Name : Gatreddi Deepthi

Class : CSE-2E

ID : S180264

**Heart disease prediction**

Topics:

* Problem statement
* Variables
* Preprocessing
* Visualization
* Removed variables
* Approach
* Accuracy values for each algorithm

Problem Statement:

Now-a-days many people are suffering from heart diseases but due to the unawareness of their disease they are losing their lives.

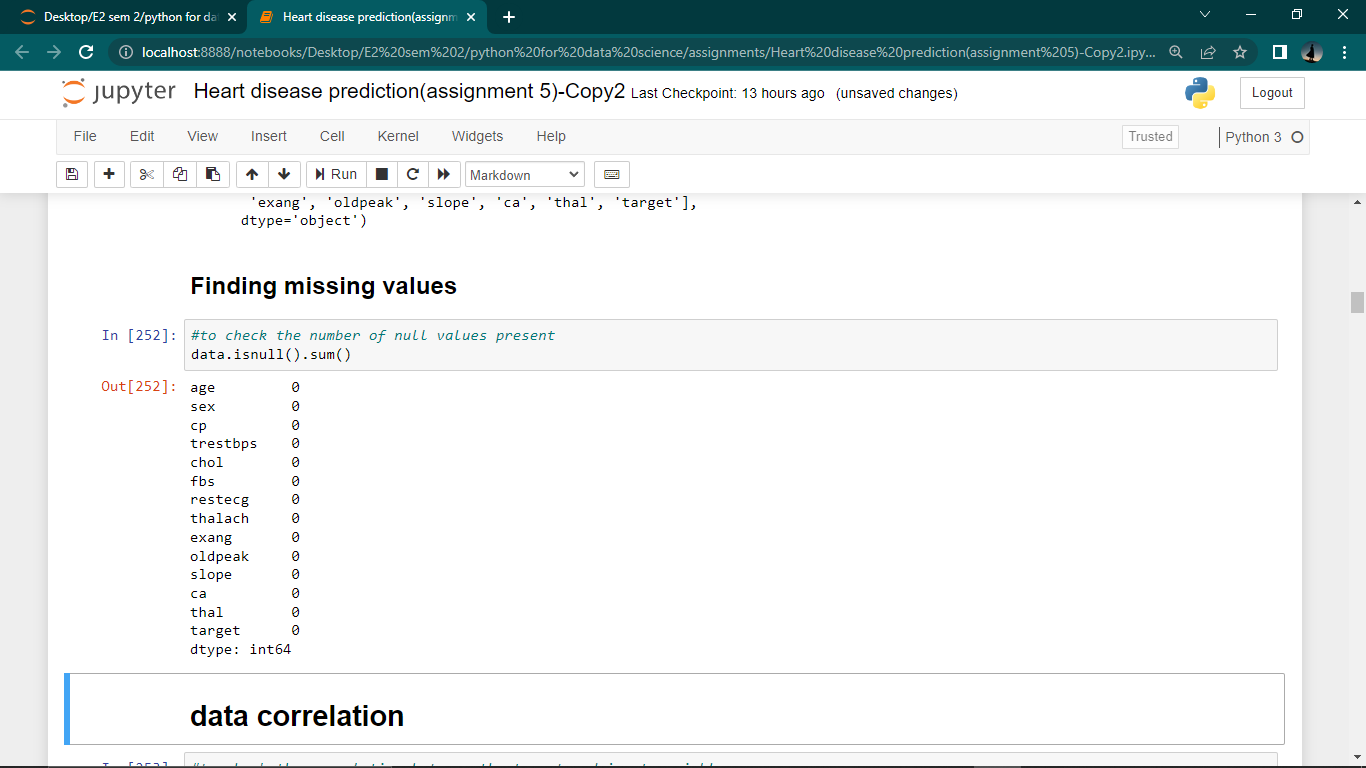
In order to save their lives we build a model based on the clinical parameters of the patient and finds out whether the person is suffering from heart disease or not.

Variables:

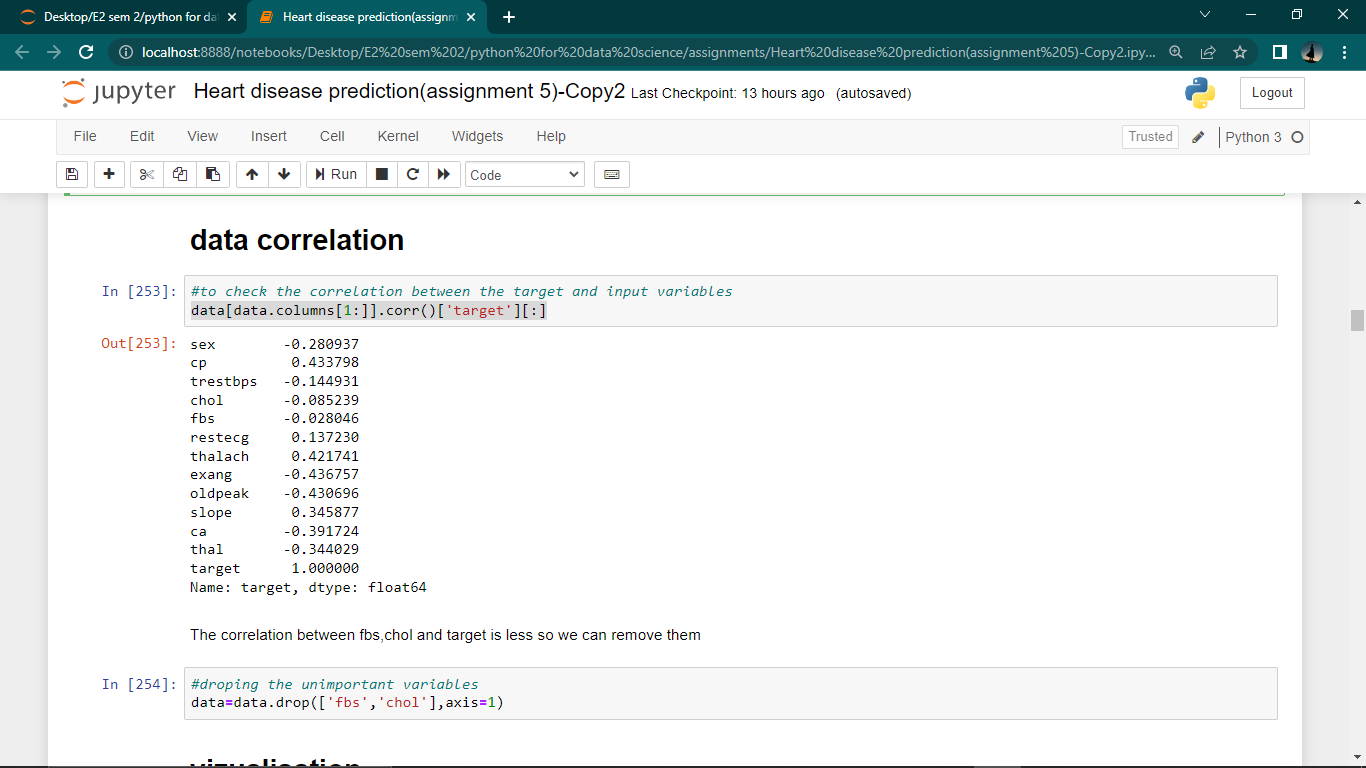
1. age: continuous
2. sex: categorical, 2 values {0: female, 1: male}
3. cp (chest pain type): categorical, 4 values {1: typical angina, 2: atypical angina, 3: non-angina, 4: asymptomatic angina}
4. restbp (resting blood pressure on admission to hospital): continuous (mmHg) //high bp high chances for heart disease
5. chol (serum cholesterol level): continuous (mg/dl) //high cloistral high danger
6. fbs (fasting blood sugar): categorical, 2 values {0: <= 120 mg/dl, 1: > 120 mg/dl} //low glucose or impaired fasting glucose are used in predicting heart disease
7. restecg (resting electrocardiography): categorical, 3 values abnormal value will make help to discover but very little {0: normal, 1: ST-T wave abnormality, 2: left ventricular hypertrophy}
8. thalach (maximum heart rate achieved): continuous //increase in the heart beat rate leads to heart disease
9. exang (exercise induced angina): categorical, 2 values {0: no, 1: yes} //if there is chest pain then it may leads to heart disease
10. oldpeak (ST depression induced by exercise relative to rest): continuous // it related to exang ,how much high the pain is
11. slope (slope of peak exercise ST segment): categorical, 3 values {1: upsloping, 2: flat, 3: downsloping} //plays an important role in determining heart disease with the help of heart beat
12. ca (number of major vessels colored by fluoroscopy): discrete (0,1,2,3)
13. thal: categorical, 3 values {3: normal, 6: fixed defect, 7: reversible defect}
14. target: (diagnosis of heart disease): categorical, 5 values {0: less than 50% narrowing in any major vessel, 1-4: more than 50% narrowing in 1-4 vessels} simply {0=not having heart disease,1=having heart disease}

preprocessing:

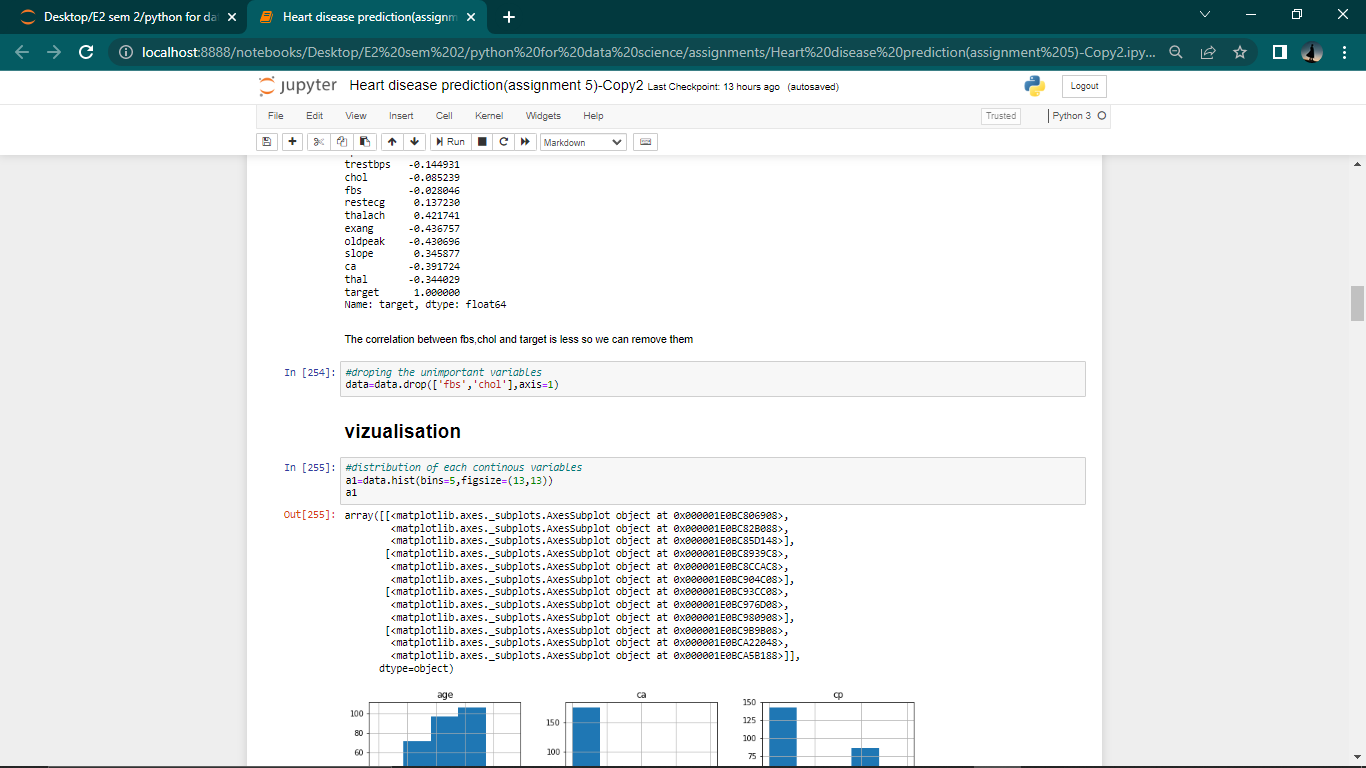
Checking for null values.

