*HEART ATTACK*

*PREDICTION*

**Problem statement:**

Cardiovascular diseases are the leading cause of death globally. It is therefore necessary to identify the causes and develop a system to predict heart attacks in an effective manner. The data below has information about the factors that might have an impact on cardiovascular health.

**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 |

**Note:**

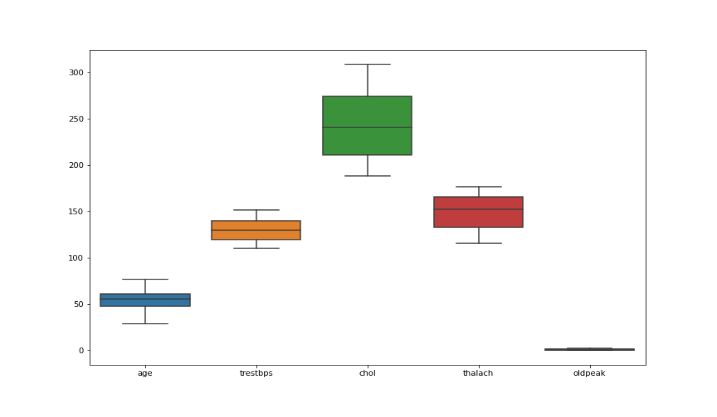
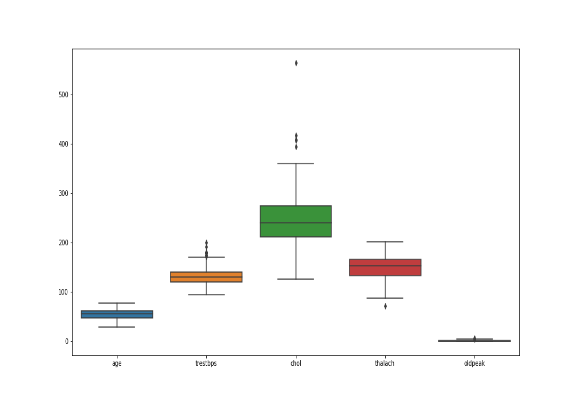
Download **CEP 1\_ Dataset.xlsx** using the link given in the **Healthcare** project problem statement

**Task to be performed:**

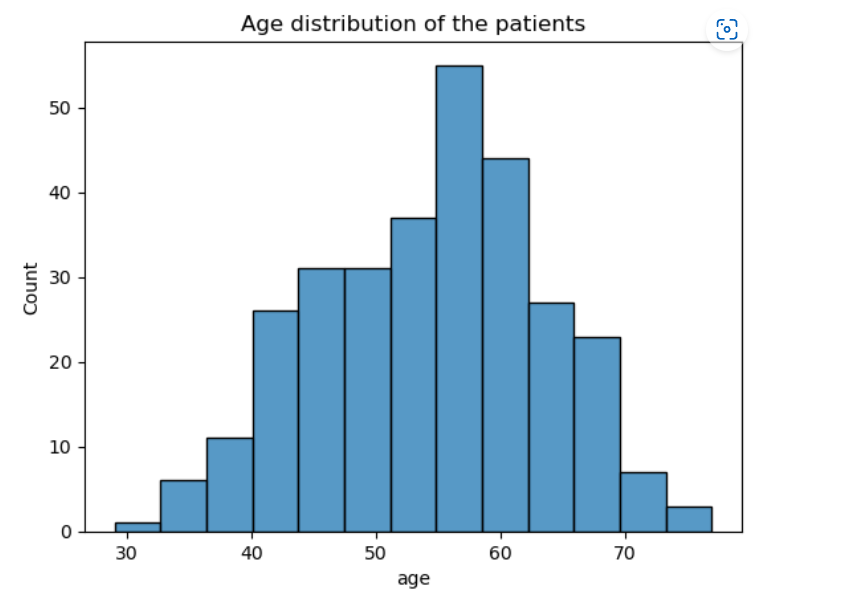
1. Preliminary analysis:
   1. Perform preliminary data inspection and report the findings on the structure of the data, missing values, duplicates, etc.
   2. Based on these findings, remove duplicates (if any) and treat missing values using an appropriate strategy
2. Prepare a report about the data explaining the distribution of the disease and the related factors using the steps listed below:
   1. Get a preliminary statistical summary of the data and explore the measures of central tendencies and spread of the data
   2. Identify the data variables which are categorical and describe and explore these variables using the appropriate tools, such as count plot
   3. Study the occurrence of CVD across the Age category
   4. Study the composition of all patients with respect to the Sex category
   5. Study if one can detect heart attacks based on anomalies in the resting blood pressure (trestbps) of a patient
   6. Describe the relationship between cholesterol levels and a target variable
   7. State what relationship exists between peak exercising and the f
   8. Check if thalassemia is a major cause of CVD
   9. List how the other factors determine the occurrence of CVD
   10. Use a pair plot to understand the relationship between all the given variables
3. Build a baseline model to predict the risk of a heart attack using logistic regression and random forest and explore the results while using correlation analysis and logistic regression (leveraging standard error and p-values from statsmodels) for feature selection

