HEART FAILURE DETECTION USING SUPERVISED ML

Gnanendar Reddy Palagiri : 200070053
Jaideep Kotani : 200070035
Jay Soni : 20D170017
Professor : Nirmal Punjabi

Problem Statement :

Data Science and medical data are a powerful combination that can help predict and prevent diseases. One of the most serious and widespread diseases is cardiovascular disease (CVD), which is responsible for 31% of all deaths worldwide, or about 17.9 million lives every year. CVD can lead to heart failure, a condition where the heart cannot pump enough blood to meet the body's needs. To avoid this, people who have or are at risk of developing CVD (due to factors such as high blood pressure, diabetes, high cholesterol, or existing heart problems) need to be diagnosed and treated early. Machine learning models can be very useful for this purpose, as they can analyze large amounts of data and identify patterns and trends that can indicate the likelihood of CVD. By using AI techniques, we can automate the process of detecting and managing CVD, and focus on solving other problems that affect human health.

Aim:

- Design a supervised ML to classify / predict whether a patient is prone to heart failure depending on multiple attributes.
- It is a binary classification with multiple numerical and categorical features.

Dataset:

This dataset was created by combining different datasets already available independently but not combined before. In this dataset, 5 heart datasets are combined over 11 common features which makes it the largest heart disease dataset available so far for research purposes. The five datasets used for its curation are:

- Cleveland: 303 observations
- Hungarian: 294 observations
- Switzerland: 123 observations
- Long Beach VA: 200 observations
- Stalog (Heart) Data Set: 270 observations

Total: 1190 observations
Duplicated: 272 observations
Final dataset: 918 observations

Dataset Attributes:

- Age : age of the patient [years]
- Sex : sex of the patient [M: Male, F: Female]
- ChestPainType : chest pain type [TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic]
- RestingBP : resting blood pressure [mm Hg]
- Cholesterol : serum cholesterol [mm/dl]
- FastingBS : fasting blood sugar [1: if FastingBS > 120 mg/dl, 0: otherwise]
- RestingECG : resting electrocardiogram results [Normal: Normal, ST: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV), LVH: showing probable or definite left ventricular hypertrophy by Estes' criteria]
- MaxHR : maximum heart rate achieved [Numeric value between 60 and 202]
- ExerciseAngina : exercise-induced angina [Y: Yes, N: No]
- Oldpeak : oldpeak = ST [Numeric value measured in depression]
- ST_Slope : the slope of the peak exercise ST segment [Up: upsloping, Flat: flat, Down: downsloping]
- HeartDisease : output class [1: heart disease, 0: Normal]

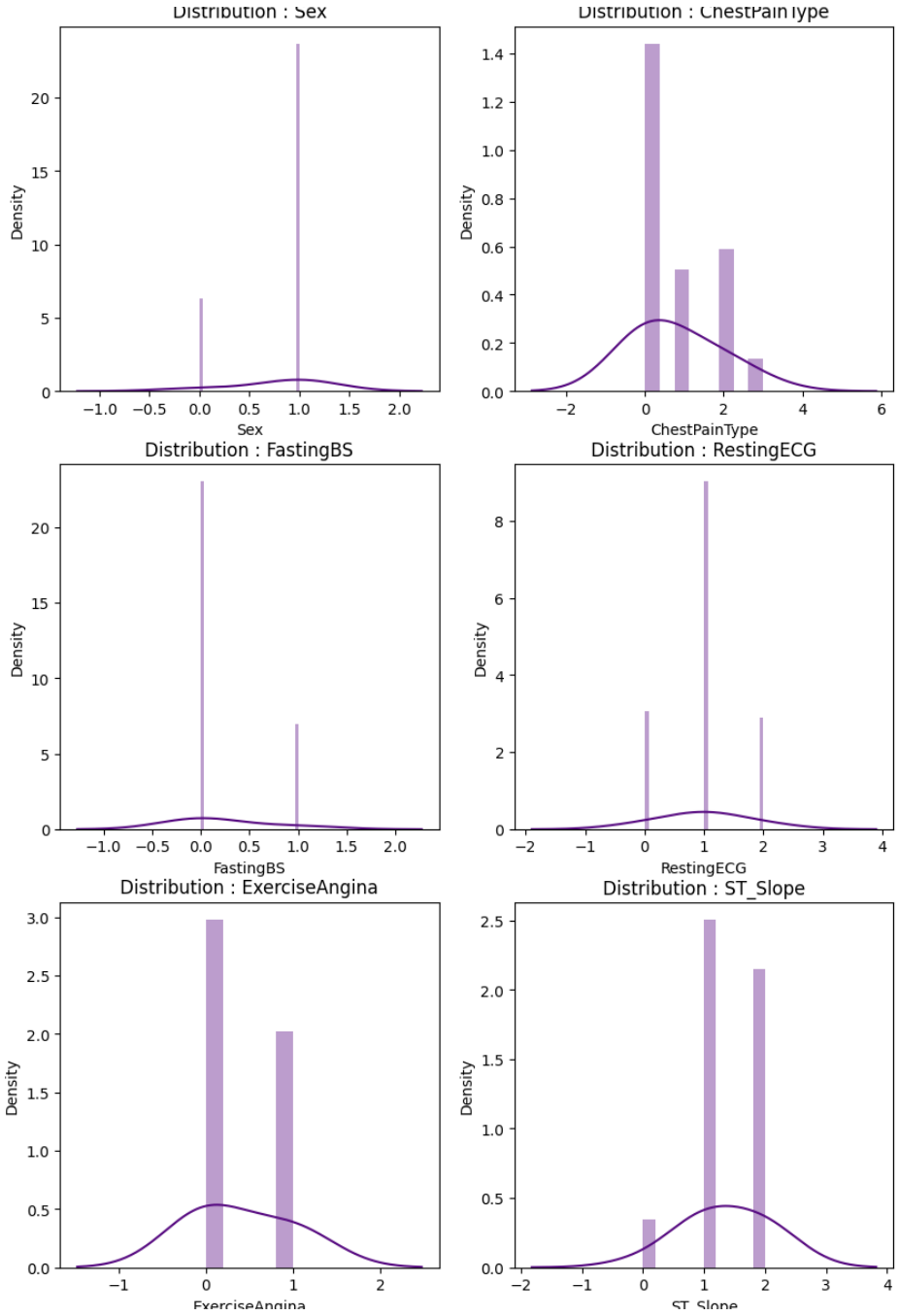
Here, *Age, RestingBP, Cholesterol, MaxHR,* are numerical values with their respective units whereas Sex, ChestPainType, FastingBS, RestingECG,ExerciseAngina, ST_Slope, HeartDisease are categorical features.Numerical features are self explanatory.

