Algorithmic Fairness from Theory to Practical Application. A Fairness Audit of NamSor′s Gender and Ethnicity Classification Algorithms.

Research Project

Linda Fernsel
s0555949

July 29, 2020

Supervisor Prof. Dr. Gefei Zhang (HTW Berlin)
Supervisor Elian Carsenat (NamSor)

Abstract

Some cases of biased algorithms have become famous, for example Amazon's recruiting tool that discriminated women [1] and face recognition software performing worst for black women [2]. But what does bias mean exactly? How is fairness measured scientifically, where does unfairness come from and what are ways to reduce bias? Going from the general answers to these questions to concrete answers, two API endpoints of "NamSor" for inferring gender and ethnicity from a name will be audited for fairness.

[1] https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

[2] http://gendershades.org/overview.html

Acknowledgments

In 2019 I wrote my Bachelor Thesis about the "State and Development of Gender Equality in Computer Science Publications". In my thesis I analyzed bibliometric data from dblp[3] by first classifying all author names into either male or female names using the onomastic gender classification API "NamSor" version 2.0.2 beta. When doing so I noticed how names classified as male where classified with a higher score than names classified as female (Fernsel 2019, p. 14). Was it possible that an algorithm performed worse identifying women as such than identifying men? I could not investigate this further in the course of my Bachelor Thesis, but my interest was ignited. In the year after my thesis, a lot of work was done to improve fairness in "NamSor"'s results, as can be seen from "NamSor"'s update log [4]. In this year I got the chance to further pursue this topic with a study project on "Algorithmic Fairness from Theory to Practical Application". I want to thank Elian Carsenat of "NamSor" and Prof. Dr. Gefei Zhang for providing me with the necessary opportunity, support and feedback.

I also want to thank Dr. Martin Bretschneider for the *LaTeX* template used in this document. The template is available for free at `https://www.bretschneidernet.de/tips/arbeit-vorlagen.zip`.

I used diagrams.net to create the diagram that can be found in this document.

Linda Fernsel, July 29, 2020

[3] `https://dblp.uni-trier.de/`

[4] "NamSor"'s update log can be found at `https://github.com/namsor/namsor-tools-v2/wiki/NamSor-Tools-V2`.

List of Figures

List of Tables

Contents

1 Introduction

Unfair algorithms have the ability to widen existing gaps between social groups and thus to increase injustice on a massive scale, as Cathy O'Neil outlines in her book "Weapons of Math Destruction" (O'Neil 2016). O'Neil identifies three perspectives for taking action. First, *governments* can regulate algorithms and their use of data (O'Neil 2016, p. 212). Secondly, any *creator of algorithms* can implement fairness in their algorithm (O'Neil 2016, p. 207) and be transparent about possible biases discovered (O'Neil 2016, p. 212). Finally, *researchers* can evaluate the impact of algorithms (O'Neil 2016, p. 208) and audit algorithms otherwise kept in a black box (O'Neil 2016, p. 218).

Approaching the matter of fairness in algorithms from a researcher's point of view is the objective in this work. The final goal is to audit gender and ethnicity classification algorithms for names created by the European company "NamSor" [1]. For the audit the algorithms' fairness will be measured for the concrete case of the "COMPAS" data set. The "COMPAS" data set contains predictions of criminal behavior by individuals in the US (Flores, Bechtel & Lowenkamp 2016, p. 7). These predictions were calculated using the tool "Correctional Offender Management Profiling for Alternative Sanctions" ("COMPAS") (Flores et al. 2016, p. 7). The fairness audit will use Python and the data audit tool "Aquitas" [2]).

To form a basic understanding of the concept of fairness, metrics for fairness, potential sources for bias and ways to mitigate bias, the first part of this report gives an introduction to these topics based mainly on the MOOC "Bias and Discrimination in AI" of the University of Montreal (*Bias and Discrimination in AI* 2020). In the second part, "NamSor"'s classification algorithms, the data set "COMPAS" and the tool "Aequitas" are explained. Then the method used for auditing the algorithms' fairness and accuracy is explained. The findings from the fairness audit are layed out in the third part. The findings are critically assessed and discussed in the fourth part. The report ends with a conclusion on the findings and their implications for the future.

[1] https://www.namsor.com/

[2] http://aequitas.dssg.io/

2 Basics

This chapter summarizes theory from the field of Fairness in Algorithms that is important for the audit of "NamSor"'s algorithms later on. This chapter will answer the questions what fairness is, how to measure fairness, where unfairness comes from and how to make algorithms fair.

2.1 The Concept of Fairness

2.1.1 Fairness in Law and Policy

Fairness can be defined as the equal treatment of people independently of one or several protected features in a specific situation. Therefore one can say that fairness is *relative* to a protected feature and to a situation or task (*Bias and Discrimination in AI* 2020). A feature, or property of a person, like race, sex or religion, can be protected by law, such as in Article 2 of the "Universal Declaration of Human Rights":

> Everyone is entitled to all the rights and freedoms set forth in this Declaration, without distinction of any kind, such as race, colour, sex, language, religion, political or other opinion, national or social origin, property, birth or other status. (*The International Bill of Human Rights* 1948, Art. 2)

Article 14 in the (European) "Convention for the Protection of Human Rights and Fundamental Freedoms" adds the "association with a national minority" to this list of protected features (*Convention for the Protection of Human Rights and Fundamental Freedoms* 1950, Art. 14). Furthermore, Article 21 of the "Charter of Fundamental Rights of the European Union" adds "ethnic or social origin, genetic features, [...], disability, age or sexual orientation" (*Charter of Fundamental Rights of the European Union* 2000, Art. 21).

As fairness is not only relative to a feature but also to a situation, law does not only define protected features such as race and gender, but it also directly defines fairness relative to specific situations. It does so in Article 23.2 of "The Universal Declaration of Human Rights" for example, where fairness is defined as "equal pay for equal work" (*The International Bill of Human Rights* 1948, Art. 23.2).

Another example can be found in Article 25.1 of "The Universal Declaration of Human Rights", where features like "unemployed", "disabled" or "elderly" are protected for the specific situation of being granted security (*The International Bill of Human Rights* 1948, Art. 25.1).

To ensure that Human Rights are respected in the field of AI, the European Commission founded the "High-Level Expert Group on Artificial Intelligence" which developped the "Ethics Guidelines for Trustworthy Artificial Intelligence" (*Ethics Guidelines for Trustworthy AI* 2019, p. 6).

Among transparency and accountability, "diversity, non-discrimination and fairness" is one requirement for trustworthy AI (*Ethics Guidelines for Trustworthy AI* 2019, p. 2). The pilot version of the AI assessment list given in (*Ethics Guidelines for Trustworthy AI* 2019) provides guidance on how to avoid unfair bias. To "establish a strategy or a set of procedures to avoid creating or reinforcing unfair bias in the AI system" one should "assess [...] limitations stemming from the composition of the used data sets" and "consider diversity [...] of users in the data" as well as "test for [...] problematic use cases" (*Ethics Guidelines for Trustworthy AI* 2019, p. 29). Additionally, it is called out to "ensure a mechanism that allows others to flag issues related to bias" (*Ethics Guidelines for Trustworthy AI* 2019, p. 30). It is also important to "ensure an adequate working definition of “fairness”", checking whether the chosen definition is "commonly used" and also "consider[ing] other definitions" (*Ethics Guidelines for Trustworthy AI* 2019, p. 30). The chosen fairness definition also needs to be measured (*Ethics Guidelines for Trustworthy AI* 2019, p. 30).

2.1.2 Two manifestations of unfairness: Bias and discrimination

Someone who is biased has an unfair point of view, while discrimination happens when unfairness, or biases, manifest themselves in the real world, impacting protected classes (*Bias and Discrimination in AI* 2020). Biases can be overcome and are no excuse for discrimination (*Bias and Discrimination in AI* 2020).

Three forms of discrimination can be observed.

First, *direct discrimination*. When directly discriminating against someone, one behaves according to one’s biases (*Bias and Discrimination in AI* 2020). For instance one might decline someone’s application because to the reviewer, the applicant seem unfit for the job due to their gender.

Opposed to direct discrimination there is *indirect discrimination*. It occurs when one is applying a rule that leads to discrimination (*Bias and Discrimination in AI* 2020). For example, an HR employee could sort out any application that has too many spelling errors, thereby discriminating against anyone who is not writing in their native language.

The third form of discrimination is *systematic discrimination*. It occurs when bias is embedded in a system leading to the reproduction of bias (*Bias and Discrimination in AI* 2020). For example, a company might require an on-place interview with candidates, excluding candidates living far away or candidates that can not afford to take a day off.

2.2 Metrics for Fairness

This section will explain different fairness metrics using the following example case. Let’s imagine there is a binary classification tool that classifies a set of input features. A binary classification tool could be an automatic CV review tool, that returns whether someone is estimated to be suited for the position or not. The model takes non-sensitive features X and sensitive features G (protected features, like gender or proxies for these). When using the tool it makes predictions D on whether the candidate is estimated to be a good match for a job. The actual outcomes are Y, whether the candidate is really a good match. A result for D or Y of "0" means "not a good match", while a result of "1" means "good match".

One may differ beween group fairness and individual fairness, where group fairness means that there is no discrimination against a group of people and individual fairness means that there is no discrimination against any person (*Bias and Discrimination in AI* 2020). A system, that is fair to a group on average, is not necessarily fair to an individual from that group (*Bias and Discrimination in AI* 2020). In order to analyze fairness in algorithms, one definition at a time has to be considered. The following metrics could be used to measure group fairness, while the last section is about individual fairness.

The summary of these metrics given in (*Bias and Discrimination in AI* 2020) is not considered to be detailed enough, which is why this section of the report is based on the paper "Fairness Definitions Explained" by Sahil Verma and Julia Rubin (Verma & Rubin 2018) that has also been referenced in (*Bias and Discrimination in AI* 2020).

The reader is expected to know about confusion matrix metrics (TP, FP, FN, TN) and the error metrics that can be derived from them (PPV, FDR, FOR, NPV, TPR, FPR, FNR and TNR). They are summarized in (Verma & Rubin 2018, p. 2f).

2.2.1 Statistical Parity

Statistical Parity is also named Demographic Parity, Equal Acceptance Rate, Independence or Group Fairness (Verma & Rubin 2018, p. 3). It measures whether the probability P of being assigned to the positive (accepting) class ($D = 1$) is the same for all distinct values of sensitive features G (Verma & Rubin 2018, p. 3). "The same" in this context means that they are not significantly different or that their difference is smaller than a chosen treshold. "The probability of being assigned to the positive class" is also called "Predicted Prevalence" and abbreviated with PPrev (*Understanding the Metrics* 2018). The probability calculated is a conditional probability: The probability for being accepted under the condition of a specific sensitive feature value.

