SENAI CIMATEC UNIVERSITY CENTER

POSTGRADUATE PROGRAMME IN COMPUTATIONAL MODELLING AND INDUSTRIAL TECHNOLOGY

Master in Computational Modelling and Industrial Technology

Master's Dissertation

Computational method for grouping and reducing representative metrics for identification and mitigation of bias and unfairness in machine learning models.

MSc.Student: Rafael Bessa Loureiro
Supervisor: Prof. Dr. Erick G. Sperandio Nascimento
Co-Supervisor: Prof.ª Dr.ª Ingrid Winkler
Co-Supervisor: Prof. Dr. Ewerton de Oliveira

September 2024

Rafael Bessa Loureiro

Computational method for grouping and reducing representative metrics for identification and mitigation of bias and unfairness in machine learning models.

Master's Dissertation presented to the Postgraduate Programme in Computational Modelling and Industrial Technology, Master in Computational Modelling and Industrial Technology Course from SENAI CIMATEC University Center, as a partial requirement for obtaining the degree of **Master in Computational Modelling and Industrial Technology**.

Supervisor: Prof. Dr. Erick G. Sperandio Nascimento

Co-supervisor: Prof.ª Dr.ª Ingrid Winkler

Co-supervisor: Prof. Dr. Ewerton de Oliveira

Salvador
2024

Ficha catalográfica elaborada pela Biblioteca do Centro Universitário SENAI CIMATEC

L892m Loureiro, Rafael Bessa

Método computacional para agrupamento e redução de métricas representativas para identificação e mitigação de viés e injustiça em modelos de aprendizado de máquina / Rafael Bessa Loureiro. – Salvador, 2024.

136 f. : il. color.

Orientador: Prof. Dr. Erick Giovani Sperandio Nascimento.
Coorientadora: Prof.ª Dra. Ingrid Winkler.
Coorientador: Prof. Dr. Ewerton de Oliveira.

Dissertação (Mestrado em Modelagem Computacional e Tecnologia Industrial) – Programa de Pós-Graduação, Centro Universitário SENAI CIMATEC, Salvador, 2024.
Inclui referências.

1. Viés. 2. Injustiça. 3. Métricas representativas. 4. Correlação. I. Centro Universitário SENAI CIMATEC. II. Nascimento, Erick Giovani Sperandio. III. Winkler, Ingrid. IV. Oliveira, Ewerton de. V. Título.

CDD 620.00113

NDI - 04

CENTRO UNIVERSITÁRIO SENAI CIMATEC

Mestrado Acadêmico em Modelagem Computacional e Tecnologia Industrial

A Banca Examinadora, constituída pelos professores abaixo listados, aprova a Defesa de Mestrado, intitulada **“Método computacional para agrupamento e redução de métricas representativas para identificação e mitigação de viés e injustiça em modelos de aprendizado de máquina**” apresentada no dia 05 de setembro de 2024, como parte dos requisitos necessários para a obtenção do Título de Mestre em Modelagem Computacional e Tecnologia Industrial.

Electronically signed by:
Erick Giovani Sperandio Nascimento
CPF: ***.666.177-**
Date: 9/7/2024 9:37:22 PM +01:00

Orientador: **Prof. Dr. Erick Giovani Sperandio Nascimento**
SENAI CIMATEC

Assinado eletronicamente por:
INGRID Winkler
CPF: ***.486.968-**
Data: 06/09/2024 16:31:34 -03:00

Coorientadora: **Prof.ª Dr.ª Ingrid Winkler**
SENAI CIMATEC

Electronically signed by:
Ewerton Lopes Silva de Oliveira
CPF: ***.266.364-**
Date: 9/6/2024 10:02:02 PM -03:00

Coorientador: **Prof. Dr. Ewerton Lopes Silva de Oliveira**
HP

Assinado eletronicamente por:
Camila de Sousa Pereira Guizzo
CPF: ***.843.378-**
Data: 10/09/2024 12:13:30 -03:00

Membro Interno: **Prof.ª Dr.ª Camila de Sousa Pereira-Guizzo**
SENAI CIMATEC

Assinado eletronicamente por:
PEDRO MARIO CRUZ E SILVA
CPF: ***.276.774-**
Data: 09/09/2024 13:53:40 -03:00

Membro Externo: **Prof. Dr. Pedro Mário Cruz e Silva**
PUC-Rio

Acknowledgments

There are many people I must thank for their invaluable guidance and support throughout the development of this thesis.

First and foremost, I am profoundly grateful to my supervisors. I would like to extend my deepest appreciation to Prof. Dr. Erick G. Sperandio Nascimento, whose insightful advice, enthusiastic support, and deep expertise consistently motivated and guided me through my research. My sincere thanks to Prof. Dr.ª Ingrid Winkler for her meticulous guidance, thoughtful suggestions, and for tirelessly reviewing my work, contributing significantly to the refinement of this manuscript. I am also deeply thankful to Prof. Dr. Ewerton de Oliveira for his steadfast support, fruitful discussions, and expert advice.

My heartfelt thanks to the entire VIALAB team, including the ones that have left for new opportunities, for their invaluable support and incentive, in particular to Prof. Dr. Oberdan Rocha Pinheiro and prof. Dr. Tiago Palmas Pagano, whose support helped me overcome numerous challenges along the way.

I am also grateful to the professors and colleagues in the MCTI for their academic and personal support. Your insightful feedback and constant encouragement have been a cornerstone of my academic progress and personal growth throughout this program. Special thanks go to my fellow researchers and friends within the MCTI. Your companionship, stimulating discussions, and shared experiences have made this journey both intellectually rewarding and personally fulfilling.

Lastly, I would like to express my deepest gratitude to my family. Your belief in me has been a constant source of strength and motivation, for which I am eternally grateful.

Salvador, Brasil
05 September 2024

Rafael Bessa Loureiro

Abstract

Bias and unfairness in machine learning models happen when they make biassed or unfair decisions that perpetuate and amplify the unfair discrimination and exclusion of people. Identifying and addressing bias and unfairness in those models in different application domains is a multifaceted challenge. While numerous unfairness metrics have been proposed, determining an optimal set of metrics for assessing a model's unfairness remains an open question in the literature due to the diverse nature of these metrics and the lack of comprehensive approaches to ensure fairness across multiple applications. Consequently, there is a pressing need to narrow down the metric space and identify representative metrics for algorithmic unfairness evaluation. The current literature presents a limited number of studies aimed at reducing the number of fairness metrics used when evaluating a model, with the available techniques facing limitations, including restriction to specific application areas, dependence on the user's understanding of the problem, and high computational cost. Therefore, this study aims to propose a computational method that allows the selection of the most representative metrics for bias and unfairness assessment in post-processing for binary classification machine learning models in different contexts. To achieve this goal, four case studies were used in the fields of criminal judgement, bank loans, demographic census, and advertisement, with unfairness identified against the sensitive attributes: race, gender, race, and age group. Furthermore, a correlation-based strategy was used as a heuristic for selecting unfairness metrics. The potential problems with the approach were then analysed, and solutions were proposed to mitigate these problems and evaluate its effectiveness. The method starts the procedure using bootstrap sampling in conjunction with the Markov chain Monte Carlo method. Modifications and validation strategies are proposed, such as transitioning to a stratified sampling method to better represent the data biases, incorporating a stopping criterion to reduce the computational cost, shifting from Pearson to Kendall correlation for more robust estimations, and validating the method by examining different aspects of the selected metrics. A substantial reduction in computational cost was noted, with an average decrease of 64.37% in the number of models required and of 20.00% in processing time. Moreover, the proposed method maintains result consistency by effectively pairing metrics with similar behaviour. The proposed experiment was able to group metrics with similar equations more frequently, making the presence of a similar term in the equation a strong indicator of a direct relationship between two metrics.

While no standout metric emerges across all contexts, within specific models or datasets, certain metrics consistently stand out. For the analysed cases, the Predictive Parity metric was highlighted in the criminal judgement, demographic census, and advertisement scenarios, while the Error Ratio metric was highlighted for the demographic census, and Equalized Odds was in evidence in criminal judgment. Overall, the proposed method successfully selects the representative metric with a considerable gain in computational costs.

Keywords: bias, unfairness, representative metric, correlation

Resumo

O viés e a injustiça em modelos de aprendizado de máquina ocorrem quando esses tomam decisões enviesadas ou injustas que perpetuam e amplificam a discriminação e exclusão injusta de pessoas. Identificar e abordar o viés e a injustiça desses modelos em diferentes domínios de aplicação é um desafio multifacetado. Apesar de várias métricas de injustiça terem sido propostas, determinar um conjunto ideal de métricas para avaliar a injustiça de um modelo continua sendo uma questão em aberto na literatura devido à natureza diversa dessas métricas e a falta de métodos abrangentes que garantam a justiça em múltiplas aplicações. Consequentemente, existe uma necessidade imediata de restringir a quantidade de métricas e identificar as métricas representativas para a avaliação da injustiça algorítmica. A literatura atual apresenta um número limitado de estudos voltados para a redução do número de métricas de injustiça utilizadas ao avaliar um modelo, com as técnicas disponíveis enfrentam limitações, incluindo a restrição a áreas específicas de aplicação, a dependência do entendimento do usuário sobre o problema, e o elevado custo computacional. Portanto, este estudo tem como objetivo propor um método computacional que permita a seleção das métricas mais representativas para avaliação de viés e injustiça, em pós-processamento, para modelos de aprendizado de máquina de classificação binária em diferentes contextos. Para alcançar esse objetivo, são utilizados quatro estudos de casos, nas áreas de julgamento criminal, empréstimo bancário, censo demográfico e publicidade, com injustiças identificadas contra os atributos sensíveis: raça, gênero, raça e faixa etária. Além disso, foi utilizado uma estratégia baseada em correlação como uma heurística para a seleção de métricas de injustiça. Em seguida, foram analisados os potenciais problemas da abordagem, propondo soluções para atenuar esses problemas e avaliar a sua eficácia. O método inicia o procedimento utilizando uma amostragem por bootstrap em conjunto com a técnica de Monte Carlo via cadeias de Markov. Modificações e estratégias de validação são propostas, como a transição para um método de amostragem estratificada para representar melhor os vieses dos dados, incorporação de um critério de parada para reduzir o custo computacional, substituição da correlação de Pearson para a de Kendall para obter estimativas mais robustas, e a validação do método por meio da análise de diferentes aspectos das métricas selecionadas. Foi constatado uma redução substancial no custo computacional, com uma diminuição média de 64,37% no número de modelos necessários e de 20,00% no tempo de processamento. Além disso, o método proposto mantém a consistência dos resultados ao agrupar efetivamente métricas com comportamento semelhante. O experimento proposto foi capaz de agrupar métricas com equações semelhantes com mais frequência, tornando a presença de um termo semelhante na equação um forte indicador de uma relação direta entre duas métricas.

Embora não surja nenhuma métrica que se destaque em todos os contextos, certas métricas se destacam em modelos ou conjuntos de dados específicos. Para os casos analisados, a métrica de Paridade Preditiva se destacou nos cenários de julgamento criminal, censo demográfico e publicidade, enquanto que a métrica de Taxa de Erro foi destaque no censo demográfico, e a métrica de Probabilidades Equalizadas foi evidenciada no julgamento criminal. De modo geral, o método proposto seleciona com sucesso as métricas mais representativas, com considerável ganho em custo computacional.

Palavras-chave: viés, injustiça, métricas representativas, correlação

Contents

List of Tables

List of Figures

List of Abbreviations

PPGMCTI	Postgraduate Programme in Computational Modelling and Industrial Technology
UN	United Nations
SDGs	Sustainable Development Goals
ML	Machine learning
FOR	False Omission Ratio
PPR	Positive Prediction Ratio
FPR	False Positive Rate
WEFE	Word Embeddings Fairness Evaluation
MLP	Multilayer-perceptron Neural Network
Logit	Logistic Regression
SVC	C-Support Vector Machine
KNN	K-nearest Neighbours
RF	Random Forest
RBF	Radial Basis Function
tanh	Hyperbolic Tangent Function
ReLU	Rectified Linear Unit
MAD	Median Absolute Deviation
TP	True Positive
FN	False Negative
FP	False Positive
TN	True Negative
LLM	Large Language Model
MCMC	Markov chain Monte Carlo

List of Symbols

$\mathcal{X}$	Set of non-sensitive independent variables
$\mathcal{Y}$	Set of dependent (target) variables
$\mathcal{S}$	Set of sensitive variables
w	Weight
s	Specific group in the sensitive attribute S
b	Bias
$*$	Dot product operation
P	Probability
k	Number of nearest points to use
Ke	Kernel function.
α	Lagrange multiplier
η	Learning rate
J	Loss function
∂	Partial derivative
σ	Width of the Gaussian function
h	Classifier
h_t	T-th decision tree
T	Total Number of Trees
N	Number of observations in the dataset
f	Unfairness metric
N	Number of observations
R	Representative subset of fairness metrics
r	Correlation Coefficient
K	Number of classifier instances
L	Number of iterations
D	Dataset
B	Bootstrapped dataset
C	Concordant pairs
DI	Discordant pairs

Chapter One

Introduction

The United Nations (UN) Sustainable Development Goals (SDGs) communicate the challenges and threats to our societies, and include 17 goals for the 2030 agenda, including those regarding social justice. Machine learning (ML) systems may assist or hinder these goals, thus it's important to be aware of the potential negative consequences (SÆTRA, 2022). ML algorithms used in decision-making systems, such as hiring processes, online advertising, loan processing, criminal pre-trial, immigrant detention, and public health, among other applications, are not free from bias and other issues related to sensitive social aspects such as race, gender, religion, and so on (PAGANO et al., 2023a).

ML systems have a significant negative impact on SDGs 5, 10, 13, and 16, corresponding respectively to: gender equality; reduced inequalities; climate action; and peace, justice, and strong institutions. Human involvement in system development and data curation unintentionally perpetuates pre-existing historical and social biases that disproportionately affect marginalised groups, compromising SDGs 5 and 10 (SÆTRA, 2022). Training large AI models on increasingly large datasets consumes a lot of electricity, contributing significantly to carbon emissions. For example, the large language model (LLM) GPT-3 is estimated to have consumed a 1287 MWh of electricity and produced 550 tons of carbon dioxide during the training phase (BOLóN-CANEDO et al., 2024), thus negatively impacting SDG 13. The lack of transparency and accountability also raises concerns about the viability of those systems, potentially leading to unexpected negative consequences and impacting SDG 16 (SÆTRA, 2022).

The importance of fair ML systems is recognised by policymakers and the academic community, with various researchers addressing the unfairness sources, impacts, and mitigation strategies (FERRARA, 2024). Identifying bias and unfairness, for example, is a difficult task because definitions can change over time and/or be different for different people and societies depending on historical, cultural, political, legal, social, and ethical factors that vary in different contexts (MITCHELL et al., 2019; PAGANO et al., 2023b).

Context refers to any information that can be used to characterise an entity's situation (ZIMMERMANN; LORENZ; OPPERMANN, 2007). The context examines the entity in a given scenario, where the place and time drive the construction of relationships and enable the exchange of information between them (ZIMMERMANN; LORENZ; OPPERMANN, 2007). In this work, context is defined as a combination of model, data, and fairness metrics. Most bias and unfairness solutions are particular to a single problem or situation, not indicating a method capable of being applied to every context (ADEL et al., 2019; PAVIGLIANITI; PASERO, 2020; SHI et al., 2020; QUADRIANTO; SHARMANSKA, 2017).

There are several metrics for detecting bias and unfairness — functions defined on key model results, such as rates of true and false positives, false omission ratio (FOR), and positive prediction ratio (PPR). There are more than 70 different fairness metrics (PAGANO et al., 2023a; MAJUMDER et al., 2023; SMITH; BEATTIE; CRAMER, 2023; CASTELNOVO et al., 2022), each exploring and exposing different aspects of bias and unfairness. There is also a lack of accepted standards or shared practice knowledge (SMITH; BEATTIE; CRAMER, 2023), yet new metrics may be proposed to tackle specific challenges. Testing all different fairness metrics and concepts is a time-consuming and difficult task that discourages ML practitioners from properly evaluating their models, resulting in either a model without concern for injustice or a poorly evaluated model with no clear fairness objective, tested with suboptimal metrics. Furthermore, exhaustively testing a model with all available fairness metrics spends a substantial computational cost, reducing the sustainability of ML model training by increasing the carbon footprint. The variety of fairness metrics demands significant expertise with algorithmic unfairness when evaluating a model (NIELSEN, 2020), resulting in a growing gap between researchers and practitioners (SMITH; BEATTIE; CRAMER, 2023).

There is also the impossibility theorem, which states that the calibration of unfairness in a group and the simultaneous balance of positive and negative classes are only possible in two simplified cases: 1) a classifier with perfect prediction, and 2) when the distribution of groups is equal and all predictions of the classifier are equal (KLEINBERG; MULLAINATHAN; RAGHAVAN, 2016). Additionally, there is no solution for building systems that are fair across multiple metrics, making it even more difficult for developers to evaluate model performance. As a result, narrowing the metric space is desirable.

Consequently, determining a representative metric from a set of metrics remains an open research question (PAGANO et al., 2023a). All due to the wide variety of metrics available and their reliance on the issue at hand. This identification is advantageous, as it could pave the way for the design of systems to help developers and practitioners better control over unfairness in their software (PAGANO et al., 2023b). Furthermore, identifying a representative set of metrics would empower ML practitioners to effectively evaluate their models, making the process more feasible and contributing to reduced carbon emissions by minimizing the need for exhaustive testing with all available metrics.

1.1 Problem definition

Each fairness metric has a different goal for bias identification, so when analysing a variety of metrics, different bias aspects appears, including contrasting results (BALAYN; LOFI; HOUBEN, 2021). There are limitations to build a fair system considering a broad number of metrics, which includes the selections of an appropriated metric to evaluate and make model selection decisions. Therefore, striving towards selecting a representative metric is ideal as it would reduce the metric space and facilitate the identification and mitigation of bias and unfairness aspects.

1.2 Research questions and assumptions

Given the critical issue of assessing unfairness in ML models and the large number of metrics, this study focuses on the following research question:

How to reduce the fairness metrics scope for a problem, maintaining their representativeness, and supporting identification and mitigation of bias and unfairness in a model?

To address this research question, the following assumptions are proposed:

1. Fairness metrics may have similar behaviour within the same context.
2. Fairness metrics, with similar behaviour, can be grouped and replaced, by a representative metric of the group, in a given context.
3. The use of a representative metric has a similar effect as using the represented ones to analyse bias and unfairness in a model.

1.3 Objective

The overarching objective of this work is to propose a computational method that allows the selection of the most representative metrics for bias and unfairness assessment in post-processing for binary classification machine learning models in different contexts.

To achieve this overarching goal, the specific objectives are to:

1. Analyse and select state-of-the-art metrics used for bias and unfairness identification in ML models.

2. Analyse cases studies in different contexts with relevant bias and unfairness phenomena.
3. Develop a correlation-based algorithm that performs the grouping and selection of the representative fairness metrics.
4. Analyse the performance of the developed algorithm for identifying bias and unfairness through the selected representative metric for different cases studies in different contexts

1.4 Significance of the research

This study significance lies on the fact that it elaborates on the use of correlation as a heuristic for fairness metric selection, highlighting important caveats, and providing a study on the efficiency of this approach by revisiting ideas from previous empirical research. The contributions are understood as follows:

- Ideas about using correlation as a heuristic tool for finding a representative fairness metric for a context (data + model) were confronted and improved;
- The required computational power needed to find the representative metrics was reduced by adding a stopping criterion;
- Experimental results demonstrating the effectiveness of the mentioned approach in different application domains are provided.

Moreover, it serves as tool to facilitate unfairness assessment in the model development process, addressing sustainability by fostering the development of more socially sustainable and trustworthy AI solutions that can help achieve UN's SDGs. Finally, having an efficient metric selection algorithm is crucial, because as the models become more complex, re-training can be expensive in terms of money, effort, and carbon footprint.

1.5 Limits and limitations

As scope limitation, we have the following:

1. Identification of bias and unfairness will only be done with post-processing approaches.

2. The ML models are treated as black boxes.
3. Simple and faster to train models will be used.
4. Only a select group of bias and fairness metrics will be used.

Limitations 1 does not consider different approaches identifying bias and unfairness in other steps of ML models development, like data-preprocessing or the training loop, which has its own set of challenges, it also guarantees that the developed method can be used without modifying the already existing training pipelines. Limitation 2 ensures that no model modification, such as hyper-parametrization or architectural changes, will occur, facilitating the model unfairness evaluation without the need to understand the targeted model intricacies. Limitations 3 and 4 are placed to guarantee that the research focus stays on develop and validate the methodology, with a feasible amount of variables to analyse.

1.6 Organisation of the Master's Dissertation

This document has 5 chapters and is structured as follows:

- **Chapter 1 - Introduction**: It contextualises the scope in which the proposed research is inserted.It therefore presents the definition of the problem, the objectives and justifications of the research and how this master's qualifying dissertation is structured;
- **Chapter 2 -Literature Review**: It describes ML ideas, bias and unfairness studies, that were used in this work. It also includes other researches that constitute the state of the art in determining a representative fairness metric.
- **Chapter 3 - Material and Methods**: This section describes the materials and methods used in master's qualifying dissertation. Materials include datasets, models, and fairness metrics, in addition to methods including bootstrap sampling, metric correlation, and stopping criteria. Used to choose representative fairness metrics and evaluate the experiments.
- **Chapter 4 - Results and Discussions**: This section gives the outcomes of the experiment using the proposed method and compares them to the base method.
- **Chapter 5 - Final considerations**: This section includes conclusions, contributions and suggestions for future research activities.

Chapter Two

Literature review

This chapter is divided into two sections: section 2.1 provides a comprehensive explanation of the theoretical aspects essential for understanding the elements presented in this dissertation, and section 2.2 reviews the current state-of-the-art for the research topic.

2.1 Theoretical background

2.1.1 Machine Learning Models

ML is a requirement for artificial intelligence, as a system in an ever-evolving environment should be able to learn, avoiding committing the same mistakes over and over again. In that sense, ML can be defined as a field of study that gives computers the ability to learn without being explicitly programmed (ALPAYDIN, 2021; MAHESH, 2020). This field is a junction of computer science, statistics, and a variety of disciplines focused on automatic improvement over time and decision-making in uncertain situations (JORDAN; MITCHELL, 2015).

ML uses many algorithms to solve issues by learning from previously collected data. There is not an absolute algorithm that is best for solving every kind of problem; the utilised approach is dependent on the problem, the data, the amount of variables, and several others complex characteristics. (MAHESH, 2020).

These algorithms can be categorised as either supervised or unsupervised. The former involves working with labelled input data to enable the classification of fresh unlabelled data. It is required to have one set of data for training and another for testing. The distinction between both is the absence of annotations for the test data, enabling the algorithm to obtain classifications from the training data. Unsupervised learning uses unlabelled data and organises it based on similarity patterns (ZHANG, 2020).

Classification algorithms attempt to predict the class of incoming data by learning about comparable data from previous observations. They are typically a subset of supervised learning algorithms, with the purpose of predicting the category of a new set of data. This work will use five supervised classification algorithms: multilayer-perceptron neural network (MLP), logistic regression (Logit), k-nearest neighbours (KNN), c-support vector machine (SVC) and random forest classifier (RF).

2.1.1.1 Multilayer-Perceptron Neural Network (MLP)

The multilayer perceptron is a popular type of neural network. Signals are transported in one direction from input to output utilising nodes or neurons in a feedforward architecture. The network typically is organised into layers, starting with the input layer, where data is introduced, followed by hidden layers where computations are performed, and ultimately, the output layer, where the final decision is made, as shown in Figure 2.1.

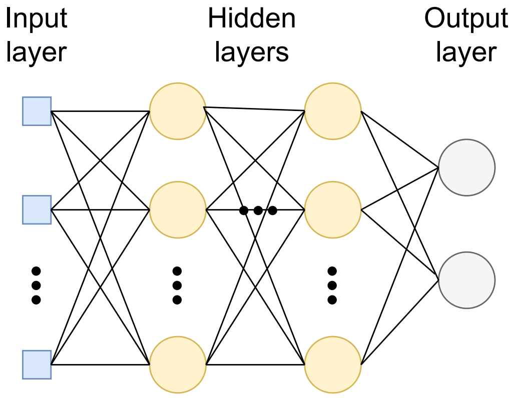


Figure 2.1: MLP typical architecture, representing the input, hidden, and output layers. Adapted from (TAUD; MAS, 2018).

The neuron makes decisions in the shape of a semi-plane, but by connecting it to another neuron, another semi-plane is added to the first, resulting in convex decision areas that enable complex solutions. The intensity of association between two neurons is defined by a weighted connection (w_i), which follows Equation 2.1 for the input ($\mathcal{X}_i$) and output ($\mathcal{Y}$), where m represents the number of inputs and weights. Another important component is the bias (b), which is a constant input to a neuron that is not associated with other neurons, and the activation function, which applies a function over the weighted sum of the neuron inputs to control the output, introducing nonlinearity into the network and allowing it to learn complex patterns in the data. Typical functions include sigmoid in Equation 2.2, hyperbolic tangent function (tanh) in Equation 2.3, Rectified Linear Unit (ReLU) in Equation 2.4, and softmax in Equation 2.5 (POPESCU et al., 2009).

$$y = f\left(\sum_{i=1}^{m} w_i \mathcal{X}_i + b\right) \tag{2.1}$$

$$sigmoid(\mathcal{X}) = \frac{1}{1+e^{-\mathcal{X}}} \tag{2.2}$$

$$\tanh(\mathcal{X}) = \frac{e^{\mathcal{X}} - e^{-\mathcal{X}}}{e^{\mathcal{X}} + e^{-\mathcal{X}}} \tag{2.3}$$

$$\mathrm{ReLU}(\mathcal{X}) = \max(0, \mathcal{X}) \tag{2.4}$$

$$\mathrm{softmax}(\mathcal{X}_i) = \frac{e^{\mathcal{X}_i}}{\sum_j e^{\mathcal{X}_j}} \tag{2.5}$$

Feedforward propagation and backpropagation are strategies for training neural networks. During feedforward propagation, input data is sent through the network layer by layer, with each layer conducting a computation based on the inputs it receives and passing the results on to the next layer. At the end, an error signal is generated. During backpropagation, the MLP adjusts network weights using a gradient descending minimization procedure, as shown in Equation 2.6, based on the error signal. η is the learning rate, which determines the step size for the weight update; J denotes the loss function, which measures the difference between the predicted output and the actual target; and $\frac{\partial J}{\partial \mathbf{w}^{(l)}}$ is the gradient of the loss function, indicating how J changes as w_i changes (KARLIK; OLGAC, 2011). This process iteratively adjusts the weights to minimise the loss function, thus improving the performance of the MLP.

$$w_i = w_i - \eta \frac{\partial J}{\partial w_i} \tag{2.6}$$

2.1.1.2 Linear Regression and Logistic Regression (Logit)

A linear regression formalises a statistical relationship between two variables, indicating that $\mathcal{Y}$ is linearly connected to $\mathcal{X}$. $\mathcal{X}$ is the input feature, and $\mathcal{Y}$ is the predicted outcome. The first step in a regression is to plot the data in a plane, as shown in Figure 2.2, with the $\mathcal{X}$ on the horizontal axis and the $\mathcal{Y}$ on the vertical axis, to approximate a linear relation between the two, defined by Equation 2.7, where b is a bias and w is a weight (HILBE, 2009). The bias represents where the line intersects with Y, whereas the weight represents the curve's slope.

$$\mathcal{Y} = b + w\mathcal{X} \tag{2.7}$$

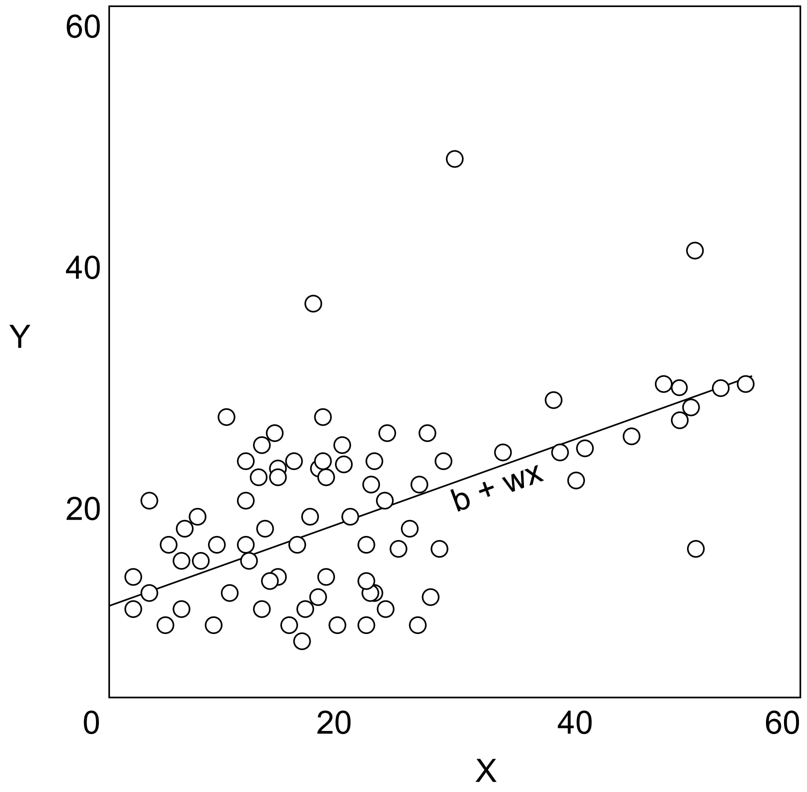


Figure 2.2: Estimated line from the linear regression of Y on X. Adapted from (AMBROSIUS, 2007).

Linear regression is useful for data with a linear connection that can be represented by first-order approximations; however, its application is limited, thus linear regression is not appropriate. Furthermore, linear regression with a continuous or binary outcome between 0 and 1 may not be ideal. Alternatively, logistic models can be used (KOMAREK, 2004). A Logit model describes the connection between a qualitative dependent variable and an independent variable. In general, a logistic regression model derives the class membership probability for one of the data set's two categories by applying the sigmoid function to a linear combination of input features. Logit uses Equation 2.8 and a representation can be seen in Figure 2.3 (HILBE, 2009; DREISEITL; OHNO-MACHADO, 2002; KOMAREK, 2004; CAMDEVIREN et al., 2007).

$$P(\mathcal{Y} = 1 \mid \mathbf{x}) = sigmoid(w * \mathcal{X} + b) \tag{2.8}$$

Where the sigmoid function is defined in Equation 2.2, and $P(y = 1 \mid \mathcal{X})$ represents the probability that the output y is 1 (positive class) given the input $\mathcal{X}$, and $*$ denotes a dot product operation.

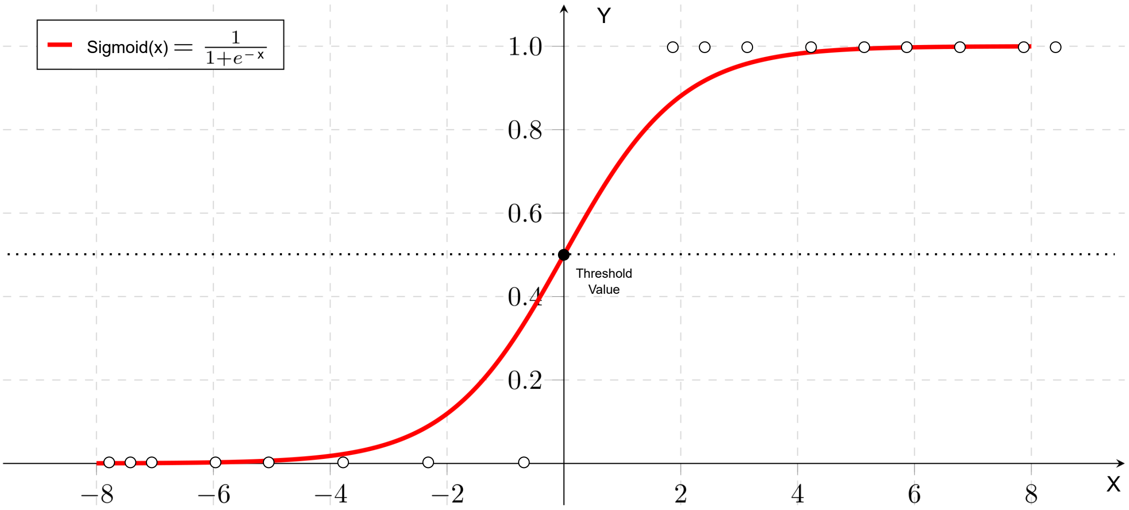


Figure 2.3: Estimated line from the Logit for a binary classification. Adapted from (KIRASICH; SMITH; SADLER, 2018).

2.1.1.3 K-Nearest Neighbours (KNN)

The KNN algorithm is a simple supervised machine learning technique that can be applied to classification tasks. Although simple to develop and understand, this method has a major drawback as it becomes slower as the data size grows (POPESCU et al., 2009). This algorithm classifies data directly, memorising the whole training dataset, and the model's only configurable parameter is the number of nearest neighbours to incorporate for class membership estimation. By changing k, the model can become more or less flexible. To classify a new data point, the algorithm computes the distance between the new point and all the other points in the training dataset, then selects the k nearest points and assigns the most prevalent classification in those points as a label to the new point (DREISEITL; OHNO-MACHADO, 2002). This method is shown in Figure 2.4.

2.1.1.4 C-Support Vector Machine (SVC)

SVCs are supervised learning models that perform both linear and nonlinear classification by implicitly mapping their inputs into high-dimensional feature spaces. The SVC utilises a'maximum-margin hyperplane' as a decision boundary to separate two different classes, defined as the ideal hyperplane that yields the maximum margin between the two classes, as depicted in Figure 2.5 for a linear case (JIANG; YAO, 2016). In general, a wider margin indicates less classification error. This approach replaces dot products with nonlinear kernel functions to match the maximum-margin hyperplane and wrap around features that are not linearly separable (NOVAKOVIC; VELJOVIC, 2011). SVC uses the decision

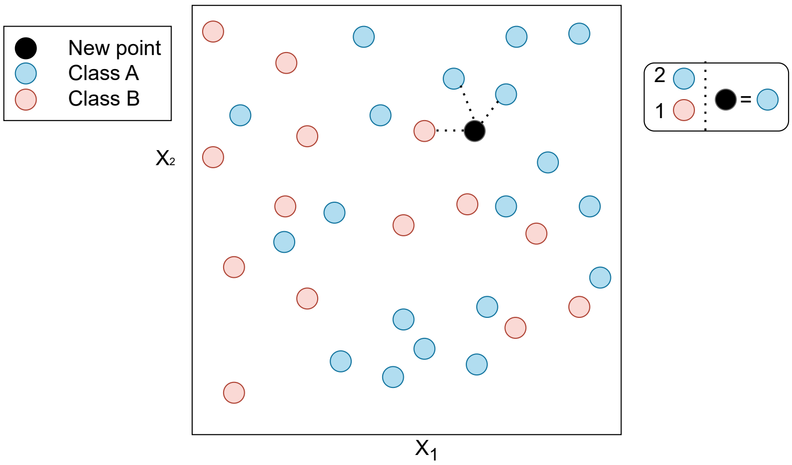


Figure 2.4: Decision process for a KNN, using three nearest neighbours (k) in a binary problem. Adapted from (GUO et al., 2003).

function in equation 2.9, where x is the input features, $alpha_i$ are Lagrange multipliers used to calculate the optimisation problem of maximising the margin, y are the labels, b is the bias, and $Ke(\mathcal{X}_i, \mathbf{x})$ is the kernel function.

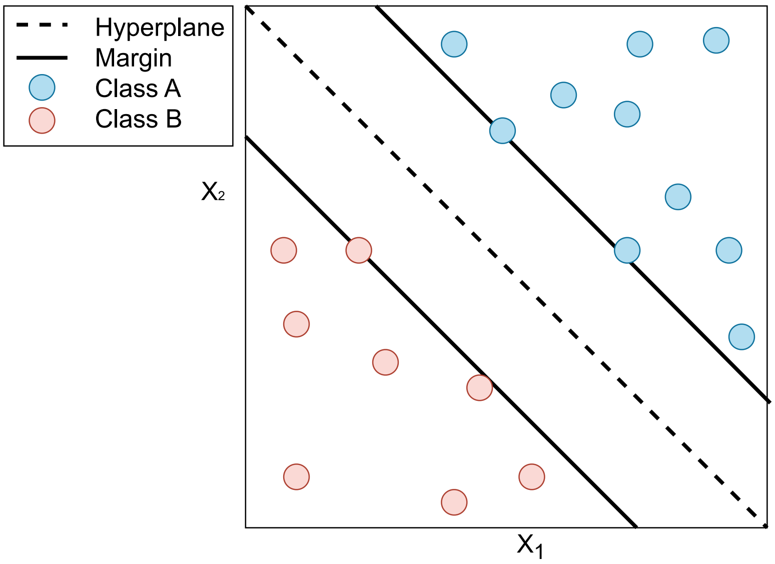


Figure 2.5: SVC hyperplane for a linearly separated scenario with two features. Adapted from (JIANG; YAO, 2016).

$$f(\mathcal{X}) = \sum_{i=1}^{n} \alpha_i y_i Ke(\mathcal{X}_i, \mathcal{X}) + b \tag{2.9}$$

The Radial Basis Function (RBF) kernel, used in this study, non-linearly maps objects into a higher-dimensional space and requires a smaller number of hyperparameters to train, alleviating the complexity of model selection (RAHMAWATI; HUANG, 2016). The

RBF is calculated using Equation 2.10, where σ is a parameter that regulates the width of the Gaussian function.

$$Ke(\mathcal{X}_i, \mathcal{X}) = e^{\left(-\frac{\|\mathcal{X}_i - \mathcal{X}\|^2}{2\sigma^2}\right)} \tag{2.10}$$

2.1.1.5 Random Forest Classifier (RF)

The decision tree classifier is the basis for the RF, which is a supervised learning method used for classification tasks. It works by recursively splitting the data space into subsets based on feature values, resulting in a tree-like model with simple decision rules inferred from the data features to predict the target classification.

RF is made up of a collection of tree-style classifiers, each of which is constructed by randomly sampling input features with identical distributions. Each tree independently casts a unit vote for the output, and the output with more votes is picked, resulting in an ensemble of decision trees to increase prediction performance, as shown in Figure 2.6. Equation 2.11 illustrates the technique, where $h_t(\mathcal{X})$ is the prediction of the t-th tree and T is the total number of trees in the forest (PAL, 2005; KULKARNI; SINHA, 2013).

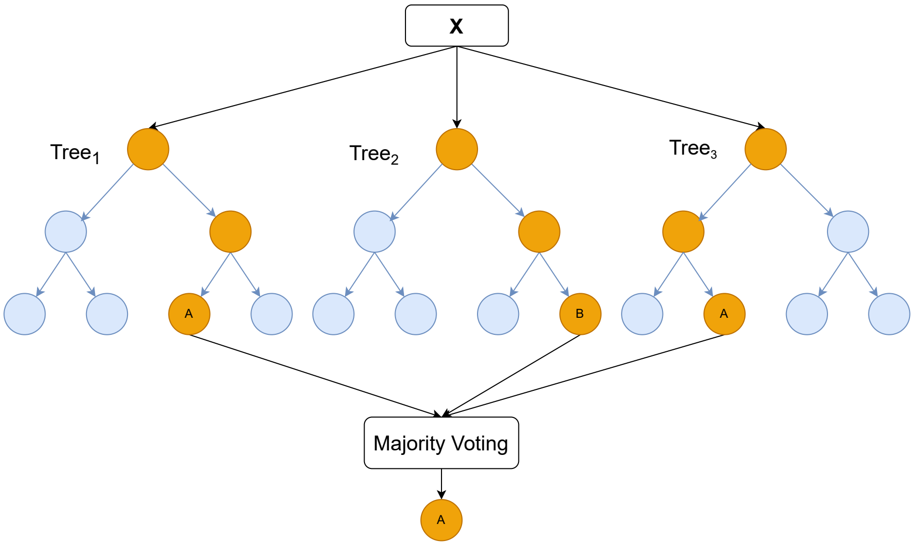


Figure 2.6: RF model representation of the voting system using three tree classifiers, with orange nodes representing the selected path for each tree. Adapted from (KIRASICH; SMITH; SADLER, 2018).

$$f(\mathcal{X}) = \sum_{i=1}^{n} \alpha_i y_i K(\mathcal{X}_i, \mathcal{X}) + b \tag{2.11}$$

2.1.2 Bias and unfairness in ML

Automated decisions can have an enormous impact in one's life. Whether considering school admittance, a job offer, or even a mortgage, arbitrary, inconsistent, or erroneous decision-making can adversely limit access to deserved opportunities (BAROCAS; HARDT; NARAYANAN, 2023).

Identifying essential characteristics for a decision might occur informally or unintentionally by associating specific features to improved performance based on prior knowledge. For example, an employer may see that those who study mathematics do better in financial occupations. ML analyses historical data to identify significant decision-making aspects that people may ignore, resulting in more complex relationships between observed data and outcomes (BAROCAS; HARDT; NARAYANAN, 2023; van Giffen; HERHAUSEN; FAHSE, 2022).

