Data Analysis Report

By- Nadimul Hasan

Website- nadimul-hasan.netlify.app

Heart Disease Prediction

Overview:

Dataset collected from kaggle, which includes 12 columns and 918 rows. Aim is to predict whether a person has a heart condition or not, based on 6 variables(Age, Sex, RestingBP, Chest pain type, Cholesterol level, MaxHR).

Cleaning data:

Dataset is loaded in python using Pandas library. Missing values are checked and system returns no missing values. Apart from out prediction variables, few of them are renamed, all other columns are dropped.

Visualization:

Using the seaborn library Kdeplots and countplots are plotted.

Encoding:

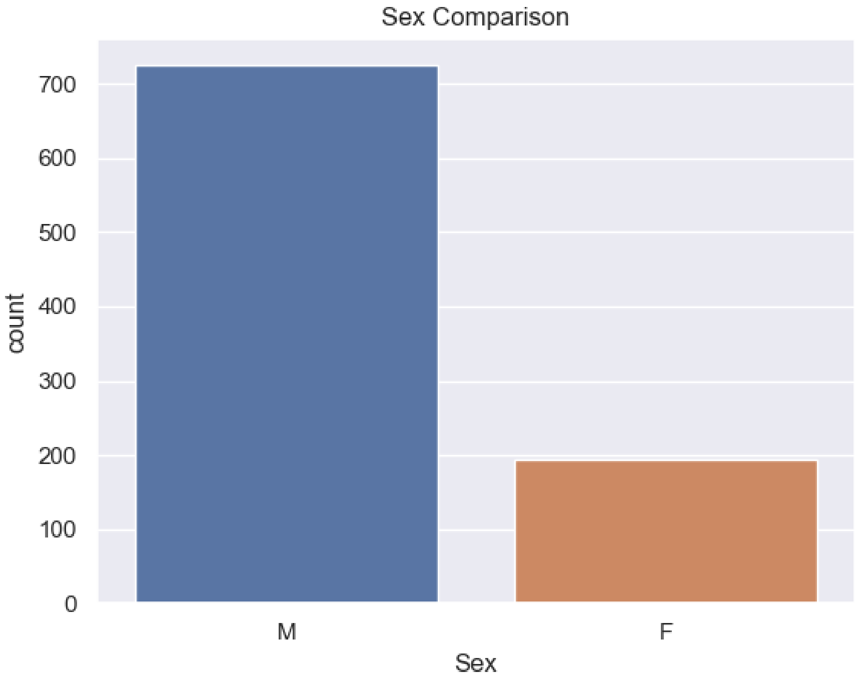
Using Sklearn, sex and chest pain type columns are encoded and converted to numbers to include in the model.

Model Creation:

From Sklearn multiple parts are imported. Data is trained and fitted in the model and tested. Next the accuracy of data is tested, using accuracy_score.

