Deepiotics

Table of contents

1. About Problem Statement and Data Set

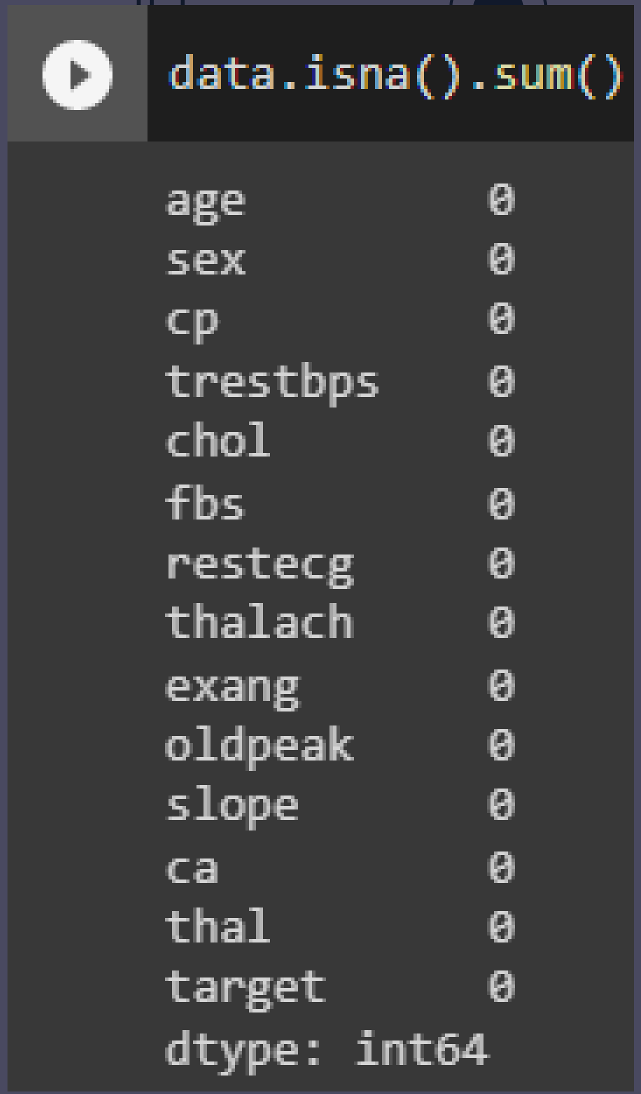
2. Exploratory Data Analysis

3. Data Preprocessing

2. Predictions & ML models Accuracy

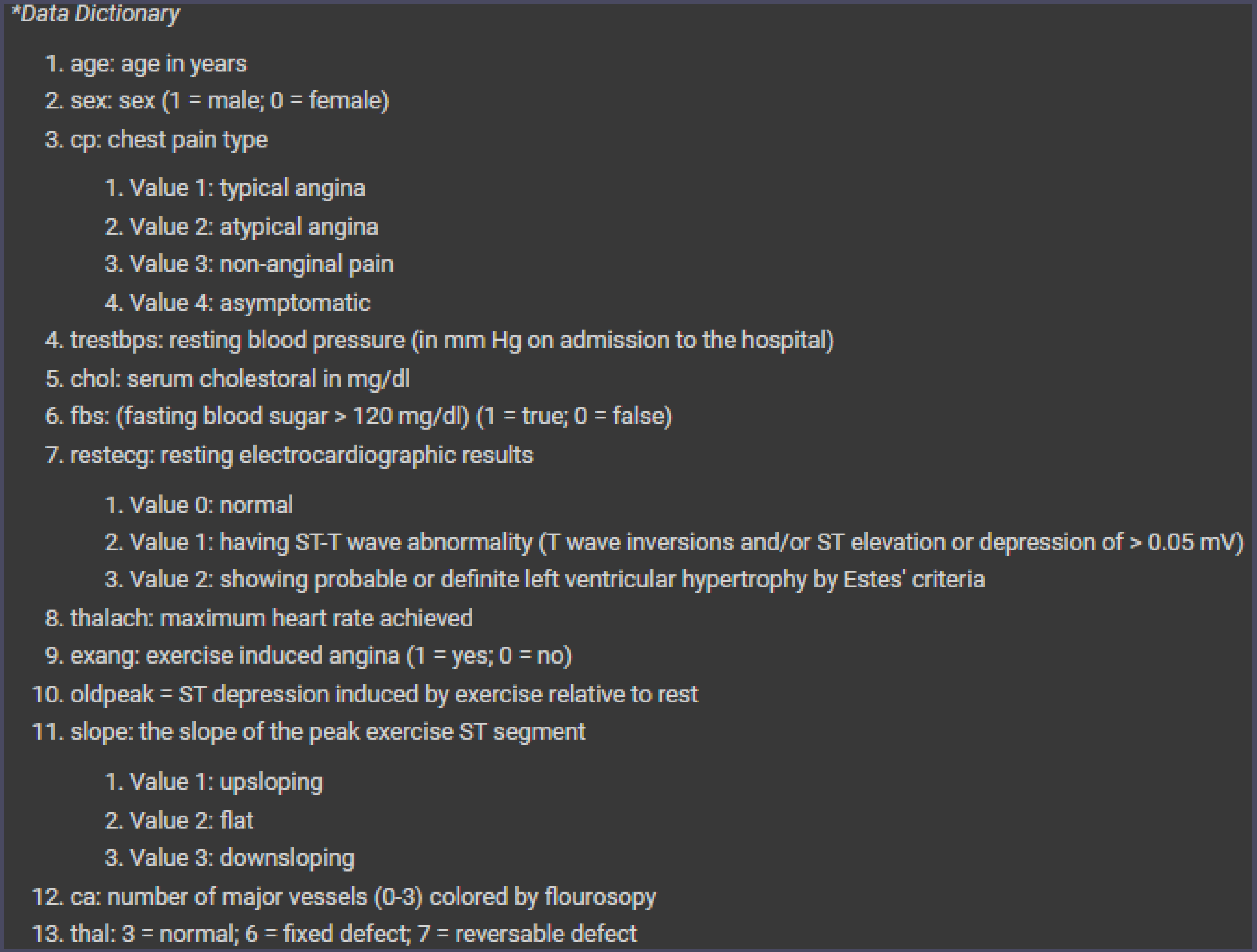
About Data Set

- The data set consists of 1025 data points with 13 features and 1 target variable.
- Out of 13 features 8 are categorical and 5 are numeric.
- The data set includes no missing values.



```
data.isna().sum()
```

```
age         0
sex         0
cp          0
trestbps    0
chol        0
fbs         0
restecg     0
thalach     0
exang       0
oldpeak     0
slope       0
ca          0
thal        0
target      0
dtype: int64
```



