D-BIAS Analytical Report

Dataset Bias Detection and Fairness Evaluation

dataset
2025-11-06
Generated by D-BIAS System

Fairness Health Score

42.00

Executive Summary

• Of course. Here is a clear, data-driven explanation of the biases detected in the `heart_disease_cleaned.csv` dataset, designed for both non-technical and technical audiences.

Analysis of Detected Biases in the Heart Disease Dataset

This report breaks down potential biases found in your dataset. Understanding these issues is crucial for building a fair, accurate, and reliable predictive model for heart disease.

1. Categorical Imbalance: `sex`

• * **Description:** The dataset is heavily skewed towards one gender, with **'Male' representing 78.9%** of the patients.

• * **Severity: High** (This is a critical issue that can fundamentally flaw the model).

| Aspect | Explanation |

• | **Meaning** | Your dataset contains nearly four times more data for males than for females. It does not reflect a balanced population and is therefore unrepresentative. |

• | **Harm** | A model trained on this data will become an "expert" on male heart health but a "novice" on female heart health. It may learn patterns specific to men and fail to capture the unique symptoms and risk factors associated with heart disease in women. |

• | **Impact** | **Real-world misdiagnosis.** A female patient's risk of heart disease could be severely underestimated. The model might miss critical warning signs because it was primarily trained to recognize male-pattern symptoms, potentially leading to delayed treatment or a missed diagnosis. |

• | **Fix** | 1. **Collect More Data:** The best solution is to gather more data on female patients to balance the dataset.
2. **Use Resampling Techniques:** Create a more balanced training set by either oversampling the minority class (synthetically creating more female records) or undersampling the majority class (using fewer male records).
3. **Apply Class Weights:** Instruct the model to pay a "heavier penalty" when it makes a mistake on a female patient, forcing it to learn their patterns more carefully. |

2. Categorical Imbalance: `fbs` **(Fasting Blood Sugar)**

• * **Description:** The 'False' value, likely indicating fasting blood sugar < 120 mg/dl, **dominates 85.0%** of the dataset.

• * **Severity: High** (This can cause the model to overlook a critical risk factor).

| Aspect | Explanation |

• | **Meaning** | Only 15% of the patients in your dataset have high fasting blood sugar. This means the model has very few examples of this condition to learn from. |

• | **Harm** | The model may learn to treat high fasting blood sugar as an anomaly rather than a significant risk factor. It might incorrectly conclude that this condition has little to no impact on

heart disease, simply due to a lack of data. |

• | **Impact** | A patient with diabetes or pre-diabetes (indicated by high fasting blood sugar) might have their heart disease risk score calculated incorrectly. The model could fail to flag them as high-risk, leading to a false sense of security and a lack of preventative care. |

• | **Fix** | 1. **Feature Engineering:** If possible, combine `fbs` with other related features to create a more robust "diabetic risk" indicator.
2. **Strategic Data Collection:** Seek out more data specifically from patients with high fasting blood sugar.
3. **Algorithmic Adjustments:** As with the `sex` feature, use class weights or sampling techniques if `fbs` is a critical predictor you want the model to focus on. |

3. Outlier Bias: `chol` **(Cholesterol)**

• * **Description:** A significant **20.0% of 'chol' values are outliers**, and the data is left-skewed.

• * **Severity: High** (This suggests a serious data quality problem that could corrupt the model).

| Aspect | Explanation |

• | **Meaning** | One in five patients has a cholesterol reading that is statistically abnormal compared to the rest. "Left-skewed" suggests many of these outliers are unusually low values, which could include placeholder values like `0` that are biologically impossible. |

• | **Harm** | Outliers can drastically distort statistical calculations and pull the model in the wrong direction. A model might learn a completely incorrect relationship between cholesterol and heart disease because its "view" is warped by these invalid data points. |

• | **Impact** | If the model is trained on data where many `0` cholesterol values are linked to patients with heart disease, it could bizarrely learn that "low cholesterol is a high-risk factor." This would lead to nonsensical and dangerous predictions for real patients with genuinely low or normal cholesterol levels. |

• | **Fix** | 1. **Data Cleaning:** Investigate these outliers immediately. If they are `0`s or other clear errors, treat them as missing values.
2. **Imputation:** Replace the erroneous values with a statistically sound estimate, such as the median cholesterol of the valid data.

