Heart Disease Prediction Model

Johnson Chuang | Eduardo Sanches | Azadeh Ramesh | Jose Davila

2025-12-04

Table of contents

Summary

Heart disease is one of the leading causes of death globally, and early detection is critical for prevention and treatment. In this project, we use the UCI Heart Disease dataset to build a machine-learning model that predicts whether a patient is likely to have heart disease based on clinical and physiological attributes. We load the dataset directly from the web, clean and wrangle the data, perform exploratory data analysis (EDA), and train a classification model (Decision Tree) to identify important predictors of heart disease. Our results highlight key risk indicators that align with well-known medical knowledge, demonstrating how machine learning can support early screening and clinical decision-making.

Introduction

The objective of this project is to develop a predictive model that determines whether a patient is at risk of heart disease using a set of clinical measurements. Heart disease diagnoses often rely on many interacting factors such as chest pain symptoms, blood pressure, cholesterol levels, and exercise response. Machine-learning models can help uncover patterns in these variables and support early identification of high-risk patients.

Our research question is:

"Given a patient's clinical and physiological attributes, can we accurately predict whether they have heart disease?"

To answer this question, we use the publicly available Heart Disease dataset from the UCI Machine Learning Repository. This dataset contains multiple medically relevant variables, making it suitable for a classification model such as a Decision Tree.

Dataset Description

We use the Heart Disease dataset from the UCI Machine Learning Repository, a widely used benchmark dataset for medical prediction tasks. The dataset includes the following 14 attributes:

- Age
- Sex
- Chest Pain Type (cp)
- Resting Blood Pressure (trestbps)
- Cholesterol (chol)
- Fasting Blood Sugar (fbs)
- Resting ECG results (restecg)

- Maximum heart rate achieved (thalach)
- Exercise induced angina (exang)
- ST depression (oldpeak)
- Slope of ST segment (slope)
- Number of major vessels (ca)
- Thalassemia result (thal)
- num (Target: the predicted attribute (0 = no heart disease, 1 = heart disease))

These variables include both continuous and categorical measurements commonly used in clinical diagnostics.

Methodology

We build a machine-learning classification model using the UCI Heart Disease dataset:

Load data from the original source on the web: https://archive.ics.uci.edu/dataset/45/heart+disease

Wrangle and clean the data

Replace missing values Assign meaningful column names Convert categorical variables to numeric where needed Ensure that the target variable is binary (0 = no heart disease, 1 = heart disease) Perform exploratory data analysis (EDA) Summary statistics for continuous variables Count plots for categorical variables Histograms and boxplots to understand feature distributions Create visualizations relevant to the classification task Pairplots to explore relationships between key features Distribution of target classes Feature correlation matrix Build a classification model A Decision Tree Classifier is trained to predict heart disease. We split the dataset into training and testing subsets and evaluate model accuracy. Visualize the model results Plot of the trained Decision Tree Feature importance bar chart

Importing the Dataset

A special note about our data download process: The following code downloads the zip file from UCI's website, unpacks them, and grabs the data of interest (Cleveland data). It is then processed minimally by adding the correct column names, and finally written out as a CSV to the data/processed folder.

In our actual analysis, we fetch the same data directly using UCI's own ucimlrepo library. The data is the same. But we include this part to show how you can download the data without UCI's own library.

Data Validation

Data validation include ALL the checks you wrote: - Data Type Check - Missing Values Check - Duplicate Check - Category Level Check - Logical Ranges Check - Train/Test Leakage Check

Exploratory Data Analysis (EDA)

Modeling Section

Column Transformations

Create the Pipeline

Crossvalidation

Fit the Model

Predict (X_test) and compatr with Actuals (y_test)

Discussion

The Decision Tree model was able to identify meaningful patterns to predict heart disease based on the data, with a test score of 0.61 and train score of 0.78. Based on these results, it might indicate that there was some overfitting based on the large difference between training and test results.

From the EDA, we see that various features such as age, sex, chol and more have clear differences in their distribution between disease and no disease which will help the model to predict between the two. For a better predictor, we may want to incorporate additional features given the complexity of heart disease.

Results and Conclusion

Our analysis shows that several clinical features differ noticeably between patients with and without heart disease. As seen in the EDA histograms, patients with heart disease tend to have higher resting blood pressure (trestbps), higher ST-depression values (oldpeak), and lower maximum heart rate achieved (thalach) compared to individuals without disease. After preprocessing the dataset using scaling for numerical variables and one-hot encoding for categorical variables, we trained a Support Vector Classifier (SVC) model. Cross-validation results

indicate an average test accuracy of 0.61, with a higher training accuracy of 0.78, suggesting some overfitting. When evaluating predictions on the unseen test set, the model correctly identified many cases but also showed several misclassifications, especially where the model predicted "0" (no disease) but the true label was "1" or "2." Overall, while the model captures meaningful patterns in the dataset, its moderate predictive performance suggests that further tuning, alternative models, or feature engineering may be needed to improve accuracy and reduce classification bias.

References