HEART-ATTACK PREDICTION

**PROBLEM STATEMENT**

The most prevalent disease in the modern world is heart attacks, often known as "myocardial infections." Our daily routines, eating patterns, and lifestyle behaviors all contribute to this. Chest discomfort is the most prevalent symptom of a heart attack. Other symptoms include high blood pressure, nausea, drowsiness, and shortness of breath. The severity of symptoms differs from individual to person. When the blood supply to the heart is restricted, a heart attack occurs. Symptoms can be subtle, severe, or even undetectable.

**OBJECTIVE**

The objective is heart attack prediction, and our goal is to investigate which supervised machine learning methods are best suited to solve it.

**OUR APPROACH**

**Dataset: Heart Disease Dataset**

This is a study on heart attacks conducted in the United States, the United Kingdom, Switzerland, and Hungary. The data collection has 1190 instances and 12 variables (11 integers and 1 decimal). These datasets were created by integrating many datasets that were previously available separately but had not been combined. Cleveland, Hungarian, Switzerland, Long Beach VA, and the Stalog (Heart) Data Set were used. It has 11 characteristics and a single target. Sex, age, chest pain, resting blood pressure, cholesterol, fasting blood pressure, resting ecg, max heart rate, exercise angina, old peak, and st slope are the characteristics.

Model Training

SUPERVISED LEARNING:

Exploration

Feature Engineering

Train-test split

Cleaning and imputation

DATA PROCESSING:

DATA COLLECTION:

Heart disease dataset from Kaggle.

OBJECTIVE:

Create a predictive model that can predict heart attack from different symptoms.

Model Evaluation

**DATA PREPROCESSING**

The dataset used here to train the prediction model was heart disease dataset from Kaggle. It contains data about all the main features that contribute to find whether the person has heart attack or not.

* **EXPLORATORY DATA ANALYSIS**

The Exploratory Data Analysis (EDA) is all about analyzing the dataset and summarizing the key insights and characteristics of the data.

**EDA Checklist:-**

1. Understanding the dataset and its shape.
2. Checking the datatype of each columns.
3. Categorical and Numerical Columns.
4. Checking for missing values.
5. Descriptive summary of the dataset.
6. Check the distribution in target variables
7. Groupby for classification problems.

* **Understanding the dataset and its shape:-**

This dataset consists of 11 features and a target variable. It has 6 nominal variables and 5 numeric variables. The detailed description of all the features are as follows:

1. **Age:** Patients Age in years (Numeric)
2. **Sex:** Gender of patient (Male - 1, Female - 0) (Nominal)
3. **Chest Pain Type:** Type of chest pain experienced by patient categorized into

1 typical,

2 typical angina,

3 non- anginal pain,

4 asymptomatic (Nominal)

1. **resting bp s:** Level of blood pressure at resting mode in mm/HG (Numerical)
2. **cholesterol:** Serum cholesterol in mg/dl (Numeric)
3. **fasting blood sugar:** Blood sugar levels on fasting > 120 mg/dl represents as 1 in case of true and 0 as false (Nominal)
4. **resting ecg :** Result of electrocardiogram while at rest are represented in 3 distinct values

0: Normal

1: Abnormality in ST-T wave

2: Left ventricular hypertrophy (Nominal)

1. **max heart rate:** Maximum heart rate achieved (Numeric)
2. **exercise angina:** Angina induced by exercise 0 depicting NO 1 depicting Yes (Nominal)
3. **oldpeak:** Exercise induced ST-depression in comparison with the state of rest (Numeric)
4. **ST slope:** ST segment measured in terms of slope during peak exercise

0: Normal

1: Upsloping

2: Flat

3: Downsloping (Nominal)

Target variable

1. **target:** It is the target variable which we have to predict 1 means patient is suffering from heart attack and 0 means patient is normal.

Csv file contain 1190 rows and 12 columns. This is a Binary-Label Classification Problem.

* **Checking the datatype of each columns:-**

Checked the datatype of each column. Out of 12 features, we have 11 int types and only one with the float data types.

* **Categorical and Numerical Columns:-**

Numerical Columns are age, resting blood pressure, cholesterol, max heart rate, old peak.

Nominal Columns are Sex, chest pain, fasting blood pressure, resting ecg, exercise angina, st slope and target.

* **Checking for missing values:-**

There is no missing values in the dataset.

Data Cleaning is the process of finding and correcting the inaccurate/incorrect data that are present in the dataset. One such process needed is to do something about the values that are missing in the dataset. In real life, many datasets will have many missing values, so dealing with them is an important step.

**Why do you need to fill in the missing data?** Because most of the machine learning models that you want to use will provide an error if you pass NaN values into it. The easiest way is to just fill them up with 0, but this can reduce your model accuracy significantly.

For filling missing values, there are many methods available. The different methods that you can use to deal with the missing data are

1. Deleting the columns with missing data
2. Deleting the rows with missing data
3. Filling the missing data with a value – Imputation
4. Imputation with an additional column
5. Filling with a Regression Model

In a missing value case, we will be filling the missing values with a certain number.

The possible ways to do this are:

1. Filling the missing data with the mean or median value if it’s a numerical variable.
2. Filling the missing data with mode if it’s a categorical value.
3. Filling the numerical value with 0 or -999, or some other number that will not occur in the data. This can be done so that the machine can recognize that the data is not real or is different.
4. Filling the categorical value with a new type for the missing values.

You can use the fillna() function to fill the null values in the dataset.

Here there is no missing values. So we don’t using any of these methods.

* **Descriptive summary of the dataset:-**

Statistical summary of the data – Descriptive Statistics.

Here check whether mean and median (50th Percentile) are close to each other.

Descriptive statistics can give you great insight into the shape of each attribute.

The *describe()* function on the Pandas DataFrame lists 8 statistical properties of each attribute:

* Count
* Mean
* Standard Devaition
* Minimum Value
* 25th Percentile
* 50th Percentile (Median)
* 75th Percentile
* Maximum Value

**SKEWNESS**

Skewness is the difference between mean and median. There are 3 types of skewness are present:-

* 1. Left skewed Data (Mean<Median) – Negatively Skewed
  2. Right skewed Data (Mean>Median) – Positively Skewed
  3. Normal Distribution (Mean=Median=Mode) – Most of the values are situated in middle.

If there is any skewness present, we need to transform the data. Mostly these are done in Regression Method. In classification, Random Forest, Decision Tree algorithms are present. They will handle this. So in regression problems we need to take care of skewness. Transform the skewed data to a normal distribution. In a normal distribution most of the values are lies on a central region, so we don’t have outliers and model can make quick predictions.