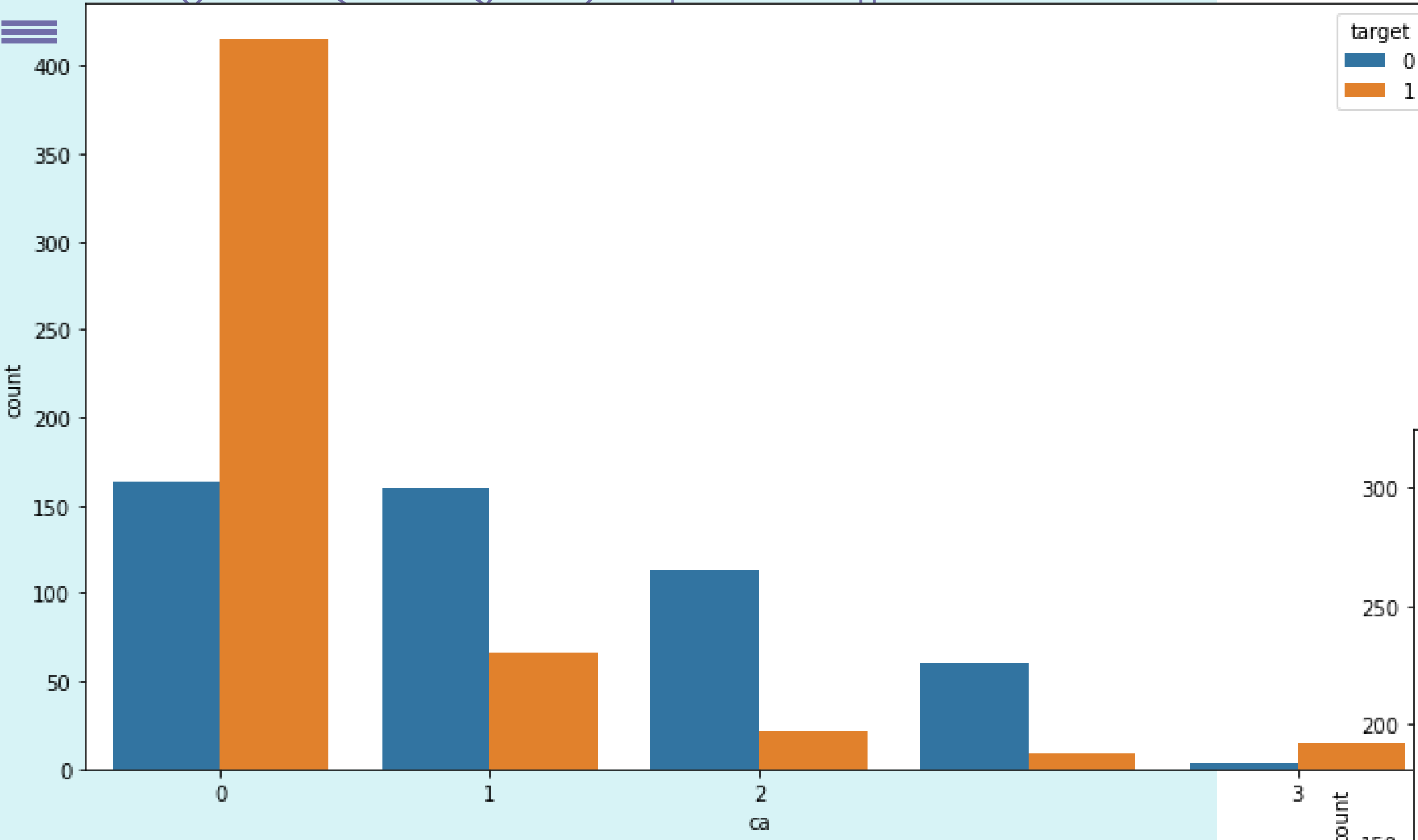
**Data Dictionary*

1. age: age in years
2. sex: sex (1 = male; 0 = female)
3. cp: chest pain type
    1. Value 1: typical angina
    2. Value 2: atypical angina
    3. Value 3: non-anginal pain
    4. Value 4: asymptomatic
4. trestbps: resting blood pressure (in mm Hg on admission to the hospital)
5. chol: serum cholestoral in mg/dl
6. fbs: (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
7. restecg: resting electrocardiographic results
    1. Value 0: normal
    2. Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)
    3. Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
8. thalach: maximum heart rate achieved
9. exang: exercise induced angina (1 = yes; 0 = no)
10. oldpeak = ST depression induced by exercise relative to rest
11. slope: the slope of the peak exercise ST segment
    1. Value 1: upsloping
    2. Value 2: flat
    3. Value 3: downsloping
12. ca: number of major vessels (0-3) colored by flourosopy
13. thal: 3 = normal; 6 = fixed defect; 7 = reversable defect

Problem Statement

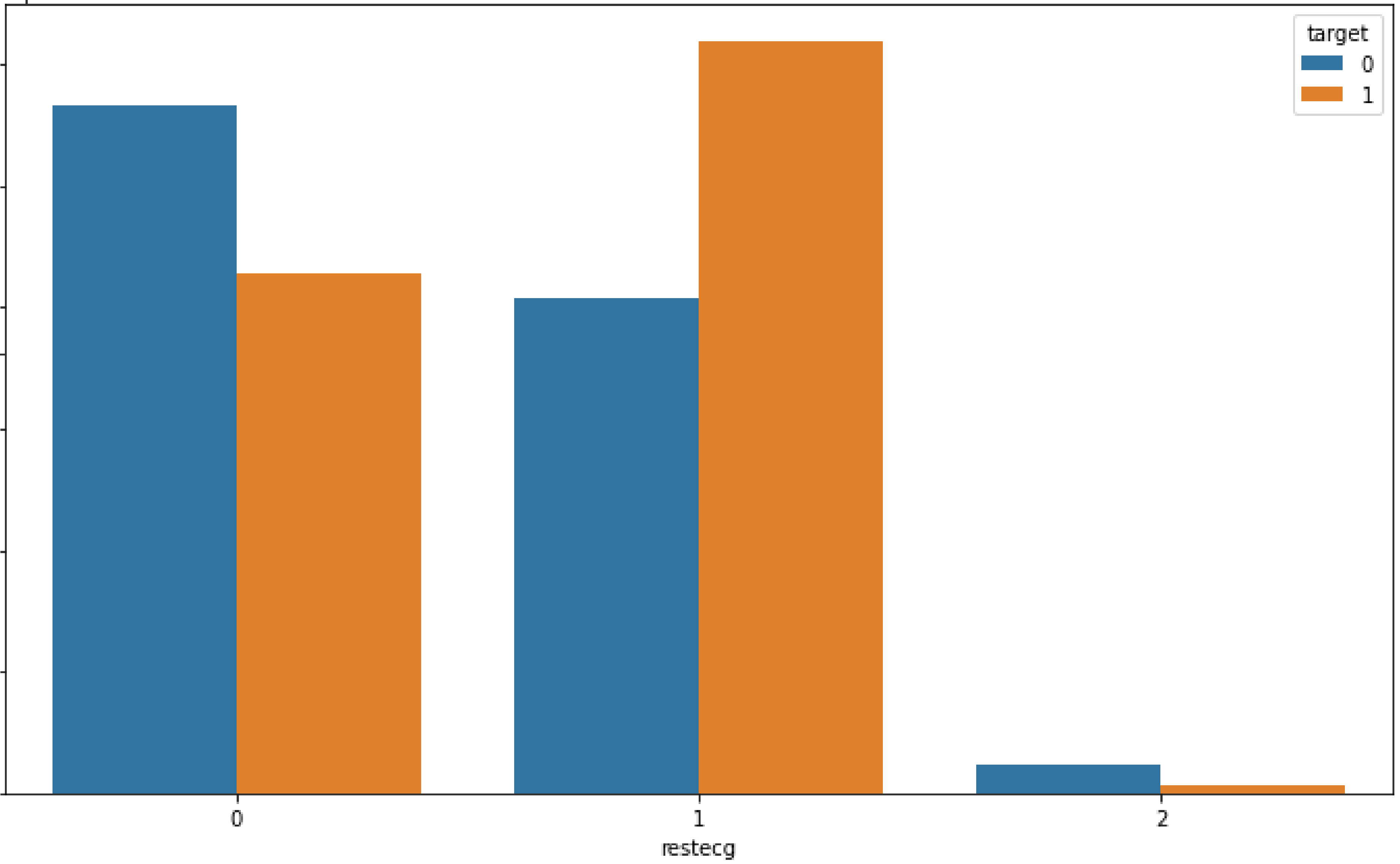
Its a Heart Disease Classification Problem.
The dataset is collected from Kaggle .
Data set original source : https://bit.ly/3Tih2sY

