**Data Mining technical report**

Table of Contents

[1. Data Set Information 1](#_Toc39692615)

[2. Data understanding and preparation 2](#_Toc39692616)

[2.1 Overview of data and Business understanding 2](#_Toc39692617)

[2.2 Data Preparation 3](#_Toc39692618)

[2.3 Data Cleaning, transformation and filtering 4](#_Toc39692619)

[3. Analysis and understanding 9](#_Toc39692620)

[3.1 Patterns 10](#_Toc39692621)

[3.2 Interpretation of Results 11](#_Toc39692622)

# Data Set Information

Data mining is a popular extraction method for KDD (knowledge discovery) and helps

in analyzing the data, understand correlations and used for prediction. Data mining plays

a very vital role in health care industry. Heart disease is one of the main reasons of

human death in the world. We will choose Heart Disease dataset and analysis them and predict

a model following stages of CRISP-DM (Cross-industry standard process for data mining).

Dataset information URL : <https://archive.ics.uci.edu/ml/datasets/Heart%20Disease>

Dataset URL : <https://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/processed.cleveland.data>

The heart disease data set is freely available online and can be downloaded.

# Data understanding and preparation

## 2.1 Overview of data and Business understanding

# information of columns

# -----------------------------------

# age: age in years

# sex: sex (0: female , 1: female)

# cp: chest pain type

# Value 1: typical angina

# Value 2: atypical angina

# Value 3: non-anginal pain

# Value 4: asymptomatic

# chol : Serum cholestoral in (mg/dl)

# trestbps : resting blood pressure

# fbs : fasting blood sugar > 120 mg/dl (1: true, 0 : false)

# resting: resting electrocardiographic results

# Value 0: normal

# Value 1: having ST-T wave abnormality

# Value 2: showing probable or definite left ventricular hypertrophy

# **exang** : exercise induced angina (1: yes, 0: no)

# thalach : maximum heart rate (thal) achieved

# **oldpeak** : ST depression induced by exercise relative to rest

# **slope** : the slope of the peak exercise ST segment

# Value 1: upsloping

# Value 2: flat

# Value 3: down sloping

# **ca** : number of major vessels colored by flouroscopy (value : 0-3)

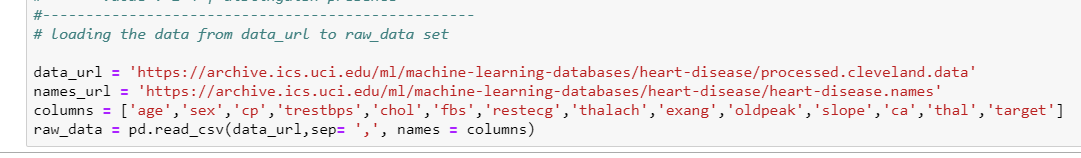
# **thal**: 3 = normal; 6 = fixed defect; 7 = reversable defect

# **target** : the presence of heart disease in the patient

# value: 0 | not present

# value: 1-4 | distinguish presence (present)

## 2.2 Data Preparation



Dataset is loaded into raw\_data from data\_url

From Figure1 (Dataset Info), it is observed that Dataset contains 303 rows and 13 input features as well as 1 (target) output feature.

It has THREE data types: float64(11), int64(1), object (2)

|  |  |  |
| --- | --- | --- |
| Datatype | count | Columns |
| float64 | 11 | Age,sex,cp,trestbps….slope |
| int64 | 1 | Target |
| object | 2 | Ca,thal |

Number of rows in dataset are 303 and number of columns are 14 (out of which 13 are features and one is target column , which we are going to predict) . However we done have any null variables in the dataset which is a good sign.