Categorical features:

- **Sex** : Whether the sample belongs to a male or a female
- **ChestPainType** : Among the 4 mentioned chest pain types which one he got (0,1,2,3)
- **FastingBS** : If the fastingBS is above the given threshold its 1 else 0
- **RestingECG** : what kind of ECG graph does the person have among mentioned
- **ExerciseAngina** : If the person gets pain/strain in heart after exercise its 1 else 0
- **ST_Slope** : It signifies if the ST part of ECG have slope up, flat, down
- **HeartDisease** : If the person is diagnosed with heart disease or not

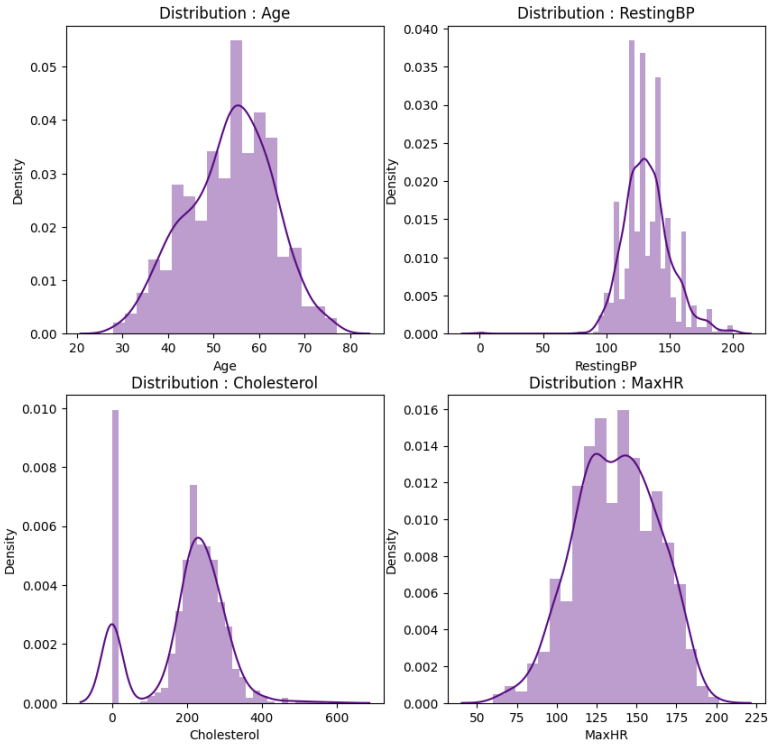
Exploratory Data Analysis:

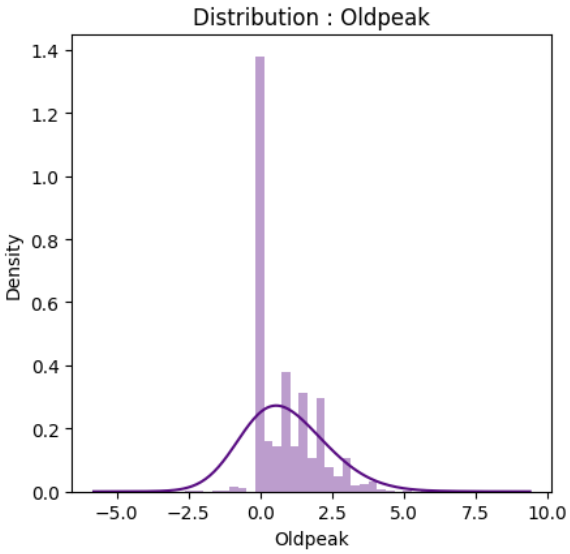
Categorical Features



- All the categorical features are near about **Normally Distributed.**

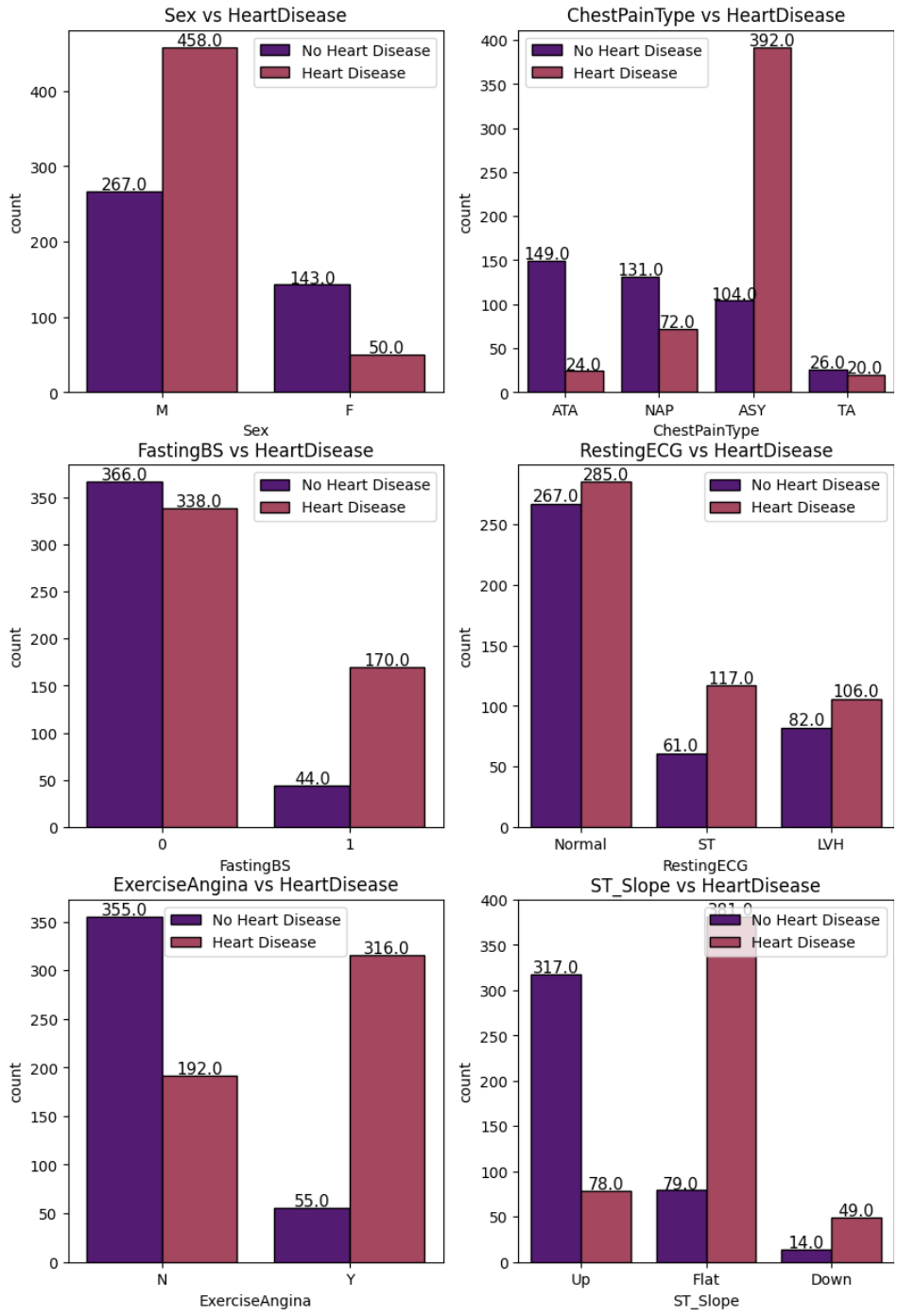
Numerical Features





- **Oldpeak's** data distribution is rightly skewed.
- **Cholestrol** has a bidmodal data distribution

Categorical Features vs Target Variable (HeartDisease)

