Assessing Fairness of Heart Disease Dataset

Yerin Lim, Phwe Thant Chay, Melissa Pax
Northeastern University
CS 6140: Machine Learning
Professor Hongyang Zhang
December 11, 2023

Abstract

Classification models in medical research are common in helping identify risk factors for different disorders. In order to be most effective, the model should be considered fair, and any bias should be mitigated before generalizing across the population. However, data collection can sometimes lead to datasets that are unbalanced, which could potentially lead to classification models that are unfair – creating more accurate predictions for one demographic group compared to the other. A popular dataset, The Heart Failure Prediction Dataset on Kaggle, is heavily skewed to male participants, making the possibility of bias in classification models highly probable. To explore this idea further, this paper uses FairLearn to calculate fairness metrics on gender using four classification models: logistic regression, random forest, decision tree, and support vector machine. Then, two bias mitigation techniques, demographic parity and equalized odds, are utilized and compared with the original models to determine if fairness could be improved. The results found that while the bias mitigation techniques increased fairness for women, the intervention caused accuracy and precision to decrease. This highlights the tradeoff between fairness and accuracy, where improvement in one can lead to performance decreases in another.

Introduction

Heart disease, a term that encompasses several kinds of heart conditions, is the leading cause of death in the United States (Centers for Disease Control and Prevention[CDC], 2023, para. 1). About 47% of the U.S. population has at least one of the three leading risk factors for developing heart disease, which includes high blood pressure, high blood cholesterol, and smoking (CDC, 2023, para. 1). Due to the fatality and the common occurrence of these risk factors, creating models that predict the onset of heart disease has the potential to save lives. However, to optimize the performance, the models need to be fair in their predictions.

Fairness studies and tries to correct bias in machine learning models. These metrics are important in determining if a model is biased toward a particular feature that should bear no weight in the prediction-making process. Sensitive attributes, such as race and gender, are those that may be at particular risk of causing harm in medical predictions, as models may be trained to be not as accurate towards different demographic groups (Haeri & Zweig, 2020, p. 2993).

The Heart Failure Prediction Dataset on Kaggle is a popular dataset that uses various features to predict the occurrence of heart disease. In this dataset, there is an imbalance of a sensitive attribute, sex, where men consist of three-fourths of the data. An imbalance to this extent could lead to models being biased, resulting in overall better performance for men compared to women. To explore this idea further, this paper aims to determine if different classification models can be considered fair using the original dataset, and, if not, if models can be developed to become more fair using two different exponentiated-gradient bias reduction methods: demographic parity and equalized odds.

Demographic parity is a fairness metric that attempts to ensure that a model's predictions are independent of belonging to a sensitive group, which would look like equal selection rates

between the two groups. With equalized odds, the metric attempts to ensure that a model performs equally well for both groups in terms of predictions being independent of a group's sensitive attribute and achieving the same false positive and true positive rates. These metrics were picked because there are no sex differences in developing heart disease with the attributes in the dataset (Gao et al., 2019, p. 2). Therefore, the model should not discriminate against women.

These fairness metrics used will be assessed with help from the Fairlearn library, an open-source Python module that helps assess and fix fairness problems. Fairness will be assessed on three different classification models: Random Forest, Logistic Regression, and Support Vector Machine. Each model will be compared with a naive implementation and the two bias reduction methods (demographic parity and equalized odds) to assess if fairness could be improved. This project is limited to the Heart Failure Prediction dataset on Kaggle and is not meant to generalize to all predictive models involving heart disease.

Methods

The dataset used in this project was obtained from Kaggle's Heart Failure Prediction Dataset. This dataset was created by combining different datasets already available independently but not combined before. In this dataset, 5 heart datasets are combined over 11 common features, making it the largest heart disease dataset available for research purposes. According to Fedesoriano(2021), The five datasets used for its curation are:

- Cleveland: 303 observations
- Hungarian: 294 observations
- Switzerland: 123 observations
- Long Beach VA: 200 observations

- Stalog (Heart) Data Set: 270 observations

The combined data has 1190 observations and 272 of them are duplicated. Therefore, in the final dataset, there are 918 observations. The dataset includes 11 features such as Age, Sex, ChestPainType, RestingBP, Cholesterol, FastingBS, RestingECG, MaxHR, ExerciseAngina, Oldpeak, ST_slope. Numerical features are Age, RestingBP, Cholesterol, MaxHR, and Oldpeak. Categorical features are ChestPainType, RestingECG, and ST_Slope. Binary features are Sex, FastingBS, and ExerciseAngina. The target variable for heart failure prediction is binary, indicating the presence (1) or absence (0) of heart failure. The sensitive feature is Sex in this project.

Table 1

Description of features of the Heart Failure Prediction dataset.

Attribute Title	***Attribute Description***	***Attribute Values***
Age	Patient's age at time of study	Years
Sex	Patient's biological sex	M: Male F: Female
ChestPainType	Type of chest pain the patient experiences	TA: Typical Angina ATA: Atypical Angina NAP: Non-Anginal Pain ASY: Asymptomatic
RestingBP	Patient's resting blood pressure	mm Hg
Cholesterol	Patient's serum cholesterol	mm/dl
FastingBS	Patient's fasting blood sugar	1: If fasting blood sugar > 120 mg/dl 0: If fasting blood sugar <= 120 mg/dl
RestingECG	Patient's resting electrocardiogram result	Normal: Normal results, ST: ST-T wave abnormality, LVH: Displaying probable or definite left ventricular hypertrophy based on Estes's criteria
MaxHR	Patient's maximum heart rate	Beats per minute
ExerciseAngina	Angina induced when patient exercises	Y: Yes N: No
Oldpeak	ST Segment	Numeric value that represents the depression of the segment
ST_Slope	Slope of the peak exercise ST segment	Up: Upsloping, Flat: Flat, Down: Downsloping
HeartDisease	Incidence of heart disease	1: Heart disease 0: Normal

*Note***.** This dataset uses twelve different features to predict the incidence of heart disease.

During the data preprocessing phase, abnormal data were found which indicated zero values for RestingBP and Cholesterol. The zero values were replaced by the median values of each feature.

Table 2

Descriptive Statistics

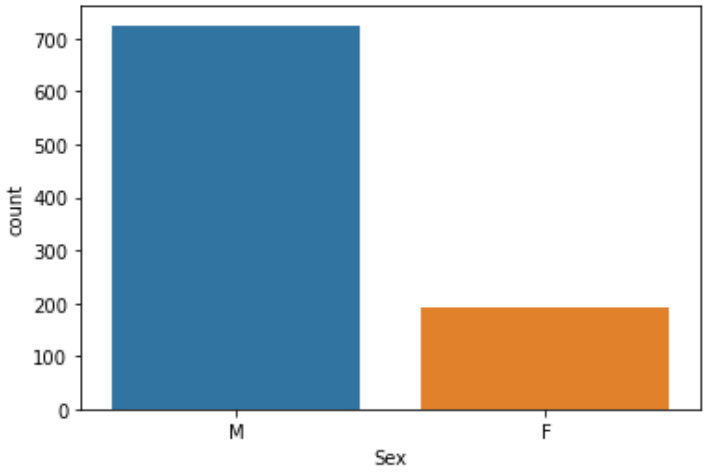
	count	mean	std	min	25%	50%	75%	max
Age	918.0	53.510893	9.432617	28.0	47.0	54.0	60.0	77.0
RestingBP	918.0	132.538126	17.990127	80.0	120.0	130.0	140.0	200.0
Cholesterol	918.0	243.204793	53.401297	85.0	214.0	237.0	267.0	603.0
FastingBS	918.0	0.233115	0.423046	0.0	0.0	0.0	0.0	1.0
MaxHR	918.0	136.809368	25.460334	60.0	120.0	138.0	156.0	202.0
Oldpeak	918.0	0.887364	1.066570	-2.6	0.0	0.6	1.5	6.2
HeartDisease	918.0	0.553377	0.497414	0.0	0.0	1.0	1.0	1.0

*Note***.** This table shows the descriptive statistics for the numerical data and binary data in the dataset. For each feature, the occurrence (count), mean, standard deviation (std), minimum (min), the 25th percentile (25), 50th percentile (50), and the 75th percentile (75), and the maximum (max) were calculated.

For categorical features, the OneHotEncoder was used to transform categorical variables into a binary matrix, where each category becomes a binary column. For binary features, the OrdinalEncoder was used to assign numerical values to each category in a way that preserves the ordinal relationship between them.

Figure 1

The number of Males and Females in the dataset

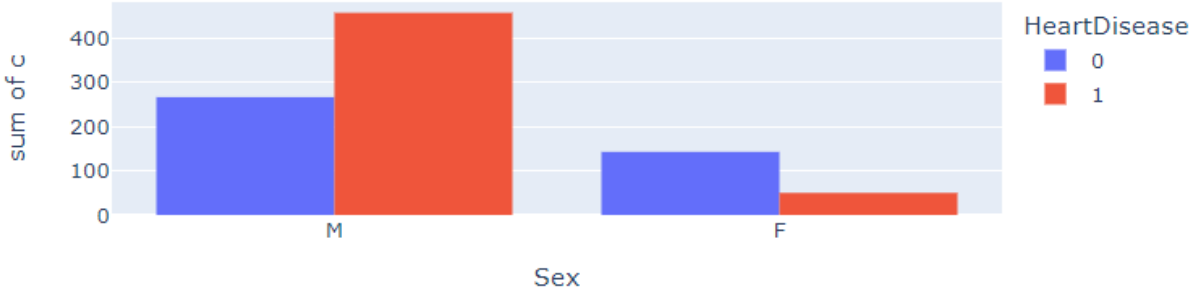


Note. This bar graph displays the occurrence of men and women in the Heart Failure Prediction Dataset. In total, there are 724 males and 193 women, meaning that males consist of 79% of the dataset.

