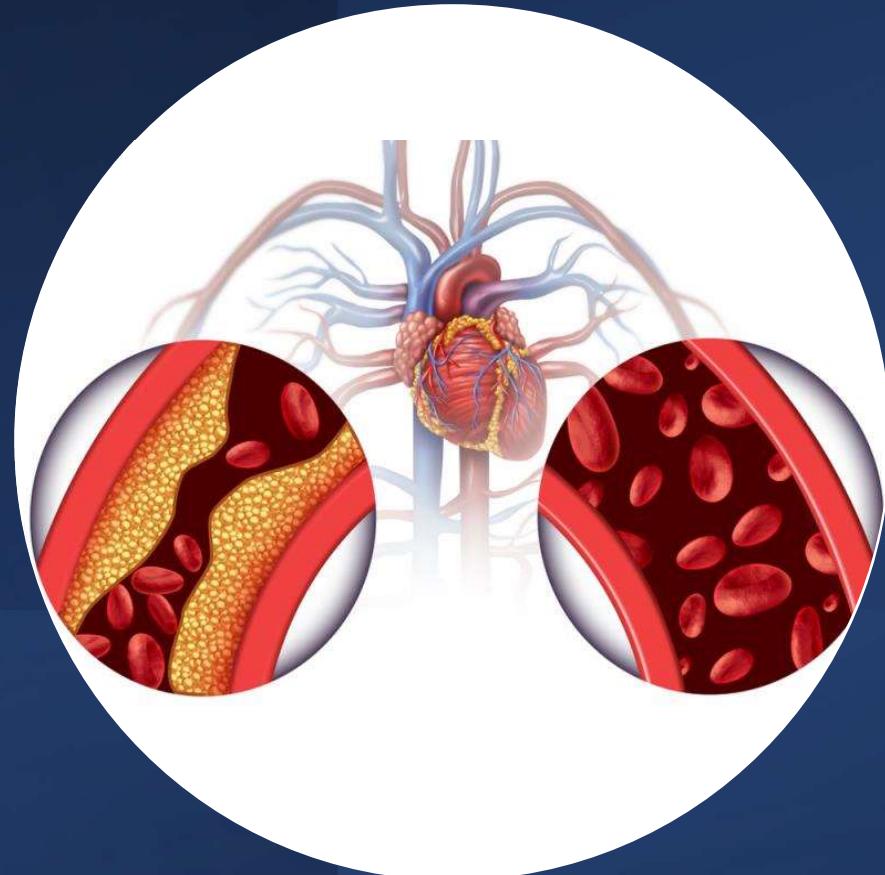


Xây dựng mô hình học máy dự đoán người có nguy cơ bệnh tim do xơ vữa động mạch (ASCVD)

Học viên: **Đồng Sỹ Huy (N7-DA32)**

Giảng viên: **Ms. Thắm**

Mentor: **Ms. Thắm**



Nội dung

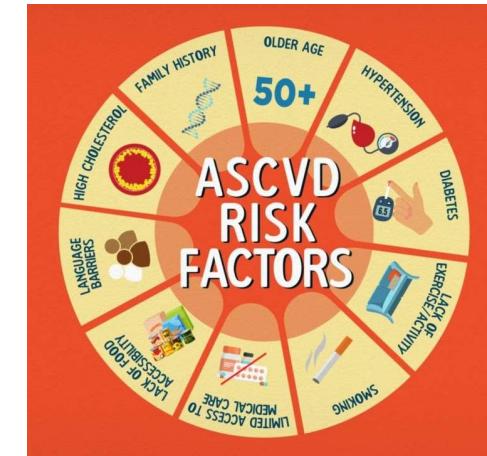
1. Giới thiệu chung
2. Xử lý và phân tích dữ liệu
3. Lựa chọn thuộc tính và xây dựng mô hình
4. Tổng kết & đề xuất



1. Giới thiệu chung

Bệnh lý tim mạch (Cardiovascular diseases - CVD) là một trong những nguyên nhân chính gây tử vong cao trên toàn cầu với gần 17,9 triệu người chết vì CVD vào năm 2016 (WHO, 2017).

Mặc dù các bất thường về tim và mạch máu khó xác định trước, nhưng vẫn có thể **dự đoán việc mắc bệnh tim mạch trong tương lai dựa trên một số kết quả kiểm tra sức khỏe thông thường cũng như lối sống.**



Bộ dữ liệu:

<https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset>

Dữ liệu bao gồm:

- 70,000 dòng
- 11 cột feature và 1 cột target

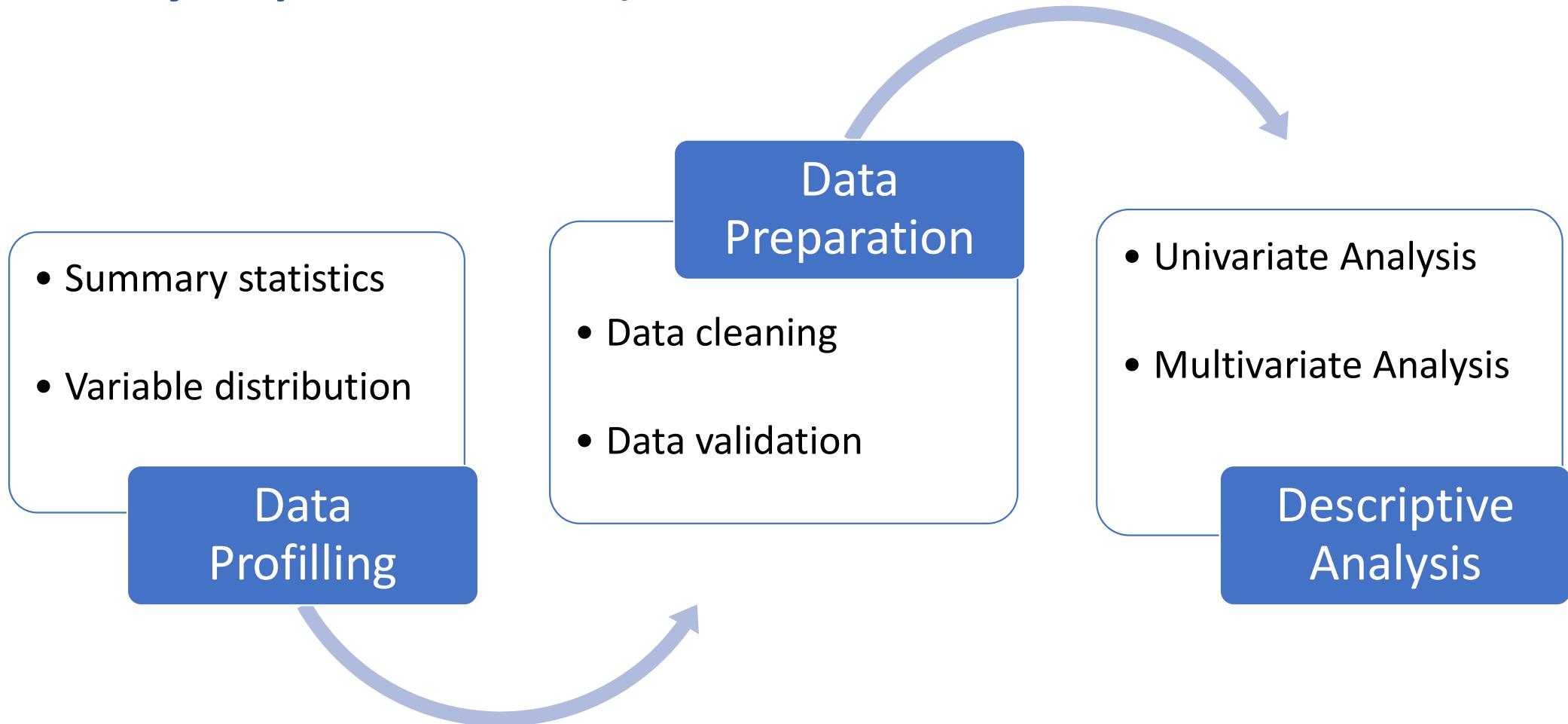
	id	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco	active	cardio	
2191	3096	20573	2	167	80.0	160	90		1	1	False	False	True	True
21070	30099	21811	1	158	98.0	110	70		3	1	False	False	True	True
3941	5572	16922	1	169	80.0	130	90		1	1	False	False	True	True
5508	7833	17095	2	164	75.0	130	90		1	1	False	False	True	False
60171	85908	21063	1	174	65.0	120	80		1	1	False	False	True	True
67471	96340	17610	1	151	70.0	100	90		2	1	False	False	True	False
59652	85165	15220	1	162	56.0	90	60		1	1	False	False	True	False
6133	8721	14822	1	165	65.0	120	90		1	1	False	False	True	False
9041	12893	20440	1	170	69.0	110	70		1	1	False	False	True	False
55393	79022	14725	1	173	80.0	12	80		2	1	False	False	True	True