Steps followed

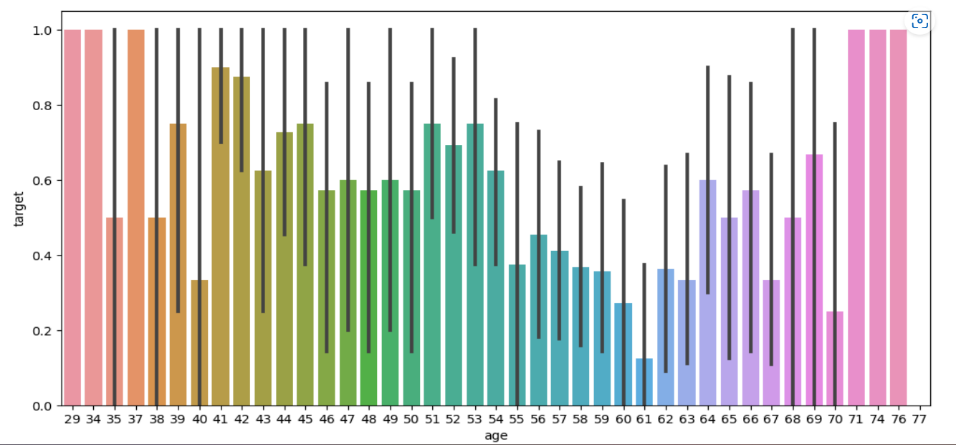
1. Required Library are imported
2. Dataset is imported as (Heart-dataset.csv)
3. Info() shows more information about the datasheet and checks null values.
4. Shape method shows there are a total of 303 rows and 14 columns.
5. Columns method shows different column attributes.
6. Describe() shows statistical information about dataset.
7. Null value checked using isnull().sum() – there is no null value found
8. Duplicated() Check any duplicate present in dataset.
9. Then checked for outliers on (age, trestbps, chol, thalach, oldpeak . And found that except age there are some outliers present in all attributes.
10. Then outliers are handled using IQR method.

