UNIVERSIDADE FEDERAL DE MINAS GERAIS
Instituto de Ciências Exatas
Programa de Pós-Graduação em Ciência da Computação

Mírian Francielle da Silva

A Study of the Nuances of AI Fairness Development in Practice: A Framework for Designing Bias Mitigation Interventions

Belo Horizonte
2024

Mírian Francielle da Silva

A Study of the Nuances of AI Fairness Development in Practice: A Framework for Designing Bias Mitigation Interventions

Final Version

Thesis presented to the Graduate Program in Computer Science of the Federal University of Minas Gerais in partial fulfillment of the requirements for the degree of Master in Computer Science.

Advisor: Ana Paula Couto
Co-Advisor: Marisa Affonso Vasconcelos

Belo Horizonte
2024

2024, Mírian Francielle da Silva.
Todos os direitos reservados.

Silva, Mírian Francielle da.

S586s A study of the nuances of AI fairness development in practice: a framework for designing bias mitigation interventions [recurso eletrônico] / Mírian Francielle da Silva – 2024.

1 recurso online (102 f. il., color.) pdf.

Orientadora: Ana Paula Couto da Silva.
Coorientadora: Marisa Affonso Vasconcelos.
Dissertação (mestrado) - Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Ciência da Computação.
Referências: f. 79-86.

1. Computação – Teses. 2. Inteligência artificial – Teses. I. Silva, Ana Paula Couto da. II. Vasconcelos, Marisa Affonso III. Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Ciência da Computação. III. Título.

CDU 519.6*82(043)

Ficha catalográfica elaborada por Célio Resende Diniz, bibliotecário CRB 6/2403 - Universidade Federal de Minas Gerais – ICEx.

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

FOLHA DE APROVAÇÃO

A Study of the Nuances of AI Fairness Development in Practice: A Framework for Designing Bias Mitigation Interventions

MÍRIAN FRANCIELLE DA SILVA

Dissertação defendida e aprovada pela banca examinadora constituída pelos Senhores:

Documento assinado digitalmente
gov.br ANA PAULA COUTO DA SILVA
Data: 01/11/2024 15:51:19-0300
Verifique em https://validar.iti.gov.br

PROFA. ANA PAULA COUTO DA SILVA - Orientadora
Departamento de Ciência da Computação - UFMG

Documento assinado digitalmente
gov.br MARISA AFFONSO VASCONCELOS
Data: 05/11/2024 10:12:05-0300
Verifique em https://validar.iti.gov.br

DRA. MARISA AFFONSO VASCONCELOS - Coorientadora
Residente Pós-Doutoral - PPGCC - UFMG

PROF. FLAVIO VINICIUS DINIZ DE FIGUEIREDO
Departamento de Ciência da Computação - UFMG

Documento assinado digitalmente
gov.br JONICE DE OLIVEIRA SAMPAIO
Data: 01/11/2024 18:11:25-0300
Verifique em https://validar.iti.gov.br

PROF. JONICE DE OLIVEIRA SAMPAIO
Departamento de Ciência da Computação - UFRJ

Belo Horizonte, 1 de novembro de 2024.

To everyone who helped me realize this goal—family, friends, professors, and mentors—your belief in me made all the difference. And to those who will come after, especially black women in science, may this work inspire you as others' work inspired me.

Acknowledgments

This research work wouldn't have been possible without the profound companionship of friends and family. It was motivated by the work of many black women researchers in the field of AI ethics, to whom I am very grateful for their great contributions to society.

First and foremost, I would like to thank my family for believing in me and providing me with everything I needed to embark on this academic journey. My parents, Geraldino Furtunato and Célia Rocha believed in me from the beginning and provided me with everything I needed to pursue my studies, even before I started this Master's degree. I want to thank my mother for all her prayers and support during the most challenging and rewarding moments. I would also like to thank my younger sister Nikolle Silva, who, despite not being a computer science student, dedicated her time to listening to me countless times, reading everything I had written, and assisting me with the formatting and structure of this work.

I am deeply grateful to Renan Bomtempo, my beloved partner and friend, who has been my constant source of support since the beginning of this journey. He was willing to study with me, read scientific papers to help me decide on my references, and listen to my endless complaints. His company has been and continues to be a constant source of strength. I thank you for your presence during the moments of sadness and joy during this period and for all your encouragement to not give up on this work. I thank you for always believing in me and my work.

I am grateful to the Black in AI organization for giving me every opportunity to attend top-tier AI conferences. Thank you for the financial support that allowed me to present my work at international conferences, for the enriching experiences of attending all the latest NeurIPS conferences, and for the academic and industrial connections I have made through Black in AI. I would not have started this Master's degree without the influence of Black in AI on my career.

I am grateful to have been selected to participate in the BEAAMO research group at the University of California, Berkeley, during this Master's research. I would like to thank Ramon Vilarino for encouraging me to enroll, Professor Rediet Abebe for her guidance throughout my time at UC Berkeley, and my fellow research colleagues with whom I've shared this experience, which greatly enriched both my work and my personal development as a researcher.

I am glad to have been part of the LOCUS research laboratory at UFMG and thankful for all the friends I've made there. There were many parties, lunches, and "*lanchinhos da tarde*" that made my experience back at UFMG more pleasant.

I would like to thank my team at IBM Research for allowing me to work and pursue my Master's degree. I am grateful for my managers' and teammates' understanding and

support during this journey. I am also thankful for contributing to research projects at IBM Research during my Master's; these experiences contributed to both my professional and academic growth.

I would also like to thank the Taws research group from ESPOL (La Escuela Superior Politécnica del Litoral) in Ecuador for welcoming me and giving me the opportunity to present my current research at the *Women in Data Science 2024* conference. It was a rewarding experience, and it greatly motivated me to continue my work, knowing that I can inspire many other students.

I want to thank Mariano Beiró from Universidad de San Andrés in Argentina, who also agreed to co-advise me during the development of this work. I am immensely grateful for all his commitment, assistance, and advice while developing this research. I would also like to thank my co-advisor, Marisa Vasconcelos, for all the detailed reviews, guidance, and monitoring during this work and my advisor, Ana Paula Couto, for the academic supervision throughout the Master's degree.

Without my friends, I wouldn't be where I am today. I thank everyone who took the time to listen to me, discuss ideas, review my texts, and help me relieve all the stress during this journey, whether it was in a bar conversation or even filling my inbox with funny videos and *memes*. This is not a comprehensive list, but to everyone who has been with me and knows how grateful I am, thank you so much. A special thanks to Gustavo, Erick, Pedrinho, Igor Iorc, Renanzinho and Larissa Batista.

I would also like to thank Julliete Nurimba for her professionalism in supporting me and helping me take care of my mental health throughout this period.

Lastly, but no less important, I thank the Federal University of Minas Gerais and the Department of Computer Science for being the catalyst for this degree.

"*Never be limited by other people's limited imaginations*"
(Dr. Mae C. Jemison)

Resumo

Intervenções para alcançar "justiça" (*fairness*) são um foco importante na maioria dos campos de pesquisa de ética em Inteligência Artificial (IA). Quando vieses relacionados a algumas características (por exemplo, raça, sexo, idade, religião) são identificados em sistemas de IA e contribuem para resultados que contém discriminação, desenvolvedores, engenheiros de IA ou partes interessadas devem escolher como e quando intervir para reduzir a disseminção de injustiças. No entanto, a literatura contém uma infinidade de noções diferentes que caracterizam as definições de *fairness* e de geração de viés. Ainda assim, a necessidade de um processo mais padronizado de intervenção, continua sendo uma tarefa desafiadora, pois exige determinar o método adequado de redução de injustiças para um determinado contexto ou tarefa. Nesta pesquisa, exploramos as definições de *fairness* com base em conceitos estatísticos com o objetivo de abordar a necessidade de desenvolvimento responsável de IA, propondo uma estrutura focada na mitigação de vieses em tarefas de classificação binária dentro do aprendizado de máquina supervisionado. A estrutura proposta é guiada por três critérios de equidade (*fairness*) estatística: Independência, Separação e Suficiência, discutidos no trabalho de Barocas et al. (2023), e visa explorar os *trade-offs* entre essas diferentes caracterizações. Nosso estudo destaca a complexidade que é definitir *fairness*, que é dependente do contexto e tarefas, ressaltando os desafios em alcançar critérios de avaliação universalmente aplicáveis, pois mostramos que essas três definições não podem ser alcançadas simultaneamente. Revisamos métodos de intervenção de viés existentes, incluindo técnicas aplicadas em diferentes estágios do desenvolvimento de um *pipeline* de aprendizado de máquina. Essa pesquisa contribui principalmente com uma estrutura que pode aprimorar o estado-da-arte do processo de desenvolvimento e aplicação de técnicas de mitigação e avaliação de viés, com base em conceitos estatísticos, comparando várias definições de imparcialidade, evitando soluções do tipo "tamanho único". Possibilitando o uso de uma abordagem que oferece múltiplas perspectivas do contexto para contemplar vários públicos, e evitar personalizações caso-a-caso (para métricas de justiça em grupo), independentemente dos vieses, objetivos ou partes interessadas envolvidas. Aplicamos nossa estrutura por meio de um estudo de caso usando dados do censo dos EUA American Community Survey (ACS) para prever níveis de renda, demonstrando a aplicação de várias intervenções de redução de viés e seu impacto no desempenho do modelo. As descobertas da pesquisa revelam que diferentes critérios de justiça levam a resultados distintos, enfatizando a importância de aplicar diferentes definições de justiça afim de selecionar a mais apropriada com base no contexto específico e nas implicações sociais em relação aos objetivos das partes interessadas. Também discutimos as limitações das ferramentas atuais de avaliação e quantificação de viéses e a ne-

cessidade de diretrizes de ética de IA mais práticas e implementáveis de acordo com diferentes noções de *fairness*. A estrutura proposta está disponível como uma ferramenta de código-aberto, com objetivo de auxiliar a comunidade de profissionais de IA a expandir as formas de aplicar mitigações de injustiças em sistemas de IA abordando uma variedade de perspectivas de *fairness*.

Palavras-chave: inteligência artificial; justiça algorítmica; mitigação de viéses; aprendizagem de máquina

Abstract

Fairness interventions are a key focus in most Artificial Intelligence (AI) ethics research fields. When biases related to some features (e.g., race, sex, age, religion) are identified in AI systems and contribute to discrimination outcomes, developers, engineers, or stakeholders must choose how and when to intervene. However, the literature contains a plethora of different notions that characterize the definitions of fairness and sources of bias. Still, with a need for more standardization in the intervention process, it remains challenging to determine the suitable option for a given context or highlight the difference between them. In this work, we explore the definitions of fairness based on statistical concepts with the goal of addressing the need for responsible AI development by proposing a framework focused on mitigating bias in binary classification tasks within supervised machine learning. The framework is guided by three statistical fairness criteria: Independence, Separation, and Sufficiency, discussed in the work by Barocas et al. (2023), and the aim is to explore the trade-offs between these different fairness definitions. Our study highlights the complexity of fairness, which is multifaceted and context-dependent, and underscores the challenges in achieving universally applicable evaluation criteria, as we show that these three definitions cannot be achieved simultaneously. We review existing bias intervention methods, including techniques applied at different stages of developing a machine learning pipeline. This research mainly contributes to a framework that can improve the state-of-the-art process of developing and applying bias mitigation and assessment techniques based on statistical concepts, comparing various definitions of fairness, and avoiding "one-size-fits-all" solutions. Enabling the use of an approach that offers multiple perspectives of the context to contemplate various audiences and avoid case-by-case customizations (for group fairness metrics), regardless of the biases, objectives, or stakeholders involved. We apply our framework through a case study using data from the American Community Survey (ACS) US census to predict income levels, demonstrating the application of various fairness interventions and their impact on model performance. The research findings reveal that different fairness criteria lead to distinct outcomes, emphasizing the importance of applying various definitions in order to select the appropriate fairness definitions based on the specific context and societal implications regarding the stakeholders' goals. We also discuss the limitations of current fairness assessment and quantification tools and the need for more practical and implementable AI ethics guidelines according to different fairness notions. The proposed framework is available as an open-source tool, intending to assist the AI practitioner community in expanding the ways to apply bias mitigating bias approaches in AI systems by addressing a variety of fairness perspectives.

Keywords: artificial intelligence; AI fairness; bias mitigation; development framework; machine learning

List of Figures

List of Tables

List of Python Scripts

List of Acronyms

AI	Artificial Intelligence
ML	Machine Learning
COMPAS	Correctional Offender Management Profiling for Alternative Sanctions
FID	Fairness in Design
AIF360	AI Fairness 360 IBM Fairness Toolkit
w.r.t	With relation to
OECD.AI	Organization for Economic Co-operation and Development for Artificial Intelligence
AIF360	AI Fairness 360 IBM Toolkit
ANOVA	Analysis of Variance
LIME	Local Interpretable Model-Agnostic Explanations
SHAP	Shapley Additive exPlanations
EDA	Exploratory Data Analysis
DAG	Directed Acyclic Graph
UI	User Interface
MLP	Multi Layer Perceptron
SVM	Support Vector Machine
PPV	Positive Predictive Value
FDR	False Discovery Rate
FOR	False Omission Rate
NPV	Negative Predictive Value
LFR	Learning Fair Representations
RDI	Removing Disparate Impact
FTU	Fairness Through Unawareness
ROpC	Reject Option Classification
EOpp	Equal Opportunity
Eq.Odds	Equalized Odds

XAI eXplainable Artificial Intelligence

CI Confidence Intervals

HITL Human-in-the-loop

Contents

Chapter 1

Introduction

We live in an era of unprecedented societal transformation, where technology is being harnessed for automation, efficiency gains, and overcoming geographical barriers. A significant part of these transformations is the rise of applications that use Artificial Intelligence (AI) in automated systems to assist decision-making processes. The ongoing development and deployment of algorithmic tools are reshaping how we live our lives. Social networking services, recommendation systems, location-based services, predictive policing, credit *bureau*, and employment platforms have all become automated integral parts of our daily routines (O'Neil, 2016). However, the emergence of each new automation without adequate ethical and legal considerations has led to a proliferation of unfairness and discrimination on a large scale, disproportionately affecting marginalized and minority groups (O'Neil, 2016; Meijer and Wessels, 2019; Raji et al., 2020a; Buolamwini and Gebru, 2018; Otterbacher et al., 2017), which underscores the need for responsible AI development and deployment.

The disparities between the development and implementation of algorithms and our comprehension of their ethical consequences can lead to outcomes that impact individuals, groups, and entire societies (O'Neil, 2016; Benjamin, 2019). Computer scientists and engineers have an essential role in this process. On the other hand, political, ethical, and social scientists and other professionals in this field are closer to studies involving the direct impact of these systems on our daily routines (Mittelstadt et al., 2016). However, detecting the impact of human subjectivity on algorithm design and configuration typically involves examining extended and multiple user scenarios in the development processes (Mittelstadt et al., 2016), many of them causing harm in different ways (O'Neil, 2016).

When it comes to AI harm, bias usually plays a central role, often resulting in unfair consequences and perpetuating inequalities (Crawford, 2013). Hence, bias can emerge at different stages of the AI development process, aggregated into training datasets, and complex feedback loops can arise when a learning model is deployed into the real world (Mahoney et al., 2020). Bias from the world is transferred to the data – whether as historical or temporal bias or through the omission or excess of certain information – and from the data, this bias is passed to Machine Learning (ML) models – as an algorithm bias – aggregated and learned by the system. It is like a cycle; since bias is intrinsic to human beings, it can be perpetuated in other phases of technology systems that involve automation. This requires attention from both creators and

those who distribute technology directly to society. Furthermore, when undergoing a human review, decision-making and actions can again reflect the social, human, and behavioral biases that may lead to discrimination (Selbst et al., 2019).

Extensive evidence has shown how AI can incorporate human and social biases (Mitchell et al., 2020), deploying them on a large scale and consequently displaying discrimination based on discriminatory features (e.g., race, gender, religion, age) against some demographic group (Crawford, 2013; O'Neil, 2016). For example, even in medical systems, where skin tone analysis is required by dermatologists or in facial recognition systems that analyze gender, these ML systems have often proven to be biased and unfair (Buolamwini and Gebru, 2018; Otterbacher et al., 2017). Overall, discriminatory behavior perpetuated by AI systems has been creating negative impacts on people's lives in many ways, such as benefit determinations, marketing strategies, credit scoring, and predictive policing (Crawford, 2013). For example, some software used to predict future criminals has been shown to exhibit significant bias against Black people, and studies by ProPublica[1] provide evidence of how biased and unfair decision-making perpetuate discrimination against marginalized and underrepresented groups. In general, evidentiary software used in law enforcement (Abebe et al., 2022), such as a facial recognition system, carries a lot of issues regarding algorithmic fairness and also individual privacy (Xiang, 2022). Brazilian researchers in the social and political sciences have also discussed how facial recognition tools have led to more incarceration in Brazil, leading mainly to ethnic-racial discrimination (Nunes, 2023). We also have cases that show algorithmic gender biases in the Brazilian court decisions (Benatti et al., 2024) and studies in financial analysis showing that credit scoring models used in Brazil perpetuate discrimination based on location information resulting in racial bias (Vilarino and Vicente, 2020). Regarding these issues, the algorithmic fairness community has advocated for participatory design methods involving stakeholders who use or are affected by technology in its design to build greater trust between users and technology creators (Raji et al., 2020b; Mitchell et al., 2019; Buolamwini and Gebru, 2018).

Concerns and problems about biased and unfair behaviors have led to advancements in both academia and industry, including the development of new tools, the establishment of best development practices, and methods to design models focused on discrimination awareness, fairness, transparency, and accountability (FAccT, 2018). Addressing potential harms, the ethical community has proposed evaluation guides (Bergman et al., 2023), assessment tools (Bellamy et al., 2019; Saleiro et al., 2018), and frameworks to assist decision-making across various fields such as healthcare (Ueda et al., 2024), law (Abebe et al., 2022), finance (Zhang and Zhou, 2019), and others (Mitchell et al., 2019).

In summary, it is widely recognized that automated systems can amplify biases presented in the data, algorithm inaccuracies, and existing social biases, resulting in biased decisions. When it comes to measuring and mitigating bias, the options are extensive, as assessing fairness usually does not come with grounded guidelines that cover all possible sources of bias. As we

[1] https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

will explore later in this work, there is still no consensus on the definitions of fairness in ML; similarly, a broad range of mitigation strategies and metrics depend on how fairness is defined.

1.1 Motivation

The literature on fairness proposes a range of notions that address various AI fairness aspects, such as: definitions focused on preventing differential treatment of individuals considered similar with respect to a specific task (Dwork et al., 2012; Yurochkin et al., 2020); definitions based on the parity of statistical metrics across groups differentiated by sensitive[2] attributes (e.g., male and female individuals, people from different ethnic groups) (Zemel et al., 2013; Kamishima et al., 2011); and definitions emphasizing the need to identify and employ causality among variables to disentangle unfair impacts on decisions (Kusner et al., 2017; Kilbertus et al., 2017), among others (Castelnovo et al., 2022). Although these notions appear internally consistent, some cannot be satisfied simultaneously, leading to mutual incompatible, as demonstrated by researchers: Barocas et al. (2023); Kleinberg et al. (2017); Mitchell et al. (2018); Friedler et al. (2021). Consequently, practitioners assessing and/or implementing fairness must choose among them.

This extensive list of definitions highlights the complexity of fairness, which can be task— or scenario-dependent, multifaceted, context-based, and involve various meanings and nuances. Some definitions focus on different intuitive notions of "unfair decisions," primarily considering the impact on individuals sharing certain personal characteristics, such as gender, age, ethnicity, and religious orientation.

Meanwhile, in terms of bias mitigation, fairness metrics can be improved through interventions[3] at different stages of machine learning model development: modifying the training data, the learning algorithm, or the model's predictions (pre-, in-, and post-processing interventions) (Mehrabi et al., 2021). However, the lack of effective methodologies to identify fairness-related issues and design approaches aligned with both social and mathematical definitions remains a challenge. Assessing these solutions is further complicated by diverse interpretations of bias and measures incompatibility (Barocas et al., 2023; Ferrara, 2024). Moreover, varied deployment scenarios complicate evaluation, as each context may require different metrics and considerations (Kleinberg et al., 2017; Binns, 2018; Kordzadeh and Ghasemaghaei, 2022). These obstacles make it difficult to establish universally applicable evaluation criteria.

Researchers are continually evolving approaches to address the nuances of fairness in

[2]In this research, we shall use the terms "protected" and "sensitive" interchangeably to refer to attributes considered in fairness and discrimination issues.

[3]By intervention we mean a method, technique or approach that will modify an existent state.

different contexts. Existing tools, such as Fairlearn (Bird et al., 2020), IBM AIF360 (Bellamy et al., 2019), and Aequitas (Saleiro et al., 2018), help developers identify and implement bias mitigation interventions in their systems. These tools offer a range of state-of-the-art algorithms and methods for assessing bias in ML models[4]. Despite the variety of options available, these tools do not provide a definitive guideline for developing, implementing, and designing solutions as pointed out by recent surveys (Jones et al., 2020; Pagano et al., 2023; Ferrara, 2024). Consequently, developers often face challenges in fully understanding the social implications, fairness concepts, limitations associated with bias mitigation, and model evaluation across various contexts (Zhang et al., 2023).

Considering the challenges of measuring fairness, choosing a clear definition of fairness at the machine learning modeling stage is important to determine if the model's results treat different groups equitably. This involves evaluating whether the outcomes are fair and balanced across various demographic groups or categories. In this aspect, this work will be developed using three fairness characterization at the modeling stage, introduced by Barocas et al. (2023) and statistically formalized as Independence, Separation, and Sufficiency. These criteria provide a framework for metrics related to fairness in machine learning based on statistical concepts.

Given these formalities and the increased need for understanding fairness evaluation from a development perspective, this research aims to explore these challenges by addressing a variety of fairness interventions under statistical concepts. Standardizing mathematical criteria across all existent methods in the literature is challenging and not feasible to address all existent scenarios (Morley et al., 2023). Instead, we proposed in this work a theoretical guidance in the development process highlighting the differences between specific implementations that fall under one of the fairness criteria. We aim to demonstrate that choosing fairness criteria a priori may lead to a loss of information on fair performance costs, as other possibilities for mitigating bias might achieve similar goals at different costs associated with distinct criteria in the design. This research addresses the problem of developing and applying bias mitigation methods and the nuances of using appropriate metrics to measure an existent mathematical definition.

1.2 Objectives and Contributions

The primary objective of this research is to propose a novel framework focused on interventions for mitigating bias. The framework will compare different fairness perspectives and analyze the trade-offs between three statistical definitions (Independence, Separation, and Sufficiency).

[4]For a more comprehensive list of these tools, see https://oecd.ai/en/catalogue/tools

To accomplish the primary objective of this work, we establish the following specific objectives:

- Analyze existing methods of bias interventions, including pre-and post-processing techniques based on three statistical fairness criteria.
- Explore how understanding statistical fairness definitions enables more appropriate choices of evaluation and intervention methods for bias mitigation in binary classification tasks.
- Implement and evaluate existing bias mitigation interventions that cover the three statistical definitions to the same classification task for comparative analysis.

Through this research, we anticipate making the following contributions that accomplish the defined objectives: (1) An analysis of the trade-offs of different statistical fairness definitions in binary classification contexts. We discuss the formalization of the statistical characterization and their equivalent fairness notions from the literature, highlighting some works and interventions that fall under each criterion. (2) An analysis of existing assessment tools used for ethical model development, including their limitations and usage scenarios. This analysis assists in the design and implementation of our proposed framework solution. (3) An open-source version of the proposed framework designed to be available for use by the AI engineering or practitioners community. This implementation accomplishes our primary objective, providing a way to compare different fairness perspectives from the same classification scenario. (4) An application of the framework in a case study using a task scenario over three fairness perspectives.