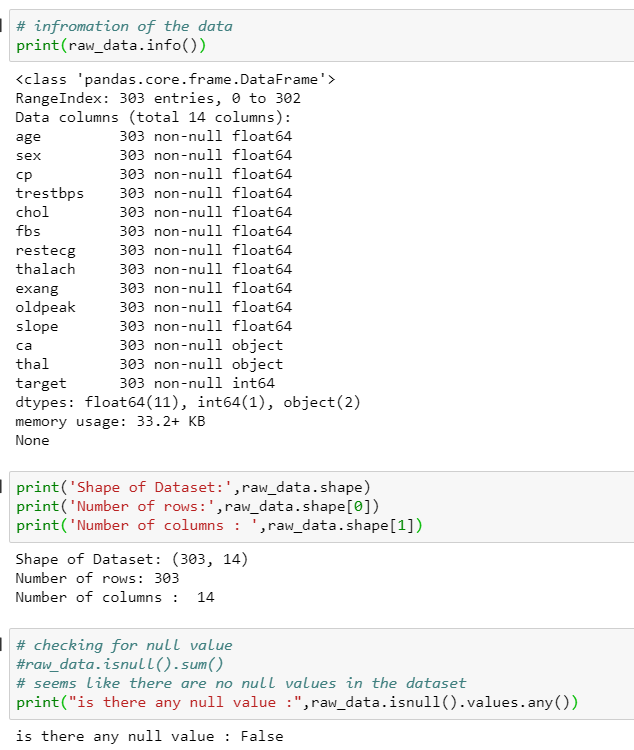


Figure 1 : Dataset Info

## 2.3 Data Cleaning, transformation and filtering

From (Figure 2: Data Description). It is observed that target value ranges from 0 (min)– 4 (max). In order to predict the we need to categorize the target value to either 0 (False) or 1 (True). We don’t need distinguish presence of heart disease. Only presence should be fine. Lets replace the target data (of value 1-4) as 1.

Convert to objects (**ca and thal**) to numeric and coercing when required (NaN to 0)

# lets change the heart disease types 1to 4 to 1

raw\_data.loc[raw\_data['target'] > 0 , 'target'] =1

raw\_data.ca = pd.to\_numeric(raw\_data.ca,errors='coerce').fillna(0)

raw\_data.thal = pd.to\_numeric(raw\_data.thal,errors='coerce').fillna(0)

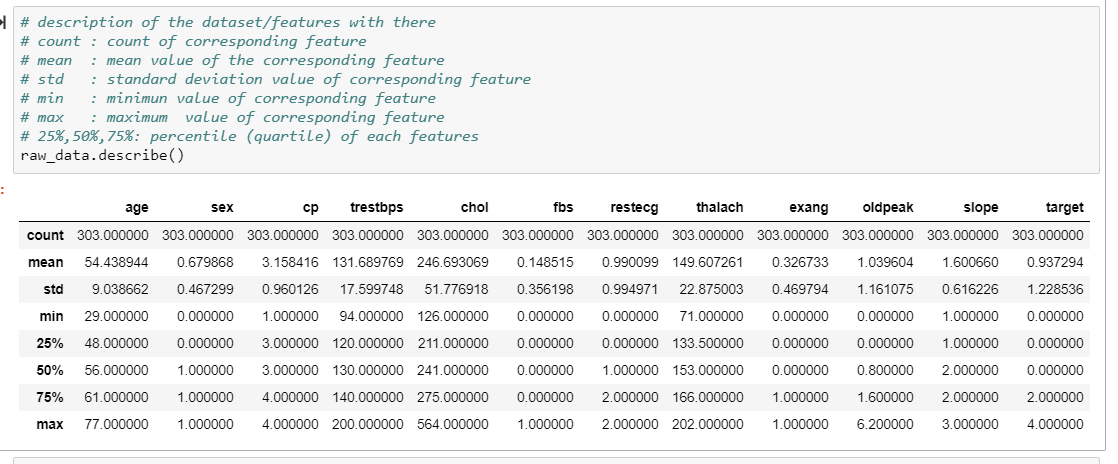
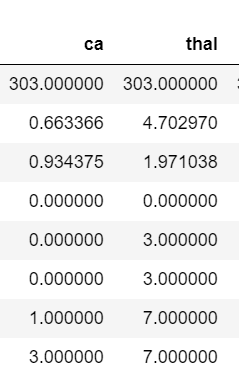
 

