UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO

DIEGO DIMER RODRIGUES

Systematic assessment of data-level bias and its impact on machine learning classifiers

Thesis presented in partial fulfillment of the requirements for the degree of Master of Computer Science

Advisor: Profa. Dra. Mariana Recamonde Mendoza

Porto Alegre
July 2026

CIP — CATALOGING-IN-PUBLICATION

Rodrigues, Diego Dimer

Systematic assessment of data-level bias and its impact on machine learning classifiers / Diego Dimer Rodrigues. – Porto Alegre: PPGC da UFRGS, 2026.

131 f.: il.

Thesis (Master) – Universidade Federal do Rio Grande do Sul. Programa de Pós-Graduação em Computação, Porto Alegre, BR–RS, 2026. Advisor: Mariana Recamonde Mendoza.

1. Machine learning. 2. Bias. 3. Pre-training bias metrics. 4. Data-level bias metrics. 5. Model evaluation. 6. Health. I. Recamonde Mendoza, Mariana. II. Título.

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
Reitora: Prof[a]. Marcia Barbosa
Vice-Reitor: Prof[a]. Pedro Costa
Pró-Reitora de Pós-Graduação: Prof[a]. Claudia Wasserman
Diretor do Instituto de Informática: Prof. Luciano Paschoal Gaspary
Coordenador do PPGC: Prof. Gabriel Luca Nazar
Bibliotecário-chefe do Instituto de Informática: Alexsander Borges Ribeiro

ABSTRACT

Machine learning models are increasingly deployed in high-stakes domains such as healthcare, criminal justice, and finance, where biased predictions can cause systematic harm to specific population groups. A critical challenge in this context is the presence of data-level bias - distributional disparities rooted in the underrepresentation of historically disadvantaged groups - which can propagate into models and produce unfair outcomes before any prediction is made. While existing fairness toolkits offer a wide range of metrics to evaluate models and mitigation strategies, there is a lack of systematic empirical evidence linking data-level statistical properties to downstream model-level disparities, leaving practitioners without reliable early-warning indicators. This thesis addresses that gap by systematically investigating how four data-level metrics - Class Imbalance, Kullback-Leibler Divergence, Kolmogorov-Smirnov, and Demographic Disparity - correlate with fairness outcomes across four supervised learning algorithms. To this end, we apply controlled manipulations to 11 diverse datasets, generating balanced and severely biased variants that enable a rigorous comparative analysis. Fairness impact is assessed both in terms of classifier performance and model-level bias metrics, including Disparate Impact, Difference in Positive Proportions in Predicted Labels, and Differences in Conditional Acceptance and Conditional Rejection. The results demonstrate that distributional metrics, particularly the divergence-based ones, outperform class imbalance as indicators of downstream unfairness; that no algorithm is inherently robust to data-level bias; and that balanced training consistently reduces model reliance on sensitive attributes, providing an evidence-based foundation for embedding fairness audits at the earliest stage of the machine learning pipeline.

Keywords: machine learning. bias. pre-training bias metrics. data-level bias metrics. model evaluation. health.

LIST OF FIGURES

Figure 2.1 Class Imbalance Example 19
Figure 2.2 KL Divergence Example 20
Figure 2.3 Kolmogorov-Smirnov Example 20
Figure 2.4 Demographic Disparity Example 21
Figure 2.5 Difference in Positive Proportions in Predicted Labels Example 22
Figure 2.6 Disparate Impact Example 23
Figure 2.7 Difference in Conditional Outcome Example 24
Figure 2.8 Example of a decision tree generated for the Heart Dataset. 27
Figure 2.9 Example of a confusion matrix for a binary classification task. 30
Figure 2.10 10-Fold cross-validation example 32

Figure 4.1 Datasets generated during the bias manipulation process 45
Figure 4.2 High-level experimental setup. Data is collected, three variations are generated and used to evaluate data-level bias, a hyper-parameter tuning step is applied before completing the model training, then performance is evaluated on the testing sets, and model-level bias metrics are computed 48

Figure 5.1 Datasets distributions per target variable. The X-axis is the target variable (1 being the positive outcome), and the Y-axis is the number of instances, where orange is the privileged class (1) and blue is the unprivileged class (0). 51
Figure 5.2 F1-score by training data variation for privileged and unprivileged groups across datasets. Each panel shows the mean F1-score ($\pm$ standard deviation) over four classifiers, evaluated on the original test distribution under three training conditions: original (Orig), equally balanced (EB), and highly biased (HB) 53
Figure 5.3 Hierarchical clustering of datasets by F1-score gap profile, using Euclidean distance and Ward.D2 linkage method. 55
Figure 5.4 F1-Score gaps across all datasets, testing on the original test set. The gap is calculated as the difference between the F1-score for the advantaged class and the F1-Score for the disadvantaged class. 65
Figure 5.5 Mean F1-Score gaps under HB and EB training, HB, Orig and EB test for all datasets. 66
Figure 5.6 Approximated mean test size (from original testing variation) in blue, compared to F1-Score gap (class advantaged - class disadvantaged), in red. 67
Figure 5.7 Comparison of DI and CI across all experiments, with different colors for each train variation. 69
Figure 5.8 Heatmap of Model Rejection Metrics (DCA and DCR) across different Train $\rightarrow$ Test variation splits, with cell background color intensity scaled by the training set's Demographic Disparity (DD). Extreme negative values (DCA $\rightarrow -\infty$) and massive positive inflation (DCR $\rightarrow +\infty$) strongly correlate with saturated dark-red rows representing high data-level bias. 71

Figure 5.9 Feature importance bump charts for HEART, IBD, COMPAS and ACTG175 under XGB, rank is the feature importance order for the given training set. The protected attribute is marked with a star and highlighted with a red arrow. 78

Figure 7.1 Box-plot results for accuracy on all the analyzed datasets. X axis is the train variation, Y axis is the test variation, in the following order: Original Dataset (blue), Equally Balanced Dataset (Orange), and Highly Biased Dataset (green). 86

Figure 7.2 Box-plot results for F1-score on all the analyzed datasets. X axis is the train variation, Y axis is the test variation, the bars, from left to right correspond to: Original Test (unprivileged group), Original Test (Privileged Group), Original Test (Overall), Equally Balanced Test (Unprivileged Group), Equally Balanced Test (Privileged Group), Equally Balanced Test (Overall), Highly Biased Test (Unprivileged Group), Highly Biased Test (Privileged Group), and Highly Biased Test (Overall). 87

LIST OF TABLES

Table 3.1 Summary of the analyzed works 39

Table 4.1 Relation of datasets used in this work. The target attribute is related to the task described in Section 4.1. 42

Table 5.1 Pre-training dataset characteristics for the original (Orig) and highly biased (HB) training variations. Arrows in column headers denote the direction associated with increased bias. 52

Table 5.2 F1-score gaps (pp) between privileged and unprivileged groups across all datasets and train/test variations. Positive values indicate higher F1 for the privileged group; negative values indicate the reverse. Dataset groupings (A–D) follow the hierarchical clustering of F1-gap profiles. 54

Table 5.3 Model-level bias metrics across test and train variations for the EMP and BANK datasets (Group A). 56

Table 5.4 Model-level bias metrics across test and train variations for the HEART, COMPAS, ADULT, and ILPD datasets (Group B). 59

Table 5.5 Model-level bias metrics across test and train variations for the PS-DAS, ACTG175, IBD, and DIAB datasets (Group C). 62

Table 5.6 Model-level bias metrics across test and train variations for the GLIOMA dataset (Group D). 64

Table 5.7 Summary of data-level bias metrics and DI for key datasets mentioned on RQ2.2 and RQ2.3. 70

Table 5.8 Protected attribute mean feature importance rank across four models (LR, DT, RF, XGBoost) by training variation. Lower rank indicates higher importance. Δ Orig$\rightarrow$HB is the change in mean rank from original to highly biased training (negative values indicate the attribute rose in importance). "HB rk. 1" reports how many of the four models assign rank 1 to the protected attribute under HB training. 74

Table 5.9 Reduction in XGBoost bias from Highly Biased (HB) to Equally Balanced (EB) training across selected datasets. 77

Table 7.1 Dataset Sizes Across Different Training and Testing Variations 84

Table 7.2 Best Hyperparameters found for RandomForestClassifier 84

Table 7.3 Best Hyperparameters found for DecisionTreeClassifier 85

Table 7.4 Best Hyperparameters found for XGBClassifier 85

Table 7.5 Complete results for all datasets, training and testing variations, and models 88

Table 7.6 Model-level fairness metrics (DPPL and DI) for all datasets, testing on the original distribution. DPPL colour bands: $|\text{DPPL}| < 0.1$ (fair), $0.1 \leq |\text{DPPL}| < 0.3$, $|\text{DPPL}| \geq 0.3$ (severe). DI colour bands: $0.8 \leq \text{DI} \leq 1.25$ (fair), otherwise (unfair). DCA and DCR are omitted due to numerical instability ($\pm\infty$) in several combinations; full values are in Appendix 7.5. 117

Table 7.7 Summary of feature importance ranks and values for the protected attribute across datasets and training variations. Orig→HB shows the change in mean rank from original to highly biased training. HB rank 1 indicates in how many of the 4 models the protected attribute was ranked 1 under highly biased training. All the values are averaged from all the models generated (10 for each variation). 117

LIST OF ABBREVIATIONS AND ACRONYMS

ADS	*Automated Decision System*
AI	*Artificial Intelligence*
AIF360	*AI Fairness 360*
AutoML	*Automated Machine Learning*
CAD	*Computer-Aided Diagnosis*
CI	*Class Imbalance*
CNN	*Convolutional Neural Network*
COVID-19	*Coronavirus Disease 2019*
CT	*Computed Tomography*
DCA	*Difference in Conditional Acceptance*
DCO	*Difference in Conditional Outcome*
DCR	*Difference in Conditional Rejection*
DD	*Demographic Disparity*
DI	*Disparate Impact*
DPPL	*Difference in Positive Proportions in Predicted Labels*
DT	*Decision Tree*
EB	*Equally Balanced*
EDA	*Exploratory Data Analysis*
EHR	*Electronic Health Records*
EMP	*Employee Dataset*
F1	*F1-Score*
FN	*False Negative*
FP	*False Positive*
GANs	*Generative Adversarial Networks*

GBM	*Glioblastoma Multiforme*
HB	*Highly Biased*
HR	*Human Resources*
IBM	*International Business Machines*
IJB-A	*IARPA Janus Benchmark A*
KL	*Kullback-Leibler*
KS	*Kolmogorov-Smirnov*
LGG	*Lower-Grade Glioma*
LR	*Logistic Regression*
LVEF	*Left Ventricular Ejection Fraction*
MIM	*Masked Image Modeling*
ML	*Machine Learning*
MLOps	*Machine Learning Operations*
NLP	*Natural Language Processing*
Orig	*Original*
PPD	*Postpartum Depression*
RF	*Random Forest*
SSL	*Self-Supervised Learning*
std	*Standard Deviation*
TN	*True Negative*
TP	*True Positive*
XGB	*Extreme Gradient Boosting*

CONTENTS

1 INTRODUCTION 12
2 THEORETICAL BACKGROUND 15
2.1 Protected Attributes 15
2.2 Definition of bias 15
2.3 Measuring bias 17
2.3.1 Data-level bias 18
2.3.1.1 Class Imbalance 18
2.3.1.2 KL Divergence 19
2.3.1.3 Kolmogorov-Smirnov 19
2.3.1.4 Demographic Disparity 21
2.3.2 Model-level bias 21
2.3.2.1 Difference in Positive Proportions in Predicted Labels 22
2.3.2.2 Disparate Impact 22
2.3.2.3 Difference in Conditional Outcome 23
2.4 Supervised learning algorithms 24
2.4.1 Logistic Regression 25
2.4.2 Decision Tree 26
2.4.3 Random Forest 27
2.4.4 XGBoost 28
2.5 Model Performance Evaluation 29
2.5.1 Accuracy 30
2.5.2 F1-Score 30
2.6 K-Fold Cross Validation 31
3 RELATED WORKS 33
3.1 Toolkits and softwares 33
3.2 Research Papers 35
4 METHODOLOGY 40
4.1 Data Collection 40
4.2 Data Preparation 42
4.3 Bias Manipulation 43
4.4 Hyperparameter Tuning 44
4.5 Model Training and Performance Evaluation 47
4.6 Bias Evaluation 47
4.7 Experimental Setup 48
5 RESULTS 50
5.1 Overall Analysis 51
5.2 Per-group Analysis 54
5.2.1 Group A: EMP and BANK 55
5.2.2 Group B: HEART, COMPAS, ADULT, and ILPD 57
5.2.3 Group C: PSDAS, ACTG175, IBD, and DIAB 60
5.2.4 Group D: GLIOMA 63
5.3 RQ1: How Does Data-Level Bias Affect Predictive Performance? . 64
5.4 RQ2: How Do Data-Level Bias Metrics Correlate With Model-Level Bias? 67
5.5 RQ3: Which Algorithms Are More Robust Against Data-Level Bias? 72
5.6 RQ4: What is the impact on Feature Importance of Protected Attributes? 75

6 CONCLUSION 79
6.1 Limitations and Future Work 82
7 APPENDIX 84
7.1 Further details about datasets 84
7.2 Best hyperparameters found by grid search 84
7.3 Accuracy and F1-Score Results 85
7.4 Complete Performance and Metrics Results 85
7.5 Feature Importances 117
7.6 Resumo Expandido 118
REFERENCES 127

1 INTRODUCTION

The recent evolution of machine learning (ML) has raised concerns about the shift from model-centric to data-centric Artificial Intelligence (AI), where data is now the focus, and we need to spend more time ensuring quality and informativeness. In the early stages of the field, research efforts were primarily focused on refining algorithmic architectures, optimization routines, and hyperparameter tuning to maximize predictive accuracy; however, as these models have transitioned from purely academic settings to high-stakes deployment in sensitive domains such as criminal justice, healthcare, and financial services, the focus has increasingly turned toward the integrity and representativeness of the underlying data distributions. This transition necessitates a rigorous evaluation of how data-level biases influence algorithmic fairness before a model is ever deployed.

Stoyanovich, Howe and Jagadish (2020) encapsulates this challenge through the "Mirror Metaphor", conceptualizing data not as an objective truth but as a "distorted mirror" of societal structures. Under this framework, bias is not a monolithic failure of code but a multi-faceted reflection of the ML lifecycle, categorized into pre-existing biases rooted in historical inequalities, technical biases arising from data processing, and emergent biases that manifest through system-environment interactions. Understanding this mirror is critical, as the deployment of biased models in public sectors can amplify existing societal disparities and cause systemic harms to specific populations, a concern that has led to an expanded call for toolkits and decision-making frameworks (FISCHER-ABAIGAR et al., 2024).

To navigate this complex landscape, researchers have developed comprehensive taxonomies to categorize the sources of algorithmic harm. Suresh and Guttag (2019) provided a seminal framework for understanding these sources, identifying distinct stages where bias enters the pipeline, including representation bias during sampling and evaluation bias during benchmarking. This is complemented by the work of Mehrabi et al. (2019), who distinguish between data-to-algorithm bias, such as representational disparities, and algorithm-to-user outcomes. Despite the existence of these taxonomies, a significant gap remains in our ability to quantify the direct relationship between pre-training statistical properties and post-training predictive behavior. While industry standards such as Amazon SageMaker Clarify have popularized pre-training metrics like Class Imbalance (CI) and the Kolmogorov-

Smirnov (KS) test to detect disparities in data distributions (HARDT et al., 2021), there is a lack of systematic evidence correlating these metrics with downstream model-level disparities across diverse datasets. The foundational study by Chouldechova (2017) on the COMPAS recidivism prediction instrument highlighted how disparate impact can arise even when predictive parity is intended, demonstrating the inherent difficulty in balancing competing fairness definitions. Similarly, in the medical domain, Larrazabal et al. (2020) demonstrated that gender imbalances in imaging datasets lead to significantly biased classifiers, underscoring the life-critical implications of data-level distribution integrity.

This thesis addresses these critical gaps by systematically investigating the impact of data-level bias on supervised learning models through granular assessment and controlled manipulation. While existing toolkits like AI Fairness 360 offer a wide range of metrics for detecting and mitigating bias (BELLAMY et al., 2018a), they often treat the model as a black box or focus on post-hoc interventions. Our research, by contrast, seeks to determine whether model-level bias can be accurately anticipated using data-level metrics such as the Class Imbalance (CI), Kullback-Leibler (KL) Divergence, Kolmogorov-Smirnov (KS) and Demographic Disparity (DD) before training begins (HARDT et al., 2021). We evaluate these relationships in classification tasks across 11 diverse datasets, ranging from the COMPAS criminal recidivism data and the Adult income dataset to specialized medical cohorts such as the Indian Liver Patient Dataset and Glioma grading records. By artificially manipulating these datasets into "Equally Balanced" (EB) and "Highly Biased" (HB) variations, we create a controlled experimental environment to observe how distributional shifts translate into predictive disparities. This study seeks to conduct a more in-depth examination of the bias-related phenomena discussed in Rodrigues and Recamonde-Mendoza (2025), Leal et al. (2025), Barros et al. (2026), Barros et al. (2025), while concurrently extending the existing literature with systematically designed experiments and empirically grounded conclusions derived from comprehensive testing and validation procedures.

The scope of this research is defined by four primary research questions: First, we investigate how data-level bias affects predictive performance, specifically whether equalizing data distributions maintains F1-scores while reducing fairness gaps. Second, we explore the correlation between data-level metrics (KL, KS, DD, CI) and model-level metrics such as Difference in Positive Proportions in Predicted

Labels (DPPL) and Disparate Impact (DI). We specifically ask whether distributional metrics such as KL and KS yield fewer false positives than simple Class Imbalance when predicting the risk of downstream unfairness. Third, we examine the robustness of different algorithms: Logistic Regression, Decision Trees, Random Forests, and XGBoost to determine which architectures are most susceptible to "absorbing" bias from the training data and to investigate if ensemble models are more prone to bias than linear models. Finally, we analyze the impact of bias on the feature importance of protected attributes, investigating whether models trained on biased data disproportionately rely on sensitive variables like gender or race for their predictions, and whether these biased models are guaranteed to use sensitive variables as proxies for the target. Through this systematic analysis, we aim to provide a comprehensive overview of how data-level audits can serve as early indicators of model bias, ultimately contributing to more responsible and transparent AI deployment in society. Recent scoping reviews have highlighted that despite the growth in fairness research, there is still a lack of evidence-based standards for clinical AI fairness (LIU et al., 2025), and our work aims to contribute to this need by bridging the gap between statistical pre-training audits and empirical model outcomes.

This work is structured as follows. In Chapter 2, we provide a theoretical background on all the concepts and methods used in this work: an overview of what protected attributes are and how to identify them, a section on how bias is defined in the literature, and what definitions are adopted throughout this work. We proceed to define all the metrics calculated on the experiments, how performance was measured, and which supervised learning algorithms were used. In Chapter 3, we define what already exists in the literature and where our work fits by bridging the gap between data-level bias and fairness outcomes from generated models. In Chapter 4, we define the methodology adopted to run the experiments, ranging from the dual approach and the variations generation. In Chapter 5, we present our results by analyzing all datasets included in the study and addressing our four research questions and their subquestions. Finally, in Chapter 6 we present our conclusions, discuss the limitations of this work, and outline directions for future research.

2 THEORETICAL BACKGROUND

In this chapter, we detail the theoretical background for this work, starting by defining what protected attributes are in Section 2.1, and then we detail how the literature defines and classifies bias in Section 2.2. In Section 2.3, we define and conceptualize the metrics used in the dual approach taken in this work: bias at the data level and at the model level. In Sections 2.4, 2.5, and 2.6, we define the theoretical baseline for the methodology: the algorithms used for classification, how performance was evaluated, and what strategy we used in splitting the data for training and testing.

2.1 Protected Attributes

In the context of machine learning fairness, protected attributes, or "sensitive attributes" are central to many definitions and mitigation strategies. These attributes refer to characteristics, such as gender or ethnicity, that are intrinsic to the population. Depending on the prediction task, these attributes can be removed from the training; however, it is essential to note that all attributes in a dataset can be correlated with the protected attributes and serve as a proxy for the sensitive characteristic (RUBACK; CARVALHO; AVILA, 2022). Furthermore, depending on the problem's domain, such as health prognosis, these attributes can be essential to the model and cannot be removed. In this work, **protected attributes** and **sensitive attributes** are used interchangeably, and are divided into two groups: **advantaged** (or privileged), the group that historically has privileges and is favored by the bias; and **disadvantaged** (or unprivileged) group, which has historically been affected by bias and may or may not be socially considered a minority.

2.2 Definition of bias

In Stoyanovich, Howe and Jagadish (2020), the author proposes a metaphor for understanding bias: conceptualizing data as a mirror reflection of the world. This reflection can be distorted, and one reason is that the mirror itself can be distorted; another is that the mirror reflects a world that is itself distorted. The

author proceeds to classify bias into three categories: Pre-existing bias, Technical Bias, and Emergent Bias.

- **Pre-existing Bias**: bias that originates in society and cannot be scientifically verified; it must be based on a belief system. One well-documented example of this bias category is the historical under-representation of women and minorities in the workforce, which can lead to their under-representation in the training data of an automated decision system (ADS). This can cause the ADS to reject more minority applicants, reproducing the pre-existing bias.
- **Technical Bias**: arises from the operation of the system, which can amplify pre-existing bias. The most common sources of technical bias are data cleansing (for example, if a form offers "male", "female", and "unspecified" gender options, women are more likely to omit their gender, using a common imputation method like mode imputation would incorrectly set all unspecified values to "male" (STOYANOVICH; HOWE; JAGADISH, 2020)), data transformation (spell correction, stemming, among others), and filtering (selecting based on a predicate, like ZIP code, that can alter the data distribution).
- **Emergent bias**: arises from the context of how a technical system is used. It differs from the other categories because it is not a property of the data or the system's internal operations, but an outcome of how the system interacts with its environment and human users. One classic example is the "rich-get-richer" phenomenon in web ranking and e-commerce. Users tend to trust that the most suitable items are at the top of a ranked list, which in turn influences their idea of a satisfactory result.

In this work, we focus primarily on the effects of what Stoyanovich, Howe and Jagadish (2020) characterized as **Technical Bias**, and how it can combine with **Pre-existing bias** to lead to **Emergent Bias** through data collection and manipulation.

In Mehrabi et al. (2019), the author presents a new taxonomy for bias, dividing it into Data to Algorithm, Algorithm to User, and User to Data. The categories themselves have various types of bias. It is essential to note that in this work, we focus on the **Data-to-Algorithm** and **Algorithm to User** types of bias, specifically on the **Representation Bias** and **Evaluation Bias**.

- Data to Algorithm: bias present in the data, when using this biased dataset for training, they can produce biased algorithmic outcomes. **Representation**

Bias arises from how a population is sampled during data collection, when the samples are not representative of the entire population. One classic example is the *ImageNet* dataset (DENG et al., 2009), which shows a strong bias toward Western cultures, due to a lack of diversity in the data collection.

- Algorithm to User: biases from algorithmic outcomes that can affect the user behavior. **Evaluation Bias** happens during model evaluation and involves the use of inappropriate benchmarks. A classic example is the use of benchmarks like Adience and IJB-A for facial recognition systems, which are biased toward light-skinned subjects.

In Suresh and Guttag (2019), the author proposes five different sources of bias that can arise from data: Historical Bias, Representation Bias, Measurement Bias, Aggregation Bias, and Evaluation Bias. In this work, we focus on the authors' definitions of **Representation bias** and **Evaluation bias**.

- **Representation bias** occurs when certain parts of the input space are under-represented in the collected data. This can also happen if the data is perfectly measured and sampled from the true population, as the minority groups will make up only a small percentage of the data, leading to a less robust model for that group. It can also arise when the sampling only reaches a portion of the population, as described in Mehrabi et al. (2019), or if the population changes over time.
- **Evaluation bias** happens when the evaluation or benchmark data used to measure the algorithms quality does not reflect the target population, as described in Mehrabi et al. (2019). An under-representative benchmark encourages the model development to only perform well on a subset, and it can also be exacerbated by using aggregate performance measures, which can hide subgroup underperformance. One example is that commercially used facial analysis algorithms (for tasks like gender or smiling detection) have been shown to have drastically worse performance on dark-skinned females.

2.3 Measuring bias

In this work, we focus primarily on detecting bias with data-level bias metrics. These metrics are statistical measures that can be derived from the data, without the

predictions from machine learning algorithms (BARROS et al., 2025). We note that, in the literature, these metrics can also be referred to as **Pre-training** bias metrics. In this work, we adopt the term **data-level bias** metrics to avoid ambiguity, as "pre-training" carries a distinct meaning in the context of transfer learning and large language models. The metrics that will be presented in this section and adopted throughout this work were defined in Hardt et al. (2021) and further expanded in Barros et al. (2025).

2.3.1 Data-level bias

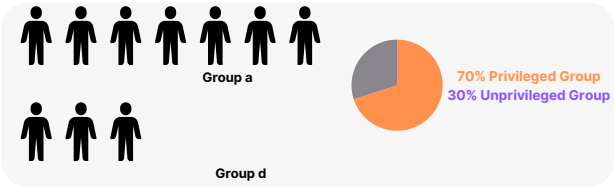
Data-level bias metrics, also referred to in the literature as *Pre-training* bias metrics, are the statistical metrics used in this work to evaluate the bias present in the *data*, prior to any training or prediction. These metrics are: Class Imbalance, KL Divergence, Kolmogorov-Smirnov and Demographic Disparity (HARDT et al., 2021; BARROS et al., 2025). It is important to note that all four metrics are designed to evaluate group fairness, as they quantify disparities between predefined demographic groups. Consequently, they do not capture notions of individual fairness, which concern whether similar individuals receive similar treatment regardless of group membership. The choice of metrics was based on evaluating available metrics in the literature and leveraging different ways of detecting data-level bias (both distributional and representational approaches).

2.3.1.1 Class Imbalance

Class Imbalance (CI) refers to situations in which a specific category of a protected attribute is underrepresented in a dataset. It extends the standard class-balancing concept by measuring the representational disparity between the privileged group (n_a) and the unprivileged group (n_d), as denoted on Equation 2.1. The metric produces a score between -1 and 1, where 0 indicates a perfect balance. As shown in the visual example in Figure 2.1, a dataset with a 70% privileged and 30% unprivileged split yields an imbalance of 0.4. This positive value demonstrates an imbalance that favors the privileged class.

$$CI = (n_p - n_d)/(n_p + n_d) \tag{2.1}$$

Figure 2.1 – Class Imbalance Example



$$\frac{(n_a - n_d)}{(n_a + n_d)} = \frac{70 - 30}{70 + 30}$$

source: the Author

2.3.1.2 KL Divergence

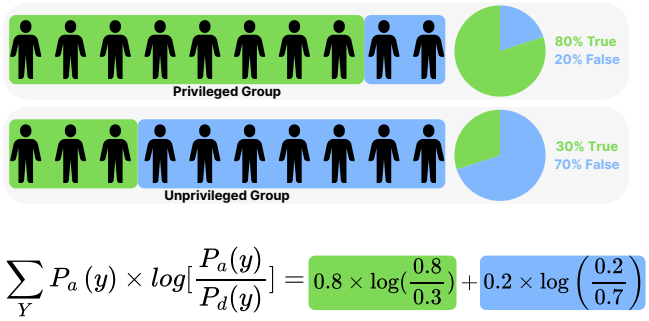
Kullback-Leibler (KL) Divergence, also known as relative entropy, evaluates the asymmetric divergence between two probability distributions. Because it lacks symmetry, the divergence from the privileged to the unprivileged group differs from the reverse calculation ($KL(P_a||P_d) \neq KL(P_d||P_a)$), which means that this metric can not be considered a distance metric. It measures how the target attribute is distributed across different demographic groups, producing values ranging from 0 (identical distributions) to infinity (maximum divergence). Based on the visual example in Figure 2.2, if the privileged class has an 80% True and 20% False distribution, while the unprivileged class has a 30% True and 70% False distribution, the calculation applying the formula on Equation 2.2, is $0.8 \times \log\left(\frac{0.8}{0.3}\right) + 0.2 \times \log\left(\frac{0.2}{0.7}\right)$. This resulting positive value indicates a divergence that favors the privileged group.

$$KL(P_p||P_d) = \sum_{Y} P_p(y) * \log[P_p(y)/P_d(y)] \tag{2.2}$$

2.3.1.3 Kolmogorov-Smirnov

The Kolmogorov-Smirnov (KS) metric utilizes a non-parametric statistical approach to detect the maximum probability divergence across all possible model predictions for the distinct protected attribute classes. KS values fall between 0 (perfect equality) and 1 (all instances belong to a single class). Applying the previ-

Figure 2.2 – KL Divergence Example

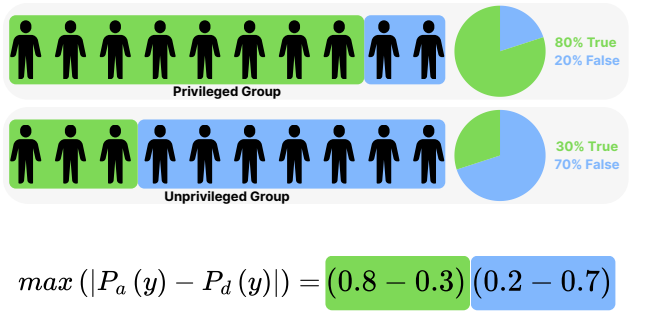


source: the Author

ous visual example's distributions (80% True / 20% False for the privileged group and 30% True / 70% False for the unprivileged group), as depicted on Figure 2.3 and applying the formula in Equation 2.3, the KS metric evaluates the maximum absolute difference: $\max(|0.8 - 0.3|, |0.2 - 0.7|)$. This yields 0.5, indicating a significant demographic imbalance.

$$KS = max(|P_p(y) - P_d(y)|) \tag{2.3}$$

Figure 2.3 – Kolmogorov-Smirnov Example



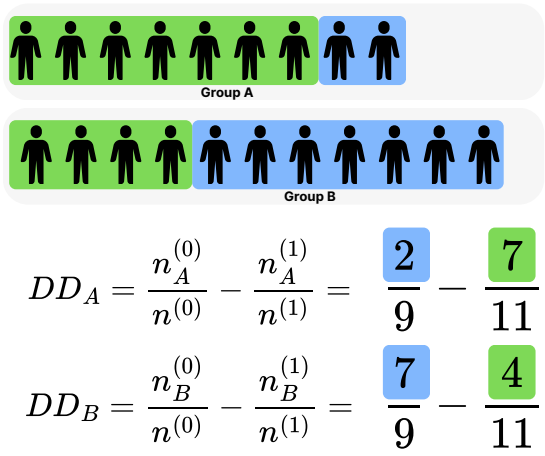
source: the Author

2.3.1.4 Demographic Disparity

Demographic Disparity (DD) assesses the proportional differences in outcomes across various demographic groups. For this work, we consider two groups: the advantaged and the disadvantaged. In the visual example, in Figure 2.4, Group A consists of 9 total individuals (2 False, 7 True) and Group B contains 11 individuals (7 False, 4 True). The disparities are determined by the ratio of favorable to unfavorable outcomes. This is calculated, according to Equation 2.4 as $\frac{2}{9} - \frac{7}{11}$ for Group A, and $\frac{7}{9} - \frac{4}{11}$ for Group B. This metric establishes a direct baseline for comparing how outcomes are distributed across different segments.

$$DD_d = \frac{n_d^{(0)}}{n^{(0)}} - \frac{n_d^{(1)}}{n^{(1)}} \tag{2.4}$$

Figure 2.4 – Demographic Disparity Example



source: the Author

2.3.2 Model-level bias

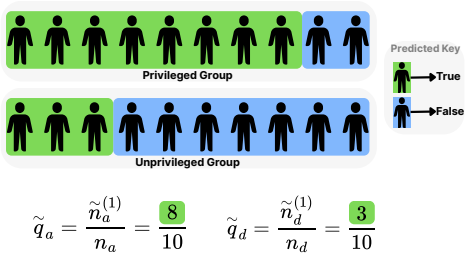
Model-level bias, also referred to as *Post-training* bias metrics, are metrics that account for the model **output** and are applied to its results to quantify its fairness. In this work, we leverage three model-level bias metrics: Difference in Positive Proportions in Predicted Labels, Disparate Impact, and Difference in Conditional Outcome (HARDT et al., 2021; BARROS et al., 2025). These metrics were chosen for their popularity and their ability to audit the outcome of machine learning models, allowing to correlate with data-level bias metrics.

