Heart Disease Prediction

Using Machine Learning Algorithm

By Aditi Goyal

Agenda

01 **Problem Statement**

02 **Understanding of Data**

03 **Basic EDA of Data**

04 **Machine Learning Algorithms**

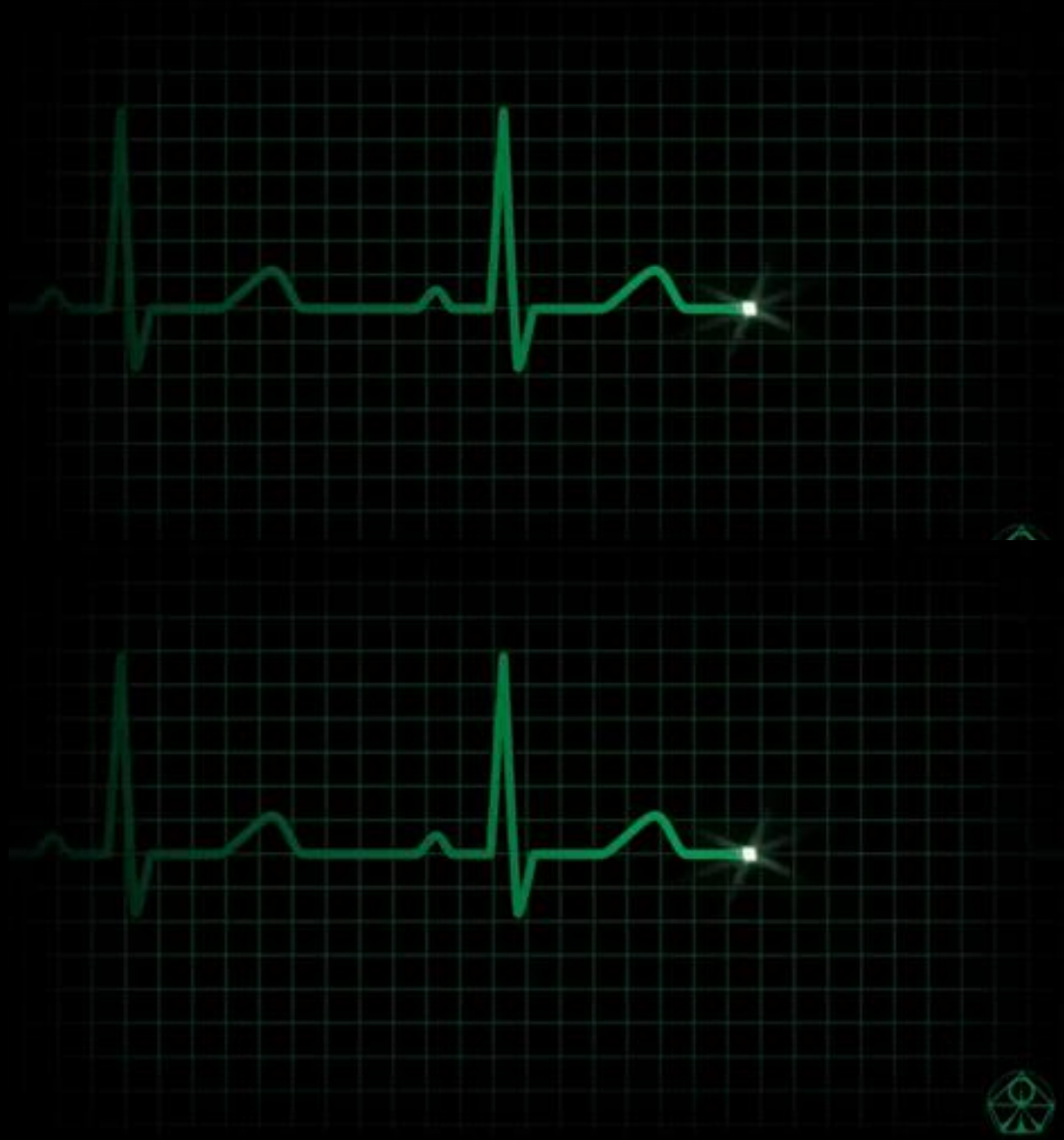
05 **References**

Problem Statement

An estimate by the World Health Organization, that over 17.9 million deaths occur every year worldwide because of cardiovascular disease, and of these deaths, 80% are because of coronary artery disease and cerebral stroke.

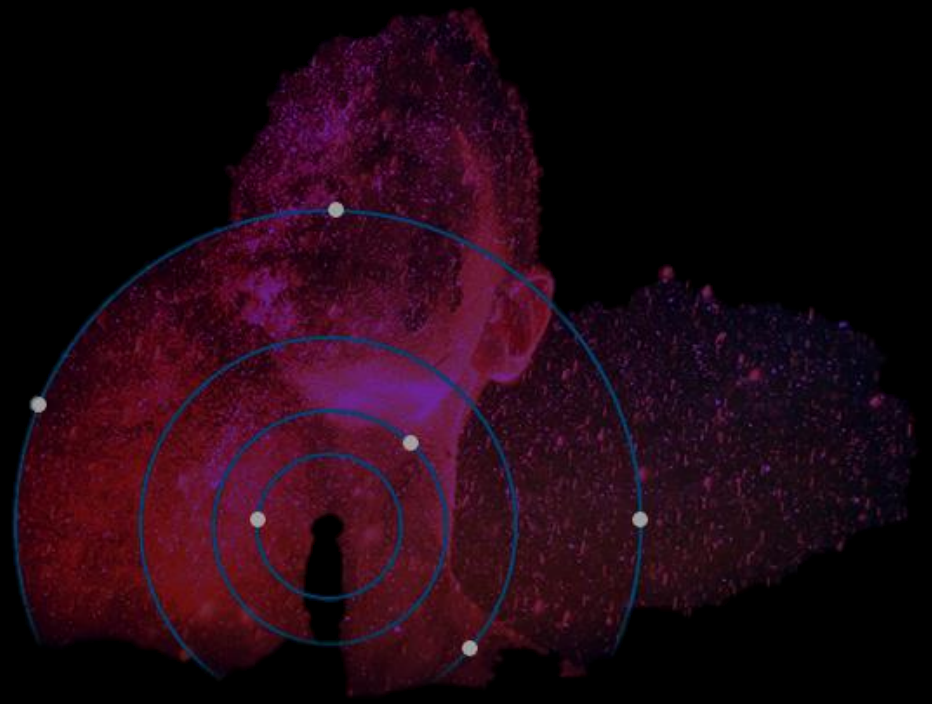
To deal with the problem there is essential need of prediction system for awareness about diseases. Machine learning provides prestigious support in predicting this event. Here, we will be taking a pseudo sample of heart disease data and will be using various ML methods such as random forest, logistic regression, decision tree in prediction of heart disease by processing patients' data and assess effectiveness of these ML methods.

