Exploring Rigorous Fairness in Machine Learning

Student: Andreia Pereira
andreia.sofia.pereira@tecnico.ulisboa.pt

Tutor: Inês Lynce
ines.lynce@tecnico.ulisboa.pt

Novos Talentos em Inteligência Artificial
Fundação Calouste Gulbenkian

September 2021

Abstract

In this project, we studied the problem of Indirect Discrimination, as a part of the larger scope of Fairness in Machine Learning. We established the importance of properly detecting the presence of Proxy Attributes in a training dataset, and developed a rigorous and detailed approach to do so, in which we looked for logical relations of Implication and Equivalence between sensitive dataset attributes. We experimented our approach in various benchmark datasets.

Keywords — Fairness, Indirect Discrimination, Proxy Attribute

1 Introduction

The current presence of Machine Learning (ML) systems in multiple settings of our lives highlights the need for an adjustment of these systems, to make them acceptable for use around sensitive topics. A rising area of research in Artificial Intelligence (AI) goes by the name of *Fair AI*, calling for a better understanding of the biases ML systems can have and how they can be detected and avoided. Fairness is particularly important in systems that have direct impact in human lives. Machine-learning bias has been found in a selection of real-life applications, ranging from racial-biased *recidivism* predictions [9, 4, 3] to disparate online advertisement of job opportunities to female users [8]. To better reinforce the relevance of this topic, the EU has recently included fairness to be a requirement of trustworthy AI systems [2].

Until recently, research on the problem of AI fairness lacked consensus in definitions and rigorous (formal) analysis. Previous work has been focused in summarizing and evaluating the more than 20 different criteria of fairness proposed in the last few years [25], concluding however that it is not clear which criterion should be used in each situation. The lack of a rigorous analysis has been shown to be a problem in various domains such as criminal risk assessment, for example [4]. As a result, first steps are being taken to rigorously analyse and build fair ML models [17].

A common approach to address this problem is to look for bias in the training data, in an attempt to find in-balances or unfairness that are not inherent to the chosen ML architecture. To do this, the consensual approach is to define a set of *protected attributes*, such as race, gender, or age, in which the final classification should not be based upon. Most work on algorithmic fairness actively attempts to find unfairness against these protected attributes, based on the definition of several fairness metrics [24, 17, 20, 9, 11].

These protected attributes are to be selected by humans, and thus the selection process may be flawed or incomplete, as unfair discrimination may arise from *unprotected attributes* [22, 6]. One way this phenomena can be mitigated is by looking for *proxies* of the already selected *protected attributes.* The main contribution of this work is the use of logic relationships of Implication and Equivalence for the purpose of proxy detection, namely by finding these logical relations between *unprotected* and *protected attributes.*

In the following sections we will describe the problem formally, report our methodologies and experiments, and then highlight the main conclusions of our work.

1.1 Motivating Example - Red Lining in the U.S.

To further explain the relevance of detecting proxy attributes, we show a real-world example of what can be considered a proxy usage. Although long time abolished in the U.S., the *red-lining* practice was one in which financial services were denied disparately in minority communities. However, banks, insurers and health practitioners would disparately target individuals from certain *neighborhoods*, rather than necessarily targeting minority individuals. We can thus say that, due to the demographic distribution of minority citizens, the individual's neighbourhood was being used as a *proxy* for their *ethnicity* [10, 22].

While the red-lining practice was intentionally discriminatory, we cannot state that it is only malicious intent that causes indirect discrimination. A ML model that learns a classification problem using real-world data will be subject to the inherent biases or injustices present in that data, and is thus prone to learning to use a proxy attribute if one exists.

2 Preliminaries

In this section, we provide a few key formal definitions and nomenclature. We also summarize prior work, used as a base for this project.

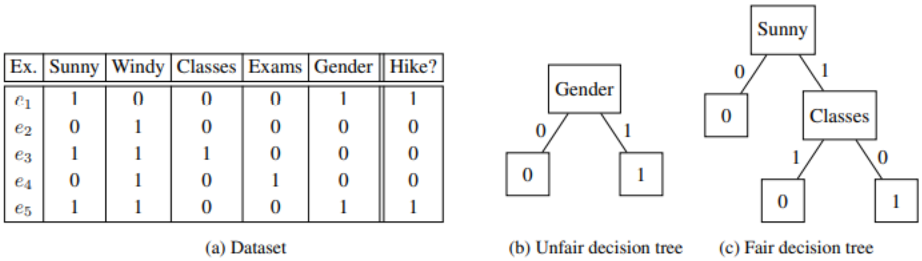
2.1 Notation - Classification Problem

Throughout this project, we will share the notation found in [17]. We consider a set of features $\mathcal{F} = \{F_1, ..., F_k\}$. Each feature F_i can take value in a domain $\mathcal{D}_i$. The space of all possible assignments of features is defined as $(F) = \Pi_i \mathcal{D}_i$ and will be referred to as *feature space.*

When building a classifier, a set of training data $\mathcal{T} = \{e_1, ..., e_M\}$, also known as a set of examples, is given. Each example is associated to a class, taken from the set of possible classes $\mathcal{C}$. Having trained the classifier, its goal is to, given an example, successfully classify it (i.e., associate it to a class in $\mathcal{C}$). Thus, a ML model $\mathbb{M}$ is represented as a function $\varphi : \mathbb{F} \rightarrow \mathcal{C}$.

Mehrabi et al. [21] consider an unfair algorithm as one whose decisions are skewed toward a particular group of people. Thus, a common concept to most fairness definitions and approaches, even though not always formalized, is the concept of *protected features* or *protected attributes* [17, 25, 21]. *Protected features* are the features in $\mathcal{F}$ for which we do not want to discriminate against.

So, we assume that $\mathcal{F}$ is further divided into two subsets: $\mathcal{P}$, which is the subset of protected features, and $\mathcal{N}$, which is the subset of non-protected features. We have that $\mathcal{P}$ and $\mathcal{N}$ are disjoint and complementary sets, thus, $\mathbb{F} = \mathbb{P} \times \mathbb{N}$.



