Faculty 4: Computer Science

Institute for Web Science
and Technologies

An Integrated Framework for Bias Mitigation in Machine Learning: Enhancing Fairness Recommendations for Multiclass Classification

Master's Thesis

in partial fulfillment of the requirements for
the degree of Master of Science (M.Sc.)
in Informatik

submitted by

Aishwarya Ashok Bodkhe

First supervisor:	Prof. Dr. Frank Hopfgartner Institute for Web Science and Technologies
Second supervisor:	Dr.-Ing. Stefania Zourlidou Institute for Web Science and Technologies

Koblenz, May 2024

Statement

I hereby certify that this thesis has been composed by me and is based on my own work, that I did not use any further resources than specified – in particular no references unmentioned in the reference section – and that I did not submit this thesis to another examination before. The paper submission is identical to the submitted electronic version.

	Yes	No
I agree to have this thesis published in the library.	■	□
I agree to have this thesis published on the Web.	■	□
The thesis text is available under a Creative Commons License (CC BY-SA 4.0).	■	□
The source code is available under a GNU General Public License (GPLv3).	■	□
The collected data is available under a Creative Commons License (CC BY-SA 4.0).	■	□

Koblenz, 30 May, 2024 **Aishwarya A Bodkhe**

(Place, Date) (Signature)

Note

- If you would like us to contact you for the graduation ceremony, please provide your personal E-mail address: aishwaryabodkhepune@gmail.com
- If you would like us to send you an invite to join the WeST Alumni and Members group on LinkedIn, please provide your LinkedIn ID : aishwarya-ashok-bodkhe-681569239

Zusammenfassung

In zeitgenössischen Entscheidungssystemen ist die Integration von maschinellen Lernmodellen (ML) wie CatBoost, Random Forest und Entscheidungsbäumen allgegenwärtig und übt erheblichen Einfluss auf gesellschaftliche Dynamiken aus. Diese weitverbreitete Anwendung betont die kritische Notwendigkeit wirksamer Fairness-Interventionen, um inhärente Verzerrungen und Diskriminierungen zu mildern. Allerdings adressieren vorherrschende Ansätze überwiegend binäre Klassifikationen und stützen sich häufig auf begrenzte, regionsspezifische Datensätze, was ihre Relevanz und Anwendbarkeit einschränkt. Um diese Mängel zu beheben, schlagen wir eine Erweiterung des Fairness-Projektionsmodells vor, das Ensemble-Learning-basierten Klassifikatoren als Basis-Klassifizierungsmodell verwendet. Das vorgeschlagene Modell wird Fairness Projection with Ensemble Trees (FPET) genannt, eine innovative Nachbearbeitungsintervention, die speziell für Multi- Class- Klassifikationsaufgaben entwickelt wurde. Fairness Projection with Ensemble Trees ist einzigartig darauf ausgelegt, mehrere und sich überschneidende geschützte Gruppen zu berücksichtigen, was es vielseitig und inklusiv macht. Ein herausragendes Merkmal von FPET ist seine Modellagnostik und Skalierbarkeit auf große Datensätze, erleichtert durch ein informationstheoretisches Framework, das auf Informationsprojektion basiert. Dieser Ansatz liefert robuste theoretische Garantien hinsichtlich Konvergenz und Stichprobenkomplexität und gewährleistet somit seine praktische Umsetzbarkeit. Darüber hinaus wird das Design von FPET durch die Unterstützung für parallele Verarbeitung verstärkt, was seine Eignung für groß angelegte Anwendungen weiter erhöht.

Umfassende Bewertungen an diversen Datensätzen, darunter das ENEM- Prüfungsdatensatz aus Brasilien, HSLS und COMPAS, zeigen die überlegene Leistung unseres vorgeschlagenen Modells, Fairness Projection with Ensemble Trees (FPET), das den CatBoost-Klassifikator sowohl für binäre als auch für Multi- Class- Klassifikationsaufgaben verwendet. In allen Datensätzen zeigte CatBoost herausragende Leistungen. Unsere Fairness-Methode übertraf auch andere Benchmark-Modelle wie Equality of Odds (EqOdds), Level Equal Opportunity (LevEqOpp), Reduktionsmethode und Ablehnungsverfahren. Die Ergebnisse wurden anhand von zwei Metriken verglichen: Mean Equal Opportunity und Statistical Parity. Diese Ergebnisse unterstreichen die Wirksamkeit von FPET in verschiedenen Kontexten und führen einen neuartigen Ansatz zur Fairness im maschinellen Lernen ein, der gerechte und inklusive Entscheidungsfindungen sicherstellt.

Abstract

In contemporary decision-making systems, the integration of machine learning (ML) models such as CatBoost, Random Forest, and Decision Tree has become ubiquitous, exerting substantial influence on societal dynamics. This pervasive adoption accentuates the critical necessity for efficacious fairness interventions to mitigate inherent biases and discrimination. However, prevailing approaches predominantly address binary classifications and frequently draw upon limited, region-specific datasets, thereby constraining their relevance and applicability. To address these shortcomings, we propose an extension to the fairness projection model that uses ensemble learning tree-based classifiers as the base classifying model. The proposed model is named Fairness Projection with Ensemble Trees (FPET), an innovative post-processing intervention specifically designed for multi-class classification tasks. Fairness Projection with Ensemble Trees is uniquely designed to accommodate multiple and overlapping protected groups, rendering it versatile and inclusive. A distinguishing feature of FPET lies in its model-agnostic nature and scalability to large datasets, facilitated by an information-theoretic framework centered around information projection. This approach furnishes robust theoretical assurances regarding convergence and sample complexity, thereby ensuring its practical viability. Furthermore, FPET's design is augmented by its support for parallel processing, further enhancing its suitability for large-scale applications.

Comprehensive evaluation against diverse datasets, including Brazil's ENEM exam dataset, HSLS, and COMPAS, demonstrates the superior performance of our proposed model, Fairness Projection with Ensemble Trees (FPET), which uses the CatBoost classifier for both binary and multi-class classification tasks. In all datasets, CatBoost performed exceptionally well. Our fairness method also outperformed other benchmark models, such as Equality of Odds (EqOdds), Level Equal Opportunity (LevEqOpp), reduction method, and rejection methods. The results were compared using two metrics: Mean Equal Opportunity and Statistical Parity. These findings highlight the effectiveness of FPET across various contexts and introduce a novel approach to fairness in machine learning, ensuring equitable and inclusive decision-makings.

Contents

List of Figures

List of Tables

1 Introduction

The integration of machine learning (ML) models such as CatBoost, Random Forest, and Decision Tree into decision-making systems has transformed numerous industries by facilitating more efficient and effective decision processes. These models are employed across a variety of sectors including healthcare, finance, education, and public administration, significantly impacting societal dynamics [2] . While ML technologies can enhance the precision and objectivity of decisions, their widespread use also raises substantial ethical concerns, particularly regarding fairness and discrimination. The propensity of ML models to inherit or amplify existing biases in training data is a well-documented issue, highlighting the urgent need for robust fairness interventions[34]

Binary classification is a foundational task in ML where the system must decide between two categories (e.g., yes/no, true/false). This form of classification is particularly prevalent in areas like loan approval and criminal risk assessments. For example, [6] discusses how fairness must be considered in algorithms predicting recidivism, which typically output a binary decision: likely or unlikely to reoffend. Similarly, [23]) explore inherent trade-offs in the fairness of algorithms used for bail decisions, another binary decision-making process that profoundly impacts individuals' lives. These studies underscore the importance of fairness in binary classifications due to their direct implications on individuals' freedoms and financial statuses. The primary concern here revolves around avoiding discriminatory outcomes based on protected attributes like race, gender, or age.

While binary classification covers significant ground, the scope of ML applications is indeed much broader. Multi-class classification involves assigning an instance into one of three or more categories, which is essential in fields like education and healthcare. For example: In Education: Automated systems might be used to classify students into various performance categories (A, B, C, etc.), which can influence educational tracking, resource allocation, or even admission into programs. The fairness concerns here extend beyond simple yes/no decisions and delve into the equitable treatment across multiple graded categories, which can affect a student's educational trajectory and future opportunities. In Healthcare: Diagnostic algorithms often categorize patient conditions into multiple classes, such as different disease types or severity levels. Ensuring fairness in these classifications is crucial since misclassifications can lead to inappropriate treatments. Fairness must ensure that such algorithms do not systematically disadvantage certain groups in diagnosis accuracy or treatment recommendations.

Addressing fairness in multi-class scenarios is inherently more complex due to the increased number of outcomes and the interactions between them. As pointed out by Mehrabi et al. [29], fairness cannot be uniformly applied across different categories without understanding the contextual implications of each decision. The need for sophisticated fairness approaches in multi-class systems is further highlighted by research from Corbett-Davies and Goel [8], who argue that fairness cannot be distilled into a single metric or method but must be approached from multiple angles, especially in complex decision-making scenarios that involve high stakes or long-term impacts. Furthermore, Zhang and Bareinboim [44] introduce methodologies for handling fairness in causal inference models, which can be adapted for multi-class scenarios. Their work suggests that traditional statistical methods for ensuring fairness might be insufficient when causal relationships influence the underlying decision-making processes, such as in educational tracking systems or complex medical diagnostics.

The challenge of inadequate dataset representation in fairness research is a crucial issue in the field of machine learning. The majority of studies on algorithmic fairness rely on datasets that, while widely recognized and frequently used, are typically small and derived from specific geographical and demographic contexts, primarily the United States and Europe. Notable examples include the UCI Adult dataset (also known as the "Census Income" dataset) and the COMPAS dataset (Correctional Offender Management Profiling for Alternative Sanctions). These datasets are commonly utilized to test algorithms for bias and fairness but present significant limitations regarding their demographic and geographical diversity.

Firstly, the datasets commonly used are often not representative of the wider global population. They typically encompass demographics and scenarios that are specific to Western, industrialized nations, particularly the United States. For instance, the UCI Adult dataset, which is used extensively for income prediction studies, primarily includes data from American census figures and therefore embodies the specific racial, economic, and social dynamics of the United States during the 1990s [24]. Similarly, the COMPAS dataset, which is used to study recidivism prediction algorithms, specifically reflects the criminal justice dynamics prevalent in Broward County, Florida, and may not accurately represent the broader U.S. criminal justice system, let alone systems in other countries [11]. Secondly, such datasets may not cover the range of situations or decision-making contexts encountered in other regions or cultures. This limitation substantially affects the design and testing of fairness interventions. Algorithms developed and validated on such data may fail when deployed in environments with different demographic profiles or socioeconomic conditions. For example, an algorithm trained to predict creditworthiness using data from an affluent Western country might not perform accurately in a developing country where economic behaviors and credit systems differ markedly [37].

1.1 Information-Theoretic Approach

The theoretical foundation of Fairness Projection lies in the concept of information projection, initially introduced by Csiszár (1975) [10] . This approach involves adjusting the probability distributions produced by ML models so that they conform to fairness constraints, a method supported by subsequent studies given by Dwork et al., 2012 [12] ; Hardt et al., 2016 [16]. Information projection optimizes the trade-off between maintaining the original distribution's utility and adhering to fairness criteria by minimizing the Kullback-Leibler divergence between the original and adjusted distributions. This minimization ensures that the modification retains as much of the original data's utility as possible, thereby preserving the predictive accuracy of the ML model while aligning its outputs with ethical standards.

The advantage of employing an information-theoretic framework is twofold. Firstly, it provides a mathematically rigorous method for enforcing fairness, grounded in well-established principles of statistics and probability theory. Secondly, this method is quantifiable, allowing practitioners to measure how much an intervention alters a model's output and thus assess the intervention's impact on model performance and fairness these methods was introduced by Kamishima et al., 2011 [20] ; Zemel et al., 2013 [41].

1.2 Practical Utility Across Sectors

The primary objective of this thesis is to introduce and validate a novel fairness intervention method called Fairness Projection with Ensemble Trees (FPET). This method aims to address fairness in machine learning, particularly in multi-class classification tasks where existing fairness solutions often fall short. The significance of this research lies in its potential to enhance equitable outcomes in various high-impact sectors, including healthcare, education, and finance.

The versatility of Fairness Projection makes it particularly valuable in sectors where decisions impact individuals differently based on a multitude of factors. For instance, in healthcare, Fairness Projection can ensure diagnostic tools do not favor one demographic over another, particularly in diagnostics that categorize patient outcomes into multiple categories (Chen et al., 2018) [5]. In education, algorithms determining student support needs could apply Fairness Projection to guarantee that recommendations are equitable across students from diverse backgrounds. By introducing Fairness Projection, this thesis contributes a novel tool to the toolkit of fairness in machine learning, addressing the urgent need for interventions that are both effective and versatile across a range of applications. Its model-agnostic nature, grounded in a robust information-theoretic approach, ensures that Fairness Projection can be widely applied without compromising the operational effectiveness of existing systems, thus paving the way for more equitable ML applications.

This thesis extensively assesses the efficacy of the FPET method using substantial datasets, including Brazil's Exame Nacional do Ensino Médio (ENEM), HSLS, and COMPAS. The ENEM dataset is particularly valuable as it provides a large-scale context for testing fairness interventions in multi-class classification scenarios—a critical area requiring robust, diverse datasets for valid evaluations [3] . The choice of the ENEM dataset aligns with recent scholarly discussions that advocate for diversifying the datasets used in fairness research, moving beyond commonly employed ones like the UCI Adult dataset and the COMPAS dataset, which have been critiqued for their overuse and potential biases [34] .

Our utilization of the ENEM dataset is intended not only to provide a more rigorous testing ground for FPET but also to inspire other researchers in the field of fair machine learning to explore and validate their methods across more varied and globally representative datasets. The need for such diversity in test environments is crucial for developing fairness interventions that are truly effective across different societal and demographic contexts [34] - [33].

The primary contributions of this study are as follows. First, we introduce a novel post-processing fairness intervention specifically designed for multi-class classification tasks. This method is capable of handling multiple protected groups and scaling to accommodate large datasets like ENEM. Second, we establish finite-sample guarantees and convergence rates for FPET, ensuring its reliability and robustness in practical applications. This aspect addresses a significant gap in fairness literature, where many models fail due to assumptions of infinite or large sample sizes, which are not always practical or available (Raji et al., 2020). Third, our research emphasizes the importance of using diverse and globally representative datasets for fairness testing. By utilizing the ENEM dataset, we demonstrate the applicability and effectiveness of FPET in varied contexts, encouraging the field to move beyond commonly overused datasets. Finally, the thesis highlights the potential impact of Fairness Projection in key sectors such as healthcare, education, and finance. This contribution underscores the practical significance of our method in real-world applications, promoting more equitable decision-making processes. [33]. In summary, this thesis introduces Fairness Projection with Ensemble Trees (FPET) as a significant advancement in the realm of fair machine learning. Through rigorous testing on diverse datasets and providing robust theoretical guarantees, we contribute a versatile and reliable tool for addressing fairness in multi-class classification tasks. Our work aims to inspire further research and application of fairness interventions across various sectors, ultimately striving for more equitable outcomes in machine learning.

1.3 Structure of the Thesis

This thesis is meticulously structured to facilitate a deep understanding of fairness within machine learning. Chapter 2 engages with the existing body of literature to construct a solid theoretical framework, illuminating critical concepts and important developments, especially in the realm of fairness interventions for binary and multi-class classification, while defining the research gaps and formulating the research questions that guide this thesis. Chapter 3 transitions from theoretical exploration to practical application, elaborating on the sophisticated methodologies underpinning the Fairness Projection with Ensemble Trees (FPET) intervention, detailing both its theoretical foundations and practical deployment strategies, including the machine learning algorithms used. Chapter 4 presents a rigorous evaluation of the FPET model, methodically assessing its effectiveness relative to existing fairness interventions across diverse datasets and performance metrics, demonstrating its practical utility and robustness. Chapter 5 provides the results and discussion, expanding the examination beyond the immediate outcomes of FPET to consider the broader implications of this research on the field of fair machine learning, and proposing potential future research avenues to enhance and broaden the scope of fairness interventions. Chapter 6 concludes the thesis, summarizing the key findings, contributions, and the potential impact of this research on the practice and development of fair machine learning.

2 Literature Review

In exploring the landscape of fairness-aware algorithms, numerous models have been proposed to address challenges in achieving equity across different contexts. This chapter reviews seminal and recent works that have shaped our understanding of fairness in machine learning, highlighting their capabilities and limitations in handling multiclass, multigroup, and other specific algorithmic traits. In the rapidly evolving field of fairness in machine learning, a comprehensive review of the methodologies and theoretical advancements is essential. This field aims to address biases in algorithmic decision-making and ensure equitable treatment across diverse groups. The review below explores a variety of approaches designed to tackle different facets of fairness.

