Measuring and Mitigating Bias in Machine Learning

Inaugural-Dissertation

zur

Erlangung des Doktorgrades der
Mathematisch-Naturwissenschaftlichen Fakultät
der Heinrich-Heine-Universität Düsseldorf

vorgelegt von

Manh Khoi Duong

aus Krefeld

Düsseldorf, November 2024

aus dem Institut für Informatik der
Heinrich-Heine-Universität Düsseldorf

Gedruckt mit der Genehmigung der
Mathematisch-Naturwissenschaftlichen Fakultät der
Heinrich-Heine-Universität Düsseldorf

Referent: Prof. Dr. Stefan Conrad

Koreferent: Prof. Dr. Michael Leuschel

Tag der mündlichen Prüfung: 07.04.2025

Erklärung

Ich versichere an Eides Statt, dass die vorliegende Dissertation von mir selbständig und ohne unzulässige fremde Hilfe unter Beachtung der „*Grundsätze zur Sicherung guter wissenschaftlicher Praxis an der Heinrich-Heine-Universität Düsseldorf*“ erstellt worden ist. Desweiteren erkläre ich, dass ich eine Dissertation in der vorliegenden oder in ähnlicher Form noch bei keiner anderen Institution eingereicht habe. Ich habe keinerlei andere Promotionsversuche unternommen.

Düsseldorf, November 2024

Manh Khoi Duong

Dedicated to my family

Acknowledgements

First of all, I would like to express my sincere gratitude to my supervisor, Prof. Dr. Stefan Conrad, for his guidance, support, and positive attitude throughout the years. I am most grateful for the trust and freedom he gave me to work on my ideas and letting me explore my own research interests. His insightful comments during our meetings were very helpful in shaping my research and finalizing this doctoral thesis. I am glad for the opportunity to have been part of his research group. I would also like to thank Prof. Dr. Michael Leuschel for his interest in my research and agreeing to be my secondary supervisor. Last but not least, regarding my supervisors, I would like to thank Dr. Martha Krakowski, who supervised me during my bachelor's and master's studies. I saw her as a mentor and role model career-wise, which motivated me to pursue a career in academia.

My gratitude extends to my colleagues from social science, with whom I had the pleasure of working on the interdisciplinary research project *Responsible Academic Performance Prediction*, funded by the *German Federal Ministry of Education and Research*. I learned a lot from their perspectives, which enriched my work and broadened my understanding of fairness in a social context, apart from a technical one. The research project helped me to develop novel methods for unidentified gaps in the literature that are relevant in real-world applications.

During the project, I had the pleasure to work closely with Jannik Dunkelau and José Andrés Cordova. My special thanks go to both of them. When I began my doctoral studies, Jannik Dunkelau introduced me to the topic of fairness in machine learning, which I appreciate till today. Our initial collaboration helped me a lot to grow as a researcher. I was able to take his tips to heart and enhance my skills. Publishing papers, programming, and traveling to Durham (UK) for a conference with him was a great experience. Regarding José Andrés Cordova, he assisted me in the project in programming tasks and was a reliable partner throughout. Deadline pressure was never an issue with him, and I am grateful for that. I still remember the time we spent together in my office, working till late hours for the project.

I am also grateful to all my colleagues in our research group for their support, valuable conversations, their effort to review my papers, and the time spent together, including Dr. Andreas Abels, Fabian Billert, Dr. Kirill Bogomasov, Thomas Germer, Philipp Grawe, Dr. Gerhard Klassen, Sergej Korlakov, Nina Liebrand, Mihir Mulye, Dr. Julia Romberg, Dr. Leonie Selbach, and Boris Thome. I also appreciate the initial discussions I had with Dr. Alexander Askinadze about the topic of academic performance prediction for the research project I was involved in.

Of course, I would like to thank Guido Königstein for standing by my side when I

encountered technical issues. Without Sabine Freese and Lisa Lorenz, the secretarial tasks would have been much more challenging. These tasks are often underestimated, and I appreciate their effort for helping me in these matters.

My family has always been the reason for my motivation to work hard for my goals. The Vietnamese work ethic and resilience that my parents, Huu Phuoc Do and Le Hang Duong, instilled in me have been a driving force for me. As war refugees from Vietnam, they had to start from scratch in Germany and worked hard to provide a better future for their children. I am grateful for their sacrifices and the values they passed on to me. My sister, Bella My Phuong Quynh Duong, has always been my biggest supporter in my life. She has always been there for me when I needed someone to talk to. I am glad that I started my academic journey with her and our cousin, Fabian Vu. We all started our minor in computer science together and supported each other during that time. Now, we all have our master's degrees and are on our way to finish our doctoral studies. I also want to thank my grandmother, Thi Chanh Nguyen († 3rd September 2023), who took care of me when I was a child. Unfortunately, my promise to finish my doctoral studies before she passed away could not be fulfilled. I hope that she was still proud of who I was and what I have achieved so far.

Finally, I would like to express my gratitude to all those who I have not mentioned here but who supported my journey in one way or another. This includes my girlfriend, Mara Kogelheide, as well as my friends and other family members.

Abstract

Machine learning has become more prevalent in recent years and is used in various applications, including supporting decision-making processes. It can act as a tool to predict outcomes based on historical data. In practice, this can range from predicting academic performances to calculating credit risk scores for potential borrowers. However, applications involving personal data can have harmful consequences if the machine learning models used are biased towards certain groups of people. To prevent discriminating subpopulations, fairness is a considerable concern.

Just recently, the European Parliament adopted the Artificial Intelligence Act (AI Act) on March 13, 2024, which aims to regulate the use of AI in the European Union. One of the concerns of the AI Act is the fairness of AI systems. Discrimination that is prohibited by the European Union or national law also applies to AI systems. Since it is expected that most of the rules will come into force on August 2, 2026, and the rules for high-risk AI systems apply earlier, research focused on making machine learning models fairer and responsible has become indispensable.

In this dissertation, we explore literature gaps on fairness in machine learning and propose novel methods to fill gaps that are relevant to the AI Act. The thesis begins by presenting works that emerged from a research project, called *Responsible Academic Performance Prediction* (RAPP), which aimed to develop a responsible AI platform for predicting academic performances. These works highlight the importance of preventing discrimination in machine learning models. Recognizing that unfair predictions often stem from biased data input, we focus on this root cause and propose methods to mitigate discrimination in the data itself.

While similar methods already exist in the literature, our methods differ in that they can handle any type of discrimination, such as intersectional discrimination or discrimination towards non-binary groups. This is very pivotal, as most prior methods can only deal with binary groups and mitigate discrimination between a privileged and an unprivileged group. But in reality, populations can be categorized into more than two groups, and discrimination can occur among any of these groups. A common approach to make former methods work with more than two groups is to merge multiple groups together. However, this leads to further marginalizing already underrepresented groups and ignoring the discrimination they face. Our methods overcome this limitation and prevent such groups from being ignored. This is done by introducing a new fairness-agnostic framework, `FairDo`, that can be used with any fairness metric. The framework itself is quite flexible, allowing users to define their own fairness metrics and objectives to optimize the data. An option to handle privacy concerns is also provided by the framework. For this, synthetic data can be used to optimize the data for fairness. With

`FairDo`, certain statistical properties of the data regarding fairness can be fulfilled and a step towards satisfying the AI Act is taken. This thesis further contributes with different aspects of the introduced framework and provides a comprehensive evaluation of the methods in the respective papers.

To strive for good scientific practice, we have made our research reproducible and accessible by publishing all of our methods and experiments on GitHub. Specifically, our fairness framework, `FairDo`, is additionally available on PyPI and comes with a documentation page[1].

[1] `https://fairdo.readthedocs.io/en/latest/`

Zusammenfassung

Maschinelles Lernen hat sich in den letzten Jahren immer mehr in diversen Anwendungen durchgesetzt, unter anderem auch zur Unterstützung von Entscheidungsprozessen. Dies geschieht dadurch, indem maschinelle Lernmodelle mit historischen Daten trainiert werden, um Vorhersagen zu treffen. In der Praxis kann dies von der Vorhersage akademischer Leistungen bis hin zur Einschätzung von Kreditrisiko für potenzielle Kreditnehmer reichen. Jedoch können solche Anwendungen, die personenbezogene Daten beinhalten, schädliche Folgen haben, wenn die verwendeten maschinellen Lernmodelle bestimmte Personengruppen benachteiligen. Um solch eine Diskriminierung zu verhindern, ist es wichtig, sich mit dem Thema Fairness im maschinellen Lernen auseinanderzusetzen.

Erst kürzlich hat das Europäische Parlament am 13. März 2024 das Artificial Intelligence Act (AI Act) verabschiedet, welches den Einsatz von künstlicher Intelligenz (KI) in der Europäischen Union regeln soll. Eines der Anliegen dieser Verordnung ist die Fairness von KI-Systemen. Im Grunde gilt, dass jede Art von Diskriminierung, die durch die Europäische Union oder nationales Recht verboten ist, auch für KI-Systeme gilt. Da die meisten Vorschriften voraussichtlich am 2. August 2026 in Kraft treten werden und die Vorschriften für hochriskante KI-Systeme bereits früher gelten, ist die Forschung über Fairness im maschinellen Lernen wichtiger denn je geworden.

In dieser Dissertation untersuchen wir relevante Forschungslücken zu diesem Thema und präsentieren Methoden, die diese Lücken füllen und für AI Act relevant sind. Zu Beginn dieser Arbeit werden Fachartikel vorgestellt, die aus einem Forschungsprojekt namens *Responsible Academic Performance Prediction* (RAPP) hervorgegangen sind. Das Projekt zielte darauf ab, einen sozialverträglichen Ansatz zu entwickeln, um akademische Leistungen mithilfe von maschinellen Lernmodellen vorherzusagen. Hierfür sollte eine Plattform für potenzielle Anwender entwickelt werden. Aus den eigenen Vorarbeiten wird ersichtlich, wie wichtig es ist, Diskriminierung in derartigen Anwendungen zu verhindern. Frühe Vorarbeiten aus der Literatur haben gezeigt, dass unfaire Vorhersagen oft auf voreingenommene Daten zurückzuführen sind. Um die Ursache zu bekämpfen, wurden im Rahmen dieser Arbeit verschiedene Ansätze entwickelt, die es ermöglichen Diskriminierung in Datensätzen zu mildern bzw. zu entfernen.

Während in der Fachliteratur bereits derartige Methoden existieren, unterscheiden sich unsere Methoden dadurch, dass sie mit jeglicher Art von Diskriminierung umgehen können, z. B. mit intersektionaler Diskriminierung oder Diskriminierung gegenüber nicht-binären Gruppen. Dieses Unterscheidungsmerkmal ist von wichtiger Bedeutung und hat relevante Auswirkungen auf die Praxis. Die meisten bisherigen Methoden können nur mit binären Gruppen umgehen und verringern die Diskriminierung zwis-

chen einer privilegierten und einer unterprivilegierten Gruppe. In der Realität können Populationen jedoch in mehr als zwei soziale Gruppen eingeteilt werden, und Diskriminierung kann zwischen jeder dieser Gruppen auftreten. Ein gängiger Ansatz, um existierende Methoden auf mehr als zwei Gruppen anzuwenden, besteht darin, mehrere Gruppen miteinander zu verschmelzen. Dies führt jedoch dazu, dass bereits unterrepräsentierte Gruppen weiter marginalisiert und die Diskriminierung, der sie ausgesetzt sind, ignoriert werden. Beispielsweise geschieht dies, wenn eine stark unterrepräsentierte Gruppe mit einer anderen Gruppe zusammengeführt wird. In diesem Fall fällt die Diskriminierung der stark unterrepräsentierten Gruppe nicht mehr auf. In den vorzustellenden Verfahren in dieser Arbeit tauchen solche Problematiken nicht auf. Hauptsächlich wird dies durch die Einführung eines neuen fairness-agnostischen Frameworks, `FairDo`, erreicht. Das Framework kann mit jeglicher Fairness-Metrik verwendet werden und ermöglicht es Daten entsprechend zu optimieren. Auch benutzerdefinierte Fairness-Metriken werden unterstützt. Eine Option zum Schutz der Privatsphäre ist ebenfalls im Framework enthalten. Hierfür werden synthetische Daten verwendet. Mit diesem Framework können Datensätze so optimiert werden, dass bestimmte statistische Eigenschaften erfüllt werden. Somit kommt man der technischen Umsetzung des AI Acts ein Stück näher. Zusätzlich werden in dieser Arbeit weitere Aspekte des vorgestellten Frameworks untersucht und die Methoden in den jeweiligen Fachartikeln umfassend evaluiert.

Da gute wissenschaftliche Praxis wichtig ist, damit zukünftige Forschung auf die vorzustellenden Arbeiten in dieser Dissertation aufbauen kann, wurden alle entsprechenden Methoden und Experimente für die Reproduzierbarkeit und Zugänglichkeit auf GitHub öffentlich gemacht. Insbesondere lässt sich das Framework `FairDo` über PyPI als Python-Paket installieren. Eine Dokumentationsseite[2] ist ebenfalls verfügbar.

[2] `https://fairdo.readthedocs.io/en/latest/`

Contents

1

Introduction

1.1 Motivation

Machine learning (ML) is a subfield of *artificial intelligence* (AI) that deals with methods to enable computers to learn patterns from structured and unstructured data [43]. Differing from traditional logic-based/symbolic approaches, ML algorithms are often based on statistical models, which allow them to be more flexible and versatile in solving complex problems. ML algorithms enable data-driven predictions, which are applicable across diverse fields such as educational data mining [8], medical diagnosis [17], and argument mining [48].

However, predictions can be harmful for many reasons. For this, some companies launched *responsible AI principles* to ensure that AI systems are fair, transparent, safe, accountable, and respect privacy [11, 42]. Regarding these aspects, the *European Commission* proposed the *Artificial Intelligence Act* (AI Act) to regulate the use of AI systems in the *European Union* (EU) [24], which the *European Parliament* recently adopted on March 13, 2024. Motivated by the AI principles and the AI Act, this dissertation mainly focuses on the fairness aspect of responsible AI. An excerpt from Recital 67 of the AI Act can be taken as motivation for this work [24]:

> *"[...] The data sets should also have the appropriate statistical properties, including as regards the persons or groups of persons in relation to whom the high-risk AI system is intended to be used, with specific attention to the mitigation of possible biases in the data sets, that are likely to affect the health and safety of persons, have a negative impact on fundamental rights or lead to discrimination prohibited under Union law, especially where data outputs influence inputs for future operations (feedback loops). [...]"*

Early works on fairness in ML have shown that ML models can discriminate against certain types of people due to the inherent biases in the data [33, 14, 36]. A quite prominent example is the COMPAS system, which is used by U.S. courts to predict the likelihoods of criminals reoffending [38, 36]. Such recidivism prediction systems help judges make informed decisions about sentencing and parole [38]. In 2016, a group of investigative journalists [38] found that the COMPAS system discriminates against African Americans and overestimates their recidivism likelihoods in comparison to White Americans. These results raise questions about the fairness of the system and the potential harm it can cause to already disadvantaged groups. These issues can be found in other domains as well. Datasets from the U.S. Census Bureau, for example, have shown to exhibit biases against females and people of color regarding income levels [37, 33]. Another example is the bank telemarketing dataset from a Portuguese banking institution [44]. Machine learning models trained on this dataset have shown biases towards age and marital status in predicting whether a customer will subscribe to a term deposit [39].

Motivated by similar findings, researchers came up with methods to counteract biases in related systems [33, 30, 19]. In response, IBM® developed the `AIF360` toolkit [15] and Microsoft® created the `Fairlearn` toolkit [16], both of which include established algorithms that mitigate bias at a particular step in the ML pipeline. Mitigation algorithms are typically grouped into three categories [13, 20, 41]: *pre-processing*, *in-processing*, and *post-processing*. Pre-processing algorithms aim to remove discrimination from the data before training the ML model. Machine learning models trained on the discrimination-reduced data are expected to be fair. In-processing algorithms are generally modified ML models that achieve fairness during training. This is often done by incorporating fairness constraints into the loss function of the model. Lastly, post-processing techniques adjust the predictions of the ML model to ensure equal outcomes for all groups.

Even with the availability of these toolkits, there are still many challenges and unsolved tasks in fairness-aware ML. The challenges are as broad, complex, and multifaceted as the problems of fairness and ML themselves. Regarding fairness, there are many definitions of it, and some of them are even contradictory [28]. Depending on the assumed worldviews, different fairness definitions are more or less suitable [28]. This alone makes it difficult to develop an estimator that is perceived as fair by all stakeholders because policies affect which worldview should be assumed [1]. Even if a fairness definition is chosen, the field of ML is very broad, and fairness is not only applicable to classification tasks but also to regression, clustering, recommendation systems, and many more [13, 10, 21, 49]. Generally, `AIF360` and `Fairlearn` revolve around classification tasks, although `Fairlearn` contains a single fairness-aware regression algorithm. Even within classification tasks, fairness for *non-binary* groups and *multiple protected attributes* has not been sufficiently addressed [2, 3, 4]. The problems have been tackled in theory [35, 27, 50, 34] but a practical solution that is used widely by practitioners and is easily accessible for the research community is still missing. Another challenge is the fairness-utility trade-off [33, 30, 22]. In many cases, improving fairness leads to a decrease in utility (performance), and vice versa. If the "*WYSIWYG*" (What You See Is What You Get) worldview [28] is assumed, then a large utility decrease is not acceptable.

Driven by the urgent need to achieve fairness in high-risk AI systems due to the AI Act in the EU, this dissertation tackles multiple understudied problems in fairness-aware ML regarding classification. The works in this thesis act as examples of how to deal with discrimination and bias theoretically and, most importantly, practically. To aim for the standards of good scientific practices, all tools, frameworks, and algorithms developed in this thesis are published as open-source software on GitHub. The links to the repositories are provided in the respective chapters.

1.2 Contributions

This thesis comes with numerous practical and theoretical contributions to the field of fairness-aware machine learning. In the following, we summarize the most significant contributions and outline novel aspects of this work.

1.2.1 Fairness-Agnostic Data Optimization Framework

Our framework, called `FairDo` [3], which stands for *Fairness-Agnostic Data Optimization*, is a novel and flexible framework for fair data pre-processing. It enables the optimization of datasets to achieve statistical fairness properties. The datasets attained can be used for training machine models that will produce fairer predictions than models trained on unprocessed datasets. The framework is released as a Python package on GitHub[1] and PyPI, and comes with a documentation page[2]. Shortly after our research was published, the European Union adopted the AI Act [24], with Recital 67 implicitly encouraging the ongoing development of methods similar to ours.

Methodologically, the framework optimizes datasets for statistical fairness properties by removing or adding data points (*under*- and *oversampling*). The framework is adaptable to different fairness metrics and can therefore be used if policies change and different fairness definitions are required. To our knowledge, `FairDo` is the first pre-processing framework that is *fairness-agnostic*, surpassing the limitations of existing methods that are implemented in `AIF360` [15] and `Fairlearn` [16]. To achieve this, we stated *multiple novel combinatorial optimization problems* to describe the under- and oversampling tasks. The objective functions of these problems are based on statistical fairness metrics. By treating them as black-box functions, heuristics can be used as solvers, making the framework applicable to a wide range of fairness definitions and hence fairness-agnostic. The framework also allows the usage of exact optimization algorithms such as brute-force search, but this is not recommended due to the NP-hardness of the problems when treating the objectives as black-boxes. Efficient and exact algorithms that are tailored to specific fairness objectives can be implemented and used in the framework, too. Additionally, privacy concerns can be addressed too, as the framework supports the use of synthetic data.

1.2.2 Tackling Discrimination in Non-binary Groups

Because non-binary groups are often understudied and mitigating discrimination in such groups is not trivial, we applied our framework to this problem [3]. For this, we first have to ask the question of how discrimination can be measured in non-binary groups. By formalizing what has been discussed by Žliobaitė [53], we are able to capture discrimination when more than two groups are present in the data. One way is to report the maximum absolute statistical disparity between any two groups.

[1] `https://github.com/mkduong-ai/fairdo`

[2] `https://fairdo.readthedocs.io/`

In AIF360 [15], datasets are processed in such a way that the protected attribute is binary by merging multiple groups into one. By following our own approach, we keep the groups as they are and treat them as separate entities. With the right objectives, we were able to show that our framework can indeed lower the maximum discrimination obtainable between any comparing groups [3] in the Adult [37], Bank [44], and COMPAS [38] datasets.

1.2.3 Achieving Fairness for Multiple Protected Attributes

The fairness problem becomes more complex when having to deal with multiple protected attributes. Yang et al. [50] defined three types of groups that can be formed by multiple protected attributes: *intersectional*, *independent*, and *gerrymandering intersectional* groups. Motivated by their work and that of Kearns et. [35], we implemented an intersectional fairness metric and introduced a new metric for independent groups [4]. By using these two metrics as objectives, we were able to show that our framework can indeed reduce discrimination in datasets when multiple protected attributes are available. This further demonstrates the flexibility of our framework. The experimentation was also done on the Adult [37], Bank [44], and COMPAS [38] datasets.

We note that, when using the intersectional definition, subgroups are formed by the Cartesian product of the protected attributes. In the Bank dataset, this leads to 48 subgroups, which is a significant increase compared to algorithms that can only deal with two groups. Our framework is able to handle this increase in groups without any issues.

1.2.4 Incorporating Data Quality in FairDo

Data quality can be a major concern when using pre-processing techniques in general. The process of collecting data can be exhaustive and expensive, and removing too many data points may not be a favorable option in this case. The upsampling technique in FairDo does not suffer from this, but the undersampling approach does. Hence, we addressed this by incorporating a constraint and considering an additional objective to regard for data quality in the optimization process when removing data points [6]. The constraint was introduced in our work, and with the additional objective, a novel multi-objective optimization problem was formulated. With this, the trade-off between fairness and data quality of the Pareto front can be visualized, and the user can choose the solution according to his/her needs. This is a novel approach to the problem.

Regarding solving the multi-objective problem, we used our own modified version of NSGA-II [23] to improve the quality of the Pareto front. Our main modification consists of developing a new initialization operator that is able to generate a more diverse set of solutions. We tested our approach against the original NSGA-II method and were able to improve the *Hypervolume indicator* [52, 26] from 0.46-0.61 to 0.77-0.90 on multiple experiments [6].

1.2.5 Fair Ranking of Decision-Makers under Uncertainty

Fairness metrics must be viewed critically, as they can be misleading if the number of samples is not taken into account. Intuitively, a decision-maker who discriminates against a smaller group should not be penalized as much as a decision-maker who discriminates against a larger group. This is because statistical conclusions are less reliable when dealing with smaller sample sizes. In cases where two decision-makers exhibit no discrimination, our preferences invert. Here, we prefer the fair decision-maker, where the uncertainty of his/her decision is lower. While this is intuitive, existing measures do not capture it.

By using Bayesian statistics, we first derived an uncertainty score for group fairness metrics [7]. The uncertainty score describes the reliability of the fairness metric when the number of samples is taken into account. This was accomplished by modeling group outcomes with Beta distributions and using the variances of the distributions to form the uncertainty score. Following this, we derived a utility score that can be used to rank decision-makers according to the pre-defined preferences mentioned above. For this, a utility function based on TOPSIS [32] was constructed. In conclusion, the utility score considers both fairness and uncertainty and can be used to rank decision-makers more reliably than solely depending on disparity reports. It allows for differing decision-makers if they exhibit the same disparity towards groups but differ in the number of samples. The stated problem and the methodology are novel and have not been approached in this manner before.

1.3 Structure of the Thesis

In this doctoral thesis, a collection of published works is presented to which the author of this thesis has contributed. Each chapter presents multiple thematically-related research papers that have been published at international conferences and workshops or have been accepted for publication. The general structure of the chapters is as follows: First, the topic of the chapter is introduced. A chapter contains multiple sections, each of which presents a research paper. In each section, the research paper is summarized, and the author's contributions are highlighted.

In Chapter 2, we present practical works that were mainly conducted for the *Responsible Academic Performance Prediction*[3] (RAPP) research project. The tasks in the project guided and motivated the development of a tool for assessing fairness in machine learning models [8]. Even if the tool was developed for use cases in higher education, it is agnostic to the domain and can be applied to any dataset where discrimination might be a problem. The second work in this chapter [1] discusses the choice of fairness metrics that suit academic performance prediction best.

Taking the motivation of knowing how to measure bias in machine learning models, Chapter 3 presents works that focus on preventing bias in the first place. By introducing multiple novel pre-processing techniques and proposing new measures [2, 3, 4] to encompass wider and more general cases, the works in this chapter are able to combat biases that have not been tackled in this way before. Specifically, our methods overcome the shortcomings of existing methods and are able to address more challenging scenarios such as dealing with *non-binary groups* and *multiple protected attributes*.

So far, measuring and mitigating bias sounds promising. Chapter 4 inspects this more closely and presents problems that can arise with it. The first work in this chapter [7] gives an example where solely reporting group disparities can result in

[3] `https://rapp.hhu.de/en/`

misleading conclusions. By deriving and incorporating the uncertainty of the fairness assessment itself, discrimination can be detected more reliably. The second work [6] criticizes our own proposed pre-processing [3] technique from Chapter 3 and discusses how the resulting fair data can be made trustworthy. By introducing a constraint and an additional objective that both aim to preserve some of the original data's properties, a fair dataset that is more reliable and trustworthy can be obtained.

In Chapter 5, a conclusion is drawn and future work regarding bias mitigation and fairness assessment is outlined. Possible future work includes improving the `FairDo`[4] framework by implementing exact algorithms for some specific fairness objectives.

[4] `https://github.com/mkduong-ai/fairdo`

2

Algorithmic Fairness in Machine Learning Pipelines

This chapter presents works that are directly related and motivated by the *Responsible Academic Performance Prediction*[1] (RAPP) research project. The aim of the RAPP project was to research whether at-risk students can be identified early on and whether machine learning models can be used responsibly for such tasks.

A similar work was done by Alexander Askinadze [12] in his PhD thesis, where he developed machine learning pipelines for dropout prediction and dashboards for educational stakeholders. Since the responsibility of machine learning applications has become a major concern in recent years, especially when sensitive data is involved, such concerns were addressed in the RAPP project. In response, the project was subdivided into three work packages, each dealing with a different aspect. The first work package was tasked with the development of responsible machine learning models, to which the following papers in this chapter contribute.

The first section in this chapter presents a tool developed for the project [8]. It was used throughout the RAPP project for feature engineering, training and testing machine learning models, and assessing their performances and fairness. By offering a graphical user interface to compare different machine learning models, the tool was used to interact with other work packages by reporting the results of the developed models. The second section presents a discussion on which fairness metric suits the RAPP project best [1]. With the specified fairness metric, the models can be evaluated and compared more specifically. By combining the two works, machine learning models can be developed and assessed in a responsible manner, thereby contributing to the goals of the RAPP project.

[1] `https://rapp.hhu.de/en/`

2.1 MLOps Tool for Responsible Academic Performance Prediction

Paper: Manh Khoi Duong, Jannik Dunkelau, José Andrés Cordova, and Stefan Conrad. A Responsible Academic Performance Prediction Tool for Decision-Making in Educational Institutes. In *BTW 2023*, volume P-331 of *Lecture Notes in Informatics*. Gesellschaft für Informatik e.V., 2023.

Personal Contribution: Manh Khoi Duong initiated the research idea. Manh Khoi Duong and Jannik Dunkelau shared the literature research work. Manh Khoi Duong wrote the paper with the support of Jannik Dunkelau and José Andrés Cordova. Manh Khoi Duong and Jannik Dunkelau equally contributed to the development of the tool presented in the paper with the assistance of José Andrés Cordova. The research underwent continuous supervision by Stefan Conrad.

Status: Published

This work was motivated by the earlier stages of the interdisciplinary research project RAPP, during which the prediction task was still in discussion. Predicting academic performance is a very vague task and can include predicting exam grades, final grades, dropout risks, study durations, or any other outcome. Besides the target label, several internal discussions were held regarding the features to be used, the fairness metrics to be considered, and the machine learning models to be applied.

In order to ease the process of developing, training, and evaluating machine learning models for this ambiguous task, we developed an *MLOps* tool as a graphical user interface (GUI) that supports the entire machine learning lifecycle. To solve the ambiguity regarding the dataset (features, target label), the tool implements an SQLite query interface with syntax highlighting to allow users to query the loaded dataset. By providing this query interface, feature engineering can be done directly inside the tool with SQLite. This makes adaptations to the dataset easy and fast for the user. Regarding the responsibility part of the research project, the GUI comes with an integrated dashboard to assess the fairness and performance of the trained models. The dashboard includes fairness metrics, confusion matrices, and other evaluation metrics. By visualizing the Pareto front of fairness and performance, the tool answers the research question of which model to deploy in real applications. When decision trees are used, the tool offers an option to visualize the tree structure to provide interpretable insights into the decision-making process of the model.

For the specific datasets that were used in the RAPP project, the tool offers inbuilt SQLite templates for feature engineering. The user can select which features to use and which target label to predict. The template engine then generates a suitable SQL query for the user. In conclusion, the tool supported the RAPP project in the development of machine learning models with its rich functionalities.

RAPP: A Responsible Academic Performance Prediction Tool for Decision-Making in Educational Institutes

Manh Khoi Duong[1] Jannik Dunkelau[2] José Andrés Cordova[3] Stefan Conrad[4]

Results
Reproduced
BTW
2023

Abstract: Due to the increasing importance of educational data mining for the early intervention of at-risk students and the growth of performance data collected in educational institutes, it becomes natural to employ machine learning models to predict student's performances based off prior data. Although machine learning pipelines are often similar, developing one for a specific target prediction of academic success can become a daunting task. In this work, we present a graphical user interface which implements a customizable machine learning pipeline which allows the training and evaluation of machine learning models for different definitions of academic success, e. g., collected credits, average grade, number of passed exams, etc. The evaluation is exported in PDF format after finishing training. As this tool serves as a decision support system for socially responsible AI systems, fairness notions were included in the evaluation to detect potential discrimination in the data and prediction space.

Keywords: educational data mining; fairness; decision making; machine learning; academic performance prediction

1 Introduction

Academic performance prediction (APP) systems can be used to identify *at-risk students* in higher education early on, allowing the university to use resources in a targeted manner to prevent them from achieving poor academic performances. The definition of at-risk students varies as it depends on the context and the purpose of prevention. It can comprise of, e.g., higher chances of dropping out, longer study durations, and worse graduation grades. In this case, the APP system acts as a supporting *artificial intelligence* (AI) system for the university at the institutional level. However, given the impact of such systems onto the student body, social challenges arise. Marcinkowski et al. [Ma20] surveyed the perception of a student body of the use of such AI-based systems and show that APP is viewed as problematic by students as far as their own data and planning are concerned. Furthermore,

[1] Heinrich Heine University, Department of Computer Science, Universitätsstraße 1, 40225 Düsseldorf, Germany manh.khoi.duong@hhu.de

[2] Heinrich Heine University, Department of Computer Science, Universitätsstraße 1, 40225 Düsseldorf, Germany jannik.dunkelau@hhu.de

[3] Heinrich Heine University, Department of Computer Science, Universitätsstraße 1, 40225 Düsseldorf, Germany jose.cordova@hhu.de

[4] Heinrich Heine University, Department of Computer Science, Universitätsstraße 1, 40225 Düsseldorf, Germany stefan.conrad@hhu.de

doi:10.18420/BTW2023-29

the notion of *fairness-aware machine learning* (FairML) [DL19, Fr19, PS20] becomes an increasingly important topic and also found its way into educational data mining systems [LMZ19, KLM22, HR20, KL20, LQN21, AC19].

Acknowledging these issues, we developed a tool for *responsible academic performance prediction* (RAPP) which tackles two main tasks: it is a tool for (1) academic performance prediction and acts as a (2) decision support system for the social responsibility when employing AI in tertiary education. The first task deals with generating multiple prediction targets and datasets for the prediction of academic performances. The goal of the second task is to find socially acceptable machine learning (ML) models and justify their use from the extensive fairness and interpretability evaluation in the tool. For the full deployment of an AI system to identify at-risk students, ethical aspects and the perception by those affected have to be researched. The fairness and interpretability evaluation plays a supportive role to disregard or regard certain ML models by, e. g., checking whether they comply with student's perception of discrimination or do not discriminate through socio-demographic features.

The source code of the RAPP tool is published under the MIT License and available online at `https://github.com/hhu-rapp/rapp-tool`.

2 Related Systems

Our proposed tool combines functionalities from two different research communities: (educational) data mining and fairness assessment. In this section, we will briefly present selected tools already available from either community.

RapidMiner [HK16], Orange [De13], and WEKA [Ha09]—to name a few—are data mining tools with a graphical user interface (GUI) just as the proposed tool in this paper. The aforementioned tools mostly include data visualization, pre-processing, feature selection, clustering, classification, regression, and evaluation metrics. The tools are modular, meaning the pipeline and its specific configurations are highly modifiable. Their aim is to enable data mining practitioners the comparison of machine learning models on custom datasets without having to write code themselves.

Although not as comprehensive and powerful, tools that were explicitly developed for educational data exist as well. They predominantly focus on a specific dataset that was provided by a particular educational institute. Especially, they analyze and predict several students' data such as programming grades [Ba16], examinations of the final school year [LMP16], students' contributions in group programming [SA20], or students' written feedback [Gr20].

Fairness and transparency in machine learning have become more important in recent years due to the awareness of potential mistreatment of AI over different demographic groups [DL19, Fr19, PS20]. As a response, authors began developing tools to audit the fairness of an ML system and to produce bias reports, to guide the selection process of a fitting fairness metric, or to apply intervening methods to reduce exhibited bias. Examples for such tools are Aequitas [Sa18], FairSight [AL19], Fairlearn [Bi20], or Fairness Compass and Fairness Library [RD22]. These topics have also been recognized by the educational data mining (EDM) community lately. To name some, Hu and Rangwala [HR20] and Kizilcec and Lee [KL20] consider prejudice and unfairness where Le Quy and Ntoutsi [LQN21] and Alonso and Casalino [AC19] acknowledge the explainability of the used models in EDM.

For the proposal, the RAPP tool aims to take on the preliminary works and combine functionalities from both communities: It is a data mining tool for educational data that includes fairness examinations and interventions to address responsibility when employing AI in educational institutes.

3 RAPP Tool

Making it possible to easily create various datasets from a single database with desired features and labels to train, save, and evaluate machine learning algorithms is the aim of the developed tool. For this, the GUI provides an intuitive way to load a particular SQLite database or a CSV file[5] and specify the initial settings for the machine learning pipeline. The demanded features and target labels can be derived by querying the database. Several settings are detected automatically such as the prediction type (classification, regression), the target variable (last column by default), and categorical features. The supported estimators for classification are *decision trees, random forest, support vector machine, naive bayes, and logistic regression* and for regression *linear regression, elastic net, bayesian ridge, decision tree regressor*, and *kernel ridge*. An *artificial neural network* with two hidden layers is also available for both of these task types. Experienced users can modify the configuration for their needs. Fig. 1 displays the user interface for the settings.

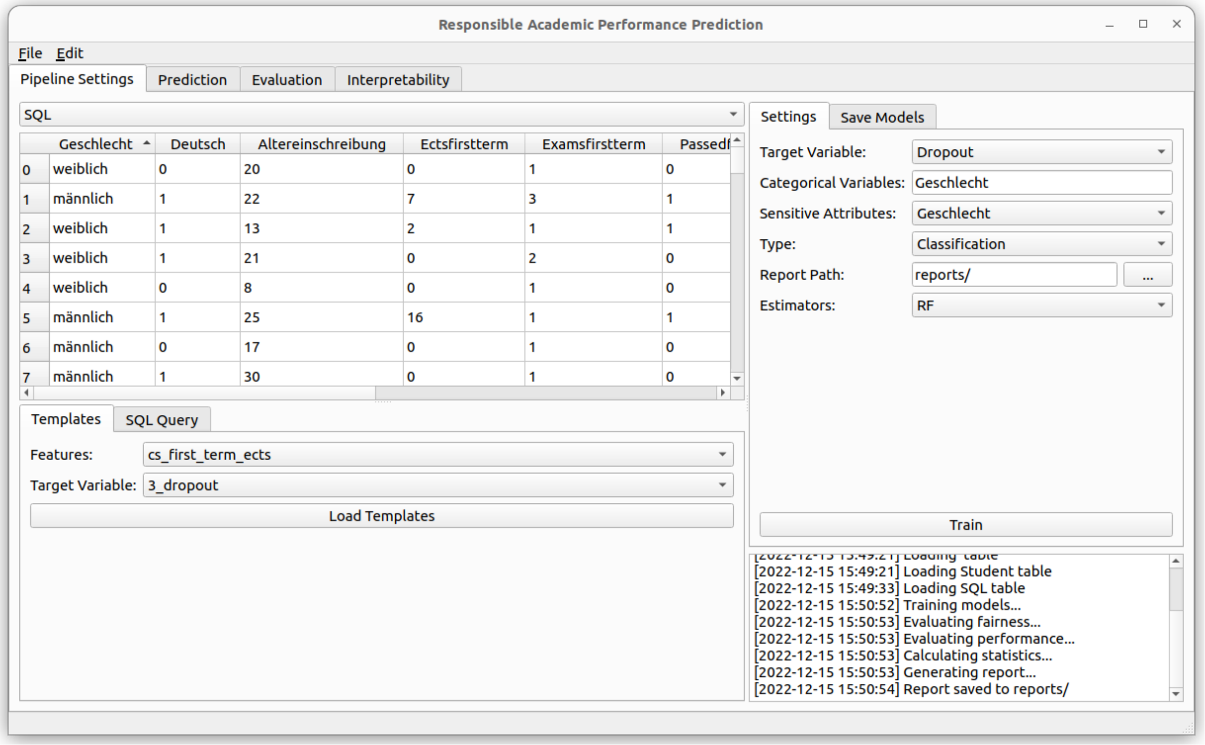
In the following, we will outline the two main uses and functionalities of the RAPP tool: APP and supporting the decision-making process whilst designing a responsible APP.

3.1 Academic Performance Prediction

3.1.1 Pipeline

At the front of the RAPP tool lies the ability to setup and train APP models over the implemented ML pipeline. The pipeline is outlined in Fig. 2. First, the pipeline's settings have to be specified. This includes the selection of a dataset to use for training as well as picking the ML algorithms to train.

[5] The CSV file is treated as a database.



	Geschlecht	Deutsch	Altereinschreibung	Ectsfirstterm	Examsfirstterm	Passedf
0	weiblich	0	20	0	1	0
1	männlich	1	22	7	3	1
2	weiblich	1	13	2	1	1
3	weiblich	1	21	0	2	0
4	weiblich	0	8	0	1	0
5	männlich	1	25	16	1	1
6	männlich	0	17	0	1	0
7	männlich	1	30	0	1	0

Features: cs_first_term_ects
Target Variable: 3_dropout

Target Variable: Dropout
Categorical Variables: Geschlecht
Sensitive Attributes: Geschlecht
Type: Classification
Report Path: reports/
Estimators: RF

```
[2022-12-15 15:49:21] Loading table
[2022-12-15 15:49:21] Loading Student table
[2022-12-15 15:49:33] Loading SQL table
[2022-12-15 15:50:52] Training models...
[2022-12-15 15:50:53] Evaluating fairness...
[2022-12-15 15:50:53] Evaluating performance...
[2022-12-15 15:50:53] Calculating statistics...
[2022-12-15 15:50:53] Generating report...
[2022-12-15 15:50:54] Report saved to reports/
```

Fig. 1: RAPP's Pipeline Settings Interface, 2022.

The data are queried over an SQLite database. While advanced users can enter custom queries on the database for feature engineering and feature selection, predefined feature and label sets were added for the given academic database to comfortably reuse and combine them in any desired pairing. The user can select, for instance, features such as credit points, grades, or number of passed exams, and target labels such as final GPA, achieved credits until semester x, or study duration. To ease working with different sets of features and labels we implemented an SQL templating engine which produces the final query based on the user's selections for a feature and a label set. This avoids combinatoric explosion which would arise if each feature-label pair's SQL query had to be implemented manually. The queried database then acts as a dataset for the machine learning pipeline.

Once the dataset is obtained, the features go through the pre-processing step of one-hot encoding any categorical features. After this, the data is split into training (80 %) and test (20 %) data.

Each of the user's selected models are trained on the training data. We also evaluate the performance over the training data via 5-fold cross-validation to capture how robust the models behave during training. The training concludes in an evaluation over various performance metrics as well as fairness metrics. Fairness is also audited directly over the dataset as well. The evaluation results are saved into a detailed PDF report file containing information over the demographics of the dataset as well as the performance and fairness results of each trained estimator.

