**Individual Project**

Table of Contents

[1 Dataset exploration 2](#_Toc153809170)

[2 Feature Pre-processing 5](#_Toc153809171)

[3 Predict 5](#_Toc153809172)

[4 Association Rule Mining 7](#_Toc153809173)

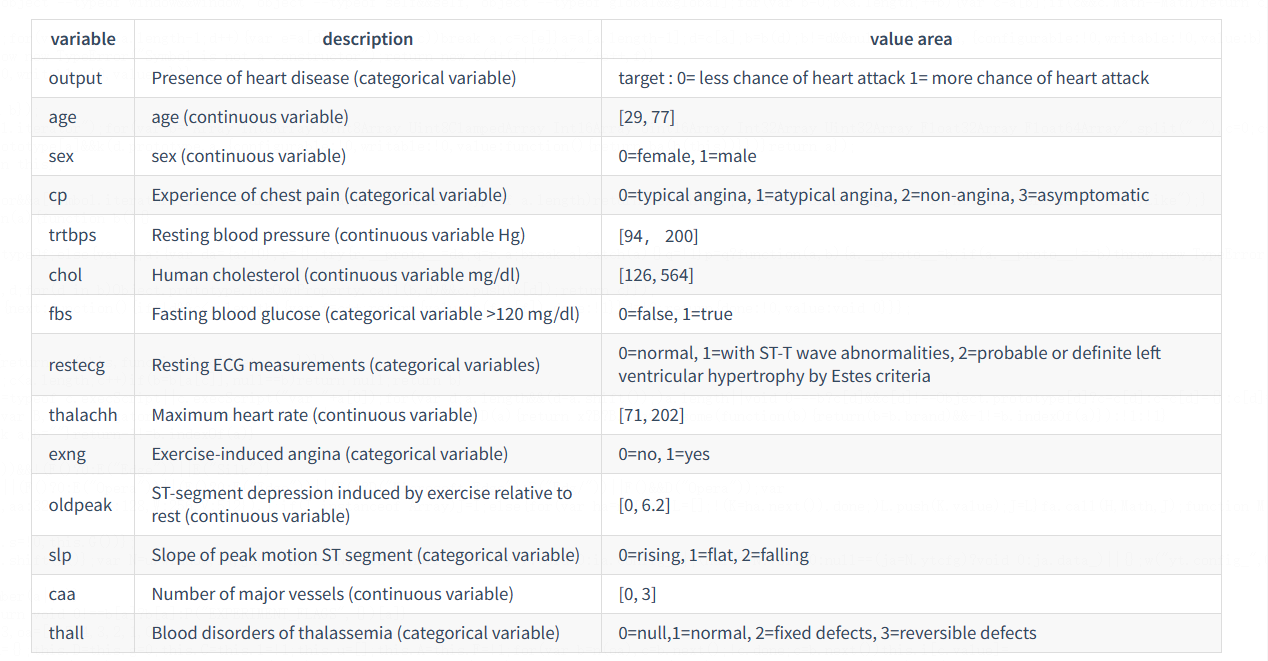
[5 Improvement 8](#_Toc153809174)

[6 Evaluation 9](#_Toc153809175)

