University of Stuttgart

Institute of Parallel and Distributed Systems
Applications of
Parallel and Distributed Systems

Pre-Processing Methods for Group Fairness in Machine Learning

Seminar Paper

Seminarbezeichung (SS2022)

Supervisor: Dennis Tschechlov

Pankhuri Verma

Stuttgart,30.08.2022

Pre-Processing Methods for Group Fairness in Machine Learning

Pankhuri Verma

st180247@stud.uni-stuttgart.de

Abstract. With an increase in the use of Machine Learning and Artificial Intelligence in the field of decision making, it is essential to consider the fairness of the outcome. Studies have proved that algorithmic decision-making is prone to biases even though there is no intention for it. Therefore, it is important to ensure that the decision-making system does not show discrimination toward a particular individual or group. This article initially discusses the different types of bias in algorithmic decision-making, their origin and ways to tackle them. The focus of this article will be on group biases and their real-world examples. Further, the different methods for group fairness i.e., Pre-Processing, In-Processing and Post-Processing are discussed. In this article, the prime focus is on the Pre-processing methods for group fairness. Therefore, the various fairness notions of algorithmic decision-making along with pre-processing methods for group fairness are discussed. In the end, various use cases of the pre-processing methods that were used to remove biases are discussed along with their results.

1 Introduction

Intelligent Systems are present in all aspects of our life. They are used as recommendation systems, hiring decisions, shopping websites etc. and have made our lives easier with their massive computational power and continuous availability. The need for an automated learning model is evident due to the belief that computers are more efficient than humans. First, algorithms have the capacity to incorporate far more data than humans can comprehend and take many more factors into account. Second, algorithms can do complicated calculations far more quickly than humans. Third, human judgments are also considered flawed, as biases can frequently be present. Consequently, it is popularly accepted idea that utilizing automated algorithms leads to unbiased results[14]. Unfortunately, this is not always the case since ML algorithms are not as accurate as one might think. It is wrong to assume that data injected into models is unbiased because this would imply that ML algorithms are free from bias. More specifically, a prediction model might actually be susceptible to bias(presence of prejudice or favouritism towards a particular individual or group [12]) since it learns from human-generated data and remembers prior biases. It is crucial to evaluate and improve the ethics of the decisions that these automated systems make because many of them have a substantial impact on people's lives. These

decision-making systems can give biased results and may lead to potential loss of an individual. Therefore, the issue of algorithm fairness has garnered enormous attention recently [12].

This report investigates the different types of biases that impact the machine learning model. It starts with a detailed description of what biases are and how they impact decision making in different aspects of life. Further, the various types of biases are explained briefly, with a focus on 'Data to Algorithm' bias. In the next section, the different fairness notions for examining the fairness quality have been discussed. These notions act as a rule book for identifying the fairness quality of the machine learning model. Furthermore, the different methods of fair machine learning—pre-processing, in-processing, and post-processing—are discussed briefly. For the scope of this seminar, only preprocessing methods are discussed in detail. The different types of preprocessing methods discussed in this report are: relabelling(massaging), resampling(reweighing), and fair representation(optimised preprocessing). The example use cases for each of the methods are discussed further in detail. The focus of the report is to use these preprocessing methods to eliminate bias in the dataset .

2 Biases in Data

Bias in data has impacted numerous systems and led to discriminatory results. One such application that has been highly discussed is the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS). The application helped judges decide whether there was a chance for the culprit to recommit a crime in future. Investigations into the fairness of the application gave biased result towards African American people [12], [5]. Similar findings have been made in other fields, such as the facial recognition software in digital cameras that overestimates Asians' blink rates or an AI system that judges beauty pageant winners but is biased towards contestants with darker skin tones [14]. These skewed the forecast results from hidden or ignored biases in the data or algorithms. There were other instances where machine learning models gave biased results against women employees in the workplace [2]. Additionally, an algorithm that would provide ads for jobs in the STEM sectors (Science, Technology, Engineering, and Math) displayed discriminating behavior. This ad was created with gender-neutral advertising delivery in mind. However, fewer women than men saw the commercial because of the gender gap in the trianing data. Although its initial and genuine intention was to be gender-neutral, this optimization system served advertisements in a discriminatory manner [14]. These biases mostly originate due to the data that is fed to the machine learning algorithm. Therefore, it is very essential that we analyse and transform the dataset so that fairness can be achieved.