3. **Robust Modeling:** Use models that are naturally less sensitive to outliers, such as tree-based models (e.g., Random Forest), but data cleaning is still the preferred first step. |

4. Numeric Correlation Bias

• * **Description:** Several numeric features show moderate to strong correlations (e.g., `ca` ↔ `num` **with r=0.574**; `oldpeak` ↔ `num` **with r=0.446**). `num` is likely the target variable for heart disease diagnosis.

• * **Severity: Moderate** (These are likely valid clinical relationships but require careful handling).

| Aspect | Explanation |

• | **Meaning** | This indicates that certain clinical measurements are strongly linked to the presence of heart disease, which is expected. For example, as `ca` (number of major vessels blocked) increases, so does the likelihood of heart disease (`num`). |

• | **Harm** | While these are strong predictors, they can cause **multicollinearity** in some models (like logistic regression), making it difficult to isolate the independent importance of each feature. The primary risk is **target leakage**: if a feature like `ca` is a diagnostic result that is only known *after* a diagnosis is made, including it in the model is "cheating." The model will perform perfectly on the test data but will be useless in the real world for predicting risk in undiagnosed patients. |

• | **Impact** | If target leakage is present, the model will be a complete failure in a real clinical setting. It would require information that is not yet available to make a prediction. If it's not leakage, the model might become over-reliant on just one or two features, making it less robust if one of those measurements is unavailable or noisy. |

• | **Fix** | 1. **Verify for Target Leakage:** Consult with a domain expert to confirm that features like `ca` and `oldpeak` are measured *before* a final diagnosis (`num`) is made. If not, they must be removed from the feature set.
2. **Feature Selection:** For models sensitive to multicollinearity, consider removing one feature from a highly correlated pair.
3. **Use Appropriate Models:** Tree-based models are generally robust to correlated features. |

Overall Reliability Assessment

The dataset's reliability is **low to moderate**. While it contains clinically relevant features, the high-severity biases detected pose a significant threat to the validity of any model built upon it. The severe gender imbalance and the massive number of outliers in the cholesterol feature are critical data quality and representation issues that must be addressed before this dataset can be trusted for developing a general-purpose clinical tool.

Fairness & Ethical Implications

The most significant ethical concern is the risk of **algorithmic bias against women**. A model trained on this data is likely to be less accurate for female patients, potentially leading to under-diagnosis and poorer health outcomes for a historically underrepresented group in cardiac research. This reinforces existing health disparities and undermines the goal of equitable healthcare. Furthermore, the data quality issues in `chol` could lead to broadly inaccurate risk assessments for all patients.

Concluding Summary: Dataset Fairness Health Score

The dataset's overall "fairness health score" is **Poor**. The critical gender imbalance, combined with significant data quality problems, makes it unsuitable for building a fair and reliable predictive model in its current state. It carries a high risk of producing a model that is both inaccurate and discriminatory.

Actionable Recommendations

To improve the dataset's fairness and reliability, follow these steps:

1. **Investigate and Clean** `chol` **Outliers:** Immediately identify the source of the 20% outliers. If they are data entry errors (e.g., `0`), convert them to null values and impute them using the median of the remaining valid data.

• 2. **Address Gender Imbalance:**

• * **Primary Goal:** Actively seek and collect more data on female patients to create a more representative sample.

• * **Technical Mitigation:** In the interim, use techniques like SMOTE (oversampling) for the female cohort or apply class weights during model training to force the algorithm to prioritize accuracy for the underrepresented group.

• 3. **Validate Features for Target Leakage:** Work with a clinical expert to confirm that strongly correlated predictors like ca (number of vessels colored) are pre-diagnostic indicators. If they are post-diagnostic results, they must be removed from the model's training data.

• 4. **Document All Findings:** Maintain clear documentation of all biases found and the specific steps taken to mitigate them. This transparency is crucial for anyone using the model in the future.

