*Presence of heart disease in the patient*

**Problem Statement:-**

**Heart Disease** refers to any condition affecting the heart. Unlike cardiovascular disease, which includes problems with the entire circulatory system, heart disease affects only the heart. According to the Centre for Disease Control and Prevention (CDC), heart disease is the leading cause of death in the United States. Around [1 in 4 deaths](https://www.cdc.gov/heartdisease/facts.htm) in the U.S. occur due to heart disease, and it affects all genders as well as all racial and ethnic groups. There are several different types of heart disease, and they affect the heart in different ways. They are as follows:-

1. Congenital heart defects

### Coronary artery disease

### [Arrhythmia](https://www.medicalnewstoday.com/articles/8887) refers

### Dilated cardiomyopathy

### Myocardial infarction

### Heart failure

The symptoms of heart disease depend on the specific type a person has. Also, some heart conditions cause no symptoms at all. The following symptoms may indicate a heart problem:

* [angina](https://www.medicalnewstoday.com/articles/8886.php), or chest pain
* difficulty breathing
* fatigue and light headedness
* swelling due to fluid retention, or [edema](https://www.medicalnewstoday.com/articles/159111)

Some signs and symptoms that [could indicate](https://www.bhf.org.uk/informationsupport/heart-matters-magazine/medical/11-signs-you-might-have-heart-disease) heart attack include:

* chest pain
* breathlessness
* heart palpitations
* nausea
* stomach pain
* sweating
* arm, jaw, back, or leg pain
* a choking sensation
* swollen ankles
* fatigue
* an irregular heartbeat

Heart attack can lead to cardiac arrest, which is when the heart stops and the body can no longer function. A person needs immediate medical attention if they have any symptoms of heart attack.

The target is to find the presence of heart disease in the patient with different attributes of the datasets. To Predict whether a patient should be diagnosed with Heart Disease is a binary outcome.

* Positive (+) = 1, patient diagnosed with Heart Disease
* Negative (-) = 0, patient not diagnosed with Heart Disease
* Experiment with various Classification Models & see which yields greatest accuracy.
* Examine trends & correlations within our data
* Determine which features are most important to Positive/Negative Heart Disease diagnosis.

**Data analysis:-**

Our Predictor (Positive or Negative diagnosis of Heart Disease) is determined by 13 features:

1. age (#)  
2. sex : 1= Male, 0= Female (Binary)  
3. (cp)chest pain type (4 values -Ordinal):

Value 1: typical angina, Value 2: atypical angina, Value 3: non-anginal pain, Value 4: asymptomatic  
4. (trestbps) resting blood pressure (#)  
5. (chol) serum cholesterol in mg/dl (#)  
6. (fbs)fasting blood sugar > 120 mg/dl(Binary)(1 = true; 0 = false)  
7. (restecg) resting electrocardiography results(values 0,1,2)  
8. (thalach) maximum heart rate achieved (#)  
9. (exang) exercise induced angina (binary) (1 = yes; 0 = no)  
10. (oldpeak) = ST depression induced by exercise relative to rest (#)  
11. (slope) of the peak exercise ST segment (Ordinal) (Value 1: up sloping , Value 2: flat , Value 3: down sloping )  
12. (ca) number of major vessels (0–3, Ordinal) colored by fluoroscopy  
13. (thal) maximum heart rate achieved — (Ordinal): 3 = normal; 6 = fixed defect; 7 = reversible defect

Our data is of the following 3 types: 1. Continuous (#): which is quantitative data that can be measured, 2.Ordinal Data: Categorical data that has an order to it (0,1,2,3, etc.), 3.Binary Data: data whose unit can take on only two possible states ( 0 &1 ).

**EDA Concluding Remarks:-**

Exploratory data analysis (**EDA**) is used by data scientists to analyse and investigate data sets and summarize their main characteristics, often employing data visualization methods. It can also help to determine if the statistical techniques you are considering for data analysis are appropriate.

**Correlation:** Correlation is a statistical matrix for measuring to what extent different variables are interdependent. Measures to what extent different variables are interdependent. Correlation does not imply causation.

* There is a **positive correlation** between chest pain (cp) & target (predictor).
* There is a **negative correlation** between exercise induced angina (exang) & the target variable.

**Pairplot:** Pairplot are also a great way to immediately see the correlations between all variables. But you will see me make it with only continuous columns from our data, because with so many features, it can be difficult to see each one. So, instead I will make a pairplot with only our continuous features. Chose to make a smaller pairplot with only the continuous variables, to dive deeper into the relationships.

**Cat plot:** ST segment depression occurs because when the ventricle is at rest and therefore repolarized. If the trace in the ST segment is abnormally low below the baseline, this can lead to this Heart Disease. This is supports the plot above because low ST Depression yields people at greater risk for heart disease. While a high ST depression is considered normal & healthy. The “slope” hue, refers to the peak exercise ST segment, with values: 0: up sloping, 1: flat , 2: down sloping). Both positive & negative heart disease patients exhibit equal distributions of the 3 slope categories.

**Violin & Box plot**: The advantages of showing the Box & Violin plots is that it shows the basic statistics of the data, as well as its distribution. These plots are often used to compare the distribution of a given variable across some categories. It shows the median, IQR, & Tukey’s fence. (Minimum, first quartile (Q1), median, third quartile (Q3), and maximum). In addition it can provide us with outliers in our data.

**Violin plot:** We can see that the overall shape & distribution for negative & positive patients differ vastly. Positive patients exhibit a lower median for ST depression level & thus a great distribution of their data is between 0 & 2, while negative patients are between 1 & 3. In addition, we don’t see many differences between male & female target outcomes.

