Heart Attack Prediction

Satish Kumar Agrawal
DSC 650 – Summer 2021

Which Domain?

Heart Disease is one the leading cause of human death. One person dies every 36 seconds in United states alone due to cardiovascular Disease. Approximately 25% of the deaths (1 out of 4 deaths) are due to heart Disease every year and it costs $219 Billions to the United States. It's not common that a person dies when they get a heart attack the very first time, but it definitely does a huge damage to the body and makes it prone to another attack and other Diseases. Heart Disease is a condition which grows over the period of time and impacts the body incrementally. If using the body conditions and blood report we can predict the potential future heart attack, the person can adjust the lifestyle or medications could started early to avoid any future heart attack. This does not eliminate the risk of future Disease but at least raises an alarm ahead of time which can help avoiding it from becoming fatal.

Which Data?

I have identified a heart attack data set from Kaggle.com so far. I am intending to search for more data set with different attributes and more observations. The data set at hand has 14 columns out of which 13 are explanatory variable and one is the dependent variable. Below are the attributes with the possible value set for categorical variables. As you can see, the categorical variables are already encoded and converted to numerical values.

Output is also encoded to 1 for chances of heart attack and 0 for no chances of heart attack.

age - Age of the patient

sex - Sex of the patient
    1 = Male,
    0 = Female

cp - Chest pain type
    0 = Typical Angina,
    1 = Atypical Angina,
    2 = Non-anginal Pain,
    3 = Asymptomatic

trtbps - Resting blood pressure (in mm Hg)

chol - Cholesterol in mg/dl fetched via BMI sensor

fbs - (fasting blood sugar > 120 mg/dl)
    1 = True,

0 = False

restecg - Resting electrocardiographic results
0 = Normal,
1 = ST-T wave normality,
2 = Left ventricular hypertrophy

thalachh - Maximum heart rate achieved

oldpeak - Previous peak

slp – Slope

caa - Number of major vessels

Research Questions? Benefits? Why analyze these data?

Heart attack is one the common Disease in the nation and the world. It is leading cause of death for the people of most of the races and ethnicity. 23.4% of all the deaths are caused by heart attack. People do go to their annual body checkups and get a chance to look at the various facts of their heart health. This research intends to be able to predict if a patient is at the risk of heart attack.

Few of the main indicators of human health are cholesterol in mg/dl (LDL), resting blood pressure [8 Understanding Blood pressure readings], electrocardiographic [10 wiki] and fasting blood sugar [9 Blood sugar to make all the difference] etc. So, research questions here are:

- Do above factors indicate potential heart attack in future?
- Can heart condition be predicted and a stroke be avoided?
- What the best health record looks like that gives a patient confident that he is free from the risk of attack?
- Is the sex of the person or age group more prone to heart attacks than others?

In this project, I intend to answer some of the questions and have a repeatable process to take in details of new patients and predict if they need to alter their lifestyle, diet or any thing else to lead a healthy life.

What Method?

The output of this project is whether a patient is prone to heart attack or not. The data set has many other attributes of a patient condition that may be strongly correlated to the target variable. I plan to explore the data looking for any anomalies like missing data, duplicate records or outliers in the data set. I would work towards imputing missing values as possible or dropping columns with excessive missing values. I consider a column should be dropped if it has more than 50% of the records missing values, which again depends upon the dataset at hand.

Looking at the dataset high level, I do not see the need of engineering other features to enhance the results but I do plan to find the attributes which are strongly correlated and drop the ones which are not correlated with the output variable. I would also look if two explanatory variables are strongly correlated to each other. It is best to drop one of them as keeping them both on is not going to be helpful.

As the output of the learning is a Boolean 1 for chances of getting heart attack and 0 for no chance of getting heart attack. I am planning to train a DecisionTree Classifier model. Apart from training Decision tree Classifier I will also train Logistic regression and K neighbors classifier models. I will use exhaustive training method using GridSearchCV from sklearn to perform model training and selection of best performing parameters. To compare models with each other and individual model performance I will use F1 score.

Before doing any of the steps listed above, I will split at least 25% of the data set and set aside for testing the final model. The training will be performed only in 75% of the total data set.

Potential Issues?

The success of quality of a machine learning project depends upon the quality of the data used for training. We have a rather smaller dataset at hand at this point. A large data set is better for the training, but small data set would also be ok if the observations are IID – Independent identically distributed. While training the model, I will be using k-fold cross validation and use shuffle = true and stratified to make sure I have good representation of both classes (1 and 0) in all the folds.

Also, it will crucial to see if data set has any bad data or missing values. Excessive missing values are not good for training the model. I do plan to implement imputation as needed

Concluding Remarks

Heart Disease takes a lot of human life every year. Our body gives us signs and indications to take care of ourselves and avoid any loss of life untimely due to this Disease. In the modern world we have accumulated enough data and are accumulating more and more every day. I am confident all these data features and observations around human health and end result can be leveraged to predict any future risk of death due to heart attack. In this project we will work with the data to train potential models with acceptable accuracy to be able to predict future heart conditions given current health status. The model trained and chosen here for production deployment will need to be retrained as time goes and we collect more data.

References:

[1] https://www.kaggle.com/kumudadk/heart-attack-prediction-and-analysis?select=heart.csv

[2] Heart Disease and Strike Statistics – 2021 Update - https://www.ahajournals.org/doi/10.1161/CIR.0000000000000950

[3] Heart Disease data set – https://archive.ics.uci.edu/ml/datasets/Heart+Disease/

[4] Heart Disease facts - https://www.cdc.gov/heartdisease/facts.htm

[5] Heart Diseases - https://medlineplus.gov/heartdiseases.html

[6] Heart and Stroke Statistics - https://www.heart.org/en/about-us/heart-and-stroke-association-statistics

[7] Heart Disease: Facts, Statistics, and you - https://www.healthline.com/health/heart-disease/statistics

[8] Understanding Blood pressure readings - https://www.heart.org/en/health-topics/high-blood-pressure/understanding-blood-pressure-readings

[9] Blood sugar to make all the difference - https://www.diabetes.org/healthy-living/medication-treatments/blood-glucose-testing-and-control

[10] Electrocardiography – wiki - https://en.wikipedia.org/wiki/Electrocardiography#:~:text=Electrocardiography%20is%20the%20process%20of,electrodes%20placed%20on%20the%20skin.