Ex.	Sunny	Windy	Classes	Exams	Gender	Hike?
e_1	1	0	0	0	1	1
e_2	0	1	0	0	0	0
e_3	1	1	1	0	0	0
e_4	0	1	0	1	0	0
e_5	1	1	0	0	1	1

(a) Dataset

(b) Unfair decision tree

(c) Fair decision tree

Figure 1: Example of protected attributes

Example 1. This example [1] illustrates the use of protected attributes. Let us consider Figure 1, where $\mathcal{P} = \{Gender\}$ and $\mathcal{N} = \{Sunny, Windy, Classes, Exams\}$.

In this classification problem, the intention is to find out what are the conditions that cause students to enjoy a hike, and thus, we have $\mathcal{C} = \{0, 1\}$. One of the goals is to have a model that does not discriminate against *Gender*, that is in this case binary.

A decision tree like Figure 1 b) can be generated and only be dependent on *Gender*, rendering it unfair. However, it is also possible to reach Fig. 1 c), in which *Gender* is not used to decide.

2.2 Fairness Through Unawareness (FTU)

We will approach fairness considering *fairness through unawareness* (FTU) [17, 14, 20] as our fairness criterion. Although other fairness metrics have been proposed [11, 20], FTU has been shown to pose ideally for formal analysis [17] and, as we demonstrate in Section 3, is particularly sensitive to indirect biases by definition.

As exposed before, we have a ML classification scenario where fairness is to be assessed and achieved. Formal analysis has mostly been done in interpretable and logic-bases ML models, like Decision-Trees. We consider that we have a training dataset $\mathcal{T}$ and a set of features $\mathcal{F}$ that is further divided into a set of protected features $\mathcal{P}$ and a set of non-protected features $\mathcal{N}$. Our interpretable and logic-based ML model $\mathbb{M}$ is trained on $\mathcal{T}$.

Definition 2.1 (Fairness Through Unawareness (FTU) [17, 14, 20])**.** An algorithm is fair if the protected features $\mathcal{P}$ are not explicitly used in the decision-making process.

[1] Presentation of CP2020 paper "Towards Formal Fairness in Machine Learning" [17] `https://www.youtube.com/watch?v=UC1_eVEqOqc&t=344s`. Accessed August 2021.

A definition which proved useful to our work was that of a **biased dataset**. Although Ignatiev et al. [17] provide two distinct definitions of dataset bias depending on the consistency (or inconsistency) of the dataset, we will use the definition for a biased inconsistent dataset[2] as it is broader and better applies to real-world datasets.

Definition 2.2 (Dataset Bias under FTU [17])**.** A dataset $\mathcal{T}$ is labeled as **biased** if the following holds:

$$\exists(\mathbf{x} \in \mathcal{N})\ \exists(\mathbf{y_1}, \mathbf{y_2} \in \mathcal{P})\ \exists(c_1, c_2 \in \mathcal{C}).[y_1 \neq y_2 \wedge c_1 \neq c_2\ \wedge \{\langle x, y_1, c_1\rangle, \langle x, y_2, c_2\rangle\} \subset \mathcal{T}]$$

Although the reading of the above definition is not the most intuitive, its written explanation is simpler. A dataset is biased if two examples $e_1, e_2 \in \mathcal{T}$ that only differ in their values within the defined *protected features*, have a *different classification*.

2.3 Previous Work in Proxy Detection

Although most work on fairness is focused on finding bias coming from protected features, there is relevant work that studies bias caused by unprotected features. Most approaches look for strong correlation between unprotected and protected attributes [26, 6, 7], which is the direction we have also taken in this project.

There have been causal approaches [18], in which it is assumed that the classification problem and training dataset are accompanied by a causal graph that models the domain. A causal graph of a domain is often hard to find, specially in the social contexts where fairness seems to be more relevant. For this reason we have not taken this route.

Despite the different methodologies, the concept of a proxy attribute is consensual amongst all approaches. A proxy attribute is an unprotected attribute that behaves similarly to a protected attribute, and that can thus be used instead of it. This lays under the notion that discrimination does not solely happen as a consequence of the use of a protected class [6], and that an attribute that is either correlated with or influential on the protected attribute can also imprint discrimination.

Datta et al. [6, 7] define the notion of an $\epsilon -$ **proxy** using the Normalized Mutual Information [5] score between two attributes. This score can be used to describe the level of association between two attributes.

Definition 2.3 (Proxy Association [6])**.** Given two random variables $X \in \mathcal{N}$ and $Y \in \mathcal{P}$, the strength of a proxy, denoted $d(X, Y)$, is given by normalized mutual information:

$$d(X, Y) = 1 - \tfrac{H(X|Y)+H(Y|X)}{H(X)+H(Y)}$$

And X is considered an $\epsilon - proxy$ of Y if $d(X, Y) \leq \epsilon$

In Section 3 we will discuss the limitations of this methodology and illustrate examples where our approach can be more beneficial. Then, in Section 4 we compare our results with the ones of Datta et al [6].

[2] Informally, the condition for a training data $\mathcal{T}$ to be consistent is that two examples differently classified are necessarily different. If that condition does not apply, the dataset is inconsistent.

3 Indirect Discrimination

As we will demonstrate next, the FTU fairness metric [17] seems to be insufficient when it comes to handling discrimination coming from non-protected features [17, 22, 20]. Previous approaches have identified this phenomena as *indirect discrimination* [15, 10].

Two concepts arise when reasoning about indirect discrimination. We will call the first the existence of *proxy attributes*, and the second the existence of *irrelevant attributes.* We will define these concepts and what kinds of discrimination may arise from them, as well as attempt to advance the work on the *automatic detection* of this phenomena.

One important note is that this project was developed with the intent of aiding in the *pre-processing* of the datasets used in training. While there are other approaches, such as *in-processing* or *post-processing*, there is indication that pre-processing - i.e., tackling dataset bias prior to model training - might be the most effective [17, 23].

3.1 Proxy Discrimination

We will assume that attribute correlation should be tested as a mean of finding proxies, this is, attributes that behave similarly to the protected attributes, and thus may print unfairness to the classification. This is a problem often found in statistics; however, in this work we have focused on finding weather logical relations can be established between *unprotected* and *protected* attributes. We took two parallel routes, and focused on finding logical Equivalence and logical Implication between attributes.

The purpose of these devices is to detect potential proxies, and a human armed with domain knowledge will have the ultimate decision on weather an attribute, even if highly correlated with a protected attribute, is a proxy. Datta et al. formally describe this by assuming that an oracle O exists and can judge on whether a given proxy use is appropriate or not [6]. Nonetheless, it seems necessary to explore the prevailing relationships between protected and non-protected features.

