**Bias Detection and Mitigation Techniques in Machine Learning Models: A Comparative Study**

**Name:** Chaudhary Usama Naeem

**Degree Program:** MSc in Computer Science

Contents

[1.1 Introduction 4](#_Toc203663520)

[Background and Motivation 4](#_Toc203663521)

[Problem Statement 5](#_Toc203663522)

[Research Aim and Objectives 6](#_Toc203663523)

[Research Questions 6](#_Toc203663524)

[Scope of the Study 7](#_Toc203663525)

[Structure of the Dissertation 8](#_Toc203663526)

[1.2 Literature Review 9](#_Toc203663527)

[1.2.1 The nature of Algorithmic bias in machine learning 9](#_Toc203663528)

[1.2.2 Measures and Defenses of Fairness 11](#_Toc203663529)

[1.2.3 Methods of Bias Mitigation: Summing them up 12](#_Toc203663530)

[1.2.4 Post-processing Techniques 14](#_Toc203663531)

[1.2.5 Toolkits for Fairness Auditing and Mitigation 15](#_Toc203663532)

[1.2.6 Previous Comparative Studies Review 16](#_Toc203663533)

[1.2.7 Study Limits and Rationale of the Research 17](#_Toc203663534)

[1.2.8 Summary 18](#_Toc203663535)

[Chapter 2: Methodology 18](#_Toc203663536)

[2.1 Research Design 18](#_Toc203663537)

[2.2 Dataset Selection 19](#_Toc203663538)

[2.3 Model Selection 20](#_Toc203663539)

[2.4 Bias Mitigation Techniques 21](#_Toc203663540)

[2.5 Evaluation Metrics 21](#_Toc203663541)

[2.6 Frameworks and Software tools 22](#_Toc203663542)

[2.7 Ethical Considerations 23](#_Toc203663543)

[2.8 Limitations of Methodology 23](#_Toc203663544)

[Chapter 3: Implementation 24](#_Toc203663545)

[3.1 Overview 24](#_Toc203663546)

[3.2 Dataset Preparation 25](#_Toc203663547)

[3.2.1 COMPAS Dataset 25](#_Toc203663548)

[3.2.2 German Credit Dataset 25](#_Toc203663549)

[3.3 Model Training 26](#_Toc203663550)

[A. Logistic Regression (LR) 26](#_Toc203663551)

[B. Random Forest (RF) 26](#_Toc203663552)

[C. Neural Network (NN) 26](#_Toc203663553)

[3.4 Application of Bias Mitigation Techniques 27](#_Toc203663554)

[A. Reweighing (Fairlearn, Pre-processing) 27](#_Toc203663555)

[B. Adversarial Debiasing (In-processing, AIF360) 27](#_Toc203663556)

[C. Equalized Odds Post-processing (AIF360) 29](#_Toc203663557)

[3.5 Evaluation Strategy 29](#_Toc203663558)

[Fairness Metrics 30](#_Toc203663559)

[Performance Metrics 30](#_Toc203663560)

[3.6 Tools and Libraries Used 30](#_Toc203663561)

[3.7 Challenges and Limitations 31](#_Toc203663562)

[References 33](#_Toc203663563)

# 1.1 Introduction

## Background and Motivation

Machine learning (ML) has rapidly become a core technology across numerous sectors such as finance, healthcare, education, human resource management, and criminal justice. These models are typically deployed to optimise predictive accuracy and automate decision-making. However, concerns have been raised over their ability to behave fairly across different segments of the population, particularly when models are trained on data that contains historical biases or structural inequalities.

One of the most cited examples of algorithmic bias in ML is the COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) tool, a proprietary risk assessment system used in the U.S. legal system. The system was designed to predict the likelihood of recidivism—that is, whether an individual who has previously been convicted of a crime is likely to reoffend. An investigative study by Angwin et al. (2016) revealed that COMPAS often assigned higher risk scores to Black defendants than White defendants with similar criminal histories. This case became central in discussions about algorithmic discrimination and has served as a critical example of how ML systems can produce unfair and potentially harmful decisions, particularly when used in sensitive domains.

Further examples have emerged in the fields of recruitment and credit scoring. For instance, Amazon discontinued an internal recruitment tool after it was found to favour male candidates over female ones for technical roles (Dastin, 2018). This bias stemmed from the training data, which was primarily composed of resumes submitted by male candidates over a ten-year period. Similarly, Cowgill and Tucker (2023) found that some commercial credit scoring systems penalise applicants from certain racial or socioeconomic backgrounds, despite the absence of explicit demographic features in the data. These cases highlight a recurring problem: even when protected attributes are removed, models may still learn discriminatory patterns through correlated proxy variables.

## Problem Statement

The above cases illustrate how bias can arise from multiple sources in the ML pipeline. According to Giffen et al. (2022), bias can be categorised into four main types: (1) historical bias, where societal inequalities are reflected in the data; (2) representation bias, which occurs when certain groups are underrepresented in the dataset; (3) measurement bias, stemming from inconsistent or subjective labelling; and (4) aggregation bias, which results when heterogeneous subpopulations are treated as a homogeneous group. These biases can compound across stages, making fairness a complex, systemic issue that extends beyond the algorithm itself.

In response, the field of fairness-aware machine learning has emerged, with the goal of developing tools and techniques to detect, evaluate, and mitigate bias in ML systems. Several academic toolkits now exist to support this work, including AI Fairness 360 (AIF360) by IBM, Fairlearn by Microsoft, and Google’s What-If Tool. These frameworks provide access to multiple bias detection metrics, fairness definitions, and mitigation strategies. Despite these developments, challenges persist.

