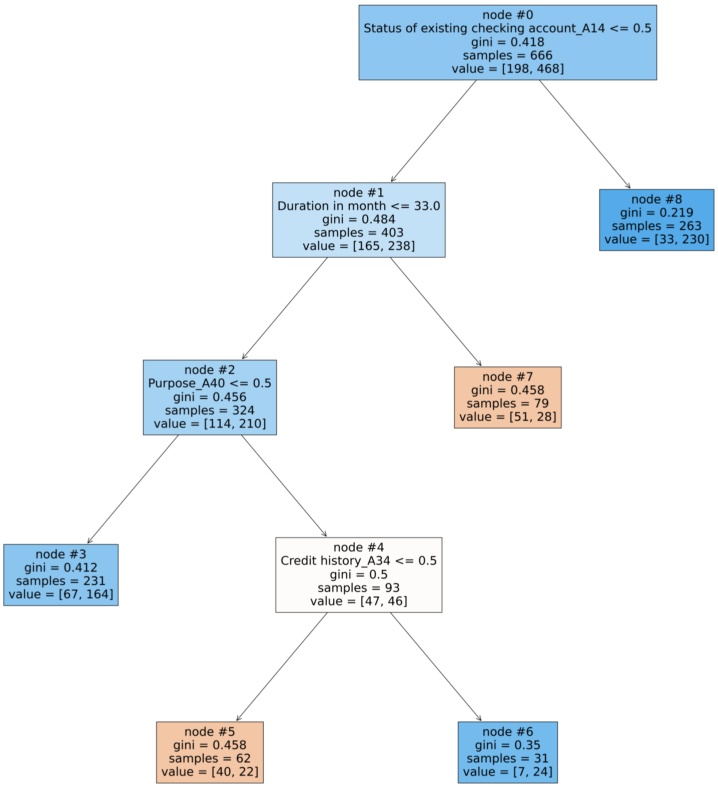

Figure A.3. The pruned DT that was used in Experiment 1 for the German dataset.

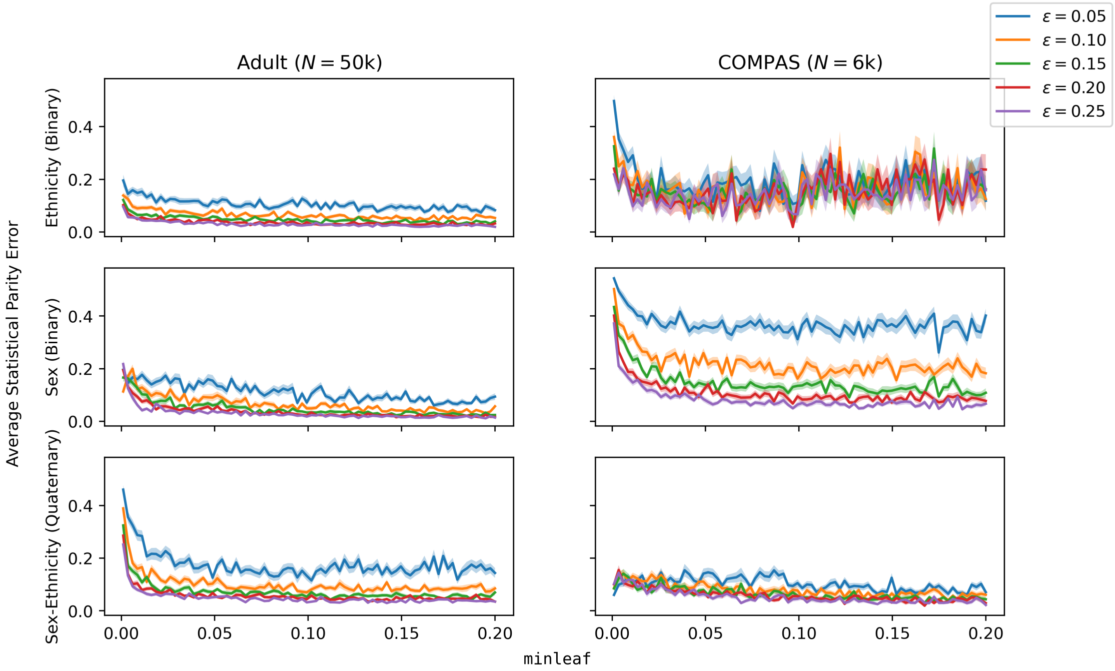

Figure A.4. The behavior of PAFER for different DTs. The uncertainty is pictured in a lighter color around the average.

A.2. Additional Experiment for RSQ1

Figure A.4 shows a clear answer to the second research question; the effect of the `minleaf` DT hyperparameter is clearly visible. The fewer instances the leaf nodes of the DT hold, the worse the performance of PAFER. Figure 4.1 alluded to this effect, showing a worse performance for the small German dataset. Figure A.4 shows that there is a trade-off between the interpretability of the DT and the performance of PAFER. When the `minleaf` value is low, the branches of the DT are generally longer and more abundant. Both more branches and longer branches in a DT negatively impact the interpretability of the DT [72]. Figure A.4 clearly shows improving performances for higher `minleaf` values.