Assignment 1

Predicting Heart Disease

Dilanka Rathnasiri

1. **Summery of the dataset**

This dataset has 302 records. This dataset has 13 features and the target field. They are,

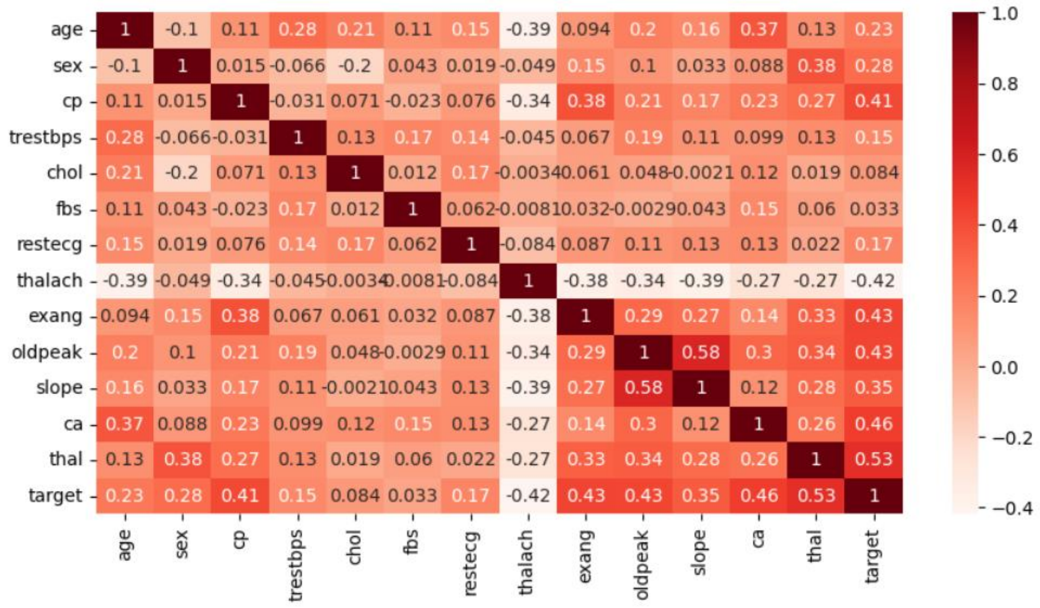
- age
- sex
- cp
- trestbps
- chol
- fbs
- restecg
- thalach
- exang
- oldpeak
- slope
- ca
- thal

“ca” and “thal” features has “?” values. They were replaced by the median of each relative feature.

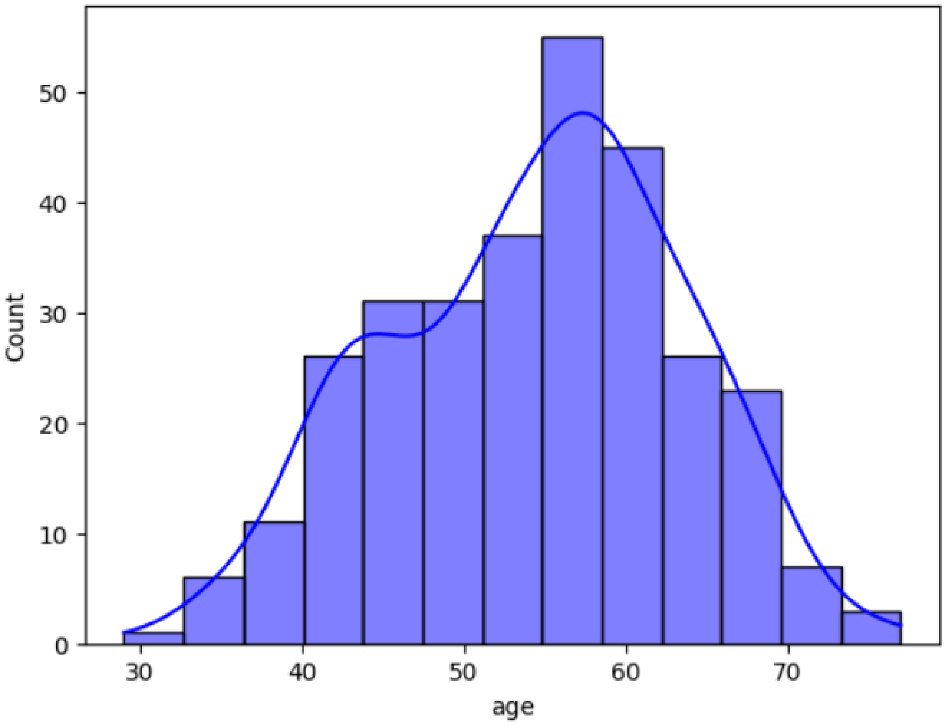
“num” field is the target of the dataset. Target has 5 categories. They are “no disease”, from 1 to 4 varying degrees of disease. Target was converted to binary categorical target. Then, 0 is no disease and 1 is has disease. So, modification was done as following,