Let's suppose G ("Gender") has only two distinct values: "f" ("female") and "m" ("male") (because whoever recorded the data failed to provide other options). Using Statistical Parity it can be measured whether men and women have the same chance of being accepted when applying for a job.

The mathematical condition for Statistical Parity in the case of the job application example is $P(D = 1|G = f) = P(D = 1|G = m)$ (Verma & Rubin 2018, p. 3).

In words: The probability for being assigned to the accepting class if the value of gender is "female" equals the probability for being assigned to the accepting class if the value of gender is "male".

To make clear that the formula above is only valid for this specific case, consider the following two cases where the formula needs to be adapted.

1. If one wanted to check for multiple protected features, eg. taking into account whether English is the first language ($L = 1$) or not ($L = 0$), one could extend the formula to the following:

$P(D = 1|G = fandL = 1) = P(D = 1|G = fandL = 0) = (D = 1|G = mandL = 1) = (D = 1|G = mandL = 0)$

2. If there are three distinct possible values for G, "d", "f" and "m", the formula needs to be extended like so:

$P(D = 1|G = f) = P(D = 1|G = m) = P(D = 1|G = d)$

2.2.2 Conditional Statistical Parity

Statistical Parity might not always be considered fair. For certain tools, like job application review tools, it might be necessary to take into account a legitimate factor L, for instance the qualification. This way, you could control that members of one group, that have the same qualifications as members from another group, are not discriminated against (Verma & Rubin 2018, p. 3).

In that case, the condition for Conditional Statistical parity is $P(D = 1|L = l, G = f) = P(D = 1|L = l, G = m)$ (Verma & Rubin 2018, p. 3).

In words: The probability for being assigned to the accepting class if the value of Qualification is l and the value of gender is "female" equals the probability for being assigned to the accepting class if the value of Qualification is l and the value of gender is "male".

2.2.3 Predictive Parity

Predictive Parity is also called the Outcome Test (Verma & Rubin 2018, p. 3). It measures whether the probability P of belonging to the

1. positive class ($I = 1$) (PPV)
2. negative class ($I = 0$) (FDR)

if having been assigned to the positive class ($D = 1$) is the same for all distinct values of sensitive features (Verma & Rubin 2018, p. 3).

In the case of job applications one could use Predictive Parity to measure whether male and female applicants where equally likely to be accepted, no matter whether they are actually a good match for the job.

Comparing male and female applicants the condition for Predictive Parity is $P(Y = I|D = 1, G = f) = P(Y = I|D = 1, G = m), I = 0, 1$ (Verma & Rubin 2018, p. 3).

In words: The probability for belonging to the accepting class if having been classified as such and the value of gender is "female" equals the probability for belonging to the accepting class if having been classified as such and the value of gender is "male". The probability for belonging to the rejecting class if having been classified as such and the value of gender is "female" equals the probability for belonging to the rejecting class if having been classified as such and the value of gender is "male".

It is enough to only test for either (1.) or (2.) (Verma & Rubin 2018, p. 3).

Test-fairness

Test-fairness, also called Calibration or Matching Conditional Frequencies, is a condition similar to Predictive Parity. But instead of looking at binary assignments (positive or negative) for the predicted outcome, it looks at a continuous probability score S, assigned by the algorithm (Verma & Rubin 2018, p. 5). For evaluation, score values can be binned, so one would check whether the condition is met for scores between 0 and (excluding) 0.1, between 0.1. and (excluding) 0.2, and so on (Verma & Rubin 2018, p. 5).

What the score is depends on the algorithm tested. A tool that automatically evaluates CVs could, instead of accepting or rejecting a person, assign scores to a person that should say how well a person is suited for a job. Using Test-fairness one could then check whether people with the same score are equally likely to actually be suited for a job, no matter what their gender is.

Mathematically, the Test-fairness condition in the case of the CV tool can be expressed as $P(Y = 1|S = s, G = m) = P(Y = 1|S = s, G = f), 0 \leq s \leq 1$ (Verma & Rubin 2018, p. 5).

In words: The probability for belonging to the accepting class if having been assigned a score s and the value of gender is "female" equals the probability for belonging to the accepting class if having been assigned a score s and the value of gender is "male".

Well-Calibration

To check whether the algorithm is well calibrated in it's assignment of prediction scores S, one can use the metric Well-Calibration (Verma & Rubin 2018, p. 5). It extends the Test-fairness metric in the following way: $P(Y = 1|S = s, G = m) = P(Y = 1|S = s, G = f) = s, 0 \leq s \leq 1$ (Verma & Rubin 2018, p. 5).

In words: The probability for belonging to the accepting class if having been assigned a score s and the value of gender is "female" equals the probability for belonging to the accepting class if having been assigned a score s and the value of gender is "male" and equals s.

In the case of the job application tool one can check for Well-Calibration by checking whether a fraction s of men and a fraction s of women that were accepted for the job are actually a good match for the job.

2.2.4 Predictive Equality

Predictive equality, or False Positive Error Rate Balance, measures whether the probability P of being

1. assigned to the positive class ($I = 1$) (FPR)
2. assigned to the negative class ($I = 0$) (TNR)

if really belonging to the negative class ($Y = 0$) is the same for all distinct values of sensitive features (Verma & Rubin 2018, p. 4).

It could be seen as fair if men and women who are not a good match for the job are equally likely to be accepted anyway and to be equally likely to be rejected correctly.

The condition for the Predictive Equality metric in the case of a CV review tool is $P(D = I|Y = 0, G = f) = P(D = I|Y = 0, G = m), I = 0, 1$ (Verma & Rubin 2018, p. 4).

In words: The probability for being assigned to the accepting class if actually belonging to the rejecting class and the value of gender is "female" equals the probability for being assigned to the accepting class if actually belonging to the rejecting class and the value of gender is "male". The probability for being assigned to the rejecting class if actually belonging to the rejecting class and the value of gender is "female" equals the probability for being assigned to the rejecting class if actually belonging to the rejecting class and the value of gender is "male".

It is enough to only test for either (1.) or (2.) (Verma & Rubin 2018, p. 4).

Balance for Negative Class

The fairness metric Balance for Negative Class measures whether groups of people with distinct sensitive features that really belong to the negative class (rejected) have an equal average predicted score $E(S)$ (Verma & Rubin 2018, p. 5). It is therefore similar to Predictive Equality (1.), except that it is looking at scores instead of assignment to a binary class (positive or negative) (Verma & Rubin 2018, p. 5).

In the case of an application review tool that assigns scores to applicants one can measure whether men and women who are not a good match for the job receive an equal average score.

The mathematical condition in this case is $E(S|Y = 0, G = f) = E(S|Y = 0, G = m)$ (Verma & Rubin 2018, p. 5).

In words: The average assigned score for women who actually belong to the rejecting class equals the average assigned score for men who actually belong to the rejecting class.

2.2.5 Equal Opportunity

Equal Opportunity, or False Negative Error Rate Balance, measures whether the probability P of being assigned to the

1. negative class ($D = I = 0$) (FNR)
2. positive class ($D = I = 1$) (TPR)

if actually belonging to the positive class ($Y = 1$) is the same for all distinct values of sensitive features (Verma & Rubin 2018, p. 4).

In the case of automatically reviewed job applications one could find out whether men and women who are a good match for the job are equally likely to be rejected anyway and to be accepted correctly.

The condition for equal opportunity in that case is $P(D = I|Y = 1, G = f) = P(d = I|Y = 1, G = m), I = 0, 1$ (Verma & Rubin 2018, p. 4).

In words: The probability for being assigned to the rejecting class if actually belonging to the accepting class and the value of gender is "female" equals the probability for being assigned to the rejecting class if actually belonging to the accepting class and the value of gender is "male". The probability for being assigned to the accepting class if actually belonging to the accepting class and the value of gender is "female" equals the probability for being assigned to the accepting class if actually belonging to the accepting class and the value of gender is "male".

It is enough to only test for either (1.) or (2.) (Verma & Rubin 2018, p. 4).

Balance for Positive Class

Balance for Positive Class is a metric similar to Equal Opportunity (2.), but it is looking at a continuous score S instead of an assignment to a binary class (Verma & Rubin 2018, p. 5). It measures whether groups of people with distinct sensitive features that really belong to the positive class (accepted) have an equal average predicted score $E(S)$ (Verma & Rubin 2018, p. 5).

Using Balance for Positive Class in the case of a tool that assigns scores to applicants one could verify that women and men who are well suited for the job receive an equal average score.

The mathematical condition for Balance for Positive Class in that case is $E(S|Y = 1, G = m) = E(S|Y = 1, G = f)$ (Verma & Rubin 2018, p. 5).

In words: The average assigned score for women who actually belong to the accepting class equals the average assigned score for men who actually belong to the accepting class.

2.2.6 Equalized Odds

Equalized Odds is also called Disparate Mistreatment metric or Conditional Procedure Accuracy Equality (Verma & Rubin 2018, p. 4). It measures whether the probability P of being assigned to the positive class ($D = 1$) if belonging to the

1. positive class ($I = 1$) (TPR)
2. negative class ($I = 0$) (FPR)

is the same for all distinct values of sensitive features (Verma & Rubin 2018, p. 4).

Using this metric one could measure whether the algorithm correctly and incorrectly accepts men and women equally.

The condition for Equalized Odds in the example of the CV tool is $P(D = 1|Y = I, G = f) = P(D = 1|Y = I, G = m), I = 0, 1$ (Verma & Rubin 2018, p. 4).

In words: The probability for being assigned to the accepting class if actually belonging to the accepting class and the value of gender is "female" equals the probability for being assigned to the accepting class if actually belonging to the accepting class and the value of gender is "male". The probability for being assigned to the accepting class if actually belonging to the rejecting class and the value of gender is "female" equals the probability for being assigned to the accepting class if actually belonging to the rejecting class and the value of gender is "male".

Equalized Odds can only be fullfilled if $P(Y = 1|G = f) = P(Y = I|G = m)$ is fullfilled (Verma & Rubin 2018, p. 4). Looking at the CV review tool example this means that being a good match for a job needs to be independent from gender in order for Equalized Odds to be fullfillable.

2.2.7 Conditional Use Accuracy Equality

Conditional Use Accuracy Equality measures whether the probability P of belonging to the

1. postive class ($Y = 1$) (PPV)
2. negative class ($Y = 0$) (NPV)

if having been assigned to the positive class ($D = 1$) is the same for all distinct values of sensitive features (Verma & Rubin 2018, p. 4).

