POLITECNICO DI TORINO

Master's Degree in Computer Engineering

Master's Degree Thesis

Towards fairness AI: A data-centric approach

Assessment and mitigation techniques to tackle bias in datasets

Supervisor

Prof. Antonio Vetrò

Candidate

Uditi Ojha

Company supervisor
Clearbox AI
Carmine D'Amico

Anno accademico 2021-2022

Summary

The consequences of bias and injustice have received more attention, even though AI is increasingly employed in delicate fields like health care, hiring, and criminal justice. We know that individual and social biases, which are frequently unconscious, affect and skew human decision-making in many ways. Although it might seem that using data to automate judgments would guarantee fairness, we now know that this is untrue. Societal bias can be incorporated into training datasets for AI, decisions made even during the machine learning development stage, and intricate feedback loops that form when a machine learning model is used in the real world.

We aim to anticipate unfairness before applying any algorithm by studying the bias associated with protected attributes such as age, ethnicity, gender, education, marital status, etc. We start by using the various bias measure metrics taken references from [1] [2] [3]. The study is carried out at different stages to evaluate how well the bias measure metrics perform on five chosen datasets from the social and financial domains.

This work, enclosed in the broader context of Data-Centric AI adopts Data Bias Assessment and mitigation techniques. As a result, we begin by developing a library for data bias assessment and comprehending several bias mitigation strategies. In particular, there are three categories of Bias Measure Metrics: Balance Measure Metrics(Gini, Simpson, Shannon, and Imbalance Ratio), Equality Measure Metrics(Generalized Entropy Index, Theil Index, Atkinson Index, and Coefficient of Variation), and Distance Measure Metrics(Infinity Norm Distance and Total

Variation Distance).The Balance measure seeks to determine if a particular class of a protected attribute is balanced. The Equality measure metrics desire to determine if a specific class of a protected attribute is handled equally, and finally, Distance measure metrics aim to determine whether the protected attribute distribution is close to the target reference distribution.

Synthetic data incorporates all the statistical and distribution properties of the original dataset.The use of synthetic data can improve AI and solve various data-related properties. In our study, we employed synthetic datasets to determine whether utilizing synthetic datasets could lessen data bias. Three distinct vendors provide the synthetic datasets (Syndata, Mostly AI, Gretel). We will evaluate the performance of our bias measure metrics on the synthetic datasets.

Finally, different bias mitigation approaches, primarily related to pre-processing bias mitigation approaches, have been applied on the original dataset. These pre-processing bias mitigation strategies are taken from Synthesized SDK and AI Fairness Toolkit 360. We will assess the performance of our bias measure metrics on the debiased datasets created by utilizing the various pre-processing bias mitigation techniques.

These experimental evaluations induced insights and considerations by comparing the different pre-processing techniques and shone a light on the possible directions that the scientific literature could take to further assess bias and mitigate data bias in the context of trustworthy and Data-centered AI.

Acknowledgements

This thesis was developed at Clearbox AI, therefore I would like to warmly thank my company supervisors Carmine D'Amico for his valuable support and availability as well as the rest of the team for giving me the opportunity to carry out this work and providing me interesting insights and different point of views which contributed to the finalization of this thesis. Moreover, I would like extend my gratitude to Prof. Antonio Vetrò for supervising this work and all the professors which I met during my attendance of the Computer Science and Engineering Master's Degree for conveying knowledge and insights through their teaching, research and expertise.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation

As mentioned in [4] one of the challenges of artificial intelligence is ensuring that model decisions are fair and without bias. Datasets, metrics, techniques, and tools are used in research to detect and mitigate unfairness and bias [4].

Prediction-based decision algorithms are widely used in industry by governments and organizations that are rapidly adopting them [5]. These techniques are already widely used in lending, contracting, and online advertising, as well as criminal pre-trial proceedings, immigration detention, and public health [6]. With the rise of these techniques came concern about the biases embedded in the models and how fair they are in defining their performance for issues pertaining to sensitive social aspects such as race, gender, class, and so on [7].

Systems that have an impact on people's lives raise ethical concerns about making fair and unbiased decisions. As a result, challenges to bias and unfairness have been thoroughly investigated, taking into account the constraints imposed by corporate practices, regulations, social traditions, and ethical obligations [8]. Recognizing and reducing bias and unfairness are difficult tasks because unfairness differs across cultures. As a result, the unfairness criterion is influenced by user experience, cultural, social, historical, political, legal, and ethical considerations [9].

In machine learning, algorithmic bias is frequently discussed, but nonetheless the underlying data, not the algorithm, is typically the primary source of bias. The most severe issue with machine learning models is that the training distribution does not always correspond to the desired distribution. If the current reality places some people at a systematic disadvantage, the training data distribution will likely replicate that disadvantage rather than reflect a more equitable future. These choices are reflected in the training data and later baked into future machine learning model decisions.

Many data scientists believe that when building machine learning models, they can simply remove protected attributes (such as race, gender, and age) to avoid unfair bias. However, many features are overly correlated with protected attributes, making it simple to reconstruct a protected attribute such as ethnicity even after removing it from your training set.

Data injustice is a representation of reality. The issue with our data may also be resolved if we address the real root cause.To begin with, we must comprehend the causes of an unjust model. They comprise proxy variables, biased datasets, and data with embedded historical injustice.

Nonetheless, the literature has identified several causes of unfairness in machine learning, according to [10]:

- Biases are already present in learning datasets, which are based on biased device measurements, historically biased human decisions, inaccurate reports, or other factors. Machine learning algorithms are essentially designed to replicate these biases;

- Missing data biases, such as missing values or sample/selection biases, result in datasets that are not representative of the target population;

- Biases resulting from algorithmic goals that aim to minimize overall aggregated prediction errors and thus benefit majority groups over minorities;

- Biases for sensitive attributes caused by "proxy" attributes. Sensitive characteristics, such as race, gender, and age, distinguish privileged and unprivileged groups and are typically not appropriate for use in decision making. Non-sensitive attributes that can be used to derive sensitive attributes are known as proxy attributes. If the dataset contains proxy attributes, the machine learning algorithm can make decisions based on the sensitive attributes while masquerading as using presumably legitimate attributes.

There are numerous approaches for recognizing bias and unfairness, known as fairness metrics [11], and this wide range makes it difficult to choose the appropriate assessment criteria for the issue at hand. Some solutions, such as AIF360 [1], FairLearn [12] and Aequitas [13] already strive to assist developers by providing particular libraries and tools to address bias and unfairness.

1.2 Context of Use

The thesis was conducted in ClearboxAI under the supervision of Carmine D'Amico and Prof. Antonio Vetro, with the goal of understanding data bias and approaches to tackle them. So we start by creating a library for data bias assessment and understanding various bias mitigation techniques. Specifically for this study we used references from other research papers and open source libraries [1] [2].

We spent time in the first half of the study developing a library which consists of ten bias measure metrics divided into three categories: Balance, Equality, and Distance measures.When selecting bias measure metrics, we must keep in mind that not all metrics can be applied to all protected attributes. For example, when there is an imbalance in a class, such as the male-female ratio, it is preferable to use the balance measure metrics, whereas it is preferable to use the equality measure metrics when there is a protected attribute associated with ethnicity.

In the second half of this study, we create synthetic datasets from various vendors, specifically Syndate, MostlyAI, and Gretel. Given that each of these vendors claims to promote ethical AI, we assess

the performance of the new generated synthetic datasets using the previously designed bias measure metrics.

Then, we study some bias mitigation algorithms during the preprocessing stage of the ML pipeline lifecycle, that is, on the dataset before any training begins. Some of these bias mitigation algorithms was applied to our chosen five datasets to generate a new debiased dataset. Finally, we evaluate the performance of bias measurement metrics on our newly created debiased dataset.

The three most important steps are listed below:

- Design a bias measure metrics library.
- Create synthetic datasets from our selected five datasets from various vendors and assess the performance of the bias measure metrics on these newly created datasets.
- Understand some bias mitigation strategies during the preprocessing stage and apply a few of them to our selected five datasets to produce a debiased dataset, as well as assess the performance of the bias measure metrics.

1.3 Limitations

The use of a few restricted metrics and the analysis of a few restricted datasets are the main limitations of this work. The urgency of the situation is the primary motivator. It takes more time to conduct appropriate research and build a more comprehensive remedy since data bias has no clear explanation.

1.4 Thesis Structure

The thesis is organized as follows:

- the first chapter introduces the motivation, the context of use and the limitations and this work;

- the second chapter reviews the state of art of machine learning in particular with a focus on the data bias its causes, effects and various strategies of mitigating it particularly synthetic data and pre-processing bias mitigation techniques.
- the third chapter introduces the research method adopted by outlining the datasets that were used, the hardware setting in which the experiments run, the synthetic data generation vendors and the open source libraries for bias assessment and mitigation ;
- the fourth chapter describes the processes and methods developed during this work. In particular, it comprehends the requirements which drove the solution design, the exploration of the dataset which shaped the data bias assessment, and the mitigation techniques.
- the fifth chapter concludes the thesis with a summary of the research study done
- the last chapter includes potential future developments.

Chapter 2

State of Art

2.1 Background

2.1.1 Introduction to AI and ML

The term Artificial Intelligence(AI) describes a broad class of software based systems that interact with their environments to provide a variety of outputs, including content, predictions, suggestions, classifications, and judgements that have an impact on those environments[14]. Today's era of rapid technological development and exponential growth in extraordinary huge data sets(also known as "big data") has allowed AI to go from theoretical study to practical application on a previously unheard-of scale[15].From analyzing incredibly enormous data sets in almost real-time,to autonomous driving cars, stream history-influenced video viewing suggestions, online purchase recommendations, ads, and fraud detection, AI has fundamentally invaded many sectors of society and frequently works silently in the background of our personal electronic gadgets. The ability to reason and take actions that have the best likelihood of reaching a certain objective is the ideal quality of artificial intelligence.

The term "Machine Learning" (ML) is more specifically used to describe the field of study that gives computers the ability to learn without being explicitly programmed, [16] or to describe computer systems that use data to recognize and apply patterns or infer statistical

relationship. Regression and classification are two examples of common ML techniques. The use of ML systems to anticipate future occurrences is debatable.Additionally, ML programs may be used to provide input for other ML systems. ML is included in the definition of AI.

2.1.2 Why AI and ML important?

These days data is becoming a valuable business asset, with the amount of data generated and stored on a global scale growing at an exponential rate. Of course, collecting data is pointless if nothing is done with it, but these massive amounts of data are simply unmanageable without the assistance of automated systems.

AI/ML enables organizations to derive value from the massive amounts of data they collect by providing actionable insights, advancing system capabilities, and automating tasks. AI/ML has the potential to help businesses by assisting them in achieving measurable results such as:

- Boosting customer satisfaction
- Providing distinct digital services
- Cost-cutting measures
- Enhancing current business services

2.1.3 Use cases of AL/ML

Here we see some real-world examples of how AI/ML is being used to transform industries.

- **Healthcare:** In healthcare applications, AI/ML is being used to improve clinical efficiency, diagnosis speed and accuracy, and patient outcomes.HCA Healthcare won the Red Hat Innovation Award for using machine learning to create SPOT (Sepsis Prediction and Optimization of Therapy). This real-time predictive analytics product can detect sepsis, a potentially fatal condition, more accurately and quickly.

- **Insurance:** AI/ML is being used in the insurance industry for various applications, including automating claims processing and providing use-based insurance services. Most insurance companies believe that modernizing their core systems is critical to differentiate their services in a crowded market, and machine learning is part of those efforts.
- **Automotive:** With the introduction of electric and autonomous vehicles, predictive maintenance models, and a slew of other disruptive trends in recent years, the automotive industry has experienced enormous change and upheaval. Of course, AI/ML plays a significant role in this transformation. It is, for example, an essential component of the BMW Group's automated vehicle initiatives.
- **Financial Services:** Similarly, financial services are using AI/ML to modernize and improve their offerings, such as personalizing customer service, improving risk analysis, and better detecting fraud and money laundering. As the amount of data that financial institutions must deal with grows, machine learning capabilities are expected to improve fraud detection models and help optimize bank service processing.

2.1.4 What is data? And it's types

Data is a collection of information,particularly facts like words, figures, measurements, and observations that are gathered for analysis, consideration, and use as a decision-supporting tool. There are primarily two ways to represent data:

1. **Qualitative:** Data that approximates and characterizes is defined as qualitative data. It is possible to observe and record qualitative data. This is a non-numerical data type. This type of data is gathered through observation,one-on-one interviews,focus groups, and other similar methods. In statistics, categorical data and qualitative data are two different terms, which is data that can be organized categorically based on the attributes and properties of a thing or a phenomenon.

(a) **Nominal:**Nominal data is data that has been "labeled" or "named" and can be divided into distinct groupings that don't cross over. Data in this case is neither measured nor assessed; it is simply assigned to multiple groups. These groups are distinct and share no characteristics.The order of the data collected cannot be established using nominal data,and thus changing the order of data has no effect. Gender, race, and relationship status are all examples of nominal data.

(b) **Ordinal:** While still belonging to the same class of values,these values have a natural ordering. When considering a clothing brand's size, it is simple to arrange them according to their name tag in the following order: small, medium, big. A+ is unquestionably better than a B grade in the grading system used to assess candidates on an exam, which is an ordinal data type.

2. **Quantitative:** Quantitative data is defined as data in the form of counts or numbers,for each data set having a distinct numerical value. Questions like "How many?" and "How long?" can be answered using quantitative data because of the ease with which mathematical derivations are provided by quantitative data,measuring various parameters becomes controllable. Surveys, polls, or questionnaires are typically used to collect quantitative data for statistical analysis. The two subcategories which describes them are :

   (a) **Discrete:** Discrete data consists of discrete variables that are non-negative,finite,numerical, and countable.Using straightforward statistical techniques like bar charts,line charts,or pie charts,discrete data can be simply displayed and demonstrated.The

   distribution of discrete data is discrete in both time and space. Analyzing discrete values is easier and more realistic with discrete distributions.The number of products in a supermarket,the number of regions in a country, and the number of books in a library are all examples of discrete data.

(b) **Continuous:** Continuous data evolves over time and can have varying values at various time intervals.Continuous data is made up of random variables that can be whole numbers or not. Data analysis methods such as line graphs,skews,and so on are used to measure continuous data. Continuous data examples include height,weight,length,time,temperature,age and so on.

2.1.5 What is training data?

Models that employ machine learning are trained, tested, and validated using data. In supervised learning, training data is enriched(labeled, tagged, or annotated)to highlight data features that are used to instruct that computer how to recognize the outcomes, or answers, that your model is intended to detect. Unsupervised ML models are trained using unlabeled data.

There are numerous methods to arrange training data. For sequential decision trees and similar algorithms, the input would be a collection of unclassified alphanumeric or text data.On the other hand, the training set for convolutional neural networks that deal with image processing and computer vision is frequently made up of a lot of images.The concept is that the machine learning program, which is quite intelligent and complex, utilizes iterative training on each of those images in order to eventually be able to identify characteristics, shapes, and even objects like humans or animals.The process cannot function without the training data, which might be described as the "food" the system consumes to function.

One of the oldest and most popular mantras in data science is "garbage in, garbage out." It still holds true despite the exponential growth in the frequency of data generation.The secret is to provide machine learning algorithms with relevant, high-quality data. As a result, model's accuracy can be greatly improved. The development of unbiased machine learning applications also requires high-quality training data.Machine learning models that are accurate rely on high-quality training data.

2.1.6 Characteristic of training data

The four main characteristics of high-quality training data are as follows:

1. **Relevant:** Data must be applicable to the current task.For example, if you're training your model to predict who will win the best employee award, you don't need data from grocery store purchases. Instead, you require relevant information about employee details.

2. **Representative:** The data must precisely represent the entire community.Most datasets are not absolute representations,but they must contain relevant attributes for proper model predictions. For example, if the model is to recognize facial images, it must be fed a diverse set of data containing people's faces of various ethnicity's. This will reduce the issue of AI bias, and the model will not be biased against a specific race, gender, or age group.

3. **Comprehensive:** The dataset should represent the majority of the model's use cases. The training data must have enough examples that'll allow the model to learn appropriately. It must contain real-world data samples as it will help train the model to understand what to expect.

4. **Uniform:** All data should be of the same form and origin. Additional information cannot be included in one part of the data.As a consequence, the training data will be inaccurate. To summarize, uniformity is an essential part of high-quality training data.

2.1.7 Importance of training data

To work with an ML algorithm, you must provide certain inputs that allow your model to understand things in its own unique way. Training data is the only source of input to your algorithms to assist your AI model to learn useful information from the data and make critical decisions.

Understanding the significance of the training set in ML will assist in obtaining the appropriate quality and quantity of training data for the

training of the model. Once you understand why it is essential and how it affects model prediction, you will select the appropriate algorithm based on the availability and compatibility of your training data set. As a result, when working with AI and ML models, prioritizing training data will undoubtedly assist you in acquiring the highest quality data sets to achieve the best results.

