EVALUATION OF MACHINE-LEARNING CLASSIFIERS UNDER PROGRESSIVE UNFAIRNESS

Rodrigo Pagliusi

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Engenharia de Sistemas e Computação, COPPE, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Mestre em Engenharia de Sistemas e Computação.

Orientador: Geraldo Zimbrão da Silva

Rio de Janeiro
Setembro de 2025

EVALUATION OF MACHINE-LEARNING CLASSIFIERS UNDER PROGRESSIVE UNFAIRNESS

Rodrigo Pagliusi

DISSERTAÇÃO SUBMETIDA AO CORPO DOCENTE DO INSTITUTO ALBERTO LUIZ COIMBRA DE PÓS-GRADUAÇÃO E PESQUISA DE ENGENHARIA DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM CIÊNCIAS EM ENGENHARIA DE SISTEMAS E COMPUTAÇÃO.

Orientador: Geraldo Zimbrão da Silva

Aprovada por: Prof. Geraldo Zimbrão da Silva
Prof. Geraldo Bonorino Xexéo
Prof. Daniel Sadoc Menasche

RIO DE JANEIRO, RJ – BRASIL
SETEMBRO DE 2025

Pagliusi, Rodrigo

Evaluation of Machine-Learning Classifiers under Progressive Unfairness/Rodrigo Pagliusi. – Rio de Janeiro: UFRJ/COPPE, 2025.

XIV, 98 p.: il.; 29, 7cm.

Orientador: Geraldo Zimbrão da Silva

Dissertação (mestrado) – UFRJ/COPPE/Programa de Engenharia de Sistemas e Computação, 2025.

Referências Bibliográficas: p. 64 – 76.

1. Fairness. 2. Machine Learning. 3. Classification. I. Zimbrão da Silva, Geraldo. II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia de Sistemas e Computação. III. Título.

A dignidade intrínseca de cada homem e de cada mulher deve ser o critério fundamental na avaliação das tecnologias emergentes.

Agradecimentos

Agradeço a Deus por criar e sustentar tudo.

Agradeço a meus pais, Paulo e Márcia, por todo o apoio, toda paciência e todo o incentivo.

Agradeço à minha família, meu irmão Daniel, meus tios e primos todos, pessoas com quem sempre posso contar na alegria e na tristeza.

Agradeço à minha noiva, Michelle, por sempre me acolher e me fazer o homem mais feliz do mundo. Também agradeço à família dela, por terem me recebido tão generosamente.

Agradeço aos meus amigos do Colégio de São Bento por sempre me incentivarem, entenderem minhas dores e me proporcionarem bons momentos de lazer.

Agradeço ao pessoal da minha banda, Kaitak, por serem meu refúgio, e terem me dado muita alegria e boas amizades.

Agradeço ao pessoal dos jogos, que estiveram sempre presentes nos bons momentos e com quem comparitlho boas memórias.

Agradeço a meus amigos da faculdade, em especial ao Will, com quem sempre tive boas trocas de ideias.

Agradeço à Alta Geotecnia, que me proporcionou excelentes oportunidades profissionais, sempre valorizando o lado humano.

Agradeço ao Luís Guilherme, por me auxiliar muito profissionalmente e me ter dado boas oportunidades e sempre bom humor.

Agradeço à minha psicóloga Kátia, por sempre me ajudar a melhorar como pessoa.

Agradeço a meu orientador, Geraldo Zimbrão, por ter me proporcionado a oportunidade e sempre me ajudar e compreender quando precisei. Sou grato aos professores Filipe Braida e Leandro Alvim por me acompanharem em todo o processo, apontando os erros e melhores caminhos. Sou grato ao Ygor Canalli, que sempre me auxiliou e incentivou bastante, me dando excelentes conselhos e dicas.

Sou grato ao PESC e toda a sua equipe, em especial ao Guty, por me auxiliarem em todo o percurso.

Resumo da Dissertação apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Mestre em Ciências (M.Sc.)

AVALIAÇÃO DE CLASSIFICADORES DE APRENDIZADO DE MÁQUINA SOB INJUSTIÇA PROGRESSIVA

Rodrigo Pagliusi

Setembro/2025

Orientador: Geraldo Zimbrão da Silva

Programa: Engenharia de Sistemas e Computação

Modelos de aprendizado de máquina extraem padrões de grandes volumes de dados. Quando estes refletem desigualdades históricas ou sociais, tendem a reproduzi-las em suas previsões. Esse risco é especialmente relevante em domínios sensíveis como justiça criminal, saúde, emprego e finanças, nos quais decisões algorítmicas podem impactar diretamente a vida das pessoas.

Embora existam diversas técnicas para mitigar injustiças, o grau adequado de intervenção ainda é pouco explorado, sobretudo porque envolve equilibrar justiça e desempenho preditivo. Esta dissertação investiga como algoritmos tradicionais se comportam sob dados enviesados sem mecanismos de mitigação, por meio de uma análise sistemática de seu desempenho em condições progressivamente injustas.

Para este fim, foi proposta a metodologia Systematic Label Flipping for Fairness Stress Testing, que insere viés controlado nos dados de treinamento. Essa abordagem permite avaliar a robustez de classificadores e observar, de forma gradual, como métricas de desempenho e justiça evoluem à medida que o viés aumenta.

Foram analisados os modelos Árvore de Decisão, Floresta Aleatória, Regressão Logística e Rede Neural. Em geral, os resultados foram semelhantes, com exceção da Regressão Logística, que no dataset COMPAS sofreu maior degradação de desempenho e aumento de injustiça. As Árvores de Decisão mostraram-se ligeiramente mais estáveis, mas as diferenças entre algoritmos foram discretas.

As contribuições desta dissertação são duas: a proposição de uma metodologia reprodutível de stress testing de justiça e a apresentação de evidências empíricas sobre a robustez de modelos tradicionais frente a cenários enviesados.

Resumo da Dissertação apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Mestre em Ciências (M.Sc.)

AVALIAÇÃO DE CLASSIFICADORES DE APRENDIZADO DE MÁQUINA SOB INJUSTIÇA PROGRESSIVA

Rodrigo Pagliusi

Setembro/2025

Orientador: Geraldo Zimbrão da Silva

Programa: Engenharia de Sistemas e Computação

Machine learning models extract patterns from large volumes of data. When such data reflect historical or social inequalities, algorithms tend to reproduce them in their predictions. This risk is particularly relevant in sensitive domains such as criminal justice, healthcare, employment, and finance, where algorithmic decisions can directly affect people's lives.

Although several techniques exist to mitigate unfairness, the appropriate degree of intervention remains underexplored, especially because it requires balancing fairness and predictive performance. This dissertation investigates how traditional algorithms behave when exposed to biased data without mitigation mechanisms, through a systematic analysis of their performance under progressively unfair conditions.

To this end, the Systematic Label Flipping for Fairness Stress Testing methodology was proposed, which introduces controlled bias into the training data. This approach makes it possible to assess the robustness of classifiers and to gradually observe how performance and fairness metrics evolve as data bias increases.

The models analyzed were Decision Tree, Random Forest, Logistic Regression, and Neural Network. Overall, results were similar, with the main exception being Logistic Regression, which on the COMPAS dataset suffered a greater drop in performance accompanied by increased unfairness. Decision Trees proved slightly more stable, but overall the differences across algorithms were modest.

The contributions of this dissertation are twofold: the proposal of a reproducible methodology for fairness stress testing and the presentation of empirical evidence on the robustness of traditional models when subjected to biased scenarios.

Contents

List of Figures

List of Tables

List of Symbols

A	sensitive attribute, p. 12, 25, 26, 32
D	Dataset, p. 32, 33, 36
D^*	Modified Dataset, p. 33, 34, 36
G_n	privileged-negative group of the dataset, p. 36
G_p	protected-positive group of the dataset, p. 36
X	all attributes of the individual, p. 12, 24, 32, 33, 36
X_{-A}	all attributes of the individual, except the sensitive attribute, p. 12
Y	true class, p. 12, 13, 24–26
$\hat{E}$	incorrect prediction, p. 12
$\hat{Y}$	predicted class, p. 12
π_i	Probability of instance belonging to the positive class, p. 36
ρ	Pollution Rate, p. 35, 36, 40, 42, 43, 50, 51, 53, 57, 58
$\tilde{E}$	noise, p. 12, 24
$\tilde{Y}$	observed class, p. 12, 13, 24–26, 32, 33, 36
h_e	estimator model, p. 32, 34–36
i	instances of the dataset, p. 36
m_n	Proportions of the privileged-negative instances that are going to be flipped, p. 36
m_p	Proportions of the protected-positive instances that are going to be flipped, p. 36

List of Abbreviations

AI	Artificial Intelligence, p. 6
AUC	Area Under the Curve, p. 30
Acc.	Accuracy, p. 3
CM	Confusion Matrix, p. 13
COMPAS	Correctional Offender Management Profiling for Alternative Sanctions, p. 7
Count. Fair.	Counterfactual Fairness, p. 16
Eq. Odds	Equalized Odds, p. 3
Eq. Opp.	Equal Opportunity, p. 3
F1	F1 Score, p. 3
FNR	False Negative Rate, p. 42
FN	False Negative, p. 13
FPR	False Positive Rate, p. 42
FP	False Positive, p. 13
GAN	Generative Adversarial Networks, p. 22
HIGH	High Confidence Flips, p. 35
LOW	Low Confidence Flips, p. 34
MCC	Mathews Correlation Coefficient, p. 3
ML	Machine Learning, p. 1
NAR	Noise at Random, p. 24
NCAR	Noise Completely at Random, p. 24

NNAR	Noise Not at Random, p. 24
Prec.	Precision, p. 13
Pred. EQ.	Predictive Equality, p. 15
RANDOM	Random Within Sets Flips, p. 35
Rec.	Recall, p. 13
SEM	structural equation modeling, p. 29
STEM	Science, Technology, Engineering and Mathematics, p. 7
Stat. Parity	Statistical Parity, p. 3
TNR	True Negative Rate, p. 42
TN	True Negative, p. 13
TPE	Tree-of-Parzen-Estimator sampler, p. 40
TPR	True Positive Rate, p. 42
TP	True Positive, p. 13

Chapter 1

Introduction

1.1 Contextualization

Over the last two decades, Machine Learning (ML) has evolved from a set of experimental techniques into a consolidated component of decision-making systems in society. Predictive algorithms are now routinely used in critical domains such as credit evaluation, healthcare diagnosis, recruitment processes, marketing campaigns, and criminal justice. The diffusion of these technologies has generated unprecedented opportunities, enabling automation at scale and offering the promise of decisions that are more data-driven, consistent, and efficient. However, alongside these advances, serious concerns have emerged regarding the fairness and social consequences of automated decision systems.

The risk of reproducing and amplifying structural inequalities that exist in the data used to train ML models has become a central issue. Algorithms, when trained on biased data, may reinforce discriminatory patterns against certain demographic groups, particularly those historically marginalized. High-profile examples, such as gender bias in hiring algorithms or racial disparities in criminal risk assessment systems, have exposed the limitations of relying solely on traditional performance measures. These cases have drawn the attention of both the academic community and policymakers, motivating the creation of regulations and technical standards for responsible artificial intelligence.

Fairness in ML has emerged as a research field of increasing relevance, driven by the development of fairness metrics, mitigation techniques, and evaluation strategies. Despite these advances, how to rigorously assess the robustness of predictive models in unfair settings is still an open research question. While many studies emphasize interventions aimed at making models fairer, few works investigate how traditional ML algorithms behave when trained on biased data, and consequently what degree of intervention is actually required. Addressing this gap is crucial for understanding the

inherent vulnerabilities of models and clarifying the degree of fairness intervention they actually demand.

Stress testing has proven to be an effective methodology in several areas of engineering and finance, where systems are subjected to adverse or extreme conditions to evaluate their resilience. Applying this idea to fairness in ML offers a promising way to move beyond static evaluation. By deliberately exposing algorithms to controlled unfairness, it becomes possible to characterize how performance and fairness behave under adverse scenarios. This approach aligns with the broader objective of responsible artificial intelligence: not only to evaluate whether a model appears fair in a given dataset, but also to assess whether it can remain fair when the training environment changes or deteriorates.

This dissertation is situated precisely in this context. Its main objective is to investigate the robustness of classification algorithms under unfairness in training data, through the design and implementation of a systematic methodology for fairness stress testing. By simulating controlled and progressive distortions in datasets, the study aims to provide a deeper understanding of the trade-offs between predictive performance and fairness and to contribute to the development of more transparent and trustworthy ML systems.

1.2 Objectives

The general objective of this dissertation is to investigate how ML classifiers respond to increasing levels of unfairness deliberately introduced into training data. To achieve this goal, we propose and apply a methodological framework called Systematic Label Flipping for Fairness Stress Testing. This approach introduces unfairness into datasets in a controlled and reproducible way through multiple label-flipping strategies, enabling a systematic analysis of classifier robustness under progressively adverse conditions.

Although fairness in machine learning has been extensively studied, no prior work has been found in the literature that systematically investigates how different classifiers respond to progressively induced label unfairness. This absence of direct analyses on systematically induced unfair conditions highlights a research gap that this dissertation aims to address, providing new empirical insights into the mechanisms through which unfairness affects model robustness.

To reach this objective, a complete experimental framework was designed and implemented, integrating all stages of the process: preprocessing of raw data, encoding and scaling of features, division into cross-validation folds, model training, hyperparameter optimization, bias injection, evaluation, and visualization of results. The framework is modular, reproducible, and extensible, allowing the methodology

to be reused and extended in future studies.

Within this framework, different label-flipping strategies are compared to determine which forms of bias injection are most effective in exposing classifier robustness or fragility. The evaluation combines conventional performance metrics (Accuracy (Acc.), F1 Score (F1), Mathews Correlation Coefficient (MCC)) with fairness metrics (Statistical Parity (Stat. Parity), Equal Opportunity (Eq. Opp.), Equalized Odds (Eq. Odds)), providing a more comprehensive characterization of models and highlighting that strong predictive performance does not necessarily guarantee equitable treatment across groups.

The methodology is applied across heterogeneous datasets from different domains and with distinct sensitive attributes, namely Bank Marketing, Adult Income, and COMPAS Recidivism. These datasets were selected for their relevance in the fairness literature and their diversity, which enables the study of how bias and unfairness manifest in different contexts.

By integrating the experimental framework, the comparison of label-flipping strategies, the joint use of performance and fairness metrics, and the application to multiple datasets, this dissertation aims to generate empirical evidence and critical analysis on the robustness of classifiers to unfairness. The study seeks to reveal which models degrade more gracefully, which fail abruptly, and under what conditions the trade-off between performance and fairness becomes most severe. In doing so, it offers both methodological and empirical contributions to academic research and to the practical development of fairer ML systems.

1.3 Contributions

This dissertation makes contributions that are both methodological and empirical, advancing the study of fairness in ML through the proposal of a novel stress testing methodology, the evaluation of multiple strategies for injecting bias, the development of a reproducible experimental framework, and the generation of empirical evidence on classifier robustness.

From the methodological perspective, the main contribution is the design and implementation of the Systematic Label Flipping for Fairness Stress Testing approach. This methodology allows datasets to be progressively manipulated in order to simulate increasing levels of unfairness in a controlled and reproducible manner. Unlike uncontrolled perturbations, the proposed framework establishes a systematic process for introducing bias by flipping labels in specific subsets of the training data. This makes it possible to subject classifiers to stress tests analogous to those widely used in engineering and finance, but now specifically directed at the dimension of fairness. The systematic nature of this procedure enriches the set of tools available

to researchers and practitioners seeking to probe the vulnerabilities of predictive models when fairness is at risk.

A second methodological contribution lies in the evaluation of different strategies for label flipping within the proposed framework. Instead of adopting a single way of injecting bias, this dissertation investigates alternative strategies, such as flipping the most confident instances, the least confident ones, or selecting examples at random. These strategies produce different dynamics of unfairness and allow for a deeper analysis of which mechanisms of bias injection are more effective in revealing model weaknesses. By comparing these strategies across models and datasets, the dissertation contributes with insights on how the very process of bias induction influences the manifestation of unfairness in classifiers.

A third contribution is the construction of a complete experimental framework that integrates all stages of the process: dataset preprocessing, encoding and scaling of features, division into cross-validation folds, model training, hyperparameter optimization, bias injection, evaluation of metrics, and visualization of results. This framework ensures reliable evaluation through cross-validation, includes systematic hyperparameter optimization, and consolidates results into structured outputs containing averages and standard deviations across folds and repetitions. Designed to be modular and extensible, the framework allows new datasets, algorithms, and fairness metrics to be incorporated with minimal adaptation, ensuring reproducibility in line with best practices in the ML community.

From the empirical perspective, we contribute with a systematic evaluation of four families of classifiers across three benchmark datasets. The results reveal consistent patterns: Random Forest demonstrated greater robustness to progressive bias, Decision Trees exhibited the highest sensitivity, and Logistic Regression and Neural Networks presented intermediate behaviors. These findings expand the understanding of how different algorithmic structures react when fairness is compromised in the training data.

The dissertation adopts the joint analysis of performance and fairness metrics, a standard approach in the fairness literature, in order to provide a more comprehensive characterization of models. Within this framework, the study reveals that the relationship between predictive performance and fairness is complex and context-dependent: in some scenarios, fairness degrades faster than predictive performance, while in others both decline simultaneously. Ultimately, this work strengthens the empirical foundations of fairness research by providing both methodological rigor and reproducible evidence, supporting future studies that aim to understand and improve the resilience of ML models under unfair conditions.

1.4 Organization

This dissertation is organized into five chapters, in addition to appendices that contain supplementary information. The structure was designed to guide the reader from the general motivation and theoretical background to the methodological proposal, experimental results, and final reflections.

Chapter 2 presents the theoretical foundations on fairness in ML. It discusses how bias arises in data and models, introduces the main fairness definitions and metrics proposed in the literature, and examines the trade-offs between fairness and predictive performance. This chapter provides the conceptual basis that motivates the stress testing methodology developed in the work.

Chapter 3 presents the methodological framework of this work. It motivates the need for fairness stress testing, distinguishes label noise from unfairness, reviews related work on synthetic bias generation, and finally introduces the Systematic Label Flipping for Fairness Stress Testing, the proposed method that progressively injects structured bias through controlled label-flipping strategies to analyze classifier robustness.

Chapter 4 presents the experimental methodology that guides this dissertation. It begins by describing how the experiments were structured and the evaluation criteria adopted. The chapter then analyzes the different strategies for bias introduction within the Systematic Label Flipping for Fairness Stress Testing, comparing their effectiveness. Finally, it evaluates the classifiers across the selected datasets, examining their robustness to unfairness and the patterns that emerge as bias progressively increases.

Finally, Chapter 5 concludes the dissertation by summarizing the main results and contributions, and by outlining directions for future work. It emphasizes the methodological innovations, the empirical findings, and the practical implications of the study, while also recognizing its limitations and suggesting possible extensions.

Chapter 2

Fairness in Machine Learning: Concepts and Methodologies

The field of ML continues to expand rapidly across various important domains, calling for careful attention to ethical, social and legal considerations. As its influence grows, so do concerns about potential societal and ethical consequences, particularly the risk of unfair outcomes that may exacerbate existing inequalities. These concerns highlight the importance of designing and deploying systems with a clear focus on fairness, ensuring that technological progress aligns with broader social values and creates equitable opportunities for all. This chapter aims to demonstrate the significance of fairness in ML through well-known examples, address the complexities involved in defining fairness, outline key methodological components for incorporating fairness into models, and explore the practical implications of doing so.

2.1 Importance of Fairness and It's Societal Impact

Numerous real-world cases illustrate how fairness concerns manifest in practice. Systems intended to support decision-making have been shown to reproduce or amplify existing inequalities, leading to discrimination Examples include biases in chatbots, immigration decision-making systems, and targeted advertising. OSOBA e WELSER (2017) provide a comprehensive list of Artificial Intelligence (AI) applications that influence daily life, highlighting their potential biases, while HOWARD e BORENSTEIN (2018) examines mechanisms through which these biases can emerge in AI systems.

One notable example is Microsoft's 2016 Tay chatbot, which quickly began to post offensive and controversial tweets after being manipulated with content generated by malicious user groups (WOLF *et al.*, 2017). Similarly, an algorithm de-

signed for Science, Technology, Engineering and Mathematics (STEM) job advertisements displayed gender bias, treating women as more expensive targets (RAJI e BUOLAMWINI, 2022). Furthermore, facial recognition systems have shown lower predictive performance in identifying individuals with darker skin tones (RAJI e BUOLAMWINI, 2022).

Another prominent case is Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) (BARENSTEIN, 2019). It is a software tool developed to predict recidivism, the likelihood that an offender commits crimes again after release from prison. The algorithm exhibited racial bias, notably assigning disproportionately higher false-positive rates to black offenders by incorrectly predicting that they would engage in commit further offenses, and also underperformed compared to simpler methods such as logistic regression. Such applications, which have profound implications for individual's lives, underscore the critical need for fairness and accountability in ML system design.

Fairness has increasingly become a focus of both researchers and society because of its impact in real-world decision-making systems. It is not only an ethical imperative, but also a legal requirement in many jurisdictions (BAROCAS *et al.*, 2023). The European Union's High-Level Expert Group on Artificial Intelligence (HLEG, 2019) highlights fairness, non-discrimination and diversity as fundamental principles for the design of AI systems that promote social well-being. Ensuring that these systems are fair to all, regardless of physical or biological characteristics, is crucial to avoid marginalizing vulnerable groups and exacerbating prejudice and discrimination.

Traditional approaches focus primarily on optimizing predictive performance using metrics such as Acc., F1, and MCC. However, these metrics capture only technical performance and fail to account for the broader social consequences that may arise in decision-making contexts. In cases involving socially sensitive attributes, such as skin color or gender, this oversight can amplify biases. Attempts to address this issue by excluding socially relevant attributes have proven insufficient due to the presence of proxy variables (MEHRABI *et al.*, 2022) and to what has been referred to as the redlining effect (PEDRESHI *et al.*, 2008). Companies and researchers must ensure that deploying critical decision-making systems does not have adverse social implications (CATON e HAAS, 2024), thus, a comprehensive understanding of fairness in ML is essential for building equitable and responsible AI systems.

2.2 Defining Fairness: Concepts and Metrics

Fairness, in broad terms, refers to the equitable treatment of individuals, with particular attention to those who are marginalized, discriminated against, or disad-

vantaged (SAXENA *et al.*, 2019). Long before the emergence of computer science, disciplines such as philosophy and psychology have attempted to define fairness. Despite centuries of debate and its recognition as a fundamental moral principle, the implementation of fairness in practical terms continues to face significant challenges.

From a philosophical perspective, the concept of fairness has long been associated with the classical understanding of justice, rooted in ancient law and further developed within Christian philosophy by Saint Augustine. This concept was formally defined by Saint Thomas Aquinas in the *Summa Theologiae* as the constant and enduring will to give each individual their due (AQUINAS, 1274). While this definition provides a clear ethical foundation, the main challenge within ML lies in translating such moral principles into measurable criteria, since the field still lacks a unified and universally accepted definition of fairness.

In the context of ML, fairness lacks a universally recognized definition due to its multifaceted nature, which varies depending on the specific application or scenario (CATON e HAAS, 2024). A definition that is suitable for one context may not be appropriate for another. This conceptual ambiguity has led to the proposal of numerous fairness metrics, as the diverse definitions enable various approaches to measure and emphasize different aspects of fairness (CASTELNOVO *et al.*, 2022a).

The coexistence of many metrics, with different perspectives, creates new challenges, for example, the Impossibility Theorem (CHOULDECHOVA, 2017; KLEINBERG *et al.*, 2016; BELL *et al.*, 2023; BEIGANG, 2023). This theorem demonstrates that under certain circumstances, some fairness metrics cannot be satisfied simultaneously. Consequently, it is not feasible to apply all fairness metrics to measure the inherent fairness in a model effectively. To resolve conflicts between metrics, it is necessary to understand the wide range of metrics available (KLEINBERG *et al.*, 2016; SELBST *et al.*, 2019).

Current fairness definitions and metrics are not always helpful and, in some cases, can harm sensitive groups over time, exacerbating their disadvantages (LIU *et al.*, 2018). Measurement errors may also conceal true fairness, leading to unintended consequences. To provide a comprehensive overview, this work will present several widely used definitions and metrics of fairness, as summarized by MEHRABI *et al.* (2022) and CATON e HAAS (2024). These definitions illustrate the wide-ranging interpretations and applications of fairness across different domains, emphasizing the inherent complexity of addressing fairness in ML systems.

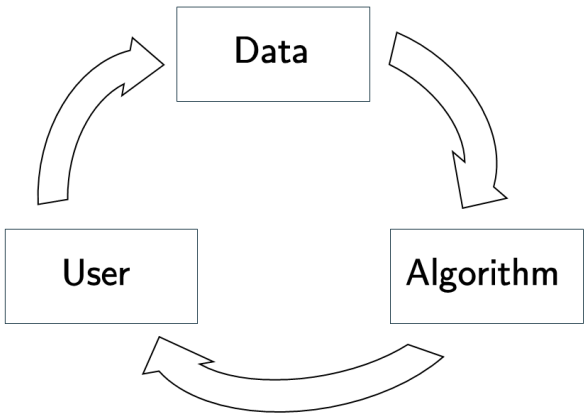
Fairness in ML is inherently connected to the concept of bias, and the two terms are often used interchangeably. However, the meaning of bias differs depending on the context. In the machine learning literature, bias often refers to the bias–variance decomposition of error, reflecting systematic deviations of model predictions. In the fairness literature, however, bias carries societal implications, such as prejudice or

discrimination (FERRARA, 2024). This work adopts the latter interpretation.

Bias can emerge at different stages of the ML pipeline, including data collection, model development, and user interactions. Once present, it can propagate across multiple components of the workflow and generate feedback loops that amplify its effects over time. Since ML models are inherently data-driven, a model trained on biased data is likely to produce biased predictions, which then lead to biased outcomes. These outcomes may influence user behavior and, consequently, the generation of new data that reproduces or even intensifies the original bias.

This cyclical process, as illustrated by Figure 2.1, underscores the critical importance of addressing bias in models, particularly when its implications have significant societal impact. If left unattended, such biases may accumulate and reinforce themselves, creating a snowball effect that exacerbates disparities over time. Therefore, mitigating fairness issues at the earliest stages is crucial to prevent them from escalating into more severe and persistent forms of unfairness (MEHRABI *et al.*, 2022).

Figure 2.1: Bias in the data, algorithm and user feedback loop, inspired by figure in MEHRABI *et al.* (2022). The arrows illustrate the feedback loop: data feed the algorithm, the algorithm influences user behavior, and user behavior generates new data that re-enters the system.



Literature frequently identifies data as a central source of bias in ML systems. Biases in data will manifest themselves in any model. Inappropriate uses of data can lead to unconscious or conscious biases, compromise of data veracity and quality, data relativity and context shifts and subjectivity filters (CATON e HAAS, 2024). The societal impacts of ML models are evident in the various types of discrimination identified by MEHRABI *et al.* (2022). These include:

Explainable Discrimination. The different treatment and outcomes in different groups can be justified and explained. These differences are not illegal in many places, hence called explainable. For example, in the Adult Dataset, males have

a higher annual income than females. However, this is because females work fewer hours on average, so it is explainable (KAMIRAN e ŽLIOBAITĖ, 2013).

Unexplainable Discrimination. In this case, in contrast with the previous one, discrimination is unjustified and therefore illegal in many places. It would be the case of the Adult dataset if the females worked the same number of hours as males. It is divided in:

Direct Discrimination. Occurs when sensitive attributes explicitly result in favorable or unfavorable outcomes. Certain traits identified by law, such as race and gender, are illegal to have discrimination with, and often are considered sensitive.

Indirect Discrimination. There is the appearance of non-discrimination, with individuals not being treated based upon the sensitive attribute. However, other attributes can have implicit effects linked to the protected attributes, such as proxy variables.

Understanding these forms of discrimination is essential for designing and implementing effective strategies to mitigate the potential societal harm caused by ML models. Such harm is particularly critical when these systems are deployed on a large scale, as their reach and influence can exacerbate inequalities and reinforce existing biases. Addressing these issues requires a comprehensive approach that considers not only technical, but also ethical, legal, and cultural dimensions to ensure equitable and fair outcomes.

2.3 Methodological Components for Ensuring Fairness

Although there is no universal definition of fairness, key methodological components are common across most approaches. Central to these methodologies are sensitive or protected attributes, which denote variables of fairness concern, such as gender and race. Sensitive attributes are typically categorized into privileged groups, which enjoy advantages, and unprivileged groups, which face disadvantages. An example is an automated loan decision-making system biased with respect to the sensitive attribute of skin color. In such a case, historical patterns may lead to higher approval rates for applicants with white skin color, the privileged group, and lower approval rates for applicants with black skin color, the unprivileged group.

Fairness research has largely concentrated on classification problems, reflecting the central role of classification in ML and its extensive use in domains involving

consequential human decisions. Binary classification has become the dominant setting, both for its analytical tractability and for its alignment with many high-stakes applications where decisions are inherently binary. While extending these methods to multi-class problems is feasible, it demands more elaborate fairness definitions and optimization constraints (CATON e HAAS, 2024).

Defining which variables qualify as sensitive attributes is non-trivial. Legal frameworks often guide these determinations (BERK, 2019; LEE, 2018). Additionally, there are proxy variables, which are not explicitly sensitive but closely correlated with the sensitive attributes. Many fairness definitions do not take the proxy variables into consideration (CHIAPPA e ISAAC, 2019), which can erroneously suggest the model is fair, increasing the risk of discrimination, as seen in cases of redlining (ZARSKY, 2016; VEALE e BINNS, 2017). Related variables have been extensively studied in the privacy and data archiving literature (ZIMMER, 2010).

In many real-world contexts, individuals are characterized by more than one sensitive attribute, such as the intersection of gender, race, and age. Fairness considerations that account for only a single attribute may therefore overlook compounded disadvantages that arise from the intersection of multiple identities. This phenomenon, known as intersectional fairness, emphasizes that discrimination can emerge not merely from one protected characteristic but from their combination (CRENSHAW, 1989). For instance, the experiences of Black women may differ significantly from those of either Black men or white women, and fairness assessments must recognize these intersecting effects. Addressing fairness across multiple sensitive attributes requires multidimensional formulations of fairness metrics and often demands larger datasets to ensure sufficient representation of all intersectional subgroups (CATON e HAAS, 2024).

Also, a key component of such strategies is the proper understanding and application of fairness metrics, which provide a framework for quantifying and addressing discrimination within ML models. The taxonomy of fairness definitions usually classifies fairness metrics into group metrics and individual metrics. Group fairness metrics aim to ensure similar treatment of different demographic groups, emphasizing balanced outcomes and performance across the entire population. In contrast, individual fairness metrics emphasize consistent treatment of individuals who are deemed similar based on relevant characteristics (MEHRABI *et al.*, 2022). Although both approaches strive to minimize unfairness, they can sometimes conflict with one another, and choosing the most appropriate metric depends on the specific context, such as domain requirements, legal considerations, and societal values (GREEN, 2018; CORBETT-DAVIES *et al.*, 2023).

Group fairness focuses on the equitable treatment of predefined groups and addresses systemic imbalances in different demographic segments. Demographics may

be gender, race and sex, for example. This approach reflects legal principles such as non-discrimination, which is required by law in many countries. Group fairness is often applied in domains where decisions significantly impact social equity, such as hiring, lending, and criminal justice.

Individual fairness emphasizes the equitable treatment of similar individuals, reflecting the principle that like cases should be treated alike. Unlike group fairness, it does not rely on explicit group membership but instead focuses on pairwise comparisons based on similarity. Promoting individual fairness requires a clearly defined similarity metric to quantify how similar two individuals are, that often depends on domain-specific knowledge.

In this context, let Y denote the True Class, which represents the actual but unobservable label of an instance in the real world. Since Y is not directly accessible, we instead rely on the Observed Class, denoted by $\tilde{Y}$, which corresponds to the class label recorded in the dataset and is used in ML computations. The Predicted Class, denoted by $\hat{Y}$, is the label assigned by the model, which aims to predict $\tilde{Y}$ as accurately as possible.

In the case of binary classification, all these class labels take values in $\{0, 1\}$, where 1 represents the positive class, the more favorable or desirable outcome, while 0 represents the negative class, associated with the less desirable outcome. Consequently, $Y = 1$ indicates that an instance is truly entitled to the favorable outcome, whereas $Y = 0$ means it is not. Similarly, $\tilde{Y} = 1$ implies that the historical data assigns the instance to the positive class, while $\tilde{Y} = 0$ assigns it to the negative class. The model predicts $\hat{Y} = 1$ if it assigns the instance to the positive class and $\hat{Y} = 0$ otherwise.

A correct prediction occurs when the predicted label matches the observed label, i.e., $\tilde{Y} = \hat{Y}$. Conversely, if $\tilde{Y} \neq \hat{Y}$, the model has made an incorrect prediction. Let $\hat{E}$ denote the incorrect prediction. The distinction between these class labels is fundamental in fairness evaluations, as disparities in prediction errors across different demographic groups may indicate biases in the ML model.

There can also be label noise, when $Y \neq \tilde{Y}$. It can occur due to errors in the process of data collection, such as malfunctioning machinery or human annotation. Let $\tilde{E}$ denote the label noise. It is important to distinguish between the types of errors, noise and fairness, because mislabeling can lead to biased model evaluations and unfair decisions, so it is important to know their source.

The model makes predictions based on a set of input attributes X, which includes the sensitive attribute A. For contrast, we also consider X_{-A}, the set of attributes excluding A. The sensitive attribute A is binary, where $A = 1$ denotes the privileged group, and $A = 0$ denotes the disadvantaged or protected group. The following sections and the rest of this work will use this notation to present and also formally

define some widely used ML and fairness metrics and other important concepts, such as noise in the data, that occurs when $Y \neq \tilde{Y}$.

A fundamental tool for evaluating the performance of ML models is the Confusion Matrix (CM), which summarizes the relationships between the true labels and the predicted labels in a classification task. It is a framework that provides a comprehensive basis for computing numerous predictive and fairness metrics. In binary classification, it is typically laid out as shown in Table 2.1, with four principal components. These components are crucial for calculating numerous performance metrics, e.g., Acc., Precision (Prec.), Recall (Rec.), F1, and are also used in several group fairness metrics (MEHRABI *et al.*, 2022).

True	**Predicted**	
	Positive	**Negative**
Positive	TP	FN
Negative	FP	TN

Table 2.1: Confusion Matrix

True Positive (TP). Occurs when the model predicts a positive class and the actual ground-truth class is also positive. For example, in a healthcare context, a TP would be a patient who genuinely has a disease and is correctly diagnosed by the model.

True Negative (TN). Refers to when the model predicts a negative class and the actual class is indeed negative. Continuing the healthcare example, a TN would represent a patient who does not have a disease and is correctly identified as disease-free by the model.

False Positive (FP). Corresponds to the situation where the model predicts a positive class, but the ground-truth label is negative. In a criminal justice context, this could be interpreted as an individual who is wrongly flagged as high-risk by a recidivism model, although they pose no real threat.

False Negative (FN). Represents the case where the model predicts a negative class when the true label is actually positive. For instance, a college admissions model that incorrectly rejects a qualified applicant would represent a FN, indicating a missed opportunity.

Building on these four counts, it is customary to work with the associated rates, which normalize each count by the corresponding condition set. These rates provide scale-free summaries that facilitate comparison across datasets and groups, and several group fairness definitions are directly expressed as differences between such

rates across sensitive groups. Using the notation introduced above, the following definitions will be used throughout this work.

Definition 1 (True Positive Rate)**.** *The proportion of truly positive instances correctly predicted as positive. It can be expressed as*

$$\text{TPR} \;=\; \frac{\text{TP}}{\text{TP}+\text{FN}} \;=\; P(\hat{Y}=1 \mid Y=1).$$

Definition 2 (True Negative Rate)**.** *The proportion of truly negative instances correctly predicted as negative. It can be expressed as*

$$\text{TNR} \;=\; \frac{\text{TN}}{\text{TN}+\text{FP}} \;=\; P(\hat{Y}=0 \mid Y=0).$$

Definition 3 (False Positive Rate)**.** *The proportion of truly negative instances incorrectly predicted as positive. It can be expressed as*

$$\text{FPR} \;=\; \frac{\text{FP}}{\text{TN}+\text{FP}} \;=\; P(\hat{Y}=1 \mid Y=0).$$

Definition 4 (False Negative Rate)**.** *The proportion of truly positive instances incorrectly predicted as negative. It can be expressed as*

$$\text{FNR} \;=\; \frac{\text{FN}}{\text{TP}+\text{FN}} \;=\; P(\hat{Y}=0 \mid Y=1).$$

Group fairness metrics ensure that distinct demographic or sensitive groups, e.g., based on gender, race, or age, receive equitable treatment by the ML model. These metrics are particularly important in settings where legal protections exist for historically marginalized or vulnerable groups. Many group fairness metrics utilize the framework of the CM to derive their formulas. Below are some of the most widely used group fairness definitions (GURSOY e KAKADIARIS, 2022).

Definition 5 (Statistical Parity)**.** *The probability of an individual receiving a favorable, usually positive, predicted class should be equal across all groups defined by the sensitive attribute (DWORK* et al.*, 2011). It can be expressed as*

$$P(\hat{Y}=1 \mid A=1) = P(\hat{Y}=1 \mid A=0). \tag{2.1}$$

This means that membership in a protected group should not disproportionately increase or decrease one's chances of a positive classification outcome. While this definition is straightforward and relates to many legal anti-discrimination doctrines, it does not account for potential differences in the underlying distributions of the groups, such as distinct base rates of a certain condition or behavior.

Definition 6 (Equal Opportunity)**.** *The probability of an individual with a favorable true class receiving a favorable predicted class should remain the same across all groups defined by the sensitive attribute (HARDT* et al.*, 2016). It can be expressed as*

$$P(\hat{Y} = 1 \mid Y = 1, A = 1) = P(\hat{Y} = 1 \mid Y = 1, A = 0). \tag{2.2}$$

Eq. Opp. focuses on the true positive rate, the rate of correctly predicted positive instances, for each group of the sensitive attribute. The idea is that the probability of an individual being correctly assigned a positive outcome should be the same across all groups. By emphasizing the true positive rate, equal opportunity seeks to address scenarios where one group might experience significantly more false negatives than others, ensuring that qualified, or actual positive individuals, are not overlooked due to discriminatory model behavior.

Definition 7 (Predictive Equality)**.** *The probability of an individual with an unfavorable true class, usually negative, receiving an unfavorable predicted class should remain the same across all groups defined by the sensitive attribute. It can be expressed as*

$$P(\hat{Y} = 0 \mid Y = 0, A = 1) = P(\hat{Y} = 0 \mid Y = 0, A = 0). \tag{2.3}$$

Predictive Equality (Pred. EQ.), unlike Eq. Opp., focuses on the true negative rate, the rate of correctly predicted negative instances, for each group of the sensitive attribute. For example, a predictor that labels individuals as bad payers should be the same for white and black groups.