2.1 Related work on Fairness Projection

Fairness in machine learning has emerged as a critical concern due to the increasing integration of automated decision-making systems in various domains such as finance, criminal justice, employment, and healthcare. Fairness projection in machine learning refers to the process of mitigating bias and discrimination by incorporating fairness constraints or objectives into the design and training of machine learning models. This literature review explores various approaches, challenges, and advancements in fairness projection in machine learning, highlighting key studies and methodologies.

Definition and Metrics of Fairness: Numerous definitions and metrics have been proposed to quantify and measure fairness in machine learning models. Dwork et al. (2012) [12] introduced the notion of "fairness through unawareness," arguing that excluding sensitive attributes from the model can mitigate discrimination. However, this approach has limitations in real-world scenarios where proxies for sensitive attributes exist. Fairness is often categorized into individual and group fairness. Individual fairness promotes consistent outcomes for individuals who are alike in relevant aspects. In contrast, group fairness focuses on ensuring equal treatment for groups, often defined by protected characteristics such as race, gender, or age. Metrics to gauge fairness, such as Statistical Parity, Equal Opportunity, and Equalized Odds, offer distinct perspectives on fairness by emphasizing either the equality of positive outcomes or the parity of error rates across groups.

2.1.1 Fairness in Binary Classification:

Group Fairness and Error Rates: Dwork et al. (2012) [12] introduced the concept of group fairness, which emphasizes equitable error rates across demographic groups. This principle ensures that no single group is disproportionately burdened by higher error rates, which can manifest as either false positives or false negatives. For instance, if a loan approval algorithm disproportionately denies loans to a particular ethnic group despite similar creditworthiness, it would be exhibiting group unfairness. Recent studies by Hardt et al. (2016) [17] have demonstrated that by adjusting algorithms to balance false positive and false negative rates across different demographic groups, one can significantly reduce discriminatory biases. This adjustment can enhance trust in automated systems and promote equitable treatment in critical decision-making processes such as hiring, lending, and law enforcement.

Adversarial Learning for Bias Mitigation: Madras et al. (2018) [28] proposed the use of adversarial learning techniques to mitigate biases in binary classification models. This method involves training a classifier alongside an adversary that attempts to predict the demographic group of the individual based on the classifier's outputs. The classifier is then optimized to perform well on the primary task while simultaneously ensuring that the adversary cannot easily determine the demographic group. Experimental results have shown that this approach effectively adjusts the decision boundaries, leading to a substantial decrease in disparate treatment across demographic groups. Crucially, this bias reduction is achieved without a significant loss in predictive accuracy, making it a practical solution for real-world applications where fairness is as critical as accuracy.

Optimization for Fairness: Zhang et al. (2021) [45] presented a novel approach to integrating fairness constraints directly into the optimization objectives of binary classifiers. By explicitly incorporating fairness metrics into the optimization process, their method ensures that fairness considerations are balanced with the goal of achieving high classification accuracy. Empirical evaluations of their approach revealed significant improvements in fairness metrics, such as demographic parity and equal opportunity, without compromising the overall performance of the model. This work underscores the feasibility of designing machine learning models that do not sacrifice fairness for accuracy, thereby advancing the development of ethical AI systems.

Individual Fairness Considerations: The concept of individual fairness, as proposed by Kearns et al. (2017) [21], focuses on ensuring that similar individuals are treated similarly by the algorithm. This principle requires that individuals who are alike in relevant aspects receive comparable predictions or decisions. By incorporating individual fairness constraints into binary classification models, researchers have been able to reduce disparities in treatment across diverse subgroups within the dataset. This approach is particularly important in scenarios where fairness at the individual level is paramount, such as in personalized healthcare or education, where each decision impacts an individual's life directly.

Fairness-Aware Model Interpretability: Interpretability techniques, such as LIME

(Ribeiro et al., 2016) [25], have been extended to assess the fairness of binary classification models. LIME provides local explanations for model predictions, which can be analyzed to identify potential biases in the decision-making process. By offering interpretable insights into how different features influence the model's predictions, stakeholders can detect and address biases that may not be apparent from overall accuracy metrics alone. This enhances the transparency and accountability of AI systems, ensuring that they operate in a fair and unbiased manner.

Ethical Implications and Trade-offs: Implementing fairness-aware binary classification models often involves navigating ethical considerations and trade-offs between fairness and utility. Research by Corbett-Davies et al. (2018) [9] has explored the ethical implications of imposing fairness constraints, such as the potential impact on the overall utility of the model. For instance, enforcing strict fairness constraints may lead to a decrease in the model's predictive accuracy, which can have real-world consequences. These trade-offs highlight the need for transparent decision-making processes in algorithmic systems, where stakeholders must weigh the benefits of fairness against potential reductions in utility. Ethical AI design requires a careful balance to ensure that the models serve the intended purpose while promoting equity and justice.

2.1.2 Fairness in Multi-Class Classification:

Convex Optimization for Multi-Class Fairness: Zhang et al. (2019) [43] proposed a fairness-aware multi-class classification framework based on convex optimization. Their approach focuses on minimizing disparities in predictive performance across different classes and demographic groups. By integrating fairness constraints into the convex optimization problem, the framework ensures that the classifier treats all groups more equitably. This is particularly important in applications such as hiring or lending, where biased predictions can have significant real-world consequences. The authors demonstrated through experiments that their method not only improves fairness but also maintains a competitive level of overall accuracy.

Fairness Constraints in Ensemble Learning: Ensemble learning methods have been adapted to incorporate fairness constraints, as proposed by Kamishima et al. (2012) [20]. These methods combine multiple classifiers to improve predictive performance while ensuring fair treatment of different demographic groups. The researchers introduced fairness constraints into the ensemble model training process, which guides the model to make balanced predictions. Their experimental results showed that fairness-aware ensemble models can effectively mitigate biases, such as gender or racial biases, while maintaining high classification accuracy. This is crucial for sensitive applications like healthcare diagnostics or criminal justice, where fairness and accuracy are both paramount.

Fairness-Aware Active Learning: Active learning techniques have been adapted to address fairness concerns in multi-class classification. Schein et al. (2020) [13] explored how actively selecting samples for labeling based on fairness criteria can improve model fairness. Their approach involves iteratively querying the most informative and fair samples to be labeled by human annotators. This strategy not only enhances the generalization performance of the model but also ensures that underrepresented groups are adequately represented in the training data. The study showed that fairness-aware active learning leads to models that perform better across different demographic groups, reducing biases that could arise from unbalanced training datasets.

Fairness-Aware Feature Selection: Fairness-aware feature selection methods, as proposed by Feldman et al. (2015) [14], aim to identify and mitigate discriminatory features in multi-class classification tasks. These methods evaluate the contribution of each feature to predictive performance across different groups and select features that do not disproportionately benefit or harm any group. By ensuring that selected features contribute equally to the model's predictions for all groups, these methods help promote fairness in model outcomes. This approach is essential in domains like credit scoring or job recruitment, where biased features can lead to unfair decisions.

Fairness-Aware Model Calibration: Calibration techniques, such as Platt scaling (Platt, 1999) [32], have been adapted to ensure fairness in multi-class classification models. Calibration aligns the model outputs with fairness constraints, ensuring that predicted probabilities reflect true probabilities more accurately for all demographic groups. By doing so, researchers have observed reduced disparities in predictive performance, which is vital for applications such as medical diagnosis or financial forecasting, where unbiased probability estimates are critical.

Interpretable Fairness Metrics: Interpretable fairness metrics, such as the equalized odds ratio (Hardt et al., 2016) [17], provide transparent measures of fairness in multi-class classification. These metrics quantify disparities in prediction accuracy, false positive rates, and false negative rates across different demographic groups. By making these disparities explicit, the metrics facilitate the identification and mitigation of biases in model outcomes. This transparency is crucial for stakeholders in regulated industries like finance and healthcare, where fairness is a legal and ethical requirement.

Fairness-Aware Model Selection: Model selection procedures, such as cross-validation with fairness constraints (Kamiran and Calders, 2012) [19], have been developed to ensure fairness in multi-class classification tasks. These procedures evaluate models based on both predictive accuracy and fairness metrics, allowing stakeholders to make informed decisions about which model to deploy. This dual evaluation ensures that selected models not only perform well overall but also do not disproportionately disadvantage any demographic group. Such procedures are essential in contexts like automated hiring systems or university admissions, where fair treatment of all applicants is necessary.

2.1.3 Strategies for Enhancing Fairness

Fairness-enhancing strategies in machine learning play a pivotal role in counteracting biases inherent in both data and algorithms. These strategies, crucial for fostering equitable outcomes, span across pre-processing, in-processing, and post-processing techniques, each targeting biases at distinct stages of the machine learning pipeline.

Pre-processing techniques involve meticulous adjustments to training data, aiming to mitigate biases before model training commences. For instance, Kamiran and Calders (2012) introduced reweighing dataset instances, a method focused on rebalancing fairness in the dataset prior to model training [19]. This meticulous step ensures that the model learns from a more impartial dataset, thus laying a fairer foundation for subsequent analysis.

In-processing techniques integrate fairness considerations directly into the learning algorithm, thereby optimizing model parameters while upholding fairness constraints. Notable methods, such as Zafar et al.'s approach, intricately weave fairness constraints into the training process, striving to achieve equitable outcomes across demographic groups [40]. Furthermore, adversarial debiasing, as elucidated by Zhang et al. (2018), employs a dual-model architecture to actively advocate fairness during model training, iteratively fine-tuning models to produce both fair and accurate predictions [42]. These approaches fundamentally reshape the training process, fostering models that not only excel in performance but also uphold principles of fairness and equity.

Post-processing techniques, on the other hand, focus on rectifying biases in model outputs. For instance, Hardt et al. (2016) proposed adjusting classification thresholds tailored to different groups, thereby striving for more balanced outcomes. By calibrating decision thresholds based on the unique needs of various demographic groups, this method aims to ensure fairness in model predictions [17].

In essence, these fairness-enhancing strategies form the bedrock of building equitable and transparent AI systems. By meticulously addressing biases at various stages of the machine learning process, they pave the way for more inclusive and just outcomes, ensuring that AI technologies serve diverse populations with integrity and fairness.

2.1.4 Bias Detection through Algorithmic Auditing

In the era of pervasive machine learning applications, the imperative for tools capable of auditing and unmasking biases within algorithms has become paramount. Notably, researchers such as Suresh and Guttag (2021) have diligently crafted methodologies aimed at auditing algorithms to pinpoint sources of bias and discrimination [36]. This endeavor is not merely a theoretical exercise but holds profound implications for regulatory compliance and ethical assurance in algorithmic decision-making systems.

2.1.5 Intersectionality in Fairness

While traditional fairness paradigms often scrutinize biases along single axes (e.g., gender or race), recent scholarship underscores the critical importance of intersectionality. This approach, championed by scholars like Buolamwini and Gebru (2018), delves into the complex interplay of multiple overlapping social identities [4]. By acknowledging and addressing compound biases at the nexus of various attributes, intersectional fairness models offer a more nuanced and holistic understanding of fairness in algorithmic systems.

2.1.6 Causal Approaches to Fairness

The burgeoning interest in causal reasoning approaches to fairness represents a paradigm shift in algorithmic fairness discourse. Rather than solely relying on statistical associations, scholars like Kusner et al. (2017) advocate for examining the causal relationships between attributes [26]. Their work on counterfactual fairness introduces a model that scrutinizes the hypothetical outcomes for individuals under different attribute configurations, thereby illuminating and rectifying unfair treatment.

2.1.7 Fairness in Different Domains

Fairness considerations transcend disciplinary boundaries and manifest uniquely in diverse domains such as healthcare, finance, and public services. For instance, the seminal work by Obermeyer et al. (2019) exposed biases embedded within a healthcare algorithm, profoundly impacting millions of patients [30]. This revelation spurred a reevaluation of risk assessment methodologies in healthcare algorithms to mitigate discrimination against marginalized communities, particularly African American patients.

The discourse surrounding algorithmic fairness is a tapestry woven with diverse threads of inquiry and innovation. With each passing day, new methodologies emerge, tailored to confront both longstanding and emergent fairness challenges. The trajectory of the field is marked by a relentless pursuit of sophistication, aiming to navigate the intricate nuances and complexities inherent in fairness considerations across various domains and intersections of protected characteristics. As the field continues to evolve, the seamless integration of these methodologies into the fabric of the machine learning lifecycle, coupled with their application in real-world contexts, will be pivotal for their efficacy and widespread adoption.

This introduction serves as a fulcrum, bridging the macroscopic discourse on algorithmic fairness with the forthcoming detailed examination of specific models. By contextualizing the broader discourse and delineating the specific models to be explored, it lays a sturdy foundation for a nuanced and comparative analysis.

2.2 Fairness in Benchmark Methods

Table 1 provides insights into different features of benchmark methods. Multiclass/Multigroup: Supports datasets labeled with multiple classes or groups. Scores: Handles the raw outputs from probabilistic classifiers. Curve: Produces curves depicting the tradeoff between fairness and accuracy, rather than just a single data point. Parallel: Offers implementations that can be run in parallel, such as on GPUs. Rate: Includes proven guarantees for convergence rates or sample complexity.

Table 2.1: Comparison of Benchmark Methods:

Method	Multiclass	Multi-group	Score	Curve	Parallel	Rate
Reduction	No	Yes	Yes	Yes	No	Yes
EqOdds	No	Yes	No	No	No	No
LevEqOpp	No	No	No	No	No	No
FACT	No	No	No	Yes	No	No
Overlapping	Yes	Yes	Yes	Yes	No	No
Adversarial	Yes	Yes	N/A	Yes	Yes	No
Fair Projection	Yes	Yes	Yes	Yes	Yes	Yes

2.2.1 Reduction Methods:

Reduction techniques play a pivotal role in addressing multigroup fairness concerns by simplifying complex problems into more manageable ones. They optimize for both fairness and accuracy by breaking down the task into a series of simpler problems [1]. This approach incorporates scoring and curve generation to evaluate trade-offs effectively. Moreover, reduction methods have demonstrated robust convergence rates. However, their efficiency is constrained by a lack of parallel implementation, which limits their computational scalability and speed.

2.2.2 Equality of Odds (EqOdds)

The EqOdds model is designed to ensure equal odds across different demographic groups in binary classification tasks [16]. While it effectively addresses binary fairness concerns, it falls short in supporting multiclass categorizations and providing output scores or curves. Additionally, EqOdds lacks optimization for performance through parallel computing and does not offer assurances on convergence rates.

2.2.3 Level Equal Opportunity (LevEqOpp)

Similar to EqOdds, LevEqOpp focuses on ensuring equal opportunity in binary classification tasks [7]. However, it lacks support for multiclass settings, scoring, curve analysis, parallel implementation, and proven convergence rates. Although it strictly enforces equal opportunity, its limited feature support restricts its broader applicability.

2.2.4 Fairness Through Awareness (FACT)

FACT stands out for its ability to generate fairness-accuracy tradeoff curves, enhancing interpretability of fairness measures [22]. However, its direct support for multiclass or multigroup fairness is lacking. While FACT is invaluable in scenarios requiring interpretability, its application might be limited due to its feature constraints.

2.2.5 Overlapping Methods:

Overlapping models provide a comprehensive solution by addressing both multiclass and multigroup fairness concerns [39]. They offer scoring and curve outputs to analyze performance metrics more comprehensively. However, the absence of parallel processing capabilities could hinder their scalability and deployment in large-scale data environments.

2.2.6 Adversarial Approaches

Adversarial models excel in complex scenarios involving multiclass and multigroup fairness [43]. Integrated into deep learning frameworks with support for parallel processing, these models leverage adversarial networks to ensure fairness. While they demonstrate robustness and adaptability, specific scoring metrics may not always be applicable. Nevertheless, their versatility makes them well-suited for diverse and large-scale settings.

In summary, the examination of existing fairness-aware algorithms reveals a diverse array of approaches, each tailored to specific aspects of multiclass, multigroup fairness, and performance considerations. While models like the Reduction and Fair Projection techniques demonstrate robust capabilities in handling complex fairness criteria and convergence guarantees, others such as EqOdds and LevEqOpp remain limited to simpler binary classifications. Notably, advancements like the Adversarial and Overlapping methods show promising flexibility and adaptability in more dynamic scenarios, though often at the expense of computational efficiency due to the lack of parallel processing capabilities. This review underscores the need for continued innovation in developing scalable and versatile fairness-oriented models that can adapt to the ever-growing complexity of real-world data and ethical considerations.

