Lebuga Bosco | 2024/HD05/25333U | 2400725333

Exploratory DAP report on the Dataset for Heart disease Prediction

Problem statement

Currently in Uganda and other parts of the word, heart disease related deaths have remained predominant challenge both in communities and health facilities, despite many ways advanced to prevent those deaths.

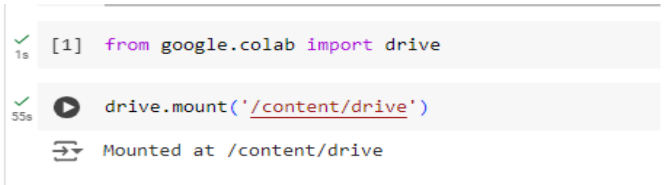
This prompts the main question, how do we detect if a person is at high risk of heart disease? Which risk factors are predominant in predicting heart disease?

Step1; Gathering the Data

I used the dataset in the link below as case study/sample to study the heat disease prediction.

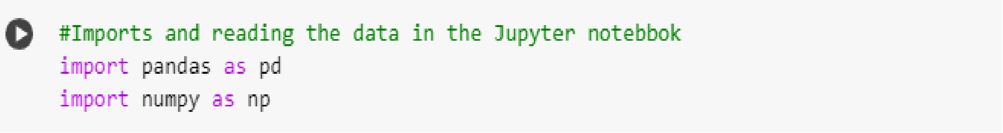
https://github.com/kb22/Heart-Disease-Prediction/blob/master/dataset.csv

The dataset is stored on my google drive and then loaded into Google Colab by importing the drive and mounting it with its contents.

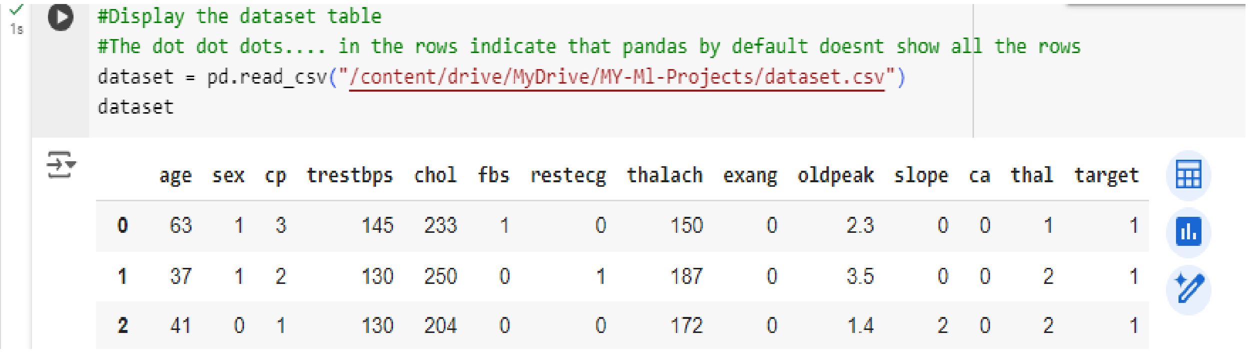


```
[1] from google.colab import drive

drive.mount('/content/drive')

Mounted at /content/drive
```

With the help of the Pandas Data frames, the gathered dataset can be displayed



```
#Imports and reading the data in the Jupyter notebbok
import pandas as pd
import numpy as np
```



```
#Display the dataset table
#The dot dot dots.... in the rows indicate that pandas by default doesnt show all the rows
dataset = pd.read_csv("/content/drive/MyDrive/MY-Ml-Projects/dataset.csv")
dataset
```

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1

Step2 Understanding the Dataset

Describing the different features of the data set and interpreting what they mean

S/n	Feature Name	Description of feature
1	age	The age of the person at the time
2	sex	The gender of the person as at birth (has 2 values)
3	cp	The chest pain type (4 values)
4	trestbps	The resting blood pressure of the person
5	chol	Serum cholesterol in mg/dL
6	fbs	Fasting blood sugar > 120 mg/dL
7	restecg	Resting electrocardiographic results (values 0,1,2)
8	thalach	Maximum heart rate achieved
9	exang	Exercise induced angina
10	oldpeak	Old peak = ST depression induced by exercise relative to rest
11	slope	The slope of the peak exercise ST segment
12	ca	Number of major vessels (0-3) colored by fluoroscopy 0 = normal; 1 = fixed defect; 2 = reversible defect.
13	thal	The inherited blood disorder caused when the body does not make enough hemoglobin (4 values)
13	target	The result/output

For precise and clear understanding of the datasets, the following tests were run as follows

dataset.shape: This displayed the dimension of the dataset having the total number of columns and rows of the table as above. The dataset has 303 rows and 14 columns.