The dataset is imbalanced as the number of positive heart failure is 508 and the number of negative heart failure is 410. Also, there is a significant difference between the number of males (725) and females (193) for the 'Sex' feature. The imbalance in the feature could introduce biases in the model because the larger category may have a more significant impact on model training.

Figure 2

Distribution of outcome (Stratified by Sex)



Note. This bar graph compares the occurrence of heart disease between men and women in the dataset.

This research used Machine Learning techniques such as logistic regression, random forest, and support vector machine for heart disease prediction. For logistic regression and support vector machine, the StandardScaler was used for numerical features to standardize and normalize features for consistency and ease of interpretation. However, decision tree and random forest models are not suitable for standard scaling because they make splits based on feature values but do not rely on the absolute scale of those values.

Logistic regression (LR) is a powerful and well-established method for supervised classification (as cited in Uddin et al., 2019, p.2). It provides straightforward interpretability and it predicts probabilities rather than class labels. LR helps in finding the probability that a new instance belongs to a certain class. Since it is a probability, the outcome lies between 0 and 1. Therefore, to use the LR as a binary classifier, a threshold needs to be assigned to differentiate two classes (as cited in Uddin et al., 2019, p.2).

Random Forests aggregate multiple decision trees, introducing an element of diversity that aids in capturing complex relationships within the dataset. This ensemble approach contributes to the burstiness of the model, accommodating both intricate patterns and potential nuances in the data. By leveraging the strengths of individual trees, the random forest promotes robustness in heart failure prediction, ensuring a nuanced and comprehensive evaluation.

Support Vector Machine (SVM) is useful for handling high-dimensional data and is particularly effective in scenarios where the margin between different classes is crucial. This model introduces an element of complexity, enhancing the perplexity of the overall predictive framework. By exploring intricate patterns in the data space, the SVM aims to capture subtle relationships that might be overlooked by simpler models, thereby contributing to the project's goal of developing a more accurate and nuanced heart failure prediction model.

Table 3

Parameters for Machine Learning Models

Machine Learning Model	Parameters
Logistic Regression	C: [0.001, 0.01, 0.1, 1, 10] penalty: ['none', 'l2']
Random Forest	n_estimators: [50, 100, 200] max_depth: [10, 20, 30] min_samples_split: [2, 5, 10] min_samples_leaf: [1, 2, 4]
SVM	C: [0.1, 1, 10, 100] gamma: [1, 0.1, 0.01, 0.001] kernel: ['rbf', 'poly', 'sigmoid']

Note. This table shows the hyperparameters each model was tested for using grid search cross-validation. C represents the inverse of the regularization strength; penalty adds a penalty term to the model; min_samples_leaf represents the number of samples required for a leaf node; max_depth represents the maximum depth of the tree; n_estimators represents how many trees exist in the forest; min_samples_split represents the number of samples for a leaf node; gamma represents the kernel coefficient; and kernel represents the kernel type to be used.

Fairness Metrics

To measure fairness in heart failure prediction, we leveraged the Fairlearn library, a powerful tool designed to facilitate fairness assessment and mitigation in machine learning models. Specifically, we employed the exponentiated gradient methodology within the fairlearn framework to measure and calculate fairness metrics.

The exponentiated gradient algorithm is an iterative optimization approach utilized for achieving demographic parity and equalized odds. It operates by iteratively adjusting the weights assigned to different instances in the dataset to minimize the chosen fairness metric. This iterative process converges towards a fairer model by dynamically assigning more importance to instances that contribute to unfairness.

Demographic parity is a fundamental fairness concept that assesses whether the model's predictions are consistent across different demographic groups. The demographic parity

difference is calculated as the absolute difference between the probabilities of positive outcomes for different groups. Mathematically, it is represented as

$$DPD = \left| P(\widehat{Y} = 1 | A = a_1) - P(\widehat{Y} = 1 | A = a_2) \right|$$

In this equation, $\widehat{Y}$ represents the predicted outcome and A represents the sensitive attribute, sex, in this project. The goal is to minimize the disparities in predicted positive outcomes across various demographic groups.

Equalized odds consider both false positives and false negatives across different groups. It measures the difference in true positive rates and false positive rates between the privileged and unprivileged groups. Mathematically, it is represented as:

$$EOD = \left| P(\widehat{Y} = 1 | A = a_1, Y = 1) - P(\widehat{Y} = 1 | A = a_2, Y = 1) \right| + \left| P(\widehat{Y} = 1 | A = a_1, Y = 0) - P(\widehat{Y} = 1 | A = a_2, Y = 0) \right|$$

In this equation, Y represents the true outcome. The goal is to minimize the disparities in both true positive and false positive rates between different demographic groups, ensuring that the model's performance is equitable across all segments of the population.

Within fairlearn, the exponentiated gradient algorithm can be applied to calculate both the demographic parity difference (DPD) and equalized odds difference (EOD) metrics. By utilizing Fairlearn's implementation of this algorithm, we ensure a systematic and efficient optimization process, aligning our model's predictions with the desired fairness objectives.

The exponentiated gradient algorithm is designed to address fairness concerns by iteratively adjusting the weights assigned to different instances in a dataset, aiming to minimize the disparity in model outcomes across sensitive attribute groups. The algorithm operates by exponentiating the gradient of a chosen fairness metric and using this information to reweight the

training instances. This iterative process allows the model to learn and adapt its predictions in a way that mitigates biases and aligns with predefined fairness objectives.

In the evaluation of our heart failure prediction model, we considered a comprehensive set of performance metrics to gain a nuanced understanding of its effectiveness. Each metric provides unique insights into different aspects of model performance.

Table 4

Fairness metric descriptions

Metric	***Definition***
Accuracy	The proportion of correctly predicted instances among the total instances
Precision	The proportion of true positive prediction among all positive predictions
False Positive Rate	The proportion of false positives among all actual negatives
False Negative Rate	The proportion of false positives among all actual negatives
True Positive Rate	The proportion of true positives among all actual positives
True Negative Rate	The proportion of true negatives among all actual negatives
Selection Rate	The proportion of instances predicted as positive by the model
Recall Score	The proportion of true positives among all actual positives

Note. This table explains the fairness metrics used in this analysis: accuracy, precision, false positive rate, false negative rate, true positive rate, true negative rate, selection rate, and recall score. These are calculated for all three models before and after bias mitigation.

Results

Parameter Choices

The dataset was segmented based on sex, with males and females forming distinct groups for analysis. Three fairness metrics were used: overall metrics before bias mitigation, metrics after bias mitigation by demographic parity, and metrics after bias mitigation by equalized odds. The choice of these metrics was to capture different aspects of fairness: overall performance, bias in selection rate, and balance in true and false positive rates. Demographic parity focuses on ensuring equal outcomes across groups, while equalized odds aim to equalize error rates across

groups. The choice of these metrics was likely driven by the nature of the task and the importance of fairness in the application context.

Fairness Metrics – Demographic Parity

This metric focuses on ensuring that a model's predictions are not influenced by membership in a sensitive group. In binary classification, it translates to equal selection rates across these groups. For instance, in a resume screening application, it would mean that the proportion of applicants shortlisted for interviews is the same across different demographic groups.

Fairness Metrics – Equalized Odds

This metric demands that all groups have identical rates for false positives and false negatives. It's a more stringent criterion than demographic parity, as it requires the model's predictions to be not just independent of sensitive group membership, but also balanced in terms of false positive and true positive rates. The significance here is that a model can achieve demographic parity—where predictions are uncorrelated with group membership—yet still produce disparate false positives among different groups.

Below are the results for all three models before bias mitigation and after bias mitigation with demographic parity and equalized odds.