Figure 2 : data description

# lets categories the age into mu

# 1: young age (29-40) ,

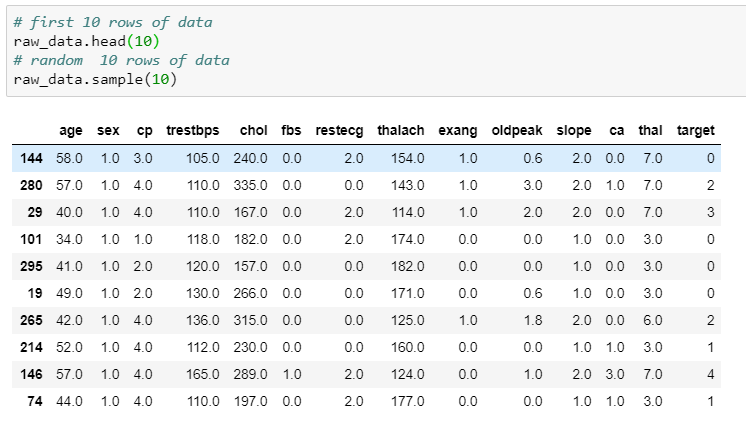
# 2: middle age (40-55) , and

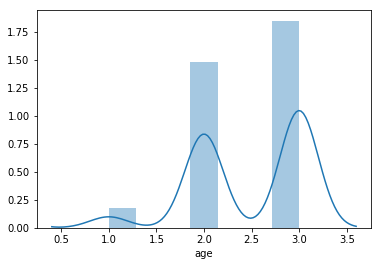
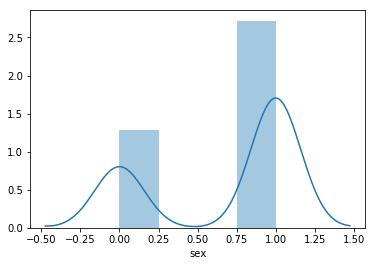
# 3: oldage(<50)

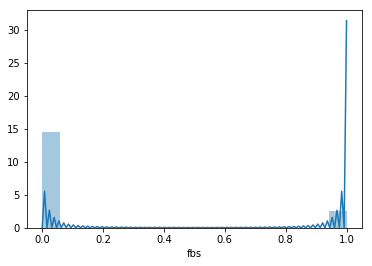
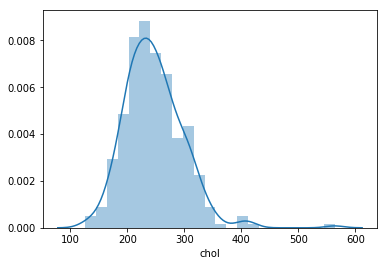
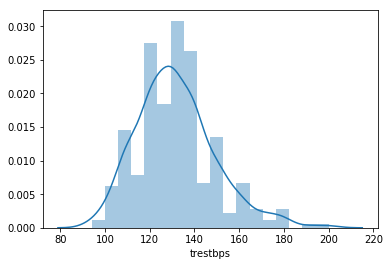
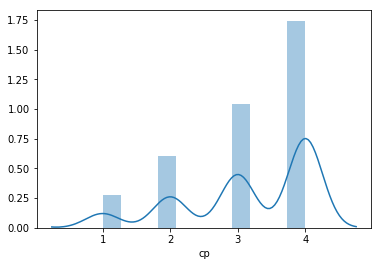
raw\_data.loc[(raw\_data.age >= 29) & (raw\_data.age < 40), 'age'] = 1

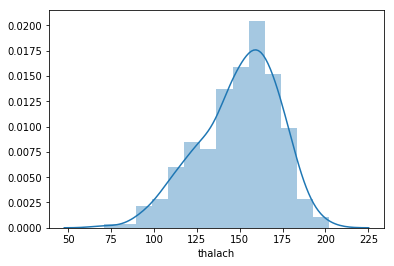
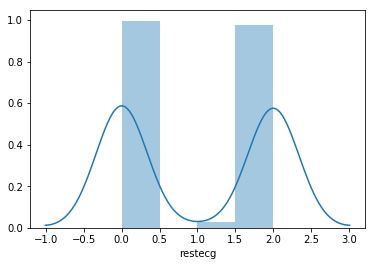
raw\_data.loc[(raw\_data.age >= 40) & (raw\_data.age < 55) , 'age'] = 2

