Bachelor's Thesis

Radboud Universiteit Nijmegen

Optimizing fairness in machine learning systems through random repair of a biased dataset.

Author:
V. Philips, s4288149
vincentwiliam.philips@student.ru.nl

Supervisor:
prof. dr. T.M. Heskes
Department of Computer Science

Nov 25, 2019

Abstract

Machine learning systems use datasets that could have biases in their data. These biases could cause unwanted biases in decision making towards groups with sensitive attribute such as race, gender and sexual orientation [5][7]. A possible method to address this problem is a preprocessing method called Random Repair which was introduced in [3] in 2018 [3].

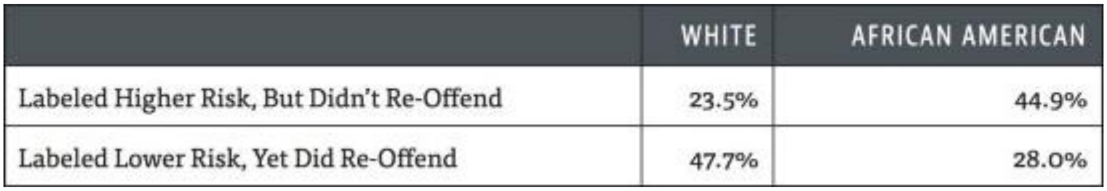
In this thesis we provide insights on the effects that random repair has on fairness and classifier accuracy in machine learning systems. We applied random repair on the adult income dataset and on the COMPAS recidivism dataset. These datasets are known to be biased datasets [5][13]. For each dataset, we compared the fairness and accuracy of a logistic regression classifier and a random forest classifier before and after preprocessing the datasets with random repair. In this thesis we measure the fairness called demographic parity of our classifiers by calculating the disparate impact index.

Our research shows that increasing the fairness through random repair results in the desired amount of fairness for both classifiers in both datasets. However, increasing the fairness trough random repair also decreases the classifier accuracy of both classifiers.

1 Introduction

Machine learning systems are increasingly being used for automated classification in areas such as insurance, employment, education, and predictive policing [1][2]. Within those areas it is assumed that classifications made by computers are objective and unbiased [2][7]. However, machine learning systems do not necessary hold this assumption, as their classifications may depend on datasets plagued with biases. Machine learnings systems use algorithms called classifiers on input data to assign inputs to chosen groups (i.e.: classification) [2].

Input data with biases in them can cause machine learning systems to make unfair classifications. These unfair classifications refer to outcomes that disproportionately hurt (or benefit) people with certain sensitive attribute values (e.g., race, gender, sexual orientation). An illustrative example is COMPAS, a machine learning algorithm used by judges in the U.S.A. for classifying a convict based on the likelihood that he or she commits a crime after being released from prison. In 2009 it was found that COMPAS had an accuracy rate of 68 percent in a sample of 2,328 people [15]. The research in 2016 revealed that COMPAS classifies African-American convicts incorrectly more often than Caucasian convicts when it comes to high risk classifications. The opposite mistake is made for Caucasian convicts: they are more likely than African-American convicts to be classified as low risk but still commit a crime after release. Details of this research are shown in figure 1.



	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Figure 1: Results showing bias decision making in COMPAS [5].

This example shows the importance of optimizing fairness in machine learning systems. To tackle the problem of fairness in machine learning system we first must define fairness. There are many different definitions of fairness that have been proposed in the literature [6][7][9]. Statistical definitions of fairness partition the world into some set of protected groups, and then ask that some statistical property be approximately equalized across these groups [14]. Typically, these protected groups are defined via sensitive attributes.

The type of fairness we are trying to achieve in this thesis is referred to as demographic parity, which is one of many definitions used for group fairness [14]. Group fairness requires that groups with sensitive attributes should be treated similarly to groups with non-sensitive attributes. Demographic parity also called Statistical Parity, requires that equal proportions of the privileged and unprivileged group receive the same classifications made by the machine learning system. In our example, this would mean that African-American convicts are just as likely as Caucasian convicts to be classified the same risk of committing a crime after release.

But how do we optimize fairness in machine learning? Removing sensitive attributes from the datasets is not enough to achieve fairness within machine learning systems. Other highly correlated features of the sensitive attribute might be present in the dataset. Biased decision making is still likely to happen if such highly correlated features are included in the prediction making process [2].

Many different methods have been considered in the machine learning literature to optimize fairness [7]. These methods are categorized by the stage of intervention during the machine learnings process. The first category known as in-processing consists of optimization at training time (e.g. changing the classifier(s) to prevent correlation with sensitive attribute(s) [7]. The second solution known as pre-processing takes place before classification and consists of transforming the input data of machine learning systems [7]. The third solution known as post-processing takes place after classification and consists of altering classifications made by the machine learning systems in such a way that they satisfy chosen fairness constraints [7].

In this thesis we will solely try to optimize fairness through random repair. Random repair is a form of pre-processing which creates a new data representation ~X by pre-processing an input dataset X. The created data representation ~X will have information correlated to the sensitive attribute removed, whilst preserving other information as much as possible [3]. We will clarify the random repair method in more detail in section 4.3 and 4.4.

We will use an index called Disparate Impact to measure the demographic parity before and after random repair. The disparate impact index has a minimum of zero and a maximum of one. Here zero implies no demographic parity and one would indicate complete demographic parity. In our example the disparate impact index would be equal to one if the proportion of African-American convicts that were classified as low risk is equal to the proportion of Caucasian convicts that were classified as low risk. We will clarify the disparate impact index in more detail in section 4.5.

Achieving a disparate impact index of one would be at a to high expense of the classifier accuracy of the machine learning system [7]. To keep both our classifier accuracy and fairness in our machine system as high as possible, we will aim for a disparate impact index of 0.8 through random repair. Aiming for a disparate impact index of 0.8 is standard within the literature [7]. Its origin lies in the four-fifths rule from the *Uniform Guidelines for Employee Selection Procedures* which are mandatory in the working industry in U.S.A since 1978 [11][12]. According to these guidelines the disparate impact index should at least be 0.8 to reasonably rule out any form of group unfairness. This means that if the proportion of the privileged group that received the positive outcome is equal to X and the proportion of the unprivileged group that received the positive outcome is less than 80 percent of that X; the positive outcome was unfairly distributed between the two groups.

2 Research formulation

Many different methods of processing datasets to increase fairness in machine learnings systems have been proposed within the literature [7]. However very little research has been done to truly reflect the effectiveness of these methods. Instead most researchers within the literature are currently creating new algorithms, extending existing algorithms or question the existing measurements of fairness. This hinders advancements in the field, since it complicates seeing which accomplishments of current methods are truly promising for the future.

An example of a method that could be promising is called random repair [3]. Although introduced in 2018 it hasn’t received any follow-up research within the literature that could reflect is effectiveness.

The aim of this project is to provide more insight on the effects that random repair has on fairness and classifier accuracy in machine learning systems.

This brings us to the research question of this thesis:

How well does pre-processing a biased dataset with random repair enhance the fairness of machine learning systems?

