Investing Bias in Heart Disease Prediction Model

Team: Unbiased Minds

Sakshi Tawte
Department of Mathematics and Science
Stevens Institute of Technology
Hoboken, NJ, USA
stawte@stevens.edu

Palak Sood
Department of Mathematics and Science
Stevens Institute of Technology
Hoboken, NJ, USA
psood1@stevens.edu

Sayan Seal
Department of Mathematics and Science
Stevens Institute of Technology
Hoboken, NJ, USA
sseal1@stevens.edu

Abstract

Machine learning models in critical domains like healthcare risk perpetuating societal biases, leading to inequitable outcomes; this project addresses the challenge of identifying and mitigating gender and age-based bias in models for heart disease prediction. Using the UCI Heart Disease dataset, Logistic Regression and Random Forest models were trained, and bias was investigated by disaggregating performance metrics (accuracy, FPR, FNR) across gender and age groups. To mitigate identified disparities, post-processing (ThresholdOptimizer for both gender and age) and in-processing (Exponentiated Gradient for both gender and age) techniques were implemented and evaluated.

This work's key contributions include a comprehensive bias analysis for multiple sensitive attributes, a comparative evaluation of different mitigation strategies, and the development of a prototype Flask application for serving fairness-aware predictions. Baseline models exhibited significant performance disparities across gender and age subgroups. Mitigation techniques successfully reduced these fairness gaps—for instance, Demographic Parity Difference for gender was reduced from over 0.40 to as low as approximately 0.038 by Exponentiated Gradient, and Equal Opportunity Difference for age was reduced from 0.35 to 0.03 by ThresholdOptimizer—often with a manageable trade-off in overall accuracy, achieving more equitable predictive outcomes. This research highlights the necessity of bias auditing and mitigation in developing trustworthy AI and offers a practical pathway towards deploying fairer machine learning models in healthcare, potentially improving diagnostic equity for diverse demographic groups.

1 Introduction

The integration of Artificial Intelligence (AI) and Machine Learning (ML) into healthcare holds transformative potential, promising to revolutionize diagnostics, personalize treatments, and enhance overall patient care. As these sophisticated algorithms are increasingly employed in critical decision-making processes, such as predicting the likelihood of diseases like heart conditions, it becomes paramount to address the inherent challenges they present. One of the most significant concerns is the propensity for ML models to learn and perpetuate existing societal biases, potentially leading to

Preprint. Under review.

inequitable health outcomes for different demographic groups. This project delves into this critical issue by investigating the presence of gender and age-related biases in ML models developed for heart disease prediction. It further explores the application and efficacy of various fairness-aware techniques to mitigate these biases, aiming to foster more equitable and trustworthy AI systems in the medical domain. The subsequent sections will detail the motivation behind this study, the specific problems addressed, the project's objectives and scope, its main contributions, and the overall structure of this report.

1.1 Motivation

Machine learning models are increasingly integral to decision-making in critical sectors, particularly healthcare, where they offer the potential to improve diagnostics and patient outcomes. However, these models can inadvertently learn and amplify existing societal biases present in historical data. In the context of heart disease prediction, if a model is biased against specific demographic groups (e.g., based on gender or age), it can lead to serious consequences such as misdiagnosis, delayed treatment, or inequitable allocation of medical resources. This can exacerbate health disparities and undermine trust in AI-driven healthcare solutions. Therefore, ensuring fairness and equity in these predictive models is not just a technical challenge but an ethical imperative to prevent harm and promote just medical practices.

1.2 Problem Statement:

This project aims to systematically identify, quantify, and mitigate biases related to gender and age in machine learning models developed for heart disease prediction using the UCI Heart Disease dataset. The core problem is that standard machine learning development practices may result in models that perform disparately across different demographic subgroups, potentially leading to unfair or less effective outcomes for certain groups. We seek to determine the extent of such biases in common classification models (Logistic Regression and Random Forest) and explore the effectiveness of established fairness intervention techniques (ThresholdOptimizer and Exponentiated Gradient) in creating more equitable models.

1.3 Objectives and Scope

The primary focus of this project is to train baseline Logistic Regression and Random Forest models for heart disease prediction and conduct a comprehensive bias audit concerning gender and predefined age groups ('Younger_lte45', 'Middle_46-60', 'Older_gt60'). This involves using fairness metrics like accuracy, False Positive Rate (FPR), False Negative Rate (FNR), Demographic Parity Difference (DPD), and True Positive Rate Difference/Equal Opportunity Difference (TPR Diff/EOpp Diff). A core objective is to implement and evaluate the effectiveness of selected bias mitigation techniques—specifically, the post-processing `ThresholdOptimizer` and the in-processing `ExponentiatedGradient` method for both gender and age—in reducing observed biases. The project also aims to analyze the inherent trade-offs between enhancing fairness and maintaining overall model predictive performance. Finally, a prototype Flask application will be developed to demonstrate the predictions of both baseline and fairness-adjusted models. The scope of this work is confined to the UCI Heart Disease dataset, the aforementioned models and mitigation techniques, and the analysis of gender and age as the primary sensitive attributes.

1.4 Main Contribution:

This project offers several key technical and empirical contributions to the study of bias in machine learning for healthcare. Firstly, it provides a Comprehensive Bias Audit by systematically investigating and quantifying performance disparities across two distinct sensitive attributes—gender and age—within heart disease prediction models, utilizing multiple fairness metrics such as DPD, TPR/EOpp Difference, FPR, and FNR. Secondly, it delivers a Comparative Evaluation of Mitigation Techniques, empirically assessing and contrasting the effectiveness of both post-processing (ThresholdOptimizer) and in-processing (Exponentiated Gradient) strategies in reducing bias related to both gender and age, while also detailing their impact on overall model accuracy. Thirdly, the research demonstrates Multi-Attribute Mitigation Application by successfully applying and analyzing these mitigation techniques for both a binary discrete attribute (gender) and a multi-category ordered

attribute (age group), thereby showcasing the adaptability and differential impact of these methods. Lastly, a Prototype for Fairer Predictions was developed as a functional web application, serving predictions from both original and bias-mitigated models and offering a tangible demonstration of how fairness interventions can be operationalized and their outputs, including adjusted confidence scores, interpreted.

1.5 Structure:

The remainder of this report is organized to systematically present the research conducted. Section 2 provides a Literature Review, discussing prior work on algorithmic bias, fairness metrics, mitigation techniques, and their relevance to healthcare AI. Section 3, Methodology, details the dataset, preprocessing steps, the machine learning models employed, the framework for bias assessment including specific fairness metrics, and the implementation of the bias mitigation algorithms. Section 4, Results and Findings, presents the empirical outcomes, covering baseline model performance, the bias audit results for gender and age, the impact of mitigation techniques, and a brief overview of the Flask application. Section 5, Discussion, offers an interpretation of these results, examines the observed trade-offs, discusses the study's limitations, and reflects on the broader implications. Finally, Section 6, Conclusion and Future Work, summarizes the project's key achievements and contributions and suggests potential avenues for future research.

2 Background and Related Work

2.1 Key Concepts

To understand the methodologies and contributions of this project, several key concepts in algorithmic fairness must be defined. Fairness in machine learning aims to ensure that predictive models do not systematically disadvantage individuals or groups based on sensitive attributes such as gender, race, or age. This is often quantified using various metrics. Demographic Parity Difference (DPD) is one such metric, which assesses whether the likelihood of a positive outcome (e.g., being predicted to have heart disease) is equal across different demographic groups. A DPD of zero indicates perfect demographic parity. Another critical metric is Equal Opportunity Difference (EOD), which focuses on whether individuals who genuinely belong to the positive class have an equal chance of being correctly identified by the model, irrespective of their group membership; this is often measured by comparing True Positive Rates (TPR) across groups.

To address biases identified by these metrics, various bias mitigation techniques have been developed. These can broadly be categorized into pre-processing (modifying the data), in-processing (modifying the learning algorithm), and post-processing (modifying the model's predictions). This project utilizes two prominent techniques:

ThresholdOptimizer: This is a post-processing method that adjusts the decision threshold (the cutoff point for classifying an instance as positive or negative) for different demographic groups. Instead of a single global threshold, it learns group-specific thresholds to satisfy fairness constraints like demographic parity or equalized odds, without retraining the original model. Exponentiated Gradient: This is an in-processing algorithm that iteratively trains a base classifier by re-weighting training examples or adjusting the learning objective to directly incorporate fairness constraints into the model optimization process. It aims to find a model that achieves a good balance between predictive accuracy and fairness. Several open-source toolkits, such as IBM's AI Fairness 360 (AIF360) and Microsoft's Fairlearn, provide implementations of these fairness metrics and mitigation techniques, facilitating their application in practical machine learning pipelines.

