HUMBOLDT-UNIVERSITÄT ZU BERLIN

MASTER'S THESIS

FairGridSearch: A Framework to Compare Fairness-Enhancing Models

---

Author:
Shih-Chi Ma
Student no.:
609131

1st Examiner:
Prof. Dr. Benjamin Fabian

2nd Examiner:
Prof. Dr. Stefan Lessmann

Submitted for acquisition of the degree Master of Science (M.Sc.)
in Economics and Management Science
at the School of Business and Economics of Humboldt-Universität zu Berlin

Berlin, March 30, 2023

HUMBOLDT-UNIVERSITÄT ZU BERLIN

Abstract

School of Business and Economics

Master's Thesis

FairGridSearch: A Framework to Compare Fairness-Enhancing Models

by Shih-Chi Ma

Machine learning (ML) models are being increasingly used in critical decision-making applications. However, these models are susceptible to replicating or even amplifying bias present in real-world data. In the relevant literature, a wide range of bias mitigation methods and base estimators exist, but selecting the optimal model for a particular dataset or application remains challenging. This paper proposes a novel framework, called FairGridSearch, to compare fairness-enhancing models. The framework enables experimentation with models using various parameter combinations and subsequently recommends the best model. The study also addresses four research questions related to metrics, base estimator, classification threshold, and accuracy-fairness trade-off by applying FairGridSearch to three popular datasets Adult, COMPAS, and German Credit. The results show that the choice of accuracy and fairness metric can significantly impact the evaluation of the model. Additionally, different base estimators and classification threshold values also affect the effectiveness of bias mitigation methods and fairness stability respectively, but the effects are not consistent across all datasets. Furthermore, the study found that there is no clear trade-off between accuracy and fairness for the datasets experimented in this study. Given findings from the choice of base estimator and classification threshold, it is recommended for future studies on fairness in machine learning to encompass a broader range of factors when building fair models, beyond bias mitigation methods alone.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Machine Learning (ML) models are increasingly utilized in numerous critical decision-making applications, such as workforce recruiting (Zhao et al., 2018, Buyl et al., 2022), justice risk assessments (Angwin et al., 2016, Tolan et al., 2019), and credit risk prediction (Kozodoi et al., 2022, Kumar et al., 2022). Despite the fact that ML algorithms are not intentionally designed to incorporate bias, recent studies have demonstrated that machine-learning models not only reproduce existing biases in the training data (Bolukbasi et al., 2016) but can also amplify them (Zhao et al., 2017, Foulds et al., 2019, Hall et al., 2022). These lines of evidence and concerns about algorithmic fairness have led to a surge of interest in the literature on defining, evaluating, and improving fairness in ML algorithms. The authors from Hort et al. (2022) found 391 bias mitigation methods with 49 base estimators such as Logistic Regression (LR) and Random Forest (RF) in the literature, where most works particularly address mitigating bias in binary classification models.

The availability of numerous base estimators and bias mitigation methods, however, poses the challenge of selecting the optimal approach for a particular dataset or application. Although several comparison studies have been conducted (Hufthammer et al., 2020, Hort et al., 2021, Chen et al., 2023), they have not provided a clear best model recommendation. In addition, such studies typically focus only on comparing different bias mitigation methods, leaving out some essential aspects of ML modeling. For instance, most studies involve only a limited set of base estimators, which are often trained using arbitrary parameters. Additionally, varying classification threshold values are typically not considered, most studies use the standard threshold 0.5, while different threshold values might be more suitable. Furthermore, evaluating bias mitigation methods necessitates taking both accuracy and fairness into account, and different fairness metrics may be required for various situations. However, many studies do not thoroughly discuss the selection of appropriate fairness metrics (Hufthammer et al., 2020, Hort et al., 2021, Chen et al., 2023). To address this research gap, this paper

proposes a grid search framework FairGridSearch[1] for comparing fairness-enhancing models. This framework enables the experimentation of models with various parameter combinations and subsequently recommends the best model. The framework supports a broad range of parameter-tuning options, including six base estimators, their corresponding parameters, classification thresholds, and nine bias mitigation methods. Furthermore, it provides flexibility in selecting various accuracy and fairness metrics to evaluate the models. The term "fairness-enhancing models" in this study refers to all models that are considered in the comparison when taking fairness into account, with or without bias mitigation methods.

Given that many previous works only consider a small set of metrics, base estimators, and classification thresholds (Hamilton and Friedler, 2017, Roth, 2018, Hufthammer et al., 2020), their impacts on model fairness may not be thoroughly discussed. To address this gap, this study conducts FairGridSearch on three prominent datasets - Adult, COMPAS, and German Credit - and, in addition to providing optimal model recommendations, performs various analyses to answer the following research questions:

1. **RQ1: (Metrics)**
   Does the choice of accuracy and fairness metric affect model evaluation?

2. **RQ2: (Base Estimator)**
   Does the choice of base estimator influence model fairness?

3. **RQ3: (Classification Threshold)**
   How do classification threshold values impact fairness?

4. **RQ4: (Trade-Off)**
   Does an empirical trade-off between accuracy and fairness exist?

The subsequent sections of this paper are organized as follows: Section 2 starts with a review of the pertinent literature that compares various bias mitigation models. Additionally, this section lays out the fundamental aspects of fairness in ML, covering the various sources and types of bias in the ML pipeline, definition and measurement of fairness, and various existing bias mitigation methods. In Section 3, the FairGridSearch framework is introduced. This section describes the framework's structure and parameters in detail. Section 4 provides a detailed description of an exemplary experiment conducted using FairGridSearch. This section commences with the selection of datasets and proceeds with descriptions and analyses of the results for each and across

[1] Implementation of the framework is written with python and can be found on GitHub repo: `https://github.com/dorisscma/FairGridSearch`

datasets. Section 5 aims to provide an interpretation of the results and answer the research questions. In Section 6, the study's limitations are discussed. Finally, Section 7 concludes the paper.

Chapter 2

Literature Review

2.1 Related Work

The growing interest in fairness in ML has sparked relevant research including definitions, measurements, and bias mitigation methods. In their survey, Hort et al. (2022) found more than 100 unique bias mitigation methods and 111 unique fairness metrics. In particular, a variety of recent work addresses mitigating bias in binary classification models (Hort et al., 2022). Several preliminary works therefore set up empirical comparisons between bias mitigation methods with the aim to understand which methods are best for use (Pessach and Shmueli, 2022). Hamilton and Friedler (2017) was one of the fist attempt in this field, the authors compare four fairness metrics using four algorithms across three datasets and conclude that none of the approaches appear to be easily and simply applicable across datasets. Across all of the datasets Roth (2018) include in the study, the author found out that both Disparate Impact Remover (DIR) (Feldman et al., 2015) and Fair Accuracy Maximizer (Zafar et al., 2017b) would be an acceptable algorithm to use for fairness maximization. Furthermore, the author also identifies Logistic Regression (LR) as base estimator is the most versatile, which yields high scores in measures of accuracy and fairness as well as a decent average run time.

Another study by Friedler et al. (2018) has provided a benchmark analysis of several fairness-aware methods with eight different fairness metrics and five datasets. They find that although different algorithms perform especially well on specific formulations of fairness, many of the metrics strongly correlate with one another. Biswas and Rajan (2020) also evaluate seven bias mitigation methods on real-world Machine Learning models from Kaggle. Based on their results, the authors claim that there exists a trade-off between accuracy and fairness, and post-processing algorithms have most competitive replacement.

Chakraborty et al. (2020) also compare three existing bias mitigation approaches with Fairway, the new method proposed by them. They exclusively rely on LR models and include five different datasets and two fairness metrics in the comparison.

TABLE 2.1: Studies with Evaluation on different Bias Mitigation Methods

Study	# Datasets	# BM	# Acc_metrics	# fair_metrics	# base
Hamilton and Friedler (2017)	3	3	2	4	2
Roth (2018)	4	3	3	2	4
Friedler et al. (2018)	5	4	4	8	4
Biswas and Rajan (2020)	5	7	2	7	varies
Chakraborty et al. (2020)	5	4	2	2	1
Hufthammer et al. (2020)	1	2	1	4	2
Hort et al. (2021)	3	8	1	2	3
Chen et al. (2023)	5	17	11	4	4
FairGridSearch	3	9	6	8	6

Hufthammer et al. (2020) conduct an empirical analysis for Reject Option Classifier (ROC) and Prejudice Remover (PR) on a binary classification task. Their preliminary results show that both methods can improve fairness with rather low cost on accuracy.

More recent works on comparing different methods include Hort et al. (2021) and Chen et al. (2023). In their paper, Hort et al. (2021) introduce Fairea, an approach to benchmark bias mitigation methods, they then report a large-scale empirical study to test the effectiveness of 12 bias mitigation methods. On top of the comparison, they also propose a strategy that quantifies the fairness-accuracy trade-off. Built on their work, Chen et al. (2023) evaluate 12 unique bias mitigation methods on five different datasets with 11 accuracy metrics and four fairness metrics, making this study one of the most comprehensive one in the existing literature.

However, prior research has limitations in terms of the number of fairness or accuracy metrics, bias mitigation methods, or machine learning base estimators evaluated. This study addresses this gap by incorporating several bias mitigation methods, base estimators and a comprehensive set of metrics. While the number of bias mitigation methods and accuracy metrics in this work is less than that of Chen et al. (2023), FairGridSearch covers more fairness metrics and allows for testing the methods with six different base estimators, including a transformer model, which has become one of the most popular machine learning models in recent years. Most importantly, after comparing all the models, FairGridSearch selects the best one based on both accuracy and group fairness metrics. An overview of previous studies and the proposed framework is presented in Table 2.1.

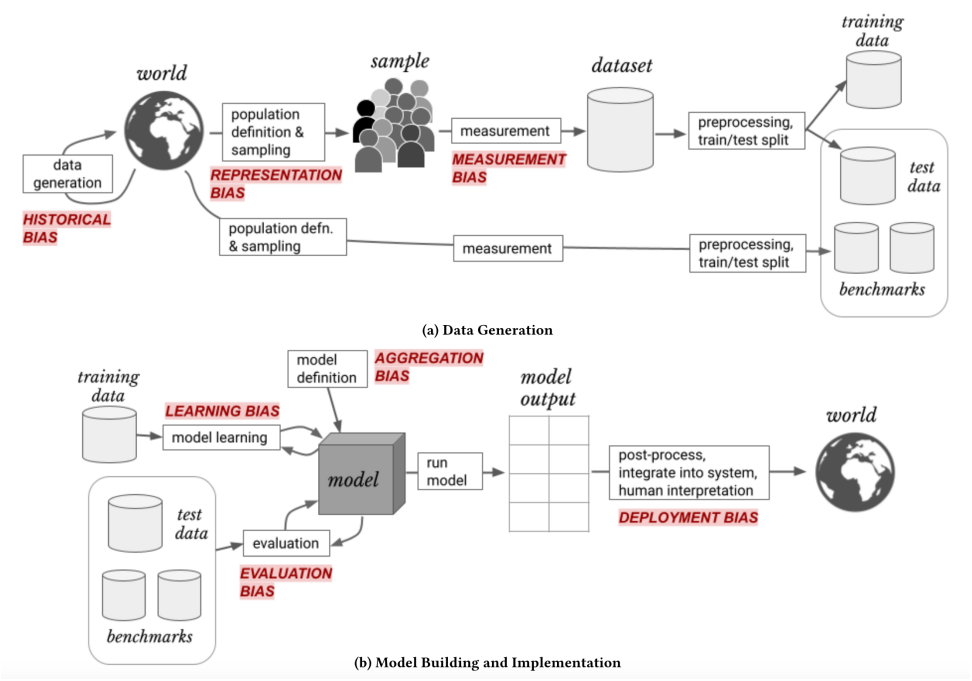
2.2 Types of Bias

Bias exist in many shapes and forms, some of which can lead to unfairness in different down-stream learning tasks (Caton and Haas, 2020). Several studies have discussed

the potential sources of harm introduced by bias in ML. For instance, Ntoutsi et al. (2020) define bias as "the inclination or prejudice of a decision made by an Artificial Intelligence (AI) system which is for or against one person or group, especially in a way considered to be unfair", and categorize causes of socio-technical bias into data generation, data collection, and institutional bias; Suresh and Guttag (2021) provide a framework that identifies seven potential sources of downstream harm arising in distinct stages of machine learning pipeline, from data generation, model building, evaluation, and deployment processes; the authors from Olteanu et al. (2019) concentrate on social data and prepare a complete list of different types of biases occurring from data origins, data collection and data processing; finally, based on the last two work from Suresh and Guttag (2021) and Olteanu et al. (2019), Mehrabi et al. (2022) introduce a different bias categorization according to the data, algorithm, and user interaction loop.

This paper follows the categorization based on different phases in ML pipeline and briefly review distinct types of bias accordingly, depicted in figure 2.1. Here, the focus lies on the ones that could lead to unfairness in classification tasks.

FIGURE 2.1: Types of Bias in ML Pipeline (Suresh and Guttag, 2021)



2.2.1 Data Collection

- **Historical Bias:** Historical bias arises even if data is perfectly measured and sampled, if the world as it is or was leads to a model that produces harmful outcomes (Suresh and Guttag, 2021). In other words, human biases already exist in the real world where data is collected from, therefore even if the data reflects the world perfectly and accurately, models built on top of it can still inflict harm. An example of such bias can be seen word embeddings. Garg et al. (2018) illustrate how word embeddings capture stereotypes toward gender and ethinic groups, for example adjectives like *honorable, ascetic, amiable* are most associated to men and *maternal, romantic, submissive* to women, or *sensitive, passive, complacent* most associated with Asians. Furthermore, based on the temporal analysis, they also find out that models trained on data from a particular decade reflect biases of that time, adjectives like *intelligent* and *logical* on average have increased in association with women over time, especially after the women's movement in the 1960s.

- **Representation Bias:** Representation bias occurs when the development sample under-represents some part of the population, and subsequently fails to generalize well for a subset of the use population (Suresh and Guttag, 2021). Non-representative samples lack the diversity of the population, with missing subgroups and other anomalies (Mehrabi et al., 2022). According to Suresh and Guttag (2021), potential causes for representational bias include: (1) when defining the target population, if it does not reflect the use population, (2) when defining the target population, if contains underrepresented groups, and (3) when sampling from the target population, if the sampling method is limited or uneven such as methods that contain self-selection bias, for example, for an opinion poll that measures enthusiasm for a political candidate, the most enthusiastic supporters are more likely to complete the poll (Mehrabi et al., 2022).

- **Measurement Bias:** Measurement bias occurs when choosing, collecting, or computing features and labels to use in a prediction problem (Suresh and Guttag, 2021). In machine learning or statistics, when a concept is not directly observable or hard to be quantified, a proxy is chosen to approximate it. For example, in the recidivism risk prediction tool COMPAS, prior arrests and friend/family arrests were used as proxy variables to measure level of "riskiness" or "crime". However, historically minority communities are controlled and policed more frequently, resulting in biased higher arrest rates for them. In general, measurement

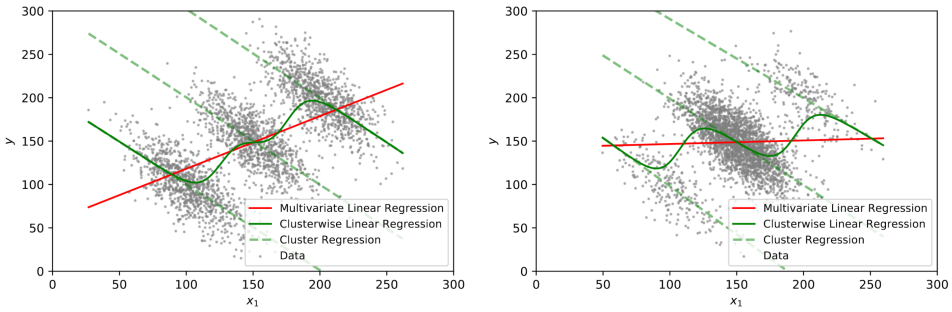
bias can arise when (1) the method of measurement is not identical for different groups, (2) the proxy is an oversimplification of a complex construct, (3) the accuracy of measurement varies across groups (Suresh and Guttag, 2021).

- **Annotator Bias:** Annotator bias, also known as label bias, arises for data sources that admit automatic labeling, manual labeling of data without adequate domain expertise and data poisoning attack by adversaries (Chatterjee et al., 2021). During manual labeling process, the annotators may transfer their prejudices to the data, and further to models trained with the data (Hellström et al., 2020).

2.2.2 Model building

- **Aggregation Bias:** Aggregation bias arises when a one-size-fits-all model is used for data in which there are underlying groups or types of examples that should be considered differently (Suresh and Guttag, 2021). As shown in figure 2.2, wrongly aggregating three sub-groups of data each with negative trend gives rise to misleading positive relationship (left) or no relationship (right), when one of the sub-groups is sampled more frequently.

FIGURE 2.2: Illustration of Aggregation Bias (Mehrabi et al., 2022).



- **Learning Bias:** In the absence of intentional interventions, a trained machine learning model can and does amplify undesirable biases in the training data (Hooker, 2021). When building a machine learning model, choosing the objective function that the algorithm learns to optimize during training is necessary, and bias can arise when the defined objective doesn't suit the problem, especially when the real objective and the chosen one are not compatible.

2.2.3 Evaluation

- **Evaluation Bias:** Evaluation bias occurs when the benchmark data used for a particular task does not represent the use population (Suresh and Guttag, 2021).

The main purpose of evaluation is to quantitatively compare models against each other, and using inappropriate benchmarks may overrate models that perform well only the benchmark data; another source of evaluation bias is the the choice of metrics that are used to report performance, looking at different metrics can induce different results.

2.2.4 Model Deployment

- **Development Bias:** Deployment bias arises when there is a mismatch between the problem a model is intended to solve and the way in which it is actually used (Suresh and Guttag, 2021). If an algorithm designed with certain assumptions made about society is deployed in unsuited applications, even if it was trained and developed free from bias, the results can be tainted by deployment biases when put into action. Selbst et al. (2019) outline the mismatch with the concept "framing trap", resulted from failure to model the entire system over which a social criterion will be enforced, and "portability trap" where repurposing algorithmic solutions designed for one social context may do harm when applied to a different one, whereas transferable code is often encouraged in the field of computer science and thus designs are first aimed to create tools independent of social context.

2.3 Algorithmic Fairness

Fighting against bias and discrimination has a long history in philosophy and psychology, and only recently in machine-learning (Mehrabi et al., 2022), and in order to identity discrimination and achieve fairness, defining fairness is of great importance. Yet, finding suitable definitions of fairness in an algorithmic context is a subject of much debate (Verma and Rubin, 2018), and the fact that no universal definition of fairness exists shows the difficulty of solving this problem (Saxena, 2019). Saxena et al. (2019), for instance, interpret fairness as the "absence of any prejudice or favoritism towards an individual or a group based on their intrinsic or acquired traits in the context of decision-making"; while Farnadi et al. (2018) believe that looking solely at attributes of individuals is not enough, and "taking into account the social, organizational, and other connections between individuals is important."

In the field of computer science, existing fairness criteria fall under different categories: group fairness and individual fairness (Zemel et al., 2013). Group fairness measures are based on statistical parity between members of different groups based

on protected attributes (or sensitive attributes), such as gender or race. It ensures that the overall positive or negative decisions across the groups are similar (Kamishima et al., 2012). Individual fairness, on the other hand, claims that if the distance between any two individuals is sufficiently small, that is, they are "similar", then they should receive the same outcomes (Dwork et al., 2012).