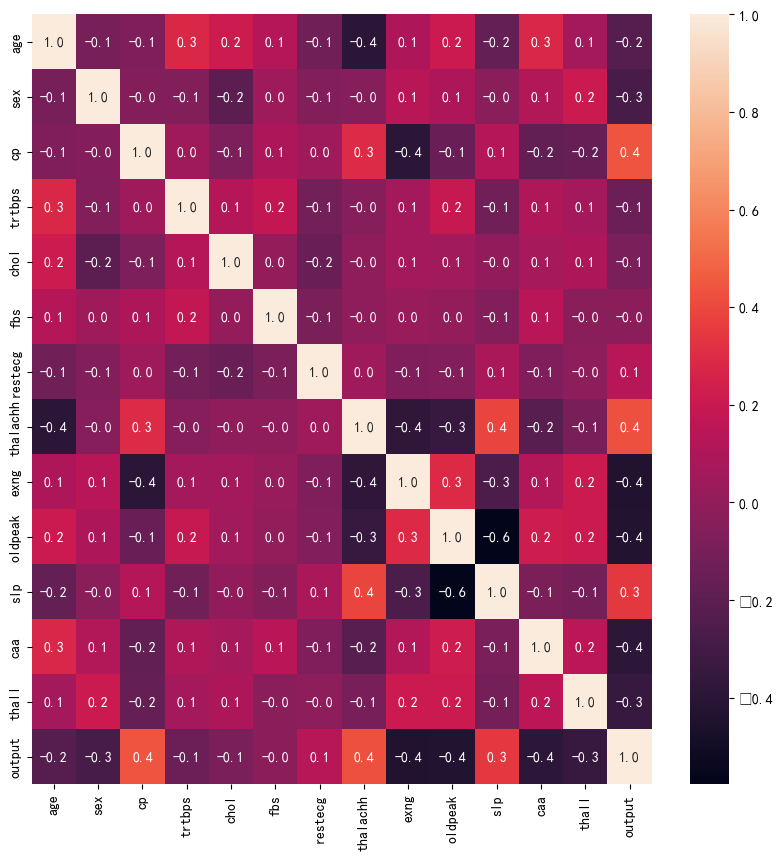
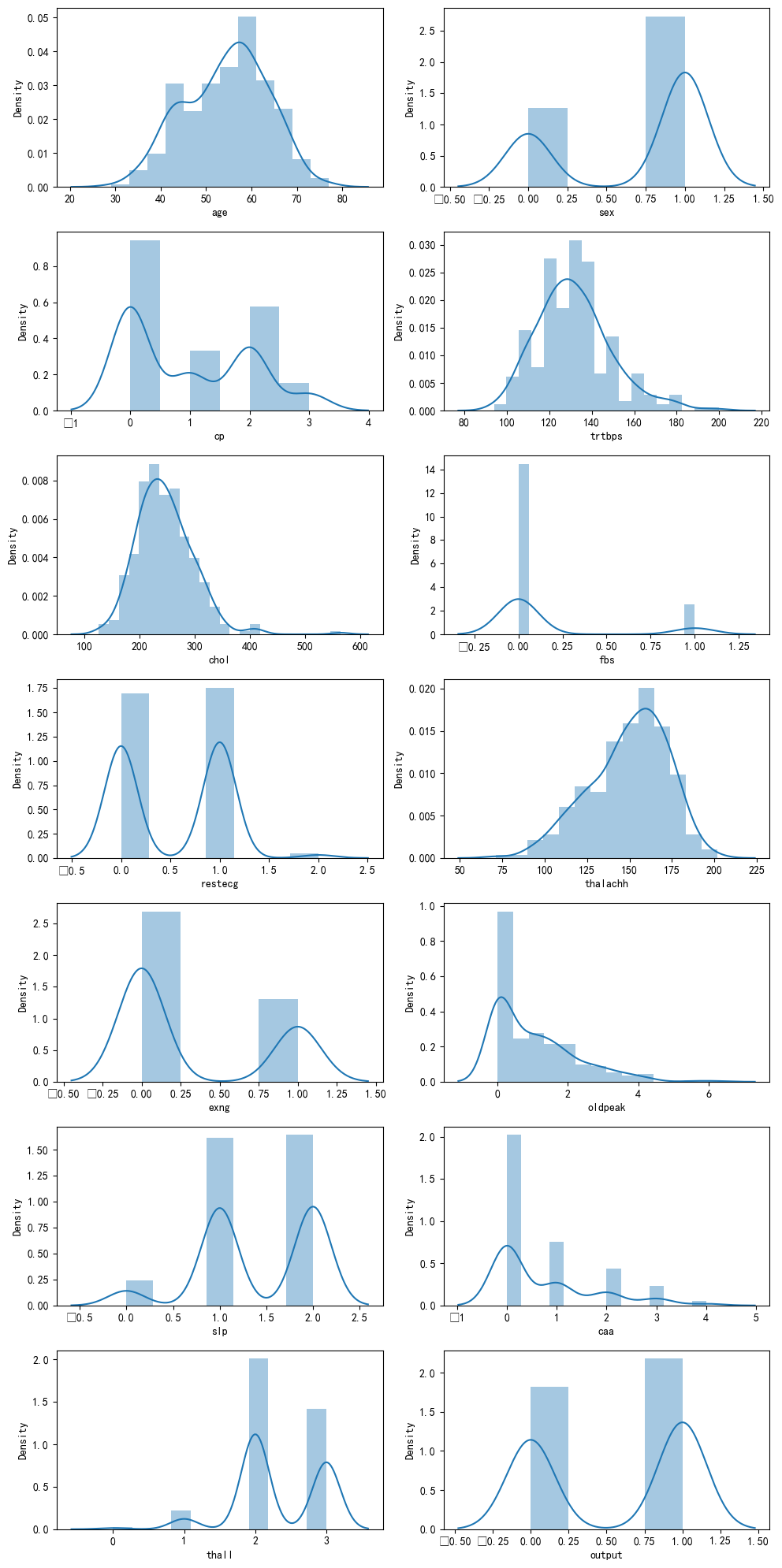
[7 References 10](#_Toc153809176)

