HEART DISEASE PREDICTION

Development of a Classifier to Diagnose a Heart Disease.

Firoj Ahmmed Patwary

Full list of author information is available at the end of the article

Abstract

Goal of the Project: To find out whether a person is suffering from heart disease or not.

Results of the Project: Neighbors Classifier scored the best score of 87% with 8 neighbors.

My Key Learning: Train Test Data, K Neighbors Classifier, Support Vector Classifier, Decision Tree Classifier, Random Forest Classifier, Current Heart Disease Rate.

Spent Time: Almost 7 Hours.

Project Evaluation Scale: 1: "Really Good Project".

Project Word: Total Words :887, Headers :13, Math Inline :1

1. Scientific Background of the Project

Nowadays, heart disease is the leading cause to die of human being. Coronary heart disease is the most common type of heart disease, killing 370,000 people per annum. That's why heart disease makes a major concern to dealt with it. But it is difficult to find out the actual reason of this disease due to several contributory risk factors such as diabetes, high blood pressure, high cholesterol and so on. Now Machine Learning is used in many problems in this globe. Machine Learning are playing an important role in healthcare industry. It can predict presence or absence of many kinds of disease by using data.

In this project we used Cleveland Heart Disease Dataset taken from UCI Machine Learning repository which was donated by David W. Aha and creators were Andras Janosi, William Steinbrunn, Matthias Pfistere and Robert Detrano.

2. Goal of the Project

In this project we will be applying some Machine Learning approaches whether a person is suffering from heart disease or not, using Cleveland Heart Disease Dataset from UCI Machine Learning repository.

3. About Data

The dataset consists of 303 individuals data. There are 14 columns in the dataset and they are: age, sex, chest-pain type, Resting Blood Pressure, Serum Cholesterol, Fasting Blood Sugar, Resting ECG, Maximum heart rate achieved, Exercise induced angina, ST depression induced by exercise relative to rest, Peak exercise ST segment, Number of major vessels (0-3) colored by fluoroscopy, Thal and Diagnosis of heart disease.

3.1. Data Preprocessing Steps

At first, I just wanted to look at the data and I used info()method. Then we can find 13 features with 1 target variable but no missing value. So, I don't need to concern about missing value. Then I used described()method to see the mean, standard deviation, quartiles and so on. After that, to understand the data, I found out correlation matrix, histogram and bar plot for target class. In correlation matrix, I saw that there has no single feature that has a very high correlation with our target value. Although, some of the features had negative and some had positive correlation. In bar plot, it is easy to identify that classes are almost balanced and further processing is possible.

For categorical variable, I changed categorical column to a dummy column and took 1 for male and 0 foe female. Here used *getdummies* method from Pandas. Now the data set is ready to training by machine learning models.

4. Description of the Used Method(s)

To achieve our goal, I used four machine learning classifiers in this project and split the data set into 67% training data and 33% testing data.

4.1. K Neighbors Classifier

In this classifier I the neighbors from 1 to 21 neighbors and calculated the test score in each case. Finally, I plotted a line graph of the numbers of neighbors and the test scores schived in each case. Here, the maximum score of 87% when the number of neighbors was chosen to be 8.

4.2. Support Vector Classifier

To forming a hyperplane, I used support vector classifier and it can separate the classes by adjusting the distance between the data points and hyperplane as much as possible. Here, I tried four kernels and they were linear, poly, rbf and sigmoid. To make different colors I used rainbow method. Here I got 83% score for linear kernel.

4.3. Decision Tree Classifier

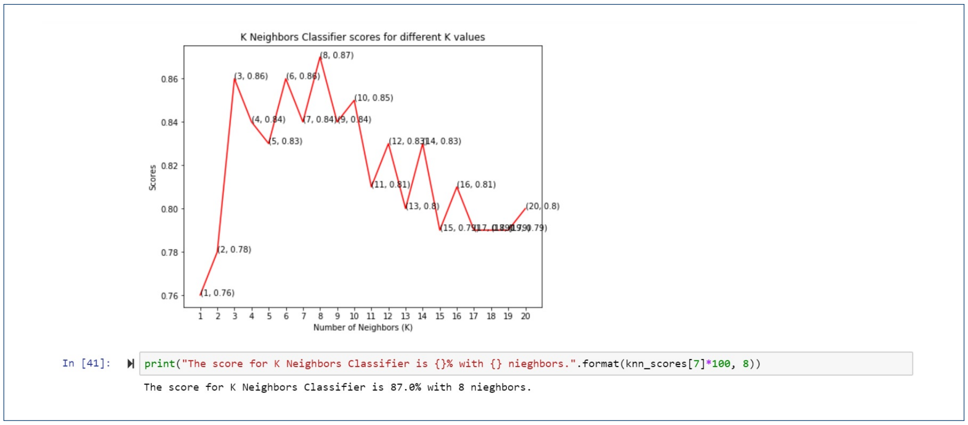
Here I used the range feature from 1 to 30. This classifier makes a decision tree based on the assigned class values of each data points. Finally, I plotted a line graph to see the effect of the number of features on the model scores. Here I found that the maximum score is 79% and it achieved for 2, 4 and 18 features.

4.4. Random Forest Classifier

Random forest classifier creates a tree forest where each tree is formed by a random selection of feature from the total features. I calculated test scores over 10, 100, 200, 500 and 1000 trees. Then I plotted this score in a bar graph to see which gave the best results. Maximum score was 84% for both 100 and 500 trees.

5. Results

K Neighbors Classifier provides 87%, Support Vector Classifier provides 83%, Decision Tree Classifier provides 79% and Random Forest Classifier provides 84% maximum scores. Hence, K Neighbors Classifier scored the best score of 87% with 8 neighbors.



```
In [41]:  print("The score for K Neighbors Classifier is {}% with {} nieghbors.".format(knn_scores[7]*100, 8))
```

```
The score for K Neighbors Classifier is 87.0% with 8 nieghbors.
```

6. Discussion (Why this is a typical project for a data-scientist or why not)

First, I used basic information and correlation matrix to know about the relation between multiple variables to assess whether some attributes are necessary for our classification. Here, I found that I can actually work with less than the normal number of parameters. This process is typical for the work of a data scientist, as a closer look is taken into the data and the attributes.

In this project, I used K Neighbors, Support Vector, Decision Tree and Random Forest to classify a given data set into whether a person is suffering from heart disease or not. As these all algorithms are machine learning techniques, this is a pivotal part of Data Science.