2.3.1 Group Fairness

Many of the proposed fair machine learning metrics have groundings in statistics (Chouldechova and Roth, 2018), and the use of statistical measures is attractive, because they are relatively simple to measure and definitions built using statistical measures can usually be achieved without having to make any assumptions on the underlying data distribution (Carey and Wu, 2022). In literature, several group fairness criteria have been proposed, most of which are concerned with properties of the joint distribution of the sensitive attribute S, the target variable Y, and the classifier score R. R represents the predicted score and falls within $[0, 1]$, with binary classifier prediction $\hat{Y} = \mathbb{1}\{R > t\}$, and t is the classification threshold. These group fairness notions typically require certain group-conditional quality metrics to be the same for all sensitive feature groups, and different choices of group-conditional quality metric led to different terminology of the corresponding group fairness notions (Speicher et al., 2018).

Abstract Concept

To simplify the landscape of fairness criteria, Barocas et al. (2019) define three representative categories that most proposed criteria in the literature fall into, based on different (conditional) independence among the random variables: independence, sufficiency and separation.

Independence Independence aims for classifiers to make their scoring independent of the group membership (Caton and Haas, 2020). Statistically, this condition is satisfied if the sensitive attribute is independent of the score (Barocas et al., 2019):

$$R \perp S \tag{2.1}$$

That is to say, the share of positive and negative classifications is equivalent among the sensitive groups. To give a practical example, consider a case that a bank is issuing a loan, and $R = 1$ means the loan is approved (acceptance), Independence demands the approval rate to be the same among all groups.

However, this is often not very realistic in certain real-world cases since it might harm the performance of the model and consequently make the business unsustainable (Kozodoi et al., 2022). Hence, a relaxation that allows a difference between the two rates, either in the form of subtraction or ratio, is introduced. A common practice is the "80%" rule (Feldman et al., 2015), which is applied in the Equal Opportunity Credit Act (ECOA) and the guideline of the Equal Employment Opportunity Commission (EEOC).

Some variants of the Independence criterion include Statistical Parity, also called Demographic Parity (Dwork et al., 2012), and Conditional Statistical Parity (Corbett-Davies et al., 2017).

Separation An extension of the Independence property is Separation. Compared to Independence, the Separation criterion acknowledges that the sensitive attribute S may be correlated with the target variable Y in many scenarios. As an instance, certain groups might have generally higher default rates on loans than others. Thus, the definition of separation demands the sensitive attribute to be "conditionally independent" of the predicted output given the true value (Barocas et al., 2019):

$$R \perp S \mid Y \tag{2.2}$$

Metrics such as Equal Opportunity (Hardt et al., 2016), Equalized Odds (Hardt et al., 2016), and Predictive Equality (Chouldechova, 2016) are all in the category of separation.

Sufficiency Sufficiency, on the other hand, formalizes that the score already subsumes the sensitive characteristic for the purpose of predicting the target (Barocas et al., 2019), and therefore requires the true output value to be conditionally independent of the sensitive attribute given the predicted output:

$$Y \perp S \mid R \tag{2.3}$$

This implies consistency of the positive and negative predictive values across all groups. Namely, the overall share of “correct” decisions should be the same for the sensitive groups. Some metrics measuring sufficiency are Conditional Procedure Accuracy (Berk et al., 2021), Predictive Parity (Chouldechova, 2016), and Calibration within Groups (Chouldechova, 2016).

Noteworthy Group Fairness Metrics

Due to the growing attention on algorithm fairness, more than twenty different notions of fairness were suggested in the last few years (Verma and Rubin, 2018), and most of them are based on the binary classification confusion matrix (Caton and Haas, 2020). The following section reviews 1) a few important concepts from confusion matrix, and afterwards, 2) some of the most widely used group fairness metrics. More criteria proposed in the literature can be found in the overview table 2.2.

Binary Classification Confusion Matrix Basics:

- True Positives (TP): Actual positives correctly predicted as positives.
- True Negatives (TN): Actual negatives correctly predicted as negatives.
- False Positives (FP): Actual positives wrongly predicted as negatives.
- False Negatives (FN): Actual negatives wrongly predicted positives

Positive Predictive Value (PPV): also known as precision, the fraction of True Positives out of all predicted positives, formulated as $\frac{TP}{TP+FP}$. PPV equals to 1-FDR, where FDR stands for False Discovery Rate.

False Omission Rate (FOR): the fraction of False Negatives out of all predicted negatives, formulated as $\frac{FN}{TN+FN}$.

Statistical Parity: one of the earliest definitions for fairness, also often called Demographic Parity. Statistical Parity requires equal probability of being classified as positive regardless of an instance's group membership (Dwork et al., 2012):

$$P(\hat{Y} = 1 \mid S = s_1) = P(\hat{Y} = 1 \mid S = s_2) \tag{2.4}$$

In other words, instances in both privileged and unprivileged groups should have the same probability of being classified to the positive class.

Disparate Impact: very similar to statistical parity, but comes in the form of ratio with a relaxation that allows a certain degree of difference in the probabilities as mentioned in section 2.3.1. According to Feldman et al. (2015), Disparate Impact exists if:

$$\frac{P(\hat{Y} = 1 \mid S = s_1)}{P(\hat{Y} = 1 \mid S = s_2)} \leq \tau, \tag{2.5}$$

with τ equals to 0.8 under the common 80% rule.

Conditional Statistical Parity: first suggested by Kamiran et al. (2013), Conditional Statistical Parity allows the use of a set of legitimate attributes in the decision-making process, formulated as follow:

$$P(\hat{Y} \mid L = \ell, S = s_1) = P(\hat{Y} \mid L = \ell, S = s_2), \tag{2.6}$$

where $\ell \in L$ is the set of legitimate features being conditioned on.

Equalized Odds: designed by Hardt et al. (2016), "a predictor $\hat{Y}$ satisfies equalized odds with respect to protected attribute S and outcome Y, if $\hat{Y}$ and S are independent conditional on Y." Formally written as:

$$\begin{cases} P(\hat{Y} = 1 \mid S = s_1, Y = 1) = P(\hat{Y} = 1 \mid S = s_2, Y = 1) \\ P(\hat{Y} = 1 \mid S = s_1, Y = 0) = P(\hat{Y} = 1 \mid S = s_2, Y = 0) \end{cases} \tag{2.7}$$

The measure requires the classifier to have equal true-positive rates (TPRs) and false-positive rates (FPRs) for both privileged and unprivileged groups.

Equal Opportunity: also designed by Hardt et al. (2016), Equal Opportunity is a relaxation of Equalized Odds, focusing only on equal true-positive rates (TPRs):

$$P(\hat{Y} = 1 \mid S = s_1, Y = 1) = P(\hat{Y} = 1 \mid S = s_2, Y = 1) \tag{2.8}$$

Predictive Parity: also known as Calibration or Test Fairness, Predictive Parity requires the positive predictive values to be similar across both privileged and unprivileged groups:

$$P(Y = 1|R = r, S = s_1) = P(Y = 1|R = r, S = s_2) \tag{2.9}$$

The main idea behind this definition is that the fraction of correct positive predictions should be the same for both genders (Verma and Rubin, 2018).

2.3.2 Individual Fairness

Except from the group fairness criteria discussed above, there exist another set of fairness metrics that consider fairness as it relates to each participating individual. Individual fairness is a comparative notion of fairness in that it asks whether there are any differences in the way that similar people are being treated (Barocas et al., 2019). First

TABLE 2.2: Group Fairness Criteria Overview

Criterion	CR	Reference
Darlington criterion (4)	IND	Darlington (1971)
Indenpendence	IND	Calders et al. (2009)
Statistical parity	IND	Dwork et al. (2012)
Group fairness	IND	Dwork et al. (2012)
Conditional statistical parity	IND	Corbett-Davies et al. (2017)
Darlington criterion (3)	SP	Darlington (1971)
Equal opportunity	SP	Hardt et al. (2016)
Equalized odds	SP	Hardt et al. (2016)
Balance for the negative class	SP	Kleinberg et al. (2016)
Balance for the positive class	SP	Kleinberg et al. (2016)
Avoiding disparate mistreatment	SP	Zafar et al. (2017a)
Predictive equality	SP	Chouldechova (2016)
Equalized correlations	SP	Woodworth et al. (2017)
Conditional procedure accuracy	SP	Berk et al. (2021)
Treatment equality	SP	Berk et al. (2021)
Cleary model	SF	Cleary (1968)
Darlington criterion (1), (2)	SF	Darlington (1971)
Predictive parity	SF	Chouldechova (2016)
Calibration within groups	SF	Chouldechova (2016)
Well calibration	SF	Kleinberg et al. (2016)
Conditional use accuracy	SF	Berk et al. (2021)

*source: *Fairness and machine learning: Limitations and Opportunities*, Barocas et al. 2019 Barocas et al. (2019), *Fairness in Credit Scoring: Assessment, Implementation and Profit Implications*, Kozodoi et al. 2021 Kozodoi et al. (2022), *The statistical fairness field guide: perspectives from social and formal sciences*, Carey et al. 2022 Carey and Wu (2022). Abbreviations: CR = Closest relative, IND = Independence, SP = Separation, SF = Sufficiency

formalized in the paper from Dwork et al. (2012), using Lipschitz conditions on the classifier, their definition requires the distributions over outcomes of two individual x and y to be indistinguishable only up to their distance $d(x, y)$. However, instead of providing a continuous measurement of fairness, this notion only returns a binary answer whether the fairness conditions are satisfied or not. In the following, some of the most common individual fairness metrics in the literature are reviewed.

Consistency (CNS) Zemel et al. (2013) proposed a notion that measures how similar the labels are for similar instances, the consistency score. The consistency score is formally defined as:

$$yNN = 1 - \frac{1}{Nk} \sum_{n} |\hat{y}_n - \sum_{j \in kNN(x_n)} \hat{y}_j|, \tag{2.10}$$

where x is the feature vector and y is the actual label. It can be seen that this notion compares the model's classification prediction of the given data x to its k-nearest neighbors, $kNN(x)$, the kNN function is applied to the full set of examples to obtain the most accurate estimate of each point's nearest neighbors.

Counterfactual Fairness: as stated in Kusner et al. (2017), a causal model with latent variables U is counterfactually fair if under any context $X = x$ and $S = s$,

$$P(\hat{Y}_{S \leftarrow s}(U) = y \mid X = x, S = s) = P(\hat{Y}_{S \leftarrow s'}(U) = y \mid X = x, S = s), \tag{2.11}$$

for all y and for any value s' attainable by S, this means changing S while holding other variables which are not causally dependent on S constant will not change the distribution of $\hat{Y}$.

Generalized Entropy Index (GEI): motivated by information theory, Speicher et al. (2018) define Generalized Entropy Index (GEI) as:

$$GEI = \frac{1}{n\alpha(\alpha - 1)} \sum_{i=1}^{n} \left[(\frac{b_i}{\mu})^{\alpha} - 1 \right], \tag{2.12}$$

where $\alpha \notin \{0, 1\}$ is a constant parameter and b_i the "benefit" standing for differences in an individual's prediction and the actual label with mean μ. Another common individual fairness criteria Theil Index (TI) is just a special case of GEI when $\alpha = 1$.

2.3.3 Summary on Fairness Criteria

Although the literature has defined a myriad of notions to quantify fairness, each measures and emphasizes different aspects of what can be considered "fair" (Caton and Haas, 2020). And several previous studies have shown that it is difficult to satisfy some of the group fairness constraints at once except in highly constrained special cases (Mehrabi et al., 2022), or even impossible (Berk et al., 2021, Kleinberg et al., 2016, Pleiss et al., 2017).

In addition, the group fairness criteria generally provide no guarantee for fairness at the individual level either, Dwork et al. (2012) demonstrate the trade-off between Statistical Parity and individual fairness in their work: in some cases, in order to maintain Statistical Parity, the classifier are forced to make positive predictions for members from the protected group even when they are under-qualified. This paper follows the instruction on the selection of fairness criteria by Saleiro et al. (2019), they design a "Fairness Tree" in collaboration with policymakers to provide guidance in linking different fairness metrics and the real-world problem in hand, more details on the selection of fairness metrics in accordance with different use cases is discussed in section 3.1.3.

Building on the metrics discussed here, the following section will continue to review some existing methods to incorporate these fairness definitions in the binary classification case.

2.4 Bias Mitigation

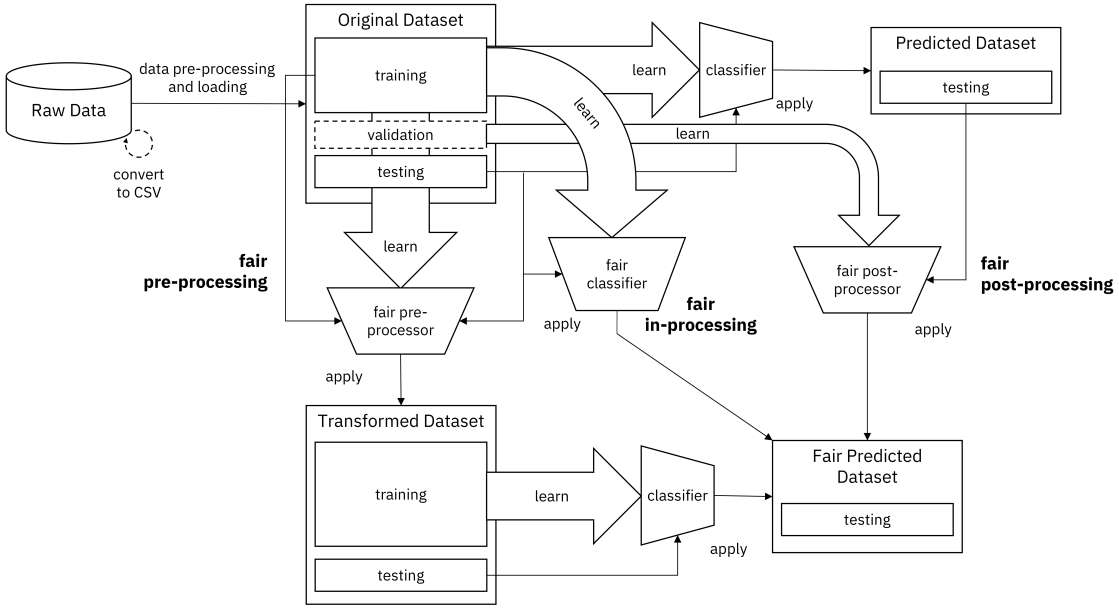
The growing demand for reducing bias that Machine Learning systems exhibit has encouraged the new field of technical research. Several attempts have been undertaken to incorporate the concept of fairness into the machine learning pipeline, with different approaches and choices of fairness criteria.

These fairness-enhancing interventions which mitigate or even eliminate certain kinds of bias are generally termed as "fairness processors" or "bias mitigators." The intervention can take place at different stages of the machine learning pipeline and, depending on the phase they alter, three categories: pre-processing, in-processing, and post-processing can be applied accordingly. The choice of the algorithm depends on both the conceptual conformity and technical feasibility: conceptual-wise, different fair processors incorporate different concepts of fairness. For example, some seek optimization towards the Independence fairness criteria, while others strive to improve Separation. Technical feasibility wise, pre-processors require the users' accessibility to the training data, in-processors ask for allowance for the change in algorithms, and

post-processors can still be employed even only the test prediction from a learned model is available.

Figure 2.3 illustrates the general fairness modelling pipeline outlined by AIF360 (Bellamy et al., 2018). More detail about the package can be found in section 2.5. The rectangles represent data output that is either generated during or after the process, and all intermediate and final outputs share the same protected attributes; the trapezoids, on the other hand, speak for models that make predictions on the test data, including the fair processors and the traditional machine learning models that are necessary for pre-, and post-processors.

FIGURE 2.3: The fairness modeling pipeline (Bellamy et al., 2018).



2.4.1 Pre-Processing

Pre-processing models modify the training data before feeding it into the model so that the model's input is fair concerning the protected attributes. The most common methods entail the transformation of instance weights or feature space from training data. For example, as a preliminary mechanism in this field, Reweighing Calders et al. (2009) adjusts the weights of the training data while keeping the features and label values unchanged. Disparate Impact Remover Feldman et al. (2015), on the other hand, focuses on alteration in feature space, modifying features in the dataset so that the distributions for both privileged and unprivileged groups become similar; another approach fair representation learning Zemel et al. (2013) pre-processes the datasets in a

latent space. The following explains the two pre-processing bias mitigators included in FairGridSearch, Reweighing and Learning Fair Representation, in more detail.

Reweighing (RW)

Reweighing assigns weights to each instance in the training data based on the overall probabilities of the group-class combinations (Calders et al., 2009), with instances in the unprivileged group having positive label receiving a higher weight since this is less likely compared to those in privileged group, weights for all kinds of instances are calculated as follow:

$$\begin{cases} W(S = s_1 \mid y = 1) = \dfrac{P(y = 1)P(S = s_1)}{P(y = 1 \cap S = s_1)} \\ W(S = s_1 \mid y = 0) = \dfrac{P(y = 0)P(S = s_1)}{P(y = 0 \cap S = s_1)} \\ W(S = s_2 \mid y = 1) = \dfrac{P(y = 1)P(S = s_2)}{P(y = 1 \cap S = s_2)} \\ W(S = s_2 \mid y = 0) = \dfrac{P(y = 0)P(S = s_2)}{P(y = 0 \cap S = s_2)} \end{cases} \tag{2.13}$$

These weights are used to balance the original biased training dataset, in the training process of classification models, a higher instance weight causes higher losses when it's mis-classified (Hort et al., 2022). Based on the computed weights, a fair training set is re-sampled with replacement such that combinations with a higher weight reappear more often, which helps to fulfill the Independence criterion (Kozodoi et al., 2022)).

Learning Fair Representations (LFR)

Representation learning aims at learning a transformation of training data such that bias is reduced while maintaining as much information as possible (Hort et al., 2022). First proposed by Zemel et al. (2013), they formulate fairness as an optimization problem of finding a good representation of the data with two objectives: (1) to encode the data as well as possible, while (2) simultaneously obfuscating any information about membership in the protected group. Overall, the learning system find the optimal latent representation Z by minimizing the following objective:

$$L = A_z \cdot L_z + A_x \cdot L_x + A_y \cdot L_y \tag{2.14}$$

where A_z, A_x, A_y are hyper-parameters governing the trade-off between the system desiderata; while the three loss functions denoting three different aims: L_z stands for

group fairness (statistical parity), L_x constrains the mapping to Z to be a good description of dataset feature X, and L_y requires the predictions to be as accurate as possible.

In other words, the aim of this new representation is to lose information that can identify whether the person belongs to the protected subgroup, while retaining as much other information as possible (Zemel et al., 2013). In general, LFR is a pre-processing technique, but it can also be used as an in-processing method by utilizing the learned target coefficients Bellamy et al. (2018). FairGridSearch hence includes both as options for bias mitigation methods.

2.4.2 In-Processing

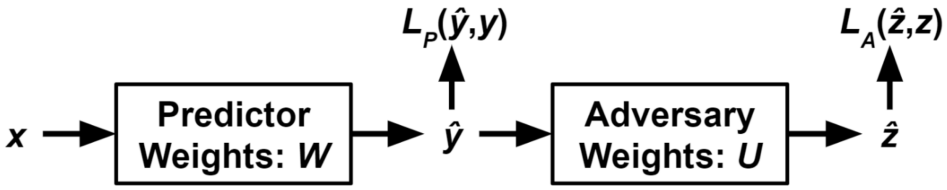
The in-processing models generally design a new learning algorithm, e.g. adding fairness regularization terms, such that bias is mitigated during the training procedure of the algorithm. By doing so, the training process involves both the minimization of the empirical loss and the optimization of the fairness criteria (Kozodoi et al., 2022). One disadvantage for in-processors in the past is that many of them are only able to handle single protected attributes, which typically limits their generality (Barocas et al., 2019). Nevertheless, it can been seen from Table 2.3 that there are more and more in-processors that could deal with multiple protected attributes proposed in recent years. Another drawback of in-processing methods is the requirement of full access to both input data and training process, which hinders its application in strictly regulated sectors like credit scoring, where regulatory approval might be necessary (Kozodoi et al., 2022). As before, the following section looks closely into the two in-processing methods included in FairGridSearch, and refer to a current survey by Wan et al. (2022) that investigate on 38 in-processing approaches for interested readers.

