Evaluation of Pre-Processing Techniques to Mitigate Bias in Machine Learning

Final report
submitted in partial fulfilment
of the requirements for the degree of
Master of Science in Data Analytics

by
Elton Grivith D'Souza
D00264329

Under the supervision of
Dr. Natalia Budarina

Department of Computing Science & Mathematics
Dundalk Institute of Technology
Louth, Ireland

September 2024

Dedication

To my family and friends, without whom I would not be here.

Acknowledgements

I would like to extend my deepest gratitude to **Dr. Natalia Budarina**, my supervisor, for her invaluable support and guidance throughout the course of this research. Her mentorship has been instrumental in the successful completion of this work.

I am also deeply grateful to the lecturers of the Data Analytics program at Dundalk Institute of Technology: **Dr. Jack McDonnell**, **Dr. Siobhan Connolly Kernan**, **Dr. Abhishek Kaushik**, **Dr. Peadar Grant**, and **Dr. Anesu Nybadza**. Their extensive professional knowledge and dedication in teaching through lectures, laboratories, and projects have greatly enhanced my understanding of data analytics.

I am truly fortunate to be part of the M**Sc Data Analytics class of 2024**. My fellow students have been incredibly supportive, helping me navigate the challenges of relocating to a new country and adapting to both the academic and cultural aspects of this journey.

I would also like to express my appreciation to the **administrative staff** and **student support services** at Dundalk Institute of Technology. Their assistance and guidance made my transition to a new country much smoother and more manageable.

Finally, I wish to thank my parents, **Gerald D'Souza** and **Reena Janet D'Souza**, for their unwavering support, encouragement, and belief in me throughout this journey. Their sacrifices and guidance have been a cornerstone of my success, and I am profoundly grateful for their constant support.

Declaration

““I hereby declare that the work described in this project is, except where otherwise stated, entirely my own work and has not been submitted as part of any degree at this or any other Institute/University”

Signed: 

Name: Elton Grivith D’Souza
Date: August 23, 2024

Abstract

This document investigates the application of various pre-processing techniques such as resampling and reweighing among others, in predictive models. Using the COMPAS dataset, which presents significant bias challenges, the study evaluates how different preprocessing approaches like resampling (SMOTE, RandomUnderSampler, and SMOTEENN), reweighting, and Disparate Impact Remover (DIR), affect fairness and model performance. Logistic regression serves as the baseline model, and the interventions are applied to mitigate bias across protected attributes, specifically sex and age.

Fairness metrics such as statistical parity difference, average odds difference, and equal opportunity difference are employed alongside traditional performance indicators like accuracy and confusion matrices. Each technique's impact on the trade-off between fairness and accuracy is critically assessed. The results demonstrate that reweighing and resampling techniques offer diverse pathways for reducing bias, with mixed effects on predictive performance. Furthermore, the incorporation of DIR highlights the potential for post-processing techniques to further enhance model fairness. These findings contribute to the broader discourse on algorithmic fairness and provide actionable insights for improving equity in machine learning systems, particularly in sensitive applications like criminal justice.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Research on artificial intelligence (AI) and machine learning (ML) has bought about significant advancements over the past few years. ML models have been increasingly used as decision-making entities that have a significant impact on the lives of many people. These models are used independently or, in some cases, in lieu with domain experts, used in sensitive domains such as healthcare, medicine, politics, and even judicial systems.

The rapid integration of ML into these crucial sectors underscores the transformative potential of AI. While these applications of ML models are impressive and noteworthy, if used in such sensitive domains without through evaluation, problems can arise Dressel & Farid (2021). This is more apparent in predictive models where the models can reinforce the bias present in the data. And when these models are deployed without a comprehensive analysis on its behaviour, it can have major impact on peoples lives Janssen & Kuk (2016). Here, the analysis of the models performance plays a vital role, and detection of these errors can help us build better performing and ethically aligned ML models.

Pre-processing is a crucial stage in machine learning, where raw data is transformed into a suitable format for the model. Effective pre-processing can significantly improve model outcomes, including fairness Bellamy, Dey, Hind, Hoffman, Houde, Kannan, Lohia, Martino, Mehta, Mojsilović et al. (2019). Various techniques are used to address bias in data, ensuring the model treats different demographic groups fairly. This is especially important in sensitive applications where decision-making can directly impact individuals' lives, such as in criminal justice or healthcare.

This research focuses on examining how various pre-processing techniques affect fairness metrics in machine learning models. Specifically, this study explores the effectiveness of these techniques on the COMPAS dataset by using fairness metrics such as statistical parity difference and average odds difference among others. The COMPAS dataset, which is widely used in fairness research, in-

volves recidivism predictions, where the fairness of the model's predictions for different demographic groups is of vital importance. The methodology used to evaluate the preprocessing techniques is discussed in further sections of the thesis.

To further provide clarity on this projects goals, we have listed the research goals and the project objectives in Table 1.1.

Research Questions	**Objectives**
How do various preprocessing techniques affect fairness metrics such as statistical parity, average odds, and equal opportunity in predictive models applied to the COMPAS dataset?	To evaluate the effectiveness of different preprocessing techniques in addressing fairness disparities in predictive models using the COMPAS dataset.
What are the impacts of these preprocessing techniques on the overall performance, including accuracy, of logistic regression models across different demographic groups?	To compare the influence of these preprocessing techniques on fairness metrics and overall model performance.
What preprocessing technique individually provides the best balance between fairness and performance in the context of the COMPAS dataset?	To identify preprocessing techniques that achieve an optimal balance between improving fairness and maintaining model accuracy.
	To offer recommendations for selecting appropriate preprocessing techniques to enhance fairness in predictive models used in sensitive applications such as criminal justice.

Table 1.1: Research Questions and Objectives

To implement and assess these objectives, a structured approach involving a range of tools and technologies has been employed. The technologies used in this project are summarized in Table 1.2. The project was developed in Python 3 with some important non-default libraries such as AIF360 among others. To help interested parties execute this code easily, we have provided a python file with all the required libraries. This can be used to obtain the same virtual environment that this project was built upon. Additionally, we also used GitHub to passively store this project and its versions while using overleaf to document findings. Dataspell was used to analyse the dataset, while JupyterLab and VS Code was used to conduct experiments.

Category	**Description**
Programming Language	Python 3
Libraries	Provided as requirements.py
IDE	JupyterLab, VS Code, DataSpell
Version Control	GitHub
Documentation	Overleaf

Table 1.2: Core Technologies used in the Project

Using these tools, this project took approximately three months for completion (Excluding literature). The life cycle of the project, shown in figure 1.1, involves multiple stages that we followed during the development of this project. Definition of the project scope was established by the research questions and key objectives, setting the boundaries of the analysis, and identifying the required resources. Identification of data sources is also crucial. We require data that is ethically sourced, which can be used for the scope of this research.

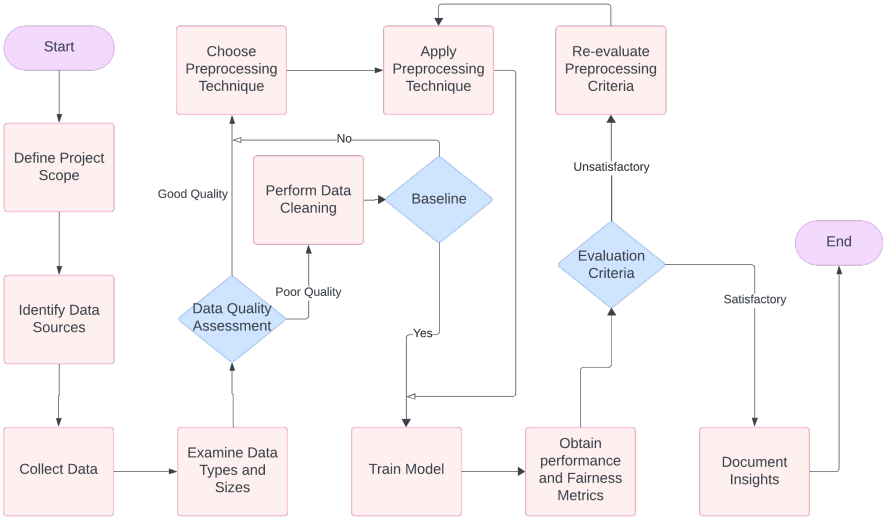

Figure 1.1: Life cycle of the project

After collecting the data and examining it, we assess the data quality to implement require formatting and cleaning procedures. We then set a baseline, gather evaluation metrics and do the same with all the preprocessing techniques. The results and insights are then documented and compiled.

The structure of the report starts off Chapter 2 (State of the Art) that describes the most relevant topics related to the research questions shown in Table 1.1. This is followed by Experimental Setup in Chapter 3. This chapter describes the experiments and its pipelines in detail. Chapter 4 (Results) explains the insights gained from the experiments and Chapter 5 (Conclusion) discusses whether the research questions were addressed and future work.

Chapter 2

State of the Art

The increasing reliance on machine learning algorithms in judicial and law enforcement systems has raised significant ethical concerns, particularly regarding bias and fairness Barenstein (2019). Recent studies have shown that these algorithms can perpetuate and even exacerbate existing biases, leading to unfair treatment of certain demographic groups Chouldechova (2017), Mehrabi, Morstatter, Saxena, Lerman & Galstyan (2021).

The literature on bias detection and mitigation is primarily focused on quantifying a measure of bias and developing algorithms to mitigate it Zandee (2021). In this chapter, we aim to critically evaluate existing studies on preprocessing techniques that address bias detection and mitigation while exploring evaluation metrics and the applicability of these algorithms in real-world scenarios.

2.1 Literature Selection Criteria

This review examines publications by decomposing our research objective and compiling information from the exploration of smaller related problems. The reviewed papers majorly consist of publications in the domain of Artificial Intelligence (AI), Machine Learning (ML) and Ethics. We explored several databases, journals, and proceedings from Elsevier, Springer, IEEE Transactions, arXiv, and key conferences such as NeurIPS and ACL among others. AI tools such as Paper Digest (PD) and Elicit were used to target solutions and explore work done in different domains. Recognising the amount of research done in this field, we have included preprints, independent and seminal works to ensure comprehensive evaluation.

2.2 Theoretical Framework

The theoretical framework for detecting and mitigating bias in data is primarily focused on FAIR principles. Raza, Ghuge, Ding & Pandya (2024) systematically explores these principles and an overview of their research is summarized in Section 2.2.2.

2.2.1 Definitions

This section has summarised some of the key terminologies used in this field of research. This work is largely based on the work of AIF360 Bellamy et al. (2019), Zandee (2021) and Raza et al. (2024).

A **protected attribute** refers to a label or value that is considered sensitive or prone to discrimination. Using these attributes without proper data augmentations could lead to unfair outcomes. Common examples of these attributes are race, gender, and age, among others.

A **privileged group** refers to a subset of the protected attribute that historically or systematically receives preferential treatment over the others. This is often contrasted by groups that are discriminated against called unprivileged groups. Examples of these factors include race, socioeconomic differences, and disability, among others.

Fairness is used to denote impartial outcomes. In the context of this research, fairness is treated as a quantifiable measure that is summarised as **bias metrics**. These metrics are described in further detail in section 3.3.

Baseline refers to a simple model that is used as a point of reference for more complex models. These are straightforward to implement and can provide an accurate view of the efficiency of preprocessing techniques while using the same model. **Ground truth** is a similar concept which describes a real-world situation. Ground truth is beneficial over baseline in certain scenarios as they benchmark the results not just by how well the algorithm mitigates bias, but it also considers how the results change compared to the situation.