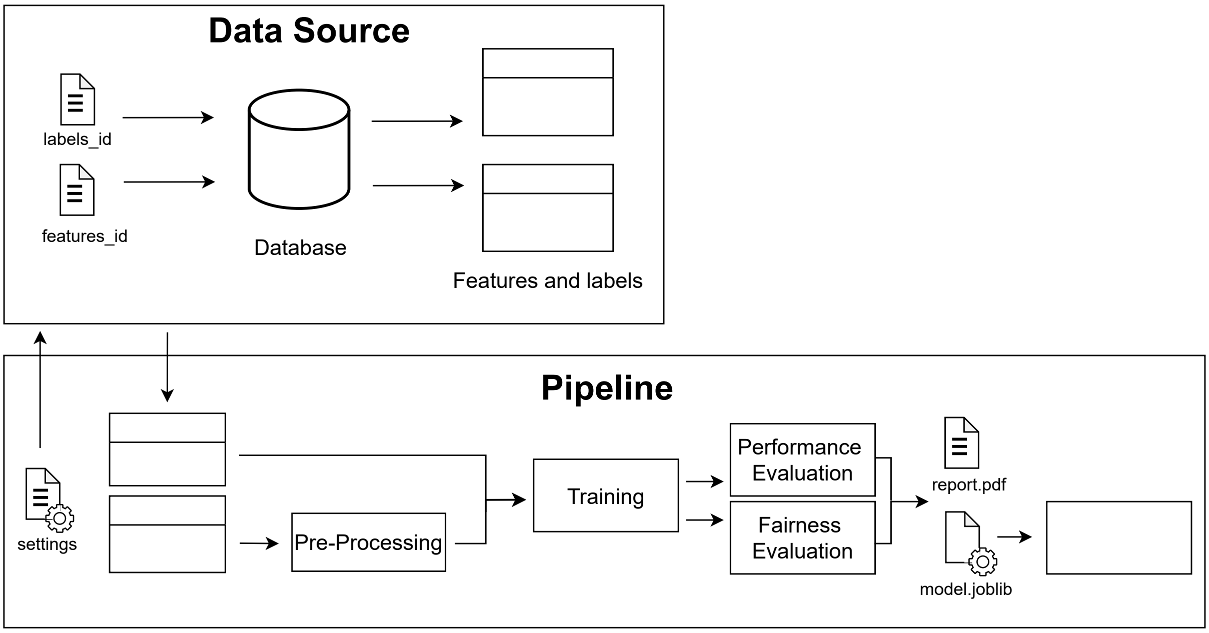


Fig. 2: RAPP's Machine Learning Pipeline, 2022.

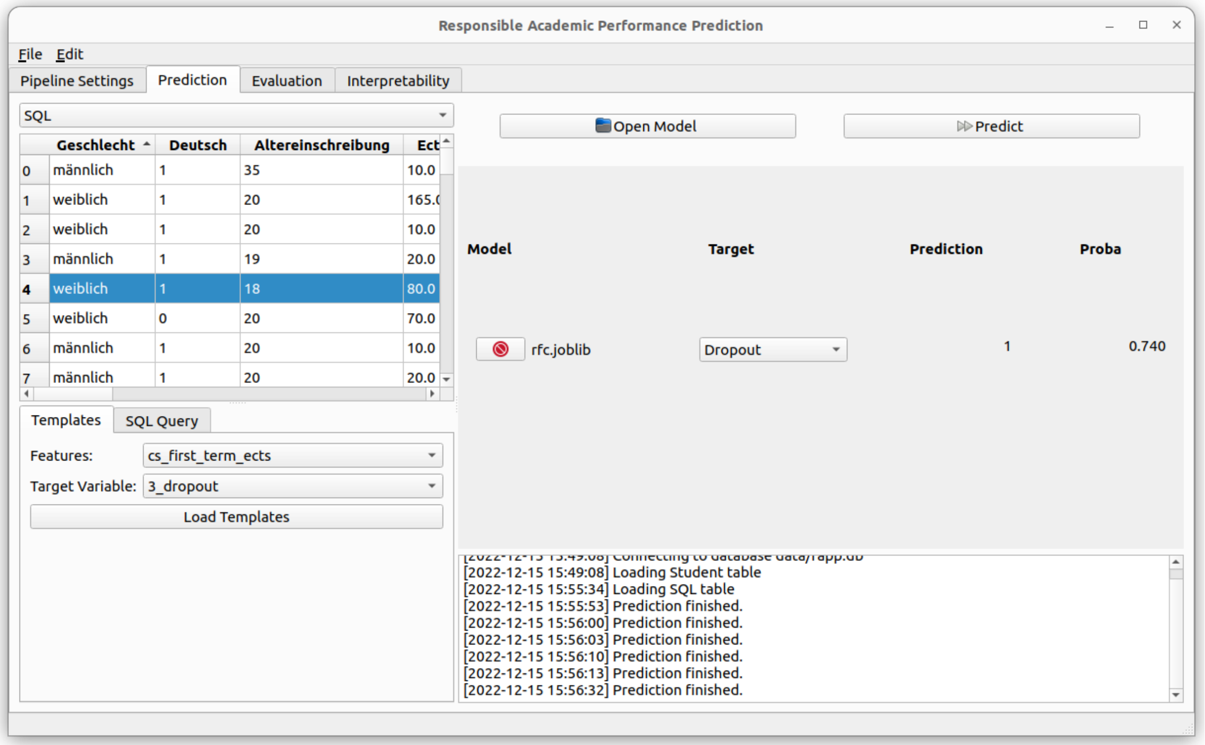
After the trained models are evaluated, the users can decide which models they want to save in order to use them later to predict on new data.

3.1.2 Prediction

To tackle the task of identifying at-risk students early, this tool includes a prediction interface as shown in Fig. 3. This interface enables the user to make predictions based on individual student's academic data. The user can then identify students who are more likely to benefit from the institution's support programs.

In order to predict the students' performances, new data from students as well as compatible models, i. e., models that have been trained with the same features, are required in the prediction interface. It is possible for the user to load various models trained for different target variables to predict several targets from the same features simultaneously. Once new data and selected models are loaded into the GUI, the features go through a pre-processing step and are then fed into the loaded models for the prediction. Fig. 3 shows an example of multiple targets being predicted with the data of a single student.

After the prediction has been run, the interface updates and displays the predictions of the models for each of the selected targets. It is also possible to load multiple models for one specific target to employ ensemble learning. In case of classification, we apply majority voting whereas in regression tasks the mean of the predicted values is used.



	Geschlecht	Deutsch	Altereinschreibung	Ect
0	männlich	1	35	10.0
1	weiblich	1	20	165.(
2	weiblich	1	20	10.0
3	männlich	1	19	20.0
4	weiblich	1	18	80.0
5	weiblich	0	20	70.0
6	männlich	1	20	10.0
7	männlich	1	20	20.0

Features: cs_first_term_ects

Target Variable: 3_dropout

Model	Target	Prediction	Proba
rfc.joblib	Dropout	1	0.740

```
[2022-12-15 15:49:08] Connecting to database data/rapp.db
[2022-12-15 15:49:08] Loading Student table
[2022-12-15 15:55:34] Loading SQL table
[2022-12-15 15:55:53] Prediction finished.
[2022-12-15 15:56:00] Prediction finished.
[2022-12-15 15:56:03] Prediction finished.
[2022-12-15 15:56:10] Prediction finished.
[2022-12-15 15:56:13] Prediction finished.
[2022-12-15 15:56:32] Prediction finished.
```

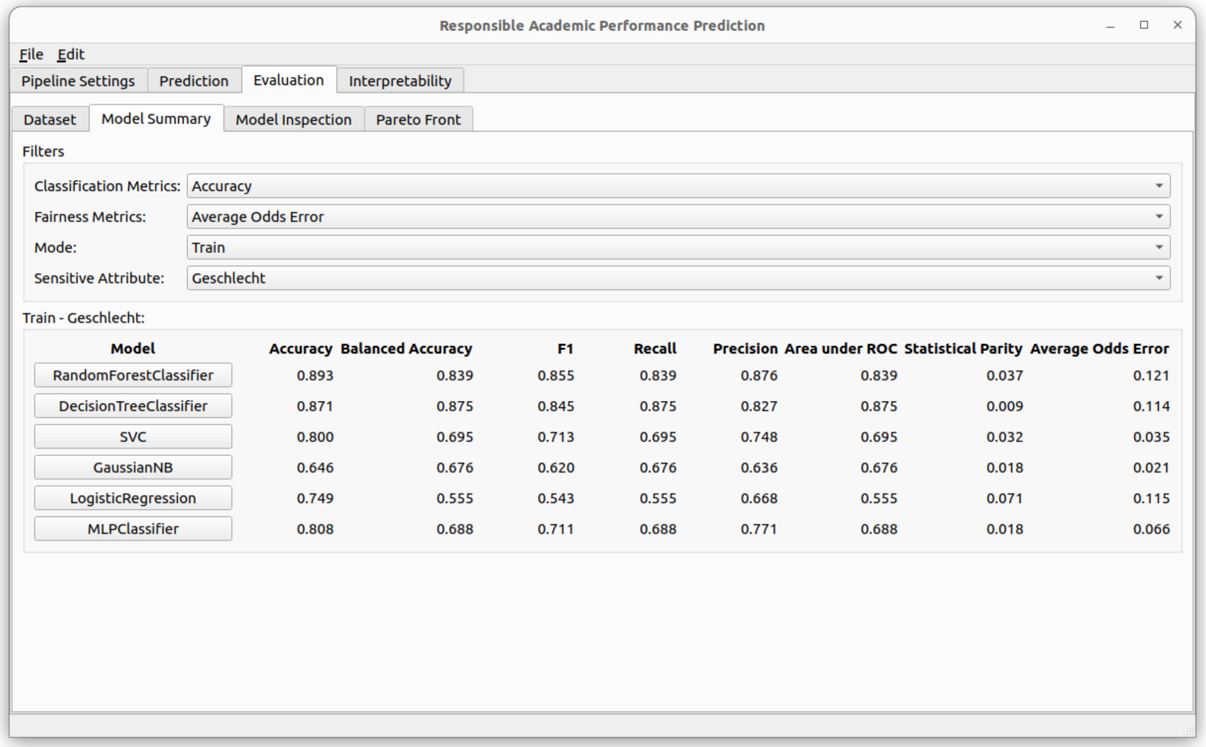
Fig. 3: RAPP’s Prediction Interface, 2022.

3.2 Decision Support System

The tool acts as a decision support system by providing the user statistical insights of the dataset as well as an extensive evaluation of the models’ performance and fairness. The models are automatically evaluated on the training and test data as they progress through the pipeline. The evaluation is displayed in the GUI, part of it is shown in Fig. 4, and is also generated as a LaTeX report, that is automatically compiled as a PDF file.

Dataset. The dataset tab contains a contingency table that displays the label $y \in \{0, 1\}$ and the sensitive attribute. This allows the user to comprehend the relationship between the sensitive attributes and the students’ performances.

Performance Metrics. As for stability reasons, the evaluation for the training data is always done with 5-fold cross-validation. The type of task that was selected beforehand determines the suitable metrics. Classification metrics included in the tool are *accuracy*, *balanced accuracy*, F_1, *recall*, *precision*, and *area under ROC*. As for regression metrics, the tool implements *mean absolute error*, *mean squared error*, *max error*, and R^2.



Classification Metrics: Accuracy
Fairness Metrics: Average Odds Error
Mode: Train
Sensitive Attribute: Geschlecht

Train - Geschlecht:

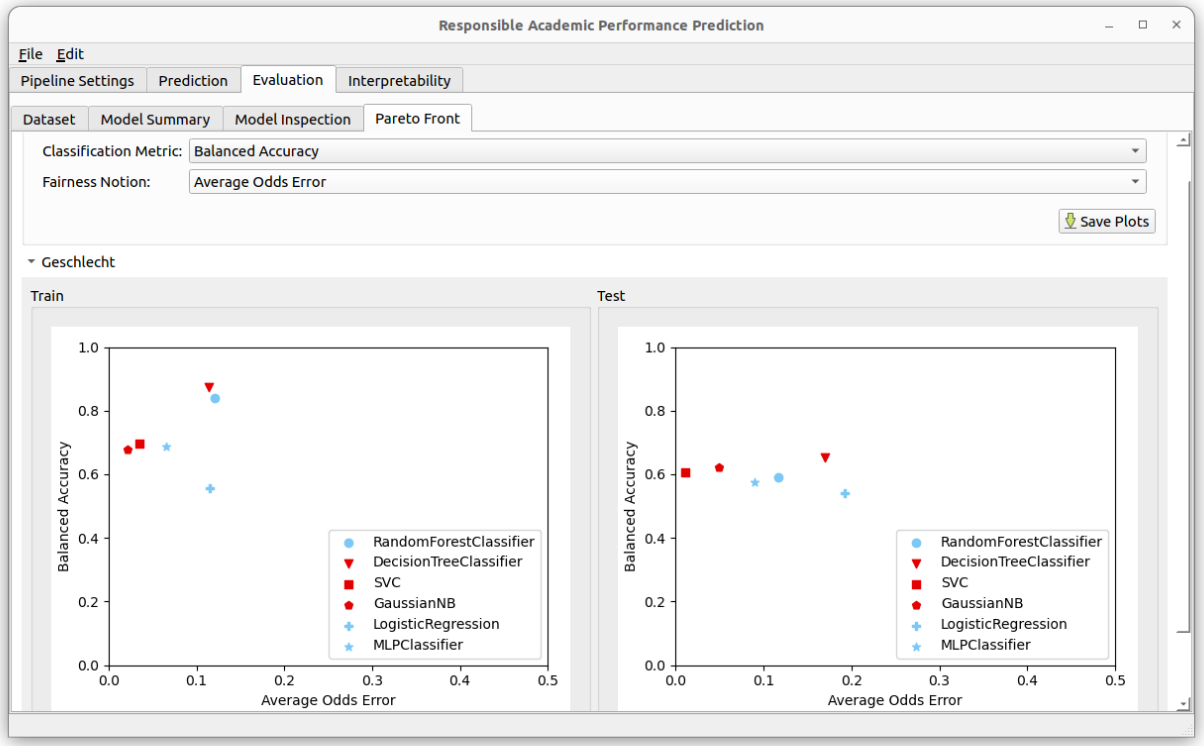
Model	Accuracy	Balanced Accuracy	F1	Recall	Precision	Area under ROC	Statistical Parity	Average Odds Error
RandomForestClassifier	0.893	0.839	0.855	0.839	0.876	0.839	0.037	0.121
DecisionTreeClassifier	0.871	0.875	0.845	0.875	0.827	0.875	0.009	0.114
SVC	0.800	0.695	0.713	0.695	0.748	0.695	0.032	0.035
GaussianNB	0.646	0.676	0.620	0.676	0.636	0.676	0.018	0.021
LogisticRegression	0.749	0.555	0.543	0.555	0.668	0.555	0.071	0.115
MLPClassifier	0.808	0.688	0.711	0.688	0.771	0.688	0.018	0.066

Fig. 4: RAPP’s Decision Support System Interface, 2022.

Fairness Notions. The fairness of the models’ predictions is assessed with regard to the sensitive attributes in order to detect potential discrimination. Similarly to the performance metrics, the notions are determined by the task type. Classification tasks implement *statistical parity*, *predictive equality*, and *equality of opportunity* [DL19, BHN19]. While statistical parity is one of the most commonly used fairness notions, recent work suggests a focus on *equalized odds* (requires predictive equality as well as equality of opportunity) as the go-to notion for APP systems [DD22]. Accordingly, the tool integrates *average odds error* [Be18] which quantifies equalized odds. For regression tasks we use the *individual fairness* and *group fairness* notion as introduced by Berk et al. [Be17].

To measure fairness criteria in classification, we use the absolute difference of the outcomes between two groups. Generally, a lower value describes less discrimination. Because group sizes greater than two (non-binary genders, multiple nationalities) might occur in the dataset, we use the maximum value of the absolute differences between all group pairs [Ž17]. This measures the maximal discrimination a classifier has achieved between two groups.

Pareto Front: Performance and Fairness Trade-Off. Due to the existence of a performance and fairness trade-off [BFT12], the trade-off can be visually examined in order to select the best trained models to use for predictions. The Pareto-efficient models, i. e., models that optimize both a particular performance metric and fairness measure, can then be identified. The fairness tab includes scatter points of the selected models in a



Classification Metric: Balanced Accuracy

Fairness Notion: Average Odds Error

Fig. 5: RAPP’s Pareto Front Evaluation, 2022.

performance-fairness plot (see Fig. 5). The Pareto front, i. e., the set of all Pareto-efficient models [JS08], is shown in a different color to differentiate them from Pareto-dominated points. Pareto-efficient models are displayed in red whereas Pareto-dominated models are displayed in lightblue. This visualization limits the decision-making space for the user as only Pareto-efficient models are of interest. Because Pareto optimal solutions are first shown and the decision-maker selects her/his preferred model afterwards, this is a posteriori method in decision-making.

In classification we aim to maximize the performance metric whereas a maximization of the performance in regression corresponds to minimizing the error. For contextual conveniences, we maximize the negative error in regression to yield for the same plot.

4 Case Study

The RAPP tool is developed as part of a research project concerning itself with designing a socially responsible framework on how to approach APP in higher education. For this, we conducted a case study over data given to us by the Heinrich Heine University Düsseldorf. The case study was concerned mostly with probing of which prediction tasks show most-promising performances and to estimate possible algorithmic fairness problems. Hereby, the prediction tasks differed in their combination of input features as well as at-risk definition for prediction.

	Prediction		
Input	Dropout	MA Adm.	SDS
ECTP + Exam stats	0.65	0.63	0.67
Grades + Exam stats	0.68	0.67	0.61
Specific modules	0.74	0.62	0.63

Tab. 1: Overview of exemplary training results over CS students in their first semester. Displaying the best performing balanced accuracy achieved by any trained model over combinations of selected feature sets and the prediction of student dropouts, master program admission (MA Adm.), and finishing in standard duration of study (SDS).

As we were interested in any combination of these predefined features and prediction goals, the RAPP tool was a great help in leveraging the combinatorial explosion problem into a manageable set of selectable templates, allowing us to quickly train and store models for each combination. Fig. 4 displays one such training result as reported within the tool, allowing comparison of the trained models over various performance and fairness measures. Tab. 1 shows exemplary results conducted with the RAPP tool over computer science (CS) students after their first semester.

5 Limitations and Future Work

Since the tool is still in development, new opportunities for future improvements present themselves constantly. These enhancements include changes to the tool's architecture, as well as making the prediction process more transparent to the user.

The tool as it currently is comes with SQL templates designed for our database in use. However, in order to allow other educational institutions to target at-risk students, the tool allows to write a different set of SQL templates and to load any SQLite database, making the tool essentially database agnostic. Still, this requires proficiency with writing SQL queries and modulating them into the templating engine, a skill that end users might not have. Here, the ease of use could be enhanced.

While allowing to inspect potentially exhibited discrimination by the trained models, it is not yet possible to train models with fairness-interventions in mind. In the future we would like to incorporate ways to train models with fairness-accounting measures such as pre-, in-, and postprocessing [DL19, Fr19, PS20]

6 Discussion

The tool helps in investigating which fairness constraints are met by any trained model and thus guides the user in their decision making of which model to employ, but by no means does the tool alone help achieving the overall goal.

Approaching RAPP includes to find a suitable definition for algorithmic fairness by involving both, the institute's stakeholders as well as the affected student body [KLM22]. While the notion of equalized odds seems to be a desirable fairness constraint [DD22], the student body appears to favor demographic parity [Ma20]. Further, the potential damage caused by misclassifications needs to be carefully considered. All these above points are not meant to be resolved by the RAPP tool but rather need to be part of the conceptualization when planning to employ such a system *before* actual employment of the system takes place. However, the RAPP tool helps to investigate whether potential concerns are dealt with appropriately by the trained models or not.

7 Conclusion

In this paper, we presented the RAPP tool for developing responsible academic performance prediction systems. The tool tackles two main tasks: designing, training, and analyzing different APP tasks, and acting as a decision support system for selecting the best suited models in a fairness-sensitive and socially responsible context.

For the setup of APP tasks, the concurrent design and direct comparison of different tasks, i. e., different input features and target labels, was a main objective as the definition of academic performances differ depending on the viewpoints of users, the student body, or the application context. In order to assist the user which model or models are socially responsible when being employed to target interventions at at-risk students, extensive performance and fairness metrics are included. The metrics are viewable in the GUI itself but are also automatically exported to a PDF file. Assessing fairness metrics and highlighting the Pareto front of classical performance metrics and achieved fairness parities guides the user in the decision-making process of finding the most suitable model for their desired task.

Overall, the tool provides an interface to non–machine learning engineers to train, evaluate, and employ models in the APP domain by providing a simplified ML pipeline configuration and highlighting crucial trade-offs of the model accuracy vs. fairness, rendering responsible APP systems a step more accessible and approachable to everyone.

Acknowledgments

This work was supported by the Federal Ministry of Education and Research (BMBF) under Grand No. 16DHB4020.

Bibliography

[AC19] Alonso, José M; Casalino, Gabriella: Explainable artificial intelligence for human-centric data analysis in virtual learning environments. In: International workshop on higher education learning methodologies and technologies online. Springer, pp. 125–138, 2019.

[AL19] Ahn, Yongsu; Lin, Yu-Ru: FairSight: Visual analytics for fairness in decision making. IEEE transactions on visualization and computer graphics, 26(1):1086–1095, 2019.

[Ba16] Badr, Ghada; Algobail, Afnan; Almutairi, Hanadi; Almutery, Manal: Predicting students' performance in university courses: a case study and tool in KSU mathematics department. Procedia Computer Science, 82:80–89, 2016.

[Be17] Berk, Richard; Heidari, Hoda; Jabbari, Shahin; Joseph, Matthew; Kearns, Michael; Morgenstern, Jamie; Neel, Seth; Roth, Aaron: A convex framework for fair regression. arXiv preprint arXiv:1706.02409, 2017.

[Be18] Bellamy, Rachel K. E.; Dey, Kuntal; Hind, Michael; Hoffman, Samuel C.; Houde, Stephanie; Kannan, Kalapriya; Lohia, Pranay; Martino, Jacquelyn; Mehta, Sameep; Mojsilovic, Aleksandra; Nagar, Seema; Ramamurthy, Karthikeyan Natesan; Richards, John T.; Saha, Diptikalyan; Sattigeri, Prasanna; Singh, Moninder; Varshney, Kush R.; Zhang, Yunfeng: AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias. CoRR, abs/1810.01943, 2018.

[BFT12] Bertsimas, Dimitris; Farias, Vivek F; Trichakis, Nikolaos: On the efficiency-fairness trade-off. Management Science, 58(12):2234–2250, 2012.

[BHN19] Barocas, Solon; Hardt, Moritz; Narayanan, Arvind: Fairness and Machine Learning. fairmlbook.org, 2019. `http://www.fairmlbook.org`.

[Bi20] Bird, Sarah; Dudík, Miro; Edgar, Richard; Horn, Brandon; Lutz, Roman; Milan, Vanessa; Sameki, Mehrnoosh; Wallach, Hanna; Walker, Kathleen: Fairlearn: A toolkit for assessing and improving fairness in AI. Microsoft, Tech. Rep. MSR-TR-2020-32, 2020.

[DD22] Dunkelau, Jannik; Duong, Manh Khoi: Towards Equalised Odds as Fairness Metric in Academic Performance Prediction. In: 2nd Workshop on Fairness, Accountability, and Transparency in Educational Data. July 2022.

[De13] Demšar, Janez; Curk, Tomaž; Erjavec, Aleš; Črt Gorup; Hočevar, Tomaž; Milutinovič, Mitar; Možina, Martin; Polajnar, Matija; Toplak, Marko; Starič, Anže; Štajdohar, Miha; Umek, Lan; Žagar, Lan; Žbontar, Jure; Žitnik, Marinka; Zupan, Blaž: Orange: Data Mining Toolbox in Python. Journal of Machine Learning Research, 14:2349–2353, 2013.

[DL19] Dunkelau, Jannik; Leuschel, Michael: Fairness-Aware Machine Learning: An Extensive Overview. Working paper, available at `https://www3.hhu.de/stups/downloads/pdf/fairness-survey.pdf`, October 2019.

[Fr19] Friedler, Sorelle A.; Scheidegger, Carlos; Venkatasubramanian, Suresh; Choudhary, Sonam; Hamilton, Evan P.; Roth, Derek: A comparative study of fairness-enhancing interventions in machine learning. In: Proceedings of the Conference on Fairness, Accountability, and Transparency. ACM, jan 2019.

[Gr20] Grönberg, Niku; Knutas, Antti; Hynninen, Timo; Hujala, Maija: An online tool for analyzing written student feedback. In: Koli Calling'20: Proceedings of the 20th Koli Calling International Conference on Computing Education Research. pp. 1–2, 2020.

[Ha09] Hall, Mark; Frank, Eibe; Holmes, Geoffrey; Pfahringer, Bernhard; Reutemann, Peter; Witten, Ian H: The WEKA data mining software: an update. ACM SIGKDD explorations newsletter, 11(1):10–18, 2009.

[HK16] Hofmann, Markus; Klinkenberg, Ralf: RapidMiner: Data mining use cases and business analytics applications. CRC Press, 2016.

[HR20] Hu, Qian; Rangwala, Huzefa: Towards Fair Educational Data Mining: A Case Study on Detecting At-Risk Students. International Educational Data Mining Society, 2020.

[JS08] Jin, Yaochu; Sendhoff, Bernhard: Pareto-Based Multiobjective Machine Learning: An Overview and Case Studies. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 38(3):397–415, 2008.

[KL20] Kizilcec, René F.; Lee, Hansol: Algorithmic Fairness in Education. arXiv, 2020.

[KLM22] Keller, Birte; Lünich, Marco; Marcinkowski, Frank: How Is Socially Responsible Academic Performance Prediction Possible? In: Strategy, Policy, Practice, and Governance for AI in Higher Education Institutions, pp. 126–155. IGI Global, may 2022.

[LMP16] Livieris, Ioannis; Mikropoulos, Tassos; Pintelas, Panagiotis: A decision support system for predicting students' performance. Themes in Science and Technology Education, 9(1):43–57, 2016.

[LMZ19] Loukina, Anastassia; Madnani, Nitin; Zechner, Klaus: The many dimensions of algorithmic fairness in educational applications. In: Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications. Association for Computational Linguistics, pp. 1–10, August 2019.

[LQN21] Le Quy, Tai; Ntoutsi, Eirini: Towards fair, explainable and actionable clustering for learning analytics. In: EDM. 2021.

[Ma20] Marcinkowski, Frank; Kieslich, Kimon; Starke, Christopher; Lünich, Marco: Implications of AI (un-)fairness in higher education admissions. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. ACM, jan 2020.

[PS20] Pessach, Dana; Shmueli, Erez: Algorithmic Fairness. volume abs/2001.09784, 2020.

[RD22] Ruf, Boris; Detyniecki, Marcin: A Tool Bundle for AI Fairness in Practice. In: CHI Conference on Human Factors in Computing Systems Extended Abstracts. pp. 1–3, 2022.

[Sa18] Saleiro, Pedro; Kuester, Benedict; Stevens, Abby; Anisfeld, Ari; Hinkson, Loren; London, Jesse; Ghani, Rayid: Aequitas: A Bias and Fairness Audit Toolkit. arXiv preprint arXiv:1811.05577, 2018.

[SA20] Sandee, Jan Jaap; Aivaloglou, Efthimia: Gitcanary: A tool for analyzing student contributions in group programming assignments. In: Koli Calling'20: Proceedings of the 20th Koli Calling International Conference on Computing Education Research. pp. 1–2, 2020.

[Ž17] Žliobaitė, Indrė: Measuring discrimination in algorithmic decision making. Data Mining and Knowledge Discovery, 31:1060–1089, 2017.

2.2 Suitable Fairness Metric for Academic Performance Prediction

Paper: Jannik Dunkelau and Manh Khoi Duong. Towards Equalised Odds as Fairness Metric in Academic Performance Prediction. In *FATED 2022: Fairness, Accountability, and Transparency in Educational Data (Educational Data Mining Workshop)*, 2022.

Personal Contribution: Jannik Dunkelau initiated the research idea. Jannik Dunkelau and Manh Khoi Duong shared the literature research work. Jannik Dunkelau wrote the paper with the support of Manh Khoi Duong.

Status: Published

In the previous work [8], we developed a tool for training machine learning models for the deployment in *academic performance prediction* (APP). Even with the provided evaluation capabilities of the tool, the question of which fairness metric to consider for the APP task remains open.

In this work, the suitability of the equalized odds [30] fairness metric for APP is analyzed. For this, the paper first discusses two contrasting *worldviews*, *We're all equal* (WAE) and *What you see is what you get* (WYSIWYG), introduced by Friedler et al. [28]. With the literature review, our paper draws the conclusion that the WYSIWYG worldview fits the APP task better than the WAE worldview. We further discuss fairness concerns known in the literature and compare fairness metrics with respect to these concerns.

The paper concludes with the recommendation to use the equalized odds fairness metric for the APP task, as it does not amplify discrimination when the WYSIWYG worldview is assumed. Long-term impacts are thereby minimized. Still, equalized odds comes with limitations as with any other fairness metric, and caution is advised when interpreting the results. A further discussion and analysis of the fairness metric in the context of APP regarding equity, individual fairness, and perceived fairness is recommended for future work.

Towards Equalised Odds as Fairness Metric in Academic Performance Prediction

Jannik Dunkelau
Heinrich-Heine-Universität Düsseldorf
D-40225 Düsseldorf, Germany
jannik.dunkelau@hhu.de

Manh Khoi Duong
Heinrich-Heine-Universität Düsseldorf
D-40225 Düsseldorf, Germany
manh.khoi.duong@hhu.de

ABSTRACT

The literature for fairness-aware machine learning knows a plethora of different fairness notions. It is however well-known, that it is impossible to satisfy all of them, as certain notions contradict each other. In this paper, we take a closer look at academic performance prediction (APP) systems and try to distil which fairness notions suit this task most. For this, we scan recent literature proposing guidelines as to which fairness notion to use and apply these guidelines onto APP. Our findings suggest equalised odds as most suitable notion for APP, based on APP's WYSIWYG worldview as well as potential long-term improvements for the population.

Keywords

Worldviews, Fairness Notion, Equalised Odds, Responsible Academic Performance Prediction

1. INTRODUCTION

Socially responsible and fairness-aware machine learning (FairML) is becoming increasingly more important to our society and aggregated a large body of research regarding how to ensure fairness and non-discrimination by artificially intelligent system [9, 13, 26, 28, 34]. As a consequence, the notion of FairML found its way into the research of educational recommender systems as well wherever a social impact onto the student body is at stake [17, 18, 21]. A major part in this plays academic performance prediction (APP). Hereby, an APP system takes data of a student as input, predicts how the student will perform in the future, and may hence induce an action based on this prediction which may itself affect the student [2]. Such predictions are usually employed as early-warning system to determine students at risk, intervene by supplying necessary help and resources, and increase graduation rates as a consequence [2, 3, 7, 17]. Although other utilisation of APP is possible, e.g. guiding university admissions, we will focus on the use case of targeted interventions. Given the need for socially responsible APP systems [17, 18], the question arises as to which notion of fairness is suitable for APP.

In the following, we review literature regarding selection of fairness notions, derive a reduced guideline to decide between two popular, parity-based fairness notions, demographic parity and equalised odds, and apply our findings onto APP. Our results and main contributions are the relationship of APPs to equalised odds and the WYSIWYG worldview which is backed by literature. Motivated by own work regarding the conceptualisation of responsible APP, we hope to narrow down the research focus for APP fairness notions, provide a base-notion for new and established APP researchers alike, and to contribute to public discourse on this matter.

2. NOTATION

In the following, let $\mathcal{D} = \{(x_i, y_i, z_i)\}_{i=1}^{n} \subset \mathbf{X} \times \mathbf{Y} \times \mathbf{Z}$, denote the training data of individuals where $\mathbf{X} \subset \mathbb{R}^d$ is the d-dimensional set of input (feature) vectors characterising each individual, $\mathbf{Y}$ denotes the set of measured true labels over the individuals, and $\mathbf{Z}$ is the set of protected attributes corresponding to each individual. Given a classifier h, we denote the set of its predictions over $\mathbf{X}$ as $\hat{\mathbf{Y}}$. Without loss of generality, we assume $y \in \mathbf{Y}$ to be binary in $\{0, 1\}$. We say an individual $(x, y, z) \in \mathcal{D}$ receives the *favourable outcome* if the prediction $\hat{y} = h(x) = 1$. Otherwise, we say the individual receives the *unfavourable outcome*. We say the individual belongs to the demographic group z. Further, let $X, Y, Z, \hat{Y}$ denote random variables describing the events that, for an individual from the training data, their features, ground truth, protected attributes, and classifier prediction take a specific value, respectively. Thus, $P(\hat{Y} = 1 \mid Y = 1)$ denotes the probability that individuals with a positive ground truth are receiving the favourable outcome.

3. PARITY-BASED FAIRNESS NOTIONS

Parity-based fairness notions are defined over the values of a classifier's confusion matrix [6]. They assume fairness once a set of predictive rates is equal for each demographic group, for instance the positive prediction rate, true positive prediction rate, or false positive prediction rate, as we will see below. For this work, we focus on two such notions which are currently prevalent in literature: demographic parity and equalised odds. We selected these notions as they seem to have higher citation counts as others [31] and are accounted for by related literature as well [6, 18, 28].

Demographic parity assumes that the distribution of the favourable outcome should be equal throughout all demographic groups. It is formally defined in Definition 1:

Definition 1 (Demographic Parity). *We say that a classifier satisfies* demographic parity *if the positive prediction rate is equal for all demographic groups, i.e.*

$$P(\hat{Y} = 1 \mid Z = z) = P(\hat{Y} = 1)\,. \tag{1}$$

While demographic parity is the most popular fairness metric in literature, it also exhibits various short comings. For instance, randomizing predictions for one demographic group while having proper predictions for another can already satisfy the notion [10]. It is however independent from any possible bias in the collection of the ground truth values $\mathbf{Y}$ which could have been assembled in a discriminatory way [4] as the notion does not rely on $\mathbf{Y}$ at all.

As an alternative, Hardt et al. [15] proposed the notion of equalised odds as given in Definition 2, which assumes fairness if $\hat{Y} \perp Z \mid Y$. As equalised odds is defined over true and false positive rates of a classifier, it is always satisfied if $\hat{\mathbf{Y}} = \mathbf{Y}$ which is not guaranteed for demographic parity.

Definition 2 (Equalised Odds). *We say that a classifier satisfies* equalised odds *if it has equal true positive rates and false positive rates for all demographic groups, i.e.*

$$P(\hat{Y} = 1 \mid Y = 1, Z = z) = P(\hat{Y} = 1 \mid Y = 1) \tag{2}$$

$$P(\hat{Y} = 1 \mid Y = 0, Z = z) = P(\hat{Y} = 1 \mid Y = 0) \tag{3}$$

4. WORLDVIEWS

Recent literature promotes accounting for the worldview that underlies the data [12, 18, 22, 32]. Worldviews were introduced by Friedler et al. [12]. To define them, we must firstly differentiate between the *observable space* $\mathcal{Y}$ and the *construct space* $\mathcal{Y}'$. The observable space $\mathcal{Y}$ represents the room of available observations and measurements. The training data $\mathcal{D}$ can only be collected from the observable space. On the other hand, the construct space $\mathcal{Y}'$ represents the theoretical space of the "true" data that is relevant to the task but not measurable. For instance, assume the task to predict whether a student will graduate within the standard duration of study. We can collect historical information of graduates to model the target variable Y and select characteristics such as grades within the first semester or number of passed courses per semester as features. These are part of the observable space that is available to us. The related construct space would contain information about how well passed courses were understood, how motivated the students will remain long-term, or how much time they will be able to invest into their studies in later semesters. This information is not accessible to us but can only be observed via assumed proxies. Further more, the construct space is free from discrimination in a sense that it would not contain the grades a student received but rather the grade a student should have received if no discrimination took place.

Worldviews model the expected differences between demographic groups in $\mathcal{Y}'$ and hence explain the presence of measured differences in $\mathcal{Y}$ [32]. Two prominent worldviews are *We're all equal* and *What you see is what you get*, for which we borrow Definitions 3 and 4 of Yeom and Tschantz [32]. Both where originally formulated by Friedler et al. [12] and seem to represent two polar ends in the fairness literature.

Definition 3 (WAE). We're all equal (WAE) *represents a world view which assumes that each demographic group is identical to each other with respect to the target variable in the construct space, i.e.* $\mathcal{Y}' \perp Z$.

Definition 4 (WYSIWYG). What you see is what you get (WYSIWYG) *is a worldview which assumes that differences in* $\mathcal{Y}$ *are explained by differences in* $\mathcal{Y}'$ *and hence that the observable space is an accurate reflection of the construct, i.e.* $\mathcal{Y} = \mathcal{Y}'$.

As we consider WAE and WYSIWYG in contexts in which we do observe discrimination in the observable space $\mathcal{Y}$ and thus assume $\mathcal{Y} \not\perp Z$, both world views contradict each other in context of this work.

5. FAIRNESS SELECTION GUIDELINES

While literature produced a great number of fairness notions to choose from, we know different fairness notions to be mutually exclusive to one another, making it impossible to satisfy all notions simultaneously [5, 12, 19, 29]. Specifically, the notions from Section 3 above are mutually exclusive in non-trivial cases. Hence, it is valuable to know which fairness metric suits the prediction task most.

Makhlouf et al. [22] collected a decision diagram guiding the fairness notion selection process. This diagram leads to the selection of demographic parity if standards exist which regulate the ratio of admission rates for the favourable outcome or we do not have a reliable ground truth or can assume historical bias or measurement bias in the data. Further, when we have a reliable ground truth or assume no historical or measurement biases in our data, the authors advocate for equalised odds if the emphasis is on the classification recall. Makhlouf et al. further advance the idea that the selection of fairness notion must be based on the explicit choice of an underlying worldview. The worldview itself is however not (explicitly) part of their guiding diagram. If we however focus on the distinction between reliability of $\mathbf{Y}$ (and existence or absence of biases), we can infer that a reliable $\mathbf{Y}$ relates to $\mathcal{Y} \approx \mathcal{Y}'$ and thus WYSIWYG, and an unreliable $\mathbf{Y}$ relates hence to WAE.

Friedler et al. [12] show in their initial conception of worldviews that individual fairness can be guaranteed under WYSIWYG while it can cause discrimination in a WAE setting. On the flip side, demographic parity is not applicable in a WYSIWYG setting while it can guarantee non-discrimination for WAE. Yeom and Tschantz [32] investigated the theoretical impact the selection of a fairness notion has on the disparity between groups. In their work, they prove that any model that satisfies demographic parity on $\hat{\mathbf{Y}}$ does not amplify existing disparity in $\mathcal{Y}'$. However, only WAE lends itself to demographic parity, as the classification performance with respect to $\mathcal{Y}'$ in WYSIWYG will always be suboptimal. A model satisfying equalised odds will not amplify any disparity in WYSIWYG but can amplify disparities if WAE holds.

Unifying the guidelines and insights from above, demographic parity should be employed when WAE holds. That is, we desire an equal distribution of the favourable outcome throughout the groups as we accredit any measurable differences in our training data to prior discrimination (historical or elsewise). Equalised odds should be favoured if WYSIWYG holds. That is, we expect differences between groups to be explainable by differences in the construct space $\mathcal{Y}'$. Some literature also promotes demographic parity when the *playing field is even* for the groups [8, 18] or the classifier is employed for one group independently [21] while promoting equalised odds otherwise [18, 21].

6. TOWARDS AN APP FAIRNESS NOTION

In this section, we will discuss the worldview that generally seems to tie to APP systems, derive equalised odds as the appropriate fairness notion, then take a closer look at the benefits and drawbacks equalised odds exhibits. We conclude with a brief overview of selected notions which we did not consider in depth.

6.1 The APP Worldview

To evaluate which worldview gives itself to APP systems, we investigate below which input features and target variables promote which worldview to conclude the related fairness notion. For this, we lean on the work of Alyahyan and Düştegör [2], who report the mostly used influential features for APP to be prior academic achievement and student demographics, accounting for 70 % of their surveyed articles.

Prior academic achievement is mostly concerned with grade-related features which are aggregated during university [2]: specific course grades, grade point average (GPA), cumulative GPA, exam results, or individual assessment grades; but also pre-university features such as high school background or study admission test results.

Taking grade-related features into account to predict on graduation level, it feels intuitive that we are in a WYSIWYG environment. Not because the grading of students can be assumed to be unbiased (which it cannot, cf. [23, 24]), but because once the grades are set, different impact onto the graduation level prediction can be solely explained by different grade distributions. For instance, assume the task to predict qualification for a subsequent master's programme. The qualification is decided by achieving a certain GPA at graduation. As the grade-based input features are already set, final GPA is rendered to a consequence and disparities can be explained by differences in cumulative grades.

The same argument can be made for using student demographics as features. Hereby, student demographics refer to protected attributes such as gender, race, religion, or socioeconomic status [2]. In a discriminatory system which grades minority groups worse than privileged groups, the protected attribute effects achieving lower grades, again rendering final GPA as a consequence. Hence, WYSIWYG holds, explaining outcome disparities due to membership in certain demographic groups. Despite this very discriminatory interpretation, WAE is not an applicable worldview in that scenario: If we assume merit to be equally distributed throughout all demographic groups, it generally will not hold that unevenly distributed cumulative GPAs should result in equally distributed final GPAs.

The above observations indicate that APP assumes WYSIWYG. This can further be supported by the following two argumentations. Firstly, due to unequal distribution of resources among demographic groups, educational disparities are to be expected [1]. Secondly, there is a difference between ideal and non-ideal fairness-perspectives [11]. The fairness ideal would imply that grade-level outcomes are equally distributed throughout groups. Our world is however non-ideal and the fairness ideal is the target state we aim to achieve. For this, we measure the deviation of our systems from the fairness ideal in FairML and attempt to minimise it [11].

As WYSIWYG for APP seems to find support in literature, consequentially APP pairs with the fairness notion of equalised odds. This aligns with (and generalises) the statement of Kizilcec and Lee [18] that equalised odds is "most appropriate in applications like student dropout prediction". Having singled out equalised odds as fairness notion, we will inspect its suitability further and discard demographic parity in the remainder of this paper.

6.2 A Closer Look at Equalised Odds

While we identified equalised odds as a fairness notion which pairs well with APP, there are further concerns in literature regarding the fairness notion of a FairML system which remain to be discussed. Fazelpour and Lipton [11] note that the approach to FairML should consider situated and system wide as well as dynamic impacts of APP intervention while Deho et al. [8] promote to focus on equity and need rather than equality.