One of the most significant gaps in the literature is the lack of standardised, comparative evaluations. Many studies focus on a single model or dataset and apply only one mitigation technique, which limits the generalisability of their findings. Moreover, fairness interventions often involve trade-offs—particularly between fairness and predictive performance—but these trade-offs are rarely analysed systematically (Siddique et al., 2024). As a result, practitioners lack clear guidance on which mitigation strategies are most effective across different contexts and how they impact core model performance metrics such as accuracy, precision, and recall.

## Research Aim and Objectives

This dissertation seeks to address this gap by evaluating the effectiveness of one mitigation strategy from each category across three common classifiers and two benchmark datasets. The project is designed as a comparative academic investigation rather than the development of a production-level tool.

The aim of the study is to conduct an empirical comparison of bias mitigation techniques and assess their effectiveness in improving fairness when applied to machine learning models trained on biased datasets.

Research Question:  
How effective are different bias mitigation techniques in improving fairness across selected machine learning models when trained on biased datasets?

Objectives:

* Identify and quantify bias in machine learning models trained on datasets with demographic imbalances.
* Apply one mitigation strategy from each major category—pre-processing (Reweighing), in-processing (Adversarial Debiasing), and post-processing (Equalized Odds)—to the selected models.
* Evaluate the models using multiple fairness metrics (SPD, EOD, DIR) and performance indicators (accuracy, precision, recall).
* Analyse the fairness-performance trade-offs associated with each mitigation strategy across three model types: logistic regression, random forest, and neural network.
* Develop a lightweight experimental prototype that enables reproducibility and comparison of results across configurations.

## Research Questions

In order to organize the research process and to ensure that the purpose does not get lost, the research question (RQ) outlined below have been formulated:

**RQ:** How effective are different bias mitigation techniques in improving fairness across selected machine learning models when trained on biased datasets?

The question will incorporate an investigation of the existence and the success of mitigation, in a variety of cases, as a more subtle and all-encompassing view of ML fairness.

## Scope of the Study

To maintain clarity and academic rigour, this study limits its scope to classification problems where sensitive group-based disparities may exist. Specifically, the research examines the performance of three well-established machine learning models—logistic regression, random forest, and a simple feed-forward neural network—on two publicly available datasets that are known to exhibit bias: the COMPAS Recidivism dataset and the German Credit dataset. These datasets were selected due to their frequent use in fairness-related research and their inclusion of protected attributes such as race, gender, and age.

Each model will be evaluated both with and without bias mitigation strategies to assess the extent of bias and the effectiveness of each technique. The mitigation strategies chosen for this study include: (1) Reweighing, a pre-processing method that adjusts instance weights to reduce bias during training; (2) Adversarial Debiasing, an in-processing method that adds a secondary adversarial model to penalise unfair predictions; and (3) Equalized Odds Post-processing, a method that adjusts predicted outputs to satisfy fairness constraints. These strategies were selected because they represent distinct points in the ML pipeline and are accessible through widely used open-source toolkits such as Fairlearn and AIF360.

The fairness of each model will be evaluated using three metrics: Statistical Parity Difference (SPD), Equal Opportunity Difference (EOD), and Disparate Impact Ratio (DIR). These metrics are widely accepted in fairness literature and provide a comprehensive view of group-level disparities. At the same time, traditional model performance metrics such as accuracy, precision, recall, and F1-score will be recorded to explore the trade-offs introduced by bias mitigation.

The ultimate aim of this dissertation is not to optimise model performance or develop new fairness algorithms, but rather to conduct a comparative academic evaluation of how different mitigation techniques perform across model types and datasets. This work is intended to contribute to the growing body of fairness-aware ML research by offering empirical insights that are reproducible, explainable, and relevant to future studies.

## Structure of the Dissertation

In this dissertation, it will take the following chapters:

**Chapter 1- Introduction and Literature Review:** An introduction to the subject, description of the research problem and objectives, and critical review of literature are presented.

**Chapter 2- Methodology:** Details of the research design, sets of data, models, mitigation techniques, and examinations.

**Chapter 3- Implementation:** This chapter describes the implementation of the experimental setup, including the preprocessing of datasets, training of machine learning models, and application of fairness mitigation techniques using Fairlearn and AIF360. It will also include brief technical details of the tools and programming environment used.

**Chapter 4- Results:** This chapter will present the results of the experiments, including performance metrics (e.g., accuracy, precision) and fairness metrics (e.g., statistical parity difference, equal opportunity difference, disparate impact ratio). Visual comparisons and tables will be used to highlight the effects of each mitigation strategy across different models and datasets.

**Chapter 5- Discussion:** This chapter will interpret the results in light of the research question and objectives. It will critically examine the trade-offs between fairness and predictive performance and reflect on the effectiveness and limitations of each bias mitigation approach.

**Chapter 6- Conclusion:** This final chapter will summarise the key findings of the study, revisit the research question, and highlight how the results contribute to the literature on algorithmic fairness. It will also discuss limitations and propose directions for future research, such as applying additional fairness techniques or testing on different datasets.

