Freie Universität Berlin

Human-Centered Data Science

Group 2:

Anuraj Suman
Egor Dubrovskii
Narek Okroyan
Saniya Nankani

Dataset 📊

- **Dataset Source:** UCI Heart Disease Dataset.
- **Key Features:** 14 Attributes including age, sex, fasting blood sugar, chest pain type, etc
- **Patient Records:** Records of 920 patients available

Target Audience 🤓

Fellow data scientists working in the field of medical research

Our Goals 💡

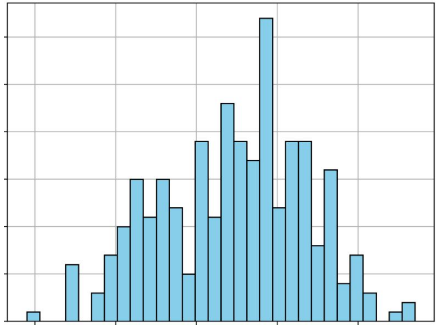
We want our target audience to be able to:

- **Explore the data** on surface level
- **Assess the performance** of various models
- **Gain insight** into models’ decision process
- **Compare** different models and their explanation techniques
- **Evaluate biases** present in data and trained models

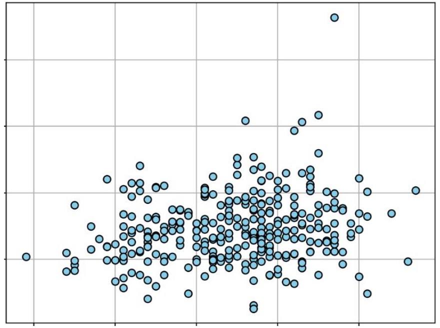
Tab: Exploring Data 🔎

Various visualizations for distributions – for surface level analysis

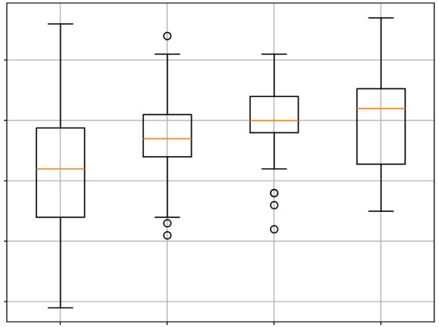
Histograms



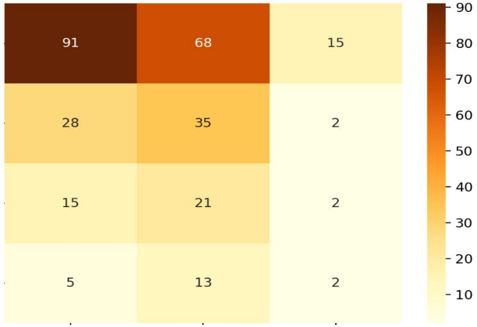
Scatterplots



Boxplots



Heatmaps

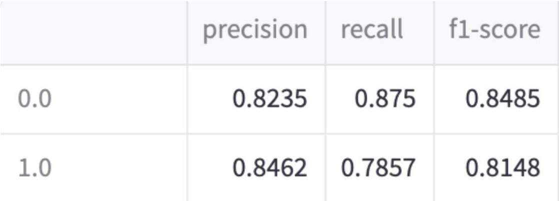


And filters apply to these, too!

Tab: Building Models 🤖

Random Forest

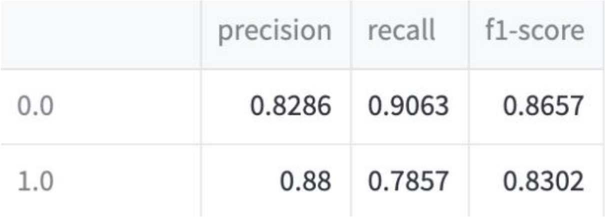
- ➔ Accuracy: 83%
- ➔ AUROC score: 0.89
- ➔ demonstrates a good balance between precision and recall for both classes, indicating reliable performance in predicting heart disease.



