Ioana-Silvia Serban, Bsc

Intersectional Bias in Public Datasets: Exploring Detection and Mitigation Strategies

Master's Thesis

to achieve the university degree of
Diplom-Ingenieur/ Master of Science

Master's degree programme:
Software Engineering and Management

submitted to

Graz University of Technology

Supervisors:
Priv.-Doz. Dipl.Ing. Dr.techn. Dominik Kowald, BSc.,
Dipl.Ing. Dr. techn. Simone Kopeinik, BSc.

Institute of Interactive Systems and Data Science
Head: Univ.-Prof. Dipl.-Ing. Dr.techn. Frank Kappe

Graz, December, 2024

EIDESSTATTLICHE ERKLÄRUNG

Ich erkläre an Eides statt, dass ich die vorliegende Arbeit selbstständig verfasst, andere als die angegebenen Quellen/Hilfsmittel nicht benutzt, und die den benutzten Quellen wörtlich und inhaltlich entnommenen Stellen als solche kenntlich gemacht habe.

Graz, am …………………………… ………………………………………………..
(Unterschrift)

STATUTORY DECLARATION

I declare that I have authored this thesis independently, that I have not used other than the declared sources / resources, and that I have explicitly marked all material which has been quoted either literally or by content from the used sources.

…………………………… ………………………………………………..
date (signature)

Acknowledgments

I would like to thank Dominik Kowald and Simone Kopeinik for guiding and supporting me throughout the work of this Master's thesis. This work was funded by the strategic FFG COMET project KnowCertiFAI of Know Center Research GmbH.

Abstract

Machine learning systems have increasingly become integrated into decision-making processes across various domains and their outcomes can significantly impact individuals' lives. These systems can perpetuate unfair bias, especially towards subpopulations belonging to certain gender, race, or ethnicities, which are referred to as sensitive attributes. While tools and algorithms have been developed to address such biases and mitigate them, a newer and less-explored area of intersectional fairness has gained more attention. This area examines bias that can occur at the intersection of multiple sensitive attributes. However, the intersection of numerous sensitive attributes can lead to an excessive number of obtained subgroups, challenging the detection of bias.

This Master's thesis investigates intersectional bias across four datasets that are publicly available and evaluates the effectiveness of bias mitigation methods in this context. To address the challenge of selecting the attributes for intersectional analysis, a detection algorithm is utilized to detect the attributes that are at the highest risk of discrimination. The results reveal that fairness is decreased in an intersectional setting compared to a non-intersectional one. Moreover, the findings demonstrate that bias mitigation methods can increase fairness, however, their effectiveness varies across datasets and tasks. Furthermore, applying bias mitigation methods can introduce trade-offs, reducing the models' accuracy, or introducing new disparities across fairness metrics.

This work highlights the importance of addressing intersectional fairness and provides insight into the complexity of achieving it. It demonstrates the benefit of using semi-automated tools to identify high-risk intersectional groups and underscores the challenges in mitigating the detected unfair biases against these groups.

Contents

List of Figures

List of Tables

1 Introduction

Recent developments and advances in Artificial Intelligence (AI) systems enable the possibility of applying them in different domains, including those where AI is responsible for making decisions that can impact individuals' lives [4]. In such domains, sensitive attributes, which are intrinsic and fundamental characteristics of an individual's identity, play a crucial role. Since AI systems have become more prevalent in everyday life, the number of unfair bias incidents towards such sensitive attributes has increased. These biases can change the outcomes of machine learning models, leading, for example, to wrongly categorized data in the case of a classification task [5]. In the research field of AI, the issue of fairness has gained more attention as much recent research has identified several machine learning applications that create unfair predictions, especially for marginalized groups [5]. Therefore, the field of identifying unfair bias and its sources, as well as addressing it, has emerged in the last years, various methods and tools [3] offering these possibilities. Furthermore, bias mitigation methods have been developed [10, 68, 14], aiming to reduce the effects of bias and improve fairness results.

However, while there has been significant progress in identifying unfair bias as well as mitigating it, most of the works focus only on ”a single sensitive axis” such as gender (Men vs. Women) or race (Caucasian vs. Black) [49]. A more recent growing issue that is often overlooked when discussing bias in machine learning systems is the challenge of addressing intersection of sensitive attributes, also referred to as intersectionality [21]. The combination of sensitive attributes (for example gender and race) creates unique subpopulations that may experience distinct forms of bias and fairness challenges. An intersectional perspective allows for a nuanced analysis of bias, revealing hidden biases that might otherwise go unnoticed. For example, recent studies have demonstrated that machine learning models exhibiting little or no unfair bias towards individual sensitive attributes can fail to obtain the same level of fairness at the intersection of these attributes [12, 43].

Despite its importance, intersectional fairness is still a topic that is relatively unexplored in the context of AI. Work on bias detection and mitigation focuses mainly on analyzing data without incorporating intersectional groups [11], therefore failing to identify this type of bias. Existing literature on this topic focuses mainly on the intersection of at most two binary sensitive attributes [30]. In practice, data typically contains multiple protected attributes that may lead to a multitude of subgroups being at risk of discrimination.

Nevertheless, incorporating intersectional groups for fairness analysis can be challenging. The intersection of non-binary sensitive attributes can lead to a large number of subgroups which can impede this process. Therefore, the decision about which attributes to intersect is crucial [77].

The aim of this work is to investigate intersectional fairness in datasets that are publicly available by using an algorithm that ranks the most sensitive attributes in a dataset and creates intersectional groups accordingly. Moreover, this work intends to check whether unfair bias exists or is amplified within these created intersectional groups. Finally, existing bias mitigation will be applied to investigate whether these mitigations can yield fairer results.

1.1 Research Questions

To address the challenges mentioned above, this section defines the following three research questions:

- ***RQ1***: *To what extent can a semi-automated approach assist in identifying sensitive attributes and forming relevant intersectional groups within public datasets?*

  One of the challenges in considering intersectionality is choosing which attributes to include. On the one hand, defining and creating new intersectional groups by combining different sensitive attributes can lead to an excessive number of subgroups within each new group. On the other hand, ignoring attributes to reduce the amount of data can exclude important attributes that are at risk of discrimination. Therefore, this research question aims to investigate whether an existing algorithm that detects and ranks the most sensitive attributes can assist in forming intersectional groups.

- ***RQ2***: *Does the consideration of intersectional groups reveal or amplify hidden biases that are not evident in non-intersectional groups?*

  There can be the case that a machine learning model exhibits no apparent bias or little bias when analyzing the dataset without intersectional groups. However, biases may emerge when considering intersectional groups. This question aims to explore the effectiveness of bias detection methods in revealing hidden bias within the intersectional groups created, as mentioned in Q1.

- ***RQ3***: *To what extent can existing bias mitigation methods reduce unfair bias in the case of intersectional groups, and what are the implications on model performance?*

  Bias mitigation methods have been integrated into existing fairness toolkits (for example, Aequitas Flow[1], Fairlearn[2], or AIF360[3]), making them accessible to a wide range of users. However, their effectiveness has been studied in non-intersectional cases, with little work focusing on their applicability on intersectional groups [18]. This research question evaluates how such methods, available through public frameworks, perform in intersectional scenarios. Moreover, the effects on the overall model performance are investigated.

1.2 Contributions

The contributions of this Master's thesis to the current state of the art of the topic are the following:

- First, this thesis uses a detection algorithm to identify and rank the most sensitive attributes in public datasets, including those that might not traditionally be recognized as sensitive. Previous works [38, 81] have used such algorithms solely to detect the most sensitive attributes. In contrast, the approach in this Master's thesis takes the top-ranked sensitive attributes and intersects them to create meaningful intersectional groups. This method ensures a balance by considering a sufficient number of attributes for intersection while avoiding an excessive number of new subgroups. Current understanding indicates that this has not been previously attempted by other works.

- Second, it highlights the importance of considering intersectional groups in unfair bias analysis, especially when datasets have more than two sensitive attributes. The results demonstrate how unfair biases can emerge or be amplified when considering intersectionality, providing insight into hidden bias that might not be detected in non-intersectional analysis.

- Third, it evaluates the effectiveness of bias mitigation methods in the intersectional case. Moreover, it investigates the trade-offs between achieving fairness and the classification performance of a model. This analysis also

[1] https://github.com/dssg/aequitas?tab=readme-ov-file
[2] https://fairlearn.org/v0.11/userguide/mitigation/index.html
[3] https://aif360.res.ibm.com/

identifies the current limitations when applying bias mitigation methods for the intersectional case.

An overview of the steps taken to answer the research questions can be observed in Figure 1.1: first, public datasets are selected and filtered, then sensitive attributes are detected using an algorithm and based on that detection, intersectional groups are created (RQ1). Next, the datasets are evaluated for biases associated with the identified intersectional groups (RQ2), and finally, bias mitigation methods are applied to reduce or eliminate detected biases (RQ3). The results are then analyzed and discussed.

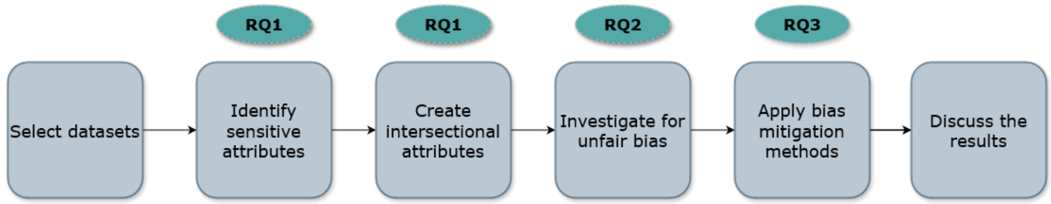

RQ1
RQ1
RQ2
RQ3
Select datasets
Identify sensitive attributes
Create intersectional attributes
Investigate for unfair bias
Apply bias mitigation methods
Discuss the results

Figure 1.1: Flowchart representing the organization of this Master's thesis.

1.3 Structure

This document is organized as follows: Chapter 2 provides foundational concepts necessary to understand the topic of unfair bias within AI, along with an overview of the current state of the art in this field. Chapter 3 outlines the approach taken to address the research questions. Next, Chapter 4 details the methodology used to investigate and answer these questions through experiments. Chapter 5 presents and discusses the results derived from these experiments, offering insights into the effectiveness of the proposed methods. Finally, Chapter 6 summarizes the findings, draws conclusions, and discusses open challenges, while in Appendix, a work-in-progress visualization tool is presented where results from this Master's thesis can be incorporated.

1.4 Terms and Definitions

To provide clarity and ensure a shared understanding, this subsection defines key terms used throughout this thesis.

Protected attributes: These are personal characteristics of individuals that are protected by laws or regulations[4] against discrimination. Examples include age, disability, race, religion, etc.

Sensitive attributes: These are characteristics of individuals that could be subject to discrimination in AI systems. These attributes may lead to biased or unfair outcomes if not handled carefully in the design or deployment of machine learning models. Sensitive attributes can overlap with protected attributes, but can also include other characteristics that, while not legally protected, are prone to unjust treatment in specific contexts (for example: education, income, etc.) [9].

Intersectional groups: These are groups obtained by intersecting sensitive attributes, such as gender and age [80]. This thesis considers only the intersection of two sensitive attributes. The categories obtained by such intersections will be referred to as intersectional subgroups (for example, men older than 50 years).

(Un)Privileged groups: Privileged and unprivileged groups refer to populations often defined by one or more sensitive attributes [15]. Privileged groups are those more likely to receive positive outcomes or classifications, while unprivileged groups are those disproportionately less likely to do so.

Binary classification: A binary classification task is a type of supervised learning problem where the goal is to categorize instances into two distinct classes [55]. These classes are usually labeled as 0 or 1 or as True or False. To classify the instances, a machine learning models is trained using labeled data, where each instance has an associated class label. The goal is to predict the class label of unseen instances, based on the learned model. For a binary classification task, the model's output represents the decision that reflects the probability of a instance belonging to a particular class (class 0 or 1). In this thesis only binary classification tasks are considered.

Bias in machine learning systems: Bias can be described as the tendency to produce unjust outcomes due to flawed assumptions. In the context of AI or machine learning, bias can reflect how models can generate errors that result in unfair or prejudiced decisions [51].

(Un)Fairness in machine learning systems: In decision-making processes, fairness refers to the impartial treatment of individuals or groups, without any favoritism based on inherent characteristics. In comparison, if a machine learning system makes decisions that favor or disadvantage a specific group, then its outcome is considered unfair [53].

Bias mitigation methods: These are methods that are developed to decrease or remove detected unfair bias from a dataset or a classifier's outcomes [53].

[4] https://www.equalityhumanrights.com/equality/equality-act-2010/protected-characteristics

2 Related Work and Background

This chapter aims to offer a better understanding of the key concepts related to fairness in AI models, as well as an overview of the current state of the art and challenges in the field. It presents topics such as sensitive attributes and their detection and selection, bias, and its sources, bias mitigation strategies, intersectionality in the context of bias in machine learning systems, and its importance.

2.1 Sensitive attributes

Investigating and mitigating unfairness in machine learning systems is centered around sensitive attributes, which are attributes that can be at risk of discrimination at any state of AI system's cycle. Common sensitive attributes include gender, age, or ethnicity, which can reflect real-world social biases or prejudice when present in datasets. The investigation of bias is based on analyzing how different bias-indicating metrics differ between the privileged group of a sensitive attribute and the unprivileged groups. Typically, the privileged group can be defined as the most represented group within an attribute, but also as the group that is socially or historically known to have more privileges (for example, white people).

Usually, the sensitive attributes are chosen by taking into consideration different anti-discrimination laws or acts [1] [2]. However, it is possible that attributes other than those typically defined as sensitive have an influence on the fairness of predictions [15]. It has been shown that even if the sensitive attributes are removed from the model training, unfairness was not eliminated entirely, due to *proxy attributes* [71, 37, 48]. Proxy attributes might not seem sensitive, but they can be related to the sensitive attributes and can contain hidden sensitive information about an individual. For example, proxy attributes can include characteristics such as income, education level, number of working hours, address, postal code, etc. [15].

One well-known term describing such cases is *redlining*, which originated in the United States after it was discovered that banks refused to invest in specific neigh-

[1] https://archive.equalityhumanrights.com/en/equality-act-2010/what-equality-act

[2] https://www.eeoc.gov/statutes/title-vii-civil-rights-act-1964

borhoods by marking them with red on maps. Although race did not directly factor into the decision making process, the neighborhoods marked with red were predominantly those that had a higher number of residents of color. Thus, attributes like address or neighborhood postal code acted as a proxy variable that indirectly influenced decisions [50].

One solution to address the issue of identifying the sensitive attributes is to detect them in an (semi-)automated way. For instance, authors from [38] proposed a method to automatically detect sensitive attributes based on user queries, assigning weights to attributes such as social security numbers, names, age, or gender. Similarly, the authors from [81] developed a system that uses regular expressions and a machine learning model to detect sensitive attributes in medical records. Despite its strengths, the authors note that their approach is limited due to both the size of the data used as well as a lack of consideration for other factors such as socioeconomic status. Moreover, the main focus of these methods is to detect and remove attributes that could identify individuals, ensuring privacy. As a result, they also prioritize identifying attributes such as names, social security numbers, or email addresses. In contrast, for the topic of bias analysis such attributes are not relevant because they pertain to individuals rather than representing characteristics shared by a group.

2.2 Unfair Bias in Machine Learning Systems

Machine learning systems have become a part of our society as they assist people in making decisions about music, movies, products, or travel. On top of that, AI systems have started to be used in domains where the decision-making action is more serious. Such domains include: hiring decisions [4], advertisements [72], bank credits [31], or public health [65]. Although one might think that the use of AI systems should make more precise and fairer decisions than human decision-makers, in reality, AI systems' results can be influenced by bias [53]. This leads to what is known as unfair bias, where the predictions made by a machine learning model unjustly differ across disadvantaged groups, such as those defined by race or gender [54].

One known example is the Correctional Offender Management Profiling for Alternative Sanctions (**COMPAS**) tool which was used in the United States to analyze the risk of a convicted person committing another crime in the future. Based on this tool, judges decided whether a person should be released earlier from prison or not. In the analysis of this software, it was discovered that for black people the false positive rates are much higher than for white people, meaning that the system considered that a black person has a higher chance of committing crime again.

For instance, the software assigned a higher risk score to a young black woman who was accused of misdemeanors than to a white man who had committed armed robberies [5].

Another case of unfairness has been identified for an algorithm used to assign grades to students. The decision was based on students' information such as previous grades and teacher-estimated grades. However, the algorithm proved to be biased against high-achieving students attending schools in poorer neighborhoods [3]. For example, students from such neighborhoods who were Spanish native speakers were predicted to fail the Spanish exam.

Unfair bias has also been identified in an AI algorithm built to assess beauty in a beauty pageant. This algorithm analyzed over 6000 uploaded pictures of people, but out of 44 winners, only one person had dark skin. This result was later followed back to the input data that was used for training which did not include enough diversity [4].

Therefore, it is important to ensure that machine learning systems do not make decisions that show discriminatory behaviors toward different populations. Given the increased use of machine learning systems in sensitive domains, numerous analyses have been undertaken to address the issue of unfair bias and to point out the importance of ensuring fairer systems [47, 17].

2.3 Metrics

To quantify unfair bias, a set of fairness metrics has been defined in literature [73]. Many of these metrics are based on the confusion matrix. The confusion matrix is a tabular representation of a classifier's performance, summarizing the relationship between predicted and actual outcomes. Table 2.1 displays the confusion matrix for binary classification tasks. For a binary classification task, the actual and the predicted classes have two values: positive and negative.

	Actual positive	**Actual negative**
Predicted positive	True Positive (TP)	False Positive (FP)
Predicted negative	False Negative (FN)	True Negative (TN)

Table 2.1: Confusion Matrix

[3] https://www.nytimes.com/2020/09/08/opinion/international-baccalaureate-algorithm-grades.html

[4] https://www.theguardian.com/technology/2016/sep/08/artificial-intelligence-beauty-contest-doesnt-like-black-people

For a binary classification task, the confusion matrix contains four elements [73]:

- **True positive (TP)**: cases that belong to the positive class and are predicted positive.
- **True negative (TN)**: cases that belong to the negative class and are predicted negative.
- **False positive (FP)**: cases that belong to the negative class but are predicted positive.
- **False negative (FN)**: cases that belong to the positive class but are predicted negative.

The confusion matrix also offers insight into the overall model performance through measures such as accuracy. Accuracy [27] is defined as the proportion of correctly classified instances (both positive and negative) to the total number of instances, calculated as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

2.3.1 Statistical Metrics

Statistical metrics are derived from the confusion matrix and can be calculated individually for each attribute (for example, gender), respectively, for each group (for example, women or men) within an attribute.

For a better understanding of the definitions that will be presented in this section, the following notations are used:

- $y \in 0, 1$: The actual true label of a binary classification problem.
- $\hat{y} \in 0, 1$: The predicted label for a binary classification problem.
- A = $a_1, a_2, ..., a_n$, sensitive attribute with multiple groups. For example A = gender, while a_1 = men, a_2 = women.
- $P(Y = 1)$, $P(Y = 0)$: probability of Y to be 1, respectively 0.
- $P(\hat{Y} = 1)$, $P(\hat{Y} = 1)$: probability of $\hat{Y}$ to be 1, respectively 0.