We mainly contribute with a framework that can enhance the state-of-the-art fairness development process by providing a structure for bias mitigation and evaluation based on statistical concepts, comparing various fairness definitions, avoiding "one-size-fits-all" solutions, but limited to group fairness approaches. This will enable the use of a development approach that has multiple perspectives of the context to contemplate various audiences and move forward with applying a unique fairness metric or bias mitigation technique to every model or dataset, regardless of the particular biases, goals, or stakeholders involved.

1.3 Outline

The following chapters are organized as follows: Chapter 2 analyzes related works, providing the necessary background to complement the motivation of this research. This includes an overview of fairness in AI, the societal consequences of AI model unfairness, the development process and challenges of applying AI fairness in practice, a summary of fairness assessment tools relevant to this study, and bias mitigation approaches applied to different stages

of ML development. Chapter 3 introduces key aspects of statistical-based fairness in machine learning, as outlined by Barocas et al. (2023), which serves as the foundation for many existing metrics in the literature and is also the foundation work behind the framework developed in this work. Chapter 4 details the methodology used in this study, which consists of (a) the framework design concept, (b) our developmental solution based on the proposed framework, (c) the experimental tools used to achieve our solution, and (c) a case study used as proof of concept, to illustrate the use of our methodology. Finally, Chapter 5 presents the dataset description, our experiment results, and discussion, followed by Chapter 6, which concludes with a summary of findings, limitations, and future research directions.

1.4 Ethics and Reproducibility

The research presented in this document does not use personal or confidential data or data from private domains. The code implementations described here use only open-source tools, and the data used are freely accessible. These data were collected via Python through the Folktables toolkit[5], which is an open platform. Additionally, there is a datasheet that covers both the prediction tasks and the underlying US Census data sources used in this research[6].

The results presented here can be reproduced using the codes available in the author's public repository and following the installation instructions for the requirements listed there[7].

[5] https://github.com/socialfoundations/folktables
[6] https://github.com/socialfoundations/folktables/blob/main/datasheet.md
[7] https://github.com/equity-ai-hub/ai-system-framework

Chapter 2

Related Work

The literature on fairness is extensive, covering evaluation metrics, algorithm bias mitigation, discrimination discovery, the cost of fairness, and ethical concepts applied to machine learning (Mahoney et al., 2020; Zliobaite, 2015; Hajian et al., 2016; Corbett-Davies et al., 2017). Other works focus on auditing, scrutinizing, proposing new methods, and raising awareness about the various problems that unfair AI systems can cause (Xiang, 2022; Abebe et al., 2022; Buolamwini and Gebru, 2018; Raji et al., 2020b). Additionally, several studies have defined multiple notions of fairness (Dwork et al., 2012; Hardt et al., 2016; Kusner et al., 2017). Although these different notions of algorithmic fairness are internally consistent, many cannot be satisfied simultaneously and are thus mutually incompatible (Barocas et al., 2023; Kleinberg et al., 2017; Mitchell et al., 2018; Chouldechova, 2017). Consequently, practitioners assessing and implementing fairness must choose among these notions. These works provide a perspective into the problems associated with data and models and highlight key areas of concern when automated systems are deployed in society, such as privacy risks, perpetuating bias, and other forms of harm.

First, we start the chapter by briefly defining some terminologies commonly used in the fairness field in machine learning and used in this work to enrich the background and knowledge about the topic. Secondly, the following sections describe the main works used as the foundation for this research. This review is not an exhaustive literature survey but focuses on publications from the past ten years. We have categorized the references into the following areas: (a) foundational fairness notions and bias mitigation techniques, with the caveat that some fundamental works fall outside this date range; (b) research, tech reports, and news highlighting instances of unfairness in society due to classification issues; (c) aspects of fairness development in practice; (d) emerging assessment tools and (e) bias mitigation approaches applied to different stages of ML development. Table 2.1 summarizes the selected related works categorized by topic.

Research Work	Foundational Fairness*	Bias Mitigation & Assessment Tools**	Fairness Surveys	Cases of Unfairness in Society
Kamiran and Calders (2011)	✔	✔		
Dwork et al. (2012)	✔	✔		
Feldman et al. (2015)	✔	✔		
Hardt et al. (2016)	✔	✔		
Kleinberg et al. (2017)	✔			
Chouldechova (2017)	✔	✔		✔
Corbett-Davies et al. (2017)	✔			
Pleiss et al. (2017)	✔	✔		
Friedler et al. (2021)	✔			
Barocas et al. (2023)	✔			
Corbett-Davies et al. (2023)	✔		✔	
Buolamwini and Gebru (2018)		✔		✔
Saleiro et al. (2018)		✔		
Bellamy et al. (2019)		✔		
Bird et al. (2020)		✔		
Raji et al. (2020a)		✔		✔
Lee and Singh (2021)		✔		
Tubella et al. (2022)		✔		
Beiró and Kalimeri (2022)		✔		✔
Morley et al. (2023)		✔		
Zhang et al. (2023)		✔		
Mitchell et al. (2018)			✔	
Verma and Rubin (2018)			✔	
Mehrabi et al. (2021)			✔	
Castelnovo et al. (2022)			✔	
Carey and Wu (2022)			✔	
Alves et al. (2023)			✔	
Pagano et al. (2023)			✔	
Caton and Haas (2024)			✔	
Obermeyer et al. (2019)				✔
Vilarino and Vicente (2020)				✔
Abid et al. (2021)				✔
Berk et al. (2021)				✔
Abebe et al. (2022)				✔
Xiang (2022)				✔
Nunes (2023)				✔
Benatti et al. (2024)				✔
Bitencourt and Ansel (2024)				✔

***Works that made significant contributions to the field, such as introducing fairness notions.**
****Works that introduce mitigation approaches through new methods, frameworks, or assessment tools.**

Table 2.1: Summary of selected related works.

2.1 Terminology

In this section, we briefly provide some key definitions related to fairness in machine learning, usually referred to in this work:

Bias: Bias can assume multiple meanings; in this work, we refer to bias as systematic

errors that result in unfair outcomes, often discriminating or disadvantaging certain individuals or groups.

Protected or Sensitive attribute: Divides a population into groups that should receive equal benefits, e.g., race, gender, and religion, among others. These attributes are context-specific rather than universal. In this work, we shall use both terms interchangeably.

Group Fairness: Seeks to ensure that members of protected groups, on average, receive the same treatment as the general population. It is often defined by the equality of certain statistical measures across different groups. As a result, group fairness prioritizes fair treatment for groups rather than focusing on individuals.

Individual fairness: Individual fairness is based on the idea that "similar individuals should be treated similarly." This concept emphasizes the comparison between individuals, ensuring that those who are alike receive equal or comparable outcomes.

Bias Mitigation algorithms or Bias Intervention: A bias mitigation algorithm is produced aimed to improve fairness metrics by modifying the training data, the learning algorithm, or the predictions, reducing unwanted bias. This work also refers to this terminology as bias mitigation: techniques, methods, or interventions.

Fairness Metrics: Quantifies unwanted bias in our models but can also be used to quantify bias in training data.

Bias mitigation usually refers to a process addressing some specific aspect of a machine learning pipeline. The techniques in the literature are categorized into three strategies:

Pre-processing methods: Approach based on removing potential biases directly from the training dataset. Then, a selected classifier model can be learned (i.e., trained) in the treated dataset.

In-processing methods: Strategy that works by adding penalties or even constraints as a way to optimize the task problem, imposing fairness at training time, and enforcing a model to produce fair outcomes.

Post-processing methods: This approach focuses on mitigating bias and achieving more fair outcomes of an already trained model. Usually, a set of unseen data is used to improve the existing model predictions.

2.2 Fairness Notions in Artificial intelligence

A plethora of definitions of fairness in machine learning have rapidly emerged in the literature over the past decade. The work of Hardt et al. (2016) defines statistical fairness across groups by presenting the concepts of *equalized odds* and *equal opportunity*. These definitions propose fairness measures aligned with supervised ML to build more accurate classifiers and

suggest efficient post-processing intervention to improve predictions, ensuring they directly capture the target while remaining independent of sensitive attributes. This definition is also covered in the book by Barocas et al. (2023), which serves as a foundation text in the fairness literature.

Barocas et al. (2023) explores the characteristics that elevate automated decision-making to a significant ethical concern. It places machine learning within a broader critical context, addressing the dangers of bureaucratic decision-making and the rigid application of formalized rules. This work explores different notions of fairness to three main mutually exclusive criteria: independence, separation, and sufficiency. Each criterion aligns with different moral intuitions and sets the stage for a wide range of fairness relaxations and equivalences.

Dwork et al. (2012) made a significant contribution to the field by proposing a framework for fair classification. This framework includes a task-specific metric for assessing the similarity of individuals concerning the classification task, as well as an algorithm for maximizing utility while ensuring that similar individuals are treated similarly. The work also explores the relationship between fairness and privacy, particularly when fairness implies privacy, and discusses how tools developed in differential privacy can be applied to fairness. Dwork et al. (2012) addresses key elements, such as the connection between individual and group fairness, and concludes that the chosen definition of fairness is a generalization of differential privacy[1]

Given the rise of various notions of non-discrimination fairness, several researchers have pointed out conflicts among these notions. Works by Kleinberg et al. (2017); Friedler et al. (2021) and Chouldechova (2017) have mathematically demonstrated that some statistical definitions are mutually incompatible in specific setups. The work of Kleinberg et al. (2017) is another foundation piece in the literature that discusses the trade-offs between different fairness aspects, focusing on calibration conditions and the balance between positive and negative class conditions in binary classification tasks. The main result of Kleinberg et al. (2017) shows that these conditions are generally incompatible; they can only be satisfied simultaneously in highly constrained scenarios. Moreover, this incompatibility also applies to approximate fairness notions — those that are relaxations of the formal statistical definitions of independence, separations, and sufficiency. Calibration holds for sufficient equivalence, while balance classes refer to separation and independence formalizations.

Additionally, Chouldechova (2017) discusses several fairness notions by applying them to assess the fairness of recidivism prediction instruments. The study highlights the concept of *disparate impact*[2] through a case study involving the COMPAS (Correctional Offender Management Profiling for Alternative Sanctions). The work explores various notions of fairness in this context, discussing the trade-offs between fairness and accuracy, as well as the inherent incompatibilities among the notions.

[1]Differential privacy is out of the scope of this research.

[2]Disparate impact occurs when members of a marginalized class are negatively affected more than others due to the use of a formally neutral rule or policy, leading to unintentional or indirect discrimination.

Likewise, Friedler et al. (2021) discusses tensions between group and individual fairness[3], also showing the incompatibilities by defining worldwide assumptions for each fairness criterion. In other words, Friedler et al. (2021) argues that any effort to design fair decision-making systems must inherently make assumptions about the observational process and/or construct space. The key assumptions include: (a) "WYSIWYG" (What You See Is What You Get), which assumes that the observational process preserves the relative position of individuals within the constructed space with respect to the task, and (b) "WAE" (We're All Equal) which implies no inherent differences between groups of individuals based on potentially protected attributes, thereby aiming to guarantee statistical parity and other group fairness notions as a metric.

As highlighted in several foundational works, assessing fairness in AI systems poses challenges due to the complexities arising from the diverse definitions of fairness. According to Narayanan (2018), there are at least 21 mathematical definitions of fairness, each attempting to provide a technical understanding of the social and political aspects underlying these characterizations. Achieving fairness often requires interventions, and techniques have been developed to be applied during an AI system's preprocessing, in-processing, and post-processing stages. Key research has established the foundation for the many state-of-the-art interventions in the literature (Kamiran and Calders, 2011; Feldman et al., 2015; Pleiss et al., 2017; Chouldechova, 2017; Hardt et al., 2016).

Kamiran and Calders (2011) explore algorithmic solutions that preprocess data to remove discrimination before a classifier is trained. The authors introduced techniques such as suppressing the sensitive attribute, massaging the dataset by changing class labels, and reweighing or resampling the data to remove discrimination without relabeling instances. Feldman et al. (2015) linked disparate impact to classification accuracy and proposed a test for disparate impact based on how well the protected class can be predicted from other attributes. Pleiss et al. (2017) conducted an empirical study investigating the tension between minimizing error disparity across different population groups while maintaining calibrated probability estimates, ultimately proposing a post-processing algorithm to achieve calibration.

Caton and Haas (2024) also discuss some fairness dilemmas, comprising (a) the balance of the trade-off between fairness and model performance, (b) the quantitative notions of fairness that permit model optimization but yet cannot balance different notions of fairness, (c) the tensions between fairness, situational, ethical, and sociocultural context and policy; (d) the recent advances to the state-of-the-art have increased the skills gap inhibiting "on-the-street" and industry uptake; and (e) the challenge of both advancing the state-of-the-art and addressing real-world data contexts.

Several survey papers showed the ethical aspects of automated decision-making, pointing out the importance of addressing societal issues. These papers have contributed to this work by summarizing state-of-the-art methods that guide our development. Mehrabi et al. (2021);

[3]Individual fairness aims to give similar individuals similar decisions.

Verma and Rubin (2018); Mitchell et al. (2018) explained fairness concepts, delineating between individual and group fairness. Pagano et al. (2023); Jones et al. (2020); Caton and Haas (2024) listed popular datasets and techniques for pre-, in-, and post-processing interventions along with summaries of various metrics. Castelnovo et al. (2022); Alves et al. (2023) discuss the challenges and complexities of fairness definitions. Corbett-Davies et al. (2023) offered a critical review of fairness metrics, categorizing the effects of decisions on disparities and how legally protected characteristics, like race and gender, influence these decisions.

Finally, Carey and Wu (2022); Binns (2018) offer insights into statistical fairness techniques from both social and formal science perspectives, explaining how fair machine learning intersects with Philosophy, Sociology, and Law. These works help practitioners understand the computer science perspective on automated fairness approaches and their societal applications.

2.3 Consequences of Bias and Unfairness of AI Models for Society

As AI becomes more integrated into our daily lives, instances of bias and discrimination within these systems are also on the rise. Several studies have highlighted the importance of addressing fairness, equity, and the societal implications of AI systems. Ferrara (2024) emphasizes the negative societal impacts of AI bias, which can hinder access to essential services and exacerbate existing inequalities. The media has also reported a variety of problems caused by unfair AI, such as Amazon's recruiting engine, which showed biases against women (Dastin, 2018); Stable Diffusion text-to-image AI models amplifying stereotypes about race and gender (Nicoletti and Bass, 2023); and a ProPublica analysis revealing that the COMPAS Recidivism Algorithm, used across the U.S. to predict future criminals was biased against Black people (Larson et al., 2016).

Other researchers have also expressed concerns and presented studies on the prevalent bias in various industries. Obermeyer et al. (2019) discussed racial bias in health algorithms in the U.S., showing how a widely used algorithm, typical of this industry-wide approach and affecting millions of patients, exhibits significant racial bias: At a given risk score, Black patients are found to be considerably sicker than White patients, as evidenced by signs of uncontrolled illnesses. Buolamwini and Gebru (2018) demonstrated disparities in the accuracy of classifying individuals based on skin tone and gender, revealing inequalities within facial analysis systems. Abid et al. (2021) shed light on persistent biases in GPT-3 models, such as the association of Muslims with violence across various applications. These biases have far-reaching consequences beyond the technology sphere, exacerbating existing inequalities and perpetuat-

ing discrimination against marginalized groups.

Benatti et al. (2024) shows the presence of algorithmic gender biases in court decisions through a case study in a Brazilian court. It shows how biased institutional responses to gender-based violence violate international human rights standards as they prevent gender minorities from accessing their rights and hamper their dignity. Vilarino and Vicente (2020) discusses credit *bureau* services, focused on bias identification and models explainability, pointing out how credit scoring can show the racial bias of Brazilian individuals, using location information without re-sourcing to protected attributes. This work draws attention to the necessity of developing a bias tracking framework that does not rely solely on protected attributes, presenting a complex and unique history of racial oppression and disparity in Brazil with strong evidence for the need for considering regional specifics when reflecting on racial issues. In the legal context, Abebe et al. (2022) proposes an audit framework for robust adversarial testing to assess the validity of evidentiary statistical software, addressing issues related to software outputs used for convicting and incarcerating people. In the media piece from Nunes (2023), the author explains how algorithmic racism works and its impact on the field of public safety — bringing to the table the Brazilian scenario, where day by day, facial recognition means more incarceration of black people. The social scientists, Bitencourt and Ansel (2024), discuss the development of modern facial recognition software that continues the rationalization of racist biases that characterize police surveillance in Brazil. In the article, they point out the most recent case resulting from this modernization of biometric criminal identification systems, which occurred in Bahia – a Brazilian state with the largest black population in the country – where the implementation of facial recognition technologies in public security that led to the biased incarceration of black people. Additionally, Xiang (2022) and Raji et al. (2020a) discuss challenges associated with computer vision, particularly the rise of facial recognition technology and its implications. They highlight the constant tension between privacy and fairness in the context of mitigating algorithmic bias. On the other hand, Beiró and Kalimeri (2022) underscores the importance of ensuring fairness in AI systems within the field of social media analytics, providing measures and interventions to address bias in this specific scenario.

In summary, releasing AI systems into society presents many challenges, and the studies listed in this section also motivated this work by focusing on the development aspects.

2.4 Development Aspects of Fairness in Practice

Designing AI systems in practice can be challenging because it depends on many factors. Each context requires different assumptions and considerations. We review several research findings highlighting the challenges of designing AI systems theoretically and in prac-

tice, especially considering the engineering aspects and the use of tools to achieve the goal, i.e., developing a system that considers bias mitigation in construction.

Morley et al. (2023) emphasizes the importance of designing AI products ethically and focuses on the need for practical tools and methods to support AI practitioners in translating ethical principles into design practices. The study revealed that many AI practitioners struggle to justify the extra time and resources required for ethical design and often lack guidance on how to incorporate ethical principles effectively. Morley et al. (2023) emphasizes the need for more practical and implementable AI ethics guidelines, diverse and inclusive approaches to AI ethics, and a cultural shift towards data-driven innovation.

Tubella et al. (2022) shows that the technical choice between in-processing and post-processing is not necessarily value-free and may have ethical implications for those who will be affected by fairness interventions. The study reveals that assessing these technical choices in terms of their ethical consequences can contribute to the design of fair models and related societal discussions. Tubella et al. emphasizes that the choice between in-processing or post-processing is not necessarily tied to the specific definition of fairness aimed to achieve. Both approaches can effectively satisfy the same group fairness criteria. So, the decision to mitigate bias through in-processing or post-processing is usually perceived as a purely technical decision, though it has deeper ethical ramifications.

Zhang et al. (2023) identifies two significant challenges in designing AI solutions considering fairness: (a) the complexity and diversity of fairness notions, which vary in meaning across different contexts and often have statistical incompatibilities. Software development teams usually struggle to understand these different notions and address blind spots during the design process. (b) The diversity of stakeholders and application scenarios, which require consideration of each stakeholder's priorities related to different notions of fairness. These factors create additional challenges for AI solution teams to consider in their design processes. To tackle this issue, Zhang et al. (2023) proposed a Fairness in Design (FID) methodology to promote discussions that bring fairness concerns in AI projects to the surface. The design framework proposed in their work will also contribute to our future work, see discussion in Section 6.2, by providing practical improvements to the solution design presented in this thesis.

Furthermore, various survey papers discuss issues regarding fairness in practice. Pagano et al. (2023) outlines some fairness assessment toolkits designed to address bias and unfairness, noting that while these toolkits propose methods for identifying and mitigating bias in ML models, the responsibility largely falls on the developers. Developers often lack adequate knowledge to tackle fairness issues effectively, making it crucial to have a methodology to guide them through this process of addressing the problem. Mehrabi et al. (2021) underscore challenges and opportunities in fairness research posed by multiple definitions of fairness and the challenge of how a solution designed for one definition might perform under another. Castelnovo et al. (2022) focuses on how the number of fairness metrics introduced in the literature has surged over the past decade. The abundance of metrics can be overwhelming for researchers or prac-

titioners new to the field, as each metric captures different aspects of fairness, and there is no comprehensive guide to help navigate these nuances.

Lee and Singh (2021) identifies a significant gap in the ability to identify risks to unintended and harmful biases in ML. They discuss how the existing frameworks and tools proposed in the literature are often not concrete enough to be fully implemented in practice. Lee and Singh (2021) highlights the need for a more systematic method to identifying and mitigating the risk of unfairness, especially given the challenge of dealing with competing definitions of fairness. Practitioners usually report difficulties explicitly considering biases and 'blind spots' that may arise while developing a machine learning model.

In summary, there are no standardized methods for applying most AI techniques, metrics, and frameworks in practice. This lack of standards becomes particularly problematic when moving beyond small-use case studies to solutions that need to be applied at the industry level. Additionally, most current solutions for addressing unfairness and bias are designed for specific problems or scenarios, and there are a variety of techniques available, often referred to as fairness metrics. However, the multitude of options makes it challenging to select the most appropriate criteria for assessing and addressing the issue. In the next subsection (2.4.1), we describe some well-accepted fairness assessment tools used in this work and highlighted in many of the survey studies mentioned earlier.

2.4.1 Fairness Assessment Tools

Machine learning researchers have developed various assessment tools to identify and measure bias in response to the ongoing examination of AI systems for bias and discrimination. The rapid development of AI technology has made it harder for policymakers, ethical practitioners, regulators, and even AI engineers to ensure that appropriate policies and governance are effectively applied.

For example, *Aequitas* (Saleiro et al., 2018) is a toolkit designed to assess model outcomes by computing various bias and fairness metrics for specific population groups. It provides valuable insights for policymakers to support fair decision-making. The toolkit also includes a decision tree[4] that guides users in selecting appropriate fairness metrics, particularly those relevant to group fairness, which could also be useful in the case study presented in Chapter 5. However, Aequitas has some limitations. Since our research work explores different types of notions of fairness, Aequitas doesn't capture all definitions. As part of our design choice to aggregate fairness interventions based on statistical definitions, we found that Aequitas could potentially lead to misinterpretations. Users might apply metrics inappropriately if they lack

[4] `http://aequitas.dssg.io/static/images/metrictree.png`

a strong understanding of the underlying fairness criteria. Although we explored the Aequitas tool, we did not integrate it into our system design because it does not fully address the trade-offs between different fairness characterizations.

IBM *AIF360* (Bellamy et al., 2019) and Microsoft *Fairlearn* (Bird et al., 2020) are the most popular[5] open-source toolkits for implementing state-of-the-art fairness algorithms and metrics at an industrial scale. Both toolkits provide access to a wide range of techniques from the literature, covering bias mitigation at pre-, in-, and post-processing stages, as well as various statistical fairness metrics. Besides their popularity, these toolkits generally require users to have some knowledge of bias mitigation to select the appropriate techniques for their specific scenario, given the variety of available options. In particular, AIF360 is best suited for narrowly defined settings, such as allocation or risk assessment problems with well-defined protected attributes, where statistical or mathematical equivalence is desired. The toolkit's algorithms[6] and metrics[7] are optimized for binary classification or regression tasks and only support two sensitive attributes per method and metric. Fairlearn shares some of these limitations but offers support for more than two sensitive attributes and includes additional algorithms for mitigating disparities using reduction approaches[8].

AIF360 and Fairlearn were adopted in this work's implementations, facilitating the addition of state-of-the-art fairness techniques within our ML pipeline since they served as functions accessible via Python programming language with useful functionalities to detect and mitigate bias. Although numerous proposed fairness methods have been in the literature over the years, far fewer have been available and ready-to-use as open-source implementations, such as AIF360 and Fairlearn.

Another tool we considered for inclusion in our framework was the pymetrics *Audit-AI* (Bias Testing for Generalized Machine Learning Applications[9]. This tool is designed to measure and mitigate discriminatory patterns in training data and the predictions made by machine learning models, particularly in socially sensitive decision-making processes. *Audit-AI* provides several methods for auditing a trained model for bias, including the 4/5th, Bayes factor, classifier posterior probabilities, z-test for classification tasks, ANOVA, group proportions at different thresholds, and 4/5th for regression tasks. However, we found that *Audit-AI* was not as comprehensive in addressing different fairness notions, as the main focus is on detecting bias that arises when models are trained on biased datasets. Because of that, this tool was not adopted in the research presented.