## 1.2 Literature Review

### 1.2.1 The nature of Algorithmic bias in machine learning

Fairness in machine learning has become a central concern as algorithms are increasingly deployed in high-impact decision-making settings. These include sectors such as criminal justice, banking, hiring, education, and healthcare, where biased predictions can have serious consequences for individuals or communities. Bias in this context refers to systematic disparities in outcomes across different demographic groups, often linked to protected characteristics such as race, gender, or age. Bias can manifest in the form of unequal prediction rates, skewed error rates, or disparities in access to opportunities and resources.

Multiple studies have shown that such disparities are not hypothetical. For example, Raji and Buolamwini (2023) demonstrated that facial recognition systems from major technology vendors showed higher error rates for darker-skinned females compared to lighter-skinned males. In credit scoring systems, Cowgill and Tucker (2023) found that applicants from minority ethnic backgrounds were more likely to be denied credit—even after controlling for relevant financial variables—due to the presence of proxy features that correlated with race or socioeconomic status.

Bias in ML systems can arise from multiple points in the pipeline. Giffen et al. (2022) provide a detailed classification of four types of bias relevant to ML:

* Historical Bias: Systemic inequalities present in society are embedded in training data.
* Representation Bias: Underrepresentation of specific groups, leading to models that generalise poorly for those populations.
* Measurement Bias: Inaccurate or subjective labels that skew the relationship between input features and output predictions.
* Aggregation Bias: A failure to account for heterogeneity across subgroups during model design and evaluation.

These biases are not mutually exclusive; they may compound over the course of model development and evaluation. For instance, in the COMPAS dataset, bias was introduced both through historical criminal justice patterns and through subjective risk scoring (Angwin et al., 2016). The risk of reoffending, or recidivism, was used as the prediction target. However, due to unequal policing and systemic bias, this label may itself be flawed, which means even a high-performing model could produce discriminatory results.

This literature makes it clear that mitigating bias cannot be achieved solely by removing protected attributes such as race or gender from the input features. As Barocas et al. (2020) note, removing sensitive features often has limited impact because ML algorithms can reconstruct such attributes using correlated variables (e.g., zip code as a proxy for ethnicity). Therefore, fairness must be approached through algorithmic design and evaluation, not just through data sanitisation.

### 1.2.2 Measures and Defenses of Fairness

To address bias in machine learning, researchers have proposed formal definitions of fairness, each capturing different notions of equity. However, these definitions are often incompatible, and selecting one depends heavily on the application context. Among the most widely used are:

* Statistical Parity Difference (SPD): This measures the difference in the rate of favourable outcomes between unprivileged and privileged groups. An SPD value close to 0 indicates that both groups receive positive predictions at similar rates.
* Equal Opportunity Difference (EOD): This focuses on true positive rates across groups, requiring that qualified individuals—those who truly belong to the positive class—have equal chances of being correctly predicted.
* Disparate Impact Ratio (DIR): This is the ratio of favourable predictions for the unprivileged group to those for the privileged group. A value below 0.8 is typically flagged as potentially discriminatory, based on the U.S. Equal Employment Opportunity Commission’s 80% rule.

Each metric captures a different fairness goal. For example, SPD is concerned with group-level parity in outcomes, whereas EOD focuses on individuals who merit a favourable decision. It is important to note that these metrics can be mathematically incompatible; when base rates differ between groups, it may be impossible to optimise for all fairness definitions simultaneously (Zliobaite, 2021). This conflict is known as the “impossibility theorem” in algorithmic fairness literature.

The choice of fairness metric has a significant influence on the evaluation of bias mitigation techniques. For instance, a model might achieve high statistical parity by randomly flipping decisions for some group members, but this could harm accuracy or lead to unfair treatment of individuals within the group. Conversely, focusing on equal opportunity might ignore broader disparities in outcome distribution. Siddique et al. (2024) recommend using multiple metrics in parallel to obtain a more balanced view of fairness performance.

This study follows that approach by evaluating each model and mitigation technique using SPD, EOD, and DIR. These three metrics offer a comprehensive perspective on group-level and error-rate-based fairness. They are also supported by open-source libraries such as Fairlearn and AIF360, which will be used in the practical component of this project.

In addition to fairness, model performance will be evaluated using standard metrics such as accuracy, precision, recall, and F1-score. This dual-metric evaluation is crucial for understanding the trade-offs between fairness and utility, which are often overlooked in academic studies but are essential in real-world applications.

### 1.2.3 Methods of Bias Mitigation: Summing them up

Bias mitigation in machine learning can be performed at different stages of the model development process. These techniques are commonly grouped into three categories: pre-processing, in-processing, and post-processing. Each category targets different components of the ML pipeline and comes with its own set of trade-offs in terms of effectiveness, model compatibility, and computational cost.

Pre-processing techniques aim to correct bias before model training. These methods operate on the input data to ensure that the distributions across protected groups are more balanced. One of the most well-known techniques is Reweighing, which adjusts the weights of training examples based on group and label combinations. For example, if one group is underrepresented in the positive class, their samples may be given more weight to ensure fairer learning outcomes (Kamiran & Calders, 2012; Wadsworth et al., 2021). These methods are typically model-agnostic, meaning they can be used with any classifier. However, they may distort the data distribution or reduce model generalisability if not applied carefully.

In-processing methods incorporate fairness constraints directly into the model training phase. One prominent example is Adversarial Debiasing, which trains a secondary adversarial model to detect the protected attribute from the primary model’s output. The main model is penalised if the adversary succeeds, encouraging predictions that are independent of protected characteristics (Bagdasaryan et al., 2020). These approaches offer fine-grained control but can be complex to implement and require significant tuning to balance accuracy and fairness.