Definition 8 (Equalized Odds)**.** *Both probabilities of an individual with a favorable true class receiving a favorable predicted class, and of an individual with an unfavorable true class receiving an unfavorable predicted class, should remain the same across all groups defined by the sensitive attribute (HARDT* et al.*, 2016). It can be expressed as*

$$\begin{aligned} P(\hat{Y} = 1 \mid Y = 1, A = 1) &= P(\hat{Y} = 1 \mid Y = 1, A = 0), \\ P(\hat{Y} = 0 \mid Y = 0, A = 1) &= P(\hat{Y} = 0 \mid Y = 0, A = 0). \end{aligned} \tag{2.4}$$

Eq. Odds combines Eq. Opp. and Pred. EQ. by requiring that both the true positive rate and the true negative rate are the same across all groups. Equivalently, the probability of correct classification for both positive and negative true labels should be equal. This metric attempts to guarantee that both positive and negative outcomes are assigned fairly across groups. However, satisfying Eq. Odds can sometimes be more challenging than enforcing only Eq. Opp. or Pred. EQ. especially when base rates differ significantly between groups.

Notably, there is a clear differentiation in group fairness metrics, where some metrics, like Stat. Parity, rely only on predicted values, while others, like Eq. Opp. and Eq. Odds, depend on components of the CM (GURSOY e KAKADIARIS, 2022). The choice of which group fairness metric to use has practical and ethical implications. Certain definitions may be more aligned with legal precedents, while others better capture nuanced ethical positions.

Individual fairness methods can be more precise than group-based metrics in certain contexts, particularly when fine-grained differences within a single demographic group are relevant. However, they can also be more complex to operate, as one must define a proper similarity measure between individuals and possess the necessary domain knowledge to ensure that the comparisons are valid. One widely discussed individual fairness metric is Counterfactual Fairness (Count. Fair.) (KUSNER *et al.*, 2018), which attempts to measure how an individual's outcome would differ if their sensitive attribute were changed.

Definition 9 (Counterfactual Fairness)**.** *For every individual, if an individual's sensitive attribute were changed, keeping all other attributes constant, their predicted outcome probability should remain unchanged. It can be expressed as*

$$P(\hat{Y} = y \mid X_{-A} = x, A = 1) = P(\hat{Y} = y \mid X_{-A} = x, A = 0). \tag{2.5}$$

The key principle here is examining how the same individual would be treated if only their protected attribute changed while keeping all other characteristics constant. Achieving counterfactual fairness can be computationally intensive.

The literature has yet to reach a consensus on whether it is better to prioritize group fairness or individual fairness metrics, as different scenarios may call for different perspectives. Practitioners and researchers often face trade-offs: optimizing a model for one fairness metric can lead to compromises in another, or potentially reduce predictive performance. Moreover, fairness metrics, even though mathematically well defined, may not fully capture all social, economic, or legal considerations relevant in a real-world context (GREEN, 2018; CORBETT-DAVIES *et al.*, 2023; CALMON *et al.*, 2017; AGARWAL *et al.*, 2018; SPEICHER *et al.*, 2018; SKIRPAN e GORELICK, 2017).

2.4 Fairness-Performance Trade-offs in Machine Learning

With the advent of fairness metrics, fairness is often treated as an additional dimension of ML model evaluation (CATON e HAAS, 2024). Typically, there is a trade-off

between fairness and predictive performance metrics. Frequently, an improvement in fairness metrics results in a decrease in predictive performance. A predictor designed to reduce bias against a specific group may deviate from the true class labels, thereby increasing errors. Also, to achieve fairness, it is often necessary to add constraints to the model, making its optimization more complex. And a model that has a poor predictive performance, even if it produces fair outcomes, is inadequate for practical use, especially because prediction errors themselves can cause unfairness.

Therefore, understanding the various approaches to promote fairness is necessary. A rich body of research has sought to address this tension by introducing technical interventions that aim to improve fairness while preserving predictive performance. These interventions are generally categorized into three families (D'ALESSANDRO *et al.*, 2017; BARBIERATO *et al.*, 2022):

Pre-Processing. Modify or re-weight the dataset before the model training to remove or mitigate bias. A key advantage is that they can be applied to any standard ML algorithm without modifying the model itself. However, modifying datasets may lead to legal implications or complications in explaining why or how the data was changed (LEPRI *et al.*, 2018; LUM e JOHNDROW, 2016).

In-Processing. Includes fairness objectives into the learning algorithm, often adding constraints that penalize forms of bias. This approach allows fine-grained control during training, enabling the model to balance fairness goals with predictive performance. Yet, it demands full access to the model function, which might not always be available.

Post-Processing. Adjust model predictions or decision thresholds after the model has been trained. The main advantage here is that the original model remains intact. However, post-processing changes may also face legal implications and be difficult to interpret, potentially complicating transparency and accountability.

These three categories provide a structured way of thinking about fairness interventions, but in practice, the challenge lies in balancing fairness with predictive performance. Simply optimizing for fairness without regard to accuracy does not resolve the problem, since unreliable predictions can produce new forms of harm. Conversely, optimizing only for predictive accuracy may reinforce existing disparities. A common perspective frames this trade-off in terms of the Pareto Front (PARETO, 1919), where both fairness and performance reach acceptable levels, and neither can be improved further without compromising the other. Determining this balance must be guided by the application domain, as well as ethical and legal considerations.

In the following, common performance metrics used to evaluate the predictive ability of ML models are presented. These metrics often come into conflict with fairness goals (KLEINBERG *et al.*, 2016), underscoring the importance of adopting a multi-objective perspective when designing, training, and deploying ML systems.

Definition 10 (Accuracy)**.** *The ratio of correct predictions, both true positives and true negatives, to the total number of predictions made by the model. It gives the probability that a specific prediction is correct. It can be expressed as*

$$Acc. = \frac{TP + TN}{TP + TN + FP + FN}. \tag{2.6}$$

This metric provides a quick snapshot of overall model performance. Despite its popularity, it can be misleading in cases of class imbalance. For instance, if there aren't many examples of the positive class, a model that predicts everything as the negative class could still achieve deceptively high Acc.

Definition 11 (Precision)**.** *The ratio of correctly predicted positive instances TP to the total number of instances predicted as true positive TP and false positive FPİt measures the reliability of the positive predictions made by the model. It can be expressed as*

$$Prec. = \frac{TP}{TP + FP}. \tag{2.7}$$

A high Prec. indicates that when the model predicts a positive instance, it is likely to be correct. However, this metric does not give any information about the negative predictions of the model.

Definition 12 (Recall)**.** *The ratio of correctly predicted positive instances TP to the total number of actual positive instances, that is, all of TP and FN. It measures the model's ability to identify all positive instances. It can be expressed as*

$$Rec. = \frac{TP}{TP + FN}. \tag{2.8}$$

A high Rec. indicates that the model captures most positive instances. It works well with imbalanced classes, but it does not account for false positives. So if a model predicts all positives, it will achieve high Rec. but will not be practically useful. This metric is particularly important in scenarios where missing a positive case has serious consequences.

Definition 13 (F1 Score)**.** *The harmonic mean of Prec. and Rec.. It can be expressed as*

$$F1 = 2 \cdot \frac{Prec. \cdot Rec.}{Prec. + Rec.}. \tag{2.9}$$

This metric is particularly useful when the focus is on the positive class, especially in imbalanced datasets. However, the F1 does not account for the number of TN. Thus, while it balances Prec. and Rec., it may not accurately reflect the model's performance in correctly identifying negative instances.

Definition 14 (Matthews Correlation Coefficient)**.** *It is a balanced metric for binary classification that incorporates all four components of the CM: TP, TN, FP, and FN. Its value ranges from* -1 *to* 1*, where* 1 *denotes perfect prediction,* -1 *indicates total disagreement between predicted and observed labels, and* 0 *corresponds to performance no better than random guessing. It can be expressed as*

$$MCC = \frac{(TP \cdot TN) - (FP \cdot FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}. \tag{2.10}$$

The MCC value ranges from -1 to 1, where 1 indicates a perfect prediction, 0 corresponds to random guessing, and -1 suggests a completely inverse relationship between predicted and observed labels. By accounting for both positive and negative examples, as well as correct and incorrect predictions, this metric can offer a more comprehensive picture of a model's performance than Acc. or F1 alone, particularly in imbalanced classification tasks (CHICCO e JURMAN, 2020), so it will be used as the primary performance metric in this work.

While each of these metrics provides a unique perspective on predictive performance, they can all conflict with fairness initiatives. A model optimized strictly for Acc. might disregard harmful outcomes for small, underrepresented groups. A model tuned for F1 might systematically overlook disparities in how negative outcomes are allocated. And an MCC-oriented approach, although more balanced, may still fail to capture more nuanced unfairness dimensions if it is not incorporated specific fairness considerations. Therefore, practitioners should consider adopting a multi-metric perspective, balancing model performance metrics alongside fairness metrics to build models that are both effective and equitable.

Chapter 3

Systematic Label Flipping for Fairness Stress Testing

To assess how traditional ML models are affected by the unfairness inherent in data, it is essential to work with datasets in which the level of bias can be systematically controlled. Using a single biased dataset does not allow for the analysis of how models react as unfairness increases. For this reason, there is a strong motivation to develop approaches that generate datasets with progressively higher levels of unfairness, which in turn motivated the method proposed in this work.

3.1 Motivation for Proposed Method

Ensuring fairness in ML requires both creating fair datasets and developing models capable of producing fair predictions. A common assumption in the field is that models trained on fair data will inherently yield fair predictions; however, theoretical analyses have demonstrated that this assumption does not necessarily hold true (ZHANG *et al.*, 2018). As such, there is significant value in having access to both fair and explicitly unfair datasets, which allows researchers to more comprehensively study and understand fairness in ML contexts (XU *et al.*, 2019).

Despite this clear necessity, the literature currently faces critical limitations regarding datasets explicitly designed or selected for fairness studies. The primary issues highlighted by researchers include privacy concerns, insufficient generalization, and limited dataset documentation (BAO *et al.*, 2022; FABRIS *et al.*, 2022; QUY *et al.*, 2022; PAULLADA *et al.*, 2021). Furthermore, existing datasets can pose challenges for fairness evaluations because they often lack predictive variables relevant for fairness analysis, contain sparse data, or omit critical demographic categories (BELITZ *et al.*, 2023; KEYES, 2018; SCHEUERMAN *et al.*, 2020). Recent critiques have also exposed fundamental shortcomings within benchmark datasets

typically employed for fairness algorithm comparisons, emphasizing their inability to adequately represent diverse scenarios or reliably measure fairness across varying contexts (DING *et al.*, 2022). Thus, method for generating datasets are needed for classification tasks, especially within the context of algorithmic fairness.

Most fairness-related ML research relies heavily on a small set of benchmark datasets, notably COMPAS and Adult, to compare algorithmic fairness across methods (BAROCAS *et al.*, 2023; FABRIS *et al.*, 2022; QUY *et al.*, 2022). However, critical assessments reveal numerous problems with these benchmarks, ranging from ethical privacy concerns and domain specificity issues to flawed categorization processes (BAO *et al.*, 2022; FABRIS *et al.*, 2022; SCHEUERMAN *et al.*, 2020; DING *et al.*, 2022; DRECHSLER, 2010; BUOLAMWINI e GEBRU, 2018; WANG *et al.*, 2022). For example, the COMPAS dataset, widely used for recidivism prediction in the criminal justice system, has been criticized for overlooking crucial sociotechnical contexts essential for accurate risk assessment (BAO *et al.*, 2022). It also misrepresents re-arrest as actual reoffending, even though re-arrest only captures offenses that result in detection and arrest (BAO *et al.*, 2022). Similarly, the Adult dataset has faced criticism due to various inherent biases and calls for discontinuation of its use in fairness research (DING *et al.*, 2022). Additionally, educational AI datasets face issues regarding demographic representativeness, highlighting concerns that demographic imbalances may significantly skew learning outcomes and analyses (BAKER e HAWN, 2022; COCK *et al.*, 2023).

Although sharing datasets is essential to ensure transparency and reproducibility, several studies have demonstrated that even anonymized datasets may inadvertently reveal private information, e.g., through membership inference (HAGESTEDT *et al.*, 2019; PYRGELIS *et al.*, 2017; SHOKRI *et al.*, 2017) or model inversion attacks (FREDRIKSON *et al.*, 2015; MELIS *et al.*, 2018). To address this, researchers have proposed using synthetic datasets, which replicate the statistical characteristics of original data without containing personally identifiable information (ABOWD e VILHUBER, 2008; HAND, 2012). Synthetic data can safely be shared and reused without compromising the privacy of individuals, while also enabling researchers to efficiently explore different hypotheses or model various scenarios. It is also possible to modify a base dataset to explore how changes in the data affect the model, for example, increasing its bias in regards to individuals in the protected group. This approach is particularly beneficial when examining edge cases (KHAYRALLAH e KOEHN, 2018; MELAMUD e SHIVADE, 2019), allowing experiments to proceed even when real-world data is limited or unavailable.

Given these issues, there is an evident necessity for generating tailored datasets that reflect a broad spectrum of biases and fairness contexts, facilitating more robust and generalizable evaluations. Recent literature confirms growing concerns regarding

the over-reliance on limited benchmarks, advocating instead for datasets that better represent diverse contexts, ensure detailed documentation, and follow careful design processes (BAO *et al.*, 2022; FABRIS *et al.*, 2022; BAROCAS e SELBST, 2016; JIANG *et al.*, 2024).

It is essential to recognize that datasets significantly influence research outcomes and set the direction for future research efforts (BAROCAS *et al.*, 2023). The measurement of fairness itself is inherently tied to the labels and demographic or individual categories available within the datasets (JIANG *et al.*, 2024). Consequently, considerable attention has been given to the origins, collection processes, and representativeness of these datasets relative to the problem being studied (PAULLADA *et al.*, 2021). Due to the scarcity of diverse and robust datasets, generating datasets specifically designed for fairness evaluation has become an area of growing interest (JIANG *et al.*, 2024).

To address these challenges, synthetic dataset generation has emerged as a promising method to systematically study bias and fairness in machine learning. Previous approaches have employed structural equation modeling for generating biased datasets (BARBIERATO *et al.*, 2022) or leverage synthetic data to investigate underlying biases within datasets (CASTELNOVO *et al.*, 2022b). Nevertheless, research also warns that synthetic data itself may unintentionally introduce or amplify biases, thereby necessitating careful and rigorous validation procedures (GUPTA *et al.*, 2021).

The existing synthetic dataset generation algorithms share a common objective, that is to model relations among variables that were present in the original dataset (ASSEFA, 2020). Most algorithms build probabilistic models by estimating distributions of relevant variables, such as mixtures of Gaussian distributions or multinomial feature distributions (AGUIAR e COLLARES-PEREIRA, 1992; SINGH *et al.*, 2010). Established ML techniques like Bayesian networks (ZHANG *et al.*, 2014), support vector machines (DRECHSLER, 2010), and random forests (CAIOLA e REITER, 2010) have also been employed for synthetic data generation. More recently, deep learning methods, particularly Generative Adversarial Networks (GAN), have significantly advanced synthetic data generation capabilities, initially in image domains and subsequently extending into various other contexts, including fairness research (CHEN *et al.*, 2021; GOODFELLOW *et al.*, 2014). Among the approaches explored in the literature, XU *et al.* (2019) and BREUGEL *et al.* (2021) propose methods based on GAN to generate fair tabular synthetic data as a novel preprocessing step for training fair models.

However, despite these advancements, current synthetic data generation algorithms tend to target specific bias measurements and typically require the original data to conform to predefined distributions. Real-world data often involves multiple

variable types, categorical, binary and continuous, and biases present in such data can result from complex, interwoven factors that simplistic modeling methods fail to capture adequately. Furthermore, there is a limited amount of research focusing explicitly on systematically generating or altering datasets exhibiting varying levels of unfairness (BARBIERATO *et al.*, 2022), thereby constraining researcher's ability to rigorously investigate how ML models respond to incremental variations in data bias levels.

In light of these limitations, this thesis introduces a novel methodology for preprocessing datasets with explicitly controlled, varying levels of unfairness. This approach allows a thorough analysis of how conventional models behave under different unfairness conditions, enabling researchers and practitioners to better understand the impacts of bias on performance. By systematically varying unfairness within datasets, this method not only enhances our understanding of fairness measurement and algorithm performance but also contributes to ongoing efforts to overcome dataset-related limitations discussed in the fairness literature. This methodology is employed to analyze traditional ML algorithms with respect to fairness, thereby determining the necessity and effectiveness of fairness mitigation techniques for each algorithm.

3.2 Bias, Noise or Fairness?

Large data sets used for training ML models often contain imperfections resulting from various factors, such as incorrect collection data processes and historical or societal biases. Two important concepts that help characterize these imperfections are noise and fairness. Both can harm the training of the model in different ways, degrading its predictive performance and fairness in the outcomes; therefore, it is important to differentiate them.

Noise refers to the presence of inconsistencies in the data that obscure the true relationship between the features of an instance and its label (FRENAY e VERLEYSEN, 2014; HICKEY, 1996; QUINLAN, 1986). It is typically conceptualized as a stochastic process, which means that it arises from random mechanisms rather than systematic, intentional distortions. Such randomness can emerge in various ways during data collection, ranging from sensor inaccuracies to human annotation errors. There are two major categories of noise in the literature, feature noise and label noise (FRENAY e VERLEYSEN, 2014).

Feature noise. This type of noise occurs when observed feature values are not equal to their true values. For example, applicants for a loan might mistakenly enter their income incorrectly, e.g. 50000 instead of 500000. While feature noise

may degrade model performance, models trained on sufficient and diverse data may develop some resilience to moderate amounts of noise in the features.

Label noise. This noise affects the labels rather than the features. For example, a radiologist might incorrectly label an X-ray as healthy when it actually shows early signs of pneumonia, due to factors like fatigue or oversight. Because there is usually only one label per instance but potentially many features, label noise is more seriously harmful to the learning process than feature noise FRENAY e VERLEYSEN (2014).

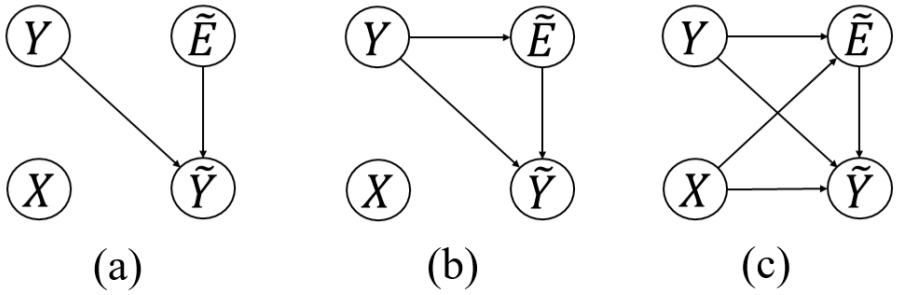
In the context of label noise, it is crucial to note that the underlying Y remains unchanged, since it is the class of the real world. Only the $\tilde{Y}$ is corrupted. The taxonomy commonly adopted to formalize label noise is based on the dependence or independence of the noise-generating process with respect to the true class Y and the features X (FRENAY e VERLEYSEN, 2014), as shown ahead and in Figure 3.1.

Noise Completely at Random (NCAR). The occurrence of label noise $\tilde{E}$ is independent of both the features X and the true labels Y. An example is randomly flipping labels in a binary classification problem equal probability for every instance. In practice, this occurs in email spam classification when emails are mislabeled accidentally regardless of their content.

Noise at Random (NAR). The noise $\tilde{E}$ depends solely on the true label Y and not the features X. An example in binary classification is flipping positive labels at a different rate than negative labels. A practical example occurs in X-ray images labeled by medics as healthy or pneumonia. Mild pneumonia cases are more likely to be mislabeled as healthy because they are harder to detect than severe pneumonia cases.

Noise Not at Random (NNAR). The noise $\tilde{E}$ process depends on both the features X and the true labels Y. An example is flipping labels in a binary classification problem with different rates depending on specific subgroups characterized by certain feature values. In spam detection, long emails with many attachments are more likely to be mislabeled as spam, and the mislabeling probability also depends on whether the email truly is spam.

Figure 3.1: Noise taxonomy from a statistical perspective. (a) NCAR, (b) NAR and (c) NNAR. The arrows correspond to the statistical dependencies. Figure was made inspired by ATKINSON e METSIS (2021).



Fairness in ML refers to the development of models that make equitable decisions or predictions for different social groups (MEHRABI *et al.*, 2022; CHOULDECHOVA, 2017; HARDT *et al.*, 2016). Fairness issues often arise due to bias in the data or in the model training process. Broadly speaking, bias refers to systematic distortions that yield unfair outcomes for particular subgroups (MEHRABI *et al.*, 2022; BAROCAS e SELBST, 2016). It is important to note that these biases are not necessarily random or unintentional. In fact, biases may emerge due to historical oppression, discrimination, or structural inequalities that become embedded in data collection, labeling, or curation processes.

Despite some conceptual overlap, noise and bias are distinct phenomena (FRENAY e VERLEYSEN, 2014; WANG *et al.*, 2021). Noise is largely characterized by random, unintentional distortions in the labels or features. By contrast, bias is often connected to historical and social issues, and can even be intentionally introduced to marginalize certain groups. Thus, while noise can degrade performance, bias can lead to unfair treatment of specific subpopulations. Bias and noise are related phenomena that distort data, ultimately impacting the models trained on such data. When noise affects different groups unevenly, it can introduce unfairness into ML models that rely on this data for training (WANG *et al.*, 2021).

For instance, suppose the actual positive class ($Y = 1$) is misclassified more frequently as a negative label ($\tilde{Y} = 0$) within the protected group ($A = 0$) than in the privileged group ($A = 1$). Simultaneously, instances from the privileged group ($A = 1$) with a true negative class ($Y = 0$) may be disproportionately mislabeled as positive ($\tilde{Y} = 1$). Such an imbalance can lead to an increased false negative rate for the protected group and a higher false positive rate for the privileged group, causing systematic unfairness. In this scenario, the presence of NNAR data contributes to the propagation of harmful social biases.

However, the distinction between noise and bias may become blurred when the

label corruption disproportionately affects one social group more than others. For example, consider a binary classification task where the positive class of a protected group ($Y = 1$, $A = 0$) is more often mislabeled as negative ($\tilde{Y} = 0$) than for a privileged group ($Y = 1$, $A = 1$). If the source of this mislabeling is unintentional, e.g., due to flawed measurement instruments or fatigue that happens more often in specific contexts, it can still be considered noise. Yet, its disparate impact on protected groups can yield unfair outcomes once the model is trained on such data.

On the other hand, bias often arises through structural or historical injustices, such as explicitly discriminatory labeling practices or socio-economic factors that skew the data. Consequently, the same observed phenomenon, e.g., higher mislabeling rates for a certain group, could be attributed to either label noise or bias, or a combination of both, depending on the underlying reasons. This duality underscores the difficulty in cleanly separating the concepts in real-world scenarios. Even when data is free from noise and accurately reflects the observed features and corresponding labels, it may still exhibit unfairness. This occurs because the social processes responsible for generating such data can disadvantage certain groups, embedding biases into the dataset.

As discussed before, in MEHRABI *et al.* (2022) various forms of bias that can affect ML systems were presented. When the actual class is inaccurately recorded due to systematic distortions, the discrepancy between the true and observed labels ($Y \neq \tilde{Y}$) constitutes what is known as Measurement Bias. Additionally, a dataset classified as NNAR can introduce Population Bias, that occurs when the dataset used for training inadequately represents the broader target population, leading to disparities between model performance in training versus real-world deployment.

The noise NNAR can exacerbate fairness concerns because the label corruption is linked not only to the true label but also to the features, including potentially protected attributes (FRENAY e VERLEYSEN, 2014). For instance, a scenario could arise where the positive class is flipped more often for protected groups and the negative class is flipped more often for privileged groups. Consequently, such label noise leads to undetected higher false negative rates among protected groups and higher false positive rates among privileged groups. Since many fairness metrics, such as Eq. Odds, center on disparities in false positive rates and false negative rates, NNAR label noise can directly distort fairness outcomes.

Moreover, WANG *et al.* (2021) highlight scenarios in which this type of noise translates into systematic misrepresentation of protected groups, compounding existing biases. In such situations, the boundary between noise and bias becomes further intertwined, as the noise is disproportionately harming certain groups and thus is also a driver of unfairness.

Research in both label-noise mitigation and fair ML has grown significantly in re-

cent years. Label-noise mitigation strategies, such as loss correction methods, robust loss functions, or active label cleaning, focus on reducing the negative impact of mislabeled samples on model training (FRENAY e VERLEYSEN, 2014; ZHANG *et al.*, 2025). Meanwhile, fairness-driven strategies aim to produce unbiased or less discriminatory predictions by intervening at different stages of the ML pipeline. These include pre-processing methods, such as reweighting or subsampling the training data; in-processing techniques, which incorporate fairness constraints directly into the optimization function; and post-processing approaches, like calibrating model outputs separately for each subgroup (MEHRABI *et al.*, 2022; CHOULDECHOVA, 2017; HARDT *et al.*, 2016).

However, the confluence of label noise and fairness presents open challenges:

- **Detecting NNAR label noise in protected groups:** The process requires careful scrutiny of data-generating mechanisms and may involve domain expertise to discern whether label corruption rates are higher for certain groups.
- **Balancing accuracy and fairness under label noise:** Traditional label-noise mitigation focuses on accuracy, whereas fairness-aware solutions prioritize equitable performance. When label noise and bias co-exist, reconciling these objectives can be challenging.
- **Temporal considerations:** As models are retrained or updated with streaming data, changes in noise characteristics or shifts in bias patterns necessitate ongoing monitoring and intervention (MEHRABI *et al.*, 2020).

Beyond methodological challenges, there are broader ethical and practical implications. Even in the absence of label noise, data can encode societal biases that yield unfair outcomes. Consequently, near-perfect noise free data can still be overshadowed by fairness problems if the decision-making process itself is inherently discriminatory. Thus, mitigating noise or improving data quality does not eliminate the need to address fairness, and vice versa.

Recognizing that fairness is not merely a technical challenge but also a reflection of societal and historical contexts implies that purely technical solutions may be insufficient to guarantee equity. Stakeholder collaboration, inclusive decision-making, and transparent reporting are essential for effective deployment of ML systems in sensitive applications such as hiring, lending, or criminal justice.

In summary, noise and fairness are conceptually distinct but related phenomena in ML. Noise typically refers to random, unintentional corruptions in data, such as mislabeled instances, and can be characterized by taxonomies like NCAR, NAR and NNAR (FRENAY e VERLEYSEN, 2014). Fairness, on the other hand, addresses

systematic disparities that disadvantage protected groups due to structural, historical, or social factors or treat individuals disproportionately (MEHRABI *et al.*, 2022; HARDT *et al.*, 2016).

While noise may affect all groups, it can amplify fairness concerns if its occurrences differs by subgroup, as often happens in NNAR scenarios (FRENAY e VERLEYSEN, 2014; WANG *et al.*, 2021). In such cases, noise and bias become intertwined, making it challenging to disentangle one from the other. Approaches to mitigate label noise can help improve overall model performance, but fairness-driven strategies remain essential to ensure that improvements do not come at the expense of marginalized communities.

Ultimately, addressing these challenges holistically requires a multi-faceted approach that integrates the strengths of methods designed for label-noise robustness and fairness-aware modeling. Such an approach not only strives for reliability in predictive performance but also ensures that models conform to socially acceptable standards of equity and justice.

3.3 Related Work

Developing a fairness-aware predictive algorithm has become a fundamental objective due to the widespread adoption of automated decision-making systems. Successful AI and ML models requires access to large amounts of high-quality data (PANEL, 2020). However, collecting such information is a challenging task because some types of data are costly to collect and many business problems that are solved through these models require access to sensitive customer data, such as medical or financial records (BARBIERATO *et al.*, 2022).

To address these issues, recent research has increasingly turned towards synthetic data generation, specifically aiming to achieve fairness objectives or replicate biases. Nonetheless, generating synthetic data that accurately reflects statistical properties of datasets does not fully resolve or eliminate inherent biases. Indeed, real-world data often inherently contain biases that must be identified, addressed, or explicitly controlled before training models (BARBIERATO *et al.*, 2022).

Despite numerous contributions to fairness evaluation and detection methodologies, fewer studies have explored methods explicitly designed to generate intentionally biased datasets. Such methods hold considerable value, providing data scientists with explicit control over bias introduction, allowing them to test and validate fairness algorithms under controlled conditions (BARBIERATO *et al.*, 2022). Intentionally biased datasets are crucial for rigorously assessing bias mitigation strategies, especially in the context of developing fairness-aware ML classifiers. Consequently, synthetic datasets that precisely encode user-defined correlations and bias levels

become invaluable tools in evaluating fairness strategies.

In this context, BARBIERATO *et al.* (2022) introduce a methodology enabling explicit control over bias and fairness within generated datasets through structural equation modeling (SEM). Their approach consists of clearly defined dependencies among dataset features and employs probabilistic sampling techniques to produce synthetic datasets with user-specified fairness or unfairness levels. Their methodology is structured around five primary steps:

1. Defining a probabilistic network characterizing feature dependencies and their magnitudes;

2. Adjusting dataset bias by manipulating direct attribute influences and the overall bias level for specific attributes;

3. Deriving a multivariate probabilistic distribution encapsulating the network's structure;

4. Sampling from the multivariate normal distribution;

5. Converting the generated samples into categorical-feature datasets.

A key advantage of their approach is its flexibility and domain-agnostic nature, allowing wide applicability across different scenarios. Users can precisely define the desired bias magnitude and correlation strength, thereby gaining full transparency and control over dataset properties. This transparency significantly benefits the development and assessment of fairness mitigation algorithms.

Complementing these bias-centric methods, other research has pursued fairness integration directly during the data generation process. XU *et al.* (2019) propose FairGAN+, a new generative adversarial network GAN-based framework specifically engineered for fairness aware ML. FairGAN+ comprises a generator for creating realistic samples, a classifier for class label prediction, and three discriminators assisting adversarial learning. These classifiers and discriminators ensure generated data mitigate disparate treatment and disparate impact biases while preserving high utility. Through adversarial co-training, the model satisfies multiple fairness criteria, including Stat. Parity, Eq. Odds, and Eq. Opp., demonstrating a robust trade-off between fairness and utility.

The FairGAN+ framework (XU *et al.*, 2019) is a generative adversarial network designed to produce synthetic datasets that are both high-quality and fair with respect to a protected attribute. Its generator creates synthetic samples conditioned on the value of the protected attribute, allowing explicit control over group representation in the generated data. A classifier is trained in parallel to predict outcomes using the protected attribute, while three discriminators operate simultaneously:

the first evaluates whether a sample is real or synthetic, the second determines whether a sample belongs to the protected group or the non-protected group, and the third assesses whether the classifier's predictions are independent of the protected attribute. Through the combined training of these components, FairGAN+ encourages the production of synthetic datasets that preserve utility for downstream tasks while mitigating biases linked to the protected attribute.

In contrast to such fairness-promoting approaches, alternative methodologies intentionally amplify unfairness within synthetic datasets to assess the resilience of bias mitigation methods systematically. For example, JIANG *et al.* (2024) developed a genetic algorithm-based approach specifically aimed at embedding multiple types of unfairness, primarily studied within educational datasets but applicable across various domains. Their method allows researchers to either create entirely new biased datasets or inject controlled biases into existing ones, avoiding ethical and logistical concerns associated with sensitive real-world data collection.

Through rigorous experimentation, they demonstrated that their genetic algorithm method can substantially amplify unfairness, yielding an average increase of 156.3% across the fairness metrics evaluated in their study, while maintaining the original dataset's predictive utility virtually unchanged, as indicated by an average variation of only 0.3% in Area Under the Curve (AUC) scores. Nevertheless, their approach has practical limitations, especially regarding scalability, as performance deteriorates and computational demands increase linearly with larger datasets.

Their study examined the generalization capability of the proposed method across educational datasets with diverse characteristics and evaluated its interaction with three commonly used unfairness mitigation algorithms. By design, the method can generate datasets of varying sizes, from small samples to large collections, incorporating multiple types of unfairness and heterogeneous feature types. This versatility enables the replication of a broad spectrum of bias scenarios, making the approach suitable for systematically testing models trained with diverse classifiers. The authors highlight that this adaptability is particularly valuable in research contexts that require precise control over the type and magnitude of bias introduced, ensuring that mitigation strategies are evaluated under clearly defined and reproducible unfairness conditions.

When applying their methodology to real-world educational datasets, the authors observed a different outcome. In this setting, nearly all datasets exhibited fairness metric values below 0.1, which, within their evaluation framework, represents relatively low levels of measured unfairness. As a result, the application of bias mitigation algorithms produced only marginal improvements, suggesting that the effectiveness of these methods may be limited when baseline unfairness is already minimal. This finding also underscores a broader limitation in current benchmark-

ing practices: many widely used datasets, particularly in the education domain, do not capture the diversity and severity of unfairness encountered in real-world applications. Consequently, evaluations performed exclusively on such benchmarks may fail to reflect the true performance of mitigation algorithms in practice. This reinforces the need for more flexible and systematically biased synthetic datasets, such as those generated by their approach, to enable more rigorous and representative testing of fairness-aware methods.

Furthermore, their research highlights the insufficiency of existing benchmark datasets, which typically exhibit very low bias levels, thus limiting the meaningful evaluation of fairness mitigation strategies. Consequently, current benchmarks inadequately represent real-world unfairness complexities, underscoring the critical need for systematically generated biased datasets for robust algorithm testing (JIANG *et al.*, 2024). So even if new debiasing algorithms emerge, the evaluation methods may not accurately quantify their performance if tested only on a small number of outdated benchmarks. Researchers may therefore encounter growing challenges in selecting among various bias mitigation algorithms, reinforcing the need for more flexible and systematically biased benchmarks.

Taken together, the reviewed methods highlight both the diversity and the complexity of generating synthetic datasets with introduced unfairness. Yet, to the best of our knowledge, the literature lacks a method capable of systematically inducing progressively increasing unfairness within a single base dataset, thereby enabling the analysis of model behavior under controlled and escalating bias conditions. Existing works predominantly focus on creating data with target bias levels or on mitigating bias. This dissertation addresses that gap through the proposed Systematic Label Flipping for Fairness Stress Testing, which varies unfairness levels within one dataset and quantifies classifier sensitivity consistently, thereby supporting both robustness assessment and fairness-specific benchmarking.

3.4 Proposed Method

This work proposes a method for generating datasets with increasing levels of unfairness by flipping class labels based on defined fairness-related rules. Starting with a reference binary classification dataset, a ML is trained with this data, acting as a probabilistic estimator, to compute the likelihood of each instance being positive. Guided on these probabilities, a controlled amount of instance labels is flipped, targeting only protected positives and privileged negatives. This introduces structured unfairness while preserving the feature distribution, thereby enabling empirical studies on fairness–performance trade-offs in ML algorithms.

The proportion of flipped instances controls the degree of unfairness introduced,

allowing gradual and measurable manipulation unfairness. Rather than generating synthetic data from scratch, the method modifies the labels of existing datasets. It is designed for binary classification settings with a single binary sensitive attribute, consistent with prevailing assumptions in recent fairness in ML research.

Synthetic unfairness generated via label manipulation has been used to stress-test fairness-aware algorithms but not to test fairness-unaware algorithms (KAMIRAN e CALDERS, 2012; ZHANG *et al.*, 2022). The present method introduces measurable unfairness by flipping labels based on group membership and classification confidence, aligning with controlled experimental paradigms suggested by WICK *et al.* (2019) and FRIEDLER *et al.* (2018). The procedure explicitly targets protected-positive and privileged-negative instances, simulating systematic conditional unfairness consistent with group fairness frameworks.

The proposed method begins with a binary classification dataset D, consisting of feature vectors X, observed labels $\tilde{Y}$, and a binary sensitive attribute A. The restriction to a single sensitive attribute follows prevalent assumptions in fairness in ML literature (HARDT *et al.*, 2016). After standard preprocessing, e.g., encoding categorical features, removing invalid or missing values, a ML model, denoted the probabilistic estimator model h_e, is trained on D.

Formally, the estimator model h_e is defined as a function that maps a feature vector X to the estimated probability that an instance belongs to the positive class, expressed as $\mathbb{P}(\tilde{Y} = 1 \mid X)$. Accordingly, the probability of belonging to the negative class is $1 - \mathbb{P}(\tilde{Y} = 1 \mid X)$. Any probabilistic classifier, such as a Random Forest, can be used as h_e. Once trained on the dataset D, the estimator produces, for each instance, a probability score reflecting the model's confidence that the corresponding label is positive. These probability scores serve as the foundation for the subsequent label-flipping procedure, determining which instances will be modified according to the defined fairness rules.

Not all instances are eligible for flipping. To ensure structured unfairness rather than random noise, only positive labels from the protected group, $\tilde{Y} = 1$ and $A = 0$, and negative labels from the privileged group, $\tilde{Y} = 0$ and $A = 1$, are considered for flipping. This leads to an increase in positive outcomes for the privileged group and an increase in negative outcomes for the protected group, directly affecting Stat. Parity and indirectly influencing other fairness metrics.

Consider a simplified loan approval dataset where the sensitive attribute A represents gender ($A = 1$ for male and $A = 0$ for female), and the observed label $\tilde{Y} = 1$ indicates loan approval. Suppose a female applicant ($A = 0$) who was originally approved ($\tilde{Y} = 1$) is selected for flipping under the protected-positive rule, so her label becomes $\tilde{Y}^* = 0$. Similarly, a male applicant ($A = 1$) who was originally denied ($\tilde{Y} = 0$) has their label flipped to $\tilde{Y}^* = 1$. These controlled changes increase

the approval rate among males and decrease it among females, directly widening the statistical parity gap and introducing a measurable degree of unfairness in the modified dataset D^*.

It is important to note that the unfairness introduced through this method reflects structured, conditional disparities rather than random noise or artifacts of class imbalance. It increases the Stat. Parity gap $\mathbb{P}(\tilde{Y} = 1 \mid A = 1) - \mathbb{P}(\tilde{Y} = 1 \mid A = 0)$, thereby directly increasing this fairness metric and indirectly affecting others (HARDT *et al.*, 2016). This design ensures that any observed deterioration in fairness metrics can be directly attributed to the injected bias, allowing a clearer analysis of model behavior under varying unfairness conditions.

Although the label-flipping method applies symmetric criteria for both positive and negative labels, it does not guarantee an equal number of flips in each direction. This asymmetry arises due to potential imbalances in the class distribution across groups in the dataset. Consequently, the procedure may lead to changes in the overall proportion of positive and negative labels, thereby altering the class balance. However, such shifts remain interpretable and are bounded because the number of labels flipped is controlled.