Some other tools focus solely on explainability or interpretability, such as LIME (Ribeiro et al., 2016) and SHAP (Lundberg and Lee, 2017). These tools help identify the features that

[5] IBM AIF360: 2.4k GitHub Stars and Microsoft Fairlearn: 1.9k GitHub Stars on Aug. 2024

[6] `https://aif360.readthedocs.io/en/stable/modules/algorithms.html`

[7] `https://aif360.readthedocs.io/en/stable/modules/metrics.html`

[8] `https://fairlearn.org/v0.10/api_reference/index.html#module-fairlearn.reductions`

[9] `https://github.com/pymetrics/audit-ai`

most impact a model's outcomes, which can be useful in identifying the sources of bias. While these aspects are orthogonal to the fairness problem, they are beyond the scope of our framework and case study.

A complete list of assessment tools for ensuring safe and trustworthy AI can be found on the OECD.AI[10] platform. This forum brings together countries and stakeholders to shape trustworthy AI, providing a catalog of tools and metrics categorized by technical, educational, and procedural approaches. It also allows filtering by various scopes. For our research, we focus on a smaller set of available toolkits to avoid increasing the complexity of the implementation and minimize the risk of misinterpretation from the wide array of options. Our goal is to assist in understanding the impact of different fairness notions in a more reproducible and scalable way, analyzing the multitude of fairness differences in practice. Given all the tools discussed, we summarize in Table 2.2 what they provide and whether they were adopted in our solution.

Assessment Tool	**Fair Metrics**	**XAI**	**Bias Auditing/Mitigation**	**Adoption**
AIF360 (Bellamy et al., 2019)	✔	✔	✔	✔
Fairlearn (Bird et al., 2020)	✔	✗	✔	✔
Aequitas (Saleiro et al., 2018)	✔	✗	✔	✗
Audit-AI	✔	✗	✔	✗
SHAP (Lundberg and Lee, 2017)	✗	✔	✗	✗
LIME (Ribeiro et al., 2016)	✗	✔	✗	✗

Table 2.2: Summary of fairness assessment tools and if the respective tool was adopted in this work.

2.5 Bias Mitigation

Methods aimed at addressing biases in algorithms can be grouped into three main categories, as highlighted by many works (Alves et al., 2023; Mehrabi et al., 2021; Verma and Rubin, 2018; Castelnovo et al., 2022):

1. **Pre-processing**: These strategies aim to modify the data before it is used in the algorithm, eliminating any underlying discrimination. This approach is applicable when the algorithm can adjust the training data. Then, a selected classifier model can be learned (i.e., trained) in the treated dataset. For example, the Removing Disparate Impact (RDI)

[10] https://oecd.ai/en/catalogue/overview

(Feldman et al., 2015) pre-processing approach uses a "repair level" to indicate how much the user wants the distributions of the groups to overlap, editing the values that will be used as a feature to increase fairness between groups. In the literature, we have an extensive list of preprocessing approaches aiming to achieve certain perspectives of fair predictions. Such as resampling the training data (Kamiran and Calders, 2011), applying suppression of the protected attributes (Dwork et al., 2012), and reweighing Kamiran and Calders (2011).

2. **In-processing:** This category involves altering advanced learning algorithms during the model training phase to reduce discrimination. In-processing can be implemented by adjusting the objective function or adding constraints within the machine learning training process. For example, the Adversarial Debiasing (Zhang et al., 2018) is based on the simultaneous training of two classifiers that will compete to achieve and accomplish a task, corresponding to a generative adversarial network. Depending on the implementation, this method can achieve two different notions of fairness: demographic parity (statistical parity) and equalized odds. (Castelnovo et al., 2022).

3. **Post-processing**: This approach takes place after the model training is complete, using a separate split of the dataset that was not part of the training data. If the algorithm operates as a black box, without the ability to change the training data or the learning algorithm, post-processing is employed to reassign the labels generated by the model based on a defined approach. For example, the Threshold Optimizer (Hardt et al., 2016) works to find an intersection between ROC Curves, minimizing the loss of classification while improving the equalized offs; another example of intervention applied in this stage of an ML pipeline is the Reject Option Classification (ROpC) (Kamiran et al., 2012) works with the probability scores, and the idea of the data instances with high uncertainty can have their labels related to the sensitive attributes switched in order to enforce fairness.

Table 2.3 summarized some of the bias mitigation interventions, their fairness notions related to bias measurement, and the open-source assessment tool that provides a code implementation of the method. This is not an extensive list but a list of possible methods that can also be addressed in our framework implementation. The table covers methods that are related to group fairness approaches since this research is limited to statistical definitions based on group approaches.

Type	Mitigation Method	Fairness Notion	Asses. Tool
Pre	LFR (Zemel et al., 2013)	Statistical parity	[1]
Pre	RDI (Feldman et al., 2015)	Statistical parity	[1]
Pre	Reweighing (Kamiran and Calders, 2011)	Independence, Statistical parity	[1]
Pre	Sampling (Kamiran and Calders, 2011)	Independence, Statistical parity	[1]
Pre	Massaging (Kamiran and Calders, 2011)	Independence, Statistical parity	[1]
In	Reductions (Agarwal et al., 2018)	Equalized odds, Separation,	[1, 2]
In	Adversarial Debiasing (Zhang et al., 2018)	Equalized odds, Statistical Parity	[1]
Post	Information Withholding (Pleiss et al., 2017)	Calibration, Sufficiency	[1]
Post	Threshold Optimizer (Hardt et al., 2016)	Equalized odds, Separation	[1, 2]
Post	ROpC (Kamiran et al., 2012)	Independence	[1]
Post	Calibration w/ Groups (Chouldechova, 2017)	Sufficiency	N/A
Post	Non-parametric Regression Estimators (DiCiccio et al., 2023)	Sufficiency, Predictive Parity	N/A

[1] AIF360 assessment tool, [2] Fairlearn assessment tool

Learning Fair Representations (LFR), *Removing Disparate Impact* (RDI)
Fairness Through Unawareness (FTU), *Reject Option Classification* (ROpC)

Table 2.3: Summary of bias mitigation interventions with the respective fairness notion achieved when using the approach and the most popular open-source assessment tool in which the method is available. [1] denoted the AIF360 assessment tool, and [2] denoted the Fairlearn.

Chapter 3

Statistical Fairness Definitions

Fairness is a complex concept that is scenario-dependent, e.g., being multifaceted and context-dependent. As highlighted by Narayanan (2018), there are at least 21 mathematical definitions of fairness. These numerous criteria represent an attempt to make technical sense of the social and political complexities associated with fairness. According to Narayanan (2018), the technical discussions about these definitions have revealed trade-offs between different mathematical notions of fairness, which deserve attention beyond the technical community.

It is important to recognize that, beyond theoretical differences, various definitions of fairness focus on different aspects, potentially leading to very different outcomes. Researchers have demonstrated that it is impossible to satisfy all the definitions at the same time (Kleinberg et al., 2017; Friedler et al., 2021). This trade-off must also be understood in practice since existing bias mitigation methods may align with specific definitions of fairness, and not all metrics are compatible when we evaluate results considering a different notion.

The majority of works on fairness focus on algorithmic classification tasks, with common measures typically concentrating on three types of attributes: (a) the predicted outcomes, (b) the predicted probabilities and actual outcomes, and (c) the predicted and actual outcomes (Verma and Rubin, 2018). These attributes derive from statistical formalities known as Independence, Separation, and Sufficiency (Barocas et al., 2023; Chouldechova and Roth, 2018). The appeal of statistical measures lies in their relative simplicity and the fact that definitions based on these measures can often be implemented without making assumptions about the distributions of the underlying data. These criteria serve as benchmarks for fairness metrics in machine learning, grounded in statistical concepts. Therefore, these concepts laid the groundwork for our framework structure to assess fairness through bias mitigation approaches. Then, it will be further formalized in the next section to enrich the understatement of these concepts and why they cannot be used simultaneously.

3.1 Statistical non-discrimination Criteria

In Barocas et al. (2023) book, the concept of statistical non-discrimination criteria is explored as a method for defining the absence of discrimination using statistical measures related to random variables that describe decision-making scenarios.

Formally, these criteria are characterized by the joint distribution of several variables: the sensitive attribute A (e.g., race, gender, religion), the target variable Y, the classifier $\hat{Y}$ or probability score R, and, in some cases, features X. By looking at the joint distribution of these random variables, we can clearly decide whether a criterion is satisfied.

Table 3.1 summarizes all symbols and notations used in the following sections. Table 3.2 lists the main statistical metrics, mainly used for binary classification predictions, that are considered as bases for other types of metrics, including independence, separation, and sufficiency, and will be used in the subsequent experiment discussions. At the end of this chapter in Table 3.3 , we also summarize other notions of fairness grouped by the closest statistical definition.

Symbol	Description
A	Random variable, sensitive (or protective) attributes.
X	Continuous or discrete variables, remaining (non-sensitive) attributes.
Y	Target: discrete variable, often binary value. Actual outcome.
$\hat{Y}$	Classification or predictor outcome for binary problems.
R	Probability score (outcome predictor, similar to $\hat{Y}$).
$\mathbb{P}\{\text{event} \mid \text{condition}\}$	Conditional probability.
$U \perp V \mid W$	Random variables U and V are conditionally independent given W.
$U \perp V$	Random variables or events, U and V are statistically independent.

Table 3.1: List of symbols and notations used in this chapter

3.1.1 Independence

Barocas et al. (2023) provides the following definition of statistical independence, which requires the sensitive characteristic to be statistically independent of the score:

Definition 3.1 (Independence)**.** Random variables (A, R) satisfy independence if $A \perp R$.

This concept has been explored through many equivalent and related definitions. It requires that the sensitive attribute, A (which can be categorized as privileged or unprivileged),

Statistical Metrics	Abbr	Formal Definition	Description
True positive	TP	When both predicted and actual outcomes are in a positive class.	$Y = 1 \cap \hat{Y} = 1$
True negative	TN	When both predicted and actual outcomes are in the negative class.	$Y = 0 \cap \hat{Y} = 0$
False positive	FP	When the predicted outcome is positive, but the actual outcome is negative.	$Y = 0 \cap \hat{Y} = 1$
False negative	FN	When the predicted outcome is negative, but the actual outcome is positive.	$Y = 1 \cap \hat{Y} = 0$
Positive predictive value/ (Precision)	PPV	Fraction of positive cases correctly predicted out of all predicted positive cases.	$\frac{TP}{TP+FP}$
False discovery rate	FDR	Fraction of negative cases incorrectly predicted as positive out of all predicted positive cases.	$\frac{FP}{TP+FP}$
False omission rate	FOR	Fraction of positive cases incorrectly predicted as negative out of all predicted negative cases	$\frac{FN}{TN+FN}$
Negative predictive value	NPV	Fraction of negative cases correctly predicted out of all predicted negative cases	$\frac{TN}{TN+FN}$
True positive rate/ (Sensitivity/Recall)	TPR	Fraction of positive cases correctly predicted out of all actual positive cases	$\frac{TP}{TP+FN}$
False positive rate	FPR	Fraction of negative cases incorrectly predicted as positive out of all actual negative cases	$\frac{FP}{TN+FP}$

Y = true value, $\hat{Y}$ = predicted value, 1 = positive class, 0 = negative class

Table 3.2: List of statistical metrics used for evaluating binary classification performance.

must be statistically independent of the classification outcome $\hat{Y}$, e.g., $\hat{Y} \perp A$.

In the context of binary classification, the definition of independence can also be expressed as:

$$\mathbb{P}\{\hat{Y} = 1|A = a\} = \mathbb{P}\{\hat{Y} = 1|A = b\}, \forall a, b \in A. \tag{3.1}$$

If we interpret the event $\hat{Y} = 1$ as a positive outcome, this definition implies that the acceptance rates across all groups must be equal. It requires that the prediction is statistically independent of the sensitive feature, e.g., $\hat{Y} \perp A$. In other words, the predicted acceptance rates for protected and unprotected groups should be equal.

This statistical fairness definition is commonly used in research, often arguing that it reflects an assumption of equality: all groups have an equal claim to acceptance, and resources should therefore be allocated proportionally (Barocas et al., 2023). A distinguishing feature of independence criteria is that they rely only on the distribution of features and decisions, unlike separation and sufficiency criteria, which also consider error rate parities and thus incorporate the target variable (Castelnovo et al., 2022).

For example, when a model is used to predict loan approval, the probability of approval should not be affected by sensitive attributes such as race or gender. Approval rates should be

based only on pertinent non-sensitive features, i.e., financial characteristics, rather than demographic attributes.

In binary classifiers, the concept of independence is often referred to as demographic parity (Kusner et al., 2017), statistical parity (Dwork et al., 2012), group fairness (Dwork et al., 2012), and equal acceptance rate (Zliobaite, 2015). Statistical parity is one of the most commonly accepted definitions of fairness that relies on the principle of independence.

3.1.1.1 Limitations of Independence

According to Barocas et al. (2023), while meeting the independence criterion ensures that acceptance rates are equal across groups, it does not guarantee fairness in all aspects. For example, consider a company that selects applicants from group a with a certain rate while applicants from group b at the same rate. Even though both groups have the same acceptance rates, the quality of selected applicants may differ significantly between the two groups. This discrepancy can lead to a false impression that members of the group b are less qualified or perform worse than members of the group a, potentially resulting in a negative reputation for the group b.

3.1.2 Separation

In a typical classification problem, there is a difference between correctly accepting a positive instance and mistakenly accepting a negative one. The target variable Y defines a way to partition the population into groups (strata) with equal claim to acceptance. However, a particular demographic group $(A = a)$ might be over- or under-represented in these different strata defined by the target variable. A decision maker might argue that in such cases, it is justified to accept more or fewer individuals from group a based on their representation.

Definition 3.2 (Separation)**.** Random variables (R, A, Y) satisfy separation if $R \perp A|Y$.

This criterion captures the notion that a sensitive attribute might correlate with the target variable in different scenarios (Barocas et al., 2023). In the context of a binary classifier, separation implies the following conditions for all groups a, b:

$$\mathbb{P}\{\hat{Y} = 1 \mid Y = 1, A = a\} = \mathbb{P}\{\hat{Y} = 1 \mid Y = 1, A = b\} \tag{3.2}$$

$$\mathbb{P}\{\hat{Y} = 1 \mid Y = 0, A = a\} = \mathbb{P}\{\hat{Y} = 1 \mid Y = 0, A = b\} \tag{3.3}$$

Separation requires that all groups have the same false negative and false positive rates, ensuring *error rate parity*.

To illustrate, consider a model used to predict whether an individual will default on a loan. Separation ensures that the false positive rate (incorrectly predicting that someone will default) and the false negative rate (incorrectly predicting that someone will not default) are consistent across different demographic groups (e.g., gender, race). This criterion ensures that errors made by the model do not disproportionately disadvantage or advantage any particular group.

In the literature, this criterion is also referred to as equalized odds (Hardt et al., 2016), alongside a relaxation to equal false negative rates, called equality of opportunity (Hardt et al., 2016), disparate mistreatment (Zafar et al., 2017), and error rate balance (Chouldechova, 2017). These variations consider both the predicted and the actual outcomes, with some serving as relaxations of the primary definition.

3.1.2.1 Separation pitfalls

Barocas et al. (2023) suggests that error rate parity can be simplified, which leads to a "natural relaxation." For example, instead of requiring equal error rates across all groups, we might only require equal false negative rates. A false negative represents a missed opportunity in situations where acceptance (positive case) is desirable, like not hiring a qualified person. However, in situations like predicting a loan default, where a positive case indicates a default, the roles of false positives and negatives are reversed. This reversal can cause confusion in terminology and interpretation.

3.1.3 Sufficiency

Sufficiency formalizes the fact that a model's predictions already incorporate the necessary information from sensitive characteristics to predict the target variable.

Definition 3.3 (Sufficiency)**.** Random variables (R, A, Y) satisfy sufficiency if

$$Y \perp A \mid R$$

According to Barocas et al. (2023), in a binary classification context, a score R is sufficient for a sensitive attribute A if and only if for all groups a, b, and for all values r that R can take, we have:

$$\mathbb{P}\{Y = 1 \mid R = r, A = a\} = \mathbb{P}\{Y = 1 \mid R = r, A = b\} \tag{3.4}$$

If we replace R by a binary predictor $\hat{Y}$, sufficiency requires that the positive and negative predictive values are equal across all groups. Unlike separation (equalized odds), sufficiency, often referred to as *conditional use accuracy equality*, focuses on ensuring parity in the predicted outcomes rather than the actual outcomes.

To illustrate this notion, using a loan approval scenario, sufficiency ensures that if a model predicts a 75% change in repayment, the actual repayment rate should be similar across all demographic groups, regardless of sensitive attributes like race or gender.

Sufficiency is closely related to the concept of *calibration* (Pleiss et al., 2017), where calibration by group implies sufficiency (Barocas et al., 2023). Other related notions include predictive parity (Chouldechova, 2017), overall accuracy equality (Berk et al., 2021), test fairness (Chouldechova, 2017) and well calibration (Kleinberg et al., 2017).

3.1.3.1 Sufficiency and Calibration

According to Barocas et al. (2023), in some applications, it is desirable to interpret score functions as probabilities. Calibration allows us to move in this direction. Sufficiency is often satisfied by the outcome of unconstrained supervised learning without requiring explicit intervention. In other words, sufficiency frequently occurs "for free" (at least approximately) due to standard ML practices. However, imposing sufficiency as a constraint on a classification system may not significantly affect current practices (Barocas et al., 2023).

3.2 Incompatibilities Between Fairness Notions

We have already seen that each fairness notion – independence, separation, and sufficiency – highlights a specific aspect of fairness. This raises the question: why not use all

definitions simultaneously to achieve all fairness conditions? The "Impossibility Theorem" – which is considered foundational in algorithmic fairness literature – stated by Kleinberg et al. (2017), demonstrates that these fairness notions are, generally, not compatible when fitting statistical models. Compatibility is only possible in trivial or degenerate scenarios, such as when (1) the algorithm is a perfect predictor or (2) there is no prevalence difference between groups. These impossibility cases introduced by Kleinberg et al. (2017) have been confirmed by similar findings in the works of Chouldechova (2017) and Friedler et al. (2021).

Below is a summary of the key incompatibilities and relationships between these fairness notions, as drawn from the literature (Chouldechova, 2017; Kleinberg et al., 2017; Barocas et al., 2023; Srivastava et al., 2019; Castelnovo et al., 2022; Friedler et al., 2021).

1. **Separation and Independence:** If Y is binary and independent of both the sensitive attribute A and the score R, then it is impossible to achieve both separation and independence. In simpler terms, achieving both requires either a model that is completely ineffective (where Y is independent of R) or a scenario where the target is independent of A, implying equal base rates across sensitive groups. Therefore, if there is an imbalance among groups identified by A, both cannot be achieved simultaneously.

2. **Sufficiency and Independence:** Similarly, if $Y \perp A$, then sufficiency and independence cannot coexist. Thus, in cases of base imbalance among groups identified by A, it is impossible to satisfy both sufficiency and independence. To satisfy sufficiency, the model must ensure that, given a score R, the probability of $Y = 1$ is the same across all groups identified by A. However, to satisfy independence, the score R should not depend on A. These two conditions conflict when there is a difference in the base rates of the target Y between the groups.

3. **Separation and sufficiency:** If Y is independent of A and the distribution A, R, Y is strictly positive, then separation and sufficiency cannot both hold true. This means that separation and sufficiency can only coexist when there is no imbalance among sensitive groups (i.e., the target is independent of sensitive attributes) or if the joint probability of A, R, Y degenerates. For binary targets, this degeneration occurs when certain values of A and R consistently correspond to either $Y = 1$ or $Y = 0$, effectively making the score a perfect predictor. In such cases, a perfect classifier $R = Y$ will satisfy both sufficiency and separation.

Fairness Metric	Description	Formal Definition
Independence		
Demographic Parity	All groups have equal probability of being assigned to the positive class	$P(\hat{Y} = 1\|A = a) = P(\hat{Y} = 1\|A = b)$
Statistical Parity	The proportion of positive outcomes is the same for all groups.	$P(\hat{Y} = 1\|A = a) = P(\hat{Y} = 1\|A = b)$
Conditional Statistical Parity	This definition extends the previous one by permitting a set of legitimate attributes L to affect the outcome.	$P(\hat{Y} = 1\|L = l, A = a) = P(\hat{Y} = 1\|L = l, A = b)$
Separation		
Equal Opportunity	Ensures that individuals in the positive class have an equal chance of being correctly classified, regardless of their group membership.	$P(\hat{Y} = 1\|Y = 1, A = a) = P(\hat{Y} = 1\|Y = 1, A = b)$
Equalized Odds	Ensure that both true positive rates and false positive rates are equal across groups.	$P(\hat{Y} = 1\|Y = y, A = a) = P(\hat{Y} = 1\|Y = y, A = b)$ for $y \in \{0, 1\}$
Treatment Equality	Ensures that the ratio of false positive and false negative rates is the same across groups.	$\frac{P(\hat{Y}=1\|Y=0,A=a)}{P(\hat{Y}=0\|Y=1,A=a)} = \frac{P(\hat{Y}=1\|Y=0,A=b)}{P(\hat{Y}=0\|Y=1,A=b)}$
False Positive Rate (FPR) Disparity	Ensures that the rate of false positives is equal across groups.	$P(\hat{Y} = 1\|Y = 0, A = a) = P(\hat{Y} = 1\|Y = 0, A = b)$
False Negative Rate (FNR) Disparity	Ensures that the rate of false negatives is equal across groups.	$P(\hat{Y} = 0\|Y = 1, A = a) = P(\hat{Y} = 0\|Y = 1, A = b)$
Sufficiency		
Positive predictive value (PPV)	Ensure that the probability of the actual positive outcome given a positive prediction is equal across groups.	$P(Y = 1\|\hat{Y} = 1, A = a) = P(Y = 1\|\hat{Y} = 1, A = b)$
Calibration	Ensure that for any predicted probability score, the actual probability of the positive outcome is the same across groups.	$P(Y = 1\|\hat{Y} = \hat{y}, A = a) = P(Y = 1\|\hat{Y} = \hat{y}, A = b)$ for all $\hat{y}$
Conditional Use Accuracy Equality	Ensure that the accuracy of predictions is equal across groups when conditioned on the predicted positive or negative class.	$P(Y = \hat{Y}\|\hat{Y} = y, A = a) = P(Y = \hat{Y}\|\hat{Y} = y, A = b)$ for $y \in \{0, 1\}$

Conditional Probability = $\mathbb{P}\{\text{event} \mid \text{condition}\}$

Table 3.3: List of some fairness metrics grouped by statistical definitions, with description and formal definitions. This is not a comprehensive list of all closest notation in the literature

Chapter 4

Methodology

This chapter outlines the methodology employed in this work. It involves detailing the design and implementation of the proposed framework and the development of a case study as an experimental method. It concludes with a discussion of the findings related to multiple fairness approaches based on the experimental outcomes. The following sections are included:

1. **Framework Design** (Section 4.1): This section covers the design concept and development solution of the framework, which includes a complete machine learning pipeline: data processing, implementation, and training of binary classification models, model evaluation, group fairness interventions, and results analysis.

2. **Experimental Approach** (Section 4.3): In this section, we covered the tools used to get our concept framework fully implemented, resulting in a tool ready-to-use for different cases of study.

3. **Case Study** (Section 4.4): This section describes the classification task goal of the case study framed as a binary classification problem. This study is used to test our framework proposal.

