Heart Disease Prediction Using Logistic Regression

Dataset: UCI Heart Disease (Cleveland subset)
https://archive.ics.uci.edu/dataset/45/heart+disease

Introduction

In this project, we aimed to build a logistic regression model to predict the presence of heart disease in patients, using the Cleveland Heart Disease dataset from the UCI Machine Learning Repository. The goal is to identify whether a patient shows signs of heart disease (1) or not (0), based on various medical indicators.

We focused on clarity, interpretability, and completeness, following a step-by-step data science process that includes data cleaning, exploratory analysis, feature preparation, modeling, and result interpretation.

Step 1 – Dataset Overview and Cleaning

The original data file (`processed.cleveland.data`) includes 14 columns. Some entries use `?` to indicate missing values. These had to be identified and addressed before any modeling could be done.

We performed the following:

- Assigned correct column names based on the dataset documentation.
- Replaced all `?` entries with actual `NaN` values for easier processing.
- Converted affected columns (`ca`, `thal`) to numeric.
- Removed all rows that contained any missing values.
- Converted the original multi-class target `num` to a binary format:

    - `0` = No heart disease
    - `1` = Presence of heart disease (from original values 1, 2, 3, or 4)

Final Dataset Summary:

- Total Rows: 297
- Total Features: 13
- Target Variable: `target` (binary: 0 = No, 1 = Yes)
- All features are numerical and clean.

Step 2 – Exploratory Data Analysis (EDA)

2.1 Target Distribution

A class balance check showed:

- 160 patients with no heart disease
- 137 patients with heart disease

This is relatively balanced, so no resampling techniques were necessary.

2.2 Correlation Matrix

We plotted a heatmap of Pearson correlations between all numeric variables. Notable findings:

- `thal`, `ca`, `cp` (chest pain type), and `slope` showed relatively high positive correlation with the target.
- `thalach` (maximum heart rate achieved) and `oldpeak` (ST depression) had negative correlations with the target.
- `fbs` (fasting blood sugar) and `chol` showed almost no relationship with the outcome.

These insights helped highlight which features might be more predictive in the modeling step.

2.3 Boxplots (Feature Distributions by Target)

We explored the relationship between selected features and heart disease using boxplots:

- **Age:** Slightly lower on average for heart disease patients.
- **Thalach (max heart rate):** Significantly lower in patients with heart disease.
- **Oldpeak:** Higher in those with heart disease.
- **Chol and Trestbps:** Distributions overlap, indicating weaker relationships.

This visual inspection confirmed and contextualized the heatmap results.

Step 3 – Feature Preparation and Scaling

Before modeling, we prepared the features:

- Defined `X` as all columns except `target`.
- Defined `y` as the binary target column.
- Split the dataset into training and test sets (80/20 split) with **stratification** to preserve target balance.
- Standardized features using `StandardScaler` for better model performance and convergence.

Step 4 – Logistic Regression Model

We initialized a logistic regression model using:

- `solver='liblinear'` (recommended for small datasets)

- `penalty='l2'` (standard regularization)
- `C=1.0` (default regularization strength)

The model was trained on the standardized training data.

Step 5 – Model Evaluation

We used the test set to generate predictions and calculated key performance metrics:

Metric	Value
Accuracy	0.8333
Precision	0.8462
Recall	0.7857
F1 Score	0.8148
ROC AUC	0.9498

The results are strong, especially the ROC AUC score, which indicates the model distinguishes between classes well. The confusion matrix also shows a balanced classification with low false positive/negative rates.

Step 6 – Feature Importance

We extracted the model coefficients to assess which features had the strongest influence on prediction. The top contributors were:

- `ca` (number of major vessels): strongest positive predictor
- `thal`: high correlation with disease presence
- `cp` (chest pain type): certain types more indicative of risk
- `sex`: male patients had higher risk
- `oldpeak`, `trestbps`, `exang`: all relevant as expected

Features like `chol` and `age` had much smaller contributions.

Step 7 – Conclusion

This project demonstrated a clear, well-structured application of logistic regression to a classic medical dataset. The results show:

- High classification performance with a simple and interpretable model.
- Real-world relevance of clinical features like `thal`, `ca`, and `chest pain type`.
- Balanced evaluation across metrics, with no overfitting observed.