Output: (303, 14)

dataset.columns; this displayed all the columns of the dataset as seen below

```
output: Index(['age', 'sex', 'cp', 'trestbps', 'chol', 'fbs', 'restecg', 'thalach',
       'exang', 'oldpeak', 'slope', 'ca', 'thal', 'target'],
      dtype='object')
```

dataset.dtypes: Displayed each column with their data types, from here it shows that all the columns are Int64 types except only ‘oldpeak’ column which is of float64 data type.

dataset.head(5); displayed the top five records

dataset.tail(5); Displayed the bottom 5 records of the dataset.

dataset.sample(5): it picked any five random or sample records from the dataset and displayed

dataset.info() ;Displayed the different features of the dataset with their data types, total no of columns =14, total no of index entries = 303 from 0 to 302, 1 column is float data type and 13 columns are int data type, memory usage = 33.3KB

dataset.describe(); Displays all the columns with their respective count, mean, std, min, max, percentile values . For example the feature "chol" has count of 303, average value of 246.264026, standard deviation of 51.830751, and minimum value of 126 and maximum of 564, 211 on the 25%, 240 on 50% range and 274 on the 75% percentile.

Observations during understanding the dataset

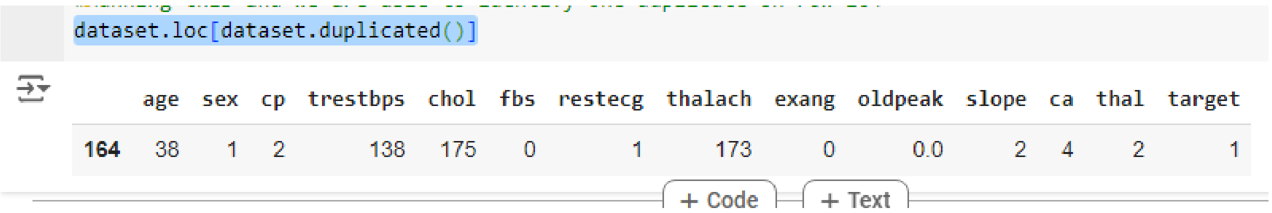
Like we can observe that some standard deviations are slightly higher than the mean values for the features of 'cp', 'exang', 'oldpeak', 'ca', and this indicates high variations between values and abnormal distribution of data. There is also need to do normalization so that all the data is in the same range to avoid much variations in relationships.

Step3: Data Preparation or Data Cleaning and Data Wrangling

Here I checked for missing/null values, duplicate values, outliers,

dataset.isna().sum() Ran the code to check for null or missing values in all the rows or entries. The result turned out that there were no missing values in my dataset. The display showed that the total number of null values for each column = 0

`dataset.loc[dataset.duplicated()]` Ran the code to check for existence of duplicates. 1 record was identified to be duplicate on row 168 as shown below



```
dataset.loc[dataset.duplicated()]
```

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
164	38	1	2	138	175	0	1	173	0	0.0	2	4	2	1

```
dataset =dataset.loc[~dataset.duplicated()]\
```

`.reset_index(drop = True).copy()` I used the code above to delete the duplicate and also created a new copy of clean dataset to be used.

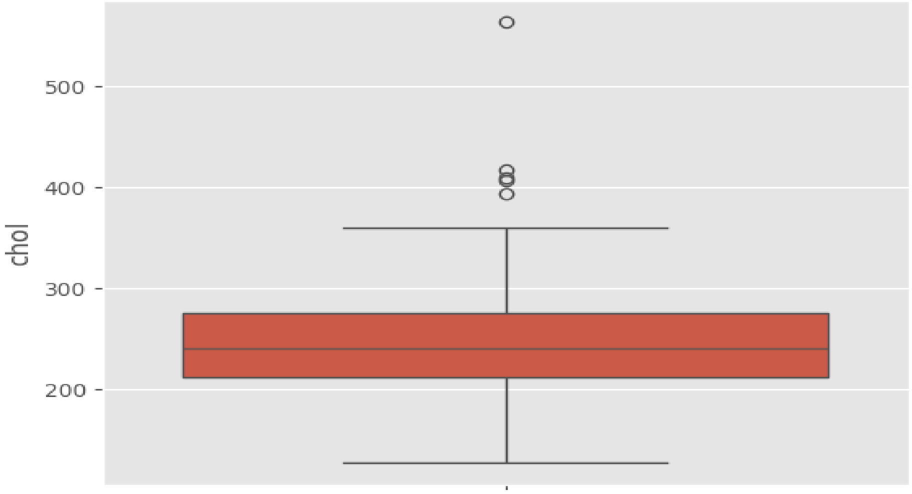
`dataset.shape` tried to check the shape of the new dataset, this displayed(302, 14). This implies that one record was deleted successfully since our original dataset has 303 rows.