4.1 Framework Design

The structure of our framework and the design decisions were based on the challenges identified in recent research, which highlights a lack of practical understanding of ethics and fairness among developers (Mehrabi et al., 2021; Pagano et al., 2023; Verma and Rubin, 2018; Carey and Wu, 2022; Morley et al., 2023). In addition, the foundational work by Barocas et al. (2023) served as a key reference in designing our framework concept and our development solution. Since our main objective focuses on the aspects of applying multiple fairness interventions to mitigate bias with an evaluation process, we aligned our design with the three established fairness criteria covered by Barocas et al. (2023) and based on statistical concepts. Figure 4.1 shows all the stages that represent our design choices to apply multiple interventions

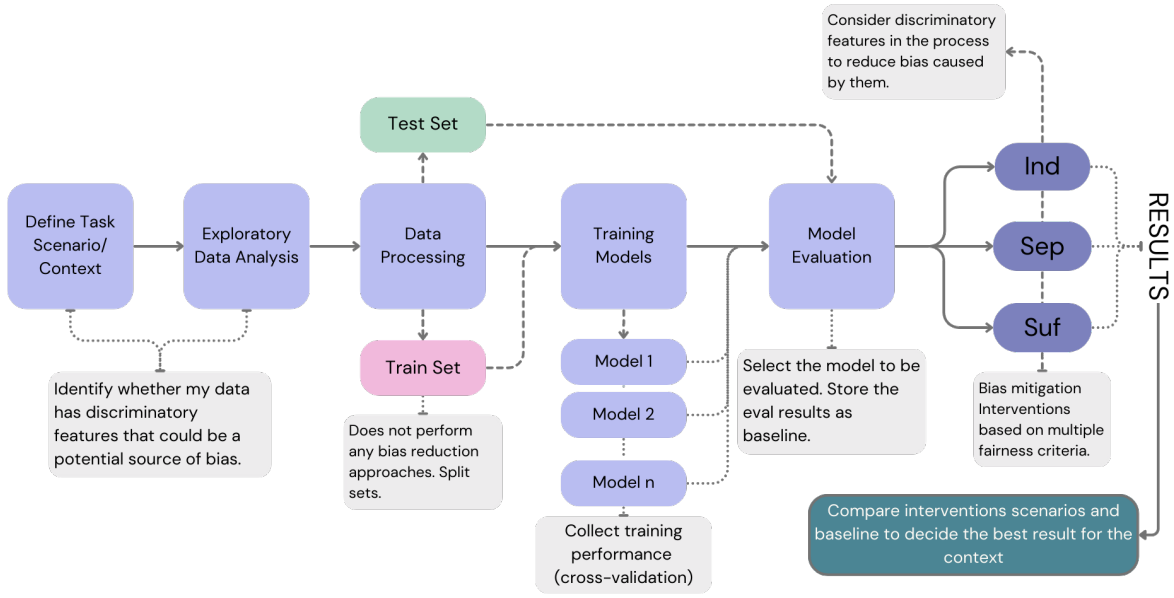


Figure 4.1: User flow through our concept design framework.

for the same context. Some steps of the framework require human intervention and cannot be fully automatized, such as defining the task context and analyzing results. Each node from the image will be described in the following sections with details about the implementation.

4.2 Developmental Framework Solution

We focus on establishing best practices in software engineering and setting standards to govern the operational aspects of ML models through fairness constraints, offering a practical and adaptable solution to our framework concept. To this end, we employ Airflow[1], an open-source workflow orchestration tool, for its user-friendly interface and flexibility. With airflow, we developed our framework using a modular architecture, where individual code blocks – referred to as *nodes* – perform specific tasks within the pipeline. These nodes connect to create an automated and dynamic machine-learning pipeline with an Airflow structure represented as a Directed Acyclic Graph (DAG).

Airflow is suitable for a wide range of orchestrations and requires only proficiency in *Python* programming language. It integrates with various cloud services and databases, operates without extensive computational resources, and features extensible components that can be customized for different settings. Additionally, Airflow is supported by a web-based User

[1] `https://airflow.apache.org`

Interface (UI) that facilitates workflow management.

The entire framework, including its design and pipeline implementation for the experiments, is open-source and publicly available[2]. Implemented using Python, Airflow, and Docker, the framework allows replicating and customizing experiments with other datasets, additional models, and evaluations using the provided code.

4.2.1 Task Scenario Definition & Exploratory Data Analyses (EDA)

Before starting any fairness analysis process, it is important to define which scenario is being evaluated and which stakeholders are involved. Since fairness is context-dependent, at this stage, the user can collect or document what each stakeholder cares most about regarding the context. In addition, being aware of the task goal includes learning about the dataset that will be used. Then, our framework includes an exploratory data analysis to be performed by a human before starting any automation process; this exploration complements the understatement of our task context. Additionally, here, we can identify the protected features that can be the potential source of bias; this is necessary knowledge to apply to the future bias mitigation process.

The Exploratory Data Analysis (EDA) stage, illustrated in Figure 4.1, is non-automated, and we considered part of the developer's or user's responsibility since EDA is challenging to standardize due to its dependence on factors such as data type, format, and classification goal. It involves understanding the classification goals, identifying sensitive attributes, analyzing data and label distribution, and detecting potential biases inherent in the data.

Conducting a fair data analysis involves various techniques, such as identifying which features contribute most to the classification target before initiating the training routine. The specific techniques used will depend on the factors mentioned above. Although EDA is not the primary focus of this framework, and we will not dig into these operational steps, interested readers are directed to the works of Pagano et al. (2023); Chen et al. (2023); Balayn et al. (2021) and Mehrabi et al. (2021) for additional resources on data analysis tools and methodologies.

4.2.2 Data Processing

The concept of data processing is to apply the necessary preprocesses to the data and split it into sets before training any model. This initial process doesn't use bias reduction.

[2] https://github.com/equity-ai-hub/ai-system-framework

In terms of the implemented solution, this node represents the initial automated steps in the pipeline. To process data, we start by doing a data loading (or downloading), which is straightforward; the dataset can be loaded as a CSV format, Pandas DataFrame[3], or through existing APIs. As part of our design framework, we process data after downloading the raw data. The data processing can be tailored to fit the specific dataset and user needs. Our standard process comprises data splitting into designated proportions for training and testing, removing any existing duplicates and entries with missing values, applying one-hot encoding to categorical features, standardizing continuous features, and saving the processed data for subsequent steps. This standard preprocessing is available for any type of dataset and task, but additional procedures in the data can be added for a user if needed.

4.2.3 Training Binary Classification Models

The processed data is ingested in this node to train various classification models. Connected to the *Data Processing* step, our framework endows concurrent training of multiple models. For our implementation, we decide to provide some standard supervised learning models as options in our solution: Logistic Regression, Multi-layer Perceptron (MLP), Random Forest, Support Vector Machine (SVM), XGBoost, and Decision tree classifiers. Listing 4.1 shows the selected models, which primarily use default parameters. The framework, as a tool, is designed to be flexible, allowing for easy expansion and modification to include additional models, adapt to other classifications, and change the model's hyperparameters.

The following Listing 4.2 provides a skeleton implementation of the Training Classification Models block, as shown in Figure 4.1, to run all the models concurrently. We highlight our implementing structure to present how our proposal is simple to understand from a technical perspective and how we ensure the direction of the arrows in our concept diagram is followed in practice, maintaining the required structure of our pipeline.

StratifiedKFold[4] cross-validation was employed during model training to evaluate performance. Metrics collected during training were implemented using the open-source AIF360[5] tool (Bellamy et al., 2019).

List of Metrics: True Positives (TP), False Positives (FP), False Negatives (FN), True Negatives (TN), True Positive Rate (TPR), True Negative Rate (TNR), False Positive Rate (FPR), False Negative Rate (FNR), Positive Predictive Value (PPV), Negative Predictive Value (NPV), False Discovery Rate (FDR), False Omission Rate (FOR), Accuracy (ACC), Balanced

[3] `https://pandas.pydata.org/docs/user_guide`
[4] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.StratifiedKFold.html
[5] `http://aif360.res.ibm.com/`

```
models = {
    "log_reg": LogisticRegression(
        solver="liblinear",
        max_iter=1000
    ),
    "mlp": MLPClassifier(
        hidden_layer_sizes=(5,),
        max_iter=1000,
        alpha=0.01,
        random_state=42 # added for reproducibility
    ),
    "random_forest": RandomForestClassifier(),
    "svm": SVC(),
    "xgboost": XGBClassifier(),
    "decision_tree": DecisionTreeClassifier()
}
```

Listing 4.1: List of classification models available in the framework with default hyperparameters.

accuracy (BAL_ACC), Average Odds Difference (AVG_ODDS_DIFF), Equalized Odds Difference (EQ_ODDS_DIFF), Equal Opportunity Difference (EQ_OPP_DIFF), Disparate Impact (DI), Statistical Parity Difference (STAT_PAR_DIFF) and Average Predictive Value Difference (AVG_PRED_VALUE_DIFF).

The complete code for training is shared in the Appendix D, covering the technical aspects, including detailed Python implementations of the model classes.

As part of our design decision, we used *balanced accuracy*[6] and *F1 macro*[7] as utility measures, to refer to the effectiveness or performance of the model, to select the best model in the baseline evaluation node. These metrics are crucial for identifying a baseline model before applying bias mitigation interventions. Developers or stakeholders can choose the most appropriate model based on their criteria for different use cases.

4.2.4 Model Evaluation

In this stage, we evaluate the best-trained model using the initial split test data (unseen data). This model served as the baseline for subsequent fairness analysis. We generate predictions and compute the same metrics listed in Section 4.2.3. At this stage, metrics implemen-

[6]The balanced accuracy in binary and multiclass classification problems is usually recommended to deal with imbalanced datasets. It is defined as the average of recall obtained in each class. This measure is equivalent to accuracy, but with class-balanced sample weights and shares desirable properties with the binary case.

[7]F1 metric measures the harmonic mean of the precision and recall; the macro version calculates metrics for each label and finds their unweighted mean without considering label imbalance.

```
# For reproducibility, use the full code available in the Appendix Section

with DAG(
    "fairness_pipeline",
    description="ML Pipeline",
) as dag:
    download = PythonOperator (...)

    process = PythonOperator(...)

    train_logreg = PythonOperator(
        # Perform logistic regression model training
        ...
    )

    train_random_forest = PythonOperator(
        # Perform logistic regression model training
        ...
    )

    train_xgboost = PythonOperator(
        # Perform XGBoost model training
        ...
    )

    train_dec_tree = PythonOperator(
        # Perform XGBoost model training
        ...
    )

    train_mlp = PythonOperator(
        # Perform MLP model training
        ...
    )
    # The following line means that the training will start
    # after the download and process nodes are finished, the
    # train nodes will run in parallel using the same data.
    download >> process >> [
        train_logreg, train_random_forest,
        train_xgboost, train_dec_tree,
        train_mlp
    ]
```

Listing 4.2: Implementation skeleton comprising the data loading, data process, and models training using the airflow required structure. Each PythonOperator performs the actions passed as arguments and respects the concept of the proposed structure in practice.

tation also considers the data's protected attributes. Protected attributes refer to discriminatory features in decision-making that could induce privileged or unprivileged to certain population groups. The protected attribute in this node can be passed as an argument since it depends on the task scenario.

In our implementation, it is possible to evaluate the sensitivity of the multiple trained models by applying the evaluation node to all the desired models instead of just a baseline

model based on training performance.

Another aspect of our evaluation process is that the model training is performed across ten stratified k-folds, and the selection of the best model considered for evaluation continues using the 10-fold of the model instead of using a unique best model over the 10-folds. This approach allows us to perform an evaluation over ten times, providing statistical confidence levels in the evaluation process. This procedure is also used when applying the interventions.

The workflow does not automate further exploration of the evaluation results. Depending on the specific needs and case scenarios, the pipeline can be extended to include automated result exploration. For this research, the result analysis is performed manually and not integrated into the automated pipeline.

4.2.5 Fairness Interventions and Results Analysis

This stage aggregates the application of fairness interventions based on three statistical non-discrimination criteria – Independence, Separation, and Sufficiency. By "aggregate intervention," we mean that the framework allows developers to apply various intervention methods, each aligned with one of these mathematical criteria. However, the framework does not include a process to validate if a selected bias mitigation method belongs to its intended criterion. Standardizing mathematical criteria across all existent methods in the literature is challenging and not feasible within this framework. Instead, as theoretical guidance, we provide three interventions to help developers understand that the nodes encapsulating specific implementations fall under one of the fairness criteria.

Figure 4.2 provides an expanded view of the concept of the fairness intervention and result analysis nodes as a sequel of Figure 4.1. Within the framework, each fairness criterion can accommodate multiple intervention methods addressing the same notion of fairness. The community can extend our standard implementation with additional methods from the literature. We suggest in Figure 4.2 some other methods as examples that can be used to achieve a selected fairness notion. In bold and more highlighted colors are the methods used in our experimental design.

Regarding how the framework evaluates the results after applying an intervention, we follow the same approach used in the baseline evaluation. The intervention method, independent of the stage (i.e., pre-, in-, or post-processing), is performed over ten stratified k-folds, and the evaluation after the intervention follows the same procedure. This gives the user all the results from ten iterations over the 10-fold models.

Following, we describe the default techniques we provide in the framework, which correspond to algorithmic methods used to achieve each definition of fairness. The AIF360 Python

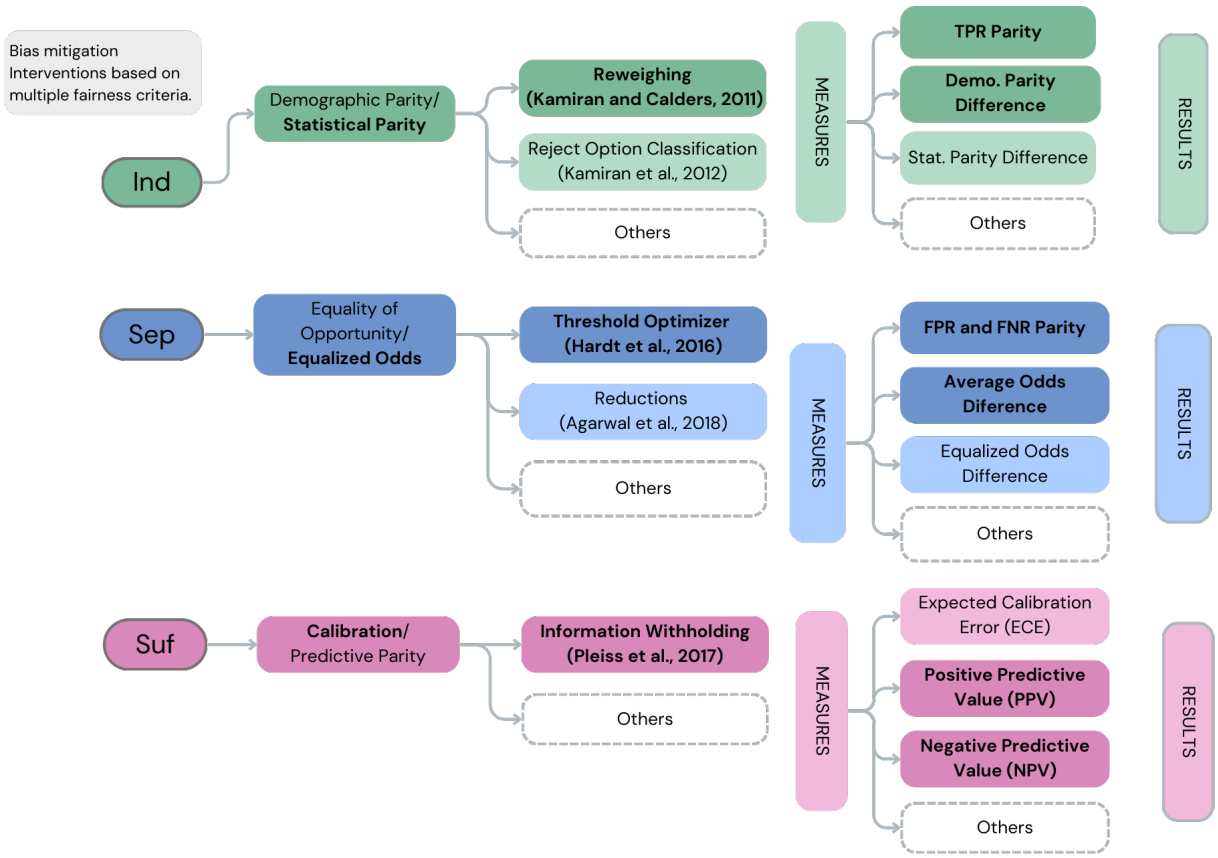


Figure 4.2: Fairness Intervention Concept. We aggregate the methods to include nodes for various interventions based on three statistical criteria. These interventions operate independently and can be executed simultaneously; we also suggest some measures based on the criteria that can be used to quantify unfairness; in the end, all the results are stored. The methods and measures used and currently implemented in our developed solution are the ones in bold text; the others are examples and suggestions.

toolkit was employed to assist the practical implementations of the selected methods that are in bold in Figure 4.2. Again, this is not an extensive list of options; the literature provides more types of interventions to achieve the same goal (Mehrabi et al., 2021; Caton and Haas, 2024).

4.2.5.1 Bias Mitigation Approach w.r.t Independence

To ensure fairness related to the Independence criteria, we provide a method referred to as **Reweighing**, proposed by Kamiran and Calders (2011). It consists of an approach that preprocesses data to remove discrimination by weighting the examples in each (group, label) combination differently to ensure fairness before the classification stage. If bias is defined following the Independence statistical criteria or an equivalent definition is not achieved after a

model evaluation, we return to the pipeline, reweigh the training data, and train the model again with the modified data. Finally, we evaluate the re-trained model using unseen testing data, and we proceed to verify if the bias mitigation approach was effective regarding the definition (Kamiran and Calders, 2011). This intervention implementation is available in the AIF360 toolkit and was adopted to be available in our framework. Independence is also referred to as Statistical Parity in the literature (Barocas et al., 2023), and a way to measure bias using this definition is measuring the acceptance rates across the protected attributes, such as *True Positive Rate Parity* and *Statistical Parity Difference*, which measures the difference between the proportions of the positive outcomes for two groups.

4.2.5.2 Bias Mitigation Approach w.r.t Separation

To provide a bias mitigation algorithm to achieve a fairness definition related to the Separation criteria, we adopted as part of our standard implementation a post-processing solution, the **Threshold Optimizer** proposed by Hardt et al. (2016), which is also available in the AIF360 toolkit (Bellamy et al., 2019). This intervention aims to achieve an equalization between rates by providing an equalized odds characterization, which states that the privileged and unprivileged groups should have similar false negatives and false positive rates. To achieve this fairness notion, a post-processor method, known as Threshold Optimizer, is employed without making changes in the training data or the trained model. The post-processor works as an optimization of the existing trained mode. It operates based solely on the joint distribution of the sensitive feature A and the predicted and actual outcomes, $\hat{Y}$ and Y, respectively, without requiring information about the non-sensitive features X. A way to measure the effectiveness of this approach and see the fairness quantification related to the definition is by comparing the rates of false negatives and false positives or comparing the true positives and negatives. In the literature, we also have a created metric often called "Average odds difference," which measures the disparity between false positives and true positive rates across sensitive groups.

4.2.5.3 Bias Mitigation Approach w.r.t Sufficiency

As a technique to be available to measure the Sufficiency criterion, we provide an approach focused on calibration, called in the literature as **Calibration via Information Withholding**, proposed by Pleiss et al. (2017) involves adjusting the model in a way that withholds

certain information from influencing the predictions, especially information that may result in unfair calibration. In other words, by reducing or ignoring some features (or modifying how they are used), the model prevents the over-exploitation of sensitive group-specific patterns that can reinforce biases. This method also relaxes the equalized odds approach and aims to maintain calibration between groups. Unlike other fairness notions, maintaining calibration and ensuring Sufficiency cannot consistently be enforced similarly to the other statistical definitions (Pleiss et al., 2017). This relaxation[8], is an optimization of our binary classification outputs to find probabilities scores changing the output labels with an equalized odds as an objective function. This optimization requires the selection of cost constraints to maintain calibration across groups, which can be the false positive rates, false negative rates, or a "weighted" cost, meaning a balance between false negatives and false positives. This means that the classifiers for different groups should lie on the same level-order curve of the cost function while maintaining calibration. Thus, we evaluate the calibration over the baseline to confirm that the fairness criterion of Sufficiency is approximately met.

4.2.5.4 Result Analysis

The final stage consist of an result analysis, and is not part of the automated workflow. In this stage, users are encouraged to build visualizations reports for each strategy to analyze the results and compare them with the baseline. These analyses help users to understand the effects of the different types of interventions, each based on different fairness assumptions.

Table 4.1 summarizes the interventions available and implemented in our solution based on the proposed framework, their corresponding closest statistical criteria, the intervention stage in which the approach is applied in the machine learning pipeline, and the fairness metrics that can be used for quantifying the bias in the model related to the definition. The metrics that we use to measure the bias related to each definition are options that come from the source reference of the method. However, our list does not cover all existent metrics, but we consider the ones mentioned at least enough to quantify bias related to the selected definition.

[8]Even as a relaxation method, this is also a post-processing technique that we will continue to refer to as an intervention in our case study to maintain the terminologies.

Intervention Method	Closest Criteria	Appr.	Fairness Measures
Reweighing (Kamiran and Calders, 2011)	Ind	Pre	**Stat. Parity Diff**, Demographic Parity, **TPR Parity**
Threshold Optimizer (Hardt et al., 2016)	Sep	Post	**Avg. Odds Diff, FPR and FNR Parity**, EOpp, Eq.Odds
Calibration via Info. Withholding (Pleiss et al., 2017)	Suf	Post	Calibration Curves, **PPV, NPV**, FDR, FOR

Approach: **Post** = post-processing, **Pre** = pre-processing
Equal Opportunity (EOpp), Equalized Odds (Eq.Odds), Positive Predictive Value (PPV)
Negative Predictive Value (NPV), False Discovery Rate (FDR), False Omission Rate (FOR)

Table 4.1: Standard intervention algorithms provided by the framework are categorized based on the application stage on the ML pipeline stage, along with the closest associated fairness criteria and metrics that can be used to quantify bias related to the definition. In bold are the measures used in the following experiments.

4.3 Experimental Tools

This research's experiments consist of applying the developed framework to assist the evaluation process and compare different fairness criteria in practice. The fairness interventions of the experiment were performed to select a baseline model after the data loading, processing, and training, following the framework concept and implementation. We analyze bias in the outcomes before and after the statistical bias mitigation interventions, considering the group fairness definitions previously described as Independence, Separation, and Sufficiency.

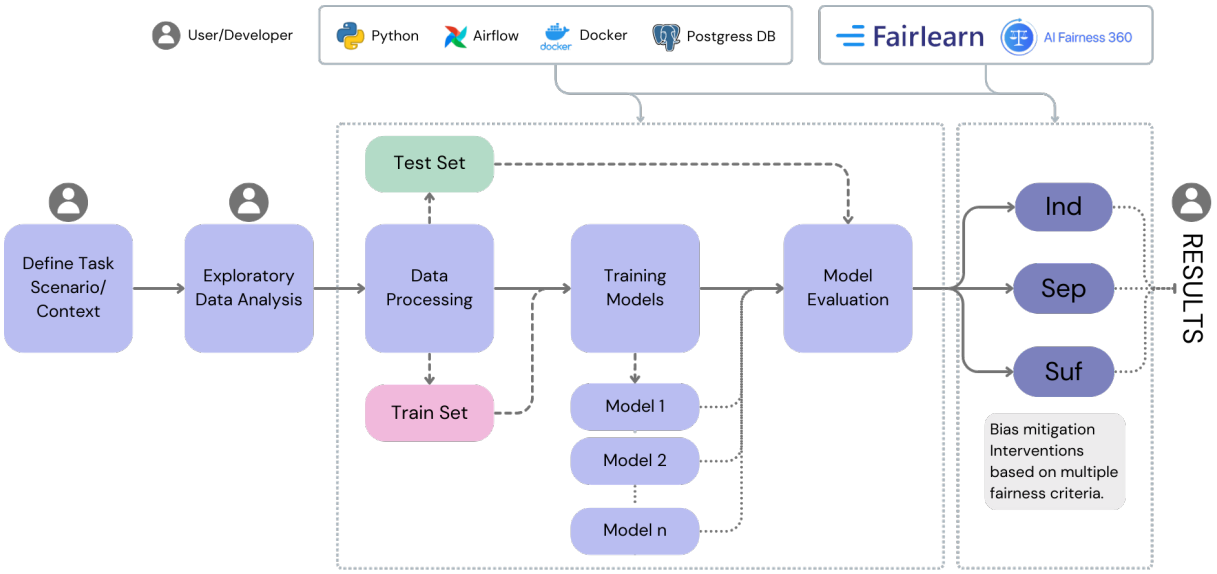


Figure 4.3: The experimental pipeline and the respective tools used.