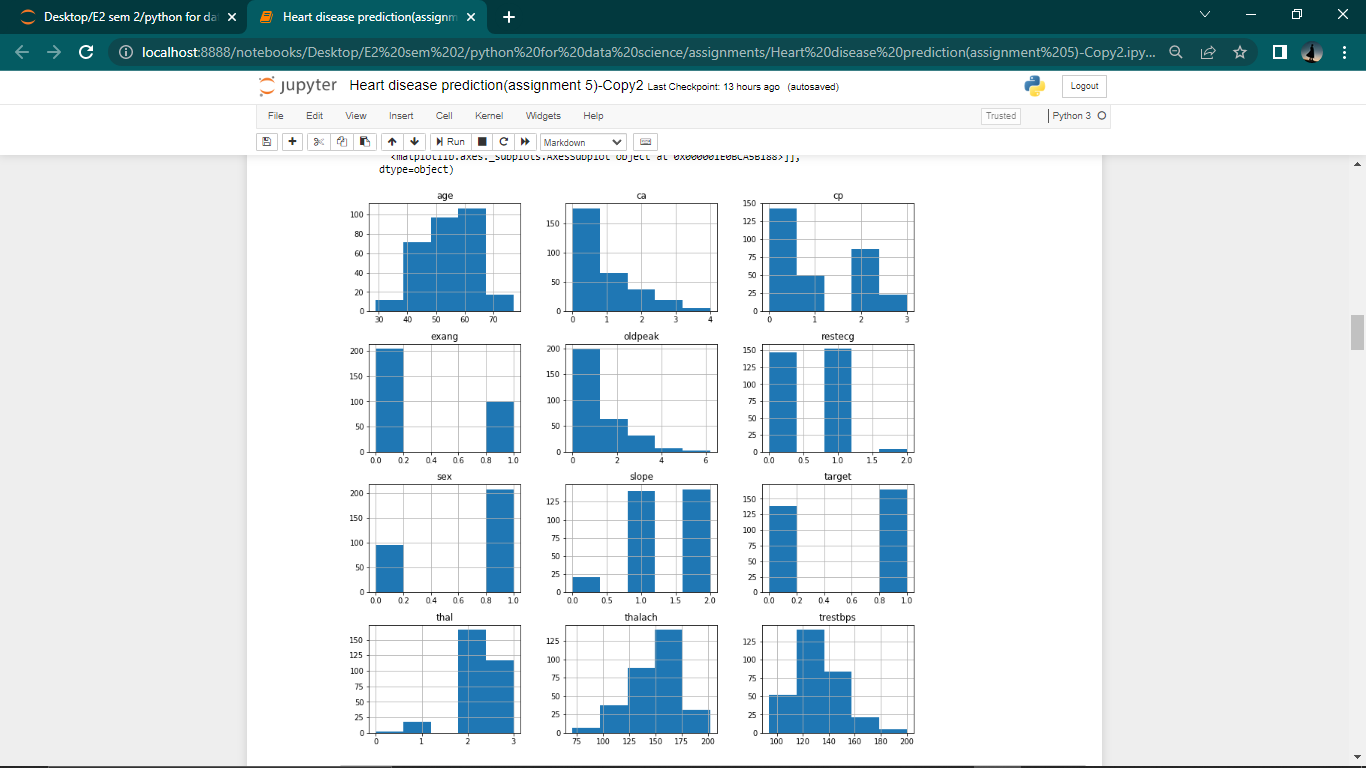


Checking the correlation between the output and input variables.

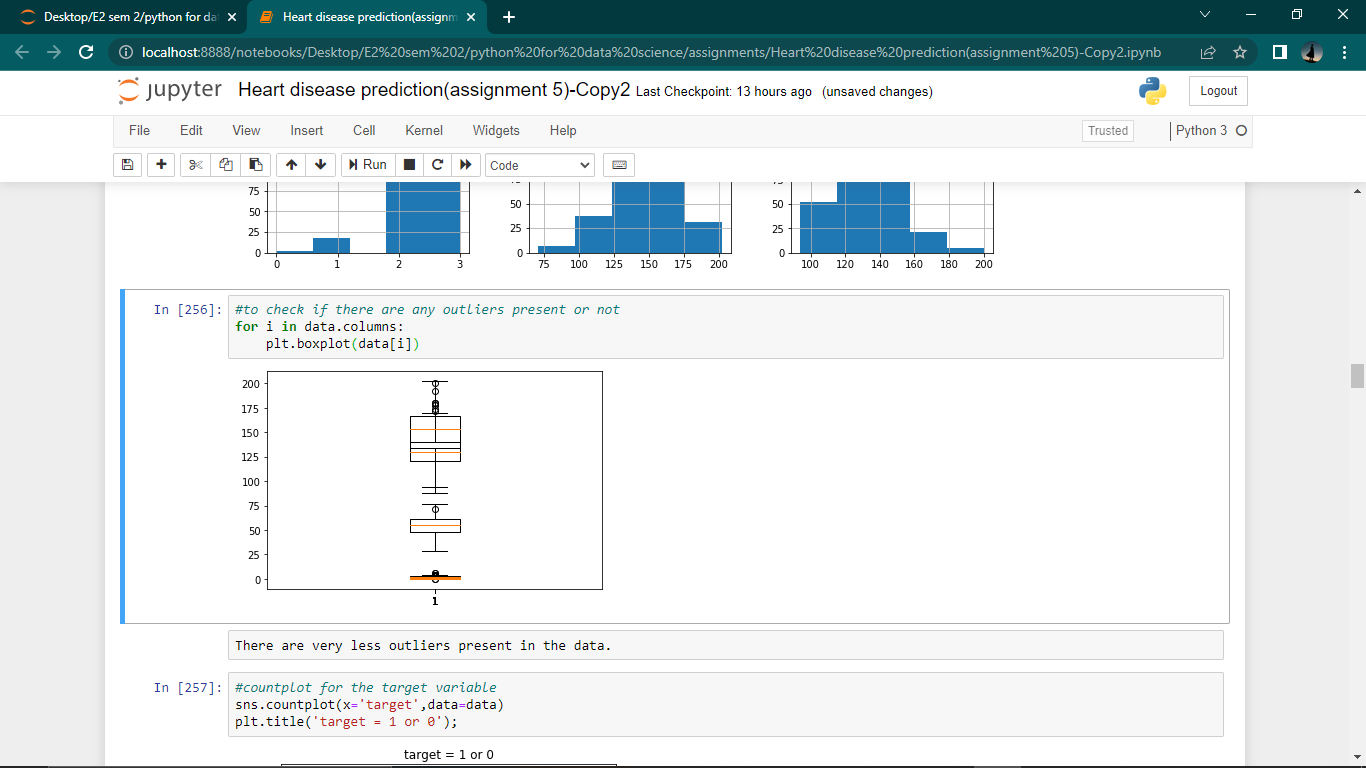


Visualization:

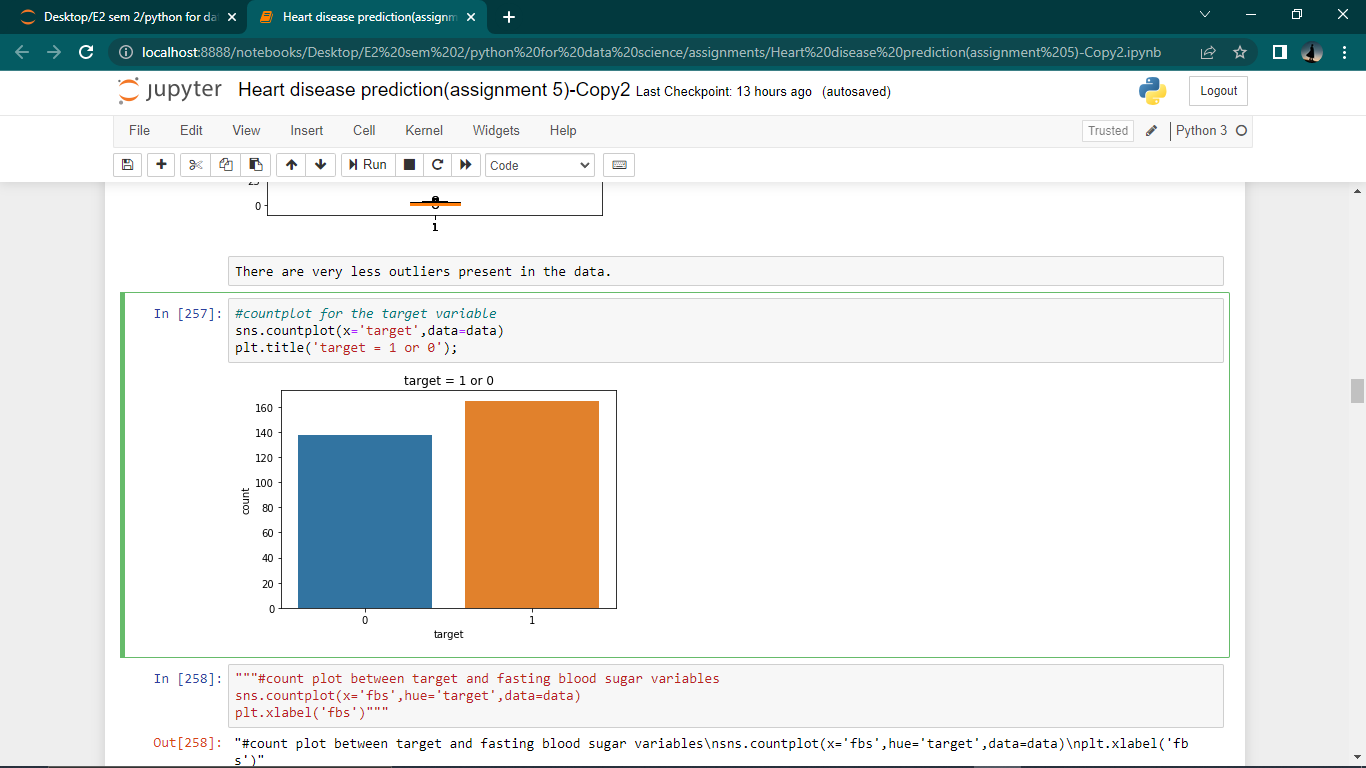




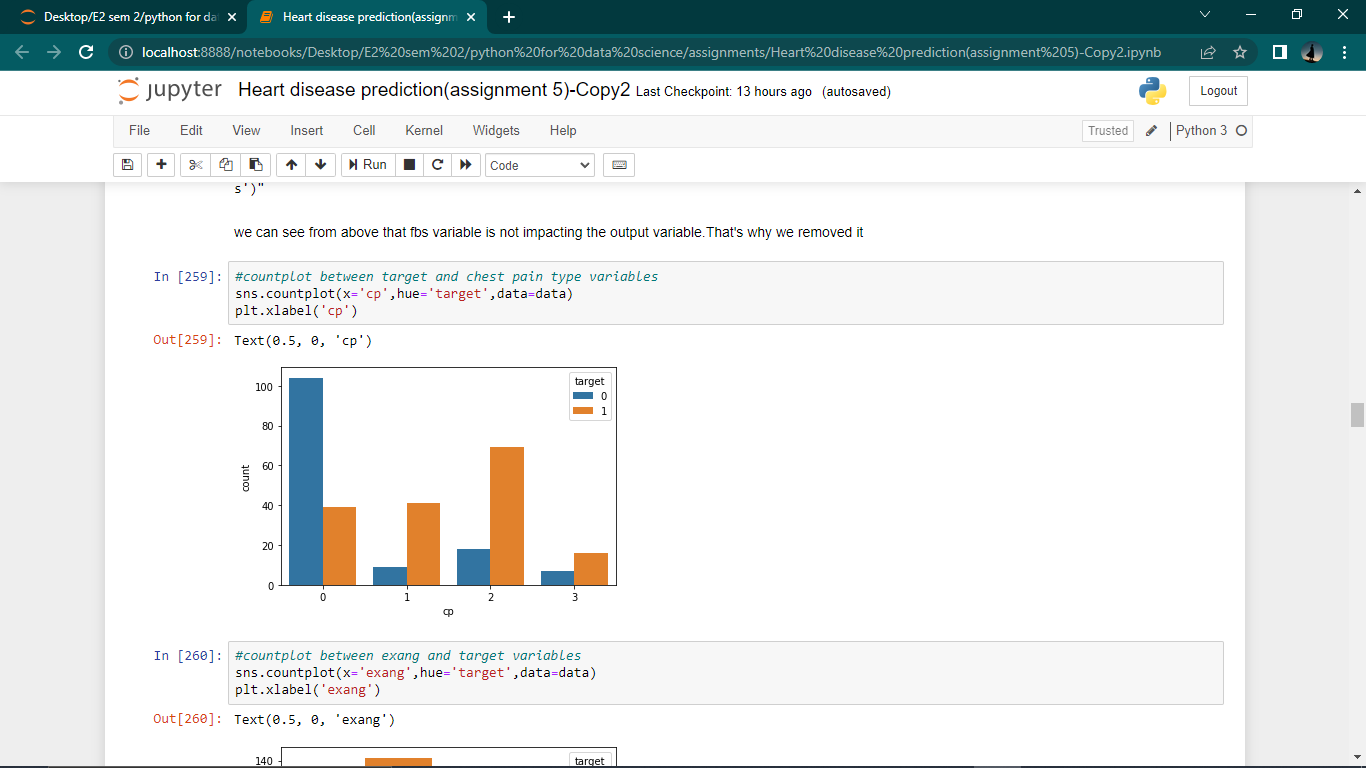