In [75], the authors provide a list of most known statistical metrics:

1. **Precision (Predictive positive value)**: represents the probability of cases predicted positive to actually belong to the predicted positive class.

$$Precision(PPV) = \frac{TP}{TP + FP} = P(Y = 1|\hat{Y} = 1)$$

2. **Predictive negative value**: represents the probability of cases predicted negative to actually belong to the predicted negative class [7].

$$NPV = \frac{TN}{TN + FN} = P(Y = 0|\hat{Y} = 0)$$

3. **Predicted Positive Rate**: represents the fraction of a group that was predicted as positive.

$$PPR = P(A = a_i|\hat{Y} = 1)$$

4. **True Positive Rate**: represents the probability of cases predicted positive to actually belong to the actual positive class.

$$TNR = \frac{TN}{TN + FP} = P(Y = 0|\hat{Y} = 0)$$

5. **True Negative Rate**: represents the probability of cases predicted negative to actually belong to the actual negative class.

$$TNR = \frac{TN}{TN + FP} = P(Y = 0|\hat{Y} = 0)$$

6. **False Discovery Rate**: represents the fraction of negative cases that were predicted positive out of all positive predicted cases.

$$FDR = \frac{FP}{FP + TP} = P(Y = 0|\hat{Y} = 1)$$

7. **False Omission Rate**: represents the fraction of positive cases that were predicted negative out of all negative predicted cases.

$$FOR = \frac{FN}{FN + TN} = P(Y = 1|\hat{Y} = 0)$$

8. **False Positive Rate**: represents the fraction of cases predicted incorrectly positive out of all negative actual cases.

$$FPR = \frac{FP}{FP + TN} = P(Y = 0|\hat{Y} = 1)$$

9. **False Negative Rate**: represents the fraction of cases predicted incorrectly negative out of all positive actual cases.

$$FNR = \frac{FN}{FN + TP} = P(Y = 1|\hat{Y} = 0)$$

2.3.2 Definitions of Fairness

Defining fairness in machine learning systems is challenging because it is difficult to quantify fairness into a single definition [75]. Depending on the task, the perspective of fairness can be different. Fairness metrics are typically computed at the level of sensitive attributes, comparing the outcomes of specific metrics between privileged and unprivileged groups within those attributes. Common definitions of fairness include [75]:

1. **Statistical Parity**: all groups (both privileged and unprivileged) should have a similar probability of being assigned to the positive class [23].

2. **Predictive parity/Precision**: both privileged and unprivileged groups should have an equal PPV, meaning that all groups should have the same probability of belonging to the positive class if they have a positive predicted value [19].

3. **Negative predictive value parity**: both privileged and unprivileged groups should have equal NPV, meaning that all groups should have a similar probability of belonging to the negative class if they have a negative predicted value [75].

4. **Equal opportunity**: the probability of correctly predicting the positive outcome should be the same across all groups [34].

5. **Equalized odds**: both privileged and unprivileged groups should have equal TPR and FPR, meaning that the probability of a subject from the positive class being assigned to a positive class and the probability of a subject from the negative class being assigned incorrectly to the positive class should be the same for both the privileged and unprivileged [34].

6. **False positive error rate balance**: both privileged and unprivileged groups should have equal FPR [19].

7. **False negative error rate balance**: both privileged and unprivileged groups should have eqaul FNR [19].

2.3.3 Bias Sources

In machine learning models, bias can occur in different phases of the AI life cycle such as the collection of training data or the model design. This will lead to obtaining unfair and biased results [61]. In [53], the authors provide an in-depth description of different types of bias that can influence machine learning models. They categorize bias definitions into three main categories: data to algorithm, algorithm to user, and user to data.

1. *Data to algorithm bias*: If an algorithm is trained on biased data, the patterns it learns will also reflect that bias. Consequently, when the algorithm is applied to new data, its predictions are likely to be biased as well. This type of bias originates from the data itself and propagates through the machine learning model. Examples of this type of bias can include:
   - Measurement bias: This happens during the collection of the features and labels that will later be used for the prediction task, especially when one category of the population is measured or observed more frequently than the others [71].
   - Omitted variable bias: Leaving out one or more important attributes can exclude significant information that the model would base its prediction on. Therefore the exclusion of these attributes can change the prediction results and can negatively affect the model's performance [20].
   - Representation bias: If the data collected does not contain enough diversity, the machine learning models cannot learn enough information about the populations that are less represented in the dataset [71].
   - Aggregation bias: Occurs when general assumptions are made for the whole dataset, failing to identify that some attributes might not have the same importance or significance for all the individuals in the dataset [71].

2. *Algorithm to user*: Bias in machine learning algorithms will affect their output which will also lead to a biased user experience and it is not necessarily connected to the input data:

  - Popularity bias: This occurs in recommendation systems and search engines when popular items are shown to users more frequently than less popular ones. However, popularity metrics can be manipulated through means such as fake reviews, meaning that the increased popularity and visibility of some items may not reflect their quality, but rather other biased factors [57].
  - Algorithmic bias: This can be caused during the algorithm design, optimization, regularization and whenever the developers evaluate the model's performance [8].
  - User interaction bias: The way the user interacts with various interfaces can be biased and it can be influenced by how the information is presented (presentation bias) or by the fact that most ranked items or results are the most attractive and the most popular (ranking bias) [8].
  - Evaluation bias: This arises during the evaluation phase of a model when benchmark datasets used to measure a model's performance do not represent real-world cases or the broader population. If these benchmarks are not representative (for example, they lack diverse data), models may perform well only on the benchmark but fail to generalize to diverse populations [71].

3. *User to data*: In many cases, the datasets on which the machine learning models are trained, are based on data generated by users (polls, web searches, reviews, etc). The bias in the user will reflect in the choices they make and therefore the data they create.
   - Historical bias: Existing societal bias in the real world can be easily reflected in data generation [71].
   - Population bias: Occurs when the attributes of a platform's user base, like demographics or user characteristics, differ from those of the intended broader population. This mismatch leads to data that does not accurately represent the original target group, which can skew findings or insights based on the platform's data [59].
   - Social bias: Being influenced by other's opinions, reviews, and behaviors can lead to changes in one's decisions or judgments [8].

2.3.4 Bias Mitigation Strategies

There are three known types of bias mitigation or reduction mechanisms, depending on which phase of the AI cycle are applied: pre-processing, in-processing, and post-processing methods [15].

- *Pre-processing methods*: Such methods will attempt to reduce the existing bias in the data before training a machine learning model on it.
  - Sampling methods: Aim to either over-sample or under-sample the training data in order to change the distribution of samples [35, 16].
  - Relabelling or massaging methods: Involve flipping or changing labels in the training data to balance the positive outcomes across the defined sensitive attributes, but without changing the class distribution [39].
  - Perturbation methods: Change the distribution of specific attributes in the training data to reduce the bias [78].
  - Reweighing methods: Assign weights to the training instances, creating a balance between the sensitive samples without changing the data directly [39].
- *In-processing methods*: Modifies the existing machine learning algorithms in such a way that the bias is minimized.
  - Constraint optimization methods: Integrate fairness constraints into the classifier's loss function, aiming to improve the fairness results. Some methods will attempt to find a balance between the accuracy and the fairness objective [6].
  - Regularization methods: Similar to the constraint optimization, such methods seek to penalize the classifier for unfair decisions, by adding penalty terms [40].
- *Post-processing methods*: This approach usually tries to modify the predicted labels to achieve certain fairness goals:
  - Thresholding methods: Adjust the decision boundaries to reduce the bias after a model was trained. To obtain the desired fairness results, a threshold value is set for the sensitive attributes based on various fairness metrics [34].

Limitations

Although bias mitigation approaches are crucial for improving fairness, they face several limitations. Modifying the data may compromise it and may create data that does not reflect real-world characteristics. For the mitigation to be effective, the modified data should be close to the original. Additionally, mitigation methods may introduce a trade-off between the accuracy of the models and balancing these two can be challenging [15].

2.3.5 Fairness Assessment Tools

Fairness assessment tools are important in analyzing and understanding detected unfair bias. They help identify unfair patterns, calculate fairness metrics and assist in mitigating the bias. Some of the most known tools are:

- IBM AI Fairness 360 (AIF360) [5]: provides calculation of various fairness metrics and mitigation methods for binary classification tasks, as well as visualization options for better understanding of the outcomes.
- Aequitas [6]: provides fairness reports evaluated for the subgroups of the protected attributes, as well as bias mitigation methods, but only for binary classification tasks.
- Microsoft Fairlearn [7]: assesses and mitigates bias for both binary classification and regression tasks.
- Responsibly: similar functionalities, but supports natural language processing (NLP) tasks [8].
- Fairlens [9]: offers bias and fairness measurements and supports sensitive attributes detection.

2.4 Intersectional Fairness

As noted in the section above, numerous works have investigated unfair bias in machine learning systems. However, a significant number of these works focus mainly on singular dimensions such as gender or race. This approach fails to address the intersectionality of these categories and therefore unfairness is not checked for individuals that might belong to two categories (black women, Asian men, etc).

The term 'intersectionality' was first introduced by [21] in an essay that explained how multiple dimensions of identity can shape an individual's experiences of racism, disadvantages, or privileges. It also highlights how focusing only on one dimension of identity can overlook intersecting ways of discrimination.

[5] https://aif360.readthedocs.io/en/stable/
[6] http://aequitas.dssg.io/
[7] https://fairlearn.org/
[8] https://github.com/ResponsiblyAI/responsibly
[9] https://github.com/synthesized-io/fairlens

One important work on unfair bias [12] analyzed three facial recognition algorithms from Microsoft, IBM, and Face++ and noted that all three algorithms perform worse on women images than men as well as on dark skinned people than lighter skinned ones. Moreover, they investigated how the algorithms performed for the intersectional groups (darker females, darker males, lighter females, and lighter males) and found that again all three algorithms gave poor performances for darker females, where the error rates were significantly higher (23%-36%) than the rates for lighter females (0% - 7%).

Such findings drew the attention that intersectionality is a topic that must be addressed in machine learning systems. There have been recent works that aimed to define notions of intersectionality and ways to identify it. In [41] the phenomenon of *Fairness Gerrymandering* is defined, which occurs when a machine learning model performs well for individual groups but fails to perform the same way for intersectional subgroups. In their paper, the authors aimed at satisfying fairness constraints for a "combinatorially large or even infinite collection of structured subgroups definable over protected attributes". The subgroups are based on the number of protected groups: if there are n number of protected groups, then the number of subgroups should be 2^n. The paper introduces a new concept called 'subgroup fairness' that considers notions of fairness such as statistical parity across many structured groups. To decide if an outcome for a subgroup is fair, the study calculates the difference between the probabilities of positive outcomes between a subgroup and the entire population. The difference is re-weighted depending on the size of the subgroup with respect to the population. The smaller the difference then the more fair the result is.

Challenges

One of the biggest challenges when considering intersectional groups is the number of new attributes obtained. If the number of original protected attributes is large, considering the approach of 2^n would be computationally challenging and impractical as it could result in many subgroups. This would lead to difficulties in evaluating fairness across every possible intersectional group and a large number of intersectional attributes can result in small subgroups with low representation in the dataset [44]. Therefore, the question of which attributes should be included in the intersectional case arises.

To address this challenge, several methods have been proposed to automatically detect intersectional subgroups at the highest risk of discrimination. For instance, [63] identifies disadvantaged subgroups focusing on groups formed by intersecting pre-defined protected attributes. In [62], the authors propose a method of detecting subgroups of a dataset that "differ from the whole dataset" with respect to defined metrics of interest such as false positive and negative rates. The visualization tool presented in [13] enables users to create subgroups by combining existing dataset attributes but also suggests the users whose subgroups might be disadvantaged. In this approach, the suggested subgroups are identified by grouping similar data points based on their features and are described by highlighting dominant features (those with values that are most common within each group).

2.5 Summary

This chapter provided an extensive overview of the current approaches to fairness in AI, providing insight into sensitive attributes, unfair bias, and the concept of intersectionality. While previous works have laid a strong foundation, several challenges remain, particularly in effectively identifying attributes at risk of discrimination and managing the complexities of intersectional analysis.

This thesis builds upon these efforts by utilizing a semi-automated algorithm to detect possible sensitive attributes in a dataset. In comparison to the methods discussed for identifying intersectional subgroups, this approach focuses on detecting the attributes at risk of discrimination, rather than the subgroups. This is done to allow the evaluation of bias both in the non-intersectional and intersectional cases and in the comparison of results. To address the computational challenges associated with intersectionality, this work constructs intersectional attributes from the highest-ranked sensitive attributes, limiting the number of groups while maintaining a focus on those most susceptible to bias. The created subgroups, in this case, are limited to the intersection of only two detected sensitive attributes at once.

The bias analysis is done with existing public fairness tools. Finally, bias mitigation techniques are applied to these intersectional cases, showcasing their effectiveness and implications in reducing unfairness.

It is important to note that while there is ongoing research on defining intersectional fairness and developing methods to mitigate bias in the intersectional case [33], many of these methods are not yet incorporated into ready-to-use toolkits. This thesis does not focus on implementing such methods but rather explores whether existing fairness toolkits can assist in identifying such bias, and more importantly, mitigating it. The findings aim to evaluate the utility and limitations of these tools in addressing the challenges of intersectional fairness.

3 Approach

This chapter outlines the overall strategy of this Master's thesis. It begins by introducing the datasets selected for analysis, providing a detailed explanation of the criteria and reasoning behind their selection. Following this, this chapter describes the approach used to address the research questions, offering insight into the techniques applied throughout this thesis.

To address RQ1 (*To what extent can a semi-automated approach assist in identifying sensitive attributes and forming relevant intersectional groups within public datasets?*) an algorithm is used to identify which attributes from a dataset are considered to be the most sensitive. Based on the intersection of "traditional" sensitive attributes, such as race or gender, and these automatically detected sensitive attributes, new intersectional groups are created.

In order to answer the next research question, RQ2 (*Does the consideration of intersectional groups reveal or amplify hidden biases that are not evident in non-intersectional groups?*), a set of statistical metrics are calculated for the defined sensitive attributes in the intersectional case to quantify and investigate unfair bias.

Finally, to answer RQ3 (*To what extent can existing bias mitigation methods reduce unfair bias in the case of intersectional groups, and what are the implications on model performance?*), existing bias mitigation methods are applied with respect to the defined intersectional groups that show unfair bias and the results are examined.

3.1 Data Selection

All the datasets analyzed for this master's thesis were downloaded from the OpenML website [74], a site that provides thousands of free machine learning datasets. Usually, these datasets come with a short description of their content, the source of the dataset, the type of data (numerical or categorical) and a target attribute. All datasets on this website are assigned a unique ID.

Categorization of Data

In order to filter through all the datasets existing on the website, only those datasets that contained a set of sensitive attributes were chosen. The sensitive attributes were defined in accordance with the Equality Act from 2010 [1] which protects people from being discriminated against the following characteristics: age, disability, gender (re)assignment, marriage and civil partnership, pregnancy and maternity, race, religion or belief, sex, and sexual orientation. Not all of these characteristics were found in the public datasets.

The selection process applied several restrictions: First, datasets were required to include gender or sex as an attribute and at least one additional characteristic from the Equality Act list. Second, datasets were required to contain a clear description of their contents and prediction tasks on the OpenML website. Finally, datasets with excessive missing values, particularly in the sensitive attributes, were excluded. After applying these criteria and removing datasets duplicates, a total of over 40 datasets remained.

The datasets were then categorized into four main topics depending on their classification task: education, job or income, health and banking or finance. From this final list, four datasets were manually selected to demonstrate the results of this thesis, one from each of the four topics. These datasets are summarized in Table 3.1.

Dataset ID	**Dataset**	**Prediction (yes or no)**	**Topic**
43141	ACS Income	Earn $\geq 39,000\$$ per year	job/income
46356	German Credit Risk	”Good” credit risks	banking/finance
45069	Diabetes180US	Readmitted to the hospital	health
43904	Law Bar School Exam	Pass bar exam	education

Table 3.1: Datasets overview with their prediction tasks and the assigned topic.

[1] https://archive.equalityhumanrights.com/en/equality-act-2010/what-equality-act

3.2 Sensitive Attributes Detection

Although the selected datasets were filtered to contain at least two sensitive attributes, other attributes that are typically not considered sensitive may still be at risk of discrimination. Attributes such as education, income level, or place of residence might not be considered sensitive attributes per se but could be linked to socioeconomic status in certain areas, thus making them indirectly sensitive. To identify such attributes, a semi-automated sensitive attribute detection algorithm was used to rank the attributes of a dataset.

The algorithm was executed on each dataset, evaluating both the pre-defined protected attributes, as well as the attributes highly correlated to them. A range of metrics was computed for these attributes, and a ranking was established by aggregating the metric-specific rankings into a single sensitivity score. Based on this aggregated ranking, new intersectional groups were created.

However, selecting an excessive number of attributes to create intersectional groups can be challenging since it may result in a large number of subgroups that may be poorly represented in the dataset. This can make it difficult, or even impossible, to calculate statistical metrics for these subgroups. For each dataset, a minimum of three sensitive attributes were used for the intersection. Where dataset size permitted, additional attributes were included, with a maximum of five attributes considered.

3.3 Intersectional Attributes and Fairness Analysis

The purpose of having such new attributes is to investigate how present bias is transferred to the intersectional level, and whether undiscovered bias can be identified. A single sensitive attribute may not show any bias on its own, but when combined with another, unfair bias can emerge (for example, when combining income level and gender).

The investigation of bias in these cases began by training different machine learning models on the datasets. Based on the predictions of these models a set of statistical metrics and fairness metrics were calculated for each defined sensitive attribute. Statistical metrics were computed individually for each subgroup within a sensitive attribute, whereas fairness metrics assessed disparities in the statistical metrics between privileged and unprivileged groups. The privileged subgroups were always pre-defined. To calculate these metrics as well as to compare and analyze results, existing fairness frameworks were used.

3.4 Bias Mitigation Methods

To mitigate the detected unfair bias, a set of existing pre-processing, in-processing, and post-processing bias mitigation methods were applied to the datasets. The selected bias mitigation methods were available through public toolkits. Pre-processing methods modify the original dataset before training, in order to obtain better fairness results. As discussed in Chapter 2, the way the data was collected can be a major influencing factor for unfair bias. Therefore, if the sensitive attributes have an imbalanced distribution within their subgroups, pre-processing methods can help counteract this. In-processing methods will attempt to control the bias while the model is training, usually by imposing certain constraints. Finally, post-processing methods will change the predicted labels after training in such a way that the unfair bias is reduced for the selected sensitive attributes.

These bias mitigation methods were applied to all selected datasets, however, their effectiveness might differ from one dataset to another. After applying each method, fairness metrics were re-evaluated and compared to the original fairness metrics before the mitigation. Additionally, trade-offs between accuracy and fairness results were taken into consideration, because although bias mitigation methods can reduce bias, this can come at the cost of affecting the model's performance.

3.5 Summary

This chapter has presented the foundational approach to addressing the research questions of this thesis. It began with the careful selection and categorization of datasets, ensuring they contained meaningful sensitive attributes for fairness analysis. Using a semi-automated detection algorithm, both predefined and indirectly sensitive attributes were identified and ranked, facilitating the creation of intersectional groups to uncover hidden biases.

By evaluating statistical and fairness metrics, the thesis aims to provide a nuanced understanding of how biases manifest at both non-intersectional and intersectional levels. The subsequent application of bias mitigation techniques offers a practical pathway for reducing these biases, while acknowledging the trade-offs between fairness and accuracy.

This approach establishes the groundwork for a comprehensive analysis of fairness across datasets and serves as the basis for answering the core research questions addressed in the following chapters.

4 Experimental Setup

This chapter outlines the experimental setup used to address the research questions. It begins by introducing the datasets selected for analysis, highlighting their key characteristics. Then, it presents the algorithm employed for detecting sensitive attributes. Finally, the tools and algorithms used for bias analysis and mitigation are presented, along with a justification for their application.

