Heart Disease-Data Analysis & Prediction

Assignment-4

By: Puspanjali Sahoo (McGill ID: X008997)

Date: 23rd April 2019

Table of Contents

[1. Introduction 3](#_Toc6918750)

[2. Dataset 3](#_Toc6918751)

[2.1 Understanding Data 3](#_Toc6918752)

[Attributes 3](#_Toc6918753)

[2.2 Visualization 4](#_Toc6918754)

[2.2.1 Categorical Variable. 4](#_Toc6918755)

[2.2.2 Continuous Variables 7](#_Toc6918756)

[3. Methodology 9](#_Toc6918757)

[4. Results 12](#_Toc6918758)

[5. Conclusion 12](#_Toc6918759)

# Introduction

The objective of this assignment is to evaluate and analyze the Heart-Disease data-set and predicting the followings.

1. The risk-factors associated with the presence of a heart disease.
2. Predicting easily the risk of having a heart disease? So that the end user would  
   like to be able on their own to know the risk for future patients.

The data has been analyzed using python.

# 2. Dataset

## 2.1 Understanding Data

This database contains 14 variables for 303 randomly selected patients. The data set contains 9 categorical and 5 continuous variables.

The target is a binary variable: 1 for presence of a heart disease and 0 otherwise.  
The other 13 variables are attributes of the patients and following are the descriptions of all the attributes.

### Attributes

|  |  | Description | Variable |
| --- | --- | --- | --- |
|  | age | age in years | continuous |
|  | sex | 1 = male, 0 = female | categorical |
|  | cp | chest pain type: 1: typical angina, 2: atypical angina, 3: non-anginal pain, 4: asymptomatic | categorical |
|  | trestbps | resting blood pressure in mm Hg | continuous |
|  | chol | serum cholestoral in mg/dl | continuous |
|  | fbs | fasting blood sugar > 120 mg/dl: 1 = true, 0 = false | categorical |
|  | restecg | 0: normal, 1: having ST-T wave abnormality, 2: left ventricular hypertrophy | categorical |
|  | thalach | maximum heart rate achieved | continuous |
|  | exang | exercise induced angina (1 = yes; 0 = no) | categorical |
|  | oldpeak | ST depression induced by exercise relative to rest | continuous |
|  | slope | the slope of the peak exercise ST segment: 0: downsloping 1: upsloping, 2: flat | categorical |
|  | ca | number of major vessels: (0-4) colored by flourosopy | categorical |
|  | thal | 1: normal, 2: fixed defect, 3: reversible defect | categorical |
|  | target | diagnosis of heart disease: (0 = false, 1 = true | categorical |

## 2.2 Visualization

In this section, the categorical and continuous variables are compared with target variable to identify the risk-factors associated with the presence of a heart disease.

### 2.2.1 Categorical Variable.

Figure 1 : This figure shows that out of 303 people, 165 are affected with heart diseases and 138 are without the heart diseases.

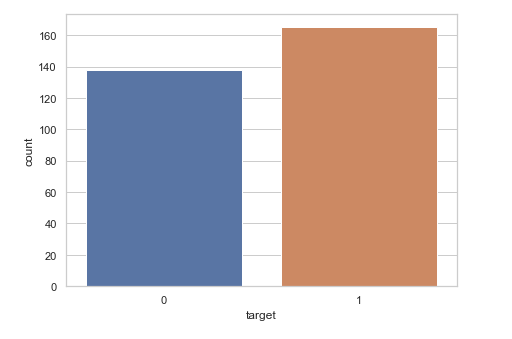


Figure 1: No of People affected with Heart Diseases

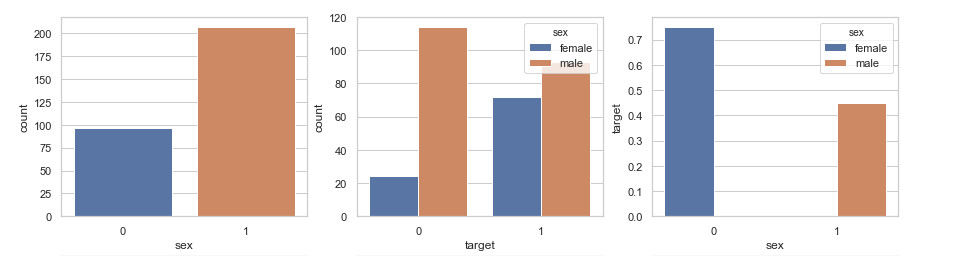


Figure 2: Heart diseases among male and female

The variation of heart disease among male and female is shown in Figure 2. It is observed that there are more male heart disease patients than female. Sex feature has a strong influence on the target variable. The last part of **Figure 2** shows that Female have a higher frequency of heart disease as compared male.

