



DS

boosting your data
exploration

HEART DISEASE PREDICTION

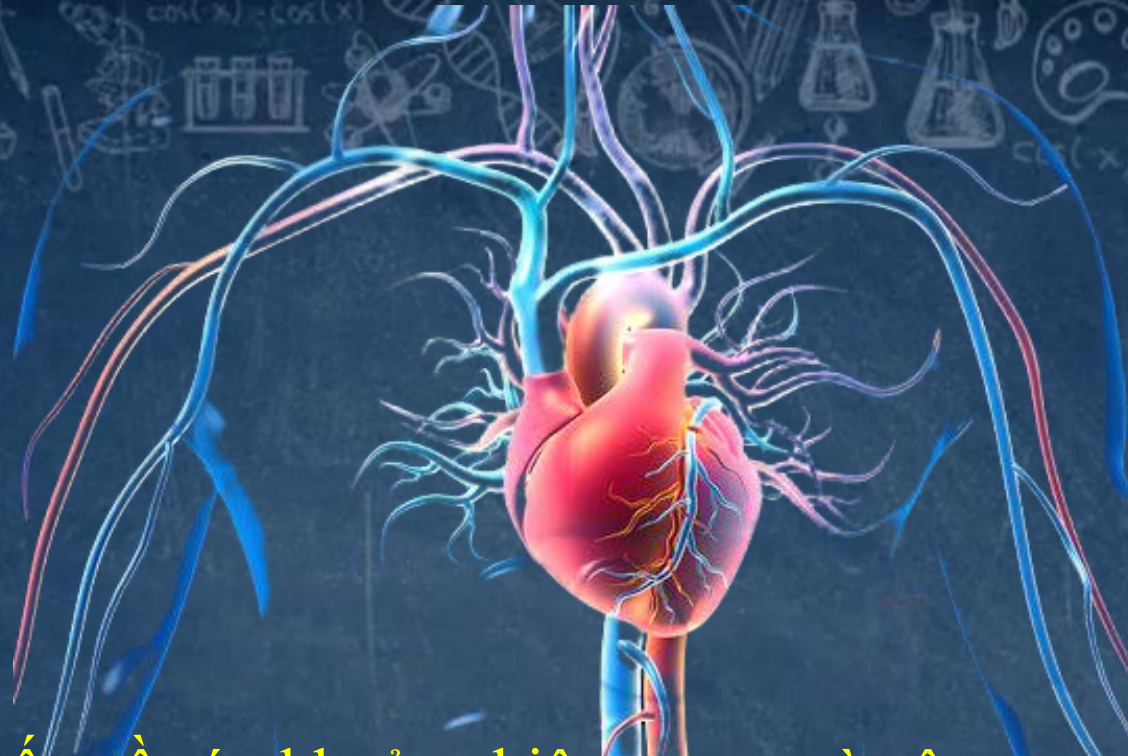


THỰC HIỆN: Nguyễn Duy Đạt

GIẢNG VIÊN: Phạm Đình Khánh

KHÓA HỌC: Data Analytics and Math for Beginner

Chủ nhật, 05/09/2021



- Đau tim là một vấn đề sức khỏe nghiêm trọng và gây nguy hiểm tới tính mạng.
- Chuẩn đoán sớm đau tim sẽ giúp điều trị bệnh tốt hơn.

Trong nhiệm vụ này chúng ta sẽ sử dụng các thông tin đầu vào là các thông tin người bệnh, chỉ số y tế và triệu chứng:

Tuổi

Giới tính

Kiểu đau ngực

Huyết áp

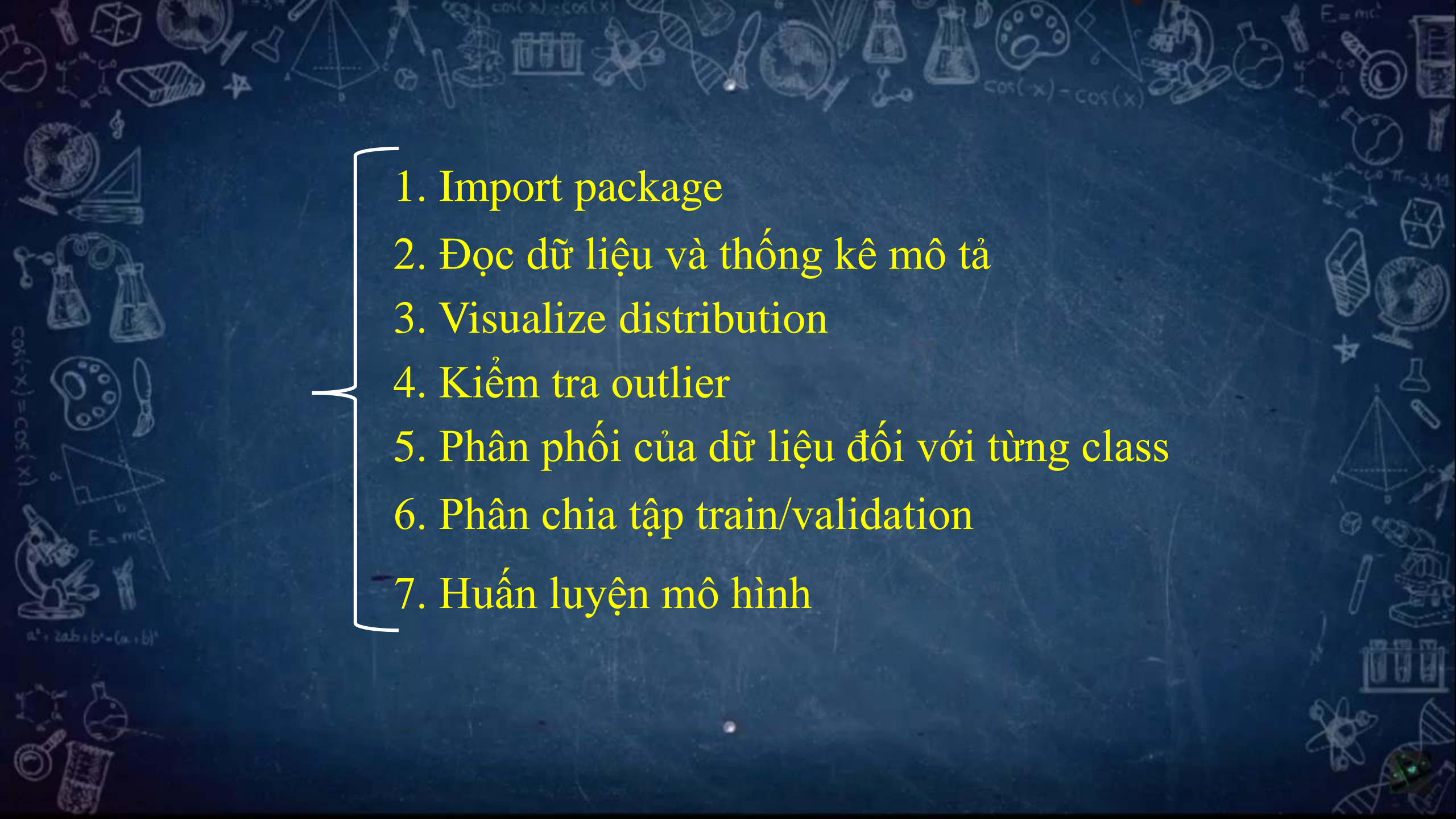
Lượng cholesterol

Đường huyết khi đói

Kq điện tâm đồ

Nhịp tim

Trầm cảm
Đau ngực do
tập TD

- 
1. Import package
 2. Đọc dữ liệu và thống kê mô tả
 3. Visualize distribution
 4. Kiểm tra outlier
 5. Phân phối của dữ liệu đối với từng class
 6. Phân chia tập train/validation
 7. Huấn luyện mô hình