Figure 4.3 shows the main tools used at each step to create the developmental design

for experimentation: Python for coding; Airflow[9]; Docker[10] and Postgres[11] relational database, for the orchestration pipeline, with User Interface. AIF360[12] and Fairlearn [13] to assess state-of-the-art bias mitigation algorithms and metrics. Besides the main tools, we also employed other tools to ensure all components work together. These include Pandas[14] for data manipulation; Scikit-learn[15] for provided predictive data analysis models, while Folktables[16] was used to fetch the dataset; and also Sweetviz[17] to generate data visualization reports for exploratory analysis.

Our practical implementation of the proposed framework has a user interface, as shown in Figure 4.4, with all the automatized nodes that were previously discussed and described in Section 4.1. All the implementations available as open-source include the UI as part of our standard framework as a tool.

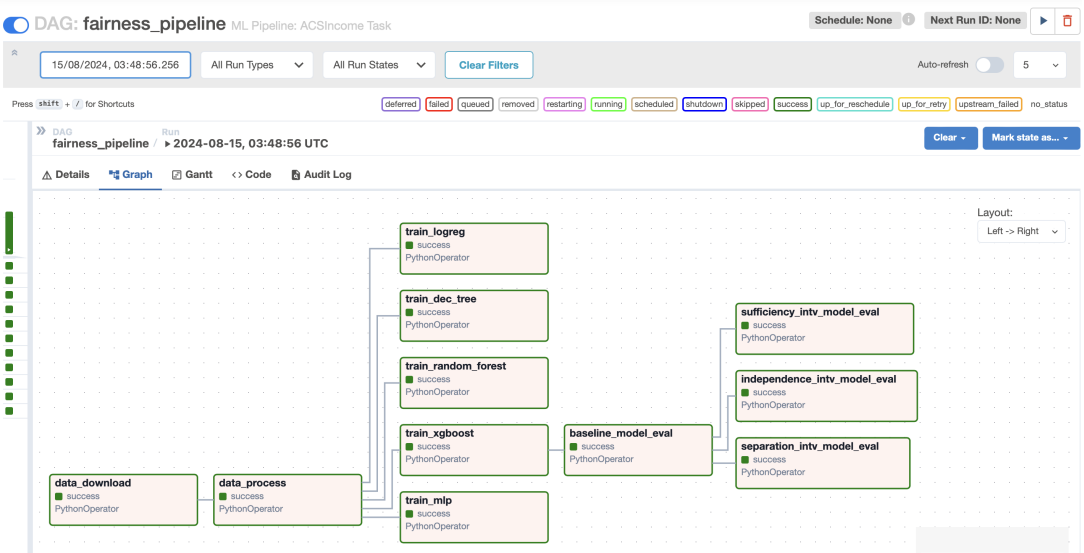


Figure 4.4: The experimental pipeline from the Airflow user interface (UI)

4.4 Case study: Income Prediction Task

The presented case study focuses on predicting whether the income of working adults in the US exceeds or falls below a certain threshold, specifically 50,000 dollars per year. The

[9] https://airflow.apache.org
[10] https://docs.docker.com
[11] https://www.postgresql.org
[12] https://github.com/Trusted-AI/AIF360
[13] https://fairlearn.org
[14] https://pandas.pydata.org
[15] https://scikit-learn.org/stable/index.html
[16] https://github.com/socialfoundations/folktables
[17] https://github.com/fbdesignpro/sweetviz

binary classification task labels individuals as high-income earners (≥ 50k) with a positive outcome or low-income earners (< 50k) with a negative outcome. The objective is to explore and narrow the gap between these income groups. Our case study serves as an application of the proposed framework. Aiming to support and demonstrate how the design choices assist the investigation of various fairness criteria regarding the same context without selecting a specific fairness definition in advance. This approach highlights the nuanced understanding of bias, presenting the trade-offs between different non-discrimination criteria and the associated performance costs in each scenario.

When predicting income with fairness approaches, different stakeholders have unique perspectives on what constitutes fair and ethical modeling. We can exemplify some applications regarding the use of a selected task – income prediction – and the use of this dataset for fairness purposes, such as: *loan approval*, where financial institutions use income predictions to assess an applicant's creditworthiness; *employment opportunities or hiring screenings*, where employers may use predicted income levels as a proxy for experience, skill level, or market value in hiring and promotion decisions; *public policies or social programs*, where governments use income predictions to identify individuals or communities in need of social assistance, subsidies, or educational grants; *housing market analysis*, where income predictions can assist in determining property affordability, market forecasting and pricing strategies for different demographic factors; etc. In each of these scenarios, fairness is essential to avoid bias and ensure ethical applications, especially when predictions influence life-changing decisions. In our case study, we will generically present results, taking into account the definition of fairness and the differences between them. However, when using the framework, the results must be interpreted according to the application scenario for more specific study cases with well-defined and mapped stakeholders. Only in this way is it possible to decide which of the definitions best suits the context.

The main body of this research discusses the results of applying the reference framework to this Income Prediction Task, with a binary feature selected as the sensitive attribute that demonstrates indicators of discrimination in the predictions. In the Appendix A, we demonstrate additional tasks, i.e., different datasets, considering different sensitive features, and executing the framework implementation similar to the main one discussed in the following chapter. We demonstrate the different perspectives using different data to give the reader more confidence about the benefits of our development framework.

Chapter 5

Dataset, Experiments, and Results

The previous chapter described the methodology for implementing the framework and the experiments for the case study, including the models used for training and the bias interventions provided for the framework that we applied in our described dataset, covering the Income prediction task.

This chapter presents the results of the algorithmic fairness experiments conducted to evaluate and compare the models after a bias mitigation intervention. The experiments were designed to illustrate the use of our framework through a case study that follows all the automated nodes until the visualization and conclusion of the results. We focus on the core aspect of the workflow: the ability to apply multiple interventions based on different fairness criteria. By employing distinct statistical definitions, we emphasize the consistency of the results derived from the same context and dataset.

5.1 Dataset

The data used in the experiments comes from the US Census, specifically from the American Community Survey (ACS), accessed via the Folktables[1] Python library (Ding et al., 2021). Folktables facilitate access to datasets for benchmarking machine learning algorithms. The ACS data was chosen mainly because of its large number of data points, which brings more confidence in relying on real-world scenarios with a more extensive range of features. The package includes a suite of pre-defined prediction tasks in domains such as income, employment, health, transportation, and housing across all US states, with options to select data from various years.

The ACS data contains over 5 million data points from all available tasks from all US states. For our experiments, we restrict the dataset to two of the most popular states, California (CA) and Texas (TX), using data from 2014. These states comprise approximately 302,000 data points, representing individuals from these regions. The filtered dataset includes ten attributes,

[1] https://github.com/socialfoundations/folktables

with five categorical and five numerical. The features are *age, class of work, education, marital status, type of relationship, place of birth (native country), occupation, weekly working hours, sex*, and *race*. For a detailed description of each feature's categorization, see Appendix E and Appendix B.1 of the source Folktables paper (Ding et al., 2021).

5.2 Experiments

In order to strengthen the reliability of our presented results and confirm that the discrepancies observed are not due to random chance, we conducted a Mann–Whitney-U[2] test for each pair of groups over the evaluations.

All plots in the following sections, especially the box plots, are labeled with a statistical significance annotation of the difference between the two compared classes, determined by the p-value, which measures the strength of evidence against the null hypothesis. We adopted this approach following some guidelines from the work of Park et al. (2022). We specify our levels of significance using the following annotations, where decreasing the p-values increases the levels of statistical significance:

[*] is p-value < .05 but > .01, meaning significant difference between the classes.

[**] is p-value < .01 but > .001, more significant difference between the classes.

[***] is p-value < .001 but > .0001, meaning highly significant differences between the classes.

n.s. no significance.

In all the following results, we provide the mean values of estimated performance across ten iterations from the 10-folds. Alongside the 95% of Confidence Intervals (CI), indicating the range in which the true mean likely falls, to reflect uncertainty in our results. These CIs were calculated by determining the standard error (a measure of how much the iteration results vary, using the standard deviation divided by the square root of the sample size) and multiplying it by a critical value from the normal distribution (for 95%, it is around 1.96 for large samples) to create the upper and lower bounds. We use the CIs to assist the understanding of the reliability of the mean, as a narrower interval suggests greater precision, while a wider interval indicates more variance in the data across our iterations.

[2]Mann–Whitney U test is a nonparametric statistical test used to assess whether two independent samples come from the same distribution.

5.2.1 Exploratory Data Analysis (EDA) & Data Processing

5.2.1.1 Data Exploration

The data exploration covered here comprises the required initial user node (**exploratory analysis**) that starts our framework conception; see Figure 4.1. We use the *sex* as our discriminating attribute in the experiment because it is one of the characteristics considered discriminatory according to some anti-discrimination laws highlighted by the literature (v. DeStefano, 2009; Ferrara, 2024). Although other features can also generate discrimination in our classification task, such as race or age, the "sex" feature is encoded using binary values, which facilitates the process of bias mitigation interventions since our standard approaches used to develop the framework initially consider binary sensitive attributes. Figure 5.1 illustrates our data distribution regarding the labels and the protected attribute. The dataset indicates that 53% of the individuals are male and 47% are female, with notable differences in the label distribution for high-income earners: 41% of males and 27% of females fall into this category. Appendix E provides a more comprehensive data report from our EDA, generated by the *sweetviz*[3] Python library.

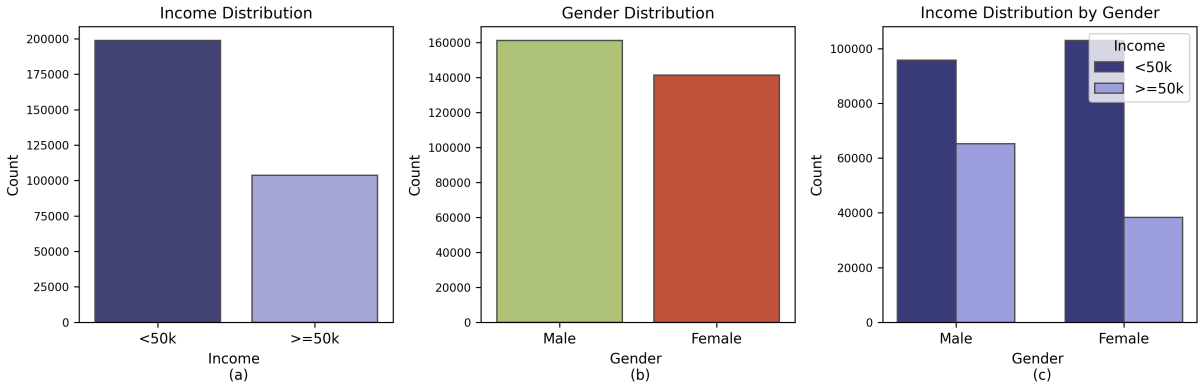


Figure 5.1: Dataset labels distribution. (a) Distribution of the label between low-income and high-income. (b) The distribution between the males and females (our protected feature). (c) Label distribution across the sensitive attribute.

[3] https://github.com/fbdesignpro/sweetviz

5.2.1.2 Data Processing

Our data analysis and explorations assist the following node in our framework, corresponding to **data loading and data processing** respectively. Our exploratory data analysis enabled us to identify if our data has missing values and duplications and comprehend the data distribution and correlations between features. It is important to remember that our preprocessing stage, described in the previous Section 4.1 and applied here, does not perform any data modification concerning bias reduction in the data or any type of fairness treatment in the dataset.

Our preprocessing standard implementation covers these previously listed aspects (removing missing values and duplications) and, the standardization of categorical and numerical features and since we have five categorical and five numerical; Neither duplications nor missing values were detected. After that, our data are splited into 80% for training and 20% for testing, and stored to be used by the other nodes.

Table 5.1 shows the distribution of labels in the dataset after processing and stratified splitting into training and testing sets.

Labels	**Training Data**	**Test Data**	**Total**
$<$ 50k (0)	158,972 (66%)	39,741 (66%)	198,713
$\geq$ 50k (1)	82,844 (34%)	20,713 (34%)	103,557
Total	241,816 (100%)	60,454 (100%)	302,270

Table 5.1: Distribution of labels in the training and test datasets.

5.2.2 Models Training and Baseline Selection

Our experiment follows the node of **training binary classification models**, and here we also **choose the best model to fit**. As described in Chapter 4, a variety of candidate models were implemented. One of these models was selected as the baseline for comparison with fairness interventions. Each model was trained using 80% of the dataset without incorporating any anti-discrimination measures or parameter tuning. We employed stratified 10-fold cross-validation to evaluate their performance, assessing the models' "out-of-sample" performance on a validation set. Table 5.2 below shows the training performance of the candidate models over ten iterations. Each iteration corresponds to an evaluation using each model from the 10-fold strategies, and the values are the means of the ten folds, with 95% confidence level. The

results for the best-performing model are highlighted in bold, and this model was selected as the baseline model for the subsequent experiments. This training was also performed using the framework implementation.

		Training Performance	
	Model	**Bal Acc ↑**	**F1 Macro ↑**
No Mitigation	Decision Tree	0.726 ± 0.003	0.726 ± 0.003
	Logistic Regression	0.757 ± 0.003	0.765 ± 0.003
	Neural Network (MLP)	0.774 ± 0.003	0.778 ± 0.003
	Random Forest	0.779 ± 0.002	0.784 ± 0.002
	XGBoost	**0.803 ± 0.002**	**0.806 ± 0.002**

Table 5.2: Mean values of the training performance metrics for candidate models, evaluated across ten cross-validation iterations, corresponding to the 10-fold models (95% confidence interval). Bold values indicate the best performance, while underlined values indicate the worst. For all metrics, higher values (up arrow ↑) represent better performance.

The differences in balanced accuracy and F1-macro between models were statistically significant (p-value < .001), leading to the selection of the XGBoost classifier as the best-performing model.

Thus, the model selected as the baseline was chosen to be evaluated on the test set (20% of the total data) using the ten-fold from the initial cross-validation approach. This model was also used to apply each fairness intervention, and its performance was evaluated on the same test set. The results of these evaluations, along with the bias mitigation methods, are discussed in the subsequent section.

5.3 Experimental Findings

In this stage, we follow the framework concept and implementation, and the results described in this section cover the **model evaluation, fairness interventions, and results analysis** nodes. The baseline was evaluated using the test data (unseen by the model) and performed across the 10-fold. The intervention results, i.e., predictions after the bias mitigation, were accomplished following the same approach.

Table 5.3 presents the average performance across four scenarios: (1) the baseline model (XGBoost), (2) the baseline model after applying a reweighing intervention (Kamiran and Calders, 2011), which aligns with the Independence criteria, (3) the baseline model with a threshold optimizer (Hardt et al., 2016), which addresses the Separation criterion, and (4) the baseline model after Calibration via Information Withholding (Pleiss et al., 2017), which is

equivalent to the Sufficiency definition. The performance results are overall values from the predictions, not filtering across the sensitive groups. Figure 5.2 illustrates the folds iterations.

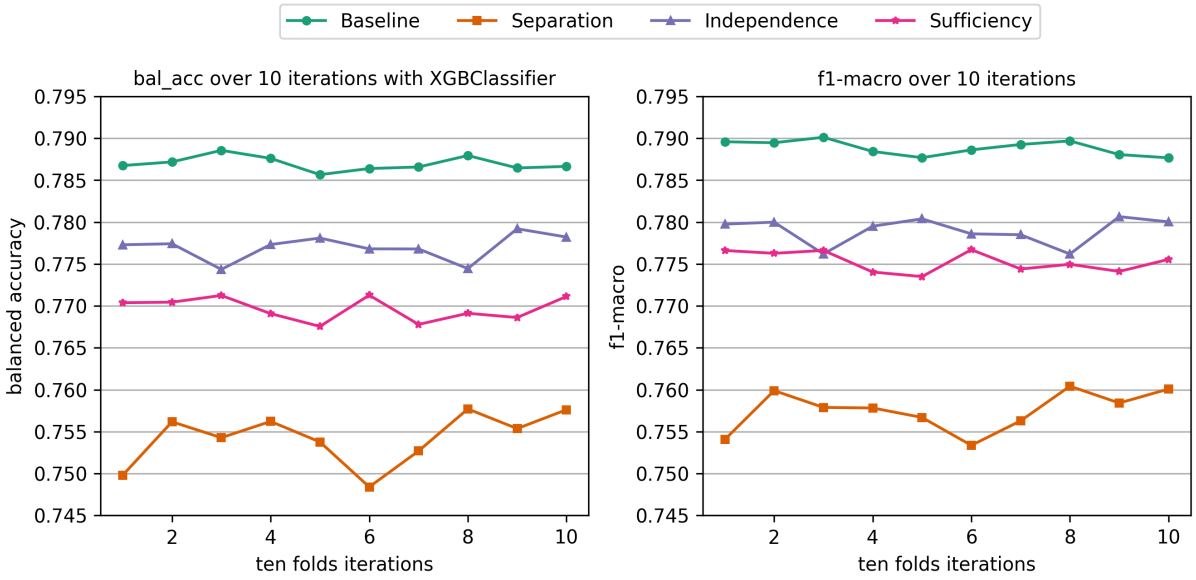


Figure 5.2: Evaluation Performance: baseline model and all intervention approaches over ten iterations. Each iteration was performed through the 10-fold, and the results presented used the models evaluated using the test data (unseen data).

			Evaluation Performance			
Appr.	**Type**		**Bal Acc ↑**	**p-value**	**F1 Macro ↑**	**p-value**
N/A	Baseline	-	0.787 ± 0.001	-	0.789 ± 0.001	-
Pre	Reweighing	IND	0.777 ± 0.001	$< .001$	0.779 ± 0.001	$< .001$
Post	Thresh. Opt.	SEP	0.754 ± 0.002	$< .001$	0.757 ± 0.002	$< .001$
Post	Calibration	SUF	0.770 ± 0.001	$< .001$	0.775 ± 0.001	$< .001$

IND: Independence, SEP: Separation, SUF: Sufficiency

Table 5.3: Evaluation performance metrics, comprising balanced accuracy and F1-macro. Mean values are presented over ten folds iterations with 95% confidence level. Statistical significance levels are compared with the baseline. Higher values (up arrow ↑) represent better performance for all metrics.

Previously, we discussed the methods of reweighing, threshold optimizer, and calibration, highlighting the statistical fairness objectives that each intervention aims to achieve. Building on that, we now explore the fairness measures associated with each approach to understand their impacts and benefits better.

5.3.1 Intervention w.r.t Independence: Reweighing

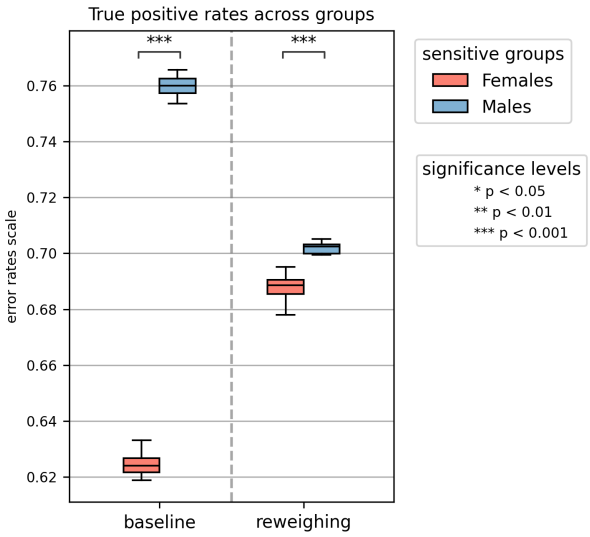
The Reweighing approach (Kamiran and Calders, 2011) was applied to achieve a fairness notion related to the Independence criteria, also known as Statistical Parity (Barocas et al., 2023). In this context, fairness implies equal acceptance rates across groups. Figure 5.3 (a) illustrates the boxplot statistics of the true positive rates (TPR) across the classes of males and females and (b) the statistical parity difference measure before and after the intervention, with statistical significance annotations. Table 5.4 shows a closer look at the values from Figure (a).

As a fairness metric related to the statistical definition, we use the Statistical Parity Difference, which measures the difference between the majority and protected classes in receiving a favorable outcome. In other words, the equal acceptance rate across all demographic groups (e.g., gender/sex) ensures that both groups have similar chances of receiving a positive outcome. A negative value of this metric indicates that the unprivileged group is at a disadvantage, while a positive value indicates that the privileged group is at a disadvantage. Ideally, the rates should be minimal. There are other variations of statistical parity as a metric of fairness; the version used here as a *difference* is often used where there are two groups of interest in the decision-making process and is useful for visualizing the acceptance proportions between them. Other ways to use this metric can include measuring individual, ratio, or pairwise statistical parity (Dwork et al., 2012).

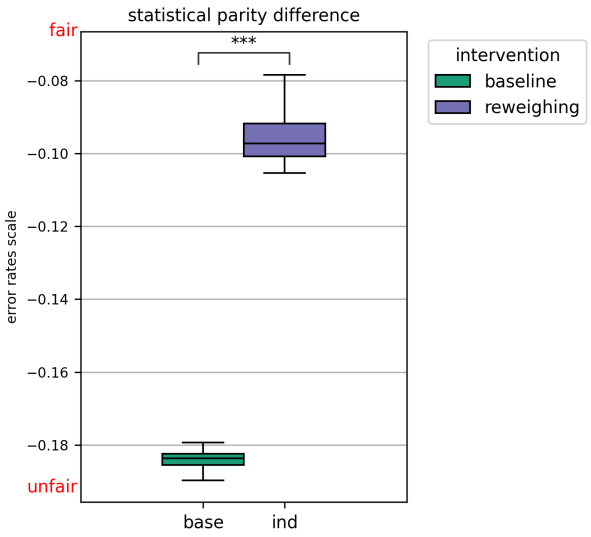
Classifier	**Measure**	**Female**	**Male**	**p-value**
Baseline	True Positive Rate	0.625 ± 0.003	0.760 ± 0.003	$< .001$
Reweighing	True Positive Rate	0.688 ± 0.004	0.701 ± 0.004	$< .001$

Table 5.4: Baseline vs. intervention closest to independence. TPR shows the fraction of positive cases correctly predicted to be in a positive class out of all actual positive cases; the values are filtered between groups (e.g., males and females). Statistical significance levels are compared between males and females. Results are over ten iterations with 95% confidence level.

Our intervention, applied in a pre-processing stage and focusing on acceptance rates, resulted in an improvement of approximately 48.27% compared to the baseline. Initially, the baseline model presented a significant negative value, indicating that the unprivileged group (e.g., female individuals) was at a disadvantage.



