Machine Learning
Heart Disease

Nissrine Hatibi

«Predict the presence of heart disease»

→ Overview

Heart disease describes a range of conditions that affect the heart.
The term "heart disease" is often used interchangeably with the term "cardiovascular disease." Cardiovascular disease generally refers to conditions that involve narrowed or blocked blood vessels that can lead to a heart attack, chest pain (angina) or stroke. Other heart conditions, such as those that affect your heart's muscle, valves or rhythm, also are considered forms of heart disease.
Many forms of heart disease can be prevented or treated with healthy lifestyle choices.

→Symptoms:

Cardiovascular disease symptoms may be different for men and women. For instance, men are more likely to have chest pain; women are more likely to have other symptoms along with chest discomfort, such as shortness of breath, nausea and extreme fatigue.

Symptoms can include:

- Chest pain, chest tightness, chest pressure and chest discomfort (angina)
- Shortness of breath
- Pain, numbness, weakness or coldness in your legs or arms if the blood vessels in those parts of your body are narrowed
- Pain in the neck, jaw, throat, upper abdomen or back

→ Risk factors :

Risk factors for developing heart disease include:

- Age.
  Aging increases your risk of damaged and narrowed arteries and weakened or thickened heart muscle.
- Sex.
  Men are generally at greater risk of heart disease. However, women's risk increases after menopause.
- Family history.
  A family history of heart disease increases your risk of coronary artery disease, especially if a parent developed it at an early age (before age 55 for a male relative, such as your brother or father, and 65 for a female relative, such as your mother or sister).
- Smoking.

Nicotine constricts your blood vessels, and carbon monoxide can damage their inner lining, making them more susceptible to atherosclerosis. Heart attacks are more common in smokers than in nonsmokers.

- Certain chemotherapy drugs and radiation therapy for cancer.
  Some chemotherapy drugs and radiation therapies may increase the risk of cardiovascular disease.
- Poor diet.
  A diet that's high in fat, salt, sugar and cholesterol can contribute to the development of heart disease.
- High blood pressure.
  Uncontrolled high blood pressure can result in hardening and thickening of your arteries, narrowing the vessels through which blood flows.
- High blood cholesterol levels.
  High levels of cholesterol in your blood can increase the risk of formation of plaques and atherosclerosis.
- Diabetes.
  Diabetes increases your risk of heart disease. Both conditions share similar risk factors, such as obesity and high blood pressure.
- Obesity.
  Excess weight typically worsens other risk factors.
- Physical inactivity.
  Lack of exercise also is associated with many forms of heart disease and some of its other risk factors, as well.
- Stress.
  Unrelieved stress may damage your arteries and worsen other risk factors for heart disease.
- Poor hygiene.
  Not regularly washing your hands and not establishing other habits that can help prevent viral or bacterial infections can put you at risk of heart infections, especially if you already have an underlying heart condition. Poor dental health also may contribute to heart disease.

Machine Learning - Heart Disease

→Objectif : Predict the presence of heart disease.

→DataSet of Cleveland

→ Read Data:

The meaning of the column headers

1.age: The person's age in years

2.sex: The person's sex (1 = male, 0 = female)

3.cp: The chest pain experienced (Value 1: typical angina, Value 2: atypical angina, Value 3: non-anginal pain, Value 4: asymptomatic)

4.trestbps: The person's resting blood pressure (mm Hg on admission to the hospital)

5.chol: The person's cholesterol measurement in mg/dl

6.fbs: The person's fasting blood sugar (> 120 mg/dl, 1 = true; 0 = false)

7.restecg: Resting electrocardiographic measurement (0 = normal, 1 = having ST-T wave abnormality, 2 = showing probable or definite left ventricular hypertrophy by Estes' criteria)

8.thalach: The person's maximum heart rate achieved

9.exang: Exercise induced angina (1 = yes; 0 = no)

10.oldpeak: ST depression induced by exercise relative to rest ('ST' relates to positions on the ECG plot. See more here)

11.slope: the slope of the peak exercise ST segment (Value 1: upsloping, Value 2: flat, Value 3: downsloping)

12.ca: The number of major vessels (0-3)

13.thal: A blood disorder called thalassemia (3 = normal; 6 = fixed defect; 7 = reversable defect)

14.target: Heart disease (0 = no, 1 = yes)

	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
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1

