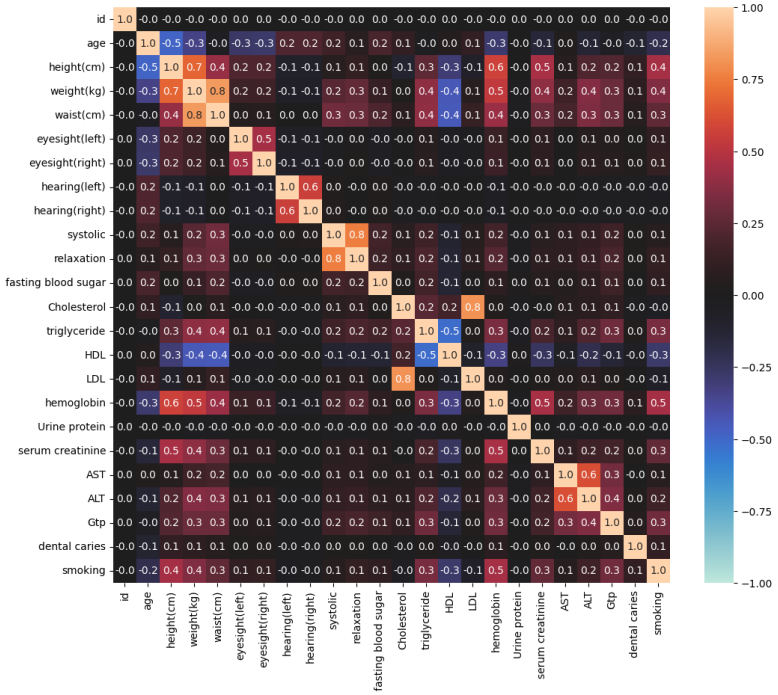
Q5. Smoke Status Recognition

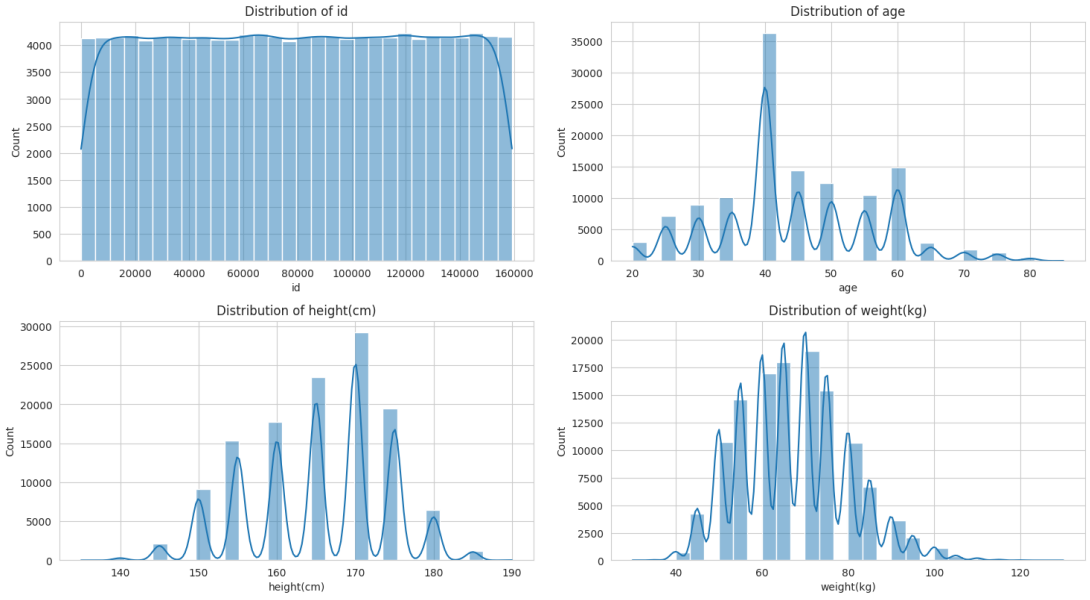
Just run the jupyter notebook of `Q5.ipynb`. It uses the given dataset. The program has one output file: * **Q5_output.csv**

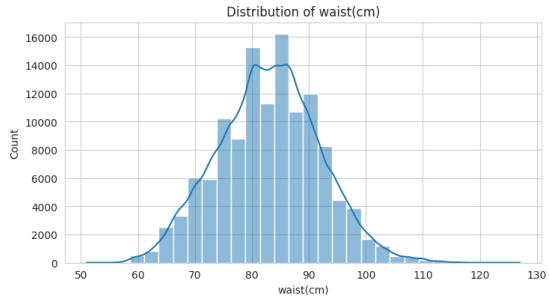
a.EDA:

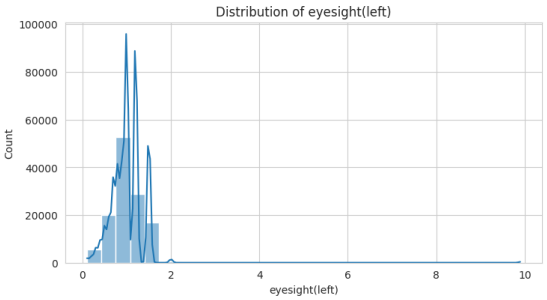
- Load the train and test data. And then view the 'discribet' of data which have the likewise distribution. So I didn't remove the outliers.
- Remove NAN 'train_df = train_df.dropna()'
- Correlation heatmap

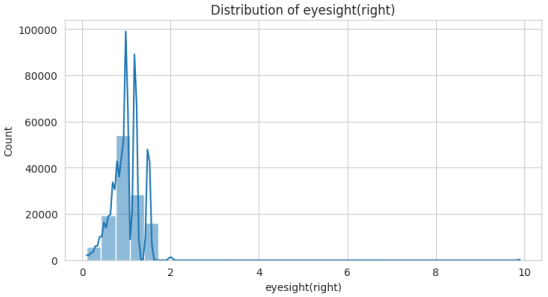


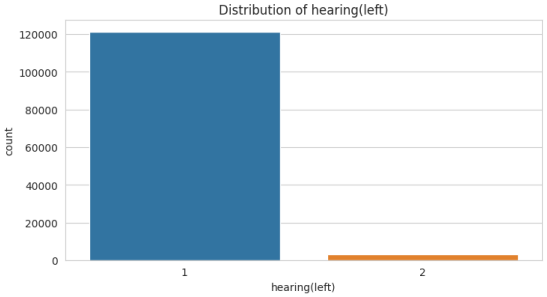
- Feature distribution

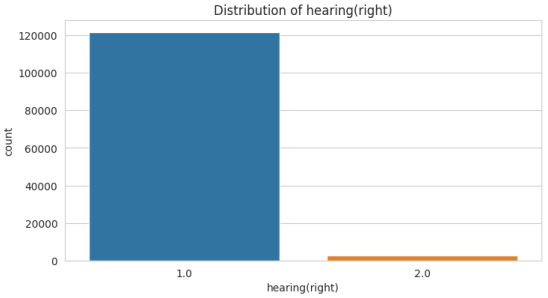


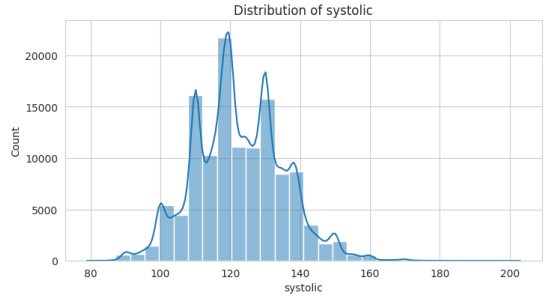

Distribution of waist(cm)
Count
waist(cm)

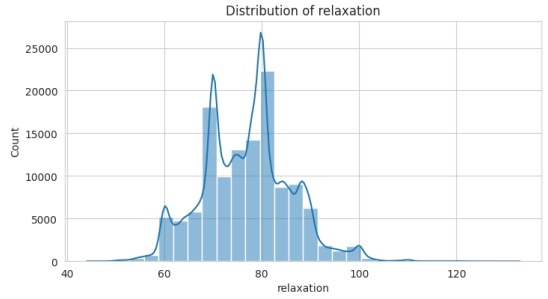

Distribution of eyesight(left)
Count
eyesight(left)

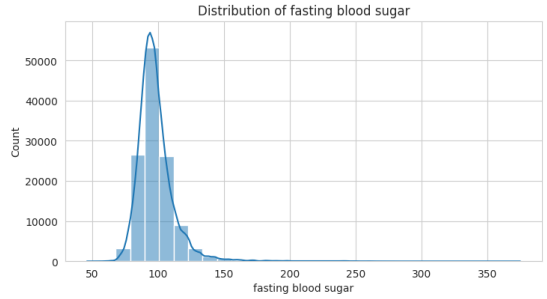

Distribution of eyesight(right)
Count
eyesight(right)

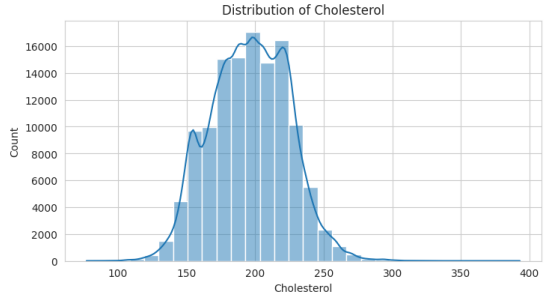

Distribution of hearing(left)
count
hearing(left)