1. Import package

Các thư viện - package sử dụng



NumPy



Matplotlib



Pandas



Seaborn



2. Đọc dữ liệu và thống kê mô tả

- Đây là bộ dữ liệu về thông tin sức khỏe của 600,000 người(train) khác nhau.
 - + Bộ dữ liệu gồm 14 biến đầu vào và 1 biến mục tiêu.
 - + Trong đó có 6 biến numeric và 7 biến category, 1 biến mục tiêu là class.
- Nhiệm vụ: Phân loại trên class ở tập test với 400,000 người

	age	sex	chest	resting_blood_pressure	serum_cholesterol	fasting_blood_sugar	resting_electrocardiographic_results	maximum_heart_rate_achieved
ID								
0	49.207124	0	4.000000	162.996167	181.108682	0	0	148.227858
1	53.628425	1	1.741596	130.233730	276.474630	0	2	152.917139
2	49.591426	1	4.000000	146.999012	223.300517	1	2	102.352090
3	58.991445	1	4.000000	112.369143	187.245501	0	0	158.164750
4	51.053602	1	1.954609	138.032047	238.482868	0	0	172.540828
exercise_induced_angina	oldpeak	slope	number_of_major_vessels	thal	class			
	1	0.944547	2	0	3	1		
	0	0.119070	2	0	3	0		
	1	1.616747	2	2	7	1		
	1	0.000000	1	1	7	1		
	0	1.150464	1	1	3	0		

2. Đọc dữ liệu và thống kê mô tả

1. **ID:** Số ID , duy nhất với mỗi người (Ko sử dụng để huấn luyện mô hình)
2. **AGE:** tuổi(numeric)
3. **SEX:** giới tính(cate) (1 = male, 0 = female)
4. **CHEST:** mức độ đau ngực(numeric):chest pain type (1 = typical angina; 2 = atypical angina; 3 = non-anginal pain; 4 = asymptomatic)
5. **RESTING_BLOOD_PRESSURE:** huyết áp(numeric):resting blood pressure (in mm Hg on admission to the hospital)
6. **SERUM_CHOLESTORAL(numeric):**chol - serum cholestoral in mg/dl
7. **FASTING_BLOOD_SUGAR(cate):**fasting blood sugar > 120 mg/dl (1 = true; 0 = false)
8. **RESTING ELECTROCARDIOGRAPHIC RESULTS:** resting electrocardiographic results (0 = normal; 1 = having ST-T; 2 = hypertrophy)(cate)
9. **MAXIMUM HEART RATE ACHIEVED(numeric)**
10. **EXERCISE INDUCED ANGINA:** đau thắt ngực do hoạt động thể dục(cate)
11. **OLDPEAK:** ST depression induced by exercise relative to rest(numeric)
12. **THE SLOPE OF THE PEAK EXERCISE ST SEGMENT:** (1 = upsloping; 2 = flat; 3 = downsloping)(cate)
13. **NUMBER OF MAJOR VESSELS (0-3)** colored by flourosopy
14. **THAL:** thal - 3 = normal; 6 = fixed defect; 7 = reversable defect(cate)
15. **CLASS:** biến mục tiêu phân loại (0 = no, 1 = yes)