4.1 Datasets

This section will describe in detail the datasets that were analyzed in this thesis, as well as data pre-processing techniques and models used for data training.

ACS Income

This dataset (OpenML id 43141) contains information about $1,664,500$ individuals' annual income. On OpenML, the dataset was already converted to numerical data with no detailed description for the attributes, but information about each attribute was provided in a link available in the dataset description [22]. Although the target attribute (yearly income) was initially continuous, it has been converted into a binary attribute using the median value of all incomes as a threshold ($\approx 39,000$ dollars/year) to convert it from a regression task to a binary classification task while achieving a balanced distribution of the target. The dataset contains 12 different attributes, including attributes that can be considered sensitive: age, race, marital status, and sex. The original race attribute contained nine different categories, however, it was extremely imbalanced: one of the categories alone represented 78% of the dataset. Therefore, the race attribute was converted into a binary column. Table 4.1 illustrates the unbalance between the races.

Race	Percentage
1 (White)	77.99%
2 (Other)	22.01%

Table 4.1: Race categories distribution, ACS Income dataset

Table 4.2 shows the distribution of the marital status attribute. It can be noted that over 50% of the whole dataset belongs to group 1 (married).

Marital status	**Percentage**
1 (Married)	54.59%
2 (Separated)	14.52%
3 (Never married)	$30,87\%$

Table 4.2: Marital status categories distribution, ACS Income dataset

In Figure 4.1, which represents the age distribution, it can be noted that the majority of individuals are between 30 and 60 years old. To make it possible to analyze the results, the age attribute was transformed into a categorical one containing three categories: 'Less than 30 years', 'Between 30 and 50', and 'More than 50 years'. The distribution of the age categories can be seen in Table 4.3.

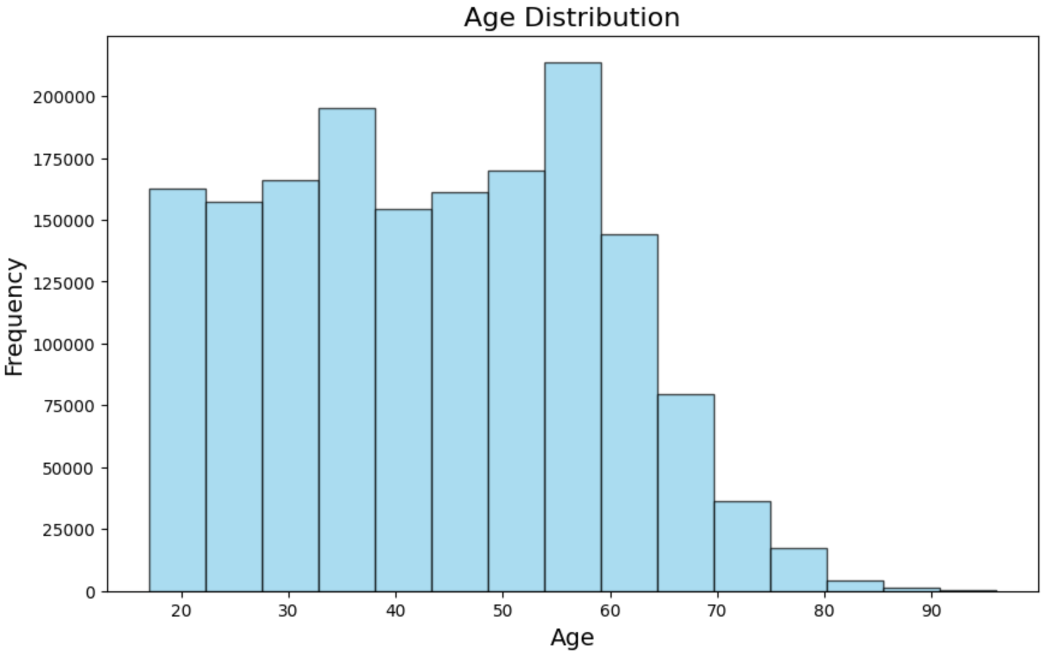


Figure 4.1: Distribution of age, ACS Income dataset

Age category	Percentage
Less than 30 years old	23.19%
Between 30 and 50 years old	38.74%
More than 50 years	38.06%

Table 4.3: Age categories, ACS Income

German Credit Risk

This dataset (OpenML id 46356) was designed to classify individuals as either ”good” or ”bad” credit risks based on a set of 21 attributes, including the target one. The distribution of the target attribute is: 70% of the individuals are considered at good risks, while 30% are not. It contains 1000 entries and includes financial information about individuals as well as personal information such as gender and age. Table 4.4 presents the gender distribution in the dataset, highlighting that the percentage of men is more than twice the number of women. The age attribute was converted from numerical data to two categories of age: 30 years old, less, and more than 30 years old. Figure 4.2 shows the initial distribution of age, while Table 4.5 shows the categories created.

Gender	Percentage
male	69%
female	31%

Table 4.4: Gender categories distribution, German credit risk dataset

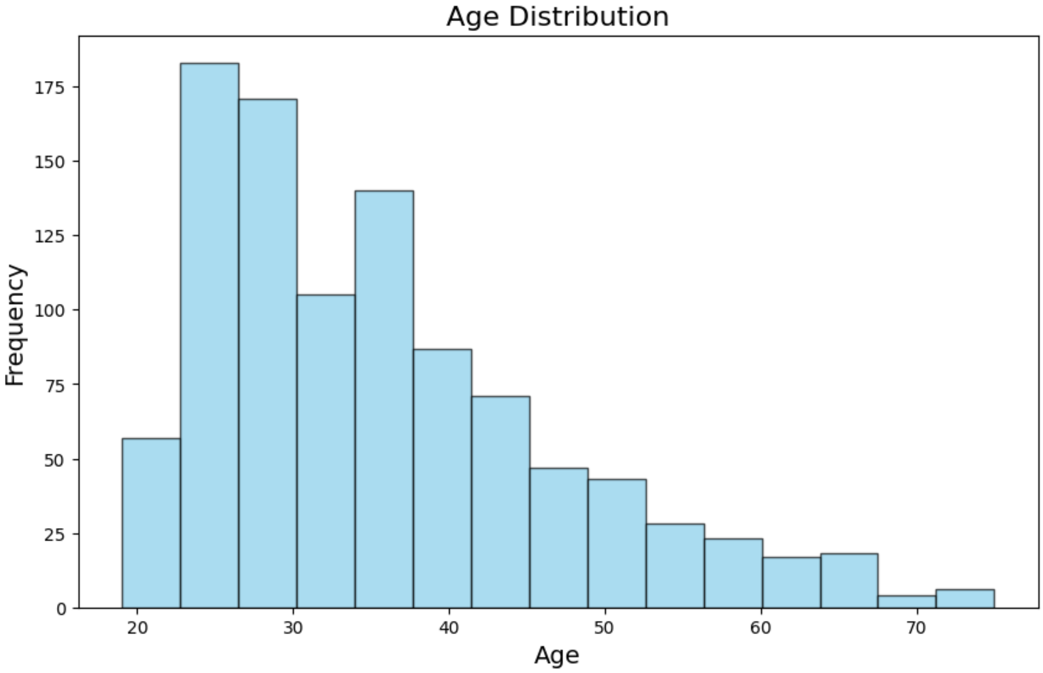


Figure 4.2: Distribution of age, German credit risk dataset

Age	Percentage
30 years old or less	62.9%
More than 30 years old	37.1%

Table 4.5: Age categories distribution, German credit risk dataset

Diabetes130US

This dataset (OpenML id 45069) has over 99,000 entries (after the cleanup of missing values in some attributes) and contains hospital records of patients with diabetes. The target attribute indicates whether a patient was readmitted to the hospital or not after hospitalization. Initially, the target attribute contained three categories: readmitted in less than 30 days, readmitted after 30 days, or not readmitted at all. Since Aequitas does not support multi-class problems, the target attribute was reorganized into two categories: readmitted to the hospital (46.6% of samples in the data set), or not (53.6% of samples). The dataset provides information about the patient's medical records, but it also contains personal information such as gender, age, and race.

The gender attribute contains balanced data. It can be seen in Table 4.6 below that the distribution of the two genders is almost equal.

Gender	Percentage
male	53.84%
female	46.15%

Table 4.6: Gender categories distribution, Diabetes130US dataset

The race attribute consisted initially of many categories (Caucasian, African-American, Hispanic, Asian), but since the Caucasian category represented over 76% as seen in Table 4.7, the other categories were combined into one ”Other” category to help with the interpretation of results.

Race	Percentage
Caucasian	76.48%
Other	23.51%

Table 4.7: Race distribution, Diabetes130US dataaset

Table 4.8 shows the distribution of the three age categories. Originally, age was categorized into nine categories ([0,10], (10,20], etc.), but since the majority of the patients were 50 years or older, these categories were consolidated into three broader groups.

Age	Percentage
Less than 50	15.69%
[50,70]	39.08%
More than 70	45.22%

Table 4.8: Age categories distribution, Diabetes130US dataset

Law School Bar Exam

The dataset (OpenML id 43904) consists of records for 20,000 law school students who attended school, with the goal of predicting whether or not they passed the bar exam. Alongside data on their grades, the dataset includes sensitive information such as race, gender, and age. The target variable is highly imbalanced: 89% of students passed the exam, while only 11% did not. The gender attribute is balanced, but the race attribute contains imbalanced data: 84% belonging to the Caucasian category. The rest of the categories were combined into one category. The age attribute was separated into two categories: less than 60 years old and 60 years old and more. The following tables present the category distribution of these three attributes.

Gender	**Percentage**
Male	56.12%
Female	43.87%

Table 4.9: Gender categories distribution, Law school admission dataset

Race	**Percentage**
White	84.1%
Other	15.9%

Table 4.10: Race categories distribution, Law school admission dataset

Age	**Percentage**
White	64.4%
Other	35.6%

Table 4.11: Age categories distribution, Law school admission dataset

Table 4.12 provides an overview of the selected datasets, including their size, prediction task, and key observations.

ID	Dataset	Size	Prediction task (Y/N)	Observations
43141	ACS Income	1,664,500	Earn $\geq 39,000\$$ per year	Balanced class and gender distribution; imbalanced race distribution.
46356	German Credit Risk	1000	”Good” credit risk	Small sized dataset, imbalanced class distribution.
45069	Diabetes180US	101,766	Readmitted to the hospital	Imbalanced race distribution.
43904	Law School Bar Exam	20,000	Pass bar exam	Imbalanced class distribution and race distribution

Table 4.12: Overview of Datasets and Observations

4.1.1 Data Cleaning

Before training machine learning models on the datasets above, each of them had to be pre-processed to ensure that they were suitable for machine learning tasks and fairness analysis. The data cleanup steps included:

- **Handling missing values**: In cases where the dataset was sufficiently large and included rows with NaNs or null values, particularly in sensitive attributes, the corresponding rows were removed. Columns with nonsensitive attributes having too many missing values (more than 50%) were either removed from the dataset or the missing values were imputed: for categorical data, the most common value of the column was used, and for numerical data, the average value of the column was used. Datasets with sensitive columns containing too many missing values were not considered for analysis.

- **Continuous target columns**: The target columns that were continuous (used for linear regression tasks) were converted to binary columns due to the limitations of the toolkits used. To convert the target column, the median value was used to divide the column into two parts of equal size. The datasets for which the target attribute could not be converted into a binary class prediction problem were also not considered, since Aequitas can only be used for binary classification problems.

- **Column cleaning**: Columns that were not important for the training, nor for fairness analysis (for example, names, IDs, etc) were removed from the dataset.

- **Data normalization**: Continuous data was normalized using the Standard Scaler[1], which transforms the data to have a mean of 0 and standard deviation of 1, ensuring that all continuous features are on a similar scale.

- **Sensitive attributes with many categories**: If a sensitive column had too many categories from which some were poorly represented, then these categories were combined into a single category. Although this step might remove information about some underprivileged groups, it was necessary in order to calculate fairness metrics. When a category is strongly underrepresented, the fairness metrics might be impossible to calculate (for example, due to divisions by 0 in statistical metrics).

- **Encoding to numerical**: for all the categorical columns, the data was encoded to integers. This was done to be able to train machine learning algorithms that do not accept categorical data.

4.1.2 Model Training

Each dataset was trained using several well-known machine learning algorithms. CatBoost [2] is a gradient boosting algorithm that automatically handles categorical features, making it efficient for large datasets and classification tasks. Random Forest [3] is an ensemble method that builds multiple decision trees and aggregates their predictions, providing robust performance and reducing overfitting. Decision

[1] https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.htmll

[2] https://catboost.ai/

[3] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

Tree [4] is a simple model that splits data based on feature values to create a tree structure, useful for both classification and regression. Logistic Regression [5] is a linear model for binary classification that estimates the probability of an outcome using a logistic function.

To optimize model performance, RandomizedSearchCV [6] was used to randomly sample hyperparameters from specified ranges. For training, the data was split into training data and test data (75% and 25%, respectively). The model's performance was evaluated using the test data. The results from all four models were compared for each dataset.

4.2 Sensitive Attributes Detection

The algorithm used to answer the first research question is semi-automated and is based on the approach in [38]. In the original paper, the authors outlined three key steps for detecting sensitive attributes in a dataset: first, computing a set of bias-indicating metrics for each attribute; second, ranking all the attributes for each metric; and finally, ranking the attributes by summing their individual metric rankings. This algorithm extends this approach by incorporating additional considerations, such as the correlations between attributes. This algorithm is applied to all four datasets presented in the previous section. The process contains the following steps:

1. **Detect protected attributes**: The first step after loading the dataset is to identify existing protected attributes in the dataset using the Fairlens toolkit [7]. This toolkit relies on a dictionary that maps potential protected attribute names to predefined categories: age, gender, ethnicity, religion, nationality, family status, disability, and sexual orientation. While this method is effective for identifying commonly recognized protected attributes, its scope is limited to the attributes explicitly listed in the dictionary for each category. The detection of protected attributes results in N_P many protected attributes.

2. **Calculate correlations between protected attributes and other attributes**: After identifying the protected attributes, the next step is to calculate the correlation between these attributes and other features in the dataset.

[4] https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html

[5] https://scikit-learn.org/stable/modules/generated/sklearn.linearmodel.LogisticRegression.html

[6] https://scikit-learn.org/1.5/modules/generated/sklearn.modelselection.RandomizedSearchCV.html

[7] https://github.com/synthesized-io/fairlens

This step helps identify attributes that may be correlated to the protected ones and could potentially carry unfair bias. Attributes with strong correlations to any protected attributes and a statistical significance level of p-value ≤ 0.05 are identified for further analysis and are referred to as *focus attributes*. This results in N_F many focus attributes.

Depending on the type of attributes being analyzed, the correlation coefficients were calculated using different correlation types:

- **Between two numerical attributes**: The Pearson correlation coefficient was used [29]. The correlation value ranges from -1 (perfect negative correlation) to 1 (perfect positive correlation), with 0 indicating no correlation:

  $$r = \frac{\sum_i^N (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_i^N (x_i - \overline{x})^2 \sum_i^N (y_i - \overline{y})^2}} \tag{4.1}$$

  Where:

  - x_i, y_i: Values of the two numerical attributes
  - $\overline{x}, \overline{y}$: Means of values x and y
  - N: Number of data points.

- **Between two categorical attributes**: The Chi-Square test with Cramér's V was applied to measure the association [42]. This is calculated based on the contingency table - a table that displays frequencies for combinations of two categorical variables- of two categorical attributes. The value ranges from 0 (no correlation) to 1 (perfect correlation).

  Chi-Square Formula:

  $$X^2 = \sum \frac{(O - E)^2}{E} \tag{4.2}$$

  Where:

  - O: Observed frequency in each category.
  - E: Expected frequency, calculated as $E = \frac{\text{row total x column total}}{\text{grand total}}$

  Cramer's V Formula:

  $$V = \sqrt{\frac{X^2}{n(k-1)}} \tag{4.3}$$

  Where:

  - n: Total number of observations.
  - k: The smaller of (number of rows - 1) or (number of columns - 1)

- **Between a numerical attribute and a binary categorical attribute**: The Point-Biserial Correlation Coefficient was computed [46]. The value ranges from 0 (no correlation) to 1 (perfect correlation).

$$r_{pb} = \frac{\overline{X_1} - \overline{X_0}}{s} \sqrt{\frac{n_1 n_0}{n^2}} \tag{4.4}$$

  Where:

  - $\overline{X_1}$: Mean of the numerical attribute for group 1 (category 1 of the binary attribute).
  - $\overline{X_0}$: Mean of the numerical attribute for group 0 (category 0 of the binary attribute).
  - s: Standard deviation of the numerical attribute.
  - n_1, n_0: Sizes of groups 1 and 0, respectively.
  - n: Total number of observations

- **Between a numerical attribute and a categorical attribute with multiple groups**: was performed to analyze the relationship between a numerical attribute and a categorical attribute with multiple groups. The Omega-Squared Effect Size is used to assess the strength of this relationship [58]. Omega-squared measures how much of the variance in the dependent variable (numerical attribute) is explained by the independent variable (categorical attribute), with values ranging from 0 (indicating no relationship) to 1 (indicating a perfect relationship). **Omega-Squared Formula**:

$$\omega = \frac{SS_{between} - df_{between} \times MSE}{SS_{total} + MSE} \tag{4.5}$$

  Where:

  - $SS_{between}$: The sum of squares between groups, which represents the variability in the data that is explained by the differences between the group means.
  - SS_{within}: The sum of squares within groups, representing the variability within each group.
  - $df_{between} = k - 1$: Degrees of freedom between groups, where k is the number of attributes analyzed.

- $df_{within} = N - k$: Degrees of freedom within groups, where N is the sum of all individual data points across all k number of groups.
- MSE: Mean Squared Error (within-group variance estimate), calculated as $\frac{SS_{within}}{df_{within}}$
- $SS_{total} = SS_{between} + SS_{within}$: Total sum of squares.

The first three correlation measures were computed using built-in methods from the SciPy library [76], while the One-Way ANOVA was implemented using the Statsmodels library [69].

3. **Metrics calculation**: A set of metrics is calculated for all selected protected and focus attributes. In total, $N_A = N_P + N_F$ many attributes are considered. The following metrics are computed:

   - **Entropy**: Measures the amount of randomness or uncertainty in a system [52]. If there is high uncertainty, the models might have difficulties in making accurate predictions.

     $$H(x) = -\sum p(x)\log(p(x)) \tag{4.6}$$

     Where:

     - $p(x)$: Probability of event x (likelihood of a specific class outcome).

   - **Imbalance Ratio (IR)**: Represents the ratio between the sample size of the most represented group in an attribute and the lowest represented group [66]. The larger the ratio, the more imbalance exists in an attribute.

     $$IR = \frac{N_{maj}}{N_{min}}, \tag{4.7}$$

     Where:

     - N_{max}: Represents the number of samples in the most represented group.
     - N_{min}: Represents the number of samples in the least represented group.

   - **Imbalance Degree (ID)**: Measures the extent of class imbalance in a dataset, its value reflecting how skewed the class distribution is [60]. In this case, to make this metric comparable across attributes with different numbers of classes, the relative imbalance degree is calculated by dividing the imbalance degree by the number of classes.

- **Statistical Parity Difference (SPD)**: Represents the difference between the probability of positive values in both the privileged and unprivileged groups [32]. A value smaller than 0 indicates better treatment for the privileged attribute.

$$SPD = P(Y = 1|G = unprivileged) - P(Y = 1|G = privileged) \tag{4.8}$$

  Where:

  - G: Unprivileged or privileged group.
  - $P(Y = 1|G = unprivileged)$: The probability that the outcome is positive, given that the individual belongs to the unprivileged group.
  - $P(Y = 1|G = privileged)$: The probability that the outcome is positive, given that the individual belongs to the privileged group.