	precision	recall	f1-score
0.0	0.8235	0.875	0.8485
1.0	0.8462	0.7857	0.8148

Logistic Regression

- ➔ Accuracy: 85%
- ➔ AUROC score: 0.91
- ➔ Robust in accurately identifying individuals with heart disease.
- ➔ Equally good as RF and maintains a balance between precision and recall.



	precision	recall	f1-score
0.0	0.8286	0.9063	0.8657
1.0	0.88	0.7857	0.8302

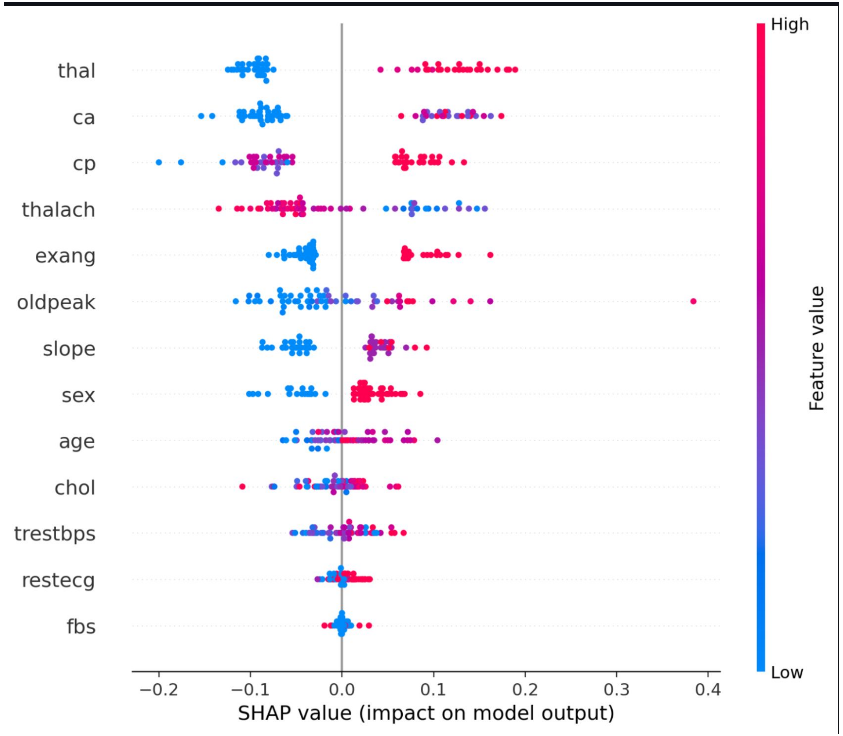
Tab: Explaining Algorithms 🤔

Random Forest: SHAP Values

For the Random Forest model, you can see the importance of each feature based on SHAP values.

The SHAP values represent the contribution of each feature to the prediction.

A higher SHAP value indicates a more significant impact on the model's output. so the thal variable has the highest SHAP values, and it has the highest effect on the prediction.