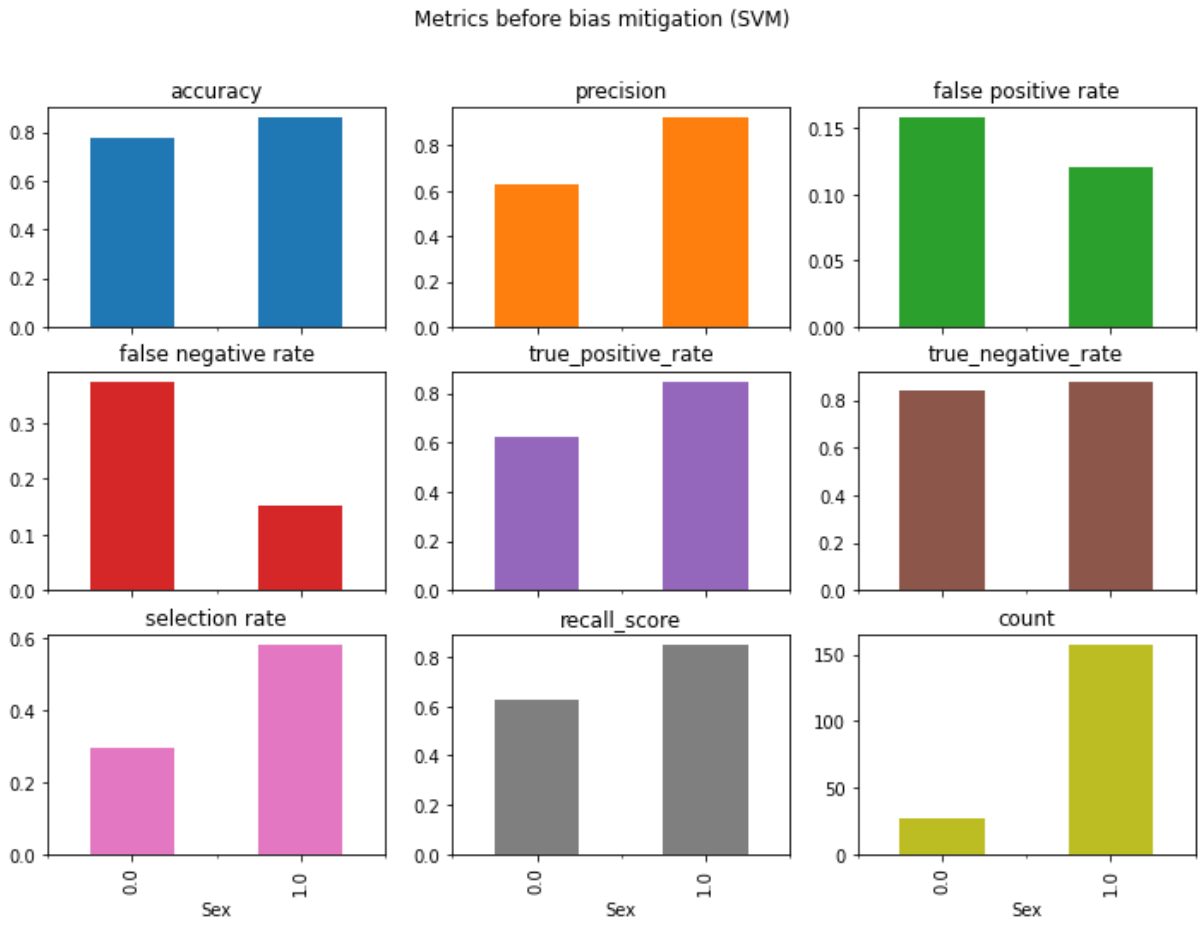
Support Vector Machine

The comprehensive results presented offer a nuanced understanding of the performance and fairness of a Support Vector Machine (SVM) model both before and after the application of

bias mitigation strategies. The goal of these strategies is to ensure that the model's decisions do not systematically disadvantage a particular group based on sensitive attributes, such as gender.

Figure 3

Support Vector Machine Before Mitigation



Note. Bar graphs comparing the metrics between sexes before implementing any bias mitigation in the support vector machine model. Women are represented by 0 and men are represented by 1.

Table 5

Support Vector Machine Before Mitigation

Metrics Type	*Overall*	*Male*	*Female*
Accuracy	84.78 %	85.99 %	77.78 %
Precision	89.90 %	92.31 %	62.50 %
False Positive Rate	12.99 %	12.07 %	15.78 %
False Negative Rate	16.82 %	15.15 %	37.50 %
True Positive Rate	83.18 %	84.85 %	62.50 %
True Negative Rate	87.01 %	87.93 %	84.21 %
Selection Rate	53.80 %	57.96 %	29.63 %
Recall Score	83.18 %	84.85 %	62.50 %
Count	184	157	27

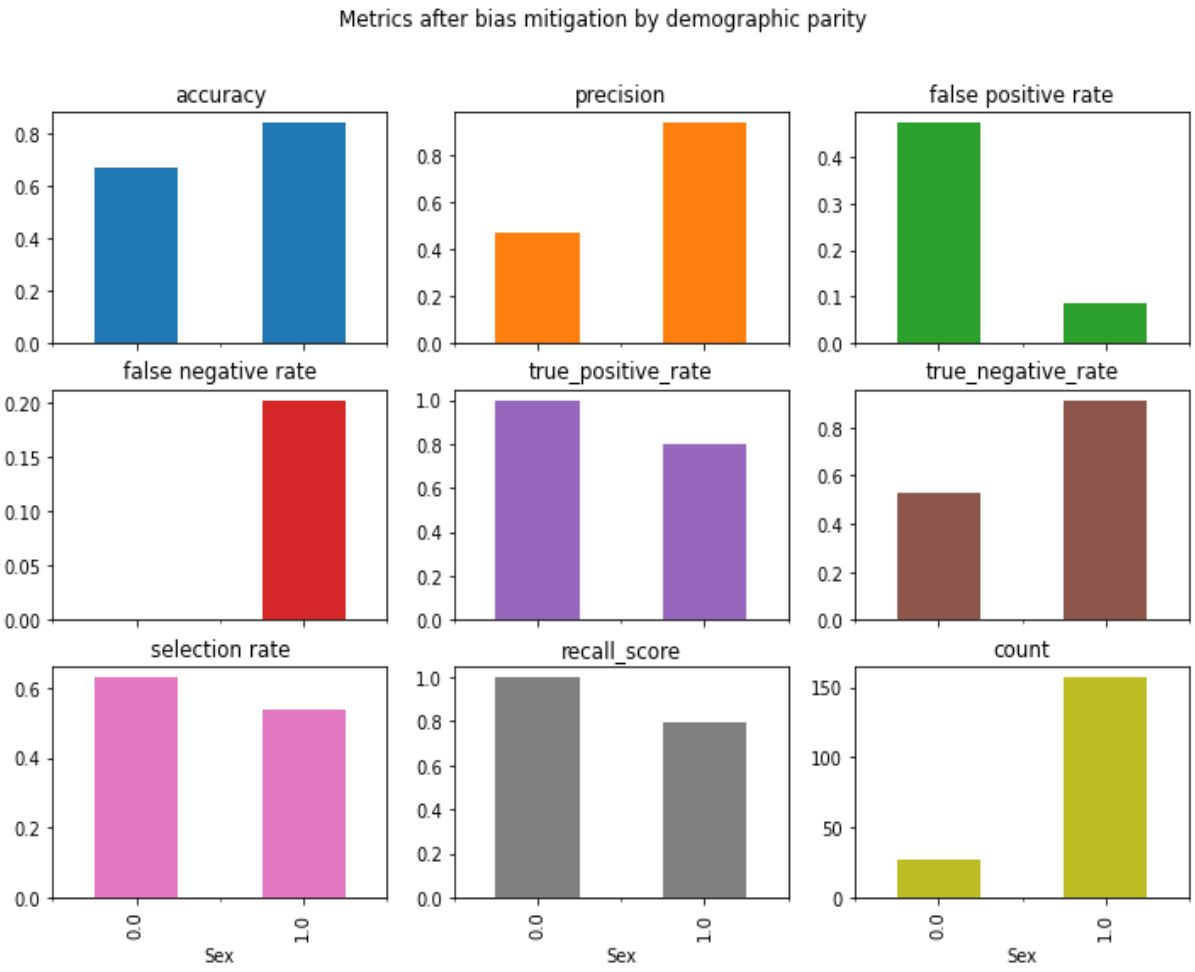
Note. The table below displays information about the metrics in men, women, and both sexes together.

The baseline Support Vector Machine (SVM) model demonstrated commendable performance, with an impressive overall accuracy of 84.78% and a high precision of 89.90%. However, the model revealed notable disparities in various metrics such as the false positive rate, false negative rate, true positive rate, and recall score when comparing outcomes between different sexes. Notably, there was a marked difference in the true positive rate between the sexes: it was significantly lower for females at 62.50%, compared to 84.85% for males. This disparity highlights a critical area for improvement in the model's ability to equally and accurately classify different sexes.

Demographic Parity

Figure 4

Demographic Parity Support Vector Machine After Mitigation.



Note. Bar graphs comparing the metrics between sexes after implementing demographic parity in the support vector machine model. Women are represented by 0 and men are represented by 1

Table 6

Demographic Parity Support Vector Machine After Mitigation.

Metrics Type	*Overall*	*Male*	*Female*
Accuracy	81.52 %	84.08 %	66.67 %
Precision	86.14 %	94.04 %	47.06 %
False Positive Rate	18.18 %	8.62 %	47.37 %
False Negative Rate	18.69 %	20.20 %	0.00 %
True Positive Rate	81.30 %	79.80 %	100 %
True Negative Rate	81.82 %	91.38 %	52.63 %
Selection Rate	54.89 %	62.96 %	53.50 %
Recall Score	81.31 %	79.80 %	100 %
Count	184	157	27

*Note***.** The table below displays information about the metrics in men, women, and both sexes together.

Upon implementing bias mitigation strategies aimed at achieving demographic parity, the model underwent a slight decline in overall accuracy and precision, dropping to 81.52% and 86.14%, respectively. This shift underscores the often-observed trade-off between achieving fairness and maintaining high-performance metrics.