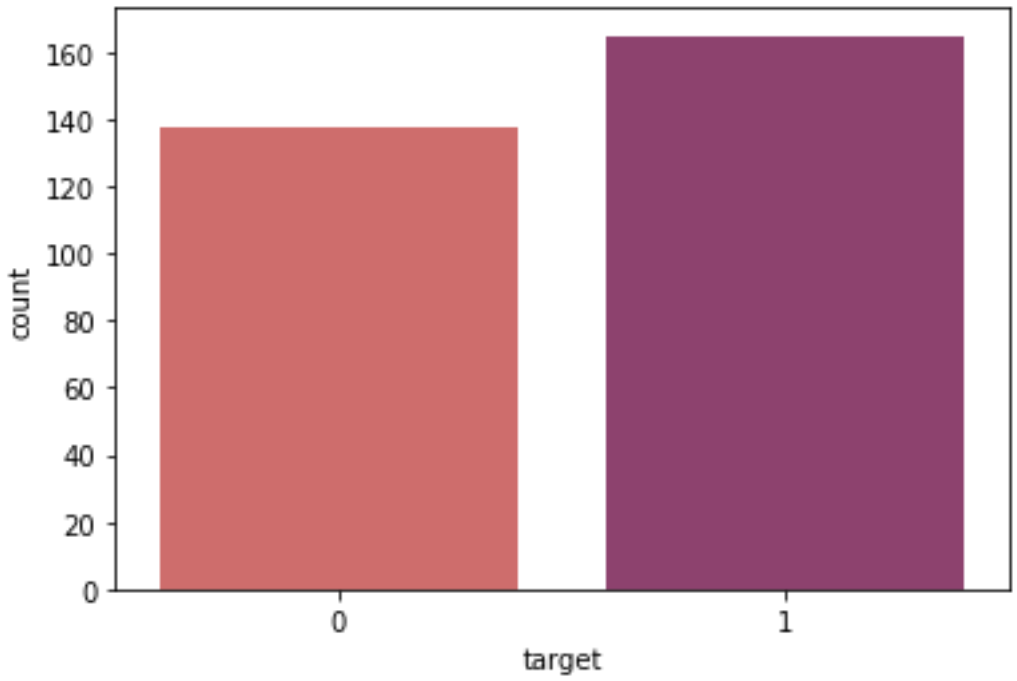
→Describe Data :

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
count	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000
mean	54.366337	0.683168	0.966997	131.623762	246.264026	0.148515	0.528053	149.646865	0.326733	1.039604	1.399340	0.729373	2.313531	0.544554
std	9.082101	0.466011	1.032052	17.538143	51.830751	0.356198	0.525860	22.905161	0.469794	1.161075	0.616226	1.022606	0.612277	0.498835
min	29.000000	0.000000	0.000000	94.000000	126.000000	0.000000	0.000000	71.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	47.500000	0.000000	0.000000	120.000000	211.000000	0.000000	0.000000	133.500000	0.000000	0.000000	1.000000	0.000000	2.000000	0.000000
50%	55.000000	1.000000	1.000000	130.000000	240.000000	0.000000	1.000000	153.000000	0.000000	0.800000	1.000000	0.000000	2.000000	1.000000
75%	61.000000	1.000000	2.000000	140.000000	274.500000	0.000000	1.000000	166.000000	1.000000	1.600000	2.000000	1.000000	3.000000	1.000000
max	77.000000	1.000000	3.000000	200.000000	564.000000	1.000000	2.000000	202.000000	1.000000	6.200000	2.000000	4.000000	3.000000	1.000000

→Explore Data :

- Target

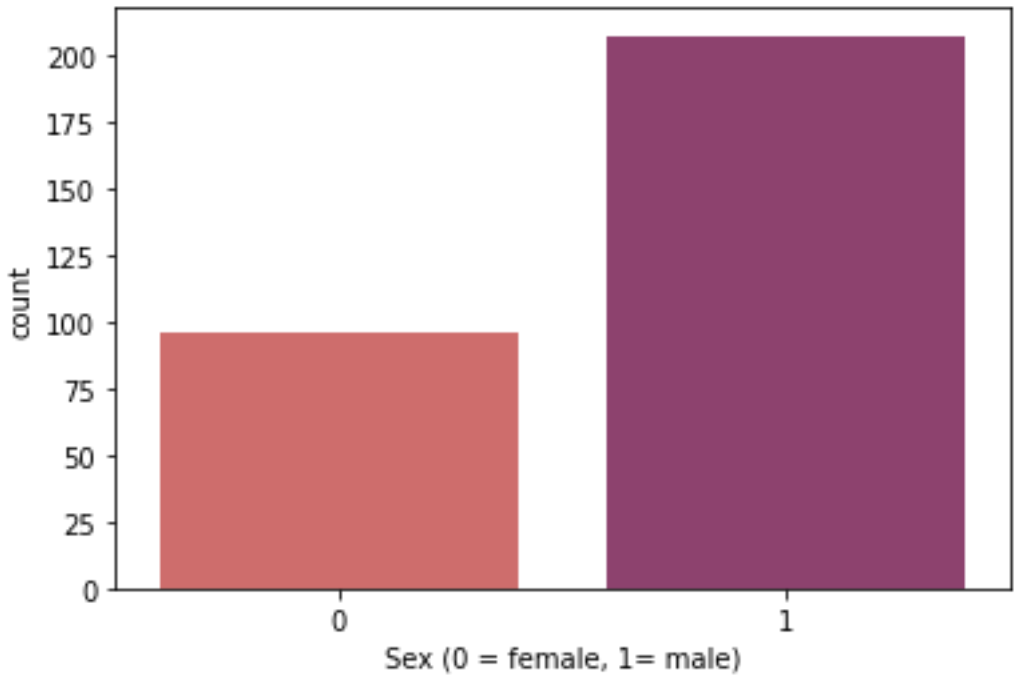
```
1    165
0    138
Name: target, dtype: int64
```




```
Percentage of Patients Haven't Heart Disease: 45.54%
Percentage of Patients Have Heart Disease: 54.46%
```

