MACHINE LEARNING – HEALTHCARE PROJECT

WRITE -UP

Introduction:

Cardiovascular diseases (CVDs) are the leading cause of death globally. Most cardiovascular diseases can be prevented by addressing behavioural risk factors such as tobacco use, unhealthy diet and obesity, physical inactivity and harmful use of alcohol. Out of the 17 million premature deaths (under the age of 70) due to noncommunicable diseases in 2019, 38% were caused by CVDs.

Project Scope:

To identify the causes and develop a system to predict heart attacks in an effective manner.

Dataset Description:

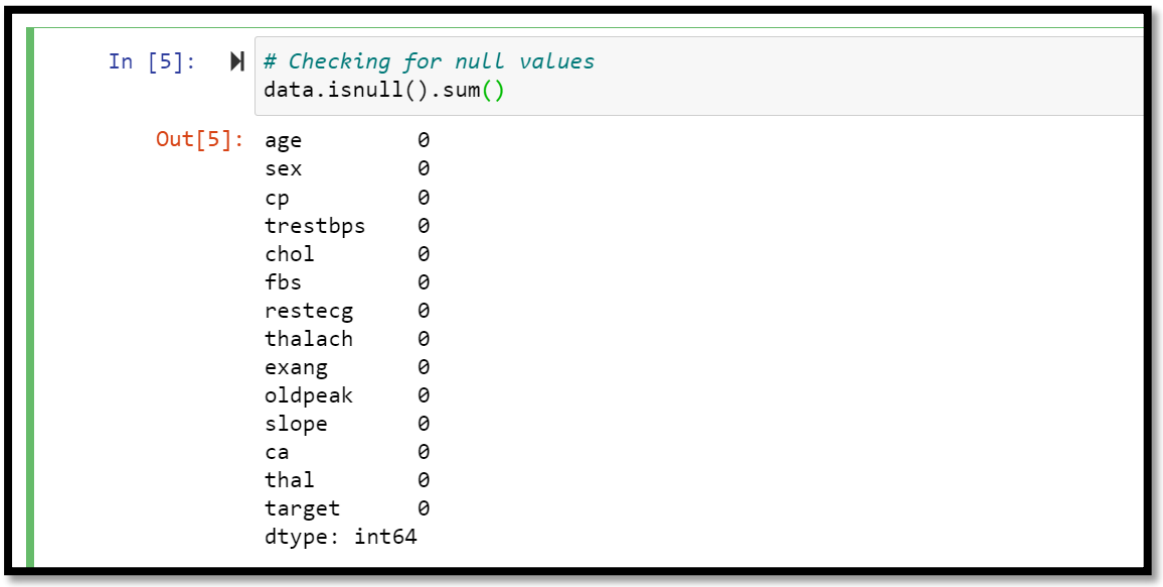
Variable	**Description**
Age	Age in years
Sex	1 = male; 0 = female
cp|	Chest pain type
trestbps	Resting blood pressure (in mm Hg on admission to the hospital)
chol	Serum cholesterol in mg/dl
fbs	Fasting blood sugar > 120 mg/dl (1 = true; 0 = false)
restecg	Resting electrocardiographic results
thalach	Maximum heart rate achieved

exang	Exercise induced angina (1 = yes; 0 = no)
oldpeak	ST depression induced by exercise relative to rest
slope	Slope of the peak exercise ST segment
ca	Number of major vessels (0-3) colored by fluoroscopy
thal	3 = normal; 6 = fixed defect; 7 = reversible defect
Target	1 or 0

Data Pre-processing

a) Null values:

Given dataset contains no null values



```
In [5]: # Checking for null values
        data.isnull().sum()

Out[5]: 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
        target      0
        dtype: int64
```

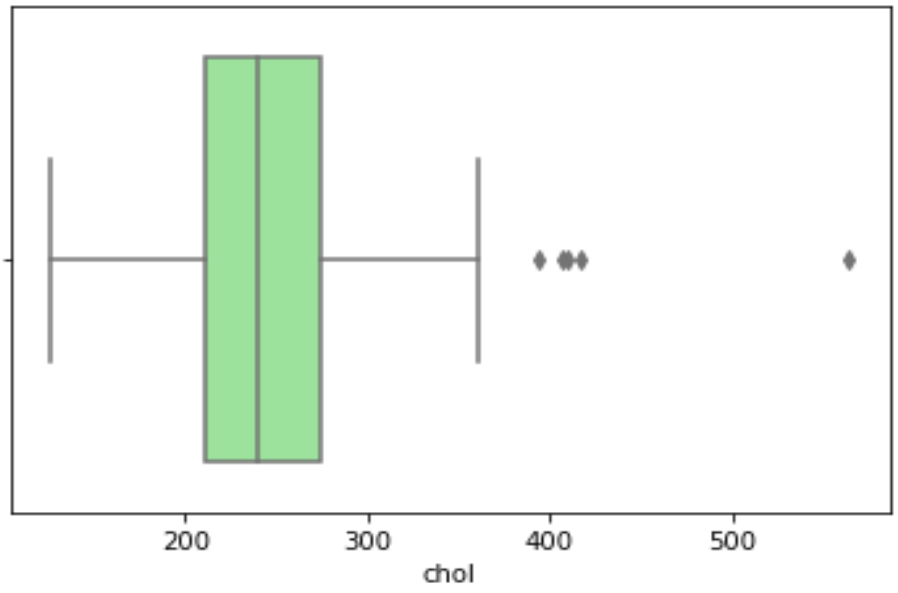
b) Duplicate Values:

One duplicate value was identified and removed

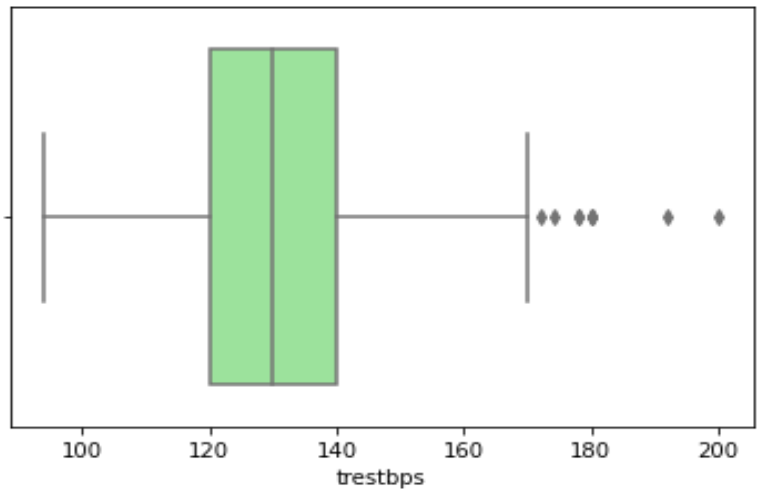
```
In [6]: # Checking for duplicate values
        duplicates=data[data.duplicated()]
        print("Duplicate values",duplicates)

Duplicate values      age  sex  cp  trestbps  chol  fbs  restecg  thalach  exang  oldpeak  \
164   38    1   2       138   175    0        1      173      0      0.0

     slope  ca  thal  target
164      2   4     2       1

In [7]: #identifying duplicate values
        dupli = data[data['age']==38]
        print("Duplicate",dupli)

Duplicate      age  sex  cp  trestbps  chol  fbs  restecg  thalach  exang  oldpeak  \
163   38    1   2       138   175    0        1      173      0      0.0
164   38    1   2       138   175    0        1      173      0      0.0
259   38    1   3       120   231    0        1      182      1      3.8

     slope  ca  thal  target
163      2   4     2       1
164      2   4     2       1
259      1   0     3       0

In [8]: # removing duplicate values
        new_data = data.drop_duplicates()
```

c) Outliers:

Outlier in Cholesterol data was identified and the values above 400 were removed.



Outlier in Blood Pressure was identified and the values above 180 were considered as outliers as per the box plot and removed.



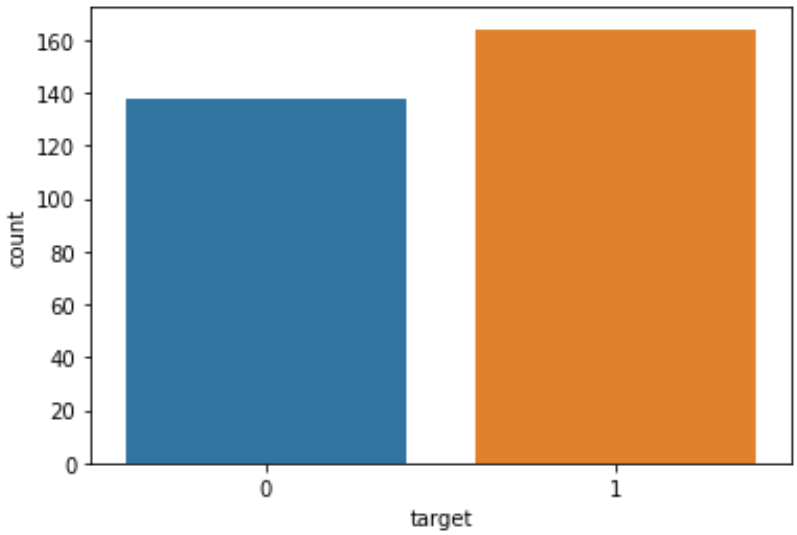
Exploratory Data Analysis (EDA)