Adversarial Debiasing (AD)

Adversarial learning makes use of an "adversary" that determines whether a model training algorithm is robust enough (Goodfellow et al., 2014), and in the context of fairness, Zhang et al. (2018) proposed Adversarial Debiasing (AD), seeking to determine whether the training process is fair. This approach utilizes two stacks of neural networks with opposite goals on top of each other, the predictor and adversary. The first stack "predictor" is trained to make predictions, and the second stack "adversary" takes the output layer of the the predictor and tries to predict the sensitive attribute S. In short, the main idea of Adversarial learning is to maximize the predictor's prediction accuracy and simultaneously reduces the adversary's ability to predict the protected attribute from the predictions. The adversary may have different inputs depending on

the fairness definition needed to be achieved, in their paper (Zhang et al., 2018), consideration of three fairness metrics were showcased: Demographic Parity, Equality of Odds, and Equality of Opportunity. Figure 2.4 shows the architecture of the adversarial network proposed by Zhang et al. (2018).

FIGURE 2.4: The architecture of Adversarial Debiasing (Zhang et al., 2018).



Exponentiated Gradient Reduction (EGR)

Exponentiated Gradient Reduction (EGR) reduces fair classification to a sequence of cost-sensitive classification problems, whose solutions yield a randomized classifier with the lowest (empirical) error subject to the desired constraints (Agarwal et al., 2018). This method utilize concepts from Game Theory (Freund and Schapire, 1996) and define a "reduction" that treats the fair classification algorithm as a sequential "game" between two players, where at each step one player maximizes accuracy and the other player imposes a particular amount of fairness (Berk et al., 2021). For the definitions of fairness, their work focuses on Statistical Parity and Equalized Odds, but also encompasses many other previously studied definitions as special cases.

2.4.3 Post-Processing

Lastly, the post-processing methods are applied after preliminary predictions from unconstrained models are obtained, these methods generally adjust the labels in accordance with the chosen fairness criteria. For example, Hardt et al. (2016) propose to flip some decisions of a classifier to enhance Equalized Odds or Equalized Opportunity, and Corbett-Davies et al. (2017) suggest to select different thresholds for privileged and unprivileged groups such that accuracy and fairness are maximized.

As mentioned before, compared to pre-, and in-processors, which demand access to input data and training process, the advantage of post-processors is that only they can be used when the original models can only be treated as a black box, that is, post-processing can still be deployed when it's not possible to modify the training data or the learning algorithm. Nonetheless, this generality comes with a price: according to

Barocas et al. (2019), post-processing algorithms are often less effective than the other methods and are accompanied by a higher cost of prediction accuracy. Below the two post-processing methods included in FairGridSearch are reviewed.

Reject Option Classifier (ROC)

Reject Option Classifier (ROC), exploits the low confidence region of a single or an ensemble of probabilistic classifiers for discrimination reduction (Kamiran et al., 2012). Traditionally, a learned probabilistic classifier assigns instances with probabilities belonging to a class, and prediction probabilities closer to 0 or 1 are considered of high certainty. In binary classification, labels are assigned to a class when the corresponding probability exceeds a certain threshold, therefore, those predictions closer to the decision boundary imply a lower certainty. Kamiran et al. (2012) believe that predictions falling into the so-called low confidence region are ambiguous and influenced by biases, hence, to reduce discrimination, ROC rejects them and gives favorable outcomes to unprivileged groups and unfavorable outcomes to privileged groups. This region is defined as:

$$\max \left[p\left(C^{+} \mid X\right), 1 - p\left(C^{+} \mid X\right) \right] \leq \theta, \tag{2.15}$$

where $\theta \in (0.5, 1)$. ROC can then be interpreted as a cost-based prediction method in which the cost or loss of misclassifying a unprivileged group instance as negative is $\theta/(1-\theta)$ times that of misclassifying it as positive (Kamiran et al., 2012).

Calibrated Equalized Odds (CEO)

Calibrated Equalized Odds (CEO) processor proposed by Pleiss et al. (2017) is another post-processing method that flips the output label to achieve fairness, but it's bound to one fairness definition, a relaxed form of Equalized Odds as suggested in the name. Since a classifier cannot satisfy both the original Equalized Odds criterion and calibration at the same time (Kleinberg et al., 2016), Pleiss et al. (2017) suggest a relaxed form of Equalized Odds and show that it can be achieved with calibration simultaneously. In practice, this approach optimizes over calibrated classifier score outputs to find probabilities with which to change output labels with an equalized odds objective Bellamy et al. (2018).

TABLE 2.3: Bias Mitigation Techniques Overview

Fairness processor	Method	Multinomial Label	Multinomial PA	Multiple PA	Reference
Reweighing	PRE	No	No	No	Kamiran and Calders (2012)
Massaging	PRE	No	No	No	Calders et al. (2009)
Classification without discrimination	PRE	No	No	No	Kamiran and Calders (2009)
Discrimination discovery K-NN	PRE	Yes	No	No	Luong et al. (2011)
Fair representation learning	PRE	Yes	No	No	Zemel et al. (2013)
Disparate impact remover	PRE	No	Yes	Yes	Feldman et al. (2015)
Variational fair autoencoder	PRE	Yes	Yes	Yes	Louizos et al. (2017)
Discrimination-free pre-processing	PRE	No	Yes	Yes	Calmon et al. (2017)
Feature adjustment	PRE	Yes	Yes	Yes	Johndrow and Lum (2017)
Prejudice remover regularizer	IN	Yes	No	No	Kamishima et al. (2012)
Fair accuracy maximizer	IN	Yes	Yes	Yes	Zafar et al. (2017b)
Non-discriminatory Learner	IN	No	No	No	Woodworth et al. (2017)
Adversarial debiasing	IN	Yes	Yes	Yes	Zhang et al. (2018)
AVD Penalizer, SD penalizer	IN	No	No	No	Bechavod and Ligett (2018)
Fair Forests	IN	Yes	Yes	Yes	Raff et al. (2017)
Exponentiated gradient reduction	IN	No	No	Yes	Agarwal et al. (2018)
Meta-fairness algorithm	IN	No	Yes	Yes	Celis et al. (2020)
Group-wise Platt scaling	POST	Yes	Yes	Yes	Platt (2000), Barocas et al. (2019)
Group-wise histogram binning	POST	Yes	Yes	Yes	Zadrozny and Elkan (2001)
Group-wise isotonic regression	POST	Yes	Yes	Yes	Niculescu-Mizil and Caruana (2005)
Fairness-aware classifier	POST	No	No	No	Calders and Verwer (2010)
Reject option classification	POST	No	Yes	Yes	Kamiran et al. (2012)
Fairness constraint optimizer	POST	Yes	Yes	Yes	Goh et al. (2016)
Equalized odds processor	POST	No	Yes	No	Hardt et al. (2016)
Calibrated equalized odds	POST	No	No	No	Pleiss et al. (2017)

*source: *Fairness in Credit Scoring: Assessment, Implementation and Profit Implications*, Kozodoi et al. (2022).
Abbreviations: PA = Protected Attributes

2.5 Assessment Tools

Several open-source libraries trying to tackle fairness in Machine Learning have been developed in recent years. Some of them were designed to detect bias in datasets and models, and some address both bias detection and bias mitigating. Libraries that can be used to detect bias include but are not limited to the followings: First of all, "FairML" (Adebayo, 2016) is an end-to-end toolbox for auditing predictive models, it quantifies the relative significance of the model's inputs; "FairTest" (Tramèr et al., 2016) checks for associations between predictions and protected attributes; "Themis" (Galhotra et al., 2017) is a package that generates efficient test suites to measure discrimination; and "Aequitas" Saleiro et al. (2019) is toolkit that covers several metrics and provides a "fairness tree" that helps to identify suitable metrics in different cases, this paper refers to the fairness tree in the proposed FairGridSearch framework as a guidance on the choice of fairness metrics.

In addition to the libraries that only identify bias, several toolkits cover bias-mitigating algorithms as well. For example, "Themis-ML" (Bantilan, 2017) provides methods like relabeling (Kamiran and Calders, 2012) and reject option classification (Kamiran et al., 2012). Another repository "Fairness Comparison" (Friedler et al., 2018) includes disparate impact remover (Feldman et al., 2015), prejudice remover (Kamishima et al., 2012), and "Fair accuracy maximizer" (Zafar et al., 2017b). "Fair-Learn" (Bird et al., 2020) provides bias mitigating algorithms such as GridSearch and Exponentiated gradient reduction, both introduced from Agarwal et al. (2018) and wrapped in the AIF360, the library used in this paper.

"AIF360" (Bellamy et al., 2018) is an open-source python library provided by IBM researchers. Containing both bias detection and bias mitigating algorithms, it allows a deep study on the effects of these distinct models on fairness processing with respect to multiple fairness criteria. AIF360 is chosen for its comprehensive set of metrics and bias mitigators. It covers various fairness metrics, including group fairness and individual fairness, and supports 13 bias mitigators. Most importantly, the new "sklearn" sub-package matches the most popular machine learning library scikit-learn for python, allowing for easier integration.

On top of fairness metrics and bias mitigators, the package also provided an interface to several popular datasets: Adult Census Income (Kohavi, 1996), German Credit (Dua and Graff, 2017), ProPublica Recidivism (COMPAS) (Larson et al., 2016), Bank Marketing (Moro et al., 2014), and different versions of Medical Expenditure Panel Surveys (AHRQ., 2015, 2016).

Chapter 3

Methodology

When working with machine learning projects, a common practice is to train multiple models on the same data and compare them to select the one with the best performances. To build a high accuracy classification model, both choosing powerful machine learning algorithms and adjusting their hyper-parameters (Syarif et al., 2016) are of high importance: choosing appropriate hyper-parameters can improve model performance significantly. However, the best values for hyper-parameters are unknown prior to training, therefore, with the same algorithm, different sets of hyper-parameters are applied in order to find the optimal values. Typically, this process of choosing a set of optimal hyper-parameters for a learning algorithm is called (hyper-) parameter tuning or optimization.

The traditional way of performing hyper-parameter tuning is grid search, which is an exhaustive search based on a pre-specified subset of the hyper-parameter space (Syarif et al., 2016). As one of the most widely used strategies in the literature of empirical machine learning, it is simple to implement and parallelization is trivial (Bergstra and Bengio, 2012). Nonetheless, the existing grid search approach does not consider bias mitigation, thereby restricting its applicability in scenarios where selecting the most suitable bias mitigation method for a given dataset is of paramount importance. To this end, this paper presents FairGridSearch, which incorporates various bias mitigation approaches as parameters, providing an effortless comparison, and consequently facilitating straightforward model recommendations. The following paragraphs first provide and overview of the general framework of FairGridSearch, and then going more into depth about which parameter-tuning opportunities are possible, and finally the best model criterion.

3.1 FairGridSearch Framework

3.1.1 General Framework

The structure of FairGridSearch closely resembles that of conventional Grid Search. In essence, the algorithm iteratively traverses a set of pre-determined parameters and fit the training dataset on the base estimator. By executing this process, it ultimately identifies and selects the optimal set among the full parameter space specified in advance. FairGridSearch offers the possibility to adjust multiple parameters, encompassing base estimators, hyper-parameters specific to the chosen base estimator, classification threshold, and finally, the bias mitigation approaches.

Moreover, according to Friedler et al. (2018), assessing model performance with a single train-test split is inadequate due to its instability, particularly for fairness metrics. The authors, therefore, suggest employing a moderate number of randomized train-test splits to account for performance instability. In this study, to address this issue, FairGridSearch enables the execution of each model with stratified k-fold cross-validation. The overall structure of the algorithm is depicted in the pseudo code 1, and the following subsections will expound on the parameter tuning for each aspect in a more comprehensive manner.

Algorithm 1: FairGridSearch

```
Input  : dataset D, base estimator base, parameter grid param_grid, k-fold k
Output: optimal set of parameters, table of all results
for hyperp in param_grid[hyperp_grid] do
    for train, test in stratified-kfold(D, k) do
        for BM in param_grid[BM_grid] do
            model = BM(base(hyperp));
            model.fit(train);
            pred_prob = model.predict_proba(test);
            for threshold in param_grid[threshold_grid] do
                Get prediction with respect to threshold;
                Calculate accuracy and fairness metrics based on prediction
            end for
        end for
    end for
    take average of all metrics from k-fold for each model
end for
return best_param, result_table
```

3.1.2 Parameter Tuning

Base Estimator

In the field of machine learning, classifiers refer to algorithms that are designed to predict the class label of a given dataset based on a set of input features. Typically, classifiers are trained on labeled datasets, where each data point is associated with a specific class label. Through this training process, the classifier learns to associate specific feature values with the corresponding class labels. And in binary classification, the class labels are limited to only two categories, often labeled as positive and negative, or as 1 and 0, respectively. Since this work focuses on binary classification, the base estimators here refer to different kinds of binary classifiers. In practice, logistic regression and decision tree are some commonly used classifiers, and in recent years, deep neural networks have also become increasingly popular.

Within the 341 publications they surveyed on, Hort et al. (2022) discovered that the most frequently used classification base estimator in fair machine learning is Logistic Regression (LR), followed by Neural Networks (NN), Random Forest (RF), Support Vector Machine (SVM), Decision Tree (DT) and Naive Bayes (NB). Additionally, Hort et al. (2022) found out the majority of publications applied their bias mitigation approaches to only one single base estimator. However, it is important to note the selection of a base estimator can impact the classification accuracy of the model, and potentially its fairness properties as well. Therefore, multiple base estimators are included in FairGridSearch, including some of the most common ones in the literature, namely LR, RF, GB, SVM, NB, and TabTransformer (TabTrans). TabTransformer is a deep tabular data modeling architecture for supervised and semi-supervised learning proposed by Huang et al. (2020). With all the base estimators mentioned above, FairGridSearch hopes to cover most categories of binary classifiers that might be of interest.

Classification Threshold

Enhancing algorithm accuracy can be achieved through multiple approaches, aside from tuning model parameters, modifying the classification threshold can also play a significant role. The classification threshold, denoted as τ, is conventionally set to 0.5, indicating that predicted probabilities above this value are assigned to the positive class, and probabilities below the threshold as negative. That is, the number of confusion matrix vary with the choice of the threshold, and there is a trade-off between the amount of false positives and the amount of false negatives (Baldi et al., 2000). Tuning classification thresholds can hence also be utilized as a means of prioritizing between errors, given that the costs of prediction errors may differ as highlighted by (Kozodoi

et al., 2022). As an input of classification threshold, the default value is set to 0.5 in FairGridSearch. However, it is possible to include any set of classification thresholds as parameters to optimize the model.

Bias Mitigation

For bias mitigation approaches, two methods each for pre-, in-, and post-processing categories are included. Since the LFR approach can also be used as an in-processing method by utilizing the learned target coefficients (Bellamy et al., 2018), FairGridSearch includes both the pre- and in- options. In addition to individual bias mitigation methods, FairGridSearch also integrates two combined approaches. Existing research has primarily focused on mitigating algorithmic bias by intervening at a specific stage of the machine learning pipeline with using only one single type of bias mitigation method. However, Ghai et al. (2022) has highlighted that algorithmic bias may still persist through other stages or components of the pipeline. Therefore, the proposed framework incorporates two approaches that address bias in two stages of the machine learning pipeline through the combination of pre-processing and post-processing methods. Specifically, the Reweighing method is combined with each of the two post-processing approaches, namely the ROC and CEO.

Due to various different reasons, the TabTransformer base estimator is not compatible with certain bias mitigation approaches. First, The LFR_pre approach necessitates the conversion of categorical variables to numerical before fitting, whereas TabTransformer requires at least one categorical variable in the dataset. Second, AD requires disabling eager execution from the machine learning library TensorFlow, while this prevents TabTransformer from functioning. And lastly EGR provided in AIF360.sklearn is limited to sklearn models, and thus TabTransformer cannot be included in conjunction with EGR.

In conclusion, the set of bias mitigation approaches considered in this study comprises RW, LFR in both pre-processing (LFR_pre) and in-processing (LFR_in) forms, AD, EGR, and two post-processing methods, ROC and CEO. Additionally two combined approaches, namely RW+ROC and RW+CEO are included. Nonetheless, certain bias mitigation approaches cannot be applied when TabTransformer is used as the base estimator due to either theoretical or technical conflicts as outlined above.

3.1.3 Best Model Criterion

In accordance with traditional GridSearch techniques, FairGridSearch employs a scoring metric to select the most optimal model from among a range of models that have

been executed with specified parameters. While conventional grid search typically uses accuracy metrics such as Accuracy (ACC), F1 score, and AUC, FairGridSearch considers both accuracy and fairness metrics. To this end, this paper adopts the method proposed by Haas (2019) as a criterion to identify the most optimal model.

The study conducted by Haas (2019) employed a cost-based analysis to identify the optimal model in relation to accuracy and fairness metrics. The authors formulated the overall cost of an approach as a linear combination of the cost associated with both objectives, which are accuracy and fairness in this study. In this approach, the levels of α and β were utilized to represent the weights for scenario-specific costs.

$$C = C_{metric_1} + C_{metric_2} = \alpha \cdot f(metric_1) + \beta \cdot f(metric_2) \tag{3.1}$$

Under the assumption that costs can be measured by the distance to the metric's optimal value, the authors then reformulated equation 3.1 from above, defined as follow for this study:

$$C = C_{acc} + C_{fair} = \alpha \cdot (1 - metric_{acc}) + \beta \cdot abs(metric_{fair}) \tag{3.2}$$

and the model that minimizes the overall cost, as defined by the cost-based analysis, will be selected as the optimal model among all the candidates.

Accuracy Metrics

Accuracy metrics are evaluation metrics used in binary classifications to measure the performance of a model in predicting the correct outcomes for a given set of data. These metrics generally take advantage of the confusion matrix, and provide information on the correctness of the predictions with respect to the ground truth values. The choice of metric is often dependent on the classification problem domain and the previous practices in the related literature (Canbek et al., 2021). FairGridSearch includes several common accuracy metrics such as Accuracy (ACC), Balanced Accuracy (BACC), F1 Score and AUC. Besides, Matthews correlation coefficient (MCC) is also included into the set of accuracy metrics as suggested from several recent research papers.

In recent years, the limitations of the commonly used metrics such as ACC and F1 Score have been noted. For instance, it has been shown that class imbalance in datasets can exert a major impact on the value of ACC, which manifests over-optimistic prediction performance towards the majority class (Luque et al., 2019, Hossin and Sulaiman, 2015). Likewise, F1 Score has been criticized as invariant to class swapping and independent from the number of true negative samples (Brown, 2018, Powers, 2019).