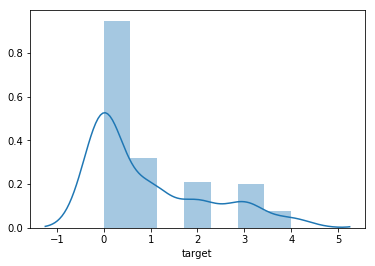
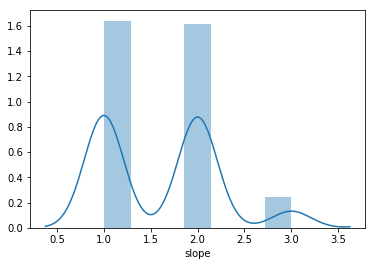
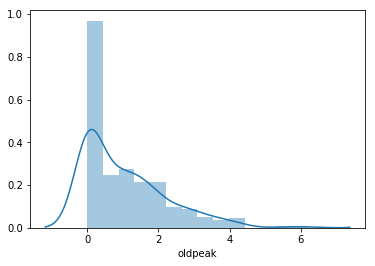
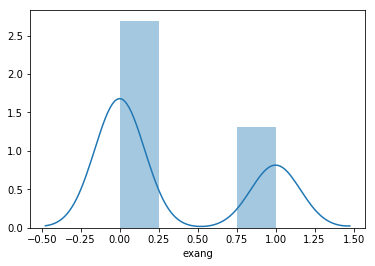
raw\_data.loc[(raw\_data.age >= 55) , 'age'] = 3









from the plot we can observe the distribution of all the features. We need to find out which are categorical features, and which are not

from the plot of ages (it’s clear that it’s a gaussian distribution) From the plot of **fbs** it’s clear that it’s a categorical value