EXPLORATORY DATA ANALYSIS

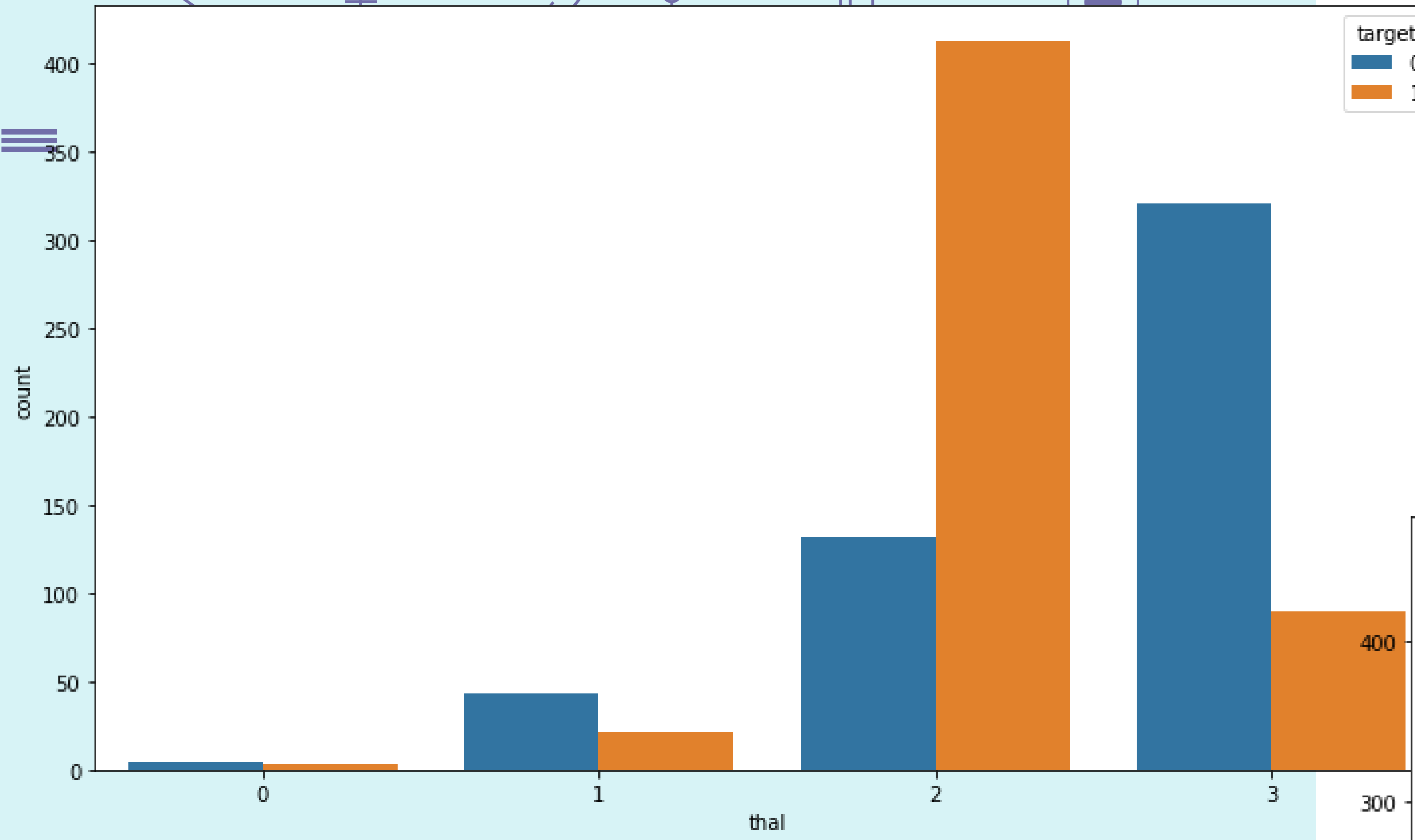


NUMBER OF MAJOR VESSELS

The below analysis is about Resting Electrocardiographic .From above analysis it is clear that those with Type 1 restecg mostly have heart disease but those with Type 0 / Type 2 restecg mostly donot have heart disease.