Post-processing methods modify the model’s predictions after training. These strategies are particularly useful when the model is a black box or cannot be retrained. Equalized Odds Post-processing is a common technique that adjusts prediction thresholds to equalise true positive and false positive rates across demographic groups (Hardt et al., 2016; Singh et al., 2022). While flexible and easy to apply, post-processing can lead to inconsistencies in individual-level decisions, especially when similar individuals receive different outcomes due to group-based adjustments.

In this study, one method from each category will be applied: Reweighing (pre-processing), Adversarial Debiasing (in-processing), and Equalized Odds Post-processing. These techniques are widely used, well-documented in fairness literature, and available through open-source toolkits such as Fairlearn and AIF360.

### 1.2.4 Post-processing Techniques

Over the past few years, several open-source toolkits have emerged to assist researchers and practitioners in assessing and mitigating bias in machine learning systems. These toolkits make it easier to apply fairness-aware algorithms, compute multiple fairness metrics, and generate visual reports.

AI Fairness 360 (AIF360) is one of the most comprehensive libraries, developed by IBM Research. It provides implementations of over 70 fairness metrics and more than 10 mitigation algorithms, covering all three stages of bias mitigation. The toolkit is available in Python and includes Jupyter notebooks for reproducibility and ease of experimentation (Bellamy et al., 2020).

Fairlearn, developed by Microsoft, focuses on reduction-based in-processing techniques and includes tools to visualise fairness-performance trade-offs. It is designed to be modular and integrates well with scikit-learn pipelines, making it accessible to researchers with standard Python ML experience (Bird et al., 2022).

Google’s What-If Tool is a visual interface for TensorFlow-based models. It allows users to conduct counterfactual testing, compare group-wise performance, and visualise decision boundaries without writing code (Wexler et al., 2021). While it offers strong visual diagnostics, it is less customisable than AIF360 and Fairlearn.

This project uses AIF360 and Fairlearn due to their flexibility, active maintenance, and compatibility with the chosen mitigation techniques. Both toolkits support the fairness metrics and model types selected for this study.

### 1.2.5 Toolkits for Fairness Auditing and Mitigation

Despite the growing body of work in fairness-aware ML, several gaps remain. First, most existing studies focus on only one model, one dataset, or one mitigation technique, limiting the generalisability of findings. Comparative studies across multiple model types—such as linear models, tree-based models, and neural networks—are still relatively rare.

Second, the interaction between fairness and performance is not sufficiently explored in the literature. Improving fairness often reduces predictive accuracy, but few studies quantify this trade-off systematically. Without such analysis, practitioners are left with little guidance on which models or mitigation methods are appropriate for their specific application needs.

Third, many studies lack reproducibility. Even though open-source toolkits now exist, experiments are often not fully documented or shared, making it difficult for other researchers to validate findings. There is also limited integration of multiple mitigation strategies within the same evaluation framework.

This dissertation aims to address these limitations by designing a focused and reproducible study that compares the effectiveness of three mitigation techniques across three ML models and two datasets. The project uses standard toolkits, evaluates across multiple fairness metrics, and documents all processes to support replicability.

By narrowing the scope and applying consistent experimental design, the study contributes to clearer understanding of fairness trade-offs in classification problems. The next chapter outlines the methodology used to implement and evaluate the proposed experiments.

### 1.2.6 Previous Comparative Studies Review

Although fairness metrics and toolkits are available, the literature does not include extensive comparisons of several types of models and datasets. Most of the existing studies are limited because they concentrate on a specific data set with no generalization or only on one mitigation technique.

To give some examples, Wang et al. (2021) used a reweighed variant of the Adult Income dataset to list the results of fairness in a logistic regression model and decision trees. Though informative, the research considered only a few of the metrics and a single pre-processing strategy. On the same note, Zhang and Gong (2020) investigated how to promote fairness in the credit scoring model but failed to advance their discussions to in-processing or post-processing. Their conclusions are very helpful but not scalable and reproducible in other realms.

More modern meta-analyses show that application effectiveness of mitigation techniques depends strongly on model type and structure of the data. A systematic survey given by Siddique et al. (2024) reveals that those mitigation techniques which are effective to linear models do not work on the neural networks, at least, having to face overfitting or being sensitive to regularization. Such inconsistency marks the necessity to have well-organized and repeatable experiments, particularly in complicated settings containing several attributes under protection.

Further, studies that also focus on fairness-performance trade-offs which are central to practitioners are very few. Fairness metrics are critical, yet none of the organizations will implement a model that is fair, but inaccurate to an unacceptable extent. Thus, in order to make an informed decision, it is important to learn how to measure and estimate the price of fairness (Bagdasaryan et al., 2020).

### 1.2.7 Study Limits and Rationale of the Research

This literature review presents a number of gaps in the body of literature:

**Insufficient holistic comparison:** The majority of studies compare fairness “in a vacuum” (i.e., they test a single type of mitigation method or a single ML model). Few studies have compared fairness performance across a variety of model architectures (e.g., linear models, tree-based models, and neural networks).

**Trade-offs that are underexplored:** There are few studies estimating the trade-off between fairness and performance that results from applying mitigation techniques. Without this comparison, practitioners are left choosing between accuracy and fairness without evidence-based guidance.

**Low reproducibility:** Although toolkits exist to support fairness analysis, they are rarely integrated into a complete and reusable pipeline that promotes reproducibility and real-world applicability.