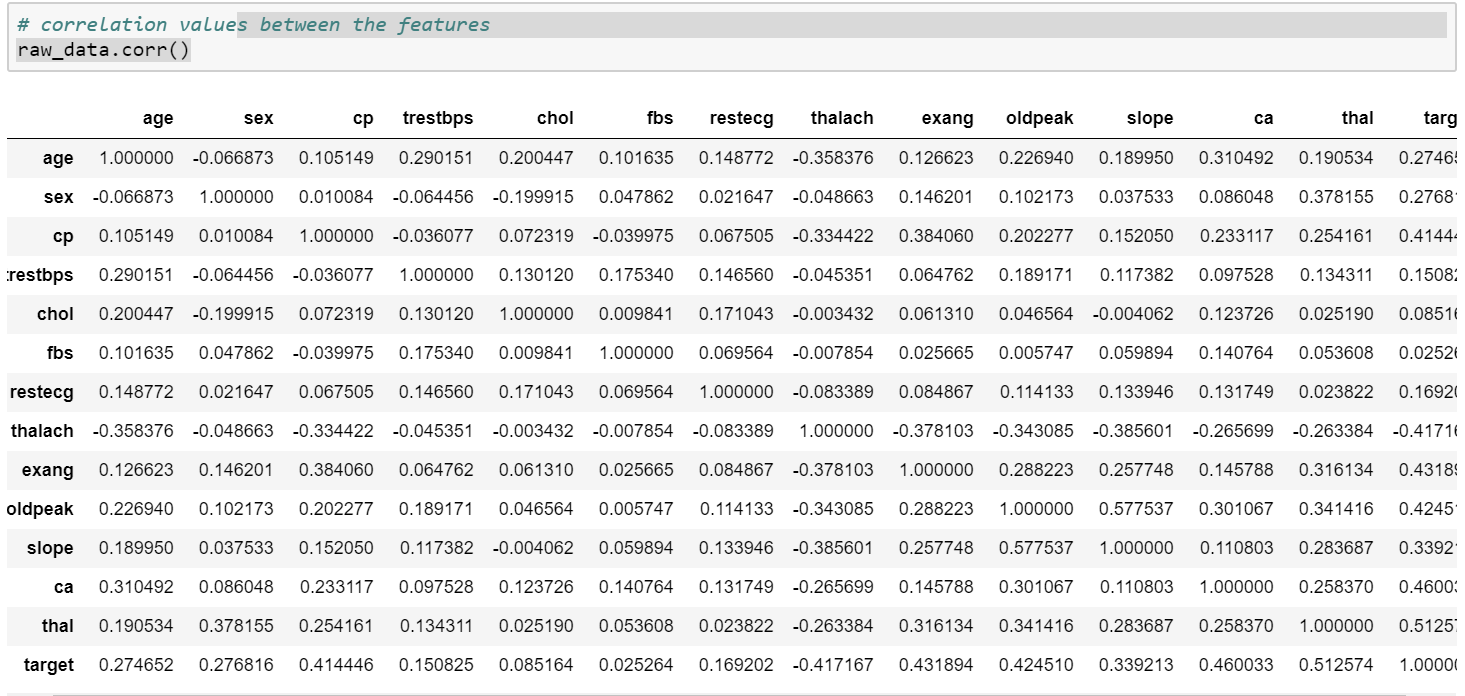


Figure 3 : Data correlation between features

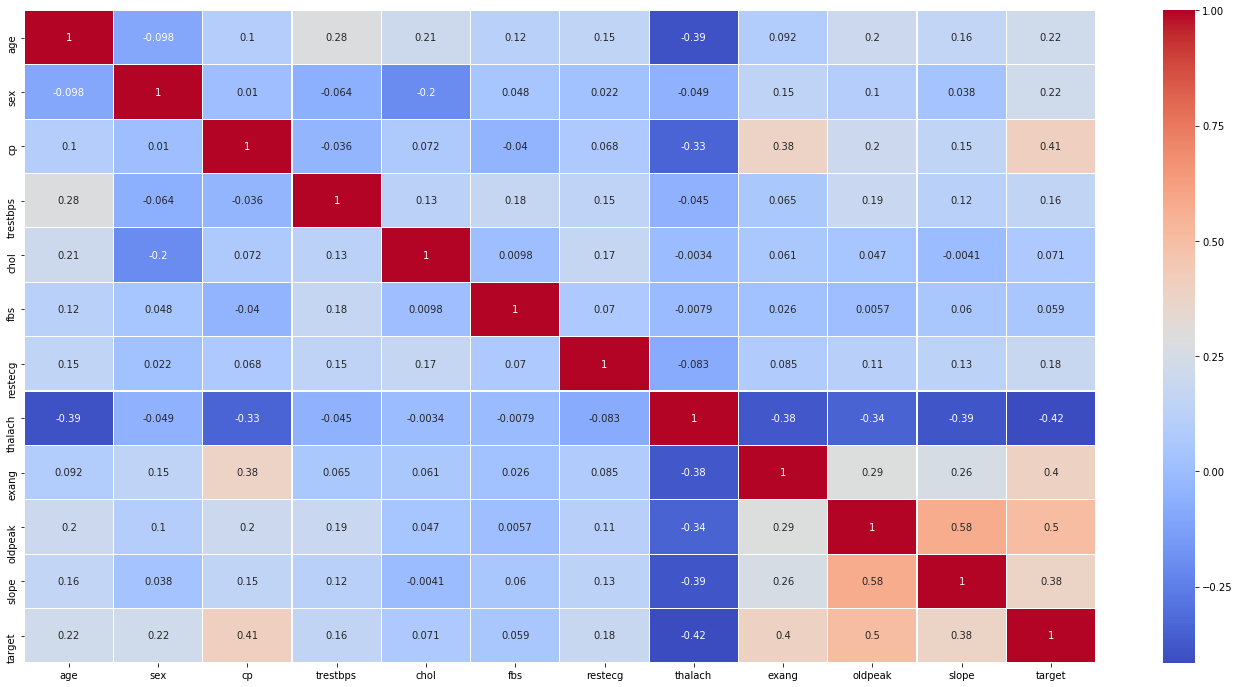


Figure 4 : Heat Map Plot of Dataset

From the Heat Map Plot (Figure 4: Heat Map Plot of Dataset) especially the bottom row (target row) it is observed that the features whose heat map color is Blue (less than 0.20) are loosely correlated with TARGET, these feature attributes need to be **dropped** from the dataset. Which are: **trestbps**(.16),chol(.071),fbs(0.059),restecg(.18),thalach(-0.42)