To make a reasonable statement regarding the effectiveness of random repair we will aim to reproduce the random repair results of [3] on the adult income dataset and on the COMPAS recidivism dataset. These datasets are known to be biased datasets [5][13]. To achieve this, we will compare the fairness and accuracy of a logistic regression classifier and a random forest classifier, on a subset of 2000 individuals from each dataset, before and after preprocessing the datasets with random repair. We want to learn what amount of random repair yields a disparate impact index of 0.8 and what accuracy our classifiers have with that amount of repair. For this reason, we will increase the amount of random repair after each repair until we reach a disparate impact index of 0.8. However, for comparison, we will continue this increase until $\lambda=1$, which is the maximum amount of random repair possible. In this thesis we will refer to random repair with $\lambda=1$ as total repair.

3 Related Work

Before elaborating random repair, this section will explain why we choose to approximate fairness by measuring demographic parity and why we try to enhance it with random repair.

3.1 Fairness in machine learning systems

Most definitions of fairness in the literature belong to the category individual fairness or group fairness[7][9].

Individual fairness requires that similar individuals should have similar outcomes [8]. That is, after deciding which metric individual similarity should be judged upon, individuals who are close in that metric should have similar probabilities of any classification. **In our example,** this would mean that convicts which differ in race but have similar other features (used for classification) should receive the same risk classification.

Group fairness requires that groups with sensitive attributes should be treated similar to groups with non-sensitive attributes. The three most commonly used criteria for group fairness within the literature are [7]:
1. Demographic Parity
2. Equalized Odds
3. Predictive Rate Parity

Demographic Parity requires that equal proportions of the privileged and unprivileged group receive the same classification(s). Its standard within the literature to take a less strict version of demographic parity by aiming for a disparate index of at least 0.8 [7]. **In our example,** the disparate impact index is equal to 0.8 when for example; 50 percent of Caucasian convicts would receive a low risk classification and 40 percent of African-American convicts also receive a low risk classification.

Equalized Odds requires that classifications made are independent of sensitive attributes conditional on the measured true outcomes. **In our example,** this would mean that an equal proportion of each group that commits a crime after release should be classified as high risk. In addition, an equal proportion of each group that doesn't commit a crime after release should be classified as low risk.

Positive and Negative Predictive Rate Parity requires that both the privileged and unprivileged group have equal PPV and NPV [16]. PPV is the proportion of correct positive outcome classification and NPV is the proportion of correct negative outcome classification. **In our example,** PPV requires that the proportion of African-American and Caucasian convicts that were correctly classified as low risk are equal. Whereas NPV requires that the proportion of African-American and Caucasian convicts that were correctly classified as high risk are equal.

We choose to measure group fairness, since defining an individual similarity metric for both datasets seems rather complicated. We choose to measure group fairness through demographic parity since this is commonly used within the literature and because it is used in [3].

3.2 Proposed solutions to achieve group fairness in machine learning

These three types of solutions are widely considered within the literature to enhance group fairness [7].
1. In-processing methods.
2. Pre-processing methods.
3. Post-processing methods.

In-processing methods optimize fairness at training time by changing the classifier or by adding a constraint or a regularization term to the existing training objective of machine learning systems. In-processing methods have the highest performance on accuracy and fairness measures [7]. However, implementing in-processing methods is task specific. That is, each task will need modifications to the classifier, which may not be possible or too time consuming in many scenarios.

Pre-processing methods sample, re-weight or alter the input data X to neutralize discriminatory effects, thereby creating a new data representation ~X [4]. After pre-processing, the generated dataset ~X should be free of biases but still contain as much other information of the original dataset as possible. The motivation lies in the assumption that removing biases present in the input data should increase fairness in classification. Unlike in-processing methods, which use input and output data, pre-processing only uses the input data and there are no changes needed to the classifier. However, pre-processing methods often have lower performance on accuracy than in-processing methods [7].

Post-processing methods alter classifications made in a way that they satisfy chosen fairness constraints. The motivation behind post-processing methods is that unfair classifications can be made fair by altering them. Unlike in-processing methods which use input and output data and pre-processing which use input data, post-processing methods only use the output data. Like pre-processing there are no changes needed to the classifier. However, post-processing methods often have lower performance on accuracy than in-processing methods [7].

We choose to use a pre-processing method called random repair since our research tries to provide insights in its effect on enhancing fairness in machine learning systems.

4 Experimental Methods

In this section we elaborate the datasets, measurements and methods that we use in this thesis.

4.1 Datasets

We will try to enhance demographic parity through Random repair on these 2 datasets:
1. The COMPAS recidivism dataset, retrieved from here.
2. The adult income dataset, retrieved from here.

Both datasets had their numerical features normalized and their categorical data made ordinal when possible or removed otherwise. The accuracy of the classifiers does not experience a dramatic decrease in accuracy because of these two steps. This was proven by running the classifiers with and without these two steps.

The COMPAS recidivism dataset contains criminal history, jail and prison time, demographics and COMPAS risk scores for defendants from Broward County. The classification task for our machine learning system is to predict risk level (low, medium or high) of how likely a convict will commit another crime. The group attribute on which we measure fairness is the race of the individual; 'Caucasian' or 'African-American'. In 2016 it was found that with this dataset COMPAS produces much higher false positive rate for African-American individuals than Caucasian individuals [5]. The subset of this dataset that we will use contains 663 Caucasian and 1337 African-American convicts.

The adult income dataset was extracted by Barry Becker from the 1994 US Census Database. This dataset contains information such as occupation, age, income, race, gender and more. The classification task for our machine learning system is to predict whether an individual earned more than 50.000 dollars in 1994. The group attribute on which we measure fairness is the gender of the individual; 'Male' or 'Female'. It was found that with this dataset, machine learning systems were three times as likely to predict a salary higher than 50.000 dollars for male individuals compared to female individuals. [13]. The subset of this dataset that we will use contains 1344 Male and 656 Female individuals.

4.1 **Classifiers**

In this thesis we use the following classifiers on the adult income dataset:
1. A logistic regression classifier.
2. A random forest classifier.

We choose to use these two classifiers since other more complex classifiers are prone to overfitting. Overfitting happens when a classifier learns the detail and noise in the training data to the extent that it negatively impacts the performance of the classifier on new data.

4.2 Random repair

In this section, we explain how random repair creates a new data representation ~X by pre-processing an input dataset X.

Let {(X_i), i=1,…,N} be an observed sample of the input dataset X, and denote by n0 the number of instances of the unprivileged group and denote by n1 the number of instances of the privileged group. Let the variable S take the value zero for the unprivileged group and the value one for the privileged group. Let the sample set be divided based on variable S. With this we can write:

$$\begin{aligned} x_{0,i} &:= X_i, && \text{if } s_i = 0,\ i = 1, \dots, n_0, \\ x_{1,j-n_0} &:= X_j, && \text{if } s_j = 1,\ j = n_0 + 1, \dots, N = n_0 + n_1. \end{aligned}$$

This yields the two sample sets: $\mathcal{X}_0 = \{x_{0,1}, \dots, x_{0,n_0}\}$ and $\mathcal{X}_1 = \{x_{1,1}, \dots, x_{1,n_1}\}$

In our example, Variable S stands for race, which is value zero for Caucasian convicts and value one for African-American convicts. The two created sample sets consist of $\mathcal{X}_0$ containing all Caucasian convicts' data and $\mathcal{X}_1$ containing all African-American convicts' data.

