Machine Learning Engineer Nanodegree

Predicting Heart Disease

Bavly AbdEl-masih fayez
December 22st, 2019

Proposal

Domain Background

Imagine getting a simple blood test to help doctors predict your risk for having a heart attack.

In recent years, these tests have become so refined that some can detect very low levels of these proteins, known as troponin. Researchers determined that troponin levels in healthy middle-aged to older adults could help predict their risk for eventually developing cardiovascular disease. Their findings were published Monday in the American Heart Association journal Circulation.

Researchers examined a group of 8,121 people, ages 54 to 74, with no history of cardiovascular disease. Troponin levels were detected in 85% of the group. Higher levels of the protein were associated with a greater chance of developing cardiovascular disease, particularly heart failure.

The study found that highly sensitive troponin tests were especially good at predicting cardiovascular events when added to the results of a special equation commonly used to calculate a person's 10-year risk of having a heart attack or stroke.

So by observing this studies heart disease can be predicted by collecting data of peoples health state history trying to designing machine learning model to predict if person will suffer from heart disease

Machine Learning is used across many spheres around the world. The healthcare industry is no exception. Machine Learning can play an essential role in predicting presence/absence of Locomotor disorders, Heart diseases and more. Such information, if predicted well in advance, can provide important insights to doctors who can then adapt their diagnosis and treatment per patient basis.

Problem Statement

In the United States, the Centers for Disease Control and Prevention is a good resource for information about heart disease. According to their website:

About 610,000 people die of heart disease in the United States every year–that's 1 in every 4 deaths.

Heart disease is the leading cause of death for both men and women. More than half of the deaths due to heart disease in 2009 were in men.

Coronary heart disease (CHD) is the most common type of heart disease, killing over 370,000 people annually.

Every year about 735,000 Americans have a heart attack. Of these, 525,000 are a first heart attack and 210,000 happen in people who have already had a heart attack.

Heart disease is the leading cause of death for people of most ethnicities in the United States, including African Americans, Hispanics, and whites. For American Indians or Alaska Natives and Asians or Pacific Islanders, heart disease is second only to cancer.

So predicting heart disease is an critical solution for this problem trying to save the life of many people by tracking the health state of people trying to predict if they are likely to be suffer from heart disease in the early future we will use kaggle dataset that collected from 300 person about their historical health state and who has suffer from heart disease to design our model

Datasets and Inputs

The data is provided on kaggle This database contains 76 attributes, but all published experiments refer to using a subset of 14 of them collected from 303 persons , where the patient_id column is a unique and random identifier. The remaining 13 features are described in the section below.

- slope_of_peak_exercise_st_segment (type: int): the slope of the peak exercise ST segment, an electrocardiography read out indicating quality of blood flow to the heart

- thal (type: categorical): results of thallium stress test measuring blood flow to the heart, with possible values normal, fixed_defect, reversible_defect

- resting_blood_pressure (type: int): resting blood pressure

- chest_pain_type (type: int): chest pain type (4 values)

- num_major_vessels (type: int): number of major vessels (0-3) colored by flourosopy

- fasting_blood_sugar_gt_120_mg_per_dl (type: binary): fasting blood sugar > 120 mg/dl

- resting_ekg_results (type: int): resting electrocardiographic results (values 0,1,2)

- serum_cholesterol_mg_per_dl (type: int): serum cholestoral in mg/dl

- oldpeak_eq_st_depression (type: float): oldpeak = ST depression induced by exercise relative to rest, a measure of abnormality in electrocardiograms

- sex (type: binary): 0: female, 1: male

- age (type: int): age in years

- max_heart_rate_achieved (type: int): maximum heart rate achieved (beats per minute)

- exercise_induced_angina (type: binary): exercise-induced chest pain (0: False, 1: True)

by analyzing the data I found that dataset is similar have a lot of similar occurrence and is more balanced

Solution Statement

A classification model based on good data-driven systems for predicting heart disease can improve the entire research and prevention process, making sure that more people can live healthy lives.