Box plot to check the outliers:



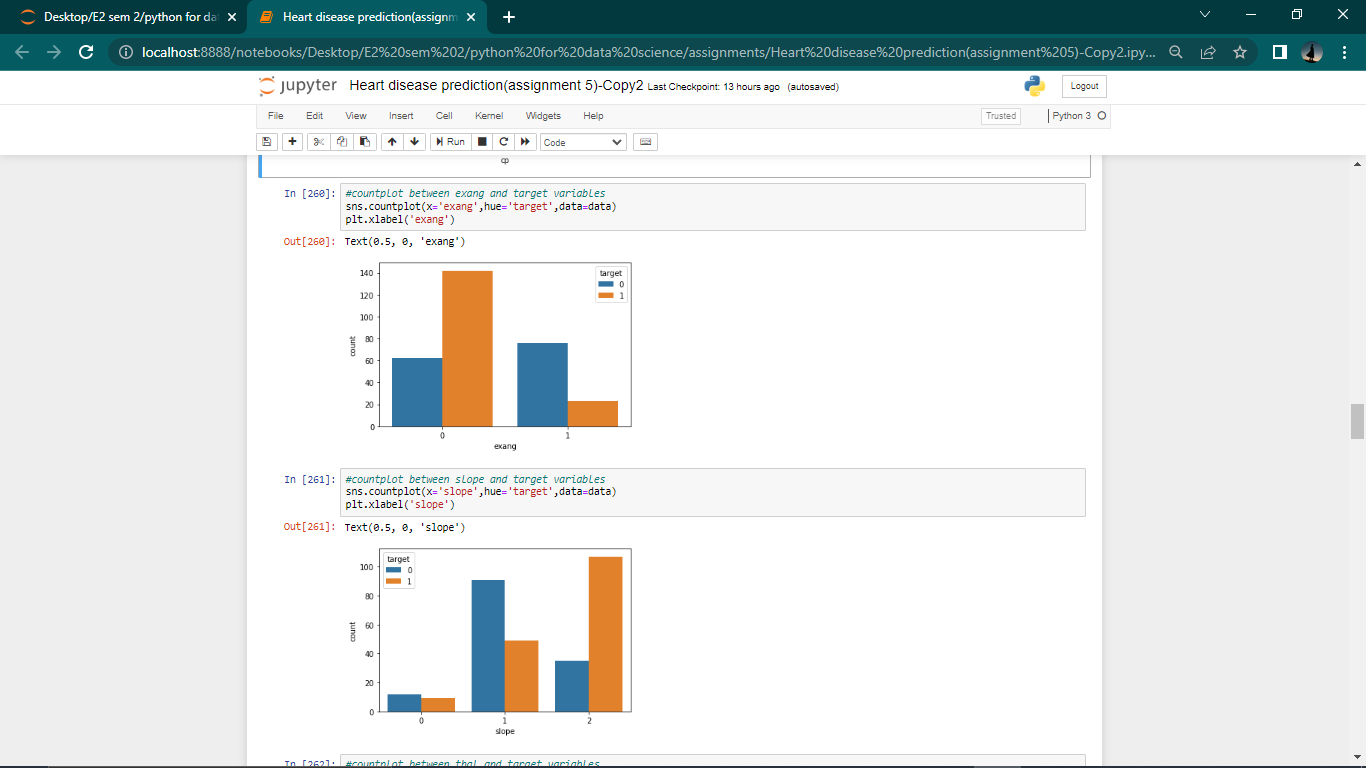
Analysing target variable:



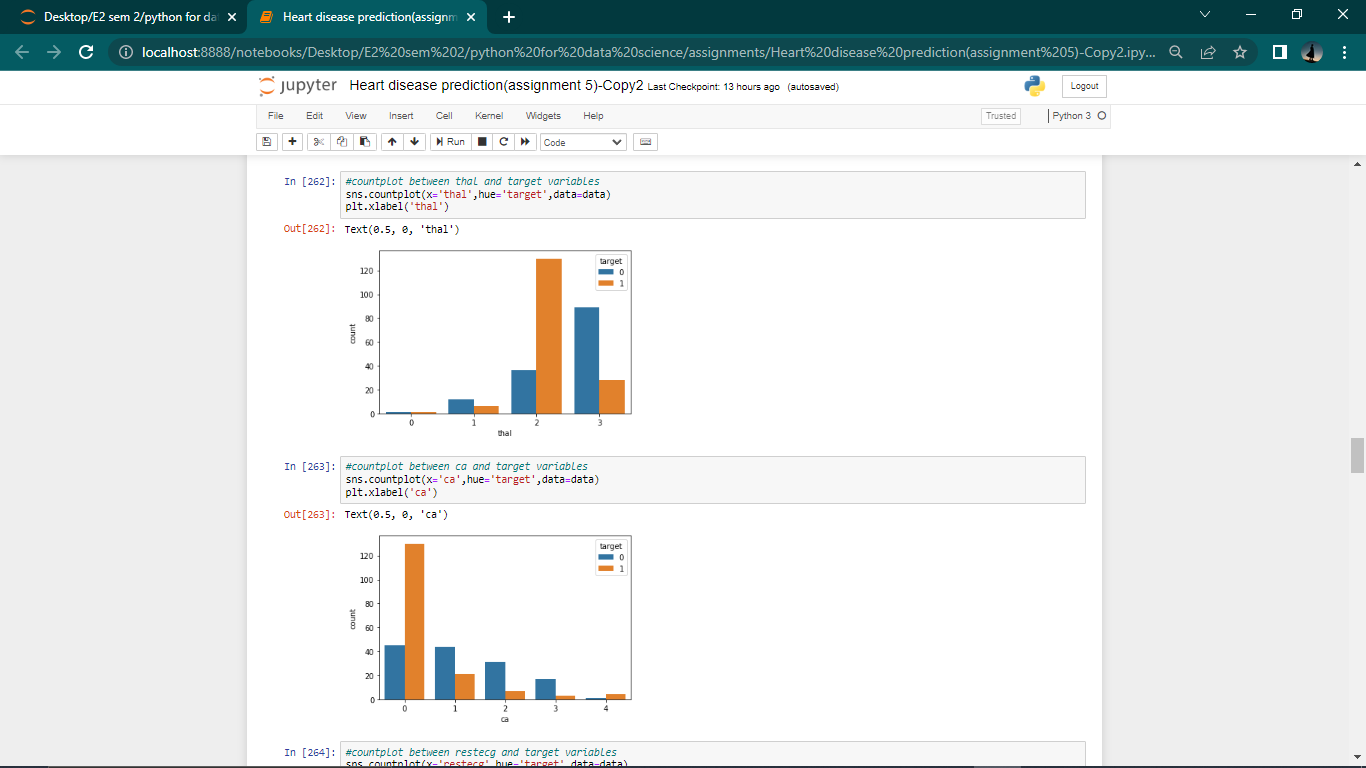
Analysing the test pain type variable:



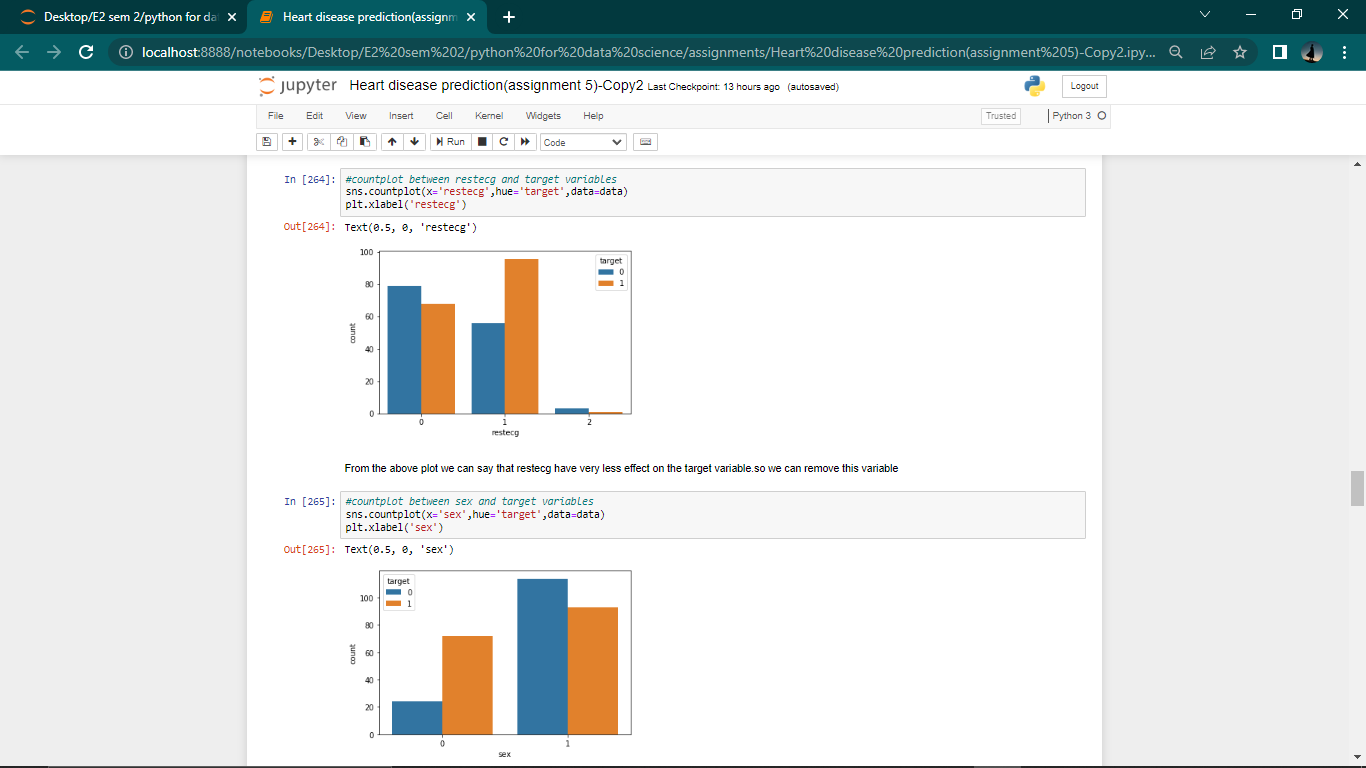
Analyzing the exang and slope variables:



Analyzing the thal and ca variables:



Analysing restecg ad sex variables:



Removed variables:

Fbs, chol and restecg variables are removed because there is less correlation between these input and output variables.

Approach:

Step 1: Data is imported and analyzed by using some methods like info, corr and isnull etc..

Step 2: There are no null values present in the data so no changes are done.

Step 3: There is less correlation between fbs,chol and target variable so they are droped.

Step 4: Data is visualized by using some plots like pairplot,boxplot,hist and countplot .

Step 5: Data is splitted into train set(70%) and test set(30%) by using train-test-split module.

Step 6: After preprocessing the data totally 8 machine-learning algorithms are applied to it

Step 6.1: 1)Linear Regression with accuracy of 60.79

Step 6.2: 2)Logistic Regression with accuracy of 92.307

Step 6.3: 3) Rondom Forest Regression with accuracy of 66.114

Step 6.4: 4) KNN classifier with accuracy of 79.125

Step 6.5: 5) decision tree with accuracy of 83.516

Step 6.6: 6) XGBoost with accuracy of 24.144

Step 6.7: 7) naive-bayes with accuracy of 90.109

Step 6.8: 8) support vactor machine with accuracy of 92.307

Accuracy values for each algorithm:

