Project Abstract

Roshan Chouhan, Yanzheng Liu

Cardiovascular disease remains the leading cause of death globally, impacting millions of lives and placing significant strain on healthcare systems. Preventative measures, early detection, and accurate risk stratification can save countless lives and resources. This project uses a publicly available dataset compiled from the 2022 Behavioral Risk Factor Surveillance System which is a large-scale health survey administered annually by the CDC. The survey gathers detailed health-related information from over 400,000 adults across the United States, making it the largest continuous health survey system in the world. The dataset captures a wide array of risk indicators, including body mass index, smoking status, alcohol consumption, physical activity, general health perception, mental and physical health days, sleep time, stroke history, and diabetes status—along with demographic features like age, sex, and race. These features provide context to research the determinants of heart disease. Importantly, the dataset includes a binary label indicating whether the respondent has had a diagnosis of heart disease, making it ideal for supervised classification tasks.

Goals for Exploratory Data Analysis:

Hypothesis 1: Individuals with high blood pressure, high cholesterol, and diabetes are significantly more likely to report having had a heart attack.
Method: Cross Validation; Logistic regression Classification; Linear Model Selection
If the hypothesis is confirmed: Screening and monitoring individuals with any combination of these conditions becomes more critical. The hospital and doctors can give better suggestions about targeted prevention.

Hypothesis 2: A small set of features, including key behavioral and self-reported variables (e.g. sleep time, BMI, smoking status, alcohol), can accurately flag individuals at risk of heart disease—supporting simple, scalable screening tools.
Method: Linear Model Selection, Moving Beyond Linearity, Resampling Methods
If the hypothesis is confirmed: This insight can be used to develop low-cost risk calculators or short surveys for non-clinical settings enabling early risk identification without lab tests or medical visits.

Hypothesis 3: Poor mental health days, inactivity, alcohol use may contribute significantly to heart disease risk.
Method: Linear Model Selection; Moving Beyond Linearity and Cross validation
If the hypothesis is confirmed: It can prevent heart disease focusing on mental health. The medical center will know how important that mental health be to a potential heart disease patient.

We aim to derive practical lessons that could inform low-cost, scalable early screening tools for heart disease prevention and triage.