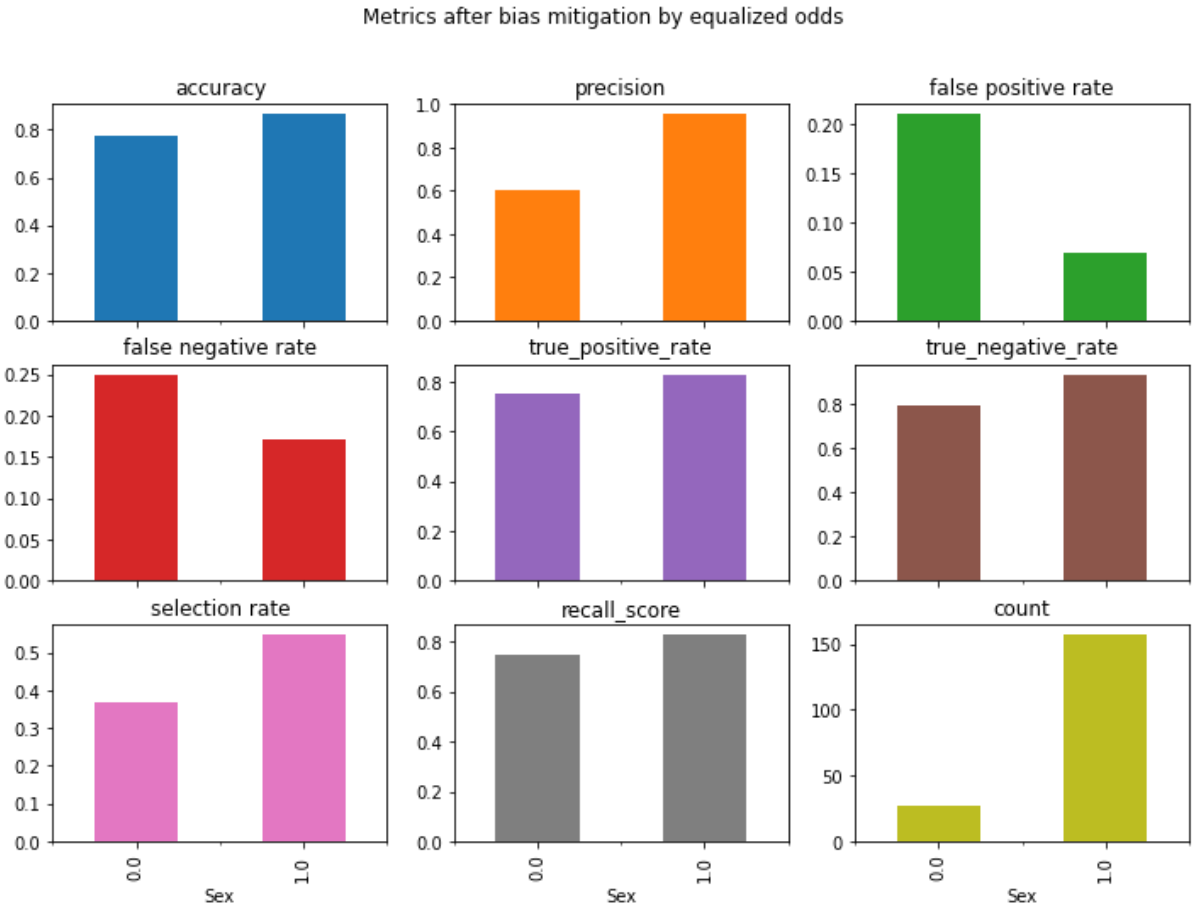
The adjustment resulted in a considerable improvement in equalizing the selection rate between sexes, demonstrating a move towards greater fairness. However, this adjustment also introduced a trade-off in error rates: there was an increase in the false positive rate for females and a rise in the false negatives for males.

Notably, the true positive rate for females saw a remarkable increase, reaching 100%. This indicates that, post-mitigation, the model is now exceptionally effective at correctly identifying all positive cases in the female group, marking a significant step forward in gender-based performance equity.

Equalized Odds

Figure 5

Equalized Odds Support Vector Machine After Mitigation



Note. Bar graphs comparing the metrics between sexes after implementing equalized odds in the support vector machine model. Women are represented by 0 and men are represented by 1.

Table 7

Equalized Odds Support Vector Machine After Mitigation

Metrics Type	*Overall*	*Male*	*Female*
Accuracy	85.33 %	86.62 %	77.78 %
Precision	91.67 %	95.35 %	60.00 %
False Positive Rate	10.39 %	6.90 %	21.05 %
False Negative Rate	17.76 %	17.17 %	25.00 %
True Positive Rate	82.24 %	82.83 %	75.00 %
True Negative Rate	89.61 %	93.10 %	78.95 %
Selection Rate	52.17 %	54.78 %	37.04 %
Recall Score	82.24 %	82.83 %	75.00 %
Count	184	157	27

Note. The table below displays information about the metrics in men, women, and both sexes together.

When bias mitigation focused on equalized odds was applied, the model showed an improvement in overall accuracy and precision, achieving 85.33% and 91.67% respectively, which is notably better than the results obtained with demographic parity-based mitigation. This approach effectively reduced disparities in false positive and negative rates, though it did not entirely eliminate these discrepancies.

Furthermore, there was a more balanced distribution in the true positive rate and recall score across different sex groups. This indicates that the model's performance is now more equitable between sexes, suggesting a significant stride towards fairer outcomes in its predictions.

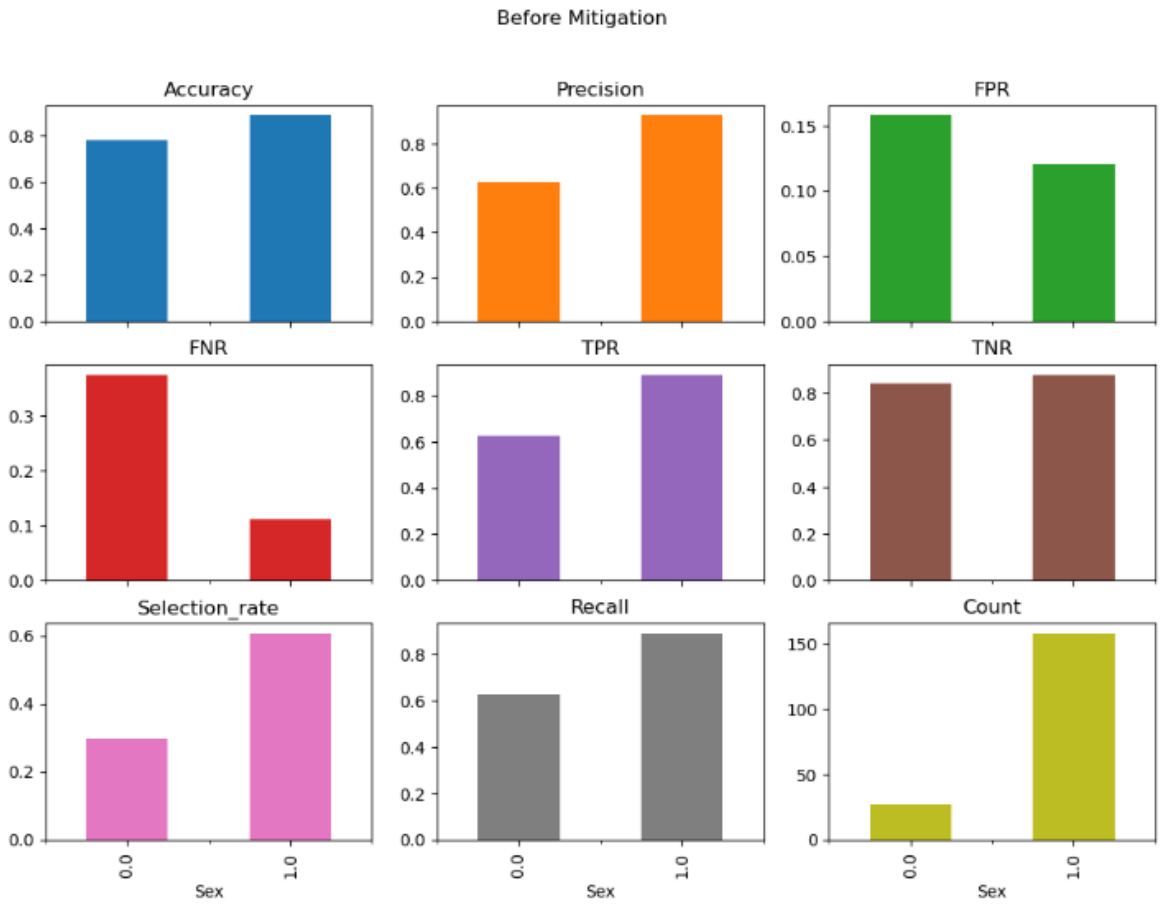
Random Forest

The detailed results give a thorough understanding of how the Random Forest model predicts heart failure and measures fairness. This analysis includes the performance evaluation before and after the implementation of bias mitigation techniques. The primary goal of these

techniques is to guarantee that the model's predictions avoid systematically disadvantaging specific groups based on sensitive attributes, such as sex.

