D-BIAS Analysis Report

heart.csv

Generated on 11/16/2025

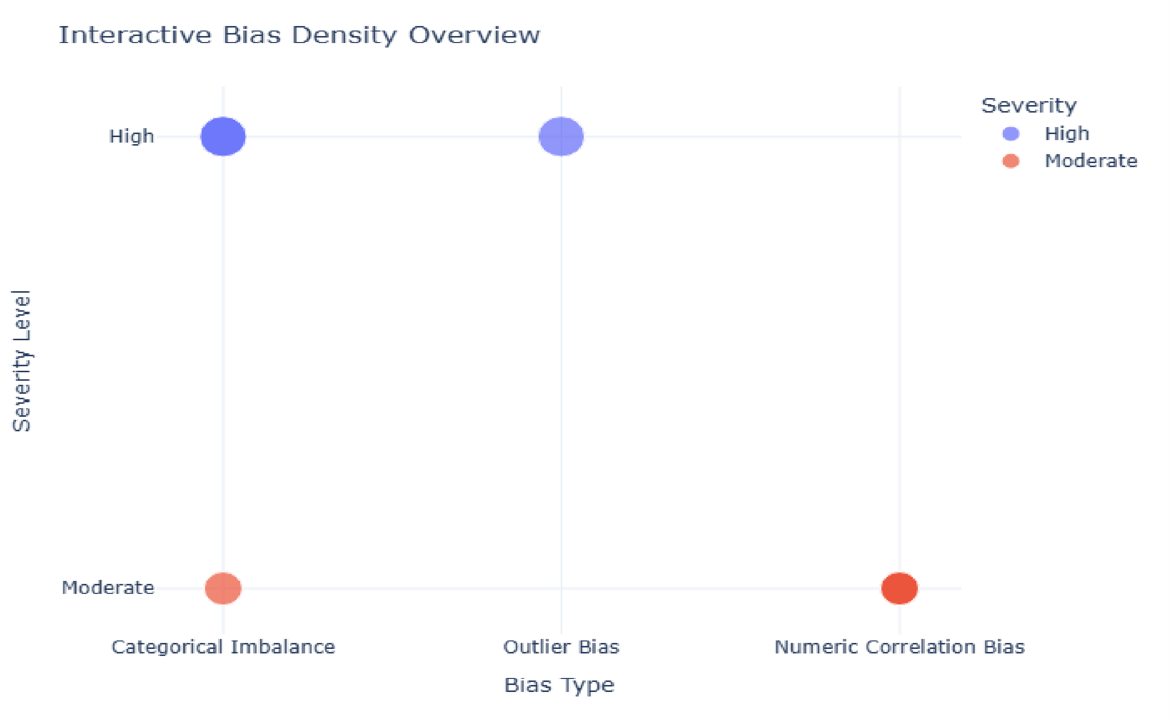
Executive Summary

Fairness Score **42/100**	Bias Risk **High**
Fairness Label **Poor**	Reliability **Moderate**

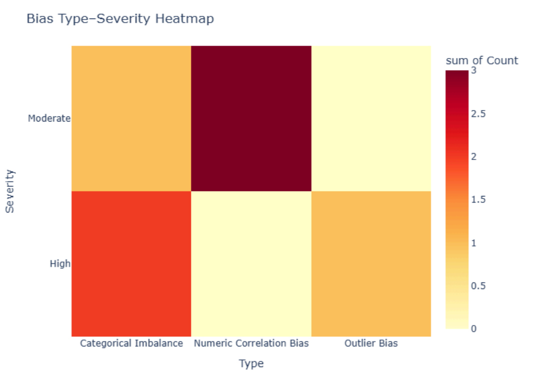
Dataset Information

Rows: 919	Columns: 16
Mean: 123.309	Median: 92

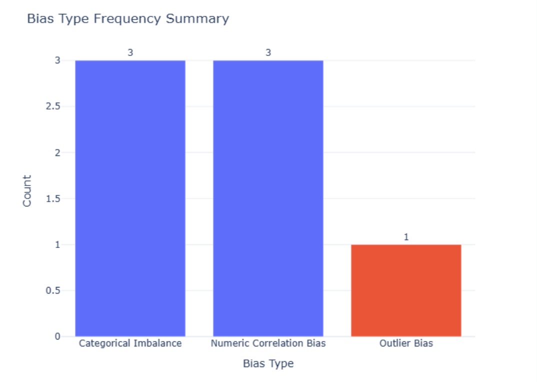
Visualizations



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 by gender. Nearly four out of every five records (**78.9%**) belong to 'Male' patients, leaving female patients significantly underrepresented. The low entropy score of 0.74 numerically confirms this lack of diversity, where a score of 1.0 would represent perfect balance.