2.1 Types of Bias

Biases exist in different forms and can impact the results of the machine learning model. They can be in the data, algorithm, or user experience. In scenarios where

the training data contains biases, the model trained on them also learns the biases, and the same is reflected in the prediction. Sometimes, even if the training data is not biased, the machine learning model gives biased results due to its design choices and learning from the previous dataset model [12].

In Reference [14], the authors prepare a complete list of different types of biases with their corresponding definitions. The major categories of biases are:

1. Data to Algorithm: This category deals with the biases in data that when fed to the machine learning algorithm results in biased outcomes [14].

2. Algorithm to User: This type of bias is a result of algorithmic outcomes and affects user behaviour [14].

3. User to Data: Most of the data collected for training ML models are user-generated. This bias originates due to the bias that is inculcated by the user in the dataset [14].

Below is a reiteration of the most important sources of Data to Algorithm bias introduced in [12].

2.1.1 Data to Algorithm:

1. Measurement Bias: This bias arises based on the way a particular feature is chosen, utilised and measured [12]. An evidence of this type of bias can be seen be seen in recidivism risk prediction tool COMPAS, which used prior arrest rates as proxy variables to quantify the level of ”riskiness” or ”crime”. This can be regarded as a mismeasured proxy. Minority groups are commonly regulated and policed, which results in greater arrest rates in those communities. There are differences in how these groups are evaluated and managed, therefore it is important to remember that just because people from minority groups have a greater rate of an arrest doesn't always imply that they are necessarily more dangerous.

2. Omitted Variable Bias: In this type of bias, essential features of the model are omitted out [12]. For example, a model that was used to predict the annual percentage rate at which customers would discontinue using a service saw an unusual behaviour in customers who were discontinuing their subscriptions even before the model's intended warning. The reason for the cancelling of subscriptions was the entry of a new competitor that provided the same facility at cheaper rates. But the model did not anticipate the appearance of this factor during the prediction. This is a typical example of omitted bias.

3. Representation Bias: The way we sample from a population when gathering data contributes to representation bias. Missing subgroups and irregularities in non-representative samples leads to lack of diversity in the population[12]. For example, Datasets like ImageNet that lack regional variety exhibit a clear bias in favor of Western cultures.

4. Aggregation Bias: This bias occurs when inaccurate generalizations are made about an individual depending on the group they belong to [12]. Clinical assistance tools are a prime illustration of this kind of prejudice. For example, think about diabetic patients who appear to have different morbidities depending on their gender and racial background. Particularly, the intricate differences across genders and ethnicities can be seen in HbA1c readings, which are frequently used to diagnose and monitor diabetes. A model that does not account for individual characteristics will therefore probably not be suitable for all racial and gender groupings in the population. Even when they are evenly represented in the training data, this is still true. Aggregation bias may emerge from any generalizations about the population's subgroups.

5. Linking Bias: Linking bias develops when network attributes derived from user connections, activities, or interactions are inconsistent and inaccurately reflect the users' actual behavior[12]. For example, writers demonstrate how social networks can be skewed toward low-degree nodes when just taking into account the network's links and ignoring the information and actions of its users.

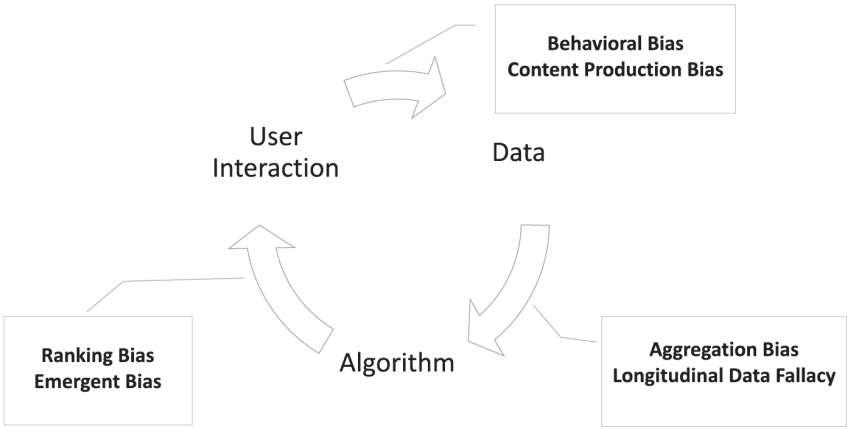

Fig. 1. Examples of bias definitions placed in the data, algorithm, and user interaction feedback loop. [12]

3 Fairness Notions for Group Fairness

Philosophy, psychology, and more recently, machine learning, all have a rich history of addressing bias and discrimination. But one must first define fairness in order to combat discrimination and bring about justice. Long before machine learning, philosophy and psychology sought to explain what fairness is. The difficulty of resolving this issue is demonstrated by the absence of a single, agreed-upon definition of fairness. It is challenging to develop a single definition of fairness that is acceptable to all parties involved because varied preferences and outlooks in various cultures favor different ways of viewing the concept. Even though fairness is a virtue that society values greatly, it may be surprisingly challenging to implement it. These notions are therefore defined by the experts in the domain and have to be agreed upon. These notions act as a rule book for defining fairness for a use case. Numerous fairness definitions have been offered in references [12], [14] to solve the various algorithmic bias and discrimination issues covered in the preceding sections.