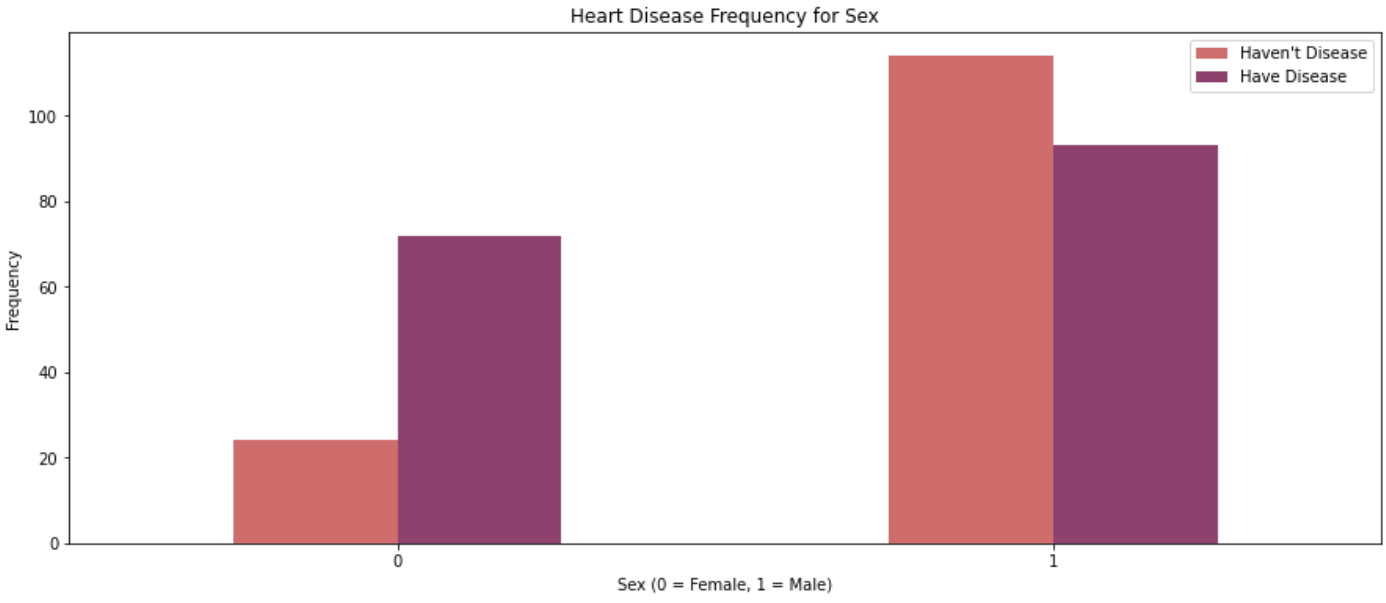
	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal
target													
0	56.601449	0.826087	0.478261	134.398551	251.086957	0.159420	0.449275	139.101449	0.550725	1.585507	1.166667	1.166667	2.543478
1	52.496970	0.563636	1.375758	129.303030	242.230303	0.139394	0.593939	158.466667	0.139394	0.583030	1.593939	0.363636	2.121212

- **Sex**

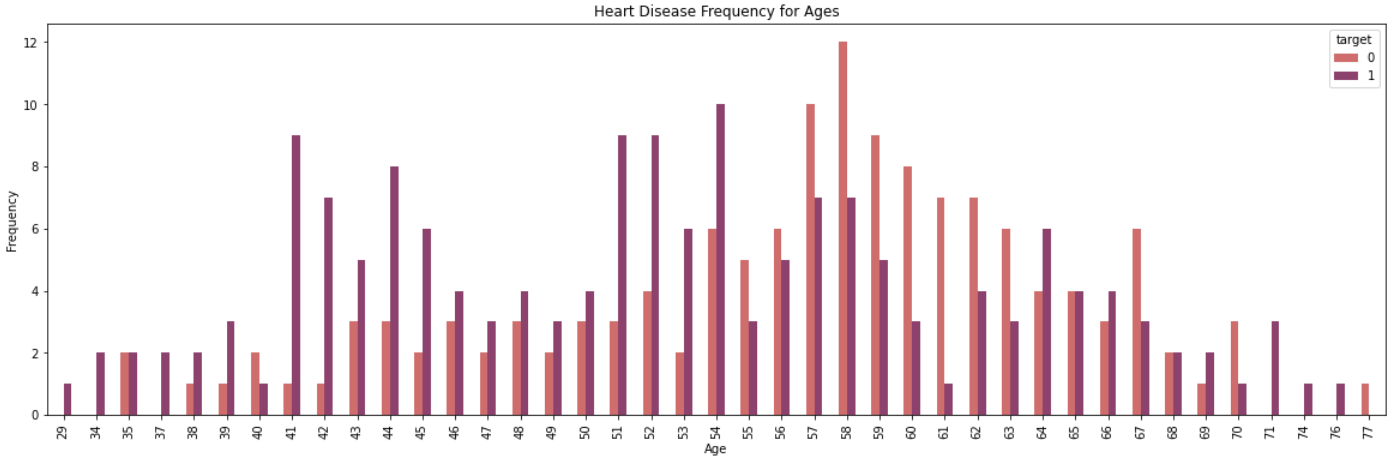



```
Percentage of Female Patients: 31.68%
Percentage of Male Patients: 68.32%
```