- **Disparate Impact Ratio (DIR)**: Represents the ratio of positive values in the dataset between the unprivileged and privileged groups [26].

$$DIR = \frac{P(Y = 1|G = unprivileged)}{P(Y = 1|G = privileged)} \tag{4.9}$$

- **Smoothed Empirical Difference (SED)**: Measures fairness by comparing smoothed probabilities of favorable and unfavorable values across different intersecting groups in a dataset [28]. It evaluates the minimum ratio of these probabilities, with values between 0 and 1, where a higher value indicates more fairness between groups.

Except for the Entropy metric all of the aforementioned metrics require that privileged and unprivileged groups are defined. The privileged groups are considered to be the attributes with higher empirical distribution in comparison to the related attribute equiprobability, while the unprivileged groups are considered to be those with lower empirical distribution. The empirical distribution of a group is calculated as the ratio between the number of instances in a group over the total number of instances of an attribute. The equiprobability of an attribute is calculated as below:

$$eqp = \frac{1}{\text{number of subgroups}} \tag{4.10}$$

These metrics are calculated using the following existing libraries: Entropy is calculated using the SciPy library [76], Imbalance Degree using a public repository[8], while Statistical Parity Difference, Disparate Impact Ratio, and Smoothed Empirical Difference are calculated using available methods in AIF360 toolkit [9]. Imbalance Ratio is computed by obtaining the sample size of the groups.

4. **Rank attributes for each metric**: Once all metrics are computed, each attribute is ranked individually for each metric based on its value relative to the other attributes. The ranking is assigned as follows:
   a) For Entropy, ID, and IR higher values indicate higher sensitivity.
   b) For SPD, greater absolute differences indicate higher sensitivity, with a difference of 0 being the least sensitive.
   c) For DIR, the higher the term $1 - DIR$ is, the more sensitive the attribute is considered, with a value of 0 being the least sensitive.
   d) SED, higher values indicate less sensitivity, 1 being the least sensitive.

   For each metric, each attribute is assigned a value from 1 to N_A, where 1 means the attribute was the least sensitive with respect to the metric, while N_A means the attribute was the most sensitive with respect to the metric.

5. **Final ranking**: The individual rankings for all metrics are summed for each attribute. Attributes with higher total rankings are considered more sensitive, resulting in a final ranked list of attributes ordered by their overall sensitivity.

The following chart depicting this described method is shown in Figure 4.3

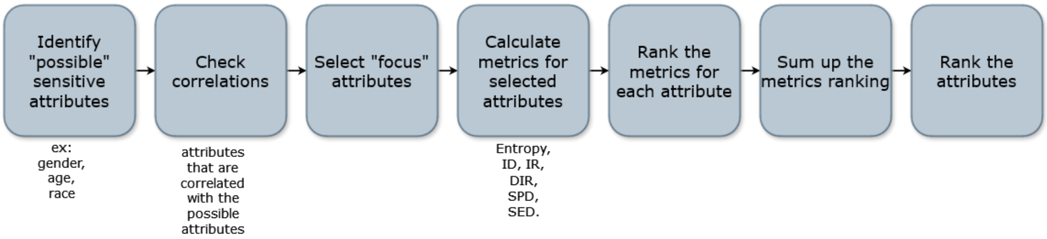

Identify "possible" sensitive attributes
ex: gender, age, race
Check correlations
attributes that are correlated with the possible attributes
Select "focus" attributes
Calculate metrics for selected attributes
Entropy, ID, IR, DIR, SPD, SED.
Rank the metrics for each attribute
Sum up the metrics ranking
Rank the attributes

Figure 4.3: Sensitive attributes detection algorithm

[8] https://github.com/mjuez/py-imbalance-degree

[9] https://aif360.readthedocs.io/en/stable/

4.3 Aequitas

This master's thesis made extensive use of the Aequitas [10] toolkit. Developed by the Center for Data Science and Public Policy at the University of Chicago, this toolkit is an open-source bias audit that can be used to analyze whether the predictions made by a machine learning model are biased or not. The tool can be accessed via the following interfaces:

- Python library
- Command Line Tool
- WebApp

The developers of Aequitas wanted to make this toolkit accessible not only for machine learning developers but also for analysts and policymakers. Therefore, the command line tool and the web audit tool do not require any programming skills. The only requirement to run these two is to have a dataset that already contains predictions made by a machine learning model [1]. For this project, only the Python library was used, as it is simple to integrate it with existing code.

4.3.1 Input Data

In order to obtain the bias report in any case, the input data must be standardized according to the requirements found on the Aequitas website [2]. The input data must contain two binary columns: a column called "*score*" for predictions and a column called "*label_value*" for the true label values. Since they are required to be binary, the Aequitas tool can only be used for binary class classifications or for linear regression problems that can be converted into a binary classification afterward using a threshold value. Besides these two columns, the sensitive columns must be added as well. The sensitive columns are those that the user wants to audit for bias, such as sex, age, race, income, education, etc. They can be either categorical or numerical, and the user is free to choose which attributes are considered sensitive. Other attributes that are not considered sensitive do not need to be added to the input data. For this thesis, the selected sensitive attributes were: 1) for the non-intersectional case, the sensitive attributes identified by the algorithm, and 2) for the intersectional case, the intersection of these sensitive attributes

[10] http://aequitas.dssg.io/

4.3.2 Output Data

After having the table containing the input data, the tool can be used to obtain the output reports with the fairness and bias metrics. The reports can include group metrics, bias reports, disparity metrics, or fairness. To generate the output, the toolkit expects a reference group for each attribute defined as sensitive. The reference group is defined by the user and can be either the group with the most representation in a category or it can be a group that is believed to be more preferred than the others. In this case, the privileged groups were defined as explained in Section 4.2.

4.3.3 Measuring Fairness with Aequitas

The tool calculates a set of statistical metrics (such as TPR, TNR, FPR, FNR, FOR, FDR, etc). To decide whether a fairness definition is met or not, the authors defined fairness criteria that decide if a group meets the parity. This fairness criteria is defined as follows:

$$(1-\tau) <= \frac{\text{statistical metric group}_{\text{i}}}{\text{statistical metric group}_{\text{privileged}}} <= \frac{1}{1-\tau} \tag{4.11}$$

This criteria is used for all statistical metrics (FPR, FNR, etc), and it calculates the parity for each group category. By default, $\tau = 20\%$, meaning that any parity that is between 0.8 and 1.25 is considered fair [67]. The parity for the privileged group will always be equal to 1.

4.3.4 Aequitas Flow

Besides the bias audit options, the latest release of the toolkit, Aequitas flow framework, contains integrated bias mitigation options, but also other functionalities such as model selection, hyperparameter optimization, and detailed plotting methods [36]. For this thesis, some of the bias mitigation methods available in this package were used, as well as some of the plotting methods that assist in understanding the bias reports.

4.3.5 Limitations

Although the toolkit provides an in-depth fairness report, there are some limitations to this toolkit that were noted during the work for this thesis. Firstly, this toolkit cannot be used for multi-class classification tasks, nor for regression tasks, which leads to either excluding tasks that may have discriminatory behaviors or transforming such tasks into binary ones. Another disadvantage is the expected format for the datasets (always renaming the target column and the predicted labels), which makes it difficult to use multiple fairness toolkits at the same time.

4.3.6 Application

This toolkit was selected for this thesis because of its capability to compute a wide range of fairness and disparity metrics for sensitive attributes. The bias analysis using this toolkit was performed after training the datasets with the machine learning models mentioned in section 4.1.2, as it requires the predicted outcomes for each task. This toolkit generates interpretable bias reports, enabling comparisons both across different sensitive attributes and between non-intersectional and intersectional cases. Its features assisted in addressing RQ2 and RQ3. In this thesis, the toolkit was used to perform the following functionalities, as found on the Aequtias website [11]:

- **Statistical metrics calculation**: Compute group-wise statistical metrics for the defined sensitive attributes. These metrics include TPR, TNR, FPR, FNR, FOR, FDR, NVP, PPR, and PPV. This is done using the *Group()* and its method *get_crosstabs()*, which calculates these metrics for each defined group.

- **Disparity metrics calculation**: Calculate for each sensitive attribute the disparity metrics between unprivileged and privileged groups for all statistical metrics. This is done using the method *get_disparity_groups()* in the *Bias* class. The method requires pre-defining the privileged group for each sensitive attribute and computes the disparity as the ratio between the metric value for the unprivileged and privileged groups. The result is a table containing the disparity values for each group.

- **Fairness criteria evaluation**: Determine whether each calculated disparity metric satisfies the defined parity criteria using Equation 4.11. This is done using the *Fairness* class and its method *get_group_attribute_fairness()* which returns a True or False value depending on whether the parity criteria was met.

- **Bias mitigation**: Apply available bias mitigation methods, selecting those suitable for scenarios with multiple sensitive attributes. The used methods will be covered in more depth in the following section.

[11] https://dssg.github.io/aequitas/usingpython.html

- **Visualization tools**: Use the built-in visualization tools to better understand and analyze the results. These methods include visual representations of the calculated metrics and the disparity values obtained during the analysis.

This workflow was implement following the guidelines provided in the Aequitas documentation [12].

4.4 Bias Mitigation Methods

For this Master's thesis, the majority of bias mitigation methods, especially the pre-processing methods, were developed within the Aequitas Flow toolkit. A short description of these methods can be found on the repository of the toolkit [13]. Aequitas Flow was selected because it is integrated within the Aequitas toolkit, ensuring consistency with the same preliminaries used throughout the thesis. Additionally, its functionality addresses RQ3, which focuses on the use of existing bias mitigation methods.

However, not all methods within Aequitas Flow could be applied, as two of its methods are restricted to binary sensitive attributes which are not applicable in this case. Furthermore, the toolkit offers a limited selection of in-processing and post-processing methods, necessitating the use of additional methods from other known toolkits. The following methods were used:

Pre-processing Methods

1. **Data repairer**: Aims to modify the data distribution to ensure that a given feature is independent of the sensitive attribute, *s*. This is accomplished by aligning the conditional distribution $P(X|s)$ with the global distribution $P(X)$ [25].

2. **Massaging**: Flips a fixed number of labels in order to reduce the prevalence disparity between the subgroups of a sensitive attribute without changing the overall class distribution [39]: This method fits a Naive Bayes Classifier to the data and then sorts the entries by the predictions. The instances with positive predicted value belonging to a privileged group are marked as ”Demotion candidates”, while the ones belonging to the unprivileged class and with negative predicted value are marked as ”Promotion candidates”. The

[12] https://dssg.github.io/aequitas/examples/compasdemo.html
[13] https://github.com/dssg/aequitas?tab=readme-ov-file

top ranked demotion candidates and the lowest ranked promotion candidates will be relabeled. The number of labels to be flipped is calculated with the following formula:

$$number\_to\_flip = \frac{((\overline{Prev}_{promotion} - \overline{Prev}_{demotion}) \times y_{postive} \times y_{negative})}{N}, \tag{4.12}$$

Where:

- N: number of instances.
- $y_{negative}$: Number of instances predicted as positive belonging to the privileged group.
- $y_{positive}$ Number of instances predicted as negative belonging to the unprivileged group.
- $\overline{Prev}_{promotion}$: Mean of all promotion candidates.
- $\overline{Prev}_{demotion}$: Mean of all demotion candidates.

3. **Prevalence sampling**: Under-samples or over-samples the original dataset in order to balance the class prevalence by changing the ratio of protected and unprotected subgroups within the sensitive attributes and the label distribution [45].

4. **Label flipping**: Flips a fixed number of labels based on the Fair Ordering-Based Noise Correction method [24], aiming to improve the demographic parity. It adjusts group label prevalence by flipping misclassified instances: negative to positive if the group's prevalence is too low and positive to negative if it is too high. Flipping stops once a target flip rate, margin threshold, or desired parity is achieved.

In-processing Methods:

1. **Fairlearn Classifier**: this method will try to train a set of different classifiers that can be chosen by the user while additionally satisfying some fairness definitions for a given sensitive attribute [6]. These definitions can be any that are available in the Reductions Package of the Fairlearn framework [14] such as Equalized Odds, TPR parity, or FPR Parity.

[14] https://fairlearn.org/v0.4.6/apireference/fairlearn.reductions.html

2. **GerryFair Classifier**: this method is developed based on the proposed idea by [41] and presented in 2, which aims to obtain subgroup fairness.

Post-processing Methods

1. **Group Threshold**: Adjusts the decision threshold of a model to achieve a specific fairness criterion independently for each group (for example, achieve FPR of 10%) [34]. It can only be applied to one sensitive attribute at a time.

2. **Threshold Optimizer**: This method attempts to achieve the same results as the above mentioned method, however, it can be applied to multiple sensitive attributes. Moreover, it can try to balance the accuracy score while also achieving the group constraints.[15]

3. **Equalized Odds Post-Processing**: Modifies the predicted labels so that the equalized odds within a group are optimized, ensuring the TPR and FPR are equalized [64] [16].

4.5 Summary

This section outlined the experimental setup and methodologies used to address the research questions. It detailed the datasets chosen for analysis, highlighting key features such as the type of prediction task, the number of entries, and any protected attributes present. The section also provided a thorough explanation of the sensitive attribute detection algorithm, outlining the specific steps taken to identify potentially sensitive attributes in the dataset. Additionally, it introduced the Aequitas fairness toolkit, explaining its functionalities and how it was used to detect unfair bias within the datasets. Finally, the section listed the bias mitigation methods employed, describing their goals in reducing bias.

[15] https://fairlearn.org/v0.8/apireference/fairlearn.postprocessing.html

[16] https://aif360.readthedocs.io/en/latest/modules/generated/aif360.algorithms.postprocessing.EqOddsPostprocessing.html

5 Results and Discussion

This chapter provides an overview of the findings obtained after undergoing all the steps mentioned in chapter 4. The results are organized as follows: for each dataset, the sensitive attributes were identified, including attributes that normally would not be considered sensitive. Based on this detection, intersectional attributes were created (RQ1). Next, biases are assessed for both non-intersectional and intersectional attributes and then compared (RQ2), followed by an evaluation of the effectiveness of the applied bias mitigation methods (RQ3). Depending on the prediction task of the dataset, the focus on the fairness metrics that need improvement changes. Finally, the last subsection of this chapter contains a discussion of the findings, which aim to answer the research questions.

5.1 RQ1: To what extent can a semi-automated approach assist in identifying sensitive attributes and forming relevant intersectional groups within public datasets?

This section addresses the first research question. The detection algorithm was applied to all four datasets to evaluate its effectiveness in identifying sensitive attributes and forming intersectional groups for bias analysis. The results include the detected sensitive attributes, their correlations, and the intersectional groups derived from the top-ranked attributes.

Overview of the detected sensitive attributes

Table 5.1 provides an overview of the findings after using the detection algorithm. The table lists the detected protected attributes using the Fairlens tool, followed by the focus attributes for each dataset (those strongly correlated to the protected attributes). Finally, the last column contains the top-ranked sensitive attributes selected for further analysis based on their sensitivity scores. This thesis focuses only on the top five most sensitive attributes, to avoid generating a large number of subgroups obtained by the intersections of the detected sensitive attributes.

For clarity, the following attributes and abbreviations are explained below:
ACS Income:

- POBC: Continuous attribute representing postal codes of the individual's place of birth.
- OCCP: Continuous attribute where each value corresponds to a specific occupation.

German Credit Risk Dataset:

- empl. time: Employment time measured in years.
- no. people maintenance: Number of people maintenance (number of people a person is liable for), with a maximum value of 2.

Diabetes130US

- diag1, diag3: Continuous attributes that represent two different diagnostics.

Law Bar School Exam

- LSAT: Law School Admission Test scores, categorized into two categories based on the median value which was equal to 37.
- family income: Categorical attribute representing five income levels, with category 5 indicating the highest income.
- cluster: Law schools grouped into six clusters. [79].

ID	Dataset	Detected Protected Attributes	Detected Focus Attributes	Top Ranked Attributes
43141	ACS Income	age, sex, race, marital status	POBC, OCCP	race, POBC, age, marital status, sex
46356	German Credit	gender, race, age	empl. time, housing, no. people maintenance	housing, gender, no. people maintenance, empl. time, age
45069	Diabetes	age, gender, race	diag1, diag3	race, diag1, diag3, gender, age
43904	Law Bar Exam	gender, age, race	LSAT, full-time, cluster, family income	race, family income, LSAT, cluster, age

Table 5.1: Overview of the selected datasets with detected protected attributes, focus attributes, and top-ranked attributes. The attributes are ranked based on their sensitivity, the first one being considered as the most sensitive. The attributes' abbreviations are described above.

The analysis of the table reveals that in all datasets, attributes beyond the predefined protected one were identified as sensitive. This suggests that non-traditional sensitive attributes may also pose a greater risk of discrimination. For example, in the Diabetes and Law School Bar Exam datasets, gender was included as a protected attribute but was not ranked among the most sensitive. Similarly, in the German Credit Risk dataset, race was not prioritized as highly sensitive. These findings emphasize the importance of considering dataset-specific attributes in fairness evaluations, as domain-specific factors such as place of birth, diagnostic types, or LSAT scores may disproportionately impact certain groups.

Correlations Between Protected and Focus Attributes

The strongest correlations between the protected attributes and focus attributes are presented in Table 5.2. The table includes the correlation values (*Corr. value*) and their associated p-values (*p-value*). Only the correlations that had a significance level $p \leq 0.05$ were considered. These correlation values determined the focus attributes' inclusion for further bias analysis.

Dataset	Protected Attribute	Focus Attributes	Corr. value	p-value
ACS Income	race	POBC	0.29	$p < 0.001$
	sex	OCCP	0.16	$p < 0.001$
German Credit	age	empl. time	0.29	$p < 0.001$
	gender	housing	0.23	$p < 0.001$
	age	no. people maintenance	0.22	$p < 0.001$
Diabetes	age	diag1	0.59	$p < 0.001$
	age	diag2	1.87	$p < 0.001$
Law Bar Exam	race	LSAT	0.25	$p < 0.001$
	age	full-time	0.2	$p < 0.001$
	race	cluster	0.14	$p < 0.001$
	age	family income	0.1	$p < 0.001$

Table 5.2: Correlation values between protected and focus attributes.

In the ACS Income dataset, the race attribute shows a strong correlation with POBC (place of birth), while the sex attribute is moderately correlated with OCCP (occupation). The German Credit Risk dataset reveals correlations such as age with employment time, respectively with number of people maintenance. Additionally, gender is correlated with housing. In the Diabetes dataset, the strongest correlations across all four datasets are observed between age and diag1, respectively diag2. The Law Bar Exam dataset shows that race is correlated with LSAT scores and with cluster (cluster of law schools). Meanwhile, age is associated with full-time and family income level. Overall, these correlations indicate that focus attributes may carry indirect biases due to their associations with protected attributes and underscore the importance of identifying and analyzing these relationships to ensure comprehensive fairness evaluations across datasets.

Intersectional Groups

The highest-ranked sensitive attributes were analyzed based on their distribution and number of groups, and then it was decided whether to keep all of them for bias investigation. Some attributes had to be excluded from the list due to their continuous format and the lack of information that could help categorize them meaningfully in subgroups. Such attributes were: POBC (ACS income), diag1, and diag3 (Diabetes). This is unfortunate, as these attributes exhibited strong correlations with protected attributes, indicating that their removal may result in the omission of potentially significant biases. Another factor in reducing the number of attributes was the size of the dataset; for example, the size of the German Credit dataset could not allow a successful creation of all the intersectional attributes because it would lead to too many under-represented subgroups, which would make the fairness analysis unreliable. Similarly, in the Law school exam dataset, the extremely imbalanced class distribution, combined with the intersection of the detected attributes, led to difficulties in calculating the fairness metrics. Therefore, for these two datasets, only the first three attributes were considered.