2. Đọc dữ liệu và thống kê mô tả

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 600000 entries, 0 to 599999
Data columns (total 14 columns):
#   Column                                Non-Null Count  Dtype
---  -
0   age                                   600000 non-null  float64
1   sex                                   600000 non-null  int64
2   chest                                600000 non-null  float64
3   resting_blood_pressure               600000 non-null  float64
4   serum_cholesterol                    600000 non-null  float64
5   fasting_blood_sugar                  600000 non-null  int64
6   resting_electrocardiographic_results 600000 non-null  int64
7   maximum_heart_rate_achieved          600000 non-null  float64
8   exercise_induced_angina              600000 non-null  int64
9   oldpeak                              600000 non-null  float64
10  slope                                600000 non-null  int64
11  number_of_major_vessels               600000 non-null  int64
12  thal                                  600000 non-null  int64
13  class                                 600000 non-null  int64
dtypes: float64(6), int64(8)
memory usage: 68.7 MB
```

=>Bộ dữ liệu không có missing value

2. Đọc dữ liệu và thống kê mô tả

-Làm tròn các cột float

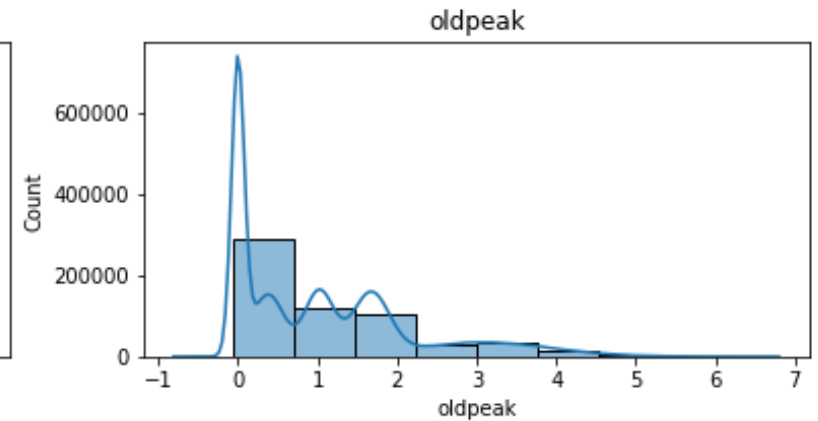
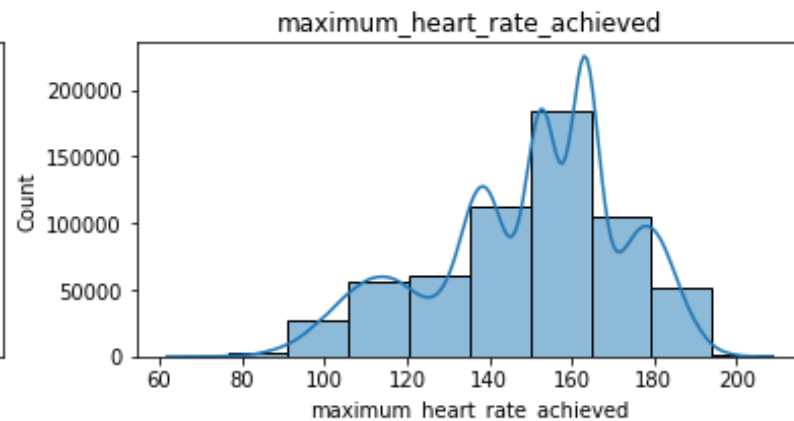
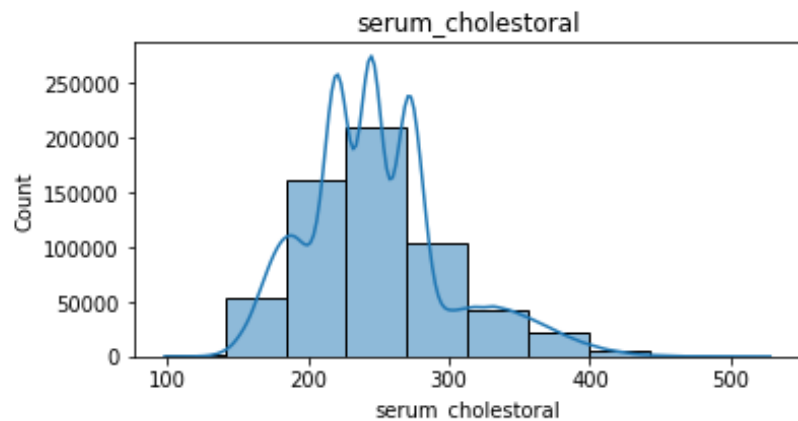
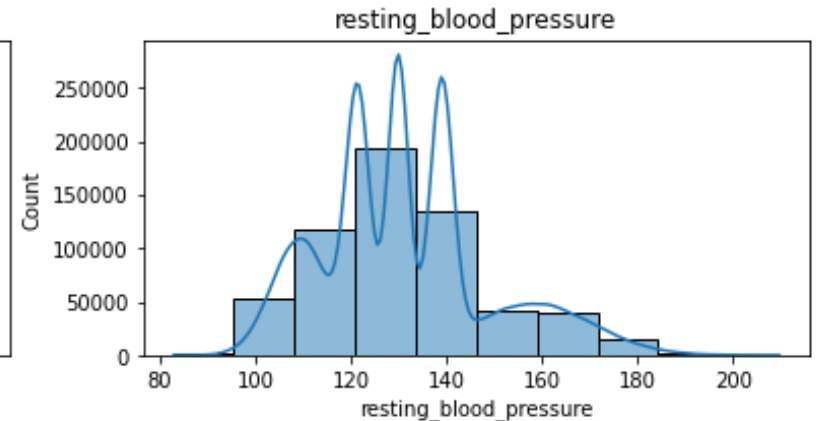
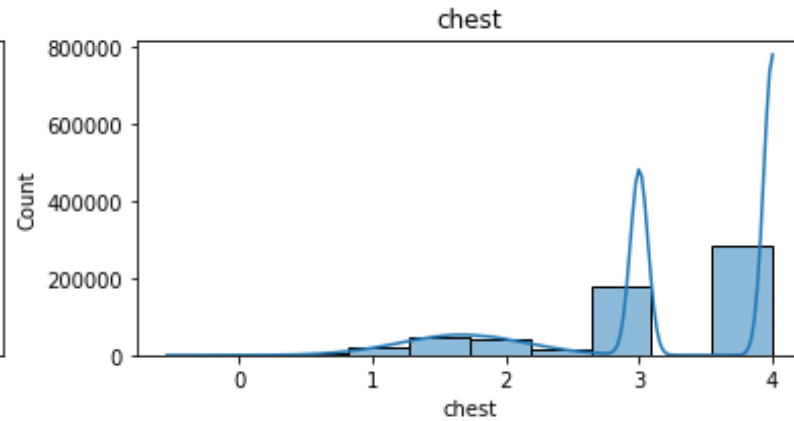
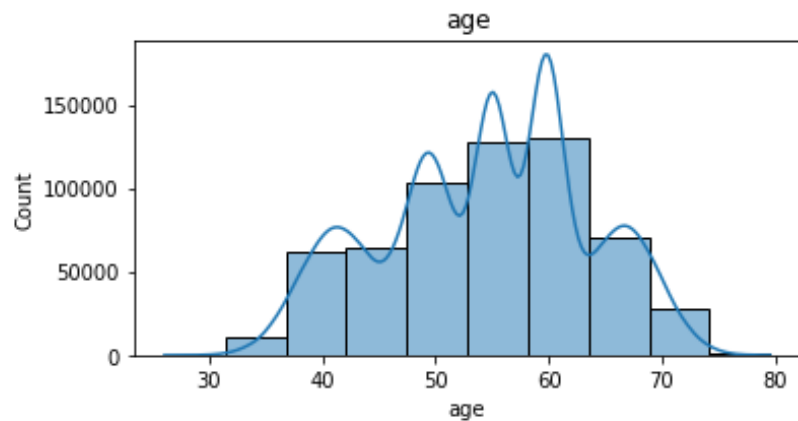
	age	sex	chest	resting_blood_pressure	serum_cholesterol	fasting_blood_sugar	resting_electrocardiographic_results	maximum_heart_rate_achieved	e
ID									
0	49.21	0	4.00	163.00	181.11	0	0	148.23	
1	53.63	1	1.74	130.23	276.47	0	2	152.92	
2	49.59	1	4.00	147.00	223.30	1	2	102.35	
3	58.99	1	4.00	112.37	187.25	0	0	158.16	