1. Giới thiệu chung

Ý nghĩa các cột

Feature	Variable Type	Variable	Value Type
Age	Objective Feature	age	int (days)
Height	Objective Feature	height	int (cm)
Weight	Objective Feature	weight	float (kg)
Gender	Objective Feature	gender	categorical code (1 - women, 2 - men)
Systolic blood pressure	Examination Feature	ap_hi	int
Diastolic blood pressure	Examination Feature	ap_lo	int
Cholesterol	Examination Feature	cholesterol	1: normal, 2: above normal, 3: well above normal
Glucose	Examination Feature	gluc	1: normal, 2: above normal, 3: well above normal
Smoking	Subjective Feature	smoke	binary
Alcohol intake	Subjective Feature	alco	binary
Physical activity	Subjective Feature	active	binary
Presence or absence of cardiovascular disease	Target Variable	cardio	binary

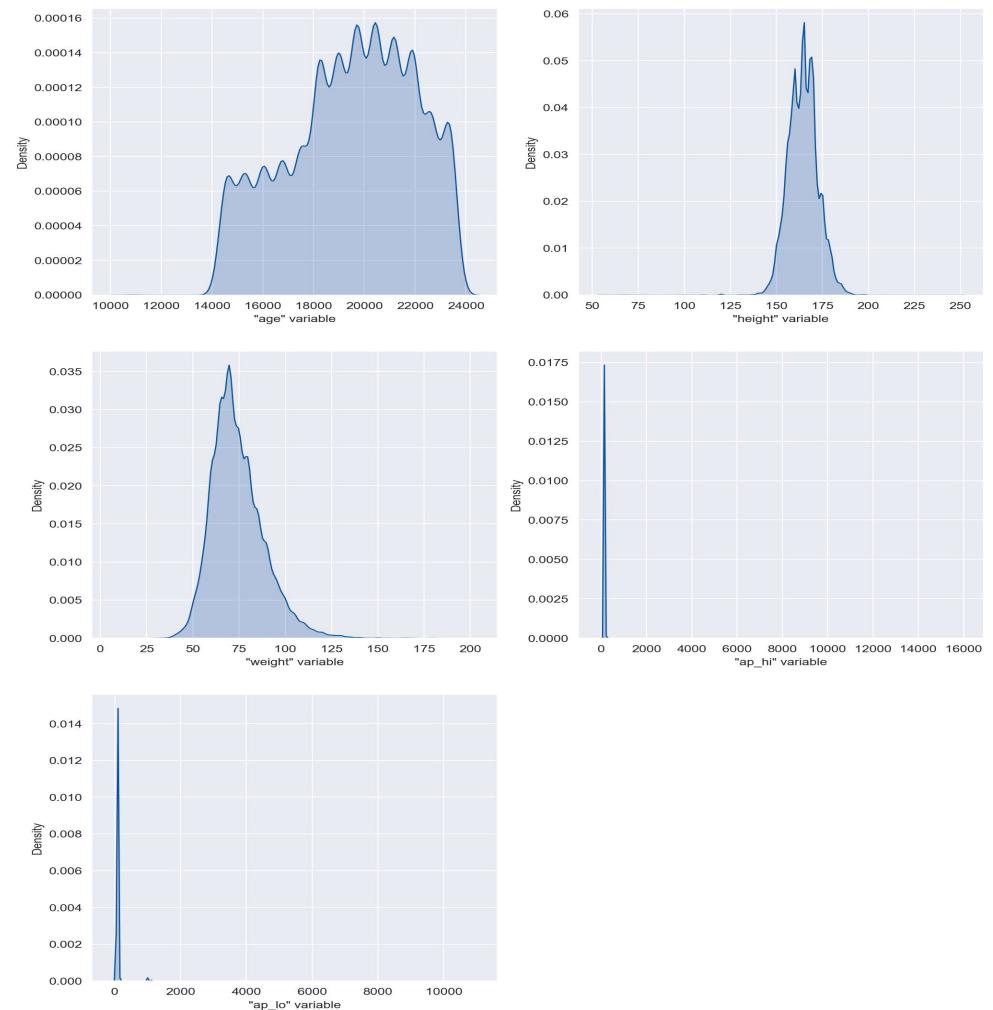
2. Xử lý và phân tích dữ liệu



2. Xử lý và phân tích dữ liệu

Data Profiling

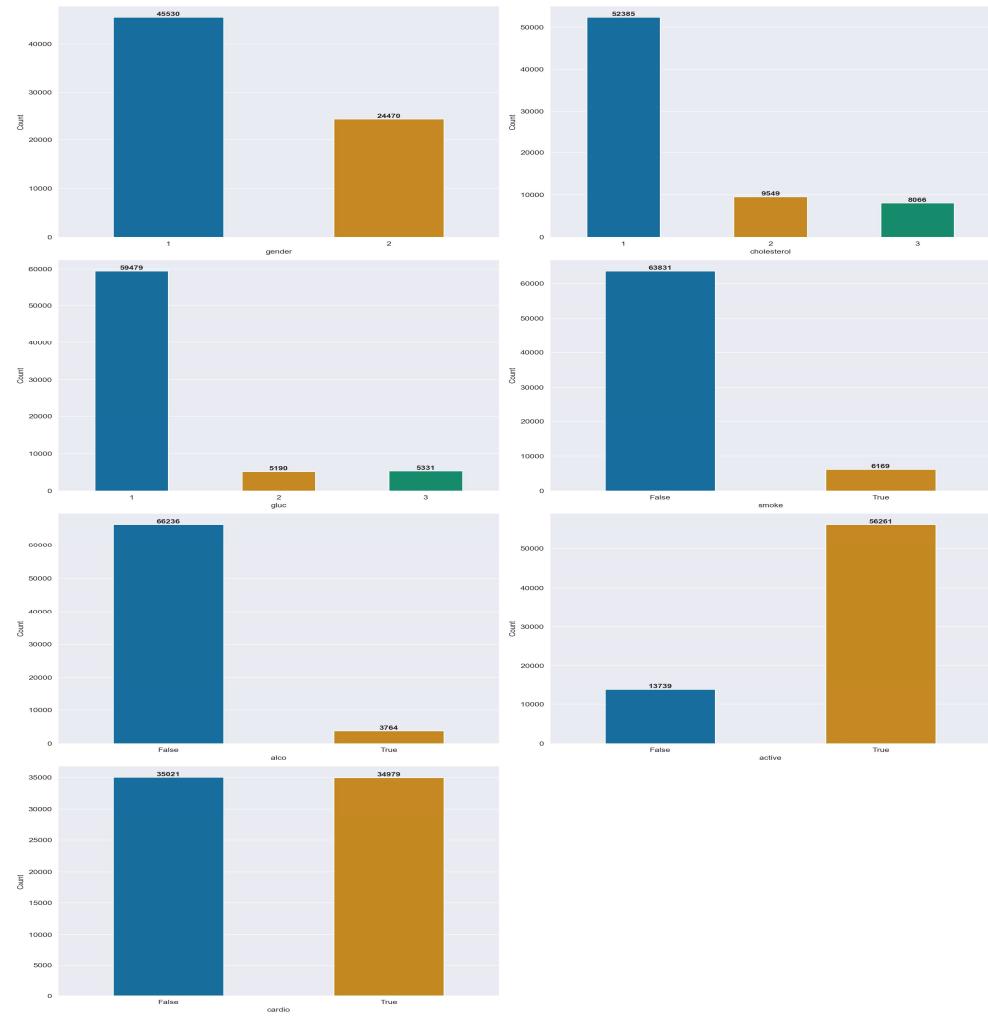
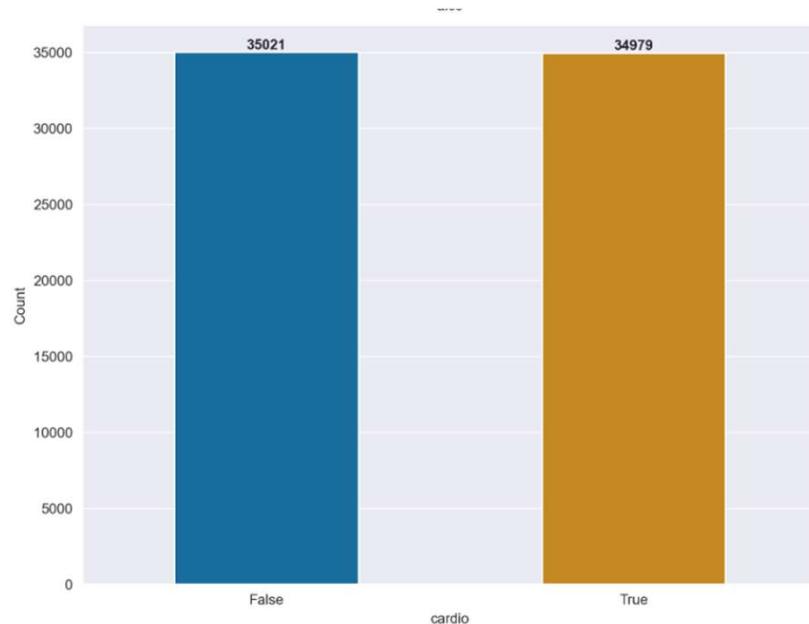
	age	height	weight	ap_hi	ap_lo
Values	70000.00	70000.00	70000.00	70000.00	70000.00
Missing	0.00	0.00	0.00	0.00	0.00
Distinct	8076.00	109.00	287.00	153.00	157.00
Max	23713.00	250.00	200.00	16020.00	11000.00
Q3	21327.00	170.00	82.00	140.00	90.00
Mean	19468.87	164.36	74.21	128.82	96.63
Median	19703.00	165.00	72.00	120.00	80.00
Q1	17664.00	159.00	65.00	120.00	80.00
Min	10798.00	55.00	10.00	-150.00	-70.00
Range	12915.00	195.00	190.00	16170.00	11070.00
IQR	3663.00	11.00	17.00	20.00	10.00
Std	2467.25	8.21	14.40	154.01	188.47
Var	6087330.79	67.41	207.24	23719.52	35521.89