After identifying the most sensitive attributes, the next step involved the intersection of these attributes, resulting in new subgroups. The subgroups were obtained by forming combinations of two attributes at a time. The following Table 5.3 summarizes the obtained intersectional groups, which are considered sensitive attributes for the intersectional fairness analysis.

Dataset	Sensitive Attribute	#Subgroups
ACS Income	race/marital status	6
	race/age	6
	race/sex	4
	marital status/age	9
	sex/age	4
	sex/marital status	6
German Credit	no. people maintenance/gender	4
	gender/housing	4
	no. people maintenance/housing	4
Diabetes	gender/race	4
	gender/age	6
	race/age	6
Law Bar Exam	race/family income	10
	race/lsat	4
	lsat/family income	10

Table 5.3: Intersectional groups obtained by the intersection of the highest ranked sensitive attributes. These groups are considered as sensitive for the fairness analysis. This table also summarizes the number of obtained subgroups for each intersectional group.

It can be noticed in the table that the intersection of sensitive attributes leads to a larger number of subgroups, the highest number of subgroups being in the Law Bar School Exam dataset, where the intersections of family income with race, respectively with LSAT resulted in ten subgroups each.

5.2 RQ2: Does the consideration of intersectional groups reveal or amplify hidden biases that are not evident in non-intersectional groups?

To address this research question, a comprehensive bias analysis was performed for both non-intersectional and intersectional scenarios using the Aequitas toolkit.

This section highlights the biases identified in both cases across all four datasets. For each dataset, the analysis identifies sensitive attributes affected by bias and examines how these biases arise in the intersectional scenario. To present the results effectively, the bias analysis was only performed on the predictions of a single model. The model with the best accuracy and balanced performance across the training and test data was selected separately for each dataset.

Figures are included to visualize group and subgroup disparities, with color-coded indicators to show whether the fairness criteria were met (green) or not (red). The fairness threshold was set to the default value of 20%. Specifically, any ratio between unprivileged and privileged groups outside the range $[0.8, 1.25]$ was considered unfair, indicating that parity was not achieved. The privileged and unprivileged groups and subgroups will be introduced for each dataset in particular.

ACS Income dataset

CatBoost emerged as the best-performing classifier for this dataset, achieving an accuracy of 79% on the test set in both intersectional and non-intersectional scenarios. The table below highlights the privileged groups identified in each case:

Attribute	Privileged group/subgroup
race	1 (white)
marital status	1 (married)
age	1 (age between 30 and 50 years old)
sex	1 (men)
race/marital status	0 (white, married)
race/age	2 (white, aged between 30 and 50 years old)
race/sex	0 (white, men)
marital status/age	2 (married, aged between 30 and 50 years old)
sex/age	1 (men, aged between 30 and 50 years old)
sex/marital status	0 (men, married)

Table 5.4: Privileged groups and subgroups, ACS Income dataset

Original dataset Disparities were found for FNR, FPR, TNR, and TPR, as illustrated in Figure 5.1. Disparities in TPR and TNR suggest that the model struggles to make correct predictions for certain sensitive groups. Notably, FNR and FPR are failing for all unprivileged groups of race, sex, and marital status.

The following attributes were affected:

- Category 3 of Marital Status (people who never married) and the youngest adults (age category 0) are the most affected groups, with all four parities failing for these groups.
- While TPR and TNR parities for race and sex were met, disparities were evident in FPR and FNR. The FNR of women (sex category 2) was nearly

double that of men, meaning that women earning more than $39,000\$$ per year were half as likely to be correctly identified compared to men. Similar disparities were observed for individuals in race category 2 (”Other”).

- TNR and TPR are unfair for the youngest adults and for the adults whose marital status belongs to category 3 (people who never married).

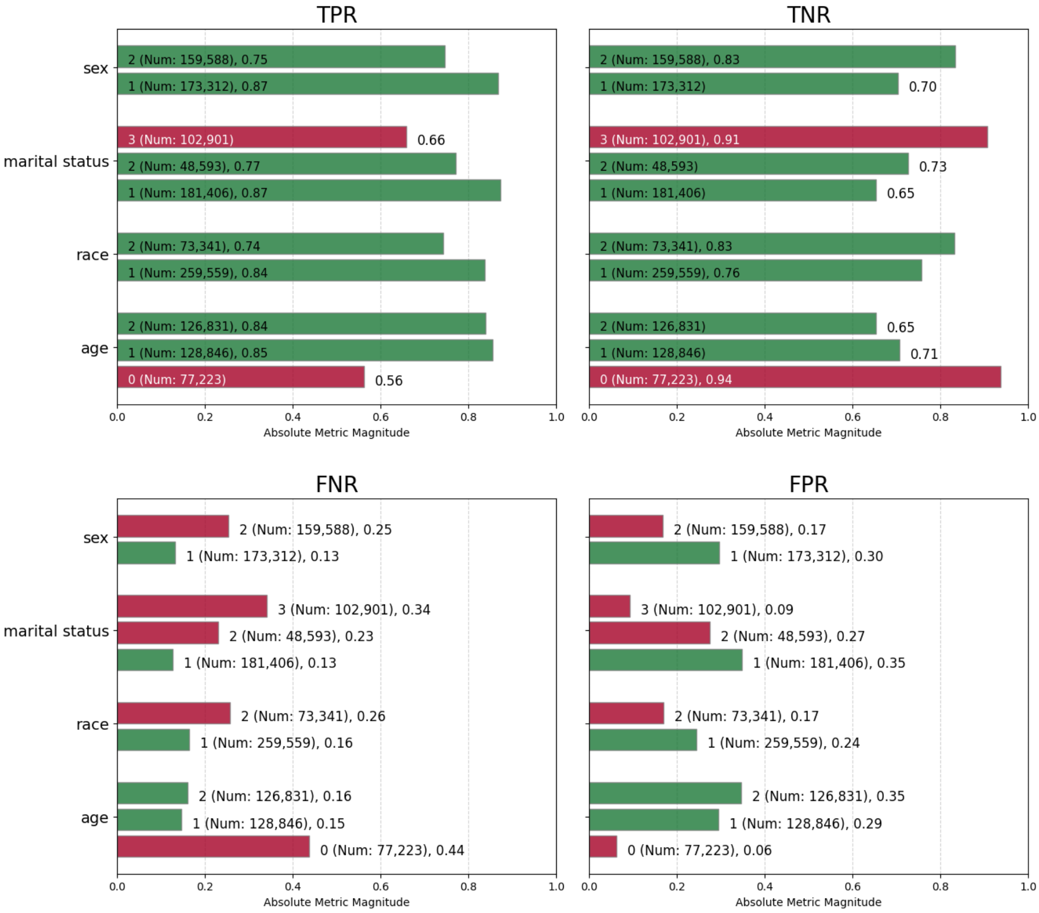


Figure 5.1: Detected disparities in the non-intersectional case. Notable are the FNR and FPR disparities that fail for almost all subgroups.

Intersectional dataset: Due to the unfairness detected in the original dataset, the disparities are intensified in the intersectional case, as seen in Figure 5.2.

- FNR is now failing for all subgroups, except for older adults from the privileged race (White) and privileged sex (Men). This indicates that the classifier struggles to predict the true label for all unprivileged subgroups in the intersectional case.

- For FPR, the highest values are for the privileged subgroups, indicating the classifier's tendency to favor these groups as positive when it is not the case. This means that individuals from the privileged group are more likely to be predicted to have a higher salary even when it is not the case.
- TPR disparities are the most pronounced for combinations involving the youngest age group (previously the most affected in the non-intersectional case) with race and sex attributes, as well as combinations involving marital status categories "divorced" or "never married" with sex and race.
- The combination of unprivileged sex (women) and race (category "Other") leads to disparities in TPR and TNR, even though these metrics passed for race and sex individually in the non-intersectional case.

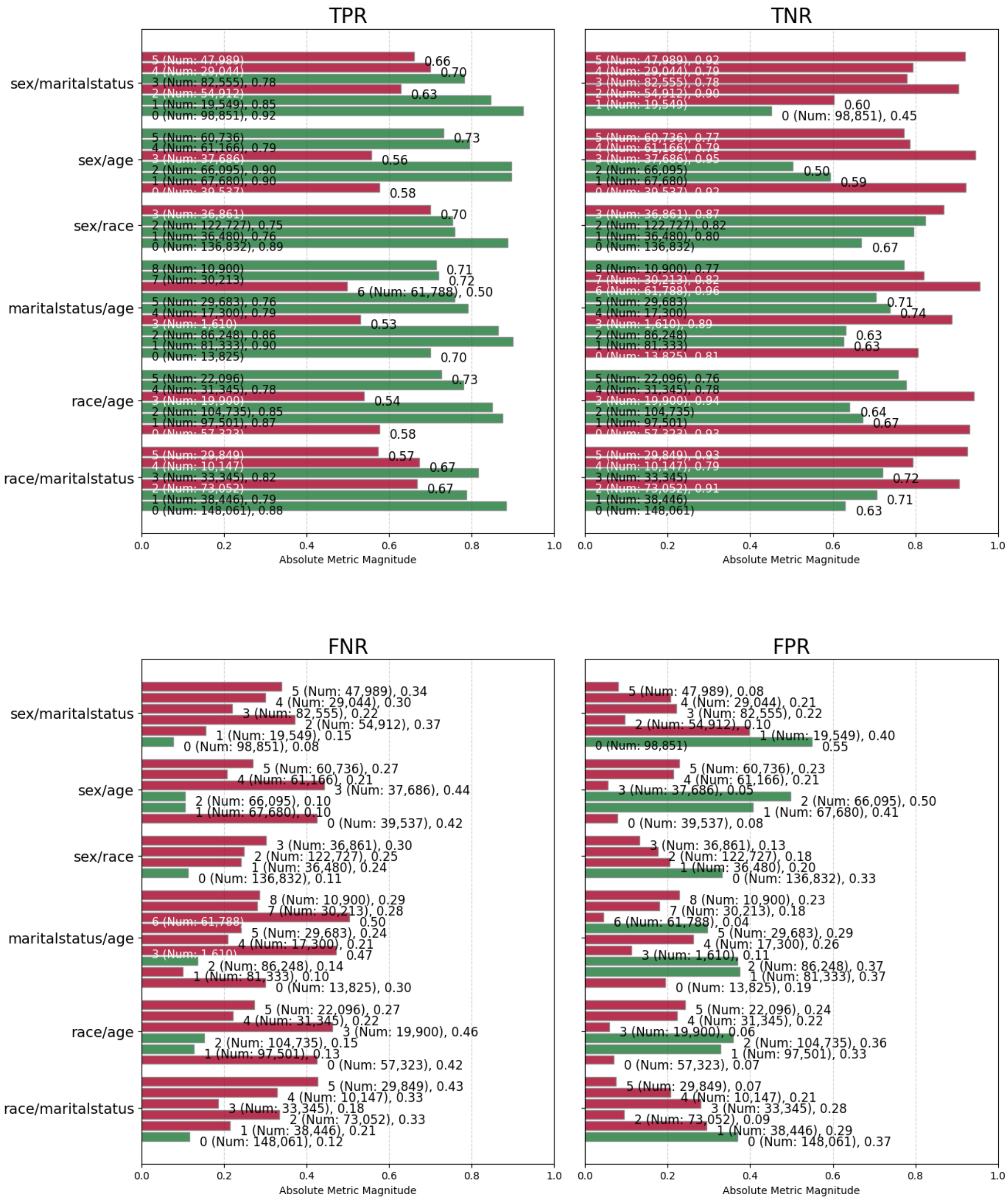


Figure 5.2: Detected disparities for the intersectional case, which highlight the amplification from the non-intersectional case.

German Credit Risk dataset

The best performing training model was Logistic Regression with an accuracy of 70% on the non-intersectional case and 71% on the intersectional case. The privileged and unprivileged groups for both cases are listed below:

Attribute	Privileged group/subgroup
housing	own
gender	male
no. people maintenance	1
gender/no. people maintenance	male, 1
gender/housing	male, own
no. people maintenance/housing	1, own

Table 5.5: Privileged groups and subgroups, German Credit Risk dataset

Original dataset: TPR and TNR parities were met for all groups. However, FNR disparities were observed for females and individuals without home ownership, as seen in 5.3. FOR is failing only for individuals without home ownership. This indicates that the model disproportionately misclassified these groups as ”bad” clients, even when they were not.

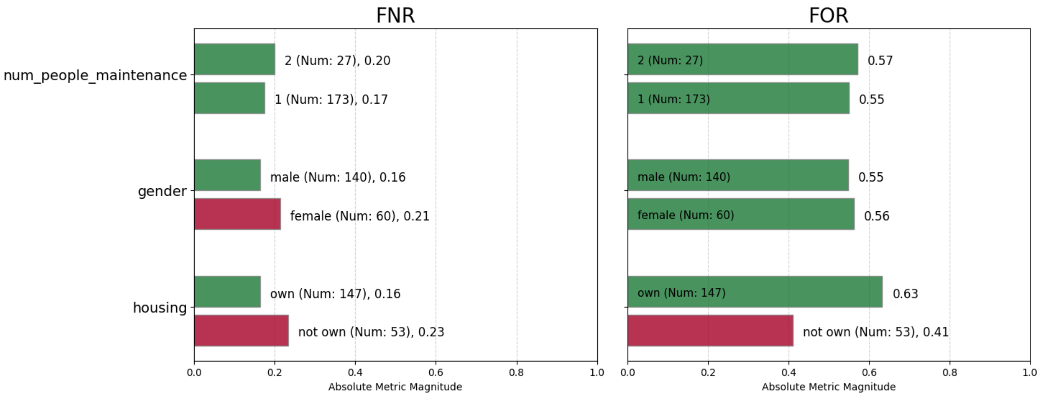


Figure 5.3: Detected FOR and FNR disparities in the non-intersectional case.

Intersectional dataset: Due to the disparities in the original dataset, in the intersectional case, the FNR disparities are amplified, being present in all three sensitive attributes. FNR disparities were observed for:

- number people maintenance/housing: for all individuals who do not own a house.

- gender/housing: for all but the privileged subgroup (men that own a house).

- number people maintenance/gender: all subgroups involving women.

Moreover FOR disparities were identified for:

- number people maintenance/housing: for the combination of 1 person in maintenance and not owning a house.

- gender/housing: for all individuals who do not own a house.

The detected disparities can be observed in 5.4

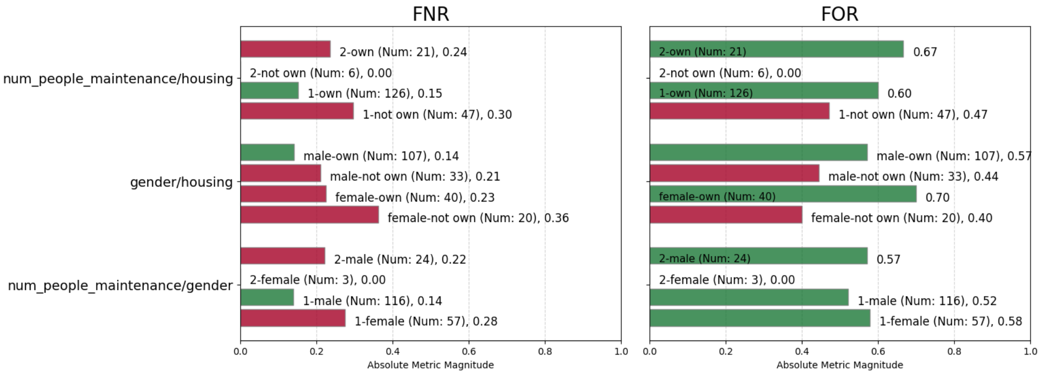


Figure 5.4: Detected disparities for the intersectional case. The FNR disparities fail for all subgroups, the exception being the privileged groups and the low-represented subgroups that had FNR=0.

Diabetes130US dataset

For this dataset, the results are based on predictions made by the Catboost classifier, which achieved an accuracy of 63% on the test sets in both cases. The table below outlines the privileged and unprivileged groups and subgroups.

Attribute	Privileged group/subgroup
gender	female
race	Caucasian
age	More than 70 years old
gender/race	Female, Caucasian
gender/age	Female, More than 70 years old
race/age	Caucasian, More than 70 years old

Table 5.6: Privileged groups and subgroups, Diabetes130US dataset

Original dataset: Overall, the model performed fairly well for this dataset, with the exception being FNR for the age attribute (showing disparities for both unprivileged age groups) as seen in Figure 5.5.

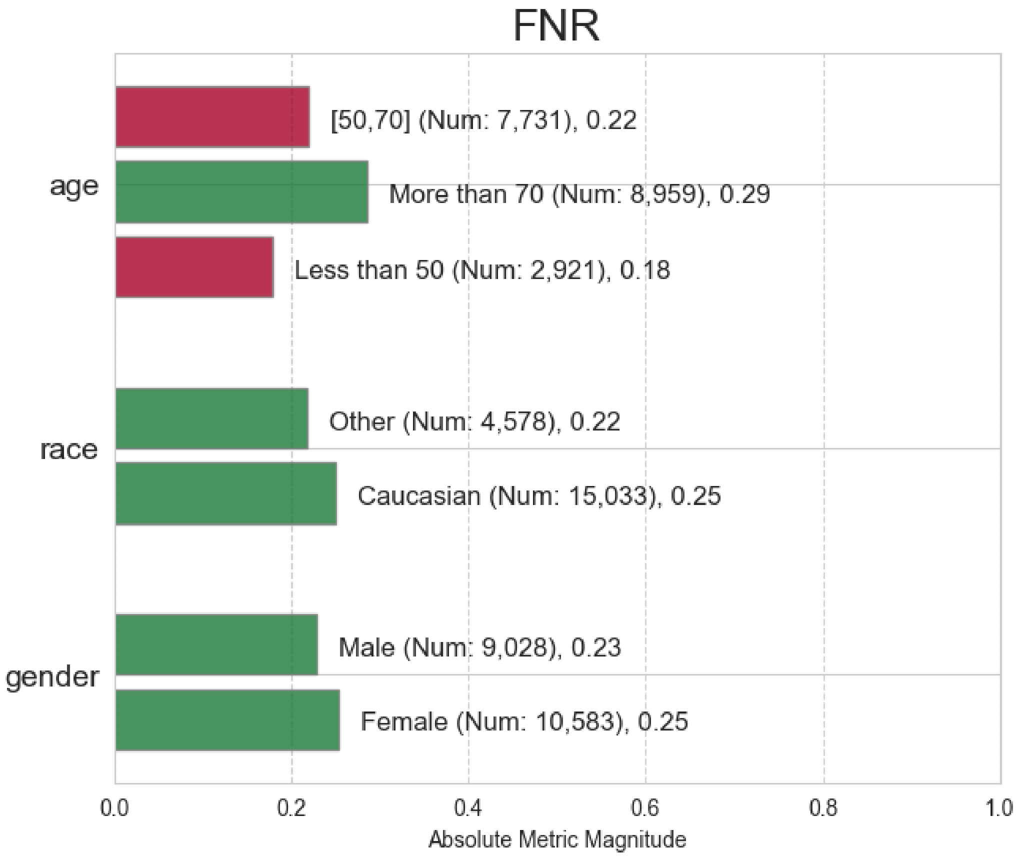


Figure 5.5: Detected disparities for the non-intersectional case. In this case, the only failing disparities found were for FNR regarding the age attribute.