Analysis of categorical variables were done using Count plot. First, the categorical variables were identified and they as follow,

- Target variable
- Sex
- Ca
- thal

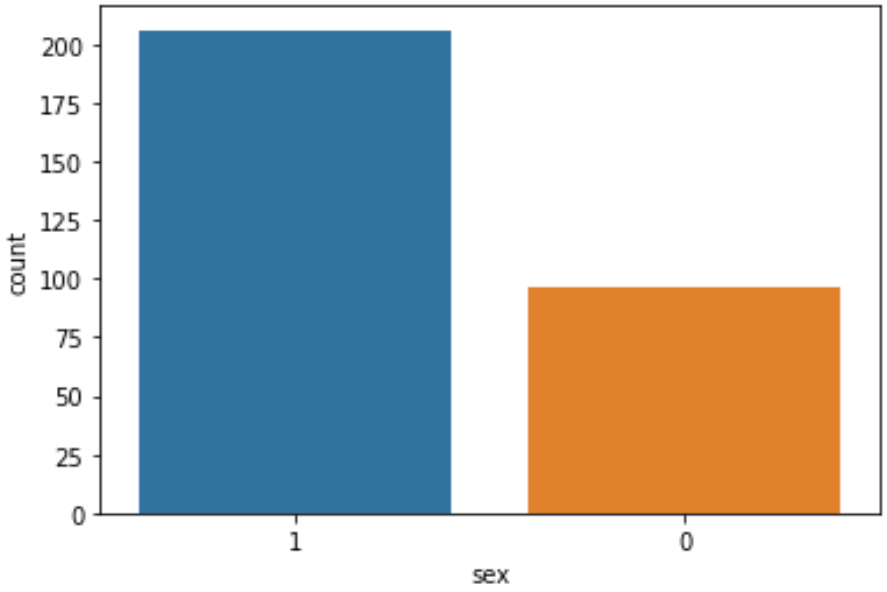
a) Target Variable:

Below graph shows the number of patients identified with cardio vascular disease and number of patients do not have CVD



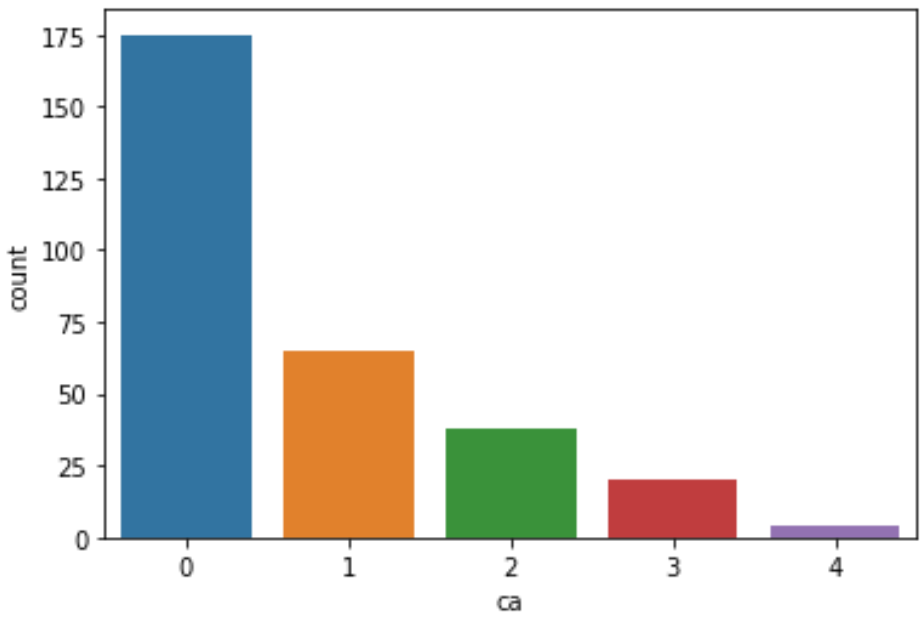
b) Gender variable:

The number of male and female members taken for the study is shown in the below graph