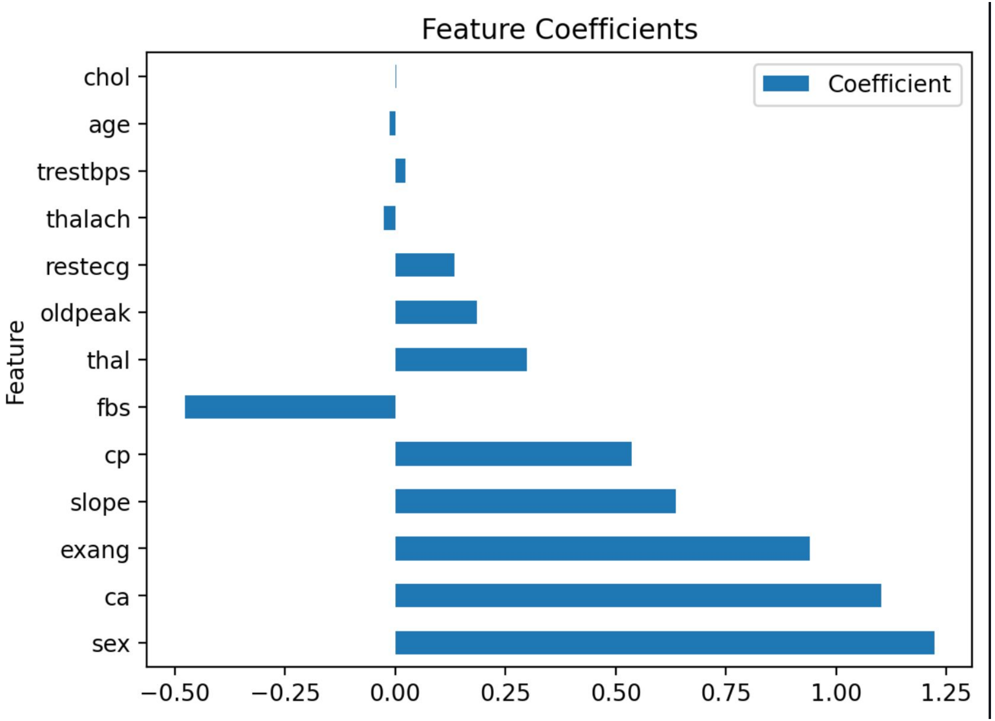


Tab: Explaining Algorithms 🤔

Logistic Regression: Feature Coefficients

For the Logistic Regression model, the tab displays the coefficients of each feature, indicating their weight in the decision-making process.

Positive coefficients suggest that higher values of the feature increase the likelihood of predicting the positive class (heart disease), while negative coefficients indicate the opposite. We can see that sex seems to have the most effect, while the chol variable has the least effect



Tab: Fairness Functions 👩‍⚖️

Q: How **fair** are the results with respect to a specific **group**? (e.g. men or women)

Several fairness metrics checked:

- **Group Fairness**
- **Conditional Statistical Fairness**
- **Predictive Parity**
- **False Positive Error Rate Balance**

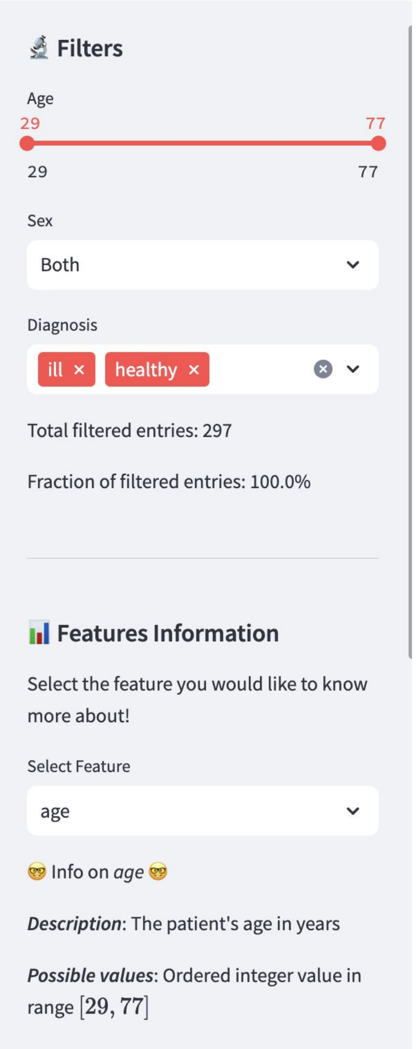
🤖(👨) = 👍

❓

🤖(👩) = 👎

Metrics provided for both random forest and logistic regression

Also, results for bias mitigation techniques included!



🔬 Filters

Age: 29 – 77

Sex: Both

Diagnosis: ill × healthy ×

Total filtered entries: 297

Fraction of filtered entries: 100.0%

📊 Features Information

Select the feature you would like to know more about!

Select Feature: age

🤓 Info on *age* 🤓

Description: The patient's age in years

Possible values: Ordered integer value in range $[29, 77]$

Live Demo 🤩