Intersectional dataset: The results were similar to the original dataset, however, disparities emerged for FNR across all three sensitive attributes, even though such disparities were absent for gender and race in the original dataset. Specifically, this parity fails for all combinations involving the ”Other” race, as well as for combinations containing the age groups ”Less than 50 years old” and ”Between 50 and 70 years old”, as shown in Figure 5.6. Overall, the combinations of attributes with the unprivileged age groups and the race category ”Other” lead to the largest FNR disparities, placing individuals in these subgroups at a higher risk of being incorrectly diagnosed as requiring hospital readmission.

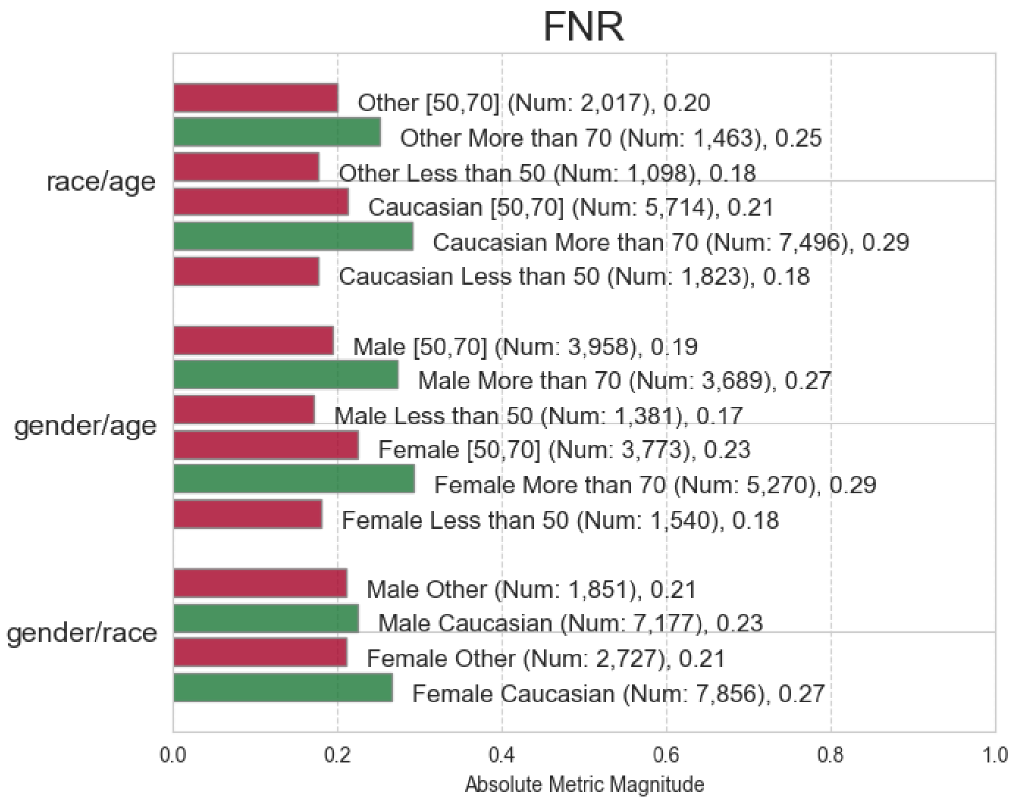


Figure 5.6: Detected disparities for the intersectional case. Notable is the new bias that emerged in the gender/race group. In the non-intersectional case, the FNR parities were met for gender, respectively for race.

Law School Bar Exam dataset

The best performing model, in this case, was the Decision Tree classifier, achieving an accuracy of 88% in both the non-intersectional and intersectional cases. The following table outlines the privileged and unprivileged groups and subgroups for each sensitive attribute.

Attribute	Privileged group/subgroup
race	white
family income	category 4
last	score ≥ 37
race/family income	white, category 4
race/lsat	white, score ≥ 37
lsat/family income	score ≥ 37, category 4

Table 5.7: Privileged groups and subgroups, Law Bar School Exam dataset

Original dataset: TPR, and precision were found to be fair across all groups. However, significant disparities were noticed in FPR and FNR, which may be influenced by the imbalanced target variable. The FNR values for all three privileged groups were very low, in some cases even 0, which made it impossible or difficult to calculate the ratio value between the unprivileged and privileged groups. In this case, the most disadvantaged subgroups were those involving the race category ”Other” and the family income categories 1 and 2 (lowest income levels). For FPR, the privileged groups had higher chances of being predicted as positive even though it was not the case. Figure 5.7 highlights the detected disparities.

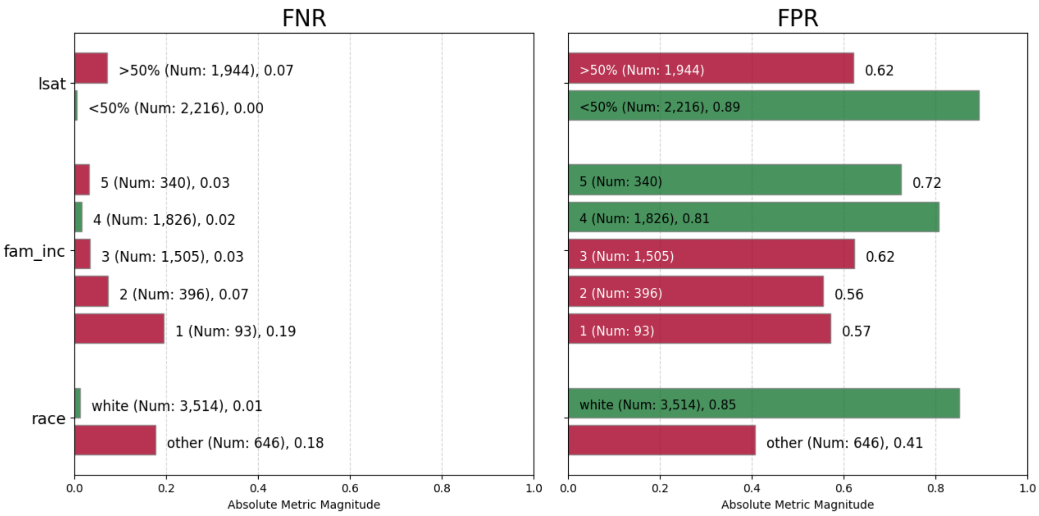


Figure 5.7: Detected disparities for non-intersectional case, highlighting the disparities in FNR and FPR. FNR fails for all subgroups but the privileged ones. Similar results are for FPR ratios.

Intersectional dataset: The large number of resulting intersectional subgroups led to under-representation in certain subgroups, as seen in 5.8, which impacted the reliability of the fairness metrics calculated for these groups. However, precision remained fair across all intersectional groups. In the case of TNR unfairness was detected in some subgroups that resulted from the combination of the "Other" race category. In contrast, in the combination of groups that contained the "White" race category, the performance remained fair. The FNR values showed the same issue, as in the original dataset, of low values for the privileged groups, again leading to poorly defined ratios.

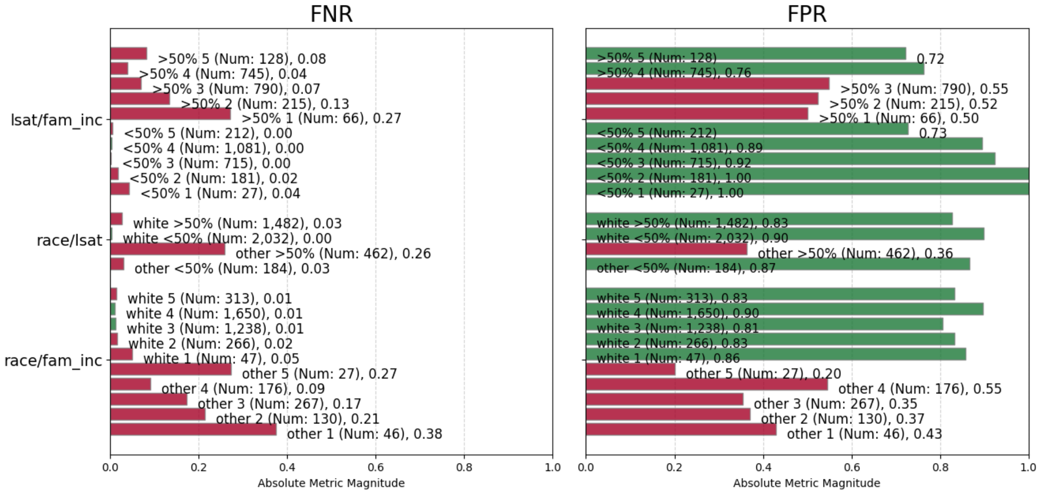


Figure 5.8: Detected FNR and FPR disparities for the intersectional case. Notable is the low values in FNR for the privileged groups (almost 0-close values), which cause high disparities in the unprivileged groups.

Table 5.8 summarizes the key observations from the bias detection analysis in the intersectional case, highlighting instances of amplified and newly detected biases across datasets.

Dataset	Intersectional Bias Observations
ACS Income	Bias amplification occurred in TPR, TNR, FPR, and FNR. New bias was detected in TPR and TNR for women in the race category "Other."
German Credit Risk	Bias amplification occurred in FNR and FOR, with smaller subgroups experiencing greater disparities.
Diabetes180US	Bias amplification occurred in FNR. New bias emerged for patients in the race category "Other," affecting both men and women.
Law School Bar Exam	Bias amplification occurred in FNR and FPR. The class imbalance caused significant disparities in FNR.

Table 5.8: Overview of the findings in terms of intersectional bias

5.3 RQ3: To what extent can existing bias mitigation methods reduce unfair bias in the case of intersectional groups, and what are the implications on model performance?

To tackle the last research question, nine different bias mitigation methods were applied to each dataset. Depending on the prediction task, the focus of the mitigation changed. This section presents the results for each dataset, highlighting the changes in various metrics. For simplicity and clarity, the disparities across the selected sensitive attributes are summarized using their average, and a comparison is made between the intersectional case and the results after applying the mitigation methods. The average of disparities is calculated as below:

$$\text{avg\_disp\_metric} = \text{mean}(\sum_{i=1}^{N_A} \frac{\text{metric\_rate\_unprivileged}}{\text{metric\_rate\_privileged}}),$$

Where:

- N_A : Number of sensitive attributes
- metric_rate: Any metrics such as FPR, FNR, etc.

Acs Income dataset

For this dataset, the focus was to improve all four metrics that showed disparities: TPR, TNR, FNR, and FPR.

Method	**TPR**	**TNR**	**FNR**	**FPR**
Intersectional	0.8	1.35	2.57	0.5
Massaging Method	0.89	1.04	4.68	0.55
Prevalence Sampling	1.02	1.01	0.92	0.95
Data Repairer	0.75	1.4	2.9	0.49
Label Flip	1	1	1	0.82
Fairlean classifier	0.83	1.21	1.5	0.64
GerryFair	0.97	2.67	3.94	0.83
Threshold Optimizer	0.99	0.99	1	1
Group Threshold	0.99	1	0.98	0.99
Equalized Odds Method	0.92	1.16	1.37	0.82

Table 5.9: The average of TPR, TNR, FPR and FNR disparities across the original intersectional dataset and all applied bias mitigation. Methods that improved these disparities the best are highlighted in bold.

Prevalence Sampling and Label Flip methods were the most effective in reducing disparities and achieving fair results, without any significant trade-offs. Due to the large size of the dataset, the Prevalence Sampling method managed to undersample the dataset and created an evenly balanced class distribution within each subgroup. This resulted in much fairer outcomes, highlighting the importance of a balanced dataset. Massaging achieved the highest accuracy by improving both TNR and TPR, but it led to higher disparities in the FNR. Similarly, Fairlearn classifier achieved TPR and TNR parity, but it affected the disparities in FPR and FNR. The Threshold Optimizer method achieved parity in all four metrics, but it reduced the accuracy. Finally, the Equalized Odds method improved both the TPR and FPR, ensuring equalized odds parity. The following figure highlights the model's accuracy across the applied bias mitigation methods.

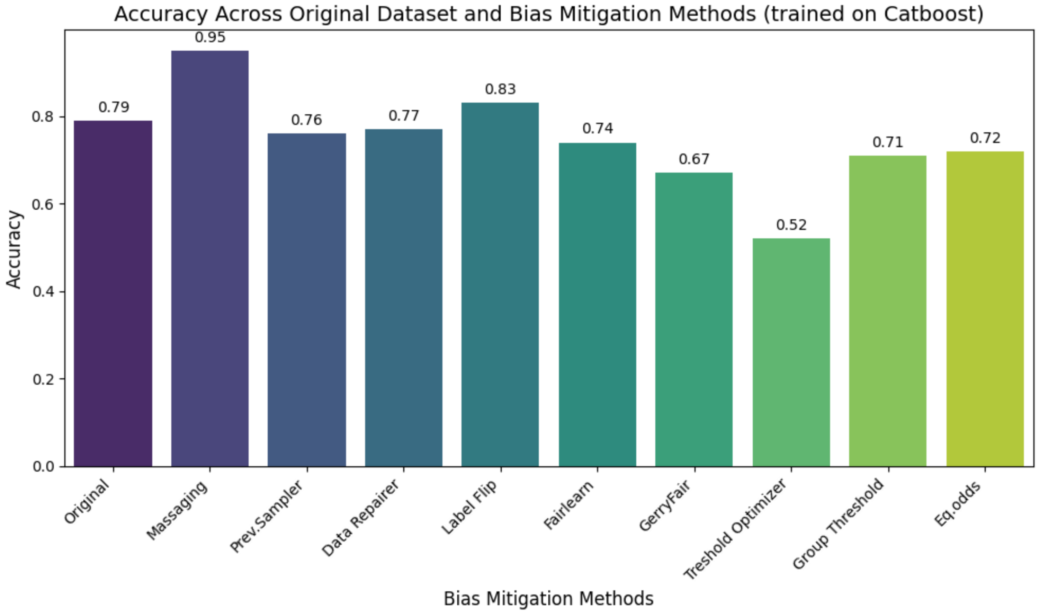


Figure 5.9: Accuracy across bias mitigation methods, ACS Income dataset

German Credit Risk dataset

The focus for improvement in this case was the FNR and FOR since disparities in these two indicates that people from unprivileged subgroups have a higher chance of not receiving a bank credit, even though they should. The following table shows the average of the FNR and FOR disparities and how they change across the bias mitigation methods.

Method	**FNR**	**FOR**
Intersectional	1.41	0.75
Massaging Method	1.41	1.52
Prevalence Sampling	0.8	0.78
Data Repairer	2.23	0.91
Label Flip	2.15	1.77
Fairlean classifier	0.94	2.35
GerryFair	8.8	1.5
Threshold Optimizer	1.87	0.62
Group Threshold	0.95	0.79
Equalized Odds Method	0.9	0.36

Table 5.10: The average of FPR and FOR disparities across the original intersectional dataset and all applied bias mitigation, using Logistic Regression as classifier. Methods that improved these disparities the most are highlighted in bold.

Due to the size of the dataset (1000 entries), it was challenging to achieve better results with the mitigation methods. In general, the methods managed to reduce the values of FNR at the cost of increasing the TNR disparities. The FairLearn classifier, as well as the threshold modifying post-processing methods, reduced the FNR disparities the most, however they came at the cost of reducing the model's accuracy. The difference between the accuracies over the applied methods can be seen in figure 5.10.

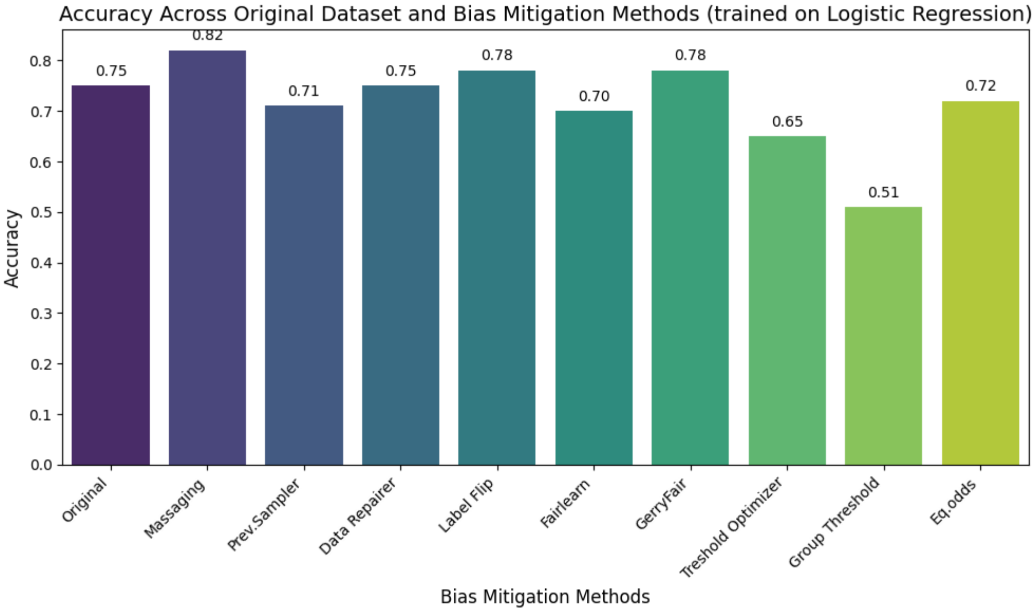


Figure 5.10: Accuracy across bias mitigation methods, German Credit Risk dataset

Diabetes130US dataset

For this dataset, the most significant disparities were observed in the FNR, making it the primary focus for improvement when applying the bias mitigation methods. Disparities in FNR indicate that for certain subpopulations, the algorithm disproportionately misclassified patients as needing readmission to the hospital, leading to unfair treatment.

The table below presents the average FNR disparities for the original dataset and the changes observed in this metric after applying the bias mitigation methods.

Method	**FNR**
Intersectional	0.74
Massaging Method	0.88
Prevalence Sampling	0.86
Data Repairer	0.77
Label Flip	0.87
Fairlean classifier	0.76
GerryFair	0.7
Threshold Optimizer	1.01
Group Threshold	0.89
Equalized Odds Method	0.87

Table 5.11: FNR average disparity across all applied bias mitigation methods using Catboost as classifier. Methods that improved the FNR the most are highlighted in bold.

Prevalence Sampling, Massaging and Label Flip methods met the FNR parity for the race/gender attribute and improved the average FNR for the other two sensitive attributes. The reduction in the FNR resulted in a slight impact on the PPR disparities. GerryFair and Threshold Optimizer reduced the FNR to values close to zero but increased the disparities in the TNR and GerryFair impacted the accuracy. Group Threshold method met the FNR parity for all subgroups but it affected the accuracy the most, with a drop of -8%. Figure 5.11 shows the accuracies obtained for each bias mitigation method.