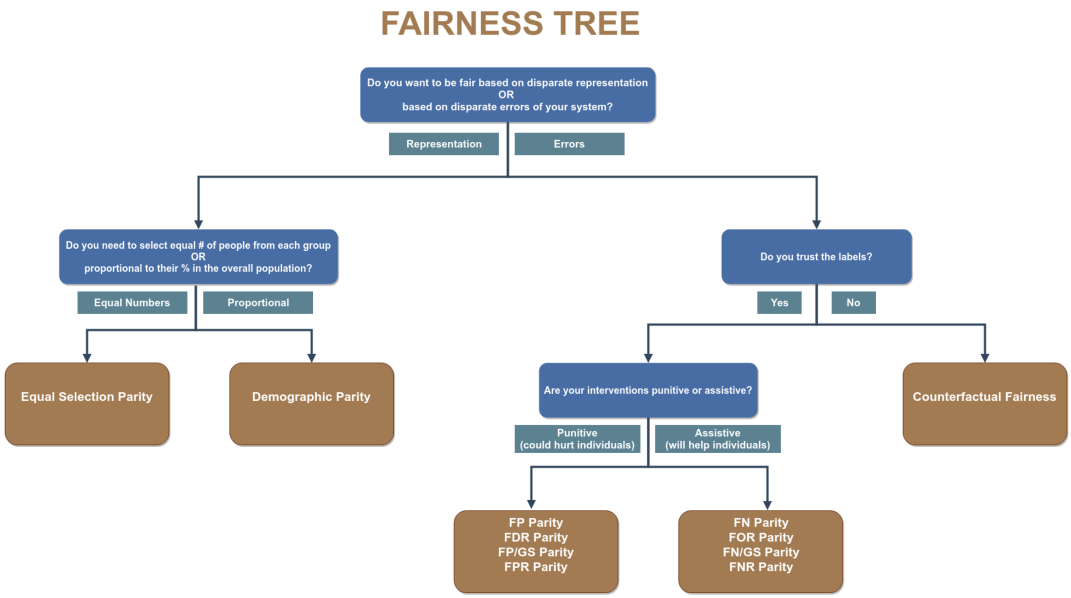
MCC arises as a better choice in recent research, Canbek et al. (2021) introduce a two-staged benchmarking method called BenchMetrics to compare the robustness of binary classification accuracy metrics. In their work, a total of thirteen different accuracy metrics were tested, and the results demonstrated that MCC is the most robust and recommended metric for binary classification performance evaluation. A series of comparison between MCC and other metrics also showed that MCC produces a more informative and truthful score than ACC, F1 Score, Diagnostic Odds Ratio (DOR), Cohen's Kappa and Brier Score (Chicco and Jurman, 2020, Chicco et al., 2021a,b), and therefore suggest machine learning practitioners to evaluate through MCC instead of the other metrics. Furthermore, Gösgens et al. (2022) conduct a systematic analysis on classification performance measures, wherein they formally defined a list of desirable properties, analyzed which measures satisfy these properties, and ultimately concluded that MCC should be preferred over ACC, F1 Score and Cohen's Kappa. In light of the literature reviewed above, this paper chooses MCC as the primary accuracy metric for the experiment presented in Section 4, but FairGridSearch also reports model accuracy measured by other commonly used metrics.

Given that MCC is bounded between -1 and 1, the normalized variant Normalized Matthews correlation coefficient (NORM_MCC) is also employed to make it more comparable to other accuracy metrics. Specifically, NORM_MCC is defined as $norm\_MCC = 0.5 * MCC + 1 \in [0, 1]$. In summary, FairGridSearch comprises six different accuracy metrics, namely ACC, BACC, F1 score, AUC, MCC, and NORM_MCC, with NORM_MCC being the primary metric employed in the best model criterion for the exemplary experiments.

Fairness Metrics

Fairness metrics play an integral part in the bias mitigation process, because adoption of different fairness metrics can produce vastly different outcomes: for instance, a model that exhibits negligible degree of Statistical Parity Difference (SPD) may at the same time be deemed highly unfair in terms of equalized odds. However, as discussed in Section 2.3, there exist various fairness metrics in the realm of fair ML, some of which proven to be incompatible with each other as reviewed in Kleinberg et al. (2016). This makes the selection of the appropriate fairness metric a challenging task, and it is necessary to consider the context and the intended use case of the fairness definitions to choose the most appropriate metrics accordingly (Mehrabi et al., 2022).

FIGURE 3.1: Fairness Tree helps to select the fairness metric(s) that are relevant to each context (Saleiro et al., 2019).



FairGridSearch framework incorporates several group and individual fairness criteria into the algorithm. However, for comparing different models, this paper exclusively relies on group fairness in the best model criterion. The selection of group fairness criteria was guided by the "fairness tree" (figure 3.1) proposed by Saleiro et al. (2019), which facilitates the determination of a suitable fairness metric based on the dataset and problem type. According to fairness tree, the choice of the metrics depends on the goals and priorities of the project: if there will be an intervention drew on model predictions, then the errors of the system become important, leaving out metrics not considering prediction such as statistical parity (or demographic parity in the graph). Subsequently, interventions are classified into punitive or assistive types, depending on the its nature. When the intervention is punitive, metrics involving false positives such as false positive rate parity and false discovery rate parity become crucial because false positives in such situations imply wrongly punishing innocent individuals. Conversely, the importance of false negatives becomes salient when the interventions are assistive, as marginalizing specific groups from support and assistance is considered more detrimental than providing additional aid to those who do not necessarily need it.

Chapter 4

Exemplary Experiment

This section demonstrate the applicability of FairGridSearch with exemplary experiments. The experiments were conducted using all six base estimators and nine bias mitigation methods provided in the framework, and five classification thresholds ranging from 0.3 to 0.7. In addition, four different combinations of base-specific parameters were considered for each base estimator as shown in Table 4.1:

TABLE 4.1: Base Estimator Parameters

Base Estimator	Parameters
LR	'C':[1, 10], 'solver':['liblinear', 'saga']
RF	'n_estimators':[10, 50], 'criterion':['gini', 'entropy']
GB	'n_estimators':[10, 50], 'max_depth':[8, 32]
SVM	'kernel':['rbf','linear','poly','sigmoid']
NB	'var_smoothing': np.logspace(0,-9, num=4)
TabTrans	'epochs':[20, 30], 'learing_rate':[1e-04, 1e-05]

Table 4.2 shows the number of models for each dataset. Notably, TabTransformer models were run with two fewer bias mitigation methods since they are incompatible with LFR_pre and EGR. The third row shows the two bias mitigation methods that are base estimator invariant, LFR_in and AD. These two in-processing methods change the entire model algorithm and therefore do not take base estimators into account. In total, 930 models were implemented for each dataset. It is worth noting that each model was run with 10-fold cross-validation.

In terms of the best model criterion, the accuracy metric chosen was NORM_MCC across all datasets, while the selection of fairness metrics were guided by the fairness tree (Saleiro et al., 2019). Specifically, the metrics SPD, PPVD, and EOD were used for the Adult, COMPAS, and German Credit datasets, respectively. It's important to note that for all fairness metrics, the absolute values were used to indicate the magnitude of bias present, regardless of the direction. A value of 0 indicated the greatest fairness, while larger values indicated greater levels of bias. Finally, the weight of both the accuracy and fairness metrics, α and β, were set to 1, meaning equal consideration for

both criteria. In the next subsection, the datasets chosen for the experiments will be introduced.

TABLE 4.2: Numbers of Models in the Exemplary Experiment

	Base	hyper-p	threshold	BM	Total
All but TabTrans	5	4	5	8	800
TabTranas	1	4	5	6	120
Base-invariant BMs	-	-	5	2	10
					930

4.1 Dataset

Fabris et al. (2022) found over two hundred datasets employed in studies of algorithmic fairness and identified the three most popular ones, Adult, COMPAS, and German Credit. On top of being the most used datasets in the relevant research, these three datasets fall into three different categories according to the fairness tree, making them suitable choices for the exemplary experiments. The following sections introduce each dataset in more details.

4.1.1 Adult

The Adult dataset was created as a resource to benchmark the performance of machine learning algorithms on socially relevant data (Fabris et al., 2022), extracted by Barry Becker from the 1994 Census database the dataset is used to predict whether or not an individual's income exceeds 50K USD a year. The original data set has 48,842 rows and 15 attributes, including demographic information such as sex, age and race, as well as attributes like education level, work class, working hours per week. The attribute "fnlwgt" is left out due to irrelevance and ambiguous definition (Fabris et al., 2022).

In the original data set, the sensitive attributes race has five groups, including "White" (85.5%), "Black" (9.6%), "Asian-Pac-Islander" (3.1%), "Amer-Indian-Eskimo" (0.97%) and "Other" (8.31%). Since most bias mitigation methods require the sensitive attributes to be binary, only "White" and "Black" in the experiment are kept, ending up with 46,447 rows and 14 attributes. For Adult dataset, group fairness metric SPD is chosen since originally there was no prediction-based intervention intended.

4.1.2 COMPAS

ProPublica compiled a dataset of criminal defendants in Broward County, Florida, from 2013 to 2014 through public records request (Angwin et al., 2016). The original dataset consists of 7,214 rows and 53 columns, containing information such as defendants' demographics (sex, race, and age), criminal histories (current and prior charge type and charge degree), the COMPAS risk scores, and the target variable shows if a recidivism within the following two years actually occurred. After selection of relevant attributes, 14 out of 53 attributes were selected, including sex, age, $race$, juv_fel_count, juv_misd_count, juv_other_count, $priors\_count$, c_charge_degree, two_year_recid.

For the COMPAS data, the protected attribute race has six groups in the original dataset, including "African-American" (51.2%), "Caucasian" (34%), "Hispanic" (8.8%), "Asian" (0.44%), "Native American" (0.25%) and "Other" (5.2%). Again, only data with "Caucasian" and "African-American" are kept, resulting in a final dataset containing 6,150 rows and 9 attributes. In the COMPAS setting, the predictions are being used to make pretrial release decisions, so FDR parity is important. In this study, Positive Predictive Value Difference (PPVD) is employed to encapsulate this notion. It is noteworthy that FDR is defined as the complement of PPV, i.e., $FDR = 1 - PPV$, hence, the absolute difference between FDR across groups is equal to that of PPV.

4.1.3 German Credit

The German Credit dataset was initially created to study computer-assisted credit decisions in a regional bank in southern Germany (Fabris et al., 2022). Specifically, the dataset comprises loan applicants who were considered creditworthy and hence approved for a loan between 1973 and 1975. The original dataset contains 1,000 rows and 21 attributes. Similar to the other two datasets, these attributes carry demographic information such as age and sex, other variables include credit history, credit amount, credit purpose etc.

Nonetheless, unlike the other two datasets, the protected attribute for German Credit is sex, the binary protected attribute is composed of "Male" (69%) and "Female" (31%). For this dataset, no specific data pre-processing was applied except from extracting protected attribute from the personal status column. As the act of issuing credit is regarded as an "assistive" intervention, achieving false negative rate parity is crucial. As per its definition, False Negative Rates (FNR) can be expressed as the complement of True Positive Rates (TPR), that is, $FNR = 1 - TPR$. Consequently, the absolute difference between FNR across various groups is equivalent to that of TPR, which has been identified by Hardt et al. (2016) as the Equal Opportunity metric.

Hence, for the German Credit dataset, Equal Opportunity Difference (EOD) is utilized as the appropriate fairness criterion. An overview of all three datasets are shown in Table 4.3.

TABLE 4.3: Datasets used in the experiments

	Size	#Features	PA	PA (priv.)	PA (unpriv.)	Favorable Label	Fairness Metric
Adult	46,447	14	Race	Caucasian	African American	High Income	SPD
COMPAS	6,150	9	Race	White	Black	No Recidivism	PPVD
German Credit	1,000	21	Sex	Male	Female	Good Credit	EOD

4.2 Results

This section reports the experimental findings and their analyses. The presentation starts with a separate discussion of each dataset (Section 4.2.1, 4.2.2, 4.2.3), followed by cross-dataset analyses. For each dataset, tables showcasing the best models identified by FairGridSearch are presented. Subsequently, two-dimensional analyses are performed to consider both accuracy and fairness from various perspectives. These include the overall results, analysis based on classification threshold, base estimators, bias mitigation methods, and bias mitigation categories.

Next, Section 4.2.4 employs the methodology proposed by Chen et al. (2023) to investigate the impact of bias mitigation methods on accuracy and fairness across datasets. Firstly, a non-parametric Mann Whitney U-test is used to determine the statistical significance of differences between models before and after bias mitigation methods at a significance level of 0.05. Secondly, Cohen's d effect size is calculated to check whether the difference has a meaningful effect. Similar to the analyses conducted for individual datasets, analyses across datasets are conducted on a general level and from different perspectives, such as base estimator and bias mitigation method.

Table 4.4 provides an overview of the run time for each dataset based on base estimators. The experiments were conducted on the CPU instance of SageMaker Studio Lab[1] with 16 GB of RAM, except for SVM models on the Adult dataset, which were executed on M1 Pro chip with CPU speed of 3228 MHz and RAM of 16 GB. According to SageMaker Studio Lab, the availability of compute instances is subject to demand, so there was no clear information on the cpu speed. From the table, it is evident that the algorithm takes significantly longer to complete all models for every estimator with the larger Adult dataset, especially for the SVM models, which take a whole week. This finding emphasizes the importance of considering computational feasibility when

[1]SageMaker Studio Lab `https://studiolab.sagemaker.aws`

choosing a suitable base estimator for a particular dataset, as certain estimators may be computationally expensive and impractical to use on large datasets. Furthermore, the findings underscore the necessity for an enhanced approach to parameter optimization, such as random search. This method may be beneficial for scenarios that face computational constraints, yet require testing of specific base estimators on larger datasets.

TABLE 4.4: Run-time for each Dataset used in the experiments

	LR	RF	GB	SVM	NB	TabTrans	Total
Adult	4.4 hr	2.2 hr	10 hr	168* hr	1.4 hr	34.5 hr	221 hr
COMPAS	5 min	5 min	12.5 min	3 hr	2.5 min	1.2 hr	4.7 hr
German Credit	5.5 min	4.5 min	50 min	5.5 min	2 min	58 min	2.1 hr

*SVM models for the Adult dataset were run on M1 Pro, while the remaining models on SageMaker Studio Lab, since SVM models exceeded the maximum runtime limit on SageMaker Studio Lab.

4.2.1 Adult

Table 4.5 presents the top ten models for the Adult dataset, which have the lowest cost in terms of both accuracy and fairness. The table reveals that two tree-based base estimators, GB and RF, are the most common among the top models. Additionally, the use of ROC as the bias mitigation method appears to be the best choice for Adult, as all of the top models utilize ROC either with or without RW. Furthermore, the majority of the top models have a lower classification threshold than the standard value of 0.5, suggesting that a lower threshold may be more appropriate for this dataset.

TABLE 4.5: Top 10 Models for Adult

Base Estimator	Param	Bias Mitigation	Threshold	Norm. MCC	Abs. SPD	Cost
GB	'criterion': 'friedman_mse', 'max_depth': 8, 'n_estimators': 50	RW+ROC	0.4	0.8115	0.0091	0.1976
GB	'criterion': 'friedman_mse', 'max_depth': 8, 'n_estimators': 50	RW+ROC	0.5	0.8028	0.0060	0.2032
RF	'criterion': 'entropy', 'max_depth': 16, 'n_estimators': 50	ROC	0.3	0.7995	0.0040	0.2046
GB	'criterion': 'friedman_mse', 'max_depth': 8, 'n_estimators': 50	RW+ROC	0.3	0.8144	0.0198	0.2054
RF	'criterion': 'gini', 'max_depth': 16, 'n_estimators': 50	ROC	0.3	0.8007	0.0065	0.2058
RF	'criterion': 'entropy', 'max_depth': 16, 'n_estimators': 10	ROC	0.3	0.7955	0.0078	0.2124
GB	'criterion': 'friedman_mse', 'max_depth': 8, 'n_estimators': 50	RW+ROC	0.6	0.7856	0.0003	0.2147
RF	'criterion': 'gini', 'max_depth': 16, 'n_estimators': 10	RW+ROC	0.4	0.7899	0.0075	0.2176
RF	'criterion': 'gini', 'max_depth': 16, 'n_estimators': 10	ROC	0.3	0.7953	0.0131	0.2178
RF	'criterion': 'gini', 'max_depth': 16, 'n_estimators': 50	RW+ROC	0.4	0.7930	0.0108	0.2178

Next on, Figure 4.1 and 4.2 provide a better overview on the performance of all models with respect to both accuracy and fairness. Normalized MCC, representing accuracy, is displayed on the x-axis, while absolute average value SPD, representing fairness, is presented on the y-axis. Models with higher accuracy and fairness are located in the bottom right quadrant of the graph.

The initial observation from Figure 4.1a) is that there is no apparent trade-off between accuracy and fairness. However, it is worth noting that models with an accuracy score ranging between 0.7 and 0.8 exhibit considerable volatility in fairness. Building on this finding, four additional figures investigate the relationship between accuracy and fairness more in-depth from different perspectives. These perspectives include classification threshold (4.1b), base estimators (4.1c), bias mitigation methods (4.2a), and bias mitigation categories (4.2b).

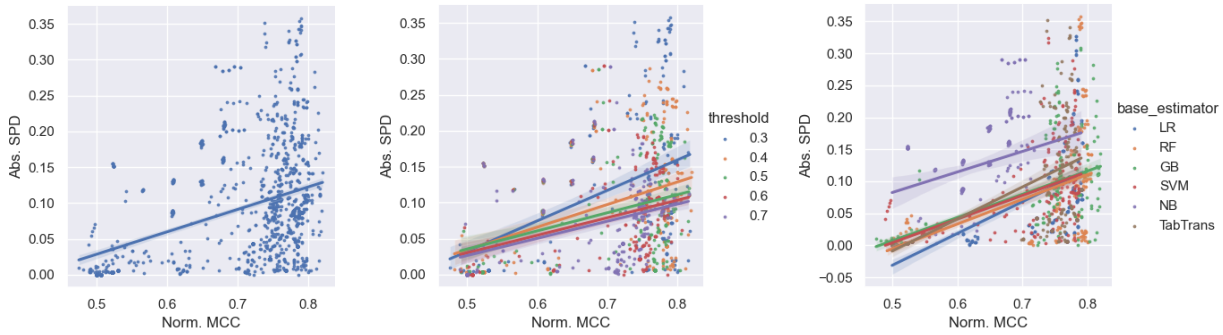
Figure 4.1b) shows that the high fairness volatility from high accuracy models is a result from distinct classification thresholds. In the case of models exhibiting NORM_MCC within the interval of 0.7 to 0.8, even though all threshold values can yield fair models with SPD close to zero, it is observed that a decrease in the classification threshold leads to an increase in fairness volatility, while an increase in the threshold value results in a greater concentration of values in the lower range of SPD. However, looking at the accuracy, fairer models with 0.7 as a threshold are less accurate. As a consequence, the best performing models are mainly composed of those with lower classification thresholds.

Based on the information presented in Figure 4.1c), it is evident that the accuracy of the NB models is generally lower than that of the other base estimators, while the other base estimators exhibit similar behavior in terms of both accuracy and fairness. However, the GB and RF base estimators appear to be slightly more accurate while maintaining fairness, which makes them the most prominent base estimators in the top-performing models.

The subsequent analysis in figure 4.2a) is conducted based on different bias mitigation methods. The graph reveals that the two LFR models exhibit lower accuracy by a large margin compared to the other models, albeit being fairer. Furthermore, among the models with higher accuracy, CEO and RW+CEO exhibit the lowest levels of fairness, while ROC and RW+ROC appear to be the most effective bias mitigation methods with high accuracy.

