Heart Disease Binary Classifier

Gangadhar Dhulipala  
DSC680 - Fall 2019  
<https://github.com/gdhulipala/DSC680>

# Which Domain?

Given my experience in the healthcare domain- I decided to go with a dataset that corresponds to the healthcare domain. Predominantly, I am focusing on the datasets where I can apply the machine learning algorithms to gain more insight in terms their application in disease prediction.

# Which Data?

Below is the URL address for the heart disease dataset.

<https://www.kaggle.com/ronitf/heart-disease-uci/downloads/heart-disease-uci.zip/1>

The dataset has about 14 attributes which are mainly the health profile attributes of different patients. For example, for each of the patients listed in the dataset, the data points include their heart rates, blood pressure in the resting versus active time heart care, whether they are smokers or non-smokers etc. Given such information about the patients, I am going to see if these algorithms are capable of predicting heart disease in patients.

# Research Questions? Benefits? Why analyze these data?

The research question in the current dataset is all about identifying the attributes that are more likely to predict the presence of heart disease in a given patient. The outcome will be really helpful and has a lot of practical applications in the real world. For example, healthcare providers and hospitals, primary care physicians can use such information to predict the disease in the patients and hence will be able to provide preventive measures to their patients.

Even though the question in the dataset is dealing about predicting the heart disease in the patients- but the end goal of the project is to build a generic model that can be used to apply and get the results in pretty much all other disease areas.

# What Method?

I am going to take a different approach in terms of choosing the method. In other words- instead of just choosing one classification algorithm to predict a disease, I will be using a combination of different algorithms in predicting the outcome. For example, since this is a classification problem, I am going to apply different classification algorithms such as logistic regression, Naïve Bayes algorithm, decision trees, random forests and SVM etc.

The main reason for taking this approach is just to get a global picture of how different algorithms analyze the same dataset and go further to understand on how to handle the results when the outcome from different algorithms is vastly different. Do we need to pick and choose or do we perform a logical elimination of one algorithm versus the other algorithm?

# Potential Issues?

As mentioned above, the potential issue might be the fact that different algorithms may give totally different output for a given dataset given their limitation associated with the individual machine learning models. This is one area where I might run into issues. Hence, I need to come up with plan of action to deal with such ambiguity as when it arises. One way to pick one best algorithm would be probably by looking at each of these algorithms in detail to understand the logic associated with the classification and does it fit well with the dataset in hand.

# Concluding Remarks

Summarizing everything we discussed above – Given my experience in the biotechnology field, I am going to work on the dataset related to the health care domain.

As a part of it I choose to go with a dataset where the data included the details about the health profile of different patients. For example, each row corresponds data corresponding to one patient which has details such as their heart rate, blood pressure, glucose levels etc. By using these details about the patient, we have to predict whether he or she has a heart disease. If a model that is able to predict the heart disease outcome based on this dataset, it would have a huge real-world application where people in the healthcare domain can use it to apply on different patients and based on the prediction the patients can be provided with the necessary preventive care. Coming to the actual model building aspect, I am going to predominantly stick with the concept of ensemble modeling where I would like test and combine different algorithm as a goal to get the best model; that can predict the outcome with decent accuracy.

References:

<https://towardsdatascience.com/predicting-presence-of-heart-diseases-using-machine-learning-36f00f3edb2c>

<https://medium.com/@dskswu/machine-learning-with-a-heart-predicting-heart-disease-b2e9f24fee84>

<https://www.researchgate.net/publication/330981991_Analysis_of_Heart_Disease_Prediction_using_Various_Machine_Learning_Techniques>

<https://www.scirp.org/html/14-1560633_88650.htm>

<https://www.sciencedaily.com/releases/2019/05/190513081412.htm>

<https://ieeexplore.ieee.org/document/8740989>

<https://pdfs.semanticscholar.org/d0a5/d4b8e8da3ee2a6bf8ac5d44196fb0365cf1c.pdf>

<https://www.hindawi.com/journals/misy/2018/3860146/>