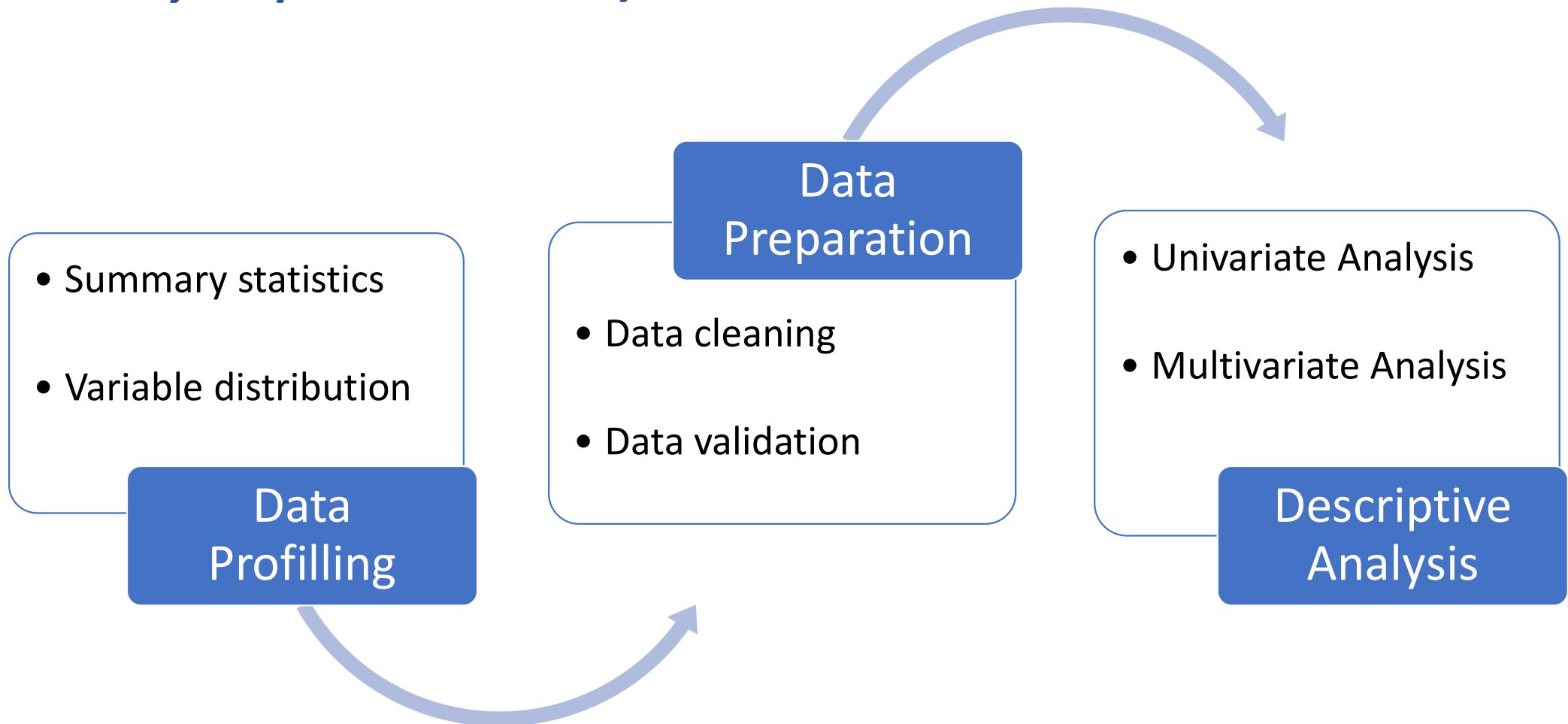
2. Xử lý và phân tích dữ liệu

Data Profiling

	gender	cholesterol	gluc	smoke	alco	active	cardio
Count	70000	70000	70000	70000	70000	70000	70000
Distinct	2	3	3	2	2	2	2
Top	1	1	1	False	False	True	False
Freq	45530	52385	59479	63831	66236	56261	35021



2. Xử lý và phân tích dữ liệu



2. Xử lý và phân tích dữ liệu

Data Preparation

Vấn đề ràng buộc dữ liệu:

- Cột app_hi: Bỏ các dữ liệu nằm ngoài 40 – 370
- Cột app_lo: Bỏ các dữ liệu nằm ngoài 20 - 360

Vấn đề về tính đồng nhất dữ liệu:

- Cột app_hi và app_lo: Bỏ những dữ liệu app_lo > app_hi
- Dữ liệu trùng lặp: Loại bỏ

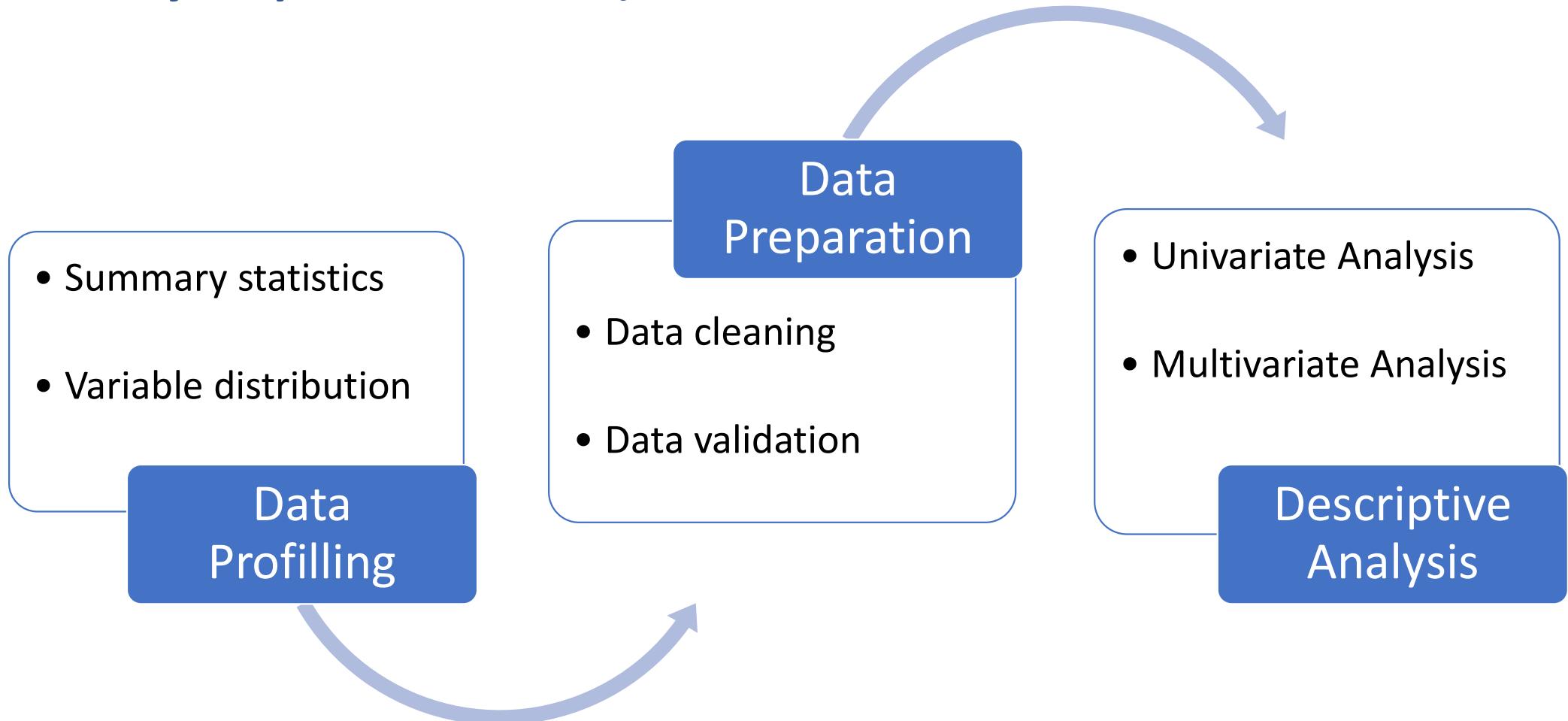
Vấn đề về đơn vị tính:

- Cột age: Chuyển đổi về đơn vị tính theo năm

Duplicate Values (all of the columns): 0									
Duplicate Values (without 'id'): 0									
	id	age_year	gender	height	weight	ap_hi	ap_lo	cholesterol	\
0	0	50	2	168	62.0	110	80	1	
1	1	55	1	156	85.0	140	90	3	
2	2	52	1	165	64.0	130	70	3	
3	3	48	2	169	82.0	150	100	1	
4	4	48	1	156	56.0	100	60	1	
...
64872	99993	53	2	168	76.0	120	80	1	
64873	99995	62	1	158	126.0	140	90	2	
64874	99996	52	2	183	105.0	180	90	3	
64875	99998	61	1	163	72.0	135	80	1	
64876	99999	56	1	170	72.0	120	80	2	
	gluc	smoke	alco	active	cardio				
0	1	False	False	True	False				
1	1	False	False	True	True				
2	1	False	False	False	True				
3	1	False	False	True	True				
4	1	False	False	False	False				
...				
64872	1	True	False	True	False				
64873	2	False	False	True	True				
64874	1	False	True	False	True				
64875	2	False	False	False	True				
64876	1	False	False	True	False				