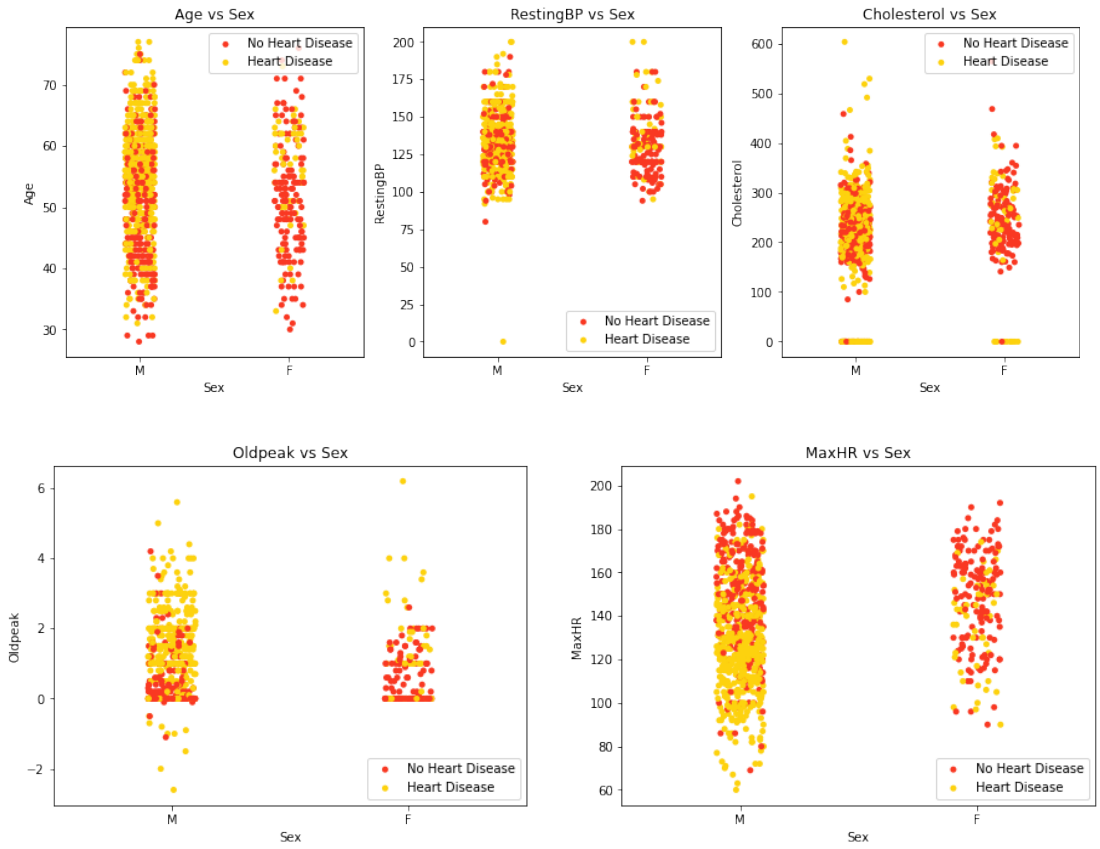


- **Male** population has more heart disease patients than no heart disease patients. In the case of Female population, heart disease patients are less than no heart disease patients.
- **ASY** type of chest pain boldly points towards major chances of heart disease.

- **Fasting Blood Sugar** is tricky! Patients diagnosed with Fasting Blood Sugar and no Fasting Blood Sugar have significant heart disease patients.
- **RestingECG** does not present with a clear cut category that highlights heart disease patients. All the 3 values consist of high number of heart disease patients.
- **Exercise Induced Angina** definitely bumps the probability of being diagnosed with heart diseases.
- With the **ST_Slope values**, **flat** slope displays a very high probability of being diagnosed with heart disease. **Down** also shows the same output but in very few data points.

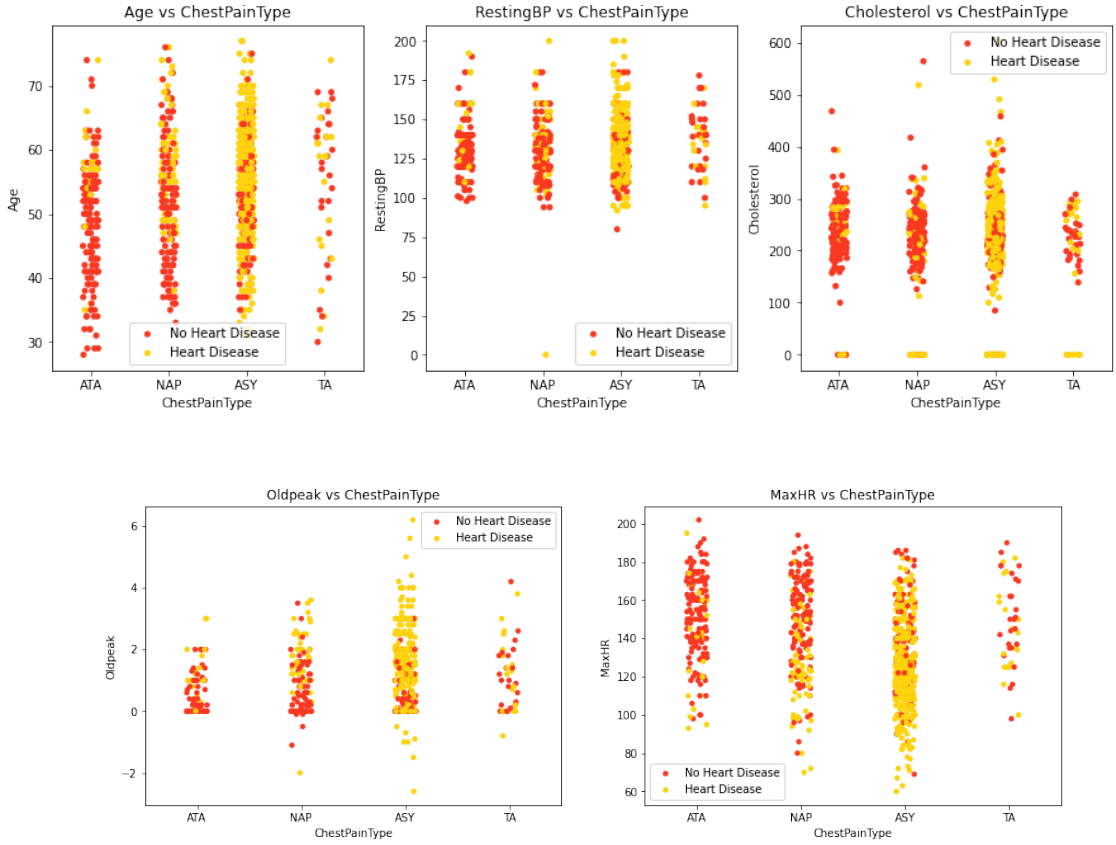
Numerical features vs Categorical features w.r.t Target variable(HeartDisease) :

Sex vs Numerical Features



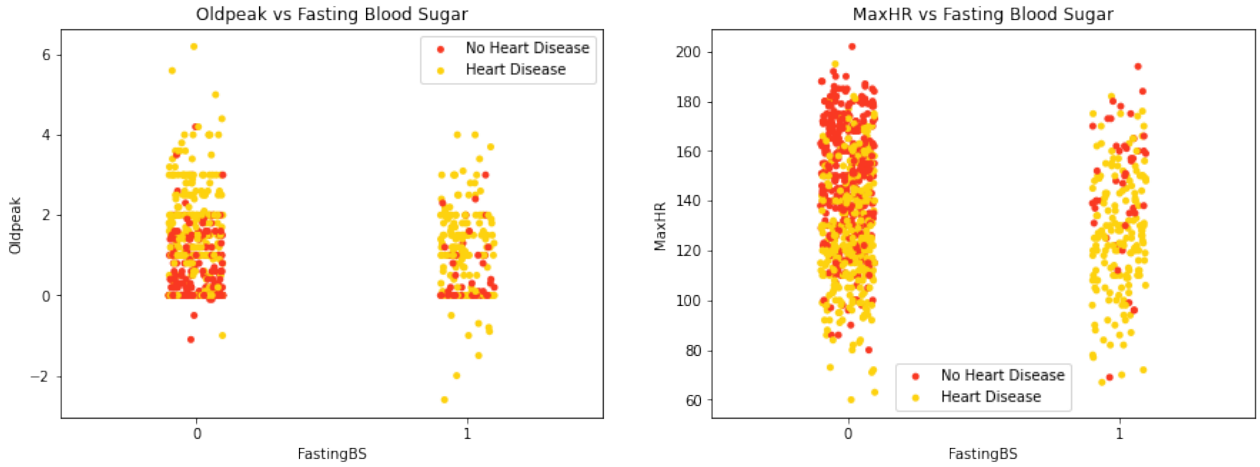
- **Male** population displays heart diseases at near about all the values of the numerical features. Above the age of 50, positive old peak values and maximum heart rate below 140, heart diseases in male population become dense.
- **Female** population data points are very less as compared to male population data points. Hence, we cannot point to specific ranges or values that display cases of heart diseases.