Using the Conditional Use Accuracy Equality metric in the case of the job application review tool one can verify whether men and women who fit the job are equally likely to be accepted and men and women who are not a good match are equally likely to be rejected.

The conditions for Conditional Use Accuracy Equality in this case are $P(Y = 1|D = 1, G = f) = P(Y = 1|d = 1, G = m) AND P(Y = 0|D = 0, G = f) = P(Y = 0|D = 0, G = m)$ (Verma & Rubin 2018, p. 4).

In words: The probability for belonging to the accepting class if having been assigned to the accepting class and the value of gender is "female" equals the probability for belonging to the accepting class if having been assigned to the accepting class and the value of gender is "male" and the probability for belonging to the rejecting class if having been assigned to the rejecting class and the value of gender is "female" equals the probability for belonging to the rejecting class if having been assigned to the rejecting class and the value of gender is "male".

2.2.8 Overall Accuracy Equality

The metric Overall Accuracy Equality is used to verify that the accuracy rate (TPR and TNR) is equal for all distinct values of sensitive features (Verma & Rubin 2018, p. 4).

In the case of a job application review tool Overall Accuracy Equality verifies that men and women are equally likely to be accepted or rejected correctly.

The mathematical condition for Overall Accuracy in that case is $P(D = Y|G = f) = P(D = Y|G = m)$ (Verma & Rubin 2018, p. 4).

In words: The probability for being assigned the same class as actually belonging to if the value for gender is "female" equals The probability for being assigned the same class as actually belonging to if the value for gender is "male".

2.2.9 Treatment Equality

This metric is used for verifying that the ratio of incorrectly rejected (FN) and incorrectly accepted (FP) cases is the same for all distinct values of sensitive features (Verma & Rubin 2018, p. 4).

Treatment Equality in the example of the job application review tool can be mathematically expressed as $FN(f)/FP(f) = FN(m)/FP(m)$ (Verma & Rubin 2018, p. 5).

In words: The ratio of incorrect assignments to the negative class if the value for gender is "female" and incorrect assignments to the positive class if the value for gender is "female" equals the ratio of incorrect assignments to the negative class if the value for gender is "male" and incorrect assignments to the positive class if the value for gender is "male".

2.2.10 Measuring Individual Fairness

Causal Discrimination

When measuring individual fairness, one strategy is to first create an additional set of predictions from the training set where sensitive features or their proxies are exchanged (eg. "male" becomes "female") (Verma & Rubin 2018, p. 6). In the second step, the predictions are compared (Verma & Rubin 2018, p. 6). If the algorithm is individually fair, the predictions have not changed for the individual data points.

Fairness through awareness

Another strategy is to check whether similar individuals are in similar classes (Verma & Rubin 2018, p. 6). Similarity between individuals can be measured by calculating the length of distance vectors between individuals (Verma & Rubin 2018, p. 6). The distance vectors are made up of chosen features (Verma & Rubin 2018, p. 6). The vector lengths are then compared with the corresponding distances between outcomes (Verma & Rubin 2018, p. 6).

In the case of a job application review tool that classifies individuals with a score one could define distance vectors between individuals to contain the features "highest degree" and "years of work experience". One would have to define a mapping of words to numbers for the feature "highest degree", eg. "Bachelors Degree" -> 1, "Masters Degree" -> 2, "PhD" -> 3 (Verma & Rubin 2018, p. 6). Now one can calculate the length of the distance vectors between integer values of features. For buckets of distances one then calculates the average prediction distance, the distance between scores (Verma & Rubin 2018, p. 6). One can then compare growth rates: Does the prediction distance grow the same way the feature distance does? One can also count how often the prediction distance is greater than the mapped feature distance.

2.3 Sources of Unfairness

Unfairness in the form of biases can evolve at different stages of a data scientific development journey, and these different kinds of bias are not mutually exclusive (*Bias and Discrimination in AI* 2020). The information about sources of unfairness given in (*Bias and Discrimination in AI* 2020) is extended mostly with information from Suresh and Guttag who summarize different forms of bias in their article "A Framework for Understanding Unintended Consequences of Machine Learning" (Suresh & Guttag 2020).

2.3.1 Pre-processing

Stereotypes can be encoded in data that is used for training in the form of proxies (*Bias and Discrimination in AI* 2020). They develop before data was even gathered, when data is gathered or when data is prepared for further training (*Bias and Discrimination in AI* 2020).

Bias can exist before gathering data, in the form of historical bias (*Bias and Discrimination in AI* 2020). With historical bias "there is a misalignment between world as it is and the values or objectives to be encoded" (Suresh & Guttag 2020, p. 2). For instance, training a sports recognition software on images of different sports games from the US might result in the software recognizing football played by black people not as football but as basketball (*Bias and Discrimination in AI* 2020).

When gathering data, one might introduce a measurement bias by defining biased labels in a way that not all necessary options are provided or that the labels or categories actually don't measure what the modeler thinks they would measure (Suresh & Guttag 2020, p. 2). For example defining only two possible values for a class "gender" will lead to data biased towards a binary definition of gender. Or trying to measure how often someone was ill might just measure how likely they are to see a doctor (*Bias and Discrimination in AI* 2020).

The bias introduced in data gathering can go even further: The way data is collected can influence how successfull the collection is with different groups because different groups might have different norms (*Bias and Discrimination in AI* 2020). When data is taken from self-assessment or observation, different people might make different decisions. Even when asked to chose a label people might chose different labels depending how they interpret them (*Bias and Discrimination in AI* 2020). Bias resulting from these two cases can be categorized as behavioral bias (*Bias and Discrimination in AI* 2020).

Another bias resulting from the way that data is collected is the collection bias (*Bias and Discrimination in AI* 2020) or sampling bias (Mehrabi, Morstatter, Saxena, Lerman & Galstyan 2019, p. 6). Some groups might refuse automatic collection or any collection right away or be excluded completely (*Bias and Discrimination in AI* 2020), for example if they don't use certain social media.

Since the way data is collected can influence the results, this effect grows when working with data collected over a large time span. For example people could chose different labels, depending on when the data was collected (*Bias and Discrimination in AI* 2020). This is called a temporal bias (Mehrabi et al. 2019, p. 6).

Finally, the bias existant in the gathered data due to temporal bias or collection bias is the representation bias (Suresh & Guttag 2020, p. 5). It exists if there is a demographic difference between the training and target population (Suresh & Guttag 2020, p. 2). For example, face recognition software might be trained on a data set of mostly faces of white people but when later applied at - say - a train station, it might fail to recognize faces of black people.

2.3.2 In-Processing

The issue of another bias, aggregation bias, lays in hiding different results for subgroups by grouping them with others and thus making individual subgroups invisible (Suresh & Guttag 2020, p. 2). For example if non-frequent ethnicities are aggregated as "other", discrimination against a specific subgroup can go unnoticed. This effect is also called Simpson's Paradox (Mehrabi et al. 2019, p. 5).

2.3.3 Post-processing

When using benchmarks to validate an algorithm, evaluation bias can occur if the benchmarks used are not representative for the whole population (Suresh & Guttag 2020, p. 6). If algorithms are modeled after this benchmark data, algorithms become biased (Suresh & Guttag 2020, p. 6). Another problem in evaluation can arise if only one metric is considered, for example the accuracy for all groups together, while other or more detailed metrics might bring more insight (Suresh & Guttag 2020, p. 6).

The way an algorithm is used can also lead to biases. If the algorithm is used to make suggestions, people who should really verify those suggestions tend to accept them as is, thanks to the confirmation bias (Suresh & Guttag 2020, p. 6). Or, verification by humans is entirely left out right away (Suresh & Guttag 2020, p. 6).

Further discrimination can occur when, following the measurement bias introduced when gathering data, proxies are not recognized as such (*Bias and Discrimination in AI* 2020). For example, one might not notice that the location based prize calculation discriminates against black people because location sometimes works as a proxy for a protected feature (*Bias* *and Discrimination in AI* 2020). If no data on protected features have been gathered, such discrimination can go unnoticed (*Bias and Discrimination in AI* 2020). This is also why the fairness strategy "Fairness through unawareness" is criticized (*Bias and Discrimination in AI* 2020).

Another major problem occurring after deployment are so called feedback loops (*Bias and Discrimination in AI* 2020). They are created when the model learns from data it produces and has pre-existing biases (*Bias and Discrimination in AI* 2020). For example, a model might rate people's CVs based on who is already working in a company. This leads to the company employing more people of the same groups, and so on [1]. Especially if an algorithm or their developer has no way to recognize false decisions, errors go unnoticed and can not be taken as a chance for improvement (*Bias and Discrimination in AI* 2020).

2.4 Implementing Fairness

Fairness can be implemented at different stages of the development process of an algorithm (*Bias and Discrimination in AI* 2020).

2.4.1 Pre-processing

Before starting the development process, a company or team should define ethical guidelines to follow in their work (*Bias and Discrimination in AI* 2020): What are fairness goals to be achieved? How will I use data, considering legal regulations like the GDPR? When doing so, thinking about atypical users and misuse is unavoidable: Who is this algorithm not for? How could it be misused?

When chosing data to work with one needs to make the task related decision between representational data sets and overdiverse data sets (*Bias and Discrimination in AI* 2020). In a representational data set the demographic distribution of training and target data are the same (*Bias and Discrimination in AI* 2020). In overdiverse data sets, there is an equal amount of data for each different group (*Bias and Discrimination in AI* 2020). If the given data does not contain sensitive information, it can be hard to check what type of data set is given. Some tools, like "NamSor", are specialized in inferring sensitive attributes like gender or ethnicity.

[1] The example named is based on the real life example of Amazon's recruiting tool discriminating women. The interested reader finds more insight at `https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G`.

2.4.2 In-processing

When using data for Machine Learning tasks, one can decide to optimize not for a common loss function, but for one combined with a fairness metric (*Bias and Discrimination in AI* 2020). A strategy in the training of neural networks is the so called "Domain-Adversarial Training of Neural Networks" (DANN) (*Bias and Discrimination in AI* 2020). Here, an adversarial neuron is trained to guess the values of sensitive features that should not influence the result and thus not be inferable (*Bias and Discrimination in AI* 2020).

2.4.3 Post-processing

Towards the end of the development process a way to receive consumer feedback needs to be established (*Bias and Discrimination in AI* 2020). It is important that complaints about unfair decisions made by the algorithm can be received so that the algorithm can be improved (*Bias and Discrimination in AI* 2020). One also needs to be transparent about the algorithm (*Bias and Discrimination in AI* 2020): One should be clear about the data used for training, about which factors play into account for the result and one should name any validity constraints or fairness concerns. Finally, one should watch the algorithm as it is put to use (*Bias and Discrimination in AI* 2020): Who uses it how? How can harm caused by the algorithm be identified?