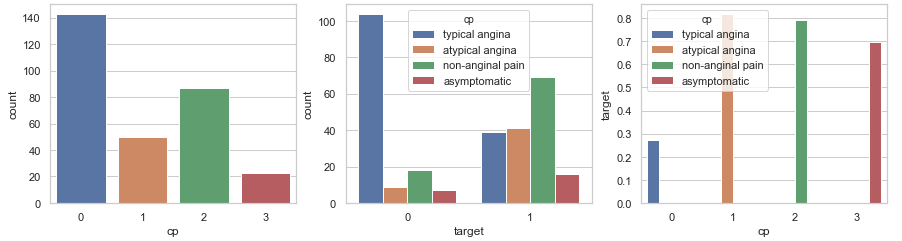


Figure 3: Observation on Heart diseases with different Chest pain type

Figure 3 shows the probability of heart diseases with different type of chest pain. The 3rd part shows People with the atypical angina and non-anginal pain, seem to have the highest probability of getting a heart disease followed by asymptomatic and typical angina type pain.

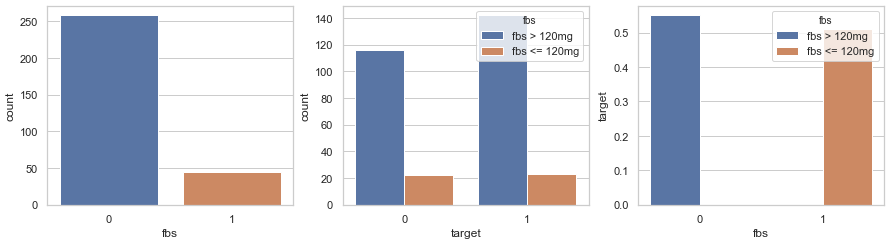


Figure 4: Effect of fasting blood sugar on Heart diseases

Figure 4 shows the effect of fasting blood sugar (fbs) on heart diseases. It is observed that fasting blood sugar level of higher or lower than 120 mg/dl is not a good predictor of the target.

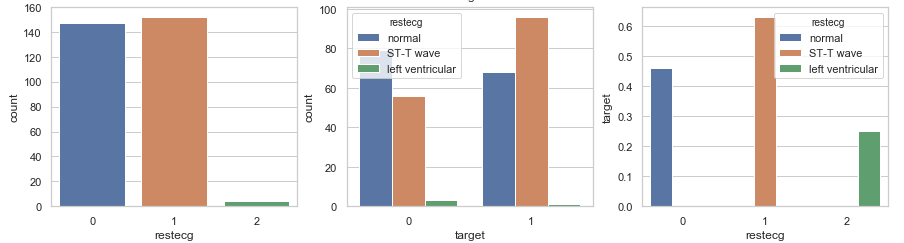


Figure 5: Effect of resting electro cardio graphic result on target

From Figure 5-3rd Part, it is observed that ST-T wave abnormality has highest impact on the target followed by normal resting electro graphic result.

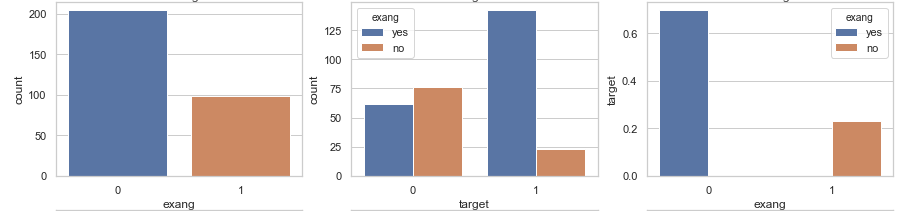


Figure 6: Effect of Exercise induced angina on Heart disease

Figure 6 show the exercise induced angina has strong impact on target.