Here is the description of the symbols that have been frequently used to define fairness notions.

A = protected attribute (e.g. race, gender), Y = class/output variable(0/1) based on class labels, $\hat{Y}$ = predicted value of Y and ϵ= threshold for probability distribution among groups(should be near to zero).

3.1 Disparate Impact

Disparate impact states that there should be a high ratio between the positive prediction rates of both groups. This will ensure that the number of positive predictions is similar across groups [14]. For example, if a positive prediction represents acceptance for a job, the proportion of accepted applicants should be similar across groups. This method is computed in the following manner:

$$\frac{P[\hat{Y} = 1|A \neq 1]}{P[\hat{Y} = 1|A = 1]} \geq 1 - \epsilon \tag{1}$$

3.2 Demographic Parity

A predictor is said to satisfy demographic parity, also known as statistical parity, if P($\hat{\text{Y}}$ |A = 0) = P($\hat{\text{Y}}$ | A = 1) is true. Regardless of the fact that a person is in the protected group, the possibility of a positive outcome should be the same [12].

3.3 Equalised Odds

A predictor $\hat{Y}$ satisfies equalized odds with respect to protected attribute A and outcome Y, if $\hat{Y}$ and A are independent conditional on Y. $P(\hat{Y} = 1|A = 0, Y =$

$y) = P(\hat{Y} = 1|A = 1, Y = y), y\epsilon\{0, 1\}$. [12] In other words, according to the equalized odds definition, the rates of true and false positives should be identical between the protected and unprotected groups.

3.4 Equal Opportunity

A binary predictor $\hat{Y}$ satisfies equal opportunity with respect to A and Y if P($\hat{Y}$=1|A=0,Y=1) = P($\hat{Y}$=1|A=1,Y=1). This implies that both members of protected and unprotected group should have a fair probability of being allocated to a positive result if they belong to a positive class [12].

4 Methods for Fair Machine Learning

The methods that were introduced previously for eliminating biases for machine learning models generally fall under three categories:

4.1 Pre-processing

Pre-processing methods strive to transform the data in a way that eliminates the underlying bias.

4.2 In-processing

Machine learning algorithms are modified as part of in-processing approaches in an effort to eliminate biases during model training. If a machine learning model's learning process can be altered, in-processing can be applied during model training by imposing a constraint or incorporating changes into the objective function.

4.3 Post-processing

After training, post-processing is carried out by utilizing a holdout set that was not used to train the model. The sole method that may be employed is post-processing, where the labels that the black-box model initially allocated are reassigned depending on a function during the post-processing stage.

For the scope of this seminar, the focus will be on pre-processing methods for group fairness in machine learning. The following section discusses the need to adopt pre-processing methods for achieving fairness. The pre-processing mechanisms involve changing the data before feeding it to the Machine Learning Model. This stage of pre-processing is essential to eliminate any bias at an initial stage so that the underlying Machine Learning model can give a fair prediction. It has been observed that even if the dataset is fair the underlying model becomes biased due to its design choices and the past data on which it is trained. Therefore, it is very important for the dataset to be fair to reduce the overall effect of bias [12].

5 Pre-Processing Methods for Group Fairness

In the following section, pre-processing techniques from several literatures have been discussed: [12], [14], [3], [**4**], [2], [10], [13] and [11]:

5.1 Relabelling

The goal of relabeling approach is to change the training set's ground truth values in a way that it satisfies the notion of fairness [6]. One of the methods of relabeling —'massaging'— will be discussed below.

Massaging: A proportion of individuals from the training data are selected, and their ground truth values are modified, using the preprocessing approach called data massaging [6], [11]. This enables any machine learning technique for classifying groups(to positive or negative class) to learn on a fair dataset with the goal of achieving group fairness. In this way the discrimination decreases, yet the overall class distribution is maintained. To do this, a ranker is used to rank the people according to how likely they are to get the result they seek. The higher a person ranks, the more likely a favorable outcome will be.