2.3.2.1 Difference in Positive Proportions in Predicted Labels

The Difference in Positive Proportions in Predicted Labels (DPPL) quantifies the gap in favorable predictions ("positive outcomes") across protected attribute classes. For binary tasks, the score ranges from -1 to 1. Positive scores favor the privileged group, negative scores favor the unprivileged group, and zero denotes parity. In the example in Figure 2.5, the privileged group receives 8 positive predictions out of 10 ($\tilde{q}_a = \frac{8}{10}$), and the unprivileged group receives 3 out of 10 ($\tilde{q}_d = \frac{3}{10}$). Applying the formula in Equation 2.5, the DPPL is calculated as $\frac{8}{10} - \frac{3}{10}$, indicating a predictive bias that favors the privileged group.

$$\hat{q}_p = \frac{\hat{n}_p^{(1)}}{n_p} \qquad \hat{q}_d = \frac{\hat{n}_d^{(1)}}{n_d} \qquad DPPL = \hat{q}_p - \hat{q}_d \tag{2.5}$$

Figure 2.5 – Difference in Positive Proportions in Predicted Labels Example



$$\tilde{q}_a = \frac{\tilde{n}_a^{(1)}}{n_a} = \frac{8}{10} \qquad \tilde{q}_d = \frac{\tilde{n}_d^{(1)}}{n_d} = \frac{3}{10}$$

$$DPPL = \frac{8}{10} - \frac{3}{10}$$

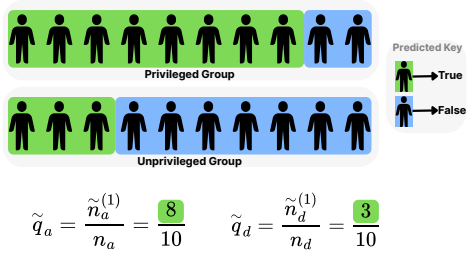
source: the Author

2.3.2.2 Disparate Impact

Disparate Impact (DI) measures the ratio of favorable prediction rates between the distinct protected attribute classes. The metric ranges from 0 to infinity. Scores below 1 show a higher proportion of positive predictions for the privileged group, while scores above 1 highlight an advantage for the unprivileged group. In the visual scenario in Figure 2.6 with 10 instances per group, the unprivileged group's positive proportion is 0.3, compared to 0.8 for the privileged group. The Disparate Impact, calculated using the formula in Equation 2.6, is $\frac{0.3}{0.8}$, confirming a clear proportional advantage for the privileged group.

$$DI = \frac{\hat{q}_d}{\hat{q}_p} \tag{2.6}$$

Figure 2.6 – Disparate Impact Example



source: the Author

2.3.2.3 Difference in Conditional Outcome

Difference in Conditional Outcome (DCO) compares the observed labels with the model's predictions, verifying if the predictions are balanced relative to the ground truth across different classes of the protected attribute. The motivation behind this metric is that a simple analysis of favorable outcomes (like DPPL) can be misleading, as it does not account for the underlying distribution of qualified candidates in the data.

For example, imagine an HR system where 60% of both male and female candidates are approved. Under DPPL, this model is considered unbiased. However, if the historical data shows that 90% of the female candidates were actually qualified ($Y = 1$) compared to only 50% of the men, the model is significantly under-delivering favorable outcomes to qualified women. In this case, the model is "fair" in its output distribution but "unfair" in its conditional accuracy.

From this concept, two metrics are derived: ***Difference in Conditional Acceptance (DCA)***, which accounts for the favorable outcome across the different groups, and ***Difference in Conditional Rejection (DCR)***, referring to the unfavorable outcome. In the following equations, $\hat{n}_g^{(x)}$ represents the number of instances in group g that the model predicted as x, while $n_g^{(x)}$ represents the number of instances in group g with true observed values of x. For binary tasks, g is either a (advantaged) or d (disadvantaged), and x is either 0 (unfavorable) or 1

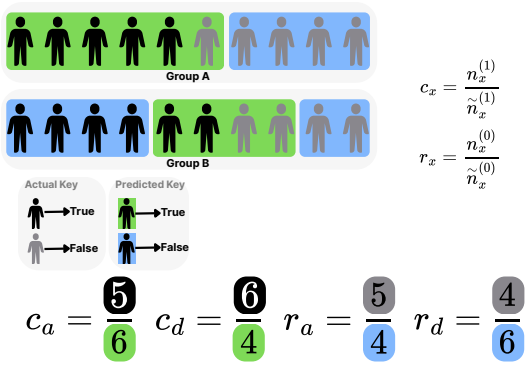
(favorable). For DCA, negative values indicate bias favoring the advantaged class (because the model over-predicts their success, shrinking the c_a ratio), and positive values indicate bias favoring the disadvantaged class. And for DCR, over-rejecting the disadvantaged class shrinks r_d, making the result of a positive value. Therefore, for DCR, positive values indicate bias against the disadvantaged class.

$$c_a = \frac{n_a^{(1)}}{\hat{n}_a^{(1)}} \qquad c_d = \frac{n_d^{(1)}}{\hat{n}_d^{(1)}} \qquad DCA = c_a - c_d \tag{2.7}$$

$$r_a = \frac{n_a^{(0)}}{\hat{n}_a^{(0)}} \qquad r_d = \frac{n_d^{(0)}}{\hat{n}_d^{(0)}} \qquad DCR = r_a - r_d \tag{2.8}$$

In the visual example in Figure 2.5, using actual and predicted outcomes, the acceptance ratios are $c_a = \frac{5}{6}$ for Group A and $c_d = \frac{6}{4}$ for Group D. The rejection ratios are $r_a = \frac{5}{4}$ and $r_d = \frac{4}{6}$. By calculating DCA ($c_a - c_d$) and DCR ($r_a - r_d$), the metric exposes nuanced underlying prediction biases.

Figure 2.7 – Difference in Conditional Outcome Example



source: the Author

2.4 Supervised learning algorithms

A dataset is a collection of labeled examples (i.e., instances with their outputs, $\{(x_i, y_i)_{i=1}^N\}$). Each element inside this example contains a value that describes the example. For instance, each labeled example in the collection might represent a person, and the first feature $x^{(1)}$ could be its height, the second its weight, and so on. For all examples in this dataset, then, the feature at position j in the features vector x always contains the same information. In classification tasks, the label

y_i identifies that the collection of elements x_i belongs to a finite set of classes C (BURKOV, 2019).

The goal of supervised learning algorithms is to use the dataset to generate a model that takes features x as inputs and outputs the information that allows deducing the class to which this input belongs. As in this work's datasets, in which binary classification tasks are derived, if the person identified by vector x has a positive or negative output. In this section, we detail the four supervised learning algorithms used in this work.

2.4.1 Logistic Regression

Logistic regression is a widely used linear model that employs a logistic function to estimate the probabilities of possible outcomes for a given dataset (FACELI et al., 2025). This model is based on the concept of odds, which represent the probability of a positive event occurring. This is defined as $\frac{p}{1-p}$, where p is the probability of the positive event. By taking the natural logarithm of the odds, we obtain the logit function described in Equation 2.9. This function allows the model to establish a linear relationship between the input features and the log-odds.

$$logit(p(y = 1|x)) = w_0x_0 + w_1x_1 + ... + w_mx_m = \sum_{i=0}^{m} w_ix_i = w^Tx \tag{2.9}$$

In this equation, $p(y = 1|x)$ is the conditional probability that a particular sample belongs to the positive class given its features x, and w represents the weight vector. To map this linear combination back to a predicted probability value constrained to [0,1], the algorithm applies the inverse of the logit function, commonly known as the sigmoid function (Equation 2.10).

$$\phi(z) = \frac{1}{1 + e^{-z}} \tag{2.10}$$

The primary objective of the logistic regression training algorithm is to find the optimal set of weights (w) that minimizes the model's cost function, often incorporating a regularization term to penalize complexity and prevent overfitting.

Based on Equation 2.9, we can also derive that, given that all the features

are in the same range (normalized to values between 0 and 1), if we analyze the weights w, we can infer that the absolute value of the coefficient is related to how important this feature is to the model, given that the changes on this value x affect most the outcome.

2.4.2 Decision Tree

The Decision Tree is a non-parametric supervised learning algorithm. The model aims to predict the value of a target variable by recursively partitioning the feature space and learning constant if-then-else decision rules inferred from the training data. While decision trees are highly interpretable, which is ideal for the medical field, and require little data preparation, they are prone to instability and can easily overfit, leading to overly complex structures that fail to generalize well. Therefore, setting hyperparameters to limit the tree's growth (such as its maximum depth) is crucial Breiman et al. (2017).

The algorithm learns by selecting splits that minimize an impurity or loss function. For a given node m representing a dataset Q_m with n_m samples, a candidate split θ partitions the data into left $Q_m^{left}(\theta)$ and right $Q_m^{right}(\theta)$. The quality of this split is computed as shown in Equation 2.11, where we calculate the weighted average of impurity of the two child nodes that would be created using this split candidate θ.

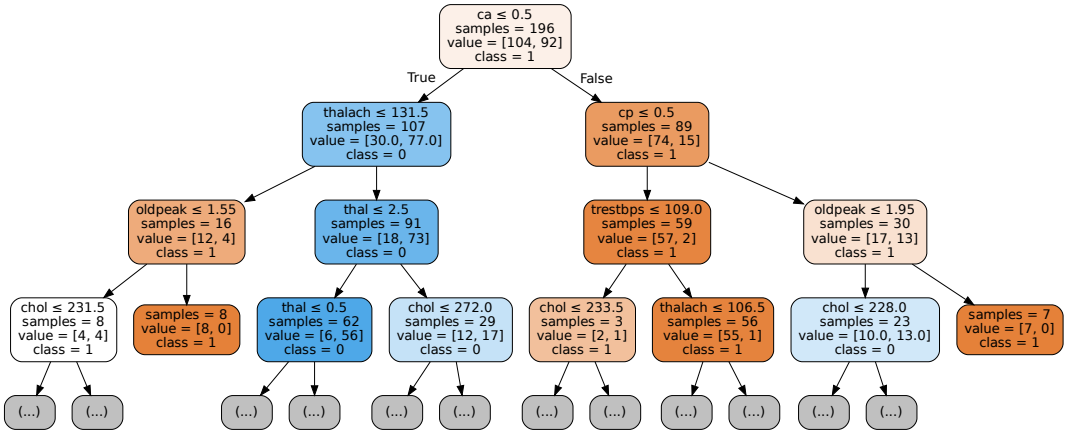
$$G(Q_m, \theta) = \frac{n_m^{left}}{n_m} H(Q_m^{left}(\theta)) + \frac{n_m^{right}}{n_m} H(Q_m^{right}(\theta)) \tag{2.11}$$

In Equation 2.11, $H()$ is the impurity function. Since we are concerned with classification, we consider the two most common metrics: the Gini index and Shannon Entropy (log loss), both of which rely on p_{mk}, the proportion of samples in node m that belong to class k. The equation 2.12 details how these metrics are calculated.

$$H_{Gini}(Q_m) = \sum_k p_{mk}(1 - p_{mk}) \quad \text{or} \quad H_{Entropy}(Q_m) = -\sum_k p_{mk} \log(p_{mk}) \tag{2.12}$$

The example in Figure 2.8 illustrates a tree generated in this work with a depth of 3 nodes from the root, for simplification. The first line in each node is the condition to be evaluated at that node, with true going to the left branch and false to the right; the second line is the number of instances (samples) in that node. The

Figure 2.8 – Example of a decision tree generated for the Heart Dataset.



source: the Author

third line, value, gives the number of instances in classes 1 and 0, respectively; the fourth line gives the majority class (the class that would be assigned if the node were a leaf node). The color relates to the class, with the opacity being the level of purity in that node. Note that leaf nodes don't have a condition; they're used to output the class for the sample.

Considering the importance of each feature, the root of the tree would be the node with the feature that most reduces impurity (the feature that most reduced impurity, using Gini or Entropy, as already defined), in this case, the feature *ca* was the most important for that tree.

2.4.3 Random Forest

While Decision Trees are well-established algorithms and a good choice when interpretability is a concern, they can easily suffer from high variance and overfitting. To overcome this issue, one possible solution is to use multiple trees so they can converge to a common solution. Based on this idea, the Random Forest algorithm was created. This algorithm uses a "perturb-and-combine" strategy to generate a diverse set of randomized trees. During construction, two sources of randomness are injected: first, each tree is built from a bootstrap sample (a sample drawn **with replacement** from the training set); second, a random subset of features is evaluated to find the best split at each node. By combining these diverse, somewhat decoupled trees, the ensemble averages out individual prediction errors, generating

a more robust and accurate model (BREIMAN, 2001).

Rather than using a simple majority vote, the Random Forest classifier aggregates the results by calculating the average of the probabilistic predictions from all the individual trees in the forest[1]. Let M be the total number of trees in the ensemble, and $\hat{p}_m(y = k|x)$ be the probability predicted by the m-th tree for class k given input x, the final probability predicted by the forest is calculated using the formula in Equation 2.13. This way, we leverage the certainty of the decision trees that the predicted class is correct, reducing noise.

$$\hat{P}(y = k|x) = \frac{1}{M} \sum_{m=1}^{M} \hat{p}_m(y = k|x) \tag{2.13}$$

Returning to the Heart Dataset example, the algorithm might draw different samples from the original data, generating different trees, as shown in Figure 2.8. These different trees might focus on different aspects of the input space, reducing overfitting and ensuring that outlier decisions do not heavily skew the classification. Analyzing how each feature contributed to the random forest model is also similar to how the decision tree does: we simply average the feature importance for each of the trees in the ensemble.

2.4.4 XGBoost

Similar to the Random Forest, Extreme Gradient Boosting (XGBoost) is also an ensemble learning method. Unlike Random Forests, which build deep trees independently and in parallel, gradient boosting is an additive model that builds shallow decision trees sequentially. Each new tree in the sequence is fitted to predict the negative gradients (or residual errors) of the loss function from the combined ensemble of the previous iterations. To prevent overfitting, the algorithm scales the contribution of each newly added tree using a shrinkage parameter known as the learning rate (that is optimized as described in Chapter 4), and directly penalizes the magnitude of the model's leaf weights using L_2 regularization. So, the goal of the XGBoost is to minimize the objective function described in Equation 2.14, where i represents a data point, n is the total number of samples, k is a tree, $l(y_i, \hat{y}_i)$ is the training loss, that for classification tasks is log loss, and $\Omega(f_k)$ is the regularization

[1] this is specific to the default implementation from Pedregosa et al. (2011)

term (in this case, L_2 regularization) (CHEN; GUESTRIN, 2016).

$$\mathcal{L}(\phi) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{2.14}$$

The final prediction $\hat{y}_i$ is the sum of K sequential weak learners (trees) predictions. The ensemble is updated at each iteration m by adding a new tree $h_m(x)$ scaled by the learning rate η. The final prediction, as described in Equation 2.15, is the sum of the initial prediction, or the "base score", that is 0.5 for binary classification, with the output of the k-th tree for that data point.

$$\hat{y}_i = \hat{y}_0 + \sum_{k=1}^{K} \eta f_k(x_i) \tag{2.15}$$

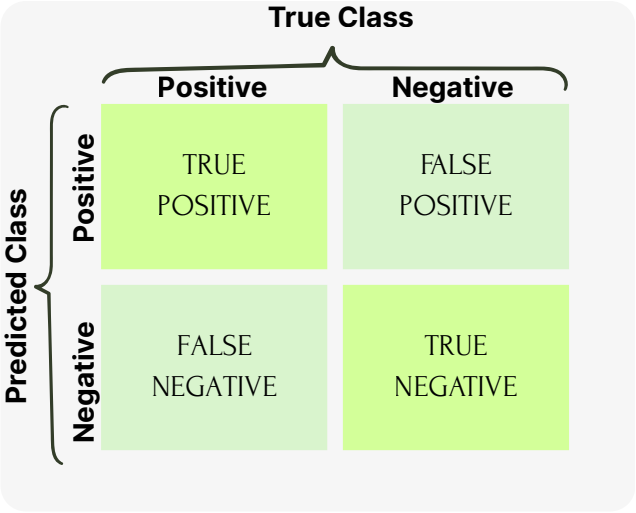
When predicting heart disease, the initial step of the XGBoost model might simply output the log-odds ratio (a constant mean prediction, 0.5 being the default value) for all patients. If the model systematically underestimates the probability of heart disease for older patients, the next tree in the sequence is explicitly trained to target those residual errors (the underestimations). The predictions are then incrementally updated by taking a small step (the learning rate) in the direction of the new tree's output, iteratively refining the prediction boundary until the loss is minimized.

To analyze the feature importance for XGBoost models, one possible approach is the average reduction in the training loss caused by the feature when it is used as a split feature. This is the closes equivalent to Gini importance for Decision Trees.

2.5 Model Performance Evaluation

Two performance metrics are used to evaluate the quality of the predictions: **Accuracy** and **F1-Score**. These metrics are well-known in the literature and can easily be applied for classification tasks. In this section, these metrics will be detailed along with the formula and examples. The base for these metrics is the confusion matrix, as illustrated in Figure 2.9.

Figure 2.9 – Example of a confusion matrix for a binary classification task.



source: the Author

2.5.1 Accuracy

Accuracy is a metric used for model evaluation, measuring the percentage of predictions that are correct. It represents the proportion of the total number of predictions that the model correctly identified, with values ranging from 0 to 1 (or, alternatively, from 0 to 100). It can be calculated using the components of the confusion matrix, specifically the true positive and true negative, divided by the total number of predictions (false positives, false negatives, true positives, and true negatives). The formula is given in Equation 2.16.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.16}$$

One of the main drawbacks of this metric is that it can be misleading, particularly when dealing with highly imbalanced datasets. In such cases, the model would get a high accuracy value when predicting the majority class for all instances.

2.5.2 F1-Score

F1-Score is calculated as the harmonic mean of precision and recall. It is known for being more appropriate for scenarios where the datasets are imbalanced, as it encompass the concept of precision (fraction of positive predictions that were actually positive) and recall (fraction of positive classes there were correctly identi-

fied). To calculate the value of F1-Score, we first define **precision** and **recall**, in Equation 2.17, and then we get the harmonic mean using Equation 2.18.

The value of F1-Score ranges between 0 and 1 (or, alternatively, from 0 to 100), with values close to 1 indicating better precision and recall, and values close to 0 indicating poor performance. In situations where the dataset is severely imbalanced, the F1-Score would typically be more reliable than the accuracy, as it would consider the model's true effectiveness on the positive class.

$$\begin{aligned} Precision &= \frac{TP}{TP + FP} \\ Recall &= \frac{TP}{TP + FN} \end{aligned} \tag{2.17}$$

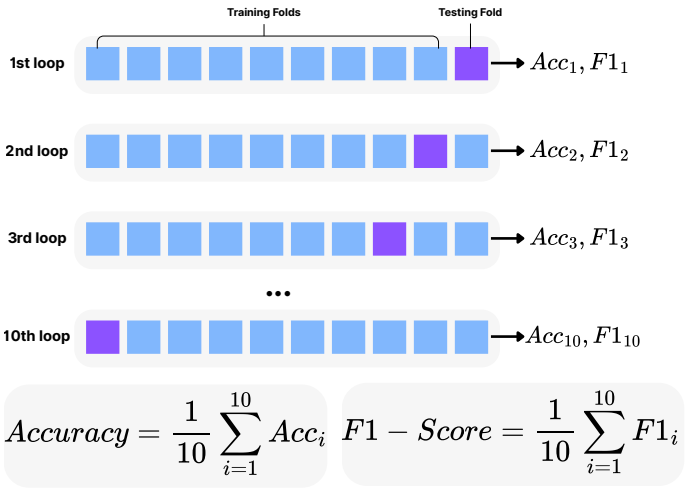
$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{2.18}$$

2.6 K-Fold Cross Validation

K-Fold cross-validation is a widely used resampling procedure for evaluating the performance of a predictive model and preventing overfitting, which can occur when a model memorizes the training data rather than generalizing to unseen observations (JAMES et al., 2021). The methodology involves randomly dividing the available set of observations into k groups, or "folds", of approximately equal size. During this procedure, the model is trained on $k - 1$ folds and evaluated in the remaining fold, as illustrated in Figure 2.10. The process is repeated k times, ensuring that each fold is used exactly once as the testing or validation fold. The model's overall performance is then calculated as the average of the k estimated validation values.

In classification tasks where datasets may exhibit rare or imbalanced classes, a variant called Stratified K-Fold is commonly used to ensure that the relative class frequencies are approximately preserved across each fold. This makes the cross-validation folds more homogeneous and helps hide some of the variability inherent to fitting models with a limited number of minority observations.

Figure 2.10 – 10-Fold cross-validation example



source: the Author

3 RELATED WORKS

In this chapter, we review related work that aims to detect, address, or study fairness in Machine Learning. In Section 3.1, we review software and toolkits for exploratory data analysis and fairness assessment as background for the literature review, our goal is to show how these metrics (data-level and model-level bias metrics) are currently available for general use in both academic and industry settings. In Section 3.2, we discuss research papers that aimed to detect, mitigate, and assess bias. We identify a gap in the literature regarding whether statistical metrics computed from the data, prior to model deployment, can consistently anticipate downstream model-level bias. Our work aims to bridge that gap by conducting an empirical study to validate the applicability of data-level bias metrics across 11 datasets from distinct domains, combining model-level bias metrics and F1-score results. Table 3.1 summarizes the analyzed works and highlights that most studies focus on post-training evaluation or non-tabular data, while only one work uses data-level metrics similar to those adopted here, but with a different objective.

3.1 Toolkits and softwares

As Artificial Intelligence evolves, it transitions from a model-centric approach to a data-centric paradigm, where the quality, representativeness, and integrity of the data determine model performance and ethical implications. Issues in the data, ranging from missing values to systematic class imbalances, can significantly impact machine learning algorithms, leading to biased architectures and models that can produce harmful outcomes in production environments. Especially in high-stakes domains, such as healthcare, preparing the data and extracting metadata information before model development has become a prerequisite for establishing trust and fairness. In the automated Exploratory Data Analysis (EDA) field, the ydata-profiling (CLEMENTE et al., 2023), initially known as pandas-profiling, provides a fast and consistent solution for understanding datasets, including a function to describe the data, that generates a comprehensive report that includes dataset overview, variable-specific statistics (like percentage of missing values, duplicated rows, class imbalance, among others), and correlation matrices. Similarly, Sweetviz (Sweetviz,) provides complete visualizations for the data distribution across

datasets, aiming to facilitate comparison between datasets or train and test data, and autoviz (SESHADRI, 2020) provides automated chart generation agnostic of the dataset, for fast exploratory analysis of data distribution, correlation between pairs of variables, and a method to fix the data quality.

Additionally, great expectations (GONG; CAMPBELL; Great Expectations,) provides high-level prescriptive guidance for data unit testing and validation, empowering users to set alerts on databases and data sources for specific scenarios, ensuring the data meets predefined quality measures. It supports various backends, such as PySpark and SQLAlchemy. For fairness evaluation and bias detection, AIF360 (BELLAMY et al., 2018b) is an open-source IBM toolkit for bias detection and mitigation. The library provides over 70 metrics to assess the fairness of the models, including generic metrics (specificity score, sensitivity score, selection rate, etc), group fairness metrics (statistical parity difference, mean difference, equal opportunity, etc), and individual fairness metrics (generalized entropy index, theil index, consistency score, etc), along with multiple methods to address and correct biases, such as Reweighting, Adversarial Debiasing, Exponentiated gradient reduction and Calibrated equalized odds post-processor, among others. Fairlearn (WEERTS et al., 2023) is an open-source toolkit focused on assessing and improving fairness. It includes metrics calculation, such as disparate impact, equalized odds, and equal opportunity. The library contains the following mitigation techniques: Exponentiated Gradient, Grid Search, Threshold Optimizer, and Adversarial Fairness, among others.

This work includes some of the metrics presented in Amazon SageMaker Clarify (HARDT et al., 2021). The software provides capabilities to detect live changes in data and model behavior, with a focus on the entire machine learning lifecycle, and integrates data-level bias detection and a model monitor for post-deployment monitoring. The metrics implemented in the software include: Class Imbalance, Kullback-Leibler Divergence, Jensen-Shannon Divergence, Kolmogorov-Smirnov, Disparate Impact, Specificity Difference, and Recall Difference, among others. These tools are presented to contextualize how metrics are currently available in the literature, since this work does not aims at proposing a new tool.

3.2 Research Papers

By treating bias detection as a step before deploying the model, some work has focused on identifying and mitigating bias in the data without the model predictions. The selected works for this analysis were chosen for their direct relevance to the proposed research questions and existing literature on analyzing bias metricas, fairness outcomes, or data-level properties in machine learning pipelines.

The work in Mandhala et al. (2022) proposes an algorithm to detect and mitigate bias in datasets, using data-level bias metrics, referred to as Pre-training bias metrics (Class Imbalance, KL Divergence, Sample Disparity, and Kolmogorov Smirnov). A weight is attributed to each pre-training metric, and the features with the maximum combined weight are identified as the protected attributes. The generated model then learns unbiased data by maximizing performance while estimating protected features, thereby reducing disparity across the protected variables identified by the weights. Similarly, in evaluating class imbalance, Larrazabal et al. (2020) examines the influence of gender imbalance in a medical imaging dataset. The work trains a deep learning-based computer-aided diagnosis (CAD) system on a gender-imbalanced dataset and inspects that the performance for minority gender groups is significantly affected. The work found that even a 25%/75% gender imbalance ratio in the training set lowers the average performance of the minority gender group compared to a balanced dataset. In the conclusion, the authors highlight the crucial need to use diverse, balanced datasets to achieve optimal performance across genders.

Furthermore, by analyzing the impact of bias before model deployment, Seth and Pai (2024) investigates the effects of two popular pre-training methods: Masked Image Modeling (MIM) and Self-Supervised Learning (SSL) on skin lesion classification models. The work found that pre-training the models can improve model performance but might introduce performance concerns for different skin tones. Including diverse skin tones improves model performance and fairness for all models, and MIM led to more substantial fairness improvements, while SSL offered greater performance gains, rendering a choice of fine-tuning approaches that can affect the performance-fairness tradeoff.

On the other side, by looking at ways to evaluate bias on deployed models using “post-training” metrics, in Openja, Laberge and Khomh (2024), the authors

propose an approach for systematically identifying bias-inducing features by using two swapping functions (single feature swapping and double feature swapping) and evaluating the impact of the swapping using four divergence measures: Hellinger distance, Jensen-Shannon divergence, total variation distance, and Wasserstein distance, on the models trained with and without the swapping functions. The work concludes that this framework can help domain experts make informed decisions to systematically identify bias-inducing features, visualize the causes of bias, and prioritize fixes. Similarly, Fabuyi (2024) investigates the impact of synthetic data as a tool to mitigate bias in model training. The authors used Generative Adversarial Networks (GANs) and validated that the generated data aligned with the original datasets across demographic features, using the Kolmogorov-Smirnov and KL Divergence tests. The work concluded, through an evaluation of bias using Demographic Parity and Equality of Opportunity, that models trained on synthetic data showed significant improvements in fairness without compromising predictive accuracy. Oliveira et al. (2024) discusses four standard fairness metrics: statistical parity difference, disparate impact, average odds difference, and equal opportunity difference. The study evaluates the performance and fairness of two machine learning algorithms across several datasets linked to credit analysis and finds that these models can achieve promising performance levels without compromising fairness in their predictions, suggesting that fairness considerations should be incorporated into credit analysis pipelines.

Other works were also dedicated to evaluating and measuring bias by looking at the models' implementation and its effects, such as Campbell, Bose and Masino (2024), where the authors present a case study on bias in electronic health records (EHR), training four machine learning models, and evaluating how the models perform for minority subgroups (African Americans, Medicaid patients) compared to the overall population. The work concludes that bias can manifest in complex ways, and understanding the context of the health condition is crucial when assessing model performance and bias. Similarly, Ladhak et al. (2023) examines how sociocultural and other types of biases translate into downstream tasks, such as text summarization, leading to hallucinations (factually incorrect information). The work concludes that more abstractive summarization models allow these biases to propagate more directly, leading to more hallucinations, and also analyzes how changes to the adaptation method and fine-tuning can affect biases. It shows that while it does

reduce the rate of hallucinations, the type of bias identified is still linked to name-nationality bias. In Cronjé et al. (2023), the article investigates whether widely used type 2 diabetes risk prediction models demonstrate racial bias between non-Hispanic whites and non-Hispanic blacks. The study finds that all models were miscalibrated with regard to race, overestimating for non-Hispanic whites and underestimating risk for non-Hispanic blacks, resulting in a larger proportion of non-Hispanic whites being prioritized for preventive interventions. The comparison was done by race and year-specific average predicted by the models, and compared with the incidences of type 2 diabetes extracted from the US Diabetes Surveillance System.

The work by Júnior et al. (2022) analyzed the COMPAS dataset using various machine learning techniques, with a focus on fairness with respect to the race attribute. The key findings of this work include detecting racial bias against the Black population, with a higher recidivism rate compared to the white population, and the finding that removing the race attribute did not impact the model performance, suggesting that the model was using other attributes as a proxy for race. Conversely, balancing the data improved statistical parity in the model's predictions across racial groups. In Gichoya et al. (2022), the authors evaluate AI models' ability to recognize a patient's racial identity from medical imagery across multiple modalities: x-ray, CT, and mammography. The work found that the high performance was not due to proxies such as body mass, breast density, or disease distribution. It was consistent across all anatomical regions and frequency spectra, suggesting that it is not linked to specific image features. Similarly, in Noseworthy et al. (2020), a convolutional neural network (CNN) trained to detect low left ventricular ejection fraction (LVEF) was able to distinguish racial/ethnic subgroups based on electrocardiogram features, indicating that these characteristics can vary by race. While not affecting the model's ability to detect LVEF, the authors recommended maintaining diverse datasets and conducting external validation to ensure responsible use of AI in medicine and to avoid perpetuating racial/ethnic disparities.

In Sjoding et al. (2020), the authors discuss racial bias in pulse oximetry measurements with a widely used medical technology to measure oxygen saturation in patients, and find that black patients had nearly three times the frequency of "occult hypoxemia" (low arterial oxygen saturation, despite a regular pulse), when compared to white patients. The authors emphasize the implications of these findings, especially in the scenario of the COVID-19 pandemic, as reliance on pulse oximetry may

place Black patients at increased risk of hypoxemia. Similarly, Obermeyer et al. (2019) dissects racial bias in a widely used algorithm that predicts the health care needs to target patients for additional care management, and verified that the algorithm was in fact predicting future health costs rather than health needs directly, and due to unequal access to care, Black patients generate lower costs than White patients with the same needs, leading to the algorithm to underestimate the health needs of black patients. Finally, Drozda et al. (2015) examined the performance of dosing algorithms in African Americans and found that the algorithms significantly over- or under-estimated warfarin doses in people with specific genotypes.

In Luccioni et al. (2023), a new method is proposed for exploring social biases in Text-to-Image systems, involving characterizing the variation on generated images by enumerating gender and ethnicity markers in the prompts, and comparing to the variation engendered by spanning different professions (with different gender assumptions for professions like clerk, singer, cleaner, among others). The work highlights that while the outputs show correlations with US labor demographics, they are also consistently under-representing marginalized identities to different extents. Similarly, in Park et al. (2021), the authors evaluate methods to reduce bias in models for predicting postpartum depression and postpartum mental health service utilization. In the key findings of this work, the authors found that white women were more likely to be diagnosed with PPD and use mental health services when compared to Black women. The work applied post-training metrics such as disparate impact and equal opportunity difference, and reduced bias using a reweighting technique without compromising model accuracy. These results were more effective than simply removing the race attribute or using the prejudice remover method.