Checking for outliers

Using the boxplot to check if there exits outliers in our dataset. Tested this on the ‘chol’ column. The result displayed showed that the chol column has an outlier as shown in the plot here.

```
base_color = sb.color_palette()[0]
sb.boxplot(data = dataset,  y ='chol', color = base_color)
plt.xticks(rotation =15);
```

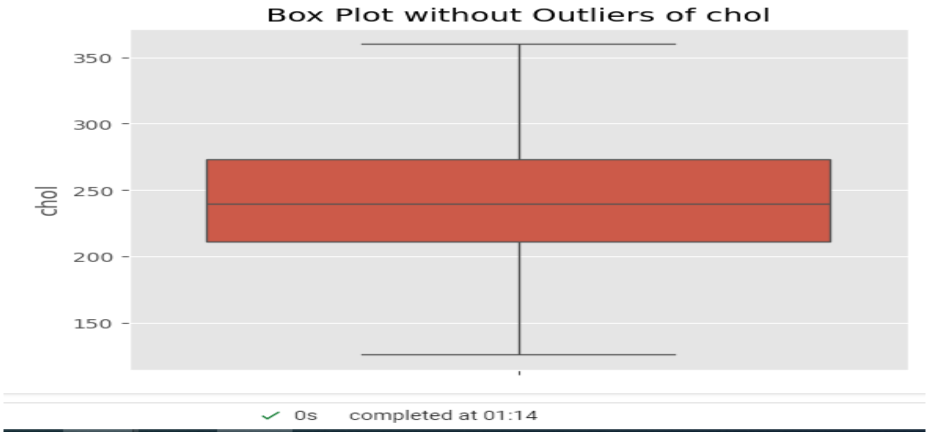
In the below plot u can clearly see that the values above 380 are acting as outliers in the dataset.



Removing outliers

Using the code below, we can remove the outliers and plot a new plot for chol column as shown

```
def removal_box_plot(dataset, column, threshold):
removed_outliers = dataset[dataset[column] <= threshold]
sb.boxplot(removed_outliers[column])
plt.title(f'Box Plot without Outliers of {column}')
plt.show()
return removed_outliers
threshold_value = 380
no_outliers = removal_box_plot(dataset, 'chol', threshold_value)
```



Deduction on data preparation/Cleaning/Wrangling

Our dataset had 303 instances, 1 duplicated record was detected and removed, there are no missing values, and 1 outlier was observed with ‘chol’ column and removed. New dataset now has 302 records and 14 features all numerical in nature.

Step4; Data Visualization,

Here I performed Univariate analysis, Bivariate and Multivariate Analysis on the dataset.

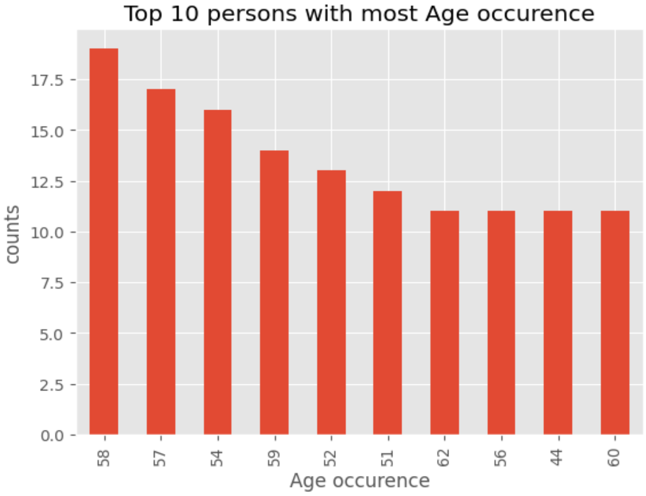
Univariate analysis

Here I was describing how each feature in dataset behaves, like the number of occurrences and distributions of values of each feature, most occurring and least occurring figures for each feature.