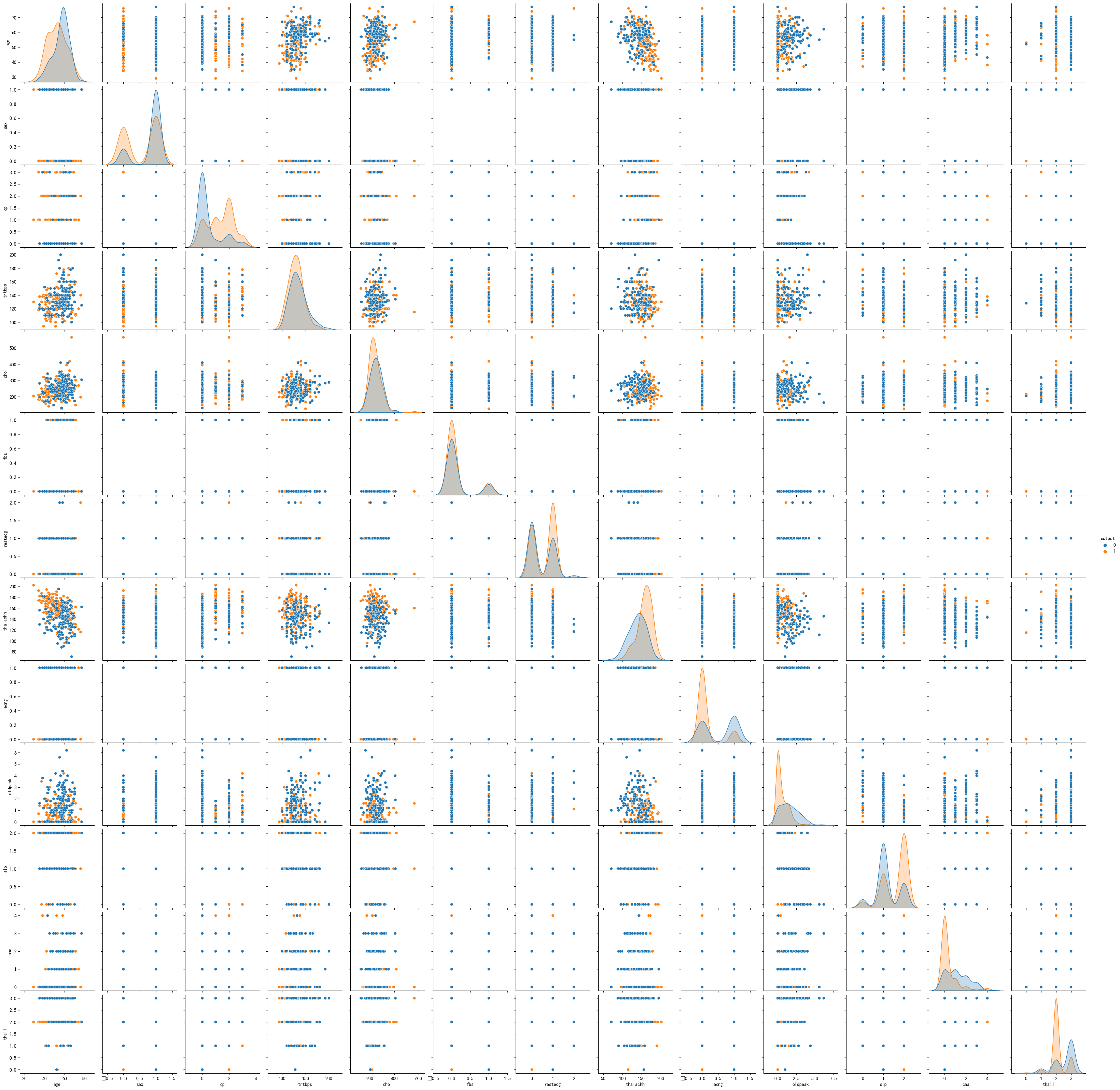
# Dataset exploration

**Data introduction**: This data comes from kaggle, contains 14 dimensions, 303 samples, the specific variables are described in the following table.



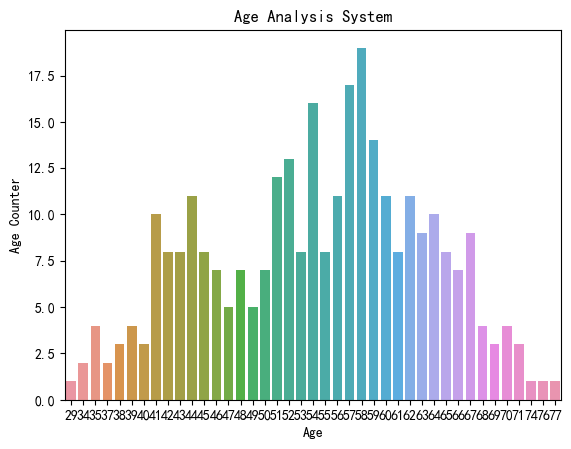
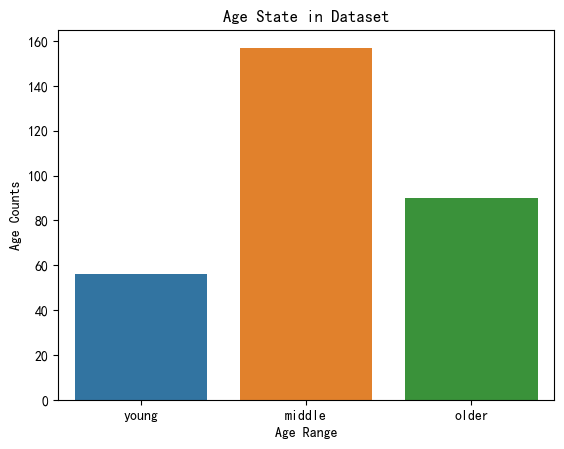
**Distribution of features, feature correlation heatmap and feature pairplot**