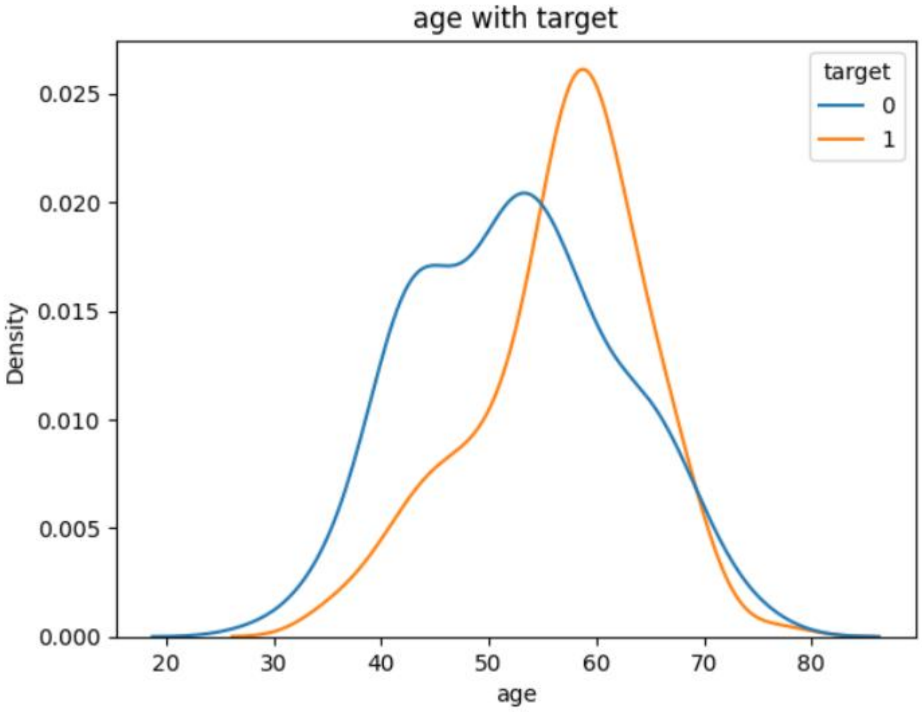
- 0 → 0
- 1,2,3,4 → 1

Correlation heat map is as follows,

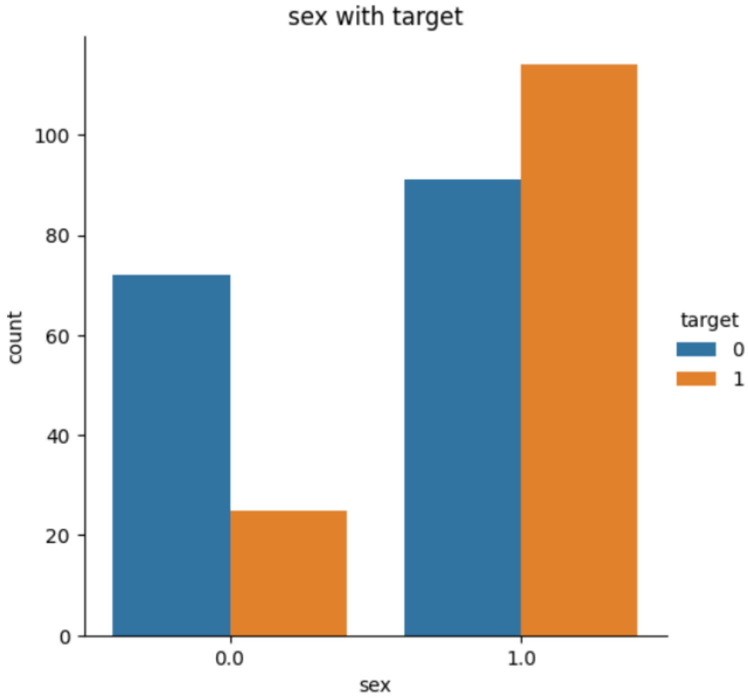


Plots for age feature is as follows,