2.2.2 FAIR Principles

The FAIR principles is an abbreviation for the attributes of Findability, Accessibility, Interoperability and Reusability. Given the constraints, the definitions for each of these principles is provided below.

1. **Findability**: This principle ensures that data and resources are easily locatable and accessible. To enhance the readability and discoverability of data, Raza et al. (2024) suggests the

use of metadata, persistent identifiers, and promoting efficient data indexing and searchability. It is also important to ensure trustworthiness of the data source to promote findability among different environments.

2. **Accessibility**: This is defined by the ease of obtaining and using the data once located. The application of FAIR principles in the context of accessibility is explored by Lamprecht, Garcia, Kuzak, Martinez, Arcila, Martin Del Pico, Dominguez Del Angel, Van De Sandt, Ison, Martinez et al. (2020). The works emphasise the significance of incorporating accessibility considerations into research software, platform support, and implementation challenges Raza et al. (2024).

3. **Interoperability**: This refers to the ability of different systems to efficiently work togetherSantos, Pinheiro & Maciel (2021). It requires standardised data formats and protocols, enabling easy data exchange and integration across diverse systems Raza et al. (2024).

4. **Reusability**: This is a key component in FAIR data principles that ensures data is stored and documented for future retrieval and reuse Raza et al. (2024). A structured approach with planning FAIRification of data in regards to reusability involves providing rich metadata, legal and ethical considerations, and potential societal impactGonzález-Cebrián, Bradford, Chis & González-Vélez (2024).

2.2.3 Measuring Fairness in ML

This section outlines some of the metrics and methodologies used to measure fairness with decision making systems. Understanding these methodologies helps us identify and mitigate bias that enables us to build an equitable model which follows the FAIR principles.

Discrimination Control

Discrimination is the prejudicial treatment of an individual based on membership in a legally protected group such as a race or gender. Direct discrimination occurs when protected attributes are explicitly used in making decisions, which is known as *disparate treatment* in law du Pin Calmon, Wei, Ramamurthy & Varshney (2017). Quantification and treatment of discrimination on an algorithmic level involves the modification of training data, the learning algorithm, or the decisions itself. These are known as pre-processing, in-processing, and post-processing, respectively. This approach utilises balanced error rates and predictive bias to focus on pre-processing that achieves equitive results.

Based on the work by du Pin Calmon et al. (2017), we arrive at a general formulation as such:

Given a dataset consisting of n i.i.d. samples $\{(D_i, X_i, Y_i)\}_{i=1}^{n}$ from a joint distribution $p_{D,X,Y}$ with domain $D \times X \times Y$. We define the following:

- D: Discriminatory variables such as gender and race.
- X: Non-protected variables used for decision-making.
- Y: Outcome random variable.

Assumptions:

1. D and X are discrete and finite domains.
2. Y is binary, i.e., $Y \in \{0, 1\}$.
3. There are no restrictions on the dimensions of D and X.

For instance, in the context of recidivism prediction using the COMPAS dataset:

- D_i could represent the demographic information (e.g., race, gender) of individual i.
- X_i could represent the prior criminal history and other non-protected attributes of individual i.
- Y_i could represent whether individual i recidivates ($Y_i = 1$) or not ($Y_i = 0$).

The objective is to analyze and model the relationship between (D, X) and Y while considering the potential impacts of discriminatory variables on the outcome.

Equal Opportunity

Equal opportunity is a metric that is used to capture the benefits of model predictions. This is done by using True Positive Rates (TPR) where true positive implies the scenario in which the model's positive decision aligns with the real observation. This can be denoted as:

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

where:

TP : Correctly predicted positive instances,

FN : Positive instances incorrectly predicted as negative.

Using this, Equal opportunity can be defined as a scenario in which

$$TPR_0 = TPR_1$$

In practice, we provide some flexibility for statistical uncertainty.

$$TPR_1 - TPR_0 < \theta$$

where:

$$\theta = \text{Threshold}$$

As a ratio, this can be represented as:

$$\frac{TPR_0}{TPR_1} > \theta$$

A statistical representation of this principle is written below:

$$P(Y = 1|A = a, \hat{Y} = t) = P(Y = 1|A = b, \hat{Y} = t)$$

$$\forall a, b \in \text{Dom}(A), \forall t \in \text{Range}(\hat{Y}),$$

where:

Y : Binary outcome variable indicating the occurrence of an event of interest,

A : Protected attribute, such as race or gender,

$\hat{Y}$: Predicted outcome from a predictive model.

Equalised Odds

Unlike equal opportunity, Equalised Odds uses False Positive Rates (FPR). In our study this can be used to represent the number of individuals that do no re-offend but are categorized as potential re-offenders. FPR is denoted as:

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}}$$

where:

FP : Negative instances incorrectly predicted as positive,

TN : Negative instances correctly predicted as negative.

Similar to Equal Opportunity, Equalised Odds can be denoted as:

$$TPR_0 = TPR_1$$

$$FPR_0 = FPR_1$$

Through this definition, we can see that Equalised Odds is a stricter definition is compared to Equal Opportunity. A statistical version of this definition is provided as a reference Bellamy et al. (2019), Feldman, Friedler, Moeller, Scheidegger & Venkatasubramanian (2015), Hardt, Price & Srebro (2016):

$$P(\hat{Y} = 1|Y = 1, A = a) = P(\hat{Y} = 1|Y = 1, A = b)$$

$$P(\hat{Y} = 0|Y = 0, A = a) = P(\hat{Y} = 0|Y = 0, A = b)$$

$$\forall a, b \in \text{Dom}(A), \forall t \in \text{Range}(\hat{Y}),$$

where:

Y : Binary outcome variable indicating the occurrence of an event of interest,

A : Protected attribute, such as race or gender,

$\hat{Y}$: Predicted outcome from a predictive model.

Disparate Impact

Disparate impact is addressed by the principles of statistical parity and group fairness du Pin Calmon et al. (2017). Also known as adversial impact, it occurs when a neutral policy disproportionately affects a group of observations Feldman et al. (2015). While disparate treatment addresses intentional discrimination, disparate impact occurs unintentionally. This metric is a result derived from the PPP (predicted as positive) rates which can be denoted as:

$$\%\text{Predicted as Positive (PPP)} = \frac{TP + FP}{N}$$

where:

$$TP : \text{True Positive},$$

$$FP : \text{False Positive},$$

$$N : \text{Number of Observations}.$$

The definition of a fair model given this PPP value is:

$$PPP_0 = PPP_1$$

Again, adjusting to statistical uncertainty,

$$\frac{PPP_0}{PPP_1} > \theta$$

In the United States, there is a legal precedent to set θ at 80% Odyssey (2023). Hence we will be using the same.

2.3 Ethical Considerations

To understand and portray the ethical implications of the research, we conducted a SWOT(Strengths, Weaknesses, Opportunities, and Threats) Analysis. This enables us to evaluate our research and determine stakeholders. Through this analysis, we aim to explore and understand the ethical landscape of the legal system that can help build towards responsible and principled decision making.

1. **Strengths**:

    - *Focus on pre-processing*: It is important to address model performance at the data level. Doing this allows researchers to address issues apparent in the data before constructing and evaluating models Chouldechova (2017). The COMPAS dataset has many issues with the raw data that ProPublica had not addressed. This makes their evaluation of the model unfit for a comparative analysis Hoffmann (2019). Issues such as these must be addressed before using model evaluation and performance strategies.
    - *Fair Decision Making*: Implementing bias mitigation strategies can lead to fairer decision-making systems, transcending individual biases and social hierarchies within the COMPAS dataset Corbett-Davies, Pierson, Feller, Goel & Huq (2017). A successful implementation of a fair model would lead to truly equitable outcomes.

2. **Weaknesses**:

- *Partial Mitigation*: While pre-processing techniques are beneficial, it can only partially address the issues within the dataset and cannot be considered as standalone solutions Chouldechova (2017). This is an inherent limitation of all research regarding preprocessing algorithms.
- *Potential for introducing new bias*: Hardt et al. (2016) elucidates upon the risk of introducing new bias within the data while trying to mitigate existing ones. Concerns that algorithms may discriminate against certain groups have led to numerous efforts to 'blind' the algorithm to race Kleinberg, Ludwig, Mullainathan & Rambachan (2018). It is crucial to have a through understanding about the research domain and methodologies used to make sure that the algorithms do not introduce non-existent bias.

3. **Opportunities**:

   - *Advancement of ML models*: The successful implementation of bias mitigation strategies can advance the performance and applicability of ML models Kleinberg et al. (2018). Our research can provide insight and new opportunities to develop and improve upon existing ML applications in the criminal justice domain.
   - *Ethical Decision Making*: The use of ethically aligned algorithms in the criminal justice system, particularly when applied to the COMPAS dataset, can potentially improve fairness, accountability, and transparency Chiao (2019). This provides opportunities to address real-world challenges with better efficiency.

4. **Threats**:

   - *Regulatory Challenges*: The lack of regulatory frameworks that defines fairness and the acceptable measures of standard errors poses significant challenges in defining and maintaining ethically fair models Chiao (2019), Hardt et al. (2016). This lack of a standard definition for fairness makes it hard for us to conduct and justify a comparative analysis with other research works.
   - *Ethical Ambiguity*: The lack of standard regulations and guidance in volatile fields of real-world impact makes it hard for organisations to adopt a new effective approach regardless of the results producedChouldechova (2017).
   - *Change of data over time*: Valavi, Hestness, Ardalani & Iansiti (2022) reviews the effects of time over the validity of data. The COMPAS dataset comprises of information from 2013 and 2014. There might be significant changes in behavioural mechanics over time.

Chapter 3

Experimental Setup

This chapter describes the experimental setup used for evaluating various pre-processing techniques using the COMPAS dataset. The chapter is divided into several sections derived from key steps of our research design.

3.1 Scope of Study

This study is focused on evaluating the impact of reweighing, resampling and disparate impact removal as preprocessing techniques for the removal of bias. It is evaluated through the use of the COMPAS dataset. This has been clearly listed below:

- **Dataset**: Utilizes the COMPAS dataset, which includes data on individuals' criminal history and demographic information.
- **Techniques**: Assesses three types of preprocessing techniques that include:
    1. **Reweighing** using the AIF360 toolkit
    2. **Resampling** strategies where
        - SMOTE is used for over-sampling.
        - RandomUnderSampler is used for under-sampling
        - SMOTE + ENN is used for a combination of Oversampling and undersampling.
    3. **Disparate Impact Remover** using AIF360 toolkit
- **Metrics**: Considers fairness metrics such as statistical parity difference, average odds difference, and equal opportunity difference, and disparate impact, along with performance metrics like accuracy and confusion matrix.
- **Protected Attributes**: Examines fairness disparities based on sex and age

- **Model**: All experiments use logistic regression as baseline and for comparison.

This study does not cover other preprocessing techniques beyond the ones mentioned above or other datasets and models outside the COMPAS and logistic regression.

3.2 Research Design

The research design serves as a blueprint for planning and executing the study. In this section, we outline our completed approach to conducting research.

We start the process by collecting and formatting our data. This process is thoroughly explained in Section 3.3 and Section 3.4. After the data formatting process, we conduct our experiment on two different protected attributes race and sex.