[64877 rows x 13 columns]

2. Xử lý và phân tích dữ liệu

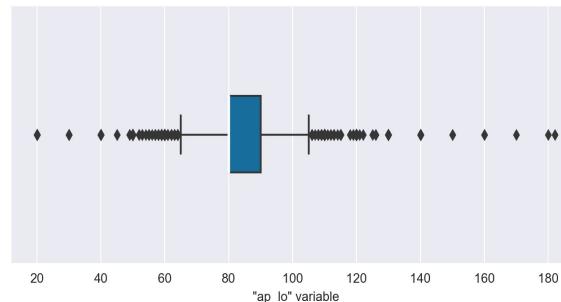
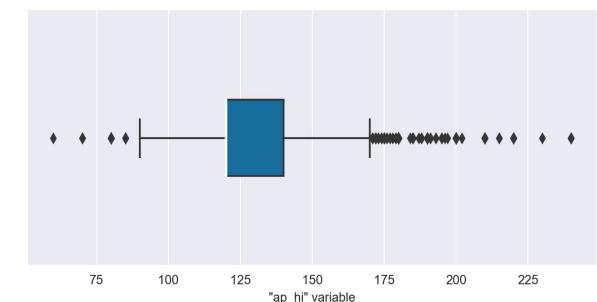
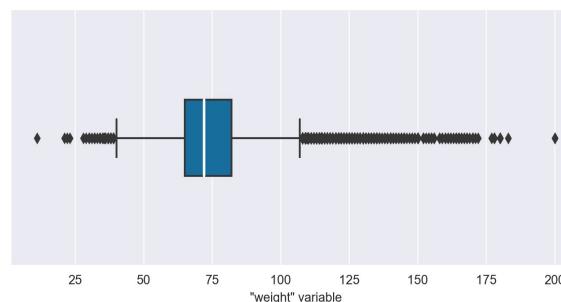
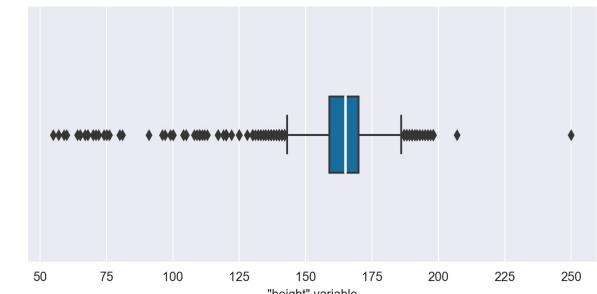
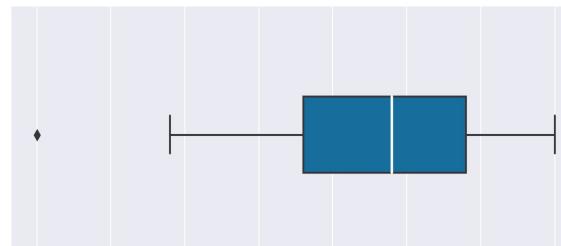


2. Xử lý và phân tích dữ liệu

Descriptive Analysis

Vấn đề outliers:

- age_year: 4 (lower 31.5, upper 75.5)
- height: 504 (lower 142.5, upper 186.5)
- weight: 1744 (lower 39.5 upper 107.5)
- ap_hi: 1012 (lower 90.0 upper 170.0)
- ap_lo: 3504 (lower 65.0 upper 105.0)



2. Xử lý và phân tích dữ liệu

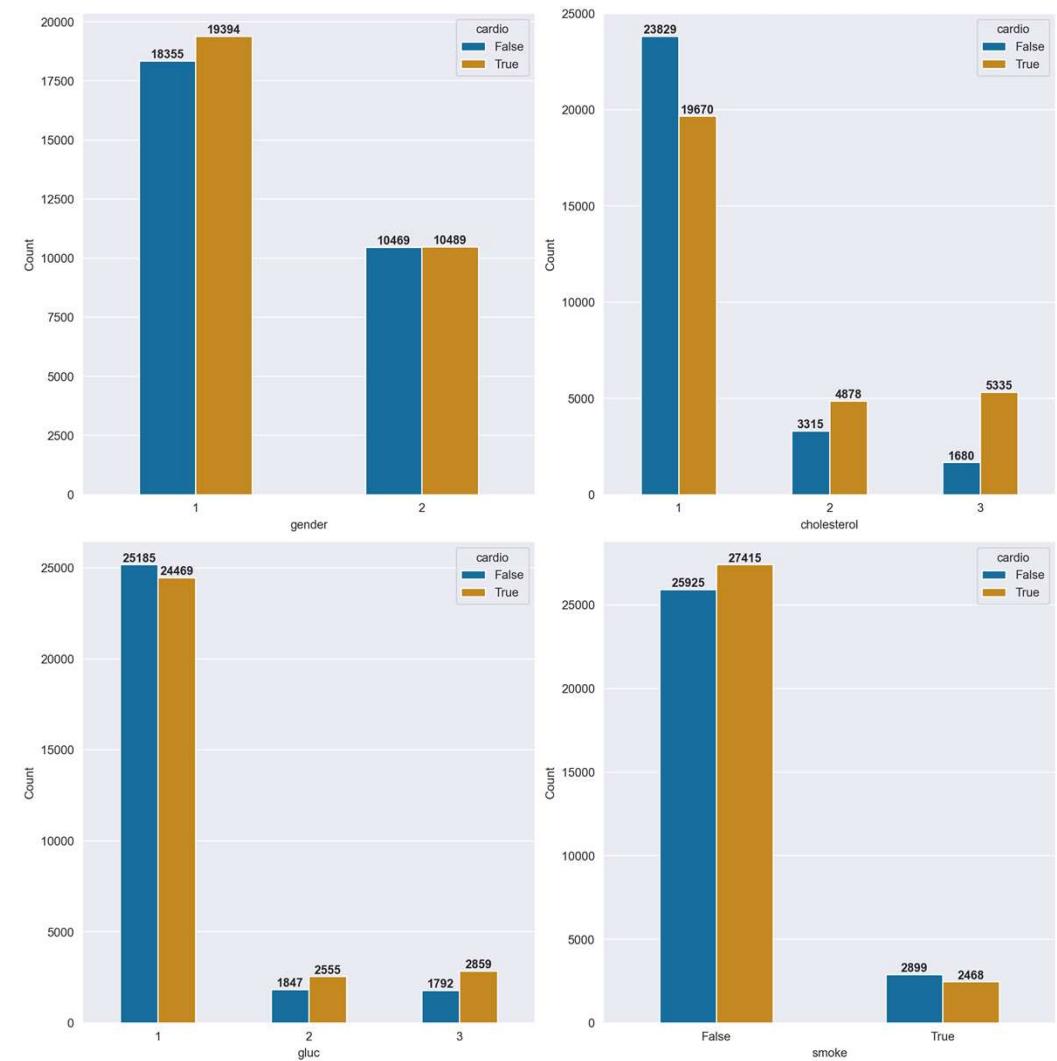
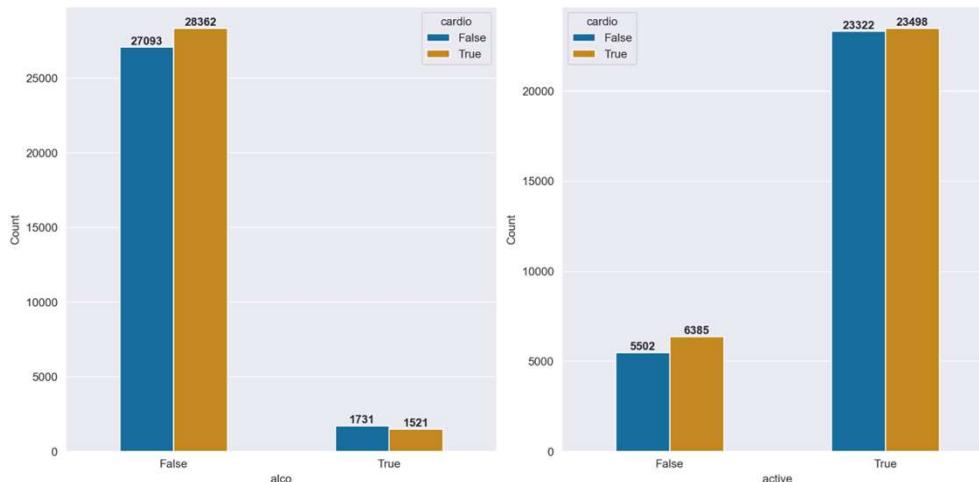
Descriptive Analysis

- Phân tích tỉ lệ người có nguy cơ mắc CVD theo:

- + Giới tính: Nữ > Nam
- + Lượng Cholesterol, Glucose : Bình thường > Cao
- + Hút thuốc, Uống rượu: Không > Có
- + Vận động: Có > Không

- Theo WHO, các yếu tố cần xem xét khi dự đoán CVD:

Giới, Tuổi, Huyết áp, Cholesterol, Hút thuốc và BMI.