3.1.1 Equivalence

Attribute equivalence or redundancy is a good sign of a proxy attribute. We started by designing two definitions that enclose this concept and can be checked in a time linear to the size of the tested dataset.

Definition 3.1 (Potential Proxy Criterion: Binary Perfect Equivalence)**.** Assuming a dataset $\mathcal{T}$ containing only binary attributes, a non-protected feature $\mathbf{u} \in \mathbb{N}$ is a potential proxy feature of a protected feature $\mathbf{v} \in \mathbb{P}$ iff

$$\forall u_i \in \mathbf{u}, v_i \in \mathbf{v}.[(u_i = \neg v_i) \lor (u_i = v_i)]$$

We extend Definition 3.1 to adapt to a dataset $\mathcal{T}$ that contains categorical attributes.

Definition 3.2 (Potential Proxy Criterion: Perfect Equivalence)**.** Assuming a dataset $\mathcal{T}$ containing both binary and categorical attributes, a non-protected feature $\mathbf{u} \in \mathbb{N}$, that takes values in the domain D_u, is a potential proxy feature of a protected feature $\mathbf{v} \in \mathbb{P}$, that takes values in the domain D_v, if

$$[\forall x \in D_u : (\mathbf{u}_i = x) \rightarrow \exists y \in D_v : (\mathbf{v}_i = y)] \land [\forall y \in D_v : (\mathbf{v}_i = y) \rightarrow \exists x \in D_u : (\mathbf{u}_i = x)]$$

Both definitions denote a type of proxy attribute that is detrimental to the FTU dataset bias test in 2.2. If a perfect proxy attribute exists to some protected attribute $\mathbf{y} \in \mathcal{P}$, then all classifications can be explained using the proxy attribute instead of the protected attribute, rendering a possibly false *FTU-fairness*.

If a potential proxy is found under this criterion and confirmed as a real proxy attribute, *FTU-fairness* arises once the model also becomes *unaware* of the proxy, either by removing it from the dataset, or by adding it to the *protected attributes* set $\mathcal{P}$.

Definitions 3.1 and 3.2 provide a good starting point but are too strict in practice. Because of this, we moved into a probabilistic approach in which we find a margin, or a percentage, of the examples in $\mathcal{T}$ that satisfy the condition in 3.2. The pseudo-code is shown in Algorithm 2, which we purposefully do not define just yet, as it will be built from Algorithm 1.

3.1.2 Implication

The Equivalence criterion seemed to have limitations regarding the real context and configuration of data, as we explain in Example 2. Because of this, we propose that attribute Implication (in both directions) should also be studied.

Example 2.

Neighborhood	Race	Class.
N1	B	
N2	B	
N3	W	
N1	B	
N3	W	
N4	W	

Table 1: An example dataset $\mathcal{T}_1$, portraying whether an applicant receives a housing loan or not. *Race* is protected and *Neighborhood* is unprotected.

Salutation	Gender	Class.
Sir	M	
Ms.	F	
Mr.	M	
Mrs.	F	
Mrs.	F	
Mr.	M	

Table 2: An example dataset $\mathcal{T}_2$, where the preferred salutation form is an attribute. Let us suppose *Gender* is protected and *Salutation* is unprotected.

In both tables we see a case where the value of the unprotected attribute perfectly predicts the value of the protected attribute. This is not true vice-versa. These are two similar cases where the equivalence test cannot possibly detect the unprotected attribute as a proxy of the protected attribute. The implication tests that we will further discuss can capture this *Many-to-One* behaviour that also should account for a proxy.

Definition 3.3 (Potential Proxy Criterion: Perfect Implication)**.** Assuming a dataset $\mathcal{T}$ containing both binary and categorical attributes, a non-protected feature $\mathbf{u} \in \mathbb{N}$, that takes values in the domain D_u, perfectly implies a protected feature $\mathbf{v} \in \mathbb{P}$, that takes values in the domain D_v, if

$$[\forall x \in D_u : (\mathbf{u}_i = x) \to \exists y \in D_v : (\mathbf{v}_i = y)]$$

Algorithm 1 Find Implication Margin

```
Input: A protected attribute $\mathbf{v}$. An unprotected attribute $\mathbf{u}$. Dataset $\mathcal{T}$
Output: Two margin values in the interval $[0, 1]$.
  function FindImplicationMargin($\mathbf{v}$, $\mathbf{u}$, $\mathcal{T}$)
      $n \leftarrow sizeof(\mathcal{T})$
      $right\_imp \leftarrow \{\}$
      $left\_imp \leftarrow \{\}$

      for $i \leftarrow 1, n$ do
          $right\_imp[\mathbf{v}_i][\mathbf{u}_i] + +$
          $left\_imp[\mathbf{u}_i][\mathbf{v}_i] + +$

      $right\_margin \leftarrow 0$
      for $x$ in $D_u$ do
          $max = $ maxValue($right\_imp[y]$)
          $right\_margin \leftarrow +max$
      $left\_margin \leftarrow 0$
      for $y$ in $D_v$ do
          $max = $ maxValue($left\_imp[y]$)
          $left\_margin \leftarrow +max$
       return $(left\_margin/n, right\_margin/n)$
```

As in the case of equivalence, perfect Implication would be hard to find in real-world datasets with thousands of examples. So, we adjust our definition:

Definition 3.4 (Potential Proxy Criterion: Margin Implication)**.** Assuming a dataset $\mathcal{T}$ containing both binary and categorical attributes, a non-protected feature $\mathbf{u} \in \mathbb{N}$ perfectly implies a protected feature $\mathbf{v} \in \mathbb{P}$ with a margin $\mathbf{m}$ if $\mathbf{u} \rightarrow \mathbf{v}$ for a fraction $\mathbf{m}$ of the examples $e_i \in \mathcal{T}$.

Similarly, we can also check for implication from the protected attribute to the unprotected attribute, by simply switching the definition. Algorithm 1 details how to find Margin Implication between two attributes in both directions.

After implementing Algorithm 1, we can develop a cycle that calls the above function for every pair of protected and unprotected attributes. This algorithm provides two margin scores, one for each possible direction of Implication. We found that these scores were not informative enough and decided to implement a visualization method that prints the detailed relationship between the domains of two attributes, D_v and D_u.

For Table 1 in Example 2, the result of this visualization is illustrated in Figure 2.

