

Capstone Project

Machine Learning Engineer Nanodegree

Heart_Disease_Prediction

Usha Sree
February 7th, 2019

Proposal

Domain Background

Heart disease is the number one killer of men and women in the United States today. The Centers for Disease Control and Prevention (CDC) estimate that heart disease causes about 1 in 4 deaths in the United States each year. That's 610,000 people per year. About 735,000 people in the United States have a heart attack each year.

Heart disease is considered one of the top preventable causes of death in the United States. Some genetic factors can contribute, but the disease is largely attributed to poor lifestyle habits. Nowadays, health disease are increasing day by day due to life style, hereditary. Especially, heart disease has become more common these days, i.e. life of people is at risk.

Among these are poor diet, lack of regular exercise, tobacco smoking, alcohol or drug abuse, and high stress. These are issues that remain prevalent in American culture, so it's no wonder that heart disease is of great concern.

I am finding out accuracy and F-score for Heart disease dataset in this project.

Problem Statement:

The main objective of this study is to build a model that can predict the heart disease occurrence, based on a combination of features (risk factors) describing the disease. Different machine learning classification techniques will be implemented and compared upon standard performance metric such as accuracy and F-score.

Heart Disease Prediction predicts the likelihood of patients getting heart disease.

There are several factors that increase the risk of heart disease, such as smoking habit, body cholesterol level, family history of heart disease, obesity, high blood pressure, and lack of physical exercise.

Datasets and Inputs:

In this data set, I have used 14 attributes and around 500 trained and test data to evaluate accuracy and F-score.

Experiments with the Cleveland database have concentrated on endeavors to distinguish disease presence (values 1, 2, 3, 4) from absence (value 0). There are several missing attribute values, distinguished with symbol '?'. The header row is missing in this dataset, so the column names have to be inserted manually.

Features information:

- age - age in years
- sex - sex(1 = male; 0 = female)
- chest_pain - chest pain type (1 = typical angina; 2 = atypical angina; 3 = non-anginal pain; 4 = asymptomatic)
- blood_pressure - resting blood pressure (in mm Hg on admission to the hospital)
- serum_cholesterol - serum cholesterol in mg/dl
- fasting_blood_sugar - fasting blood sugar > 120 mg/dl (1 = true; 0 = false)
- electrocardiographic - resting electrocardiographic results (0 = normal; 1 = having ST-T; 2 = hypertrophy)
- max_heart_rate - maximum heart rate achieved
- induced_angina - exercise induced angina (1 = yes; 0 = no)
- ST_depression - ST depression induced by exercise relative to rest
- slope - the slope of the peak exercise ST segment (1 = upsloping; 2 = flat; 3 = downsloping)
- no_of_vessels - number of major vessels (0-3) colored by flourosopy
- thal - 3 = normal; 6 = fixed defect; 7 = reversable defect
- diagnosis - the predicted attribute - diagnosis of heart disease (angiographic disease status) (Value 0 = < 50% diameter narrowing; Value 1 = > 50% diameter narrowing)

Types of features:

1. **Categorical features** (Has two or more categories and each value in that feature can be categorised by them): sex, chest_pain
2. **Ordinal features** (Variable having relative ordering or sorting between the values): fasting_blood_sugar, electrocardiographic, induced_angina, slope, no_of_vessels, thal, diagnosis
3. **Continuous features** (Variable taking values between any two points or between the minimum or maximum values in the feature column): age, blood_pressure, serum_cholesterol, max_heart_rate, ST_depression

Solution Statement:

In this project I am trying to predict the rate of heart disease for the selected data set. For this we want to use different classification models like Support Vector Machine, Random forest, Logistic Regression, Naïve Bayes etc. Then we will find the accuracy and F-score for each classification model. I will explore the data set with matplotlib.py library to plot. Visualization helps me to better understand the result.

Benchmark Model:

Here we compare the final model with the remaining models to see if it got better or same or worse. The accuracy and F-scores are compared among the models and the optimal one is selected.

Evaluation Metrics:

Accuracy:

It is the number of correct predictions made by the model over all kinds of predictions made

$$\text{Accuracy} = (\text{TP} + \text{TN}) / (\text{TP} + \text{FP} + \text{FN} + \text{TN})$$

Accuracy is a good measure when the target variable classes in the data are nearly balanced.

F-score:

It is used to measure a test's accuracy and it balances the use of precision and recall to do it. It can provide a realistic measure of test's performance.

Project Design:

The project will be composed of the following steps:

Data Exploration: Visualizing the dataset, detect outliers, replacing a missing value and cleaning the dataset, splitting training dataset into training and testing sets etc.

Training and Testing: Testing and training conducted on data sets by using various models such as KNN, Decision Trees, Logistic Regression, Naive Bayes, SVM, Random Forests.

The collected data were used to create a structured database system. The pre-processing was done by identifying the associated fields and removing all the duplications. After that, all the missing values were filled, and the data were coded according to the domain value.

Finally, I declare that the model with the highest accuracy and F-score on both training and testing dataset will be concluded as the best model for predicting the rate of heart disease