Let

$$\epsilon = P_1(Y = 1) - P_0(Y = 1) \tag{2}$$

denote the measured discrimination of the training data. The number M of required number of modifications in the dataset is calculated as:

$$M = \epsilon \times \frac{|D_1| \times |D_2|}{|D_1| + |D_2|} \tag{3}$$

where $D_1 = X|Z = 1$ and $D_0 = X|Z = 0$ denote the sets of privileged and unprivileged individuals respectively. Two candidate sets, called promotion(pr) and demotion (dem), are made that divides the dataset based on their class label and discriminatory variables. The promotion set is arranged in descending order, whereas the demotion set is ordered in ascending order. This is done in order the achieve fairness among groups. The people belonging to promotion set are the unprivileged ones and their class label needs to be changed to a positive class(class label should be promoted from negative to positive). Whereas, the people belonging to the demotion set are privileged and their class label should be changed to negative class(class label should be demoted from positive to negative)[11], [2]. This is done in order to achieve fairness for the unprivileged group.

By sorting both sets according to their ranks- the sets $pr = \{X \in D_0|Y = 0\}$ and $dem = \{X \in D_1|Y = 1\}$ are massaged. The top-M persons' labels in each sets are reversed (i.e. massaged), respectively, which are the M people who are closest to the decision border[8].

Example Usecase for Massaging Technique: Consider the example in Figure 2, which depicts a sample job-application relation with a positive class probability for both men and women. The goal is to learn a classifier that can forecast the class of objects for which predictions are made without discrimination to Sex = f(female). A Naive Bayesian classification model's positive class probability is used in this example to rank the objects. The data is arranged separately in the second phase (Figure 3) for male applicants with class + and female applicants with class - with respect to their positive class likelihood. This forms the ordered promotion and demotion candidates[11], [2], [10].

Here, $P_1(Y = 1)$ = Probability of males having a positive class = 4/5 and $P_0(Y = 1)$ = Probability of females having a positive class = 2/5.

We know that,

$$\epsilon = P_1(Y = 1) - P_0(Y = 1) \tag{4}$$

$$\epsilon = \frac{4}{5} - \frac{2}{5} = \frac{2}{5} \tag{5}$$

The number M of labels of promotion and demotion candidates that needs to be changed are:

D_1 = Set of privileged individuals(males) = 5 and D_2 = Set of unprivileged individuals(females) = 5

$$M = \epsilon \times \frac{|D_1| \times |D_2|}{|D_1| + |D_2|} \tag{6}$$

$$M = \frac{2}{5} \times \left(\frac{5 \times 5}{5 + 5}\right) = 1 \tag{7}$$

Since M =1, for the data to be discrimination-free, one change will be needed from the list of promotion candidates and one change from the list of demotion candidates. The top candidates for promotions and demotions (rows in Figure 3 marked in strong font) have their names changed. The discrimination in the dataset is reduced once the labels for this data are altered. The dataset then becomes discrimination-free and will be used for future classifier learning[11], [2], [10].

5.2 Resampling

Re-sampling techniques change the sampling rate of the training data by omitting or increasing certain samples, or by changing their relative importance at training time [6], [11], [2], [10]. One of the methods of resampling- 'Reweighing' is discussed below in detail.

Reweighing: The massaging method modifies the labels of the items, which causes disruption in the dataset. This drawback is not reflected in the second method-reweighing. In this method, the class labels are not renamed, but new weights are applied to them [11], [2], [10].

For instance, consider the objects with discriminatory variable D. D can take both privileged($\bar{d}$) and unprivileged values(d). If D = d and their predicted Class

Sex	Ethnicity	Highest Degree	Job Type	Cl.	Prob
m	native	h. school	board	+	98%
m	native	univ.	board	+	89%
m	native	h. school	board	+	98%
m	**non-nat.**	**h. school**	**healthcare**	**+**	69%
m	non-nat.	univ.	healthcare	-	30%
f	non-nat.	univ.	education	-	2%
f	**native**	**h. school**	**education**	**-**	40%
f	native	none	healthcare	+	76%
f	non-nat.	univ.	education	-	2%
f	native	h. school	board	+	93%

Fig. 2. Sample job-application relation with positive class probability. [11]

Sex	Ethnicity	Highest Degree	Job Type	Cl.	Prob
f	**native**	**h. school**	**education**	**-**	**40%**
f	non-nat.	univ.	education	-	2%
f	non-nat.	univ.	education	-	2%

