POLITECNICO DI TORINO

MASTER's Degree in COMPUTER ENGINEERING

Artificial Intelligence and Data Analytics

MASTER's Degree Thesis

Addressing Gender and Racial Bias in AI: A Data-Centric Approach for Fairer Outcomes

Supervisors

Prof. ANTONIO VETRÒ

Dr. LUCA GILLI

Dr. SIMONA MAZZARINO

Candidate

FESTA SHABANI

2024/2025

Abstract

This thesis tackles the issue of bias in Artificial Intelligence (AI) systems, specifically focusing on the mitigation of gender and racial biases through a data-centric approach that seeks to achieve more equitable results. As AI systems become more prevalent in critical sectors such as healthcare, finance, and criminal justice, they risk unintentionally reinforcing and amplifying societal biases that are present in the data they learn from. To address this, this research explores two main techniques for bias mitigation: preprocessing bias correction methods using the AI Fairness 360 (AIF360) [1] toolkit and synthetic data generation with Clearbox AI's Synthetic Kit [2] to augment underrepresented groups. Specifically, the Adult [3] and Medical Expenditure [4] datasets, which involve sensitive attributes such as sex and race, are used to demonstrate how bias manifests differently in socio-economic and healthcare domains. Various preprocessing methods, including Reweighing, Disparate Impact Remover, Learning Fair Representations, and Optimized Preprocessing, are applied to mitigate bias, while synthetic data is generated to balance demographic disparities. The effectiveness of these methods is evaluated based on fairness metrics like Statistical Parity Difference, Disparate Impact, Average Odds Difference, Equal Opportunity Difference, and Theil Index, alongside performance metrics such as Balanced Accuracy. The results highlight the potential of preprocessing bias mitigation techniques, especially synthetic data generation as a form of dataset augmentation, in reducing bias without significantly sacrificing model performance. This work contributes to the growing field of responsible AI by demonstrating how a data-centric approach can improve fairness in AI models, ensuring fairer outcomes across diverse groups. The findings have practical implications for AI deployment in sensitive applications, providing strategies to improve fairness and accountability in AI-driven decision-making systems. Future work may explore additional bias mitigation techniques, including in-processing and post-processing methods, and further investigate the role of synthetic data generation to reduce bias in real-world AI systems.

Dedications

To the Albanian women who came before me, whose access to education was never granted. This thesis is a tribute to your strength, resilience, and unrecognized potential. Your sacrifices and unspoken dreams paved the way for my journey. May this work stand as a symbol of what is possible, in honor of all that you could have been.

Festa Shabani
MSc in Computer Engineering
Artificial Intelligence and Data Analytics
Turin, Italy, 2025

ACKNOWLEDGMENTS

The completion of this thesis marks the culmination of a journey filled with growth, challenges, and opportunities, and I am deeply indebted to those who have walked alongside me during this time.

First and foremost, I want to extend my sincerest gratitude to my thesis supervisors, Prof. Antonio Vetrò, Dr. Luca Gilli, and Dr. Simona Mazzarino. Their wisdom, guidance, and support have been instrumental in shaping this work.

I am incredibly grateful to my family and friends for their patience, encouragement, and understanding throughout this process. Their support has been a constant source of strength and motivation.

Lastly, I wish to express my heartfelt appreciation to my partner, Lorik, for the profound impact on both my personal and academic life and for always treating me with the utmost respect, setting a benchmark for what fairness and equality should look like.

Table of Contents

Dedications **II**

1 Introduction **1**
1.1 Motivation . . . 1
1.2 Ethical Considerations . . . 2
1.3 Problem Statement and Research Objectives . . . 4
1.4 Scope and Limitations . . . 4

2 Literature Review **6**
2.1 Overview of Bias in AI . . . 6
2.2 Bias Detection and Mitigation . . . 7
2.3 Synthetic Data . . . 9
2.3.1 Types of Synthetic Data Generation . . . 9
2.3.2 Applications of Synthetic Data . . . 9
2.3.3 Synthetic Data for Bias Mitigation . . . 10

3 Methodology **12**
3.1 Datasets . . . 12
3.1.1 Adult Dataset . . . 12
3.1.2 Medical Expenditure Dataset . . . 12
3.1.3 Key Features of Both Datasets . . . 13
3.1.4 Dataset Insights . . . 13
3.2 Bias Detection Metrics . . . 14
3.3 Bias Mitigation Techniques . . . 15
3.3.1 Reweighing . . . 15
3.3.2 Disparate Impact Remover . . . 16
3.3.3 Learning Fair Representations (LFR) . . . 16
3.3.4 Optimized Preprocessing . . . 17
3.3.5 Dataset Augmentation through Synthetic Data . . . 17
3.4 Classifiers . . . 18
3.4.1 Logistic Regression . . . 18
3.4.2 Random Forest . . . 18
3.4.3 Gradient Boosting . . . 18
3.5 Performance evaluation . . . 19

3.6 Tools and Frameworks 19

4 Implementation **21**
4.1 Data Overview and Insights 21
4.1.1 Adult 22
4.1.2 Medical Expenditure 24
4.2 Preprocessing of Datasets 29
4.2.1 Adult Dataset Preprocessing 29
4.2.2 Medical Expenditure Dataset Preprocessing 33
4.3 Initial Bias Assessment 34
4.4 Bias Mitigation 34
4.4.1 Bias Mitigation Using AIF360 34
4.4.1.1 Reweighing 35
4.4.1.2 Disparate Impact Remover 35
4.4.1.3 Learning Fair Representations (LFR) 35
4.4.1.4 Optimized Preprocessing 35
4.4.2 Bias Mitigation Using Synthetic Data Augmentation 36
4.4.2.1 Identifying Underrepresented Groups 36
4.4.2.2 Synthetic Data Generation Process 37
4.4.2.3 Synthetic datasets 37
4.5 Fairness Evaluation Before and After Transformation 38
4.6 Classifier Training and Evaluation 38
4.6.1 Classifier Training 39
4.6.2 Evaluation 39
4.6.3 Plotting Fairness Metrics 40

5 Results **43**
5.1 Adult 44
5.2 Medical Expenditure 44

6 Discussion **46**
6.1 Raw Performance and Fairness Interpretations 48
6.1.1 Adult 48
6.1.1.1 Logistic Regression (LR): 48
6.1.1.2 Random Forest (RF): 49
6.1.1.3 Gradient Boosting (GB): 50
6.1.2 Medical Expenditure 51
6.1.2.1 Logistic Regression (LR): 51
6.1.2.2 Random Forest (RF): 52
6.1.2.3 Gradient Boosting (GB): 53
6.2 Relative Improvements Over Baselines 54
6.2.1 Adult 55
6.2.1.1 Balanced Accuracy Differences 55
6.2.1.2 Disparate Impact Differences 56

6.2.2 Medical Expenditure 57
6.2.2.1 Balanced Accuracy Differences 58
6.2.2.2 Disparate Impact Differences 58
6.3 Key Trends Across Datasets 59
6.3.1 Insights and Implications 60

7 Conclusion **61**

Bibliography **63**

List of Figures

3.1 The fairness pipeline followed in this thesis, which focuses on the preprocessing phase. The diagram illustrates the process of transforming the original dataset into a fairer version using fairness preprocessing algorithms. This transformed dataset is then used for training a classifier, and fairness metrics are evaluated on both the original and transformed datasets. (Taken from [1]). 15

4.1 Dataset, Bias Mitigation Techniques, and Classifiers Structure 21
4.2 Gender distribution of individuals in Adult dataset. 23
4.3 Rate of positive outcomes by gender in Adult dataset. 24
4.4 Race distribution in Adult dataset. 25
4.5 Rate of positive outcomes by race in Adult dataset. 26
4.6 Age distribution of individuals in the Adult dataset. 26
4.7 Positive outcome rates by age group and gender in Adult dataset. . . 27
4.8 Distribution of years of education in Adult dataset. 27
4.9 Positive outcome rates by education level and gender in Adult dataset. 28
4.10 Race distribution in MEPS dataset. 28
4.11 Positive outcome rates by race in MEPS dataset. 29
4.12 Gender distribution in MEPS dataset. 30
4.13 Age distribution in MEPS dataset. 30
4.14 Positive outcome rates across age groups in MEPS dataset. 31
4.15 Positive outcome rates by poverty category in MEPS dataset. 31
4.16 Positive outcome rates by race and poverty category in MEPS dataset. 32
4.17 Positive outcome rates by insurance coverage and race in MEPS dataset. 32
4.18 Preprocessed Adult dataset 33
4.19 Preprocessed MEPS dataset 33
4.20 Disparate Impact vs. Threshold for the Original Test Data of the `Adult` Dataset using Logistic Regression and Reweighing. 40
4.21 Average Odds Difference vs. Threshold for the Original Test Data of the `Adult` Dataset using Logistic Regression and Reweighing. 41
4.22 Disparate Impact vs. Threshold for the Transformed Test Data of the `Adult` Dataset using Logistic Regression and Reweighing. 41
4.23 Average Odds Difference vs. Threshold for the Transformed Test Data of the `Adult` Dataset using Logistic Regression and Reweighing. . . . 42

6.1 Bar plot showing performance and fairness metric differences for the **Adult dataset**. 55
6.2 Bar plot showing performance and fairness metric differences for the **MEPS dataset**. 57

List of Tables

5.1 Performance and Fairness Metrics for Bias Mitigation Methods (**Adult Dataset**) 44
5.2 Results for accuracy and fairness differences across methods for the **Adult dataset**. 44
5.3 Performance and Fairness Metrics for Bias Mitigation Methods (**Medical Expenditure Dataset**) 44
5.4 Results for accuracy and fairness differences across methods for the **Medical Expenditure dataset**. 45

Acronyms

AI	Artificial Intelligence.
AIF360	AI Fairness 360.
DIR	Disparate Impact Remover.
LFR	Learning Fair Representations.
OptimPreproc	Optimized Preprocessing.
VAE	Variational Autoencoder.
MEPS	Medical Expenditure Panel Survey.
TPR	True Positive Rate.
TNR	True Negative Rate.
FPR	False Positive Rate.
F1	F1 Score.
SPD	Statistical Parity Difference.
DI	Disparate Impact.
AOD	Average Odds Difference.
EOD	Equal Opportunity Difference.
TI	Theil Index.
PCS42	Physical Component Summary Score 42.
MCS42	Mental Component Summary Score 42.

POVCAT	Poverty Category.
INSCOV	Insurance Coverage.
LogReg	Logistic Regression.
RF	Random Forest.
GB	Gradient Boosting.
GAN	Generative Adversarial Network.

Chapter 1

Introduction

1.1 Motivation

An AI model is only as fair as the data it learns from, yet history rarely tells an unbiased story. AI's reliance on vast datasets and complex algorithms means that these systems can unintentionally reflect and amplify existing societal patterns, including historical inequalities. This raises a fundamental challenge: ensuring that AI systems preserve fairness, accountability, and transparency.

Bias, which refers to systematic patterns within AI models that result in unequal or unfair outcomes by favoring or disadvantaging certain groups, plays a central role in this challenge. Bias can emerge at multiple stages of the AI development pipeline, from data collection and preprocessing to model training, algorithm design, and deployment. Each of these stages holds the potential to introduce or increase bias, leading to skewed predictions, classifications, or recommendations. Understanding and addressing bias within AI systems is thus essential to building trustworthy and inclusive technologies that serve society equitably.

Effectively addressing bias in AI requires a careful balance between technical precision and ethical considerations. While rarely introduced intentionally, bias often arises from gaps or patterns that were previously overlooked or deemed acceptable. In real-world applications, these biases can have serious consequences, especially in high-stakes fields like healthcare, finance, and law enforcement, where biased decisions may disproportionately impact certain individuals or communities.

Data-centric AI focuses on tackling these issues at the data level as a foundational step toward fairness. While bias can originate anywhere in the pipeline, a data-centric approach emphasizes ensuring that the data itself is fair, representative, and balanced. This involves refining datasets through profiling, auditing, and augmentation techniques, making data more reflective of diverse populations. By concentrating on the data as a core component of model fairness, data-centric AI enables more robust and unbiased outcomes, reducing the need for corrective measures later in the process.

1.2 Ethical Considerations

The integration of AI systems into critical areas such as healthcare, finance, and criminal justice, has sparked important ethical debates about the implications of these technologies. One of the core ethical concerns surrounding AI is the potential for **biased decision-making**. Algorithms, while often seen as objective, can unintentionally reinforce or worsen present societal inequalities when trained on biased data. This can result in discriminatory outcomes that disproportionately affect marginalized or underrepresented groups, raising serious questions about fairness in AI systems.

The concept of **fairness** itself presents another ethical challenge in AI. Fairness is not a one-size-fits-all concept; what is considered fair can vary greatly depending on the context and the stakeholders involved. In some cases, fairness might mean equal treatment across all groups, while in others, it might involve ensuring that historically disadvantaged groups are given additional support to achieve equitable outcomes. This divergence in defining fairness complicates the task of creating universally accepted fairness standards for AI systems.

The ethical dilemma of how fairness should be implemented is further compounded when the AI system's decision-making processes lack **transparency**. In many instances, AI systems are seen as 'black boxes' that make decisions without offering clear explanations. This lack of transparency highlights the need for **explainability** in AI. Users and affected individuals must be able to understand how and why decisions are made, especially when they lead to potentially life-changing outcomes.

Moreover, **accountability** remains a crucial ethical issue. When AI systems make biased or unfair decisions, determining who is responsible is a complex task. Is it the developers who built the model? The organizations that deploy the system? Or perhaps the data itself, which may be flawed or incomplete? This lack of clear accountability is a significant ethical concern, as individuals and communities affected by biased AI decisions may not have a clear path to take action or receive compensation for the harm they may have suffered.

Data privacy and protection are also fundamental aspects of responsible AI development. The use of sensitive data, such as race, gender, and health information, raises important privacy concerns, especially when such data is used to train machine learning models [5]. Even when efforts are made to anonymize or de-identify data, there is always the risk of re-identification, leading to the potential misuse of personal information [6]. In this context, the use of synthetic data has been proposed as a way to mitigate privacy concerns by generating data that mimics the statistical properties of real data without risking the exposure of sensitive personal information. While synthetic data can help preserve privacy, its use must also be examined from an ethical perspective, as it can introduce new biases if not carefully generated and validated [7].

Another ethical concern in the realm of AI is the potential erosion of **human autonomy**. As AI systems take on more decision-making roles, particularly in areas

such as loan approval or hiring, they can reduce human agency, leading individuals to feel as though their fate is controlled by algorithms rather than their own choices or actions [8]. This shift in decision-making power raises questions about the balance between human judgment and machine-driven processes. When some bias mitigation techniques are applied, they may change the nature of the data, which can reduce the model's overall accuracy. This trade-off between fairness and accuracy presents a further ethical challenge; how to ensure fairness without sacrificing the effectiveness of the system or introducing new forms of harm [9].

The ethical implications of AI also extend to **power dynamics and inequality**. The development and deployment of AI technologies are often controlled by a small number of powerful institutions or corporations, leading to concerns about whether these entities act in the best interests of society [10]. In some cases, AI systems may be designed to reinforce existing power structures or commercial interests, rather than promote fairness and equity [11]. Additionally, marginalized communities may have limited access to AI technologies or may lack the resources to challenge biased systems [12]. This concentration of power in the hands of a few raises fundamental questions about whose interests are being served by AI, and whether the benefits of AI are distributed equitably across society.

Finally, **ethical frameworks and regulations** are needed to ensure that AI is developed and deployed responsibly. Initiatives such as the European Union's AI Act [13] and the OECD's AI Principles [14] aim to provide guidance on how AI systems should be governed, with a focus on ensuring fairness, transparency, and accountability. These regulatory frameworks are an essential step toward addressing the ethical challenges of AI, ensuring that systems are designed to minimize harm and maximize benefits for all stakeholders. However, even with these frameworks in place, researchers and developers must remain vigilant in addressing the ethical risks associated with bias.

In conclusion, the ethical implications of bias detection and mitigation in AI systems are multifaceted and require careful consideration of fairness, transparency, accountability, privacy, and autonomy. As AI systems continue to shape critical aspects of society, it is essential to ensure that these systems are developed and deployed in ways that align with ethical standards and contribute to a fairer, more just society. While this thesis focuses on technical methods for detecting and mitigating bias, it is important to acknowledge that bias is not merely a technical problem but also a social and institutional issue. AI systems do not exist in a vacuum and must be contextualized within the broader societal structures that shape and influence them. Addressing data-specific biases and mitigating their impact through techniques like dataset augmentation is a crucial step in this process, but it must be done with a commitment to transparency, accountability, and a broader understanding of the social consequences of AI decision-making.

1.3 Problem Statement and Research Objectives

This research addresses the pervasive bias in datasets used to train AI models, focusing on how biased datasets lead to biased predictions and unfair treatment of certain groups. In particular, the study highlights the impact of these biases in critical sectors such as healthcare and socio-economic contexts, where AI decisions can result in serious real-world consequences.

To tackle this issue, the research explores methods to measure and mitigate bias in datasets. The role of synthetic data generation is investigated as a potential solution to augment underrepresented groups, balance dataset distributions, and improve fairness in AI models.

The specific objectives of this thesis are:

- To detect and measure bias in datasets using the **AIF360** [1] toolkit, particularly those that involve sensitive attributes like race and gender.
- To explore methods of mitigating bias using the **AIF360** toolkit, through preprocessing techniques such as Disparate Impact Remover, Learning Fair Representations (LFR), Optimized Preprocessing, and Reweighing.
- To evaluate the effectiveness of **AIF360**-based bias mitigation methods in improving fairness and reducing bias in AI models.
- To explore methods of mitigating bias through the generation of synthetic data using **Clearbox AI's Synthetic Kit**, aiming to augment underrepresented groups and balance dataset distributions.
- To evaluate the effectiveness of synthetic data generation with **Clearbox AI's Synthetic Kit** in improving fairness and mitigating bias in AI models.

By achieving these objectives, the thesis aims to contribute to the development of more equitable AI systems, ensuring fairer outcomes across various domains.

1.4 Scope and Limitations

This study specifically concentrates on data-specific biases, such as those arising from historical inequalities, underrepresentation, and demographic imbalances within datasets. The datasets used for analysis include socio-economic data (Adult dataset) and healthcare data (Medical Expenditure dataset), both of which are commonly used for fairness-related experiments.

The effectiveness of the proposed methods is evaluated within the scope of these selected datasets, and results may vary when applied to other types of data or more complex real-world scenarios.

While this research aims to provide valuable insights into the application of dataset augmentation for bias mitigation, it does not cover all possible bias mitigation

techniques. For instance, the study does not focus on in-processing or post-processing methods, nor does it explore all forms of synthetic data generation.

Additionally, while the study addresses ethical considerations in the development and deployment of AI systems, it does not delve into the social or policy implications of these technologies in depth, focusing primarily on the technical aspects of bias detection and mitigation.

Chapter 2

Literature Review

2.1 Overview of Bias in AI

Bias in AI can stem from several sources, and its impact can undermine the fairness and trustworthiness of AI-driven decisions, thereby affecting individuals and society at large. According to NIST Special Publication 1270 [15], bias management in AI requires not only technical interventions but also a comprehensive socio-technical approach that considers the societal, institutional, and human factors that shape AI systems throughout their lifecycle.

The NIST framework categorizes AI bias into three primary types: systemic, statistical, and human biases. Each of these plays a role in perpetuating inequality or unintended harm through AI systems:

- **Systemic Bias** arises from institutional and historical contexts, where societal structures or norms favor certain groups over others. This bias is often reflected in the training data used for AI systems, which may inherit historical inequalities. For instance, ProPublica's 2016 investigation revealed that risk assessment software used in the U.S. criminal justice system was biased against Black defendants. The software's predictions about recidivism risk were more likely to falsely label Black defendants as higher risk, leading to unfair sentencing decisions and reinforcing racial disparities [16]. Another example is Amazon's automated recruiting system, which was found to systematically downgrade resumes that contained terms associated with women, such as participation in "women's chess club" or graduation from all-women's colleges. The system, trained on historical hiring data, learned patterns that favored male candidates, ultimately reinforcing gender disparities [17].

- **Statistical or Computational Bias** is rooted in the data and algorithms used to develop AI models. When datasets are non-representative or skewed toward specific demographics, the resulting models may fail to generalize across different population groups. For example, facial recognition systems trained predominantly on lighter-skinned individuals often perform poorly when identifying people with darker skin tones. This reflects a fundamental imbalance

in the underlying data, leading to disproportionate error rates across different demographic groups [18].

- **Human Bias** refers to the biases introduced by AI developers, data annotators, and end-users. These biases may be implicit or explicit and can affect decisions made during data collection, model development, or system deployment. For example, developers' assumptions about which variables are important for predicting outcomes can inadvertently reflect their own cognitive biases, influencing the fairness and equity of AI systems.

2.2 Bias Detection and Mitigation

The rapid proliferation of machine learning models in critical decision-making areas has raised significant concerns about inherent biases and fairness. To address these issues, researchers have developed a variety of tools for bias detection and mitigation that help assess fairness in predictive models and datasets.

Effective bias detection serves as a crucial initial step, motivating the development of various open-source libraries designed to evaluate and uncover biases in predictive models. *Aequitas* [19] is a comprehensive toolkit designed for both data scientists and policymakers. It offers a Python library along with a web platform for uploading datasets for bias analysis. Aequitas includes fairness metrics such as demographic parity and disparate impact, as well as a "fairness tree" to guide users in selecting the appropriate metric for their specific case. Similarly, *Fairness Measures* [20] provides metrics like the difference of means, disparate impact, and odds ratio, though its dataset offerings are more limited, with some datasets requiring explicit permission for access. *FairTest* [21] is another framework designed to identify and test for unwarranted associations between an algorithm's outputs and specific user subpopulations defined by protected features. This methodology highlights areas within the input space where models tend to make disproportionately high error rates, thus identifying potential sources of bias. *FairML* [22] also serves a tool designed to audit machine learning models by quantifying the significance of model inputs to evaluate their fairness. It employs model compression and input ranking algorithms to facilitate the detection of bias in predictive models, enabling analysts to assess discriminatory tendencies effectively. Additionally, *Themis* [23] is a testing framework designed to measure software discrimination by generating efficient, automated test suites based on valid input schemas. Unlike traditional methods, it does not rely on a predefined reference point (oracle), allowing for a thorough evaluation of predictive models for fairness in diverse decision-making processes.