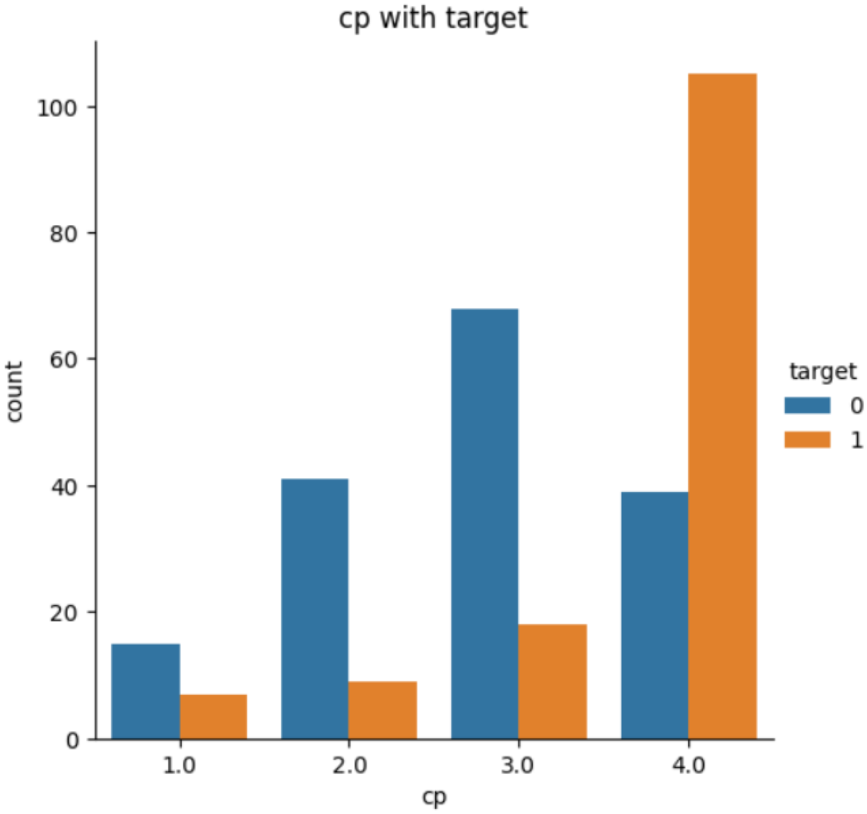




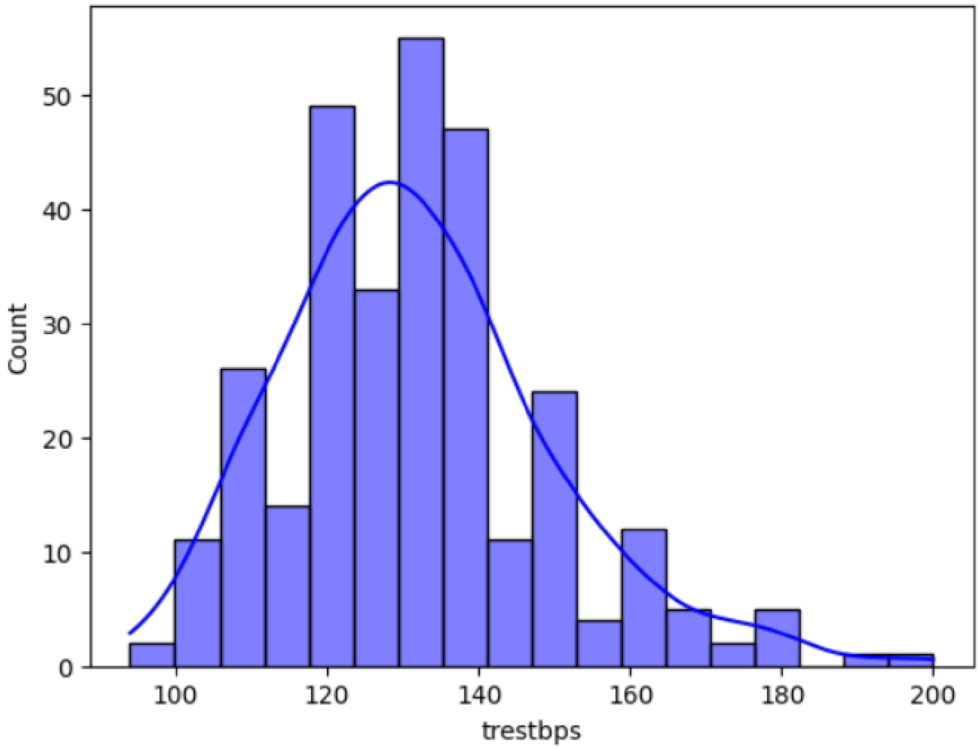
Plots for sex feature is as follows,

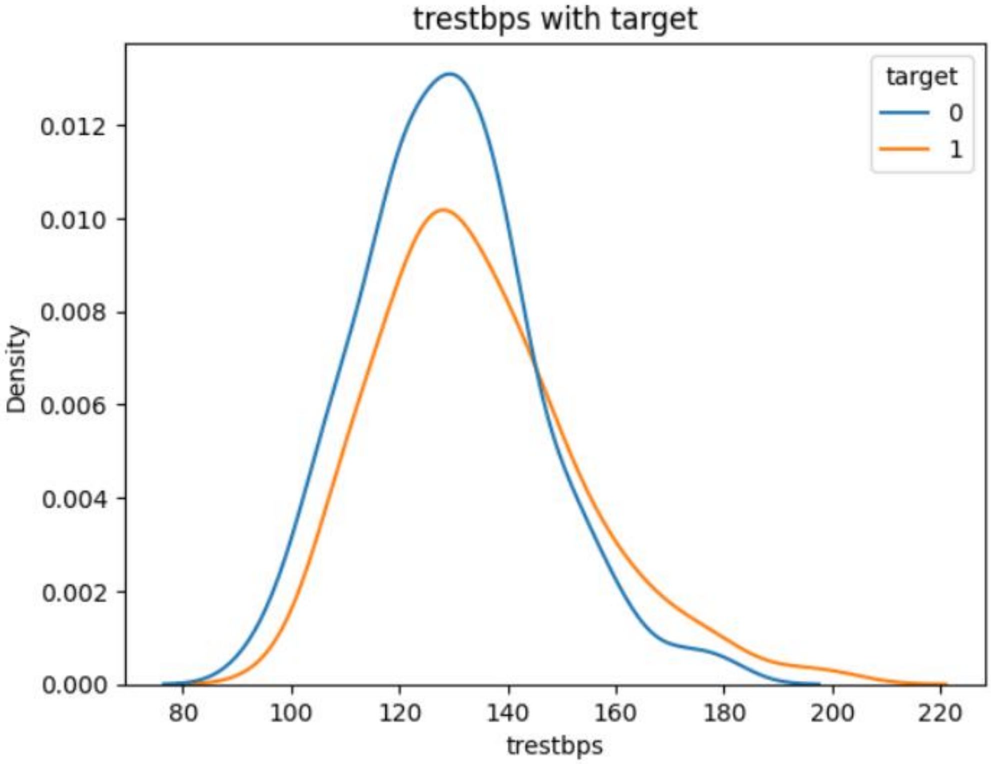