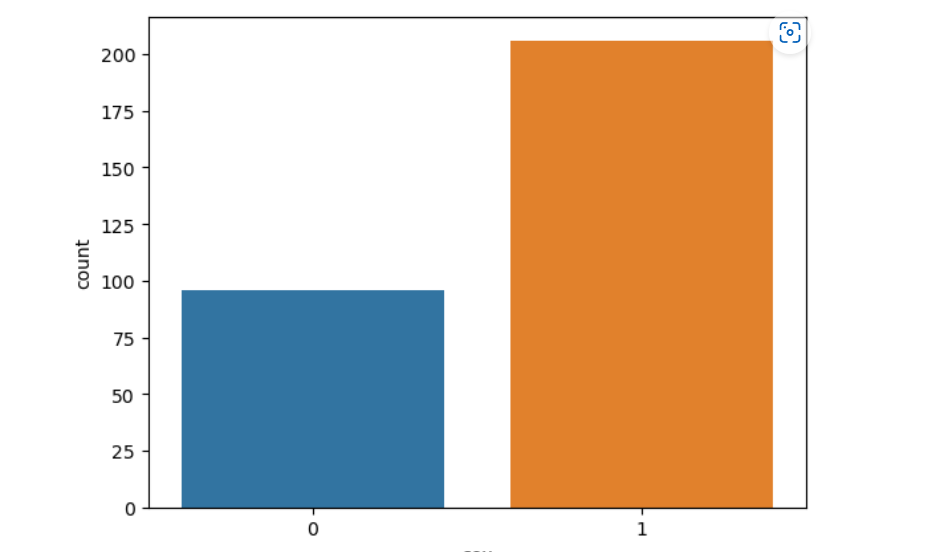


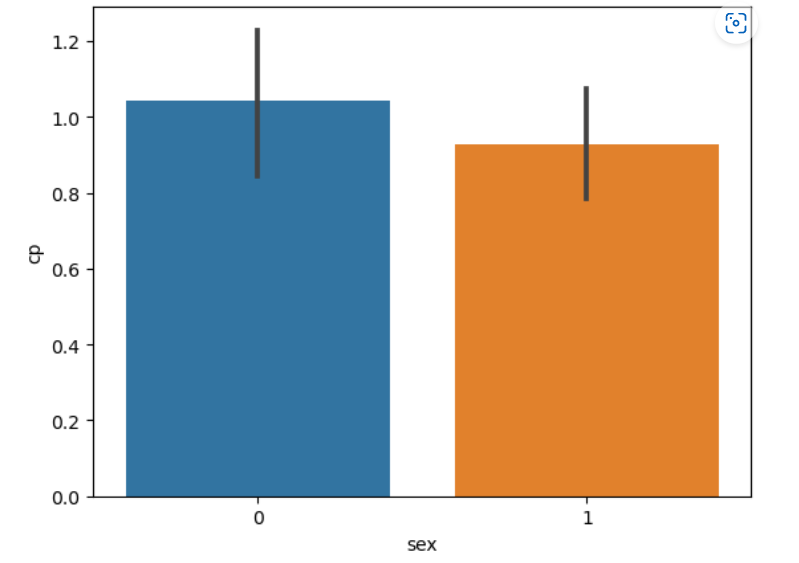
1. A histogram was drawn – it shows The below diagram give insight about the age category of patient. The majority of the patients are older than 40. The peak is at 56-57 years old.



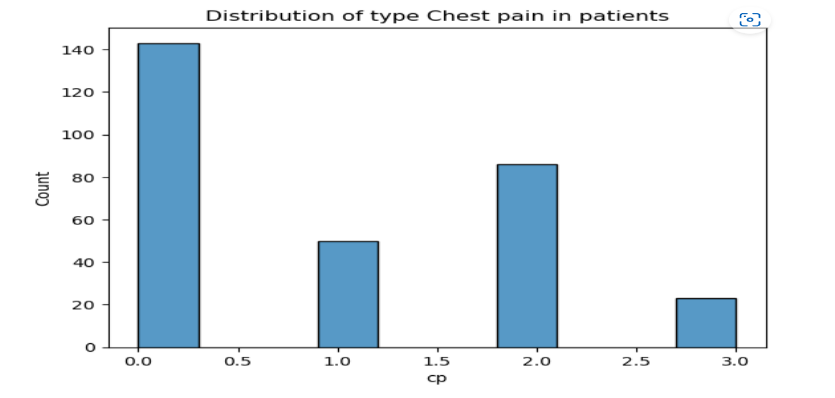
1. The subplot shows occurrence of CVD across the Age category.



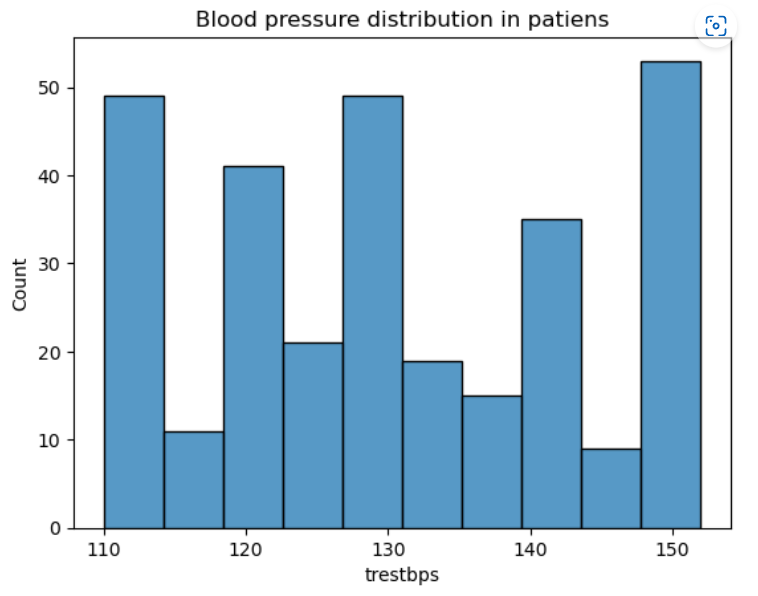
1. Here shows that there are more males than females in the patients category. (0=female, 1= male) 
2. Chest pain among male=1 and female=0