Even while ML achieved success in different areas of application, such as object recognition, where humans are unable to provide the whole set of rules to fulfil the goal, systematic biases were uncovered in several commercial ML models after they were put into use (MITCHELL et al., 2019). Examples include computer vision (PERERA, 2024), attribute detection (BUOLAMWINI; GEBRU, 2018), criminal justice (TRAVAINI et al., 2022), and toxic comment detection (PAGANO et al., 2023b). The problem originates from the fundamental manner that ML models learn by example, which involves generalising from many inputs rather than committing examples to memory. Learning patterns, details, and features to accomplish the required task on data samples similar, but never seen before (BAROCAS; HARDT; NARAYANAN, 2023).

These issues highlight the need of having a sufficient number of instances, a diverse set of examples, and a sufficiently well-annotated set of examples. These are required to train accurate, trustworthy, and fair ML models, but such data is rarely available and are costly to acquire (HE et al., 2022; BAROCAS; HARDT; NARAYANAN, 2023; KILKENNY; ROBINSON, 2018; GEIGER et al., 2020). When utilising machine learning to predict human conduct and traits, the unfairness problem is underlined since the historical examples presented will most often represent past prejudices against specific social groups, dominant cultural stereotypes, and demographic inequalities. As a result, the ML model will mimic the same dynamics identified in data patterns (BAROCAS; HARDT; NARAYANAN, 2023; LESLIE et al., 2023; BOOTH et al., 2021).

Fairness is a social term that is based on value judgements, hence there is no standardised definition. Fairness is a subjective appraisal that differs between cultures and societies (BOOTH et al., 2021). Consensus on a universal definition of fairness is a challenge in AI ethics and governance. It must be treated as a contextual and multivalent concept, manifesting from a diversity of factors, such as social, technical, sociotechnical, legislation, societal traditions, and ethical commitments, among others (LESLIE et al., 2023; JONES et al., 2020). To provide a quantitative approach to ML unfairness, the literature usually uses two broader definitions:

- Group unfairness.
- Individual unfairness.

For group unfairness, statistical metrics are employed to compare two separate groups based on different protected attributes. A protected attribute is any information that cannot be used to justify disparities in model results. In other words, a feature for which the model cannot be biassed. Protected attributes, often known as sensitive attributes, includes race, gender, religion, national origin, citizenship, pregnancy, disability status, genetic information, and many more characteristics. Each ML model has a set of protected attributes based on the application context (CASTELNOVO et al., 2022; KIM et al., 2021; PAGANO et al., 2023a). Hiding the protected attribute from the model is insufficient to ensure fairness, as the data may contain proxy variables for those attributes. For example, credit insurers may use credit information to price the insurance, leading to unintentional proxy discrimination against low-income or minority groups (PRINCE; SCHWARCZ, 2019).

Individual unfairness focuses on ensuring fairness at the individual level. Ensuring that people with comparable relevant qualifications have comparable system outcomes. To demonstrate how similar two people are in regard to a particular task, this method requires a similarity metric (LESLIE et al., 2023; CASTELNOVO et al., 2022; CHEN et al., 2023). Although the idea of individual unfairness is more powerful than that of group unfairness, both concepts often originate from the same fundamental principle: that people who are similar should be treated similarly (KLEINBERG; MULLAINATHAN; RAGHAVAN, 2016). However, individual unfairness cannot be applied when the objective is to address biases in the data; rather, it has relevance only when discrimination occurs during the decision-making process. On the other hand, group unfairness is quantifiable in a variety of ways and is dependent on the outcome statistics for the subgroups which are indexed in the data, allowing for the correction of data bias (PAGANO et al., 2023a).

Counterfactual unfairness is a variety of individual unfairness that characterises a result as fair if a choice made about a member of a sensitive group would have been the same if the member of the group had come from a different group in an alternate world. By incorporating causality into the analysis, it becomes transparent which elements affect the result (LESLIE et al., 2023). Because a change in a sensitive attribute could result in an unrealistic sample, the main disadvantage of this technique is the requirement to understand the causal linkages and relationships among the variables underlying the problem. For example, a male with the gender label reversed can have unrealistic (or less probable) features, such as height or hair length, therefore a change in these two variables with a specific probability is also required. It may be feasible if the number of relevant variables is small and the phenomenon is well understood, which is unlikely in complex social settings (CASTELNOVO et al., 2022).

2.1.2.1 Types of bias

Bias and unfairness are related concepts, nevertheless, they are distinct topics in machine learning. Bias occurs as a consequence of an unintentional error or systematic deviation in a system, and is not always a negative thing. Unfairness refers to the systematic prejudice or unequal treatment of individuals or groups (BOOTH et al., 2021; FERRARA, 2024). Both terms can be used together as bias and unfairness to express the biases that result in unequal or prejudicial treatment of individuals or groups.

Bias in an ML model's lifecycle can arrive from multiple places and exist in many forms. Two complementary views are the biases arising from the interactions of data, users, and algorithms (models) (MEHRABI et al., 2021), and the other view sees bias as four different and generalist groups defined as world, data, design, and ecosystem (LESLIE et al., 2023). Both representations are correct and exhibit comparable types of bias. Only the most pertinent bias types will be discussed here because there is a wide range of bias types and this study does not focus on redefining or proposing new types. It's also critical to remember that, because bias is ingrained in all machine learning models, bias is not equal to injustice. A model is considered unfair when there is bias towards a protected attribute.

Historical bias: already existing bias in the real world, normally in the form of social patterns of discrimination, social injustice, and discriminatory attitudes. The model propagates those historical biases that are present in the dataset.

Structural and institutional bias: is a type of discrimination that exists within an organisation and is brought about by the laws, guidelines, policies, and practices of a government or agency.

Temporal bias: show how cultures evolve over time due to variations in population and conduct; these variations may result in the favouring of one temporal behaviour over another.

Representation bias: associated with the demographic sample used in the data collection process. Minor details or subgroups that never make it into the dataset characterise non-representative samples, lacking diversity. It can also be caused by any broad presumptions regarding population groupings.

Selection bias: when bias develops as a result of the selection procedure used during the creation of the model or data. The selection process could involve the following: the study subjects choosing their own candidates; an inadequate model to address the issue; or an ambiguous label (target) choice.

Measurement bias: discusses the option for measuring or reporting the features that the model uses. It appears when the concepts being measured are not fairly and equally represented by the measuring criteria.

Emergent bias: occurs as a result of use and interaction with real users. This bias arises as a result of changes in population, cultural values, or societal knowledge. It's common in systems with user feedback, where complex interactions between the system, its users, and the environment often manifest due to unforeseen dynamics that occur post-deployment.

2.1.2.2 Protected Attributes

Protected attributes are demographic or personal features of individuals that cannot be used to make a decision. The protected attribute could be defined and protected as part of a legal mandates or because of organizational values. These attributes include, but are not limited to, race, religion, national origin, gender, marital status, age, disability status, and socioeconomic status. Ensuring fairness in algorithmic decision-making requires careful consideration of these attributes to prevent models from perpetuating or exacerbating existing disparities (BAROCAS; HARDT; NARAYANAN, 2023).

In many jurisdictions, the use of sensitive attributes in decision-making processes is heavily regulated. Including regulations and laws as: Penalties for discriminating in housing (U.S. Congress, 1968), Convention against discrimination in education (United Nations Educational, Scientific and Cultural Organization (UNESCO), 1960), Convention on the Elimination of All Forms of Racial Discrimination (United Nations General Assembly, 1965), Convention on the Elimination of All Forms of Discrimination against Women (United Nations General Assembly, 1979), and Convention on the Rights of Persons with Disabilities (United Nations General Assembly, 2006).

Furthermore, discrimination against members of a protected group could happen indirectly, when persons appear to be treated indifferently to the protected attributes but are still treated unfairly as a result of implicit protected attributes. For example, using a person's residential postcode can result in racial discrimination because the population of residential regions may be correlated with race. These indirect variables are also known as proxy variables (MEHRABI et al., 2021).

2.1.2.3 Concerns involving bias and unfairness

Different concerns concerning bias and unfairness in ML models have been raised in a variety of studies. Issues include a lack of transparency, explainability, regulation, and accountability. All those concerns are included in the broader concept of trustworthy AI, which can be defined as a set of overlapping properties: reliability, safety, security, privacy, availability, usability, accuracy, robustness, fairness, accountability, transparency, interpretability, explainability, and ethics. These properties apply to all stages of an AI system's life cycle. Improving trustworthiness in any one aspect demands improvements at several stages, while a violation of trust in any single property might damage the overall trustworthiness of the system. (LI et al., 2023; WING, 2021; LIANG et al., 2022).

Transparency and explainability in ML models is exceedingly difficult to achieve because the models contain millions of parameters with no clear indication of which parameter is responsible for the final result. The problem is exposed when huge organisations, such as Meta and Telegram, are not committed to publishing how the employed systems operate and what the constraints are. The models can only be studied by the team that produced them (AMMAR, 2019).

Transparency is defined as a combination of process and result transparency, referring to disclosing information regarding the model entire lifecycle, while explainability which refers to understanding the algorithm's basic characteristics as well as the judgements and patterns that led to the final categorization process. Local explanation can identify the most essential features for a specific decision, whereas global explanation assesses all decisions based on specific metrics (SEYMOUR, 2018; LI et al., 2023).

Models can be classified as white-box or black-box, depending on their availability and constraints:

- White-box: ML models produce straightforward results for application domain expertise. Typically, these models achieve an appropriate mix between accuracy and explainability. The structure and functionality of this model category are simple to change and analyse (LOYOLA-GONZALEZ, 2019).
- Black-box: ML models that are incredibly difficult for experts in the field to describe and understand. Changes to the structure of models in this category are limited,

making it difficult to grasp their structure and operation (LOYOLA-GONZALEZ, 2019).

There is still no clear and globally accepted definition of responsibility for AI systems, but it should include: fairness, safety, privacy, explainability, security, and reproducibility (NOIA et al., 2022).

To address security concerns in automated decision systems, data engineers are urged to develop a more fair and inclusive procedure. Automated decision systems must be responsible in development, design, application, and use, as well as strictly regulated and monitored to avoid perpetuating inequality (STOYANOVICH; HOWE; JAGADISH, 2020). The regulation should emphasise an obligation to minimise the risk of erroneous or biassed decisions in critical areas (NOIA et al., 2022).

Accountability addresses the regulation on AI systems improving legal and institutional norms on AI governance, requiring that the stakeholders of an AI system to be obligated to justify their design, implementation, and results. It could also require the auditability of systems, requires the justification of a system to be reviewed, assessed, and audited by a third party (LI et al., 2023).

Opposition to the use of models for decision-making emphasises the existence of advantages and disadvantages. Implementation of these tools can introduce new uncertainties, disruptions, and risks to already critical scenarios, such as strategic governance, with concern about ethical aspects (DWIVEDI et al., 2021; BOOTH et al., 2021; KÖNIG; WENZELBURGER, 2021).

Governments are experimenting with ML to boost efficiency in large-scale personalisation of services based on citizen profiles, such as predicting viral outbreaks, crime hotspots, and food safety inspections (DWIVEDI et al., 2021). Bias can produce governance challenges in those circumstances, endangering society and sustaining previous prejudice and unwanted habits. The discussion should include technology diplomacy as a facilitator of global policy alignment and governance (FEIJÓO et al., 2020). Finally, bias in ML systems can be regarded as a censoring act because the algorithms usually neglect unusual information; for example, religious content is censored as an unintentional consequence of counterterrorism models (AMMAR, 2019).

2.2 *Literature review*

Few works try to select representative fairness metrics, or to reduce the number of metrics to evaluate a model.

Correlation between metrics was used in a word embedding application to evaluate and compare fairness metrics, in the proposed Word Embeddings Fairness Evaluation (WEFE) framework. Using pre-trained word embedding models, WEFE calculates an unfairness-based ranking of these models by encapsulating existing fairness metrics. After ranking the metrics in different models and sensitive attributes, the correlation between rankings is used, and similarities are shown only for the gender attribute. As a result, the correlation is substantially weaker when we consider ethnicity and religion bias, indicating that more work is needed to propose fairness metrics able to consistently rank embeddings for dimensions beyond gender (BADILLA; BRAVO-MARQUEZ; PéREZ, 2020). Furthermore, the application does not provide a method for selecting a representative metric, and only four fairness metrics associated with the word embedding context are employed.

A decision-tree style framework for choosing fairness metrics relevant to recommendation and ranking systems was used to assist in the decision-making process of choosing the appropriate metric within a certain context. The framework is a decision-making structure that is specifically intended to scope measurable harms and corresponding metrics from a predefined potential harm to the system. To develop the framework, capturing the complex challenges practitioners face when scoping and identifying fairness metrics, a series of interview and literature reviews were conducted, using the decision tree to correspond fairness metrics to the possible harms that could emerge from a conceptual system (SMITH; BEATTIE; CRAMER, 2023). This strategy relies solely on the literature regarding the validation of the chosen metrics through interviews and is restricted to selecting a single metric through a series of questions concerning the problem's unfairness, based on the user's knowledge about the system.

The analyses of the relationship between sensitive attributes and fairness metrics is explored by evaluating models in three different application areas using the same sensitive attribute. Namely, a computer vision, natural language processing, and recommendation systems were trained and evaluated to the gender attribute. It used a facial recognition detection, message toxicity detection, and movies recommendations cases studies. Most of the fairness metrics had the same behaviour in all models, only two metrics expressed a different measurement between the cases studies. The results reveal that, regardless of the model, the sensitive attribute predicts a similar behaviour for the metrics, with the sensitive attribute serving as an indicator of the measure to be employed (PAGANO et al., 2023b). The analyses lack testing with different sensitive attributes, and needs uniformity in the classes for the sensitive attribute, since one of the datasets, in addition to the male and female groups, also had transgender and other genders.

Another strategy attempts to group the metrics together for every possible scenario, using a threshold to transform unfairness in a binary problem, calculating the dissimilarity between them using a hierarchical clustering technique. In experiments with 26 metrics in 4 datasets, they aim to analyse if the fairness metrics agree with each other, if they can be grouped, how sensitive to change they are, and if it's possible to achieve fairness in all of them at the same time. Proposing a framework to run metrics on real-world data, find clusters of correlated metrics and prune insensitive clusters. Agglomerative clustering is used across all the data to grouped together pairs of metrics, creating a dendrogram of the between-cluster dissimilarity. The finding showed that the metrics disagree when labelling a model as fair or unfair, metrics can be clustered together based on how they measure bias, clusters could be ignored when changes to the data did not change the unfairness scores, and it is not possible to satisfy all the clusters (MAJUMDER et al., 2023). This approach has potential drawbacks when applied to other areas, since it overlooks the context created by the interaction of data, model and fairness metrics.

Some ideas for aggregating the metrics into representative groups rely on similarities calculated using correlation, where a low variance estimator is used to group tightly related metrics under model and data dependencies. A framework to identify a small subset of fairness metrics that are representative of the broader set is proposed, using correlation-Based Selection of metrics, through a Monte-Carlo Sampling algorithm, responsible to estimate the fairness metrics in a sampled ML model (ANAHIDEH; NEZAMI; ASUDEH, 2021). Although this approach is relatively straightforward, there are lots of caveats, which include the inability to capture fine-grained aspects such as confounding (SU et al., 2022). This work improves over this approach, providing a solution for the caveats, with a more reliable validation and a more efficient method.

Chapter Three

Material and Methods

Compiling with specific objective 1, this work focus on *group unfairness*, a broad type of unfairness that is measured via the distribution of a model outcome over different, often dichotomous, groups of interest (CATON; HAAS, 2024). In this context, for a classification problem with a dataset of N observations, $(\mathcal{X}, \mathcal{S}, \mathcal{Y})$, defined via a model h mapping $(\mathcal{X}, \mathcal{S}) \rightarrow \mathcal{Y}$, we have that $\mathcal{X}$ is the set of non-sensitive independent variables[1]; $\mathcal{S}$ is the set of sensitive ones, and $\mathcal{Y}$, in turn, is the set of dependent (target) variables. The set $\mathcal{S}$ acts as the conceptual enabler for obtaining a measure of unfairness across the groups. For simplicity, assuming a dichotomous property of a single sensitive variable $S \in \mathcal{S}$, i.e., $S \in \{0, 1\}$, the data can be divided into *Privileged* ($S = 1$) and *Unprivileged* ($S = 0$) sensitive groups (e.g., gender = {male, female}). For each such group s, a metric f_s^j is chosen from a set $\mathcal{F} = \{f_i\}_{i=1}^{M}$ of M fairness metrics.

A common structure across metrics, particularly non-causal ones, is the strong reliance on misclassification rates such as True Positive (TP), False Negative (FN), False Positive (FP), and True Negative (TN) — all elements of a confusion matrix diagnostic common, but not limited to, binary classification settings. The rates are then combined to produce informative diagnostics of the model h tendency towards the sensitive groups (PAGANO et al., 2023a). For example, a notion of positive outcome parity can be achieved when equating the rates w.r.t of the total group observations $N_{S=0}$ and $N_{S=1}$:

$$\frac{TP_{S=1} + FP_{S=1}}{N_{S=1}} = \frac{TP_{S=0} + FP_{S=0}}{N_{S=0}} \tag{3.1}$$

Special combinations of the error rates give rise to different unfairness perspectives (Table 3.2). We acknowledge that dataset properties (sampling scheme, organic relationship of entities being modelled, etc.) and the choice of model (affected by the model's inductive bias) are key factors impacting unfairness. Therefore, in our analysis, following the prior report (ANAHIDEH; NEZAMI; ASUDEH, 2021), we rely on the concept of *context*, which is the environment specified by at least three components: (a) the available data, (b) the standing notion of data generation mechanisms, and (c) the model type under use.

Deciding which metric should be used in a problem requires thoughtful considerations and often the support of practitioners with deep experience in algorithmic unfairness, which may be a rare skill. Next, we detail a basic correlation-based strategy, from which we

[1] Variables are often referred to as attributes in the unfairness literature.

provide further development.

We investigate the usefulness of choosing a representative subset of fairness metrics based on similarities measured via correlation in a given context (*model + data*). Formally, for a set of fairness metrics $\mathcal{F}$, one wants to find a representative subset $R_{\mathcal{F}}$ of much lower cardinality, i.e. $|R_{\mathcal{F}}| << |\mathcal{F}|$. *High* correlation is the element bounding the metrics in the two sets. That is, $\forall f_i \in \mathcal{F}$, *there exists a fairness metric* $f_j \in R_{\mathcal{F}}$ *such that the correlation between* f_i *and* f_j *is high* (ANAHIDEH; NEZAMI; ASUDEH, 2021).

3.1 Base Experiment

Next, we describe a base approach for selecting a subset of representative metrics through sampling and correlation, highlighting the potential problems with the approach.

3.1.1 Bootstrap sampling via Markov chain Monte Carlo (MCMC)

Correlation is a common idea for modelling the relationship between different metrics. In (ANAHIDEH; NEZAMI; ASUDEH, 2021), the authors propose a simple sampling strategy to estimate the correlation between the metrics in $\mathcal{F}$. The idea anchors on the ability to sample different model performances for the same dataset and, ultimately, from such performances, on the computation of aggregated values of each $f_i \in \mathcal{F}$ under consideration. The estimation process takes place following the well-known *Pearson Correlation Coefficient*:

$$r_{ij} = \frac{\sum_{k=1}^{K}(f^s_{i,k} - \overline{f}^s_i)(f^s_{j,k} - \overline{f}_j)}{\sum_{k=1}^{K}(f^s_{i,k} - \overline{f}^s_i)^2 \sum_{k=1}^{K}(f^s_{j,k} - \overline{f}_j)^2} \tag{3.2}$$

Where the sampling process samples N classifiers instances h_k for $k \in \{1, ..., N\}$, from which fairness metric i values, $f_{i,k}$, are calculated for each classifier h_k.

For obtaining robust estimation, the authors repeat the computation in (3.2) L times ($L = 30$), producing $r^{\ell}_{i,j}$ results. The final correlation estimate is then averaged following standard guarantees given by the central limit theorem:

$$r^*_{i,j} = \frac{1}{L}\sum_{\ell=1}^{L} r^{\ell}_{i,j} \tag{3.3}$$

The performance of a model for the different (un/privileged) groups depends on the ratio of available samples (the class balance) and the distribution of their label values. To tackle this factor and mitigate the effects stemming from the noise in the dataset, an option taken by the authors was to consider training the models based on subsamples from the dataset in a process similar to cross-validation. Specifically,(ANAHIDEH; NEZAMI; ASUDEH, 2021) used bootstrapping sampling, a process similar to random sampling with replacement. Assuming K to be the number of drawn bootstraps samples given a dataset D. They aimed to construct smaller representative subsets of D for model training. For a binary classification problem with two groups g_1 and g_2, a vector $\boldsymbol{w} = \{w_1, w_2, w_3, w_4\}$, such that $\sum_1^4 w_i = 1$, of the proportion of samples in each cell can be constructed. For completeness, we recall the process as follows:

1. Draw a vector from w uniformly from $\boldsymbol{w}$.
2. Bootstrap $w_i \times K$ samples from the samples of D that belong to the cell i of the table to form the bootstrapped dataset B_j.
3. Use the dataset B_j to train the sampled classifier h_j.
4. Evaluate the model to compute the values f_{kj}, for each fairness metric, $f_k \in \mathcal{F}$ and return the vector $\{f_{k1}, ..., f_{km}\}$.

The computed fairness metric in the above step 4 is then used to compute the correlations via Equations (3.2) and (3.3).

3.1.2 Base Experiment Caveats

In the base experiment, the fairness metrics are grouped by their Pearson's correlation using a computationally costly method, which triggers some considerations:

1. The base experiment generates different probabilities for each combination of samples at random and uniformly, which does not ensure that data group biases are properly represented in the samples.
2. The computational efficiency of Monte Carlo methods is widely criticised for their high computational cost (SHIELDS et al., 2015), so any reduction in the quantity of models to be retrained is ideal.
3. Pearson's correlation cannot be used with non-normal data (SNEDECOR; COCHRAN, 1980), so another coefficient must be used to calculate the correlation.

4. More explicit tests to validate the experiment are needed to prove the relevance and veracity of the results obtained.

The proposed solutions for caveats 1 to 3 are adopted in the proposed experiment, detailed in the following section. Caveat 4 is tackled in Section 3.2.4.4.

3.2 Proposed Experiment

In this section, we describe our proposed approach for selecting a subset of representative metrics, starting with the base experiment and modifying it to address the aforementioned caveats, and compiling with specific objective 3.

3.2.1 Stratified Sample

Addressing the first correlation caveat in Section 3.1.2, we divided the dataset based on their sensitive attribute ($\mathcal{S}$) and the dependent (target) variables ($\mathcal{Y}$), using stratified sampling, to avoid problems with underrepresented minority classes and ensure credibility for the used fairness metrics.

The stratified sampling technique divides the population into distinct groups, using features that express the overall situation of each group. The dataset is then sampled proportionally to the size of each group, having less variance than random sampling, since samples within a group are more similar than samples from different ones (SHIELDS et al., 2015; MOORE; NOTZ; NOTZ, 2006).

Stratified sampling has been effectively applied to classification problems involving minority and infrequent classes. For example, stratified sampling designs are used in land cover classification map models to handle classes with little coverage (such as water and wetlands), lowering the margin of error for the area determined by pixel counting (PUERTAS; BRENNING; MEZA, 2013). When categorising texts with imbalances, stratified sampling is used to ensure that the features chosen in the subspace are more balanced and informative for both the minority and the majority classes. This improves the classification performance of Random Forest models when compared to the random sampling version, especially for minority classes with a very small number of instances (WU et al., 2014).

For extremely unbalanced datasets with few minority samples, the random sampling method would result in a small, potentially zero, number of minority samples in the test set, rendering the evaluation metric applied to it invalid. The stratified sampling method for splitting the data is used in order to preserve the imbalance ratio of the original dataset in the test set. This increases the assessment metric's trustworthiness and prevents a situation in which it cannot be calculated from the test set (HU et al., 2022).

3.2.2 Stopping Criterion

Addressing the second correlation caveat in Section 3.1.2, we propose a stopping criterion for the Monte Carlo method that reduces the amount of model training required to find the representative metrics, thus reducing the computational cost.

The computational cost reduction matters because Monte Carlo is costly yet a reliable and efficient method of assessing uncertainty in computer analysis. As a result, there has been a great deal of interest in improving the effectiveness of Monte Carlo techniques (SHIELDS et al., 2015). To mitigate the problem, it is necessary to improve the sample routine's convergence rate and determine when to terminate the MCMC algorithm (ROY, 2020; SHIELDS et al., 2015).

The proposed stopping criteria employ the median absolute deviation (MAD) metric, which is a robust estimator for obtaining the standard deviation of a non-normal distribution (ASLAM et al., 2019), shown in equation 3.4, where f_i represents a metric value. Given our experimental results in Chapter 4, there is no guarantee that the metrics will follow a normal distribution; therefore, MAD is used as a non-parametric statistical metric. The MAD of each fairness metric from the trained models reaches the early stopping criteria when two conditions are met, stopping the model creation process. The first condition is the maximum MAD threshold, achieved when the MAD of a fairness metric is less than the defined threshold. The second condition is the minimum number of fairness metrics that achieve the first condition. To determine the experiment thresholds, different values were tested, and the pair with the highest mean correlation from a represented metric to its representative was chosen, with the conditional that it did not stop on the first or last model possible, as this is an indication that the threshold is too rigorous or lenient.

To extrapolate the threshold values for different contexts, the selected values are a good starting point, but they should be fine-tuned to the desired application while analysing the mean correlation.

$$MAD = \frac{1}{n}\sum_{i=1}^{n}|f_i - median| \tag{3.4}$$

The stopping criteria are considered from the third iteration onward, where the MAD metric presents a more stable behaviour.

3.2.3 Correlation Method

To address the third correlation caveat in section 3.1.2, we used Kendall's *tau* correlation to fill the Pearson's correlation gaps shown in the previous subsection. This type of coefficient is well-known for being recommended for non-normal data and small datasets (CHOK, 2010). It is based on the rank of the data, allowing it to deal with non-linearity (BRUCE; BRUCE, 2018). Kendall's *tau* calculation is shown in Equation 3.5, where C are concordant pairs and DI are discordant pairs, so τ will range between -1 and +1.

$$\tau = \frac{C - DI}{n(n-1)/2} \tag{3.5}$$

3.2.4 Experiments Setup

Our experimental setup generates the experimental result using a pre-selected set of datasets, models, and metrics, tackling the specific objectives 1 and 2, as well as four approaches to validate different aspects of the chosen metrics, compiling with specific objective 4.

3.2.4.1 Dataset

The results are based on the benchmark datasets commonly used for unfairness problems (ANAHIDEH; NEZAMI; ASUDEH, 2021; MAJUMDER et al., 2023), involving four case studies in the fields of criminal judgement, bank loans, demographic census, and advertisement. The obtained data were pre-processed prior to the experiments, transforming the values of each protected attribute into categorical numeric-type data. Figure 3.1 shows the data distribution in each dataset used.

The Adult (DUA; GRAFF, 2017), also known as Census Income, has 48,842 total data samples and attempts to predict whether annual income exceeds $50,000. Individual information such as age, education, occupation, relationship, race, gender, capital gain, and capital loss were extracted from the 1994 Census database. In this case, the target attribute is *income*, and the protected attribute under consideration is *race*. Figure 3.1 depicts the distribution of groups in this dataset in terms of output classes.

The Bank Marketing (DUA; GRAFF, 2017) contains data directly related to market campaigns in phone calls from a Portuguese bank between 2008 and 2013 (MORO; CORTEZ; RITA, 2014) and attempts to predict whether the customer has signed a term deposit. This dataset contains 45,211 tuples with individual data such as age, job type, marital status, education, and personal loan. The target attribute is y, a binary attribute describing if the client subscribed to a term deposit, and the sensitive attribute in analysis is *age cat*. The distribution of the sensitive groups in this dataset in relation to the output classes is depicted in 3.1.

The Compas was published by ProPublica and is a database that contains criminal and demographic information, including criminal history, length of time in prison, gender, age group, and race. Analysing a study by ProPublica the algorithm developed with this database became biassed in favour of white defendants and against black defendants when taking into account the *two year recid* attribute (LARSON et al., 2016). The distribution of the sensitive groups in this dataset with respect to the output classes is depicted in 3.1

The German Credit Data (DUA; GRAFF, 2017) is a dataset designed to determine whether a person has good or bad credit risk. This dataset contains 1000 tuples of information gathered by the University of Hamburg, including personal information such as age, job, credit history, loan purpose, marital status, and gender. The attributes *risk* and *sex* are to be considered sensitive in analysis. Figure 3.1 shows how the sensitive groups in this dataset are distributed in relation to the output classes.

All datasets have data imbalances and disparities, including sensitive attributes with fewer data samples than others and samples with varying proportions of the target class. These disparities represent inherent biases in the data and do not necessarily imply that the data is unfair, as long as they accurately reflect the true context from which the data originated. Nonetheless, these disparities should not be replicated by the ML models trained on them.

3.2.4.2 Models

We experimented with five classifiers: MLP, Logit, SVC, KNN, RF.

The developed experiment is sensitive to several hyperparameters that are required for experiment replication. For the model implementation, we used the popular `scikit-learn` library (`v1.1.1`) (PEDREGOSA et al., 2011). Table 3.1 highlights the hyperparameters used in the models; default parameters are used for the ones not described in it.

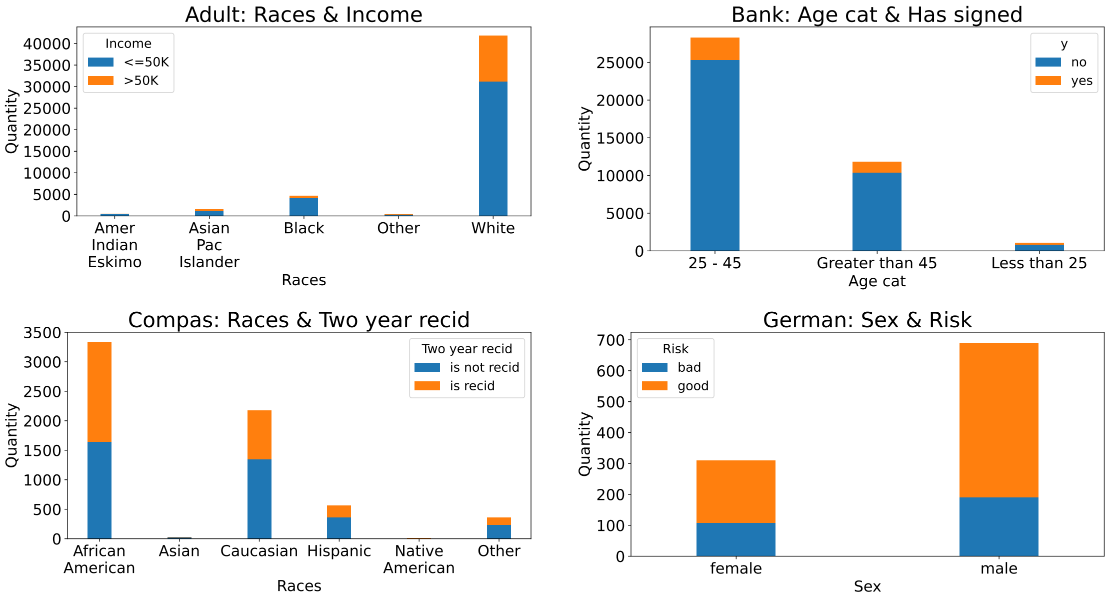


Figure 3.1: Distribution of the sensitive groups in each dataset in relation to their target values. The target values are: Adult: Income over or under $50000; Bank: signed or not the bank deposit; Compas: Relapse or not into criminal behaviour in two years; German: Has a good or bad credit risk score.

Table 3.1: Models parameters for each classifier

Logit	RF	SVC	MLP	KNN

solver='liblinear'	max_depth=2	gamma='auto'
		kernel='rbf'
			hidden_layer_sizers=(15,)
			warm_start=True
			activation='relu'
			solver='adam'
				n_neighbors=2

3.2.4.3 Metrics

Table 3.2 lists the fairness metrics used in the experiments, the list contains commonly used fairness metrics, however, any metrics based on misclassification rates might be adopted.

3.2.4.4 Validation

To address the fourth correlation caveat in Section 3.1.2, the experiments were validated using the four criteria listed below:

1. Analysis of the number of models and subsets: The experiment was run for each combination of datasets and classifiers, using both the base and proposed experiment, for a total of 40 executions, with the upper bound for the number of models

Table 3.2: Metrics used as unfairness criteria with the respective formulation.

Metric	*Formulation*
Equalized Odds Values (HARDT; PRICE; SREBRO, 2016)	$\frac{1}{2} * \left(\left| \frac{FP_0}{FP_0+TN_0} - \frac{FP_1}{FP_1+TN_1} \right| + \left| \frac{TP_0}{TP_0+FN_0} - \frac{TP_1}{TP_1+FN_1} \right| \right)$
Error Difference (BERK et al., 2021)	$\frac{FP_0+FN_0}{N_1+N_0} - \frac{FP_1+FN_1}{N_1+N_0}$
Error Ratio (BERK et al., 2021)	$\frac{\frac{FP_0+FN_0}{N_1+N_0}}{\frac{FP_1+FN_0}{N_1+N_0}}$
Discovery Difference (VERMA; RUBIN, 2018)	$\frac{FP_0}{TP_0+FP_0} - \frac{FP_1}{TP_1+FP_1}$
Discovery Ratio (VERMA; RUBIN, 2018)	$\frac{\frac{FP_0}{TP_0+FP_0}}{\frac{FP_1}{TP_1+FP_1}}$
Predictive Equality (CORBETT-DAVIES et al., 2017)	$\frac{FP_0}{FP_0+TN_0} - \frac{FP_1}{FP_1+TN_1}$
False Positive Rate (FPR) Ratio (VERMA; RUBIN, 2018)	$\frac{\frac{FP_0}{FP_0+TN_0}}{\frac{FP_1}{FP_1+TN_1}}$
False Omission Rate (FOR) Difference (VERMA; RUBIN, 2018)	$\frac{FN_0}{TN_0+FN_0} - \frac{FN_1}{TN_1+FN_1}$
False Omission Rate (FOR) Ratio (VERMA; RUBIN, 2018)	$\frac{\frac{FN_0}{TN_0+FN_0}}{\frac{FN_1}{TN_1+FN_1}}$
Disparity Impact (PPR Ratio) (FELDMAN et al., 2015)	$\frac{\frac{TP_0+FP_0}{N_0}}{\frac{TP_1+FP_1}{N_1}}$
Statistical Parity (DWORK et al., 2012)	$\frac{TP_0+FP_0}{N_0} - \frac{TP_1+FP_1}{N_1}$
Equal Opportunity (HARDT; PRICE; SREBRO, 2016)	$\frac{TP_0}{TP_0+FN_0} - \frac{TP_1}{TP_1+FN_1}$
False Negative Rate (FNR) Difference (VERMA; RUBIN, 2018)	$\frac{FN_0}{FN_0+TP_0} - \frac{FN_1}{FN_1+TP_1}$
False Negative Rate Ratio (VERMA; RUBIN, 2018)	$\frac{\frac{FN_0}{FN_0+TP_0}}{\frac{FN_1}{FN_1+TP_1}}$
Average Odd Difference (MANDHALA; BHATTACHARYYA; MIDHUNCHAKKARAVARTHY, 2022)	$\frac{1}{2} * \left(\frac{FP_0}{FP_0+TN_0} - \frac{FP_1}{FP_1+TN_1} + \frac{TP_0}{TP_0+FN_0} - \frac{TP_1}{TP_1+FN_1} \right)$
Predictive Parity (VERMA; RUBIN, 2018)	$\frac{TP_0}{TP_0+FP_0} - \frac{TP_1}{TP_1+FP_1}$

created being 1200 (30 × 40), if not for the early stop in the proposed method.

The number of models and subsets analysed compares the impact of the proposed modification over the base experiment on those quantitative measures. Furthermore, the experiment's stability is investigated by repeating it 30 times and analysing the results' box plot, both for the number of models and the number of subsets. Additionally, the mean execution time between both experiments are compared.

2. Analysis of metric subsets: The analysis of established metrics subsets seeks to validate each group of metrics formed by examining the correlation between the represented metrics and the group's representative, their similarities in relation to their equations, and, finally, whether the grouped metrics are incompatible.

   Definitions from the impossibility theorem were used to validate their compatibility (KLEINBERG; MULLAINATHAN; RAGHAVAN, 2016) and were deepened by the studies of (KIM; CHEN; TALWALKAR, 2020) and (GARG; VILLASENOR; FOGGO, 2020). These works talked about metrics that cannot be satisfied simultaneously, resulting in a trade-off in their performances; therefore, those metrics should not be grouped together. The following incompatible metric sets were used:

   (a) Statistical Parity, Equalized Odds and Predictive Parity.

   (b) Predictive Equality, False Negative Ratio, Predictive Parity.

3. Subset recurrence analysis: Seeks to identify common behaviours, grouping patterns, and generalisations among the experiments by examining the number of times two metrics appear in the same group, the number of times each single group appears, the frequency at which a metric is used as a representative, and the frequency at which a metric represents another one. We also analyse the similarity in the metrics equations to find underlying behaviours that can be justified by them.

4. Group similarity analysis: Used to determine how similar the metrics within a group are and to compare experiments. This is performed by determining the mean absolute distance between the representative and represented metrics in a group. To further examine the results, we apply alternative contextual viewpoints to the model and datasets, fixing one while measuring the mean value of the other.

   To calculate the similarity, each classifier must be trained on the complete dataset. Then, we measure the classifier's fairness metrics and calculate similarity using the previously picked groups, which indicates how close the metrics in the group are to the chosen representative. Equation 3.6 represents the mean absolute distance between two metrics. Where n_s is the number of sensitive groups (excluding the benefited), $f_s^k(representative)$ is an fairness metric k chosen as the representative metric, and $f_s^j(represented)$ is an fairness metric j, in the same group as k.

$$\underset{Similarity}{Metric} = \frac{\sum_{s=0}^{n_s} \left| f_s^k(representative) - f_s^j(represented) \right|}{n_s} \tag{3.6}$$

Chapter Four

Results and Discussions

In the following sections, we present our experiment results in terms of the evaluation points detailed in section 3.2.4.4.

4.1 Number of models and subsets

4.1.1 Results

The goal of this comparison is to examine the differences from the base experiment to the proposed experiment, which added the stratified sampling strategy, the Kendall correlation, and the early stopping criteria using MAD. To define the early stop threshold, different pairs of the maximum MAD value and the minimum number of metrics under the MAD value were used, ranging respectively from 0.05 to 0.95 with a step of 0.1 and from 4 to 16 with a step of 1, resulting in 130 combinations that were chosen based on the highest mean correlation from all the represented to representative metrics in the experiment. Pairs that resulted in a one model difference from the minimum or maximum mean number of models were also excluded, as they were too rigorous or lenient. The chosen pair was 0.35 and 10, as shown in Table 4.1, and they were used in all the results. As a reference, the mean correlation from the base experiment was 0.760. The metric groups for the proposed and base experiment can be seen from figures S.1 to S.20 and from figures S.21 to S.40 respectively.

Tables 4.2 and 4.3 compare the results based on the number of models generated and the number of subsets. Figure 4.1 illustrates the number of models for each of the proposed method's cases, while Figure 4.2 illustrates the number of subsets produced for the base and proposed method's cases, both figures include the quartile interval, median, lowest, and maximum values, based on the results of 30 executions. Finally, Table 4.4 displays the average time required to choose the representative metrics for each case study, as well as the difference between the proposed and base method, taking into account 30 repeated executions.

Table 4.1: Top 10 results of different threshold parameters for the early stop. Sorted by mean correlation and with the values within one under or over the maximum and minimum number of models excluded. In green is the chosen parameter.

Ranking	Maximum MAD	Minimun Number of Metrics	Mean Models Used	Variation Models Used	Mean Subsets Created	Variation Subset created	Mean Correlation	Variation Correlation
1	0.350	10	11.350	119.924	5.800	3.011	0.755	0.004
2	0.250	14	28.000	41.684	6.350	2.766	0.750	0.003
3	0.350	6	4.300	0.642	5.600	2.147	0.750	0.003
4	0.250	4	5.050	2.997	5.800	3.221	0.749	0.003
5	0.450	14	14.900	116.305	6.000	2.842	0.748	0.002
6	0.150	4	15.950	128.050	6.350	2.976	0.746	0.002
7	0.550	8	4.250	0.829	5.400	3.410	0.744	0.005
8	0.050	5	28.700	33.800	6.350	2.450	0.743	0.003
9	0.350	13	20.850	103.503	5.950	3.629	0.742	0.004
10	0.150	8	22.200	108.168	6.050	3.313	0.742	0.003

Table 4.2: The number of models used in each dataset for the proposed experiment, and the percentage reduction*.

Dataset	Number of models ↓ (reduction%↑)				
	KNN	MLP	SVC	RF	Logit
Adult	4(86.66%)	4(86.66%)	4(86.66%)	4(86.66%)	5(83.33%)
Bank	30(0.00%)	27(10.00%)	27(10.00%)	30(0.00%)	30(0.00%)
COMPAS	4(86.66%)	4(86.66%)	4(86.66%)	4(86.66%)	30(0.00%)
German	8(73.33%)	6(80.00%)	5(83.33%)	6(80.00%)	4(86.66%)

*The number of models in the base experiment is always 30.

Table 4.3: Number of subsets achieved in each experiment, separated by model, comparing the proposed and base experiment.