Here,

**Positively skewed data (mean>median) 5 columns**

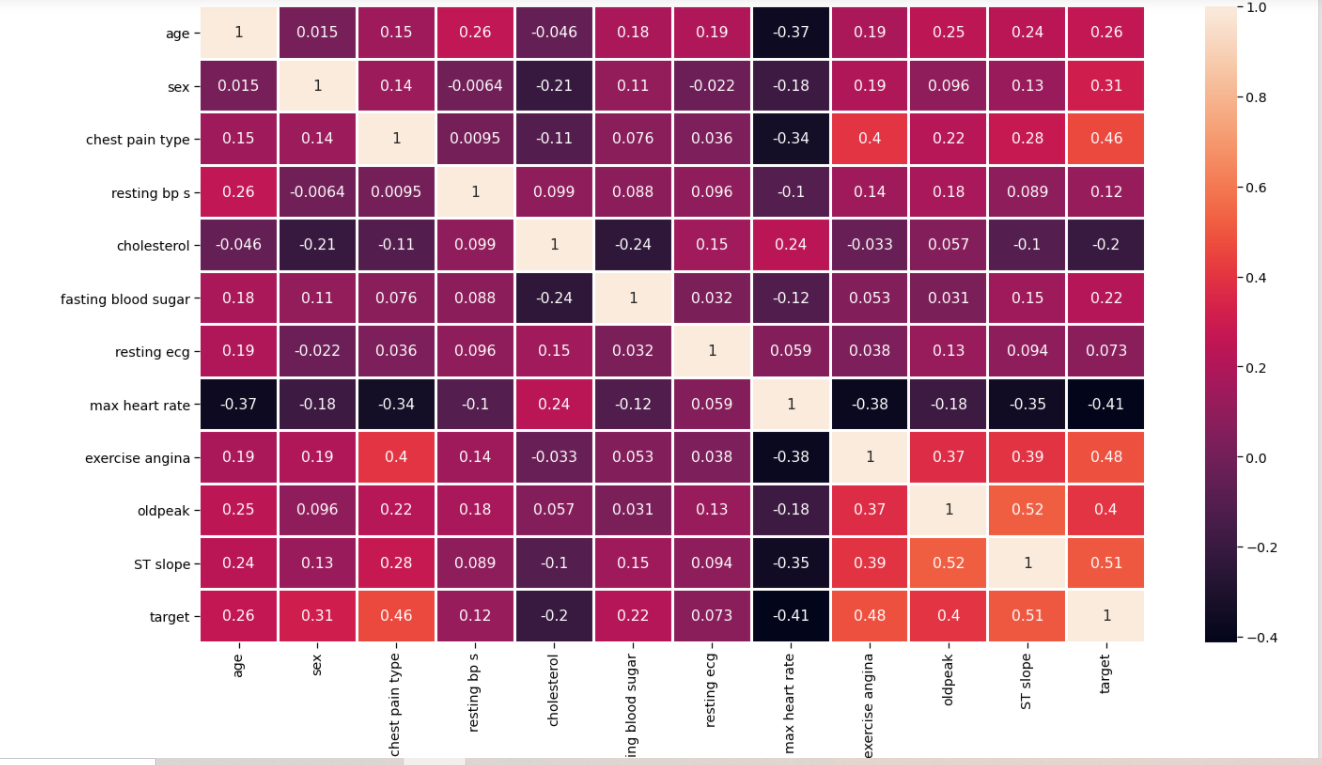
resting bp s, fasting blood sugar, resting ecg, excercise angina, old peak

**Negatively skewed data (mean<median) 7 column**

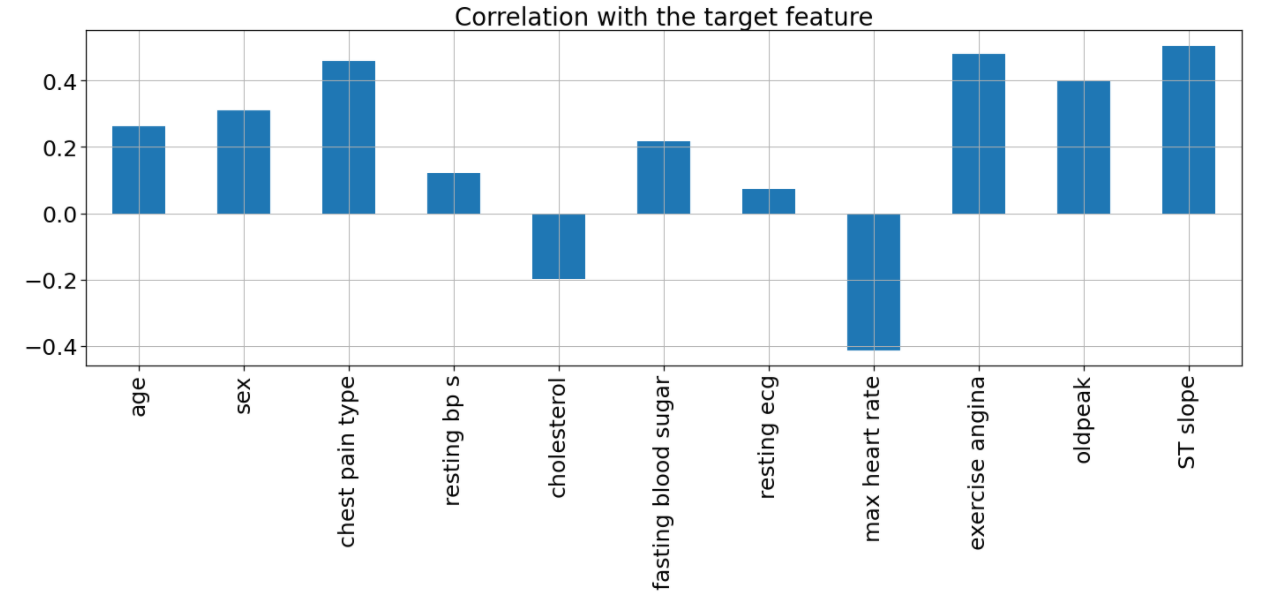
age, sex, chest pain type, cholesterol, max heart rate, ST slope, target

Inference :- The dataset contains more Negatively skewed data

* **Check the correlation between various features. - CORRELATION MATRIX**

.