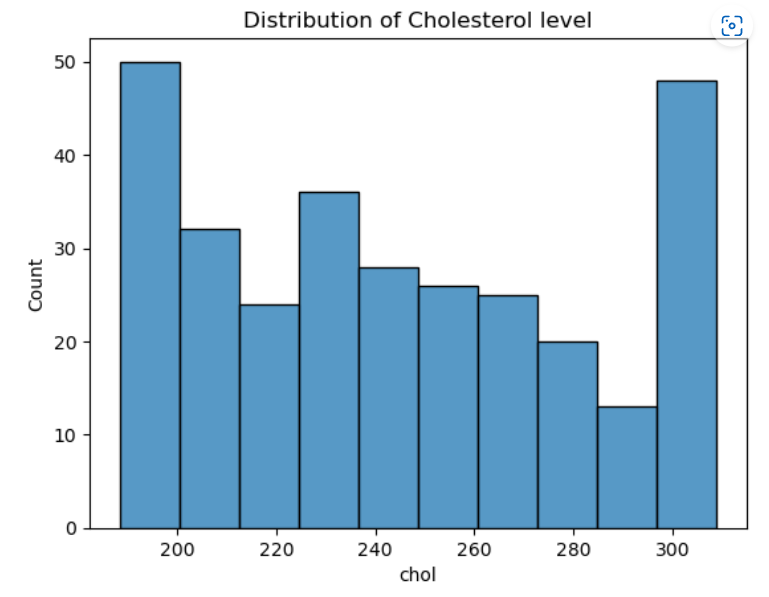
1. Distribution types of chest pain. 1 = typical angina; 2 = atypical angina; 3 = non-anginal pain; 0 = asymptomatic. Almost half of the patients have no pain in the chest



1. Distribution of blood pressure in patients. In our dataset, the resting blood pressure distribution has a peak at a value of approx. 150



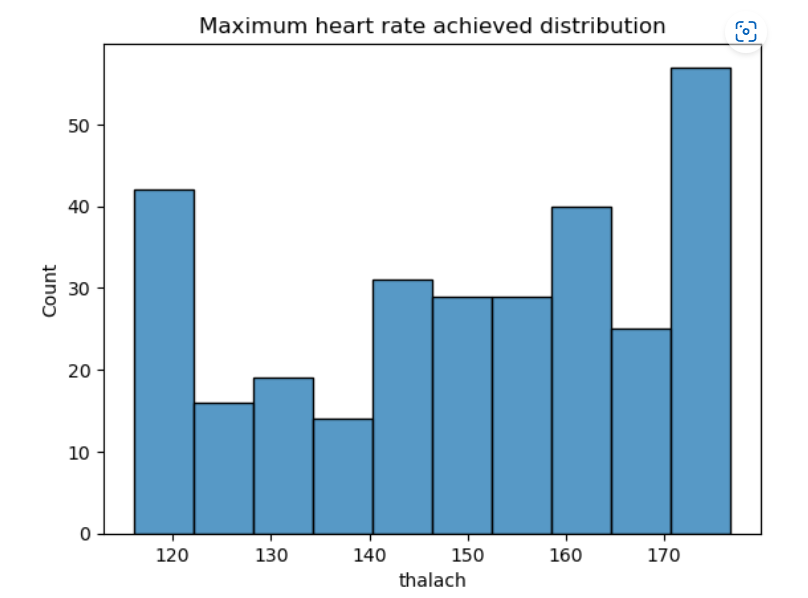
1. Distribution of Cholesterol level among patients. In this dataset some patients have an extremely high level of cholesterol.