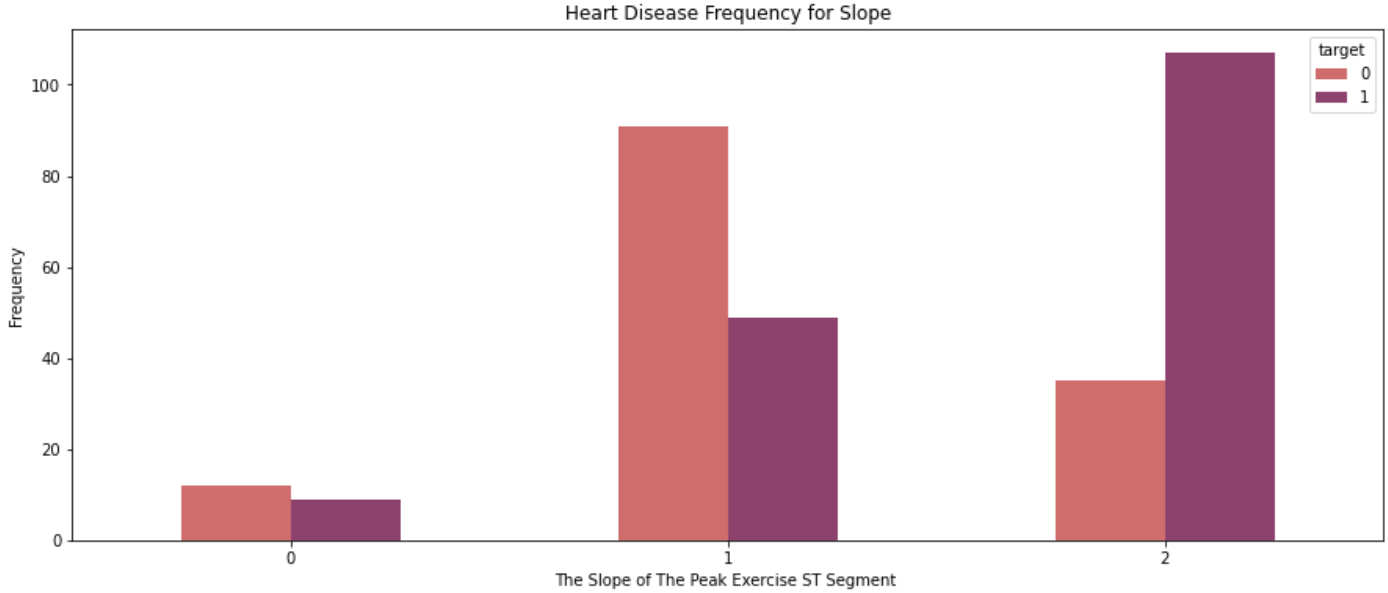
➔ **Heart Disease Frequency for Sex**



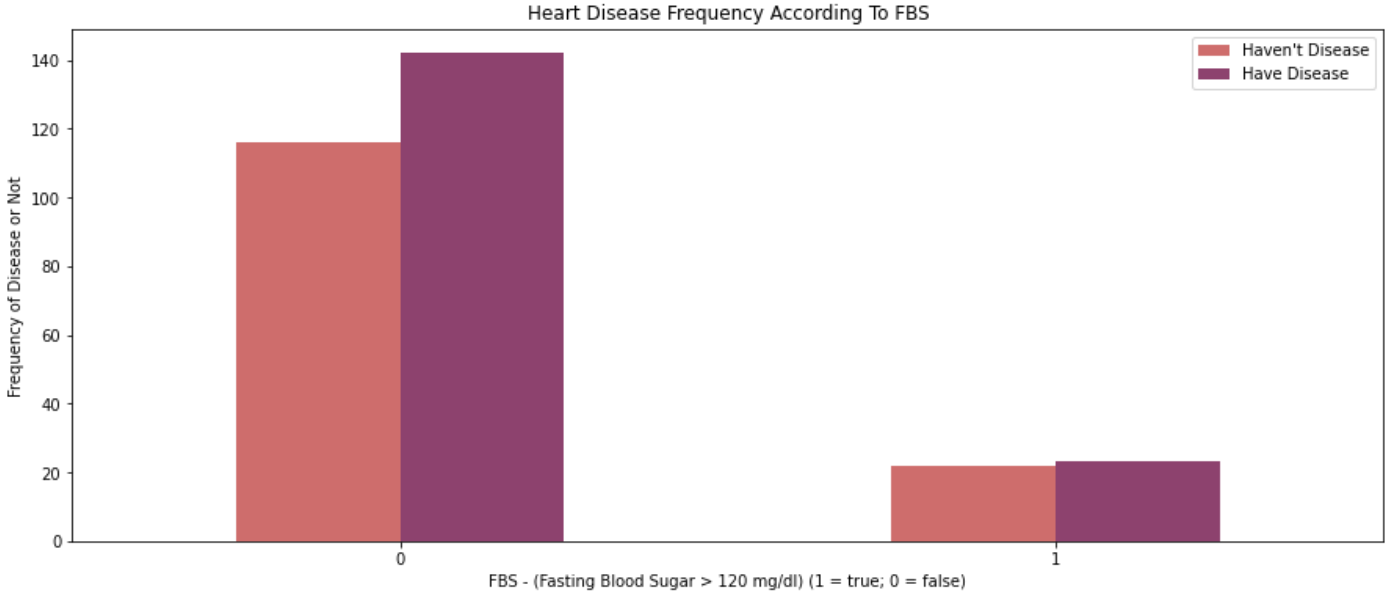
➔ Heart Disease Frequency for Ages



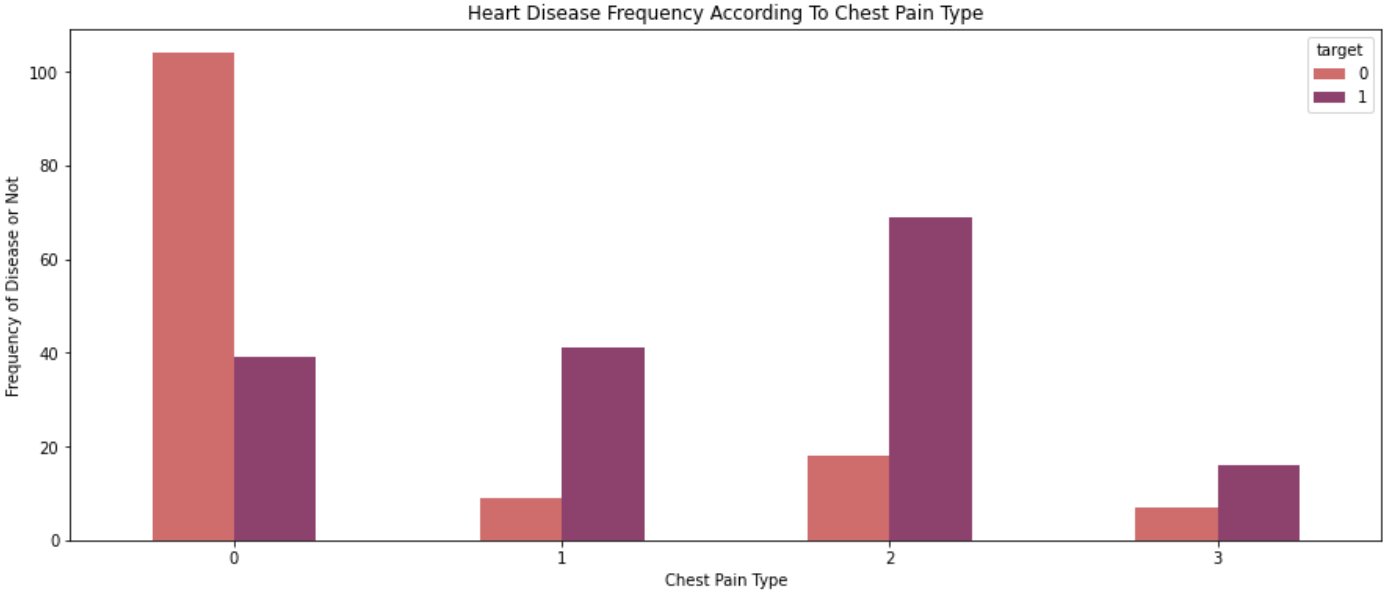
➔ Heart Disease Frequency for Slope



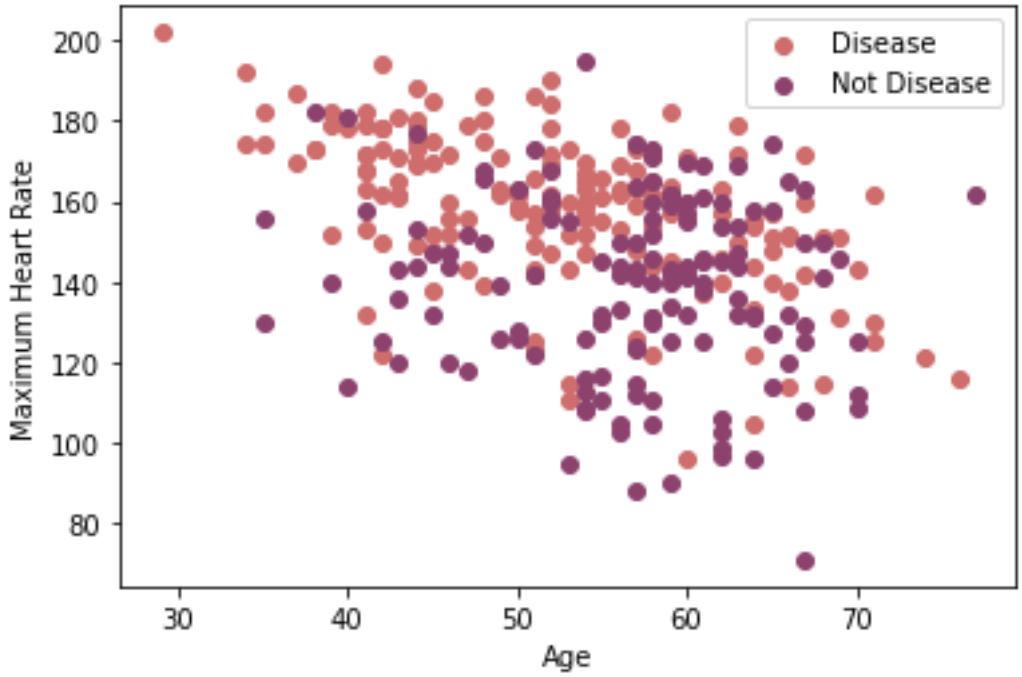
➔ Heart Disease Frequency According To FBS



➔ Heart Disease Frequency According To Chest Pain Type



➔ Maximum Heart Rate



➔ Creating Dummy Variables
From the dataset 'cp', 'thal' and 'slope' are categorical variables I'll turn them into dummy variables.