(a) TPR across Sensitive Groups



(b) Fairness Metric: Baseline vs. Reweighing (Independence)

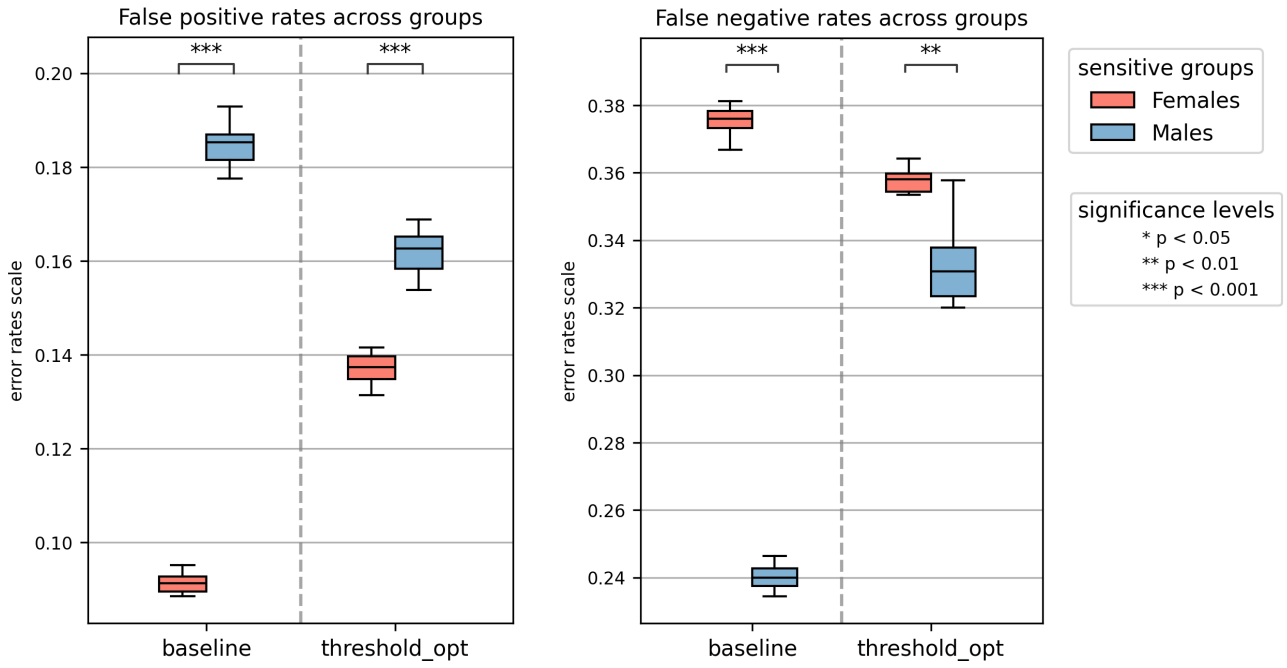
Figure 5.3: (a) True positive error rates across sensitive groups. After applying the reweighing, a comparison is made between the baseline model and the results, with averages over ten iterations. (b) Statistical Parity Difference measures the difference between the proportions of positive outcomes for two groups. A more negative value of Statistical Parity Difference indicates that the unprivileged group is at a disadvantage, while a more positive value indicates that the privileged group is at a disadvantage.

5.3.2 Intervention w.r.t Separation: The Threshold Optimizer

The threshold optimizer approach (Hardt et al., 2016) was applied to enforce equalized odds concerning the sensitive feature "sex." This fairness notion, aligned with the statistical definition of Separation, aims to ensure that false negatives and false positive rates are similar across groups. To evaluate the effectiveness of this intervention, we compare the average rates of false negatives and false positives between sensitive groups, e.g., males and females.

In the baseline model, males were incorrectly predicted to have a high-income 18% of the time, while this rate was only 9% for females. At the same time, the baseline model incorrectly predicts that males have a low-income 24% of the time, compared to 37.5% for females (highlighted in bold in Table 5.5). Overall, the baseline model shows significant disparities, with males benefiting from better identification of high earners, while females face more misclassification when they earn $\geq 50k$ per year. The threshold optimizer intervention equalized these disparities, resulting in closer error rates for females, who were previously disadvantaged. Figure 5.4 illustrates the boxplot statistics of the false error rates across groups.

Figure 5.5 shows the average odds difference for both the baseline model and after applying the mitigation approach. This metric involves two metrics, measuring bias based on false positive rates (false favorable label rate) and true positive rates (true favorable label rate) across



(a) FPR across sensitive groups

(b) FNR across sensitive groups

Figure 5.4: (a) False positive and (b) negative error rates across sensitive groups. A comparison is made between the baseline model and the results after applying the threshold optimizer, which averages over ten iterations.

Classifier	Measure	Female	Male	p-value
Baseline	**False Positive Rate**	**0.091 ± 0.001**	**0.185 ± 0.003**	< .001
	True Positive Rate	0.625 ± 0.003	0.760 ± 0.003	< .001
	False Negative Rate	**0.375 ± 0.003**	**0.240 ± 0.003**	< .001
	True Negative Rate	0.909 ± 0.001	0.815 ± 0.003	< .001
Threshold Optimizer	**False Positive Rate**	**0.138 ± 0.004**	**0.162 ± 0.004**	< .001
	True Positive Rate	0.643 ± 0.004	0.666 ± 0.009	< .01
	False Negative Rate	**0.357 ± 0.004**	**0.334 ± 0.009**	< .01
	True Negative Rate	0.862 ± 0.004	0.838 ± 0.004	< .001

Table 5.5: Baseline vs. intervention closest to Separation. Equalized odds are achieved across the sensitive attribute. Values of false positives, false negatives, true positives, and true negatives across groups are presented over ten iterations at 95% confidence level. Statistical significance levels are compared between males and females.

sensitive groups, with a target value of zero indicating perfect fairness according to this definition. The intervention achieved effective results, showing approximately 80% improvement over the baseline.

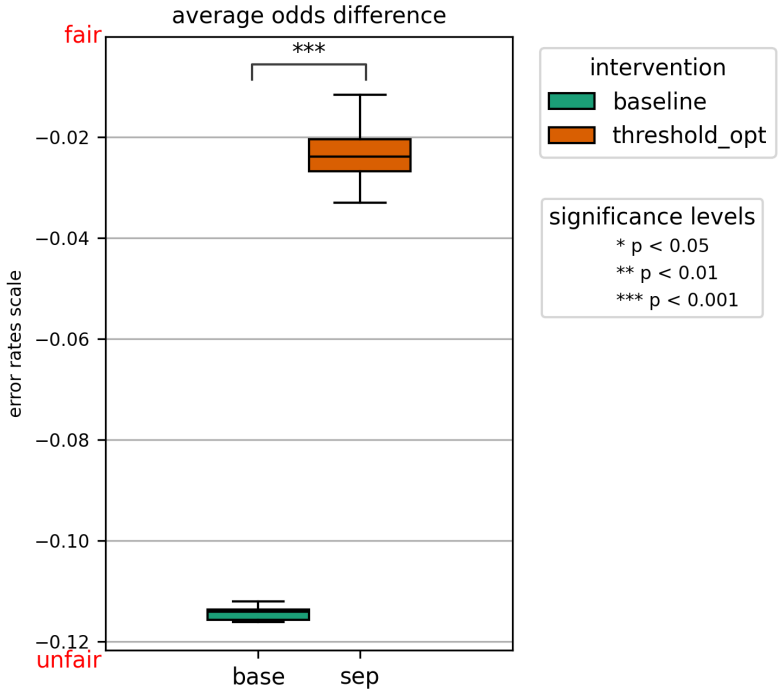


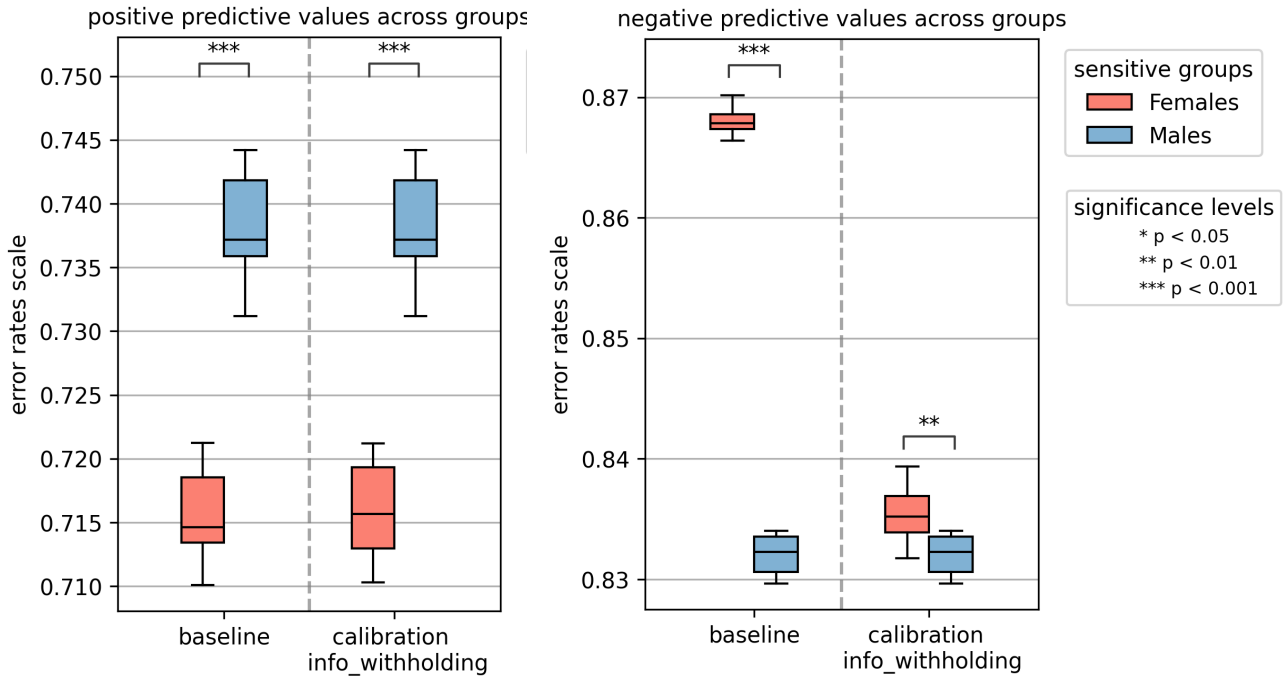
Figure 5.5: Average odds difference between unprivileged and privileged groups. This fairness metric measures the disparity between false positives and true positives rates across sensitive groups. A lower absolute value, closer to zero, indicates better fairness.

5.3.3 Intervention w.r.t Sufficiency: Calibration via Information Withholding

To access a fairness notion related to Sufficiency, our framework implementation provides a calibration approach. According to Barocas et al. (2023), the concept of calibration is intuitive; if our outcome scores are calibrated by group, a given score value indicates the same rate of positive outcomes across all groups. As previously discussed (Section 3.1.3), calibration and the definition of Sufficiency are directly equivalent and are often satisfied by the outcomes of unconstrained supervised learning without requiring explicit intervention.

The method proposed by Pleiss et al. (2017) ensures that the predicted probabilities are meaningful and consistent across different groups. As our cost constraint to preserve calibration, the method assigns a weight to each type of error to balance false-positive (FP) and false-negative (FN) rates. A calibration attempt using this cost constraint can achieve a more nuanced form of fairness that takes into account the relative importance of different types of errors rather than focusing solely on one type of error.

To measure the effectiveness of our weighted calibration, we measure the positive and predictive values across the sensitive groups. Figure 5.6 shows our measures used to quantify the sufficiency definition, and Table 5.6 has a closer look at the mean values from the plots.



(a) PPV across sensitive groups

(b) NPV across sensitive groups

Figure 5.6: (a) Positive predictive values and (b) negative predictive values across sensitive groups. A comparison is made between the baseline and the weighted calibrated models after the attempt to maintain calibration between groups, which averages over ten iterations.

Regarding PPV, after calibration, males are even more advantaged in terms of correctly identifying individuals who earn $\geq$ 50k per year. The calibration seems to have favored males and slightly deteriorated the disparity between the classes in predicting a higher income. Regarding the NPV, at the baseline, females are significantly higher than males. This means the model was better at correctly predicting low income for females compared to males. The NPV for females decreased after calibration, suggesting that females are now less advantaged in terms of correctly identifying individuals earning $< 50k$ per year. The NPV values for males and females are now closer, meaning the disparity between genders in predicting low income has been reduced, achieving certain calibration but at the cost of balancing.

In summary, males benefited more by reducing false positives (higher PPV), and females paid the cost, as their false negatives increased (lower NPV), resulting in a loss of performance in predicting low income. However, our cost constraint to preserve calibration was balancing FPs and FNs, which was achieved, but consequently, the fairness between groups in terms of performance was not preserved.

Classifier	Measure	Female	Male	p-value
Baseline	Positive Pred. Values	0.7157 ± 0.003	0.7160 ± 0.003	$< .001$
	Negative Pred. Values	0.8680 ± 0.001	0.8320 ± 0.001	$< .001$
Calibration	Positive Pred. Values	0.7160 ± 0.003	0.7380 ± 0.003	$< .001$
	Negative Pred. Values	0.8356 ± 0.002	0.8320 ± 0.001	$< .01$

Table 5.6: Baseline vs. intervention closest to sufficiency. Statistical significance levels are compared between males and females. Results are over ten iterations with 95% confidence level.

5.4 Discussion

In our case study, we evaluated a machine learning model after applying the framework in our development fairness assessment, using three types of interventions based on different statistical group fairness approaches. To highlight the trade-offs between these criteria from a practical perspective, we summarize the caveats and recommendations, as well as our observed gains from attempts to mitigate discrimination, as follows:

- **Independence (or Statistical Parity):**

  Measures/Compare: This concept compares predictions across groups while disregarding the true values.

  Motivation To Use: This definition is a good choice when it is known that the data contains demographic bias.

  Intervention Effectiveness: Our attempt to mitigate bias concerning this definition showed that our dataset presented disparities across groups. After applying the intervention, we achieved an improvement of approximately 48% over the baseline, reducing discrimination in predictions against unprivileged groups.

  Caveats and Recommendations: Since we are only concerned with predicting values, we may lose some information during optimization, and the selection rate can create a misleading impression that the acceptance rates are fair.

- **Separation (or Equalized Odds):**

  Measures/Compare: This definition compares true and false positive rates (error rate parity) across groups.

  Motivation To Use: This fairness approach is suitable when the dataset does not contain historical or measurement bias and the error rates between classes are equally important to achieve.

Intervention Effectiveness: The intervention method applied concerning this definition improved by 80% over the baseline, revealing significant disparities in false positive and negative rates before the mitigation process.

Caveats and Recommendations: This definition has its limitations. If a dataset is biased, this fairness measure cannot capture the entire picture, as it relies on the availability of ground truth from the test data. A higher imbalance between negative and positive classes will highlight issues related to privileged and unprivileged groups.

- **Sufficiency (or Calibration):**

  Measures/Compare: This criterion compares predicted outcomes to true outcomes.

  Motivation To Use: Sufficiency is a good choice when predictions are reliable for everyone, regardless of whether they belong to unprivileged or privileged groups.

  Intervention Effectiveness: Our results showed that our objective to maintain calibration with a cost constraint balancing the false positives and false negatives was achieved and reliable for different populations. However, in terms of performance, the fairness between the groups didn't improve.

  Caveats and Recommendations: If historical biases exist in the data, ensuring Sufficiency could inadvertently reinforce these biases, as it only ensures that predictions are fair within the context of the predictions themselves. However, while this approach can help balance different types of errors, it may still result in performance degradation and increased disparity in other fairness metrics.

Nevertheless, it is important to take into consideration the **specific context** and **societal implications** when selecting a fairness definition, as each has its strengths and limitations. For example, stakeholders' perspectives when my context is a real state and housing market, where income prediction is used to determine property affordability and pricing strategies. The results can be interpreted in different ways under each fairness criterion.

From an Independence (or statistical parity/demographic parity) point of view, this means each demographic group (e.g., based on race, gender, religion, or other characteristics) has an equal chance of being predicted as able to afford a property regardless of actual income differences. This can promote certain accessibility to housing for historically disadvantaged groups by balancing representation across demographics in housing qualification. On the other hand, this can also potentially overestimate affordability for certain groups if income disparities exist, leading to inaccurate predictions for those who may struggle with property costs, as actual income levels are not fully reflected.

From a Separation (or equalized odds) point of view, the income predictions should have similar error rates for each group, meaning that people from each demographic group who

actually qualify for a property are predicted accurately at similar rates, and those who do not qualify are also correctly identified at similar rates. Considering this definition as a measure of what is fair, this approach can help avoid disparities in the model's accuracy for different groups, meaning that groups with lower average income levels are not disproportionately affected by false positives or false negatives. However, it may be challenging to maintain both fairness and predictive power if the base income distributions across groups vary significantly.

From a Sufficiency (or calibration/predictive parity) point of view, if our model is well-calibrated, then the prediction matches the actual ability to afford a property for every demographic group. This means that if a certain income is predicted, it will reflect reality for each group without bias. Considering this approach means that the calibration can provide realistic affordability predictions and help ensure that groups are neither overqualified nor underqualified for properties, reflecting an accurate economic capacity. On the other hand, it may not directly address equal representation or error distribution across groups and could potentially reflect existing income disparities if those differences are present.

The trade-offs between different fairness definitions must be carefully evaluated to ensure that the chosen approach aligns with the stakeholders' goals and the broader societal impact. Each criterion provides different interpretations; for our housing market example of application using our income prediction task, we can see the pros and cons of each definition. The decision-making process should be able to take into account all the possible scenarios when creating an algorithmic solution. So, selecting the appropriate fairness criterion will depend on your specific goals: for equal access (independence), for balanced error rates (separation), or for accurate affordability predictions (sufficiency); therefore, it is necessary to choose among them since they cannot be achieved simultaneously.

In summary, our framework facilitated the implementation of multiple interventions via the orchestration pipeline. Knowing the goal of the fairness definition facilitates the knowledge of how the user can quantify the changes and the unfairness after an intervention. Since each chosen approach changed the baseline model's results from the statistical perspective of the respective fairness criteria. The metrics to quantify unfairness were straightforwardly selected, considering that the notion of fairness and the intervention were defined and introduced in advance. Regarding our quantitative results, the performance measures (balanced accuracy and f1-macro) were reduced in all three perspectives; this comes because of the trade-off between fairness and utility since all *group fairness* definitions usually become more fair in terms of modifying some error rate distribution that can cause changes in the measures of the overall performance.

Chapter 6

Conclusion

Ensuring algorithmic fairness has become essential in all applications involving automated decision-making. As highlighted by Morley et al. (2023), operationalizing AI ethics is complicated, especially when translating ethical principles into practice. The literature offers an extensive list of codes of conduct, frameworks, and standards for different subjects, serving as guides for diverse services. However, as argued by Morley et al. (2023); Caton and Haas (2024) and Pagano et al. (2023), there is a gap in the application of these toolkits and guides, as most resources remain abstract for those practitioners that lack practical applicability of responsible and nondiscriminatory algorithms design. Motivated by this gap, our research focuses on creating a framework for employing bias mitigation approaches in classification tasks. This framework explores various fairness criteria and provides an open-source implementation workflow that can be adapted for other case studies and scenarios.

In most works on fairness and bias mitigation, defining what fairness means in the specific context of study or application is an essential first step before applying any method to reduce discrimination. Without a clear definition, it becomes challenging to select what we want to measure in the outcomes and determine what constitutes unfairness. The literature presents an extensive list of different fairness-related metrics, each applicable under specific circumstances. Not all metrics are appropriate for every context, as the concept of "fairness" can vary significantly across different stakeholders and mathematical definitions.

Our framework proposes a flexible approach that does not tie the development of a solution to a single fairness definition. This flexibility allows for diverse outcomes from varying perspectives. We focus our design on discrimination reduction, leading to a better understanding of the statistical criteria that define the three main categories in the literature: Independence, Separation, and Sufficiency. We discussed that most existing fairness notions are equivalences or relaxations from these foundational definitions, as extensively discussed by Barocas et al. (2023). The design choices of our practical framework are built upon this foundation.

In our experiments, the framework facilitated the application of multiple fairness interventions, each based on different statistical criteria and considering the same task context, allowing multiple outcomes perspectives when the goal is to reduce discrimination between groups. The Reweighing approach improved the statistical parity difference by approximately 48% over the baseline, reducing the gender disparity between the sensitive groups, which ini-

tially showed males as more advantaged, resulting in the intervention an improvement in fairness performance considering the true positive rates (with statistical significance <.001). The Threshold Optimizer intervention improves fairness between females and males by bringing the *error rate parity* closer together. Suggesting that the intervention aimed to reduce bias in the model significantly reduced the disparity in how often females are misclassified when earning $\geq 50k$ per year, achieving an 80% improvement regarding the average odds difference. The p-values (<.001 and <.01) indicate that these improvements significantly reduce unfairness across groups. The Calibration method preserved some aspects of the initial calibration, as expected since we cannot enforce a perfect calibration, also achieving a balance of false negatives and positive predictive values (our cost constraint); however, this calibration cost performance of fairness related to the sensitive groups' outcomes.

Having different perspectives on achieving some notions of fairness in the same context, with a knowledge of the definition and unfairness quantification (selecting metrics), demonstrates the framework's capability to address various fairness perspectives effectively. Since some statistical concepts group them, the final user can choose the most appropriate aspect of fairness with a bias mitigation approach. We highlight some key contributions that our framework can assist when adopted as a tool based on different actors; however, to validate our assumptions, it is necessary to apply a human assessment with the variety of roles that can be target users or diverse stakeholders:

- Improved clarity and understanding (for practitioners and researchers): By categorizing definitions, we can simplify complex ideas, making it easier to understand differences among definitions (e.g., demographic parity vs. equalized odds).
- Consistency across applications (for industry practitioners): Supporting standardized fairness practices and creating replicable, reliable systems. Reducing the need for case-by-case metric customization.
- Enhanced comparability and benchmarking (for researchers and policy-makers): Researchers can compare models or datasets on a shared statistical basis, which is essential for benchmarking and effectively assessing fairness trade-offs.
- Targeted bias mitigation strategies (for developers and fairness practitioners): Developers and practitioners can tailor their interventions more precisely, allowing for bias mitigation that targets specific statistical definitions and their respective ways to quantify unfairness.

Our research comprises an orchestration tool that provides a complete workflow for a machine learning pipeline, focusing on training different models as baseline candidates and applying different fairness interventions to a selected model, with each intervention based on a distinct fairness notion. This framework can contribute to enhancing state-of-the-art fairness research by providing a structured, statistical basis for evaluating and comparing various fairness definitions, avoiding "one-size-fits-all" solutions. Differing from Aequitas (Saleiro et al.,

2018) and Audit-AI[1] tools, that provide auditing methodology to reduce bias, but not development guidance across multiples definitions. Concerning other open-source tools, our framework adopts the resources from Fairlearn (Bird et al., 2020) and AIF360 (Bellamy et al., 2019) since they provide stand-alone state-of-the-art algorithmic methods and metrics that, with our methodology, can be applied to cover multiple fairness and bias notions. Researchers can establish a common language to explore trade-offs and benchmark models across different contexts by standardizing metrics selections and categorizing them through a statistical lens. This approach supports the development of more targeted, evidence-based bias mitigation techniques and encourages experimentation with hybrid models or new fairness definitions that bridge gaps between existing concepts. We can move forward by applying a unique fairness metric or bias mitigation technique to every model or dataset, regardless of the particular biases, goals, or stakeholders involved, and start using the development approach of having multiple perspectives of the context to contemplate multiple audiences. Further, it promotes reproducibility, making research findings more accessible to validate and apply, thus advancing applicable development methods in fairness research. We also provide a case study illustrating the diverse outcomes that can be accomplished using different fairness measures based on statistical concepts, offering users the flexibility to continue future experiments within the same context.

We also acknowledge that there are many types of bias (Ferrara, 2024; Kordzadeh and Ghasemaghaei, 2022), and mitigating discrimination that comes from any automated decision-making is not a straightforward task. In this research, we focused on mitigating bias in an existing solution that requires correction, specifically improving a baseline model that produced unfair outcomes based on the selected fairness definition. Each chosen approach modified the baseline model's results from the statistical perspective of the respective fairness criteria. We summarized and discussed the nuances, caveats, and recommendations in more detail for our case study, which should be taken into account by decision-makers evaluating fairness from different perspectives. From the experiments conducted, our framework facilitated the implementation of multiple interventions via the orchestration pipeline. To implement the orchestrated pipeline of our framework, we adopted Airflow and Docker, which enabled a flexible, intuitive, and adaptable workflow with a user interface that provided visibility into the experiment stages. We hope this approach can lead to a more extensive way to assess distinct fairness notions, with the aim of reducing unwanted bias.