Favourable outcome revisited. In classical FairML, we assume $\hat{y} = 1$ to denote a favourable outcome, such as an approved credit loan or being hired at a new job. Intuitively, the favourable outcome in APP for a student is to be predicted as a successful student. However, the real classification task behind APP is rather to predict the need of intervention to help the student achieve a higher performance. The emphasis from a stakeholder's perspective lies on the need of action. Thus we can reframe the favourable outcome in APP as dependent on Y. For at-risk students with $y = 1$ the favourable outcome is indeed $\hat{y} = 1$ so they receive the intervention. For students who will graduate without further intervention and thus $y = 0$ the favourable outcome would be to not get flagged as at-risk, i.e. $\hat{y} = 0$. Thus, for APP, the favourable outcome would be a perfect predictor with $\hat{y} = y$. Such a predictor would always satisfy equalised odds [15]. This differs from classical FairML as the students did not apply for the interventions, contrasting loan or job applications where we assume an approved application to be favoured by the individual.

Long-term impacts. Liu et al. [20] show that both, demographic parity as well as equal true positive rates (only Equation 2 from equalised odds satisfied), are able to cause improvement, stagnation, or even decline in the long-term well-being of disadvantaged groups, depending on the settings. While not considering the stricter notion of equalised odds, their results still suggest that further inspection of respectively underlying distributions of Y needs to be accounted for. Contrasting this with the results of Yeom and Tschantz [32] however, that equalised odds will not amplify discrimination when WYSIWYG holds, gives at least some kind of (theoretical) reassurance of the selection of equalised odds as fairness notion. Further, due to the intervening nature of APP as well as the favourable outcome being dependent on Y, we can illustrate at least a partial improvement over time. As stated above, educational disparities are to be expected due to resources being unequally distributed and our world being non-ideal [1, 11]. Hence, we can assume a proportionally higher rate of $y = 1$ in minority groups. For an APP system satisfying equalised odds, this would result in a higher proportion of minority students receiving the intervention. Assuming the intervention increases graduation rate and/or graduation quality, it should increase the availability of resources for these groups long-term. Thus, the divergence from the fairness-ideal should be reduced. This however only narrows the gap but will be unable to close it, as for instance biases in grading may not be cured in this process.

From Equality to Equity. Instead of promoting equal treatment as measure of fairness, Deho et al. [8] propose to focus on equity and needed treatment instead. However, it is unclear from their paper whether they regard equalised odds to be a measure of equity, whereby Jiang and Pardos [16] apply data rebalancing techniques to boost equity for an APP system in terms of true positive and true negative rates, hence they use equalised odds as measure for equity. This makes sense for APP, as the intervening nature inherently attempts to target students at risk. However, Naggita and Aguma [27] show that a system satisfying equalised odds can still promote inequity. This is conditioned over the accessibility of the system towards the demographic groups. *Accessibility* is hereby defined over the notion of obstacles which obstruct an individual to exhibit their true feature vector towards the prediction system. Such obstructions could be due to biased grading processes which APP alone is unable to solve.

Limitations. Corbett-Davies and Goel [6] have shown that equalised odds, as well as all parity based notions, is subject to the problem of infra-marginality as a unified classification threshold is not sensible if the underlying risk distributions are unequal for two demographic groups. In such cases, the error scores will differ and parity cannot be achieved. Furthermore, equalised odds is usually only satisfiable when different classification thresholds for the demographic groups are employed in the first place [14, 18]. In such cases, the use of the protected attribute is needed at classification time, which might not everywhere be legally feasible. However, Yu et al. [33] argue that APP systems such as dropout detection should include protected attributes, albeit the authors only report a limited benefit in terms of fairness and performance.

Students' Perceived Fairness. First work analyses the implications and perceptions of fairness in APP systems [25, 30], however a more thorough investigation regarding equalised odds needs yet to be conducted. While Smith et al. [30] report student's focus on relational and stake fairness, which equalised odds could cater to, Marcinkowski et al. [25] report focus on distributional and procedural fairness dimensions. Although equalised odds fits procedural fairness, it fails to do so for distributional fairness which would rather be satisfied by demographic parity instead. This could be overcome by a weighted trade-off between both notions as suggested by Kizilcec and Lee [18]. However, it is unclear whether the benefits of equalised odds remain unaffected in this case or whether the student body is willing of such a compromise.

6.3 Undiscussed Notions

We only described two fairness notions in Section 3, but current literature provides a plethora of further notions [9, 22, 26, 31] While it is not possible for us to talk about all of them, we will highlight selected notions and outline their relevance for APP or why we discarded them in our work.

Next to demographic parity and equalised odds, calibration [29] and predictive parity are also popular notions in literature. However, Yeom and Tschantz [32] showed that neither WAE nor WYSIWYG motivate either notion.

Work that considers worldviews usually promotes individual fairness [10] as suitable for a WYSIWYG setting [12, 18]. Individual fairness is strictly not parity based, but we intended to review parity based notions specifically. However, as both, equalised odds and individual fairness, are promoted for WYSIWYG settings, an investigation of their relationship should be followed up in future work.

Gardner et al. [14] introduced ABROCA scores as measure for fairness, which rely on different ROC curves of the demographic groups. While equalised odds is satisfied at intersections of ROC curves, slicing analysis with ABROCA allows for a more nuanced evaluation of the overall fairness trends for different classification thresholds. Specifically, if one does not require equality for the demographic groups in Equations 2 and 3 but only requires an absolute difference of at most ϵ, ABROCA might allow for easier selection of classification thresholds. Whether guarantees regarding disparity amplification under WYSIWYG stay true is left for future work.

Yeom and Tschantz [32] define the notion of an α-hybrid worldview which assumes that discrimination in $\mathcal{Y}$ is partially explained in $\mathcal{Y}'$ to a ration of $\alpha \in [0, 1]$ and thus positions itself between WAE and WYSIWYG. While the authors present the α-disparity test as a fairness measure, the value of α needs to be approximated by social research as well as public discourse.

7. CONCLUSIONS

In this work we reviewed recent literature in search of finding a suitable fairness notion to employ in responsible APP systems. The consensus of our search favours equalised odds over demographic parity, calibration, or predictive parity. After highlighting APPs relation to WYSIWYG, we further found support of equalised odds in terms of reframing the favourable outcome, inspecting possible long-term impacts and partly relating to equity notions. While equalised odds still shows limitations in its applicability, we emphasise the need of further analysis regarding equalised odds in APP contexts specifically: in terms of equity, relation to individual fairness, and perceived fairness.

8. ACKNOWLEDGMENTS

This work was supported by the Federal Ministry of Education and Research (BMBF) under Grand No. 16DHB4020.

References

[1] AERA. *Standards for educational and psychological testing.* American Educational Research Association, 1999.

[2] E. Alyahyan and D. Düştegör. Predicting academic success in higher education: literature review and best practices. *International Journal of Educational Technology in Higher Education*, 17(1), feb 2020. doi: 10.1186/s41239-020-0177-7.

[3] M. Attaran, J. Stark, and D. Stotler. Opportunities and challenges for big data analytics in US higher education. *Industry and Higher Education*, 32(3):169–182, apr 2018. doi: 10.1177/0950422218770937.

[4] S. Barocas, E. Bradley, V. Honavar, and F. Provost. Big data, data science, and civil rights. *arXiv preprint arXiv:1706.03102*, 2017.

[5] A. Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.

[6] S. Corbett-Davies and S. Goel. The measure and mismeasure of fairness: A critical review of fair machine learning. July 2018.

[7] B. Daniel. Big data and analytics in higher education: Opportunities and challenges. *British Journal of Educational Technology*, 46(5):904–920, dec 2014. doi: 10.1111/bjet.12230.

[8] O. B. Deho, C. Zhan, J. Li, J. Liu, L. Liu, and T. D. Le. How do the existing fairness metrics and unfairness mitigation algorithms contribute to ethical learning analytics? *British Journal of Educational Technology*, apr 2022. doi: 10.1111/bjet.13217.

[9] J. Dunkelau and M. Leuschel. Fairness-aware machine learning: An extensive overview. Working paper, available at https://www3.hhu.de/stups/downloads/pdf/fairness-survey.pdf, Oct. 2019.

[10] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226. ACM, 2012.

[11] S. Fazelpour and Z. C. Lipton. Algorithmic fairness from a non-ideal perspective. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society.* ACM, feb 2020. doi: 10.1145/3375627.3375828.

[12] S. A. Friedler, C. Scheidegger, and S. Venkatasubramanian. On the (im)possibility of fairness. *arXiv preprint arXiv:1609.07236*, 2016.

[13] S. A. Friedler, C. Scheidegger, S. Venkatasubramanian, S. Choudhary, E. P. Hamilton, and D. Roth. A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency.* ACM, jan 2019. doi: 10.1145/3287560.3287589.

[14] J. Gardner, C. Brooks, and R. Baker. Evaluating the fairness of predictive student models through slicing analysis. In *Proceedings of the 9th International Conference on Learning Analytics and Knowledge.* ACM, mar 2019. doi: 10.1145/3303772.3303791.

[15] M. Hardt, E. Price, N. Srebro, et al. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pages 3315–3323, 2016.

[16] W. Jiang and Z. A. Pardos. Towards equity and algorithmic fairness in student grade prediction. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society.* ACM, jul 2021. doi: 10.1145/3461702.3462623.

[17] B. Keller, M. Lünich, and F. Marcinkowski. How is socially responsible academic performance prediction possible? In *Strategy, Policy, Practice, and Governance for AI in Higher Education Institutions*, pages 126–155. IGI Global, may 2022. doi: 10.4018/978-1-7998-9247-2.ch006.

[18] R. F. Kizilcec and H. Lee. Algorithmic fairness in education. arXiv, 2020. doi: 10.48550/ARXIV.2007.05443.

[19] J. Kleinberg, S. Mullainathan, and M. Raghavan. Inherent trade-offs in the fair determination of risk scores. In C. H. Papadimitriou, editor, *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*, volume 67 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 43:1–43:23, Dagstuhl, Germany, 2017. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

[20] L. T. Liu, S. Dean, E. Rolf, M. Simchowitz, and M. Hardt. Delayed impact of fair machine learning. In J. Dy and A. Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3150–3158. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/liu18c.html.

[21] A. Loukina, N. Madnani, and K. Zechner. The many dimensions of algorithmic fairness in educational applications. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–10. Association for Computational Linguistics, Aug. 2019. doi: 10.18653/v1/w19-4401.

[22] K. Makhlouf, S. Zhioua, and C. Palamidessi. Machine learning fairness notions: Bridging the gap with real-world applications. *Information Processing & Management*, 58(5):102642, sep 2021. doi: 10.1016/j.ipm.2021.102642.

[23] J. M. Malouff and E. B. Thorsteinsson. Bias in grading: A meta-analysis of experimental research findings. *Australian Journal of Education*, 60(3):245–256, sep 2016. doi: 10.1177/0004944116664618.

[24] J. M. Malouff, S. J. Stein, L. N. Bothma, K. Coulter, and A. J. Emmerton. Preventing halo bias in grading the work of university students. *Cogent Psychology*, 1(1): 988937, dec 2014. doi: 10.1080/23311908.2014.988937.

[25] F. Marcinkowski, K. Kieslich, C. Starke, and M. Lünich. Implications of AI (un-)fairness in higher education admissions. In *Proceedings of the 2020 Conference on*

Fairness, Accountability, and Transparency. ACM, jan 2020. doi: 10.1145/3351095.3372867.

[26] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54(6):1–35, jul 2021. doi: 10.1145/3457607.

[27] K. Naggita and J. C. Aguma. The equity framework: Fairness beyond equalized predictive outcomes. arXiv, 2022. doi: 10.48550/ARXIV.2205.01072.

[28] D. Pessach and E. Shmueli. Algorithmic fairness. volume abs/2001.09784, 2020.

[29] G. Pleiss, M. Raghavan, F. Wu, J. Kleinberg, and K. Q. Weinberger. On fairness and calibration. In *Advances in Neural Information Processing Systems*, pages 5680–5689, 2017.

[30] L. M. Smith, L. Todd, and K. Laing. Students' views on fairness in education: the importance of relational justice and stakes fairness. *Research Papers in Education*, 33(3):336–353, mar 2017. doi: 10.1080/02671522.2017.1302500.

[31] S. Verma and J. Rubin. Fairness definitions explained. In *FairWare'18: IEEE/ACM International Workshop on Software Fairness*. ACM, May 2018.

[32] S. Yeom and M. C. Tschantz. Avoiding disparity amplification under different worldviews. In *Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 273–283, New York, NY, USA, 2021. ACM. doi: 10.1145/3442188.3445892.

[33] R. Yu, H. Lee, and R. F. Kizilcec. Should college dropout prediction models include protected attributes? In *Proceedings of the Eighth ACM Conference on Learning @ Scale*. ACM, jun 2021. doi: 10.1145/3430895.3460139.

[34] I. Žliobaitė. Measuring discrimination in algorithmic decision making. *Data Mining and Knowledge Discovery*, 31(4):1060–1089, July 2017.

3

Measuring and Mitigating Bias in Machine Learning Datasets

In the last chapter, we have presented works [1, 8] about assessing fairness in machine learning models. The studies were carried out as part of the RAPP project and accordingly focused on educational use cases. The studies highlight the importance of incorporating fairness assessment when evaluating machine learning models, as biased models can have severe societal consequences.

Many methods exist to mitigate bias in machine learning pipelines. They can be categorized into three groups, namely *pre-processing*, *in-processing*, and *post-processing* [13, 20, 41]. Each comes with its own advantages and disadvantages [20] regarding the user's needs. Because pre-processing acts on the data directly, it can be expected that multiple machine learning models will benefit from the same pre-processed dataset. This advantage was especially important for the RAPP project, as the decision on which machine learning model to use was not fixed and could change over time due to new research findings, stakeholders' needs, or legal requirements.

Following this motivation, we present three works [2, 3, 4] that focus on measuring and mitigating bias in datasets. All works contribute with new methods and findings for various scenarios: (1) single protected attribute with two groups [2], (2) single protected attribute with non-binary groups [3], and multiple protected attributes [4] that can contain any number of groups. The last two papers are built upon their previous ones and encompass more complex cases. With these works, problems that have not been addressed thoroughly before are tackled. All the scenarios that we have dealt with cover all possible cases that can occur in datasets with sensitive information. Methodologically, our methods are under- and oversampling strategies that are formulated as combinatorial optimization problems. Due to the high dimensionality of the stated problems, the following works employ heuristics to find approximate solutions. Notably, the framework introduced in the second work [3] can be seen as a generalization of the first work [2].

3.1 Improving Fairness for Binary Groups with Generated Data

Paper: Manh Khoi Duong and Stefan Conrad. Dealing with Data Bias in Classification: Can Generated Data Ensure Representation and Fairness? In *Big Data Analytics and Knowledge Discovery*, volume 14148 of *Lecture Notes in Computer Science*. Springer Cham, 2023.

Personal Contribution: The idea, methodology, and implementation of the research were developed solely by Manh Khoi Duong. All parts of the paper were written by Manh Khoi Duong under the supervision of Stefan Conrad.

Status: Published

Algorithms are only as good as the data that is provided to them [14]. If the data is biased, then the decisions made by algorithms are most likely biased as well. Hence, mitigating bias in data can ensure fair decisions. Many techniques [51, 25, 19] alter the data by editing features, protected attributes, or labels. We argue in this paper that these approaches are less interpretable and cannot be used in real-life applications safely. Another aspect is that many fairness definitions exist and fulfill different purposes. The user should be able to choose the fairness definition that fits the application best, and the pre-processing method should be able to handle the chosen fairness definition.

Motivated by the rise of generative models in recent years [46, 47], we deal with the research question of how synthetic data can be used for fairness. To make the pre-processed data trustworthy, we formulated a combinatorial optimization problem where synthetic data is added to the original dataset to ensure representation and fairness. To solve the optimization problem, we developed a simple but effective heuristic. The user is left with the choice of how many synthetic points are added to the original dataset. Depending on the modality of the data, the user can also decide which generative model to use. Because only synthetic data is added and features are not modified, our approach can be considered more interpretable: The resulting data contains the original but has additional synthetic data, where the amount is controlled by the user.

To compare our approach against popular pre-processing methods [33, 51, 25], we used the implementations and datasets [38, 37, 31, 44] given by `AIF360` [15]. All datasets contain a protected attribute with a privileged and an unprivileged group, where the machine learning task is to predict the binary target label. Pre-processing methods aim to make the two groups more equal in terms of the target label. The experiments showed that our approach is promising compared to existing methods [33, 51, 25]. The pre-processed datasets, attained with our approach, are fairer than the original and do not negatively impact machine learning classifiers in terms of accuracy. We also showed that our approach is indeed *fairness-agnostic* and can handle any fairness definition. These results confirm that bias can be mitigated by adding synthetic data points and alteration is not necessary.

Dealing with Data Bias in Classification: Can Generated Data Ensure Representation and Fairness?

Manh Khoi Duong(✉) and Stefan Conrad

Heinrich Heine University, Universitätsstraße 1, 40225 Düsseldorf, Germany
{manh.khoi.duong,stefan.conrad}@hhu.de

Abstract. Fairness is a critical consideration in data analytics and knowledge discovery because biased data can perpetuate inequalities through further pipelines. In this paper, we propose a novel pre-processing method to address fairness issues in classification tasks by adding synthetic data points for more representativeness. Our approach utilizes a statistical model to generate new data points, which are evaluated for fairness using discrimination measures. These measures aim to quantify the disparities between demographic groups that may be induced by the bias in data. Our experimental results demonstrate that the proposed method effectively reduces bias for several machine learning classifiers without compromising prediction performance. Moreover, our method outperforms existing pre-processing methods on multiple datasets by Pareto-dominating them in terms of performance and fairness. Our findings suggest that our method can be a valuable tool for data analysts and knowledge discovery practitioners who seek to yield for fair, diverse, and representative data.

Keywords: fairness · bias · synthetic data · fairness-agnostic · machine learning · optimization

1 Introduction

Data analytics has grown in popularity due to its ability to automate decision-making through machine learning. However, real-world data can contain biases that produce unfair outcomes, making fairness in data pipelines involving machine learning a pressing concern. Fairness in machine learning typically deals with intervening algorithms providing equitable outcomes regardless of protected characteristics such as gender, race, or age group.

The existing related works can be divided into three categories [5,8,20]. The first category of methods are pre-processing methods, which aim to reduce bias in the data. Examples of such methods include data augmentation and data

This work was supported by the Federal Ministry of Education and Research (BMBF) under Grand No. 16DHB4020.

R. Wrembel et al. (Eds.): DaWaK 2023, LNCS 14148, pp. 176–190, 2023.
https://doi.org/10.1007/978-3-031-39831-5_17

balancing [2]. The second category of methods are in-processing methods, which aim to enforce fairness constraints during the training procedure [15]. Examples of in-processing methods include regularization techniques and constrained optimization [31]. The last category are post-processing methods that allow the improvement of fairness after training by correcting the outputs of the trained model [14].

The goal of this paper is to introduce a pre-processing method that achieves fairness by including generated data points. This is done by utilizing a statistical model that learns the distribution of the dataset, enabling the generation of synthetic samples. Additionally, a discrimination measure is employed to evaluate the fairness when incorporating the generated data points. Our method treats the discrimination measure as a black-box, making it able to optimize any discrimination measure defined by the user. We refer to this property of our algorithm as *fairness-agnostic*. This makes it suitable for cases where a specific fairness notion is required.

For the experimentation, multiple datasets known to be discriminatory were used. The experiments were performed by firstly loading the datasets and then pre-processing them using different pre-processing techniques. The pre-processed datasets were then fed into several classifiers. The performance of each classifier was then evaluated in terms of performance and fairness to assess the effectiveness of the pre-processing methods. Our experiments have empirically shown that our technique effectively lessens discrimination without sacrificing the classifiers' prediction qualities. Moreover, it is compatible with any machine learning model. Of the pre-processors tested, none were able to meet all of these conditions. The scope and application of our method is not necessarily limited to tabular data and classification tasks, even though experiments were conducted on them. The method is more broadly suitable for supervised learning tasks where the data, label, and protected attribute are available. Only the appropriate discrimination measures have to be derived for the right task. Generally, our primary contributions are:

- The introduction of a novel pre-processing technique that can optimize any given fairness metric by pre-selecting generated data points to include into the new fair dataset.
- We carry out a comprehensive empirical study, comparing our method against three widely recognized pre-processors [9,13,31], using multiple datasets commonly found in fairness literature.
- We present interesting and valuable properties, such as the empirical evidence that our method consistently improved fairness in comparison to the unprocessed data.

2 Related Work

Many pre-processing algorithms in literature alter the dataset to achieve fairness [4,9,31]. Because the methods simply return a fair dataset, they can be used with any estimator. However, such approaches cannot be used with ease: They often require a parameter setting that sets how aggressive the change should be. As the approaches differ in their methodology, it is hard to interpret the parameter's setting and their unexpected effects on the data. Data alteration methods also have a higher risk of producing data that do not resemble the original data distribution in any ways.

Other approaches return a weight for each sample in the dataset that the estimator should account for when fitting the data [1,13]. While the approaches seem promising [1,13], they require estimators to be able to handle sample weights. A way to account for this is to replicate samples based on their sample weights. However, this is not computationally scalable for larger datasets or for larger differences between the sample weights.

Another related approach is removing data samples that influence estimators in a discriminatory way [28]. Nevertheless, this approach does not seem feasible for smaller datasets.

Differently from related works, we present an algorithm that does not come with the above mentioned drawbacks. Further, our approach is able to satisfy any fairness notion that is defined for measuring discrimination or bias in the dataset. While the work of Agarwal et al. [1] also features this property, the fairness definitions must be formalizable by linear inequalities on conditional moments. In contrast, our work requires the fairness definitions to quantify discrimination in a numeric scale where lower values indicate less discrimination. This can be as simple as calculating the differences of probabilistic outcomes between groups.

While there exist works that train fair generative models to produce data that is fair towards the protected attribute on images [7,24,27] or tabular data [12,23], our approach can be seen as a framework that employs generative models and can therefore be used for any data where the protected attribute is accessible. Specifically, our research question is not *"How can fair generative models be constructed?"*, we instead deal with the question *"Using any statistical or generative model that learns the distribution of the dataset, how can the samples drawn from the distribution be selected and then included in the dataset such that fairness can be guaranteed?"*. Other works that generate data for fairness include generating counterfactuals [26] and generating pseudo-labels for unlabeled data [6].

3 Measuring Discrimination

In this section, we briefly present *discrimination measures* that assess the fairness of data. For that, we make use of following notation [5,8,20]: A *data point* or *sample* is represented as a triple (x, y, z), where $x \in X$ is the *feature*, $y \in Y$ is the ground truth *label* indicating favorable or unfavorable outcomes, and $z \in Z$ is the *protected attribute*, which is used to differentiate between groups. The sets X, Y, Z typically hold numeric values and are defined as $X = \mathbb{R}^d$, $Y = \{0, 1\}$, and $Z = \{1, 2, \ldots, k\}$ with $k \geq 2$. For simplicity, we consider the case where protected attributes are binary, i.e., $k = 2$. Following the preceding notation, a *dataset* is defined as the set of data points, i.e., $\mathcal{D} = \{(x_i, y_i, z_i)\}_{i=1}^{n}$. Machine learning models $\phi : X \times Z \to Y$ are trained using these datasets to predict the target variable $y \in Y$ based on the input variables $x \in X$ and $z \in Z$. We call the output $\hat{y} := \phi(x, z)$ *prediction*.

Based on the work of [32], we derive *discrimination measures* to the needs of the pre-processing method in this paper. To make our algorithm work, a *discrimination measure* must satisfy certain properties which we introduce in the following.

Definition 1. *A* discrimination measure *is a function $\psi : \mathbb{D} \to \mathbb{R}^+$, where $\mathbb{D}$ is the set of all datasets, satisfying the following axioms:*

1. *The discrimination measure $\psi(\cdot)$ is bounded by [0, 1]. (Normalization)*
2. *Minimal and maximal discrimination are captured with 0, 1 by $\psi(\cdot)$, respectively.*

The first and second axiom together assure that the minimal or maximal discrimination can be assessed by this measure. Furthermore, through normalization it is possible to evaluate the amount of bias present and its proximity to the optimal solution. As achieving no discrimination is not always possible, i.e., $\psi(\mathcal{D}) = 0$, we consider lower discrimination as better and define a fairer dataset as the one with the lower discrimination measure among two datasets.

Literature [2,5,8,19,20,32] on fairness-aware machine learning have classified fairness notions to either representing group or individual fairness. We subdivide the most relevant fairness notions into two categories which are *dataset* and *prediction notions* and derive discrimination measures from it as suggested by [32]. From now on, we denote x, y, z as random variables describing the events of observing an individual from a dataset $\mathcal{D}$ taking specific values.

Dataset notions typically demand the independency between two variables. When the protected attribute and the label of a dataset are independent, it is considered fair because it implies that the protected attribute does not influence or determine the label. An example to measure such dependency would be the *normalized mutual information* (NMI) [29] where independency can be concluded if and only if the score is zero. Because it is normalized as suggested by the name, it is a discrimination measure.

Definition 2 (Normalized mutual information). *Let $H(\cdot)$ be the entropy and $I(y; z)$ be the mutual information [25]. The normalized mutual information score is defined in the following [30]:*

$$\psi_{NMI}(\mathcal{D}) = 2\frac{I(y; z)}{H(y) + H(z)}.$$

Statistical parity [15,31] and *disparate impact* [9] are similar notions that also demand independency, except they are specifically designed for binary variables. Kang et al. [16] proved that zero mutual information is equivalent to statistical parity. To translate statistical parity to a discrimination measurement, we make use of differences similarly to Žliobaitė [32].

Definition 3 (Statistical parity). *Demanding that each group has the same probability of receiving the favorable outcome is statistical parity, i.e.,*

$$p(y = 1 \mid z = 1) = p(y = 1 \mid z = 0).$$

Because we want to minimize discrimination towards any group, we measure the absolute difference between the two groups to assess the extent to which the dataset fulfills statistical parity. This is also known as (absolute) statistical disparity (SDP) [8]. A value of 0 indicates minimal discrimination:

$$\psi_{SDP}(\mathcal{D}) = |p(y = 1 \mid z = 1) - p(y = 1 \mid z = 0)|. \tag{1}$$

Because disparate impact [9] essentially demands the same as statistical parity but contains a fraction, dividing by zero is a potential issue that may arise. Therefore, its use should be disregarded [32]. Note that dataset notions can also be applied to measure the fairness on predictions by exchanging the data label with the prediction label.

Parity-based notions, fulfilling the *separation* or *sufficiency* criterion [2], require both prediction and truth labels to evaluate the fairness. Contrary to the category before, measuring solely on datasets is not possible here. Despite this, it is still essential to evaluate on such measures to account for algorithmic bias. Here, the discrimination measure takes an additional argument, which is the prediction label $\hat{y}$ as a random variable. According fairness notions are, for example, *equality of opportunity* [10], *predictive parity* [2], and *equalized odds* [2].

Definition 4 (Equalized odds). *Equalized odds is defined over the satisfaction of both* equality of opportunity *and* predictive parity *[10],*

$$p(\hat{y} = 1 \mid y = i, z = 1) = p(\hat{y} = 1 \mid y = i, z = 0) \;\; \forall i \in \{0, 1\},$$

where equality of opportunity is the case of $i = 1$ and predictive parity is the case of $i = 0$, correspondingly. Making use of the absolute difference, likewise to SDP (1), we denote the measure of equality of opportunity as $\psi_{EO}(\mathcal{D}, \hat{y})$ and predictive parity as $\psi_{PP}(\mathcal{D}, \hat{y})$.

To turn equalized odds into a discrimination measure, we can calculate the average of the absolute differences for both equality of opportunity and predictive parity. This is referred to as *average odds error* [3]:

$$\psi_{ODDS}(\mathcal{D}, \hat{y}) = \frac{\psi_{EO}(\mathcal{D}, \hat{y}) + \psi_{PP}(\mathcal{D}, \hat{y})}{2}. \tag{2}$$

4 Problem Formulation

Intuitively, the goal is to add an amount of synthetic datapoints to the original data to yield for minimal discrimination. With the right discrimination measure chosen, it can be ensured that the unprivileged group gets more exposure and representation in receiving the favorable outcome. Still, the synthetic data should resemble the distribution of the original data. The problem can be stated formally in the following: Let $\mathcal{D}$ be a dataset with cardinality n, let $\tilde{n}$ be the number of samples to be added to $\mathcal{D}$. The goal is to find a set of data points $S = \{d_1, d_2, \ldots, d_{\tilde{n}}\}$ that can be added to the dataset, i.e., $\mathcal{D} \cup S$ with $\mathcal{S} \sim P(\mathcal{D})$, that minimizes the discrimination function $\psi(\mathcal{D} \cup S)$. Hence, we consider the following constrained problem:

$$\begin{aligned} \min \quad & \psi(\mathcal{D} \cup \mathcal{S}) \\ \text{subject to} \quad & \mathcal{S} \sim P(\mathcal{D}) \\ & |\mathcal{S}| = \tilde{n}. \end{aligned} \tag{3}$$

The objective (3) suggests that the samples d_i that are added to the dataset $\mathcal{D}$ are drawn from $P(\mathcal{D})$. To draw from $P(\mathcal{D})$, a statistical or generative model P_G that learns the data distribution can be used. Therefore generating data samples and bias mitigation are treated as sequential tasks where the former can be solved by methods from literature [22]. Because the discrimination measure ψ can be of any form, the optimization objective is treated as a black-box and is solved heuristically.

5 Methodology

Our algorithm relies on a statistical model, specifically the Gaussian copula [22], to learn the distribution of the given dataset $P(\mathcal{D})$. Gaussian copula captures the relationship between variables using Gaussian distributions. While assuming a Gaussian relationship, the individual distributions of the variables can be any continuous distribution, providing flexibility in modeling the data.

Still, the type of model for this task can be set by the user as long as it can sample from $P(\mathcal{D})$. Because discrimination functions are treated as black-boxes, the algorithm does not require the derivatives of ψ and optimizing for it leads to our desired *fairness-agnostic* property: It is suitable for any fairness notion that can be expressed as a discrimination function. Our method handles the size constraint in Eq. (3) as an upper bound constraint, where a maximum of $\tilde{n}$ samples are added to $\mathcal{D}$.

Our method, outlined in Algorithm 1, begins by initializing $\hat{\mathcal{D}}$ with the biased dataset $\mathcal{D}$. Then $\hat{n}$ is set as a multiplicative $r > 1$ of the original dataset's size. Lastly in the initialization, the distribution of $P(\mathcal{D})$ is learned by a generative model P_G. The algorithm then draws m samples from the generative model P_G which are referred to as the set of candidates C. The next step is decisive for the optimization (Line 9): The candidate which minimizes the discrimination most when included in the dataset $\hat{\mathcal{D}}$ is added to $\hat{\mathcal{D}}$. The steps of drawing samples and adding the best candidate to the dataset is repeated till $\hat{\mathcal{D}}$ has a cardinality of $\hat{n}$ or the discrimination is less than the fairness threshold ϵ. Because ϵ is set to 0 by default, the algorithm can stop earlier before the dataset reaches its requested size if the discrimination cannot be further reduced, i.e., $\psi(\hat{\mathcal{D}}) = 0$. Because calculating $\psi(\hat{\mathcal{D}} \cup \{c\})$ (Line 9) does not involve retraining any classifier and solely

Algorithm 1. Pseudocode of MetricOptGenerator

```
Input: $\mathcal{D}, r = 1.25, m = 5, \epsilon = 0$
Output: $\hat{\mathcal{D}}$
  Initialization:
  $\hat{\mathcal{D}} \leftarrow \mathcal{D}$
  $\hat{n} \leftarrow \lfloor r \cdot |\mathcal{D}| \rfloor$
  $P_G \leftarrow$ learn distribution of $P(\mathcal{D})$
  Generating fair samples:
  for $i = 1$ to $\hat{n} - |D|$ do
      if $(\psi(\hat{\mathcal{D}}) \leq \epsilon)$ then
          return $\hat{\mathcal{D}}$
      end if
      $C \leftarrow$ sample $m$ candidates from $P_G$
      $\hat{\mathcal{D}} \leftarrow \hat{\mathcal{D}} \cup \{\text{argmin}_{c \in C}\, \psi(\hat{\mathcal{D}} \cup \{c\})\}$
  end for
  return $\hat{\mathcal{D}}$
```

evaluates the dataset, this step is practically very fast. In our implementation, we generate a set of synthetic data points prior to the for-loop, eliminating the sampling cost during the optimization step. We refer to Appendix A for the proof outlining the polynomial time complexity of the presented method.

6 Evaluation

To evaluate the effectiveness of the presented method against other pre-processors in ensuring fairness in the data used to train machine learning models, we aim to answer following research questions:

- **RQ1** What pre-processing approach can effectively improve fairness while maintaining classification accuracy, and how does it perform across different datasets?
- **RQ2** How stable are the performance and fairness results of classifiers trained on pre-processed datasets?
- **RQ3** How does pursuing for statistical parity, a data-based notion, affect a prediction-based notion such as average odds error?
- **RQ4** Is the presented method fairness-agnostic as stated?

To especially address the first three research questions, which deal with effectiveness and stability, we adopted the following experimental methodology: We examined our approach against three pre-processors on four real-world datasets (see Table 1). The pre-processors we compare against are *Reweighing* [13], *Learning Fair Representation* [31] (LFR), and *Disparate Impact Remover* [9] (DIR). The data were prepared such that categorical features are one-hot encoded and rows containing empty values are removed from the data. We selected sex, age, race, and foreign worker as protected attributes for the respective datasets. Generally, the data preparation was adopted from AIF360 [3].

Table 1. Overview of datasets.

Dataset	Protected Attribute	Label	Size	Description
Adult [17]	Sex	Income	45 222	Indicates individuals earning over $50 000 annually
Bank [21]	Age	Term Deposit	30 488	Subscription to a term deposit
COMPAS [18]	Race	Recidivism	6 167	Arrested again for a new offense within a period of 2 years after initial arrest
German [11]	Foreign Worker	Credit Risk	1 000	Creditworthiness of loan applicants

All hyperparameter settings of the pre-processsors were kept as they are, given the implementation provided by AIF360 [3]. For the case of LFR, we empirically had to lower the hyperparameter of optimizing for fairness. It was initially set too high which led to identical predictions for all data points. For our approach, we set $r = 1.25$ which returns a dataset consisting of additional 25% samples of the dataset's initial size. The discrimination measure chosen was the absolute difference of statistical parity (1), which all other methods also optimize for. Further, we set $m = 5$ and $\epsilon = 0$ as shown in Algorithm 1.

The experimental methodology for a single dataset is visualized in Fig. 1 as a pipeline. The given dataset is firstly split into a training (80%) and test set (20%). Afterwards, the training set is then passed into the available pre-processors. Then, all debiased data are used to train several classifiers. We employed three different machine learning algorithms—*k-nearest neighbors* (KNN), *logistic regression* (LR), and *decision tree classifier* (DT)—to analyze the pre-processed datasets and the original, unprocessed dataset for comparison. The unprocessed dataset is referred to as the baseline. Finally, the performance and fairness is evaluated on the prediction of the test set. It is noteworthy to mention that the test sets were left untouched to demonstrate that by pre-processing the training data, unbiased results can be achieved in the prediction space even without performing bias mitigation in the test data. Due to stability reasons (and to handle **RQ2**), we used Monte Carlo cross-validation to shuffle and split the dataset. This was done 10 times for all datasets. The results from it set the performance-fairness baseline. While our optimization focuses on SDP, we address **RQ3** by assessing the error of average odds. To answer **RQ4**, we refer to Sect. 6.2 for the experimentation and discussion.

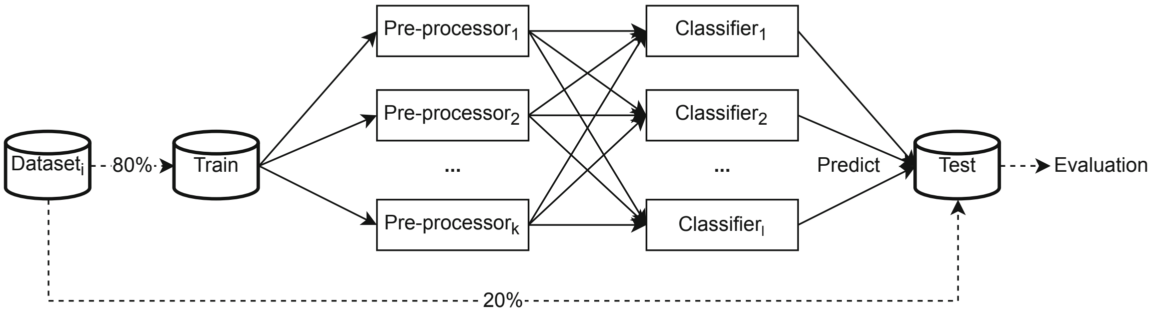


Fig. 1. Experimental methodology visualized as a pipeline

6.1 Comparing Pre-processors

Table 2 presents the performance-fairness test results of pre-processors on different datasets (**RQ1**). For the discrimination, the table displays SDP and average odds error of the predictions on the test sets. To assess the classifier's performances, we used *area under the receiver operating characteristic curve* (AUC). An estimator that guesses classes randomly would produce an AUC score of 0.5. Here, higher scores imply better prediction performances. Means and standard deviations of the Monte Carlo cross-validation results are also displayed to evaluate the robustness (**RQ2**). We note that all classifiers except of KNN were able to handle sample weights in training, which are required for Reweighing. Therefore Reweighing was not able to mitigate bias in KNN and performed as well as the baseline in contrast to other approaches including ours.

Because all pre-processors aim to reduce statistical disparity (or the equivalent formulation), we compare the SDP scores between the pre-processors: In most cases, our approach produced Pareto optimal solutions with respect to both SDP and AUC. Generally, only Reweighing and our approach appear to consistently improve fairness without sacrificing notable prediction power. In direct comparison, LFR improved the fairness at most across all experiments but at the same time sacrifices prediction quality of all classifiers to such a great extent that the predictions become essentially useless. In experiments where LFR attained standard deviations of 0 across all scores (Table 2b, 2d), we investigated the pre-processed data and found that LFR had modified almost all labels to a single value. As a result, the estimators were unable to classify the data effectively, as they predicted only one outcome. The results of DIR are very inconsistent. DIR sometimes even worsens the fairness, as seen in the COMPAS and German datasets, where SDP and average odds error are increased in most settings. This situation arises when there is an excessive correction of the available discrimination for the unprivileged group, leading to discrimination against the privileged group. If the discrimination measures are defined such that the privileged or unprivileged groups do not matter (similarly to this paper), reverse discrimination would not mistakenly occur by our approach. This extra property renders our method more suitable for responsible use cases.

When comparing the average odds error rates (**RQ3**), our approach has successfully reduced algorithmic bias without aiming for it under nearly all experiments. The increase in the average odds error rate (mean), albeit negligible, was observed only when training DT on the Banking data and LR in the German dataset. In all other ten model and dataset configurations, our approach did reduce the error rate without particularly optimizing for it. This can be expected in practice as the independency of the label with the protected attribute (SDP) is a sufficient condition for average odds.

Table 2. The tables displays each classifier's mean test performance and discrimination when trained on different pre-processed training sets. The best performing statistic for each classifier is marked in bold. Minimal standard deviations are marked bold, too. All values displayed are percentages.