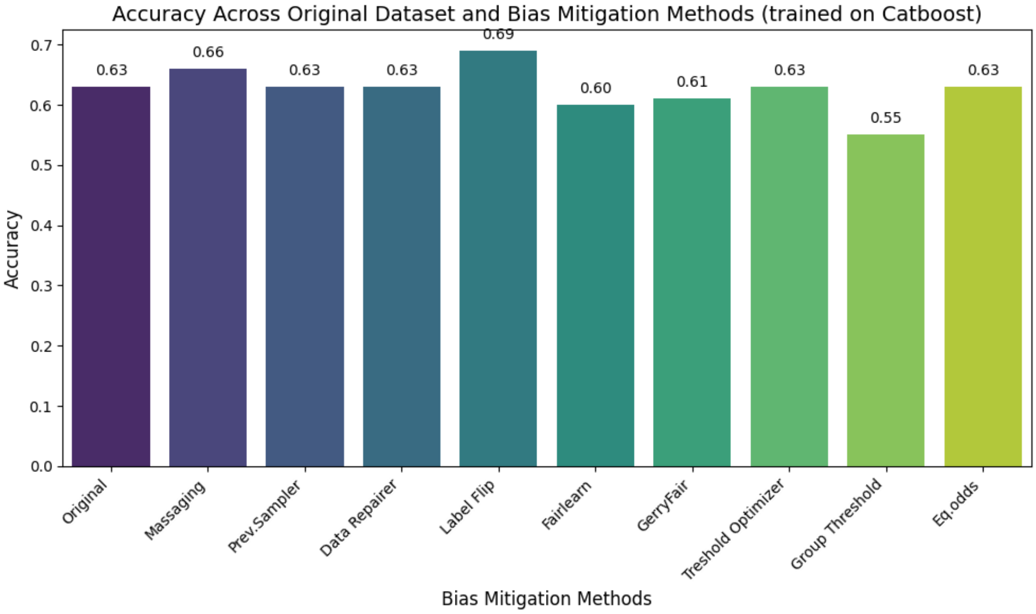


Figure 5.11: Accuracy across bias mitigation methods, Diabetes130US dataset

Law School Bar Exam dataset

The main focus of bias mitigation for this dataset was the reduction of FNR and FPR disparities. For this prediction task, minimizing FNR is crucial to ensure that the sensitive groups are not unfairly disadvantaged by being incorrectly labeled as failing the exam when they should be labeled as passing. Improving the FPR parities aligns with obtaining the equalized odds parity which requires equal TPR and FPR across all groups. For TPR, no improvements were needed, since all four models were capable of correctly identifying positives for all groups. The following tables present the modified average disparities for FNR, respectively, FPR after applying the bias mitigation methods.

Method	FNR	FPR
Intersectional	24.07	0.64
Massaging Method	7.6	0.72
Prevalence Sampling	0.96	1.07
Data Repairer	48,4	0.8
Label Flip	10.48	1.04
Fairlearn classifier	3.2	0.99
GerryFair	inf	1
Threshold Optimizer	0.99	1.07
Group Threshold	1.57	1.04
Equalized Odds Method	4.4	0.99

Table 5.12: FNR and FPR average disparities across all applied bias mitigation methods using Decision Tree as a classifier. Values equal to infinity (inf) represent the cases of division to 0 (FNR of privileged group = 0). Methods that improved the best results are highlighted in bold.

Due to the imbalance in the class target, the majority of bias mitigation methods struggled to improve the FNR parity. The Massaging method and Data Repairer methods reduced the values for FNR to 0-close values, the exception being for the attributes resulted from intersecting the family income of category 1 with race of category ”other”, respectively with last of category $Score > 37$. For these two, the FNR values were much larger, hence the high average for FNR. GerryFair model predicted no negative values at all, therefore the ∞ value in the FNR, respectively the value of 1 in FPR disparities. The goal of improving FPR was to meet the equalized odds parity and the majority of methods reduced the disparity. Equalized Odds method improved the FPR parity and it reduced the values of FNR to < 0.1 for all subgroups. Threshold Optimizer and Group Threshold achieved the FPR parities, however, the accuracy with Group Threshold was reduced as can be noted in Figure 5.12 below:

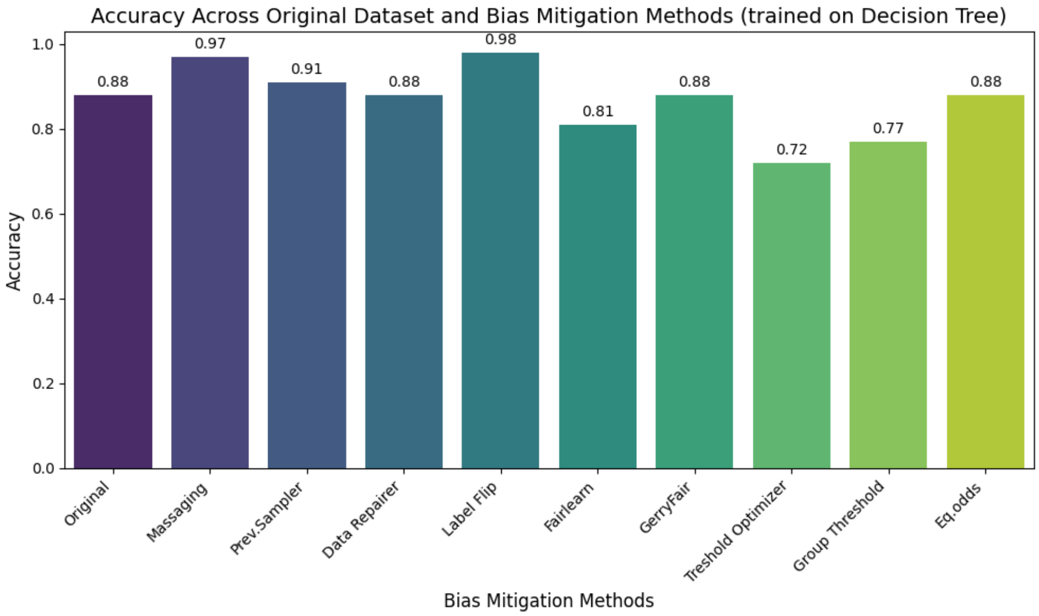


Figure 5.12: Accuracy across bias mitigation methods, Law school bar exam dataset

Table 5.13 summarizes the key observations from the bias mitigation analysis, highlighting the challenges encountered in addressing disparities across datasets.

Dataset	Challenges and Observations from Bias Mitigation
ACS Income	The large dataset size contributed to achieving the best results. However, simultaneously improving all four metrics often led to trade-offs between them.
German Credit Risk	Bias mitigation was challenging due to the small dataset size and the high number of subgroups, particularly in reducing FNR disparities.
Diabetes180US	FNR disparities were successfully improved with five bias mitigation methods without significantly affecting other metrics.
Law School Bar Exam	Bias mitigation was highly challenging due to the imbalanced class distribution, resulting in near-zero FNR values for privileged subgroups and high FNR values for unprivileged ones.

Table 5.13: Summary of challenges and outcomes in bias mitigation efforts across datasets.

5.4 Discussion

5.4.1 RQ1: To what extent can a semi-automated approach assist in identifying sensitive attributes and forming relevant intersectional groups within public datasets?

The detection algorithm successfully identified sensitive attributes in the datasets, including attributes that normally are not considered sensitive. In the German Credit Risk dataset, *housing* and the *number of people maintenance* were detected as sensitive attributes. These attributes can be seen as a sign of financial stability and significant disparities were revealed in the intersectional case.

Notable are the results regarding the Law School Bar Exam dataset, where the second highest sensitive attribute is *family income*. This attribute introduces a social-economic bias, with the lowest family income (category 1) showing the highest FNR disparities in both the non-intersectional and intersectional cases.

Together, these findings emphasize the need to expand bias detection beyond traditional protected attributes. Attributes like housing, income, or others that correlate with social or economic factors can also carry biases. When combined with protected attributes, they may amplify unfairness, leading to intersectional bias that would otherwise go undetected.

The detection algorithm also enabled the creation of intersectional groups by ranking attributes based on their likelihood of disadvantage. This approach facilitated a more focused analysis of bias, prioritizing the most critical intersections. Such a strategy is particularly valuable when the creation of a large number of intersectional groups is impractical due to imbalanced or insufficient data. While this method does not fully resolve the challenges associated with numerous intersectional attributes, it ensures attention is directed toward those most prone to amplified bias.

One challenge encountered during attribute detection was handling continuous attributes like the place of birth (*POBC*) attribute in the ACS Income dataset. Such an attribute could introduce unfair regional social-economic bias, but in this case, there is no additional information on how to aggregate this attribute into regions or states, hence it was not considered for bias analysis. Similarly, in the Diabetes180US dataset, two medical attributes (*diag_1, diag_3*) detected as sensitive could not be utilized due to their continuous nature. Such features, however, combined with gender or age, could lead to hidden bias.

Findings regarding RQ1*: The algorithm proved to be an effective tool for identifying sensitive attributes. Furthermore, it facilitated the formation of intersectional attributes, creating a balance between prioritizing attributes and avoiding having an excessive number of intersectional subgroups. Nevertheless, the ranking of the attributes requires a thorough analysis, as the number of attributes, as well as the attributes that can be considered for bias analysis, differ from one dataset to another.*

5.4.2 RQ2: Does the consideration of intersectional groups reveal or amplify hidden biases that are not evident in non-intersectional groups?

The analysis of intersectional bias revealed that unfair bias becomes more pronounced when considering intersectional cases. Across all four datasets, existing biases were amplified. Furthermore, new biases emerged for the DiabetesUS180 and ACS Income datasets. In the case of the DiabetesUS180 dataset, the FNR parity, which was fair for both race and gender in the original analysis, failed to be met for their intersectional combination. Similarly, for the ACS Income dataset, the subgroup formed by the intersection of the unprivileged sex group (women) and the race category ”Other” exhibited disparities in both TPR and TNR, even though these metrics had passed for race and sex individually in the non-intersectional case.

Figure 5.13 highlights the average disparities amplifications for each dataset.

This was calculated by averaging the detected and discussed disparities in Chapter 5. Their values were averaged separately for the non-intersectional and intersectional cases. The absolute difference between these averages was then computed to quantify the overall increase in disparities introduced by considering intersectional groups.

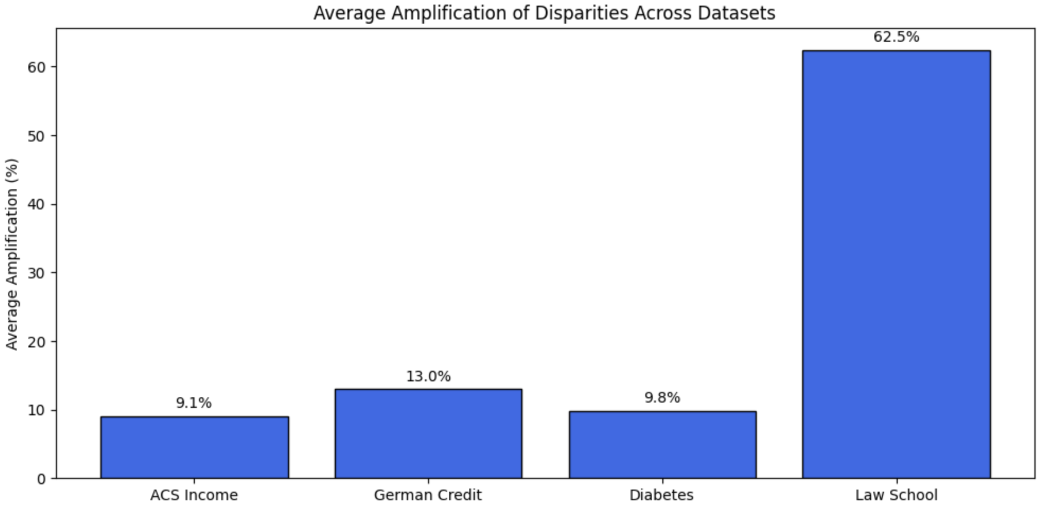


Figure 5.13: Average disparities amplifications per dataset.

Notable is the high spike in amplifications in the Law Bar School Exam dataset. This was mainly due to its imbalanced class distribution, which created higher metric disparities for smaller intersectional subgroups compared to those that had higher representation in the dataset.

Dataset size significantly influenced the intersectional bias analysis. German Credit Risk dataset contains only 1000 entries and combined with unequal distribution within the sensitive attributes groups, lead to intersectional subgroups being under-represented. This limited representation made it difficult to evaluate fairness, as disparities were magnified in these small subgroups.

These two challenges (imbalanced class distribution and data size), motivated the merging of under-represented subgroups into broader categories: For instance, in some cases, the race attribute was categorized into ”Caucasian/white” and ”Other” due to under-representations of other races. This approach allowed for fairness evaluations, but it obscured the unfair bias that occurred in specific subpopulations.

Findings regarding RQ2*: This analysis revealed that unfair bias emerges when* *evaluating fairness metrics in the intersectional case, that could be otherwise overlooked. Additionally, imbalanced class distribution and too small dataset sizes seem to be a large driver of increased disparities in the fairness metrics.*

5.4.3 RQ3: To what extent can existing bias mitigation methods reduce unfair bias in the case of intersectional groups, and what are the implications on model performance?

The applied bias mitigation methods yielded different results depending on the dataset and the metrics for which an improvement was sought. In general, Prevalence Sampling and Label Flip proved to be the most effective, without any notable trade-offs. The Massaging Method predominantly increased the accuracy and improved metrics such as TPR. However, it failed to improve FNR or FPR disparities. The in-processing methods showed limited success; FairLearn has a limited number of constraints that can be improved. Similarly, GerryFair improved metrics like TPR. However, it had little impact on improving other disparities. Among post-processing methods, the Threshold Optimizer and Equalized Odds methods were the most successful, although their performance varied by the dataset. Data size, imbalanced target variable distribution and subgroup representations affected the efficiency of the applied methods, as also noted in other works [18].

Looking at the effects of these applied methods, two types of trade-offs were identified:

- Accuracy trade-offs: FairLearn, GerryFair, Threshold Optimizer, and Group Threshold affected the model's accuracy. This effect was especially strong for the Group Threshold, which only met the desired parities at the high cost of accuracy.

- Metrics trade-offs: Were observed for various methods, depending on the mitigation task. For the Law School Bar Exam, GerryFair model predicted no negative predictions, leading to a model impractical in real-world use. The Massaging Method and FairLearn successfully improved TPR values but at the cost of affecting the disparities in FNR and FPR. This trade-off can be best noticed in the ACS Income results table 5.3.

Findings regarding RQ3*: Overall, their results underscore that bias mitigation methods can effectively reduce disparities, however, their performance depends on the nature of the datasets, as well as the method's improvement target. Moreover, it is also crucial to investigate the trade-offs between fairness metrics and overall accuracy to ensure trustworthy and useful results.*

5.5 Summary

This chapter provided a comprehensive analysis of the results. It began by presenting the sensitive attributes identified for each dataset, along with the intersectional attributes. The detection algorithm uncovered attributes that might not typically be considered sensitive, although a manual review of these detected attributes is necessary. Next, a detailed bias analysis was conducted for both cases, revealing that in intersectional scenarios, biases can be amplified, and in some cases, new biases can emerge. Finally, bias mitigation methods were applied to address the identified unfair biases. The results showed that while existing bias mitigation methods can improve outcomes, their effectiveness varies depending on the dataset.

6 Conclusions and Open Challenges

This Master's thesis analyzed intersectional unfair bias across four publicly available datasets, utilizing a semi-automatic algorithm to identify sensitive attributes at the highest risk of disadvantage and to form intersectional groups. The fairness analysis revealed that biases are amplified in the intersectional case compared to the non-intersectional groups, particularly for smaller and under-represented subgroups. Nine existing bias mitigation methods were applied and evaluated across the datasets, revealing varying levels of effectiveness. Smaller datasets, imbalanced target distributions, and under-represented subgroups posed significant challenges to achieving fairness improvements without introducing trade-offs.

These findings highlight the importance of addressing intersectional fairness in machine learning models, underscoring the need for careful selection of mitigation methods to balance fairness improvements with model performance. This work contributes insights into the possibilities of using a semi-automated detection algorithm to identify intersectional groups that might otherwise be overlooked, as well as the challenges in mitigating unfair bias detected within these groups.

6.1 Research Questions

- *RQ1: To what extent can a semi-automated approach assist in identifying sensitive attributes and forming relevant intersectional groups within public datasets?*

  The detection algorithm proved effective in identifying sensitive attributes, including those not traditionally recognized as sensitive, such as housing or family income level. Additionally, the algorithm facilitated the formation of intersectional groups by ranking attributes by their likelihood to be disadvantaged, enabling a more focused bias analysis. However, challenges remain in handling continuous attributes and balancing the number of intersectional subgroups, especially in datasets with imbalanced or insufficient data.

- *RQ2: Does the consideration of intersectional groups reveal or amplify hidden biases that are not evident in non-intersectional groups?*

  The analysis revealed that intersectional attributes amplify existing biases and can uncover new biases that are not evident in non-intersectional cases. For example, disparities in TPR and TNR emerged in specific subgroups, such as women in the "Other" race category in the ACS Income dataset. Small dataset sizes and imbalanced distributions were also significant drivers of disparities.

- RQ3: *To what extent can existing bias mitigation methods reduce unfair bias in the case of intersectional groups, and what are the implications on model performance?*

  Bias mitigation methods showed varying degrees of success depending on the dataset, the metrics targeted for improvement, and the mitigation approach. Prevalence Sampling and Label Flip were generally the most effective without major trade-offs. In contrast, some methods, such as FairLearn and GerryFair, improved specific metrics like TPR but introduced trade-offs in accuracy or other disparities. These findings underscore the need to evaluate the effectiveness of bias mitigation methods on a case-by-case basis, balancing fairness improvements with potential accuracy trade-offs. Task-specific considerations remain crucial for achieving practical and trustworthy outcomes.

6.2 Open Challenges and Future Work

6.2.1 Open Challenges

During the work for this Master's thesis, several open problems have been identified:

- Datasets availability: Finding datasets that contain sufficient and diverse entries to ensure a balanced distribution across intersectional groups is challenging. Out of the four datasets investigated in this work, only ACS Income dataset was large enough to provide a more proportional distribution of the attributes.

- Data sparsity: Under-representation in certain subpopulations leads to difficulties in accurately analyzing bias. This issue highlights the critical need to collect data as inclusively as possible, especially when evaluating fairness in intersectional groups.

- Integrating bias mitigation methods into toolkits: While the bias mitigation methods selected for this work are available and incorporated into toolkits, other bias mitigation methods developed in research are not yet integrated into user-friendly frameworks [70], [56]. This limits their practicality and applicability.

- Applicability of bias mitigation methods: Selecting the most effective bias mitigation method is challenging, as their performance often varies depending on the dataset and task. In this work, nine methods were evaluated, with results demonstrating that their effectiveness is context-specific. Even the methods that demonstrated positive results in this work may fail to reproduce similar outcomes in different contexts. This emphasizes the need for task-specific considerations when applying bias mitigation methods.

6.3 Future Work

A possible direction for future work could involve integrating a broader range of bias mitigation methods into ready-to-use toolkits, with a particular focus on methods that address intersectional fairness. This integration would streamline the process of testing and comparing various mitigation approaches under consistent conditions, providing clearer insights into their effectiveness across different datasets and tasks. Moreover, such toolkits would enhance accessibility by lowering technical barriers, enabling more practitioners to incorporate fairness considerations into their workflows. Furthermore, they would raise awareness about the importance of addressing intersectional biases, encouraging their adoption in real-world applications.

Reproducibility

The code developed as part of this thesis is made available to support reproducibility[1].

[1] https://github.com/IoanaSil14/Intersectional-fairness-in-public-datasets

Bibliography

[1] Aequitas. http://www.datasciencepublicpolicy.org/our-work/tools-guides/aequitas/. Accessed: 2023-08-20.

[2] Aequitas data requirements. http://aequitas.dssg.io/upload.html. Accessed: 2023-08-20.

[3] Ai fairness 360 (aif360). https://github.com/Trusted-AI/AIF360. Accessed: 2023-10-05.

[4] Can an algorithm hire better than a human? https://www.nytimes.com/2015/06/26/upshot/can-an-algorithm-hire-better-than-a-human.html. Accessed: 2024-09-05.