[1] https://github.com/pymetrics/audit-ai

6.1 Limitations

In addition to the challenges of mitigating bias using different fairness perspectives, several caveats should be considered regarding our approach. Below are the main limitations of this work:

- **Other types of fairness notions:** All experiments and state-of-the-art algorithms in this research focused solely on group fairness notions, addressing unfair outcomes that affect a group within the dataset population (e.g., race, gender, religion). We did not explore individual or causal fairness notions, which could be addressed in future work.
- **Multiclass classification problems:** Our experiments were limited to binary classification tasks. Despite the flexibility of our framework, the metrics and algorithms available in AIF360 and Fairlearn do not fully cover fairness or multiclass classification tasks, which would require additional evaluation procedures.
- **Multiples sensitive attributes:** The standard implementation and current fairness interventions are not adapted to handle multiples protected attributes or multiclass tasks. However, they can be expanded by designating one class as favorable and others as unfavorable or by using sensitive attributes with more than two categories (e.g., ethnic group, socio-economic level, age).
- **Different source of bias:** While we acknowledge the numerous sources of bias, addressing all existing types was beyond the scope of this research. We focused on a manageable subset, excluding aspects such as data or societal bias. However, interventions addressing these types of bias can be incorporated into our pipeline without significant difficulties. Appendix B provides more details on the tool's usability.
- **Need for human assessment and validation:** Although our framework shows a promising approach, we did not validate our design choices with large-scale AI system developers beyond the motivations derived from the literature. It can be helpful to collect the strengths and weaknesses of our structure, with both AI fairness experts and developers working on fairness issues.
- **Fairness definitions with societal perspectives:** In the framework design, we focused on fairness in risk allocation or assessment problems using protected attributes and statistical or mathematical notions, incorporating social perspectives throughout the development process rather than just in the result analysis is crucial. This study analyzes the trade-offs of different fairness definitions with assumptions about justice based solely on experimental outcomes. This highlights the need for human assessment and insights from domain experts to better align fairness definitions with societal viewpoints.

6.2 Future Research Directions

We proposed several key future directions to further expand this work and address some limitations previously mentioned in Section 6.1. These are not exhaustive and may evolve as the project develops:

- **Improving user experience with built-in visualization tools:** Improving operationalization of algorithmic fairness includes offering built-in visualizations of outcomes. Providing appropriate visual resources for different fairness definitions will improve interpretability for both developers and non-developers, making results more accessible.
- **Interpretability and explainability resources:** A significant focus can be placed on improving interpretability and explainability[2] across all stages of the pipeline. This is especially important for measuring unfairness, identifying bias sources, and understanding trade-offs between fairness concepts. Expert insights will be crucial here.
- **Supporting multiclass models and sensitive attributes:** As discussed in the limitations, this work focuses on binary classification problems with binary sensitive attributes. Expanding the framework to include multiclass classification tasks and non-binary sensitive attributes, such as *race* that can be found in the dataset into more than two categories: white, Black, Asian, American Indian, and others.
- **Addressing individual and causal fairness:** In addition to group statistical fairness definitions, we plan to extend the framework to incorporate definitions that address individual fairness and causal fairness.
- **Community feedback and validation:** As highlighted in Section 6.1, gathering community feedback will be essential to refining and expanding the framework.
- **Adapting to distribution shifts:** To increase the framework's robustness in real-world scenarios, we will explore how distribution shifts — where the target distribution in data-driven policy applications diverge from the training distribution — affect fairness and bias. This will enhance the reliability of our framework in practical machine-learning deployments.
- **Addition of a Human-in-the-loop (HITL) into the development lifecycle:** Besides, some of the steps in the framework are performed manually (such as results analysis and data exploration). We plan to refine our human-in-the-loop design to enhance our tool's judgment, contextual understanding, transparency, explainability, and adaptability. This

[2]We will use the terms interpretability and explainability (XAI) interchangeably, referring to "the ability to explain or to present in understandable terms to a human" (Doshi-Velez and Kim, 2017).

can be achieved after applying community assessment and getting feedback on the current approach.

- **Adapting our solution to meet the FAIR principles**: The FAIR principles provide guidelines to improve the **F**indability, **A**ccessibility, **I**nteroperability, and **R**euse of digital assets (Wilkinson et al., 2016). Besides our open-source implementation being able to reproduce all experiments covered in this research, to enhance our implementation as a tool for more complex datasets and contexts, we plan to extend our result assets by providing metadata to assist the workflows for analysis, storage, and processing, improving the solution traceability.

Bibliography

Abebe, R., Hardt, M., Jin, A., Miller, J., Schmidt, L., and Wexler, R. (2022). Adversarial scrutiny of evidentiary statistical software. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 1733–1746, New York, NY, USA. Association for Computing Machinery.

Abid, A., Farooqi, M., and Zou, J. (2021). Persistent anti-muslim bias in large language models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, page 298–306, New York, NY, USA. Association for Computing Machinery.

Agarwal, A., Beygelzimer, A., Dudik, M., Langford, J., and Wallach, H. (2018). A reductions approach to fair classification. In Dy, J. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 60–69. PMLR.

Alves, G., Bernier, F., Couceiro, M., Makhlouf, K., Palamidessi, C., and Zhioua, S. (2023). Survey on fairness notions and related tensions. *EURO Journal on Decision Processes*, 11:100033.

Balayn, A., Lofi, C., and Houben, G.-J. (2021). Managing bias and unfairness in data for decision support: a survey of machine learning and data engineering approaches to identify and mitigate bias and unfairness within data management and analytics systems. *The VLDB Journal*, 30:1–30.

Barocas, S., Hardt, M., and Narayanan, A. (2023). *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press.

Beiró, M. G. and Kalimeri, K. (2022). Fairness in vulnerable attribute prediction on social media. *Data Min. Knowl. Discov.*, 36(6):2194–2213.

Bellamy, R. K. E., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., Nagar, S., Ramamurthy, K. N., Richards, J., Saha, D., Sattigeri, P., Singh, M., Varshney, K. R., and Zhang, Y. (2019). AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *IBM Journal of Research and Development*, 63(4/5):4:1–4:15.

Benatti, R., Severi, F., Avila, S., and Colombini, E. L. (2024). Gender bias detection in court decisions: A brazilian case study. In *Proceedings of the 2024 ACM Conference on Fair-*

ness, Accountability, and Transparency, FAccT '24, page 746–763, New York, NY, USA. Association for Computing Machinery.

Benjamin, R. (2019). *Race After Technology: Abolitionist Tools for the New Jim Code*. Polity.

Bergman, A. S., Hendricks, L. A., Rauh, M., Wu, B., Agnew, W., Kunesch, M., Duan, I., Gabriel, I., and Isaac, W. S. (2023). Representation in ai evaluations. *Conference on Fairness, Accountability and Transparency*.

Berk, R., Heidari, H., Jabbari, S., Kearns, M., and Roth, A. (2021). Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 50(1):3–44.

Binns, R. (2018). Fairness in machine learning: Lessons from political philosophy. In Friedler, S. A. and Wilson, C., editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 149–159. PMLR.

Bird, S., Dudík, M., Edgar, R., Horn, B., Lutz, R., Milan, V., Sameki, M., Wallach, H., and Walker, K. (2020). Fairlearn: A toolkit for assessing and improving fairness in ai. Technical Report MSR-TR-2020-32, Microsoft.

Bitencourt, L. and Ansel, P. (2024). Novas tecnologias, antigos interesses: sistema de reconhecimento facial e genocídio negro na bahia. Nexo Políticas Públicas. [Online; accessed 12/05/2024].

Buolamwini, J. and Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In Friedler, S. A. and Wilson, C., editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 77–91. PMLR.

Carey, A. N. and Wu, X. (2022). The statistical fairness field guide: perspectives from social and formal sciences. *AI and Ethics*, 3(1):1–23.

Castelnovo, A., Crupi, R., Greco, G., Regoli, D., Penco, I. G., and Cosentini, A. C. (2022). A clarification of the nuances in the fairness metrics landscape. *Scientific Reports*, 12(1).

Caton, S. and Haas, C. (2024). Fairness in machine learning: A survey. *ACM Comput. Surv.*, 56(7).

Chen, P., Wu, L., and Wang, L. (2023). Ai fairness in data management and analytics: A review on challenges, methodologies and applications. *Applied Sciences*, 13(18).

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 5(2):153–163. PMID: 28632438.

Chouldechova, A. and Roth, A. (2018). The frontiers of fairness in machine learning. *ArXiv*.

Corbett-Davies, S., Gaebler, J. D., Nilforoshan, H., Shroff, R., and Goel, S. (2023). The measure and mismeasure of fairness. *Journal of Machine Learning Research*, 24(1).

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., and Huq, A. (2017). Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, page 797–806, New York, NY, USA. Association for Computing Machinery.

Crawford, K. (2013). The hidden biases in big data. Harvard Business Review. [Online; accessed 04/25/2024].

Dastin, J. (2018). Amazon scraps secret ai recruiting tool that showed bias against women. Reuters.com. [Online; accessed 04/25/2024].

DiCiccio, C., Hsu, B., Yu, Y., Nandy, P., and Basu, K. (2023). Detection and mitigation of algorithmic bias via predictive parity. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, page 1801–1816, New York, NY, USA. Association for Computing Machinery.

Ding, F., Hardt, M., Miller, J., and Schmidt, L. (2021). Retiring adult: New datasets for fair machine learning. *Advances in Neural Information Processing Systems*, 34.

Doshi-Velez, F. and Kim, B. (2017). Towards a rigorous science of interpretable machine learning.

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

FAccT, A. (2018). Acm conference on fairness, accountability, and transparency (acm facct).

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., and Venkatasubramanian, S. (2015). Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, page 259–268, New York, NY, USA. Association for Computing Machinery.

Ferrara, E. (2024). Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies. *Sci*, 6(1).

Friedler, S. A., Scheidegger, C., and Venkatasubramanian, S. (2021). The (im)possibility of fairness: different value systems require different mechanisms for fair decision making. *Commun. ACM*, 64(4):136–143.

Hajian, S., Bonchi, F., and Castillo, C. (2016). Algorithmic bias: From discrimination discovery to fairness-aware data mining. *KDD*.

Hardt, M., Price, E., Price, E., and Srebro, N. (2016). Equality of opportunity in supervised learning. In Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Jones, G. P., Hickey, J. M., Stefano, P. G. D., Dhanjal, C., Stoddart, L. C., and Vasileiou, V. (2020). Metrics and methods for a systematic comparison of fairness-aware machine learning algorithms. *ArXiv*.

Kamiran, F. and Calders, T. (2011). Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33:1 – 33.

Kamiran, F., Karim, A., and Zhang, X. (2012). Decision theory for discrimination-aware classification. In *2012 IEEE 12th International Conference on Data Mining*, pages 924–929.

Kamishima, T., Akaho, S., and Sakuma, J. (2011). Fairness-aware learning through regularization approach. In *2011 IEEE 11th International Conference on Data Mining Workshops*, pages 643–650.

Kilbertus, N., Rojas-Carulla, M., Parascandolo, G., Hardt, M., Janzing, D., and Schölkopf, B. (2017). Avoiding discrimination through causal reasoning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 656–666, Red Hook, NY, USA. Curran Associates Inc.

Kleinberg, J., Mullainathan, S., and Raghavan, M. (2017). Inherent Trade-Offs in the Fair Determination of Risk Scores. In Papadimitriou, C. H., editor, *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*, volume 67 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 43:1–43:23, Dagstuhl, Germany. Schloss Dagstuhl – Leibniz-Zentrum für Informatik.

Kordzadeh, N. and Ghasemaghaei, M. (2022). Algorithmic bias: review, synthesis, and future research directions. *European Journal of Information Systems*, 31(3):388–409.

Kusner, M. J., Loftus, J., Russell, C., and Silva, R. (2017). Counterfactual fairness. *Advances in neural information processing systems*, 30.

Larson, J., Mattu, S., Kirchner, L., and Angwin, J. (2016). Machine bias - how we analyzed the compas recidivism algorithm. ProPublica.org. [Online; accessed 08/22/2024].

Lee, M. S. A. and Singh, J. (2021). Risk identification questionnaire for detecting unintended bias in the machine learning development lifecycle. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, page 704–714, New York, NY, USA. Association for Computing Machinery.

Lundberg, S. and Lee, S.-I. (2017). A unified approach to interpreting model predictions.

Mahoney, T., Varshney, K., and Hind, M. (2020). *AI Fairness; How to Measure and Reduce Unwanted Bias in Machine Learning*. O'Reilly Media Company.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Comput. Surv.*, 54(6).

Meijer, A. and Wessels, M. (2019). Predictive policing: Review of benefits and drawbacks. *International Journal of Public Administration*, 42(12):1031–1039.

Mitchell, M., Baker, D., Moorosi, N., Denton, E., Hutchinson, B., Hanna, A., Gebru, T., and Morgenstern, J. (2020). Diversity and inclusion metrics in subset selection. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20, page 117–123, New York, NY, USA. Association for Computing Machinery.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., and Gebru, T. (2019). Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 220–229, New York, NY, USA. Association for Computing Machinery.

Mitchell, S., Potash, E., Barocas, S., D'Amour, A., and Lum, K. (2018). Prediction-based decisions and fairness: A catalogue of choices, assumptions, and definitions. *arXiv: Applications*.

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., and Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, 3(2):2053951716679679.

Morley, J., Kinsey, L., Elhalal, A., Garcia, F., Ziosi, M., and Floridi, L. (2023). Operationalising ai ethics: barriers, enablers and next steps. *AI & SOCIETY*, pages 1–13.

Narayanan, A. (2018). 21 fairness definitions and their politics. In *tutorial at Conference on Fairness, Accountability, and Transparency, New york, USA*, volume 1170, page 3.

Nicoletti, L. and Bass, D. (2023). Humans are biased. generative ai is even worse. Bloomberg.com. [Online; accessed 08/22/2024].

Nunes, P. (2023). O que é racismo algorítmico? qual seu impacto no campo da segurança pública? Nexo Políticas Públicas, Centro de Estudos de Segurança e Cidadania (CESeC). [Online; accessed 12/05/2024].

Obermeyer, Z., Powers, B., Vogeli, C., and Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453.

O'Neil, C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown Publishing Group, USA.

Otterbacher, J., Bates, J., and Clough, P. (2017). Competent men and warm women: Gender stereotypes and backlash in image search results. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, page 6620–6631, New York, NY, USA. Association for Computing Machinery.

Pagano, T. P., Loureiro, R. B., Lisboa, F. V. N., Peixoto, R. M., Guimarães, G. A. S., Cruz, G. O. R., Araujo, M. M., Santos, L. L., Cruz, M. A. S., Oliveira, E. L. S., Winkler, I., and Nascimento, E. G. S. (2023). Bias and unfairness in machine learning models: A systematic review on datasets, tools, fairness metrics, and identification and mitigation methods. *Big Data and Cognitive Computing*, 7(1).

Park, J. H., Lee, D. K., Kang, H., Kim, J. H., Nahm, F. S., Ahn, E., In, J., Kwak, S. G., and Lim, C.-Y. (2022). The principles of presenting statistical results using figures. *Korean journal of anesthesiology*, 75(2):139–150.

Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., and Weinberger, K. Q. (2017). On fairness and calibration. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Raji, I. D., Gebru, T., Mitchell, M., Buolamwini, J., Lee, J., and Denton, E. (2020a). Saving face: Investigating the ethical concerns of facial recognition auditing. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20, page 145–151, New York, NY, USA. Association for Computing Machinery.

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., and Barnes, P. (2020b). Closing the ai accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 33–44, New York, NY, USA. Association for Computing Machinery.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 1135–1144.

Saleiro, P., Kuester, B., Stevens, A., Anisfeld, A., Hinkson, L., London, J., and Ghani, R. (2018). Aequitas: A bias and fairness audit toolkit. *arXiv preprint arXiv:1811.05577*.

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

Srivastava, M., Heidari, H., and Krause, A. (2019). Mathematical notions vs. human perception of fairness: A descriptive approach to fairness for machine learning. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, page 2459–2468, New York, NY, USA. Association for Computing Machinery.

Tubella, A. A., Barsotti, F., Koçer, R. G., and Mendez, J. A. (2022). Ethical implications of fairness interventions: what might be hidden behind engineering choices? *Ethics and Information Technology*, 24(1):12.

Ueda, D., Kakinuma, T., Fujita, S., Kamagata, K., Fushimi, Y., Ito, R., Matsui, Y., Nozaki, T., Nakaura, T., Fujima, N., Tatsugami, F., Yanagawa, M., Hirata, K., Yamada, A., Tsuboyama, T., Kawamura, M., Fujioka, T., and Naganawa, S. (2024). Fairness of artificial intelligence in healthcare: review and recommendations. *Japanese Journal of Radiology*, 42(1):3–15.

v. DeStefano, R. (2009). Ricci v. destefano. `https://supreme.justia.com/cases/federal/us/557/557/`.

Verma, S. and Rubin, J. (2018). Fairness definitions explained. In *Proceedings of the International Workshop on Software Fairness*, FairWare '18, page 1–7, New York, NY, USA. Association for Computing Machinery.

Vilarino, R. and Vicente, R. (2020). An experiment on the mechanisms of racial bias in ml-based credit scoring in brazil. *CoRR*, abs/2011.09865.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., et al. (2016). The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3(1):1–9.

Xiang, A. (2022). Being 'Seen' vs. 'Mis-Seen': Tensions between Privacy and Fairness in Computer Vision. *Harvard Journal of Law & Technology*.

Yurochkin, M., Bower, A., and Sun, Y. (2020). Training individually fair ml models with sensitive subspace robustness. In *International Conference on Learning Representations*.

Zafar, M. B., Valera, I., Gomez Rodriguez, M., and Gummadi, K. P. (2017). Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, page 1171–1180, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., and Dwork, C. (2013). Learning fair representations. In Dasgupta, S. and McAllester, D., editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 325–333, Atlanta, Georgia, USA. PMLR.

Zhang, B. H., Lemoine, B., and Mitchell, M. (2018). Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, page 335–340, New York, NY, USA. Association for Computing Machinery.

Zhang, J., Shu, Y., and Yu, H. (2023). Fairness in design: A framework for facilitating ethical artificial intelligence designs. *International Journal of Crowd Science*, 7(1):32–39.

Zhang, Y. and Zhou, L. (2019). Fairness assessment for artificial intelligence in financial industry. *Robust AI in FS 2019 : NeurIPS 2019 Workshop on Robust AI in Financial Services: Data, Fairness, Explainability, Trustworthiness, and Privacy*.

Zliobaite, I. (2015). On the relation between accuracy and fairness in binary classification. In *The 2nd workshop on Fairness, Accountability, and Transparency in Machine Learning (FATML) at ICML'15*.

Appendix A

Additional Experimental Results

These additional experiments were performed using the same implementation as the main case study. Considering "race" as a sensitive attribute. The data for this task was also obtained from the ACS Census Data. Our sensitive feature, *race*, has nine classes in all types of data from the census, but for our extra experiment, we considered only the largest sensitive groups (*White* and *Black*) to have binary sensitive values. The following task aims to predict employability, and we provide the three fairness perspectives from the framework outcomes. For these additional experiments, the model used for training and our baseline were reduced to an XGBoost Classifier. The preprocessing follows the same as our main case study, using 10-fold cross-validation, and the evaluation and interventions follow the ten-fold as iterations.

The dataset includes various demographic, socioeconomic, and employment-related features for individuals in the US. Comprising the attributes such as *age, gender, race, education level, and marital status*, alongside with employment-related variables, such as *employment status, class of worker, occupation, hours worked per week, personal income, place of birth*. Figure A.1 shows the label distribution in the data and the label distribution concerning our sensitive feature. This dataset comprised 63.677 data points.

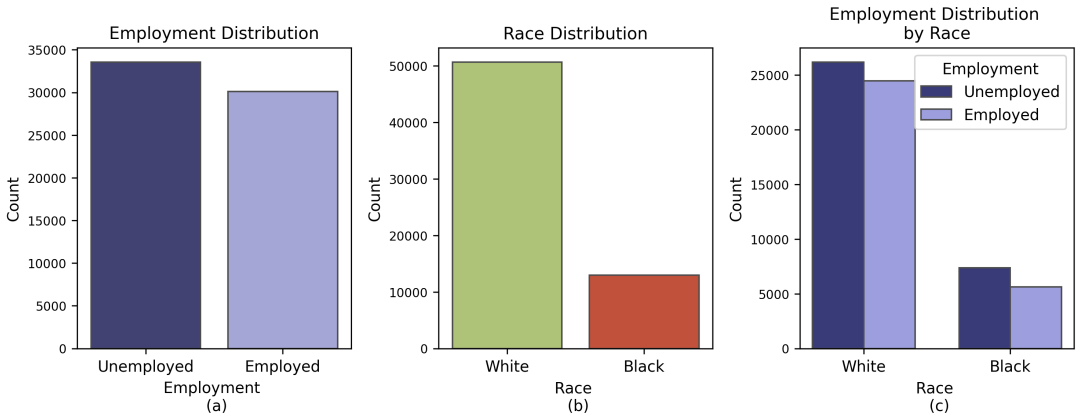


Figure A.1: Dataset labels distribution. (a) Distribution of the label between employed and unemployed. (b) The distribution between the White and Black individuals (our protected feature). (c) Label distribution across the sensitive attribute.

A.1 Comparison Between Fairness Perspectives: Employment Task

Following, we briefly contrast the results obtained regarding the prediction of employability with respect to discrimination based on race attributes, black and white individuals.

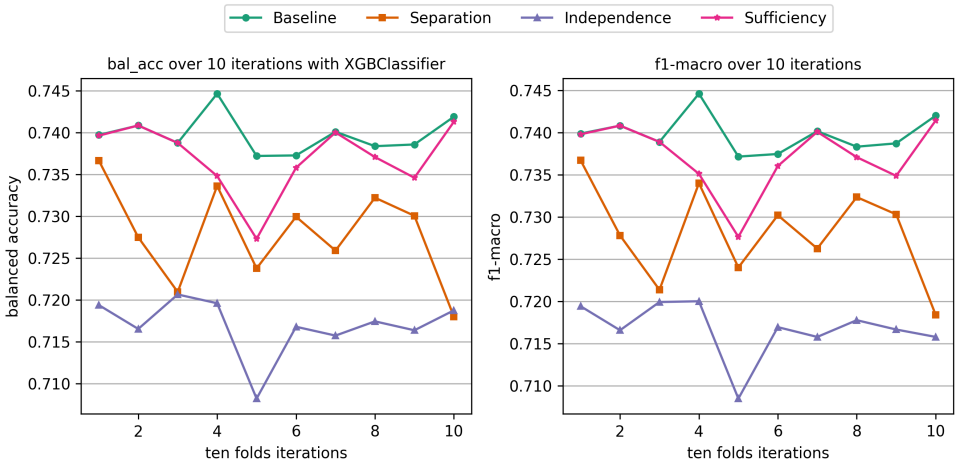
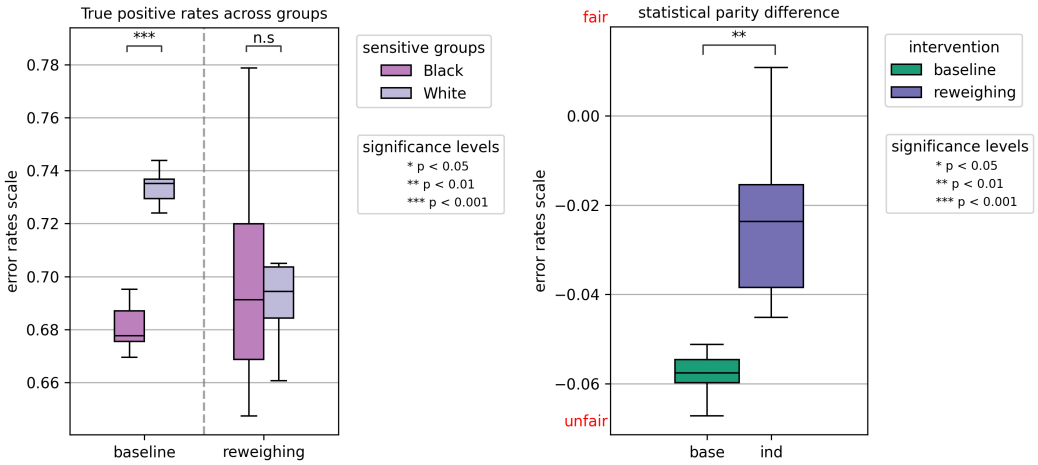


Figure A.2: Evaluation Performance: baseline model and all intervention approaches over ten iterations. Each iteration was performed through the 10-fold, and the results presented used the models evaluated using the test data (unseen data).