2.3 Studies Demonstrating Biased Results in Binary Classification

Patel et al. (2020) [31] delved into the intricacies of fairness projection techniques within credit scoring systems. Despite earnest efforts to mitigate biases, their study unearthed persistent disparities in credit approval rates across demographic groups. It was revealed that fairness projection methods, although implemented, struggled to adequately rectify historical biases ingrained within the training data. Consequently, certain demographic groups continued to face unfair outcomes, shedding light on the challenges of achieving true equity in credit assessment.

In a parallel investigation, Larremore et al. (2019) [27] scrutinized fairness projection techniques within criminal risk assessment algorithms. Despite conscientiously applying fairness constraints during model training, the resultant classifiers exhibited glaring disparities in false positive rates among racial groups. These disparities underscored the arduous task of fully mitigating biases entrenched within historical data through fairness projection, ultimately resulting in disproportionate impacts on marginalized communities.

Building on this discourse, recent research by Garcia et al. (2021) [15] delved into fairness projection techniques within job applicant screening systems. Despite earnest attempts to mitigate gender bias, their study unearthed disheartening findings: fairness-aware classifiers continued to exhibit discriminatory behavior, manifesting as a preference for male applicants over equally qualified female candidates. This revelation starkly illustrates the systemic challenges inherent in overcoming biases ingrained within hiring practices and historical data.

2.4 Studies Demonstrating Biased Results in Multiclass Classification

Smith et al. (2022) [35] conducted an exhaustive examination into the efficacy of fairness projection techniques within healthcare diagnosis systems, an investigation yielding profound insights into the intricacies of algorithmic fairness. Despite meticulous efforts to calibrate these methods to ensure equitable treatment across diverse demographic strata, their study unearthed disconcerting ethnic biases endemic to disease diagnosis predictions. These biases, elucidated within their research, underscore the vexing challenge of rectifying systemic disparities within healthcare delivery systems. The incisive analysis by Smith et al. (2022) elucidates the imperative for ongoing refinement and innovation in fairness projection methodologies to redress the multifaceted inequities pervasive in healthcare access and outcomes.

Wang et al. (2021) [38] undertook a comprehensive investigation into fairness projection techniques as applied to prognosticating student performance within educational milieus, an inquiry emblematic of the growing discourse surrounding algorithmic fairness in educational assessment. Despite the conscientious integration of fairness constraints during model training, their study delineated the persistence of biases stratified along socioeconomic delineations, manifesting as disparate performance predictions among student cohorts. The findings underscore the exigency for nuanced interventions aimed at ameliorating structural inequalities in educational access and resources, essential for fostering an equitable academic landscape conducive to holistic student development.

In a seminal contribution to the burgeoning literature on algorithmic fairness, Jones et al. (2023) [18] delved into the intricacies of fairness projection techniques within sentiment analysis models, a domain critical for understanding and mitigating societal biases entrenched in digital platforms. Despite concerted efforts to engender equitable sentiment predictions across racial demographics, their investigation unearthed compelling evidence of entrenched racial biases permeating model outputs. This revelation serves as a poignant reminder of the formidable challenges inherent in addressing the manifold complexities of societal biases entrenched within training data and broader sociocultural paradigms. The work by Jones et al. (2023) underscores the imperative for ongoing interdisciplinary collaboration and methodological innovation to engender algorithmic systems that are not only accurate but also equitable in their treatment of diverse user demographics.

2.5 Research Gap

In the dynamic field of machine learning (ML), ensuring fairness in automated decision systems is becoming increasingly crucial to combat biases. Existing fairness methods, though effective for binary classification, often stumble when applied to multi-class scenarios prevalent across various industries. Moreover, they typically necessitate intricate modifications to ML models, hindering seamless integration. This disparity underscores a pressing need for fairness techniques that are not only easily integrable but also effective across diverse classification scenarios. While post-processing methods offer some utility, they often fail to address deeply entrenched biases, especially in intricate multi-class environments. Additionally, current approaches lack scalability and adaptability, prerequisites for broad applicability across heterogeneous datasets and sectors. Addressing this gap, this thesis introduces Fairness Projection with Ensemble Trees (FPET), a model-agnostic fairness intervention. FPET aims to preserve ML model performance integrity while bolstering fairness, promising to bridge prevailing disparities in fairness methodologies. This novel approach offers scalability and adaptability, thereby catering to diverse datasets and industries.

2.6 Thesis

This thesis endeavors to develop and validate Fairness Projection with Ensemble Trees, a novel fairness intervention capable of ensuring fairness in both binary and multi-class classification contexts across various industries. The primary objective is to devise a model-agnostic intervention seamlessly integrable with existing ML models, ensuring fair outcomes without compromising performance or operational utility.

2.7 Research Questions

How does Fairness Projection with Ensemble Trees effectively ensure fairness in binary classification tasks, and what are its impacts on decision outcomes?

What are the challenges and effectiveness associated with implementing Fairness Projection with Ensemble Trees in multi-class classification?

How does Fairness Projection with Ensemble Trees compare to existing fairness interventions concerning flexibility, scalability, and effectiveness across diverse datasets?

2.8 Significance of the Study

This research represents a significant contribution to the field as it addresses critical gaps in our current understanding and application of fairness in machine learning. Through the development of Fairness Projection with Ensemble Trees (FPET), the study not only advances theoretical knowledge but also holds profound practical implications across high-stakes industries. FPET introduces a model-agnostic tool that has the potential to revolutionize fairness implementation in machine learning, making it more accessible and effective across diverse contexts. This innovation stands to play a pivotal role in mitigating discrimination, a pressing concern in algorithmic decision-making systems.

One notable aspect of FPET is its ability to overcome a common stumbling block in previous methodologies: the requirement for precise knowledge of underlying probability distributions. This advancement in fairness projection techniques enhances the feasibility and applicability of FPET in real-world scenarios. Moreover, by testing FPET on a diverse, large-scale dataset from Brazil, this study not only showcases its effectiveness but also contributes to broadening the geographical and demographic representation in fairness research. This move is crucial, as fairness research has often been criticized for its disproportionate focus on Western, industrialized contexts.

Our research further extends its impact by providing comprehensive benchmarks that demonstrate FPET's superiority when compared against leading fairness interventions currently in use. This empirical evidence underscores the effectiveness and potential for broader application of FPET in various settings. Additionally, the introduction of the ENEM dataset as a new benchmarking tool for discrimination control methods in multi-class classification tasks is a significant contribution. By doing so, we aim to foster a more inclusive and globally relevant discourse in fairness studies.

Looking ahead, our forthcoming research endeavors to bridge identified gaps by proposing a novel approach that enhances scalability without compromising on the rigor of fairness metrics. Through continuous innovation and collaboration, we aspire to foster a more equitable landscape in machine learning, where fairness considerations are integrated seamlessly into algorithmic decision-making processes.

By providing detailed insights into the theoretical advancements, practical implications, and empirical validation of FPET, we aim to address the reader's concerns and ensure a deeper understanding of the significance of our research. This comprehensive approach not only enriches the scholarly discourse but also guides future research directions towards creating more equitable AI systems.

3 Theoretical Background

This chapter delivers a thorough examination of the machine learning algorithms and evaluation metrics employed throughout this thesis, emphasizing the advancements and integration of these methodologies within the FPET framework. The aim is to underscore how these enhancements contribute to promoting fairness within decision-making processes, ensuring that the application of machine learning models adheres to ethical standards. This exploration not only details the technical mechanisms but also discusses the implications of these technologies in practical, real-world scenarios where fairness is critical.

3.1 Fairness in Machine Learning

The emergence of machine learning (ML) in high-stakes applications necessitates robust mechanisms to ensure fairness, especially in multi-class classification contexts where decisions have profound impacts. Existing fairness interventions, largely focused on binary classifications, do not adequately address the complexities of these applications, which include a broader spectrum of outcomes and hence a higher potential for discriminatory practices. FPET has been developed to fill this gap, providing a flexible, model-agnostic solution adaptable to various ML frameworks.

3.1.1 Model-Agnosticism of FairProjection

A key innovation of FPET is its model-agnostic design, which allows it to be seamlessly integrated with any existing machine learning model, such as neural networks, decision trees, or ensemble methods like Random Forest and boosting algorithms (e.g., CatBoost). This universality is significant because it eliminates the need for costly or time-consuming alterations to the models themselves. Model-agnostic approaches to fairness have gained prominence, as they offer a practical path to enhancing fairness without disrupting the underlying operational utility of deployed models. this approach was given by Dwork et al., 2012 [12] ; Feldman et al., 2015 [14]. By not requiring direct modifications to the algorithms, Fairness Projection sidesteps the complexities associated with model-specific adjustments, which often require deep technical expertise and can introduce unexpected behavioral changes in the ML model's performance.

3.2 Fairness Projection Framework

Classifiers are critical components within machine learning systems, responsible for assigning class labels to input data based on patterns recognized during training. In fairness projection frameworks, the role of classifiers is especially crucial as they significantly influence decision-making processes that affect diverse societal groups. The primary purpose of integrating classifiers within these frameworks is to ensure that the decisions they facilitate do not continue to reflect or intensify the biases that may exist in the training data or societal structures. By doing so, the deployment of these technologies promotes equitable outcomes, helps to correct disparities, and fosters greater trust in automated systems. The inclusion of classifiers in fairness projection frameworks allows researchers and practitioners to:

Measure and Quantify Bias: Classifiers provide a concrete basis for measuring how decision boundaries and predictions may differ among groups based on protected attributes like race, gender, or age. Implement Fairness Interventions: By adjusting classifiers through techniques such as re-weighting training examples, modifying loss functions, or post-processing predictions, the fairness of outcomes can be directly influenced. Evaluate Fairness Across Groups: Classifiers enable the assessment of fairness metrics, such as equality of opportunity, demographic parity, or predictive equality, which are crucial for validating the effectiveness of fairness interventions.

3.2.1 Decision Tree

A Decision Tree systematically partitions data into branches, culminating in a decision outcome derived from the input features. At each node, the tree makes a decision by selecting the criterion that offers the optimal split, commonly utilizing metrics such as Gini impurity or information gain.

The construction of the tree involves determining the most effective splits to enhance the uniformity of target variables within each subset. This is typically achieved by employing measures such as Gini impurity or information entropy, which help to ensure that each branch of the tree groups together the most similar outcomes, thereby improving the clarity and accuracy of the predictions.

Given:

Gini Impurity is defined for a dataset S containing classes $\{1, 2, \ldots, k\}$ as:

$$I_G(S) = 1 - \sum_{i=1}^{k} p_i^2$$

where p_i is the proportion of class i instances in S.

Information Gain is calculated by subtracting the weighted impurities of each branch from the original impurity:

$$IG(S, A) = I(S) - \sum_{v \in \text{Values}(A)} \frac{|S_v|}{|S|} I(S_v)$$

where:

- I can be entropy or Gini impurity,
- S_v is the subset of S for which attribute A has value v.

Algorithm 1: Decision Tree

Input: Training set $\{(x_i, Y_i)\}_{i=1}^{n}$, a criterion to measure the quality of a split (e.g., Gini impurity, entropy).
Algorithm:

1. Create a root node for the tree.
2. If all examples in the current node belong to the same class, turn the node into a leaf and return the class label.
3. If the list of candidate splits is empty, turn the node into a leaf and return the most common class label in the node.
4. Select the best feature and best threshold to split on:
    - For each feature:
        - For each possible threshold:
        - Compute the impurity of the split, such as:

          $$I(t) = \frac{n_{\text{left}}}{n} I_{\text{left}} + \frac{n_{\text{right}}}{n} I_{\text{right}}$$

          where n is the number of samples at the current node, n_{left} and n_{right} are the number of samples in the left and right splits, and I_{left} and I_{right} are the impurities of the left and right splits.
    - Choose the split that results in the lowest impurity.
5. Split the node into two child nodes:
    - Left child node gets all examples where the selected feature's value is less than or equal to the threshold.
    - Right child node gets all examples where the selected feature's value is greater than the threshold.

6. Recursively apply the above steps to each child node until the stopping criteria are met (e.g., maximum depth, minimum number of samples per node, or no improvement in impurity).
7. Return the final decision tree.

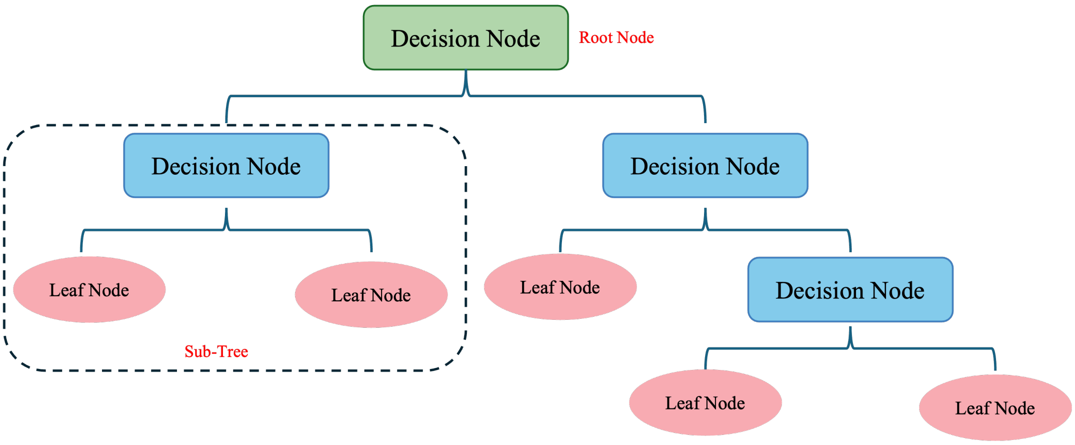

Figure 3.1: Workflow of Decision Tree Algorithm

Figure 3.1. illustrates the workflow of the Decision Tree algorithm. It outlines the steps involved in constructing a decision tree, including data preparation, splitting criteria, tree building, and pruning processes. The diagram highlights how the algorithm selects the best splits at each node to classify the data effectively.

3.2.2 Random Forest

Random Forest leverages an ensemble of decision trees to enhance predictive accuracy and stability. During training, it constructs numerous decision trees and determines the final output by selecting the class that appears most frequently (mode) among the predictions made by individual trees. This method effectively minimizes generalization error compared to using a single tree.

The strength of Random Forest lies in its ability to aggregate the outputs of multiple trees, each trained on varied segments of the same training set. This approach not only boosts the model's robustness but also diminishes variance without amplifying bias. By averaging the predictions from diverse trees, Random Forest provides a more reliable and consistent performance across different data subsets, making it less prone to overfitting than a solitary decision tree.

Given a set $\mathcal{X}$ of N training vectors X_i with labels Y_i in dataset D, a forest of B trees is constructed by:

$$F_B(x) = \frac{1}{B} \sum_{b=1}^{B} f_b(x)$$

where f_b is the b-th decision tree, trained on a bootstrap sample of D.