2.2 What is bias?

2.2.1 Introduction to data bias

While AI holds remarkable promise, the comfort of computerized classification and discovery within massive datasets can come with sizable downsides to folks and society through the amplification of current biases[17]. Bias is a aspect of human thinking process, and records amassed from people consequently inherently displays that bias. This makes it exceedingly hard to accumulate and modify information so that it omits bias whilst maintaining its accuracy-especially in view that the dedication of what bias is frequently subjective.When an end-user is introduced with data on-line that stigmatizes them primarily based on race,age,or gender or would not precisely pick out their identification it reasons harm.When humans experience that they are now not being pretty judged when making use of for jobs or loans it can limit public have faith in AI technology.

Most AI systems are driven by data and require huge data to be trained on. Thus, data is tightly coupled to the performance of these algorithms and systems. In the instances where the training data contains the bias,the algorithms educated on them will study these biases and reflect them into their predictions. As a result biases can have an effect on the algorithms using the data, producing biased outcomes.

AI is neither developed nor applied in a world isolated from societal realities like prejudice or unfair treatment.The concept of AI as a socio-technical systems emphasizes that the development of technology involves more than only its mathematical and computational blocks.

The values and behavior modeled from the datasets, the people that interact with them, and the intricate organizational aspects that go into their commission, design, development, and ultimate deployment are all considered in a socio-technical approach to AI.

2.2.2 Different types of data bias

There are numerous distinct types of bias, some of which might results in unfairness in various learning tasks. In this section we discuss about various different types of bias which can influence the performance of our AI model.Moreover,once we are familiar with the various biases, we are better able to discuss about various mitigation strategies. In this section, we discuss data biases that, when taken into account by ML training algorithms, may lead to biased results.

1. **Measurement Bias:** Measurement, or reporting, bias arises from how we choose,utilize, and measure particular features [18]. This may occur if we are using proxy variables and their quality differs from group to group. An illustration of this type of bias may be found in the recidivism risk prediction tool COMPAS, where prior arrests and arrests of friends and family members were employed as proxies to gauge the degree of "riskiness" or "crime"—-which can be seen as mismeasured proxies on their own. This is partially because minority populations experience more frequent policing and control, which results in greater arrest rates. People from minority groups are more likely to be arrested, but this does not mean that they are necessarily more dangerous because there are differences in how these groups are evaluated and managed [18].

2. **Omitted variable Bias:** When one or more significant variables are excluded from the model, omitted variable bias results [19] [20] [21]. An illustration of this situation would be if a model was created to predict, with a fair amount of accuracy, the annual percentage rate at which customers would discontinue using a service, however it was soon realized that the majority of users were discontinuing their subscriptions without the model's intended

forewarning. Imagine that a new, formidable competitor has entered the market and is offering the same solution for half the price, leading to the cancellation of memberships. The competitor's appearance was something for which the model was not prepared, so it is regarded as an omitted variable.

3. **Representation Bias:** When gathering data, how we sample from a population results in representation bias [18]. Non-representative samples lack the diversity of the population, with missing subgroups and other anomalies. Lack of geographical diversity in datasets like ImageNet results in a demonstrable bias towards Western cultures.

4. **Aggregation Bias:** When incorrect inferences are made about specific individuals based on studying the entire community, this is known as aggregate bias(or ecological fallacy) [22]. Clinical assistance tools are a prime illustration of this kind of prejudice. Think about diabetic patients with different morbidities depending on their gender and racial background. Particularly, the intricate differences across genders and ethnicities can be seen in HbA1c readings, which are frequently used to diagnose and monitor diabetes. A model that does not account for individual characteristics would therefore probably not be suitable for all racial and gender groupings in the population [18]. Even when they are evenly represented in the training data, this is still true. Aggregation bias may emerge from any generalizations about the population's subgroups [22].

5. **Sampling Bias:** Sample bias, which is related to representation bias,results from the non-random selection of subgroups for sampling [22].Because of sample bias, it's possible that the trends identified for one demographic won't apply to information gathered from a different population.

6. **Historical Bias:** Even with flawless sample and feature selection,historical bias, which is an existing prejudice and socio-technical problems in the world,can contaminate the data generating process [18]. An illustration of this type of bias can be seen in a 2018 image

search result when looking for female CEOs ultimately produced fewer female CEO photos because only 5% of Fortune 500 CEOs were female, skewing the search results in favor of male CEOs [18] .

7. **Population Bias:** Population bias occurs when the user population of the platform differs from the initial target population in terms of statistics,demographics,representatives,and end user attributes [23]. False data are produced through population bias. Different user demographics on various social media sites, such as women's higher usage of Pinterest, Facebook, and Instagram compared to men's higher activity on online forums like Reddit or Twitter, are examples of this type of prejudice. In accordance with gender, color, ethnicity, and parental educational background, additional examples and data on young adults' use of social media are available in [24]

8. **Self-selection Bias:** Self-selection is a form of sampling or selection bias in which study participants choose themselves. One instance of this kind of bias is seen in an opinion poll to measure enthusiasm for a political candidate, where the most enthusiastic supporters are more likely to complete the poll [22].

9. **Social Bias:** When the behavior of others influences our judgment,social bias occurs [25]. When we intend to give something a low rating or review, but are persuaded by other high ratings to change it, we may be guilty of this form of bias since we may feel that we are being too harsh [25] [26].

10. **Behavioral Bias:** Different user behavior across platforms, situations, or datasets is what causes behavioral bias [23]. An illustration of this type of prejudice is shown in [27], where the authors discuss how variations in emoji representation across platforms can cause people to respond and behave in various ways, sometimes even resulting in communication problems.

11. **Temporal Bias:** Due to variations in populations and behaviors across time, there is a temporal bias [23]. On Twitter, for instance,

you can see how people chatting about a certain issue will use a hashtag to draw attention before moving on to talk about an event without using the hashtag [23].

12. **Content Production Bias:** Content production bias results from disparities in the structure, lexicon, semantics, and syntax of user-generated contents [23]. One instance of this kind of bias is seen in [28] in where the differences in the use of language across different gender and age groups is discussed. The differences in the use of language use can be observed within and among different nations and groups.[22].

2.2.3 Use cases of data bias

The development of a software system by Amazon[29] to evaluate resumes of potential employees gathered from the internet is a well-known illustration of this problem. The initiative, which was initiated in 2014 to leverage word patterns taken from CVs from the preceding ten years to predict successful future employees,was discontinued in 2017 due to the systematic devaluation of female professionals. Since men make up the bulk of employees in the technology sector, the difficulty stemmed from the fact that training data was primarily composed of them.

Another study found that Facebook job advertisements[30] were significantly biased toward gender and ethnic groups, resulting in unequal job opportunities and persistent discriminatory treatment for the duration of the advertisement.People are denied opportunities because of their personal characteristics, which contradicts the statements in Article 21 of the EU Charter of Human Rights[31]. As a result, the Department of Housing and Urban Development sued Facebook in March 2019 for violating the Fair Housing Act due to the discriminatory actions of its advertisements,as housing ads were disproportionately targeted based on race,gender„and other personal traits[32].

Similarly, a scientific experiment on the search engine Common Crawl[33] revealed discriminatory treatment due to gender imbalance in the input data(almost 4000.000 biographies): authors compared three

machine learning approaches for occupational classification and demonstrated that in each case, the rate of correct classification accompanied that existing gender biases of the occupational groups, even without explicitly using gender indicators.

The investigation into COMPASS(Correctional Offender Management Profiling for Alternative Sanctions), a formula that judges use to evaluate the likelihood of recidivism of defendants,is the most famous case in the criminal justice system for data bias. The non-profit organization Pro Publica discovered that the algorithm was biased in favor of white defendants[34]: in fact, those were arrested and convicted were nearly twice as likely as black defendants who were not rearrested were nearly twice as likely as white defendants to be misclassified as higher risk(false positive). The main cause of thus distorted effect was the large number of records pertaining to white defendants.

In the medical field, a recent study[35] reported on the case of widely used commercial system for determining which patients should be admitted to an intensive care unit. Medical professionals risk assessments developed by a machine learning algorithm using previous data on medical spending and health-care utilization. In cases of comparable health status, white patients were found to be significantly more highly probable than black patients to be appointed to the intensive care unit. In this case, the system was also affected by ethnic discrimination, as the risk score reflected the expected cost of treatment rather compared to the real health issues, the former being strongly associated with the patients economic wealth.

The examples given above explicitly demonstrate how bias in data can propagate and be reflected in Model output. This is becoming a socio-technical issue, especially in the public sector, where high-stakes decision are increasingly delegated to such systems.

2.2.4 Consequences of data bias

Concerns about data bias are growing. DataRobot surveyed 350 technology leaders in the United States and the United Kingdom,including CIOs, IT managers, IT directors, data scientists, and development

leaders, among others, who use or plan to use AI. According to the survey results, 54% of technology leaders are very or highly concerned about AI bias. It is 12% higher than in 2019.(42% shared such a sentiment). Simultaneously, an overwhelming majority(81%) calls for more AI regulation. As a result, the primary concerns about bias in AI are the loss of customer trust, reputation, and exposure to detailed compliance checks.

For example, as per the Consumer Federation of America, women will be charged more for car insurance in Oregon. The average annual premium for basic coverage is $976.05 for women and $ 876.20 for men. Everything else being equal, the gender gap is $ 100, or 11.4%. AI bias can easily lead to a loss of consumer trust and , as an outcome, revenue. Every customer who is furious about the injustice may be able to find a more equitable insurance company.

The true impact of data bias strikes hard at their companies' most valuable assets - specialists. A bias like this could harm your bottom line and your employer's branding strategy. Biased data creates biased ML models which directly impact your company's reputation. For example, a study of the racial bias in pulse oximetry measurement discovered that black patients had nearly three times the occurrence of occult hypoxemia that was not detected by pulse oximetry as white patients. This may endanger the lives of some patients. Bias jeopardizes your healthcare entity's reputation, loss of future patients, revenue, and legal issues.

AI bias tends to result in a vicious cycle. It's because all of the negative effects of bias are intertwined. To maintain the company's good reputation, we must ensure that bias can be mitigated. This, in turn, will increase customer trust by promoting equality.

2.2.5 How to prevent bias?

Machine learning bias can be avoided by education and sound governance; after it has been identified, an organization can use best practices to address it, such as the ones listed below.

- Choose training data that is sufficiently large and representative

to mitigate biases that commonly affect machine learning, such as sample bias and prejudice bias.

- To ensure that bias resulting from algorithms or data sets does not appear in the results of machine learning systems, test and validate them.

- Keep an eye on machine learning systems as they function to ensure that biases don't gradually seep in as the algorithms learn more and more.

- Examine and inspect models using additional tools, such as IBM's AI Fairness 360 Open Source Toolkit or Google's What-if Tool.

- Maintain consistency throughout your organization.Empowering marketers to use data is terrific for organizational velocity, but don't discount the insight that data scientists can offer. Schedule regular interdepartmental meetings to ensure everyone has the resources they require and that your data is being generated and processed as efficiently as possible.

2.3 What is fairness?

2.3.1 Understanding "fairness"

ML fairness [36]is a recently established area of machine learning that studies how to ensure that biases in the data and model inaccuracies do not lead to models that treat individuals unfavorably on the basis of characteristics such as e.g., race, gender, disabilities, and sexual or political orientation. Fairness is defined differently across disciplines[37].

- **Law:** Fairness in the law entails protecting individuals and groups from discrimination or mistreatment, focusing on prohibiting behaviors and biases, and making decisions based on certain protected factors.

- **Quantitative Fields:** Quantitative fields (math, computer science, statistics, and economics):fairness issues are viewed as mathematical problems. For a specific task or problem, fairness usually corresponds to some sort of criterion, such as equal or equitable allocation, representation, or error rates.

- **Social Science:** Fairness is frequently considered in social science in light of social relationships.Members of certain groups (or identities) are more likely to benefit.

- **Philosophy:** Philosophically, ideas of fairness are based on the assumption that what is fair is also morally correct.Fairness is linked to concepts of justice and equity.

Definitions can differ even within disciplines. It's no surprise, then, that fairness in machine learning systems lacks a standardized definition.

2.3.2 Previous related work

In recent years, remarkable research has been conducted to ensure that model outcomes are free of bias. The primary research has concentrated on techniques for detecting and mitigating systematic discrimination based on various definitions of unfairness.Among the most comprehensive works we remind the book by Barocas et al. [38] (from which we derived the unfairness measures used here), the survey on bias and fairness in machine learning by Mehrabi et al. [22] as well as the review of discrimination measures for algorithm decision making by Zliobaite [39]. One significant limitation of determining whether a software output is fair or not is the formal impossibility of satisfying multiple mathematical notations of fairness at the same time [40] [41].This is an ontological limitation: there can be no universally acceptable notion of fairness because to define a "fair impact," several political, economic, and cultural factors must be considered [42].The ACM Conference on Fairness, Accountability, and Transparency[43] 5 has recognized this issue and has been designed and promoted not only for computer scientists working in the area, but also for scholars and practitioners

from the law, social sciences, and humanities to investigate and tackle issues in this emerging area."

Finally, it is important to mention tool development: in recent years, researchers both in the public and private sectors have created toolkits for bias detection and mitigation[44].As an example:

- IBM's AI Fairness 360 Open Source Toolkit [1] is an open-source library designed to examine and mitigate bias in machine learning model output. It includes several metrics for analyzing model unfairness as well as pre-processing algorithms for transforming the dataset.

- The LinkedIn Fairness Toolkit (LiFT) [2] is a Scala/Spark library that allows for fairness measurement and bias mitigation in large-scale machine learning workflows. Measuring biases in training data, evaluating fairness metrics for ML models, and detecting statistically significant differences in their performance across different subgroups are all part of the measurement module. It is also useful for ad hoc fairness analysis. A post-processing method for transforming model scores to ensure so-called equality of opportunity for rankings (in the presence/absence of position bias) is included in the mitigation section. This method can be applied directly to model-generated scores without requiring any changes to the existing model training pipeline.

- We introduce Fairlearn [12], a set of open-source tools that enable developers and data scientists to assess and improve their AI systems' fairness. An interactive visualization dashboard and unfairness mitigation algorithms make up Fairlearn's two halves. These parts are made to make balancing the trade-offs between fairness and model performance easier. We underline that it is a socio-technical problem to prioritize justice in AI systems. It is impossible to completely "debias" a system or to guarantee fairness due to the numerous complicated sources of unfairness—some societal and some technical—so the objective is to reduce unfairness-related harms as much as feasible.

- The What-If Tool [45], by Google, which can be used to analyze the characteristics of a dataset and of the models derived from it. These models can also be examined for their unfairness w.r.t. various measures, and an interactive graphical user interface let the user perform a sensitivity analysis by moving classification thresholds for the selected features.

2.3.3 What is fairness assessment?

The word "fairness assessment" primarily refers to determining how accurately your dataset can reflect all of the populations. Thus, it doesn't exhibit any bias against individuals, organizations, or people of any certain gender, race, or other category. In this study, we use a variety of statistical formulae to determine whether any dataset features are contributing to any bias or unfairness.We will speak specifically about tabular datasets and supervised learning since the work was based on these parameters. The steps involved in fairness assessment:

- Looking for possible biases: We must first examine the dataset and make an effort to comprehend numerous unintentional biases, including exclusion bias, observer bias, measurement bias, recall bias, and other biases. When evaluating fairness, all of these biases should be taken into consideration.

- Choose the protected attribute: We need to carefully select the protected attributes. The protected attributes are usually considered as the sensitive information associated with individuals or groups. A few examples of protected attributes are sex, age, marital status, race, and others. The choice of the protected attributes needs to be done in order to understand if it's imbalanced or treated not evenly.

- Establishing fairness parameters: We must employ a variety of metrics that offer the mathematical definition of fairness.The choice of metric is specific to the protected attribute chosen. A few categories in which this metrics can be classified are :

- Balance measure metrics
- Equality measure metrics
- Distance measure metrics
- Individual fairness metrics
- Group fairness metrics

We will be able to assess whether there is bias in the data by using the aforementioned three stages in some scenarios, which is the first step toward fair ML.

2.3.4 Importance of fair dataset

The key lesson here is that we must consider data variety. We need to stop believing that gathering a ton of raw data will be sufficient to advance our cause. We must be extremely careful while collecting datasets from scratch.

For example, if we are creating a model that determines whether or not a person should be given a loan, we need to ensure that we can train our model with a variety of data. This requires that our model must be universal and not be skewed by certain characteristics, such as being white, married, and others. If we utilize solely white, privileged, and married people to train our model, it will be prejudiced against black and single people, which will encourage racial prejudice and mistrust among users. Therefore, we must make our dataset as fair as feasible.

A fair dataset will foster user trust as it will fairly represents the society, will produces good results, and accurately captures the situation you are seeking to study.

2.4 Fairness assessment using synthetic data

2.4.1 What is synthetic data?

Synthetic data is generated by using AI algorithms instead of collecting data from real-world cases. It incorporates all the statistical and distribution properties of the original dataset. The use of synthetic data can improve AI and solve various data-related properties. ML engineers typically require a large amount of data that must be properly cleaned and labeled; this data cleaning and collection operation is quite costly. Furthermore, diverse training data will be able to demonstrate more accurate AI models.

2.4.2 Advantages of synthetic data

The following are the primary benefits of using synthetic data:

- **Data Protection and Privacy Preservation:** Rather than masking or anonymizing the original data, synthetic data can be used to protect data privacy. It prevents the disclosure of sensitive user information.

