Machine Learning
Assignment 4

Eman Moustafa Ismail

25 December, 2023

❖ Task 1: Feature engineering → these are the possible options. However, I will only consider the options that improve the output of the algorithm.

➢ Check if there are any missing values → the data has no missing values.

```
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
```

➢ Getting the data summary statistics → showed that the features have a different variance scale. This suggests that maybe normalization could be beneficial.

```
              age         sex          cp     trestbps         chol         fbs  \
count  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
std      9.082101    0.466011    1.032052   17.538143   51.830751    0.356198
min     29.000000    0.000000    0.000000   94.000000  126.000000    0.000000
25%     47.500000    0.000000    0.000000  120.000000  211.000000    0.000000
50%     55.000000    1.000000    1.000000  130.000000  240.000000    0.000000
75%     61.000000    1.000000    2.000000  140.000000  274.500000    0.000000
max     77.000000    1.000000    3.000000  200.000000  564.000000    1.000000

           restecg     thalach       exang      oldpeak       slope          ca  \
count   303.000000  303.000000  303.000000  303.000000  303.000000  303.000000
mean      0.528053  149.646865    0.326733    1.039604    1.399340    0.729373
std       0.525860   22.905161    0.469794    1.161075    0.616226    1.022606
min       0.000000   71.000000    0.000000    0.000000    0.000000    0.000000