Finally, in the topic of bias detection, it is equally important to provide data and tools for the models to be evaluated on. For this purpose, in Daneshjou et al. (2022), the authors create a diverse, curated clinical image dataset called Diverse Dermatology Images. The proposed dataset assesses potential biases in algorithm performance by containing diverse skin tones and pathologically confirmed lesions. The work also highlights that dermatologist consensus labeling can lead to noisy predictions, amplifying concerns about bias against diverse skin tones. Similarly, Jesus et al. (2022) presents a Bank Account Fraud dataset suite, a set of datasets for evaluating machine learning bias, generated from a real-world bank account opening fraud detection dataset. The work demonstrates how models trained on the base

	Stage of Bias Detection	Domain	Data Type	Bias Mitigation Strategy
Mandhala et al. (2022)	Pre-Training	General	Tabular	Yes
Larrazabal et al. (2020)	Pre-Training	Health	Image	No
Seth and Pai (2024)	Pre-training	Health	Image	Yes
Openja, Laberge and Khomh (2024)	Post-Training	General	Tabular	Yes
Fabuyi (2024)	Post-Training	General	Tabular	Yes
Oliveira et al. (2024)	Post-Training	Finance	Tabular	No
Campbell, Bose and Masino (2024)	Post-Training	Health	Tabular (EHR)	No
Ladhak et al. (2023)	Post-Training	NLP	Text	Yes
Cronjé et al. (2023)	Post-Training	Health	Tabular	No
Júnior et al. (2022)	Post-Training	General	Tabular	No
Gichoya et al. (2022)	Post-Training	Health	Image	No
Noseworthy et al. (2020)	Post-Training	Health	Image/Signal	No
Sjoding et al. (2020)	Post-Training	Health	Hardware/Signal	No
Obermeyer et al. (2019)	Post-Training	Health	Tabular	No
Drozda et al. (2015)	Post-Training	Health	Tabular	No
Luccioni et al. (2023)	Post-Training	Generative AI	Image	No
Park et al. (2021)	Post-Training	Health	Tabular	Yes
Daneshjou et al. (2022)	Not Applicable	Health	Image	No
Jesus et al. (2022)	Not Applicable	Finance	Tabular	No
This work	Both	Multiple Domains	Tabular	No

Table 3.1 – Summary of the analyzed works

dataset and its variants exhibit different fairness-accuracy tradeoffs, highlighting the importance of evaluating in dynamic environments.

Taken together, the works summarized in Table 3.1 show that, although fairness has been studied from several complementary perspectives, the literature still leaves open an important question: whether data-level bias metrics computed before model training can serve as reliable indicators of downstream unfairness across different datasets and learning algorithms. Our work aims to address this gap by systematically evaluating such metrics across 11 datasets from distinct domains, and by comparing their behavior with model-level bias measures and F1-score outcomes. In doing so, we provide empirical evidence on the importance of data-level signals and when they are informative, and when it fails to capture the fairness implications of the training data.

4 METHODOLOGY

This chapter outlines the systematic approach employed to investigate the impact of data-level bias on machine learning models. The methodology is structured into several key phases, from data acquisition and preparation to model training and evaluation. The primary objective was to develop a reproducible, transparent workflow, primarily implemented in Python, leveraging the *scikit-learn* package (PEDREGOSA et al., 2011).

4.1 Data Collection

The primary focus of this work is to provide a comparative analysis of the impact of the data-level bias on model performance for datasets originating from various domains. The complete list of datasets used in this work, along with the number of instances, sensitive attributes, analyzed attributes, and target variables for the prediction tasks, is presented in Table 4.1, comprising a total of 11 datasets with varying sizes.The datasets used in this study were selected to provide a diverse and representative evaluation across different domains. The main selection criteria were public availability, the presence of protected attributes, and sufficient variety in application context. In addition, some widely used datasets were included to support comparability with previous fairness studies and to ground the empirical analysis in familiar benchmarks. The number of datasets was chosen to provide a qualitative analysis that covers different scenarios while still allowing manual analysis. The list of datasets analyzed includes:

- **Predict Students' Dropout and Academic Success (PSDAS):**Predicts the academic outcome of a student, specifically whether they will drop out or graduate. The original task is to predict three categories, but in this work we simplify to predict whether a student will graduate or not (REALINHO VALENTIM; BAPTISTA, 2021).
- **Employee Dataset (EMP):** Contains information about employees in a company, HR, and demographic information. The classification task is to predict whether an employee will leave the company (ELMETWALLY, 2021).
- **COMPAS (COMPAS):** Well-known in the literature, this is a dataset used

in algorithms to score the likelihood of reoffending for a criminal. In this work, the task is to predict criminal recidivism (whether a defendant will reoffend) (LARSON et al., 2016).

- **Adult Dataset (ADULT):** Census data, the task is to predict whether an individual's annual income exceeds 50K/yr (BECKER; KOHAVI, 1996).
- **Bank Marketing (BANK):** Data extracted from marketing campaigns of a Portuguese banking institution. The task is to predict whether the client will subscribe to a term deposit (MORO S.; CORTEZ, 2014).
- **Heart Dataset (HEART):** Hospital data from patients (clinical data), the target is the presence of heart disease (JANOSI ANDRAS; DETRANO, 1989).
- **Indian Liver Patient Dataset (ILPD):** This dataset contains age, gender and clinical information about patients from Andhra Pradesh, India. The target variable is whether the patient has liver disease (RAMANA; VENKATESWARLU, 2022).
- **AIDS Clinical Trials Group Study 175 (ACTG175):** This dataset is a collection of healthcare statistics and categorical information about patients diagnosed with AIDS; the target variable is whether the patient died in the time window (Kaggle, 1996).
- **Intersectional Bias Dataset (IBD):** This is an artificial dataset created to train a classifier in predicting a diagnosis (of schizophrenia or depression), and assess the presence of intersectional bias (MASLEJ, 2022).
- **Diabetes Dataset (DIAB):** The dataset contains records from patients collected by an automatic electronic recording device and paper records, with clinical measurements. The target is for the patient to have no diabetes or pre-diabetes/diabetes (MORO, 2014).
- **Glioma Dataset (GLIOMA):** In this dataset, each instance is the record of a patient with brain glioma. Each record contains 20 molecular features and 3 demographic features. The classification task is to determine the glioma grade (LGG (Lower-Grade Glioma) or GBM (Glioblastoma Multiforme)) (TASCI; CAMPHAUSEN, 2022).

All datasets were pre-processed as described in Section 4.2. The processed data is publicly available on the project's GitHub[1].

[1]https://github.com/mlab-inf-ufrgs/fairnessinsight/tree/main/datasets

Dataset Name	Protected Attribute	Advantaged Class	Target	# of instances	# of columns
PSDAS	Gender	Male	DropsOut	4424	37
EMP	Gender	Male	LeaveOrNot	2764	14
COMPAS	sex	Male	WillRecid	11757	33
ADULT	sex	Male	Income	32537	108
BANK	marital	Married	WillSubscribe	45211	50
HEART	sex	Male	HasDisease	302	14
ILPD	Sex	Male	HasDisease	570	11
ACTG175	homo	NotHomo	Died	2139	24
IBD	Sex	Male	Diagnosis	11000	24
DIAB	Sex	Male	Diabetes_binary	229474	22
GLIOMA	Race	White	Grade	838	24

Table 4.1 – Relation of datasets used in this work. The target attribute is related to the task described in Section 4.1.

4.2 Data Preparation

Before analysis, all datasets were standardized to ensure structural compatibility with the supervised learning algorithms and to support a consistent automated pipeline. This process involved the following tasks:

- **Data Cleaning:** Basic cleaning operations were performed to handle inconsistencies or missing values, ensuring the data quality fed into the models.
- **Handling Categorical Features:** The protected attributes and target variables were mapped to binary values. For instance, a 'Gender' attribute with values 'Male' and 'Female' would be converted to 1 and 0, respectively. All favorable outcomes and privileged groups were assigned a value of 1, while unfavorable outcomes and unprivileged groups were assigned a value of 0. This binary encoding simplifies the analysis of bias and model performance.
- **Target Mapping:** As the supervised learning algorithms used in this work and the metrics used to evaluate bias require binary classification, all prediction tasks from the datasets were mapped to binary classification, with positive outcome mapping to 1 and negative mapping to 0.
- **Timestamp Features:** Any timestamp features were converted to numerical representations, to avoid one-hot encoding of high cardinality features.
- **One-Hot Encoding:** Categorical features with more than two categories were transformed using one-hot encoding, creating binary columns for each category. This approach prevents the model from assuming an ordinal relationship between categories.
- **Feature Scaling:** Numerical features were scaled using z-score normalization

to ensure that all features contributed equally to the model training process.

This preprocessing step is crucial as most machine learning models, including those used in this thesis, require numerical input. With this standardization, we ensure that the datasets are in a suitable format for subsequent analysis.

4.3 Bias Manipulation

Two variations of each original dataset were created using a resampling methodology to systematically study the effect of class imbalance related to protected attributes. The idea is that by varying the data in each dataset to increase or decrease the data-level bias metrics, we can correlate the performance with the level of bias in each generated model. The two variations are:

- **Equally Balanced:** The dataset was resampled to have an equal number of instances for each class of the protected attribute (e.g., an equal number of male and female individuals) for each outcome (the target of the prediction). This would represent an "ideal" scenario in which both classes of the protected attribute are equally represented.
- **Highly Imbalanced:** The dataset was resampled to create a significant disparity between the classes of the protected attribute, simulating a scenario of severe representation bias. To generate this version, we removed 80% of the instances that satisfied the condition: either from the unprivileged group with a favorable outcome or from the privileged group with an unfavorable outcome. This would represent extreme cases of bias, where the disadvantaged group would be severely underrepresented in the data.

These manipulations are applied during the model training. For each experiment, we train the models with each of the **three** variations and test on each of the **three** variations, generating **nine** different experiments for dataset. This way, we aim to systematically study the effects of the bias introduced by the **data** (RUBACK; CARVALHO; AVILA, 2022), or the **Representation Bias** (SURESH; GUTTAG, 2019). The Figure 4.1 represents how the data is transformed to create these variations. The pseudocode for the algorithm used to generate the equally balanced variation is in Algorithm 1, and the pseudocode used for generating the highly biased variation is in Algorithm 2. We note that these algorithms are called for each experiment.

After splitting the data into N-folds, we call the function for the training and testing sets, depending on the experiment variation.

Algorithm 1 Function to generate Equally Balanced data version

1: Identify the protected and unprotected groups.
2: **for** each group (privileged/unprivileged) and outcome (favorable/unfavorable) **do**
3: Find all samples in that cohort.
4: **end for**
5: **for** each outcome (favorable/unfavorable) **do**
6: Find the minimum count between privileged and unprivileged groups.
7: **end for**
8: **for** each cohort **do**
9: Randomly sample the minimum count of rows.
10: **end for**
11: Concatenate all balanced cohorts into a new DataFrame.
12: Shuffle the new DataFrame.
13: **return** features (without target), target column.

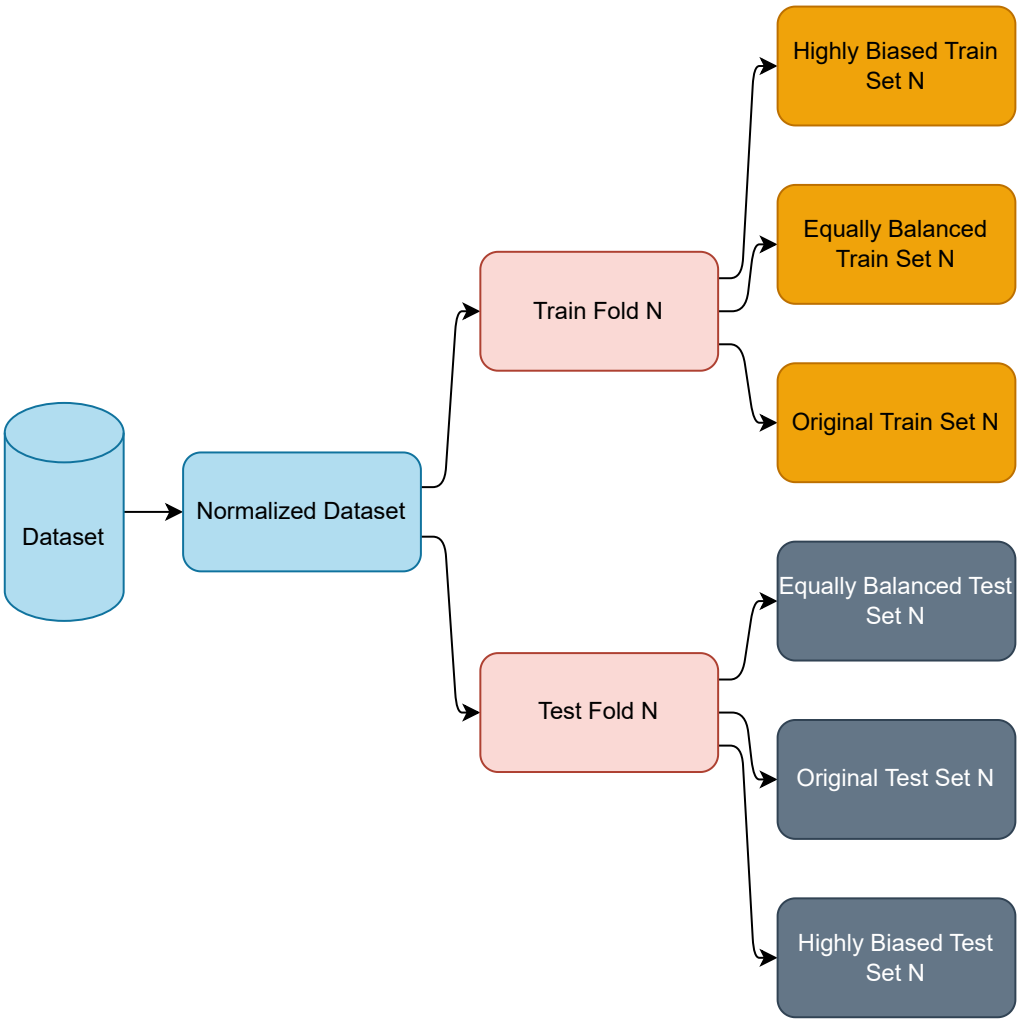
Algorithm 2 Function to generate Highly Biased data version

1: Identify the protected attribute and target column.
2: Find all indices where:
3: (disadvantaged group and positive outcome) OR (advantaged group and negative outcome)
4: Calculate 80% of these indices.
5: Randomly select and remove these indices from the DataFrame.
6: **return** features (without target), target column.

4.4 Hyperparameter Tuning

A hyperparameter tuning phase was conducted to ensure that model evaluations were based on the strongest possible versions of each algorithm. Using the `GridSearchCV` utility from the `scikit-learn` library (PEDREGOSA et al.,

Figure 4.1 – Datasets generated during the bias manipulation process



source: the Author

2011), an exhaustive search was performed to find the optimal combination of hyperparameters for the Decision Tree, Random Forest, and XGBoost models. The hyperparameters considered for each model are as follows:

- **Decision Tree:**
  - `max_depth`: [None, 5, 7, 10, 50, 100, 200]
  - `criterion`: ['gini', 'entropy']
- **Random Forest:**
  - `n_estimators`: [10, 20, 50, 100, 200, 600]
  - `max_depth`: [None, 5, 7, 10, 100]
  - `criterion`: ['gini', 'entropy']
- **XGBoost:**
  - `n_estimators`: [10, 20, 50, 100, 300, 500]
  - `max_depth`: [None, 5, 7, 10, 100, 200, 500, 700]
  - `learning_rate`: [0.001, 0.01, 0.1, 0.2, 0.3, 0.5]

The search was conducted for the original datasets, and the Grid Search goal was to maximize F1-Score. The best-performing set of parameters for each model and dataset was identified and recorded for use in the main experiment. It was empirically verified that using any of the 10-folds to tune the hyperparameters, and splitting the data with different random seeds, produced very similar output. Therefore, for hyperparameter tuning, only the first fold of the 10-fold cross-validation with a fixed random seed is used.

The process starts by creating an artificial stratification key by concatenating the target of the prediction with the protected attribute, ensuring the distribution is constant across both the train and test sets. We then call the stratified K-Fold to split the data into train-test sets using this stratification key. In the first train-test split of this K-Fold (K=10), we perform a grid search with 10-fold cross-validation to maximize F1-Score. This step ensures that comparisons between models are fair and not skewed by arbitrary or default parameter choices.

It is noted that this approach can suffer from data leakage, as the instances used for fine-tuning the hyperparameters may be included in the testing set for a different fold. This is an accepted risk, as the parameters did not deviate significantly when using different folds, and the computational cost would impact overall

performance of our experiments. The complete list of hyperparameters used in this work for each dataset variation and algorithm trained is in the Appendix 7.2.

4.5 Model Training and Performance Evaluation

The core of the experimental evaluation involved training four supervised learning algorithms: Logistic Regression, Decision Tree, Random Forest, and XGBoost. For each algorithm, we performed 10-fold cross-validation for each dataset variation. The models were trained on the training set and evaluated on the three hold-out test sets. The choice of algorithms was based on their prevalence in the literature and their explainability. The idea was to evaluate tree-based models from the simplest to the more robust, and a linear model to cover different algorithmic approaches to supervised learning.

Model performance was assessed using two standard classification metrics:

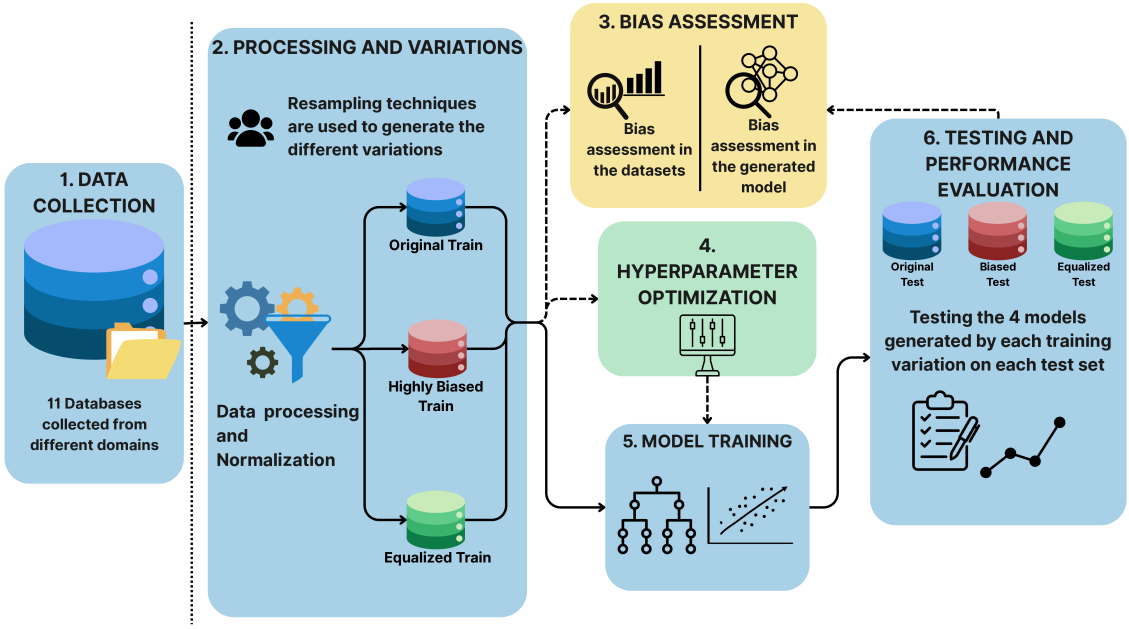
- **Accuracy:** The proportion of correctly classified instances, providing a general measure of model correctness.
- **F1-Score:** The harmonic mean of precision and recall, which is particularly useful for evaluating model performance on imbalanced datasets where accuracy alone can be misleading.

The final performance scores for each model on a given dataset were calculated by averaging the results from the 10-fold iterations.

4.6 Bias Evaluation

The evaluation of bias was conducted at two stages: before model training (pre-training, or **data-level bias**) and after (post-training, or **model-level bias**). The data-level bias metrics were calculated to quantify the bias inherent in the data itself. The post-training metrics were used to evaluate how this initial bias was reflected in the model's predictions, assessing fairness in terms of outcomes. Both data-level and model-level metrics were defined in Section 2.3. This dual approach enables a deeper understanding of the impact of biased data on the model's prediction and whether it is reflected in the generated model.

Figure 4.2 – High-level experimental setup. Data is collected, three variations are generated and used to evaluate data-level bias, a hyper-parameter tuning step is applied before completing the model training, then performance is evaluated on the testing sets, and model-level bias metrics are computed



source: the Author

4.7 Experimental Setup

The main experiment followed an automated pipeline to manage the training, evaluation, and analysis process for all datasets and their variations. A high-level diagram outlining the methodology adopted in this work is presented in Figure 4.2, where each step is defined in this section. The workflow for each experiment is as follows:

1. **Load Data and Parameters:** The process begins by loading the dataset and the corresponding optimal hyperparameters identified during the parameter tuning phase.
2. **Apply Data Variation Functions**: Depending on the experiment, we pass the functions described in Section 4.3 to perturb the train and test sets; these functions will be called prior to model training or model evaluation.
3. **Repetitive Training and Testing:** For statistical robustness, we use K-Fold cross-validation with 10 folds, to ensure lower bias relative to single train/test and variance. Each data point is guaranteed to be in the training set 90% of the folds. The models trained are Logistic Regression, Decision Tree, Random

Forest, and XGBoost.

4. **Metric Calculation:** In each repetition, both data-level bias metrics on the training data and model-level bias on the generated model were evaluated to extract performance metrics. Accuracy and F1-score are also reported for each algorithm.
5. **Result Aggregation and Saving:** The results from all repetitions, including metrics and model predictions, were aggregated. These results, along with visualizations of the data distributions and error analyses, were recorded for subsequent analysis.

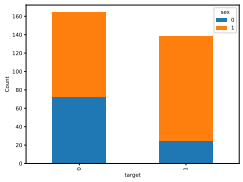
This setup ensures a systematic and comprehensive evaluation, generating all the necessary data for the results and discussion chapters.

5 RESULTS

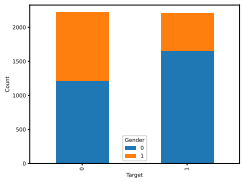
In this chapter, we structure the analysis of our experiments following three approaches: analyzing overall behavior, grouping datasets by similar patterns observed in the results, and, finally, through a set of research questions. In Section 5.1, we first introduce the dataset collection and discuss its data-level bias metrics, along with the F1-scores obtained on the original test set. This initial analysis serves as the foundation for identifying four distinct dataset groups, which are further examined in Section 5.2. Sections 5.2.1 through 5.2.4 then describe Group A as datasets in which HB training produces severe bias amplification, while the original and EB variations yield smaller F1-score gaps; Group B as datasets showing a clear separation among training variations; Group C as datasets that are comparatively less affected by these variations and do not exhibit substantial increases in F1-score gap values across conditions; and Group D as a single outlier dataset that differs substantially from the others.

Next, we analyze the four research questions presented on this work. First, RQ1 (Section 5.3) examines how data-level bias affects predictive performance, revealing that equalizing data distributions successfully reduces fairness gaps while maintaining overall model F1-scores, provided that the balancing process does not result in severe data scarcity. Second, RQ2 (Section 5.4) explores the correlation between data-level and model-level bias metrics, demonstrating that distributional metrics like KL Divergence and KS are far more reliable early-warning indicators of downstream model unfairness than simple Class Imbalance. Third, RQ3 (Section 5.5) investigates which algorithms are more robust against biased data, concluding that there is no consistent algorithmic hierarchy; complex ensemble methods (like Random Forest and XGBoost) are just as vulnerable to absorbing severe bias as simpler linear models like Logistic Regression. Finally, RQ4 (Section 5.6) analyzes the impact of bias on feature importance, proving that highly biased training data systematically forces all tested algorithms to elevate the protected attribute to their primary decision-making feature, a critical vulnerability that Equally Balanced (EB) training consistently mitigates. We note that while our discussion regarding performance estimates will be focused in F1-score, accuracy results are provided in the Appendix, Table 7.4. Moreover, information regarding train and test sizes for each dataset under all variations is provided in Table 7.1.

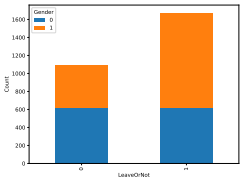
Figure 5.1 – Datasets distributions per target variable. The X-axis is the target variable (1 being the positive outcome), and the Y-axis is the number of instances, where orange is the privileged class (1) and blue is the unprivileged class (0).



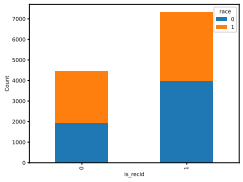
(a) Heart



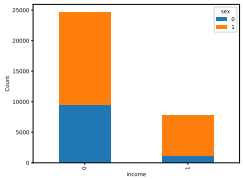
(b) Students' Dropout



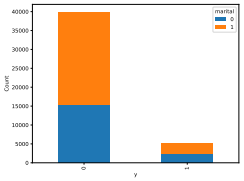
(c) Employee



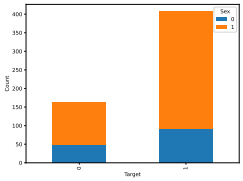
(d) COMPAS



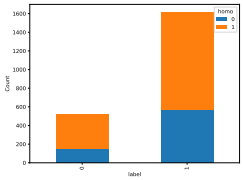
(e) Adult



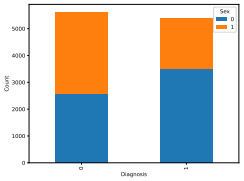
(f) Bank Marketing



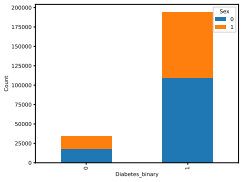
(g) Indian Liver Patient



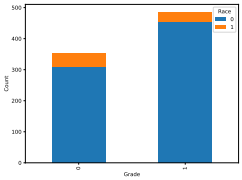
(h) AIDS Clinical Trials



(i) Intersectional Bias



(j) Diabetes



(k) Glioma Dataset

source: the Author

5.1 Overall Analysis

Figure 5.1 presents the class distributions for all 11 datasets across the target variable and the protected attribute. The datasets were selected to provide diversity in domain, size, and distributional characteristics, ranging from small clinical cohorts (Glioma, $n \approx 900$; Indian Liver Patient, $n \approx 600$) to large population surveys (Diabetes, $n > 250{,}000$; Adult, $n > 48{,}000$). This diversity is intentional: it allows us to assess whether the relationships between data-level bias metrics and downstream fairness outcomes are consistent across different data regimes, or whether they are can be influenced by dataset size, domain, or baseline distributional properties.

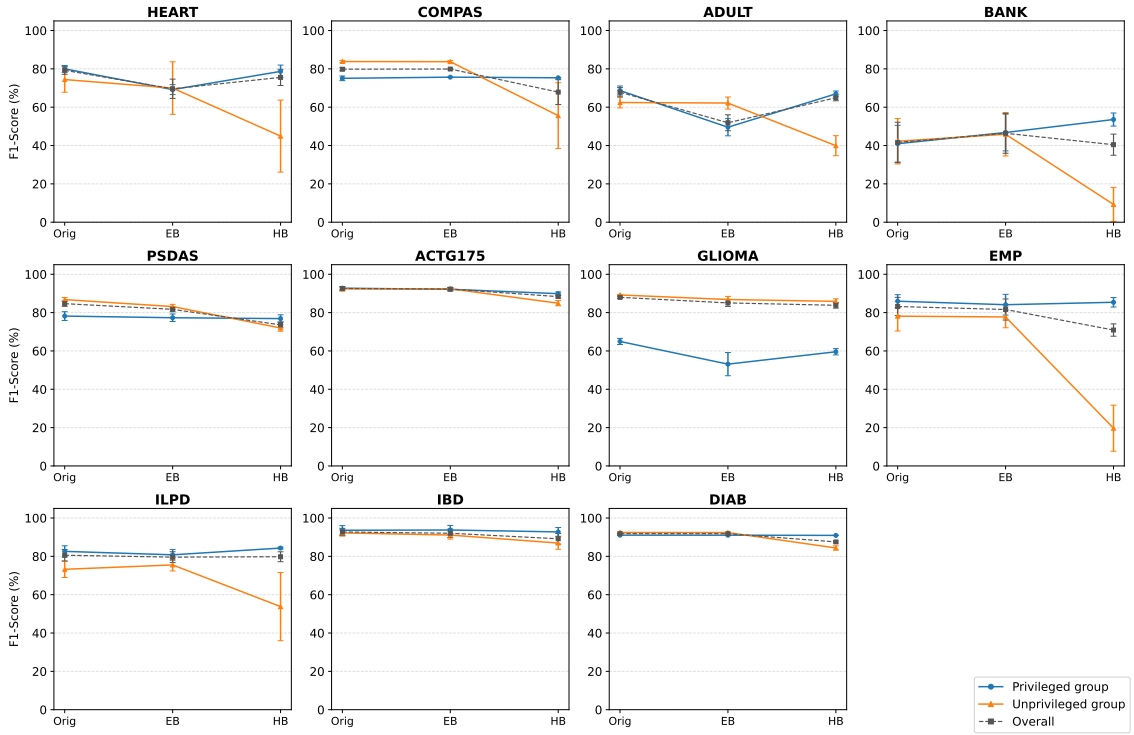
Table 5.1 – Pre-training dataset characteristics for the original (Orig) and highly biased (HB) training variations. Arrows in column headers denote the direction associated with increased bias.

Dataset	**Class Imb.** $\lvert\uparrow\rvert$		**KL Div.** $\uparrow$		**KS** $\uparrow$		**Dem. Disp.** $\lvert\uparrow\rvert$	
	Orig	HB	Orig	HB	Orig	HB	Orig	HB
HEART	0.364	0.280	0.208	2.238	0.303	0.806	0.265	0.754
COMPAS	-0.011	0.168	0.023	0.727	0.102	0.577	-0.109	0.600
ADULT	0.338	-0.007	0.141	1.941	0.196	0.664	0.238	0.726
BANK	0.204	-0.352	0.007	0.613	0.040	0.329	-0.092	0.605
PSDAS	-0.297	-0.336	0.104	0.597	0.227	0.510	-0.207	0.480
ACTG175	0.322	0.614	0.007	0.586	0.049	0.506	-0.060	0.575
GLIOMA	-0.823	-0.827	0.074	0.711	0.191	0.564	-0.063	0.221
EMP	0.106	0.218	0.072	1.364	0.187	0.749	0.194	0.761
ILPD	0.509	0.663	0.016	0.972	0.083	0.653	0.075	0.631
IBD	-0.102	-0.133	0.076	0.668	0.194	0.541	-0.192	0.538
DIAB	-0.122	0.367	0.002	0.449	0.024	0.416	-0.045	0.638

Table 5.1 reports the four data-level bias metrics computed for the original and highly biased variations of each dataset. The equally balanced variation is omitted from this table since all metrics were zero. For the original variation, KL Divergence and KS remain below 0.3 across almost all datasets (except for KS = 0.303 in Heart), and Demographic Disparity stays below 0.27 in absolute value, indicating that the original datasets exhibit limited distributional bias with respect to the protected attribute. Class Imbalance is the exception: ILPD reaches 0.509, and ACTG175 reaches 0.322, reflecting a skewed ratio between favorable and unfavorable instances in the original data. The HB variation consistently elevates KL Divergence, KS, and Demographic Disparity across all datasets. KL Divergence exceeds 0.50 in every case and surpasses 1.0 in Heart, Adult, Employee, and ILPD, while Class Imbalance behaves inconsistently, decreasing or changing sign in several datasets (Heart, Adult, Bank, PSDAS, IBD).

Figure 5.2 presents the F1-score trajectories for the privileged and unprivileged groups, alongside the overall F1-score, across the three training variations, with all models tested on the original distribution of the datasets. Across most datasets, the magnitude of the performance gap (*i.e.,* the area between the two groups' curves) remains narrow under original and equally balanced training and widens under highly biased training, with the privileged group's curve rising while the unprivileged group's curve falls. This divergence indicates that highly biased training does not merely degrade overall performance, but redistributes it unevenly between groups, systematically favoring the privileged group at the unprivileged

Figure 5.2 – F1-score by training data variation for privileged and unprivileged groups across datasets. Each panel shows the mean F1-score (± standard deviation) over four classifiers, evaluated on the original test distribution under three training conditions: original (Orig), equally balanced (EB), and highly biased (HB)



group's expense. A smaller subset of datasets shows comparatively narrow gaps across all three training conditions, suggesting lower sensitivity to distributional manipulation regardless of the training variation used. The Glioma dataset is an exception: the unprivileged group consistently outperforms the privileged group across all training variations.

Finally, we examine F1-score gaps for all datasets across all train/test variations. That is, in addition to evaluating models on the original test set, we also assess their performance when tested on the equally balanced (EB) and highly biased (HB) variations. Results are summarized in Table 5.2. Positive values indicate a higher F1-score for the privileged group, whereas negative values indicate the reverse. The notation *Train Original* $\rightarrow$ *HB*, for instance, means that the model was trained in the original data and applied to classify the HB variation of the dataset.

Observing the results across all datasets we note that, first, the test distribution is the single strongest determinant of gap severity. Testing on the HB variation yields a higher mean absolute gap regardless of training condition: 34.05 pp, compared to 12.81 pp when testing on the original distribution and 11.31 pp when

Table 5.2 – F1-score gaps (pp) between privileged and unprivileged groups across all datasets and train/test variations. Positive values indicate higher F1 for the privileged group; negative values indicate the reverse. Dataset groupings (A–D) follow the hierarchical clustering of F1-gap profiles.