These libraries focus primarily on detecting bias in machine learning systems, but they often do not include techniques for bias mitigation. To address this, several fairness toolkits have been developed that provide both bias detection and mitigation capabilities. *AI Fairness 360 (AIF360)* [1] is one such comprehensive toolkit that offers a full suite for bias detection and mitigation. It includes various fairness

measures, such as statistical parity difference, equalized odds, and disparate impact, and integrates several mitigation strategies. AIF360 stands out for its ability to combine multiple techniques found in other libraries, making it a powerful resource for addressing fairness issues across the machine learning pipeline.

Another well-established toolkit is *Fairness Comparison* [24], which includes a broad collection of fairness metrics and mitigation methods. This toolkit provides approaches like the disparate impact remover and prejudice remover, as well as a two-Naive Bayes method. It functions as a test-bed to compare different algorithms, ensuring consistency across datasets and fairness metrics. Similarly, *Themis-ML* [25] provides fairness metrics like mean difference and offers mitigation methods such as relabeling, the additive counterfactually fair estimator, and reject option classification. These techniques address fairness concerns in machine learning models with a more targeted approach.

Building on the foundation established by frameworks above, bias mitigation methods operate at three critical stages of the machine learning pipeline: preprocessing, in-processing, and post-processing. These methods address bias systematically, targeting data preparation, model training, and output adjustment to promote fairness.

Preprocessing methods intervene before the learning process by modifying the input data to reduce inherent biases. Techniques like *Disparate Impact Removal* [26] adjust data distributions to balance representation while maintaining predictive performance. *Learning Fair Representations* [27] encodes data in a way that retains essential information while obfuscating protected attributes, thus preventing models from exploiting these attributes in decision-making. *Optimization-based preprocessing* frameworks [28] strategically modify data distributions to minimize discrimination while preserving the utility of the dataset. For this thesis, we primarily utilized these preprocessing methods to address bias in the input data prior to training the models.

In-processing methods, on the other hand, intervene during the model training process to ensure fairness constraints are met. *Adversarial Learning* [29] is one such technique that incorporates an adversarial network to reduce the dependency between sensitive attributes and predictions, thus mitigating bias during the learning phase. Similarly, *Fairness-Aware Classifiers with Prejudice Removers* [30] apply fairness regularizers to penalize discriminatory behaviors during training, ensuring that fairness is incorporated into the learning process without sacrificing model accuracy.

Finally, post-processing techniques adjust the model's outputs after training to ensure fairness. *Equalized Odds Postprocessing* [31] modifies decision thresholds to ensure equal true positive and false positive rates across different groups, thereby promoting fairness in the final model outputs. *Reject Option Classification* [32] reassigns outcomes in uncertain cases, allowing the model to correct biased predictions.

By combining robust bias detection tools with targeted mitigation methods at different stages of the machine learning pipeline, researchers continue to drive progress toward developing more equitable and trustworthy AI systems.

2.3 Synthetic Data

Synthetic data refers to artificially generated data that mimics the statistical properties of real data while ensuring privacy, fairness, and scalability. Unlike traditional data anonymization techniques, which modify existing datasets, synthetic data is created from scratch using algorithms that capture the underlying distributions and correlations of real data. This ensures that synthetic datasets maintain utility while preventing privacy risks associated with re-identification [7].

The use of synthetic data is particularly valuable in scenarios where real data is scarce, privacy-sensitive, or imbalanced. In the context of bias mitigation, synthetic data offers a powerful tool for balancing datasets and enhancing fairness by generating realistic, diverse, and representative samples.

2.3.1 Types of Synthetic Data Generation

There are three primary methods for generating synthetic data [7]:

- **Synthesis from Real Data**: A generative model is trained on an existing dataset to learn its statistical properties, then used to produce new, realistic samples. This ensures that the synthetic data retains similar distributions while reducing the risk of privacy violations. This is the method employed in this thesis.
- **Synthesis Without Real Data**: In this approach, synthetic data is generated based on predefined rules, theoretical distributions, or simulations. This method is useful when real data is unavailable or unreliable.
- **Hybrid Synthesis**: This technique combines elements of both approaches, incorporating real-world patterns while introducing modifications to enhance fairness and privacy.

2.3.2 Applications of Synthetic Data

The use of synthetic data has grown significantly in recent years, as it helps address several data access challenges across multiple industries [7]. Synthetic data generation allows organizations to simulate realistic data in situations where obtaining real data is difficult, expensive, or privacy-sensitive.

In manufacturing, synthetic data supports training industrial robots for complex tasks by creating diverse and realistic training datasets without the need for manual data collection. An example is NVIDIA's use of a graphics-rendering engine to simulate images for training robots to play dominoes, demonstrating the power of synthetic data in cost-effectively building training models [33].

In healthcare, synthetic data has resolved issues related to data privacy and access. For instance, data used for cancer research has been made publicly available through synthetic datasets, helping researchers comply with privacy regulations while

still enabling valuable analyses [34]. Synthetic data enables the use of complex, open data, which would otherwise be difficult to share due to re-identification risks [6].

Synthetic data also plays a crucial role in the financial services industry. It helps with tasks such as testing fraud detection algorithms and creating standardized benchmarks for evaluating software and hardware solutions. For example, the STAC-A2 benchmark in financial market risk modeling uses synthetic data to allow companies to compare solutions on a consistent basis, offering a more cost-effective and privacy-preserving method for evaluating new technologies.

In transportation, synthetic data is used in microsimulation models to evaluate the impact of infrastructure changes, such as new bridges or malls [35, 36, 37]. It is also essential for training autonomous vehicles, where real-world data cannot cover all edge cases. Using synthetic data, engineers can generate diverse and customizable driving scenarios, enabling thorough training and testing without the high cost and risk of using real-world environments. [1]

Overall, synthetic data is being increasingly adopted across industries such as manufacturing, healthcare, financial services, and transportation. As data-access challenges persist, the use of synthetic data is expected to expand, providing effective solutions to privacy, cost, and data availability issues while enhancing the capabilities of AI and machine learning systems.

2.3.3 Synthetic Data for Bias Mitigation

Synthetic data generation has become an essential approach for mitigating bias in machine learning systems. By creating artificial datasets that mirror the statistical patterns of real-world data, researchers can address disparities and ensure fairness while preserving data utility and privacy.

One of the methods in this field is BayesBoost, introduced by Draghi et al. [38]. This approach combines Bayesian probabilistic models with synthetic data generation to address biases in datasets. It identifies underrepresented groups in data and oversamples them using Bayesian networks, creating synthetic datasets that better reflect the original population's distribution.. BayesBoost has shown significant improvements in fairness metrics such as AUC and ROC curves, making it an effective tool for privacy preservation and bias identification. Despite its potential, the method depends heavily on high-quality input data, which may require extensive preprocessing for optimal performance.

Building upon the understanding of causal relationships, DECAF, developed by Van Breugel et al. [39], introduces a GAN-based method that generates fair synthetic data by leveraging Structural Causal Models (SCMs). It embeds structural causal models into the generator, enabling inference-time debiasing through the removal of biased edges in the causal graph. This approach ensures fairness in downstream machine learning models by eliminating bias at the data generation stage while maintaining data utility. This framework ensures a theoretically robust approach

[1] Some of the references are secondary sources cited in *Practical Synthetic Data Generation* by Khaled El Emam et al. [7], where relevant papers are discussed within the book.

to bias mitigation but requires expert knowledge in causal modeling, presenting challenges for broader adoption. Another notable framework is GenEthos, introduced by Gujar et al. [40]. Integrating Generative Adversarial Networks (GANs) with an interactive Graphical User Interface (GUI), GenEthos combines bias detection with synthetic data generation. It has demonstrated significant fairness improvements, including reductions in Statistical Parity Difference (SPD) by up to 93%. However, its scope is limited to datasets like the German Credit and Adult datasets, raising concerns about generalizability across diverse applications. Adding experimental control to synthetic data generation, Baumann et al. [41] propose a toolkit for generating synthetic data with predefined biases. This open-source framework allows researchers to model various types of biases and analyze their impact on AI systems. While highly beneficial for controlled studies, the generated data may lack the complexity required to reflect real-world scenarios, limiting its utility in broader contexts. In conclusion, these methods illustrate the versatility of synthetic data generation in bias mitigation. Each technique offers unique advantages for addressing disparities and improving fairness. However, their effectiveness is influenced by factors such as the quality of input data, the computational resources required, and the expertise needed for implementation. Continued advancements in these methodologies will play a crucial role in fostering equitable and reliable AI systems.

Chapter 3

Methodology

This research was conducted in collaboration with Clearbox AI, a company specializing in synthetic data generation and data-centric AI solutions. This chapter provides an overview of the datasets, classifiers, bias detection, mitigation techniques, and evaluation metrics employed, aiming to assess the effectiveness of various approaches to promoting fairness in machine learning.

3.1 Datasets

For this research, two datasets were chosen to explore bias in machine learning models: one representing the socio-economic domain (*Adult dataset*) [3] and the other representing the healthcare domain (*Medical Expenditure dataset*) [4]. These datasets were chosen for their real-world relevance and their ability to illustrate how bias manifests in different fields.

3.1.1 Adult Dataset

The *Adult dataset*, also known as the *Census Income dataset*, was originally compiled by Barry Becker from the 1994 U.S. Census database. This dataset consists of 48,842 instances and 14 attributes, with the goal of predicting whether an individual's annual income exceeds \$50K based on demographic and work-related factors. The dataset classifies individuals into two categories: those earning more than \$50K (positive class) and those earning less than or equal to \$50K (negative class).

The dataset includes several important features related to demographics, education, work, and relationships. Key demographic features include age, sex, and race, while work-related attributes such as hours worked per week and employment type provide context for income predictions. Other variables, such as marital status and relationship status, are also included.

3.1.2 Medical Expenditure Dataset

The *Medical Expenditure dataset* used in this study is derived from the *2015 Full Year Consolidated Data File*. The data comes from Panel 19 of the *Medical Expenditure*

Panel Survey (MEPS), which collects comprehensive information on healthcare utilization, costs, and demographic factors.

The primary target variable is *UTILIZATION*, a composite feature representing the total number of medical visits across various categories, including office visits, outpatient visits, emergency room visits, inpatient nights, and home health visits. A classification task is to predict whether a person will have high utilization ($UTILIZATION \geq 10$).

In addition to *UTILIZATION*, the dataset includes key health-related features such as physical health scores (*PCS42*) and mental health scores (*MCS42*). Socio-economic factors like poverty categories (*POVCAT*) and insurance coverage (*INSCOV*) are also included. The dataset further encompasses features related to medical history (chronic diseases and mental health conditions), activity limitations (walking, social, cognitive), and marital and family status (e.g., presence of children living with the individual).

3.1.3 Key Features of Both Datasets

In both datasets, a few key features were selected to provide critical insight into the factors contributing to bias and fairness in predictive models. For the *Adult dataset*, the selected features include demographic attributes such as *race*, *sex*, and *age*, as well as *education years*. The latter two features were preprocessed by grouping *age* into decades and one-hot encoding the *education years* categories. These features were specifically chosen as they are central to understanding income disparities.

For the *Medical Expenditure dataset*, the selected features include *physical health scores* (*PCS42*), *mental health scores* (*MCS42*), *age* (grouped into decades), and socio-economic factors such as *poverty categories* (*POVCAT*) and *insurance coverage* (*INSCOV*). These features provide a comprehensive view of the factors affecting healthcare utilization. Both datasets share common sensitive attributes, namely *race* and *sex*, which are crucial for investigating fairness in predictions related to income and healthcare.

3.1.4 Dataset Insights

Both datasets exhibit key challenges relevant to the study of fairness. In the *Adult dataset*, there is a significant **demographic skew**, with 85.5% of the population being White and 66.8% being Male. This imbalance in race and sex distribution contributes to potential bias in income predictions. Furthermore, the **class imbalance** in the dataset—76.1% of individuals earning less than \$50K—poses another challenge for training accurate and fair models.

The *Medical Expenditure dataset* also has its challenges, notably **class imbalance**, with only 17% of the data representing individuals with high healthcare utilization (UTILIZATION ≥ 10). This imbalance highlights the need for strategies that ensure fairness, especially when dealing with underrepresented groups in predictive models.

3.2 Bias Detection Metrics

Bias detection metrics are designed to assess fairness in machine learning models by quantifying how different groups (defined by sensitive attributes like race, sex, etc.) are treated in the model's predictions. These metrics evaluate whether the outcomes for privileged and unprivileged groups are equitable. The goal is to detect disparities that could indicate bias in the model's predictions, thus ensuring that the model behaves fairly across different demographic groups. In this research, the following fairness metrics, sourced from the AI Fairness 360 (AIF360) toolkit [1], are used to assess and quantify biases:

- **Statistical Parity Difference:** This metric calculates the difference in the probability of favorable outcomes between the unprivileged and privileged groups.

$$\Pr(\hat{Y} = 1 \mid D = \text{unprivileged}) - \Pr(\hat{Y} = 1 \mid D = \text{privileged})$$

- **Disparate Impact:** This metric quantifies the ratio of favorable outcomes between the unprivileged and privileged groups.

$$\frac{\Pr(\hat{Y} = 1 \mid D = \text{unprivileged})}{\Pr(\hat{Y} = 1 \mid D = \text{privileged})}$$

- **Average Odds Difference:** This metric evaluates the average difference in False Positive Rate (FPR) and True Positive Rate (TPR) between the unprivileged and privileged groups. The formula is:

$$\frac{1}{2}\left[(FPR_{\text{unprivileged}} - FPR_{\text{privileged}}) + (TPR_{\text{unprivileged}} - TPR_{\text{privileged}})\right]$$

- **Equal Opportunity Difference:** This metric computes the difference in True Positive Rate (TPR) scores between unprivileged and privileged groups. It is defined as:

$$TPR_{\text{unprivileged}} - TPR_{\text{privileged}}$$

- **Theil Index:** This is an inequality metric used to measure the disparity in outcomes across groups. The Theil index is defined as:

$$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{b_i}{\mu}\ln\frac{b_i}{\mu}\right)$$

  where b_i are the values for the group, and μ is the mean value.

These metrics offer a way to evaluate and ensure fairness by quantifying potential biases. When the differences between groups are large according to these metrics, it suggests the need for adjustments in the model or dataset to mitigate these disparities.

3.3 Bias Mitigation Techniques

Bias mitigation algorithms are designed to improve fairness in machine learning models by addressing disparities in data or predictions. These algorithms can intervene at different stages of the machine learning pipeline: pre-processing, in-processing, and post-processing, as illustrated in Figure 3.1. Pre-processing algorithms focus on modifying the training data before it is used to train a model, which is the focus of this thesis. These techniques aim to mitigate bias by transforming the data in a way that reduces unfairness, while maintaining its overall utility for learning tasks.

In this research, the following bias mitigation techniques are used:

- Reweighing
- Disparate Impact Remover
- Learning Fair Representations (LFR)
- Optimized Preprocessing
- Dataset Augmentation through Synthetic Data

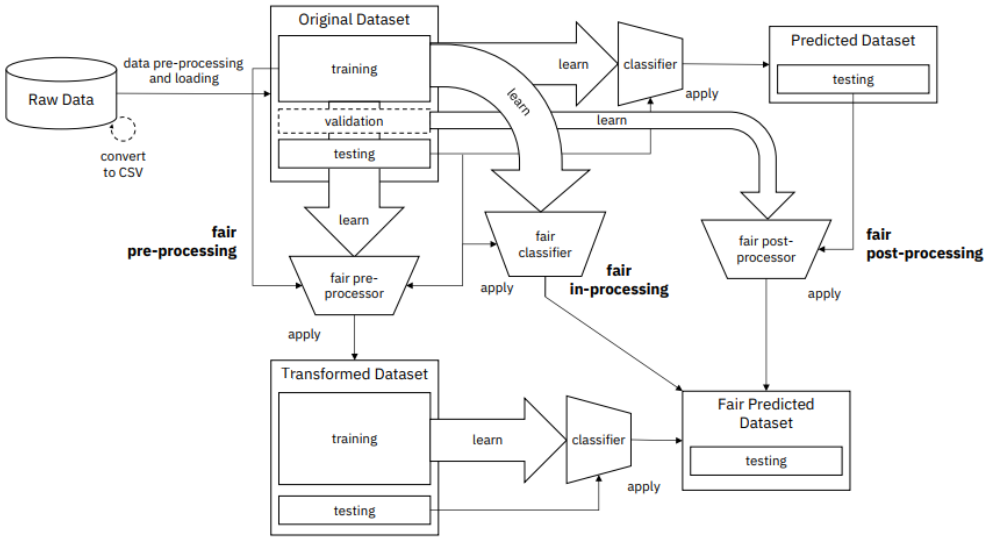

Figure 3.1: The fairness pipeline followed in this thesis, which focuses on the pre-processing phase. The diagram illustrates the process of transforming the original dataset into a fairer version using fairness preprocessing algorithms. This transformed dataset is then used for training a classifier, and fairness metrics are evaluated on both the original and transformed datasets. (Taken from [1]).

3.3.1 Reweighing

Reweighing is a preprocessing technique designed to ensure fairness in the dataset by adjusting the weights of examples across different (group, label) combinations. The approach, proposed by Kamiran and Calders (2012) [42], aims to balance the

dataset by assigning different weights to privileged and unprivileged groups for each class label. By doing so, the technique helps mitigate bias stemming from underrepresentation or overrepresentation of certain groups, thus creating a fairer dataset for subsequent classification tasks.

The main parameters for the Reweighing algorithm include:

- **unprivileged_groups**: A list of dictionaries defining the unprivileged groups in the dataset. This typically includes groups that have historically faced disadvantages, such as specific racial or gender demographics.
- **privileged_groups**: A list of dictionaries representing the privileged groups in the dataset, often those that have traditionally held advantages in societal structures.

3.3.2 Disparate Impact Remover

The Disparate Impact Remover (DIR) is a preprocessing technique that seeks to enhance group fairness by editing feature values associated with sensitive attributes, as proposed by Feldman et al. (2015) [26]. This method modifies specific feature values to reduce the disparate impact across groups while maintaining the relative rank order within each group. By doing so, the technique preserves meaningful feature relationships while improving fairness, reducing the risk of unintentional bias influencing the model.

The Disparate Impact Remover includes the following key parameters:

- **repair_level**: A parameter controlling the extent of repair applied to the sensitive attribute values. A repair level of 0.0 indicates no modification, while a level of 1.0 represents full repair, maximizing fairness adjustments.
- **sensitive_attribute**: Specifies the protected attribute in the dataset (e.g., race or gender) that the algorithm will modify to achieve fairness.

3.3.3 Learning Fair Representations (LFR)

Learning Fair Representations (LFR) is a pre-processing technique designed to mitigate bias by transforming data into a latent representation that is both predictive of the target variable and independent of sensitive attributes. This method was first introduced by Zemel et al. (2013) [27] and remains a foundational approach to achieving fairness in machine learning systems.

The LFR algorithm seeks to encode the input data in a way that obfuscates information about protected attributes (e.g., gender or race) while retaining the utility of the data for downstream tasks. This is accomplished by balancing three key objectives:

- **Input Reconstruction Quality (A_x)**: Ensuring the latent representation retains sufficient information to reconstruct the original input data accurately.

- **Fairness Constraint (A_z)**: Penalizing the model for retaining discriminatory information about protected attributes.
- **Prediction Accuracy (A_y)**: Maintaining the model's ability to predict the target variable effectively.

3.3.4 Optimized Preprocessing

Optimized Preprocessing (OptimPreproc) is a pre-processing technique that learns a probabilistic transformation to adjust features and labels for group fairness. Proposed by Calmon et al. (2017) [28], it minimizes disparities between privileged and unprivileged groups without significantly distorting individual data points. The method uses an optimization framework to find the best transformation based on fairness metrics and constraints.

Some of the key parameters involved in the OptimPreproc algorithm are:

- **optimizer**: The optimizer class used to perform the optimization. The optimization framework seeks to minimize unfairness while ensuring data fidelity.
- **optim_options**: A dictionary of options used to configure the optimization process, such as hyperparameters and constraints for fairness and distortion minimization.
- **unprivileged_groups**: A representation of the unprivileged groups in the dataset.
- **privileged_groups**: A representation of the privileged groups in the dataset.

3.3.5 Dataset Augmentation through Synthetic Data

In this research, we utilize **Clearbox AI's Tabular Engine** from Synthetic Kit [2] for synthetic data generation. This tool is designed to generate and evaluate synthetic data, particularly for tabular datasets, ensuring that it preserves the statistical properties of the original data while safeguarding privacy. At the core of the tool is the **TabularEngine**, which leverages a **Variational Autoencoder (VAE)** model to learn the underlying distributions of the original dataset. This engine performs encoding and decoding tasks, enabling the generation of synthetic data that mirrors the original dataset without revealing sensitive information.

Before synthetic data generation, the tool preprocesses the data using the **Preprocessor** class, which transforms the dataset into a suitable format for the VAE model. The preprocessing step handles various feature types such as ordinal, categorical, and datetime, ensuring that the relationships between features are preserved. Ordinal features are discretized, categorical features are one-hot encoded, and datetime features are converted to numerical representations.

Once the data is preprocessed, the **LabeledSynthesizer** class uses the trained model to generate synthetic instances. These synthetic data points are created in

such a way that they retain the statistical properties of the original dataset, while ensuring that they do not re-identify individuals or reveal sensitive details. This is particularly important in use cases where data privacy is a concern.

Overall, ClearBox AI provides a comprehensive framework for generating high-quality synthetic data. By combining advanced machine learning models with robust privacy and utility evaluations, the tool ensures that synthetic data can be safely used without compromising on its usefulness for machine learning tasks.

3.4 Classifiers

In this research, three different classifiers were employed to assess the effectiveness of bias mitigation techniques on model performance:

3.4.1 Logistic Regression

Logistic Regression is a simple yet powerful classifier commonly used for binary classification tasks. It models the relationship between a binary target variable and one or more predictor variables using the logistic function. In this study, Logistic Regression outputs probabilities for each class, and a decision threshold was identified using the validation set to achieve the best trade-off between fairness and performance. While a threshold of 0.5 is traditionally used, here it was adjusted dynamically based on validation data to improve results. This classifier was chosen for its interpretability and efficiency, especially for datasets with linear decision boundaries.

3.4.2 Random Forest

Random Forest is an ensemble learning method that constructs multiple decision trees and aggregates their predictions (via majority voting) for classification tasks. This classifier is effective at modeling non-linear relationships and robust to overfitting. The decision thresholds were fine-tuned using validation data to account for potential imbalances and to improve fairness metrics. Random Forest was selected for its versatility and ability to handle complex, high-dimensional datasets.