ChestPainType vs Numerical Features



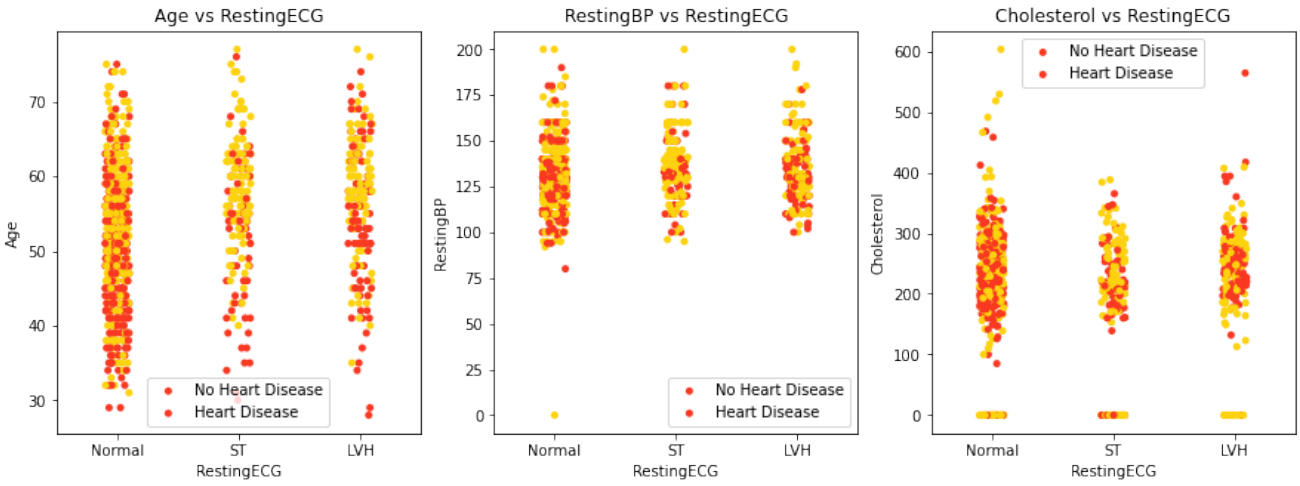
- **ASY** type of chest pain dominates other types of chest pain in all the numerical features by a lot

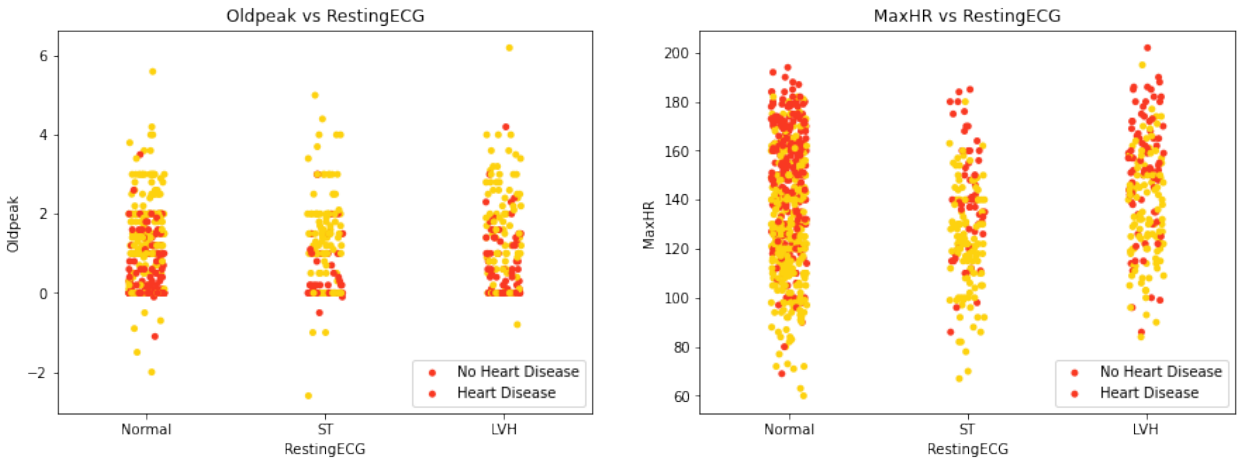
FastingBS vs Numerical features



- Above the **age** 50, heart diseases are found throughout the data irrespective of the patient being diagnosed with Fasting Blood Sugar or not.
- **Fasting Blood Sugar** with **Resting BP** over 100 has displayed more cases of heart diseases than patients with no fasting blood sugar.
- **Cholesterol** with **Fasting Blood Suga**r does not seem to have an effect in understanding reason behind heart diseases.
- Patients that have not been found positive with **Fasting Blood Sugar** but have maximum heart rate below 130 are more prone to heart diseases.

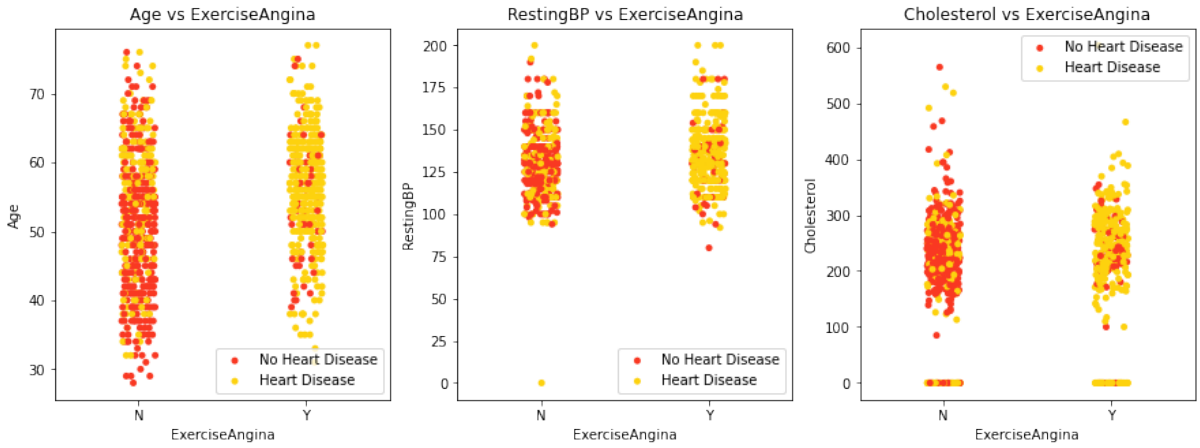
RestingECG vs Numerical Features

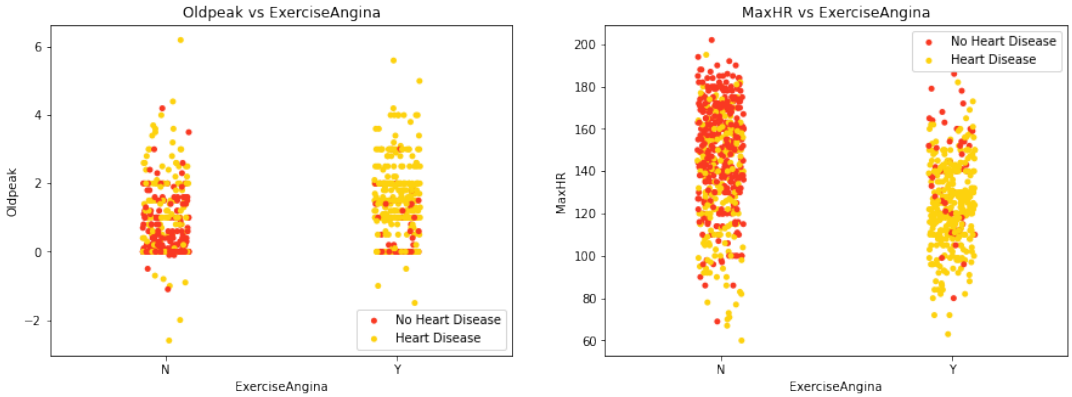




- Heart diseases with **RestingECG** values of **Normal**, **ST** and **LVH** are detected starting from 30,40 & 40 respectively. Patients above the age of 50 are more prone than any other ages irrespective of **RestingECG values**.
- Heart diseases are found consistently throughout any values of **RestingBP** and **RestingECG**.
- **Cholesterol** values between 200 - 300 coupled with **ST** value of **RestingECG** display a patch of patients suffering from heart diseases.
- For **maximum Heart Rate values**, heart diseases are detected in dense below 140 points and **Normal** RestingECG. **ST** & **LVH** throughout the maximum heart rate values display heart disease cases.