The random repair method consists of randomly altering the original conditional distribution of each datapoint in the sample sets, by either keeping the original conditional distribution or transforming it to a target distribution μ (the latter will be explained later). The choice between both is governed by the Bernoulli variable B with parameter λ, B ~ B(λ). Here the Bernoulli variable takes the value one with probability λ and the value zero with probability 1 – λ. Whereby λ ∈ [0,1] represent the amount of repair desired for a given input dataset X, since λ = 0 leaves the conditional distributions unchanged and λ = 1 yields a demographic parity equal to one. For each datapoint in the sample sets a value for B is generated. If the variable B is equals to zero, the original conditional distribution remains unchanged. If the variable B equals to one, the distribution is "repaired" by replacing it with the target distribution μ.

To compute the target distribution μ we must firsts calculate a transport distribution matrix $\hat{\gamma}$ which is a matrix that can be used to transform the distribution of dataset X_0 to X_1 with minimal transport cost. With this $\hat{\gamma}$ a new distribution $\mu_{B,n}$ is computed. This $\mu_{B,n}$ is a Wasserstein barycenter of the distribution of X_0 to X_1 with respect to weights π_0 and π_1. The wasserstein barycenter is a distribution which has the lowest possible sum of wasserstein distances to each datapoint in the input dataset X_0 and X_1. The wasserstein distance is a mathematical distance which describes the distance between two probability distributions. See [17] for more details on how the transport distribution matrix $\hat{\gamma}$ and the barycenter $\mu_{B,n}$ are computed.

Each datapoint in X_0 to X_1 that is repaired will split its mass to be transported into their target distribution(s) μ which are computed by:

$$\tilde{x}_{0,i,j} = \tilde{x}_{1,j,i} := \pi_0 x_{0,i} + \pi_1 x_{1,j},$$

The original code used in this thesis, containing both classifiers and the random repair method for the adult income dataset can be found here and here. For this thesis the original code has been altered for both datasets and can be found here.

4.3 Example of random repair

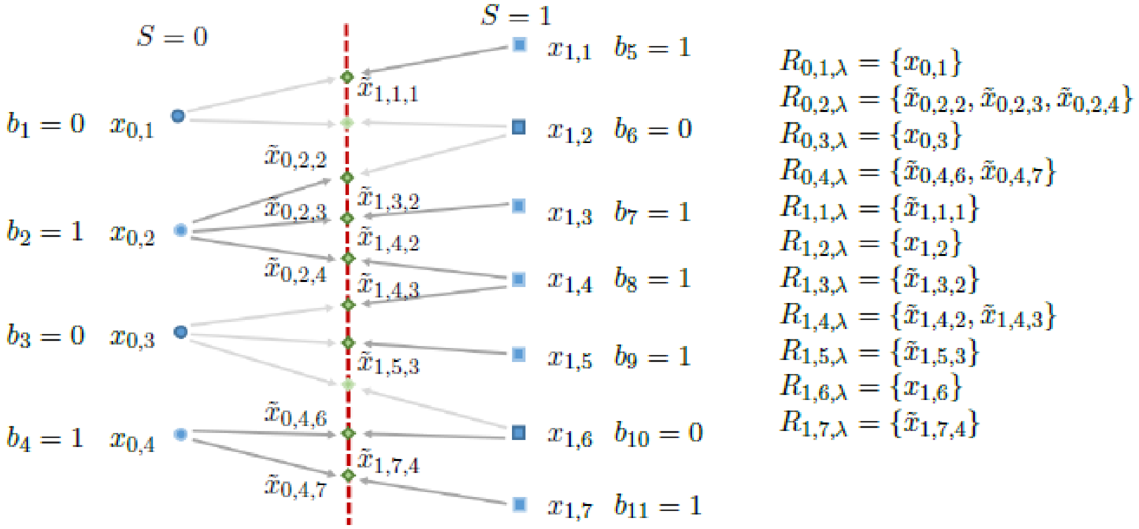
We have an observed sample set of the adult income dataset of 11 individuals, which after dividing on variable S (gender) yields two sample sets X_0 and X_1 of sizes $n_0 = 4$ and $n_1 = 7$, respectively:

$$\mathcal{X}_{0,\lambda} = \{x_{0,1}, x_{0,2}, x_{0,3}, x_{0,4}\}$$

$$\mathcal{X}_{1,\lambda} = \{x_{1,1}, x_{1,2}, x_{1,3}, x_{1,4}\ x_{1,5}, x_{1,6}, x_{1,7}\}$$

The two created sample sets consist of $\mathcal{X}_0$ containing the data of 4 male individuals and $\mathcal{X}_1$ containing the data of 7 female individuals. Because there are more female than male individuals and using the random repair method requires an equal amount of privileged and unprivileged datapoints [3]; some male individual datapoints will be copied and repaired towards a different female datapoint target. After classification the classifications made for the copied datapoints will be converted to one for each original copied datapoint.

Figure 2 below shows us that the distribution of each individual (blue) is either unchanged (indicated by light arrow) or “repaired”(indicated by dark arrow) by transforming the individuals’ distribution to a target distribution µ (green) based on its value of B.



$S=0$
$S=1$
$b_1=0$ $x_{0,1}$
$b_2=1$ $x_{0,2}$
$b_3=0$ $x_{0,3}$
$b_4=1$ $x_{0,4}$
$\tilde{x}_{0,2,2}$
$\tilde{x}_{0,2,3}$
$\tilde{x}_{0,2,4}$
$\tilde{x}_{0,4,6}$
$\tilde{x}_{0,4,7}$
$\tilde{x}_{1,1,1}$
$\tilde{x}_{1,3,2}$
$\tilde{x}_{1,4,2}$
$\tilde{x}_{1,4,3}$
$\tilde{x}_{1,5,3}$
$\tilde{x}_{1,7,4}$
$x_{1,1}$ $b_5=1$
$x_{1,2}$ $b_6=0$
$x_{1,3}$ $b_7=1$
$x_{1,4}$ $b_8=1$
$x_{1,5}$ $b_9=1$
$x_{1,6}$ $b_{10}=0$
$x_{1,7}$ $b_{11}=1$
$R_{0,1,\lambda}=\{x_{0,1}\}$
$R_{0,2,\lambda}=\{\tilde{x}_{0,2,2},\tilde{x}_{0,2,3},\tilde{x}_{0,2,4}\}$
$R_{0,3,\lambda}=\{x_{0,3}\}$
$R_{0,4,\lambda}=\{\tilde{x}_{0,4,6},\tilde{x}_{0,4,7}\}$
$R_{1,1,\lambda}=\{\tilde{x}_{1,1,1}\}$
$R_{1,2,\lambda}=\{x_{1,2}\}$
$R_{1,3,\lambda}=\{\tilde{x}_{1,3,2}\}$
$R_{1,4,\lambda}=\{\tilde{x}_{1,4,2},\tilde{x}_{1,4,3}\}$
$R_{1,5,\lambda}=\{\tilde{x}_{1,5,3}\}$
$R_{1,6,\lambda}=\{x_{1,6}\}$
$R_{1,7,\lambda}=\{\tilde{x}_{1,7,4}\}$

Figure 2: Example random repair with λ = 0,5 and its resulting sets R [3].