The current dissertation addresses these gaps by designing and implementing a comparative framework to evaluate bias mitigation strategies across different model types and datasets. Specifically, the study uses two publicly available datasets that are commonly used in fairness research: the COMPAS Recidivism dataset and the German Credit dataset. These datasets include protected attributes such as race, age, and gender, and serve as benchmarks in algorithmic fairness studies.

The study applies three mitigation techniques—Reweighing (pre-processing), Adversarial Debiasing (in-processing), and Equalized Odds Post-processing (post-processing)—to three machine learning models: logistic regression, random forest, and a neural network. All configurations are evaluated using standard fairness and performance metrics and implemented using reproducible pipelines supported by AIF360 and Fairlearn.

This dual focus—on rigorous, multi-model comparison and reproducibility—makes the study well-aligned with current research needs in algorithmic fairness. The scope is intentionally narrow to ensure results are meaningful and interpretable within the limits of a taught Master’s project.

### 1.2.8 Summary

Machine learning bias is an ethically, socially and technically challenging problem. There are several definitions of fairness but none can provide the best solution. Pre-, in-, and post-processing mitigation strategies are good tools to have but effective on a model-specific and domain-specific basis. Fairness tools, such as AIF360 and Fairlearn, now allow access to fairness assessments, but the empirical evidence of the use in real life applications has not been satisfied.

In this regard, this dissertation would address this gap by carrying out a reproducible and systematic comparative analysis of biases in ML models. It also tests the performance of various mitigations on a regular measure of differing algorithms and datasets. It is hoped that the results will form both the theoretical insights and practical models of how to scale equitable machine learning systems to be applied in practice.

# Chapter 2: Methodology

## 2.1 Research Design

This project adopts a comparative, experiment-based methodology that aligns with the objectives of a taught Master's dissertation. The primary aim is to investigate the effectiveness of different bias mitigation techniques when applied to machine learning models trained on datasets containing demographic imbalances. The study does not attempt to develop new algorithms or frameworks but instead evaluates existing techniques using publicly available data, common classifiers, and reproducible workflows.

A comparative experimental approach is suitable for this investigation as it allows for controlled testing of fairness metrics, model behaviours, and mitigation effects across multiple configurations. The work is exploratory and quantitative in nature, guided by a single focused research question and a defined set of evaluation criteria.

The implementation and experiments will be conducted using a lightweight Python-based prototype. This prototype will include modules for data preprocessing, model training, mitigation application, and metric calculation. The aim is to enable systematic comparisons across model types and techniques without building an industrial-grade system.

## 2.2 Dataset Selection

Two datasets will be used for experimentation. These datasets were chosen based on their frequent use in fairness-related research, presence of protected attributes, and accessibility through open-source repositories.

A. COMPAS Recidivism Dataset  
This dataset contains information about criminal defendants, including demographic attributes and whether they reoffended within two years. It has been widely used to highlight racial disparities in algorithmic decision-making. The protected attribute in this case is race (Black vs. White), and the target variable is two-year recidivism (Yes/No).

B. German Credit Dataset  
This dataset includes information about loan applicants, including age, gender, and credit status. It has been used to study fairness in financial decision-making. The protected attributes selected are age (under 25 vs. 25 and older) and gender (Male vs. Female), and the target variable is creditworthiness (Good vs. Bad credit).

Both datasets are publicly available and will be sourced from well-known repositories such as the UCI Machine Learning Repository and ProPublica. Before use, each dataset will be reviewed for missing values, inconsistent labels, and skewed distributions.

## 2.3 Model Selection

To prepare the data for model training, the following steps will be applied:

* Handling Missing Data: Any null or missing values will be imputed using the median for numerical features and the mode for categorical features.
* Encoding Categorical Variables: One-hot encoding will be used for nominal variables, while ordinal features (if any) will be encoded accordingly.
* Feature Scaling: Numeric features will be standardised using z-score normalisation.
* Train-Test Split: Each dataset will be split into training (70%) and testing (30%) sets using stratified sampling on the target variable to maintain class distribution.

Throughout preprocessing, the protected attributes will be retained for fairness evaluation but excluded from model inputs unless required by a mitigation method.

## 2.4 Bias Mitigation Techniques

Three machine learning classifiers will be used in this study:

* Logistic Regression (LR): A linear model that provides interpretable coefficients and serves as a baseline.
* Random Forest (RF): A tree-based ensemble model known for its robustness and performance on tabular data.
* Feedforward Neural Network (NN): A basic multilayer perceptron model used to represent more complex non-linear decision functions.

These models were selected to reflect a range of algorithmic complexities and training behaviours. All models will be implemented using the scikit-learn and TensorFlow libraries. Hyperparameters will be tuned using cross-validation on the training set.

## 2.5 Evaluation Metrics

To assess the impact of mitigation, one method from each major category will be applied:

A. Pre-processing – Reweighing  
Implemented via Fairlearn, this technique adjusts sample weights in the training data based on group-label combinations to equalise representation. It is effective when the dataset itself exhibits historical or representational bias (Kamiran & Calders, 2012).

B. In-processing – Adversarial Debiasing  
Implemented via AI Fairness 360, this method trains an adversary to predict the protected attribute from the model’s output. The classifier is penalised for outputs that allow such predictions, thereby encouraging fairer representations (Bagdasaryan et al., 2020).

