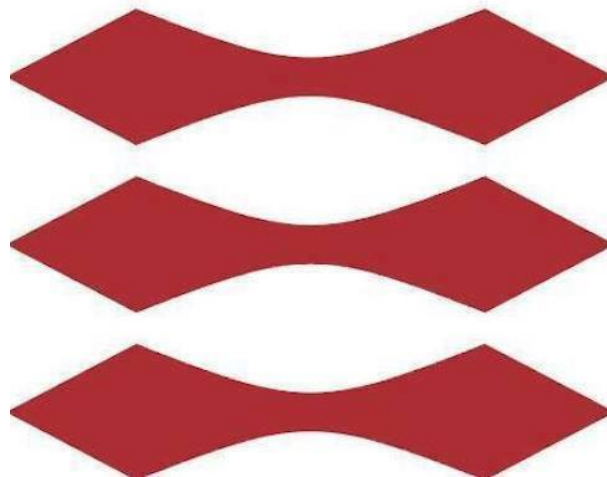


# 02446 - Project in Artificial Intelligence and Data Science

---

## Fairness in classification

DTU



Rasmus Stokholm Bryld, s183898  
Matilde Maria De Place, s183960  
Sunniva Olsrud Punsvik, s183924

## Contents

<b>1</b>	<b>Introduction</b>	<b>2</b>
<b>2</b>	<b>Data Description</b>	<b>2</b>
2.1	How the Data is Acquired . . . . .	2
2.2	Description of Attributes . . . . .	3
2.3	Distribution . . . . .	5
<b>3</b>	<b>Method</b>	<b>5</b>
3.1	Bias in Data . . . . .	5
3.2	Bias in Algorithms . . . . .	5
3.3	Constructing a Classifier . . . . .	6
<b>4</b>	<b>Results</b>	<b>6</b>
<b>5</b>	<b>Discussion</b>	<b>6</b>
5.1	Bias in Data . . . . .	6
<b>6</b>	<b>Conclusion</b>	<b>6</b>

## 1 Introduction

Fairness and bias are relatively new concepts in the realm of computer science, and over the last decade, the quality and accuracy of different AI models (especially machine and deep learning) have increased significantly. The outcome of this success has led us to introduce neural networks as part of various decision making tasks. In the U.S. there has been a rise in the use of software for risk assessments to help in court and every stage of the justice process[1], which has received heaps of criticism because this software, which is used in several states already, allegedly is biased against African Americans. The-state-of-the-art in this domain is quite interesting, some papers on fairness in classification often work with the notion of parity, that is; equality of outcome for protected groups. However, correcting a classifier in order to achieve for example demographic parity brings about a significant disadvantage, namely accuracy. Additionally, this method raises an important ethical question: which attributes are not protected? An alternative method was proposed by M. Hardt in 2016 [2] which introduces an equalised odds classifier to maximise the accuracy of the predictions. Another state-of-the-art paper which combined computer science, criminology, and statistics concluded that in most non-trivial cases, it is impossible to maximise accuracy simultaneously with fairness because one needs to consider challenging trade-offs[3]. Due to the fact that these are new concepts, the consensus is not clear and therefore an interesting topic to inspect.

This project will investigate the meaning of bias in data and algorithms, how to detect and examine them, and how biases contribute to risk assessment scores. This project will specifically delve into the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) dataset, to see if earlier findings of bias are reproducible. Additionally, by implementing a classifier trained on the COMPAS dataset by following the method used in [2]. Then by judging the importance of different features and the direction they influence the decile score, this will be done by considering some selected data, that is, looking into all defendants who committed misdemeanours and comparing their respective risk assessment scores with their race. By inspecting whether they re-offended or not, one can look for which combination of features are mostly responsible for the decile score. Lastly the report will consider the ethical aspects of using such technology and ask the ethical question of why we sentence people to jail and when this should be done.

## 2 Data Description

This project takes advantages of two, out of several, data sets obtained by ProPublica: `compas-scores.csv` and `compas-scores-two-years.csv`[4]. `compas-scores.csv` consists of 47 attributes and 11,757 observations, all with COMPAS scores accessed at pretrial. `compas-scores-two-years.csv` is a subset of `compas-scores.csv`, consisting of 7,214 observations 53 attributes. One of the additional attributes being a binary variable with information of two years recidivism. Hereto, one shall know Northpoints definition of recidivism as well as the interpretation made by ProPublica: *"finger-printable arrest involving a charge and a filing for any uniform crime reporting (UCR) code. We interpreted that to mean a criminal offence that resulted in a jail booking and took place after the crime for which the person was COMPAS scored."*[4].