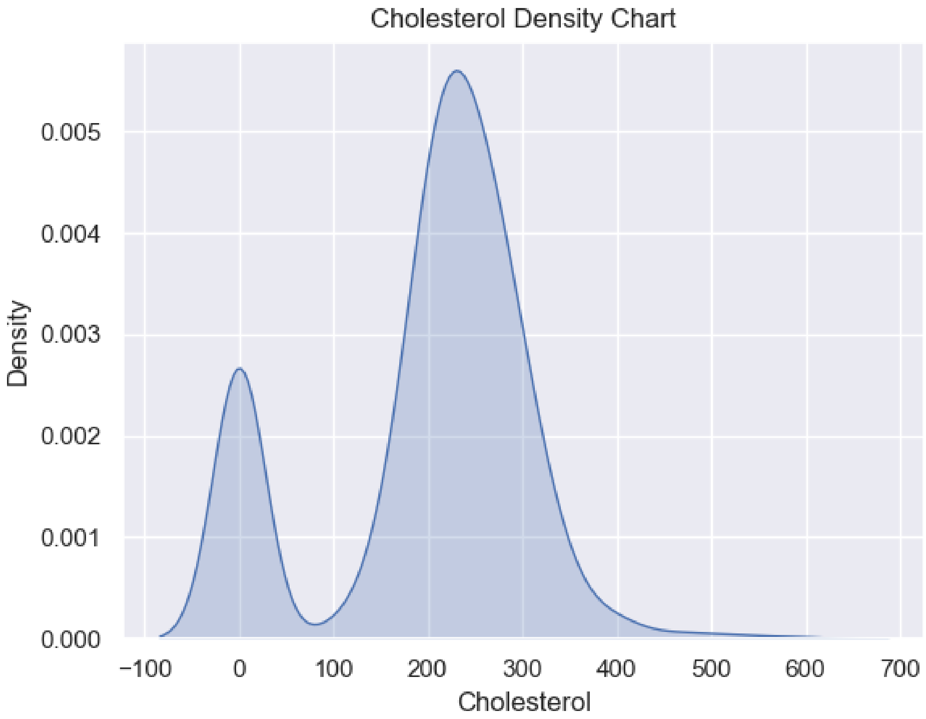
Proceedings

The main target is to make a model which accurately predicts if a person has a heart condition or not.

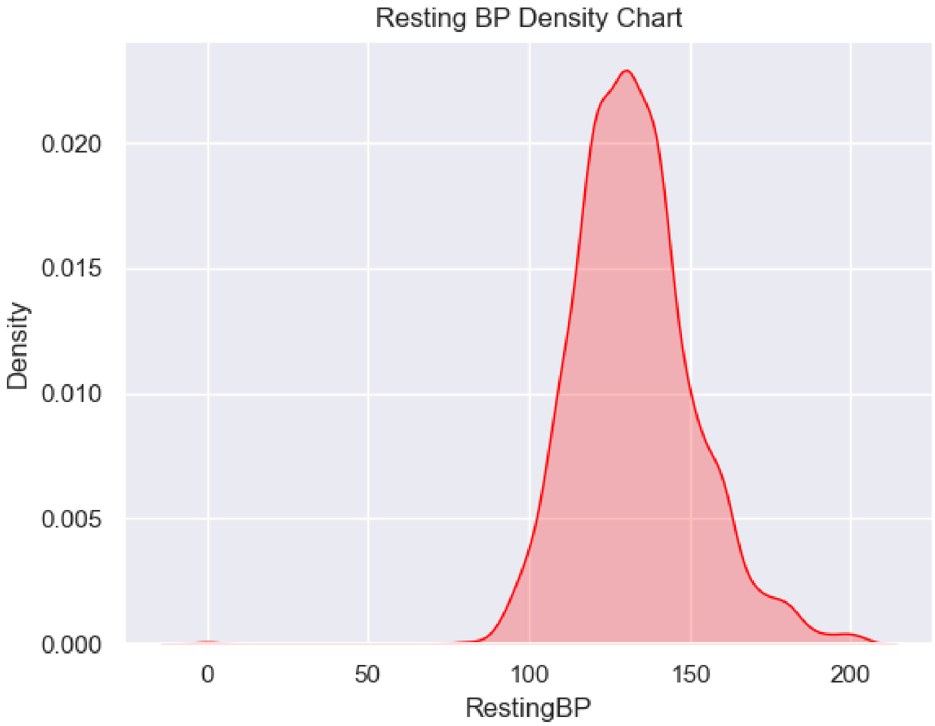


From the above **Countplot** we can see that majority of people in the survey consist of males(over 700), while women cover less than one-fourth of the total number.