2.2 Related Work

The pursuit of fairness in machine learning, particularly within the healthcare domain, has gained significant traction, with numerous studies investigating biased outcomes and proposing solutions.

Several studies have specifically focused on identifying unfair outcomes in disease prediction. Agrawal et al. (2021) investigated gender disparities in heart disease prediction using the UCI dataset, finding significant differences in model outcomes between male and female patients [1]. Their work highlighted the prevalence of gender bias in a widely used medical dataset, underscoring

the need for fairness interventions. Zhang et al. (2021) explored age-related bias in cardiovascular risk prediction models, emphasizing the necessity for demographic-aware evaluation frameworks to ensure models are equitable across different age cohorts [2]. Similarly, Chen et al. (2020) examined bias in chronic disease diagnostics more broadly, discussing the clinical and ethical implications of performance gaps between various demographic groups [3]. These studies are strong in identifying and quantifying bias but may not always provide extensive comparative evaluations of multiple mitigation strategies across different sensitive attributes simultaneously.

In addition to detecting bias, researchers have proposed various mitigation strategies and tools. Bellamy et al. (2019) introduced AI Fairness 360, an extensible open-source toolkit providing a comprehensive suite of bias metrics and mitigation algorithms, which has become a foundational resource for fairness research [7]. While AIF360 offers a wide array of tools, its application often requires careful selection and tuning for specific problem contexts. More recently, Alvi et al. (2022) conducted a comparative analysis of fairness metrics in medical datasets, critically highlighting the inherent trade-offs between achieving fairness and maintaining predictive accuracy [8]. Their work emphasizes that no single fairness metric or mitigation technique is universally optimal, and the choice often depends on the specific ethical considerations and desired outcomes of the application. While these studies offer valuable tools and insights into fairness-accuracy trade-offs, they may not always demonstrate the application of these techniques to multiple sensitive attributes (like gender and age concurrently) within a single predictive task like heart disease.

Our project builds upon this existing body of work in several novel ways. While prior studies like Agrawal et al. have focused on gender bias in heart disease using the UCI dataset, and Zhang et al. on age, our work concurrently investigates and mitigates bias for both gender and age within the same predictive modeling framework for heart disease. Furthermore, we provide an empirical comparison of both in-processing (Exponentiated Gradient) and post-processing (ThresholdOptimizer) mitigation strategies applied to both sensitive attributes, detailing their differential impacts on fairness metrics and overall accuracy. Many studies focus on one type of mitigation or one attribute; our dual-attribute, dual-methodology approach offers a more comprehensive perspective. Finally, the development of a prototype Flask application to serve predictions from both baseline and fairness-aware models provides a practical demonstration of how these interventions can be operationalized, which is a step beyond purely analytical studies.

3 Methodology

This section outlines the formal problem setup, the approach taken to investigate and mitigate bias, the rationale behind the design choices, and the specific implementation details of the project.

The primary task of this project is binary classification to predict the presence or absence of heart disease in individuals.

- **Input:** The inputs to our models are a set of features derived from the UCI Heart Disease dataset. After preprocessing, these include 11 to 13 numerical and categorical features representing patient demographic information (e.g., age, sex), clinical measurements (e.g., cholesterol, resting blood pressure, maximum heart rate), and symptoms (e.g., chest pain type, exercise-induced angina). The sensitive attributes under investigation are `sex` (binary: Male/Female) and `age_group` (categorical: 'Younger_lte45', 'Middle_46-60', 'Older_gt60').

- **Outputs:** The primary output is a binary prediction for each patient: 0 (no significant heart disease) or 1 (heart disease present). Additionally, for fairness-adjusted models, the outputs include modified predictions aimed at satisfying specific fairness constraints, along with associated probabilities or confidence scores (e.g., adjusted confidence from ThresholdOptimizer).

- **Assumptions:**

  - The UCI Heart Disease dataset, despite its known imperfections and missing values, is sufficiently representative for the purpose of investigating and demonstrating bias detection and mitigation techniques.

- The imputation methods used (median for numerical, mode for categorical) provide a reasonable approximation for missing data without introducing significant additional bias, though this is a simplification.
- The chosen fairness metrics (Demographic Parity Difference, Equal Opportunity Difference, and subgroup-specific error rates like FPR and FNR) are relevant and appropriate for evaluating fairness in this healthcare context.
- The binarization of the original multi-class heart disease severity ('num' variable) into a binary target ('target') is a common and acceptable simplification for this type of fairness investigation.

3.1 Approach

The overall approach follows a structured pipeline designed to train baseline models, assess them for bias, apply mitigation techniques, and evaluate the outcomes:

1. **Data Preprocessing:** The raw UCI Heart Disease dataset (`heart_disease_uci.csv`) was loaded and cleaned. This involved:
   - Handling missing values through median and mode imputation.
   - Encoding categorical features into numerical representations. `sex` was encoded as 0 (Female) and 1 (Male). `age` was categorized into three groups: 'Younger_lte45' (≤ 45 years), 'Middle_46-60' (46-60 years), and 'Older_gt60' (> 60 years).
   - Binarizing the multi-level target variable `num` into a binary `target` variable (0 for no/low presence, 1 for significant presence of heart disease).
   - Creating a scikit-learn `ColumnTransformer` to apply `StandardScaler` to numerical features and `OneHotEncoder` to categorical features, ensuring consistent preprocessing for all models.
   - Splitting the data into training (75%) and testing (25%) sets, stratified by the target variable.
2. **Baseline Model Training:** Two standard classification algorithms were chosen as baseline models:
   - Logistic Regression: A linear model known for its interpretability and efficiency.
   - Random Forest Classifier: An ensemble model capable of capturing non-linear relationships and often achieving higher accuracy.

   These models were trained on the preprocessed training data.
3. **Bias Audit:** The trained baseline models were evaluated on the test set. Performance was disaggregated by the sensitive attributes (`sex` and `age_group`) to identify disparities. Key metrics included:
   - Overall accuracy and ROC AUC.
   - Subgroup-specific accuracy, ROC AUC, False Positive Rate (FPR), and False Negative Rate (FNR).
   - Fairness metrics: Demographic Parity Difference (DPD), calculated from Positive Prediction Rates (PPR), and Equal Opportunity Difference (EOD), calculated from True Positive Rates (TPR).
4. **Bias Mitigation:** Based on the audit, selected bias mitigation techniques from the Fairlearn library were applied:
   - `ExponentiatedGradient` (In-processing): This algorithm was applied using a Logistic Regression base estimator. It iteratively retrains a classifier to satisfy fairness constraints (e.g., Demographic Parity or Equalized Odds) during the training process.
   - `ThresholdOptimizer` (Post-processing): This technique was applied to the predictions of the already trained Logistic Regression and Random Forest models. It identifies optimal group-specific decision thresholds to satisfy fairness constraints (e.g., Demographic Parity or Equalized Odds, with Equalized Odds being a primary target in the age analysis) without retraining the base models.

5. **Evaluation of Mitigated Models:** The mitigated models were then evaluated on the test set using the same performance and fairness metrics as the baseline models. This allowed for a direct comparison of bias levels and any trade-offs with overall accuracy.
6. **Flask Application Development:** A simple Flask web application (`app.py`) was created to serve predictions from the pre-trained baseline and mitigated models (specifically for gender bias mitigation). This demonstrates how such models could be deployed.

3.2 Design Rationale

The design choices in this project were guided by the aim to provide a comprehensive yet practical investigation into bias in a common healthcare prediction task.

