**Heart Disease Dataset Analysis** Sankalp Gupta

**Contents**

**1 Introduction**

**1.1 Problem Statement . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 2**

**1.2 Data . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 2**

**2 Methodology. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3**

**2.1 EDA(Exploratory Data Analysis) . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3**

**2.1.1 Bivariate Analysis . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .3**

**2.1.2 Multivariate Analysis . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 5**

**2.2 Data pre-processing. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 6**

**2.2.1 Missing Value Analysis. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .. .6**

**2.2.2 Outlier Analysis. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 7**

**2.2.3 Feature scaling. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .8**

**2.3 Modeling . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 9**

**2.3.1 Model Selection . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 9**

**2.3.2 Logistic Regression . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 9**

**2.3.3 Decision Tree Classifier. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 9**

**2.3.4 KNeighbours classifier . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 9**

**2.3.5 SVC . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 9**

**3 Conclusion . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .10**

**3.1 Model Evaluation . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 10**

**3.1.1 Accuracy . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .10**

**3.1.2 Cross Validation Score . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 10**

**Appendix A – Python Code . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 11**

**1. Introduction:**

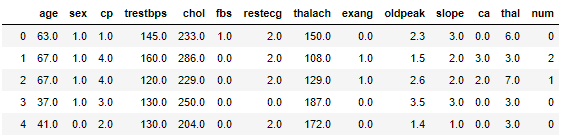
**1.1 Problem Statement:**

This database contains 14 attributes . In particular, the Cleveland database is the only one that has been used by ML researchers to this date. The "goal" field refers to the presence of heart disease in the patient. It is integer valued from 0 (no presence) to 4. Experiments with the Cleveland database have concentrated on simply attempting to distinguish presence (values 1,2,3,4) from absence (value 0).

**1.1 Data:**

We are supposed to build a Classification model as our target variable ‘num’ is categorical, which gives us the data whether a person is having a heart disease or not.

Given below is a sample of the data set, we are using to predict the num.

**Table 1.1: Cleveland data set**

Below are the predictor variables, will help us to predict total number of counts.

**Table 1.3: Predictor variables**

|  |  |
| --- | --- |
| Sr.no. | Variables |
| 1 | age |
| 2 | sex |
| 3 | cp |
| 4 | trestbps |
| 5 | chol |
| 6 | fbs |
| 7 | restecg |
| 8 | thalach |
| 9 | exang |
| 10 | oldpeak |
| 11 | slope |
| 12 | ca |
| 13 | thal |
| 14 | num |

**2. Methodology:**

**2.1 EDA(Exploratory Data Analysis):**

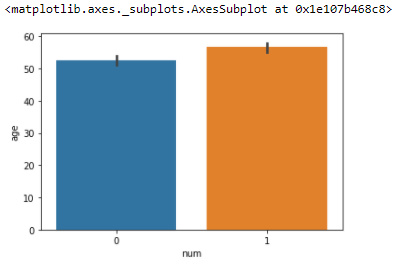
Any predictive modeling requires that we look at the data before we start modeling. However, in data mining terms *looking at data* refers to so much more than just looking. Looking at data refers to exploring the data, cleaning the data as well as visualizing the data through graphs and plots. This is often called as **Exploratory Data Analysis**. To start this process we will first try and look at all the distributions of the Numeric variables.

**2.1.1 Bivariate Analysis:**

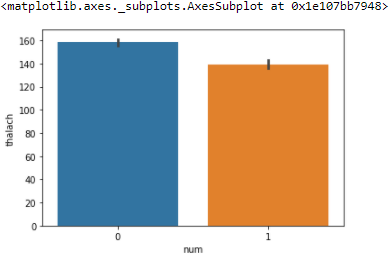
Bivariate analysis is done within features to analyse the contribution of the dataset features with the target variable.

Figures below describes the same.