The margin implication scores of this example, according to our definition, are listed in Table 3 bellow. As we can see, the detailed domain implication visualization in Figure 2 give us more information about where the margin scores come from and why does *Neighbourhood* ($D_N = \{N1, N2, N3, N4\}$) perfectly implies *Race* ($D_R = \{B, W\}$), but the same does not happen vice versa.

Algorithm 2 serves to obtain the Equivalence Margin between two attributes, and is built using the previous Margin Implication algorithm.

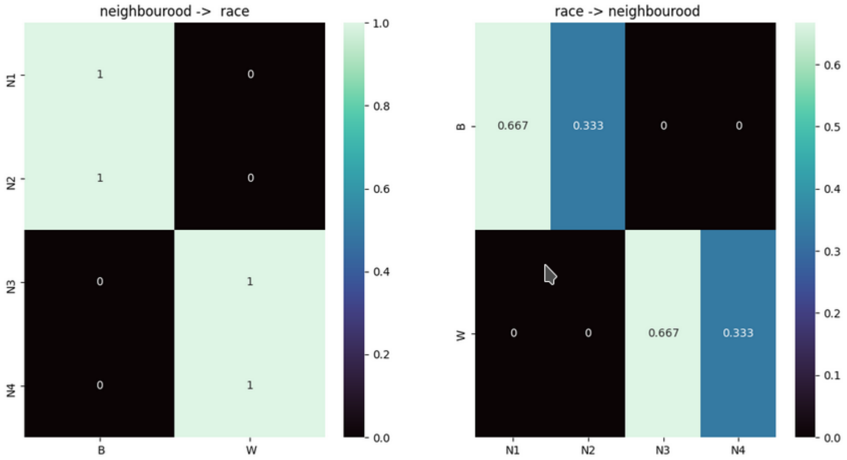


Figure 2: Detailed visualization of the domain implication of attributes *Neighborhood* and *Race*, regarding Example 2.1.

	Neighbourhood → Race	Race → Neighbourhood
Margin Implication Score	1.00	0.66

Table 3: The margin implication scores of the redlining Example 2.1.

3.2 Irrelevant Features

Although we have not gone much deep into the study of irrelevant features, one thing that is important to notice is that the definition 2.2 proposed by Ignatiev et al. [17] on dataset bias will not detect any unfairness if the data contains some sort of identification attribute, as is common practice.

This is because definition 2.2 can only detect unfairness when there exist two or more examples $e_1, e_2, ..., e_N \in \mathcal{T}$ that only differ in the protected attributes. If a dataset $\mathcal{T}$ contains an attribute that takes different values in each example in $\mathcal{T}$, then the previous stated criterion will not work.

Theorem 3.1. The FTU dataset bias criterion cannot be applied if the following holds:

$$\exists(\mathbf{x} \in \mathcal{N}).[\forall x_i, x_j \in \mathbf{x} : x_i \neq x_j]$$

The above condition may hold if the dataset contains an *id* attribute, for example, but it might also happen in smaller datasets with attributes such as date of birth, name, address, or any other highly variable attribute. The solution to this problem is to test the above condition for all *non-protected* attributes, and exclude those that verify the condition when applying the dataset bias test. This procedure also will not hurt the classification process, since a highly variable attribute will not be useful for the classification task - making it therefore safe to be removed.

Algorithm 2 Find Equivalence Margin

Input: A protected attribute $\mathbf{v}$. An unprotected attribute $\mathbf{u}$. Dataset $\mathcal{T}$
Output: A margin value in the interval $[0, 1]$

```
function FindEquivalenceMargin($\mathbf{v}$, $\mathbf{u}$, $\mathcal{T}$)
    $left\_margin, right\_margin \leftarrow$ FindImplicationMargin($\mathbf{v}, \mathbf{u}, \mathcal{T}$)
    $eq\_margin \leftarrow$ max($left\_margin$, $right\_margin$)
     return $eq_{margin}$
```

4 Experiments

We developed a set of Python scripts [3] that implement the algorithms described in Section 3. A few Python libraries were used to handle the datasets, as well as to provide some of the visualizations, as we show next. All experiments were run in a Linux environment, namely on Ubuntu 20.10, and the scripts were developed using Python 3.8.10, with help of the Pandas and Seaborne libraries.

4.1 Data Pre-Processing

There is indication that fluctuations in dataset composition, namely in the separation of categorical features and in the selection of protected features, may incur in different conclusions between different fairness-aware approaches and metrics [13].

This was important when deciding which version of the datasets to use in the experiments. We opted for not making the attributes binary, and so used the raw datasets. Most of the used datasets are in the categorical domain, and numerical attributes, such as age, are a clear limitation of this work. We also came across a few datasets with categorical *Age* or *Salary* attributes, that had already set certain intervals as categories. This is a good example on how dataset composition can change the results.

4.2 Benchmark Datasets

In this section we briefly explain each tested dataset. When counting the number of attributes, we excluded the classification or output attribute.

- *Compas* [3]. It is taken from a COMPAS algorithm on recidivism prediction, and is known to discriminate against African-American individuals [17]. The dataset has 11 attributes, and the protected ones are *African American*, *Asian*, *Hispanic*, *Native American*, *Other* and *Female*. Calmon et al. [10] found Indirect Discrimination to exist.

- *Adult* [1, 19]. It was collected from the 1994 Census database, and the classification task is to predict whether an individual's income exceeds $50K per year. It has 12 categorical attributes, where *Sex* and *Race* are protected. The dataset was found to be FTU-unfair with respect to both protected attributes [17]. Datta et al. [6] point *Relationship* as a proxy attribute of *Sex*.

[3] Source Code developed for this project. `https://github.com/decasppereira/Fair-Proxy-Detect`.

- *German Credit*[16]. This dataset classifies people as a good or bad credit risk, given a set of 21 attributes [12, 17], both categorical and numeric, with *Sex* and *Age* being protected.
- *Titanic*. In this dataset, the challenge is to predict whether an individual survived the sinking of Titanic [27]. It includes information about the passengers aboard the ship, contained in 7 attributes, being *Sex* the only protected attribute.

From early on, one conclusion was that datasets with few attributes are not ideal for proxy detection, as the probability of finding redundancy or correlation is low. This happened with the Ricci and the COMPAS datasets. The results we got can be consulted in the appendix section. We will go into detail in the Adult dataset, as it is the one providing the most interesting results.