* **Check the correlation of the target variable.**

****

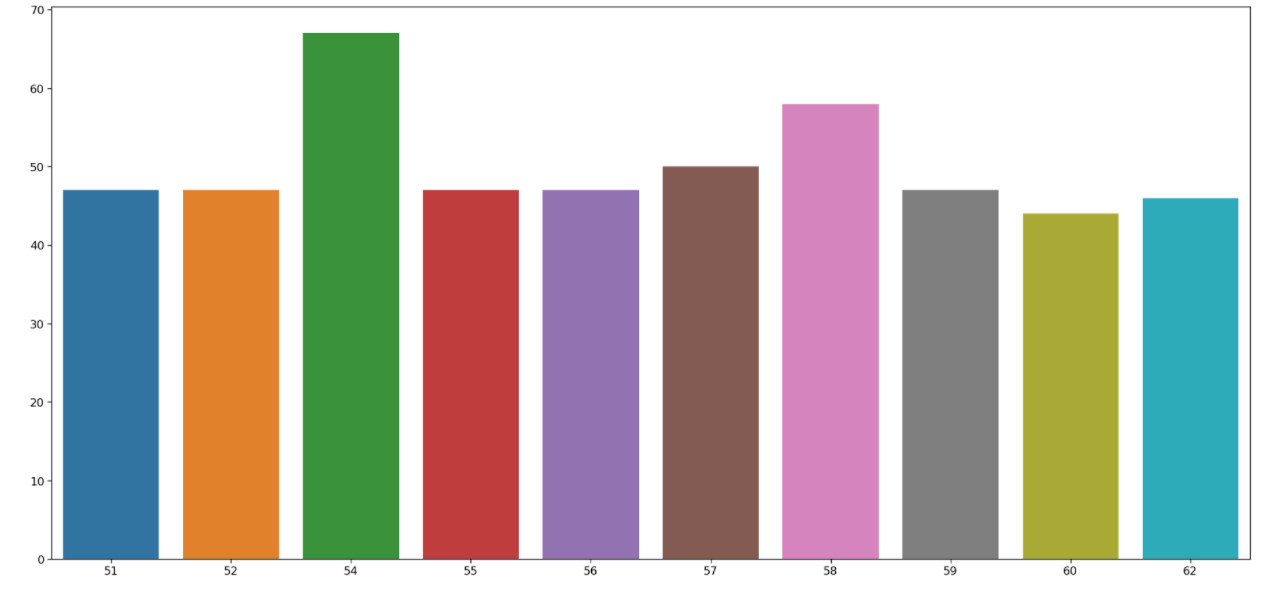
Inference: Insights from the above graph are:

1. Two features ( “cholesterol”, “max heart rate” ) are negatively correlated with the target feature.
2. Other features are positively correlated with the target feature.

# **Analysis of the individual features which comprises both univariate and bivariate analysis**.

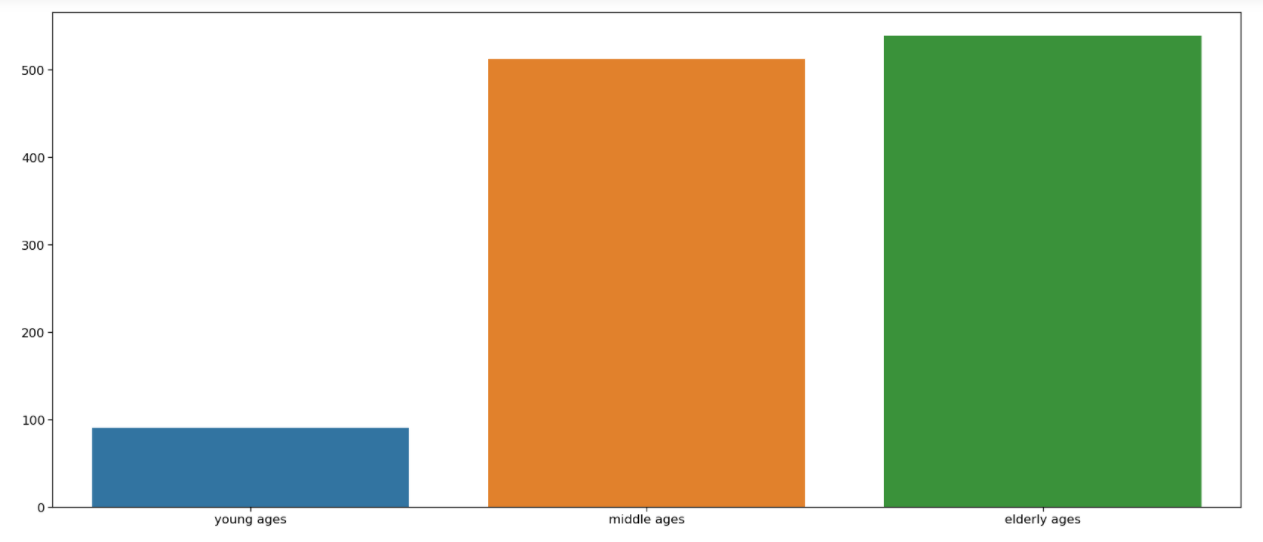
# **Age (“age”) Analysis**

Here we will be checking the 10 ages and their counts.



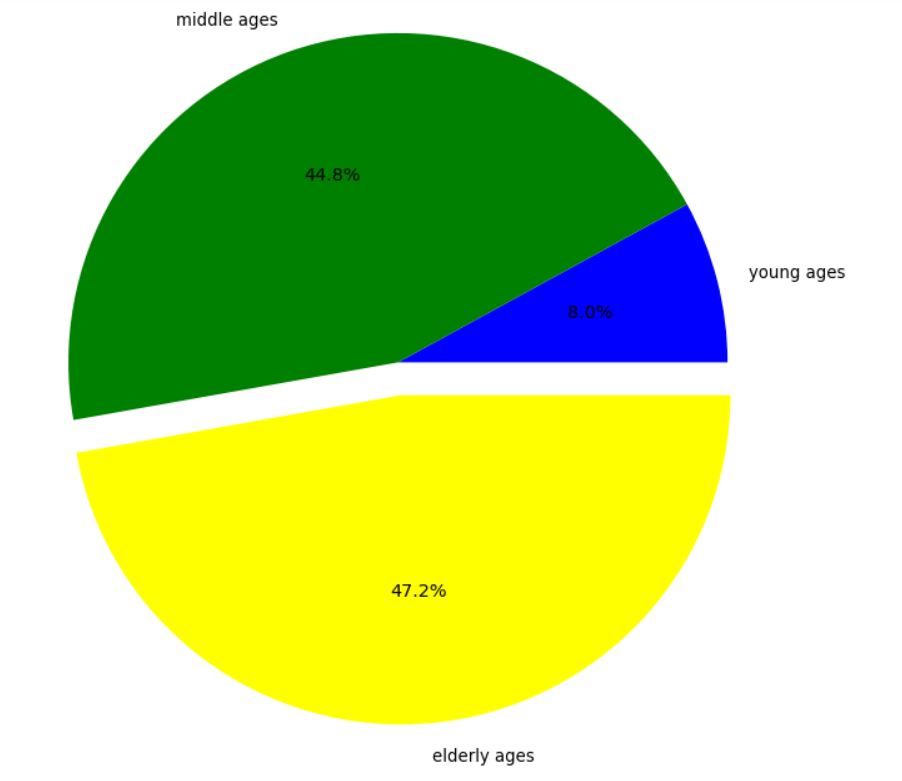
Inference: Here we can see that the 54 age column has the highest frequency.