The method with maximum accuracy can be used to predict diseases and take precautionary steps thus saving many lives.



Understanding of Data

Column Name	Description
age	age
sex	sex (1 = male, 0 = female)
cp	chest pain type (4 values) [0: Typical Angina, 1: Atypical Angina, 2: Non-Anginal Pain, 3:Asymptomatic]
trestbps	resting blood pressure
chol	serum cholesterol in mg/dl
fbs	fasting blood sugar ; if fbs> 120 mg/dl then [0 = no, 1 = yes]
restecg	resting electrocardiographic results [0: normal, 1: having ST-T wave abnormality , 2: showing probable or definite left ventricular hypertrophy]
thalach	maximum heart rate achieved
exang	exercise induced angina [1 = yes, 0 = no]
oldpeak	ST depression induced by exercise relative to rest
slope	the slope of the peak exercise ST segment
ca	number of major vessels (0-3) colored by flourosopy
thal	displays the thalassemia 1 = normal; 2 = fixed defect; 3 = reversable defect
target	0 = disease, 1 = no disease



Data Normalization

- Duplicate data
- Aligning data types
- Missing data and its imputation
- Outliers and their removal
- Grouping

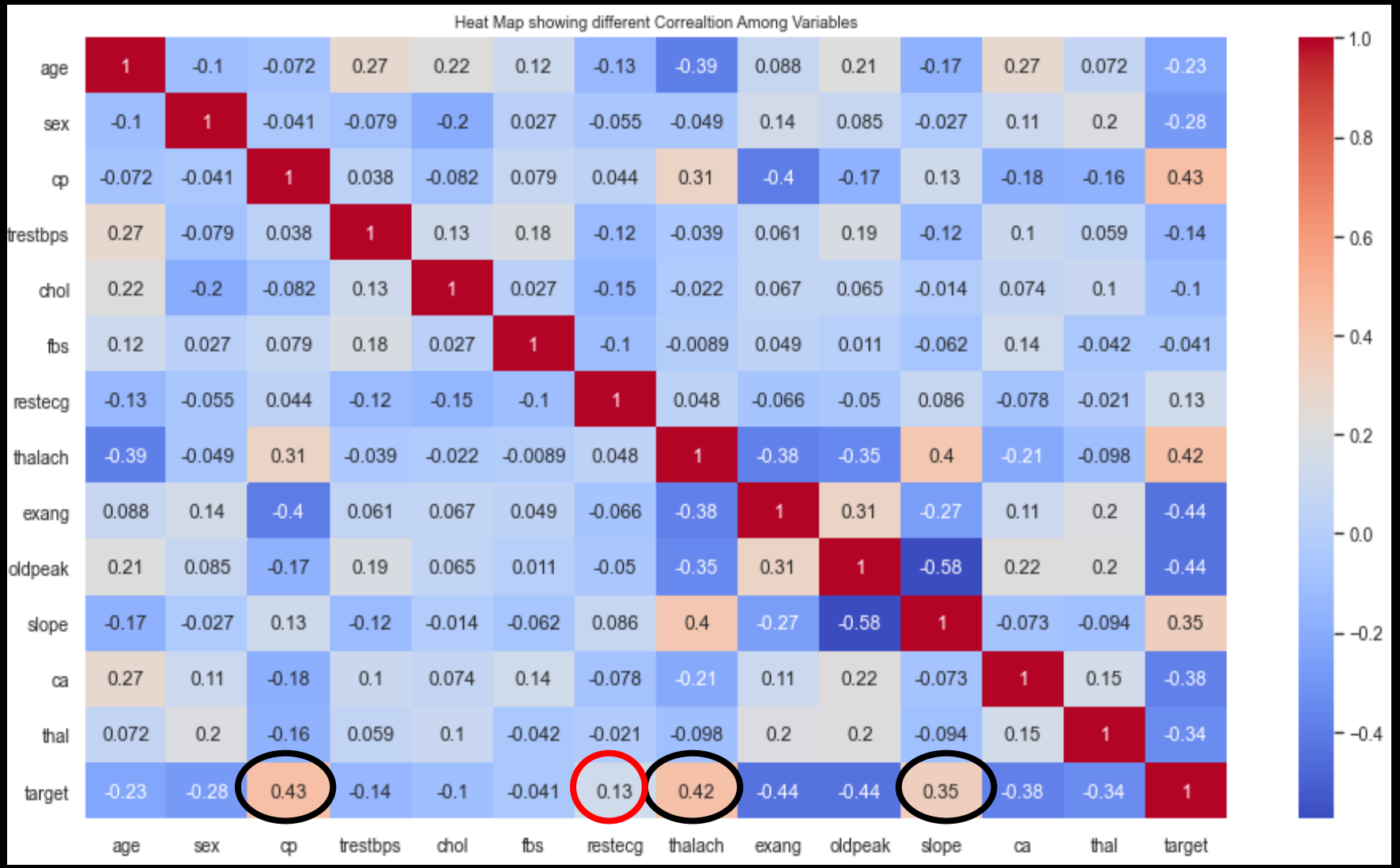
```
For the feature age, No of Outliers is 0
For the feature trestbps, No of Outliers is 9
For the feature chol, No of Outliers is 5
For the feature thalach, No of Outliers is 1
For the feature oldpeak, No of Outliers is 5
```

Feature 'ca' ranges from **0–3**, however, df.nunique() listed **0–4**.

We observe that '4' is basically null value hence we have replaced it with **NAN**.