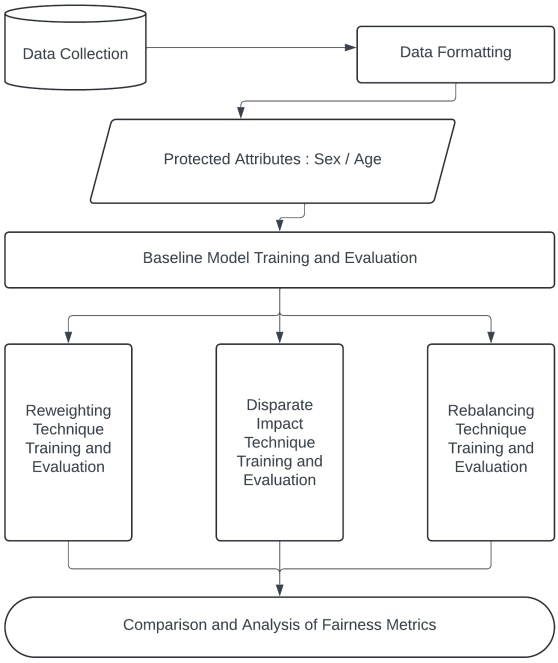

Figure 3.1: Research Design

Following the definition of the protected attribute, We define the baseline by training this raw data on a logistic regression model. We chose this model as Barenstein (2019) mentioned that a logistic regression model fitted on the data with the attributes that we have chosen, provides similar results to the actual model. This model is then evaluated for fairness and performance metrics. We refer to this model as the baseline model and the metrics as baseline metrics. It serves as a reference for future comparison.

We then apply different preprocessing techniques such as reweighing, resampling and disparate impact removal to gain similar metrics as the baseline model. The exact procedure of execution is elaborated in further sections. The obtained metrics are then compared against the baseline metrics to evaluate different preprocessing techniques and gain insight into their behavior. This research design is illustrated in Figure 3.1.

3.3 Data Collection

The selection of data needed several considerations. This comprised of using or obtaining data that was proven to be biased while consisting of qualities that could be addressed by this research.

The Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) dataset is a comprehensive repository of information that is used by judges and parole officers in the United States of America to predict a criminal defendant's likelihood of re-offending (recidivism). This data consists of extensive details such as demographics, prison and jail time, type and details of the offense among others.

Through a public records request, ProPublica obtained two years worth of COMPAS scores from the Broward County Sheriff's Office in Florida Larson, Mattu, Kirchner & Angwin (2016). This data consists of all the criminal defendants processed through the COMPAS framework during 2013 and 2014. This data was then made open source, and we were able to obtain a version of it through Kaggle.

The choice of data is justified for our research for several reasons:

1. **Widespread Usage**: Automations withing the criminal justice system can have significant real-world impact. Addressing these issues can lead to fair decision making that can improve and obtain equitable outcomes for varied demographic groups Dressel & Farid (2021).

2. **Existing concerns of bias**: Larson et al. (2016) have investigated the data to provide substantial evidence regarding the bias that exists within the data. By focusing on reducing these racial disparities, researchers can help contribute to fairer justice administration.

3. **Comprehensive Data**: The richness of data on criminal defendants allows us to explore multiple pre-processing techniques in various dimensions Mehrabi et al. (2021) such as racial and gender bias, among others.

Variable Name	Variable Type	Description
Sex	Binary	M: Male, F: Female
Race	Binary	Race of inmate: Caucasian or Black
age_cat	Categorical	Categorical representation of inmate's age: • Less than 25 • 25-45 • Greater than 45
priors_count	Numeric	Number of prior offenses
c_charge_degree	Categorical	Charge degree: felony (F) or misdemeanor (M)
two_year_recid	Binary	Recidivism within two years: 1 or 0

Table 3.1: Description of COMPAS dataset variables used in the analysis.

By decomposing our problem statement, we chose a core data subset from the original version provided by ProPublica. This provisional decision allows us to focus on the core features that lead to the decision outcome. The data card for our subset is shown in Table 3.1.

3.4 Data formatting

AIF360, a comprehensive toolkit for fairness analysis, provides various examples in their documentation using the COMPAS dataset. This was vital in procuring and cleaning the core dataset. The original data gathered from our data source was loaded into '`compas.datasets`'. The `load_preproc_data_compas` function was then used to gather the core subset. The resultant subset is mentioned in the data card provided in Table 3.1.

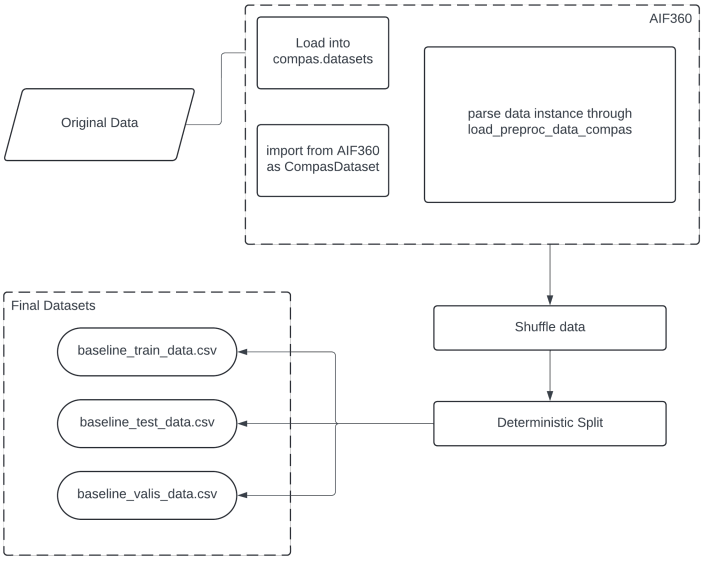

Figure 3.2: Data formatting process pipeline

To ensure the precision and reliability of the results, we followed a deterministic approach to splitting the data. This ensures that each run of the experiment provides the same results and that it is not affected by random influences such as data shuffles and stratification. To further reduce errors

that occur from the underlying patterns in the original dataset, we shuffled our data before the deterministic split. The results were stored in separate CSV files called `baseline_train_data.csv`, `baseline_test_data.csv` and `baseline_valis_data.csv`. The process has been visualized in Figure 3.2.

3.5 Exploratory Analysis

The formatted data consists of 5278 observations and 6 columns.Refer Table 3.1 for feature details. We have chosen to present the formatted data as this closely aligns to the experiment. An initial assessment shows the following insights:

Metric	**Sex**	**Race**	**two_year_recid**
Mean	0.195339	0.398446	0.470443
Standard Deviation	0.396499	0.489625	0.499173
Count	5278	5278	5278

Table 3.2: Mean, standard deviation and Count of numerical columns

- **Mean**: The mean value for the **sex** variable is 0.195, indicating that approximately 19.5% of the observations are coded as '1'. Since it represents females, this suggests a significantly lower proportion of females compared to males in the dataset. The mean of the **race** variable is 0.398, which shows that approximately 40% of the individuals belong to the race coded as '1'. In this case, 1 represents African-American. This data has 60% of Caucasian samples. The detailed discussions of distributions is written in further sections. The mean for recidivism is at 47% which suggests nearly equal distribution of labels.
- **Standard deviation**: Since the features are mostly binary, this value signifies the distribution more than variation. The observed standard deviations for race, sex and two_year_recid is 0.489, 0.396 and 0.499 respectively. This supports our observations and insights from the mean value.

Table 3.2 summarises the values seen above. Since most of our attributes are binary/categorical in nature, visualisations would make it more convenient to gather insights.

From the graphs in Figure 3.3, we can observe significant class imbalance between privileged and unprivileged groups in the attributes sex and race. This coincides with our findings from the mean and standard deviation and shows signs that could lead to biased outcomes. There is a slight difference in group sizes for two_year_recid, but it is relatively balanced.

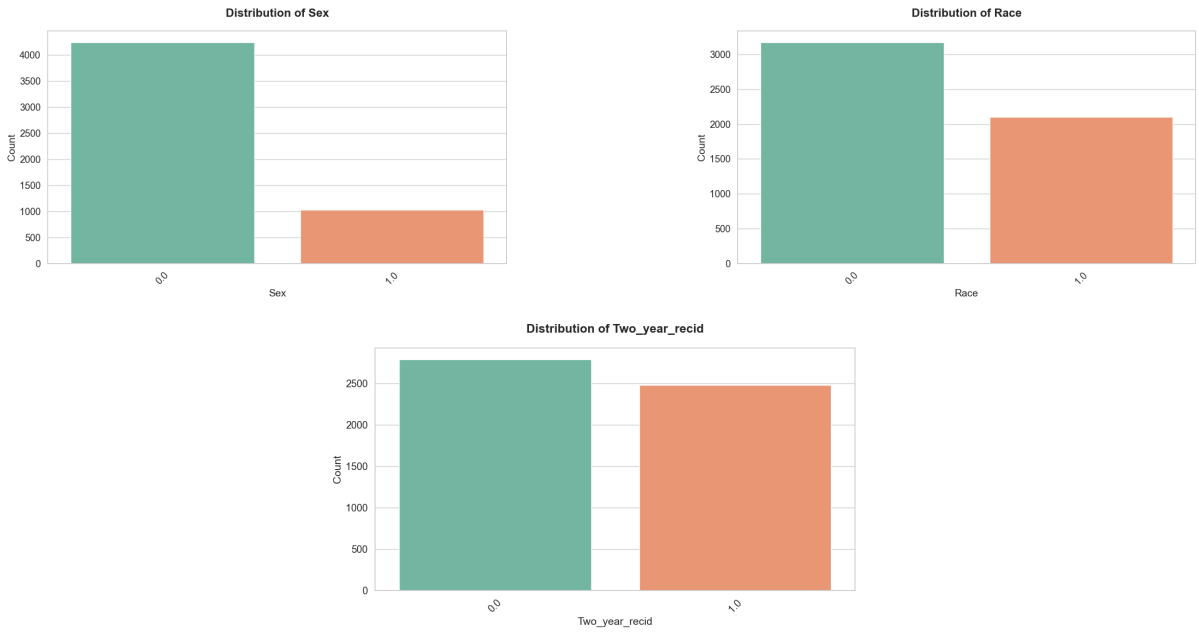

Figure 3.3: Distributions of race, sex, and two-year recidivism. The top left plot shows the distribution of *Sex* (0: Female, 1: Male). The top right plot illustrates the distribution of *Race* (0: Caucasian, 1: African-American). The bottom plot displays the distribution of *Two-Year Recidivism* (0: Non-recidivist, 1: Recidivist).

Following the graphs of age categories, charge degree and priors count in Figure 3.4, we see that most of the samples belong to the age category of 25-45. the other two age categories have similar counts. The charge degree specifies the type of crime. Here, F stands for Felony and M for Misdemeanor. Observing the distributions, we can conclude that there is an imbalance between the two categories. we have approximately two times the number of samples for felony as compared to misdemeanor. The distribution of priors count is relatively balanced.

Figure 3.5 analyses the distributions of our protected attributes (race, sex) against the target feature (two_year_recid). We can observe that in general, most samples fall into the '*No recidivism*' category. This might imply the effectiveness of rehabilitation measures.

Also, the imbalance between groups is more pronounced in the unprivileged groups. This imbalance might play a crucial role in amplifying the bias inherent in the dataset. Form our literature, it is apparent that imbalances in the representation of different groups can exacerbate biased outcomes, especially when using machine learning models. This imbalance introduces potential risks of the model systematically disadvantaging unprivileged groups, further entrenching existing inequalities.