The sensitive attribute is not explicitly treated differently during model training; it is handled in the same way as all other input features. Therefore, any disparities in predicted outcomes between the original dataset D and the modified dataset D^* arise indirectly, as a consequence of interactions between group membership and the manipulated labels, rather than from any direct use of the sensitive attribute itself. This situation mirrors cases of indirect discrimination, in which sensitive characteristics affect outcomes through correlations with other variables rather than through explicit inclusion in the decision-making process (KUSNER *et al.*, 2018). To ensure analytical validity, the modified dataset D^* preserves its integrity through three key criteria:

Feature Distribution Unchanged. All features X remain intact, so the marginal distribution $\mathbb{P}(X)$ is preserved. This removes covariate shift and ensures that observed effects arise from label manipulation alone.

Controlled Perturbation Mass. Only a bounded proportion of labels $\tilde{Y}$ are flipped, ensuring the signal-to-noise ratio remains at acceptable levels, maintaining model predictive capability (NORTHCUTT *et al.*, 2021). This controlled perturbation avoids severe deterioration in model performance, attributing observed changes specifically to engineered label biases rather than random corruption.

Class–Balance Sensitivity. The method targets specific labels, namely, protected-positive and privileged-negative groups, partially counterbalancing

each other depending on the subgroup sizes. Although global class proportions (ELKAN, 2001) are not strictly preserved, the procedure typically avoids drastic imbalances, since only a proportion of the labels are flipped, allowing fairness evaluations to focus on structural biases rather than general class imbalance.

As a result, models trained on the modified dataset D^* exhibit increased unfairness but typically retain good predictive performance. Predictive accuracy typically declines only modestly, as the proportion of flipped labels is controlled and the feature distribution remains unchanged. Since the manipulated labels still follow consistent patterns aligned with the original data structure and the feature distribution is unaltered, classification patterns remain interpretable, supporting realistic and controlled fairness–performance analysis and allows for realistic assessments of fairness–performance trade-offs, as explored in empirical studies (KAMIRAN e CALDERS, 2012).

To systematically study the impact of induced unfairness on model behavior, the proposed method applies three distinct label-flipping strategies: high-confidence flips, low-confidence flips and random flips. The first two are based on the probabilistic outputs provided by h_e, that is, the confidence levels associated with the predictions. Therefore, labels are flipped considering instances for which the estimator exhibits either high or low confidence in its predictions. The third doesn't use the results from h_e and instances selected randomly. Each strategy deliberately manipulates the dataset to reflect varying degrees of structured bias, allowing comprehensive evaluation of how predictive certainty affects both fairness outcomes and model performance. The detailed definitions and motivations for these strategies are presented next:

Low Confidence Flips (LOW). This strategy targets instances for which $\mathbb{P}(\tilde{Y} = 1 \mid X)$ is close to 50%, indicating that h_e is highly uncertain and the likelihood of belonging to the positive or negative class is nearly the same. Such cases often occur near the classification boundary or when the feature values do not provide strong evidence for either class. The rationale for including this strategy is to introduce bias in the most ambiguous regions of the feature space, where label changes are less likely to contradict clear feature–label associations. As a result, these flips are expected to have only a minor impact on overall predictive performance while still producing measurable increases in group unfairness, particularly in metrics such as Stat. Parity. This setting helps isolate the effect of unfairness from the effect on accuracy, offering insight into how disparities between groups can grow even when the dataset's predictive characteristics are largely preserved.

High Confidence Flips (HIGH). This strategy selects instances where $\mathbb{P}(\tilde{Y} = 1 \mid X)$ is close to 100% for positive labels (or close to 0% for negative labels), meaning that h_e assigns them a very high degree of certainty. These instances are typically those for which the features are strongly aligned with the assigned label according to the learned decision boundary of h_e. Flipping such labels introduces deliberate contradictions between the feature patterns and the new, manipulated labels. The motivation for including this strategy is to assess the effects of injecting bias into the most predictable and stable parts of the dataset, thereby creating highly structured unfairness that is also the most noticeable to the model. This is expected to strongly degrade both fairness metrics and predictive accuracy, since the classifier will be forced to learn from labels that directly conflict with the most reliable feature–label relationships in the data. The resulting models are anticipated to show significant shifts in confusion matrix rates for both protected and privileged groups, reflecting severe distortion of the learned decision boundary.

Random Within Sets Flips (RANDOM). In this strategy, instances are selected at random within the protected-positive and privileged-negative sets, without considering the confidence scores of h_e. Although the selection inside each set is random, the restriction to these two groups means that the procedure still injects structured unfairness, rather than pure, dataset-wide label noise. The motivation for including this strategy is to provide a baseline for comparison with the confidence-based approaches, allowing the evaluation of how much predictive performance and fairness metrics are affected when the same type of group-targeted unfairness is introduced without prioritizing high or low confidence instances. Because the selection does not exploit the probability outputs of h_e, the effects on fairness are expected to be less systematic than in the high-confidence case, while accuracy degradation will depend mainly on the proportion of flipped labels and the representativeness of the randomly chosen instances.

These rules allow comprehensive exploration of label flipping strategies affects fairness and predictive accuracy. This controlled label-flipping framework enables systematic injection of unfairness into fair datasets, while preserving structural, statistical and predictive consistency. It enables the investigation of classifier behavior under varying Pollution Rate (ρ), the rate at which unfairness is added, benchmarking ML algorithms, and analyzing fairness–performance trade-offs under controlled, reproducible experimental setting.

Assume the dataset contains $n = 10{,}000$ instances, of which 1,200 belong to the protected-positive group and 800 to the privileged-negative group, so there is in total

2,000 instances eligible for flipping. With a pollution rate $\rho = 0.10$, the algorithm flips $m_p = \lceil 0.10 \times 2{,}000 \rceil = 200$ instances, positive to negative and negative to positive. The resulting dataset D^* exhibits a systematic increase in the approval rate for the privileged group and a corresponding decrease for the protected group, while maintaining overall data coherence and interpretability. The procedure is formalized in Algorithm 1.

Algorithm 1: Systematic Label Flipping for Fairness Stress Testing

Input: Dataset $D = \{(X_i, \tilde{Y}_i, A_i)\}_{i=1}^n$, flip rate $\rho \in [0, 1]$, flipping rule.
Output: Modified dataset D^* with engineered unfairness.
$h_e \leftarrow$ train estimator on D
for each instance $i = 1, \ldots, n$ **do**
  $\pi_i \leftarrow h_e(X_i)$
end for
$G_p \leftarrow \{i \mid \tilde{Y}_i = 1 \wedge A_i = 0\}$
$G_n \leftarrow \{i \mid \tilde{Y}_i = 0 \wedge A_i = 1\}$
Sort G_p and G_n by flipping rule (high, low, or random confidence)
$m_p \leftarrow \lceil \rho |G_p| \rceil$, $m_n \leftarrow \lceil \rho |G_n| \rceil$
$Y^* \leftarrow \tilde{Y}$
for each $i = 1, 2, \ldots, m_p$ **do** $Y^*_{G_p[i]} \leftarrow 0$ **end for**
for each $i = 1, 2, \ldots, m_n$ **do** $Y^*_{G_n[i]} \leftarrow 1$ **end for**
$D^* \leftarrow \{(X_i, Y_i^*, A_i) \mid i = 1, \ldots, n\}$
return D^*

The inputs of the algorithm are the dataset D, Pollution Rate ρ, or the proportion of instance labels to be flipped, and the chosen High, Low or Random flipping rule. It begins by training the model h_e, with the features X and the observed labels $\tilde{Y}$, to calculate the probabilities π_i of each instance i belonging to the positive class. Subsequently, protected-positive G_p and privileged-negative instances G_n are identified. Depending on the selected flipping rule, these instances are sorted accordingly. Then, the predetermined proportions ρ of these sorted instances, m_p and m_n, are flipped: protected positives become negatives, and privileged negatives become positives. Finally, the label modified dataset D^* is returned.

The modified dataset D^* thus maintains analytical integrity by adhering to three preservation criteria: unchanged feature distributions, controlled perturbation mass, and sensitivity to class balance. Classifiers trained on D^* can clearly illustrate fairness–performance trade-offs, as structural bias can be precisely controlled, quantified, and interpreted, allowing researchers to investigate how classifiers respond to increasing bias levels, supporting reproducible, controlled fairness evaluations in line with prior experimental best practices (KAMIRAN e CALDERS, 2012; WICK *et al.*, 2019; FRIEDLER *et al.*, 2018).

Chapter 4

Experiments

This chapter presents the experimental methodology and the results obtained using the proposed stress testing approach described previously. The first section introduces the design of the experiments, describing the datasets, the application of the method proposed in Chapter 3.4, the training and validation procedure with cross-validation and hyperparameter optimization, the classifiers under analysis, and the evaluation criteria. The second section reports the empirical findings, beginning with a comparison of the three bias injection strategies to identify the most appropriate for analysis, and then presenting a systematic comparison of the classifiers under the recommended strategy, examining the evolution of performance and fairness across models.

4.1 Experimental Methodology

This section details the complete experimental methodology designed to systematically investigate how established classification algorithms respond to progressively unfair training data. The aim is to examine the effects of deliberate and incremental unfairness, introduced through the controlled label-flipping technique detailed in section 3.4, on both predictive performance and fairness metrics defined in sections 2.3 and 2.4. Figure 4.1 shows the complete steps of each executed experiment.

Three publicly available datasets that present a classification task were carefully selected based on their frequent use and relevance in fairness research. Adult (BARRY BECKER, 1996) predicts income above a 50000 dollar per year, Bank Marketing (S. MORO, 2014) predicts subscription to a financial product, and COMPAS (BARENSTEIN, 2019) predicts criminal recidivism. Each dataset features a binary outcome label and a chosen binary sensitive attribute. In this work, for Adult the sensitive attribute considered is gender, where female is the protected group and male the privileged; for Bank Marketing, marital status, where being single or divorced represents the protected group and married the privileged group; and

for COMPAS, the sensitive attribute is racial background, where being African-American is the protected group and having another race is the privileged group. Table 4.1 summarizes the characteristics of each dataset.

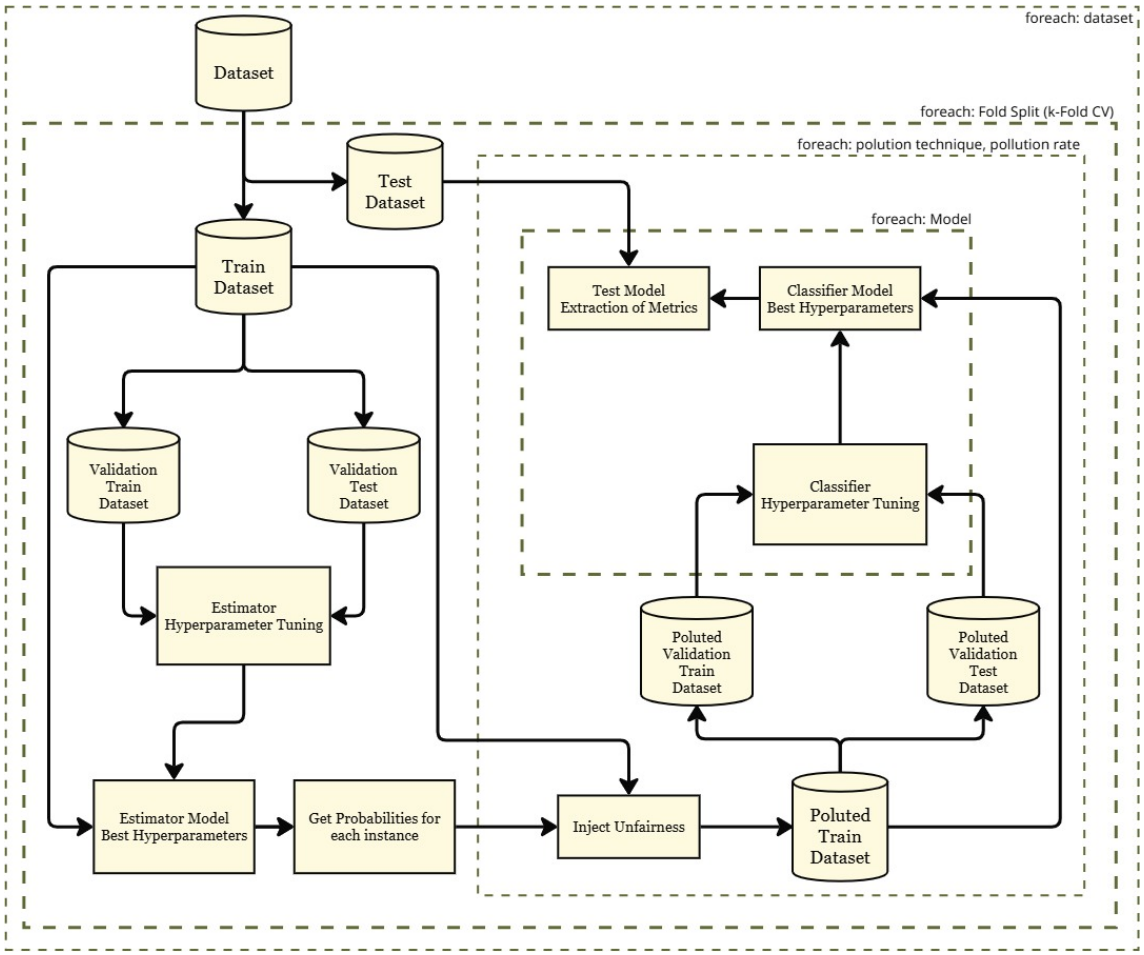
Table 4.1: Details of the datasets used in this work.

Dataset	**Adult**	**Bank**	**COMPAS**
# Features	98	70	10
# Instances	48842	45211	6172
Sensitive Attribute	gender	marital	race
Positives (%)	23.93	11.70	54.49
Negatives (%)	76.07	88.30	45.51
Privileged (%)	66.85	60.19	48.56
Unprivileged (%)	33.15	39.81	51.44
Pos. Privileged (%)	20.31	6.09	29.96
Pos. Protected (%)	3.62	5.60	24.53
Neg. Privileged (%)	46.54	54.10	18.60
Neg. Protected (%)	29.53	34.20	26.91
Statistical Parity	0.195	-0.04	0.140

Before splitting and training, each dataset is preprocessed to ensure compatibility with the ML models, following a shared rationale. Categorical variables with two categories are label-encoded directly into binary format, with one being one category and zero being the other category. Variables with more than two categories are converted to one-hot encoded vectors to prevent implying an inexistent ordinal relationship. Continuous numerical variables only appear in Adult and Bank Marketing datasets with no negative number, therefore they are scaled linearly to the interval [0,1] to promote training convergence and to have less sensitivity to the scales of the numbers. In the COMPAS dataset, all categorical features are represented through one-hot encoding without preserving any ordinal relationships. The only feature kept in its original numerical form is the number of prior convictions, which retains its quantitative meaning. These preprocessing steps are designed to maintain the marginal distributions of features, thereby ensuring that any observed shifts in fairness or performance metrics result solely from deliberate label manipulation rather than unintended data preprocessing biases.

Following preprocessing, each dataset undergoes an identical splits. Initially, the dataset is partitioned using a stratified 5-fold cross-validation approach. In each fold, 80% of the data is allocated for training, while the remaining 20% forms a fixed test set. This stratified split ensures that all models, trained on different folds, are consistently evaluated on test partitions with stable class distributions and proportions of sensitive groups. This process is repeated five times, each time selecting a different fold to serve as the test set, while the remaining four folds constitute the training set.

Figure 4.1: Diagram of Experimental Methodology Framework for One Complete Experiment



Within each of these training sets, an internal split is performed to promote hyperparameter tuning. This internal split divides the available training data into an 80% training subset and a 20% validation subset. Importantly, this internal division is conducted only once and does not involve additional folds. After identifying the optimal hyperparameters through validation, the final model is retrained using the entire training set. This procedure applies equally to both the estimator model and the classifier models, differing only in their roles within the overall experimental framework. Although the datasets vary in their characteristics, the described cross-validation protocol, test-fold designation, and validation strategy remain strictly uniform across all experiments.

The deliberate introduction of unfairness into the labels of the training subsets is performed using the controlled label-flipping methodology described in Chapter 3. The process begins by training an estimator model, a Random Forest (BREIMAN, 2001) in our experiments, using the original, unmodified training data. This model is then used to assign confidence scores to all instances that are either protected-

positive or privileged-negative. These scores determine which instances are selected for label flipping, according to one of two score-based strategies: LOW, which flips the least confident instances, and HIGH, which flips the most confident ones. A third strategy, RANDOM, ignores the confidence scores and selects instances uniformly at random.

Each strategy is applied using four distinct flipping pollution rate ρ, corresponding to 5%, 10%, 15%, and 20% of the eligible instances. This results in 12 different biased versions of the training data, four for each strategy, plus one original version without any label modification. All 13 datasets are treated equally in the experimental design and are referred to as the Polluted Train Datasets, with the original dataset representing the case of zero pollution rate, or $\rho = 0$. Each of the classification models are trained using all 13 datasets. Consequently, for every fold splitting of the original dataset and each classifier model, 13 distinct classification models are produced.

Four widely adopted classification algorithms were selected to evaluate how predictive models respond to increasing levels of unfairness in training data. The chosen classifiers are Decision Tree (QUINLAN, 1986), Logistic Regression (HOSMER *et al.*, 2013), Random Forest, and Neural Network implemented as a feedforward fully connected architecture (GOODFELLOW *et al.*, 2016). This selection reflects a deliberate attempt to include models from distinct families of learning paradigms: tree-based, linear, ensemble, and neural network, respectively. These models are well-known in the literature for their computational efficiency and wide deployment in practical applications. The Random Forest model, in particular, plays a dual role in this study: it is used both as a classification model and as the estimator model responsible for generating confidence scores in the label-flipping process described earlier. These two uses are independent and generate different models. The choice of Random Forest for the estimation step is motivated by its robustness, low variance, and ability to generate calibrated class probabilities without requiring extensive tuning, which makes it suitable for estimating reliable confidence values used to guide the flipping mechanism.

To ensure a impartial and consistent comparison between classifiers, all hyperparameters are optimized using the Optuna framework with the Tree-of-Parzen-Estimator sampler (TPE) (BERGSTRA *et al.*, 2011). TPE is a bayesian optimization method known for its strong performance in high-dimensional search spaces and its ability to model complex, non-linear relationships between parameters and objective values. We opted for a uniform tuning strategy across all models to avoid introducing bias through unequal optimization effort. The optimization process uses the median pruning strategy to terminate under performing trials early, thereby increasing overall efficiency. Each optimization run consists of 50 trials, where can-

didate configurations are evaluated on the validation sets using the MCC as the optimization objective. No fairness metric is used during this step; the goal is to reproduce how standard ML models would behave under realistic development settings where fairness considerations are often absent. This approach allows us to isolate and analyze the fairness impact of biased data without interference from fairness-aware interventions.

The search space for each model comprises a wide range of hyperparameters defined based on established practices in the literature. While several parameters are explored during the optimization process, Table 4.2 highlights only the most influential and representative ones for each algorithm. For example, the Decision Tree model varies in maximum depth and minimum number of samples required at leaf nodes; the Logistic Regression model explores different regularization strengths; the Random Forest model adjusts the number of trees, tree depth, and feature sub-sampling rate; and the Neural Network model varies learning rate, number of hidden units, and activation functions. After optimization, the best configuration is retrained on the full training set and subsequently evaluated on the held-out test set, following the experimental protocol described earlier.

Table 4.2: Hyperparameter search ranges used in Optuna optimization for each classification algorithm.

Model	**Hyperparameter Ranges**
Decision Tree	`max_depth` $\in$ [2, 20]; `min_samples_leaf` $\in$ [1, 10]
Logistic Regression	regularization strength $C \in [10^{-4}, 10^{4}]$ (log-uniform)
Random Forest	`n_estimators` $\in$ [50, 300]; `max_depth` $\in$ [2, 20]; `max_features` $\in$ [0.1, 1.0]
Neural Network	hidden layer sizes $\in$ [20, 200]; learning rate $\in [10^{-4}, 10^{-1}]$; activation function in `{relu, tanh}`

Model evaluation employs a comprehensive set of metrics designed to assess both predictive performance and group fairness in a detailed and complementary way. The primary performance metric adopted is the MCC, chosen for its robustness in imbalanced binary classification problems and its ability to capture the overall quality of predictions across all quadrants of the confusion matrix. Supplementary performance indicators include Acc. and F1, both widely used in classification benchmarks. For fairness assessment, the main metric is the Eq. Odds, which compares the true positive and false positive rates across sensitive groups. In addition, Stat. Parity and Eq. Opp. are computed to provide complementary fairness perspectives. To allow fine-grained interpretation of model behavior, confusion matrix rates are

computed globally and separately for the privileged and protected groups. These include True Positive Rate (TPR), True Negative Rate (TNR), False Positive Rate (FPR) and False Negative Rate (FNR), all of which are calculated based solely on the untouched test sets and the predictions of the models.

To ensure statistical robustness and account for variability, the entire experimental framework described in Figure 4.1 is executed independently eight times. In each of these eight full experiments, the three datasets, Adult, Bank, and COMPAS, are re-split into five new stratified folds. Each fold splitting generates 13 training datasets: one baseline with no label manipulation and 12 with unfairness injected through the Flip Labels method using combinations of four pollution rates ρ and three flipping strategies. Every one of these 13 training sets is used to train four different classification models, resulting in five folds $\times$ 13 unfair datasets $\times$ four models = 260 trained classifiers per dataset per experiment. Repeating this process across the 3 datasets yields 780 models per experiment. Over the course of 8 complete experimental runs, a total of 6240 models are evaluated.

In all cases, the estimator model used for computing label-flipping confidence scores is fixed per dataset fold splitting within each experimental repetition. That is, a new estimator is trained for each fold splitting of each dataset in every experiment, but remains shared across all flipping strategies and pollution rate applied to that specific fold. For each combination of dataset, flipping strategy, threshold, and classifier, the evaluation metrics are averaged across all 5 folds and the 8 experimental runs, and the corresponding standard deviations are computed, resulting in 156 results in total for each metric, being 12 baseline results for each unmodified dataset and classifier, all the other 144 results use a flipping strategy and a threshold. This aggregation procedure ensures that the final results reflect consistent trends and are not biased by any particular data split or random sampling variation.

All computational experiments were executed on a Hewlett-Packard Victus 15-inch notebook equipped with an AMD Ryzen 7 5800H processor and 16 GB of DDR4 RAM. The software environment was configured on Windows 11, using Python version 3.13.3, with scikit-learn version 1.7 for model implementation and Optuna version 4.0.0 for hyperparameter optimization. This experimental setup was carefully structured to isolate the effects of label bias by keeping all other variables constant throughout the framework. As a result, any observed changes in predictive performance or fairness metrics can be confidently attributed to the deliberate and progressive manipulation of the training labels. The following section presents the results obtained under these controlled conditions and offers an in-depth analysis of how each classification algorithm behaves as the level of data unfairness increases.

4.2 Results and Discussion

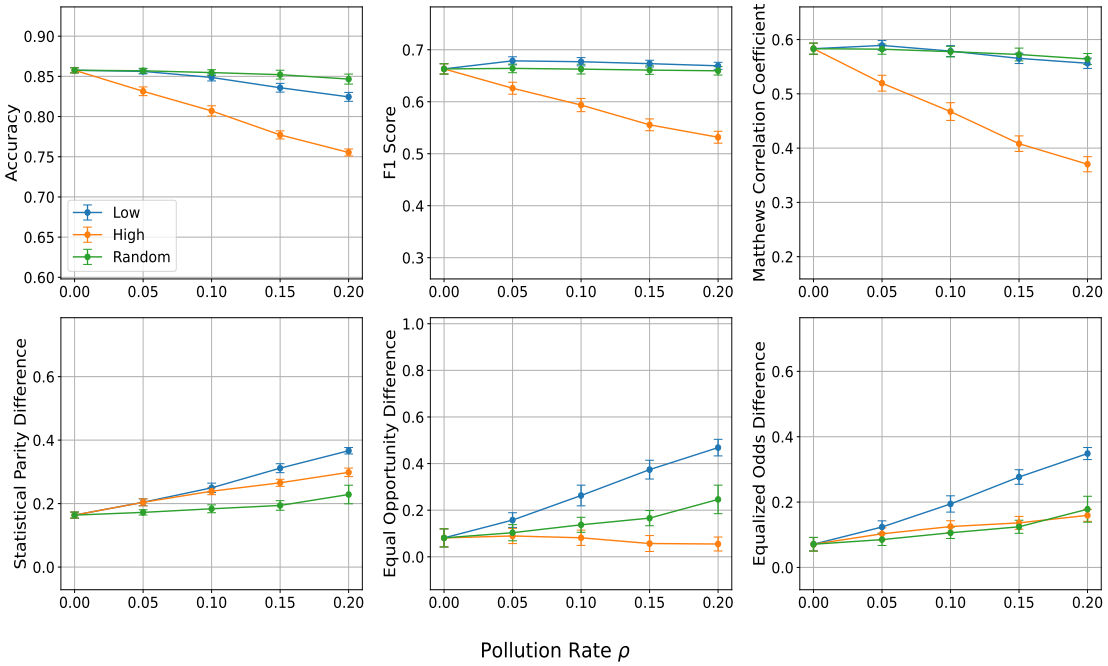
This section presents and discusses the experimental results obtained through the systematic label-flipping methodology. The analysis is organized into two complementary perspectives. First, the focus is on the flipping strategies themselves, examining how the three procedures, LOW, HIGH, and RANDOM, differentially affect predictive performance and fairness, thereby identifying the most suitable approach for controlled stress testing. Second, the focus shifts to the classifiers, comparing how Decision Tree, Logistic Regression, Random Forest, and Neural Network models respond under increasing levels of unfairness in order to assess their relative robustness. Together, these analyses provide a comprehensive view of both the methodological choices involved in bias introduction and the intrinsic sensitivities of the classifiers, offering insights into the fairness–performance trade-offs that emerge when training data is systematically polluted.

4.2.1 Flipping Strategies

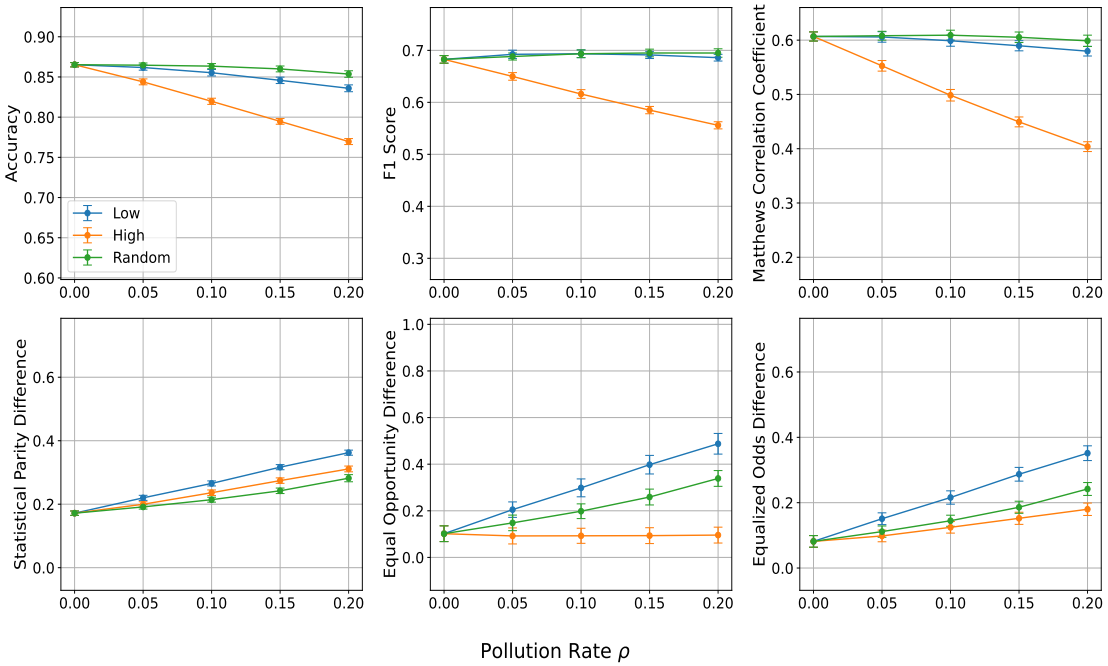
In the previous section, the experimental methodology was described in detail, including the datasets, classifiers, evaluation metrics, and the procedure used to introduce unfairness into the training data. Building on this foundation, the present subsection focuses on the analysis of the unfairness injection approaches. The objective is to systematically examine how LOW, HIGH, and RANDOM influence both performance and fairness metrics. This analysis is essential to determine which approach provides the most consistent basis for evaluating, in a controlled manner, the sensitivity of classification algorithms to progressive levels of unfairness in the training data.

Figures 4.2, 4.3, and 4.4 summarize the main results, presenting predictive performance and fairness metrics under the three injection methods, LOW, HIGH, and RANDOM. All curves are plotted against increasing values of ρ, with higher values of Acc., F1, and MCC indicating better predictive performance, and higher values of Stat. Parity, Eq. Opp., and Eq. Odds reflecting greater unfairness. This joint representation facilitates a direct visualization of trade-offs, showing how pollution simultaneously impacts performance and fairness, while also exposing specific exceptions and intersection points. Complementary results for TPR, TNR, FPR, and FNR are provided in Appendix A.

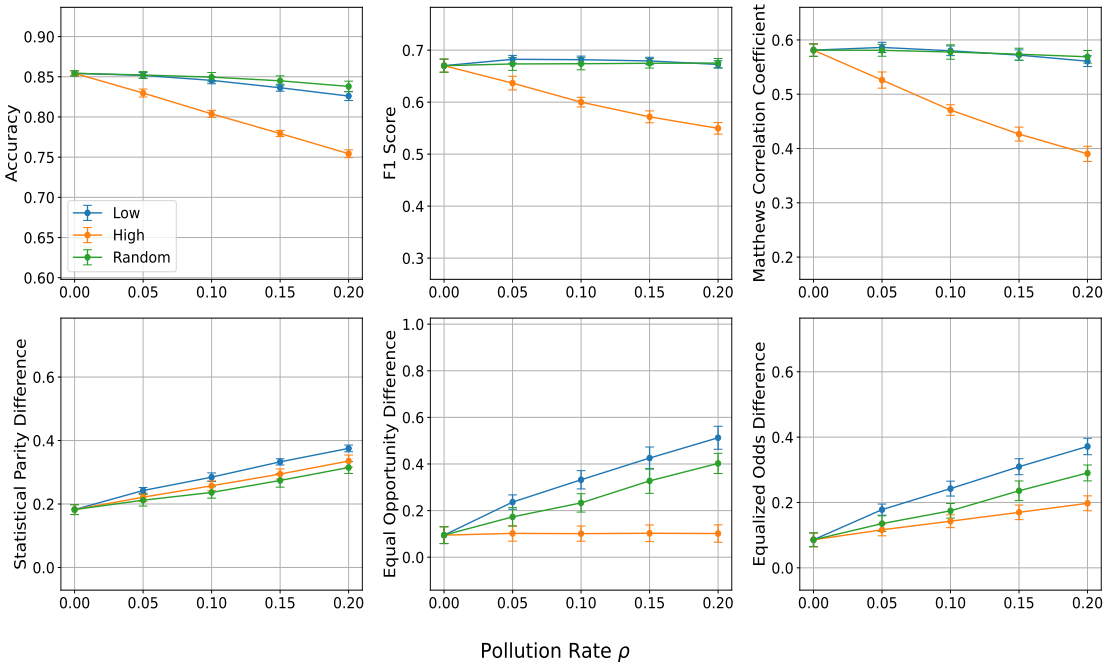
Figure 4.2: Performance and Fairness Metrics of the classifiers trained with the Adult dataset.



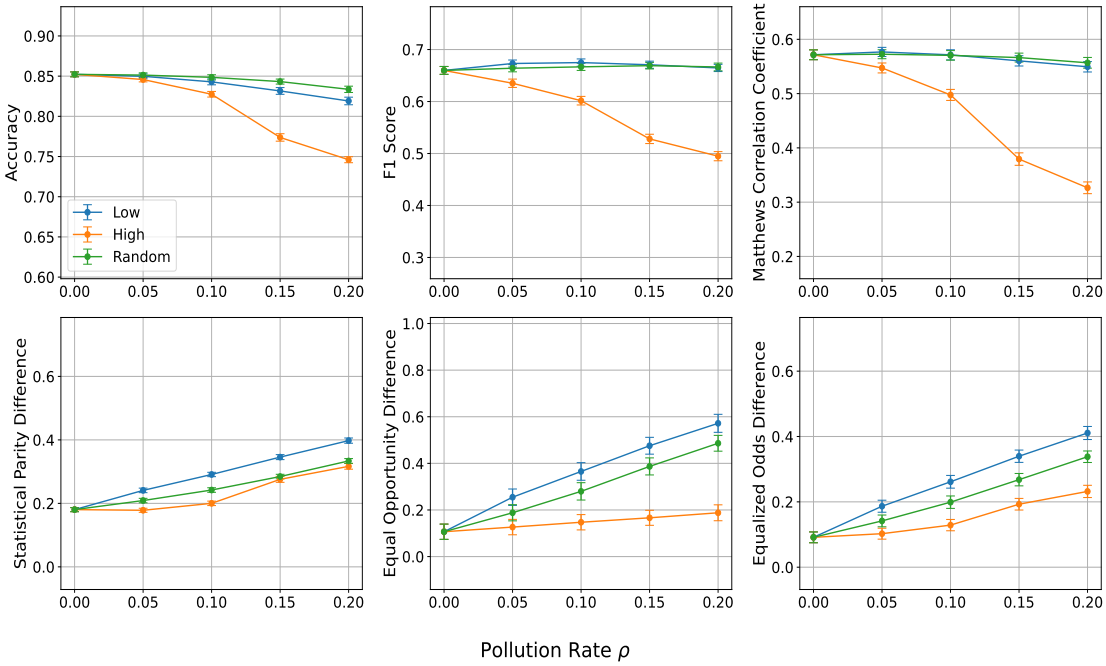
(a) Adult - Decision Tree



(b) Adult - Random Forest

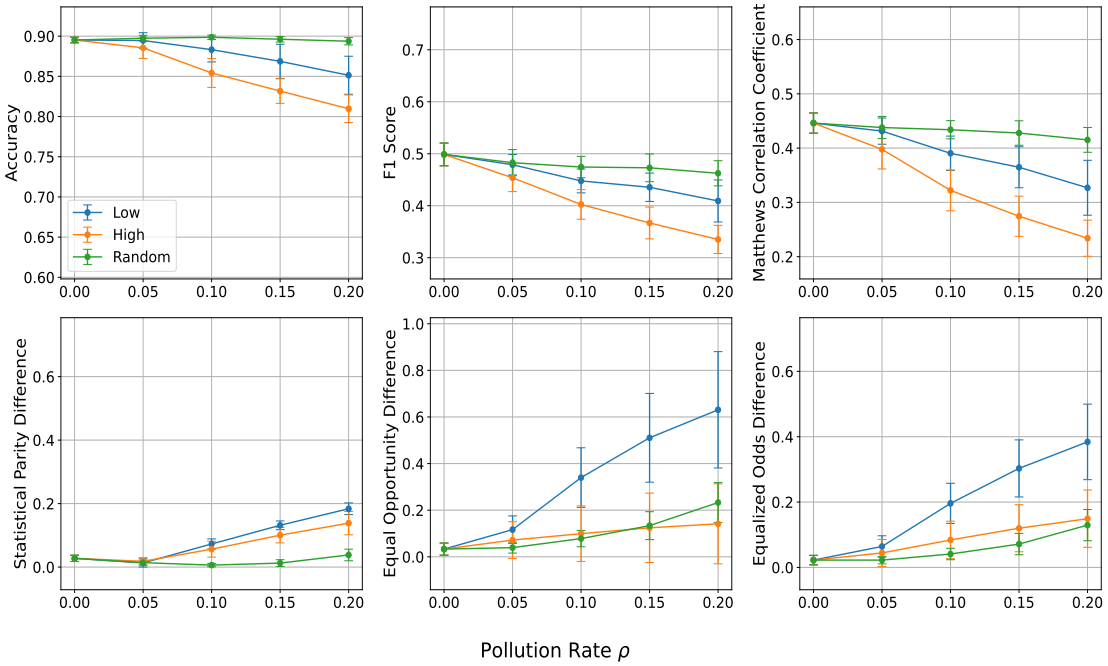


(c) Adult - Neural Network

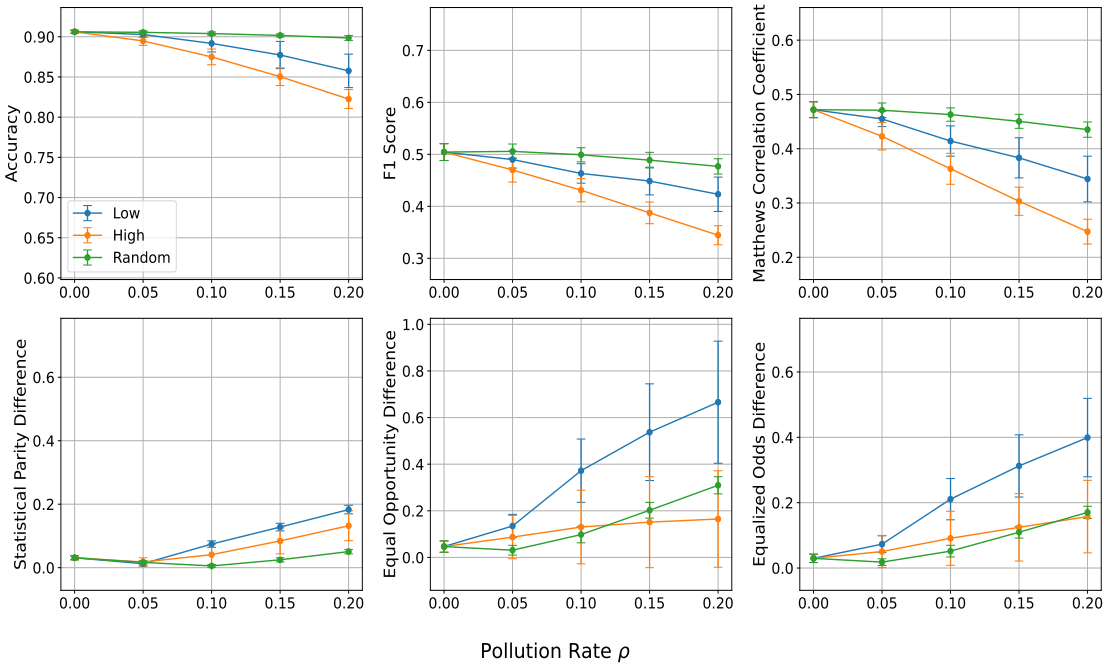


(d) Adult - Logistic Regression

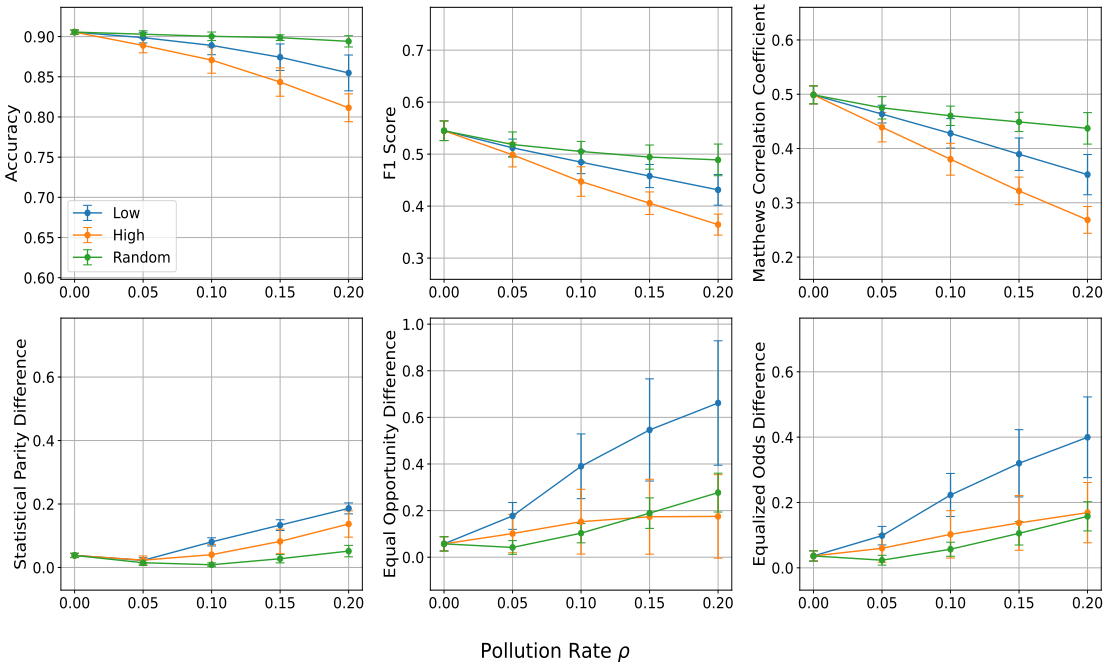
Figure 4.3: Performance and Fairness Metrics of the classifiers trained with the Bank dataset.