		Train Original			**Train EB**			**Train HB**		
Group	**Dataset**	*→Orig*	*→EB*	*→HB*	*→Orig*	*→EB*	*→HB*	*→Orig*	*→EB*	*→HB*
A	EMP	+7.8	−1.8	+44.1	+6.4	−3.9	+45.9	+65.7	+57.8	+77.3
	BANK	−1.3	+0.2	+18.0	+0.9	+3.4	+25.2	+44.3	+50.4	+62.3
B	HEART	+5.6	−16.5	+58.9	−0.8	−24.3	+47.3	+33.8	+7.8	+78.6
	COMPAS	−8.8	−3.1	+36.9	−8.1	−2.0	+39.2	+19.8	+24.6	+48.3
	ADULT	+6.3	−10.1	+34.5	−12.6	−17.6	+10.4	+27.0	−2.6	+52.7
	ILPD	+9.4	+3.0	+52.4	+5.3	+0.6	+43.8	+30.5	+25.0	+61.6
C	PSDAS	−8.6	+1.5	+25.1	−5.8	+1.4	+19.8	+4.9	+9.6	+33.9
	ACTG175	+0.4	+1.7	+17.1	−0.2	+1.4	+18.1	+5.0	+7.2	+24.0
	IBD	+1.4	+5.0	+17.2	+2.6	+5.3	+14.8	+5.8	+8.1	+16.3
	DIAB	−1.4	−0.2	+24.9	−1.4	−0.2	+24.3	+6.6	+6.9	+21.6
D	GLIOMA†	−24.3	−19.9	−9.5	−33.7	−29.3	−19.0	−26.3	−20.9	+0.6

Colour bands: $|\text{gap}| < 10$, $10 \leq |\text{gap}| < 30$, $|\text{gap}| \geq 30$.

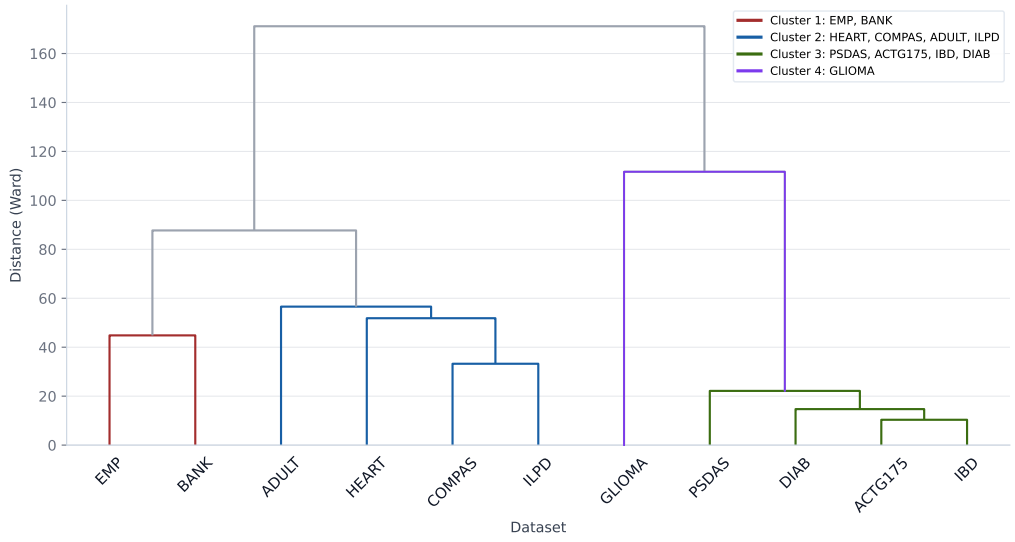
testing on the EB variation. This indicates that exposure to a biased population at deployment time degrades fairness substantially more than any property of the training data alone. Moreover, as expected, in several scenarios the performance gap increases as more bias is present in the training dataset. This is generally observed, for instance, comparing the testing in the original data across the training in the three variations (especially Original *vs* EB; and HB *vs* Original). However, looking at the mean absolute gaps, we note that these values are very close under Train-Original (14.45 pp) and Train-EB (14.39 pp), while Train-HB doubles this average value (29.33 pp).

To support the analysis of these results, we run a hierarchical clustering algorithm, using Euclidean distance and the Ward.D2 linkage method, on the F1-score gap profiles extracted from this table for each dataset. The goal is to identify patterns among datasets that can aid the discussion of results. This analysis yields four clusters, shown in Figure 5.3, which are indicated in the first column of Table 5.2. The following sections will explore and discuss the datasets per group.

5.2 Per-group Analysis

This section discusses the trends related to each group of datasets, according to Figure 5.3, as well as results associated to each dataset.

Figure 5.3 – Hierarchical clustering of datasets by F1-score gap profile, using Euclidean distance and Ward.D2 linkage method.



5.2.1 Group A: EMP and BANK

This cluster groups the two datasets exhibiting the most extreme bias amplification under HB training. Both datasets show low-to-negligible gaps under original and EB training (all values below |8| pp) but significant increases under HB, with Train-HB→Test-HB gaps exceeding +62 pp in both cases and Train-HB→Test-Orig reaching +44 to +66 pp — the two highest values observed for this combination in the entire study. This pattern distinguishes Group A from Group B not by the presence of bias transfer, but by its magnitude. Even the comparatively favorable Train-EB→Test-Orig condition fails to fully neutralize the effect, remaining moderate (BANK, +0.9) to more severe impact (EMP, +6.4) .

In the Employee dataset (EMP), the original dataset also had very low pre-training bias reported (less than 0.2 for all metrics), whereas in the biased version, all metrics exceeded 0.7 (except class imbalance), as shown in Table 5.1. Looking at the performance gaps between the three variations, we can see that the equally balanced version holds smaller gaps when compared to the other two variations (as seen in Table 5.2), and the overall F1-Score was also kept high under this variation.

For the Bank Marketing dataset (BANK), the central issue observed was that the distribution of the samples was very imbalanced for the unfavorable outcome, but was somewhat evenly distributed on the protected attribute, giving a class imbalance value of 0.20 for the original dataset, and all other metrics with very low values, below 0.1 (Table 5.1. The highly biased version was also challenging

Table 5.3 – Model-level bias metrics across test and train variations for the EMP and BANK datasets (Group A).

Dataset	**Test Variation**	**Train Variation**	DPPL	DI	DCA	DCR
EMP	Test EB	Train EB	0.017	0.978	-0.022	0.073
		Train HB	0.809	0.095	-inf	inf
		Train Original	0.191	0.768	-0.232	1.419
	Test HB	Train EB	0.358	0.585	0.736	-1.050
		Train HB	0.930	0.049	-inf	inf
		Train Original	0.534	0.421	0.557	-0.188
	Test Original	Train EB	0.106	0.862	0.138	-0.137
		Train HB	0.846	0.084	-inf	inf
		Train Original	0.273	0.679	-0.072	1.058
BANK	Test EB	Train EB	0.001	0.983	-0.090	0.001
		Train HB	0.148	0.058	-inf	0.155
		Train Original	-0.009	1.151	0.286	-0.009
	Test HB	Train EB	0.121	0.247	1.529	-0.243
		Train HB	0.276	0.015	-inf	-0.077
		Train Original	0.090	0.273	2.096	-0.271
	Test Original	Train EB	-0.013	1.180	-0.369	0.029
		Train HB	0.128	0.067	-inf	0.176
		Train Original	-0.019	1.396	-0.092	0.022

to generate, with the KS value not exceeding 0.32, and the KL Divergence and demographic disparity exceeding 0.6. In the F1-Scores obtained on this dataset, we see that the gap between the two groups of the protected attribute was considerably low in the original and EB variation, and very high in the HB one, as shown in Table 5.2. The overall performance on this dataset, however, showed an F1-score below 50%, while maintaining an accuracy of over 80% for all classes, corroborating the presence of a class imbalance problem.

The model-level bias metrics for these datasets are reported on Table 5.3. This analysis reveals a consistent and substantial contrast between training conditions for both EMP and BANK. Under Train-HB, DPPL increases sharply relative to Train-EB across all three test variations — most notably for Test-Original, where DPPL rises from 0.106 to 0.846 in EMP (an almost eightfold increase) and from 0.013 to 0.128 in BANK. This pattern is mirrored by DI, which falls to severely unfair levels under Train-HB (EMP: 0.084; BANK: 0.067), well below the 0.80–1.25 fairness range, while remaining close to or above 1.0 under Train-Original and Train-EB. DCA and DCR, by contrast, exhibit systematic numerical instability under Train-HB: DCA collapses to $-\infty$ and DCR to ∞ in every Test-Original and Test-EB combination for both datasets, reflecting near-zero conditional acceptance or rejection rates for one of the groups.

5.2.2 Group B: HEART, COMPAS, ADULT, and ILPD

This group is characterized by a clear and consistent contrast between EB and HB training: Train-EB→Test-Orig gaps remain small (between −12.6 and +5.3 pp), while Train-HB→Test-Orig gaps are consistently severe (+19.8 to +33.8 pp). This is the canonical bias-transfer signature: EB training tends to neutralize the gap, while HB training reliably reintroduces it, making Group B the clearest empirical demonstration that the choice of training distribution, not the algorithm, drives the fairness outcome on a balanced deployment population. The four datasets reach this outcome through different routes, detailed below.

In HEART, Train-EB→Test-Orig gaps are below 1%, near-ideal fairness, suggesting the model trained on balanced data learned a representation that transfers well to the original population. In contrast, Train-HB produces gaps of 33% on the original test set and 78% on the highly biased one. Model-level metrics, shown in Table 5.4, suggests that the best fairness in terms of DI is achieved under Train-Original→Test-EB, while the HB version shows strong bias against the unprivileged group. Model-level metrics, shown in Table 5.4, confirm this pattern: Train-EB→Test-Orig yields DPPL =0.129 and DI =0.686, indicating residual but moderate unfairness even in the best F1-gap scenario – a reminder that F1-gap and DPPL measure complementary aspects of fairness rather than the same quantity. Train-HB→Test-Orig reaches DPPL =0.661 and DI =0.144, the most extreme values in the group for that deployment scenario.

COMPAS is notable for not reproducing the strong fairness issues commonly associated with this dataset in the literature: data-level metrics on the original version are low (Table 5.1), and the resulting gap under original training is among the smallest in the group (36%) — consistent with those low metric values. The biased version, in contrast, pushes all metrics above 0.5 except Class Imbalance. One distinguishing feature of COMPAS is a pronounced disparity in training set size across variations: the original training set averaged 10,581.3 samples, against 9,469.8 for EB and 5,904.9 for HB (Table 7.1), a consequence of the dataset's large original size and balanced distribution across both the target and protected attribute.

ADULT is the partial exception within the group: EB training overcompensates toward the unprivileged group rather than achieving full neutrality, producing a moderate negative gap under Train-EB→Test-Orig (−12.6 pp) — the only such

inversion in Group B. This behavior is likely explained by the target variable's distribution, which is predominantly negative for both groups (Figure 5.1e), making the underprivileged class disproportionately sensitive to rebalancing. Original data-level metrics remain below the 0.35 threshold for all four measures (Table 5.1), in line with the comparatively modest original-training gap. An interesting point regarding model-level metrics for COMPAS is the moderation of DPPL/DI under non-HB conditions (reflecting the low original data-level metrics) and the extreme DCA/DCR values observed under HB training. This demonstrates that a dataset with very low data bias can still yield severe conditional errors when subjected to aggressive manipulation of the training distribution.

ILPD mirrors BANK's class structure in reverse: most instances fall in the favorable outcome class, yielding high F1-scores but lower accuracy (60–70%). Despite a comparatively high Class Imbalance, the remaining data-level metrics stay below 0.1 in the original version, and EB training achieves the smallest gaps of all four test scenarios (Table 5.2, ILPD with train EB under Orig and EB test) using only ≈ 252 training instances — fewer than both HB (365.1) and the original (513.0) — without any loss in F1-score or accuracy. This makes ILPD the clearest case in the group where fairness improves with less data, not despite it. Model-level metrics (Table 5.4) corroborate this: Train-EB→Test-Orig yields DPPL =0.004 and DI =1.000, the closest values to perfect fairness observed in Group B. Train-HB→Test-Orig, by contrast, reaches DPPL =0.543 and DI =0.450, consistent with the severe bias-transfer pattern of the group. , and `NaN` values across most combinations.

Table 5.4 confirms, at the level of individual model-level metrics, the bias-transfer signature already observed in the F1-gap analysis for Group B. Train-HB→Test-Original consistently produces severe unfairness across all four datasets, with DPPL ranging from 0.177 (ADULT) to 0.661 (HEART) and DI falling below 0.46 in every case. DPPL and DI do not always agree on the threshold for fairness, however: ADULT's Train-EB→Test-Original condition shows the lowest DPPL in the group (0.034) yet still falls outside the 0.8–1.25 DI range (0.716), reflecting the overcompensation toward the unprivileged group already noted in the F1-gap discussion.

Table 5.4 – Model-level bias metrics across test and train variations for the HEART, COMPAS, ADULT, and ILPD datasets (Group B).

Dataset	**Test Variation**	**Train Variation**	DPPL	DI	DCA	DCR
HEART	Test EB	Train EB	-0.076	inf	inf	-0.099
		Train HB	0.557	0.176	-inf	1.897
		Train Original	0.066	0.955	-0.221	0.113
	Test HB	Train EB	0.381	0.246	-inf	-0.717
		Train HB	0.853	0.046	-inf	inf
		Train Original	0.569	0.156	-inf	-0.472
	Test Original	Train EB	0.129	0.686	0.314	-0.287
		Train HB	0.661	0.144	-inf	1.346
		Train Original	0.285	0.458	-0.204	-0.034
COMPAS	Test EB	Train EB	0.010	0.989	-0.009	0.515
		Train HB	0.600	0.367	-1.601	6.808
		Train Original	-0.087	1.121	0.091	-0.793
	Test HB	Train EB	0.151	0.832	0.577	-0.917
		Train HB	0.717	0.264	-0.574	5.551
		Train Original	0.067	0.921	0.673	-1.953
	Test Original	Train EB	-0.016	1.020	-0.107	0.663
		Train HB	0.581	0.382	-1.731	6.777
		Train Original	-0.113	1.159	-0.011	-0.588
ADULT	Test EB	Train EB	-0.032	1.689	0.908	-0.032
		Train HB	0.392	0.073	-3.483	0.627
		Train Original	0.070	0.538	-0.648	0.080
	Test HB	Train EB	0.204	0.153	2.242	-0.601
		Train HB	0.736	0.011	-2.131	0.246
		Train Original	0.429	0.073	0.793	-0.428
	Test Original	Train EB	0.034	0.716	1.321	-0.184
		Train HB	0.503	0.058	-3.169	0.571
		Train Original	0.177	0.309	-0.198	-0.033
ILPD	Test EB	Train EB	-0.014	1.034	0.011	-0.167
		Train HB	0.544	0.450	-1.776	inf
		Train Original	0.066	0.932	-0.083	NaN
	Test HB	Train EB	0.052	0.941	0.748	-inf
		Train HB	0.593	0.400	-inf	inf
		Train Original	0.110	0.882	0.667	-inf
	Test Original	Train EB	0.004	1.000	0.086	-0.755
		Train HB	0.543	0.450	-1.691	inf
		Train Original	0.081	0.912	-0.012	-inf

5.2.3 Group C: PSDAS, ACTG175, IBD, and DIAB

Group C comprises datasets with consistently low sensitivity to training variation. Across all nine train/test combinations, performance gaps remain below 25 pp, and most are below 10 pp (Table 5.2). Even Train-HB→Test-Orig, which produces large gaps in Groups A and B, results in only small differences (+4.9 to +6.6 pp) in three of the four datasets. PSDAS presents the largest gap in the group, reaching +25.1 pp under Train-HB→Test-HB. Although all four datasets show low sensitivity to training variation, the reasons for this behavior appear to differ across datasets.

PSDAS shows the largest variation within the group. Data-level metrics are low in the original dataset (around the 0.3 threshold) and increase to approximately 0.5 for all metrics except Class Imbalance in the biased version (Table 5.1). Even so, EB training produces smaller gaps than those observed in the original setting, reducing the Train-HB→Test-HB gap to about 20% while using fewer training samples (2800.8 under EB versus 3981.6 in the original dataset, as shown in Table 7.1) and maintaining overall F1-score.

ACTG175 follows a pattern similar to ILPD in Group B, with a majority-favorable and majority-privileged distribution resulting in a moderate Class Imbalance (0.322). However, all other data-level metrics remain below 0.1, even in the biased setting (Table 5.1). Unlike ILPD, performance gaps remain small across all train/test combinations. The smallest gaps are observed under Original and EB training, with no meaningful reduction in predictive performance (Table 5.2).

IBD is an artificial dataset designed to study intersectional bias, that is, the combined effect of multiple sensitive attributes. In this work, however, only the **sex** attribute is considered, and its distribution remains stable across the dataset. As in the other datasets in this group, the original data-level metrics remain below the 0.3 threshold, while the biased variation increases all metrics except Class Imbalance to values above 0.5 (Table 5.1). IBD is the only dataset in the group for which equalization does not provide a clear benefit. Performance gaps remain largely unchanged after balancing, with no relevant improvement in predictive performance (Table 5.2). At the same time, training set sizes decrease considerably under both EB (8019.0) and HB (5188.0) compared with the original dataset (9900.0), as shown in Table 7.1. This result suggests that balancing a single sensitive attribute may

not be enough when bias is related to interactions among multiple attributes.

DIAB is the largest dataset in the study and presents a strong imbalance in the target variable toward the positive class (Figure 5.1j). Nevertheless, all data-level metrics remain low, and even the biased variation stays below the 0.5 threshold (Table 5.1). Original training already produces small performance gaps, and differences increase only under HB testing conditions (Table 5.2). Equalization has little effect on F1-score or accuracy, except for a change in accuracy when highly biased training is combined with Original or EB testing (see Appendix for complete results). With average training set sizes of 206,526.6 (Original), 181,366.2 (EB), and 114,998.8 (HB) samples, DIAB suggests that large sample sizes may reduce sensitivity to distributional changes, regardless of the dataset characteristics.

The model-level bias metrics shown in Table 5.5 are generally consistent with the low performance gaps observed in Group C, but they also reveal some differences that are not visible in the F1-gap analysis. ACTG175 and DIAB show some of the lowest bias values observed in the study under Train-EB and Train-Original evaluated on the original distribution. ACTG175 achieves DPPL values between -0.056 and -0.039 and DI values between 1.050 and 1.074, while DIAB reaches near-neutral values under Train-Original→Test-Original (DPPL = 0.001, DI = 0.999). IBD, in contrast, consistently favors the unprivileged group when evaluated on the original distribution, regardless of the training variation. DPPL ranges from -0.054 to -0.206, and DI remains above 1.1 in all cases. This pattern suggests that balancing only one attribute may not be sufficient to remove disparities in a dataset designed around intersectional bias. PSDAS shows a similar pattern under Train-EB and Train-Original→Test-Original, with DPPL values reaching -0.248 and DI values up to 1.655. These results indicate that small F1-score gaps do not necessarily imply neutral behavior according to model-level fairness metrics.

Finally, DCA and DCR remain relatively stable across most Group C configurations, unlike what was observed in Group B. Only a few `inf` values appear under Train-HB in ACTG175 and DIAB, suggesting that these datasets are less affected by the distribution changes introduced in this study.

Table 5.5 – Model-level bias metrics across test and train variations for the PSDAS, ACTG175, IBD, and DIAB datasets (Group C).

Dataset	**Test Variation**	**Train Variation**	DPPL	DI	DCA	DCR
PSDAS	Test EB	Train EB	-0.035	1.114	0.092	-0.059
		Train HB	0.255	0.513	-0.668	0.476
		Train Original	-0.107	1.286	0.202	-0.226
	Test HB	Train EB	0.304	0.492	0.471	-0.418
		Train HB	0.585	0.239	-0.263	0.268
		Train Original	0.244	0.624	0.599	-0.519
	Test Original	Train EB	-0.177	1.530	-0.082	0.103
		Train HB	0.140	0.735	-0.853	0.679
		Train Original	-0.248	1.655	0.006	-0.102
ACTG175	Test EB	Train EB	-0.011	1.015	0.013	-0.093
		Train HB	0.256	0.719	-0.348	2.083
		Train Original	-0.023	1.031	0.028	-0.118
	Test HB	Train EB	0.262	0.720	0.315	-0.873
		Train HB	0.536	0.448	-0.124	inf
		Train Original	0.266	0.710	0.314	-0.907
	Test Original	Train EB	-0.039	1.050	-0.016	-0.065
		Train HB	0.236	0.735	-0.383	1.974
		Train Original	-0.056	1.074	0.006	-0.096
IBD	Test EB	Train EB	-0.003	1.008	0.006	-0.005
		Train HB	0.104	0.774	-0.274	0.170
		Train Original	-0.043	1.104	0.095	-0.078
	Test HB	Train EB	0.466	0.344	0.189	-0.186
		Train HB	0.557	0.258	-0.097	0.022
		Train Original	0.426	0.398	0.310	-0.246
	Test Original	Train EB	-0.164	1.437	-0.052	0.068
		Train HB	-0.054	1.129	-0.312	0.260
		Train Original	-0.206	1.549	0.025	-0.023
DIAB	Test EB	Train EB	0.011	0.989	-0.009	1.550
		Train HB	0.312	0.688	-0.379	inf
		Train Original	0.003	0.997	-0.002	1.077
	Test HB	Train EB	0.062	0.937	0.386	-3.609
		Train HB	0.447	0.552	-0.028	inf
		Train Original	0.044	0.954	0.396	-4.947
	Test Original	Train EB	0.008	0.992	-0.031	1.854
		Train HB	0.302	0.698	-0.396	inf
		Train Original	0.001	0.999	-0.025	1.481

5.2.4 Group D: GLIOMA

Group D contains only the GLIOMA dataset and differs from the other groups not because of the magnitude of the gaps, but because of their direction. In eight of the nine train/test combinations, the gap is negative, meaning that the unprivileged group achieves a higher F1-score than the privileged group. The only exception is Train-HB→Test-HB, which produces a near-zero gap (+0.6 pp). This pattern is not necessarily a fairness advantage. Instead, it appears to be related to the dataset's distribution, in which most instances belong to the unprivileged group, combined with the substantial reduction in training set size after equalization.

In GLIOMA, most instances belong to the unprivileged group (Figure 5.1k), resulting in a strongly negative Class Imbalance value. As the data manipulations increased and decreased the bias metrics, the average training set size dropped from 754.2 samples in the original dataset to 394.2 in the highly biased version and only 133.2 in the equally balanced version (Table 7.1). When these training sizes are considered together with the performance gaps shown in Table 5.2, the results suggest that equalization comes at a high cost for this dataset. Using only about 17% of the original training data leads to poorer performance for both groups, even though this effect is not clearly reflected in accuracy, which changes by less than 10% between the original and equally balanced versions. The F1-score boxplots (Appendix, Figure 7.2) further show that the performance shift affects both groups.

The model-level bias metrics for GLIOMA are presented in Table 5.6. DI is equal to ∞ in seven of the nine train/test combinations, making it difficult to interpret in this dataset. Because the unprivileged group is the majority group, the model often produces very few positive predictions for the privileged group, causing the DI denominator to approach zero. DPPL is the only metric that remains stable across all combinations. Its values follow the same pattern observed in the F1-score gaps, remaining negative under Train-EB and Train-Original when evaluated on the original distribution (-0.212 and -0.190, respectively), and becoming positive only under HB training. Together, these results suggest that GLIOMA's behaviour is mainly driven by its underlying class and group distributions rather than by the same effects observed in the other dataset groups.

Table 5.6 – Model-level bias metrics across test and train variations for the GLIOMA dataset (Group D).

Dataset	**Test Variation**	**Train Variation**	DPPL	DI	DCA	DCR
GLIOMA	Test EB	Train EB	-0.033	inf	inf	-0.003
		Train HB	0.356	0.684	-0.717	inf
		Train Original	-0.001	inf	inf	0.035
	Test HB	Train EB	0.198	inf	inf	inf
		Train HB	0.572	inf	inf	inf
		Train Original	0.305	inf	inf	-0.789
	Test Original	Train EB	-0.212	inf	inf	0.070
		Train HB	0.170	0.997	-0.403	inf
		Train Original	-0.190	inf	inf	0.087

5.3 RQ1: How Does Data-Level Bias Affect Predictive Performance?

Bias manipulation consistently degraded fairness metrics and F1-score gaps across all 11 datasets, but the direction and magnitude of effects depended on the dataset, variation, and test condition. Training on Highly Biased (HB) data inflated gaps, while Equally Balanced (EB) training largely – though not universally – improved fairness. And in some cases, adding a tradeoff by increasing the fairness gap at the expensing of neutralizing data-level bias metrics.

RQ1.1 – Does EB maintain predictive performance while reducing fairness gaps?

Verdict: Mostly yes, with notable exceptions.

This claim holds across most datasets, as denoted in Figure 5.4 for all the datasets under original training. For the Heart Dataset, EB training on the original test set produced F1-gaps below 1% (−0.79%), versus 33.77% under HB training, while maintaining overall F1-scores around 73–80% for LR. For Predict Students' Dropout, EB training on original test yielded gaps of −4.35 %–−5.12% on LR and XGB, versus −10.51% under original training for DT, without major loss in overall F1. The Employee Dataset showed 7.06% to 5.21% gaps under EB training (Original test), compared to 14.31% – 5.14 under original training, while maintaining comparable F1. AIDS Clinical Trials and Diabetes similarly maintained comparable overall performance while reducing gaps.

However, two datasets **disagree**: the Intersectional Bias Dataset showed that EB training did not reduce gaps in the EB or original test settings (5.32%

Figure 5.4 – F1-Score gaps across all datasets, testing on the original test set. The gap is calculated as the difference between the F1-score for the advantaged class and the F1-Score for the disadvantaged class.



source: the Author

EB-test gap vs. 4.97% original-test gap under original training), and even increased gaps in some configurations. The Glioma dataset exhibited severe degradation: EB training used only 133.2 average samples (17% of the original 754.2), resulting in large F1 gaps of −29.30% to −33.73% – worse than the original variation – despite maintaining accuracy within 10%.

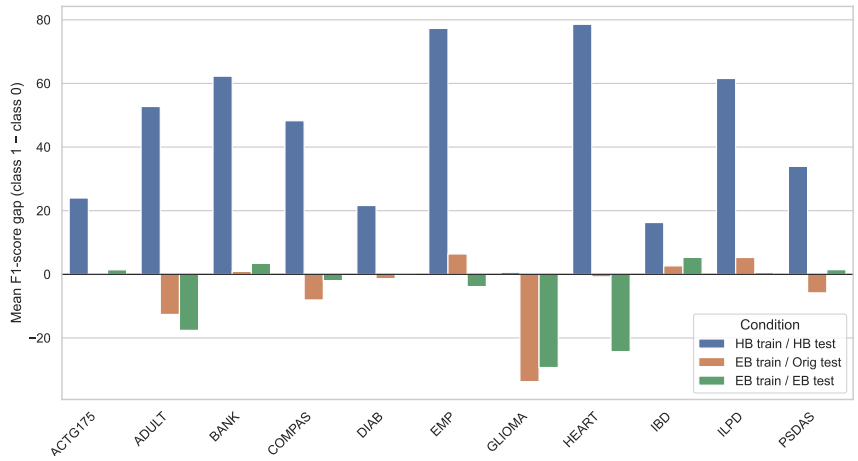
RQ1.2 – Do HB variations significantly inflate F1-score gaps vs. EB?

Verdict: Yes, consistently across 10 of 11 datasets, with the Glioma Dataset as the main exception.

HB training produced dramatically inflated gaps across nearly all datasets. The Heart Dataset reached a 78.59% gap (HB train/HB test), compared to −24.30% under EB training. The Employee Dataset showed a 77.28% gap under HB (HB train/HB test), compared to −3.86% under EB. The Bank Marketing dataset showed 62.27% vs. 3.44% under HB and EB training respectively. COMPAS reached 48.26% HB gap vs. −1.98% under EB. Indian Liver Patient showed 61.55% vs. 0.55%. AIDS Clinical Trials showed 23.98% vs. 1.40%. The Diabetes dataset showed 21.62% vs. −0.18%.

The Glioma Dataset explicitly **disagrees**: EB training tended to favor the

Figure 5.5 – Mean F1-Score gaps under HB and EB training, HB, Orig and EB test for all datasets.



source: the Author

unprivileged class, yielding an average −29.30% gap (EB train/EB test), not a reduction toward zero, likely due to the dataset's predominantly negative-target distribution for both groups, and average 0.62 gap on HB/HB variation, which can be caused by data scarcity under balancing, not due to bias amplification. In Figure 5.5, we illustrate how the EB training variation most notably reversed the bias trend on some datasets, and closed the gap on others, and how Glioma collapsed under EB variation.

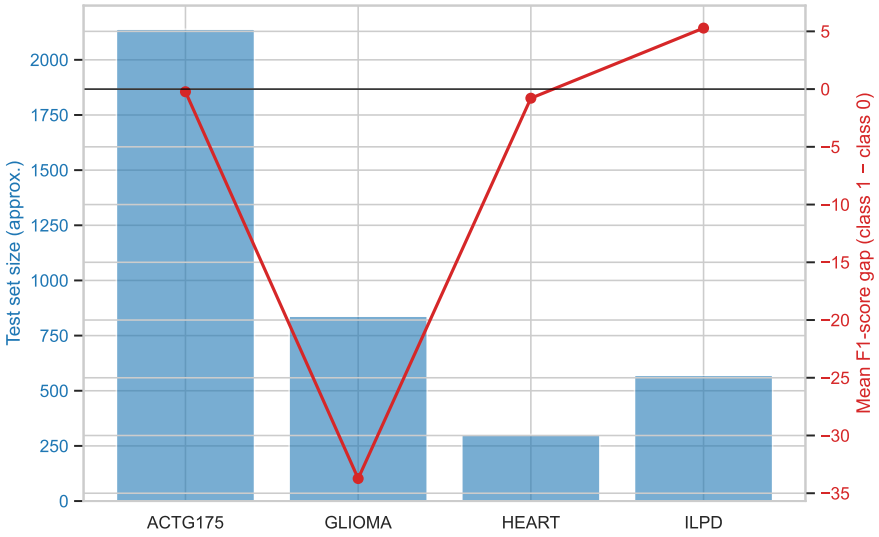
Considering that the EB variation consistently reduces the dataset size, we also investigate how this characteristic impacts model performance.

RQ1.3 – Does EB, with reduced data compared to the Orig, yield better results on Heart, Indian Liver Patient, AIDS, Glioma?

Verdict: Yes for Heart, Indian Liver Patient, and AIDS. No for Glioma.

For Heart Dataset, EB used 172.8 mean training samples (vs. 271.8 original) and achieved near-perfect fairness on the original test (−0.79% gap), outperforming the original variation (5.60% gap). For Indian Liver Patient, EB used 252 samples (vs. 513 original) and reduced original-test gaps from 9.36% to 5.28%, with comparable overall F1. For AIDS Clinical Trials, EB used 1,305 samples (vs. 1,925 original) and reduced original-test gaps from 0.36% to −0.23%, with DI improving

Figure 5.6 – Approximated mean test size (from original testing variation) in blue, compared to F1-Score gap (class advantaged - class disadvantaged), in red.



from 1.074 to 1.050.

Glioma is the clear counterexample: EB reduced training data to only 133.2 mean samples (vs. 754.2 original), causing model instability. F1-gaps under EB were −29.30% to −33.73% on original test – larger than the original variation's −24.25% – and DI remained at infinity (a sign that one group had zero predictions). This suggests a minimum data threshold below which balancing causes more harm than good. Figure 5.6 illustrates this scenario: models trained on the EB variation for Glioma, using a much smaller training set than in the original configuration, exhibit severely degraded F1-score gaps when evaluated on the original test set, highlighting the instability caused by aggressive balancing under data scarcity.

5.4 RQ2: How Do Data-Level Bias Metrics Correlate With Model-Level Bias?

This research question is decomposed into five sub-questions that progressively narrow in scope: RQ2.1 asks whether data-level metrics are predictive of downstream bias in general; RQ2.2 and RQ2.3 compare the reliability of distributional metrics (KL, KS, DD) against Class Imbalance, examining false positives, false negatives, and divergence cases; RQ2.4 focuses on whether extreme Demographic Disparity values translate into higher rejection rates for the unprivileged group; and RQ2.5 tests whether simultaneous crossing of the 0.4 threshold by all four metrics

universally results in unacceptably low Disparate Impact.

RQ2.1 – Is model-level bias accurately anticipated by data-level bias metrics?