An odd trend is that we have more samples recidivists in the privileged race group. Understanding such unexpected observations is important to accurately interpret the fairness implications of the dataset and ensuring that any deployed model does not inadvertently propagate bias.

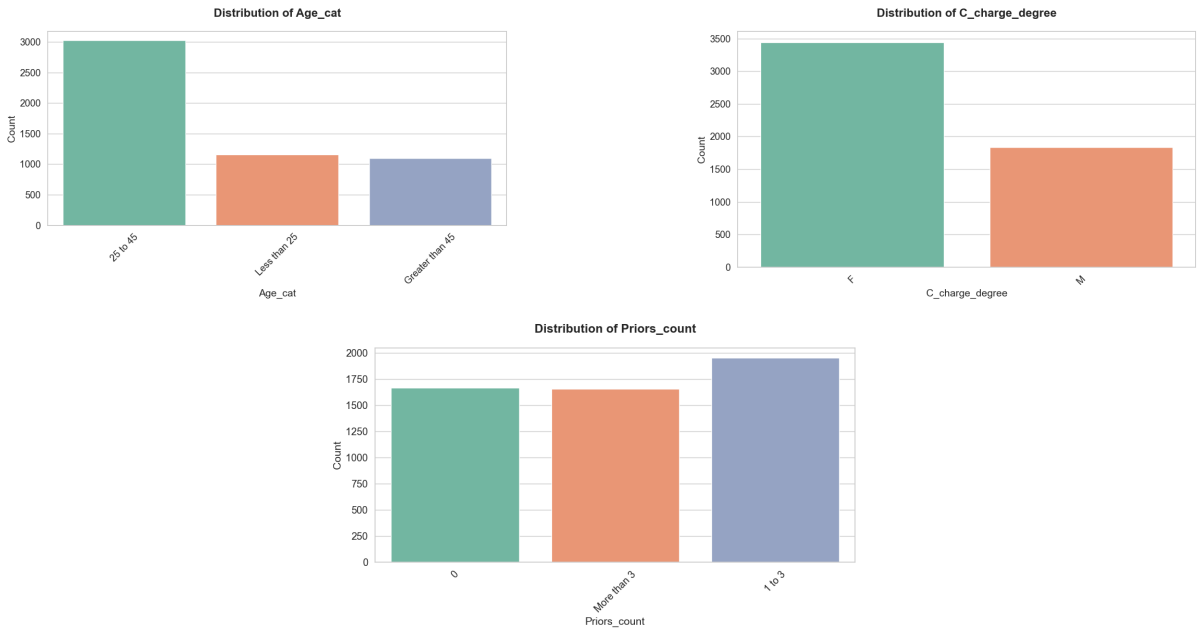

Figure 3.4: Distributions of age categories, charge degree and priors count.

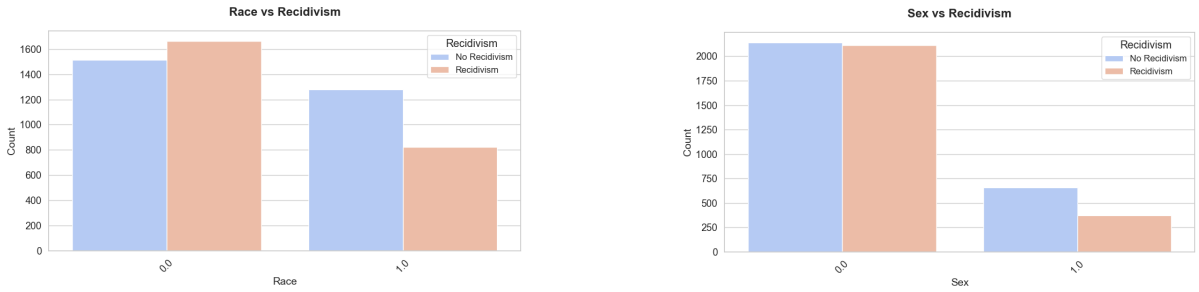

Figure 3.5: Attributes Race and Sex plotted against target two_year_recid.

3.6 Baseline Setup

A baseline is a reference point that can be used for comparison purposes. It represents a starting point against which changes can be compared and measured. We have considered two factors of measurements for this:

1. **Baseline Model**: The baseline model is the foundational algorithm against which all other modifications will be assessed. In this case, we employ a logistic regression model that is trained directly on `baseline_train_data.csv` with no amputations or changes.

2. **Baseline Measurements**: The initial measurements taken from the baseline include:

   - **Accuracy**: Accuracy of the model on `baseline_test_data.csv`. This is a general indicator of the models performance.
   - **Confusion Matrix**: The confusion matrix provides a detailed breakdown of the model's performance by presenting the values for True Positives (TP), True Negatives (TN), False

Positives (FP), and False Negatives (FN). These values are important to understand the model's prediction capabilities and error distribution.

- **Statistical Parity Difference**: This is the difference in the probability of prediction between the two groups.
- **Average Odds Difference**: This is the average of the absolute difference in the false positive rate and the true positive rate for the monitored and reference groups.
- **Equal Opportunity Difference**: It is a fairness metric used to assess whether a model is predicting the desirable outcome equally well for all values of a sensitive attribute.
- **Disparate Impact**: This measure quantifies the deviation from statistical parity in the predictions.

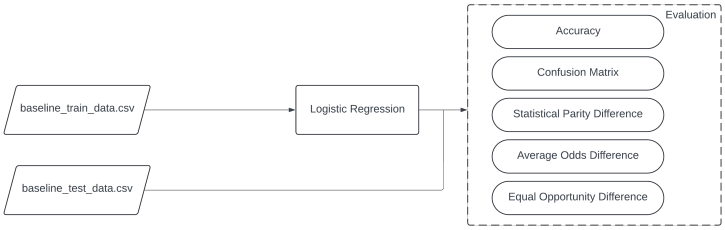

Figure 3.6: Baseline Setup Pipeline

By evaluating how the baseline performs, we can quantify the effectiveness of different pre-processing techniques. Understanding the effectiveness can help us determine if the changes being made to the data is bringing meaningful improvements. This also helps us maintain consistency and reproducibility across different experiments. A visualisation of this process is shown in figure 3.6.

3.7 Reweighing Setup

Reweighing is used to address bias in the model by adjusting the influence of different samples based on their protected attributes. Each `(group, label)` combination is weighed differently to ensure fairness before classification.

In our experiment, we used AIF360 library's Reweighing function to achieve this. Configuration of this process is done by defining the privileged and unprivileged groups based on the protected attribute.

```
reweigher = Reweighing(
    unprivileged_groups = unprivileged_groups,
    privileged_groups = privileged_groups
)
```

In this setup, unprivileged_groups and privileged_groups are lists of dictionaries specifying which groups are considered privileged and unprivileged, respectively. This classification helps in applying appropriate weights to balance the impact of different groups. This transformation is then applied to the dataset by using the method `fit_transform`. This method calculates the weights for each sample of the dataset and produces a new dataset with the applied weights.

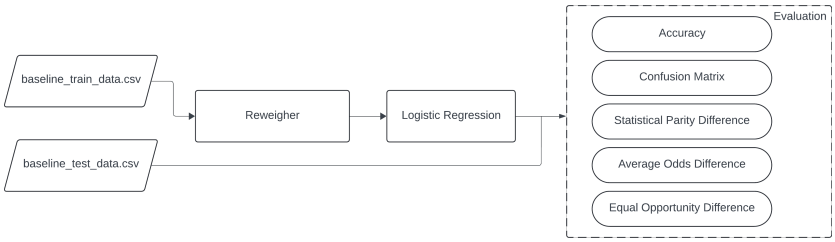

Figure 3.7: Reweighing Setup Pipeline

The reweighed data is then used to train a logistic regression model to derive similar metrics as the baseline setup which includes accuracy, confusion matrix, and fairness metrics such as statistical parity difference, average odds difference, and equal opportunity difference. This pipeline visualisation is illustrated in figure 3.7.

3.8 Resampling setup

Resampling techniques are use to address class imbalances and improve representative fairness for machine learning models. In this setup, we have employed and evaluated three different strategies that include SMOTE, RandomUnderSampler and SMOTE + ENN. This section describes how each strategy works and elucidates upon the process pipeline that we have used.

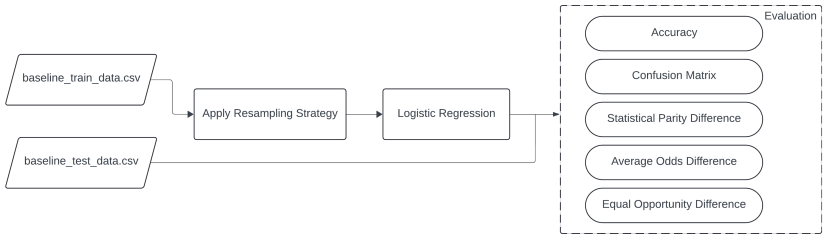

Figure 3.8: Resampling Setup Pipeline

Brief explanations of each strategy is provided below:

1. **SMOTE (Synthetic Minority Over-sampling Technique)**: This algorithm is designed to address class imbalance by creating synthetic samples for the minority class. For each instance of the minority class, SMOTE identifies its k-nearest neighbours. This information is then used to generate new instances by interpolating between them SMO (n.d.). By generating these

instances, SMOTE helps reduce the class imbalance of the minority class that can potentially improve the models performance.

2. **RandomUnderSampler**: This algorithm helps address class imbalance by under-sampling the majority class. It is done by randomly removing a subset of samples from the majority class till the distribution of classes is approximately equal. While this algorithm helps balance the classes, it might reduce the model's reliability as the removed samples might contain important variations or patterns.

3. **SMOTE + ENN (Edited Nearest Neighbors)**: This approach combines oversampling and under-sampling approaches to improve the data quality and class balance. SMOTE is used to generate synthetic samples as described before and ENN is used to clean the dataset by removing noisy data by using nearest neighbors. This method helps attain a clean and representative dataset.

The approach is similar to baseline setup discussed in section 3.6. With the sampling strategies defined, we call a function `evaluate_fairness`. This function uses the defined sampling strategy and fits a logistic regression model to that data. This is then evaluated for the same metrics as previously mentioned. Figure 3.8 illustrates this pipeline.

3.9 Disparate Impact Remover Setup

Disparate Impact Remover (DIR) is a preprocessing technique that is used to mitigate bias. This algorithm works by analysing the distribution of positive outcomes among the privileged and unprivileged groups and tries to equalise these outcomes. Ideally, the proportion of outcomes of unprivileged group should be the same as that of the privileged group.

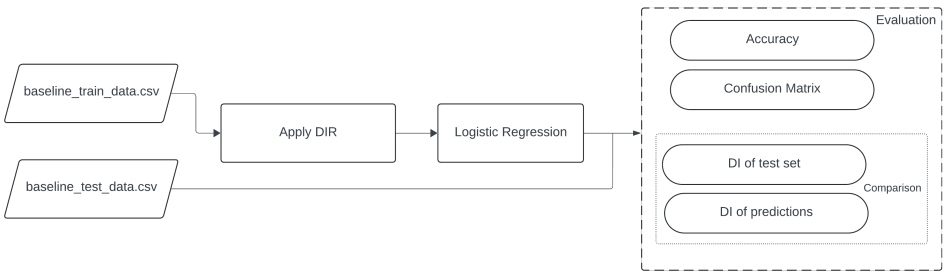

Figure 3.9: Disparate Impact Remover Setup Pipeline

In our experiment, we begin the process by calculating the disparate impact (DI) in the training dataset. This measure is quantified by dividing the proportion of positive outcomes for the unprivileged group by that of the privileged group. A DI value close to 1 indicates that the proportions are

nearly equal, signifying a lack of bias. This provides a comparative reference for DIR implementation.

