D-BIAS Analysis Report

heart.csv

Generated on 11/16/2025

Executive Summary

Fairness Score	Bias Risk
42/100	**High**

Fairness Label	Reliability
Poor	**Moderate**

Dataset Information

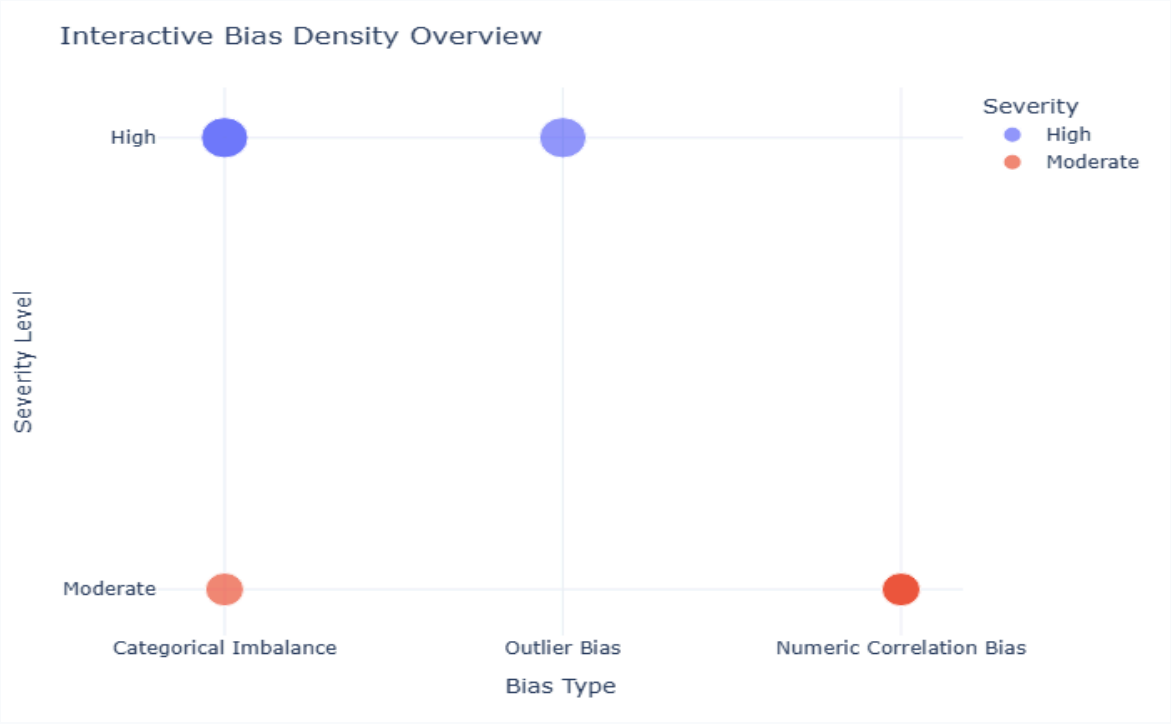
Rows: 919

Columns: 16

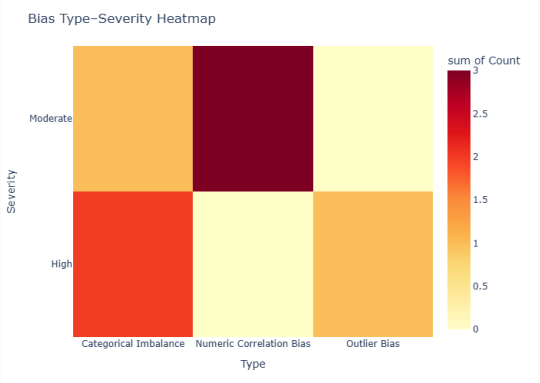
Mean: 123.309

Median: 92

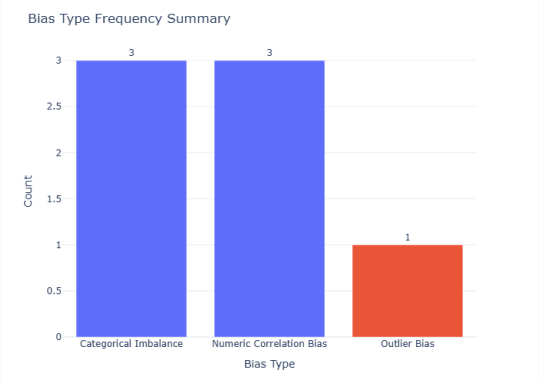
Visualizations



Bias Distribution Overview



Feature Correlation & Fairness Indicators



Severity Heatmap Across Attributes

Detected Biases

Categorical Imbalance

High

Column: sex

Description: 'Male' dominates 78.9% of 'sex' values (entropy=0.74).

AI Explanation

Feature(s): sex

Bias Type: Categorical Imbalance

Severity: High

Meaning: The dataset is heavily skewed towards one gender. Out of 919 patient records, **78.9%** (approximately 725 individuals) are identified as 'Male', while only 21.1% are 'Female'. This represents a nearly 4-to-1 ratio of males to females.

Harm: A machine learning model trained on this data will learn predominantly from male patient profiles. It may fail to capture the unique symptoms and risk factors for heart disease in females, leading to a model that is significantly less accurate for the underrepresented group.

Impact: In a real-world clinical setting, an AI tool built on this data could be less reliable at diagnosing or predicting heart disease in women. This could lead to missed diagnoses, delayed treatment, and poorer health outcomes for female patients.

Severity Explanation: High severity indicates that this imbalance is critical. It is very likely to cause the model's performance to be unequal across genders, directly impacting fairness and reliability. This issue requires immediate attention before model deployment.

Fix:

Data Augmentation: Collect more data on female patients to balance the dataset.
Resampling Techniques: Use oversampling methods like SMOTE to create synthetic data points for the female category or undersample the male category.
Algorithmic Approach: Implement class weights during model training to give more importance to predictions for female patients, forcing the model to pay closer attention to them.

Definition: Categorical Imbalance