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	...	cp_1	cp_2	cp_3	thal_0	thal_1	thal_2	thal_3	slope_0	slope_1	slope_2
0	63	1	3	145	233	1	0	150	0	2.3	...	0	0	1	0	1	0	0	1	0	0
1	37	1	2	130	250	0	1	187	0	3.5	...	0	1	0	0	0	1	0	1	0	0
2	41	0	1	130	204	0	0	172	0	1.4	...	1	0	0	0	0	1	0	0	0	1
3	56	1	1	120	236	0	1	178	0	0.8	...	1	0	0	0	0	1	0	0	0	1
4	57	0	0	120	354	0	1	163	1	0.6	...	0	0	0	0	0	1	0	0	0	1

➔ Feature selection
Spliting the 80% of the dataset into train_data and 20% of the dataset into test_data.

➔ Creating Different Machine Learning Model

The machine learning algorithms :

1.Logistic Regression

The meaning of the term regression is very simple: any process that attempts to find relationships between variables is called regression. Logistic regression is regression because it finds relationships between variables. It is logistic because it uses logistic function as a link function.

2.Support Vector Machine (SVM)

A support vector machine (SVM) is a supervised machine learning model that uses classification algorithms for two-group classification problems.

3.K Nearest Neighborhood (kNN)

K nearest neighbors is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure (e.g., distance functions). KNN has been used in statistical estimation and pattern recognition.

4.Gradient Boosting Classifier

Gradient boosting is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models

5.Random Forest Classifier

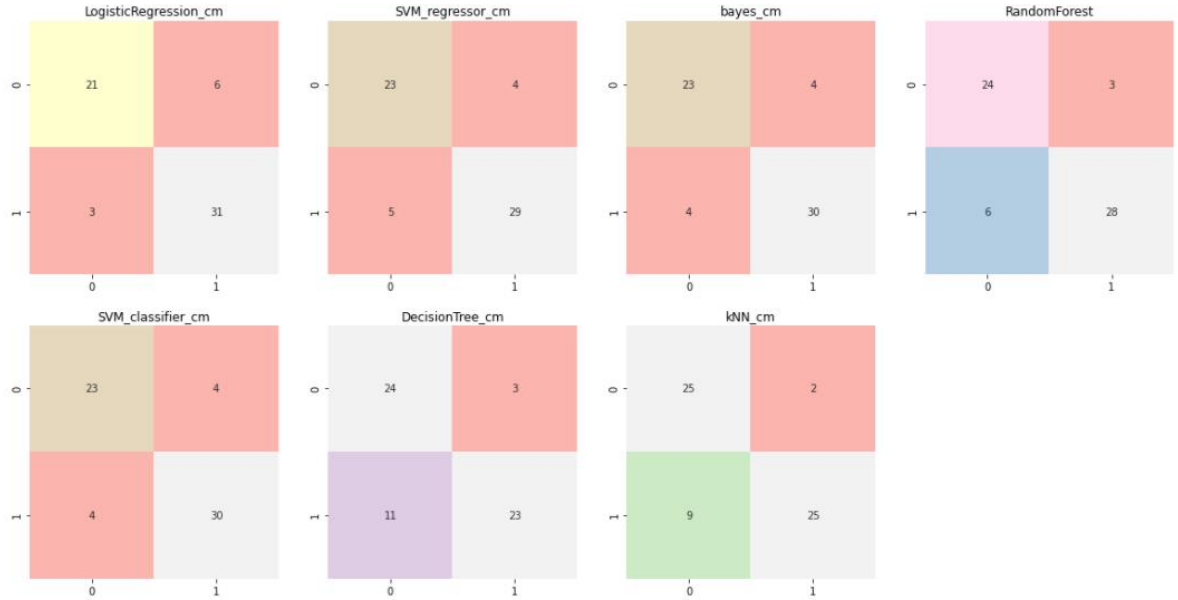
Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes or mean prediction of the individual trees

6.Gaussian NB

Other functions can be used to estimate the distribution of the data, but the Gaussian (or Normal distribution) is the easiest to work with because we only need to estimate the mean and the standard deviation from the training data.