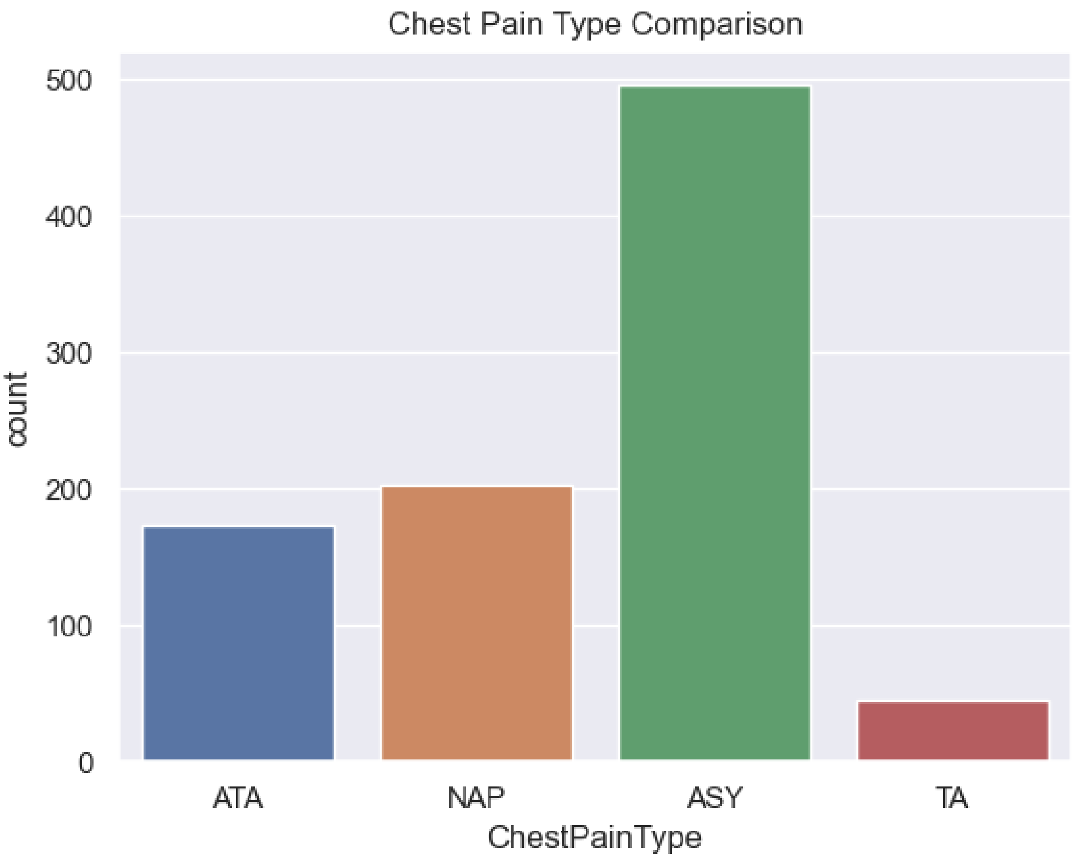
Coming onto the second plot, a **Kdeplot** is plotted which shows that the most of the people have a cholesterol level between 220-280 and around 0. The two large peaks signifies that either people have negligible cholesterol or a high level of fat in their body. There are few cases where the cholesterol is almost 600.



The next graph consist of peoples resting Blood Pressure. As we can clearly observe that most of the patients have a mean BP of around 130. The highest BP of a patient is over 200, which is quite abnormal.



Lastly, the different chest pain types are compared. Almost 500(over 50%) people have **ASY** heart condition. The next most common type is **NAP**(200 people) followed by **ATA**(19.6%) and most rare is **TA**(less than 30 people).



After analyzing all charts, the model is prepared. Firstly, all the columns containing strings are changed to numeric values using **Label Encoder**, because **Sklearn** models cannot process only numbers.

The Chest pain type column is encoded as follows:

- 'ASY'=0
- 'ATA'=1
- 'NAP'=2
- 'TA'=3

The Sex column is changed to:

- Male=1
- Female=0

After this, ‘X’ and ‘Y’ are set to be trained and tested, by importing train_test_split from sklearn.model_selection. Next the data is divided into 70-30 to be trained and tested. Model is selected as **DecisionTreeClassifie**r with criterion set to ‘Entropy’(to increase accuracy). Then, model is fitted with x_train and y_train and tested wih x_test. Lastly, accuracy is measured using accuracy_score and y_test which gives a max score of **78.26%**.

In my opinion, the accuracy can be further improved with a much larger dataset and by reducing the number of variables. Another way could be to include more females and people with age over 20 years.

Python file can be found on my personal website.

To use custom variable values, input numbers in the following order-

[[Age, RestingBP, Cholesterol, Pain Type, Sex, MaxHR]]

By- Nadimul Hasan

Website- nadimul-hasan.netlify.app