4.3 Adult Dataset

We begin by showing the detailed visualization for two unprotected attributes - *Relationship* and *Country* - that we suspected could be proxies of the protected attributes *Sex* and *Race*, respectively. These results are shown in Figures 3 and 4.

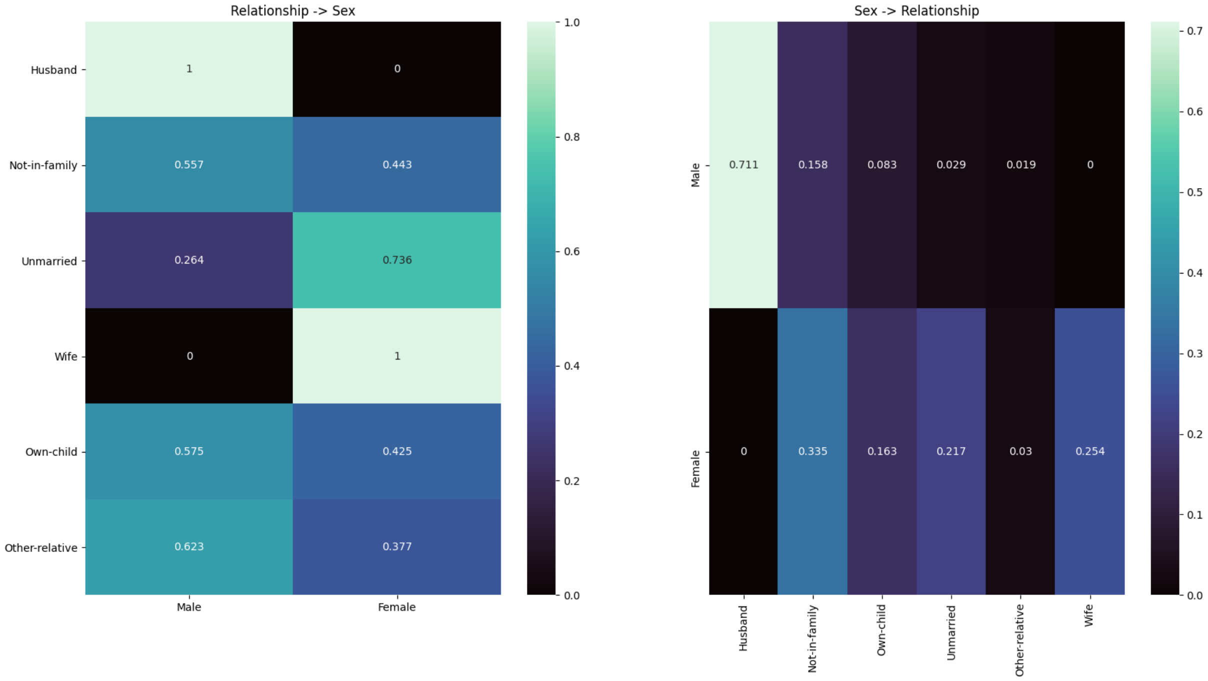


Figure 3: Detailed Implication results for the *Relationship* and *Sex* attributes of the Adult dataset.

We can see that, as suspected, there is a high correlation between the domain values of the attributes *Relationship* and *Sex*, as seen in Figure 3. We can see that:

$$
\begin{aligned}
Relationship = Husband &\rightarrow Sex = Male \\
Relationship = Wife &\rightarrow Sex = Female.
\end{aligned}
$$

Furthermore, due to dataset composition and an imbalance in *Female* and *Male* individuals in the dataset, all other possible values of the *Relationship* attribute are also imbalanced.

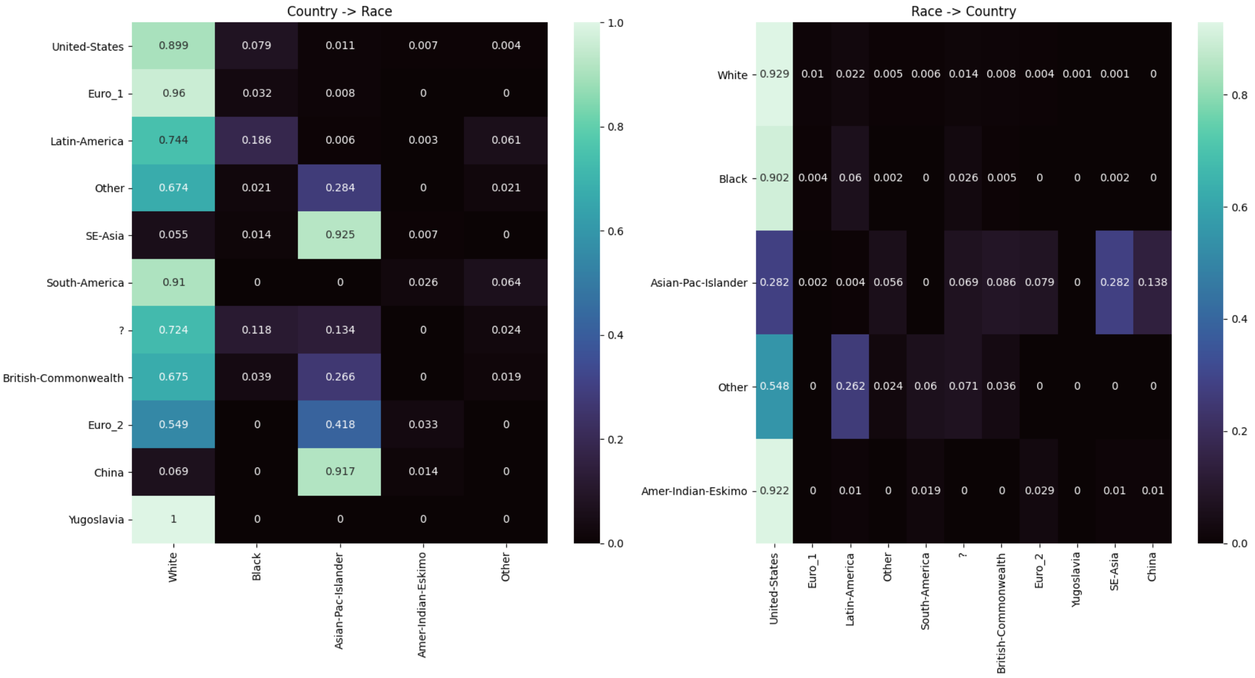


Figure 4: Detailed Implication results for the *Country* and *Race* attributes of the Adult dataset.