### 2.1 How the Data is Acquired

Before diving into a description of the remaining attributes, an elaboration of ProPublicas data aggregation will be stated. From Broward County Sheriff's Office in Florida, ProPublica received and downloaded COMPAS scores of 18,610 people obtained in the period 2013 to 2014,

as well as jail records from January 2013 o April 2016. They only kept the 11,757 COMPAS scores assessed at pretrial stage. From Broward County Clerk’s Office website, April 1, 2016, ProPublica obtained 80,000 public criminal records and from the Florida Department of Corrections website, Propublica downloaded incarceration records aw well. In order to build each relevant profile, Propublica matched COMPAS scores with criminal history using date of birth and first and last name. A persons race was defined according to Broward County Sheriffs Office classification[4].

## 2.2 Description of Attributes

For simplicity, one ought to divide the attributes into four and five categories for `compas-scores.csv` and `compas-scores-two-years.csv` respectively. One shall notice that only `compas-scores-two-years.csv` contains the fifth category.

1. *Personal information of defendant*
2. *Case of which COMPAS scores are assessed*
3. *Recidivism*: Information of whether or not recidivism occurs. In case of occurence, one can obtain information of the relevant case.
4. *Violent recidivism*: Information of whether or not violent recidivism occurs. In case of occurence, one can obtain information of the relevant case.
5. *2 years recidivism*: Whether or not recidivism occurs within 2 years.

Once again for simplicity, attributes of *Recidivism* and *Violent recidivism* with definition equivalent a attributes of *Case of which COMPAS scores are assessed* are not included in the respective table but listed below with their equivalent attribute. One ought to notice that attributes staring woth `r_` belongs to *Recidivism* and `v_` and `vr_` belongs to *Violent recidivism*.

- `r_case_number` and `vr_case_number` are equivalent to `c_case_number`
- `r_days_from_arrest` equivalent to `c_days_from_arrest`
- `r_offense_date` and `vr_offense_date` equivalent to `c_offense_date`
- `r_jail_in` equivalent to `c_jail_in`
- `r_jail_out` equivalent to `c_jail_out`
- `r_charge_desc` and `vr_charge_desc` are equivalent to `c_charge_desc`

## Attributes in category "Personal information of defendant"

Attribute	Description of attribute
name, first, last	Name of defendant (full and divided into first and last name)
sex	Male or Female
age	Age of defendant in years
age_cat	'Greater than 45', '25 - 45', 'Less than 25'
race	'Other', 'African-American', 'Caucasian', 'Hispanic', 'Native American', 'Asian'
dob	Date of birth of defendant (year-month-date)

## Attributes in category "Case of which COMPAS scores are assessed"

days_b_screening_arrest	Days between c_jail_in and (compas_screening_date) (integer)
c_jail_in	Jail time begins (year-month-date, hour:minute:second)
c_jail_out	Jail time begins (year-month-date, hour:minute:second)
c_case_number	Case number of case of relevant case
c_offense_date / c_arrest_date	Offence/arrest date of relevant case (only one completed)
c_days_from_compas	Days between compas_screening_date and c_offense_date or c_arrest_date
c_charge_degree	Classification of relevant crime (misdemeanor or felony)[5]
c_charge_desc	Description of crime

## Attributes in category "Recidivism"

is_recid	Binary, 1 if recidivism occurs 0 otherwise
r_charge_degree	Possible values: (F3), (M1), (F2), (M2), (MO3), (F1), (F6),(F7), (CO3), (F5) <sup>1</sup>
decile_score.1 and decile_score	COMPAS score "Risk of Recidivism", integers from 1-10
score_text	Categorical "Risk of Recidivism". Low: 1-4 Medium: 5-7 High: 8-10
screening_date	Date of which "Risk of Recidivism" is assessed

## Attributes in category "Violent recidivism"

is_violent_recid	Binary, 1 if violent recidivism occurs 0 otherwise
vr_charge_degree	Possible values: (F3), (F2), (F1), (M1), (MO3), (M2), (F6), (F7), (F5) <sup>2</sup>
v_decile_score	COMPAS score "Risk of Violence", integers from 1-10
v_score_text	Categorical "Risk of Violence". Low: 1-4 Medium: 5-7 High: 8-10
v_screening_date	Date of which "Risk of Recidivism" is assessed

## Attributes in category "2 years recidivism"

two_years_recid	Binary, 1 if recidivism occurs within two years, 0 otherwise
-----------------	--