Categorical Imbalance

High

Column: fbs

Description: 'False' dominates 85.0% of 'fbs' values (entropy=0.61).

AI Explanation

Feature(s): fbs (fasting blood sugar > 120 mg/dl)

Bias Type: Categorical Imbalance

Severity: High

Meaning: The vast majority of individuals in this dataset (**85.0%**, or about 781 people) have a normal fasting blood sugar level (labeled as 'False'). Consequently, only 15.0% of the data represents individuals with high fasting blood sugar ('True'), a key indicator for diabetes and a risk factor for heart disease.

Harm: With so few examples of patients with high fasting blood sugar, the model may struggle to learn its true impact on heart disease. It might incorrectly conclude that this factor is not significant or misinterpret its relationship with other features.

Impact: A diagnostic model might underestimate the risk of heart disease for patients with high fasting blood sugar. This could lead to a false sense of security for diabetic or pre-diabetic individuals, causing them to miss opportunities for preventive care.

Severity Explanation: High severity means there is a critical lack of data for one category ('True'). This can significantly skew the model's understanding of this important health metric and compromise its predictive power for that patient subgroup.

Fix:

Collect More Data: If possible, gather more records of patients with fasting blood sugar levels above 120 mg/dl.
Algorithmic Mitigation: Use oversampling techniques to increase the representation of the 'True' class or apply class weights to penalize the model more for misclassifying this rare but important group.

Definition: Categorical Imbalance

Categorical Imbalance

Moderate

Column: exang

Description: 'False' dominates 58.0% of 'exang' values (entropy=0.98).

AI Explanation

Feature(s): exang (exercise-induced angina)

Bias Type: Categorical Imbalance

Severity: Moderate

Meaning: The dataset shows a moderate imbalance regarding exercise-induced angina (chest pain during exercise). **58.0%** of individuals did not experience this symptom ('False'), while 42.0% did ('True'). While not as extreme as other imbalances, there is still a noticeable majority of 'False' cases.