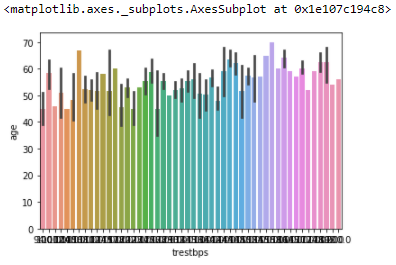
**Fig 2.5 Barplot of age with the target variable**



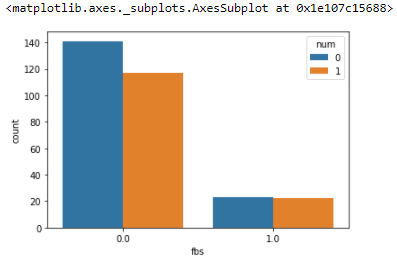
**Fig 2.6 Barplot analysis for thalach with num**



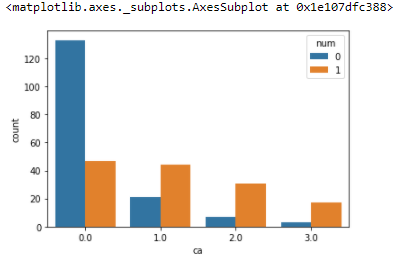
**Fig 2.7 Barplot analysis of age with trestbps**



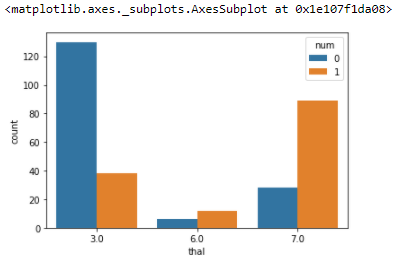
**Fig 2.8 Countplot of fbs with num**



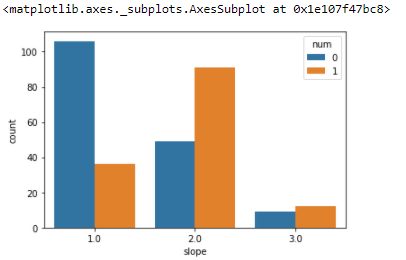
**Fig 2.9 Countplot of ca with num**



**Fig 2.10 Countplot of thal with num**



**Fig 2.11 Countplot of Slope with num**

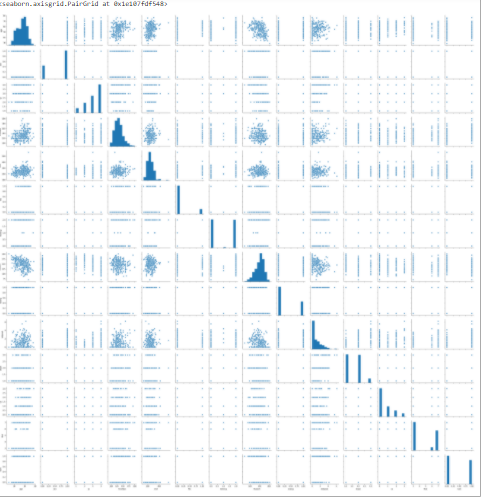


**2.1.2 Multivariate Analysis:**

We are doing bi-variate analysis over here, trying to get relationship between each continuous feature amongst each other. Specially, relationship with target variable ‘num’.

Below figure is pair-plot of all the features.

**Fig 2.12**



**2.2 Data Pre-Processing**

**2.2.1 Missing Value Analysis**

Missing values in data is a common real world problem, we face while analysing and fitting the model.

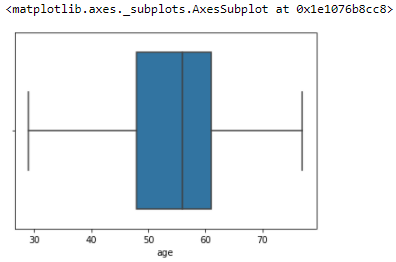
The table below shows us that we don’t have any missing value in our data.

|  |  |
| --- | --- |
| **Features** | **Counts** |
| age | 0 |
| sex | 0 |
| cp | 0 |
| trestbps | 0 |
| chol | 0 |
| fbs | 0 |
| restecg | 0 |
| thalach | 0 |
| exang | 0 |
| oldpeak | 0 |
| slope | 0 |
| ca | 0 |
| thal | 0 |
| num | 0 |

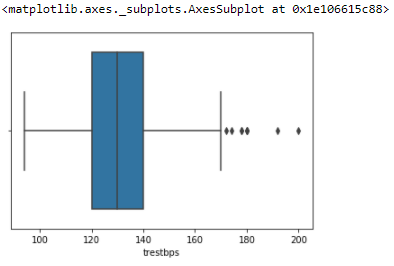
**2.2.2 Outlier Analysis:**

In Figures below I have plotted box plot to analyse the outliers in the features of the dataset.