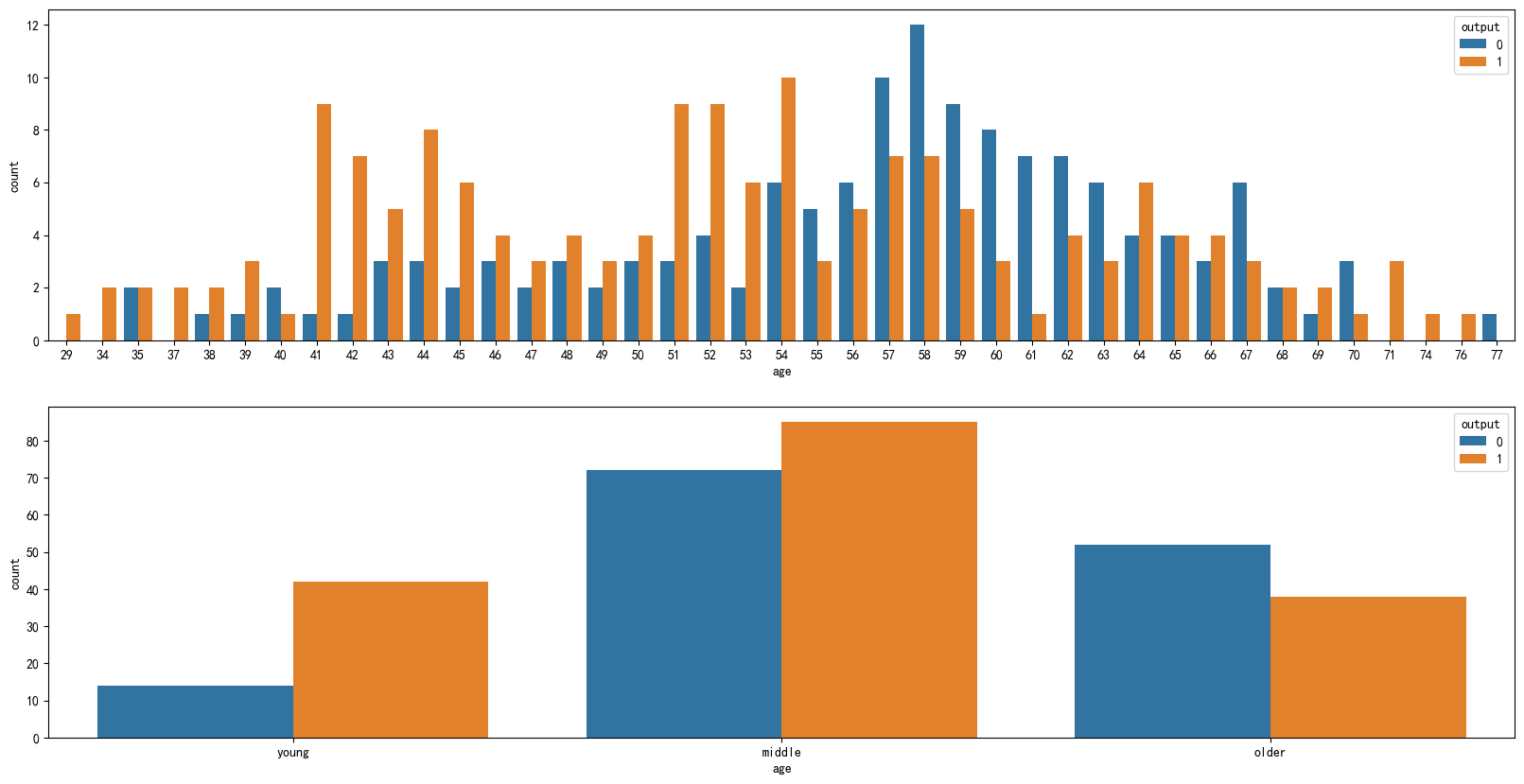




**Analysis**: In the feature distribution bar chart, we can find feature age, trtbps, chol, thalachh follow a normal distribution. In the correlation heatmap figure, we can know the feature cp, thalachh, exng, oldpeak, thall, caa, thall has highest correlation with feature output. In the pairplot figure, we can find the distributions of features age, cp, tahlachh, oldpeak, caa with respect to the two values of output are more differentiated. These features are more discriminatory for outputs.