Experiment	Dataset	Number of subsets ↓ *				
		KNN	MLP	SVC	RF	Logit
Base	Adult	6	**6**	**7**	7	**6**
Proposed	Adult	6	7	9	7	7
Base	Bank	**3**	**6**	**4**	**3**	**3**
Proposed	Bank	7	8	7	6	7
Base	COMPAS	**4**	**3**	**5**	**3**	6
Proposed	COMPAS	5	6	7	5	**5**
Base	German	3	6	3	6	5
Proposed	German	3	**3**	3	**4**	**4**

*Values considered for tau = 0.5.

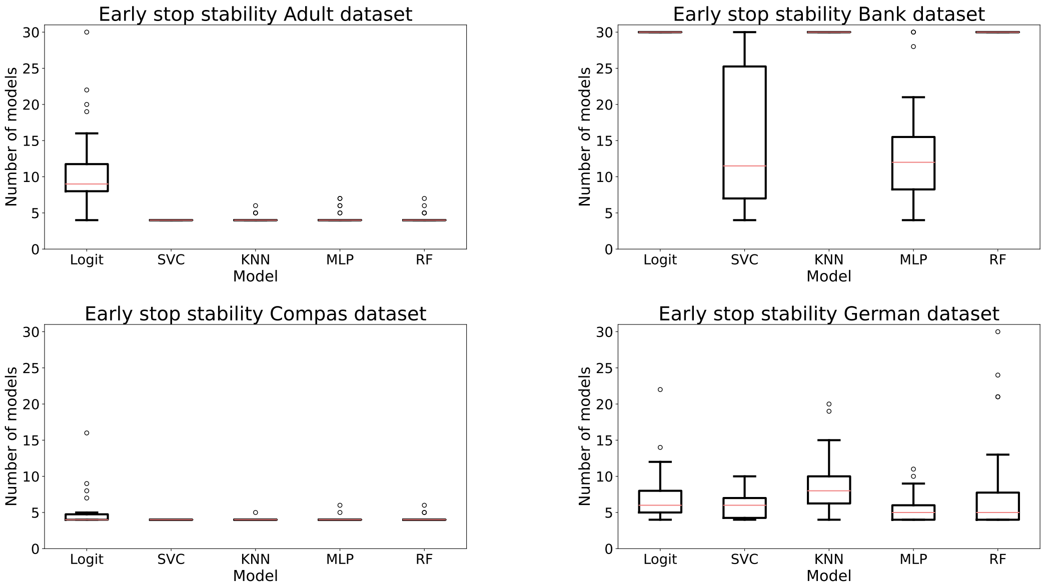


Figure 4.1: Box plot of the stopping criteria for each dataset and model.

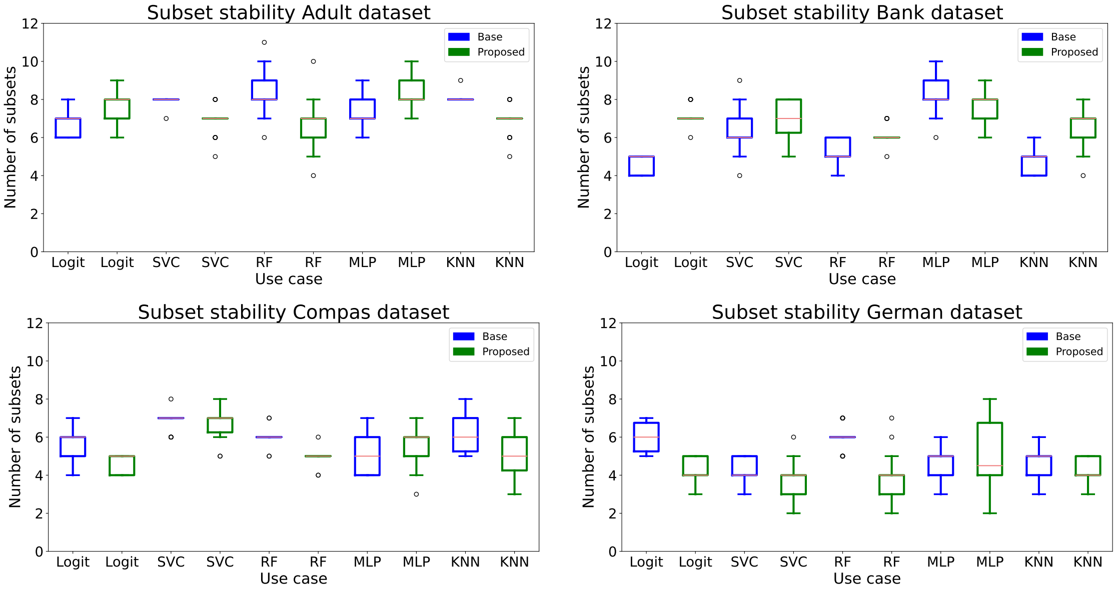


Figure 4.2: Box plot of subsets for each dataset and model.

Context	Base mean execution time(s)	Proposed mean execution time(s)	Percent Difference	*Legend
Adult_KNN	13.799	7.807	-43.41%	-65.00%
Adult_Logit	6.714	7.531	12.15%	-30.00%
Adult_MLP	14.028	5.628	-59.87%	0.000%
Adult_RF	11.266	6.616	-41.27%	30.00%
Adult_SVC	13.112	6.382	-51.32%	65.00%
Bank_KNN	17.955	22.093	23.04%	
Bank_Logit	6.767	11.133	64.51%	
Bank_MLP	10.671	14.276	33.78%	
Bank_RF	11.864	8.340	-29.701%	
Bank_SVC	14.054	11.573	-17.65%	
Compass_KNN	5.810	3.699	-36.32%	
Compass_Logit	5.679	4.679	-17.61%	
Compass_MLP	11.487	4.621	-59.76%	
Compass_RF	8.742	4.632	-47.02%	
Compass_SVC	7.625	4.077	-46.52%	
German_KNN	5.051	4.605	-8.82%	
German_Logit	5.230	4.644	-11.19%	
German_MLP	7.109	6.119	-13.93%	
German_RF	7.235	4.544	-37.20%	
German_SVC	6.958	4.988	-28.31%	
Mean	9.558	7.399	-20.82%	

Table 4.4: Comparison of mean execution times over 30 executions for metric selection. The Percent Difference shows the relative change in execution time for the proposed method compared to the base method. Negative values indicate a decrease in execution time, while positive values indicate an increase.

4.1.2 Discussion

Table 4.2 outlines that the stopping criteria optimised the number of models generated in most cases, with the worst results for all models using the Bank dataset and the Logit model using the Compas dataset. The average reduction of models generated was 64.37%. Furthermore, Table 4.3 shows the number of subsets for each dataset and model combination, with the proposed experiment presenting an overall higher number of subsets for the Bank and Compas datasets while remaining close to the base result for the Adult and being smaller for the German datasets.

When the number of models was reduced, less computational work was required, resulting in competitive outcomes when compared to the baseline study, which always needs 30 models for every execution, as seen in Figure 4.1. The stopping criterion varied considerably across the cases, with the Adult and Compas having low variation across the models, with only some outliers, except for the Logit, which had higher quartiles, with a variation of 5 to 10 models in Adult dataset. The German model expressed a similar variation across the models; the RF outliers are more notable, reaching 30 models. For the Bank dataset with the Logit, KNN and RF architectures, 30 models were always needed, showing that the stopping criteria were not effective in those cases, while the SVC and MLP had the highest variation, floating from 4 to 30 models. In general, the number of models stayed below the base experiment, ensuring a computational gain when estimating the representative metrics without considerably increasing the variability of the results, as Figure 4.2 depicts.

Furthermore, related to the number of subsets the proposed experiment findings are generally similar to the base, with certain examples having a lower number of subsets and less variation, such as Logit model for the Compas and German datasets, and others having a higher number of subsets or more variation, such as the MLP model for the Adult and German datasets. Another critical point is the dataset's dependence on the outcome because, even though the same metrics were used, the number of subsets formed differs in each case.

Finally, Table 4.4 indicates an average reduction of 20.82% in execution time. An important note about execution time is that the models used in the tests are small and fast to train; as a result, when the reduction of the number of models is small, the execution time may increase since the early stop adds another phase of computations to the process. This increase should not be relevant when training larger models, which take longer to train, because the increase in model size has no effect on the early stop calculation. In other words, the proposed experiment will increase processing time by a few seconds, whereas reducing a single model might reduce time by hours.

4.2 *Analysis of the Metrics subsets*

4.2.1 *Results*

The examination of established metric groups is a central focus within this study, encompassing the different contexts and experiments. Specifically, Tables S.1 through S.40 show the representations of the formed metric groups, highlighting the correlation values and the similar terms (TP, FP, TN, FN) from the equations between each represented metric and its corresponding representative, facilitating the analysis of how equation similarity influences the constitution of these metric groups, the effects on correlation, and distinctive patterns in the clustering of metrics for each experiment.

4.2.2 Discussion

The expected results are that fairness metrics, with similar equations, would be grouped together and have a high correlation between them. Nevertheless, the presence of similar terms does not exhibit a straightforward, direct relationship with correlation values. In certain instances, metrics with a greater number of similar terms may display either lower or higher correlations. An exception is for metric pairs, metrics with similar formulas that share the same base calculation but differ in the operation, where the correlation is constantly high (over 70%), for example in tables for example in tables S.2,S.6,S.25, S.33. This behaviour shows that the structure of the equation as a whole, considering terms and operations, is more impactful for the metric similarity than only the similar terms. Most metric pairs are always together in the same cluster, but the proposed experiment clusters them together more often than the base experiment, with the Error Ratio and Difference being the ones that are in many instances separated, 6 and 10 respectively, for the proposed and base experiment. Within the majority of groups, one or two similar terms predominate in both experiments.

In the base experiment, there are instances where a metric is represented by another metric devoid of any related terms in their respective equations. This recurrent pattern is notably discernible in Tables discernible in Tables S.22-S.25, S.28,S.29, S.34-S.36, S.38, S.40. Such behaviour is generally less prevalent in the proposed experiment, with an exception being the German dataset, where this phenomenon occurs in Tables S.16-S.18, S.20. Showing that the proposed experiment have more coherent groups.

4.3 Subset Recurrence Analysis

4.3.1 Results

The results of the cluster recurrence and repetition analysis are reported in two parts. Figure 4.3 illustrates the first part, which shows the frequency with which two metrics occur together in a heat-map style. Table 4.5 presents the second part, which displays the repeated clusters that were created during the experiments.

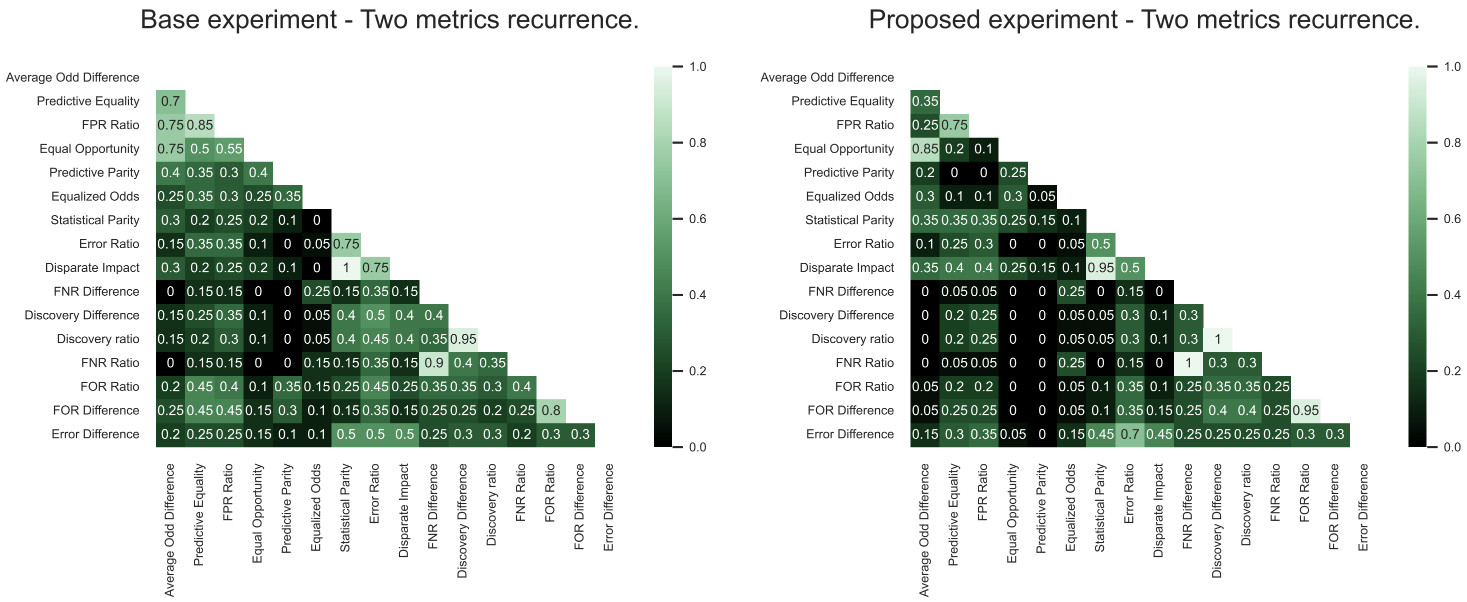


Figure 4.3: The frequency at which two metrics are in the same group across all contexts.

Table 4.5: Repetition of clusters across contexts.

Experiment	Metrics	Repetitions
Proposed	Predictive Parity	14
	FNR Difference, FNR Ratio	9
	FOR Difference, FOR Ratio	8
	Equalized Odds	6
	Disparate Impact, Error Difference, Error Ratio, Statistical Parity	5
	Average Odd Difference, Equal Opportunity, Equalized Odds	5
	Predictive Equality, FPR Ratio,	5
	Average Odd Difference, Equal Opportunity	4
	Error Difference	3
	Error Ratio	3
Baseline	Equalized Odds	7
	Error Difference	6
	Discovery Ratio, Discovery Difference	6
	FNR Difference, FNR Ratio	6
	Disparate Impact, Error Ratio, Statistical Parity	5
	Predictive Parity, FOR Difference, FOR Ratio	4
	Equal Opportunity	3
	Predictive Parity	3

4.3.2 Discussion

The heat-maps in Figure 4.3 indicate similarities in the metrics clustering for both experiments. As an illustration, consider the Predictive Parity metric, which is isolated in the majority of clusters and displays an individual bias measure that, aside from rare instances in which metrics containing the True Positive term could appear related to it.

Additionally, we observed that metric pairs are often clustered together. Figure 4.3 illustrates this effect in both experiments for the metric pairs: FOR Difference and Ratio, FNR Difference and Ratio, Disparity Impact and Statistical Parity, Error Difference and Ratio and the FPR Ratio and Predictive Equality metrics.

The exceptions are the Average Odds Difference and Equalized Odds pair, which have quite similar formulations but are infrequently grouped together, differing solely by the absolute value presence. Equalized Odds, in both experiments, were paired at most seven times in the base experiment and six times in the proposed experiment, suggesting that it's a relevant metric and reflecting a unique aspect of bias. Predictive Parity is the only other metric in a comparable circumstance. The Average Odds Difference, had a high recurrence with Equal Opportunity across the experiments, and with Predictive Equality and FPR Ratio in the base experiment, metrics focused on TP or FP which are terms present in Average Odds Difference.

Certain metrics are never combined in either experiment, indicating that they measure unfairness in different ways. For instance, the Error Ratio and Predictive Parity do not appear together in any case and do not share any elements in their equations. However, FNR Difference and Average Odds Difference share the FN and TP but do not appear together. As with the Average Odds Difference and Equalized Odds, these results show that while some equation similarity is required for correlated behaviour, it does not ensure that the metrics will measure unfairness in the same way.

Comparing the results illustrated by Figure 4.3, the proposed experiment presents a higher recurrence of metric pairs when the correlation is strong, while also presenting less infrequent groups. Indicating that the proposed experiment managed to group the metrics more consistently, which may lead to lower context dependence.

Table 4.5 displays the metric groups for each experiment that appear in at least three contexts. 50 and 58 distinct groups in total were discovered for the proposed and base experiments, respectively. Of these, 6 repeated twice for the proposed experiment and 4 for the base experiment. The overall cluster recurrence of the proposed experiment was the highest, corroborating the results displayed in Figure 4.3 and the group consistency.

Clusters containing a single metric and two similar metrics had the highest repetition in both experiments. This indicates that while some metrics, such as isolated metrics, represent unique perspectives on unfairness, others are repetitive and can be substituted, as would be expected from similar metrics like FNR Ratio and Difference. Except for the lack of the Error Difference in the base experiment, the following groups are present in both experiments: Disparate Impact, Error Difference, Error Ratio, and Statistical Parity. The Error Ratio and Difference, along with the Statistical Parity and Disparate Impact, are two metric pairs. Both use N in the denominator and FP in the numerator; the only distinction is whether TP or FN is present in the numerator. Among the most frequent groups, the impossibility theorem is followed by not grouping together any incompatible metrics.

4.4 *Frequency of representative metrics*

4.4.1 *Results*

Metrics that are frequently chosen as representative are the best for analysing unfairness because they will either have a distinct definition of unfairness (when they are isolated) or will cover an aspect of unfairness more broadly. The outcomes are displayed as follows: 4.6 for the isolated metrics; Table 4.7 for overall frequencies; Tables 4.8 and 4.9 for the dataset view; Tables 4.10 and 4.11 for the model view.

Table 4.6: Normalised frequency of isolated metrics for each case (20).

Metrics	**Base**	**Proposed**
Disparate Impact	0.000	0.000
Discovery Ratio	0.050	0.000
FOR Ratio	0.000	0.050
FPR Ratio	0.000	0.100
FNR Ratio	0.050	0.000
Equalized Odds	0.350	0.300
Error Difference	0.300	0.150
Error Ratio	0.050	0.150
Discovery Difference	0.000	0.000
Average Odd Difference	0.000	0.000
FOR Difference	0.100	0.000
Predictive Equality	0.050	0.050
Statistical Parity	0.000	0.050
Equal Opportunity	0.150	0.100
FNR Difference	0.000	0.000
Predictive Parity	0.150	0.700
Mean	0.078	0.103
Variance	0.012	0.032

Table 4.7: Normalised frequency of representative metrics for each case (20). A higher value indicates a more unique measure*.

Metrics	Base	Proposed	*Legend
Disparate Impact	0.250	0.250	0.000
Discovery Ratio	0.200	0.200	0.250
FOR Ratio	0.300	0.350	0.500
FPR Ratio	0.250	0.400	0.750
FNR Ratio	0.400	0.300	1.000
Equalized Odds	0.500	0.500	
Error Difference	0.300	0.300	
Error Ratio	0.200	0.450	
Discovery Difference	0.300	0.400	
Average Odd Difference	0.400	0.450	
FOR Difference	0.300	0.350	
Predictive Equality	0.150	0.300	
Statistical Parity	0.250	0.150	
Equal Opportunity	0.250	0.300	
FNR Difference	0.350	0.350	
Predictive Parity	0.350	0.750	
Mean:	0.297	0.363	
Variance	0.008	0.019	

Table 4.8: Normalised frequency of representative metrics for each dataset in the base experiment (5). A higher value indicates a more unique measure*.

Metrics	Adult	Bank	Compas	German	*Legend
Disparate Impact	0.200	0.400	0.000	0.400	0.000
Discovery Ratio	0.000	0.000	0.400	0.400	0.250
FOR Ratio	0.200	0.400	0.200	0.400	0.500
FPR Ratio	0.400	0.000	0.600	0.000	0.750
FNR Ratio	0.400	0.600	0.400	0.200	1.000
Equalized Odds	0.600	0.200	0.400	0.800	
Error Difference	0.800	0.200	0.000	0.200	
Error Ratio	0.400	0.200	0.000	0.200	
Discovery Difference	0.800	0.000	0.200	0.200	
Average Odd Difference	0.400	0.600	0.400	0.200	
FOR Difference	0.400	0.600	0.200	0.000	
Predictive Equality	0.000	0.200	0.200	0.200	
Statistical Parity	0.400	0.000	0.200	0.400	
Equal Opportunity	0.400	0.200	0.000	0.400	
FNR Difference	0.600	0.200	0.400	0.200	
Predictive Parity	0.400	0.000	0.600	0.400	
Mean	0.400	0.238	0.263	0.288	
Variance	0.053	0.049	0.041	0.037	

Table 4.9: Normalised frequency of representative metrics for each dataset in the proposed experiment (5). A higher value indicates a more unique measure*.

Metrics	Adult	Bank	Compas	German	*Legend
Disparate Impact	0.400	0.200	0.200	0.200	0.000
Discovery Ratio	0.400	0.400	0.000	0.000	0.250
FOR Ratio	0.600	0.400	0.200	0.200	0.500
FPR Ratio	0.400	0.600	0.200	0.400	0.750
FNR Ratio	0.000	0.600	0.600	0.000	1.000
Equalized Odds	0.600	0.400	0.800	0.200	
Error Difference	0.600	0.200	0.200	0.200	
Error Ratio	1.000	0.600	0.000	0.200	
Discovery Difference	0.200	0.600	0.600	0.200	
Average Odd Difference	0.400	0.600	0.600	0.200	
FOR Difference	0.400	0.400	0.200	0.400	
Predictive Equality	0.200	0.200	0.400	0.400	
Statistical Parity	0.200	0.200	0.200	0.000	
Equal Opportunity	0.200	0.200	0.400	0.400	
FNR Difference	0.600	0.400	0.200	0.200	
Predictive Parity	1.000	1.000	0.800	0.200	
Mean	0.450	0.438	0.350	0.213	
Variance	0.077	0.049	0.067	0.019	

Table 4.10: Normalised frequency of representative metrics for each model in the base experiment (4). A higher value indicates a more unique measure*.

Metrics	KNN	Logit	MLP	RF	SVC	*Legend
Disparate Impact	0.500	0.000	0.250	0.000	0.500	0.000
Discovery Ratio	0.250	0.250	0.250	0.250	0.000	0.250
FOR Ratio	0.500	0.500	0.250	0.000	0.250	0.500
FPR Ratio	0.000	0.250	0.250	0.500	0.250	0.750
FNR Ratio	0.250	0.250	0.000	0.750	0.750	1.000
Equalized Odds	0.500	0.500	0.250	0.500	0.750	
Error Difference	0.250	0.250	0.750	0.250	0.000	
Error Ratio	0.000	0.500	0.250	0.250	0.000	
Discovery Difference	0.000	0.250	0.250	0.250	0.750	
Average Odd Difference	0.750	0.500	0.500	0.250	0.000	
FOR Difference	0.250	0.250	0.250	0.250	0.500	
Predictive Equality	0.000	0.250	0.000	0.250	0.250	
Statistical Parity	0.250	0.250	0.250	0.250	0.250	
Equal Opportunity	0.250	0.250	0.250	0.000	0.500	
FNR Difference	0.250	0.250	1.000	0.250	0.000	
Predictive Parity	0.000	0.500	0.500	0.750	0.000	
Mean	0.250	0.313	0.328	0.297	0.297	
Variance	0.050	0.021	0.064	0.052	0.085	

Table 4.11: Normalised frequency of representative metrics for each model in the proposed experiment (4). A higher value indicates a more unique measure*.

Metrics	KNN	Logit	MLP	RF	SVC	*Legend
Disparate Impact	0.000	0.500	0.250	0.250	0.250	0.000
Discovery Ratio	0.250	0.250	0.500	0.000	0.000	0.250
FOR Ratio	0.500	0.250	0.500	0.500	0.000	0.500
FPR Ratio	0.750	0.250	0.250	0.750	0.000	0.750
FNR Ratio	0.000	0.500	0.000	0.500	0.500	1.000
Equalized Odds	0.250	0.250	1.000	0.750	0.250	
Error Difference	0.500	0.500	0.000	0.000	0.500	
Error Ratio	0.250	0.500	0.500	0.750	0.250	
Discovery Difference	0.500	0.250	0.250	0.250	0.750	
Average Odd Difference	0.250	0.500	0.250	0.500	0.750	
FOR Difference	0.250	0.250	0.250	0.000	1.000	
Predictive Equality	0.000	0.250	0.500	0.000	0.750	
Statistical Parity	0.250	0.000	0.000	0.250	0.250	
Equal Opportunity	0.500	0.000	0.500	0.250	0.250	
FNR Difference	0.500	0.500	0.500	0.000	0.250	
Predictive Parity	0.500	1.000	0.750	0.750	0.750	
Mean	0.328	0.359	0.375	0.344	0.406	
Variance	0.048	0.058	0.075	0.091	0.099	

4.4.2 Discussion

Table 4.6 highlights two metrics: Predictive Parity in the proposed experiment and Equalized Odds for both scenarios, as they are isolated more frequently. Equalized Odds is satisfied when both of its terms, TPR and FPR, are also satisfied. It is a distinct metric in terms of equation, and the absolute value sets it apart from the Average Odd Difference. Predictive Parity is the only metric that looks only at the positive miss-classification values, justifying the isolation, and is satisfied when the classes have an equal chance of receiving a positive classification.

The most isolated metric is also the most representative metric in table 4.7 since an isolated representative is still a representative metric. The idea that the fairness metrics depend on context and no single metric predominates over the others is supported by the fact that the frequency of the metrics remained relatively consistent across all experiments and metrics, except for Predictive Parity.

The proposed modifications have a significant impact on the metric chosen as representative, as shown by a comparison of tables 4.8 and 4.9, as well as tables 4.10 and 4.11, with dataset and model vision, respectively. The proposed experiment demonstrates a dataset- or model-specific metric that would be effective in most cases where one of those factors is constant, showing consistency with metrics that are always or never selected. The same can be said in a smaller degree for the base experiment, as there is less of those cases.

Furthermore, when analysing the dataset element for the proposed experiment, the metrics highlighted in Table 4.9 for each application scenario were: Error Ratio and Predictive Parity for demographic census in Adult dataset, Predictive Parity for advertisement in Bank dataset, and Equalized Odds and Predictive Parity for criminal judgement in Compas dataset. The Predictive Parity, in all three examples, confirms that one of the most important issues in those circumstances is the proportion of correct positive predictions between groups, while the Error Ratio for the demographic census indicates that a problem is the difference in miss classification between groups, encompassing both positive and negative classes over the total. Finally, the Equalized Odds in the criminal judgement scenario show that the balance between correct and incorrect predictions of the positive class is a concern. Conversely, the German dataset seems to be highly dependent in the model used, as no single metric stood out.

4.5 Group Similarity Analysis

4.5.1 Results

Tables 4.12, 4.13, 4.14, and 4.15 display the group similarity analysis with various views. Each value measures the average distance between the represented metrics and the representing one. The overall average results for each experiment are displayed using the metric shown in Table 4.12.

Table 4.12: Similarity between each metric and its representative, averaging each metric. Lower values indicate better representation*.

Metrics	**Base**	**Proposed**	*Legend
Disparate Impact	0.349	0.565	0.000
Discovery Ratio	0.323	0.301	0.250
FOR Ratio	0.479	0.348	0.500
FPR Ratio	0.599	0.662	0.750
FNR Ratio	0.249	0.105	1.000
Equalized Odds	0.169	0.178	
Error Difference	0.411	0.387	
Error Ratio	0.401	0.561	
Discovery Difference	0.364	0.328	
Average Odd Difference	0.201	0.132	
FOR Difference	0.364	0.313	
Predictive Equality	0.232	0.408	
Statistical Parity	0.275	0.376	
Equal Opportunity	0.172	0.125	
FNR Difference	0.210	0.111	
Predictive Parity	0.328	0.353	
Mean	0.320	0.328	
Variance	0.014	0.029	

Table 4.13: Similarity between each metric and its representative, averaging each context. Lower values indicate better representation*.

Context	**Base**	**Proposed**	*Legend
Adult_KNN	0.134	0.122	0.000
Adult_Logit	0.147	0.492	0.250
Adult_MLP	0.161	0.229	0.500
Adult_RF	0.078	0.160	0.750
Adult_SVC	0.075	0.051	1.000
Bank_KNN	0.301	0.374	
Bank_Logit	0.369	0.265	
Bank_MLP	0.479	0.305	
Bank_RF	0.336	0.372	
Bank_SVC	0.283	0.286	
Compas_KNN	0.398	0.333	
Compas_Logit	0.456	0.367	
Compas_MLP	0.465	0.330	
Compas_RF	0.486	0.464	
Compas_SVC	0.373	0.467	
German_KNN	0.218	0.434	
German_Logit	0.374	0.208	
German_MLP	0.237	0.315	
German_RF	0.215	0.383	
German_SVC	0.678	0.460	
Mean	0.3132	0.321	
Variance	0.025	0.015	

Table 4.14: Similarity between each metric and its representative, averaging each model. Lower values indicate better representation*.

Model	**Base**	**Proposed**	*Legend
KNN	0.263	0.316	0.000
Logit	0.336	0.333	0.250
MLP	0.336	0.295	0.500
RF	0.279	0.345	0.750
SVC	0.352	0.316	1.000
Mean	0.313	0.321	
Variance	0.002	0.000	

Table 4.15: Similarity between each metric and its representative, averaging each dataset. Lower values indicate better representation*.

Dataset	**Base**	**Proposed**	*Legend
Adult	0.119	0.211	0.000
Bank	0.354	0.320	0.250
Compas	0.436	0.392	0.500
German	0.344	0.360	0.750
Mean	0.313	0.321	1.000
Variance	0.018	0.006	

4.5.2 Discussion

The practically identical mean values across the tables for the base and proposed models demonstrate that the reduction of models has little impact on the group quality as a whole. Although certain metrics perform better in one experiment than another, the majority of metrics show similar results, with differences of about 0.2, indicating that the results are within the same range. Average Odd Difference and Equalized Odds, two metrics that include more elements than the others, Equal Opportunity, which is included in the first two metrics, and the pair FNR ratio/difference, which represents the miss-classification for the positive values, are the metrics that yielded the best results across all experiments. This indicates that those metrics behaviours are well represented.

FPR Ratio, which shows miss-classification for negative values, was the least effective metric in both trials. Predictive Equality, on the other hand, which has an equation that is similar, produced a better result. The Error Ratio and Error Difference were two other metrics that under-performed. They both contain terms that are used in other equations, but they do it differently because they use the total number of elements from both classes in the denominator. Even though they are not always well represented, they should be examined on an individual basis since they constitute important metrics for model evaluation.

Table 4.14 shows the models which demonstrate that the differences in the models had no discernible effect on how similar the formed groups were in any experiment, since all values are close. Demonstrating that despite group composition changing across models, group quality remains constant. However, the dataset view in table 4.15 reveals that, in both experiments, the Adult dataset had more distinct groups, whereas the Compas dataset had closer groups. Race is used as a sensitive attribute in both datasets, which makes this behaviour especially intriguing. However, when compared to the Compas dataset, the Adult dataset has more data instances and a more uneven distribution. German and Bank had comparable and average results, with gender and age sensitive attributes, respectively.

The group similarity analysis is significantly impacted by the dataset; a more detailed analysis is shown in table 4.13. While most results in a given context are similar, there are some variations and outliers that cannot be entirely attributed to either the model or the dataset. Tables 4.2 and 4.3 provide comparisons of the reduction percentage and number of subsets between the two experiments, which can be used to further investigate cases where there is a variation of more than 0.1 between the two experiments. The following are the cases: German KNN, German Logit, German RF, German SVC, Adult Logit, Bank Logit, Bank MLP, Compas MLP.

The proposed experiment in Bank Logit and Bank MLP cases showed no or minimal reduction in the number of models (0% and 10%, respectively), an increase in the number of subsets (by 4 and 2, respectively), and improved similarity results. The increased number of subsets is expected to improve the distance results since the metrics will be more evenly distributed, bringing the subsets closer together.

The Adult Logit showed an increase of one subset, with a higher model reduction of 83.33%; however, the grouping fared considerably worse, suggesting that insufficient models were used to accurately correlate the fairness metrics. Still, while reducing the number of models between 73.33% and 86.66%, the cases of Compas MLP, German KNN, German Logit, German RF, and German SVC had varying results.

4.6 Overall Discussion

The primary objective of this dissertation was to propose a computational method that allows the selection of the most representative metrics for bias and unfairness assessment in post-processing for binary classification machine learning models in different contexts.

To accomplish this objective, 16 state-of-the-art metrics were analysed and selected for detecting bias and unfairness, which were used in four separate datasets and five different

ML models. A correlation-based algorithm was developed, improving on various aspects of a base algorithm, including better data bias representation, increased computational efficiency, a more accurate and robust correlation method, and explicit tests and validation. Finally, we compare the proposed algorithm's performance in selecting the representative metrics to the base method in terms of the number of models and subsets, analysis of the formed subsets, recurrence of the subset, frequency of the representative metrics, and group similarity.

In summary, the analysis demonstrated that the number of metrics could be reduced for both the base and proposed experiments. Despite this, the proposed experiment managed to maintain competitive performance while lowering both the computational cost and the quantity of metrics. Aside from the computational cost, the proposed experiment delivers a more consistent result across contexts, suggesting that in similar contexts it may be viable to reuse the chosen metrics; however, a new choice of metrics is still preferred. This context consistency is caused by a better representation of the data, coming from the stratified samples and from the appropriated correlation method, considering the non-normality of the data.

The reduction in the number of models has significant implications for computational efficiency, allowing for faster processing without compromising the quality of the results and supporting the development of fair ML models that are also sustainable with a reduced carbon footprint. The proposed experiment is especially desirable for large-scale tasks, such as text generation, where retraining and fine-tuning the model is expensive, and any reduction is advantageous.

Analysing the underlined behaviour of the fairness metrics, metrics with similar equations that only differ in the operation and use the same terms generally exhibit similar behaviour. The proposed experiment was able to pair them more frequently than the base experiment, and the correlation between them is consistently high, making it the first option for reducing the number of metrics by eliminating one of the pair's elements. Two metric pairs had different behaviour: the Equalized Odds and Average Odd Difference pair, and the Error Ratio and Error Difference pair. Equalized Odds showed a unique unfairness view, frequently being isolated in a cluster, with the presence of the absolute value in the equation being enough to differentiate it from Average Odd Difference. Both metrics were well represented, with a small distance between their groups. Error Ratio and Error Difference, on the other hand, were two of the worst performance metrics, distance wise, likely caused by the amount of information in the equation being higher than the other metrics. The correlation is enough to group them, but a case-by-case analysis may be necessary.

In cases that don't involve a metric pair, the presence of similar terms is a strong indication of a direct relationship between the metrics, but it's not enough to determine the groups and the correlation, as a single term can change the metric results depending on the context.

Analysing the metrics chosen as representatives, if all contexts are considered, there is not a metric to highlight, as they all have around the same pick rate. But when looking closer into only the models, or only the datasets, while keeping the other one the same, we can see that more consistency arises, as there are metrics that are always or never picked; this is particularly true in the proposed experiment. For the analysed cases, the Predictive Parity metric was highlighted in the criminal judgement, demographic census, and advertisement scenarios, while the Error Ratio metric was highlighted for the demographic census, and Equalized Odds was in evidence in criminal judgment. The sensitive attribute analysed didn't show any underlying property kept in different datasets, as the two using race as a sensitive attribute had different results in relation to the group distance, pointing out that the size of the dataset and the distribution of the data have a bigger impact in the context than the sensitive attribute.

Chapter Five

Conclusions

This work Research Question is 'How to reduce the fairness metrics scope for a problem, maintaining their representativeness, and supporting identification and mitigation of bias and unfairness in a model?', leading to the overarching objective to propose a computational method that allows the selection of the most representative metrics for bias and unfairness assessment in post-processing for binary classification machine learning models in different contexts.

The research question can be addressed by implementing a method that uses bootstrap sampling via MCMC to estimate correlations between fairness metrics. Following with a clustering process to group similar metrics, allowing for the selection of a representative metric from each cluster. This approach, referred to as the base experiment, selects a subset of metrics to be used for the identification and mitigation of bias and unfairness.

However, to ensure the representativeness of the selected metrics, it's crucial to address certain caveats in the method, including the use of random sampling, the high computational cost, the choice of correlation methods, and the validation of results. Each of these issues has been explored, and solutions have been integrated into the proposed experiment.

The proposed experiment included three novel modifications to the base experiment: a MAD-based stop criterion to reduce the computational cost of running the Monte Carlo method for estimating the fairness metrics correlation, a stratified sample approach rather than random samples to better reflect the data bias, and the Kendall correlation method instead of Pearson to better work with the new set of data.

To achieve the overarching objective, we confronted and improved ideas about using correlation as a heuristic tool for finding representative fairness metrics for a context (dataset + model), using five binary classification ML models: KNN, Logit, MLP, RF, and SVC, trained with four datasets: Adult, Bank, Compas, and German, encompassing respectively the fields of criminal judgement, bank loans, demographic census, and advertisement. Finally, we provided experimental results demonstrating the effectiveness of the mentioned approach in different contexts.

As a result of the modifications, the number of models decreased by an average of 64.37% from the base experiment, which always calls for 30 models, and the training time was reduced by 20.82%. This ensures a computational gain when estimating the representative metrics without suffering significant losses. The proposed method did not lead to an increase in the variability of the results; no group that violates the impossibility theorem was formed; the mean distance between a represented and representative metric remains largely unchanged; and similar metrics are paired together more frequently.

Our proposed method successfully reduced the number of models needed to estimate the representative metrics in a given context while keeping the chosen metrics and formed relevant subsets, with improvements in the grouping consistency. We show that, with the appropriate method, correlation can be used to estimate the most representative fairness metrics if the dataset's class imbalance is respected and a more robust correlation, such as Kendall, is applied, taking into account the number of samples and the data's non-normality.

The results of this work have significant practical implications for future research and opportunities for industry and society. By enabling interdisciplinary dialogue to address ethical and social sustainability challenges, this research directly contributes to several SDGs, including:

- SDG 5: Achieve gender equality and empower all women and girls.

  Our work contributes by evaluating a case study focused on gender unfairness in the context of bank loans. Additionally, we provide a set of representative metrics that are highlighted in this context.

- SDG 10: Reduce inequality within and among countries.

  By focussing on reducing inequalities, our research ensures that advantaged groups do not maintain disproportionate advantages over marginalised groups. This directly supports the goal by promoting fairness and equity.

- SDG 13: Take urgent action to combat climate change and its impacts.

  We contribute to SDG 13 by reducing the computational cost of selecting a representative fairness metric, leading to a more computationally efficient and environmentally sustainable approach that has the potential to reduce carbon footprint.

- SDG 16: Promote peaceful and inclusive societies for sustainable development, provide access to justice for all and build effective, accountable and inclusive institutions at all levels.

  The method proposed supports and enables the development of fair decision-making in machine learning, stepping closer to trustworthy AI systems with the potential to prevent the propagation of inequities and contributing to more inclusive institutions.

Finally, it is important to note that the method created is not limited to the investigated case studies or the established fairness metrics and could be used for other relevant topics for unfairness assessment in post-processing for binary classification. For example, tackling race, gender, and disability discrimination in automated recruiting systems (MINATEL et al., 2023); addressing age, gender, and ethnicity discrimination in automated immigration and refugee applications (MOLNAR, 2019); and issues of hate speech against minorities and the production of sexist, racist, or xenophobic responses by LLM chatbot (MINATEL et al., 2023).