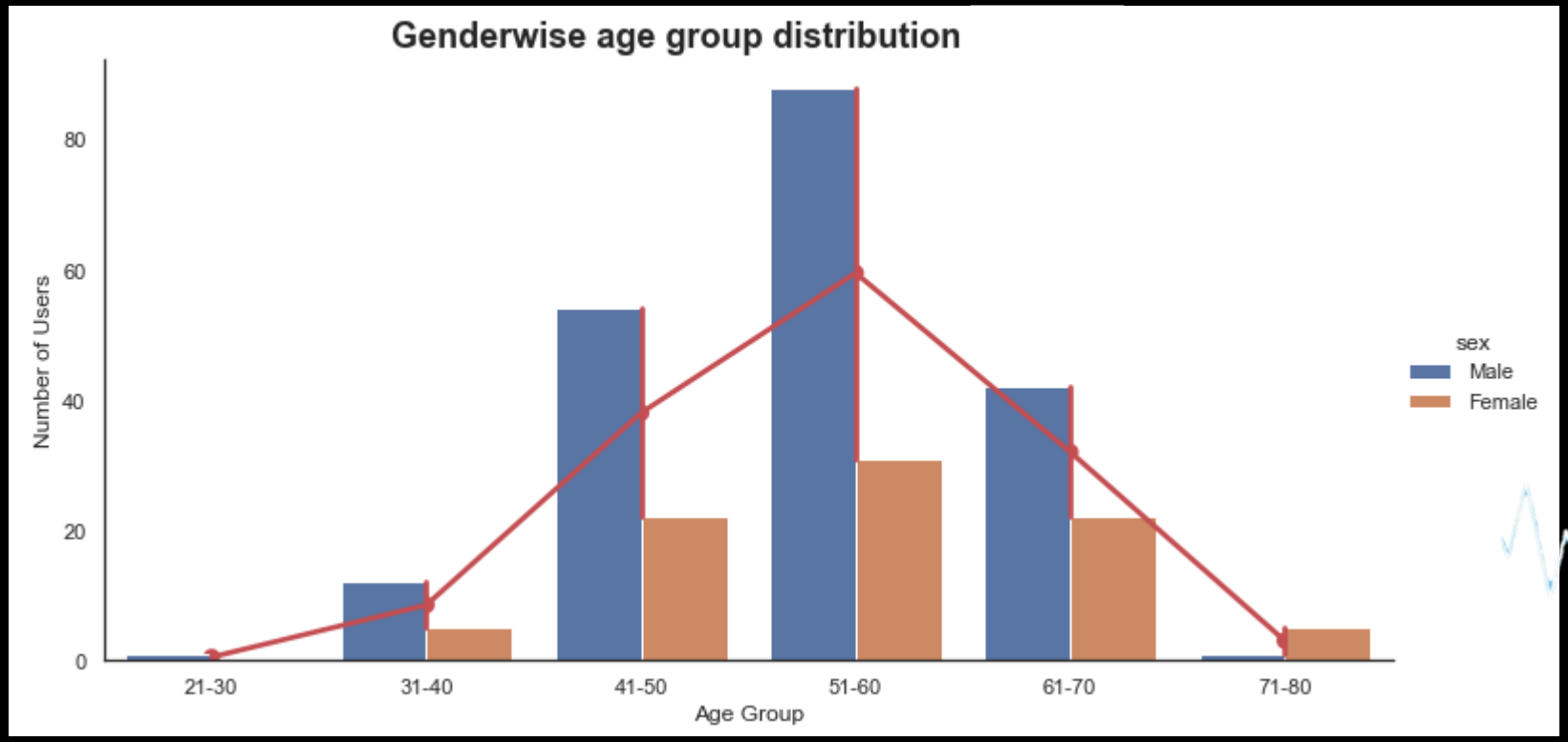
	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	52	Male	typical_angina	125	212	False	1	168	No	1.0	downsloping	2	normal	No_disease
1	53	Male	typical_angina	140	203	True	0	155	Yes	3.1	upsloping	0	normal	No_disease
2	70	Male	typical_angina	145	174	False	1	125	Yes	2.6	upsloping	0	normal	No_disease
3	61	Male	typical_angina	148	203	False	1	161	No	0.0	downsloping	1	normal	No_disease
4	62	Female	typical_angina	138	294	True	1	106	No	1.9	flat	3	reversable_defect	No_disease

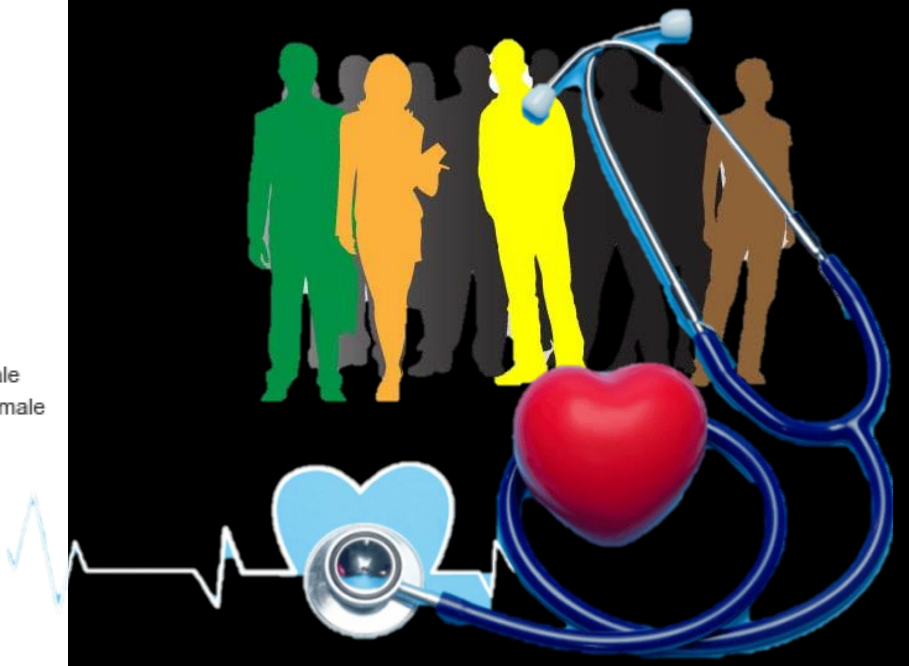
Correlation Matrix



- ‘cp’, ‘thalach’, ‘slope’ are positively correlated with target
- ‘restecg’ has low correlation with target..
- Other variables are negatively corelated