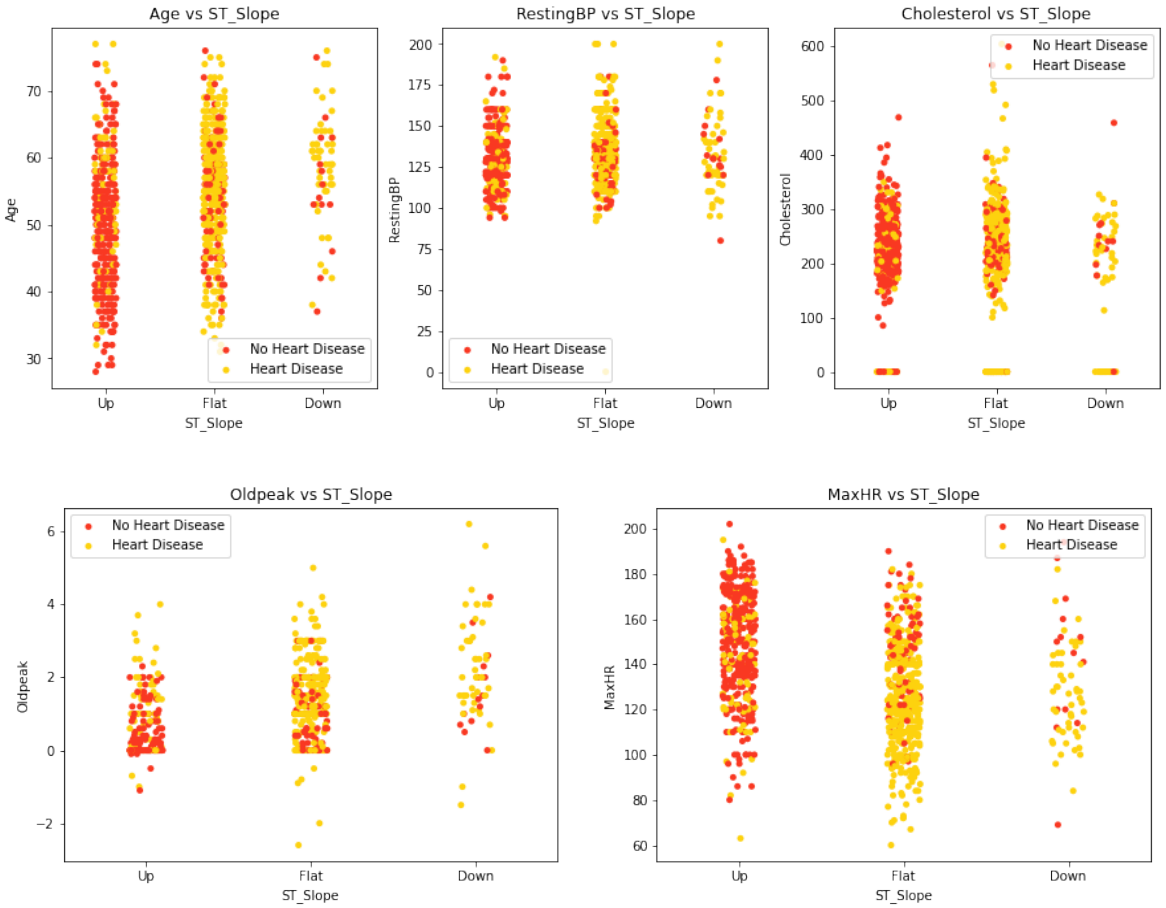
ExerciseAngina vs Numerical Features





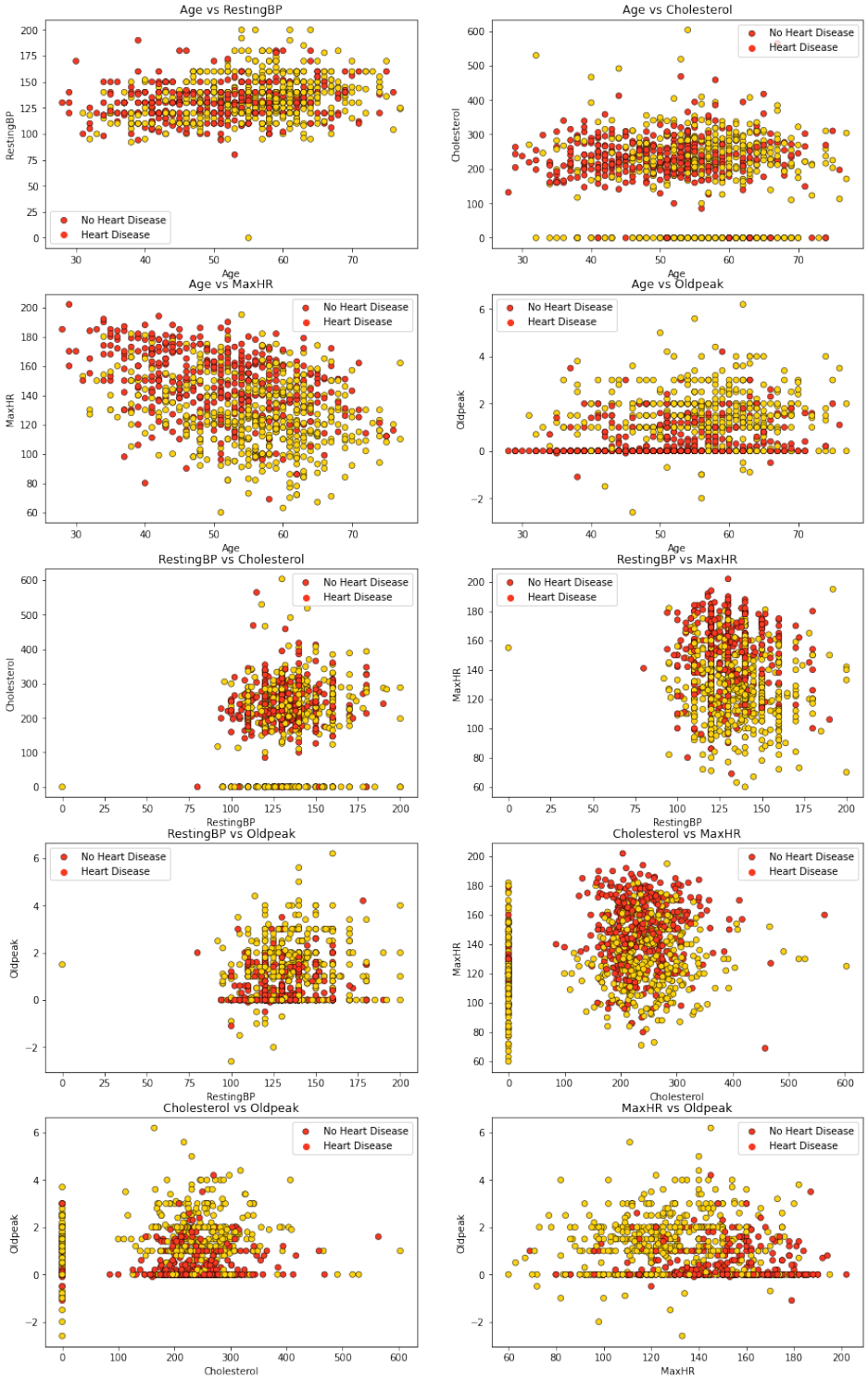
- A crsytal clear observation can be made about the relationship between **heart disease** case and **Exercise induced Angina**. A positive correlation between the 2 features can be concluded throughout all the numerical features.