Harm:

A model trained on this data will learn primarily from the health patterns of males. It will have insufficient data to understand the nuances of heart disease in females, potentially leading to a model that is systematically less accurate for one gender.

Impact:

In a real-world scenario, this could lead to a higher rate of misdiagnosis or inaccurate risk assessments for female patients. A diagnostic tool built from this data might fail to recognize female-specific symptoms or risk factors, potentially delaying life-saving treatment.

Severity Explanation:

High severity indicates that this imbalance is critical. It will almost certainly create a biased model that performs poorly for the underrepresented group (females) if not addressed directly.

Fix:

1. **Data Collection:** The best solution is to gather more data for female patients to create a more balanced dataset.

2. **Sampling Techniques:** Use methods like oversampling (duplicating female records) or SMOTE (creating synthetic female records) to balance the dataset.

3. **Weighted Models:** Assign a higher weight to the female class during model training, forcing the model to pay more attention to their data.

---

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 dataset is overwhelmingly composed of individuals with normal fasting blood sugar levels. **85.0%** of the entries are marked as 'False', meaning their blood sugar was not high. The very low entropy of 0.61 highlights this extreme imbalance, showing very little variation in this feature.

Harm:

A machine learning model may struggle to learn the importance of high fasting blood sugar as a risk factor because it encounters it so infrequently. The model might treat it as a rare anomaly rather than a significant clinical indicator, diminishing its predictive power.

Impact:

A predictive tool might systematically underestimate the risk of heart disease for patients with high fasting blood sugar (diabetic or pre-diabetic individuals), as the model has not learned to associate this condition strongly with the outcome. This could lead to false reassurances and delayed preventative care.

Severity Explanation:

High severity means that the minority class ('True') is so rare that a model might ignore it entirely to achieve high overall accuracy, making it unreliable for that specific patient group.

Fix:

1. **Oversampling:** Increase the number of 'True' cases in the training data so the model has more examples to learn from.

2. **Cost-Sensitive Learning:** Modify the model's learning algorithm to penalize misclassifications of the minority class more heavily than the majority class.

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:

There is a noticeable imbalance in whether patients experienced chest pain (angina) during exercise. **58.0%** of patients did not experience this symptom ('False'), while 42.0% did. While not as extreme as other imbalances, this still represents a majority-minority split.

Harm:

The model may become slightly better at making predictions for the majority group (patients without exercise-induced angina). Its performance on the minority group might be less reliable due to having fewer examples.

Impact:

This could translate to a model that is slightly less confident or accurate in assessing risk for patients who do present with exercise-induced angina. While the impact may be subtle, it can contribute to a cumulative loss of accuracy.

Severity Explanation:

Moderate severity indicates that the imbalance is noticeable and could influence model performance. It warrants attention but is less critical than a 'High' severity bias.

Fix:

1. **Monitor Performance:** First, evaluate if the model's performance is indeed worse for the 'True' class.

2. **Apply Mild Balancing:** If needed, use gentle balancing techniques or class weights to correct for the skew without drastically altering the data distribution.

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:

This bias indicates a moderate positive relationship (**r=0.417**) between a patient's `age` and `ca`, the number of major blood vessels blocked. In simple terms, the data shows that as patients get older, they tend to have more blocked vessels.

Harm:

When two features are correlated, a model can struggle to distinguish their individual contributions to the outcome (a problem called multicollinearity). This can make the model's internal logic unstable and difficult to interpret. The model might overemphasize the combined effect of age and blocked vessels.

Impact:

A model might find it hard to determine if a high risk score is due to advanced age alone, the number of blocked vessels, or both. This ambiguity can reduce the clinical interpretability of the model's predictions.

Severity Explanation:

Moderate severity suggests the correlation is strong enough to potentially interfere with model interpretation and feature importance calculations but is not so strong as to be redundant.

Fix:

1. **Feature Selection:** Use regularization methods (e.g., Lasso) that can automatically select one feature over the other if they are highly correlated.

2. **Domain Knowledge:** Consult with a medical expert to decide if both features provide unique, essential information. If not, one could potentially be removed.

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:

There is a moderate positive correlation (**r=0.446**) between `oldpeak` (a measure of ST depression on an EKG during exercise) and `num` (the presence of heart disease). This means higher `oldpeak` values are strongly associated with a positive heart disease diagnosis. This is an expected and clinically relevant finding.

Harm:

The "harm" here is not that the correlation exists, but that it might create a dominant feature. The model could become overly reliant on `oldpeak` for its predictions, potentially ignoring other, more subtle risk factors that are important in complex cases.