Figure 6

Random Forest Model Before Mitigation



Note. Bar graphs comparing the metrics between sexes before implementing bias reduction in the random forest model. Women are represented by 0 and men are represented by 1.

Table 8

Random Forest Model Before Mitigation

Metrics Type	*Overall*	*Male*	*Female*
Accuracy	86.96 %	88.54 %	77.78 %
Precision	90.29 %	92.63 %	62.50 %
False Positive Rate	12.99 %	12.07 %	15.79 %
False Negative Rate	13.08 %	11.11 %	37.50 %
True Positive Rate	86.92 %	88.89 %	62.50 %
True Negative Rate	87.01 %	87.93 %	84.21 %
Selection Rate	55.98 %	60.51 %	29.62 %
Recall Score	86.92 %	88.89 %	62.50 %
Count	184	157	27

Note. The table below displays information about the metrics in men, women, and both sexes together.

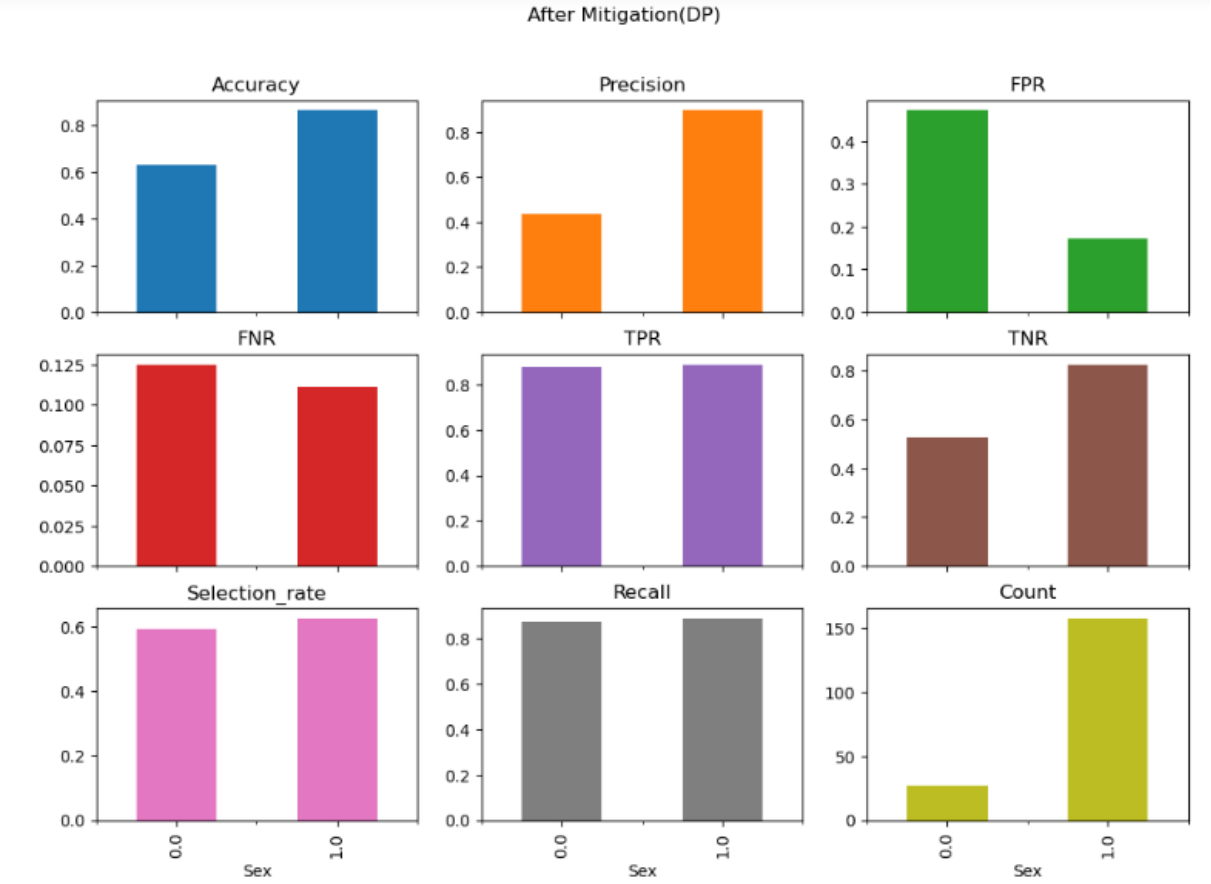
The initial evaluation of the heart failure prediction model shows a strong overall performance, with an accuracy of 86.96%, highlighting the model's proficiency in correctly classifying instances. The precision, recall, and true negative rate all surpassed 85%, indicating a balanced capacity to identify positive cases while minimizing false positives. However, a notable discrepancy emerged in gender-specific metrics, with males consistently outperforming females in accuracy, precision, recall, and true negative rate.

For the fairness metrics, the heart failure prediction model exhibited noticeable disparities, with a Demographic Parity Difference of 0.31 and an Equalized Odds Difference of 0.26 as demonstrated in Figure 16. These metrics indicated existing imbalances in predictions across different demographic groups.

Demographic Parity

Figure 7

Demographic Parity Random Forest Model After Mitigation



Note. Bar graphs comparing the metrics between sexes after implementing demographic parity in the random forest model. Women are represented by 0 and men are represented by 1.

Table 9

Demographic Parity Random Forest Model After Mitigation

Metrics Type	*Overall*	*Male*	*Female*
Accuracy	83.15 %	86.62 %	62.96 %
Precision	83.33 %	89.80 %	43.75 %
False Positive Rate	24.68 %	17.24 %	12.50 %
False Negative Rate	11.22 %	11.11 %	12.50 %
True Positive Rate	88.79 %	88.89 %	87.50 %
True Negative Rate	75.32 %	82.76 %	52.63 %
Selection Rate	61.96 %	62.42 %	59.26 %
Recall Score	88.79 %	88.89 %	87.50 %
Count	184	157	27

Note. The table below displays information about the metrics in men, women, and both sexes together.

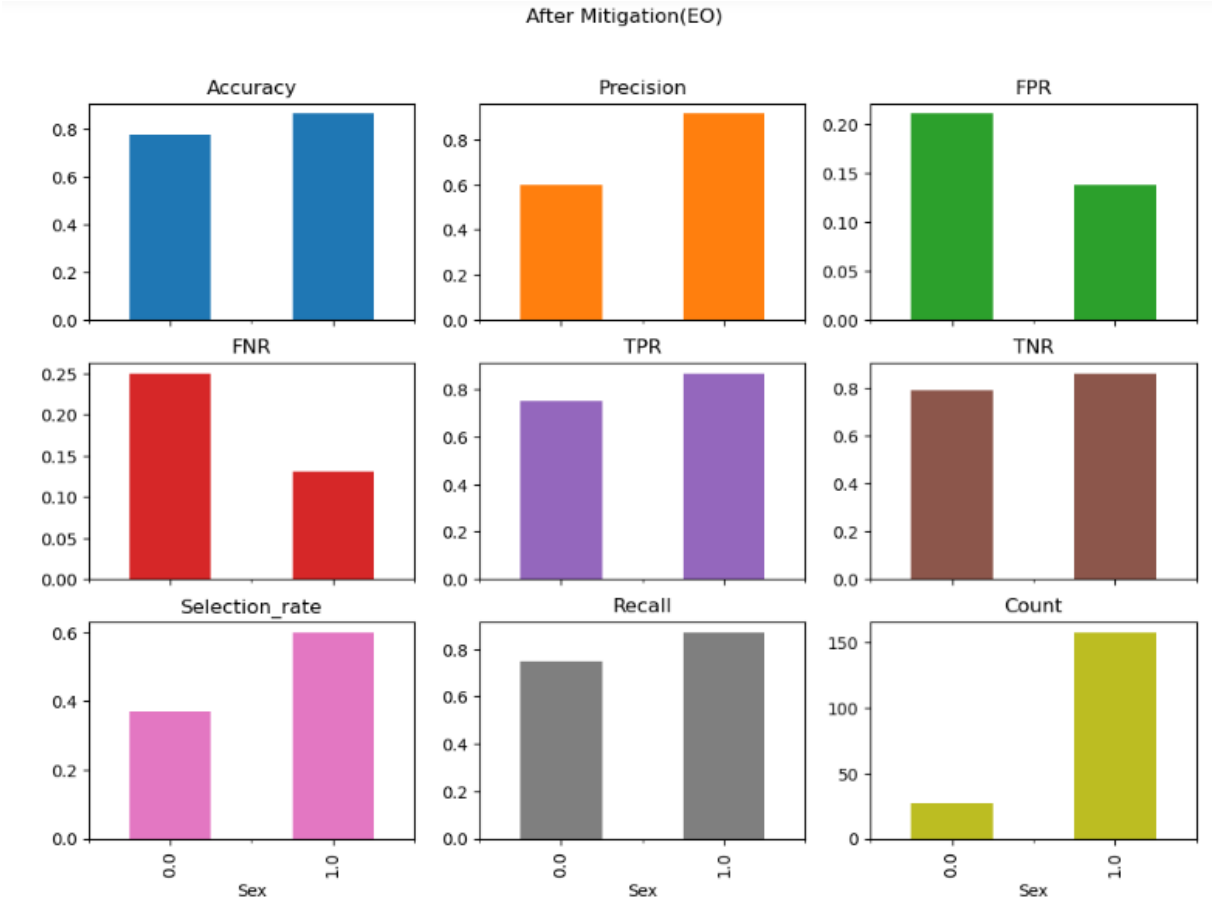
This algorithm aimed to reduce gender-related disparities by equalizing the selection rates between males and females. The result was a slight decrease in overall accuracy to 83.15%, accompanied by improvements in precision and recall for females. While the model's overall accuracy decreased, the trade-off reflected a conscious effort to mitigate gender-based biases, enhancing fairness considerations.