Harm: A model might develop a slight preference for the majority class ('False'). This could make it slightly less sensitive in identifying the patterns associated with patients who do experience exercise-induced angina, a critical symptom of heart disease.

Impact: The model's predictions for patients who experience angina might be slightly less confident or accurate. While the impact is less severe than the 'sex' or 'fbs' imbalance, it could still contribute to minor inaccuracies in risk assessment.

Severity Explanation: Moderate severity implies the imbalance is noticeable and could influence the model's behavior. It is not a critical flaw but should be monitored and potentially addressed to ensure robust performance.

Fix:

Monitor Performance: Pay close attention to model performance metrics (e.g., precision, recall, F1-score) for the 'True' class to ensure it is being classified correctly.
Consider Mild Weighting: If performance is lacking for the minority group, apply a gentle class weight to balance the model's focus.

Definition: Categorical Imbalance

Numeric Correlation Bias

Moderate

Column: age ↔ ca

Description: Strong correlation r=0.417.

AI Explanation

Feature(s): age ↔ ca (number of major vessels colored by flourosopy)

Bias Type: Numeric Correlation Bias

Severity: Moderate

Meaning: There is a moderate positive correlation (**r=0.417**) between a patient's 'age' and 'ca', the number of major blood vessels blocked. This suggests that, in this dataset, older individuals tend to have more blocked vessels.

Harm: This is not a fairness bias but a statistical one. Having two correlated features can introduce redundancy. It might make the model harder to interpret, as the importance of 'age' and 'ca' could be conflated. It can also slightly increase the risk of overfitting, where the model learns noise from the data.

Impact: A model might attribute a patient's risk incorrectly. For instance, it might over-emphasize age as a factor when the more direct cause is the number of blocked vessels, which happens to be correlated with age. This can affect the "explainability" of the model's predictions.

Severity Explanation: Moderate severity indicates the relationship is strong enough to be aware of during model development and feature selection. It doesn't require removing a feature but warrants consideration.

Fix:

Awareness: Keep this correlation in mind when interpreting feature importance scores from the model.

Regularization: Use modeling techniques like Lasso (L1) regularization, which can help by automatically reducing the influence of one of the correlated features.

Definition: Numeric Correlation Bias

Numeric Correlation Bias

Moderate

Column: oldpeak ↔ num

Description: Strong correlation r=0.446.

AI Explanation

Feature(s): oldpeak ↔ num (diagnosis of heart disease)

Bias Type: Numeric Correlation Bias

Severity: Moderate

Meaning: The analysis found a moderate positive correlation (**r=0.446**) between 'oldpeak' (ST depression induced by exercise) and 'num' (the presence of heart disease). This means that higher 'oldpeak' values are associated with a greater likelihood of heart disease in this dataset.

Harm: This correlation is not harmful; in fact, it is clinically expected and indicates that 'oldpeak' is a strong predictive feature. The "bias" here is that the model will be heavily influenced by this feature, which is desirable for accuracy as long as it doesn't cause the model to ignore other useful information.

Impact: The model will likely learn to use 'oldpeak' as a key indicator for predicting heart disease, leading to more accurate predictions. This aligns with medical knowledge and is a positive characteristic of the data.

Severity Explanation: Moderate severity highlights a significant and useful predictive relationship. It confirms the feature's value for the modeling task.

Fix:

No Fix Needed: This is a valuable signal.
Recommendation: During model evaluation, use feature importance plots to confirm that the model is leveraging this feature appropriately without completely ignoring other relevant predictors.

Definition: Numeric Correlation Bias

Numeric Correlation Bias

Moderate

Column: ca ↔ num

Description: Strong correlation r=0.574.