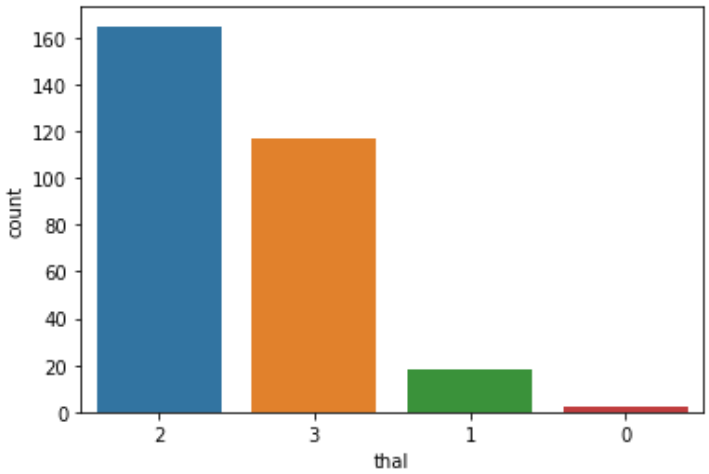
c) Blood vessels (ca)

This column shows the major blood vessels coloured by fluoroscopy and their values are 0,1,2,3. The below graph shows the number of persons under each blood vessel



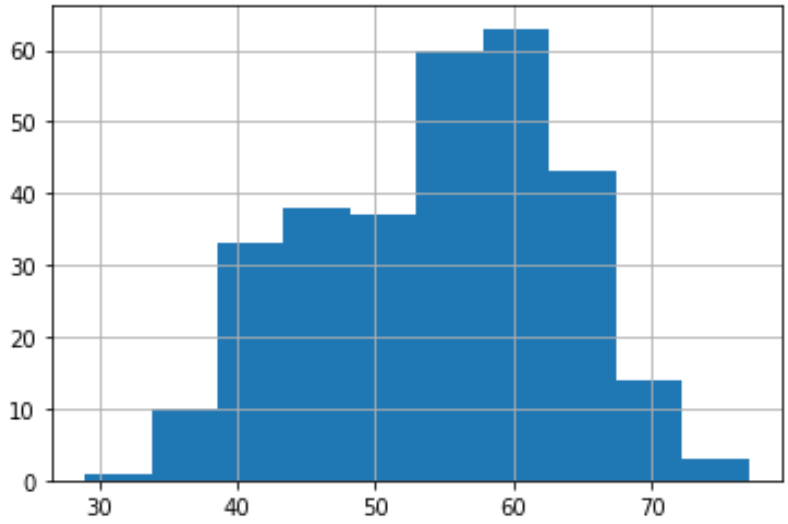
d) Thalassemia(thal)

Thalassemia is a blood disorder and can be categorised into 4 categories based on intensity. The below graph shows the classification of number of persons suffering from this disorder under each stage.



Analysis of Age Variable

The general distribution of Age variable was studied using histogram. Majority of person taken for study lies between 40 – 70 years of age.bas