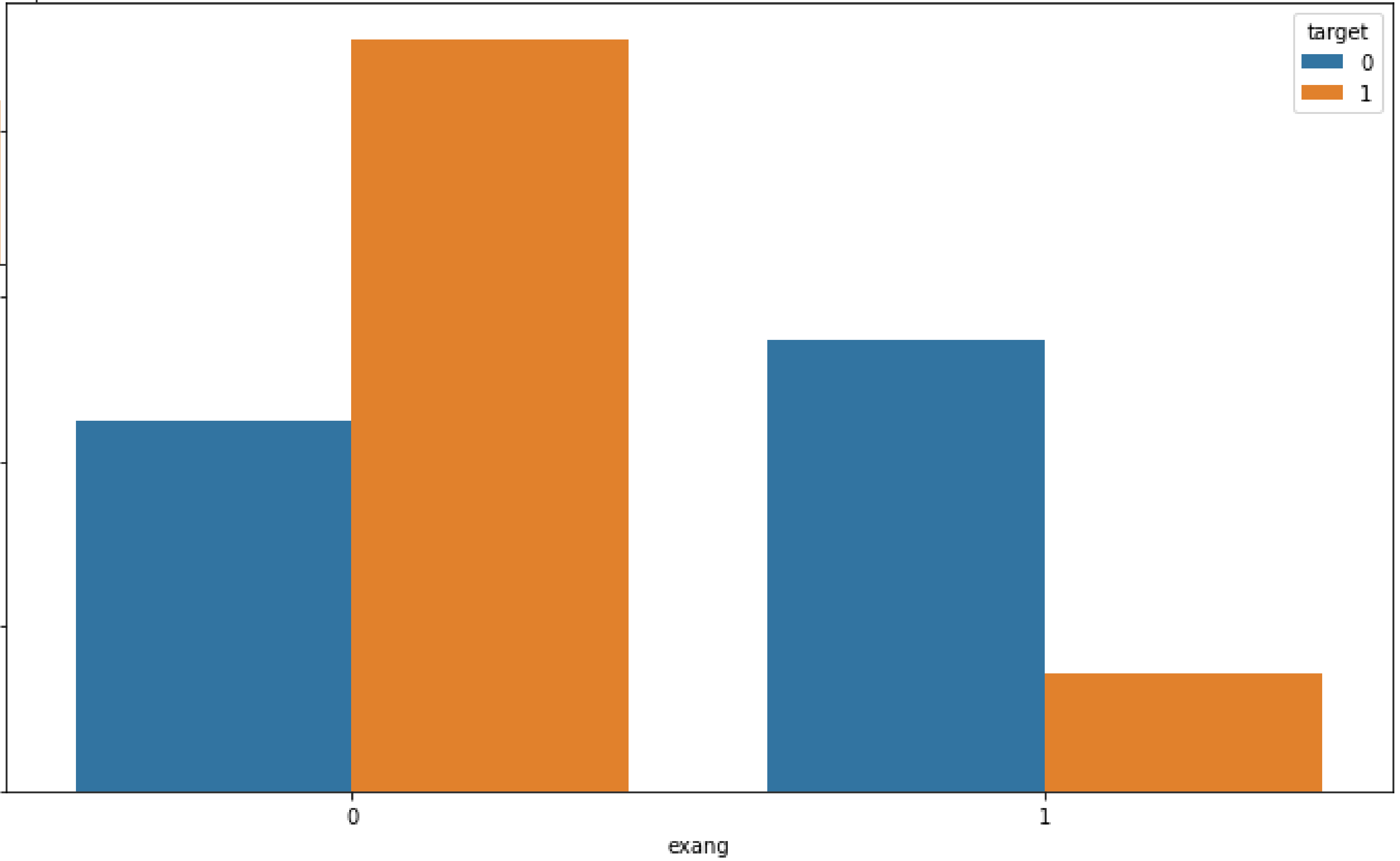


EXPLORATORY DATA ANALYSIS

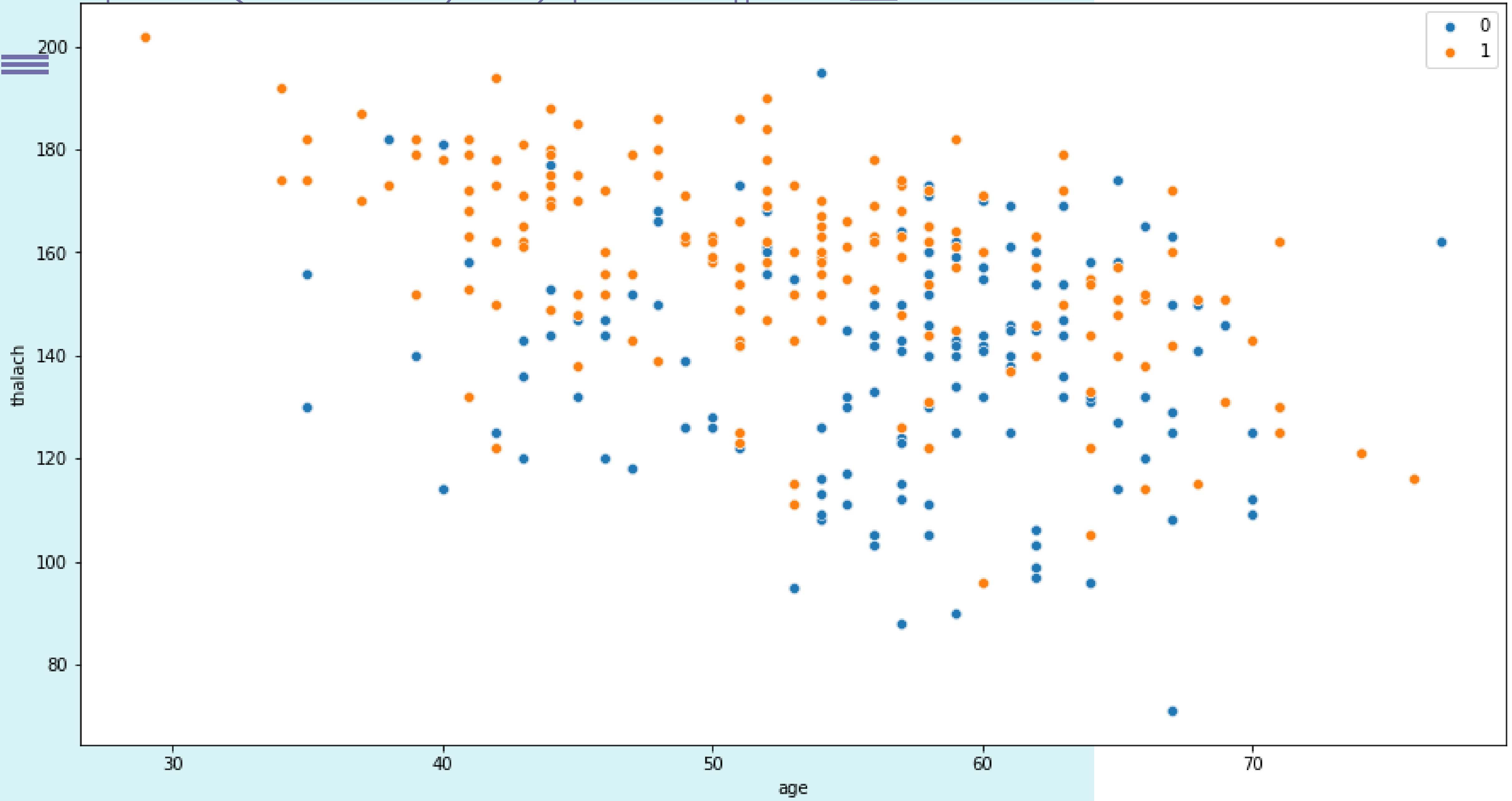


THALH

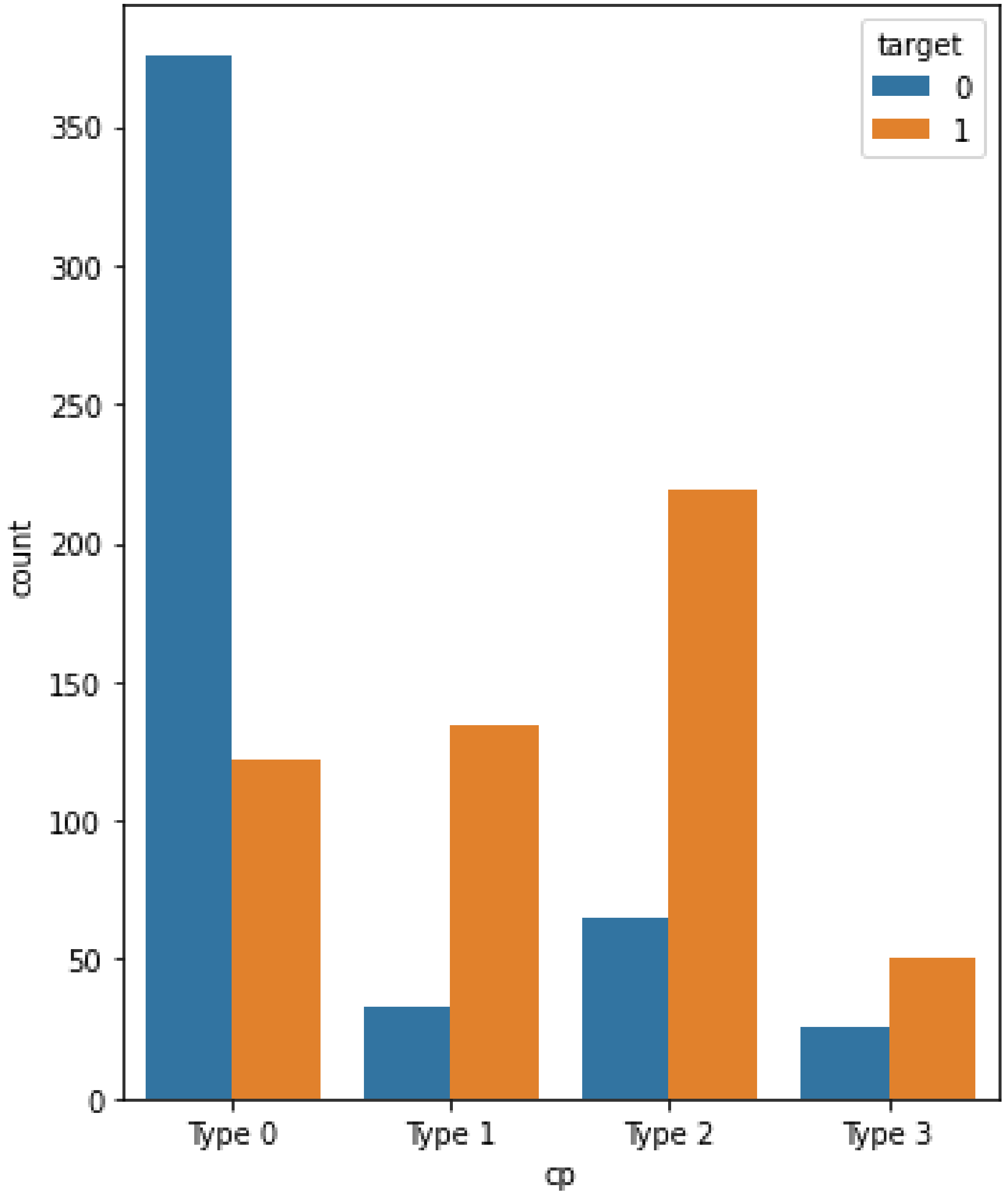
EXERCISE INDUCED ANGINA



EXPLORATORY DATA ANALYSIS

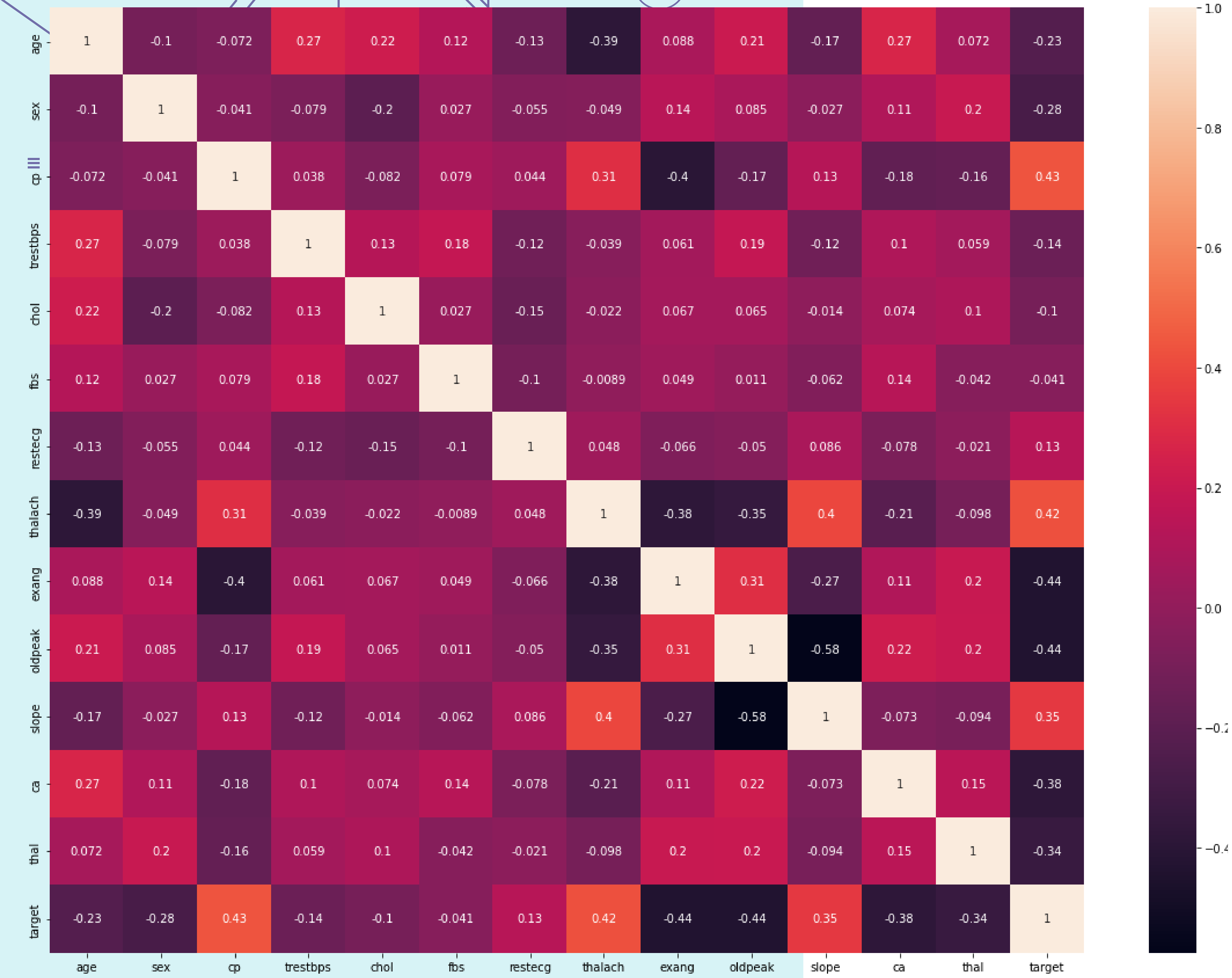


Based on age, patients with and without heart diseases mostly between 50-70 years old. Patients with heart diseases tend to have high heart rate\ncompared to patients with no heart diseases.'



The above analysis is about Chest pain. It is clear that those with Type 0 / Value 1 type of chest pain mostly do not have heart disease but those with Type 2 / Value 3 type of chest pain mostly do have heart disease.