AI Explanation

Feature(s): ca ↔ num (diagnosis of heart disease)

Bias Type: Numeric Correlation Bias

Severity: Moderate

Meaning: A strong positive correlation (**r=0.574**) exists between 'ca' (the number of major blocked vessels) and 'num' (the presence of heart disease). This is a very strong signal that as the number of blocked vessels increases, so does the probability of a heart disease diagnosis.

Harm: Similar to the previous bias, this is a powerful and medically relevant signal, not a harmful issue. A model trained on this data will be strongly "biased" to associate a higher 'ca' value with higher risk, which is correct.

Impact: This strong relationship will likely make the predictive model more accurate and reliable, as it can lean on a clear, quantifiable indicator of disease.

Severity Explanation: Moderate severity (despite the strong 'r' value) is used here to flag a very influential feature. It's a key driver in the dataset that will heavily influence any predictive model.

Fix:

No Fix Needed: This is a key predictive feature.
Recommendation: Ensure the model is validated using robust methods like cross-validation to prevent it from simply "memorizing" this strong relationship and failing to generalize to new, unseen data.

Definition: Numeric Correlation Bias

Outlier Bias

High

Column: chol

Description: 20.0% of 'chol' values are outliers (left-skewed).

AI Explanation

Feature(s): chol (serum cholesterol)

Bias Type: Outlier Bias

Severity: High

Meaning: A very large portion of the cholesterol data—**20.0% of all values**—are identified as outliers. The distribution is left-skewed, meaning these outliers are unusually low values. Many of these may be zero-values, which are physiologically impossible for cholesterol and likely represent missing or incorrectly entered data.

Harm: These outliers can severely corrupt the model's learning process. The model might learn a completely wrong relationship between cholesterol and heart disease, or it may treat these invalid data points as genuine, leading to nonsensical patterns. Statistical measures like the average cholesterol level are rendered meaningless.

Impact: A predictive tool built on this data would be fundamentally unreliable. It might wrongly flag low cholesterol as a common or even risky trait, leading to dangerously

inaccurate patient risk assessments. The model's overall performance will be degraded.

Severity Explanation: High severity indicates this is a critical data quality issue. With one-fifth of the data for a key feature being suspect, any model trained without addressing this will be untrustworthy.

Fix:

Investigate and Clean: Examine these outliers. If they are 0s or other placeholder values, treat them as missing data.
Imputation: Replace the cleaned missing values using a sensible strategy, such as imputing with the median cholesterol value (which is less sensitive to outliers than the mean) or using a more advanced imputation model.
Robust Modeling: If some extreme values are determined to be genuine, use models that are naturally robust to outliers, such as Random Forests or Gradient Boosting Machines.

***

Overall Assessment and Recommendations

Definition: Outlier Bias

Recommendations

- **Prioritize Data Cleaning:** Immediately investigate and correct the **20.0% outliers** in the `chol` feature. Replace placeholder values (like 0) with a proper missing data representation and then impute them.
- **Address Gender Imbalance:** Implement a strategy to mitigate the **78.9% male dominance**. The preferred method is to collect more data on female patients. If not possible, use a combination of SMOTE (oversampling) and class weighting during model training.
- **Balance Key Health Metrics:** Apply similar balancing techniques (resampling or class weights) for the `fbs` feature to ensure the model learns from the underrepresented group of patients with high fasting blood sugar.
- **Validate Fairness:** After building a model, do not rely solely on overall accuracy. Evaluate its performance separately for males and females (and for high vs. low `fbs`) to ensure it performs equitably across all subgroups.
- **Document Limitations:** Clearly document all biases found and the steps taken to mitigate them. This transparency is crucial for any clinical tool used in a real-world setting.

Conclusion

The dataset's "fairness health score" is **Poor**. While it contains medically relevant features, the high-severity biases related to representation and data quality pose critical risks to fairness, accuracy, and reliability. Proceeding without addressing these issues would likely result in a biased and untrustworthy model.