C. Post-processing – Equalized Odds Post-processing  
Also from AI Fairness 360, this method adjusts predicted labels to balance true positive and false positive rates across groups. It is applied after training, allowing fairness corrections without modifying the model itself (Hardt et al., 2016; Singh et al., 2022).

Each mitigation method will be applied separately to all three classifiers across both datasets.

## 2.6 Frameworks and Software tools

The effectiveness of mitigation strategies will be assessed using a combination of fairness and performance metrics.

Fairness Metrics:

* Statistical Parity Difference (SPD): Measures difference in favourable outcome rates between protected groups.
* Equal Opportunity Difference (EOD): Compares true positive rates between groups.
* Disparate Impact Ratio (DIR): Ratio of favourable predictions between unprivileged and privileged groups (ideal ≈ 1.0).

Performance Metrics:

* Accuracy
* Precision
* Recall
* F1-Score

All metrics will be computed using AIF360 and Fairlearn’s metric APIs. Results will be reported in both tabular and visual formats to highlight trade-offs..

## 2.7 Ethical Considerations

The implementation will be conducted in Python 3.11 using Jupyter notebooks for transparency and reproducibility. Key libraries will include:

* scikit-learn (model training and evaluation)
* TensorFlow/Keras (neural network implementation)
* Fairlearn (preprocessing and metric analysis)
* AIF360 (in-/post-processing and fairness metrics)
* pandas, numpy (data handling)
* matplotlib, seaborn (visualisation)

Each experimental run will be logged and saved in structured formats (e.g., CSV) to support reproducibility. Results will be analysed using grouped bar charts and fairness-performance trade-off plots.

## 2.8 Limitations of Methodology

This study uses only publicly available datasets that are fully anonymised and do not contain personally identifiable information. Ethical considerations focus on responsible interpretation of fairness results and avoidance of overgeneralisation. Fairness metrics will be reported with contextual discussion to ensure findings are not misrepresented.

This study is limited by its scope: only two datasets, three classifiers, and three mitigation strategies will be used. While this design enables comparative analysis, it may not generalise to other domains or models. In addition, fairness definitions themselves can be contested, and no single metric can fully capture what is “fair” in all settings.

# Chapter 3: Implementation

## 3.1 Overview

In this chapter, the practical application of the methodology provided in Chapter 2 is outlined. The main aim of the implementation was to perform a comparative empirical assessment of the success rate of the cherry-picked bias mitigation techniques of various machine learning models trained on known demographically biased datasets. In order to accomplish this, a Python prototype was designed to enable the three steps (1) dataset preprocessing, (2) model training and evaluation, and (3) use of fairness mitigation in the standard libraries.

Three models were employed in the implementation which include the Logistic Regression (LR), Random Forest (RF), and the simple Feedforward Neural Network (NN). Every model was tested in its unmitigated version and subsequently re-tested after three forms of fairness mitigation; Reweighing (pre-processing), Adversarial Debiasing (in-processing), and Equalized Odds Post-processing were employed on two publicly available datasets, COMPAS Recidivism and German Credit.

The implementation within the environment consisted of Python 3.11, and the extraordinary libraries were scikit-learn, Fairlearn, AIF360, NumPy, Pandas, Matplotlib, and TensorFlow/ Keras. Jupyter Notebooks Jupyter Notebooks were used to develop and test the code locally.

## 3.2 Dataset Preparation

### 3.2.1 COMPAS Dataset

The ProPublica GitHub repository was used as a source of the COMPAS dataset that represents processed criminal justice data applied in the previous fairness studies. The attribute (or group of attributes) that was chosen to be the protected one in this research was race, namely the comparison of the outcomes between Black and White defendants, whereas the target one was the two-year recidivism (a binary variable that represented whether the person reoffended or not within a two-year term of release or not).

Preliminary cleanup activities were:

* Sifting records holding valid data on the chosen properties.
* Dropping features of high multi-collinearity or strong proxy (e.g. charge description).
* Cattle conversion of categorical variables like sex and the degree of charge.
* Converting the race situation into a binary character (Black = 1, White = 0).

### 3.2.2 German Credit Dataset

Data of German Credit has been taken by processing the UCI Machine Learning Repository. It contains a population and monetary information of 1000 of the people who were assessed regarding their credit risk. Age (young: <25, older: 25 and over) and gender, (male/female) were the characteristics selected to make the protection on and the target was credit risk (good = 1, bad = 0).

To do this, the following steps were prepared:

* Encoding of categories to one-hot encodings.
* Standardizing Numerial Features with MinMaxScaler.
* Generating binary flags with the regards to the protected attributes.
* Deletion of all the personal identifiable features and redundant features.

Both data-sets were partitioned into 70 percent training and 30 percent test split where the stratified sampling was used to maintain a balance of classes. The outcomes obtained in the data form were saved in CSV files to reuse on different experiments.

## 3.3 Model Training

On each of the dataset, three models were trained in their unmitigated form as follows:

### A. Logistic Regression (LR)

This model run with the LogisticRegression algorithm of scikit-learn was a simple model, but effective as a method to have a first look at the data and provided a concrete baseline. Through 5-fold cross-validation hyperparameters like regularization strength (C) were optimized.

### B. Random Forest (RF)

This ensemble method ranked higher in the overfitting resistant and higher performance in the non-linear results using the RandomForestClassifier scikit-learn. Optimization of the number of estimators, the maximum depth, and the criteria of selecting the features to split were performed on the basis of the validation accuracy.