Figure 4 shows us that the domain values of the attributes *Race* and *Country* are also highly correlated and that a high percentage of examples follows a certain Implication rule. For example,

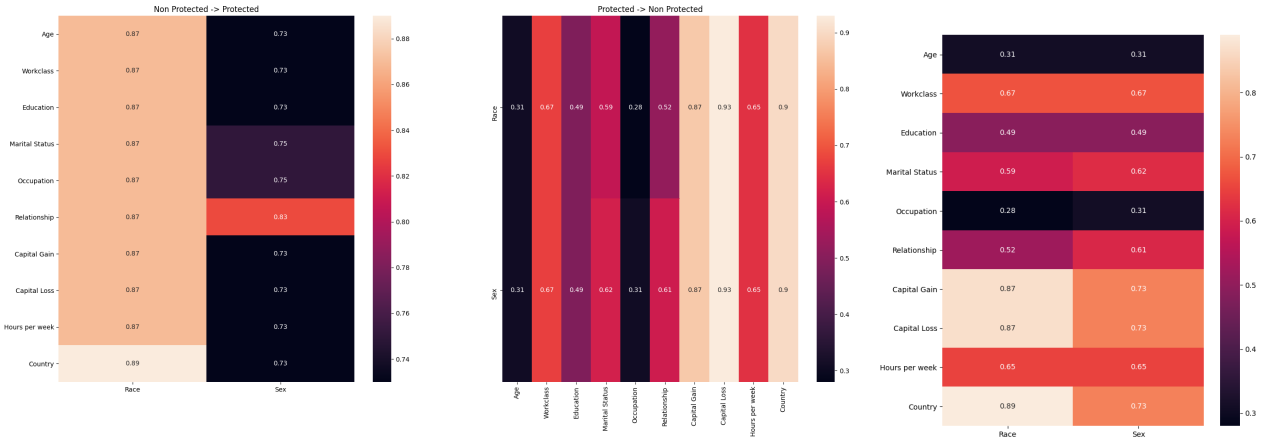
- In 100% of examples where $Country = Jugoslavia$, $Race = White$.
- In 91.7% of examples where $Country = China$, $Race =$ *Asian/Pacific Islander*.

Figure 5 shows the results of running Algorithms 1 and 2 for all pairs of protected and unprotected attributes in the Adult dataset. With these results, it becomes clear that *Relationship* is a proxy for *Sex*, as the score between *Relationship* and *Sex* stands out. The same thing happens for *Country* as a proxy for *Race*.

These two attributes show us a good example of a *Many-to-One* relationship that we were able to detect using the Implication test. As we see in the *Equivalence* plot of Figure 5, our equivalence test does not allow us to make any solid conclusions. However, we end up reaching the same conclusions as Datta et al. regarding the *Relationship* attribute, and additionally, we suggest *Country* as a proxy of *Race* in the Adult dataset.

Once we detected those two proxy attributes, we applied the FTU dataset bias criterion that Ignatiev et al. [17] propose and that is shown in Definition 2.2. We found that the Adult dataset is *FTU-unfair* regarding both proxy attributes *Relationship* and *Country*.

Two important considerations are to be made.



(a) Implication margins.

(b) Equivalence margins..

Figure 5: Results for the Adult dataset.

The first is that these scores are not normalized within the dataset, meaning that the composition of each group in the dataset examples (e.g. *Female* or *Male* in the *Sex* attribute) does impact the scores between all atribute pairs. This is noticeable in Figure 5, where we see that even the *Marital Status* and *Race* attributes have an implication score of 0.87 (in the *right* direction). If we were to take the scale $[0.73 - 0.89]$ and normalize it to $[0 - 1]$ the differences in scores would be more visible, but formal Definitions 3.3 and 3.4 would not be applicable.

The second is that that we do not assume a threshold value to determine whether an attribute is a proxy or not, as this seems to be out of our judgment abilities and should depend on context.

5 Conclusions

In this project, we extended the search for unfairness beyond protected attributes and studied the problem of proxy attributes in Machine Learning fairness. Namely, we studied how two logical operators - Implication and Equivalence - can aid in finding those attributes.

We proposed formal definitions for perfect Equivalence and Implication, and then adjusted them to fit real-world datasets, accepting that an attribute might still be a proxy even when not 100% of examples pass the tests. We developed several scripts to detect these proxies and tested them on several benchmark datasets, finding two proxy attributes in the Adult dataset. Furthermore, our option to visualize the behaviour of attribute's domains gave us a detailed view that was useful in confirming those proxies.

The study of the Implication operation proved useful as it captures a type of behaviour that previous work does not directly find or even look for. This was the case for the *Country* attribute in the Adult dataset, that had not previously been labeled as a proxy for attribute *Race*.

Overall, the studied logical operators emerge as valuable methods to find proxies and detect attribute relationships in a more detailed and transparent form.

6 Acknowledgements

I would like to thank the Calouste Gulbenkian Foundation for awarding me a grant in the fourth edition of the ”Novos Talentos em Inteligência Artificial” program, for all their support and the opportunity of being inserted in a group of outstanding students and academics in an area I am deeply interested in.

I would also like to thank and show my gratitude for Professor Inês Lynce, for the valuable advice, guidance and overall encouragement she provided me with throughout this project.

References

[1] Adult. UCI Machine Learning Repository, 1996.

[2] European union high-level expert group on artificial intelligence: Ethics guidelines for trustworthy AI. `https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai`, April 2019.

[3] J. Angwin, J. Larson, S. Mattu, and L. Kirchner. Machine bias. `http://tiny.cc/a3b3iz`, May 2016.

[4] R. Berk, H. Heidari, S. Jabbari, M. Kearns, and A. Roth. Fairness in criminal justice risk assessments: The state of the art, 2017.

[5] T. M. Cover and J. A. Thomas. *Elements of information theory (2. ed.).* Wiley, 2006.

[6] A. Datta, M. Fredrikson, G. Ko, P. Mardziel, and S. Sen. Proxy non-discrimination in data-driven systems. *CoRR*, abs/1707.08120, 2017.