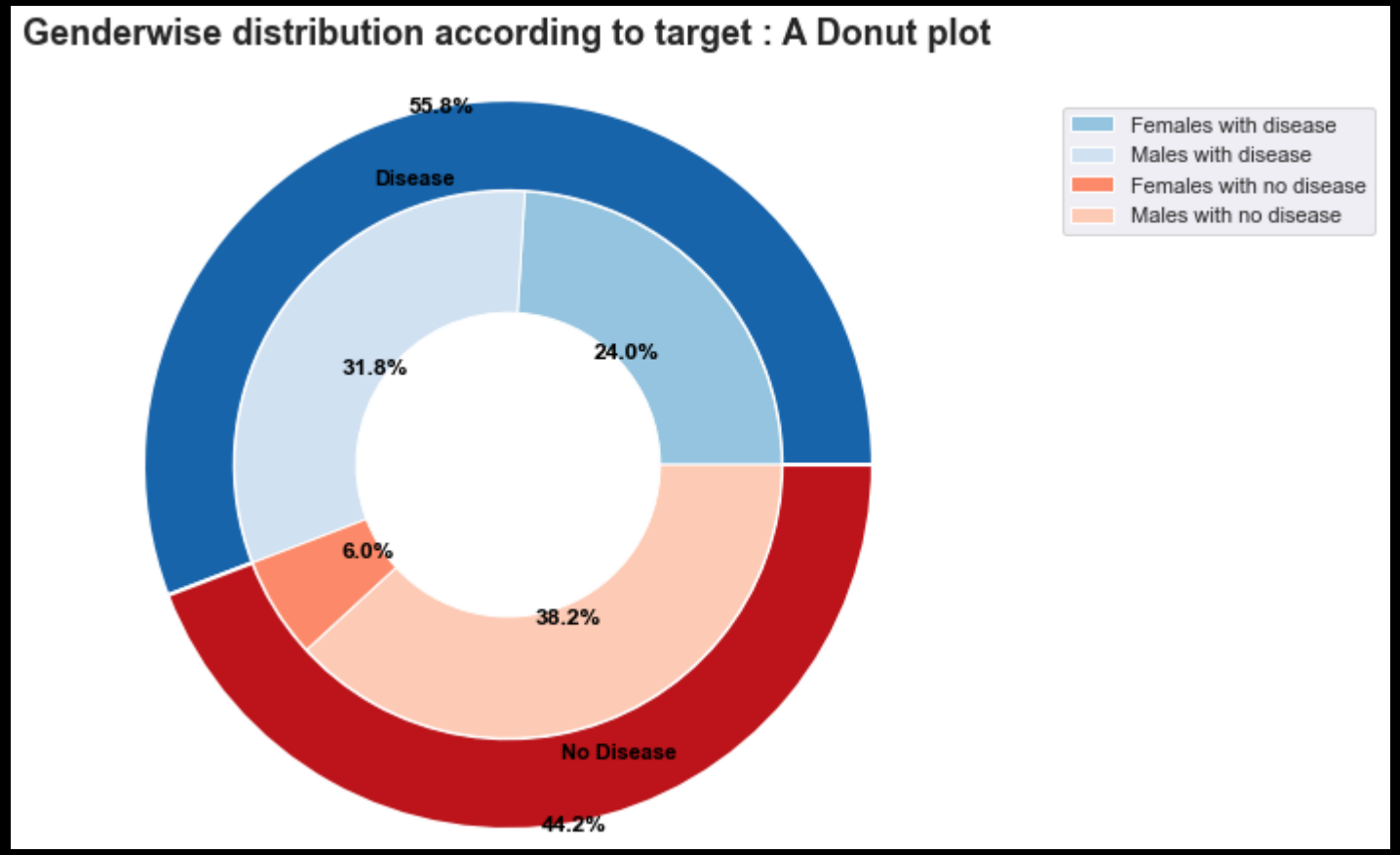
Gender wise age group distribution





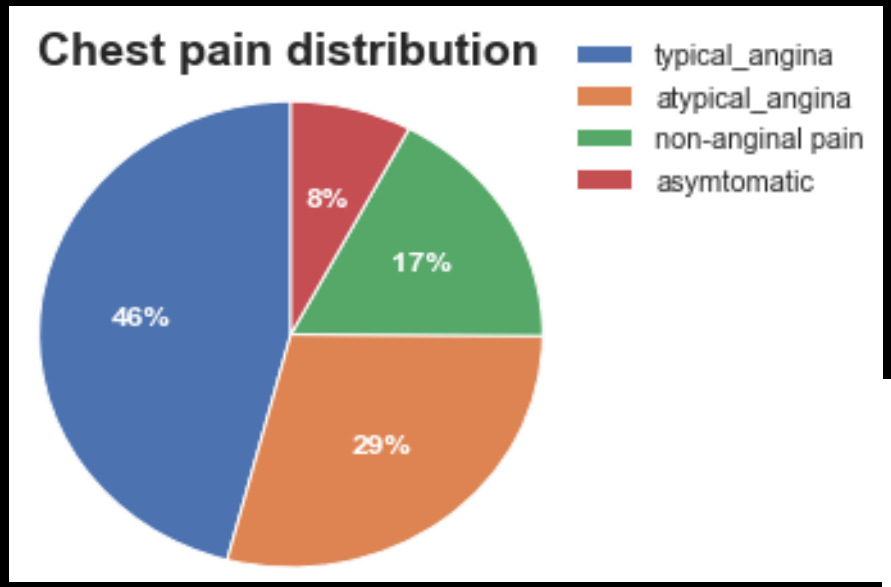
- Maximum users are of age group 51-60 yrs which indicates that risk of heart disease is quite common in this age group.
- For every age group, number of males are higher than female.

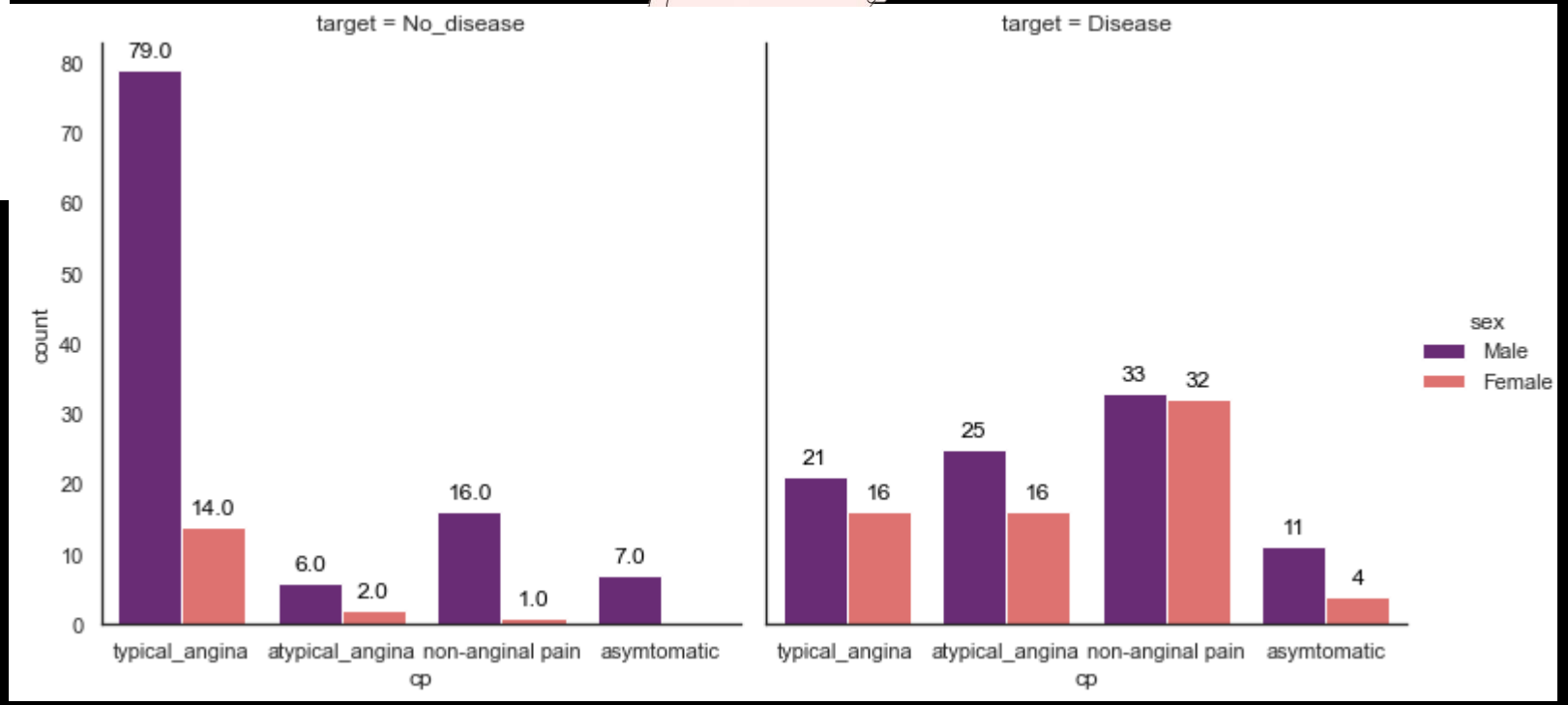
Gender wise distribution according to target : A Donut plot