(a) Adult

		AUC		SDP		AVG Odds Error	
		mean	std	mean	std	mean	std
Model	Preprocessor						
DT	DIR	81.28	0.53	19.65	1.04	24.50	1.27
	LFR	50.18	1.20	**0.18**	**0.46**	**0.16**	**0.37**
	Our	78.91	0.52	9.68	1.19	10.55	1.26
	Original	**81.35**	0.52	19.77	0.96	24.66	1.21
	Reweighing	78.95	**0.48**	4.96	1.22	1.37	0.76
KNN	DIR	75.35	0.86	20.94	2.35	22.31	3.32
	LFR	51.80	4.96	**1.14**	3.26	**0.70**	**2.02**
	Our	75.26	**0.60**	18.84	2.91	19.80	3.54
	Original	**75.53**	0.85	21.09	**2.16**	22.33	3.05
	Reweighing	**75.53**	0.85	21.09	**2.16**	22.33	3.05
LR	DIR	80.12	0.59	17.84	0.46	22.80	0.49
	LFR	55.35	8.64	**1.33**	3.05	**0.80**	2.02
	Our	76.96	0.52	3.60	0.82	1.30	0.63
	Original	**80.13**	0.59	17.75	**0.45**	22.71	**0.48**
	Reweighing	77.29	**0.51**	4.63	0.57	1.90	0.71

(b) Bank

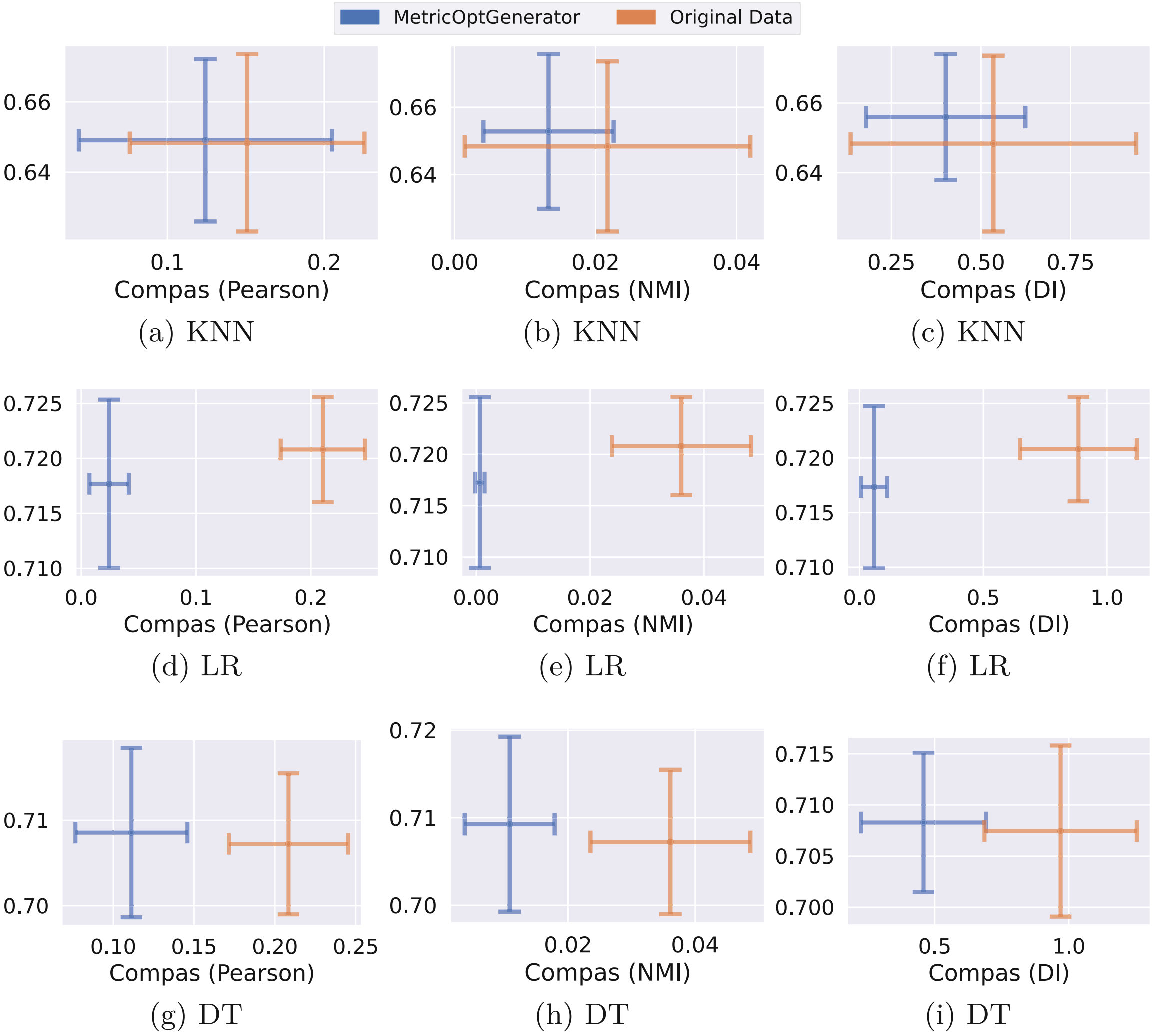
		AUC		SDP		AVG Odds Error	
		mean	std	mean	std	mean	std
Model	Preprocessor						
DT	DIR	67.66	1.24	3.30	2.03	7.81	3.66
	LFR	50.00	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
	Our	72.68	0.96	10.36	3.75	6.93	3.42
	Original	**72.94**	1.15	10.69	1.70	6.49	2.97
	Reweighing	72.81	1.16	9.58	1.99	5.93	3.07
KNN	DIR	81.42	0.82	8.43	3.69	6.56	3.16
	LFR	50.00	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
	Our	**86.98**	0.66	9.05	3.55	5.00	3.74
	Original	**86.98**	0.65	9.05	3.55	5.00	3.75
	Reweighing	**86.98**	0.65	9.05	3.55	5.00	3.75
LR	DIR	91.48	0.42	3.90	1.08	3.60	2.59
	LFR	50.00	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
	Our	91.25	0.51	6.72	2.24	2.87	2.00
	Original	**92.14**	0.34	7.00	2.85	4.37	2.57
	Reweighing	92.11	0.36	5.82	2.47	3.37	1.81

(c) COMPAS

		AUC		SDP		AVG Odds Error	
		mean	std	mean	std	mean	std
Model	Preprocessor						
DT	DIR	70.75	**0.53**	23.58	4.70	22.06	4.71
	LFR	50.26	3.97	**8.32**	21.12	**8.05**	21.17
	Our	**70.91**	0.85	10.55	4.31	8.63	4.06
	Original	70.76	0.82	21.16	**3.64**	19.67	3.77
	Reweighing	70.35	0.97	10.22	4.02	8.98	**2.80**
KNN	DIR	**65.78**	2.88	21.74	7.64	20.58	7.03
	LFR	53.58	6.15	**2.29**	**3.67**	**3.00**	**4.20**
	Our	65.13	**1.49**	12.56	7.98	11.94	7.48
	Original	64.84	2.52	15.62	7.88	14.55	8.16
	Reweighing	64.84	2.52	15.62	7.88	14.55	8.16
LR	DIR	**72.28**	0.54	23.20	3.41	21.31	3.77
	LFR	56.94	9.07	**1.95**	3.44	**2.55**	3.57
	Our	71.78	0.68	2.31	**1.51**	5.38	**1.79**
	Original	72.08	**0.48**	21.74	3.76	20.01	4.06
	Reweighing	71.52	0.76	3.89	2.46	5.61	2.17

(d) German

		AUC		SDP		AVG Odds Error	
		mean	std	mean	std	mean	std
Model	Preprocessor						
DT	DIR	61.15	4.02	23.22	13.70	28.23	14.30
	LFR	50.00	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
	Our	**63.08**	3.47	16.61	10.90	26.78	11.09
	Original	62.76	3.95	15.07	11.29	27.12	10.83
	Reweighing	62.71	4.65	17.53	15.03	33.15	6.73
KNN	DIR	**55.42**	4.44	18.70	9.28	23.91	7.47
	LFR	50.00	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
	Our	54.21	4.34	12.95	2.95	13.81	3.26
	Original	54.08	3.86	16.99	4.63	18.27	4.95
	Reweighing	54.08	3.86	16.99	4.63	18.27	4.95
LR	DIR	78.05	2.15	20.05	11.84	29.79	14.16
	LFR	50.00	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
	Our	77.60	1.73	15.49	10.46	31.00	9.50
	Original	**78.10**	1.67	16.79	11.10	29.83	10.54
	Reweighing	78.05	1.88	16.42	11.10	30.32	10.56



(a) KNN (b) KNN (c) KNN

(d) LR (e) LR (f) LR

(g) DT (h) DT (i) DT

Fig. 2. Results of optimizing different discrimination objectives with our method on the COMPAS dataset. Objectives are ordered by columns, classifiers by rows. The y-axis displays AUC.

6.2 Investigating the Fairness-Agnostic Property

To demonstrate the fairness-agnostic property of our algorithm (**RQ4**), we evaluated our method against the baseline dataset on multiple measures and examine whether the objective was improved (see Fig. 2). The COMPAS dataset was used for this experiment. The chosen objectives are: the absolute value of Pearson's ρ, NMI (2), and the objective of disparate impact (DI) as given by [9]. All other experimental settings remained the same as described prior, except that other pre-processing methods were not used.

It can be observed that all discrimination measures were lowered significantly. Generally, our method was able to optimize on any fairness notion, as evidenced here and Sect. 6.1. It was even able to outperform algorithms that were specifically designed for a single metric, demonstrating its adaptability.

7 Conclusion

Machine learning can be utilized for malicious purposes if estimators are trained on data that is biased against certain demographic groups. This can have an incredibly negative impact on the decisions made and the groups that are being discriminated against.

The presented pre-processing method in this work is a sampling-based optimization algorithm that firstly uses a statistical model to learn the distribution of the given dataset, then samples points from this distribution, and determines which one to add to the data to minimize the discrimination. This process continues until the predefined criteria set by the user are satisfied. The method can optimize any discrimination measure as it is treated as a black-box, making it more accessible for wider use cases.

The results of our experiments demonstrate that our technique is reliable and significantly reduces discrimination while not compromising accuracy. Although a few other methods performed similarly in a few experiments, they were not compatible with certain estimators or even added bias to the original data. Because fairness was improved among the experiments and our method adds samples, it indicates that representativeness can be achieved with our method. Our research underscores the importance of addressing bias in data and we hope to contribute such concerns in data analytics and knowledge discovery applications.

8 Discussion and Future Work

The results of our approach demonstrate that it is possible to achieve fairness in machine learning models using generated data points. Despite our approach showing promise, it is important to acknowledge that our results rely heavily on the quality of the statistical model used to generate synthetic data. For tabular data, Gaussian copula [22] seems to be a good choice.

In future work, we aim to explore the potential of our method in making pre-trained models fairer with our method. While retraining large models using debiased datasets may not always be feasible from a cost-effective perspective, our approach allows using generated data to fine-tune the model for fairness, which provides a more efficient alternative.

Additionally, our evaluation deals with datasets where the protected attribute is a binary variable, which leaves some use cases untreated. Neglecting to recognize non-binary groups can lead to overlooking those who are most in need of attention. Similarly, research on dealing with multiple protected attributes at the same time could be done. This is to make sure that no protected group is being disadvantaged. Previous studies have touched on this subject [1,4,32], but we hope to reformulate these issues as objectives that work with our approach.

A Proof of Time Complexity

Theorem 1 (Time complexity). *If the number of candidates m and fraction r are fixed and calculating the discrimination $\psi(\mathcal{D})$ of any dataset $\mathcal{D}$ takes a linear amount of time, i.e., $\mathcal{O}(n)$, Algorithm 1 has a worst-case time complexity of $\mathcal{O}(n^2)$ where n is the dataset's size when neglecting learning the data distribution.*

Proof. In this proof, we will focus on analyzing the runtime complexity of the for-loop within our algorithm as the steps before such as learning the data distribution depends heavily on the used method. The final runtime of the complete algorithm is simply the sum of the runtime complexities of the for-loop that is focus of this analysis and the step of learning the data distribution.

Our algorithm firstly checks whether the discrimination of the dataset $\hat{\mathcal{D}}$ is already fair. The dataset grows at each iteration and runs for $\lfloor rn \rfloor - n = \lfloor n(r-1) \rfloor$ times. For simplicity, we use $n(r-1)$ and yield,

$$\begin{aligned}\sum_{i=0}^{n(r-1)-1} n+i &= \sum_{i=1}^{n(r-1)} n+i+1 \\ &= \sum_{i=1}^{n(r-1)} n + \sum_{i=1}^{n(r-1)} i + \sum_{i=1}^{n(r-1)} 1 \\ &= n^2(r-1) + \frac{(n(r-1))^2 + (n(r-1)+1)}{2} + n(r-1) \in \mathcal{O}(n^2),\end{aligned}$$

making the first decisive step for the runtime quadratic.

The second step that affects the runtime is returning the dataset that minimizes the discrimination where each of the m candidates $c \in C$ is merged with the dataset, i.e., $\psi(\hat{\mathcal{D}} \cup \{c\})$. The worst-case time complexity of it can be expressed by

$$\begin{aligned}\sum_{i=1}^{n(r-1)} m(n+i) = m \cdot \sum_{i=1}^{n(r-1)} n+i &= m \cdot \left(\sum_{i=1}^{n(r-1)} n + \sum_{i=1}^{n(r-1)} i \right) \\ &= m \cdot \left(n^2(r-1) + \frac{(n(r-1))^2 + (n(r-1))}{2} \right) \in \mathcal{O}(n^2),\end{aligned}$$

which is also quadratic. Summing both time complexities makes the overall complexity quadratic. □

Although the theoretical time complexity of our algorithm is quadratic, measuring the discrimination, which is a crucial part of the algorithm, is very fast and can be assumed to be constant for smaller datasets. Conclusively, the complexity behaves nearly linearly in practice.

In our experimentation, measuring the discrimination of the Adult dataset [17], which consists of 45 222 samples, did not pose a bottleneck for our algorithm.

References

1. Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., Wallach, H.: A reductions approach to fair classification. In: International Conference on Machine Learning, pp. 60–69. PMLR (2018)
2. Barocas, S., Hardt, M., Narayanan, A.: Fairness and Machine Learning. fairmlbook.org (2019). http://www.fairmlbook.org
3. Bellamy, R.K.E., et al.: AI fairness 360: an extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. CoRR arxiv:1810.01943 (2018)
4. Calmon, F., Wei, D., Vinzamuri, B., Natesan Ramamurthy, K., Varshney, K.R.: Optimized pre-processing for discrimination prevention. In: Guyon, I., et al. (eds.) Advances in Neural Information Processing Systems, vol. 30. Curran Associates, Inc. (2017). https://proceedings.neurips.cc/paper/2017/file/9a49a25d845a483fae4be7e341368e36-Paper.pdf
5. Caton, S., Haas, C.: Fairness in machine learning: a survey. arXiv preprint arXiv:2010.04053 (2020)
6. Chakraborty, J., Majumder, S., Tu, H.: Fair-SSL: building fair ML software with less data. arXiv preprint arXiv:2111.02038 (2022)
7. Choi, K., Grover, A., Singh, T., Shu, R., Ermon, S.: Fair generative modeling via weak supervision. In: III, H.D., Singh, A. (eds.) Proceedings of the 37th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 119, pp. 1887–1898. PMLR, 13–18 July 2020
8. Dunkelau, J., Leuschel, M.: Fairness-aware machine learning (2019)
9. Feldman, M., Friedler, S.A., Moeller, J., Scheidegger, C., Venkatasubramanian, S.: Certifying and removing disparate impact. In: proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 259–268 (2015)
10. Hardt, M., Price, E., Srebro, N.: Equality of opportunity in supervised learning. In: Advances in Neural Information Processing Systems 29 (2016)
11. Hofmann, H.: German credit data (1994). https://archive.ics.uci.edu/ml/datasets/Statlog+%28German+Credit+Data%29
12. Jang, T., Zheng, F., Wang, X.: Constructing a fair classifier with generated fair data. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, pp. 7908–7916 (2021)
13. Kamiran, F., Calders, T.: Data preprocessing techniques for classification without discrimination. Knowl. Inf. Syst. **33**(1), 1–33 (2012)
14. Kamiran, F., Karim, A., Zhang, X.: Decision theory for discrimination-aware classification. In: 2012 IEEE 12th International Conference on Data Mining, pp. 924–929. IEEE (2012)
15. Kamishima, T., Akaho, S., Asoh, H., Sakuma, J.: Fairness-aware classifier with prejudice remover regularizer. In: Flach, P.A., De Bie, T., Cristianini, N. (eds.) ECML PKDD 2012. LNCS (LNAI), vol. 7524, pp. 35–50. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33486-3_3
16. Kang, J., Xie, T., Wu, X., Maciejewski, R., Tong, H.: InfoFair: information-theoretic intersectional fairness. In: 2022 IEEE International Conference on Big Data (Big Data), pp. 1455–1464. IEEE (2022)
17. Kohavi, R.: Scaling up the accuracy of Naive-Bayes classifiers: a decision-tree hybrid. In: KDD 1996, pp. 202–207. AAAI Press (1996)

18. Larson, J., Angwin, J., Mattu, S., Kirchner, L.: Machine bias, May 2016. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
19. Makhlouf, K., Zhioua, S., Palamidessi, C.: Machine learning fairness notions: bridging the gap with real-world applications. Inf. Process. Manage. **58**(5), 102642 (2021). https://doi.org/10.1016/j.ipm.2021.102642. https://www.sciencedirect.com/science/article/pii/S0306457321001321
20. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. ACM Comput. Surv. (CSUR) **54**(6), 1–35 (2021)
21. Moro, S., Cortez, P., Rita, P.: A data-driven approach to predict the success of bank telemarketing. Decis. Support Syst. **62**, 22–31 (2014)
22. Patki, N., Wedge, R., Veeramachaneni, K.: The synthetic data vault. In: 2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA), pp. 399–410, October 2016. https://doi.org/10.1109/DSAA.2016.49
23. Rajabi, A., Garibay, O.O.: TabfairGAN: fair tabular data generation with generative adversarial networks. Mach. Learn. Knowl. Extr. **4**(2), 488–501 (2022)
24. Sattigeri, P., Hoffman, S.C., Chenthamarakshan, V., Varshney, K.R.: Fairness GAN: generating datasets with fairness properties using a generative adversarial network. IBM J. Res. Dev. **63**(4/5), 1–3 (2019)
25. Shannon, C.E.: A mathematical theory of communication. Bell Syst. Tech. J. **27**(3), 379–423 (1948)
26. Sharma, S., Henderson, J., Ghosh, J.: CERTIFAI: a common framework to provide explanations and analyse the fairness and robustness of black-box models. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, pp. 166–172 (2020)
27. Tan, S., Shen, Y., Zhou, B.: Improving the fairness of deep generative models without retraining. arXiv preprint arXiv:2012.04842 (2020)
28. Verma, S., Ernst, M.D., Just, R.: Removing biased data to improve fairness and accuracy. CoRR arXiv:2102.03054 (2021)
29. Vinh, N.X., Epps, J., Bailey, J.: Information theoretic measures for clusterings comparison: is a correction for chance necessary? In: Proceedings of the 26th Annual International Conference on Machine Learning, pp. 1073–1080 (2009)
30. Witten, I.H., Frank, E., Hall, M.A., Pal, C.J., DATA, M.: Practical machine learning tools and techniques. In: Data Mining, vol. 2 (2005)
31. Zemel, R., Wu, Y., Swersky, K., Pitassi, T., Dwork, C.: Learning fair representations. In: International Conference on Machine Learning, pp. 325–333. PMLR (2013)
32. Žliobaitė, I.: Measuring discrimination in algorithmic decision making. Data Min. Knowl. Disc. **31**, 1060–1089 (2017)

3.2 Mitigating Bias for Non-binary Protected Attributes with `FairDo`

Paper: Manh Khoi Duong and Stefan Conrad. Towards Fairness and Privacy: A Novel Data Pre-processing Optimization Framework for Non-binary Protected Attributes. In *Data Science and Machine Learning*, volume 1943 of *Communications in Computer and Information Science*. Springer Singapore, 2023.

Personal Contribution: Tackling the limitations of the previous work [2], Manh Khoi Duong solely developed a novel and generalized framework to account for fairness and privacy in data pre-processing. Manh Khoi Duong single-handedly implemented the framework, published it as a Python package on GitHub[1] and PyPI under the name `FairDo` [3], and conducted the experiments. All parts of the paper were written by Manh Khoi Duong. Stefan Conrad supervised the work.

Status: Published

In this paper, we addressed the limitations of our previous work [2] by extending the methodology to cover *non-binary* protected attributes. For this, a discrimination measure that handles this case was used. Furthermore, we introduced a novel optimization framework, `FairDo` [3], which is more flexible and generalizable. Due to its flexibility, we can employ any type of heuristic to solve the optimization problem, making our prior work [2] a special case of this framework. The two main optimization problems introduced in our framework involve removing and adding data points to minimize discrimination. Each serves a different purpose.

Existing works often focus on binary protected attributes. Specifically, all implemented pre-processing methods [33, 51, 25, 19] in `AIF360` [15] are only able to deal with binary protected attributes. But in reality, protected attributes often contain more than two groups, such as nationality or age groups. This is a major limitation of existing fairness libraries. To make the pre-processing methods work, `AIF360`'s sample datasets are modified to have binary protected attributes. This is done by merging multiple groups together so that only two groups remain. The two groups are then considered the privileged and unprivileged groups.

Our framework avoids these limitations by allowing for non-binary protected attributes. When a protected attribute is non-binary, measuring the discrimination becomes non-trivial. Since multiple ways to measure such discrimination exist [53], our introduced optimization framework can account for any fairness metric. We call this property *fairness-agnostic*. This is achieved by treating the optimization objective as a black-box function. This makes our framework flexible and generalizable. To solve the optimization problem, heuristics are used. In our experiments, genetic algorithms have been shown to be most effective. Our framework can also employ synthetic data for fairness and therefore handles privacy concerns as well.

[1] `https://github.com/mkduong-ai/fairdo`

Towards Fairness and Privacy: A Novel Data Pre-processing Optimization Framework for Non-binary Protected Attributes

Manh Khoi Duong(✉) and Stefan Conrad

Heinrich Heine University, Universitätsstraße 1, 40225 Düsseldorf, Germany
{manh.khoi.duong,stefan.conrad}@hhu.de

Abstract. The reason behind the unfair outcomes of AI is often rooted in biased datasets. Therefore, this work presents a framework for addressing fairness by debiasing datasets containing a (non-)binary protected attribute. The framework proposes a combinatorial optimization problem where heuristics such as genetic algorithms can be used to solve for the stated fairness objectives. The framework addresses this by finding a data subset that minimizes a certain discrimination measure. Depending on a user-defined setting, the framework enables different use cases, such as data removal, the addition of synthetic data, or exclusive use of synthetic data. The exclusive use of synthetic data in particular enhances the framework's ability to preserve privacy while optimizing for fairness. In a comprehensive evaluation, we demonstrate that under our framework, genetic algorithms can effectively yield fairer datasets compared to the original data. In contrast to prior work, the framework exhibits a high degree of flexibility as it is metric- and task-agnostic, can be applied to both binary or non-binary protected attributes, and demonstrates efficient runtime.

Keywords: Fairness · Data privacy · Non-binary · Fairness-agnostic · Genetic algorithms

1 Introduction

Machine learning has become an increasingly important tool for decision-making in various applications, ranging from income [17] to recidivism prediction [18]. However, the use of these models can perpetuate existing biases in the data and result in unfair treatment of certain demographic groups. One of the key concerns in the development of fair machine learning models is the prevention of discrimination regarding protected attributes such as race, gender, and religion.

This work was supported by the Federal Ministry of Education and Research (BMBF) under Grand No. 16DHB4020.

D. Benavides-Prado et al. (Eds.): AusDM 2023, CCIS 1943, pp. 105–120, 2024.
https://doi.org/10.1007/978-981-99-8696-5_8

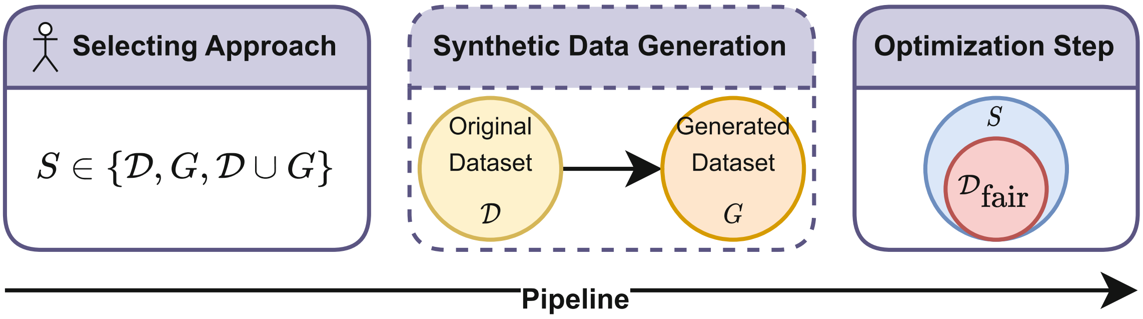

Selecting Approach
$S \in \{\mathcal{D}, G, \mathcal{D} \cup G\}$
Synthetic Data Generation
Original Dataset $\mathcal{D}$
Generated Dataset G
Optimization Step
S
$\mathcal{D}_{\text{fair}}$
Pipeline

Fig. 1. The pipeline consists of three steps: (1) The user sets the sample set S and other settings, including the objective, discrimination measure, and protected attribute; (2) Synthetic data is generated if needed; (3) A solver optimizes the fairness objective to yield a biased-reduced subset $\mathcal{D}_{\text{fair}}$ from the user-selected set S. If $S = G$ was chosen, the user obtains a bias-reduced synthetic dataset that does not leak privacy-related information.

While most of the existing literature focuses on classification problems where the protected attribute is binary [2,4,6,7,10,20,24,28], the real world presents a more complex scenario where the protected attribute can consist of more than two social groups, making it non-binary. While works that discuss and deal with non-binary protected attributes exist, and we do not neglect their existence [5, 14,29], we view it as a necessity to contribute further to this field by providing a flexible framework that accommodates various fairness notions and applications, including data privacy, to strive for the employment of responsible artificial intelligence in practice.

Since bias is rooted in data, we introduce an optimization framework that pre-processes data to mitigate discrimination. In the context of fairness, pre-processing ensures the generation of a fair, debiased dataset. We address the challenges associated with non-binary protected attributes by deriving appropriate discrimination measures. To prevent discrimination, we formulate a combinatorial optimization problem to identify a subset from a given sample dataset that minimizes a specific discrimination measure, as depicted in Fig. 1. Depending on the provided sample dataset, which may also include synthetically generated data, the framework allows for the removal of such data points or the inclusion of synthetic ones to achieve equitable outcomes. By using generated data, we can utilize our method in applications where data privacy is a concern. Since the discrimination objective is stated as a black box, heuristics, which do not assess the analytical expression of the discrimination measure during optimization, are needed to solve our stated problem. Our formulation makes the framework *fairness-agnostic*, allowing it to be used to pursue any fairness objective.

The experimentation was carried out on the Adult [17], Bank [22], and COMPAS [18] datasets, all known to exhibit discrimination. We compared the discrimination of the datasets before and after pre-processing them with different heuristics on various discrimination measures. The results show that genetic algorithms [12] were most effective in reducing discrimination for non-binary protected attributes. To summarize, the primary contributions of this paper are:

- We present an optimization framework that renders different approaches for yielding fair data. The approaches include removing, adding generated data, or solely using generated data.
- We underscore the framework's ability to handle cases where data privacy is a significant concern.
- Our methodology is designed to handle a protected attribute that can be non-binary, offering broader applicability.
- We carry out an extensive evaluation of the proposed techniques on three biased datasets. The evaluation focuses on their effectiveness in reducing discrimination and their runtimes.
- We publish our implementation at https://github.com/mkduong-ai/fairdo as a documented Python package and distribute it over PyPI.

2 Related Work

Recently, related works have equivalently formulated subset selection problems to achieve fairness goals [7,26]. While in the work of Tang et al. [26], a distribution is generated that represents the selection probability of each feasible set to maximize the global utility on average, our work aims to return a definite subset. To achieve fairness according to any defined criteria, our formulation treats discrimination measures as black boxes. These measures can encompass both group and individual fairness notions, distinguishing our work from that of Tang et al. [26], whose framework is limited to group fairness.

Previous studies have also utilized synthetic data to address fairness and privacy concerns [7,19]. Both of these studies employed heuristics similar to our approach. In particular, Liu et al. [19] specialized on generating synthetic data using a genetic algorithm to satisfy specific privacy definitions [3,8]. While our framework does not generate privacy-preserving data specifically, it utilizes synthetic data, which can be generated with such methods. Similarly to our work, Duong et al. [7] leveraged synthetic data by introducing a sampling-based heuristic for selecting a subset of such data points to minimize discrimination. Our work generalizes the work of Duong et al. [7] as their approach can be viewed as a special case of ours. Additionally, our formulation offers greater flexibility compared to the approach of Duong et al. [7], as it allows for any heuristic to tackle the task and is also not limited to binary protected attributes.

3 Measuring Discrimination

In this section, we introduce the notation used to derive discrimination measures for assessing dataset fairness: A *data point* or *sample* is represented as a triple (x, y, z), where $x \in X$ is the *feature*, $y \in Y$ is the ground truth *label* indicating favorable or unfavorable outcomes, and $z \in Z$ is the *protected attribute*, which is used to differentiate between groups. The sets X, Y, Z typically hold numeric values and are defined as $X = \mathbb{R}^d$, $Y = \{0, 1\}$, and $Z = \{1, 2, \ldots, k\}$ with $k \geq 2$. For instance, in the context of predicting personal attributes, we can use X to represent numeric values that encode particular aspects of a person. Y typically describes the positive or negative outcome that we aim to predict for the person. Z can denote any protected attribute, such as race, which can be used to identify the person as Caucasian, Afro-American, Latin American, or Asian. We assume that z is not included as a feature in x. To be able to differentiate between groups, $k \geq 2$ must hold. If $k > 2$, the protected attribute Z is said to be non-binary. Following the definition, a *dataset*, denoted as $\mathcal{D} = \{d_i\}_{i=1}^n$, consists of data points, where a single sample is defined as $d_i = (x_i, y_i, z_i)$. Machine learning models are trained using these datasets to predict the target variable y based on the input variables x and z. Finally, we denote a discrimination measure with $\psi \colon \mathbb{D} \rightarrow [0, 1]$, where $\mathbb{D}$ is the set of all datasets.

In the following, x, y, z are noted as random variables that can take on specific values.

3.1 Absolute Measures

To deal with non-binary groups, Žliobaitė [29] suggested in her work to compare groups pairwise. For this, she presented three possible ways which are comparing each group with another, one against the rest for each group, and all groups against the unprivileged group. The author further discussed options to aggregate the results. Although Žliobaitė [29] stated textually how to measure discrimination for more than two groups, we express them mathematically in this work. To treat groups equally without presuming which group is unprivileged and to get the full picture, we choose to make use of comparing each group with another. We first introduce the common fairness notion *statistical parity* [16,28], which demands equal positive outcomes for different groups in $Z = \{1, 2, \ldots, k\}$. It is usually defined for binary groups, but we present the non-binary cases [29].

Definition 1 (Statistical parity). *Demanding that each of the k groups have the same probability of receiving the favorable outcome is statistical parity, i.e.,*

$$
\begin{aligned}
& P(y = 1 \mid z = 1) = \ldots = P(y = 1 \mid z = k) \\
\iff \quad & P(y = 1 \mid z = i) = P(y = 1 \mid z = j) \quad \forall i, j \in Z.
\end{aligned}
$$

As the group size k grows, the satisfaction of statistical parity becomes less probable. Because of this, the equality constraints are treated softly by deriving differences between the groups. Consequently, smaller differences imply more equality. For binary groups, the difference is often referred to as statistical disparity (SDP) [6].

Definition 2 (Sum of absolute statistical disparities). *Let there be k groups, then the sum of absolute statistical disparities is calculated as follows [29]:*

$$\begin{aligned}\psi_{SDP\text{-}sum}(\mathcal{D}) &= \sum_{\substack{i,j \in Z\\ i \neq j}} |P(y=1 \mid z=i) - P(y=1 \mid z=j)| \\ &= \sum_{i=1}^{k} \sum_{j=i+1}^{k} |P(y=1 \mid z=i) - P(y=1 \mid z=j)|.\end{aligned}$$

Because the total number of comparisons is $\frac{k(k-1)}{2}$ [29], the average discrimination between all groups becomes:

$$\begin{aligned}\psi_{SDP\text{-}avg}(\mathcal{D}) = \frac{2}{k(k-1)} \cdot \sum_{i=1}^{k} \sum_{j=i+1}^{k} |P(y=1 \mid z=i) \\ - P(y=1 \mid z=j)|.\end{aligned}$$

Definition 3 (Maximal absolute statistical disparity). *Maximal absolute statistical disparity measures the absolute statistical disparity between all pairs $i, j \in Z$ and returns the maximum value. Specifically, it is given by:*

$$\begin{aligned}\psi_{SDP\text{-}max}(\mathcal{D}) = \max_{i,j \in Z} |P(y=1 \mid z=i) \\ - P(y=1 \mid z=j)|.\end{aligned}$$

Žliobaitė [29], after consulting with legal experts, recommends using the maximum function to aggregate disparities, though the choice depends on the ethical context of the specific use case. Discrimination measures can be seen as social welfare functions. Minimizing the sum of absolute statistical disparities is analogous to the utilitarian viewpoint [21], which aims to maximize the general utility of the population. If one decides to care for the least well-off group, then minimizing the maximal absolute statistical disparity corresponds to the Rawlsian social welfare [25].

4 Optimization Framework

Inspired by related works that identify unfair data samples [15,27], we propose a method to remove such samples for fairness. The task is formulated as a combinatorial problem where the aim is to determine a subset $\mathcal{D}_{\text{fair}}$ of a given set S such that the discrimination of the subset $\psi(\mathcal{D}_{\text{fair}})$ is minimal, as shown in Fig. 1. Depending on the application, set S can be the original data $\mathcal{D}$, a synthetic set G with the same distribution as $\mathcal{D}$, or their union $\mathcal{D} \cup G$.

4.1 Problem Formulation

To state the problem mathematically, let note $S = \{s_1, s_2, \ldots, s_{\tilde{n}}\}$ and further introduce a binary vector b with the same length as S, i.e., $b = (b_1, b_2, \ldots, b_{\tilde{n}})$. To define the combinatorial optimization problem, each entry b_i in b is examined whether it is 1 ($b_i = 1$), in which case the corresponding sample s_i in S is included in the subset $\mathcal{D}_{\text{fair}}$. Therefore, the fair set is defined with

$$\mathcal{D}_{\text{fair}} = \{s_i \in S \mid b_i = 1, i = 1 \ldots \tilde{n}\}. \tag{1}$$

The objective $f\colon 0, 1^{\tilde{n}} \to [0, 1]$ can then be expressed by:

$$\begin{aligned} & f_{S,\psi}(b) = \psi(\mathcal{D}_{\text{fair}}) \\ \iff \quad & f_{S,\psi}(b) = \psi(\{s_i \in S \mid b_i = 1, i = 1 \ldots \tilde{n}\}), \end{aligned} \tag{2}$$

where $f_{S,\psi}$ is defined as the discrimination of a subset $\mathcal{D}_{\text{fair}}$ of the given set S and ψ evaluates the level of discrimination on $\mathcal{D}_{\text{fair}}$. Note that the decision variable is b, for which $\mathcal{D}_{\text{fair}}$ can be obtained. The subindices S and ψ of $f_{S,\psi}$ can be seen as settings for the objective. Ignoring the subindices, we write out the combinatorial optimization problem as follows:

$$\begin{aligned} \min_{b} \quad & f(b) \\ \text{subject to} \quad & b_i \in \{0, 1\} \quad \forall i = 1, \ldots, \tilde{n}. \end{aligned} \tag{3}$$

Because the set of feasible subsets $\mathcal{P}(S)$ grows exponentially regarding the cardinality of S, we employ heuristics to solve our stated problem.

In the following subsections, we discuss different and useful settings of S that serve different purposes with their corresponding advantages and disadvantages.

4.2 Removing Samples ($S = \mathcal{D}$)

By setting $S = \mathcal{D}$, it is intended to determine data points in the training set that can be removed to prevent discrimination. Intuitively, having an overexposure of certain types of samples that fulfill stereotypes can result in a discriminatory dataset. In such situations, the most practical step is to remove the affected samples.

However, this method is not recommended if the given dataset is small. Likewise, some could argue that minorities can be easily removed by this method. Luckily, this can be prevented by choosing the right discrimination measure.

4.3 Employing only Synthetic Data ($S = G$)

To employ synthetic data, this method relies on a statistical model. The statistical model is used to learn the distribution of the original data $P(\mathcal{D})$. By doing so, synthetic samples G can be drawn from the learned distribution $G \sim P(\mathcal{D})$.

Relying solely on synthetic data is particularly important in use cases where data privacy and protection are major concerns and the use of real data is prohibited. Of course, synthetic data is not necessarily disjoint from the original dataset and can therefore be a privacy breach itself. For tabular and smaller datasets, this can be naively mitigated by removing such privacy breaching points from the synthetic data by setting $S = G \setminus \mathcal{D}$. Other ways include populating differential privacy techniques in the data generation process [1,8,13,19].

When generally using synthetic data, one cannot easily ensure that the corresponding label of the features is correct. Training machine learning models on synthetic data can therefore lead to higher error rates when predicting on real data. Despite the distribution of the synthetic data following the distribution of the real dataset, it depends heavily on the method used when it comes to generating qualitative, faithful data.

4.4 Merging Real and Synthetic Data ($S = \mathcal{D} \cup G$)

Another approach to generate a non-discriminatory dataset is to merge the original dataset $\mathcal{D}$ with synthetic data G that has been generated with a statistical model as described in Sect. 4.3. By combining the two sets $S = \mathcal{D} \cup G$, it is possible to increase the size of the resulting dataset while avoiding over-representation of discriminatory samples.

One advantage of this method is that it can improve the quality of the data by utilizing both the real $\mathcal{D}$ and synthetic data G. The resulting dataset can be larger and more diverse, which can lead to greater robustness when training machine learning models. If the dataset is too small to apply removal techniques ($S = \mathcal{D}$) or relying solely on synthetic data ($S = G$) appears unreliable, merging the two sets may be a viable option.

However, this method is not without its limitations and comes with disadvantages when generally using synthetic data, e.g., quality and faithfulness. Different from the method described in Sect. 4.3, this method is not applicable for purposes with privacy concerns as samples from the real data are not omitted.

4.5 Adding Synthetic Data

A different approach that requires a new formulation of the objective is to include synthetic data points without deleting any samples from the real data. As well, a set of generated data points G must be given, and the research question is which of the generated points can lead to a fairer distribution when including them in the original dataset. The possible use case for this problem is to fine-tune machine learning models that have already learned from an unfair dataset. This is mostly useful for large machine learning models where resources are scarce to retrain the whole model. Following the preceding notation, the fair dataset becomes:

$$\mathcal{D}_{\text{fair}}^{\text{add}} = \mathcal{D} \cup \{s_i \in S \mid b_i = 1, i = 1 \dots \tilde{n}\} \tag{4}$$

and we express the corresponding objective $f_{S,\psi}^{\text{add}}$ by:

$$\begin{aligned} & f_{S,\psi}^{\text{add}}(b) = \psi(\mathcal{D}_{\text{fair}}^{\text{add}}) \\ \iff \quad & f_{S,\psi}^{\text{add}}(b) = \psi(\mathcal{D} \cup \{s_i \in S \mid b_i = 1, i = 1 \dots \tilde{n}\}), \end{aligned} \tag{5}$$

where S is set to G to achieve the described approach. Certainly, S can also be set to $\mathcal{D}$ or any other set operation on $\mathcal{D}$ with G. Although such settings are possible, they do not serve any meaningful purposes. However, one could argue that setting $S = \mathcal{D}$ can act as a reweighing method. Still, we argue against facilitating duplicates in a dataset with intent, as no additional information is provided.

As seen, our framework offers many advantages due to its versatility and therefore potential use in a broad range of applications. By choosing the appropriate objective function, discrimination measure, and sample set, the formulation is tailored to the specific intent and use case. Because the formulation is agnostic to the solver, it can serve multiple purposes without modifying solvers.

Table 1. Overview of Datasets

Dataset	Entries	Cols.	Label	Protected Attribute	Description
Adult [17]	32 561	22	Income	Race: White, Black, Asian-Pacific-Islander, American-Indian-Eskimo, Other	Indicates individuals earning over $50,000 annually
Bank [22]	41 188	53	Term deposit subscription	Job: Admin, Blue-Collar, Technician, Services, Management, Retired, Entrepreneur, Self-Employed, Housemaid, Unemployed, Student, Unknown	Shows whether the client has subscribed to a term deposit.
COMPAS [18]	7 214	8	2-year recidivism	Race: African-American, Caucasian, Hispanic, Other, Asian, Native American	Displays individuals that were rearrested for a new crime within 2 years after initial arrest

5 Heuristics

This section presents heuristics that specifically solve combinatorial optimization problems. These include: a baseline method that returns the original dataset, a simple random heuristic, and genetic algorithms with different operators.

1. **Original**: Uses the original data by returning a vector of ones $b = \mathbf{1}_{\tilde{n}}$.
2. **Random Heuristic**: Generates a user-defined number of random vectors, with each entry having a 50% chance of being zero or one, and then returns the best solution.
3. **Genetic Algorithm (GA)**: The workflow of GAs [9] involves generating an initial population of candidate solutions and then repeatedly performing *selection*, *crossover*, and *mutation* operations over several generations. In our implementation, the GA terminates earlier if improved solutions were not found within 50 generations. Following operators were used in our experimentation [11]:
   - Selection: *Elitist*, *Tournament*, *Roulette Wheel* (see [11] for more details)
   - Crossover: *Uniform* (each entry of the offspring has the same probability of either inheriting the entry from the first or second parent)
   - Mutation: *Bit Flip* (flips a fixed amount of random bits for each vector, that is $\lfloor p_m \cdot \tilde{n} \rfloor$, where $p_m \in [0, 1]$ is the mutation rate)

6 Evaluation

In our evaluation, we conducted multiple experiments to address the following research questions:

- **RQ1** How do the heuristics perform in making the datasets fairer?
- **RQ2** How does runtime vary among the heuristics?
- **RQ3** How stable are the results across the runs?
- **RQ4** Is there a clear winner? If yes, which method is recommended for practical use?

To answer these research questions, we specifically designed experiments for the Adult [17], Bank [22], and COMPAS [18] datasets. Both the Adult and COMPAS datasets include race as a non-binary protected attribute, whereas the Bank dataset utilizes the job as a non-binary protected attribute. All datasets were prepared and cleansed in the same manner: Categorical features were one-hot encoded, with the exception of the protected attribute and the label. Additionally, rows containing missing values were excluded from all datasets. Table 1 shows details about the datasets used in our experiments after the preparation and cleansing steps.

Following the dataset preparation, we executed two distinct experiments. The first experiment (Sect. 6.1) was dedicated to hyperparameter tuning of the GAs, adjusting both population sizes and the number of generations to pinpoint optimal configurations. Armed with these optimal settings, our second experiment (Sect. 6.2) focused on comparing different selection operators within GAs (**RQ1**). Our aim was to determine which operator yielded the best performance. This experiment included comparisons to several baseline methods, one of which simply returned the original data. By expanding our evaluation to multiple discrimination measures in this phase, we can comprehensively assess the effectiveness of GAs in reducing discrimination in datasets.