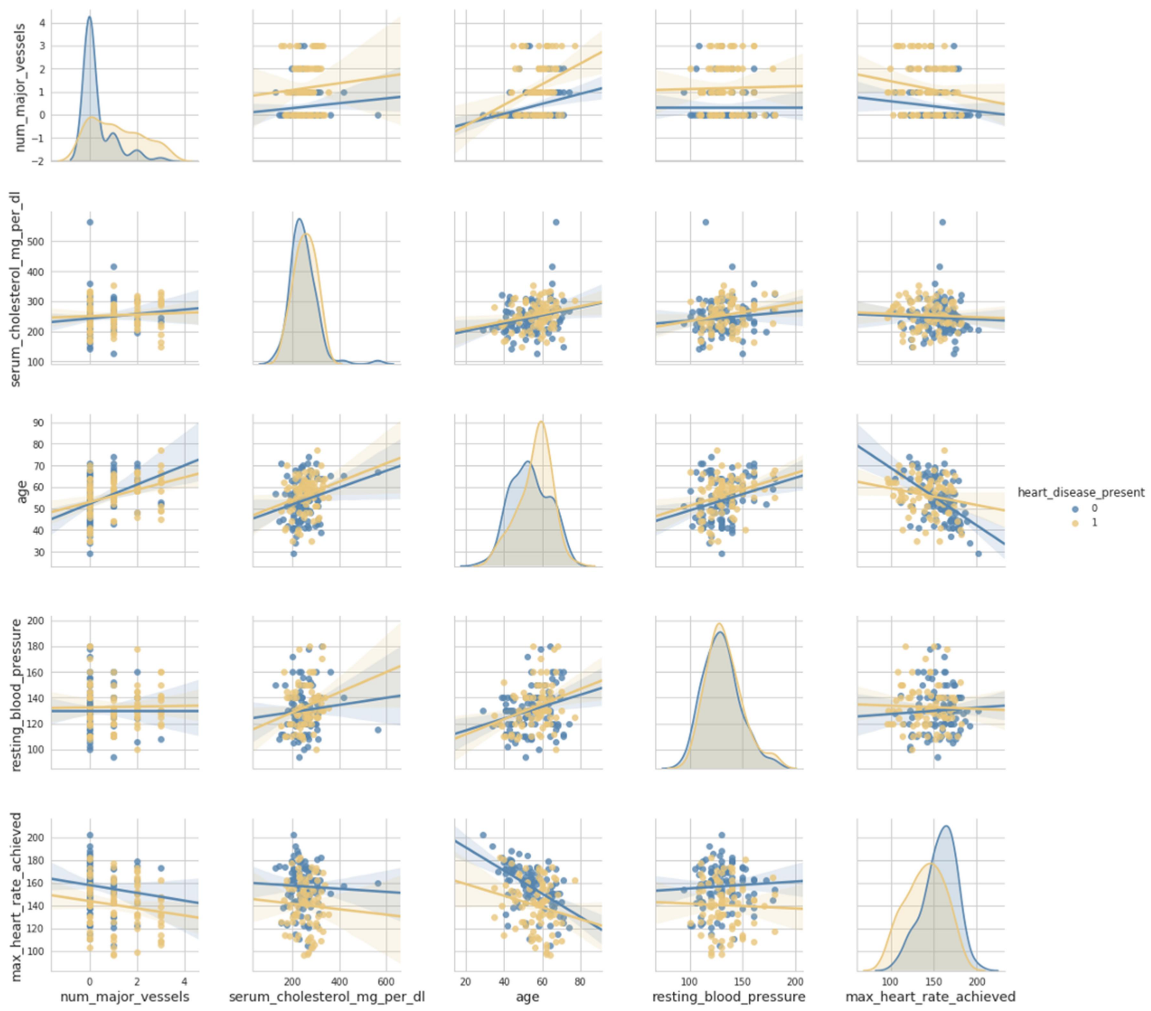
Benchmark Model

A predicting model that whether or not the patient have heart disease or not with accuracy more than 80 percent on test dataset .

