**Analyzing Heart Disease Key Factors**

**By:** Arash Ahmadi

**Data:**

<https://www.kaggle.com/ronitf/heart-disease-uci>

**Source:**

This data comes from UCI.

**Context:**

Heart disease is the leading cause of death for men, women, and people of most racial and ethnic groups in the United States. One person dies every 37 seconds in the United States from cardiovascular disease. About 647,000 Americans die from heart disease each year—that's 1 in every 4 deaths. Heart disease costs the United States about **$219 billion**each year from 2014 to 2015. This includes the cost of health care services, medicines, and lost productivity due to death. This database contains 76 attributes, but all published experiments refer to using a subset of 14 of them. In particular, the Cleveland database is the only one that has been used by ML researchers to this date. The "goal" field refers to the presence of heart disease in the patient. It is integer valued from 0 (no presence) to 4. Experiments with the Cleveland database have concentrated on simply attempting to distinguish presence (values 1,2,3,4) from absence (value 0). [High blood pressure](https://www.cdc.gov/bloodpressure/index.htm), [high blood cholesterol](https://www.cdc.gov/cholesterol/index.htm), and smoking are key [risk factors](https://www.cdc.gov/heartdisease/risk_factors.htm) for heart disease. About **half of Americans** (47%) have at least one of these three risk factors. By studying this dataset, I am trying to find the most key factors of heart disease which might be a good study to reduce the number of people with heart disease.

This dataset gives several variables along with a target condition of having or not having heart disease. Clearly this is a classification problem, therefore various techniques such as SVM, random forest model, and finally all the models are investigated using Machine Learning explainability tools and techniques. Surprisingly, this dataset can also be used as an unsupervised learning and then the results will be compared to the target labels to calculate the accuracy of clustering. Different clustering techniques such as K-means, Gaussian Mixture Models, and DBSCAN can be implemented. Additionally, Keras and Neural networks will be used as well.