We should divide the Age feature into three parts – “Young”, “Middle” and “Elder”

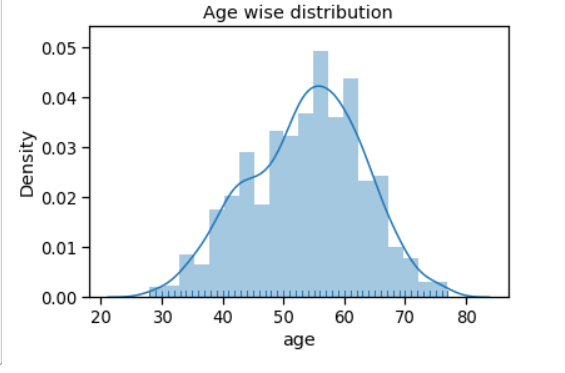


Inference: Here we can see that elder people are the most affected by heart disease and young ones are the least affected.

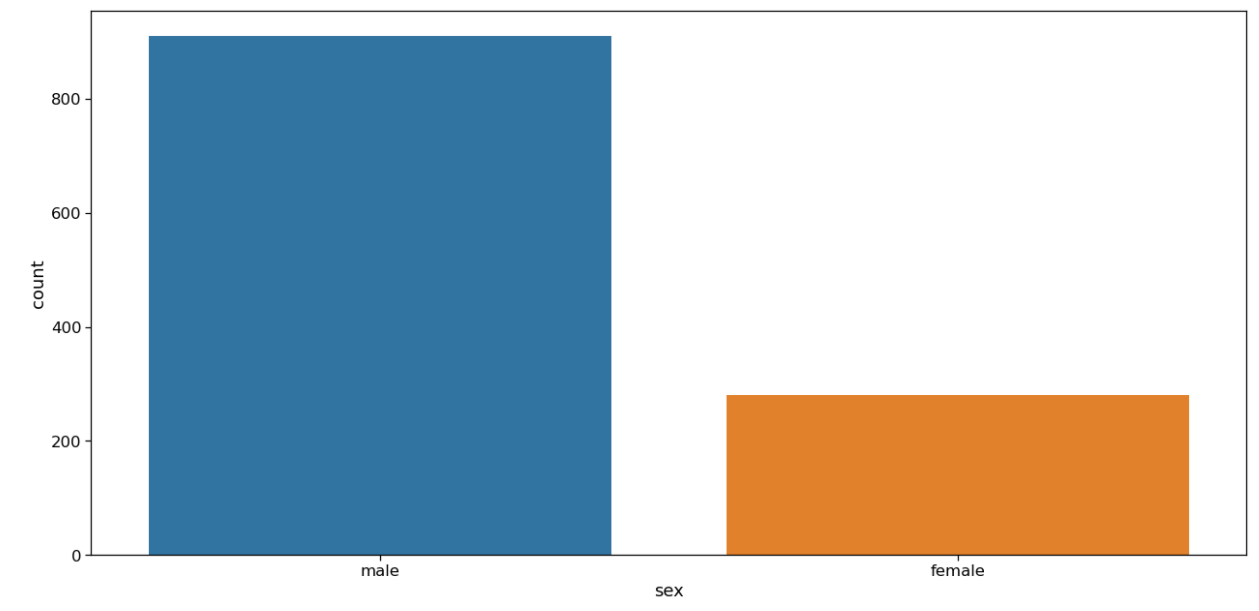
To prove the above inference we will plot the pie chart.