ST_Slope vs Numerical Features



- Another crystal clear positive observation can be made about the positive correlation between **ST_Slope** value and **Heart Disease** cases.
- **Flat**, **Down** and **Up** in that order display high, middle and low probability of being diagnosed with heart diseases respectively

Numerical features vs Numerical features w.r.t Target variable(HeartDisease)



- For **age** 50+, **RestingBP** between 100 - 175, **Cholesterol** level of 200 - 300,Max Heart Rate below 160 and positive **oldpeak** values displays high cases of heart disease.
- For **RestingBP** values 100 - 175, highlights too many heart disease patients for all the features.
- **Cholesterol** values 200 - 300 dominates the heart disease cases.
- Similarly, **Max** Heart Rate values below 140 has high probability of being diagnosed with heart diseases.

Summary of EDA

Order / Values of features for positive cases of heart disease

Categorical Features (Order) :

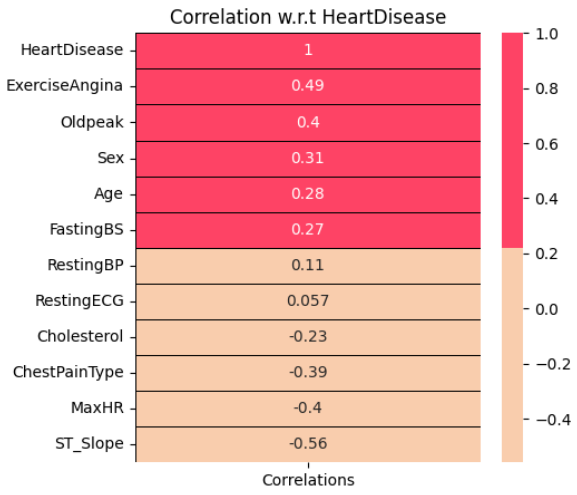
- Sex : Male > Female
- ChestPainType : ASY > NAP > ATA > TA
- FastingBS : (FBS < 120 mg/dl) > (FBS > 120 mg/dl)
- RestingECG : Normal > ST > LVH
- ExerciseAngina : Angina > No Angina
- ST_Slope : Flat > Up > Down

Numerical Features (Range) :

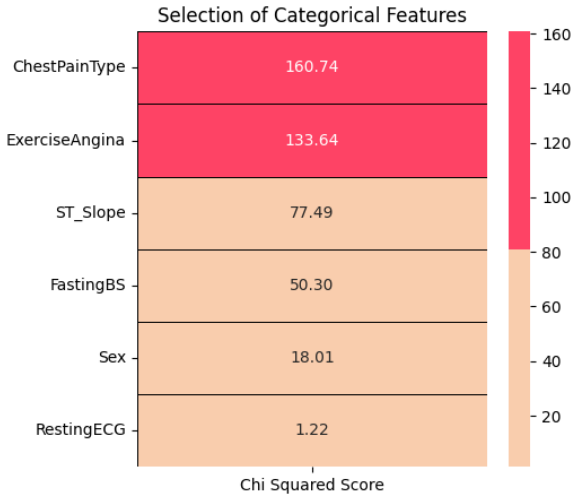
- Age : 50+
- RestingBP : 95 - 170
- Cholesterol : 160 - 340
- MaxHR : 70 - 180
- Oldpeak : 0 - 4

Feature Engineering

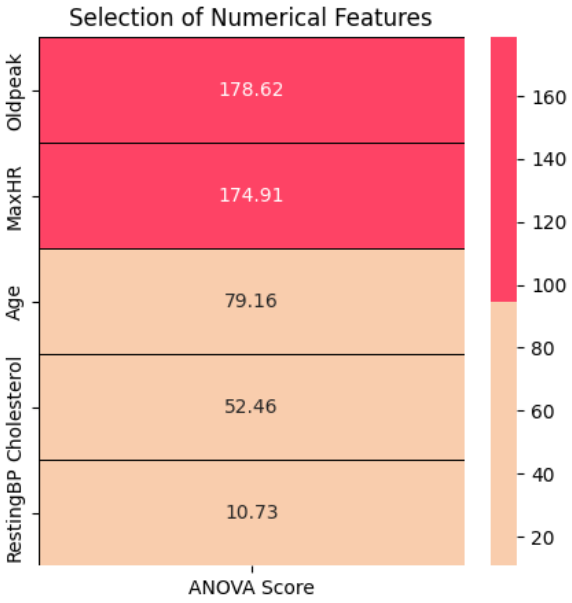
Correlation Matrix :



Chi Squared Test :



ANOVA Test :



Observations:

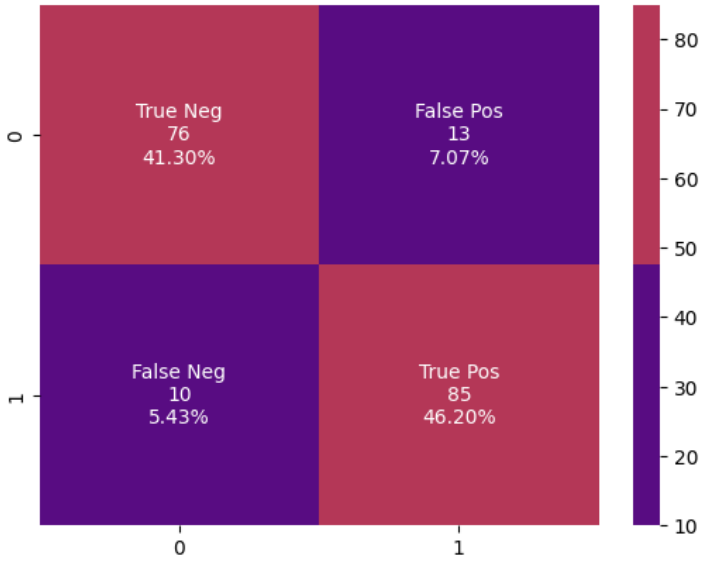
- We will leave **RestingECG**, **RestingBP** because they both have lower correlation with other features

Modeling

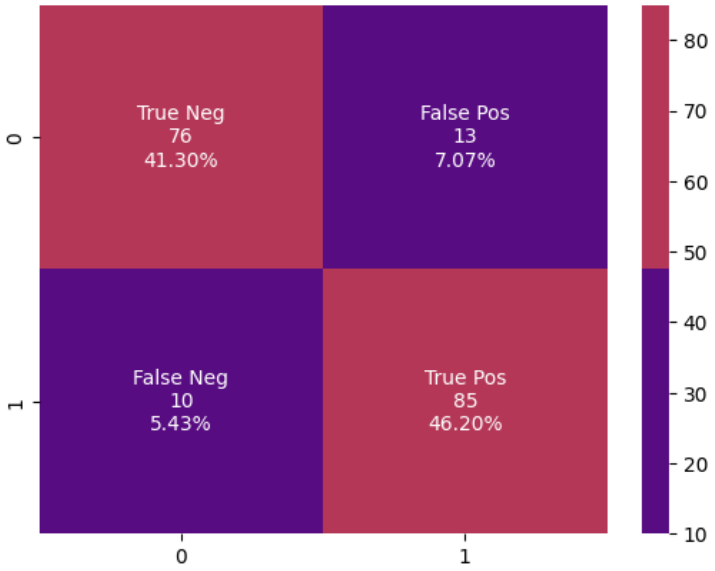
- Selecting the features from the above conducted tests and splitting the data into 80 - 20 train - test groups.
- We have 2 options for data scaling : 1) **Normalization** 2) **Standardization**. As most of the algorithms assume the data to be normally (Gaussian) distributed, **Normalization** is done for features whose data does not display normal distribution and **standardization** is carried out for features that are normally distributed where their values are huge or very small as compared to other features.
- **Normalization** : **Oldpeak** feature is normalized as it had displayed a right skewed data distribution.
- **Standardizarion** : **Age**, **RestingBP**, **Cholesterol** and **MaxHR** features are scaled down because these features are normally distributed.
- We are using 1)Logistic Regression 2)Support Vector Classifier 3)Decision Tree Classifier 4)Random Forest Classifier 5) K-nearest Neighbors Classifier