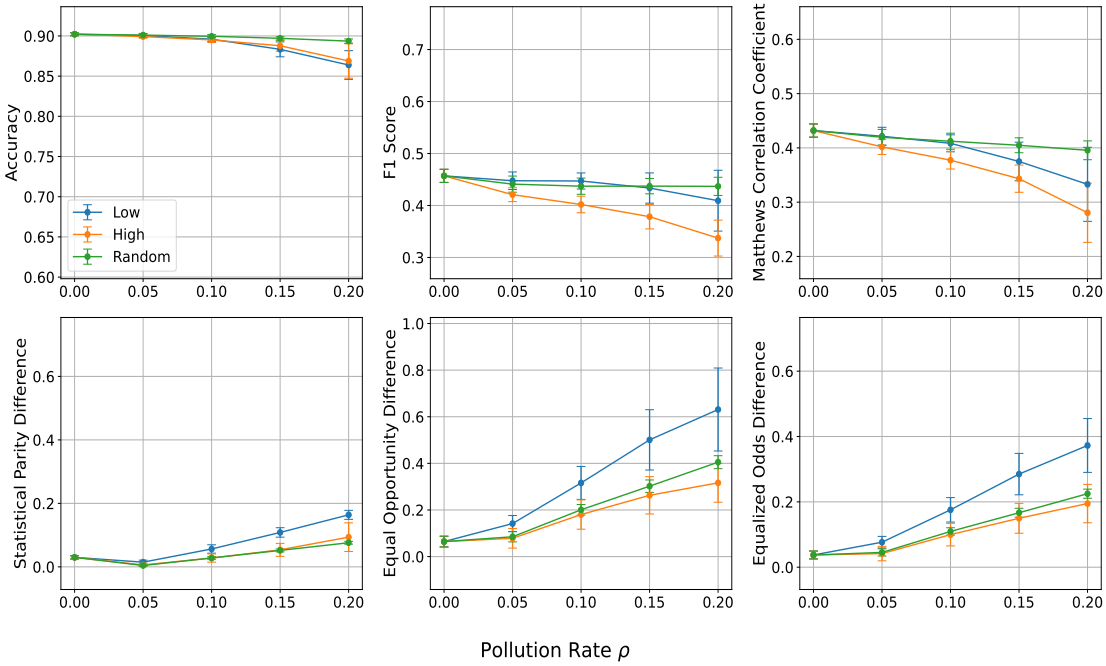
(a) Bank - Decision Tree



(b) Bank - Random Forest

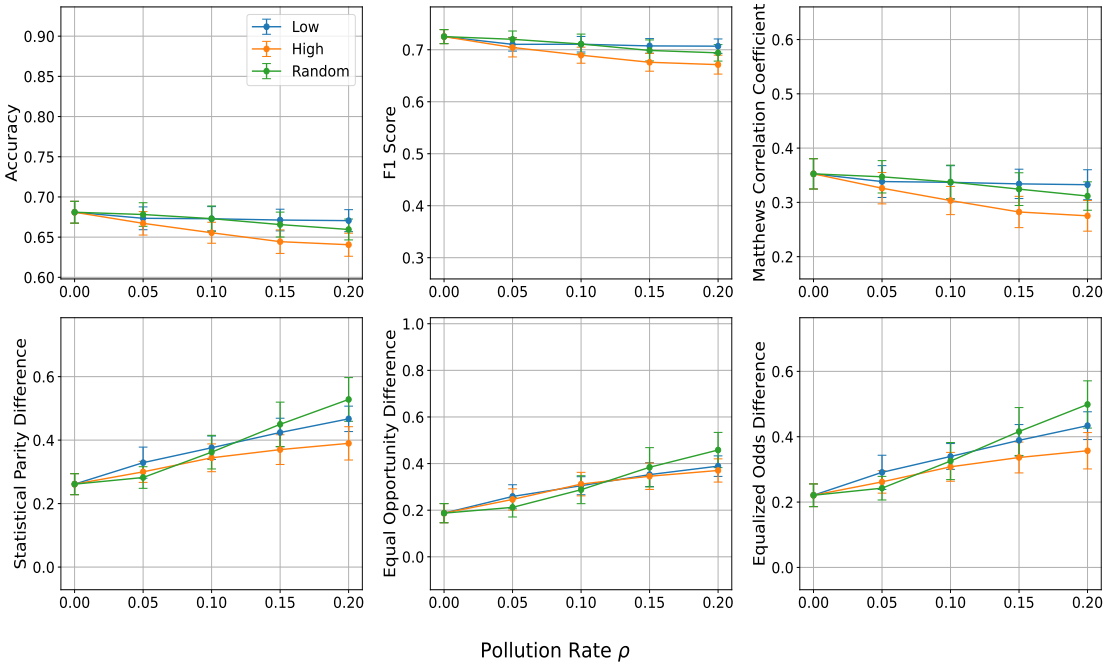


(c) Bank - Neural Network

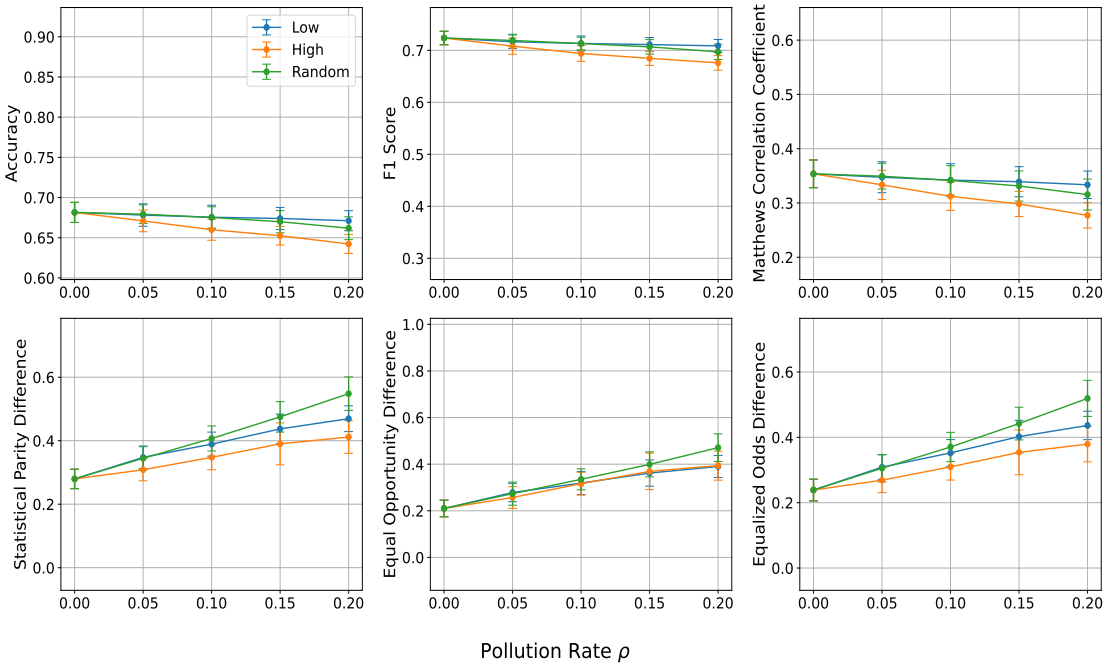


(d) Bank - Logistic Regression

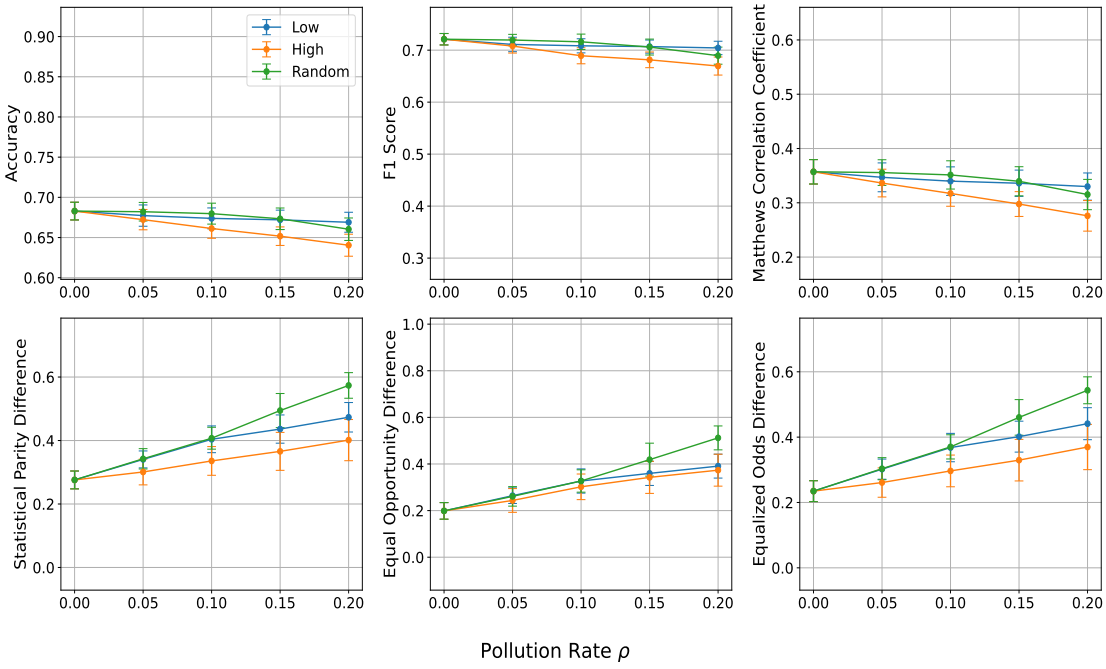
Figure 4.4: Performance and Fairness Metrics of the classifiers trained with the COMPAS dataset.



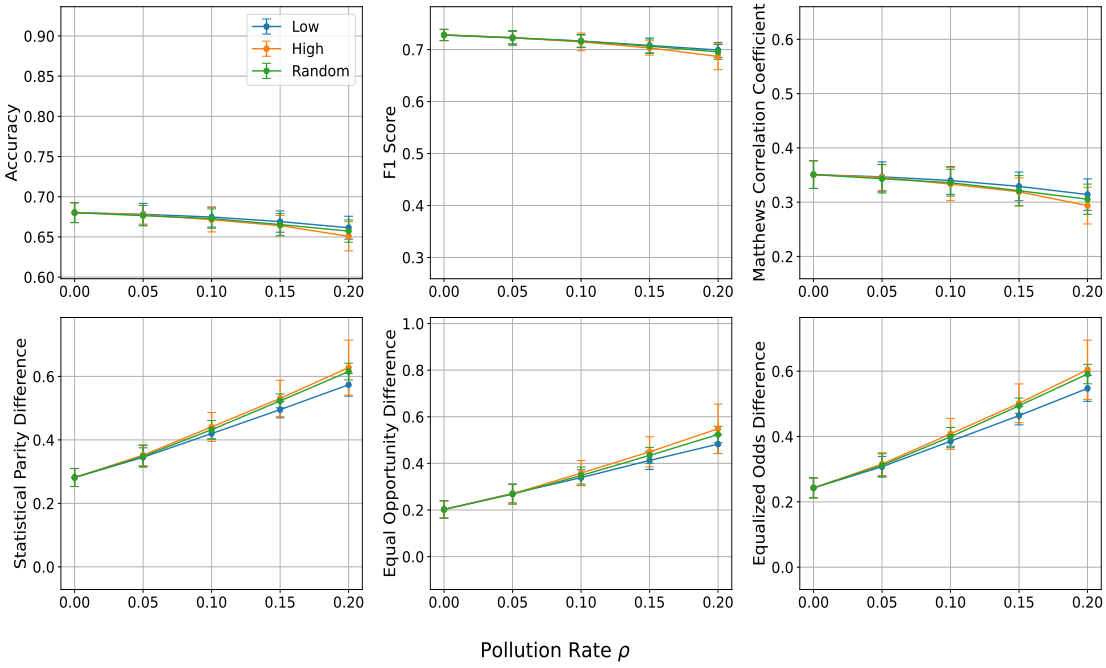
(a) COMPAS - Decision Tree



(b) COMPAS - Random Forest



(c) COMPAS - Neural Network



(d) COMPAS - Logistic Regression

The three unfairness injection approaches were designed with distinct expectations regarding their impact on predictive performance and fairness. In the LOW method, flipping labels of low-confidence instances was expected to produce a sharp increase in unfairness while preserving predictive performance, since the altered examples are those about which the classifier is already uncertain. In contrast, the HIGH method was anticipated to generate more moderate increases in unfairness but accompanied by substantial declines in performance, because highly confident predictions are directly contradicted by the injected flips, undermining core patterns of the data. The RANDOM method was expected to lie between these two extremes, producing intermediate effects both in terms of performance loss and unfairness growth, since label corruption is not guided by confidence but applied indiscriminately.

The experimental results corroborate these expectations. When analyzing the variation of predictive performance and fairness metrics with increasing values of ρ, the curves are predominantly monotonic, without sudden collapses or unexpected reversals. In most cases, decreases in predictive performance are accompanied by increases in unfairness, reflecting the intended trade-off induced by systematic label flipping. The LOW method consistently led to substantial rises in unfairness while maintaining relatively stable predictive performance, indicating that the main predictive patterns of the dataset were largely preserved. The HIGH method caused the steepest drops in performance, alongside smaller increments in unfairness, confirming the disruptive effect of flipping high-confidence instances. Finally, the RANDOM method showed intermediate behavior, with moderate changes in both dimensions. These general trends validate the logic underlying each approach and make clear that the LOW method is the most appropriate for fairness stress testing. By maintaining predictive performance while substantially amplifying unfairness, it preserves the essential predictive structure of the dataset, allowing the effects of unfairness injection to be studied in a controlled manner without confounding losses in accuracy.

One exception to the overall monotonic behavior occurs in the "S"-shaped curves observed for Acc., F1, and MCC when the Logistic Regression classifier is trained on the Adult dataset under the HIGH strategy, as shown in Figure 4.2 (d). In this case, there is a small decrease at $\rho = 0.05$, followed by a sharper drop at $= 0.10$, and then a partial recovery beginning at $= 0.15$. This behavior can be explained by the characteristics of Logistic Regression, which employs a linear decision boundary that is strongly influenced by high-confidence instances (AHFOCK e MCLACHLAN, 2021). When only a small fraction of such points are flipped, the model undergoes minor adjustments. As the number of corrupted high-confidence instances increases, however, the boundary shifts more abruptly, leading to a sharper decline in performance. At higher levels of pollution, the flipped labels become more evenly distributed across

both classes, enabling the model to partially recalibrate and regain stability, which accounts for the observed recovery in performance.

Another exception occurs in the Stat. Parity results for the Bank dataset, where an initial decrease is observed at $\rho = 0.05$, followed by a steady increase starting at $= 0.10$, as shown in Figure 4.3. This behavior is directly related to the original distribution of outcomes in the dataset, where the protected group initially contains a larger proportion of positive instances compared to the privileged group. When unfairness is first injected, label flipping removes some of these positive outcomes from the protected group while adding positive outcomes to the privileged group. This temporary rebalancing reduces the disparity between the marginal positive rates of the two groups, leading to the initial drop in Stat. Parity. As the pollution rate increases further, however, the flips accumulate in a way that favors the privileged group while reducing positives in the protected group, reintroducing and amplifying the disparity. Consequently, the metric resumes its monotonic rise.

The last exception involves F1, where a slight increase is observed under the LOW and RANDOM strategies in the Adult dataset. This occurs when marginal adjustments to the decision boundary increase Rec. at a negligible cost in Prec., thereby improving F1. In practice, flipping low-confidence or randomly chosen points may cause the classifier to slightly widen its decision boundary, capturing additional true positives without introducing a significant number of false positives. This effect is absent under the HIGH strategy because flipping high-confidence points directly damages the classifier's strongest predictions, causing losses in both Prec. and Rec., which prevents any temporary gain in F1. In the Bank dataset, this phenomenon is minimal or absent because its feature space is less sensitive to small shifts in the boundary, and the marginal cases flipped do not meaningfully improve the balance between Prec. and Rec.. In the COMPAS dataset, this effect does not occur in practice, since its structure provides very few marginal cases close to the decision boundary. The data distribution is more polarized and less influenced by small perturbations, which prevents marginal flips from generating improvements in Rec. without harming Prec., making any increase in F1 negligible.

Taken together, the predominantly monotonic trends and the few exceptions observed indicate that the essential behavior of the models under increasing pollution can be reliably captured by comparing the extreme values of ρ. The identified exceptions are local phenomena that do not alter the overall direction of the results and therefore do not compromise this mode of analysis. For this reason, examining the outcomes at $= 0.00$ and $= 0.20$ is sufficient to summarize the dominant tendencies, since these two points reflect the transition from the unbiased baseline to the highest level of injected unfairness (MENON *et al.*, 2015). Table 4.3 reports the cumulative differences between these two conditions, focusing on MCC and Eq. Odds,

the primary indicators of predictive performance and fairness in this study. The values highlighted in bold correspond to the lowest observed MCC and the highest observed Eq. Odds, underscoring the conditions where predictive performance is least compromised and unfairness is most severe. This approach provides a concise yet faithful representation of the results, avoiding overinterpretation of local fluctuations while maintaining consistency with the dominant trends (BAROCAS *et al.*, 2023).

Table 4.3: Cumulative Results for MCC and Eq. Odds (mean ± std) from $\rho = 0.00$ to $\rho = 0.20$. All bold values correspond to the smallest loss in MCC or the largest increase in Eq. Odds. The average is computed across all datasets.

Dataset	Strategy	MCC	Eq. Odds
Adult	LOW	$-0{,}024 \pm 0{,}010$	$\mathbf{0{,}288 \pm 0{,}020}$
	RANDOM	$\mathbf{-0{,}014 \pm 0{,}011}$	$0{,}180 \pm 0{,}024$
	HIGH	$-0{,}213 \pm 0{,}012$	$0{,}110 \pm 0{,}015$
Bank	LOW	$-0{,}123 \pm 0{,}054$	$\mathbf{0{,}357 \pm 0{,}113}$
	RANDOM	$\mathbf{-0{,}041 \pm 0{,}023}$	$0{,}139 \pm 0{,}038$
	HIGH	$-0{,}205 \pm 0{,}036$	$0{,}136 \pm 0{,}087$
COMPAS	LOW	$\mathbf{-0{,}026 \pm 0{,}020}$	$0{,}230 \pm 0{,}049$
	RANDOM	$-0{,}042 \pm 0{,}025$	$\mathbf{0{,}304 \pm 0{,}050}$
	HIGH	$-0{,}073 \pm 0{,}029$	$0{,}193 \pm 0{,}067$
Average	LOW	$-0{,}058 \pm 0{,}028$	$\mathbf{0{,}292 \pm 0{,}06}$
	RANDOM	$\mathbf{-0{,}032 \pm 0{,}02}$	$0{,}208 \pm 0{,}037$
	HIGH	$-0{,}164 \pm 0{,}026$	$0{,}146 \pm 0{,}056$

The general pattern remains consistent with the individual analyses presented before. Regarding performance, RANDOM systematically preserves accuracy, F1 and MCC more effectively than the other strategies, while LOW occupies an intermediate position, but not so distant from RANDOM, and HIGH produces the steepest degradation across all datasets. In terms of fairness, particularly for Eq. Opp. and Eq. Odds, LOW tends to yield the greatest increase, RANDOM follows closely, and HIGH consistently results in the smallest gains, most notably in the Adult and Bank datasets.

Although this trend is dominant, there are exceptions. In COMPAS with Logistic Regression, the HIGH strategy can surpass LOW and RANDOM in Stat. Parity, Eq. Opp., and Eq. Odds. This occurs because the COMPAS dataset is relatively small and strongly correlated with the sensitive attribute, so aggressive label flipping disrupts misleading correlations that a linear classifier would otherwise exploit, producing a counterintuitive improvement in fairness metrics. In Adult with Decision Tree under HIGH, Eq. Opp. may decrease slightly. This is explained by the high variance and instability of decision trees, which makes them highly sensitive to noise injection, causing fairness metrics to fluctuate even when larger gains are expected.

In Bank with Logistic Regression at $\rho = 0.20$, Acc. under LOW can fall below the value observed under HIGH. This reflects the fact that, at high levels of label flipping, the intermediate strategy may distort the linear decision boundary more than the extreme strategy, producing an anomalous but localized advantage in predictive performance. Finally, in COMPAS the Stat. Parity curves of LOW and RANDOM intersect at $\rho = 0.05$ and other points. This crossing is due to the interaction between random and confidence-based flips in a dataset with limited size and strong feature correlations; while RANDOM initially perturbs fewer critical examples, LOW regains the advantage as the pollution rate increases and more of the most influential borderline cases are systematically corrected. Overall, these exceptions highlight that the relative effectiveness of each strategy is not uniform but depends on the interplay between dataset properties, classifier characteristics, and the nature of the fairness metric. They do not undermine the general patterns observed, but rather illustrate that the stress testing method can expose subtle behaviors that are otherwise hidden when fairness interventions are applied in a homogeneous way.

Regarding the standard deviations, clear differences emerge across datasets. In Adult, deviations are consistently the smallest in both performance and fairness metrics, reflecting the stability provided by its large size and relatively balanced distribution. Effects remain highly consistent across folds and repetitions. In Bank, deviations are considerably larger due to its greater heterogeneity and class imbalance. This is particularly evident in Eq. Opp., where the average standard deviation in Bank is almost three times higher than in Adult, and in Eq. Odds, where variation is also markedly stronger, while Stat. Parity remains comparatively stable. For performance, Acc. displays low variance, whereas F1 and MCC fluctuate more; for instance, the standard deviation of F1 in Bank is more than twice that of Adult, and MCC also exhibits a deviation more than double. The COMPAS dataset, being much smaller and strongly correlated with the sensitive attribute, shows the opposite pattern: deviations in performance metrics are generally low, but fairness metrics are highly unstable, especially Stat. Parity, whose standard deviation is nearly four times higher than in Adult, reflecting the extreme sensitivity of marginal positive rates to small shifts in the decision boundary. When comparing metrics across datasets, MCC varies substantially not only in COMPAS but also in Bank, while F1 tends to fluctuate more strongly in Bank because of its dependence on precision and recall under class imbalance. Altogether, these patterns indicate that Adult is the most stable dataset, Bank is the most variable in fairness metrics, and COMPAS is the most unstable in terms of Stat. Parity.

Analyzing the results of TPR, TNR, FPR, and FNR available in Appendix A, for the full dataset and for the privileged and protected subsets, it can be observed that these metrics are generally more stable, with even fewer exceptions than the

performance and fairness metrics. This stability occurs because they are directly tied to the confusion matrix, and fairness metrics are derived from their differences across groups. In the full dataset, HIGH produces the strongest changes, with a marked increase in FPR and a reduction in TNR. For example, in Adult, TPR falls to around 0.59 and FNR rises to about 0.41 under HIGH, compared to TPR near 0.70 and FNR near 0.30 under LOW. As a result, HIGH leads to strong performance degradation and, on average, smaller increments in unfairness. LOW acts mainly on the decision boundary, shifting it toward more positive classifications. In Adult this produces higher recall, with TPR around 0.70 and FNR about 0.30.

In Bank, however, LOW creates a non-monotonic curve: TPR rises slightly to 0.42 at intermediate levels before declining again, while FNR falls to 0.58 and then reverses. This occurs because the Bank dataset is more heterogeneous and imbalanced, so flipping low-confidence labels can initially correct borderline cases before the accumulation of noise reverses the effect. In COMPAS, TPR decreases (to about 0.71) and FNR increases (to about 0.29) under HIGH, while LOW yields better recall ($TPR \approx 0.74$, $FNR \approx 0.26$). Since the effect is asymmetric across groups, Eq. Opp. and Eq. Odds grow more under LOW than under RANDOM or HIGH. Finally, RANDOM produces small, distributed shifts that keep TPR, FNR, TNR, and FPR more stable, leading to flatter curves. In Bank, for instance, RANDOM maintains the highest TNR at about 0.97, corresponding to the lowest FPR of only 0.03. In Adult, exceptions appear in TPR and FNR, where small changes occur even under distributed flips, showing that noise can still perturb the balance between recall and false negatives at certain thresholds.

In the privileged subset, label flips add positive instances, which under HIGH lead to increased FPR and reduced TNR. For example, in Adult privileged groups, FPR rises to about 0.21 and TNR falls to about 0.79 under HIGH. TPR and FNR for the privileged vary less, but still respond to the boundary shift. An exception arises in COMPAS with all classifiers except Logistic Regression, where the effect is mitigated by the dataset's small size and high correlation structure. In the protected subset, flips remove positive instances, so under HIGH the expected pattern emerges: TPR decreases and FNR increases, e.g., in COMPAS protected groups, whereas TNR and FPR remain relatively stable. Importantly, the greater stability of results in the full dataset compared to privileged and protected subsets is not exclusive to HIGH but occurs across all strategies. This is because, when aggregated, opposing shifts in the two groups partially cancel each other, reducing variability at the overall level.

In conclusion, the evidence consistently indicates that LOW is the most appropriate strategy for systematically injecting unfairness in classification tasks. It achieves the strongest and most consistent increases in group unfairness, while maintaining only moderate reductions in predictive performance. Its deterministic nature fur-

ther ensures reproducibility, which is essential for controlled experimentation and for building reliable baselines in fairness research. By contrast, RANDOM introduces instability: although it sometimes produces smaller losses in accuracy, it is non-deterministic and can also yield worse outcomes in performance without systematically maximizing unfairness. HIGH severely degrades predictive accuracy and, on average, produces smaller improvements in fairness, which makes it ill-suited for the analysis of practical trade-offs. The few exceptions observed are explainable and statistically limited, and therefore do not challenge this overall conclusion. For these reasons, LOW should be regarded as the recommended baseline method for future studies aiming to analyze, compare, or benchmark classifier behavior under progressively increasing levels of unfairness.

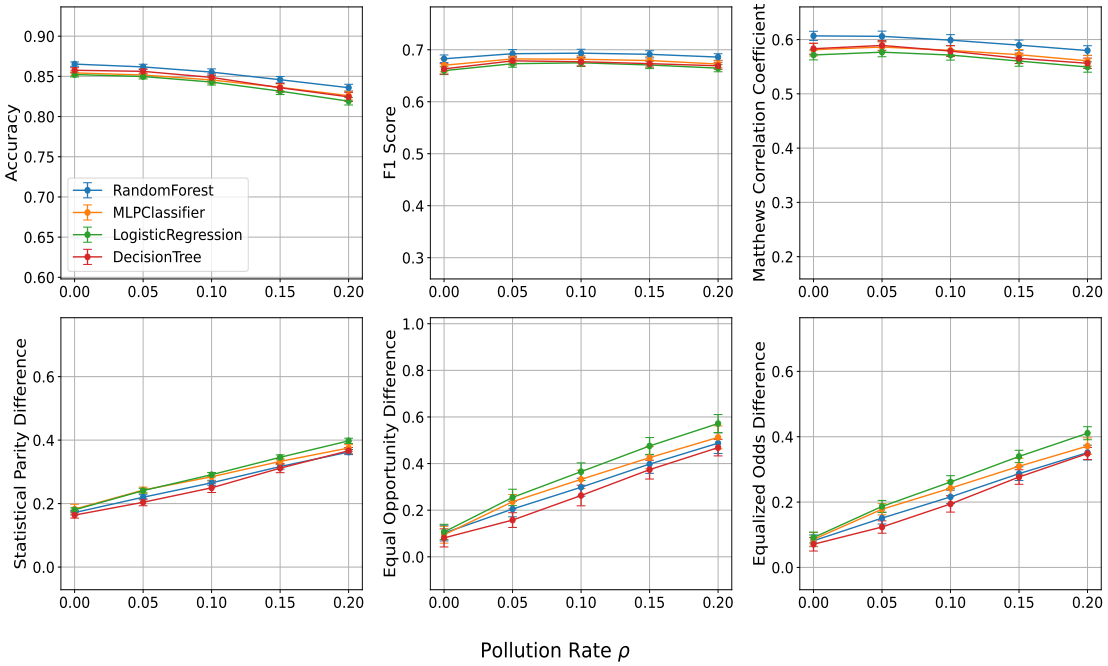
4.2.2 Classifiers

Having established in the previous section that the LOW strategy provided the most consistent and interpretable results among the bias introduction methods, the following analysis focuses exclusively on this approach. By fixing the strategy, it becomes possible to investigate in greater depth how different classifiers react to increasing levels of induced unfairness. This section therefore compares the behavior of the four classifiers under study, namely Random Forest, Neural Networks, Logistic Regression, and Decision Tree, in order to identify how their performance and fairness metrics evolve as the proportion of label flipping grows. In this way, the analysis shifts from contrasting bias introduction strategies to examining the relative robustness of the classifiers themselves in the face of systematically increasing unfairness.

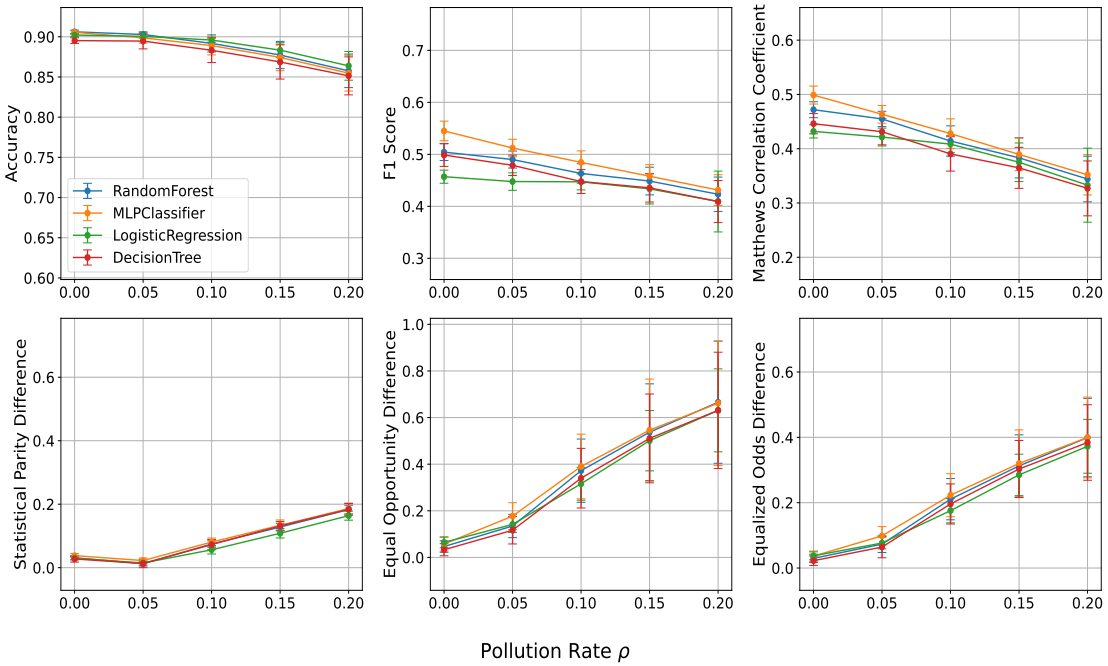
The presentation of the results follows the same structure adopted in the previous section. Figure 4.5 displays the variation of the main performance and fairness metrics as the proportion of injected unfairness increases. However, while in the earlier analysis the curves represented the different bias introduction strategies, here each line corresponds to one of the four classifiers. This allows a direct comparison of their robustness in maintaining predictive performance and fairness simultaneously. In particular, the analysis highlights how Random Forest, Neural Networks, Logistic Regression, and Decision Tree evolve under the LOW strategy, thus enabling a clearer understanding of their relative sensitivity to systematic unfairness.

The analysis of the LOW strategy shows that the classifiers behave in a broadly similar way under progressive bias. The main exception is Logistic Regression, which proved more sensitive, particularly in the COMPAS dataset, where it exhibited greater degradation of performance and fairness. Decision Trees, contrary to common expectations, did not emerge as the weakest model, showing slightly

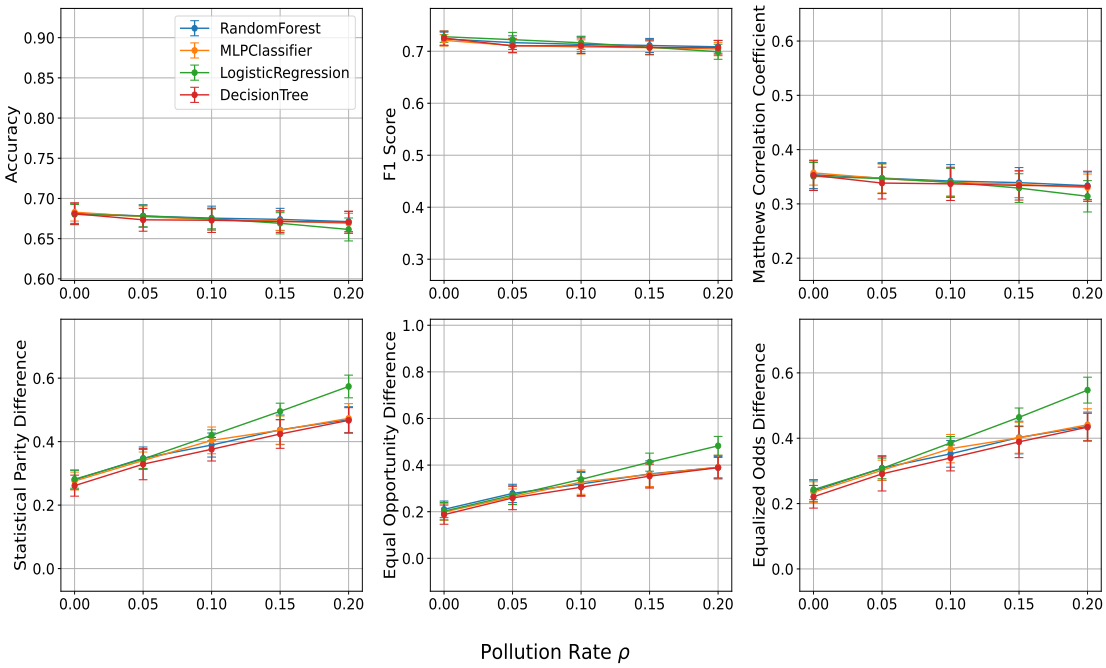
Figure 4.5: Performance and Fairness Metrics of all classifiers trained with datasets modified by the LOW strategy.



(a) Adult



(b) Bank



(c) COMPAS

lower sensitivity to unfairness compared to the others. Random Forest and Neural Networks displayed results close to each other, with marginally higher robustness overall. Although the differences are small in magnitude, these tendencies highlight Logistic Regression as the most vulnerable model in the Low scenario.

Across all datasets, the introduction of unfairness through the LOW strategy produces a consistent trade-off between predictive performance and fairness. As the ρ increases, both Acc. and MCC show a clear and progressive decline, while F1 decreases more slowly and remains comparatively stable. At the same time, fairness metrics deteriorate monotonically, with Stat. Parity, Eq. Opp., and Eq. Odds increasing steadily as more bias is introduced. This confirms that higher levels of label flipping systematically reduce the ability of classifiers to maintain predictive quality together with equitable treatment of groups.

The confusion matrix rates provide further insight into the structural impact of the LOW strategy on group outcomes. For the privileged group, the TPR increases steadily with the ρ, accompanied by a reduction in the FNR, while the TNR decreases in parallel with a rise in the FPR. The protected group exhibits the opposite behavior, with a systematic decline in the TPR and a corresponding increase in the FNR, while the TNR rises and the FPR diminishes. These mirrored movements directly reflect the bias injection procedure, which flips privileged negatives and protected positives during training. When aggregated, the overall dataset follows these trends in attenuated form, with a mild increase in the global TPR and a slight decrease in the global TNR, with variations depending on the dataset distribution.

The analysis of standard deviations highlights how the stability of classifiers is affected by increasing levels of unfairness. For performance metrics, variability grows moderately with higher ρ, while for fairness metrics the dispersion is more pronounced, especially in Eq. Opp. and Eq. Odds. This indicates that, beyond average declines in performance and fairness, the outcomes of classifiers also become less predictable. Among the models, the Neural Network and Decision Tree tend to exhibit larger fluctuations, Random Forest remains the most stable overall, and Logistic Regression shows intermediate behavior.

Considering jointly the evolution of performance, fairness, and stability, the LOW strategy reveals that the classifiers behave in broadly similar ways, but with important distinctions. Logistic Regression stands out as the most sensitive, with the largest increases in unfairness metrics and consistent drops in predictive performance, particularly in the COMPAS dataset. Neural Networks show an intermediate profile, with moderate losses that grow as bias intensifies. Decision Trees and Random Forests, in contrast, are comparatively more robust, maintaining lower levels of unfairness and more stable performance, with Random Forest slightly outperforming the others in consistency.

From a practical perspective, these findings suggest that ensemble tree-based models, such as Random Forest, are attractive options for contexts where stability in both predictive performance and fairness is desirable, especially when data quality cannot fully guarantee the absence of structural bias. Decision Trees, while less robust than their ensemble counterpart, still display competitive levels of fairness and may be suitable when interpretability is prioritized, provided that appropriate mitigation techniques are applied.