References

ADEL, T.; VALERA, I.; GHAHRAMANI, Z.; WELLER, A. One-network adversarial fairness. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. [S.l.: s.n.], 2019. v. 33, n. 01, p. 2412–2420. 1

ALPAYDIN, E. *Machine learning*. [S.l.]: MIT press, 2021. 2.1.1

AMBROSIUS, W. T. *Topics in biostatistics*. [S.l.]: Springer, 2007. (document), 2.2

AMMAR, J. Cyber gremlin: social networking, machine learning and the global war on al-qaida-and is-inspired terrorism. *International Journal of Law and Information Technology*, Oxford University Press, v. 27, n. 3, p. 238–265, 2019. 2.1.2.3

ANAHIDEH, H.; NEZAMI, N.; ASUDEH, A. On the choice of fairness: Finding representative fairness metrics for a given context. *arXiv preprint arXiv:2109.05697*, 2021. 2.2, 3, 3.1.1, 3.1.1, 3.2.4.1

ASLAM, M.; RAO, G. S.; AL-MARSHADI, A. H.; AHMAD, L.; JUN, C.-H. Control charts for monitoring process capability index using median absolute deviation for some popular distributions. *Processes*, MDPI, v. 7, n. 5, p. 287, 2019. 3.2.2

BADILLA, P.; BRAVO-MARQUEZ, F.; PéREZ, J. Wefe: The word embeddings fairness evaluation framework. In: BESSIERE, C. (Ed.). *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*. International Joint Conferences on Artificial Intelligence Organization, 2020. p. 430–436. Main track. Disponível em: <https://doi.org/10.24963/ijcai.2020/60>. 2.2

BALAYN, A.; LOFI, C.; HOUBEN, G.-J. Managing bias and unfairness in data for decision support: a survey of machine learning and data engineering approaches to identify and mitigate bias and unfairness within data management and analytics systems. *The VLDB Journal*, 2021. 1.1

BAROCAS, S.; HARDT, M.; NARAYANAN, A. *Fairness and Machine Learning: Limitations and Opportunities*. [S.l.]: MIT Press, 2023. 2.1.2, 2.1.2.2

BERK, R.; HEIDARI, H.; JABBARI, S.; KEARNS, M.; ROTH, A. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, v. 50, n. 1, p. 3–44, 2021. Disponível em: <https://doi.org/10.1177/0049124118782533>. 3.2

BOLóN-CANEDO, V.; MORáN-FERNáNDEZ, L.; CANCELA, B.; ALONSO-BETANZOS, A. A review of green artificial intelligence: Towards a more sustainable future. *Neurocomputing*, v. 599, p. 128096, 2024. ISSN 0925-2312. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0925231224008671>. 1

BOOTH, B. M. et al. Integrating psychometrics and computing perspectives on bias and fairness in affective computing: A case study of automated video interviews. *IEEE Signal Processing Magazine*, v. 38, n. 6, p. 84–95, 2021. 2.1.2, 2.1.2.1, 2.1.2.3

BRUCE, P. C.; BRUCE, A. Exploratory data analysis. In: ______. *Practical Statistics for Data Scientists: 50 essential concepts*. 1. ed. [S.l.]: O'Reilly Media, 2018. p. 38–41. 3.2.3

BUOLAMWINI, J.; GEBRU, T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In: FRIEDLER, S. A.; WILSON, C. (Ed.). *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*. PMLR, 2018. (Proceedings of Machine Learning Research, v. 81), p. 77–91. Disponível em: <https://proceedings.mlr.press/v81/buolamwini18a.html>. 2.1.2

CAMDEVIREN, H. A.; YAZICI, A. C.; AKKUS, Z.; BUGDAYCI, R.; SUNGUR, M. A. Comparison of logistic regression model and classification tree: An application to postpartum depression data. *Expert Systems with Applications*, v. 32, n. 4, p. 987–994, 2007. ISSN 0957-4174. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0957417406000753>. 2.1.1.2

CASTELNOVO, A. et al. A clarification of the nuances in the fairness metrics landscape. *Scientific Reports*, v. 12, n. 1, p. 4209, Mar 2022. ISSN 2045-2322. Disponível em: <https://doi.org/10.1038/s41598-022-07939-1>. 1, 2.1.2

CATON, S.; HAAS, C. Fairness in machine learning: A survey. *ACM Comput. Surv.*, Association for Computing Machinery, New York, NY, USA, v. 56, n. 7, apr 2024. ISSN 0360-0300. Disponível em: <https://doi.org/10.1145/3616865>. 3

CHEN, Z.; ZHANG, J. M.; SARRO, F.; HARMAN, M. A comprehensive empirical study of bias mitigation methods for machine learning classifiers. *ACM Trans. Softw. Eng. Methodol.*, Association for Computing Machinery, New York, NY, USA, v. 32, n. 4, may 2023. ISSN 1049-331X. Disponível em: <https://doi.org/10.1145/3583561>. 2.1.2

CHOK, N. S. Pearson's versus spearman's and kendall's correlation coefficients for continuous data. 2010. Disponível em: <http://d-scholarship.pitt.edu/8056/>. 3.2.3

CORBETT-DAVIES, S.; PIERSON, E.; FELLER, A.; GOEL, S.; HUQ, A. Algorithmic decision making and the cost of fairness. In: *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: Association for Computing Machinery, 2017. (KDD '17), p. 797–806. ISBN 9781450348874. Disponível em: <https://doi.org/10.1145/3097983.3098095>. 3.2

DREISEITL, S.; OHNO-MACHADO, L. Logistic regression and artificial neural network classification models: a methodology review. *Journal of Biomedical Informatics*, v. 35, n. 5, p. 352–359, 2002. ISSN 1532-0464. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1532046403000340>. 2.1.1.2, 2.1.1.3

DUA, D.; GRAFF, C. *UCI Machine Learning Repository*. 2017. Disponível em: <http://archive.ics.uci.edu/ml>. 3.2.4.1

DWIVEDI, Y. K. et al. Artificial intelligence (ai): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. *International Journal of Information Management*, Elsevier, v. 57, p. 101994, 2021. 2.1.2.3

DWORK, C.; HARDT, M.; PITASSI, T.; REINGOLD, O.; ZEMEL, R. Fairness through awareness. In: *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*. New York, NY, USA: Association for Computing Machinery, 2012. (ITCS '12), p. 214–226. ISBN 9781450311151. Disponível em: <https://doi.org/10.1145/2090236.2090255>. 3.2

FEIJÓO, C. et al. Harnessing artificial intelligence (ai) to increase wellbeing for all: The case for a new technology diplomacy. *Telecommunications Policy*, Elsevier, v. 44, n. 6, p. 101988, 2020. 2.1.2.3

FELDMAN, M.; FRIEDLER, S. A.; MOELLER, J.; SCHEIDEGGER, C.; VENKATASUBRAMANIAN, S. Certifying and removing disparate impact. In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: Association for Computing Machinery, 2015. (KDD '15), p. 259–268. ISBN 9781450336642. Disponível em: <https://doi.org/10.1145/2783258.2783311>. 3.2

FERRARA, E. Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies. *Sci*, v. 6, n. 1, 2024. ISSN 2413-4155. Disponível em: <https://www.mdpi.com/2413-4155/6/1/3>. 1, 2.1.2.1

GARG, P.; VILLASENOR, J.; FOGGO, V. Fairness metrics: A comparative analysis. In: . [S.l.: s.n.], 2020. 2

GEIGER, R. S. et al. Garbage in, garbage out? do machine learning application papers in social computing report where human-labeled training data comes from? In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. New York, NY, USA: Association for Computing Machinery, 2020. (FAT* '20), p. 325–336. ISBN 9781450369367. Disponível em: <https://doi.org/10.1145/3351095.3372862>. 2.1.2

GUO, G.; WANG, H.; BELL, D.; BI, Y.; GREER, K. Knn model-based approach in classification. In: SPRINGER. *On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE: OTM Confederated International Conferences, CoopIS, DOA, and ODBASE 2003, Catania, Sicily, Italy, November 3-7, 2003. Proceedings*. [S.l.], 2003. p. 986–996. (document), 2.4

HARDT, M.; PRICE, E.; SREBRO, N. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, v. 29, 2016. 3.2

HE, X.; NASSAR, I.; KIROS, J.; HAFFARI, G.; NOROUZI, M. Generate, Annotate, and Learn: NLP with Synthetic Text. *Transactions of the Association for Computational Linguistics*, v. 10, p. 826–842, 08 2022. ISSN 2307-387X. Disponível em: <https://doi.org/10.1162/tacl_a_00492>. 2.1.2

HILBE, J. M. *Logistic regression models*. [S.l.]: Chapman and hall/CRC, 2009. 2.1.1.2, 2.1.1.2

HU, J.; ZHANG, H.; LIU, Y.; SUTCLIFFE, R.; FENG, J. Bbw: a batch balance wrapper for training deep neural networks on extremely imbalanced datasets with few minority samples. *Applied Intelligence*, v. 52, n. 6, p. 6723–6738, Apr 2022. ISSN 1573-7497. Disponível em: <https://doi.org/10.1007/s10489-021-02623-9>. 3.2.1

JIANG, L.; YAO, R. Modelling personal thermal sensations using c-support vector classification (c-svc) algorithm. *Building and Environment*, Elsevier, v. 99, p. 98–106, 2016. (document), 2.1.1.4, 2.5

JONES, D.; SNIDER, C.; NASSEHI, A.; YON, J.; HICKS, B. Characterising the digital twin: A systematic literature review. *CIRP Journal of Manufacturing Science and Technology*, Elsevier, v. 29, p. 36–52, 2020. 2.1.2

JORDAN, M. I.; MITCHELL, T. M. Machine learning: Trends, perspectives, and prospects. *Science*, American Association for the Advancement of Science, v. 349, n. 6245, p. 255–260, 2015. 2.1.1

KARLIK, B.; OLGAC, A. V. Performance analysis of various activation functions in generalized mlp architectures of neural networks. *International Journal of Artificial Intelligence and Expert Systems*, v. 1, n. 4, p. 111–122, 2011. 2.1.1.1

KILKENNY, M. F.; ROBINSON, K. M. Data quality: "garbage in – garbage out". *Health Information Management Journal*, v. 47, n. 3, p. 103–105, 2018. PMID: 29719995. Disponível em: <https://doi.org/10.1177/1833358318774357>. 2.1.2

KIM, H. et al. Genetic discrimination: introducing the asian perspective to the debate. *npj Genomic Medicine*, v. 6, n. 1, p. 54, Jul 2021. ISSN 2056-7944. Disponível em: <https://doi.org/10.1038/s41525-021-00218-4>. 2.1.2

KIM, J. S.; CHEN, J.; TALWALKAR, A. FACT: A diagnostic for group fairness trade-offs. In: III, H. D.; SINGH, A. (Ed.). *Proceedings of the 37th International Conference on Machine Learning*. PMLR, 2020. (Proceedings of Machine Learning Research, v. 119), p. 5264–5274. Disponível em: <https://proceedings.mlr.press/v119/kim20a.html>. 2

KIRASICH, K.; SMITH, T.; SADLER, B. Random forest vs logistic regression: binary classification for heterogeneous datasets. *SMU Data Science Review*, v. 1, n. 3, p. 9, 2018. (document), 2.3, 2.6

KLEINBERG, J.; MULLAINATHAN, S.; RAGHAVAN, M. Inherent trade-offs in the fair determination of risk scores. 09 2016. 1, 2.1.2, 2

KOMAREK, P. *Logistic regression for data mining and high-dimensional classification*. [S.l.]: Carnegie Mellon University, 2004. 2.1.1.2

KÖNIG, P. D.; WENZELBURGER, G. When politicization stops algorithms in criminal justice. *The British Journal of Criminology*, Oxford University Press UK, v. 61, n. 3, p. 832–851, 2021. 2.1.2.3

KULKARNI, V. Y.; SINHA, P. K. Random forest classifiers: a survey and future research directions. *Int. J. Adv. Comput*, v. 36, n. 1, p. 1144–1153, 2013. 2.1.1.5

LARSON, J.; MATTU, S.; KIRCHNER, L.; ANGWIN, J. How we analyzed the compas recidivism algorithm. *ProPublica (5 2016)*, v. 9, n. 1, p. 3–3, 2016. 3.2.4.1

LESLIE, D. et al. Ai fairness in practice. *The Alan Turing Institute*, 2023. 2.1.2, 2.1.2.1

LI, B. et al. Trustworthy ai: From principles to practices. *ACM Comput. Surv.*, Association for Computing Machinery, New York, NY, USA, v. 55, n. 9, jan 2023. ISSN 0360-0300. Disponível em: <https://doi.org/10.1145/3555803>. 2.1.2.3

LIANG, W. et al. Advances, challenges and opportunities in creating data for trustworthy ai. *Nature Machine Intelligence*, Nature Publishing Group UK London, v. 4, n. 8, p. 669–677, 2022. 2.1.2.3

LOYOLA-GONZALEZ, O. Black-box vs. white-box: Understanding their advantages and weaknesses from a practical point of view. *IEEE access*, IEEE, v. 7, p. 154096–154113, 2019. 2.1.2.3

MAHESH, B. Machine learning algorithms-a review. *International Journal of Science and Research (IJSR).[Internet]*, v. 9, n. 1, p. 381–386, 2020. 2.1.1

MAJUMDER, S.; CHAKRABORTY, J.; BAI, G. R.; STOLEE, K. T.; MENZIES, T. Fair enough: Searching for sufficient measures of fairness. *ACM Trans. Softw. Eng. Methodol.*, Association for Computing Machinery, New York, NY, USA, v. 32, n. 6, sep 2023. ISSN 1049-331X. Disponível em: <https://doi.org/10.1145/3585006>. 1, 2.2, 3.2.4.1

MANDHALA, V. N.; BHATTACHARYYA, D.; MIDHUNCHAKKARAVARTHY, D. Mitigating bias by optimizing the variance between privileged and deprived data using post processing method. *Revue d'Intelligence Artificielle*, International Information and Engineering Technology Association (IIETA), v. 36, n. 1, p. 87, 2022. 3.2

MEHRABI, N.; MORSTATTER, F.; SAXENA, N.; LERMAN, K.; GALSTYAN, A. A survey on bias and fairness in machine learning. *ACM Comput. Surv.*, Association for Computing Machinery, New York, NY, USA, v. 54, n. 6, jul 2021. ISSN 0360-0300. Disponível em: <https://doi.org/10.1145/3457607>. 2.1.2.1, 2.1.2.2

MINATEL, D. et al. Unfairness in machine learning for web systems applications. In: *Proceedings of the 29th Brazilian Symposium on Multimedia and the Web*. New York, NY, USA: Association for Computing Machinery, 2023. (WebMedia '23), p. 144–153. ISBN 9798400709081. Disponível em: <https://doi.org/10.1145/3617023.3617043>. 5

MITCHELL, M. et al. Model cards for model reporting. In: *Proceedings of the conference on fairness, accountability, and transparency*. [S.l.: s.n.], 2019. p. 220–229. 1, 2.1.2

MOLNAR, P. Technology on the margins: Ai and global migration management from a human rights perspective. *Cambridge International Law Journal*, Edward Elgar Publishing Ltd, Cheltenham, UK, v. 8, n. 2, p. 305 – 330, 2019. Disponível em: <https://www.elgaronline.com/view/journals/cilj/8/2/article-p305.xml>. 5

MOORE, D. S.; NOTZ, W. I.; NOTZ, W. *Statistics: Concepts and controversies*. [S.l.]: Macmillan, 2006. 3.2.1

MORO, S.; CORTEZ, P.; RITA, P. A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, v. 62, p. 22–31, 2014. ISSN 0167-9236. 3.2.4.1

NIELSEN, A. *Practical Fairness: Achieving Fair and Secure Data Models*. O'Reilly Media, Incorporated, 2020. ISBN 9781492075738. Disponível em: <https://books.google.com.br/books?id=aNlazQEACAAJ>. 1

NOIA, T. D.; TINTAREV, N.; FATOUROU, P.; SCHEDL, M. Recommender systems under european ai regulations. *Communications of the ACM*, ACM New York, NY, USA, v. 65, n. 4, p. 69–73, 2022. 2.1.2.3

NOVAKOVIC, J.; VELJOVIC, A. C-support vector classification: Selection of kernel and parameters in medical diagnosis. In: IEEE. *2011 IEEE 9th international symposium on intelligent systems and informatics*. [S.l.], 2011. p. 465–470. 2.1.1.4

PAGANO, T. P. et al. Bias and unfairness in machine learning models: A systematic review on datasets, tools, fairness metrics, and identification and mitigation methods. *Big data and cognitive computing*, MDPI, v. 7, n. 1, p. 15, 2023. 1, 2.1.2, 3

PAGANO, T. P. et al. Context-based patterns in machine learning bias and fairness metrics: A sensitive attributes-based approach. *Big Data and Cognitive Computing*, v. 7, n. 1, 2023. ISSN 2504-2289. Disponível em: <https://www.mdpi.com/2504-2289/7/1/27>. 1, 2.1.2, 2.2

PAL, M. Random forest classifier for remote sensing classification. *International journal of remote sensing*, Taylor & Francis, v. 26, n. 1, p. 217–222, 2005. 2.1.1.5

PAVIGLIANITI, A.; PASERO, E. Vital-ecg: a de-bias algorithm embedded in a gender-immune device. In: IEEE. *2020 IEEE International Workshop on Metrology for Industry 4.0 & IoT*. [S.l.], 2020. p. 314–318. 1

PEDREGOSA, F. et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, v. 12, p. 2825–2830, 2011. 3.2.4.2

PERERA, M. T. Investigating ethical considerations and challenges for real-time computer vision machine learning applications in urban environments. *International Journal of Social Analytics*, v. 9, n. 3, p. 1–10, Mar. 2024. Disponível em: <https://norislab.com/index.php/ijsa/article/view/76>. 2.1.2

POPESCU, M.-C.; BALAS, V. E.; PERESCU-POPESCU, L.; MASTORAKIS, N. Multilayer perceptron and neural networks. *WSEAS Transactions on Circuits and Systems*, World Scientific and Engineering Academy and Society (WSEAS) Stevens Point . . . , v. 8, n. 7, p. 579–588, 2009. 2.1.1.1, 2.1.1.3

PRINCE, A. E.; SCHWARCZ, D. Proxy discrimination in the age of artificial intelligence and big data. *Iowa L. Rev.*, HeinOnline, v. 105, p. 1257, 2019. 2.1.2

PUERTAS, O. L.; BRENNING, A.; MEZA, F. J. Balancing misclassification errors of land cover classification maps using support vector machines and landsat imagery in the maipo river basin (central chile, 1975–2010). *Remote Sensing of Environment*, v. 137, p. 112–123, 2013. ISSN 0034-4257. 3.2.1

QUADRIANTO, N.; SHARMANSKA, V. Recycling privileged learning and distribution matching for fairness. Neural Information Processing Systems Foundation, Inc., 2017. 1

RAHMAWATI, D.; HUANG, Y.-P. Using c-support vector classification to forecast dengue fever epidemics in taiwan. In: IEEE. *2016 International Conference on System Science and Engineering (ICSSE)*. [S.l.], 2016. p. 1–4. 2.1.1.4

ROY, V. Convergence diagnostics for markov chain monte carlo. *https://doi.org/10.1146/annurev-statistics-031219-041300*, Annual Reviews, v. 7, p. 387–412, 3 2020. ISSN 2326831X. 3.2.2

SÆTRA, H. S. *AI for the sustainable development goals*. [S.l.]: CRC Press, 2022. 1

SEYMOUR, W. Detecting bias: does an algorithm have to be transparent in order to be fair? *BIAS 2018*, BIAS 2018, 2018. 2.1.2.3

SHI, S. et al. Algorithm bias detection and mitigation in lenovo face recognition engine. In: SPRINGER. *CCF International Conference on Natural Language Processing and Chinese Computing*. [S.l.], 2020. p. 442–453. 1

SHIELDS, M. D.; TEFERRA, K.; HAPIJ, A.; DADDAZIO, R. P. Refined stratified sampling for efficient monte carlo based uncertainty quantification. *Reliability Engineering & System Safety*, v. 142, p. 310–325, 2015. ISSN 0951-8320. 2, 3.2.1, 3.2.2

SMITH, J. J.; BEATTIE, L.; CRAMER, H. Scoping fairness objectives and identifying fairness metrics for recommender systems: The practitioners' perspective. In: *Proceedings of the ACM Web Conference 2023*. New York, NY, USA: Association for Computing Machinery, 2023. (WWW '23), p. 3648–3659. ISBN 9781450394161. Disponível em: <https://doi.org/10.1145/3543507.3583204>. 1, 2.2

SNEDECOR, G. W.; COCHRAN, W. G. *Statistical methods*. 7. ed. [S.l.]: The Iowa State University Press, 1980. 3

STOYANOVICH, J.; HOWE, B.; JAGADISH, H. Responsible data management. *Proceedings of the VLDB Endowment*, v. 13, n. 12, 2020. 2.1.2.3

SU, C.; YU, G.; WANG, J.; YAN, Z.; CUI, L. A review of causality-based fairness machine learning. *Intelligence & Robotics*, v. 2, p. 244–74, 2022. ISSN 2770-3541 (Online). Disponível em: <http://dx.doi.org/10.20517/ir.2022.17>. 2.2

TAUD, H.; MAS, J.-F. Multilayer perceptron (mlp). In: . [s.n.], 2018. Disponível em: <https://api.semanticscholar.org/CorpusID:67464997>. (document), 2.1

TRAVAINI, G. V.; PACCHIONI, F.; BELLUMORE, S.; BOSIA, M.; MICCO, F. D. Machine learning and criminal justice: A systematic review of advanced methodology for recidivism risk prediction. *Int. J. Environ. Res. Public Health*, MDPI AG, v. 19, n. 17, p. 10594, ago. 2022. 2.1.2

United Nations Educational, Scientific and Cultural Organization (UNESCO). *Convention against Discrimination in Education*. 1960. Adopted by the General Conference of UNESCO at its 11th session on 14 December 1960. Disponível em: <https://www.unesco.org/en>. 2.1.2.2

United Nations General Assembly. *International Convention on the Elimination of All Forms of Racial Discrimination*. 1965. Adopted and opened for signature and ratification by General Assembly resolution 2106 (XX) of 21 December 1965. Disponível em: <https://www.ohchr.org/en/instruments-mechanisms/instruments/international-convention-elimination-all-forms-racial>. 2.1.2.2

United Nations General Assembly. *Convention on the Elimination of All Forms of Discrimination against Women*. 1979. Adopted by the General Assembly on 18 December 1979. Disponível em: <https://www.ohchr.org/en/instruments-mechanisms/instruments/convention-elimination-all-forms-discrimination-against-women>. 2.1.2.2

United Nations General Assembly. *Convention on the Rights of Persons with Disabilities*. 2006. Adopted by the General Assembly on 13 December 2006. Disponível em: <https://www.ohchr.org/en/instruments-mechanisms/instruments/convention-rights-persons-disabilities>. 2.1.2.2

U.S. Congress. *Fair Housing Act*. 1968. Title VIII of the Civil Rights Act of 1968. Disponível em: <https://www.justice.gov>. 2.1.2.2

van Giffen, B.; HERHAUSEN, D.; FAHSE, T. Overcoming the pitfalls and perils of algorithms: A classification of machine learning biases and mitigation methods. *Journal of Business Research*, v. 144, p. 93–106, 2022. ISSN 0148-2963. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0148296322000881>. 2.1.2

VERMA, S.; RUBIN, J. Fairness definitions explained. In: *Proceedings of the International Workshop on Software Fairness*. New York, NY, USA: Association for Computing Machinery, 2018. (FairWare '18), p. 1–7. ISBN 9781450357463. Disponível em: <https://doi.org/10.1145/3194770.3194776>. 3.2

WING, J. M. Trustworthy ai. *Commun. ACM*, Association for Computing Machinery, New York, NY, USA, v. 64, n. 10, p. 64–71, sep 2021. ISSN 0001-0782. Disponível em: <https://doi.org/10.1145/3448248>. 2.1.2.3

WU, Q.; YE, Y.; ZHANG, H.; NG, M. K.; HO, S.-S. Forestexter: An efficient random forest algorithm for imbalanced text categorization. *Knowledge-Based Systems*, v. 67, p. 105–116, 2014. ISSN 0950-7051. 3.2.1

ZHANG, X.-D. A matrix algebra approach to artificial intelligence. Springer, 2020. 2.1.1

ZIMMERMANN, A.; LORENZ, A.; OPPERMANN, R. An operational definition of context. In: SPRINGER. *International and interdisciplinary conference on modeling and using context*. [S.l.], 2007. p. 558–571. 1

Appendix A

S. Supplementary Materials

Group	Metric	Correlation	Similar Terms
1	FNR Difference		
	FNR Ratio	0.9333	TP, FN
	Equalized Odds	0.6667	TP, FN
	Discovery ratio	0.5941	TP
	Discovery Difference	0.5774	TP
2	Predictive Parity		
	Equal Opportunity	0.5774	TP
	Average Odd Difference	0.5105	FP, TP
3	Error Difference		
4	Error Ratio		
	Disparate Impact	0.7395	FP
	Statistical Parity	0.7395	FP
5	FOR Ratio		
	FOR Difference	0.9	TN, FN
	Predictive Equality	0.5333	TN
6	FPR Ratio		

Table S.1: Terms similarity and correlation comparison for each metric group in Adult KNN Proposed experiment.

Group	Metric	Correlation	Similar Terms
1	Disparate Impact		
	Statistical Parity	0.7474	FP, TP
	FPR Ratio	0.6526	FP
	Average Odd Difference	0.5684	FP, TP
	Equal Opportunity	0.5426	TP
	Predictive Equality	0.5158	FP
2	Error Ratio		
3	FOR Difference		
	FOR Ratio	0.9263	TN, FN
4	Discovery ratio		
	Discovery Difference	0.9263	FP, TP
5	Predictive Parity		
6	Error Difference		
7	FNR Difference		
	FNR Ratio	0.9032	TP, FN
	Equalized Odds	0.7554	TP, FN

Table S.2: Terms similarity and correlation comparison for each metric group in Adult Logit Proposed experiment.

Group	Metric	Correlation	Similar Terms
1	Predictive Parity		
2	Disparate Impact		
	Statistical Parity	0.8905	FP, TP
	FPR Ratio	0.7	FP
3	Discovery ratio		
	Discovery Difference	0.8519	FP, TP
4	Error Ratio		
5	Equalized Odds		
	FNR Ratio	0.6637	TP, FN
	Error Difference	0.6612	FP, FN
	FNR Difference	0.6283	TP, FN
6	FOR Ratio		
	FOR Difference	0.9167	TN, FN
7	Average Odd Difference		
	Equal Opportunity	0.8761	TP, FN
	Predictive Equality	0.6336	FP, TN

Table S.3: Terms similarity and correlation comparison for each metric group in Adult MLP Proposed experiment.

Group	Metric	Correlation	Similar Terms
1	FPR Ratio		
	Predictive Equality	0.6565	FP, TN
	Discovery ratio	0.6	FP
	Disparate Impact	0.5649	FP
	Discovery Difference	0.5344	FP
	FOR Difference	0.5344	TN
2	Predictive Parity		
3	FOR Ratio		
4	Equalized Odds		
	Error Difference	0.5992	FP, FN
	FNR Ratio	0.5897	TP, FN
	FNR Difference	0.5726	TP, FN
5	Average Odd Difference		
	Equal Opportunity	0.9316	TP, FN
6	Statistical Parity		
7	Error Ratio		

Table S.4: Terms similarity and correlation comparison for each metric group in Adult RF Proposed experiment.

Group	Metric	Correlation	Similar Terms
1	Error Ratio		
	Disparate Impact	0.9496	FP
	Statistical Parity	0.7815	FP
2	Discovery Difference		
	Discovery ratio	0.9167	FP, TP
3	FOR Difference		
	FOR Ratio	0.8652	TN, FN
4	Equalized Odds		
5	Equal Opportunity		
6	Predictive Parity		
7	Predictive Equality		
	FPR Ratio	0.9123	FP, TN
	Average Odd Difference	0.6182	FP, TN
8	FNR Difference		
	FNR Ratio	0.707	TP, FN
9	Error Difference		

Table S.5: Terms similarity and correlation comparison for each metric group in Adult SVC Proposed experiment.

Group	Metric	Correlation	Similar Terms
1	Error Difference		
	Error Ratio	0.6603	FP, FN
	Statistical Parity	0.598	FP
	Disparate Impact	0.5448	FP
2	Discovery ratio		
	Discovery Difference	0.9503	FP, TP
3	FOR Difference		
	FOR Ratio	0.9017	TN, FN
4	Predictive Parity		
5	Average Odd Difference		
	Equalized Odds	0.7661	FP, TN, TP, FN
	Equal Opportunity	0.7243	TP, FN
6	FPR Ratio		
	Predictive Equality	0.9081	FP, TN
7	FNR Difference		
	FNR Ratio	0.9638	TP, FN

Table S.6: Terms similarity and correlation comparison for each metric group in Bank KNN Proposed experiment.

Group	Metric	Correlation	Similar Terms
1	FPR Ratio		
	Predictive Equality	0.9119	FP, TN
2	Discovery Difference		
	Discovery ratio	0.9695	FP, TP
3	FNR Ratio		
	FNR Difference	0.9638	TP, FN
	FOR Ratio	0.5989	FN
	FOR Difference	0.504	FN
4	Predictive Parity		
5	Error Ratio		
	Error Difference	0.7882	FP, FN
6	Average Odd Difference		
	Equal Opportunity	0.7356	TP, FN
	Equalized Odds	0.5593	FP, TN, TP, FN
7	Disparate Impact		
	Statistical Parity	0.9345	FP, TP

Table S.7: Terms similarity and correlation comparison for each metric group in Bank Logit Proposed experiment.

Group	Metric	Correlation	Similar Terms
1	Error Ratio		
	Disparate Impact	0.7756	FP
	Statistical Parity	0.7738	FP
	Error Difference	0.5538	FP, FN
2	Discovery ratio		
	Discovery Difference	0.8689	FP, TP
3	FNR Difference		
	FNR Ratio	0.89	TP, FN
4	FPR Ratio		
	Predictive Equality	0.8236	FP, TN
5	FOR Ratio		
	FOR Difference	0.6402	TN, FN
6	Equal Opportunity		
	Average Odd Difference	0.7244	TP, FN
7	Predictive Parity		
8	Equalized Odds		

Table S.8: Terms similarity and correlation comparison for each metric group in Bank MLP Proposed experiment.

Group	Metric	Correlation	Similar Terms
1	Error Ratio		
	Statistical Parity	0.6328	FP
	Disparate Impact	0.6188	FP
	Error Difference	0.5201	FP, FN
2	Discovery Difference		
	Discovery ratio	0.9288	FP, TP
	FPR Ratio	0.6826	FP
	Predictive Equality	0.5729	FP
3	FOR Ratio		
	FOR Difference	0.8746	TN, FN
4	Equalized Odds		
	Average Odd Difference	0.9785	FP, TN, TP, FN
	Equal Opportunity	0.8723	TP, FN
5	Predictive Parity		
6	FNR Ratio		
	FNR Difference	0.9458	TP, FN

Table S.9: Terms similarity and correlation comparison for each metric group in Bank RF Proposed experiment.

Group	Metric	Correlation	Similar Terms
1	Discovery Difference		
	Discovery ratio	0.709	FP, TP
2	Average Odd Difference		
	Equal Opportunity	0.85	TP, FN
	Equalized Odds	0.648	FP, TN, TP, FN
3	FNR Ratio		
	FNR Difference	0.7927	TP, FN
4	FOR Difference		
	FOR Ratio	0.7479	TN, FN
5	Statistical Parity		
	Disparate Impact	0.9814	FP, TP
	Error Difference	0.6698	FP
	Error Ratio	0.5418	FP
6	Predictive Equality		
	FPR Ratio	0.8452	FP, TN
7	Predictive Parity		

Table S.10: Terms similarity and correlation comparison for each metric group in Bank SVC Proposed experiment.

Group	Metric	Correlation	Similar Terms
1	Discovery Difference		
	Discovery ratio	0.9349	FP, TP
	FNR Ratio	0.6963	TP
	FNR Difference	0.6815	TP
2	Statistical Parity		
	FPR Ratio	0.9063	FP
	Predictive Equality	0.8833	FP
	Disparate Impact	0.8754	FP, TP
	Error Ratio	0.8541	FP
	Error Difference	0.8011	FP
	Average Odd Difference	0.7333	FP, TP
3	Equalized Odds		
4	Equal Opportunity		
	Predictive Parity	0.6815	TP
5	FOR Ratio		
	FOR Difference	0.9895	TN, FN

Table S.11: Terms similarity and correlation comparison for each metric group in Compas KNN Proposed experiment.

Group	Metric	Correlation	Similar Terms
1	FNR Ratio		
	FNR Difference	0.9604	TP, FN
2	Equalized Odds		
3	Average Odd Difference		
	Equal Opportunity	0.7677	TP, FN
	Predictive Equality	0.5939	FP, TN
	Disparate Impact	0.5631	FP, TP
	Statistical Parity	0.5616	FP, TP
4	Error Difference		
	Error Ratio	0.8835	FP, FN
	FPR Ratio	0.8686	FP
	Discovery Difference	0.8341	FP
	Discovery ratio	0.8049	FP
	FOR Difference	0.6639	FN
	FOR Ratio	0.6608	FN
5	Predictive Parity		

Table S.12: Terms similarity and correlation comparison for each metric group in Compas Logit Proposed experiment.

Group	Metric	Correlation	Similar Terms
1	Predictive Equality		
	Error Ratio	0.75	FP
	Disparate Impact	0.75	FP
	FPR Ratio	0.7342	FP, TN
	Statistical Parity	0.7167	FP
	Error Difference	0.7	FP
	Average Odd Difference	0.65	FP, TN
	FOR Difference	0.5833	TN
	FOR Ratio	0.5667	TN
2	Predictive Parity		
3	FNR Difference		
	FNR Ratio	0.8918	TP, FN
4	Equal Opportunity		
5	Equalized Odds		
6	Discovery Difference		
	Discovery ratio	0.9394	FP, TP

Table S.13: Terms similarity and correlation comparison for each metric group in Compas MLP Proposed experiment.

Group	Metric	Correlation	Similar Terms
1	FNR Ratio		
	FNR Difference	0.9306	TP, FN
2	FPR Ratio		
	Disparate Impact	0.9234	FP
	Error Ratio	0.9063	FP
	Statistical Parity	0.8379	FP
	Error Difference	0.8156	FP
	Discovery ratio	0.7719	FP
	Predictive Equality	0.7695	FP, TN
	Discovery Difference	0.7182	FP
	FOR Difference	0.5814	TN
	FOR Ratio	0.5814	TN
3	Average Odd Difference		
	Equal Opportunity	0.8167	TP, FN
4	Predictive Parity		
5	Equalized Odds		

Table S.14: Terms similarity and correlation comparison for each metric group in Compas RF Proposed experiment.

Group	Metric	Correlation	Similar Terms
1	Discovery Difference		
	Discovery ratio	0.9734	FP, TP
2	Predictive Equality		
	FPR Ratio	1.0	FP, TN
3	FOR Difference		
	FOR Ratio	0.9667	TN, FN
	Equalized Odds	0.6364	TN, FN
4	Disparate Impact		
	Error Ratio	0.8737	FP
	Statistical Parity	0.8632	FP, TP
	Error Difference	0.8179	FP
5	FNR Ratio		
	FNR Difference	0.9092	TP, FN
6	Average Odd Difference		
	Equal Opportunity	0.9167	TP, FN
7	Predictive Parity		

Table S.15: Terms similarity and correlation comparison for each metric group in Compas SVC Proposed experiment.

Group	Metric	Correlation	Similar Terms
1	Discovery Difference		
	Discovery ratio	1.0	FP, TP
	FOR Difference	1.0	
	FOR Ratio	1.0	
	FNR Difference	0.7857	TP
	FNR Ratio	0.7857	TP
	Error Ratio	0.6429	FP
	Error Difference	0.6183	FP
2	Equal Opportunity		
	Predictive Parity	0.7857	TP
	Average Odd Difference	0.7143	TP, FN
3	FPR Ratio		
	Predictive Equality	0.982	FP, TN
	Disparate Impact	0.6183	FP
	Statistical Parity	0.6183	FP
	Equalized Odds	0.5455	FP, TN

Table S.16: Terms similarity and correlation comparison for each metric group in German KNN Proposed experiment.

Group	Metric	Correlation	Similar Terms
1	Predictive Parity		
2	Predictive Equality		
	FPR Ratio	1.0	FP, TN
	Disparate Impact	1.0	FP
	Statistical Parity	1.0	FP
	Error Difference	0.6667	FP
	Equal Opportunity	0.6667	
	Average Odd Difference	0.6667	FP, TN
3	FNR Difference		
	Equalized Odds	1.0	TP, FN
	FNR Ratio	1.0	TP, FN
4	FOR Ratio		
	Discovery Difference	1.0	
	Discovery ratio	1.0	
	FOR Difference	1.0	TN, FN
	Error Ratio	0.6667	FN

Table S.17: Terms similarity and correlation comparison for each metric group in German Logit Proposed experiment.

Group	Metric	Correlation	Similar Terms
1	FOR Difference		
	Discovery Difference	1.0	
	Discovery ratio	1.0	
	FPR Ratio	1.0	TN
	Error Ratio	1.0	FN
	FOR Ratio	1.0	TN, FN
	FNR Difference	1.0	FN
	Error Difference	0.7143	FN
	FNR Ratio	0.7143	FN
2	Equalized Odds		
	Disparate Impact	1.0	FP, TP
	Statistical Parity	1.0	FP, TP
	Equal Opportunity	1.0	TP, FN
	Predictive Parity	1.0	FP, TP
	Average Odd Difference	0.7778	FP, TN, TP, FN
3	Predictive Equality		

Table S.18: Terms similarity and correlation comparison for each metric group in German MLP Proposed experiment.

Group	Metric	Correlation	Similar Terms
1	Error Ratio		
	FOR Difference	0.8281	FN
	FOR Ratio	0.8281	FN
	Error Difference	0.7143	FP, FN
	Discovery Difference	0.6667	FP
	Discovery ratio	0.6667	FP
	FNR Difference	0.6447	FN
	FNR Ratio	0.6447	FN
	Predictive Equality	0.6429	FP
2	Disparate Impact		
	Statistical Parity	1.0	FP, TP
	Predictive Parity	0.527	FP, TP
3	Equal Opportunity		
	Average Odd Difference	0.8563	TP, FN
	Equalized Odds	0.7006	TP, FN
4	FPR Ratio		

Table S.19: Terms similarity and correlation comparison for each metric group in German RF Proposed experiment.

Group	Metric	Correlation	Similar Terms
1	Error Difference		
	Equalized Odds	0.6	FP, FN
	Error Ratio	0.6	FP, FN
	Predictive Equality	0.6	FP
	FPR Ratio	0.6	FP
2	FOR Difference		
	FOR Ratio	1.0	TN, FN
	Discovery Difference	0.8	
	Discovery ratio	0.8	
	FNR Difference	0.8	FN
	FNR Ratio	0.6	FN
3	Average Odd Difference		
	Disparate Impact	1.0	FP, TP
	Statistical Parity	1.0	FP, TP
	Equal Opportunity	0.8	TP, FN
	Predictive Parity	0.8	FP, TP

Table S.20: Terms similarity and correlation comparison for each metric group in German SVC Proposed experiment.

Group	Metric	Correlation	Similar Terms
1	Average Odd Difference		
	Equal Opportunity	0.8942	TP, FN
	Predictive Equality	0.6891	FP, TN
	FPR Ratio	0.6337	FP, TN
	Predictive Parity	0.5031	FP, TP
2	Equalized Odds		
3	Statistical Parity		
	Disparate Impact	0.9592	FP, TP
	Error Ratio	0.9333	FP
4	FNR Difference		
	FNR Ratio	0.9527	TP, FN
	Discovery Difference	0.5217	TP
	Discovery ratio	0.5143	TP
5	FOR Ratio		
	FOR Difference	0.963	TN, FN
6	Error Difference		

Table S.21: Terms similarity and correlation comparison for each metric group in Adult KNN Base experiment.

Group	Metric	Correlation	Similar Terms
1	FOR Difference		
	FOR Ratio	0.9567	TN, FN
	Predictive Parity	0.6996	
	Predictive Equality	0.637	TN
	FPR Ratio	0.5654	TN
	Average Odd Difference	0.5135	TN, FN
2	FNR Difference		
	FNR Ratio	0.6518	TP, FN
	Equalized Odds	0.5569	TP, FN
3	Equal Opportunity		
4	Discovery Difference		
	Discovery ratio	0.9742	FP, TP
5	Error Ratio		
	Disparate Impact	0.9903	FP
	Statistical Parity	0.9389	FP
6	Error Difference		

Table S.22: Terms similarity and correlation comparison for each metric group in Adult Logit Base experiment.

Group	Metric	Correlation	Similar Terms
1	Predictive Parity		
	FOR Ratio	0.5947	
	FOR Difference	0.5639	
2	Error Ratio		
	Disparate Impact	0.9881	FP
	Statistical Parity	0.8685	FP
3	Error Difference		
4	FNR Difference		
	Equalized Odds	0.6944	TP, FN
	FNR Ratio	0.5544	TP, FN
5	Average Odd Difference		
	Predictive Equality	0.8308	FP, TN
	FPR Ratio	0.8069	FP, TN
	Equal Opportunity	0.7772	TP, FN
6	Discovery Difference		
	Discovery ratio	0.9315	FP, TP

Table S.23: Terms similarity and correlation comparison for each metric group in Adult MLP Base experiment.

Group	Metric	Correlation	Similar Terms
1	Equalized Odds		
	FNR Difference	0.6882	TP, FN
2	Statistical Parity		
	Disparate Impact	0.9037	FP, TP
	Error Ratio	0.8876	FP
3	Error Difference		
4	FNR Ratio		
	FOR Ratio	0.5025	FN
5	Discovery Difference		
	Discovery ratio	0.8733	FP, TP
6	FPR Ratio		
	Average Odd Difference	0.7705	FP, TN
	Predictive Equality	0.6784	FP, TN
	FOR Difference	0.6151	TN
	Equal Opportunity	0.5244	
7	Predictive Parity		

Table S.24: Terms similarity and correlation comparison for each metric group in Adult RF Base experiment.

Group	Metric	Correlation	Similar Terms
1	Disparate Impact		
	Error Ratio	0.9976	FP
	Statistical Parity	0.962	FP, TP
2	Equalized Odds		
	Error Difference	0.5819	FP, FN
	FNR Difference	0.5229	TP, FN
3	FPR Ratio		
	Predictive Equality	0.978	FP, TN
	Average Odd Difference	0.9235	FP, TN
4	FOR Difference		
	FOR Ratio	0.6697	TN, FN
	Predictive Parity	0.5429	
5	Equal Opportunity		
6	Discovery Difference		
	Discovery ratio	0.998	FP, TP
7	FNR Ratio		

Table S.25: Terms similarity and correlation comparison for each metric group in Adult SVC Base experiment.

Group	Metric	Correlation	Similar Terms
1	Disparate Impact		
	Statistical Parity	0.9983	FP, TP
	Error Ratio	0.998	FP
	Discovery ratio	0.765	FP, TP
	Discovery Difference	0.7583	FP, TP
	Error Difference	0.7541	FP
2	Average Odd Difference		
	Equalized Odds	0.919	FP, TN, TP, FN
	Predictive Equality	0.8561	FP, TN
	FPR Ratio	0.7605	FP, TN
	Equal Opportunity	0.7001	TP, FN
	Predictive Parity	0.6926	FP, TP
	FOR Ratio	0.5725	TN, FN
	FOR Difference	0.5672	TN, FN
3	FNR Ratio		
	FNR Difference	0.8847	TP, FN

Table S.26: Terms similarity and correlation comparison for each metric group in Bank KNN Base experiment.