Based on the above data, age variable was grouped as follow,

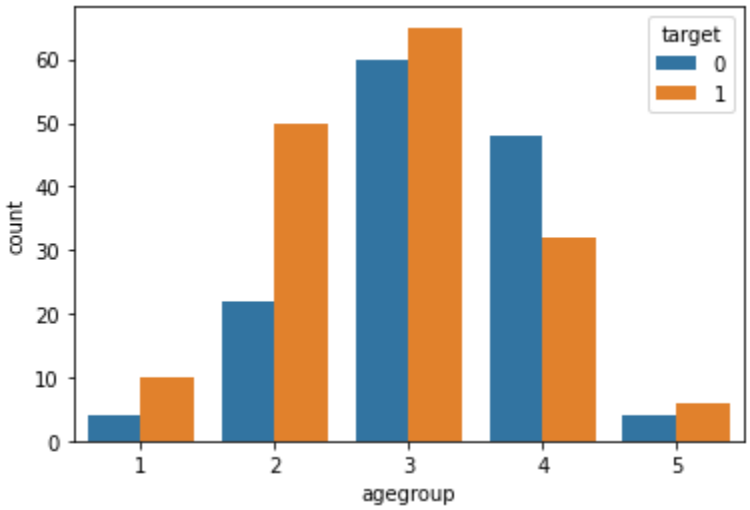
Group 1 - 30 to 40

Group 2 – 40 to 50

Group 3 – 50 to 60

Group 4– 60 to 70

Group 5– above 70

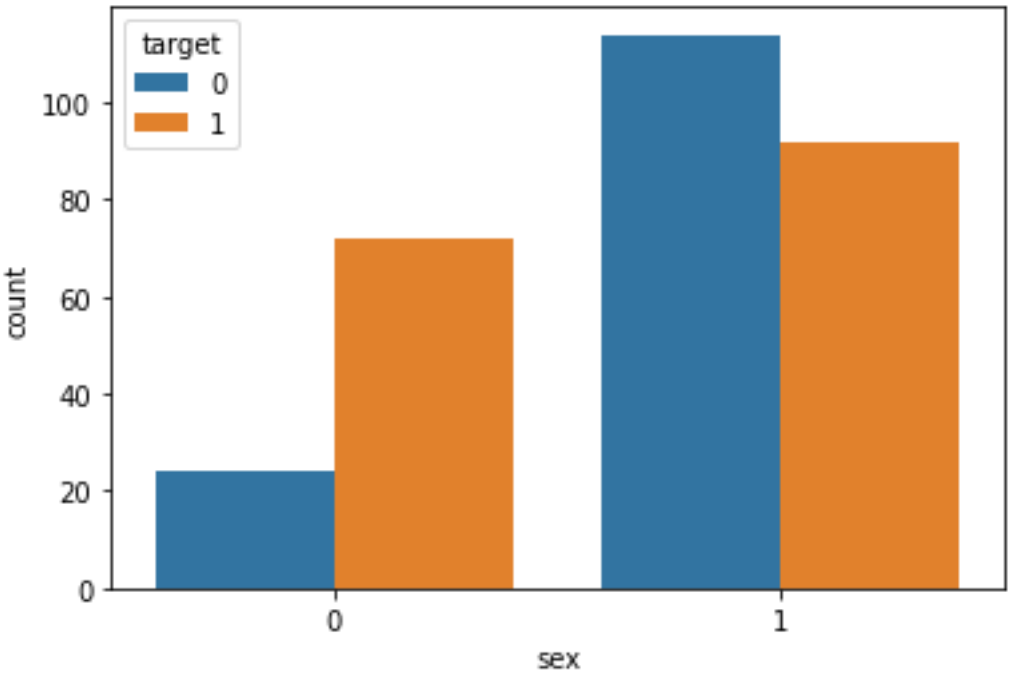


A count plot was created to showcase the age group and count of members suffering from CVD under each age group. The graph shows that people under age group of '40 – 50' and '50 – 60' have higher CVD positive cases that compared to other age groups. This shows that age is one of the contributing factor for CVD

Analysis of Gender Variable

The composition of male and female members taken under study are,

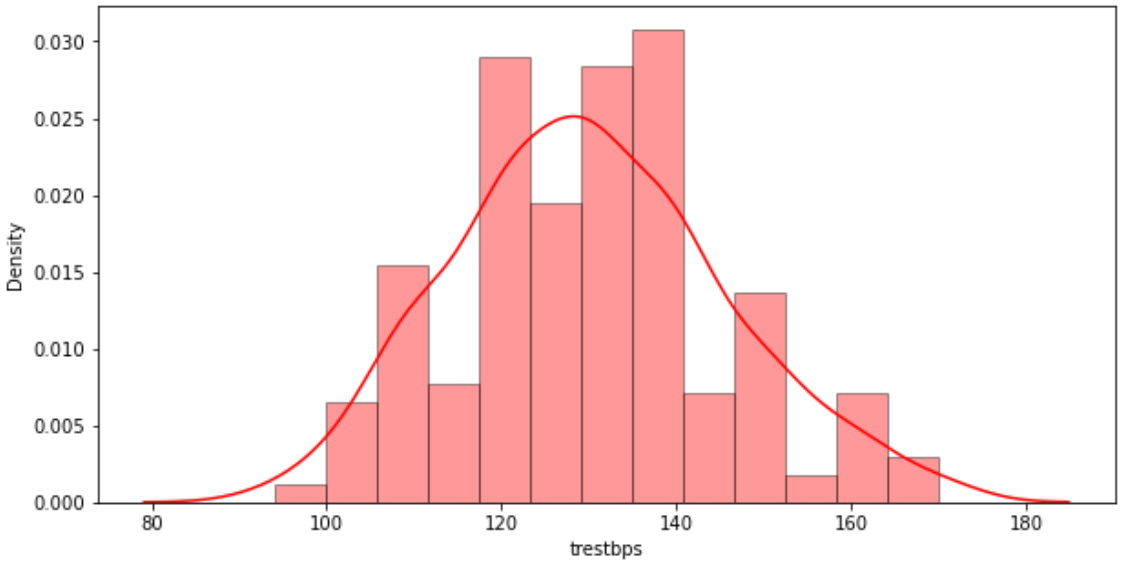
- Male – 206
- Female – 96



A study to identify which gender has more CVD cases was done using above graph and it shows each gender and the number of persons suffering from CVD in each gender. It is observed that Female gender is more likely to have higher rate compared to opposite gender. But still Gender alone cannot be considered as a contributing factor to CVD.