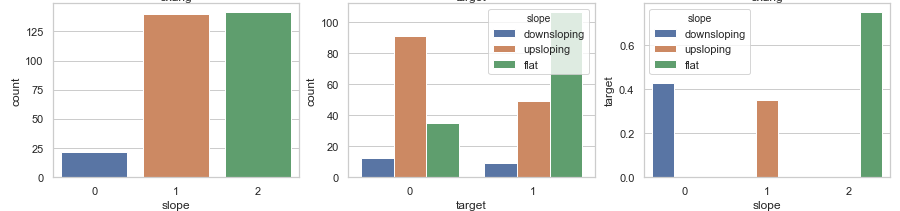


Figure 7: Probability of heart disease based on different slope of peak exercise

Figure 7 shows the probability of having heart disease due to the different slope of peak exercise ST segment. There is a high decrease in the probability of having a heart disease where respondents have an up sloping peak exercise slope.

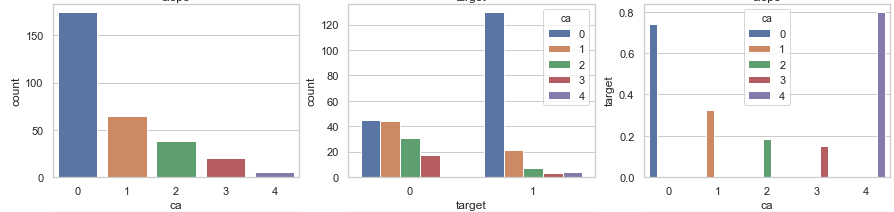


Figure 8: Effect of number of major vessels on Heart disease.

Figure 8 depicts the effect of number of major vessel on heart disease. It is observed that ca variable has a strong impact on the target variable.

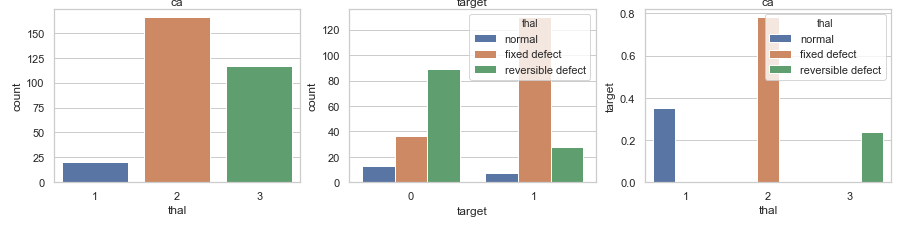


Figure 9: Effect of thal on Heart Disease.

Figure 9 shows the thal may have strong impact on target variable. The probability of heart disease increases with fixed defect thal.