Group	Metric	Correlation	Similar Terms
1	Average Odd Difference		
	Equalized Odds	0.993	FP, TN, TP, FN
	Predictive Equality	0.9647	FP, TN
	Equal Opportunity	0.9166	TP, FN
	FPR Ratio	0.8715	FP, TN
	Predictive Parity	0.738	FP, TP
2	Error Ratio		
	Disparate Impact	0.999	FP
	Statistical Parity	0.9969	FP
	Discovery ratio	0.893	FP
	Discovery Difference	0.8213	FP
	FNR Difference	0.7471	FN
	FNR Ratio	0.7123	FN
	Error Difference	0.6813	FP, FN
3	FOR Ratio		
	FOR Difference	0.9524	TN, FN

Table S.27: Terms similarity and correlation comparison for each metric group in Bank Logit Base experiment.

Group	Metric	Correlation	Similar Terms
1	Disparate Impact		
	Error Ratio	0.995	FP
	Statistical Parity	0.9932	FP, TP
	Discovery Difference	0.7904	FP, TP
	Discovery ratio	0.7001	FP, TP
2	FNR Difference		
	FNR Ratio	0.7926	TP, FN
3	FOR Ratio		
	Equalized Odds	0.5753	TN, FN
	Predictive Equality	0.5395	TN
	Predictive Parity	0.5348	
4	FOR Difference		
5	Average Odd Difference		
	Equal Opportunity	0.7969	TP, FN
	FPR Ratio	0.6131	FP, TN
6	Error Difference		

Table S.28: Terms similarity and correlation comparison for each metric group in Bank MLP Base experiment.

Group	Metric	Correlation	Similar Terms
1	FNR Ratio		
	Statistical Parity	0.8127	TP
	Error Ratio	0.7827	FN
	Disparate Impact	0.7802	TP
	Discovery Difference	0.623	TP
	Error Difference	0.6094	FN
	Discovery ratio	0.6057	TP
	FOR Ratio	0.5517	FN
	FNR Difference	0.5243	TP, FN
2	Predictive Equality		
	Average Odd Difference	0.9271	FP, TN
	Equalized Odds	0.9139	FP, TN
	Equal Opportunity	0.7557	
	FPR Ratio	0.6299	FP, TN
	Predictive Parity	0.503	FP
3	FOR Difference		

Table S.29: Terms similarity and correlation comparison for each metric group in Bank RF Base experiment.

Group	Metric	Correlation	Similar Terms
1	FNR Ratio		
	FOR Ratio	0.5831	FN
	Discovery Difference	0.5696	TP
	Error Ratio	0.5575	FN
	Statistical Parity	0.5516	TP
	Disparate Impact	0.547	TP
	FNR Difference	0.5386	TP, FN
	Discovery ratio	0.5022	TP
2	Equalized Odds		
	Average Odd Difference	0.9528	FP, TN, TP, FN
	Predictive Equality	0.9527	FP, TN
	FPR Ratio	0.9182	FP, TN
	Predictive Parity	0.6264	FP, TP
3	FOR Difference		
	Error Difference	0.6618	FN
4	Equal Opportunity		

Table S.30: Terms similarity and correlation comparison for each metric group in Bank SVC Base experiment.

Group	Metric	Correlation	Similar Terms
1	Discovery ratio		
	Discovery Difference	0.9965	FP, TP
	FNR Difference	0.6603	TP
	FNR Ratio	0.6128	TP
2	FOR Difference		
	FOR Ratio	0.9983	TN, FN
	Predictive Parity	0.6582	
3	Average Odd Difference		
	Equal Opportunity	0.9159	TP, FN
	Predictive Equality	0.6572	FP, TN
	FPR Ratio	0.6235	FP, TN
	Disparate Impact	0.5435	FP, TP
	Statistical Parity	0.5429	FP, TP
	Error Ratio	0.5152	FP, FN
	Error Difference	0.5136	FP, FN
4	Equalized Odds		

Table S.31: Terms similarity and correlation comparison for each metric group in Compas KNN Base experiment.

Group	Metric	Correlation	Similar Terms
1	FPR Ratio		
	Disparate Impact	0.8494	FP
	Error Ratio	0.7843	FP
	Predictive Equality	0.7626	FP, TN
	FOR Ratio	0.647	TN
	FOR Difference	0.6431	TN
	Statistical Parity	0.6199	FP
	Error Difference	0.5851	FP
2	Equalized Odds		
3	Average Odd Difference		
	Equal Opportunity	0.6672	TP, FN
4	Discovery ratio		
	Discovery Difference	0.887	FP, TP
5	FNR Ratio		
	FNR Difference	0.8558	TP, FN
6	Predictive Parity		

Table S.32: Terms similarity and correlation comparison for each metric group in Compas Logit Base experiment.

Group	Metric	Correlation	Similar Terms
1	FPR Ratio		
	Predictive Equality	0.9486	FP, TN
	Error Ratio	0.9011	FP
	Disparate Impact	0.8832	FP
	Statistical Parity	0.8677	FP
	Error Difference	0.8656	FP
	Average Odd Difference	0.6749	FP, TN
	FOR Difference	0.6579	TN
	FOR Ratio	0.6334	TN
	Discovery ratio	0.5801	FP
	Discovery Difference	0.5776	FP
2	FNR Difference		
	FNR Ratio	0.9458	TP, FN
3	Predictive Parity		
	Equal Opportunity	0.7185	TP
	Equalized Odds	0.6248	FP, TP

Table S.33: Terms similarity and correlation comparison for each metric group in Compas MLP Base experiment.

Group	Metric	Correlation	Similar Terms
1	FPR Ratio		
	Predictive Equality	0.9637	FP, TN
	Error Ratio	0.9235	FP
	Average Odd Difference	0.9131	FP, TN
	Disparate Impact	0.9046	FP
	Error Difference	0.9027	FP
	Statistical Parity	0.8969	FP
	Discovery ratio	0.6337	FP
	Discovery Difference	0.6317	FP
	FOR Difference	0.6262	TN
	Equal Opportunity	0.5979	
	FOR Ratio	0.5835	TN
2	Predictive Parity		
	Equalized Odds	0.6012	FP, TP
3	FNR Difference		
	FNR Ratio	0.9584	TP, FN

Table S.34: Terms similarity and correlation comparison for each metric group in Compas RF Base experiment.

Group	Metric	Correlation	Similar Terms
1	Statistical Parity		
	Disparate Impact	0.9993	FP, TP
	Error Ratio	0.9846	FP
	Error Difference	0.9813	FP
2	FNR Ratio		
	FNR Difference	0.7614	TP, FN
3	Predictive Equality		
	FPR Ratio	0.9377	FP, TN
	Average Odd Difference	0.9268	FP, TN
	Equal Opportunity	0.7735	
	Equalized Odds	0.6134	FP, TN
4	Discovery Difference		
	Discovery ratio	0.9999	FP, TP
5	FOR Ratio		
	FOR Difference	0.9939	TN, FN
	Predictive Parity	0.8784	

Table S.35: Terms similarity and correlation comparison for each metric group in Compas SVC Base experiment

Group	Metric	Correlation	Similar Terms
1	FOR Ratio		
	FOR Difference	0.9843	TN, FN
	Discovery Difference	0.8222	
	Error Ratio	0.8204	FN
	FNR Ratio	0.8197	FN
	Discovery ratio	0.8125	
	FNR Difference	0.7918	FN
	Predictive Equality	0.6993	TN
	Error Difference	0.6344	FN
	Equalized Odds	0.6188	TN, FN
	FPR Ratio	0.5714	TN
2	Disparate Impact		
	Statistical Parity	0.9814	FP, TP
	Average Odd Difference	0.7235	FP, TP
3	Equal Opportunity		
	Predictive Parity	0.5766	TP

Table S.36: Terms similarity and correlation comparison for each metric group in German KNN Base experiment.

Group	Metric	Correlation	Similar Terms
1	Statistical Parity		
	Disparate Impact	0.9917	FP, TP
	FPR Ratio	0.9314	FP
	Discovery ratio	0.8567	FP, TP
	Discovery Difference	0.6053	FP, TP
	Average Odd Difference	0.5382	FP, TP
	Equal Opportunity	0.5058	TP
2	FOR Ratio		
	FOR Difference	0.9914	TN, FN
	FNR Ratio	0.8865	FN
	Error Ratio	0.7588	FN
	FNR Difference	0.7439	FN
	Error Difference	0.6792	FN
3	Predictive Equality		
4	Predictive Parity		
5	Equalized Odds		

Table S.37: Terms similarity and correlation comparison for each metric group in German Logit Base experiment.

Group	Metric	Correlation	Similar Terms
1	FNR Difference		
	FNR Ratio	0.7641	TP, FN
	FOR Ratio	0.7618	FN
	FOR Difference	0.7479	FN
	Error Ratio	0.7319	FN
	FPR Ratio	0.54	
	Discovery Difference	0.5317	TP
	Predictive Equality	0.5285	
2	Discovery ratio		
3	Equalized Odds		
4	Equal Opportunity		
	Average Odd Difference	0.9008	TP, FN
	Predictive Parity	0.5317	TP
5	Statistical Parity		
	Disparate Impact	0.9753	FP, TP
6	Error Difference		

Table S.38: Terms similarity and correlation comparison for each metric group in German MLP Base experiment.

Group	Metric	Correlation	Similar Terms
1	Average Odd Difference		
	Equal Opportunity	0.9646	TP, FN
	Predictive Equality	0.5142	FP, TN
2	Predictive Parity		
	Statistical Parity	0.8788	FP, TP
	Disparate Impact	0.8407	FP, TP
	Error Difference	0.5445	FP
3	Equalized Odds		
4	FNR Ratio		
	FOR Ratio	0.9862	FN
	FNR Difference	0.9175	TP, FN
	FOR Difference	0.8848	FN
5	Error Ratio		
6	Discovery ratio		
	Discovery Difference	0.9892	FP, TP
	FPR Ratio	0.8519	FP

Table S.39: Terms similarity and correlation comparison for each metric group in German RF Base experiment.

Group	Metric	Correlation	Similar Terms
1	Equalized Odds		
2	Discovery Difference		
	Discovery ratio	0.9983	FP, TP
	Error Ratio	0.9089	FP
	FNR Difference	0.89	TP
	FOR Difference	0.8073	
	FOR Ratio	0.6759	
	FPR Ratio	0.5958	FP
	Predictive Equality	0.5944	FP
	FNR Ratio	0.5279	TP
3	Disparate Impact		
	Statistical Parity	0.9983	FP, TP
	Average Odd Difference	0.9154	FP, TP
	Equal Opportunity	0.8276	TP
	Error Difference	0.6288	FP
	Predictive Parity	0.5938	FP, TP

Table S.40: Terms similarity and correlation comparison for each metric group in German SVC Base experiment.

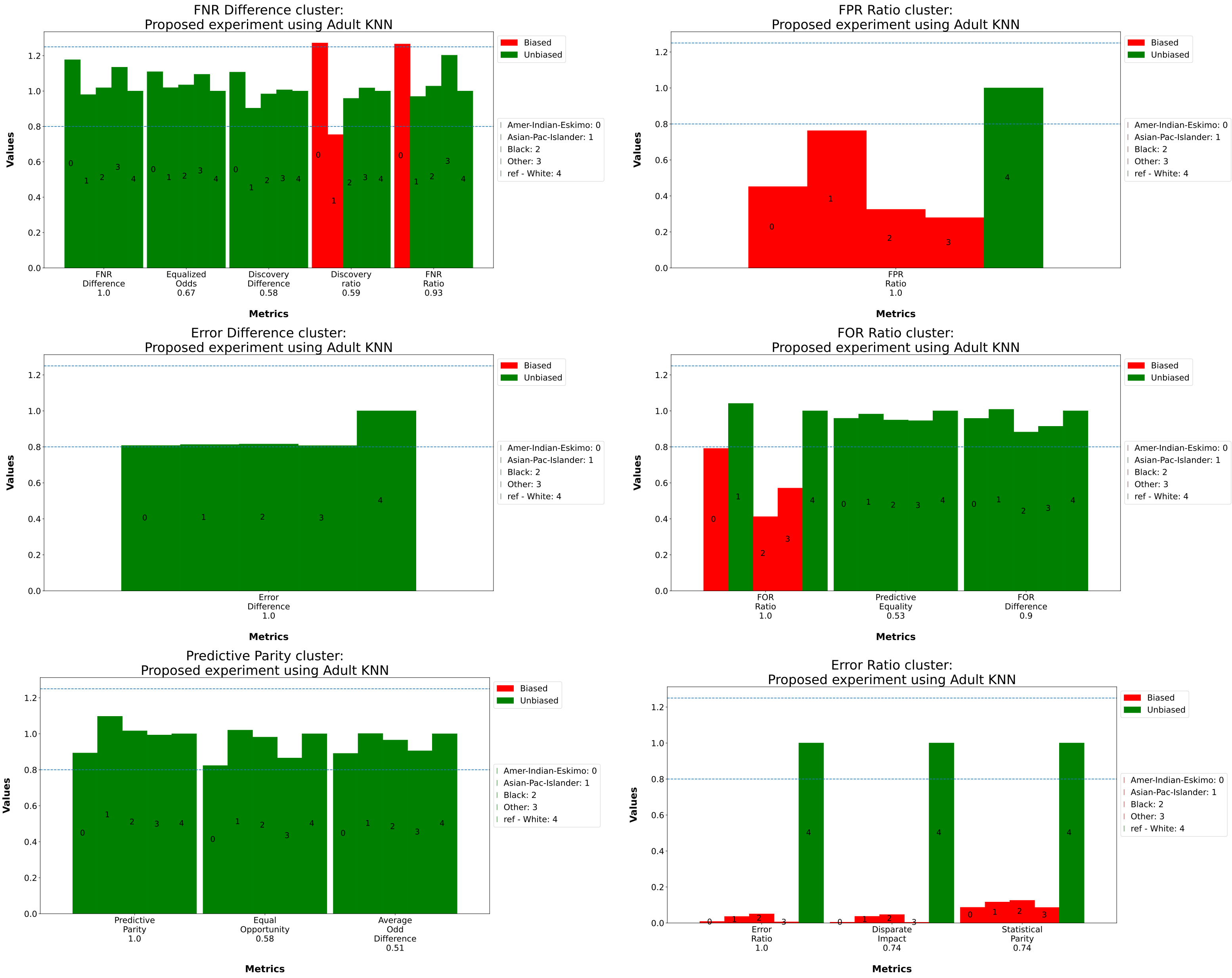


Figure S.1: Proposed Experiment clusters using Adult dataset and KNN model.

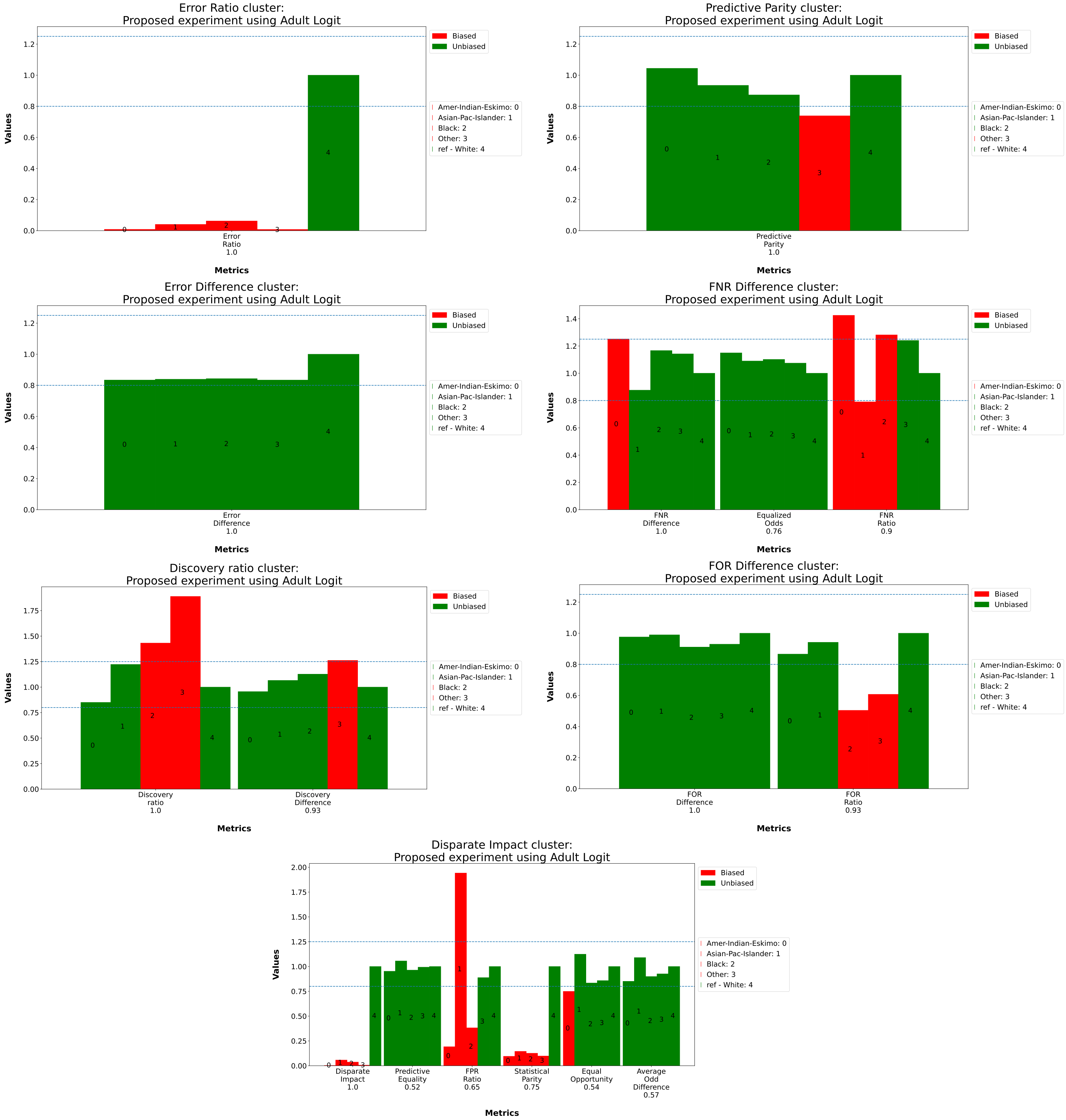


Figure S.2: Proposed Experiment clusters using Adult dataset and Logit model.

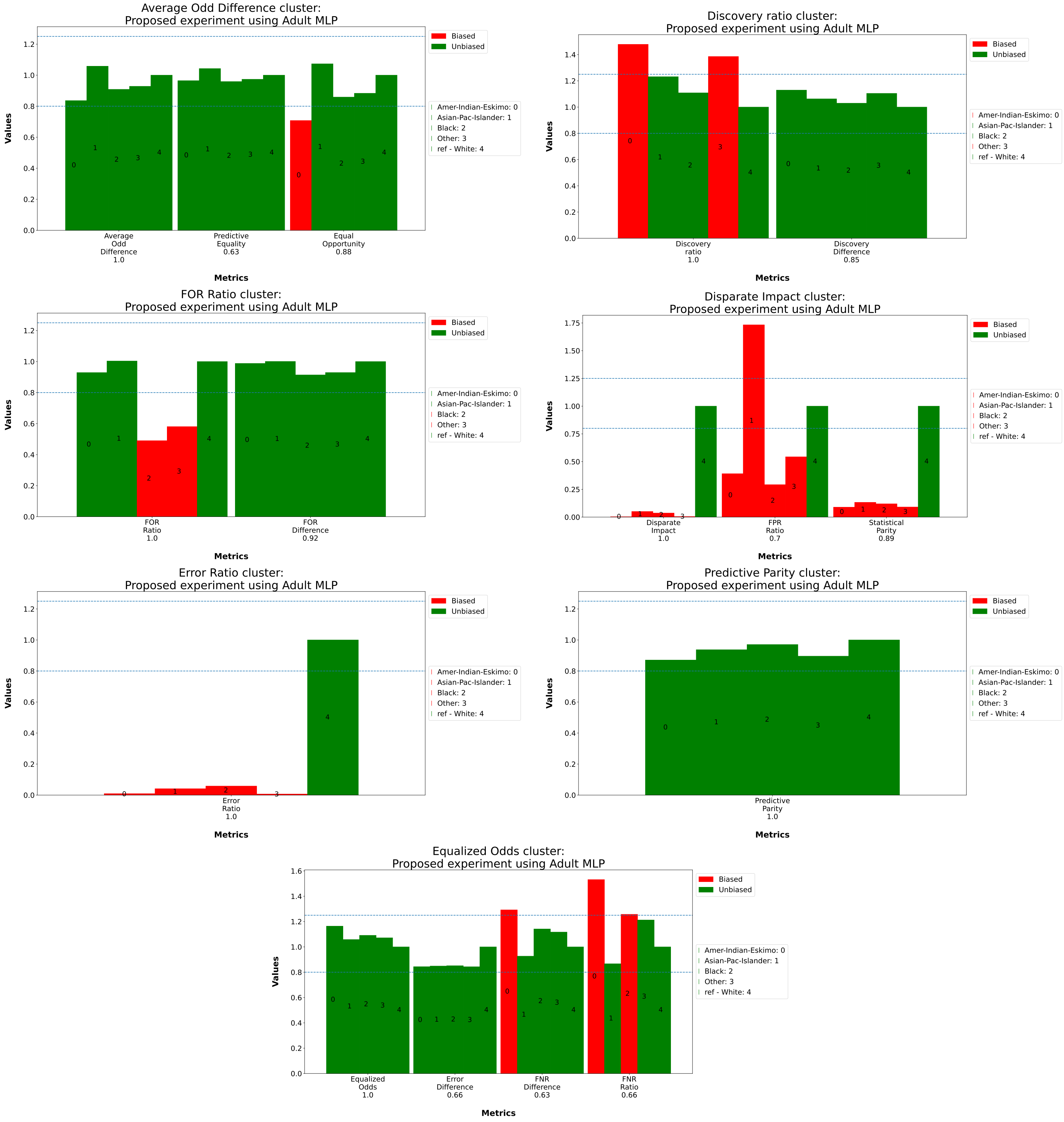


Figure S.3: Proposed Experiment clusters using Adult dataset and MLP model.

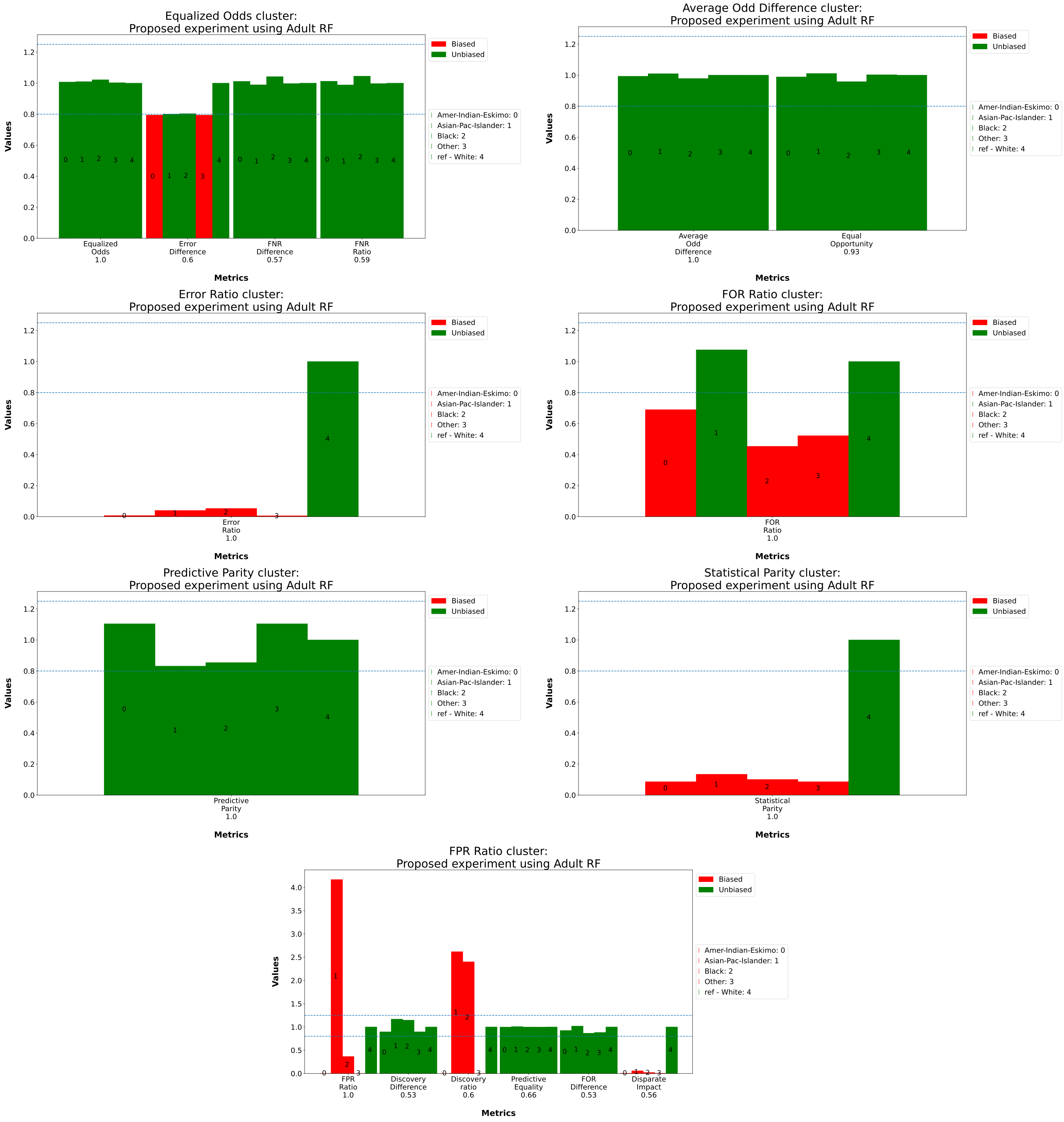


Figure S.4: Proposed Experiment clusters using Adult dataset and RF model.

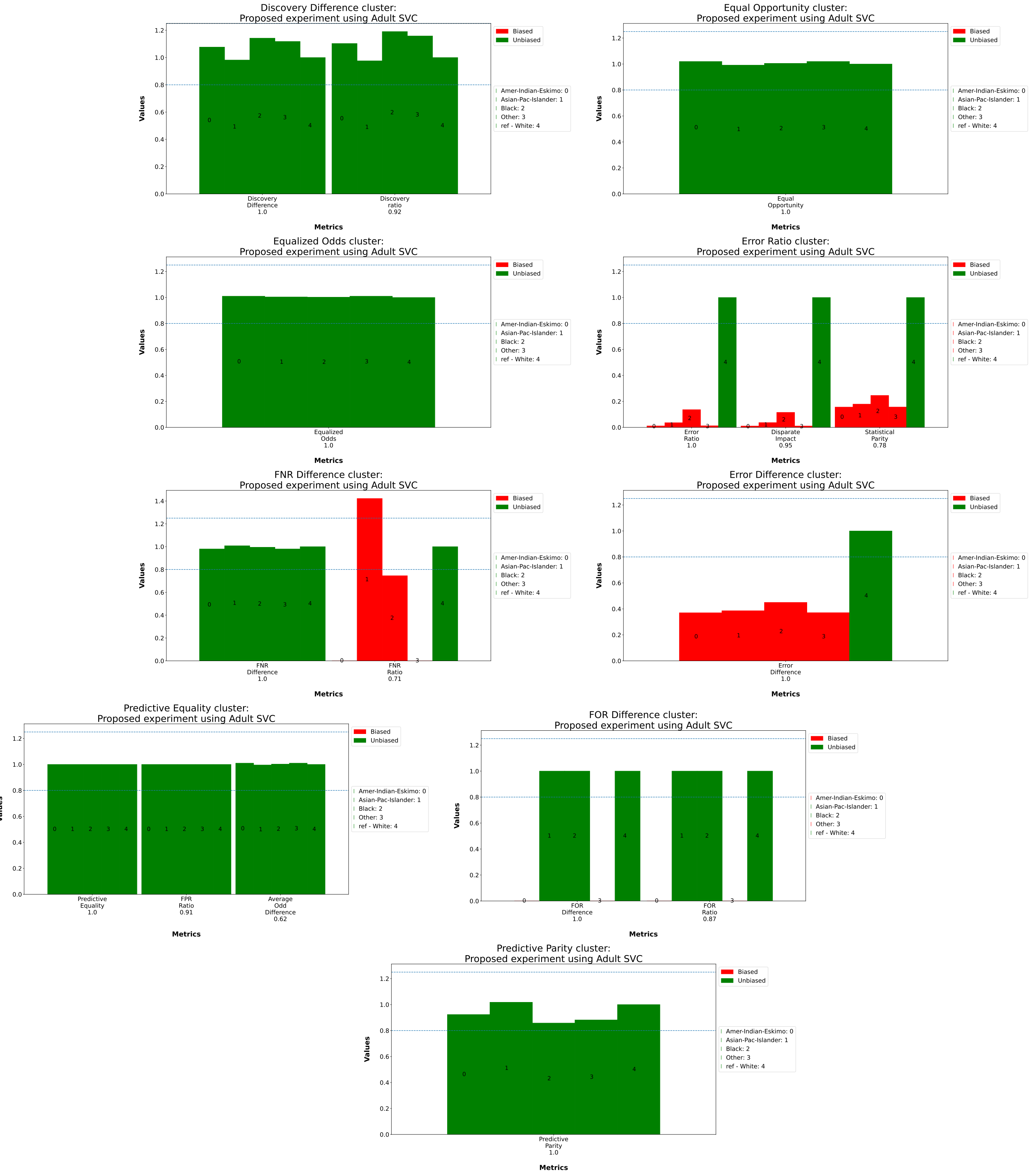


Figure S.5: Proposed Experiment clusters using Adult dataset and SVC model.

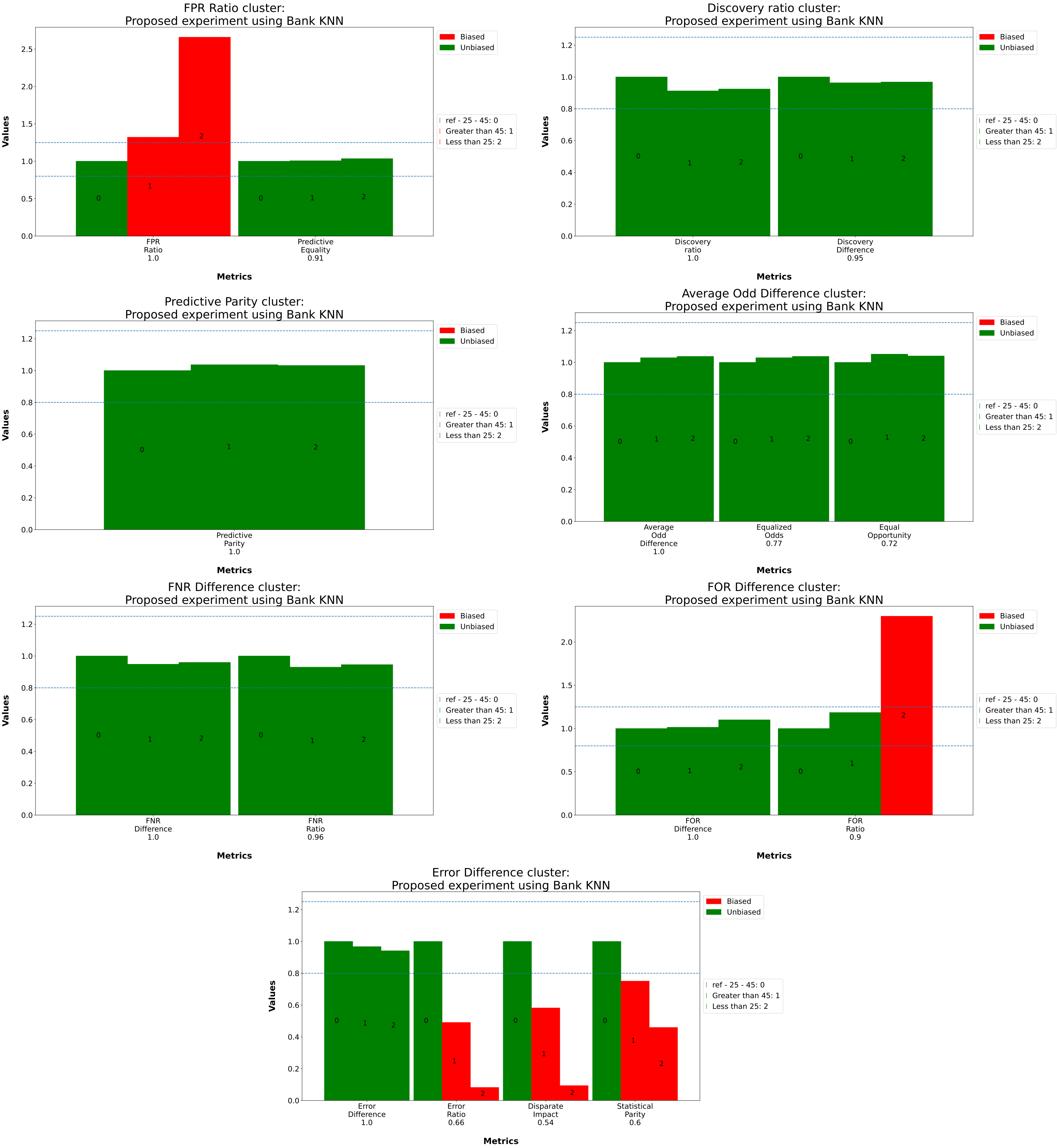


Figure S.6: Proposed Experiment clusters using Bank dataset and KNN model.

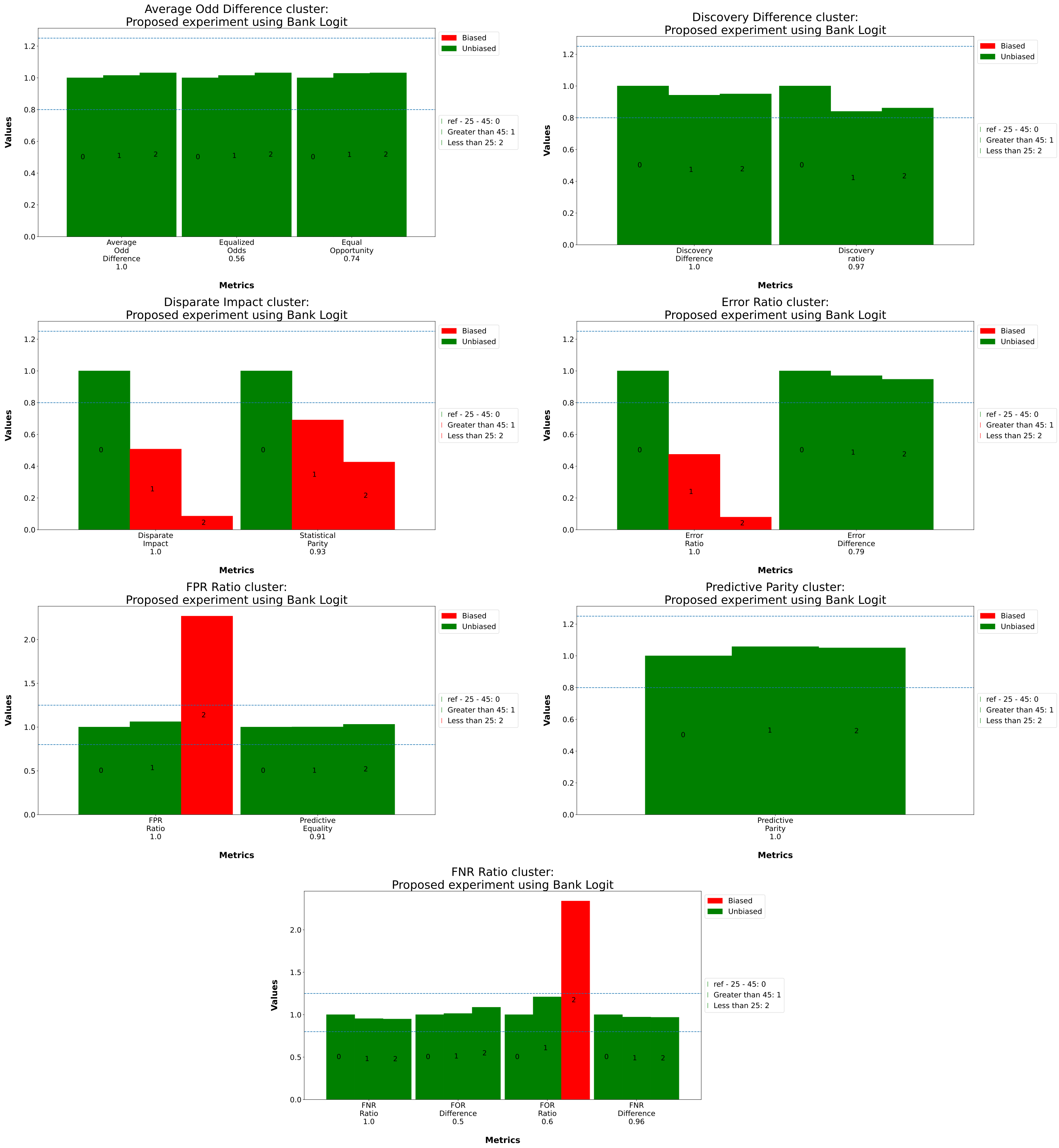
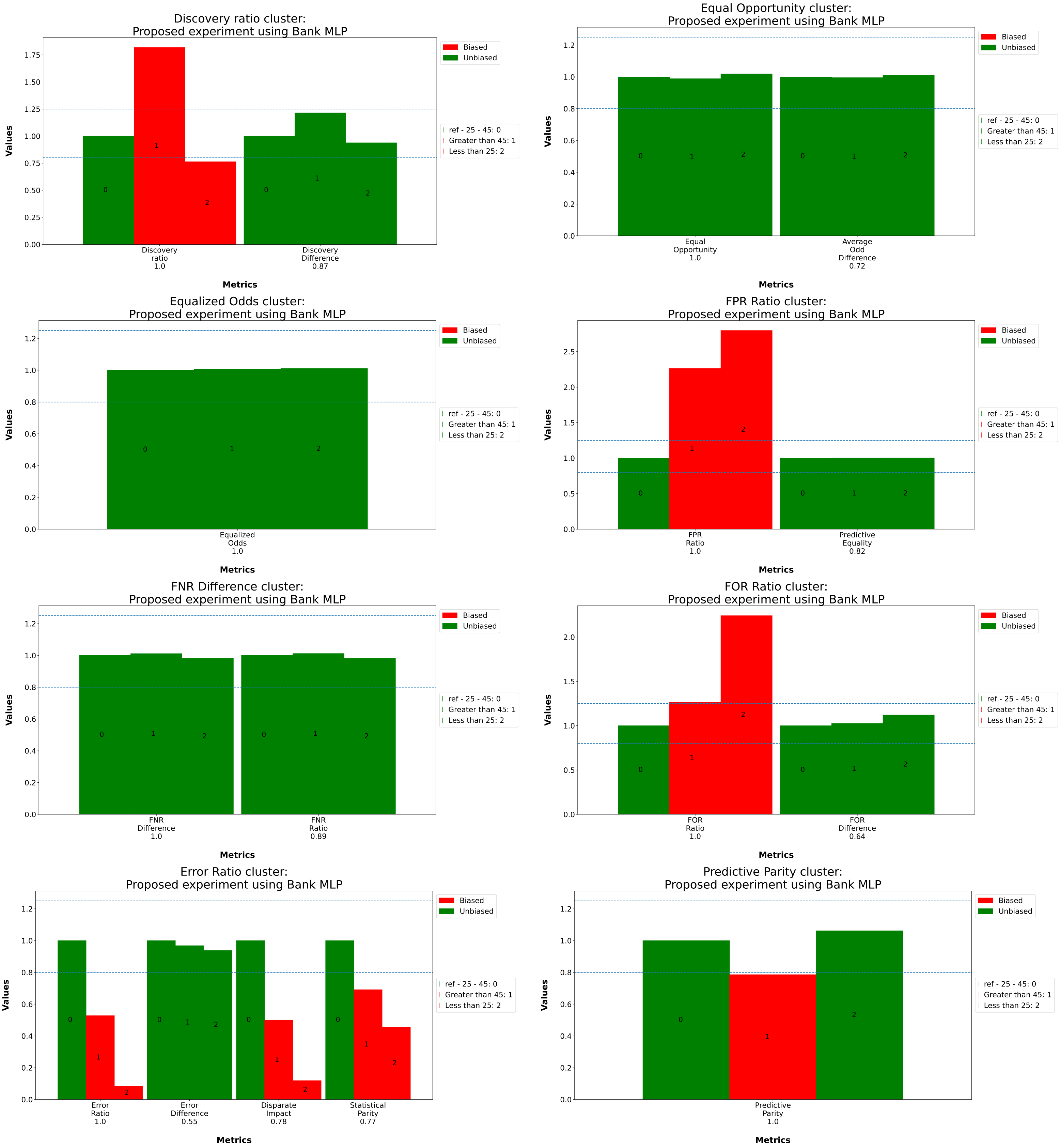


Figure S.7: Proposed Experiment clusters using Bank dataset and Logit model.