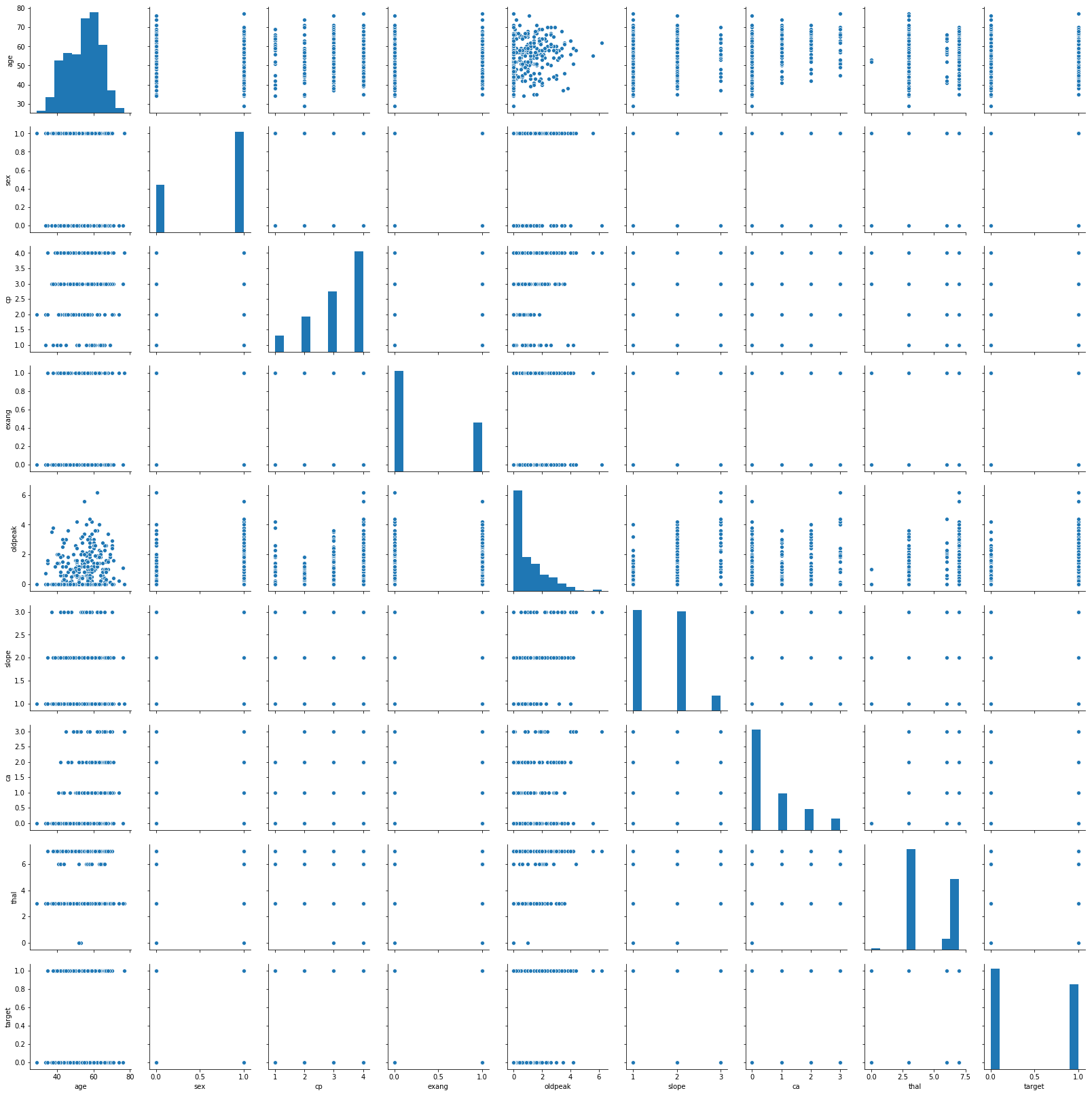
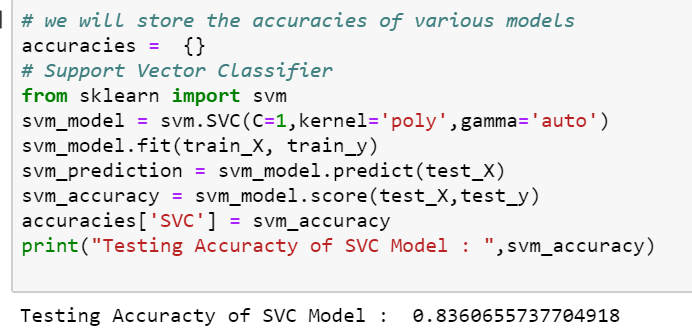