- **Easy Labeling:** Labeling is simple with fully synthetic data. For example, if a picture of a classroom is generated, it is simple to assign labels to the backboard, people, table, and chairs automatically. We don't need to hire people to manually label these objects.

- **Data Augmentation:** Synthetically generated data reduces data scarcity and improves the generalization of ML models. It resolves various issues such as an imbalanced dataset,missing values and remove duplicate records.

- **Promotes AI Fairness:** Data synthetization can be used to correct data bias by ensuring data diversity to represent the real world.

2.4.3 Types of synthetic data

- **Tabular data:** Tabular data contains far more sensitive and private information than other types. For these reasons, it must not only be anonymized, but also synthesized. Anonymization of data entails removing characteristics from a data set that can be used to identify a person. To anonymize the data, some identifying points in the set, such as name, address, and gender, should be removed. The more data we delete, the less valuable the information for future analysis becomes. Even deleting a large amount of a person's private data makes it possible to identify someone using the limited information available.

  Tabular data synthesis has many applications, including fraud detection and economic forecasting in finance, medical applications, and marketing and advertising campaigns for customer behavior and reactions.

- **Time series data:** Time series synthetic data is similar to tabular data in some ways, but the main difference is that time series is focused on data that is time-related. Anyone can use autoregressive models (AR) to generate it with models because they specialize in time series data.

  Most frequent applications of time series synthetic data are in the fields of financial predictions, demand forecasting, trade, market predictions, transaction recording, nature forecasts, and component monitoring in machines and robotics

- **Image Data:** Image data that has been synthesized can be used for a variety of applications, the most well-known of which are computer vision and face generation. A machine learns how and where to understand what it perceives during the computer vision process in order to perform a specific action. It is widely valued in the robotics and automotive industries. Both require a computer to distinguish between objects and backgrounds, as well as the distances and sizes among them.

Face generation is the process of creating human faces from scratch. Produced human faces can be used to train machine learning models to identify human faces for security or robotics applications.

- **Text Data:** Textual data can be used to train chatbots, algorithms that check email boxes for spam, and machine learning models that detect abuse. The text generation model GPT-3 is used to generate synthesised text. It is an abbreviation for Generative Pre-trained Transformer 3. GPT-3 is an autoregressive model that generates human-like language written in text that can be used to train machine learning models for text recognition or understanding.

2.4.4 Use cases of synthetic data

- **Fraud Prevention:** Synthetic data was used for the training of AI fraud prevention models by American Express. Since they didn't have enough data on fraudulent cases, they used GANs to synthesize enough data. The goal was to create a balanced dataset with different fraud variations for a better AI model for different scenarios.

- **Face Analysis:** Fake it until you make it: facial analysis in the world using synthetic data alone [46]. Researchers created a variety of human 3D faces, including labeling, to use as training material computer vision, landmark localization, and face parsing machine learning models. The project's findings demonstrated that synthetic data could accurately match real data.

- **Chatbot Development:** Moveworks, a startup, developed a chatbot trained on synthetic data to answer customers' questions about HR, finance, and, most notably, IT. It used synthetic data to train on cases where real data was scarce. Customers can use the chatbot to help with tasks such as password reset, software installation, and device connection.

- **Dialog Prepossessing:** Amazon trained Alexa to recognize requests in multiple languages using synthetic data. When a new

language is added to the system, the data pool of requests for the machine learning model is greatly diminished. In this case, Amazon is using synthetic data alongside real data to enrich their sample set and train Alexa's natural-language understanding (NLU) models.

- **Self Driving Vehicles:** Waymo trains self-driving cars with synthetic data. Waymo has developed its own deep recurrent neural network (RNN) called ChaufferNet. In this environment, a vehicle is trained on labeled synthetic data as well as real data to drive safely while recognizing objects and following traffic rules.

2.4.5 Can synthetic data address data bias?

A solution to data bias is synthetic data. The quality of the raw real data has an effect on the quality of the raw synthetic data. The potential of synthetic data lies in the ability to manage the output, which enables the creation of a more balanced, pure, and valuable synthetic dataset. Real datasets do not offer this amount of control, in contrast to synthetic data.A strong synthetic data generator must also have the intelligence to recognize faults in the real data and offer solutions.

By supplementing your existing data with things you haven't seen, synthetic data can help reduce bias. Synthetic data can help fill in the data gaps that result from data bias, such as when there is not enough data, it is too expensive, or there is no consent for usage in ML projects.

Additionally, synthetic data can assist in balancing out a dataset that is out of balance, such as one where the sample is predominately made up of members of a particular social group.

For ML models, having enough high-quality data is crucial. Your team may not always be confident of the data they will need to train the model at the outset of a project. Synthetic data can shed light on the type of data used to build the model. And by being transparent, machine learning algorithms are less likely to develop prejudice.

2.4.6 Limitation of synthetic data

While using synthetic data has many advantages, there are times when it is better not to. Data synthesis is a faster and less expensive process than data collection, but it is a complex procedure and requires an experienced persons. Data that has been incorrectly synthesized may not accurately represent events in the real world or may still contain bias.

In addition, depending on the goal, collecting real data can be more profitable or useful. For example, sociological research that gathers primarily differences in opinions about new events is far more reliable and valid than machine-generated data.

While there is still a significant dependence on human-annotated and real-world data, synthetic data is still frequently used since it is simple to produce, inexpensive, and very valuable in specific situations. Testing a model's performance using well-understood, human-annotated validation data is the only method to ensure that it is producing accurate, realistic outputs. Real-world human-annotated data continues to be a crucial component of machine learning training data, even though creating realistic synthetic data has been easier over time.

2.5 Fairness assessment using various bias mitigation techniques

2.5.1 What is bias mitigation?

Bias mitigation is the method we'll employ to try to reduce the bias brought on by numerous outside influences. We may lessen bias with the use of bias mitigation techniques, which is beneficial for creating a model that accurately depicts society without any forms of racism or sexism.They can be used either individually or in groups.

In most cases, we address group-level bias reducing measures where we observe that the privileged group is not always favored. Instead, we must ensure that the privileged group and the underprivileged group is treated equally. In this study, we have applied pre-processing bias

mitigation strategies, which are bias reduction procedures at the dataset level.

2.5.2 What are the different stages of bias mitigation?

We need to use various bias mitigation algorithms to reduce data bias. Bias mitigation algorithms can be used at various stages of the ML pipeline. It is broadly divided into three stages:

- **Pre-processing bias mitigation:** Pre-processing bias mitigation begins with training data, which is used in the first phase of the AI development process and frequently introduces underlying bias.The analysis of model performance train on this data may result in various types of biased treatment, such as a specific gender (male) receiving more income than females, or a specific race receiving more government benefits. We must treat individuals fairly regardless of their gender, race, or age. The manner in which data is used to train the learner shapes the results. We must now ensure that our data is free of bias, or it will have a negative impact and will be unable to represent the diverse population.

- **In-processing bias mitigation:** When training a machine learning model, in-processing algorithms provide unique opportunities for increasing fairness and reducing bias. For example, before approving a loan, a bank may attempt to calculate a customer's "ability to repay." Based on sensitive variables such as race, gender, or proxy variables that may correlate, the AI system may predict someone's ability. For in-processing, we can implement algorithms such as adversarial debiasing and prejudice remover.

- **Post-processing bias mitigation:** Post-processing mitigation is useful after the model has been trained but the user wants to reduce bias in predictions. Equalized or calibrated equalized odds are two among many other algorithms that can be used. While mitigating bias after model predictions, the model's accuracy may

suffer. As a result, we must strike a balance between bias reduction and model accuracy.

2.5.3 Do mitigating bias in training datasets helpful?

As we know that the training dataset is a crucial component of any model's learning process, it is important to ensure that the data used to train the model is as bias-free as possible. There are various pre-processing bias mitigation techniques which can be applied to the dataset like Reweighing [47], Disparate Impact Remover [48], Optimized Preprocessing [49] and others.

If we apply these strategies are training dataset will be more ideal for making a better model. Experimental analysis carried out by [1] have chosen a better results with the above mentioned algorithms.

Therefore, we can state that applying some pre-processing bias mitigation measures to the training dataset will be beneficial in later stages of the machine learning pipeline.

Chapter 3

Methodology

3.1 Datasets

We look at five datasets from two different application domains: financial services - credit, payment, loan risk and social topics – personal earnings and education. We chose a few protected attributes from each of these datasets and will now examine how bias measure metrics perform on them. Furthermore, we will test the performance of some bias mitigation algorithms on the above-mentioned datasets. The datasets chosen are described in more detail below.

3.1.1 UCI Adult Income

Barry Becker extracted these data from the 1994 Census database; the prediction task is to determine whether a person makes more than $50,000 per year based on that set of reasonably clean records, also known as the "Census Income" dataset [50]. Thus, income represents the target variable, which can take one of two values: = $ 50,000 or > $ 50,000.

- features: "age","work_class","education", "marital_status", "occupation", "relationship","race", "sex","capital_gain","capital_loss", "hours_per_week", "native_country"
- target: "income"

- protected attributes: "sex", "race","income" "marital_status",

The adult income dataset can be used to reflect the real-world skewness in terms of the percentage of the minority class earning more than $50,000, which is 24%. As a result, the dataset is imbalanced. The large number of rows representing the majority class can lead to bias.

3.1.2 German Credit Card

The German professor Hans Hofmann provided this widely used German credit dataset from the UCI Machine Learning Repository [51] as part of a collection of datasets from a European project called "Statlog." When a bank obtains a loan application, it must decide whether or not to proceed with the loan approval based on the applicant's profile. Each entry in this dataset represents a person who obtains credit from a bank. According to the set of attributes, each person is classified as having good or bad credit risk.

- features:"month","credit_amount","sex","status","housing", "investment_as_income_percentage", "residence_since", "age", "number_of_credits", "people_liable_for", "credit_history", "purpose","savings", "employment", "other_debtors", "property", "installment_plans", "housing", "skill_level", "telephone", "foreign_worker"
- target: credit
- protected attribute: "sex" ,"age",

The data are a stratified sample of 1000 credits (700 good and 300 bad) collected between 1973 and 1975 from a large regional bank in southern Germany with about 500 urban and rural branches.

3.1.3 Default Credit Card

From April 2005 to September 2005, this dataset contains information on default payments, demographic factors, credit data, payment history,

and bill statements for credit card clients in Taiwan. It was obtained from the UCI Machine Learning Repository [52]. The dataset contains 25 variables, but only a subset of them are used as predictors, in keeping with the base paper.

- features: "limit_bal","sex", "education", "marriage" ,"age" , "pay", "bill_amt", "pay_amt"
- target: "default.payment.next.month"
- protected attribute: "education", "marriage"

The credit card company has collected data on 30000 customers. The data was collected with the goal to figure out how the issuer decides who gets a credit card based on the various attributes.

3.1.4 Lending Club

Lending Club is the biggest online loan marketplace, offering personal loans, business loans, and medical procedure financing. Borrowers can easily obtain lower interest rate loans by using a quick online interface. These datasets contain complete loan data for all loans issued between 2007 and 2011, including the current loan status (Current, Charged-off, Fully Paid, etc.) and most recent payment information.

- features:"loan_amount","payments_term", "monthly_payment", "grade", "working_years", "home", "annual_income", "verification", "purpose", "debt_to_income", "inquiries", "open_credit_lines", "derogatory_records", "revolving_balance", "revolving_rate", "total_accounts", "bankruptcies", "fico_average"
- target: "loan_risk"
- protected attribute: "home" , "purpose",

There are total of 31150 records present in this dataset.

3.1.5 Student-Por

This data [53] examines secondary school student achievement in two Portuguese schools. The data attributes include student grades, demographics, and social and school-related features), and it was gathered through the use of school reports and questionnaires.Two datasets are provided for performance in two distinct subjects: mathematics (mat) and Portuguese language (por). We used the student-por dataset specifically in our study.

- features:"sex", "age", "address","famsize", "Pstatus","Medu", "Fedu", "Mjob", "Fjob", "reason","guardian","traveltime","studytime", "failures", "schoolsup", "famsup", "paid", "activities","nursery","higher", "internet", "romantic", "famrel", "freetime", "goout", "Dalc", "Walc" , "health", "absenses", "G1", "G2"
- target: G3
- protected attribute: "sex", "mother's job", "father's job"

There are total of 650 records in the dataset.

3.2 Hardware

As for the purpose of this thesis a local setup has been used, outlined in Table 3.1, this was allowed by the tabular format of the data and the lightness of the models in use.

Hardware Configuration		
CPU	GPU	RAM
11þGen Intel(R)Core(TM)i5-1135G7	intel(R) Iris(R) Xe Graphics	8GB 3200MHz

Table 3.1: Local Configuration

3.3 Libraries

Different open source libraries and frameworks have been integrated into the thesis. Besides the most popular and general purpose packages, the most important ones as for the development of this work are:

- Lift [2] , used to design the bias measure metrics.
- Aif360 [1], utilized to create the bias measuring metrics and the bias mitigation pre-processing techniques.
- Synthesized SDK, utilized to create the bias mitigation pre-processing techniques.

3.4 Bias measure metrics

Bias measure metrics are used to determine whether a specific attribute in a dataset is biased. In this study, we will attempt to comprehend three distinct bias factors: balance, inequality, and distance.We use the following metrics to assess balance: Gini, Simpson, Shannon, and Imbalance Ratio.Similarly, for inequality, we employ the following metrics:Generalized entropy index, Theil index, Atkinson index, and Coefficient of variations and lastly for distance measures we use the metrics: Infinity Norm Distance and Total Variation of Distance.

3.4.1 Balance measure metrics

- **Gini index:** It is a measure of heterogeneity [54] that is used in many disciplines and is frequently discussed under different names: examples include political polarization, market competition, ecological diversity, and racial discrimination.Heterogeneity refers to the number of different types (such as protected groups) that are represented. The heterogeneity of a discrete random variable with m categories in statistics with frequency f_i (with i=1,....,m) can vary between a degenerate case(=minimum value of heterogeneity) and an equiprobable case(= maximum value of heterogeneity, since

categories are all equally represented). This means that for a given m, the heterogeneity increases if probabilities become as equal as possible, i.e. the different protected groups have similar representations.

The Gini index is computed as follows:

$$G = \frac{m}{m-1}\left(1 - \sum_{i=1}^{m} f_i^2\right) \tag{3.1}$$

Where we added the multiplication factor $\frac{m}{m-1}$ in order to normalize the index between 0 and 1.

- **Simpson index:** Another indicator of diversity is the Simpson index, which calculates the likelihood that two individuals randomly selected from a sample belong to the same species (i.e., the same class or category). It is used in social and economic sciences to measure wealth, uniformity, and equity, as well as in ecology to measure the diversity of living beings in a given location.Suppose we consider a discrete random variable which assumes m categories with frequency f_i where i=1,...,m (that is, the proportion f_i of the species i with respect to the total number of species):

$$D = \frac{1}{m-1}\left(\frac{1}{\sum_{i=1}^{m} f_i^2 - 1}\right) \tag{3.2}$$

- **Shannon index:** Diversity indices are a useful tool for measuring imbalance because they provide information about community composition as well as the relative amounts of different species (classes). The Shannon index, which is a measure of species diversity in a community, is a widely used concept in biology, phylogenetics, and ecology. The index was calculated as follows:

$$S = -\left(\frac{1}{\ln m}\right)\sum_{i=1}^{m} f_i \ln f_i \tag{3.3}$$

In order to normalize the index we divide by $\ln m$. In addition since $\ln 0 = -\infty$, to deal with empty classes -i.e when f_i =0 we resort to the notable limit:

$$\lim_{x \to \infty} x \ln x = 0$$

- **Imbalance ratio:** The Imbalance Ratio (IR) is a widely used measure that calculates the ratio of the highest and lowest frequency. We use the inverse to normalize it in the range [0,1] and convert it to a balance measure:

$$IR = \frac{min(\{f_1..m\})}{max(\{f1..m\})} \tag{3.4}$$

3.4.2 Equality measure metrics

- **Generalized Entropy Index:** The generalized entropy index assesses inequality across a population [55].It derives from information theory as a measure of data redundancy.A measure of redundancy in information theory can be interpreted as non-randomness or data compression; thus, this interpretation also applies to this index.

$$GE = \epsilon(\alpha) = \begin{cases} \frac{1}{n\alpha(\alpha-1)} \sum_{i=1}^{n} [(\frac{b_i}{\mu})^{\alpha} - 1] & \text{for } \alpha \neq 0, 1, \\ \frac{1}{n} \sum_{i=1}^{n} \frac{b_i}{\mu} \ln \frac{b_i}{\mu} & \text{for } \alpha = 1, \\ -\frac{1}{n} \sum_{i=1}^{n} \ln \frac{b_i}{\mu} & \text{for } \alpha = 0. \end{cases} \tag{3.5}$$

  - Parameters: b (array-like) – The parameter used to compute the entropy index.
  - alpha (scalar) – A parameter that governs how much weight is given to distances between values in different parts of the distribution.In the case of the generalized entropy index, alpha is equal to 2.