### C. Neural Network (NN)

A typical feedforward neural network was carried out with Keras (TensorFlow backend). Its architecture was:

* input layer match feature size,
* Two ReLU hidden layers (e.g. 32 and 16 units).
* A binary classification successively sigmoid-activated output layer.

The model is put together with the binary cross-entropy loss parameter and Adam optimizer and trained using the batch size of 32 epochs. Dropout (rate = 0.2) was used to avoid over fitting.

The models were trained and tested without the application of any fairness interventions to form a baseline of performance as well as fairness metrics.

## 3.4 Application of Bias Mitigation Techniques

The mitigation of bias was used to work along three steps, and everyone was another point in the ML process.

### A. Reweighing (Fairlearn, Pre-processing)

The Reweighing module of Fairlearn was used to perform the Reweighing. The method uses weights of the instances in the training set so as to provide parity in the distribution of classes based on groups that are being protected. The denomination of the mitigated dataset was then applied to retrain the three models. It is less invasive and effective on the different types of models.

from fairlearn.preprocessing import Reweighing

rw = Reweighing(prot\_attr\_names=['race'], dataset=X\_train)

X\_train\_rw, y\_train\_rw = rw.fit\_transform(X\_train, y\_train)

This prepped data found its way to the model training pipeline and the same testing was evaluated.

### B. Adversarial Debiasing (In-processing, AIF360)

The implementation of Adversarial Debiasing was performed through AIF360s AdversarialDebiasing model and it is built on TensorFlow. This was a method of simultaneously training a classifier and an adversarial network in order to accurately predict which of the model outputs had the sensitive attribute. The primary classifier will be penalised in the event that the adversary discovers the membership of the group successfully, and lean the model towards generating group-invariant predictions. This is TensorFlow specific and restricts the application of the method to TensorFlow models, as well as needs to transform data into the BinaryLabelDataset model of AIF360. The label and Protected attributed fields were clearly stated.

Example setup:

from aif360.algorithms.inprocessing import AdversarialDebiasing

sess = tf.Session()

ad\_model = AdversarialDebiasing(

privileged\_groups=[{'race': 0}],

unprivileged\_groups=[{'race': 1}],

scope\_name='adv\_debiasing',

debias=True,

sess=sess

)

ad\_model.fit(train\_bld)

This method has been used in both datasets (COMPAS and German Credit) and after model training, the model was tested on the test data using the same fairness measures to have consistency. Adversarial debiasing approach was more flexible in terms of fairness tuning, but took more resources to both set up and train than reweighing.

### C. Equalized Odds Post-processing (AIF360)

The Equalized Odds Post-processing approach seeks to alter the predicted labels of a trained model in a manner that tries to equalize the rates at which the model gives false positive and true positive across the protected groups. It accomplishes this without re-training the model, an aspect that makes it particularly applicable in cases where fixed or third-party models are to be used, where internal retraining is impossible. Implemented using:

from aif360.algorithms.postprocessing import EqOddsPostprocessing

eq\_post = EqOddsPostprocessing(

privileged\_groups=[{'race': 0}],

unprivileged\_groups=[{'race': 1}]

)

eq\_post = eq\_post.fit(train\_bld, pred\_bld)

preds\_eq = eq\_post.predict(test\_bld)

This procedure was carried out following the training of every model on the initial dataset. It came in handy in explaining the minimal-intrusion bias correction approach. The adjustments, however, usually translated to some loss of accuracy, particularly, when such complicated models as neural networks are in the picture.

## 3.5 Evaluation Strategy

In order to make a solid and fair comparison between the models and mitigation techniques the following fairness and performances measures were directly calculated per configuration (baseline and mitigated):

### Fairness Metrics

* Statistical Parity Difference (SPD): desired value is 0. Differentiates selected proportion between privileged and unprivileged groups.
* Equal Opportunity Difference (EOD): Preferably, this value is 0. Comparative measure of the true positive rates.
* Disparate Impact Ratio (DIR): The value should be 1. A DIR lower than 0.8 is viewed as unlawfully discriminative in the U.S. court.

### Performance Metrics

* Accuracy
* Precision
* Recall
* F1-score

The test set was used to calculate metrics with custom scoring functions and AIF360 embedded evaluation tools. In some instances, multiple runs were averaged to reflect the variability in training the model (especially the case in the neural network).

All the results were saved in CSV format and visualised as comparative bar chart using Matplotlib and Seaborn along with fairness-performance trade-off plot.

## 3.6 Tools and Libraries Used

The following tools were used during implementation:

|  |  |
| --- | --- |
| Tool / Library | Purpose |
| Python 3.11 | Base programming environment |
| Jupyter Notebook | Interactive coding and visualisation |
| scikit-learn | Training baseline classifiers (LR, RF) |
| Keras / TensorFlow | Neural network model + adversarial debiasing |
| Fairlearn | Pre-processing bias mitigation (reweighing) |
| AIF360 | In-processing (adversarial debiasing) and post-processing (equalized odds) |
| NumPy / Pandas | Data preprocessing and transformation |
| Matplotlib / Seaborn | Plotting results and comparisons |

Code was modularised into separate scripts and notebooks for data preparation, model training, mitigation application, and metric evaluation. Folder structure was kept organised by dataset and model type to enable future reproducibility and debugging.