EXPLORATORY DATA ANALYSIS



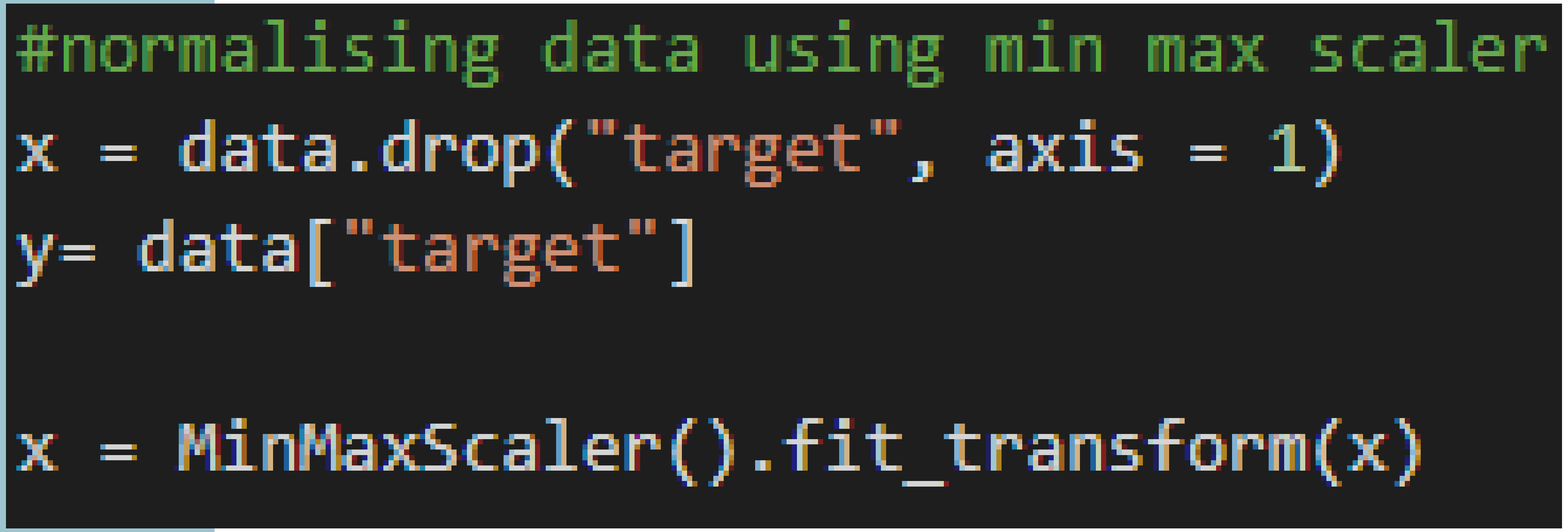
	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
age	1	-0.1	-0.072	0.27	0.22	0.12	-0.13	-0.39	0.088	0.21	-0.17	0.27	0.072	-0.23
sex	-0.1	1	-0.041	-0.079	-0.2	0.027	-0.055	-0.049	0.14	0.085	-0.027	0.11	0.2	-0.28
cp	-0.072	-0.041	1	0.038	-0.082	0.079	0.044	0.31	-0.4	-0.17	0.13	-0.18	-0.16	0.43
trestbps	0.27	-0.079	0.038	1	0.13	0.18	-0.12	-0.039	0.061	0.19	-0.12	0.1	0.059	-0.14
chol	0.22	-0.2	-0.082	0.13	1	0.027	-0.15	-0.022	0.067	0.065	-0.014	0.074	0.1	-0.1
fbs	0.12	0.027	0.079	0.18	0.027	1	-0.1	-0.0089	0.049	0.011	-0.062	0.14	-0.042	-0.041