`dataset['age'].value_counts()`The code was used to display the number of occurrences of unique values for the ‘age’ feature. It showed that age 58 occurred most times (19 times) while 77, 76, 74, 29 occurred least times (1 time)

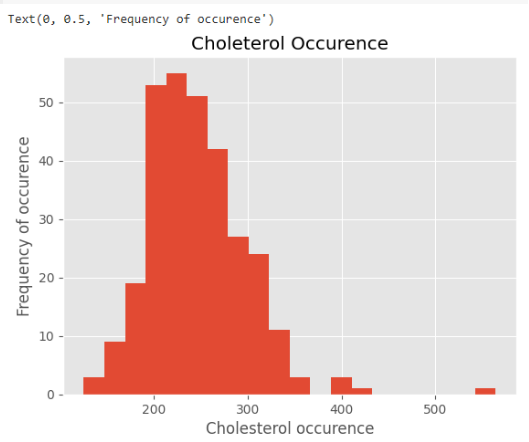
Performing another univariate analysis on the ‘chol’ feature by plotting a bar graph to display the top 10 most occurring values in the ‘chol’ feature

```
dataset['age'].value_counts()\
                   .head(10)\
                   .plot(kind ='bar', title ='Top 10 persons with most Age
occurence')
plt.xlabel('Age occurence')
plt.ylabel('counts')
```



We can also use the above code to check for cholesterol level occurrence and distribution by plotting a histogram

```
ax =dataset['chol'].plot(kind ='hist' bins =20,title ='Choleterol
Occurence')
ax.set_xlabel('Cholesterol occurence')
ax.set_ylabel('Frequency of occurence')
```



```
Text(0, 0.5, 'Frequency of occurence')
```

Deduction on univariate Analysis:

Accordingly, the univariate analysis on the column ‘age’ depicts that age 58 has the most occurrences while 62, 56, 44, 60 have least occurrences as shown in the bar graph. Simarly a univariate analysis on ‘chol’ depicts that the range of values between 200 to 290 are mostly occurring.

Bivariate Analysis

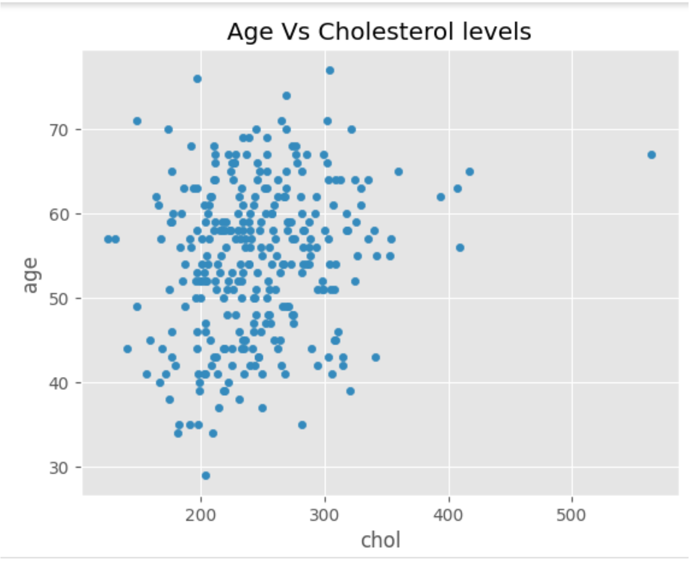
Understanding how two features of the data set and their distributions relate to each other.

I used Scatter plot, Heatmap plot correlation, Pair plot and groupBy comparisons

Checking the relationship between age and cholesterol levels using the code below