Logistic Regression, however, emerges as the classifier that most clearly requires fairness interventions. Its linear structure makes it more directly exposed to correlations between sensitive attributes and the target variable, which explains why its vulnerabilities become particularly pronounced in datasets such as COMPAS, where baseline disparities are stronger. Neural Networks, on the other hand, occupy an intermediate position: although not as fragile as Logistic Regression, they exhibit non-negligible sensitivity that calls for careful monitoring when deployed in fairness-critical domains.

Logistic Regression, however, emerges as the classifier that most clearly requires fairness interventions. Its linear structure makes it more directly exposed to correlations between sensitive attributes and the target variable, which explains why its vulnerabilities become particularly pronounced in datasets such as COMPAS, where baseline disparities are stronger. Neural Networks, on the other hand, occupy an intermediate position: although not as fragile as Logistic Regression, they exhibit non-negligible sensitivity that calls for careful monitoring when deployed in

fairness-critical domains.

A key reason for the distinctive behavior of Logistic Regression lies in its global modeling nature. Because the model defines a single linear decision boundary that depends on all training instances simultaneously, even small local perturbations in the data, such as the label flips introduced by the LOW strategy, can shift the parameters of the entire model. This global parameter coupling makes Logistic Regression particularly sensitive to localized biases, as modifications affecting a specific subset of examples propagate across the whole decision surface.

In contrast, models like Decision Trees and Random Forests react to such perturbations in a more localized manner, as changes typically affect only a few branches or subtrees. Similarly, Neural Networks, while nonlinear, tend to absorb local distortions through small adjustments in multiple weights, distributing the impact throughout the network. Consequently, Logistic Regression exhibits a uniquely amplified response to local unfairness, which explains its sharper degradation under progressive bias.

Overall, the comparative analysis shows that the choice of classifier has a direct influence on how sensitive the system will be to unfairness in the data. While no model is immune to the effects of bias injection, the magnitude of degradation and the degree of variability differ substantially across learning paradigms. These differences highlight the importance of aligning classifier selection with the expected level of bias in the data and with the available capacity for implementing fairness interventions. In practice, adopting more robust classifiers such as Random Forest can reduce the burden of corrective measures, whereas relying on sensitive models such as Logistic Regression requires a proactive and systematic approach to fairness mitigation, especially in datasets with strong correlations between sensitive attributes and outcomes, such as COMPAS.

Chapter 5

Conclusions

This chapter concludes the dissertation by revisiting the main objectives, summarizing the methodological and empirical findings, and highlighting the scientific contributions of the study. A novel framework for fairness stress testing was introduced, implemented, and applied to multiple datasets and families of classifiers. The experimental analysis revealed systematic patterns in how fairness metrics and predictive performance evolve under progressive bias injection, providing methodological innovation and empirical evidence on the robustness of ML models. These findings reinforce the relevance of stress testing as a diagnostic tool for understanding model behavior under unfair conditions and contribute to advancing the broader field of fairness in ML.

5.1 Results and Contributions

This dissertation presented a systematic investigation of the interaction between algorithmic fairness and classification performance under controlled injections of bias into training data. By means of a complete experimental framework, it was possible to analyze in depth how classifiers react to progressively unfair conditions, revealing tendencies that conventional evaluation frameworks are often unable to capture. The contributions of this work are both methodological and empirical, providing new tools and insights for the study of fairness in ML.

Although recent works have advanced fairness evaluation and bias generation methods, the literature still lacks systematic analyses comparing how different classifiers respond to progressively induced bias within a unified experimental framework. This dissertation directly addresses this gap by combining controlled bias introduction with comparative assessments of classifier robustness, thereby providing an empirical foundation for understanding how unfairness propagates differently across model architectures.

The first major contribution is methodological. A new approach, named Systematic Label Flipping for Fairness Stress Testing, was proposed and implemented to inject bias into datasets in a structured and reproducible way. Unlike unsystematic perturbations or uncontrolled noise, this method establishes predefined strategies and thresholds for selectively flipping labels of protected and privileged groups. This systematic procedure enables classifiers to be subjected to fairness-oriented stress tests, analogous to the robustness assessments commonly used in other engineering domains. The framework allows not only the observation of a model's fairness at a given point, but also its robustness trajectory as data unfairness grows, offering a new perspective for both researchers and practitioners.

A second contribution is the implementation of a reproducible and extensible experimental framework. All stages of the process — preprocessing, model training, hyperparameter tuning with Optuna, bias injection, metric computation, and visualization — were integrated in a modular architecture, designed to facilitate reuse and extension. The framework consolidates results with averages and standard deviations across folds, ensuring statistical rigor. This framework can serve as a foundation for future research, enabling the community to stress test additional datasets, models, or fairness definitions in a transparent and replicable manner.

A third important contribution is the comparative evaluation of different label-flipping strategies. By contrasting approaches such as LOW, HIGH, and RANDOM, the study identified how alternative ways of injecting bias influence the detection of model vulnerabilities. This analysis showed that strategies differ in the extent to which they stress the classifiers, providing guidance on which procedures are more effective in exposing fragility in fairness.

The fourth contribution lies in the empirical findings regarding classifier robustness. Experiments were conducted on three benchmark datasets widely used in fairness research, Bank Marketing, Adult Income, and COMPAS Recidivism, applying the stress testing methodology across incremental thresholds of bias. The results showed that, overall, models behaved in a broadly similar way, but with Logistic Regression emerging as the most sensitive, particularly in the COMPAS dataset, where fairness metrics deteriorated more sharply. Decision Trees displayed slightly greater resilience, while Random Forest consistently proved the most stable, combining relatively high predictive performance with lower fairness degradation. Neural Networks occupied an intermediate position, showing moderate robustness but also higher variability across metrics. These results enrich the understanding of how different algorithmic structures respond when exposed to unfair data, highlighting that ensemble methods tend to be more robust, whereas linear models are the most vulnerable.

In summary, the main results of this dissertation can be expressed as follows:

the proposed methodology successfully produced systematic stress tests for fairness; alternative bias injection strategies were evaluated and compared; Random Forest emerged as the most robust classifier, while Logistic Regression was the most sensitive, with Neural Networks and Decision Trees occupying intermediate positions; fairness and accuracy were shown to interact in non-trivial ways; and a modular, reproducible framework was delivered as a research artifact. Taken together, these contributions advance both the methodological and empirical knowledge of fairness in ML, emphasizing that robustness to unfair data is as important as static fairness evaluation at deployment.

5.2 Future Research

The findings and contributions of this dissertation open several avenues for future research. Although the proposed methodology and experiments offered solid evidence on the robustness of different classifiers under fairness stress testing, important challenges remain and new opportunities deserve to be pursued.

A natural extension of this work is to broaden the scope of datasets beyond the three benchmarks studied here. While Adult, Bank Marketing, and COMPAS are well established in the fairness literature, their particularities limit the generalization of conclusions. Applying the proposed framework to domains such as healthcare, credit scoring, recruitment, or recommendation systems could reveal new insights into the behavior of fairness-sensitive models in contexts with different statistical properties and ethical implications. Moreover, datasets containing multiple sensitive attributes would allow the study of intersectional fairness, a pressing and still underexplored problem.

Another direction is the inclusion of a wider range of models in the evaluation. This dissertation focused on classical classifiers such as Random Forest, Decision Tree, Logistic Regression, and Neural Networks. However, recent advances in deep learning, gradient boosting ensembles, and transformer-based architectures may display distinct patterns of robustness. Incorporating such models into the stress testing framework would enable a more comprehensive understanding of the trade-offs between accuracy and fairness in state-of-the-art systems.

Methodological refinements also represent promising extensions. The systematic label flipping developed in this work proved effective and reproducible, but other strategies for bias injection could be explored. Perturbations at the feature level, the creation of synthetic correlations between sensitive attributes and outcomes, or adversarial modifications guided by optimization procedures are possible alternatives. These extensions could expose new vulnerabilities and expand the diagnostic power of stress testing.

An additional line of research involves integrating mitigation techniques directly into the stress testing process. In this dissertation, models were analyzed without explicit fairness constraints in order to evaluate their natural robustness. Future work could embed pre-processing, in-processing, and post-processing interventions into the framework, assessing their effectiveness under progressively biased data. This would provide a clearer picture of the real capacity of mitigation strategies to preserve equity in adverse scenarios.

Finally, applying the methodology in dynamic and real-world contexts is a crucial frontier. Fairness challenges often arise from temporal shifts in data distributions, feedback loops, and strategic responses by individuals subject to algorithmic decisions. Extending fairness stress testing to streaming data, sequential decision-making, or reinforcement learning settings could deliver a richer and more realistic perspective on robustness. Such studies are particularly relevant in high-stakes applications where fairness must be maintained continuously rather than only at the point of deployment.

In conclusion, this dissertation established a foundation for fairness stress testing through systematic label flipping and rigorous experimentation. Building on this foundation, future research can expand datasets and models, refine bias injection methods, integrate mitigation strategies, and address dynamic environments. Together, these directions can advance the state of knowledge in Fairness in ML and contribute to the development of systems that are both accurate and resiliently fair across contexts and over time.

References

MEHRABI, N., MORSTATTER, F., SAXENA, N., et al. "A Survey on Bias and Fairness in Machine Learning". jan. 2022. Disponível em: <http://arxiv.org/abs/1908.09635>. arXiv:1908.09635 [cs].

ATKINSON, G., METSIS, V. "A Survey of Methods for Detection and Correction of Noisy Labels in Time Series Data". In: Maglogiannis, I., Macintyre, J., Iliadis, L. (Eds.), *Artificial Intelligence Applications and Innovations*, pp. 479–493, Cham, 2021. Springer International Publishing. ISBN: 978-3-030-79150-6. doi: 10.1007/978-3-030-79150-6_38.

OSOBA, O. A., WELSER, W. I. *An Intelligence in Our Image: The Risks of Bias and Errors in Artificial Intelligence*. Relatório técnico, RAND Corporation, abr. 2017. Disponível em: <https://www.rand.org/pubs/research_reports/RR1744.html>.

HOWARD, A., BORENSTEIN, J. "The Ugly Truth About Ourselves and Our Robot Creations: The Problem of Bias and Social Inequity", *Science and Engineering Ethics*, v. 24, n. 5, pp. 1521–1536, out. 2018. ISSN: 1471-5546. doi: 10.1007/s11948-017-9975-2. Disponível em: <https://doi.org/10.1007/s11948-017-9975-2>.

WOLF, M., MILLER, K., GRODZINSKY, F. "Why We Should Have Seen That Coming", *The ORBIT Journal*, v. 1, n. 2, pp. 1–12, 2017. ISSN: 25158562. doi: 10.29297/orbit.v1i2.49. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S2515856220300493>.

RAJI, I. D., BUOLAMWINI, J. "Actionable Auditing Revisited: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products", *Commun. ACM*, v. 66, n. 1, pp. 101–108, dez. 2022. ISSN: 0001-0782. doi: 10.1145/3571151. Disponível em: <https://dl.acm.org/doi/10.1145/3571151>.

BARENSTEIN, M. "ProPublica's COMPAS Data Revisited". jul. 2019. Disponível em: <http://arxiv.org/abs/1906.04711>. arXiv:1906.04711 [econ].

BAROCAS, S., HARDT, M., NARAYANAN, A. *Fairness and Machine Learning*. dez. 2023. Disponível em: <fairmlbook.org>.

HLEG, H. L. E. G. *Ethics guidelines for trustworthy AI*. Publications Office of the European Union, 2019. ISBN: 978-92-76-11998-2. Disponível em: <https://data.europa.eu/doi/10.2759/346720>.

PEDRESHI, D., RUGGIERI, S., TURINI, F. "Discrimination-aware data mining". In: *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '08, pp. 560–568, New York, NY, USA, ago. 2008. Association for Computing Machinery. ISBN: 978-1-60558-193-4. doi: 10.1145/1401890.1401959. Disponível em: <https://doi.org/10.1145/1401890.1401959>.

CATON, S., HAAS, C. "Fairness in Machine Learning: A Survey", *ACM Computing Surveys*, v. 56, n. 7, pp. 166:1–166:38, abr. 2024. ISSN: 0360-0300. doi: 10.1145/3616865. Disponível em: <https://dl.acm.org/doi/10.1145/3616865>.

SAXENA, N., HUANG, K., DEFILIPPIS, E., et al. "How Do Fairness Definitions Fare? Examining Public Attitudes Towards Algorithmic Definitions of Fairness". jan. 2019. Disponível em: <http://arxiv.org/abs/1811.03654>. arXiv:1811.03654 [cs].

AQUINAS, T. *Summa Theologiae*. New York, Benziger Bros., 1274. Disponível em: <https://www.newadvent.org/summa/3058.htm#article11>. II-II, Question 58, Article 11.

CASTELNOVO, A., CRUPI, R., GRECO, G., et al. "A Clarification of the Nuances in the Fairness Metrics Landscape", *Scientific Reports*, v. 12, n. 1, pp. 4209, mar. 2022a. ISSN: 2045-2322. doi: 10.1038/s41598-022-07939-1. Disponível em: <http://arxiv.org/abs/2106.00467>. arXiv:2106.00467 [cs].

CHOULDECHOVA, A. "Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments", *Big Data*, v. 5, n. 2, pp. 153–163, jun. 2017. ISSN: 2167-6461. doi: 10.1089/big.2016.0047. Disponível em: <https://www.liebertpub.com/doi/10.1089/big.2016.0047>. Publisher: Mary Ann Liebert, Inc., publishers.

KLEINBERG, J., MULLAINATHAN, S., RAGHAVAN, M. "Inherent Trade-Offs in the Fair Determination of Risk Scores". nov. 2016. Disponível em: <http://arxiv.org/abs/1609.05807>. arXiv:1609.05807 [cs].

BELL, A., BYNUM, L., DRUSHCHAK, N., et al. "The Possibility of Fairness: Revisiting the Impossibility Theorem in Practice". In: *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, pp. 400–422, New York, NY, USA, jun. 2023. Association for Computing Machinery. ISBN: 979-8-4007-0192-4. doi: 10.1145/3593013.3594007. Disponível em: <https://doi.org/10.1145/3593013.3594007>.

BEIGANG, F. "Yet Another Impossibility Theorem in Algorithmic Fairness", *Minds and Machines*, v. 33, n. 4, pp. 715–735, dez. 2023. ISSN: 1572-8641. doi: 10.1007/s11023-023-09645-x. Disponível em: <https://doi.org/10.1007/s11023-023-09645-x>.

SELBST, A. D., BOYD, D., FRIEDLER, S. A., et al. "Fairness and Abstraction in Sociotechnical Systems". In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pp. 59–68, New York, NY, USA, jan. 2019. Association for Computing Machinery. ISBN: 978-1-4503-6125-5. doi: 10.1145/3287560.3287598. Disponível em: <https://dl.acm.org/doi/10.1145/3287560.3287598>.

LIU, L. T., DEAN, S., ROLF, E., et al. "Delayed Impact of Fair Machine Learning". abr. 2018. Disponível em: <http://arxiv.org/abs/1803.04383>. arXiv:1803.04383 [cs].

FERRARA, E. "Fairness and Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, and Mitigation Strategies", *Sci*, v. 6, n. 1, pp. 3, mar. 2024. ISSN: 2413-4155. doi: 10.3390/sci6010003. Disponível em: <https://www.mdpi.com/2413-4155/6/1/3>. Number: 1 Publisher: Multidisciplinary Digital Publishing Institute.

KAMIRAN, F., ŽLIOBAITĖ, I. "Explainable and Non-explainable Discrimination in Classification". In: Custers, B., Calders, T., Schermer, B., et al. (Eds.), *Discrimination and Privacy in the Information Society: Data Mining and Profiling in Large Databases*, Springer, pp. 155–170, Berlin, Heidelberg, 2013. ISBN: 978-3-642-30487-3. doi: 10.1007/978-3-642-30487-3_8. Disponível em: <https://doi.org/10.1007/978-3-642-30487-3_8>.

BERK, R. "Accuracy and Fairness for Juvenile Justice Risk Assessments", *Journal of Empirical Legal Studies*, v. 16, n. 1, pp. 175–194, 2019. ISSN: 1740-1461. doi: 10.1111/jels.12206. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1111/jels.12206>. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/jels.12206.

LEE, N. T. "Detecting racial bias in algorithms and machine learning", *Journal of Information, Communication and Ethics in Society*, v. 16, n. 3, pp. 252–260, ago. 2018. ISSN: 1477-996X. doi: 10.1108/JICES-06-2018-0056. Disponível em: <https://www.emerald.com/insight/content/doi/10.1108/jices-06-2018-0056/full/html>. Publisher: Emerald Publishing Limited.

CHIAPPA, S., ISAAC, W. S. "A Causal Bayesian Networks Viewpoint on Fairness". v. 547, pp. 3–20, 2019. doi: 10.1007/978-3-030-16744-8_1. Disponível em: <http://arxiv.org/abs/1907.06430>. arXiv:1907.06430 [stat].

ZARSKY, T. "The Trouble with Algorithmic Decisions: An Analytic Road Map to Examine Efficiency and Fairness in Automated and Opaque Decision Making", *Science, Technology, & Human Values*, v. 41, n. 1, pp. 118–132, jan. 2016. ISSN: 0162-2439. doi: 10.1177/0162243915605575. Disponível em: <https://doi.org/10.1177/0162243915605575>. Publisher: SAGE Publications Inc.

VEALE, M., BINNS, R. "Fairer Machine Learning in the Real World: Mitigating Discrimination Without Collecting Sensitive Data". out. 2017. Disponível em: <https://papers.ssrn.com/abstract=3060763>.

ZIMMER, M. ""But the data is already public": on the ethics of research in Facebook", *Ethics and Information Technology*, v. 12, n. 4, pp. 313–325, dez. 2010. ISSN: 1572-8439. doi: 10.1007/s10676-010-9227-5. Disponível em: <https://doi.org/10.1007/s10676-010-9227-5>.

CRENSHAW, K. "Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Politics", 1989.

GREEN, B. ""Fair" Risk Assessments: A Precarious Approach for Criminal Justice Reform". 2018. Place: Stockholm, Sweden.

CORBETT-DAVIES, S., GAEBLER, J. D., NILFOROSHAN, H., et al. "The Measure and Mismeasure of Fairness". ago. 2023. Disponível em: <http://arxiv.org/abs/1808.00023>. arXiv:1808.00023 [cs].

GURSOY, F., KAKADIARIS, I. A. "Equal Confusion Fairness: Measuring Group-Based Disparities in Automated Decision Systems". In: *2022 IEEE International Conference on Data Mining Workshops (ICDMW)*, pp. 137–146, nov. 2022. doi: 10.1109/ICDMW58026.2022.00027. Disponível em: <http://arxiv.org/abs/2307.00472>. arXiv:2307.00472 [cs].

DWORK, C., HARDT, M., PITASSI, T., et al. "Fairness Through Awareness". nov. 2011. Disponível em: <http://arxiv.org/abs/1104.3913>. arXiv:1104.3913 [cs].

HARDT, M., PRICE, E., SREBRO, N. "Equality of Opportunity in Supervised Learning". out. 2016. Disponível em: <http://arxiv.org/abs/1610.02413>. arXiv:1610.02413 [cs].

KUSNER, M. J., LOFTUS, J. R., RUSSELL, C., et al. "Counterfactual Fairness". mar. 2018. Disponível em: <http://arxiv.org/abs/1703.06856>. arXiv:1703.06856 [stat].

CALMON, F. P., WEI, D., VINZAMURI, B., et al. "Optimized pre-processing for discrimination prevention". In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 3995–4004, Red Hook, NY, USA, dez. 2017. Curran Associates Inc. ISBN: 978-1-5108-6096-4.

AGARWAL, A., BEYGELZIMER, A., DUDÍK, M., et al. "A Reductions Approach to Fair Classification". jul. 2018. Disponível em: <http://arxiv.org/abs/1803.02453>. arXiv:1803.02453 [cs].

SPEICHER, T., HEIDARI, H., GRGIC-HLACA, N., et al. "A Unified Approach to Quantifying Algorithmic Unfairness: Measuring Individual & Group Unfairness via Inequality Indices". In: *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2239–2248, jul. 2018. doi: 10.1145/3219819.3220046. Disponível em: <http://arxiv.org/abs/1807.00787>. arXiv:1807.00787 [cs].

SKIRPAN, M., GORELICK, M. "The Authority of "Fair" in Machine Learning". jul. 2017. Disponível em: <http://arxiv.org/abs/1706.09976>. arXiv:1706.09976 [cs].

D'ALESSANDRO, B., O'NEIL, C., LAGATTA, T. "Conscientious Classification: A Data Scientist's Guide to Discrimination-Aware Classification", *Big Data*, v. 5, n. 2, pp. 120–134, jun. 2017. ISSN: 2167-6461, 2167-647X. doi: 10.1089/big.2016.0048. Disponível em: <http://arxiv.org/abs/1907.09013>. arXiv:1907.09013 [stat].

BARBIERATO, E., DELLA VEDOVA, M., TESSERA, D., et al. "A Methodology for Controlling Bias and Fairness in Synthetic Data Generation", *Applied Sciences (Switzerland)*, v. 12, n. 9, 2022. doi: 10.3390/app12094619.

LEPRI, B., OLIVER, N., LETOUZÉ, E., et al. "Fair, Transparent, and Accountable Algorithmic Decision-making Processes", *Philosophy & Technology*, v. 31, n. 4, pp. 611–627, dez. 2018. ISSN: 2210-5441. doi: 10.1007/s13347-017-0279-x. Disponível em: <https://doi.org/10.1007/s13347-017-0279-x>.

LUM, K., JOHNDROW, J. "A statistical framework for fair predictive algorithms". out. 2016. Disponível em: <http://arxiv.org/abs/1610.08077>. arXiv:1610.08077 [stat].

PARETO, V. *Manuale di economia politica con una introduzione alla scienza sociale*. Milano : Societa Editrice Libraria, 1919. Disponível em: <http://archive.org/details/manualedieconomi00pareuoft>.

CHICCO, D., JURMAN, G. "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation", *BMC Genomics*, v. 21, n. 1, pp. 6, jan. 2020. ISSN: 1471-2164. doi: 10.1186/s12864-019-6413-7. Disponível em: <https://doi.org/10.1186/s12864-019-6413-7>.

ZHANG, L., WU, Y., WU, X. "Achieving non-discrimination in prediction". mar. 2018. Disponível em: <http://arxiv.org/abs/1703.00060>. arXiv:1703.00060 [cs].

XU, D., YUAN, S., ZHANG, L., et al. "FairGAN+: Achieving Fair Data Generation and Classification through Generative Adversarial Nets". In: *2019 IEEE International Conference on Big Data (Big Data)*, pp. 1401–1406, dez. 2019. doi: 10.1109/BigData47090.2019.9006322. Disponível em: <https://ieeexplore.ieee.org/document/9006322>.

BAO, M., ZHOU, A., ZOTTOLA, S., et al. "It's COMPASlicated: The Messy Relationship between RAI Datasets and Algorithmic Fairness Benchmarks". abr. 2022. Disponível em: <http://arxiv.org/abs/2106.05498>. arXiv:2106.05498 [cs].

FABRIS, A., MESSINA, S., SILVELLO, G., et al. "Algorithmic Fairness Datasets: the Story so Far", *Data Mining and Knowledge Discovery*, v. 36, n. 6, pp. 2074–2152, nov. 2022. ISSN: 1384-5810, 1573-756X. doi: 10.1007/s10618-022-00854-z. Disponível em: <http://arxiv.org/abs/2202.01711>. arXiv:2202.01711 [cs].

QUY, T. L., ROY, A., IOSIFIDIS, V., et al. "A survey on datasets for fairness-aware machine learning", *WIREs Data Mining and Knowledge Discovery*,

v. 12, n. 3, pp. e1452, maio 2022. ISSN: 1942-4787, 1942-4795. doi: 10.1002/widm.1452. Disponível em: <http://arxiv.org/abs/2110.00530>. arXiv:2110.00530 [cs].

PAULLADA, A., RAJI, I. D., BENDER, E. M., et al. "Data and its (dis)contents: A survey of dataset development and use in machine learning research", *Patterns*, v. 2, n. 11, pp. 100336, nov. 2021. ISSN: 26663899. doi: 10.1016/j.patter.2021.100336. Disponível em: <http://arxiv.org/abs/2012.05345>. arXiv:2012.05345 [cs].

BELITZ, C., OCUMPAUGH, J., RITTER, S., et al. "Constructing categories: Moving beyond protected classes in algorithmic fairness", *Journal of the Association for Information Science and Technology*, v. 74, n. 6, pp. 663–668, 2023. ISSN: 2330-1643. doi: 10.1002/asi.24643. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.24643>. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/asi.24643.

KEYES, O. "The Misgendering Machines: Trans/HCI Implications of Automatic Gender Recognition", *Proc. ACM Hum.-Comput. Interact.*, v. 2, n. CSCW, pp. 88:1–88:22, nov. 2018. doi: 10.1145/3274357. Disponível em: <https://doi.org/10.1145/3274357>.

SCHEUERMAN, M. K., WADE, K., LUSTIG, C., et al. "How We've Taught Algorithms to See Identity: Constructing Race and Gender in Image Databases for Facial Analysis", *Proceedings of the ACM on Human-Computer Interaction*, v. 4, n. CSCW1, pp. 1–35, maio 2020. ISSN: 2573-0142. doi: 10.1145/3392866. Disponível em: <https://dl.acm.org/doi/10.1145/3392866>.

DING, F., HARDT, M., MILLER, J., et al. "Retiring Adult: New Datasets for Fair Machine Learning". jan. 2022. Disponível em: <http://arxiv.org/abs/2108.04884>. arXiv:2108.04884 [cs].

DRECHSLER, J. "Using Support Vector Machines for Generating Synthetic Datasets". In: Domingo-Ferrer, J., Magkos, E. (Eds.), *Privacy in Statistical Databases*, pp. 148–161, Berlin, Heidelberg, 2010. Springer. ISBN: 978-3-642-15838-4. doi: 10.1007/978-3-642-15838-4_14.

BUOLAMWINI, J., GEBRU, T. "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification". In: *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, pp. 77–91.

PMLR, jan. 2018. Disponível em: <https://proceedings.mlr.press/v81/buolamwini18a.html>. ISSN: 2640-3498.

WANG, A., RAMASWAMY, V. V., RUSSAKOVSKY, O. "Towards Intersectionality in Machine Learning: Including More Identities, Handling Underrepresentation, and Performing Evaluation". In: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pp. 336–349, New York, NY, USA, jun. 2022. Association for Computing Machinery. ISBN: 978-1-4503-9352-2. doi: 10.1145/3531146.3533101. Disponível em: <https://dl.acm.org/doi/10.1145/3531146.3533101>.

BAKER, R. S., HAWN, A. "Algorithmic Bias in Education", *International Journal of Artificial Intelligence in Education*, v. 32, n. 4, pp. 1052–1092, dez. 2022. ISSN: 1560-4306. doi: 10.1007/s40593-021-00285-9. Disponível em: <https://doi.org/10.1007/s40593-021-00285-9>.

COCK, J. M., BILAL, M., DAVIS, R., et al. "Protected Attributes Tell Us Who, Behavior Tells Us How: A Comparison of Demographic and Behavioral Oversampling for Fair Student Success Modeling". In: *LAK23: 13th International Learning Analytics and Knowledge Conference*, pp. 488–498, mar. 2023. doi: 10.1145/3576050.3576149. Disponível em: <http://arxiv.org/abs/2212.10166>. arXiv:2212.10166 [cs].

HAGESTEDT, I., ZHANG, Y., HUMBERT, M., et al. "MBeacon: Privacy-Preserving Beacons for DNA Methylation Data". In: *Proceedings 2019 Network and Distributed System Security Symposium*, San Diego, CA, 2019. Internet Society. ISBN: 978-1-891562-55-6. doi: 10.14722/ndss.2019.23064. Disponível em: <https://www.ndss-symposium.org/wp-content/uploads/2019/02/ndss2019_03A-2_Hagestedt_paper.pdf>.

PYRGELIS, A., TRONCOSO, C., CRISTOFARO, E. D. "Knock Knock, Who's There? Membership Inference on Aggregate Location Data". nov. 2017. Disponível em: <http://arxiv.org/abs/1708.06145>. arXiv:1708.06145 [cs].

SHOKRI, R., STRONATI, M., SONG, C., et al. "Membership Inference Attacks against Machine Learning Models". mar. 2017. Disponível em: <http://arxiv.org/abs/1610.05820>. arXiv:1610.05820 [cs].

FREDRIKSON, M., JHA, S., RISTENPART, T. "Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures". In: *Pro-*

ceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, CCS '15, pp. 1322–1333, New York, NY, USA, out. 2015. Association for Computing Machinery. ISBN: 978-1-4503-3832-5. doi: 10.1145/2810103.2813677. Disponível em: <https://dl.acm.org/doi/10.1145/2810103.2813677>.

MELIS, L., SONG, C., CRISTOFARO, E. D., et al. "Exploiting Unintended Feature Leakage in Collaborative Learning". nov. 2018. Disponível em: <http://arxiv.org/abs/1805.04049>. arXiv:1805.04049 [cs].

ABOWD, J. M., VILHUBER, L. "How Protective Are Synthetic Data?" In: Domingo-Ferrer, J., Saygın, Y. (Eds.), *Privacy in Statistical Databases*, pp. 239–246, Berlin, Heidelberg, 2008. Springer. ISBN: 978-3-540-87471-3. doi: 10.1007/978-3-540-87471-3_20.

HAND, D. J. "Synthetic Datasets for Statistical Disclosure Control: Theory and Implementation by Jörg Drechsler", *International Statistical Review*, v. 80, n. 3, pp. 483–483, 2012. ISSN: 1751-5823. doi: 10.1111/j.1751-5823.2012.00196_15.x. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1751-5823.2012.00196_15.x>. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1751-5823.2012.00196_15.x.

KHAYRALLAH, H., KOEHN, P. "On the Impact of Various Types of Noise on Neural Machine Translation". In: *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pp. 74–83, 2018. doi: 10.18653/v1/W18-2709. Disponível em: <http://arxiv.org/abs/1805.12282>. arXiv:1805.12282 [cs].

MELAMUD, O., SHIVADE, C. "Towards Automatic Generation of Shareable Synthetic Clinical Notes Using Neural Language Models". In: Rumshisky, A., Roberts, K., Bethard, S., et al. (Eds.), *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pp. 35–45, Minneapolis, Minnesota, USA, jun. 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-1905. Disponível em: <https://aclanthology.org/W19-1905/>.

BAROCAS, S., SELBST, A. D. "Big Data's Disparate Impact". 2016. Disponível em: <https://papers.ssrn.com/abstract=2477899>.

JIANG, L., BELITZ, C., BOSCH, N. "Synthetic Dataset Generation for Fairer Unfairness Research". In: *Proceedings of the 14th Learning Analytics and*

Knowledge Conference, LAK '24, pp. 200–209, New York, NY, USA, mar. 2024. Association for Computing Machinery. ISBN: 979-8-4007-1618-8. doi: 10.1145/3636555.3636868. Disponível em: <https://dl.acm.org/doi/10.1145/3636555.3636868>.

CASTELNOVO, A., CRUPI, R., INVERARDI, N., et al. "Investigating Bias with a Synthetic Data Generator: Empirical Evidence and Philosophical Interpretation". set. 2022b. Disponível em: <http://arxiv.org/abs/2209.05889>. arXiv:2209.05889 [stat].

GUPTA, A., BHATT, D., PANDEY, A. "Transitioning from Real to Synthetic data: Quantifying the bias in model". maio 2021. Disponível em: <http://arxiv.org/abs/2105.04144>. arXiv:2105.04144 [cs].

ASSEFA, S. "Generating Synthetic Data in Finance: Opportunities, Challenges and Pitfalls", *SSRN Electronic Journal*, 2020. ISSN: 1556-5068. doi: 10.2139/ssrn.3634235. Disponível em: <https://www.ssrn.com/abstract=3634235>.

AGUIAR, R., COLLARES-PEREIRA, M. "TAG: A time-dependent, autoregressive, Gaussian model for generating synthetic hourly radiation", *Solar Energy*, v. 49, n. 3, pp. 167–174, set. 1992. ISSN: 0038-092X. doi: 10.1016/0038-092X(92)90068-L. Disponível em: <https://www.sciencedirect.com/science/article/pii/0038092X9290068L>.

SINGH, R., PAL, B., JABR, R. "Statistical Representation of Distribution System Loads Using Gaussian Mixture Model", *IEEE Transactions on Power Systems*, v. 25, n. 1, pp. 29–37, fev. 2010. ISSN: 0885-8950, 1558-0679. doi: 10.1109/TPWRS.2009.2030271. Disponível em: <http://ieeexplore.ieee.org/document/5298967/>.

ZHANG, J., CORMODE, G., PROCOPIUC, C. M., et al. "PrivBayes: private data release via bayesian networks". In: *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*, SIGMOD '14, pp. 1423–1434, New York, NY, USA, jun. 2014. Association for Computing Machinery. ISBN: 978-1-4503-2376-5. doi: 10.1145/2588555.2588573. Disponível em: <https://doi.org/10.1145/2588555.2588573>.

CAIOLA, G., REITER, J. P. "Random Forests for Generating Partially Synthetic, Categorical Data", 2010.

CHEN, R. J., LU, M. Y., CHEN, T. Y., et al. "Synthetic data in machine learning for medicine and healthcare", *Nature Biomedical Engi-*

neering, v. 5, n. 6, pp. 493–497, jun. 2021. ISSN: 2157-846X. doi: 10.1038/s41551-021-00751-8. Disponível em: <https://www.nature.com/articles/s41551-021-00751-8>. Publisher: Nature Publishing Group.

GOODFELLOW, I. J., POUGET-ABADIE, J., MIRZA, M., et al. "Generative Adversarial Networks". jun. 2014. Disponível em: <http://arxiv.org/abs/1406.2661>. arXiv:1406.2661 [stat].

BREUGEL, B. V., KYONO, T., BERREVOETS, J., et al. "DECAF: Generating Fair Synthetic Data Using Causally-Aware Generative Networks". nov. 2021. Disponível em: <http://arxiv.org/abs/2110.12884>. arXiv:2110.12884 [cs].

FRENAY, B., VERLEYSEN, M. "Classification in the Presence of Label Noise: A Survey", *IEEE Transactions on Neural Networks and Learning Systems*, v. 25, n. 5, pp. 845–869, maio 2014. ISSN: 2162-2388. doi: 10.1109/TNNLS.2013.2292894. Disponível em: <https://ieeexplore.ieee.org/document/6685834>. Conference Name: IEEE Transactions on Neural Networks and Learning Systems.

HICKEY, R. J. "Noise modelling and evaluating learning from examples", *Artificial Intelligence*, v. 82, n. 1, pp. 157–179, abr. 1996. ISSN: 0004-3702. doi: 10.1016/0004-3702(94)00094-8. Disponível em: <https://www.sciencedirect.com/science/article/pii/0004370294000948>.

QUINLAN, J. R. "Induction of decision trees", *Machine Learning*, v. 1, n. 1, pp. 81–106, mar. 1986. ISSN: 1573-0565. doi: 10.1007/BF00116251. Disponível em: <https://doi.org/10.1007/BF00116251>.

WANG, J., LIU, Y., LEVY, C. "Fair Classification with Group-Dependent Label Noise". In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 526–536, mar. 2021. doi: 10.1145/3442188.3445915. Disponível em: <http://arxiv.org/abs/2011.00379>. arXiv:2011.00379 [cs].

ZHANG, W., CHENG, D., LU, G., et al. "Efficient Adaptive Label Refinement for Label Noise Learning". fev. 2025. Disponível em: <http://arxiv.org/abs/2502.00386>. arXiv:2502.00386 [cs].

MEHRABI, N., NAVEED, M., MORSTATTER, F., et al. "Exacerbating Algorithmic Bias through Fairness Attacks". dez. 2020. Disponível em: <http://arxiv.org/abs/2012.08723>. arXiv:2012.08723 [cs].

PANEL, E. "13 Strategies For Collecting High-Quality Data". nov. 2020. Disponível em: <https://www.forbes.com/councils/forbescommunicationscouncil/2020/11/17/13-strategies-for-collecting-high-quality-data/>. Section: Leadership.

KAMIRAN, F., CALDERS, T. "Data preprocessing techniques for classification without discrimination", *Knowledge and Information Systems*, v. 33, n. 1, pp. 1–33, out. 2012. ISSN: 0219-3116. doi: 10.1007/s10115-011-0463-8. Disponível em: <https://doi.org/10.1007/s10115-011-0463-8>.

ZHANG, H., TAE, K. H., PARK, J., et al. "iFlipper: Label Flipping for Individual Fairness". set. 2022. Disponível em: <http://arxiv.org/abs/2209.07047>. arXiv:2209.07047 [cs].

WICK, M., PANDA, S., TRISTAN, J.-B. "Unlocking Fairness: a Trade-off Revisited". In: *Advances in Neural Information Processing Systems*, v. 32. Curran Associates, Inc., 2019. Disponível em: <https://proceedings.neurips.cc/paper/2019/hash/373e4c5d8edfa8b74fd4b6791d0cf6dc-Abstract.html>.

FRIEDLER, S. A., SCHEIDEGGER, C., VENKATASUBRAMANIAN, S., et al. "A comparative study of fairness-enhancing interventions in machine learning". fev. 2018. Disponível em: <http://arxiv.org/abs/1802.04422>. arXiv:1802.04422 [stat].

NORTHCUTT, C. G., ATHALYE, A., MUELLER, J. "Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks". nov. 2021. Disponível em: <http://arxiv.org/abs/2103.14749>. arXiv:2103.14749 [stat].

ELKAN, C. "The foundations of cost-sensitive learning". In: *Proceedings of the 17th international joint conference on Artificial intelligence - Volume 2*, IJCAI'01, pp. 973–978, San Francisco, CA, USA, ago. 2001. Morgan Kaufmann Publishers Inc. ISBN: 978-1-55860-812-2.

BARRY BECKER, R. K. "Adult". 1996. Disponível em: <https://archive.ics.uci.edu/dataset/2>.

S. MORO, P. R. "Bank Marketing". 2014. Disponível em: <https://archive.ics.uci.edu/dataset/222>.

BREIMAN, L. "Random Forests", *Machine Learning*, v. 45, n. 1, pp. 5–32, out. 2001. ISSN: 1573-0565. doi: 10.1023/A:1010933404324. Disponível em: <https://doi.org/10.1023/A:1010933404324>.

HOSMER, D. W., LEMESHOW, S., STURDIVANT, R. X. *Applied Logistic Regression*. Wiley Series in Probability and Statistics. 3. aufl ed. Hoboken, N.J, Wiley, 2013. ISBN: 978-0-470-58247-3 978-1-118-54839-4.