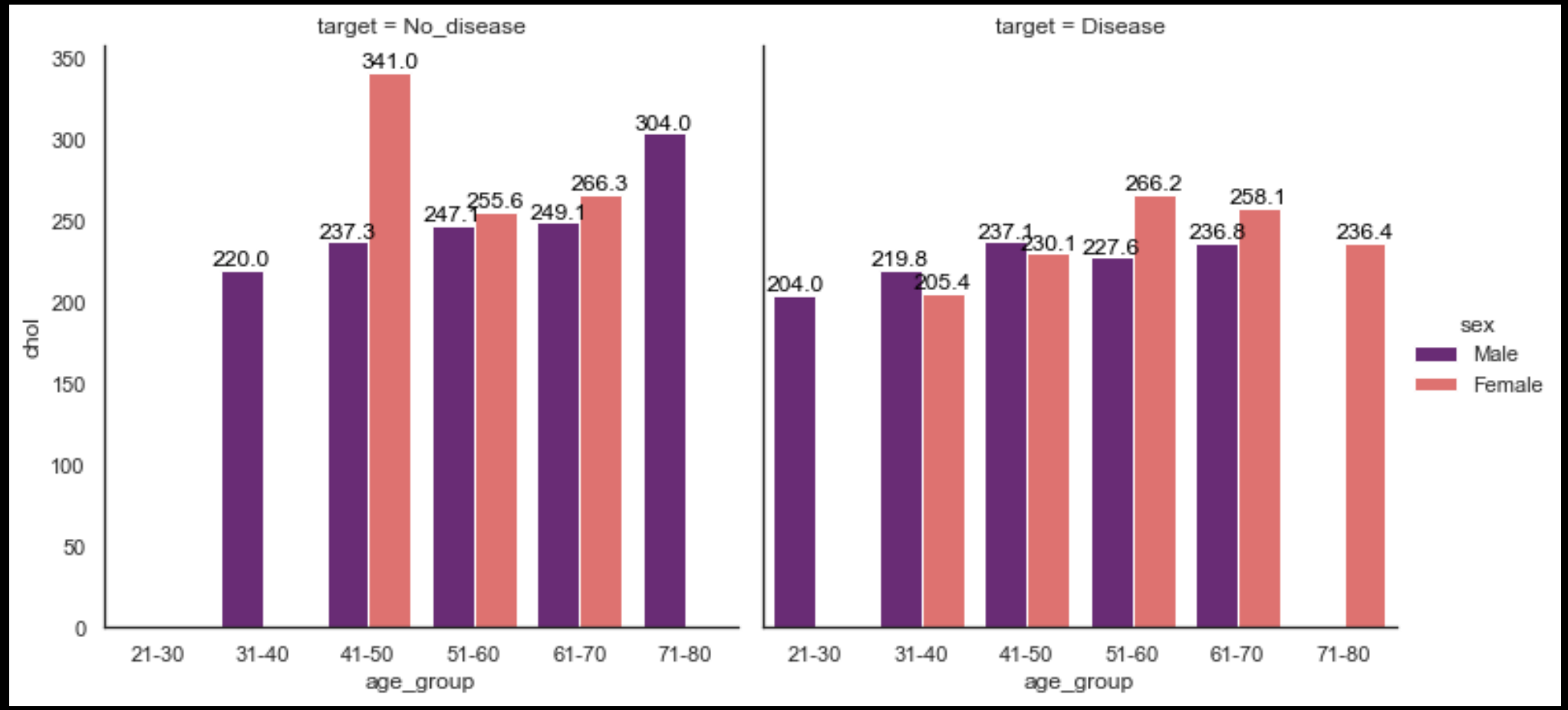
- 55.8% of people are targeted to have heart disease out of which 31.8% are males.