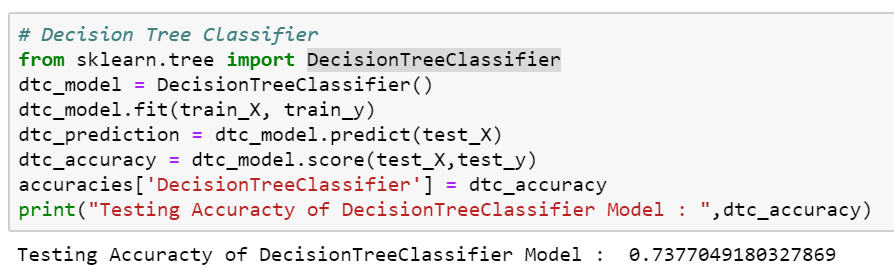
Figure 5 : Pair plot of all features

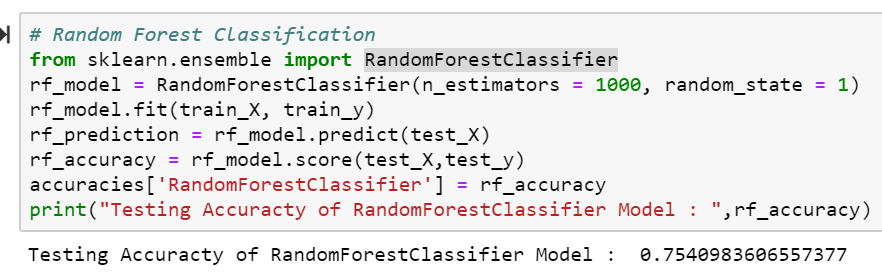
# Analysis and understanding

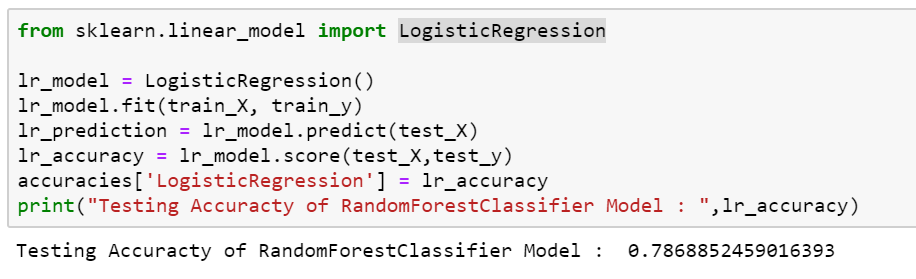
We will use 4 algorithms including **SVC (Support Vector Classifier), Random Forest Classifier, Logistic Regression and Decision Tree Classifier** to predict the heart disease.

## 3.1 Patterns









## 3.2 Interpretation of Results

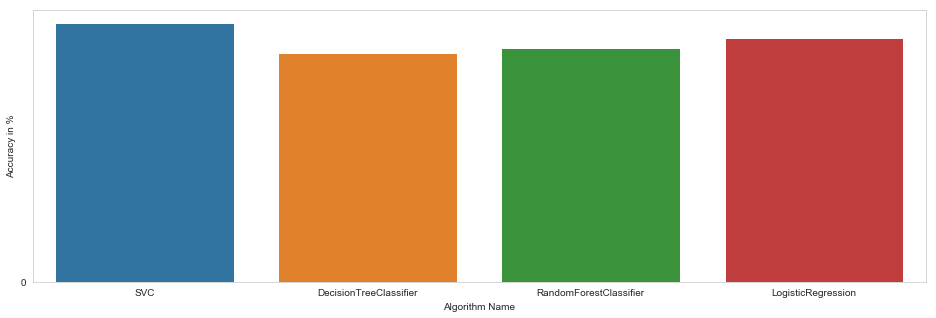


Figure 6 : Accuracy plot of algorithm

From the accuracy plot its is found that **SVM** algorithm gives more accurate **prediction (83% )** followed **Logistic Regression** (78%) , **Random Forest Classifier (75%) and Decision Tree Classifier** **(73%)**



Figure 7 : Confusion Matrix Plots

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | **Actual Class** | |
|  |  | Positive | Negative |
| **PREDICTED CLASS** | Positive | TP | FP |
| Negative | FN | TN |

|  |  |
| --- | --- |
|  |  |

where: TP = True Positive, TN = True Negative, FP = False Positive, , FN = False Negative.

**From the confusion matrix plot and Table,** it is observed that the **SVM model** or **SVC** (**Support Vector Classifier)** has better **True Positive (34)**, **True Negative** (17). as compared to other models. And it can be concluded that **SVC** is the winner in our analysis models.