3.4.3 Gradient Boosting

Gradient Boosting is an ensemble learning method that sequentially builds decision trees, with each tree trained to minimize the errors of the previous one. This iterative process enables Gradient Boosting to capture intricate patterns in the data. Similar to the other classifiers, the decision thresholds for Gradient Boosting were determined using validation data to enhance both fairness and predictive performance. This classifier is particularly suited for capturing non-linear relationships between features and target variables.

These three classifiers were selected to compare how different model architectures perform on the same tasks, enabling a comprehensive evaluation of the impact of bias mitigation techniques across multiple approaches.

3.5 Performance evaluation

The performance of the classifiers is evaluated using **Balanced Accuracy**, a metric that is particularly useful when dealing with imbalanced datasets. Balanced accuracy helps to ensure that the model performs well on both the positive and negative classes, providing a more equitable measure of model performance than traditional accuracy.

Balanced accuracy is defined as the average of the **True Positive Rate (TPR)** and the **True Negative Rate (TNR)**:

$$\text{Balanced Accuracy} = \frac{TPR + TNR}{2}$$

where:

$$TPR = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}, \quad TNR = \frac{\text{True Negatives}}{\text{True Negatives} + \text{False Positives}}$$

In the context of imbalanced datasets, where one class significantly outnumbers the other, traditional accuracy can be misleading. For instance, a model that always predicts the majority class can still achieve high accuracy but fail to predict the minority class correctly. Balanced accuracy addresses this by considering the model's performance on both classes, ensuring that the classifier does not disproportionately favor the majority class.

3.6 Tools and Frameworks

The research relied on several key software libraries and tools, which enabled efficient data processing, model training, and bias mitigation:

- **Python** was the primary programming language used throughout the project, providing flexibility for implementing various techniques and algorithms.
- **AIF360 (AI Fairness 360)** [1], an open-source toolkit developed by IBM, was central to the analysis, offering a comprehensive set of fairness metrics and bias mitigation strategies, which were crucial for evaluating and addressing fairness in the datasets.
- **Scikit-learn** served as the foundation for machine learning tasks, providing a variety of classifiers such as Logistic Regression, Random Forest, and Gradient Boosting, used for both training and evaluating the models.
- **Pandas** and **NumPy** were essential for data manipulation, cleaning, and transformation, enabling the preprocessing of the datasets for further analysis.
- **Matplotlib** was employed for visualizations, helping to illustrate the performance of the bias mitigation methods and provide insights into the results.

- **Jupyter Notebooks** facilitated an interactive workflow, allowing for iterative experimentation, code documentation, and presentation of results.

Synthetic data generation was performed using **Docker** to run the open-source **ClearBox Synthetic Kit** [2], which includes the Engine and Synthesizer modules.

The remaining experiments were conducted on a **local environment**, as the hardware resources were sufficient for processing the data and running the models effectively.

All the code used for data preprocessing, bias detection, bias mitigation techniques, synthetic data generation and their evaluation is available in this GitHub repository.

Chapter 4

Implementation

This chapter provides a detailed overview of the steps taken to implement bias mitigation techniques and evaluate fairness in the datasets. The implementation process is structured around the two main methods of bias mitigation: AIF360-based techniques and synthetic data augmentation. The chapter is organized into sections that cover dataset preprocessing, initial bias assessment, bias mitigation methods, and classifier training and evaluation.

Figure 4.1 provides a visual representation of the implementation workflow used in this research. It highlights the relationships between the datasets, bias mitigation techniques, and classifiers. The process begins with two distinct datasets—Adult and Medical Expenditure—which undergo preprocessing and are then evaluated using three classifiers: Logistic Regression, Random Forest, and Gradient Boosting. Additionally, bias mitigation methods such as Reweighing, Disparate Impact Remover (DIR), Learning Fair Representations (LFR), Optimized Preprocessing (OptimPreproc), and Synthetic Data Augmentation are applied to the datasets, creating enhanced or transformed versions. Both the original unenhanced datasets and the enhanced datasets are assessed using the classifiers, enabling a comparative analysis to determine the effectiveness of bias mitigation in improving fairness and maintaining model performance.

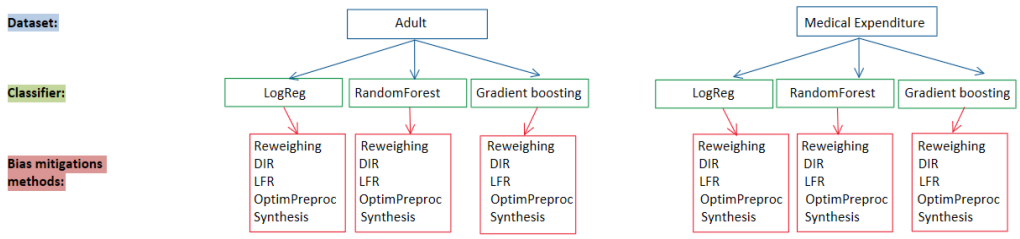

Figure 4.1: Dataset, Bias Mitigation Techniques, and Classifiers Structure

4.1 Data Overview and Insights

This section aims to uncover key patterns, distributions, and potential biases within the Adult and Medical Expenditure datasets. By analyzing demographic and socioe-

conomic characteristics alongside positive outcome rates, we provide insights that may influence predictive modeling and decision-making processes. Below, we present our findings for each dataset.

4.1.1 Adult

The gender distribution within the Adult dataset exhibits a significant imbalance, with male individuals being overrepresented compared to female individuals (Figure 4.2). Specifically, males account for approximately two-thirds of the dataset, with 30,527 male individuals and 14,695 female individuals. This imbalance may introduce biases in predictive models, particularly for outcomes like income and employment opportunities.

The disparity in gender representation correlates strongly with differences in positive outcome rates (Figure 4.3). Males exhibit a significantly higher positive outcome rate of approximately 0.31, whereas females show a much lower rate of around 0.11. This threefold difference suggests that gender is a key determinant influencing these results. Such disparities may reflect systemic factors, including biases in data collection, unequal opportunities, or broader societal influences. Addressing these imbalances is crucial for ensuring that models trained on this dataset do not perpetuate existing inequalities. By examining the root causes behind these trends, we can better understand how to account for gender disparities in predictions and improve model fairness.

The race distribution in the Adult dataset reveals a striking imbalance, with 38,903 individuals identified as "White" and only 6,319 categorized as "Non-white" (Figure 4.4). This skewed demographic composition highlights a disproportionate overrepresentation of White individuals, which may introduce biases in model predictions. The underrepresentation of Non-white individuals could result in less accurate or equitable outcomes for these groups, potentially perpetuating disparities if not adequately addressed.

This imbalance is further reflected in positive outcome rates across racial groups, as shown in Figure 4.5. White individuals exhibit a higher positive outcome rate of approximately 0.26, compared to a lower rate of around 0.16 for Non-white individuals. This disparity suggests that race is a significant factor influencing outcomes within the dataset. Understanding the root causes of these differences is critical for mitigating potential biases. Structural elements, such as unequal access to resources, opportunities, or systemic advantages, may contribute to the observed disparities. Careful analysis and interventions are required to ensure that predictive models trained on this dataset promote fairness and inclusivity.

The age distribution in the Adult dataset exhibits notable trends, with the majority of individuals falling within the 20 to 50 age range and a high concentration between 30 and 40 years (Figure 4.6). Beyond the age of 50, the number of individuals steadily declines, especially after the age of 70. This pattern may reflect both natural demographic trends, where fewer individuals belong to older age groups, and potential

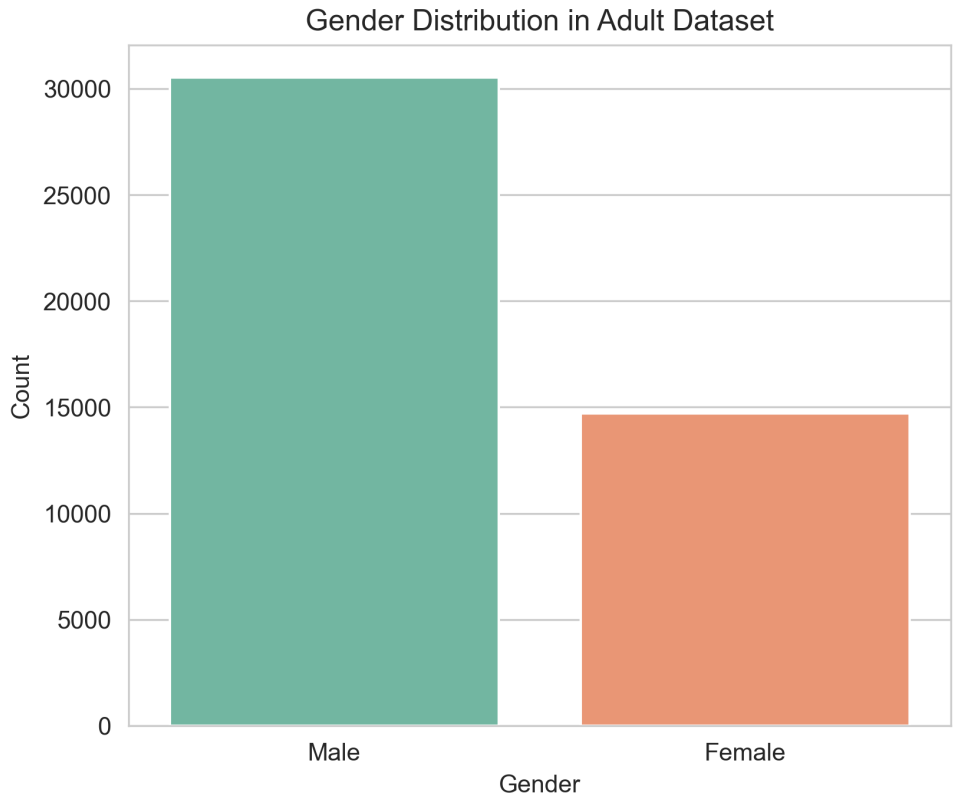

Figure 4.2: Gender distribution of individuals in Adult dataset.

biases introduced during data collection.

The relationship between age and positive outcomes reveals distinct patterns across age groups and genders (Figure 4.7). Positive outcomes steadily increase during early adulthood, peaking in the 40–50 age range. Beyond this peak, outcomes decline across all genders, with the sharpest drops observed among females in later years. Males consistently outperform females in positive outcome rates across all age groups, with the gender gap most pronounced in the 20–50 age range. While positive outcomes for both genders are lowest in the 0–20 age group, males exhibit slightly higher rates. This disparity becomes more evident as individuals enter their 30s and 40s, highlighting the intersection of age and gender as critical factors influencing outcomes in this dataset.

In addition to age, education levels in the dataset reveal important trends that influence socioeconomic factors. The distribution of education levels in the dataset follows a structured pattern, with distinct peaks at 9, 10, and 13 years of education (Figure 4.8). These peaks correspond to common educational milestones, particularly high school completion (typically 9-10 years) and some college or undergraduate-level education (13+ years). The presence of these patterns suggests that the dataset includes a diverse range of educational backgrounds but is primarily concentrated around standard academic progressions. Understanding the education distribution is crucial, as it directly correlates with employment opportunities and income levels. If

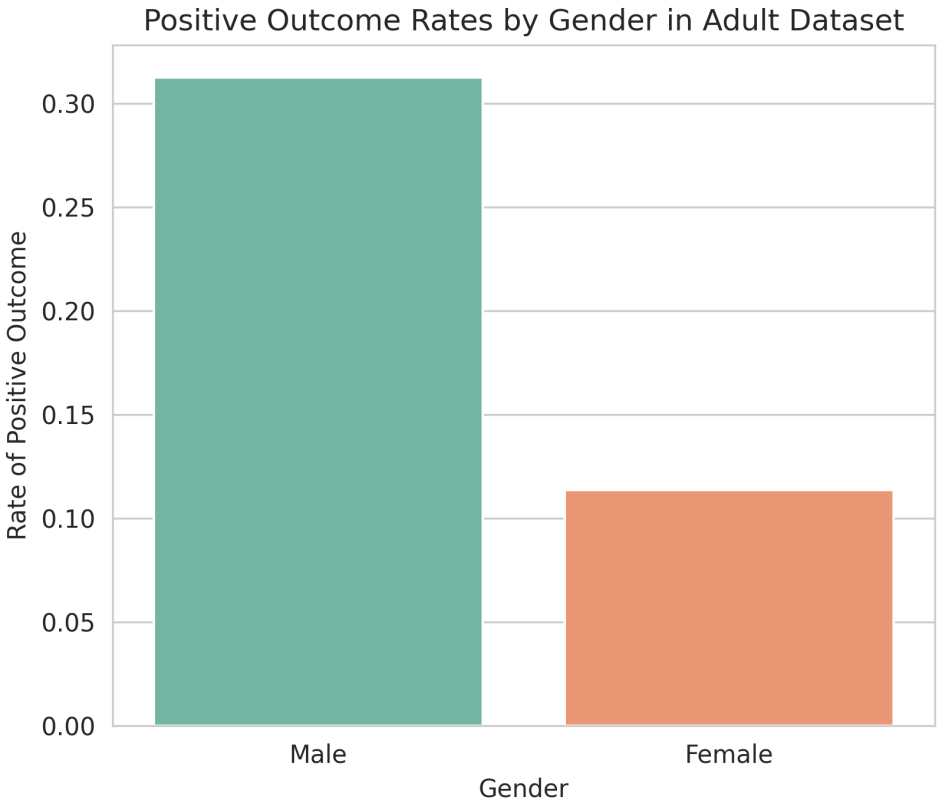

Figure 4.3: Rate of positive outcomes by gender in Adult dataset.

the dataset lacks sufficient representation of individuals with lower or higher education levels, it could introduce biases in models predicting socioeconomic outcomes.

The positive outcome rates across education levels reinforce the significance of educational attainment (Figure 4.9). Positive outcomes increase consistently with higher education levels, underscoring the critical role of education in achieving favorable results. Males outperform females in positive outcome rates at all education levels, with the disparity widening at higher levels of education. This pattern highlights the intersection of education and gender as key factors influencing outcomes in the dataset.

4.1.2 Medical Expenditure

The race distribution in the MEPS dataset reveals that Non-white individuals make up a larger portion of the dataset, with 10,174 Non-white individuals compared to 5,656 White individuals (Figure 4.10). Despite this, disparities in positive medical utilization rates persist across racial groups. As shown in Figure 4.11, White individuals exhibit a significantly higher rate of medical service utilization (25.5%) compared to Non-white individuals (12.5%). This difference may stem from factors such as healthcare accessibility, socioeconomic conditions, or systemic biases. These trends underline a disparity that disproportionately affects Non-white individuals in

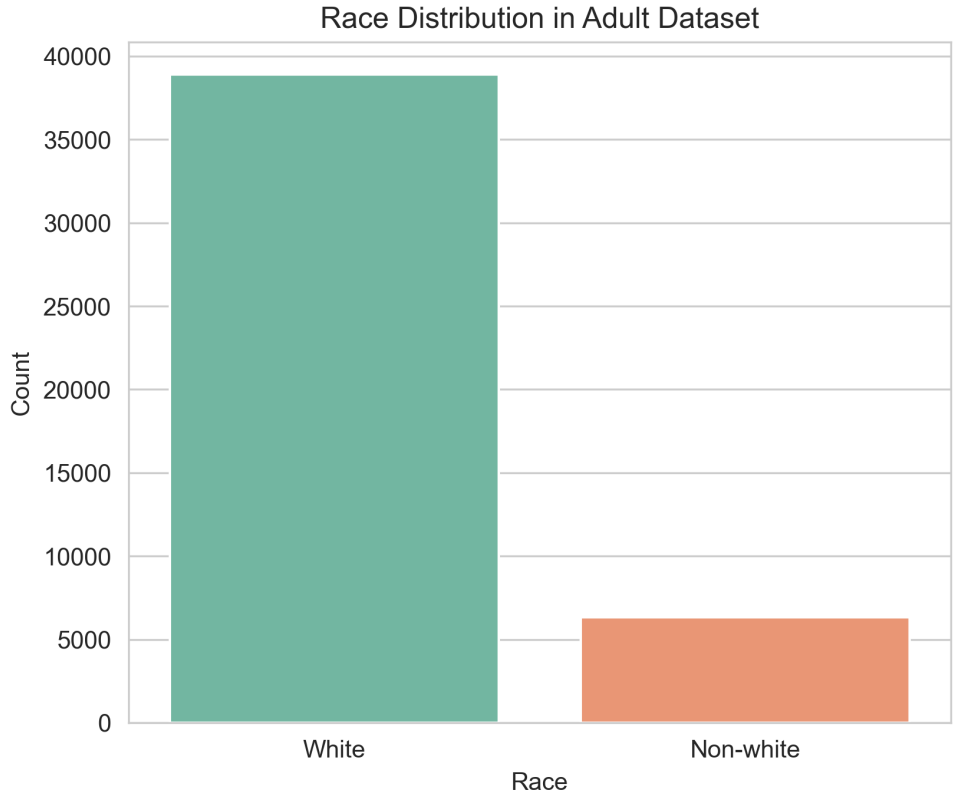

Figure 4.4: Race distribution in Adult dataset.

terms of healthcare utilization.

The MEPS dataset demonstrates a relatively balanced gender distribution, with females comprising 52.1% (8,250 individuals) and males 47.9% (7,580 individuals) (Figure 4.12). This balance ensures adequate representation of both genders for meaningful analysis.

The age distribution within the dataset reflects a relatively even spread, with higher concentrations in childhood, early adulthood, and middle age (Figure 4.13). Older age groups show a gradual decline in representation. Positive medical utilization rates increase with age, as depicted in Figure 4.14. The youngest age group (0–20 years) has the lowest rate of utilization (0.1), while the highest rate (0.5) is observed in the 70+ age group. This reflects the general trend that older individuals tend to use healthcare services more frequently due to the increased likelihood of chronic conditions, age-related health issues, and the need for regular medical attention.

Poverty categories in the dataset are divided into five levels, ranging from "Negative or Poor" (less than 100% of the poverty line) to "High Income" (greater than or equal to 400%). Figure 4.15 illustrates the rate of positive medical utilization across these categories. The utilization rate is consistent across Categories 1 to 4, ranging from 15.5% to 15.7%. However, individuals in the "High Income" category exhibit a significantly higher utilization rate (21.7%), reflecting the positive impact of financial stability and access to healthcare services.

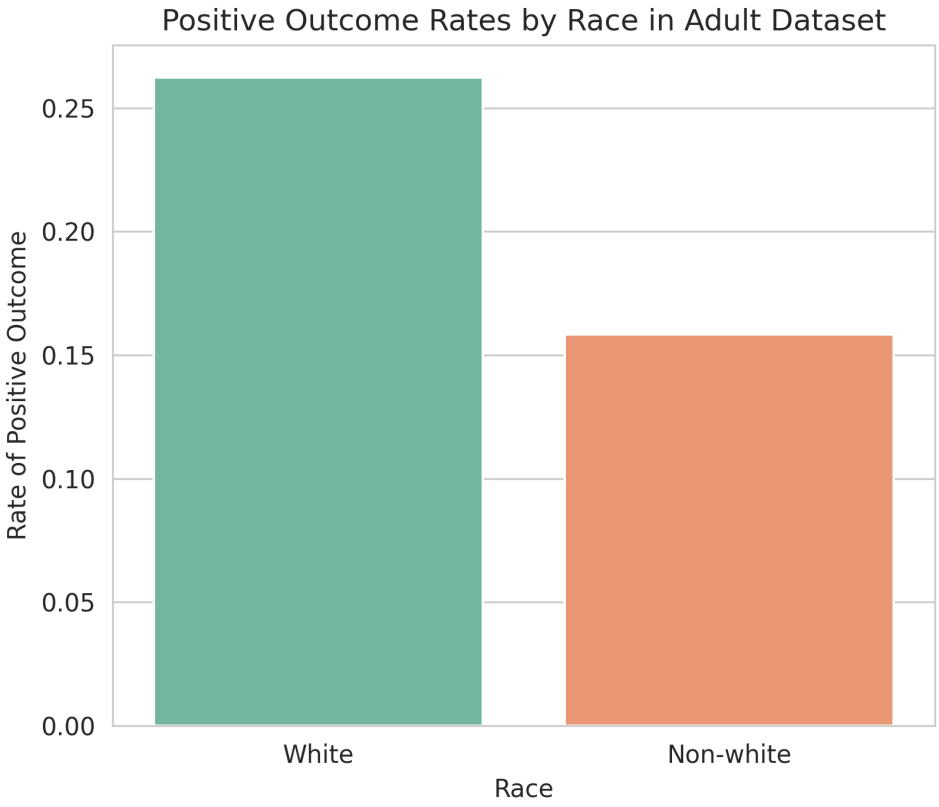

Figure 4.5: Rate of positive outcomes by race in Adult dataset.

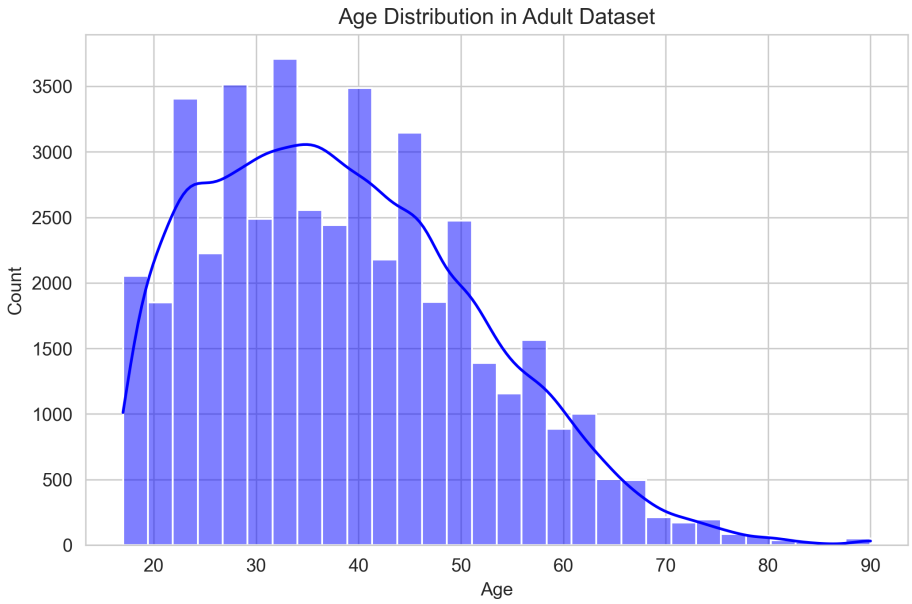

Figure 4.6: Age distribution of individuals in the Adult dataset.