Sex	Ethnicity	Highest Degree	Job Type	Cl.	Prob
m	**non-nat.**	**h. school**	**healthcare**	**+**	**69%**
m	native	univ.	board	+	89%
m	native	h. school	board	+	98%
m	native	h. school	board	+	98%

Fig. 3. Promotion candidates (negative objects with Sex = f in descending order) and demotion candidates (positive objects with Sex = m in ascending order). [11]

= +, then they will get higher weights than objects with D = d and Class = - . This is done because unprivileged people with positive label do not need any modification in class label. Similarly, objects with D = $\bar{d}$ and Class = + will get lower weights than objects with D = $\bar{d}$ and Class = - . This is done because privileged people with negative label do not need any modifications to their class label. The goal is to reduce the discrimination and maintain the overall positive class probability.

If the dataset X is not discriminant, i.e. D and Class are independent of each other, the expected probability is: $P_{exp}(d \wedge +)$

$$P_{exp}(d \wedge +) = d \times + \tag{8}$$

where d is the fragment of tuples with D = d and + the fragment of tuples having Class = +. But the actual probability is:

$$P_{act}(d \wedge +) = d \wedge + \tag{9}$$

where $d \wedge +$ represents the fragment of items with D = d and Class = +. In case the expected probability is more than the actual probability, we can incure that the bias is towards class - and D = d. Subsequently, weights are assigned to d with respect to class +. For a tuple x in the dataset X:

$$W(x(D) = d|x(Class) = +) = \frac{P_{exp}(d \wedge +)}{P_{act}(d \wedge +)} \tag{10}$$

This weight of d for class + will increase the weightage of items with D = d for the class +.

Further, the weight of d for class - will be

$$W(x(D) = d|x(Class) = -) = \frac{P_{exp}(d \wedge -)}{P_{act}(d \wedge -)} \tag{11}$$

and the weights of $\bar{d}$ for class + and - will be:

$$W(x(D) = \bar{d}|x(Class) = +) = \frac{P_{exp}(\bar{d} \wedge +)}{P_{act}(\bar{d} \wedge +)} \tag{12}$$

$$W(x(D) = \bar{d}|x(Class) = -) = \frac{P_{exp}(\bar{d} \wedge -)}{P_{act}(\bar{d} \wedge -)} \tag{13}$$

In this way a weight is assigned to every tuple according to its D and Class-values. The dataset with these weights becomes balanced. A discrimination-free classifier is learned based on this balanced dataset [11], [2], [10].

Example Usecase for Reweighing Technique: Consider the table in Figure 4, the weight for each data object is computed according to its D and Class-value. The weight of a data object with D = f(female) and Class + is calculated. We know that 50 percent objects have D = f and 60 percent objects have Class-value +, so the expected probability of the object should be [11]:

$$P_{exp}(Sex = f|x(Class) = +) = 0.5 \times 0.6 \tag{14}$$

but its actual probability is 20 percent.

$$P_{act}(Sex = f|x(Class) = +) = \frac{2}{10} \tag{15}$$

So the weight W will be:

$$W(Sex = f|x(Class) = +) = \frac{0.5 \times 0.6}{0.2} = 1.5 \tag{16}$$

Similarly the weights of the other combinations are:

$$W(Sex = f|x(Class) = -) = 0.67 \tag{17}$$

$$W(Sex = m|x(Class) = +) = 0.75 \tag{18}$$

$$W(Sex = m|x(Class) = -) = 2 \tag{19}$$

These weights are assigned to the individual tuples to eliminate discrimination, and, furthermore, based on these new weights a fair classifier is learned.

Sex	Ethnicity	Highest Degree	Job Type	Cl.	Weight
m	native	h. school	board	+	0.75
m	native	univ.	board	+	0.75
m	native	h. school	board	+	0.75
m	non-nat.	h. school	healthcare	+	0.75
m	non-nat.	univ.	healthcare	-	2
f	non-nat.	univ.	education	-	0.67
f	native	h. school	education	-	0.67
f	native	none	healthcare	+	1.5
f	non-nat.	univ.	education	-	0.67
f	native	h. school	board	+	1.5

Fig. 4. Sample job-application relation with weights. [11]

5.3 Fair Representations

Finding fair representations is a strategy that is connected to the idea of representation learning. For a biased dataset D, an alternative, cleaned(reduced discrimination) dataset $\hat{D}$ is built that masks the original bias while being as similar to the original data as possible [3].

Optimized Pre-Processing: The above mentioned preprocessing methods do not provide generic, principled optimization frameworks for balancing discrimination control and data utility, nor do they provide linkages to be formed via probabilistic descriptions to the larger statistical learning and information theory literature. Individual distortion or fairness is not explicitly stated, which is another problem [3].