After the random repair method, we have 2 randomly repaired sets:

$$\tilde{\mathcal{X}}_{0,\lambda}=\{x_{0,1},\tilde{x}_{0,2,2},\tilde{x}_{0,2,3},\tilde{x}_{0,2,4},x_{0,3},\tilde{x}_{0,4,6},\tilde{x}_{0,4,7}\}$$

$$\tilde{\mathcal{X}}_{1,\lambda}=\{\tilde{x}_{1,1,1},x_{1,2},\tilde{x}_{1,3,2},\tilde{x}_{1,4,2},\tilde{x}_{1,4,3},\tilde{x}_{1,5,3},x_{1,6},\tilde{x}_{1,7,4}\}$$

Both datasets can be combined to create a new data representation ~X which can be used by machine learning systems. Using data representation ~X should yield more fair classification than using dataset X.

4.5 Measuring disparate impact

To measure the disparate impact, we separate individuals in a dataset into two groups. One with ‘negative’ and one with ‘positive’ sensitive attributes, which we will respectively refer to as the unprivileged and the privileged group. Then the proportion of the unprivileged group that received the positive outcome is divided by the proportion of the privileged group that received the positive outcome [10][11].

$$\textbf{Disparate Impact} = \frac{\textbf{Pr}\textbf{\textit{(positive outcome | unprivileged)}}}{\textbf{Pr}\textbf{\textit{(positive outcome | privileged)}}}$$

Figure 3 Disparate Impact Formula.

In our example, the unprivileged group contains African-American convicts and the privileged group contains the Caucasian convicts while the positive outcome is being classified as a low risk. Thus:

$$\textbf{Disparate Impact} = \frac{\textbf{Pr}\textbf{\textit{(low risk classification | African-American)}}}{\textbf{Pr}\textbf{\textit{(low risk classification | Caucasian)}}}$$

Figure 4 Disparate Impact Formula for the COMPAS recidivism dataset.

By the adult income dataset, the unprivileged group contains the female individuals and the privileged group contains the male individuals while the positive outcome is being classified with a salary higher than 50.000 dollars. Thus:

$$\text{Disparate Impact} = \frac{Pr(\textit{higher salary} \mid \textit{female})}{Pr(\textit{higher salary} \mid \textit{male})}$$

Figure 5 Disparate Impact Formula for the adult income dataset

4.6 Measuring accuracy

The accuracy of both classifiers can be measured by calculating the classification errors made by each classifier. We calculate the classification error by dividing the number of incorrect classifications with the number of total classifications. Thus:

$$\textbf{Classification Error} = \frac{\textbf{incorrect classifications}}{\textbf{total classifications}}$$

Figure 6 Disparate Impact Formula for the adult income dataset

5 Results with the COMPAS recidivism dataset

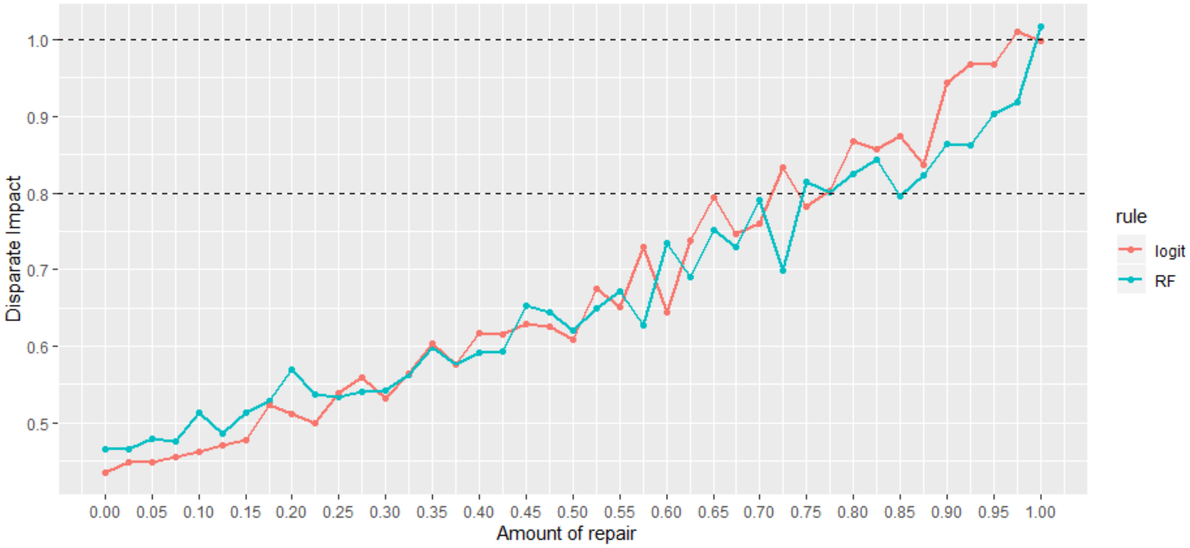


Figure 7: Disparate impact with random repair for the logistic regression (red) and the random forest classifier(blue).

Disparate Impact	None (λ=0)	Random (DI≈0.8)	Total (λ =1)
Logistic Regression classifier	0,435	0,834 (λ=0,725)	0,998
Random Forest classifier	0,465	0,815 (λ=0,75)	1,017

Figure 8: Amount of Disparate Impact for each form of repair.