Further analysis reveals disparities in medical utilization rates between White and Non-white individuals across poverty categories (Figure 4.16). Within each

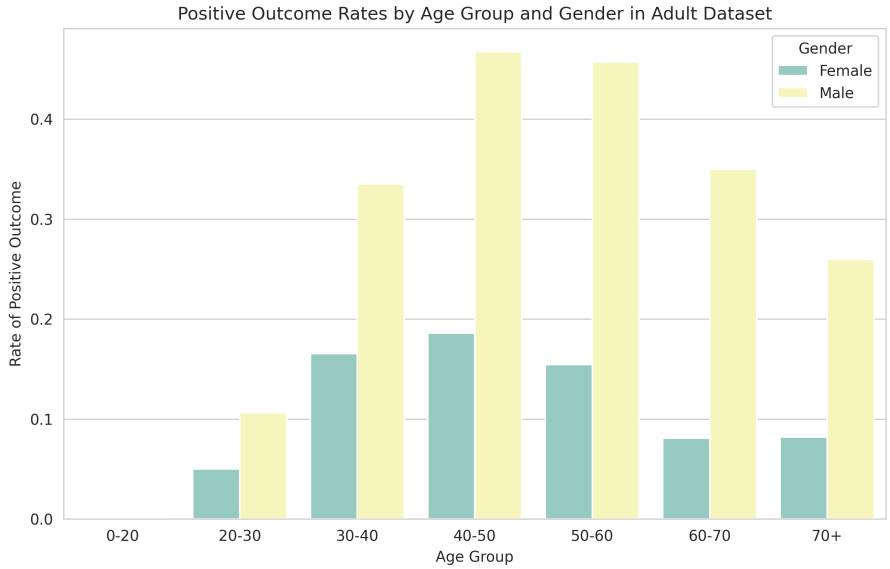

Figure 4.7: Positive outcome rates by age group and gender in Adult dataset.

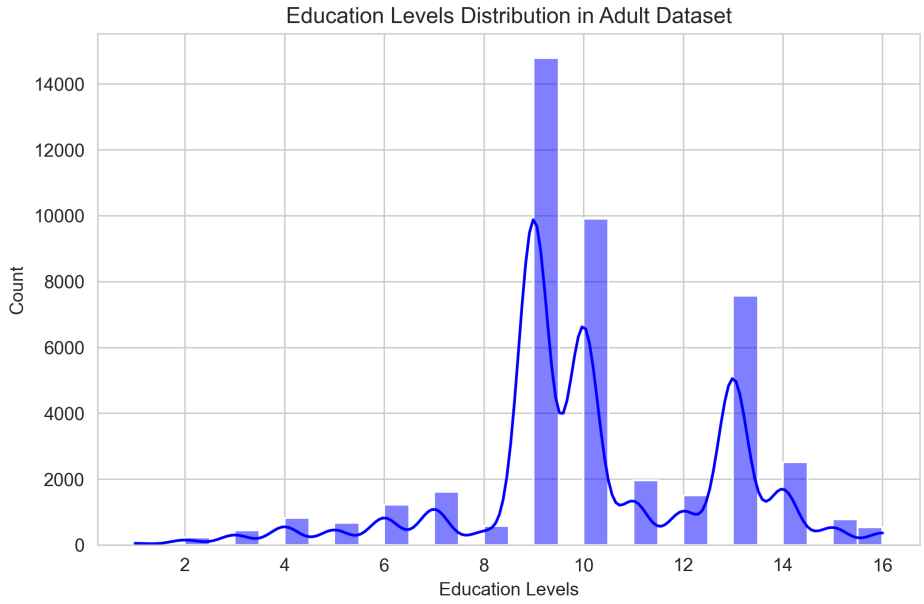

Figure 4.8: Distribution of years of education in Adult dataset.

category, White individuals consistently exhibit higher utilization rates than Non-white individuals. For instance, in the poorest category, utilization rates are 25.6% for White individuals compared to 12.7% for Non-white individuals. This pattern persists across all categories, with the largest disparity observed in Category 5.

Similarly, disparities are evident across insurance coverage types (Figure 4.17). White individuals utilize medical services at higher rates than Non-white individuals, regardless of insurance status. The largest disparity occurs among those with public insurance, with utilization rates of 33.5% for White individuals compared to 16.3%

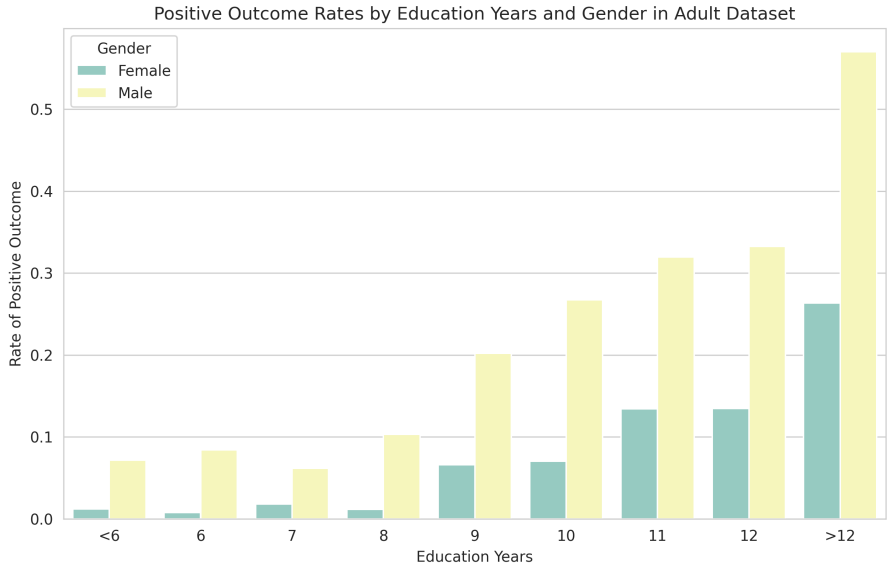

Figure 4.9: Positive outcome rates by education level and gender in Adult dataset.

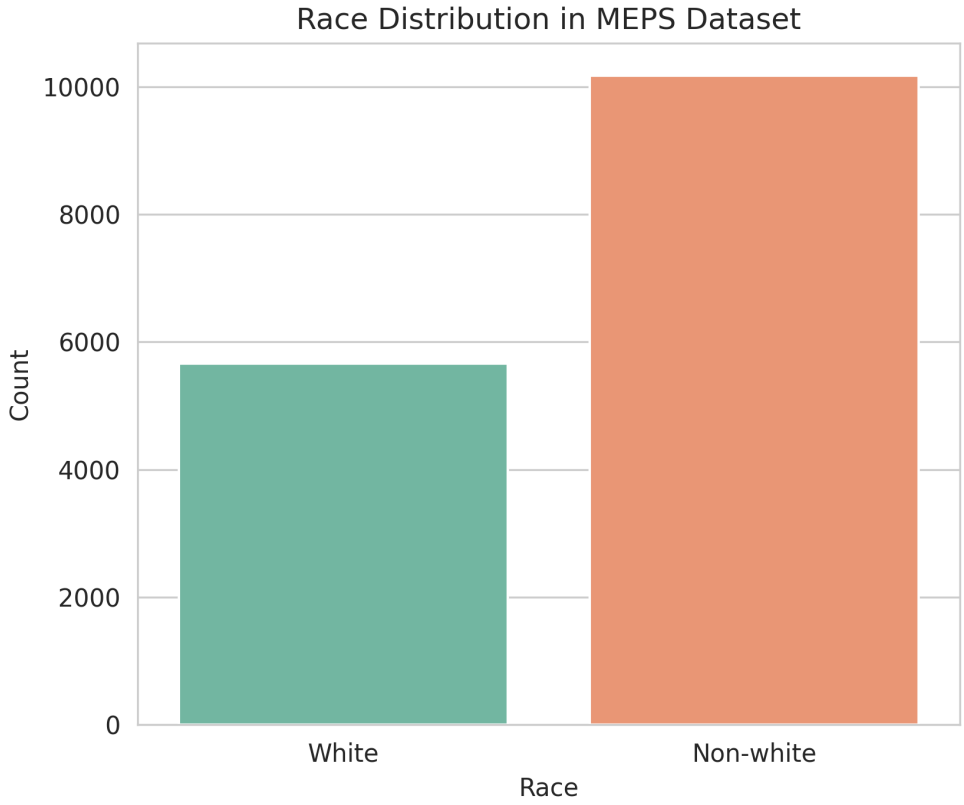

Figure 4.10: Race distribution in MEPS dataset.

for Non-white individuals. These findings highlight systemic inequities that persist even when income and insurance coverage are accounted for.

The analysis of the Adult and Medical Expenditure datasets reveals significant dis-

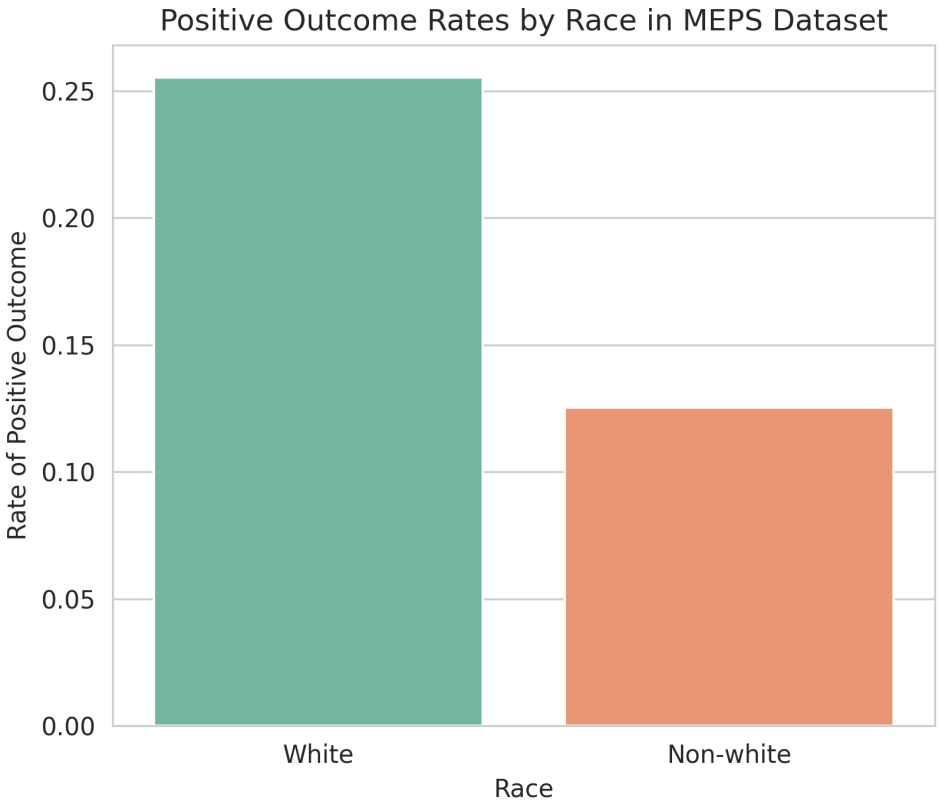

Figure 4.11: Positive outcome rates by race in MEPS dataset.

parities across demographic groups, such as gender, race, age, education, income, and insurance coverage. These disparities underscore the importance of addressing biases and systemic inequities when designing predictive models. A deeper understanding of these patterns can inform fair and inclusive decision-making frameworks.

4.2 Preprocessing of Datasets

4.2.1 Adult Dataset Preprocessing

The `AdultDataset` from AIF360 was used, and several preprocessing steps were performed using the `load_preproc_data_adult` function from AIF360 to prepare the data for bias detection and mitigation while preserving the sensitive attributes.

Figure 4.18 illustrates the preprocessed Adult dataset, showcasing transformations applied to enable effective bias mitigation.

The preprocessing steps included the following: Age attributes were grouped into decades, resulting in the following categories: `Age (decade)=10`, `Age (decade)=20`, `Age (decade)=30`, `Age (decade)=40`, `Age (decade)=50`, `Age (decade)=60`, and `Age (decade)=>=70`. The education attribute (`education-num`) was transformed into categories, including `Education Years=<6`, `Education Years=6`, `Education Years=7`, `Education Years=8`, `Education Years=9`, `Education Years=10`, `Education`

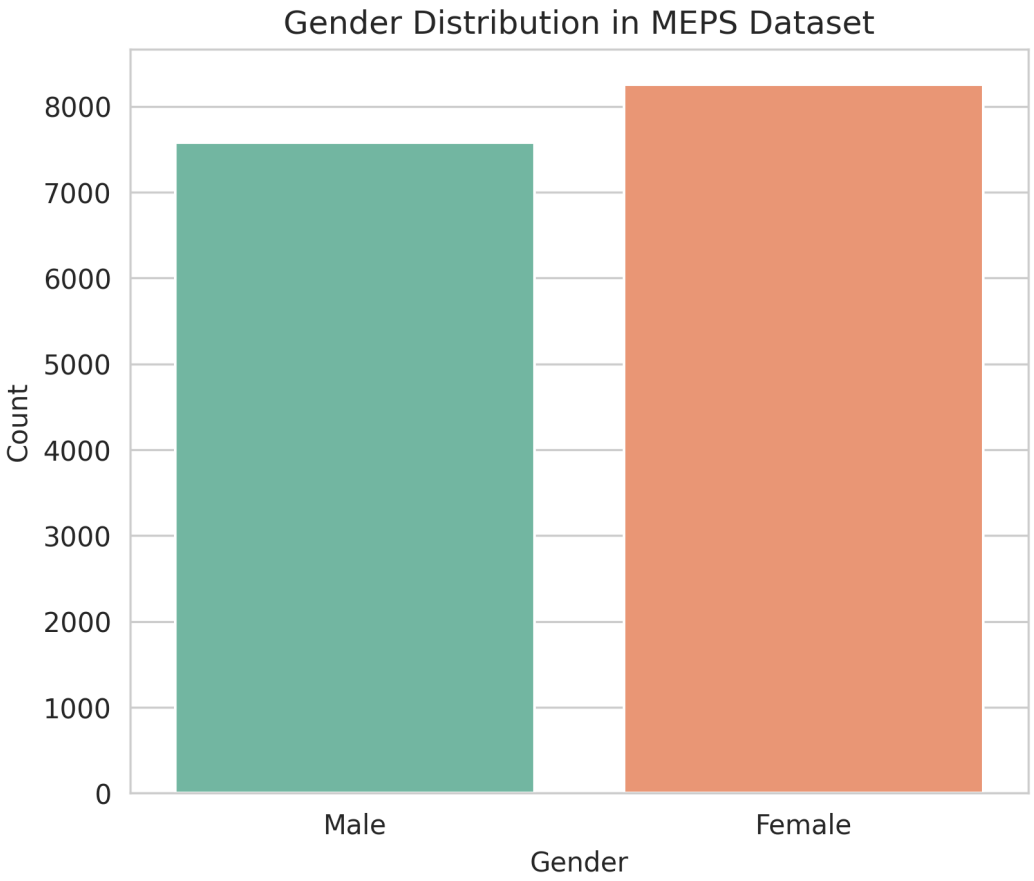

Figure 4.12: Gender distribution in MEPS dataset.

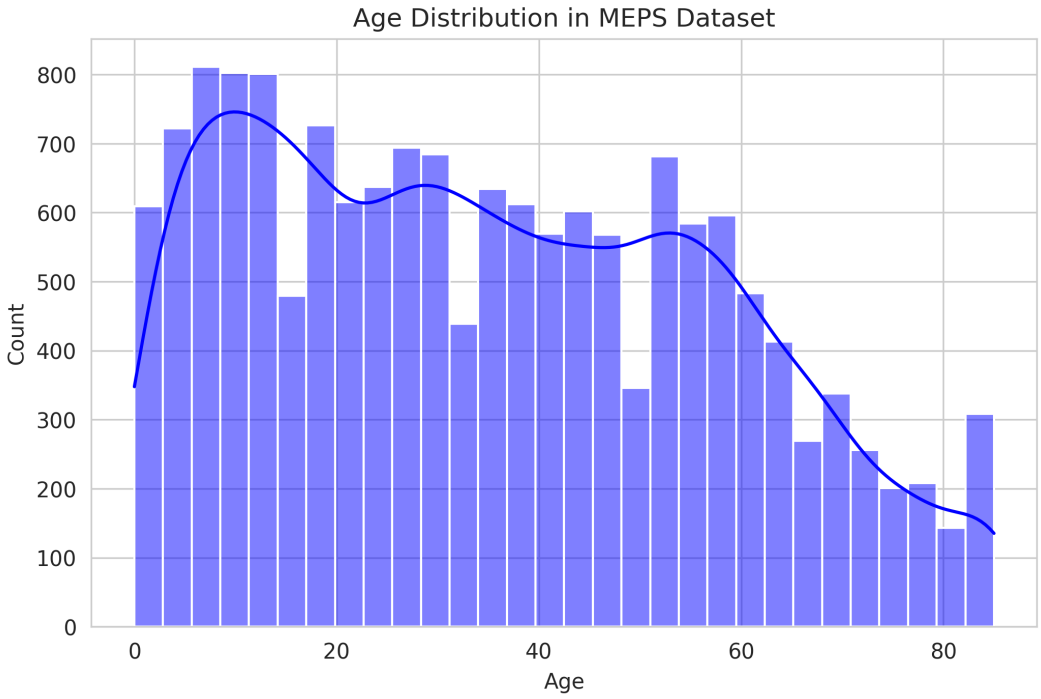

Figure 4.13: Age distribution in MEPS dataset.

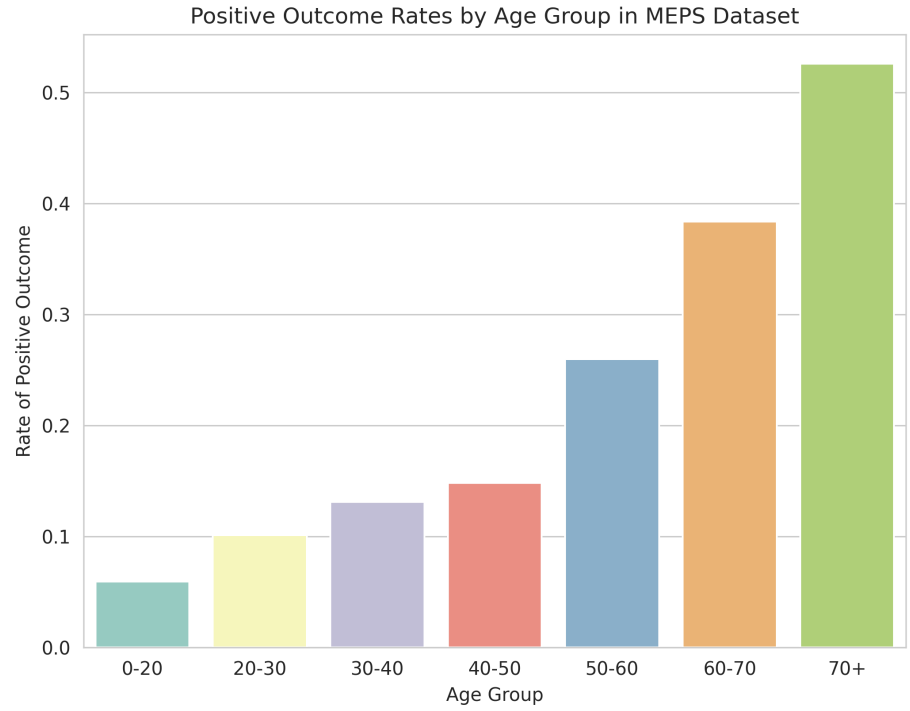

Figure 4.14: Positive outcome rates across age groups in MEPS dataset.

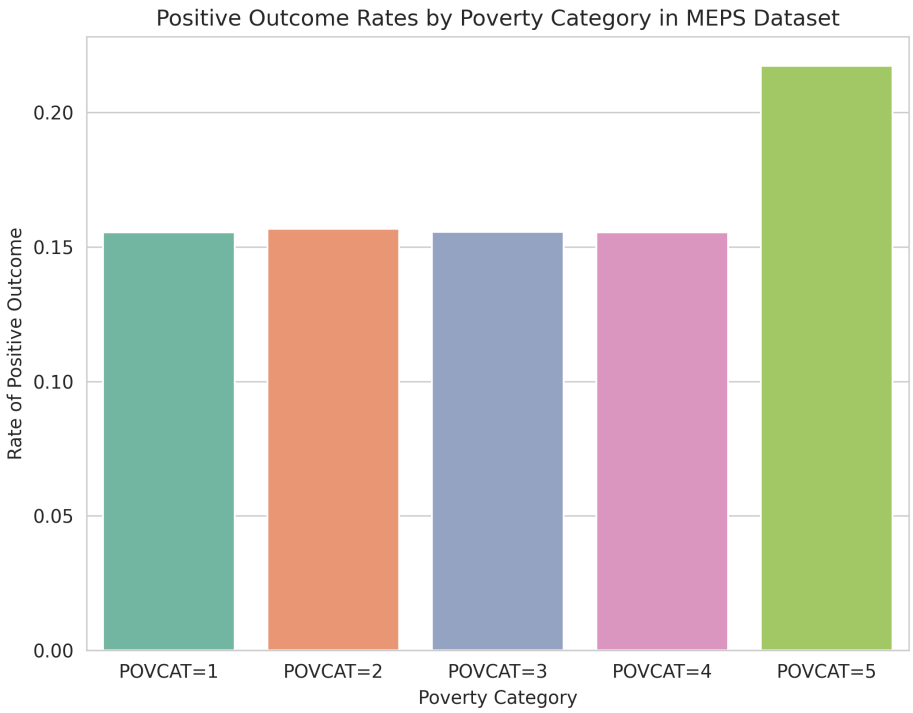

Figure 4.15: Positive outcome rates by poverty category in MEPS dataset.