4	51.05	1	1.95	138.03	238.48	0	0	172.54	

exercise_induced_angina	oldpeak	slope	number_of_major_vessels	thal	class
1	0.94	2	0	3	1
0	0.12	2	0	3	0
1	1.62	2	2	7	1
1	0.00	1	1	7	1
0	1.15	1	1	3	0

3. Visualize distribution

Biến liên tục

number of numeric field: 6

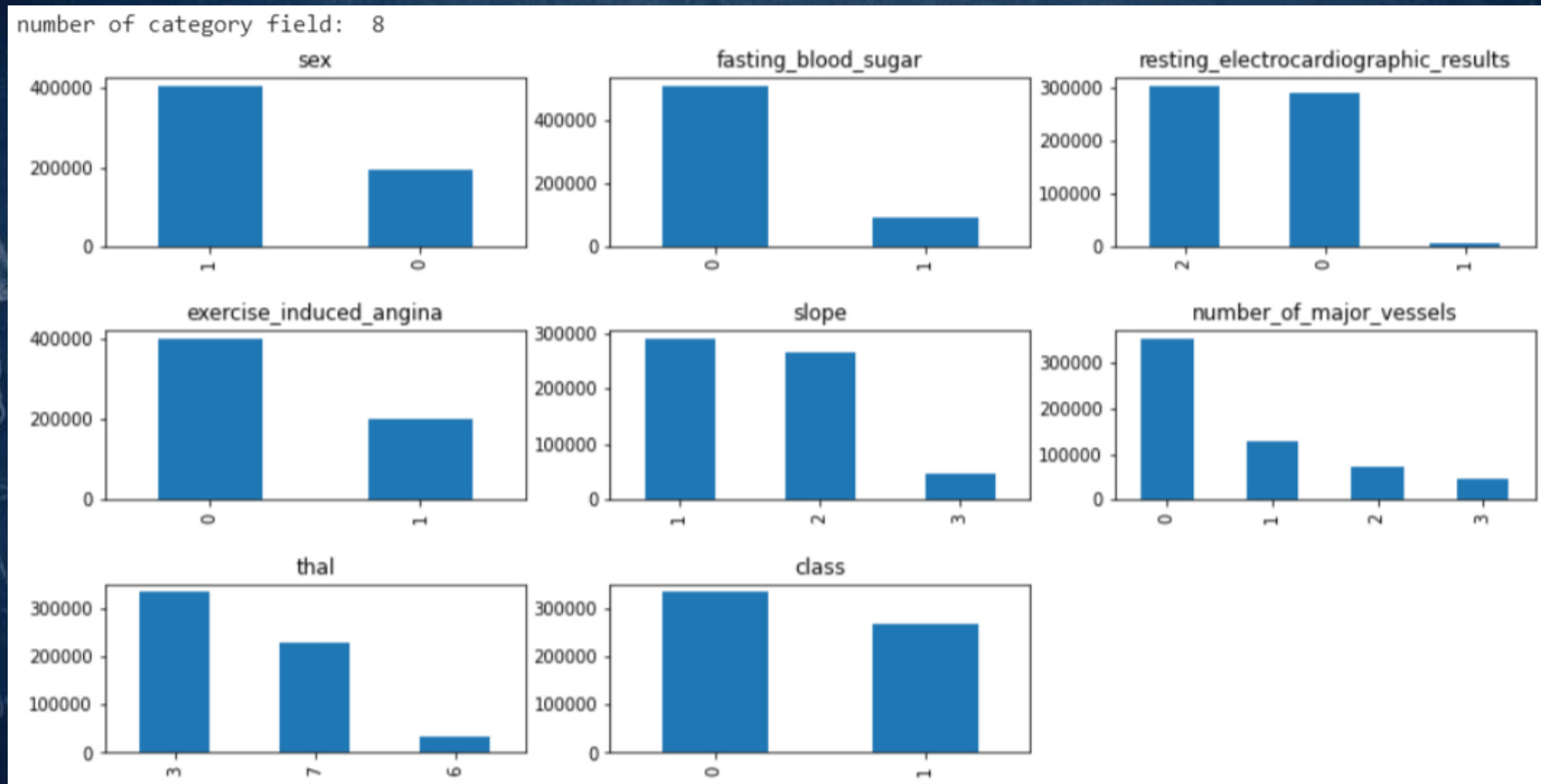


1) age: Có hơn 80% người ở độ tuổi khoảng 38-> 67.

2) serum_cholesterol, maximum_heart_rate_achieve lần lượt dao động trong các khoảng 190->310 và 135->180 là chủ yếu.

3. Visualize distribution

Biến phân loại



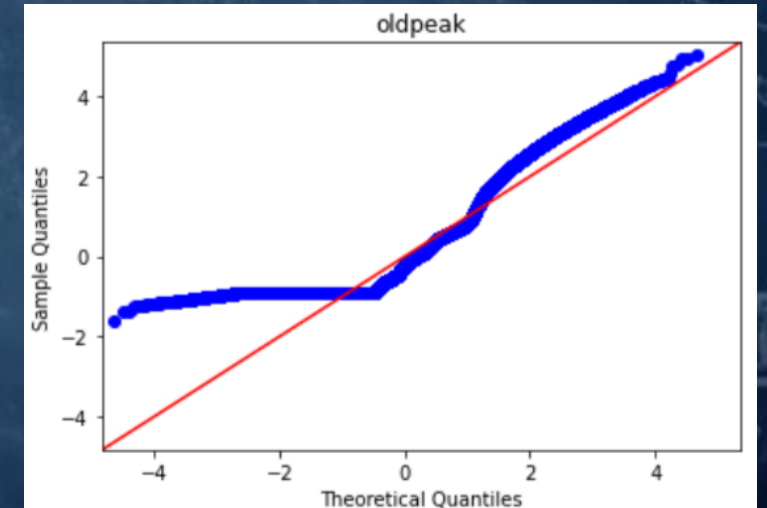
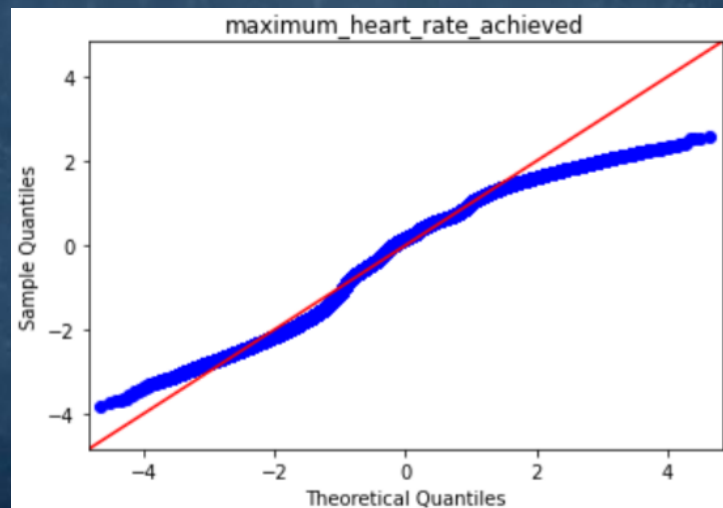
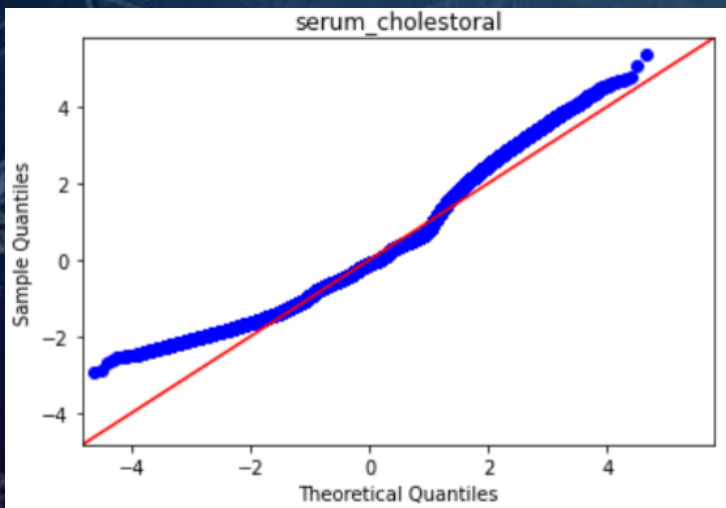
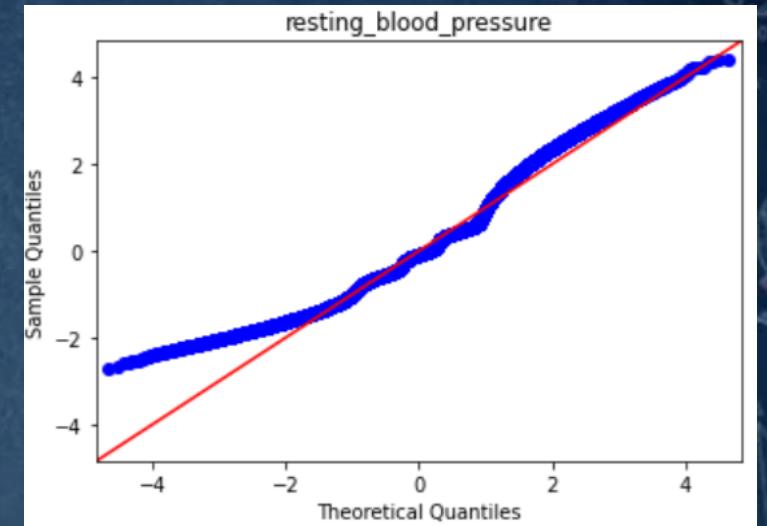
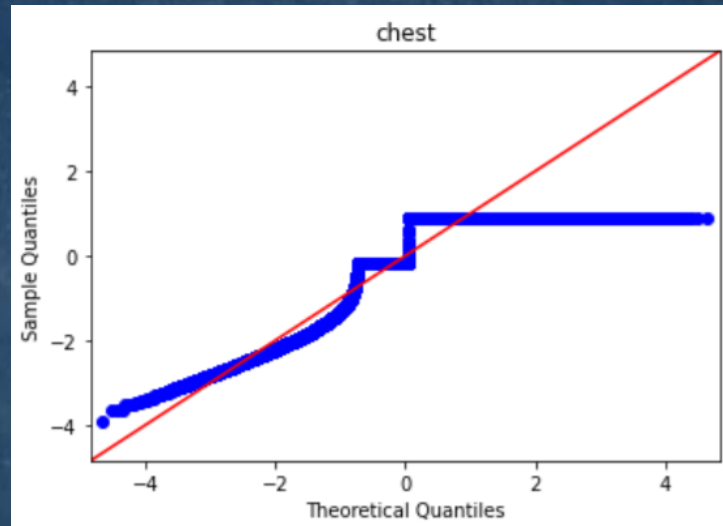
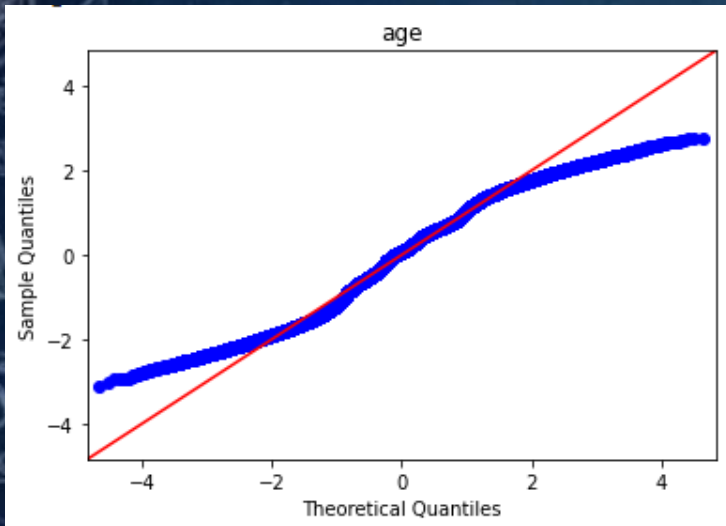
1) Male và exercise_induced_angina là 0 chiếm 2/3

2) Number_of_major_vessels và thal giảm đều theo thứ tự tương ứng là 0->1->2->3 và 3->7->6