A probabilistic framework for discrimination-preventing pre-processing in supervised learning is introduced by the optimization method described in the works of [3], which also fills in the gaps in the above mentioned pre-processing literature. This type of pre-processing optimizes the trade-off between discrimination control, data utility, and individual distortion. Unlike distortion, which is controlled per-sample, discrimination and utility are controlled at the level of probability distributions, that increases fairness and limits the impact of the modification on individuals. Fig. 5 shows the proposed preprocessing method as part of a supervised learning pipeline. As part of the pipeline for predictive learning with discrimination prevention, the learn mode is applied to training data and the apply mode is applied to novel test data. This method will be further used in detail to overcome the discrimination issue in COMPAS Tool [3],[8].

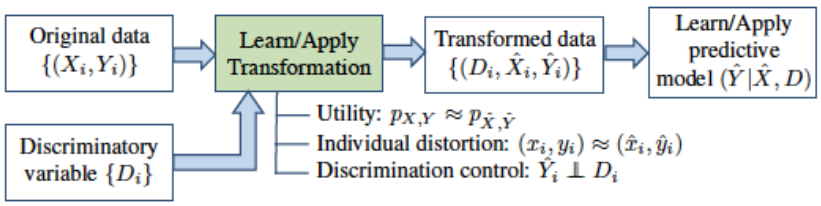

Fig. 5. Pipeline for discrimination prevention using predictive learning. Learn mode is applied to training data and apply mode to test data. [3].

General Formulation A dataset consisting of n i.i.d. samples $\{(D_i, X_i, Y_i)\}_{i=1}^n$ from a joint distribution $p_{D,X,Y}$ with domain $D \times X \times Y$ is used for this method. Here, D is a set of discriminating factors like race and gender, X is a set of other non discriminating factors utilized in decision-making, and Y is the outcome variable.

A randomized mapping $p_{\hat{X},\hat{Y}|X,Y,D}$ is determined that (i) transforms the existing dataset into a new dataset $\left\{\left(D_i, \hat{X}, \hat{Y}\right)\right\}_{i=1}^n$, which is utilised to train a model, and (ii) transforms test data also. Here, $\hat{X}$ is the set of transformed discriminatory variables and $\hat{Y}$ is the set of outcome variables. For test data, Y_i is

not available at the input and $\hat{Y}_i$ is not needed at the output. Therefore, a reduced mapping $p_{\hat{X}|X,D}$ is used, which is obtained from $p_{\hat{X},\hat{Y}|X,Y,D}$ by marginalizing over $\hat{Y}$ and Y after weighting by $p_{Y|X,D}$ [3]. Each $\left(\hat{X},\hat{Y}\right)$ is taken independently from the domain $X \times Y$ as X, Y by applying $p_{\hat{X},\hat{Y}|X,Y,D}$ to the matching/ corresponding triplet (D_i, X_i, Y_i). Since D_i is retained as-is, it is not included in the mapping that has to be determined.

Discrimination Control: Discrimination control limits the dependence of $\hat{Y}$ on the discriminatory variable D, expressed by the conditional distribution $p_{\hat{Y}|D}$. There are two formulations for this. The first one requires $p_{\hat{Y}|D}$ to be close to p_{Y_T} for all values of D,

$$J\left(p_{\hat{Y},D}\left(y|d\right), p_{Y_T}\left(y\right)\right) \leq \epsilon_{y,d} \forall d \in D, y \in \{0,1\} \tag{20}$$

where $J\left(.,.\right)$ denotes some distance function.The second formulation constrained $p_{\hat{Y},D}$ to be similar for the two values of D.

$$J\left(p_{\hat{Y},D}\left(y|d_1\right), p_{\hat{Y},D}\left(y|d_2\right)\right) \leq \epsilon_{y,d_1,d_2} \tag{21}$$

for all $d_1, d_2 \in D, y \in \{0,1\}$. d1 and d2 are two different discriminatory variables like gender or race for which discrimination control should be checked. The choice of target p_{Y_T} in (20), and distance J and thresholds ϵ in (20) and (21) should be decided by the experts in the domain. The instantiation of (20) should involve consultation with domain experts and stakeholders before being put into practice. Here, the value of J is as follows [3]:

$$J\left(p, q\right) = \left|\frac{p}{q} - 1\right| \tag{22}$$

where, p and q are two different probability distributions.