Evaluation Metrics

From the frequency plot of heart disease below, we see that the two classes (‘Heart Disease’ and ‘No Heart Disease’) are approximately balanced, with 45% of observations having heart disease and the remaining population not having heart disease.

From this observation we could use accuracy as our evaluation metrics



Project Design

Programming language: *Python 3.6*

Import libraries

I imported several libraries for the project:

1. **numpy**: To work with arrays
2. **pandas**: To work with csv files and dataframes
3. **matplotlib**: To create charts using pyplot, define parameters using rcParams and color them with cm.rainbow
4. **warnings**: To ignore all warnings which might be showing up in the notebook due to past/future depreciation of a feature
5. **train_test_split**: To split the dataset into training and testing data

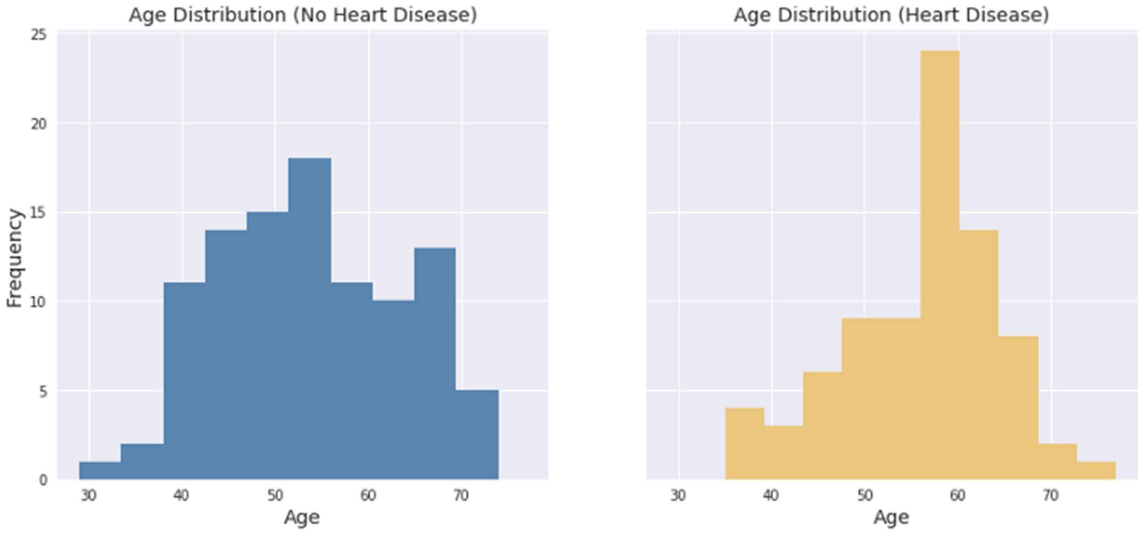
6. **StandardScaler**: To scale all the features, so that the Machine Learning model better adapts to the dataset

Workflow:

Our project will be divided into 3 steps :

1- ***Understanding the data :*** *begin with, let's see the correlation matrix of features and try to analyze it and here is an example*

- *Resting blood pressure tends to increase with age regardless of heart disease.*
- *We can see that max heart rates are significantly lower for people without heart disease.*



2- ***Data Processing:***

- *Splitting data to train and test dataset*
- *To work with categorical variables, we should break each categorical column into dummy columns with 1s and 0s.*

3- ***Machine Learning:*** *I took 4 algorithms(knn – svm - Decision Tree Classifier - Random Forest Classifier) and varied their various parameters and compared the final models. I split the dataset into 70% training data and 30% testing data.*

Also use grid search algorithm to find the best hyper parameters