**Exploration on age features**

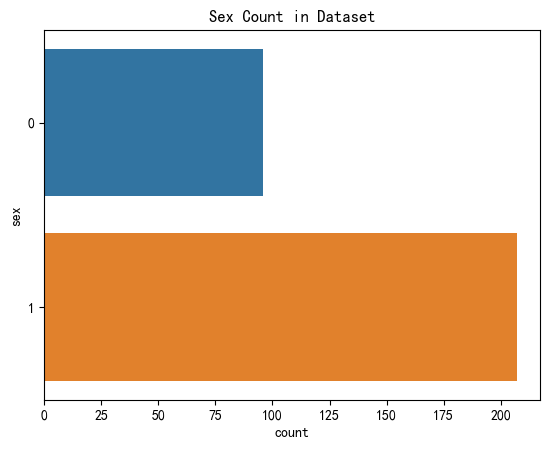
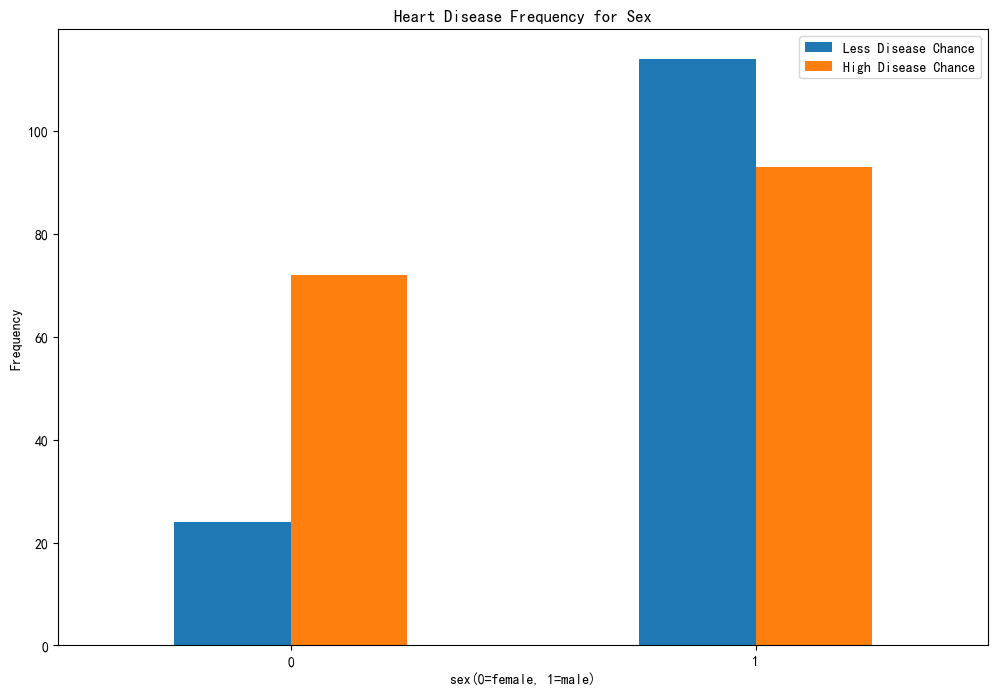
 

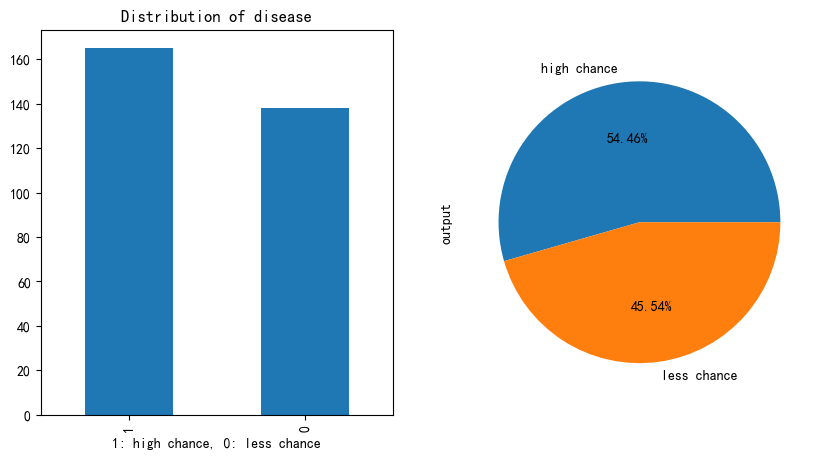


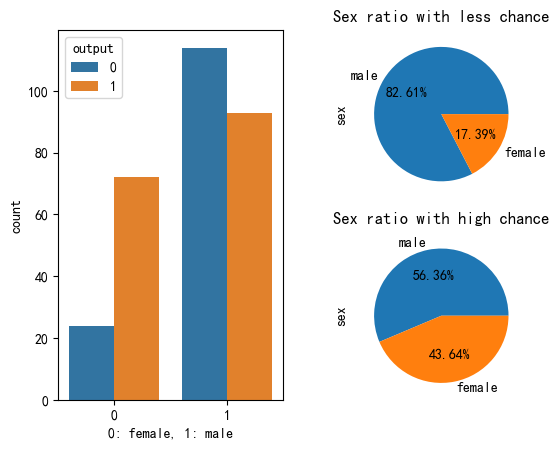
**Analysis:** The above figure shows us the age distribution of the different datasets, we can see that the number of people and the percentage of illnesses are not the same for different ages, in order to better represent the groups, we discretize the ages, [0,45) for young people, [45,60) for middle-aged people, and [60,100) for elderly people. We can also observe that the proportion of probability of disease varies in different groups.

The young people and middle people have more chance to have a heart attack than the old people.

**Exploration on sex features**





**Analysis**: According the above figure, we can find the there is no gender balance in the given data set and also can discover the male ratio is bigger than female under the condition of less chance of heart disease but roughly eauql under the condition of high chance of heart disease

# Feature processing

Step 1:Convert

Converts discrete data, from ordinary 0,1,2, to real strings. The rules are as follow.