• Top detected bias types: Categorical Imbalance (3), Numeric Correlation Bias (3), Outlier Bias (1)

• Reliability: {'fairness_score': 42, 'message': 'Moderate dataset size — reliability good but not fully stable. High bias risk — dataset requires correction before modeling.', 'n_columns': 16, 'n_rows': 919, 'reliability_level': 'moderate'}

Mapped Bias Analysis

Categorical Imbalance

Feature: **sex** ● High

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

Feature: **fbs** ● High

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

Feature: **exang** ● Moderate

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

Numeric Correlation Bias

Feature: **age** ↔ **ca**

Moderate

Strong correlation r=0.417.

Feature: **oldpeak** ↔ **num**

Moderate

Strong correlation r=0.446.

Feature: **ca** ↔ **num**

Moderate

Strong correlation r=0.574.

Outlier Bias

Feature: **chol**

High

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

Visualizations

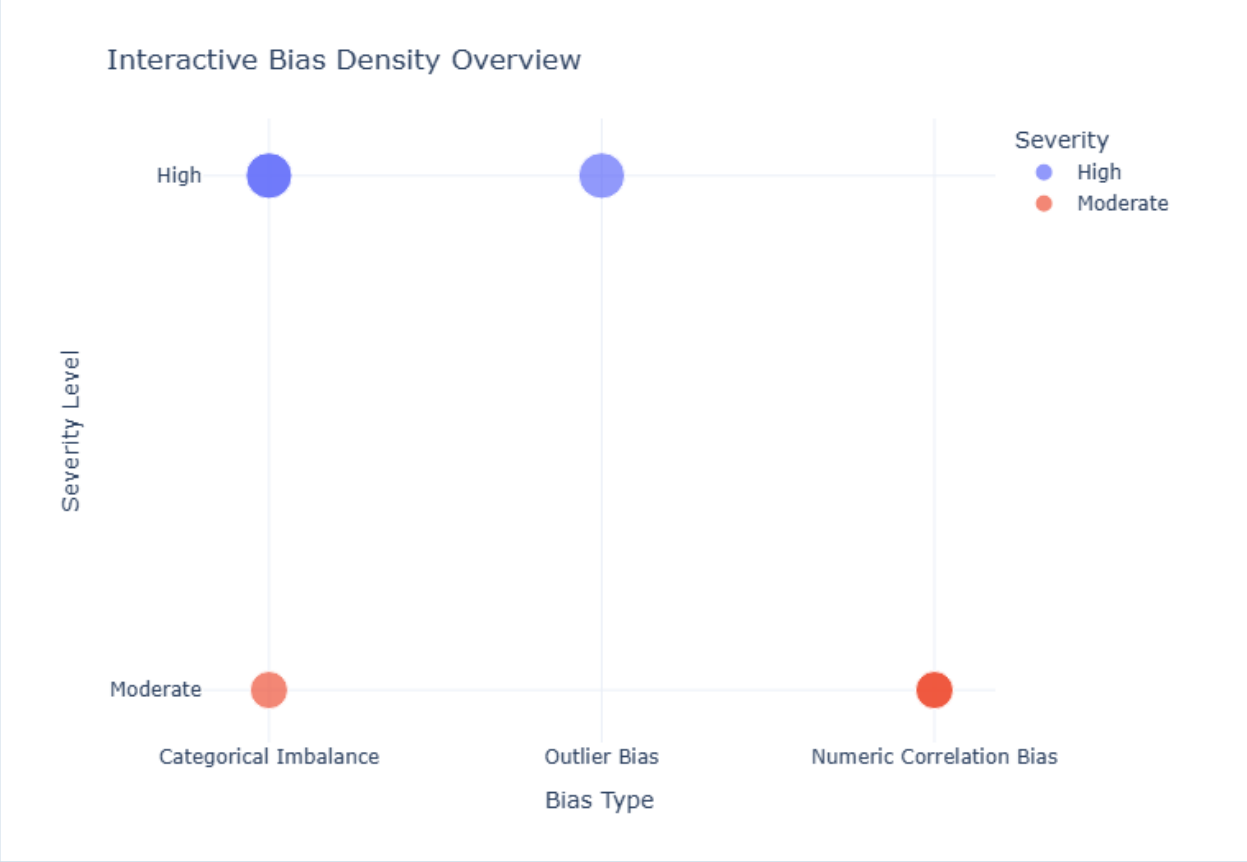

Figure 1 — Bias Distribution by Type

Density and distribution of detected bias types and their severity levels.