Distortion Control: Distortion Control states that the mapping $p_{\hat{X},\hat{Y}|X,Y,D}$ should satisfy the distortion constraints for the domain $X \times Y$. These constraints restrain the mapping from reducing or avoiding some significant modifications. Given a distortion metric $\delta : (X \times Y)^2 \rightarrow R_+$, the conditional expectation of the distortion is as follows [3]:

$$E\left[\delta\left(\left(x, y\right), \left(\hat{X},\hat{Y}\right)\right)|D = d, X = x, Y = y\right] \leq c_{d,x,y} \forall (d, x, y) \in D \times X \times Y \tag{23}$$

It is assumed that $\delta(x, y, x, y) = 0 \ \forall (x, y) \in (X, Y)$ and $c_{d,x,y}$ is the level of control.

Utility Preservation: Utility preservation describes that the distribution of $(\hat{X}, \hat{Y})$ is statistically close to the distribution of (X,Y). This makes sure that a model learned from the transformed dataset is identical to the one learned from the original dataset. And the dissimilarity measure $\Delta(p_{\hat{X},\hat{Y}}, p_{X,Y})$(e.g. KL Divergence) between both the probability distributions should be minimal.[3].

While transforming the dataset, a good tradeoff should be maintained between the three parameters-Discrimination Control, Distortion Control, Utility Preservation-in order to get fair results [3].

Example Usecase for Optimised Preprocessing: The section deals with the example usecase of optimised preprocessing on COMPAS dataset. The term "recidivism" describes someone who relapses to criminal activity. According to research, about two-thirds of US prisoners who are released are later arrested. Therefore, it is crucial to comprehend the recidivistic tendencies of those behind bars who are being considered for release at various stages of the criminal justice system (bail hearings, parole, etc.). For this purpose, automated risk scoring systems have been created and are currently utilized in US courtrooms, particularly the exclusive COMPAS tool.

An article on racial prejudice in the COMPAS algorithm was recently published by ProPublica, along with a dataset that contains information on COMPAS risk scores, recidivism rates, and other pertinent factors. The COMPAS algorithm tends to give African-American people higher scores, which is a reflection of the group's higher rate of recidivism as a whole. The article goes on to show that African Americans and Caucasian Americans have different false positive and false negative rates [3]. The charts in figure 6 show that scores for white defendants were skewed toward lower-risk categories. Scores for black defendants were not. Figure 8 shows the false positive and false negative rates for black defendants and white defendants, respectively. It is observed that black defendants have a high false positive rate of 44.85 percent, and white defendants have a high false negative rate of 47.72 percent.

For investigation, severity of crime, number of previous crimes, and age were selected as the decision variables from ProPublica's dataset. The outcome variable was a binary variable with values reoffended or not reoffended. And race and gender were set as the discriminatory variables. The encoding of the decision and discrimination variables is described in Figure 7. The dataset contained around 5k records [3], [7], [9], [4] and [1].

For getting optimal results, c(level of control) was set as 0.5 and ϵ(distance measure threshold) as 0.1 for the experiment. The optimal value of utility measure (KL divergence) was 0.021. In order to evaluate if discrimination control was achieved as expected, the dependence of the outcome variable on the discrimination variable before and after the transformation was checked. Figure 9 shows the results before and after the transformation. It can be seen that the probability of a person re-offending a crime given the discriminatory variables has been reduced after the transformation(highlighted in the figure) [3].

Fig. 6. These charts show that scores for white defendants were skewed toward lower-risk categories. Scores for black defendants were not. (Source: ProPublica analysis of data from Broward County, Fla.) [7]

Feature	Values	Comments
Recidivism (binary)	$\{0, 1\}$	1 if re-offended, 0 otherwise
Gender	{Male, Female}	
Race	{Caucasian, African-American}	Races with small samples removed
Age category	$\{< 25, 25 - 45, > 45\}$	years of age
Charge degree	{Felony, Misdemeanor}	For the current arrest
Prior counts	$\{0, 1 - 3, > 3\}$	Number of prior crimes

Fig. 7. ProPublica dataset features. [3]

All Defendants			Black Defendants			White Defendants		
	Low	High		Low	High		Low	High
Survived	2681	1282	Survived	990	805	Survived	1139	349
Recidivated	1216	2035	Recidivated	532	1369	Recidivated	461	505
FP rate: 32.35			FP rate: 44.85			FP rate: 23.45		
FN rate: 37.40			FN rate: 27.99			FN rate: 47.72		

Fig. 8. False Positive and False Negative rates of Black Defendants and White Defendants [3].