[7] A. Datta, M. Fredrikson, G. Ko, P. Mardziel, and S. Sen. Use privacy in data-driven systems: Theory and experiments with machine learnt programs. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, CCS ’17, page 1193–1210, New York, NY, USA, 2017. Association for Computing Machinery.

[8] A. Datta, M. C. Tschantz, and A. Datta. Automated experiments on ad privacy settings: A tale of opacity, choice, and discrimination. *CoRR*, abs/1408.6491, 2014.

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

[10] F. du Pin Calmon, D. Wei, B. Vinzamuri, K. N. Ramamurthy, and K. R. Varshney. Optimized pre-processing for discrimination prevention. In I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 3992–4001, 2017.

[11] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. S. Zemel. Fairness through awareness. In S. Goldwasser, editor, *Innovations in Theoretical Computer Science 2012, Cambridge, MA, USA, January 8-10, 2012*, pages 214–226. ACM, 2012.

[12] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian. Certifying and removing disparate impact. In L. Cao, C. Zhang, T. Joachims, G. I. Webb, D. D. Margineantu, and G. Williams, editors, *Proceedings of the 21th ACM SIGKDD*

International Conference on Knowledge Discovery and Data Mining, Sydney, NSW, Australia, August 10-13, 2015, pages 259–268. ACM, 2015.

[13] S. A. Friedler, C. Scheidegger, S. Venkatasubramanian, S. Choudhary, E. P. Hamilton, and D. Roth. A comparative study of fairness-enhancing interventions in machine learning. In danah boyd and J. H. Morgenstern, editors, *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* 2019, Atlanta, GA, USA, January 29-31, 2019*, pages 329–338. ACM, 2019.

[14] N. Grgic-Hlaca, M. B. Zafar, K. P. Gummadi, and A. Weller. Beyond distributive fairness in algorithmic decision making: Feature selection for procedurally fair learning. In S. A. McIlraith and K. Q. Weinberger, editors, *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 51–60. AAAI Press, 2018.

[15] S. Hajian and J. Domingo-Ferrer. A methodology for direct and indirect discrimination prevention in data mining. *IEEE Trans. Knowl. Data Eng.*, 25(7):1445–1459, 2013.

[16] H. Hofmann. Statlog (German Credit Data). UCI Machine Learning Repository, 1994.

[17] A. Ignatiev, M. C. Cooper, M. Siala, E. Hebrard, and J. Marques-Silva. Towards formal fairness in machine learning. In H. Simonis, editor, *Principles and Practice of Constraint Programming - 26th International Conference, CP 2020, Louvain-la-Neuve, Belgium, September 7-11, 2020, Proceedings*, volume 12333 of *Lecture Notes in Computer Science*, pages 846–867. Springer, 2020.

[18] N. Kilbertus, M. Rojas-Carulla, G. Parascandolo, M. Hardt, D. Janzing, and B. Schölkopf. Avoiding discrimination through causal reasoning. In I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 656–666, 2017.

[19] R. Kohavi. Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In E. Simoudis, J. Han, and U. M. Fayyad, editors, *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96), Portland, Oregon, USA*, pages 202–207. AAAI Press, 1996.

[20] M. J. Kusner, J. R. Loftus, C. Russell, and R. Silva. Counterfactual fairness. In I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 4066–4076, 2017.

[21] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan. A survey on bias and fairness in machine learning. *CoRR*, abs/1908.09635, 2019.

[22] A. E. Prince and D. Schwarcz. Proxy discrimination in the age of artificial intelligence and big data. *Iowa Law Review*, (3), 2020.

[23] K. H. Tae, Y. Roh, Y. H. Oh, H. Kim, and S. E. Whang. Data cleaning for accurate, fair, and robust models: A big data - AI integration approach. In S. Schelter, N. Polyzotis, S. Seufert, and M. Vartak, editors, *Proceedings of the 3rd International Workshop on Data*

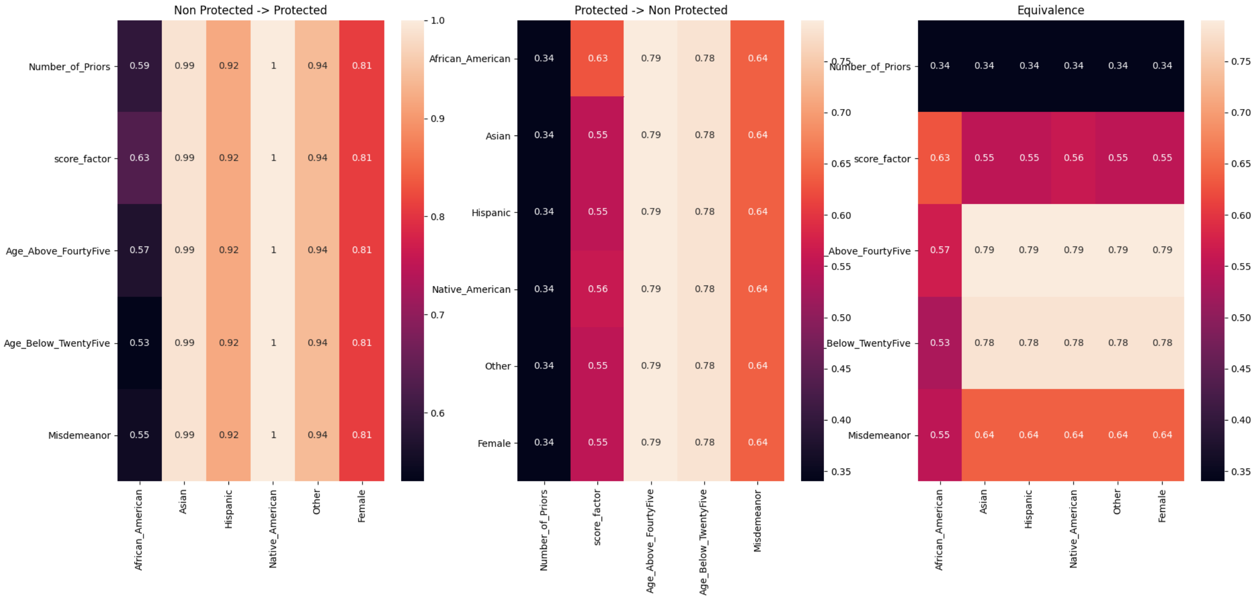


Figure 6: Implication and Equivalence margins for the COMPAS dataset.