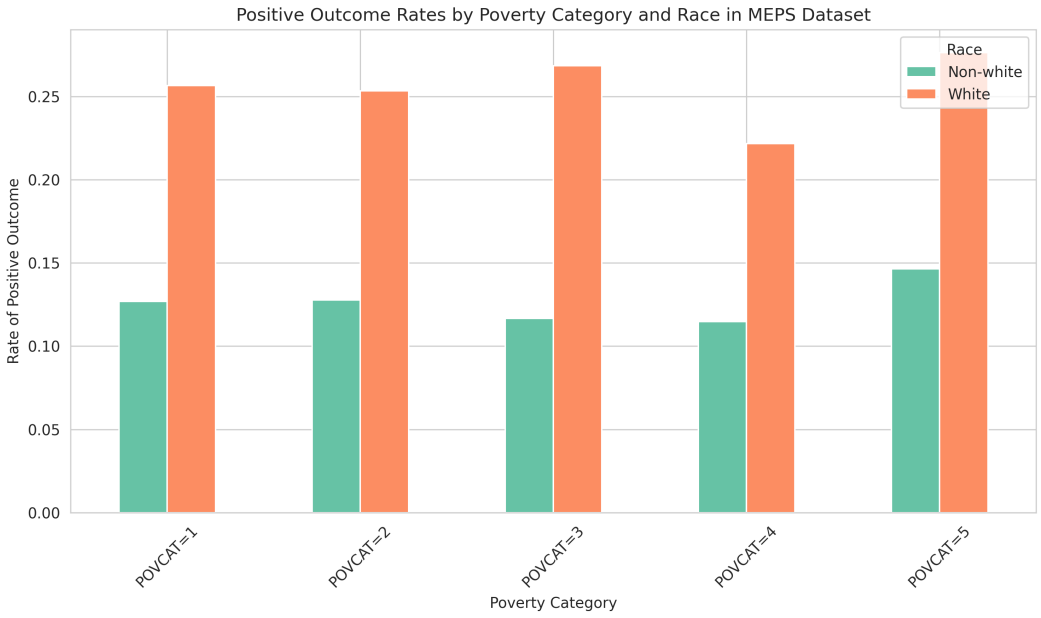

Figure 4.16: Positive outcome rates by race and poverty category in MEPS dataset.

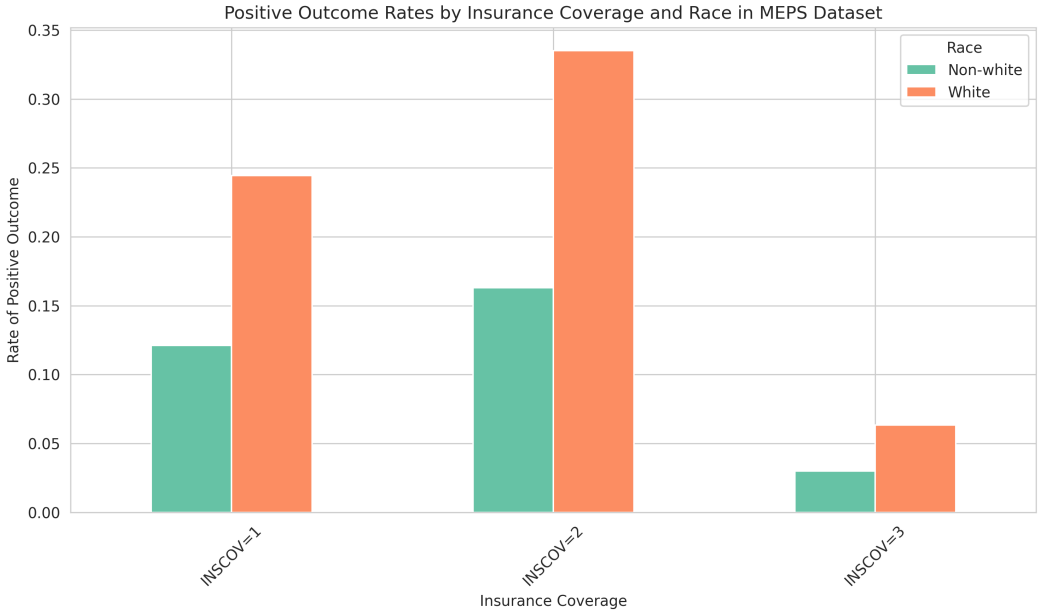

Figure 4.17: Positive outcome rates by insurance coverage and race in MEPS dataset.

`Years=11`, `Education Years=12`, and `Education Years=>12`. Income classification was transformed into binary values (`>50K` and `<=50K`) to simplify the target variable. Sensitive attributes, such as gender and race, were recoded numerically (`{0: Female, 1: Male}` for gender and `{0: Non-white, 1: White}` for race).

Feature selection retained both numerical and categorical features (`race`, `sex`, `Age (decade)`, and `Education Years`) for most bias mitigation methods. However, the DIR method required a subset of purely numerical attributes (`age`, `education-num`, `capital-gain`, `capital-loss`, and `hours-per-week`) due to its incompatibility with

	race	sex	Age (decade)=10	Age (decade)=20	Age (decade)=30	Age (decade)=40	Age (decade)=50	Age (decade)=60	Age (decade)=>=70	Education Years=6
0	0.0	1.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0
1	1.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0
2	1.0	1.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0
3	0.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0
4	1.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...
48837	1.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0
48838	1.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0
48839	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0
48840	1.0	1.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0
48841	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0

Figure 4.18: Preprocessed Adult dataset

categorical data.

4.2.2 Medical Expenditure Dataset Preprocessing

The Medical Expenditure dataset required preprocessing to address data quality issues and prepare it for analysis using AIF360 tools. Figure 4.19 depicts the preprocessed dataset, highlighting adjustments made to the features.

	RACE	SEX	PCS42	MCS42	Age (decade)_0	Age (decade)_10	Age (decade)_20	Age (decade)_30	Age (decade)_40	Age (decade)_50	...	Age (decade)_70
0	1.0	1.0	25.930	58.47	0.0	0.0	0.0	0.0	0.0	1.0	...	0.0
1	1.0	0.0	20.420	26.57	0.0	0.0	0.0	0.0	0.0	1.0	...	0.0
3	1.0	0.0	53.120	50.33	0.0	0.0	1.0	0.0	0.0	0.0	...	0.0
4	1.0	1.0	53.435	54.37	1.0	0.0	0.0	0.0	0.0	0.0	...	0.0
5	0.0	1.0	53.435	54.37	0.0	0.0	1.0	0.0	0.0	0.0	...	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...
16573	0.0	1.0	56.710	62.39	0.0	0.0	1.0	0.0	0.0	0.0	...	0.0
16574	0.0	0.0	56.710	62.39	0.0	0.0	1.0	0.0	0.0	0.0	...	0.0
16575	1.0	0.0	53.435	54.37	1.0	0.0	0.0	0.0	0.0	0.0	...	0.0
16576	0.0	0.0	43.970	42.45	0.0	0.0	0.0	0.0	0.0	1.0	...	0.0
16577	0.0	0.0	42.680	43.46	0.0	0.0	0.0	0.0	0.0	0.0	...	1.0

Figure 4.19: Preprocessed MEPS dataset

Key preprocessing steps included cleaning the columns (`PCS42` and `MCS42`) by replacing invalid values with NaN and imputing missing values using the column median. Age attributes were grouped into decades, creating the feature `Age (decade)`. The `RACE` attribute was recoded as {0: Non-white, 1: White}.

Feature selection was tailored to ensure compatibility with the requirements of each bias mitigation method. Both categorical and numerical features (`RACE`, `SEX`, `PCS42`, `MCS42`, `POVCAT`, `INSCOV`, and `Age (decade)`) were initially retained. However, categorical features such as `Age (decade)` were excluded from the dataset for compatibility with the `DIR` and `OptimPreproc` methods, which require numerical

attributes. As a result, only numerical features (`RACE`, `SEX`, `PCS42`, `MCS42`, and `UTILIZATION`) were retained for these methods. Additionally, for the `OptimPreproc` method, numerical features `PCS42` and `MCS42` were discretized into bins to enhance computational efficiency.

4.3 Initial Bias Assessment

Before applying any bias mitigation techniques, an initial bias assessment was conducted on both datasets using fairness metrics. These metrics were calculated to evaluate the disparity between privileged and unprivileged groups with respect to the target variable.

The fairness metrics were computed using the `BinaryLabelDatasetMetric` class from AIF360 [1], which provides the following metrics:

- **Statistical Parity Difference**: Measures the difference in favorable outcomes between the unprivileged and privileged groups. A negative value indicates a bias favoring the privileged group, while a positive value indicates bias favoring the unprivileged group. A value of 0 signifies perfect balance.
- **Disparate Impact**: Computes the ratio of favorable outcomes between the unprivileged and privileged groups. A ratio of 1 represents perfect fairness, while values significantly below or above 1 indicate unequal representation.

These initial metrics establish a baseline for evaluating the effectiveness of bias mitigation techniques applied in subsequent stages of analysis.

4.4 Bias Mitigation

After assessing the initial bias, the next step was to apply bias mitigation techniques to the datasets. Two main approaches were used for bias mitigation: AIF360-based techniques and synthetic data augmentation.

4.4.1 Bias Mitigation Using AIF360

This section outlines the practical implementation of four bias mitigation methods: Reweighing, Disparate Impact Remover, Learning Fair Representations (LFR), and Optimized Preprocessing (OptimPreproc). These methods were implemented using the AIF360 [1] framework. The parameters for LFR and OptimPreproc were initially adopted from the library defaults as implemented in the AIF360 GitHub repository and described in the related research papers. Adjustments were made only when necessary to ensure that the methods functioned correctly for the datasets used in this study.

4.4.1.1 Reweighing

The `Reweighing` algorithm was implemented to adjust the instance weights based on the privileged and unprivileged groups. The groups were defined as follows:

- For the `Adult` dataset: `{sex=0: Female, sex=1: Male}`
- For the `MEPS` dataset: `{race=0: Non-white, race=1: White}`

Weights were assigned to ensure equal representation of groups in subsequent analyses, and the transformed datasets were generated by fitting and applying the `Reweighing` object to the train and test sets.

4.4.1.2 Disparate Impact Remover

For the `Disparate Impact Remover` algorithm, the `repair_level` parameter was tested over 10 evenly spaced values between 0 and 1 to determine the optimal trade-off between fairness and performance. The sensitive attribute was set as `sex` for the `Adult` dataset and `race` for the `MEPS` dataset. This approach modified the datasets to reduce bias while preserving rank order within each group.

4.4.1.3 Learning Fair Representations (LFR)

The `LFR` algorithm was applied with the following parameters:

- `k=10`: Number of prototype points in the latent representation.
- `Ax=0.1`: Weight for reconstruction loss.
- `Ay=1`: Weight for prediction loss.
- `Az=1.5`: Weight for fairness loss.
- `seed=42`: Ensures reproducibility.
- `verbose=1`: Provides detailed output.

For the `Adult` dataset, the sensitive attribute was defined as `sex`, with privileged groups assigned as `{sex=1: Male}` and unprivileged groups as `{sex=0: Female}`. For the `MEPS` dataset, the sensitive attribute was `race`, with privileged groups defined as `{race=1: White}` and unprivileged groups as `{race=0: Non-white}`. The transformed datasets were generated to effectively balance fairness, prediction accuracy, and reconstruction quality, enabling bias mitigation without significant loss of utility.

4.4.1.4 Optimized Preprocessing

The `OptimPreproc` algorithm was implemented with dataset-specific parameters and distortion functions. Due to computational constraints stemming from the larger size of the `MEPS` dataset, fewer parameter combinations were explored compared to the `Adult` dataset. The parameters were as follows:

- **Adult Dataset**:
  - Distortion function: `get_distortion_adult`, penalizing significant changes in education years (greater than 1 year), large age jumps (greater than one decade), and income reductions.
  - Parameters: `epsilon=0.05`, `clist=[0.99, 1.99, 2.99]`, `dlist=[0.1, 0.05, 0]`.
- **MEPS Dataset**:
  - Distortion function: `get_distortion_medical`, penalizing large changes in `PCS42` and `MCS42` bins and undesirable reductions in `UTILIZATION`.
  - Parameters: `epsilon=0.01`, `clist=[0.99]`, `dlist=[0.1]`.

The algorithm leveraged these distortion functions to minimize group disparities while retaining the integrity of the original data.

After applying each bias mitigation technique, fairness metrics were recalculated on the transformed datasets to assess the effectiveness of the mitigation. These recalculated metrics served as a benchmark for evaluating and comparing the performance of the different techniques.

4.4.2 Bias Mitigation Using Synthetic Data Augmentation

To address bias and enhance the fairness of the machine learning model, synthetic data generation was employed using **Clearbox AI's Tabular Engine** [2]. This approach generated synthetic data for underrepresented groups, ensuring better representation within the dataset and improving fairness metrics such as Disparate Impact and Statistical Parity Difference.

4.4.2.1 Identifying Underrepresented Groups

The underrepresented groups in this research were identified based on their sensitive attributes (sex and race) combined with a positive outcome in the dataset. These groups, highlighted in the **Data Overview and Insights** 4.1, were selected to address disparities in representation and ensure fairness through synthetic augmentation. Each group reflects a population underrepresented relative to others in the dataset. Intersectional groups, combining multiple attributes, were also considered to address more nuanced forms of underrepresentation.

Below is an overview of the identified underrepresented groups:

- **Groups from the Adult Dataset**:
  - **Women_50K**: Women earning more than $50K, representing gender-based underrepresentation in higher income brackets.
  - **Women_Non-white_50K**: Non-white women earning more than $50K, addressing intersectional disparities in income outcomes.

- **Women_above60_50K**: Women over 60 years old earning more than $50K, highlighting age-based underrepresentation intersecting with gender in higher income brackets.

- **Groups from the MEPS Dataset**:
  - **Non-white_UTILIZATION10**: Non-white individuals with high healthcare utilization (UTILIZATION $\geq$ 10), addressing racial disparities in healthcare access and usage.
  - **Non-white_InsuranceCov2_UTILIZATION10**: Non-white individuals with public insurance coverage and high healthcare utilization, tackling intersectional disparities related to insurance and healthcare usage.
  - **Non-white_PovertyCat1_UTILIZATION10**: Non-white individuals in poverty category 1 with high healthcare utilization, combining poverty and race attributes to address disparities.

These groups represent populations that are disproportionately underrepresented and were selected to ensure fairer representation across sensitive attributes. Synthetic data was generated to provide a balanced dataset before proceeding to the classifier training phase.

4.4.2.2 Synthetic Data Generation Process

The synthetic data generation process involved filtering the dataset to isolate underrepresented groups, followed by using the **Clearbox AI Tabular Engine** [2] to generate synthetic samples. The engine was trained on the filtered data to learn group-specific features and patterns, ensuring consistency with the original dataset. The synthesized data points were then combined with the original dataset, creating an augmented dataset for training and testing machine learning models. This augmentation improved the dataset's balance and representation of diverse groups.

4.4.2.3 Synthetic datasets

The number of synthetic datasets generated varied depending on the size of the underrepresented group and its impact on fairness metrics. Smaller underrepresented groups required a larger number of synthetic datasets to reach fair metrics, as their representation in the original dataset was disproportionately low. These synthetic datasets were iteratively generated, increasing the dataset count until Disparate Impact approached 1 and Statistical Parity Difference neared 0 for the transformed training dataset before the classifier training phase.

The final number of **synthetic datasets** generated for each group is outlined below:

- **Adult Dataset:**

  - **Women_50K**: 2.65 datasets.
  - **Women_Non-white_50K**: 22.5 datasets.
  - **Women_above60_50K**: 50.5 datasets.

- **Medical Expenditure Dataset:**
  - **Non-white_UTILIZATION10**: 1.5 datasets.
  - **Non-white_InsuranceCov2_UTILIZATION10**: 2.5 datasets.
  - **Non-white_PovertyCat1_UTILIZATION10**: 5.0 datasets.

By targeting these underrepresented groups, the synthetic data augmentation process aims to improve fairness in the dataset while maintaining the predictive capability of the machine learning models.

4.5 Fairness Evaluation Before and After Transformation

After applying the bias mitigation techniques, it was essential to assess their effectiveness in reducing bias.

To summarize the process, the following steps were taken for each dataset and bias mitigation technique:

1. **Fairness Metric Calculation (Original Data)**: Fairness metrics were first calculated on both the training and test sets of the original (untransformed) data to evaluate the initial bias.

2. **Transformation Application**: The bias mitigation technique was performed on both the training set and the test set.

3. **Fairness Metric Calculation (Transformed Data)**: After applying the bias mitigation technique, fairness metrics were recalculated on both the transformed training and test set to evaluate the impact of the transformation on bias reduction.

4.6 Classifier Training and Evaluation

As discussed in previous sections, once the bias mitigation techniques were applied, classifier training and evaluation were performed on both the original and transformed datasets for each bias mitigation technique and dataset.

The classifiers used in this study were:

- **Logistic Regression**: Implemented with `class_weight='balanced'` to address imbalances in the datasets, `solver='liblinear'` for efficiency with smaller datasets, and `random_state=1` to ensure reproducibility.

- **Random Forest**: Implemented using default hyperparameters and `random_state=1` for reproducibility, with no additional optimization to maintain comparability across datasets and mitigation methods.
- **Gradient Boosting**: Implemented with default parameters, including `random_state=1` for reproducibility, to ensure consistent evaluations without favoring specific methods or datasets.

4.6.1 Classifier Training

The following steps were taken for classifier training:

1. **Training on Original Dataset**: Each classifier was trained on the original training dataset (before any bias mitigation). This provided a baseline for comparison across fairness metrics and classification performance.
2. **Threshold Selection**: The validation set was used to select the optimal classification threshold based on balanced accuracy.
3. **Training on Transformed Dataset**: Each classifier was trained on the transformed training dataset (after bias mitigation). The same threshold, selected based on the untransformed validation set, was applied to the transformed test set for evaluation. However, for Learning Fair Representations (LFR), fairness-aware scores are generated as part of its optimization process. These scores are designed to balance reconstruction accuracy, prediction accuracy, and fairness. Since these scores inherently account for fairness, we only used the decision thresholds predicted by the classifiers, eliminating the need to train separate classifiers on the transformed dataset.

To maintain simplicity and enable consistent comparisons across datasets and bias mitigation techniques, hyperparameter optimization was deliberately omitted. Optimizing hyperparameters for classifiers such as Random Forest and Gradient Boosting could improve results but would require customizing the process for each method, given their differences. Instead, the default hyperparameters provided by the respective libraries were utilized. This approach ensures uniformity in evaluation and facilitates direct comparison of classification performance and fairness metrics across all methods and datasets.

4.6.2 Evaluation

After training the classifiers on both the original and transformed datasets, the next step was to evaluate their performance on both datasets. This evaluation considered two critical aspects: fairness, which measures how the model treats different groups, and performance, which measures the overall predictive effectiveness of the model.

As described in Section 3.2, fairness metrics were used to evaluate disparities between privileged and unprivileged groups. These metrics included Statistical

Parity Difference, Disparate Impact, Average Odds Difference, Equal Opportunity Difference, and Theil Index. For each classifier, these metrics were calculated on both the original and transformed test sets to assess whether the bias mitigation techniques successfully reduced disparities.

To evaluate model performance, Balanced Accuracy was used as the primary metric, as outlined in Section 3.5. Balanced Accuracy was calculated for both the original and transformed test sets to determine whether the bias mitigation strategies impacted the model's predictive performance. This metric ensured a fair comparison by accounting for class imbalances.

The ultimate goal of the evaluation was to determine the effectiveness of bias mitigation techniques in reducing disparities between privileged and unprivileged groups while maintaining acceptable levels of model performance. This trade-off analysis provided valuable insights into the practicality and limitations of each mitigation approach.

4.6.3 Plotting Fairness Metrics

To visualize the impact of bias mitigation techniques, fairness metrics were plotted against classification thresholds for both the original and transformed test datasets. For this analysis, two key metrics were considered: **Disparate Impact (DI)** and **Average Odds Difference (AOD)**. These plots enable an intuitive assessment of how fairness evolves as the classification threshold changes, and they aid in understanding the trade-offs between fairness and performance.

Original Test Data

The first two plots illustrate the fairness metrics for the original test data from the `Adult` dataset using Logistic Regression and the Reweighing method:

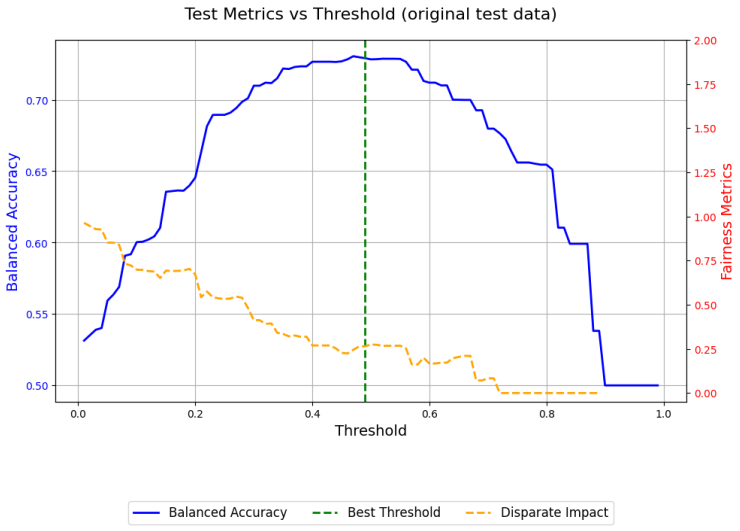

Figure 4.20: Disparate Impact vs. Threshold for the Original Test Data of the `Adult` Dataset using Logistic Regression and Reweighing.

From the first plot (Figure 4.20), it is evident that Disparate Impact decreases as the threshold increases, moving farther away from the ideal value of 1. This

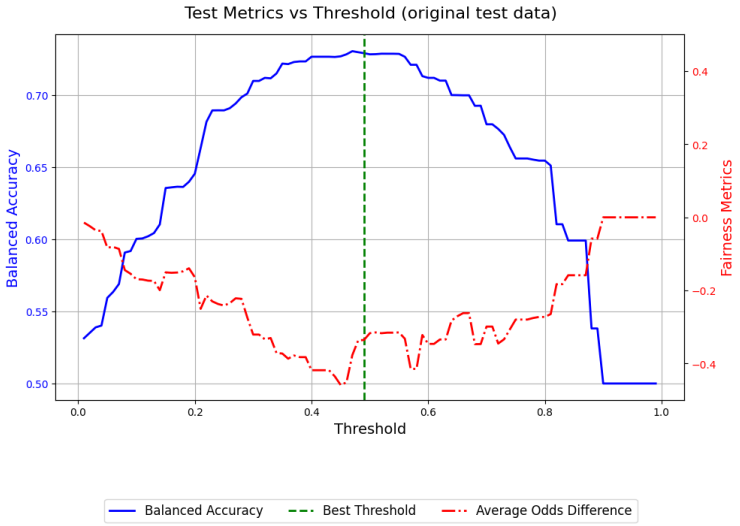

Figure 4.21: Average Odds Difference vs. Threshold for the Original Test Data of the `Adult` Dataset using Logistic Regression and Reweighing.

demonstrates that fairness, as measured by Disparate Impact, diminishes at higher thresholds, highlighting a trade-off between fairness and performance as the classification threshold changes. From the second plot (Figure 4.21), it is evident that Average Odds Difference (AOD) increases in magnitude as the threshold approaches the **Best Threshold**, signifying greater unfairness that favors the privileged group. Beyond the **Best Threshold**, however, AOD begins to decrease toward zero, indicating an improvement in fairness as the threshold continues to increase. Balanced Accuracy peaks at this threshold, reflecting optimal predictive performance despite the fairness challenges observed around the **Best Threshold**.

