Capstone II Problem Statement

Background

Cardiovascular diseases (CVDs) are the leading cause of death worldwide, accounting for about 18 million, or about 1 in 3 deaths yearly. Heart diseases are a subcategory of CVD, and include things like coronary heart disease and heart failure. We are interested in developing a model to help predict the incidence of heart disease, which can be used to help inform decisions to save lives. To that end, we have a dataset consisting of 908 individuals with information across twelve categories, including incidence of heart disease. Other categories of information include age, chest pain type, and cholesterol.

Problem Statement

Our specific goal is this: Of the tracked attributes in our dataset, which are the best predictors of heart disease?

Criteria For Success

We will consider our project successful if we can develop a model that identifies some factors as greater predictors of heart disease than others. Though it may be the case that the variance in predictability of heart disease is small, and none of the factors predict heart disease very well, we will not consider such an outcome to be a failure as long as our model produces numbers with reasonable confidence.

Scope of Solution Space

Our dataset is not particularly large, and it only includes individuals from a handful of places in Europe and North America. Therefore, it should not be considered representative of heart disease trends in the whole world, but only the western world. Furthermore some categories of data may be misleading. For example, the cholesterol feature is measurements of serum cholesterol, not dietary cholesterol.

Constraints

Conclusions derived in this project cannot be taken as a general solution for heart disease. Many factors that probably do contribute to heart disease are not tracked here, notably genetic and lifestyle factors.

It should also not be forgotten that heart disease is only one kind of CVD. Conclusions from this project do not speak to other kinds of CVDs.

Data Sources

The individuals in our case study come from a combination of different datasets already available independently. The five datasets used for its curation have been contributed from

- Cleveland: 303 observations
- Hungarian: 294 observations
- Switzerland: 123 observations
- Long Beach VA: 200 observations
- Stalog (Heart) Data Set: 270 observations

By the following contributors

- Hungarian Institute of Cardiology. Budapest: Andras Janosi, M.D.
- University Hospital, Zurich, Switzerland: William Steinbrunn, M.D.
- University Hospital, Basel, Switzerland: Matthias Pfisterer, M.D.
- V.A. Medical Center, Long Beach and Cleveland Clinic Foundation: Robert Detrano, M.D., Ph.D.

https://www.kaggle.com/fedesoriano/heart-failure-prediction