### Continuous Variables

The variation of heart disease among age group is shown in Figure 10

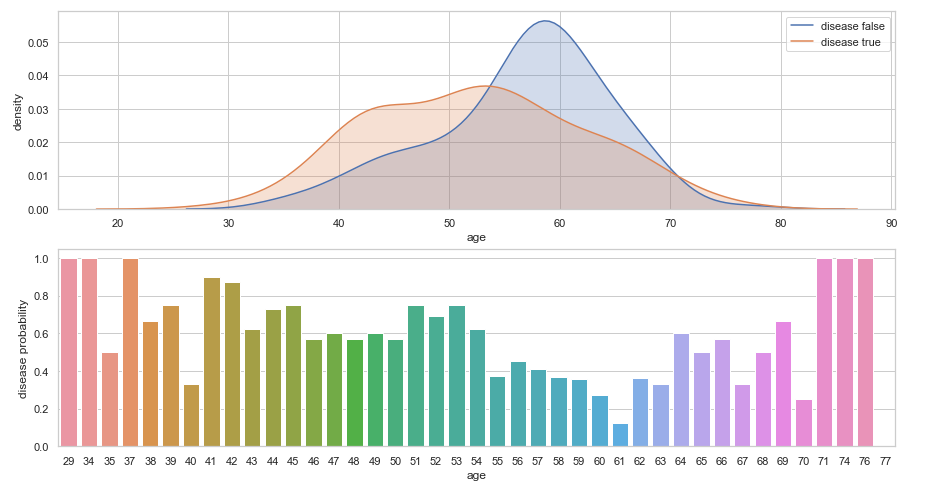


Figure 10: Variation of Heart diseases with age group

As shown in Figure 10, age has a symmetric and unimodal distribution for both target outcomes. The data seems to be centered around 40 to 70 years old. The data also shows that respondents with a heart disease peaked between 50 and 60 years, with highest density.

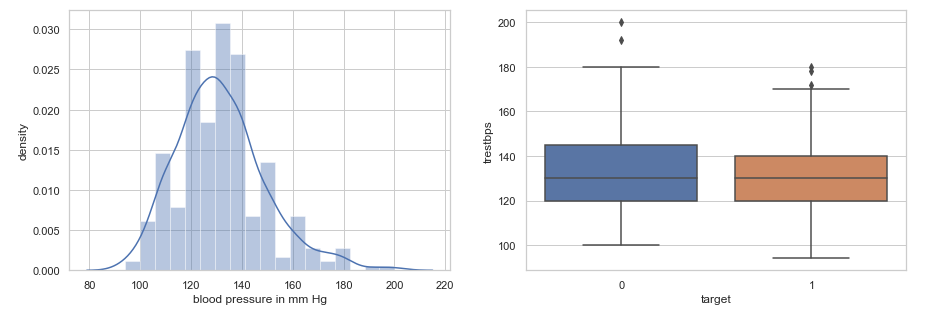


Figure 11: Effect of blood pressure on target variable

Figure 11 shows the blood pressure distribution is symmetric and unimodal. The distribution of serum cholesterol (Figure 12) and maximum heart rate (Figure 13) also show symmetric and unimodal. Therefore, the probability of heart disease due to blood pressure, serum cholesterol and heart rate seems to be weak.