Transformed Test Data

The next two plots depict the same metrics but focus on the transformed test data after applying the Reweighing bias mitigation method:

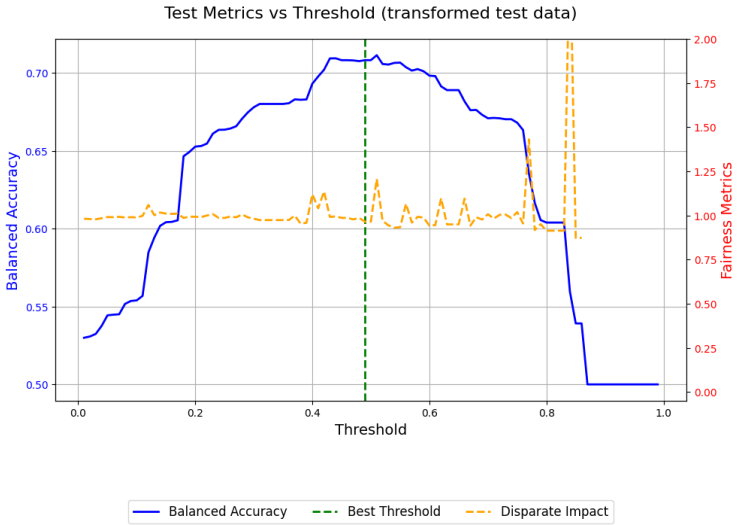

Figure 4.22: Disparate Impact vs. Threshold for the Transformed Test Data of the `Adult` Dataset using Logistic Regression and Reweighing.

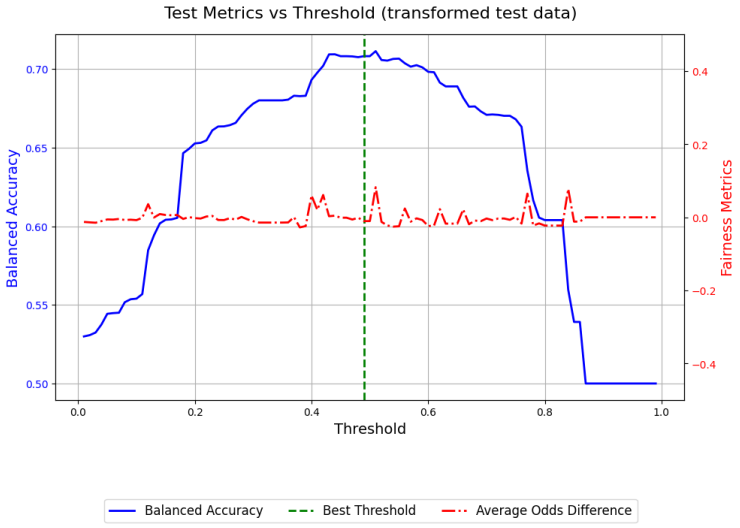

Figure 4.23: Average Odds Difference vs. Threshold for the Transformed Test Data of the `Adult` Dataset using Logistic Regression and Reweighing.

As observed in the Disparate Impact plot (Figure 4.22), mitigation efforts significantly reduce disparities, with Disparate Impact stabilizing near the ideal value of 1. Similarly, the Average Odds Difference plot (Figure 4.23) highlights improved fairness, as AOD approaches 0 at the **Best Threshold**. Balanced Accuracy remains stable across thresholds, indicating that the bias mitigation method does not compromise predictive performance, even as fairness improves.

These visualizations illustrate the trade-offs between fairness and performance achieved through bias mitigation techniques. By comparing the original and transformed data, the effectiveness of each method in reducing disparities while maintaining predictive performance is made evident. For a detailed view of fairness metrics across other combinations of datasets, classifiers, and mitigation methods—analyzing both original and transformed test data—please refer to the project's GitHub repository. These plots provide comprehensive insights into the effects of bias mitigation across configurations.

Chapter 5

Results

This chapter presents the results from evaluating different bias mitigation strategies applied to the datasets, which were then used to train the classifiers in the experiments. The classifiers used in this study include **Logistic Regression**, **Random Forest**, and **Gradient Boosting**, while the datasets are the Adult Dataset and the Medical Expenditure Dataset.

Each dataset was transformed using multiple bias mitigation methods: **Reweighing**, **Disparate Impact Remover (DIR)**, **Labeled Fair Representation (LFR)**, **OptimPreproc**, and **Synthetic Data Generation**, before being used to train the classifiers.

Each table summarizes performance and fairness metrics, specifically focusing on **Balanced Accuracy**, **Statistical Parity Difference**, and **Disparate Impact**. Performance metrics assess the classifiers' overall effectiveness, while fairness metrics evaluate the impact of bias mitigation methods on fairness across different groups.

All fairness metrics, including Average Odds Difference, Equal Opportunity Difference, Theil Index, Statistical Parity Difference, and Disparate Impact, were calculated in the analysis. However, only **Statistical Parity Difference** and **Disparate Impact** are presented in the results. These two metrics were selected for their clarity and ease of interpretation: Statistical Parity Difference measures the difference between the probability of a positive outcome of privileged and unprivileged groups, while Disparate Impact measures the ratio of favorable outcomes between these groups. Although other metrics were calculated, they involve more complex interpretations and may not convey the disparities as directly or clearly. For a comprehensive view of the analysis, including the results of all calculated metrics, readers can refer to the respective code uploaded to the GitHub repository.

The results for each classifier are presented in two forms: absolute metrics and differences.

Tables 5.1 and 5.3 display the absolute performance and fairness metrics for the Adult and Medical Expenditure datasets, respectively. The results for each classifier are presented individually, starting with baseline performance (trained on the original untransformed data). The subsequent rows detail performance and fairness metrics following the application of each bias mitigation method to transform the datasets.

Tables 5.2 and 5.4, on the other hand, illustrate the differences in performance and fairness metrics across bias mitigation methods for each classifier. These tables focus on relative changes, highlighting how each method impacts **Balanced Accuracy** and **Disparate Impact (DI)** compared to their respective baselines. Together, the absolute metrics and relative differences provide a comprehensive perspective on the effectiveness of the bias mitigation methods.

The following sections further analyze and interpret these findings to provide meaningful insights into the impact of bias mitigation strategies.

5.1 Adult

Bias Mitigation Method	Logistic Regression			Random Forest			Gradient Boosting		
	Balanced Acc.	SPD	DI	Balanced Acc.	SPD	DI	Balanced Acc.	SPD	DI
Baseline (Original Data for DIR)	0.74	-0.41	0.24	0.75	-0.30	0.26	0.77	-0.36	0.17
DIR	0.71	-0.04	0.89	0.73	-0.13	0.61	0.74	-0.12	0.58
Baseline (Original Data)	0.72	-0.39	0.26	0.72	-0.39	0.26	0.72	-0.39	0.27
Reweighing	0.70	-0.01	0.96	0.69	0.06	1.14	0.69	0.06	1.13
LFR	0.60	-0.04	0.58	0.68	-0.07	0.86	0.68	-0.07	0.86
OptimPreproc	0.70	-0.08	0.80	0.71	-0.20	0.60	0.71	-0.21	0.57
Synthetic Data (Women_50K)	0.72	0.00	1.01	0.66	-0.02	0.96	0.66	-0.39	0.38
Synthetic Data (Women_Non-white_50K)	0.74	0.00	1.01	0.74	-0.16	0.70	0.71	-0.20	0.61
Synthetic Data (Women_above60_50K)	0.78	-0.09	0.79	0.76	-0.15	0.70	0.67	-0.31	0.42

Table 5.1: Performance and Fairness Metrics for Bias Mitigation Methods (**Adult Dataset**)

Methods	LR AccDiff	RF AccDiff	GB AccDiff	LR SPD Diff	RF SPD Diff	GB SPD Diff	LR DI Diff	RF DI Diff.	GB DI Diff.
DIR	-0.03	-0.02	-0.03	0.37	0.17	0.24	0.65	0.35	0.41
Reweighing	-0.02	-0.03	-0.03	0.38	0.45	0.45	0.70	0.88	0.86
LFR	-0.12	-0.04	-0.04	0.35	0.32	0.32	0.32	0.60	0.59
OptimPreproc	-0.02	-0.01	-0.01	0.31	0.19	0.18	0.54	0.34	0.30
Synthetic Data (Women_50K)	0.00	-0.06	-0.06	0.39	0.37	0.00	0.75	0.70	0.11
Synthetic Data (Women_Non-white_50K)	0.02	0.02	-0.01	0.39	0.23	0.19	0.75	0.44	0.34
Synthetic Data (Women_above60_50K)	0.06	0.04	-0.05	0.30	0.24	0.08	0.53	0.44	0.15

Table 5.2: Results for accuracy and fairness differences across methods for the **Adult dataset**.

5.2 Medical Expenditure

Bias Mitigation Method	Logistic Regression			Random Forest			Gradient Boosting		
	Balanced Acc.	SPD	DI	Balanced Acc.	SPD	DI	Balanced Acc.	SPD	DI
Baseline (Original Data for DIR)	0.71	-0.43	0.27	0.71	-0.43	0.27	0.71	-0.29	0.42
DIR	0.70	-0.08	0.72	0.65	-0.03	0.90	0.70	-0.1	0.72
Baseline (Original Data for OptimPreproc)	0.68	-0.51	0.26	0.71	-0.29	0.39	0.71	-0.30	0.41
OptimPreproc	0.59	-0.17	0.72	0.68	-0.07	0.80	0.68	-0.13	0.70
Baseline (Original Data)	0.75	-0.31	0.40	0.72	-0.23	0.48	0.74	-0.27	0.44
Reweighing	0.74	-0.03	0.89	0.55	-0.00	0.99	0.53	-0.05	0.94
LFR	0.63	-0.02	0.81	0.68	-0.07	0.75	0.68	-0.09	0.72
Synthetic Data (Non-white_UTILIZATION10)	0.81	-0.11	0.70	0.78	-0.10	0.76	0.79	-0.16	0.68
Synthetic Data (Non-white_InsuranceCov2_UTILIZATION10)	0.80	-0.11	0.70	0.77	-0.12	0.71	0.79	-0.18	0.63
Synthetic Data (Non-white_PovertyCat1_UTILIZATION10)	0.80	-0.08	0.76	0.77	-0.11	0.73	0.78	-0.19	0.63

Table 5.3: Performance and Fairness Metrics for Bias Mitigation Methods (**Medical Expenditure Dataset**)

Methods	LR AccDiff	RF AccDiff	GB AccDiff	LR SPD Diff	RF SPD Diff	GB SPD Diff	LR DI Diff	RF DI Diff	GB DI Diff
DIR	-0.01	-0.06	-0.01	0.35	0.40	0.19	0.45	0.63	0.30
OptimPreproc	-0.09	-0.03	-0.03	0.34	0.22	0.17	0.46	0.41	0.29
Reweighing	-0.01	-0.17	-0.21	0.28	0.23	0.22	0.49	0.51	0.50
LFR	-0.12	-0.04	-0.06	0.29	0.16	0.18	0.41	0.27	0.28
SynData (Non-white_UTILIZATION10)	0.06	0.06	0.05	0.20	0.13	0.11	0.30	0.28	0.24
SynData (Non-white_InsuranceCov2_UTILIZATION10)	0.05	0.05	0.05	0.20	0.11	0.09	0.30	0.23	0.19
SynData (Non-white_PovertyCat1_UTILIZATION10)	0.05	0.05	0.04	0.23	0.12	0.08	0.36	0.25	0.19

Table 5.4: Results for accuracy and fairness differences across methods for the **Medical Expenditure dataset**.

Chapter 6

Discussion

In this chapter, we focus on discussing the results of various bias mitigation methods applied to the two datasets. These methods are evaluated based on the performance and fairness of the transformed datasets when used to train classifiers. Key metrics, including balanced accuracy, Statistical Parity Difference (SPD), and Disparate Impact (DI), are used to assess whether the transformed datasets maintain or improve fairness and predictive performance. The goal is to analyze how effectively each bias mitigation method achieves these objectives while retaining the utility of the datasets.

Before diving into the results, it's important to clarify the ideal values for the metrics used to evaluate classifier performance and fairness:

- **Balanced accuracy** ranges from 0 to 1, with the perfect value being 1;
- **Statistical Parity Difference** (SPD) has a range of -1 to 1, where 0 indicates perfect balance, meaning no disparity between groups;
- **Disparate Impact** (DI) can range from 0 to infinity, but a value of 1 represents perfect fairness, signifying equal treatment of different groups.

Additionally, DI values above 0.8 are considered acceptable, adhering to the *80% rule* commonly used in fairness evaluations. These benchmarks will guide the interpretation of the results presented below.

In fairness evaluations, the relationship between Statistical Parity Difference (SPD) and Disparate Impact (DI) depends on the probability of positive outcomes. When these probabilities are low, DI fluctuates significantly even if SPD remains stable, as DI is more sensitive to small absolute differences. Conversely, when probabilities are high, SPD and DI behave more similarly. For example, when Group A has a probability of 0.10 and Group B has 0.06, SPD = 0.10 - 0.06 = 0.04 (seems fair), but DI = 0.06 / 0.10 = 0.60 (very unfair).

SPD alone should not be relied upon for fairness, as it can be misleading, particularly with low probabilities. While SPD measures the absolute difference in positive outcome rates, DI captures the relative disparity between groups, which is crucial for fairness assessments. Therefore, DI should be prioritized over SPD, as it better reflects the extent of bias and disparity.

Moreover, it is essential to outline the key methodological decisions and constraints that influenced the analysis. These considerations ensure clarity in interpreting the outcomes of various bias mitigation methods applied to the datasets.

For the Adult dataset, two baselines were employed due to limitations of the Disparate Impact Remover (DIR) in handling categorical data. DIR applies linear transformations to adjust feature values and reduce their dependence on sensitive attributes (e.g., sex) while preserving the relative order of data within each group. While this approach is effective for continuous numerical features, it faces challenges with categorical features, necessitating their removal. Consequently, the resulting dataset was slightly different. This adjustment was necessary to evaluate DIR alongside other methods effectively. The repair levels used for DIR were carefully selected by testing ten evenly spaced intervals. The repair level yielding the optimal balance between accuracy and fairness was chosen to ensure robust results.

In the MEPS dataset, three baselines were utilized, including an additional baseline for OptimPreproc. To address computational complexity, categorical features were removed, and specific numerical features, such as PCS42 and MCS42, were discretized. High-dimensional data and numerous continuous features can make preprocessing techniques like OptimPreproc computationally expensive. Discretizing features reduces the number of distinct values, thereby simplifying mathematical operations and ensuring computational feasibility while enabling meaningful comparisons between methods.

On the other hand, the synthetic data generation process is inherently stochastic, meaning its effectiveness may vary across underrepresented groups. Despite this variability, we used the same augmented dataset across all classifiers to ensure consistency in comparisons.

For the sake of producing comparable results with consistent baselines, certain methodological constraints were applied. A fixed random seed (set to 42) was used to ensure consistent train-test splits across methods. However, using a specific random seed can result in variations in the distribution of data between training and testing sets. For instance, certain groups or sensitive attributes may become overrepresented or underrepresented in one set, leading to potential imbalances that can negatively impact the performance of some bias mitigation methods such as LFR, which rely on fair representation of groups within the data.

Additionally, hyperparameter optimization was deliberately omitted to maintain simplicity and comparability. While optimizing hyperparameters for classifiers like Random Forest and Gradient Boosting could yield improved results, doing so would favor certain bias mitigation methods over others. Instead, default hyperparameters were employed to ensure consistency in evaluation. Similarly, consistent hyperparameter settings were applied to bias mitigation methods like LFR and OptimPreproc, acknowledging that this may benefit some classifiers while penalizing others.

In the current section, we interpret the results using the raw performance and fairness values of the metrics for each bias mitigation method, as presented in Table 5.1 and in Table 5.3. This approach provides an immediate understanding of the absolute

outcomes for each method. However, since some methods are evaluated against different baselines (e.g., DIR operates on a modified dataset), it can be slightly challenging to directly compare their effectiveness across all metrics. To address this limitation, in the subsequent section, we will interpret the results through relative differences from their respective baselines. By analyzing the changes in performance and fairness metrics relative to each method's starting point, we can more clearly identify which methods demonstrate superior improvements or trade-offs. Therefore, for a straightforward analysis, readers may skip directly to Section 6.2.

6.1 Raw Performance and Fairness Interpretations

6.1.1 Adult

The results outlined in Table 5.1 illustrate the performance and fairness metrics for bias mitigation methods applied to the Adult dataset.

6.1.1.1 Logistic Regression (LR):

The **baseline for the original data used for DIR** shows a reasonably good accuracy (0.74), but there is considerable bias, as indicated by the negative SPD (-0.41), which suggests that the classifier is unfairly biased against certain groups. The DI of 0.24 also indicates a potential imbalance in the outcomes for different groups. **DIR** results in a slight drop in accuracy (0.71) but significantly reduces the SPD and improves DI, bringing fairness metrics closer to ideal. The DI value of 0.89 and SPD of -0.04 indicates a much more balanced outcome between different groups.

The **baseline of the original data** used for other methods (0.72) has slightly lower accuracy than that of the data used for DIR, but the fairness metrics (SPD=-0.39 and DI=0.26) are quite similar, showing that the dataset still has bias, though slightly less pronounced.

Reweighing also results in a small drop in accuracy (0.70) but achieves excellent fairness improvements. The SPD value of -0.01 shows almost no disparity, and the DI value of 0.96 is very close to ideal fairness, demonstrating that this method mitigates bias very effectively.

LFR produces the lowest accuracy (0.60) among all methods and still has some disparity, as seen in the SPD and DI values (SPD=-0.04 and DI=0.58). This suggests that while LFR attempts to reduce bias, it is less effective in terms of performance, making it less suitable for this dataset. Even though the SPD appears relatively small, the DI highlights a significant disparity in outcomes, demonstrating that SPD alone does not fully capture the extent of bias.

OptimPreproc results in a small drop in accuracy (0.70) and good improvements in fairness. The SPD of -0.08 and DI of 0.80 show that the method helps reduce bias but is less effective than DIR or Reweighing.

The synthetic data method (Women_50K) maintains the same balanced accuracy (0.72) as the baseline. It achieves perfect SPD (0.00), indicating no disparity

between groups, and the DI of 1.01 shows that the outcome for different groups is nearly equal, making this method very effective for reducing bias without sacrificing performance.

Synthetic Data (Women_Non-white_50K) shows good accuracy (0.74) and eliminates the disparity in SPD (0.00). The DI of 1.01 shows that the model performs almost perfectly for fairness.

Synthetic Data (Women_above60_50K) shows the highest accuracy (0.78) among all methods, though it has a slightly negative SPD of -0.09. The DI of 0.76 indicates that it is very effective in reducing bias, though not as much as the other synthetic datasets.

6.1.1.2 Random Forest (RF):

The **baseline for the original data used for DIR** shows a good accuracy (0.75), but there is considerable bias, as indicated by the negative SPD (-0.30), which suggests that the classifier is biased against certain groups. The DI of 0.26 further highlights an imbalance in outcomes for different groups. The baseline for the original data used for other methods (0.72) has slightly lower accuracy, but the fairness metrics (SPD=-0.39 and DI=0.26) are similar, confirming the presence of bias in both cases. **DIR** results in a slight drop in accuracy (0.73) but achieves notable improvements in fairness metrics. The SPD improves to -0.13, indicating reduced disparity, and the DI increases to 0.61, reflecting some progress toward proportional fairness. However, these results suggest that while DIR mitigates bias to some extent, it does not achieve the level of fairness observed with other methods in this setting.

Reweighing shows a larger drop in accuracy (0.69) but achieves notable fairness results. The SPD value increases to 0.06, indicating minimal disparity, and the DI improves beyond the ideal value of 1 to 1.14. This suggests a slight overcompensation in favor of the unprivileged group. While this outcome demonstrates the method's ability to mitigate bias effectively, it also highlights that the fairness improvements can lean disproportionately toward one group. In absolute terms, the DI achieved by Reweighing (1.14) is comparable to the DI achieved by LFR (0.86), as both methods achieve similar fairness results. However, the difference lies in the direction of fairness adjustments: Reweighing favors the unprivileged group, while LFR leans more toward the privileged group.

LFR performs reasonably well in fairness but compromises accuracy (0.68). The SPD improves to -0.07, and the DI value of 0.86 demonstrates much better proportional fairness than the baseline.

OptimPreproc results in an accuracy of 0.71, showing better performance than LFR and Reweighing. The SPD of -0.20 and DI of 0.60 indicate moderate fairness improvements compared to the baseline but are least impressive compared to those achieved by other methods.

The synthetic data method (Women_50K) exhibits a substantial drop in accuracy (0.66) compared to the baseline of 0.72. The SPD improves to -0.02, and

the DI reaches 0.96, reflecting excellent bias mitigation. However, the decline in accuracy makes it a less optimal choice for this classifier.

Synthetic Data (Women_Non-white_50K) shows good improvement in accuracy (0.74). It also achieves fairness improvements with SPD improving to -0.16 and DI to 0.70, making this method a balanced option.

Synthetic Data (Women_above60_50K) achieves the highest accuracy (0.76) among all methods for Random Forest. The fairness metrics show good improvements, with SPD at -0.15 and DI at 0.70. While the fairness outcomes are not perfect, this method strikes a strong balance between accuracy and fairness, making it one of the most effective approaches for achieving a favorable trade-off.