restecg	-0.13	-0.055	0.044	-0.12	-0.15	-0.1	1	0.048	-0.066	-0.05	0.086	-0.078	-0.021	0.13
thalach	-0.39	-0.049	0.31	-0.039	-0.022	-0.0089	0.048	1	-0.38	-0.35	0.4	-0.21	-0.098	0.42
exang	0.088	0.14	-0.4	0.061	0.067	0.049	-0.066	-0.38	1	0.31	-0.27	0.11	0.2	-0.44
oldpeak	0.21	0.085	-0.17	0.19	0.065	0.011	-0.05	-0.35	0.31	1	-0.58	0.22	0.2	-0.44
slope	-0.17	-0.027	0.13	-0.12	-0.014	-0.062	0.086	0.4	-0.27	-0.58	1	-0.073	-0.094	0.35
ca	0.27	0.11	-0.18	0.1	0.074	0.14	-0.078	-0.21	0.11	0.22	-0.073	1	0.15	-0.38
thal	0.072	0.2	-0.16	0.059	0.1	-0.042	-0.021	-0.098	0.2	0.2	-0.094	0.15	1	-0.34
target	-0.23	-0.28	0.43	-0.14	-0.1	-0.041	0.13	0.42	-0.44	-0.44	0.35	-0.38	-0.34	1