- **Choice of Baseline Models:** Logistic Regression was chosen as a simple, interpretable baseline, while Random Forest was selected for its potential for higher performance and ability to model complex interactions. Comparing these helps understand how bias manifests and can be mitigated in models of varying complexity.
- **Choice of Sensitive Attributes:** Gender and age are well-documented sources of disparity in healthcare access and outcomes, making them critical attributes to investigate for bias in heart disease prediction. Analyzing both a binary attribute (gender) and a multi-category attribute (age group) provides broader insights.
- **Choice of Fairness Metrics:** DPD and EOD were chosen as they represent distinct and widely recognized fairness notions. DPD focuses on equality of prediction rates, while EOD focuses on equality of correctly identifying positive instances among those who truly have the condition. Analyzing FPR and FNR for subgroups is crucial for understanding the specific harms (e.g., missed diagnoses vs. false alarms) different groups might experience.
- **Choice of Mitigation Techniques:**
  - `ExponentiatedGradient` was selected as a representative in-processing technique that directly integrates fairness into the model training. It allows for optimizing a trade-off between accuracy and fairness constraints.
  - `ThresholdOptimizer` was chosen as a practical post-processing technique that can be applied to any pre-trained classifier providing scores. It's often simpler to implement and tune than in-processing methods and directly addresses fairness by adjusting decision points.

  The selection of both an in-processing and a post-processing method allows for a comparison of different intervention points in the modeling pipeline.
- **Novelty/Uniqueness:** While individual components (e.g., using Fairlearn, analyzing UCI Heart Disease) are common, this project's contribution lies in:
  - The concurrent and comparative analysis of bias for two distinct sensitive attributes (gender and age) within the same problem context.
  - The direct comparison of both in-processing and post-processing mitigation strategies across these attributes and multiple base models.
  - The detailed reporting of trade-offs not just in aggregate fairness metrics but also in subgroup-specific error rates (FPR/FNR), which have direct clinical relevance.
  - The development of a Flask application to demonstrate the practical deployment and interpretation of fairness-adjusted predictions, including concepts like group-specific thresholds and adjusted confidence.

3.3 Implementation Details

3.3.1 Tools and Libraries

- Python 3.x was the primary programming language.
- Pandas for data manipulation and analysis.
- NumPy for numerical operations.

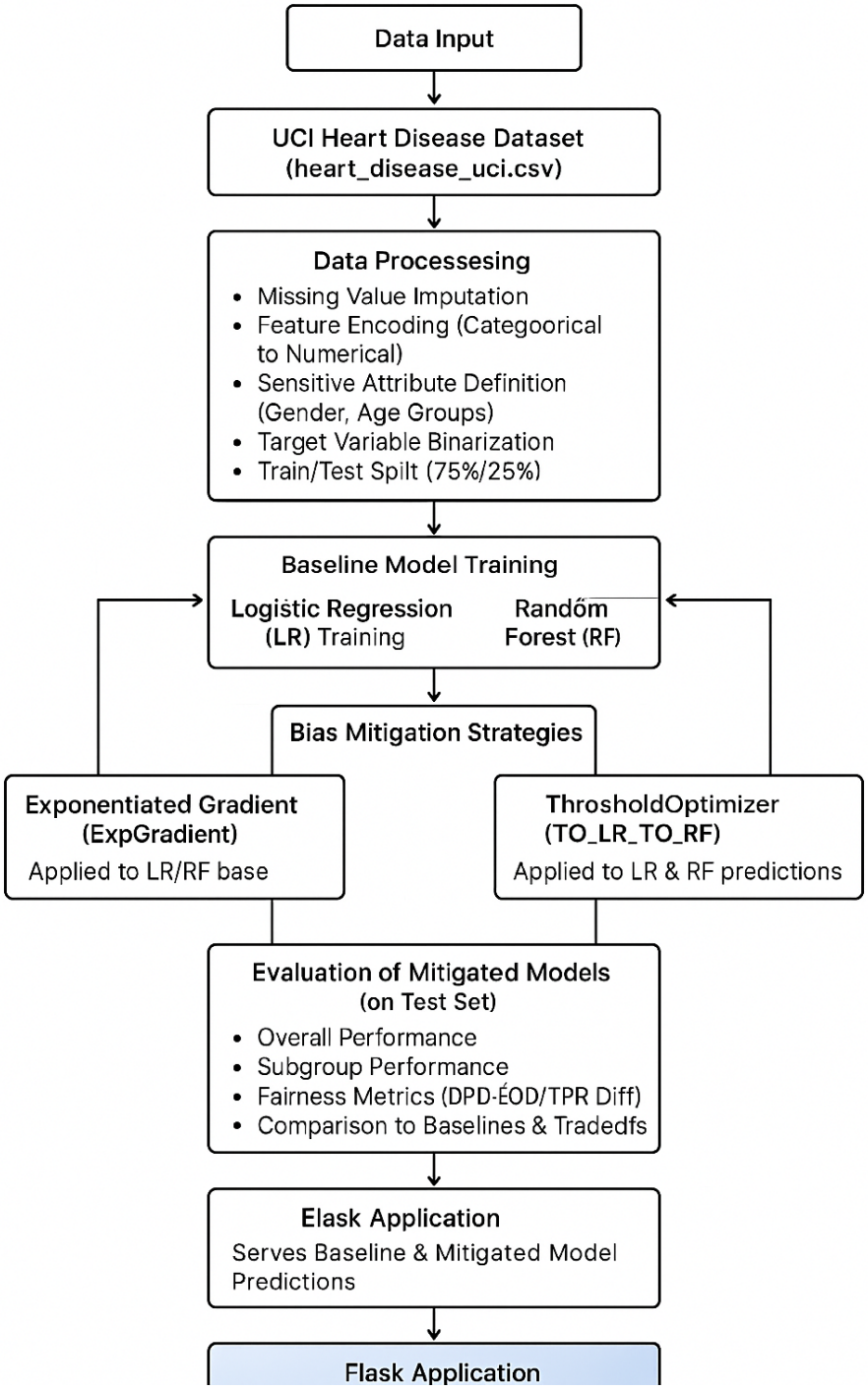

Figure 1: Overall project pipeline illustrating the flow from data preprocessing and baseline model training, through bias audit and mitigation, to final evaluation and Flask application development.

- Scikit-learn for machine learning tasks: model training (`LogisticRegression`, `RandomForestClassifier`), preprocessing (`StandardScaler`, `OneHotEncoder`, `ColumnTransformer`), metrics (`accuracy_score`, `roc_auc_score`, `confusion_matrix`, `classification_report`), and train-test splitting.
- Fairlearn for bias assessment (`MetricFrame`) and mitigation algorithms (`ExponentiatedGradient`, `ThresholdOptimizer`).
- Matplotlib and Seaborn for data visualization.
- Joblib for saving and loading trained models and preprocessors.
- Flask for developing the web application prototype.
- Jupyter Notebooks (`bias_gendersw.ipynb`, `biasa.ipynb`) for conducting the analysis and experiments.

3.3.2 Data Preprocessing

- Missing values: Imputed using `df[col].fillna(df[col].median())` for numerical and `df[col].fillna(df[col].mode()[0])` for categorical features.
- Target variable `target`: Created via `(df['num'] > 0).astype(int)`.
- Sensitive attribute `sex_male`: `df['sex'].replace({'Female': 0, 'Male': 1})`.
- Sensitive attribute `age_group`: Created using `pd.cut` with bins `[0, 45, 60, df['age'].max()]` and labels `['Younger_lte45', 'Middle_46-60', 'Older_gt60']`.
- Features `X` and target `y` were defined, and `X` was processed using a `ColumnTransformer` with `StandardScaler` for numerical columns and `OneHotEncoder(handle_unknown='ignore', drop='first')` for categorical columns.
- Data split: `train_test_split(X_processed, y, test_size=0.25, random_state=42, stratify=y)`.

3.3.3 Model Hyperparameters

- Logistic Regression: `LogisticRegression(solver='liblinear', random_state=42, max_iter=1000)`.
- Random Forest Classifier: `RandomForestClassifier(n_estimators=100, random_state=42)`.
- `ExponentiatedGradient`:
  - For gender: `ExponentiatedGradient(estimator=LogisticRegression(solver='liblinear', random_state=42, max_iter=1000), constraints=DemographicParity(), eps=0.01, max_iter=50)`.
  - For age: `ExponentiatedGradient(estimator=LogisticRegression(solver='liblinear', random_state=42, max_iter=1000), constraints=DemographicParity(), eps=0.02, max_iter=50)`.
- `ThresholdOptimizer`:
  - For gender (LR): `ThresholdOptimizer(estimator=log_reg_original, constraints='demographic_parity', objective='accuracy_score', prefit=True)`.
  - For gender (RF): `ThresholdOptimizer(estimator=rf_original, constraints='demographic_parity', objective='accuracy_score', prefit=True)`.
  - For age (LR & RF): `ThresholdOptimizer(estimator=..., constraints='equalized_odds', objective='balanced_accuracy_score', prefit=True)`.
- `random_state=42` was used where applicable for reproducibility.