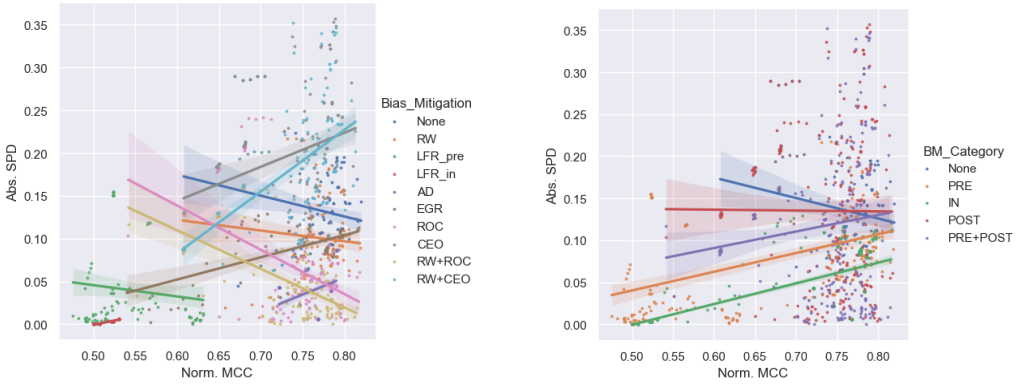
Finally, in Figure 4.2b), it can be observed that among the high accuracy models, those utilizing in-processing methods exhibit greater stability in fairness, whereas models incorporating post-processing methods exhibit high volatility in fairness, ranging from minimal bias to the least fair outcomes. However, as indicated by the previous

figure (Figure 4.2a), bias mitigation methods involving ROC and CEO, the two post-processing methods, show significant differences in terms of fairness performance. Thus, consolidating them into a single category and drawing a unified conclusion may be misleading.



A) Overall Results. B) Based on Thresholds. C) Based on Base Estimators.

FIGURE 4.1: Adult: Accuracy versus Fairness



A) Based on BMs. B) Based on BM Categories.

FIGURE 4.2: Adult: Accuracy versus Fairness (continued)

4.2.2 COMPAS

Table 4.6 displays the top models for the COMPAS dataset. The table indicates that the ten best models are composed of three distinct base estimators, with LR being the most commonly used. It is noteworthy that the second and third best models in this dataset are generated without any bias mitigation methods. Both models are LR with a classification threshold of 0.4, and they produce predictions that are considerably fair even without bias mitigation. And among the top-performing models that incorporate bias mitigation methods, all of them employ either of the two post-processing

methods. Furthermore, the table shows that only two values, 0.5 and 0.4, are used as classification thresholds in the top-performing models.

TABLE 4.6: Top 10 Models for COMPAS

Base Estimator	Param	Bias Mitigation	Threshold	Norm. MCC	Abs. PPVD	Cost
GB	'criterion': 'friedman_mse', 'max_depth': 8, 'n_estimators': 10	RW+ROC	0.5	0.6506	0.0019	0.3512
LR	'C': 1, 'penalty': 'l2', 'solver': 'saga'	None	0.4	0.6539	0.0054	0.3515
LR	'C': 10, 'penalty': 'l2', 'solver': 'liblinear'	None	0.4	0.6541	0.0056	0.3515
LR	'C': 10, 'penalty': 'l2', 'solver': 'saga'	ROC	0.4	0.6543	0.0058	0.3516
SVM	'gamma': 'scale', 'kernel': 'rbf'	RW+ROC	0.5	0.6711	0.0241	0.3530
LR	'C': 10, 'penalty': 'l2', 'solver': 'saga'	CEO	0.4	0.6497	0.0037	0.3540
LR	'C': 10, 'penalty': 'l2', 'solver': 'liblinear'	CEO	0.4	0.6495	0.0039	0.3544
SVM	'gamma': 'scale', 'kernel': 'poly'	CEO	0.5	0.6470	0.0015	0.3545
LR	'C': 1, 'penalty': 'l2', 'solver': 'liblinear'	CEO	0.4	0.6493	0.0041	0.3548
SVM	'gamma': 'scale', 'kernel': 'rbf'	ROC	0.4	0.6612	0.0164	0.3553

Based on the Fairness Tree, PPVD has been identified as the more appropriate fairness metric for the COMPAS dataset. Consequently, in the analyses of this dataset, PPVD is employed as the fairness metric. Figure 4.3a) provides an overview of the performance of all models and again indicates that there is no discernible trade-off between accuracy and fairness. Unlike the Adult dataset, where models with high accuracy demonstrate volatility in fairness, variability in fairness exist among low accuracy models for the COMPAS dataset.

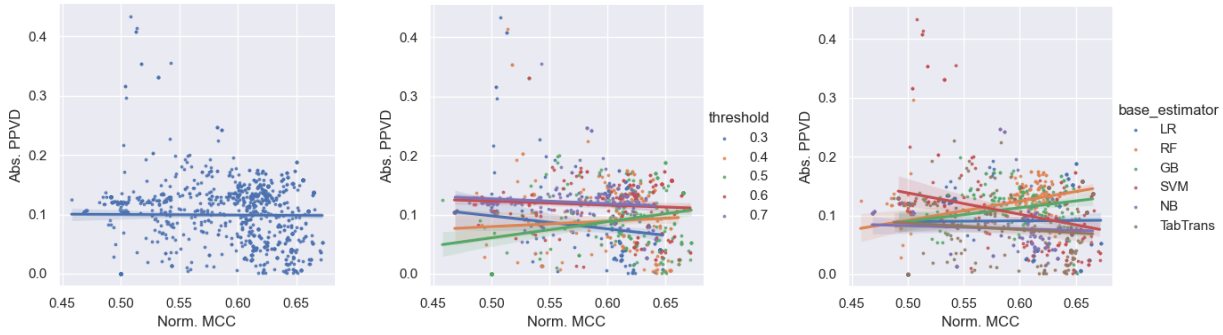
The analysis of different values of classification threshold from Figure 4.3b) does not reveal a clear difference in fairness as observed in the Adult dataset. Nonetheless, the classification thresholds of 0.5 and 0.6 generate more accurate predictions than the other threshold values.

Based on the information presented in Figure 4.3c), it can be concluded that SVM models are the primary contributors to the high variability in fairness among low accuracy models. However, it is noteworthy that some SVM models are also among the highest accuracy models, alongside LR and TabTrans. On the other hand, the well-performed base estimators RF and GB in the COMPAS dataset generally fail to achieve fairness while maintaining high accuracy. While GB has a few exceptions, RF models have lower accuracy and fairness compared to other base estimators.

The analysis presented in Figure 4.4a) pertains to the examination of various bias mitigation methods employed in the COMPAS dataset. The graph suggests again that the two LFR models exhibit remarkably lower accuracy than the other models, particularly LFR_in. Moreover, the remaining in-processing methods, namely AD and EGR, fail to mitigate bias to the same extent as other methods as well. Models employing

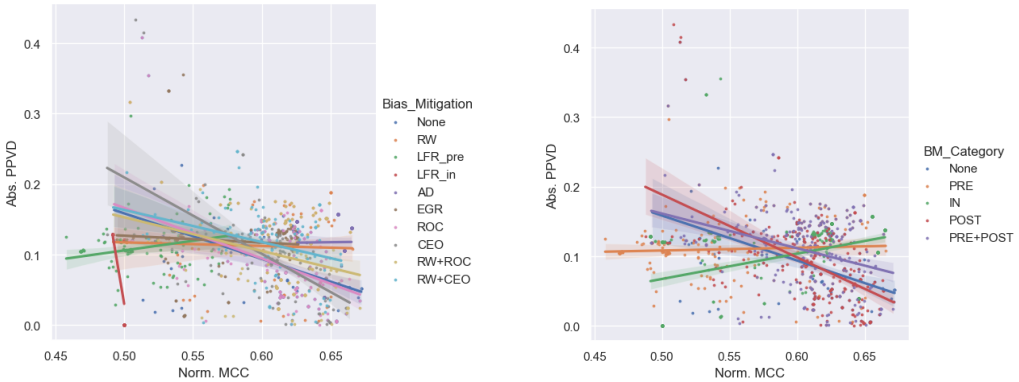
these techniques tend to produce less fair predictions, even when compared to models without any bias mitigation.

Expanding on the previous finding, Figure 4.4b) demonstrates that among high accuracy models, those utilizing in-processing methods are not as successful in completely mitigating bias compared to the other methods. In contrast, models with post-processing techniques, or a combination of post-processing and pre-processing techniques, dominate the bottom right corner of the graph where models have both higher accuracy and fairness.



A) Overall Results. B) Based on Thresholds. C) Based on Base Estimators.

FIGURE 4.3: COMPAS: Accuracy versus Fairness



A) Based on BMs. B) Based on BM Categories.

FIGURE 4.4: COMPAS: Accuracy versus Fairness (continued)

4.2.3 German Credit

Top models for German Credit are shown in Table 4.7. As shown in the first column, the ten best models comprise four different base estimators, with LR being the most frequently used. Notably, RW emerges as the preferred bias mitigation method for this dataset, as nine out of the top ten models utilized this approach, with EGR being

the only exception. Moreover, in contrast to the Adult dataset, the majority of the top models employed a classification threshold higher than the conventional value of 0.5, suggesting that a higher threshold might be more appropriate in this context.

TABLE 4.7: Top 10 Models for German_Credit

Base Estimator	Param	Bias Mitigation	Threshold	Norm. MCC	Abs. EOD	Cost
RF	'criterion': 'gini', 'max_depth': 16, 'n_estimators': 50	EGR	0.7	0.7062	0.0003	0.2941
LR	'C': 1, 'penalty': 'l2', 'solver': 'liblinear'	RW	0.6	0.7096	0.0065	0.2968
LR	'C': 10, 'penalty': 'l2', 'solver': 'liblinear'	RW	0.6	0.7091	0.0085	0.2993
LR	'C': 1, 'penalty': 'l2', 'solver': 'liblinear'	RW	0.7	0.6974	0.0008	0.3034
LR	'C': 1, 'penalty': 'l2', 'solver': 'saga'	RW	0.5	0.6968	0.0008	0.3040
LR	'C': 1, 'penalty': 'l2', 'solver': 'saga'	RW	0.7	0.6967	0.0011	0.3044
SVM	'gamma': 'scale', 'kernel': 'linear'	RW	0.6	0.7091	0.0146	0.3056
SVM	'gamma': 'scale', 'kernel': 'linear'	RW	0.7	0.6969	0.0045	0.3076
SVM	'gamma': 'scale', 'kernel': 'rbf'	RW	0.5	0.7067	0.0158	0.3091
NB	'var_smoothing': 1.0	RW	0.7	0.7000	0.0103	0.3103

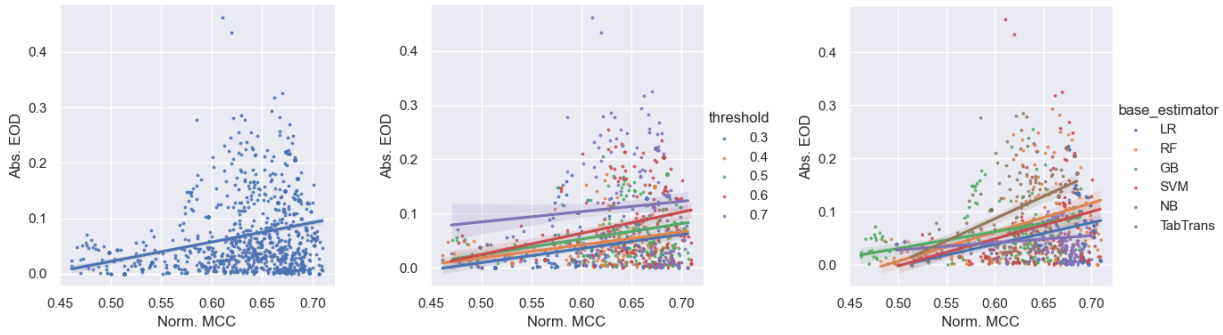
For the analyses of the German Credit dataset, EOD is chosen as the fairness metric according to the Fairness Tree. Starting with examining the overall performance of all models, 4.5a) shows no clear linear relationship between accuracy and fairness as in the other two datasets. However, similar to the Adult dataset, volatility in fairness again appear among models with high accuracy.

The analysis conducted on the classification threshold, as demonstrated in Figure 4.5b), reveals a similar phenomenon to the Adult dataset, whereby threshold values are the primary contributors to the instability in fairness. However, unlike the Adult dataset, where fairness volatility is observed for lower values of the threshold, here, the fairness volatility appears at higher threshold values. It is worth noting that threshold values, which yield more fairness stability in both datasets, tend to come with lower accuracy, while the values causing high volatility of fairness could have models that achieve both high accuracy and fairness. Hence, such threshold values appear in the top-performing models for both datasets.

As indicated in Figure 4.5c), SVM models with a threshold value of 0.7 generate the least fair predictions, with both EOD higher than 0.4. On the other hand, NB models with high accuracy produce constantly fair results regardless of the threshold values. It is also worth noting that, except for a few exceptions, NB models are relatively concentrated in the bottom right quadrant of the graph, indicating higher accuracy and fairness. Another observation worth mentioning is that the TabTrans models are absent from the bottom right corner where the best-performing models are located, indicating poor performance from this base estimator.

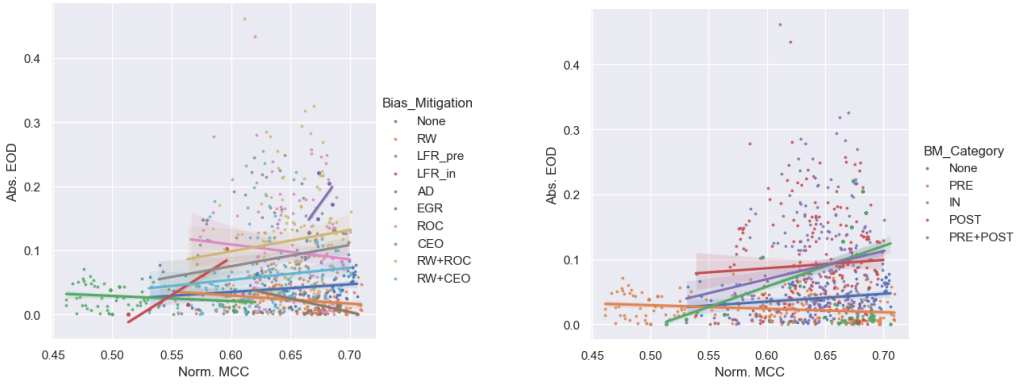
The subsequent analysis in figure 4.6a) is conducted based on different bias mitigation methods. Similar to all the other datasets, both LFR models exhibit relatively

higher fairness but markedly lower accuracy compared to the others. Furthermore, among the models with higher accuracy, ROC and RW+ROC exhibit the highest instability regarding fairness, followed by CEO, RW+CEO and AD. Combining Figure 4.5b) and 4.6a), it can be observed that models utilizing post-processing methods have a tendency to display unpredictable fairness outcomes when used in conjunction with specific threshold values. In contrast, models with pre-processing and in-processing methods tend to show more consistent fairness performance across different threshold values. This phenomenon is further supported by the evidence presented in Figure 4.6b), particularly for pre-processing methods where models consistently exhibit EODs of less than 0.1.



A) Overall Results. B) Based on Thresholds. C) Based on Base Estimators.

FIGURE 4.5: German Credit: Accuracy versus Fairness



A) Based on BMs. B) Based on BM Categories.

FIGURE 4.6: German Credit: Accuracy versus Fairness (continued)

4.2.4 Cross Dataset

This section To investigate the impact of bias mitigation methods on accuracy and fairness across datasets, the first two parts of the section adopts the methodology proposed

by Chen et al. (2023). The analysis proceeds in two stages. First, a non-parametric Mann Whitney U-test is utilized to determine the statistical significance of differences between models before and after applying bias mitigation methods, with a significance level of 0.05. Second, Cohen's d effect size is calculated to assess whether the difference has a substantive effect. Chen et al. (2023) suggest that a difference with $d \in [0, 0.5)$ should be regarded as a small effect, $d \in [0.5, 0.8)$ as a medium effect, and $d \in [0.8, \infty)$ as a large effect. General analyses are conducted across datasets from various perspectives, such as base estimator and bias mitigation method, using a similar approach as the analyses conducted for individual datasets.

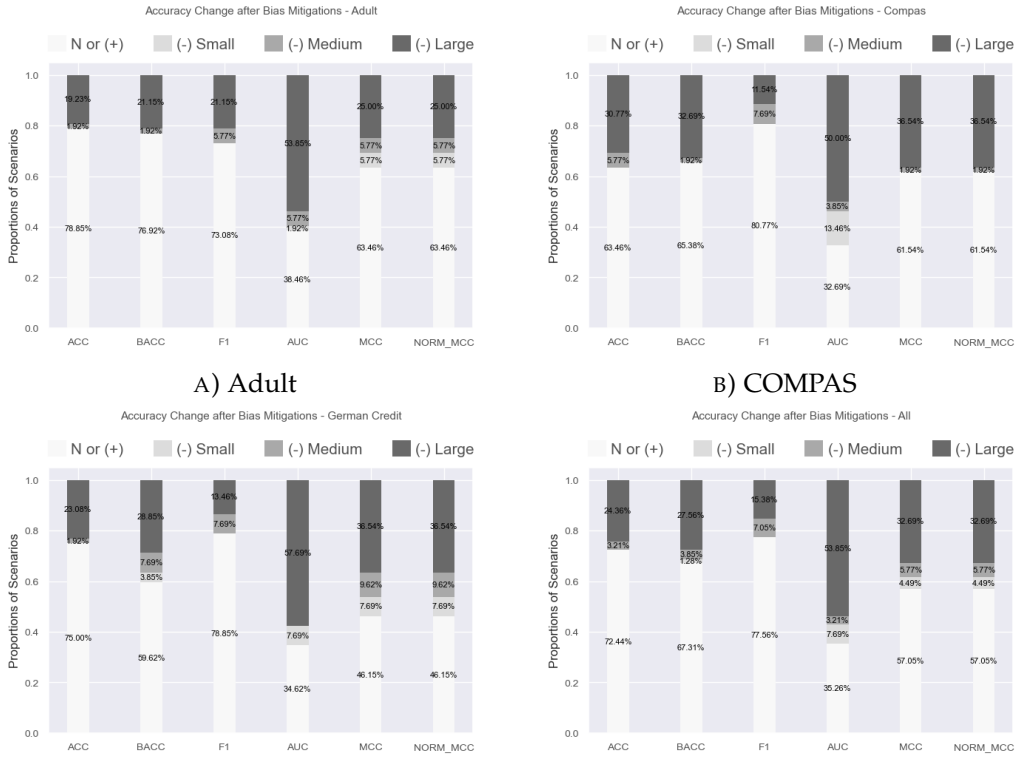
Accuracy Change

This section aims to investigate the impact of bias mitigation methods on model accuracy. Initially, the study employs six different accuracy metrics to investigate whether bias mitigation methods significantly alter the accuracy values of ML models. Subsequently, the analysis concentrates on NORM_MCC and conducts the investigation based on different base estimators. Finally, the research compares various bias mitigation methods based on their impact on accuracy, again with a particular focus on normalized MCC.

Accuracy Change Among Metrics In order to examine the impact of bias mitigation methods on model accuracy, models utilizing bias mitigation methods are compared with their corresponding baselines. For example, a LR model with RW is compared to the same LR model with no bias mitigation. The non-parametric Mann Whitney U-test and Cohen's d are employed to assess both the significance of the differences between these models and the size of the effect. The results are then categorized as "significantly decreased (large)", "significantly decreased (medium)", "significantly decreased (small)", or "not significant or increased". The following graphs illustrate the proportions of each scenario for each dataset, as well as for all datasets combined.

Figure 4.7d) reveals that the values of the six accuracy metrics decrease significantly in an average of 40% of all scenarios (ranging from 22.44% to 64.74% depending on the metric). This suggests that different accuracy metrics exhibit varying degrees of sensitivity to the application of bias mitigation methods. For instance, AUC experiences the greatest reduction in accuracy among all the metrics used in the experiments. This trend remains consistent across all datasets analyzed. However, other than AUC, most of the accuracy metrics were not significantly decreased after bias mitigation was applied in the majority of cases. The only exception is (normalized) MCC in the German

Credit dataset, where significant decreases of the metrics are observed in 53.85% of the scenarios.