[5] Machine bias there's software used across the country to predict future criminals. and it's biased against blacks. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. Accessed: 2023-10-05.

[6] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna M. Wallach. A reductions approach to fair classification. *CoRR*, abs/1803.02453, 2018.

[7] Guilherme Alves, Fabien Bernier, Miguel Couceiro, Karima Makhlouf, Catuscia Palamidessi, and Sami Zhioua. Survey on fairness notions and related tensions. *EURO Journal on Decision Processes*, 11:100033, 2023.

[8] Ricardo Baeza-Yates. Bias on the web. *Communications of the ACM*, 61:54–61, 05 2018.

[9] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press, 2023.

[10] Richard Berk, Hoda Heidari, Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. A convex framework for fair regression, 2017.

[11] Sumon Biswas and Hridesh Rajan. Do the machine learning models on a crowd sourced platform exhibit bias? an empirical study on model fairness. In *Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, ESEC/FSE '20. ACM, November 2020.

[12] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Sorelle A. Friedler and Christo Wilson, editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 77–91. PMLR, 23–24 Feb 2018.

[13] Angel Alexander Cabrera, Will Epperson, Fred Hohman, Minsuk Kahng, Jamie Morgenstern, and Duen Horng Chau. Fairvis: Visual analytics for discovering intersectional bias in machine learning. In *2019 IEEE Conference on Visual Analytics Science and Technology (VAST)*, page 46–56. IEEE, October 2019.

[14] Toon Calders and Sicco Verwer. Three naive bayes approaches for discrimination-free classification. *Data Min. Knowl. Discov.*, 21:277–292, 09 2010.

[15] Simon Caton and Christian Haas. Fairness in machine learning: A survey. *CoRR*, abs/2010.04053, 2020.

[16] L. Elisa Celis, Amit Deshpande, Tarun Kathuria, and Nisheeth K. Vishnoi. How to be fair and diverse? *CoRR*, abs/1610.07183, 2016.

[17] Joymallya Chakraborty, Suvodeep Majumder, Zhe Wu, and Tim Menzies. Fairway: SE principles for building fairer software. *CoRR*, abs/2003.10354, 2020.

[18] Zhenpeng Chen, Jie M. Zhang, Federica Sarro, and Mark Harman. Fairness improvement with multiple protected attributes: How far are we?, 2024.

[19] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments, 2016.

[20] KEVIN A. CLARKE. The phantom menace: Omitted variable bias in econometric research. *Conflict Management and Peace Science*, 22(4):341–352, 2005.

[21] Kimberle Crenshaw. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *The University of Chicago Legal Forum*, 140:139–167, 1989.

[22] Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring adult: New datasets for fair machine learning, 2022.

[23] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard S. Zemel. Fairness through awareness. *CoRR*, abs/1104.3913, 2011.

[24] Inês Oliveira e Silva, Sérgio Jesus, Hugo Ferreira, Pedro Saleiro, Inês Sousa, Pedro Bizarro, and Carlos Soares. Fair-obnc: Correcting label noise for fairer datasets, 2024.

[25] Michael Feldman, Sorelle Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact, 2015.

[26] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, page 259–268, New York, NY, USA, 2015. Association for Computing Machinery.

[27] C. Ferri, José Hernández-Orallo, and R. Modroiu. An experimental comparison of performance measures for classification. *Pattern Recognit. Lett.*, 30:27–38, 2009.

[28] James R. Foulds and Shimei Pan. An intersectional definition of fairness. *CoRR*, abs/1807.08362, 2018.

[29] David Freedman, Robert Pisani, and Roger Purves. Statistics (international student edition). *Pisani, R. Purves, 4th edn. WW Norton & Company, New York*, 2007.

[30] Sorelle A. Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P. Hamilton, and Derek Roth. A comparative study of fairness-enhancing interventions in machine learning, 2018.

[31] Andreas Fuster, PAUL GOLDSMITH-PINKHAM, Tarun Ramadorai, and ANSGAR WALTHER. Predictably unequal? the effects of machine learning on credit markets. *The Journal of Finance*, 77, 10 2021.

[32] Pratyush Garg, John D. Villasenor, and Virginia Foggo. Fairness metrics: A comparative analysis. *CoRR*, abs/2001.07864, 2020.

[33] Usman Gohar and Lu Cheng. A survey on intersectional fairness in machine learning: Notions, mitigation, and challenges. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, IJCAI-2023. International Joint Conferences on Artificial Intelligence Organization, August 2023.

[34] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. *ArXiv*, abs/1610.02413, 2016.

[35] Vasileios Iosifidis, Besnik Fetahu, and Eirini Ntoutsi. FAE: A fairness-aware ensemble framework. *CoRR*, abs/2002.00695, 2020.

[36] Sérgio Jesus, Pedro Saleiro, Beatriz M Jorge, Rita P Ribeiro, João Gama, Pedro Bizarro, Rayid Ghani, et al. Aequitas flow: Streamlining fair ml experimentation. *arXiv preprint arXiv:2405.05809*, 2024.

[37] Kory D. Johnson, Dean P. Foster, and Robert A. Stine. Impartial predictive modeling and the use of proxy variables, 2022.

[38] P. Kamakshi and A. Vinaya Babu. Automatic detection of sensitive attribute in ppdm. In *2012 IEEE International Conference on Computational Intelligence and Computing Research*, pages 1–5, 2012.

[39] Faisal Kamiran and Toon Calders. Data pre-processing techniques for classification without discrimination. *Knowledge and Information Systems*, 33, 10 2011.

[40] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. Fairness-aware classifier with prejudice remover regularizer. pages 35–50, 09 2012.

[41] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2564–2572. PMLR, 10–15 Jul 2018.

[42] Hae-Young Kim. Statistical notes for clinical researchers: Chi-squared test and fisher's exact test. *Restorative Dentistry and Endodontics*, 42:152–155, 05 2017.

[43] Hannah Rose Kirk, Yennie Jun, Filippo Volpin, Haider Iqbal, Elias Benussi, Frederic Dreyer, Aleksandar Shtedritski, and Yuki Asano. Bias out-of-the-box: An empirical analysis of intersectional occupational biases in popular generative language models. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 2611–2624. Curran Associates, Inc., 2021.

[44] Youjin Kong. Are ?intersectionally fair? ai algorithms really fair to women of color? a philosophical analysis. *Facct: Proceedings of the Acm Conference on Fairness, Accountability, and Transparency*, pages 485–494, 2022.

[45] Hemank Lamba, Kit T. Rodolfa, and Rayid Ghani. An empirical comparison of bias reduction methods on real-world problems in high-stakes policy settings. *CoRR*, abs/2105.06442, 2021.

[46] Vincent LeBlanc and Michael Cox. Interpretation of the point-biserial correlation coefficient in the context of a school examination. *The Quantitative Methods for Psychology*, 13:46–56, 01 2017.

[47] Michael Lohaus, Michael Perrot, and Ulrike Von Luxburg. Too relaxed to be fair. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 6360–6369. PMLR, 13–18 Jul 2020.

[48] Kristian Lum and James Johndrow. A statistical framework for fair predictive algorithms, 2016.

[49] Gaurav Maheshwari, Aurélien Bellet, Pascal Denis, and Mikaela Keller. Fair without leveling down: A new intersectional fairness definition, 2023.

[50] Denton NA Massey DS. *American Apartheid : Segregation and the Making of the Underclass.* Harvard University Press, 1993.

[51] Konstantinos Mavrogiorgos, Athanasios Kiourtis, Argyro Mavrogiorgou, Andreas Menychtas, and Dimosthenis Kyriazis. Bias in machine learning: A literature review. *Applied Sciences*, 14(19), 2024.

[52] Robert McEliece. *The theory of information and coding*, volume 3. Cambridge University Press, 2002.

[53] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Comput. Surv.*, 54(6), jul 2021.

[54] Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application*, 8(1):141–163, March 2021.

[55] Tom M Mitchell. *Machine learning*, volume 1. McGraw-hill New York, 1997.

[56] Giulio Morina, Viktoriia Oliinyk, Julian Waton, Ines Marusic, and Konstantinos Georgatzis. Auditing and achieving intersectional fairness in classification problems, 2020.

[57] Azadeh Nematzadeh, Giovanni Luca Ciampaglia, Filippo Menczer, and Alessandro Flammini. How algorithmic popularity bias hinders or promotes quality. *CoRR*, abs/1707.00574, 2017.

[58] S.F. Olejnik and James Algina. Generalized eta and omega squared statistics: Measures of effect size for some common research designs. *Psychological methods*, 8:434–47, 12 2003.

[59] Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kıcıman. Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data*, 2, 2019.

[60] Jonathan Ortigosa-Hernández, Iñaki Inza, and Jose A. Lozano. Measuring the class-imbalance extent of multi-class problems. *Pattern Recognition Letters*, 98:32–38, 2017.

[61] Tiago Palma Pagano, Rafael Bessa Loureiro, Fernanda Vitória Nascimento Lisboa, Gustavo Oliveira Ramos Cruz, Rodrigo Matos Peixoto, Guilherme Aragão de Sousa Guimarães, Lucas Lisboa dos Santos, Maira Matos Araujo, Marco Cruz, Ewerton Lopes Silva de Oliveira, Ingrid Winkler, and Erick Giovani Sperandio Nascimento. Bias and unfairness in machine learning models: a systematic literature review, 2022.

[62] Eliana Pastor, Elena Baralis, and Luca de Alfaro. A hierarchical approach to anomalous subgroup discovery. In *2023 IEEE 39th International Conference on Data Engineering (ICDE)*, pages 2647–2659, 2023.

[63] Eliana Pastor and Francesco Bonchi. Intersectional fair ranking via subgroup divergence. *Data Mining and Knowledge Discovery*, 38:1–37, 05 2024.

[64] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon M. Kleinberg, and Kilian Q. Weinberger. On fairness and calibration. *CoRR*, abs/1709.02012, 2017.

[65] Eric Potash, Joe Brew, Alexander Loewi, Subhabrata Majumdar, Andrew Reece, Joe Walsh, Eric Rozier, Emile Jorgenson, Raed Mansour, and Rayid Ghani. Predictive modeling for public health: Preventing childhood lead poisoning. 08 2015.

[66] Ronaldo Prati, Gustavo Batista, and Diego Silva. Class imbalance revisited: A new experimental setup to assess the performance of treatment methods. *Knowledge and Information Systems*, 45:245–279, 10 2014.

[67] Pedro Saleiro, Benedict Kuester, Abby Stevens, Ari Anisfeld, Loren Hinkson, Jesse London, and Rayid Ghani. Aequitas: A bias and fairness audit toolkit. *arXiv preprint arXiv:1811.05577*, 2018.

[68] Samira Samadi, Uthaipon Tantipongpipat, Jamie Morgenstern, Mohit Singh, and Santosh Vempala. The price of fair pca: One extra dimension, 2018.

[69] Skipper Seabold and Josef Perktold. statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*, 2010.

[70] Changjian Shui, Gezheng Xu, Qi Chen, Jiaqi Li, Charles Ling, Tal Arbel, Boyu Wang, and Christian Gagné. On learning fairness and accuracy on multiple subgroups, 2022.

[71] Harini Suresh and John V. Guttag. A framework for understanding unintended consequences of machine learning. *CoRR*, abs/1901.10002, 2019.

[72] Latanya Sweeney. Discrimination in online ad delivery. *CoRR*, abs/1301.6822, 2013.

[73] Kai Ming Ting. *Confusion Matrix*, pages 209–209. Springer US, Boston, MA, 2010.

[74] Joaquin Vanschoren, Jan N. van Rijn, Bernd Bischl, and Luis Torgo. Openml: networked science in machine learning. *SIGKDD Explorations*, 15(2):49–60, 2013.

[75] Sahil Verma and Julia Rubin. Fairness definitions explained. page 1–7, New York, NY, USA, 2018. Association for Computing Machinery.

[76] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones,

Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020.

[77] Angelina Wang, Vikram V Ramaswamy, and Olga Russakovsky. Towards intersectionality in machine learning: Including more identities, handling underrepresentation, and performing evaluation. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 336–349. ACM, June 2022.

[78] Hao Wang, Berk Ustun, and Flávio P. Calmon. Repairing without retraining: Avoiding disparate impact with counterfactual distributions. *CoRR*, abs/1901.10501, 2019.

[79] L.F. Wightman, H. Ramsey, and Law School Admission Council. *LSAC National Longitudinal Bar Passage Study*. LSAC research report series. Law School Admission Council, 1998.

[80] Forest Yang, Moustapha Cisse, and Sanmi Koyejo. Fairness with overlapping groups, 2020.

[81] Kai Zhang and Xiaoqian Jiang. Sensitive data detection with high-throughput machine learning models in electrical health records, 2023.

Appendix

Appendix

Visualization Tool for AI Trustworthiness

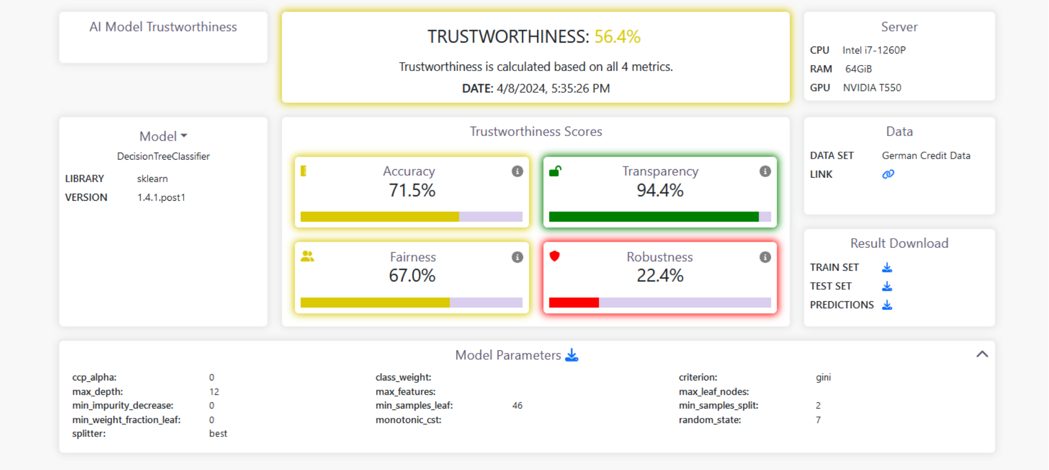
This section presents a work-in-progress visualization tool designed for evaluating the trustworthiness of machine learning models based on four key performance indicators: *accuracy, fairness, robustness and transparency.* The primary aim of this toolkit is to provide a comprehensive and reproducible way to evaluate and compare the trustworthiness of machine learning models.

The tool allows users to upload their datasets (train, test, and prediction data), machine learning models that were trained on the data, and associated parameters for analysis and comparison. This supports reproducibility and ensures a transparent evaluation process across the four defined dimensions.

Screenshots and Examples

Figure 1 displays the visualization tool. This example shows the results for the German Credit Risk dataset, for which the prediction task is to decide if a person represents a ”good” or a ”bad” credit risk. The four dimensions represent the following:

- Accuracy: The model’s predictive accuracy based on the test set.
- Fairness: Evaluated in this example as the highest FNR between unprivileged groups and privileged group within the selected sensitive attributes.
- Robustness: The fraction of features that need to be changed to alter the classifier’s decision.
- Transparency: The explainability of the model, measured by the number of important features that contribute to the decisions.



AI Model Trustworthiness

TRUSTWORTHINESS: 56.4%

Trustworthiness is calculated based on all 4 metrics.

DATE: 4/8/2024, 5:35:26 PM

Server

CPU	Intel i7-1260P
RAM	64GiB
GPU	NVIDIA T550

Model

DecisionTreeClassifier

LIBRARY	sklearn
VERSION	1.4.1.post1

Trustworthiness Scores

Accuracy 71.5%

Transparency 94.4%

Fairness 67.0%

Robustness 22.4%

Data

DATA SET	German Credit Data
LINK	

Result Download

TRAIN SET

TEST SET

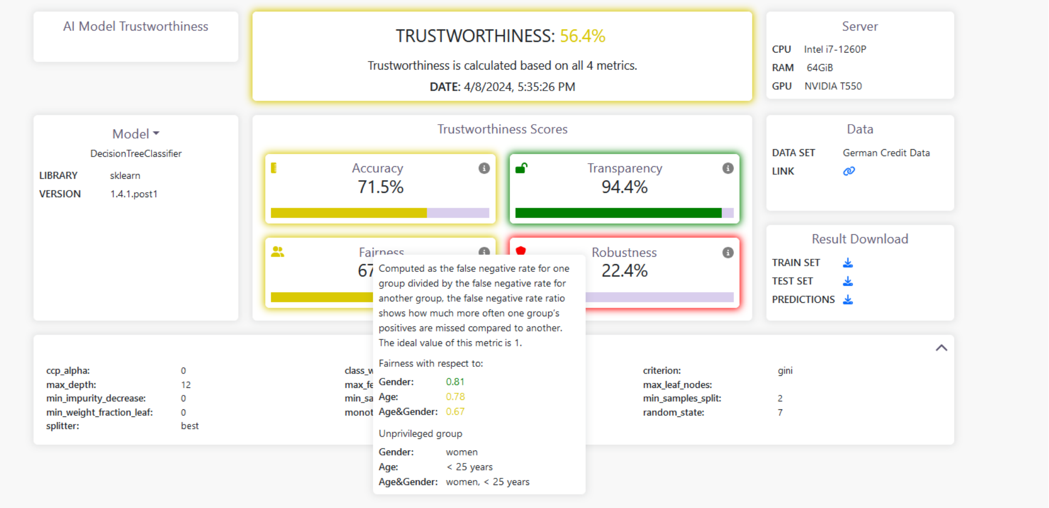
PREDICTIONS

Model Parameters

ccp_alpha:	0	class_weight:		criterion:	gini
max_depth:	12	max_features:		max_leaf_nodes:	
min_impurity_decrease:	0	min_samples_leaf:	46	min_samples_split:	2
min_weight_fraction_leaf:	0	monotonic_cst:		random_state:	7
splitter:	best				

Figure 1: Overview of the visualization tool

Figure 2 displays in more detail the fairness domain of the visualization tool. The selected sensitive attributes for analysis were gender, age, and their intersection (gender/age). The unprivileged groups for each sensitive attribute is mentioned in the details window. The overall fairness is calculated in this case as the average of the three FNR disparity values.



AI Model Trustworthiness

TRUSTWORTHINESS: 56.4%

Trustworthiness is calculated based on all 4 metrics.

DATE: 4/8/2024, 5:35:26 PM

Server

CPU	Intel i7-1260P
RAM	64GiB
GPU	NVIDIA T550

Model

DecisionTreeClassifier

LIBRARY	sklearn
VERSION	1.4.1.post1

Trustworthiness Scores

Accuracy 71.5%

Transparency 94.4%

Fairness 67

Robustness 22.4%

Computed as the false negative rate for one group divided by the false negative rate for another group, the false negative rate ratio shows how much more often one group's positives are missed compared to another. The ideal value of this metric is 1.

Fairness with respect to:

Gender:	0.81
Age:	0.78
Age&Gender:	0.67

Unprivileged group

Gender:	women
Age:	< 25 years
Age&Gender:	women, < 25 years

Data

DATA SET	German Credit Data
LINK	

Result Download

TRAIN SET

TEST SET

PREDICTIONS

ccp_alpha:	0	class_w		criterion:	gini
max_depth:	12	max_fe		max_leaf_nodes:	
min_impurity_decrease:	0	min_sa		min_samples_split:	2
min_weight_fraction_leaf:	0	monot		random_state:	7
splitter:	best				

Figure 2: Detail of the presented fairness metrics