Table 1

## 2.3 Distribution

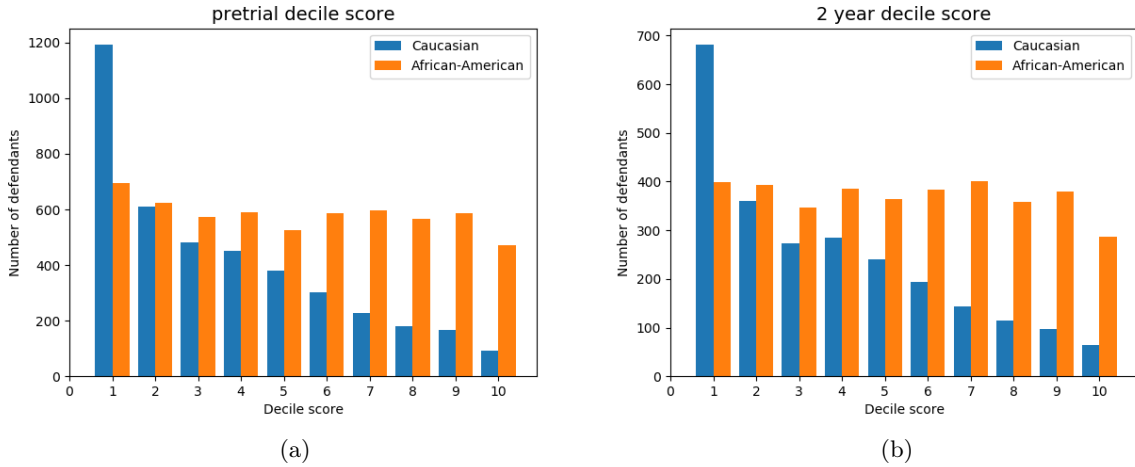


Figure 1

## 3 Method

The first point in question is to comprehend the concept of bias in data and algorithms, to search for these biases in the COMPAS dataset and precisely what makes the algorithm biased. One ought to make a distinct definition of what bias entails in this context, to have a clear-cut interpretation surrounding bias and fairness for consistency reasons, as well as it is important to distinguish that bias in data may not be the same as bias in a classifier. The literature on this is quite foggy and some papers have different standards as to what bias means, where one typically designates a mathematical definition when it comes to classifiers, however, first one needs to tackle what bias means in the domain of data analysis. Second point in question is how to detect bias and consequently reduce them.

### 3.1 Bias in Data

This paper will use the following definition for bias in data: *"Bias is taken to mean interference in the outcomes of research by predetermined ideas, prejudice or influence in a certain direction. Data can be biased but so can the people who analyse the data. When data is biased, we mean that the sample is not representative of the entire population"*[6].

### 3.2 Bias in Algorithms

To obtain a definition of bias in classification, this paper will use the definition used in [2], which states two criterion on what bias is in supervised learning (notice that the cited paper uses the terms "discriminatory" and "non-discriminatory" and this paper will use "biased" and "unbiased"). The two criterion suggested in [2] are *equalised odds* and *equal opportunity*. This method is a probabilistic approach to define bias, and quantifiable able to reduce them. The goal here is to predict risk assessment scores, based on the available features in the COMPAS dataset, whilst assuring the prediction is as unbiased as possible with respect to a protected attribute, which this case in point is "race".

### 3.3 Constructing a Classifier

## 4 Results

## 5 Discussion

### 5.1 Bias in Data

The problem which occurs when investigating bias in data, is how to search for it? By looking at the current demographics is a good starting point; the racial makeup of the U.S. population compared to the U.S.' adult correctional population. The U.S. is renowned for having the largest incarcerated population in the world, accumulating 21% of the world's prisoners[7]. African Americans accounts for 12.7% of the U.S. population[8], but comprises 34% of the correctional population[7] (when adding Hispanics, the correctional populations accumulates to 54%).

## 6 Conclusion

## References

- [1] J. Angwin, L. Kirchner, J. Larson, and S. Mattu, "Machine bias." <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>, May 2016. Accessed on 20.02.2020.
- [2] M. Hardt, "Equality of opportunity in supervised learning," *NIPS 2016*, Oct 2016. Accessed on 14.03.2020.
- [3] R. Berk, H. Heidari, S. Jabbari, M. Kearns, and A. Roth, "Fairness in criminal justice risk assessments: The state of the art," *SAGE Journals*, July 2018. Accessed on 14.03.2020.
- [4] J. Angwin, L. Kirchner, J. Larson, and S. Mattu, "How we analyzed the compas recidivism algorithm." <https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm>, May 2016. Accessed on 14.03.2020.
- [5] Missing, "Pennsylvania crime classification." <https://www.davidcohenlawfirm.com/pennsylvania-crime-classification?fbclid=IwAR1tGFrdZOZT7qeBFw2X1jlonmyKJxs7w5d1bTo5FBUJe-pC2I0x-ext2YM>, Missing Missing. Accessed on 14.03.2020.
- [6] S. d. Vries, "5 types of bias in data analytics." <https://cmotions.nl/en/5-typen-bias-data-analytics/>, Nov 2017. Accessed on 14.03.2020.
- [7] NAACP, "Criminal justice fact sheet." <https://www.naacp.org/criminal-justice-fact-sheet/>, 2019. Accessed on 14.03.2020.
- [8] U. S. C. Bureau, "The united states census quick search." <https://www.census.gov/quickfacts/fact/table/US/PST045218>, 2019. Accessed on 14.03.2020.