The experimental methodology involves the application of heuristics to produce a binary mask, which yields fair data. We then measure the discrimination of the resulting dataset. To ensure stability in our findings (**RQ3**), each experiment was repeated 15 times. We additionally recorded the runtime of each trial to tackle **RQ2**. Depending on the experiment, we employed suitable heuristics that aim to solve each objective with the associated discrimination measure, as listed in Table 2. For instance, each heuristic either optimizes $f_{S,\psi}$ or $f^{\text{add}}_{S,\psi}$ with varying settings of S and ψ as given in the table. In order to perform experiments with synthetic data, we generated data that has the same size as the original dataset, i.e., $|G| = |\mathcal{D}|$. The statistical model used to generate synthetic data is Gaussian copula [23] which is fast and performs well on tabular data. For privacy-sensitive use cases, we advise utilizing privacy-preserving techniques [1,8,13,19]. All experiments were conducted on an Intel(R) Xeon(R) Gold 5120 processor clocking at 2.20 GHz.

Table 2. Configuration details of heuristics, objectives, and discrimination measures for each experiment.

Experiment	Heuristics	Objectives (f, S)	Disc. Measures (ψ)
Hyperparam	GA	Remove, Merge, Add	Sum SDP
Comparison	Original, Random, GA (Elitist, Tournament, Roulette Wheel)	Remove, Merge, Add	Sum SDP, Max SDP

6.1 Hyperparameter Tuning

For the genetic algorithm, we performed hyperparameter tuning, exploring various population sizes [20, 50, 100, 200] and generations [50, 100, 200, 500], all using tournament selection, uniform crossover, and bit flip mutation at a rate of 5%. These configurations are described in Sect. 5. We evaluated the algorithm on three distinct objectives and set $\psi_{\text{SDP-sum}}$ as the discrimination measure.

Discrimination. As seen in Fig. 2, the heatmaps display the average discrimination (including the standard deviation) of GAs solving various objectives on different datasets. Each heatmap shows hyperparameters that were set for the experimentation. Across the different objectives and datasets, there is a consistent trend indicating that utilizing larger populations combined with a higher number of generations typically results in less discrimination. This is particularly evident when contrasting scenarios with a population size of 20 and 50 generations, which, on average, have discrimination scores higher by 0.1. However, the improvements in discrimination plateau beyond certain thresholds. Specifically, once the number of generations surpasses 200 or when the population size exceeds 100, there is no significant further decrease in discrimination observable.

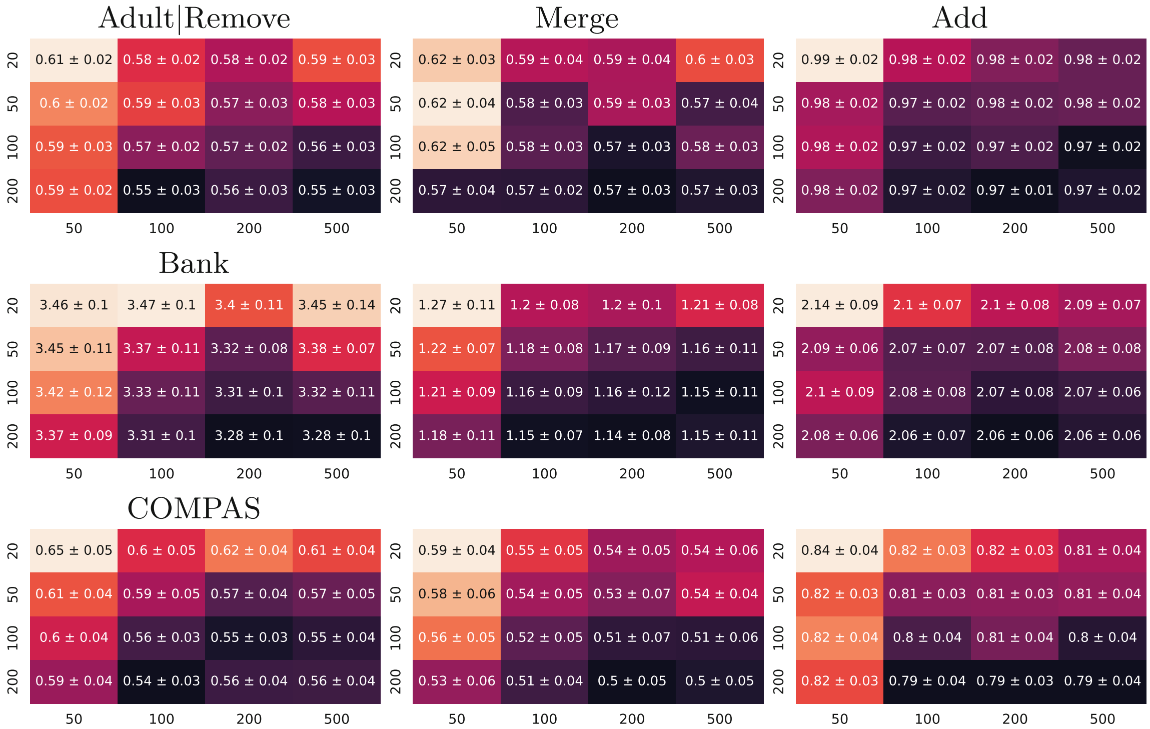


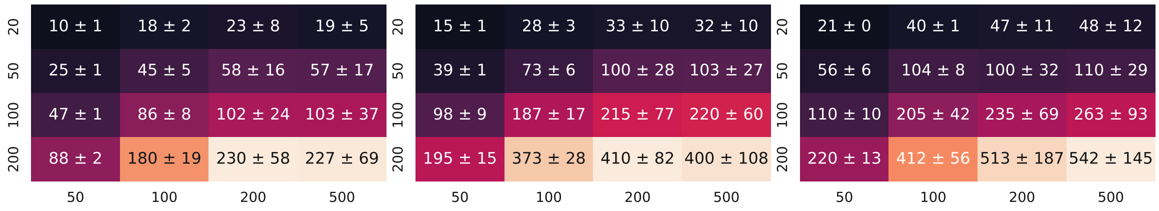
Fig. 2. Heatmaps showing discrimination scores ($\psi_{\text{SDP-sum}}$) after pre-processing with genetic algorithms using different population sizes (y-axis) and generations (x-axis). Rows depict the results of Adult, Bank, and COMPAS datasets, while columns represent the objectives.

Runtime. For brevity reasons, we display the runtimes solely for the Bank dataset in Fig. 3, given its larger size and the similarity of the results across other datasets. The outcome of this analysis pointed towards an optimal setting of a population size of 100 combined with 500 generations. Under our specifications, executing the GA with these settings takes, on average, between 1.5 and 4.5 min. While increasing the population size further did not show significant improvements in reducing the bias in the datasets, it proved to be more efficient in terms of the runtime.

6.2 Comparing Heuristics

After determining that a population size of 100 with 500 generations offered optimal results w.r.t. discrimination and time, this configuration was maintained for all subsequent experiments. Here, three GAs were compared, each differing by their selection operator: elitist, tournament, and roulette wheel selection. All GAs were set with uniform crossover and bit flip mutation at a rate of 5% to perform the experiments. Additionally, we established both the original dataset and the random heuristic as baselines.

Discrimination. Table 3 presents the discrimination results of our experiments. It is evident that all tested algorithms are stable, as reflected by the low standard



(a) Remove (b) Merge (c) Add

Fig. 3. Heatmaps showing runtimes in seconds for the Bank dataset after preprocessing with genetic algorithms using different population sizes (y-axis) and generations (x-axis).

Table 3. Displayed are the mean discrimination scores, accompanied by standard deviations, from 15 runs. The heuristics were evaluated across multiple objectives using varying discrimination measures on the Adult, Bank, and COMPAS datasets. Best results are marked bold.

Objective	Method	Sum SDP			Max SDP		
		Adult	Bank	COMPAS	Adult	Bank	COMPAS
Add	1. Original	1.07 ± 0.02	1.83 ± 0.09	1.17 ± 0.06	0.23 ± 0.00	0.09 ± 0.00	0.17 ± 0.01
	2. Random	1.03 ± 0.02	2.27 ± 0.07	0.94 ± 0.03	0.21 ± 0.00	0.11 ± 0.00	0.15 ± 0.01
	3. Elitist	**0.82 ± 0.02**	**1.54 ± 0.06**	**0.59 ± 0.03**	**0.16 ± 0.00**	**0.07 ± 0.00**	**0.10 ± 0.00**
	4. Tournament	0.97 ± 0.02	2.06 ± 0.06	0.80 ± 0.03	0.20 ± 0.00	0.10 ± 0.00	0.13 ± 0.00
	5. Roulette	1.03 ± 0.02	2.31 ± 0.08	0.94 ± 0.05	0.21 ± 0.00	0.11 ± 0.00	0.15 ± 0.01
Merge	1. Original	1.07 ± 0.02	1.83 ± 0.09	1.17 ± 0.06	0.23 ± 0.00	0.09 ± 0.00	0.17 ± 0.01
	2. Random	0.80 ± 0.03	1.46 ± 0.09	0.76 ± 0.08	0.16 ± 0.01	0.07 ± 0.00	0.12 ± 0.01
	3. Elitist	**0.21 ± 0.04**	**0.42 ± 0.07**	**0.11 ± 0.05**	**0.04 ± 0.01**	**0.02 ± 0.00**	**0.01 ± 0.00**
	4. Tournament	0.58 ± 0.04	1.17 ± 0.09	0.51 ± 0.04	0.11 ± 0.01	0.05 ± 0.00	0.09 ± 0.01
	5. Roulette	0.85 ± 0.05	1.49 ± 0.09	0.79 ± 0.09	0.16 ± 0.01	0.07 ± 0.00	0.12 ± 0.01
Remove	1. Original	0.97 ± 0.00	4.81 ± 0.00	1.89 ± 0.00	0.17 ± 0.00	0.25 ± 0.00	0.27 ± 0.00
	2. Random	0.71 ± 0.02	4.07 ± 0.07	0.72 ± 0.03	0.12 ± 0.00	0.19 ± 0.00	0.12 ± 0.01
	3. Elitist	**0.25 ± 0.02**	**1.41 ± 0.12**	**0.20 ± 0.07**	**0.05 ± 0.00**	**0.07 ± 0.01**	**0.01 ± 0.00**
	4. Tournament	0.57 ± 0.02	3.29 ± 0.08	0.56 ± 0.04	0.11 ± 0.00	0.15 ± 0.01	0.09 ± 0.01
	5. Roulette	0.75 ± 0.03	4.15 ± 0.10	0.75 ± 0.08	0.13 ± 0.00	0.20 ± 0.01	0.12 ± 0.01

deviations (**RQ3**). All heuristics were able to reduce the discrimination available in the datasets in most cases. Elitist selection consistently outperformed other methods, offering notable improvements in fairness compared to the original datasets (**RQ1**). We emphasize that the measures handle non-binary attributes, providing flexibility in targeting various fairness goals. Further, by the range of discrimination measures utilized, our methodology can aim for diverse fairness goals, be it the enhancement of the utilitarian social welfare ($\psi_{\text{SDP-sum}}$) or Rawlsian social welfare ($\psi_{\text{SDP-max}}$), as evidenced. An interesting observation from our study is the varied discrimination levels based on the specific measure used, as seen in the Bank dataset, where its discrimination is either highest or lowest when compared with other datasets. This is due to the higher number of groups, leading to more group comparisons that affect the overall discrimination score. When examining the objectives, removing both the synthetic and original data tends to outperform others. This observation is particularly evident in the

Merge objective. Given the consistent performance of the elitist selection in our tests, we strongly recommend its use for those aiming to achieve the best fairness outcomes (**RQ4**).

Table 4. Mean runtimes in seconds of different methods solving different objectives with varying discrimination measures on the Adult, Bank, and COMPAS datasets.

Objective	Method	Sum SDP			Max SDP		
		Adult	Bank	COMPAS	Adult	Bank	COMPAS
Add	1. Original	**0 ± 0**	**0 ± 0**	**0 ± 0**	**0 ± 0**	**0 ± 0**	**0 ± 0**
	2. Random	50 ± 1	107 ± 12	14 ± 0	51 ± 6	103 ± 7	13 ± 0
	3. Elitist	320 ± 105	605 ± 224	53 ± 21	334 ± 80	636 ± 179	79 ± 23
	4. Tournament	122 ± 38	209 ± 50	39 ± 17	119 ± 37	216 ± 74	34 ± 12
	5. Roulette	82 ± 26	131 ± 46	26 ± 9	82 ± 40	132 ± 48	26 ± 12
Merge	1. Original	**0 ± 0**	**0 ± 0**	**0 ± 0**	**0 ± 0**	**0 ± 0**	**0 ± 0**
	2. Random	46 ± 3	67 ± 1	15 ± 2	44 ± 4	66 ± 1	15 ± 3
	3. Elitist	283 ± 103	359 ± 143	79 ± 25	286 ± 111	397 ± 161	75 ± 28
	4. Tournament	127 ± 39	185 ± 69	36 ± 11	131 ± 61	169 ± 51	44 ± 19
	5. Roulette	69 ± 21	127 ± 53	28 ± 9	83 ± 33	118 ± 31	29 ± 14
Remove	1. Original	**0 ± 0**	**0 ± 0**	**0 ± 0**	**0 ± 0**	**0 ± 0**	**0 ± 0**
	2. Random	23 ± 1	44 ± 1	11 ± 0	22 ± 1	47 ± 11	11 ± 0
	3. Elitist	138 ± 66	281 ± 119	52 ± 18	176 ± 68	290 ± 80	50 ± 15
	4. Tournament	78 ± 13	119 ± 25	24 ± 9	73 ± 27	132 ± 46	25 ± 7
	5. Roulette	58 ± 27	72 ± 19	22 ± 8	52 ± 24	71 ± 24	18 ± 7

Runtime. An analysis of the runtimes is presented in Table 4. The original method consistently took 0 s (rounded) to finish. At second comes the random method and lastly GAs. The elitist operator took the longest, with runtimes approximately three times slower than the quickest operator, the roulette wheel. Tournament selection comes in between. Most experiments were finished in 5 min or less, which is still very efficient. Regarding the measures, the runtimes when optimizing $\psi_{\text{SDP-max}}$ appeared negligibly higher compared to $\psi_{\text{SDP-sum}}$, so it can be disregarded. Generally, larger datasets yielded longer runtimes, revealing a linear relationship between dataset size and runtime. In addressing the research question posed in **RQ4**, it becomes evident that the elitist operator is superior among the tested methods. Despite being the slowest method, it is still very efficient at reducing discrimination on datasets consisting of up to 41 188 samples, as seen in our experimentation.

7 Conclusion

We introduced a novel and flexible optimization framework to reduce discrimination and preserve privacy in datasets. The framework accommodates various intents such as data removal, synthetic data addition, and exclusive use of synthetic data for privacy reasons. Notably, the objectives in our framework are designed to be independent of specific discrimination measures, allowing users and stakeholders to address any form of discrimination without modifying the solvers.

Due to the relatively sparse work existing on dealing with non-binary attributes, particularly regarding established methods, we tackled non-binary protected attributes in our experiments by deriving discrimination measures based on the work of Žliobaitė [29] and showed that our framework allowed the effective and fast reduction of discrimination by employing heuristics.

8 Future Work and Discussion

Future work could include extending the usability of this framework by deriving different discrimination measurements. Thus, handling multiple protected attributes as well as regression tasks can be done without modifying the general methodology. Additionally, formulating and integrating constraints into the objective function can also be done, which further enhances the responsibility of our approach. For instance, we could consider constraints such as group sizes and add penalties if samples of minorities get removed.

Although we aim for fairness and data privacy with our framework, it is still important to engage with diverse stakeholders to identify unintended consequences and address possible ethical implications. Particularly, an extensive discussion and analysis of the used objective and discrimination measure for a specific application should be done to ensure that the data aligns with the desired fairness goals.

References

1. Abay, N.C., Zhou, Y., Kantarcioglu, M., Thuraisingham, B., Sweeney, L.: Privacy preserving synthetic data release using deep learning. In: Berlingerio, M., Bonchi, F., Gärtner, T., Hurley, N., Ifrim, G. (eds.) ECML PKDD 2018. LNCS (LNAI), vol. 11051, pp. 510–526. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-10925-7_31
2. Barocas, S., Hardt, M., Narayanan, A.: Fairness and machine learning. fairmlbook.org (2019). http://www.fairmlbook.org
3. Bun, M., Steinke, T.: Concentrated differential privacy: simplifications, extensions, and lower bounds. In: Hirt, M., Smith, A. (eds.) TCC 2016. LNCS, vol. 9985, pp. 635–658. Springer, Heidelberg (2016). https://doi.org/10.1007/978-3-662-53641-4_24
4. Caton, S., Haas, C.: Fairness in machine learning: a survey. arXiv preprint arXiv:2010.04053 (2020)
5. Celis, L.E., Huang, L., Keswani, V., Vishnoi, N.K.: Fair classification with noisy protected attributes: a framework with provable guarantees. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning. Proceedings of Machine Learning Research, 18–24 July 2021, vol. 139, pp. 1349–1361. PMLR (2021). https://proceedings.mlr.press/v139/celis21a.html

6. Dunkelau, J., Leuschel, M.: Fairness-aware machine learning (2019)
7. Duong, M.K., Conrad, S.: Dealing with data bias in classification: can generated data ensure representation and fairness? In: Wrembel, R., Gamper, J., Kotsis, G., Tjoa, A.M., Khalil, I. (eds.) Big Data Analytics and Knowledge Discovery, DaWaK 2023. LNCS, vol. 14148, pp. 176–190. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-39831-5_17
8. Dwork, C.: Differential privacy. In: Bugliesi, M., Preneel, B., Sassone, V., Wegener, I. (eds.) ICALP 2006. LNCS, vol. 4052, pp. 1–12. Springer, Heidelberg (2006). https://doi.org/10.1007/11787006_1
9. Eiben, A.E., Smith, J.E.: Introduction to Evolutionary Computing. NCS, Springer, Heidelberg (2015). https://doi.org/10.1007/978-3-662-44874-8
10. Friedrich, F., et al.: Fair diffusion: instructing text-to-image generation models on fairness. arXiv preprint at arXiv:2302.10893 (2023)
11. Goldberg, D.E.: Genetic Algorithms in Search, Optimization and Machine Learning, 1st edn. Addison-Wesley Longman Publishing Co., Inc, USA (1989)
12. Holland, J.: Adaptation in Natural and Artificial Systems (1975)
13. Jordon, J., Yoon, J., Van Der Schaar, M.: PATE-GAN: generating synthetic data with differential privacy guarantees. In: International Conference on Learning Representations (2019)
14. Kamani, M.M., Haddadpour, F., Forsati, R., Mahdavi, M.: Efficient fair principal component analysis. Mach. Learn. **111**, 3671–3702 (2022). https://doi.org/10.1007/s10994-021-06100-9
15. Kamiran, F., Calders, T.: Data preprocessing techniques for classification without discrimination. Knowl. Inf. Syst. **33**(1), 1–33 (2012)
16. Kamishima, T., Akaho, S., Asoh, H., Sakuma, J.: Fairness-aware classifier with prejudice remover regularizer. In: Flach, P.A., De Bie, T., Cristianini, N. (eds.) ECML PKDD 2012. LNCS (LNAI), vol. 7524, pp. 35–50. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33486-3_3
17. Kohavi, R.: Scaling up the accuracy of Naive-Bayes classifiers: a decision-tree hybrid. In: KDD 1996, pp. 202–207. AAAI Press (1996)
18. Larson, J., Angwin, J., Mattu, S., Kirchner, L.: Machine bias, May 2016. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
19. Liu, T., Tang, J., Vietri, G., Wu, S.: Generating private synthetic data with genetic algorithms. In: International Conference on Machine Learning, pp. 22009–22027. PMLR (2023)
20. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. ACM Comput. Surv. (CSUR) **54**(6), 1–35 (2021)
21. Mill, J.S.: Utilitarianism. Parker, Son, and Bourn (1863)
22. Moro, S., Cortez, P., Rita, P.: A data-driven approach to predict the success of bank telemarketing. Decis. Support Syst. **62**, 22–31 (2014)
23. Patki, N., Wedge, R., Veeramachaneni, K.: The synthetic data vault. In: 2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA), October 2016, pp. 399–410 (2016). https://doi.org/10.1109/DSAA.2016.49
24. Prost, F., Qian, H., Chen, Q., Chi, E.H., Chen, J., Beutel, A.: Toward a better trade-off between performance and fairness with kernel-based distribution matching. CoRR abs/1910.11779 (2019). http://arxiv.org/abs/1910.11779
25. Rawls, J.: A Theory of Justice. Belknap Press (1971)
26. Tang, S., Yuan, J.: Beyond submodularity: a unified framework of randomized set selection with group fairness constraints. J. Comb. Optim. **45**(4), 102 (2023)

27. Verma, S., Ernst, M.D., Just, R.: Removing biased data to improve fairness and accuracy. CoRR abs/2102.03054 (2021). https://arxiv.org/abs/2102.03054
28. Zemel, R., Wu, Y., Swersky, K., Pitassi, T., Dwork, C.: Learning fair representations. In: International Conference on Machine Learning, pp. 325–333. PMLR (2013)
29. Žliobaitė, I.: Measuring discrimination in algorithmic decision making. Data Min. Knowl. Disc. **31**(4), 1060–1089 (2017). https://doi.org/10.1007/s10618-017-0506-1

3.3 Dealing with Bias in Datasets with Multiple Protected Attributes

Paper: Manh Khoi Duong and Stefan Conrad. Measuring and Mitigating Bias for Tabular Datasets with Multiple Protected Attributes. In *Proceedings of the 2nd Workshop on Fairness and Bias in AI co-located with 27th European Conference on Artificial Intelligence (ECAI 2024), Santiago de Compostela, Spain, October 20th, 2024*, volume 3808 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2024.

Personal Contribution: Manh Khoi Duong developed and conducted the whole research to fill the literature gap. Manh Khoi Duong wrote all parts of the paper. Stefan Conrad supervised the work.

Status: Published

In our earlier work, we introduced a fairness-agnostic pre-processing framework [3]. We showed that our framework can be used to mitigate bias in datasets with a single protected attribute, even if it is non-binary, proving the superiority of our framework over existing methods. Since it is common for datasets to have multiple protected attributes and our framework should be able to handle them, we aim to experiment on this problem with this paper. From the promises of our framework, we expect it to work seamlessly as long as the proper fairness metric is defined.

The following paper contributes in mainly two aspects [4]. Firstly, we found that there is no established notion to cover the fairness of datasets with multiple protected attributes. Further, choosing a fairness metric for datasets generally can become overwhelming, especially when multiple attributes come into play. Hence, we introduce a notation that covers multiple protected attributes and provide a guideline for choosing the right fairness metric depending on specific cases. Secondly, we propose a new fairness metric, opposed to the existing ones, that can be used to measure the fairness of datasets with multiple protected attributes but is less prone to overestimating bias. Related works often use an *intersectional* approach, which leads to exponentially more groups to consider. When aggregating the results, the bias can be overestimated [40]. In contrast, our proposed metric treats each protected attribute independently and aggregates the results afterwards.

Finally, the paper concludes with experiments on real-world datasets to show the effectiveness and adaptability of our framework to minimize discrimination for multiple protected attributes using several metrics. We also discuss the shortcomings of the existing framework `Fairlearn` [16] for comparison. We argue that our framework is superior, as data integrity constraints are kept intact and not violated. The pre-processing technique `CorrelationRemover` in `Fairlearn` transforms discrete target labels to continuous values. By changing the statistical data type, the same fairness metrics and classification algorithms can no longer be used, which makes comparing `CorrelationRemover` with our framework infeasible.

Measuring and Mitigating Bias for Tabular Datasets with Multiple Protected Attributes*

Manh Khoi Duong[1],*, Stefan Conrad[1]

[1]*Heinrich Heine University, Universitätsstraße 1, 40225 Düsseldorf, Germany*

Abstract

Motivated by the recital (67) of the current corrigendum of the AI Act in the European Union, we propose and present measures and mitigation strategies for discrimination in tabular datasets. We specifically focus on datasets that contain multiple protected attributes, such as nationality, age, and sex. This makes measuring and mitigating bias more challenging, as many existing methods are designed for a single protected attribute. This paper comes with a twofold contribution: Firstly, new discrimination measures are introduced. These measures are categorized in our framework along with existing ones, guiding researchers and practitioners in choosing the right measure to assess the fairness of the underlying dataset. Secondly, a novel application of an existing bias mitigation method, `FairDo`, is presented. We show that this strategy can mitigate any type of discrimination, including intersectional discrimination, by transforming the dataset. By conducting experiments on real-world datasets (Adult, Bank, COMPAS), we demonstrate that de-biasing datasets with multiple protected attributes is possible. All transformed datasets show a reduction in discrimination, on average by 28%. Further, these datasets do not compromise any of the tested machine learning models' performances significantly compared to the original datasets. Conclusively, this study demonstrates the effectiveness of the mitigation strategy used and contributes to the ongoing discussion on the implementation of the European Union's AI Act.

Keywords

Machine Learning, Bias Mitigation, Intersectional Discrimination, Fairness, AI Act

1. Introduction

Discrimination in artificial intelligence (AI) applications is a growing concern since the adoption of the *AI Act* by the European Parliament on March 13, 2024 [1]. It still remains a significant challenge across numerous domains [2, 3, 4, 5]. To prevent biased outcomes, *pre-processing* methods are often used to mitigate biases in datasets before training machine learning models [6, 7, 8, 9]. The current corrigendum of the *AI Act* [1] emphasizes this in Recital (67):

> *"[...] The data sets should also have the appropriate statistical properties, including as regards the persons or groups of persons in relation to whom the high-risk AI system is intended to be used, with specific attention to the mitigation of possible biases in the data sets [...]"*

Since datasets often consist of multiple protected attributes, pre-processing methods should be able to handle these cases. However, only a few works have addressed this issue [7, 10, 11, 12, 13] and de-biasing such datasets is still an ongoing research topic. In addition, there is no straightforward approach to managing multiple protected attributes, as shown in Figure 1.

Our paper mainly focuses on how to measure and mitigate discrimination in datasets where multiple protected attributes are present. In our first contribution, we provide a comprehensive categorization of discrimination measuring methods. Besides introducing new measures for some of these cases, we also categorize existing measures from the literature. Some of the listed measures specifically address *intersectional discrimination* and *non-binary groups*. The second contribution deals with bias mitigation. For this, we use our published pre-processing framework, `FairDo` [9], that is *fairness-agnostic*. The fairness-agnostic property makes it possible to define any discrimination measure that should be

AEQUITAS 2024: Workshop on Fairness and Bias in AI | co-located with ECAI 2024, Santiago de Compostela, Spain

*Corresponding author.

manh.khoi.duong@hhu.de (M. K. Duong); stefan.conrad@hhu.de (S. Conrad)

0000-0002-4653-7685 (M. K. Duong); 0000-0003-2788-3854 (S. Conrad)

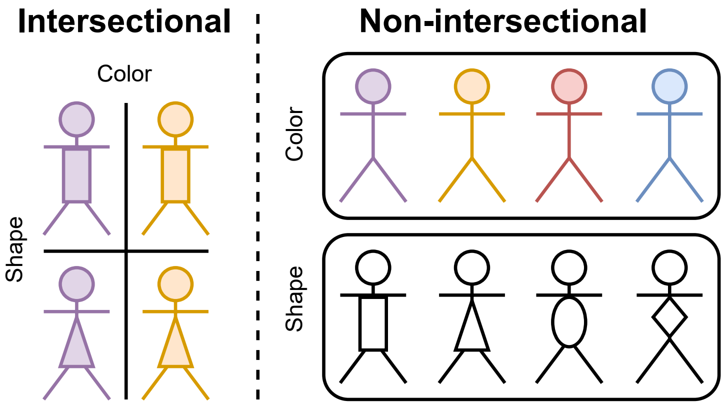


Figure 1: Stick figures can be differentiated by their color and shape. In intersectional discrimination, attributes are intersected, which leads to new subgroups. In non-intersectional, each attribute is treated independently, i.e., colors and shapes are not intersecting in this case.

minimized. By implementing the introduced measures, we can therefore mitigate biases for multiple protected attributes. Another advantage of `FairDo` is that it preserves data integrity and does not modify the features of individuals during the optimization process, unlike other methods [14, 3, 7].

We evaluated our methodology on popular tabular datasets with fairness concerns, such as Adult [15], Bank [16], and COMPAS [17]. We used different discrimination measures to evaluate the effectiveness of the bias mitigation process. Because a successful mitigation process does not guarantee that the outcomes of machine learning models are fair, we trained machine learning models on the transformed datasets and evaluated their predictions regarding fairness and performance. The code for the experiments can be found in the accompanying repository: https://github.com/mkduong-ai/fairdo/evaluation.

The results of the bias mitigation process as well as the performance of the machine learning models are promising. They indicate that achieving fairness in datasets with multiple protected attributes is possible, and `FairDo` is a proper framework for this task. Overall, our work contributes technical solutions for stakeholders to enhance the fairness of datasets and machine learning models, aiming for compliance with the *AI Act* [1].

2. Preliminaries

To handle multiple protected attributes, we define $\mathcal{Z} = \{Z_1, \ldots, Z_p\}$ as a set of protected attributes. It can represent the set of sociodemographic features such as age, gender, and ethnicity. These factors may make individuals vulnerable to discrimination. Each protected attribute $Z_k \in \mathcal{Z}$ is formally a *discrete random variable* that can take on values from the sample space g_k. In this context, we refer g_k to groups that describe distinct social categories of a protected attribute. For example, let Z_k represent gender; then g_k is a set containing the genders male, female, and non-binary. To avoid limitations to a particular group fairness notion, we introduce a generalized notation based on the works of Žliobaitė [2], Duong and Conrad [9] in the following.

Definition 2.1 (Treatment). *Let E_1, E_2 be events and Z_k be a random variable that can take on values from g_k, then we call the conditional probability*

$$P(E_1 \mid E_2, Z_k = i)$$

treatment, *where $i \in g_k$. E_1 describes some favorable outcome, such as getting accepted for a job, while E_2 often represents some additional information about the individual, such as their qualifications.*

Definition 2.2 (Fairness Criteria). *With the definition of treatment, we can define fairness criteria that demand equal treatment for different groups. Let $P(E_1 \mid E_2, Z_k = i)$ and $P(E_1 \mid E_2, Z_k = j)$ be* *treatments, then we call the following equation:*

$$P(E_1 \mid E_2, Z_k = i) = P(E_1 \mid E_2, Z_k = j)$$

a fairness criterion*, for all* $i, j \in g_k$*.*

Definition 2.2 allows us to define various group fairness criteria, including *statistical parity* [18], *predictive parity* [3], *equality of opportunity* [19], etc. They all demand some sort of equal outcome for different groups and can be defined by configuring the events E_1, E_2. For instance, statistical parity [18] requires that two different groups have an equal probability of receiving a favorable outcome $(Y = 1)$.

Example 2.1 (Statistical Parity [18])**.** *To define statistical parity for the attribute* Z_k *using our notation, we set* $E_1 := (Y = 1)$ *and* $E_2 := \Omega$*. By setting* E_2 *to the sample space* Ω*, we compare the probabilities of the event* $Y = 1$ *across different groups without conditioning on any additional event:*

$$\begin{aligned} &P(Y = 1 \mid \Omega, Z_k = i) = P(Y = 1 \mid \Omega, Z_k = j) \\ \iff &P(Y = 1 \mid Z_k = i) = P(Y = 1 \mid Z_k = j), \end{aligned}$$

where $i, j \in g$ *represent different groups.*

In real-world applications, achieving equal probabilities for certain outcomes is not always possible. Due to variations in sample sizes in the groups, it is common to yield unequal treatments, even when they are similar. Thus, existing literature [2] uses the absolute difference to quantify the strength of discrimination.

Definition 2.3 (Disparity)**.** *Let* $P(E_1 \mid E_2, Z_k = i)$ *and* $P(E_1 \mid E_2, Z_k = j)$ *be two treatments, then we refer to*

$$\delta_{Z_k}(i, j, E_1, E_2) = |P(E_1 \mid E_2, Z_k = i) - P(E_1 \mid E_2, Z_k = j)|$$

as the disparity*, for all* $i, j \in g_k$*. Trivially,* δ_{Z_k} *is commutative regarding* i, j*. In practice, it prevents reverse discrimination due to the absolute value.*

Definition 2.4 (Discrimination)**.** *We use* $\psi\colon \mathbb{D} \to \mathbb{R}$ *to denote some discrimination measure that quantifies the discrimination inherent in any dataset* $\mathcal{D} \in \mathbb{D}$*. A dataset* $\mathcal{D}$ *consists of features, protected attributes, and labels for each individual. The explicit form of* ψ *depends on the cases introduced in Section 3.*

3. Measuring Discrimination for Multiple Attributes

We found that numerous scenarios arise when dealing with multiple protected attributes. We categorize these scenarios based on the number of groups, denoted as $|g|$, and the number of protected attributes, denoted as $|\mathcal{Z}|$. By going through all cases, we present possible approaches from the literature as well as our own suggestions to measure discrimination.

3.1. Single Protected Attribute $(|\mathcal{Z}| = 1)$

In the case of having only one protected attribute, i.e., $|\mathcal{Z}| = |\{Z_1\}| = 1$, we distinguish between cases by the number of available groups $|g|$ in the dataset. We categorize the cases by $|g| = 0, 1, 2$, and $|g| > 2$.

3.1.1. No Groups ($|g| = 0$)

When there are no groups, the measurement of discrimination is impossible if no assumptions are being made. Discrimination can be assessed through proxy variables [20]; however, this approach can be imprecise and may introduce new biases. This case is equivalent to having no protected attribute, i.e., $|\mathcal{Z}| = 0$.

3.1.2. Single Group ($|g| = 1$)

Similarly to the case of having no groups, discrimination cannot be measured when having only one group. For this, we propose practices where prior information can be incorporated:

1. *No discrimination*: As no difference towards any other group can be measured, returning a discrimination score of 0 is one viable option.
$$\psi(\mathcal{D}) = 0. \tag{1}$$
2. *Difference to optimal treatment*: Another way is to return the absolute difference of the group's outcome to the optimal treatment. For example, group i has an 80% chance of receiving the favorable treatment. Ideally, having a 100% chance would represent the optimal scenario. Therefore, the discrimination score is 20% in this case. It is given by:
$$\psi(\mathcal{D}) = |P(E_1 \mid E_2, Z_1 = i) - 1|. \tag{2}$$
3. *Difference to expected treatment*: We can use the expected treatment as a reference point. For example, we know that a company has a 50% acceptance rate for job applications. Now a machine learning classifier is trained to predict whether an applicant will be accepted and the model's predictions result in a 60% acceptance rate for group i. Hence, the model is positively biased towards group i by 10%. This can be formulated as:
$$\psi(\mathcal{D}) = |P(E_1 \mid E_2, Z_1 = i) - p_{\text{expect.}}|, \tag{3}$$
where $p_{\text{expect.}}$ is the expected treatment. It can describe the average treatment across all groups [21] or some other prior information that is not included in the dataset.

3.1.3. Binary Groups ($|g| = 2$)

Without using any prior information, we can calculate the discrimination score by taking the absolute difference between the treatments of the two groups, as advised by Žliobaitė [2]. The discrimination measure ψ is then simply given by the disparity as mentioned in Definition 2.3.

3.1.4. Non-binary Groups ($|g| > 2$)

While the case for binary attributes is straightforward, it becomes non-trivial for non-binary attributes that arise naturally in real-world data. We can fall back to $|g| = 2$ by calculating the absolute difference between every distinct group $i, j \in g$. Because the discrimination between i and j is the same as between j and i, only $\binom{|g|}{2}$ pairs need to be compared and we use an aggregation function $\text{agg}^{(1)}$ to report the differences [2]. Lum et al. [22] refers to measures that aggregate or summarize discrimination scores as *meta-metrics*. The aggregate can be the *sum* or *maximum* function, depending on the use case. The result for a single protected attribute Z_k with two or more groups can be computed as follows:

$$\psi(\mathcal{D}) = \underset{i,j \in g_k, i<j}{\text{agg}^{(1)}} \delta_{Z_k}(i, j, E_1, E_2), \tag{4}$$

where δ_{Z_k} is the disparity as defined in Definition 2.3 and $i < j$ ensures that each pair is considered only once (assuming label-encoded groups). According to Žliobaitė [2] and her personal discussions with legal experts, she advocates using the maximum function, i.e.,

$$\begin{align} \psi(\mathcal{D}) &= \max_{i,j \in g_k, i<j} \delta_{Z_k}(i, j, E_1, E_2) \tag{5} \\ &= \max_{i \in g_k} P(E_1 \mid E_2, Z_k = i) - \min_{j \in g_k} P(E_1 \mid E_2, Z_k = j). \tag{6} \end{align}$$

Equation (5) describes the maximum discrimination obtainable between two groups. An alternative and equivalent formulation is given in Equation (6) [7]. The latter is computationally more efficient as it requires $\mathcal{O}(2|g|)$ operations compared to $\mathcal{O}(|g|^2)$ operations for the former.

A more general approach to measuring discrimination is to calculate some form of *correlation coefficient* between the protected attribute and the outcome. The correlation coefficient can be calculated using Pearson's correlation [23], Spearman or Kendall's rank correlation [24, 25]. The discrimination measure can then be defined as the absolute value of the correlation coefficient:

$$\psi(\mathcal{D}) = |\mathrm{Corr}(E_1, Z_k)|. \tag{7}$$

This approach can be applied to any number of groups. `Fairlearn` provides a pre-processing method that removes the correlation between the protected attribute and the outcome by transforming the data [7]. However, the given approach violates data integrity constraints as categorical attributes are transformed into continuous values. Moreover, zero correlation does not imply independence between two variables.

3.2. Multiple Protected Attributes ($|\mathcal{Z}| > 1$)

There are several ways to measure discrimination for multiple protected attributes ($|\mathcal{Z}| > 1$). Based on the works of Kearns et al. [21], Yang et al. [11] and Kang et al. [13], we categorize them into two approaches: *intersectional* and *non-intersectional* (see Figure 1). Intersectional approaches consider the intersection of identities. The overlapping of such identities forms *subgroups* [21]. Non-intersectional approaches treat each protected attribute independently [11].

3.2.1. Intersectional Discrimination

The central idea of intersectionality is that individuals experience overlapping forms of oppression or privilege based on the combination of multiple social categories they belong to. In the following, we will introduce definitions to formulate intersectional discrimination, which is based on the work of Kearns et al. [21].

Definition 3.1 (Subgroup [21])**.** *Let $\mathcal{Z} = \{Z_1, \dots, Z_p\}$ be a set of discrete random variables representing protected attributes that can take on values from corresponding groups $g_1, \dots, g_p$. A* subgroup *i is defined as $i = (i_1, \dots, i_p) \in g_1 \times \dots \times g_p$. In other words, a* subgroup *encompasses multiple groups from different protected attributes.*

Definition 3.2 (Subgroup Treatment)**.** *Let i be a subgroup as defined in Definition 3.1 and let $\mathcal{Z} = \{Z_1, \dots, Z_p\}$ be a set of discrete random variables.* Subgroup treatment *is then defined as:*

$$P(E_1 \mid E_2, Z_1 = i_1, \dots, Z_p = i_p).$$

Definition 3.3 (Subgroup Disparity)**.** *Let $\mathcal{Z} = \{Z_1, \dots, Z_p\}$ be a set of discrete random variables. Let $i, j \in g_1 \times \dots \times g_p$ be two subgroups with $i = (i_1, \dots, i_p)$ and $j = (j_1, \dots, j_p)$. The disparity between two subgroups is denoted as $\hat{\delta}_{\mathcal{Z}}$ and is given by:*

$$\hat{\delta}_{\mathcal{Z}}(i, j, E_1, E_2) = |P(E_1 \mid E_2, Z_1 = i_1, \dots, Z_p = i_p) - P(E_1 \mid E_2, Z_1 = j_1, \dots, Z_p = j_p)|.$$

Similarly to Equation (4), we can calculate the discrimination score for multiple protected attributes by aggregating disparities across all subgroups. A subgroup can be treated like a normal group. According to Definition 3.1, there are theoretically at least 2^p subgroups, where p is the number of protected attributes. However, not all subgroups may be available in the dataset. For unavailable subgroups, the disparity cannot be calculated as the corresponding treatment is undefined.

Let us denote the set of available subgroups as $G_{\text{avail}} \subseteq g_1 \times \dots \times g_k$. To finally capture the discrepancies across all available subgroup pairs, an aggregation function $\mathrm{agg}^{(1)}$ is applied to the subgroup disparities $\hat{\delta}_{\mathcal{Z}}$:

$$\psi_{\text{intersect}}(\mathcal{D}) = \underset{i,j \in G_{\text{avail}}}{\mathrm{agg}^{(1)}} \hat{\delta}_{\mathcal{Z}}(i, j, E_1, E_2). \tag{8}$$

Table 1

Example dataset of individuals receiving a favorable ($Y = 1$) or unfavorable ($Y = 0$) outcome. The dataset shows four individuals with their respective age group and sex.