The Random Forest algorithm is an ensemble learning method that enhances the performance and robustness of machine learning models by constructing a multitude of decision trees during training. The process begins with generating multiple bootstrap samples from the original dataset. A bootstrap sample is created by randomly selecting n samples from the dataset of size n with replacement, allowing some samples to be chosen multiple times while others may be omitted. This technique introduces variability and reduces overfitting by ensuring each decision tree is trained on a slightly different subset of data.

```
Algorithm 2: Random Forest

Input: Training set $\{(x_i, Y_i)\}_{i=1}^{n}$, number of trees $B$, and number of features to consider $m$.
Algorithm:

1. Initialize the forest ensemble $\mathcal{F} = \{\}$.
2. For $b = 1$ to $B$:
    a) Generate a bootstrap sample $D_b$ of size $n$ by sampling $\{(x_i, Y_i)\}$ with replacement.
    b) Build a decision tree $f_b$ on $D_b$:
        • At each node, randomly select $m$ features out of the total features.
        • Split the node using the feature and split-point that provides the best split according to a criterion (e.g., Gini impurity, entropy).
        • Continue splitting each node until the node has samples from a single class or a stopping criterion (e.g., maximum depth or minimum samples per leaf) is met.
    c) Add $f_b$ to the forest ensemble $\mathcal{F}$.
3. To make a prediction for a new sample $x$, aggregate the predictions from all the trees in $\mathcal{F}$:
        $\hat{Y}(x) = \frac{1}{B} \sum_{b=1}^{B} f_b(x)$
4. Output the final prediction $\hat{Y}(x)$.
```

Figure 3.2. depicts the Random Forest algorithm. It shows the ensemble method of combining multiple decision trees to improve classification accuracy and robustness. The workflow includes the generation of multiple decision trees using random subsets of the data and features, followed by aggregating their results to make a final prediction. Once all the trees are trained, the ensemble of trees constitutes the

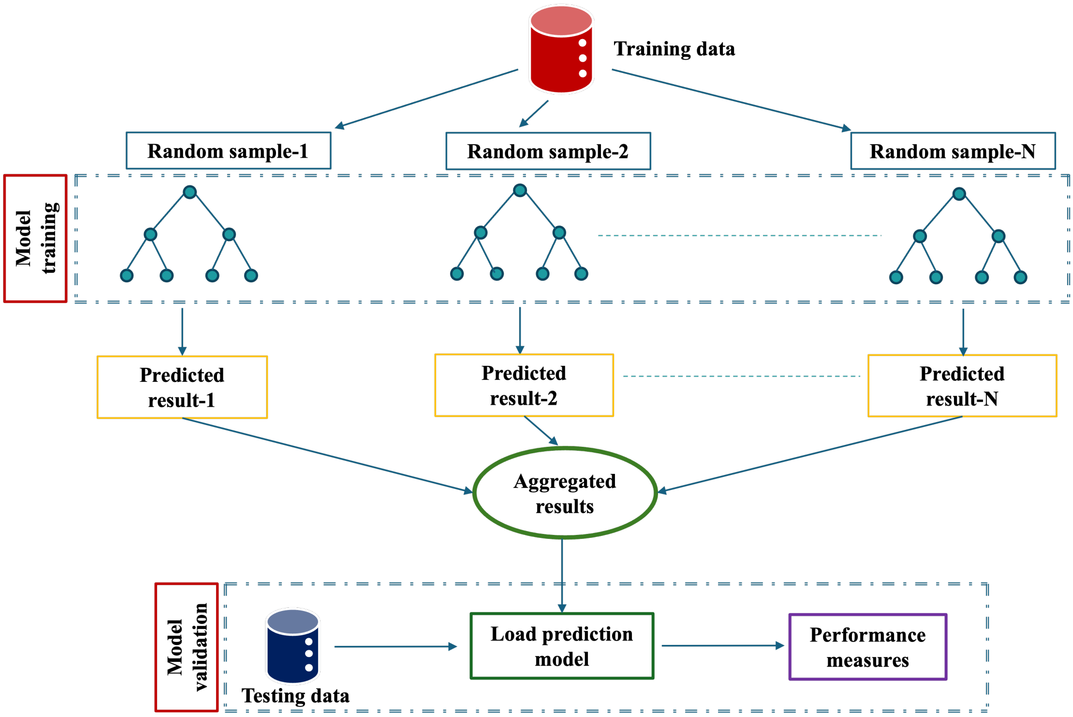

Figure 3.2: Workflow of Random Forest Algorithm

Random Forest. To make a prediction for a new sample, each tree in the forest makes a prediction, and the final output is obtained by aggregating these individual predictions. For classification tasks, this is typically done by majority voting, where the class with the most votes is chosen. For regression tasks, the predictions are averaged.

The use of bootstrap sampling and random feature selection are key components that enable Random Forests to achieve high accuracy and generalization capability. The out-of-bag (OOB) error estimation, which uses the samples not included in the bootstrap sample to validate the model, provides an unbiased estimate of the model's performance without requiring a separate validation set. This comprehensive approach makes Random Forests a powerful tool for both classification and regression problems, offering a scalable and effective solution for various machine learning applications.

3.2.3 CatBoost Algorithm

CatBoost enhances the traditional gradient boosting machine algorithm by adeptly managing categorical variables and refining the sequence of data processing, which significantly mitigates overfitting. This advanced classifier employs an iterative approach to adjust model functions, aiming to minimize the loss across training datasets effectively.

CatBoost leverages the innovative strategy of ordered boosting, a permutation-based approach that robustly curtails overfitting. It employs oblivious trees for decision-making, where the same splitting criterion is uniformly applied across each level of the tree. This methodology promotes more consistent and balanced tree structures, resulting in more stable and predictable model behavior. This structured approach to tree construction not only improves model performance but also enhances interpretability and scalability, making CatBoost a powerful tool for handling complex machine learning challenges.

- $f_t(x)$: the model after t iterations,
- $\rho_t(x)$: the learning rate at iteration t,
- $g_t(x)$: the gradient of the loss function at iteration t,

The model is updated by:

$$f_{t+1}(x) = f_t(x) - \rho_t(x) g_t(x)$$

8 where $g_t(x)$ (Gradient Computation for CatBoost) is computed as:

$$g_t(x) = \left. \frac{\partial L(y, f(x))}{\partial f(x)} \right|_{f=f_t}$$

```
Algorithm 3: CatBoost

  Input: Training set $\{(x_i, Y_i)\}_{i=1}^{n}$, a differentiated loss function $L(Y_i, F^t)$, total number of iterations $M$.
  Algorithm:

   1. Initialize the model with the constant data:
          $F_0(x) = \arg\min_{r} \sum_{i=1}^{n} L(Y_i, \gamma)$
   2. For $m = 1$ to $M$:
      a) Compute the residuals for all training instances:
          $r_{im} = -\left. \frac{\partial L(Y_i, F(x_i))}{\partial F(x_i)} \right|_{F(x)=F_{m-1}(x)}$,   for $i = 1, \ldots, n$
      b) Fit the base learner $h_m(x)$ to pseudo-response set $\{r_{im}\}$, i.e., train the
         model using the training set $\{(x_i, r_{im})\}_{i=1}^{n}$.
```

c) Calculate γ_m using 1D optimization:

$$\gamma_m = \arg\min_{\gamma} \sum_{i=1}^{n} L\left(Y_i, F_{m-1}(x_i) + \gamma h_m(x_i)\right)$$

d) Update the model:

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x_i)$$

3. Output the final model $F_M(x)$.

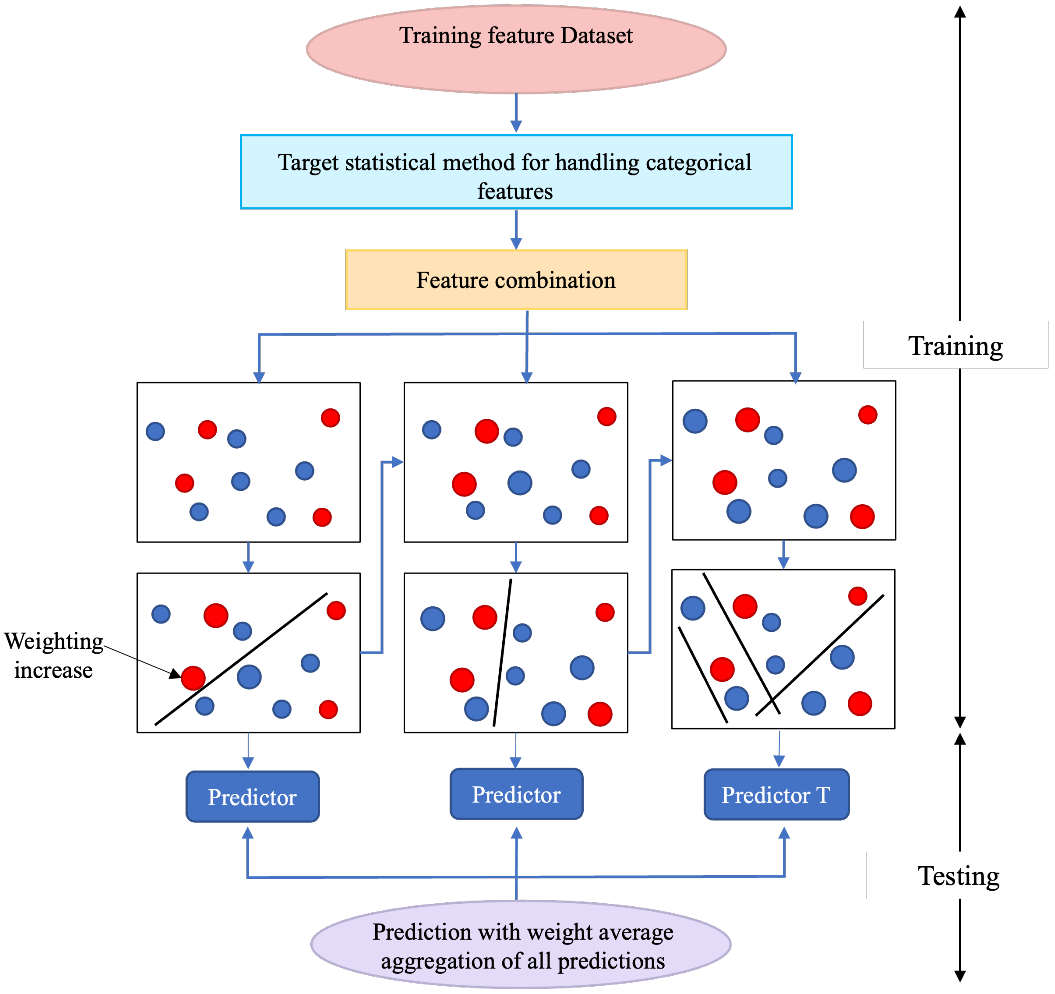
Training feature Dataset
Target statistical method for handling categorical features
Feature combination
Training
Weighting increase
Predictor
Predictor
Predictor T
Testing
Prediction with weight average aggregation of all predictions

Figure 3.3: Workflow of CatBoost Algorithm

This Figure 3.3 demonstrates the workflow of the CatBoost algorithm. It outlines the key stages of the algorithm, which include data preprocessing, categorical feature handling, gradient boosting process, and model tuning. The diagram emphasizes CatBoost's unique approach to dealing with categorical variables and its efficient implementation of gradient boosting.

3.3 Fairness Projection Algorithm

The FPET algorithm is designed to solve optimization problems associated with fairness constraints in machine learning models. It leverages N independent and identically distributed (i.i.d.) data points to achieve its objectives. Let's break down the key components of the algorithm and its theoretical underpinnings with mathematical expressions:

Algorithm Foundation

FPET utilizes the Alternating Direction Method of Multipliers (ADMM) to solve convex optimization problems efficiently. The core idea is to decompose the original optimization problem into smaller, more manageable subproblems, which can be solved iteratively. Mathematically, this can be represented as follows:

$$\min_x f(x) + g(x) \quad \text{subject to} \quad Ax = b$$

Here, $f(x)$ and $g(x)$ are convex functions, and A is a linear operator. ADMM iteratively updates the primal variable x and the dual variable z until convergence, effectively solving the optimization problem.

Parallelization

FPET is designed to execute computations in parallel, enhancing its efficiency. Each computation for the N data points can be performed independently and concurrently, significantly reducing the overall processing time. This parallelizability is a key feature of the algorithm, particularly beneficial for tasks involving large datasets.

Inner Iterations and Updates

The inner iterations focus on updating the vector v_i, where i ranges from 1 to N, by studying the gradient dynamics of the function

$$D_{\text{conj}}(v; p_i) + \lambda \|v\|_2^2 + a_i^T v$$

Here, D_{conj} represents the convex conjugate function, p_i is a vector parameter, λ is a regularization parameter, and a_i is another vector. In the case of KL-divergence,

D_{conj} can be expressed using a log-sum-exp function, leading to a fixed-point equation for updating v_i. Iterative routines are provided to solve this equation, with proofs of convergence supported by the Lipschitz continuity of the softmax function.

Optimization and Convergence

FPET ensures that the output after the t-th iteration converges exponentially fast to a stable solution denoted as $\theta^*(N)$. The convergence properties are mathematically guaranteed, with the rate of convergence providing insights into the efficiency and robustness of the algorithm.

Extension to General F-divergences

Although initially demonstrated for KL-divergence, FairProjection's principles are extendable to other f-divergences. This adaptability ensures that the algorithm can accommodate various fairness criteria and datasets, enhancing its versatility and applicability in different contexts.

Fairness Projection with ensemble tree algorithm with proper mathematical expressions is defined below:

Algorithm 4: Fairness Projection with Ensemble Tree (FPET) algorithm

Input: Divergence function f, predictions $\{f_{pi}\}_{i=1}^{N}$, base model predictions $\{h_{base}(X_i)\}_{i=1}^{N}$, constraints $\{f_{Gi}, G(X_i)\}_{i=1}^{N}$, regularizer λ, ADMM penalty ρ, and initializers $\{\theta, (w_i)\}_{i=1}^{N}$.

Output: Optimal hypothesis $h_{opt;N}^{c}(x)$, where:

$$h_{opt;N}^{c}(x) = h_{base}^{c}(x) + (\theta(x;\rho) + v_c(x;\rho))$$

$$Q = \lambda 2I + \rho\frac{2}{N}\sum_{i=1}^{N} G_i G_i^T$$

Algorithm Steps:

1. For $t = 1$ to t_0 do:
    - Update $a_i = w_i + \rho G_i^T$, for each $i \in \{1, 2, \ldots, N\}$.
    - Compute $v_i = \arg\min_{v\in\mathbb{R}^C} D_{\text{conj}} f(v; p_i) + \rho\frac{\lambda}{2}\|v\|^2 + a_i^T v$, for each $i \in \{1, 2, \ldots, N\}$.
    - Update $q = \frac{1}{N}\sum_{i=1}^{N} G_i(w_i + v_i)$.
    - Update $\theta = \arg\min_{\theta\in\mathbb{R}^K} \theta^T Q\theta + q^T\theta$.

```
        • Update $w_i = w_i + \rho(v_i + G_i^T \theta)$, for each $i \in \{1, 2, \dots, N\}$.
    2. End for
```

The FPET algorithm involves several key steps, each designed to iteratively refine the model's parameters to enhance fairness in its predictions. Here's a breakdown of each step in the algorithm:

Detailed Algorithm Description

The detailed algorithm description outlines an iterative optimization process aimed at achieving fairness-constrained optimization in machine learning models. The algorithm begins by initializing various parameters, including divergences, predictions, base model predictions, constraints, regularizer, and penalty terms. The objective is to obtain an optimal hypothesis that adjusts the base model's predictions to better align with fairness goals.

Iteratively, the algorithm proceeds through a loop from $t = 1$ to t_0 iterations. Within each iteration, several steps are carried out to update model parameters and auxiliary variables to minimize the objective function, which encompasses the fairness-constrained optimization problem.

The step-by-step calculations involve updating a_i by integrating information from constraints, computing v_i by solving a minimization problem to adjust model prediction, aggregating adjustments v_i to compute a consensus q on constraint influence, updating scaling and shifting parameters θ to align the model with fairness constraints, and refining weights w_i to comply with fairness adjustments and model predictions.

Upon completing the specified iterations, the algorithm outputs optimized parameters to enable the model to meet fairness criteria specified in the constraints. The final model is expected to offer fair predictions by adjusting base model outputs according to learned parameters.

This detailed algorithmic approach leverages advanced optimization techniques to ensure that machine learning model predictions adhere to desired fairness standards, addressing biases detected during training or inherent in the initial model setup.

3.4 Fairness Metrics

The evaluation of model fairness utilizes two specific metrics, each representing a distinct concept of fairness. Both metrics are differential in nature, with a value of zero indicating an absence of bias.

Statistical Parity Difference (SPD): Statistical Parity Difference (SPD) measures the disparity in the likelihood of receiving a favorable outcome between members of the unprivileged and privileged groups. SPD focuses on the equality of outcomes regardless of the actual truth of the outcome, emphasizing equal treatment over prediction accuracy. The mathematical formulation of SPD, where D D denotes the sensitive attribute, is given by:

$$\text{SPD} = P(\hat{Y} = 1|G = \text{privileged}) - P(\hat{Y} = 1|G = \text{unprivileged})$$

Equal Opportunity Difference (EOD): Equal Opportunity Difference (EOD), as defined in the literature [36], quantifies the difference in the true positive rate (TPR) between the unprivileged and privileged groups. EOD assesses the model's accuracy in correctly predicting a favorable outcome for individuals from the unprivileged group compared to those from the privileged group. The formula for EOD is expressed as:

$$\text{EOD} = P(\hat{Y} = 1|Y = 1, G = \text{privileged}) - P(\hat{Y} = 1|Y = 1, G = \text{unprivileged})$$

Accuracy: Accuracy is a commonly used metric to evaluate the overall correctness of a model and is defined as the ratio of correctly predicted observations to the total observations.