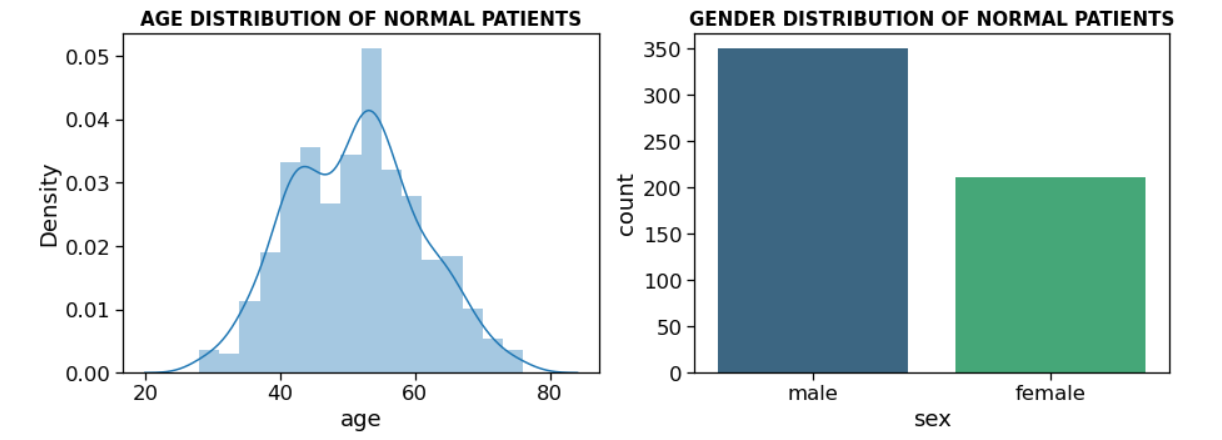
****

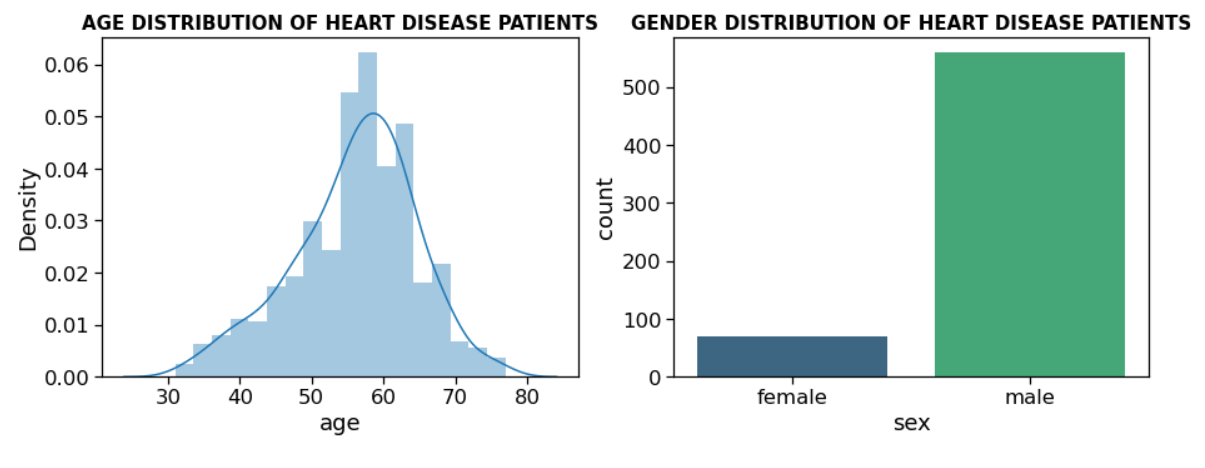
**Age wise Distribution**

****

# **Sex (“sex”) Feature Analysis**

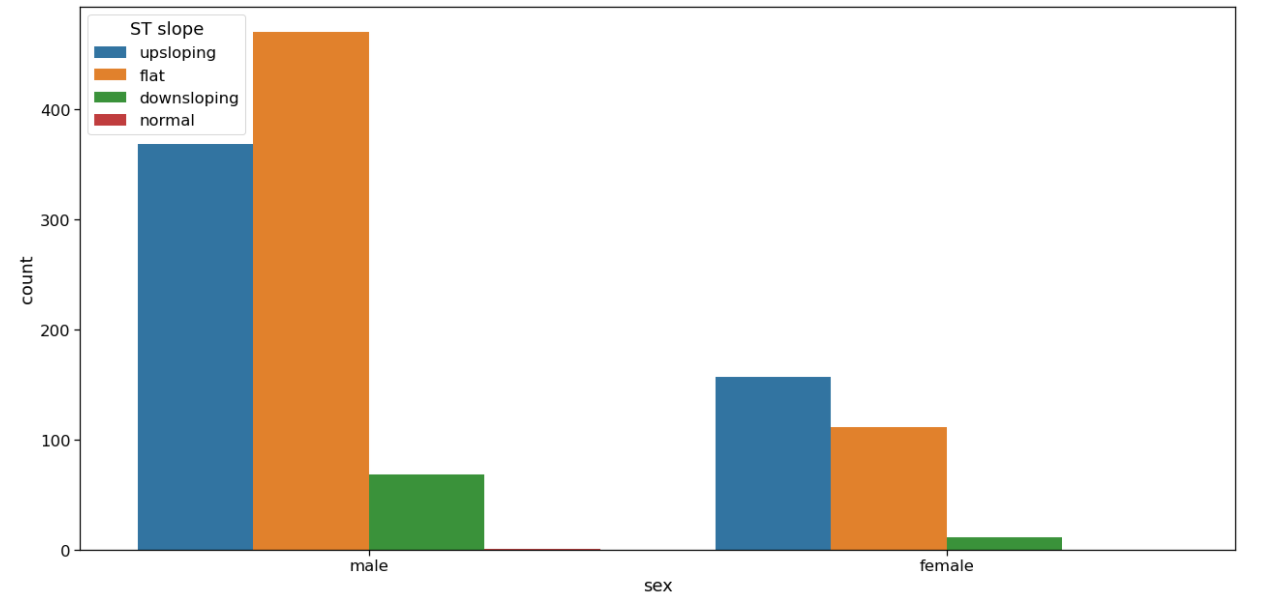




****

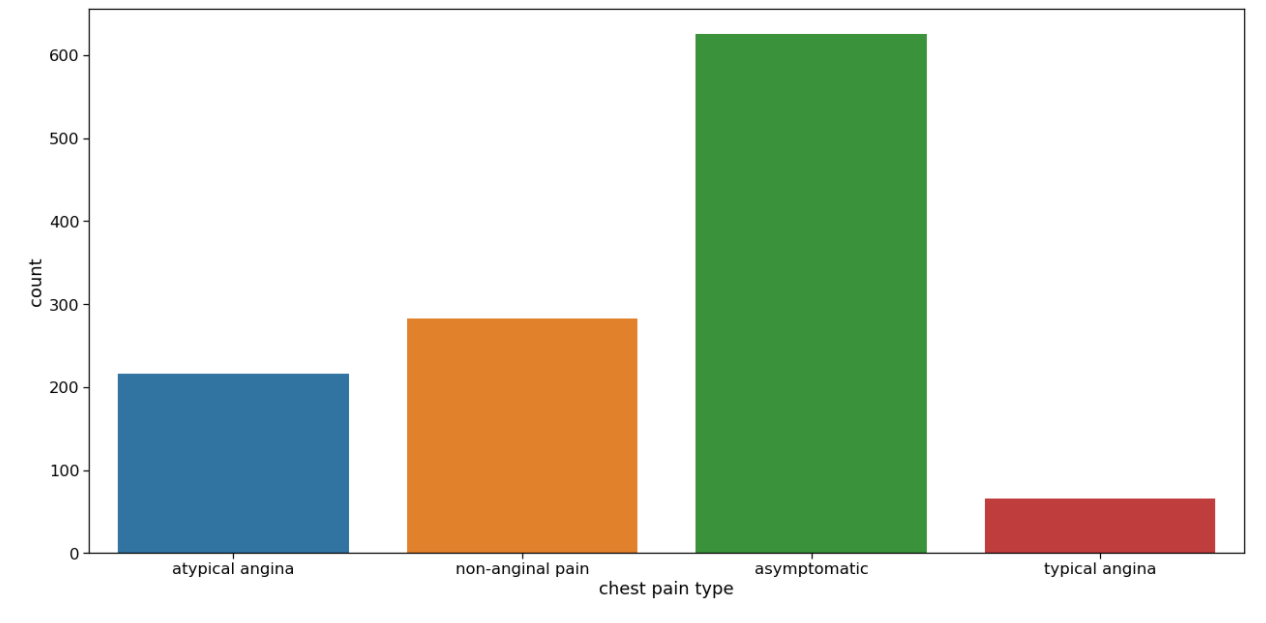
Inference: Males (1) have higher cases compared to Females(0)

# **Plot the relation between sex and slope.**



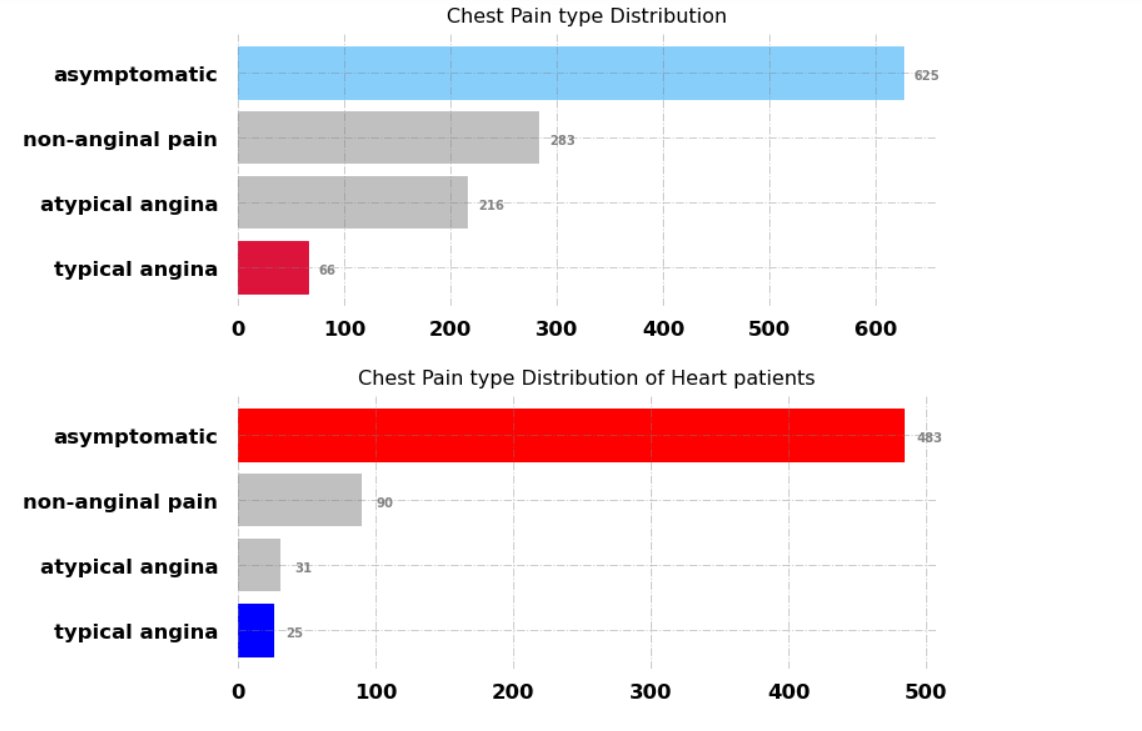
Inference: Here it is clearly visible that the slope value is higher in the case of males(1).