3 Preparation

This chapter explains technology and data used for the fairness audit of "NamSor"'s classification algorithms. Technologies used will be presented first, then follows a description of the data set and finally the applied methodology is explained.

3.1 The "NamSor" API

"NamSor" is an API that classifies names by different categories. In other words: It infers categories from names. In this work the categories - or protected attributes - gender and ethnicity will be audited for fairness.

The Python SDK of "NamSor" 2.0.9 [1] was used.

3.1.1 Motivation

If "NamSor"'s fairness can be validated, "NamSor"'s gender and ethnicity inference API can be a tool that supports fairness audits by making protected attributes (of which decision makers or tools are unaware) visible. The API can thereby help to find discrimination. For example in HR an applicant does not give information about their gender or ethnicity. However, that information might still be encoded in their resume and thus lead to discrimination by an automatic application review tool or a recruiter. "NamSor" combined with a fairness audit could be used in HR to verify whether applicants gathered in the first step of the recruiting process are diverse and thus the HR's job openings and ads target a diverse group of people. "NamSor" could then be used after each step of the hiring process to help check whether the remaining group of applicants is (still) diverse. It could also be used to help find out whether the finally chosen candidates contribute to the overall diversity of the workforce already employed.

[1] `https://github.com/namsor/namsor-tools-v2/releases/tag/namsor-java-tools_v2.0.9`

3.1.2 Use

To classify names by gender the API endpoint "genderGeoBatch" [2] has been selected. It takes 1000 names composed of an id, a first name, a last name and the ISO2 code of a country, "US" in the case of the "COMPAS" data set. The outputs used in this work are a guess for the gender category the name could fall into ("likelyGender") and the probability for the guess being correct ("probabilityCalibrated"). "NamSor" only differs between two categories for gender, "Male" and "Female". "NamSor" explicitly sais their API infers gender, not sex, acknowledging that gender identities other than male and female can not be inferred by their API (*NamSor Tools V2* 2020).

To classify names by ethnicity the API endpoint "usRaceEthnicityBatch" [3] has been chosen. It specializes on names from people located in the US. The input is the same as for "GenderGeoBatch". In this work, the returned guess for an ethnicity category the name belongs to ("raceEthnicity") and the probability for the guess being correct ("probabilityCalibrated") are used. The predictions for ethnicity can be "Black but not Latino", "White but not Latino", "Hispano Latino" and "Asian", following the US Census Taxonomy (*NamSor Tools V2* 2020). The US Census Taxonomy makes a difference between ethnicity and race: Ethnicity is either "Hispanic" or "not Hispanic" while race can be "White, Black or African-American, Asian" and other races(*Race and Ethnicity* 2017). "NamSor" thus counts every person of Hispanic ethnicity as such and divides everyone who is not Hispanic up into White, Black or Asian.

"NamSor" uses a Naive Bayes classification algorithm to calculate the most probable gender and ethnicity category for a name [4].

The basic idea of using Naive Bayes to classify names can be inferred from the implementation. The key is to treat each name like a group of features. Using training data, "NamSor" registers for each category the observed features and how often they occured. If an unknown name is to be classified, "NamSor" calculates the conditional probabilities for the name belonging to a certain group under the condition of containing the features it is made up of. Then the category with the greater conditional probability is selected.

Let's look at the idea in more detail: "NamSor" should classify a name n into one of the categories C. From n, the features F are extracted. These features can be the first name, the last name, the ending of the first name and more (Carsenat 2019, p. 5). Now for each of the features

[2] `https://v2.namsor.com/NamSorAPIv2/apidoc.html/personal/genderGeoBatch`

[3] `https://v2.namsor.com/NamSorAPIv2/apidoc.html/personal/usRaceEthnicityBatch`

[4] The implementation can be found at `https://github.com/namsor/Java-Naive-Bayes-Classifier-JNBC/tree/68603ae5d5d44d00c26729cb5cc6a64a5d68d566`.

the relative amount of counted f in a category c is calculated, or mathematically speaking: $P(f|c)$. Then, for each category, the conditional probability for the name belonging to c under the condition of containing features f, or, mathematically speaking $P(c|N)$, is calculated. This is done for each category c by multiplying the product of all $P(f|c)$ with the relative amount of features per category $P(c)$. The category with the highest value for $P(c|N)$ is returned as well as $P(c|N)$, which is the "probabilityCalibrated".

3.1.3 Current Method Applied by "NamSor" for Checking for Bias

Currently, for testing whether their API endpoints are biased, "NamSor" tests them using open source data such as WikiData (Carsenat 2020, p. 2). Precision and recall are calculated for each of the groups identified by "NamSor", such as country (Carsenat 2020, p. 4). Then, the algorithms are calibrated so that each group has approximately the same precision and recall as the other groups (Carsenat 2020, p. 7).

3.2 The "Aequitas" Tool

Created by Pedro Saleiro et al. in 2018 (Saleiro, Kuester, Stevens, Anisfeld, Hinkson, London & Ghani 2018, p. 2), "Aequitas" is a tool that evaluates fairness of algorithms' results in relation to protected classes (Saleiro et al. 2018, p. 4). "Aequitas" is one of many tools that provide analysis of algorithmic bias (Bellamy, Dey, Hind, Hoffman, Houde, Kannan, Lohia, Martino, Mehta, Mojsilovic, Nagar, Ramamurthy, Richards, Saha, Sattigeri, Singh, Varshney & Zhang 2018, p. 2). "Aequitas" has been chosen for this project because its functions appear to be straight forward and "Aequitas" provides an elaborate example on how it can be used[5]. "Aequitas" can be used with Python, via the command line (Saleiro et al. 2018, p. 4) or via their website[6]. In this project, the "Aequitas" 38.1 library was used with Python 3.7.3 in a Jupyter Notebook.

3.2.1 Data preparation for "Aequitas"

To be able to use "Aequitas" in a meaningful way, one has to prepare a data set fitting the requirements:

[5] See (*COMPAS Analysis* 2017)

[6] `http://aequitas.dssg.io/`

- There needs to be at least one column that contains a protected attribute, such as "gender" (*Understanding Input Data* 2018). Groups are defined by having distinct values for protected features.
- There needs to be one column providing a "score" between 0 and 1, which corresponds to a prediction made by an algorithm (*Understanding Input Data* 2018). If the value for "score" is continuous, one needs to chose a treshold that sorts contunuous scores into one of two values, 0 or 1 (*Understanding Input Data* 2018).
- There needs to be one column named "label_value" which corresponds to the actual result (*Understanding Input Data* 2018).
- One group is to be chosen as a reference group. This could be the group that have been favored historically, such as "male". It could also be determined by selecting the biggest or smallest group in the data set or for instance by selecting the group with highest precision (Saleiro et al. 2018, p. 7). Other selection processes are also possible.

3.2.2 What "Aequitas" does measure

"Aequitas" measures absolute and relative distribution metrics and absolute and relative error metrics. As a reminder, with the exception of Predicted Positive Rate (PPR) they are all defined in (Verma & Rubin 2018, p. 2f).

PPR is "the fraction of the entities predicted as positive that belong to a certain group" (*Understanding the Metrics* 2018). It can be written as $P(G = g|D = 1)$ where P means Probability, $D = 1$ means predicted as positive and $G = g$ means belonging to group g (*Understanding the Metrics* 2018).

Absolute Distribution Metrics

"Aequitas" can measure the following relative distribution metrics per group (Saleiro et al. 2018, p. 6f). The abbreviation used in "Aequitas" is given in brackets.

- Positives (P, "group_label_pos")
- Negatives (N, "group_label_neg")
- Predicted Positives (PP)
- Predicted Negatives (PN)

Relative Distribution Metrics

"Aequitas" can measure the following relative distribution metrics per group (Saleiro et al. 2018, p. 6f):

- Prevalence (Prev)
- Predicted Prevalence (PPrev)
- Predicted Positive Rate (PPR)

Absolute Error Metrics

The following absolute error metrics, also called confusion matrix, can be calculated by "Aequitas" (Saleiro et al. 2018, p. 6f):

- False Positives (FP)
- True Positives (TP)
- False Negatives (FN)
- True Negatives (TN)

Relative Error Metrics

"Aequitas" can calculate the following relative error metrics per group (Saleiro et al. 2018, p. 7):

- True Positive Rate (TPR)
- True Negative Rate (TNR)
- False Discovery Rate (FDR)
- False Omission Rate (FOR)
- False Positive Rate (FPR)
- False Negative Rate (FNR)
- Negative Predictive Value (NPV)
- Positive Predictive Value / Precision (PPV)

Fairness Metrics

"Aequitas" defines and calculates fairness metrics that are partly similar, partly different from those summarized in (Verma & Rubin 2018). For it's fairness metrics, "Aequitas" returns disparity ratios instead of boolean values that say whether parity is given or not (Saleiro et al. 2018, p. 7), therefore when using "Aequitas", one has to evaluate the resulting ratios using a fairness threshold or one can use "Aequitas"' default fairness threshold of 0.8 (*COMPAS Analysis* 2017, p. 9). A fairness threshold of 0.8 means that a disparity ratio (eg. $FPR(somegroup)/FPR(referencegroup)$) above 0.8 or below 1.2 is considered to be fair.

"Aequitas" provides the following group fairness metrics (*COMPAS Analysis* 2017, p. 9). "Aequitas" calculates...

- ... *Statistical Parity* (PPR Parity), while (Verma & Rubin 2018) define it as PPrev Parity ("Aequitas" calls Statistical Parity as defined by (Verma & Rubin 2018) "Impact Parity")
- ... *Equalized Odds* (both TPR Parity and FPR Parity) as defined in (Verma & Rubin 2018)
- ... both conditions for *Predictive Parity* seperately, PPV Parity (which it calls Precision Parity) and FDR Parity, while it is enough to calculate only one to check for Predictive Parity, according to (Verma & Rubin 2018, p. 3)
- ... both conditions for *Predictive Equality* separately, FPR Parity and TNR Parity, while it is enough to calculate only one to check for Predictive Equality (Verma & Rubin 2018, p. 4).
- ... both conditions for *Equal Opportunity*, TPR Parity and FNR Parity, while it is enough to test only one (Verma & Rubin 2018, p. 4)
- ... both conditions for *Overall Accuracy Equality*, TPR and TNR Parity, but without combining them into one logical-and statement
- ... both conditions for *Conditional Use Accuracy Equality*, NPV Parity and PPV Parity, but without combining them into one logical-and statement

Additionally to the metrics mentioned in (Verma & Rubin 2018, p. 4), "Aequitas" also calculates the following metrics:

- *Impact Parity*, or PPrev Parity
- *Unsupervised Fairness*, which is fullfilled if Statistical Parity and Impact Parity are fullfilled

- FOR Parity
- *Type I* Parity, which is fullfilled if there is FDR Parity and FPR Parity
- *Type II* Parity, which is fullfilled if there is FNR Parity and FOR Parity
- *Supervised Fairness*, which is fullfilled if there is Type I and Type II Parity

3.2.3 What "Aequitas" does not measure

"Aequitas" does not directly measure Treatment Equality. Test-Fairness, Well-Calibration, Balance for Positive Class and Balance for Negative Class that use a score without threshold are not measured by "Aequitas" either.