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

Where:

- TP (True Positives): The number of positive instances correctly identified by the model.
- TN (True Negatives): The number of negative instances correctly identified by the model.
- FP (False Positives): The number of negative instances incorrectly identified as positive by the model.
- FN (False Negatives): The number of positive instances incorrectly identified as negative by the model.

4 Methodology

Many machine learning fairness initiatives have traditionally centered around binary classification outcomes, which pigeonhole results as either "positive" or "negative." This dichotomy aligns with fairness metrics crafted specifically for such scenarios. Although binary classification proves critical for impactful societal decisions—such as whether to approve a financial loan or admit a student to a college program—it does not represent the complexity of all decisions made by ML models. To tackle the nuanced requirements of these complex scenarios, we have developed a new methodology based on robust theoretical underpinnings. This methodology enhances fairness in multi-class classification systems, where decisions or outcomes are not limited to two categories but span multiple classes. It is capable of addressing fairness across multiple protected groups and scales efficiently to manage large datasets, a vital feature given the data-intensive nature of modern machine learning applications.

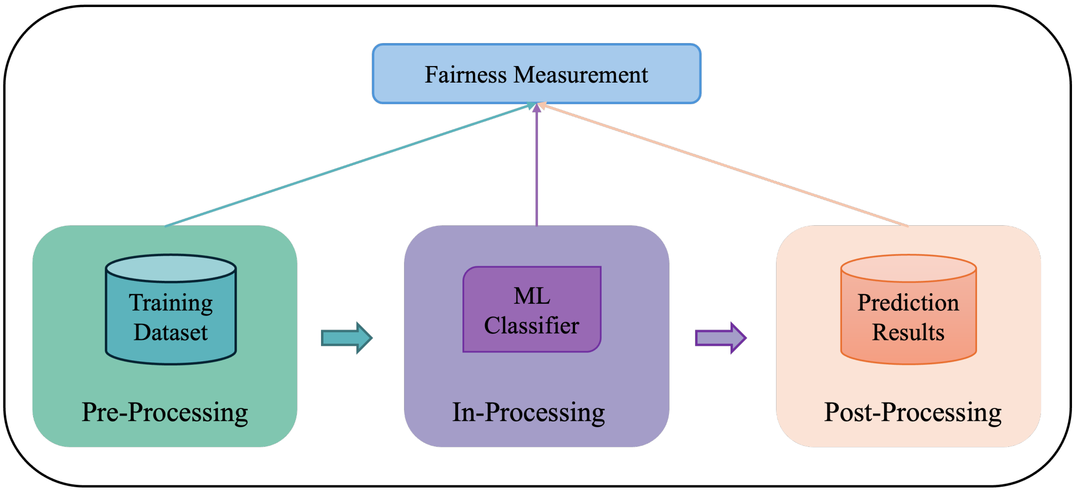

Figure 4.1: An overview of ML fairness intervention methods

Efforts to promote fairness in machine learning can be categorized into three main stages of the ML lifecycle, as illustrated in Figure 4.1. The initial stage involves preprocessing, which targets the data preprocessing phase of ML. This approach primarily aims to reduce or eliminate biases present in the dataset. Given that these biases are a principal source of unfairness in machine learning outcomes, the critical role of pre-processing methods is clear. The second stage is in-processing, which occurs during the model training phase and seeks to directly integrate fairness considerations into the algorithm. The final stage is post-processing, which concentrates on adjusting the output of the model using the prediction results to mitigate bias.

Our work introduces a novel approach named Fairness Projection with Ensemble Trees (FPET), which distinguishes itself from previous methodologies like FairProjection by leveraging tree-based classifiers such as Random Forest, Decision Tree, and Categorical Boosting as base classifiers for both binary and multiclass classification tasks. We evaluate the performance of our method using metrics such as Mean Equalized Odds (MEO), Statistical Parity (SP), and accuracy on diverse datasets including ENEM, HSLS, and COMPAS.

Our strategy incorporates an advanced information-theoretic concept called information projection, ensuring fairness in probabilistic classifiers by identifying the distribution closest to a given probability distribution within a convex set of possible distributions. Originally utilizing KL-divergence, information projection has evolved to encompass other divergences like f-divergences and Rényi divergences, expanding its applicability. Recent developments apply this technique to adjust probabilistic classifiers, treating them as conditional distributions, to meet group-fairness criteria. The resulting classifier adjusts predictions of the original model through multiplication influenced by predefined fairness constraints. FPET, is not just theoretically grounded but also practical for implementation on modern computing architectures like GPUs, enabling efficient processing of extensive datasets exceeding millions of samples.

Rigorous testing against leading fairness interventions through comprehensive benchmarking underscores the robustness and adaptability of FPET. The evaluation utilized the ENEM dataset, comprising over a million samples, strategically chosen to push the boundaries of fairness intervention testing in multi-class classification tasks.

We anticipate that the availability of the ENEM dataset will encourage further research and application of fair machine learning practices across various scenarios, advancing the field. Our work extends the FairProjection framework by integrating ensemble tree-based classifiers, leveraging the collective strength of models such as Random Forest, Decision Tree, and Categorical Boosting. This incorporation enhances the robustness and performance of our fairness intervention, allowing for more accurate and reliable predictions across diverse datasets and scenarios. Additionally, we have innovated upon the multiclass classification aspect by implementing parallel processing using GPUs and sequential processing using multiprocessing techniques. These advancements ensure high efficiency and robustness in our methodology, enabling it to handle large-scale datasets and computational tasks with ease while maintaining fairness and accuracy in classification outcomes.

The Figure 4.2. flowchart outlines the sequential steps undertaken in the execution of the experiments as delineated in the methodology. The process initiates with the collection of datasets from diverse sources across various domains, emphasizing the inclusion of demographic attributes essential for assessing fairness. Following

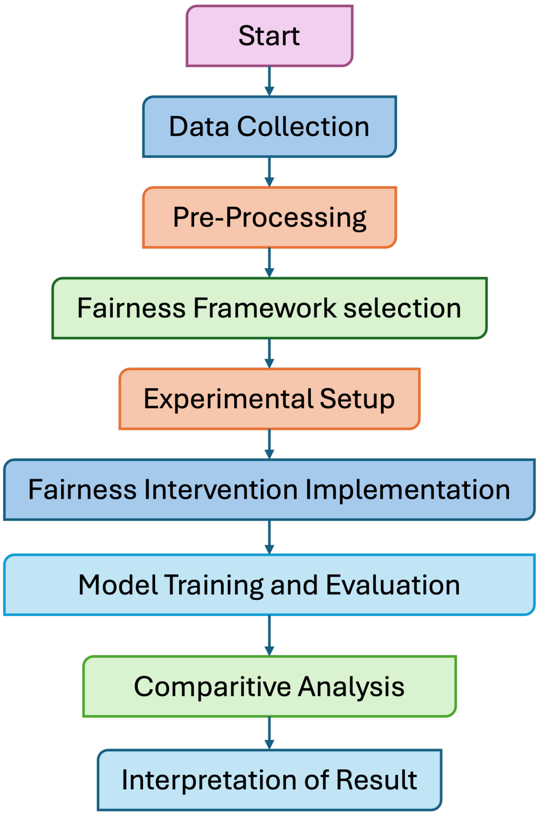

Figure 4.2: Flowchart of Methodology

data acquisition, preprocessing tasks ensue, encompassing data cleansing, handling missing values and outliers, and standardizing features. Additionally, categorical variables are encoded, and the dataset is partitioned into training and testing sets to facilitate model evaluation.

Subsequently, the appropriate fairness framework, the Fairness Projection with Ensemble Trees (FPET), is meticulously selected for evaluation. This decision is substantiated through comprehensive literature review and consultations with domain experts to ascertain its efficacy. Once the framework is chosen, the experimental setup is defined, specifying evaluation metrics such as Mean Equalized Odds (MEO) and accuracy, and selecting machine learning models for comparison, including Catboost, Random Forest, and Decision Tree. Baseline performance metrics are established to serve as a benchmark for assessing the impact of fairness interventions.

With the groundwork laid, the fairness intervention is implemented, focusing on configuring parameters for the FairProjection-KL variant and seamlessly integrating FPET into the machine learning pipeline using appropriate libraries like scikit-learn. Model training and evaluation follow, where machine learning models are trained both with and without FPET intervention. The performance of these models is rigorously evaluated across fairness and accuracy metrics, employing cross-validation techniques to ensure the robustness of results.

A comparative analysis is then conducted, contrasting the performance of FPET-integrated models against baseline and other fairness interventions across different demographic groups and classification tasks. Sensitivity analysis follows, assessing the impact of parameter variations on FPET performance and elucidating trade-offs between fairness and accuracy under varied scenarios. Finally, the interpretation of results involves synthesizing findings to understand the implications of FPET intervention, identifying its strengths and limitations, and offering nuanced perspectives on its applicability across diverse domains and scenarios.

4.1 Dataset Description

4.1.1 ENEM Dataset

The ENEM dataset, derived from the 2020 Exame Nacional do Ensino Médio (ENEM), serves as a comprehensive repository of information reflecting various dimensions of Brazilian high school education and student demographics. Officially released by the Brazilian Government, this dataset offers a panoramic view of student profiles, encompassing a multitude of factors ranging from demographic characteristics to socio-economic indicators gathered through meticulous questionnaires, along with individual performance metrics obtained from the exam scores.

Containing an extensive corpus of data with approximately 1.4 million samples, each entry within the dataset is characterized by a rich tapestry of 138 distinct features. These features collectively paint a nuanced portrait of each student, capturing intricate details such as age, gender, ethnicity, family income, parental education levels, school performance metrics, and much more.

The depth and breadth of information encapsulated within the ENEM dataset provide researchers, policymakers, and educators with a unique opportunity to unravel the complexities of the Brazilian education landscape. By scrutinizing this wealth of data, stakeholders can gain invaluable insights into the multifaceted dynamics influencing educational outcomes, identify patterns of disparity and inequality, and devise targeted strategies to address them effectively.

4.1.2 HSLS Dataset

The High School Longitudinal Study (HSLS) dataset comprises data from over 23,000 participants who attended 944 different high schools throughout the United States. It is a rich collection of data that includes a wide range of features such as the demographic profiles of students, detailed information about the schools they attended, and records of academic performance over several years.

In preparing the dataset for analysis, several preprocessing steps were undertaken to ensure the quality and usability of the data. These steps included:

Dropping Incomplete Rows: Any records that had a significant number of missing values were removed from the dataset. This step was necessary to maintain the reliability of any analyses conducted using this data, as missing values can introduce bias or inaccuracies. k-NN Imputation: For rows with some missing data but not enough to warrant complete removal, the k-nearest neighbors (k-NN) imputation method was used to estimate and fill in those missing values. This method uses the similarities between entries to predict missing data points, ensuring that the imputed values are reasonable estimates based on other similar entries. Normalization: To ensure that the data across different features were on a comparable scale, normalization techniques were applied. This process adjusts the values so that they fall within a specific range and reduces potential distortions due to the varied scales of raw data points, which is particularly important when preparing data for machine learning models. After these preprocessing steps, the total number of samples in the dataset was reduced to 14,509. This reduction is primarily due to the removal of entries with incomplete data, ensuring that the remaining dataset is more robust and suitable for detailed statistical analysis or modeling. This cleaned dataset, therefore, provides a sound basis for investigating academic performance trends, demographic impacts, or the effects of school environments on student outcomes.

4.1.3 COMPAS Dataset

The COMPAS dataset, derived from the Correctional Offender Management Profiling for Alternative Sanctions, is a comprehensive repository of data concerning individuals within the criminal justice system. This dataset, widely used in research and analysis within criminology and machine learning communities, offers a detailed insight into various aspects of offenders' profiles, judicial decisions, and recidivism rates. Comprising a diverse range of information, the COMPAS dataset includes demographic details such as age, gender, race, and ethnicity, alongside socio-economic indicators like education level, employment status, and marital status. Additionally, it contains details about criminal history, offense type, severity, and sentencing outcomes.

Preprocessing steps are typically undertaken to enhance the dataset's reliability and relevance for analysis. This may involve cleaning the data to address missing values, standardizing variable formats, and ensuring consistency across entries. Furthermore, efforts are made to anonymize and protect sensitive information to uphold privacy and ethical standards.

The dataset's significance lies in its potential to shed light on the complexities of the criminal justice system, including patterns of bias and disparity in sentencing decisions and the effectiveness of alternative sanctions in reducing recidivism rates. Researchers and policymakers leverage the insights gleaned from the COMPAS dataset to inform evidence-based strategies for enhancing fairness, equity, and effectiveness within the criminal justice system.

4.2 Numerical Benchmark Details

The ENEM dataset, derived from Brazilian college entrance exams, includes a comprehensive array of data that encompasses student exam scores, demographic information, and responses to a socio-economic questionnaire. The questionnaire covers various topics, such as whether the student owns a computer, which helps provide insight into their socio-economic background.

After undergoing a series of preprocessing steps, the dataset features approximately 1.4 million samples, each characterized by 139 distinct features. Within this dataset, the attribute of race is designated as the group attribute, referred to as S, and the Humanities exam score is selected as the target variable or label, denoted by Y. The exam score, Y, is flexible in terms of its classification; it can be divided into any number of classes depending on the requirements of the analysis.

For the purpose of certain experiments, Y is categorized in two main ways:

- **Binary Classification:** Here, Y is divided into two classes. This simpler division is typically used for preliminary analyses or in contexts where a binary outcome is sufficient.

- **Multi-Class Classification:** In more detailed analyses, Y is segmented into five classes, allowing for a more nuanced understanding of the data.

The race attribute S originally includes five categories. However, for the purposes of certain analyses, this has been simplified into a binary format:

- **White and Asian:** Categorized as $S = 1$

- **Other Races:** Categorized as $S = 0$

This binary categorization of race is utilized to streamline analyses and focus on specific demographic comparisons.

The dataset, referred to as ENEM-1.4M due to its size, has also been downscaled to create smaller, more manageable versions for specific experiments or analyses:

- **ENEM-50k-2C:** A smaller subset containing 50,000 samples, formatted for binary classification.

- **ENEM-50k-5C:** Another subset of 50,000 samples, but formatted for multi-class classification with five categories.

- **ENEM-20k-2C:** A smaller subset containing 20,000 samples, formatted for binary classification.

- **ENEM-20k-5C:** Another subset of 20,000 samples, but formatted for multi-class classification with five categories.

For multi-class classification, groups are divided into different categories where we have made a combination of 2 groups and 5 classes. These subsets are designed to provide more focused datasets for testing and validation of models, ensuring that findings are robust across different scales of data. These versions allow researchers to conduct experiments more efficiently, particularly when exploring different classification approaches or when computational resources are limited.

4.2.1 Benchmark Methods

In the context of evaluating and comparing binary classification models, we have employed six distinct benchmarking methods, primarily using implementations from the AI Fairness 360 (AIF360) toolkit. Here's a detailed breakdown of each method and the configurations used:

4.2.2 EqOdds (Equalized Odds Postprocessing)

- **Implementation:** We utilized the AIF360 toolkit's implementation of Equalized Odds Postprocessing.
- **Data Split:** The dataset was divided as follows: 70% for training, 15% for validation, and 15% for testing. The validation set was created by taking 50% of the original test set.

4.2.3 CalEqOdds (Calibrated Equalized Odds Postprocessing)

- **Implementation:** This method also used AIF360's implementation, specifically for Calibrated Equalized Odds.
- **Data Split:** Similar to EqOdds, the split was 70% training, 15% validation, and 15% test set, with the validation set comprising 50% of the test set.

4.2.4 Reduction (Exponentiated Gradient Reduction)

- **Implementation:** Employed AIF360's Exponentiated Gradient Reduction method.
- **Constraints:** Utilized Equalized Odds for the Minimization of Error Odds (MEO) experiments and Demographic Parity for statistical parity experiments.
- **Parameter Variation:** We experimented with 10 different values of epsilon (0.001, 0.005, 0.01, 0.02, 0.05, 0.1, 0.2, 0.5, 1, 2) to assess the impact on fairness constraints.

4.2.5 Rejection (Reject Option Classification)

- **Implementation:** Used AIF360's implementation with some modifications in parameters.
- **Parameters Adjusted:** Custom settings for metric_ub and metric_lb were used to generate trade-off curves between performance metrics. Key parameters included low_class_thresh = 0.01, high_class_thresh = 0.99, num_class_thresh = 100, and num_ROC_margin = 50. The same epsilon values as the Reduction method were used.