Verdict: Generally yes, but with meaningful false positives and false negatives.

In datasets where all four data-level metrics remained below 0.3, the resulting models were consistently fair. Diabetes (original: CI=-0.122, KL=0.002, KS=0.024, DD=-0.045) produced DI of 0.995–1.010 across models. AIDS Clinical Trials (original: CI=0.322, KL=0.007, KS=0.049, DD=-0.060) produced DI of 1.044–1.092. Conversely, HB variations with extreme distributional metrics universally produced biased models across Heart, Employee, COMPAS, Predict Students' Dropout, Bank Marketing, and Adult.

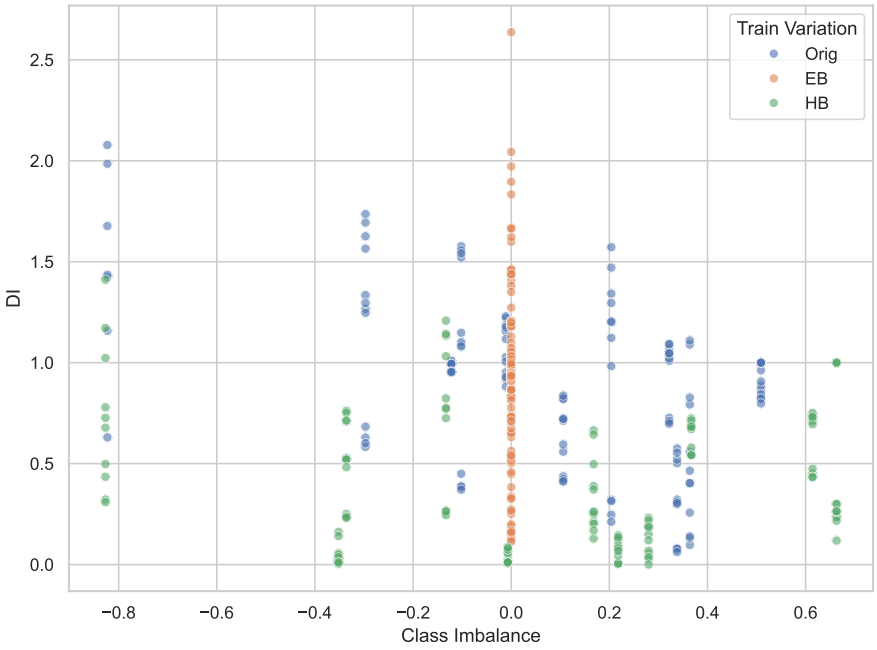
False negatives were observed in: Intersectional Bias Dataset (original metrics low, but DI reached 1.549 on original test, indicating bias against the privileged group), Bank Marketing (original metrics very low, DI=1.151–1.396 on original test), and Predict Students' Dropout (original data metrics low, yet DPPL showed consistent negative values -0.237 to -0.265 on test). Moreover, false positives appeared in Indian Liver Patient (CI=0.509 originally, but DI $\approx$ 1.0 and DPPL $\approx$ 0).

Figure 5.7 provides a global view of how the CI metric relates to model-level bias as measured by DI across all datasets, training variations, and models. Points concentrated near CI $\approx$ 0 tend to cluster around DI $\approx$ 1.0, reinforcing that low class imbalance is usually associated with fair models. At the same time, the figure also reveals meaningful false positives and false negatives: there are runs with large |CI| but DI close to 1.0 (e.g., Indian Liver Patient), and runs with CI near zero but DI far from 1.0 (e.g., Intersectional Bias, Bank Marketing, and Predict Students' Dropout),

RQ2.2 – Do distributional metrics (KL, KS, DD) exhibit fewer false positives/negatives than Class Imbalance?

Verdict: Yes, this is strongly supported by the data.

Figure 5.7 – Comparison of DI and CI across all experiments, with different colors for each train variation.



source: the Author

The most compelling case is the Heart Dataset HB variation: CI *dropped* from 0.364 to 0.280 (suggesting less bias), while KL surged from 0.208 to 2.238, KS from 0.303 to 0.806, and DD from 0.265 to 0.754. The resulting model had DI=0.046 (Heart HB train/HB test) and F1-gaps up to 78.59% – CI generated a false negative while distributional metrics correctly anticipated severe bias.

Similarly, for Indian Liver Patient, CI=0.509 (original) flags severe imbalance, but KL=0.016, KS=0.083, DD=0.075 – all near zero. The resulting models were fair (DI $\approx$ 1.0, DPPL $\approx$ 0). CI was a false positive while KL and KS correctly anticipated fairness. This pattern is consistent across Predict Students' Dropout, Employee, AIDS, Diabetes, and COMPAS. The partial exceptions are Adult and Glioma original variations, where CI still correctly co-flagged bias that was also confirmed by distributional metrics. Table 5.7 synthesizes the key dataset–variation combinations discussed, contrasting the magnitude of Class Imbalance with the distributional metrics (KL, KS, DD) and the resulting DI. The Indian Liver Patient original setting illustrates a clear false positive of CI: despite $CI = 0.509$, all three distributional metrics remain close to zero and the model's DI stays near 1.0, indicating essentially fair behavior. In contrast, the Heart HB condition shows the opposite pattern: CI drops to 0.280, while KL, KS and DD become extreme and

Table 5.7 – Summary of data-level bias metrics and DI for key datasets mentioned on RQ2.2 and RQ2.3.

Case	|CI|	KL	KS	|DD|	DI
ILPD orig	0.509	0.016	0.083	0.075	0.912
HEART HB	0.280	2.238	0.806	0.754	0.046
BANK orig	0.204	0.007	0.040	0.092	1.396
PSDAS orig	0.297	0.104	0.227	0.207	1.655
ADULT orig	0.338	0.141	0.196	0.238	0.309
GLIOMA orig	0.823	0.074	0.191	0.063	inf

DI collapses towards zero, revealing severe model-level bias that CI alone fails to anticipate.

RQ2.3 – When metrics diverge, do KL/KS serve as more reliable predictors than Class Imbalance?

Verdict: Yes, supported by multiple datasets.

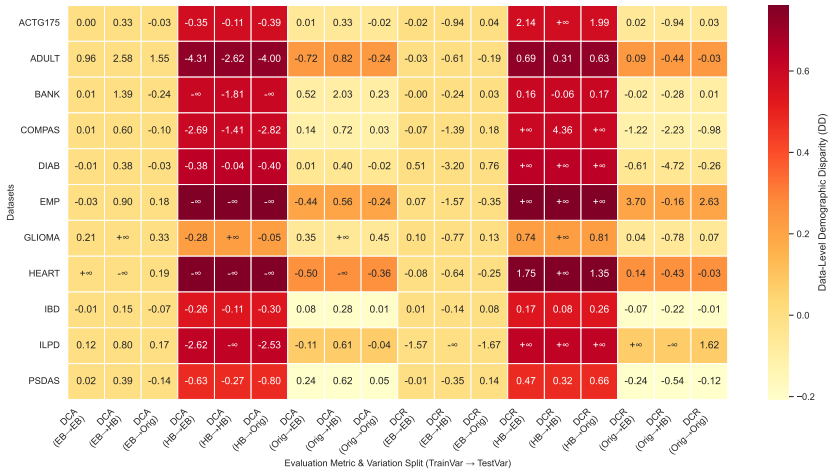
The Indian Liver Patient case provides the cleanest evidence: CI=0.509 (HB: 0.663) flagged strong bias, while KL=0.016 and KS=0.083 in the original variation indicated fair distributions – and the models confirmed fairness (DI $\approx$ 1.0). In the Heart HB variation, CI alone was misleading (dropped to 0.280), while KL=2.238 and KS=0.806 correctly predicted catastrophic model bias. For Bank Marketing, all original distributional metrics were nearly zero (KL=0.007, KS=0.040, DD=-0.092), and the original model maintained DI between 1.200–1.572 – a marginal deviation the distributional metrics underestimated, making this a partial exception. The Adult Dataset represents a case where CI (0.338) and distributional metrics (KL=0.141, KS=0.196) diverge and CI is the metric that anticipated bias (DI 0.3), according to Table 5.7.

RQ2.4 – Do extreme DD scores correlate with models unfairly rejecting the unprivileged class?

Verdict: Yes, with strong empirical support.

In Figure 5.8, there is a structural propagation from severe data-level demographic disparity to severe downstream model-level rejection behavior, noted by

Figure 5.8 – Heatmap of Model Rejection Metrics (DCA and DCR) across different Train → Test variation splits, with cell background color intensity scaled by the training set's Demographic Disparity (DD). Extreme negative values (DCA → $-\infty$) and massive positive inflation (DCR → $+\infty$) strongly correlate with saturated dark-red rows representing high data-level bias.



source: the Author

DCA and DCR metrics.

For the Heart HB variation (DD=0.754), DCA reached $-\infty$ across all model,s and DCR was large and positive in average (e.g., 1.346(EB test)–1.897(HB test)), indicating systematic non-acceptance of the unprivileged group and over-rejection. For the Employee HB variation (DD=0.761), DCA=-∞ and DCR reached $+\infty$ for Logistic Regression, confirming zero favorable predictions for the unprivileged group. For the COMPAS HB variation (DD=0.600), DCA reached -1.601 and DCR=6.808 (average on all models), also confirming the pattern. Bank Marketing HB (DD=0.605) showed DCA= $-\infty$ for DT and RF.

A partial exception is the Glioma HB variation (DD=0.221, below the extreme range), where DI was infinite due to zero privileged predictions, not unprivileged ones – the bias direction reversed because the dataset's unprivileged group was the majority.

RQ2.5 – Does agreement of all four metrics crossing 0.4 universally result in unacceptably low DI?

Verdict: Mostly yes, but the strict condition of all four crossing 0.4 is rarely met because specific metrics often lag behind.

While DI consistently drops to unacceptable levels under severe bias, it is

rare for all four metrics to simultaneously cross 0.4. In most cases, the three distributional metrics (KL, KS, DD) easily crossed the threshold, but Class Imbalance (CI) disagreed and remained lower. Despite CI failing to cross 0.4, unacceptable DI values were consistently observed: Heart HB (CI=0.280, KL=2.238, KS=0.806, DD=0.754) → DI=0.046 (HB test, averaged); Employee HB (CI=0.218, KL=1.364, KS=0.749, DD=0.761) → DI=0.004(LR HB test)–0.080(DT HB test); COMPAS HB (CI=0.168, KL=0.727, KS=0.577, DD=0.600) → DI=0.129(LR HB test)–0.497(XGB HB test); and Diabetes HB (CI=0.367, KL=0.449, KS=0.416, DD=0.638) → DI=0.542(DT HB test)–0.695(LR original test).

The Intersectional Bias Dataset perfectly illustrates this trend: all distributional metrics were high under the HB variation (KL=0.668, KS=0.541, DD=0.538), but CI severely disagreed, remaining at just 0.133. Yet, DI still collapsed to 0.245(RF HB test)–0.267(DT HB test), which is an unacceptable fairness failure.

However, the Glioma HB variation proves that distributional metrics do not always universally agree either. In Glioma, CI was extremely high (0.827) alongside KL (0.711) and KS (0.564), but the distributional metric DD disagreed and fell well below the threshold at 0.221. Despite this lack of agreement, the DI hit inf (or very high) because the unprivileged group entirely dominated predictions, meaning the direction of bias was inverted rather than absent. Strict universality of all four metrics crossing 0.4 is therefore not required to guarantee fairness failure.

5.5 RQ3: Which Algorithms Are More Robust Against Data-Level Bias?

RQ3 examines whether algorithm choice moderates the impact of data-level bias on fairness outcomes. RQ3.1 asks whether ensemble methods (Random Forest and XGBoost) absorb distributional bias more severely than Logistic Regression in terms of model-level fairness metrics, testing whether greater model expressiveness amplifies bias under skewed training conditions. RQ3.2 extends this comparison to feature importance, asking whether ensemble algorithms more frequently elevate the protected attribute to the top-ranked feature under HB training — a behavioral signature of bias absorption that LR, through its linear decision boundary, may be less susceptible to.

RQ3.1 – Do ensemble algorithms absorb bias more severely than Logistic Regression?

Verdict: No clear universal pattern – all algorithms are similarly affected.

Analyzing ensemble versus linear algorithms, we identified that no clear universal pattern could be detected. Under HB conditions, all four algorithms suffered comparably severe fairness degradation (complete results are provided in the Appendix, Table 7.5). For Employee HB (original test), Logistic Regression produced DI=0.006, Decision Tree=0.128, Random Forest=0.082, XGBoost=0.120 – LR actually performed worst. For Heart HB (original test), LR produced DI=0.180, DT=0.193, RF=0.149, XGBoost=0.054 – XGBoost was worst. For COMPAS HB, LR produced DI=0.212, DT=0.260, RF=0.389, XGBoost=0.665 – here RF and XGBoost were notably more robust than LR. For Bank Marketing HB, LR=0.162, DT=0.027, RF=0.022, XGBoost=0.056 – tree-based models performed worse here.

The results show no consistent hierarchy: LR was least robust in Employee HB, while DT and RF were least robust in Bank Marketing HB. XGBoost was most robust in COMPAS HB but least in Heart HB. The conclusion is that no algorithm is universally more robust, though XGBoost showed marginally better behavior in some high-bias settings due to its regularization, but this did not hold consistently.

RQ3.2 – Do ensemble algorithms elevate the protected attribute to #1 feature importance more frequently than LR under HB?

Verdict: No. The elevation of the protected attribute is a universal vulnerability across both ensemble and linear algorithms.

Under HB conditions, the protected attribute systematically surged to the top of the feature importance rankings regardless of the algorithm's complexity. The feature importance logs confirm that Logistic Regression (LR) absorbed the bias just as severely as tree-based models, and in some cases, ranked the protected attribute even higher than the ensembles did when evaluated by the absolute value of its coefficients.

- For the **Heart HB** dataset, all models elevated `sex` to the #1 position. XG-

Table 5.8 – Protected attribute mean feature importance rank across four models (LR, DT, RF, XGBoost) by training variation. Lower rank indicates higher importance. Δ Orig→HB is the change in mean rank from original to highly biased training (negative values indicate the attribute rose in importance). "HB rk. 1" reports how many of the four models assign rank 1 to the protected attribute under HB training.

			Mean rank (4 models)				
Dataset	**Prot. attr.**	**Group**	*Orig*	*EB*	*HB*	**Δ O→HB**	**HB rk. 1**
Employee	*gender*	A	3.2	7.2	1.0	−2.2	4/4
Bank Mktg	*marital*	A	29.0	34.5	1.0	−28.0	4/4
Heart	*sex*	B	6.8	12.0	1.0	−5.8	4/4
COMPAS	*race*	B	6.2	5.8	1.2	−5.0	3/4
Adult	*sex*	B	13.5	32.8	2.2	−11.3	2/4
Ind. Liver	*sex*	B	9.8	8.8	1.0	−8.8	4/4
PSDAS	*gender*	C	24.5	13.5	1.2	−23.3	3/4
ACTG175	*homo*	C	13.0	13.2	1.8	−11.2	1/4
Intersect.	*sex*	C	8.8	7.2	2.0	−6.8	1/4
Diabetes	*sex*	C	9.0	11.0	1.0	−8.0	4/4
Glioma	*race*	D	10.2	9.5	2.2	−8.0	1/4

Rank colour bands: rank 1–3 (dominant), rank 4–10 (moderate), rank >10 (low importance).

Boost placed its importance at 0.570, Random Forest at 0.333, and Logistic Regression reported a massive coefficient importance of 3.336 for `sex`, putting it firmly at rank #1.

- For the **Employee HB** dataset, `Gender` reached the #1 rank across the board: XGBoost (0.833), RF (0.593), DT (0.669), and Logistic Regression (3.967).
- For the **Adult HB** dataset, LR actually elevated the protected attribute *more* aggressively than XGBoost. While XGBoost placed `sex` at rank #2 (0.140) behind `relationship_Husband`, Logistic Regression placed `sex` at rank #1 with an extreme importance score of 3.958.
- For **COMPAS HB**, XGBoost placed `race` at #1 (0.897) [16, 17]. Logistic Regression mirrored this strong reliance, placing `race` at rank #2 (3.090), just fractionally behind `c_charge_degree_O` (3.337).

This pattern extends beyond these examples. Under HB training, Logistic Regression also elevated the protected attribute to the absolute #1 rank in the Predict Students' Dropout dataset (`Gender`: 2.827), the Glioma dataset (`Race`: 2.490), and the Bank Marketing dataset (`marital`: 3.037).

Therefore, ensemble methods do not elevate the protected attribute more frequently than Logistic Regression; this behavior is a consistent reflection of severe data-level bias rather than an algorithm-specific flaw.

5.6 RQ4: What is the impact on Feature Importance of Protected Attributes?

RQ4 investigates the relationship between feature importance rankings of the protected attribute and observed fairness outcomes. This analysis is summarized in Table 5.8. RQ4.1 asks whether datasets where the protected attribute is naturally prominent (top-6 rank under original training) are disproportionately affected by bias, while RQ4.2 examines the converse: whether severe F1-gaps can occur even when the protected attribute does not dominate feature importance under HB training. RQ4.3 and RQ4.4 shift focus to the effect of EB training on feature reliance, asking first whether EB strictly suppresses the protected attribute's rank relative to original training, and then whether this suppression translates into a quantifiable and consistent reduction in model dependence on the sensitive attribute across datasets and classifiers.

RQ4.1 – Are models more severely affected when the sensitive attribute naturally ranks in the top 6?

Verdict: Partially supported, but not a strict rule.

In the Heart Dataset, `sex` originally ranked **#6** in XGBoost importance (0.059). Under HB it rose to **#1** (0.570), representing a dramatic amplification. In the Employee Dataset, `Gender` ranked **#6** originally (0.078) and rose to **#1** under HB (0.833). In the Intersectional Bias Dataset, `Sex` ranked **#4** originally (0.060) and rose to **#2** under HB (0.147) – here the attribute was already in the top 4, and the amplification was more modest.

Conversely, in COMPAS, `race` ranked **#10** (0.004) in the original XGBoost model, yet rose to **#1** (0.897) under HB – the severe amplification happened despite a low original rank, **contradicting** the hypothesis. In AIDS Clinical Trials, `homo` ranked **#7** originally (0.041) and rose to **#2** under HB (0.191) – again a low-starting attribute experiencing large amplification. The data therefore does **not confirm** that top-6 original rank is a prerequisite for severe amplification.

RQ4.2 – Can a model exhibit F1-gap > 30% without the protected attribute reaching #1 feature importance?

Verdict: Yes, confirmed by multiple datasets.

The COMPAS original training variation produced average F1-gaps of 36.86% on HB test, yet `race` ranked **#10** in XGBoost importance (0.004) under original training. The AIDS Clinical Trials HB variation produced an F1-gap of ∼23.98% on HB test under HB training, while `homo` was ranked **#2** (0.191), not #1 – `time` dominated at 0.262. The Predict Students' Dropout HB variation produced F1-gaps of 33.93% (HB train/HB test, on average), while `Gender` was **#1** at 0.206 – this case agrees with the hypothesis.

For the Bank Marketing HB variation, F1-gap reached an average of 62.27% while `marital` (the protected attribute) was indeed **#1** (0.453), so this case does not support the question's premise. The COMPAS example is the cleanest evidence: large gaps (>30%) arose from the test data distribution rather than from the model having internalized the protected attribute as a primary feature during original training.

RQ4.3 – Does EB training strictly push the protected attribute's importance rank lower than the original?

Verdict: Not strict – often lower than HB, but not always lower than original.

For the Heart Dataset, `sex` dropped from rank #6 (original, XGBoost=0.059) to rank #12 (EB, XGBoost=0.013) – clearly lower than original. For Predict Students' Dropout, `Gender` moved from rank ∼31 (original, XGBoost=0.012) to rank ∼12 (EB, XGBoost=0.020) – this is actually **higher** than the original rank, not lower. For the Employee Dataset, `Gender` dropped from rank #6 (original, 0.078) to rank #9 (EB, 0.059) – slightly lower. For COMPAS, `race` remained near the bottom in both original (#10, 0.004) and EB (#7, 0.007) – roughly comparable.

For the Intersectional Bias Dataset, `Sex` was rank #4 (original, 0.060) and remained rank #3 (EB, 0.053) – essentially unchanged. For Glioma, `Race` was rank #15 (original, 0.010) and rank #10 (EB, 0.040) – **higher** under EB, not lower. The claim is therefore **not strictly true**: EB reliably brings importance below the HB

level, but does not consistently push it strictly below the original unmanipulated rank.

RQ4.4 – Do EB-trained models exhibit a quantifiable reduction in reliance on protected attributes?

Verdict: Yes, compared to HB, consistently. Compared to original, mostly but not universally.

The reduction compared to HB is consistent and quantifiable across all datasets where HB dramatically elevated the protected attribute. This is illustrated on Table 5.9

Table 5.9 – Reduction in XGBoost bias from Highly Biased (HB) to Equally Balanced (EB) training across selected datasets.

Dataset	HB	EB	Reduction
Heart	0.570	0.013	−97.7%
Employee	0.833	0.059	−92.9%
COMPAS	0.897	0.007	−99.2%
Predict Students' Dropout	0.206	0.020	−90.3%
Bank Marketing	0.453	0.002	−99.6%
Adult (sex)	0.140	0.008	−94.3%

Compared to the original baseline, the reduction is less uniform. For Indian Liver Patient, `Sex` was ≈0.0 in both original and EB XGBoost - no change. For Glioma, `Race` increased from 0.010 (original) to 0.040 (EB) – a slight increase. For Intersectional Bias, `Sex` decreased slightly from 0.060 to 0.053 – a modest but quantifiable reduction.

Overall, the results from the four research questions analyzed in this chapter indicate that addressing bias at the data level is often more effective than relying on model-level corrections alone. Across the analyzed datasets, balancing the data generally reduced fairness gaps without necessarily degrading predictive performance, although the impact depended on dataset size and the severity of rebalancing. More importantly, the findings show that distributional metrics provide stronger and cheaper early warning signals than class imbalance alone, making them useful for audit and data curation before model training. The results also reinforce the idea that no algorithm can fully compensate for biased input data, since all models were affected to some degree by the training distribution. In this sense, the thesis con-

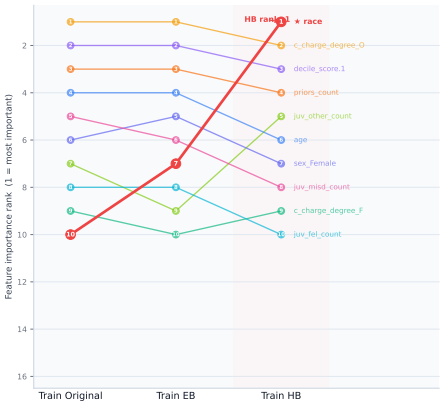
Figure 5.9 – Feature importance bump charts for HEART, IBD, COMPAS and ACTG175 under XGB, rank is the feature importance order for the given training set. The protected attribute is marked with a star and highlighted with a red arrow.



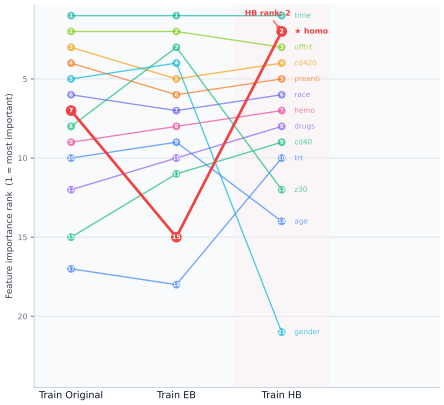
(a) HEART



(b) IBD



(c) COMPAS



(d) ACTG175

source: the Author

firms the “garbage in, garbage out” principle and suggests that fairness-oriented machine learning workflows should prioritize representativeness and balance over simply increasing data volume.

6 CONCLUSION

The goal of this work was to investigate a critical and underexplored dimension of fairness considerations: the extent to which data-level bias metrics, that can be computed entirely before model training, can anticipate downstream model-level unfairness. This work addressed four primary research questions: the effects of distributional manipulation on predictive performance; the correlation between data-level and model-level bias metrics; algorithm robustness; and the role of protected attributes in learned feature representations.

To investigate these questions, a controlled experimental pipeline was constructed in which 11 diverse datasets, sourced from multiple domains, such as healthcare, criminal justice, finance, and education, and three variations from each dataset: the original data, an Equally Balanced (EB) variation, and a Highly Biased (HB) variation. Four supervised learning algorithms were trained and evaluated across all combinations, yielding 396 distinct experimental configurations. Bias was measured at the data level through Class Imbalance (CI), KL Divergence (KL), Kolmogorov-Smirnov (KS), and Demographic Disparity (DD), and at the model level through Difference in Positive Proportions in Predicted Labels (DPPL), Disparate Impact (DI), and Difference in Conditional Outcome (DCO).

To characterize the diversity of responses across datasets, a hierarchical clustering analysis of F1-gap profiles revealed four behavioral groups: Group A exhibiting the most extreme HB-induced gaps; Group B showing the canonical bias-transfer signature where EB reliably neutralizes gaps that HB reliably reintroduces; Group C characterized by uniformly low sensitivity to distributional manipulation; and Group D representing a structural outlier driven by majority-unprivileged composition rather than bias dynamics.

The evidence gathered in response to the first research question is unambiguous: training on Highly Biased data consistently and severely inflated F1-score gaps between privileged and unprivileged groups across 10 of the 11 evaluated datasets. The Employee dataset illustrates this point exhibiting gaps of 77.28 under HB training versus -3.86 under EB. Equally important, EB training largely maintained overall predictive performance while simultaneously reducing these fairness gaps. A critical exception was the Glioma dataset, where EB training reduced available training samples to only 133.2 on average (approximately 17% of the original 754.2)

causing model instability and F1-gaps of 29.30 to 33.73 training on EB variation, suggesting the existence of a minimum data threshold below which distributional balancing becomes counterproductive. Bias direction was also changed in some datasets, suggesting that there's a tradeoff between fairness metrics and F1-Score.

With respect to the second research question, the findings broadly confirm that pre-training distributional metrics can serve as reliable early-warning indicators of downstream unfairness, while also revealing important limitations of Class Imbalance as a standalone indicator. While Highly Biased (HB) variations with extreme distributional metrics universally produced biased models, the data firmly rejects the converse assumption that low data-level metrics guarantee model fairness. For example, in the Adult original dataset, all four metrics remained comfortably below the 0.4 threshold (|CI|=0.338, KL=0.141, KS=0.196, |DD|=0.238), yet the resulting models still exhibited catastrophic bias with an average Disparate Impact (DI) around 0.30 for all models. This reveals a critical "false negative" blind spot in data-level metrics alone.

However, when anticipating severe bias, the most compelling evidence for the superiority of distributional metrics over Class Imbalance came from two contrasting cases. In the Heart Dataset HB variation, CI actually decreased while KL, KS, and DD correctly anticipated the model's fairness collapse. Conversely, in the Indian Liver Patient original dataset, CI flagged strong imbalance, yet KL, KS, and DD remained near zero, and the models correctly yielded fair outcomes. These cases, corroborated across multiple datasets, provide strong empirical support that KL Divergence and the Kolmogorov-Smirnov statistic yield materially fewer false positives than Class Imbalance, and should be preferred in pre-deployment fairness audits when the two types of metrics diverge. Furthermore, extreme Demographic Disparity scores reliably predicted conditional unfairness: across the Heart, Employee, and COMPAS HB variations, extreme |DD| values correlated with systematic over-rejection of the unprivileged group, as confirmed by massive spikes in the Difference in Conditional Rejection (DCR) metric.

The third research question tested the hypothesis that ensemble methods would be more susceptible to bias than Logistic Regression, given their greater representational capacity. The results refute this hypothesis entirely. No algorithm demonstrated consistent, universal superiority in fairness robustness: under HB conditions, all four algorithms suffered comparably severe fairness degradation, with the

relative ordering varying by dataset. For the Employee dataset, Logistic Regression produced the worst DI, outperforming even tree-based models in severity; for COMPAS, Random Forest and XGBoost were notably more robust than Logistic Regression; for Bank Marketing, tree-based models performed worst. The conclusion is that vulnerability to data-level bias is fundamentally a property of the data distribution, not of the algorithmic architecture. This finding has practical significance: it implies that switching model families cannot serve as a substitute for data-level fairness interventions.

The fourth and final research question examined the effect of bias manipulation on the role of protected attributes in learned feature representations. Under HB training, the protected attribute was systematically elevated to the top position in feature importance rankings across all four algorithms and nearly all datasets evaluated. For the Heart dataset HB condition, the sex attribute ranked 1 across all models, with XGBoost and Logistic Regression assigning a coefficient higher than any ensemble model. Similarly, COMPAS dataset showed race increasing in feature importance under HB XGBoost. This confirms that the elevation of protected attributes under biased training is a universal phenomenon, not an artifact of any particular model family. Conversely, EB training consistently reduced reliance on protected attributes relative to the HB condition by quantifiable margins: Heart, Employee, COMPAS, and Bank Marketing. These reductions were not always consistent relative to the unmanipulated original baseline; however, with datasets such as Glioma and Predict Students' Dropout showed non-monotonic behavior, indicating that the effect of balancing on feature importance is most reliable as a guard against actively amplified bias rather than low-magnitude organic disparities.

The practical implications of these findings extend to any organizational context in which machine learning is used to support consequential decisions. The overarching contribution is the demonstration that a compact set of pre-training metrics — particularly KL Divergence, Kolmogorov-Smirnov, and Demographic Disparity — can function as reliable, computationally inexpensive screening tools in data auditing pipelines. In healthcare settings, where datasets such as the Indian Liver Patient, AIDS Clinical Trials, and Diabetes exemplify the risks of biased clinical models, these metrics can prospectively identify risks before any model is deployed, enabling data curation as an effective first-line fairness intervention. In criminal justice applications such as recidivism prediction, experiments showed that DI de-

graded from near-parity to very low values under HB conditions in the COMPAS dataset, indicating a systematic and severe disadvantage for the unprivileged group. The practical implication is that institutions deploying risk assessment instruments should mandate pre-training audits using distributional metrics, with the evidence here suggesting that convergent agreement of all four metrics above 0.4 constitutes a strong signal of unacceptable downstream DI. We also note that, while these metrics can serve as a powerful indicator of bias, there is still a need to periodically assess the deployed model using model-level bias metrics or bias mitigation techniques.

6.1 Limitations and Future Work

While this work provides a thorough investigation of the impact of data-level bias metrics, it still has limitations that circumscribe the direct generalizability of these conclusions. The experimental framework is restricted to binary classification on tabular data, and the findings do not directly extend to multi-class, regression, or unstructured data settings (some of the metrics do not support multi-class classification problems). The bias manipulation strategy aimed to be controlled and reproducible, which produces synthetic distributional shifts that may not fully capture the organic complexity of real-world bias generation, which often emerges from the interplay of historical inequities, measurement artifacts, and feedback loops. The Glioma edge case additionally reveals that minimum sample thresholds need to be formally characterized before recommending EB resampling in small-N clinical cohorts. Future work should examine whether the threshold relationships identified here generalize across broader dataset collections and intersectional protected attributes, and whether these metrics can be effectively integrated into AutoML and MLOps platforms as automated fairness gates at the data ingestion stage.

In sum, this thesis provides systematic empirical evidence that data-level bias metrics are not merely retrospective descriptors of data quality, but prospective predictors of model unfairness. By demonstrating that KL Divergence and the Kolmogorov-Smirnov test outperform Class Imbalance as early-warning indicators, that no algorithm is inherently resilient to data-level bias, and that distributional balancing can reduce model reliance on protected attributes, this work bridges the gap between statistical pre-training audits and empirical model outcomes. It provides the research community and practitioners with an evidence-based foundation

for embedding fairness considerations at the earliest and most tractable stage of the machine learning pipeline: before a single prediction is made.

7 APPENDIX

7.1 Further details about datasets

In Table 7.1, we report the mean train and test sizes used for model training and evaluation. This is the average of the 10-fold cross-validation and the four models used for each dataset.

Table 7.1 – Dataset Sizes Across Different Training and Testing Variations

	Train Highly Biased	Test Highly Biased	Train Original	Test Original	Train Equally Balanced	Test Equally Balanced
PSDAS	2060.50	229.50	3981.60	442.40	2800.80	311.20
EMP	1699.60	188.90	2487.60	276.40	1971.00	219.00
COMPAS	5904.90	656.10	10581.30	1175.70	9469.80	1052.20
ADULT	17551.80	1950.30	29283.30	3253.70	19371.60	2152.40
BANK	21255.20	2361.80	40689.90	4521.10	32394.60	3599.40
HEART	188.40	21.60	271.80	30.20	172.80	19.20
ILPD	365.10	40.90	513.00	57.00	252.00	28.00
ACTG175	1249.00	138.90	1925.10	213.90	1305.00	145.00
IBD	5188.00	576.50	9900.00	1100.00	8019.00	891.00
DIAB	114998.80	12778.20	206526.60	22947.40	181366.20	20151.80
GLIOMA	394.20	43.80	754.20	83.80	133.20	14.80

7.2 Best hyperparameters found by grid search

The following tables summarize the best hyperparameters found for each algorithm, in each dataset analyzed.