3) fasting_blood_sugar và resting_electrocardiographic_results đều có chỉ số 1 là thấp nhất, các chỉ số còn lại cao hơn hẳn.

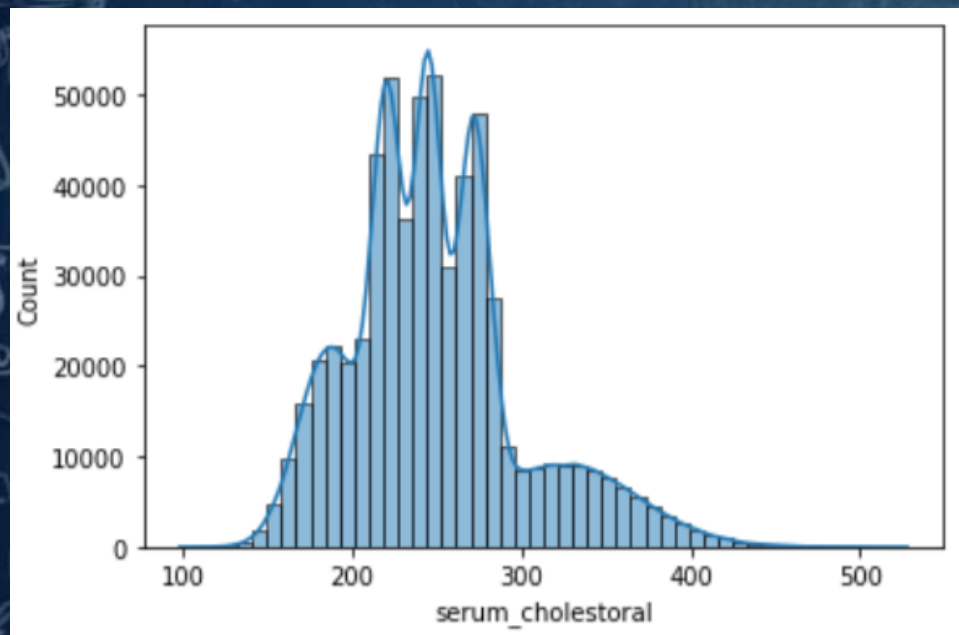
4. Kiểm tra outlier

Dùng biểu đồ Q-Q Plot

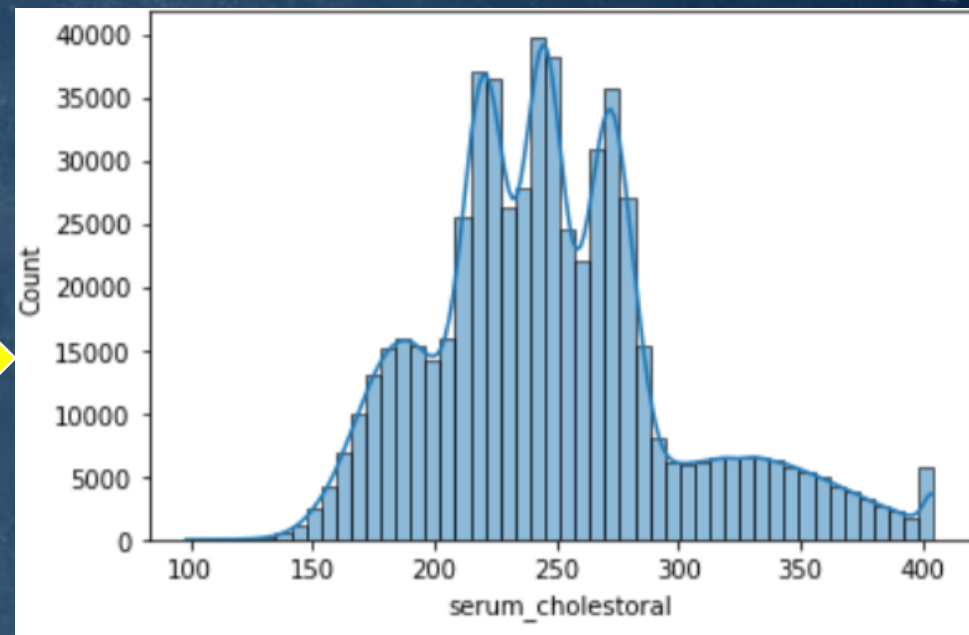
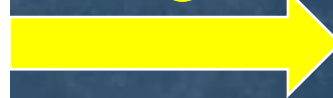


Các feature đều có outlier

4. Kiểm tra outlier



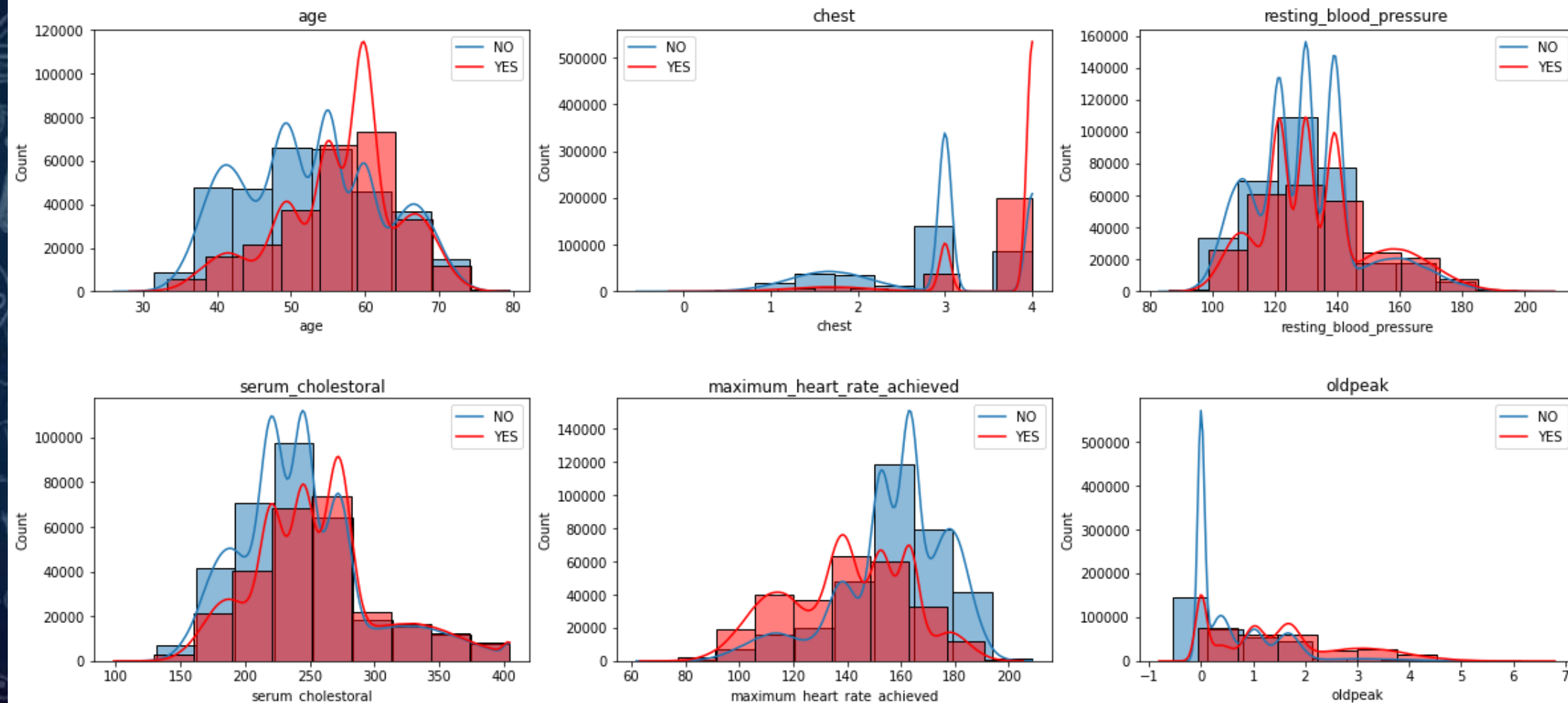
3-sigma



5. Phân phối của dữ liệu đối với từng class

Biến liên tục

number of numeric field: 6



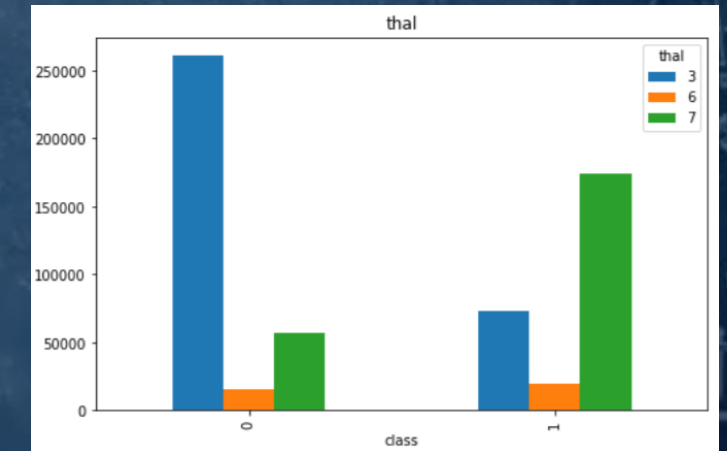
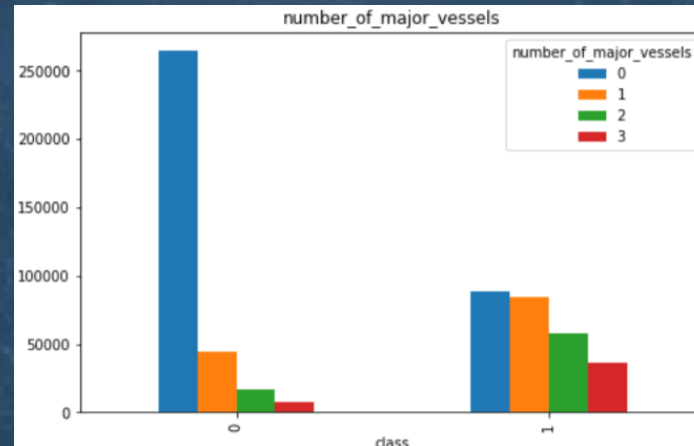
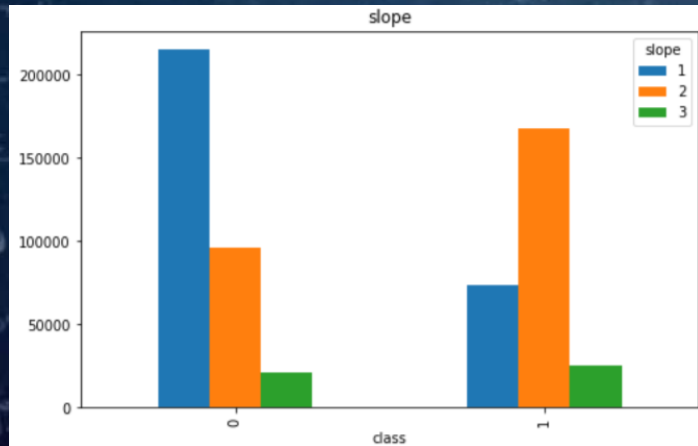
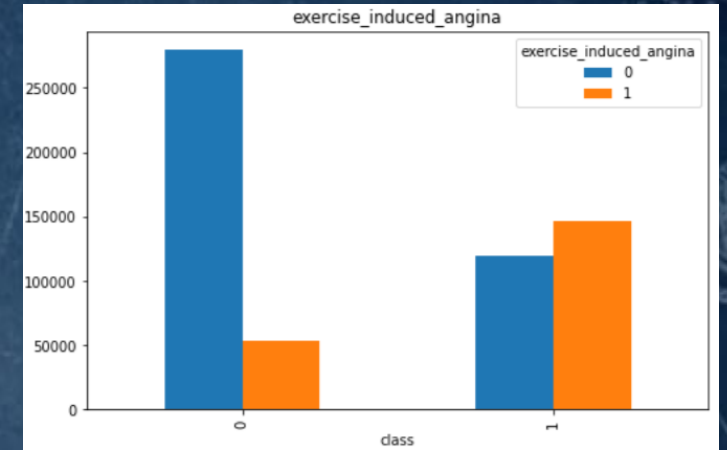
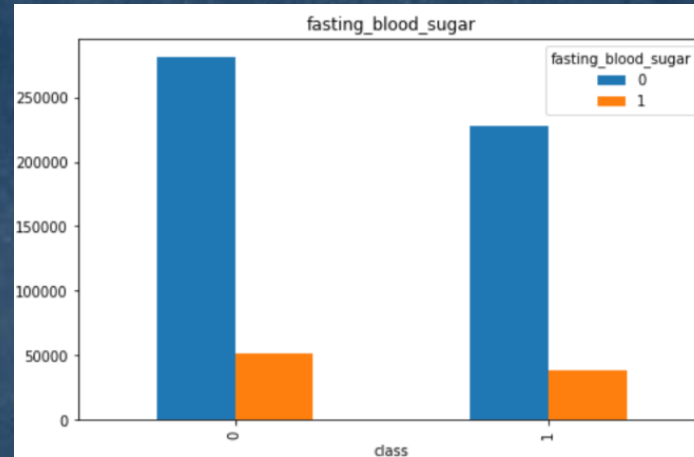
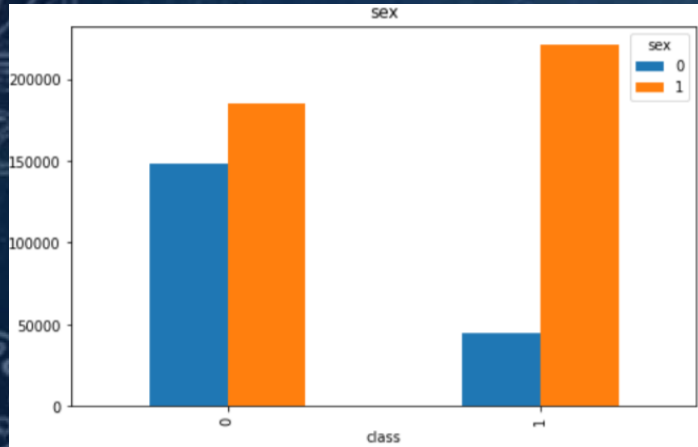
5. Phân phối của dữ liệu đối với từng class

Biến liên tục

- 1) age: 30->55 NO cao hơn YES, từ 55 trở đi YES cao hơn NO.
- 2) chest: -0.5 -> 3.5 NO cao hơn YES, từ 3.5 trở đi YES cao hơn NO.