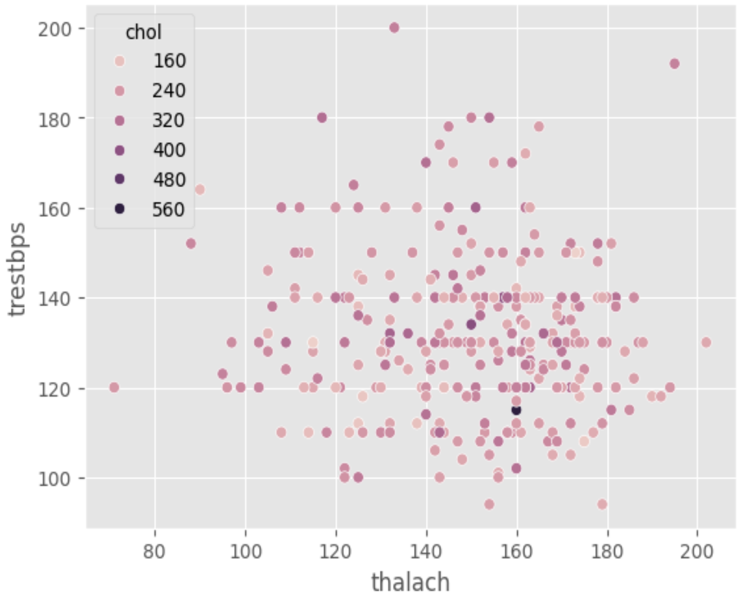
```
dataset.plot(kind ='scatter', x='chol', y = 'age', title = 'Age Vs Cholesterol levels')
```

The plot implies that between ages of 40 and 70, there are high concentrations of cholesterol levels.



Established the relation between resting blood pressure(trestbps) and maximum heart rate (thalach) and adding another third feature to color the two features(hue =chol)

```
sb.scatterplot(x='thalach',
                y ='trestbps',
                hue = 'chol',
               data = dataset)
```



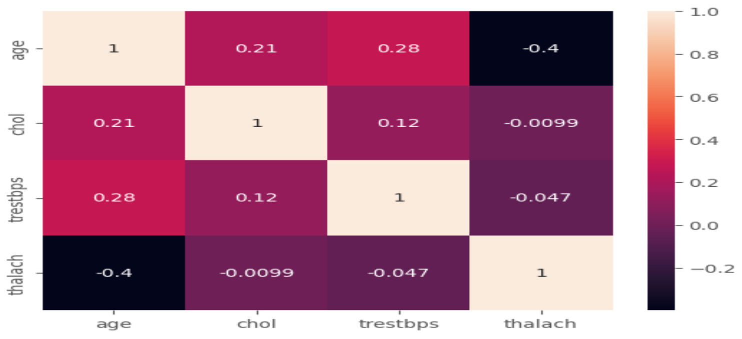
Observations on bivariate analysis;

For age vs cholesterol, as age is around 40 to 60, cholesterol levels are in high concentrations of from 200 to 300. Similarly as the maximum heart rate (thalach) increases, the resting blood pressure (trespbps) also increases. But heart rates of range 120 to 170 also correspond to increased rest blood pressure in the range of 120 to 140 as shown in the scatter plot.

Multivariate Analysis

Here I established correlation between many features by using HeatMaps coorelation method as shown using the code the code below

```
sb.heatmap(dataset_corr, annot =True)
```



	age	chol	trestbps	thalach
age	1	0.21	0.28	-0.4
chol	0.21	1	0.12	-0.0099
trestbps	0.28	0.12	1	-0.047
thalach	-0.4	-0.0099	-0.047	1

Observations from the Multivariate analysis

As seen, some features have negative correlations and others have big correlations for example, there exists a positive correlation between age and cholesterol implying an increase in age leads to increase in cholesterol but it's a weak correlations also a negative correlation between age and maximum bps implies an increase in age leads to decrease in blood pressure but the correlation is weak one (<-0.3>)

Step5: Questions

Qn1. Which age bracket has the highest cholesterol levels (minimum of 200 cholesterol and above 5 count)?

Answered this question using the statistics generated by the code below, the age bracket of 41-67 has the highest average cholesterol levels of above 200 and most occurring

```
dataset.groupby('age')['chol']\
       .agg(['mean', 'count'])\
       .query('mean >=200 and count >=5')\
       .sort_values('mean')
```

Qn2. Which features of the dataset contribute to heart disease most?

QN3. What is the correlation between age, cholesterol, blood pressure and heart rate?

Conclusions

In summary, we can conclude that out original dataset wasn't all that accurate to perform the analysis. It had 303 instances of data and 14 features. But 1 instance of the record was duplicated and hence removed. The total number of records in the cleaned dataset after cleaning and data wrangling becomes 302. The new dataset look substantially enough to predict the risk of heart disease.

The strong features in dataset which can be used to predict the heart disease include the cholesterol levels (chol), maximum blood pressure (thalach), the rest blood pressure (trestbps) and age also affects.