# **Chest Pain Type(“chest pain type”) Analysis**

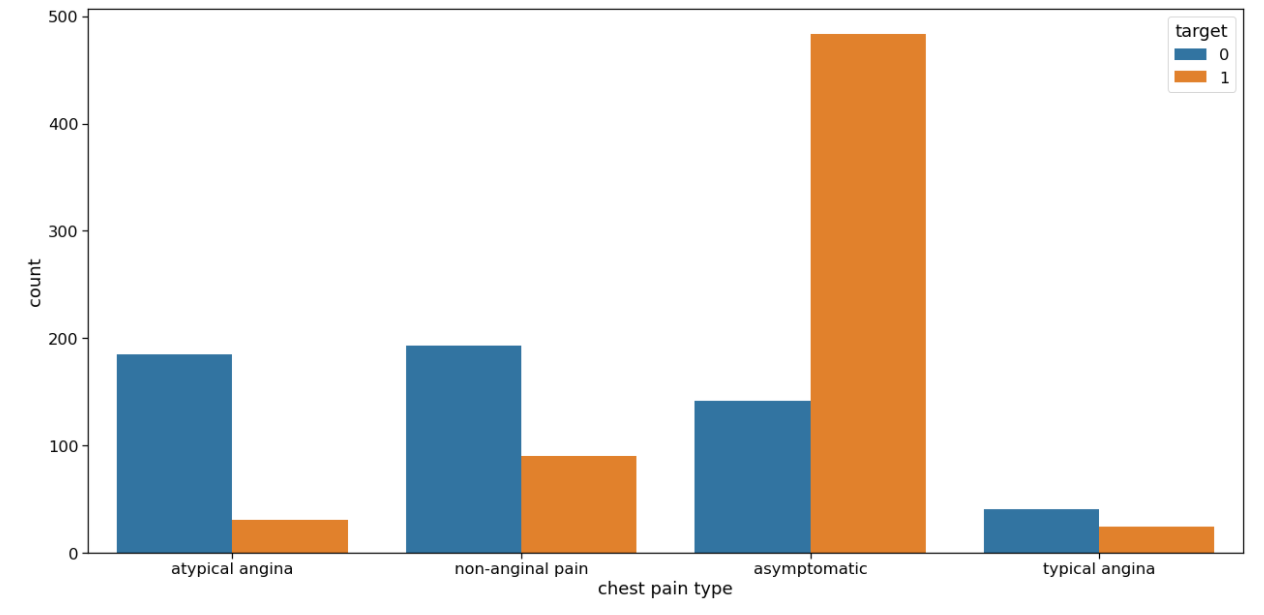
****

Inference: As seen, there are 4 types of chest pain

1. status at least
2. condition slightly distressed
3. condition medium problem
4. condition too bad

****

**Analyzing chest pain type vs target column**

****

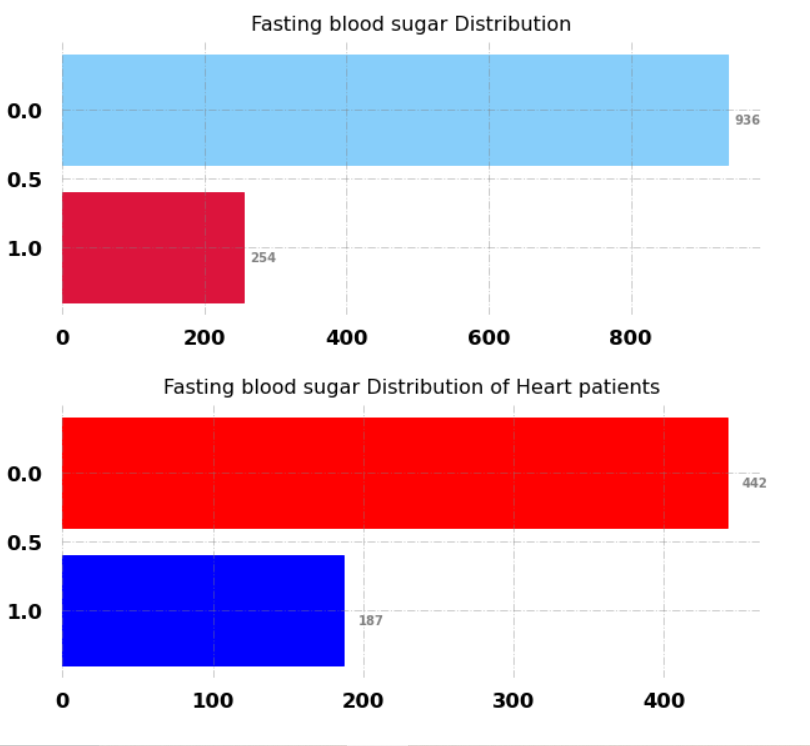
Inference: From the above graph we can make some inferences,

People having the least chest pain are not likely to have heart disease.