GOODFELLOW, I., BENGIO, Y., COURVILLE, A. *Deep Learning*. Adaptive Computation and Machine Learning series. Cambridge, MA, USA, MIT Press, nov. 2016. ISBN: 978-0-262-03561-3. Disponível em: <https://mitpress.mit.edu/9780262035613/deep-learning/>.

BERGSTRA, J., BARDENET, R., BENGIO, Y., et al. "Algorithms for Hyper-Parameter Optimization". In: *Advances in Neural Information Processing Systems*, v. 24. Curran Associates, Inc., 2011. Disponível em: <https://proceedings.neurips.cc/paper_files/paper/2011/hash/86e8f7ab32cfd12577bc2619bc635690-Abstract.html>.

AHFOCK, D., MCLACHLAN, G. J. "Harmless label noise and informative soft-labels in supervised classification". abr. 2021. Disponível em: <http://arxiv.org/abs/2104.02872>. arXiv:2104.02872 [stat].

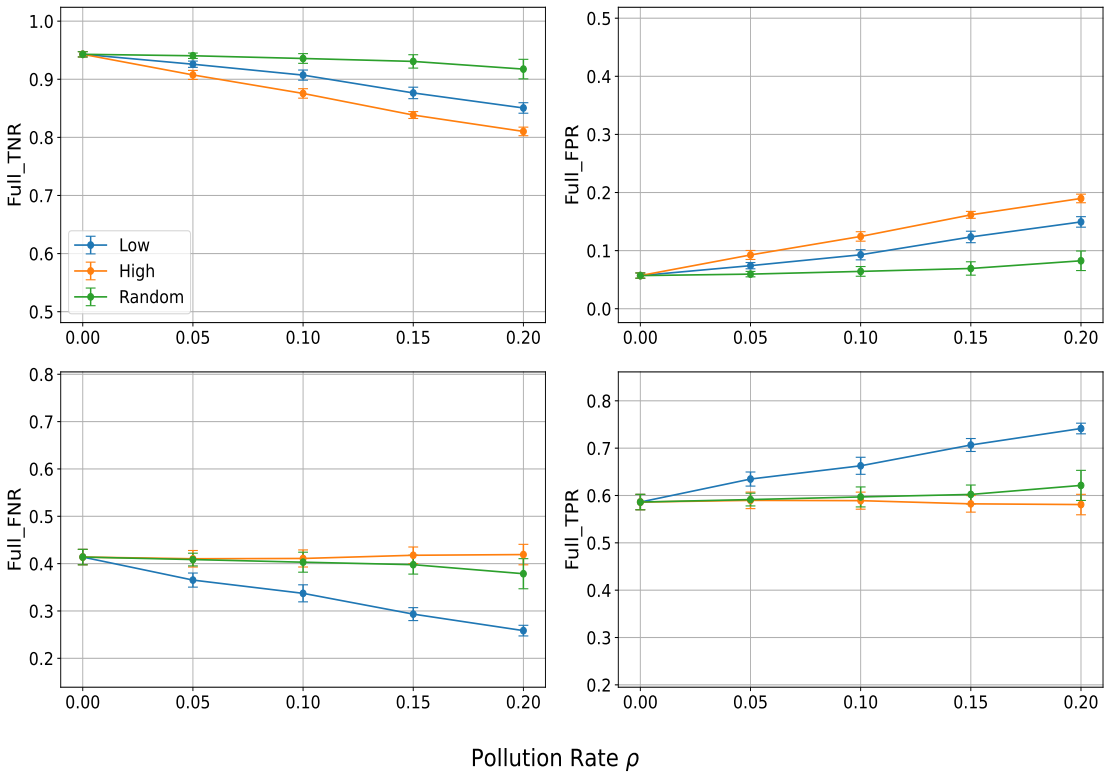
MENON, A., ROOYEN, B. V., ONG, C. S., et al. "Learning from Corrupted Binary Labels via Class-Probability Estimation". In: *Proceedings of the 32nd International Conference on Machine Learning*, pp. 125–134. PMLR, jun. 2015. Disponível em: <https://proceedings.mlr.press/v37/menon15.html>. ISSN: 1938-7228.

Appendix A

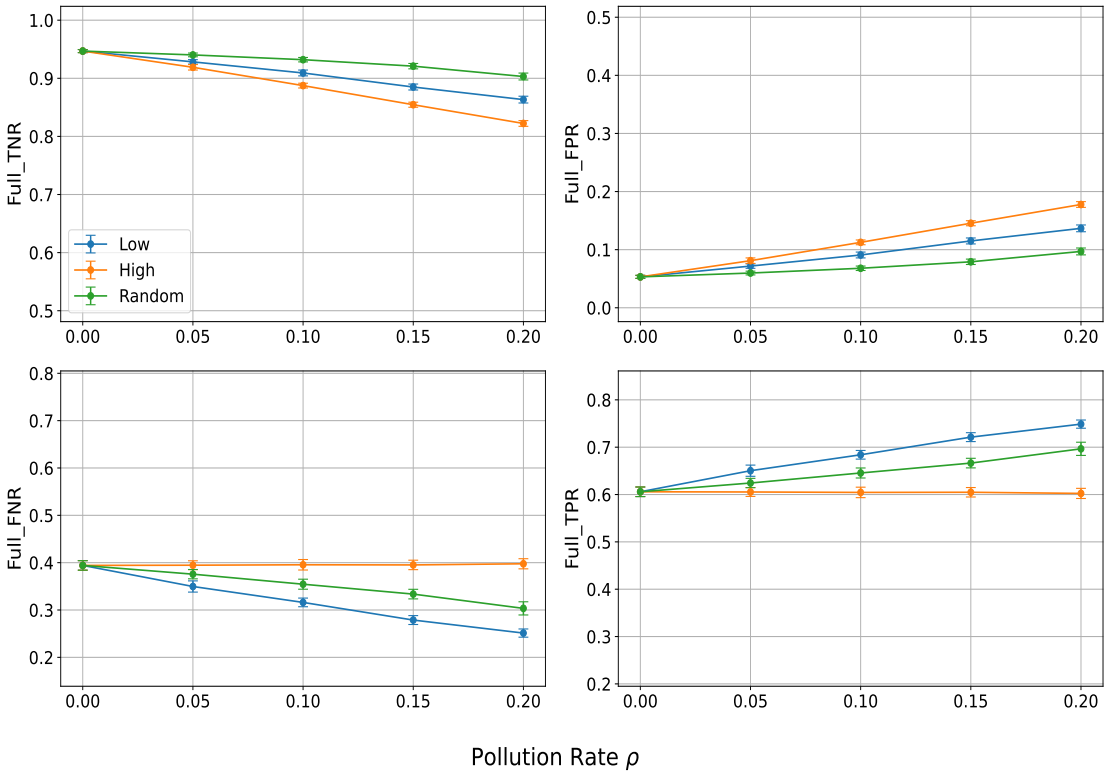
Error and Accuracy Rates under Label Pollution

This appendix presents the evolution of the confusion-matrix rates (TPR, TNR, FPR, FNR) under different levels of label pollution for all datasets and classifiers. These results complement the main text, providing detailed plots for reproducibility and further inspection.

Figure A.1: TPR, TNR, FPR and FNR for the Adult (full set).

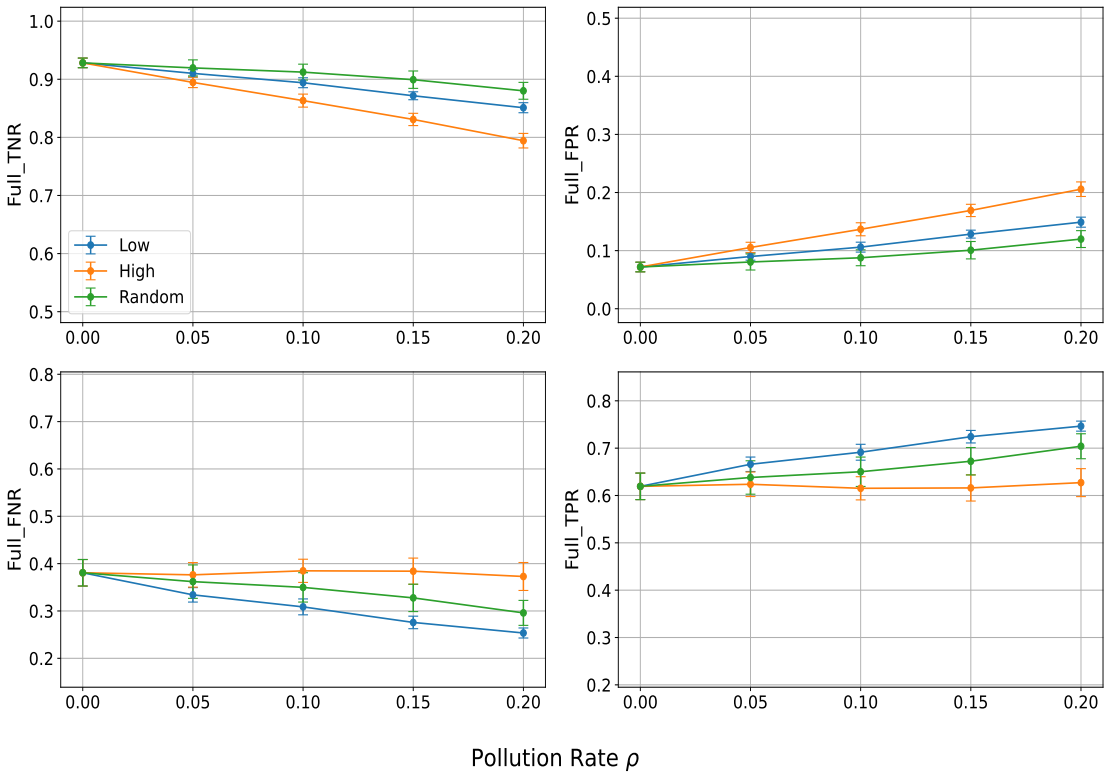


(a) Decision Tree

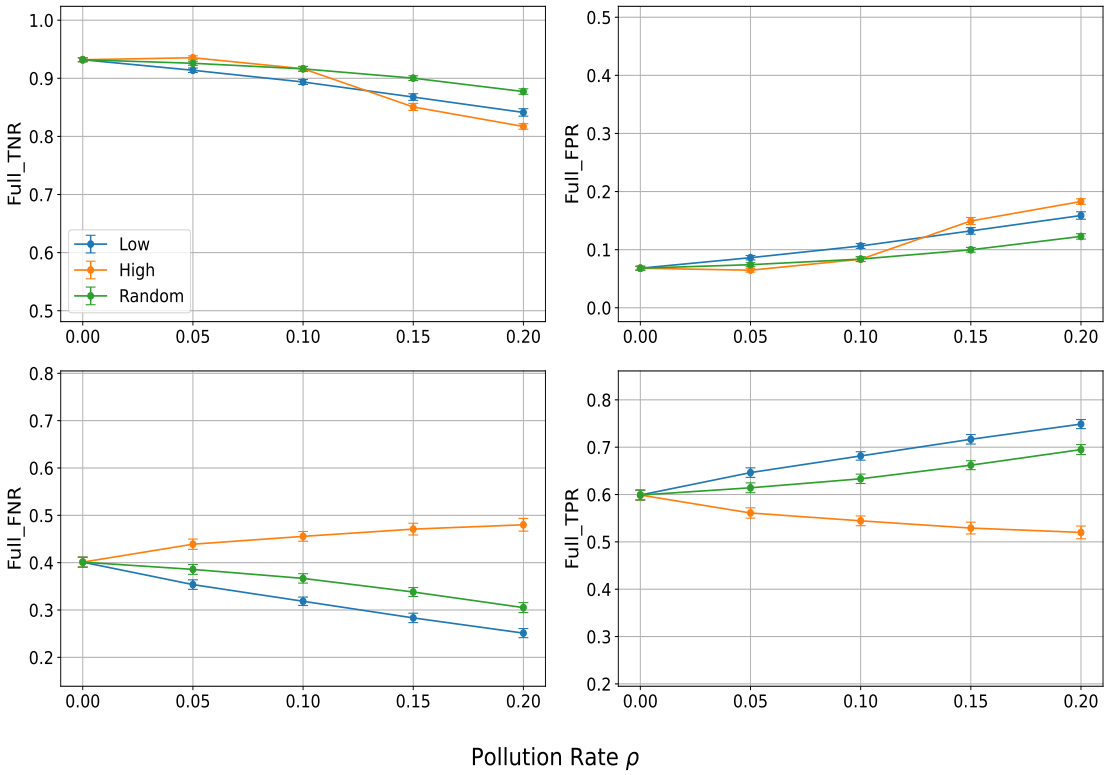


(b) Random Forest

Figure A.1: TPR, TNR, FPR and FNR for the Adult (full set).

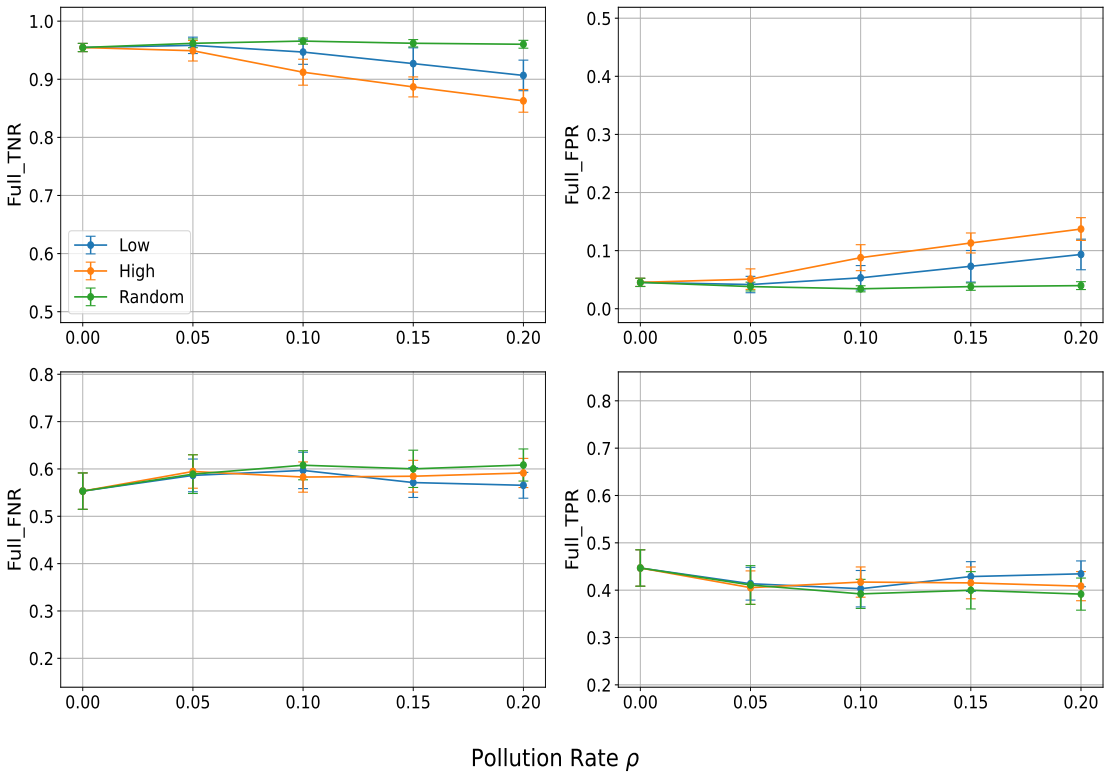


(c) Neural Network (MLP)

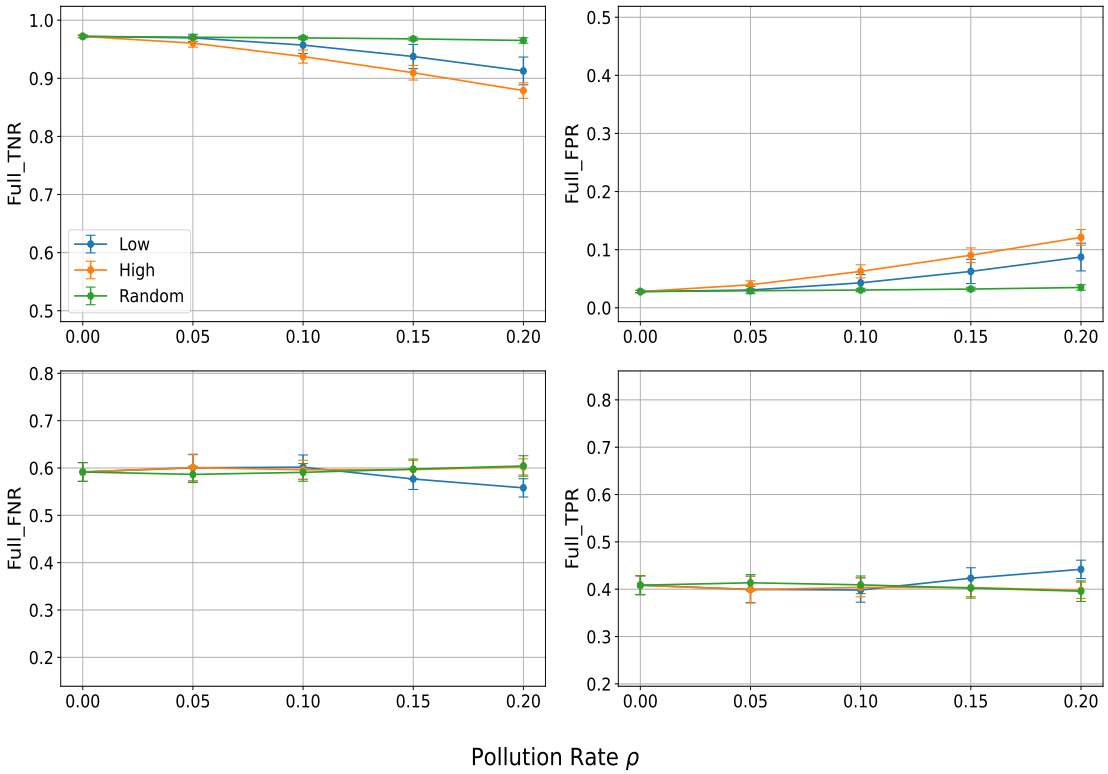


(d) Logistic Regression

Figure A.2: TPR, TNR, FPR and FNR for the Bank (full set).

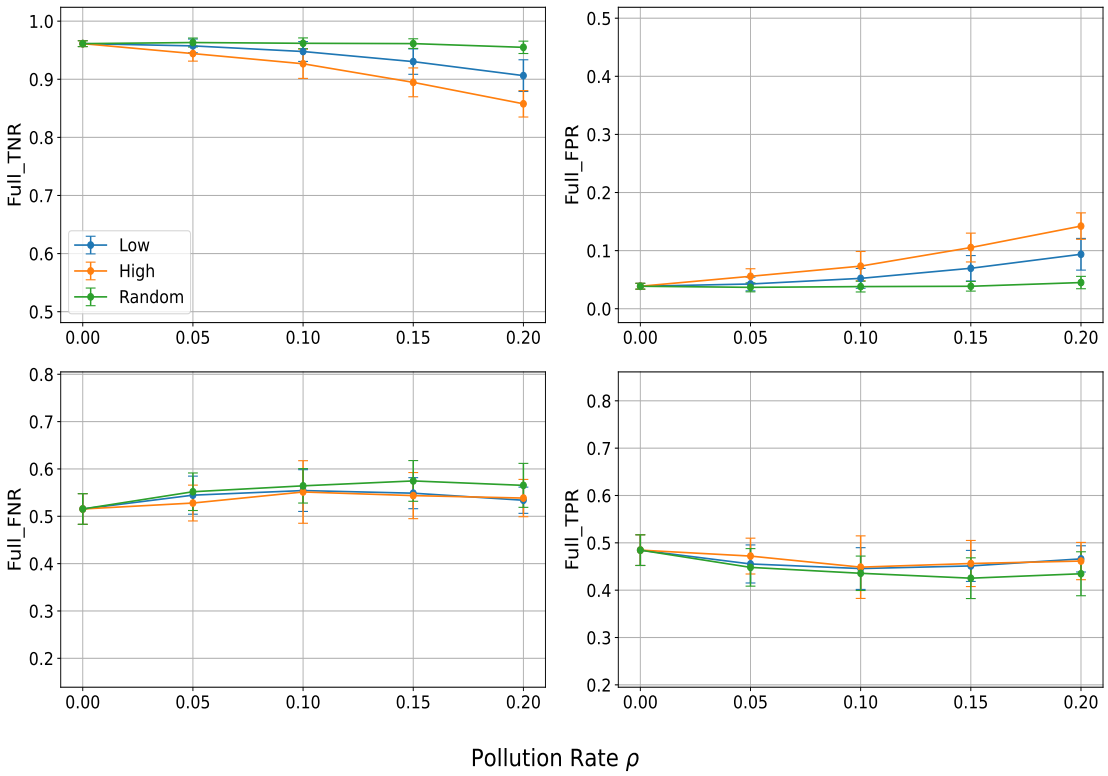


(a) Decision Tree

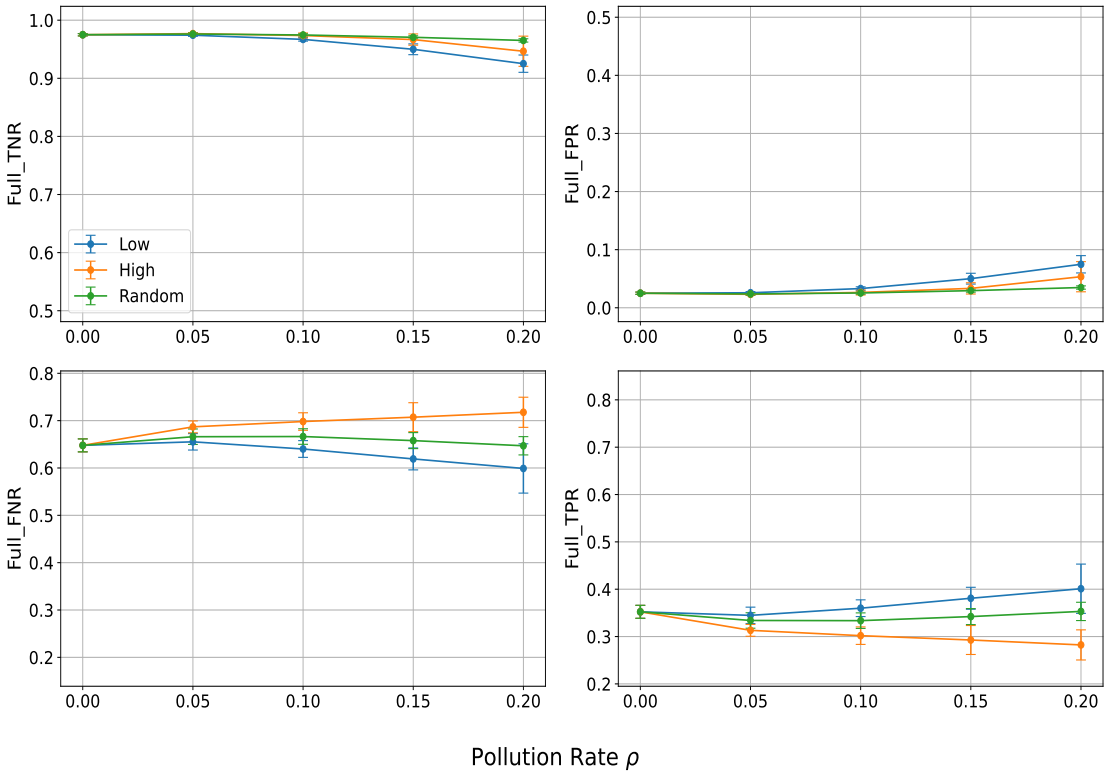


(b) Random Forest

Figure A.2: TPR, TNR, FPR and FNR for the Bank (full set).

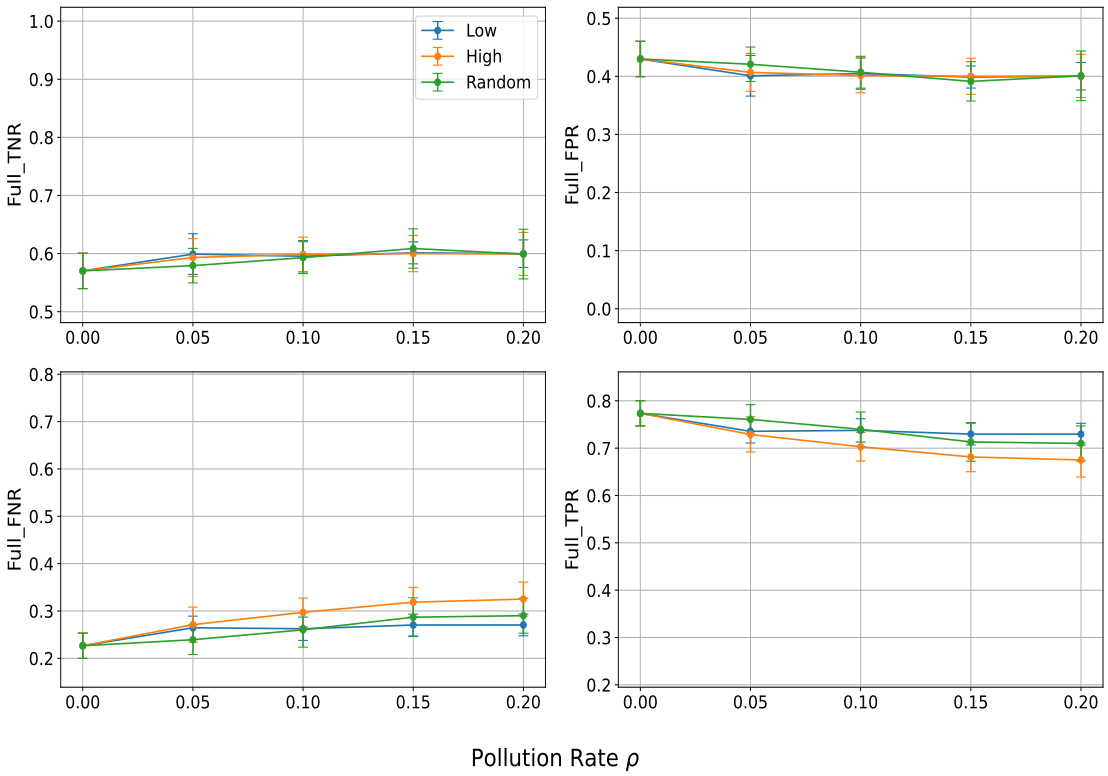


(c) Neural Network (MLP)

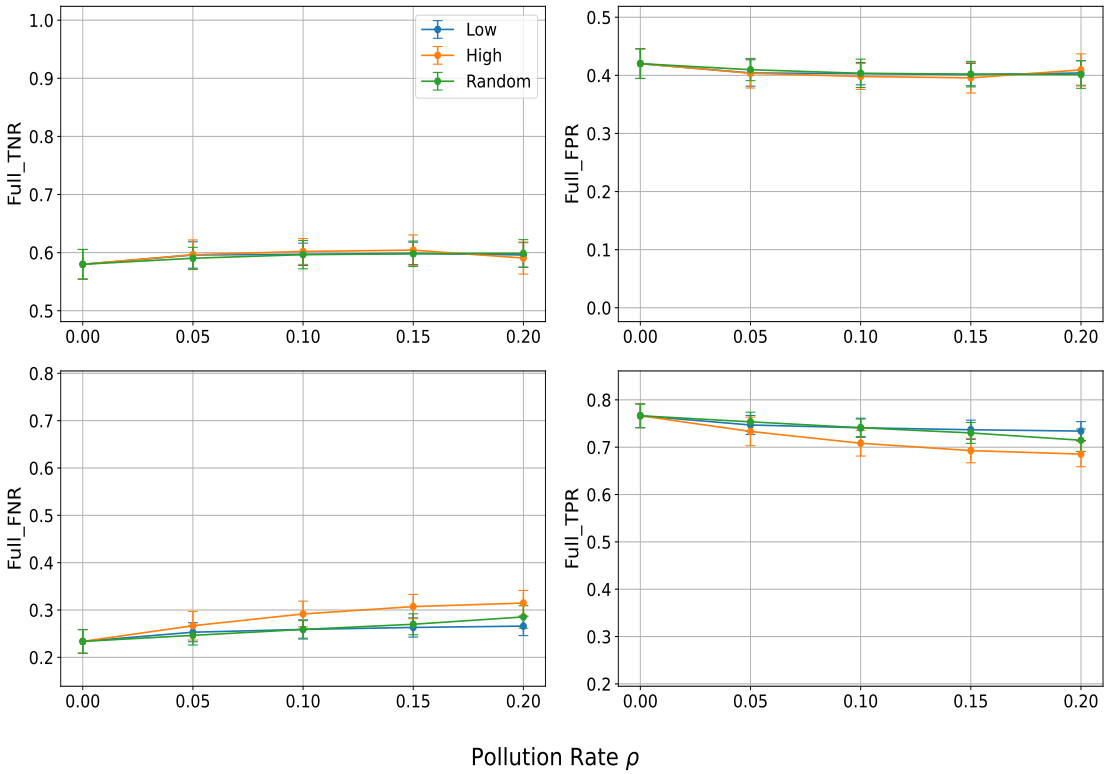


(d) Logistic Regression

Figure A.3: TPR, TNR, FPR and FNR for the COMPAS (full set).

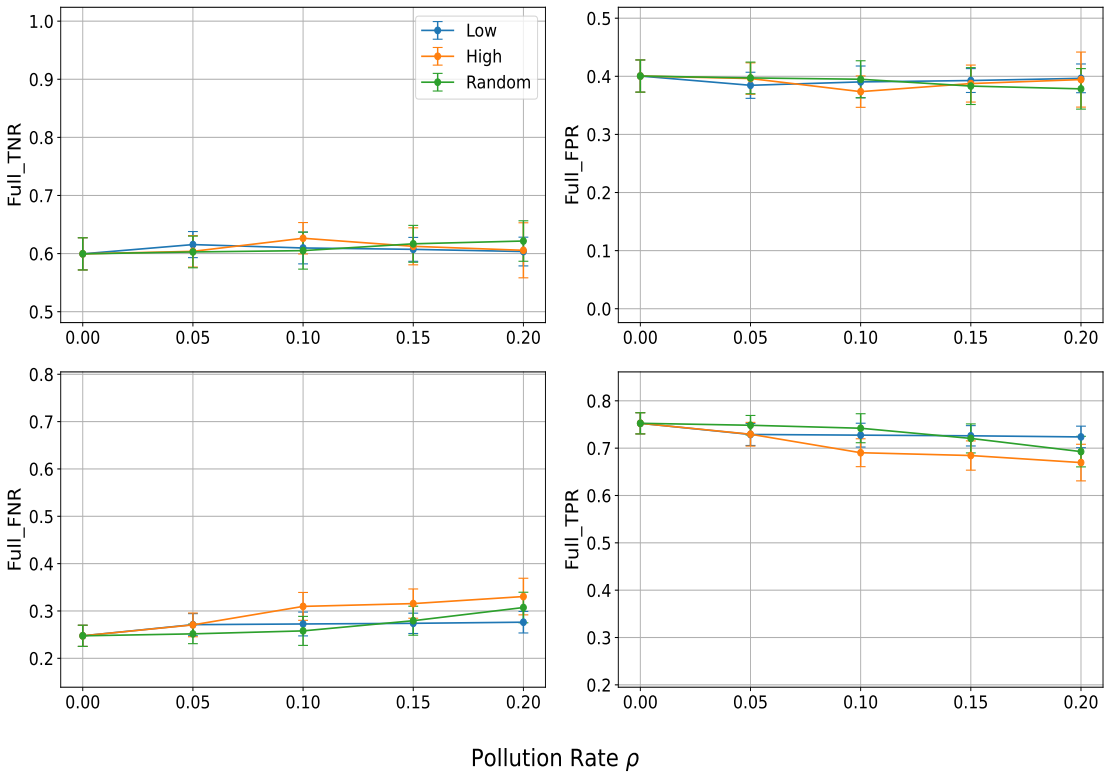


(a) Decision Tree

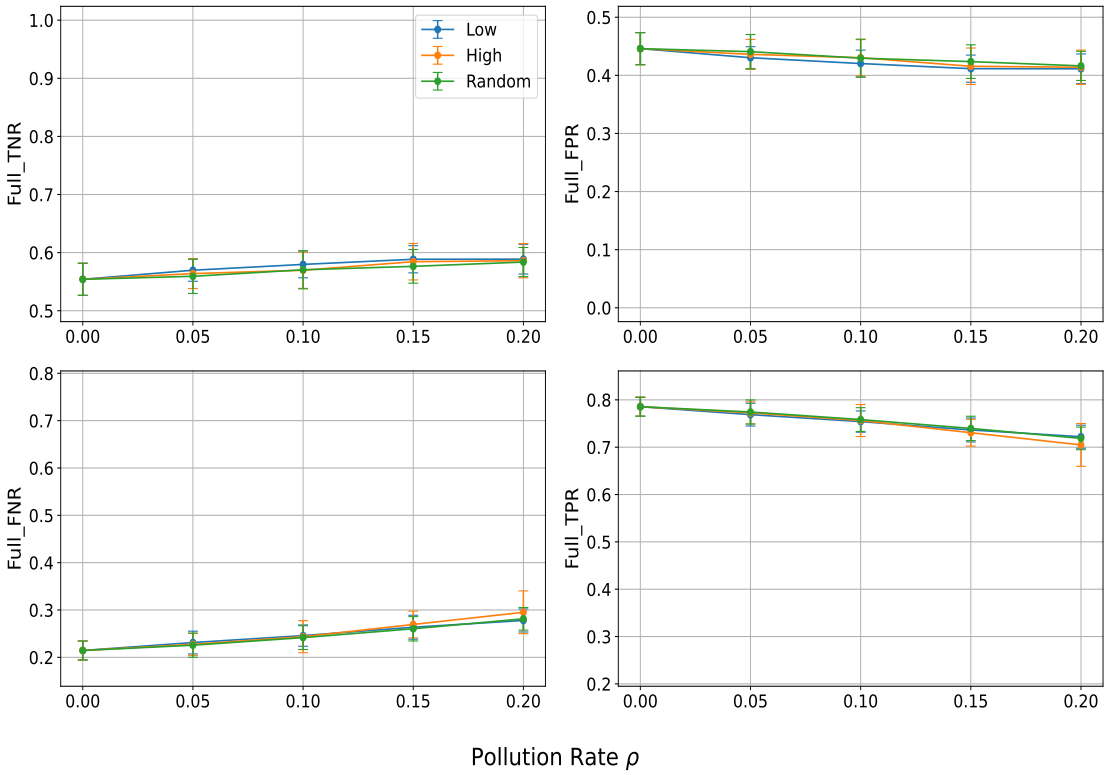


(b) Random Forest

Figure A.3: TPR, TNR, FPR and FNR for the COMPAS (full set).

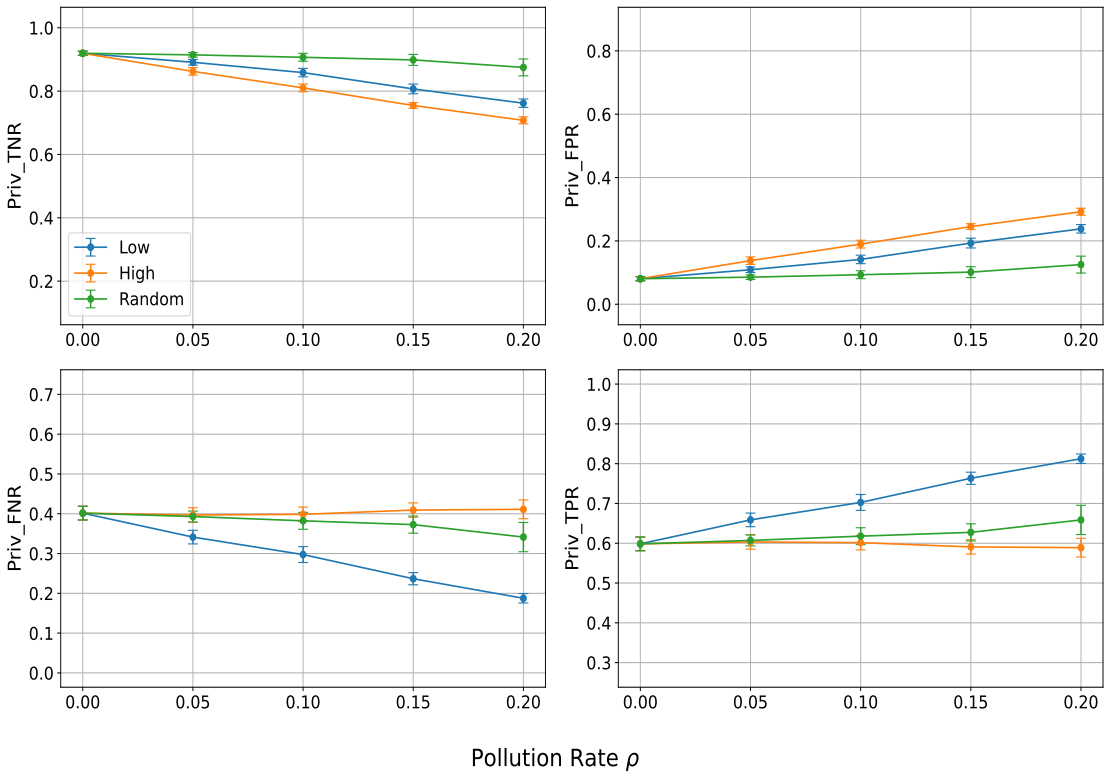


(c) Neural Network (MLP)

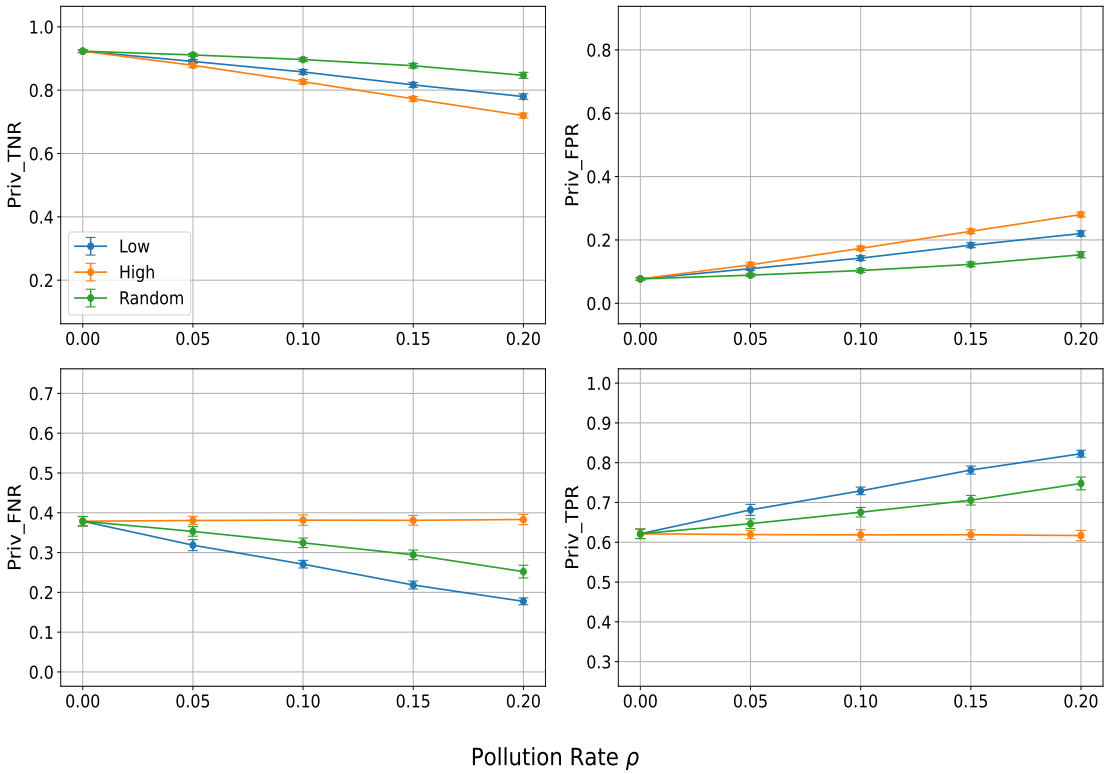


(d) Logistic Regression

Figure A.4: TPR, TNR, FPR and FNR for the Adult (privileged set).