Results

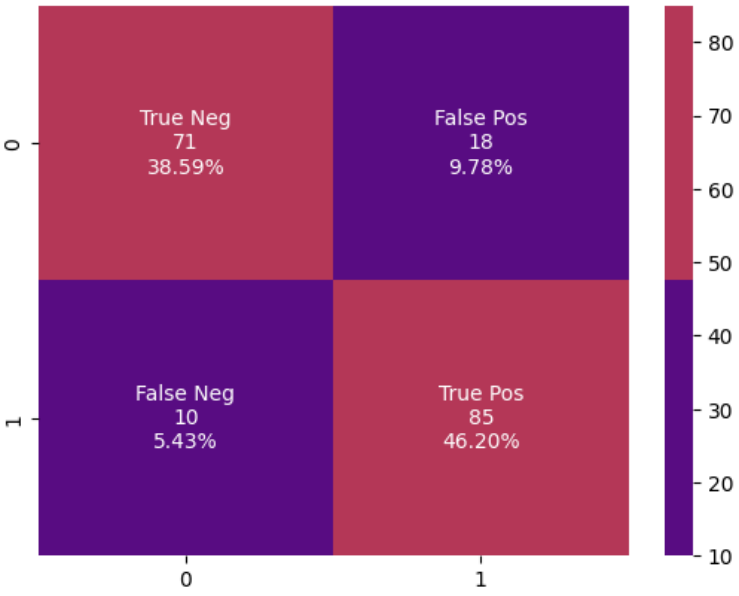
Confusion Metric



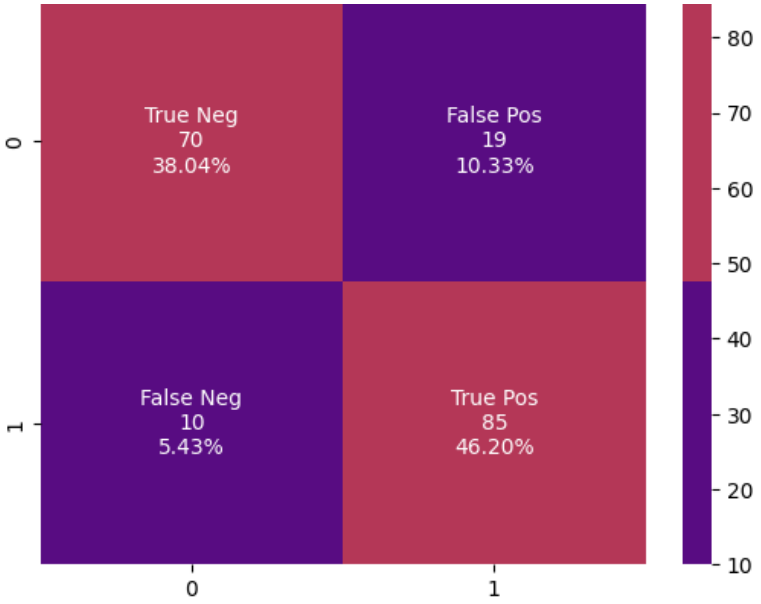
Logistic Regression



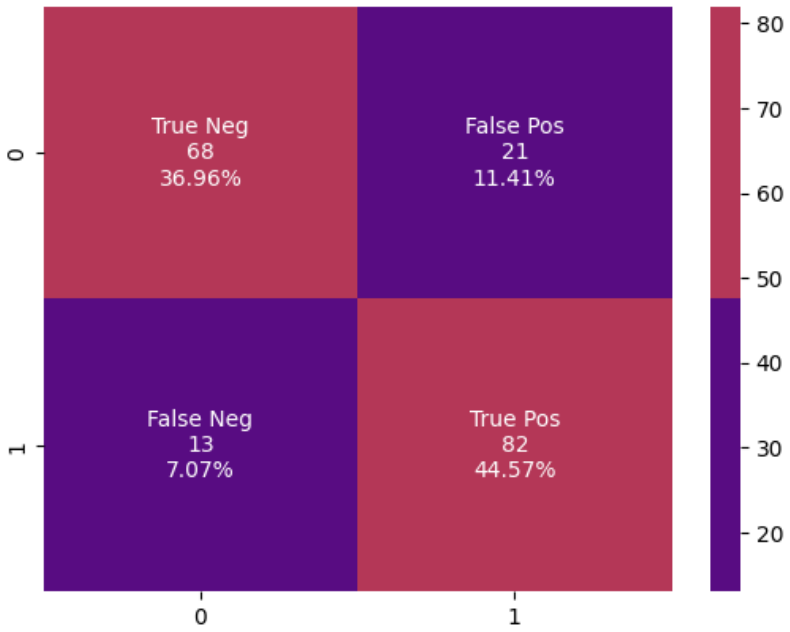
SVC



Decision Tree



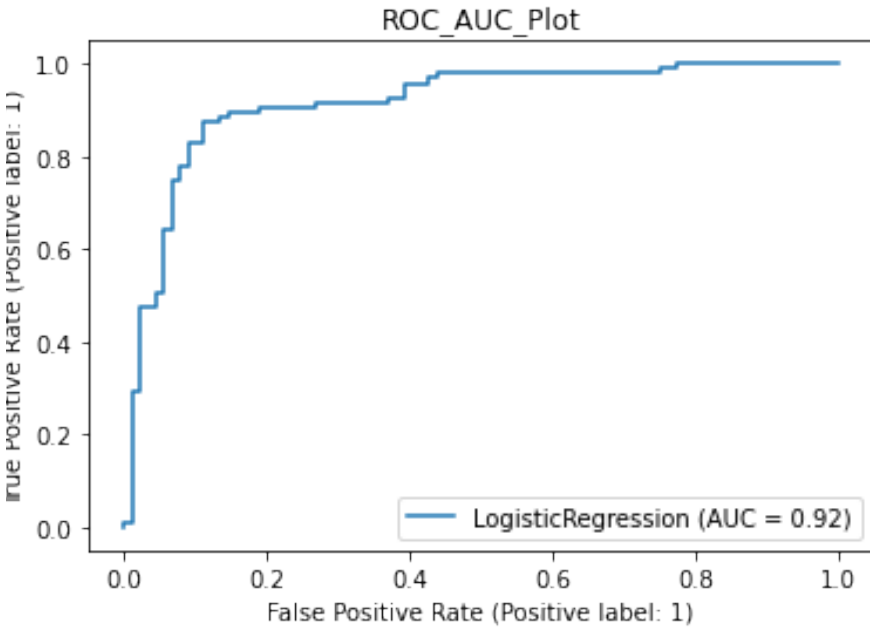
Random Forest

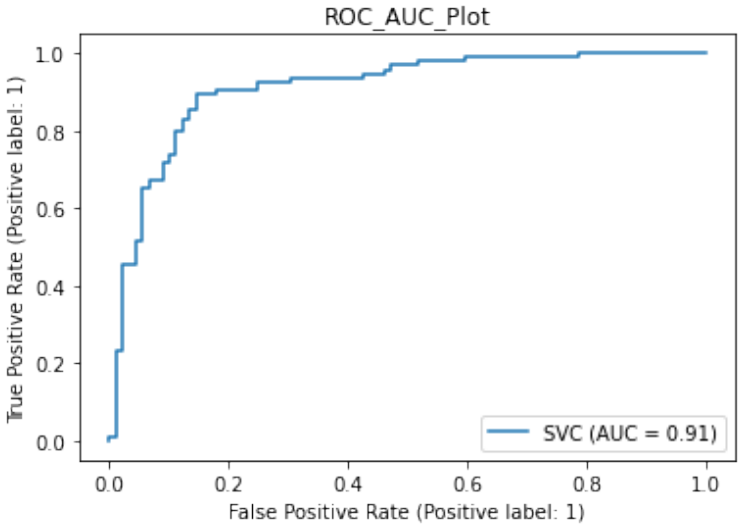


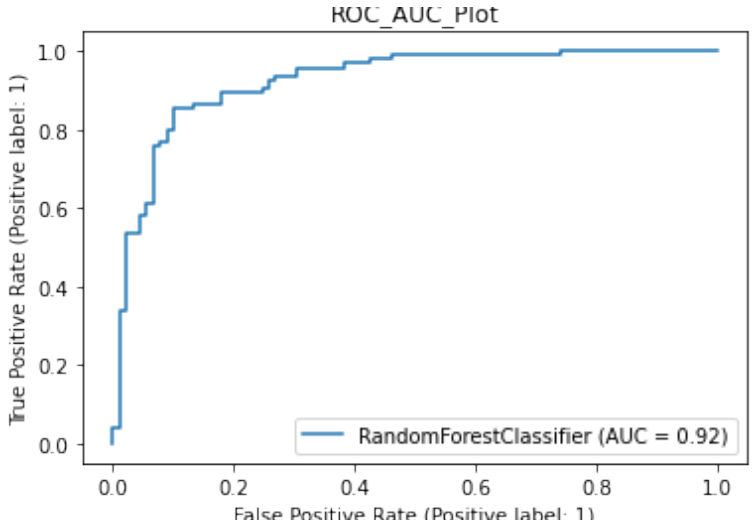
KNN

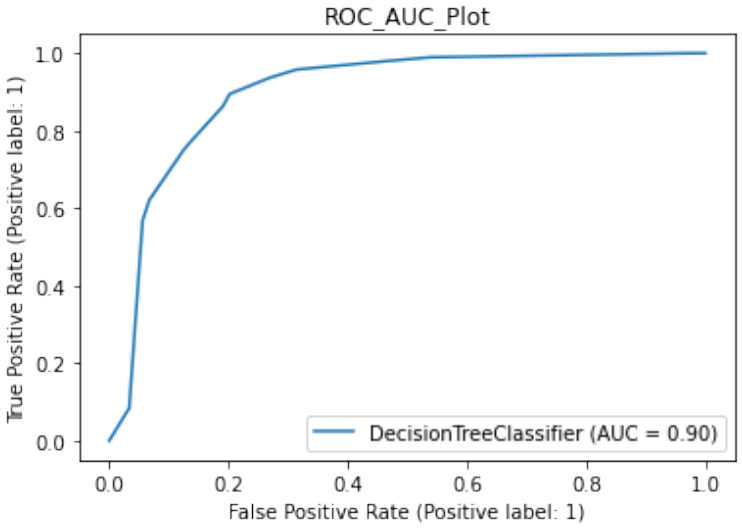
ROC AOC PLOT:

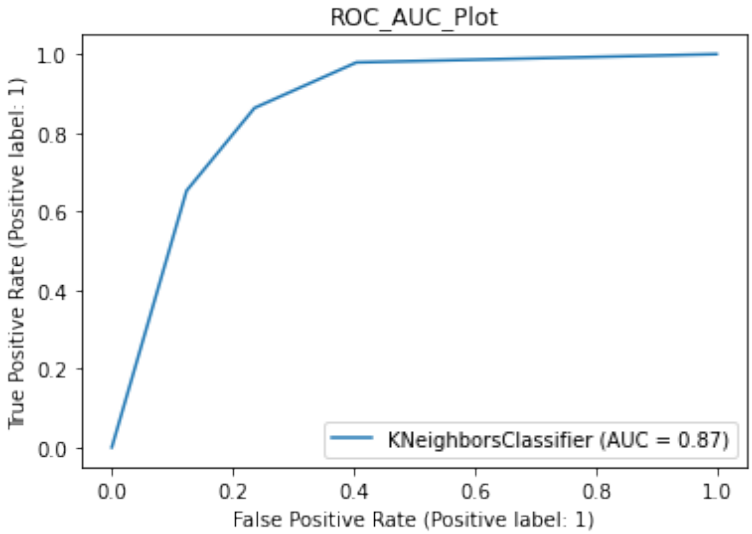
- Since we were using predefined models we weren't able to extract data from model to plot all ROC_AUC curves in single plot.(here, plotting ROC_AUC was possible because its inbuilt into model)




ROC_AUC_Plot
True Positive Rate (Positive label: 1)
False Positive Rate (Positive label: 1)
SVC (AUC = 0.91)
0.0
0.2
0.4
0.6
0.8
1.0


ROC_AUC_Plot
True Positive Rate (Positive label: 1)
False Positive Rate (Positive label: 1)
RandomForestClassifier (AUC = 0.92)
0.0
0.2
0.4
0.6
0.8
1.0


ROC_AUC_Plot
True Positive Rate (Positive label: 1)
False Positive Rate (Positive label: 1)
DecisionTreeClassifier (AUC = 0.90)
0.0
0.2
0.4
0.6
0.8
1.0


ROC_AUC_Plot
True Positive Rate (Positive label: 1)
False Positive Rate (Positive label: 1)
KNeighborsClassifier (AUC = 0.87)
0.0
0.2
0.4
0.6
0.8
1.0

Conclusion

Sr. No.	ML Algorithm	Accuracy	Cross Validation Score	ROC AUC Score
1	Logistic Regression	87.50%	91.12%	87.43%
2	Support Vector Classifier	87.50%	90.53%	87.43%
3	Decision Tree Classifier	84.78%	89.09%	84.62%
4	Random Forest Classifier	84.24%	92.91%	84.06%
5	K-Nearest Neighbors Classfier	81.52%	89.34%	81.36%