Following this assessment, We apply DIR to the training set. DIR adjusts the features of the dataset to mitigate the detected bias while preserving as much of the original data distribution as possible. The amount of changes done can be modified using the parameter `repair_level`. This parameter is a floating point variable ranging from 0 (no repair) to 1 (aggressive repair). We used a repair level of 0.5 to maintain data integrity while reducing bias.

The transformed data, like in all other experiments is then used to train a logistic regression model. Evaluation is conducted by comparing the disparate impact of the test set to that of the model predictions. This comparison allows us to assess whether the application of DIR has effectively reduced bias in the model's predictions. We have also included the accuracy and confusion matrix of the model to provide a comprehensive view of the models performance. This process is illustrated in figure 3.9.

Chapter 4

Results

This section presents the findings from the various experimental setups discussed in chapter 3. In summary, we have discussed the results of the baseline model and metrics while comparing them to various pre-processing techniques such as reweighing, resampling and disparate impact remover (DIR). The section is organised to highlight key metrics such as accuracy, Statistical Parity difference (SPD), Equal Opportunity Difference (EOD), Average Odds Difference (AOD) among others. This analysis aims to comprehensively understand each individual preprocessing technique in regards to its effectiveness in reducing bias across two different protected attributes i.e. sex and race.

4.1 Baseline Results

The baseline results, as shown in Table 4.1, provides a snapshot of the models performance across the protected attributes sex and race. The accuracy obtained is 65.02% which while less suggests the need for improvement particularly in regards to fairness. The accuracy remain the same across the two models as the training set and approach remains the same for both models. Only the protected attributes and calculation of fairness metrics differ between the two experiments.

Metric	**Race**	**Sex**
Baseline Accuracy	65.02%	65.02%
Statistical Parity Difference	0.3214	0.2378
Equal Opportunity Difference	0.355	0.22
Average Odds Difference	0.2829	0.224

Table 4.1: Summary of the Baseline Results

Insights from the fairness metrics are described below:

1. **Statistical Parity Difference (SPD)**: The SPD value of 0.3214 (race) and 0.2378 (sex) indicates notable disparity of assignment rates of positive outcomes for the privileged and unprivileged groups. In an ideal scenario, this value would be exactly zero. So the SPD values of our baseline suggests significant bias especially in the case of race.

2. **Equal Opportunity Difference (EOD)**: The EOD for race (0.355) and sex (0.22) shows imbalance in true positive rates (TPR) for the unprivileged groups. This indicates that the model struggles in correctly predicting positive outcomes for the unprivileged groups of race more than sex. It also shows the presence of fairness issues that needs to be addressed through bias mitigation techniques.

3. **Average Odds Difference (AOD)**: The AOD values of 0.2829 (race) and 0.224 (sex) show that on average, the False Positive Rates (FPR) and False Negative Rates (FNR) differ between the privileged and unprivileged groups. The scale of these differences portray that the model shows signs of fairness-related challenges that impact the quality of its predictions.

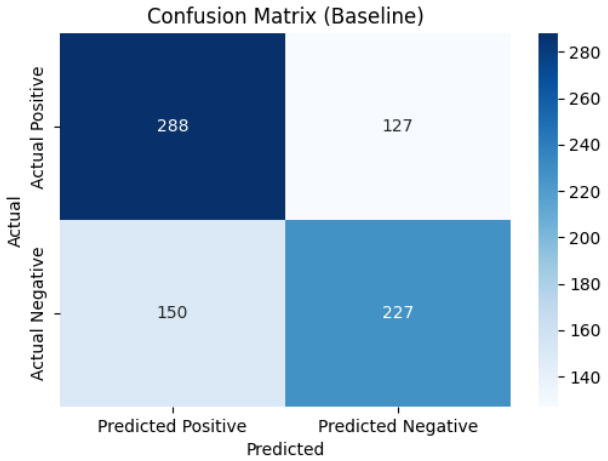

Figure 4.1: Confusion matrix of the baseline logistic regression model

To further understand our models behaviour, we have also included the confusion matrix shown in Figure 4.1. This confusion matrix remains the same across our different baseline experiments as the data that the model is trained on and the model training strategy remains the same. From figure 4.1, we can infer that the baseline model has correctly classified 288 samples as positive (True Positive) and 227 samples as negative (True negative). The model has also misclassified 127 samples as positive (False Positive) and 150 as negative (False Negative).

These values of the confusion matrix complement or findings from Table 4.1. The relatively high False Positive Rate (FPR) and False Negative Rate (FNR) suggest that the model faces challenges in maintaining fairness across both protected attributes. With a precision of 69.4%, the model shows that it is relatively effective in predicting positive samples but could still reduce false positives. The FPR is approximately 35.9%. This could be an issue, as errors in FP is costly in this domain.

4.2 Reweighing Results

The reweighing method, as shown in Table 4.2, shows a substantial improvement over the baseline results. This method balances the weights of different groups within the protected attribute without changing the underlying data. This helps to relatively maintain the prediction capabilities while addressing fairness concerns.

Protected Attribute	acc_bf	acc_af	spd_bf	spd_af	eo_bf	eo_af	ao_bf	ao_af
Race	65.02%	63%	0.3213	0.0669	0.355	0.0985	0.2829	0.0265
Sex	65.02%	65.15%	0.2377	0.0594	0.22	0.0852	0.224	0.0495

Table 4.2: Comparison of reweighing and baseline Results.

Insights derived from Table 4.2 are described below:

1. **Accuracy**: The accuracy of the models decreased slightly for the protected attribute of race (from 65. 02% to 63%), but it showed a slight improvement for sex (from 65.02% to 65.15%). The reweighing algorithm helped us balance accuracy with fairness, which led to slight trade-offs in performance for one attribute while improving others.

2. **Statistical Parity Difference (SPD)**: The SPD values have improved significantly for both race and sex. The protected attribute race has shown slightly better improvements over sex with a difference of 0.2544 and the SPD value of sex has decreased by 0.1783. These results indicate that the reweighing treatment can reduce bias in our data by equalizing the rate of positive outcomes between the privileged and unprivileged groups, indicated by the values moving closer to zero.

3. **Equal Opportunity Difference (EOD)**: The EOD values have also seen a significant improvement. EOD of race has decreased from 0.355 to 0.0985, and for sex, from 0.22 to 0.0852. This shows that reweighing improves the model's ability to correctly classify the positive outcomes for the unprivileged groups. It has helped to reduce differences in true positive rates (TPR).

4. **Average Odds Difference (AOD)**: The protected attribute race has shown an AOD drop of 0.2564 from the original 0.2829, and the AOD of sex has decreased by 0.1745 from 0.224 to 0.0495. Reweighing helps to balance the error rates between privileged and unprivileged groups.

In addition to the fairness metrics, we have also presented the confusion matrices for both race and sex in Figure 4.2. This provides deeper insight into the classification performance.

For the protected attribute race, the confusion matrix shows that the model has correctly classified 276 samples as positive (True Positive) showing a slight decrease from the baseline of 288. 223

samples were classified as negative (True negative) which is similar to the baseline. The model has also misclassified 139 samples as positive (False Positive) and 154 as negative (False Negative).

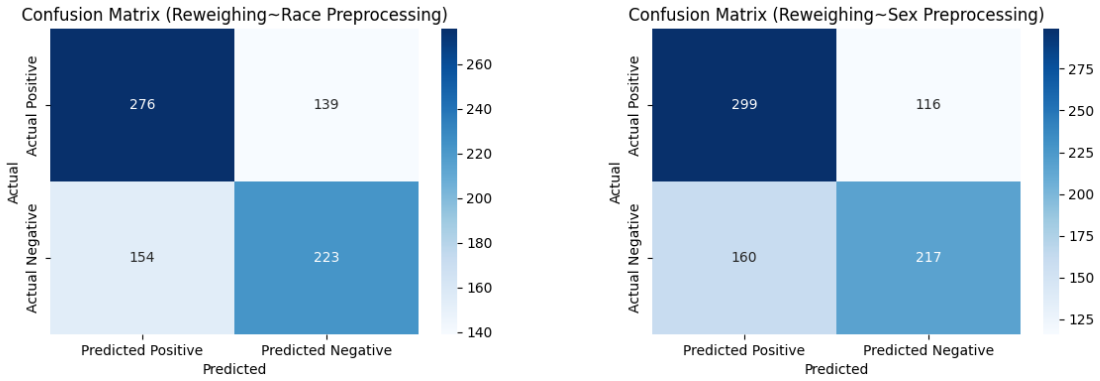

Figure 4.2: Confusion matrices of the reweighing approach for the protected attributes of race and sex

For the protected attribute sex, the confusion matrix shows that the model has correctly classified 299 samples as positive (True Positive) showing an improvement over the baseline of 288. 217 samples were classified as negative (True negative) which is less that the baseline's 227. The model has also misclassified 116 samples as positive (False Positive) and 160 as negative (False Negative).The increase in TP shows that reweighing improves the model's performance for sex without a significant increase in false negatives.

In summary, the reweighing technique shows significant decrease in bias while keeping the metrics relatively stable. These results indicate that reweighing is an effective technique for mitigating fairness issues in our model, and this is supported by the decrease in the values of fairness metrics shown in Table 4.2.

4.3 Resampling Results

SMOTE (Synthetic Minority Over-sampling Technique), RandomUnderSampler, and SMOTE + ENN were used as resampling strategies to evaluate the effectiveness of resampling as a preprocessing techniques to mitigate bias. Sampling strategies are commonly used to rebalance the dataset and improve the representation of the dataset. The workings of each strategies has been described in Section 3.8 and the results are summarized in Table 4.3. Below, we analyze the results across race and sex for each resampling technique.

4.3.1 Accuracy Comparison Before and After Resampling

A models' accuracy refers to the degree of closeness between the predictions made by a model and the actual values. With our previously observed accuracy values of the baseline model (65.02%), the following patterns were observed after applying various resampling strategies:

- **SMOTE**: The accuracy after using the SMOTE technique remains the same as the baseline accuracy at 65.02%. This shows that while the data was rebalanced to improve fairness, the models performance did not degrade. But since our baseline models performance was moderate at best, this result might not be favourable. Ideally, we seek to improve fairness while also improving model performance. But, this is beyond the scope of our research.
- **RandomUnderSampler**: After using this strategy, the accuracy slightly decreased to 64.65%. Under-sampling is known to cause the loss of useful data so this change was expected. We can see that the model maintains its ability to generalise even with some loss of data while balancing the labels.
- **SMOTE + ENN**: This combined approach resulted in the largest decrease in accuracy. With the accuracy dropping to 63.51%, it indicates that while the model's fairness improved (as seen in later metrics), the trade-off was a slight reduction in overall predictive power. SMOTE + ENN combines the approach of over-sampling and under-sampling. The removal of noise through ENN could account for the decrease of the models' performance in accuracy especially if relevant data points was removed along with noise.

4.3.2 Statistical Parity Difference Before and After Resampling

Statistical Parity Difference (SPD) is a measure of the difference in the probability of favourable outcomes among privileged and unprivileged groups. Ideally, the value should be exactly zero which indicates that there is no disparity between the groups.