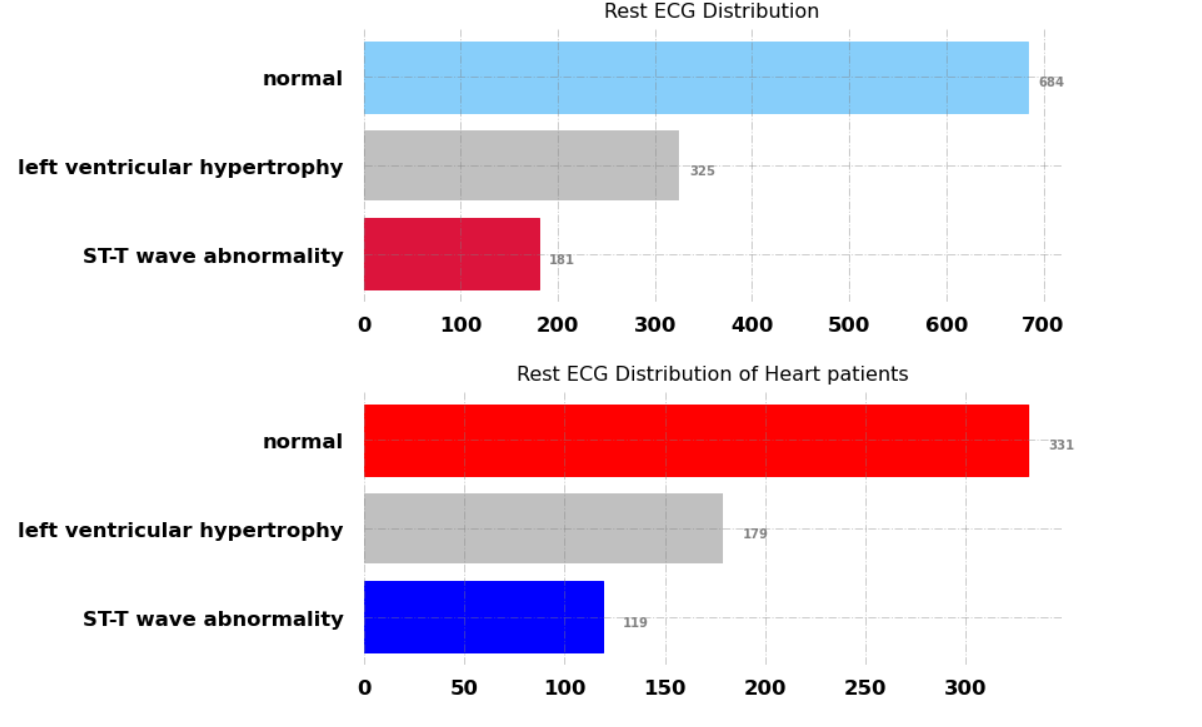
People having severe chest pain are likely to have heart disease.

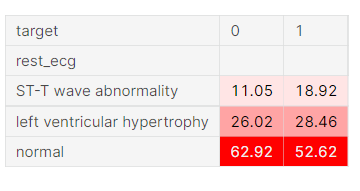
Elderly people are more likely to have chest pain.

# **Fasting blood sugar Analysis**

****

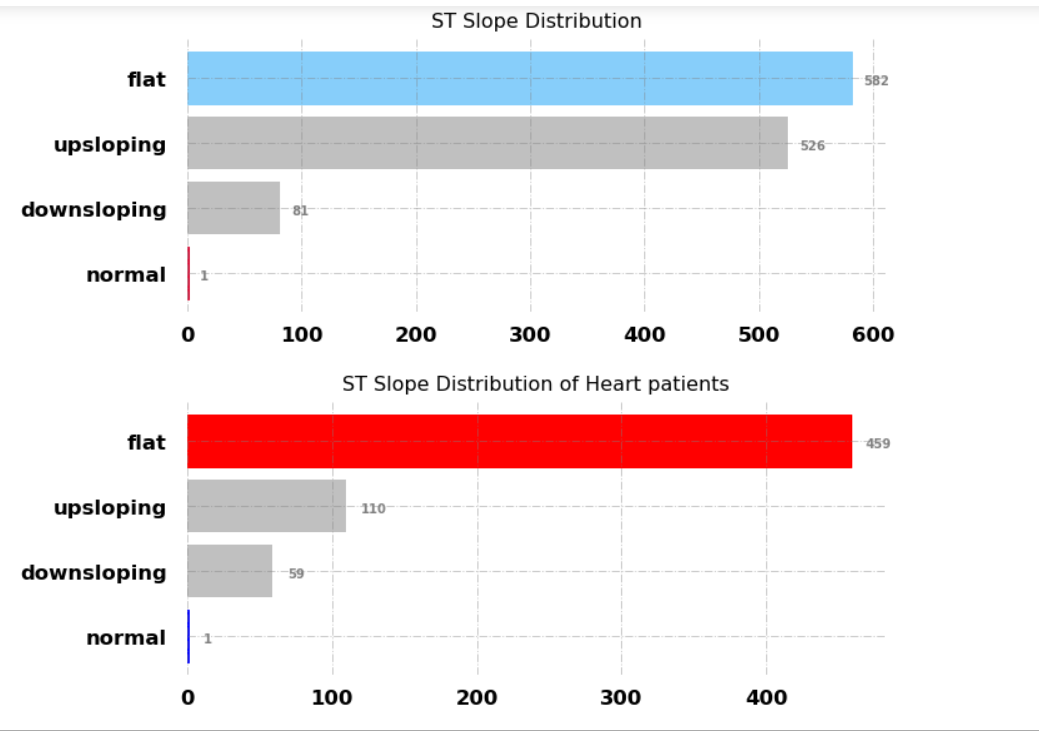
# **Resting ECG Analysis**

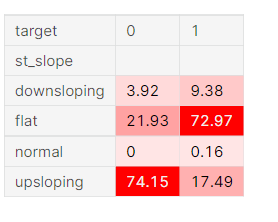




An electrocardiogram records the electrical signals in your heart. It's a common test used to detect heart problems and monitor the heart's status in many situations. Electrocardiograms — also called ECGs or EKGs. But ECG has limits. It measures heart rate and rhythm—but it doesn’t necessarily show blockages in the arteries. That’s why in this dataset around 52% heart disease patients have normal ECG

# **ST Slope Analysis**





The ST segment /heart rate slope (ST/HR slope), has been proposed as a more accurate ECG criterion for diagnosing significant coronary artery disease (CAD) in most of the research papers.

As we can see from above plot upsloping is positive sign as 74% of the normal patients have upslope where as 72.97% heart patients have flat sloping.