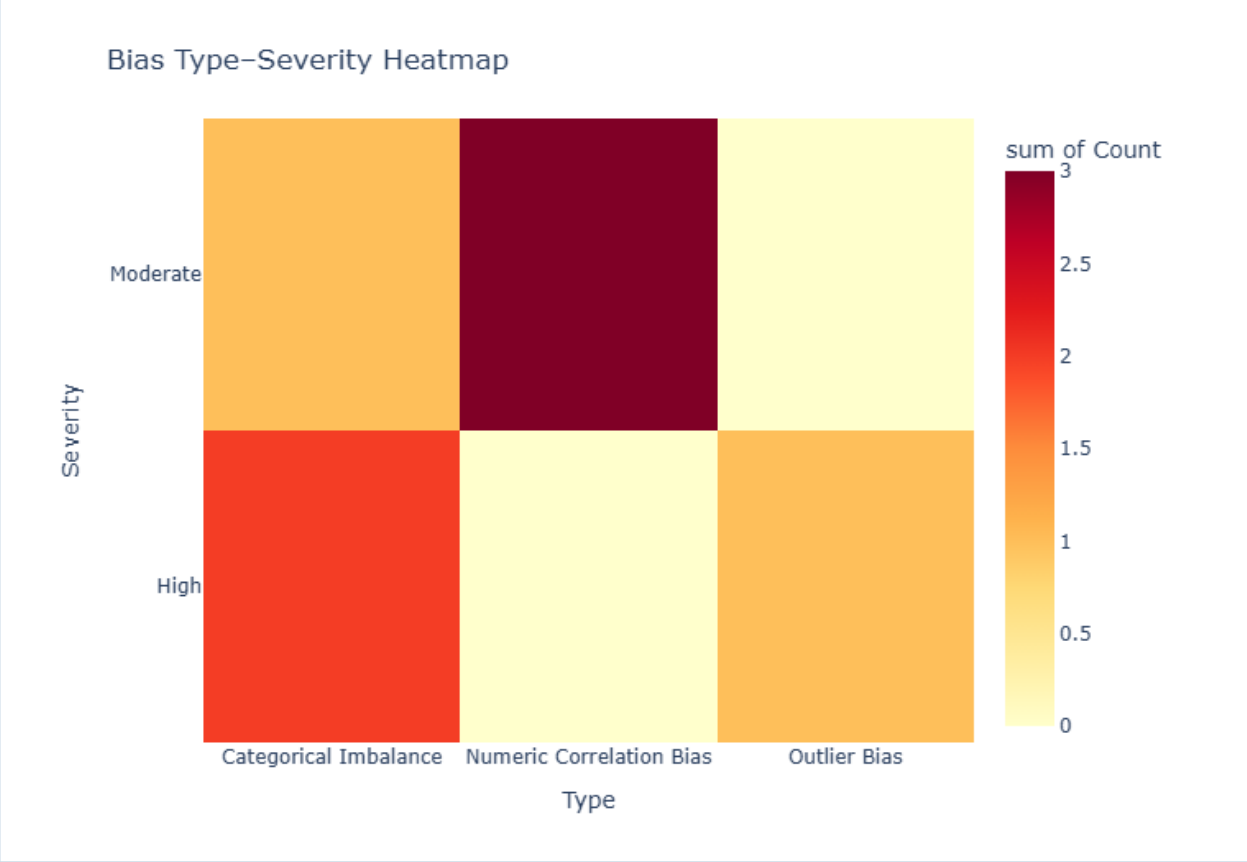

Figure 2 — Fairness Trend Over Time

Temporal trend or aggregated pattern indicating how fairness metrics evolve.