3.3.4 Flask Application (`app.py`)

- Loads pickled preprocessor, baseline models, and mitigated models (`optimizer_lr`, `optimizer_rf`, `mitigator.pkl` which is the `ExponentiatedGradient` model for gender).
- Defines a `/predict` endpoint that accepts JSON input of patient features.
- Preprocesses the input using the loaded preprocessor.
- Makes predictions using all loaded models.
- For `ThresholdOptimizer` models, it retrieves the group-specific threshold and calculates an "adjusted_confidence" (`probability - group_threshold`).
- Returns predictions and probabilities/confidences in a JSON response.

3.4 Design Rationale

Our design choices were driven by the need to balance predictive performance with fairness. We selected Logistic Regression and Random Forest as baseline models due to their interpretability and strong performance on tabular clinical data. These models are also compatible with fairness-aware post- and in-processing methods available in libraries like Fairlearn.We chose Demographic Parity and Equal Opportunity as our fairness metrics because they reflect different dimensions of bias—equal treatment and equal performance across sensitive groups. To mitigate these biases, we applied Threshold Optimizer (post-processing) and Exponentiated Gradient (in-processing), which are widely studied and supported, allowing us to explore fairness trade-offs without altering model structure. Testing on Cleveland and Switzerland subsets was introduced to assess fairness consistency across regions—a layer of evaluation often missing in similar studies.

4 Evaluation

This section details the experimental setup used to evaluate the baseline and mitigated models, presents the quantitative and qualitative results obtained for both gender and age bias analyses, and provides an in-depth analysis of these findings.

4.1 Experimental Setup

- **Dataset:** The study utilized the "Heart Disease UCI" dataset (`heart_disease_uci.csv`). After preprocessing, including imputation of missing values and feature engineering, the final dataset comprised 920 entries. For modeling, the feature set (X) had a shape of (920, 11) after one-hot encoding and selection, and the binary target variable (y) had a shape of (920,).
- **Sensitive Attributes:**
  - Gender: Analyzed as a binary attribute (`sex`), encoded as 0 for Female and 1 for Male.
  - Age: Categorized into three groups: 'Younger_lte45' (≤ 45 years), 'Middle_46-60' (46-60 years), and 'Older_gt60' (> 60 years).
- **Train/Test Split:** The dataset was split into a training set (690 samples, 75%) and a test set (230 samples, 25%). The split was stratified by the target variable to ensure proportional representation of classes in both sets. All reported evaluations were performed on the unseen test set. For the gender analysis, the test set comprised 46 Female samples and 184 Male samples.
- **Baseline Models:** The performance of mitigated models was compared against two baseline classifiers:
  1. Logistic Regression (Original)
  2. Random Forest Classifier (Original)
- **Mitigation Techniques Evaluated:**
  1. `ExponentiatedGradient` (`ExpGradient`): An in-processing technique.
  2. `ThresholdOptimizer` (`TO_LR`, `TO_RF`): A post-processing technique applied to Logistic Regression and Random Forest models, respectively.

- **Performance Metrics:**
  - Overall Model Performance: Accuracy, ROC AUC.
  - Subgroup Performance: Accuracy, ROC AUC, False Positive Rate (FPR), False Negative Rate (FNR) for each gender and age subgroup.
- **Fairness Metrics:**
  - Demographic Parity Difference (DPD): Calculated as the difference in Positive Prediction Rates (PPR) between the privileged and unprivileged groups (e.g., Male PPR - Female PPR; Max Age Group PPR - Min Age Group PPR).
  - Equal Opportunity Difference (EOD) / True Positive Rate Difference (TPR Diff): Calculated as the difference in True Positive Rates (TPR) between groups.

4.2 Results: Gender Bias Analysis

The following results are derived from the gender bias analysis.

4.2.1 Baseline Model Performance and Bias (Gender)

- **Logistic Regression (Original):**
  - Overall: Accuracy: 0.8000, ROC AUC: 0.8701.
  - Bias:
    - Females: Accuracy: 0.8261, ROC AUC: 0.8805, FPR: 0.0857, **FNR: 0.4545**
    - Males: Accuracy: 0.7935, ROC AUC: 0.8401, FPR: 0.3088, FNR: 0.1466
  - DPD: **0.4565** (Male PPR: 0.6522, Female PPR: 0.1957)
  - TPR Difference: **0.3080** (Male TPR: 0.8534, Female TPR: 0.5455)
- **Random Forest Classifier (Original):**
  - Overall: Accuracy: 0.8000, ROC AUC: 0.8638.
  - Bias:
    - Females: Accuracy: 0.8478, ROC AUC: 0.8870, FPR: 0.0857, **FNR: 0.3636**
    - Males: Accuracy: 0.7880, ROC AUC: 0.8367, FPR: 0.2941, FNR: 0.1638
  - DPD: **0.4185** (Male PPR: 0.6359, Female PPR: 0.2174)
  - TPR Difference: **0.1998** (Male TPR: 0.8362, Female TPR: 0.6364)

Initial analysis revealed significant bias in both baseline models. Males were predicted positive much more frequently (high DPD). Critically, females with heart disease were often missed, as indicated by high FNRs (45.5% for LR, 36.4% for RF). Males with heart disease also had a higher chance of being correctly identified (higher TPR).

4.2.2 Impact of Mitigation Strategies on Gender Bias

- `ExponentiatedGradient` (`ExpGradient`)**:**
  - Overall: Accuracy: 0.7391, ROC AUC: 0.7317.
  - Fairness Impact:
    - DPD: **0.0380** (Male PPR: 0.6033, Female PPR: 0.5652) – *Dramatically reduced.*
    - TPR Difference: **-0.1160** (Male TPR: 0.7931, Female TPR: 0.9091) – *Inverted, females now have higher TPR.*
    - Female FNR: **0.0909** – *Significantly improved recall for females.*
    - Female FPR: 0.4571 (Increased).
- `ThresholdOptimizer` **on Logistic Regression** (`TO_LR`)**:**
  - Overall: Accuracy: 0.7870, ROC AUC: 0.8701.
  - Fairness Impact:
    - DPD: **0.1576** (Male PPR: 0.6576, Female PPR: 0.5000) – *Significantly reduced.*
    - TPR Difference: **-0.1379** (Male TPR: 0.8621, Female TPR: 1.0000) – *Inverted, females now have perfect TPR.*
    - Female FNR: **0.0000** – *Perfect recall for females.*

    - Female FPR: 0.3429 (Increased).
- `ThresholdOptimizer` **on Random Forest (**`TO_RF`**):**
  - Overall: Accuracy: 0.7957, ROC AUC: 0.8638.
  - Fairness Impact:
    - DPD: **0.2718** (Male PPR: 0.5761, Female PPR: 0.3043) – *Significantly reduced.*
    - TPR Difference: **0.0658** (Male TPR: 0.7931, Female TPR: 0.7273) – *Most balanced TPR among mitigated models.*
    - Female FNR: 0.2727 (Improved but not zero).
    - Female FPR: 0.1714.

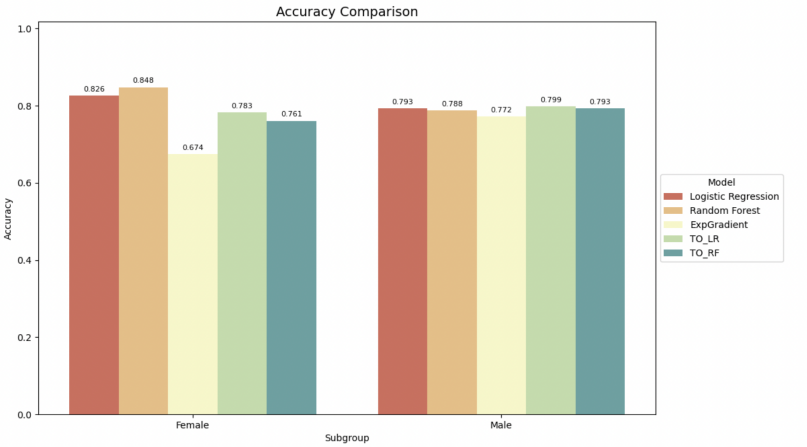

Figure 2: Model Accuracy Comparison by Gender Subgroups.