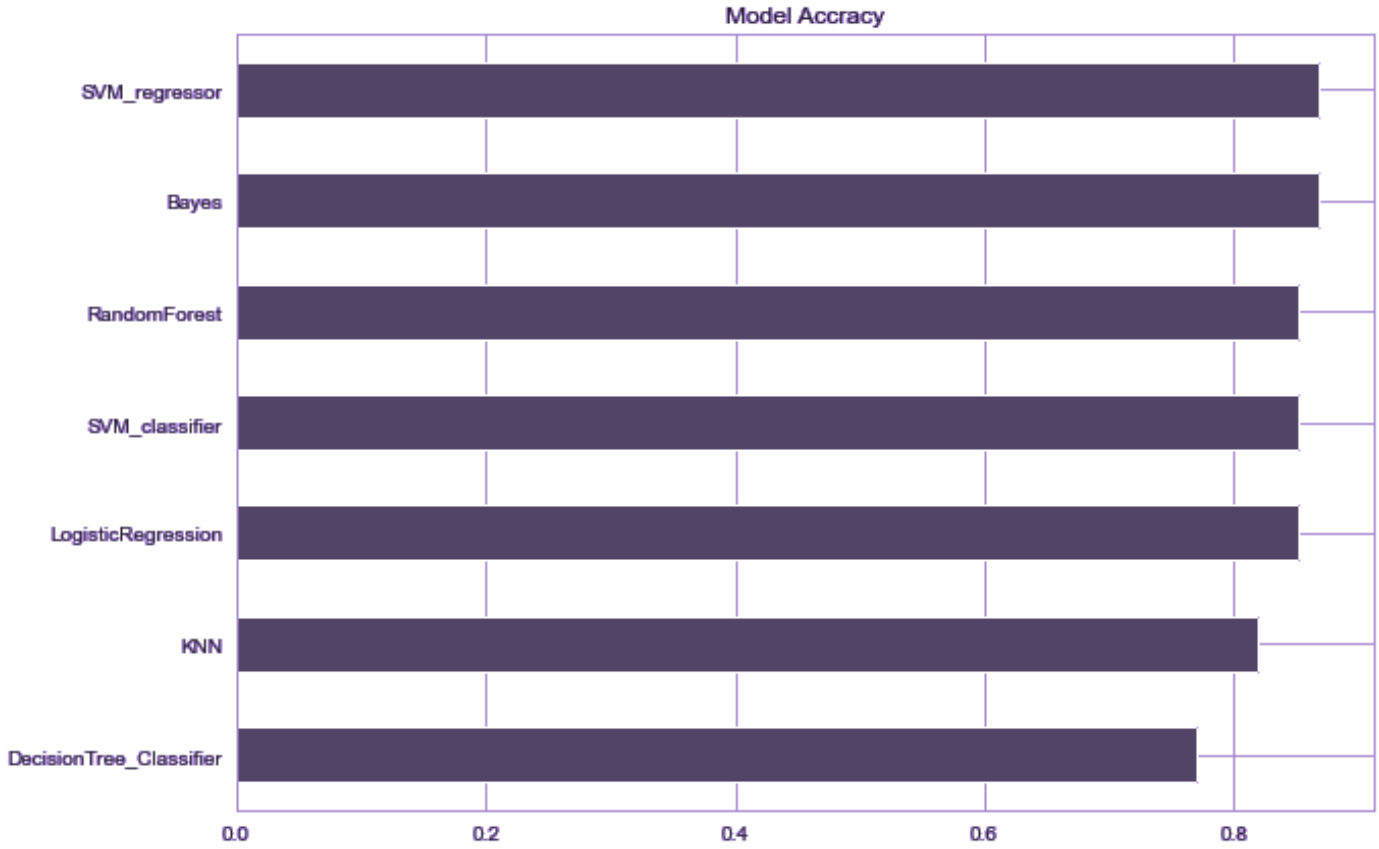
7.DecisionTree Classifier algorithms.

A decision tree classifier is a tree in which internal nodes are labeled by features. The classifier categorizes an object xi by recursively testing for the weights that the features labeling the internal nodes have in vector xi, until a leaf node is reached. The label of this node is then assigned to xi



➔ **Accuracy of the models**

```
LogisticRegression_accuracy:      0.8524590163934426
SVM_regressor_accuracy:           0.8688524590163934
Bayes_accuracy:                   0.8688524590163934
RandomForest_accuracy:            0.8524590163934426
SVM_classifier_accuracy:          0.8524590163934426
DecisionTree_accuracy:            0.7704918032786885
KNN_accuracy:                     0.819672131147541
```



➔ **TRAIN score**

```
LogisticRegression TRAIN score:   0.8677685950413223
SVM_regressor TRAIN score:        0.9090909090909091
Bayes TRAIN score:                0.8471074380165289
Random Forest TRAIN score:        0.9793388429752066
SVM_classifier TRAIN score:       0.8677685950413223
DecisionTree TRAIN score:         1.0
KNN TRAIN score:                  0.8801652892561983
```

➔ **TEST score**

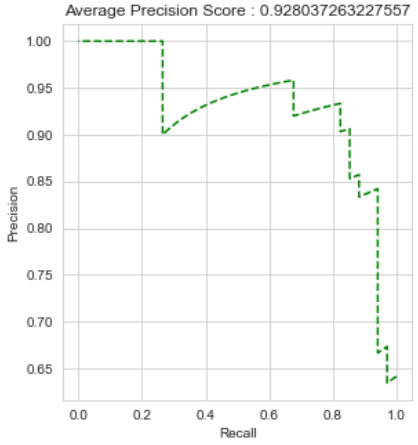
```
LogisticRegression TRAIN score:   0.8524590163934426
SVM_regressor TRAIN score:        0.8688524590163934
Bayes TRAIN score:                0.8688524590163934
Random Forest TRAIN score:        0.8524590163934426
SVM_classifier TRAIN score:       0.8524590163934426
DecisionTree TRAIN score:         0.7704918032786885
KNN TRAIN score:                  0.819672131147541
```