Following, we summarize the outcomes regarding each perspective of fairness provided by the framework application, using the respective metrics to quantify the expected definition of fairness.

The **Reweighting** (Kamiran and Calders, 2011) technique addresses fairness by balancing the acceptance rates between the groups, aiming to achieve the statistical parity/Independence definition. Figure A.3 shows the results of this approach. At baseline, the model systematically favors the *White* group in terms of the true positive rate (more likely to be correctly classified as employed), with statistically significant ($p < .001$), indicating a notable disparity in treatment between the groups. The overall likelihood of being classified as employed (as noted in the statistical parity difference) also shows the same disadvantages for the unprivileged group, with lower chances of being classified as employed. After applying reweighting intervention, the TPRs for both groups converge to more balanced outcomes, without a statistically significance difference (n.s.).

In the **Threshold Optmization** (Hardt et al., 2016), the intervention aimed to achieve the separation criteria. Figure A.4 shows the results of this fairness approach. In the baseline, the *Black* individuals were more at a disadvantage. At the same time, the *White* group received the privilege of receiving more positive labels (as employed) when they were actually unemployed. After the intervention, the false positives become nearly the same and more bal-



(a) TPR across Sensitive Groups

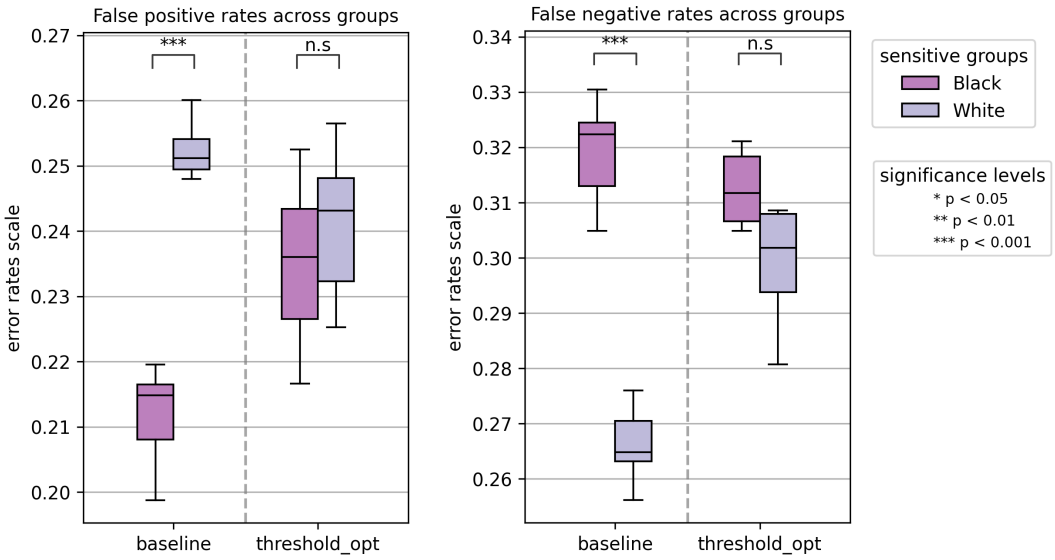
(b) Fairness Metric: Baseline vs. Reweighing (Independence)

Figure A.3: (a) True positive error rates across sensitive groups. After applying the reweighing, a comparison is made between the baseline model and the results, with averages over ten iterations. (b) Statistical Parity Difference measures the difference between the proportions of positive outcomes for two groups. A more negative value of Statistical Parity Difference indicates that the unprivileged group is at a disadvantage, while a more positive value indicates that the privileged group is at a disadvantage.

anced, indicating a minimization in the disparity across the groups, with the difference in the labels no longer statistically significant. Regarding the false negatives, the black individuals in the baseline were more likely misclassified as unemployed when they were actually employed, again showing that white groups were receiving more privilege of lower false negative errors. After the intervention, the black individuals reduced the false negative, treating both groups more equally, with no longer statistical significance (n.s.), indicating an effective reduction of unfairness.

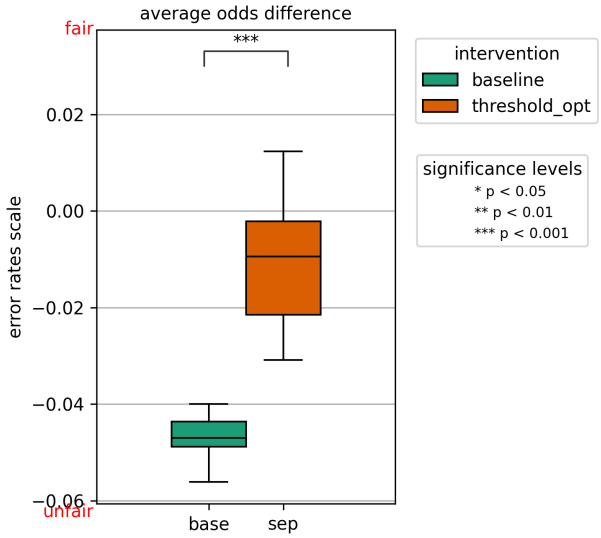
Regarding the **Calibration via Information Withholding** (Pleiss et al., 2017) to preserve Calibration, achieving Sufficiency. Figure A.5. Both groups receive similar accuracy in predicting employment at the baseline and after the calibration attempt with no statistically significant difference (n.s.). Since the intervention was used as a cost constraint *weight* to maintain a balance of the false positives and false negatives, the PPV values were already balanced and well-calibrated, and the intervention did not affect the positive predictive values. But, when it comes to the negative predictive values (NPV), the model more accurately predicts unemployment for the group of *White* individuals, and the difference is statistically significance ($p < .01$), showing a bias favoring the White group in predicting unemployment. After the intervention method, the NPVs between the *Black* and *White* groups converge with no significant difference (n.s.). This indicates that the calibration method successfully reduced the unfair advantage previously held by the *White* group in predicting unemployment, making this fairness definition also resulting in more equitable outcomes for both groups.

All three interventions for this task, concerning race as a sensitive attribute, show im-



(a) FPR across sensitive groups

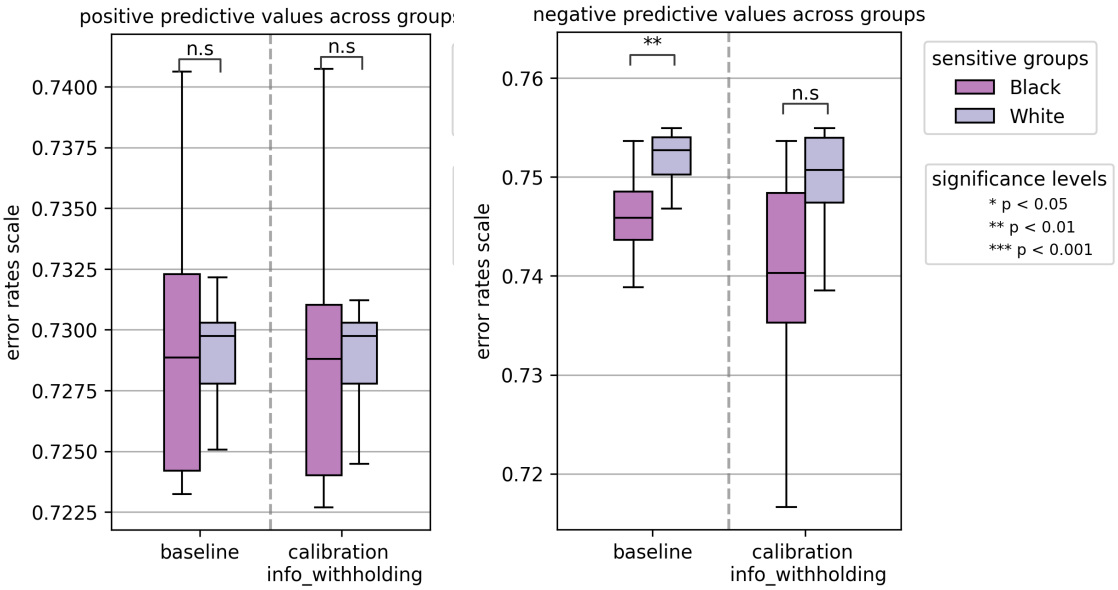
(b) FNR across sensitive groups



(c) Average odds difference between unprivileged and privileged groups.

Figure A.4: (a) False positive and (b) negative error rates across sensitive groups. (c) This fairness metric measures the disparity between false positives and true positives rates across sensitive groups. A lower absolute value, closer to zero, indicates better fairness. A comparison is made between the baseline model and the results after applying the threshold optimizer, which averages over ten iterations.

provements in mitigating bias. Each of them follows a different perspective, providing for the final user options to be considered.



(a) PPV across sensitive groups

(b) NPV across sensitive groups

Figure A.5: (a) Positive predictive values and (b) negative predictive values across sensitive groups. A comparison is made between the baseline and the weighted calibrated models after the attempt to maintain calibration between groups, which averages over ten iterations.

Appendix B

Framework as a Tool: Usability

- The code itself is straightforward to follow but requires a Python understatement.
- In the open-source repository[1], the user can find other ways to add or read new datasets documented in the code.
- The implemented interventions for each fairness notion can be used in any scenario or task where the data is tabular.
- There is a folder called *dags* in the public repository, and to add a new workflow for a different dataset, the example of dag available should be duplicated. The source of data should be provided as a parameter.
- At the same `example_pipeline` file, the user can decide the models to be trained and the models to be evaluated; the implemented list comprises five supervised machine learning models and can be expanded following the same path as the existing ones.
- To broaden the notion of fairness, the user can follow the same structure used for Independence, Separation, and Sufficiency.

[1] `https://github.com/equity-ai-hub/ai-system-framework`

Appendix C

ACS Dataset Class Python Implementation

```python
import pandas as pd
from folktables import ACSDataSource, ACSEmployment, ACSIncome
from sklearn.preprocessing import StandardScaler
import src.utils.data_helper as helper

# 1.White, 2.Black, 3.Asian, 4.America Native and Alaska Native,
# 5.Some Other, 6.Two or More
def group_race(x):
    if x == 3.0 or x == 4.0 or x == 5.0 or x == 7.0:
        return 4.0  # America Native and Alaska Native
    if x == 6.0:
        return 3.0  # Asian
    if x == 8.0:
        return 5.0  # Some Other
    if x == 9.0:
        return 6.0  # Two or More
    else:
        return x

class ACSDataset:
    def __init__(self, survey_year="2014", US_states=["CA", "TX"],
        horizon="1-Year", survey="person",
    ):
        self.survey_year = survey_year
        self.horizon = horizon
        self.survey = survey
        self.states = US_states

    def task_task(self, task_name: str):
        if task_name == "employment":
            return ACSEmployment
        elif task_name == "income":
            return ACSIncome
        else:
            raise AttributeError("Attribute not found")
```

```
36
37      def get_data(self, download=True,
38          task_name="employment", return_type="csv",
39      ):
40          data_source = ACSDataSource(
41              survey_year=self.survey_year,
42              horizon=self.horizon,
43              survey=self.survey
44          )
45          states_data = data_source.get_data(
46              states=self.states, download=download
47          )
48
49          acs_task = self.task_task(task_name)
50          features, labels, _ = acs_task.df_to_pandas(states_data)
51          features["RACE"] = features["RAC1P"].apply(
52              lambda x: group_race(x)
53          )
54          features = features.drop(columns=["RAC1P"])
55
56          # raw labels are boolean, convert them to int
57          labels = labels.astype(int)
58          features["LABELS"] = labels
59
60          # check if there is duplications and remove them
61          features = features.drop_duplicates()
62
63          if return_type == "dataframe":
64              return features
65          elif return_type == "csv":
66              features_obj = features.to_csv(
67                  index=False
68              ).encode("utf-8")
69              return features_obj
70          else:
71              return "return_type not found"
72
73      def split_data(self, df: pd.DataFrame, test_size: float = 0.2,
74          random_state: int = 42, stratify=None, dtype=None
75      ):
76          from sklearn.model_selection import train_test_split
77          train, test = train_test_split(df, test_size=test_size,
78              random_state=random_state, stratify=stratify
79          )
80
81          if dtype == "csv":
82              return train.to_csv(index=False).encode(
```

```
                "utf-8"
            ), test.to_csv(index=False).encode("utf-8")
        else:
            return train, test

    def preprocess_data(
        self,
        df: pd.DataFrame,
        categorical_features: list = [],
        dtype=None,
    ):
        features_list_not_labels = df.columns.to_list()[:-1]
        df = df.drop_duplicates()

        if categorical_features and not set(
            categorical_features
        ).issubset(set(features_list_not_labels)):
            raise ValueError(
                "Categorical features not found in the dataset"
            )

        if categorical_features:
            df = helper.one_hot_encode(
                df, categorical_features
            )
            df = df.drop(columns=categorical_features)

        continuous_features = []
        for feature in features_list_not_labels:
            if feature not in categorical_features:
                continuous_features.append(feature)

        scale = StandardScaler()
        df[continuous_features] = scale.fit_transform(
            df[continuous_features]
        )

        if dtype == "csv":
            return df.to_csv(index=False).encode("utf-8")
        else:
            return df
```

Listing C.3: ACS Dataset class implementation in Python

Appendix D

Model Class Python Implementation

```python
from sklearn.ensemble import RandomForestClassifier
from sklearn.linear_model import LogisticRegression
from sklearn.neural_network import MLPClassifier
from sklearn.svm import SVC
from sklearn.tree import DecisionTreeClassifier
from xgboost import XGBClassifier

class Model:
    def __init__(self, model_name, dataset_id, df_one_hot,
        df_preproc, target, sensitive_attr,
    ):
        self.df_one_hot = df_one_hot
        self.df_preproc = df_preproc
        self.scores = None
        self.model = None
        self.evalution_scores = None
        self.target = target
        self.sensitive_attr = sensitive_attr
        self.dataset_id = dataset_id
        self._initialize_model(model_name)

    def _initialize_model(self, model_name: str):
        models = {
            "logistic_regression": LogisticRegression(
                solver="liblinear", max_iter=1000
            ),
            "mlp": MLPClassifier(
                hidden_layer_sizes=(5,), max_iter=1000,
                alpha=0.01, random_state=42,
            ),
            "random_forest": RandomForestClassifier(),
            "svm": SVC(),
            "xgboost": XGBClassifier(),
            "decision_tree": DecisionTreeClassifier(),
        }
```

```
36          if model_name in models:
37              self.model = models[model_name]
38
39      def train(self, n_folds: int = 10, data_dir: str = None,
40          sample_weight: Any = None,):
41          """Train the model using StratifiedKFold cross validation
42          and export the model and training scores.
43          Return the best model based on the average accuracy score.
44          """
45          # get the original preprocessed dataset without one-hot
46          # encoding for folds data visualization
47          X_preproc, y_preproc = (
48              self.df_preproc.drop(self.target, axis=1),
49              self.df_preproc[self.target],
50          )
51          X, y = (
52              self.df_one_hot.drop(self.target, axis=1),
53              self.df_one_hot[self.target],
54          )
55
56          skf = StratifiedKFold(n_splits=n_folds,shuffle=True,random_state=42)
57
58          if self.model is None:
59              raise ValueError("Model has not been initialized")
60
61          for train_idx, val_idx in skf.split(X, y):
62              n_folds = n_folds - 1
63              X_fold_train, X_fold_val = (X.iloc[train_idx], X.iloc[val_idx])
64              y_fold_train, y_fold_val = (y.iloc[train_idx], y.iloc[val_idx])
65
66              X_preproc_fold_train, X_preproc_fold_val = (
67                  X_preproc.iloc[train_idx], X_preproc.iloc[val_idx]
68              )
69              y_preproc_fold_train, y_preproc_fold_val = (
70                  y_preproc.iloc[train_idx], y_preproc.iloc[val_idx]
71              )
72
73              self.model.fit(X_fold_train, y_fold_train,
74                  sample_weight=sample_weight,
75              )
76
77              y_fold_pred = self.model.predict(X_fold_val)
78              try:
79                  y_fold_predict_proba = self.model.predict_proba(
80                      X_fold_val
81                  )[:, 1]
```

```
            except AttributeError:
                y_fold_predict_proba = np.zeros(len(y_fold_pred))

            y_pred = pd.Series(y_fold_pred, name="y_pred",
                index=X_fold_val.index
            )
            y_pred_proba = pd.Series(y_fold_predict_proba,
                name="y_pred_proba", index=X_fold_val.index
            )

            # save all models artifacts for each fold if
            # data_dir is not None
            val_fold_preproc = pd.concat(
                [
                    X_preproc_fold_val,
                    y_preproc_fold_val,
                    y_pred,
                    y_pred_proba,
                ],
                axis=1,
            )
            val_fold_oh = pd.concat(
                [X_fold_val, y_fold_val, y_pred], axis=1
            )

            train_fold_preproc = pd.concat(
                [
                    X_preproc_fold_train,
                    y_preproc_fold_train,
                ],
                axis=1,
            )
            train_fold_oh = pd.concat(
                [X_fold_train, y_fold_train], axis=1
            )

            # metrics = Metrics.calculate_metrics(y_fold_val, y_fold_pred)
            scores = Metrics.metrics_scores_aif360(
                df=val_fold_preproc, y_pred=y_fold_pred
            )
            self.scores = {
                "model_name": self.model.__class__.__name__,
                "kfold": n_folds,
                # "sampling_method": sampling_method,
                "scores": scores,
            }
```

```
            if data_dir is not None:
                print(
                    f"Exporting model artifacts to {data_dir}"
                )
                self._export_training_artifacts(
                    data_dir,
                    train_fold_oh,
                    train_fold_preproc,
                    val_fold_oh,
                    val_fold_preproc,
                    n_folds,
                )
```

Listing D.4: Model Class Python Implementation

Appendix E

Dataset Features & EDA Report

```
"COW": {
    1.0:"Employee of a private for-profit company or"
        "business, or of an individual, for wages,"
        "salary, or commissions",
    2.0:"Employee of a private not-for-profit, tax-exempt,"
        "or charitable organization",
    3.0: "Local government employee (city, county, etc.)",
    4.0: "State government employee",
    5.0: "Federal government employee",
    6.0:"Self-employed in own not incorporated business,"
        "professional practice, or farm",
    7.0:"Self-employed in own incorporated business,"
        "professional practice or farm",
    8.0: "Working without pay in family business or farm",
    9.0: "Unemployed and last worked 5 years ago or earlier or never worked",
},
"SCHL": {
    1.0: "No schooling completed",
    2.0: "Nursery school, preschool",
    3.0: "Kindergarten", 4.0: "Grade 1", 5.0: "Grade 2",
    6.0: "Grade 3", 7.0: "Grade 4", 8.0: "Grade 5", 9.0: "Grade 6",
    10.0: "Grade 7", 11.0: "Grade 8", 12.0: "Grade 9",
    13.0: "Grade 10", 14.0: "Grade 11", 15.0: "12th grade - no diploma",
    16.0: "Regular high school diploma",
    17.0: "GED or alternative credential",
    18.0: "Some college, but less than 1 year",
    19.0: "1 or more years of college credit, no degree",
    20.0: "Associate's degree",
    21.0: "Bachelor's degree",
    22.0: "Master's degree",
    23.0: "Professional degree beyond a bachelor's degree",
    24.0: "Doctorate degree",
},
"MAR": {
    1.0: "Married", 2.0: "Widowed", 3.0: "Divorced",
```

```
    4.0: "Separated", 5.0: "Never married or under 15 years old",
},
"SEX": {1.0: "Male", 2.0: "Female"},
"RAC1P": {
    1.0: "White alone",
    2.0: "Black or African American alone",
    3.0: "American Indian alone",
    4.0: "Alaska Native alone",
    5.0:"American Indian and Alaska Native tribes specified;"
        "or American Indian or Alaska Native, not specified and no other",
    6.0: "Asian alone",
    7.0: "Native Hawaiian and Other Pacific Islander alone",
    8.0: "Some Other Race alone",
    9.0: "Two or More Races",
}
```

Listing E.5: Dataset Features Dictionary

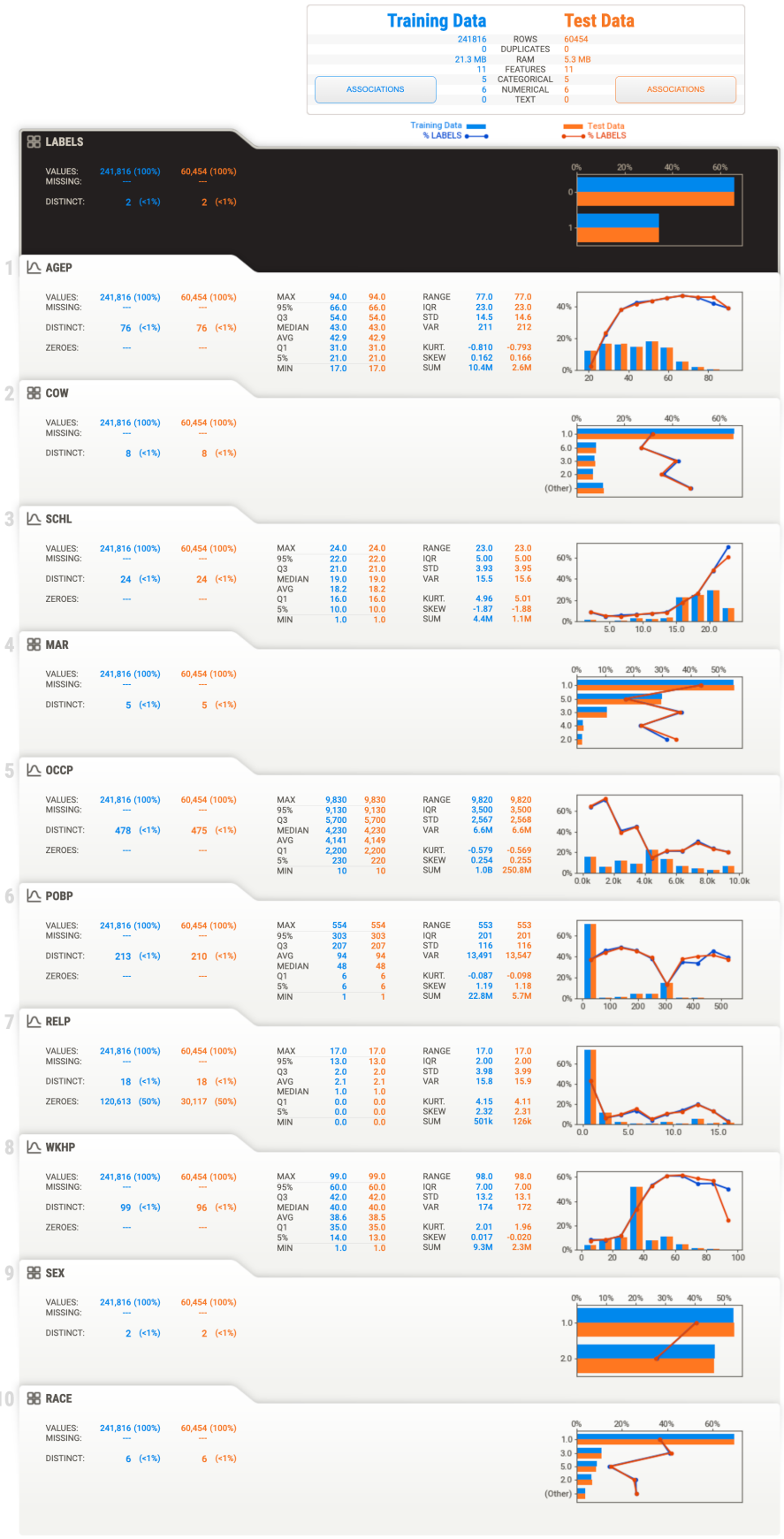


Figure E.1: Dataset exploratory data analysis full report generate via SweetViz tool