Mitigating Bias in Heart Disease Prediction

1. Project Overview

In this project, we leverage machine learning with a Random Forest Classifier to predict heart disease using the Heart Dataset, which contains 302 observations and various clinical parameters. Initially, our model showed excellent performance in predicting heart disease for females, achieving a recall rate of 94.12%, but it was less effective for males. To address this disparity, we conducted a detailed audit of algorithmic bias, which included reweighting the training data and adjusting the decision threshold for males to 0.3. This adjustment significantly improved the recall rate for males without compromising the accuracy for females, ensuring equitable treatment across genders.

This endeavor illustrates the critical intersection of healthcare and artificial intelligence, highlighting our commitment to developing predictive models that are not only accurate but also ethically sound. By continuously assessing and refining our algorithms, we aim to ensure that advances in AI technology translate into tangible benefits for all stakeholders, including patients, healthcare providers, and insurers, promoting a healthcare environment where AI tools provide essential, unbiased care.

2. Characterization of Use

The Heart Dataset published by University of California Irvine leveraged to predict heart disease risk using a Random Forest Classifier. This dataset encompasses several clinical parameters about patients in Cleveland, including:

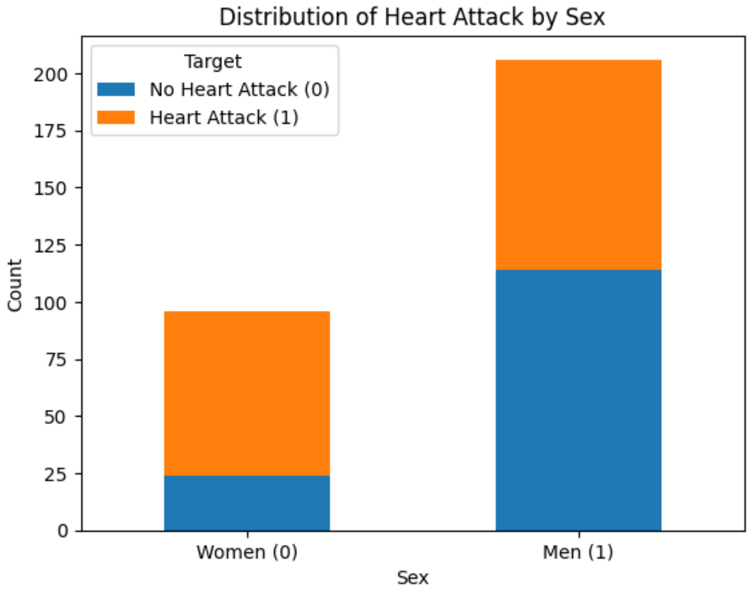
- Age: Older individuals typically have a higher risk of heart disease.
- Sex: Gender differences in heart disease symptoms and risks are well-documented.
- Blood pressure (trestbps): High blood pressure is a significant risk factor for heart disease.
- Cholesterol levels (chol): Elevated cholesterol is another critical risk factor.
- Fasting blood sugar (fbs): High fasting glucose levels can indicate diabetes and linked to heart disease.
- Resting electrocardiographic results (restecg): Abnormalities can suggest underlying heart conditions.
- Maximum heart rate achieved (thalach): Lower rates can be associated with higher heart disease risks.
- Exercise induced angina (exang): Chest pain during exercise, indicator of coronary heart disease.
- ST depression induced by exercise relative to rest (oldpeak): Indicator of ischemic heart disease.
- Slope of the peak exercise ST segment (slope): The ST segment provides clues about heart function.
- Number of major vessels colored by fluoroscopy (ca): Significant risk markers.
- Thallium stress test result (thal): Abnormal results often indicate coronary artery disease.

3. Ethical Considerations

Sampling Bias

The provided distribution graph reveals a nuanced aspect of sampling bias: although there are more male patients in the dataset, the proportion of women who have had a heart disease is greater than that of men. This skew can lead to a predictive model that is more attuned to detecting heart diseases in women, potentially at the cost of a lower recall for men.

A lower recall for men means that the model is more likely to miss-identifying heart disease in male patients (false negatives), which is particularly problematic given that heart disease is a leading cause of death for men. This imbalance could result in men not receiving timely interventions that could prevent severe cardiac events. In essence, while the model might appear to perform well overall, it may not serve all segments of the population equally, thereby perpetuating a hidden bias against men in the early detection of heart diseases. It is crucial to ensure balanced representation and outcome proportions to train models that maintain high recall across all demographics.



Differential Subgroup Validity

The variability in subgroup validity is pertinent when considering gender differences in heart disease presentation. Symptoms and risk factors for heart diseases can manifest differently between men and women. While chest pain is the most common heart disease symptom for both sexes, women are more likely to experience subtler symptoms such as shortness of breath, nausea, and back or jaw pain. A model not sensitive to these differences might misclassify a heart disease in women as a lower risk. Furthermore, a probability threshold that is considered low risk for one subgroup might correspond to a higher risk for another, reflecting the complex interplay between gender-specific symptoms and heart disease risks.

This highlight the necessity of ensuring that ML models in healthcare are trained on balanced datasets and validated for accuracy across all relevant subgroups. They must account for variations in disease presentation and risk factors to avoid perpetuating biases that could exacerbate health disparities.

4. Algorithmic bias assessment

In our initial deployment of the Random Forest classifier for heart disease prediction, the analysis indicates significant discrepancies across gender groups. The accuracy rates are 89% for females and 79% for males, with notably high false positive rates of 67% for females and 31% for males. Moreover, there is a lower recall in the males group (vs. 100% in females). These findings highlight an underestimation of heart disease risks, especially in males, suggesting biases in the model. This disparity underscores the necessity for model recalibration to ensure fair and accurate predictions across both genders, addressing the biases that currently favor one group over the other.