- **Baseline**: The baseline SPD for race was 0.3213, and for sex, the observed value was 0.2377. these values indicate significant bias in the models predictions. This was more apparent in race where the disparity between the privileged and unprivileged groups was more pronounced.
- **SMOTE**: With the application of SMOTE, SPD increased for both sex (0.3279) and race (0.4430).This might have occurred due to SMOTE balancing the distribution by creating synthetic samples of the privileged groups along with unprivileged groups. This could potentially further amplify the structural bias in the dataset.

Metric	Technique	Race	Sex
acc_bf	-	65.02%	65.02%
acc_af	SMOTE	65.02%	65.02%
	RandomUnderSampler	64.65%	64.65%
	SMOTE + ENN	63.51%	63.51%
spd_bf	-	0.3213	0.2377
spd_af	SMOTE	0.4430	0.3279
	RandomUnderSampler	0.4124	0.3423
	SMOTE + ENN	0.1820	0.1736
eo_bf	-	0.355	0.22
eo_af	SMOTE	0.4750	0.3286
	RandomUnderSampler	0.4456	0.33
	SMOTE + ENN	0.2169	0.18
ao_bf	-	0.2829	0.224
ao_af	SMOTE	0.4072	0.3157
	RandomUnderSampler	0.3769	0.33
	SMOTE + ENN	0.1459	0.1636

Table 4.3: Comparison of Resampling Techniques with Baseline. Abbreviations: **acc_bf** - accuracy before resampling, **acc_af** - accuracy after resampling, **spd_bf** - statistical parity difference before resampling, **spd_af** - statistical parity difference after resampling, **eo_bf** - equal opportunity difference before resampling, **eo_af** - equal opportunity difference after resampling, **ao_bf** - average odds difference before resampling, **ao_af** - average odds difference after resampling.

- **RandomUnderSampler**: With the RandomUnderSampler strategy, we see slight improvements over smote with the SPD of race being 0.4124 and 0.3423 for sex. This is still performing worse when compared to the baseline. A probable cause for this could be the loss of random samples from the unprivileged group. These results indicate that this strategy is insufficient by itself.

- **SMOTE + ENN**: This combined strategy is the only one that was successful in reducing the statistical parity difference. Observed values include 0.1820 for race and 0.1736 for sex. The drop of difference compared to baseline is 0.1393 and 0.0641 respectively. This indicates that the strategy is comparatively more effective by not only generating more samples, but by also removing noisy and redundant samples that could reinforce bias.

4.3.3 Equal Opportunity Difference Before and After Resampling

Equal Opportunity Difference (EOD) focuses on the true positive rate (TPR) between the privileged and unprivileged groups. A lower EOD indicates that the model is performing similarly well for both groups in identifying true positive outcomes.

- **Baseline**: The baseline EOD was 0.355 for race and 0.22 for sex, indicating that the model was particularly biased toward race, with significant differences in the true positive rates between the privileged and unprivileged groups.

- **SMOTE**: After resampling with SMOTE, EOD increased for both race (0.4750) and sex (0.3286), suggesting that while the dataset became more balanced, the model became worse at predicting positive outcomes for the unprivileged groups. This is similar to what we observed previously for SPD. This shows that the complexities of the unprivileged group is not adequately represented.

- **RandomUnderSampler**: This technique reduced EOD slightly compared to SMOTE, with values of 0.4456 for race and a higher 0.33 for sex. Although still higher than the baseline, this indicates that RandomUnderSampler performs somewhat better in maintaining a balance of true positive rates across groups of race.

- **SMOTE + ENN**: SMOTE + ENN demonstrated the best performance in reducing bias, with EOD dropping to 0.2169 for race and 0.18 for sex. This improvement suggests that SMOTE + ENN is particularly effective at balancing the model's ability to correctly classify positive outcomes for both privileged and unprivileged groups. Removing noise through ENN ensures that the synthetic data generated by SMOTE does not overly skew the model's true positive rate.

4.3.4 Average Odds Difference Before and After Resampling

Average Odds Difference (AOD) evaluates both the true positive rate (TPR) and false positive rate (FPR) to assess the overall fairness of the model. AOD close to zero indicates that the error rates are similar between privileged and unprivileged groups.

- **Baseline**: The baseline AOD for race was 0.2829, and for sex, it was 0.224. These values indicate moderate bias in the model's predictions, with a notable difference in error rates across both protected attributes.

- **SMOTE**: Using this technique, AOD increased to 0.4072 for race and 0.3157 for sex. This suggests that while SMOTE helped in creating synthetic samples, it led to greater disparities in both TPR and FPR, making the model more prone to fairness-related issues. Resampling the protected attributes instead of labels might have proved to be a better solution.

- **RandomUnderSampler**: AOD improved slightly compared to SMOTE, with values of 0.3769 for race and 0.33 for sex. Although this technique reduces the impact of the majority class, it still introduces a significant disparity in error rates, indicating that under-sampling alone is insufficient to address fairness concerns in terms of average odds.

- **SMOTE + ENN**: This combined technique yielded the best reduction in AOD, bringing it down to 0.1459 for race and 0.1636 for sex. This again shows that removing samples based on

neighbors to filter noise is an effective approach. This technique provided the best balance for TPR and FPR.

4.3.5 Insights Derived from the Confusion Matrices of Resampling

The confusion matrices for the three resampling techniques offer valuable insights into how each method affects the model's performance, particularly regarding its ability to correctly classify both positive and negative outcomes. The visualisations of these matrices are provided in Figure 4.3.

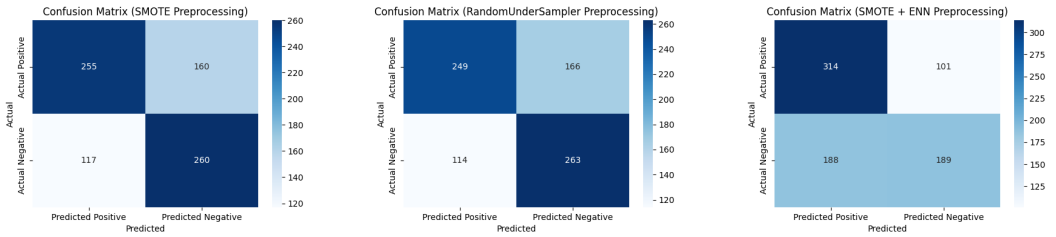

Figure 4.3: Confusion Matrix of resampling technique

1. **SMOTE**

   - **True Positives (TP)**: SMOTE resulted in 255 true positives, indicating that the model correctly identified a good number of positive instances after resampling. But, this is lower than the baseline TP of 288, indicating that SMOTE may have slightly reduced the model's ability to predict positive outcomes.
   - **False Positives (FP)**: The number of false positives increased to 160, showing that the model incorrectly classified 160 negative samples as positive. This increase in FP is a cause for concern because, in sensitive domains, such as criminal justice, false positives can be costly.
   - **False Negatives (FN)**: With 117 false negatives, the model missed 117 actual positive samples. This is lower than the baseline FN of 150, indicating that SMOTE helped the model become more conservative in its negative predictions, reducing the likelihood of false negatives.
   - **True Negatives (TN)**: SMOTE led to the classification of 260 true negatives, an increase from the baseline of 227. This suggests that the model improved in correctly identifying negative samples, which is a positive outcome.

2. **RandomUnderSampler**

   - **True Positives (TP)**: The TP count decreased slightly to 249, lower than both SMOTE and the baseline. This suggests that RandomUnderSampler made the model less likely to identify positive outcomes.

- **False Positives (FP)**: FP increased further to 166, a higher value than even SMOTE's 160. This rise in FP implies that the model became more prone to misclassifying negative samples as positive, which could lead to undesirable consequences.
- **False Negatives (FN)**: The number of false negatives decreased to 114, showing that the model missed fewer positive samples compared to SMOTE (117) and the baseline (150). This improvement suggests that RandomUnderSampler has been effective in reducing missed positive outcomes.
- **True Negatives (TN)**: This increased to 263, the highest among all resampling methods. This suggests that the model becomes more accurate in identifying negative outcomes after under-sampling.

3. **SMOTE + ENN**

   - **True Positives (TP)**: SMOTE + ENN generated 314 true positives, the highest among all the methods, significantly higher than the baseline (288) and other resampling techniques (255 for SMOTE and 249 for RandomUnderSampler). This demonstrates that the model became far more aggressive in identifying positive instances, improving its recall for positive outcomes.
   - **False Positives (FP)**: this decreased to 101, the lowest across all methods. This is an excellent result in terms of reducing the number of negative samples misclassified as positive, showing that SMOTE + ENN produces a more reliable model when predicting positive outcomes.
   - **False Negatives (FN)**: The number of false negatives increased significantly to 188, the highest of all the methods, which indicates that while SMOTE + ENN improved the model's positive predictions, it also became more prone to missing actual positive samples. This increase could be due to the ENN component, which eliminates noisy or borderline samples and may have inadvertently removed informative data points.
   - **True Negatives (TN)**: At 189, lower than both the baseline (227) and the other resampling techniques (260 for SMOTE and 263 for RandomUnderSampler), it indicates that SMOTE + ENN compromised the model's ability to correctly classify negative outcomes.

While it is important to analyse the confusion matrix to gather insights for the models performance, it is also important to consider a trade-off for fairness. SMOTE+ENN was quite capable of reducing bias. Even while sacrificing accuracy, it is still a more reliable model.

4.4 Disparate Impact Remover Results

Disparate Impact (DI) is a measure used to evaluate whether different groups within a dataset are treated fairly by a model. It compares the proportion of favorable outcomes between different groups. A lower Disparate Impact suggests greater disparity between groups, while a value closer to 1 indicates more equitable treatment. Disparate Impact Remover (DIR) is a preprocessing bias mitigation technique designed to remove disparate impact by transforming the data while preserving its utility.

The results for Disparate Impact of the Original Test Set (OTS) and Disparate Impact (Predictions) for the protected attributes of race and sex are summarized in Table 4.4. The test accuracy for both attributes stands at 0.6503, indicating similar performance across these attributes. An evaluation of the experiment results is provided below:

Protected Attribute	Test Accuracy	Disparate Impact (OTS)	Disparate Impact (Predictions)
Race	0.6503	0.6857	0.5744
Sex	0.6503	0.9181	0.6791

Table 4.4: Comparison of Test Accuracy and Disparate Impact for Race and Sex

1. **Protected Attribute: Race**

    - **Test Accuracy**: We observed a test accuracy of 65.03%. This is slightly better than the baseline accuracy after the DIR transformation. It shows that DIR is preserving the nature of the data even after the transformations. The models' performance remains similar.
    - **Disparate impact (Original Test Set)**: Before the application of DIR, the original test set (OTS) had a DI value of 0.6857 for race, indicating an existing bias where unprivileged racial groups receive fewer favorable outcomes than the privileged group.
    - **Disparate Impact (Predictions)**: After the application of DIR, the DI value on the model's predictions decreases to 0.5744. Although DIR is expected to remove or reduce bias, this result reveals that the post-processing predictions still show significant bias. A DI value significantly less than 1 indicates that the model is still treating unprivileged racial groups unfavorably.

2. **Protected Attribute: Sex**

    - **Test Accuracy**: The accuracy for the protected attribute sex remains at 65.03%, consistent with the performance observed across the other protected attribute - race. This again shows that the application of DIR does not harm the overall predictive performance of the model.