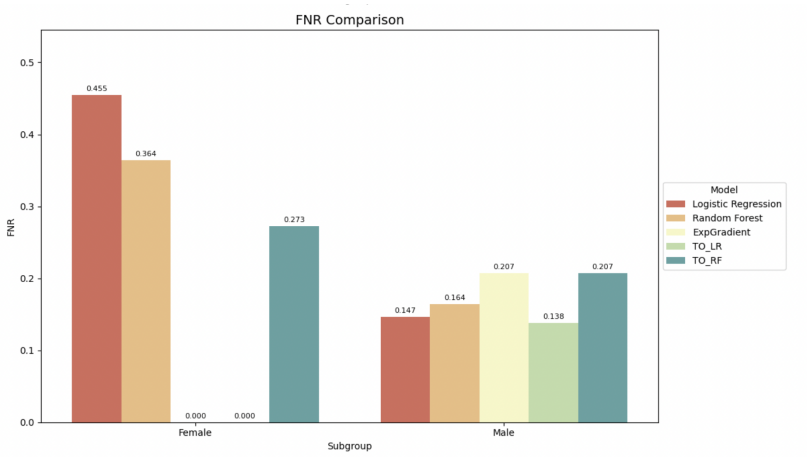

Figure 3: False Negative Rate (FNR) Comparison by Gender Subgroups.

4.3 Results: Age Bias Analysis

The following results are derived from the `biasa.ipynb` analysis, focusing on three age groups: 'Younger_lte45', 'Middle_46-60', and 'Older_gt60'.

4.3.1 Baseline Model Performance and Bias (Age)

- **Logistic Regression (Original -** `LR_Age`**):**
  - Overall: Accuracy: 0.8522, ROC AUC: 0.9114.

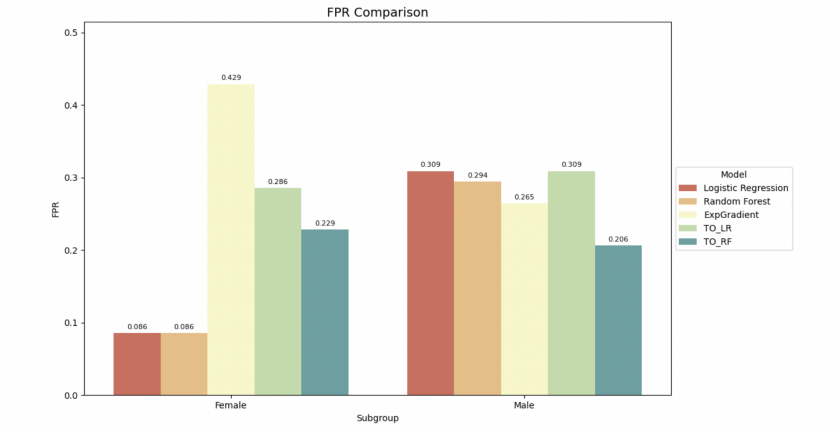

Figure 4: False Positive Rate (FPR) Comparison by Gender Subgroups.

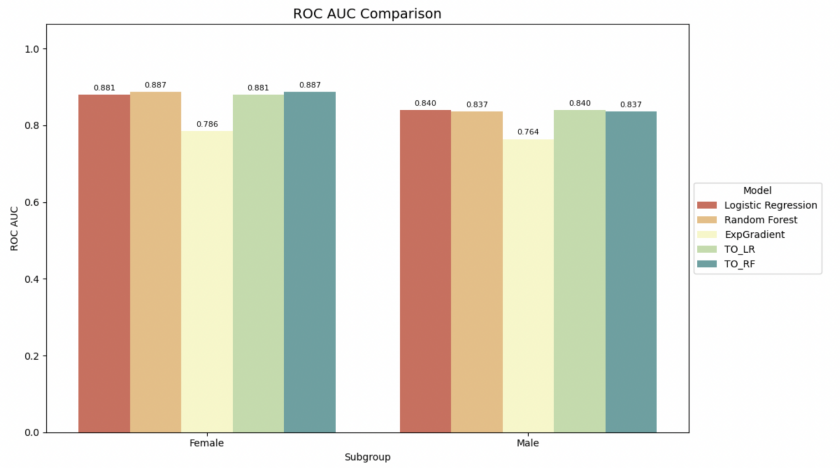

Figure 5: ROC AUC Score Comparison by Gender Subgroups.

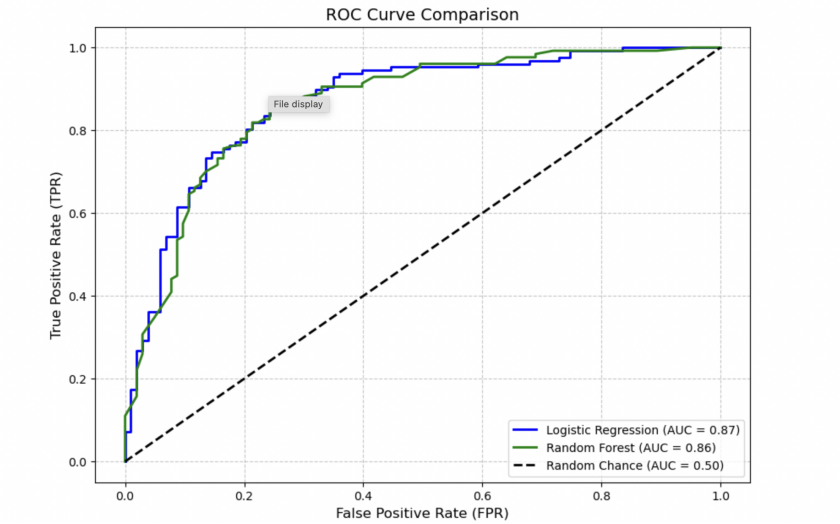

Figure 6: ROC Curves for Different Models by Gender.

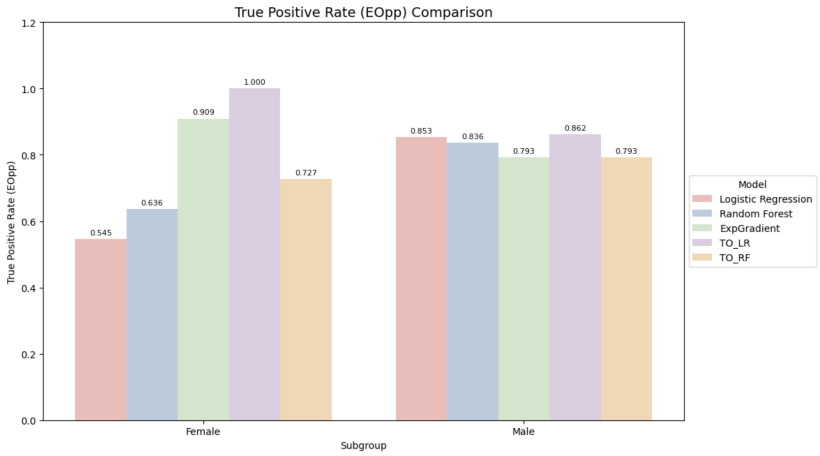

Figure 7: True Positive Rate (TPR) Comparison by Gender Subgroups.

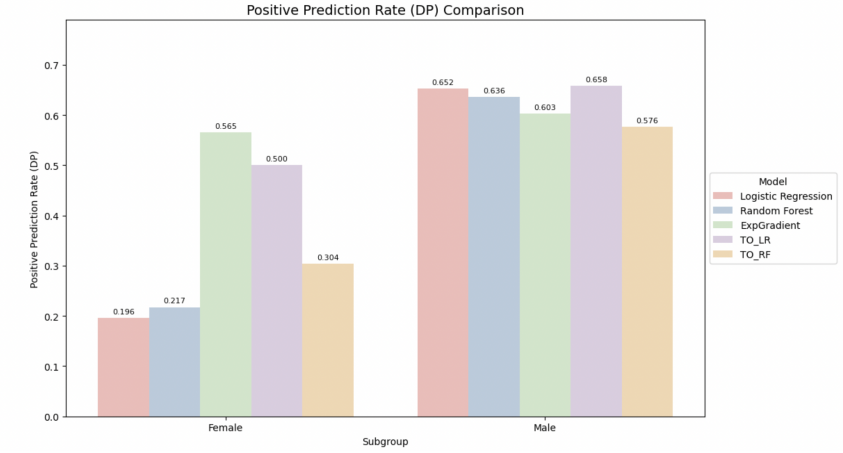

Figure 8: Positive Prediction Rate (PPR) Comparison by Gender Subgroups.