Individual	Age	Sex	Outcome (Y)
1	Old	Male	1
2	Old	Female	0
3	Young	Male	0
4	Young	Female	1

Equation (8) represents the aggregated discrimination between all available subgroups in the dataset. When using the maximum function as the aggregator, the calculations are equivalent to Equation (5) and Equation (6). The only difference is that the conditionals are now subgroups instead of groups:

$$\begin{aligned}\psi_{\text{intersect}}(\mathcal{D}) &= \max_{i,j \in G_{\text{avail}}} \hat{\delta}_{Z_k}(i, j, E_1, E_2) \\ &= \max_{i \in G_{\text{avail}}} P(E_1 \mid E_2, Z_1 = i_1, \ldots, Z_p = i_p) - \min_{j \in G_{\text{avail}}} P(E_1 \mid E_2, Z_1 = j_1, \ldots, Z_p = j_p).\end{aligned} \tag{9}$$

Kang et al. [13] also dealt with intersectional discrimination in their work by introducing a multivariate random variable Z where each dimension represents a protected attribute. Their fairness objective is to minimize the mutual information between the outcome and the multivariate random variable. By minimizing the mutual information, the outcome is independent of the protected attributes, which is a desirable property for fairness [14, 26]. In this context, zero mutual information implies the absence of intersectional discrimination [13]. However, this approach relies on expensive techniques to approximate the mutual information. Using our notation, their formulation can be written as [13]:

$$\psi_{\text{MI}}(\mathcal{D}) = \text{MI}(E_1, Z), \tag{10}$$

where MI denotes the mutual information.

3.2.2. Non-intersectional Discrimination

The problem with measuring discrimination for intersectional groups is that it has an upward bias when using meta-metrics [22]. This is because the number of subgroups grows exponentially with the number of protected attributes. This leads to many subgroups where the number of samples in each subgroup is possibly small, resulting in larger noise in the treatment estimates [22].

Besides intersectional groups, Yang et al. [11] listed a non-intersectional definition of groups, called *independent groups*. Building on the definition of *independent groups*, we propose an appropriate approach to measure discrimination for this type of groups. It is more suitable when dealing with a large number of subgroups or when intersectional discrimination is not deemed important. Our non-intersectional approach treats each protected attribute independently and aggregates the discrimination scores across all protected attributes. For this, a second aggregate function with $\text{agg}^{(2)}$ is introduced, yielding the following equation:

$$\psi_{\text{indep}}(\mathcal{D}) = \underset{Z_k \in \mathcal{Z}}{\text{agg}^{(2)}} \left\{ \underset{i,j \in g_k, i<j}{\text{agg}^{(1)}} \delta_{Z_k}(i, j, E_1, E_2) \right\}. \tag{11}$$

The first-level aggregator $\text{agg}^{(1)}$ aggregates disparities within a protected attribute, considering unique pairs of groups i and j. The second-level aggregator $\text{agg}^{(2)}$ then combines the results across all protected attributes. By applying both operators, we obtain a discrimination measure that captures disparities between groups across multiple attributes.

3.2.3. Example

Let us consider a dataset with two protected attributes, age and sex (see Table 1). The set of protected attributes is $\mathcal{Z} = \{Z_1, Z_2\} = \{\text{Age}, \text{Sex}\}$ and the set of available subgroups in the dataset is $G_{\text{avail}} = \{\text{Old}, \text{Young}\} \times \{\text{Male}, \text{Female}\}$. We measure discrimination using *statistical disparity.* For simplicity, all aggregation functions are set to the maximum function. The *intersectional approach* yields the following discrimination score:

$$\begin{aligned}\psi_{\text{intersect}}(\mathcal{D}) &= \max_{i,j \in G_{\text{avail}}} \hat{\delta}_{\mathcal{Z}}(i, j, (Y = 1), \Omega) \\ &= \max_{i,j \in G_{\text{avail}}} \hat{\delta}_{\{\text{Age, Sex}\}}(i, j, (Y = 1), \Omega) \\ &= \max_{i \in G_{\text{avail}}} P(Y = 1 \mid Z_1 = i_1, Z_2 = i_2) - \min_{j \in G_{\text{avail}}} P(Y = 1 \mid Z_1 = j_1, Z_2 = j_2) \\ &= |P(Y = 1 \mid \text{Age} = \text{Old}, \text{Sex} = \text{Male}) - P(Y = 1 \mid \text{Age} = \text{Young}, \text{Sex} = \text{Male})| = 1,\end{aligned} \tag{12}$$

while the discrimination score for the *non-intersectional approach* is given by:

$$\begin{aligned}\psi_{\text{indep}}(\mathcal{D}) &= \max_{Z_k \in Z} \left\{ \max_{i,j \in g_k, i<j} \delta_{Z_k}(i, j, (Y = 1), \Omega) \right\} \\ &= \max \left\{ \delta_{\text{Age}}(\text{Old}, \text{Young}, (Y = 1), \Omega), \delta_{\text{Sex}}(\text{Male}, \text{Female}, (Y = 1), \Omega) \right\} \\ &= \max\{|0.5 - 0.5|, |0.5 - 0.5|\} = \max\{0, 0\} = 0.\end{aligned} \tag{13}$$

The non-intersectional approach yields a discrimination score of 0 because the disparities for both protected attributes are 0. This is quite different from the intersectional approach, which reports a discrimination score of 1. As seen, the results can differ depending on the approach.

4. Experiments

Our experimentation follows a pipeline consisting of *data pre-processing*, *bias mitigation*, *model training*, and *evaluation.* To mitigate bias in tabular datasets with multiple protected attributes, we used the sampling method, `FairDo` [9], that constructs fair datasets by selectively sampling data points. The method is very flexible and only requires the user to define the discrimination measure that should be minimized. In our case, we are interested in a dataset that has minimal bias across multiple protected attributes. The experiments revolve around the following research questions:

- **RQ1** Is it possible to yield a fair dataset with `FairDo`, where bias for multiple protected attributes is reduced?
- **RQ2** Are machine learning models trained on fair datasets more fair in their predictions than those trained on original datasets?

4.1. Experimental Setup

Datasets and Pre-processing The tabular datasets employed in our experiments include the Adult [15], Bank [16], and COMPAS [17] datasets. They are known for their use in fairness research and contain multiple protected attributes. We pre-processed the datasets by applying one-hot encoding to categorical variables and label encoding to protected attributes. Table 2 shows important characteristics of the datasets after pre-processing.

Each dataset was divided into training and testing sets using an 80/20 split, respectively. We ensured that the split was stratified (if possible) based on protected attributes to maintain representativeness across different groups in both sets.

Table 2
Overview of Datasets

Dataset	Samples	Feats.	Label	Protected Attributes	Description
Adult [15]	32 561	21	Income	**Race**: White, Black, Asian-Pacific-Islander, American-Indian-Eskimo, Other **Sex**: Male, Female	Indicates individuals earning over $50,000 annually
Bank [16]	41 188	50	Term deposit subscription	**Job**: Admin, Blue-Collar, Technician, Services, Management, Retired, Entrepreneur, Self-Employed, Housemaid, Unemployed, Student, Unknown **Marital Status**: Divorced, Married, Single, Unknown	Shows whether the client has subscribed to a term deposit.
COMPAS [17]	7 214	13	2-year recidivism	**Race**: African-American, Caucasian, Hispanic, Other, Asian, Native American **Sex**: Male, Female **Age Category**: <25, 25-45, >45	Displays individuals that were rearrested for a new crime within 2 years after initial arrest.

Bias Mitigation Applying the bias mitigation method `FairDo` [9] to the datasets can be regarded as a pre-processing step, too. This is because the method simply returns a dataset that is fair with respect to the given discrimination measure. `FairDo` [9] offers a variety of options to mitigate bias, and we chose the *undersampling* method that removes samples. In this option, the optimization objective is stated as [9]:

$$\min_{\mathcal{D}_{\text{fair}} \subseteq \mathcal{D}} \quad \psi(\mathcal{D}_{\text{fair}}), \tag{14}$$

where $\mathcal{D}$ is the training set of Adult, Bank, or COMPAS, and ψ is the fairness objective function. We experimented with both $\psi_{\text{intersect}}$ and ψ_{indep} as objectives functions. Bias mitigation is only applied to the training set and the testing set remains unchanged. `FairDo` internally uses genetic algorithms to select a subset of the training set that minimizes the objective function. We used the same settings and operators as provided in the package and only adjusted the population size (200) and the number of generations (400).

Model Training We utilized the `scikit-learn` library [27] to train various machine learning classifiers, namely *Logistic Regression* (LR), *Support Vector Machine* (SVM), *Random Forest* (RF), and *Artificial Neural Network* (ANN). These classifiers were trained on both the original and fair datasets. Classifiers trained on the original datasets serve as a baseline for comparison. We used the default hyperparameters given by `scikit-learn` package for each classifier.

Evaluation Metrics We evaluated the models' predictions on fairness and performance using the test set. For fairness, we assessed $\psi_{\text{intersect}}$ and ψ_{indep}. For the classifiers' performances, we report the *area under the receiver operating characteristic curve* (AUROC) [28], where higher values indicate better performances. Because removing data points can compromise the overall quality of the data, we also report the number of subgroups before and after bias mitigation to check for representativeness.

Trials For each dataset and discrimination measure combination, the bias mitigation process was repeated 10 times. The results were averaged over the trials to obtain a more robust evaluation.

4.2. Results

Fair Dataset Generation Table 3 shows the average discrimination before and after mitigating bias in the training sets. On all datasets, discrimination was reduced after applying `FairDo`. Without

Table 3
Average discrimination and number of subgroups before and after pre-processing the training sets with `FairDo`.

Dataset	Metric	Disc. Before	**Disc. After**	Subgroups Before	Subgroups After
Adult	ψ_{indep}	20%	13%	10	10
	$\psi_{\text{intersect}}$	31%	16%	10	10
Bank	ψ_{indep}	24%	5%	48	48
	$\psi_{\text{intersect}}$	33%	15%	48	46.2
COMPAS	ψ_{indep}	30%	5%	34	34
	$\psi_{\text{intersect}}$	100%	17%	34	28.8

considering group intersections, discrimination was reduced by 7%, 19%, and 25% for Adult, Bank, and COMPAS, respectively. When considering intersectionality, the discrimination was reduced by 15%, 18%, and 83%. Hence, discrimination was reduced by 28% on average across all datasets, thus answering **RQ1** positively. When comparing the discrimination scores, it can be observed that the intersectional discrimination scores are generally higher. This is because in the intersectional setting, more subgroups are considered, which potentially leads to larger differences between them [21].

We also report the number of subgroups before and after bias mitigation to assess the impact of the undersampling method on the dataset. The removal of subgroups can only be observed in the intersectional setting. In the COMPAS dataset 5.2 out of 34 subgroups were removed on average, indicating the largest amount of subgroups removed across all datasets. While the Bank dataset consists of 48 subgroups, only 1.8 subgroups were removed on average. Because the COMPAS dataset's initial intersectional discrimination score is 100%, removing more subgroups seems inevitable to reduce bias.

Model Performance and Fairness Figure 2 shows the results of the classifiers' performances on the test set. The classifiers' performances are displayed on the y-axis, while the discrimination values are shown on the x-axis. We note that the axes do not share the same scale across the subfigures for analytical purposes.

Classifiers trained on fair datasets did not suffer a significant decline in performance compared to those trained on original datasets. In all cases, only a slight decrease of 1%-3% in performance can be noted. This indicates that the bias mitigation process does not compromise the dataset's fidelity and, therefore, the classifiers' performances. Regarding discrimination, a significant reduction is evident. The x-axis scales are much larger than the y-axis scales, suggesting that changes in discrimination are larger than changes in performance. For example, the RF classifier trained on the Bank dataset (Figure 2g) shows a decrease in intersectional discrimination from 38% to 15%, while the performance only decreases by 2%. Similar results can be observed for the other classifiers and datasets as well, successfully addressing **RQ2**. The results suggest that `FairDo` can be reliably used to mitigate bias in tabular datasets for various measures that consider multiple protected attributes. Still, we advise users to carefully perform similar analyses when applying the method to their datasets.

5. Discussion

The results of our experiments show that the presented measures detect discrimination in datasets with multiple protected attributes differently. When using the intersectional discrimination measure, more groups are identified and compared to each other. While subgroups are not ignored by this measure, measuring higher discrimination scores by random chance becomes more likely [21, 22]. In contrast, treating each protected attribute separately prevents this issue but may lead to overlooking discrimination. The choice of measure is up to the stakeholders and depends on the context of the dataset and the regulations that apply to the AI system. We generally recommend using the intersectional discrimination measure if the number of individuals in each subgroup is large enough to draw statistically significant conclusions. Otherwise, treating each protected attribute separately is more suitable.

By using the mitigation strategy `FairDo` [9], the resulting datasets in the experiments have improved

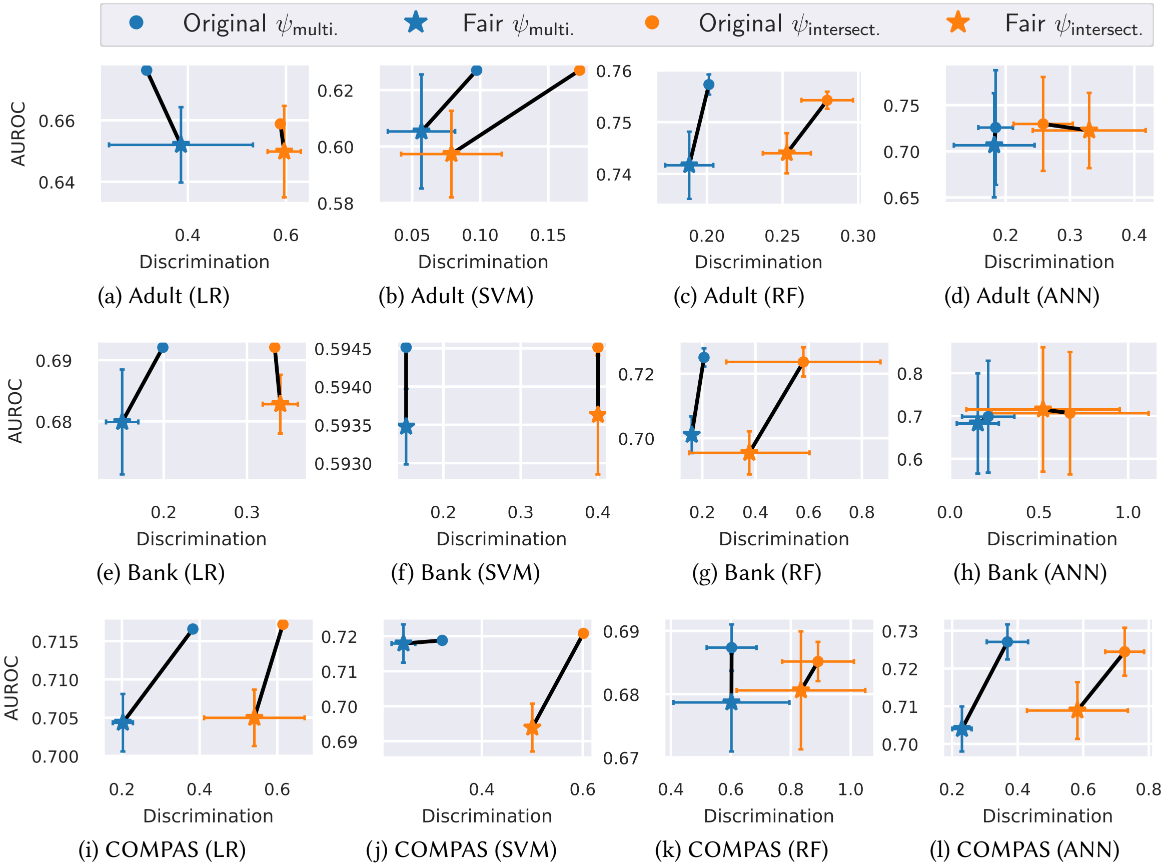


Figure 2: Results on the test set. The x-axis represents the discrimination values (legend indicates used measure) and the y-axis represents the classifiers' performances. We compare the pre-processed (fair) data with the original data. The points/stars represent averages, and the error bars display the standard deviations of the AUROC and discrimination values over 10 trials.

statistical properties regarding fairness. Whether intersectionality was considered or not, reducing discrimination in datasets was possible. At the current state, the AI Act [1] does not explicitly mention *intersectional discrimination* nor how to deal with multiple protected attributes generally. While recital (67) states that datasets *"should [...] have the appropriate statistical properties"*, it does not specify what these properties are. Hence, our work serves as an initial guideline for what these properties could be and how to achieve them in practice.

6. Conclusion

Datasets often come with multiple protected attributes, which makes measuring and mitigating discrimination more challenging. Most existing studies only deal with a single protected attribute, and works that consider multiple protected attributes often focus on intersectionality. In opposition to this, we proposed a new non-intersectional measure that treats each protected attribute separately. This is more suitable when the number of subgroups is too large or the number of individuals in each subgroup is small. We used both intersectional and non-intersectional measures as objectives and applied the `FairDo` framework to mitigate discrimination in multiple datasets. The experiments show that discrimination was reduced in all datasets and on average by 28%. Machine learning models trained on the bias-mitigated datasets also improved their fairness while maintaining performance compared to models trained on the original datasets.

References

[1] European Commission, Artificial Intelligence Act, Corrigendum, 19 April 2024, Available online: https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138-FNL-COR01_EN.pdf, 2024. Accessed: 17 May 2024.

[2] I. Žliobaitė, Measuring discrimination in algorithmic decision making, Data Mining and Knowledge Discovery 31 (2017) 1060–1089.

[3] M. B. Zafar, I. Valera, M. Gomez Rodriguez, K. P. Gummadi, Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment, in: Proceedings of the 26th International Conference on World Wide Web, 2017. doi:10.1145/3038912.3052660.

[4] S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, A. Huq, Algorithmic decision making and the cost of fairness, in: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2017, pp. 797–806.

[5] S. Barocas, M. Hardt, A. Narayanan, Fairness and Machine Learning, fairmlbook.org, 2019.

[6] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, S. Venkatasubramanian, Certifying and removing disparate impact, in: proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining, 2015, pp. 259–268.

[7] S. Bird, M. Dudík, R. Edgar, B. Horn, R. Lutz, V. Milan, M. Sameki, H. Wallach, K. Walker, Fairlearn: A toolkit for assessing and improving fairness in AI, Technical Report MSR-TR-2020-32, Microsoft, 2020. URL: https://www.microsoft.com/en-us/research/publication/fairlearn-a-toolkit-for-assessing-and-improving-fairness-in-ai/.

[8] A. Agarwal, A. Beygelzimer, M. Dudík, J. Langford, H. Wallach, A reductions approach to fair classification, in: International Conference on Machine Learning, PMLR, 2018, pp. 60–69.

[9] M. K. Duong, S. Conrad, Towards fairness and privacy: A novel data pre-processing optimization framework for non-binary protected attributes, in: D. Benavides-Prado, S. Erfani, P. Fournier-Viger, Y. L. Boo, Y. S. Koh (Eds.), Data Science and Machine Learning, Springer Nature Singapore, Singapore, 2024, pp. 105–120.

[10] J. R. Foulds, R. Islam, K. N. Keya, S. Pan, Bayesian Modeling of Intersectional Fairness: The Variance of Bias, 2020, pp. 424–432. doi:10.1137/1.9781611976236.48.

[11] F. Yang, M. Cisse, S. Koyejo, Fairness with overlapping groups, in: Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS '20, Curran Associates Inc., Red Hook, NY, USA, 2020.

[12] L. E. Celis, L. Huang, V. Keswani, N. K. Vishnoi, Fair classification with noisy protected attributes: A framework with provable guarantees, in: M. Meila, T. Zhang (Eds.), Proceedings of the 38th International Conference on Machine Learning, volume 139 of *Proceedings of Machine Learning Research*, PMLR, 2021, pp. 1349–1361.

[13] J. Kang, T. Xie, X. Wu, R. Maciejewski, H. Tong, Infofair: Information-theoretic intersectional fairness, 2022 IEEE International Conference on Big Data (Big Data) (2021) 1455–1464.

[14] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, C. Dwork, Learning fair representations, in: International conference on machine learning, PMLR, 2013, pp. 325–333.

[15] R. Kohavi, Scaling up the Accuracy of Naive-Bayes Classifiers: A Decision-Tree Hybrid, KDD'96, AAAI Press, 1996, p. 202–207.

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

[17] J. Larson, J. Angwin, S. Mattu, L. Kirchner, Machine bias, 2016. URL: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[18] T. Calders, F. Kamiran, M. Pechenizkiy, Building classifiers with independency constraints, in: 2009 IEEE International Conference on Data Mining Workshops, 2009, pp. 13–18. doi:10.1109/ICDMW.2009.83.

[19] M. Hardt, E. Price, N. Srebro, Equality of opportunity in supervised learning, Advances in neural information processing systems 29 (2016).

[20] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, A. Galstyan, A survey on bias and fairness in

machine learning, ACM Computing Surveys (CSUR) 54 (2021) 1–35.

[21] M. Kearns, S. Neel, A. Roth, Z. S. Wu, Preventing fairness gerrymandering: Auditing and learning for subgroup fairness, in: J. Dy, A. Krause (Eds.), Proceedings of the 35th International Conference on Machine Learning, volume 80 of *Proceedings of Machine Learning Research*, PMLR, 2018, pp. 2564–2572.

[22] K. Lum, Y. Zhang, A. Bower, De-biasing "bias" measurement, in: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, FAccT '22, Association for Computing Machinery, New York, NY, USA, 2022, p. 379–389. doi:10.1145/3531146.3533105.

[23] K. Pearson, Notes on regression and inheritance in the case of two parents, Proceedings of the Royal Society of London 58 (1895) 240–242.

[24] C. Spearman, The proof and measurement of association between two things, American Journal of Psychology 15 (1904) 72–101.

[25] M. G. Kendall, A new measure of rank correlation, Biometrika 30 (1938) 81–93.

[26] A. Ghassami, S. Khodadadian, N. Kiyavash, Fairness in supervised learning: An information theoretic approach, in: 2018 IEEE International Symposium on Information Theory (ISIT), IEEE Press, 2018, p. 176–180. doi:10.1109/ISIT.2018.8437807.

[27] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830.

[28] T. Fawcett, An introduction to ROC analysis, Pattern Recognition Letters 27 (2006) 861–874. doi:10.1016/j.patrec.2005.10.010.

4

Enhancing Trust in Assessing and Mitigating Bias

So far, we have covered works that deal with assessing and mitigating bias in datasets and machine learning models [2, 3, 4]. Assessing bias is an essential step to detect whether a dataset or a model is biased. Based on the assessment, mitigation methods such as `FairDo` [3] can be applied to reduce the measured bias in the dataset. However, assessment and mitigation methods do not come without challenges regarding their trustworthiness. Specifically, when dealing with underrepresented minority groups, assessing bias becomes less reliable and prone to estimation errors due to small sample sizes. This makes it difficult to trust the results and the decisions made based on them. Regarding the mitigation methods, they may introduce new biases or reduce the data quality and therefore impact the performance of machine learning models negatively.

Early works [35, 27] have pointed out the challenges of assessing discrimination for underrepresented groups. Kearns et al. [35]'s subgroup fairness definition underweights the discrimination of small groups, and Foulds et al. [27] proposed a Bayesian approach to estimate two intersectional fairness metrics. However, these works do not address the trustworthiness aspect and do not deal with disparity metrics that are used in practice and advised by Žliobaitė [53].

To address these challenges, we propose a new fairness assessment method that incorporates the uncertainty of the bias estimation [7]. In addition to the disparity, an uncertainty metric is provided which quantifies the reliability of the disparity metric. This allows the comparison between different models or decision-makers that exhibit the same amount of bias but come with a different amount of samples. Regarding our method `FairDo`, we consider the amount of data that is removed during the mitigation process and provide users with the option to control the trade-off between fairness and data lost [6]. By visualizing the Pareto frontier of the trade-off, we ease the decision-making process for users.

4.1 Enhancing Trust in Disparity Measures

Paper: Manh Khoi Duong and Stefan Conrad. (Un)certainty of (Un)fairness: Preference-Based Selection of Certainly Fair Decision-Makers. In *ECAI 2024: 27th European Conference on Artificial Intelligence*, volume 392 of *Frontiers in Artificial Intelligence and Applications*. IOS Press, 2024.

Personal Contribution: The research idea was developed solely by Manh Khoi Duong. All written parts of the paper were authored by Manh Khoi Duong with the supervision of Stefan Conrad.

Status: Published

Disparity measures are commonly used to assess the discrimination inherent in datasets or machine learning models. They particularly report differences between the outcomes of different social groups. These outcomes can be, for example, the acceptance rates of applicants for a job. However, disparity measures do not account for the number of samples in each group. Most often, they are calculated based on the number of positive outcomes in each group. This *frequentist approach* can lead to sampling errors if the number of samples in a group is small. Drawing conclusions from samples that do not represent the population can lead to under- or overestimation of discrimination.

In our paper, we refer to the outcomes of specific groups as group treatments. Using *Bayesian statistics* [29], we first model the posterior distributions of group treatments, which represent the uncertainties in their estimations. With the posteriors, it is possible to display the 95% credible intervals of group treatments. This already allows for a more comprehensive assessment of discrimination. Our paper goes one step further by introducing a utility function that maps the disparity and its uncertainty to a scalar value. This allows for the ranking of multiple machine learning models not only by their disparity but also by the certainty of their disparity.

Our paper comes with theoretical results and guarantees. The practical implications of our approach are demonstrated in the paper with a job hiring example. Essentially, we differentiate between decision-makers (e.g., machine learning models, humans) that have the same disparity but different uncertainties. For example, a recruiter who exhibits a disparity of 100% but discriminates against only a few applicants is preferred over a recruiter with the same disparity who discriminates against more applicants. This is because we are more certain about the disparity of the latter. The utility value of the former is higher than the utility value of the latter in our approach. The utility values represent the preferences we have over decision-makers.

ECAI 2024
U. Endriss et al. (Eds.)

doi:10.3233/FAIA240575

(Un)certainty of (Un)fairness: Preference-Based Selection of Certainly Fair Decision-Makers

Manh Khoi Duong[a,*] **and Stefan Conrad**[a]

[a]Heinrich Heine University, Universitätsstraße 1, 40225 Düsseldorf, Germany
ORCID (Manh Khoi Duong): https://orcid.org/0000-0002-4653-7685, ORCID (Stefan Conrad): https://orcid.org/0000-0003-2788-3854

Abstract. Fairness metrics are used to assess discrimination and bias in decision-making processes across various domains, including machine learning models and human decision-makers in real-world applications. This involves calculating the disparities between probabilistic outcomes among social groups, such as acceptance rates between male and female applicants. However, traditional fairness metrics do not account for the uncertainty in these processes and lack of comparability when two decision-makers exhibit the same disparity. Using Bayesian statistics, we quantify the uncertainty of the disparity to enhance discrimination assessments. We represent each decision-maker, whether a machine learning model or a human, by its disparity and the corresponding uncertainty in that disparity. We define preferences over decision-makers and utilize brute-force to choose the optimal decision-maker according to a utility function that ranks decision-makers based on these preferences. The decision-maker with the highest utility score can be interpreted as the one for whom we are most certain that it is fair.

1 Introduction

Traditional fairness metrics have played an important role in quantifying disparities between different social groups in data, machine learning predictions, and decision-making systems [27, 26, 5, 1]. However, they fail to address the inherent uncertainty present in real-world data, i.e., *aleatoric uncertainty*, particularly when minorities or generally data samples are underrepresented. Our work is motivated by comparing machine learning models regarding their fairness in any socially responsible application. We use the umbrella term *decision-maker* which can refer to any system or human that makes decisions based on data. Therefore, our work deals with both human and algorithmic decision-makers and is not limited to either of them. Still, for simplicity, our examples only involve humans.

We consider an illustrative scenario (see Figure 1) in a hiring setting in which two different companies, labeled A and B, sought to hire applicants. We also assume that all applicants in this scenario have equal qualifications and do not differ in any way except for the social group they belong to. Company A notably only accepted yellow candidates and rejected all blue candidates. Company B acted in the same way but received significantly fewer applications. When using *statistical disparity* [6, 4] to assess discrimination from both companies, we obtain the same score, which is 100%, signifying the disparity of the chances between yellow and blue candidates of getting accepted. Intuitively, we are more certain about the decisions being made by company A than company B. In the case of company B, the rejection of blue candidates can be attributed to random circumstances. In this case, we would judge company A as more discriminatory than company B because we are more certain that A is unfair and very uncertain about the unfairness of B. But if both companies accepted all applicants, the disparity would be 0%, and we would conversely judge B as more discriminatory than A. This is because we are certain that A is fair, while we are uncertain about the fairness of B. Lastly, when comparing between uncertain fair and uncertain unfair decision-makers, we would prefer the former over the latter. These examples underscore the importance of quantifying and assessing uncertainty in discrimination evaluations.

In the context of this example, we use the notation $A \succ B$ to signify a preference relation, indicating that company A is preferred over company B. The preferences we obtain are as follows:

$$\text{fair certain} \succ \text{fair uncertain}, \tag{1}$$
$$\text{fair uncertain} \succ \text{unfair uncertain}, \tag{2}$$
$$\text{unfair uncertain} \succ \text{unfair certain}, \tag{3}$$

where unfair and fair refer to a disparity of 100% and 0%, respectively. With these *trivial preferences*, following *research questions* arise:

- **RQ1:** How do we quantify the uncertainty of a decision-maker's (un)fairness?
- **RQ2:** How do we compare decision-makers that exhibit different levels of disparity and uncertainty on a continuous scale? How do we express preferences over them and rank them accordingly?
- **RQ3:** How do we select the optimal decision-maker according to our preferences?

We note that the task of selecting the most preferable decision-maker cannot be done by determining the Pareto front because uncertain cases can seem more or less fair than certain cases depending on the circumstances. This can be observed in the preferences (2) and (3). Furthermore, disparity and uncertainty are not necessarily discrete values, making it *non-trivial* to compare between decision-makers that are represented by them. To answer the research questions, our paper's structure and contributions are as follows:

- We first introduce a notation generalizing various group fairness criteria, eliminating the limitation to a particular group fairness criterion in our work.

[*] Corresponding Author. Email: duong@hhu.de.

- Using our notation, we demonstrate how to quantify uncertainties of group disparities exhibited by decision-makers using *Bayesian statistics* [13] (**RQ1**).
- Representing decision-makers by their disparate treatments of groups and the uncertainty of it, we formally define preferences over decision-makers (**RQ2**). By introducing a utility function, that assigns a value to each decision-maker, we are able to select the optimal decision-maker from a set of decision-makers (**RQ3**). The utility values allow ranking decision-makers according to our preferences.
- We evaluate our methodology on synthetic and real-world datasets to demonstrate its practical usability and necessity.
- We draw ethical conclusions by discussing the implications of our work and the importance of incorporating uncertainty in discrimination assessments.
- The implementation of the proposed scores and experiments can be found at https://github.com/mkduong-ai/fairness-uncertainty-score.

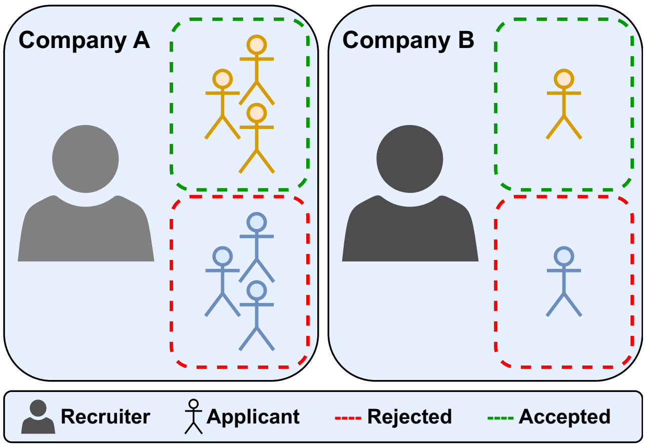


Figure 1. Both companies, A and B, discriminate against applicants by only accepting yellow candidates and rejecting all blue candidates. The statistical disparity score for both companies is 100%. Nevertheless, company B received fewer applicants, making its case more uncertain. Which decision-maker do we favor regarding fairness in such a case? What if both companies accepted all applicants? Who do we favor then?

2 Related Work

Fairness metrics have been widely studied in the literature [27, 26, 5, 1] and have been used to assess discrimination in various domains. Common are group fairness metrics that report *disparities* between social groups. In its simplest form, the disparity is calculated as the (absolute) difference between the outcomes of two groups, e.g., the acceptance rates between males and females in a job application process [4]. Any other probabilistic outcome can be used as well [6]. It gets more complex when more than two groups are involved. In this case, aggregating pairwise differences [27, 8] or similarly using *meta-metrics* [20] are common approaches. For example, the maximum disparity possible can be reported in such a case [27]. Depending on the aggregation method, the intended *social welfare* is different [8]. When dealing with multiple protected attributes, *subgroups* (white male, black woman, etc.) can be formed by the cross-product of the protected attributes. However, exponentially many subgroups can be formed in this way, and any classifier can be accused of discriminating against some subgroup. To prevent this, Kearns et al. [17] proposed to ignore subgroups that represent a small fraction of the population. Foulds et al. [11] generally criticized ignoring small subgroups, as minorities are often vulnerable to discrimination. Still, both works [17, 11] employ a disparity calculation to measure fairness. Hence, the disparity serves as a base for discrimination assessment in several problem settings.

However, relying solely on the disparity to assess discrimination can be problematic. Such a measurement can be uncertain, for instance, when samples underrepresent a population due to data sparsity [16, 22, 12]. Lum et al. [20] showed that meta-metrics are statistically biased upwards when more groups are involved. This effect is attributed to the increased number of comparisons between groups, which raises the likelihood of observing greater disparities. The authors combated this by deriving a correction term to *de-bias* the disparity. Foulds et al. [11] addressed a similar problem where they used a Bayesian approach to estimate the fairness of underrepresented, small subgroups. They did it as follows: Subgroups are essentially intersections of protected attributes, which can be represented by joint and marginal distributions. Under the frequentist perspective, empirical counts can be used to estimate such probabilities. This comes with disadvantages when the counts are small or when the subgroups are not present in the data. In such cases, the estimates are uncertain or undefined due to division by zero. Foulds et al. [11] proposed learning the marginal distribution with probabilistic models, allowing for uncertainty quantification. Singh et al. [24] and Tahir et al. [25] shared a similar concern about uncertainty in fairness assessments and argued that uncertainty can lead to unfairness.

We follow a similar strategy to Foulds et al. [11] in our work. We differ by allowing for a more general uncertainty quantification of fairness that is not limited to subgroups. Additionally, our uncertainty measure is normalized to ensure comparability. Upon quantifying the uncertainty, we express preferences over pairs of disparities and uncertainties, which is not done in the work by Foulds et al. [11]. Further, we combine the disparity and uncertainty into a single utility score, allowing for a straightforward comparison and ranking of decision-makers. The ranking reflects preferences over decision-makers, which we introduce in this work.

3 Preliminaries

Protected attributes such as ethnicity, nationality, and gender make individuals vulnerable to discrimination. We define Z, which represents a protected attribute, as a discrete random variable that can take on values from the set g. We refer to g as groups that are distinct categories an individual can belong to. For example, let Z represent the gender, then g is a set containing the genders male, female, and non-binary. Further, Y denotes the outcome of an individual, which is a binary random variable and $\hat{Y}$ is the predicted outcome. For both, Y and $\hat{Y}$, we use the values 1 and 0 to indicate positive and negative outcomes, respectively. We define E_1 and E_2 as events, which are subsets of the sample space Ω. The sample space Ω is the set of all possible outcomes of an experiment.

With the intention of avoiding limitations on a particular group fairness criterion, we introduce a generalized framework through the following definition:

Definition 1 (Treatment)**.** *We refer to the conditional probability of E_1 given that E_2 occurs and Z takes on the value $i \in g$, i.e.,*

$$P(E_1 \mid E_2, Z = i),$$

as the treatment *of group i.*

Using this notation, we can generalize group fairness notions that are based on conditional probabilities, including statistical parity, equality of opportunity, predictive parity etc. [4, 14, 26]. These criteria typically demand equal outcomes E_1 for different groups $i, j \in g$ given the same events E_2. Expressed with our notation, we yield:

$$P(E_1 \mid E_2, Z = i) = P(E_1 \mid E_2, Z = j). \tag{4}$$

In our case, E_1 often represents a dichotomous outcome, such as $Y = 1$ or $Y = 0$. Therefore, $P(E_1 \mid E_2, Z = i)$ can be interpreted as the success probability of a Binomial distribution. In the following, we demonstrate examples of common group fairness criteria expressed with our notation by only specifying E_1 and E_2.

Example 1 (Statistical Parity [4]). *Statistical parity requires equal positive outcomes between groups:*

$$P(Y = 1 \mid Z = i) = P(Y = 1 \mid Z = j),$$

where $i, j \in g$ represent different groups. To equivalently express it with our notation, we set $E_1 := (Y = 1)$ and $E_2 := \Omega$. By setting E_2 equal to the sample space, we compare the probabilities of the event $Y = 1$ across different groups without conditioning on any additional event.

Example 2 (Equality of Opportunity [14]). *To achieve equality of opportunity, we have to set $E_1 := (\hat{Y} = 1)$ and $E_2 := (Y = 1)$, which results in:*

$$P(\hat{Y} = 1 \mid Y = 1, Z = i) = P(\hat{Y} = 1 \mid Y = 1, Z = j).$$

This is equivalent to equal true positive rates across groups.

Example 3 (Predictive Parity [26]). *Predictive parity aims for equal predictive accuracy across different groups. To achieve this with our notation, we set $E_1 := (Y = 1)$, $E_2 := (\hat{Y} = 1)$ and yield:*

$$P(Y = 1 \mid \hat{Y} = 1, Z = i) = P(Y = 1 \mid \hat{Y} = 1, Z = j).$$

This is equivalent to equal positive predictive values across groups.

Because achieving equal probabilities for certain outcomes is not always possible due to variations in sample sizes in the groups, it is common to yield unequal probabilistic outcomes, even when the outcomes are similar. Hence, existing literature [27] use the absolute difference between the probabilities to quantify the strength of discrimination.

Definition 2 (Disparity). *We define the difference between the treatments of the groups $i, j \in g$ in the following:*

$$\delta_Z(i, j, E_1, E_2) = |P(E_1 \mid E_2, Z = i) - P(E_1 \mid E_2, Z = j)|,$$

and refer to it as disparity. *The* disparity *satisfies all properties of a mathematical metric regarding i, j and is also referred to as* fairness metric.

Higher differences indicate increased discrimination. Trivially, δ_Z is commutative regarding i, j. Establishing δ_Z provides a fundamental foundation for various scenarios. For instance, it allows us to aggregate pairwise differences between groups, particularly when dealing with attributes that are non-binary [27, 10, 7, 9].

4 Quantifying Uncertainty

As shown in Equation (4), we can describe fairness criteria by demanding equal treatments. However, the treatment of a group $i \in g$ can often exhibit uncertainty due to the limited number of samples. In this section, we contrast frequentist and Bayesian approaches to estimate the treatment probabilities $P(E_1 \mid E_2, Z = i)$. We then model the uncertainty of the disparity δ_Z using the variances of the posterior distributions. Finally, we define a decision-maker by its disparity and the corresponding uncertainty, enabling an enhanced discrimination assessment.

4.1 Estimating Treatment Probabilities

Earlier, we defined treatment as the probability of group $i \in g$ receiving some specific event E_1 given E_2. Let us consider the hiring process as an example again, then $P(E_1 \mid E_2, Z = i)$ could represent the chances of group i receiving a job offer E_1 under the condition of having a certain qualification E_2. This example depicts a Binomial distribution, where the outcome is binary. When having samples from the hiring process, we can denote the number of applicants in group i as:

$$n_i = |\{Z = i\} \cap E_2|, \tag{5}$$

and those of group i who received a job offer as:

$$k_i = |E_1 \cap \{Z = i\} \cap E_2|. \tag{6}$$

4.1.1 Frequentist Approach

In frequentist statistics, the probability of a Binomial distribution is estimated using empirical counts[1]. For shorthand, let's denote $p_i := P(E_1 \mid E_2, Z = i)$, then the estimate is given by:

$$\hat{p_i} = \frac{|E_1 \cap \{Z = i\} \cap E_2|}{|\{Z = i\} \cap E_2|} = \frac{k_i}{n_i}. \tag{7}$$

With more samples, the estimate becomes more accurate, i.e., $\lim_{n_i \to \infty} \hat{p_i} = p_i$. In practice, n_i can be small and therefore the estimate $\hat{p_i}$ can be quite different from the true probability p_i.

4.1.2 Bayesian Approach

In Bayesian statistics [13], the quantification of uncertainty involves modeling p_i as a random variable rather than setting it to a fixed constant as in Equation (7). We start with a *prior distribution* $p(p_i)$ that represents our beliefs before observing any data $\mathcal{D}$. When estimating parameters for a Binomial event, the Beta distribution, denoted with $\mathcal{B}(\alpha, \beta)$, is commonly used as the prior distribution [13]. Similarly to the Binomial distribution, it models binary outcomes. It does this with two shape parameters, α and β. To yield a non-informative uniform prior [13], both parameters are usually set with

$$\begin{aligned} \alpha_{\text{prior}} &= 1, \\ \beta_{\text{prior}} &= 1. \end{aligned} \tag{8}$$

This setting is motivated by the principle of indifference in Bayesian statistics and aligns with Laplace's rule of succession. In the next step, the prior distribution

$$p(p_i) = \mathcal{B}(\alpha_{\text{prior}}, \beta_{\text{prior}}) \tag{9}$$

[1] Maximum likelihood estimation

is updated. The updated distribution is known as the *posterior distribution* $p(p_i|\mathcal{D})$, which models the distribution of p_i after observing data $\mathcal{D}$ and represents our current beliefs.