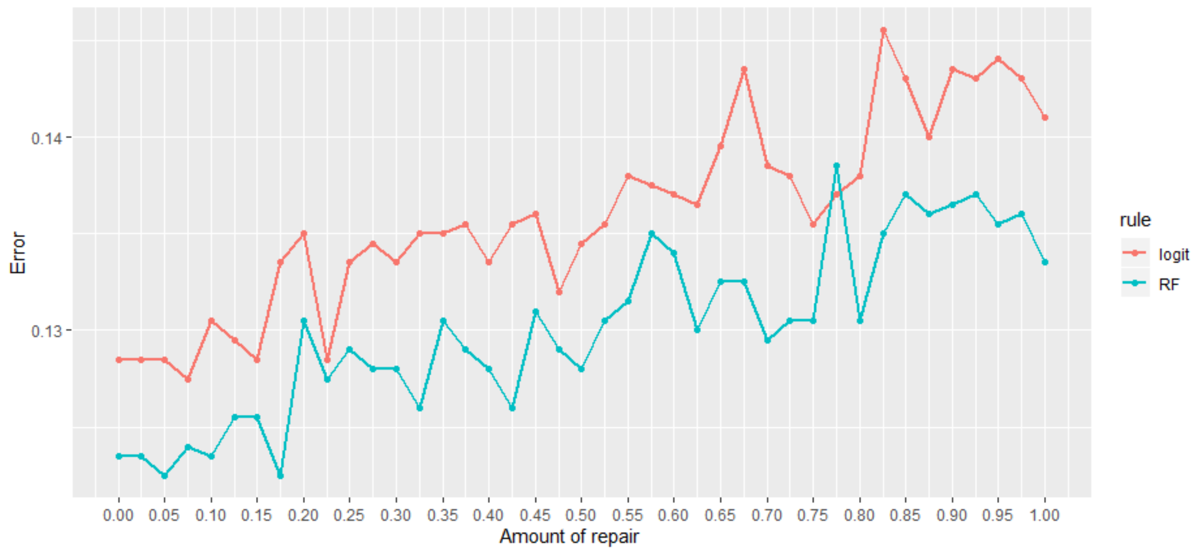


Figure 9: Classification errors of the classifications with logistic regression (red) and the random forest classifier(blue).

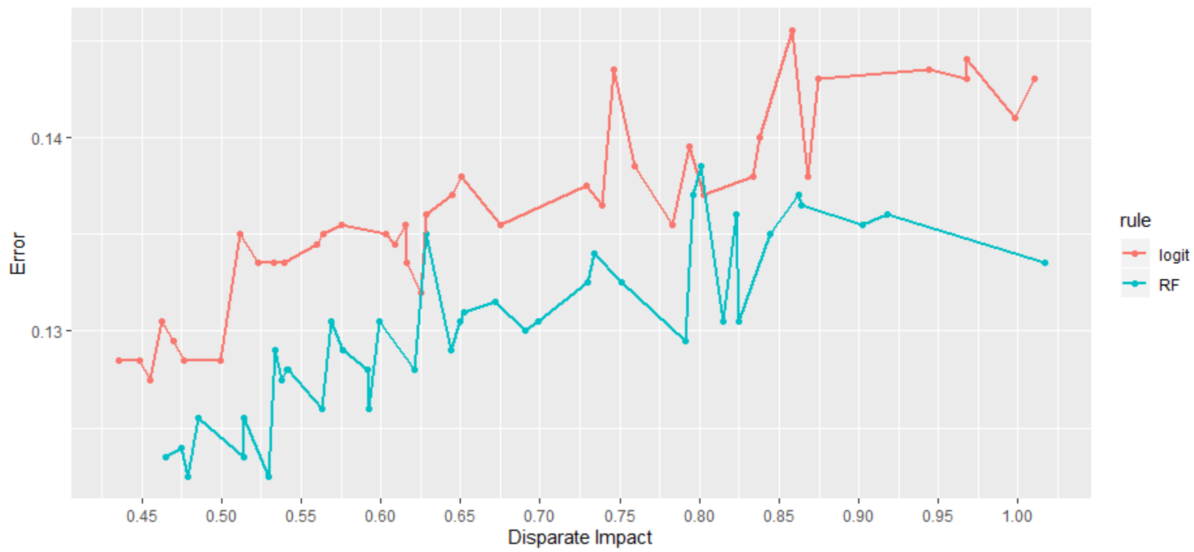


Figure 10: Classification errors of the classifications with logistic regression (red) and the random forest classifier(blue).

Logistic Regression classifier	None	Random	Total
Error rate	0,129	0,138	0,141
Amount of repair	λ=0	λ=0,725	λ=1
Disparate Impact	0,435	0,834	0,998

Figure 11: Error rate, amount of repair and Disparate impact of the logistic regression classifier.

Random Forest classifier	None	Random	Total
Error rate	0,124	0,131	0,134
Amount of repair	λ=0	λ=0,75	λ=1
Disparate Impact	0,465	0,815	1,017

Figure 12: Error rate, amount of repair and Disparate impact of the random forest classifier.

The tradeoff in accuracy for fairness with the COMPAS recidivism dataset

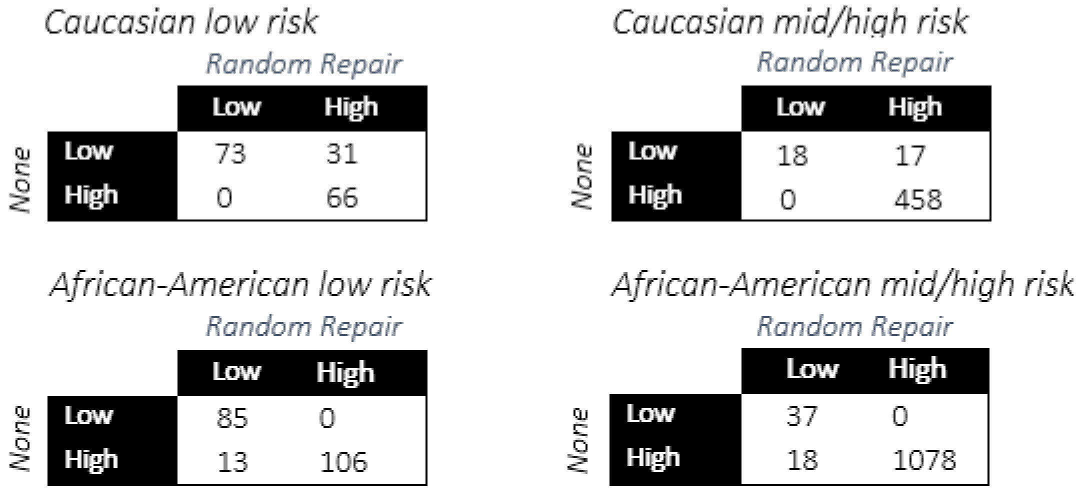
Logistic Regression classifier	None	Random	Trade-off
Disparate Impact	0,435	0,834	+0,399
Error rate	0,129	0,138	+0,009

Figure 13: The tradeoff in accuracy for fairness with random repair for the logistic regression classifier.

Random Forest classifier	None	Random	Trade-off
Disparate Impact	0,465	0,815	+0,350
Error rate	0,124	0,131	+0,007

Figure 14: The tradeoff in accuracy for fairness with random repair for the random forest classifier.