## 3.7 Challenges and Limitations

Implementation process also showed some useful limitations and constraints:

1. **Toolkit Incompatibilities:** The data structures and APIs are different between Fairlearn and AIF360 and therefore it was not easy to integrate all methods in a single structured pipeline. Format bridging required wrappers and converters to be written.
2. **Adversarial Debiasing Overhead:** This was also the most resource-burdening method in that it was necessary to control the use of TensorFlow sessions manually and format data in a more complex manner. The training process also consumed very much time as compared to the other processes.
3. **Constraints on Datasets:** Both the combined COMPAS and German Credit data sets are small and this is a limiting factor to generalizability. In addition, the sensitive features are binary; it may be investigated in the future whether it can be done in the multi-class setting.
4. **Performance-Fairness Trade-offs:** In a number of the cases, enhancing fairness measures was obtained at the expense of accuracy specifically in post-processing. The significance of domain-sensitive definitions of fairness is pointed out.
5. **Time and Scope Shortcuts:** The three mitigation methods and two datasets considered can be attributed to this project given the nature of an MSc project on time constraints. This was good in order to conduct a more thorough comparison but there is still room to deepen the study by adding more models or datasets.

# References

Bagdasaryan, E., Veit, A., Hua, Y., Estrin, D. and Shmatikov, V. (2019). Differential Privacy Has Disparate Impact on Model Accuracy. *Advances in Neural Information Processing Systems (NeurIPS)*. [online] Available at: https://neurips.cc/virtual/2019/poster/14443 [Accessed 30 Jun. 2025].

Cowgill, B. and Tucker, C. (2023). Algorithmic Exclusion: The Fragility of Algorithms to Sparse and Missing Data. *Brookings Institution Working Paper*. [online] Available at: https://www.brookings.edu/wp-content/uploads/2023/02/Algorithmic-exclusion-FINAL.pdf [Accessed 30 Jun. 2025].

Giffen, B.V., Herhausen, D. and Fahse, T. (2022). Overcoming the Pitfalls and Perils of Algorithms: A Classification of ML Biases and Mitigation Methods. *Journal of Business Research*, 144, pp.93–106. [online] Available at: https://www.sciencedirect.com/science/article/pii/S0148296322000023 [Accessed 30 Jun. 2025].

Pagano, T.P. et al. (2023). Bias and Unfairness in ML Models: A Systematic Review. *Big Data and Cognitive Computing*, 7(1), p.15. [online] Available at: https://www.mdpi.com/2504-2289/7/1/15 [Accessed 30 Jun. 2025].

Raji, I.D. and Buolamwini, J. (2023). Actionable Auditing Revisited: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products. *Communications of the ACM*, 66(1), pp.42–49. [online] Available at: https://dl.acm.org/doi/10.1145/3571151 [Accessed 30 Jun. 2025].

Siddique, S. et al. (2024). Survey on Machine Learning Biases and Mitigation Techniques. *Digital*, 4(1), pp.1–68. [online] Available at: https://www.mdpi.com/2673-6470/4/1/1 [Accessed 30 Jun. 2025].

Wadsworth, C., Vera, F. and Piech, C. (2018). Achieving Fairness through Adversarial Learning: An Empirical Analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1). [online] Available at: https://ojs.aaai.org/index.php/AAAI/article/view/11771 [Accessed 30 Jun. 2025].

Zliobaite, I. (2017). Measuring Discrimination in Algorithmic Decision Making. *Data Mining and Knowledge Discovery*, 31, pp.1060–1089. [online] Available at: https://link.springer.com/article/10.1007/s10618-017-0506-1 [Accessed 30 Jun. 2025].

Bellamy, R.K.E. et al. (2019). AI Fairness 360: A Robust Toolkit for Algorithmic Fairness. *IBM Journal of Research and Development*, 63(4/5), pp.4:1–4:15. [online] Available at: https://ieeexplore.ieee.org/document/8851337 [Accessed 30 Jun. 2025].

Bird, S. et al. (2020). Fairlearn: A Toolkit for Assessing and Improving Fairness in AI. *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT)*. [online] Available at: https://dl.acm.org/doi/10.1145/3351095.3372878 [Accessed 30 Jun. 2025].

Lohaus, M., Sterbak, P. and Zopf, M. (2021). Evaluating Post-processing Bias Mitigation with Uncertainty. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15), pp.13257–13265. [online] Available at: https://ojs.aaai.org/index.php/AAAI/article/view/17691 [Accessed 30 Jun. 2025].

Singh, J., Gummadi, K.P. and Weller, A. (2022). Fairness Tradeoffs in Machine Learning: A Quantitative Survey. *Journal of Machine Learning Research*, 23(112), pp.1–45. [online] Available at: https://www.jmlr.org/papers/volume23/21-0342/21-0342.pdf [Accessed 30 Jun. 2025].

Wang, T., Gupta, S. and Arya, V. (2021). Comparative Fairness Analysis of Machine Learning Models. *IEEE Transactions on Technology and Society*, 2(3), pp.123–135. [online] Available at: https://ieeexplore.ieee.org/document/9442263 [Accessed 30 Jun. 2025].

Wexler, J., Pushkarna, M., Bolukbasi, T., Wattenberg, M. and Viegas, F. (2021). The What-If Tool: Interactive Probing of Machine Learning Models. *Distill*. [online] Available at: https://distill.pub/2019/what-if-tool/ [Accessed 30 Jun. 2025].

Zhang, B. and Gong, N.Z. (2020). Fairness-Aware Machine Learning for Credit Risk Assessment. *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT)*. [online] Available at: https://dl.acm.org/doi/10.1145/3351095.3372859 [Accessed 30 Jun. 2025].