- This dataset is great for understanding how to handle binary classification problems with the combination of numerical and categorical features.
- Subject matter experts, in this case doctors or nurses, can be assisted by providing insights that enables them to take the next line of action.
- For feature engineering, it might feel confusing about the order of the processes. In this case, data scaling was executed before the feature selection test. We might feel like we are tampering the data before passing it to the tests but the results are same irrespective of the order of the process. (Try it out!)
- For this problem, outlier detection was not done as I was not able to read any papers about heart diseases. It becomes a pivotal part to understand the subject before removing outliers even though the outlier detection tests come out positive.
- Visualization is key. It makes the data talkative. Displaying the present information and results of any tests or output through visualization becomes crucial as it makes the understanding easy.
- For modeling, hyperparameter tuning is not done. It can push the performances of the algorithms. Overall the algorithm performances are good.

References

- Every dataset used can be found under the Index of heart disease datasets from UCI Machine Learning Repository on the following link

Acknowledgments for data creators:

1. Hungarian Institute of Cardiology. Budapest: Andras Janosi, M.D.
2. University Hospital, Zurich, Switzerland: William Steinbrunn, M.D.
3. University Hospital, Basel, Switzerland: Matthias Pfisterer, M.D.
4. V.A. Medical Center, Long Beach and Cleveland Clinic Foundation: Robert Detrano, M.D., Ph.D.