Classifications before and after random repair with the logistic regression classifier



Caucasian low risk

		Random Repair	
		Low	High
None	Low	73	31
	High	0	66

Caucasian mid/high risk

		Random Repair	
		Low	High
None	Low	18	17
	High	0	458

African-American low risk

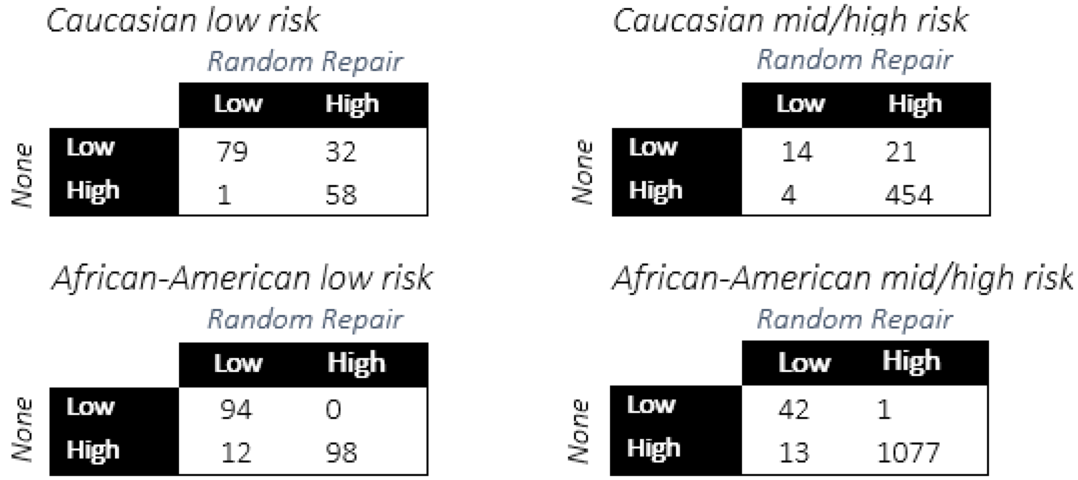
		Random Repair	
		Low	High
None	Low	85	0
	High	13	106

African-American mid/high risk

		Random Repair	
		Low	High
None	Low	37	0
	High	18	1078

Figure 15: Classifications before and after random repair with the logistic regression classifier.

Classifications before and after random repair with the random forest classifier



Caucasian low risk

		Random Repair	
		Low	High
None	Low	79	32
	High	1	58

Caucasian mid/high risk

		Random Repair	
		Low	High
None	Low	14	21
	High	4	454

African-American low risk

		Random Repair	
		Low	High
None	Low	94	0
	High	12	98

African-American mid/high risk

		Random Repair	
		Low	High
None	Low	42	1
	High	13	1077

Figure 16: Classifications before and after random repair with the random forest classifier.

5.1 Results with the adult income dataset

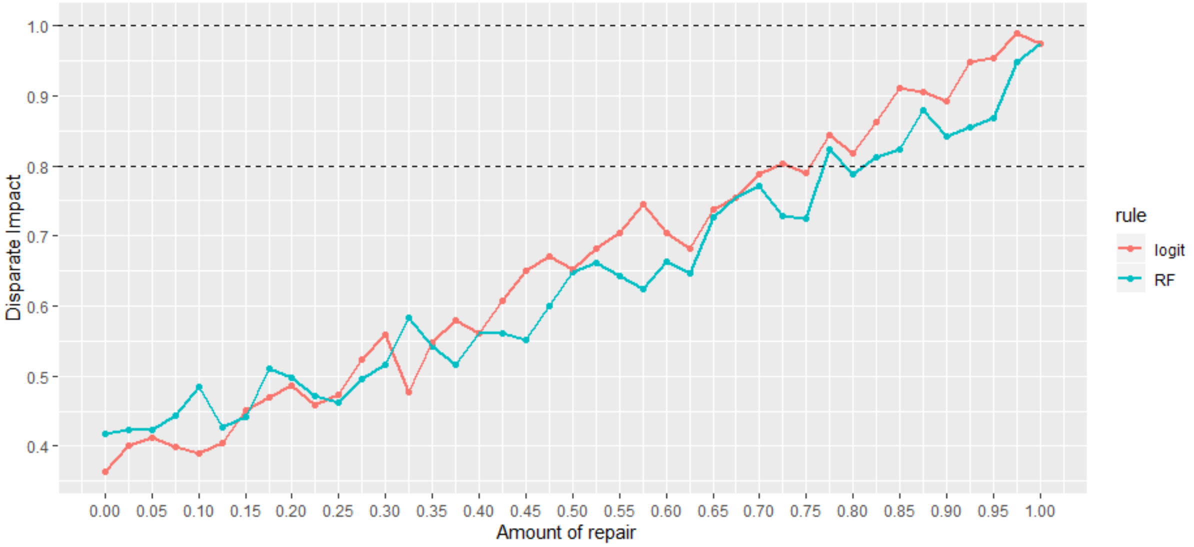


Figure 17: Disparate impact with random repair for the logistic regression (red) and the random forest classifier(blue).

Disparate Impact	None (λ=0)	Random (DI≈0.8)	Total (λ=1)
Logistic Regression classifier	0,364	0,803 (λ=0,725)	0,974
Random Forest classifier	0,418	0,824 (λ=0,775)	0,975

Figure 18: Amount of Disparate Impact for each form of repair.

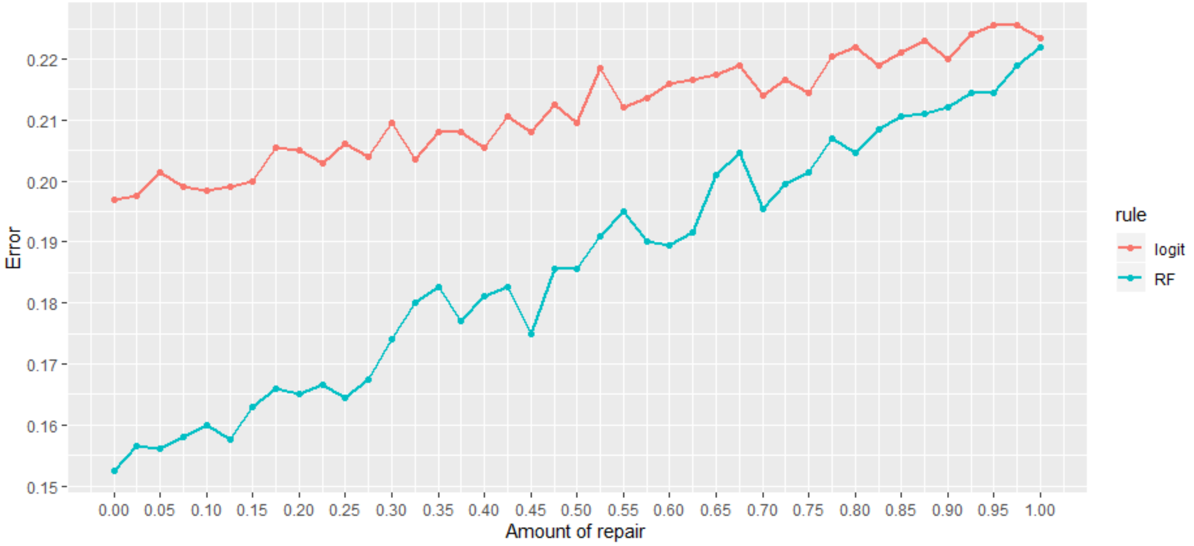


Figure 19: Classification errors of the classifications with logistic regression (red) and the random forest classifier(blue).