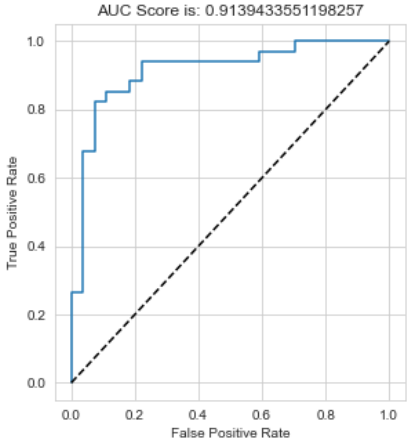
➔ **ROC and Precision Recall Curve for the model which gives the heighest accuracy**

- **A receiver operating characteristic curve, or ROC curve, is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied.**

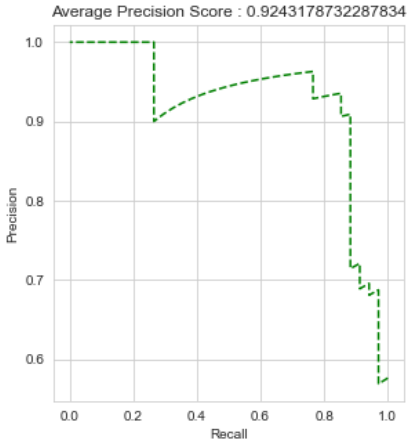
- A precision-recall curve is a plot of the precision (y-axis) and the recall (x-axis) for different thresholds, much like the ROC curve. A no-skill classifier is one that cannot discriminate between the classes and would predict a random class or a constant class in all cases

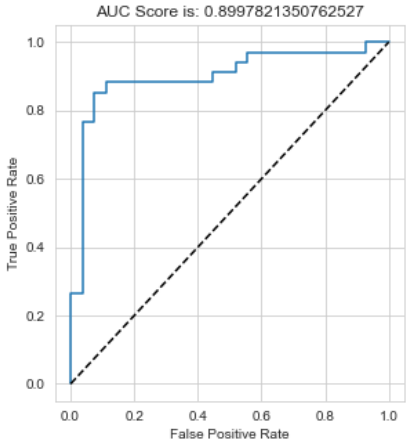
- Logistic Regression



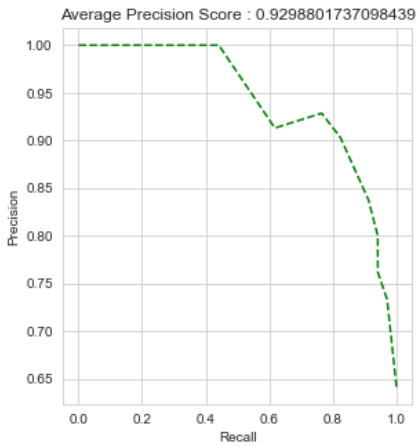


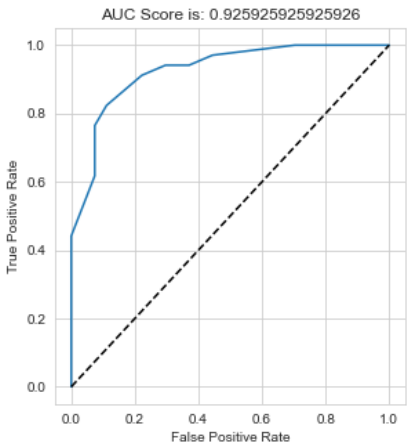
- Bayes



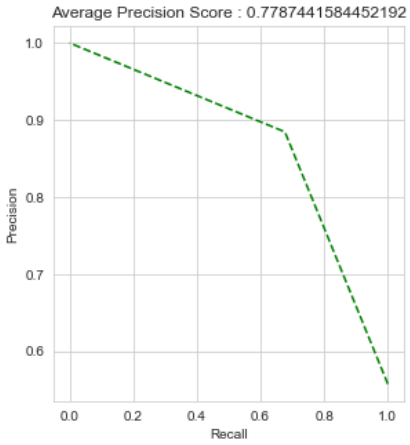


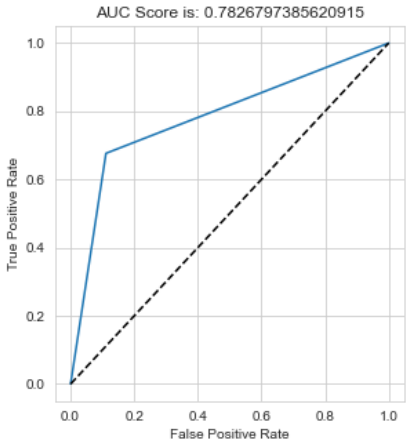
- Random Forest





- **Decision Tree**





- **KNN**