After implementing the Random Forest model with demographic parity mitigation, there was a significant reduction in bias as shown in Figure 16. The Demographic Parity Difference decreased substantially to 0.03, suggesting a noteworthy improvement in equalizing predictions across demographic segments. However, the Equalized Odds Difference increased slightly to 0.3, indicating a potential trade-off in achieving equalized outcomes for positive and negative instances.

Equalized Odds

Figure 8

Equalized Odds Random Forest Model After Mitigation



Note. Bar graphs comparing the metrics between sexes after implementing equalized odds in the random forest model. Women are represented by 0 and men are represented by 1.

Table 10

Equalized Odds Random Forest Model After Mitigation

Metrics Type	***Overall***	***Male***	***Female***
Accuracy	85.33 %	86.62 %	77.78 %
Precision	88.46 %	91.49 %	60.00 %
False Positive Rate	15.58 %	13.79 %	21.05 %
False Negative Rate	14.02 %	13.13 %	25.00 %
True Positive Rate	85.98 %	86.87 %	75.00 %
True Negative Rate	84.42 %	86.21 %	78.95 %
Selection Rate	56.52 %	59.87 %	37.03 %
Recall Score	85.98 %	86.87 %	75.00 %
Count	184	157	27

*Note***.** The table below displays information about the metrics in men, women, and both sexes together.

The application of the equalized odds mitigation strategy yielded a more refined outcome. The overall accuracy increased by 2.32% compared to the demographic parity mitigation, suggesting that this strategy achieved a more balanced performance across gender groups. Precision exhibited a notable 5.96% increase from the demographic parity stage, indicating a more precise identification of positive cases. However, there is a trade-off, with a slight decrease in recall and an increase in false negative rates. The changes in performance metrics indicate that the equalized odds strategy aimed to find a balance between minimizing biases and maintaining a reasonable level of overall accuracy.

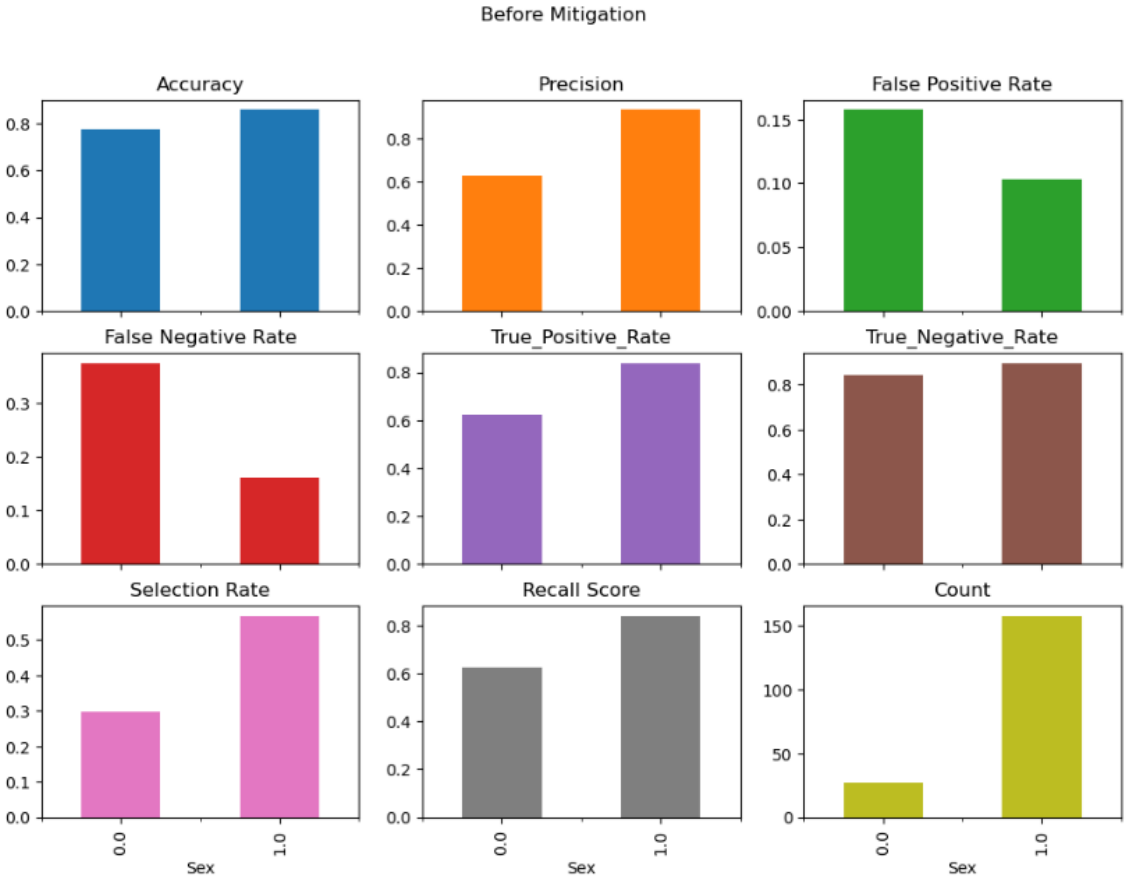
In the application of the Random Forest model with equalized odds mitigation, the Demographic Parity Difference was reduced further to 0.23, which indicates ongoing improvements in minimizing disparities. Additionally, the Equalized Odds Difference decreased to 0.12, indicating a successful reduction in differences between demographic groups concerning both false positives and false negatives. These differences are shown in Figure 16.

Logistic Regression

The same procedures as described above were applied to the logistic regression. Bias mitigation with demographic parity and equalized odds aimed to make the model more fair in regard to sex.

Figure 9

Logistic Regression Before Mitigation



Note. Bar graphs comparing the metrics between sexes before implementing bias reduction in the logistic regression model. Women are represented by 0 and men are represented by 1.

Table 11

Logistic Regression Before Mitigation

Metrics Type	*Overall*	*Male*	*Female*
Accuracy	84.78 %	85.99 %	77.78 %
Precision	90.72 %	93.26 %	62.50 %
False Positive Rate	11.69 %	10.34 %	15.79 %
False Negative Rate	17.75 %	16.16 %	37.50 %
True Positive Rate	82.24 %	83.83 %	62.50 %
True Negative Rate	88.31 %	89.66 %	84.21 %
Selection Rate	52.71 %	56.68 %	29.63 %
Recall Score	82.24 %	83.84 %	62.50 %
Count	184	157	27

*Note***.** The table below displays information about the metrics in men, women, and both sexes together.

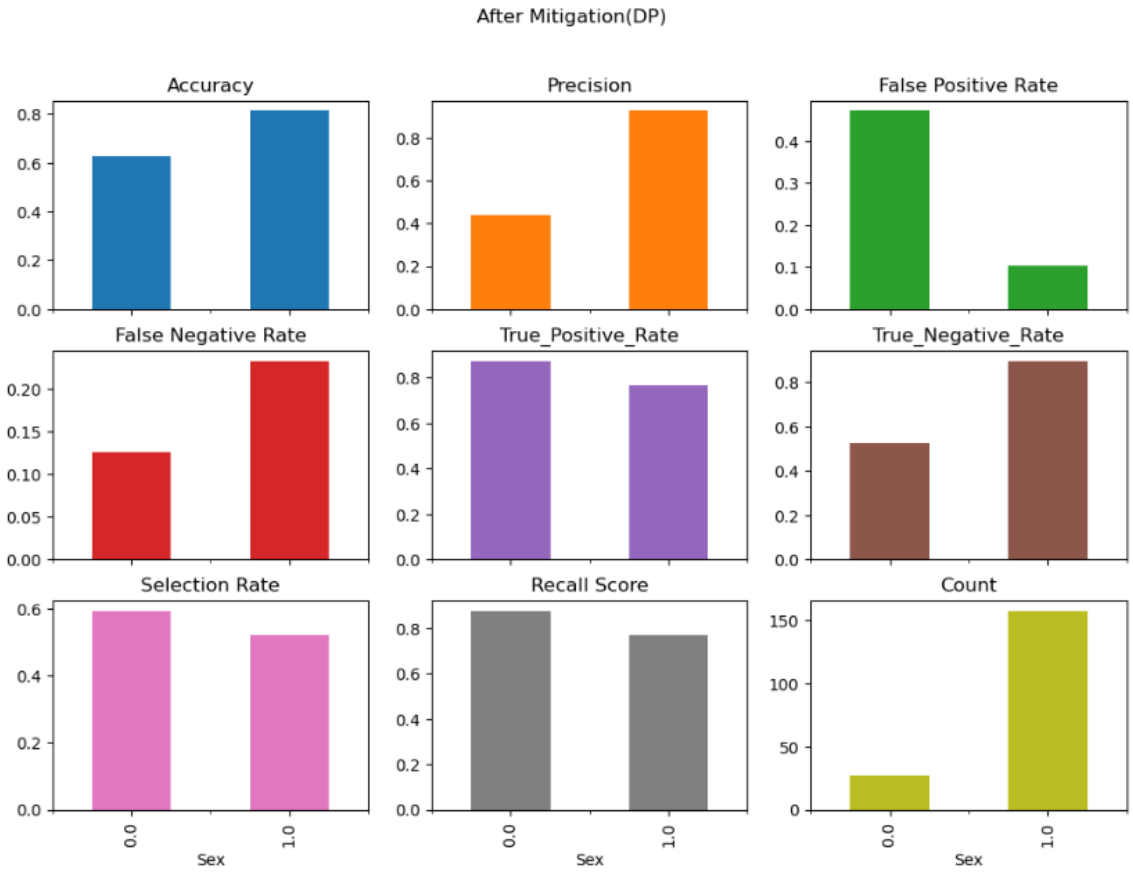
Logistic regression before mitigation appeared to be effective in generating predictions for heart disease in this population. Accuracy, precision, true positive rate, and true negative rate were all above 80%, and the false positive rate and false negative rate were below 20%.

