Capstone Project: Heart Disease Prediction using Machine Learning

By

Rezwanur Rahman (Rez)

Background: Heart disease is very common among different ages. People with certain blood sugar level, cholesterol level etc., may exhibit symptoms for heart disease. It is important to find a smart way to speculate possibility of heart disease given certain physiological status. UCI ML database has a large, collected data for heart disease will be used to develop a ML model. Historically great effort has been made to detect prime arbitrator of heart disease given various parameters [1]. This work highlights on few analytical metrics without proposing a ML prediction methodology(s). There has been a necessity to do a systematic analysis on "primary" causes of heart disease to trigger alarm for a person. Overall analytics and prediction can be compiled as a streamlined process to work on any probable future datasets. As a part of the work author will be creating a ML pipeline to predict possibility of heart disease using multiple candidate ML methods. This will be done to ensure reliability of overall prediction.

Problem Statement: This data has several parameters that encompasses a relationship between person's general health status and potential heart risk. A ML model can be used to extract important parameter (aka features) while analyzing the data. Eventually ML model can be used to predict heart risk.

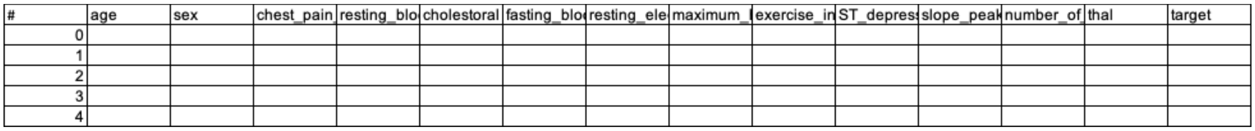
Dataset and Input: In this project UCI ML dataset on heart disease has been used [2]. From the dataset has a wide range of input features:

- age: The person's age in years
- sex: The person's sex
- trestbps: The person's resting blood pressure (mm Hg on admission to the hospital)
- chol: The person's cholesterol measurement in mg/dl
- fbs: The person's fasting blood sugar
- restecg: resting electrocardiographic results
- thalach: The person's maximum heart rate achieved
- exang: Exercise induced
- oldpeak: ST depression induced by exercise relative to rest
- slope: the slope of the peak exercise ST segment
- ca: The number of major vessels
- thal: A blood disorder called thalassemia

As part of pre-processing following steps will be done to cleanup the data:

1. Find out missing values by looking for NULL values
2. Find out duplicate attributes
3. Look for outliers in the data
4. Rename features if necessary

Once the data is pre-processes and prepared for ML study it will be ingested in the ML pipeline. Final goal is to create a dataset with tabular structure:



#	age	sex	chest_pain	resting_blo	cholestoral	fasting_blo	resting_ele	maximum_	exercise_in	ST_depres	slope_peak	number_of	thal	target
0														
1														
2														
3														
4														

Solution Statement: To predict heart risk given dataset an exploratory data analysis will be done. Important features will be identified. These will be used to classify possible risk of heart disease or not.

The overall solution process is outlined as follows:

1. Perform exploratory data analysis to develop an understanding on relationship between heart disease and a given attribute. For example, a histogram in Fig. 1 depicts a good idea on relationship between heart disease and age.

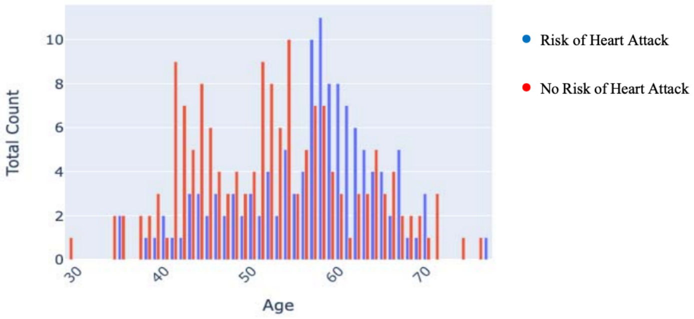


Fig. 1

2. Develop clustering models to use unsupervised learning in finding a pattern of self-similarity in the data. For example, can we say "Young females have tendency to have low heart disease"?
3. Develop a regression model(s) to conclude a relationship between most significant features and risk of heart disease.

Benchmark Model: Data can be split into train and test parts. Model can be benchmarked based on its performance with test dataset. Model performance can be compared with literature data [1, 3].

Evaluation Metrics: Performance of the ML model can be determined in terms of ROC curve and confusion matrix.

Project Design: For this project ML package Scikit-learn will be used for both supervised and unsupervised ML methods. Sklearn is a rich an highly applied python package. As ML methodologies for regression we will use models like: RandomForest Regression, Logistic regression, XGboost regression etc. Idea is to use effective regression method capable of capturing both linear and non-linear relationships in the data. To avoid overfitting we will be using k-fold cross validation method on training set. K-nearest neighbor based clustering will be used to find any self similar pattern in the data. Model will be deployed on GitHub repo.

Reference:

1. Jindal, Harshit, et al. "Heart disease prediction using machine learning algorithms." IOP Conference Series: Materials Science and Engineering. Vol. 1022. No. 1. IOP Publishing, 2021.
2. https://archive.ics.uci.edu/ml/datasets/Heart+Disease
3. Albahr, Abdulaziz, et al. "Computational Learning Model for Prediction of Heart Disease Using Machine Learning Based on a New Regularizer." Computational Intelligence and Neuroscience 2021 (2021).