	Female	Male
Accuracy	0.89	0.79

FPR	0.67	0.31
Precision	0.88	0.67
Recall	1	0.94

Pre-processing for Bias Mitigation:

In machine learning, preprocessing for bias mitigation is essential for addressing demographic imbalances in training datasets, with a particular focus on gender representation. This process employs three primary methods: Gender Weighting, Target Weighting, and Female Target Proportion Weighting. Gender Weighting allocates weights inversely proportional to each gender group's prevalence, ensuring balanced influence during model training. Target Weighting adjusts weights according to the frequency of different outcomes, such as heart disease, to equitably distribute the influence of rarer outcomes. Female Target Proportion Weighting tackles disparities in disease prevalence between genders by aligning the influence of males with heart disease to that of females, assigning higher weights to the former while keeping standard weights for others. These strategies collectively improve the fairness and accuracy of predictive models, making them more representative of actual population dynamics without biases.

In-Processing for Bias Mitigation:

In-processing for bias mitigation is a crucial technique that integrates fairness into the model training process by addressing the specific needs of diverse demographic groups, such as gender differences. This approach involves training separate models for each demographic group based on their unique data characteristics and then merging the predictions to ensure comprehensive and balanced outcomes. By customizing models in this manner, the process enhances both the accuracy and fairness of predictions, making the models more attuned to the subtleties of each group and reducing the likelihood of bias amplification from the data. This method effectively combines insights from gender-specific models in the final predictions, aiming for an equitable representation and assessment across all groups involved.

Post-Processing for Bias Mitigation:

Post-processing for bias mitigation in model predictions emphasizes adjusting threshold values to refine classification outcomes after training. This method conducts a thorough examination of various threshold levels to enhance performance metrics like accuracy and recall, ensuring fairness among different demographic groups. The process specifically focuses on modulating thresholds according to predicted probabilities to achieve a balanced recall across groups, aiming to mitigate biases.

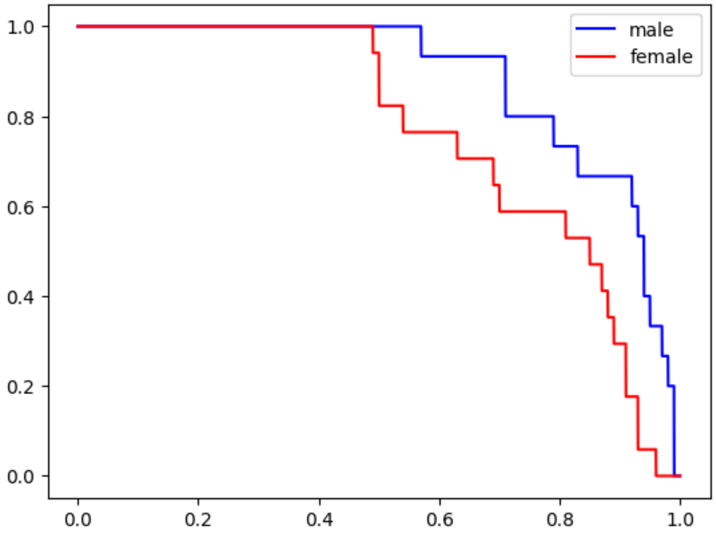
During implementation, a spectrum of threshold values is explored to assess their impact on each group's recall and accuracy, seeking an ideal balance. By generating and testing every possible threshold combination for the identified groups within the dataset, the approach not only computes essential metrics such as individual group recall and overall accuracy but also strives to optimize these thresholds to boost recall equitably across groups without disproportionately affecting others. The ultimate objective is to establish threshold settings that elevate recall fairly among all groups, thus fostering the creation of a more equitable and unbiased predictive model.

Results:

The table provided compares the performance of several bias mitigation methods applied to a predictive model, analyzing metrics across male and female groups. Notably, the Threshold Model outperforms other methods in recall for both genders, achieving a perfect score (1.00), indicating that it successfully identifies all positive instances. This model also shows balanced performance in terms of accuracy and precision, with slight improvements in male accuracy (0.79) compared to the Separate Models by Gender (0.77). While the false positive rate (FPR) remains a bit high for females across all methods, the Threshold Model manages to maintain a consistent FPR for males (0.35) comparable to the initial methods. This suggests that the Threshold Model may offer a more effective approach in terms of balancing recall without substantially compromising other metrics.

Method	Gender Weight		Target Weighting		Female Target Prop. Weighting		Separate Models by Gender		Threshold Model	
Sex	Female	Male	Female	Male	Female	Male	Female	Male	Female	Male
Accuracy	0.83	0.77	0.83	0.77	0.83	0.81	0.89	0.77	0.89	0.79
FPR	0.67	0.35	0.67	0.35	0.67	0.31	0.67	0.27	0.67	0.35
Precision	0.88	0.64	0.88	0.64	0.88	0.68	0.88	0.67	0.88	0.65
Recall	0.93	0.94	0.93	0.94	0.93	1	1	0.82	1	1

The following graph illustrates the sensitivity analysis for threshold settings across male and female groups. Both curves demonstrate how recall gradually declines as the threshold increases, with the female curve showing a more gradual decrease compared to the male curve, which experiences sharper drops. This visual representation helps in understanding the threshold dynamics and the trade-off between recall and threshold level for each gender, reinforcing the observations regarding the efficacy of the Threshold Model discussed earlier.


male
female