Age: 0-45 young, 45-60 middle, 60-100 older

Sex: 0 female, 1male

Cp: 0 typical, 1 atypical, 2 non-anginal, 3 asymptomatic

Fbs: 0 false, 1 true

Exng: 0 no,1 yes

Slp: 0 rising, 1flat, 2 falling

Thall: 0 null, 1 normal, 2 fixed, 3 reversed

Restecg: 0 normal, 1 ST-T-abnormal, 2 Left-ventricular-hypertrophy

Caa: 0 zero, 1 one, 2 two, 3 three

Step 2:

One-hot encoding

Step 3:

z-score normalization

# Predict

Clustering-based Approaches:

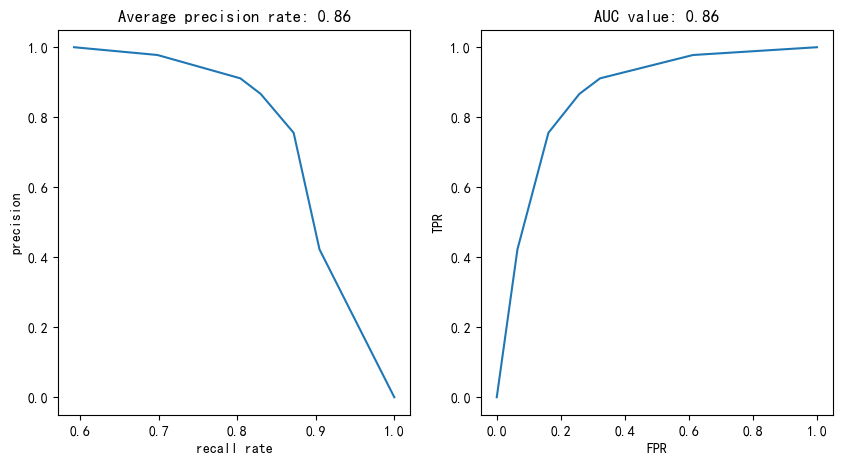
KNN:

accuracy: 0.8481420765027323

precision: 0.8297872340425532

recall: 0.8666666666666667

F1 score: 0.8478260869565217



**Classification-based**

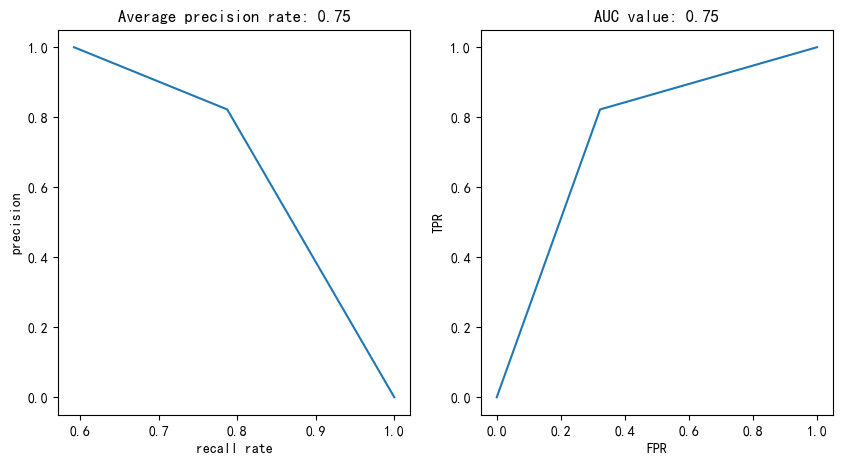
Decision Tree(Only set max\_depth=10):

accuracy: 0.7189617486338797

precision: 0.7872340425531915

recall: 0.8222222222222222

F1 score: 0.8043478260869565



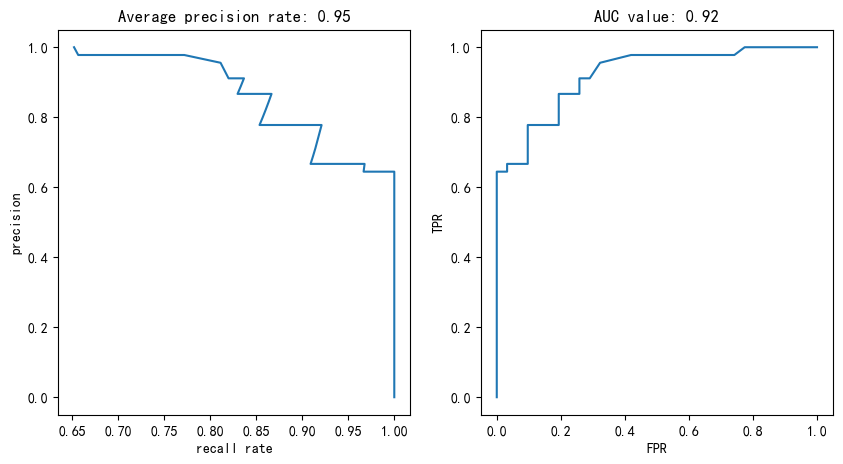
Random forest(only set n\_estimators=100 ):

accuracy: 0.8283060109289616

precision: 0.8636363636363636

recall: 0.8444444444444444

F1 score: 0.853932584269663



**Analysis:** From the above results, we can find that among the classification methods, the knn performance of the clustering method is better than the decision tree, but not as good as the random deep forest of the integration method. In the above results, we can also find that knn has the smoothest curve.

# Association Rule Mining

**Further discretization:**

In order to use association rule mining, we need to discretize all data. So add the follow converting rules. These rule can be found in reference.