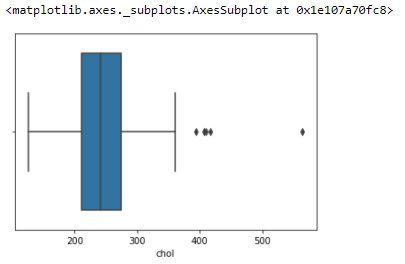
**Fig 2.1 Box plot analysis of age feature.**



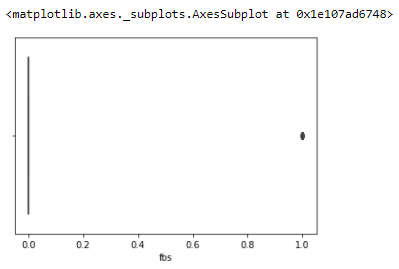
**Fig 2.2 Boxplot analysis of trestbps feature**



**Fig 2.3 Boxplot analysis of chol feature**



**Fig 2.4 Boxplot analysis of fbs feature**



**2.2.3 Feature Scaling**

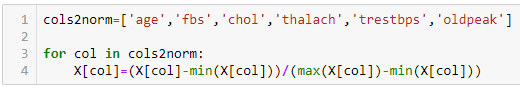
Feature scaling is one of the most important pre-processing technique. It majorly involves two techniques named as Normalization and standardization.

**Link to Random forest classifier based feature importance is given below, https://towardsdatascience.com/running-random-forests-inspect-the-feature-importance s-with-this-code-2b00dd72b92e**

It’s important to rescale features else it may lead to wrong predictions, especially in the case of regression problems.

Rescaling data between 0 and 1 is known as feature scaling. We need to normalize ‘age’, ’fbs’, ‘chol’, ‘thalach’, ‘trestbps’ and ‘oldpeak’ features.

**Code snippet:**



**2.3 Modeling**

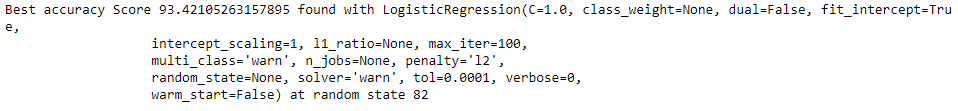
**2.3.1 Model Selection**

Since our target variable is categorical so we will go for classification methodology to go for model training.

We will start our model building from the most simplest to more complex. Therefore we use Logistic Regression at first.

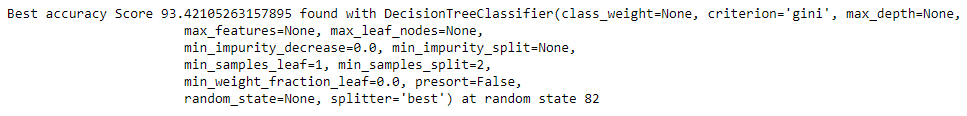
**2.3.2 Logistic Regression**

Logistic regression model gave 93% of accuracy at random state 82.



**2.3.3 Decision Tree Classifier**

Decision tree classifier gave us 93% of accuracy at random state 82.



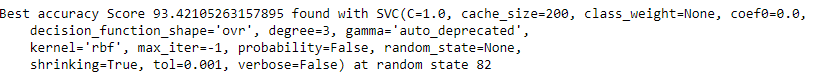
**2.3.4 KNeighbours Classifier**

KNeighbours classifier model gave us 93% accuracy at random state 82.



**2.3.5 SVC**

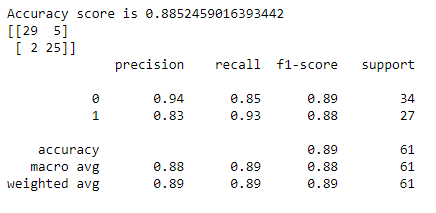
SVC model gave us 93% of the accuracy at random state 82.



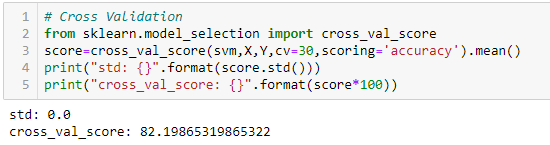
**3. Conclusion**

**3.1 Model valuation:**

**a. Accuracy:** Since all the model gave us the same accuracy I decided to go with SVC model as it was having an accuracy of 93% at random state 82. So checking the accuracy with the confusion matrix.



**b. Cross Validation Score:** After validating the cross val score it says that we are neither overfitting nor underfitting.



**Appendix A- Python Code**