6.1.1.3 Gradient Boosting (GB):

The **baseline for the original data used for DIR** shows the highest accuracy (0.77) among all classifiers, but there is considerable bias, as indicated by the negative SPD (-0.36), which suggests significant disparity between groups. The DI of 0.17 further highlights the imbalance in outcomes for different groups. **The baseline for the original data** used for other methods (0.72) has slightly lower accuracy, but the fairness metrics (SPD=-0.39 and DI=0.27) are similar, confirming the presence of bias. **DIR** results in a slight drop in accuracy (0.74) but shows moderate improvements in fairness metrics. The SPD improves to -0.12, indicating reduced disparity, and the DI increases to 0.58, suggesting some progress toward proportional fairness. However, similar to Random Forest, DIR's fairness improvements are relatively less pronounced compared to some other methods.

Reweighing shows a larger drop in accuracy (0.69) but significantly improves the results. The SPD value increases to 0.06, indicating minimal disparity, and the DI improves significantly to 1.13, suggesting slight overcompensation in favor of the unprivileged group. This method demonstrates strong bias mitigation capabilities but may result in imbalances favoring one group.

LFR shows reasonable fairness improvements but at the cost of accuracy, which drops to 0.68. The SPD improves to -0.07, and the DI reaches 0.86, reflecting much better proportional fairness than the baseline. However, the trade-off in predictive performance makes LFR less optimal for this classifier. Although the accuracy is not the highest among the methods, this trade-off allows LFR to strike the best balance, compared to the other methods.

OptimPreproc achieves an accuracy of 0.71, which is better than Reweighing and LFR. The SPD improves to -0.21, and the DI reaches 0.57, showing moderate fairness improvements. However, its results are less impressive compared to other methods like Reweighing or LFR.

The synthetic data method (Women_50K) results in a substantial drop in accuracy (0.66) compared to the baseline (0.72). While there is a slight improvement in DI, which increases from 0.27 to 0.38, the SPD remains unchanged at -0.39. These results indicate minimal progress in reducing disparity and suggest that the method is

not particularly effective in improving fairness outcomes, making it a less satisfactory option overall.

Synthetic Data (Women_Non-white_50K) achieves a better accuracy (0.71) and fairness improvements, with SPD improving to -0.20 and DI reaching 0.61. This method provides a balanced trade-off between accuracy and fairness outcomes.

Synthetic Data (Women_above60_50K) achieves an accuracy of 0.67, which is slightly lower than the baseline. The SPD improves modestly to -0.31, and the DI reaches 0.42, showing moderate fairness improvements but much less pronounced than those of some other methods like Reweighing and LFR.

6.1.2 Medical Expenditure

The findings presented in Table 5.3 provide an overview of the performance and fairness outcomes for the Medical Expenditure dataset.

6.1.2.1 Logistic Regression (LR):

The **baseline** model, using the **original dataset** for the Disparate Impact Remover (**DIR**) method, achieves a Balanced Accuracy of 0.71. This indicates a relatively strong overall performance. However, the fairness metrics show room for improvement: the SPD is -0.43, suggesting a significant disparity in model performance between different groups. Additionally, the DI is 0.27, implying a moderate level of disparate impact, favoring the privileged group. After applying the **DIR** method for bias mitigation, there is a slight decrease in Balanced Accuracy, which drops to 0.70. However, this method significantly improves fairness: the SPD improves to -0.08, indicating a much smaller disparity between groups. The DI increases to 0.72, demonstrating a better balance in the treatment of different groups.

The baseline for the original data used for OptimPreproc method starts with a lower Balanced Accuracy and a similar DI to the baseline of DIR. However, when applying the **OptimPreproc** method for preprocessing, Balanced Accuracy drops significantly to 0.59. The SPD improves to -0.17, suggesting that the model's fairness has increased but is still not fully balanced. The DI reaches a relatively high fairness of 0.72, indicating that fairness is prioritized, though at the cost of model performance.

The baseline for the original data used remaining methods is notably higher than those for DIR and OptimPreproc in terms of both performance and fairness (with values of 0.75, -0.31, and 0.40, respectively). This is because the subsequent methods utilize additional features, including categorical variables. Therefore, it is important to keep in mind that these methods operate with different baselines, which should be carefully considered when analyzing and comparing the following results.

The **Reweighing** method improves fairness metrics, as indicated by the SPD of -0.03, which is very close to zero, indicating minimal disparity. The DI is 0.89, showing a very high level of fairness. The Balanced Accuracy is slightly reduced from 0.75 to 0.74, indicating a balance between fairness and predictive performance.

The **LFR** method results in a Balanced Accuracy of 0.63, which is a significant reduction compared to the baseline of 0.75, signaling a decline in predictive performance. The SPD is -0.02, indicating a great improvement in fairness. The DI is 0.81, showing a substantial improvement in fairness, though this comes at the cost of a reduced model performance.

The **Synthetic Data (Non-white_UTILIZATION10)** method results in a Balanced Accuracy of 0.81, which is the highest among the methods discussed, indicating strong predictive performance. The SPD is -0.11, showing reduction in disparity. The DI is 0.70, suggesting an acceptable balance of fairness, with slight favor toward the privileged group.

Using **Synthetic Data (Non-white_InsuranceCov2_UTILIZATION10)** variant results in a Balanced Accuracy of 0.80, slightly lower than the previous method but still very strong. The SPD is -0.11, which is similar to the previous synthetic method, and the DI is 0.70, showing a notable improvement in fairness, similar to the previous variant.

Finally, the **Synthetic Data (Non-white_PovertyCat1_UTILIZATION10)** variant yields a Balanced Accuracy of 0.80, which is still quite good, but lower than the first synthetic data variant. The SPD is -0.08, and the DI is 0.76, indicating a balanced approach to fairness and with higher fairness metrics compared to the other synthetic methods.

6.1.2.2 Random Forest (RF):

The **baseline using the original dataset for DIR** method achieves a Balanced Accuracy of 0.71. This indicates good predictive performance. However, the fairness metrics suggest significant disparities, with an SPD of -0.43 and a DI of 0.27, showing a notable disparate impact that favors the privileged group. After applying **DIR**, there is a marked reduction in Balanced Accuracy, which drops to 0.65, indicating a significant performance trade-off. On the fairness side, the SPD improves dramatically to -0.03, showing minimal disparity between groups, and the DI improves substantially to 0.90, reflecting near parity in outcomes. Despite the accuracy reduction, DIR demonstrates strong fairness improvements.

The **baseline for the original data used for the OptimPreproc method** begins with higher fairness metrics than DIR, with an SPD of -0.29 and DI of 0.39, while maintaining a Balanced Accuracy of 0.71. After applying the **OptimPreproc** method, Balanced Accuracy drops to 0.68. However, fairness metrics improve further, with the SPD improving to -0.07 and the DI increasing to 0.80, showing substantial progress toward fairness.

The **baseline for the original data used for the remaining methods** shows both higher predictive performance and fairness (Balanced Accuracy = 0.72, SPD = -0.23, DI = 0.48) compared to the baselines used by DIR and OptimPreproc. This difference is due to the inclusion of categorical variables, as noted earlier. However, it still remains unfair.

The **Reweighing** method significantly improves fairness metrics, with an SPD of -0.00, indicating almost no disparity, and a DI of 0.99, which is close to parity. However, this comes at a steep cost to accuracy, which drops significantly to 0.55, suggesting that Reweighing sacrifices predictive performance for fairness.

The **LFR** method achieves a Balanced Accuracy of 0.68, maintaining performance closer to the baseline. The SPD improves to -0.07, and the DI reaches 0.75, indicating good fairness improvements, though not as much as with Reweighing. This method strikes a reasonable balance between fairness and accuracy for this classifier.

The **Synthetic Data (Non-white_UTILIZATION10)** method results in a Balanced Accuracy of 0.78, the highest among the methods, showing excellent predictive performance. The SPD improves to -0.10, and the DI increases to 0.76, reflecting good fairness improvements alongside strong performance.

The **Synthetic Data (Non-white_InsuranceCov2_UTILIZATION10)** variant achieves a Balanced Accuracy of 0.77, slightly lower than the previous synthetic method. The SPD is -0.12, and the DI is 0.71, showing consistent fairness improvements, though less pronounced compared to the previous variant.

Finally, the **Synthetic Data (Non-white_PovertyCat1_UTILIZATION10)** variant also achieves a Balanced Accuracy of 0.77, which is on par with other synthetic data approaches. The SPD improves to -0.11, and the DI reaches 0.73, offering a balanced approach to fairness and accuracy.

6.1.2.3 Gradient Boosting (GB):

The **baseline** model, using the **dataset tailored DIR**, achieves a Balanced Accuracy of 0.71, reflecting solid predictive performance. However, the fairness metrics are problematic, with an SPD of -0.29, showing noticeable group disparity, and a DI of 0.42, indicating an uneven allocation of positive outcomes between groups. After applying the **DIR** method, Balanced Accuracy decreases slightly at 0.70. Fairness metrics, however, improve significantly, with the SPD improving to -0.10 and the DI increasing to 0.72. These results show substantial fairness improvements, making DIR a strong option for reducing bias in this classifier.

The baseline for the OptimPreproc method has similar fairness metrics compared to DIR's baseline, with an SPD of -0.30 and a DI of 0.41, while the Balanced Accuracy remains at 0.71. After applying **OptimPreproc**, the Balanced Accuracy drops modestly to 0.68, but fairness improves, with the SPD reaching -0.13 and the DI increasing to 0.70. This suggests a focus on fairness, though with some trade-off in model performance.

The **general baseline**, which includes categorical features, yields higher overall fairness and performance values, with a Balanced Accuracy of 0.74, SPD of -0.27, and DI of 0.44. These serve as a useful benchmark when analyzing subsequent methods.

The **Reweighing** method emphasizes fairness, achieving an SPD of -0.05 and a DI of 0.94, which are close to ideal values. However, these gains are achieved at the cost of Balanced Accuracy, which drops substantially to 0.53, indicating a significant

sacrifice in predictive power for fairness.

The **LFR** method results in a Balanced Accuracy of 0.68, which is lower than the baseline (0.75) but much higher than Reweighing (0.53), indicating a more moderate trade-off between fairness and performance. SPD improves from -0.31 to -0.09, reducing disparity significantly. DI increases from 0.40 to 0.72, indicating that fairness has improved, though not as dramatically as with Reweighing.

The Synthetic Data (Non-white_UTILIZATION10) method stands out for its strong Balanced Accuracy of 0.79, the highest among the Gradient Boosting methods. The SPD improves to -0.16, and the DI reaches 0.68, showing solid fairness improvements alongside strong predictive performance.

The **Synthetic Data (Non-white_InsuranceCov2_UTILIZATION10)** variant also performs well, achieving a Balanced Accuracy of 0.79. The fairness metrics are slightly less favorable, with an SPD of -0.18 and a DI of 0.63, but it still demonstrates considerable improvements over the baseline.

Finally, the **Synthetic Data (Non-white_PovertyCat1_UTILIZATION10)** variant maintains a Balanced Accuracy of 0.78. Fairness metrics improve further, with the SPD improving to -0.19 and the DI reaching 0.63, which remains less favorable for proportional fairness outcomes compared to other methods.

6.2 Relative Improvements Over Baselines

In this section, we will focus primarily on interpreting the results of Balanced Accuracy to evaluate model performance and Disparate Impact (DI) to assess fairness. This choice is based on the fact that DI measures the relative disparity between groups, offering a more nuanced understanding of proportional fairness. By concentrating on these two metrics, we aim to provide a clearer understanding of how the models perform in terms of both accuracy and fairness.

To ensure clarity in the interpretation, the results are presented in terms of the difference from the baseline rather than raw values, especially as the DIR method has a different baseline compared to the other methods due to its exclusion of categorical data. Negative values indicate a reduction in the metric compared to the baseline, while positive values signal an improvement.

The analysis is structured into two parts. First, we compare Balanced Accuracy across classifiers to evaluate the relative effectiveness of the bias mitigation methods in maintaining predictive performance. Then, we compare Disparate Impact across classifiers to assess the relative improvements in fairness outcomes achieved by the different methods. These interpretations are based on the results summarized in Table 5.2 for the Adult dataset and Table 5.4 for the Medical Expenditure dataset, as well as the corresponding visualizations in Figure 6.1 and Figure 6.2.

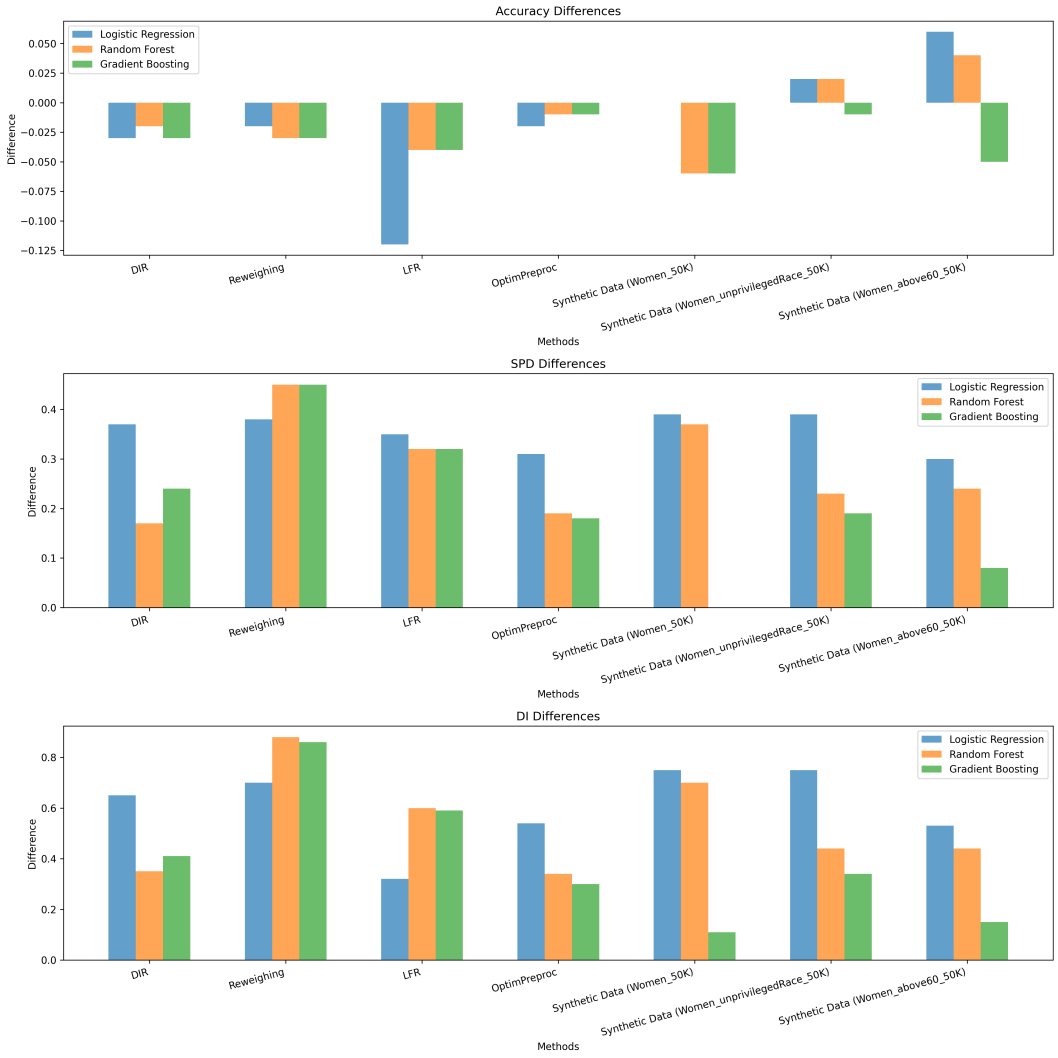

Figure 6.1: Bar plot showing performance and fairness metric differences for the **Adult dataset**.

6.2.1 Adult

6.2.1.1 Balanced Accuracy Differences

The accuracy differences for **Logistic Regression** range from -0.12 to +0.06. This suggests that most methods either have a minimal impact or slightly reduce the accuracy of the model, with the exception of synthetic data methods which improve the accuracy, especially Synthetic Data (Women_above60_50K), which shows the highest increase in accuracy (+0.06). Methods such as Reweighing, OptimPreproc and DIR show small reduction in accuracy, in this order, while LFR shows the biggest drop in accuracy. Synthetic Data (Women_50K) maintains the same performance, so there is no drop in accuracy, and the difference from the baseline is 0, which is why we cannot see it in the plot.

Meanwhile, the accuracy differences for **Random Forest** range from -0.06 to +0.04, indicating a slight reduction in accuracy with most methods, with a few methods leading to small improvements. Synthetic Data (Women_Non-white_50K) shows

the biggest accuracy improvement (+0.04), followed by Synthetic Data (Women_Non-white_50K) with a smaller improvement (+0.02). On the other hand, methods like Synthetic Data (Women_50K) and LFR result in the largest accuracy reductions (-0.06 and -0.04, respectively). Reweighing, DIR, and OptimPreproc also show slight decreases in accuracy, with the reductions ordered as follows: Reweighing (-0.03), DIR (-0.02), and OptimPreproc (-0.01).

Last, the accuracy differences for **Gradient Boosting** range from -0.06 to -0.01, indicating that all methods lead to a reduction in accuracy, with no improvements observed. The largest accuracy drop is seen with Synthetic Data (Women_50K) (-0.06), followed by Synthetic Data (Women_above60_50K) (-0.05) and LFR (-0.04). Other methods, such as Reweighing and DIR, show smaller reductions in accuracy, with differences of -0.03 for both. Synthetic Data (Women_Non-white_50K) shows the smallest accuracy drop of -0.01, similar to OptimPreproc.

6.2.1.2 Disparate Impact Differences

In the context of Disparate Impact (DI), a higher difference indicates a more significant positive change, meaning that the method has a larger effect on improving fairness. The baseline DI value is approximately 0.26, and the ideal value for perfect fairness is 1.0. Therefore, the maximum possible improvement in DI is approximately 0.74 (i.e., 1.0 - 0.26), as any increase beyond this would shift the fairness towards favoring the advantaged group, rather than improving fairness for the disadvantaged group.

The DI differences for **Logistic Regression** range from 0.32 to 0.75, with the highest values seen in Synthetic Data (Women_50K) and Synthetic Data (Women_Non-white_50K), both showing a substantial increase (+0.75). This indicates that these synthetic data methods significantly improve fairness, particularly by increasing representation of underrepresented groups. In comparison, other methods such as Reweighing (+0.70), DIR (+0.65), and OptimPreproc (+0.54) also show positive DI differences, indicating improvements in fairness, though to a lesser extent. LFR yields the lowest increase (+0.32), demonstrating some improvement in fairness but less pronounced than the other methods.

Moving to **Random Forest**, the DI differences range from 0.34 to 0.88, with the largest improvement seen in Reweighing (+0.88). However, it is important to note that Reweighing exceeds the ideal fairness difference threshold of 0.74, reaching 0.88. This indicates that the method may be improving fairness too much, potentially favoring the other group (in this case, the disadvantaged group). This would be equivalent to improving absolute fairness by a difference of 0.6, which is still a significant improvement. Synthetic Data (Women_50K) shows a strong improvement (+0.7), followed by LFR (+0.6). On the other hand, Synthetic Data (Women_above60_50K) and Synthetic Data (Women_50K) show a smaller but still notable improvement (+0.44 for both), indicating a moderate positive impact on fairness for Random Forest. Finally, OptimPreproc and DIR show the smallest improvements (+0.34 and +0.35, respectively), indicating a smaller impact on fairness for Random Forest.

Finally, for **Gradient Boosting**, the DI differences range from 0.11 to 0.86, with the largest improvement observed in Reweighing (+0.86). However, similar to Random Forest, it is important to note that Reweighing exceeds the ideal fairness difference threshold of 0.74. Other methods like LFR (+0.59) and DIR (+0.41) show moderate improvements, indicating positive but less dramatic changes in fairness. Synthetic Data (Women_Non-white_50K) shows a slight improvement (+0.34), while Synthetic Data (Women_above60_50K) shows an even smaller improvement (+0.15). On the other hand, Synthetic Data (Women_50K) shows the smallest improvement (+0.11), indicating a minimal impact on fairness for Gradient Boosting.

6.2.2 Medical Expenditure

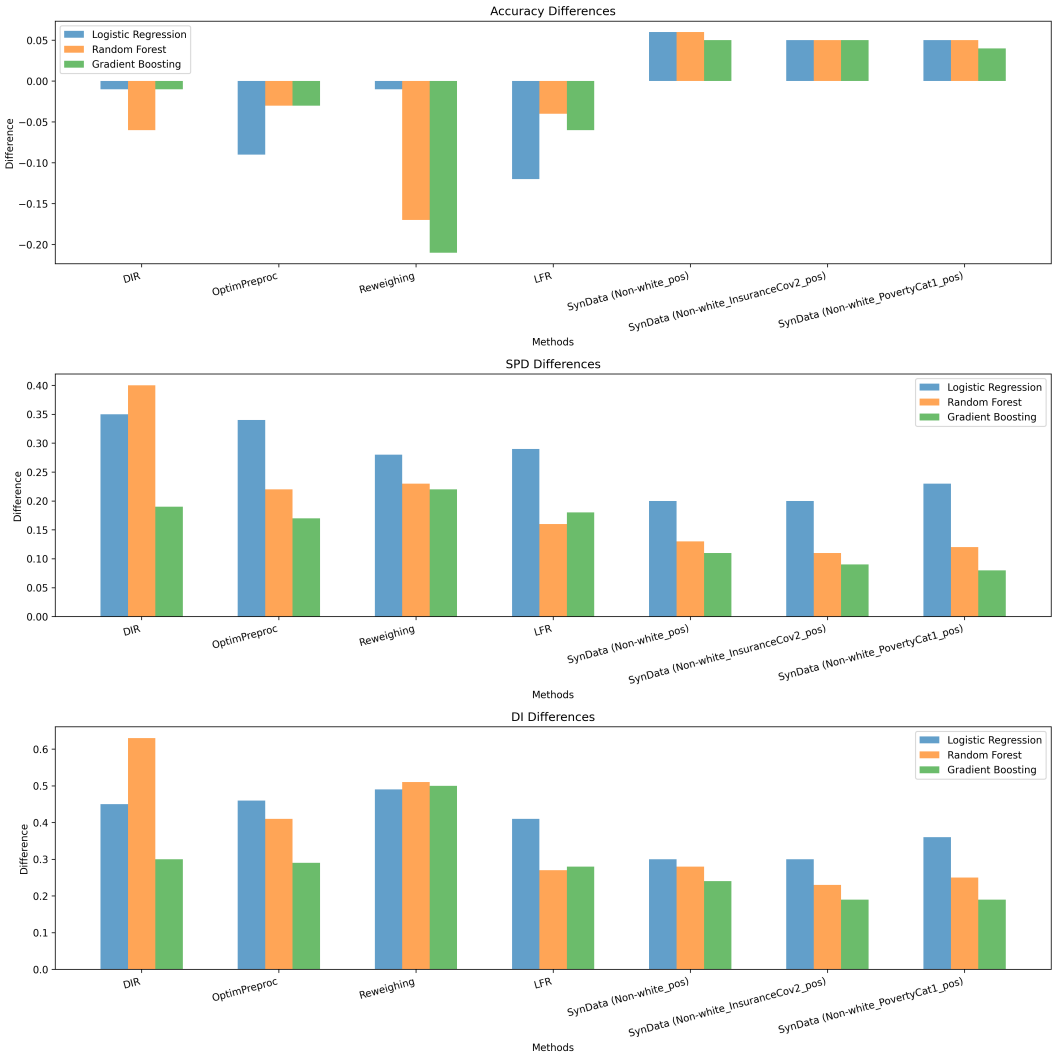

Figure 6.2: Bar plot showing performance and fairness metric differences for the **MEPS dataset**.