According to Gelman et al. [13], the posterior can be obtained by adding the corresponding number of successes and failures to the shape parameters of the prior distribution. Specifically, the parameters for the posterior are:

$$\alpha^{(i)}_{\text{post.}} = \alpha_{\text{prior}} + k_i, \tag{10}$$

$$\beta^{(i)}_{\text{post.}} = \beta_{\text{prior}} + n_i - k_i.$$

With the posterior distributions:

$$p(p_i|\mathcal{D}) = \mathcal{B}(\alpha^{(i)}_{\text{post.}}, \beta^{(i)}_{\text{post.}}) \tag{11}$$

for each group $i \in g$ in hand, we can compare the group disparities more comprehensively. The posterior distributions allow us to derive alternative definitions for treatment and disparity. Since p_i and p_j are not single probabilities under this paradigm, the definitions of treatment and disparity undergo notational modifications.

Definition 3 (Bayesian Treatment)**.** *We denote the expected value of the posterior $p(p_i|\mathcal{D})$ as $\mathbb{E}(p_i|\mathcal{D})$. It is given by [13]:*

$$\mathbb{E}(p_i|\mathcal{D}) = \frac{\alpha^{(i)}_{\textit{post.}}}{\alpha^{(i)}_{\textit{post.}} + \beta^{(i)}_{\textit{post.}}}$$

and is the Bayesian estimate of p_i.

Definition 4 (Bayesian Disparity)**.** *Denoting p_i and p_j with the expected value, the Bayesian disparity $\delta^{\textit{(Bay.)}}_Z$ becomes:*

$$\delta^{\textit{(Bay.)}}_Z(i, j, E_1, E_2) = |\mathbb{E}(p_i|\mathcal{D}) - \mathbb{E}(p_j|\mathcal{D})|.$$

$\delta^{\text{(Bay.)}}_Z$ differs from δ_Z marginally if the number of samples is small. We leave the choice of the disparity definition to the user. We suggest using Bayesian disparity, if there is an initial belief that both groups have a 50% chance of receiving the favorable outcome. If such a belief is not present, the frequentist disparity is more suitable and less biased. We use the frequentist disparity in our work.

4.2 Modeling (Un)certainty of (Un)fairness

As seen in Figure 2, even with same frequentist treatments for group i and j (80%), the posterior distributions are vastly different. This is due to the different group sizes n_i and n_j and is signified by the variances of the posteriors. Hence, the variances of the posterior distributions describe the underlying uncertainties. We denote the variance with $\sigma^2_{\mathcal{B}}$ and it is defined by [13]:

$$\sigma^2_{\mathcal{B}}(\alpha, \beta) = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}. \tag{12}$$

Due to interpretability reasons, we aim to normalize the variance to the closed interval $[0, 1]$, where 0 represents no uncertainty and 1 represents maximum uncertainty. For this, it is essential to consider a few characteristics of the variance. Notably, $\sigma^2_{\mathcal{B}}$ is monotonically decreasing with respect to the shape parameters α and β, i.e., larger parameters lead to a smaller variance. Given that these shape parameters are natural numbers, the largest achievable variance of the posterior distribution, derived from Equation (11), is given by $\sigma^2_{\mathcal{B}}(1, 2)$, or equivalently $\sigma^2_{\mathcal{B}}(2, 1)$. We employ this maximum variance as a scaling factor, resulting in the following normalized variance $\hat{\sigma}^2_{\mathcal{B}}$:

$$\hat{\sigma}^2_{\mathcal{B}}(\alpha, \beta) := \frac{\sigma^2_{\mathcal{B}}(\alpha, \beta)}{\sigma^2_{\mathcal{B}}(1, 2)}. \tag{13}$$

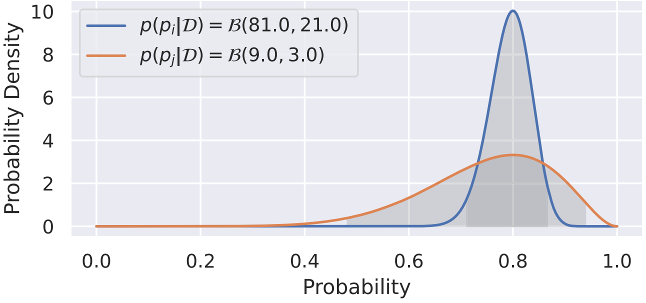


Figure 2. Group i comprises $n_i = 100$ individuals, with $k_i = 80$ receiving the favorable outcome, while group j consists of $n_j = 10$ individuals, of which $k_j = 8$ experience the favorable outcome. The figure displays the probability density functions of the posteriors. The filled areas mark the 95% credible intervals of each distribution. Noticeable, we are less certain about the data from group j. In frequentist statistics, both groups are treated equally, but the Bayesian approach enables differentiating the groups.

When comparing the disparities between two groups $i, j \in g$, we can use both normalized variances of the posteriors to obtain the uncertainty of the disparity and answer research question **RQ1** with the following definition.

Definition 5 (Uncertainty)**.** *We define the uncertainty of the disparity between two groups $i, j \in g$ as the mean of the normalized variances of their posterior distributions:*

$$\bar{\sigma}^2_{\delta_Z}(i, j, E_1, E_2) = \frac{\hat{\sigma}^2_{\mathcal{B}}(\alpha^{(i)}_{\textit{post.}}, \beta^{(i)}_{\textit{post.}}) + \hat{\sigma}^2_{\mathcal{B}}(\alpha^{(j)}_{\textit{post.}}, \beta^{(j)}_{\textit{post.}})}{2}.$$

By taking the average, the uncertainties from both groups are combined. A higher uncertainty score indicates a lower precision of the disparity estimate and vice versa. A maximum uncertainty of 1 is achieved if both groups consist of a single individual. We can now define a decision-maker by its disparity and the corresponding uncertainty in the following definition.

Definition 6 (Decision-Maker)**.** *A decision-maker $D \in [0, 1]^2$ is defined by its disparity and the corresponding uncertainty:*

$$D = (\delta_Z, \bar{\sigma}^2_{\delta_Z}).$$

5 Ranking Decision-Makers

In this section, we begin by defining preferences over decision-makers, establishing the criteria for what is deemed to be more or less fair. Subsequently, we formulate a utility function that maps decision-makers to values that represent the preferences and enables ranking, thus answering research question **RQ2**. A higher utility value indicates a more preferred decision-maker. To autonomously select the optimal decision-maker, we iterate through all candidates to find the decision-maker with the maximal utility value (**RQ3**). Additionally, we introduce the concept of indifference curves, offering insights into cases where two different decision-makers are equally preferred. The preference definitions in this section are mainly inspired by the work of Levin and Milgrom [19] and were adapted to fit our context.

5.1 Preferences

We recall the preferences (1)-(3) from Section 1 we have over decision-makers. We first introduce the definition of a preference relation and then define the preferences (1)-(3) formally using the definition of a decision-maker.

Definition 7 (Preference Relation)**.** *We denote a strict preference relation with $\succ$ or $\prec$ and write $D_1 \succ D_2$ to signify that decision-maker D_1 is preferred over D_2. The symbol $\sim$ denotes indifference, i.e., $D_1 \sim D_2$ means that D_1 and D_2 are equally preferred. The strict preference relation is transitive, while the indifference relation is reflexive and transitive.*

Definition 8 (Trivial Preferences)**.** *We have following preferences over decision-makers:*

$$\begin{aligned} \textit{fair certain} \succ \textit{fair uncertain}: &\quad (0,0) \succ (0,1)\\ \textit{fair uncertain} \succ \textit{unfair uncertain}: &\quad (0,1) \succ (1,1)\\ \textit{unfair uncertain} \succ \textit{unfair certain}: &\quad (1,1) \succ (1,0)\end{aligned}$$

Due to transitivity, we can derive additional preferences:

$$\begin{aligned} \textit{fair certain} \succ \textit{unfair uncertain}: &\quad (0,0) \succ (1,1)\\ \textit{fair certain} \succ \textit{unfair certain}: &\quad (0,0) \succ (1,0)\\ \textit{fair uncertain} \succ \textit{unfair certain}: &\quad (0,1) \succ (1,0)\end{aligned}$$

The listed preferences are trivial and extreme cases, where a decision-maker is characterized by extreme instances of (un)fairness and (un)certainty, i.e., $D \in \{0,1\}^2$. We note that listing all preferences over decision-makers, as defined in Definition 8, is impossible because infinite decision-makers exist in the continuous space, thus making the preference relation incomplete. We call any preference that is not trivial a *non-trivial preference*.

Definition 9 (Non-Trivial Preference)**.** *$D_1 \succ D_2$ is a non-trivial preference if and only if $D_1, D_2 \in]0,1[^2$.*

Modeling non-trivial preferences can be challenging as we are comparing decision-makers that are neither extremely fair, unfair, certain, nor uncertain. However, it is possible to infer non-trivial preferences from the trivial ones, as we will show in the next section.

5.2 Ranking with Utility Values

By introducing a utility function u, we can translate preferences over decision-makers into utility values that enable proper comparison, i.e.,

$$D_1 \succ D_2 \implies u(D_1) > u(D_2). \tag{14}$$

Importantly, the utility function must satisfy all trivial preferences from Definition 8. However, this still leaves us open with infinitely many decision-makers that are not covered by the defined preferences, specifically for any $D \in]0,1[$. Therefore, we need to define a utility function that is able to assign a value to all possible decision-makers. By doing so, we can rank all decision-makers accordingly to the defined preferences and the undefined, non-trivial preferences. For the latter, we assume that these preferences can be implied from the utility:

$$D_1 \succ D_2 \impliedby u(D_1) > u(D_2). \tag{15}$$

Definition 10 (Utility Function)**.** *Let $\mathbb{D} = [0,1]^2$ be the set of all decision-makers, a utility function $u : \mathbb{D} \to \mathbb{R}$ is total and must fulfill all preferences from Definition 8, that is:*

$$\begin{aligned} u(0,0) &> u(0,1)\\ u(0,1) &> u(1,1)\\ u(1,1) &> u(1,0),\end{aligned}$$

including all derived preferences due to transitivity.

By demanding totality, we ensure that the utility function is able to assign a value to every decision-maker $D \in [0,1]^2$. A possible utility function is given by the following example.

Example 4 (TOPSIS Utility)**.** *Motivated by TOPSIS [15], we calculate the utility of decision-makers based on their distances to the ideal solution $(0,0)$ and the worst solution $(1,0)$. Because utility is to be maximized, distances should be penalized accordingly. We define $u_{topsis} : [0,1]^2 \to [-1,1]$ with:*

$$\begin{aligned} u_{topsis}(\delta_Z, \bar{\sigma}^2_{\delta_Z}) &= \left\|(\delta_Z, \bar{\sigma}^2_{\delta_Z}) - (1,0)\right\|_2 - \left\|(\delta_Z, \bar{\sigma}^2_{\delta_Z}) - (0,0)\right\|_2\\ &= \sqrt{(\delta_Z - 1)^2 + (\bar{\sigma}^2_{\delta_Z})^2} - \sqrt{(\delta_Z)^2 + (\bar{\sigma}^2_{\delta_Z})^2}.\end{aligned}$$

Theorem 1. *u_{topsis} is a utility function as it is total, fulfills all preferences from Definition 8, and preserves the transitive preferences.*

Proof. Trivial. u_{topsis} is total by definition, i.e., a value $u_{\text{topsis}}(D)$ exists for all $D \in \mathbb{D}$. Next, input the values from Definition 8 and show that all preferences including the transitive ones hold. □

The idea behind u_{topsis} is that the decision-maker that is closer to the ideal decision-maker $(0,0)$ and farther away from the worst decision-maker $(1,0)$ is rewarded with a higher utility value. The utility function is not unique and can be replaced by any other function fulfilling the requirements from Definition 10. Since we modeled the utility function in Example 4 to favor certainly fair decision-makers and disfavor certainly unfair ones, we can be sure that any decision-maker with a higher utility value is more preferred than any other by rational users that have the same preferences as in Definition 8.

Because a normalized score is more intuitive, stakeholders might prefer to use the utility function from the following example.

Example 5 (Normalized Utility)**.** *We define a normalized utility function $u_{norm} : [0,1]^2 \to [0,1]$ with:*

$$u_{norm}(\delta_Z, \bar{\sigma}^2_{\delta_Z}) = \frac{u_{topsis}(\delta_Z, \bar{\sigma}^2_{\delta_Z}) + 1}{2}. \tag{16}$$

Theorem 2. *u_{norm} is a utility function as it is total, fulfills all preferences from Definition 8, and preserves the transitive preferences.*

Proof. Trivial. Apply the same steps as in the proof of Theorem 1. □

5.3 Objective Function and Selecting Optimal Decision-Maker

Let us have a set of decision-makers $D = \{D_1, D_2, \ldots, D_m\}$, then the approach to choose the optimal decision-maker D^* is given by solving the following optimization problem:

$$D^* = \underset{D_i \in D}{\operatorname{argmax}} \quad u(D_i).$$

For a finite set of decision-makers, this can be solved efficiently with brute-force search in $\mathcal{O}(m)$.

5.4 Indifference Curve

When two decision-makers have the same utility, they are indifferent to each other, i.e., $D_1 \sim D_2$. In such cases, the user is left with free choices to select their optimal decision-maker. All points having the same utility value lie on an indifference curve. It can be derived by solving the following equation:

$$u(D_1) = u(D_2). \tag{17}$$

Let us denote $D_1 = (a_1, a_2)$, $D_2 = (b_1, b_2)$, then we specifically solve:

$$\sqrt{(a_1-1)^2 + a_2^2} - \sqrt{a_1^2 + a_2^2} = \sqrt{(b_1-1)^2 + b_2^2} - \sqrt{b_1^2 + b_2^2}. \tag{18}$$

Depending on which variable (a_1, a_2, b_1, b_2) is treated as a constant, the analytical solution can become excessively long. We did find such solutions for the indifference curve with symbolic computation [21], but they are not insightful. We found a trivial solution with:

$$u(D_1) = u(D_2) = 0. \tag{19}$$

For this case, the curve is given when $a_1 = b_1 = 0.5$ and a_2, b_2 can be any value in $[0, 1]$. This means that decision-makers are indifferent as long as their disparities are both 50%. Utility values are also negative if the disparity is higher than 50% and positive if it is lower.

6 Experiments

Before diving into the experiments, we revisit the example from Figure 1. We calculate the disparity and uncertainty for the two recruiters, A and B, and list the utility values using u_{topsis} in Table 1. When comparing the disparities, both recruiters are indifferent as they are equally unfair towards group j. According to the utility values, recruiter B has a higher utility than A and is therefore more preferred. This aligns with the intuition that we are more uncertain about B's unfairness than A's.

Table 1. Revisiting example given in Figure 1.

Recruiter	n_i	k_i	n_j	k_j	$\hat{p}_i$	$\hat{p}_j$	DM $(\delta_Z, \bar{\sigma}^2_{\delta_Z})$	Utility
A	3	3	3	0	100%	0%	(1.000, 0.480)	-0.629
B	1	1	1	0	100%	0%	(1.000, 1.000)	-0.414

To explore our methodology more extensively, we conduct experiments on synthetic and real-world datasets. We use synthetic data to have full control over the disparities and uncertainties of decision-makers. This is done by setting different group treatments and varying the group sizes.

6.1 *Synthetic Data*

We first generate group sizes $(n_i, n_j) \in \{1, 5, 10, 50\}^2$. Each group $i \in g$ can receive any number of favorable outcomes k_i based on its size n_i. For example, if $n_i = 5$, then k_i can be any natural number in $[0, 5]$. Decision-makers are then created by calculating the disparity and uncertainty through all possible combinations of group sizes and treatments. This results in 4 900 decision-makers. We then calculate the utility value using u_{topsis} for each decision-maker.

Table 2. Four decision-makers with the highest and lowest utility values from the synthetic data created in the experiments.

Rank	n_i	k_i	n_j	k_j	$\hat{p}_i$	$\hat{p}_j$	DM $(\delta_Z, \bar{\sigma}^2_{\delta_Z})$	Utility
1	50	50	50	50	100%	100%	(0.000, 0.006)	0.994
2	50	0	50	0	0%	0%	(0.000, 0.006)	0.994
3	50	49	50	49	98%	98%	(0.000, 0.013)	0.988
4	50	1	50	1	2%	2%	(0.000, 0.013)	0.988
4897	50	0	50	49	0%	98%	(0.980, 0.009)	-0.958
4898	50	49	50	0	98%	0%	(0.980, 0.009)	-0.958
4899	50	50	50	0	100%	0%	(1.000, 0.006)	-0.994
4900	50	0	50	50	0%	100%	(1.000, 0.006)	-0.994

We list four decision-makers with the highest and lowest utility values from the synthetic data in Table 2. The most favorable decision-makers, with the same highest utility values, are those where all individuals from both groups either receive the favorable or unfavorable outcome, i.e., $k_i, k_j \in \{0, n_i\}$ with $k_i = k_j$. Groups are essentially treated equally and consist of large sample sizes. The least favorable decision-makers are the ones, where the disparity is maximized and the uncertainty is lowest. This aligns with the intuition that decision-makers, where we know that they are without a doubt unfair, are less preferred.

6.2 *COMPAS Dataset*

We use the COMPAS [18] dataset to evaluate decision-makers. The dataset contains information about defendants and their criminal histories. We compare different machine learning models, namely *Logistic Regression* (LR), *Support Vector Machine* (SVM), *Random Forest* (RF), and *k-Nearest Neighbors* (KNN), that predict whether a defendant will be rearrested within two years. These models act as decision-makers in our context. The dataset consists of 7 214 samples, and we use an 80/20 split for training and testing. Different from the processed versions of COMPAS in other fairness libraries [2, 3], the protected attribute '*race*' has not been reduced to two categories but is utilized in its original form. To calculate the disparity for this, we report the following difference [27, 3]:

$$\delta_Z = \max_{i \in g} P(\hat{Y} = 0 \mid Z = i) - \min_{j \in g} P(\hat{Y} = 0 \mid Z = j), \tag{20}$$

where $\hat{Y} = 0$ is the predicted outcome on the test set, noting that it is considered the favorable outcome as it indicates that a defendant will not be rearrested. Using this formula, the most and least privileged groups can differ for each model.

Table 3 displays the results of the experimentation on the COMPAS dataset. The models are ranked based on their utility values with u_{topsis}. We also report the accuracy of each model. The Logistic Regression model has the highest utility value and is therefore the most preferred. Interestingly, we observed that Asians are always the most privileged group, while Native Americans are always the least privileged group. Nearly all Asians receive a favorable outcome, while only a few Native Americans do. Both groups come with a small sample size and are therefore associated with high uncertainty. In this real-world scenario, ranking models by their utility values aligns with ranking them by the disparity δ_Z. This is because the utility function is designed to favor decision-makers with lower disparities. However, utility values contain information about the uncertainty of the disparities. Moreover, as illustrated in the example from Table 3, utility values are essential for distinguishing between decision-makers who exhibit the same level of disparity. In cases where both disparity and uncertainty are equal, the utility values are also the same. This is the case for the SVM and RF models in our experiment. For this, we advise considering the accuracy of the models as well. To conclude, LR has the highest utility value and accuracy, making it the most suitable model for recidivism prediction in this case.

6.3 *Summary of Results*

Our work addresses three key research questions. Firstly, we establish a method to distinguish between decision-makers exhibiting

Table 3. Results from the COMPAS dataset.

Model	Most Privileged (i)	Least Privileged (j)	n_i	k_i	n_j	k_j	$\hat{p}_i$	$\hat{p}_j$	DM ($\delta_Z, \bar{\sigma}^2_{\delta_Z}$)	Utility	Accuracy
LR	Asian	Native American	6	6	4	2	100%	50%	(0.500, 0.431)	0	72%
KNN	Asian	Native American	6	5	4	0	83.33%	0%	(0.833, 0.366)	-0.508	66.81%
SVM	Asian	Native American	6	6	4	0	100%	0%	(1.000, 0.288)	-0.753	71.10%
RF	Asian	Native American	6	6	4	0	100%	0%	(1.000, 0.288)	-0.753	70.20%

the same levels of discrimination by integrating uncertainty into our analysis (**RQ1**). This involves modeling the uncertainty of the measured disparity of outcomes between groups. Using both disparity and uncertainty, we define a decision-maker and establish our preferences among them. Secondly, to compare decision-makers within the continuous space of preferences, we introduce a utility function that evaluates each candidate. The utility values are then used to rank all decision-makers according to the defined preferences (**RQ2**). Lastly, to identify the optimal decision-maker, we introduce an optimization objective, allowing us to select the most suitable candidate, thus addressing **RQ3**. The synthetic and real-world experiments demonstrate the practical usability and necessity of our methodology to reliably assess the fairness of decision-makers.

7 Discussion

While we answered all research questions prior, we want to discuss several aspects of our methodology, including the scope of our work, in this section.

It is important to model the utility function in such a way that it reflects the user's preferences. This is because non-trivial preferences are implied by the utility function. Here, we refer the reader to methods that map multiple criteria to a single value, such as TOPSIS [15] or the Analytic Hierarchy Process [23]. Ranking decision-makers based on the utility function is a good starting point to check if the preferences are correctly modeled.

Another important aspect is the indifference curve. We found that decision-makers are indifferent to each other as long as their disparities are both 50%. Here, the utility function is not sufficient to differentiate between decision-makers, and the choice is left to the user. We discourage choosing such a decision-maker where the uncertainty is close to zero. This is because 50% disparity is quite high in practice. Decision-makers with a higher level of uncertainty are more preferred in such cases.

Our methodology is not invulnerable to manipulation. For example, if a human decision-maker is aware of the internal workings of our method, he or she could artificially increase the uncertainty of their disparity to appear less discriminatory. In a hiring scenario, this can be done by generally rejecting candidates coming from a very marginalized group where the number of samples is small. In such a case, minority groups should be grouped together into one large group to avoid this kind of manipulation.

8 Conclusion

When dealing with small sample sizes, particularly in the case of minority groups, we are often uncertain about the collected data and the information derived from it. Group fairness metrics aim to report how different groups are treated based on some specified events and outcomes, disregarding uncertainty. Therefore, we first introduce a method utilizing Bayesian statistics to quantify the uncertainty of the disparity of group treatments and employ them to enhance the assessment of discrimination. With both the disparity and the uncertainty, we define decision-makers and derive preference relations over them. By introducing a utility function that aligns with these preferences and is defined for every possible decision-maker, we are able to select the most preferred decision-maker with the largest utility from a set of candidates using brute-force. Our methodology comes with proven guarantees, and we have demonstrated its behavior on synthetic and real-world datasets.

The implications of our work are noteworthy, as we are able to differentiate between systematic discrimination and random outcomes and have defined preferences in such cases. Decision-makers exhibiting discrimination on fewer samples are more preferred than those exhibiting discrimination on larger sample sizes. Similarly, a certainly fair decision-maker is preferred over an uncertainly fair decision-maker. The latter is when the decision-maker receives fewer samples. Our methodology can be used for a wide range of applications, including evaluating machine learning models as well as hiring and admission processes at companies and universities. Additionally, the utility function can also be incorporated into the loss function of a machine learning model to penalize decisions that are certainly unfair.

Ethics Statement

With our proposed utility score, we address the issue of reporting discrimination in uncertain cases. The proposed score can protect decision-makers from discrimination accusations when the disparity they exhibited is uncertain, while also ensuring that those who are clearly discriminatory are appropriately penalized. Consequently, the societal impact of our work is positive. Still, further research is needed to investigate the impact of our method on several real-world applications.

References

[1] S. Barocas, M. Hardt, and A. Narayanan. *Fairness and Machine Learning*. fairmlbook.org, 2019. URL http://www.fairmlbook.org.

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

[3] S. Bird, M. Dudík, R. Edgar, B. Horn, R. Lutz, V. Milan, M. Sameki, H. Wallach, and K. Walker. Fairlearn: A toolkit for assessing and improving fairness in AI. Technical report, Microsoft, May 2020. URL https://www.microsoft.com/en-us/research/publication/fairlearn-a-toolkit-for-assessing-and-improving-fairness-in-ai/.

[4] T. Calders, F. Kamiran, and M. Pechenizkiy. Building classifiers with independency constraints. In *2009 IEEE International Conference on Data Mining Workshops*, pages 13–18, 2009. doi: 10.1109/ICDMW.2009.83.

[5] S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, and A. Huq. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 797–806, 2017.

[6] N. Duan, X.-L. Meng, J. Y. Lin, C.-n. Chen, and M. Alegria. Disparities in defining disparities: statistical conceptual frameworks. *Statistics in Medicine*, 27(20):3941–3956, 2008.

[7] M. K. Duong and S. Conrad. Towards fairness and privacy: A novel data pre-processing optimization framework for non-binary protected attributes. In D. Benavides-Prado, S. Erfani, P. Fournier-Viger, Y. L. Boo, and Y. S. Koh, editors, *Data Science and Machine Learning*, pages 105–120. Springer Nature Singapore, 2023. ISBN 978-981-99-8696-5.

[8] M. K. Duong and S. Conrad. Towards fairness and privacy: A novel data pre-processing optimization framework for non-binary protected attributes. In D. Benavides-Prado, S. Erfani, P. Fournier-Viger, Y. L. Boo, and Y. S. Koh, editors, *Data Science and Machine Learning*, pages 105–120, Singapore, 2023. Springer Nature Singapore. ISBN 978-981-99-8696-5.

[9] M. K. Duong and S. Conrad. Measuring and mitigating bias for tabular datasets with multiple protected attributes. In *Proceedings of the 2nd Workshop on Fairness and Bias in AI co-located with 27th European Conference on Artificial Intelligence (ECAI 2024), Santiago de Compostela, Spain, October 20th, 2024*, CEUR Workshop Proceedings. CEUR-WS.org, 2024.

[10] M. K. Duong, J. Dunkelau, J. A. Cordova, and S. Conrad. Rapp: A responsible academic performance prediction tool for decision-making in educational institutes. In *BTW 2023*, pages 595–606. Gesellschaft für Informatik e.V., 2023. ISBN 978-3-88579-725-8. doi: 10.18420/BTW2023-29.

[11] J. R. Foulds, R. Islam, K. N. Keya, and S. Pan. *Bayesian Modeling of Intersectional Fairness: The Variance of Bias*, pages 424–432. SIAM, 2020. doi: 10.1137/1.9781611976236.48.

[12] P. Ganesh, H. Chang, M. Strobel, and R. Shokri. On the impact of machine learning randomness on group fairness. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 1789–1800, 2023.

[13] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. *Bayesian Data Analysis*. CRC press, 1995. doi: 10.1201/b16018.

[14] M. Hardt, E. Price, and N. Srebro. Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems*, 29, 2016.

[15] C.-L. Hwang and K. Yoon. *Methods for Multiple Attribute Decision Making*, pages 58–191. Springer Berlin Heidelberg, Berlin, Heidelberg, 1981. ISBN 978-3-642-48318-9. doi: 10.1007/978-3-642-48318-9_3.

[16] D. Ji, P. Smyth, and M. Steyvers. Can i trust my fairness metric? assessing fairness with unlabeled data and bayesian inference. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 18600–18612. Curran Associates, Inc., 2020.

[17] M. Kearns, S. Neel, A. Roth, and Z. S. Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In J. Dy and A. Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2564–2572. PMLR, July 2018.

[18] J. Larson, J. Angwin, S. Mattu, and L. Kirchner. Machine bias, May 2016. URL https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[19] J. Levin and P. Milgrom. Introduction to choice theory. https://web.stanford.edu/~jdlevin/Econ%20202/Choice%20Theory.pdf, 2004. Working Note, accessed on August 15, 2024.

[20] K. Lum, Y. Zhang, and A. Bower. De-biasing "bias" measurement. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 379–389, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393522. doi: 10.1145/3531146.3533105.

[21] A. Meurer, C. P. Smith, M. Paprocki, O. Čertík, S. B. Kirpichev, M. Rocklin, A. Kumar, S. Ivanov, J. K. Moore, S. Singh, T. Rathnayake, S. Vig, B. E. Granger, R. P. Muller, F. Bonazzi, H. Gupta, S. Vats, F. Johansson, F. Pedregosa, M. J. Curry, A. R. Terrel, v. Roučka, A. Saboo, I. Fernando, S. Kulal, R. Cimrman, and A. Scopatz. Sympy: symbolic computing in python. *PeerJ Computer Science*, 2017. ISSN 2376-5992. doi: 10.7717/peerj-cs.103.

[22] R. Rastogi and T. Joachims. Fairness in ranking under disparate uncertainty. *arXiv preprint arXiv:2309.01610*, 2024.

[23] T. L. Saaty. *The Analytic Hierarchy Process: Decision Making in Complex Environments*, pages 285–308. Springer US, Boston, MA, 1984. ISBN 978-1-4613-2805-6. doi: 10.1007/978-1-4613-2805-6_12.

[24] A. Singh, D. Kempe, and T. Joachims. Fairness in ranking under uncertainty. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 11896–11908. Curran Associates, Inc., 2021.

[25] A. Tahir, L. Cheng, and H. Liu. Fairness through aleatoric uncertainty. *CoRR*, abs/2304.03646, 2023.

[26] M. B. Zafar, I. Valera, M. Gomez Rodriguez, and K. P. Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 2017. ISBN 9781450349130. doi: 10.1145/3038912.3052660.

[27] I. Žliobaitė. Measuring discrimination in algorithmic decision making. *Data Mining and Knowledge Discovery*, 31:1060–1089, 2017.

4.2 Trusting Fair Data by Incorporating Data Quality Measures

---

Paper: Manh Khoi Duong and Stefan Conrad. Trusting Fair Data: Leveraging Quality in Fairness-Driven Data Removal Techniques. In *Big Data Analytics and Knowledge Discovery*, volume 14912 of *Lecture Notes in Computer Science*. Springer Cham, 2024.

Personal Contribution: Manh Khoi Duong developed the idea to enhance trust of `FairDo` [3]. The methodology of the research was implemented completely by Manh Khoi Duong. Manh Khoi Duong wrote all parts of the paper. Stefan Conrad supervised the work.

Remark: The following paper [5] is an extended version of the conference paper, containing additional details and results.

Status: Published

Pre-processing data for fairness is a double-edged sword. While such methods are able to mitigate bias, data quality and integrity often come with compromises. One of the most popular fairness Python libraries, `AIF360`, implements four pre-processing techniques to mitigate bias [33, 51, 25, 19], of which three transform and edit the features. Only one technique, `Reweighing` [33], assigns weights to the samples that are used in the training process. `Fairlearn` [16] is another Python library that provides a set of algorithms to mitigate bias. It includes a pre-processing technique called `CorrelationRemover` that transforms the features to remove the correlation with the sensitive attribute.

All of the aforementioned techniques (with the exception of `Reweighing`) have one common drawback: They alter the data in an uninterpretable way [2]. We argue that our methods [2, 3] are more interpretable because they are over- and undersampling techniques. Data is either added or removed, but features are not altered. This is a huge advantage because data points exist in the original or synthetic dataset (used for oversampling). This is not the case for feature transformation techniques. Still, specifically when removing data, a lot can happen, and certain guarantees are needed to ensure that the data can be trusted.

In this paper, we propose a modification of a fairness metric and introduce several novel optimization problems to enhance trust in fairness-driven data removal techniques. Specifically, our work guarantees that no protected group is completely removed from the dataset during the removal process. Additionally, users are introduced to a new objective, which is to minimize the number of removed samples. Similarly to the fairness-utility trade-off [18], users can now decide the extent to which they are willing to compromise fairness to keep the data intact. All methods are implemented and extended in the `FairDo` [3] package. The updated package includes new solvers for the stated multi-objective optimization problems.

Trusting Fair Data: Leveraging Quality in Fairness-Driven Data Removal Techniques*

Manh Khoi Duong[1] [0000−0002−4653−7685] and
Stefan Conrad[1] [0000−0003−2788−3854]

Heinrich Heine University, Universitätsstraße 1, 40225 Düsseldorf, Germany
{manh.khoi.duong, stefan.conrad}@hhu.de

Abstract. In this paper, we deal with bias mitigation techniques that remove specific data points from the training set to aim for a fair representation of the population in that set. Machine learning models are trained on these pre-processed datasets, and their predictions are expected to be fair. However, such approaches may exclude relevant data, making the attained subsets less trustworthy for further usage. To enhance the trustworthiness of prior methods, we propose additional requirements and objectives that the subsets must fulfill in addition to fairness: (1) group coverage, and (2) minimal data loss. While removing entire groups may improve the measured fairness, this practice is very problematic as failing to represent every group cannot be considered fair. In our second concern, we advocate for the retention of data while minimizing discrimination. By introducing a multi-objective optimization problem that considers fairness and data loss, we propose a methodology to find Pareto-optimal solutions that balance these objectives. By identifying such solutions, users can make informed decisions about the trade-off between fairness and data quality and select the most suitable subset for their application. Our method is distributed as a Python package via PyPI under the name `FairDo`[1].

Keywords: Fairness · Bias mitigation · Data quality · Coverage · AI Act.

1 Introduction

Machine learning models are often trained on biased data, which can lead to biased predictions [5]. A common approach to addressing fairness concerns is to use bias mitigation techniques. They can be categorized into pre-processing, in-processing, and post-processing [16]. Pre-processing techniques aim to remove bias from the training data before training a machine learning model, in-processing techniques modify the learning algorithm, and post-processing techniques adjust the predictions.

[*] Manuscript submitted to the 26th International Conference on Big Data Analytics and Knowledge Discovery (DaWaK 2024).

[1] https://github.com/mkduong-ai/fairdo

Because bias can be introduced at various stages during data preparation steps, pre-processing techniques can be integrated into the data preparation pipeline to ensure that the training data is fair. Some of the pre-processing techniques involve the removal of certain data points from the training set [20,8,9]. By removing certain data points, the machine learning model is trained on a fair subset and its predictions are expected to be fair as well. These techniques aim to fix representation bias in the data. While they tackle the root cause of the problem, they are deemed problematic as they may lead to the exclusion of relevant data.

In this paper, we explore multiple problems that can arise from the removal of data points and propose a decision-making methodology for selecting a subset that is more trustworthy for the user. One particular problem is the *removal of groups* as a whole, i.e., lack of *coverage*. For instance, the elimination of non-privileged groups can be considered fair by various fairness metrics, yet it fails to align with our main objective, which is to represent every group fairly in the resulting dataset. Removing entire groups can be seen as equivalent to underreporting discrimination. Another problem we tackle is the *amount of data removed*. When removing too many data points, the resulting dataset may not accurately represent the original data and data quality is compromised. Hence, our second objective is to retain as much data as possible. With this additional objective, there is a trade-off between fairness and the amount of data removed.

In summary, our main contributions are as follows:

- We present two additional criteria, *coverage* and *data loss*, to enhance trustworthiness for fairness-driven data removal techniques.
- We propose a multi-objective optimization problem that considers fairness and data loss. Using NSGA-II [6], we find Pareto-optimal solutions that the user can choose from.
- We provide an extensive and empirical evaluation of our proposed methodology on three real-world datasets (Adult [14], Bank [17], COMPAS [15]) and evaluate the attained subsets by training machine learning models on them. We assess the models' fairness and performances by comparing them to models trained on the original datasets.
- We publish our methods in an open-source and documented Python package `FairDo` that can be used on-the-fly to pre-process datasets. It comes with several tutorials and examples.

2 Related Work

While there are many bias mitigation techniques, and some of them are implemented in popular Python packages such as `AIF360` [1] and `Fairlearn` [2], the included pre-processing techniques are not able to deal with *non-binary groups* and at the same time transform the data in an *uninterpretable* way by editing features and labels. The package `FairDo` [9] aims to address these issues. It is a highly adaptive framework for removing data points to achieve fairness. In this framework, it is possible to deal with binary, non-binary groups, and multiple protected attributes if the fairness metric is defined accordingly. While the resulting fair datasets are more interpretable, as they are subsets of the original dataset, data removal can also be viewed critically. One can argue that their framework offers other solutions, such as adding synthetic data points to the original data for fairness. However, all solutions are based on data removal within their framework.

Drawing inspiration from prior research that has addressed group *representativeness* [7,19,4], we extend the methodology of Duong et al. [9] accordingly. There are many ways to define representativeness. The work of Stoyanovich et al. [19] explores two main definitions: *proportional representation* and *coverage*. Proportional representation ensures that the dataset contains a representative number of data points from each group. Coverage, on the other hand, ensures that the dataset covers the entire population. We aim for coverage in our work and argue that aiming for proportional representation is not feasible because the group counts are given beforehand and not much can be done about it during the data removal process. Coverage only requires each group to be represented at least once in the dataset and is hence a more relaxed constraint.

The work of Catania et al. [4] is also related to our work. They propose a constraint-based optimization approach for their problem, which is to mitigate biases in datasets during a selection-based query. We deal with selecting a subset that optimizes a certain fairness objective and do not consider being in a query setting.

3 Preliminaries

Following definitions are primarily based on the work of Žliobaitė [22] and taken from Duong et al. [9]. We applied minor modifications to fit the definitions to the context of this paper.

3.1 Measuring Discrimination

Protected attributes such as race, gender, and nationality make individuals vulnerable to discrimination. Generally, we use Z to represent a protected attribute and Y to denote the outcome for an individual. Formally, we define Z and Y as discrete random variables. Z can take on values from the sample space g, which represents social groups such as male, female, and non-binary. For Y, we use the values 1 and 0 to indicate positive and negative outcomes, respectively. Further, we denote z_i and y_i to refer to the values of the i-th individual.

Definition 1 (Dataset). *We define a dataset $\mathcal{D}$ as a set of data points d_i:*

$$\mathcal{D} = \{d_i\}_{i=1}^{n},$$

where n is the number of data points in the dataset. A data point can be defined as a triplet (x_i, z_i, y_i), where x_i is the feature vector, z_i is the protected attribute, and y_i is the outcome.

Fairness criteria are often based on conditional probabilities, and typically demand some equal outcome between groups [13,21]. One of the most common fairness criteria is *statistical parity* [3].

Definition 2 (Statistical Parity [3]). *Statistical parity requires equal positive outcomes between groups:*

$$P(Y = 1 \mid Z = i) = P(Y = 1 \mid Z = j),$$

where $i, j \in g$ represent different groups.

Typically, the probabilities are estimated using sample statistics. Because achieving equal probabilities for certain outcomes is not always possible, existing literature [22] present measures to quantify the level of discrimination.

Definition 3 (Statistical Disparity [22]). Statistical disparity *is defined as the absolute difference between the probabilities of the positive outcome $Y = 1$ between two groups $i, j \in g$:*

$$\delta_Z(i, j) = |P(Y = 1 \mid Z = i) - P(Y = 1 \mid Z = j)|.$$

Establishing δ_Z provides a fundamental foundation for various scenarios. For instance, it allows us to aggregate pairwise differences between groups, particularly when dealing with attributes that are non-binary [22]. This allows us to quantify discrimination for more than two groups.

Definition 4 (Disparity for Non-binary Groups). *We introduce an aggregate function* $\text{agg}^{(1)}$ *as a function that takes a set of values and returns a single value. The function* $\text{agg}^{(1)}$ *can represent, for example, the sum or maximum function. With* $\text{agg}^{(1)}$*, we can compute the discrimination for a single protected attribute Z with any amount of groups. Simplifying notation, we write $\psi(\mathcal{D})$ to represent the discrimination measure for a dataset $\mathcal{D}$:*

$$\psi(\mathcal{D}) = \underset{i,j \in g, i \neq j}{\text{agg}^{(1)}} \delta_Z(i, j).$$

Example 1 (Maximal Statistical Disparity). The maximal statistical disparity is defined as:

$$\psi_{\text{SDP-max}}(\mathcal{D}) = \max_{i,j \in g, i \neq j} \delta_Z(i, j).$$

It describes the maximum discrimination obtainable between two groups.

There are many ways to measure discrimination in a dataset. In this paper, we focus on the maximal statistical disparity as it provides an interpretable measure of discrimination and is recommended in the work of Žliobaitė [22]. Still, our framework can be extended to other measures as well, that is, any ψ that maps a dataset to a positive value. Our framework only assumes that the objective is to minimize ψ. We use the term *discrimination measure* and *fairness metric* interchangeably to refer to ψ.

3.2 Fair Subset Selection

Duong et al. [9] proposed a framework for removing discriminating data points from a given dataset $\mathcal{D} = \{d_i\}_{i=1}^{n}$. They stated the problem as finding a subset $\mathcal{D}_{\text{fair}} \subseteq \mathcal{D}$, which minimizes the discrimination in that subset. This describes following combinatorial optimization problem:

$$\min_{\mathcal{D}_{\text{fair}} \subseteq \mathcal{D}} \quad \psi(\mathcal{D}_{\text{fair}}). \tag{1}$$

To make the problem solvable, the authors [9] introduced a binary decision variable b_i for each sample $d_i \in \mathcal{D}$, where $b_i = 1$ indicates if the sample d_i is included in the fair subset $\mathcal{D}_{\text{fair}}$ and 0 otherwise. More formally $\mathcal{D}_{\text{fair}}$ is defined as:

$$\mathcal{D}_{\text{fair}} = \{d_i \in \mathcal{D} \mid b_i = 1\}. \tag{2}$$

Defining $\mathbf{b} = (b_1, b_2, \ldots, b_n) \in \{0,1\}^n$ as a solution vector, finding the optimal subset $\mathcal{D}_{\text{fair}}$ is equivalent to solving for the optimal binary vector $\mathbf{b}^*$:

$$\mathbf{b}^* = \operatorname*{argmin}_{\mathbf{b} \in \{0,1\}^n} \quad \psi(\{d_i \in \mathcal{D} \mid b_i = 1\}) \tag{3}$$

Because finding the exact optimal solution $\mathbf{b}^*$ is an NP-hard problem if ψ is treated as a black-box, the authors [9] employed genetic algorithms to heuristically solve the problem.