Management for End-to-End Machine Learning, DEEM@SIGMOD 2019, Amsterdam, The Netherlands, June 30, 2019, pages 5:1–5:4. ACM, 2019.

[24] F. Tramèr, V. Atlidakis, R. Geambasu, D. J. Hsu, J. Hubaux, M. Humbert, A. Juels, and H. Lin. Fairtest: Discovering unwarranted associations in data-driven applications. In *2017 IEEE European Symposium on Security and Privacy, EuroS&P 2017, Paris, France, April 26-28, 2017*, pages 401–416. IEEE, 2017.

[25] S. Verma and J. Rubin. Fairness definitions explained. In Y. Brun, B. Johnson, and A. Meliou, editors, *Proceedings of the International Workshop on Software Fairness, FairWare@ICSE 2018, Gothenburg, Sweden, May 29, 2018*, pages 1–7. ACM, 2018.

[26] S. Yeom, A. Datta, and M. Fredrikson. Hunting for discriminatory proxies in linear regression models. In S. Bengio, H. M. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 4573–4583, 2018.

[27] T. Zhang, T. Zhu, M. Han, J. Li, W. Zhou, and P. S. Yu. Fairness constraints in semi-supervised learning. *CoRR*, abs/2009.06190, 2020.

A Full Datasets Results

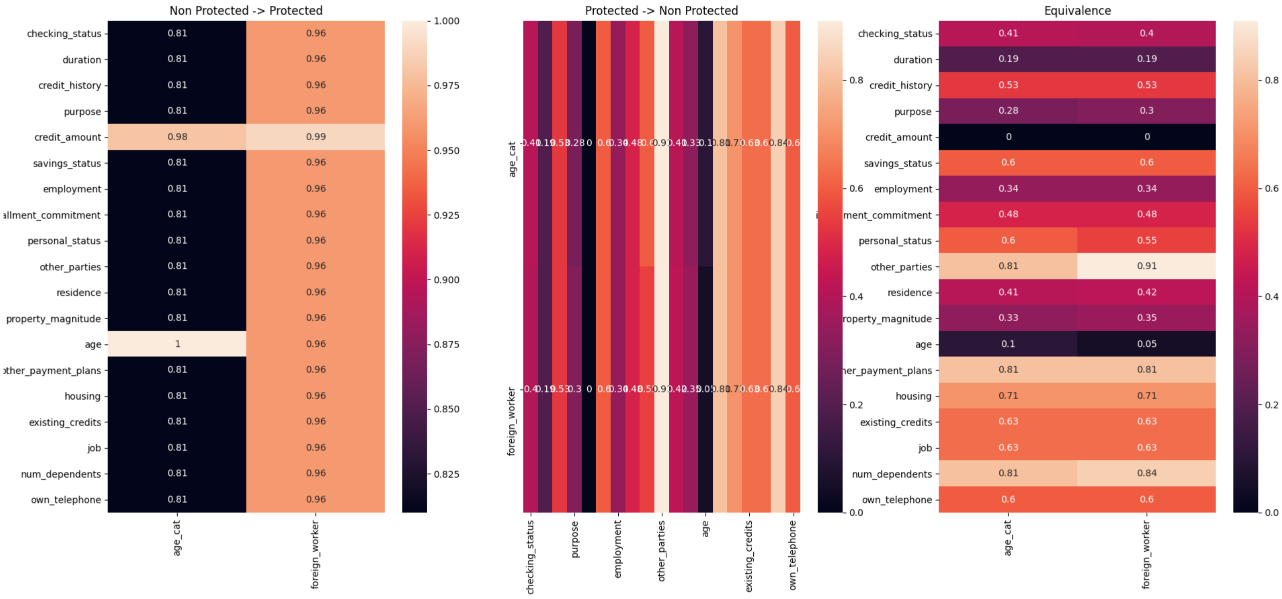


Figure 7: Implication and Equivalence margins for the German dataset.

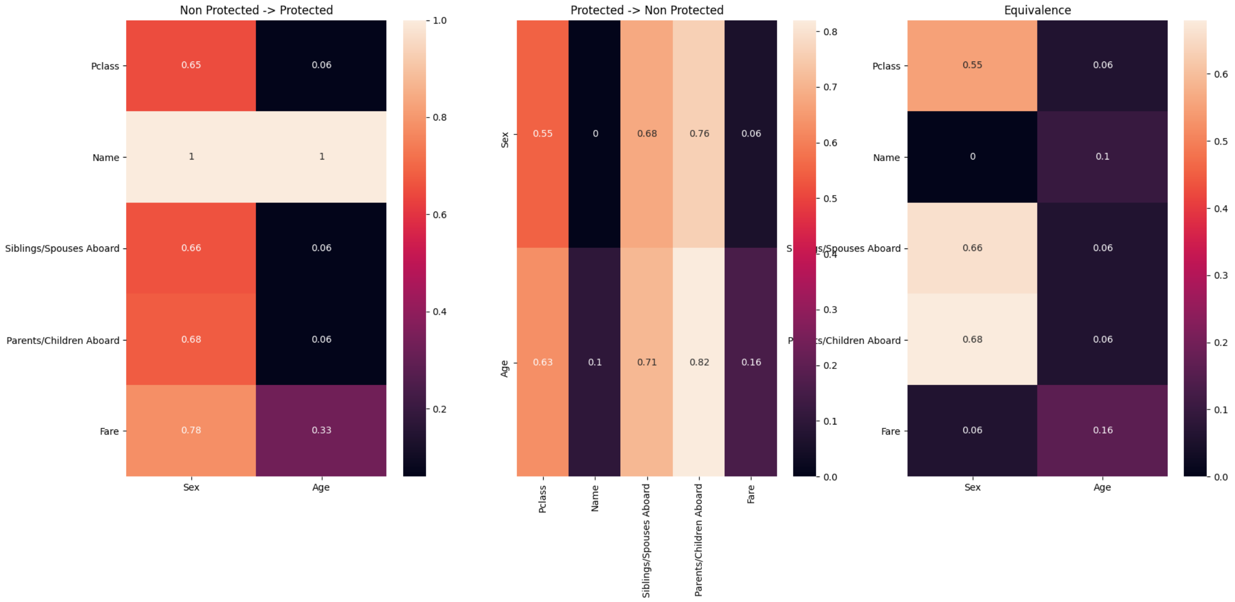


Figure 8: Implication and Equivalence margins for the Titanic dataset.