3.2.4 How "Aequitas" calculates the significants of its results

"Aequitas" can supposedly calculate the significance of the disparity metrics. It does so by calculating whether the reference group and the compared group have significantly different means. From "Aequitas"' source code (*Aequitas Source Code* 2019), the following steps have been summarized:

1. The significance for each disparity metric is calculated either on the basis of false positives (PPV, FPR, TNR, FDR) or on the basis of false negatives (NPV, FNR, TPR, FOR) or on the basis of the score (PPrev, PPR).
2. For each part of the data that belongs to the reference or a compared group, each entry is determined as belonging to the significance basis (being a false positive or false negative) or not. This determination is binary encoded except when the raw score values are used. This results in multiple lists of values, one for each group.
3. For each list is determined whether its values are normally distributed.
4. It is determined whether the variances of each compared group and the reference group are significantly different.
    - If either the reference list or the compared list are not normally distributed, the variances of both lists are compared using 's test implemented in the "SciPy" library for Python, using the median as a center [7].

[7] The documentation of Levene's test method of "SciPy" can be found at `https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.levene.html`

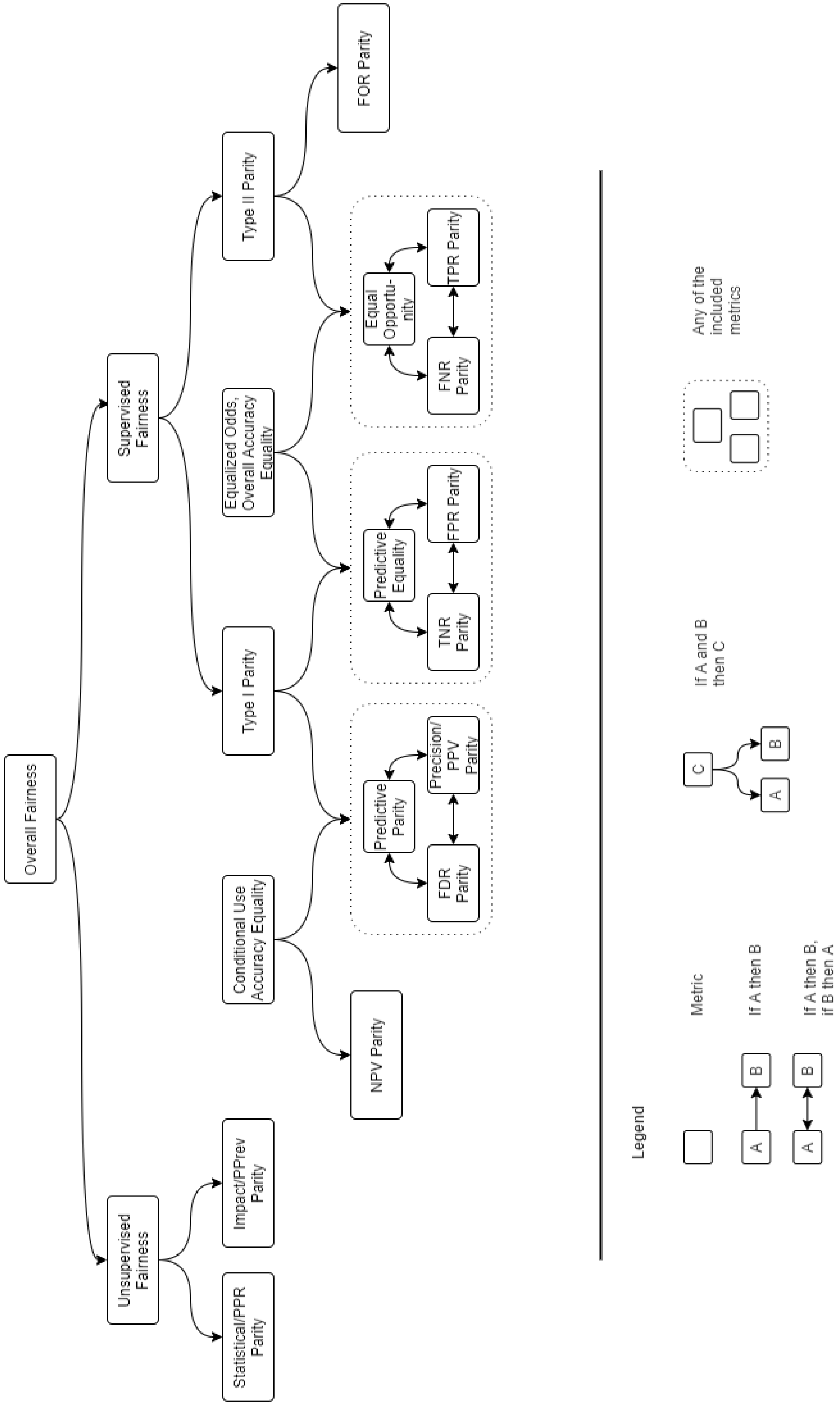


Figure 3.1: The relationships between the group fairness metrics as measured by "Aequitas".

- If both lists are normally distributed, the variances of both lists are compared using Bartlett's test implemented in the "SciPy" library [8]

5. It is determined whether the means of each compared group and the reference group are significantly different using "SciP"y's t-test method.
   - If the variances of reference group and compared group are not significantly different, a 2 sample t-test is performed (*SciPy v1.5.2 Reference Guide* 2020).
   - If the variances are significantly different, Welch's t-test is executed (*SciPy v1.5.2 Reference Guide* 2020).

The p-value returned by the t-test method is returned. Based on that p-value one can now judge whether a disparity metric is significant or not, depending on the chosen significance alpha.

3.3 The "COMPAS" Data Set

The personal data used in this project originally comes from a data set compiled by "ProPublica" with data from the instrument "Correctional Offender Management Profiling for Alternative Sanctions", or "COMPAS" for short [9]. In this case the instrument "COMPAS" was used to obtain predictions of how likely pre-trial defendents in Broward County, Florida, are to be re-arrested withing two years (Flores et al. 2016, p. 7). This data set contains 7214 recordings.

"ProPublica" originally used their version of the "COMPAS" data set to analyze whether the tool "COMPAS" made biased predictions (Larson, Mattu, Kirchner & Angwin 2016a). This analysis was later redone by "Aequitas" as an example for how to use the "Aequitas" library for Python (*COMPAS Analysis* 2017). In this project the data will be used to analyze whether "NamSor" performs equally well by ethnicity and gender.

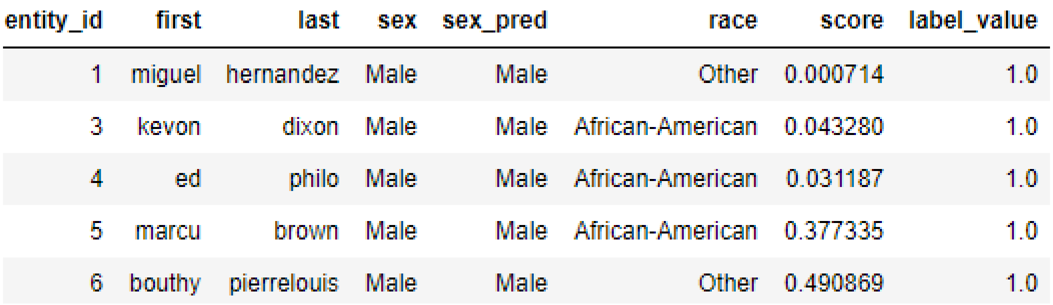
Data interesting for this research has been obtained from (Larson et al. 2016a) using a Python script taken from "Aequitas"[10]. The original script was modified here so it keeps first and last names[11].

[8] The documentation of Bartlett's test method of "SciPy" can be found athttps://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.bartlett.html.

[9] See also (*compas-scores-two-years* 2016)

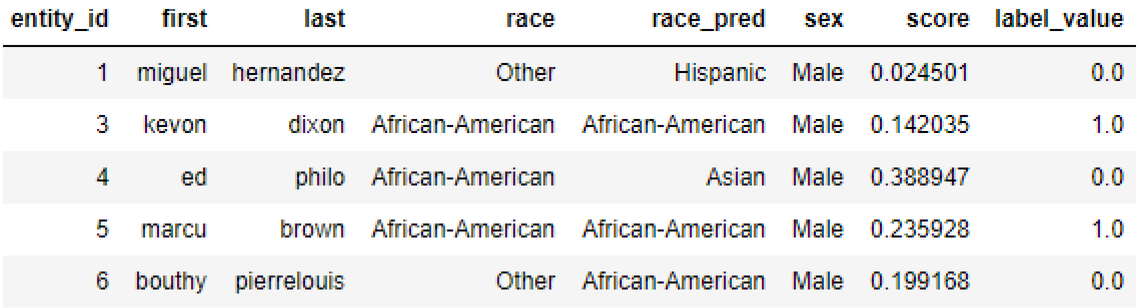
[10] "Aequitas"' original script can be found at https://github.com/dssg/aequitas/blob/master/examples/compas_data_for_aequitas.py

[11] The new script can be found at https://github.com/namsor/namsor-compass/blob/master/01-Data-Preparation-01-Data-Formatting.ipynb



entity_id	first	last	sex	sex_pred	race	score	label_value
1	miguel	hernandez	Male	Male	Other	0.000714	1.0
3	kevon	dixon	Male	Male	African-American	0.043280	1.0
4	ed	philo	Male	Male	African-American	0.031187	1.0
5	marcu	brown	Male	Male	African-American	0.377335	1.0
6	bouthy	pierrelouis	Male	Male	Other	0.490869	1.0