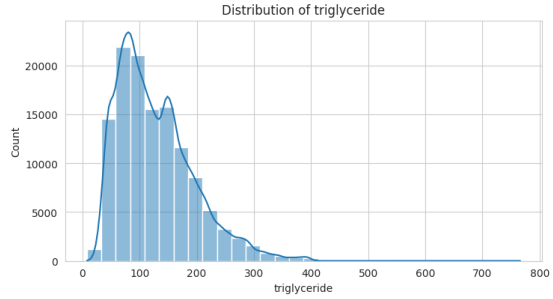

Distribution of hearing(right)
count
hearing(right)

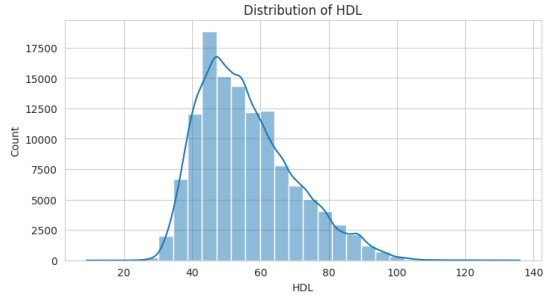

Distribution of systolic
Count
systolic

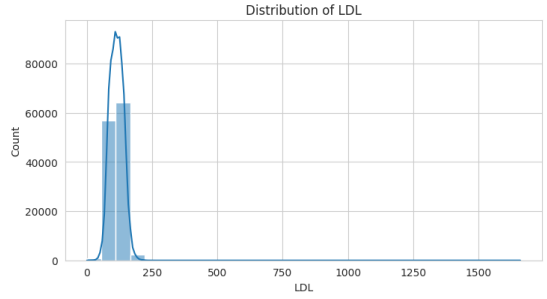

Distribution of relaxation
Count
relaxation


Distribution of fasting blood sugar
Count
fasting blood sugar

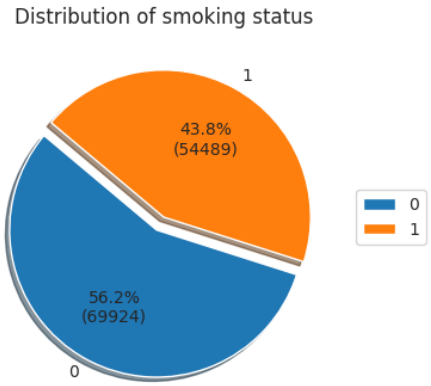

Distribution of Cholesterol
Count
Cholesterol


Distribution of triglyceride
Count
triglyceride

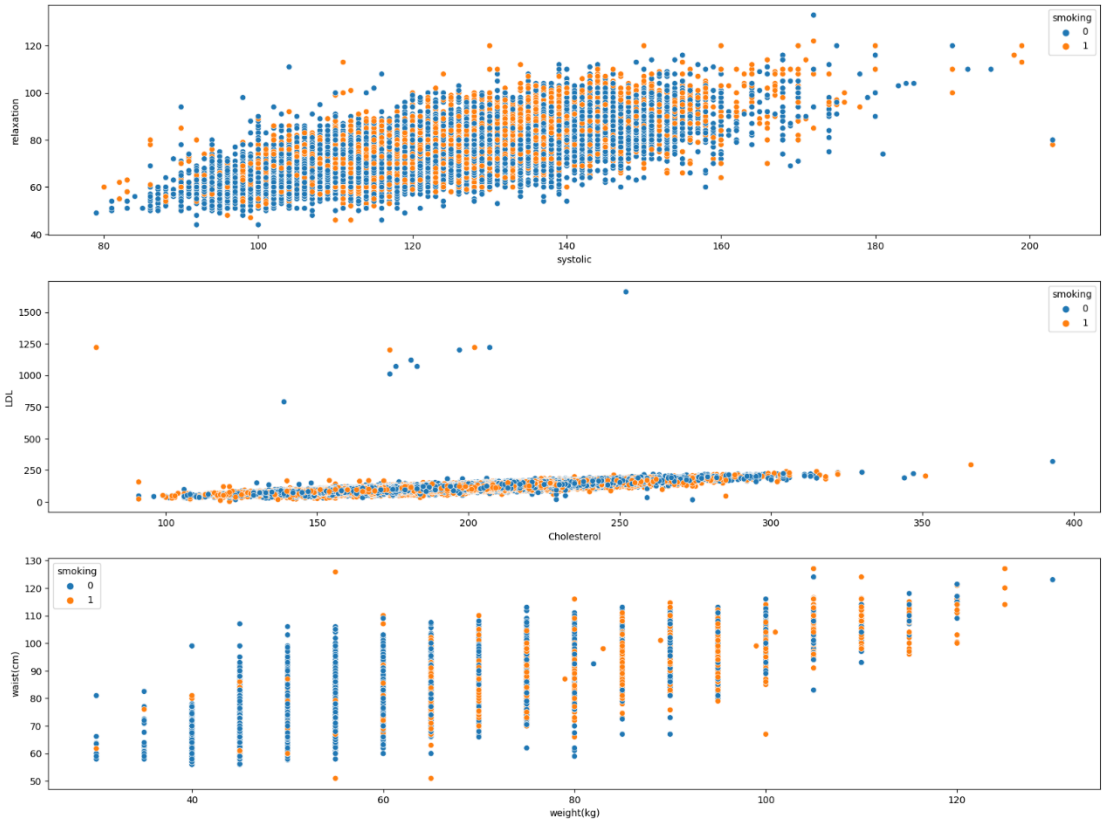

Distribution of HDL
Count
HDL


Distribution of LDL
Count
LDL

- Label distribution



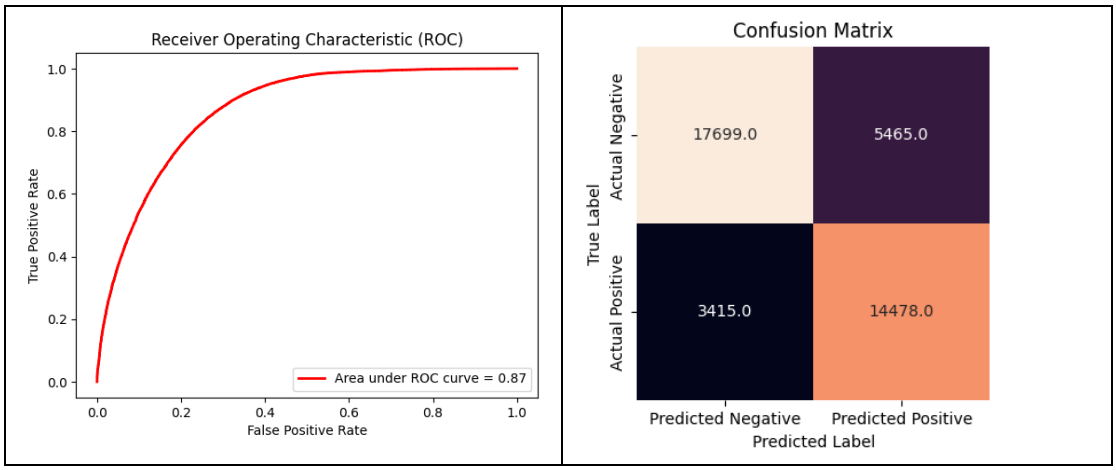
- Scatter relationship



b. choose model

- XGBoost: cross_val_score= 0.8695248062789621
- LightGBM: cross_val_score= 0.8691387378052922
- CatBoost: cross_val_score= 0.8662158517499438