Plots for cp feature is as follows,

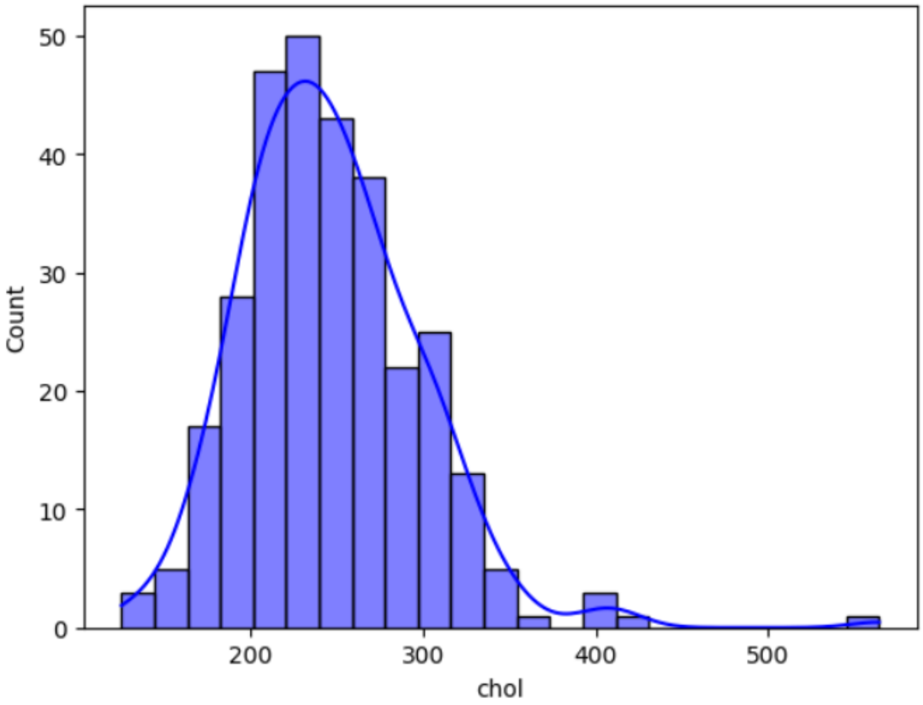


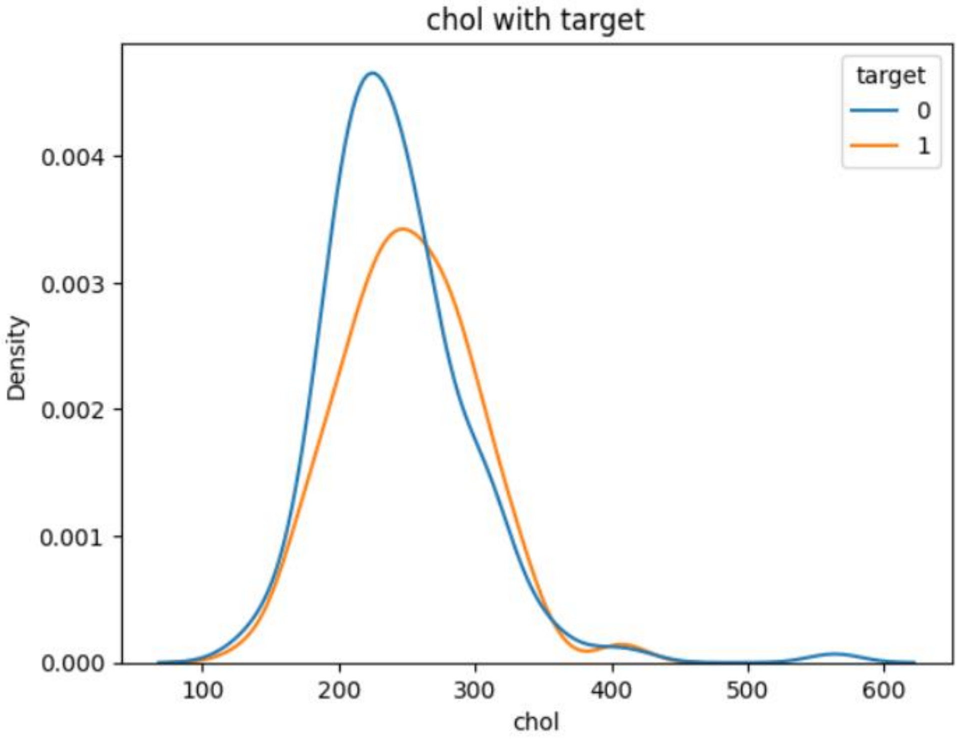
Plots for trestbps feature is as follows,