Discovery ratio cluster: Proposed experiment using Bank MLP

Equal Opportunity cluster: Proposed experiment using Bank MLP

Equalized Odds cluster: Proposed experiment using Bank MLP

FPR Ratio cluster: Proposed experiment using Bank MLP

FNR Difference cluster: Proposed experiment using Bank MLP

FOR Ratio cluster: Proposed experiment using Bank MLP

Error Ratio cluster: Proposed experiment using Bank MLP

Predictive Parity cluster: Proposed experiment using Bank MLP

Figure S.8: Proposed Experiment clusters using Bank dataset and MLP model.

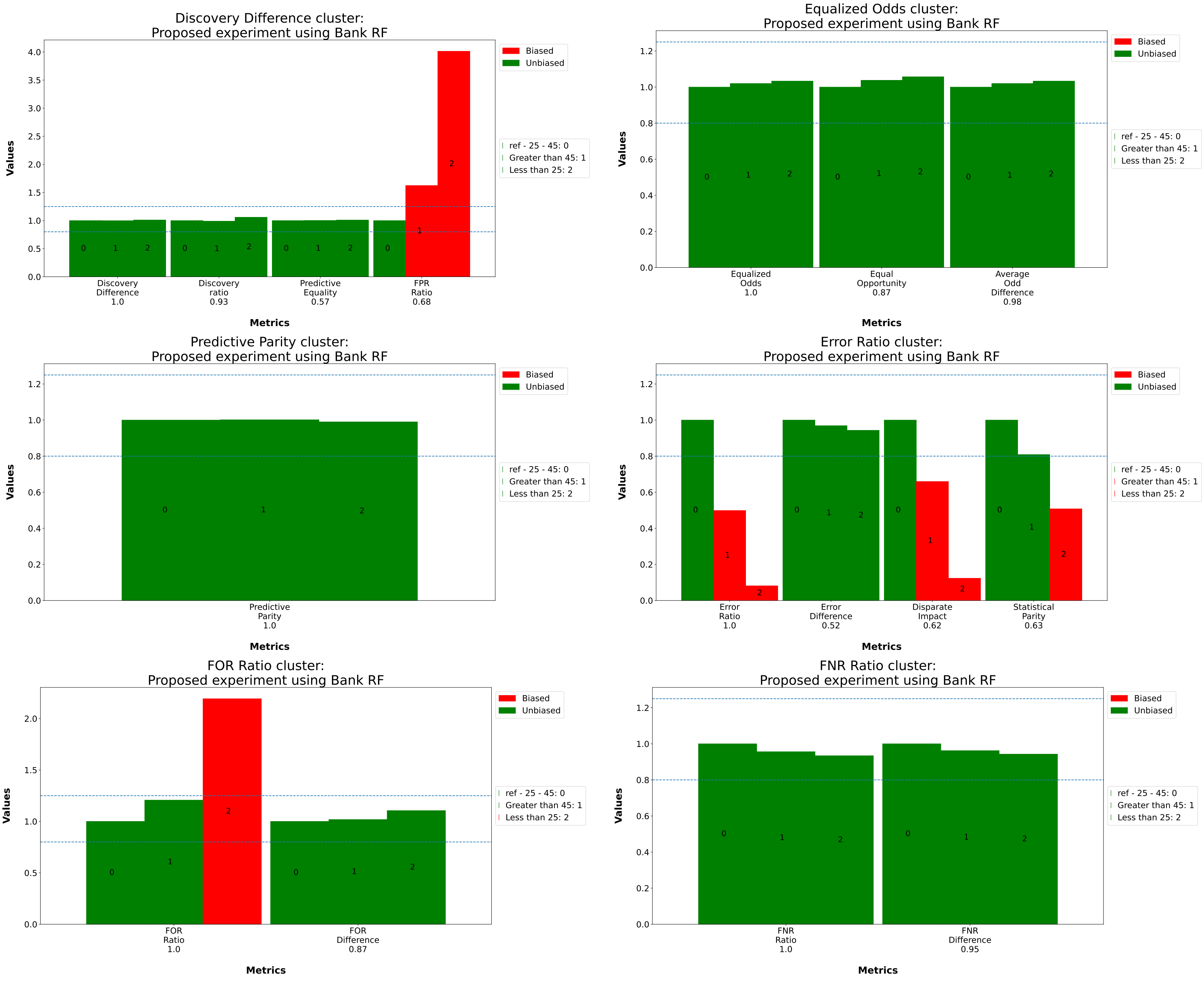


Figure S.9: Proposed Experiment clusters using Bank dataset and RF model.

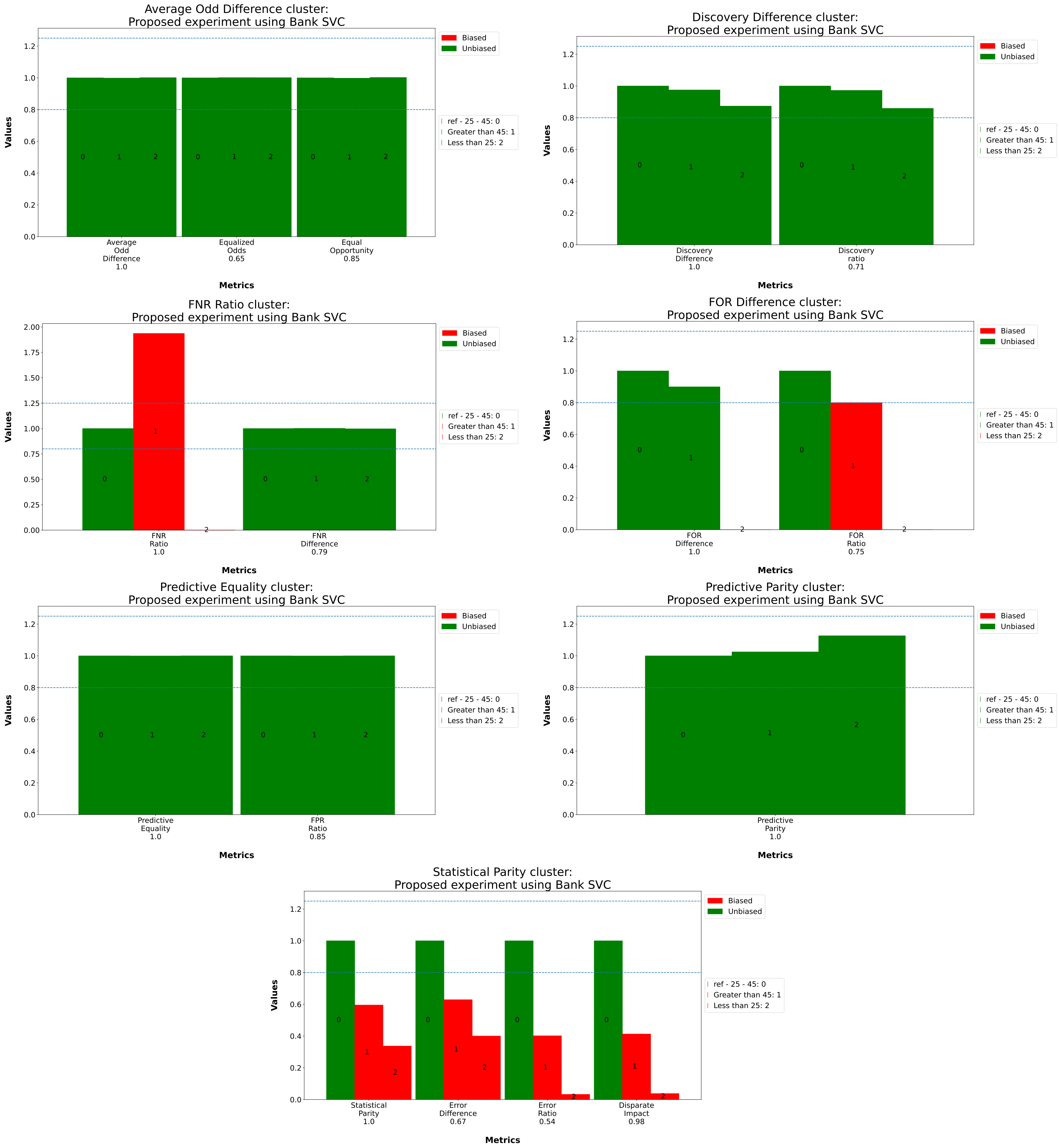


Figure S.10: Proposed Experiment clusters using Bank dataset and SVC model.

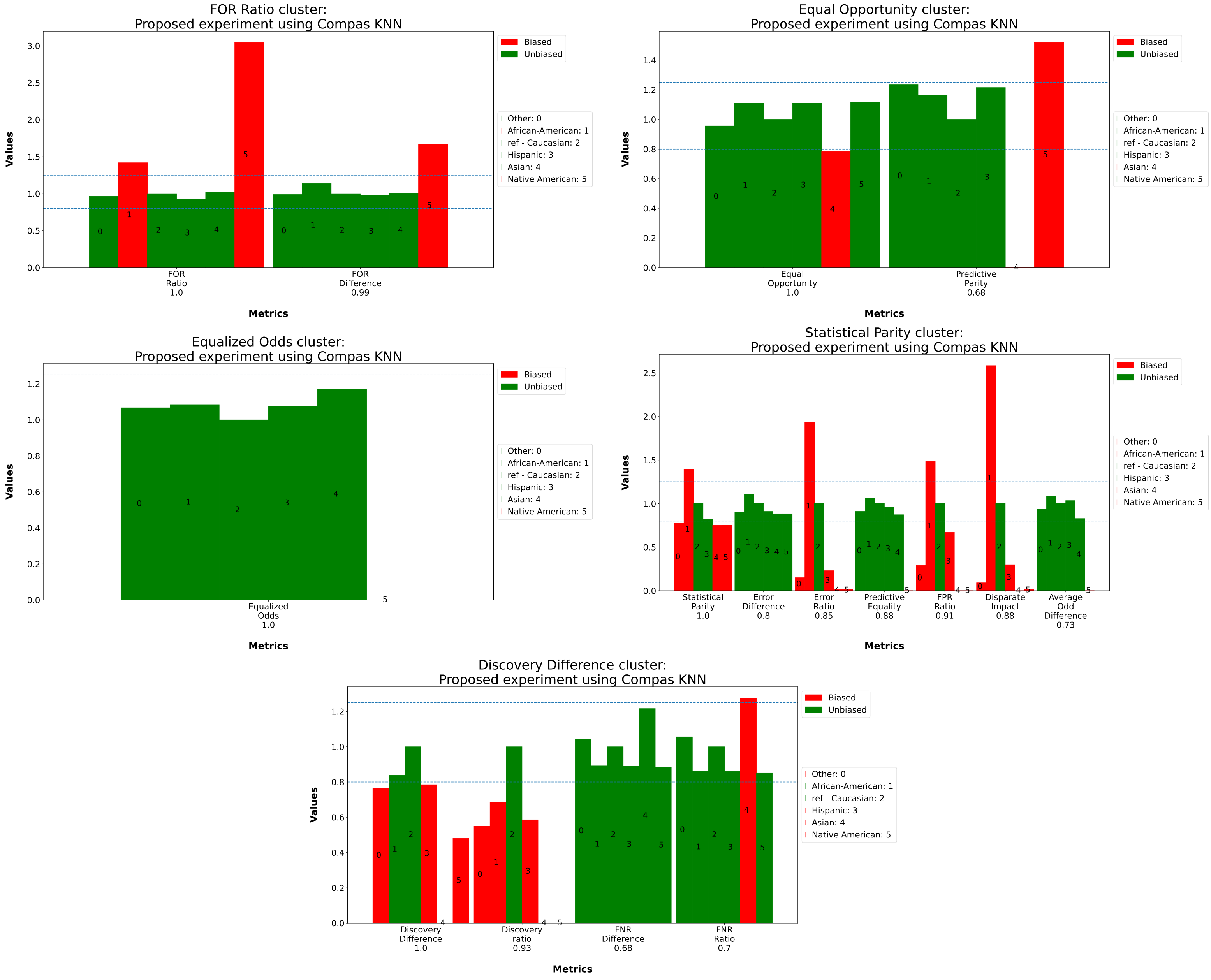


Figure S.11: Proposed Experiment clusters using Compas dataset and KNN model.

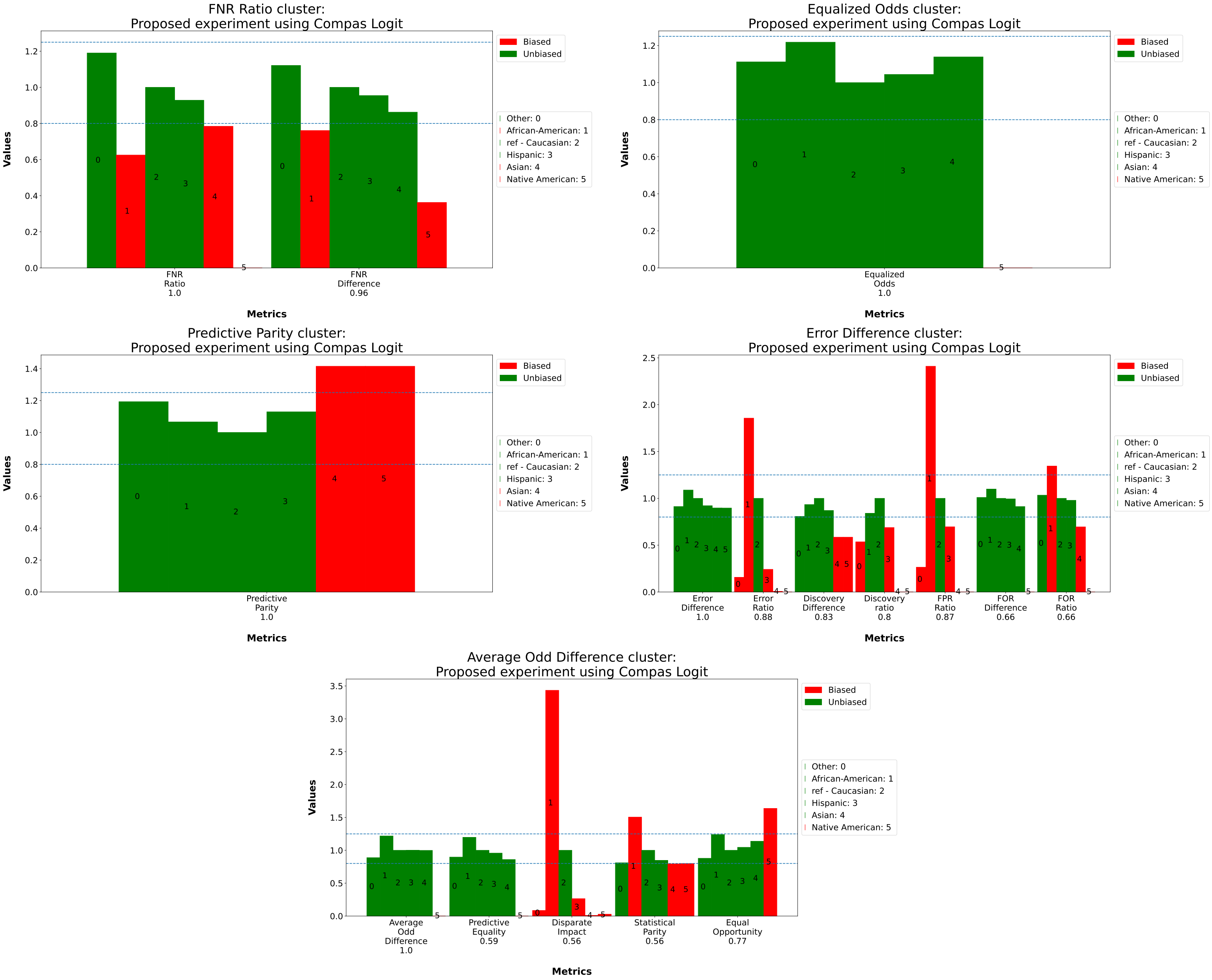


Figure S.12: Proposed Experiment clusters using Compas dataset and Logit model.

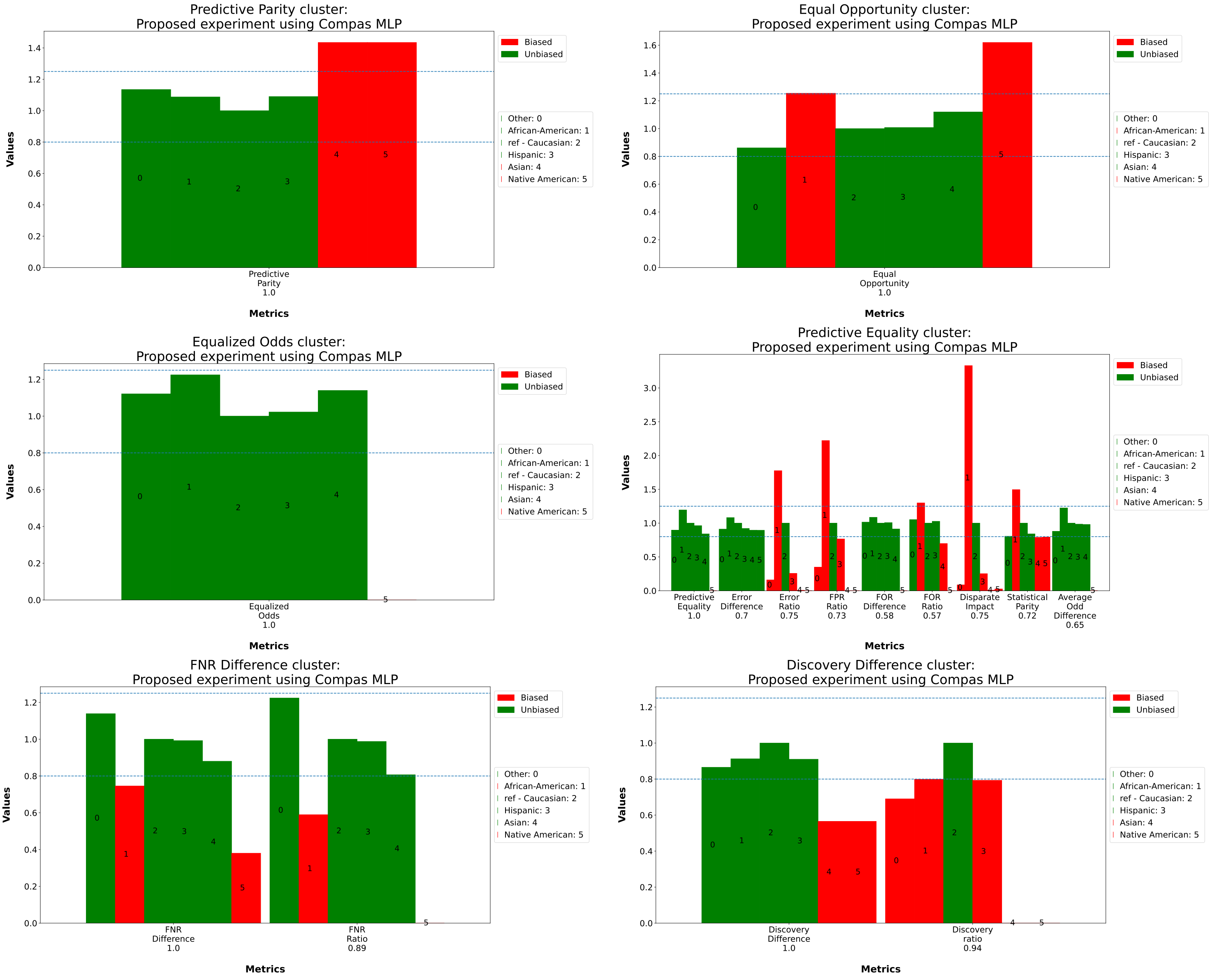


Figure S.13: Proposed Experiment clusters using Compas dataset and MLP model.

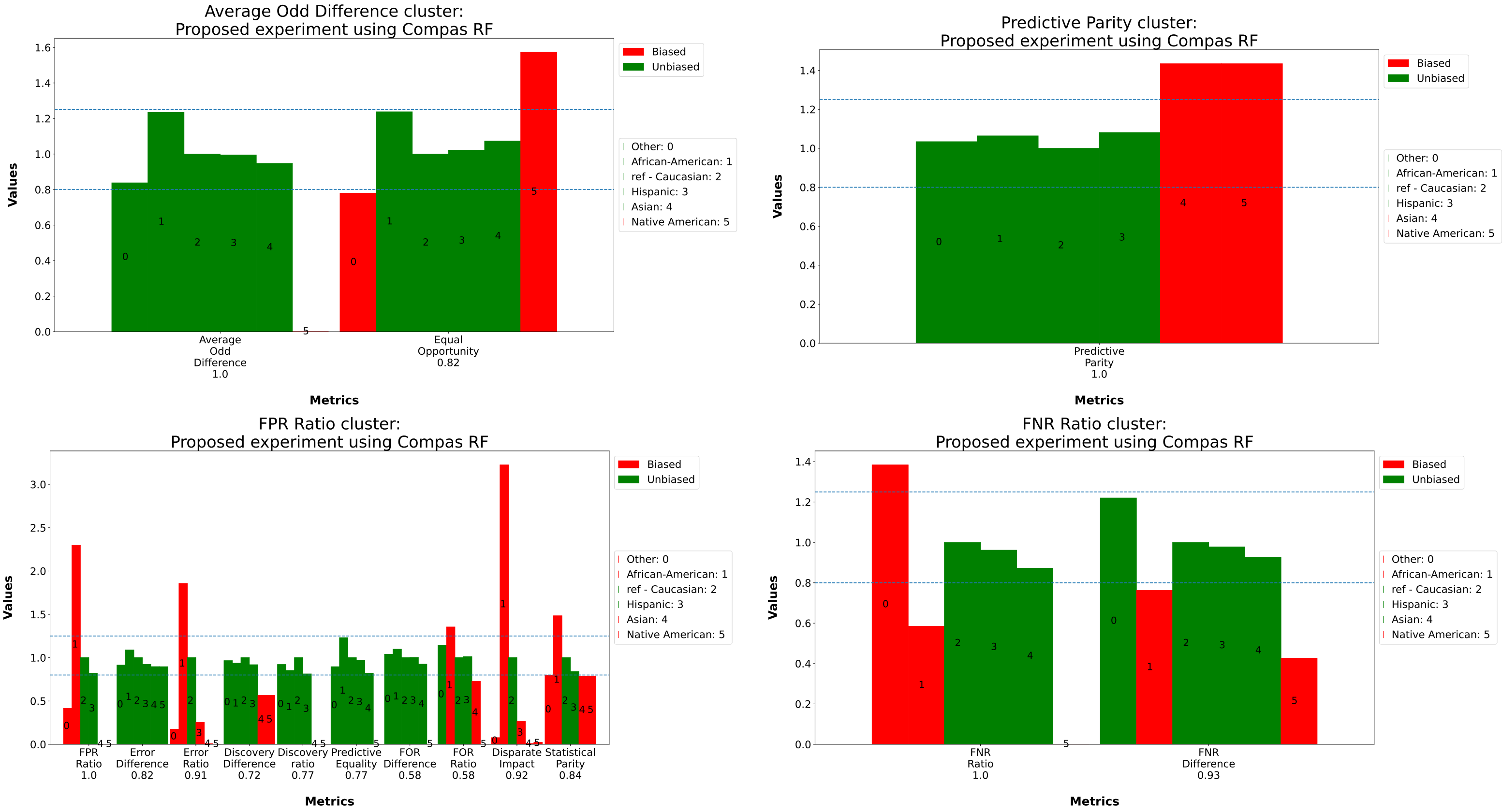


Figure S.14: Proposed Experiment clusters using Compas dataset and RF model.

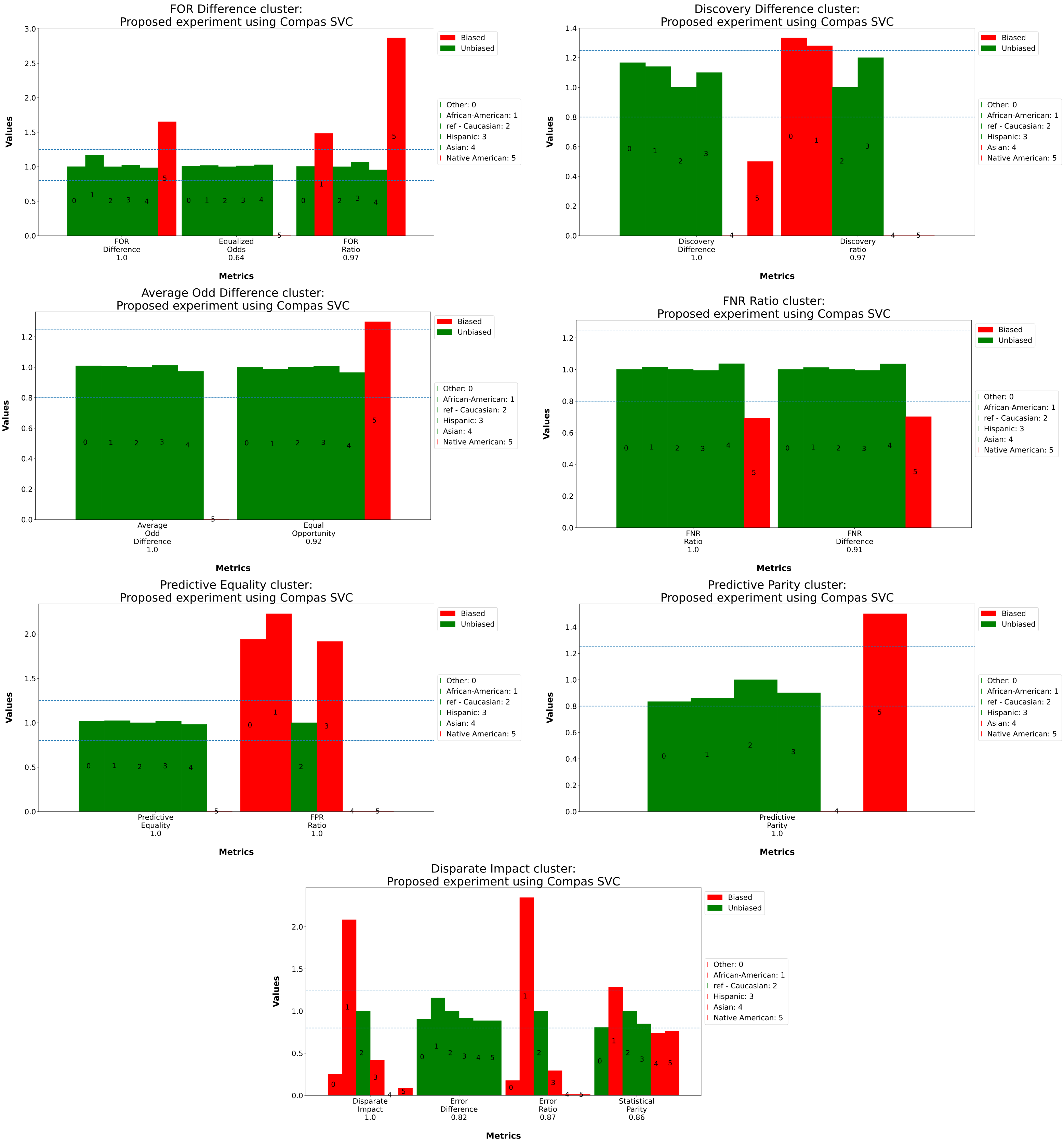


Figure S.15: Proposed Experiment clusters using Compas dataset and SVC model.

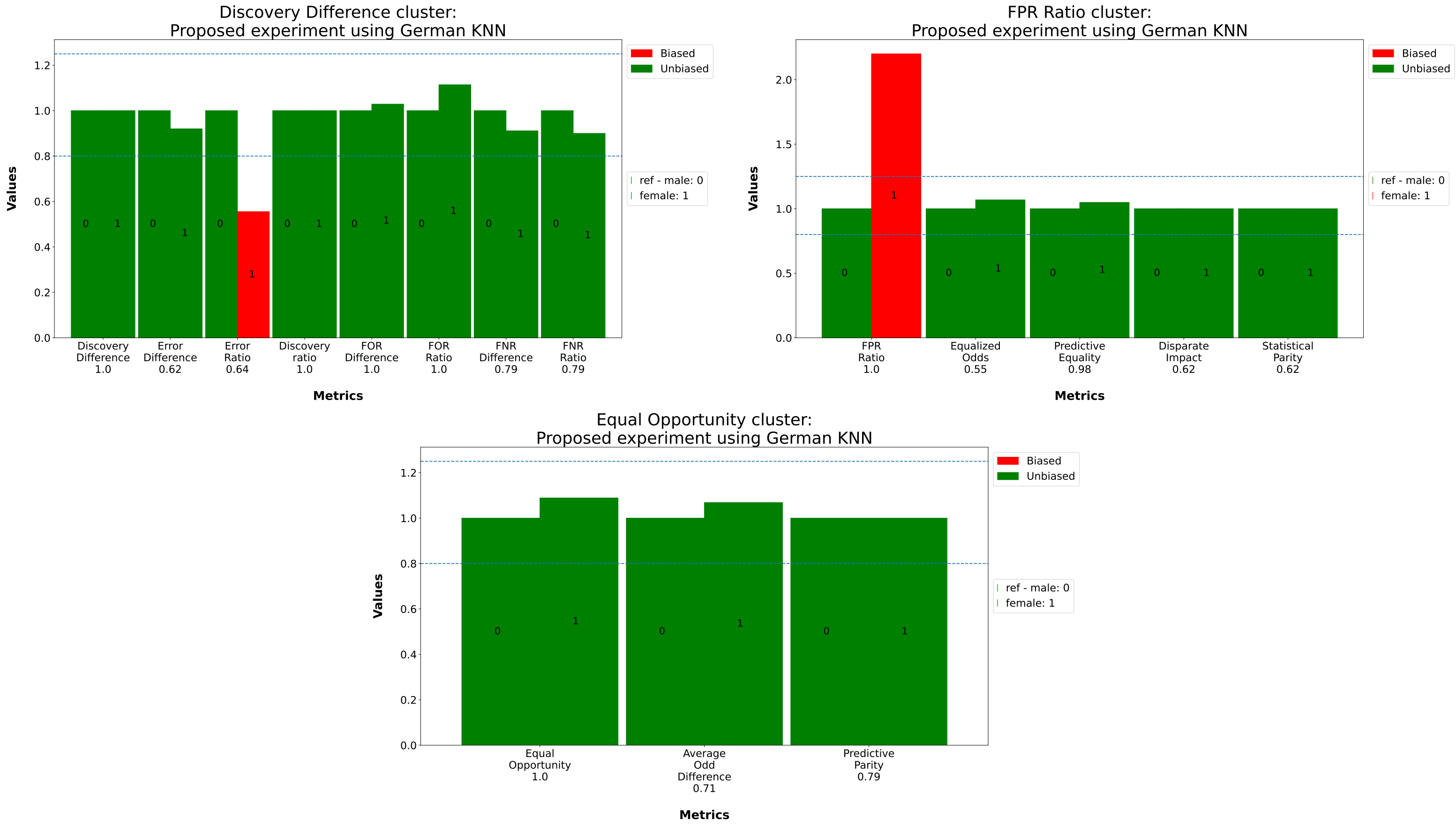


Figure S.16: Proposed Experiment clusters using German dataset and KNN model.

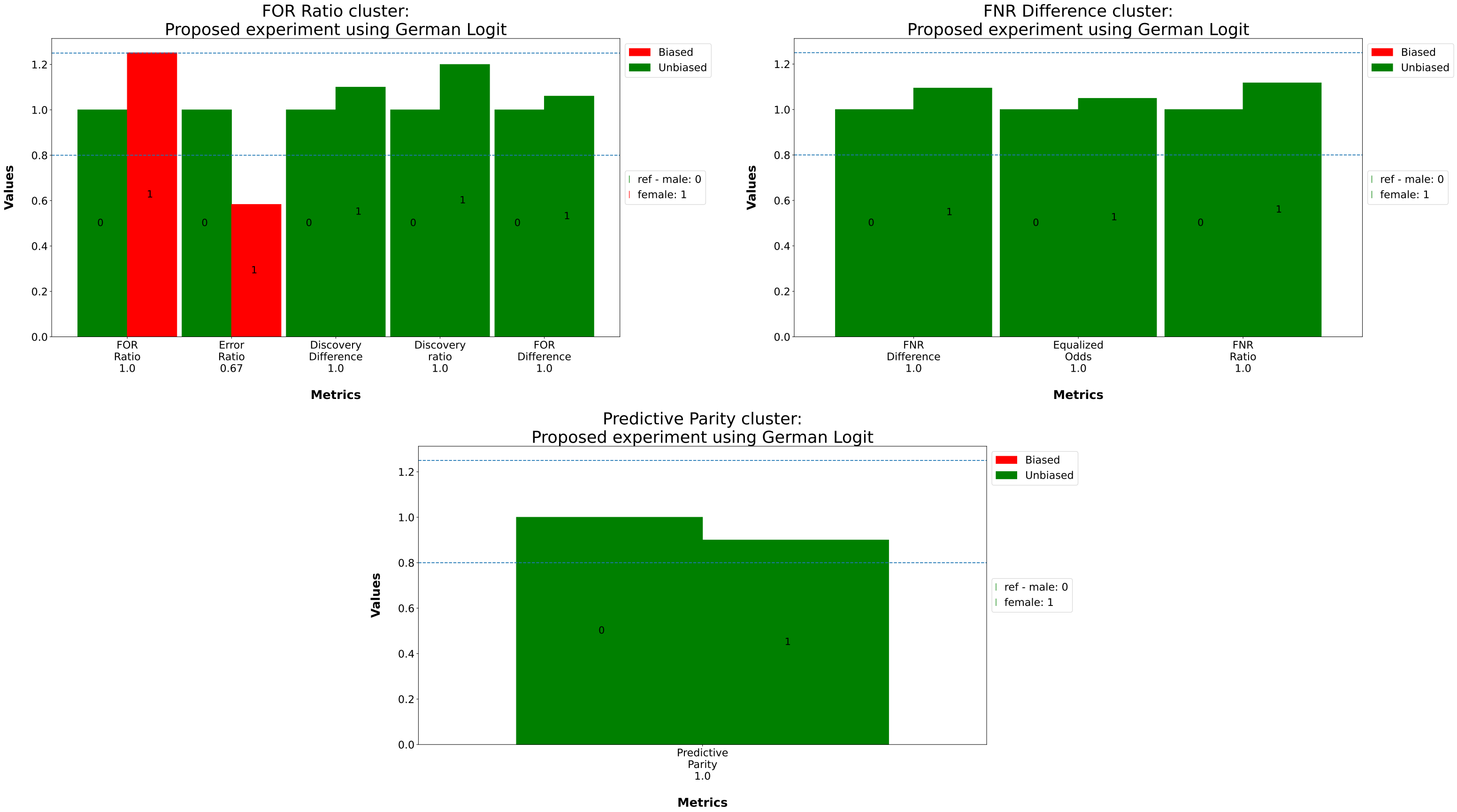


Figure S.17: Proposed Experiment clusters using German dataset and Logit model.

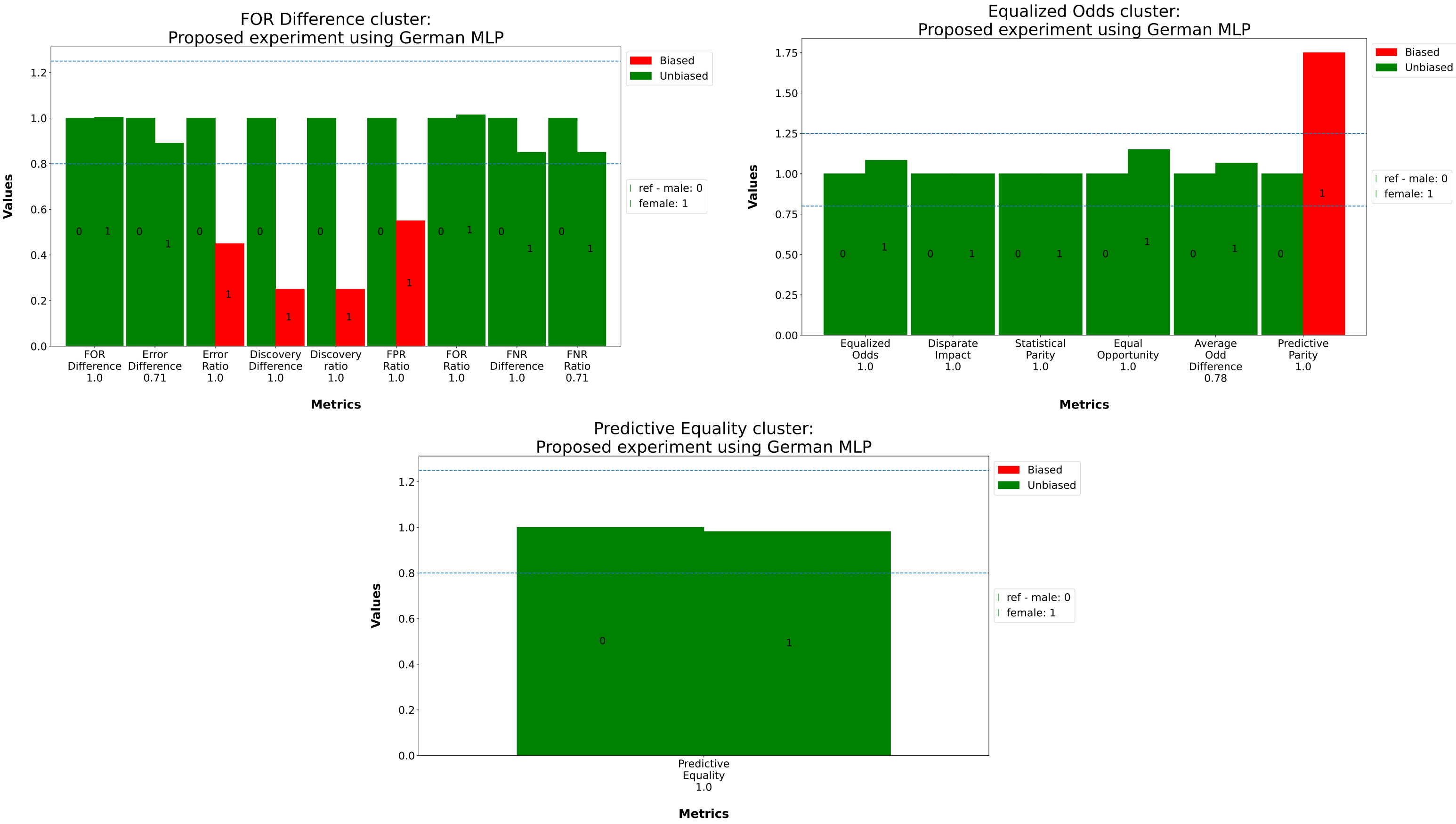


Figure S.18: Proposed Experiment clusters using German dataset and MLP model.

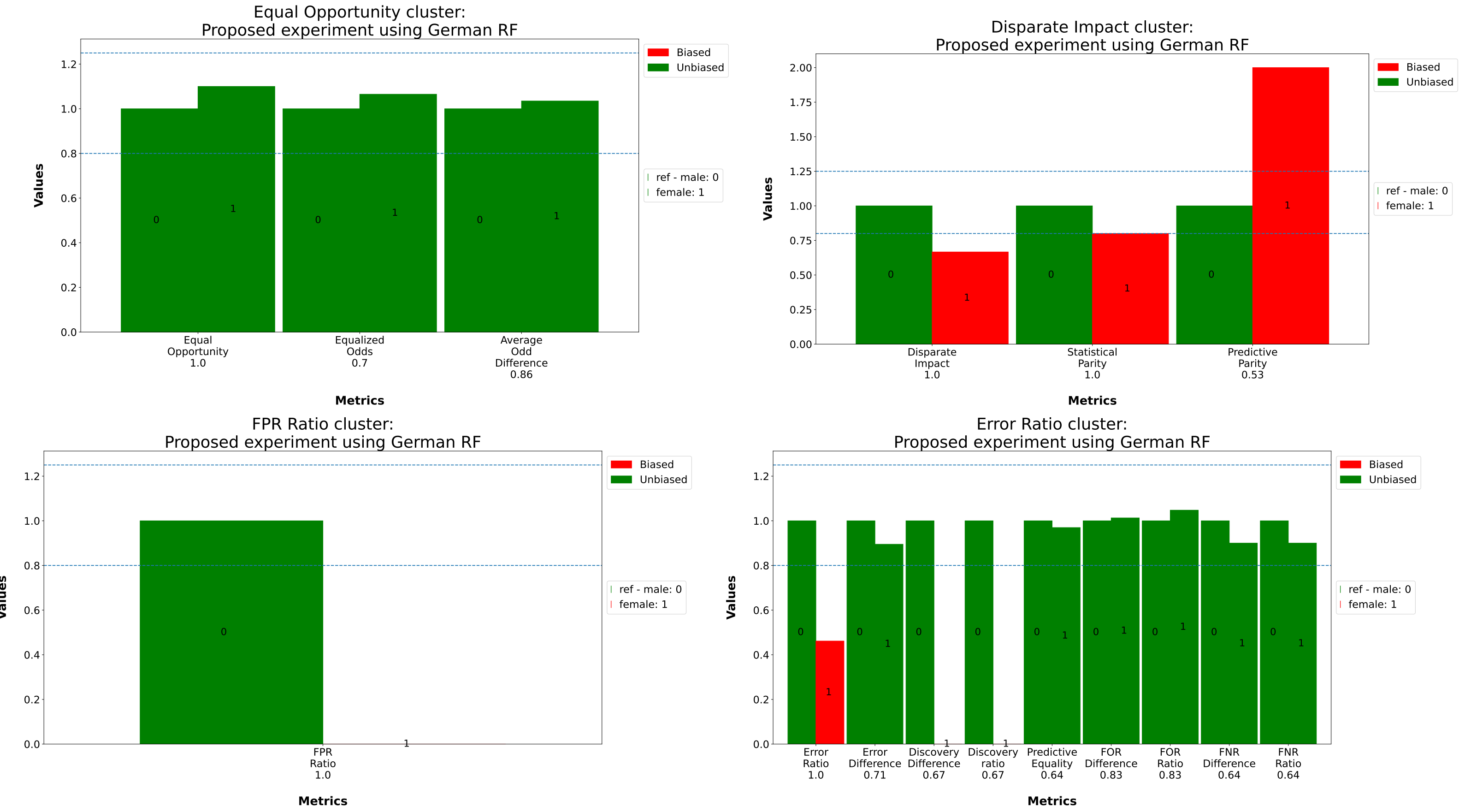


Figure S.19: Proposed Experiment clusters using German dataset and RF model.