- **Disparate Impact (Original Test Set)**: In the original test set, the DI value for sex is 0.9181, which is relatively close to 1, indicating a much smaller degree of bias when compared to race. This suggests that the dataset is more balanced in terms of sex, with favorable outcomes being more equitably distributed between privileged and unprivileged groups.
- **Disparate Impact (Predictions)**: After applying DIR, the DI on predictions for sex falls to 0.6791. Like in the experiment with race, DIR while expected to reduce bias, has amplified it instead. Although the bias present in the predictions is less severe than for race, this drop shows that the model still treats the unprivileged sex group less favorably than the privileged group.

Chapter 5

Conclusion and Future Work

This study aimed to evaluate the effectiveness of various bias mitigation techniques in reducing unfairness in a machine learning model while maintaining its predictive performance. The baseline model exhibited notable bias, particularly in the protected attribute of race, as evidenced by the high values of fairness metrics such as Statistical Parity Difference (SPD), Equal Opportunity Difference (EOD), and Average Odds Difference (AOD).

The reweighing technique showed significant improvements in reducing bias across all fairness metrics for both race and sex, while maintaining a relatively stable accuracy. This suggests that reweighing is an effective method for mitigating unfairness in the model.

Among the resampling techniques, SMOTE + ENN demonstrated the best performance in reducing bias, with SPD, EOD, and AOD values improving substantially compared to the baseline and other resampling methods. However, this came at the cost of a slight decrease in overall accuracy.

The Disparate Impact Remover (DIR) technique, while expected to reduce bias, surprisingly amplified it in the model's predictions for both race and sex. The Disparate Impact (DI) values decreased significantly after applying DIR, indicating that the model still treated unprivileged groups less favorably than privileged groups.

5.1 Limitations

While this study provides valuable insights into bias mitigation techniques, it is not without its limitations.

1. **Dataset Limitations**: The analysis relied on a specific dataset that may not be representative of broader populations. Biases inherent in the dataset could affect the generalizability of the findings.

2. **Model Complexity**: The models used in this study were relatively simple logistic regression models. More complex models may exhibit different behaviors regarding bias and fairness, which warrants further investigation.

3. **Evaluation Metrics**: The study focused on specific fairness metrics (SPD, EOD, AOD) which, while informative, do not capture all aspects of fairness. Other metrics, such as calibration and overall model interpretability, were not assessed.

4. **Trade-offs**: The trade-offs between fairness and accuracy were evident, particularly with certain techniques like SMOTE + ENN, which improved fairness metrics at the cost of accuracy. This raises questions about the practical applicability of these techniques in real-world scenarios where both fairness and accuracy are crucial.

5. **Inherent Limitations of Preprocessing Techniques**: While pre-processing techniques can help reduce the bias, It does not completely mitigate it. Reliable and systematic approaches of in-processing and post-processing techniques need to be paired with this to effectively mitigate bias.

6. **Limited scope of Research**: While the study comprehensively analyses the pre-processing techniques individually, interaction between different techniques as a pipeline has not been evaluated.

5.2 Future Work

Future research should focus on addressing the limitations identified in this study and expanding the scope of bias mitigation techniques.

1. **Diverse Datasets**: Utilizing a more diverse range of datasets can enhance the generalizability of the findings. This includes datasets from various domains and demographic groups.

2. **Advanced Models**: Investigating the effectiveness of bias mitigation techniques on more complex models, such as deep learning architectures and fair learning models, can provide insights into their robustness and applicability.

3. **Comprehensive Metrics**: Future studies should incorporate a broader set of fairness metrics to evaluate models comprehensively. This includes assessing model calibration and interpretability alongside traditional fairness metrics.

4. **Exploration of New Techniques**: Research into additional bias mitigation techniques, particularly those that combine multiple approaches, could lead to more effective solutions. Techniques that balance fairness and accuracy without significant trade-offs should be prioritized.

5. **Longitudinal Studies**: Conducting longitudinal studies to assess the long-term impacts of bias mitigation techniques on model performance and fairness in dynamic environments can provide valuable insights into their sustainability and effectiveness over time.

By addressing these research areas, the findings can contribute to the broader discourse on algorithmic fairness and provide actionable insights for improving equity in machine learning systems, particularly in sensitive applications like criminal justice.

Appendices

Appendix A

Ethical Approval Form

DUNDALK INSTITUTE OF TECHNOLOGY
INSTITIÚID TEICNEOLAÍOCHTA DHÚN DEALGAN

DUNDALK INSTITUTE OF TECHNOLOGY
School of Informatics & Creative Arts
Ethical Approval Form for Research Projects

Researcher Name: ELTON GRIVITH D'SOUZA Year: 5 Course: MSc Data Analytics

Title of project: Pre-processing Techniques for Mitigation of Bias in Machine Learning

Name of supervisor/s: Natalia Budarina Date: 15 APRIL 2024
(if applicable)

This application is to be completed by the researcher and where appropriate, in conjunction with the project supervisor. The lead researcher/supervisor is responsible for submitting the completed form to the appropriate Research Ethics Committee (details below)

Please note: If your submission is incomplete or unclear, your application will be returned to you and your project may be delayed.

Section 1

Type of Researcher

Please tick the appropriate box below to indicate the type of researcher you are:

Undergraduate ☐
(proceed to section 2)

Completed Ethical Approval forms for undergraduate research should be submitted to the relevant Departmental Research Ethics Committee (DREC).

Drec.dcamm@dkit.ie – Department of Creative Arts Media and Music
Drec.dcsm@dkit.ie – Department of Computing Science and Mathematics
Drec.dvhcc@dkit.ie – Department of Visual and Human Centred Computing

Postgraduate ☑
(proceed to section 3)

Completed Ethical Approval forms for Postgraduate research should be submitted directly to the School Research Ethics Committee (SREC).

Srec.ica@dkit.ie

Staff ☐
(proceed to section 3)

Completed Ethical Approval forms for Staff research should be submitted directly to the School Research Ethics Committee (SREC).

Srec.ica@dkit.ie

Section 2

Please complete questions 1-4 listed below.

	Human and / or Animal Research	YES	NO
1	**Does your research involve human participants other than the following[1]?** • Research using exclusively secondary sources. • Research using materials legally accessible to the public that have legal protection, e.g., record of court judgements, data archives. • Research using materials that are publicly accessible and where there is no reasonable expectation for privacy, e.g., books, published third party interviews. • Observations of human behaviour in public where (i) those being observed have no reasonable expectation of privacy, (ii) there is no intervention on the part of the researcher nor any interaction between the researcher and those observed, and (iii) individuals are not identifiable in the results. If '**YES**', please complete **B and C** below.		
2	**Does your project involve working with animals?** – If '**YES**', please complete **B and C** below. – Please note that for ethical consideration: 'Animals' are classed as vertebrate animals including cyclostomes and cephalopods (DIRECTIVE 2010/63/EU)		
3	**Does your project involve working with participants from any of the following categories?** • Minors (under 18 years of age) • People with learning or communication difficulties • Patients • People in custody • People engaged in illegal activities If '**YES**', please complete **D** below.		
4	**Does your project have any possible ethical implications other than those outlined in questions 1, 2 and 3?** If '**YES**', please complete **E** below.		

A. **I consider that this project has no significant ethical implications[2] to be brought through the ICA School Ethics Review Process**
Please complete Section 4

☐

[1] If there is any doubt, researchers should contact the Chair of the SREC.
[2] In determining significant implications, please consider all potential risks attached to this project. If there is any doubt, researchers should contact the Chair of the SREC.

B.

I. **Is this study part of a larger project that already has ethical clearance?**

NO

If **YES** please answer question B II.
If **NO** please answer question C.

II. If this study is part of a larger project that already has ethical clearance, are you proposing any changes to the operational plan already ethically approved?

NO

If **YES,** please complete *Sections 3 and 4.*
If **NO,** then please provide the project details below and complete *Section 4.*

Title of project with ethical clearance: _____

C. Could this project have ethical implications that should be brought before the appropriate ICA Departmental Ethics Review Committee as it will be carried out with human participants?

NO

If **YES,** please complete *Sections 3 and 4.*
If **NO,** please complete *Section 4.*

D. I consider that this project may have ethical implications that should be brought before the School Research Ethics Committee as it will be carried out with human participants in a "vulnerable" category. ☐
Please complete Sections 3 and 4

All research carried out with human participants in a vulnerable category must be referred by the Departmental Research Ethics Committee to the School Research Ethics Committee for approval.

E. Could this project have ethical implications, other than those previously outlined, that should be brought before the appropriate ICA Departmental Ethics Review Committee?

NO

If **YES,** please complete *Sections 3 and 4.*
If **NO,** please complete *Section 4.*

Section 3

3.1 Application Form Checklist

Please complete Section 3 and provide additional information as attachments.

My application includes the following documentation:	INCLUDED (mark as YES)	NOT APPLICABLE (mark as N/A)
Recruitment advertisement		N/A
Participant Information Leaflet		N/A
Participant Informed Consent form		N/A
Questionnaire/Survey		N/A
Interview/Focus Group Questions		N/A
Debriefing material		N/A
Evidence of approval to gain access to off-site location		N/A
Ethical approval from external organizations. If ethical approval from external organizations is pending give details below		N/A
Details		

3.2 Project Details

a) Lay description (Maximum 200 words)

Please outline, in terms that any non-expert would understand, what your research project is about, including what participants will be required to do. Please explain any technical terms or discipline-specific phrases.

Bias is a concept where certain groups or types of data are favored over others, leading to unfair outcomes. The proposed project focuses on analyzing preprocessing techniques with an aim to find better ways to prepare data for machine learning tasks. We hope to make machine learning models that are fairer and accurate while reducing bias by going beyond conventional approaches used for these tasks.

This project involves the study of different analysis techniques, bias detection techniques, and various methods of sampling data to create a comparative study. Instead of using the traditional methods, we'll try new and innovative approaches to make sure that the data being used by the machine learning models are balanced and representative. This work aids in the development of ethically responsible AI systems. We have planned to use the **COMPAS Recidivism Racial Bias** dataset that is publicly available on Kaggle for this task.

The COMPAS Dataset is a popular public-domain commercial algorithm used by judges and parole officers to score a criminal defendant on the individuals' likelihood of reoffending. A two year follow up study on this data has already proved that the data is biased in favor of white defendants and against black inmates. Studies have shown that the pattern of mistakes in the algorithm as measured by precision/sensitivity is notable which makes it optimal for our proposed research.

This data is a public dataset available on Kaggle. The observations in the data includes names of inmates,case id, dates and type of crime among others. We intend to use these to find the metrics for bias for example, how do African-american names cause bias when compared to white names etc.

b) Research objectives (Maximum 150 words)

Please summarise briefly the objectives of the research.

- To investigate multiple pre-processing methods for mitigation of bias (such as suppression, massaging the dataset, reweighting techniques, sampling)
- To examine the performance of machine learning models based on different performance metrics (Example: specificity and accuracy) and different bias metrics (Examples: statistical parity difference, average odds difference)
- Comparative analysis of machine learning algorithms meant for detection and mitigation of bias (Examples: Fairness-aware algorithms, dimensionality-reduction and fair-representation learning, anonymization among others)
- To try to use our inferences for pre-processing techniques to novel approaches in mitigation of bias.

c) Research location and duration

Location(s)/Population*	DkIT campus
Research start date	March 2024
Research end date	August 2024
Approximate duration	6 months

* If location/Population other than DkIT campus/population, provide details of the approval to gain access to that location/population as an appendix.