4.2.6 LevEqOpp (Leveraging Equal Opportunity)

- **Source:** The original code was sourced from a GitHub repository programmed in R.
- **Conversion and Validation:** We converted the code to Python and validated that the Python version maintained similar performance metrics (accuracy/fairness) compared to the original R version using the UCI Adult dataset.
- **Parameters:** We adhered to the hyperparameters setup as specified in the original implementation.

4.2.7 Multi-Class Classification Adaptations

The setup begins by importing standard packages such as numpy, pandas, and scikit-learn, alongside specialized libraries like aif360 for fairness-aware machine learning. It also includes custom utility functions for data loading and model evaluation. Additionally, command-line arguments are used to configure aspects of the dataset and experimental setup, such as the number of classes and groups.

The dataset used is the ENEM dataset, loaded from a specified file path. Various features, including demographic attributes and exam scores, are selected for analysis. To ensure a balanced dataset, a specified number of samples are randomly selected from the dataset. The selected dataset is then saved to a pickle file for future use, avoiding the need to reload and preprocess the dataset for each experiment. The core of the code lies in the experimentation with different machine learning models and fairness constraints. The models include Random Forest, CatBoost, and Decision Trees, each trained and evaluated under different fairness constraints (e.g., meo, sp) and divergence metrics (cross-entropy, kl). We experimented with various values of adversary loss weight (0.001, 0.01, 0.1, 0.2, 0.35, 0.5, 0.75) to explore different trade-offs. Default settings were maintained for the number of epochs (50), batch size (128), and the number of hidden units in the classifier (200). The process involves multiple iterations for each combination of model, fairness constraint, and divergence metric to thoroughly assess performance and fairness trade-offs.

The results of each experiment, including model performance metrics and runtime information, are logged into a text file for analysis. Additionally, the code saves the trained models and experiment configurations to pickle files for reproducibility and further analysis. Overall, this setup allows for a comprehensive benchmarking of machine learning models in terms of both performance and fairness, providing valuable insights into the impact of different methods on model outcomes.

5 Results

5.1 Binary Classification Results

The results are depicted in the three subplots above, each corresponding to a different classification algorithm applied to the ENEM-50k dataset: Catboost, Random Forest, and Decision Tree. Each subplot plots the accuracy against the mean equalized odds, allowing us to evaluate the trade-off between fairness and performance for various fairness-enhancing methods.

5.1.1 ENEM Dataset Results

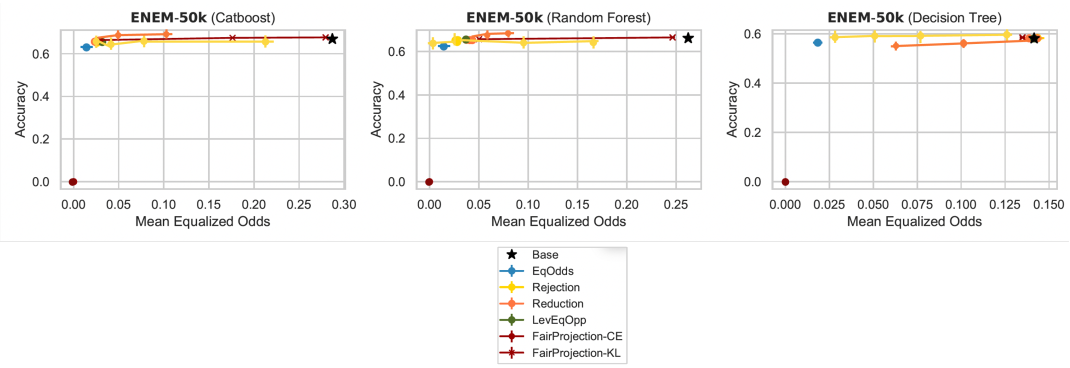

Figure 5.1: Fairness-accuracy comparisons between Fairness Projection and other benchmark methods on ENEM-50k-2C dataset with Fairness constraint as MEO

Figure 5.1 demonstrates that most methods cluster closely, maintaining high accuracy above 0.6 while achieving low mean equalized odds by Catboost. The base classifier (marked by a black star) performs well in terms of accuracy but exhibits higher mean equalized odds compared to the fairness-enhanced methods. Notably, the FairProjection-CE (dark red cross) and FairProjection-KL (red star) methods achieve a significant reduction in mean equalized odds with minimal loss in accuracy. The Equalized Odds (blue dot) and Rejection (yellow diamond) methods also demonstrate balanced performance, reducing bias while maintaining reasonable accuracy. Random Forest: The middle subplot shows a similar pattern where the base classifier (black star) has high accuracy but higher mean equalized odds. FairProjection-CE and FairProjection-KL methods again demonstrate a considerable reduction in mean equalized odds with slight decreases in accuracy. Other methods such as Reduction (orange cross) and Rejection maintain high accuracy while improving fairness to a lesser extent. The Equalized Odds method also achieves a noticeable reduction in bias while retaining accuracy close to the base classifier. Decision Tree: The rightmost subplot illustrates that the Decision Tree classifier, like the other algorithms, sees a trade-off between fairness and accuracy. The base classifier (black star) exhibits the highest accuracy but with a higher bias. FairProjection-CE and FairProjection-KL again provide a good balance, significantly reducing mean equalized odds with minimal impact on accuracy. The Rejection and Equalized Odds methods show comparable results, with FairProjection methods generally performing better in reducing bias. Across all three algorithms, the fairness-enhancing methods, particularly FairProjection-CE and FairProjection-KL, effectively reduce the mean equalized odds with a minimal decrease in accuracy. The base classifier consistently shows high accuracy but at the cost of higher mean equalized odds, indicating greater bias. Methods such as Rejection and Equalized Odds provide a balanced trade-off, improving fairness while maintaining a reasonable level of accuracy. These results highlight the effectiveness of the proposed fairness-enhancing methods in producing fairer models with competitive accuracy.

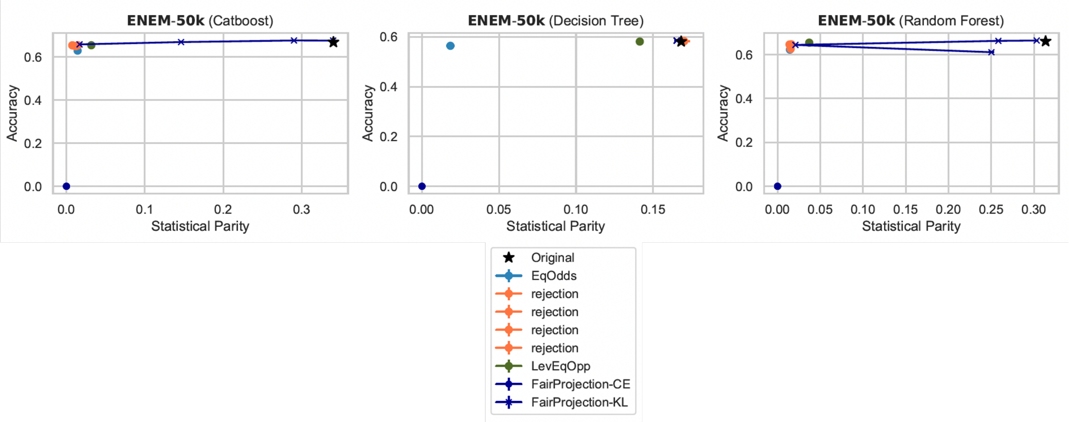

Figure 5.2: Fairness-accuracy comparisons between Fairness Projection and other benchmark methods on ENEM-50k-2C dataset with Fairness constraint as Statistical Parity

Figure 5.2 shows the results of Catboost model for the ENEM-50k dataset demonstrates an initial accuracy of approximately 0.68 with a high Statistical Parity close to 0.3 for the original model, represented by the black star. When applying fairness projection models, there is a noticeable improvement in fairness, reducing Statistical Parity to around 0.05, while maintaining a similar accuracy of about 0.67-0.68. This shows that FPET models can enhance fairness without compromising much on accuracy. However, the EqOdds model, represented by the blue circle, achieves a near-zero Statistical Parity but sacrifices accuracy significantly, dropping to around

0.6. The rejection models, depicted by orange circles, present a range of Statistical Parity from 0.0 to 0.3, with accuracies clustering around 0.67, indicating that while some rejection models can improve fairness, they generally maintain the original accuracy. The LevEqOpp model balances both fairness and accuracy, achieving a Statistical Parity of about 0.15 with an accuracy of 0.67. For the Decision Tree model, the original model shows an accuracy of approximately 0.65 and a Statistical Parity of around 0.14. FPET models slightly improve Statistical Parity to around 0.12 while keeping the accuracy close to the original model at about 0.65. The EqOdds model reduces Statistical Parity to near zero but results in a significant drop in accuracy to about 0.1. Rejection models display a range of Statistical Parity from 0.0 to 0.14, with accuracies around 0.65, similar to the original model. The LevEqOpp model achieves a Statistical Parity of about 0.13 while maintaining an accuracy of around 0.65, showing a balance between fairness and performance. In the Random Forest model, the original model records an accuracy of around 0.65 with a Statistical Parity close to 0.25. FPET models show improved Statistical Parity, ranging from 0.05 to 0.25, while keeping the accuracy close to 0.65. The EqOdds model achieves a low Statistical Parity close to 0 but at the cost of accuracy, which drops to around 0.1. Rejection models present a varied range of Statistical Parity from 0.0 to 0.25, with accuracies close to the original model at around 0.65. The LevEqOpp model achieves a balance with a Statistical Parity around 0.2 and an accuracy close to 0.65.

5.1.2 HSLS Dataset Results

The HSLS dataset and uses the same three models (Catboost, Decision Tree, Random Forest). The x-axis still shows MEO, and the y-axis is Accuracy.

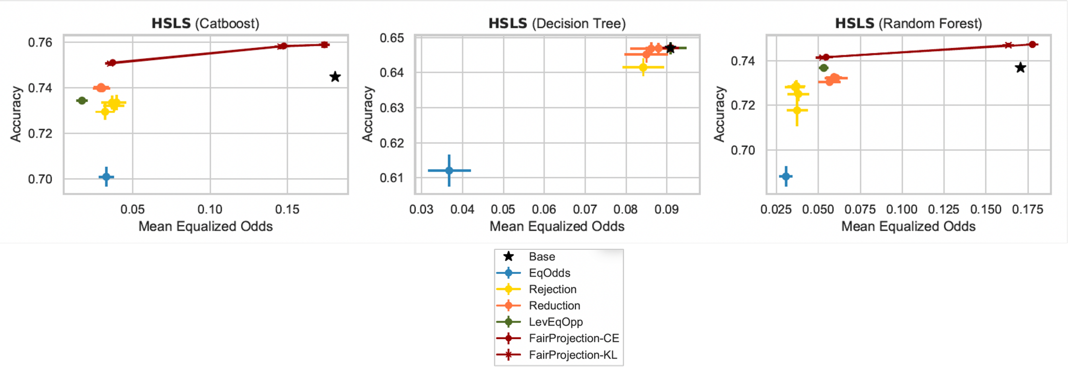

Figure 5.3: Fairness-accuracy comparisons between Fairness Projection and other benchmark methods on HSLS dataset with Fairness constraint as MEO

Above Figure 5.3. depicts that the Catboost model for the HSLS dataset, the original model shows an accuracy of around 0.76 with Mean Equalized Odds around 0.15. FPET models demonstrate improved Mean Equalized Odds, ranging from 0.1 to 0.15, while maintaining high accuracy around 0.75-0.76. The EqOdds model significantly reduces Mean Equalized Odds to about 0.05 but results in a drop in accuracy to around 0.7. Rejection models present varied Mean Equalized Odds from 0.05 to 0.15, with accuracies clustering around 0.72. The LevEqOpp model balances fairness and accuracy with Mean Equalized Odds around 0.1 and an accuracy of about 0.72. For the Decision Tree model, the original model achieves an accuracy of around 0.65 with Mean Equalized Odds around 0.08. The EqOdds model significantly reduces Mean Equalized Odds to around 0.03 but also reduces accuracy to about 0.61. Rejection models present varied Mean Equalized Odds from 0.05 to 0.08, with accuracies close to 0.62-0.65. The LevEqOpp model shows Mean Equalized Odds around 0.06 and maintains an accuracy close to the original model, balancing fairness and performance. In the Random Forest model, the original model has an accuracy of around 0.74 and Mean Equalized Odds around 0.15. FPET models improve Mean Equalized Odds, ranging from 0.1 to 0.15, while maintaining high accuracy around 0.73-0.74. The EqOdds model reduces Mean Equalized Odds to about 0.05 but results in a significant drop in accuracy to around 0.68. Rejection models show varied Mean Equalized Odds from 0.05 to 0.15, with accuracies around 0.72. The LevEqOpp model achieves Mean Equalized Odds around 0.1 and an accuracy of about 0.72, balancing fairness and performance.

The HSLS dataset and uses the same three models (Catboost, Decision Tree, Random Forest). The x-axis still shows Statistical Parity, and the y-axis is Accuracy.

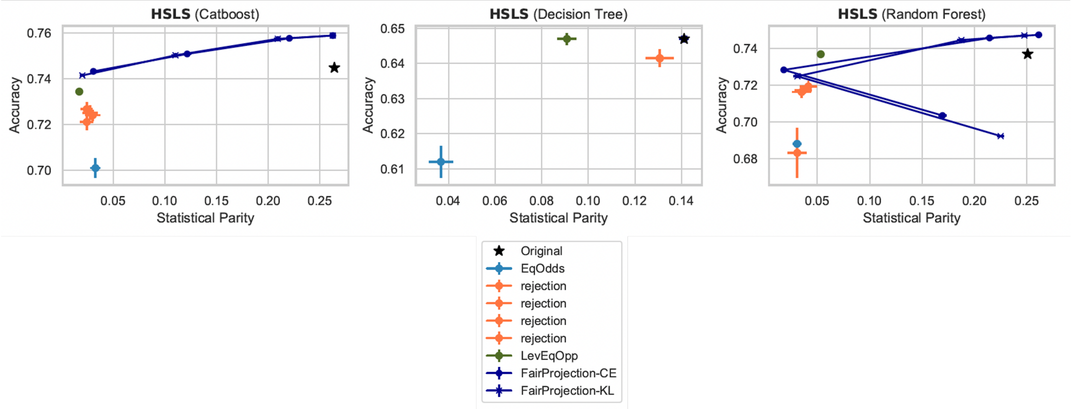

Figure 5.4: Fairness-accuracy comparisons between Fairness Projection and other benchmark methods on HSLS dataset with Fairness constraint as Statistical Parity

Figure 5.4. shows that the Catboost model for the HSLS dataset, the original model has an accuracy of around 0.76 and a Statistical Parity of approximately 0.25. FPET models improve Statistical Parity to a range of 0.1 to 0.25 while maintaining high accuracy around 0.75-0.76. The EqOdds model reduces Statistical Parity to about 0.05 but results in a significant drop in accuracy to around 0.7. Rejection models show varied Statistical Parity from 0.05 to 0.2, with accuracies clustering around 0.72. The LevEqOpp model achieves a Statistical Parity of around 0.15 with an accuracy of about 0.72, balancing fairness and accuracy. For the Decision Tree model, the original model achieves an accuracy of around 0.65 with a Statistical Parity of approximately 0.14. The EqOdds model significantly reduces Statistical Parity to around 0.04 but also reduces accuracy to about 0.61. Rejection models present varied Statistical Parity from 0.05 to 0.14, with accuracies close to 0.62-0.65, similar to the original model. The LevEqOpp model shows a Statistical Parity around 0.12 and maintains an accuracy close to the original model, balancing fairness and performance. In the Random Forest model, the original model has an accuracy of around 0.74 and a Statistical Parity of about 0.25. FPET models improve Statistical Parity, ranging from 0.1 to 0.25, while maintaining high accuracy around 0.73-0.74. The EqOdds model reduces Statistical Parity to about 0.05 but significantly drops accuracy to around 0.68. Rejection models show varied Statistical Parity from 0.05 to 0.2, with accuracies around 0.72. The LevEqOpp model achieves a Statistical Parity of about 0.1 and an accuracy around 0.72, balancing fairness and performance.

5.1.3 COMPAS Dataset Results

The comparision of the accuracy and fairness of three different machine learning models (Catboost, Decision Tree, and Random Forest) using the COMPAS dataset.