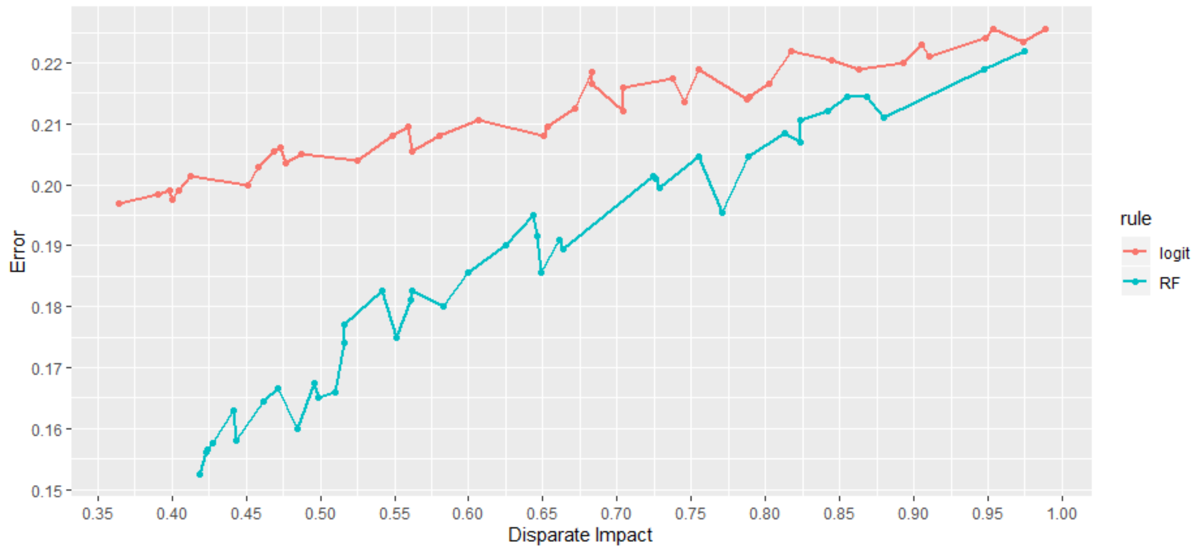


Figure 20: Classification errors of the classifications with logistic regression (red) and the random forest classifier(blue).

Logistic Regression classifier	None	Random	Total
Error rate	0,197	0,217	0,224
Amount of repair	λ=0	λ=0,725	λ=1
Disparate Impact	0,364	0.803	0,974

Figure 21: Error rate, amount of repair and Disparate impact of the logistic regression classifier.

Random Forest classifier	None	Random	Total
Error rate	0,153	0,207	0,222
Amount of repair	λ=0	λ=0,775	λ=1
Disparate Impact	0,418	0,824	0,975

Figure 22: Error rate, amount of repair and Disparate impact of the random forest classifier.

The tradeoff in accuracy for fairness with the adult income dataset

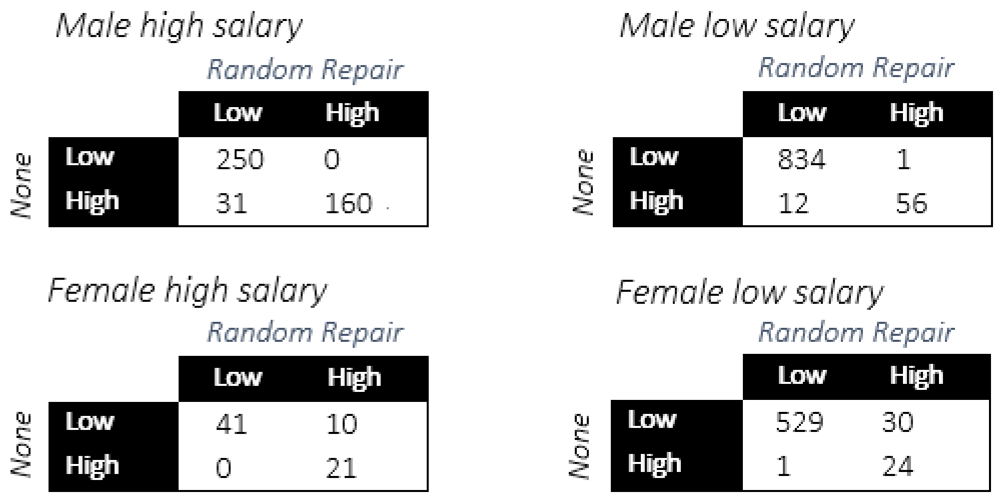
Logistic Regression classifier	None	Random	Trade-off
Disparate Impact	0,364	0,803	+0,439
Error rate	0,197	0,217	+0,020

Figure 23: The tradeoff in accuracy for fairness with random repair for the logistic regression classifier.

Random Forest classifier	None	Random	Trade-off
Disparate Impact	0,418	0,824	+0,406
Error rate	0,153	0,207	+0,054

Figure 24: The tradeoff in accuracy for fairness with random repair for the random forest classifier.

Classifications before and after random repair with the logistic regression classifier



Male high salary

		Random Repair	
		Low	**High**
None	**Low**	250	0
	High	31	160

Male low salary

		Random Repair	
		Low	**High**
None	**Low**	834	1
	High	12	56

Female high salary

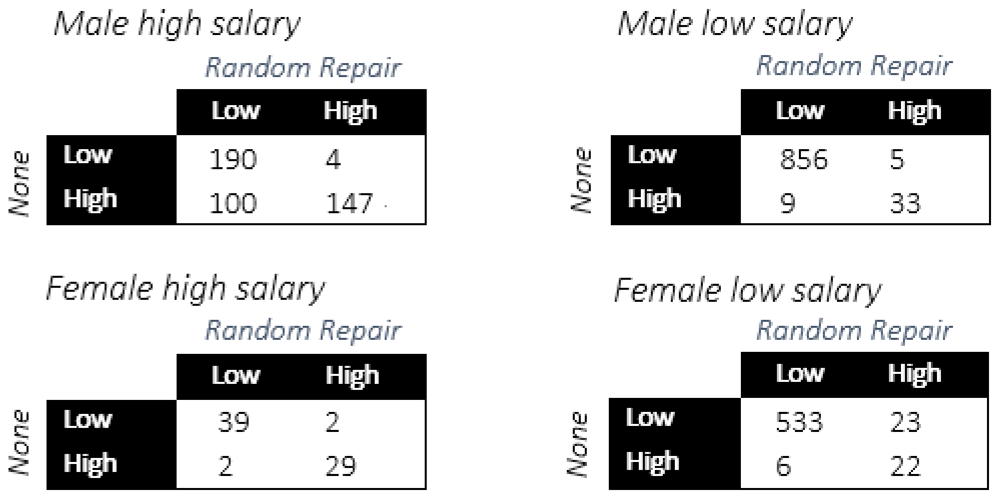
		Random Repair	
		Low	**High**
None	**Low**	41	10
	High	0	21

Female low salary

		Random Repair	
		Low	**High**
None	**Low**	529	30
	High	1	24

Figure 25: Classifications before and after random repair with the logistic regression classifier.