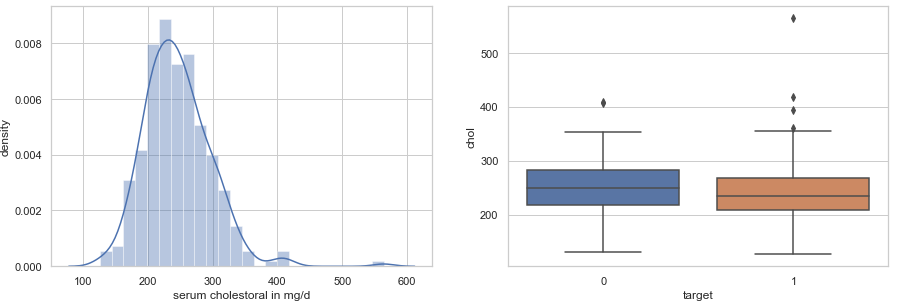


Figure 12: Effect of serum cholesterol on target variable

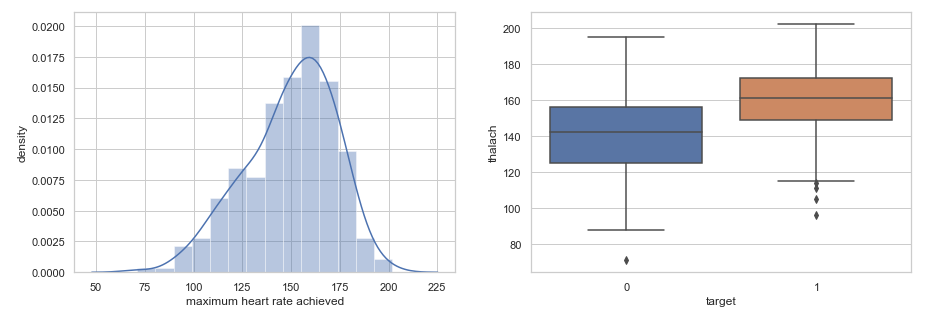


Figure 13: Effect of maximum heart rate on heart disease

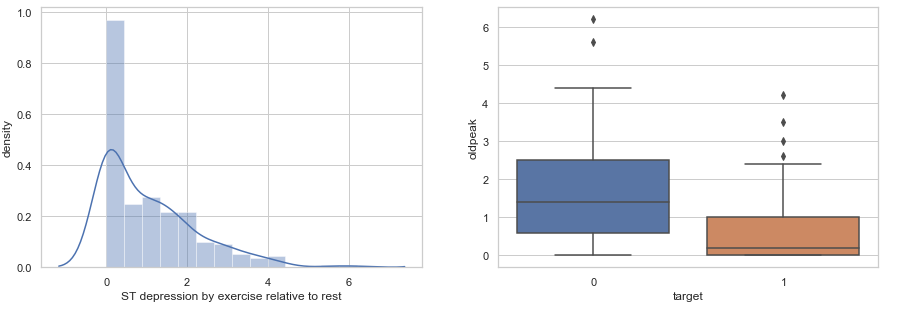


Figure 14: Impact of ST depression on heart disease

Figure 14 shows the oldpeak distribution is asymmetric. Therefore, ST depression induced by exercise relative to rest has great impact on heart disease.

From the above discussion, it is concluded that followings are the risk factors associated with heart diseases.

1. Sex
2. cp
3. restecg
4. exang
5. slope
6. ca
7. thal
8. oldpeak

The effect of above factors on heart disease is quantitavily estimated through different models are discussed in the next section.

# Methodology

The dataset discussed in section 2 is splited into training (70%) and testing (30%) data for validation. The results are predicted using the following four different methods with different tuning parameters and compared to consider the best model to find out the important parameters predicting the heart-disease.

1. Logistic Regression
2. Random Forest classifier
3. Decision tree classifier
4. Linear Discriminant analysis

The results are stored in variable AUC and the accuracy of each model is also calculated. The confusion matrixes are also calculated for different model. The accuracy of different models is summarized in

Table 1.

|  |  |
| --- | --- |
| Model | Accuracy |
| Logistic Regression | 82.41758241758241 |
| Random Forest | 82.41758241758241 |
| Decision tree classifier | 72.52747252747253 |
| Linear Discriminant analysis | 80.21978021978022 |

Table 1: Accuracy predicted by different Model

From the above **Table 1** we can consider Random-Forest is a better model compared to others in predicting the heart-disease. For comparison purpose I have tried the same data in Liner Discriminant Analysis (LDA) also. Here are the results for LDA, followed by Random-Forest.

The accuracy predicted by training and test data is 81.13 and 80.22 respectively.

Using “**SequentialFeatureSelector”** From **mlxtend** I have tried to reduce the features to important feature only. With the linear discriminant analysis, it is identified that followings are the variables affect the heart disease more as compared to others.

1. cp
2. thalach
3. oldpeak
4. slope
5. thal

Again using only the above 5 features (reducing the features) we have calculated the confusion matrix (predicted by linear discriminant analysis) & the result are shown below along with ROC.

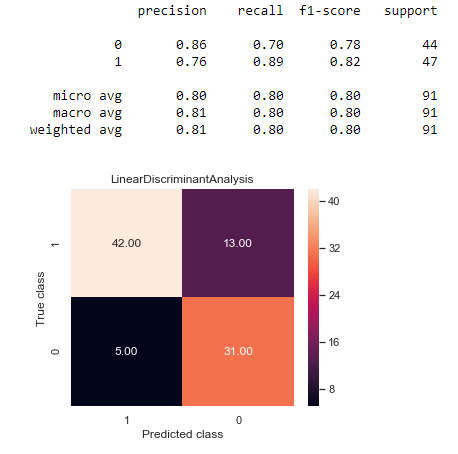


Figure 15: Confusion Matrix, linear discriminant analysis

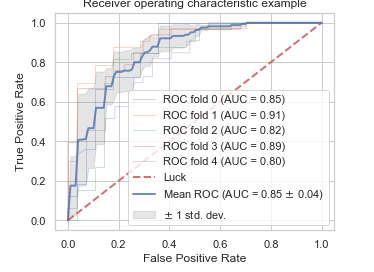


Figure 16: ROC, linear discriminant analysis

Now Using the Random Forest (as it has the highest accuracy in comparison Table 1), the accuracy score is: 100.0 & 83.55 for the train and test data respectively.