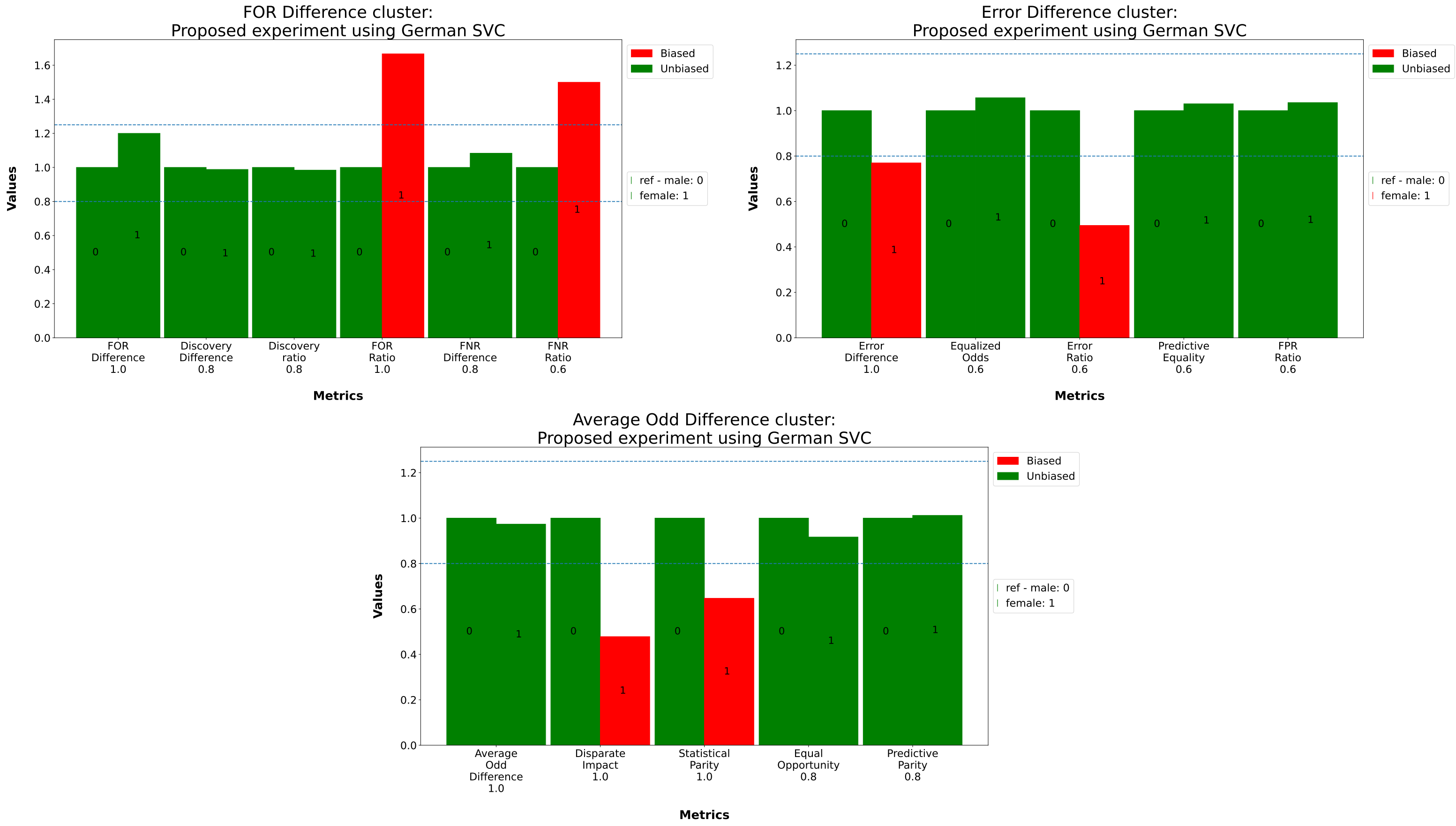


Figure S.20: Proposed Experiment clusters using German dataset and SVC model.

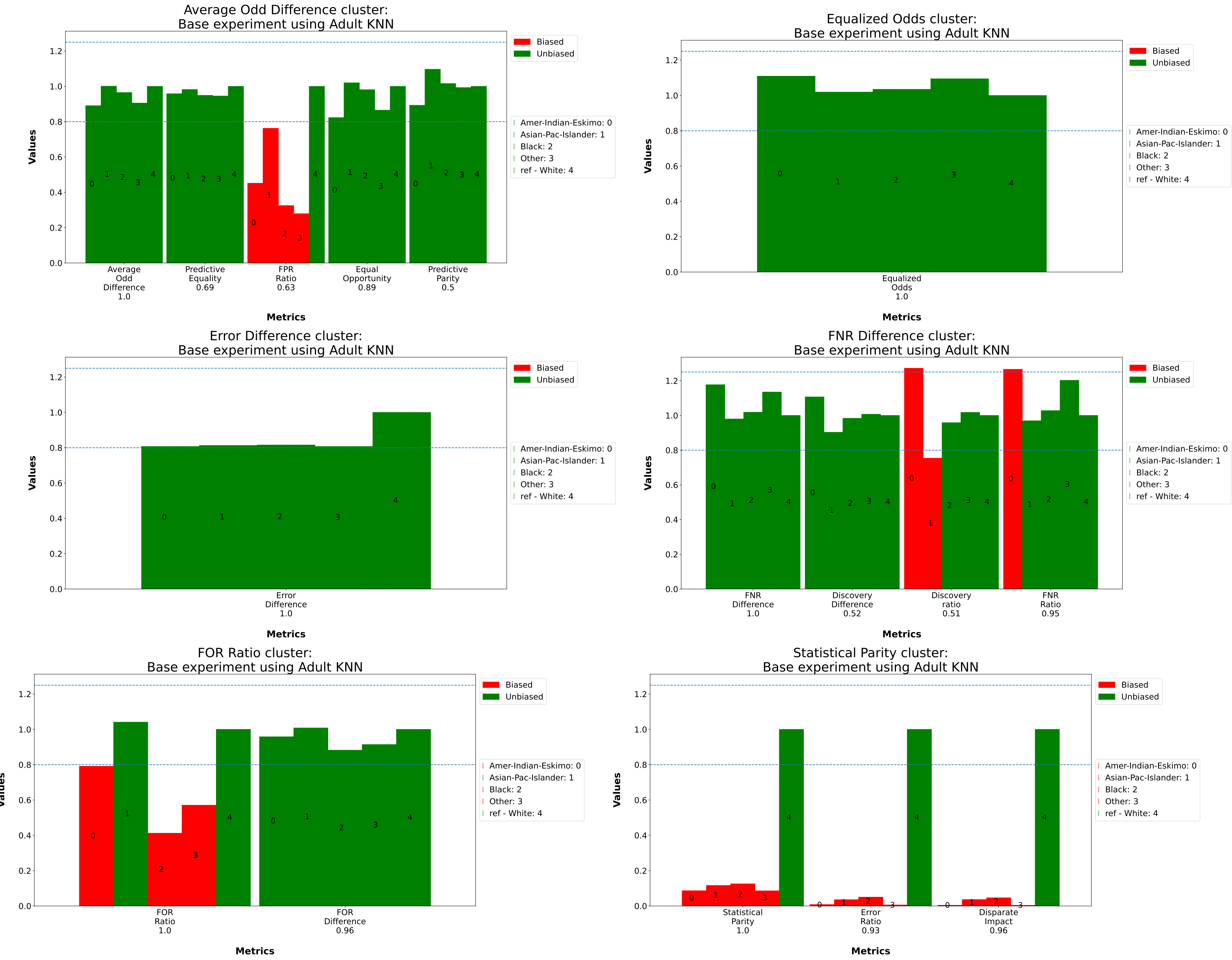


Figure S.21: Base Experiment clusters using Adult dataset and KNN model.

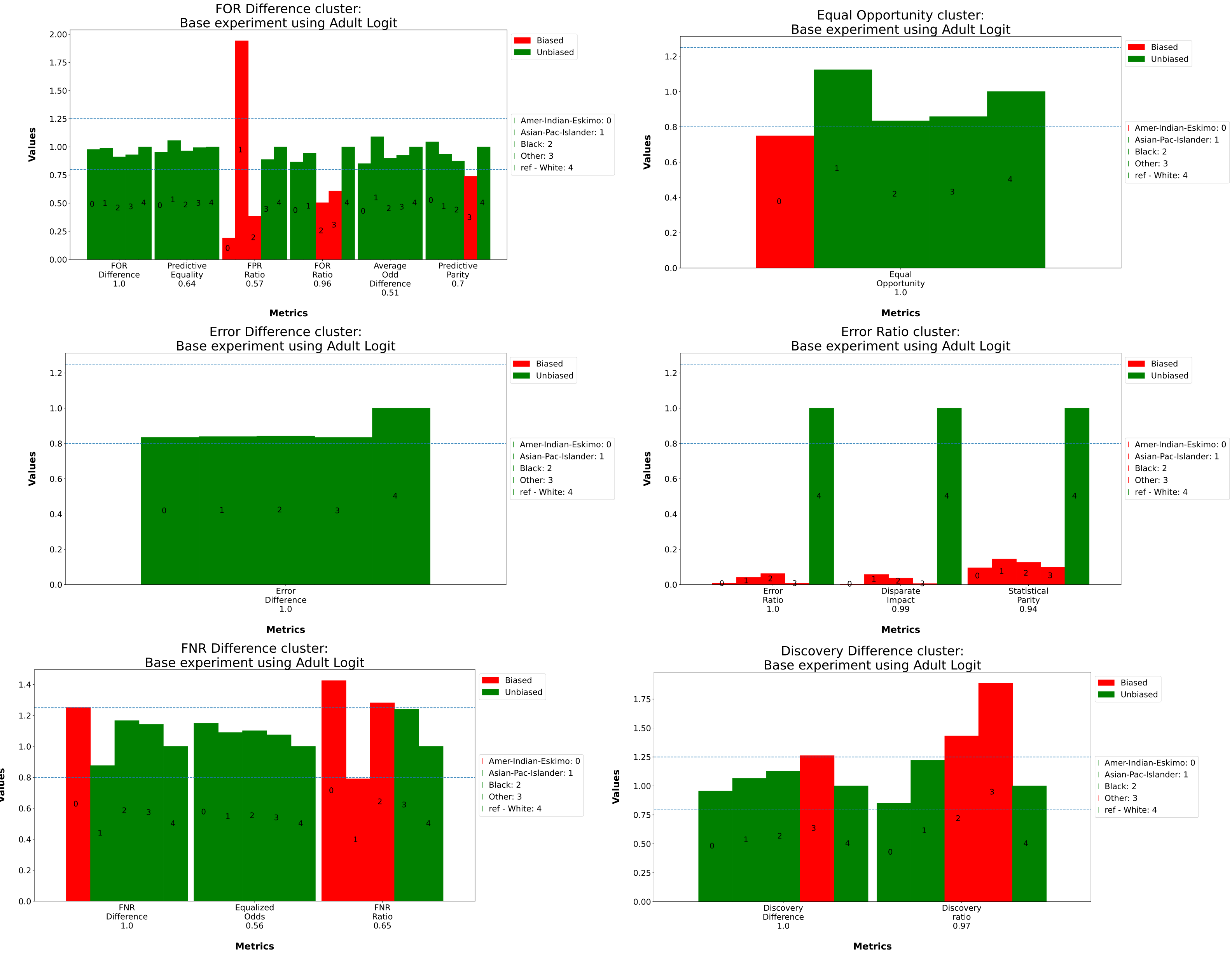


Figure S.22: Base Experiment clusters using Adult dataset and Logit model.

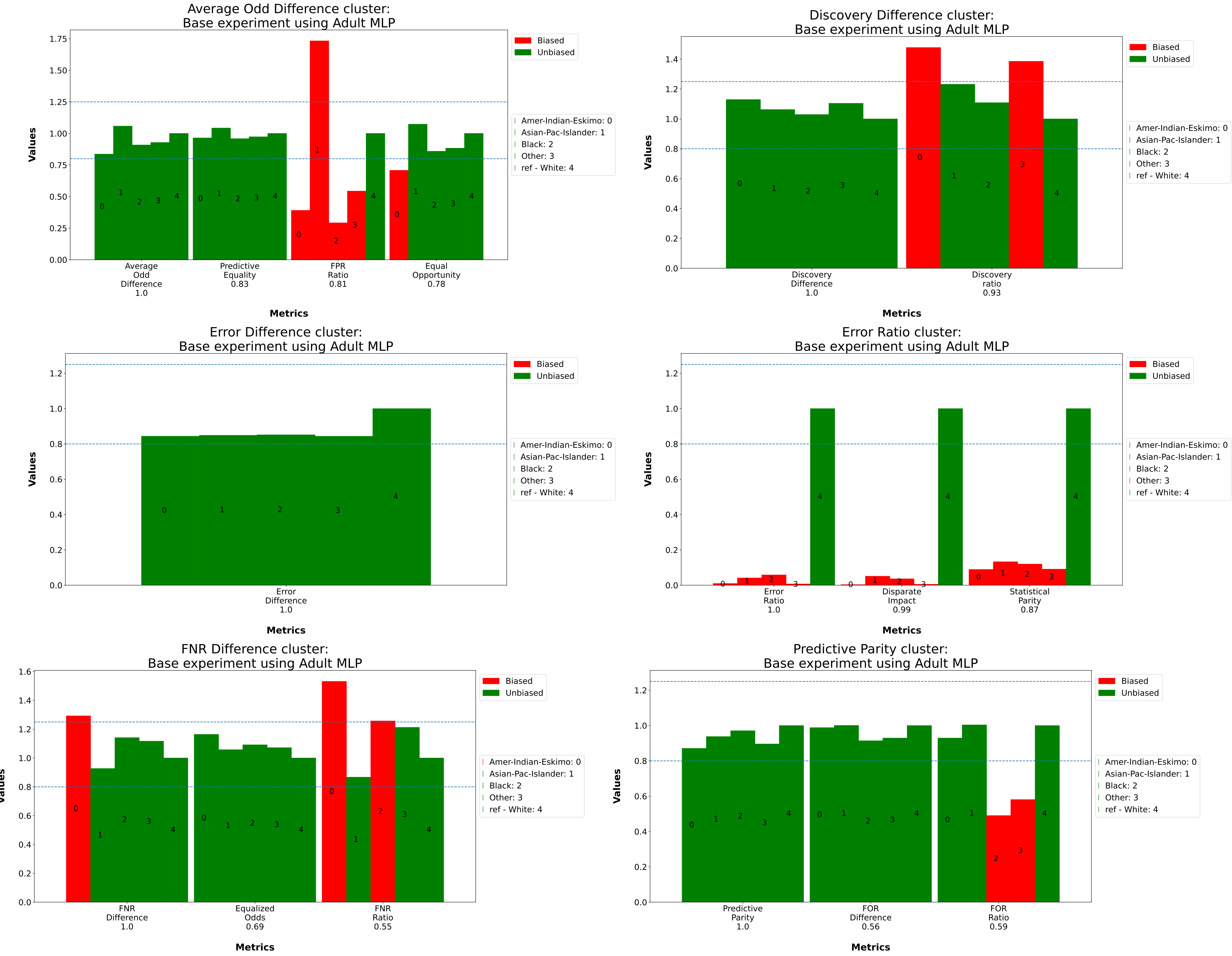


Figure S.23: Base Experiment clusters using Adult dataset and MLP model.

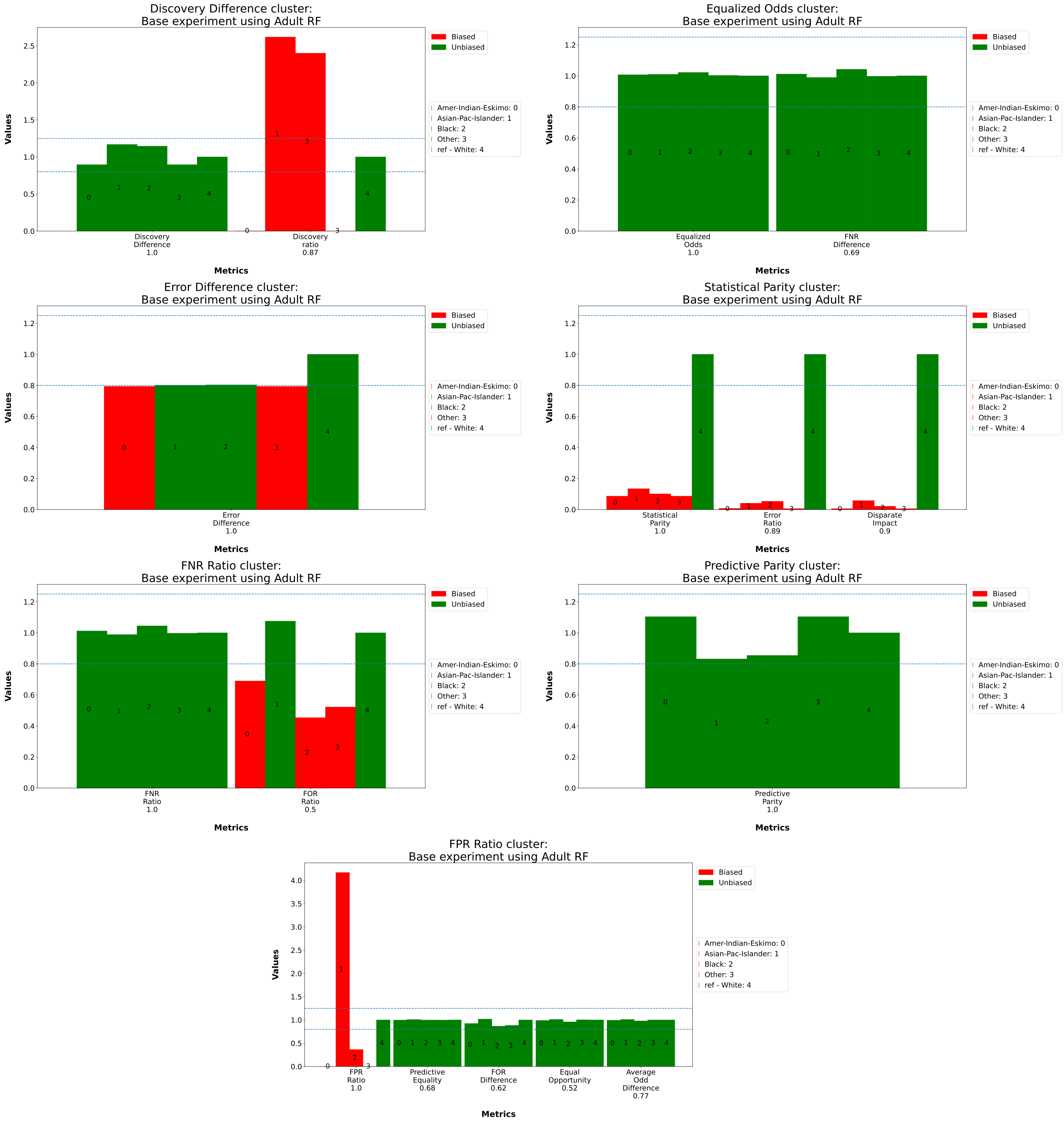


Figure S.24: Base Experiment clusters using Adult dataset and RF model.

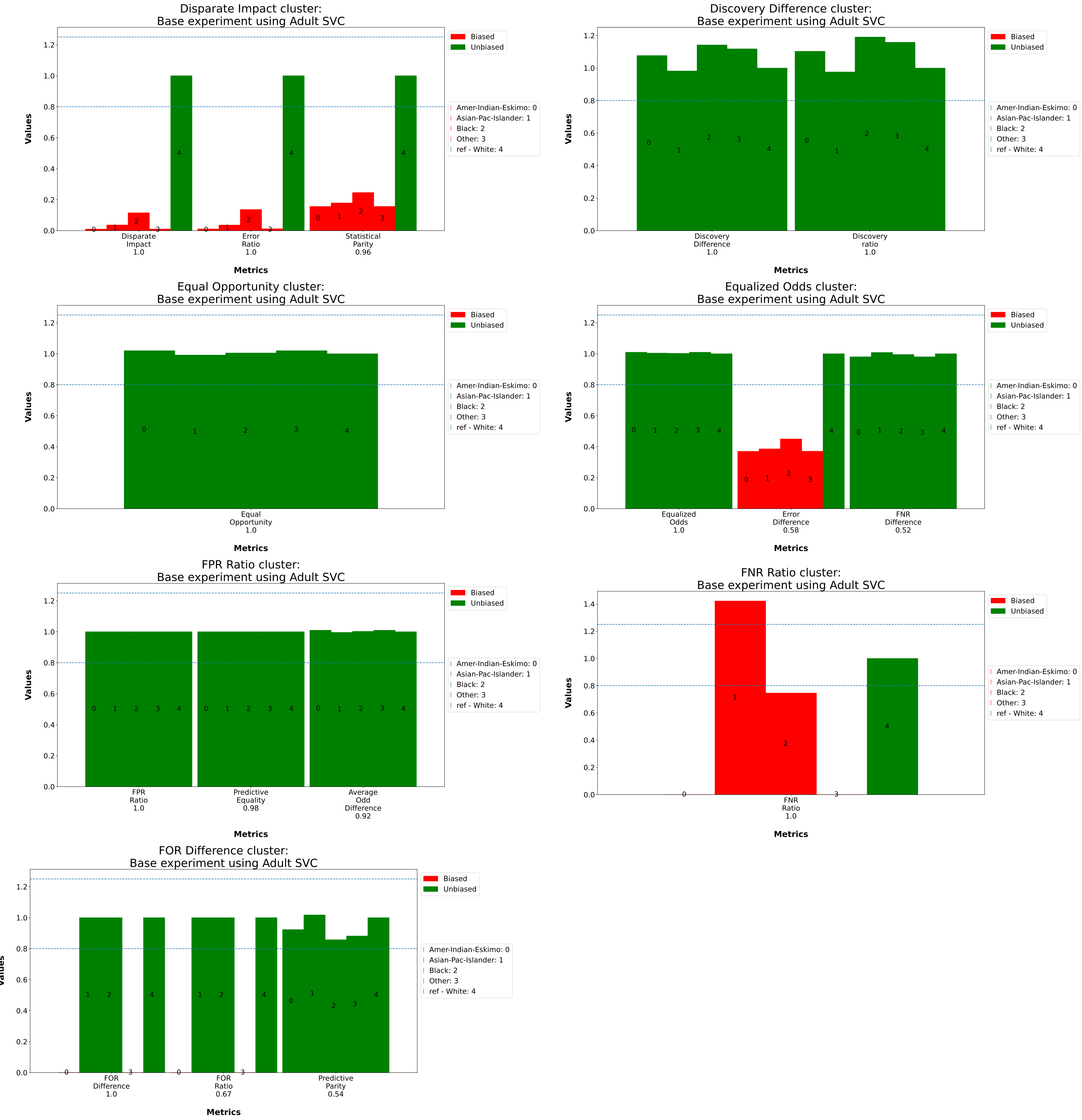


Figure S.25: Base Experiment clusters using Adult dataset and SVC model.

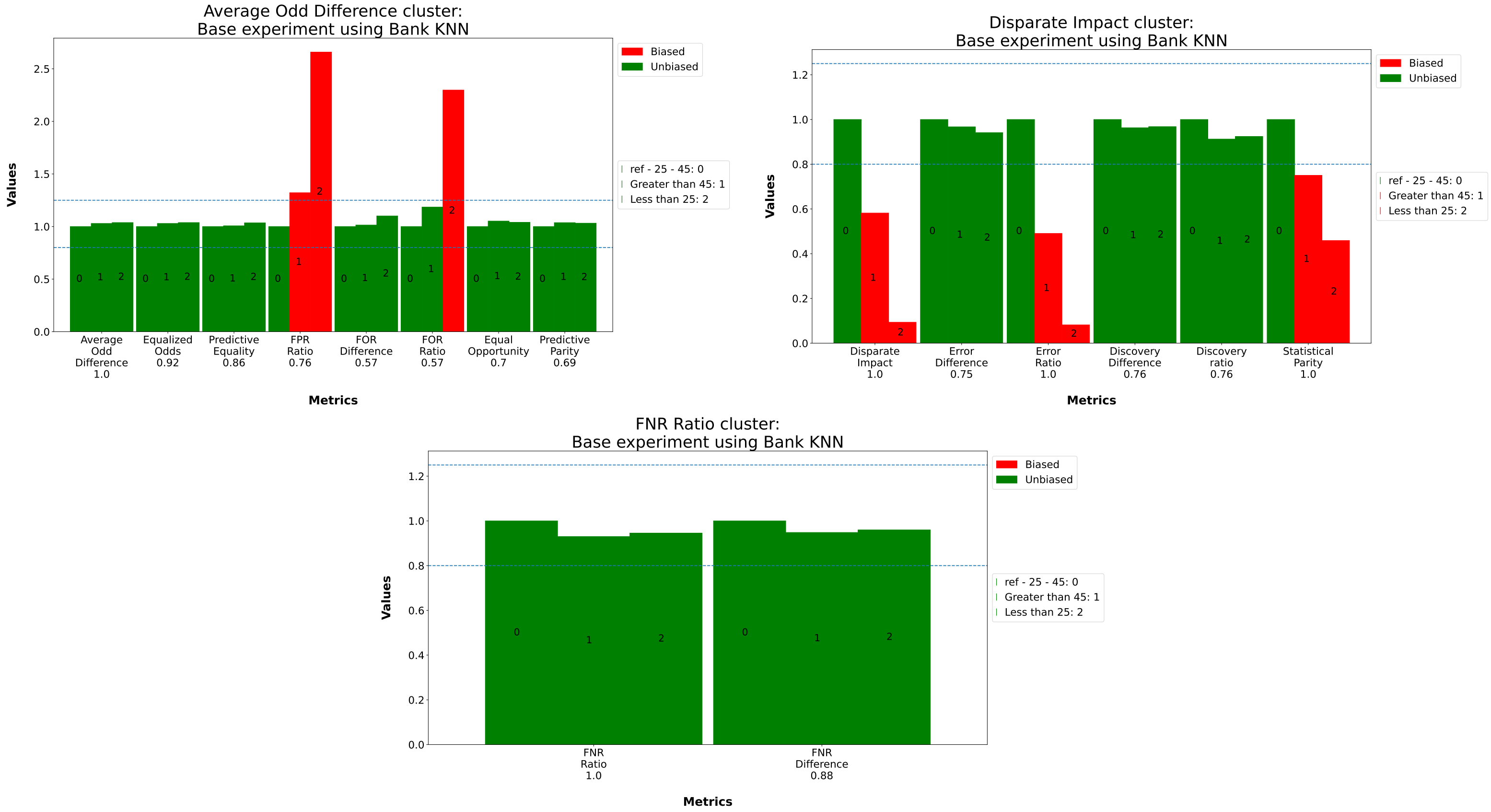


Figure S.26: Base Experiment clusters using Bank dataset and KNN model.

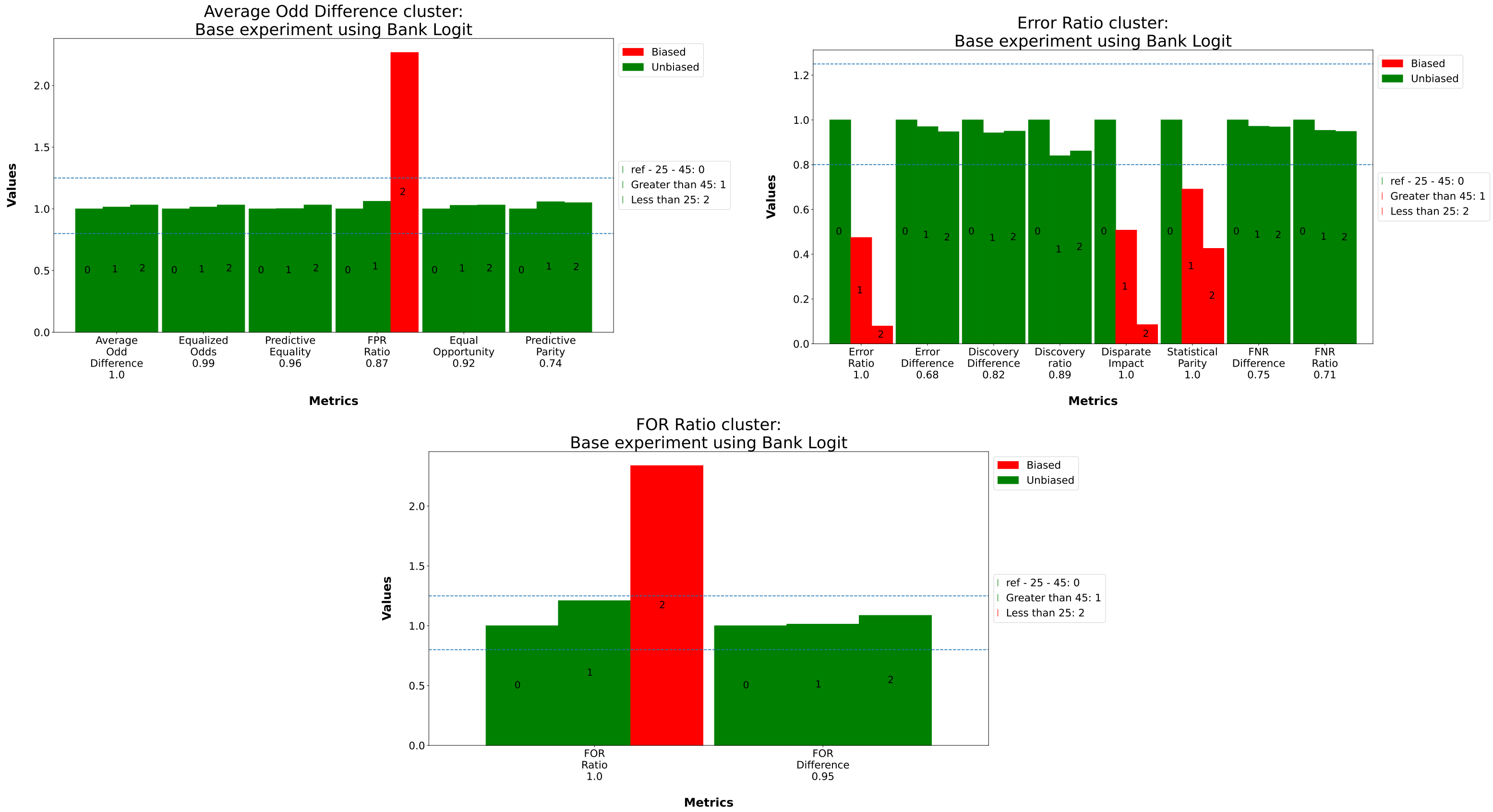


Figure S.27: Base Experiment clusters using Bank dataset and Logit model.

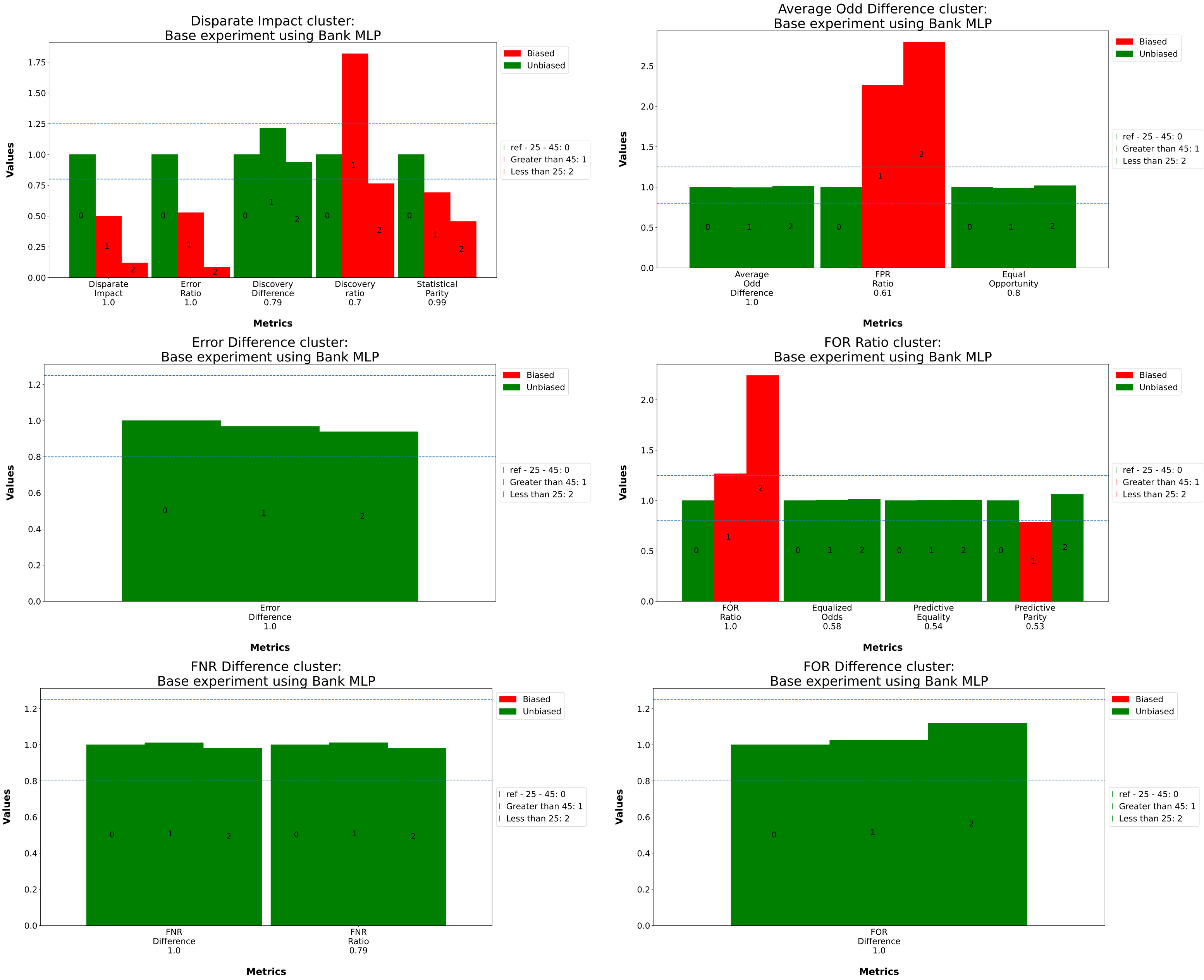


Figure S.28: Base Experiment clusters using Bank dataset and MLP model.

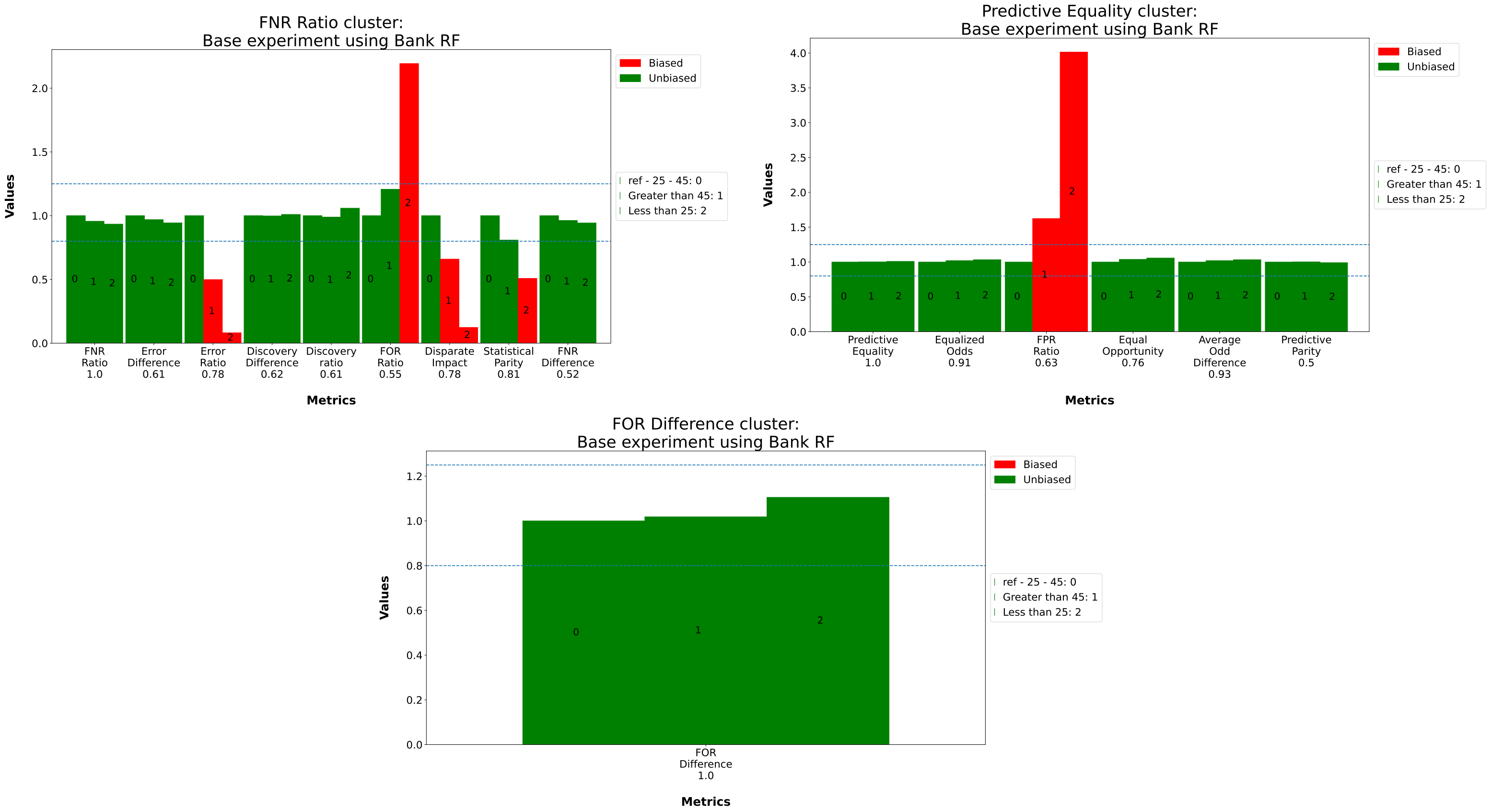


Figure S.29: Base Experiment clusters using Bank dataset and RF model.