- 3) resting_blood_pressure: 90 ->145 NO cao hơn YES, từ 145 trở đi YES cao hơn NO.
- 4) serum_cholesterol: 140->250 NO cao hơn YES, từ 250 trở đi YES cao hơn NO.
- 5) maximum_heart_rate_achieved:80-> 145 YES cao hơn NO, từ 145 trở đi NO cao hơn YES
- 6) oldpeak: từ -0.5->0.8 NO cao hơn YES, từ 0.8 trở đi YES cao hơn NO

5. Phân phối của dữ liệu đối với từng class

Biến phân loại



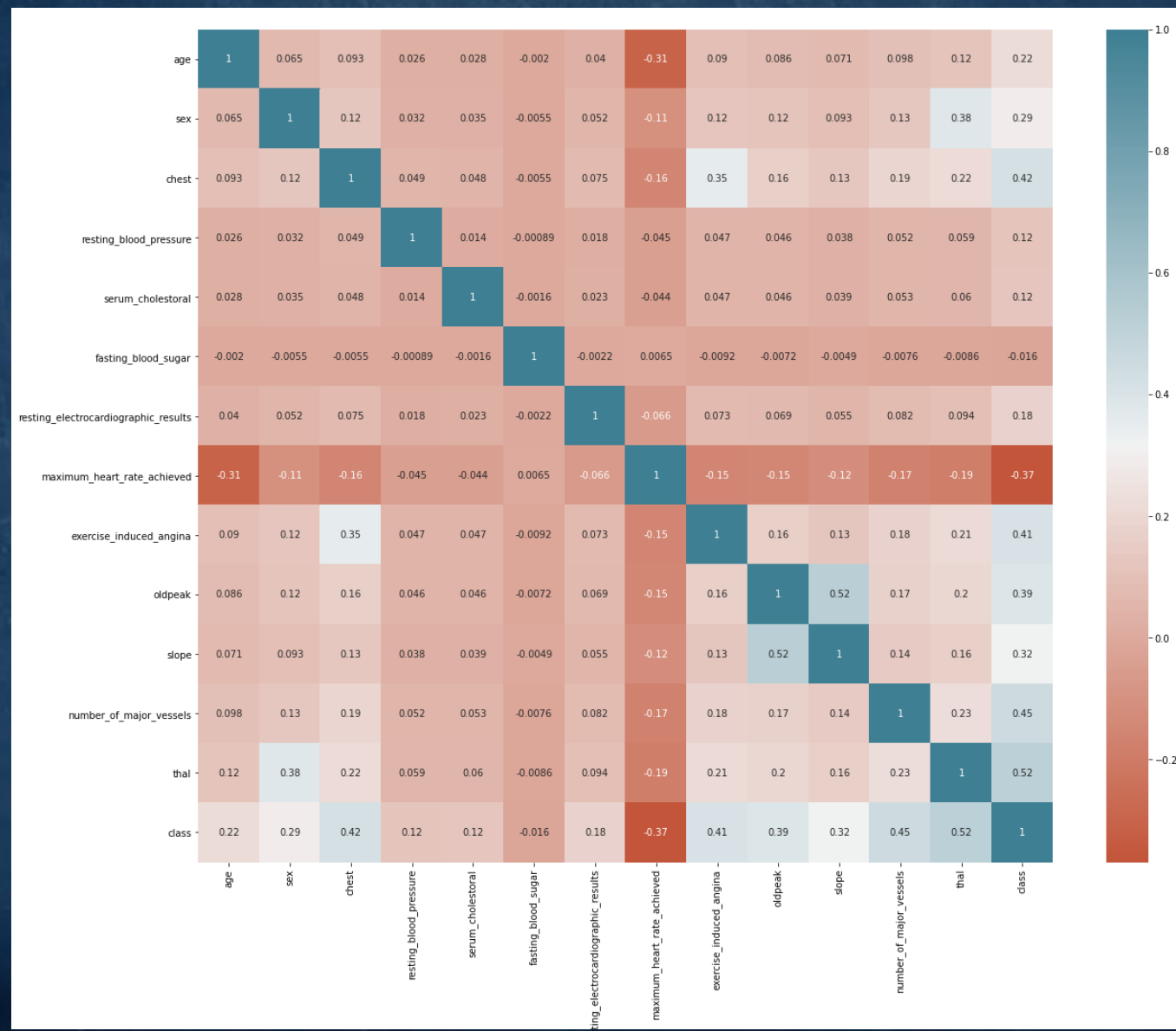
5. Phân phối của dữ liệu đối với từng class

Biến phân loại

- 1) sex: male không có sự chênh lệch lớn về bị hay không bị bệnh tim, chủ yếu là chênh lệch ở female
- 2) fasting_blood_sugar: những người bị và không bị bệnh đều có chỉ số 0 cao hơn hẳn chỉ số 1
- 3) number_of_major_vessels: những người không bị bệnh thì có chỉ số 0 rất cao
- 4) resing_electrocardiographic_results và thal đều có chỉ số 1 rất thấp ở cả người bị và không bị bệnh.

5. Phân phối của dữ liệu đối với từng class

Biểu đồ correlation tương quan giữa các biến