D	Before transformation		After transformation	
(gender, race)	$p_{Y\|D}(0\|d)$	$p_{Y\|D}(1\|d)$	$p_{\hat{Y}\|D}(0\|d)$	$p_{\hat{Y}\|D}(1\|d)$
F, A-A	0.607	0.393	0.607	0.393
F, C	0.633	0.367	0.633	0.367
M, A-A	0.407	0.593	0.596	0.404
M, C	0.570	0.430	0.596	0.404

Fig. 9. Dependence of the outcome variable on discriminatory variable before and after the transformation. F and M indicate Female and Male, and A-A, and C indicate African-American and Caucasian respectively[3].

6 Conclusion

In this report, the problems that can adversely affect AI systems in terms of bias and unfairness were discussed. Further examples of the potential real-world problems that unfairness can have on society are elaborated. The different biases and fairness notions were further discussed in detail. Different methods i.e. pre-processing, in-processing and post processing are introduced briefly. The main focus of this report is on pre-processing methods for group fairness. Different pre-processing methods like relabelling, resampling and fair representation have been discussed. All the methods have been supported with examples and implemented on real world problems. One of the methods of fair representation(Optimised Pre-Processing) has been discussed in detail. COMPAS datasets have been analyzed using a data-driven optimization framework to reduce algorithmic discrimination. During the analysis of the original and transformed datasets, interesting discrimination patterns were revealed, and corrective adjustments were made to control discrimination while maintaining data utility. To conclude, since machine learning algorithms are integral to our everyday lives, demanding that these intelligent systems are fair is difficult to achieve. In addition, there should be a constant effort not only to develop more accurate algorithms but also to reduce bias in the data.

References

[1] Julia Angwin and Jeff Larson. *Bias in Criminal Risk Scores Is Mathematically Inevitable, Researchers Say*. https://www.propublica.org/article/bias-in-criminal-risk-scores-is-mathematically-inevitable-researchers-say. Dec. 2016.

[2] Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. "Building Classifiers with Independency Constraints". In: *2009 IEEE International Conference on Data Mining Workshops*. 2009, pp. 13–18. DOI: 10.1109/ICDMW.2009.83.

[3] Flavio Calmon et al. "Optimized Data Pre-Processing for Discrimination Prevention". In: (Apr. 2017). URL: https://arxiv.org/pdf/1704.03354.pdf.

[4] *COMPAS Recidivism Risk Score Data and Analysis*. https://www.propublica.org/datastore/dataset/compas-recidivism-risk-score-data-and-analysis. Aug. 2022.

[5] Sam Corbett-Davies et al. "Algorithmic Decision Making and the Cost of Fairness". In: *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2017, pp. 797–806. DOI: 10.1145/3097983.3098095. URL: https://doi.org/10.1145/3097983.3098095.

[6] Jannik Dunkelau and Michael Leusche. *Fairness-aware machine learning: An Extensive Overview*. Universität Düsseldorf, 2019.

[7] Lauren Kirchner Eff Larson Surya Mattu and Julia Angwin. *How We Analyzed the COMPAS Recidivism Algorithm*. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm. May 2016.

[8] Moritz Hardt, Eric Price, and Nathan Srebro. "Equality of Opportunity in Supervised Learning". In: (Oct. 2016).

[9] Surya Mattu Julia Angwin Jeff Larson and ProPublica Lauren Kirchner. *Machine Bias*. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. May 2016.

[10] Faisal Kamiran and Toon Calders. "Classifying without discriminating". In: *2009 2nd International Conference on Computer, Control and Communication*. 2009, pp. 1–6. DOI: 10.1109/IC4.2009.4909197.

[11] Faisal Kamiran and Toon Calders. "Data Pre-Processing Techniques for Classification without Discrimination". In: *Knowledge and Information Systems* 33 (Oct. 2011). DOI: 10.1007/s10115-011-0463-8.

[12] Ninareh Mehrabi et al. "A Survey on Bias and Fairness in Machine Learning". In: *ACM Comput. Surv.* 54.6 (July 2021). URL: https://doi.org/10.1145/3457607.

[13] Hussein Mouzannar, Mesrob I. Ohannessian, and Nathan Srebro. "From Fair Decision Making To Social Equality". In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, 2019, pp. 359–368. ISBN: 9781450361255. DOI: 10.1145/3287560.3287599. URL: https://doi.org/10.1145/3287560.3287599.

[14] Dana Pessach and Erez Shmueli. "A Review on Fairness in Machine Learning". In: *ACM Comput. Surv.* 55.3 (Feb. 2022). URL: https://doi.org/10.1145/3494672.