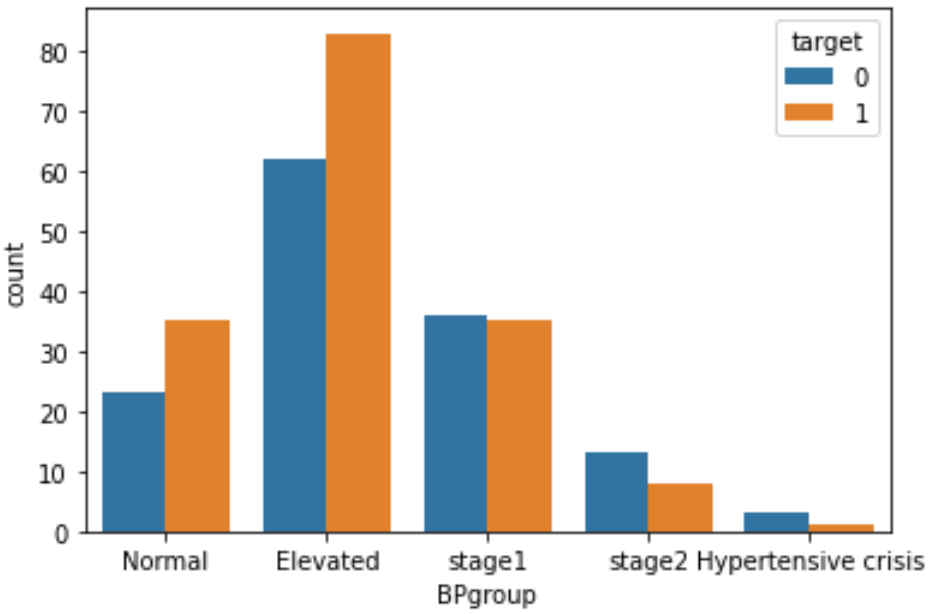
Study on BP:

The normal blood pressure of a person is 120 mm Hg for systolic movement as per the general guideline of healthcare department. First, the spread of the data was studied using distribution plot. The plot shows that most of the people have BP between 120 – 140 which a normal one.



Based on the table given by the guidelines of health care, data are grouped in to following category for ease of analysis.

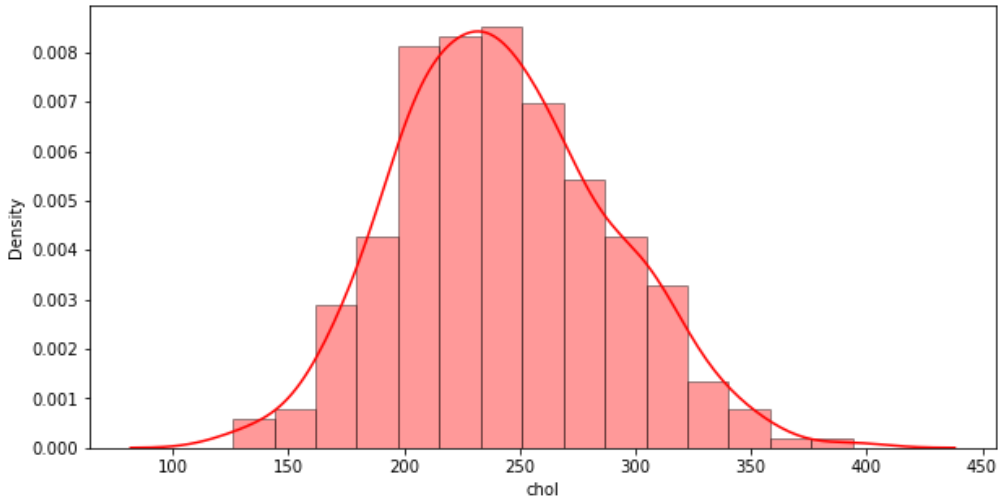
- NORMAL - below 120
- ELEVATED - 120-140
- HYPERTENSION STAGE1 - 140-160
- HYPERTENSION STAGE2 - 160-180
- HYPERTENSIVE CRISIS - above 180



Since most of the people fall under 'Elevated' group as per their blood pressure reading the above graph also shows that people under that group has higher positive CVD cases. Blood pressure is surely a contributing factor for cardio vascular disease and the above graph proves them.