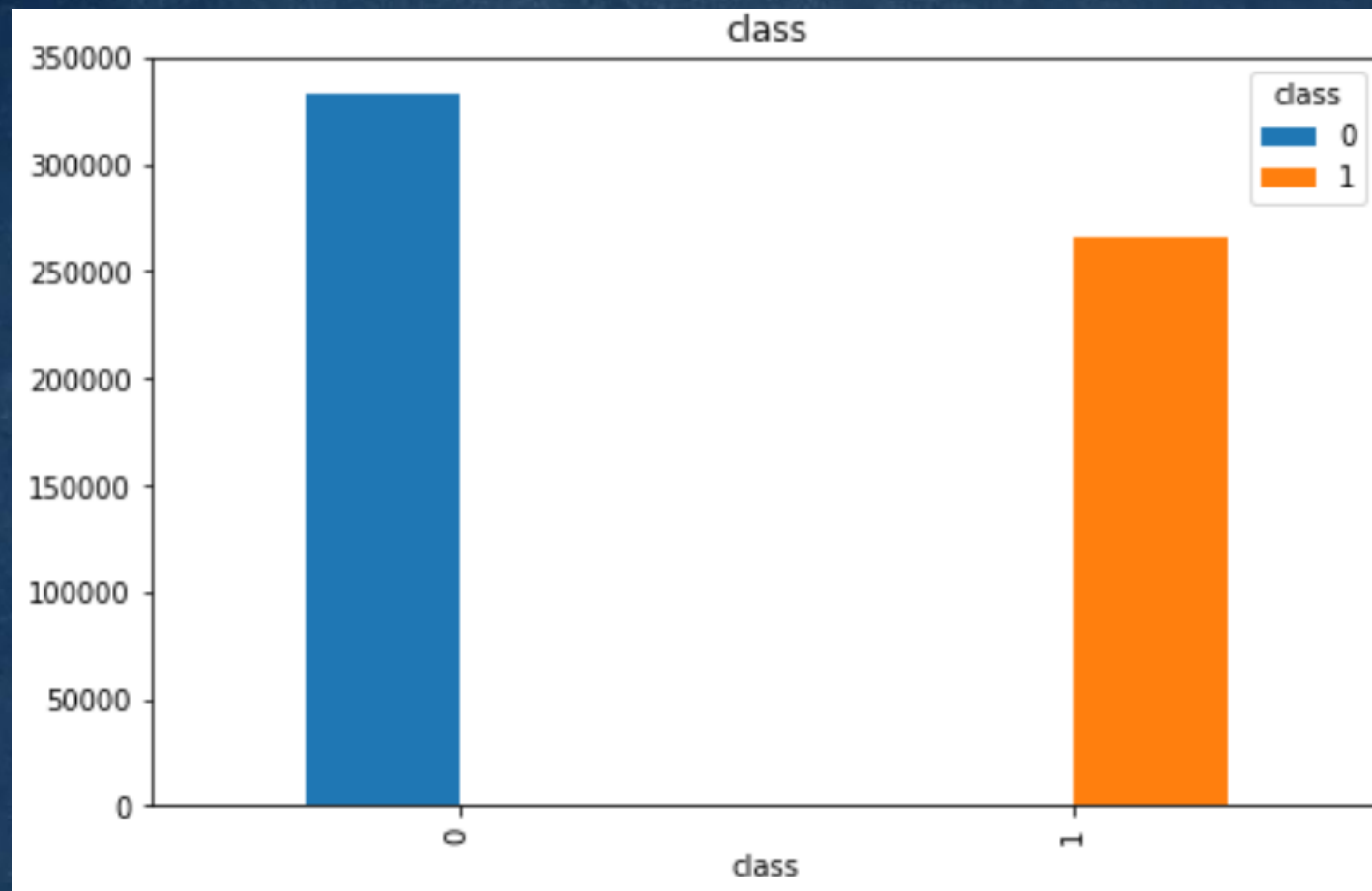


5. Phân phối của dữ liệu đối với từng class

Biến phân loại

- 1) sex: male không có sự chênh lệch lớn về bị hay không bị bệnh tim, chủ yếu là chênh lệch ở female
- 2) fasting_blood_sugar: những người bị và không bị bệnh đều có chỉ số 0 cao hơn hẳn chỉ số 1
- 3) number_of_major_vessels: những người không bị bệnh thì có chỉ số 0 rất cao
- 4) resing_electrocardiographic_results và thal đều có chỉ số 1 rất thấp ở cả người bị và không bị bệnh.

6. Phân chia tập train/validation



Biến mục tiêu class không bị mất cân bằng
nên ta sẽ dùng thang đo accuracy để đánh giá!

7. Huấn luyện mô hình

1 { LogisticRegression
Accuracy_train: 0.8716
Accuracy_val: 0.8713

2 { Decision Tree
Accuracy_train: 1.0
Accuracy_val: 0.8504

3 { KNeighborsClassifier
Accuracy_train: 0.8713
Accuracy_val: 0.8131

4 { RandomForestClassifier
Accuracy_train: 1.0
Accuracy_val: 0.8978