Chest Pain distribution according to target and gender



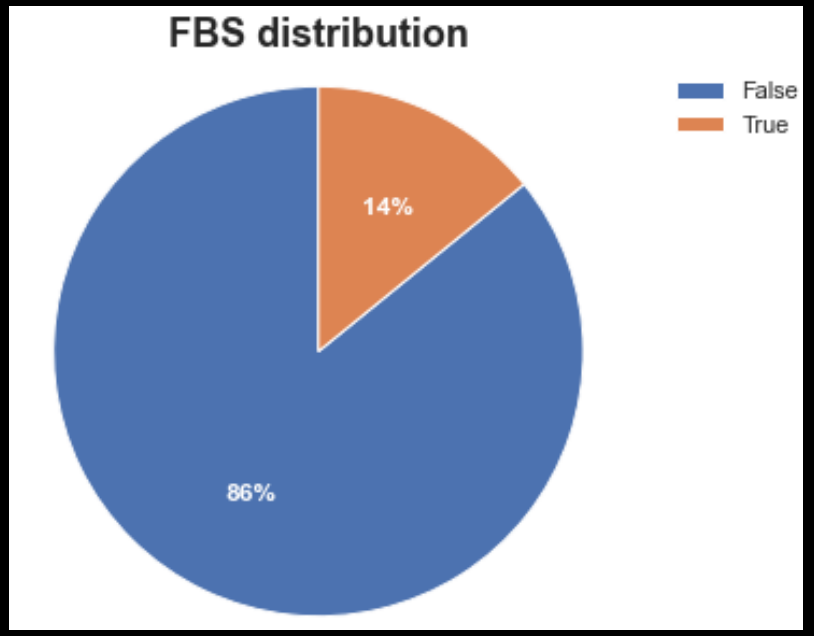


Cholesterol Distribution according to target, age and gender



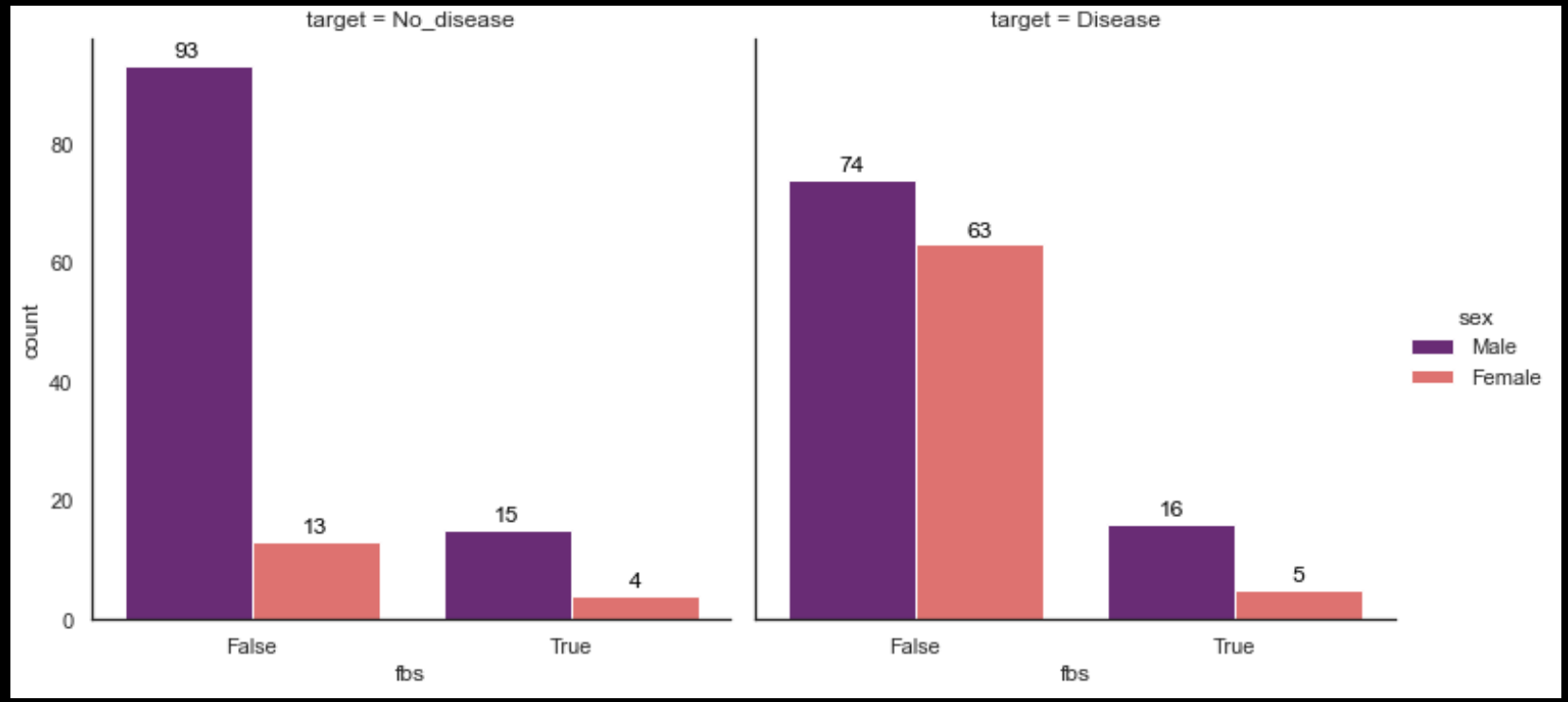
- Cholesterol is maximum for age group 51-60 yrs but these people doesn't have t disease.
- This means that high cholesterol is not the only factor. Your body still needs a little cholesterol for healthy digestion and to make vitamin D and certain hormones hence detailed analysis of cholesterol is required.

Diabetic!! So are you heart patient as well?