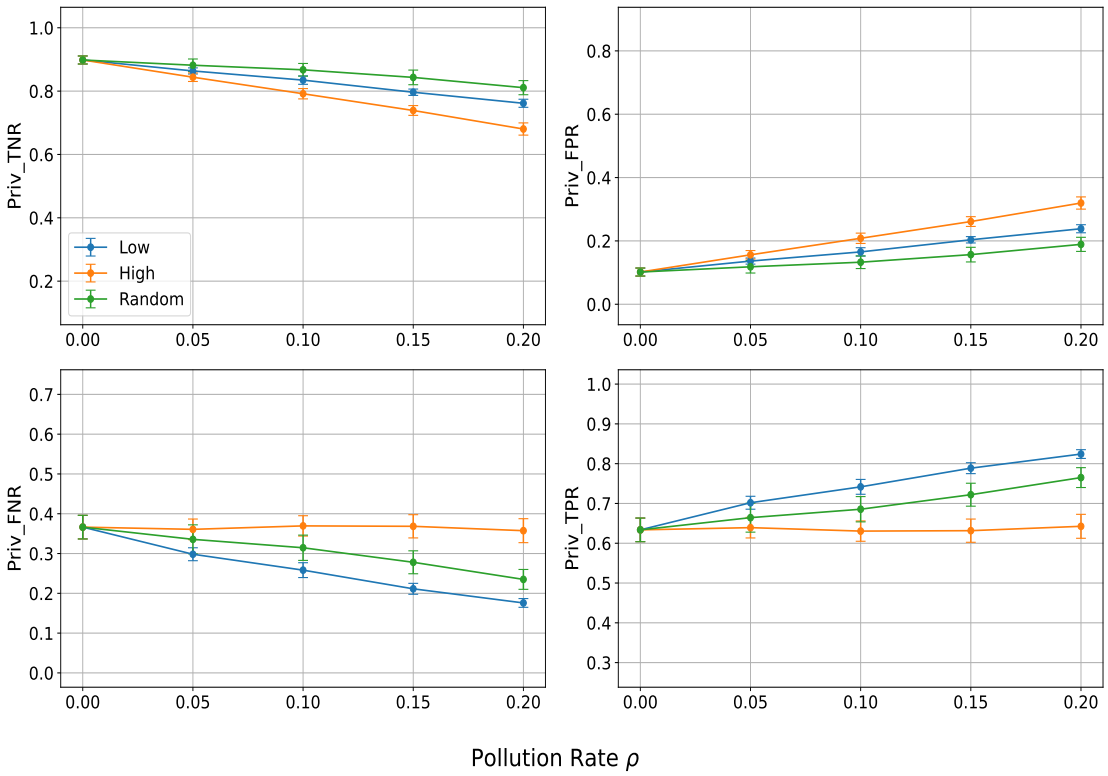


(a) Decision Tree

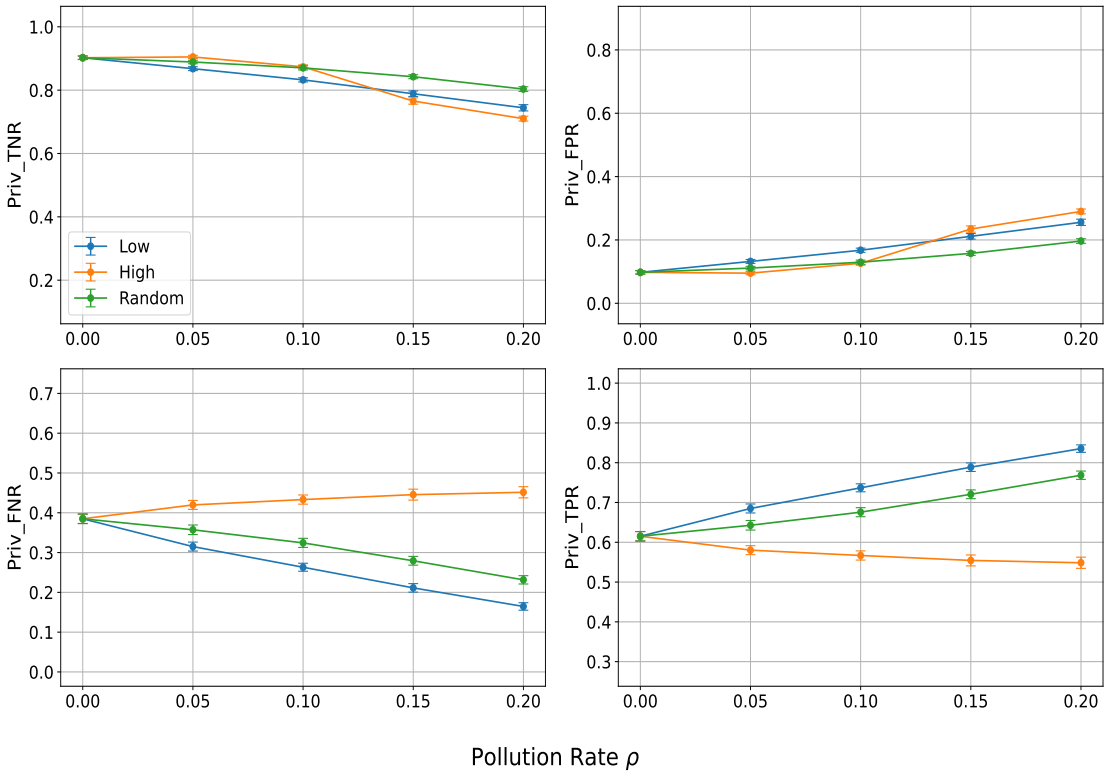


(b) Random Forest

Figure A.4: TPR, TNR, FPR and FNR for the Adult (privileged set).

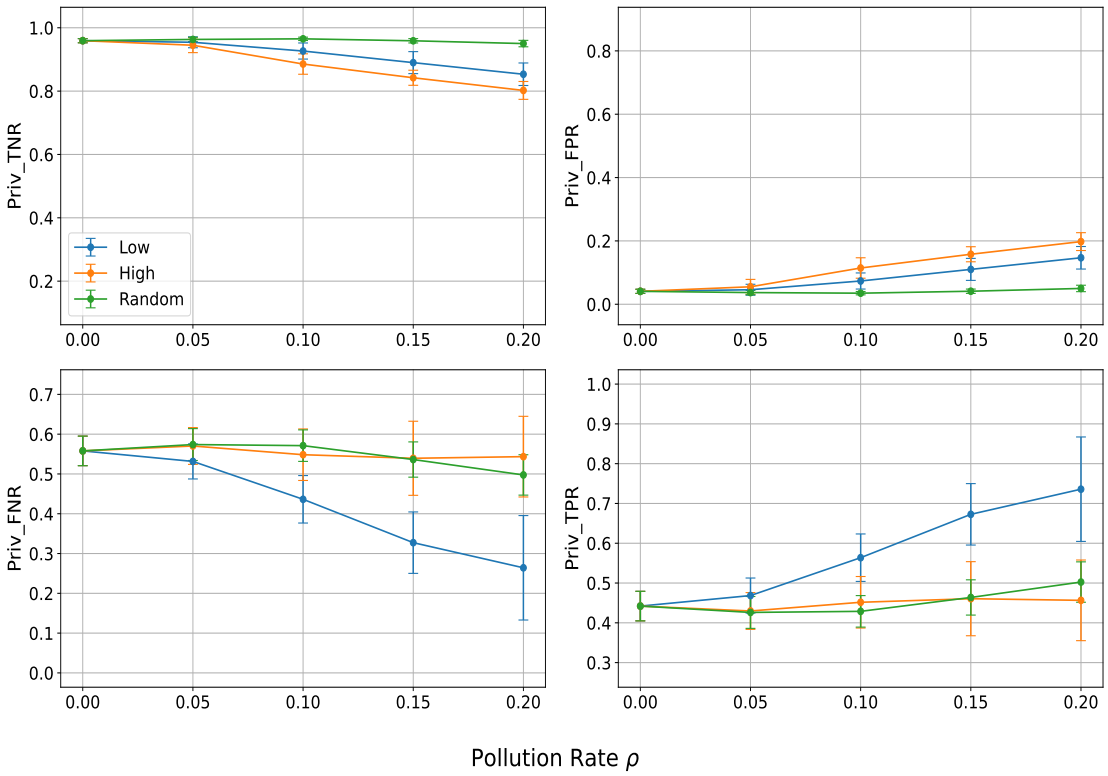


(c) Neural Network (MLP)

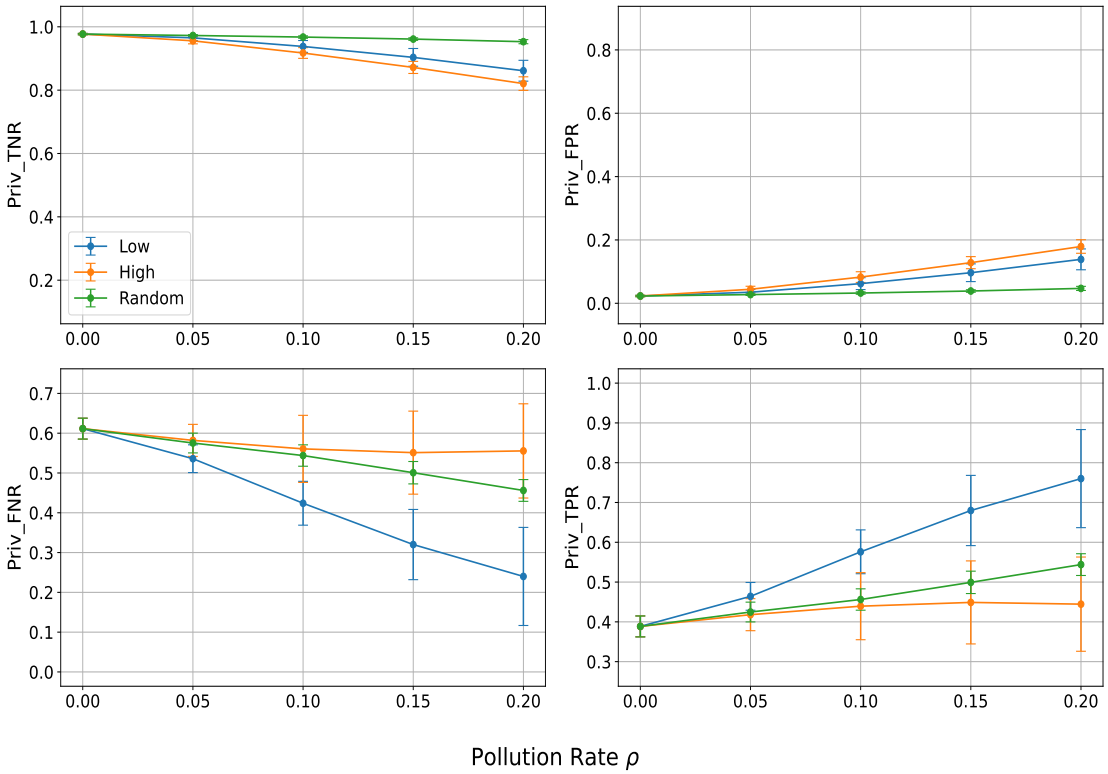


(d) Logistic Regression

Figure A.5: TPR, TNR, FPR and FNR for the Bank (privileged set).

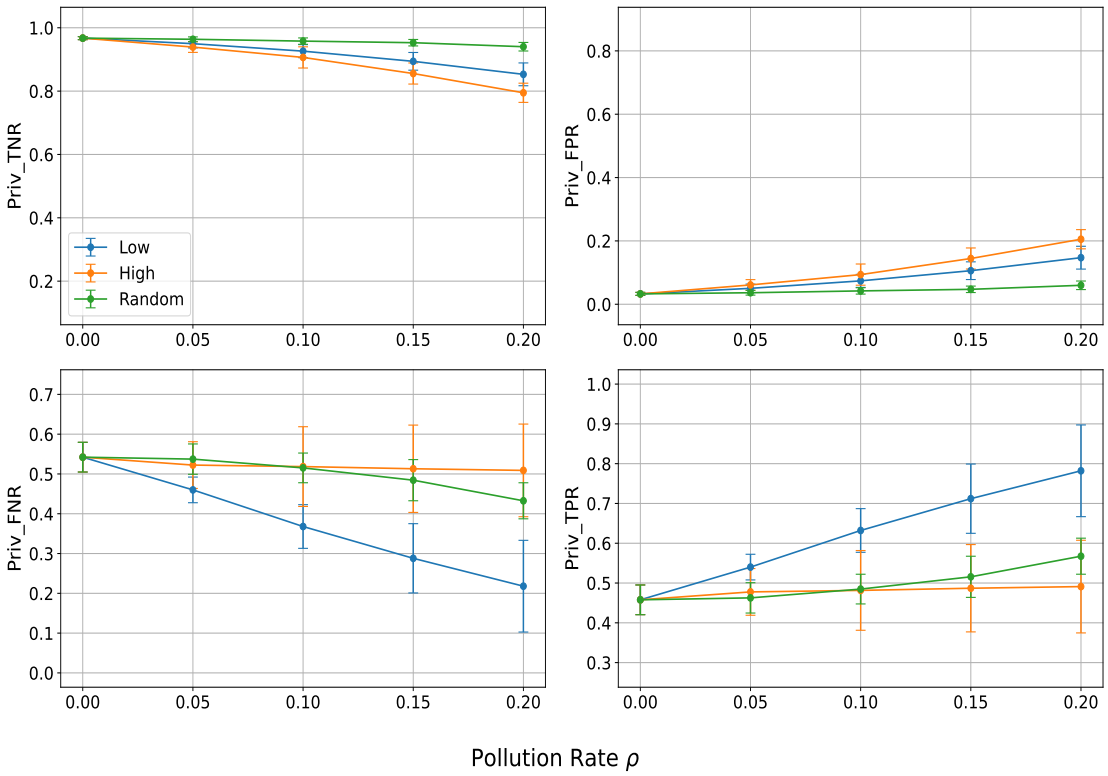


(a) Decision Tree

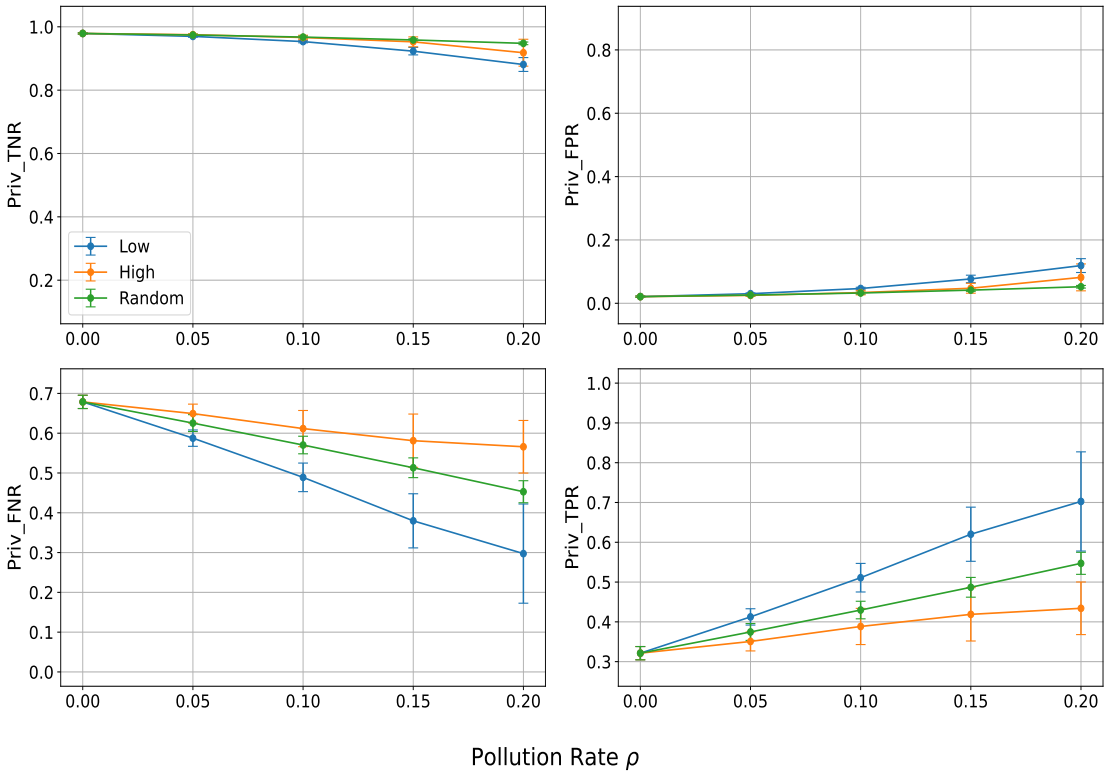


(b) Random Forest

Figure A.5: TPR, TNR, FPR and FNR for the Bank (privileged set).

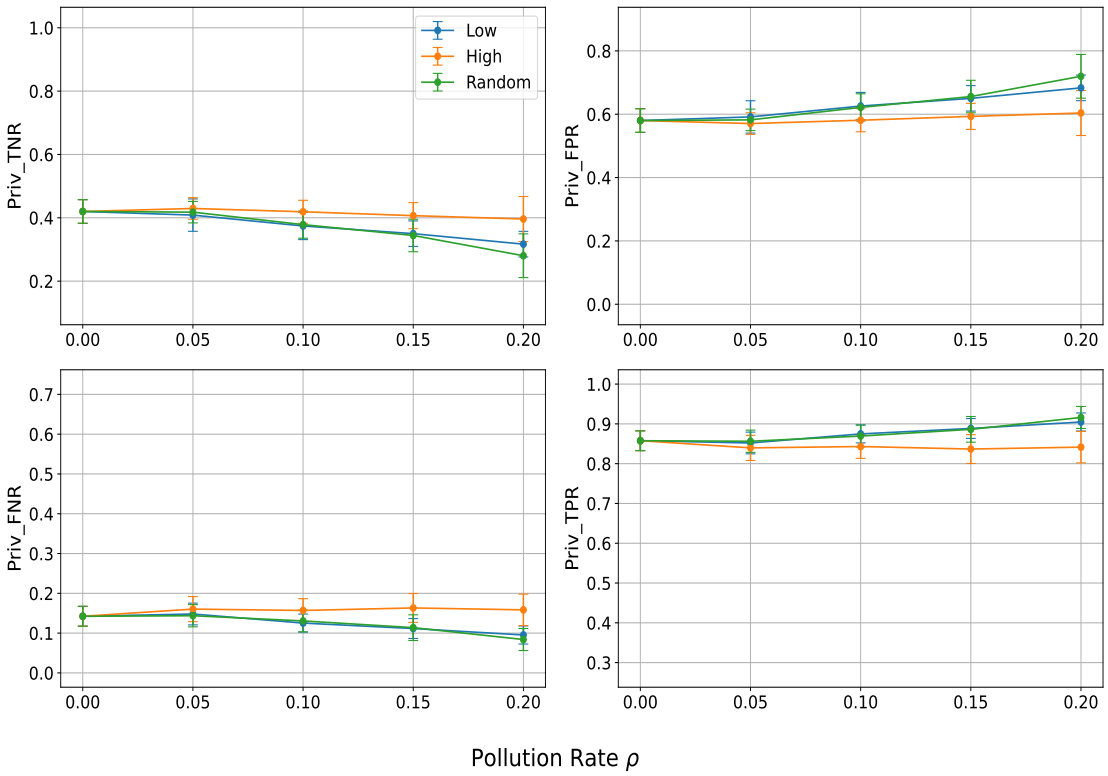


(c) Neural Network (MLP)

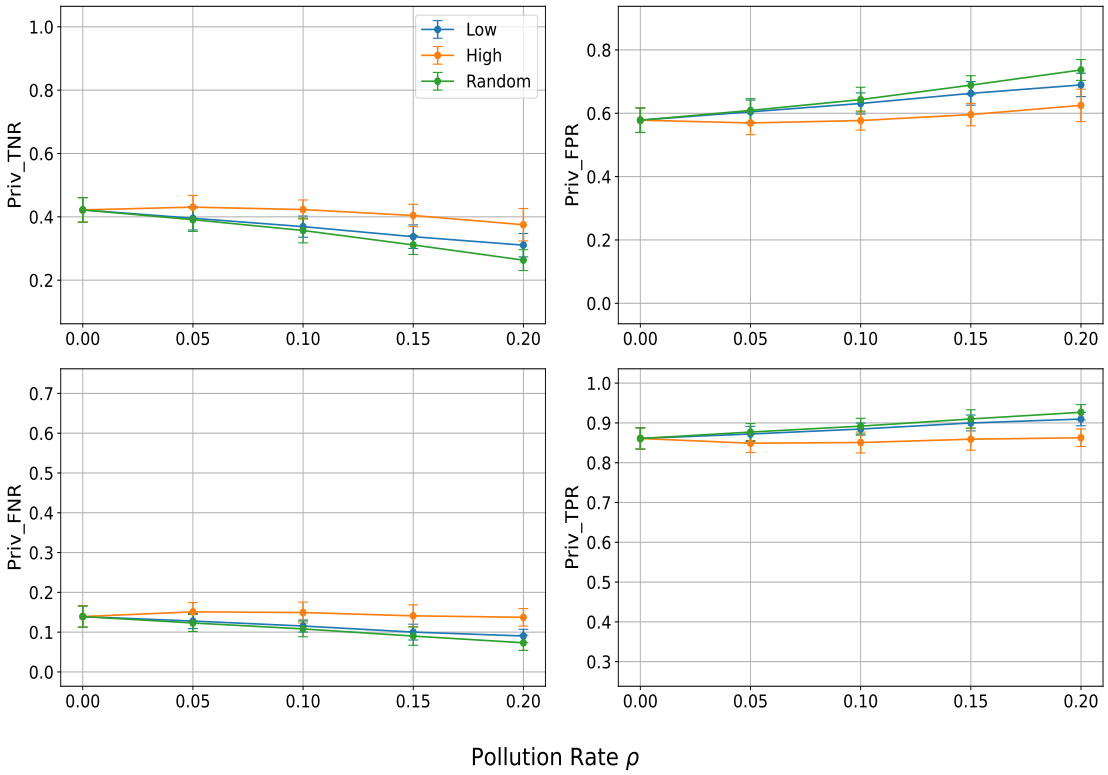


(d) Logistic Regression

Figure A.6: TPR, TNR, FPR and FNR for the COMPAS (privileged set).

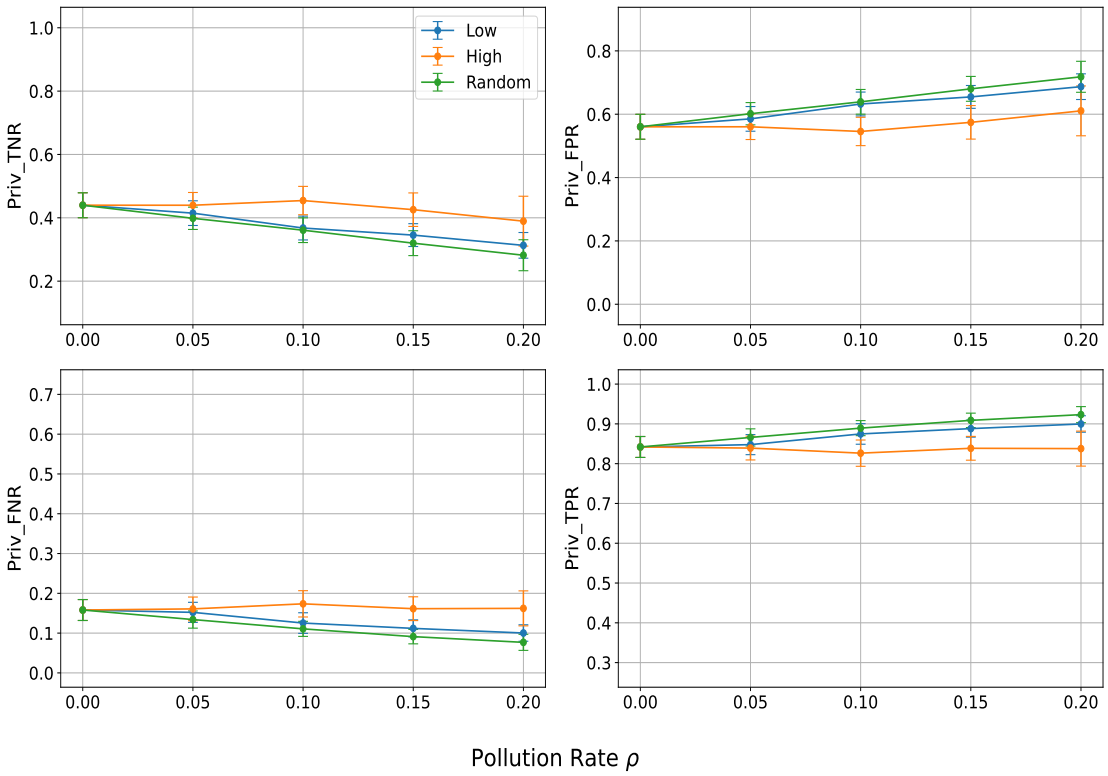


(a) Decision Tree

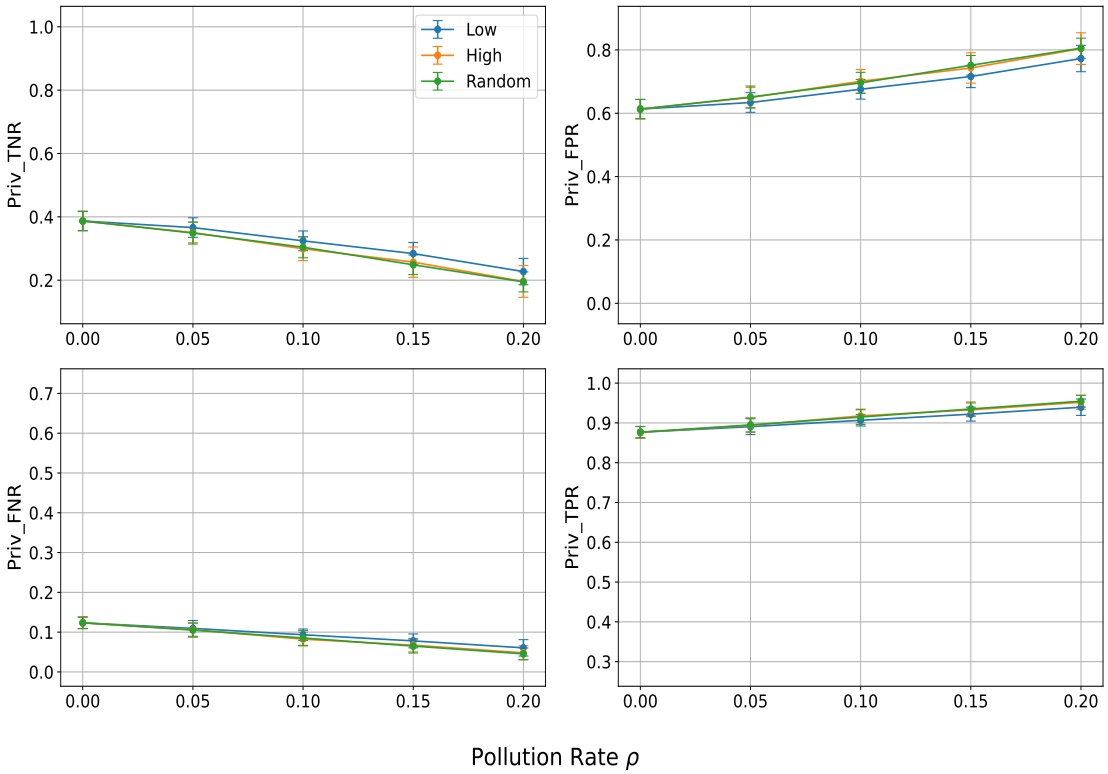


(b) Random Forest

Figure A.6: TPR, TNR, FPR and FNR for the COMPAS (privileged set).

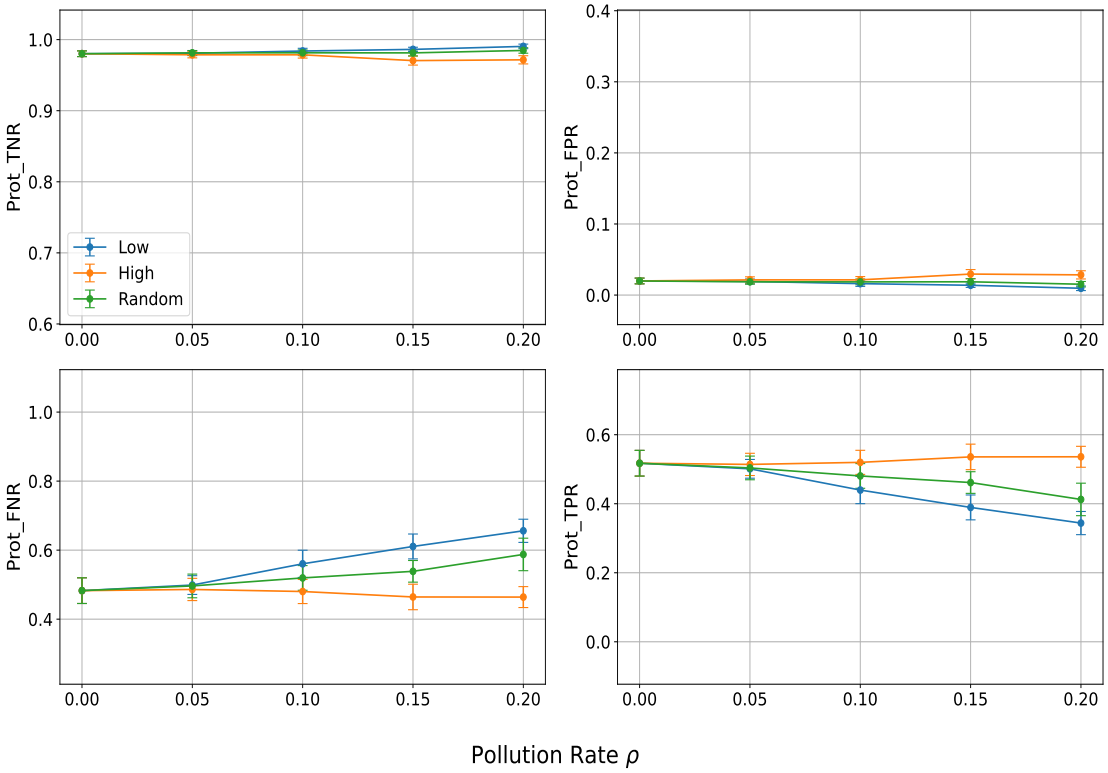


(c) Neural Network (MLP)

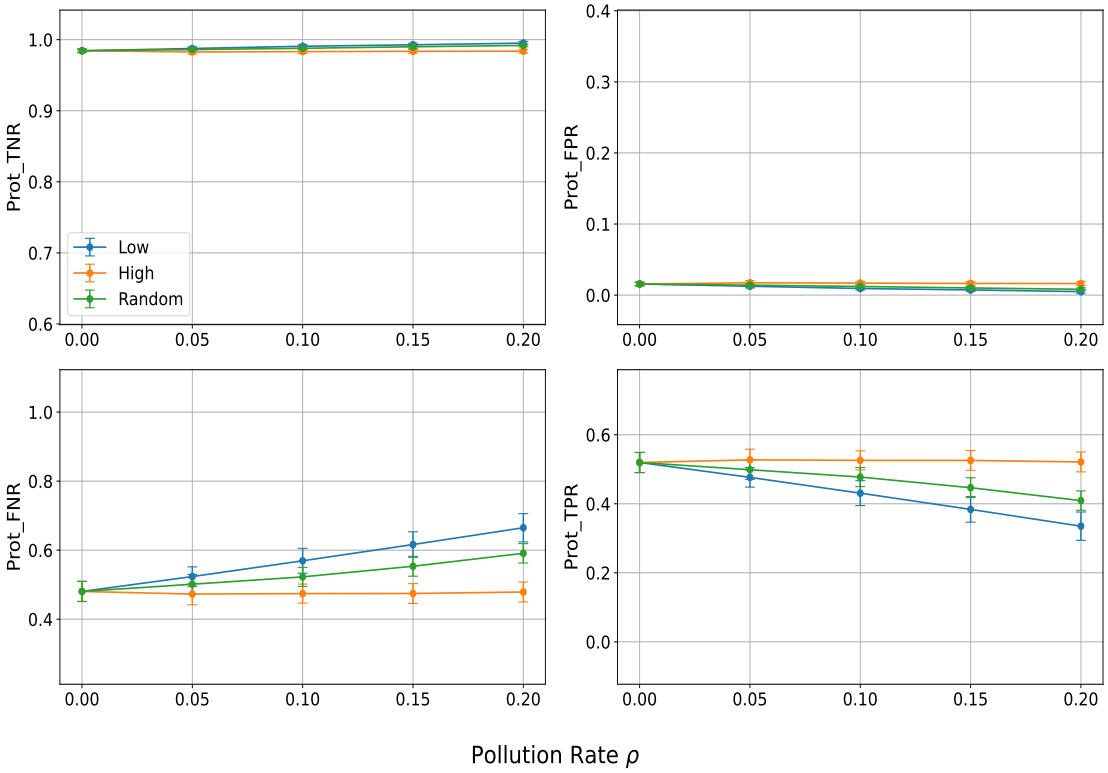


(d) Logistic Regression

Figure A.7: TPR, TNR, FPR and FNR for the Adult (protected set).

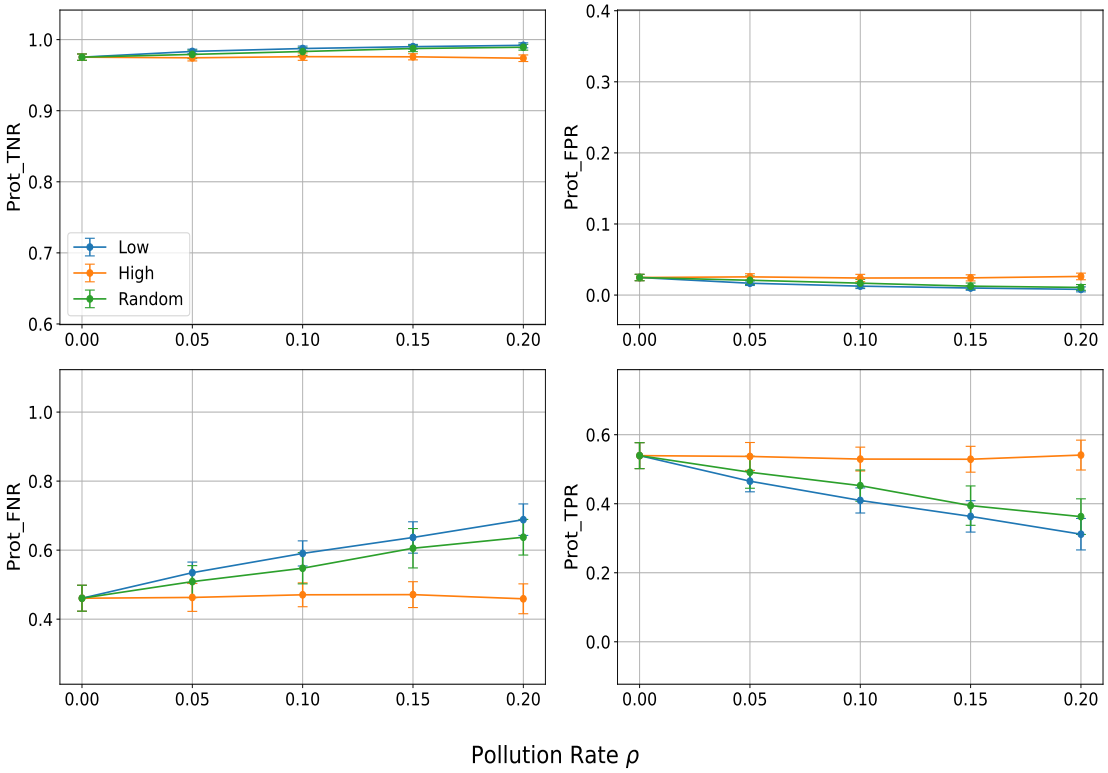


(a) Decision Tree

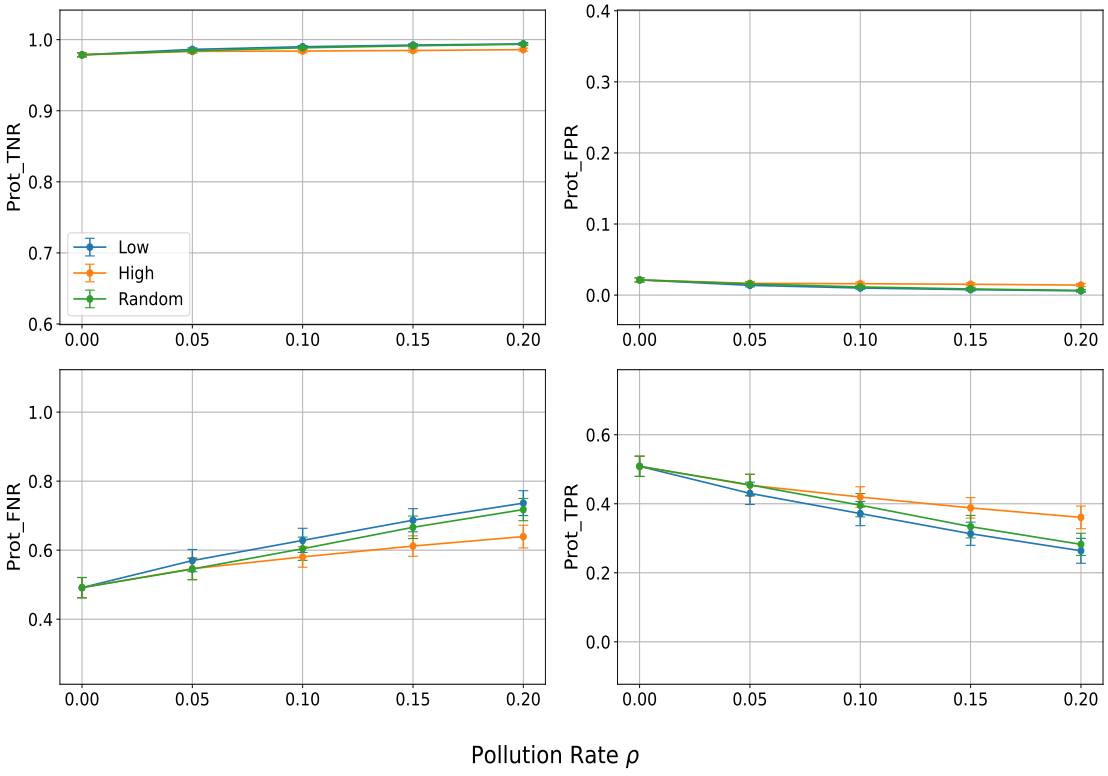


(b) Random Forest

Figure A.7: TPR, TNR, FPR and FNR for the Adult (protected set).