* **Check the distribution in target variable:-**

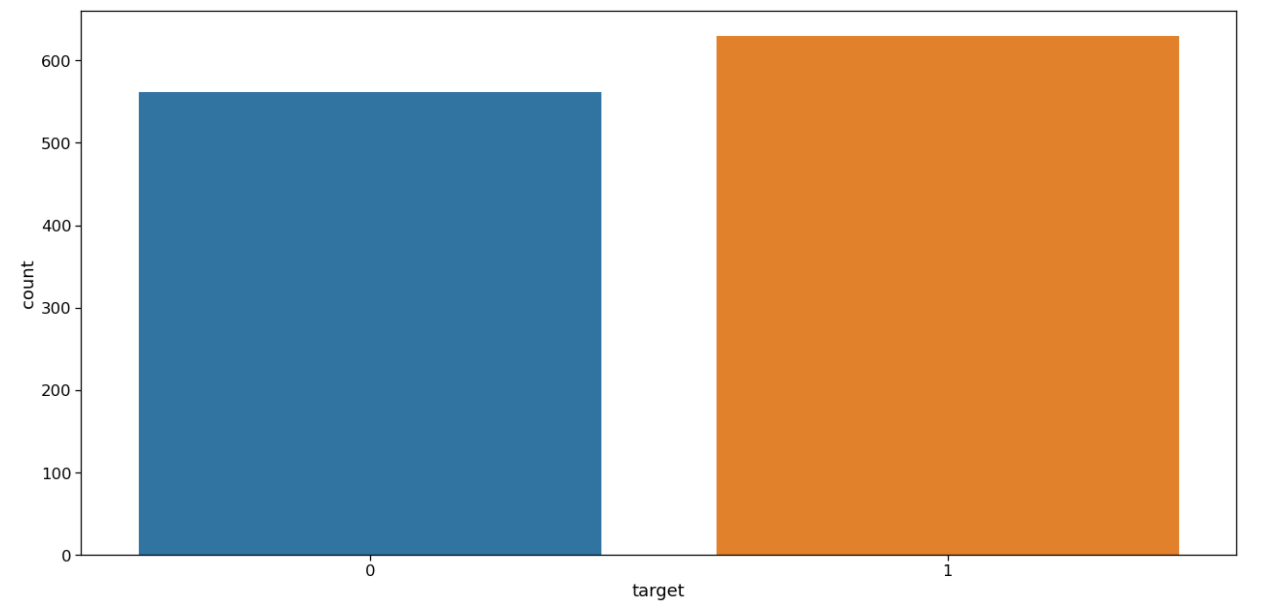
Checked the distribution in target variable and got

**1 629**

**0 561**

* **Groupby for classification problems:-**

Grouped the data based on Target.



INFERENCE:- For most of the Features, the mean values are higher on patient is suffering from heart risk cases (1) and lower for NORMAL (0) cases.

The ratio between 1 and 0 is much less than 1.5 which indicates that the target feature is not imbalanced. So for a balanced dataset, we can use accuracy\_score as evaluation metrics for our model.

**SUMMARY OF EDA**

1. No missing values

2. All are numerical values

3. Median is slightly more than mean for most of the features. So it is left skewed.

4. Not imbalanced dataset

5. Mean of most features are clearly larger for patients suffering from heart risks cases compared to Normal cases.

# Feature Engineering

First we will be remove the target column from our set of features,

then we will categorize all the categorical variables using the get dummies method which will create a separate column for each category suppose X variable contains 2 types of unique values,

then it will create 2 different columns for the X variable.

Now we will use the standard scaler method to scale down the data so that it won’t raise the outliers also dataset which is scaled to general units leads to having better accuracy.

**Standard Scaler**

StandardScaler comes into play when the characteristics of the input dataset differ greatly between their ranges, or simply when they are measured in different units of measure.

StandardScaler removes the mean and scales the data to the unit variance. However, outliers have an influence when calculating the empirical mean and standard deviation, which narrows the range of characteristic values.

These differences in the initial features can cause problems for many machine learning models.

In Machine Learning, StandardScaler is used to resize the distribution of values ​​so that the mean of the observed values ​​is 0 and the [standard deviation](https://en.wikipedia.org/wiki/Standard_deviation) is 1.

**CHOOSE RIGHT MODEL**

1. The data has splits into 2 sections and prepared for Feature Selection:

* First section is used for training our model i.e. Train Set, will be the majority of the dataset (70% of dataset)
* Second section will be used for the evaluation of the trained model’s performance [Also called Validation Set] (30% of dataset)

1. Tried different types of algorithms to train the classifier and choose the best one.

**Algorithms used: KNN, Random Forest, Decision Tree.**

|  |  |
| --- | --- |
| **ALGORITHMS** | **ACCURACY** |
| **Random Forest** | **87%** |
| **KNN** | **86%** |
| **Decision Tree** | **85%** |

**MODEL EVALUATION**

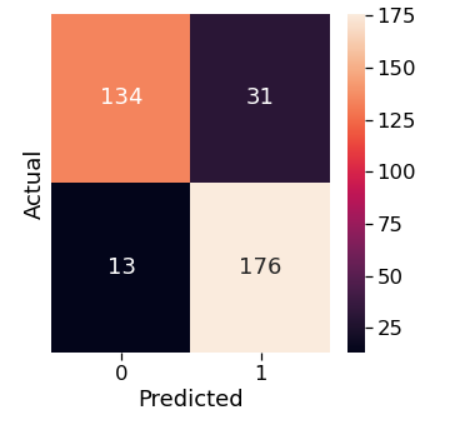
Random Forest Classifier is a tree style algorithm and it has been very effective for many machine learning problems.

1. **ACCURACY**

RF Train\_data : 0.847457627118644

RF Validation Data : 0.8757062146892656

1. **CONFUSION MATRIX**



**CLASSIFICATION REPORT**

Classification Report :

precision recall f1-score support

0 0.91 0.81 0.86 165

1 0.85 0.93 0.89 189

accuracy 0.88 354

macro avg 0.88 0.87 0.87 354

weighted avg 0.88 0.88 0.87 354

**BUILD PREDICTIVE SYSTEM**

The predictive model which successfully predicted whether the given data is Heart Risk or Normal Condition. The predictive model successfully predicted 2 labels.

# Conclusion on Heart Attack Prediction

1. We did data visualization and data analysis of the target variable, age features, and whatnot along with its univariate analysis and bivariate analysis.
2. We also did a complete feature engineering part in this article which summons all the valid steps needed for further steps i.e. model building.
3. From the above model accuracy, RandomForest is giving us the accuracy which is 87%.
4. We have also interpreted second best performing algo i.e., random forest algorithm
5. The top 5 most contribution features are **st\_slope\_upsloping, st\_slope\_flat, excercise\_induced\_angina, sex** and **cholesterol**
6. We find that the ***higher value of st\_slope\_upsloping reduces the risk of heart disease while lower values increases the chances of heart disease where as in case of st\_slope\_flat and excercise\_induced\_angina their higher values are riskier for heart disease while lower values are safe***

Research questions:

1. Which age group and gender are most likely to suffer a heart attack?

The elder people who has 54 years are the most affected by heart disease and young ones are the least affected.

1. What are the most prevalent health problems among those who are predisposed to heart attacks?

 High blood pressure, high cholesterol, high fasting blood sugar, and physical inactivity.

1. Is there a link between physical inactivity and the risk of heart disease?

Approximately 35% of coronary heart disease mortality is due to physical inactivity.

1. What is the highest heart rate a patient has ever had?

202

1. What are the rates of co-morbidity in patients with diabetes and high cholesterol?

Having both high blood pressure and diabetes can greatly increase your risk for heart disease

1. How far does the ecg indication proves positive result?

An electrocardiogram records the electrical signals in your heart. It's a common test used to detect heart problems and monitor the heart's status in many situations. Electrocardiograms — also called ECGs or EKGs. but ECG has limits. It measures heart rate and rhythm—but it doesn’t necessarily show blockages in the arteries.Thats why in this dataset around 52% heart disease patients have normal ECG

1. What is the prevalence of asymptomatic heart attacks?

**50% to 80%** of all heart attacks are silent.