HEAT MAP

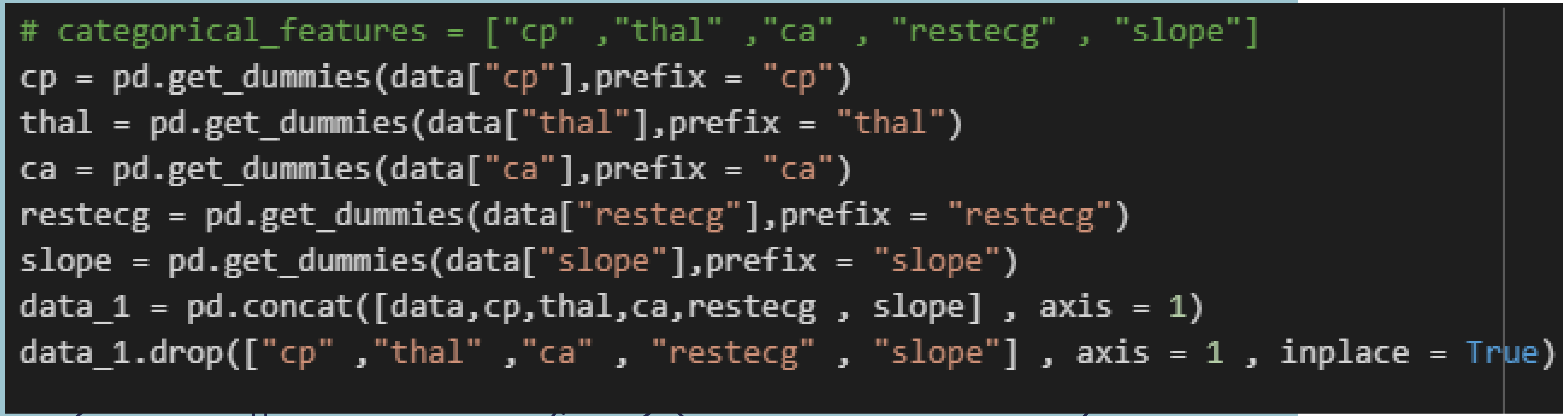
All features correlation is plotted in the form of heat map

Data Preprocessing

- Dependent and independent variables are seprated.
- One Hot Encoding is done on all categorical features with more than 2 categories
- Independent variables are **Normalized** using Min Max Scalar .
- Splitting the data into train and test data.



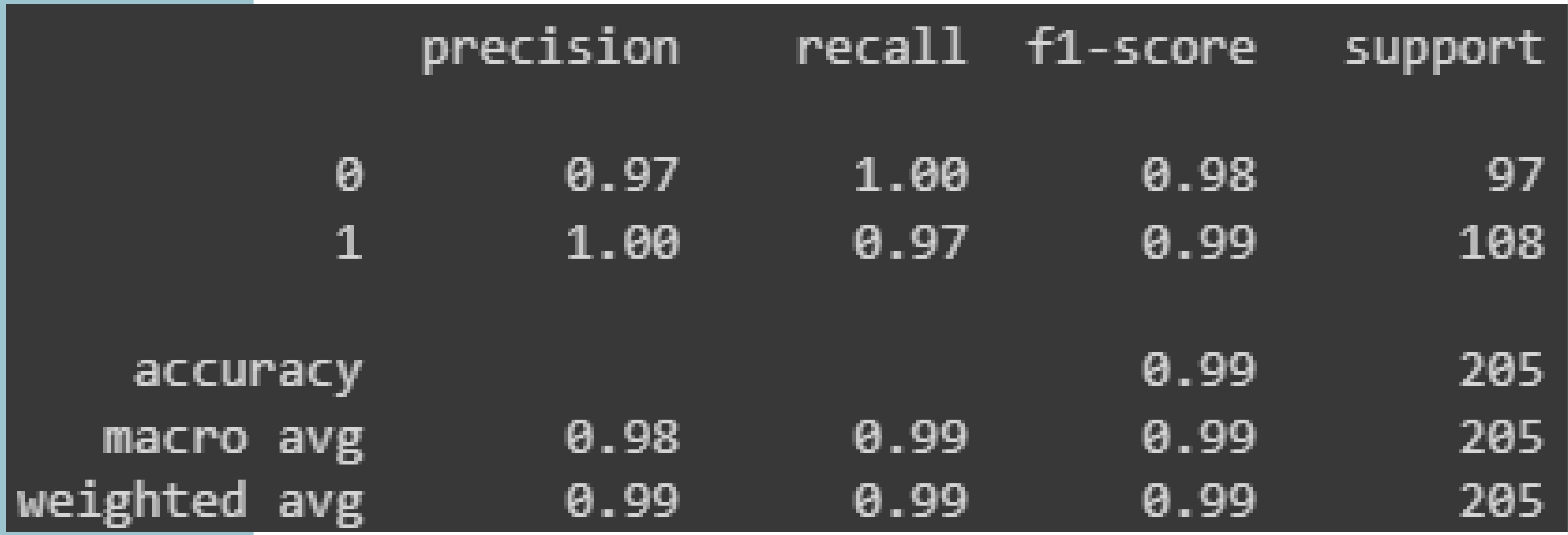
```
#normalising data using min max scaler
x = data.drop("target", axis = 1)
y= data["target"]

x = MinMaxScaler().fit_transform(x)
```



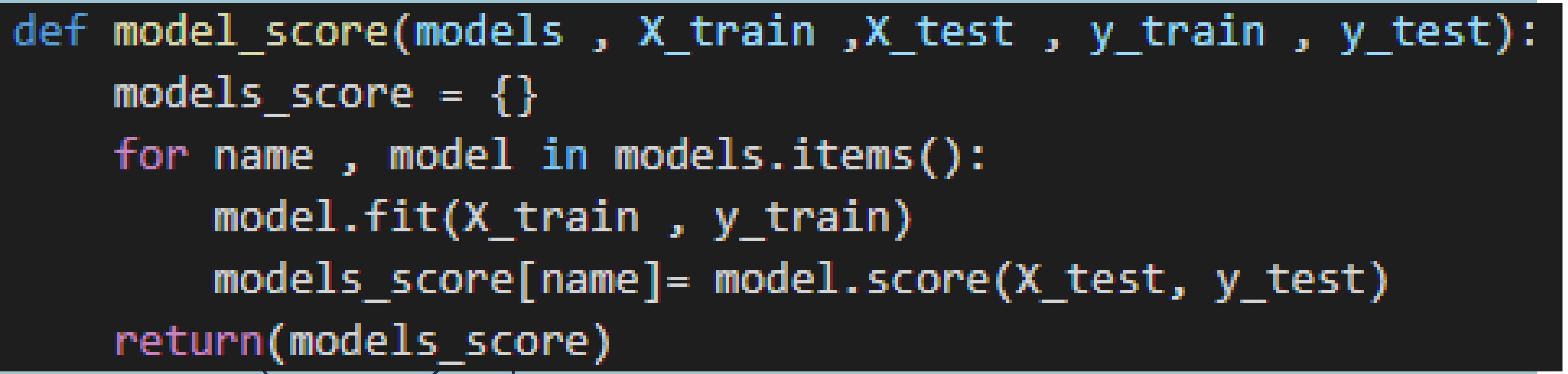
```
# categorical_features = ["cp" ,"thal" ,"ca" , "restecg" , "slope"]
cp = pd.get_dummies(data["cp"],prefix = "cp")
thal = pd.get_dummies(data["thal"],prefix = "thal")
ca = pd.get_dummies(data["ca"],prefix = "ca")
restecg = pd.get_dummies(data["restecg"],prefix = "restecg")
slope = pd.get_dummies(data["slope"],prefix = "slope")
data_1 = pd.concat([data,cp,thal,ca,restecg , slope] , axis = 1)
data_1.drop(["cp" ,"thal" ,"ca" , "restecg" , "slope"] , axis = 1 , inplace = True)
```