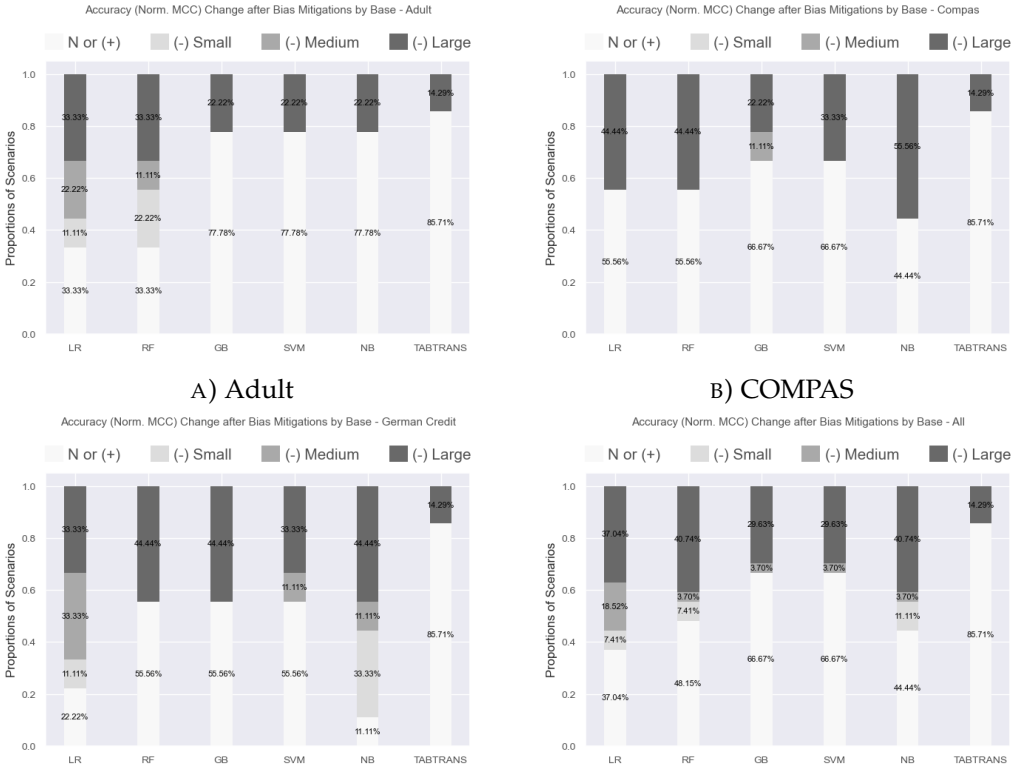
A) Adult B) COMPAS

C) German Credit D) All

FIGURE 4.7: Accuracy Change after Bias Mitigations

Accuracy Change Among Base Estimators In this section, the analyses focus on one single accuracy metric, namely the NORM_MCC. The aim is to investigate whether the accuracy changes differently among different base estimators after applying bias mitigation methods.

The results from Figure 4.8 indicate that TabTransformer models experience the least decrease in accuracy after the application of bias mitigation. Conversely, LR and RF base estimators are more impacted than the others, especially in the Adult dataset, and LR for the German Credit dataset. Notably, the effect of NB base estimator differs greatly across the three datasets in the experiment. For example, around 90% of the scenarios in the German Credit dataset where models are trained with NB as the base estimator exhibit a significant decrease in accuracy, while only 55.56% of scenarios in the COMPAS dataset show a significant decrease in accuracy, and only 22.22% of scenarios in the Adult dataset.

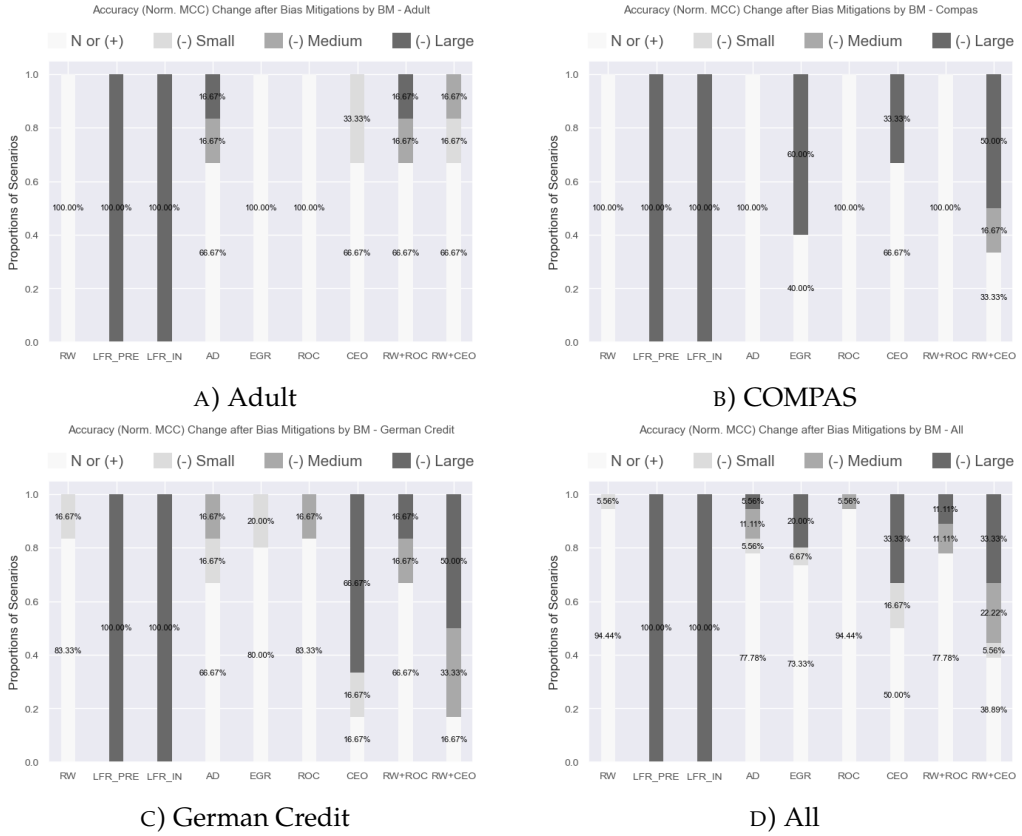


A) Adult B) COMPAS

C) German Credit D) All

FIGURE 4.8: Accuracy Change after Bias Mitigations by Base Estimators

Accuracy Change Among Bias Mitigations Figure 4.9 compares different bias mitigation methods in terms of decrease in accuracy, indicated by normalized MCC. The empirical findings reveal that the impact of bias mitigation methods on accuracy varies widely depending on the method and the dataset. LFR, for example, consistently results in significant decreases in accuracy, while RW appears to have little impact on accuracy in most cases. Moreover, EGR and CEO show a more varied effect on accuracy depending on the dataset. EGR results in minimal accuracy reduction in the Adult and the German Credit datasets. On the other hand, in the COMPAS dataset, EGR decreases accuracy significantly in 60% of the cases. Another examined bias mitigation method, CEO, yields a minor decrease in accuracy in only a third of the cases for the Adult dataset. However, in the COMPAS dataset, CEO results in a considerable reduction in accuracy. Notably, the accuracy decrease is substantial in the German Credit dataset, where CEO significantly lowers the normalized MCC metric by 83.37%, with 66.67% being a large reduction and 16.67% small reduction.



A) Adult B) COMPAS

C) German Credit D) All

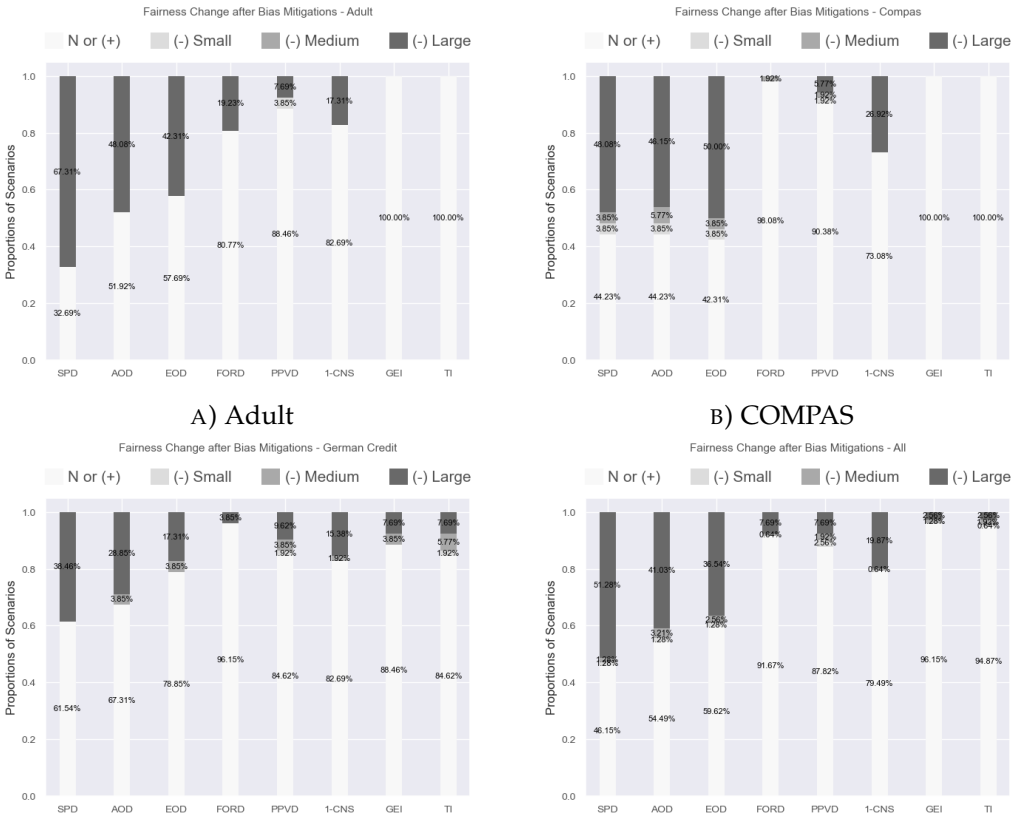
FIGURE 4.9: Accuracy Change after Bias Mitigations by BM

Fairness Change

The present section examines the impact of bias mitigation methods on fairness. The analysis in this section adopts the methodology presented in Section 4.2.4, with a focus on fairness metrics. The subsequent analyses begin by comparing the changes in various fairness metrics, including five group fairness metrics and three individual fairness metrics. Subsequently, with the chosen group fairness metric for each dataset, the analyses are carried out based on base estimators and bias mitigation methods. Unlike accuracy metrics, where the optimal value is 1, for fairness metrics, the optimal value is zero. Therefore, a significant decrease in fairness metrics is desirable as it indicates a reduction in bias.

Fairness Change Among Metrics As depicted in Figure 4.10, the efficacy of bias mitigation methods varies across different fairness metrics. Notably, the three individual fairness metrics on the right side show no or minimal improvement after implementing bias mitigation methods, which is anticipated as the methods are primarily designed to improve group fairness, and enhancing group fairness does not guarantee improvement in individual fairness Dwork et al. (2012). However, FORD and PPVD, which are

group fairness criteria, remain largely unimproved after bias mitigation, particularly in the COMPAS dataset. Fairness measured by other criteria such as SPD, AOD, and EOD exhibit comparatively better improvement than other metrics, but also not much. On average, bias mitigation methods fail to enhance fairness in 53.42% of the scenarios, except for SPD in the Adult dataset. Notably, the COMPAS dataset displays the least improvement in fairness after bias mitigation, with more than 60% of the scenarios not showing any improvement, regardless of the fairness metric.



A) Adult B) COMPAS

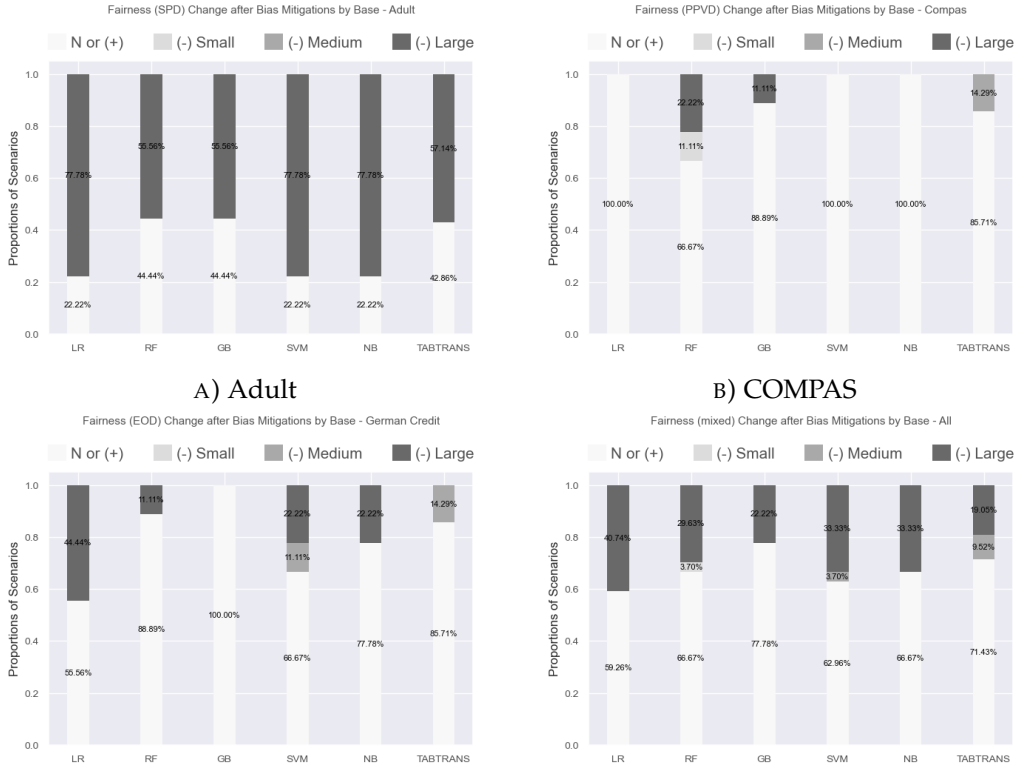
C) German Credit D) All

FIGURE 4.10: Fairness Change after Bias Mitigations

Fairness Change Among Base Estimators For this section and next section, the analyses focus on one fairness metric for each dataset, the metrics are SPD, PPVD and EOD for the Adult, COMPAS and German Credit dataset respectively. The aim is to investigate whether the fairness changes differently among different base estimators after applying bias mitigation methods.

The findings depicted in Figure 4.11a) indicate that bias mitigation methods demonstrate a high level of effectiveness in enhancing fairness for almost all base estimators in the Adult dataset. Conversely, the results for the COMPAS dataset display a contrary picture, with no statistically significant improvement in fairness observed

for LR, SVM, and NB, and only minimal enhancement for RF, GB, and TabTrans. The efficacy of bias mitigation varies based on the choice of base estimator for the German Credit dataset, with LR models demonstrating improvement in fairness in almost half of the scenarios, while other base estimators show little or no improvement.



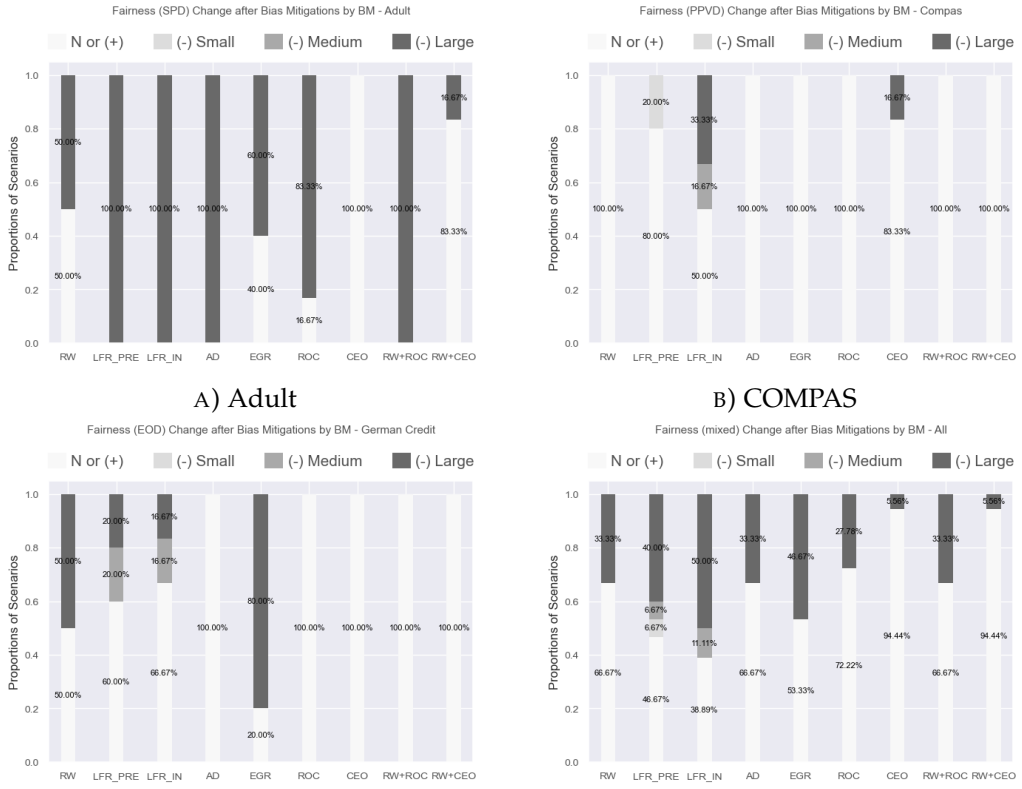
A) Adult B) COMPAS

C) German Credit D) All

FIGURE 4.11: Fairness Change after Bias Mitigations by Base Estimators

Fairness Change Among Bias Mitigations The results in Figure 4.12 demonstrate that LFR models exhibit the highest efficacy among the BMs, with EGR, RW, AD, and RW+ROC following suit. CEO and RW+CEO, on the other hand, appear to be the least effective. For the Adult dataset, most BMs are quite effective, except for CEO and RW+CEO, which have no notable effect on fairness. In contrast, bias mitigation methods fail to enhance fairness on the COMPAS dataset, where only LFRs and CEO exhibit a statistically significant effect in some few cases. For the German Credit dataset, AD and methods involving ROC and CEO have no significant effect on fairness, whereas EGR and RW are rather effective, with LFRs also mitigating bias in some cases. In contrast to the accuracy analysis, where mixed approaches consistently decrease accuracy more than single methods, mixed approaches do not consistently affect fairness more. For instance, RW+ROC performs better than both RW and ROC on the Adult dataset, but RW+CEO performs worse than RW. Similarly, RW+CEO performs worse than CEO

on the COMPAS dataset, and both RW+ROC and RW+CEO perform worse than RW on the German Credit dataset.



A) Adult B) COMPAS

C) German Credit D) All

FIGURE 4.12: Fairness Change after Bias Mitigations by BM

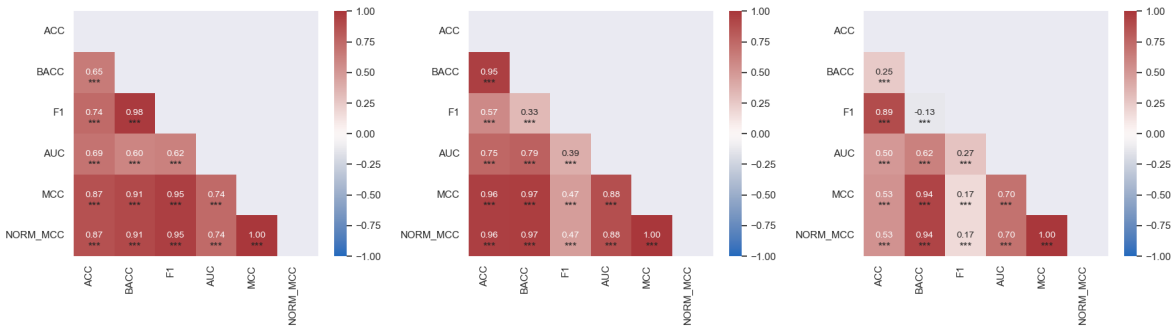
Correlation between Metrics

In order to better understand the relationship between model accuracy and fairness, as well as group fairness and individual fairness, the following sections illustrate the Spearman's rank correlation coefficient ρ between every pair of the metrics including in the experiments with heat maps.