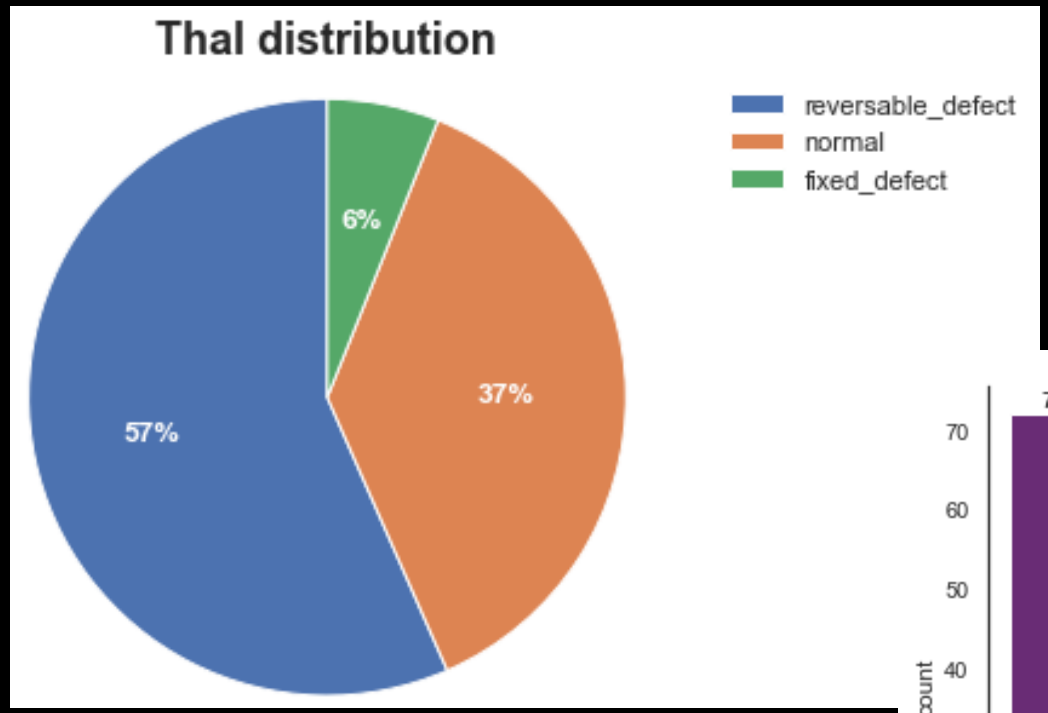


- Only 14% of the people as per data set have fasting blood sugar

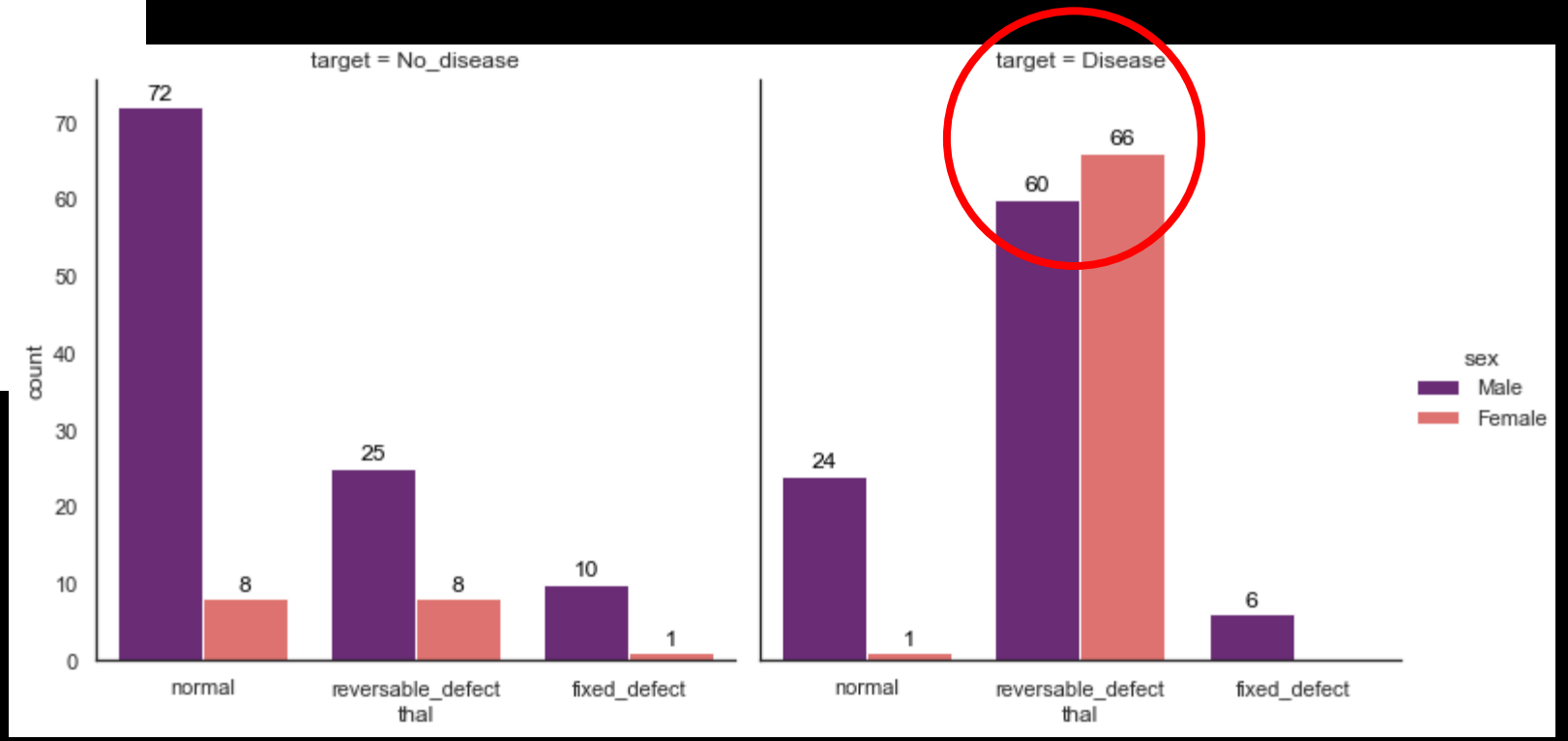
- Higher number of heart disease patient are without diabetes.
- Hence fbs won't be good indicator