6.2.2.1 Balanced Accuracy Differences

The accuracy differences for **Logistic Regression** range from -0.12 to +0.06. These results show that most methods result in a slight reduction in accuracy compared to the baseline, with the largest drop seen for the LFR method (-0.12). Synthetic data methods show the largest improvement in accuracy, particularly Synthetic Data (Non-white_UTILIZATION10), which shows a modest increase of +0.06, followed closely by other synthetic data methods with a smaller improvement of +0.05. DIR and Reweighing both show a minimal reduction in accuracy (-0.01), while OptimPreproc shows a larger reduction in accuracy (-0.09).

The accuracy differences for **Random Forest** range from -0.17 to +0.06. Most methods lead to a slight reduction in accuracy compared to the baseline, with the largest decrease observed for the Reweighing method (-0.17). Synthetic data methods show positive improvements, with Synthetic Data (Non-white_UTILIZATION10) increasing accuracy by +0.06, while Synthetic Data (Non-white_InsuranceCov2_UTILIZATION10) and Synthetic Data (Non-white_PovertyCat1_UTILIZATION10) both result in a +0.05 improvement. In contrast, methods such as OptimPreproc (-0.03), LFR (-0.04), and DIR (-0.06) result in slight reductions in accuracy.

However, the accuracy differences for **Gradient Boosting** range from -0.21 to +0.05. Most methods result in a slight reduction in accuracy compared to the baseline, with the largest decrease observed for the Reweighing method (-0.21). LFR also leads to a moderate accuracy drop (-0.06), while OptimPreproc (-0.03) and DIR (-0.01) show smaller reductions. On the other hand, synthetic data methods improve accuracy, with Synthetic Data (Non-white_UTILIZATION10) and Synthetic Data (Non-white_InsuranceCov2_UTILIZATION10) both showing an increase of +0.05, while Synthetic Data (Non-white_PovertyCat1_UTILIZATION10) results in a slightly smaller improvement of +0.04.

6.2.2.2 Disparate Impact Differences

The Disparate Impact (DI) differences for **Logistic Regression** range from 0.30 to 0.49, indicating that all methods contribute to improving fairness to some extent. The highest increase is observed with the Reweighing method (+0.49), followed closely by OptimPreproc (+0.46) and DIR (+0.45), suggesting that these methods are particularly effective in mitigating bias. LFR also shows a considerable improvement (+0.41), though slightly lower than the top-performing methods. In contrast, synthetic data methods yield smaller differences in fairness, with Synthetic Data (Non-white_UTILIZATION10) and Synthetic Data (Non-white_InsuranceCov2_UTILIZATION10) both at +0.30, while Synthetic Data (Non-white_PovertyCat1_UTILIZATION10) achieves a slightly higher improvement of +0.36. Moving on to the next classifier, the DI differences for **Random Forest** range from 0.23 to 0.63, indicating varying degrees of fairness improvement across methods. The highest increase is observed with DIR (+0.63), suggesting that it has the strongest

effect in reducing bias within Random Forest. Reweighing and OptimPreproc also show notable improvements (+0.51 and +0.41, respectively), though to a lesser extent than DIR. LFR, on the other hand, results in a smaller DI increase (+0.27). Synthetic data methods show the least impact on fairness, with DI increases ranging from +0.23 to +0.28. Among them, Synthetic Data (Non-white_UTILIZATION10) achieves the highest improvement (+0.28), while the other two synthetic data methods, Synthetic Data (Non-white_InsuranceCov2_UTILIZATION10) and Synthetic Data (Non-white_PovertyCat1_UTILIZATION10), yield slightly lower increases (+0.23 and +0.25, respectively). Lastly, the DI differences for **Gradient Boosting** range from 0.19 to 0.50, indicating that while some methods improve fairness more, others have a more limited effect. The largest increase is observed with Reweighing (+0.50), suggesting that it has the strongest impact on mitigating bias in this model. DIR and OptimPreproc show moderate improvements in fairness, with DI increases of +0.30 and +0.29, respectively. LFR has a slightly smaller effect (+0.28), indicating that its ability to enhance fairness in Gradient Boosting is somewhat weaker compared to other methods. Synthetic data methods exhibit the lowest DI improvements, ranging from +0.19 to +0.24. Among them, Synthetic Data (Non-white_UTILIZATION10) achieves the highest fairness gain (+0.24), while Synthetic Data (Non-white_InsuranceCov2_UTILIZATION10) and Synthetic Data (Non-white_PovertyCat1_UTILIZATION10) show the least impact (+0.19 each).

6.3 Key Trends Across Datasets

This section highlights key observations regarding the performance and fairness impacts of synthetic data methods and other bias mitigation techniques on both the **Adult** and **Medical Expenditure (MEPS)** datasets. Insights are grouped by dataset for clarity.

In the **Adult** dataset, synthetic data methods work best with Logistic Regression, which benefits from the increased representation of underrepresented groups, while Random Forest and Gradient Boosting show limited improvement. These tree-based models already excel at handling complex relationships within the data and therefore do not benefit as much from synthetic data augmentation. LFR performs significantly better with Random Forest and Gradient Boosting because these models can handle nonlinear relationships, which LFR is designed to optimize, while Logistic Regression struggles to effectively model such complexities. Reweighing works well with Logistic Regression but shows an unusually large impact on fairness in Random Forest and Gradient Boosting, likely due to the linear nature of the method, which might not always align with the nonlinear characteristics of these models OptimPreproc and DIR tend to work better with Logistic Regression, as their adjustments have more noticeable effects in linear models. This analysis highlights the need to carefully consider classifier and method compatibility to achieve an optimal balance between fairness and predictive performance.

A notable observation is that the **Medical Expenditure** dataset consistently

exhibits a smaller increase in fairness compared to the Adult dataset. Across all classifiers in MEPS, synthetic data methods show the smallest improvements in fairness. The contribution of synthetic data to reducing bias gets progressively lower with each classifier. However, synthetic data show positive accuracy improvements for all classifiers (while in the Adult dataset they were mainly performing better with LR). Reweighing stands out as the most effective fairness mitigation method across all classifiers, yielding the highest DI increases. However, it introduces a trade-off, significantly reducing accuracy for Random Forest and Gradient Boosting. In contrast, Reweighing achieves a good balance for Logistic Regression, with an accuracy of 0.74 and a DI of 0.89 (differences of -0.01 and +0.49 from their respective baselines). An intriguing observation is that LFR performs better in terms of DI with Logistic Regression than with more complex models like Random Forest or Gradient Boosting. This was unexpected, as its design for handling nonlinear relationships is better suited for tree-based models. However, this improvement in fairness comes at a notable cost to accuracy, particularly for Logistic Regression, while smaller accuracy drops are observed for the other classifiers.

6.3.1 Insights and Implications

This analysis highlights a critical point: the effectiveness of bias mitigation methods is not one-size-fits-all but depends on the dataset, classifier, and method. It underscores the importance of carefully selecting mitigation strategies tailored to the context. Logistic Regression benefits the most across all fairness techniques due to its linear nature, while Random Forest and Gradient Boosting pose unique challenges due to their inherent complexity. Furthermore, the variability in results between the Adult and Medical Expenditure datasets highlights the importance of testing methods across diverse datasets to derive generalizable insights. Such insights emphasize the need for further exploration into the nuanced interactions between methods, models, and datasets, paving the way for more adaptable and effective bias mitigation strategies in the future.

Chapter 7

Conclusion

The increasing integration of Artificial Intelligence (AI) in decision-making processes across various sectors, including healthcare, finance, and criminal justice, brings significant opportunities but also raises critical concerns regarding fairness and bias. This thesis has extensively explored the issue of gender and racial biases in AI systems, emphasizing a data-centric approach to achieve fairer outcomes.

A key aspect of this study involved evaluating preprocessing methods for bias mitigation using the AI Fairness 360 (AIF360) toolkit [1] and synthetic data generation through Clearbox AI's Synthetic Kit [2]. The experimental analysis was conducted on two datasets, the Adult dataset [3] and the Medical Expenditure dataset [4], both containing sensitive attributes such as gender and race. Preprocessing methods such as Reweighing, Disparate Impact Remover, Learning Fair Representations, and Optimized Preprocessing demonstrated their potential to reduce bias effectively, while synthetic data augmentation emerged as a promising strategy for enhancing fairness through targeted augmentation of underrepresented groups. These techniques improved fairness metrics, including Statistical Parity Difference and Disparate Impact, without significantly compromising predictive accuracy. This indicates that fairness and performance need not be mutually exclusive.

One of the key findings of this research is that no single bias mitigation technique is universally effective across all models and datasets. Different classifiers exhibit varying degrees of sensitivity to bias correction, with Logistic Regression benefiting the most, while more complex models like Random Forest and Gradient Boosting present additional challenges. The study underscores the need for a nuanced, context-specific approach to bias mitigation, rather than relying on generic solutions.

Beyond the technical aspects, this thesis highlights the broader implications of bias in AI and the necessity for an interdisciplinary approach to addressing fairness issues. While technical solutions provide important tools for mitigating bias, they must be complemented by regulatory frameworks, organizational policies, and ethical guidelines to ensure AI systems operate transparently and equitably.

Despite these advancements, the study acknowledges certain limitations. The scope was restricted to preprocessing methods and datasets like Adult and Medical Expenditure, which may not fully represent the complexities of real-world applica-

tions. Additionally, the dependency on predefined sensitive attributes for fairness evaluations could miss nuanced bias patterns inherent in diverse datasets. Moreover, understanding the interplay between various bias mitigation approaches across different classifiers offers another valuable avenue for exploration. Future work may expand on this research by exploring in-processing and post-processing bias mitigation techniques, as well as testing across more varied and dynamic datasets. Further investigation into advanced synthetic data generation methods, could uncover deeper insights into reducing bias in AI systems. By integrating these approaches, the vision of truly fair, accountable, and transparent AI systems can move closer to reality.

In conclusion, this thesis contributes meaningfully to the growing field of responsible AI by providing practical solutions to mitigate bias in machine learning pipelines. The insights derived from this work lay a solid foundation for deploying fairer AI technologies, ensuring inclusivity and equity in decision-making processes across critical domains such as healthcare and socio-economic applications.

Bibliography

[1] Rachel K. E. Bellamy et al. *AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias*. GitHub: `https://github.com/Trusted-AI/AIF360`. 2018. arXiv: `1810.01943 [cs.AI]`. URL: `https://arxiv.org/abs/1810.01943` (cit. on pp. I, 4, 7, 14, 15, 19, 34, 61).

[2] Clearbox AI. *Clearbox Synthetic Kit: The synthetic data generation library*. `https://github.com/Clearbox-AI/clearbox-synthetic-kit`. 2025 (cit. on pp. I, 17, 20, 36, 37, 61).

[3] M. Lichman. *UCI Machine Learning Repository: Adult dataset*. `https://archive.ics.uci.edu/dataset/2/adult`. 2013 (cit. on pp. I, 12, 61).

[4] Agency for Healthcare Research and Quality. *Medical Expenditure Panel Survey (MEPS)*. `https://meps.ahrq.gov/mepsweb/`. 2015 (cit. on pp. I, 12, 61).

[5] Mohammad Al-Rubaie and J. Morris Chang. *Privacy Preserving Machine Learning: Threats and Solutions*. 2018. arXiv: `1804.11238 [cs.CR]`. URL: `https://arxiv.org/abs/1804.11238` (cit. on p. 2).

[6] Khaled El Emam. *A De-identification Protocol for Open Data*. `https://bit.ly/33AetZq`. IAPP Privacy Tech, May 16, 2016. 2016 (cit. on pp. 2, 10).

[7] Khaled El Emam, Lucy Mosquera, and Richard Hoptroff. *Practical Synthetic Data Generation: Balancing Privacy and the Broad Availability of Data*. O'Reilly Media, 2020. ISBN: 978-1-492-07274-4 (cit. on pp. 2, 9, 10).

[8] Kate Crawford. *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. New Haven, CT: Yale University Press, 2021. ISBN: 9780300264630. URL: `https://yalebooks.yale.edu/book/9780300264630/atlas-of-ai/` (cit. on p. 3).

[9] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. *Fairness Constraints: Mechanisms for Fair Classification*. 2017. arXiv: `1507.05259 [stat.ML]`. URL: `https://arxiv.org/abs/1507.05259` (cit. on p. 3).

[10] Cathy O'Neil. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. USA: Crown Publishing Group, 2016. ISBN: 0553418815 (cit. on p. 3).

[11] Safiya Umoja Noble. *Algorithms of oppression: How search engines reinforce racism*. New York: NYU Press, 2018. ISBN: 978-1479837243 (cit. on p. 3).

[12] Virginia Eubanks. *Automating inequality: How high-tech tools profile, police, and punish the poor*. New York: St. Martin's Press, 2018. ISBN: 978-1250074317 (cit. on p. 3).

[13] European Commission. *Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on Artificial Intelligence (AI Act) and amending certain Union legislative acts*. COM(2021) 206 final, 2021/0106 (COD), Brussels, 21.4.2021. 2021. URL: https://eur-lex.europa.eu/resource.html?uri=cellar:e0649735-a372-11eb-9585-01aa75ed71a1.0001.02/DOC_1&format=PDF (cit. on p. 3).

[14] Organisation for Economic Co-operation and Development (OECD). *OECD Principles on Artificial Intelligence*. 2019. URL: https://www.oecd.org/going-digital/ai/principles/ (cit. on p. 3).

[15] Reva Schwartz, Apostol Vassilev, Kristen Greene, Lori Perine, Andrew Burt, and Patrick Hall. *Managing Bias in Artificial Intelligence*. https://doi.org/10.6028/NIST.SP.1270. NIST Special Publication 1270, National Institute of Standards and Technology. 2022 (cit. on p. 6).

[16] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. "Machine Bias". In: *ProPublica* (2016). URL: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (cit. on p. 6).

[17] Jeffrey Dastin. "Amazon scraps secret AI recruiting tool that showed bias against women". In: *Reuters* (2018). URL: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G (cit. on p. 6).

[18] Joy Buolamwini and Timnit Gebru. "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification". In: *Proceedings of Machine Learning Research* 81 (2018), pp. 1–15 (cit. on p. 7).

[19] Pedro Saleiro, Benedict Kuester, Loren Hinkson, Jesse London, Abby Stevens, Ari Anisfeld, Kit T. Rodolfa, and Rayid Ghani. *Aequitas: A Bias and Fairness Audit Toolkit*. GitHub: https://github.com/dssg/aequitas. 2019. arXiv: 1811.05577 [cs.LG]. URL: https://arxiv.org/abs/1811.05577 (cit. on p. 7).

[20] Meike Zehlike, Carlos Castillo, Francesco Bonchi, Ricardo Baeza-Yates, Sara Hajian, and Mohamed Megahed. *FAIRNESS MEASURES: A Platform for Data Collection and Benchmarking in Discrimination-Aware ML*. https://fairnessmeasures.github.io. June 2017. URL: https://fairnessmeasures.github.io (cit. on p. 7).

[21] Florian Tramer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, Jean-Pierre Hubaux, Mathias Humbert, Ari Juels, and Huang Lin. "FairTest: Discovering Unwarranted Associations in Data-Driven Applications". In: *arXiv preprint arXiv:1510.02377* (2015). GitHub: https://github.com/columbia/fairtest (cit. on p. 7).

[22] J. A. Adebayo. “FairML: Toolbox for Diagnosing Bias in Predictive Modeling”. MA thesis. Massachusetts Institute of Technology, 2016. URL: https://github.com/adebayoj/fairml (cit. on p. 7).

[23] Sainyam Galhotra, Yuriy Brun, and Alexandra Meliou. “Fairness testing: testing software for discrimination”. In: *Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering*. ESEC/FSE'17. GitHub: https://github.com/LASER-UMASS/Themis. ACM, Aug. 2017, pp. 498–510. DOI: 10.1145/3106237.3106277. URL: http://dx.doi.org/10.1145/3106237.3106277 (cit. on p. 7).

[24] Sorelle A. Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P. Hamilton, and Derek Roth. *A Comparative Study of Fairness-Enhancing Interventions in Machine Learning*. GitHub: https://github.com/algofairness/fairness-comparison. 2018. arXiv: 1802.04422 [stat.ML]. URL: https://arxiv.org/abs/1802.04422 (cit. on p. 8).

[25] Niels Bantilan. *Themis-ml: A Fairness-aware Machine Learning Interface for End-to-end Discrimination Discovery and Mitigation*. GitHub: https://github.com/cosmicBboy/themis-ml. 2017. arXiv: 1710.06921 [cs.CY]. URL: https://arxiv.org/abs/1710.06921 (cit. on p. 8).

[26] 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. Sydney, NSW, Australia: Association for Computing Machinery, 2015, pp. 259–268. ISBN: 9781450336642. DOI: 10.1145/2783258.2783311. URL: https://doi.org/10.1145/2783258.2783311 (cit. on pp. 8, 16).

[27] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork. “Learning Fair Representations”. In: *Proceedings of the International Conference on Machine Learning*. JMLR.org, 2013, pp. 325–333 (cit. on pp. 8, 16).

[28] F. P. Calmon, D. Wei, B. Vinzamuri, K. Natesan Ramamurthy, and K. R. Varshney. “Optimized Pre-Processing for Discrimination Prevention”. In: *Conference on Neural Information Processing Systems (NeurIPS)*. 2017. URL: https://github.com/fair-preprocessing/nips2017 (cit. on pp. 8, 17).

[29] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. *Mitigating Unwanted Biases with Adversarial Learning*. 2018. arXiv: 1801.07593 [cs.LG]. URL: https://arxiv.org/abs/1801.07593 (cit. on p. 8).

[30] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. “Fairness-Aware Classifier with Prejudice Remover Regularizer”. In: *Machine Learning and Knowledge Discovery in Databases*. Ed. by Peter A. Flach, Tijl De Bie, and Nello Cristianini. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 35–50. ISBN: 978-3-642-33486-3 (cit. on p. 8).

[31] Moritz Hardt, Eric Price, and Nathan Srebro. *Equality of Opportunity in Supervised Learning*. 2016. arXiv: 1610.02413 [cs.LG]. URL: https://arxiv.org/abs/1610.02413 (cit. on p. 8).

[32] Faisal Kamiran, Asim Karim, and Xiangliang Zhang. "Decision Theory for Discrimination-Aware Classification". In: *2012 IEEE 12th International Conference on Data Mining*. 2012, pp. 924–929. DOI: 10.1109/ICDM.2012.45 (cit. on p. 8).

[33] Jonathan Tilley. *Automation, Robotics, and the Factory of the Future*. https://oreil.ly/L27Ol. McKinsey, September 2017. 2017 (cit. on p. 9).

[34] Health Data Insight. *Simulacrum: Artificial Patient-like Cancer Data to Help Researchers Gain Insights*. https://simulacrum.healthdatainsight.org.uk/. 2023 (cit. on p. 10).

[35] William Edwards Deming and Frederick F. Stephan. "On a Least Squares Adjustment of a Sampled Frequency Table When the Expected Marginal Totals are Known". In: *Annals of Mathematical Statistics* 11 (1940), pp. 427–444. URL: https://api.semanticscholar.org/CorpusID:121777010 (cit. on p. 10).

[36] Richard J. Beckman, Keith A. Baggerly, and Michael D. McKay. "Creating synthetic baseline populations". In: *Transportation Research Part A: Policy and Practice* 30.6 (1996), pp. 415–429. ISSN: 0965-8564. DOI: https://doi.org/10.1016/0965-8564(96)00004-3. URL: https://www.sciencedirect.com/science/article/pii/0965856496000043 (cit. on p. 10).

[37] Zengyi Huang. "A Comparison of Synthetic Reconstruction and Combinatorial Optimization Approaches to the Creation of Small-Area Micro Data". In: 2002. URL: https://api.semanticscholar.org/CorpusID:39354642 (cit. on p. 10).

[38] B. Draghi, Z. Wang, P. Myles, and A. Tucker. "Identifying and Handling Data Bias within Primary Healthcare Data Using Synthetic Data Generators". In: *Heliyon* 10 (2024), e24164. DOI: 10.1016/j.heliyon.2024.e24164 (cit. on p. 10).

[39] B. Van Breugel, T. Kyono, J. Berrevoets, and M. van der Schaar. "DECAF: Generating Fair Synthetic Data Using Causally-Aware Generative Networks". In: *Advances in Neural Information Processing Systems*. 2021, pp. 22221–22233 (cit. on p. 10).

[40] S. Gujar, T. Shah, D. Honawale, V. Bhosale, F. Khan, D. Verma, and R. Ranjan. "GenEthos: A Synthetic Data Generation System with Bias Detection and Mitigation". In: *Proceedings of the International Conference on Computing, Communication, Security and Intelligent Systems (IC3SIS)*. Kochi, India, June 2022, pp. 23–25 (cit. on p. 11).

[41] J. Baumann, A. Castelnovo, A. Cosentini, R. Crupi, N. Inverardi, and D. Regoli. “Bias On Demand: Investigating Bias with a Synthetic Data Generator”. In: *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence (IJCAI-23) Demonstrations Track*. Macao, China, Aug. 2023, pp. 19–25 (cit. on p. 11).

[42] F. Kamiran and T. Calders. “Data Preprocessing Techniques for Classification without Discrimination”. In: *Knowledge and Information Systems* 33.1 (2012), pp. 1–33. DOI: 10.1007/s10115-011-0463-8 (cit. on p. 15).