- **Theil Index:** The Theil index is a statistic that has been used to measure racial segregation as well as economic inequality. It is a subset of the generalized entropy index with α .It can be interpreted as a measure of redundancy, lack of diversity, isolation, segregation, inequality, non-randomness, and compressibility.

$$TI = \epsilon(\alpha) = \begin{cases} \frac{1}{n\alpha(\alpha-1)} \sum_{i=1}^{n} [(\frac{b_i}{\mu})^{\alpha} - 1] & \text{for } \alpha \neq 0, 1, \\ \frac{1}{n} \sum_{i=1}^{n} \frac{b_i}{\mu} \ln \frac{b_i}{\mu} & \text{for } \alpha = 1, \\ -\frac{1}{n} \sum_{i=1}^{n} \ln \frac{b_i}{\mu} & \text{for } \alpha = 0. \end{cases} \tag{3.6}$$

– Parameters: b (array-like) – The parameter used to compute the entropy index.In the case of the Theil index, alpha is equal to 1.

- **Atkinsons Index:** A derivative of the Generalized Entropy Index under the restriction that $\epsilon = 1 - \alpha$ .Specifically, an Atkinson index with high inequality aversion is derived from a GE index with low α.

$$AI = COV = \epsilon(\alpha) = \begin{cases} [\epsilon(\epsilon-1)GE]^{\left(\frac{1}{(1-\epsilon)}\right)} & \text{for } \epsilon \neq 1 \\ 1 - e^{-GE} & \text{for } \epsilon = 1 \end{cases} \tag{3.7}$$

- **Coefficient of variation:** A generalized entropy index derivative. It calculates the standard deviation divided by the benefit vector's mean.It is two times of generalized entropy index.

$$COV = \epsilon(\alpha) = \begin{cases} \frac{1}{n\alpha(\alpha-1)} \sum_{i=1}^{n} [(\frac{b_i}{\mu})^{\alpha} - 1] & \text{for } \alpha \neq 0, 1, \\ \frac{1}{n} \sum_{i=1}^{n} \frac{b_i}{\mu} \ln \frac{b_i}{\mu} & \text{for } \alpha = 1, \\ -\frac{1}{n} \sum_{i=1}^{n} \ln \frac{b_i}{\mu} & \text{for } \alpha = 0. \end{cases} \tag{3.8}$$

– Parameters: b (array-like) – The parameter used to compute the entropy index.In the case of the COV, alpha is equal to 2.

3.4.3 Distance measure metrics

- **Infinity Norm Distance:** Calculates the Chebyshev distance between the observed and reference distributions. It is equal to the maximum difference between the two distributions.The Infinity Norm Distance:

$$D_Chebyshev(x, y) := \max_i(|x_i - |y_i) \tag{3.9}$$

– Parameters: observed distribution (x_i) – The observed distribution is the distribution of the protected attribute chosen.

– reference distribution (y_i) - The reference distribution is the target distribution.

- **Total Variation Distance:** Calculates the Total Variation Distance between the observed and reference distributions. It is half the L1 distance between the two distributions. The Total Variation Distance:

$$d_L(\bar{x}\bar{y}) = polynom_abs(\bar{x}) = \frac{1}{2}\sum_{i=1}^{I}|x_i - y_i| \tag{3.10}$$

  - Parameters: observed distribution (x_i) – The observed distribution is the distribution of the protected attribute chosen.
  - reference distribution (y_i) - The reference distribution is the target distribution.

3.5 Explanatory study design

3.5.1 Measures details

The following are a few of the selected measures:

- In this study we work both with categorical and numerical attribute.
- For categorical attribute we use label encoding because to convert the labels into numerical form.
- For numerical attribute we use Kbin Discretizer because to map numerical variables onto discrete values. Values for the variable are grouped together into discrete bins and each bin is assigned a unique integer such that the ordinal relationship between the bins is preserved.
- Range in the interval [0, 1]. Results close to 1 are considered better in all cases

3.5.2 Details of the bias measure metrics

The bias measures metrics are designed using references from AI Fairness 360 [1], also known as AIF360, and The LinkedIn Fairness Toolkit (LiFT)

[2], also known as Lift. There are ten metrics in total, divided into three groups: distance, equality, and balance. Each of these measures is capable of computing both continuous and categorical protected attributes.

The full dataframe can be passed in, and all of the dataset's attributes will have their bias measures metrics generated. Additionally, we have the ability to send either one protected attribute or multiple protected attributes. As an illustration, we can pass age to our Gini function alone or pass age and sex to our Gini function combined. The Gini value will be calculated in both scenarios.

While we employed KBins discretizer for the continuous property, label encoding was used for the categorical attribute. The process of label encoding only requires that each value in a column be changed from a letter to a number. Consider the attribute "race," which has three categories: European, Asian, and American. The column will be modified in the event of label encoding, for example, European: 1, Asian: 2, and American: 3.

With the intention of transforming the numerical variable into a discrete distribution of probabilities where each numerical value is provided a label and the labels have an ordered relationship, the KBins discretizer is used for attributes like age. The properties of KBins discretizer chosen are: n_bins=10, encode="ordinal", strategy="uniform"

Here are some images that illustrate them:

3.5.3 Vendors of synthetic data generation

There are several vendors that offer the generation of synthetic data with privacy guarantees, which means that mechanisms in the synthetic data are designed to prevent the re-identification of an individual from the original data.

Betterdata, Datomize, Gretel, MostlyAI, Generatrix, and others are a few examples. While some vendors provide a solution tailored to a specific industry, such as finance, others allow you to generate synthetic data for a single dataset in their free version.

Considering all of these external factors, I decided on Syndate, Gretel,

```
df = pd.read_csv(path_to_data)
target_attribute = "income"
m = M.Metric(df)
# To pass the value individually of few attribute
gini = m.gini(attributes=['sex'])
```

Figure 3.1: Passing Categorical attribute:'sex'

```
df = pd.read_csv(path_to_data)
target_attribute = "income"
m = M.Metric(df)
# To pass the value individually of few attribute
gini = m.gini(attributes=['age'])
```

Figure 3.2: Passing Continuous attribute:'age'

```
df = pd.read_csv(path_to_data)
target_attribute = "income"
m = M.Metric(df)
#To pass the entire dataframe
gini = m.gini()
```

Figure 3.3: Passing df: Entire Dataframe

```
df = pd.read_csv(path_to_data)
target_attribute = "income"
m = M.Metric(df)
# To pass the value individually of few attribute
gini = m.gini(attributes=['work_class'])
```

Figure 3.4: Passing One attribute: "work_class"

```
df = pd.read_csv(path_to_data)
target_attribute = "income"
m = M.Metric(df)
# To pass the value individually of few attribute
gini = m.gini(attributes=['age']['sex'])
```

Figure 3.5: Passing Two attribute: "age","sex"

and MostlyAI for a few reasons:

- They are free and easily accessible from their website or GitHub.
- More than one dataset can be used in their free version.
- There is a possibility to customize the number of rows.
- Each of these companies tries to generate ethical data i.e. take into consideration the bias problem.

Furthermore, because each of these vendors works with structured data is best suited to our needs.

3.5.4 Syndate

- **About Syndate:** Syndata AB has developed "Syndapp," which they refer to as a Synthetic engine.Syndapp can generate large amounts of data that match the statistical attributes of real data but are completely synthetic by using Machine Learning, and Artificial Intelligence. This means that the data can be used for a variety of purposes, including predictive modeling, analytics, software testing, bias mitigation, and more.

  The output is simply a simulated representation of real-world data, but without exposing private information and with the same utility for innovation.As a result, the generated synthetic data can be used by a variety of organizations and business sectors. The entire procedure and experiment can be found in [56].

- **Steps to take in order to generate synthetic data for this study** Following are the steps to generate synthetic data:

  – Upload the dataset and configure the attributes you want in your synthetic data. In this study, we kept the attribute in the synthetic dataset is same as it was in the real dataset.

  – Then choose train and generate, which gives you two options: random and realistic.The random model will synthesize by

randomly sampling from each attribute's distribution, whereas the realistic model will synthesize using a CTGAN model.

– In this study, we use the realistic approach and specify the number of epochs and rows in our dataset.

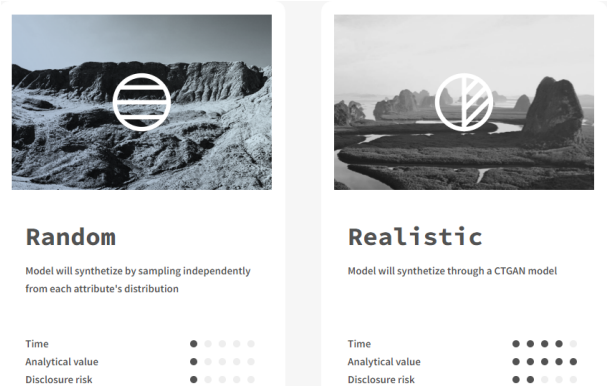

Figure 3.6: Two approaches used by Syndapp to generate synthetic data.

3.5.5 MostlyAI

- **About MostlyAI:** MostlyAI is an AI-powered big data analytics platform for understanding client behavior. It enables users to track and manage large amounts of data from various sources. It investigates and provides insights into client behaviors, as well as develops predictive standards, by utilizing artificial intelligence technologies , machine learning, and deep learning . It operates a synthetic data engine that allows users to simulate data, develop marks, and comprehend data deviations using deep neural networks.
- **How MostlyAI contributes to ethical AI** Synthetic data that has been bias-corrected can effectively address fairness issues. Using synthetization, one can correct embedded injustices within the data, right at the heart of the problem.When one can synthesize fair versions of your data, the end result is accurate, privacy-safe, and, most importantly, fair synthetic data. For training machine learning models, fair synthetic data outperforms real data.

MostlyAI generates bias-corrected synthetic data that can address both privacy and fairness concerns, allowing for the use and democratization of big data assets while minimizing risks. They add fairness constraints, such as statistical parity, to obtain fair data.

For example, suppose the Adult Income dataset is trained. In that case, we penalize statistical parity violations by a number proportional to the difference between the split of women and men in the high-income segment. The model parameters are then modified to minimize both the accuracy loss and the fairness constraint.Using this method, they could eliminate gender income inequality from the synthetic version of the Adult Income data set.

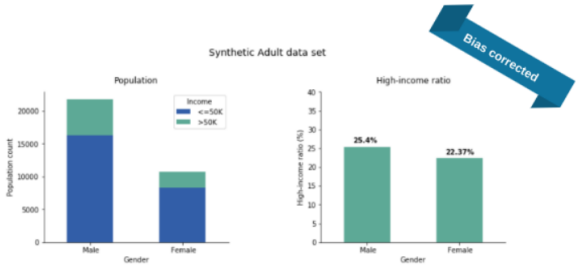

Figure 3.7: MostlyAI is creating a bias-free synthetic dataset in which the proportion of high-income individuals is equal across genders.

In general, adding the fairness constraint broadens the company's software's objective from generating accurate and private synthetic data to generating accurate, private, and fair synthetic data.

- **Steps to take in order to generate synthetic data** Following are the steps to be taken:
  - **Upload your CSV files:** Select Ad hoc jobs that can synthesize subject tables and subject table-linked table datasets. A subject table is a CSV file in which each row contains profile information about the entities whose privacy you want to protect, such as names, email addresses, birthdates, and so on. In addition to the subject table, a linked table with sequential

data is included in a subject table-linked table dataset. You can use this feature to combine historical activities, customer journeys, and transactional records. MOSTLYAI can produce precise, privacy-protected synthetic copies of all of them.

- **Configure the synthetization procedure:** MostylAI is fully automated. It analyzes the first 1000 rows of the tables and recommends the best job configuration.You can optimize the job for accuracy or speed by adjusting various synthetization parameters such as encoding types and rare category protection.
- **Train and generate:** Our AI model learns data patterns, statistical distributions, correlations, and time dependencies. MOSTLYAI will use this model to create a synthetic version of the uploaded data.

3.5.6 Gretel

- **About Gretel:** Gretel.ai was founded with the goal of empowering developers to unlock innovation through safe, efficient collaboration with sensitive data. Gretel was the first to offer Privacy Engineering as a Service and created a synthetic data tool suite based on their open-sourced AI-based core. These tools facilitate and accelerate faster to generate data that protects privacy and can be safely shared.

  Gretel created simple, accessible APIs for developers to generate high-quality synthetic data, and transform and anonymize data - tools that quickly remove privacy-related bottlenecks and accelerate business innovation for organizations in life sciences, financial services, technology, healthcare, gaming, and other industries.

- **Gretel reducing AI bias with synthetic data** A key use case at Gretel is the generation of bias-free synthetic data. The steps taken by Gretel are mentioned as :

  - To get started, log into the Gretel Console with a GitHub or Google account, and create a new project.

- Select "From Blueprint" and then "Automatically balance your data" from the "Recommended" section.
- Following that, a Gretel Project will be created with sample data.
- After you've uploaded your data, go to the "Transform" tab. Copy the Connection URI for the Project from the "Integration" menu in the top right.
- Start the Notebook "Automatically balance your data."

For example, to begin, we must select the protected attribute from which we want to remove bias. Following the selection of your fields,

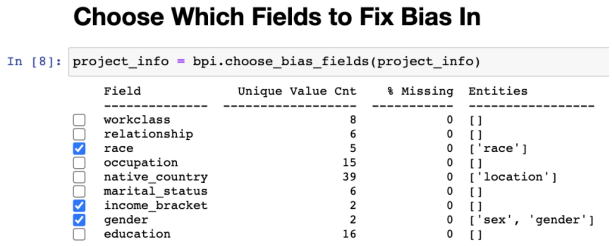

Figure 3.8: Gretel's example uses Race, Gender, and Income bracket as protected attributes to eliminate bias.

the blueprint helps to guide you through the protected attribute cells to train your synthetic data model. Following synthetic generation, we seed the model with the classes that require boosting in order to generate additional records. The new data will be unbiased, with protected attributes balanced out. The blueprint concludes with the new synthetic data being saved to a CSV file.

- **Steps to take in order to generate synthetic data for this study** Following are the steps to be taken:
  - Navigate to the dashboard and choose the file. There are two file formats that can be accepted: JSON and CSV.
  - After selecting your file, create a project and give it a name.

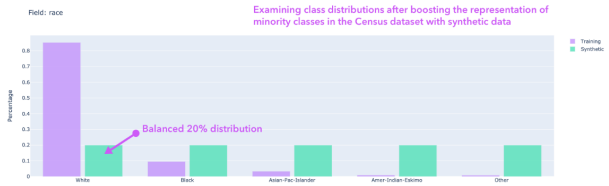

Figure 3.9: The minority class is balanced out after increasing the representation of the race-protected attribute in the adult income dataset.

- Select the option to generate synthetic data after clicking on the option to create a model.
- This will generate 5000 rows of synthetic data.
- Once the sample synthetic data is generated we have the option to decide the number of rows and create a new synthetic data.

3.5.7 Using open source libraries for mitigating bias

Both open source methodologies and open source technology have enormous potential to aid in the battle against bias.Open source software dominates modern artificial intelligence, from TensorFlow to IBM Watson to packages like scikit-learn. Since the open source community has already proven to be extremely successful in creating robust and extensively tested machine-learning tools, it seems reasonable that the same community could also effectively build anti-bias tests.

The open source community should design tools for detecting data bias, and those techniques should be applied to the wide range of accessible training data sets.The open source methodology is also well-suited to creating processes to combat bias. Making software conversations open, democratized, and aligned with social good is critical to countering a problem caused in part by reverse conversations, private software development, and undemocratized decision-making. Fighting bias should become easier if corporations, online communities, and

academics embrace these open source communities when approaching machine learning.

Finally, we should all work together to create and strengthen an open source community centered on ethical AI. Whether it's making a contribution to software tools, stress-testing machine learning models, or combing through gigabytes of training data, it's time to harness the power of open-source methodology to combat one of our digital age's most serious threats. . Numerous studies have been conducted in the last few years, and various bias mitigating open source libraries have been released.Here, we discuss five such tools and frameworks that are widely used to detect and remove bias in AI and ML models.

- **FairML:** FairML is a framework for detecting bias in machine learning models. It works by determining the relative importance and significance of features used in the machine learning model to detect bias.It can search for data that is biased based on attributes such as sexual identity, ethnic background, religious practice, and others. It works by auditing predictive models and quantifying the relative importance of the model's input, which aids in determining the model's fairness [57].

- **Synthesized SDK:** The Synthesized SDK is an open-source library that generates high-quality, privacy-preserving datasets for machine learning and data science applications. They offer the following features: bootstrap datasets, rebalance, and impute missing values.To deal with data bias, the SDK attempts to upsample rare groups of data in order to detect any hidden biases.

- **Google's What-If Tool:** Google's interactive open-source tool allows users to visually investigate machine learning models. It is a component of the open-source TensorBoard and can analyze datasets as well as trained TensorFlow models. It explains how models work in different scenarios and generates rich visualizations to describe model performance. Its bias detection feature enables the user to manually process edit samples from a dataset and examines the impact of these changes using the associated model.

- **Microsoft's Fairlearn:** Microsoft's open-source toolkit enables AI researchers and data scientists to identify and improve the fairness of their AI systems. This tool, which consists of two components — an interactive visualization dashboard and a bias mitigation algorithm — significantly improves the fairness and model performance. According to the company, emphasizing fairness in AI systems is a social and technical challenge, This tool's primary goal is to mitigate as many fairness-related harms as possible [12].

