Chad Huntebrinker’s Final Project Proposal

Research Question and Motivation

Each year there are about 17.9 million lives lost to cardiovascular diseases, which is the leading cause of death globally (World Health Organization, 2023). One such type of cardiovascular disease is heart disease. In the United States, heart disease is the leading cause of death for men and women; in 2022, 702,880 people died from heart disease (Centers for Disease Control and Prevention, 2024). One of the problems with heart disease is that it can be difficult to have early detection due to it being related to multiple different medical and lifestyle factors. This results in many cases that might’ve been easily prevented or cured instead go undetected. Fortunately, with the help of data-modeling and using them to make predictions, we can create models that assist in forecasting and understanding heart disease. For this final project, I hope to answer the following research question:

What are the most significant indicators that predict the presence of heart disease?

Understanding which variables contribute most to the likelihood of heart disease could help in early screening, prevention, and treatment plans. I hypothesize that exercise-induced angina, chest pain type, and thallium stress test results are leading factors of heart disease. The CDC also states that some key factors for heart disease are high cholesterol and high blood pressure, so I will be exploring those variables in the model as well (Centers for Disease Control and Prevention, 2024).

Dataset Description

I will be using the Heart Disease dataset from the [UCI Machine Learning Repository](https://archive.ics.uci.edu/dataset/45/heart+disease). More specifically, I will be using the data collected at the Cleveland clinic as that’s the only one that has been used by machine learning researchers to date. The database was collected in the following way:

The reference group used to derive the model consisted of 303 consecutive patients referred for coronary angiography at the Cleveland Clinic between May 1981 and September 1984. No patient had a history or electrocardiographic evidence of prior myocardial infarction or known valvular or cardiomyopathic disease. All 303 patients provided a history and underwent physical examination, electrocardiogram at rest, serum cholesterol determination and fasting blood sugar determination as part of their routine evaluation. Historical data were recorded and coded without knowledge of noninvasive or angiographic test data. In addition, after giving informed consent, the patients underwent 3 noninvasive tests as part of a research protocol. The results of these tests (exercise electrocardiogram, thallium scintigraphy and cardiac fluoroscopy) were not interpreted until after the coronary angiograms had been read. These tests were analyzed and the results recorded without knowledge of the historical or angiographic results… The mean age of these patients was 54 years; 206 were men. Angiograms were interpreted by a cardiologist without knowledge of other test data (Deterano, Robert, 1998).

The dataset I will be using includes 303 observations with the following 14 variables:

|  |  |
| --- | --- |
| **Variable** | **Description** |
| Age | Age of the patient in years |
| Sex | Male or female |
| Cp | Chest pain type (1 = typical angina, 2 = atypical angina, 3 = non-anginal pain, 4 = asymptomatic) |
| Trestbps | Resting blood pressure in mmHg |
| Chol | Serum cholesterol in mg/dL |
| Fbs | Fasting blood sugar > 120 mg/dl (1 = true, 0 = false) |
| Restecg | Resting electrocardiographic results (0 = normal, 1 = ST-T wave abnormality, 2 = probable left ventricular hypertrophy) |
| Thalach | Maximum heart rate during exercise |
| Exang | Exercise-induced angina (1 = yes, 0 = no) |
| Oldpeak | ST depression induced by exercise related to rest |
| Slope | Slope of the peak exercise ST segment (1 = upsloping, 2 = flat, 3 = downsloping) |
| Ca | Number of major vessels colored by fluoroscopy |
| Thal | Result from a Thallium stress test (3 = normal, 6 = fixed defect, 7 = reversible defect). |
| num | Presence of heart disease (0 = no heart disease, 1–4 = heart disease of increasing severity). |

Dependent and Independent Variables

The dependent variable I will be using is num. Num tracks whether or not heart disease is present. Its values range from 0 – 4 (based on how severe the heart disease is), but I will be converting that to binary values 0 (when heart disease is not present) and 1 (when heart disease is present). This is due to the goal of this model being to predict if there is heart disease present, not how severe the heart disease is.

The independent variables I will be using include the following: cp (chest pain type), trestbps (resting blood pressure), chol (serum cholesterol), thalach (maximum heart rate during exercise), exang (exercise-induced angina) and thal (results from a Thallium stress test). However, I will be exploring all the variables (besides num) within the data as some may have a direct effect on predicting heart disease or may assist the model with prediction when included with other variables.

As for the methods of analysis, I will be fitting multiple logistic regression models to the data and doing assorted tests and validation methods to verify that the best model is being used. Validation methods include p-value and hypothesis testing with coefficients, ROC curve, goodness-of-fit test (like AIC), along with other methods.

Works Cited

Centers for Disease Control and Prevention. “Heart Disease Facts.” *Centers for Disease Control and Prevention*, 11 Jan. 2024, <https://www.cdc.gov/heart-disease/data-research/facts-stats/index.html>.

Detrano, Robert, et al. “International Application of a New Probability Algorithm for the Diagnosis of Coronary Artery Disease.” *The American Journal of Cardiology*, vol. 64, no. 5, 1989, pp. 304–310. [https://doi.org/10.1016/0002-9149(89)90524-9](https://www.ajconline.org/article/0002-9149(89)90524-9/pdf).

Dua, Dheeru, and Casey Graff. “Heart Disease Data Set.” *UCI Machine Learning Repository*, University of California, Irvine, School of Information and Computer Sciences, 2019, <https://archive.ics.uci.edu/ml/datasets/heart+Disease>.

World Health Organization. “Cardiovascular Diseases.” *World Health Organization*, 2023, [https://www.who.int/health-topics/cardiovascular-diseases](https://www.who.int/health-topics/cardiovascular-diseases?utm_source=chatgpt.com#tab=tab_1).