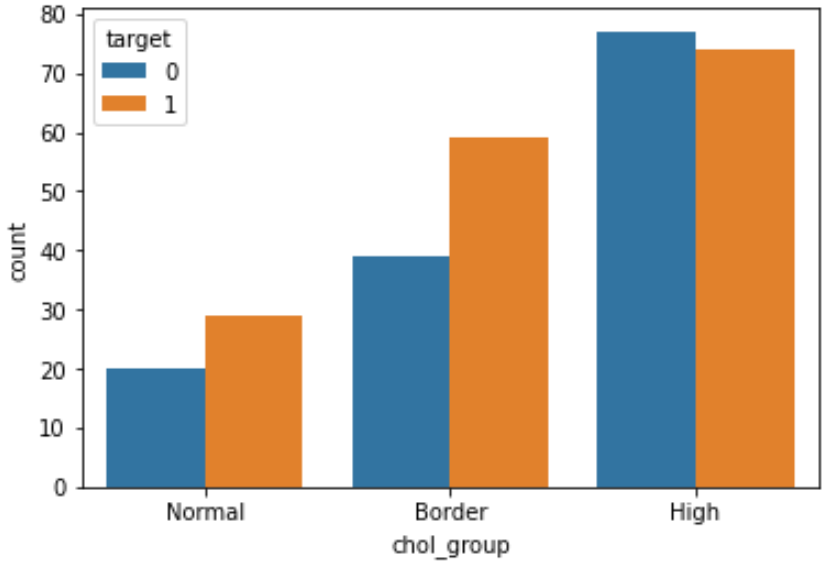
Analysis of Cholesterol:

The cholesterol levels of a normal person is 200 for total cholesterol as per the general guideline of healthcare department. First, the spread of the data was studied using distribution plot. The plot shows that most of the people have their cholesterol levels between 200 -250.



Based on the table given by the guidelines of health care, data are grouped in to following category for ease of analysis.

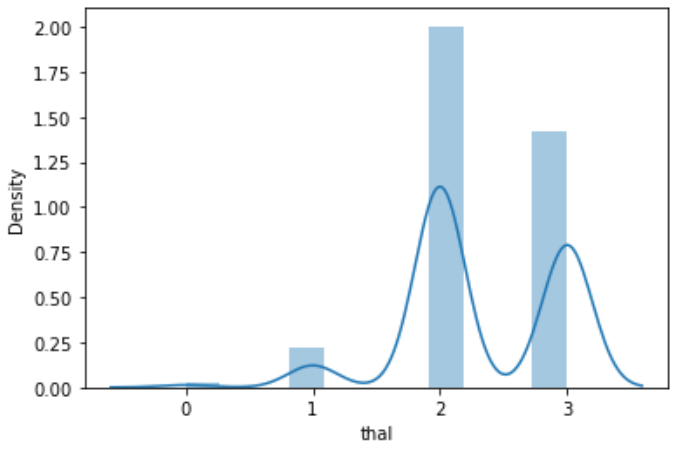
- NORMAL - below 200
- BORDER - 200 -250
- HIGH - above 250



The above graph clearly shows that people with high cholesterol have more risk of having CVD and definitely this is a contributing factor for heart disease. More positive cases are seen under 'High' category and cholesterol levels are above 250 for the people falling under this category.