- **IBM AI Fairness 360**: This open-source toolkit from IBM aids in the mitigation of bias from large datasets because it is centered on more than 70 fairness metrics and ten bias mitigation algorithms. These bias algorithms are used in sectors such as re-weighting and optimized preprocessing. These bias mitigation algorithms can be used by a developer to recognize fairness and compare it to the original model. It is an open-source toolkit for examining, reporting, and mitigating discriminatory treatment in ML models all throughout the AI application lifecycle [58].

In this study, we will primarily discuss two open source libraries: Synthesized SDK and IBM AI Fairness 360 because they address the issue of bias mitigation at the data level for ethical AI.

3.5.8 Synthesized SDK

- **About SDK:** For machine learning and data science use cases, the SDK generates high-quality, privacy-preserving datasets.After data is loaded from a data source, the process of data transferring by Synthesized can be divided into three steps:

  - **Annotate and preprocess data:** The software automatically recognizes data formats and types. This step it can deal with missing data and incorrect values.
  - **Create a data generative mathematical model:** The software creates a generative representation, which is a mathematical equation that describes how data properties should

appear. Internally, this equation enables the user to transform pure data noise into output data with the properties of original data.

- **Create a new dataset using the generative model:** Finally, once trained, the generative model can be used to generate new data samples on demand.Moreover, the software supports data manipulation, which is used to balance out some of the dataset's protected attributes so that the output data has the desirable characteristics.

In this study, we will use SDK to create three distinct versions of the dataset. Furthermore, each technique will be demonstrated in greater detail. The following sections will demonstrate the steps taken for this study.

- **Installation:** For installation, there are basically two steps required:
  - **Installing the package:** Synthesized can then be installed using pip by issuing the following command: "pip install synthesized".
  - **Setting the licence key:** After installing the package, you'll need a license key to use the software.The quickest way to see if the SDK is operational is to run the command: "synth-validate".
- **Data synthesis:** To synthesize the date, we must first perform the following steps:-
  - **Fetch the dataset:** We use the utility function "synthesized.util.get example data()" to retrieve the dataset, which returns a small dataframe.
  - **Extract metadata information:** Before we can create a Synthesizer object, we must first extract metadata about the data. This begins to look at the dataframe and attempts to conclude things like:
    * Is this a continuous or categorical column?

* What is the scope of this information?
* Is it a special kind? (date, address, etc..)

This information can then be used to instruct the Synthesizer on how to model the data.The primary method of extracting this information is using the "synthesized.MetaExtractor class".Particularly, the.extract(df) method.

- **Create Synthesizer model:**The following stage creates a blank generative model of the data that is ready for the learning process. One of the main generator objects in the SDK is the "HighDimSynthesizer."
- **Learn the original data:** After that, run the command "synthesizer.learn(df)" to learn the data.The "HighDimSynthesizer" will learn patterns in the data in order to generate them later.

  The "num iterations" argument in "synthesizer.learn" can be fixed to a specific value to limit the number of Synthesizer learning steps.This is especially useful for testing pipelines that contain the "HighDimSynthesizer" before attempting to Synthesize data properly.

  If a large value for "num iterations" is provided, the Synthesizer may decide to end training early irrespective, so rising training time is not possible in this way.Since the Synthesizer is designed to learn the dataset in a single call, this should not be required in most cases.
- **Synthesize new data:** Using the command "synthesizer.synthesize(num rows=1000)," the Synthesizer can be used to generate data finally. This will produce a dataframe with the specified number of rows.

```
import synthesized # import synthesized package

df = synthesized.util.get_example_data() ①
df_meta = synthesized.MetaExtractor.extract(df, annotations=...) ②
synthesizer = synthesized.HighDimSynthesizer(df_meta) ③
synthesizer.learn(df) ④
synthesizer.synthesize(num_rows=len(df)) ⑤
```

Figure 3.10: The Synthesized SDK's steps.

- **Different techniques for generating synthetic data**

  - **Bootstrapping:** In order to generate synthetic data by using the technique bootstrapping we simply need to follow the 5 steps as mentioned above. The code is :-

```
import pandas as pd
from synthesized import HighDimSynthesizer,
    MetaExtractor
     df1 = pd.read_csv(Pathname)
     df1_meta = MetaExtractor.extract(df=df1)

     synth1 = HighDimSynthesizer(df1_meta)
     synth1.learn(df_train=df1)

     df1_synth = synth1.synthesize(1000)
     df1_balanced = df1_synth
```

  - **Reshaping:** When developing a predictive model for imbalanced classification, a number of pitfalls can occur: some models are unsuitable, model precision suffers, and unwanted biases are propagated.

    To address these issues, the Synthesized SDK enables fast and accurate dataset rebalancing via conditional sampling of the generative model. The code is :

```
import pandas as pd
from synthesized import HighDimSynthesizer,
    MetaExtractor
from synthesized import ConditionalSampler

     df2 = pd.read_csv(Pathname)
     df2_meta = MetaExtractor.extract(df=df2)

     synth2 = HighDimSynthesizer(df2_meta)
     synth2.learn(df_train=df2)

     sampler = ConditionalSampler(synth2)
```

```
    df2_synth = synth2.synthesize(1000)
    df2_balanced = sampler.synthesize(num_rows=len(df2))
```

– **Data Bias:** SDK's most recent version introduced the concept of "Data Bias." SDK uses the fairlens library to upsample the protected attribute of the dataset. Moreover, it also tries to find any hidden biases which can be present.

```
import fairlens as fl
import pandas as pd
from synthesized import HighDimSynthesizer, MetaExtractor
from synthesized import ConditionalSampler

    df3 = pd.read_csv(Pathname)
    fs = fl.FairnessScorer(df3, 'RawScore')
    fs.demographic_report()
    fs.plot_distributions()

    df3_meta = MetaExtractor.extract(df=df3)
    synth2 = HighDimSynthesizer(df2_meta)
    synth2.learn(df_train=df3)
    sampler = ConditionalSampler(synth2)
    df3_synth = synth2.synthesize(1000)
    df3_balanced = sampler.synthesize(num_rows=len(df3))

    fs_balanced = fl.FairnessScorer(df3_balanced, 'RawScore')
    fs_balanced.demographic_report()
    fs_balanced.plot_distributions()
```

3.5.9 AI Fairness 360

The AI Fairness 360 toolkit developed by IBM is an open-source, extensible library that contains techniques developed by the researcher to detect and mitigate bias in ML models throughout the AI application

lifecycle. Both Python and R versions of the AI Fairness 360 package are available.

The AI Fairness 360 package includes the following components:

- A comprehensive set of metrics for testing the biases of datasets and models.
- These metrics' explanations,and
- Algorithms for reducing bias in datasets and models. It is to convert laboratory-based algorithmic research into real-world applications in fields as diverse as human capital management,finance, healthcare, and education.

This library attempts to mitigate bias at three different stages of the ML pipeline's lifecycle: Pre-processing, In-processing, and Post processing.

In this study , we will exclusively discuss the pre-processing techniques as we talk to mitigate bias at the data level.The algorithms which are used at preprocessing level are :

- **Reweighing:** To ensure fairness before classification, reweighing is a preprocessing technique that weights the examples in each (group, label) combination differently [47]. The class is represented below:

  Class: classaif360.algorithms.preprocessing.Reweighing(unprivileged_groups, privileged_groups)

  – unprivileged_groups (list(dict)) – Representation for unprivileged group.

  – privileged_groups (list(dict)) – Representation for privileged group.

- **Disparate Impact Remover:** Disparate impact remover is a data preprocessing technique that modifies feature values to improve group fairness while maintaining rank-ordering within groups [48]. The class is represented below:

  classaif360.algorithms.preprocessing.DisparateImpactRemover (repair_level=1.0,sensitive_attribute=")

  - repair_level (float) – Repair amount. 0.0 is no repair while 1.0 is full repair.
  - sensitive_attribute (str) – Single protected attribute with which to do repair.

- **LFR: Learning Fair Representation:** Learning fair representations is a data pre-processing technique that finds a latent representation that encodes the data well but conceals information about protected attributes [59]. The class is represented below:

  classaif360.algorithms.preprocessing.LFR(unprivileged_groups, privileged_groups, k=5, Ax=0.01, Ay=1.0, Az=50.0, print_interval=250, verbose=0, seed=None)

  - unprivileged_groups (tuple) – Representation for unprivileged group.
  - privileged_groups (tuple) – Representation for privileged group.
  - k (int, optional) – Number of prototypes.
  - Ax (float, optional) – Input recontruction quality term weight.
  - Az (float, optional) – Fairness constraint term weight.
  - Ay (float, optional) – Output prediction error.
  - print_interval (int, optional) – Print optimization objective value every print_interval iterations.
  - verbose (int, optional) – If zero, then no output.
  - seed (int, optional) – Seed to make predict repeatable.

- **Optimized Preprocessing:** Optimized preprocessing is a data preprocessing technique that learns a probabilistic transformation that edits the features and labels in the data while adhering to constraints and objectives such as individual distortion, group fairness, and data fidelity [60]. The class is represented below as :

  classaif360.algorithms.preprocessing.OptimPreproc(optimizer, optim_options,unprivileged_groups=None,privileged_groups=None, verbose=False, seed=None)

- optimizer (class) – Optimizer class.
- optim_options (dict) – Options for optimization to estimate the transformation.
- unprivileged_groups (dict) – Representation for unprivileged group.
- privileged_groups (dict) – Representation for privileged group.
- verbose (bool, optional) – Verbosity flag for optimization.
- seed (int, optional) – Seed to make fit and predict repeatable.

Each of the aforementioned bias mitigation methods will result in a fresh, debiased dataset. The unprivileged_groups is represented by the class which is less in number whereas the privileged_groups is represented by the class which is more in number.In the case of the class "sex" attribute, for instance, the privileged group might be male and the unprivileged group might be female with regard to the target attribute.For example, if we are considering the class "sex" attribute in that case the privileged group may be male whereas the unprivileged group is female w.r.t to the target attribute.

Chapter 4

Analysis

We shall proceed in this section by examining each dataset separately. The performance of the bias measure metrics will be examined first on the original dataset, then on a synthetic dataset, and finally on a dataset created utilizing pre-processing bias mitigation techniques.

We'll go forward with our analysis by beginning with UCI Adult Income and moving on to all the other remaining datasets. Different attribute cardinalities are considered to better evaluate each index in different contexts. Different forms of bias measure metrics will be used, depending on the protected properties selected. In particular m = 2,5,7 are reported, with m number of attribute's categories for the balance measure metrics. Different target references will be selected for the distance measure metrics depending on the dataset.

4.1 UCI Adult Income

The first dataset under analysis is "UCI Adult Income" the protected attributes chosen are: "sex", "race", "income", "martial_status", "age". For the UCI Adult Income dataset we will study all the three stages : Original, Synthetic and Debiased datasets. The metrics used are:

- sex: balance measure metrics;
- race: equality measure metrics;

- income: equality measure metrics;
- martial_status: distance measure metrics;
- age: equality measure metrics

Protected Attribute: "SEX"

We begin with the "sex" attribute, which has a cardinality m = 2 i.e. male and female. The distribution of the attribute "sex" is represented below.

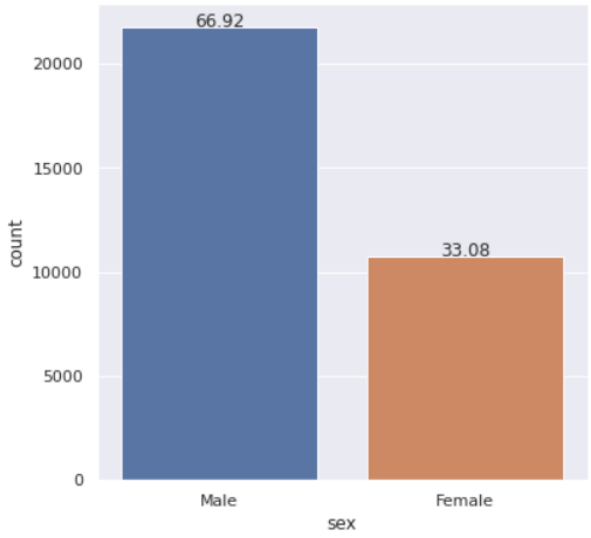

Figure 4.1: dataset: 'Adult Income', attribute: 'sex'

According to the following diagram, gender bias could result from an uneven representation of male and female demographics. So we decided to investigate the effectiveness of the balance measures metrics as a result.

At first we will assess the value of the balance measure metrics on the original dataset:

Balance Measure Metrics	
Metrics	Value
Gini	0.8854
Shannon	0.9157
Simpson	0.7944
Imbalance	0.4943

Table 4.1: dataset:'Original Adult Income', attribute: 'sex'

Shannon Index displays a number that is the closest to 1 of the four balance measure criteria.

Now, using synthetic data produced by three different vendors, we assess the performance of the balance measure metrics on the sex attribute.

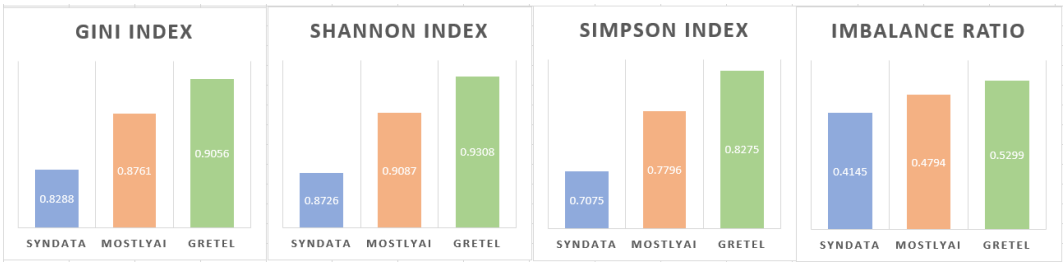

Figure 4.2: dataset: 'Synthetic Adult Income', attribute: 'sex'

We can observe that Gretel is performing better than the other synthetic data generator vendors in every instance. In this situation, we might assert that Gretel is faithful to its claim in order to provide synthetic data in a morally upright manner.

The effectiveness of the balance measure metrics on the dataset that was debiased by SDK utilizing pre-processing bias mitigation

techniques for the "sex" attribute Similarly here we see the value of the bias measure metrics produced on the aif360 dataset using pre-processing bias mitigation techniques for the "sex" attribute:

Balance Measure Metrics			
Metrics	Bootstrapping	Reshapping	Data Bias
Gini	0.8656	0.7971	0.8648
Shannon	0.9008	0.8482	0.9002
Simpson	0.7631	0.6627	0.7619
Imbalance Ratio	0.4635	0.3789	0.4624

Table 4.2: dataset: 'Debiased Adult Income', attribute: 'sex'

The effectiveness of the balance measure metrics on the dataset that was debiased by aif360 utilizing pre-processing bias mitigation techniques for the "sex" attribute.

Balance Measure Metrics			
Metrics	Disparate Impact Remover	Optimized Preprocessing	Learning Fair Representation
Gini	0.9376	0.8860	0.8856
Shannon	0.8465	0.9161	0.9158
Simpson	0.6005	0.7953	0.7946
Imbalance Ratio	0.0109	0.4951	0.4945

Table 4.3: dataset: 'Debiased Adult Income', attribute: 'sex'

We can see that the aif360 performs well with two of its algorithms, "Learning Fair Representation" and "Optimized Preprocessing," among the pre-processing bias mitigation solutions.

Protected Attribute: "RACE"

Let's proceed with the 'race' attribute and use our equality measure metrics. Description about the distribution :

- There are 5 uniques categories in the race attribute.
- Most of them is "white" which is roughly 85%
- This dataset is totally bias towards the "white" race.
- Second major race is black which is around 10%

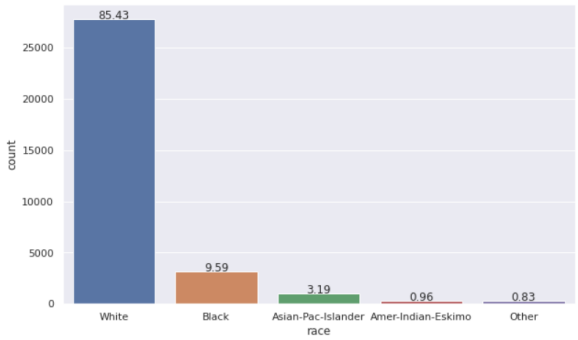

Figure 4.3: dataset:'Adult Income', attribute:'race'

Metrics for equality measures are chosen since, in most cases, they were created particularly for attributes like race or income. These measures are useful in determining which end of the distribution contributed most to the observed inequality. It is useful to recognize the dominant categories.

The representation of the equality measure metrics on the original dataset is shown in the table below.

Equality Measure Metrics	
Metrics	Value
GEI	0.9998
Atkinson	0.9946
Theil	0.9966
COV	0.9946

Table 4.4: dataset:'Original Adult Income', attribute: 'race'

From the above value of the equality measure we can see that there is no randomness as one race is dominant.

Following we have the value of the equality measure metrics on the synthetic datasets:

Equality Measure Metrics			
Metrics	Syndata	MostlyAI	Gretel
GEI	1.0000	0.9997	0.9996
Atkinson	1.0000	0.9941	0.9895
Theil	1.0000	0.9968	0.9912
COV	1.0000	0.9941	0.9895

Table 4.5: dataset: 'Synthetic Adult Income', attribute: 'race'.

We can see that the synthetic data does not significantly outperform the real dataset for the race attribute.

Similarly, we apply two pre-processing bias reduction strategies from the aif360 package, namely "Learning Fair Representation" and "Optimized Preprocessing," that are well matched along with "Data Bias" techniques from SDK.

Equality Measure Metrics			
Metrics	Learning Fair Representation	Optimized Preprocessing	Data Bias
GEI	0.0000	0.0000	0.0043
Atkinson	0.0000	0.0000	0.0310
Theil	0.0000	0.0000	0.01137
COV	0.0000	0.0000	0.03102

Table 4.6: dataset: " Debiased Adult Income", attribute: 'race'.

Race: In the case of race, a value near to 1 reflects higher inequality since it indicates that one race is dominant in the dataset, meaning that there is no diversity and little randomization. In the bias mitigation strategies we can see the value is 0 , it means it is able to remove inequality in races specially the strategies LFR and Optimized Preprocessing.

Protected Attribute: 'Income'

Since the "income attribute" is used to quantify income disparity, we employ equality measure metrics that are similar to it.

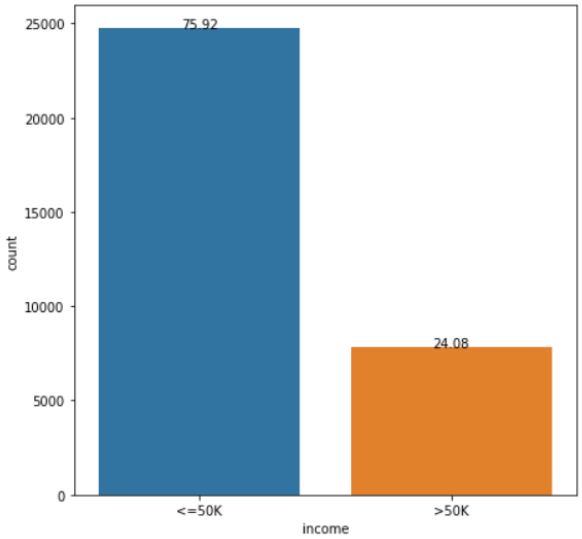

Figure 4.4: dataset:'Adult Income', attribute:'income'

Description about the distribution :

- 25% of the population belong to income>50k
- 75% of the population belong to income <50k

The equality measure metrics from the original dataset are depicted in the table below:

Equality Measure Metrics	
Metrics	Value
GEI	0.9792
Atkinson	0.8672
Theil	0.7158
COV	0.8672

Table 4.7: dataset:'Original Adult Income', attribute: "income"

The value of the equality measure metrics on the synthetic datasets:

Equality Measure Metrics			
Metrics	Syndata	MostlyAI	Gretel
GEI	0.0969	0.9705	0.9881
Atkinson	0.0505	0.8418	0.9014
Theil	0.0141	0.6985	0.7852
COV	0.0505	0.8418	0.9014

Table 4.8: dataset: 'Synthetic Adult Income', attribute: 'income'.

The value of the equality measure metrics on the debiased dataset using the pre-processing bias mitigation strategies.

Equality Measure Metrics			
Metrics	Learning Fair Representation	Optimized Preprocessing	Data Bias
GEI	0.1000	0.03807	0.3279
Atkinson	0.1000	0.1601	0.5479
Theil	0.1000	0.2952	0.5968
COV	0.1000	0.1601	0.5479

Table 4.9: dataset: "Debiased Adult Income", attribute: 'income'.

Income: A value near to 1 in the case of income indicates greater equality because it shows that a greater proportion of the dataset's participants have the same income.

Protected Attribute: "martial_status"

Now, using the target attribute, namely income, we check the marital status.

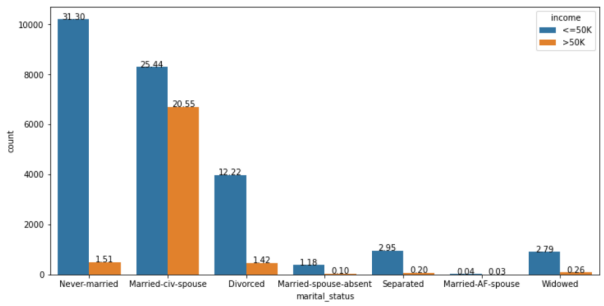

Figure 4.5: dataset: 'Original Adult Income', attribute:'marital_status'

Description of the distribution:

- There are 7 distinct groups for this marital-status attribute.
- Two of them—Never-married (33%) and Married-Civil Spouse (45.82%)—dominate other groups.

- There are a maximum number of samples for married-civ-spouse.
- There is a minimum number of samples for married-AF-spouse.

We check the value of the distance measure metrics as they are calculated by taking reference the target attribute. The below table represent the value taken by evaluating the original dataset.

Distance Measure Metrics	
Metrics	Value
Total Variation Distance	0.2479
Infinity Norm Distance	0.9998

Table 4.10: dataset" 'Adult Income', reference:' marital_status', target:'income'

Below is the representation of the distance measure metrics on the synthetic dataset:

Distance Measure Metrics			
Metrics	Syndate	MostlyAI	Gretel
Total Variation Distance	0.0229	0.1965	0.2818
Infinity Norm Distance	0.9992	0.9998	0.9999

Table 4.11: dataset: "Synthetic Adult Income", attribute: 'martial_status'.

The value of distance measure metrics produced on the SDK dataset using bias mitigation techniques for the "martial_status" attribute.

Distance Measure Metrics			
Metrics	Bootstrapping	Reshapping	Data Bias
Total Variation Distance	0.2105	0.4441	0.2035
Infinity Norm Distance	0.9996	0.9999	0.9999

Table 4.12: dataset: "Debiased Adult Income", attribute: 'martial_status'.

Protected Attribute: 'age'

We then examine the continuous attribute "age."

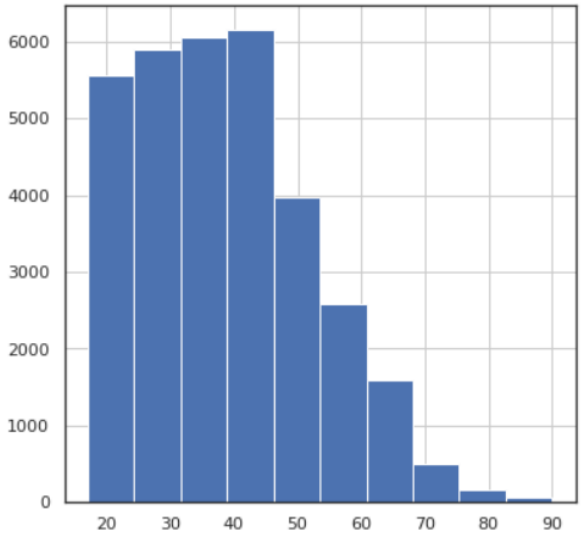

Figure 4.6: dataset:'Adult Income', attribute:'age'

The graphic above demonstrates that:

- The "age" property is not balanced.
- It has a right skew (But this is totally fine as younger adult earn wages not the older ones)
- The persons range in age from 17 to 90, respectively.
- After a certain age, or 70 years, there are fewer observations of people's ages in this dataset (868).

The table below is the measure of the equality measure metrics on the original dataset.

Equality Measure Metrics	
Metrics	Value
GEI	0.9943
Atkinson	0.9364
Theil	0.8965
COV	0.9364

Table 4.13: dataset:'Original Adult Income', attribute: "age"

In fact, Gini Index in this context shows a better result followed by Shannon and Simpson.

The value of the equality measure metrics on the synthetic datasets:

Equality Measure Metrics			
Metrics	Syndate	MostlyAI	Gretel
GEI	0.9658	0.9918	0.9954
Atkinson	0.8287	0.9238	0.9443
Theil	0.7865	0.8895	0.9098
COV	0.8287	0.9238	0.9443

Table 4.14: dataset: "Synthetic Adult Income", attribute: 'age'.

The Optimized Preprocessing and Learning Fair Represenation, which can be seen in the image below, splits the age range into ranges from 0 to 70 because that is the range with the highest income.

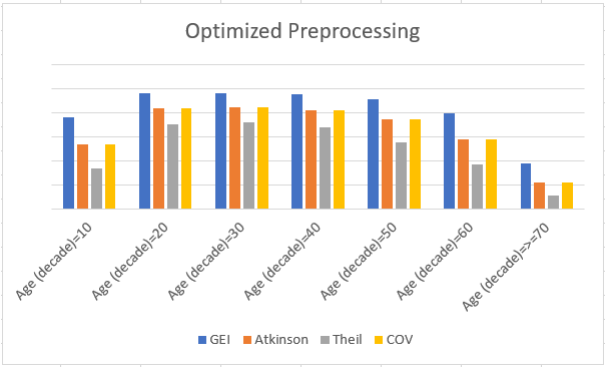

Figure 4.7: dataset:'Debiased Adult Income', attribute:'age'

Equality Measure Metrics				
Metrics	GEI	Atkinson	Theil	COV
Age (decade)=10	0.7638	0.5383	0.3344	0.5383
Age (decade)=20	0.9614	0.8385	0.7024	0.8385
Age (decade)=30	0.9648	0.8476	0.7181	0.8476
Age (decade)=40	0.9553	0.8233	0.6770	0.8233
Age (decade)=50	0.9168	0.7440	0.5588	0.7440
Age (decade)=60	0.7998	0.5785	0.3714	0.5785
Age (decade)=>=70	0.3814	0.2237	0.1129	0.2237

Table 4.15: dataset: "Debiased Adult Income", attribute: 'age'.

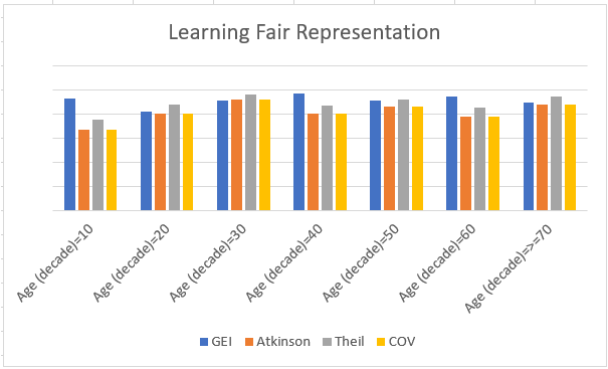

Figure 4.8: dataset:'Debiased Adult Income', attribute:'age'

Equality Measure Metrics				
Metrics	GEI	Atkinson	Theil	COV
Age (decade)=10	0.9276	0.6714	0.7534	0.6714
Age (decade)=20	0.8214	0.8385	0.8800	0.8073
Age (decade)=30	0.9131	0.8476	0.9614	0.9241
Age (decade)=40	0.9718	0.8233	0.8759	0.8036
Age (decade)=50	0.9110	0.7440	0.9190	0.8644
Age (decade)=60	0.9441	0.5785	0.8559	0.7807
Age (decade)=>=70	0.8971	0.2237	0.9460	0.8813

Table 4.16: dataset: "Debiased Adult Income", attribute: 'age'.

From the above value from the synthetic datasets and the debiased dataset, we can observe that comparable age groups are present more in the dataset which suggests less variety.

4.2 German Credit Card

The next dataset under analysis is the "German Credit Card". The protected attribute chosen are : sex and age. For the German Credit Card dataset we will study two stages : Original and Debiased datasets.

At first we see the distribution of the dataset w.r.t to the target attribute.

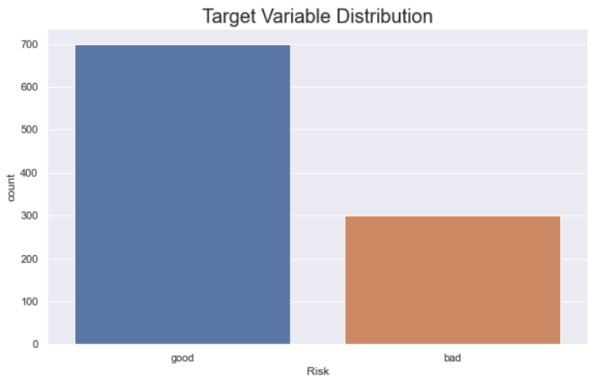

Figure 4.9: Representation of the target distribution

The details about the distribution:

- There have been 700 applicants that have been classified as good applicants
- There have been 300 applications that have been classified as bad applicants

The protected attributes chosen are: "sex", "age". The metrics used are:

- sex: balance measure metrics;
- age: balance measure metrics;

Protected Attribute: "Sex"

We now turn to the "sex" attribute and attempt to comprehend how it is distributed throughout the dataset.

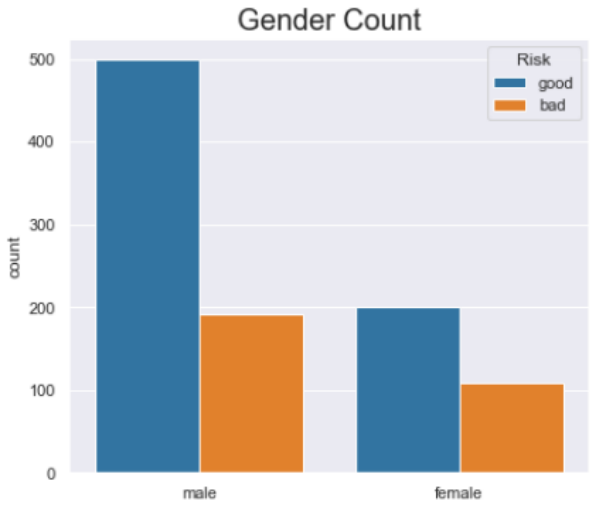

Figure 4.10: Representation of the sex distribution

We can observe that there are twice as many males as female applicants. About two-fifths of male applicants and one-third of female applicants are deemed bad. Most prominently, we can see that the

dataset is sex-biased. Therefore, we decide to study the metrics for the balance measure.

Balance Measure Metrics	
Metrics	Value
Gini	0.8556
Shannon	0.8931
Simpson	0.7476
Imbalance Ratio	0.4492

Table 4.17: dataset:'Original German Credit Card', attribute: "sex"

We can see that the Gini and the Shannon Index is avaliable to the value most close to 1.

Now, we use a variety of bias mitigation strategies to assess the effectiveness of the balance measure metrics. The pre-processing bias mitigation strategies from the aif360 library on the 'sex' attribute

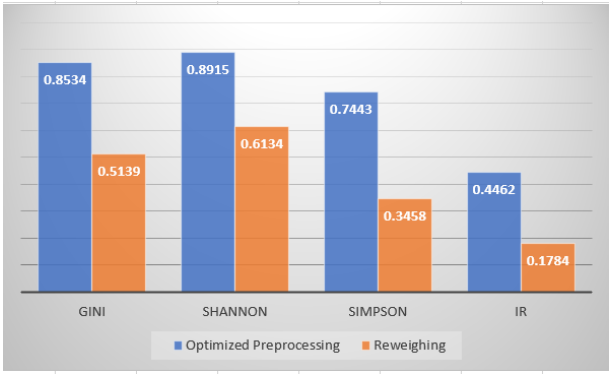

Figure 4.11: dataset:'Debiased German Credit Card', attribute: "sex"

We can see that "Optimized Preprocessing" pre-processing bias mitigation improves the performance of our balance measure metrics specifically for the "Gini" and "Shannon" Index.

Protected Attribute: 'Age'

The next protected attribute chosen from this dataset is "age" and the value of m=7

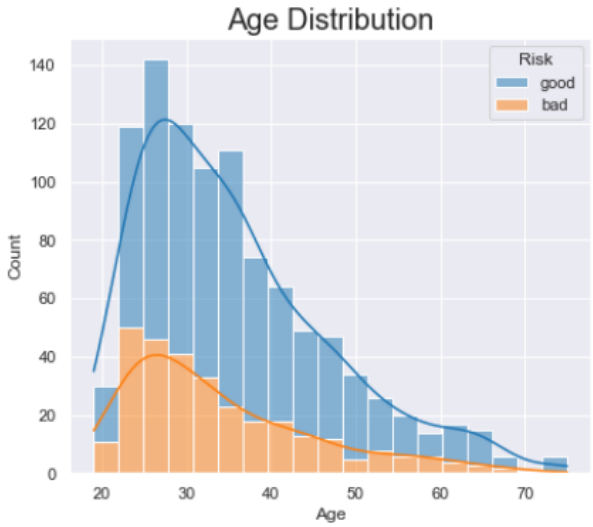

Figure 4.12: Representation of the age distribution

The details about the distribution:

- The positive skew in every graph indicates that the mean is higher than the median.
- Ages 20 to 30 are the most typical range for applicants to submit a loan application.

The value of the balance measure metrics on the original dataset:

Balance Measure Metrics	
Metrics	Value
Gini	0.6156
Shannon	0.7014
Simpson	0.4446
Imbalance Ratio	0.2345

Table 4.18: dataset:'Original German Credit Card', attribute: "age"

The pre-processing bias mitigation strategies from the aif360 library on the 'age' attribute.

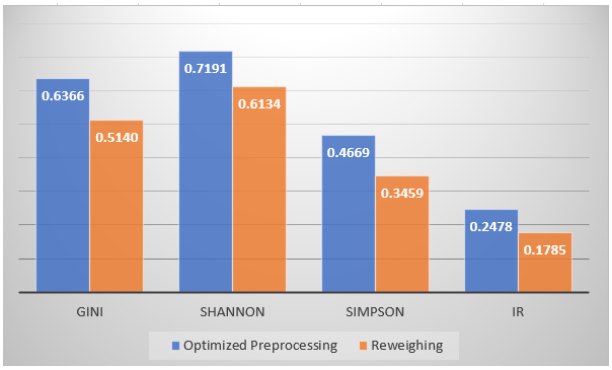

Figure 4.13: dataset:'Debiased German Credit Card', attribute: "age"

We can see that "Optimized Preprocessing" pre-processing bias mitigation improves the performance of our balance measure metrics specifically the Shannon Index is enhanced.

4.3 Default Credit Card

The third dataset under analysis is 'Default Credit Card' and the protected attributes chosen are: 'marriage', 'education'. For the Default Credit Card dataset we will study two stages : Original and Synthetic datasets.

The metrics used are:

- marriage: balance measure metrics;
- education: equality measure metrics;

Protected Attibute: 'marriage'

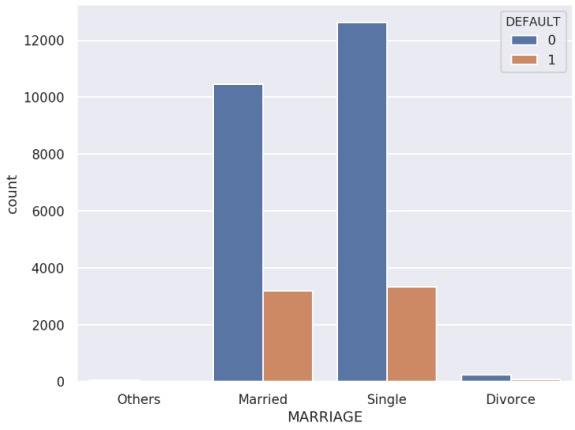

Figure 4.14: Representation of the marriage distribution

We can see that there are enormous difference between the default payments. Additionally, the attribute marriage is unbalanced because it primarily includes married and single people.

Below is the value of the balance measure metrics on the original dataset:

Balance Measure Metrics	
Metrics	Value
Gini	0.6792
Shannon	0.5439
Simpson	0.3461
Imbalance Ratio	0.0033

Table 4.19: dataset:'Original Default Credit Card', attribute: 'marriage'

We can see in this case the Gini Index shows a better results as it is more close to 1.

The value of balance measure metrics produced on the synthetic datasets for the "marriage" attribute.

Balance Measure Metrics			
Metrics	Syndata	Mostly AI	Gretel
Gini	0.2192	0.7561	0.9583
Shannon	0.2317	0.6681	0.9697
Simpson	0.0655	0.5089	0.9201
Imbalance Ratio	0.0019	0.0057	0.6611

Table 4.20: dataset:'Synthetic Default Credit Card', attribute: 'marriage'

Gretel is able to artificially produce a balanced dataset when compared to other synthetic data generating vendors, as can be seen. Following Gretel, we can observe that MostlyAI is situated close to its output.

Protected Attribute: Education

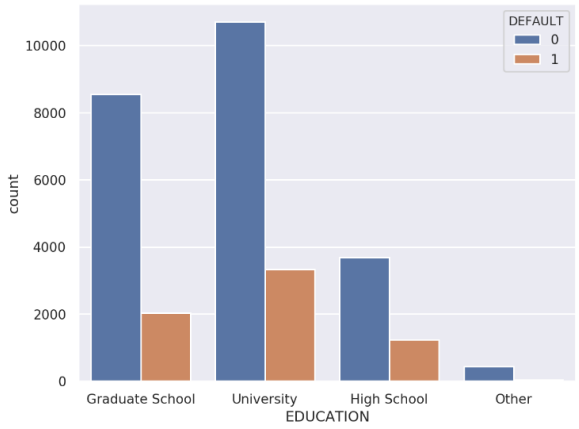

Figure 4.15: Representation of the education distribution

While there is a higher risk of default in high school and more balance there, there is greater inequity in graduate and post-secondary education due to the differences in default payments.

Below table is the representation of the equality measure metrics on the original dataset:

Equality Measure Metrics	
Metrics	Value
GEI	0.9996
Atkinson	0.9846
Theil	0.9342
COV	0.9846

Table 4.21: dataset: 'Original Default Credit Card', attribute: 'education'

The value of equality measure metrics produced on the synthetic datasets for the "education" attribute.

Equality Measure Metrics		
Metrics	Syndata	Gretel
GEI	0.9993	0.9995
Atkinson	0.9823	0.9866
Theil	0.9628	0.9461
COV	0.9823	0.9866

Table 4.22: dataset: 'Synthetic Default Credit Card', attribute: "education"

This value shows that people have similar levels of education. As other times Gretel is better performing vendor.

4.4 Lending Club

The fourth dataset under analysis is 'Lending Club' and the protected attributes chosen are: "home", "purpose". For the Lending Club dataset we will all the three stages : Original, Synthetic and Debiased datasets.

The metrics used are:

- home: equality measure metrics;
- purpose: balance measure metrics;

Protected Attribute: 'Home'

The "home" characteristic makes it possible to determine whether the present person has a "home" or not. The two dominant categories are "rent" and "mortgage".

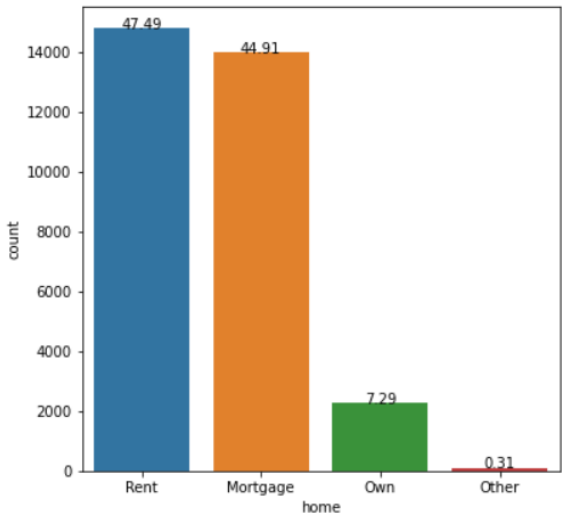

Figure 4.16: Representation of the home distribution

Below table is the representation of the equality measure metrics on the original dataset:

Equality Measure Metrics	
Metrics	Value
GEI	0.7566
Atkinson	0.6650
Theil	0.4373
COV	0.0065

Table 4.23: dataset: 'Original Lending Club', attribute: "home"

Below diagram is the representation of the equality measure metrics on the synthetic data generated:

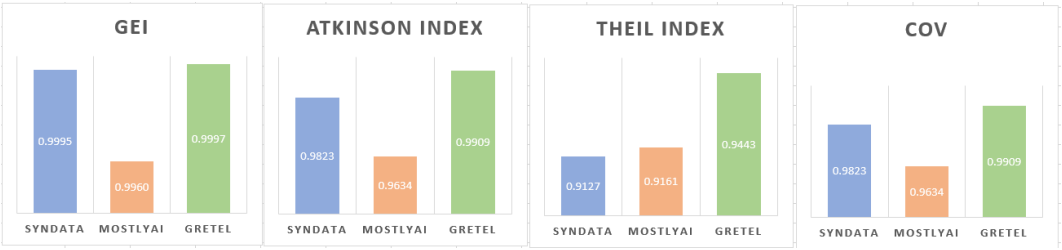

Figure 4.17: dataset: 'Synthetic Lending Club' attribute:'home'

The value of equality measure metrics produced on the SDK dataset using bias mitigation techniques for the "home" attribute.

Equality Measure Metrics			
Metrics	Bootstrapping	Reshapping	Data Bias
GEI	0.9998	0.9998	0.9998
Atkinson	0.9941	0.9953	0.9930
Theil	0.9523	0.9564	0.9516
COV	0.9941	0.9953	0.9930

Table 4.24

Protected Attribute: "Purpose"

As it a categorical feature that says what's the purpose to the loan, would be interesting to start by Purpose.

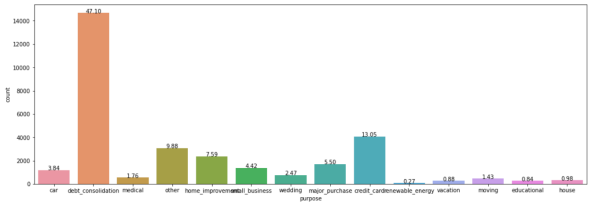

Figure 4.18: Representation of the home distribution

Below table is the representation of the balance measure metrics on the original dataset:

Balance Measure Metrics	
Metrics	Value
Gini	0.7945
Simpson	0.6947
Shannon	0.2164
Imbalance Ratio	0.0057

Table 4.25: dataset: 'Original Lending Club', attribute: 'purpose'

Following that we have the representation of the equality measure metrics on the synthetic datasets generated.

Balance Measure Metrics			
Metrics	Syndate	Mostly AI	Gretel
Gini	0.7409	0.7931	0.7772
Shannon	0.5424	0.6948	0.6391
Simpson	0.1696	0.2150	0.1703
Imbalance Ratio	0.0003	0.0056	0.0000

Table 4.26: dataset: 'Synthetic Lending Club', attribute: 'purpose'

Among the balance measure metrics we can say that Gini Index got an edge.

Next we check the value of the balance measure metrics on the data bias mitigation strategies taken by open source library SDK:

Balance Measure Metrics			
Metrics	Bootstrapping	Reshapping	Data Bias
Gini	0.8589	0.8543	0.8229
Shannon	0.7573	0.7490	0.7218
Simpson	0.3031	0.2953	0.2492
Imbalance Ratio	0.0048	0.0051	0.0171

Table 4.27: dataset: 'Debiased Lending Club', attribute: 'purpose'

4.5 Student-Por

The last data under analysis is the Student-por dataset and the protected attribute chosen are : "sex", "mother's_job" , "father's_job". For the Student-Por dataset we will study two stages : Original and Synthetic datasets. The metrics used are:

- sex: balance measure metrics;
- mother's_job: equality measure metrics;
- father's_job: equality measure metrics;

Protected Attribute: 'sex'

The gender ratio is 59% to 40%, which is not drastically out of balance but not not quite balanced either.

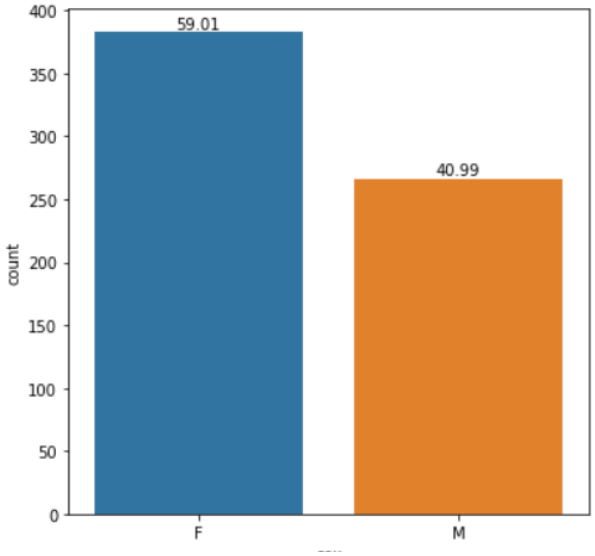

Figure 4.19: Representation of the sex distribution

The value of the balance measure metrics on the original dataset:

Balance Measure Metrics	
Metrics	Value
Gini	0.9675
Simpson	0.9764
Shannon	0.9370
Imbalance Ratio	0.6945

Table 4.28: dataset:'Original Student-Por', attribute: 'sex'

All the three balance measure metrics shows a greater value Gini, Shannon, Simpson.

The value of the balance measure metrics on the synthetic datasets

generated:

Balance Measure Metrics			
Metrics	Syndate	MostlyAI	Gretel
Gini	0.7597	0.7597	0.9197
Shannon	0.6950	0.8850	0.8960
Simpson	0.3873	0.6522	0.6962
Imbalance Ratio	0.0730	0.1654	0.1616

Table 4.29: dataset:'Synthetic Student-Por', attribute: 'sex'

The results by the balance measure metrics on the synthetic datasets is the same interpretation of the original datasets.

Protected Attribute: 'mother's_job'

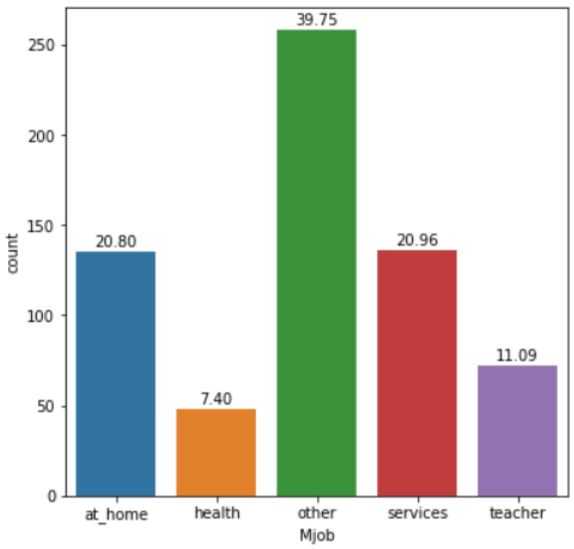

Figure 4.20: Representation of the mother's_job distribution

There are a total of 5 categories for the mother's job class, with "other" being the most prevalent followed by "services" and "at home."
The value of the equality measure metrics on the original dataset:

Equality Measure Metrics	
Metrics	Value
GEI	0.9792
Atkinson	0.9065
Theil	0.9112
COV	0.9065

Table 4.30: dataset:'Original Student-Por', attribute: 'mother's_job'

The value of the GEI shows a value close to followed by Atkinson and Theil.

Moving forward we analyze the protected attribute: **"mothers_job"** for three different vendors.

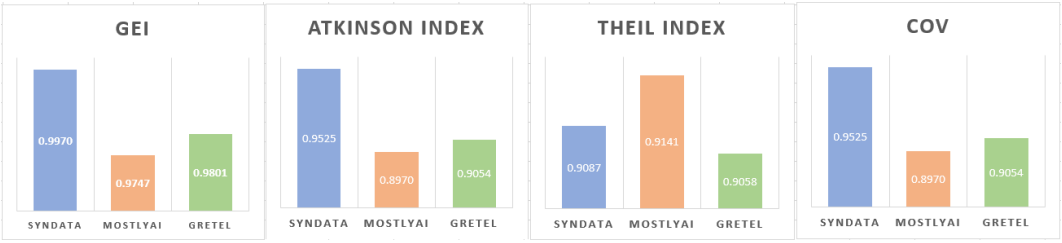

Figure 4.21: dataset:'Synthetic Student-Por', attribute:'mothers_job'

As we see that Syndata is showing high values it means that there is one category which might be significant for mother's_job.

Protected Attribute: 'father's_job'

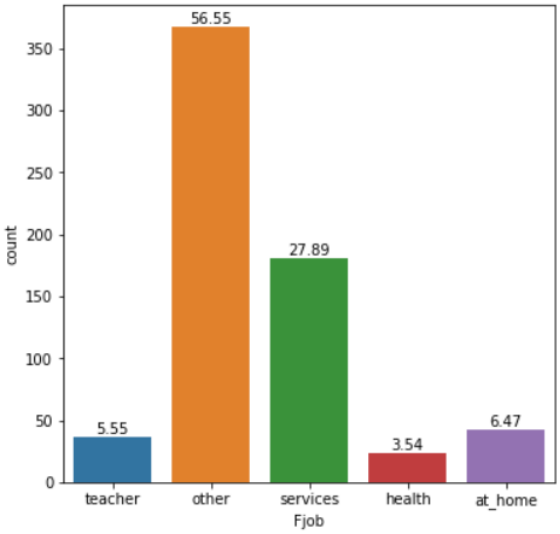

Figure 4.22: Representation of the father's_job distribution

In case of father's job "others" is dominant categories followed by all categories "services", "educators", "health care", and "homes".

The value of the equality measure metrics on the original dataset:

Equality Measure Metrics	
Metrics	Value
GEI	0.9961
Atkinson	0.9762
Theil	0.9796
COV	0.9762

Table 4.31: dataset:'Original Student-Por', attribute: 'father's_job'

The value of equality measure metrics for the dataset: "Student-Por",

attribute: 'father's_job'.

Equality Measure Metrics			
Metrics	Syndate	MostlyAI	Gretel
GEI	0.9995	0.9945	0.9968
Atkinson	0.9838	0.9701	0.9783
Theil	0.9888	0.9821	0.9818
COV	0.9838	0.9701	0.9783

Table 4.32: dataset:'Synthetic Student-Por', attribute: 'father's_job'

In every other instance, the outcome is essentially a representation of the original dataset.

Chapter 5

Conclusions and Future Works

5.1 Balance measure metrics

Gini and Shannon generally bring to lower penalization, independently on attribute cardinality or the type of disproportion. They are followed by Simpson and Imbalance Ratio.

One of the better balancing measures is considered to be Shannon. The Shannon index is calculated by taking into consideration all the categories of the specific class, and a number that is near 1 suggests that frequencies are more evenly distributed.

A dominance index, the Simpson index assigns more weight to the common or dominant category within the class. In this situation, a few rare categories with a small number of rare representatives won't have an impact on the diversity.