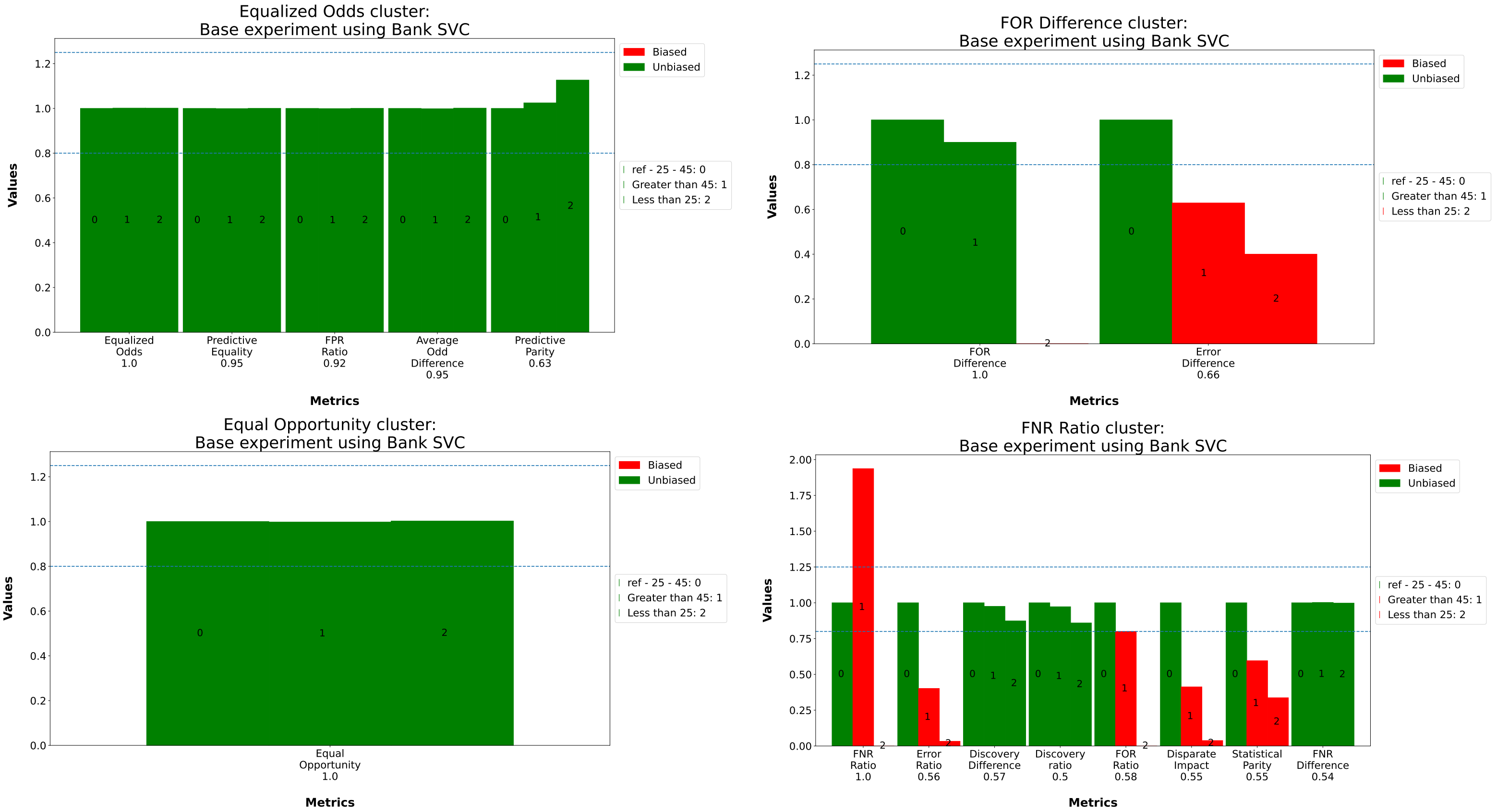
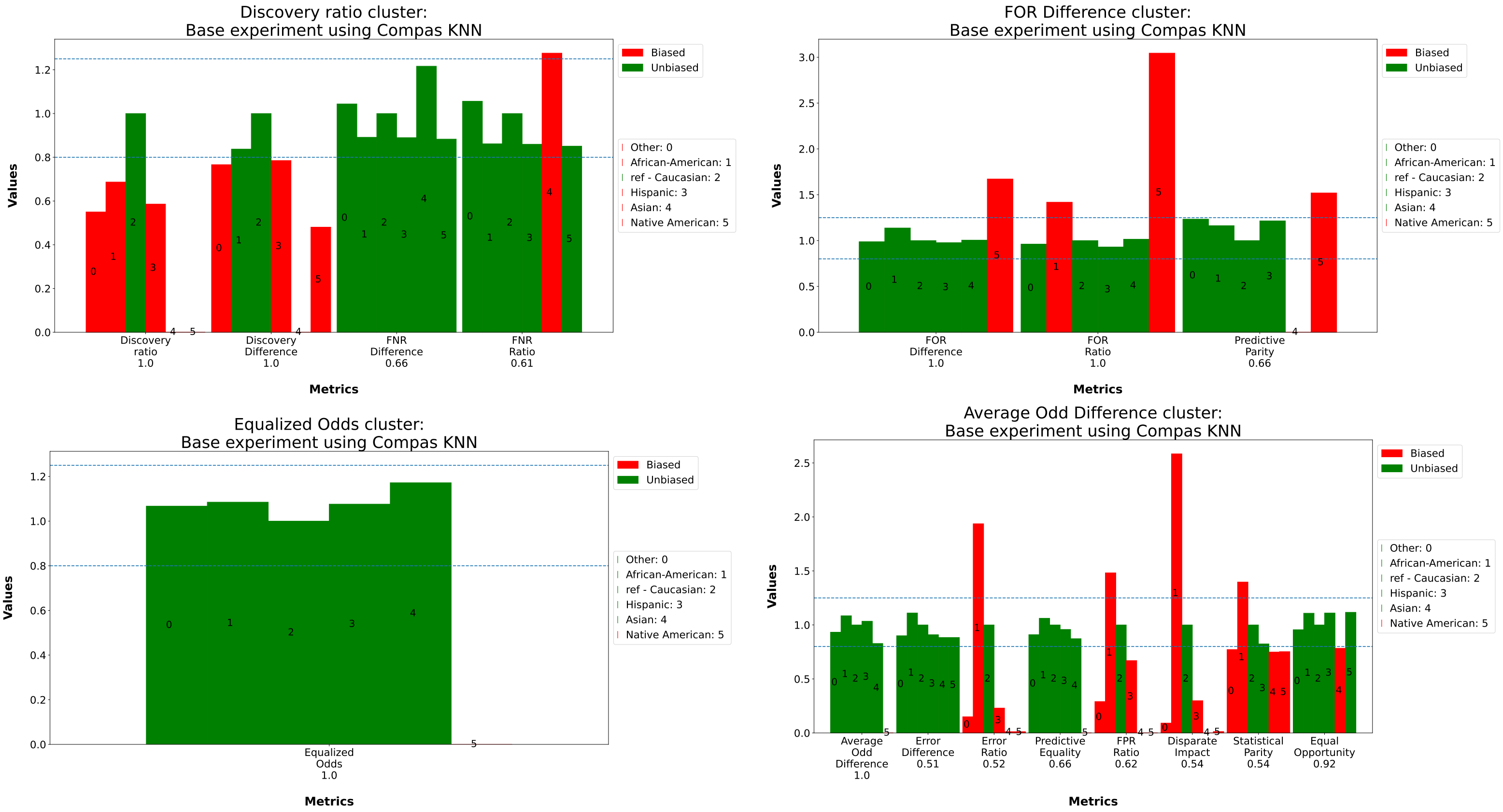


Figure S.30: Base Experiment clusters using Bank dataset and SVC model.



Discovery ratio cluster:
Base experiment using Compas KNN

FOR Difference cluster:
Base experiment using Compas KNN

Equalized Odds cluster:
Base experiment using Compas KNN

Average Odd Difference cluster:
Base experiment using Compas KNN

Figure S.31: Base Experiment clusters using Compass dataset and KNN model.

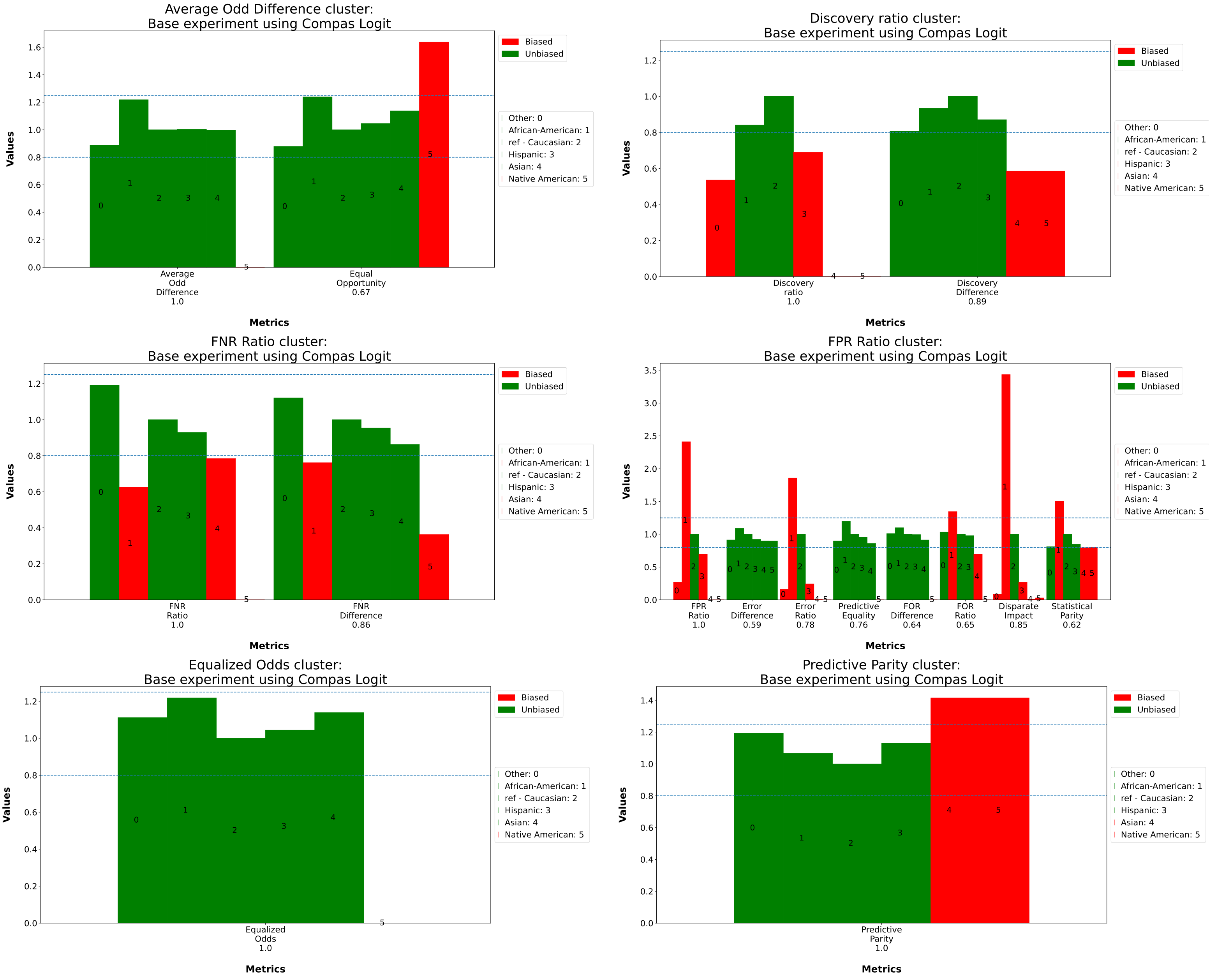
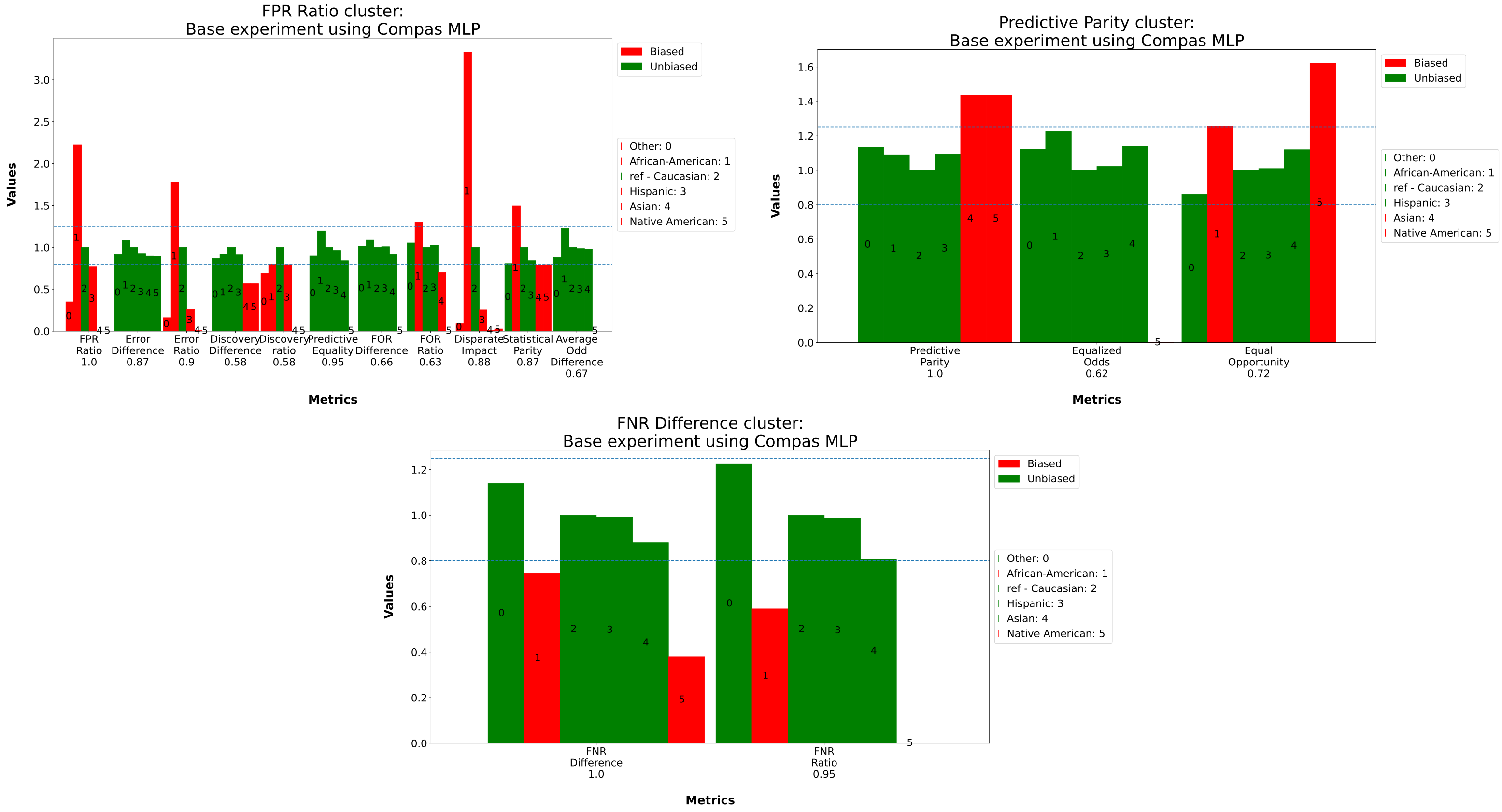


Figure S.32: Base Experiment clusters using Compass dataset and Logit model.

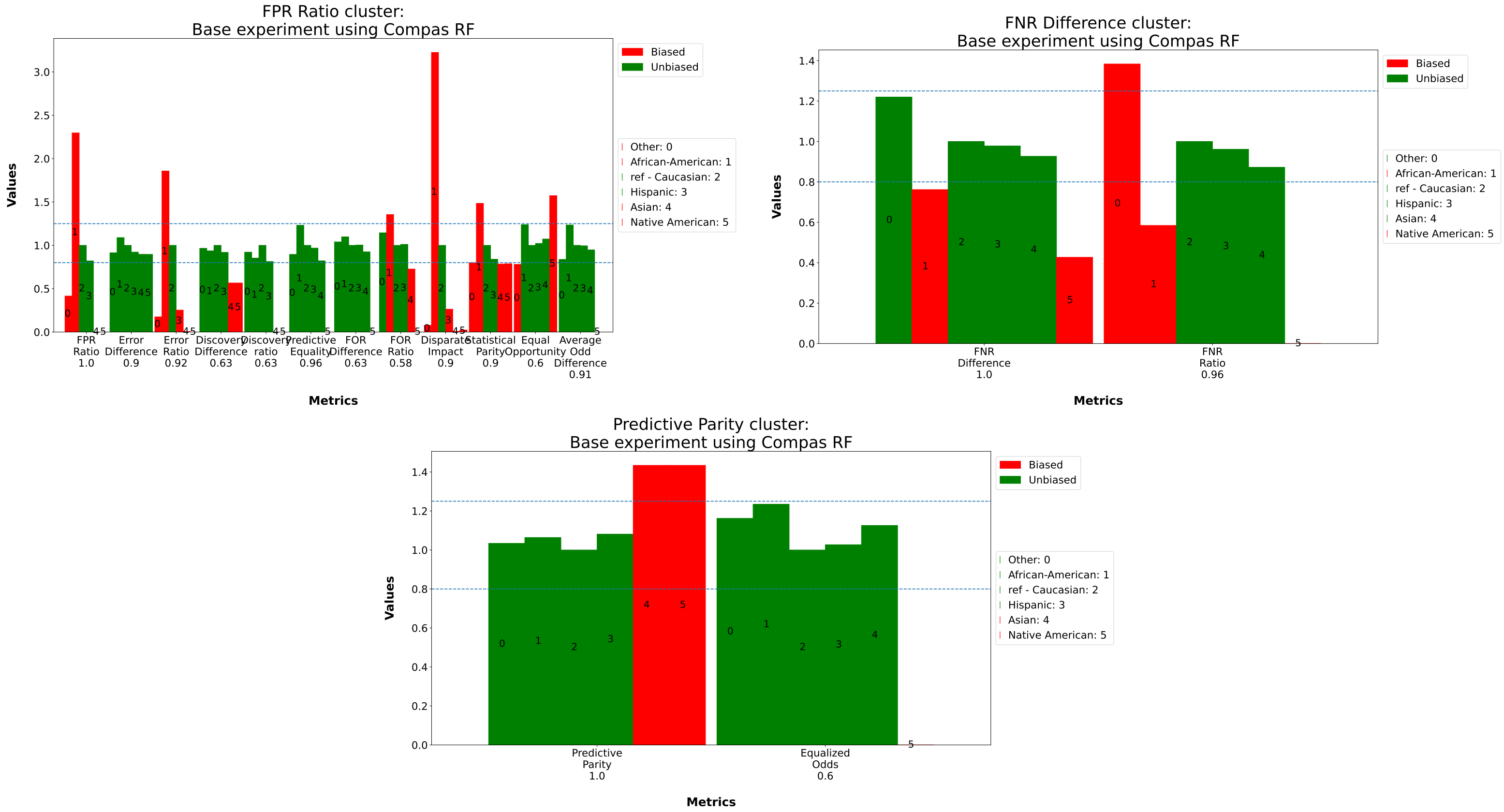


FPR Ratio cluster:
Base experiment using Compas MLP

Predictive Parity cluster:
Base experiment using Compas MLP

FNR Difference cluster:
Base experiment using Compas MLP

Figure S.33: Base Experiment clusters using Compass dataset and MLP model.

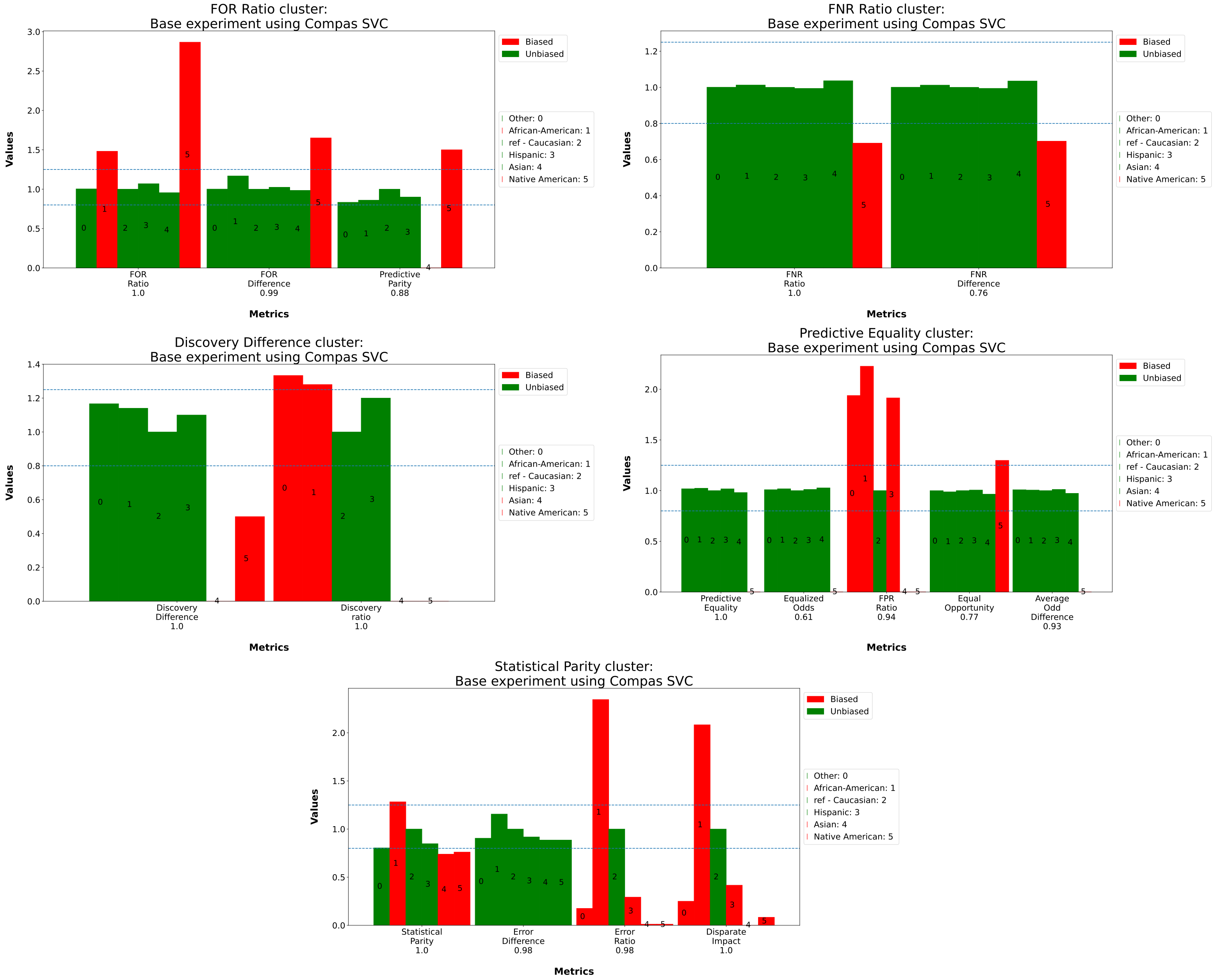


FPR Ratio cluster:
Base experiment using Compas RF

FNR Difference cluster:
Base experiment using Compas RF

Predictive Parity cluster:
Base experiment using Compas RF

Figure S.34: Base Experiment clusters using Compass dataset and RF model.



FOR Ratio cluster:
Base experiment using Compas SVC

FNR Ratio cluster:
Base experiment using Compas SVC

Discovery Difference cluster:
Base experiment using Compas SVC

Predictive Equality cluster:
Base experiment using Compas SVC

Statistical Parity cluster:
Base experiment using Compas SVC

Figure S.35: Base Experiment clusters using Compass dataset and SVC model.

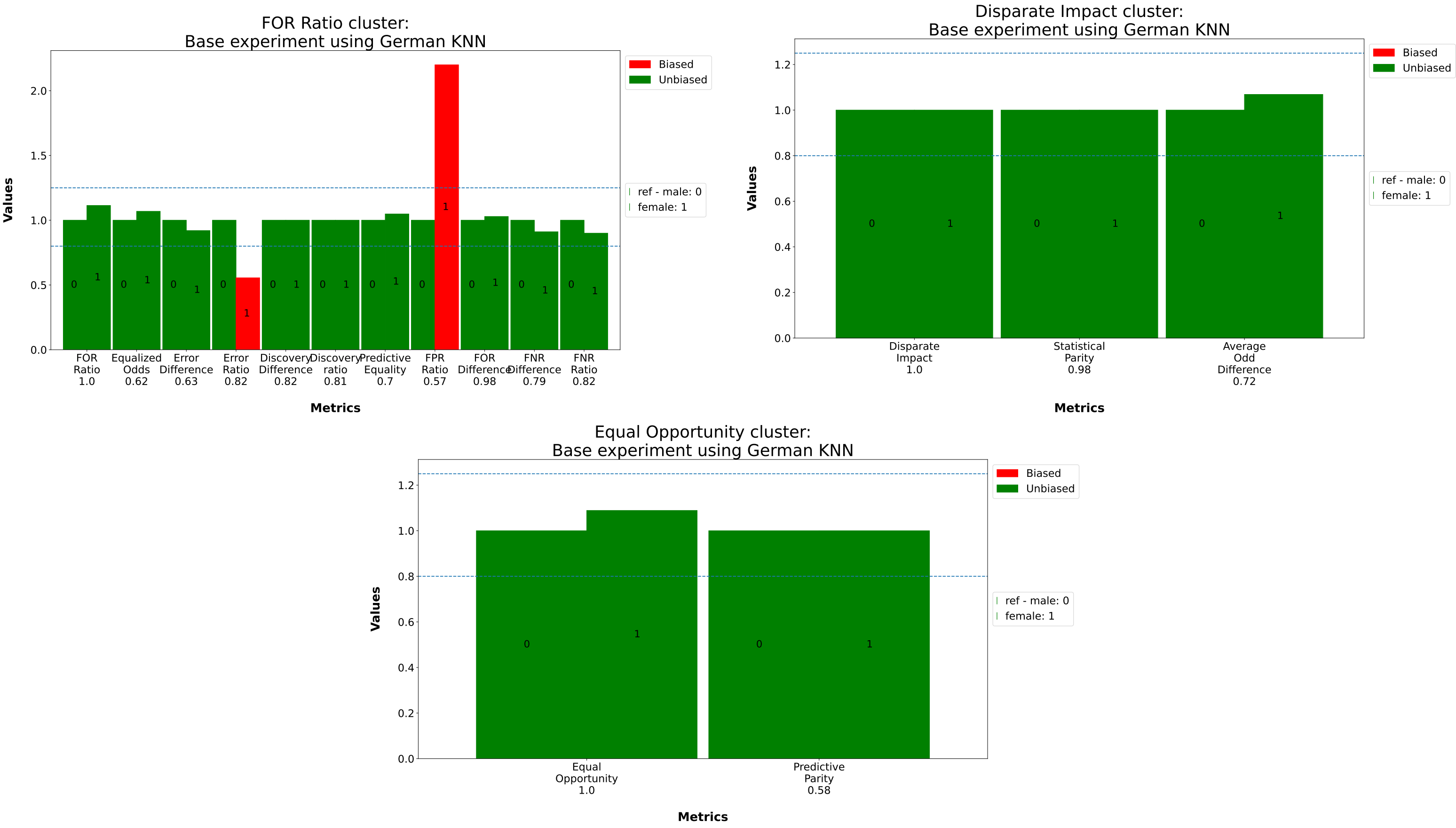


Figure S.36: Base Experiment clusters using German dataset and KNN model.

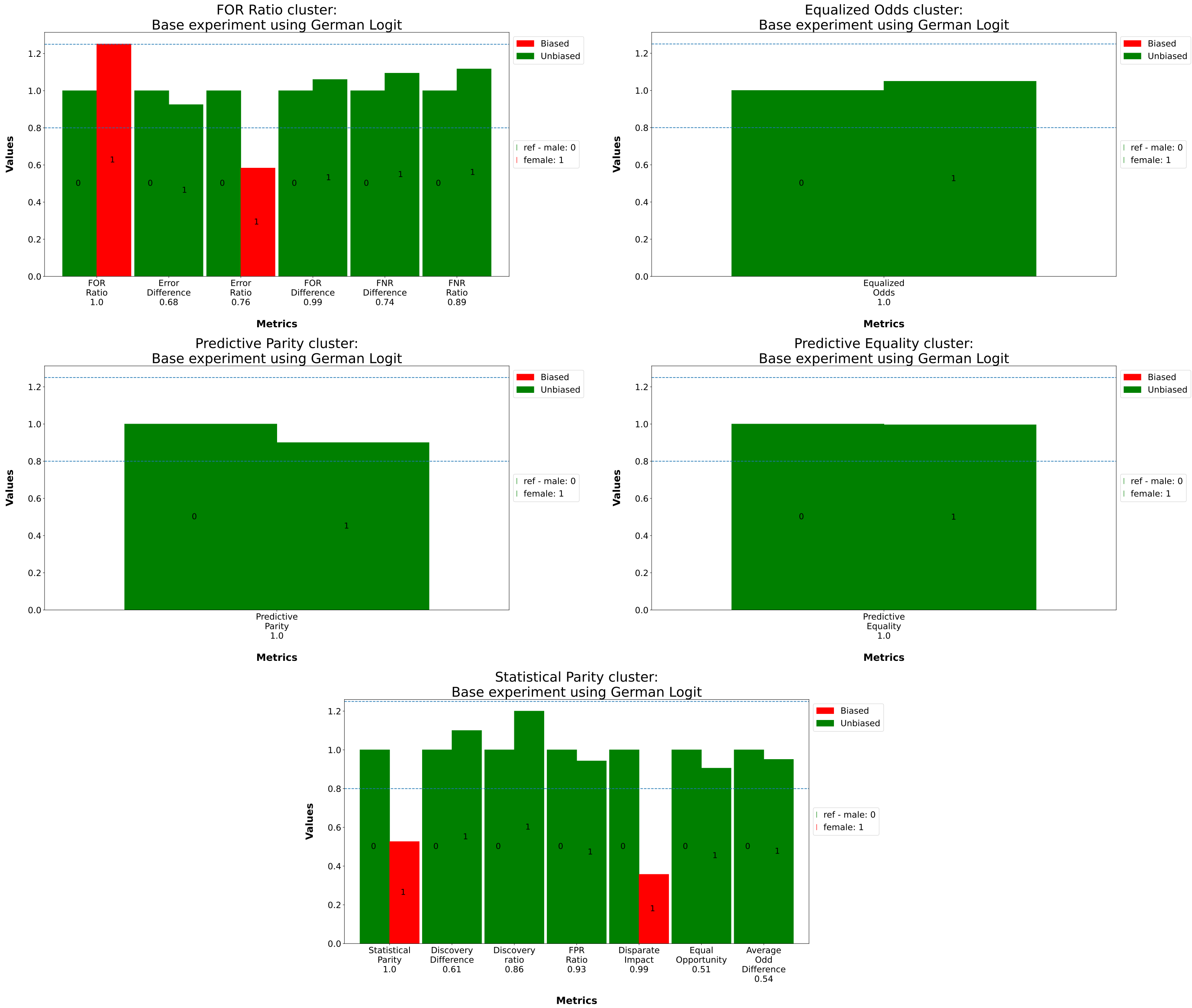


Figure S.37: Base Experiment clusters using German dataset and Logit model.

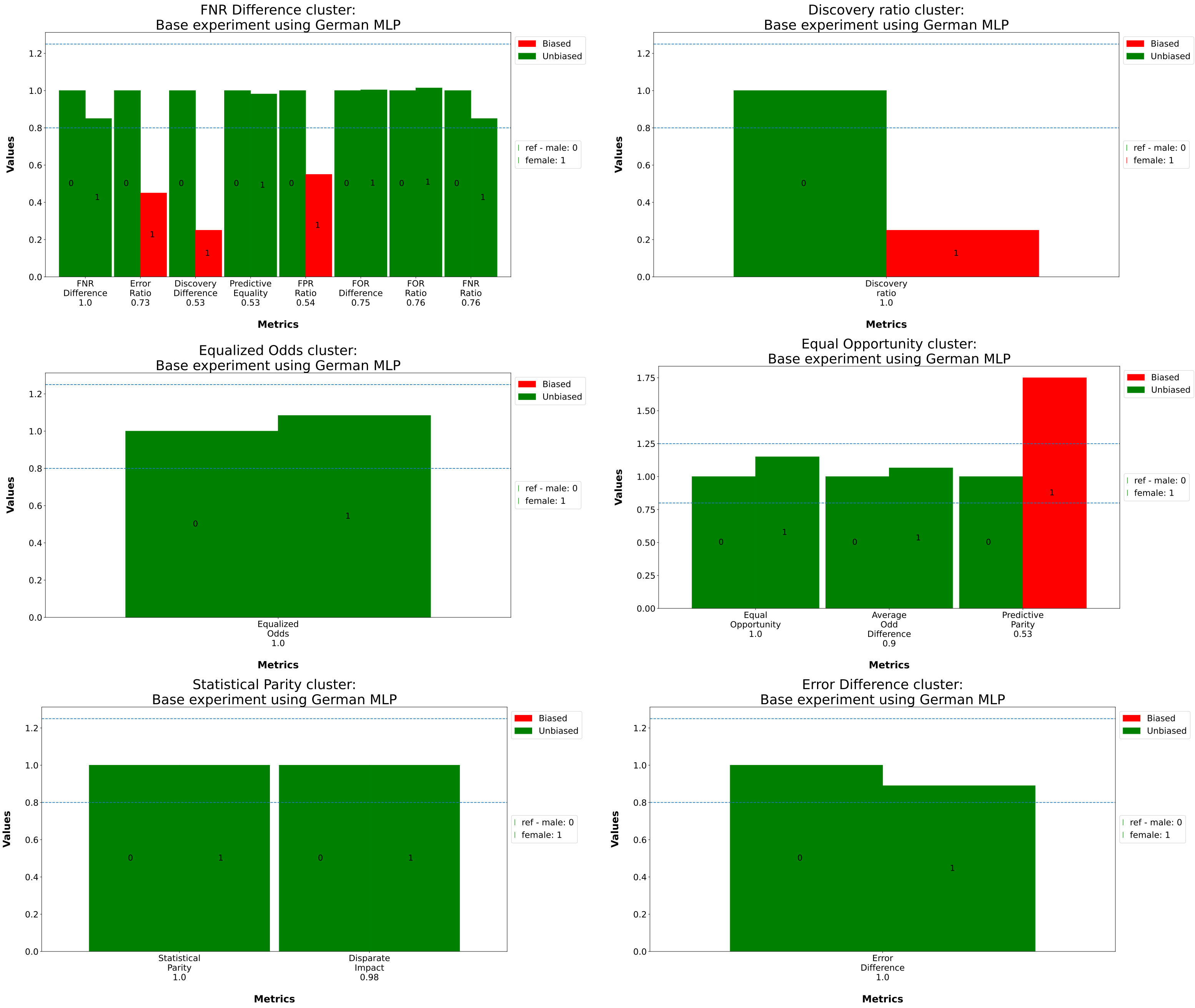


Figure S.38: Base Experiment clusters using German dataset and MLP model.

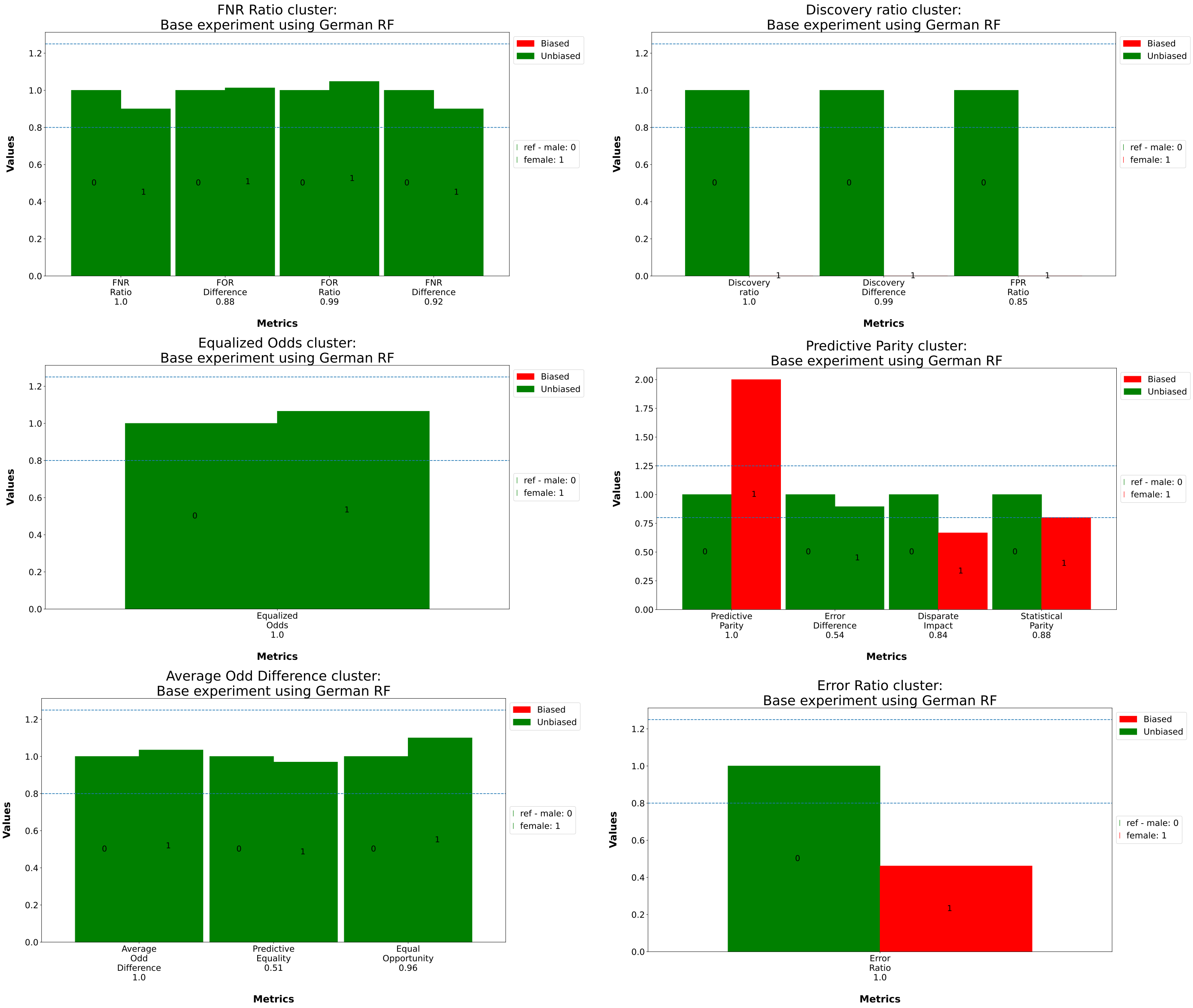


Figure S.39: Base Experiment clusters using German dataset and RF model.

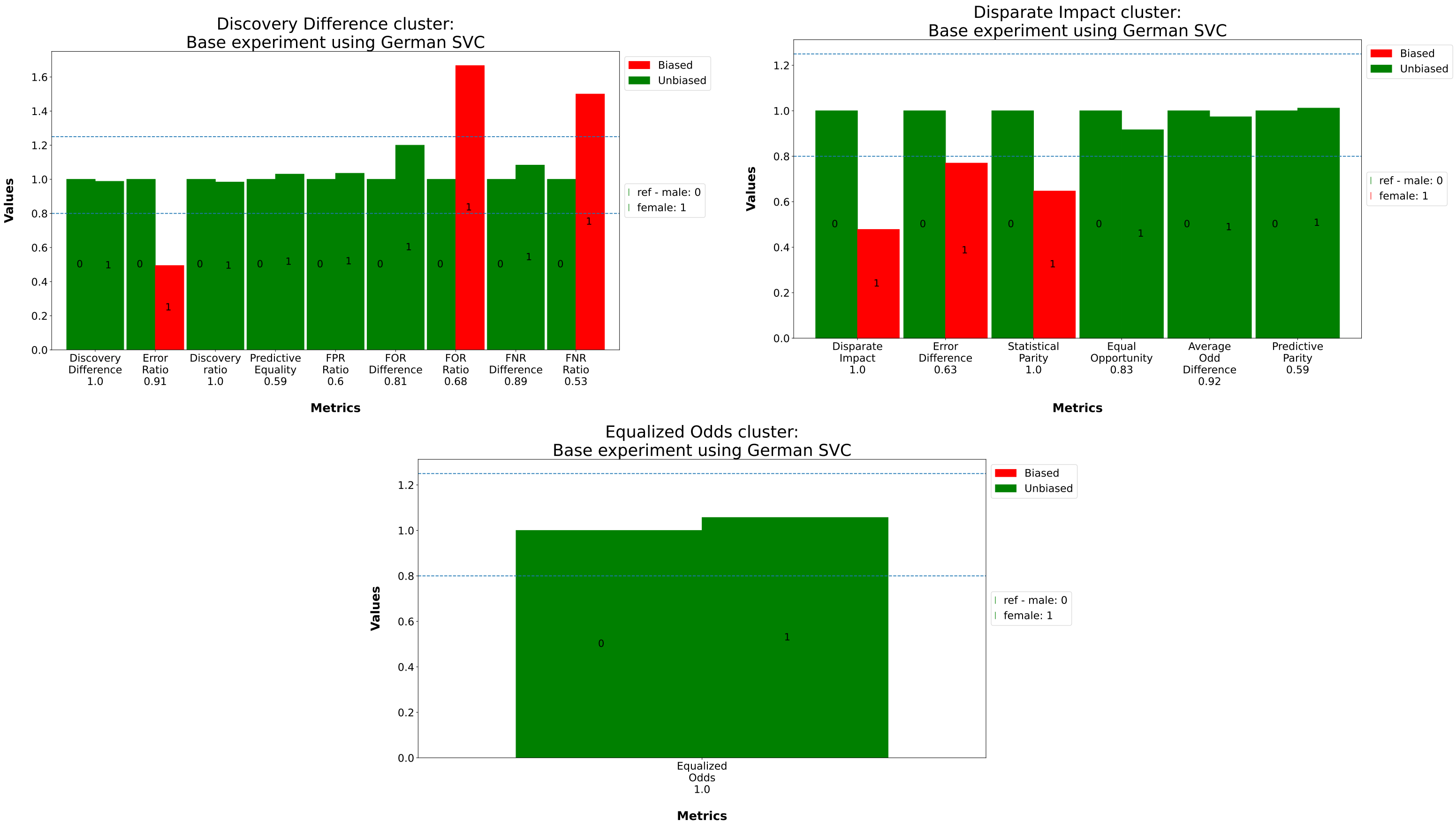


Figure S.40: Base Experiment clusters using German dataset and SVC model.

Computational method for grouping and reducing representative metrics for identification and mitigation of bias and unfairness in machine learning models.

Rafael Bessa Loureiro

Salvador, September 2024.