Trtbps: [90,139] normal, and [139,200] high

Chol: <200 mg/dl low risk, [200,240] borderline high risk, >240 high risk

Tthalachh [150,200] normal, others are abnormal

Oldpeak: normal <=1, >1 abnormal

**Normalization**

Then use z-score normalization

**ARM:**

Use mlxtend package to do the apriori and association\_rules. Set the min\_support=0.2 and min\_threshold=0.8.

**Result**:

There are so many rules so I only list some of them. More results can be found in folder csv file.

Less chance of heart attack

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Antecedents | consequents | support | confidence | lift |
| frozenset({'exng\_yes', 'sex\_male'}) | frozenset({'output\_less'}) | 0.20462 | 0.805195 | 1.767928 |
| frozenset({'cp\_typical', 'thalachh\_abnormal'}) | frozenset({'output\_less'}) | 0.244224 | 0.822222 | 1.805314 |
| frozenset({'exng\_yes', 'cp\_typical'}) | frozenset({'output\_less'}) | 0.231023 | 0.875 | 1.921196 |

High chance of heart attack

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| antecedents | consequents | support | confidence | lift |
| frozenset({'age\_middle','thall\_fixed'}) | frozenset({'output\_more'}) | 0.211221 | 0.8 | 1.469091 |
| frozenset({'exng\_no','sex\_female'}) | frozenset({'output\_more'}) | 0.211221 | 0.864865 | 1.588206 |
| frozenset({'thall\_fixed','sex\_female'}) | frozenset({'output\_more'}) | 0.227723 | 0.873418 | 1.603913 |

**Analysis:**

In this part, I only choose two rule to do analysis.

For (exng\_yes,sex\_male->output\_less), It means a people who is male and have exercise-induced angina have less chance of less chance of heart attack that this rule has 20.5% support, 80.5% confident and 1.76 lift (>1 Positive correlation exists).

For (exng\_no,sex\_female->output\_more), It means a people who is female and doesn’t have exercise-induced angina have less chance of less chance of heart attack that this rule has 21.1% support, 86.5% confident and 1.58 lift (>1 Positive correlation exists).

# Improvement

I only made minor modifications i.e. categorizing on the ARM dataset. Among them, the performance of knn and decison tree has some improvement, and knn effect enhancement is more obvious. and random forest has some decrease. The results are shown below

**Clustering-based Approaches:**

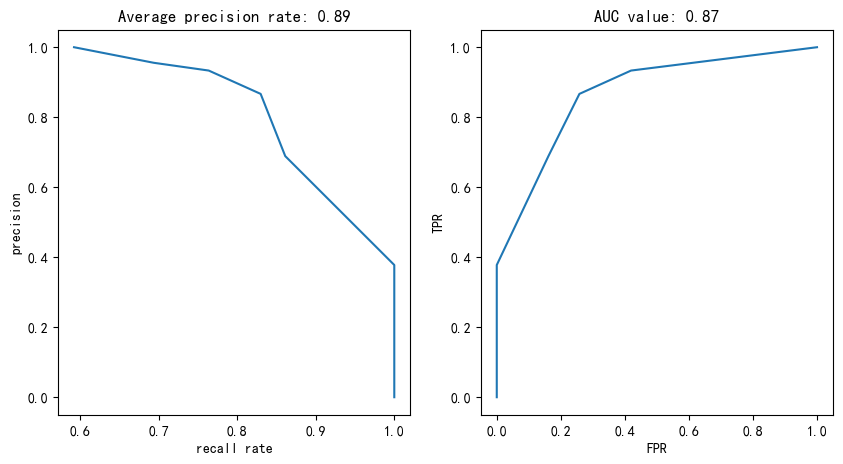
KNN:

accuracy: 0.8349726775956284

precision: 0.8297872340425532

recall: 0.8666666666666667

F1 score: 0.8478260869565217



**Classification-based**

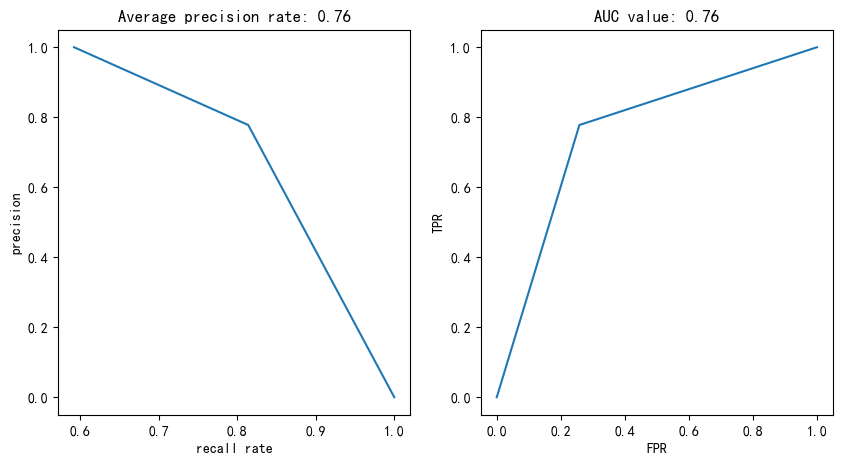
Decision Tree(Only set max\_depth=10):