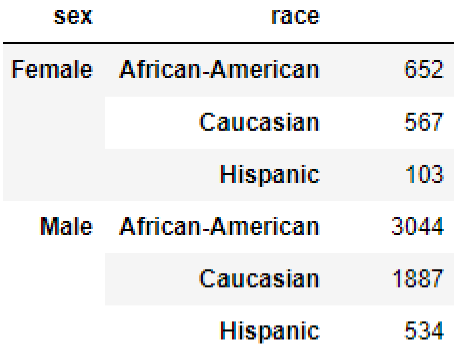
Figure 3.2: Extract from the data set prepared to analyze the "NamSor"'s gender predictions



entity_id	first	last	race	race_pred	sex	score	label_value
1	miguel	hernandez	Other	Hispanic	Male	0.024501	0.0
3	kevon	dixon	African-American	African-American	Male	0.142035	1.0
4	ed	philo	African-American	Asian	Male	0.388947	0.0
5	marcu	brown	African-American	African-American	Male	0.235928	1.0
6	bouthy	pierrelouis	Other	African-American	Male	0.199168	0.0

Figure 3.3: Extract from the data set prepared to analyze the "NamSor"'s ethnicity predictions

The data extracted from "ProPublica"'s data set was annotated with "NamSor"'s predictions for gender and ethnicity and with updates to the columns describing prediction and actual outcome. Finally, each recording contains a person's first and last name, their sex and race according to their own statement (equivalent to the "COMPAS" data set provided by "Aequitas") and their gender and ethnicity as predicted by "NamSor". From there, two separate data sets (one for each of the two different API endpoints that are being analyzed here) are derived, each with two columns "score" and "label_value". "Aequitas"' "score" almost corresponds to the "probabilityCalibrated" values returned by "NamSor". Correspondence can be achieved by setting the score to $1 - probabilityCalibrated$. This way we get the probability for the prediction being incorrect. The minimum "probabilityCalibrated" returned by "NamSor" for the gender prediction being correct is just just above 0.5 while the maximum is 1. The minimum probability for the race prediction being correct is (rounded) 0.35 while the maximum is (rounded) 0.98. The "label_value" encodes whether "NamSor's" prediction made for gender or ethnicity was actually correct (1) or not (0).



sex	race	
Female	African-American	652
	Caucasian	567
	Hispanic	103
Male	African-American	3044
	Caucasian	1887
	Hispanic	534

Figure 3.4: How many people of each ethnicity category are there per gender category?

3.3.1 Label taxonomy mapping between COMPAS and NamSor

The original "COMPAS" data set recognizes the categories "African-American", "Asian", "Caucasian", "Hispanic" and "Native American" as races and summarizes all other possible categories as "Other". This classification used by "ProPublica" was taken from the Broward County Sheriff's Office, but no further explanation on the taxonomy is given (Larson, Mattu, Kirchner & Angwin 2016b).

"COMPAS"' race and "NamSor"'s ethnicity are mapped in the following way:

"NamSor"	"COMPAS"
Asian	Asian
Black but not Latino	African-American
Hispano Latino	Hispanic
White but not Latino	Caucasian

No ethnicity category from "NamSor" could be mapped to Native American or Other. These groups are therefore removed. The data set contains 3696 recordings for African-Americans, 2454 recordings for Caucasians, 637 for Hispanics and 32 for Asians. "Asian" appears to be too small a group, which is why it is also removed. This leaves 6787 recordings.

Like "NamSor", the "COMPAS" data set only knows binary gender. It is unclear whether prisoners entered their biological sex or their gender identity. In the data set, there are 1322 recordings of women and 5465 of men. The table of amount of recordings by race and gender shows, that the amount of Hispanic women is also rather small.

3.4 Methodology

First, the distribution of groups in the "COMPAS" data set was explored. Because of the few amount of Asian people in the data set, fairness relative to Asians will not be studied in this work. Neither can the fairness relative to those races not fitting any "NamSor" ethnicity be studied. As mentioned before, these groups are therefore removed from the data set. It is also to note that the three remaining ethnicities and the two available genders are not equally distributed, again: see the previous sections.

The data set was then annotated with "NamSor"'s results as described in the previous chapter, while also deciding on how to map label taxonomies between "NamSor" and "COMPAS". Original information and predictions made by "NamSor" where compared and their equality encoded in "0" (incorrect prediction) and "1" (correct prediction).

For the score threshold per API endpoint the maximum available score was chosen. That means that no predictions for the classification being incorrect are counted - it is always assumed that the classification is correct, no matter the score.

Now, "Aequitas" is used in the following way:

1. A list of group defining attributes is defined: Gender, ethnicity and both combined. When auditing fairness it is important to look at combinations of groups. It is only when we split the group "Hispanic" up into "Female Hispanic" and "Male Hispanic" that we see that there are only 103 Hispanic women and 534 Hispanic men. That means we can not generalize fairness towards the group "Hispanic" because that, in this case, means in general fairness towards Hispanic men.

2. A table $xtabs$ of calculations of the confusion matrix (eg. amount of false positives) and group error metrics (eg. FPR) is retrieved by using $Group().get\_crosstabs()$, passing the data set, the list of group defining attributes and the threshold as arguments.

3. The eference groups are defined to be "Male" for gender and "Caucasian" for ethnicity (and then of course "Male Caucasian" for both combined).

4. A table bdf of the calculated relative error metric disparities between groups and the reference groups is retrieved by calling $Bias().get\_disparity\_predefined\_groups()$, passing $xtabs$, the original data set and the reference groups as arguments.

5. A table fdf of the evaluated fairness metrics is retrieved by calling $Fairness().get\_group\_value\_fairness()$, passing bdf as a parameter.

6. A table gaf of the evaluated higher level fairness metrics Unsupervised Fairness, Supervised Fairness and Overall Fairness is retrieved

To analyze one starts at table gaf. If Overall Fairness is given, everything is fair, even the data set is representative. If not, one can continue down the tree presented in figure 3.1: Is Unsupervised Fairness given? If not, check the table fdf. Is Impact Parity given? If not, check the table bdf to find out how big the disparities between groups are exactly. And so on. It is important to note that Statistical Parity - and cascading also Unsupervised Fairness and Overall Fairness - can be ignored in this case since Statistical Parity is highly dependent from the amount of members of a group, and it was already concluded that groups are not equally distributed in the data set [12].

The results will be presented in the following chapter.

[12] This decision should be made in each use case individually. It might be that even though a data set with not equally distributed groups is given, an outcome with equally distributed groups is wanted.

4 Findings

4.1 Audit results

4.1.1 Gender Classification Endpoint

"Aequitas" returns the data presented in the tables 4.1 and 4.2[1].

Conditional Use Accuracy Equality could not be measured due to missing NPV values - since there are no predictions for the negative class. Type II Parity was not measured due to missing FOR values and, cascading, Supervised Fairness could not be measured either. Thanks to the relationships between fairness metrics (shown in figure 3.1) all other fairness metrics could still be measured. As said priorly, Statistical Parity is ignored as well as Unsupervised Fairness, because of the distribution of groups in the "COMPAS" data set. For the record, Statistical Parity and Unsupervised Fairness are not given when using the "COMPAS" data set.

[1] The complete audit for the gender classification endpoint can be found at `https://github.com/namsor/namsor-compass/blob/master/05_Data_Analysis_01_Fairness_Analysis_of_Gender_Endpoint.ipynb`

group	tpr	tnr	for	fdr	fpr	fnr	npv	ppv	ppr	pprev	prev
Female	1.0	0.0	NaN	0.07	1.0	0.0	NaN	0.93	0.19	1.0	0.93
Male	1.0	0.0	NaN	0.06	1.0	0.0	NaN	0.94	0.81	1.0	0.94
African-American	1.0	0.0	NaN	0.08	1.0	0.0	NaN	0.92	0.54	1.0	0.92
Caucasian	1.0	0.0	NaN	0.04	1.0	0.0	NaN	0.96	0.36	1.0	0.96
Hispanic	1.0	0.0	NaN	0.03	1.0	0.0	NaN	0.97	0.09	1.0	0.97
Female African-American	1.0	0.0	NaN	0.10	1.0	0.0	NaN	0.90	0.10	1.0	0.90
Female Caucasian	1.0	0.0	NaN	0.06	1.0	0.0	NaN	0.94	0.08	1.0	0.94
Female Hispanic	1.0	0.0	NaN	0.01	1.0	0.0	NaN	0.99	0.02	1.0	0.99
Male African-American	1.0	0.0	NaN	0.07	1.0	0.0	NaN	0.93	0.45	1.0	0.93
Male Caucasian	1.0	0.0	NaN	0.03	1.0	0.0	NaN	0.97	0.28	1.0	0.97
Male Hispanic	1.0	0.0	NaN	0.03	1.0	0.0	NaN	0.97	0.08	1.0	0.97

Table 4.1: Group metrics of the Gender Endpoint with a Maximum Score Threshold.

Group	ppr	pprev	ppv	fdr	fpr	tpr
Female	0.24	1.0	0.98	1.32	1.0	1.0
Male	1.00	1.0	1.00	1.00	1.0	1.0
African-American	1.51	1.0	0.96	1.94	1.0	1.0
Caucasian	1.00	1.0	1.00	1.00	1.0	1.0
Hispanic	0.26	1.0	1.01	0.65	1.0	1.0
Female African-American	0.35	1.0	0.93	2.81	1.0	1.0
Female Caucasian	0.30	1.0	0.97	1.71	1.0	1.0
Female Hispanic	0.05	1.0	1.03	0.28	1.0	1.0
Male African-American	1.61	1.0	0.96	2.14	1.0	1.0
Male Caucasian	1.00	1.0	1.00	1.00	1.0	1.0
Male Hispanic	0.28	1.0	1.01	0.86	1.0	1.0

Table 4.2: Disparities by Group of the Gender Endpoint with a Maximum Score Threshold.

Taking 0.8 as a fairness threshold, "Aequitas" falsely claims that FDR Parity is not given for any group except Hispanic men, as can be seen in 4.2. However, it is indeed given for every group because Precision Parity is given for every group. Cascading, Type I Parity is also given for every group even though "Aequitas" claims it is not.

Equalized Odds, Predictive Equality, Equal Opportunity, Overall Accuracy Equality and Impact Parity are inherently given if the threshold is set to a maximum, because this way, no negative predictions exist.

