Udacity Machine Learning Engineer Nanodegree

Capstone proposal – Predicting heart disease

Estevam Ribeiro do Valle Donnabella Santos

July 2019

Domain background

Heart diseases are a huge problem in our modern society. It accounts for one third of the total of death cause in the U.S., yet Google and media coverage is about only 2-3%.

In order to make society more careful about this matter, we could use machine learning to predict whether people are prone to develop heart disease in the future and alert them in order to prevent deaths. This Dataset was used into the article "Heart Disease Prediction Using Machine learning and Data Mining Technique".

Problem statement

The goal is to create a prediction model that predicts with a minimum of 70% of accuracy whether a given person with determined attributes will develop a heart disease or not. The 14 attributes contained in the dataset are well defined and all measureable with a good degree of precision.

Dataset and inputs

In 1988 four institutes collected data from 1025 people and made a dataset of 76 attributes about heart disease. The institutes are listed below

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

Historically, from the 76 attributes, only a subset of 14 are been used and considered relevant for hear disease prediction studies, thus this capstone project will use the 14 features subset as the dataset. The target attribute refers to either 0 (no heart disease presence) or 1 (heart disease presence). This dataset was cited in this article

"Heart Disease Prediction Using Machine learning and Data Mining Technique" but was collected in Kaggle.

All the features are number quantities as shown below

1. Age
2. Sex
3. Chest pain type (4 values)
4. Resting blood pressure
5. Serum cholestoral in mg/dl
6. Fasting blood sugar > 120 mg/dl
7. Resting electrocardiographic results (values 0,1,2)
8. Maximum heart rate achieved
9. Exercise induced angina
10. Oldpeak = ST depression induced by exercise relative to rest
11. The slope of the peak exercise ST segment
12. Number of major vessels (0-3) colored by flourosopy
13. Thal: 3 = normal; 6 = fixed defect; 7 = reversable defect
14. Target: 0 = no presence of heart disease; 1 = presence of heart disease

The target feature is well distributed (499 for “no presence of heart disease“ and 526 for “presence of heart disease“)

Solution statement

The solution is to use different supervised machine learning techniques and compare them to see which one would be more accurate to solve the problem. After determining the fine-tuned supervised method, the trained model will be run against the test dataset for validation.

Benchmark model

The article "Heart Disease Prediction Using Machine learning and Data Mining Technique" reach a best result of 56.76% accuracy using Decision Tree with algorithm J48 with Reduced Error Pruning. This model is going to be used as the Benchmark model.

Evaluation metrics

The evaluation metric is the accuracy score, since the target feature is well balanced.

Project design

The project design will be as follow:

1. Data analysis – See if there is outliers, odd data and so on and treat them. Boolean features with no answer will be discarded. Outliers from wide range

numeric features like age, trestbps, chol and so on will be erased if the 1.5*IQR rule is reached.

2. Data randomly split in 3 subsets of 70% (training set), 15%(validation) and 15% (testing).
3. The training dataset will be run against at least 3 different supervised machine learning methods, in principle Logistic regression, Decision Trees and SVM.
4. Grid-search the methods to fine tune them and compare the maximized models.
5. Evaluate the model against the test subset.