So I choose **XGBoost** to train the model.
Valuate the model:



The result looks well. But I want to do some improvemnets.

c. Add Preprocessing

- Tranform the feature
- Classify the feature to [scale] and [one hot]
- Use **Robustscaler** to perform data scaling
- One Hot Encoding the categorical columns
- Tomek Links : Downsampling for getting balanced dataset

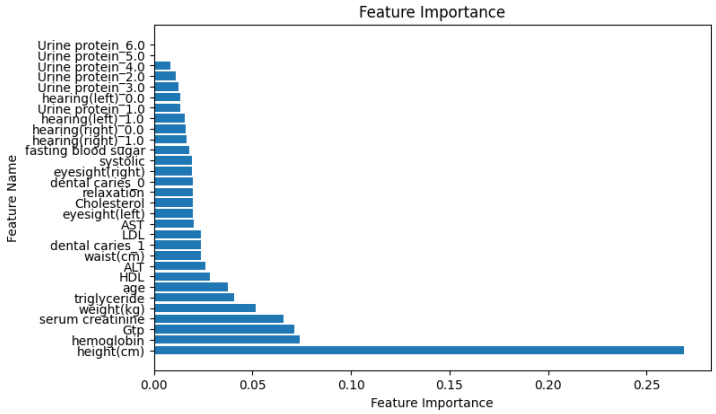
d. Model Training

- performing cross-validation using XGBoost (Extreme Gradient Boosting) for a binary classification task

```
1   print(best_auc)
```

```
0.8938614576110752
```

- feature importance



Reference：

[1]https://www.kaggle.com/code/anthonynam/smoke-status-prediction
[2]https://www.kaggle.com/code/mostafamohammednouh/smoker-status-prediction-eda
[3]https://www.kaggle.com/code/arunklenin/ps3e24-eda-feature-engineering-ensemble