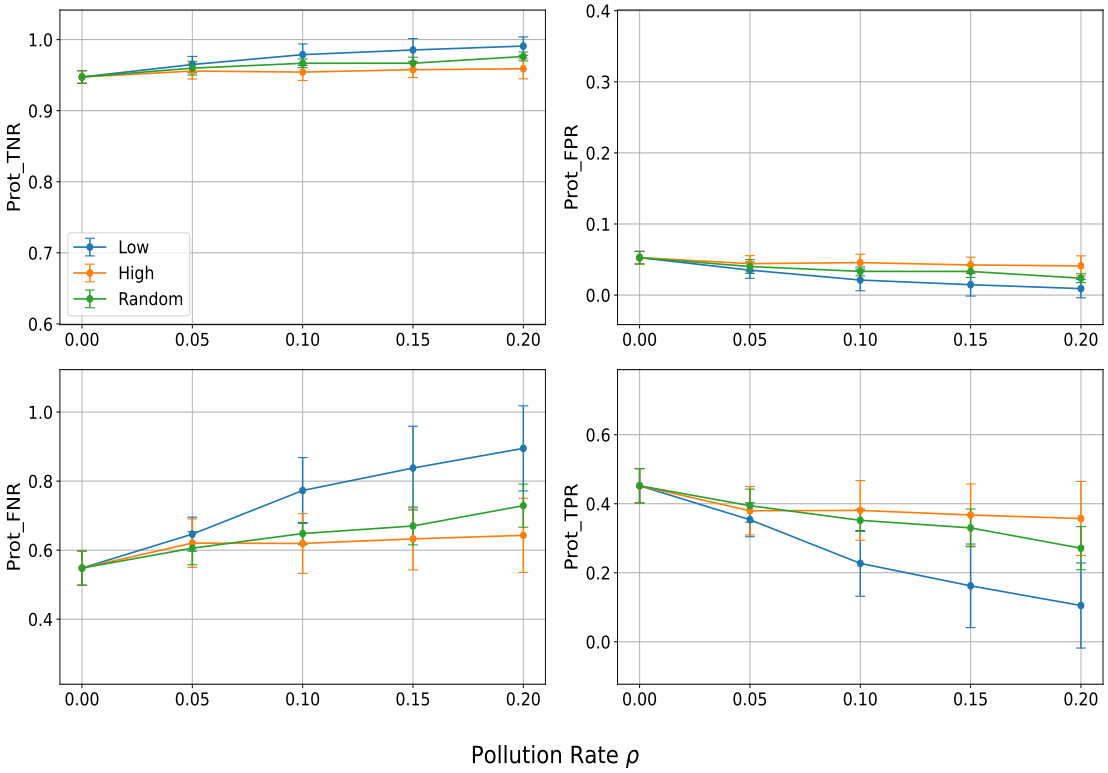


(c) Neural Network (MLP)

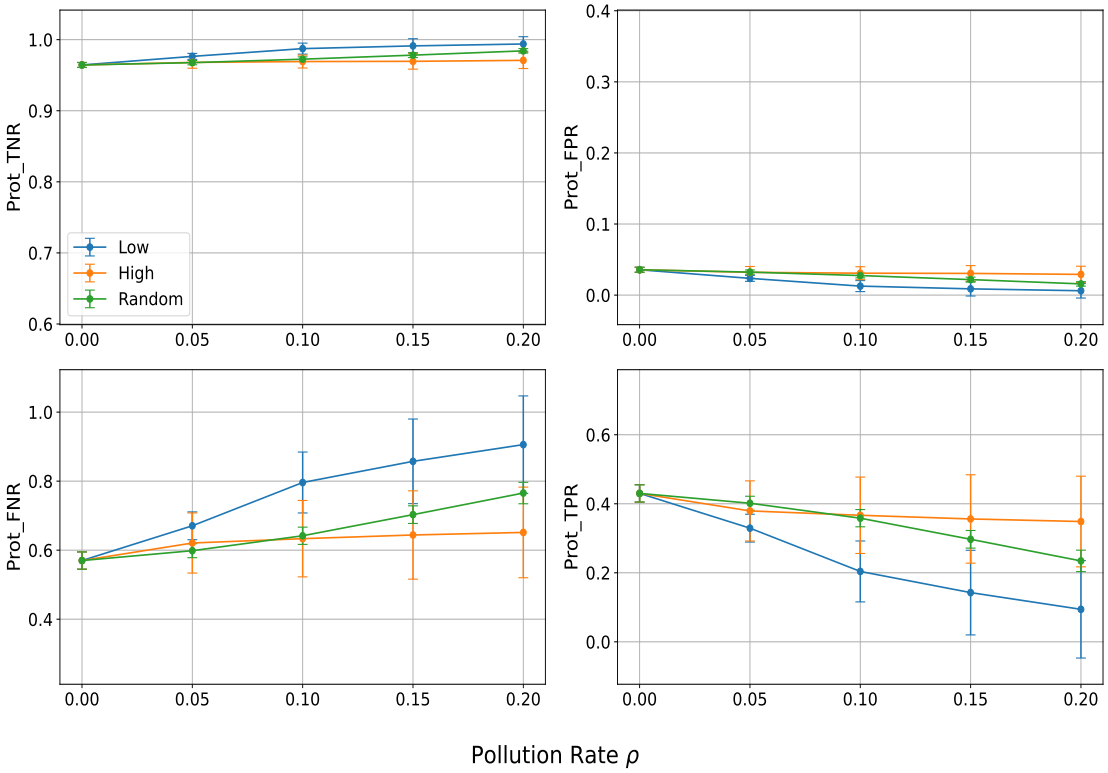


(d) Logistic Regression

Figure A.8: TPR, TNR, FPR and FNR for the Bank (protected set).

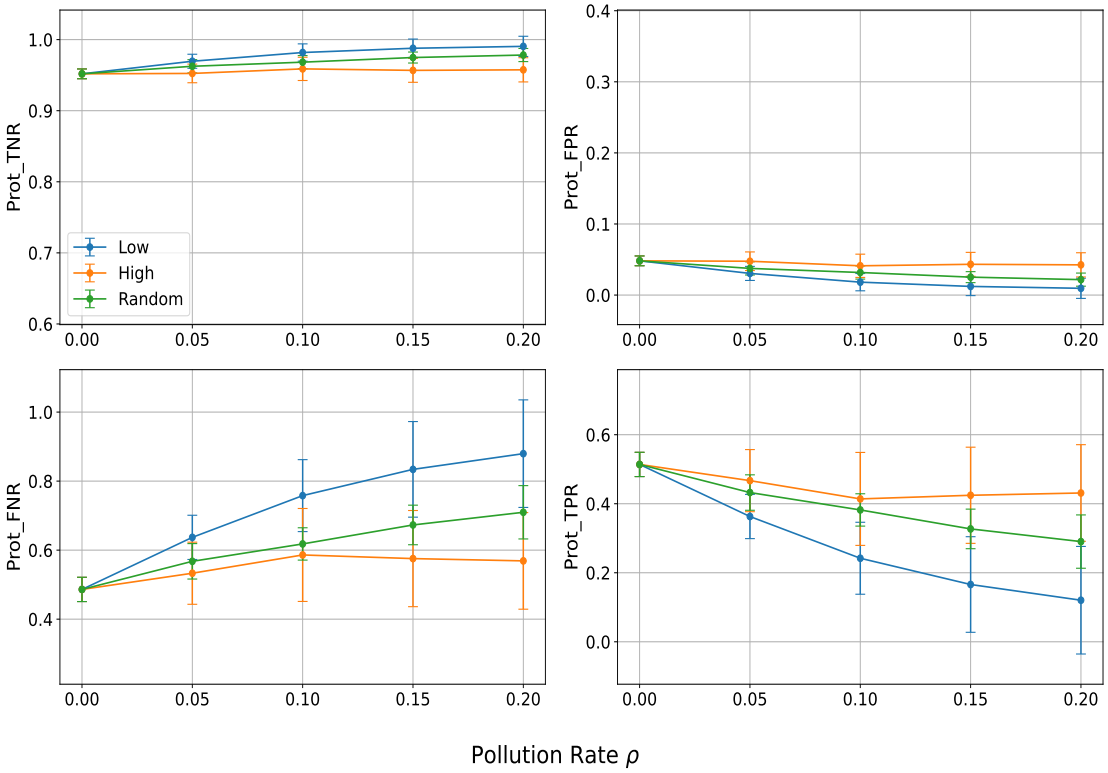


(a) Decision Tree

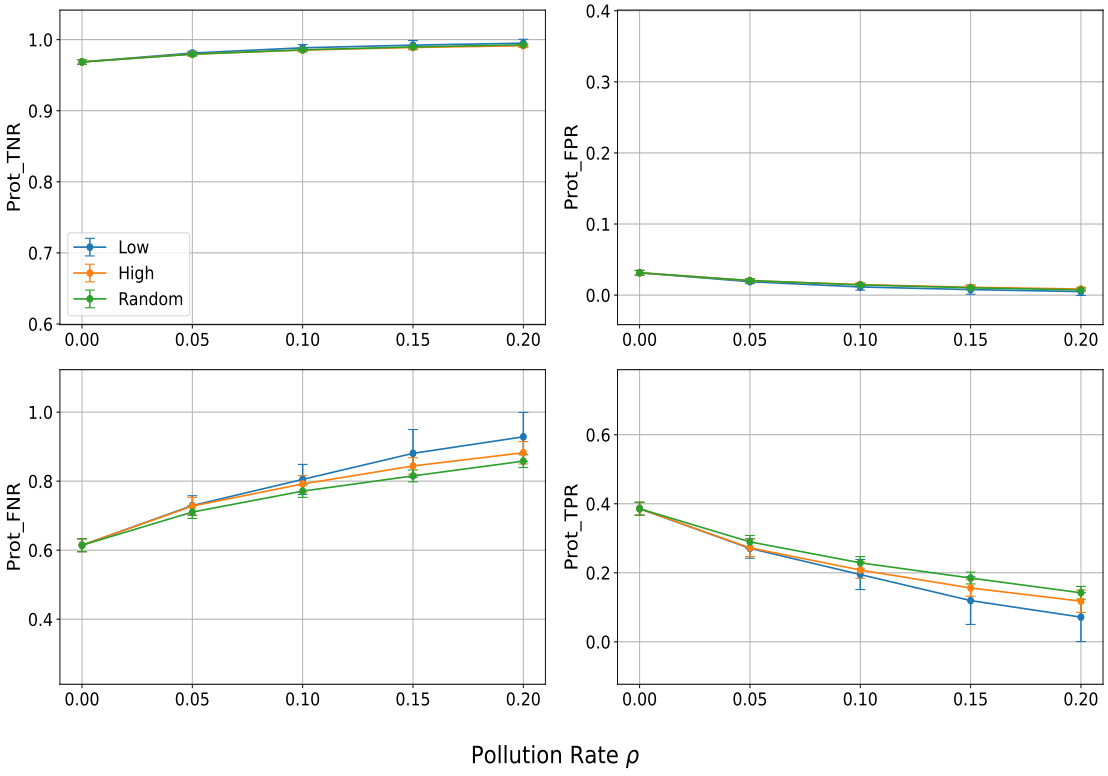


(b) Random Forest

Figure A.8: TPR, TNR, FPR and FNR for the Bank (protected set).

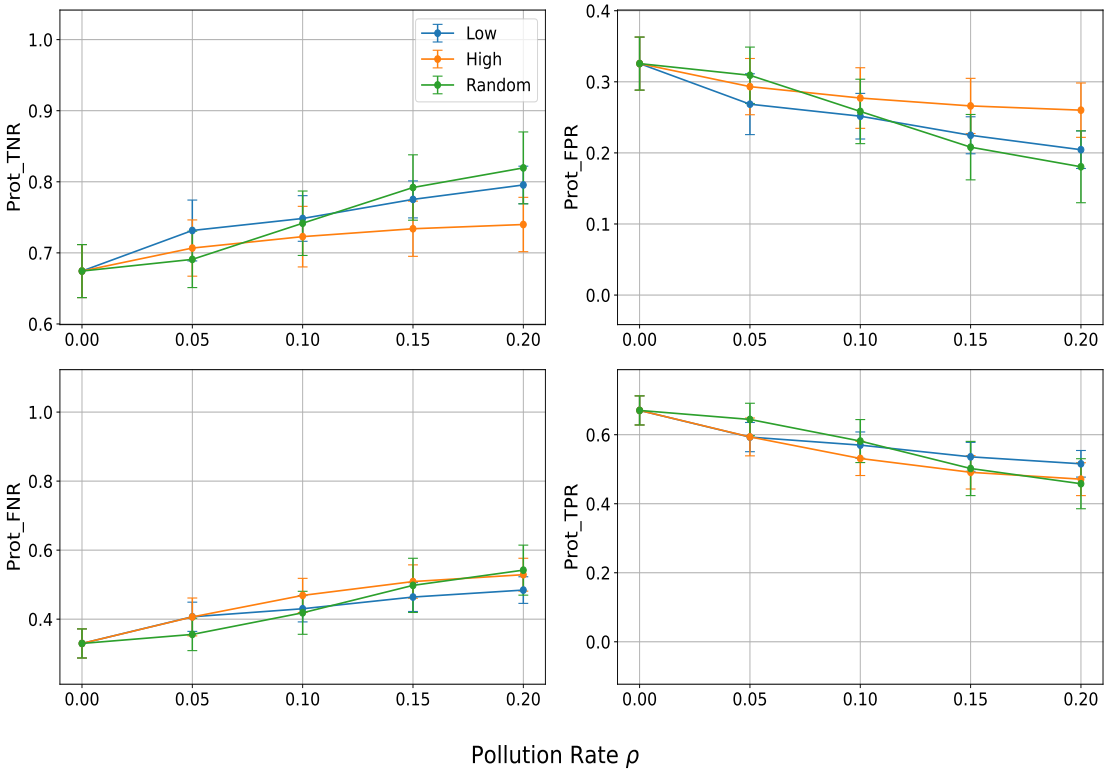


(c) Neural Network (MLP)

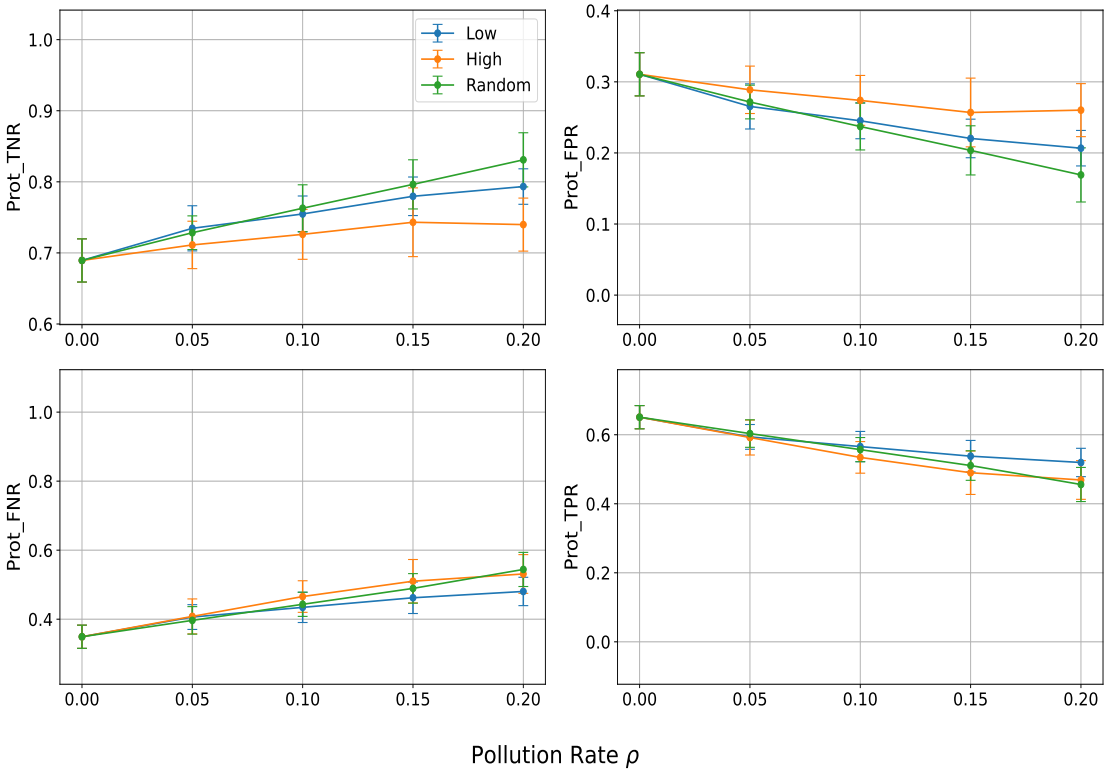


(d) Logistic Regression

Figure A.9: TPR, TNR, FPR and FNR for the COMPAS (protected set).

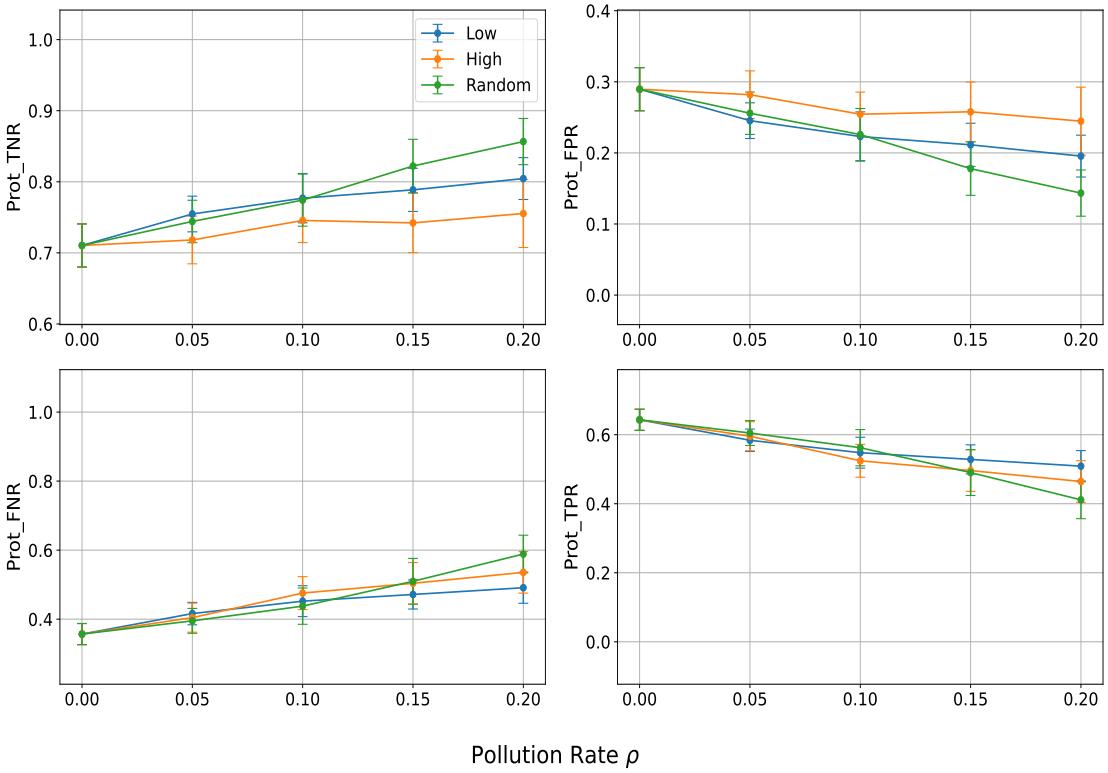


(a) Decision Tree

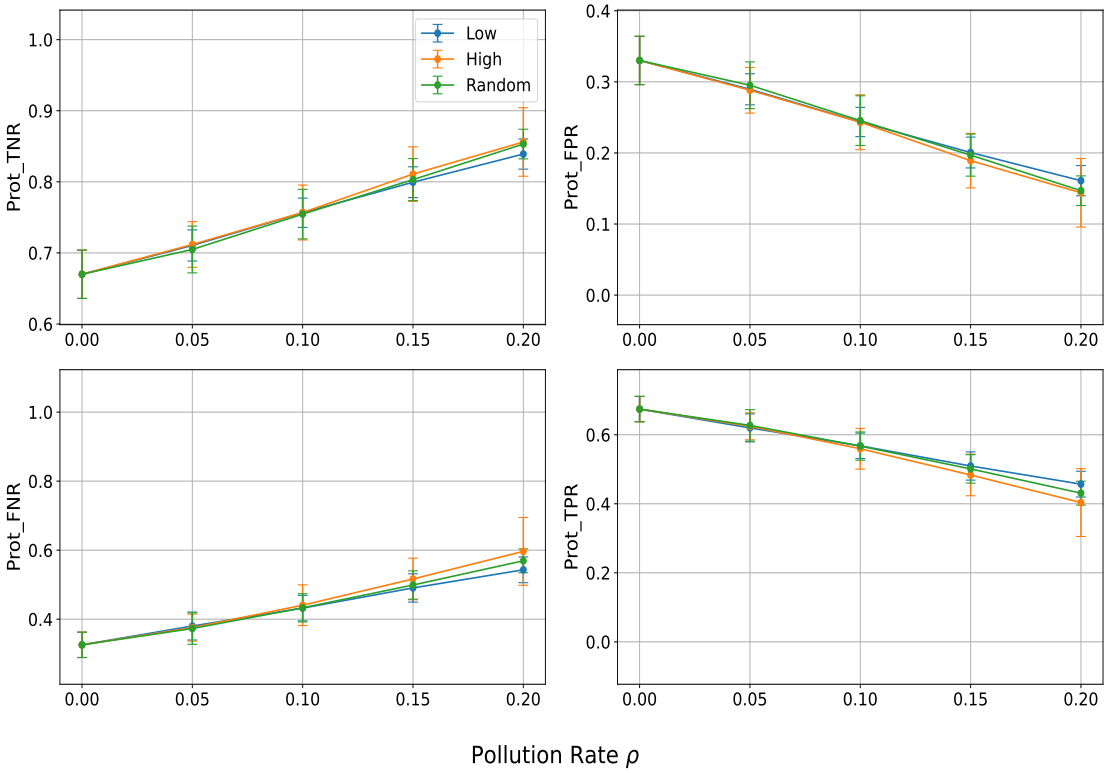


(b) Random Forest

Figure A.9: TPR, TNR, FPR and FNR for the COMPAS (protected set).

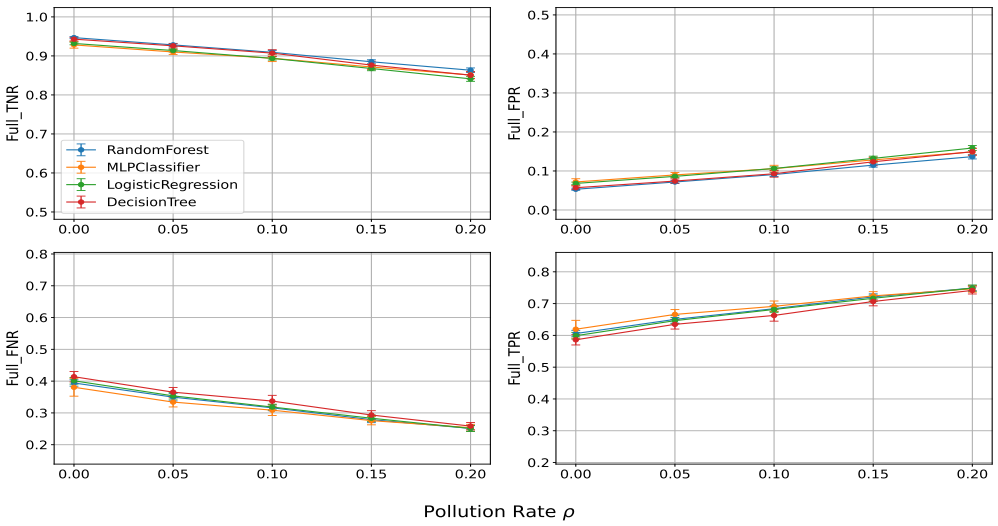


(c) Neural Network (MLP)

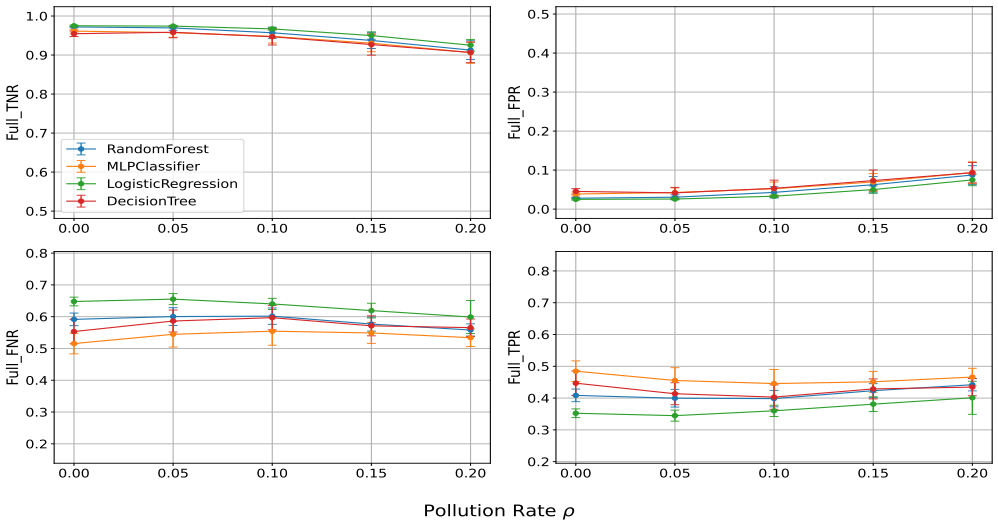


(d) Logistic Regression

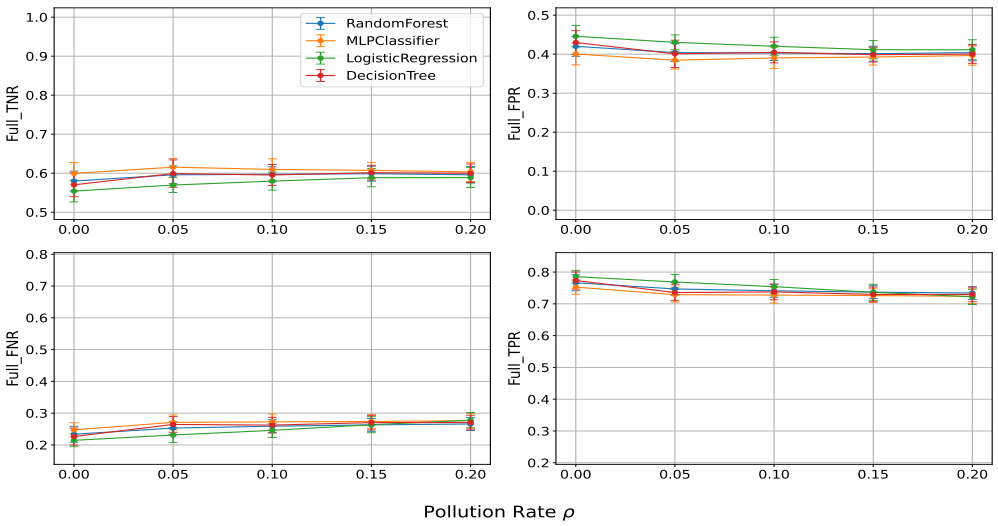
Figure A.10: TPR, TNR, FPR and FNR for the LOW strategy (Full Datasets).



(a) Adult

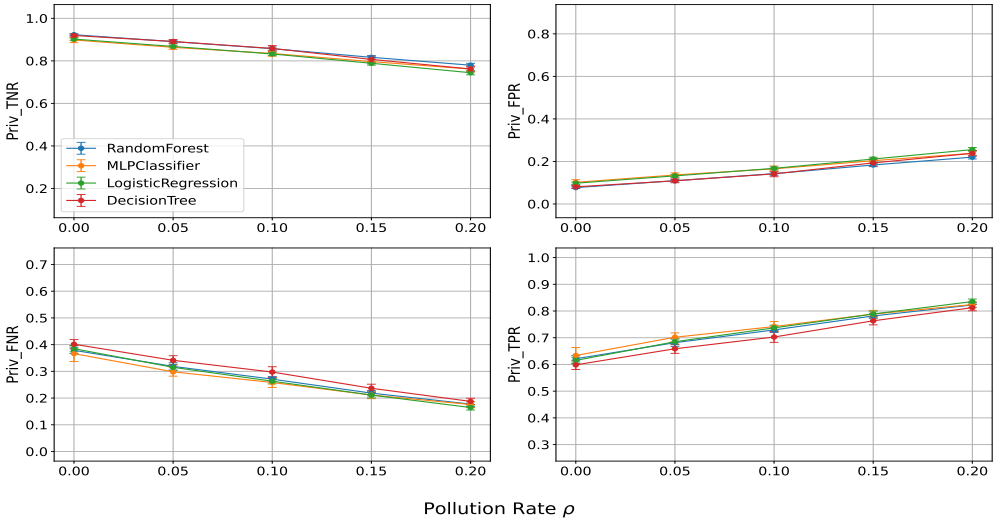


(b) Bank

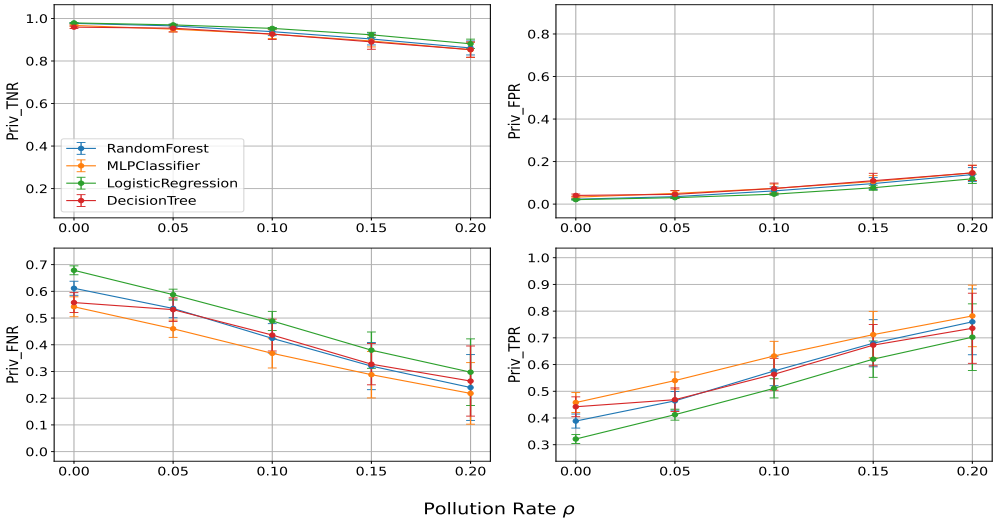


(c) COMPAS

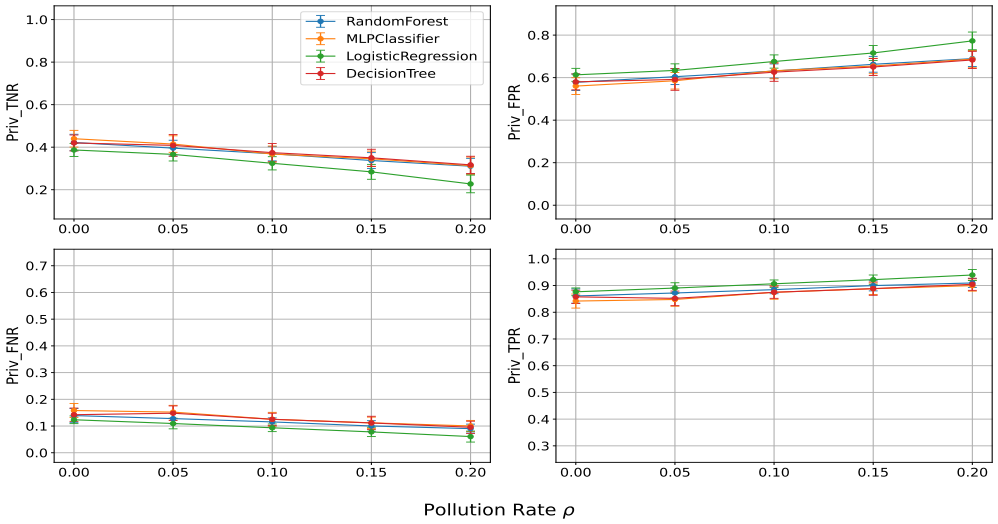
Figure A.11: TPR, TNR, FPR and FNR for the LOW strategy (Privileged Subsets).



(a) Adult

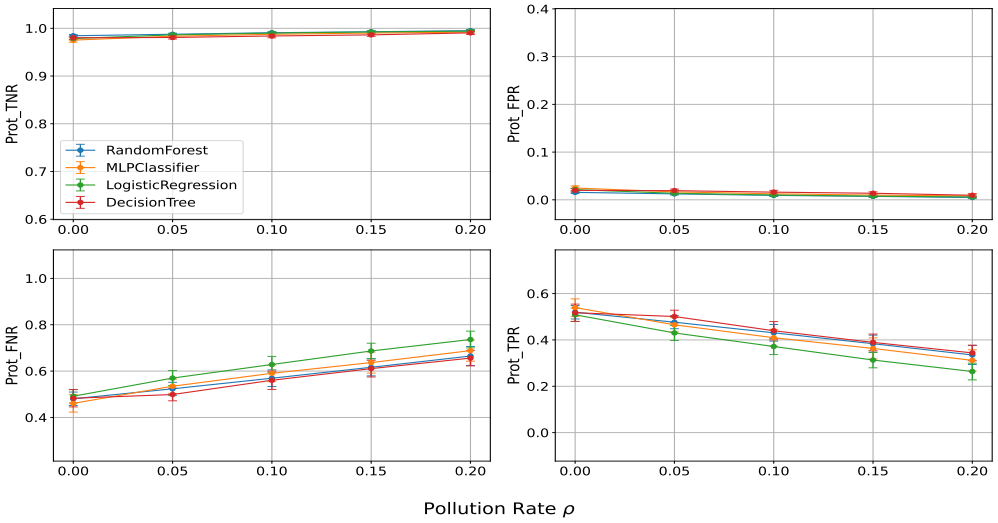


(b) Bank

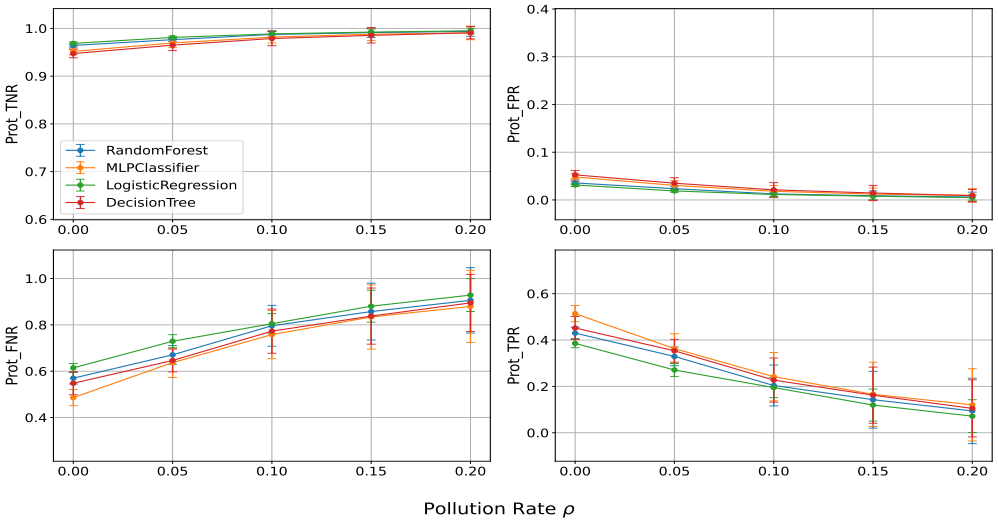


(c) COMPAS

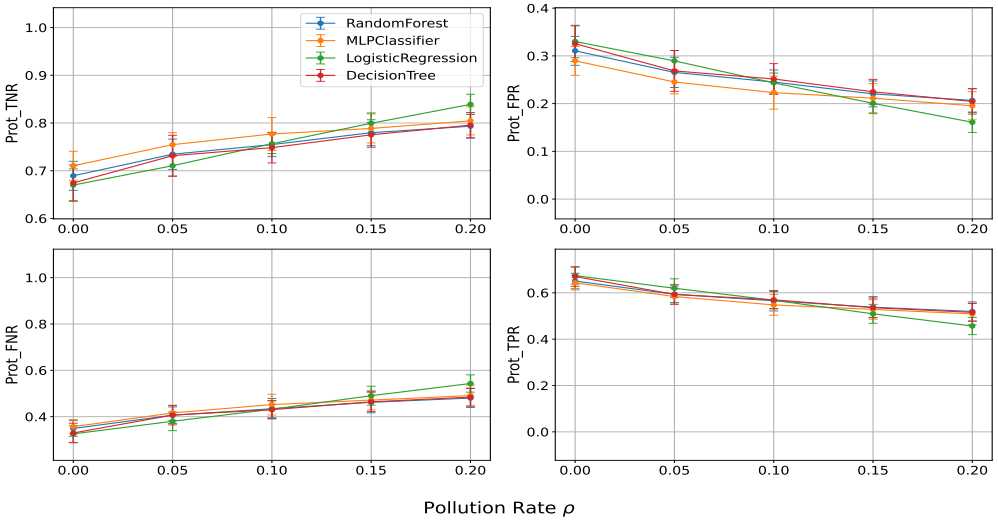
Figure A.12: TPR, TNR, FPR and FNR for the LOW strategy (Protected Subsets).



(a) Adult



(b) Bank



(c) COMPAS