When separated by gender, the model performed worse for women across all metrics. The model with only men performed the best, with all metrics improving compared to the model with both sexes. Some of these metrics have large differences between the sexes, including precision (93.26% for men compared to 62.50% for women), false negative rate (16.16% for men compared to 37.50% for women), true positive rate (83.83% for men compared to 62.50% for women), and selection rate (56.68% for men compared to 29.63% for women). Figure 13 has more details on each of the metrics for both sexes.

Demographic Parity

Figure 10

Demographic Parity Logistic Regression Model After Mitigation



*Note***.** Bar graphs comparing the metrics between sexes after implementing demographic parity in the logistic regression model. Women are represented by 0 and men are represented by 1.

Table 12

Demographic Parity Logistic Regression Model After Mitigation

Metrics Type	*Overall*	*Male*	*Female*
Accuracy	78.80 %	81.53 %	62.96 %
Precision	84.69 %	92.68 %	43.75 %
False Positive Rate	19.48 %	10.34 %	47.37 %
False Negative Rate	22.43 %	23.23 %	12.50 %
True Positive Rate	77.57 %	76.77 %	87.50 %
True Negative Rate	80.52 %	89.66 %	52.63 %
Selection Rate	53.26 %	52.22 %	59.26 %
Recall Score	77.57 %	76.77 %	87.50 %
Count	184	157	27

Note. The table below displays information about the metrics in men, women, and both sexes together.

Demographic Parity bias mitigation technique was applied to logistic regression in an attempt to rectify the disparities between men and women. With this implementation, the model performed worse overall but still exhibited above-chance accuracy (78.80%) and precision (84.69%) in this binary classification. All metrics exhibited similar decreases, but none indicate that the model performs poorly.

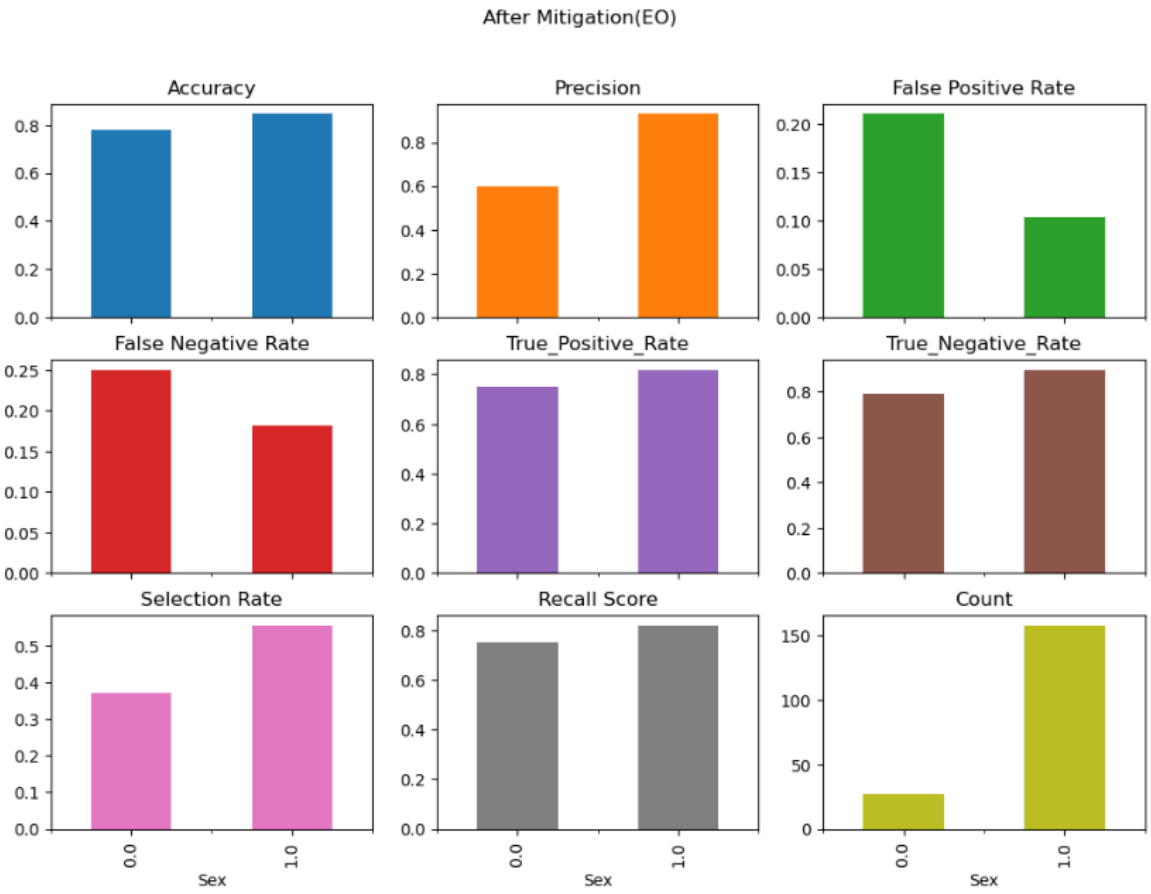
The gap between sexes was not reduced and grew. The accuracy dropped from 77.78% to 62.96% in women compared to the drop from 85.99% to 81.55% in men. Similar decreases were observed in precision and true negative rate, and there was an increase in false positive rate in women.

There were some metrics where women outperformed men, which were recall score, selection rate, true positive rate, and false negative rate. Figure 14 goes further in depth into all the metrics obtained in this analysis.

Equalized Odds

Figure 11

Equalized Odds Logistic Regression Model After Mitigation



Note. Bar graphs comparing the metrics between sexes after implementing demographic parity in the logistic regression model. Women are represented by 0 and men are represented by 1.

Table 13

Equalized Odds Logistic Regression Model After Mitigation

Metrics Type	*Overall*	*Male*	*Female*
Accuracy	83.69 %	84.71 %	77.78 %
Precision	89.69 %	93.10 %	60.00 %
False Positive Rate	12.99 %	10.34 %	21.05 %
False Negative Rate	18.69 %	18.18 %	25.50 %
True Positive Rate	81.31 %	81.82 %	75.00 %
True Negative Rate	87.01 %	89.66 %	78.95 %
Selection Rate	52.72 %	55.41 %	37.04 %
Recall Score	81.31 %	81.82 %	75.00 %
Count	184	157	27

Note. The table below displays information about the metrics in men, women, and both sexes together

In another attempt to fix the disparities between men and women, equalized odds bias mitigation was applied to the model. This bias mitigation strategy performed similar to the model before mitigation. Though equalized odds made the model perform slightly worse, there was minimal change across the metrics with this implementation.

Comparing sexes, women continued to perform worse across all metrics. However, the differences between the sexes were reduced with accuracy, precision, false negative rate, true positive rate, selection rate and recall score. This reduction indicates that applying this mitigation strategy creates a more equitable model between the sexes.

Table 14

Fairness metrics comparison

		Support Vector Machine
	Demographic Parity Difference	*Equalized Odds Difference*
Before Mitigation	0.28	0.22
After Mitigation (Demographic Parity)	0.09	0.39
After Mitigation (Equalized Odds)	0.18	0.14
		Random Forest
	Demographic Parity Difference	*Equalized Odds Difference*
Before Mitigation	0.31	0.26
After Mitigation (Demographic Parity)	0.03	0.30
After Mitigation (Equalized Odds)	0.23	0.12
		Logistic Regression
	Demographic Parity Difference	*Equalized Odds Difference*
Before Mitigation	0.27	0.21
After Mitigation (Demographic Parity)	0.07	0.37
After Mitigation (Equalized Odds)	0.18	0.11

*Note***.** This table records the demographic parity difference and equalized odds difference of all three models before mitigation, after mitigation with demographic parity, and after mitigation with equalized odds.