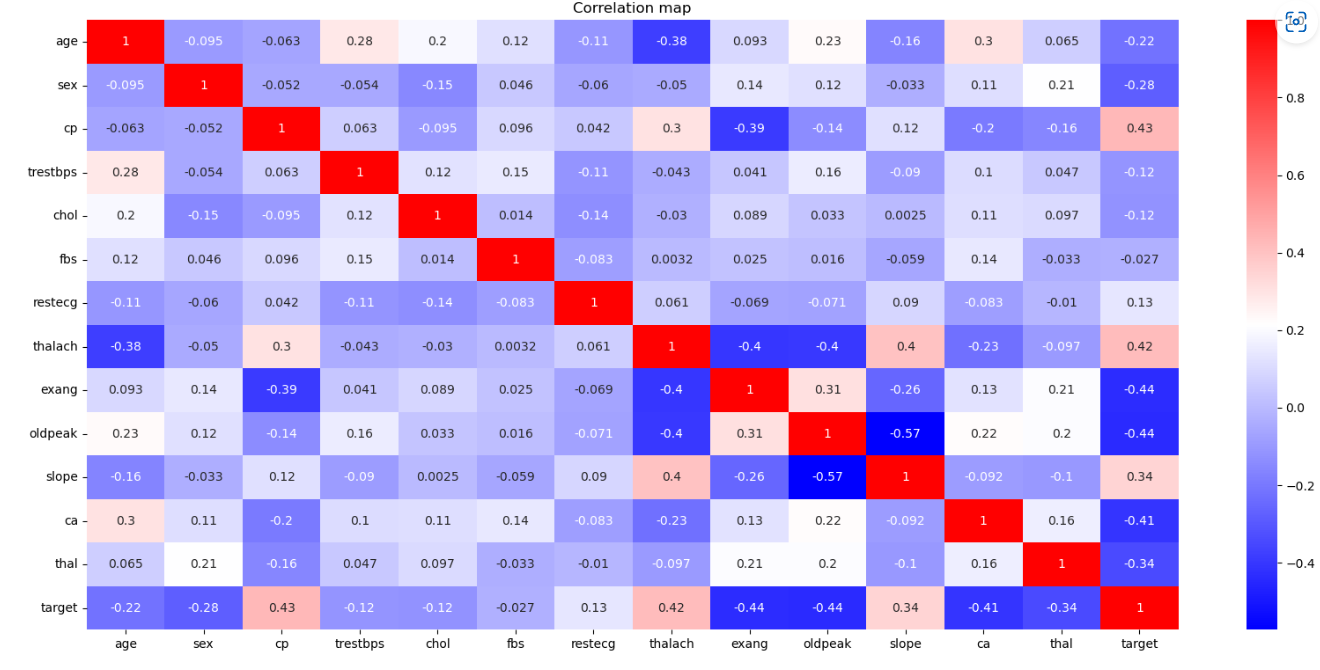
1. The slope of the peak exercise ST segment distribution. 0 = downsloping; 1 = flat; 2 = upsloping  
   The slope of peak exercise ST segment (slope) is mostly flat and upsloping.



1. Distribution of maximum heart rate achieved.



1. Correlation . From the above heatmap we can observe that the features "cp", "thalach", "exng", "oldpeak", and "ca" are highly correlated to target that is possible heart attack.



Conclusion.

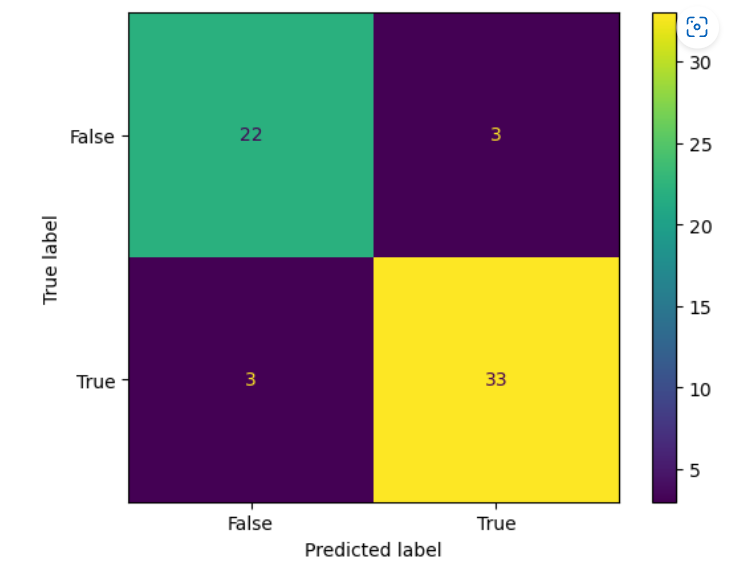
1.The majority of the patients are older than 40.

2.In the dataset there are 2 times more males than females.

3.54.3% of the patients have a high risk of heart attack, 45.7% - low risk of a heart attack.

4. Age distribution of patients from the high-risk group has a peak at 55-56 ages.

5. The major factors influencing heart attacks are: Chest pain, Maximum heart rate achieved, Exercise-induced angina, ST depression induced by exercise relative to rest and Number of major vessels

1. Feature scaling done on the dataset using StandardScaler from the sklearn library.
2. Data Preparation is done using train\_test\_split.
3. Two models were built using logistic regression. And Random Forest Classifier. The output shows 88.52% in logistic regression and 93. 44 % in random forest classifier.
4. Classification report on logistic regression shows 89% accuracy in the f1-score. And in the random forest, it is 93% accuracy in the f1-score.
5. By using GridSearchCV in the random forest classifier classification report shows a total 90% accuracy in the f1-score.
6. Confusion matrix generates a plot for evaluating the performance of a binary classification model. The confusion matrix provides valuable insights into the model’s ability to correctly classify instances into their respective classes. 
7. Finally last plot shows model comparison on bar plot. 