3.3 Participants

		YES	NO	N/A
Do participants fall into any of	**Minors (under 18 years of age)**			✓
	People with learning or communication difficulties			✓

the following special groups?	Patients			✓
	People in custody			✓
	People engaged in illegal activities (e.g. drug-taking)			✓
Have you given due consideration to the need for satisfactory Garda clearance?				✓

3.4 Sample Details

Approximate number	N/A
Where will participants be recruited from?	N/A
Inclusion Criteria	N/A
Exclusion Criteria	N/A

Will participants be remunerated, and if so in what form?

N/A

Justification for proposed sample size and for selecting a specific gender, age, or any other group if this is done in your research.

N/A

3.5 Risk to Participants

a) Please describe any risks to participants that may arise due to the research. Such risks could include physical stress, emotional distress, perceived coercion e.g. lecturer interviewing own students. Detail the measures and considerations you have put in place to minimize these risks

The names will be used during the research process as it contains valid metrics for identifying bias. The case id will be used to identify and link the observation to previous offences. GDPR regulations will be followed during the overall process of data handling.

b) What will you communicate to participants about any identified risks? Will any information be withheld from them about the research purpose or procedure? If so, please justify this decision.

The data will be obtained from a data hosting platform called Kaggle. The dataset in the platform is listed as `*COMPAS Recidivism Racial Bias*`. This data will be stored in DkIT OneDrive.

3.6 Informed Consent

	YES	NO	N/A
Will you obtain active consent for participation?			✓
Will you describe the main experimental procedures to participants in advance?			✓
Will you inform the participants that their participation is voluntary and may be withdrawn at any point?			✓
If the research is observational, will you ask for their consent to being observed?			✓
With questionnaires, will you give participants the option of omitting questions they do not want to answer?			✓
Will you tell participants that their data will be treated with full confidentiality and that, if published, it will not be identifiable as theirs?			✓
Will the data be anonymous?			✓
Will you debrief participants at the end of their participation?			✓
Will your project involve deliberately misleading participants in any way, or will information be withheld? **If you answer yes, give details and justification for doing this below.**			✓

a)	**Please outline your approach to ensuring the confidentiality of data (that is, that the data will only be accessible to agreed upon parties and the safeguarding**

mechanisms you will put in place to achieve this.) You should include details on how and where the data will be stored, and who will have access to it.

According to the DkIT data retention policy.

b) Please outline how long the data will be retained for, if it will be destroyed and how it will be destroyed.

1.Storage: All data will be retained for 6-12 months after the completion of the project. It will be stored for access in DkIT OneDrive.

2. Access: The dataset is open-source. Access to the research files will be restricted to the researcher and the supervisor.

3. Communication: All communication will be held through DkIT email system and Microsoft Teams meetings.

4. Digital Platform Usage: Since the dataset is opensource, no specialised protection is needed. The files related to the research will be stored as per the researchers needs while ensuring that no party other than the supervisors have access to it.

5. Data maintenance: Data will be maintained on the institute's cloud system as a compressed password protected file.

Section 4

Researcher I have read and I understand the DkIT Ethics Policy available from: https://www.dkit.ie/assets/uploads/documents/Research/Policies/DkIT%20Research%20Ethics%20Policy.pdf

Signed: (Researcher) Print Name: Elton Grivith D'Souza Date: 15 Apr 2024

Supervisor: Applications for Ethical Approval of Undergraduate projects are forwarded to the Departmental Research Ethics Committee for approval or referral to the School Research Ethics Committee. Applications for Ethical Approval of Postgraduate and Staff projects are sent to the School Research Ethics Committee for Approval.

I have read and approved this form & information:

Signed: Print Name: Natalia Budarina Date: 15 Apr 2024
(Supervisor/Head of Department/ Research Centre Director/ Head of School)

There is an obligation on the researcher and/or supervisor to bring to the attention of the Departmental/School Research Ethics Committee(s): (a) Any issues with ethical implications not clearly covered by this form (b) Any ethical issues which may arise during the carrying out of the research; (c) Any ethically significant change made to the project after approval.

Section 5 (For office use only)

STATEMENT OF ETHICAL APPROVAL (FOR UNDERGRADUATE PROJECTS ONLY)

This project has been considered using agreed department procedures and is now:

Approved: ☐

Referred to the School Ethics Committee: ☐

Signed: _____ Print Name: _____ Date: _____
(Chair of Departmental Research Ethics Committee/Head of Department)

STATEMENT OF ETHICAL APPROVAL

This project has been considered using agreed School procedures and is now:

Approved: ☐

Rejected (further information sought): ☐

Chair of School Research Ethics Committee
This project has been considered by the Ethics Committee and ethical approval is granted.

Signed: _____ Print Name: _____ Date: _____
Chair of School Research Ethics Committee

Appendix B

Code

Link to repository: Redirect to GitHub Repository.

We have attached a few preliminary code samples and syntax references below:

1. **Description**: Syntax and Parameters for Classification Metric
   - **Source**: AIF360.
   - **Syntax**:

```
aif360.metrics.ClassificationMetric(dataset,
                                    classified_dataset,
                                    unprivileged_groups=None,
                                    privileged_groups=None)
```

   - **Parameter description**:
     - **dataset (**`BinaryLabelDataset`**)** – Dataset containing ground-truth labels.
     - **classified_dataset (**`BinaryLabelDataset`**)** – Dataset containing predictions.
     - **privileged_groups (*list*(*dict*))** – Privileged groups. Format is a list of *dicts* where the keys are *protected_attribute_names* and the values are values in *protected_attributes*. Each dict element describes a single group.
     - **unprivileged_groups (*list*(*dict*))** – Unprivileged groups in the same format as privileged_groups.
2. **Description**: Syntax and Parameters for `BinaryLabelDatasetMetric`. Used for computing metrics based on a single `BinaryLabelDataset`.
   - **Source**: AIF360.
   - **Syntax**:

```
aif360.metrics.BinaryLabelDatasetMetric(dataset,
                                        unprivileged_groups=None,
                                        privileged_groups=None)
```

   - **Parameter description**:
     - **dataset (**`BinaryLabelDataset`**)** – Dataset containing ground-truth labels.
     - **privileged_groups (*list*(*dict*))** – Privileged groups. Format is a list of *dicts* where the keys are *protected_attribute_names* and the values are values in *protected_attributes*. Each dict element describes a single group.
     - **unprivileged_groups (*list*(*dict*))** – Unprivileged groups in the same format as privileged_groups.

Appendix C

AI Queries Utilized Throughout the Research Period

The following queries were used during the course of this research. GPT tools include ChatGPT, Perplexity, and Merlin. Long format queries have not been listed directly but the actionable commands in those queries have been listed.

- General structure of an academic interim report focused on literature review
- Most of the formulas were translated to LaTeX code using ChatGPT
- Privileged groups in machine learning
- Is the COMPAS model used anywhere other than America?
- Tools to automate and search for literature review
- Simplify the systematic review process and provide task list
- Difference between baseline and ground truth
- What is a theoretical framework and how to structure literature inside it?
- Metrics commonly used for bias detection
- Odds ratio interpretation
- Set theory fundamentals
- Survival analysis and its applications in debiasing
- Kaplan-Meier explanation
- Logit assumptions in Statsmodels
- accuracy remained the same after SMOTE
- Do you resample the protected groups or the target feature when analysing fairness
- How does ENN work?
- Automating EDA for univariate, bivariate, and multivariate analysis in Python
- Justify deterministic splits in words
- how to read disparate impact
- Circumstances in which DIR would fail

- (provided a paragraph as template): replace the values with the new confusion matrix values.
- how to attach pdf documents in overleaf journals.
- Write an outline for abstract for the research of comparing different pre-processing techniques in the context of fair learning and mitigating bias.
- (Attached the results section): Write a brief colclusion and future works for my thesis.
- (Provided lifecycle flowchart and research design): Use this to define the scope of study for documentation.

AI provided by Anaconda's JupyterLab was used in several instances to debug code. LucidCharts AI features were used to create project lifecycle flowchart (This was then edited manually).

Bibliography

Barenstein, M. (2019), 'Propublica's compas data revisited', *arXiv preprint arXiv:1906.04711* .

Bellamy, R. K., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilović, A. et al. (2019), 'Ai fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias', *IBM Journal of Research and Development* **63**(4/5), 4–1.

Chiao, V. (2019), 'Fairness, accountability and transparency: notes on algorithmic decision-making in criminal justice', *International Journal of Law in Context* **15**(2), 126–139.

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

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S. & Huq, A. (2017), Algorithmic decision making and the cost of fairness, *in* 'Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining', pp. 797–806.

Dressel, J. & Farid, H. (2021), 'The dangers of risk prediction in the criminal justice system'.

du Pin Calmon, F., Wei, D., Ramamurthy, K. N. & Varshney, K. R. (2017), 'Optimized data preprocessing for discrimination prevention', *CoRR, abs/1704.03354* .

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C. & Venkatasubramanian, S. (2015), Certifying and removing disparate impact, *in* 'proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining', pp. 259–268.

González-Cebrián, A., Bradford, M., Chis, A. E. & González-Vélez, H. (2024), 'Standardised versioning of datasets: a fair–compliant proposal', *Scientific Data* **11**(1), 358.

Hardt, M., Price, E. & Srebro, N. (2016), 'Equality of opportunity in supervised learning', *Advances in neural information processing systems* **29**.

Hoffmann, A. L. (2019), 'Where fairness fails: data, algorithms, and the limits of antidiscrimination discourse', *Information, Communication & Society* **22**(7), 900–915.

Janssen, M. & Kuk, G. (2016), 'The challenges and limits of big data algorithms in technocratic governance'.

Kleinberg, J., Ludwig, J., Mullainathan, S. & Rambachan, A. (2018), Algorithmic fairness, *in* 'Aea papers and proceedings', Vol. 108, American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203, pp. 22–27.

Lamprecht, A.-L., Garcia, L., Kuzak, M., Martinez, C., Arcila, R., Martin Del Pico, E., Dominguez Del Angel, V., Van De Sandt, S., Ison, J., Martinez, P. A. et al. (2020), 'Towards fair principles for research software', *Data Science* **3**(1), 37–59.

Larson, J., Mattu, S., Kirchner, L. & Angwin, J. (2016), 'How we analyzed the compas recidivism algorithm', *ProPublica (5 2016)* **9**(1), 3–3.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K. & Galstyan, A. (2021), 'A survey on bias and fairness in machine learning', *ACM computing surveys (CSUR)* **54**(6), 1–35.

Odyssey, A. D. (2023), ‘Definitions of fairness in machine learning | equal opportunity, equalized odds disparate impact’, *YouTube* .

Raza, S., Ghuge, S., Ding, C. & Pandya, D. (2024), ‘Fair enough: How can we develop and assess a fair-compliant dataset for large language models’ training?’, *arXiv preprint arXiv:2401.11033* .

Santos, K. S. S., Pinheiro, L. B. L. & Maciel, R. S. P. (2021), Interoperability types classifications: A tertiary study, *in* ‘Proceedings of the XVII Brazilian Symposium on Information Systems’, pp. 1–8.

SMO (n.d.). SMOTE Documentation. Retrieved from https://imbalanced-learn.org/stable/references/generated/imblearn.over_sampling.SMOTE.html on September 7, 2024.

Valavi, E., Hestness, J., Ardalani, N. & Iansiti, M. (2022), ‘Time and the value of data’, *arXiv preprint arXiv:2203.09118* .

Zandee, R. (2021), ‘A comparative study of bias mitigation methods applied on a multitude of classification algorithms’, *no. June* .