2. Xử lý và phân tích dữ liệu

Descriptive Analysis

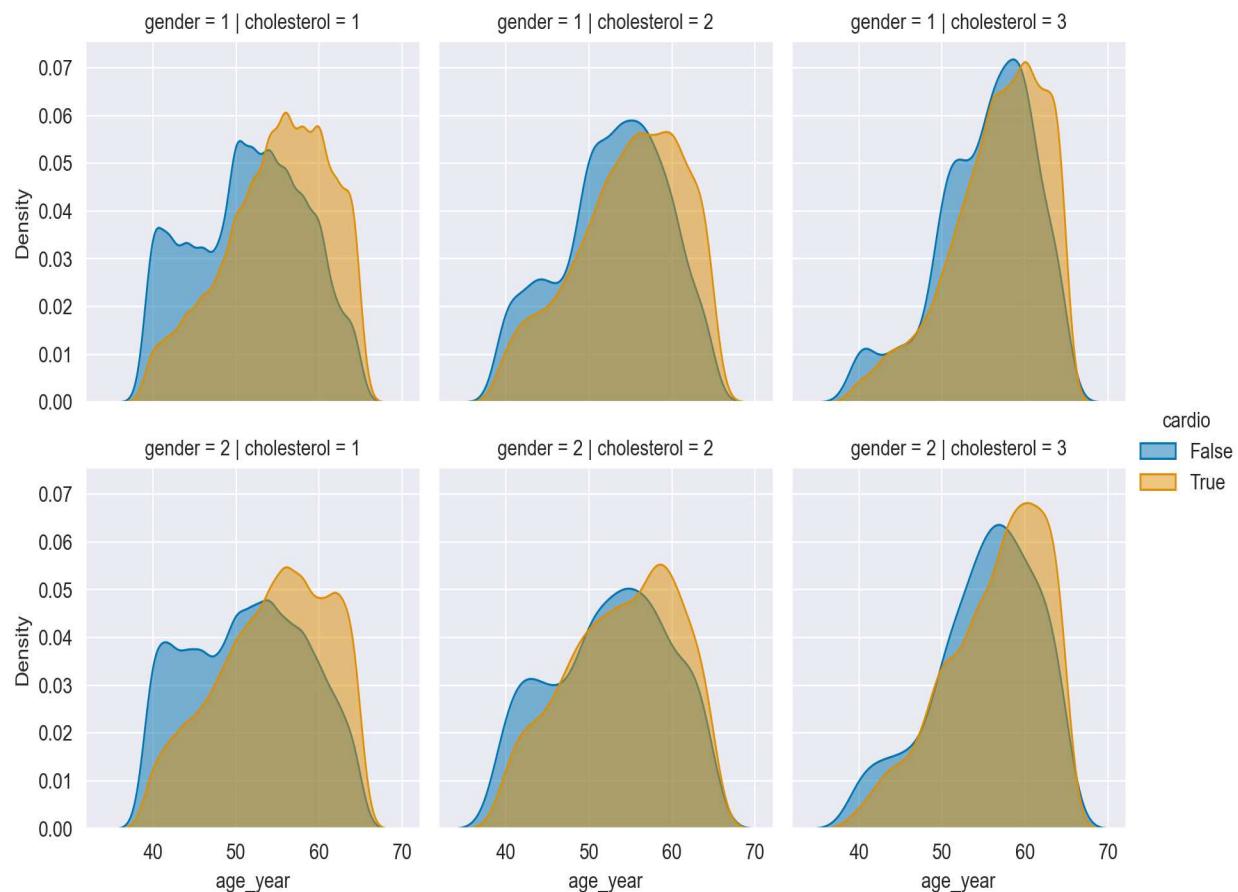
Tỷ lệ mắc bệnh tim mạch của nam giới và nữ giới theo tuổi:

- Mức cholesterol bình thường:

Ở độ tuổi 40 tỷ lệ là tương đương. Từ độ tuổi từ 50, tỷ lệ mắc CVD ở nam giới bắt đầu tăng, ở nữ giới vẫn ổn định hoặc bắt đầu giảm.

- Mức cholesterol từ cao đến rất cao:

Tỷ lệ mắc bệnh tim mạch của nam giới và nữ giới cũng tăng lên theo độ tuổi, nhưng tốc độ tăng của nam giới nhanh hơn nữ giới.

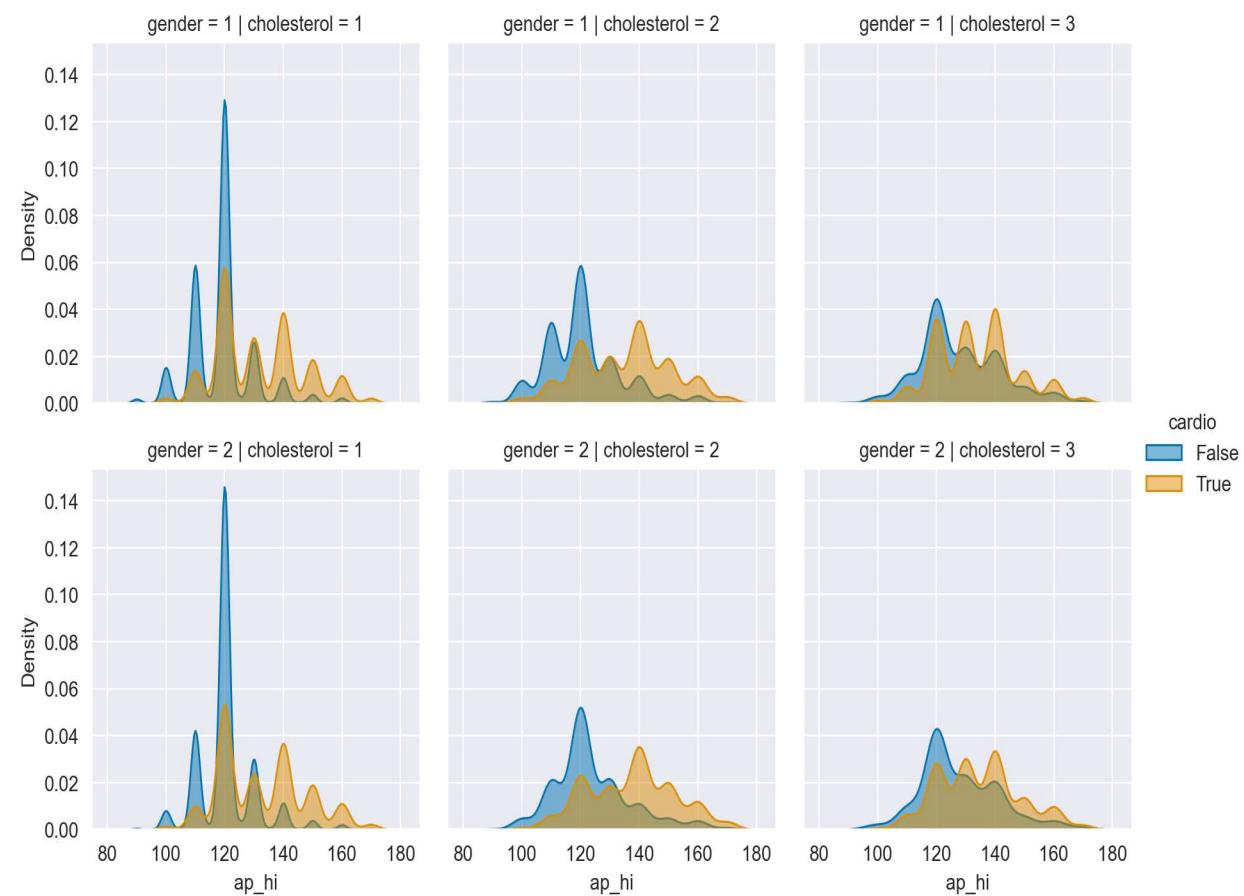


2. Xử lý và phân tích dữ liệu

Descriptive Analysis

Tương quan giữa giới tính, mức cholesterol, huyết áp và nguy cơ bệnh tim mạch :

- Người có huyết áp cao, đặc biệt là huyết áp cao ở mức cholesterol 3, có nguy cơ mắc bệnh tim mạch cao hơn.
- Phụ nữ có huyết áp cao và mức cholesterol cao, có nguy cơ mắc bệnh tim mạch cao hơn nam giới