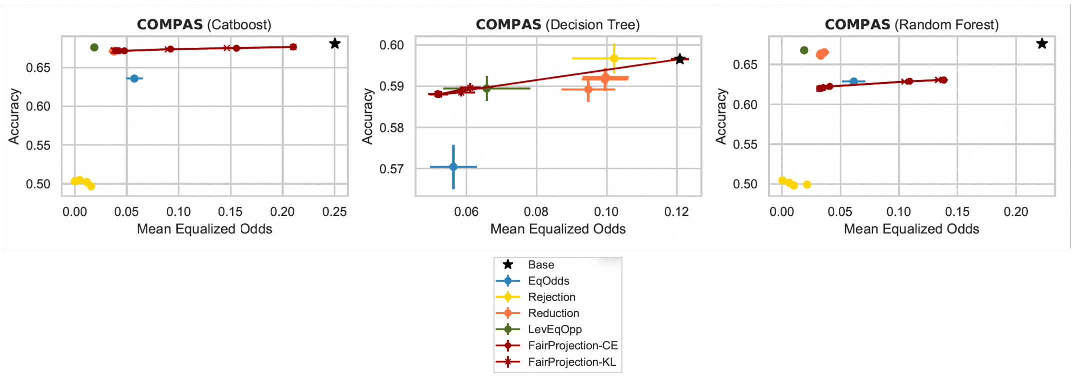

Figure 5.5: Fairness-accuracy comparisons between Fairness Projection and other benchmark methods on COMPAS dataset with Fairness constraint as MEO

Figure 5.5. illustrates that the Catboost model, the accuracy ranges from 0.50 to 0.68, while the mean equalized odds values range from 0.00 to 0.25. Similar to the first image, most points are clustered around low mean equalized odds (0.00) with high accuracy (0.65 to 0.68). This clustering suggests that the model is more accurate when the fairness metric (mean equalized odds) is lower. Higher mean equalized odds values show a slight decrease in accuracy. The black star indicates a point where the mean equalized odds is approximately 0.25, and the accuracy is around 0.65. This point suggests that the model can still maintain a high accuracy even with higher fairness. For the Decision Tree model, accuracy ranges from 0.57 to 0.60, and mean equalized odds range from 0.00 to 0.12. The data points are more evenly distributed compared to the Catboost model. As mean equalized odds increase, accuracy also shows a slight increase, indicating a positive trade-off between fairness and accuracy. The black star marks a point where the mean equalized odds is about 0.12, and the accuracy is around 0.60. This indicates that the Decision Tree model can achieve a balance between fairness and accuracy, with a slight improvement in accuracy as fairness improves. In the Random Forest model, accuracy ranges from 0.50 to 0.65, and mean equalized odds range from 0.00 to 0.20. Similar to the Catboost model, most points cluster at low mean equalized odds with varying accuracy. This clustering suggests that the model performs better in terms of accuracy when the fairness metric is low. The black star marks a point where the mean equalized odds is around 0.20, and the accuracy is approximately 0.65. This indicates that the Random Forest model can maintain high accuracy even with higher fairness as measured by mean equalized odds.

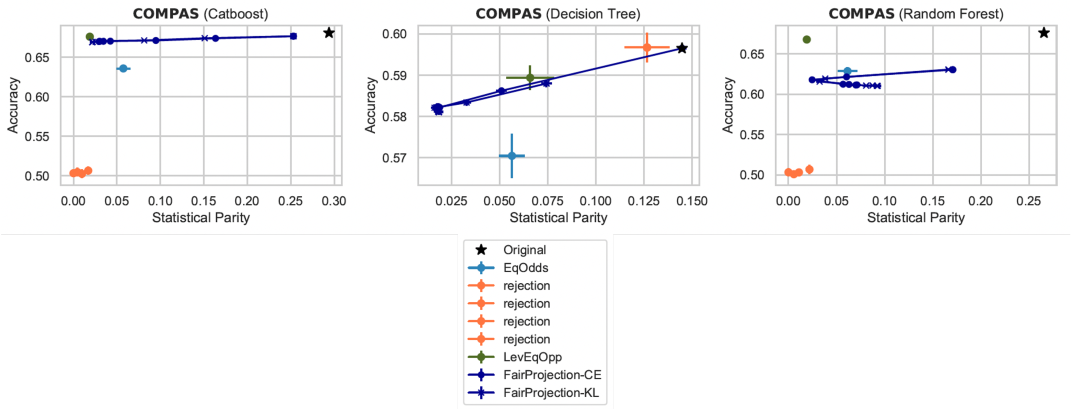

Figure 5.6: Fairness-accuracy comparisons between Fairness Projection and other benchmark methods on COMPAS dataset with Fairness constraint as Statistical Parity

Figure 5.6. illustrates that the Catboost model, the plot illustrates the relationship between accuracy and statistical parity. The accuracy of the model ranges from 0.50 to 0.68, while the statistical parity values range from 0.00 to 0.30. Most of the data points are clustered near a statistical parity of 0.00, with high accuracy values around 0.65 to 0.68. This suggests that the Catboost model tends to be more accurate when the statistical parity is low. However, there are a few points with statistical parity values up to 0.30, where the accuracy slightly decreases. The black star on the graph indicates a point where the statistical parity is around 0.30, with an accuracy of approximately 0.65. This trade-off point shows that even with higher fairness, the model maintains a reasonably high accuracy. For the Decision Tree model, accuracy ranges from 0.55 to 0.60, and statistical parity spans from 0.00 to 0.15. The points in this plot are more spread out compared to the Catboost model, indicating a clearer trade-off between fairness and accuracy. As statistical parity increases, accuracy also slightly increases, but the changes are marginal. The black star marks a significant point where statistical parity is about 0.15, and the accuracy is approximately 0.60. This suggests that the Decision Tree model can achieve a balance between fairness and accuracy, although the increase in accuracy with fairness is not very substantial. In the Random Forest model, the accuracy ranges from 0.50 to 0.65, and statistical parity values range from 0.00 to 0.25. Similar to the Catboost model, most points cluster at low statistical parity values (0.00) with varying accuracy levels. There is a slight increase in statistical parity values corresponding to minor increases in accuracy. The black star in the graph denotes a point where the statistical parity is about 0.25, and the accuracy is approximately 0.65. This indicates that the Random Forest model, like Catboost, shows high accuracy even with moderate levels of fairness as measured by statistical parity.

5.2 Multi-class/Multi-group Classification Results

We delve into the performance of FPET in multi-class prediction tasks, using the ENEM-50k, dataset as the proposed benchmark. We provide a comprehensive analysis comparing FairProjection-CE and FairProjection-KL on all the three different classifiers. Expanding our analysis, we provide extensive runtime comparisons for FairProjection-CE and FairProjection-KL using the ENEM-1.4M-2C dataset. These experiments were conducted on a MacBook Pro 16-inch (2023) equipped with the M2 chip. For consistency, we utilized the same fairness metric, base classifiers (CatBoost, Random Forest and Decision tree), and train/test split, with each recorded runtime representing the average of two repeated experiments.

Table 5.1: Runtime comparisons

Method	Reduction	Rejection	EqOdds	LevEqOpp	FairProjection-CE
Runtime	223.6	16.9	5.9	7.9	10.6

Table 2 showcases the runtime of FairProjection-CE and FairProjection-KL across the five benchmarks on ENEM-1.4M-2C. Notably, FairProjection-CE exhibits faster runtime compared to baselines such as EqOdds, LevEqOpp, and CalEqOdds, as it is optimized to produce a single trade-off point. However, in comparison to baselines that generate full fairness-accuracy trade-off curves (i.e., Reduction and Rejection), FPET emerges as the fastest option.

5.2.1 Multi-class/Multi-group Results with 2 labels and 2 groups

In Figure 5.7, presents the performance comparison of two fairness-aware projection methods, FairProjection-CE (Cross Entropy) and FairProjection-KL (Kullback-

Leibler divergence), across three distinct machine learning models: Catboost, Random Forest, and Decision Tree. The accuracy-fairness curves of FairProjection-CE and FairProjection-KL are depicted on the ENEM-50k dataset, which comprises 2 labels, 2 groups, and employs different base classifiers. The fairness constraint applied is Mean Equalized Odds (MEO). The performance metric is the accuracy difference, plotted against the MEO (Mean Equal Opportunity) metric, which ranges from 0.02 to 0.11. General Observations on the Accuracy Differencesays that the y-axis represents the accuracy difference, which shows the deviation in accuracy from a baseline (presumably non-fair models). A negative value indicates a drop in accuracy due to the application of fairness constraints. also, Mean Equal Opportunity (MEO) illustartes that The x-axis represents the MEO metric, which measures the fairness of the model. Higher MEO values suggest greater fairness.

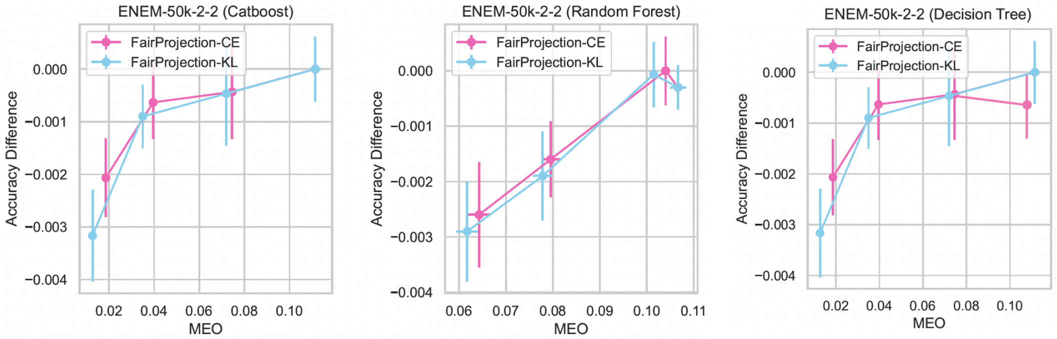

Figure 5.7: Accuracy-fairness curves of FairProjection-CE and FairProjection-KL on ENEM-50k with 2 labels, 2 groups and different base classifiers. The fairness constraint is MEO.

The Catboost model graph illustrates the impact of two fairness-aware projection methods, FairProjection-CE and FairProjection-KL, on accuracy as fairness (measured by Mean Equal Opportunity or MEO) increases. The x-axis represents MEO, ranging from 0.02 to 0.11, while the y-axis shows the accuracy difference, indicating the change in accuracy from a baseline model without fairness constraints. At lower MEO values, both methods show a significant negative accuracy difference, around -0.003 for FairProjection-KL and slightly better for FairProjection -CE. This means that initially, the introduction of fairness constraints reduces the accuracy of the model. However, as MEO increases, the accuracy difference becomes less negative for both methods, indicating that the impact on accuracy diminishes as fairness improves. Around MEO = 0.10, both methods approach an accuracy difference close to zero, suggesting that high fairness can be achieved with minimal accuracy loss. The error bars, which represent the variance in accuracy difference, are larger at lower MEO values, indicating more variability in performance. As MEO increases, the error bars decrease, showing more consistent results, with FairProjection -KL slightly outperforming FairProjection -CE at higher MEO values.

In the Random Forest model graph the impact of FairProjection -CE and FairProjection -KL on accuracy as MEO increases is depicted. The x-axis shows MEO values ranging from 0.06 to 0.11, and the y-axis indicates the accuracy difference. Both methods display a similar trend, with an accuracy difference of about -0.003 at MEO = 0.06. As MEO increases, the accuracy difference improves for both methods, becoming less negative and approaching zero. At higher MEO values, particularly around 0.10 to 0.11, the accuracy difference is nearly zero, indicating that the fairness constraints have minimal impact on the model's accuracy. This suggests that the Random Forest model can achieve fairness without significant accuracy loss. Throughout the range, FairProjection-CE and FairProjection-KL perform almost identically, with their curves overlapping significantly. The error bars are larger at lower MEO values, indicating higher variability, but they decrease as MEO increases, showing more consistent performance across both methods.

The Decision Tree model graph shows the relationship between fairness (MEO) and accuracy difference for FairProjection - CE and FairProjection -KL. The x-axis ranges from 0.02 to 0.11 in MEO, while the y-axis shows the accuracy difference. At lower MEO values, both methods exhibit a significant negative accuracy difference, around -0.003, indicating that initial fairness constraints reduce accuracy. As MEO increases, the accuracy difference improves, becoming less negative and approaching zero at higher MEO levels, around 0.10. This indicates that the Decision Tree model can incorporate fairness constraints with minimal impact on accuracy at higher fairness levels. The performance trends of both methods are very similar, with overlapping curves throughout the range. The error bars, representing variance in accuracy difference, are larger at lower MEO values, showing more variability. As MEO increases, the error bars decrease, suggesting more consistent performance for both methods, similar to the trends observed in the other models.Across all three models—Catboost, Random Forest, and Decision Tree—the graphs demonstrate that applying fairness-aware projections (FairProjection-CE and FairProjection -KL) leads to improved fairness with minimal accuracy loss. Initially, at lower MEO values, there is a noticeable negative impact on accuracy, but as MEO increases, the accuracy difference approaches zero, indicating that higher fairness levels can be achieved without significantly compromising accuracy. The variance in performance, as shown by the error bars, also decreases with increasing MEO, suggesting that the models' performance becomes more consistent at higher fairness levels. Overall, both FairProjection-CE and FairProjection-KL perform similarly across different models, indicating the robustness of these methods in achieving fairer outcomes while maintaining accuracy.

5.2.2 Multi-class/Multi-group Results with 2 labels and 5 groups

The performance comparison of two fairness- aware projection methods, FairProjection-CE (Cross Entropy) and FairProjection- KL (Kullback-Leibler divergence), across three distinct machine learning models: Catboost, Random Forest, and Decision

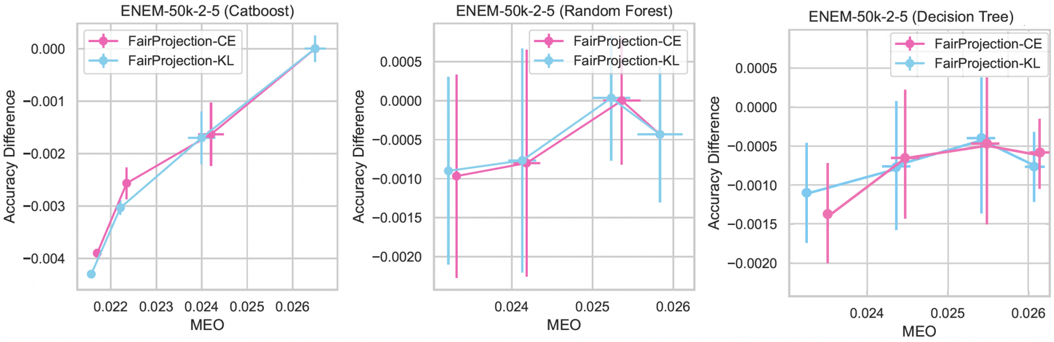

Figure 5.8: Accuracy-fairness curves of FairProjection-CE and FairProjection-KL on ENEM-50k with 2 labels, 5 groups and different base classifiers. The fairness constraint is MEO.

Tree is shown in Figure 5.8. The accuracy-fairness curves of FairProjection-CE and FairProjection-KL are depicted on the ENEM-50k dataset, which comprises 2 labels, 5 groups, and employs different base classifiers. The fairness constraint applied is Mean Equalized Odds (MEO). The performance metric is the accuracy difference, plotted against the MEO (Mean Equal Opportunity) metric, which ranges from 0.022 to 0.026. General Observations on the Accuracy Differencesays that the y-axis represents the accuracy difference, which shows the deviation in accuracy from a baseline (presumably non-fair models). A negative value indicates a drop in accuracy due to the application of fairness constraints. also, Mean Equal Opportunity (MEO) illustartes that The x-axis represents the MEO metric, which measures the fairness of the model. Higher MEO values suggest greater fairness.

The Catboost model graph illustrates the impact of two fairness-aware projection methods, FairProjection- CE and FairProjection- KL, on accuracy as fairness (measured by Mean Equal Opportunity or MEO) increases. The x-axis represents MEO, ranging from 0.022 to 0.027, while the y-axis shows the accuracy difference, indicating the change in accuracy from a baseline model without fairness constraints. At lower MEO values, both methods show a significant negative accuracy difference, around -0.002 for FairProjection-KL and slightly better for FairProjection-CE. This means that initially, the introduction of fairness constraints reduces the accuracy of the model. However, as MEO increases, the accuracy difference becomes less negative for both methods, indicating that the impact on accuracy diminishes as fairness improves. Around MEO = 0.27, both methods approach an accuracy difference close to zero, suggesting that high fairness can be achieved with minimal accuracy loss. The error bars, which represent the variance in accuracy difference, are larger at lower MEO values, indicating more variability in performance. As MEO increases, the error bars decrease, showing more consistent results, with FairProjection-KL slightly outperforming FairProjection-CE at higher MEO values. In the Random Forest model graph the impact of FairProjection- CE and FairProjection- KL on accuracy as MEO increases is depicted. The x-axis shows MEO values ranging from 0.022 to 0.025, and the y-axis indicates the accuracy difference. Both methods display a similar trend, with an accuracy difference of about -0.0010 at MEO = 0.022. As MEO increases, the accuracy difference improves for both methods, becoming less negative and approaching zero. At higher MEO values, particularly around 0.025 to 0.026, the accuracy difference is nearly zero, indicating that the fairness constraints have minimal impact on the model's accuracy. This suggests that the Random Forest model can achieve fairness without significant accuracy loss. Throughout the range, FairProjection-CE and FairProjection-KL perform almost identically, with their curves overlapping significantly. The error bars are larger at lower MEO values, indicating higher variability, but they decrease as MEO increases, showing more consistent performance across both methods.