The Spearman's rank correlation coefficient is a non-parametric measure of the correlation between two variables, it assesses the monotonic relationship between the variables without making assumptions about their distribution. Spearman's ρ ranges from -1 to 1, where -1 indicates a perfect negative monotonic correlation, 0 indicates no monotonic correlation, and 1 indicates a perfect positive monotonic correlation. A positive ρ implies that as one variable increases, the other also tends to increase. Conversely, a negative ρ indicates an inverse relationship, signifying that as one variable increases, the other tends to decrease. For each metric pair, the overall ρ along with

stars indicating the statistical significance level is presented (i.e., *, **, *** indicating p-value < 0.05, 0.01, 0.001, respectively).

Accuracy Metrics Figure 4.13 illustrates the Spearman's rank correlation coefficient ρ between accuracy metrics. The results indicate that almost all metric-pairs exhibit a positive correlation, implying that accuracy metrics tend to move in the same direction; when one metric increases, the others are likely to increase as well. However, the degree of correlation differs considerably across the datasets. For the Adult dataset, a high correlation is observed between most metric-pairs, particularly between (F1, BACC) and MCC with all other metrics except AUC. The COMPAS dataset exhibits similar behavior, albeit with a larger reduction in the correlation between F1 and the other metrics. In contrast, for the German Credit dataset, only a few exceptions, such as (F1, ACC) and (MCC, BACC), exhibit a high correlation between accuracy metrics, and other metric-pairs show no such relationship. It is noteworthy that the degree of correlation from most metric-pairs varies across different datasets, indicating that the relationship between accuracy metrics is influenced by the specific characteristics of the dataset. The only exception to this trend is the (MCC/NORM_MCC, BACC) pair, which consistently shows a high positive correlation across all datasets. Figure 4.14



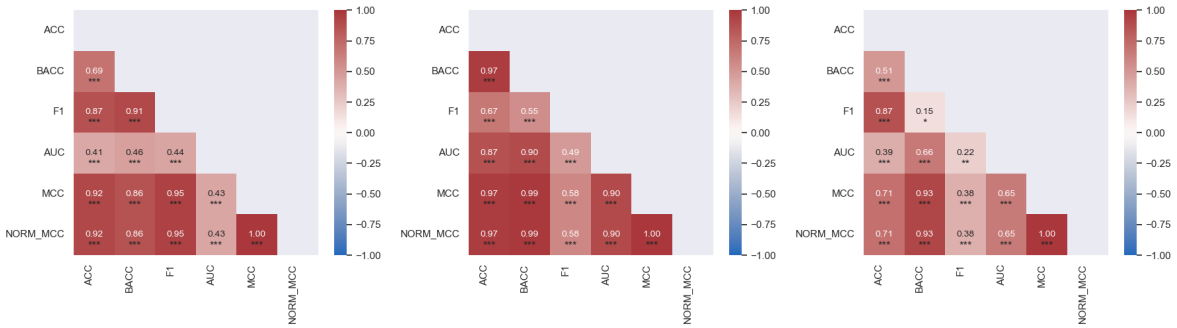
A) Adult B) COMPAS C) German Credit

FIGURE 4.13: Correlation between Accuracy Metrics

depicts the Spearman's rank correlation coefficient ρ between the changes in accuracy metrics resulting from the application of bias mitigation techniques. Specifically, the differences in the values of each accuracy metric before and after bias mitigation are calculated, and the overall correlation between the differences of each metric-pair is computed.

The overall trend of the correlation is similar to the findings reported in the previous analysis. Positive correlation, in this context, indicates that the metrics respond similarly to the implementation of bias mitigation methods. However, for the Adult dataset, the correlation of changes in accuracy metrics after bias mitigation decreases

compared to the correlation of the metrics themselves. This observation suggests that, although the metrics are somewhat correlated, the AUC metric reacts differently to bias mitigation compared to the other metrics for this dataset. In contrast, for the German Credit dataset, the correlation of changes in accuracy metrics between the (MCC, ACC) pair increases compared to the correlation of the metrics themselves. This indicates that although the metrics themselves are not highly correlated, they respond similarly to the implementation of bias mitigation. Again, (MCC/NORM_MCC, BACC) is the only metric-pair that consistently shows high correlation across all datasets.



A) Adult B) COMPAS C) German Credit

FIGURE 4.14: Correlation between Changes of Accuracy Metrics after BM

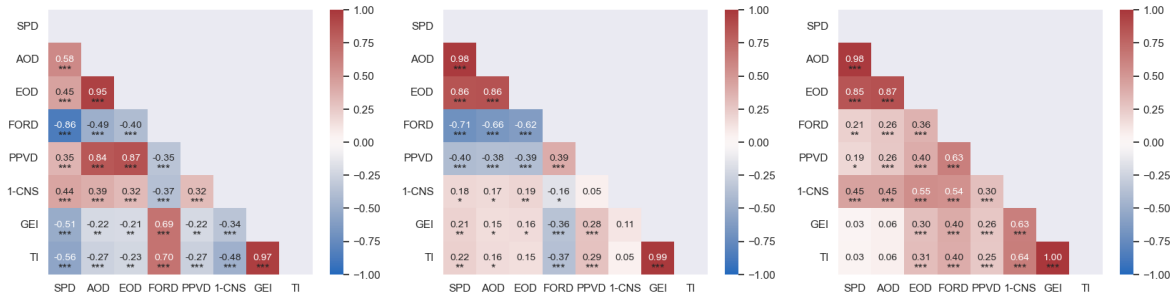
Fairness Metrics Figure 4.13 depicts the Spearman's rank correlation coefficient ρ between fairness metrics, revealing that the correlation between these metrics varies substantially across different datasets. The metrics of SPD, EOD, and AOD generally exhibit higher correlation with each other, particularly in the COMPAS and German Credit datasets. In the Adult dataset, SPD is not as strongly correlated with the other two, but EOD and AOD are still highly correlated. Furthermore, the FORD metric generally shows negative correlation with other fairness metrics in Adult and COMPAS datasets, indicating that when FORD increases, the other fairness tend to metrics decrease. This negative correlation is especially high for the (FORD, SPD) metric pair in the Adult dataset. In contrast, there is no apparent negative correlation in the German Credit dataset as observed in the other two datasets.

In this study, the same methodology as in the previous analysis is adopted to examine the impact of bias mitigation methods on fairness metrics. The outcomes of this analysis are presented in Figure 4.16. As observed in the findings for accuracy metrics, the general trend from the correlation of changes in fairness metrics after bias mitigation is similar to the correlations between the metrics themselves. However, the response of fairness metrics to the application of bias mitigation methods varies across datasets, and the correlations are not consistent across all metrics.

A) Adult B) COMPAS C) German Credit

FIGURE 4.15: Correlation between Fairness Metrics

The FORD metric, which displays a negative correlation with other fairness metrics in the Adult and COMPAS datasets, also exhibits a negative correlation in terms of changes after bias mitigation. This implies that bias mitigation methods that result in a decrease in FORD value often increase bias in terms of other metrics. Notably, the negative correlation between FORD and SPD is particularly high in the Adult dataset.

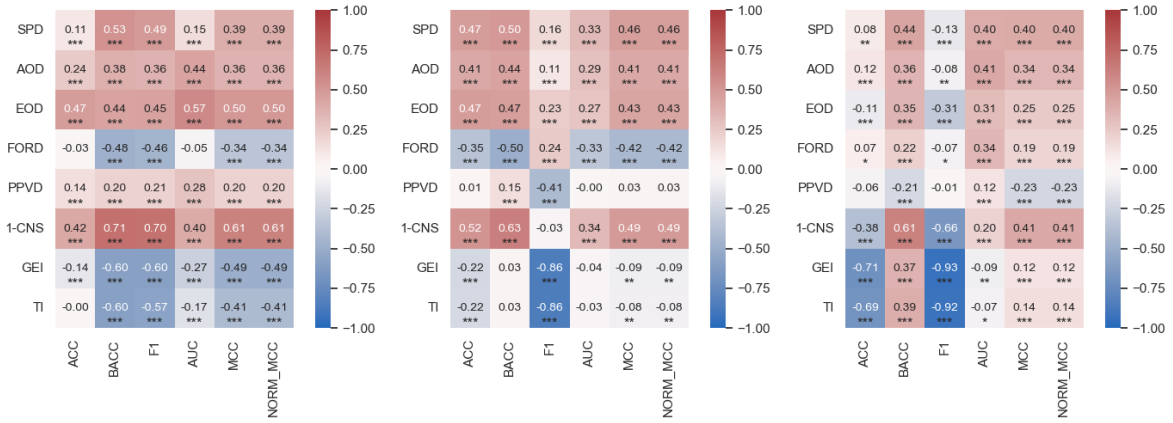


A) Adult B) COMPAS C) German Credit

FIGURE 4.16: Correlation between Changes of Fairness Metrics after BM

Accuracy and Fairness Metrics Figure 4.17 illustrates the correlation between accuracy and fairness metrics, most of the correlation coefficients which are not close to zero are statistically significant. It should be noted that the optimal value for fairness metrics is zero, and thus a positive relationship here indicates a trade-off between accuracy and fairness. Although there exist several positive correlations between accuracy and fairness metrics, the values are all quite small, indicating weak correlation. Nonetheless, two exceptions exist in the Adult dataset, where (1-CNS, BACC) and (1-CNS, F1) exhibit a correlation coefficient of 0.7. Furthermore, some metric-pairs exhibit high negative correlation, such as (GEI/TI, F1) in both COMPAS and German Credit datasets, and (GEI/TI,ACC) in the German Credit dataset. These results indicate that when accuracy in terms of F1 or ACC improves, individual fairness GEI/TI also improves.

In summary, while there is some evidence of a negative relationship between accuracy and fairness for certain fairness metrics, the overall correlation coefficients are small, indicating that the relationship is not strong. Furthermore, for some of the fairness metrics, it appears that it is possible to achieve both high accuracy and fairness simultaneously. This observation is contrary to the common belief in the existing literature that often assumes a trade-off between accuracy and fairness.



A) Adult B) COMPAS C) German Credit

FIGURE 4.17: Correlation between Accuracy and Fairness Metrics

Chapter 5

Discussion

The proposed FairGridSearch approach provides two main benefits. Firstly, it facilitates an easy implementation and comparison of different fairness-enhancing models. Secondly, it suggests the best model for a given dataset or application.

The literature has recently seen a multitude of proposals for fairness metrics and bias mitigation methods. While studies comparing different methods exist, their findings are typically limited to a smaller set of model combinations and specific datasets. Additionally, they do not provide suggestions on the most suitable models among all experimented ones.

To further investigate the fairness of ML models and the impacts of different bias mitigation methods, a unified framework that enables researchers to try out multiple fairness-enhancing models is needed. FairGridSearch addresses this need by incorporating multiple base estimators with bias mitigation methods, focusing on the most common methods provided by the AIF360 package. This allows practitioners to implement and analyze different models with varying base estimators in combination with bias mitigation methods. Additionally, the approach includes base estimator-specific parameters such as penalty functions for logistic regression algorithms and alternative classification thresholds, which are not considered in most of the previous research.

Furthermore, FairGridSearch offers the advantage of suggesting the most suitable model based on the accuracy and fairness of the evaluated models. Given the availability of multiple base estimators and bias mitigation methods, it can be challenging to identify the model that optimally balances accuracy and fairness for a particular dataset or application. In fact, the results of this study show that no single base estimator, threshold value or bias mitigation method is universally effective across all datasets, let alone one that optimally balances accuracy and fairness. Consequently, determining the best model for a certain dataset requires trying out multiple models, which can be a cumbersome task. By utilizing the FairGridSearch approach, researchers and practitioners can receive a recommendation for the most suitable fairness-enhancing model for their specific application. This not only streamlines

the model selection process but also enhances the overall efficiency and effectiveness of the decision-making process.

Based on the experiment conducted in this study using FairGridSearch, the research questions raised in Section 1 can be answered as follows:

RQ1: (Metrics) For both accuracy and fairness metrics, the choice of metric affects model evaluation since they either are not correlated with each other or react differently towards bias mitigation methods.

Accuracy: There is a wide range of accuracy metrics available in the literature, yet numerous studies in the field of fairness in ML tend to concentrate on a limited set of metrics, primarily ACC and F1. In contrast, FairGridSearch incorporates multiple accuracy metrics, providing the opportunity to investigate and compare the accuracy of models measured by different metrics and to understand if the choice of metric affects the results. The results from Figure 4.7 demonstrate that different accuracy metrics may produce different outcomes after applying bias mitigation methods. In particular, AUC exhibited large decreases most frequently in all three datasets. Additionally, while some accuracy metrics are highly correlated with each other in the Adult and COMPAS dataset, only a few are correlated in German Credit, as illustrated in Figure 4.13. Furthermore, the changes in metrics after applying bias mitigation methods are not necessarily correlated with each other, as demonstrated in Figure 4.14. For instance, in the Adult dataset, the change in AUC after bias mitigation is not correlated with the changes in other metrics. Similarly, in COMPAS, F1 is not correlated with most of the other metrics, and in German Credit, most of the metrics are not correlated with each other. This means, the metrics don't react similarly after applying bias mitigation. The results suggest that the strength and pattern of correlation between accuracy metrics vary considerably across different contexts, highlighting the importance of carefully selecting and interpreting the appropriate metrics for each use case.

Fairness: Similar to accuracy metrics, FairGridSearch incorporates multiple fairness metrics. From Figure 4.10, it is evident that model fairness measured by different metrics varies after applying bias mitigation, especially in the Adult and COMPAS datasets. Furthermore, as shown in Figure 4.15, correlation between different fairness metrics varies significantly between metric-pairs and across different datasets. For instance, SPD exhibits a high correlation with AOD and EOD in COMPAS and German Credit datasets, but not in the Adult dataset. It is noteworthy that FORD exhibits a negative correlation with the other metrics in the Adult dataset, suggesting that its response to bias mitigation employed in the study is in contrast to that of the other fairness metrics. Specifically, the mitigation techniques that improve the values of other fairness metrics decrease the value of FORD. Moreover, as demonstrated in Figure

4.16, the changes in fairness metrics after bias mitigation are not consistently correlated across different metric-pairs. Even in cases where certain metric-pairs exhibit high correlation in a particular dataset, such as PPVD and AOD in the Adult dataset, this high correlation doesn't exist in other datasets. Notably, the only metric-pair that shows consistent high positive correlation across all datasets is AOD and EOD.

These results emphasize the criticality of selecting appropriate metrics that align with the application under consideration, as selecting an unsuitable metric may lead to a distorted understanding of a model's accuracy and fairness. To this end, this study exclusively employs NORM_MCC as the accuracy metric for selecting the best model, as it has been shown to be the more robust choice in recent research. As for fairness metrics, the selection is guided by fairness tree (Saleiro et al., 2019). However, in scenarios where the most suitable accuracy and fairness metrics for the application is uncertain, it is recommended to incorporate a diverse set of metrics in the evaluation process to obtain a more comprehensive assessment of the models.

RQ2: (Base Estimator) No single base estimator yields consistently fairer models than the others across datasets. However, the sensitivity of base estimators to bias mitigation can vary depending on the dataset.

The influence of base estimators on model fairness has not been extensively discussed in the literature. Therefore, with multiple base estimators used to construct models in FairGridSearch, this study attempts to investigate their impact on fairness. From Figure 4.1c), 4.3c), 4.5c), it can be observed that no base estimator consistently yields fairer models than the others across all datasets. Nevertheless, the choice of base estimator has been found to exhibit varying degrees of sensitivity to bias mitigation in different datasets. Figure 4.11 highlights the inconsistency of the effectiveness of bias mitigation on different base estimators across datasets. Specifically, while all base estimators in the Adult dataset improve model fairness after bias mitigation, half of the base estimators in COMPAS show no improvement in fairness. Furthermore, certain base estimators such as GB have been found to improve fairness in COMPAS but not in German Credit, making it evident that the effectiveness of bias mitigation is not consistent across different base estimators and datasets.

RQ3: (Classification Threshold) Classification threshold values could yield different volatility in model fairness, but the ones yielding high volatility could reach high fairness and accuracy at the same time, while threshold values that have more stability in fairness across models tend to have lower accuracy.

The impact of the classification threshold on model accuracy has been discussed in the literature, but its effect on fairness has has received less attention. The examination of the top-performing models for each dataset suggests that there is a preferred

threshold value for each dataset. Specifically, the top models in the Adult dataset are mainly composed of models with a lower threshold value, while for the COMPAS dataset, the preferred threshold value is around 0.5. In contrast, the top models for the German Credit dataset tend to have higher threshold values. However, it should be noted that the threshold values that yield the highest accuracy may also generate high volatility in fairness, as shown in the Adult and German Credit datasets. In contrast, threshold values that have greater stability in fairness tend to prevent the model from reaching its highest accuracy potential. Given the variability in optimal classification threshold across datasets, and the significant impact of this threshold and other model parameters on fairness, it is advisable to include a comprehensive set of models in the comparison to identify the most suitable one.

RQ4: (Trade-Off) There's no clear trade-off between accuracy and fairness for the datasets experimented in this study.

In the field of fairness in ML, it is often assumed that improving model fairness comes at the expense of accuracy, suggesting a trade-off between the two objectives (Berk et al., 2021, Haas, 2019, Janssen and Sadowski, 2021). However, the findings of this study challenge this belief, as demonstrated by Figures 4.1a), 4.3a), 4.5a), and 4.17. While there is some evidence of a negative relationship between accuracy and fairness for certain fairness metrics, the overall correlation coefficients are small, indicating that the relationship is not strong. In fact, for some of the fairness metrics, there exists a slightly positive relationship with accuracy, such as FORD with multiple accuracy metrics in the Adult and COMPAS dataset, or individual fairness GEI with F1 in all three datasets. These results suggest that achieving high accuracy and fairness simultaneously is possible, at least for the specific datasets and models studied in this research. This observation is in contrast to the common assumption in the literature that accuracy and fairness are mutually exclusive objectives.

Chapter 6

Limitations

This study presents a novel grid search framework for fairness-enhancing models, but several limitations should be acknowledged. Firstly, the framework is designed only for binary classification problems and does not currently extend to multi-class classification problems. Secondly, the TabTransformer approach is incompatible with several of the base estimators included in the study. Thirdly, while the framework includes seven commonly used bias mitigation methods and two mixed approaches, there are numerous other methods available in the literature that have not been considered. Finally, while grid search is an effective approach for identifying optimal parameter combinations, the computational resources required to evaluate all possible combinations may be substantial. To address this limitation, alternative parameter optimization methods such as random search could be beneficial, since it's more efficient in scenarios where computational constraints are a concern.

In addition to these framework limitations, it is also important to acknowledge the limitations of the datasets used in the exemplar experiments. Specifically, the three datasets used - Adult, COMPAS, and German Credit - may not be representative of all possible datasets, and the generalizability of the results to other datasets should be carefully considered. Future work could address this limitation by expanding the evaluation to a wider range of datasets.

Chapter 7

Conclusion

The field of fairness in ML is rapidly evolving, with researchers and practitioners developing new metrics and methods to ensure that machine learning models are fair. However, the task of selecting the optimal approach from a plethora of base estimators and bias mitigation methods remains challenging.