Predictions and ML Models

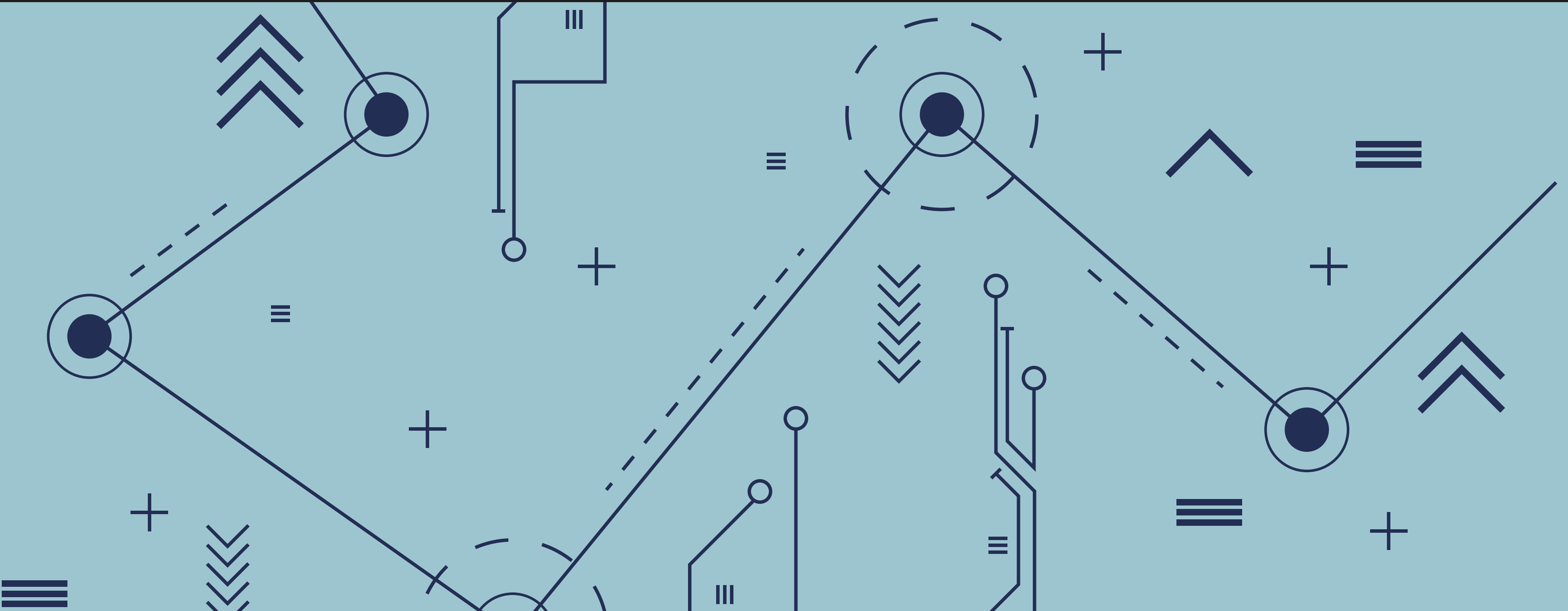
- I have trained data on 3 different machine learning models Logistic Regression , K Nearest Neighbors and Random Forest classifier.
- Out of 3 Random Forest Classifier outperformed with 98% Accuracy .

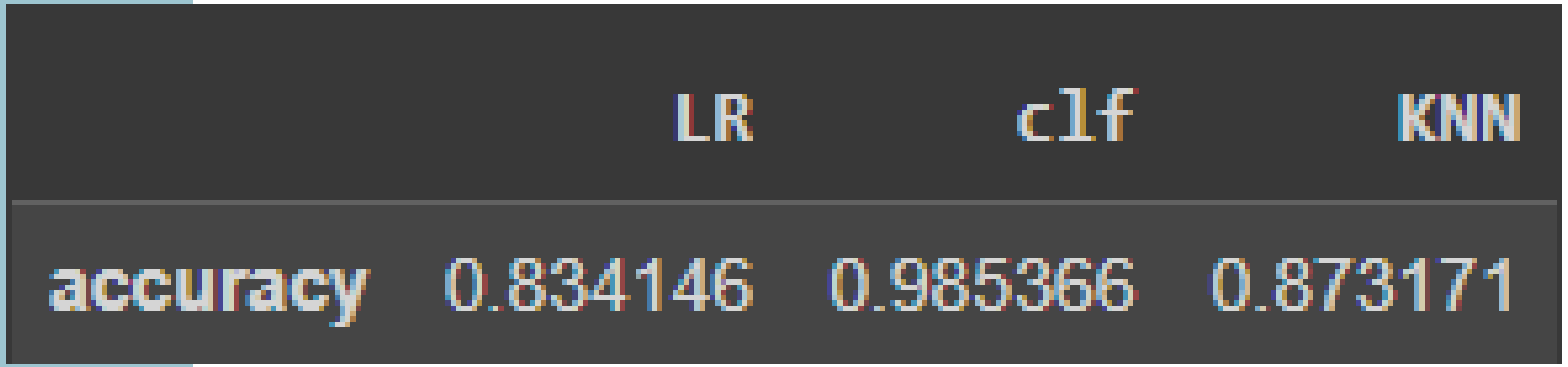


```
              precision    recall  f1-score   support

           0       0.97      1.00      0.98        97
           1       1.00      0.97      0.99       108

    accuracy                           0.99       205
   macro avg       0.98      0.99      0.99       205
weighted avg       0.99      0.99      0.99       205
```



```
def model_score(models , X_train ,X_test , y_train , y_test):
    models_score = {}
    for name , model in models.items():
        model.fit(X_train , y_train)
        models_score[name]= model.score(X_test, y_test)
    return(models_score)
```





	LR	clf	KNN
accuracy	0.834146	0.985366	0.873171