Table 7.2 – Best Hyperparameters found for RandomForestClassifier

Dataset	RandomForestClassifier
ACTG175	‘criterion’: ‘entropy’, ‘max_depth’: 10, ‘n_estimators’: 600
ADULT	‘criterion’: ‘entropy’, ‘max_depth’: 100, ‘n_estimators’: 600
BANK	‘criterion’: ‘gini’, ‘max_depth’: 7, ‘n_estimators’: 10
COMPAS	‘criterion’: ‘gini’, ‘max_depth’: 5, ‘n_estimators’: 50
DIAB	‘criterion’: ‘gini’, ‘max_depth’: 10, ‘n_estimators’: 50
EMP	‘criterion’: ‘entropy’, ‘max_depth’: 7, ‘n_estimators’: 200
GLIOMA	‘criterion’: ‘gini’, ‘max_depth’: 7, ‘n_estimators’: 50
HEART	‘criterion’: ‘entropy’, ‘max_depth’: 5, ‘n_estimators’: 100
ILPD	‘criterion’: ‘entropy’, ‘max_depth’: 5, ‘n_estimators’: 100
IBD	‘criterion’: ‘gini’, ‘max_depth’: 100, ‘n_estimators’: 50
PSDAS	‘criterion’: ‘entropy’, ‘max_depth’: 100, ‘n_estimators’: 600

Table 7.3 – Best Hyperparameters found for DecisionTreeClassifier

Dataset	DecisionTreeClassifier
ACTG175	'criterion': 'entropy', 'max_depth': 5
ADULT	'criterion': 'gini', 'max_depth': 10
BANK	'criterion': 'entropy', 'max_depth': 5
COMPAS	'criterion': 'gini', 'max_depth': 7
DIAB	'criterion': 'gini', 'max_depth': 5
EMP	'criterion': 'entropy', 'max_depth': 7
GLIOMA	'criterion': 'gini', 'max_depth': 5
HEART	'criterion': 'gini', 'max_depth': 50
ILPD	'criterion': 'gini', 'max_depth': 5
IBD	'criterion': 'entropy', 'max_depth': 7
PSDAS	'criterion': 'entropy', 'max_depth': 5

Table 7.4 – Best Hyperparameters found for XGBClassifier

Dataset	XGBClassifier
ACTG175	'learning_rate': 0.1, 'max_depth': 5, 'n_estimators': 50
ADULT	'learning_rate': 0.2, 'max_depth': 5, 'n_estimators': 100
BANK	'learning_rate': 0.5, 'max_depth': 5, 'n_estimators': 10
COMPAS	'learning_rate': 0.01, 'max_depth': None, 'n_estimators': 50
DIAB	'learning_rate': 0.01, 'max_depth': None, 'n_estimators': 300
EMP	'learning_rate': 0.2, 'max_depth': 5, 'n_estimators': 20
GLIOMA	'learning_rate': 0.5, 'max_depth': 5, 'n_estimators': 10
HEART	'learning_rate': 0.5, 'max_depth': None, 'n_estimators': 10
ILPD	'learning_rate': 0.001, 'max_depth': 5, 'n_estimators': 10
IBD	'learning_rate': 0.1, 'max_depth': None, 'n_estimators': 300
PSDAS	'learning_rate': 0.1, 'max_depth': 100, 'n_estimators': 300

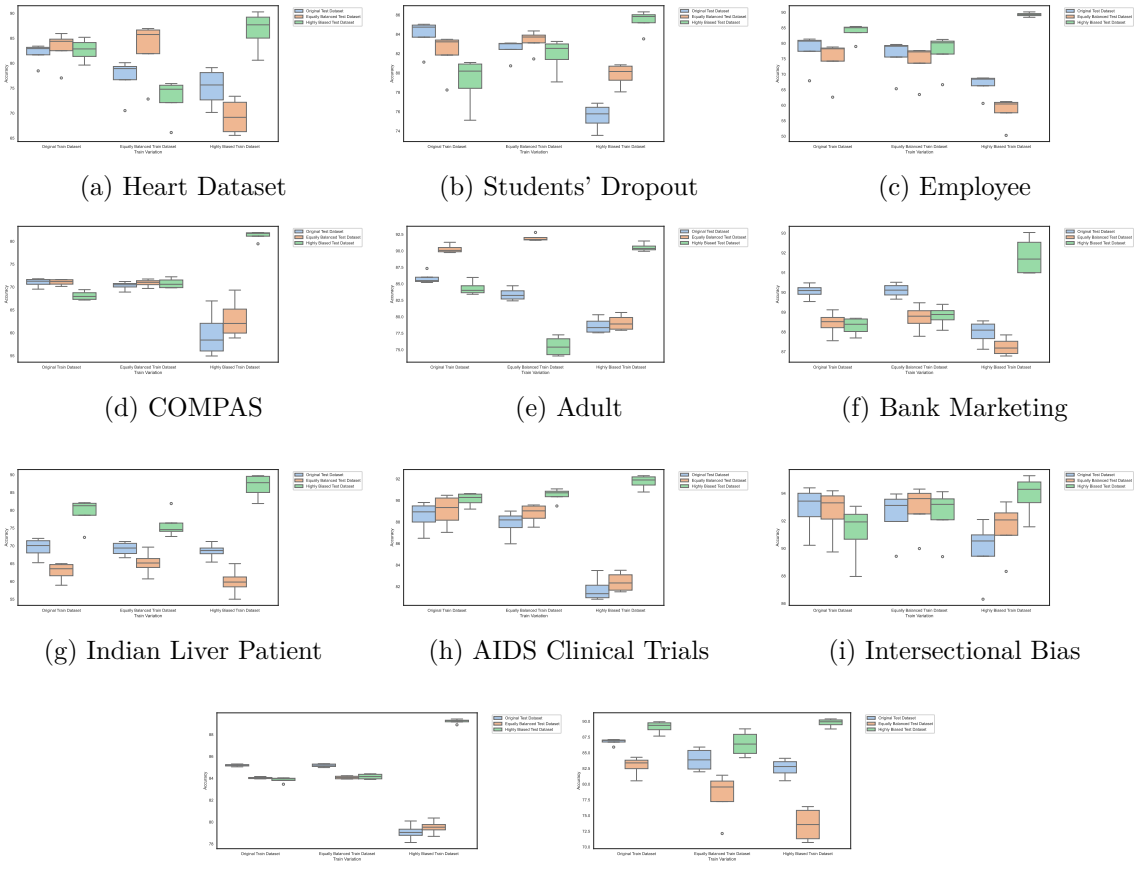
7.3 Accuracy and F1-Score Results

The complete relation of accuracy results for each dataset is presented in Figure 7.1. For the F1-Score, results are presented in Figure 7.2.

7.4 Complete Performance and Metrics Results

In Table 7.5, we present all the results obtained from this experiment, listing all the datasets grouped by its training and testing variations.

Figure 7.1 – Box-plot results for accuracy on all the analyzed datasets. X axis is the train variation, Y axis is the test variation, in the following order: Original Dataset (blue), Equally Balanced Dataset (Orange), and Highly Biased Dataset (green).



(a) Heart Dataset (b) Students' Dropout (c) Employee

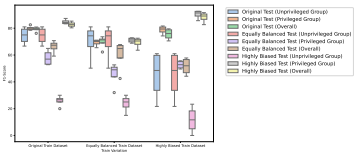
(d) COMPAS (e) Adult (f) Bank Marketing

(g) Indian Liver Patient (h) AIDS Clinical Trials (i) Intersectional Bias

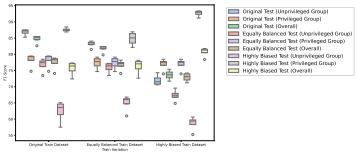
(j) Diabetes (k) Glioma Dataset

source: the Author

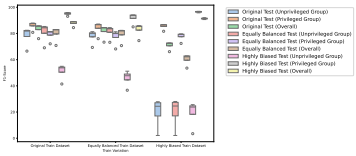
Figure 7.2 – Box-plot results for F1-score on all the analyzed datasets. X axis is the train variation, Y axis is the test variation, the bars, from left to right correspond to: Original Test (unprivileged group), Original Test (Privileged Group), Original Test (Overall), Equally Balanced Test (Unprivileged Group), Equally Balanced Test (Privileged Group), Equally Balanced Test (Overall), Highly Biased Test (Unprivileged Group), Highly Biased Test (Privileged Group), and Highly Biased Test (Overall).



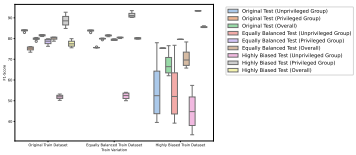
(a) Heart



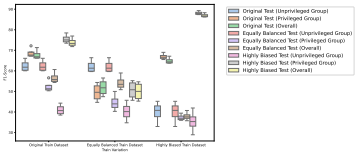
(b) Students' Dropout



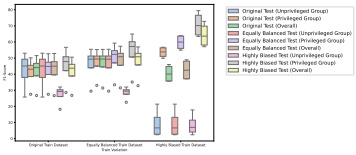
(c) Employee



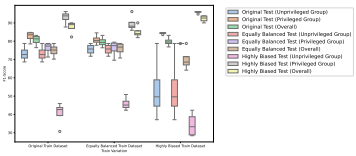
(d) COMPAS



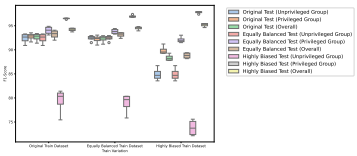
(e) Adult



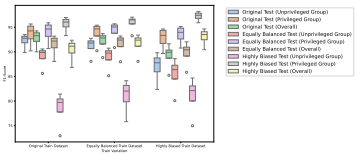
(f) Bank Marketing



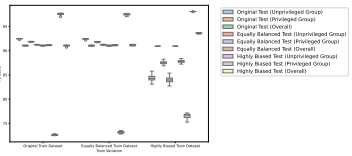
(g) Indian Liver Patient



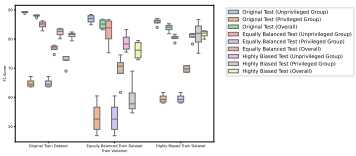
(h) AIDS Clinical Trials



(i) Intersectional Bias



(j) Diabetes



(k) Glioma Dataset

source: the Author

Table 7.5 – Complete results for all datasets, training and testing variations, and models

Dataset	Train Variation	Test Variation	Model	F1-Score	Accuracy	CI	KL	KS	DD	DPPL	DI	DCA	DCR
ACTG175	Train EB	Test EB	DT	92.990	88.623	0.000	0.000	0.000	0.000	0.001	0.999	-0.001	0.034
			LR	92.374	87.520	0.000	0.000	0.000	0.000	-0.001	1.003	0.002	-0.023
			RF	93.538	89.446	0.000	0.000	0.000	0.000	-0.021	1.029	0.024	-0.232
			XGB	93.542	89.580	0.000	0.000	0.000	0.000	-0.021	1.029	0.025	-0.153
		Test HB	DT	94.638	90.781	0.000	0.000	0.000	0.000	0.277	0.704	0.298	-0.704
			LR	93.935	89.486	0.000	0.000	0.000	0.000	0.249	0.734	0.327	-0.944
			RF	94.823	91.073	0.000	0.000	0.000	0.000	0.253	0.731	0.325	-0.915
			XGB	94.528	90.639	0.000	0.000	0.000	0.000	0.267	0.712	0.311	-0.927
		Test O	DT	92.325	87.986	0.000	0.000	0.000	0.000	-0.029	1.037	-0.027	0.050
			LR	91.164	85.976	0.000	0.000	0.000	0.000	-0.024	1.030	-0.033	0.042
			RF	92.648	88.406	0.000	0.000	0.000	0.000	-0.048	1.061	-0.007	-0.219
			XGB	92.928	89.013	0.000	0.000	0.000	0.000	-0.055	1.073	0.004	-0.131
	Train HB	Test EB	DT	88.162	81.520	0.614	0.586	0.506	0.575	0.279	0.695	-0.386	2.119
			LR	88.352	81.724	0.614	0.586	0.506	0.575	0.261	0.714	-0.355	2.143

Continued on next page

Table 7.5 – Complete results for all datasets, training and testing variations, and models

Dataset	Train Variation	Test Variation	Model	F1-Score	Accuracy	CI	KL	KS	DD	DPPL	DI	DCA	DCR
			RF	89.296	82.958	0.614	0.586	0.506	0.575	0.244	0.737	-0.312	2.567
			XGB	89.383	83.518	0.614	0.586	0.506	0.575	0.241	0.729	-0.337	1.504
		Test HB	DT	95.114	91.648	0.614	0.586	0.506	0.575	0.553	0.432	-0.167	inf
			LR	94.616	90.784	0.614	0.586	0.506	0.575	0.534	0.451	-0.113	inf
			RF	95.443	92.153	0.614	0.586	0.506	0.575	0.515	0.475	-0.061	inf
			XGB	95.455	92.296	0.614	0.586	0.506	0.575	0.544	0.434	-0.154	0.572
		Test O	DT	87.852	81.021	0.614	0.586	0.506	0.575	0.260	0.710	-0.423	2.206
			LR	87.774	80.833	0.614	0.586	0.506	0.575	0.241	0.730	-0.391	1.986
			RF	88.469	81.673	0.614	0.586	0.506	0.575	0.228	0.750	-0.350	2.406
			XGB	89.290	83.497	0.614	0.586	0.506	0.575	0.215	0.749	-0.369	1.298
	Train O	Test EB	DT	92.872	88.547	0.322	0.007	0.049	-0.060	-0.014	1.021	0.019	-0.058
			LR	91.998	87.037	0.322	0.007	0.049	-0.060	-0.006	1.010	0.008	0.021
			RF	94.069	90.480	0.322	0.007	0.049	-0.060	-0.036	1.048	0.044	-0.230
			XGB	93.836	90.137	0.322	0.007	0.049	-0.060	-0.035	1.046	0.042	-0.206

Continued on next page

Table 7.5 – Complete results for all datasets, training and testing variations, and models

Dataset	Train Variation	Test Variation	Model	F1-Score	Accuracy	CI	KL	KS	DD	DPPL	DI	DCA	DCR
		Test HB	DT	94.108	89.990	0.322	0.007	0.049	-0.060	0.271	0.706	0.309	-0.827
			LR	93.718	89.198	0.322	0.007	0.049	-0.060	0.253	0.727	0.326	-0.939
			RF	94.490	90.640	0.322	0.007	0.049	-0.060	0.264	0.711	0.316	-0.958
			XGB	94.444	90.566	0.322	0.007	0.049	-0.060	0.277	0.698	0.303	-0.903
		Test O	DT	92.546	88.500	0.322	0.007	0.049	-0.060	-0.052	1.070	0.002	-0.064
			LR	91.379	86.490	0.322	0.007	0.049	-0.060	-0.034	1.044	-0.021	0.027
			RF	93.123	89.387	0.322	0.007	0.049	-0.060	-0.068	1.091	0.020	-0.183
			XGB	93.372	89.809	0.322	0.007	0.049	-0.060	-0.069	1.092	0.022	-0.163
ADULT	Train EB	Test EB	DT	52.474	91.688	0.000	0.000	0.000	0.000	-0.042	2.044	1.280	-0.043
			LR	50.859	91.614	0.000	0.000	0.000	0.000	-0.030	1.667	0.964	-0.030
			RF	54.432	91.772	0.000	0.000	0.000	0.000	-0.022	1.381	0.507	-0.022
			XGB	58.975	92.799	0.000	0.000	0.000	0.000	-0.032	1.664	0.881	-0.033
		Test HB	DT	46.958	74.332	0.000	0.000	0.000	0.000	0.181	0.189	2.543	-0.616
			LR	45.279	74.055	0.000	0.000	0.000	0.000	0.176	0.160	2.585	-0.615

Continued on next page

Table 7.5 – Complete results for all datasets, training and testing variations, and models

Dataset	Train Variation	Test Variation	Model	F1-Score	Accuracy	CI	KL	KS	DD	DPPL	DI	DCA	DCR
			RF	53.072	76.465	0.000	0.000	0.000	0.000	0.226	0.138	1.976	-0.589
			XGB	54.687	77.260	0.000	0.000	0.000	0.000	0.233	0.124	1.864	-0.584
		Test O	DT	49.490	82.795	0.000	0.000	0.000	0.000	0.020	0.831	1.623	-0.198
			LR	47.448	82.420	0.000	0.000	0.000	0.000	0.026	0.744	1.547	-0.190
			RF	54.057	83.689	0.000	0.000	0.000	0.000	0.048	0.631	1.009	-0.172
			XGB	56.540	84.707	0.000	0.000	0.000	0.000	0.043	0.657	1.103	-0.177
	Train HB	Test EB	DT	37.296	78.178	-0.007	1.941	0.664	0.726	0.409	0.077	-3.085	0.678
			LR	35.743	77.955	-0.007	1.941	0.664	0.726	0.418	0.056	-4.313	0.685
			RF	37.920	79.641	-0.007	1.941	0.664	0.726	0.377	0.074	-3.509	0.584
			XGB	40.729	80.654	-0.007	1.941	0.664	0.726	0.365	0.085	-3.027	0.561
		Test HB	DT	86.228	89.940	-0.007	1.941	0.664	0.726	0.740	0.013	-1.595	0.270
			LR	87.005	90.468	-0.007	1.941	0.664	0.726	0.751	0.010	-2.616	0.314
			RF	86.420	90.243	-0.007	1.941	0.664	0.726	0.720	0.011	-2.131	0.167
			XGB	88.299	91.499	-0.007	1.941	0.664	0.726	0.734	0.011	-2.181	0.234

Continued on next page

Table 7.5 – Complete results for all datasets, training and testing variations, and models

Dataset	Train Variation	Test Variation	Model	F1-Score	Accuracy	CI	KL	KS	DD	DPPL	DI	DCA	DCR
		Test O	DT	63.749	77.576	-0.007	1.941	0.664	0.726	0.513	0.062	-2.774	0.613
			LR	63.911	77.718	-0.007	1.941	0.664	0.726	0.526	0.045	-4.005	0.631
			RF	64.847	79.002	-0.007	1.941	0.664	0.726	0.488	0.058	-3.187	0.522
			XGB	67.134	80.312	-0.007	1.941	0.664	0.726	0.484	0.066	-2.710	0.519
	Train O	Test EB	DT	55.038	90.239	0.338	0.141	0.196	0.238	0.060	0.575	-0.606	0.068
			LR	54.364	89.746	0.338	0.141	0.196	0.238	0.077	0.502	-0.718	0.088
			RF	56.497	89.909	0.338	0.141	0.196	0.238	0.078	0.520	-0.638	0.090
			XGB	60.656	91.303	0.338	0.141	0.196	0.238	0.065	0.554	-0.631	0.073
		Test HB	DT	71.818	83.418	0.338	0.141	0.196	0.238	0.407	0.078	0.889	-0.450
			LR	72.669	83.736	0.338	0.141	0.196	0.238	0.419	0.075	0.823	-0.438
			RF	74.052	84.285	0.338	0.141	0.196	0.238	0.434	0.079	0.830	-0.423
			XGB	76.954	85.971	0.338	0.141	0.196	0.238	0.455	0.062	0.631	-0.400
		Test O	DT	66.308	85.558	0.338	0.141	0.196	0.238	0.165	0.322	-0.151	-0.048
			LR	66.193	85.208	0.338	0.141	0.196	0.238	0.179	0.300	-0.240	-0.031

Continued on next page

Table 7.5 – Complete results for all datasets, training and testing variations, and models

Dataset	Train Variation	Test Variation	Model	F1-Score	Accuracy	CI	KL	KS	DD	DPPL	DI	DCA	DCR
			RF	67.417	85.420	0.338	0.141	0.196	0.238	0.184	0.312	-0.177	-0.024
			XGB	71.330	87.334	0.338	0.141	0.196	0.238	0.182	0.303	-0.224	-0.027
BANK	Train EB	Test EB	DT	51.469	88.654	0.000	0.000	0.000	0.000	0.002	0.979	-0.033	0.002
			LR	50.172	88.940	0.000	0.000	0.000	0.000	-0.001	1.010	0.006	-0.001
			RF	31.481	87.781	0.000	0.000	0.000	0.000	0.003	0.935	-0.340	0.002
			XGB	57.362	89.476	0.000	0.000	0.000	0.000	-0.001	1.008	0.008	-0.001
		Test HB	DT	51.602	88.792	0.000	0.000	0.000	0.000	0.134	0.274	1.336	-0.235
			LR	50.012	88.974	0.000	0.000	0.000	0.000	0.125	0.251	1.387	-0.242
			RF	32.856	88.085	0.000	0.000	0.000	0.000	0.069	0.199	2.273	-0.285
			XGB	56.930	89.398	0.000	0.000	0.000	0.000	0.158	0.266	1.119	-0.211
		Test O	DT	49.879	89.943	0.000	0.000	0.000	0.000	-0.014	1.178	-0.236	0.029
			LR	48.895	90.286	0.000	0.000	0.000	0.000	-0.014	1.208	-0.237	0.028
			RF	31.394	89.660	0.000	0.000	0.000	0.000	-0.004	1.129	-0.823	0.037
			XGB	55.371	90.513	0.000	0.000	0.000	0.000	-0.018	1.205	-0.179	0.024

Continued on next page

Table 7.5 – Complete results for all datasets, training and testing variations, and models

Dataset	Train Variation	Test Variation	Model	F1-Score	Accuracy	CI	KL	KS	DD	DPPL	DI	DCA	DCR
	Train HB	Test EB	DT	37.080	86.784	-0.352	0.613	0.329	0.605	0.133	0.024	-inf	0.133
			LR	48.952	87.851	-0.352	0.613	0.329	0.605	0.147	0.141	-5.152	0.156
			RF	37.169	87.420	-0.352	0.613	0.329	0.605	0.115	0.019	-97.771	0.113
			XGB	47.779	86.951	-0.352	0.613	0.329	0.605	0.198	0.049	-14.215	0.217
		Test HB	DT	58.426	91.003	-0.352	0.613	0.329	0.605	0.237	0.008	-inf	-0.129
			LR	68.815	92.370	-0.352	0.613	0.329	0.605	0.295	0.036	-1.814	-0.057
			RF	57.062	90.986	-0.352	0.613	0.329	0.605	0.222	0.004	-inf	-0.145
			XGB	72.797	93.031	-0.352	0.613	0.329	0.605	0.352	0.012	-6.900	0.021
		Test O	DT	35.607	87.844	-0.352	0.613	0.329	0.605	0.114	0.027	-inf	0.157
			LR	45.999	88.341	-0.352	0.613	0.329	0.605	0.125	0.162	-5.296	0.175
			RF	35.810	88.560	-0.352	0.613	0.329	0.605	0.099	0.022	-97.969	0.139
			XGB	44.343	87.125	-0.352	0.613	0.329	0.605	0.173	0.056	-14.340	0.232
	Train O	Test EB	DT	43.896	88.437	0.204	0.007	0.040	-0.092	0.002	0.983	-0.061	0.002
			LR	46.059	88.604	0.204	0.007	0.040	-0.092	-0.018	1.296	0.519	-0.018

Continued on next page

Table 7.5 – Complete results for all datasets, training and testing variations, and models

Dataset	Train Variation	Test Variation	Model	F1-Score	Accuracy	CI	KL	KS	DD	DPPL	DI	DCA	DCR
			RF	26.695	87.556	0.204	0.007	0.040	-0.092	-0.003	1.123	0.395	-0.003
			XGB	52.849	89.120	0.204	0.007	0.040	-0.092	-0.016	1.204	0.290	-0.017
		Test HB	DT	44.635	88.644	0.204	0.007	0.040	-0.092	0.103	0.247	1.682	-0.260
			LR	42.786	88.132	0.204	0.007	0.040	-0.092	0.088	0.321	2.029	-0.276
			RF	26.847	87.696	0.204	0.007	0.040	-0.092	0.051	0.212	3.143	-0.297
			XGB	50.553	88.691	0.204	0.007	0.040	-0.092	0.117	0.314	1.531	-0.251
		Test O	DT	43.099	90.029	0.204	0.007	0.040	-0.092	-0.010	1.200	-0.322	0.032
			LR	45.072	90.159	0.204	0.007	0.040	-0.092	-0.029	1.572	0.225	0.013
			RF	26.754	89.538	0.204	0.007	0.040	-0.092	-0.008	1.342	-0.352	0.034
			XGB	51.529	90.482	0.204	0.007	0.040	-0.092	-0.031	1.471	0.081	0.011
COMPAS	Train EB	Test EB	DT	80.011	71.270	0.000	0.000	0.000	0.000	0.036	0.959	-0.035	0.355
			LR	80.485	71.764	0.000	0.000	0.000	0.000	-0.006	1.007	0.005	-0.074
			RF	80.664	70.813	0.000	0.000	0.000	0.000	-0.016	1.019	0.014	-0.354
			XGB	80.620	69.702	0.000	0.000	0.000	0.000	0.025	0.973	-0.018	2.133

Continued on next page

Table 7.5 – Complete results for all datasets, training and testing variations, and models

Dataset	Train Variation	Test Variation	Model	F1-Score	Accuracy	CI	KL	KS	DD	DPPL	DI	DCA	DCR
		Test HB	DT	80.358	72.246	0.000	0.000	0.000	0.000	0.196	0.777	0.570	-1.105
			LR	79.796	71.316	0.000	0.000	0.000	0.000	0.161	0.813	0.596	-1.390
			RF	79.745	69.852	0.000	0.000	0.000	0.000	0.117	0.871	0.588	-1.875
			XGB	80.468	69.898	0.000	0.000	0.000	0.000	0.128	0.868	0.552	0.701
		Test O	DT	79.414	70.630	0.000	0.000	0.000	0.000	0.007	0.995	-0.134	0.591
			LR	79.975	71.217	0.000	0.000	0.000	0.000	-0.033	1.042	-0.095	0.182
			RF	80.250	70.401	0.000	0.000	0.000	0.000	-0.042	1.049	-0.083	-0.162
			XGB	80.006	68.912	0.000	0.000	0.000	0.000	0.006	0.993	-0.115	2.043
	Train HB	Test EB	DT	67.484	60.310	0.168	0.727	0.577	0.600	0.708	0.250	-2.190	6.287
			LR	65.758	58.886	0.168	0.727	0.577	0.600	0.756	0.202	-2.688	6.470
			RF	71.814	63.798	0.168	0.727	0.577	0.600	0.598	0.372	-1.158	7.329
			XGB	78.353	69.340	0.168	0.727	0.577	0.600	0.339	0.644	-0.368	7.145
		Test HB	DT	85.584	81.816	0.168	0.727	0.577	0.600	0.806	0.169	-0.937	3.959
			LR	85.309	81.680	0.168	0.727	0.577	0.600	0.846	0.129	-1.409	4.363

Continued on next page

Table 7.5 – Complete results for all datasets, training and testing variations, and models

Dataset	Train Variation	Test Variation	Model	F1-Score	Accuracy	CI	KL	KS	DD	DPPL	DI	DCA	DCR
			RF	86.141	81.924	0.168	0.727	0.577	0.600	0.722	0.262	-0.247	7.013
			XGB	85.321	79.455	0.168	0.727	0.577	0.600	0.492	0.497	0.297	6.867
		Test O	DT	63.899	56.409	0.168	0.727	0.577	0.600	0.692	0.260	-2.366	6.383
			LR	62.056	54.946	0.168	0.727	0.577	0.600	0.738	0.212	-2.821	6.436
			RF	68.974	60.449	0.168	0.727	0.577	0.600	0.577	0.389	-1.267	7.234
			XGB	76.559	66.973	0.168	0.727	0.577	0.600	0.317	0.665	-0.470	7.054
	Train O	Test EB	DT	78.728	70.813	-0.011	0.023	0.102	-0.109	-0.117	1.175	0.139	-0.661
			LR	80.028	71.650	-0.011	0.023	0.102	-0.109	-0.133	1.185	0.137	-1.217
			RF	80.677	71.593	-0.011	0.023	0.102	-0.109	-0.092	1.117	0.083	-1.421
			XGB	80.650	70.149	-0.011	0.023	0.102	-0.109	-0.004	1.005	0.003	0.126
		Test HB	DT	75.548	67.170	-0.011	0.023	0.102	-0.109	0.060	0.921	0.746	-1.741
			LR	76.598	67.368	-0.011	0.023	0.102	-0.109	0.038	0.952	0.718	-2.227
			RF	78.294	68.542	-0.011	0.023	0.102	-0.109	0.059	0.930	0.656	-2.529
			XGB	79.931	69.441	-0.011	0.023	0.102	-0.109	0.112	0.882	0.572	-1.316

Continued on next page

Table 7.5 – Complete results for all datasets, training and testing variations, and models

Dataset	Train Variation	Test Variation	Model	F1-Score	Accuracy	CI	KL	KS	DD	DPPL	DI	DCA	DCR
		Test O	DT	78.667	70.970	-0.011	0.023	0.102	-0.109	-0.146	1.224	0.032	-0.433
			LR	79.986	71.813	-0.011	0.023	0.102	-0.109	-0.161	1.230	0.034	-0.984
			RF	80.520	71.591	-0.011	0.023	0.102	-0.109	-0.120	1.156	-0.015	-1.210
			XGB	80.140	69.550	-0.011	0.023	0.102	-0.109	-0.024	1.028	-0.094	0.276
DIAB	Train EB	Test EB	DT	91.098	83.962	0.000	0.000	0.000	0.000	0.006	0.993	-0.006	1.324
			LR	90.987	83.946	0.000	0.000	0.000	0.000	0.008	0.991	-0.008	0.512
			RF	91.217	84.168	0.000	0.000	0.000	0.000	0.014	0.986	-0.012	2.433
			XGB	91.235	84.235	0.000	0.000	0.000	0.000	0.014	0.986	-0.012	1.929
		Test HB	DT	91.037	83.897	0.000	0.000	0.000	0.000	0.050	0.949	0.393	-4.572
			LR	90.978	83.972	0.000	0.000	0.000	0.000	0.072	0.926	0.384	-3.200
			RF	91.293	84.347	0.000	0.000	0.000	0.000	0.060	0.939	0.385	-3.422
			XGB	91.307	84.410	0.000	0.000	0.000	0.000	0.065	0.933	0.383	-3.240
		Test O	DT	91.807	85.120	0.000	0.000	0.000	0.000	0.004	0.996	-0.028	1.648
			LR	91.638	84.981	0.000	0.000	0.000	0.000	0.005	0.995	-0.029	0.760

Continued on next page

Table 7.5 – Complete results for all datasets, training and testing variations, and models

Dataset	Train Variation	Test Variation	Model	F1-Score	Accuracy	CI	KL	KS	DD	DPPL	DI	DCA	DCR
			RF	91.903	85.290	0.000	0.000	0.000	0.000	0.011	0.989	-0.034	2.774
			XGB	91.908	85.332	0.000	0.000	0.000	0.000	0.011	0.989	-0.034	2.234
	Train HB	Test EB	DT	87.234	78.692	0.367	0.449	0.416	0.638	0.329	0.671	-0.409	inf
			LR	87.754	79.487	0.367	0.449	0.416	0.638	0.315	0.685	-0.384	159.651
			RF	88.388	80.369	0.367	0.449	0.416	0.638	0.286	0.714	-0.335	inf
			XGB	87.804	79.573	0.367	0.449	0.416	0.638	0.318	0.682	-0.388	inf
		Test HB	DT	93.411	88.894	0.367	0.449	0.416	0.638	0.458	0.542	-0.047	inf
			LR	93.632	89.260	0.367	0.449	0.416	0.638	0.452	0.547	-0.037	59.556
			RF	93.679	89.271	0.367	0.449	0.416	0.638	0.421	0.579	0.018	inf
			XGB	93.725	89.422	0.367	0.449	0.416	0.638	0.458	0.542	-0.046	inf
		Test O	DT	86.886	78.130	0.367	0.449	0.416	0.638	0.319	0.681	-0.426	inf
			LR	87.464	79.005	0.367	0.449	0.416	0.638	0.305	0.695	-0.401	159.711
			RF	88.238	80.097	0.367	0.449	0.416	0.638	0.276	0.724	-0.351	inf
			XGB	87.516	79.092	0.367	0.449	0.416	0.638	0.307	0.693	-0.404	inf