Classifications before and after random repair with the random forest classifier



Male high salary

		Random Repair	
		Low	**High**
None	**Low**	190	4
	High	100	147

Male low salary

		Random Repair	
		Low	**High**
None	**Low**	856	5
	High	9	33

Female high salary

		Random Repair	
		Low	**High**
None	**Low**	39	2
	High	2	29

Female low salary

		Random Repair	
		Low	**High**
None	**Low**	533	23
	High	6	22

Figure 26: Classifications before and after random repair with the random forest classifier.

6 Discussion

We have shown that random repair can be used on the adult income dataset and on the COMPAS recidivism dataset. But the classifiers used in our research where relatively simple compared to most classifiers used within the literature. The reason we didn't use more complex classifiers was to prevent overfitting. Its therefore unknown to us if similar results in tradeoff between fairness and classifier accuracy could be obtained on more complex classifiers.

There are multiple directions through which this research could be expanded to increase insight on the effects that random repair has on the fairness of machine learning systems:

- Applying random repair on other datasets to show its compatibility and usefulness in other cases.
- A comparison between more complex classifiers (e.g. support vector machines) to measure if their complexity allows a positive tradeoff between fairness and accuracy trough random repair.

7 Conclusion

Our research shows that random repair can effectively increase the fairness of machines learning systems to the desired amount of fairness. However, increasing the fairness trough random repair also decreases the classifier accuracy.

The increase in fairness for the COMPAS recidivism dataset resulted in more incorrect classified convicts for both classifiers. In a similar way the increase in fairness for the adult income dataset resulted in more incorrect classified individuals for both classifiers.

Since biased machine learnings systems could have serious negative effects for people with certain sensitive attributes, one could argue that the generally big increase in fairness is worth the generally small decrease in classifier accuracy. However, from our research, no clear conclusion can be drawn whether the increase in fairness is worth the decrease in classifier accuracy.

Although random repair can effectively increase the fairness in machine learning systems. It will be up to scientists, companies and others to decide whether they are willing to trade a certain loss in classifier accuracy against a certain increase in fairness.

References

1. R. S. Zemel, Y. Wu, K. Swersky, T. Pitassi, C. Dwork; *Learning Fair Representations.* Pages 1-2 ICML 2013.

2. Lydia T. Liu, Sarah Dean, Esther Rolf, Max Simchowitz, Moritz Hardt. *Delayed Impact of Fair Machine Learning.* Pages 1-2. ICML 2018.

3. Eustasio Del Barrio, Fabrice Gamboa, Paula Gordaliza, Jean-Michel Loubes. *Obtaining fairness using optimal transport theory.* Pages 1-17. 2018

4. F. Kamiran, T. Calders. *Data preprocessing techniques for classification without discrimination.* Pages 3-4 & 21-28. Knowledge and Information Systems, 2011.

5. J. Angwin, J. Larson, S. Mattu and L. Kirchner, ProPublica. *Machine Bias.* Larson et al. ProPublica, 2016

6. Pratik Gajane and Mykola Pechenizkiy. *On Formalizing Fairness in Prediction with Machine Learning.* Pages 1-5. 2017

7. Z. Zhong. *A Tutorial on Fairness in Machine Learning. 2018*

8. C. Dwork, M. Hardt, T. Pitassi, O. Reingold, R. Zemel. *Fairness Through Awarenes. Pages 1-5. 2011.*

9. Sahil Verma and Julia Rubin, FairWare. *Fairness Definitions Explained.* Pages 1-6. *ACM/IEEE* 2018.

10. M.B. Zafar, I. Valera, M. Gomez Rodriguez, K. Gummadi Fairness Constraints: Mechanisms for Fair Classification Pages 3-4 MPI-SWS 2017.

11 S. Ronagan, *AI Fairness Explanation of Disparate Impact Remover.* 2019.

12. *Adverse Impact Analysis of the Four-Fifths Rule*. Workplace Dynamics, LLC 2009.

13. P.T. *Artificial intelligence tool can identify gender and racial bias* 2019.

14. Aaron Roth. *Between "statistical" and "individual" notions of fairness in Machine Learning.* 2019.

15. Tim Brennan, William Dieterich and Beate Ehret. *Evaluating the predictive validity of the COMPAS risk and needs assessment system*. Pages 30-32. Northpointe Institute for Public Management Inc, 2009.

16. **Sahil Verma** and **Julia Rubin.** *Fairness Definitions Explained.* Pages 2-4. ACM/IEEE International Workshop on Software Fairness, 2018

17. Marco Cuturi and Arnaud Doucet. *Fast Computation of Wasserstein Barycenters*. Page 685-693 International Conference on Machine Learning, 2014.

18. James E Johndrow and Kristian Lum. An algorithm for removing sensitive information: application to race-independent recidivism prediction. arXiv:1703.04957, 2017.

Appendix

Classifications made in the COMPAS recidivism subset.

Classifications Logistic Regression Classifier				None	Random λ=0,725	Total	Difference None and Random repair
Caucasian	correct	classified	mid/high risk	458	475	479	+17
Caucasian	correct	classified	low risk	104	73	67	-31
Caucasian	incorrect	classified	mid/high risk	66	97	103	+31
Caucasian	incorrect	classified	low risk	35	18	14	-17
African-American	correct	classified	mid/high risk	1096	1078	1071	-18
African-American	correct	classified	low risk	85	98	101	+13
African-American	incorrect	classified	mid/high risk	119	106	103	-13
African-American	incorrect	classified	low risk	37	55	62	+18

Classifications Random Forest classifier				None	Random λ=0,75	Total	Difference None and Random repai
Caucasian	correct	classified	mid/high risk	458	475	481	+17
Caucasian	correct	classified	low risk	111	80	67	-31
Caucasian	incorrect	classified	mid/high risk	59	90	103	+31
Caucasian	incorrect	classified	low risk	35	18	12	-17
African-American	correct	classified	mid/high risk	1090	1078	1078	-12
African-American	correct	classified	low risk	94	106	107	+12
African-American	incorrect	classified	mid/high risk	110	98	97	-12
African-American	incorrect	classified	low risk	43	55	55	+12

Classifications made in the adult income subset.

Classifications Logistic Regression Classifier			None	Random λ=0,725	Total	Difference None and Random
Male	correct	classified low salary	835	846	853	+9
Male	correct	classified high salary	191	160	152	-31
Male	incorrect	classified low salary	250	281	289	+31
Male	incorrect	classified high salary	68	57	50	-9
Female	correct	classified low salary	559	530	518	-29
Female	correct	classified high salary	21	31	30	+10
Female	incorrect	classified low salary	51	41	42	-10
Female	incorrect	classified high salary	25	54	66	+29

Classifications Random Forest classifier			None	Random λ=0,775	Total	Difference None and Random
Male	correct	classified low salary	861	865	867	+4
Male	correct	classified high salary	247	151	130	-96
Male	incorrect	classified low salary	194	290	311	+96
Male	incorrect	classified high salary	42	38	36	-4
Female	correct	classified low salary	556	539	532	-17
Female	correct	classified high salary	31	31	27	-0
Female	incorrect	classified low salary	41	41	45	+0
Female	incorrect	classified high salary	28	45	52	+17