25%       0.000000  133.500000    0.000000    0.000000    1.000000    0.000000
50%       1.000000  153.000000    0.000000    0.800000    1.000000    0.000000
75%       1.000000  166.000000    1.000000    1.600000    2.000000    1.000000
max       2.000000  202.000000    1.000000    6.200000    2.000000    4.000000

              thal      target
count   303.000000  303.000000
mean      2.313531    0.544554
std       0.612277    0.498835
min       0.000000    0.000000
25%       2.000000    0.000000
50%       2.000000    1.000000
75%       3.000000    1.000000
max       3.000000    1.000000
```
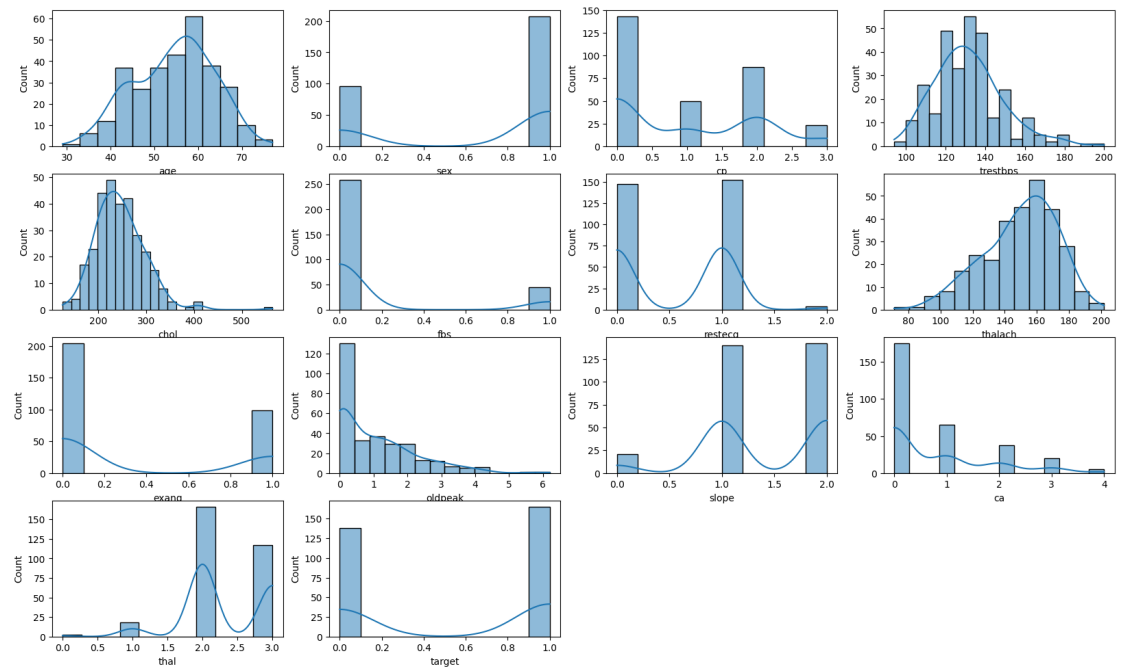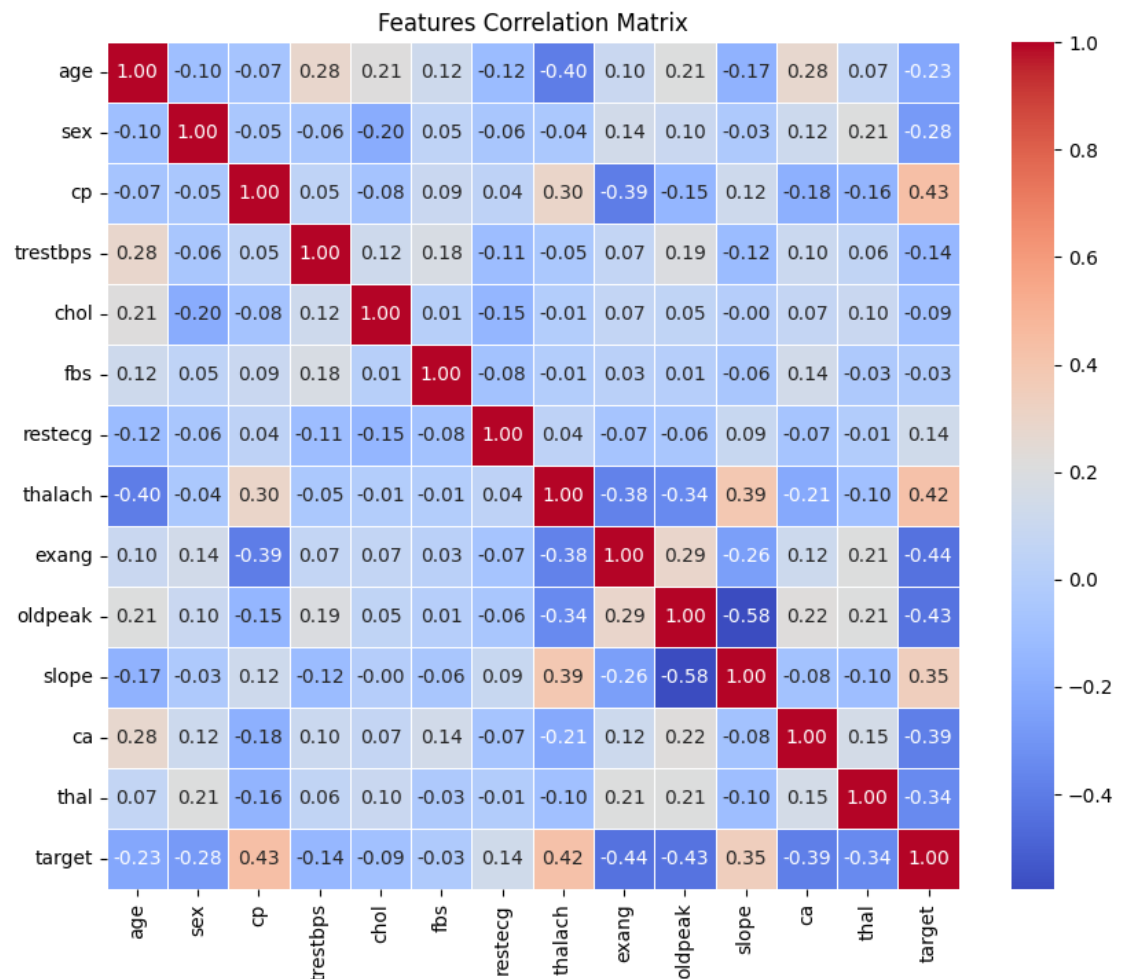
➢ Checking the skewness and feature distribution in the data.

➢ For the feature selection process. I got the feature correlation matrix → as shown, some features could be ignored.

Features Correlation Matrix

➢ Combine&tweaking features:
- One famous approach is to convert age to age groups.
- Another famous technique is categorizing → this can be used with blood pressure. Furthermore, encoding should be applied so that the logistic regression algorithm would understand the input.

❖Task 2: Applying data processing → data processing will not be applied to the ignored features (fbs, chol, restecg, and trestbps), which were ignored due to having a small correlation with the target. However, training the data with these features will be attempted as well, just to confirm the feature selection impact.