Continued on next page

Table 7.5 – Complete results for all datasets, training and testing variations, and models

Dataset	Train Variation	Test Variation	Model	F1-Score	Accuracy	CI	KL	KS	DD	DPPL	DI	DCA	DCR
	Train O	Test EB	DT	91.143	83.985	-0.122	0.002	0.024	-0.045	0.007	0.993	-0.006	1.738
			LR	91.031	83.953	-0.122	0.002	0.024	-0.045	-0.007	1.008	0.007	-0.606
			RF	91.185	84.047	-0.122	0.002	0.024	-0.045	0.007	0.993	-0.006	2.064
			XGB	91.232	84.164	-0.122	0.002	0.024	-0.045	0.005	0.995	-0.004	1.111
		Test HB	DT	91.081	83.919	-0.122	0.002	0.024	-0.045	0.044	0.955	0.395	-5.123
			LR	90.733	83.458	-0.122	0.002	0.024	-0.045	0.045	0.953	0.401	-4.724
			RF	91.132	83.989	-0.122	0.002	0.024	-0.045	0.043	0.957	0.395	-5.072
			XGB	91.141	84.039	-0.122	0.002	0.024	-0.045	0.046	0.953	0.394	-4.868
		Test O	DT	91.855	85.158	-0.122	0.002	0.024	-0.045	0.004	0.995	-0.028	2.159
			LR	91.715	85.051	-0.122	0.002	0.024	-0.045	-0.010	1.010	-0.016	-0.258
			RF	91.900	85.226	-0.122	0.002	0.024	-0.045	0.005	0.995	-0.028	2.491
			XGB	91.931	85.310	-0.122	0.002	0.024	-0.045	0.003	0.997	-0.027	1.532
EMP	Train EB	Test EB	DT	82.374	77.536	0.000	0.000	0.000	0.000	-0.002	1.006	0.002	-0.014
			LR	70.717	63.472	0.000	0.000	0.000	0.000	0.020	0.973	-0.028	0.067

Continued on next page

Table 7.5 – Complete results for all datasets, training and testing variations, and models

Dataset	Train Variation	Test Variation	Model	F1-Score	Accuracy	CI	KL	KS	DD	DPPL	DI	DCA	DCR
			RF	82.164	77.670	0.000	0.000	0.000	0.000	0.026	0.965	-0.032	0.125
			XGB	81.689	76.941	0.000	0.000	0.000	0.000	0.025	0.967	-0.029	0.115
		Test HB	DT	85.169	79.832	0.000	0.000	0.000	0.000	0.387	0.563	0.698	-0.935
			LR	74.539	66.648	0.000	0.000	0.000	0.000	0.205	0.733	0.898	-1.575
			RF	85.963	81.261	0.000	0.000	0.000	0.000	0.428	0.512	0.668	-0.847
			XGB	85.540	80.520	0.000	0.000	0.000	0.000	0.413	0.533	0.679	-0.844
		Test O	DT	84.400	79.162	0.000	0.000	0.000	0.000	0.098	0.874	0.147	-0.164
			LR	73.358	65.340	0.000	0.000	0.000	0.000	0.068	0.906	0.182	-0.350
			RF	84.540	79.632	0.000	0.000	0.000	0.000	0.132	0.830	0.107	-0.018
			XGB	84.161	78.981	0.000	0.000	0.000	0.000	0.127	0.838	0.115	-0.018
	Train HB	Test EB	DT	62.852	61.198	0.218	1.364	0.749	0.761	0.719	0.145	-6.592	2.283
			LR	53.553	50.273	0.218	1.364	0.749	0.761	0.993	0.007	-inf	inf
			RF	61.520	60.003	0.218	1.364	0.749	0.761	0.783	0.094	-8.264	2.830
			XGB	62.814	60.826	0.218	1.364	0.749	0.761	0.743	0.135	-5.078	2.720

Continued on next page

Table 7.5 – Complete results for all datasets, training and testing variations, and models

Dataset	Train Variation	Test Variation	Model	F1-Score	Accuracy	CI	KL	KS	DD	DPPL	DI	DCA	DCR
		Test HB	DT	91.263	89.148	0.218	1.364	0.749	0.761	0.884	0.080	-2.868	1.772
			LR	90.612	88.407	0.218	1.364	0.749	0.761	0.996	0.004	-inf	inf
			RF	92.037	90.154	0.218	1.364	0.749	0.761	0.935	0.044	-inf	3.822
			XGB	91.484	89.412	0.218	1.364	0.749	0.761	0.907	0.067	-2.396	2.614
		Test O	DT	72.662	68.781	0.218	1.364	0.749	0.761	0.769	0.128	-6.361	2.116
			LR	66.103	60.600	0.218	1.364	0.749	0.761	0.994	0.006	-inf	inf
			RF	72.080	68.200	0.218	1.364	0.749	0.761	0.829	0.082	-8.014	2.764
			XGB	72.770	68.670	0.218	1.364	0.749	0.761	0.790	0.120	-4.727	2.603
	Train O	Test EB	DT	82.512	78.130	0.106	0.072	0.187	0.194	0.138	0.819	-0.172	0.664
			LR	70.826	62.606	0.106	0.072	0.187	0.194	0.364	0.595	-0.442	3.701
			RF	82.986	78.355	0.106	0.072	0.187	0.194	0.139	0.822	-0.161	0.762
			XGB	83.005	78.812	0.106	0.072	0.187	0.194	0.122	0.837	-0.155	0.551
		Test HB	DT	88.703	84.966	0.106	0.072	0.187	0.194	0.537	0.413	0.554	-0.104
			LR	84.468	78.982	0.106	0.072	0.187	0.194	0.525	0.437	0.564	-0.161

Continued on next page

Table 7.5 – Complete results for all datasets, training and testing variations, and models

Dataset	Train Variation	Test Variation	Model	F1-Score	Accuracy	CI	KL	KS	DD	DPPL	DI	DCA	DCR
			RF	88.968	85.125	0.106	0.072	0.187	0.194	0.536	0.424	0.557	-0.147
			XGB	89.067	85.444	0.106	0.072	0.187	0.194	0.538	0.412	0.552	-0.339
		Test O	DT	85.239	80.610	0.106	0.072	0.187	0.194	0.234	0.712	-0.024	0.506
			LR	76.030	67.910	0.106	0.072	0.187	0.194	0.402	0.559	-0.242	2.631
			RF	85.653	80.825	0.106	0.072	0.187	0.194	0.231	0.723	-0.009	0.642
			XGB	85.819	81.368	0.106	0.072	0.187	0.194	0.226	0.721	-0.015	0.452
GLIOMA	Train EB	Test EB	DT	70.375	78.929	0.000	0.000	0.000	0.000	-0.100	inf	inf	-0.081
			LR	74.571	81.429	0.000	0.000	0.000	0.000	0.018	1.273	0.210	0.098
			RF	71.000	80.179	0.000	0.000	0.000	0.000	-0.093	1.833	0.750	-0.104
			XGB	61.706	72.143	0.000	0.000	0.000	0.000	0.043	1.200	-0.100	0.075
		Test HB	DT	72.907	84.223	0.000	0.000	0.000	0.000	0.151	inf	inf	-0.794
			LR	78.561	87.643	0.000	0.000	0.000	0.000	0.266	inf	inf	-0.768
			RF	79.463	88.811	0.000	0.000	0.000	0.000	0.172	inf	inf	-0.803
			XGB	73.766	85.132	0.000	0.000	0.000	0.000	0.202	inf	inf	inf

Continued on next page

Table 7.5 – Complete results for all datasets, training and testing variations, and models

Dataset	Train Variation	Test Variation	Model	F1-Score	Accuracy	CI	KL	KS	DD	DPPL	DI	DCA	DCR
		Test O	DT	83.182	81.974	0.000	0.000	0.000	0.000	-0.242	inf	inf	0.027
			LR	86.429	85.191	0.000	0.000	0.000	0.000	-0.178	1.896	0.332	0.133
			RF	86.788	85.911	0.000	0.000	0.000	0.000	-0.273	2.637	0.940	-0.031
			XGB	83.823	82.562	0.000	0.000	0.000	0.000	-0.157	1.972	0.397	0.151
	Train HB	Test EB	DT	68.608	70.714	-0.827	0.711	0.564	0.221	0.450	0.435	-0.990	inf
			LR	70.826	76.429	-0.827	0.711	0.564	0.221	0.229	1.023	-0.285	0.745
			RF	70.913	75.625	-0.827	0.711	0.564	0.221	0.305	0.779	-0.657	inf
			XGB	68.625	71.518	-0.827	0.711	0.564	0.221	0.441	0.498	-0.937	inf
		Test HB	DT	82.831	90.180	-0.827	0.711	0.564	0.221	0.607	0.322	0.232	inf
			LR	81.468	89.709	-0.827	0.711	0.564	0.221	0.468	inf	inf	inf
			RF	82.599	90.391	-0.827	0.711	0.564	0.221	0.551	inf	inf	inf
			XGB	79.957	88.795	-0.827	0.711	0.564	0.221	0.661	0.309	0.124	inf
		Test O	DT	83.689	82.216	-0.827	0.711	0.564	0.221	0.261	0.678	-0.650	inf
			LR	84.504	83.405	-0.827	0.711	0.564	0.221	0.043	1.411	-0.046	0.814

Continued on next page

Table 7.5 – Complete results for all datasets, training and testing variations, and models

Dataset	Train Variation	Test Variation	Model	F1-Score	Accuracy	CI	KL	KS	DD	DPPL	DI	DCA	DCR
			RF	85.256	84.122	-0.827	0.711	0.564	0.221	0.101	1.172	-0.263	inf
			XGB	81.809	80.548	-0.827	0.711	0.564	0.221	0.274	0.727	-0.652	inf
	Train O	Test EB	DT	77.039	84.286	-0.823	0.074	0.191	-0.063	-0.039	inf	inf	-0.019
			LR	77.867	83.661	-0.823	0.074	0.191	-0.063	0.002	1.425	0.350	0.043
			RF	77.476	83.125	-0.823	0.074	0.191	-0.063	0.016	1.435	0.235	0.086
			XGB	74.677	80.536	-0.823	0.074	0.191	-0.063	0.018	1.158	-0.050	0.032
		Test HB	DT	81.744	89.715	-0.823	0.074	0.191	-0.063	0.273	inf	inf	-0.756
			LR	81.388	89.022	-0.823	0.074	0.191	-0.063	0.323	inf	inf	-0.783
			RF	82.410	89.942	-0.823	0.074	0.191	-0.063	0.338	inf	inf	-0.760
			XGB	79.409	87.659	-0.823	0.074	0.191	-0.063	0.285	0.630	1.122	-0.858
		Test O	DT	88.105	87.104	-0.823	0.074	0.191	-0.063	-0.235	inf	inf	0.026
			LR	88.239	86.984	-0.823	0.074	0.191	-0.063	-0.194	2.078	0.454	0.072
			RF	88.155	86.987	-0.823	0.074	0.191	-0.063	-0.157	1.985	0.393	0.169
			XGB	87.316	85.909	-0.823	0.074	0.191	-0.063	-0.174	1.677	0.165	0.083

Continued on next page

Table 7.5 – Complete results for all datasets, training and testing variations, and models

Dataset	Train Variation	Test Variation	Model	F1-Score	Accuracy	CI	KL	KS	DD	DPPL	DI	DCA	DCR
HEART	Train EB	Test EB	DT	42.510	72.833	0.000	0.000	0.000	0.000	-0.143	inf	inf	-0.215
			LR	67.810	86.944	0.000	0.000	0.000	0.000	-0.070	inf	inf	-0.079
			RF	66.857	86.556	0.000	0.000	0.000	0.000	-0.060	inf	inf	-0.066
			XGB	62.833	84.944	0.000	0.000	0.000	0.000	-0.030	1.350	0.283	-0.035
		Test HB	DT	63.605	66.126	0.000	0.000	0.000	0.000	0.239	0.542	1.568	-0.916
			LR	72.545	75.931	0.000	0.000	0.000	0.000	0.435	0.167	-inf	-0.636
			RF	69.479	74.113	0.000	0.000	0.000	0.000	0.413	0.159	-inf	-0.668
			XGB	71.475	75.476	0.000	0.000	0.000	0.000	0.437	0.114	-inf	-0.648
		Test O	DT	62.261	70.516	0.000	0.000	0.000	0.000	0.054	0.919	0.780	-0.416
			LR	73.614	80.108	0.000	0.000	0.000	0.000	0.148	0.649	0.189	-0.249
			RF	71.259	79.097	0.000	0.000	0.000	0.000	0.134	0.676	0.317	-0.266
			XGB	71.333	78.763	0.000	0.000	0.000	0.000	0.180	0.501	-0.031	-0.218
	Train HB	Test EB	DT	47.747	65.556	0.280	2.238	0.806	0.754	0.542	0.219	-inf	1.792
			LR	57.807	73.389	0.280	2.238	0.806	0.754	0.492	0.232	-inf	1.753

Continued on next page

Table 7.5 – Complete results for all datasets, training and testing variations, and models

Dataset	Train Variation	Test Variation	Model	F1-Score	Accuracy	CI	KL	KS	DD	DPPL	DI	DCA	DCR
			RF	56.245	71.778	0.280	2.238	0.806	0.754	0.564	0.186	-inf	2.085
			XGB	44.192	66.556	0.280	2.238	0.806	0.754	0.629	0.068	-inf	1.958
		Test HB	DT	82.674	80.606	0.280	2.238	0.806	0.754	0.763	0.120	-inf	inf
			LR	90.249	88.874	0.280	2.238	0.806	0.754	0.875	0.028	-inf	inf
			RF	91.616	90.281	0.280	2.238	0.806	0.754	0.880	0.037	-inf	inf
			XGB	87.962	86.515	0.280	2.238	0.806	0.754	0.892	0.000	-inf	inf
		Test O	DT	70.555	70.151	0.280	2.238	0.806	0.754	0.609	0.193	-inf	0.997
			LR	79.532	79.118	0.280	2.238	0.806	0.754	0.623	0.180	-inf	1.350
			RF	78.562	77.796	0.280	2.238	0.806	0.754	0.674	0.149	-inf	1.561
			XGB	73.516	73.505	0.280	2.238	0.806	0.754	0.737	0.054	-inf	1.477
	Train O	Test EB	DT	59.202	77.056	0.364	0.208	0.303	0.265	0.041	1.090	-0.012	0.092
			LR	70.725	85.944	0.364	0.208	0.303	0.265	0.094	0.793	-0.502	0.145
			RF	66.455	84.333	0.364	0.208	0.303	0.265	0.044	1.110	-0.002	0.082
			XGB	68.167	84.444	0.364	0.208	0.303	0.265	0.084	0.827	-0.368	0.133

Continued on next page

Table 7.5 – Complete results for all datasets, training and testing variations, and models

Dataset	Train Variation	Test Variation	Model	F1-Score	Accuracy	CI	KL	KS	DD	DPPL	DI	DCA	DCR
		Test HB	DT	80.301	79.632	0.364	0.208	0.303	0.265	0.503	0.257	-inf	-0.577
			LR	83.429	83.788	0.364	0.208	0.303	0.265	0.592	0.097	-inf	-0.428
			RF	81.919	81.948	0.364	0.208	0.303	0.265	0.583	0.130	-inf	-0.469
			XGB	85.292	85.195	0.364	0.208	0.303	0.265	0.599	0.139	-inf	-0.413
		Test O	DT	76.196	78.473	0.364	0.208	0.303	0.265	0.248	0.559	0.079	-0.066
			LR	79.998	83.097	0.364	0.208	0.303	0.265	0.300	0.403	-0.362	-0.029
			RF	79.903	82.763	0.364	0.208	0.303	0.265	0.275	0.465	-0.191	-0.061
			XGB	80.893	83.419	0.364	0.208	0.303	0.265	0.318	0.403	-0.344	0.020
ILPD	Train EB	Test EB	DT	70.841	60.714	0.000	0.000	0.000	0.000	0.029	0.971	-0.043	0.762
			LR	78.247	69.643	0.000	0.000	0.000	0.000	-0.107	1.180	0.119	-1.567
			RF	75.327	65.357	0.000	0.000	0.000	0.000	0.021	0.986	-0.030	0.304
			XGB	78.768	65.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000	NaN
		Test HB	DT	81.972	72.640	0.000	0.000	0.000	0.000	0.076	0.910	0.811	-inf
			LR	84.037	74.579	0.000	0.000	0.000	0.000	0.017	0.989	0.800	-inf

Continued on next page

Table 7.5 – Complete results for all datasets, training and testing variations, and models

Dataset	Train Variation	Test Variation	Model	F1-Score	Accuracy	CI	KL	KS	DD	DPPL	DI	DCA	DCR
			RF	84.276	74.579	0.000	0.000	0.000	0.000	0.116	0.864	0.728	-2.337
			XGB	90.034	81.902	0.000	0.000	0.000	0.000	0.000	1.000	0.652	NaN
		Test O	DT	76.428	66.667	0.000	0.000	0.000	0.000	0.025	0.961	0.059	-0.487
			LR	79.989	70.526	0.000	0.000	0.000	0.000	-0.061	1.098	0.167	-1.666
			RF	78.724	68.246	0.000	0.000	0.000	0.000	0.050	0.940	0.034	-0.111
			XGB	83.194	71.228	0.000	0.000	0.000	0.000	0.000	1.000	0.083	NaN
	Train HB	Test EB	DT	64.179	55.000	0.663	0.972	0.653	0.631	0.714	0.262	-2.814	inf
			LR	68.010	59.643	0.663	0.972	0.653	0.631	0.764	0.236	-2.617	inf
			RF	69.196	60.000	0.663	0.972	0.653	0.631	0.700	0.300	-1.672	inf
			XGB	78.768	65.000	0.663	0.972	0.653	0.631	0.000	1.000	0.000	NaN
		Test HB	DT	91.569	86.061	0.663	0.972	0.653	0.631	0.706	0.265	-inf	inf
			LR	93.853	89.726	0.663	0.972	0.653	0.631	0.881	0.119	-inf	inf
			RF	93.766	89.482	0.663	0.972	0.653	0.631	0.783	0.217	-inf	inf
			XGB	90.034	81.902	0.663	0.972	0.653	0.631	0.000	1.000	0.652	NaN

Continued on next page

Table 7.5 – Complete results for all datasets, training and testing variations, and models

Dataset	Train Variation	Test Variation	Model	F1-Score	Accuracy	CI	KL	KS	DD	DPPL	DI	DCA	DCR
		Test O	DT	76.882	65.439	0.663	0.972	0.653	0.631	0.708	0.262	-2.724	inf
			LR	79.382	68.596	0.663	0.972	0.653	0.631	0.764	0.236	-2.534	inf
			RF	79.720	68.772	0.663	0.972	0.653	0.631	0.700	0.300	-1.589	inf
			XGB	83.194	71.228	0.663	0.972	0.653	0.631	0.000	1.000	0.083	NaN
	Train O	Test EB	DT	70.225	58.929	0.509	0.016	0.083	0.075	0.007	0.999	-0.033	-inf
			LR	76.176	64.643	0.509	0.016	0.083	0.075	0.107	0.886	-0.111	inf
			RF	74.212	62.500	0.509	0.016	0.083	0.075	0.150	0.844	-0.188	inf
			XGB	78.768	65.000	0.509	0.016	0.083	0.075	0.000	1.000	0.000	NaN
		Test HB	DT	82.345	72.384	0.509	0.016	0.083	0.075	0.079	0.907	0.795	NaN
			LR	89.609	82.134	0.509	0.016	0.083	0.075	0.168	0.823	0.607	-inf
			RF	88.515	80.683	0.509	0.016	0.083	0.075	0.193	0.798	0.616	-inf
			XGB	90.034	81.902	0.509	0.016	0.083	0.075	0.000	1.000	0.652	NaN
		Test O	DT	76.494	65.263	0.509	0.016	0.083	0.075	0.033	0.962	0.019	-inf
			LR	82.413	72.105	0.509	0.016	0.083	0.075	0.124	0.866	-0.037	1.623

Continued on next page

Table 7.5 – Complete results for all datasets, training and testing variations, and models

Dataset	Train Variation	Test Variation	Model	F1-Score	Accuracy	CI	KL	KS	DD	DPPL	DI	DCA	DCR
			RF	80.232	68.947	0.509	0.016	0.083	0.075	0.167	0.822	-0.114	1.666
			XGB	83.194	71.228	0.509	0.016	0.083	0.075	0.000	1.000	0.083	NaN
IBD	Train EB	Test EB	DT	87.985	89.989	0.000	0.000	0.000	0.000	-0.014	1.035	0.034	-0.024
			LR	92.093	93.344	0.000	0.000	0.000	0.000	0.006	0.987	-0.014	0.009
			RF	92.702	93.895	0.000	0.000	0.000	0.000	-0.006	1.016	0.014	-0.011
			XGB	93.223	94.310	0.000	0.000	0.000	0.000	0.003	0.992	-0.010	0.005
		Test HB	DT	88.062	89.402	0.000	0.000	0.000	0.000	0.419	0.383	0.295	-0.292
			LR	92.151	92.975	0.000	0.000	0.000	0.000	0.484	0.330	0.149	-0.142
			RF	92.620	93.426	0.000	0.000	0.000	0.000	0.473	0.337	0.172	-0.175
			XGB	93.413	94.120	0.000	0.000	0.000	0.000	0.487	0.326	0.140	-0.135
		Test O	DT	88.839	89.427	0.000	0.000	0.000	0.000	-0.162	1.439	-0.051	0.075
			LR	92.498	92.800	0.000	0.000	0.000	0.000	-0.157	1.408	-0.072	0.079
			RF	93.140	93.445	0.000	0.000	0.000	0.000	-0.171	1.461	-0.035	0.055
			XGB	93.702	93.955	0.000	0.000	0.000	0.000	-0.166	1.438	-0.051	0.062

Continued on next page

Table 7.5 – Complete results for all datasets, training and testing variations, and models

Dataset	Train Variation	Test Variation	Model	F1-Score	Accuracy	CI	KL	KS	DD	DPPL	DI	DCA	DCR
	Train HB	Test EB	DT	85.849	88.328	-0.133	0.668	0.541	0.538	0.129	0.725	-0.348	0.211
			LR	90.234	91.829	-0.133	0.668	0.541	0.538	0.104	0.778	-0.262	0.175
			RF	90.603	92.312	-0.133	0.668	0.541	0.538	0.102	0.772	-0.280	0.162
			XGB	92.034	93.379	-0.133	0.668	0.541	0.538	0.080	0.823	-0.206	0.132
		Test HB	DT	90.434	91.570	-0.133	0.668	0.541	0.538	0.545	0.267	-0.052	-0.007
			LR	93.145	93.912	-0.133	0.668	0.541	0.538	0.570	0.255	-0.109	0.078
			RF	93.901	94.675	-0.133	0.668	0.541	0.538	0.563	0.245	-0.152	0.010
			XGB	94.656	95.282	-0.133	0.668	0.541	0.538	0.552	0.263	-0.076	0.008
		Test O	DT	85.230	86.309	-0.133	0.668	0.541	0.538	-0.014	1.032	-0.411	0.323
			LR	89.941	90.473	-0.133	0.668	0.541	0.538	-0.057	1.134	-0.298	0.257
			RF	89.867	90.609	-0.133	0.668	0.541	0.538	-0.058	1.142	-0.303	0.250
			XGB	91.629	92.100	-0.133	0.668	0.541	0.538	-0.085	1.208	-0.235	0.208
	Train O	Test EB	DT	88.064	89.742	-0.102	0.076	0.194	-0.192	-0.060	1.148	0.131	-0.111
			LR	91.753	92.929	-0.102	0.076	0.194	-0.192	-0.036	1.087	0.080	-0.065

Continued on next page

Table 7.5 – Complete results for all datasets, training and testing variations, and models

Dataset	Train Variation	Test Variation	Model	F1-Score	Accuracy	CI	KL	KS	DD	DPPL	DI	DCA	DCR
			RF	92.643	93.693	-0.102	0.076	0.194	-0.192	-0.042	1.102	0.094	-0.077
			XGB	93.189	94.186	-0.102	0.076	0.194	-0.192	-0.033	1.080	0.074	-0.059
		Test HB	DT	86.842	87.962	-0.102	0.076	0.194	-0.192	0.376	0.450	0.413	-0.356
			LR	90.757	91.570	-0.102	0.076	0.194	-0.192	0.439	0.385	0.285	-0.221
			RF	91.544	92.281	-0.102	0.076	0.194	-0.192	0.439	0.387	0.289	-0.215
			XGB	92.365	93.062	-0.102	0.076	0.194	-0.192	0.452	0.371	0.252	-0.192
		Test O	DT	90.001	90.227	-0.102	0.076	0.194	-0.192	-0.214	1.577	0.043	-0.040
			LR	92.840	93.000	-0.102	0.076	0.194	-0.192	-0.198	1.521	0.007	-0.008
			RF	93.736	93.873	-0.102	0.076	0.194	-0.192	-0.208	1.556	0.030	-0.028
			XGB	94.265	94.400	-0.102	0.076	0.194	-0.192	-0.203	1.542	0.021	-0.017
PSDAS	Train EB	Test EB	DT	74.066	81.456	0.000	0.000	0.000	0.000	-0.062	1.194	0.160	-0.113
			LR	78.192	84.348	0.000	0.000	0.000	0.000	-0.007	1.029	0.017	-0.012
			RF	77.084	83.770	0.000	0.000	0.000	0.000	-0.056	1.182	0.154	-0.090
			XGB	77.234	83.672	0.000	0.000	0.000	0.000	-0.013	1.052	0.036	-0.021

Continued on next page

Table 7.5 – Complete results for all datasets, training and testing variations, and models

Dataset	Train Variation	Test Variation	Model	F1-Score	Accuracy	CI	KL	KS	DD	DPPL	DI	DCA	DCR
		Test HB	DT	72.642	79.080	0.000	0.000	0.000	0.000	0.260	0.544	0.558	-0.499
			LR	78.139	83.265	0.000	0.000	0.000	0.000	0.343	0.447	0.388	-0.348
			RF	76.475	82.175	0.000	0.000	0.000	0.000	0.276	0.520	0.536	-0.469
			XGB	77.639	82.915	0.000	0.000	0.000	0.000	0.338	0.457	0.401	-0.357
		Test O	DT	79.814	80.741	0.000	0.000	0.000	0.000	-0.195	1.599	-0.034	0.063
			LR	82.361	83.047	0.000	0.000	0.000	0.000	-0.154	1.436	-0.145	0.145
			RF	82.338	83.092	0.000	0.000	0.000	0.000	-0.201	1.622	-0.020	0.068
			XGB	82.309	83.002	0.000	0.000	0.000	0.000	-0.159	1.461	-0.131	0.137
	Train HB	Test EB	DT	71.218	78.051	-0.336	0.597	0.510	0.480	0.258	0.525	-0.606	0.515
			LR	74.219	80.656	-0.336	0.597	0.510	0.480	0.248	0.526	-0.633	0.469
			RF	73.864	80.846	-0.336	0.597	0.510	0.480	0.243	0.519	-0.673	0.430
			XGB	72.359	79.657	-0.336	0.597	0.510	0.480	0.271	0.482	-0.761	0.490
		Test HB	DT	78.385	83.529	-0.336	0.597	0.510	0.480	0.575	0.251	-0.200	0.252
			LR	81.437	85.926	-0.336	0.597	0.510	0.480	0.590	0.236	-0.272	0.319

Continued on next page

Table 7.5 – Complete results for all datasets, training and testing variations, and models

Dataset	Train Variation	Test Variation	Model	F1-Score	Accuracy	CI	KL	KS	DD	DPPL	DI	DCA	DCR
			RF	81.787	86.320	-0.336	0.597	0.510	0.480	0.577	0.238	-0.273	0.191
			XGB	81.208	85.753	-0.336	0.597	0.510	0.480	0.597	0.232	-0.307	0.310
		Test O	DT	71.718	73.553	-0.336	0.597	0.510	0.480	0.159	0.710	-0.872	0.738
			LR	75.368	76.877	-0.336	0.597	0.510	0.480	0.125	0.762	-0.795	0.664
			RF	74.228	76.311	-0.336	0.597	0.510	0.480	0.125	0.753	-0.843	0.629
			XGB	73.098	75.248	-0.336	0.597	0.510	0.480	0.150	0.714	-0.903	0.686
	Train O	Test EB	DT	74.097	78.245	-0.297	0.104	0.227	-0.207	-0.110	1.267	0.175	-0.272
			LR	78.210	83.061	-0.297	0.104	0.227	-0.207	-0.119	1.335	0.240	-0.237
			RF	78.827	83.482	-0.297	0.104	0.227	-0.207	-0.093	1.247	0.179	-0.188
			XGB	78.519	83.383	-0.297	0.104	0.227	-0.207	-0.105	1.296	0.215	-0.206
		Test HB	DT	72.406	75.119	-0.297	0.104	0.227	-0.207	0.219	0.683	0.624	-0.546
			LR	75.714	79.519	-0.297	0.104	0.227	-0.207	0.236	0.629	0.617	-0.539
			RF	77.324	81.088	-0.297	0.104	0.227	-0.207	0.266	0.583	0.563	-0.488
			XGB	77.067	80.870	-0.297	0.104	0.227	-0.207	0.254	0.601	0.590	-0.503

Continued on next page

Table 7.5 – Complete results for all datasets, training and testing variations, and models

Dataset	Train Variation	Test Variation	Model	F1-Score	Accuracy	CI	KL	KS	DD	DPPL	DI	DCA	DCR
		Test O	DT	82.618	81.126	-0.297	0.104	0.227	-0.207	-0.239	1.565	-0.039	-0.144
			LR	85.117	84.562	-0.297	0.104	0.227	-0.207	-0.265	1.736	0.049	-0.120
			RF	85.434	84.923	-0.297	0.104	0.227	-0.207	-0.237	1.626	-0.012	-0.060
			XGB	85.491	85.036	-0.297	0.104	0.227	-0.207	-0.251	1.694	0.026	-0.083

Table 7.6 – Model-level fairness metrics (DPPL and DI) for all datasets, testing on the original distribution. DPPL colour bands: $|\text{DPPL}| < 0.1$ (fair), $0.1 \leq |\text{DPPL}| < 0.3$, $|\text{DPPL}| \geq 0.3$ (severe). DI colour bands: $0.8 \leq \text{DI} \leq 1.25$ (fair), otherwise (unfair). DCA and DCR are omitted due to numerical instability ($\pm\infty$) in several combinations; full values are in Appendix 7.5.

		Train Original		**Train EB**		**Train HB**	
Type	**Dataset**	*DPPL*	*DI*	*DPPL*	*DI*	*DPPL*	*DI*
A	Heart	+0.285	0.458	−0.041	1.240	+0.661	0.144
	Employee	+0.273	0.679	+0.106	0.862	+0.846	0.084
	COMPAS	−0.113	1.159	−0.016	1.020	+0.581	0.382
	Ind. Liver	+0.081	0.912	+0.004	1.000	+0.543	0.450
B	Adult	+0.177	0.309	+0.034	0.716	+0.503	0.058
	Bank Mktg	−0.019	1.396	−0.013	1.180	+0.128	0.067
C	PSDAS	−0.248	1.655	−0.177	1.530	+0.140	0.735
	ACTG175	−0.056	1.074	−0.039	1.050	+0.236	0.735
	Intersect.	−0.206	1.549	−0.164	1.437	−0.054	1.129
	Diabetes	+0.001	0.999	+0.008	0.992	+0.302	0.698
D	Glioma†	−0.190	∞	−0.212	∞	+0.170	0.997

7.5 Feature Importances

Table 7.7 – Summary of feature importance ranks and values for the protected attribute across datasets and training variations. Orig→HB shows the change in mean rank from original to highly biased training. HB rank 1 indicates in how many of the 4 models the protected attribute was ranked 1 under highly biased training. All the values are averaged from all the models generated (10 for each variation).