5.2 Equality measure metrics

In case of inequality measure metrics they also fall into the [0,1] range, where 0 denotes equality and 1 denotes inequality; thus, a number around 1 indicates greater inequality because it indicates that one class dominates the protected attribute, reducing randomness and eliminating

diversity.

As the equality measure moves closer to 1, as in the case of protected attributes like race, age, and relationship, it indicates that a particular class of the protected attribute is predominating.

Coefficient of variations tells about the distribution of that category, if it's close to 1 it indicates its near the mean whereas close to 0 represents more dispersed. So we should use cov to understand the distribution of the data.

5.3 Distance measure metrics

As we can see, the distance measure metrics are helpful because they make it simple for us to interpret the target attribute distribution in relation to the reference attribute distribution. So it tells us the difference between the two distribution. If the difference approaches 0, it indicates that the reference and target distributions are not distributed equally. However, if the difference is less than Infinity Norm Distance is closer to 1, indicating that the distribution is more equal. It performs calculations across a certain class of the protected attribute. When analyzing the link between "marital_status" and "income" using distance measure metrics, as mentioned in the section 4.1 We can observe that the value is quite near 1, which indicates that one particular class from the "martial_status" earns more than the remaining other.

Total Variation Distance is calculated over a group, which means it will precisely consider all classes for each protected attribute. It is preferable to have a value close to 1, as this indicates that various classes of the protected attribute have the same distribution as the target distribution. When analyzing "marital_status" versus "income", a value close to 0 is shown in the section 4.1 because it suggests that one class of the "marital_status" earns more than the remaining.

5.4 Performance by synthetic datasets

The synthetic dataset that functions well is the one that can upsample the rare class of each protected attribute. The synthetic dataset should be able to duplicate the original dataset while also producing a new dataset more evenly distributed. The performance of Gretel, one of the three vendors, is the best because it generates synthetic datasets with more balanced ratios for each class of a protected attribute. Despite the fact that the value of the balance measure metrics is typically close to that of the original dataset.Nevertheless, we can reduce bias in the dataset by using synthetic datasets.

Instead of attempting to balance out the ratio of the class of a protected attribute, Syndata just duplicates the original dataset to create synthetic datasets, which prevents it from considerably improving its performance.

5.5 Performance by bias mitigation strategies

To understand bias associated with a specific class of the protected attribute, "Optimized pre-processing" and "Learning Fair Representation" should be used, whereas "Reweighing" and "Disparate Impact Remover" should be used when we want to understand bias associated with the protected attribute as a whole (group, label). As it reduces bias, "Learning Fair Representation" should be employed only when we want to conceal some sensitive data, as in the case of income- and race-protected attributes in section 4.1. If we employ "Optimized Pre-processing" and "Learning Fair Representation" in the case of a continuous protected attributes, it will divide the attribute into a particular range and attempt to balance the ranges as mentioned in section 4.1.

In contrast, every technique in the SDK library can be used for a particular cause the Bootstrap technique should be used when the density of data is low. The Reshaping technique when building a predictive model for imbalanced classification as it aids in rebalancing the datasets

through conditional sampling. The Data Bias technique should also be used to upsample the rare class of the protected attribute after identifying biases in structured data. Among all the three techniques, Data Bias is the most ideal one as it is more suited for this study.

In order to conclude, the upcoming research and techniques have greater potential to handle bias at various levels: from generating balanced and ethically right synthetic dataset , using various tools and metrics for measuring bias before training a model, using in-processing bias mitigation strategies along with pre-processing. In the end we can use post-processing bias mitigation strategies before putting a product in market.

5.6 Future Works

The conducted study can be expanded by first including more datasets (consequently attributes) from various other domains like criminal justice, health care, and others among the proposed ones in section 3.1. It is feasible to construct more biased measures metrics at the dataset level, Demographic Parity and others among the proposed ones in section 3.4 to understand the skewness of the data. Moreover, our library can compute the bias measure metrics on one protected attribute at a time in this study, but in the future, it might be possible to compute the bias measure metrics on two attributes concurrently. For instance, we can extend our research to compute race and sex instead of only computing either race or sex.

It might be possible to leverage more synthetic data generation vendors as proposed in the section 3.5. Specifically few selected vendors, especially depending on the domain of the dataset. For example Hazy, might be used in the financial domain dataset. Similar to that, we could use vendors like Octopize MD, MD Clone, and others if the dataset was typical of the healthcare domain.

As all of our measures are being conducted at the dataset level, we are only using the pre-processing bias measure metrics. However, there is potential to utilize bias mitigation techniques at other phases, such as in-processing and post-processing, in the future. Nonetheless, we have

the possibility to train our model using the bias mitigated dataset as this can improve the performance.However, the proposed bias measure metrics are still valid and can be used as source material for further investigations.

References

[1] Rachel KE Bellamy et al. «AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias». In: *IBM Journal of Research and Development* 63.4/5 (2019), pp. 4–1 (cit. on pp. ii, 3, 21, 30, 35, 39).

[2] Sriram Vasudevan and Krishnaram Kenthapadi. «Lift: A scalable framework for measuring fairness in ml applications». In: *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 2020, pp. 2773–2780 (cit. on pp. ii, 3, 21, 35, 40).

[3] Mariachiara Mecati, Antonio Vetro, and Marco Torchiano. «Detecting Discrimination Risk in Automated Decision-Making Systems with Balance Measures on Input Data». In: *2021 IEEE International Conference on Big Data (Big Data)*. IEEE. 2021, pp. 4287–4296 (cit. on p. ii).

[4] Tiago Palma Pagano et al. «Bias and unfairness in machine learning models: a systematic literature review». In: *arXiv preprint arXiv:2202.08176* (2022) (cit. on p. 1).

[5] Artificial Intelligence. «Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy». In: *International Journal of Information Management* 57 () (cit. on p. 1).

[6] Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. «Algorithmic fairness: Choices, assumptions, and definitions». In: *Annual Review of Statistics and Its Application* 8 (2021), pp. 141–163 (cit. on p. 1).

[7] Sahil Verma and Julia Rubin. «Fairness definitions explained». In: *2018 ieee/acm international workshop on software fairness (fairware)*. IEEE. 2018, pp. 1–7 (cit. on p. 1).

[8] David Jones, Chris Snider, Aydin Nassehi, Jason Yon, and Ben Hicks. «Characterising the Digital Twin: A systematic literature review». In: *CIRP Journal of Manufacturing Science and Technology* 29 (2020), pp. 36–52 (cit. on p. 1).

[9] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. «Model cards for model reporting». In: *Proceedings of the conference on fairness, accountability, and transparency*. 2019, pp. 220–229 (cit. on p. 1).

[10] Dana Pessach and Erez Shmueli. «Algorithmic fairness». In: *arXiv preprint arXiv:2001.09784* (2020) (cit. on p. 2).

[11] Aileen Nielsen. *Practical Fairness*. O'Reilly Media, 2020 (cit. on p. 3).

[12] Sarah Bird, Miro Dudık, Richard Edgar, Brandon Horn, Roman Lutz, Vanessa Milan, Mehrnoosh Sameki, Hanna Wallach, and Kathleen Walker. «Fairlearn: A toolkit for assessing and improving fairness in AI». In: *Microsoft, Tech. Rep. MSR-TR-2020-32* (2020) (cit. on pp. 3, 21, 49).

[13] Pedro Saleiro, Benedict Kuester, Loren Hinkson, Jesse London, Abby Stevens, Ari Anisfeld, Kit T Rodolfa, and Rayid Ghani. «Aequitas: A bias and fairness audit toolkit». In: *arXiv preprint arXiv:1811.05577* (2018) (cit. on p. 3).

[14] Stuart J Russell. *Artificial intelligence a modern approach*. Pearson Education, Inc., 2010 (cit. on p. 6).

[15] Eric J Topol. «High-performance medicine: the convergence of human and artificial intelligence». In: *Nature medicine* 25.1 (2019), pp. 44–56 (cit. on p. 6).

[16] Arthur L Samuel. «Some studies in machine learning using the game of checkers. II—Recent progress». In: *IBM Journal of research and development* 11.6 (1967), pp. 601–617 (cit. on p. 6).

[17] Reva Schwartz, Apostol Vassilev, Kristen Greene, Lori Perine, Andrew Burt, Patrick Hall, et al. «Towards a Standard for Identifying and Managing Bias in Artificial Intelligence». In: (2022) (cit. on p. 12).

[18] Harini Suresh and John V Guttag. «A framework for understanding unintended consequences of machine learning». In: *arXiv preprint arXiv:1901.10002* 2 (2019), p. 8 (cit. on pp. 13–15).

[19] Kevin A Clarke. «The phantom menace: Omitted variable bias in econometric research». In: *Conflict management and peace science* 22.4 (2005), pp. 341–352 (cit. on p. 13).

[20] David B Mustard. «Reexamining criminal behavior: the importance of omitted variable bias». In: *Review of Economics and Statistics* 85.1 (2003), pp. 205–211 (cit. on p. 13).

[21] Stephanie K Riegg. «Causal inference and omitted variable bias in financial aid research: Assessing solutions». In: *The Review of Higher Education* 31.3 (2008), pp. 329–354 (cit. on p. 13).

[22] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. «A survey on bias and fairness in machine learning». In: *ACM Computing Surveys (CSUR)* 54.6 (2021), pp. 1–35 (cit. on pp. 14–16, 20).

[23] Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kıcıman. «Social data: Biases, methodological pitfalls, and ethical boundaries». In: *Frontiers in Big Data* 2 (2019), p. 13 (cit. on pp. 15, 16).

[24] Eszter Hargittai. «Whose space? Differences among users and non-users of social network sites». In: *Journal of computer-mediated communication* 13.1 (2007), pp. 276–297 (cit. on p. 15).

[25] Ricardo Baeza-Yates. «Bias on the web». In: *Communications of the ACM* 61.6 (2018), pp. 54–61 (cit. on p. 15).

[26] Ting Wang and Dashun Wang. «Why Amazon's ratings might mislead you: The story of herding effects». In: *Big data* 2.4 (2014), pp. 196–204 (cit. on p. 15).

[27] Hannah Jean Miller, Jacob Thebault-Spieker, Shuo Chang, Isaac Johnson, Loren Terveen, and Brent Hecht. «"Blissfully happy" or "ready tofight": Varying interpretations of emoji». In: *Tenth international AAAI conference on Web and social media*. 2016 (cit. on p. 15).

[28] Dong Nguyen, Rilana Gravel, Dolf Trieschnigg, and Theo Meder. «" How old do you think I am?" A study of language and age in Twitter». In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 7. 1. 2013, pp. 439–448 (cit. on p. 16).

[29] Jeffrey Dastin. «Amazon scraps secret AI recruiting tool that showed bias against women». In: *Ethics of Data and Analytics*. Auerbach Publications, 2018, pp. 296–299 (cit. on p. 16).

[30] Muhammad Ali, Piotr Sapiezynski, Miranda Bogen, Aleksandra Korolova, Alan Mislove, and Aaron Rieke. «Discrimination through optimization: How Facebook's Ad delivery can lead to biased outcomes». In: *Proceedings of the ACM on human-computer interaction* 3.CSCW (2019), pp. 1–30 (cit. on p. 16).

[31] Antonio Vetrò, Marco Torchiano, and Mariachiara Mecati. «A data quality approach to the identification of discrimination risk in automated decision making systems». In: *Government Information Quarterly* 38.4 (2021), p. 101619 (cit. on p. 16).

[32] T Jan and E Dwoskin. «Facebook is sued by HUD for housing discrimination». In: *The Washington Post* (2021) (cit. on p. 16).

[33] Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. «Bias in bios: A case study of semantic representation bias in a high-stakes setting». In: *proceedings of the Conference on Fairness, Accountability, and Transparency*. 2019, pp. 120–128 (cit. on p. 16).

[34] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. «Machine bias». In: *Ethics of Data and Analytics*. Auerbach Publications, 2016, pp. 254–264 (cit. on p. 17).

[35] Ziad Obermeyer and Sendhil Mullainathan. «Dissecting racial bias in an algorithm that guides health decisions for 70 million people». In: *Proceedings of the conference on fairness, accountability, and transparency*. 2019, pp. 89–89 (cit. on p. 17).

[36] Luca Oneto and Silvia Chiappa. «Fairness in machine learning». In: *Recent Trends in Learning From Data*. Springer, 2020, pp. 155–196 (cit. on p. 19).

[37] Deirdre K Mulligan, Joshua A Kroll, Nitin Kohli, and Richmond Y Wong. «This thing called fairness: Disciplinary confusion realizing a value in technology». In: *Proceedings of the ACM on Human-Computer Interaction* 3.CSCW (2019), pp. 1–36 (cit. on p. 19).

[38] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning. fairmlbook. org*. 2019 (cit. on p. 20).

[39] Indrė Žliobaitė. «Measuring discrimination in algorithmic decision making». In: *Data Mining and Knowledge Discovery* 31.4 (2017), pp. 1060–1089 (cit. on p. 20).

[40] Sorelle A Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. «On the (im) possibility of fairness». In: *arXiv preprint arXiv:1609.07236* (2016) (cit. on p. 20).

[41] Jon Kleinberg. «Inherent trade-offs in algorithmic fairness». In: *Abstracts of the 2018 ACM International Conference on Measurement and Modeling of Computer Systems*. 2018, pp. 40–40 (cit. on p. 20).

[42] Elena Beretta, Antonio Santangelo, Bruno Lepri, Antonio Vetrò, and Juan Carlos De Martin. «The invisible power of fairness. how machine learning shapes democracy». In: *Canadian Conference on Artificial Intelligence*. Springer. 2019, pp. 238–250 (cit. on p. 20).

[43] Antonio Vetro. «Imbalanced data as risk factor of discriminating automated decisions: a measurement-based approach». In: *J. Intell. Prop. Info. Tech. & Elec. Com. L.* 12 (2021), p. 272 (cit. on p. 20).

[44] Michelle Seng Ah Lee and Jat Singh. «The landscape and gaps in open source fairness toolkits». In: *Proceedings of the 2021 CHI conference on human factors in computing systems*. 2021, pp. 1–13 (cit. on p. 21).

[45] James Wexler, Mahima Pushkarna, Tolga Bolukbasi, Martin Wattenberg, Fernanda Viégas, and Jimbo Wilson. «The what-if tool: Interactive probing of machine learning models». In: *IEEE transactions on visualization and computer graphics* 26.1 (2019), pp. 56–65 (cit. on p. 22).

[46] Erroll Wood, Tadas Baltrušaitis, Charlie Hewitt, Sebastian Dziadzio, Thomas J Cashman, and Jamie Shotton. «Fake it till you make it: face analysis in the wild using synthetic data alone». In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 3681–3691 (cit. on p. 26).

[47] Faisal Kamiran and Toon Calders. «Data preprocessing techniques for classification without discrimination». In: *Knowledge and information systems* 33.1 (2012), pp. 1–33 (cit. on pp. 30, 54).

[48] 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*. 2015, pp. 259–268 (cit. on pp. 30, 54).

[49] Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. «Optimized preprocessing for discrimination prevention». In: *Advances in neural information processing systems* 30 (2017) (cit. on p. 30).

[50] *Adult Income Dataset*. URL: `https://archive.ics.uci.edu/ml/datasets/adult` (cit. on p. 31).

[51] *German Credit Data*. URL: `https://archive.ics.uci.edu/ml/datasets/Statlog+(German+Credit+Data)` (cit. on p. 32).

[52] *Default Credit Card Data*. URL: `https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients` (cit. on p. 33).

[53] *Student Performance Data*. URL: `https://archive.ics.uci.edu/ml/datasets/student+performance` (cit. on p. 34).

[54] Stefania Capecchi and Maria Iannario. «Gini heterogeneity index for detecting uncertainty in ordinal data surveys». In: *Metron* 74.2 (2016), pp. 223–232 (cit. on p. 35).

[55] Till Speicher, Hoda Heidari, Nina Grgic-Hlaca, Krishna P Gummadi, Adish Singla, Adrian Weller, and Muhammad Bilal Zafar. «A unified approach to quantifying algorithmic unfairness: Measuring individual &group unfairness via inequality indices». In: *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*. 2018, pp. 2239–2248 (cit. on p. 37).

[56] Andre Goncalves, Priyadip Ray, Braden Soper, Jennifer Stevens, Linda Coyle, and Ana Paula Sales. «Generation and evaluation of synthetic patient data». In: *BMC medical research methodology* 20.1 (2020), pp. 1–40 (cit. on p. 42).

[57] Julius A Adebayo et al. «FairML: ToolBox for diagnosing bias in predictive modeling». PhD thesis. Massachusetts Institute of Technology, 2016 (cit. on p. 48).

[58] Knut T Hufthammer, Tor H Aasheim, Sølve Ånneland, Håvard Brynjulfsen, and Marija Slavkovik. «Bias mitigation with AIF360: A comparative study». In: *Norsk IKT-konferanse for forskning og utdanning*. 1. 2020 (cit. on p. 49).

[59] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. «Learning fair representations». In: *International conference on machine learning*. PMLR. 2013, pp. 325–333 (cit. on p. 55).

[60] Dennis Wei, Karthikeyan Natesan Ramamurthy, and Flavio P Calmon. «Optimized score transformation for fair classification». In: *Proceedings of Machine Learning Research* 108 (2020) (cit. on p. 55).