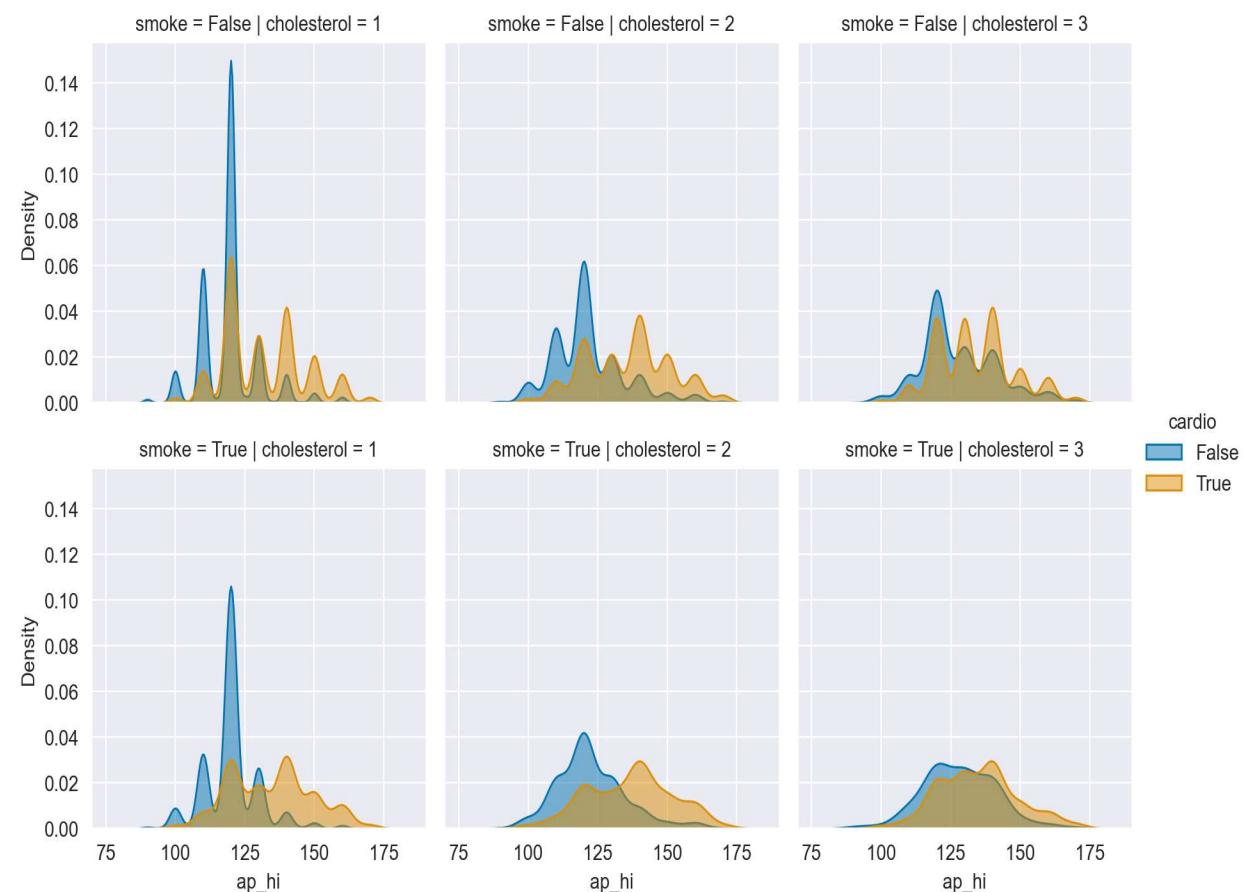


2. Xử lý và phân tích dữ liệu

Descriptive Analysis

Tương quan giữa hút thuốc, huyết áp, mức cholesterol và nguy cơ bệnh tim mạch:

- Hút thuốc, cholesterol cao và huyết áp cao làm tăng tỉ lệ mắc bệnh tim mạch.

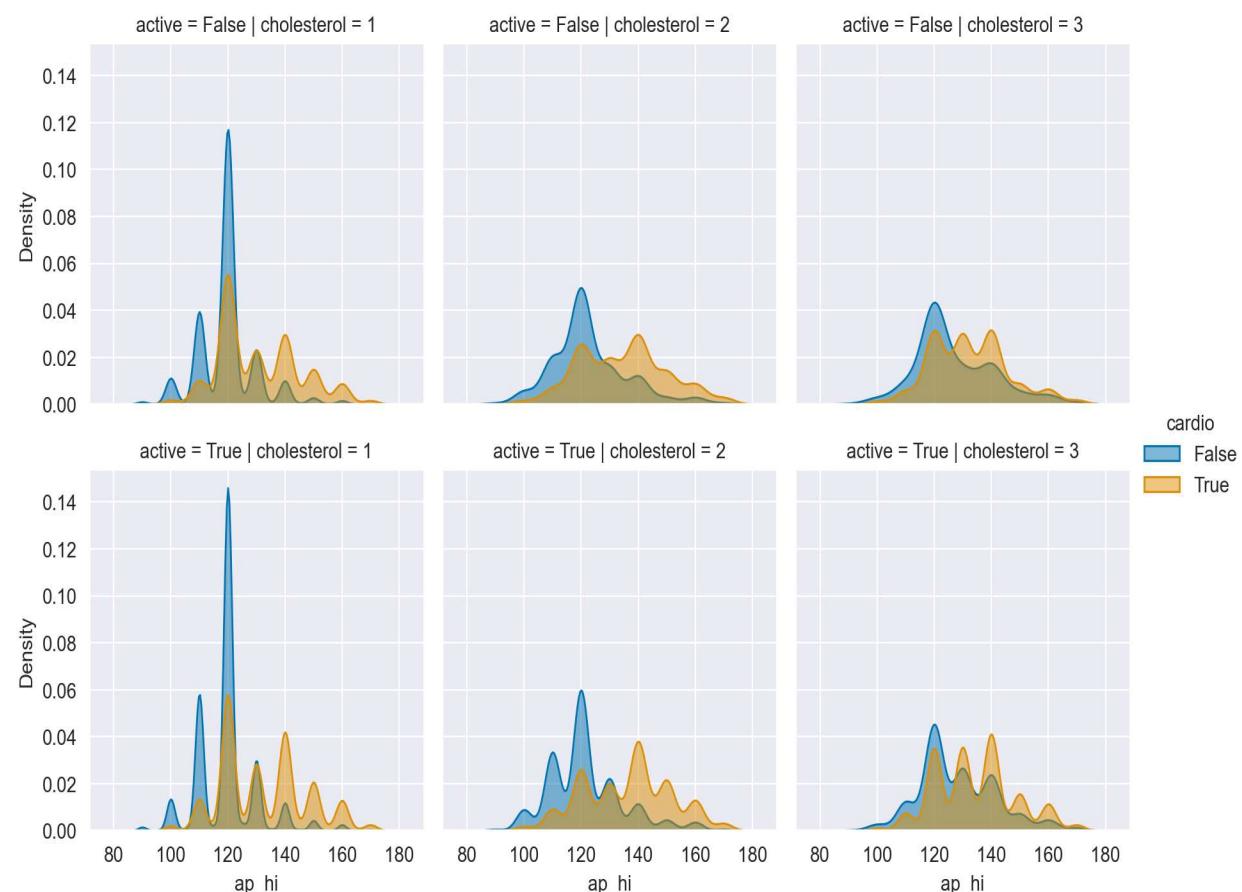


2. Xử lý và phân tích dữ liệu

Descriptive Analysis

Tương quan giữa hoạt động thể chất, huyết áp, mức cholesterol và nguy cơ bệnh tim mạch:

- Người hoạt động thể chất, đặc biệt là người hoạt động thể chất thường xuyên, có nguy cơ mắc bệnh tim mạch thấp hơn.
- Sự khác biệt về nguy cơ mắc bệnh tim mạch giữa người hoạt động thể chất và không hoạt động thể chất càng lớn khi cholesterol và huyết áp càng cao.

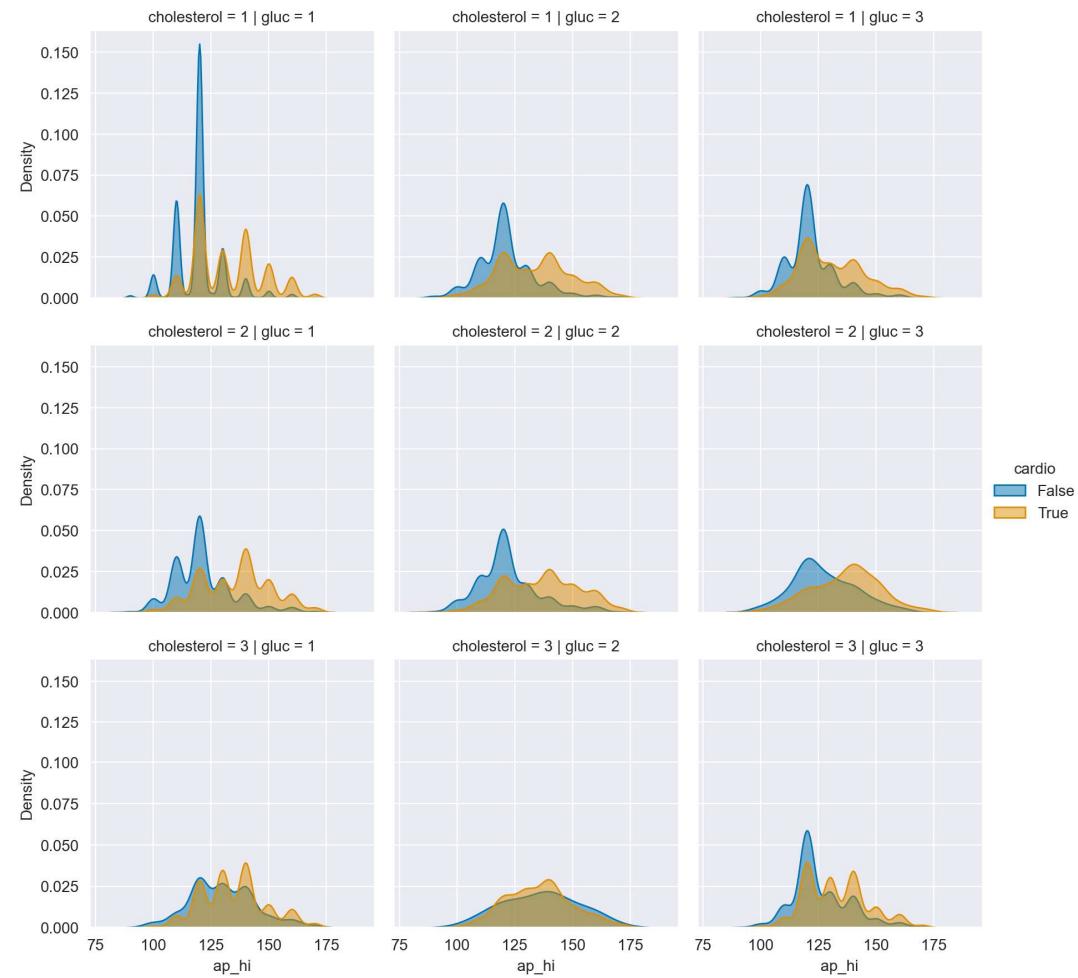


2. Xử lý và phân tích dữ liệu

Descriptive Analysis

Tương quan giữa tiểu đường, huyết áp, mức cholesterol và nguy cơ bệnh tim mạch:

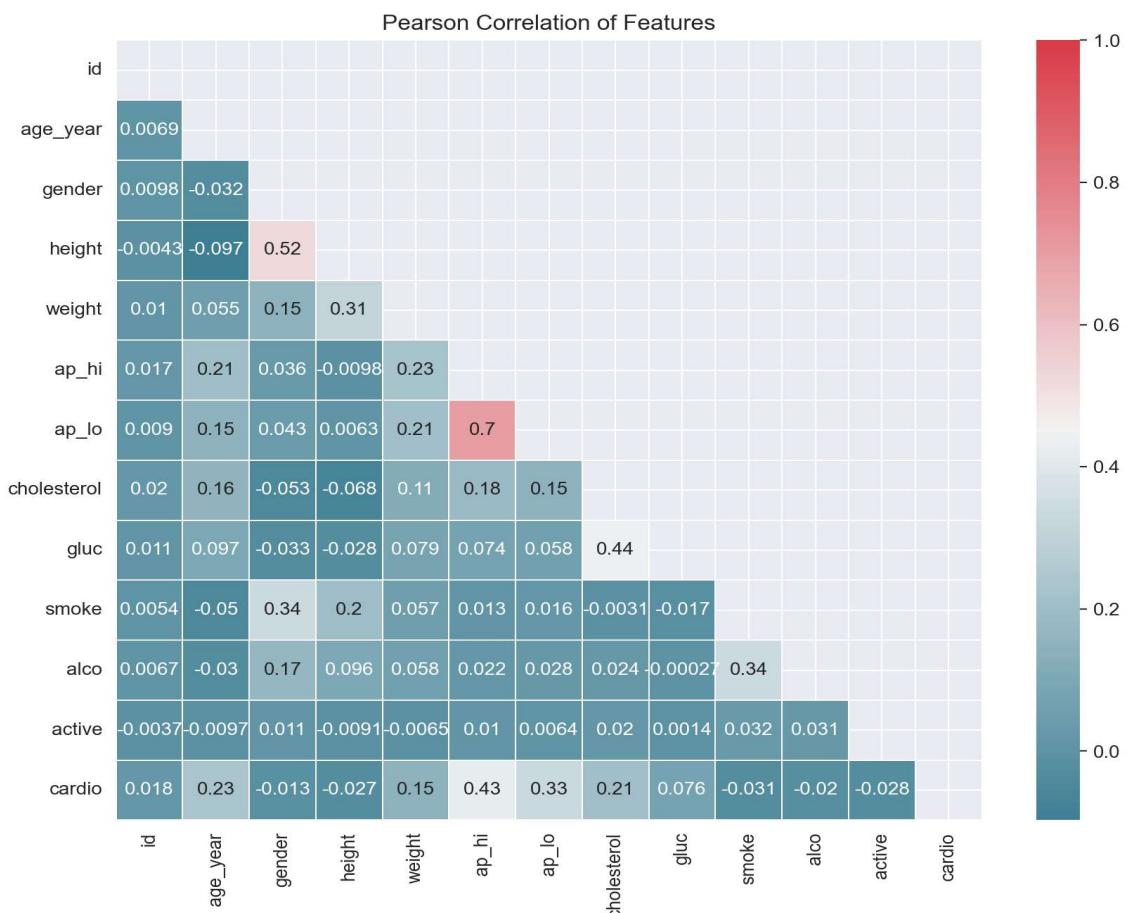
- Tiểu đường, huyết áp và cholesterol đều là những yếu tố nguy cơ quan trọng đối với bệnh tim mạch.
- Sự khác biệt về nguy cơ mắc bệnh tim mạch giữa các mức tiểu đường khác nhau càng tăng khi cholesterol và huyết áp càng cao.



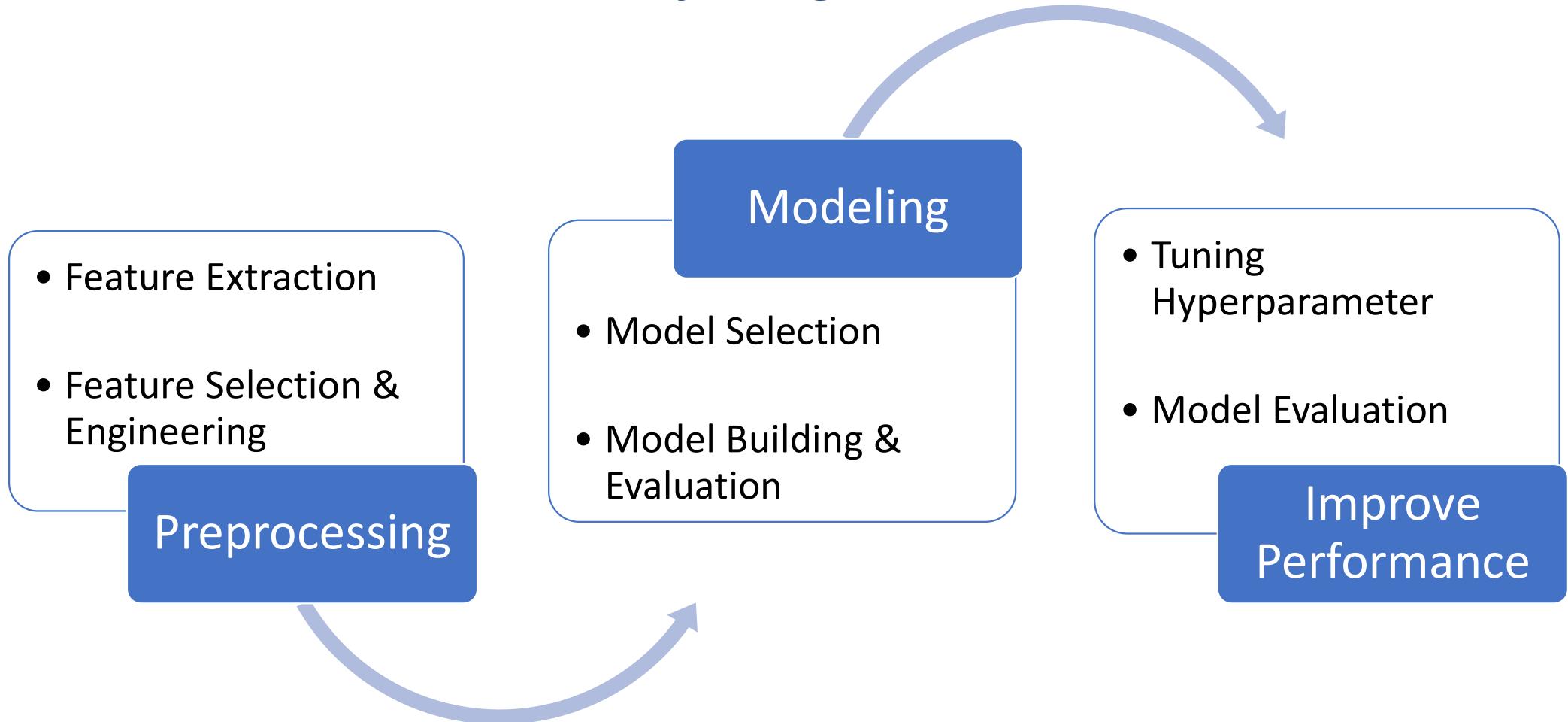
2. Xử lý và phân tích dữ liệu

Descriptive Analysis

- Huyết áp tâm thu có tương quan thuận mạn với huyết áp tâm trương
 - Cholesterol với tiểu đường có tương quan thuận với nhau
 - Hút thuốc với uống rượu có tương quan thuận với nhau
 - Tuổi, Cân nặng, Huyết áp, Cholesterol, Tiểu đường có tương quan thuận với tỉ lệ mắc bệnh tim mạch
- Có thể loại ap_lo, biến height và weight thành BMI.
- Vậy còn gender, gluc, smoke, alco, active?