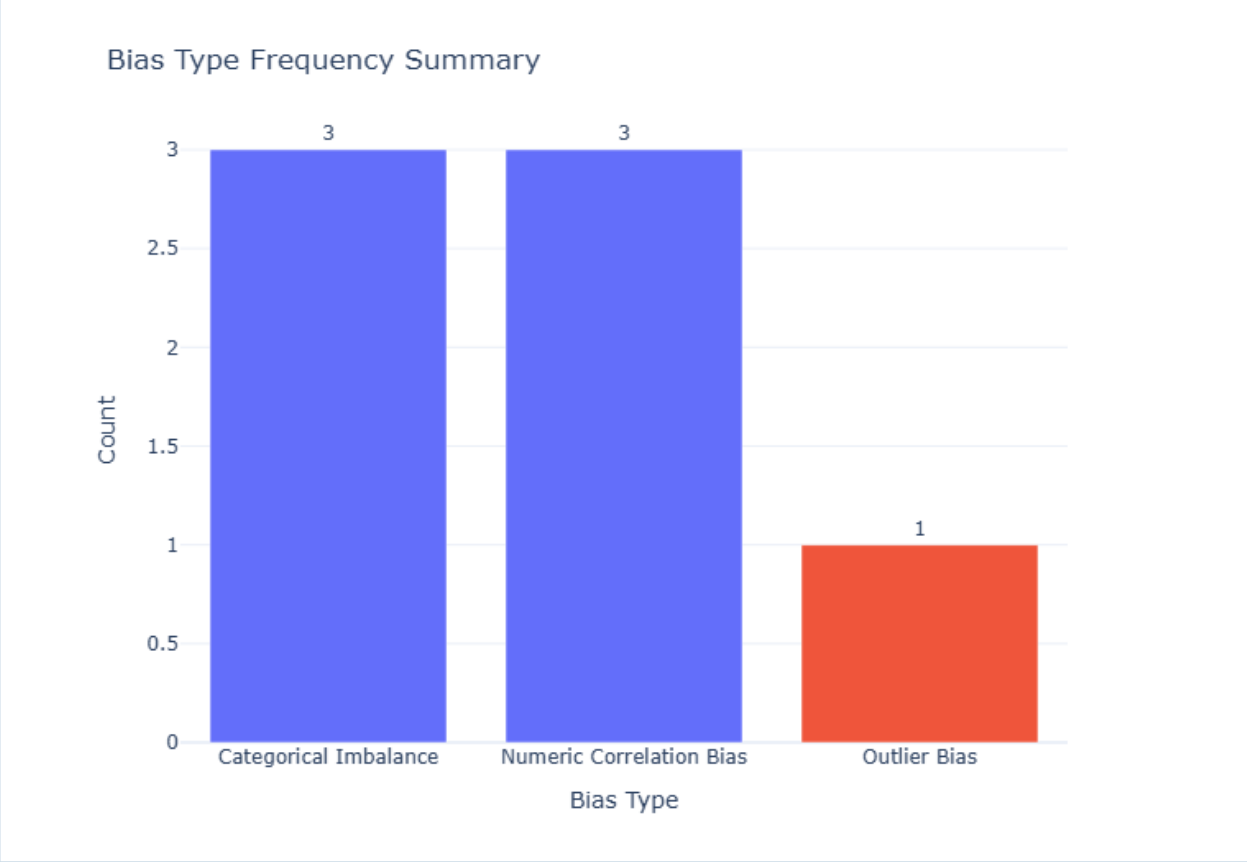

Figure 3 — Bias Severity Breakdown

Counts and proportions of bias severities across detected types.

Recommendations & Actions

Other Recommendations

1. To improve the dataset's fairness and reliability, follow these steps:

2. **Investigate and Clean `chol` Outliers:** Immediately identify the source of the 20% outliers. If they are data entry errors (e.g., `0`), convert them to null values and impute them using the median of the remaining valid data.

3. **Address Gender Imbalance:**

4. **Primary Goal:** Actively seek and collect more data on female patients to create a more representative sample.

5. **Technical Mitigation:** In the interim, use techniques like SMOTE (oversampling) for the female cohort or apply class weights during model training to force the algorithm to prioritize accuracy for the underrepresented group.

6. **Validate Features for Target Leakage:** Work with a clinical expert to confirm that strongly correlated predictors like `ca` (number of vessels colored) are pre-diagnostic indicators. If they are post-diagnostic results, they must be removed from the model's training data.