This study focuses on binary classification and proposes the FairGridSearch approach, which facilitates the implementation and comparison of various fairness-enhancing models and suggests the most suitable model for a given application. The study also addresses several research questions regarding the impact of metrics, base estimators, classification threshold values, and the trade-off between accuracy and fairness with three popular datasets in the field. Specifically, the findings reveal that (1) both accuracy and fairness metrics' selection affects model evaluation since they may not be correlated or react differently towards bias mitigation methods; (2) no single base estimator clearly produces fairer models than others in all three datasets. However, the sensitivity of base estimators to bias mitigation may vary depending on the dataset; (3) classification threshold values may result in different volatility in model fairness, but those yielding high volatility may achieve high fairness and accuracy simultaneously. Conversely, threshold values with more fairness stability across models tend to have lower accuracy; (4) there is no clear trade-off between accuracy and fairness for the datasets examined in this study. Findings (2) and (3) highlight the need of taking base estimators and classification threshold values into account when determining the most optimal model that balances accuracy and fairness.

In light of these results, it is recommended that future studies on fairness in machine learning should broaden their focus to encompass a wider range of factors when building fair models, beyond bias mitigation methods alone. In the future, more bias mitigation methods should be included in FairGridSearch and more datasets should be analyzed to further explore the research questions. Furthermore, given the computational expense of full grid search, integrating options like random search to expedite the algorithm could increase its feasibility.

Chapter 8

References

Adebayo, J. A. (2016). *FairML : ToolBox for Diagnosing Bias in Predictive Modeling*. Thesis, Massachusetts Institute of Technology.

Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., and Wallach, H. (2018). A Reductions Approach to Fair Classification.

AHRQ. (2015). Medical expenditure panel survey data: 2015.

AHRQ. (2016). Medical expenditure panel survey data: 2016.

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. (2016). Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Baldi, P., Brunak, S., Chauvin, Y., Andersen, C. A. F., and Nielsen, H. (2000). Assessing the accuracy of prediction algorithms for classification: An overview. *Bioinformatics*, 16(5):412–424.

Bantilan, N. (2017). Themis-ml: A Fairness-aware Machine Learning Interface for End-to-end Discrimination Discovery and Mitigation.

Barocas, S., Hardt, M., and Narayanan, A. (2019). Fairness and Machine Learning: Limitations and Opportunities. *fairmlbook.org*.

Bechavod, Y. and Ligett, K. (2018). Penalizing Unfairness in Binary Classification.

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., Saha, D., Sattigeri, P., Singh, M., Varshney, K. R., and Zhang, Y. (2018). AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias.

Bergstra, J. and Bengio, Y. (2012). Random Search for Hyper-Parameter Optimization.

Berk, R., Heidari, H., Jabbari, S., Kearns, M., and Roth, A. (2021). Fairness in Criminal Justice Risk Assessments: The State of the Art. *Sociological Methods & Research*, 50(1):3–44.

Bird, S., Dudík, M., Edgar, R., Horn, B., Lutz, R., Milan, V., Sameki, M., Wallach, H., Walker, K., and Design, A. (2020). Fairlearn: A toolkit for assessing and improving fairness in AI. page 7.

Biswas, S. and Rajan, H. (2020). Do the machine learning models on a crowd sourced platform exhibit bias? an empirical study on model fairness. In *Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, ESEC/FSE 2020, pages 642–653, New York, NY, USA. Association for Computing Machinery.

Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V., and Kalai, A. (2016). Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings.

Brown, J. B. (2018). Classifiers and their Metrics Quantified. *Molecular Informatics*, 37(1-2):1700127.

Buyl, M., Cociancig, C., Frattone, C., and Roekens, N. (2022). Tackling Algorithmic Disability Discrimination in the Hiring Process: An Ethical, Legal and Technical Analysis. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1071–1082, Seoul Republic of Korea. ACM.

Calders, T., Kamiran, F., and Pechenizkiy, M. (2009). Building Classifiers with Independency Constraints. In *2009 IEEE International Conference on Data Mining Workshops*, pages 13–18, Miami, FL, USA. IEEE.

Calders, T. and Verwer, S. (2010). Three naive Bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, 21(2):277–292.

Calmon, F., Wei, D., Vinzamuri, B., Natesan Ramamurthy, K., and Varshney, K. R. (2017). Optimized Pre-Processing for Discrimination Prevention. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Canbek, G., Taskaya Temizel, T., and Sagiroglu, S. (2021). BenchMetrics: A systematic benchmarking method for binary classification performance metrics. *Neural Computing and Applications*, 33(21):14623–14650.

Carey, A. N. and Wu, X. (2022). The statistical fairness field guide: Perspectives from social and formal sciences. *AI and Ethics*.

Caton, S. and Haas, C. (2020). Fairness in Machine Learning: A Survey.

Celis, L. E., Huang, L., Keswani, V., and Vishnoi, N. K. (2020). Classification with Fairness Constraints: A Meta-Algorithm with Provable Guarantees.

Chakraborty, J., Majumder, S., Yu, Z., and Menzies, T. (2020). Fairway: A way to build fair ML software. In *Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, ESEC/FSE 2020, pages 654–665, New York, NY, USA. Association for Computing Machinery.

Chatterjee, A., Paul, S., and Agneeswaran, V. (2021). A Model-Agnostic Framework to Correct Label-Bias in Training Data Using a Sample of Trusted Data. In Arai, K., editor, *Advances in Information and Communication*, Advances in Intelligent Systems and Computing, pages 431–443, Cham. Springer International Publishing.

Chen, Z., Zhang, J. M., Sarro, F., and Harman, M. (2023). A Comprehensive Empirical Study of Bias Mitigation Methods for Machine Learning Classifiers.

Chicco, D. and Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, 21(1):6.

Chicco, D., Starovoitov, V., and Jurman, G. (2021a). The Benefits of the Matthews Correlation Coefficient (MCC) Over the Diagnostic Odds Ratio (DOR) in Binary Classification Assessment. *IEEE Access*, 9:47112–47124.

Chicco, D., Warrens, M. J., and Jurman, G. (2021b). The Matthews Correlation Coefficient (MCC) is More Informative Than Cohen's Kappa and Brier Score in Binary Classification Assessment. *IEEE Access*, 9:78368–78381.

Chouldechova, A. (2016). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments.

Chouldechova, A. and Roth, A. (2018). The Frontiers of Fairness in Machine Learning.

Cleary, T. A. (1968). Test Bias: Prediction of Grades of Negro and White Students in Integrated Colleges. *Journal of Educational Measurement*, 5(2):115–124.

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., and Huq, A. (2017). Algorithmic decision making and the cost of fairness.

Darlington, R. B. (1971). Another Look at "Cultural Fairness"1. *Journal of Educational Measurement*, 8(2):71–82.

Dua, D. and Graff, C. (2017). UCI machine learning repository.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, ITCS '12, pages 214–226, New York, NY, USA. Association for Computing Machinery.

Fabris, A., Messina, S., Silvello, G., and Susto, G. A. (2022). Algorithmic Fairness Datasets: The Story so Far. *Data Mining and Knowledge Discovery*, 36(6):2074–2152.

Farnadi, G., Babaki, B., and Getoor, L. (2018). Fairness in Relational Domains. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 108–114, New Orleans LA USA. ACM.

Feldman, M., Friedler, S., Moeller, J., Scheidegger, C., and Venkatasubramanian, S. (2015). Certifying and removing disparate impact.

Foulds, J., Islam, R., Keya, K. N., and Pan, S. (2019). An Intersectional Definition of Fairness.

Freund, Y. and Schapire, R. E. (1996). Game theory, on-line prediction and boosting. In *Proceedings of the Ninth Annual Conference on Computational Learning Theory - COLT '96*, pages 325–332, Desenzano del Garda, Italy. ACM Press.

Friedler, S. A., Scheidegger, C., Venkatasubramanian, S., Choudhary, S., Hamilton, E. P., and Roth, D. (2018). A comparative study of fairness-enhancing interventions in machine learning.

Galhotra, S., Brun, Y., and Meliou, A. (2017). Fairness testing: Testing software for discrimination. In *Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering*, ESEC/FSE 2017, pages 498–510, New York, NY, USA. Association for Computing Machinery.

Garg, N., Schiebinger, L., Jurafsky, D., and Zou, J. (2018). Word Embeddings Quantify 100 Years of Gender and Ethnic Stereotypes. *Proceedings of the National Academy of Sciences*, 115(16).

Ghai, B., Mishra, M., and Mueller, K. (2022). Cascaded Debiasing: Studying the Cumulative Effect of Multiple Fairness-Enhancing Interventions.

Goh, G., Cotter, A., Gupta, M., and Friedlander, M. P. (2016). Satisfying Real-world Goals with Dataset Constraints. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative Adversarial Networks.

Gösgens, M., Zhiyanov, A., Tikhonov, A., and Ostroumova Prokhorenkova, L. (2022). *Good Classification Measures and How to Find Them*.

Haas, C. (2019). The Price of Fairness - A Framework to Explore Trade-Offs in Algorithmic Fairness. *ICIS 2019 Proceedings*.

Hall, M., van der Maaten, L., Gustafson, L., Jones, M., and Adcock, A. (2022). A Systematic Study of Bias Amplification.

Hamilton, E. and Friedler, S. (2017). Benchmarking Four Approaches to Fairness-Aware Machine Learning.

Hardt, M., Price, E., Price, E., and Srebro, N. (2016). Equality of Opportunity in Supervised Learning. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Hellström, T., Dignum, V., and Bensch, S. (2020). Bias in Machine Learning – What is it Good for?

Hooker, S. (2021). Moving beyond "algorithmic bias is a data problem". *Patterns*, 2(4):100241.

Hort, M., Chen, Z., Zhang, J. M., Sarro, F., and Harman, M. (2022). Bias Mitigation for Machine Learning Classifiers: A Comprehensive Survey.

Hort, M., Zhang, J. M., Sarro, F., and Harman, M. (2021). Fairea: A model behaviour mutation approach to benchmarking bias mitigation methods. In *Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, ESEC/FSE 2021, pages 994–1006, New York, NY, USA. Association for Computing Machinery.

Hossin, M. and Sulaiman, M. (2015). A Review on Evaluation Metrics for Data Classification Evaluations. *International Journal of Data Mining & Knowledge Management Process*, 5(2):01–11.

Huang, X., Khetan, A., Cvitkovic, M., and Karnin, Z. (2020). TabTransformer: Tabular Data Modeling Using Contextual Embeddings.

Hufthammer, K. T., Aasheim, T. H., Ånneland, S., Brynjulfsen, H., and Slavkovik, M. (2020). Bias mitigation with AIF360: A comparative study. *Norsk IKT-konferanse for forskning og utdanning*, (1).

Janssen, P. and Sadowski, B. M. (2021). Bias in Algorithms: On the trade-off between accuracy and fairness.

Johndrow, J. E. and Lum, K. (2017). An algorithm for removing sensitive information: Application to race-independent recidivism prediction.

Kamiran, F. and Calders, T. (2009). *Classifying without Discriminating*.

Kamiran, F. and Calders, T. (2012). Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1):1–33.

Kamiran, F., Karim, A., and Zhang, X. (2012). Decision Theory for Discrimination-Aware Classification. In *2012 IEEE 12th International Conference on Data Mining*, pages 924–929, Brussels, Belgium. IEEE.

Kamiran, F., Žliobaitė, I., and Calders, T. (2013). Quantifying explainable discrimination and removing illegal discrimination in automated decision making. *Knowledge and Information Systems*, 35(3):613–644.

Kamishima, T., Akaho, S., Asoh, H., and Sakuma, J. (2012). Fairness-Aware Classifier with Prejudice Remover Regularizer. In Hutchison, D., Kanade, T., Kittler, J., Kleinberg, J. M., Mattern, F., Mitchell, J. C., Naor, M., Nierstrasz, O., Pandu Rangan, C., Steffen, B., Sudan, M., Terzopoulos, D., Tygar, D., Vardi, M. Y., Weikum, G., Flach, P. A., De Bie, T., and Cristianini, N., editors, *Machine Learning and Knowledge Discovery in Databases*, volume 7524, pages 35–50. Springer Berlin Heidelberg, Berlin, Heidelberg.

Kleinberg, J., Mullainathan, S., and Raghavan, M. (2016). Inherent Trade-Offs in the Fair Determination of Risk Scores.

Kohavi, R. (1996). Scaling Up the Accuracy of Naive-Bayes Classifiers: A Decision-Tree Hybrid. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*.

Kozodoi, N., Jacob, J., and Lessmann, S. (2022). Fairness in Credit Scoring: Assessment, Implementation and Profit Implications. *European Journal of Operational Research*, 297(3):1083–1094.

Kumar, I. E., Hines, K. E., and Dickerson, J. P. (2022). Equalizing Credit Opportunity in Algorithms: Aligning Algorithmic Fairness Research with U.S. Fair Lending Regulation. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pages 357–368, Oxford United Kingdom. ACM.

Kusner, M. J., Loftus, J., Russell, C., and Silva, R. (2017). Counterfactual Fairness. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Larson, J., Angwin, J., Kirchner, L., and Mattu, S. (2016). Machine bias: There's software used across the country to predict future criminals. And its biased against blacks.

Louizos, C., Swersky, K., Li, Y., Welling, M., and Zemel, R. (2017). The Variational Fair Autoencoder.

Luong, B. T., Ruggieri, S., and Turini, F. (2011). K-NN as an implementation of situation testing for discrimination discovery and prevention. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 502–510, San Diego California USA. ACM.

Luque, A., Carrasco, A., Martín, A., and de las Heras, A. (2019). The impact of class imbalance in classification performance metrics based on the binary confusion matrix. *Pattern Recognition*, 91:216–231.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. (2022). A Survey on Bias and Fairness in Machine Learning.

Moro, S., Cortez, P., and Rita, P. (2014). A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, 62:22–31.

Niculescu-Mizil, A. and Caruana, R. (2005). Predicting good probabilities with supervised learning. In *Proceedings of the 22nd International Conference on Machine Learning - ICML '05*, pages 625–632, Bonn, Germany. ACM Press.

Ntoutsi, E., Fafalios, P., Gadiraju, U., Iosifidis, V., Nejdl, W., Vidal, M.-E., Ruggieri, S., Turini, F., Papadopoulos, S., Krasanakis, E., Kompatsiaris, I., Kinder-Kurlanda, K., Wagner, C., Karimi, F., Fernandez, M., Alani, H., Berendt, B., Kruegel, T., Heinze, C., Broelemann, K., Kasneci, G., Tiropanis, T., and Staab, S. (2020). Bias in data-driven artificial intelligence systems—An introductory survey. *WIREs Data Mining and Knowledge Discovery*, 10(3):e1356.

Olteanu, A., Castillo, C., Diaz, F., and Kıcıman, E. (2019). Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries. *Frontiers in Big Data*, 2.

Pessach, D. and Shmueli, E. (2022). A Review on Fairness in Machine Learning. *ACM Computing Surveys*, 55(3):51:1–51:44.

Platt, J. (2000). Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. *Adv. Large Margin Classif.*, 10.

Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., and Weinberger, K. Q. (2017). On Fairness and Calibration.

Powers, D. M. W. (2019). What the F-measure doesn't measure: Features, Flaws, Fallacies and Fixes.

Raff, E., Sylvester, J., and Mills, S. (2017). Fair Forests: Regularized Tree Induction to Minimize Model Bias.

Roth, D. (2018). A Comparison of Fairness-Aware Machine Learning Algorithms.

Saleiro, P., Kuester, B., Hinkson, L., London, J., Stevens, A., Anisfeld, A., Rodolfa, K. T., and Ghani, R. (2019). Aequitas: A Bias and Fairness Audit Toolkit.

Saxena, N., Huang, K., DeFilippis, E., Radanovic, G., Parkes, D., and Liu, Y. (2019). How Do Fairness Definitions Fare? Examining Public Attitudes Towards Algorithmic Definitions of Fairness.

Saxena, N. A. (2019). Perceptions of Fairness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '19, pages 537–538, New York, NY, USA. Association for Computing Machinery.

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., and Vertesi, J. (2019). Fairness and Abstraction in Sociotechnical Systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pages 59–68, New York, NY, USA. Association for Computing Machinery.

Speicher, T., Heidari, H., Grgic-Hlaca, N., Gummadi, K. P., Singla, A., Weller, A., and Zafar, M. B. (2018). A Unified Approach to Quantifying Algorithmic Unfairness: Measuring Individual &Group Unfairness via Inequality Indices. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2239–2248, London United Kingdom. ACM.

Suresh, H. and Guttag, J. V. (2021). A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. In *Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–9.

Syarif, I., Prugel-Bennett, A., and Wills, G. (2016). SVM Parameter Optimization using Grid Search and Genetic Algorithm to Improve Classification Performance. *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, 14(4):1502–1509.

Tolan, S., Miron, M., Gómez, E., and Castillo, C. (2019). Why Machine Learning May Lead to Unfairness: Evidence from Risk Assessment for Juvenile Justice in Catalonia. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law*, pages 83–92, Montreal QC Canada. ACM.

Tramèr, F., Atlidakis, V., Geambasu, R., Hsu, D., Hubaux, J.-P., Humbert, M., Juels, A., and Lin, H. (2016). FairTest: Discovering Unwarranted Associations in Data-Driven Applications.

Verma, S. and Rubin, J. (2018). Fairness definitions explained. In *Proceedings of the International Workshop on Software Fairness*, pages 1–7, Gothenburg Sweden. ACM.

Wan, M., Zha, D., Liu, N., and Zou, N. (2022). In-Processing Modeling Techniques for Machine Learning Fairness: A Survey. *ACM Transactions on Knowledge Discovery from Data*, page 3551390.

Woodworth, B., Gunasekar, S., Ohannessian, M. I., and Srebro, N. (2017). Learning Non-Discriminatory Predictors.

Zadrozny, B. and Elkan, C. (2001). Obtaining Calibrated Probability Estimates from Decision Trees and Naive Bayesian Classifiers. *ICML*, 1.

Zafar, M. B., Valera, I., Gomez Rodriguez, M., and Gummadi, K. P. (2017a). Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, pages 1171–1180, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Zafar, M. B., Valera, I., Rodriguez, M. G., and Gummadi, K. P. (2017b). Fairness Constraints: Mechanisms for Fair Classification.

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., and Dwork, C. (2013). Learning Fair Representations. In *Proceedings of the 30th International Conference on Machine Learning*, pages 325–333. PMLR.

Zhang, B. H., Lemoine, B., and Mitchell, M. (2018). Mitigating Unwanted Biases with Adversarial Learning.

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., and Chang, K.-W. (2017). Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints.

Zhao, J., Zhou, Y., Li, Z., Wang, W., and Chang, K.-W. (2018). Learning Gender-Neutral Word Embeddings.

Declaration of Authorship

Hiermit erkläre ich, Shih-Chi Ma, dass ich die vorliegende Arbeit noch nicht für andere Prüfungen eingereicht habe. Ich habe die Arbeit selbständig ver- fasst. Sämtliche Quellen einschließlich Internetquellen, die ich unverändert oder abgewandelt wiedergegeben habe, insbesondere Quellen für Texte, Grafiken, Tabellen und Bilder, habe ich als solche kenntlich gemacht. Ich bin mir darüber bewusst, dass bei Verstößen gegen diese Grundsätze ein Ver- fahren wegen Täuschungsversuchs bzw. Täuschung eingeleitet wird.

I, Shih-Chi Ma, hereby declare that I have not previously submitted the present work for other examinations. I wrote this work independently. All sources, including sources from the Internet, that I have reproduced in either an unaltered or modified form (particularly sources for texts, graphs, tables and images), have been acknowledged by me as such. I understand that violations of these principles will result in proceedings regarding deception or attempted deception.

_____

Location, Date:

_____

Signature