Dataset	Protected attribute	Rank Orig	Rank EB	Rank HB	Importance Orig	Importance EB	Importance HB	Orig→HB	HB rank 1
HEART	sex	6.900	10.900	1.000	0.384	0.078	1.212	-5.900	4/4
PSDAS	gender	21.000	14.400	1.300	0.084	0.166	0.869	-19.600	1/4
EMP	gender	3.300	6.700	1.000	0.283	0.073	1.516	-2.300	4/4
COMPAS	race	9.300	8.200	1.200	0.029	0.095	1.312	-8.100	3/4
ADULT	sex	13.200	31.000	2.200	0.221	0.110	1.072	-11.000	2/4
BANK	marital	23.600	33.600	1.100	0.065	0.038	1.070	-22.600	3/4
ILPD	sex	9.600	9.100	1.000	0.013	0.066	1.168	-8.600	4/4
ACTG175	homo	13.500	12.500	1.800	0.028	0.091	0.900	-11.800	1/4
IBD	sex	9.100	7.000	2.000	0.137	0.286	0.952	-7.100	1/4
DIAB	sex	9.500	12.400	1.000	0.069	0.023	1.241	-8.500	4/4
GLIOMA	race	9.800	9.300	2.400	0.129	0.066	0.686	-7.400	1/4

7.6 Resumo Expandido

Esta dissertação parte de uma questão que se tornou cada vez mais incontornável à medida que sistemas de aprendizado de máquina passaram a ocupar espaços decisórios mais sensíveis: se os dados utilizados para treinamento já carregam assimetrias relevantes entre grupos sociais, em que medida tais assimetrias tendem a se refletir, mais tarde, nas decisões produzidas pelos modelos. Embora a literatura disponha de múltiplas taxonomias para descrever fontes de viés e de um conjunto já consolidado de métricas para avaliar unfairness após o treinamento, ainda são relativamente escassos os trabalhos que investigam, de forma sistemática e comparativa, o quanto certas propriedades estatísticas observáveis diretamente nos dados podem funcionar como indícios confiáveis do comportamento futuro dos classificadores, antes do esforço computacional ser empregado no treinamento dos modelos. Esse trabalho visa expandir a literatura incluindo essa análise. Em vez de tomar o viés apenas como um problema a ser diagnosticado depois que o modelo já foi ajustado, a dissertação procura olhar para um momento anterior do pipeline e perguntar se parte desse problema já não está dada, de forma mensurável, na própria organização dos dados de entrada.

A relevância dessa discussão decorre, em larga medida, do deslocamento recente de uma inteligência artificial centrada em modelos para uma inteligência artificial centrada em dados. Se, durante muito tempo, o foco esteve concentrado na escolha da arquitetura, no refinamento de hiperparâmetros e na melhoria do desempenho preditivo, a consolidação de aplicações em áreas como saúde, justiça criminal, educação, finanças e recursos humanos tornou mais evidente que a qualidade de um sistema não pode ser julgada apenas pela sua capacidade de acertar mais. Em contextos dessa natureza, não basta que o modelo apresente bom desempenho médio; é necessário também compreender quem se beneficia desse desempenho, quem suporta seus erros e de que maneira a distribuição dos dados de treinamento condiciona ou aumenta a produção de desigualdades entre grupos da sociedade. Assumir que a questão da equidade emerge apenas no momento da predição significa, nesse sentido, restringir o problema a uma etapa tardia demais, quando parte importante da estrutura que sustenta o comportamento do modelo já foi definida.

É justamente por essa razão que a dissertação concentra sua atenção nas chamadas métricas de viés em nível de dados. Tais métricas, também referidas

em parte da literatura como métricas pré-treinamento, procuram mensurar disparidades entre grupos antes da indução do modelo, isto é, sem depender das saídas de um classificador. No trabalho, foram utilizadas quatro métricas dessa natureza: Class Imbalance, Kullback-Leibler Divergence, Kolmogorov-Smirnov e Demographic Disparity. Em paralelo, a análise pós-treinamento foi conduzida com métricas que capturam o viés no comportamento efetivo do modelo, como Difference in Positive Proportions in Predicted Labels, Disparate Impact e Difference in Conditional Outcome. A questão subjacente a essa escolha metodológica é: se os dados apresentam sinais de distorção entre grupos, até que ponto essas distorções antecipam a direção e a intensidade das disparidades produzidas pelo classificador? Mais do que propor uma substituição entre métricas pré e pós-treinamento, o trabalho procura estabelecer uma ponte entre esses dois momentos, sustentando que uma auditoria conduzida cedo o suficiente pode fornecer informação valiosa sobre o risco de disparidade no modelo gerado.

A investigação foi estruturada a partir de quatro questões de pesquisa que atravessam toda a dissertação. A primeira delas trata do efeito do viés em nível de dados sobre o desempenho preditivo, com especial interesse em avaliar se a equalização das distribuições entre grupos reduz as lacunas de desempenho sem comprometer, de forma significativa, o F1-score global dos modelos. A segunda volta-se à correlação entre viés a nível de dados e modelo, buscando verificar se todas elas possuem o mesmo poder explicativo ou se algumas se mostram mais adequadas como indicadores precoces, focando também na diferença entre métricas distribucionais e o desbalanceamento absoluto das classes. A terceira examina a robustez relativa de diferentes algoritmos supervisionados frente a dados enviesados. A quarta, por fim, analisa o quanto a presença de viés na base altera a importância atribuída aos atributos protegidos durante o processo de aprendizado. Embora distintas entre si, essas questões convergem para um mesmo problema de fundo: entender se o viés em nível de dados deve ser tratado como uma irregularidade estatística lateral ou como um elemento central na explicação do comportamento dos modelos.

Para responder a essas questões, foi conduzido um estudo empírico com 11 conjuntos de dados provenientes de domínios bastante distintos: Heart, Predict Students' Dropout and Academic Success, Employee, COMPAS, Adult, Bank Marketing, Indian Liver Patient Dataset, AIDS Clinical Trials Group Study 175, Intersectional Bias Dataset, Diabetes e Glioma. A escolha dessas bases não teve apenas

o propósito de ampliar a diversidade amostral do estudo, mas também de evitar que as conclusões dependessem excessivamente de um único benchmark ou de um único domínio de aplicação. Cada conjunto apresenta suas próprias especificidades quanto ao atributo protegido, ao desfecho de interesse, ao volume de dados e ao tipo de assimetria inicialmente presente. Em alguns casos, o atributo protegido é sexo ou gênero; em outros, raça, estado civil ou relacionado à orientação sexual. Em todos eles, entretanto, as tarefas foram tratadas como problemas de classificação binária, o que permitiu aplicar o mesmo conjunto de métricas e organizar os experimentos sob um protocolo metodológico comum.

Durante a análise, identificou-se também que os conjuntos de dados analisados, durante os experimentos, se encaixavam em quatro classes, chamadas de A, B, C, e D. A primeira classe foca em dois datasets com amplificação mais extrema do viés na variação altamente desbalanceada. A classe B é caracterizada por um claro contraste entre treinar com dados igualmente balanceados e desbalanceados, corroborando a ideia de como a distribuição dos dados de treino afeta diretamente os resultados, e não apenas a escolha do algoritmo. O grupo C identifica os conjuntos de dados em que as variações nos conjuntos de treino não são tão sensíveis às variações do treinamento, e, por fim, o grupo D inclui apenas um dataset com comportamento completamente diferente dos demais.

A preparação dos dados seguiu uma lógica de padronização necessária para tornar os experimentos comparáveis. Os atributos protegidos e as variáveis-alvo foram mapeados para valores binários, estabelecendo-se 1 para grupos privilegiados e desfechos favoráveis, e 0 para grupos não privilegiados e desfechos desfavoráveis. Variáveis categóricas com mais de duas classes foram transformadas por one-hot encoding; variáveis temporais, quando presentes, foram convertidas para representações numéricas; e os atributos contínuos foram padronizados por normalização z-score.

O ponto mais característico da metodologia está na criação deliberada de diferentes versões de cada conjunto de dados. Para além da base original, foram geradas duas variações: uma versão igualmente balanceada, na qual a distribuição dos grupos privilegiado e não privilegiado foi equalizada dentro de cada classe do desfecho, e uma versão altamente desbalanceada, produzida pela remoção de 80 por cento das instâncias correspondentes ao grupo não privilegiado com desfecho favorável ou ao grupo privilegiado com desfecho desfavorável. Esse procedimento

não foi adotado com a pretensão de reproduzir a complexidade total das formas pelas quais o viés emerge em cenários reais, mas como estratégia de controle experimental. Ao manipular de maneira sistemática a estrutura distributiva dos dados, tornou-se possível observar com maior nitidez como alterações nesse nível se traduzem em variações posteriores de desempenho e fairness. A partir dessas três versões foram definidos cenários em que os modelos eram treinados e testados em todas as combinações possíveis, totalizando nove configurações por conjunto de dados.

Os algoritmos avaliados foram Regressão Logística, Árvore de Decisão, Floresta Aleatória e XGBoost. A escolha desses modelos buscou contemplar tanto abordagens lineares quanto não lineares, bem como arquiteturas simples e ensembles mais sofisticados. Para os algoritmos cuja parametrização exigia ajuste, foi utilizada busca em grade com GridSearchCV, adotando-se o F1-score como critério de otimização. A etapa principal do experimento foi conduzida com validação cruzada em 10 folds, calculando-se, em cada repetição, tanto as métricas associadas ao viés em nível de dados quanto as métricas de desempenho e fairness em nível de modelo. Esse desenho metodológico permitiu que a análise não se restringisse ao comportamento agregado de um único treinamento, mas considerasse a estabilidade dos resultados ao longo de diferentes partições dos dados, o que confere maior robustez às interpretações formuladas ao longo do trabalho.

Os resultados obtidos mostram, de forma consistente, que o viés presente nos dados de treinamento não permanece como uma característica inerte da base, mas se projeta de maneira concreta sobre o comportamento dos classificadores. Em termos gerais, o treinamento sobre a variação altamente desbalanceada esteve associado ao aumento dos gaps de desempenho entre grupos e ao agravamento das métricas pós-treinamento de fairness. Isso foi observado em 10 dos 11 conjuntos analisados, o que sugere que o fenômeno não está restrito a uma base ou domínio específico. O caso do conjunto Heart é especialmente elucidativo: quando os modelos foram treinados com a versão igualmente balanceada e avaliados no teste original, os gaps de F1-score entre grupos ficaram muito próximos da paridade. Já sob treinamento desbalanceado, esses valores cresceram de maneira acentuada, alcançando mais de 33 pontos no teste original e mais de 78 pontos quando o teste também se encontrava altamente enviesado. Esse contraste evidencia, de maneira difícil de ignorar, que as disparidades distributivas introduzidas na fase de treinamento têm consequências diretas sobre a forma como o modelo distribui qualidade preditiva entre os grupos

analisados.

O conjunto Employee apresenta comportamento semelhante e reforça essa leitura. Na condição de treinamento balanceado, os gaps observados foram relativamente contidos, ao passo que, na condição de treinamento altamente enviesado, alcançaram valores superiores a 77 pontos. Em outros conjuntos, como Predict Students' Dropout and Academic Success, ACTG175, Diabetes e Indian Liver Patient, o padrão geral se repete, ainda que com intensidades distintas. O que varia de uma base para outra é menos a direção do efeito do que sua magnitude, modulada por fatores como tamanho do conjunto, natureza do desfecho e estrutura interna das variáveis. Isso permite afirmar, com algum grau de segurança, que a presença de viés em nível de dados exerce influência sistemática sobre os resultados, ainda que essa influência não se manifeste de maneira idêntica em todos os cenários.

Ao mesmo tempo, os experimentos mostram que a equalização das distribuições frequentemente reduz disparidades sem impor perdas expressivas de desempenho global. Essa constatação aparece com clareza em diferentes bases, nas quais o treinamento com a versão balanceada preserva F1-scores médios competitivos enquanto diminui a distância entre grupos. Em outras palavras, a melhoria em fairness não se dá, nesses casos, à custa de uma degradação generalizada da capacidade preditiva do modelo. Esse é um ponto importante, porque parte do debate aplicado em machine learning costuma opor desempenho e equidade como grandezas inevitavelmente tensionadas em toda situação. O que os resultados desta dissertação sugerem é algo mais nuançado: em muitos contextos, reduzir a assimetria distributiva na base pode melhorar substancialmente a distribuição dos erros sem comprometer de forma relevante o desempenho agregado.

Ainda assim, as métricas de viés nos aos não antecipam a performance para todos os casos observados: o conjunto Glioma aparece no cenário contrário. Nesse caso, a versão balanceada implicou uma redução importante no tamanho do conjunto de treinamento, e a consequência foi a deterioração do desempenho e o aumento dos gaps entre grupos. A relevância desse resultado está em implicar que o comportamento do modelo incide sobre estruturas concretas de dados, e seus efeitos dependem da quantidade de instâncias preservadas, da relação entre grupos e desfechos e da complexidade do problema preditivo em questão. Em bases pequenas ou particularmente sensíveis à perda de amostras, o custo estatístico do balanceamento pode superar seus benefícios distributivos.

Entre as contribuições esse trabalho, destaca-se a comparação entre as métricas de viés em nível de dados e sua capacidade de antecipar as métricas a nível e modelo. Os resultados indicam que nem todas as métricas pré-treinamento são igualmente informativas: de modo bastante consistente, métricas distributivas como Kullback-Leibler Divergence e Kolmogorov-Smirnov mostraram-se mais adequadas do que Class Imbalance para sinalizar risco de disparidade do modelo. O conjunto Heart torna isso particularmente evidente. Na comparação entre a base original e a versão Highly Biased, o valor de Class Imbalance diminui, o que, em uma leitura superficial, poderia ser tomado como indício de melhora. No entanto, Kullback-Leibler Divergence cresce de 0,208 para 2,238, Kolmogorov-Smirnov sobe de 0,303 para 0,806 e Demographic Disparity aumenta de 0,265 para 0,754. Posteriormente, os modelos treinados sobre essa variação passam a apresentar valores muito baixos de Disparate Impact e forte desvantagem para o grupo não privilegiado. Nesse cenário, a métrica de Class Imbalance falha em captar a deterioração da estrutura dos dados, enquanto as métricas sensíveis à distribuição conseguem indicar, de forma muito mais precisa, o agravamento do risco.

O caso do Indian Liver Patient Dataset ajuda a completar esse argumento, agora pelo contraste. Nessa base, o valor de Class Imbalance na versão original é relativamente alto, o que poderia sugerir, por si só, um problema importante de representatividade. Entretanto, Kullback-Leibler, Kolmogorov-Smirnov e Demographic Disparity permanecem baixos, e os modelos resultantes apresentam comportamento comparativamente mais equilibrado. O que essa oposição entre Heart e Indian Liver Patient mostra é que contar quantas instâncias há em cada grupo, embora possa ser útil em algumas situações, não basta para caracterizar adequadamente o potencial de viés de uma base. O ponto decisivo está em compreender como o atributo protegido se articula com a distribuição do alvo, isto é, como se organizam, dentro da base, as combinações entre pertencimento ao grupo e desfecho favorável ou desfavorável. Métricas capazes de incorporar essa dimensão relacional acabam oferecendo um retrato mais fiel do risco efetivamente carregado pelos dados.

Essa constatação tem consequências importantes para a prática de auditoria. Se o objetivo é detectar precocemente bases com maior propensão a induzir desigualdades, parece pouco prudente depender apenas de métricas representacionais simples. A dissertação sugere que métricas baseadas em divergência ou distância entre distribuições fornecem sinais mais úteis, justamente porque capturam mudanças

mais sutis e estruturalmente relevantes na relação entre grupos e resultados. Isso não significa que exista um valor de corte universal ou uma regra rígida aplicável a qualquer domínio, mas reforça a ideia de que auditorias pré-treinamento podem ser mais informativas quando se apoiam em um conjunto de métricas complementares, e não em um único indicador tomado isoladamente.

No que se refere aos algoritmos, os resultados não sustentam a hipótese de que alguma das famílias analisadas seja intrinsecamente robusta ao viés presente nos dados. Regressão Logística, Árvore de Decisão, Floresta Aleatória e XGBoost responderam de maneira diferente a depender da base e da configuração experimental, mas nenhum deles se mostrou sistematicamente imune às distorções introduzidas nas distribuições de treinamento. Em alguns cenários, os ensembles apresentaram melhor desempenho agregado, mas ainda assim reproduziram disparidades relevantes. Em outros, modelos mais simples se mostraram tão suscetíveis quanto os mais complexos. O que emerge dessa comparação é menos uma hierarquia entre algoritmos e mais a percepção de que a estrutura dos dados pesa de forma decisiva. Não se trata, portanto, de um problema que possa ser resolvido apenas pela substituição do classificador. Quando a base carrega assimetrias suficientemente fortes, essas assimetrias tendem a ser absorvidas, em maior ou menor medida, por qualquer um dos modelos testados. Outro aspecto importante diz respeito à forma como o viés em nível de dados altera a importância atribuída aos atributos protegidos. Em diversos conjuntos, o treinamento sobre versões altamente enviesadas fez com que variáveis sensíveis passassem a ocupar posições centrais no processo decisório do modelo. No conjunto Heart, por exemplo, o atributo sex se tornou uma das variáveis mais importantes sob treinamento Highly Biased; padrões semelhantes foram observados em Employee, COMPAS, Adult e Bank Marketing. Já nas versões balanceadas, essa dependência tende a diminuir de maneira expressiva. Isso sugere que o viés não afeta apenas os resultados finais observados nas métricas de fairness, mas a própria maneira como o modelo organiza internamente seus critérios de decisão. Em termos interpretativos, trata-se de um achado relevante, porque mostra que a mitigação no nível dos dados pode alterar não só a distribuição das saídas, mas também a centralidade de características sensíveis na estrutura aprendida pelo classificador.

É importante notar, entretanto, que a presença de unfairness não depende exclusivamente de o atributo protegido aparecer como a variável mais importante do modelo. Em alguns casos, disparidades relevantes foram observadas mesmo quando

esse atributo não ocupava a primeira posição no ranking de importância. Isso indica que o problema pode ser mediado por proxies e correlações indiretas, por meio das quais outras variáveis passam a carregar parte da informação associada ao atributo sensível. Essa observação reforça a necessidade de uma análise combinada: não basta remover o atributo protegido e tampouco é suficiente olhar apenas para as importâncias de variáveis.

No conjunto, os resultados sustentam a ideia de que auditorias em nível de dados podem desempenhar um papel mais relevante do que geralmente se lhes atribui. Elas não substituem a avaliação pós-treinamento, nem esgotam a complexidade do problema da justiça algorítmica, mas oferecem uma camada inicial de diagnóstico que se mostrou empiricamente informativa ao longo desta dissertação. Ao estabelecer relações consistentes entre propriedades estatísticas observáveis antes do treinamento e disparidades verificadas depois dele, o trabalho contribui para aproximar dois momentos que muitas vezes aparecem dissociados na prática: a curadoria dos dados e a avaliação de fairness do modelo. Essa aproximação desloca parte da atenção para uma etapa em que intervenções ainda são relativamente menos custosas e podem evitar que problemas mais sérios se consolidem nas fases seguintes do pipeline.

Naturalmente, essa contribuição precisa ser lida à luz de algumas limitações. Os experimentos foram realizados com tarefas de classificação binária em dados tabulares estruturados, o que restringe a generalização imediata dos resultados para contextos como regressão, classificação multiclasse ou modalidades não tabulares, como texto, imagem e áudio. Além disso, a manipulação controlada das distribuições, embora metodologicamente útil, não pretende reproduzir toda a complexidade sociotécnica envolvida na produção de viés em sistemas reais, onde fatores históricos, institucionais e operacionais interagem de forma menos transparente. Também se verificou que estratégias de balanceamento podem se tornar problemáticas em bases pequenas, quando a redução do conjunto compromete a estabilidade do aprendizado. Essas limitações, contudo, não anulam o alcance do trabalho; antes, ajudam a situá-lo com maior precisão e a indicar direções possíveis para continuidade da pesquisa.

Entre essas direções, parecem especialmente promissoras a ampliação do estudo para outros conjuntos de dados, a incorporação de cenários interseccionais mais complexos, a experimentação com outras métricas pré-treinamento e a integração dessas análises em pipelines automatizados de governança e desenvolvimento

de modelos. Se parte do risco de unfairness já pode ser percebida na distribuição dos dados, então faz sentido que mecanismos de auditoria sejam incorporados desde as etapas iniciais de preparação e seleção das bases, e não apenas ao final do processo. Em última instância, é essa a implicação mais ampla: modelos mais justos dependem, em grande medida, de dados mais bem compreendidos, mais cuidadosamente auditados e menos assimetricamente organizados.

REFERENCES

BARROS, B. et al. The impact of model selection metrics during hyperparameter tuning on algorithmic fairness: An empirical study. In: *Anais do LIII Seminário Integrado de Software e Hardware*. Porto Alegre, RS, Brasil: SBC, 2026. p. 710–721. ISSN 2595-6205. Available from Internet: <https://sol.sbc.org.br/index.php/semish/article/view/43553>.

BARROS, B. M. de et al. Construindo modelos justos: Fundamentos, estratégias e desafios para uma ia Ética e equitativa na saúde. In: RECAMONDE-MENDOZA, M.; GARCÉS, L. (Ed.). *Minicursos do XXV Simpósio Brasileiro de Computação Aplicada à Saúde*. Porto Alegre, RS, Brasil: Sociedade Brasileira de Computação, 2025. chp. 4, p. 134–183. Available from Internet: <https://doi.org/10.5753/sbc.16731.5.4>.

BECKER, B.; KOHAVI, R. *Adult*. 1996. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5XW20.

BELLAMY, R. K. E. et al. Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. 10 2018. Available from Internet: <http://arxiv.org/abs/1810.01943>.

BELLAMY, R. K. E. et al. *AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias*. 2018. Available from Internet: <https://arxiv.org/abs/1810.01943>.

BREIMAN, L. Random forests. *Machine Learning*, v. 45, p. 5–32, 10 2001. ISSN 0885-6125. Available from Internet: <https://link.springer.com/10.1023/A:1010933404324>.

BREIMAN, L. et al. *Classification And Regression Trees*. Routledge, 2017. ISBN 9781315139470. Available from Internet: <http://dx.doi.org/10.1201/9781315139470>.

BURKOV, A. *The hundred-page machine learning book*. [S.l.]: Andriy Burkov, 2019. 141 p. ISBN 9781999579517.

CAMPBELL, E. A.; BOSE, S.; MASINO, A. J. Conceptualizing bias in ehr data: A case study in performance disparities by demographic subgroups for a pediatric obesity incidence classifier. *PLOS Digital Health*, Public Library of Science, v. 3, 10 2024. ISSN 27673170.

CHEN, T.; GUESTRIN, C. Xgboost: A scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: Association for Computing Machinery, 2016. (KDD '16), p. 785–794. ISBN 9781450342322. Available from Internet: <https://doi.org/10.1145/2939672.2939785>.

CHOULDECHOVA, A. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, v. 5, p. 153–163, 6 2017. ISSN 2167-6461. Available from Internet: <http://www.liebertpub.com/doi/10.1089/big.2016.0047>.

CLEMENTE, F. et al. ydata-profiling: Accelerating data-centric ai with high-quality data. *Neurocomputing*, v. 554, p. 126585, 2023. ISSN 0925-2312. Available from Internet: <https://www.sciencedirect.com/science/article/pii/S0925231223007087>.

CRONJé, H. T. et al. Assessing racial bias in type 2 diabetes risk prediction algorithms. *PLOS Global Public Health*, Public Library of Science, v. 3, 5 2023. ISSN 27673375.

DANESHJOU, R. et al. Disparities in dermatology ai performance on a diverse, curated clinical image set. *Science Advances*, v. 8, 8 2022. ISSN 2375-2548. Available from Internet: <https://www.science.org/doi/10.1126/sciadv.abq6147>.

DENG, J. et al. Imagenet: A large-scale hierarchical image database. In: IEEE. *2009 IEEE conference on computer vision and pattern recognition*. [S.l.], 2009. p. 248–255.

DROZDA, K. et al. Poor warfarin dose prediction with pharmacogenetic algorthms that exclude genotypes important for african americans. *Pharmacogenetics and Genomics*, Lippincott Williams and Wilkins, v. 25, p. 73–81, 2 2015. ISSN 17446880.

ELMETWALLY, T. *Employee Dataset*. 2021. Kaggle Dataset. Accessed: 2026-05-16. Available from Internet: <https://www.kaggle.com/datasets/tawfikelmetwally/employee-dataset>.

FABUYI, J. A. Leveraging synthetic data as a tool to combat bias in artificial intelligence (ai) model training. *Journal of Engineering Research and Reports*, Sciencedomain International, v. 26, p. 24–46, 11 2024.

FACELI, K. et al. *Inteligência artificial: uma abordagem de aprendizado de máquina*. [S.l.]: LTC, 2025.

FISCHER-ABAIGAR, U. et al. Bridging the gap: Towards an expanded toolkit for ai-driven decision-making in the public sector. *Government Information Quarterly*, v. 41, p. 101976, 12 2024. ISSN 0740624X. Available from Internet: <https://linkinghub.elsevier.com/retrieve/pii/S0740624X24000686>.

GICHOYA, J. W. et al. Ai recognition of patient race in medical imaging: a modelling study. *The Lancet Digital Health*, Elsevier Ltd, v. 4, p. e406–e414, 6 2022. ISSN 25897500.

GONG, A.; CAMPBELL, J.; Great Expectations. *Great Expectations*. Available from Internet: <https://github.com/great-expectations/great_expectations>.

HARDT, M. et al. Amazon sagemaker clarify: Machine learning bias detection and explainability in the cloud. 9 2021. Available from Internet: <http://arxiv.org/abs/2109.03285 http://dx.doi.org/10.1145/3447548.3467177>.

JAMES, G. et al. *An Introduction to Statistical Learning*. Springer US, 2021. ISSN 2197-4136. ISBN 978-1-0716-1417-4. Available from Internet: <https://link.springer.com/10.1007/978-1-0716-1418-1>.

JANOSI ANDRAS, S. W. P. M.; DETRANO, R. *Heart Disease*. 1989. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C52P4X.

JESUS, S. et al. *Turning the Tables: Biased, Imbalanced, Dynamic Tabular Datasets for ML Evaluation*. [S.l.], 2022. Available from Internet: <https://github.com/feedzai/bank-account-fraud>.

JúNIOR, R. L. I. et al. Justiça nas previsões de modelos de aprendizado de máquina: um estudo de caso com dados de reincidência criminal. In: *Anais do XIX Encontro Nacional de Inteligência Artificial e Computacional (ENIAC 2022)*. Sociedade Brasileira de Computação - SBC, 2022. p. 636–647. Available from Internet: <https://sol.sbc.org.br/index.php/eniac/article/view/22819>.

Kaggle. *AIDS Clinical Trials Group Study 175*. 1996. Kaggle Dataset. Accessed: 2026-05-16. Available from Internet: <https://www.kaggle.com/datasets/tanshihjen/aids-clinical-trials>.

LADHAK, F. et al. When do pre-training biases propagate to downstream tasks? a case study in text summarization. In: *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 2023. p. 3206–3219. Available from Internet: <https://aclanthology.org/2023.eacl-main.234>.

LARRAZABAL, A. J. et al. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proceedings of the National Academy of Sciences of the United States of America*, National Academy of Sciences, v. 117, p. 12592–12594, 6 2020. ISSN 10916490.

LARSON, J. et al. *How We Analyzed the COMPAS Recidivism Algorithm*. 2016. ProPublica. Accessed: 2026-05-16. Available from Internet: <https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm>.

LEAL, G. et al. Can data complexity measures detect pre-training bias in machine learning? a case-study with health data. In: *Anais do XXV Simpósio Brasileiro de Computação Aplicada à Saúde*. Porto Alegre, RS, Brasil: SBC, 2025. p. 1011–1016. ISSN 2763-8952. Available from Internet: <https://sol.sbc.org.br/index.php/sbcas/article/view/35568>.

LIU, M. et al. A scoping review and evidence gap analysis of clinical ai fairness. *npj Digital Medicine*, Nature Research, v. 8, 12 2025. ISSN 23986352.

LUCCIONI, A. S. et al. Stable bias: Analyzing societal representations in diffusion models. 11 2023. Available from Internet: <http://arxiv.org/abs/2303.11408>.

MANDHALA, V. N. et al. Detecting and mitigating bias in data using machine learning with pre-training metrics. *Ingénierie des systèmes d information*, International Information and Engineering Technology Association, v. 27, p. 119–125, 2 2022. ISSN 16331311. Available from Internet: <https://www.iieta.org/journals/isi/paper/10.18280/isi.270114>.

MASLEJ, M. *Intersectional-Bias-Assessment-(Testing-Data)*. [S.l.]: OpenML, 2022. Https://openml.org.

MEHRABI, N. et al. A survey on bias and fairness in machine learning. 8 2019. Available from Internet: <http://arxiv.org/abs/1908.09635>.

MORO, P. R. S. *Bank Marketing*. UCI Machine Learning Repository, 2014. Available from Internet: <https://archive.ics.uci.edu/dataset/222>.

MORO S., R. P.; CORTEZ, P. *Bank Marketing*. 2014. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5K306.

NOSEWORTHY, P. A. et al. Assessing and mitigating bias in medical artificial intelligence. *Circulation: Arrhythmia and Electrophysiology*, Lippincott Williams and Wilkins, v. 13, 3 2020. ISSN 1941-3149. Available from Internet: <https://www.ahajournals.org/doi/10.1161/CIRCEP.119.007988>.

OBERMEYER, Z. et al. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, American Association for the Advancement of Science, v. 366, p. 447–453, 10 2019. ISSN 10959203.

OLIVEIRA, T. A. et al. Estudo experimental sobre justiça algorítmica aplicada em modelos de análise de crédito. In: *Anais Estendidos do XXXIX Simpósio Brasileiro de Banco de Dados (SBBD Estendido 2024)*. Sociedade Brasileira de Computação - SBC, 2024. p. 29–36. Available from Internet: <https://sol.sbc.org.br/index.php/sbbd$_e$*stendido/article/view*/30767>.

OPENJA, M.; LABERGE, G.; KHOMH, F. Detection and evaluation of bias-inducing features in machine learning. *Empirical Software Engineering*, Springer, v. 29, 2 2024. ISSN 15737616.

PARK, Y. et al. Comparison of methods to reduce bias from clinical prediction models of postpartum depression. *JAMA Network Open*, American Medical Association, v. 4, p. e213909, 4 2021. ISSN 2574-3805. Available from Internet: <https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2778568>.

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

RAMANA, B.; VENKATESWARLU, N. *ILPD (Indian Liver Patient Dataset)*. 2022. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5D02C.

REALINHO VALENTIM, V. M. M. M. J.; BAPTISTA, L. *Predict Students' Dropout and Academic Success*. 2021. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5MC89.

RODRIGUES, D.; RECAMONDE-MENDOZA, M. Bias propagation in health ai: Measuring pre-training bias and its effect on machine learning model outcomes. In: *Anais do XXV Simpósio Brasileiro de Computação Aplicada à Saúde*. Porto Alegre, RS, Brasil: SBC, 2025. p. 341–352. ISSN 2763-8952. Available from Internet: <https://sol.sbc.org.br/index.php/sbcas/article/view/35508>.

RUBACK, L.; CARVALHO, D.; AVILA, S. Mitigando vieses no aprendizado de máquina. *iSys - Brazilian Journal of Information Systems*, Sociedade Brasileira de Computacao - SB, v. 15, 12 2022.

SESHADRI, R. *GitHub - AutoViML/AutoViz: Automatically Visualize any dataset, any size with a single line of code*. 2020. Available from Internet: <https://github.com/AutoViML/AutoViz">.

SETH, P.; PAI, A. K. Does the fairness of your pre-training hold up? examining the influence of pre-training techniques on skin tone bias in skin lesion classification. In: *2024 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*. IEEE, 2024. p. 580–587. ISBN 979-8-3503-7028-7. Available from Internet: <https://ieeexplore.ieee.org/document/10495612/>.

SJODING, M. W. et al. Racial bias in pulse oximetry measurement. *New England Journal of Medicine*, Massachusetts Medical Society, v. 383, p. 2477–2478, 12 2020. ISSN 0028-4793.

STOYANOVICH, J.; HOWE, B.; JAGADISH, H. V. Responsible data management. *Proceedings of the VLDB Endowment*, VLDB Endowment, v. 13, p. 3474–3488, 8 2020. ISSN 2150-8097. Available from Internet: <https://dl.acm.org/doi/10.14778/3415478.3415570>.

SURESH, H.; GUTTAG, J. V. A framework for understanding sources of harm throughout the machine learning life cycle. 1 2019. Available from Internet: <http://arxiv.org/abs/1901.10002 http://dx.doi.org/10.1145/3465416.3483305>.

Sweetviz. Available from Internet: <https://github.com/fbdesignpro/sweetviz>.

TASCI, E.; CAMPHAUSEN, K. *Glioma Grading Clinical and Mutation Features*. UCI Machine Learning Repository, 2022. Accessed: 2026-05-16. Available from Internet: <https://archive.ics.uci.edu/dataset/759>.

WEERTS, H. et al. *Fairlearn: Assessing and Improving Fairness of AI Systems*. 2023. 1–8 p. Available from Internet: <http://jmlr.org/papers/v24/23-0389.html>.