Again using the same “**SequentialFeatureSelector”** From **mlxtend** I have tried to reduce the features to important feature only. With the Random forest analysis, it is identified that followings are the variables affect the heart disease most.

1. cp
2. ca
3. thal

With the best selected features, the accuracy predicted by training and test data is 86.32 and 81.32 respectively. Again using the above 3 features only (reducing the features), I have calculated the confusion matrix (predicted by random forest) & the result are shown below along with ROC.

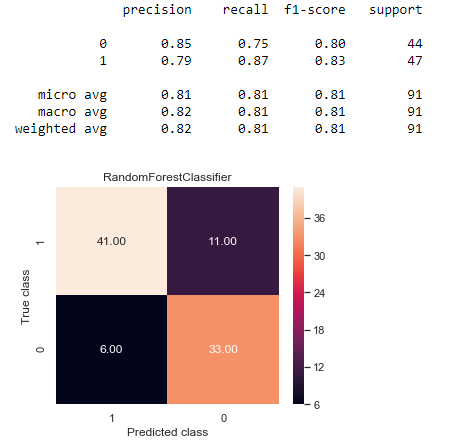


Figure 17: Confusion Matrix, Random Forest

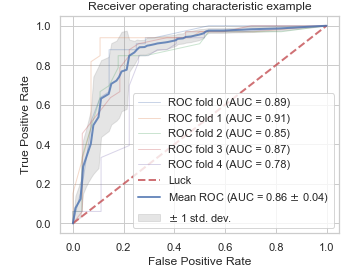


Figure 18: ROC, Random Forest

# 4. Results

The dataset was composed of 207 (68.31%) men and 96(31.68%) women. A total of 165 (54.45%) ones were patients with heart disease and 138(45.54%) ones were patients without heart disease. The relative importance of each variable in estimating the model is associated with the importance of each feature in making a prediction, and it does not relate to model accuracy. **Random-Forest and LDA** algorithms were used to compare the accuracy. As shown in Figure 15, features with great impact on target (presence of Heart Disease) were listed in order of variable importance. Even though each algorithm has assigned different relative weight to features, the most significant factor influencing heart-disease (target) obtained from Random Forest are **ca, cp and thal** It should be noted that end-user can use these information to analyze the strengths and weaknesses of medical attributes associated with heart-disease.

With Random forest model, the importance of all the features/variables for heart failure prediction are identified & are as follows:

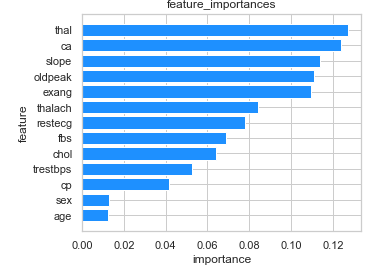


Figure 19: Importance of various features on predicting Heart-Disease

# 5. Conclusion

Most of the time, clinical decisions are made by physician’s experience, while all physicians are not experienced. Hence, systems with diagnosis support would be a guideline for clinical decision making.As a result, we believe that general physicians can use this information to perform medical screening just on important attributes instead of doing that on all attributes of patients who are likely to be diagnosed with heart diseases. This will reduce wasting time, medical expenses, administrative costs, and diagnosis time.

Therefore.

In this project, I have tried to analyze all the variables or features provided in the heart-disease data set for efficient prediction of the disease based on. Applying various visualization techniques, it can be concluded that out of the 13 variables (14th one is the target), Following variables or risk-factors are closely associated with the presence of a heart disease (**Sex, cp, restecg, exang, slope, ca, thal, oldpeak**).

Again on further analyzing algorithmic model (based on Random-Forest) we can conclude that the features such as: **1.cp (chest pain type), 2.ca (number of major vessels) and 3.thal** are measure contributor in predicting the heart-disease.

As a result, we believe that general physicians can use this information to perform medical screening just on above important attributes instead of doing that on all attributes of patients who are likely to be diagnosed with heart diseases.