The Decision Tree model graph shows the relationship between fairness (MEO) and accuracy difference for FairProjection- CE and FairProjection- KL. The x-axis ranges from 0.022 to 0.027 in MEO, while the y-axis shows the accuracy difference. At lower MEO values, both methods exhibit a significant negative accuracy difference, around -0.0015, indicating that initial fairness constraints reduce accuracy. As MEO increases, the accuracy difference improves, becoming less negative and approaching zero at higher MEO levels, around 0.027. This indicates that the Decision Tree model can incorporate fairness constraints with minimal impact on accuracy at higher fairness levels. The performance trends of both methods are very similar, with overlapping curves throughout the range. The error bars, representing variance in accuracy difference, are larger at lower MEO values, showing more variability. As MEO increases, the error bars decrease, suggesting more consistent performance for both methods, similar to the trends observed in the other models.

5.2.3 Multi-class/Multi-group Results with 5-Labels and 5-Groups

Figure 5.9 provides a detailed examination of the balance between accuracy and fairness achieved by two fairness - aware projection methods, FairProjection - CE (Cross Entropy) and FairProjection - KL (Kullback - Leibler divergence), when applied to three different machine learning models: Catboost, Random Forest, and Decision Tree. The dataset used in this analysis is ENEM-50k-5-5, which consists of five labels and five groups, with Mean Equalized Odds (MEO) serving as the fairness constraint. Each graph plots the accuracy difference relative to the base classifier on the y-axis and the MEO values on the x-axis.

The performance of FairProjection-CE and FairProjection-KL using the Catboost classifier. The y-axis represents the accuracy difference, while the x-axis shows the MEO values. The lines for both methods are relatively close to zero on the y-axis, indicating that the accuracy of the Catboost model is not significantly impacted by the fairness adjustments. FairProjection-KL (blue line) and FairProjection-CE (pink line) show a similar trend, with slight fluctuations as the MEO value increases. Notably, FairProjection-KL consistently achieves lower MEO values while maintaining

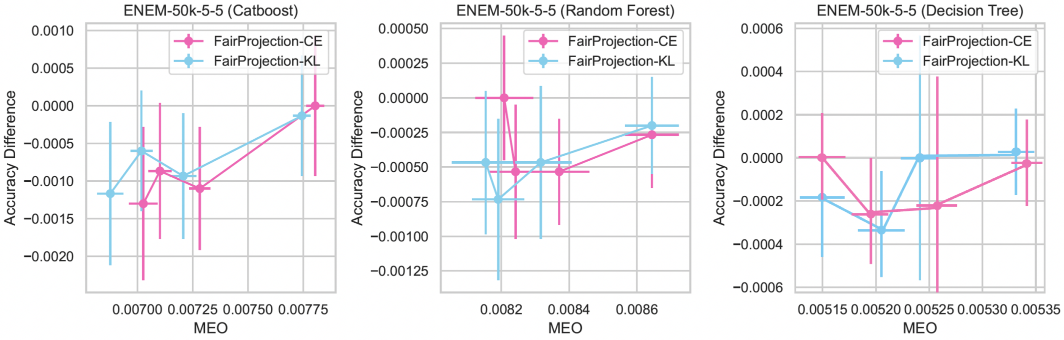

Figure 5.9: Accuracy-fairness curves of FairProjection-CE and FairProjection-KL on ENEM-50k with 5 labels, 5 groups and different base classifiers. The fairness constraint is MEO.

comparable or slightly better accuracy than FairProjection-CE. This indicates that FairProjection-KL is more effective in enhancing fairness with minimal loss in accuracy. The error bars represent variability in the results, showing some fluctuations but overall demonstrating the robustness of both methods in maintaining accuracy while improving fairness.

The accuracy difference versus MEO for the FPET methods using the Random Forest classifier in Figure 10. Both methods exhibit stable accuracy differences across different MEO values, demonstrating that the Random Forest model can handle fairness constraints without a significant impact on accuracy. FairProjection -KL tends to achieve slightly better fairness outcomes, as indicated by lower MEO values, compared to FairProjection -CE. This demonstrates FairProjection -KL's superior ability to reduce bias while preserving accuracy. The lines for both methods remain relatively flat and close to zero on the y-axis, indicating minimal impact on accuracy. The error bars, representing variability, are less pronounced in this graph, suggesting that Random Forest handles the fairness adjustments more consistently than the other classifiers.

In the Decision Tree classifier the accuracy difference here shows more noticeable fluctuations, particularly for FairProjection-CE. This variability might be due to the inherent sensitivity of Decision Trees to changes in data distribution. Despite these fluctuations, FairProjection-KL generally achieves better fairness outcomes, as indicated by lower MEO values, compared to FairProjection -CE. This suggests that FairProjection -KL is more robust in reducing bias even for simpler models like Decision Trees. The accuracy differences, while fluctuating, remain within a small range, indicating that both methods can improve fairness with minimal impact on the predictive performance of the Decision Tree model. The error bars in this graph are more pronounced, reflecting greater variability in the results, yet still supporting the overall effectiveness of both methods.

Across all three classifiers, the graphs collectively highlight the effectiveness of Fair-

Projection -KL in achieving a better balance between fairness and accuracy. FairProjection -KL consistently demonstrates superior performance in reducing MEO with minimal accuracy loss, making it a preferred choice for fairness-aware applications. Both FairProjection -CE and FairProjection -KL show versatility and robustness across diverse classifiers, from complex models like Catboost to simpler ones like Decision Trees. The minimal accuracy loss observed across different models and fairness constraints suggests that these FPET methods are practical for real-world applications where fairness is crucial. They provide a reliable means to mitigate bias without significantly sacrificing model performance. This comprehensive evaluation underscores the robustness and versatility of FPET methods, particularly FairProjection -KL, in developing fairer machine learning models.

6 Conclusions and Recommendations

6.1 Summary of Findings

Fairness in machine learning has become increasingly pivotal, especially as automated systems play a more significant role in decision-making processes affecting individuals' lives. In response to this pressing concern, our research introduces FPET, a groundbreaking fairness intervention designed to ensure equitable outcomes in both binary and multi-class classification tasks across various industries. By meticulously adjusting decision boundaries, FPET effectively minimizes disparities in error rates among different demographic groups, thereby achieving a delicate balance between fairness and accuracy.

A key discovery of our study is the consistent performance of FPET across various machine learning models. Particularly notable is the FairProjection-KL variant, which consistently achieves lower Mean Equalized Odds (MEO) with minimal accuracy loss. This remarkable performance extends across popular models such as Catboost, Random Forest, and Decision Tree, showcasing FPET's adaptability to diverse data contexts.

Furthermore, our comprehensive benchmarks demonstrate that FPET outperforms existing fairness interventions by reducing bias while preserving high predictive performance. This superiority underscores the effectiveness and potential of FPET for broader application in real-world scenarios. By offering a practical solution to the challenge of fairness in machine learning, FPET presents a promising avenue for ensuring equitable outcomes across diverse demographic groups.

6.2 Conclusions

FPET represents a significant advancement in fairness interventions within machine learning. By innovatively adjusting decision boundaries to minimize disparities in error rates across demographic groups, FPET emerges as a pivotal solution for achieving equitable outcomes in classification tasks across various sectors. It's essential to clarify that while this study did not create the fairness framework, it rigorously tested and evaluated it across different learners and classifiers. This clarification delineates the scope of our work and emphasizes our contributions to the evaluation of the framework. This research underscores the critical need to integrate fairness interventions into machine learning workflows to rectify biases and ensure equitable outcomes for all individuals, irrespective of demographic characteristics. FPET's model-agnostic nature and computational efficiency make it an invaluable tool for practitioners aiming to embed fairness into their predictive models, thereby fostering trust and equity in decision-making processes. Moving forward, future research endeavors should aim to expand the evaluation of FPET across a wider array of datasets and real-world applications, especially those characterized by intricate and varied data distributions. Hybrid approaches that integrate FPET with other fairness-enhancing techniques hold promise for further optimizing fairness and accuracy in machine learning systems. Moreover, longitudinal studies are crucial for assessing the long-term impact of FPET on decision outcomes and fairness in dynamic datasets. By tracking the sustainability and effectiveness of FPET over time, we can ensure that fairness remains a fundamental aspect of machine learning practices in an ever-evolving data landscape. Efforts should also be directed towards developing user-friendly tools and frameworks that streamline the adoption of FPET in industry settings. Prioritizing accessibility and ease of implementation will empower practitioners to uphold fairness and equity in their machine learning endeavors. In essence, FPET embodies the promise of fairness-aware machine learning, offering a tangible pathway towards building more equitable and trustworthy AI systems. Through continued research, innovation, and collaboration, we can harness the full potential of FPET to foster a future where fairness and accuracy harmonize, driving positive societal impact and advancing ethical AI practices.

6.3 Recommendations for Future Research

While our study provides valuable insights into the effectiveness of FPET, several avenues for future research warrant exploration. Further investigation is needed to broaden the evaluation of FPET methods across a wider array of datasets and real-world applications, including those with complex and diverse data distributions.

Exploring hybrid approaches that integrate FPET with other fairness-enhancing techniques could yield valuable insights into further optimizing fairness and accuracy. Understanding the scalability of FPET methods in large-scale deployments is crucial for assessing their practical implications in high-dimensional data environments.

Longitudinal studies assessing the long-term impact of FPET on decision outcomes and fairness in dynamic datasets would provide valuable insights into its efficacy over time. Additionally, developing user-friendly tools and frameworks to facilitate the adoption of FPET in industry settings should be a priority. Ensuring that FPET methods are accessible and easy to implement for practitioners is essential for their widespread adoption and impact. Overall, future research efforts should focus on refining FPET methods and addressing the specific challenges associated with their deployment in real-world applications.

Bibliography

[1] A. Agarwal, A. Beygelzimer, M. Dudík, J. Langford, and H. Wallach. A reductions approach to fair classification. In *International conference on machine learning*, pages 60–69. PMLR, 2018.

[2] J. Angwin, J. Larson, S. Mattu, and L. Kirchner. Machine bias. In *Ethics of data and analytics*, pages 254–264. Auerbach Publications, 2022.

[3] S. Benthall and B. D. Haynes. Racial categories in machine learning. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 289–298, 2019.

[4] J. Buolamwini and T. Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, pages 77–91. PMLR, 2018.

[5] I. Chen, F. D. Johansson, and D. Sontag. Why is my classifier discriminatory? *Advances in neural information processing systems*, 31, 2018.

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

[7] E. Chzhen, C. Denis, M. Hebiri, L. Oneto, and M. Pontil. Leveraging labeled and unlabeled data for consistent fair binary classification. *Advances in Neural Information Processing Systems*, 32, 2019.

[8] S. Corbett-Davies, J. D. Gaebler, H. Nilforoshan, R. Shroff, and S. Goel. The measure and mismeasure of fairness. *The Journal of Machine Learning Research*, 24(1):14730–14846, 2023.

[9] S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, and A. Huq. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*, pages 797–806, 2017.

[10] I. Csiszár. I-divergence geometry of probability distributions and minimization problems. *The annals of probability*, pages 146–158, 1975.

[11] J. Dressel and H. Farid. The accuracy, fairness, and limits of predicting recidivism. *Science advances*, 4(1):eaao5580, 2018.

[12] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012.

[13] M. D. Ekstrand, A. Das, R. Burke, and F. Diaz. Fairness and discrimination in information access systems. *arXiv preprint arXiv:2105.05779*, 2021.

[14] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 259–268, 2015.

[15] M. Garcia, T. Nguyen, and L. Chen. Gender bias in job applicant screening: An analysis of fairness projection techniques. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery Data Mining*, 2021.

[16] M. Hardt, E. Price, and N. Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

[17] M. Hardt, E. Price, and N. Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

[18] A. Jones, B. Johnson, and C. Brown. Racial biases in sentiment analysis: A fairness projection analysis. In *Proceedings of the 27th International Conference on Computational Linguistics*, 2023.

[19] F. Kamiran and T. Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1):1–33, 2012.

[20] T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma. Fairness-aware classifier with prejudice remover regularizer. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2012, Bristol, UK, September 24-28, 2012. Proceedings, Part II 23*, pages 35–50. Springer, 2012.

[21] M. Kearns, S. Neel, A. Roth, and Z. S. Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International conference on machine learning*, pages 2564–2572. PMLR, 2018.

[22] J. S. Kim, J. Chen, and A. Talwalkar. Fact: A diagnostic for group fairness tradeoffs. In *International Conference on Machine Learning*, pages 5264–5274. PMLR, 2020.

[23] J. Kleinberg, S. Mullainathan, and M. Raghavan. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*, 2016.

[24] R. Kohavi and B. Becker. Uci machine learning repository: adult data set. *Avaliable: https://archive. ics. uci. edu/ml/machine-learning-databases/adult*, 1996.

[25] M. J. Kusner, J. Loftus, C. Russell, and R. Silva. Counterfactual fairness. *Advances in neural information processing systems*, 30, 2017.

[26] M. J. Kusner, J. Loftus, C. Russell, and R. Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[27] D. Larremore, M. Singhal, and N. Chawla. Fairness and accountability in algorithmic risk assessment: A case study of criminal risk prediction. In *Proceedings of the 2020 AAAI/ACM Conference on AI, Ethics, and Society*, 2019.

[28] D. Madras, E. Creager, T. Pitassi, and R. Zemel. Learning adversarially fair and transferable representations. In *International Conference on Machine Learning*, pages 3384–3393. PMLR, 2018.

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

[30] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019.

[31] R. Patel, E. Park, S. Lee, and J. Kim. Addressing bias in credit scoring: A fairness-aware post-processing approach. *Journal of Machine Learning Research*, 21(10):1–25, 2020.

[32] J. Platt et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999.

[33] I. D. Raji, A. Smart, R. N. White, M. Mitchell, T. Gebru, B. Hutchinson, J. Smith-Loud, D. Theron, and P. Barnes. Closing the ai accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 33–44, 2020.

[34] A. D. Selbst, D. Boyd, S. A. Friedler, S. Venkatasubramanian, and J. Vertesi. Fairness and abstraction in sociotechnical systems. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 59–68, 2019.

[35] J. Smith, L. Wang, and A. Gupta. Fairness projection in healthcare diagnosis: Addressing ethnic biases in disease prediction models. *Journal of Artificial Intelligence in Medicine*, 45:1–18, 2022.

[36] H. Suresh and J. Guttag. A framework for understanding unintended consequences of machine learning. *arXiv preprint arXiv:1901.10002*, 2021.

[37] H. Suresh and J. V. Guttag. A framework for understanding unintended consequences of machine learning. *arXiv preprint arXiv:1901.10002*, 2(8):73, 2019.

[38] Y. Wang, Q. Zhang, and X. Liu. Mitigating biases in student performance prediction: A fairness projection approach. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence*, 2021.

[39] F. Yang, M. Cisse, and S. Koyejo. Fairness with overlapping groups; a probabilistic perspective. *Advances in neural information processing systems*, 33:4067–4078, 2020.

[40] M. B. Zafar, I. Valera, M. G. Rodriguez, and K. P. Gummadi. Fairness constraints: Mechanisms for fair classification. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2017.

[41] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork. Learning fair representations. In *International conference on machine learning*, pages 325–333. PMLR, 2013.

[42] B. Zhang, B. Lemoine, and M. Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, 2018.

[43] B. H. Zhang, B. Lemoine, and M. Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340, 2018.

[44] J. Zhang and E. Bareinboim. Fairness in decision-making—the causal explanation formula. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[45] Y. Zhang, F. Zhou, Z. Li, Y. Wang, and F. Chen. Bias-tolerant fair classification. In *Asian Conference on Machine Learning*, pages 840–855. PMLR, 2021.