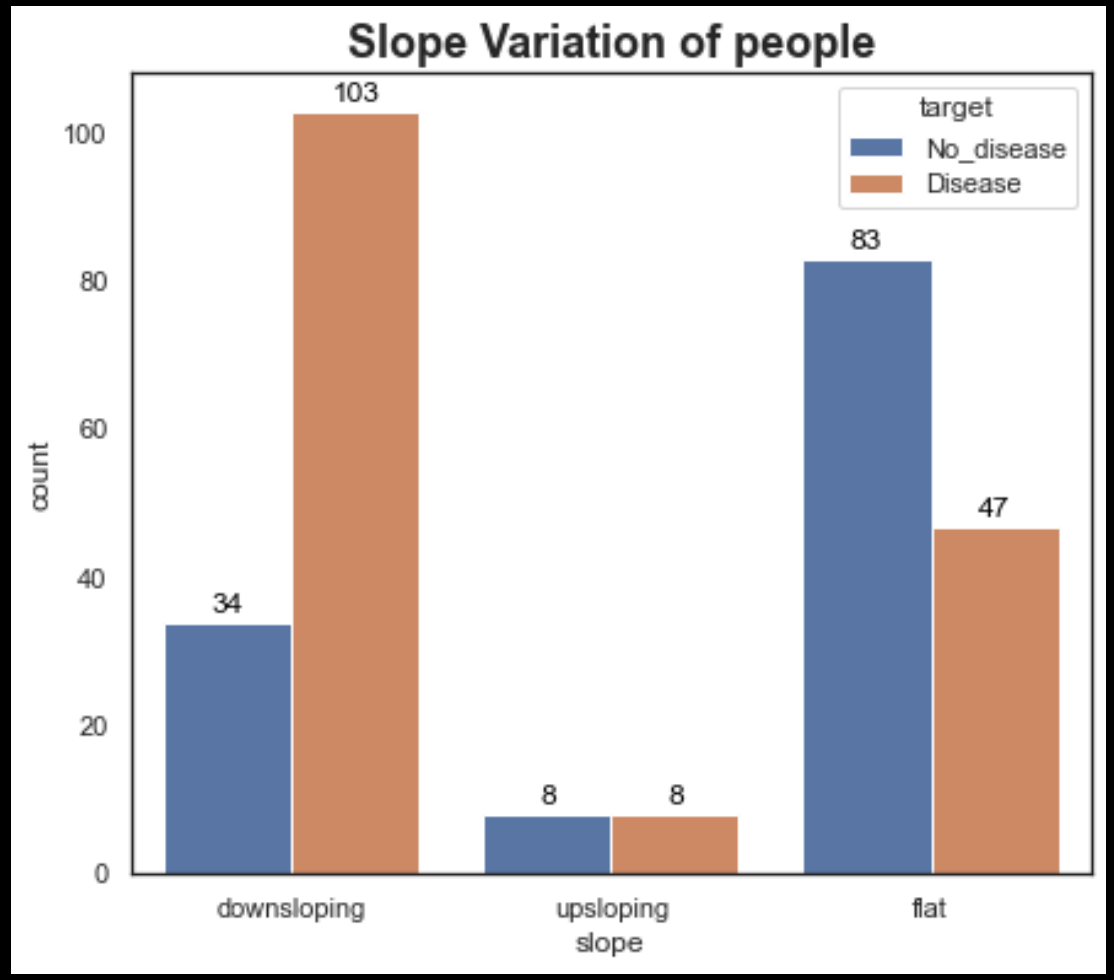
Thal distribution and it's impact



- Patients diagnosed with a reversible defect were more prevalent in the cardiac disease cohort while most non-disease patients exhibited a normal phenotype. This difference may contribute to the prediction of heart disease.



Slope Variation and its impact



- Number of people in down sloping is more thus these people with down sloping have high probability of heart disease

Machine Learning

Libraries for Algorithms

```
from sklearn.preprocessing import StandardScaler
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score
from sklearn.metrics import confusion_matrix
from sklearn import tree
from sklearn.tree import DecisionTreeClassifier
from sklearn.model_selection import GridSearchCV
from sklearn.ensemble import RandomForestClassifier
```

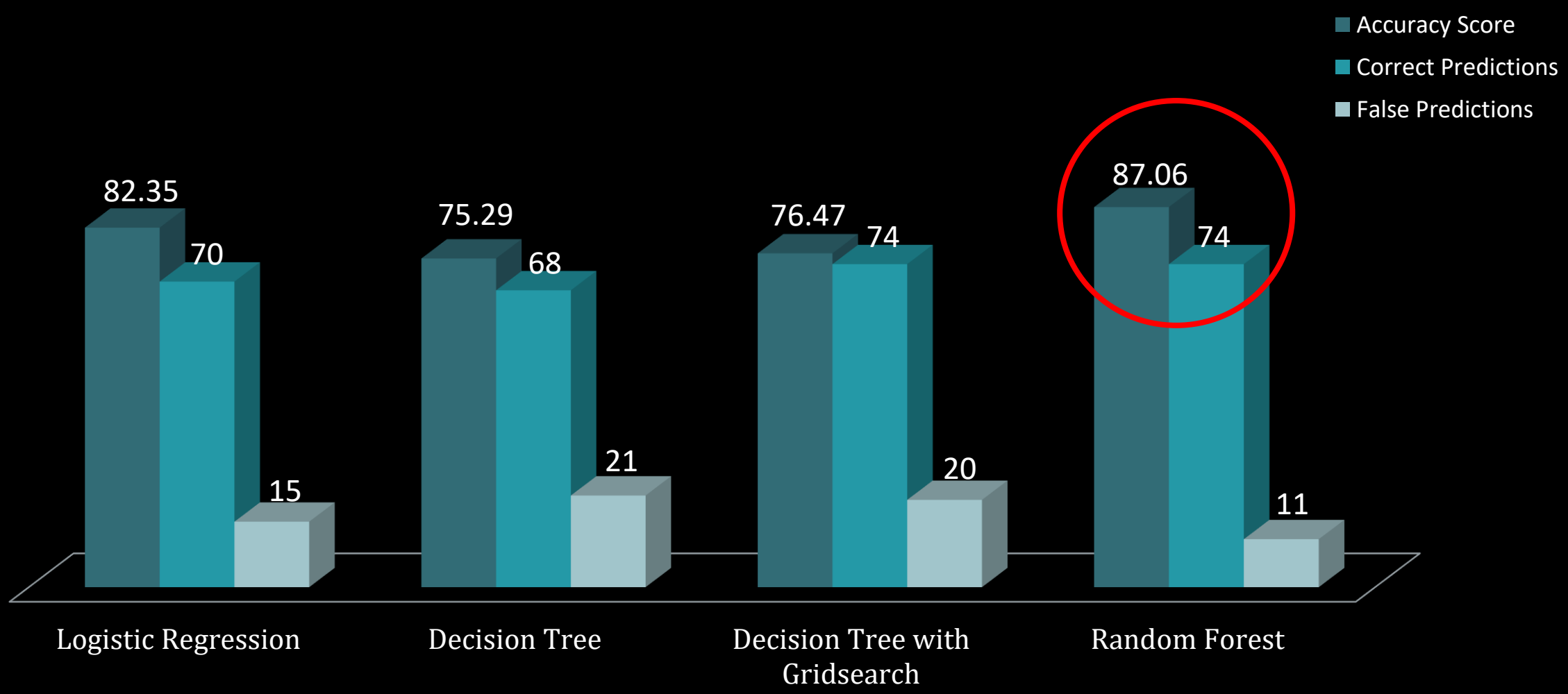
Standardization

	age	trestbps	chol	thalach	oldpeak	target	sex_Female	sex_Male
0	-0.224919	-0.317917	-0.680075	0.796365	0.046903	No_disease	0	1
1	-0.115174	0.658608	-0.881981	0.222452	2.096513	No_disease	0	1
2	1.750489	0.984117	-1.532566	-1.101964	1.608511	No_disease	0	1
3	0.762785	1.179422	-0.881981	0.487335	-0.929101	No_disease	0	1
4	0.872530	0.528405	1.159512	-1.940760	0.925307	No_disease	1	0

Train-Test Split

```
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size = .30, random_state = 0)
```

Algorithm Results



Values obtained for confusion matrix using different algorithm

Confusion Matrix describes the performance of the model. The confusion matrix obtained by the proposed model for different algorithms is shown below in Table

Algorithm	True Positive	True Negative	False Positive	False Negative	True Predictions	False Predictions
Logistic Regression	42	6	9	28	70	15
Decision Tree	36	12	9	28	68	21
Decision Tree with Grid Search	35	13	7	30	65	20
Random Forest	43	5	6	31	74	11

References

- UCI, Heart Disease Data Set : https://www.kaggle.com/ronitf/heart-disease-uci.
- Avinash Golande, Pavan Kumar T, Heart Disease Prediction Using Effective Machine Learning Techniques, International Journal of Recent Technology and Engineering, Vol 8,2019
- Nagaraj M Lutimath,Chethan C,Basavaraj S Pol.,Prediction Of Heart Disease using Machine Learning, International journal Of Recent Technology and Engineering,8,(2S10), 2019.
- Images from www.google.com


Thank you