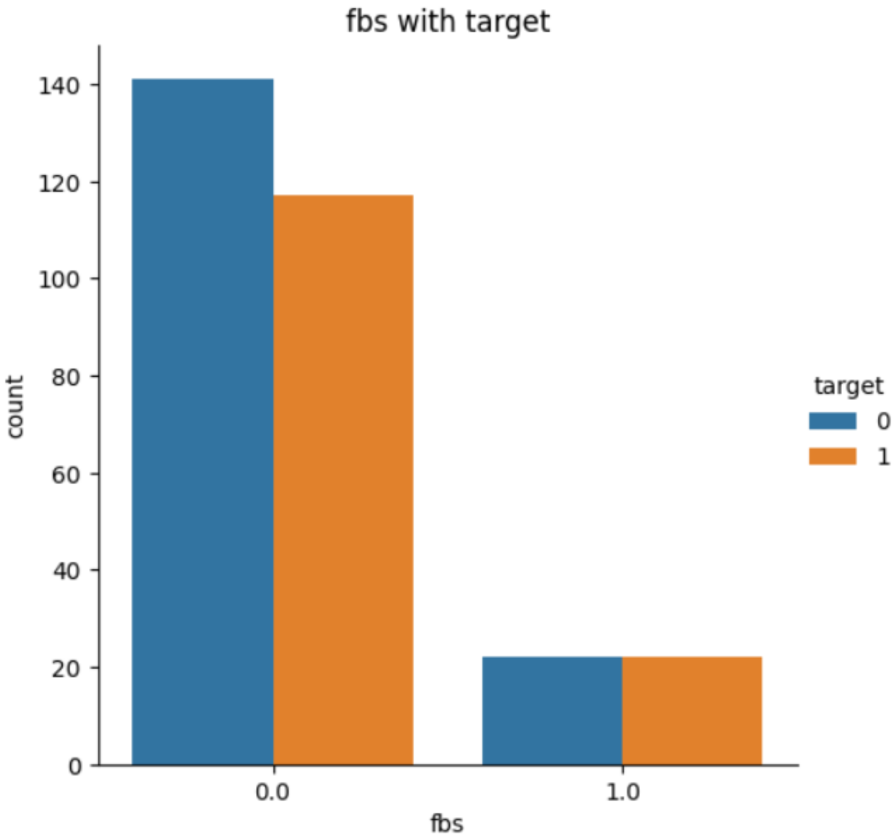


Plots for chol feature is as follows,

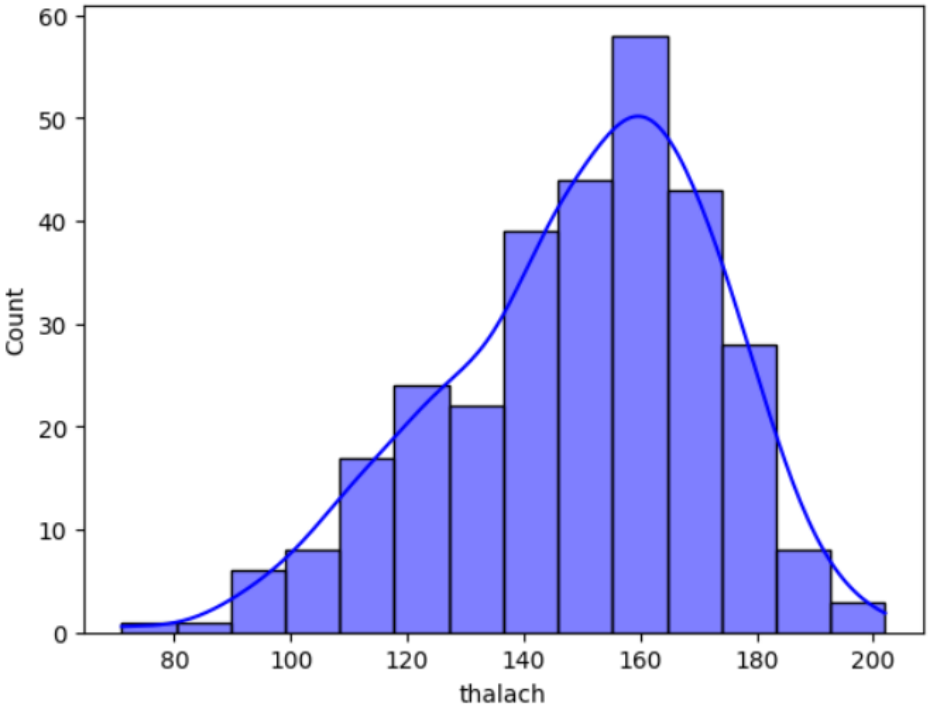


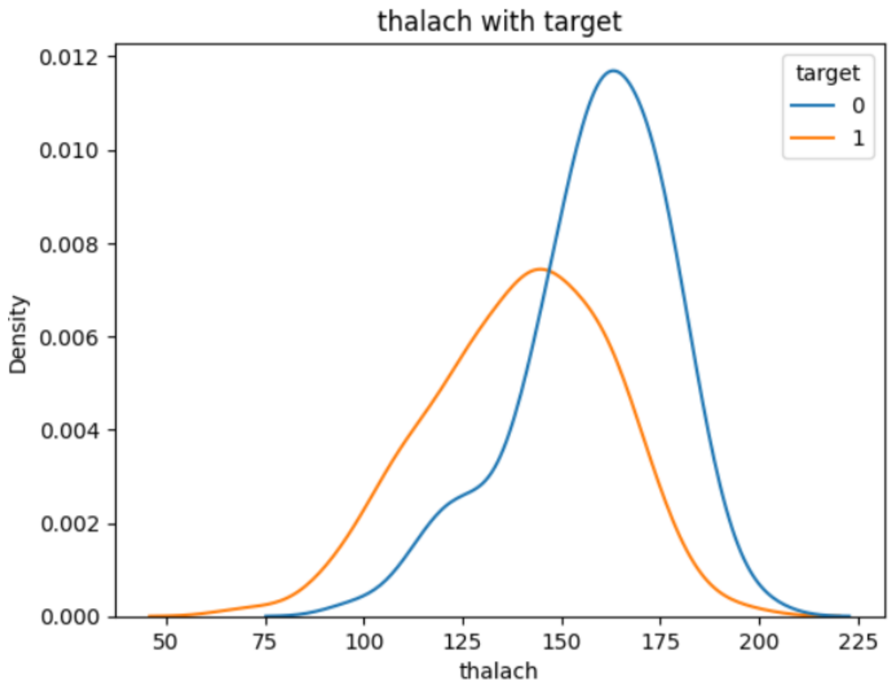


Plots for fbs feature is as follows,

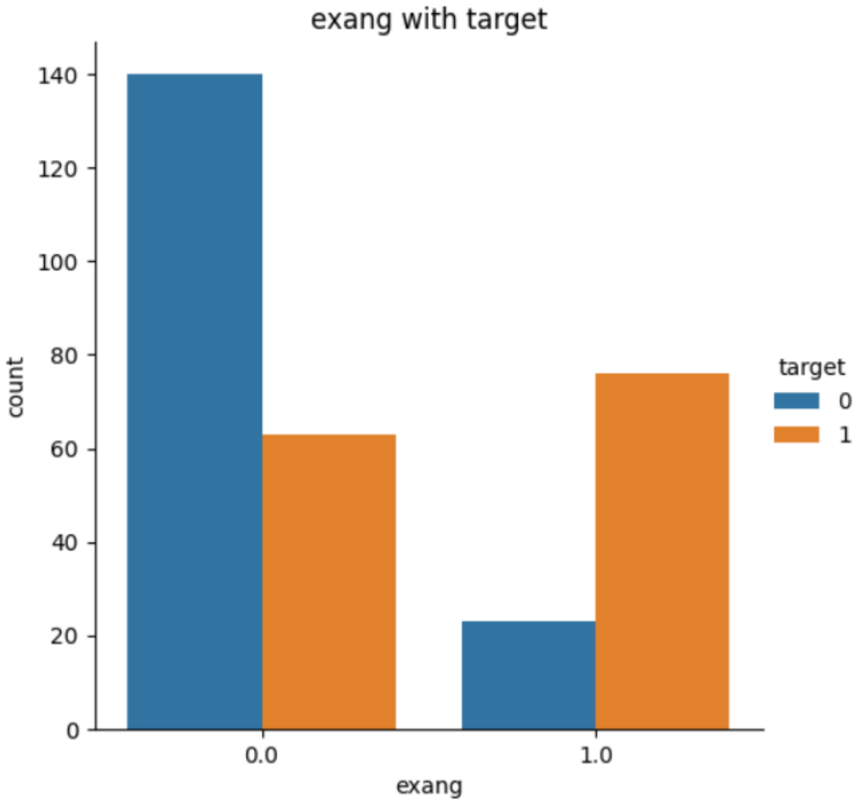


Plots for thalach feature is as follows,

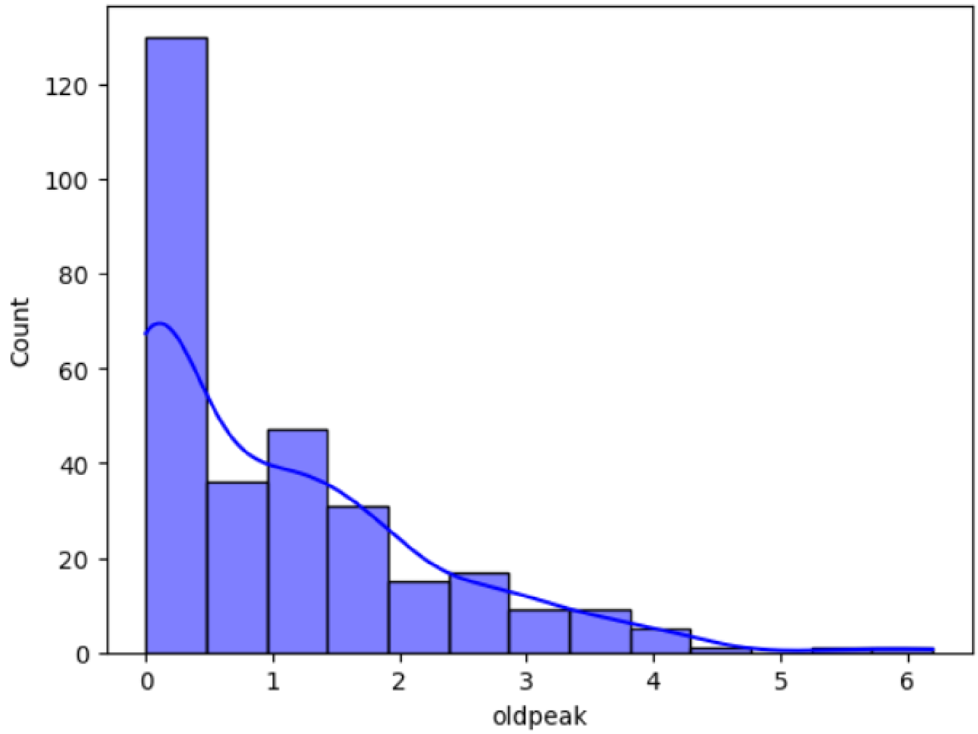