We can thus conclude:

- Predictive Parity: Given
- Type I Parity: Given
- Equalized Odds: Inherently Given
- Predictive Equality: Inherently Given
- Equal Opportunity: Inherently Given
- Overall Accuracy Equality: Inherently Given
- Impact Parity (PPrev Parity): Inherently Given
- FOR Parity: Not measurable
- Conditional Use Accuracy Equality: Not measurable
- Type II Parity: Not measurable

group	tpr	tnr	for	fdr	fpr	fnr	npv	ppv	ppr	pprev	prev
Female	1.0	0.0	NaN	0.29	1.0	0.0	NaN	0.71	0.19	1.0	0.71
Male	1.0	0.0	NaN	0.29	1.0	0.0	NaN	0.71	0.81	1.0	0.71
African-American	1.0	0.0	NaN	0.25	1.0	0.0	NaN	0.75	0.54	1.0	0.75
Caucasian	1.0	0.0	NaN	0.41	1.0	0.0	NaN	0.59	0.36	1.0	0.59
Hispanic	1.0	0.0	NaN	0.07	1.0	0.0	NaN	0.93	0.09	1.0	0.93
Female African-American	1.0	0.0	NaN	0.16	1.0	0.0	NaN	0.84	0.10	1.0	0.84
Female Caucasian	1.0	0.0	NaN	0.48	1.0	0.0	NaN	0.52	0.08	1.0	0.52
Female Hispanic	1.0	0.0	NaN	0.12	1.0	0.0	NaN	0.88	0.02	1.0	0.88
Male African-American	1.0	0.0	NaN	0.26	1.0	0.0	NaN	0.74	0.45	1.0	0.74
Male Caucasian	1.0	0.0	NaN	0.40	1.0	0.0	NaN	0.60	0.28	1.0	0.60
Male Hispanic	1.0	0.0	NaN	0.06	1.0	0.0	NaN	0.94	0.08	1.0	0.94

Table 4.3: Group metrics of the Ethnicity Endpoint with a Maximum Score Threshold.

group	ppr	pprev	ppv	fdr	fpr	tpr
Female	0.24	1.0	0.99	1.01	1.0	1.0
Male	1.00	1.0	1.00	1.00	1.0	1.0
African-American	1.51	1.0	1.29	0.59	1.0	1.0
Caucasian	1.00	1.0	1.00	1.00	1.0	1.0
Hispanic	0.26	1.0	1.58	0.17	1.0	1.0
Female African-American	0.35	1.0	1.39	0.41	1.0	1.0
Female Caucasian	0.30	1.0	0.86	1.21	1.0	1.0
Female Hispanic	0.05	1.0	1.46	0.29	1.0	1.0
Male African-American	1.61	1.0	1.22	0.67	1.0	1.0
Male Caucasian	1.00	1.0	1.00	1.00	1.0	1.0
Male Hispanic	0.28	1.0	1.55	0.16	1.0	1.0

Table 4.4: Disparities by Group of the Ethnicity Endpoint with a Maximum Score Threshold.

- Supervised Fairness: Not measurable

All in all, no valid unfairness was found in the results of the gender classifying endpoint of "NamSor" when auditing it using "Aequitas" and chosing a fairness threshold of 0.8 and a maximum score threshold.

4.1.2 Ethnicity Classification Endpoint

"Aequitas" returns the data presented in the tables 4.3 and 4.4[2].

[2] The complete audit for the ethnicity classification endpoint can be found at `https://github.com/namsor/namsor-compass/blob/master/05_Data_Analysis_02_Fairness_Analysis_of_Ethnicity_Endpoint.ipynb`

Like with the gender endpoint, some metrics could not be measured because no negative predictions where made. These metrics are: Conditional Use Accuracy Equality, Type II Parity and Supervised Fairness. All other fairness metrics could still be derived based on the relationships between them. Again, Statistical Parity and Unsupervised Fairness are not given when using the "COMPAS" data set.

Precision and thus Predictive Parity and thus Type I Parity to Caucasian men are not given for any group other than Caucasian women as can be seen in table 4.4. Looking closely at the disparities reveals that the API performs better for all groups except Caucasians.

Again, Equalized Odds, Predictive Equality, Equal Opportunity, Overall Accuracy Equality and Impact Parity are inehrently given since the threshold is set to a maximum and therefore no negative predictions exist.

We can thus conclude:

- Equalized Odds: Inherently Given
- Predictive Equality: Inherently Given
- Equal Opportunity: Inherently Given
- Overall Accuracy Equality: Inherently Given
- Impact Parity (PPrev Parity): Inherently Given
- Predictive Parity: Not given
- Type I Parity: Not given
- FOR Parity: Not measurable
- Conditional Use Accuracy Equality: Not measurable
- Type II Parity: Not measurable
- Supervised Fairness: Not measurable

All in all, unfairness against Caucasians seems to exist in the results of the ethnicity classifying endpoint of "NamSor" when auditing it using "Aequitas" and chosing a fairness threshold of 0.8 and a maximum score threshold. The API endpoint returns better results for groups that are not Caucasian.

4.2 Generalized Method for Auditing a NamSor API endpoint

The following general rules can be applied to audit a "NamSor" API endpoint for fairness using "Aequitas".

1. Select a fitting data set. The data set should be a listing of people, containing for each person their first and last names as well as the information that should be tested for, for instance gender or ethnicity.

2. Use "NamSor" to predict the information that should be tested for, and also store the probability of the result being incorrect ("NamSor" returns the probability for being correct).

3. Decide on how the original information maps to the prediction values. Do the given information and the prediction define the same categories? How reliable is the given information?

4. Compare original information and predictions. If the prediction was correct, note that in a table column "label_value" as "1". Otherwise put a "0".

5. Chose a treshold for the score for when should be assumed that the prediction is correct.

6. Chose attributes that define a group, eg. gender or ethnicity.

7. Select a reference group, eg. men.

8. Use "Aequitas" to calculate disparities and fairness metrics.

9. Analyze starting at the most restrictive fairness definition and continue investigating whenever a fairness is not given. The fairness metric overview diagram in figure 3.1 can be used; the uppermost metric is the most restrictive one. Think about which metrics make sense to include.

5 Discussion

The following choices that define limitations and implications of the findings presented in the previous chapter will be discussed:

- Choice of protected features and their options
- Choice of reference group
- Choice of score threshold
- Choice of fairness threshold
- Choice of fairness metrics
- Choice of relative disparity (instead of absolute difference)

Finally, the significance of the findings is briefly discussed.

5.1 Choice of protected features and their available options

Race and ethnicity have been chosen as the protected features to be analyzed. However, more protected features could be analyzed in the future, such as age.

The options (or categories) for race and ethnicity that make the groups that are present in "COMPAS" and "NamSor" are greatly simplified. If taking the "126 distinct racial/ethnic combinations" (Schilling 2002, p. 29) from the US census of 2000 as a reference, those are being merged down into very few categories and in the case of "COMPAS" in the aggregate group "Other". "NamSor" does not allow for an "other" categorie and sorts all names into one of four categories. Hispanics of White, Black and other races are merged into one ethnicity, making it impossible to find out whether black Hispanics are treated differently that white Hispanics. This is in so far relevant that according to numbers given in (Schilling 2002, p. 30) based on the US Census of 2000, 13.4% of the US population identify as Hispanic, but less than half of them also identify as White. This issue might thus lead to distortions when comparing the treatment of Hispanics to those acually classified as Black, White or Asian, since the mixture of races classified as "Hispanic" could even out any treatment disparities.

Additionally, since the exact taxonomy for race in the "COMPAS" data set is unknown, how people who identify as multiple "categories", eg. Latino and White or Latino and Black chose to identify can not be inferred. Therefore it can not be verified that the mapping between "COMPAS" and "NamSor" taxonomies as done in this work is correct.

5.2 Choice of reference group

When calculating fairness using "Aequitas", one always has to chose a reference group. In this audit, "Caucasian" and "Male" are used as the reference group. Depending on which reference group one choses, the results could change. For example, if looking at the FPR disparities for the gender endpoint between each group and the reference group with any threshold smaller than 0.4, one can see that all groups except Hispanic women have lower FPRs than Caucasian men. As FPR counts the relative amount of negatives predicted to be positive out of all negatives(Verma & Rubin 2018, 3), a lower FPR is actually good, because that means less errors when predicting to be correct. So one could say that it is actually Caucasian men who get unfair results here, with almost everyone else getting better results. This effect is even stronger when looking at the ethnicity endpoint. Because "Male Caucasian" has been chosen as the reference group it seems like the API is unfair towards all groups. However, looking at the disparities, it seems like the API only returns worse results for Caucasians.

5.3 Choice of score threshold

In this audit the value chosen as a threshold for "Aequitas"' score was the minimum available score because this is what was high enough for "NamSor" to decide for one result to return. However, developers could decide to chose a different threshold and for example ignore results that do not meet the minimum probability of being correct.

score threshold	group	pprev	npv	ppv	fpr	fnr	for
0.1	Female African-American	0.84	1.52	0.97	0.64	3.13	0.90
	Female Caucasian	0.98	1.96	0.99	0.78	1.07	0.81
	Female Hispanic	0.99	0.75	1.03	0.00	1.19	1.05
	Male African-American	0.86	1.24	0.98	0.66	2.96	0.95
	Male Caucasian	1.00	1.00	1.00	1.00	1.00	1.00
	Male Hispanic	0.98	1.53	1.02	0.37	1.17	0.89
0.2	Female African-American	0.87	1.15	0.96	0.72	4.68	0.95
	Female Caucasian	0.99	1.82	0.98	0.93	0.87	0.74
	Female Hispanic	1.00	0.00	1.02	1.32	1.40	1.31
	Male African-American	0.90	1.16	0.98	0.68	3.63	0.95
	Male Caucasian	1.00	1.00	1.00	1.00	1.00	1.00
	Male Hispanic	0.99	1.92	1.02	0.41	0.90	0.71
0.3	Female African-American	0.88	1.08	0.96	0.75	4.99	0.97
	Female Caucasian	0.99	1.70	0.98	0.91	1.00	0.75
	Female Hispanic	0.99	0.00	1.02	1.29	1.74	1.37
	Male African-American	0.91	1.07	0.98	0.71	4.02	0.97
	Male Caucasian	1.00	1.00	1.00	1.00	1.00	1.00
	Male Hispanic	0.98	1.71	1.02	0.40	1.11	0.74
0.4	Female African-American	0.94	1.20	0.95	0.84	3.87	0.90
	Female Caucasian	0.99	1.80	0.98	0.90	0.85	0.60
	Female Hispanic	1.00	0.00	1.02	1.22	1.49	1.50
	Male African-American	0.95	1.00	0.97	0.86	3.64	1.00
	Male Caucasian	1.00	1.00	1.00	1.00	1.00	1.00
	Male Hispanic	1.00	1.91	1.01	0.69	0.59	0.55

Table 5.1: Disparity metrics of the Gender Endpoint when using different score thresholds.

Depending on the score chosen, the fairness audit can produce different results, as can be seen in table 5.1 for the gender API endpoint and in table 5.2 for the ethnicity API endpoint.