**Box plot**: Positive patients exhibit a heightened median for ST depression level, while negative patients have lower levels. In addition, we don’t see many differences between male & female target outcomes, expect for the fact that males have slightly larger ranges of ST Depression

**Filtering data by positive & negative Heart Disease patient**: From comparing positive and negative patients we can see there are vast differences in means for many of our 13 Features. From examining the details, we can observe that positive patients experience heightened maximum heart rate achieved (thalach) average. In addition, positive patients exhibit about 1/3rd the amount of ST depression induced by exercise relative to rest (oldpeak).

**Pre-processing Pipeline:-**

Helps us get an idea of what data we are working with.

**Description of Datasets:** In description of the datasets, we determine the count, mean, standard deviation, minimum, 25%, 50%, 75% and maximum of each column in the datasets.

There are no missing value for each column. Hence we don’t need to do any data cleaning or data wrangling process.

**Building Machine Learning Models:-**

The **Model Development:** A model can be thought of as a mathematical equation used to predict a value given one or more other values.

Relating one or more independent variables to dependent variables.

Usually the more relevant data you have the more accurate your model will be.

To prepare data for modelling, just remember **ASN (Assign, Split, Normalize).**

**1. Assign**:  The 13 features to X, & the last column to our classification predictor, Y

**2**. **Split:** The data set into the Training set and Test set

**3. Normalize:** Standardizing the data will transform the data so that its distribution will have a mean of 0 and a standard deviation of 1.

**Logistic Regression:** Logistic Regression is a Machine Learning algorithm which is used for the classification problems, it is a predictive analysis algorithm and based on the concept of probability. The hypothesis of logistic regression tends it to limit the cost function between 0 and 1.

precision recall f1-score support

0 0.77 0.67 0.71 30

1 0.71 0.81 0.76 31

accuracy 0.74 61

macro avg 0.74 0.74 0.74 61

weighted avg 0.74 0.74 0.74 61

**KNN (K-Nearest Neighbours):** K-Nearest Neighbours (KNN) is one of the simplest algorithms used in Machine Learning for regression and classification problem. KNN algorithms use data and classify new data points based on similarity measures (e.g. distance function). Classification is done by a majority vote to its neighbours.

precision recall f1-score support

0 0.78 0.70 0.74 30

1 0.74 0.81 0.77 31

accuracy 0.75 61

macro avg 0.76 0.75 0.75 61

weighted avg 0.76 0.75 0.75 61

**SVM (Support Vector Machine):** A support vector machine (SVM) is a supervised machine learning model that uses classification algorithms for two-group classification problems. After giving an SVM model sets of labelled training data for each category, they're able to categorize new text.

precision recall f1-score support

0 0.80 0.67 0.73 30

1 0.72 0.84 0.78 31

accuracy 0.75 61

macro avg 0.76 0.75 0.75 61

weighted avg 0.76 0.75 0.75 61

**Naïve Bayes algorithm:** It is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems. Naïve Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions.

precision recall f1-score support

0 0.79 0.73 0.76 30

1 0.76 0.81 0.78 31

accuracy 0.77 61

macro avg 0.77 0.77 0.77 61

weighted avg 0.77 0.77 0.77 61

**Decision Trees:** A Decision Tree determines the predictive value based on series of questions and conditions. For instance, this simple Decision Tree determining on whether an individual should play outside or not. The tree takes several weather factors into account, and given each factor either makes a decision or asks another question.

precision recall f1-score support

0 0.68 0.70 0.69 30

1 0.70 0.68 0.69 31

accuracy 0.69 61

macro avg 0.69 0.69 0.69 61

weighted avg 0.69 0.69 0.69 61

**Random Forest Regression:** Random Forest Models can be thought of as Bagging, with a slight tweak. Random Forest models decide where to split based on a random selection of features. Rather than splitting at similar features at each node throughout, Random Forest models implement a level of differentiation because each tree will split based on different features. This level of differentiation provides a greater ensemble to aggregate over, ergo producing a more accurate predictor.

precision recall f1-score support

0 0.88 0.70 0.78 30

1 0.76 0.90 0.82 31

accuracy 0.80 61

macro avg 0.82 0.80 0.80 61

weighted avg 0.81 0.80 0.80 61

**XGBoost:** XGBoost is a decision-tree-based ensemble Machine Learning algorithm that uses a gradient boosting framework. In prediction problems involving unstructured data (images, text, etc.) artificial neural networks tend to outperform all other algorithms or frameworks.

precision recall f1-score support

0 0.69 0.67 0.68 30

1 0.69 0.71 0.70 31

accuracy 0.69 61

macro avg 0.69 0.69 0.69 61

weighted avg 0.69 0.69 0.69 61

**Concluding Remarks:-**

**Confusion Matrix:** A good rule of thumb is that any accuracy above 70% is considered good, but be careful because if your accuracy is extremely high, it may be too good to be true (an example of Over fitting). Thus, 80% is the ideal accuracy.

**Feature Importance:** Feature Importance provides a score that indicates how helpful each feature was in our model. The higher the Feature Score, the more that feature is used to make key decisions & thus the more important it is.

**1.** Out of the 14 features we examined, the top 4 significant features that helped us classify between a positive & negative Diagnosis were chest pain type (cp), maximum heart rate achieved (thalach), number of major vessels (ca), and ST depression induced by exercise relative to rest (oldpeak).

**2.** Our machine learning algorithm can now classify patients with Heart Disease. Now we can properly diagnose patients, & get them the help they needs to recover. By diagnosing detecting these features early, we may prevent worse symptoms from arising later.

**3.** Our Random Forest algorithm yields the highest accuracy, 80%. Any accuracy above 70% is considered good, but be careful because if your accuracy is extremely high, it may be too good to be true (an example of Over fitting). Thus, 80% is the ideal accuracy.