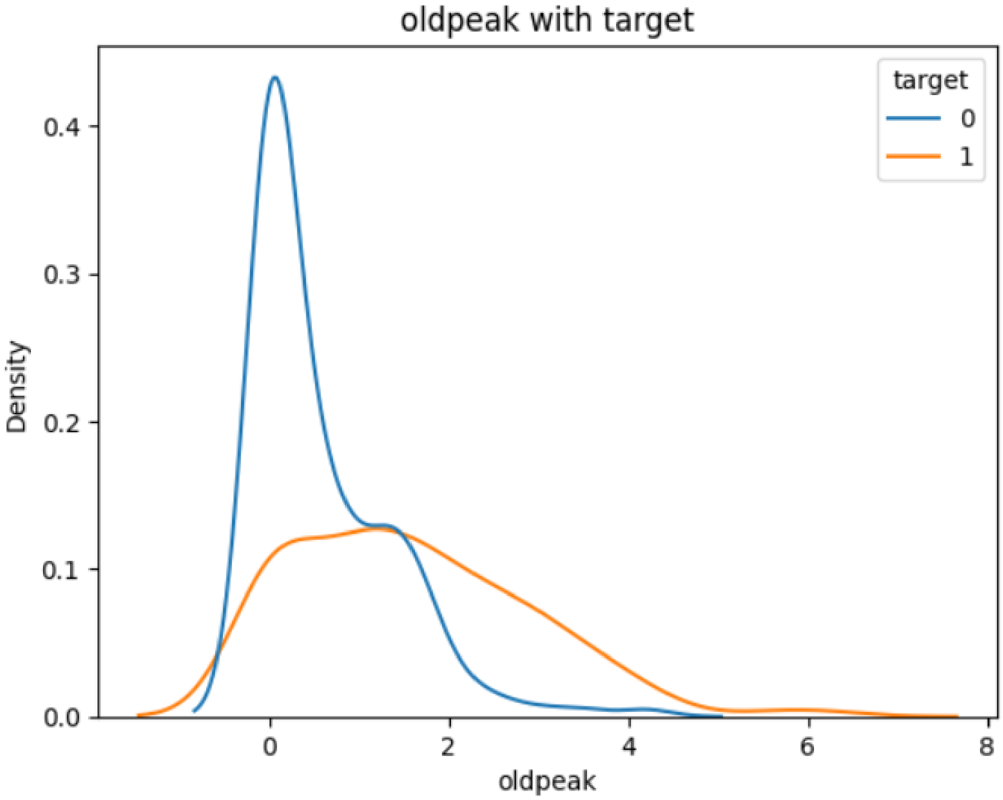


Plots for exang feature is as follows,

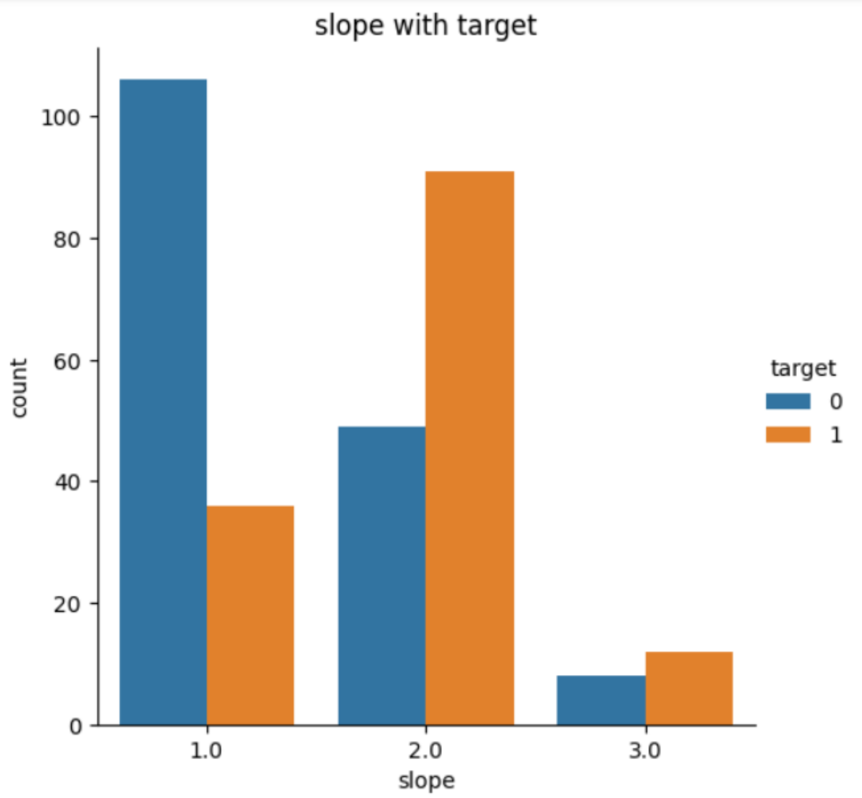


Plots for oldpeak feature is as follows,

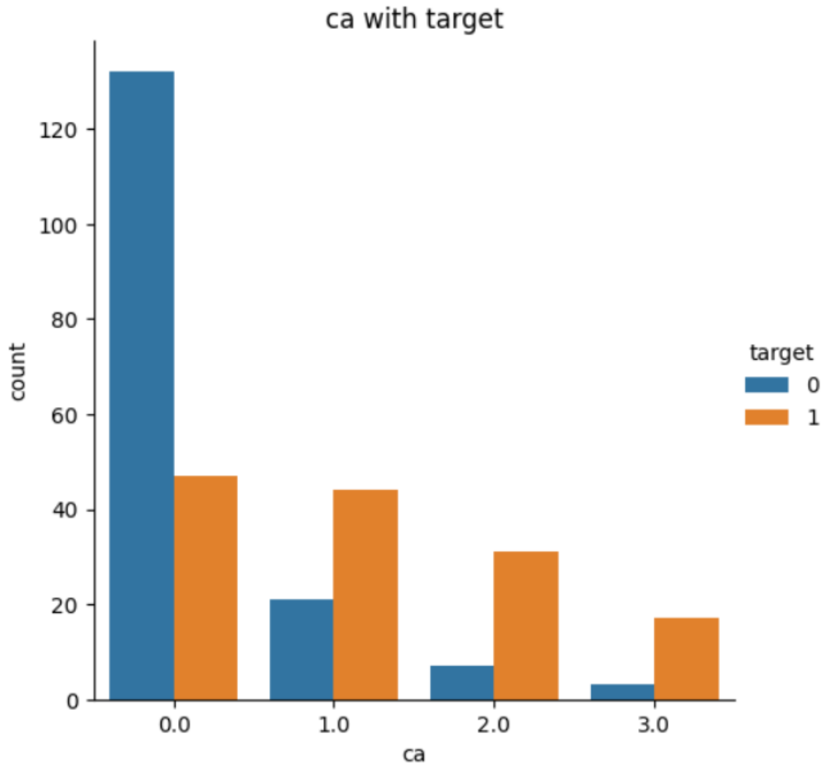




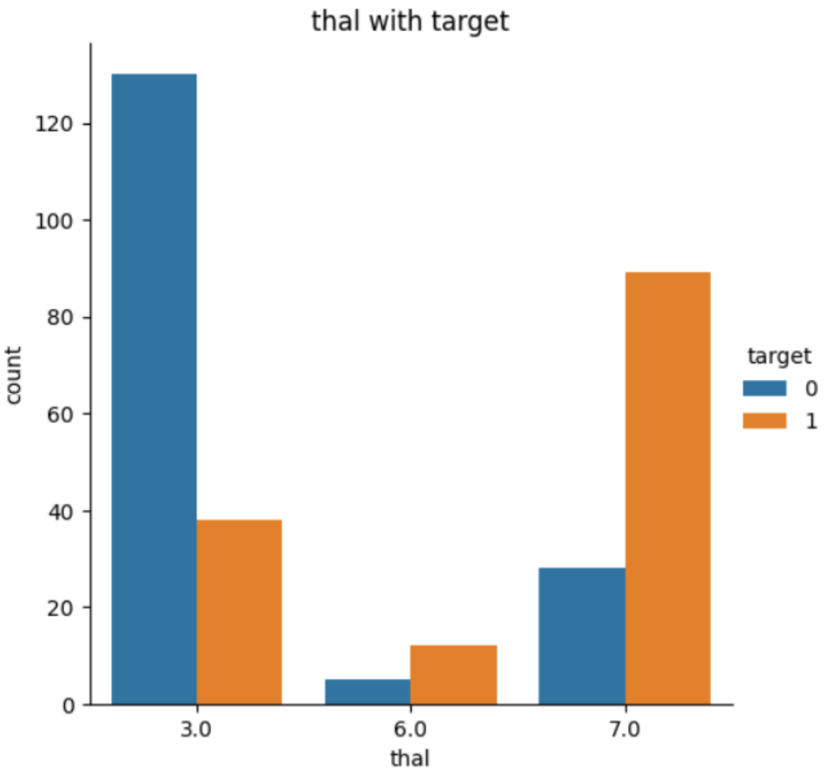
Plots for slope feature is as follows,



Plots for ca feature is as follows,



Plots for thal feature is as follows,



2. Data modelling

Following Machine Learning models was used for predictions.

- Logistic Regression
- Support Vector Machine
- Random Forest Classifier
- Naive Bayes
- K-Nearest Neighbour
- Extreme Gradient Boost

Accuracies for each model are shown in the following table.

Model	Accuracy
Logistic Regression	88.52459016393442%
Support Vector Machine	83.60655737704919%
Random Forest Classifier	85.24590163934425%
Naive Bayes	85.24590163934425%
K-Nearest Neighbour	81.9672131147541%
Extreme Gradient Boost	86.88524590163934%

Therefore, we can observe **Logistic Regression** model has the highest accuracy.

So, we can use logistics regression model to predict the heart disease.