accuracy: 0.7556284153005464

precision: 0.813953488372093

recall: 0.7777777777777778

F1 score: 0.7954545454545455



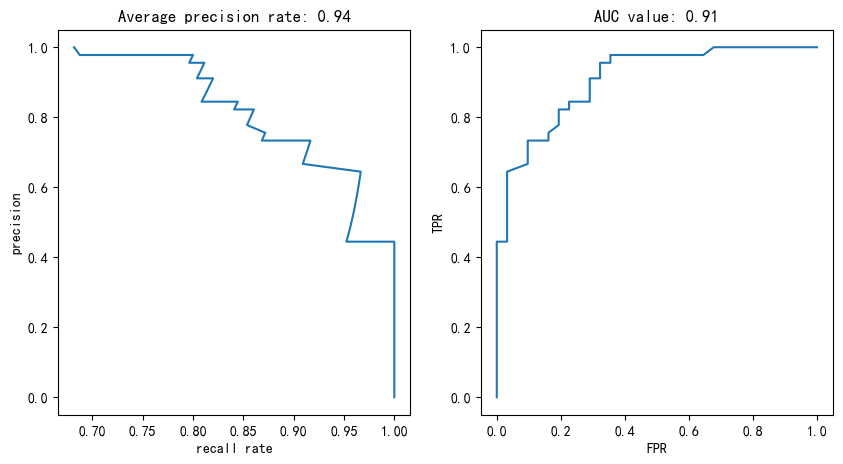
Random forest(only set n\_estimators=100 ):

accuracy: 0.8281967213114754

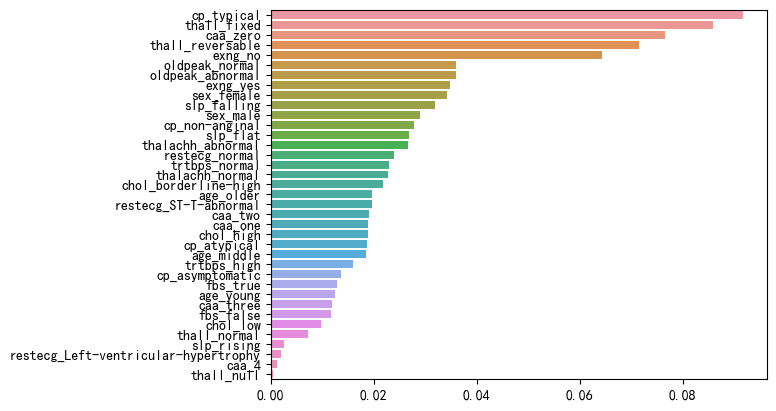
precision: 0.8444444444444444

recall: 0.8444444444444444

F1 score: 0.8444444444444444



Heart Disease Prediction-Characteristic Importance Analysis based on random forest



Analysis: Through this figure we can find that the three features that have the greatest importance in predicting the outcome of the random forest are cp, thall, caa. This also validates the results of feature distribution and feature correlation.

# Evaluation

Throughout this project, I have successfully explored the power of data mining in solving a practical problem and demonstrated my understanding of various data mining techniques. I learned and developed an effective data mining solution by utilizing seaborn (sns) library to create distribution plots, heatmaps, and pairplots for data exploration.

In terms of prediction, I implemented three different algorithms - K-Nearest Neighbors (KNN), Decision Tree, and Random Forest - to predict the target variable. By employing these algorithms, I was able to make accurate predictions based on the given dataset.

Furthermore, I utilized association data mining to discover and extract rules between the features and the target variable. This helped me identify significant associations and patterns in the data, providing valuable insights into the relationships between different variables.

To further enhance the predictive power of the models, I applied discretization techniques based on association data mining. By discretizing the data, I was able to improve the performance and accuracy of the prediction methods.

Overall, this project allowed me to gain hands-on experience in data mining techniques. I learned how to effectively explore and visualize data using sns distribution plots, heatmaps, and pairplots. I successfully implemented KNN, Decision Tree, and Random Forest algorithms for prediction, and leveraged association data mining to discover valuable rules between features and the target variable. Lastly, I applied discretization techniques to further enhance the predictive performance of the models.

This project has not only deepened my understanding of data mining but also provided me with practical experience in solving real-world problems through data analysis and prediction.

# References

[Heart Attack Analysis & Prediction Dataset (kaggle.com)](https://www.kaggle.com/datasets/rashikrahmanpritom/heart-attack-analysis-prediction-dataset/discussion/329925)

[Blood pressure - Wikipedia](https://en.wikipedia.org/wiki/Blood_pressure)

[Cholesterol - Wikipedia](https://en.wikipedia.org/wiki/Cholesterol)

[Normal Heart Rate By Age: Range, Charts And More – Forbes Health](https://www.forbes.com/health/healthy-aging/normal-heart-rate-by-age/)

[Diagnostic and prognostic value of ST segment depression limited to the recovery phase of exercise stress test - PMC (nih.gov)](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1768611/)

[Python可视化 | Seaborn5分钟入门(七)——pairplot - 知乎 (zhihu.com)](https://zhuanlan.zhihu.com/p/98729226)

[基于Kaggle心脏病数据集的数据分析和分类预测-StatisticalLearning统计学习实验报告\_kaggle医学数据分析-CSDN博客](https://blog.csdn.net/qq_40605313/article/details/120061220)