4 Enhancing Trust in Fair Data

When having a fair dataset, we want to ensure that the dataset is still faithful to its original version. To enhance the trustworthiness, we introduce two additional criteria. The two criteria roughly describe a form of quality assurance for the fair subset. Overall, we have three criteria:

- **Fairness**: We want to minimize the discrimination in the attained subset [9].
- **Coverage**: All groups must be included in the fair subset.
- **Data Loss**: The fair subset should resemble the original dataset by retaining as much data as possible.

4.1 Coverage

When we compare the discrimination scores between two datasets, one could naively assume that the dataset with the lower score should be preferred over the other. However, simply comparing the discrimination scores is not sufficient.

Table 1: A dataset and two of its possible subsets $\mathcal{D}_{\text{fair}}, \mathcal{D}_{\text{cov.}} \subseteq \mathcal{D}$. Both subsets achieve perfect fairness scores but only $\mathcal{D}_{\text{cov.}}$ satisfies coverage.

Table 2: $\mathcal{D}$

d_i	$\mathbf{z}$	$\mathbf{y}$
1	0	1
2	1	1
3	2	0
4	2	1

Table 3: $\mathcal{D}_{\text{fair}}$

d_i	$\mathbf{z}$	$\mathbf{y}$
1	0	1
2	1	1
3	2	0
4	2	1

Table 4: $\mathcal{D}_{\text{cov.}}$

d_i	$\mathbf{z}$	$\mathbf{y}$
1	0	1
2	1	1
3	2	0
4	2	1

Example Table 2 represents the original dataset $\mathcal{D}$, and Table 3 depicts a subset $\mathcal{D}_{\text{fair}} \subseteq \mathcal{D}$, purposely selected to achieve fairness. The discrimination scores are $\psi_{\text{SDP-max}}(\mathcal{D}) = 0.5$ and $\psi_{\text{SDP-max}}(\mathcal{D}_{\text{fair}}) = 0$. Despite $\mathcal{D}_{\text{fair}}$ yielding a perfect fairness score, group 2 is missing in that set. Another fair subset $\mathcal{D}_{\text{cov.}} \subseteq \mathcal{D}$ is shown in Table 4, which includes all groups and hence satisfies coverage. It also achieves a perfect fairness score, $\psi_{\text{SDP-max}}(\mathcal{D}_{\text{cov.}}) = 0$. We even argue that any dataset $\mathcal{D}_{\text{cov.}}$ satisfying coverage is more preferred than any other dataset that does not, regardless of their fairness scores.

Incorporating Coverage We want to construct a fairness metric $\hat{\psi}$ that reflects our preference regarding coverage: The subset that satisfies coverage is always preferred over the subset that does not. However, if both subsets satisfy coverage, we want to compare them based on the fairness metric ψ. Thus, a penalty is only applied if a group is missing. In this case, $\hat{\psi}$ must have a higher value than the maximum discrimination achievable in ψ to enforce the preference.

Definition 5 (Penalized Discrimination). *The highest disparity possible for $\psi_{SDP\text{-}max}$ is 1. Let $|g_m|$ be the number of missing groups and $\epsilon > 0$, then we penalize $\psi_{SDP\text{-}max}$ as follows to enforce preferring coverage over non-coverage:*

$$\hat{\psi}_{SDP\text{-}max}(\mathcal{D}) = \max(\psi_{SDP\text{-}max}(\mathcal{D}), [|g_m| > 0] \cdot (1 + \epsilon)),$$

where $[|g_m| > 0]$ is an indicator function, which returns 1 if $|g_m| > 0$ and 0 otherwise. Setting ϵ to any positive value ensures that the penalty is higher than the maximum discrimination score.

4.2 Data Loss

By data loss, we refer to the similarity between the fair subset and its original. There are several ways to measure data loss by this means. But some methods require knowledge of the true underlying distribution of the data, which has to be expensively estimated. Therefore, we use an efficient measure, which is the relative amount of data removed.

Definition 6 (Data Loss). *The relative amount of the data removed is given by:*

$$\mathcal{L}(\mathcal{D}, \mathcal{D}_{fair}) = 1 - \frac{|\mathcal{D}_{fair}|}{|\mathcal{D}|},$$

where $\mathcal{D}_{fair} \subseteq \mathcal{D}$. A lower value indicates less data is removed, and therefore is better.

5 Optimization Objectives

We have three objectives to optimize for: *fairness*, *coverage*, and *data loss.* Two of the objectives, fairness and coverage, can be combined into a single objective, as shown in Definition 5. The third objective, data loss, can be treated in multiple ways, which we will discuss in the following.

5.1 Multi-objective Optimization

If there are no preferences provided regarding fairness and data loss, we have to treat the problem as a multi-objective optimization problem. The aim is to minimize both discrimination $\hat{\psi}$ and data loss $\mathcal{L}$. The optimization problem is written as follows:

$$\min_{\mathcal{D}_{\text{fair}} \subseteq \mathcal{D}} \quad (\hat{\psi}(\mathcal{D}_{\text{fair}}), \mathcal{L}(\mathcal{D}, \mathcal{D}_{\text{fair}})). \tag{4}$$

Solvers for multi-objective optimization problems aim to find the Pareto front, which is the set of solutions that are not dominated by any other solution. A solution is dominated if there is another solution that is better in at least one objective and not worse in any other objective.

5.2 Single-objective Optimization

If the importance of fairness and data loss is known beforehand, the problem can be transformed into a single-objective optimization problem by using the weighted sum of the objectives:

$$\min_{\mathcal{D}_{\text{fair}} \subseteq \mathcal{D}} \quad \alpha\hat{\psi}(\mathcal{D}_{\text{fair}}) + (1-\alpha)\mathcal{L}(\mathcal{D}, \mathcal{D}_{\text{fair}}), \tag{5}$$

where $\alpha \in [0, 1]$ is a weighting factor that determines the importance of fairness over data loss. A value of $\alpha = 0.5$ indicates that both objectives are equally important and is set as the default value in our experiments. When α values are set lower, the user prioritizes data fidelity more than fairness.

However, both objectives do not necessarily map to the same scale and therefore require normalization to make them comparable. Specifically, $\hat{\psi}$ requires normalization depending on the fairness metric used. Introducing β as the normalization factor, the single-objective optimization problem is then:

$$\min_{\mathcal{D}_{\text{fair}} \subseteq \mathcal{D}} \quad \frac{\alpha\hat{\psi}(\mathcal{D}_{\text{fair}})}{\beta} + (1-\alpha)\mathcal{L}(\mathcal{D}, \mathcal{D}_{\text{fair}}). \tag{6}$$

There are two meaningful choices for β: We either care about the *absolute* or *relative discrimination* score compared to the original dataset. For the absolute score, β is set as the theoretical maximum value of $\hat{\psi}$. In the case of $\hat{\psi}_{\text{SDP-max}}$, setting $\beta = 1$ or $\beta = 1+\epsilon$ are both viable and similar options if ϵ is small enough. For the relative score, β is set as the discrimination score of the original dataset, $\beta = \hat{\psi}(\mathcal{D})$. Hence, any discrimination score under 1 implies a reduction of discrimination. The score can be interpreted as the percentage of discrimination removed or added.

We note that normalizing is not required in the multi-objective optimization approach as the objectives are treated separately and the used heuristic compares candidate solutions based on the Pareto order. Only selecting a single solution from the Pareto front requires weighting the objectives.

6 Heuristics

To make the framework flexible and agnostic to the fairness metric, we have to use heuristics that only require function evaluations. For this, we use genetic algorithms to solve the optimization problems in Equations (4) and (6). To solve the single-objective problem, we use the genetic algorithm borrowed from Duong et al. [9]. For the multi-objective optimization problem, we use our own modified version of the NSGA-II [6] algorithm as described in the following.

6.1 NSGA-II Modification

We added the possibility to select between methods that initialize the population in our modified NSGA-II [6] algorithm. We created our own initializer that initializes the population with variable bias.

Generally, all employed algorithms operate on a population of solutions. In our implementation, the population is encoded as a binary matrix $\mathbf{P} \in \{0,1\}^{M \times n}$, where each row $\mathbf{b}_i \in \{0,1\}^n$ represents a solution, i.e.,

$$\mathbf{P} = (\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_M)^\top, \tag{7}$$

and M is the population size.

Random Initializer A common approach to initialize a population is to randomly assign each entry in the binary matrix $\mathbf{P}$ to 0 or 1 with a certain probability p [10]. Usually, the probability is set to 0.5. Trivially, for all $\mathbf{b}_i \in \mathbf{P}$, the expected number of 1s is:

$$\mathbb{E}(\mathbf{b}_i) \to \frac{n}{2}, \tag{8}$$

as n approaches infinity. This implies that each subset is expected to be half the size of the original dataset.

Hence, this initialization method is not suitable for our problem, as we aim to minimize the number of removed data points and require a diverse population.

Variable Initializer To address the issues with the prior method, we propose an initializer, which creates individuals with varying probabilities p. The user can specify the range of the probabilities $[p_{\min}, p_{\max}]$, and the initializer will create individuals with probabilities evenly distributed within the specified range. This leads to a more diverse population, where some individuals have less data removed than others. The default range is set to $[p_{\min}, p_{\max}] = [0.5, 0.99]$.

7 Evaluation

To evaluate the proposed framework, we conducted multiple experiments. Following research questions guided our evaluation:

- **RQ1** Which configuration of genetic operators is best suited for the NSGA-II algorithm in the context of bias mitigation in datasets?
- **RQ2** What is the impact of pre-processed datasets on the fairness and performance of machine learning models, as compared to models trained on unprocessed data?

Each following subsection corresponds to one of the research questions. Each experiment was conducted on the same objectives, datasets, and other settings as listed below. The specific details for each experiment are detailed in the corresponding subsections.

Objectives The objectives are $\hat{\psi}_{\text{SDP-max}}$ (Definition 5) with $\epsilon = 0.01$ and $\mathcal{L}$ (Definition 6). Both are to be minimized.

Datasets We conducted all experiments on three popular datasets in the fairness literature: Adult [14], Bank [17], and Compas [15], providing a comprehensive examination across various domains. They all serve as baselines for our experiments.

Trials We conducted 10 trials for each configuration in the experiments to ensure the reliability of our results.

7.1 Hyperparameter Optimization

The aim is to find solutions for the optimization problem in Equation (4) using the NSGA-II algorithm. To find the best genetic operators for it, we conducted hyperparameter optimization. We use grid search to go through all configurations of genetic operators. For each operator combination, we evaluated the resulting Pareto front using the *hypervolume indicator* (HV) [11]. Due to the stochastic nature of GAs, we conducted 10 trials for each combination. We used a population size of 100 and the number of generations was set to 200. This was done on all given datasets.

Table 5: Hyperparameter optimization results showing hypervolume indicator values for different genetic operators. Best results are highlighted in bold.

Initializer	Selection	Crossover	Mutation	Adult	Bank	Compas
Random	Elitist	1-Point	Bit Flip	0.49 ± 0.00	0.47 ± 0.00	0.53 ± 0.00
			Shuffle	0.47 ± 0.01	0.44 ± 0.01	0.52 ± 0.03
		Uniform	Bit Flip	0.48 ± 0.00	0.45 ± 0.00	0.51 ± 0.01
			Shuffle	0.46 ± 0.01	0.43 ± 0.01	0.49 ± 0.01
	Tournament	1-Point	Bit Flip	0.49 ± 0.00	0.47 ± 0.00	0.53 ± 0.00
			Shuffle	0.50 ± 0.01	0.46 ± 0.00	0.57 ± 0.02
		Uniform	Bit Flip	0.49 ± 0.00	0.47 ± 0.00	0.54 ± 0.00
			Shuffle	0.52 ± 0.01	0.47 ± 0.01	0.61 ± 0.02
Variable	Elitist	1-Point	Bit Flip	**0.87** ± 0.00	**0.89** ± 0.00	**0.90** ± 0.00
			Shuffle	0.85 ± 0.00	0.78 ± 0.00	0.83 ± 0.01
		Uniform	Bit Flip	0.85 ± 0.00	0.78 ± 0.00	0.85 ± 0.00
			Shuffle	0.84 ± 0.00	0.77 ± 0.00	0.82 ± 0.01
	Tournament	1-Point	Bit Flip	**0.87** ± 0.00	0.86 ± 0.00	0.89 ± 0.00
			Shuffle	0.85 ± 0.00	0.78 ± 0.00	0.84 ± 0.01
		Uniform	Bit Flip	**0.87** ± 0.00	0.81 ± 0.00	0.88 ± 0.00
			Shuffle	0.85 ± 0.00	0.78 ± 0.00	0.84 ± 0.01

Hyperparameters There are several known methods we can choose from for each genetic operator (*initializer, selection, crossover, mutation*) [12,10]. We only considered those that return binary vectors, as our solutions are binary. For all selection methods, we used two parents. Furthermore, the bit flip mutation rate was set to 5%.

Metric To goal is to maximize the *hypervolume indicator* (HV) [11]. It measures the volume between the Pareto front and a reference point. A higher hypervolume indicates more coverage of the solution space and hence a better Pareto front. We use the nadir point (1, 1) as the reference point. A value of 1 indicates the theoretically best possible solution and a value of 0 indicates the worst.

Results The results are displayed in Table 5. We display the mean and standard deviation of the hypervolume indicator for each dataset and genetic operator combination from the trials. To answer **RQ1**, we observed best results with variable initializer, elitist selection, 1-point crossover, and bit flip mutation in our experiments. Remarkably, the results confirm that the variable initializer suits our problem better than the random initializer as it outperformed the latter in all datasets by a significant margin. We also observed that elitist selection slightly outperformed the binary tournament selection method. Bigger differences can be observed in the crossover and mutation operators. We note that all results were very consistent, as the maximum standard deviation was 0.01 and most often 0.00 when rounding to two decimals.

7.2 Bias Mitigation and Classification Performance

The aim of this experiment is to assess the performances of various machine learning classifiers trained on datasets pre-processed for fairness using our methodology. Specifically, we aim to evaluate the effectiveness in mitigating bias while maintaining classification accuracy. For this, we also trained the models on the unprocessed datasets, serving as a baseline.

Train and Test Split For each dataset, we split the data into training and testing sets, ensuring stratification based on sensitive attributes to preserve representativeness across groups. We used a 80-20 split for training and testing, respectively.

Bias Mitigation After splitting, we either applied the multi-objective optimization approach or the single-objective optimization approach to mitigate bias in the training data. The test data was left completely unprocessed.

For both approaches, we experimented with different normalization factors $\beta \in \{(1+\epsilon), \hat{\psi}(\mathcal{D})\}$. The population size was set to 200 and the number of generations was set to 400.

For the multi-objective approach, we used the NSGA-II algorithm with the best genetic operators identified in the former experiment. Because we get a Pareto front of solutions, we need to select one solution for the evaluation. We selected the solution from the Pareto front $\mathcal{PF}$ based on β as follows:

$$\underset{\mathcal{D}_{\text{fair}} \in \mathcal{PF}}{\operatorname{argmin}} \; \frac{\hat{\psi}_{\text{SDP-max}}(\mathcal{D}_{\text{fair}})}{\beta} + \mathcal{L}(\mathcal{D}, \mathcal{D}_{\text{fair}}).$$

For the single objective, we used the genetic algorithm from Duong et al. [9] with the same genetic operators as the NSGA-II algorithm for comparison. We used the solution that is returned when solving the optimization problem in Equation (6). An alpha value of 0.5 was set to equally weigh the objectives. However, we note that β also influences the selection of the solution.

Machine Learning Models We trained several machine learning classifiers implemented by the `scikit-learn` library [18], including *Logistic Regression* (LR), *Support Vector Machines* (SVM), *Random Forest* (RF), and *Artificial Neural Networks* (ANNs), on both the pre-processed fair data and the original, unprocessed data.

Metrics Using the test set, we evaluated the models' predictions on fairness and performance. For fairness, we used the maximal statistical disparity (see Example 1). We used this instead of the penalized version because there is no need to penalize the test set for coverage. We note that the test set automatically contains all groups due to the stratification in the train-test split. For the classifiers' performances, we report the *area under the receiver operating characteristic curve* (AUROC), where higher values indicate better performance.

Table 6: Fairness and relative size of the pre-processed training sets compared to the original training sets. Best results are in bold.

Dataset	Approaches	$\hat{\psi}_{\text{SDP-max}}(\mathcal{D})$	$\hat{\psi}_{\text{SDP-max}}(\mathcal{D}_{\text{fair}})$	$\lvert\mathcal{D}_{\text{fair}}\rvert/\lvert\mathcal{D}\rvert$
Adult	Single $\beta = \hat{\psi}(\mathcal{D})$	17%	**5%**	54%
	Single $\beta = (1+\epsilon)$	17%	16%	**99%**
	Multi $\beta = \hat{\psi}(\mathcal{D})$	17%	8%	54%
	Multi $\beta = (1+\epsilon)$	17%	17%	**99%**
Bank	Single $\beta = \hat{\psi}(\mathcal{D})$	25%	**6%**	51%
	Single $\beta = (1+\epsilon)$	25%	25%	**99%**
	Multi $\beta = \hat{\psi}(\mathcal{D})$	25%	11%	53%
	Multi $\beta = (1+\epsilon)$	25%	25%	**99%**
Compas	Single $\beta = \hat{\psi}(\mathcal{D})$	21%	**1%**	54%
	Single $\beta = (1+\epsilon)$	21%	15%	94%
	Multi $\beta = \hat{\psi}(\mathcal{D})$	21%	3%	53%
	Multi $\beta = (1+\epsilon)$	21%	15%	**96%**

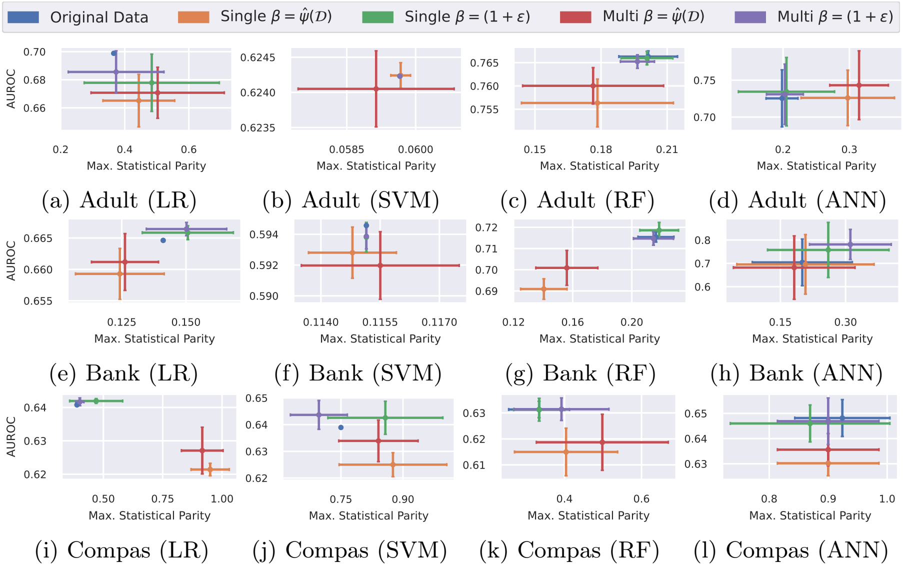
Results Table 6 shows the maximal statistical disparity values of the original and pre-processed training sets, as well as the relative size of the pre-processed training sets compared to the original training sets. Notably, setting $\beta = \hat{\psi}(\mathcal{D})$ results in a much lower discrimination than setting $\beta = (1+\epsilon)$, but also in a higher data loss. We observe only small differences between the single- and multi-objective approaches in the discrimination values when setting $\beta = \hat{\psi}(\mathcal{D})$. The classifiers' results are displayed in Fig. 1. We compare the classifiers' performances on the test set using both the single- and multi-objective optimization approaches. For each approach, we varied the normalization factor β as described above. We use error bars to display the mean and standard deviation of the AUROC and the maximal statistical disparity values from the 10 trials.

We do not observe a clear trend in the results regarding improving or worsening fairness and performance. This emphasizes that the pre-processed datasets can indeed be used reliably as training data. Only the experiment shown in Fig. 1i stands out, where the predictions on the pre-processed data are significantly less fair than on the original data.

When comparing the classifiers, SVM seems to be least affected by the pre-processed datasets, mostly apparent in Fig. 1b and Fig. 1f. Further, fairer classifiers tend to have lower performances, indicating a trade-off between fairness and performance.

Contrasting the approaches with different β parameters, we do not see a clear winner. Notable is the result of the single-objective approach with $\beta = (1+\epsilon)$ in Fig. 1i, where its fairness did not worsen. The experiment in Fig. 1h shows an interesting result, where less fair predictions achieve better performances. This also indicates that a fairness-performance trade-off is indeed available.

We can conclude that the pre-processed datasets can be used reliably as training data, but an improvement in fairness can only be guaranteed in the training



(a) Adult (LR) (b) Adult (SVM) (c) Adult (RF) (d) Adult (ANN)

(e) Bank (LR) (f) Bank (SVM) (g) Bank (RF) (h) Bank (ANN)

(i) Compas (LR) (j) Compas (SVM) (k) Compas (RF) (l) Compas (ANN)

Fig. 1: Results on the test set using two approaches with varying β parameters. x-axis and y-axis represent discrimination and performance metrics, respectively.

set and not necessarily in the test set. Evidence of a fairness-performance trade-off was found in the results. This explains why fairness was rarely improved in the experiments and why the classifiers' performances were not negatively affected. The protected attributes seem to correlate with the target variable, making it difficult to remove the discrimination without changing the data significantly.

8 Discussion

In this section, we reflect on key aspects of our approach, particularly focusing on data quality and the trade-off between fairness and data loss.

8.1 Data Quality

We extend the work of Duong et al. [9] by introducing new constraints and objectives that the resulting dataset must fulfill. To our knowledge, we are the first to specifically aim for data quality while pre-processing datasets for fairness. Some techniques in the literature [2,1] violate several data integrity constraints and transform the dataset in a way that makes it unusable. For example, `CorrelationRemover` in `Fairlearn` [2] projects discrete features into continuous features. Categorical features such as the label are also affected by it, making it impossible to train classifiers for comparative purposes. The pre-processed datasets from our work do not come with these issues, making them more useful in practice.

8.2 Trade-off

We proposed a multi-objective optimization problem where the objectives are fairness and data loss. Solving this problem results in a set of Pareto-optimal solutions where each solution is a fair subset. The user can then choose the most suitable subset for their application using the weighted sum of the objectives. Here, we proposed a parameter β that weights the fairness objective. The choice of β has to be made by the user and depends on the size of the dataset and the importance of fairness in the application. Our suggested values for β serve as initial guidance, allowing users to further adjust based on their needs.

9 Conclusion

In this paper, we developed a data pre-processing technique that aims to remove discriminating data points for fairness while maintaining data quality. We introduced two additional criteria, *coverage* and *data loss*, to enhance the trustworthiness of the resulting dataset. Using our methods, the fairness of the dataset can be improved without compromising the quality of the data. By evaluating our methodology on machine learning models using three real-world datasets, the results show that the models' fairness and performances were not affected significantly by the data removal process compared to models trained on the original datasets. This indicates that the pre-processed datasets are reliable and can be used for further analysis.

References

1. Bellamy, R.K.E., Dey, K., Hind, M., Hoffman, S.C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., Nagar, S., Ramamurthy, K.N., Richards, J.T., Saha, D., Sattigeri, P., Singh, M., Varshney, K.R., Zhang, Y.: AI fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. CoRR **abs/1810.01943** (2018)
2. Bird, S., Dudík, M., Edgar, R., Horn, B., Lutz, R., Milan, V., Sameki, M., Wallach, H., Walker, K.: Fairlearn: A toolkit for assessing and improving fairness in AI. Tech. Rep. MSR-TR-2020-32, Microsoft (May 2020), https://www.microsoft.com/en-us/research/publication/fairlearn-a-toolkit-for-assessing-and-improving-fairness-in-ai/
3. Calders, T., Kamiran, F., Pechenizkiy, M.: Building classifiers with independency constraints. In: 2009 IEEE International Conference on Data Mining Workshops. pp. 13–18 (2009). https://doi.org/10.1109/ICDMW.2009.83
4. Catania, B., Guerrini, G., Janpih, Z.: Mitigating representation bias in data transformations: A constraint-based optimization approach. In: 2023 IEEE International Conference on Big Data (BigData). pp. 4127–4136. IEEE Computer Society, Los Alamitos, CA, USA (2023). https://doi.org/10.1109/BigData59044.2023.10386181
5. Catania, B., Guerrini, G., Accinelli, C.: Fairness & friends in the data science era. AI & SOCIETY **38**(2), 721–731 (2023)
6. Deb, K., Pratap, A., Agarwal, S., Meyarivan, T.: A fast and elitist multiobjective genetic algorithm: Nsga-ii. IEEE transactions on evolutionary computation **6**(2), 182–197 (2002)

7. Drosou, M., Jagadish, H., Pitoura, E., Stoyanovich, J.: Diversity in big data: A review. Big data **5**(2), 73–84 (2017). https://doi.org/10.1089/big.2016.0054
8. Duong, M.K., Conrad, S.: Dealing with data bias in classification: Can generated data ensure representation and fairness? In: Big Data Analytics and Knowledge Discovery - 25th International Conference, DaWaK 2023, Penang, Malaysia, August 28-30, 2023, Proceedings. Lecture Notes in Computer Science, vol. 14148, pp. 176–190. Springer Nature Switzerland (2023)
9. Duong, M.K., Conrad, S.: Towards fairness and privacy: A novel data preprocessing optimization framework for non-binary protected attributes. In: Data Science and Machine Learning. pp. 105–120. No. CCIS 1943 in AusDM: Australasian Conference on Data Science and Machine Learning, Springer Nature Singapore (2023)
10. Eiben, A.E., Smith, J.E.: Introduction to Evolutionary Computing. Springer Publishing Company, Incorporated, 2nd edn. (2015)
11. Fonseca, C., Paquete, L., Lopez-Ibanez, M.: An improved dimension-sweep algorithm for the hypervolume indicator. In: 2006 IEEE International Conference on Evolutionary Computation. pp. 1157–1163 (2006). https://doi.org/10.1109/CEC.2006.1688440
12. Goldberg, D.E.: Genetic Algorithms in Search, Optimization and Machine Learning. Addison-Wesley Longman Publishing Co., Inc., USA, 1st edn. (1989)
13. Hardt, M., Price, E., Srebro, N.: Equality of opportunity in supervised learning. Advances in neural information processing systems **29** (2016)
14. Kohavi, R.: Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. p. 202–207. KDD'96, AAAI Press (1996)
15. Larson, J., Angwin, J., Mattu, S., Kirchner, L.: Machine bias (May 2016), https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
16. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. ACM Computing Surveys (CSUR) **54**(6), 1–35 (2021)
17. Moro, S., Cortez, P., Rita, P.: A data-driven approach to predict the success of bank telemarketing. Decision Support Systems **62**, 22–31 (2014)
18. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011)
19. Stoyanovich, J., Yang, K., Jagadish, H.: Online set selection with fairness and diversity constraints. In: Proceedings of the EDBT Conference (2018)
20. Verma, S., Ernst, M.D., Just, R.: Removing biased data to improve fairness and accuracy. CoRR **abs/2102.03054** (2021)
21. Zafar, M.B., Valera, I., Gomez Rodriguez, M., Gummadi, K.P.: Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In: Proceedings of the 26th International Conference on World Wide Web. International World Wide Web Conferences Steering Committee (2017). https://doi.org/10.1145/3038912.3052660
22. Žliobaitė, I.: Measuring discrimination in algorithmic decision making. Data Mining and Knowledge Discovery **31**, 1060–1089 (2017)

5

Conclusion

In this thesis, multiple novel contributions to the field of fairness in machine learning have been made. The following section provides a summary of the key findings gained from this work and discusses their implications in practice. In the last section, future research directions are outlined.

5.1 Concluding Remarks

The increasing reliance on machine learning models in decision-making processes can have significant implications for individuals and society as a whole. Specifically, the problem of fairness in machine learning has been addressed from multiple perspectives in this dissertation.

In Chapter 2, we have explored these critical issues in a specific case study within an academic setting [8, 1]. As such, it is crucial to ensure that machine learning models are fair and unbiased to prevent discrimination against vulnerable subpopulations.

With the identified problems and challenges in mind, Chapter 3 presents several novel methods [2, 3] to mitigate discrimination in datasets, aiming to train fairer machine learning models. The introduced algorithms address the limitations of existing methods and fill the gaps in the literature. For example, most pre-processing methods transform and edit the data [33, 51, 25, 19]. To maintain the integrity of the dataset, `MetricOptimizer` [2] and the `FairDo` framework [3] are able to utilize synthetic data to incorporate it into the biased original dataset to improve fairness. This is particularly useful when the original dataset is small and changes to it are unwanted. The use of synthetic data can also help to protect the privacy of individuals by obfuscating the original data. Further, we were able to deal with types of discrimination that prior methods were not able to handle, such as discrimination towards non-binary and intersectional groups [2, 3, 4].

For completeness, Chapter 4 aims to enhance trust by critically examining the reliability of group fairness metrics and the introduced `FairDo` framework [3]. We show that fairness metrics can be misleading and propose a problem where solely considering group disparities is not sufficient to conclude the fairness of a decision-making system [7]. We address this by introducing a new measure that considers the uncertainty when measuring discrimination. In another study [6], we improved the `FairDo` framework when users decide to remove samples from the dataset. In such cases, there is a risk of removing too many samples or entire groups, which is undesirable. For this reason, we introduced a new multi-objective optimization problem to balance fairness and data removal; users can decide how much fairness they want to achieve and how much data they are willing to remove.

The methods introduced in the papers in Chapter 3 and Chapter 4 deal with different aspects that are part of our fairness framework `FairDo`. These methods are designed to be used in conjunction with each other to provide flexibility in solving fairness problems. For example, it is possible to aim for fairness for multiple protected attributes and at the same time consider the quality of the data by using the discrimination measures from Section 3.3 and the multi-objective optimization problem from Section 4.2.

To summarize, this dissertation has made several contributions to the field of fairness in machine learning. We motivated the need for fair machine learning models, solved it by introducing novel pre-processing methods, and finally enhanced prior methods to raise trust. These contributions are particularly relevant as they directly tackle some aspects of the AI Act [24], which will soon come into force [45]. Our work thus aligns with upcoming regulatory requirements by addressing current challenges, emphasizing its practical importance.

5.2 Future Work

While this thesis has made novel and important contributions to fill the literature gap regarding fairness in machine learning, there are still many areas that can be further explored.

Our pre-processing methods [2, 3, 4, 5] come with the advantage that any fairness metric can be used as the objective function. This is realized by treating the objective function as a black-box. However, this flexibility implies that only heuristics can be used in practice. The only exact method to deal with this type of optimization problem is brute-force search, which is not feasible for large datasets. Future work could focus on developing exact solvers tailored to specific fairness metrics to improve the efficiency of the optimization process and to guarantee the global optimality of the solution.

Regarding the `FairDo` package [3], there is room to expand the package by incorporating additional optimization algorithms that address both single-objective and multi-objective optimization problems as stated in Chapter 3. Furthermore, the package can be extended with more metrics, datasets, and any other functionality. Because we follow a functional programming paradigm, the package can be easily extended and maintained.

Generally, the proposed methods are not limited to classification tasks, as shown in the experiments, but are also applicable to regression tasks. By using fairness metrics specifically designed for regression tasks, such as individual fairness, additional experiments could be conducted to show the effectiveness of the proposed methods. Additionally, experiments on other datasets can be conducted as well, besides using popular and well-known datasets from the fairness literature.

The possibilities are endless, and we hope that this thesis will inspire future researchers to continue to address one of these aspects and make use of our tools [8], methods [2, 3, 4, 5], and fairness measures [7].

6

Publications

Contributing Publications

[1] Jannik Dunkelau and Manh Khoi Duong. Towards Equalised Odds as Fairness Metric in Academic Performance Prediction. *Fairness, Accountability, and Transparency in Educational Data 2022 Workshop*, 2022.

[2] Manh Khoi Duong and Stefan Conrad. Dealing with Data Bias in Classification: Can Generated Data Ensure Representation and Fairness? In Robert Wrembel, Johann Gamper, Gabriele Kotsis, A. Min Tjoa, and Ismail Khalil, editors, *Big Data Analytics and Knowledge Discovery*, volume 14148 of *Lecture Notes in Computer Science*, pages 176–190. Springer Cham, 2023.

[3] Manh Khoi Duong and Stefan Conrad. Towards Fairness and Privacy: A Novel Data Pre-processing Optimization Framework for Non-binary Protected Attributes. In Diana Benavides-Prado, Sarah Erfani, Philippe Fournier-Viger, Yee Ling Boo, and Yun Sing Koh, editors, *Data Science and Machine Learning*, volume 1943 of *Communications in Computer and Information Science*, pages 105–120. Springer Singapore, 2023.

[4] Manh Khoi Duong and Stefan Conrad. Measuring and Mitigating Bias for Tabular Datasets with Multiple Protected Attributes. In Roberta Calegari, Virginia Dignum, and Barry O'Sullivan, editors, *Proceedings of the 2nd Workshop on Fairness and Bias in AI co-located with 27th European Conference on Artificial Intelligence (ECAI 2024), Santiago de Compostela, Spain, October 20th, 2024*, volume 3808 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2024.

[5] Manh Khoi Duong and Stefan Conrad. Trusting Fair Data: Leveraging Quality in Fairness-Driven Data Removal Techniques. *CoRR*, abs/2405.12926v3, 2024. Extended version.

[6] Manh Khoi Duong and Stefan Conrad. Trusting Fair Data: Leveraging Quality in Fairness-Driven Data Removal Techniques. In Robert Wrembel, Silvia Chiusano, Gabriele Kotsis, A Min Tjoa, and Ismail Khalil, editors, *Big Data Analytics and Knowledge Discovery*, volume 14912 of *Lecture Notes in Computer Science*, pages 375–380. Springer Cham, 2024.

[7] Manh Khoi Duong and Stefan Conrad. (Un)certainty of (Un)fairness: Preference-Based Selection of Certainly Fair Decision-Makers. In *ECAI 2024 - 27th European Conference on Artificial Intelligence*, volume 392 of *Frontiers in Artificial Intelligence and Applications*, pages 882–889. IOS Press, 2024.

[8] Manh Khoi Duong, Jannik Dunkelau, José Andrés Cordova, and Stefan Conrad. RAPP: A Responsible Academic Performance Prediction Tool for Decision-Making in Educational Institutes. In *BTW 2023*, volume P-331 of *Lecture Notes in Informatics*, pages 595–606. Gesellschaft für Informatik e.V., 2023.

Other Publications

[9] Manh Khoi Duong. Automated Architecture-Modeling for Convolutional Neural Network. In *BTW 2019 – Workshopband*, volume P-290 of *Lecture Notes in Informatics*, pages 163–172. Gesellschaft für Informatik e.V., 2019.

Bibliography

[10] Alekh Agarwal, Miroslav Dudik, and Zhiwei Steven Wu. Fair regression: Quantitative definitions and reduction-based algorithms. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 120–129. PMLR, Jun 2019.

[11] Google AI. AI principles progress update. `https://ai.google/static/documents/ai-principles-2023-progress-update.pdf`, 2023. Accessed: 01 October 2024.

[12] Alexander Askinadze. *From Collecting, Integrating, and Visualizing Student Data to Predicting Student Dropout and Performance.* PhD thesis, Heinrich Heine University Düsseldorf, 2020.

[13] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning.* fairmlbook.org, 2019.

[14] Solon Barocas and Andrew D. Selbst. Big data's disparate impact. *California Law Review*, 104(3):671–732, 2016.

[15] Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John T. Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang. AI fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *CoRR*, abs/1810.01943, 2018.

[16] Sarah Bird, Miro Dudík, Richard Edgar, Brandon Horn, Roman Lutz, Vanessa Milan, Mehrnoosh Sameki, Hanna Wallach, and Kathleen Walker. Fairlearn: A toolkit for assessing and improving fairness in AI. Technical report, Microsoft, May 2020.

[17] Kirill Bogomasov, Daniel Braun, Andreas Burbach, Ludmila Himmelspach, and Stefan Conrad. Feature and deep learning based approaches for automatic report generation and severity scoring of lung tuberculosis from CT images. In *Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum*, volume 2380 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2019.

[18] Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. Building classifiers with independency constraints. In *2009 IEEE International Conference on Data Mining Workshops*, pages 13–18, 2009.

[19] Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. Optimized pre-processing for discrimination prevention. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[20] Simon Caton and Christian Haas. Fairness in machine learning: A survey. *ACM Computing Surveys*, 56(7), 2024.

[21] Flavio Chierichetti, Ravi Kumar, Silvio Lattanzi, and Sergei Vassilvitskii. Fair clustering through fairlets. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 5036–5044, Red Hook, NY, USA, 2017. Curran Associates Inc.

[22] Sam Corbett-Davies and Sharad Goel. The measure and mismeasure of fairness: A critical review of fair machine learning. *CoRR*, abs/1808.00023, 2018.

[23] Kalyanmoy Deb, Amrit Pratap, Sameer Agarwal, and TAMT Meyarivan. A fast and elitist multiobjective genetic algorithm: Nsga-ii. *IEEE transactions on evolutionary computation*, 6(2):182–197, 2002.

[24] European Commission. Artificial Intelligence Act, Corrigendum, 19 April 2024. `https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138-FNL-COR01_EN.pdf`, April 2024. Accessed: 01 October 2024.

[25] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, page 259–268. Association for Computing Machinery, 2015.

[26] C.M. Fonseca, L. Paquete, and M. Lopez-Ibanez. An improved dimension-sweep algorithm for the hypervolume indicator. In *2006 IEEE International Conference on Evolutionary Computation*, pages 1157–1163, 2006.

[27] James R. Foulds, Rashidul Islam, Kamrun Naher Keya, and Shimei Pan. Bayesian modeling of intersectional fairness: The variance of bias. In *Proceedings of the 2020 SIAM International Conference on Data Mining (SDM)*, pages 424–432, 2020.

[28] Sorelle A. Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. On the (im)possibility of fairness. *CoRR*, abs/1609.07236, 2016.

[29] Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. *Bayesian Data Analysis*. CRC press, 1995.

[30] Moritz Hardt, Eric Price, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

[31] Hans Hofmann. German credit data. `https://archive.ics.uci.edu/ml/datasets/Statlog+%28German+Credit+Data%29`, 1994. Accessed: 01 October 2024.

[32] Ching-Lai Hwang and Kwangsun Yoon. *Methods for Multiple Attribute Decision Making*, pages 58–191. Springer Berlin Heidelberg, Berlin, Heidelberg, 1981.

[33] Faisal Kamiran and Toon Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1):1–33, 2012.

[34] Jian Kang, Tiankai Xie, Xintao Wu, Ross Maciejewski, and Hanghang Tong. Infofair: Information-theoretic intersectional fairness. In *2022 IEEE International Conference on Big Data (Big Data)*, pages 1455–1464, 2022.

[35] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2564–2572. PMLR, Jul 2018.

[36] Keith Kirkpatrick. It's not the algorithm, it's the data. *Communications of the ACM*, 60(2):21–23, Jan 2017.

[37] Ron Kohavi. Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, page 202–207. AAAI Press, 1996.

[38] Jeff Larson, Julia Angwin, Surya Mattu, and Lauren Kirchner. Machine bias. `https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing`, May 2016. Accessed: 01 October 2024.

[39] Tai Le Quy, Arjun Roy, Vasileios Iosifidis, Wenbin Zhang, and Eirini Ntoutsi. A survey on datasets for fairness-aware machine learning. *WIREs Data Mining and Knowledge Discovery*, 12(3):e1452, 2022.

[40] Kristian Lum, Yunfeng Zhang, and Amanda Bower. De-biasing "bias" measurement. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 379–389, New York, NY, USA, 2022. Association for Computing Machinery.

[41] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54(6), 2021.

[42] Microsoft. Responsible AI Standard, v2. https://cdn-dynmedia-1.microsoft.com/is/content/microsoftcorp/microsoft/final/en-us/microsoft-brand/documents/Microsoft-Responsible-AI-Standard-General-Requirements.pdf?culture=en-us&country=us, 2022.

[43] Tom Michael Mitchell. *Machine Learning*, volume 1. McGraw-Hill Education, 1997.

[44] Sérgio Moro, Paulo Cortez, and Paulo Rita. A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, 62:22–31, 2014.

[45] Future of Life Institute. AI act historic timeline. https://artificialintelligenceact.eu/developments/, 2024. Accessed: 01 October 2024.

[46] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

[47] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

[48] Julia Romberg. *Machine-assisted Text Classification of Public Participation Contributions.* PhD thesis, Heinrich Heine University Düsseldorf, 2023.

[49] Yifan Wang, Weizhi Ma, Min Zhang, Yiqun Liu, and Shaoping Ma. A survey on the fairness of recommender systems. *ACM Transactions on Information Systems*, 41(3), Feb 2023.

[50] Forest Yang, Moustapha Cisse, and Sanmi Koyejo. Fairness with overlapping groups. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc.

[51] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 325–333. PMLR, 2013.

[52] E. Zitzler and L. Thiele. Multiobjective evolutionary algorithms: a comparative case study and the strength pareto approach. *IEEE Transactions on Evolutionary Computation*, 3(4):257–271, 1999.

[53] Indrė Žliobaitė. Measuring discrimination in algorithmic decision making. *Data Mining and Knowledge Discovery*, 31:1060–1089, 2017.