Impact:

If a patient's `oldpeak` measurement is flawed or unavailable (e.g., they couldn't complete a stress test), a model that heavily depends on it might produce a very unreliable or inaccurate risk score, missing other critical signs.

Severity Explanation:

Moderate severity highlights a strong predictive signal that could lead to model over-reliance. The risk is not in the data itself, but in how a simplistic model might use it.

Fix:

1. **Feature Scaling:** Standardize all numeric features so that `oldpeak` doesn't dominate simply because of its scale.

2. **Regularization:** Use techniques that prevent any single feature from having an excessive influence on the final prediction.

3. **Analyze Feature Importance:** After training, check how much the model relies on `oldpeak` compared to other features.

Definition: Numeric Correlation Bias

Numeric Correlation Bias

Moderate

Column: ca ↔ num

Description: Strong correlation r=0.574.

AI Explanation

Feature(s): ca (Number of major vessels colored) ↔ num (diagnosis of heart disease)

Bias Type: Numeric Correlation Bias

Severity: Moderate

Meaning:

This is the strongest correlation detected (**r=0.574**), indicating a powerful relationship between the number of blocked vessels (`ca`) and the final heart disease diagnosis (`num`). This aligns with clinical knowledge: more blocked vessels are a direct and serious indicator of heart disease.

Harm:

Similar to `oldpeak`, the primary risk is creating an overly simplistic model. A model might learn a rule like "if `ca` > 1, predict disease" and ignore other crucial information like cholesterol, blood pressure, or patient history.

Impact:

The model's predictions could be heavily swayed by the `ca` value alone. A patient with many other risk factors but a low `ca` count (perhaps from an inconclusive scan) might be incorrectly classified as low-risk, delaying necessary intervention.

Severity Explanation:

Moderate severity is assigned despite the high correlation value because this is a clinically valid and powerful predictor. The concern is about ensuring it works in concert with other features, not by itself.

Fix:

1. **Holistic Modeling:** Ensure the model is complex enough (e.g., using ensemble methods like Random Forest) to capture interactions between features, rather than relying on a single predictor.

2. **Interpretability:** Use tools like SHAP or LIME to understand *why* the model makes a certain prediction and ensure `ca` is not the only factor being considered.

Definition: Numeric Correlation Bias

Outlier Bias

High

Column: chol

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

AI Explanation

Feature(s): chol (Cholesterol)

Bias Type: Outlier Bias

Severity: High

Meaning:

A very large portion of the cholesterol data—**20.0% of all entries**—is flagged as outliers. The data is "left-skewed," which means there is a long tail of unusually low values. Often in medical datasets, a value of `0` is used to indicate missing data, which could explain these extreme low values.

Harm:

Outliers can severely distort the training process of many machine learning models. The model will try to accommodate these extreme values, which can skew its understanding of a "normal" cholesterol range and its relationship with heart disease.

Impact:

A model trained on this data might learn that a cholesterol level of `0` is a possible and meaningful value, leading to nonsensical predictions. This could corrupt the model's ability to accurately assess risk for patients with genuinely low, normal, or high cholesterol levels.

Severity Explanation:

High severity is warranted because 1 in 5 data points is an outlier. This is a substantial data quality issue that will significantly degrade model performance and reliability if left unaddressed.

Fix:

1. **Investigate Outliers:** Determine the cause of the outliers. If `0` values represent missing data, they must be treated as such.

2. **Imputation:** Replace the erroneous/missing values using a sound strategy, such as filling them with the median or using a more advanced imputation model.

3. **Robust Scaling:** If the outliers are deemed genuine but extreme, use a scaling method (like `RobustScaler`) that is less sensitive to their influence.

Overall Summary and Recommendations

Definition: Outlier Bias

Recommendations

- **Prioritize Data Quality:** The **20% outliers in `chol`** must be investigated and cleaned first. Determine if they are missing values and impute them correctly.
- **Address Severe Imbalances:** Implement a robust strategy to handle the imbalances in `sex` and `fbs`. A combination of collecting more diverse data and using advanced techniques like SMOTE or weighted classes is recommended.
- **Build a Balanced Model:** When training, use regularization techniques (L1/L2) to prevent over-reliance on the highly correlated features (`ca`, `oldpeak`) and ensure the model learns from a wide range of indicators.
- **Audit for Fairness:** After building a model, rigorously test its performance across different demographic groups (male vs. female). The goal is not just high overall accuracy, but equitable accuracy for all subgroups.

Conclusion

The dataset has a **poor "fairness health score."** It requires significant pre-processing to address critical representation biases and data quality errors. Without careful mitigation, any model trained on this data would likely be both inaccurate and unfair.