The advantage of these differences is that one can do a fairness audit to evaluate which is the most fair threshold for assuming a result to be correct. "Fair" here is adaptable to an individual situation and one's own priorities. For example if, when using the gender endpoint, one is interested in FNR Parity a threshold for the score to be 0.1 (so, the probability returned by "NamSor" needs to be at least 0.9) leads to the FNR being three times as high for African-American Men than for Caucasian Men. But if one selects a higher score threshold of 0.3 (so the probability returned by "NamSor" needs to be at least 0.7) the FNR is four times as high for African-American Men than for Caucasian Men. One might thus chose a lower score threshold to increase fairness. Or, as another example, using the gender endpoint one might be especially interested in FPR Parity. Here, parity is given for Caucasian women for any

		pprev	npv	ppv	fpr	fnr	for
score threshold	group						
0.1	Female African-American	0.62	0.52	9.32	0.32	0.93	1.21
	Female Caucasian	0.63	1.40	0.00	0.57	1.02	0.83
	Female Hispanic	3.55	0.85	11.56	0.00	0.39	1.07
	Male African-American	0.38	0.87	8.46	0.17	0.96	1.06
	Male Caucasian	1.00	1.00	1.00	1.00	1.00	1.00
	Male Hispanic	4.12	0.57	11.56	0.00	0.33	1.18
0.2	Female African-American	0.91	0.73	1.98	0.30	0.86	1.11
	Female Caucasian	0.74	1.45	0.89	0.67	1.12	0.81
	Female Hispanic	1.92	1.78	2.08	0.29	0.15	0.67
	Male African-American	0.70	1.09	1.86	0.25	0.97	0.96
	Male Caucasian	1.00	1.00	1.00	1.00	1.00	1.00
	Male Hispanic	1.93	1.20	2.13	0.00	0.18	0.92
0.3	Female African-American	0.94	0.85	1.64	0.37	0.85	1.08
	Female Caucasian	0.81	1.41	0.91	0.75	1.20	0.80
	Female Hispanic	1.43	2.07	1.71	0.37	0.10	0.46
	Male African-American	0.79	1.17	1.55	0.34	1.00	0.91
	Male Caucasian	1.00	1.00	1.00	1.00	1.00	1.00
	Male Hispanic	1.46	1.44	1.74	0.26	0.14	0.78
0.4	Female African-American	0.99	1.02	1.52	0.48	0.73	0.99
	Female Caucasian	0.85	1.37	0.93	0.79	1.27	0.75
	Female Hispanic	1.19	1.83	1.58	0.44	0.13	0.45
	Male African-American	0.86	1.11	1.38	0.55	1.07	0.92
	Male Caucasian	1.00	1.00	1.00	1.00	1.00	1.00
	Male Hispanic	1.23	1.26	1.60	0.59	0.15	0.83

Table 5.2: Disparity metrics of the Ethnicity Endpoint when using different score thresholds.

threshold, but Caucasian women get less fair result with higher htresholds. However, a score of 0.4 or more would benefit more groups overall, leading to parity for all except Hispanic men and women. One might thus chose a threshold depending on the groups one needs to protect most.

5.4 Choice of fairness threshold

It is evident that the choice of the fairness threshold can influence whether the audit finds unfairness or not, depending on the disparity ratios. In this case, 0.8 was chosen as the fairness threshold and chosing a more restrictive threshold of 0.9 would not have made a difference for the results, neither for the gender nor for the ethnicity endpoint.

5.5 Choice of fairness measures

Only some of the available fairness measures have been tested in this work. More fairness metrics could be calculated without "Aequitas". With Conditional Statistical Parity, one could check whether legitimate factors make it harder for "NamSor" to infer gender or ethnicity, for example if the name contains certain syllables or has a certain length. Using Well-Calibration one could check whether "NamSor"'s probabilityCalibrated is equally well calibrated for all group. Additionally, the following can be calculated:

- Treatment Equality
- Test Fairness
- Balance for Neative Class
- Balance for Positive Class

5.6 Relative disparity versus absolute difference

"Aequitas" defines error disparity ratios instead of absolute error differences. This means that if both values which the disparity ratio is calculated from are very small, their absolute differences and thus their variance would be small but the ratio can still be high. This can be seen in the fairness audit of the gender endpoint with FDR values. FDR values range from

0.01 to 0.1 and have a variance of 0.001. The disparity ratios however range from 0.3 to 1.9. In this case we know that FDR Parity is fullfilled because precision (or PPV) parity is fullfilled and according to (Verma & Rubin 2018, p. 3), if one of both is fullfilled, the other is fullfilled as well. However, "Aequitas" comes to the conclusion that FDR parity is not fullfilled. In cases like these "Aequitas"' results need to be manually corrected if one opts for prioritizing absolute difference over relative disparity. This problem cascades to fairness metrics that are a combination of other fairness metrics, like Type I Parity.

5.7 Significance

While "Aequitas" is supposed to have a built-in significance check for it's fairness metrics, this feature does not appear to be working in the current version, 38.1 [1]. Therefore, significance of results was not checked in this work, except that it was excluded right away from the few amount of Asians in the "COMPAS" data set that results concerning the fairness towards Asians could not be significant. One way to check the results significance without "Aequitas" is to validate the results using more test data sets.

[1] A GitHub Issue was opened at `https://github.com/dssg/aequitas/issues/86` but to date received no reply from the "Aequitas" creators.

6 Conclusion

Fairness is required by law and can be mathematically defined in many ways. When creating a data based algorithm, unfairness can emerge when gathering data, processing data and when using the algorithm. It is therefore important to audit algorithms that can influence people for fairness and to be transparent about fairness limitations. This project contributed to the transparency about the fairness of "NamSor"'s gender and ethnicity inferring API endpoints.

No unfairness of "NamSor"'s gender inferring API endpoint could be detected between the groups "Female" and "Male" and "African-American", "Caucasian" and "Hispanic" and their subgroups in the "COMPAS" data set. Being used on the "COMPAS" data set, "NamSor"'s ethnicity inferring API endpoint has worse results for Caucasians. This result needs to be confirmed using more data sets, since if it was true, it could be a disadvantage when using "NamSor" to complement protected attribute data to data sets that should be analyzed for the representative or equal distribution of contained groups. If, in that case, the API more often miscategorizes Caucasians, the analyzer could come to the conclusion that the group of Caucasians in the data set was smaller than it actually is. In the worst case, this could lead to unfairness being overlooked. Thus, improving the results for the group "Caucasian" could benefit all other groups.

Bibliography

Aequitas Source Code (2019).
URL: *https://github.com/dssg/aequitas/blob/a61ef33a55a8e21611425f13c5688bae6743f04 src/aequitas/bias.py*

Bellamy, R. K. E., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., Nagar, S., Ramamurthy, K. N., Richards, J. T., Saha, D., Sattigeri, P., Singh, M., Varshney, K. R. & Zhang, Y. (2018). AI fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias, *CoRR* **abs/1810.01943**.
URL: *http://arxiv.org/abs/1810.01943*

Bias and Discrimination in AI (2020).
URL: *https://www.edx.org/course/bias-and-discrimination-in-ai*

Carsenat, E. (2019). Inferring gender from names in any region, language, or alphabet.

Carsenat, E. (2020). Back testing gender inference from names using wikidata.

Charter of Fundamental Rights of the European Union (2000). 2000/C 364/01.
URL: *https://www.europarl.europa.eu/charter/pdf/text_en.pdf*

COMPAS Analysis (2017).
URL: *https://github.com/propublica/compas-analysis/blob/master/CompasAnalysis.ipynb*

compas-scores-two-years (2016).
URL: *https://github.com/propublica/compas-analysis/blob/master/compas-scores-two-years.csv*

Convention for the Protection of Human Rights and Fundamental Freedoms (1950).
URL: *https://www.echr.coe.int/Documents/Convention_ENG.pdf*

Ethics Guidelines for Trustworthy AI (2019).
URL: *https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419*

Fernsel, L. (2019). Gender equality in computer science publications.
URL: *https://github.com/LiFaytheGoblin/Gender-Equality-in-CS-Publications/blob/master/Bachelors-Thesis-Gender-Equality-in-CS.pdf*

Flores, A., Bechtel, K. & Lowenkamp, C. (2016). False positives, false negatives, and false analyses: A rejoinder to "machine bias: There's software used across the country to predict future criminals. and it's biased against blacks.", *Federal probation* **80**.

Larson, J., Mattu, S., Kirchner, L. & Angwin, J. (2016a). How we analyzed the compas recidivism algorithm.
URL: *https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algo*

Larson, J., Mattu, S., Kirchner, L. & Angwin, J. (2016b). How we analyzed the compas recidivism algorithm.
URL: *https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algo*

Mazieres, A. & Roth, C. (2018). Large-scale diversity estimation through surname origin inference, *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique* **139**.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K. & Galstyan, A. (2019). A survey on bias and fairness in machine learning.

NamSor Tools V2 (2020).
URL: */urlhttps://github.com/namsor/namsor-tools-v2/wiki/NamSor-Tools-V2#classifiers*

O'Neil, C. (2016). *Weapons of Math Destruction*.

Race and Ethnicity (2017).
URL: *https://web.archive.org/web/20200603231927/https://www.census.gov/mso/www/training/pdf/race-ethnicity-onepager.pdf*

Saleiro, P., Kuester, B., Stevens, A., Anisfeld, A., Hinkson, L., London, J. & Ghani, R. (2018). Aequitas: A bias and fairness audit toolkit, *CoRR* **abs/1811.05577**.
URL: *http://arxiv.org/abs/1811.05577*

Schilling, M. (2002). Measuring diversity in the united states, *Math Horizons* **9**(4): 29–30.
URL: *http://www.jstor.org/stable/25678371*

SciPy v1.5.2 Reference Guide (2020).
URL: *https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ttest_ind.html*

Suresh, H. & Guttag, J. V. (2020). A framework for understanding unintended consequences of machine learning.

The International Bill of Human Rights (1948). Adopted and proclaimed by General Assembly resolution 217 A (III).
URL: *https://www.ohchr.org/Documents/Publications/Compilation1.1en.pdf*

Understanding Input Data (2018).
URL: *https://dssg.github.io/aequitas/input_data.html#Input-data-for-Python-package*

Understanding the Metrics (2018).
URL: *https://dssg.github.io/aequitas/metrics.html#Bias-Metrics*

Verma, S. & Rubin, J. (2018). Fairness definitions explained, *Proceedings of the International Workshop on Software Fairness*, FairWare '18, Association for Computing Machinery, New York, NY, USA, p. 1–7.
URL: *https://doi.org/10.1145/3194770.3194776*