Discussion

Support Vector Machine

Our initial findings showed a model that was proficient in terms of accuracy and precision, yet it demonstrated a bias towards Males, highlighting the necessity for fairness. Post-intervention, the model achieved a more equitable balance. While demographic parity notably improved the true positive rate and recall score for Females, it also increased the false positive rate for this group. The equalized odds approach, on the other hand, yielded a more balanced enhancement in fairness metrics, including recall score, with a marginal compromise on overall performance.

For the equalized odds method, the Demographic Parity Difference was 0.18, and the Equalized Odds Difference was 0.14. This contrasted with the results from the demographic parity approach, where the Demographic Parity Difference was 0.09, and the Equalized Odds Difference was 0.39. These stats indicate the trade-offs between the two approaches in balancing fairness metrics.

The demographic parity intervention notably improved recall scores for Females, enhancing the identification of positive cases. However, this led to an increase in false positives. The equalized odds approach provided a harmonized balance, improving model fairness across sexes with marginal performance compromise.

It's essential to highlight that an increase in the recall score signifies the model's improved ability to correctly identify true positive cases. This is particularly crucial in scenarios where missing a positive case (false negative) could have significant consequences, such as in medical diagnoses, fraud detection, or credit risk assessment. In the context of bias mitigation, achieving a balanced increase in recall across different groups, such as gender, is pivotal. It ensures that the model is equally effective for all groups in identifying true positives. A disproportionate increase in recall for one group over another could still indicate bias. Therefore, the ultimate goal is to attain high recall while maintaining fairness across all groups.

Random Forest

Before mitigation, the heart failure prediction model, utilizing a Random Forest algorithm, displayed gender-related disparities, particularly with the sensitive attribute of sex. The evaluation revealed a Demographic Parity Difference of 0.31 and an Equalized Odds Difference of 0.26, indicating imbalances in predictions across different gender groups.

After applying the exponentiated gradient using the demographic parity approach with the Random Forest model, there was significant progress in terms of fairness mitigation. The Demographic Parity Difference decreased to 0.03, reflecting a substantial improvement in equalizing predictions across male and female groups. However, this improvement came with a trade-off, as the Equalized Odds Difference increased slightly to 0.3, suggesting a potential compromise in achieving equalized outcomes for positive and negative instances within gender groups. The application of demographic parity notably improved the true positive rate and recall score for females, but also led to an increase in the false positive rate for this group.

The equalized odds approach with the Random Forest model led to further improvements. The Demographic Parity Difference reduced to 0.23, indicating ongoing success in minimizing disparities based on gender. Simultaneously, the Equalized Odds Difference decreased to 0.12, signaling an effective reduction in differences between male and female groups concerning both false positives and false negatives. This approach effectively enhanced the fairness of the model, particularly in reducing disparities in true positive and false positive rates between genders.

In addition to these points, it's important to discuss the selection rate changes. Before the bias mitigation, there was a significant disparity in selection rates between males and females. The demographic parity intervention managed to equalize these rates effectively, bringing them closer together across genders. This is crucial because a disproportionate selection rate can lead to unfair advantages or disadvantages for certain groups. The equalized odds approach also made strides in balancing the selection rates, though not as dramatically as the demographic parity method. The ability to balance selection rates while maintaining acceptable levels of accuracy and precision is a key challenge in creating fair and effective models.

The demographic parity approach effectively reduced disparity in selection rates and recall scores, making the model more equitable across genders. However, this came at the cost of decreased accuracy and precision, particularly for females. In the equalized odds approach, it balanced performance metrics better and reduced disparity in error rates. However, it did not equalize outcomes as effectively as the demographic parity approach.

The decision between these approaches depends on what aspect of fairness is prioritized. If the goal is to equalize the rate of positive outcomes (e.g., selection rates, recall scores) across genders, then the demographic parity approach is more effective. If the focus is on minimizing error rates and maintaining overall performance, then the equalized odds approach is preferable.

Logistic Regression

The initial logistic regression model without any mitigation performed well, yet there was significant bias in favor of men. Both the metrics results and the demographic parity difference of 0.27 and equalized odds difference of 0.21 reflect this observation. In attempts to improve the performance between sexes, an exponentiated gradient using demographic parity and equalized odds were performed respectively.

Of the two bias mitigation methods, equalized odds performed better than demographic parity. Initially, the demographic parity difference of 0.07 appears to be a significant improvement from the naive implementation, the equalized odds difference increased to 0.37. Meanwhile, with equalized odds, both the demographic parity difference and the equalized odds difference decreased to 0.18 and 0.11 respectively.

The metrics also demonstrated that equalized odds were superior to demographic parity, as the gap between sexes decreased with a minimal tradeoff of slightly poorer overall accuracy. In order to compare to the naive implementation of logistic regression, one would need to assess

the tradeoff between accuracy and bias reduction. While equalized odds did create a fairer model, the performance is still worse than the naive implementation.

The equalized odds method proved effective in enhancing fairness metrics, particularly in reducing the disparity in error rates between genders, while maintaining a reasonable balance in performance metrics. This approach appears to be more successful in addressing gender disparities in the logistic regression model compared to the demographic parity method, which improved selection rates but at the expense of overall performance and a significant increase in false positives for females.

Addressing Limitations in Bias Mitigation Techniques

Our project successfully utilized the in-processing ExponentiatedGradient algorithms from Fairlearn's reductions module to address bias in our machine learning model. This algorithm, designed to optimize a fairness criterion while maintaining accuracy, showed effectiveness in enhancing fairness. However, its impact varied across different model aspects, indicating room for further refinement.

Future Directions for Enhanced Impact

In our future work, we plan to incorporate Fairlearn's CorrelationRemover into our pre-processing bias mitigation strategy. This technique is designed to minimize the correlation between sensitive features and other attributes in the dataset before model training. By integrating CorrelationRemover with our ExponentiatedGradient framework, we aim to achieve a more nuanced and effective approach to fairness. This combination promises to enhance our model's ability to handle fairness challenges more effectively, leading to more consistent and impactful outcomes in various modeling contexts.

Alternatively, we plan to enhance the fairness of our machine-learning models by incorporating advanced preprocessing algorithms from AIF360. These include the Disparate Impact Remover, Learning Fair Representations (LFR), Optimized Preprocessing, and Reweighing. These techniques are not just diverse in their approach but are crucial in tackling bias at the most foundational level - the preprocessing stage. Addressing bias early is imperative; it ensures that the data feeding into our models is as unbiased as possible, setting a robust foundation for fair and equitable outcomes. By proactively mitigating bias at this initial stage, we align with best practices in fairness and significantly elevate the overall integrity and reliability of our models across various applications.

Conclusion

This project emphasizes the trade-offs between fairness and accuracy in machine learning models and highlights the essential need for ongoing evaluation and fine-tuning to achieve both optimal predictive performance and fairness across sensitive groups. Specifically, the application of fairness metrics like Equalized Odds and Demographic Parity has led to more equitable recall scores across genders. Yet, this advancement in fairness comes with significant trade-offs. Notably, there's a marked decrease in accuracy and precision, particularly for female individuals in the dataset.

This situation exemplifies a broader issue in fairness interventions in machine learning. When efforts are concentrated on enhancing fairness through methods like ensuring demographic parity (equal selection rates among different groups) or achieving equalized odds (equal false and true positive rates across groups), there's often an unintended consequence on other performance indicators. In this case, the focus on demographic parity and equalized odds resulted in lowered accuracy and precision for female subjects.

Moreover, the selected fairness metric significantly influences which aspects of model performance are emphasized. For example, demographic parity focuses on ensuring equal outcomes, yet this comes with the trade-off of higher error rates. On the other hand, equalized odds aim at balancing error rates, which in turn leads to marginally improved accuracy. This underscores the need to experiment with both metrics to understand their distinct effects and find a balanced approach that ensures both fairness and high-quality predictive outcomes.

This dilemma underscores the complex interplay between various performance metrics in machine learning and the need for a balanced approach to ensure both fairness and high-quality predictive outcomes.

References

CDC. (2023, March 21). *Know Your Risk for Heart Disease*. Centers for Disease Control and Prevention.
https://www.cdc.gov/heartdisease/risk_factors.htm#:~:text=About%20half%20of%20all%20Americans,%2C%20high%20cholesterol%2C%20and%20smoking.&text=Some%20risk%20factors%20for%20heart,the%20factors%20you%20can%20control.

CDC. (2023, May 15). *About Heart Disease*. Centers for Disease Control and Prevention.
https://www.cdc.gov/heartdisease/about.htm

Fedesoriano. (2021). Heart Failure Prediction Dataset [Data set].
https://www.kaggle.com/fedesoriano/heart-failure-prediction.

Gao, Z. Chen, Z. Sun, A. & Deng, X. (2019). Gender differences in cardiovascular disease. *Medicine in Novel Technology and Devices,* 4, 2-4.
https://doi.org/10.1016/j.medntd.2019.100025.

Haeri. A. M. & Zweig. A. K. (2020). The Crucial Role of Sensitive Attributes in Fair Classification. *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*. 2993-3002. doi: 10.1109/SSCI47803.2020.9308585.

Uddin, S., Khan, A., Houssain, E. M., & Moni, A. M. (2019). Comparing different supervised machine learning algorithms for disease prediction. *BMC Med Inform Decis Mak,* 19, 281.
https://doi.org/10.1186/s12911-019-1004-8