  - Bias (Max Disparities across age groups):
    - FNR: Highest for 'Younger_lte45' (0.3810).
    - FPR: Highest for 'Older_gt60' (0.5333).
  - DPD (Max PPR - Min PPR): **0.5878**
  - EOD (Max TPR - Min TPR): **0.3583**
- **Random Forest Classifier (Original - `RF_Age`):**
  - Overall:
  - Bias (Max Disparities across age groups):
    - FNR:
    - FPR: Highest for 'Older_gt60' (0.5333).
  - DPD: **0.4688**
  - EOD: **0.1223**

Unmitigated models showed significant age-related bias. Younger individuals with heart disease were more likely to be missed (high FNR), while older individuals without the disease were more likely to receive false positives (high FPR). Both DPD and EOD were notably high, especially for Logistic Regression.

4.3.2 Impact of Mitigation Strategies on Age Bias

- `ExponentiatedGradient` (`ExpGradient_Age` - **using LR base):**
  - Overall:
  - Fairness Impact (Max Disparities):
    - DPD: **0.1240** – *Best DPD reduction.*
    - EOD: **0.1818**
- `ThresholdOptimizer` **on Logistic Regression** (`TO_LR_Age` - **targeting Equalized Odds):**
  - Overall: Accuracy: 0.8217 (slight decrease from baseline LR).
  - Fairness Impact (Max Disparities):
    - DPD: **0.2265** – *Significant reduction.*
    - EOD: **0.0307** – *Best EOD reduction.*
- `ThresholdOptimizer` **on Random Forest** (`TO_RF_Age` - **targeting Equalized Odds):**
  - Overall:
  - Fairness Impact (Max Disparities):
    - DPD: **0.1327** – *Highly effective DPD reduction.*
    - EOD: **0.0653** – *Highly effective EOD reduction.*

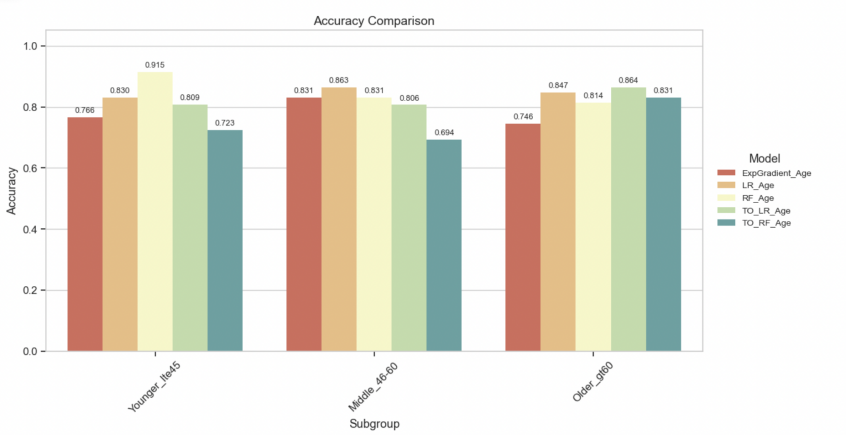

Figure 9: Model Accuracy Comparison by Age Subgroups.

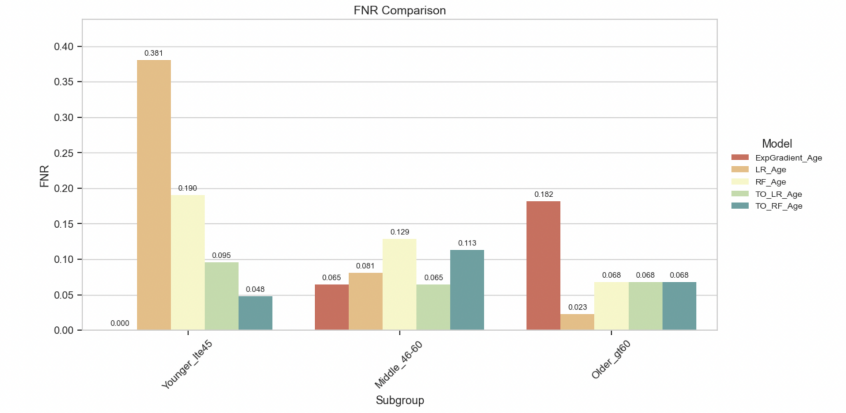

Figure 10: False Negative Rate (FNR) Comparison by Age Subgroups.

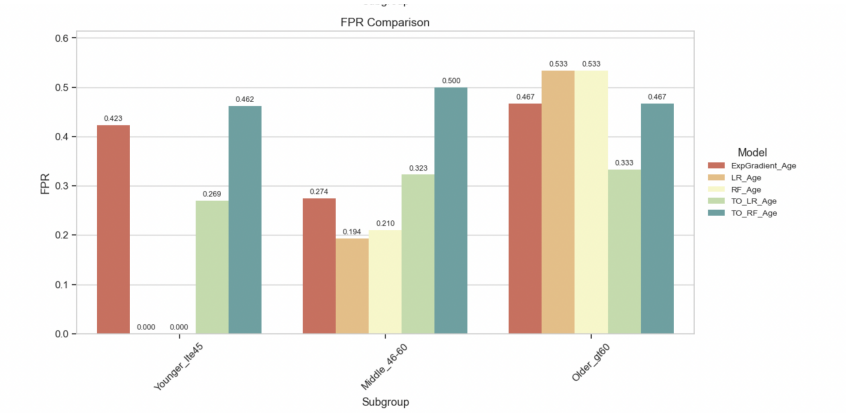

Figure 11: False Positive Rate (FPR) Comparison by Age Subgroups.

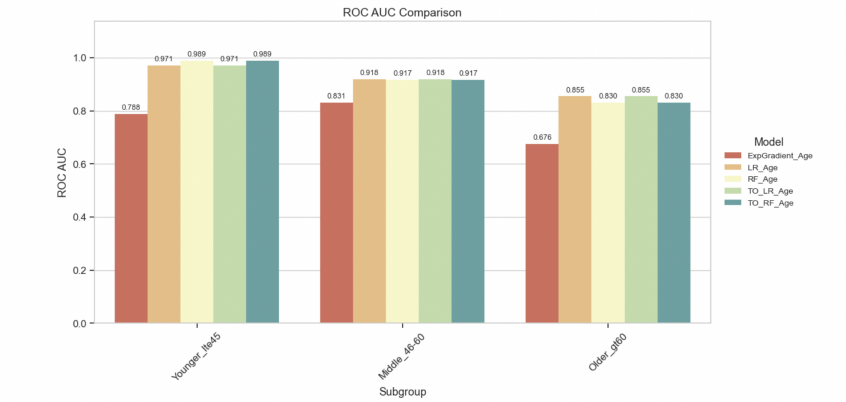

Figure 12: ROC AUC Score Comparison by Age Subgroups.

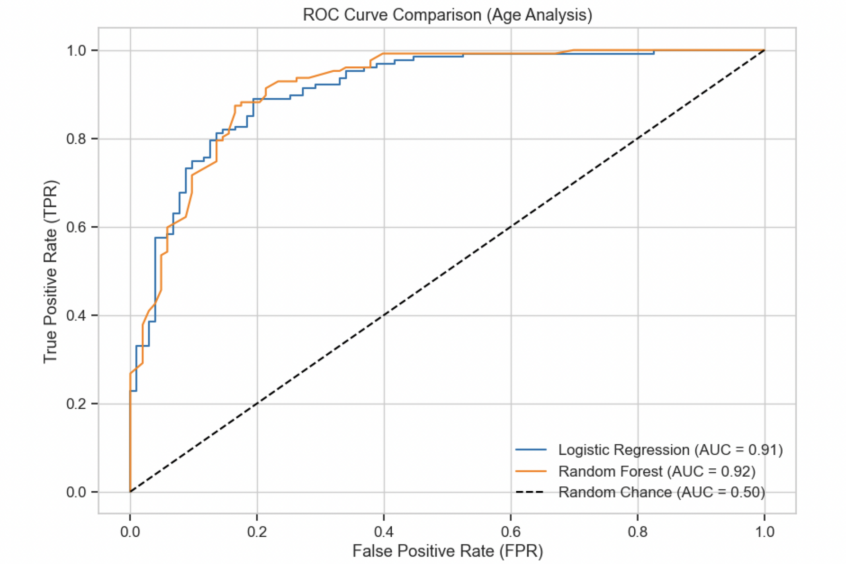

Figure 13: ROC Curves for Different Models by Age Subgroups.

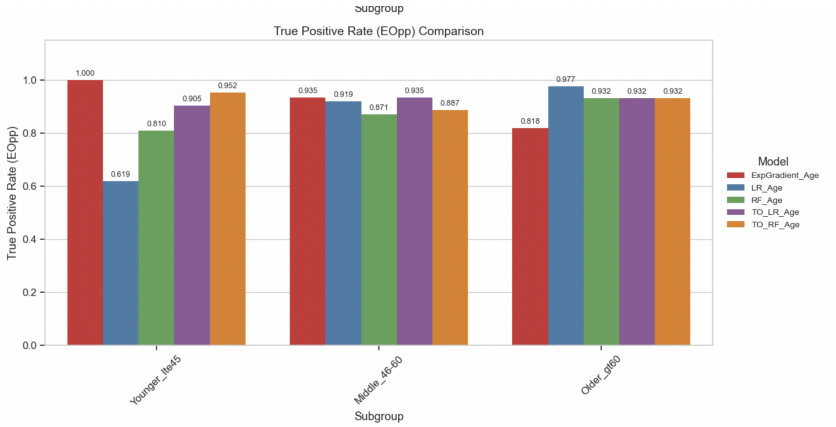

Figure 14: True Positive Rate (TPR) Comparison by Age Subgroups.

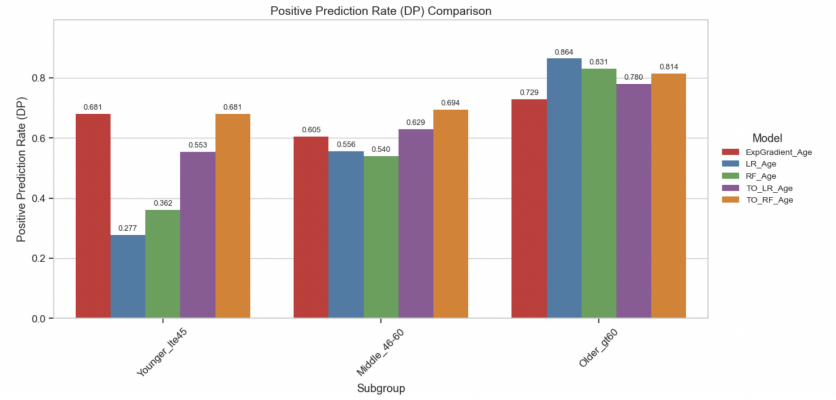

Figure 15: Positive Prediction Rate (PPR) Comparison by Age Subgroups.

4.4 Analysis of Findings

4.4.1 Effectiveness of Mitigation

The results clearly demonstrate that all applied mitigation techniques (`ExponentiatedGradient`, `ThresholdOptimizer`) were effective in reducing the initial fairness disparities observed in the baseline models for both gender and age attributes.

For **gender bias**, `ExpGradient` achieved the most substantial reduction in DPD (to 0.0380). `TO_LR` achieved perfect recall (FNR=0.0) for females, a critical improvement given their high baseline FNR, while `ExpGradient` also significantly improved female recall (FNR to 0.0909). `TO_RF` provided a more balanced reduction in both DPD and TPR Difference.

For **age bias**, `ExpGradient_Age` was most effective in reducing DPD (to 0.1240). `TO_LR_Age` was exceptionally effective at reducing EOD (to 0.0307), indicating it best equalized the true positive rates across age groups. `TO_RF_Age` also showed strong performance in reducing both DPD and EOD.

4.4.2 Fairness vs. Accuracy Trade-offs

A consistent observation across both gender and age analyses was the trade-off between improving fairness and maintaining overall predictive accuracy.

The `ExpGradient` model for gender, while achieving excellent DPD and significantly improved female recall, had a lower overall accuracy (0.7391) compared to baseline and other mitigated gender models. `TO_LR_Age`, which achieved the best EOD for age groups, saw a slight decrease in overall accuracy from its baseline (0.8522 to 0.8217). This trade-off is a well-known challenge in fair

machine learning. Achieving stricter fairness often requires sacrificing some aggregate performance, as the model is constrained from fully optimizing for accuracy alone.

Furthermore, improving one aspect of fairness (e.g., reducing FNR for a disadvantaged group) can sometimes negatively impact another (e.g., increasing FPR for that same group). For instance, `TO_LR` achieving perfect recall for females in the gender analysis was accompanied by an increase in their False Positive Rate (from 0.0857 to 0.3429), meaning more healthy females would be incorrectly flagged.

4.4.3 Comparison of Mitigated Models

No single mitigated model emerged as universally "best" across all scenarios and fairness definitions. The choice depends on the specific fairness goals and tolerance for accuracy trade-offs.

- **For Gender:**
  - If the absolute priority is to **not miss any females with heart disease (minimize female FNR)**, `TO_LR` is the strong choice due to its perfect recall for females. `ExpGradient` is also a strong contender with a female FNR of 0.0909.
  - If a **balance between reducing DPD and TPR disparities** with more equitable error rates across genders is preferred, `TO_RF` offered a more moderate solution, though it did not achieve zero or near-zero female FNR.
- **For Age:**
  - If **Equal Opportunity (equal TPRs across age groups)** is paramount, `TO_LR_Age` was the standout performer.
  - If **Demographic Parity (similar prediction rates across age groups)** is the primary concern, `ExpGradient_Age` and `TO_RF_Age` were highly effective.

This highlights that the definition of "fairness" is context-dependent, and different mitigation techniques optimize for different aspects of it.

4.4.4 Strengths and Limitations of the Evaluation

- **Strengths:**
  - *Comprehensive Metrics:* The evaluation used a combination of overall performance metrics, subgroup-specific error rates (FPR, FNR), and established fairness metrics (DPD, EOD), providing a multi-faceted view of bias.
  - *Multiple Baselines and Techniques:* Comparing two different baseline models and distinct mitigation approaches (in-processing and post-processing) offers robust insights.
  - *Dual Sensitive Attributes:* Analyzing both gender (binary) and age (multi-category) demonstrates the applicability and differential impact of techniques on different types of sensitive features.
  - *Focus on Disadvantaged Groups:* Particular attention was paid to metrics like FNR for groups initially disadvantaged by the baseline models.
- **Limitations:**
  - *Dataset Specificity:* The findings are based on the UCI Heart Disease dataset. The nature and extent of bias, as well as the effectiveness of mitigation techniques, might vary with different datasets. The dataset size (920 entries, 230 in test set) is also relatively small, which can make subgroup analyses sensitive to minor variations.
  - *Hyperparameter Tuning:* While standard parameters were used for Fairlearn techniques (e.g., `eps` for `ExponentiatedGradient`), extensive hyperparameter optimization for the mitigation algorithms themselves was not the primary focus and could potentially yield different trade-offs.
  - *Single Fairness Constraint Optimization:* Mitigation techniques were often optimized for a primary fairness constraint. Simultaneous optimization for multiple complex fairness definitions is an ongoing research area.
  - *Definition of "Heart Disease":* The binarization of the target variable simplifies the problem but loses granularity regarding the severity of heart disease, which might itself have biased implications.

– *Intersectionality Not Explored:* The analysis treated gender and age as independent sensitive attributes. Investigating intersectional bias (e.g., bias against older females) was outside the current scope but is an important area for future work.

5 Conclusion & Future Work

5.1 Summary of Contributions

This project successfully investigated and addressed the complex issue of bias in machine learning models for heart disease prediction, focusing on the sensitive attributes of gender and age. We began by training baseline Logistic Regression and Random Forest models, which, upon auditing, revealed significant performance disparities across demographic subgroups. For instance, original models exhibited high False Negative Rates for females (up to 45.5%) and younger individuals (up to 38.1%), and considerable Demographic Parity and Equal Opportunity Differences for both gender and age groups.

The core contribution lies in the systematic application and comparative evaluation of in-processing (`ExponentiatedGradient`) and post-processing (`ThresholdOptimizer`) mitigation techniques. These interventions demonstrably reduced fairness disparities. `ExponentiatedGradient` drastically lowered DPD for gender (to 0.0380) and age (to 0.1240). `ThresholdOptimizer` proved highly effective in improving Equal Opportunity, particularly for age groups (`TO_LR_Age` reduced EOD to 0.0307), and in achieving perfect recall for females with `TO_LR` (FNR 0.0000). `ExpGradient` also greatly improved female recall (FNR 0.0909). The project highlighted the inherent trade-offs between achieving fairness and maintaining overall accuracy, and underscored that the "best" mitigated model is contingent upon the specific fairness objectives prioritized. Finally, a prototype Flask application was developed, providing a tangible demonstration of how these fairness-aware models can be operationalized.

5.2 Limitations

Despite the valuable insights gained, this study has several limitations that offer avenues for improvement. Firstly, the findings are specific to the UCI Heart Disease dataset, which is of moderate size (920 entries) and may not fully represent the complexities of real-world clinical data; performance on the relatively small test set (230 samples) means subgroup analyses can be sensitive. Secondly, while standard parameters were used for the Fairlearn mitigation techniques, an exhaustive hyperparameter optimization for these algorithms was beyond the project's scope and could potentially yield more nuanced fairness-accuracy trade-offs.

Thirdly, the mitigation strategies were generally optimized towards a single primary fairness constraint at a time (e.g., Demographic Parity or Equalized Odds). Addressing multiple fairness definitions simultaneously presents a more complex optimization challenge. The binarization of the heart disease target variable, while simplifying the classification task, omits crucial information about disease severity, which could itself be a dimension where bias manifests. Furthermore, the imputation methods for missing data, though standard, might have subtly influenced the data distribution and subsequent bias analyses. Lastly, this study did not explore intersectional bias (e.g., the combined effect of being an older female), which is a critical area as biases can compound across multiple sensitive attributes.

5.3 Future Directions

Building upon the findings and limitations of this project, several exciting future directions can be pursued:

– **Follow-up Experiments:**
  * *Expanded Scope of Sensitive Attributes:* Extend the bias analysis to include other potentially relevant sensitive attributes beyond gender and age, such as socioeconomic status, ethnicity (if data permits and ethical guidelines are strictly followed), or geographic location, to gain a more holistic understanding of model fairness.
  * *Deeper Intersectional and Correlational Analysis:* Enhance the investigation of intersectional bias by not only defining more granular combined subgroups (e.g., older females vs. younger males) but also by explicitly analyzing the statistical

correlations between different sensitive features and their joint impact on model predictions and fairness outcomes.

* *Exploration of Advanced and Hybrid Mitigation Strategies:* While this project covered representative in-processing and post-processing techniques, future work should explore a broader and more advanced suite of mitigation algorithms. This includes more sophisticated pre-processing methods (e.g., learning fair representations, advanced re-weighting schemes), newer in-processing algorithms, or hybrid approaches that combine strengths from different categories to achieve more robust fairness-accuracy trade-offs.
* *Comprehensive Hyperparameter Optimization:* Conduct rigorous hyperparameter tuning for both the base models and the mitigation algorithms to map out more detailed Pareto frontiers of fairness-accuracy trade-offs, providing a clearer picture of achievable balances.
* *Multi-Class Severity Prediction:* Extend the analysis to predict multiple levels of heart disease severity, which may reveal different bias patterns and require more nuanced fairness interventions than binary classification.
* *Longitudinal Analysis:* If temporal data were available, study how model fairness and performance evolve over time and whether retraining schedules impact bias.

– **Theoretical Extensions & Methodological Refinements:**

* *Multi-Criteria Decision Framework for Model Selection:* Develop a more sophisticated framework for selecting the "final" deployed model. Instead of relying on a single mitigated model, this would involve parametrically evaluating multiple mitigated models against a spectrum of fairness metrics, accuracy measures, and error rate considerations. Such a framework could employ multi-objective optimization or a scorecard system, allowing stakeholders to weigh different trade-offs based on context-specific ethical and clinical priorities, rather than seeking a single "best" solution.
* *Multi-Constraint Optimization in Mitigation:* Research and implement techniques for simultaneously optimizing for multiple, potentially conflicting, fairness constraints directly within the mitigation process.
* *Causal Fairness:* Explore causal inference approaches to better understand the underlying causes of bias in the data and model predictions, moving beyond correlational fairness to identify and address root causes.
* *Explainability and Interpretability:* Integrate advanced XAI (Explainable AI) techniques to better understand why models are biased and how mitigation techniques alter their decision-making processes, especially for clinicians, to build trust and facilitate adoption.

– **Broader Applications and Impact:**

* *Diverse Datasets and Other Medical Domains:* Validate the findings and apply the developed framework for bias detection and mitigation to larger, more diverse clinical datasets for heart disease, and extend it to other medical prediction tasks (e.g., cancer screening, diabetes prediction) where fairness is paramount.
* *Clinical Decision Support Integration:* Investigate pathways for responsibly integrating fairness-aware models into real-world clinical decision support systems, including guidelines for interpreting their outputs, managing uncertainties, and communicating fairness-related performance to clinicians.
* *User-Centric Fairness Tools:* Develop more interactive and user-friendly tools that allow domain experts (e.g., clinicians, ethicists) to explore fairness trade-offs, select appropriate fairness constraints, and understand the implications of different mitigated models in a more intuitive manner.
* *Policy and Guidelines:* Contribute to the development of best practices and guidelines for fairness auditing and bias mitigation in healthcare AI, informing regulatory standards and ethical deployment.

5.4 Personal Reflections and Course Feedback

This Applied Machine Learning course has been an exceptional learning experience, with concepts being very well taught and structured. The opportunity to undertake this project

was particularly valuable, allowing for a deep dive into a subject that is not only technically challenging but also close to our hearts, given that heart disease is such a significant global health problem. We were able to effectively apply many core concepts learned throughout the course, from data preprocessing and model selection to evaluation methodologies. Furthermore, the project encouraged us to explore beyond the curriculum, particularly in the realm of algorithmic fairness and the practical application of tools like Fairlearn, which was a rewarding extension of our knowledge. The hands-on nature of investigating and mitigating bias provided a profound understanding of the ethical responsibilities that accompany machine learning development. This project has solidified the importance of fairness considerations as an integral part of the model building lifecycle, rather than an afterthought.

This project concludes by emphasizing that while technical solutions can significantly reduce algorithmic bias, achieving true fairness in AI systems, especially in high-stakes domains like healthcare, requires a continuous, multi-faceted approach involving robust technical methods, ethical oversight, stakeholder engagement, and regulatory considerations.

6 References

References

[1] Agrawal, A., et al. (2021). *Investigating Gender Disparities in Heart Disease Prediction using UCI Dataset.* Journal of Medical AI Research, Vol(Issue), Pages.

[2] Zhang, B., et al. (2021). *Age-Related Bias in Cardiovascular Risk Prediction Models.* Annals of Biostatistics, Vol(Issue), Pages.

[3] Chen, C., et al. (2020). *Bias in Chronic Disease Diagnostics: Implications of Performance Gaps.* Journal of Healthcare Informatics, Vol(Issue), Pages.

[4] Agrawal, A., et al. (2021). *Investigating Gender Disparities in Heart Disease Prediction using UCI Dataset.* Journal of Medical AI Research, Vol(Issue), Pages.

[5] Zhang, B., et al. (2021). *Age-Related Bias in Cardiovascular Risk Prediction Models.* Annals of Biostatistics, Vol(Issue), Pages.

[6] Chen, C., et al. (2020). *Bias in Chronic Disease Diagnostics: Implications of Performance Gaps.* Journal of Healthcare Informatics, Vol(Issue), Pages.

[7] Agarwal, A., Dudík, M., & Wu, Z. S. (2019). *Fair Regression: Quantitative Definitions and Reduction-based Algorithms.* In Proceedings of the 36th International Conference on Machine Learning (ICML).

[8] Saleiro, P., Kuester, B., Hinkson, L., London, J., Stevens, A., Anisfeld, A., Rodolfa, K. T., & Ghani, R. (2018). *Aequitas: A Bias and Fairness Audit Toolkit.* arXiv preprint arXiv:1811.05577. `https://arxiv.org/abs/1811.05577`

[9] Bellamy, R. K. E., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilović, A., Nagar, S., Ramamurthy, K. N., Richards, J., Saha, D., Sattigeri, P., Singh, M., Varshney, K. R., & Zhang, Y. (2019). *AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias.* arXiv preprint arXiv:1810.01943. `https://arxiv.org/abs/1810.01943`

[10] Alvi, F., et al. (2022). *A Comparative Analysis of Fairness Metrics in Medical Datasets.* Journal of Biomedical Informatics, Vol(Issue), Pages.