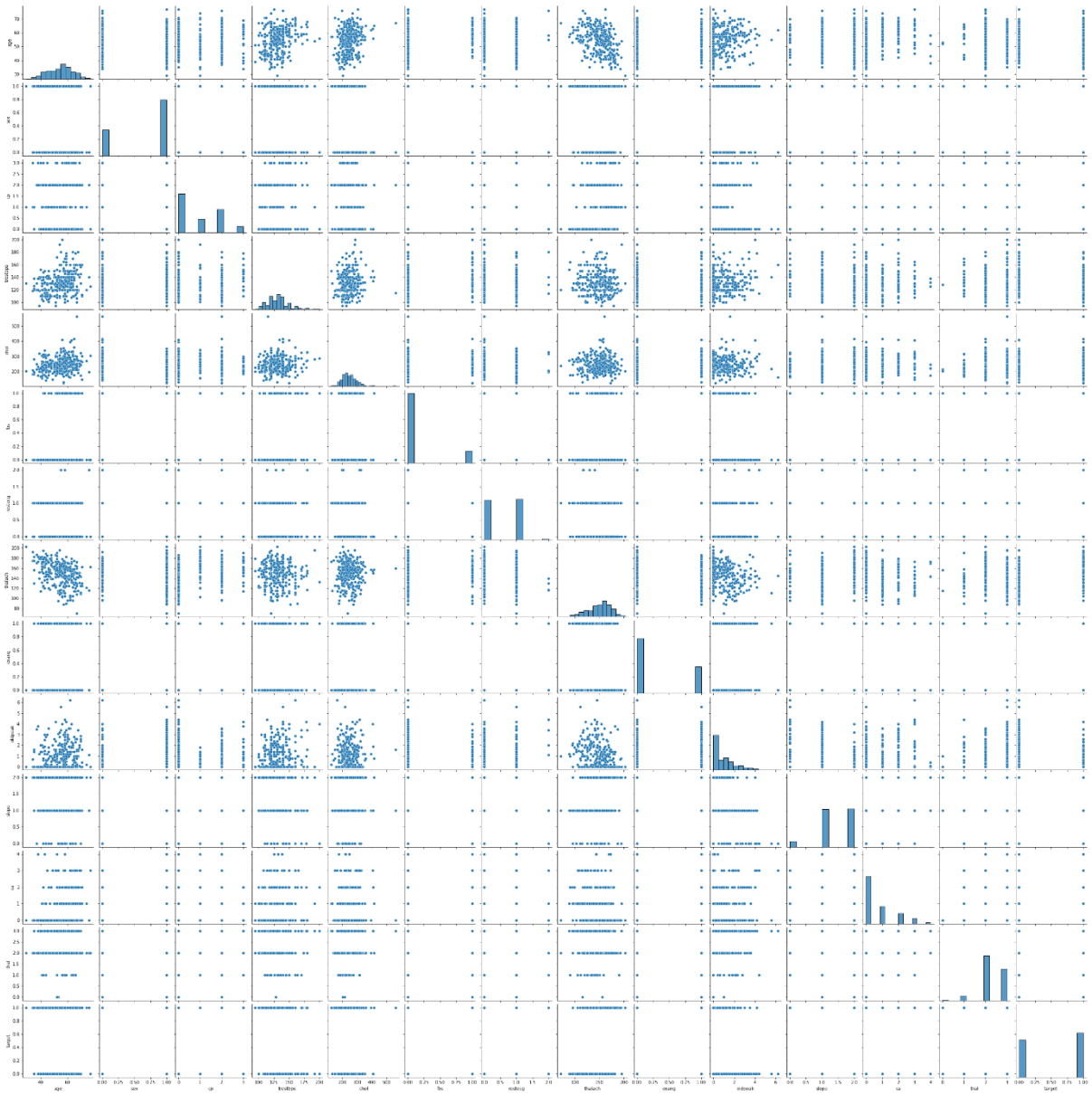
Study on Thalassemia:

General distribution of thal variable is studied using distribution graph. Thalassemia is a blood disorder and can be categorised into 4 categories based on intensity. The below graph shows the distribution of data across 4 categories. It is observed that most of the people fall under stage2. This is a contributing factor to heart disease.



Pair Plot Analysis:

To study the pairwise relationship between the variables in the given dataset Pair plot is used. The pairs plot builds on two basic figures, the histogram and the scatter plot. The histogram on the diagonal allows us to see the distribution of a single variable while the scatter plots on the upper and lower triangles show the relationship between two variables.



Above plot shows that ‘age’ is positively correlated with ‘BP’ and ‘cholesterol’ and negatively correlated with ‘maximum heart rate achieved(thalach)’. The exact relationship between other variables were not able to conclude using this method.

Fitting models:

a) Train – Test Split:

To build a model that best fits the given dataset, First the data was split into test and training dataset in 70:30 ratio. The target and features were identified and assigned to X and Y variables

b) Logistic Regression

Logistic regression is used when there is one dependent variable and one or more independent variables. The target variable is of categorical variable. The Accuracy thus obtained while applying logistic regression to the given dataset is 81.31%

Logistic Regression

```
In [101]: classifier = LogisticRegression(solver='lbfgs',max_iter=10000)

In [102]: classifier.fit(X_train, y_train)
Out[102]: LogisticRegression(max_iter=10000)

In [103]: y_pred=lr.predict(X_test)

In [104]: accuracy_score(y_test,y_pred)
Out[104]: 0.8131868131868132
```

c) Random Forest

It is an ensemble learning method for classification and operates by constructing multitude of constructing decision trees. Applying random forest to the given dataset shows an accuracy of 85.71% which a good fit.

```
Out[86]: RandomForestClassifier(max_depth=50, n_estimators=900, random_state=0)

In [87]: y_pred=rfc.predict(X_test)

In [91]: accuracy_score(y_test,y_pred)
Out[91]: 0.8571428571428571

In [92]: #Confusion Matrix
         confusion_matrix(y_test,y_pred)
Out[92]: array([[36,  4],
                [ 9, 42]], dtype=int64)

In [95]: print(classification_report(y_test,y_pred))
                    precision    recall  f1-score   support

                 0       0.80      0.90      0.85        40
                 1       0.91      0.82      0.87        51

          accuracy                           0.86        91
         macro avg       0.86      0.86      0.86        91
      weighted avg       0.86      0.86      0.86        91
```

Hence, we can conclude that both the models fit the given dataset and Random Forest is the best fit for predicting cardio vascular disease.