3. Lựa chọn thuộc tính và xây dựng mô hình



3. Lựa chọn thuộc tính và xây dựng mô hình

Preprocessing: Feature Extraction

```
Drop columns done!
Duplicate Values (all of the columns): 0
Duplicate Values (without 'cardio'): 1565

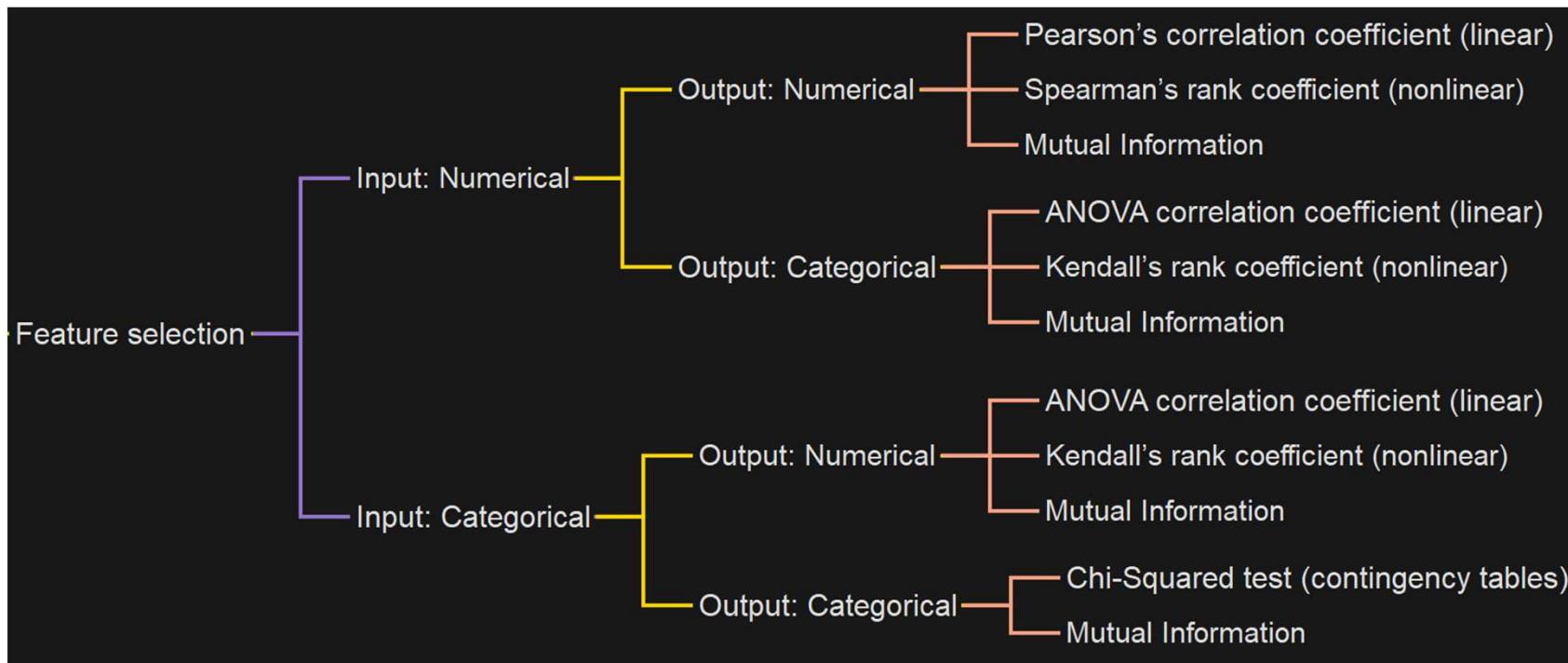
      age_year  gender  height  weight  ap_hi  ap_lo  cholesterol  gluc \
0            50       2     168    62.0    110     80           1      1
1            55       1     156    85.0    140     90           3      1
2            52       1     165    64.0    130     70           3      1
3            48       2     169    82.0    150    100           1      1
4            60       1     151    67.0    120     80           2      2
...
58702        54       1     172    70.0    130     90           1      1
58703        58       1     165    80.0    150     80           1      1
58704        53       2     168    76.0    120     80           1      1
58705        61       1     163    72.0    135     80           1      2
58706        56       1     170    72.0    120     80           2      1

      smoke  alco  active  cardio
0   False  False   True  False
1   False  False   True   True
2   False  False  False   True
3   False  False   True   True
4   False  False  False  False
...
58702  False  False   True   True
58703  False  False   True   True
58704   True  False   True  False
58705  False  False  False   True
58706  False  False   True  False

[58707 rows x 12 columns]
```

3. Lựa chọn thuộc tính và xây dựng mô hình

Preprocessing: Feature Selection & Engineering



Categorical/Numerical Input, Categorical Output: Mutual Information, RFE, SHAP

3. Lựa chọn thuộc tính và xây dựng mô hình

Preprocessing: Feature Selection & Engineering

	Prediction_contribution	Error_contribution	SHAP_Error_Rank_Valid
32	ap_hi	0.959975	-0.093749
	cholesterol	0.293781	-0.013508
	age_year	0.214346	-0.002639
	alco	0.029704	-0.000195
	gluc	0.072033	-0.000135
	smoke	0.029194	0.000402
	bmi	0.153060	0.000492
	active	0.073863	0.000540
	gender	0.030299	0.001270

SHAP:

- “Error Contribution” & “Prediction Contribution”
- “Positive, then the presence of the feature leads to an increase in the prediction error, so the feature is bad for that observation.”

	MF_Rank	RFE_Rank	SHAP_Error_Rank_Valid
31	ap_hi	1	1
	cholesterol	2	4
	bmi	3	3
	age_year	4	2
	gluc	5	5
	alco	6	9
	smoke	7	6
	active	8	8
	gender	9	9

- ➔ Dataset 31 : Top 4 theo RFE, MF (“WHO mini”)
- ➔ Dataset 32: Top 4 theo SHAP
- ➔ Dataset 33: “WHO mini”+ ‘gender’, ‘smoke’ (“WHO full”)
- ➔ Dataset 34: Đề nguyên 9 features

3. Lựa chọn thuộc tính và xây dựng mô hình

Preprocessing: Feature Selection & Engineering

➔ **Dataset 31 : Top 4 theo RFE, MF (“WHO mini”)**

X dataset: 10241 rows x 4 columns (['age_year', 'ap_hi', 'cholesterol', 'bmi'])

Y dataset: 10241 rows x 1 columns (cardio)

Test dataset: 2049 rows. Percentage of each class: {1: '60.0%', 0: '40.0%'}

Train dataset: 8192 rows. Percentage of each class: {1: '60.0%', 0: '40.0%'}

➔ **Dataset 32: Top 5 theo SHAP**

X dataset: 3343 rows x 5 columns (['age_year', 'ap_hi', 'cholesterol', 'gluc', 'alco'])

Y dataset: 3343 rows x 1 columns (cardio)

Test dataset: 669 rows. Percentage of each class: {1: '60.0%', 0: '40.0%'}

Train dataset: 2674 rows. Percentage of each class: {1: '60.0%', 0: '40.0%'}

➔ **Dataset 33: “WHO mini”+ gender, smoke (“WHO full”)**

X dataset: 16546 rows x 6 columns (['age_year', 'gender', 'ap_hi', 'cholesterol', 'smoke', 'bmi'])

Y dataset: 16546 rows x 1 columns (cardio)

Test dataset: 3310 rows. Percentage of each class: {1: '59.0%', 0: '41.0%'}

Train dataset: 13236 rows. Percentage of each class: {1: '59.0%', 0: '41.0%'}

➔ **Dataset 34: Đề nguyên 9 features**

X dataset: 26393 rows x 9 columns (['age_year', 'gender', 'ap_hi', 'cholesterol', 'gluc', 'smoke', 'alco', 'active', 'bmi'])

Y dataset: 26393 rows x 1 columns (cardio)

Test dataset: 5279 rows. Percentage of each class: {1: '58.0%', 0: '42.0%'}

Train dataset: 21114 rows. Percentage of each class: {1: '58.0%', 0: '42.0%'}

3. Lựa chọn thuộc tính và xây dựng mô hình Modeling: Model Selection

X dataset: 57142 rows x 11 columns

```
train_test_split(  
    X, y,  
    test_size=0.2,  
    random_state=1981,  
    stratify=y  
)  
  
StandardScaler()  
  
KFold Cross Validation  
    X_scaled,  
    y,  
    scoring="f1",  
    cv=KFold(n_splits=5),  
    n_jobs=-1
```

Algorithm	ascvd-train2 - Metric	ascvd-train2 - Mean	ascvd-train2 - Errors	ascvd-train2 - Rank
GradientBoostingClassifier	f1	0.731	0.008	1
LGBMClassifier	f1	0.730	0.009	2
XGBClassifier	f1	0.725	0.010	3
SVC	f1	0.723	0.007	4
LogisticRegression	f1	0.719	0.008	5
LinearSVC	f1	0.716	0.008	6
AdaBoostClassifier	f1	0.715	0.007	7
RandomForestClassifier	f1	0.710	0.008	8
SGDClassifier	f1	0.706	0.016	9
KNeighborsClassifier	f1	0.698	0.008	10
ExtraTreesClassifier	f1	0.697	0.009	11
GaussianNB	f1	0.692	0.007	12

- ➔ Hist Gradient Boosting Classifier
- ➔ LGBM Classifier
- ➔ SVC

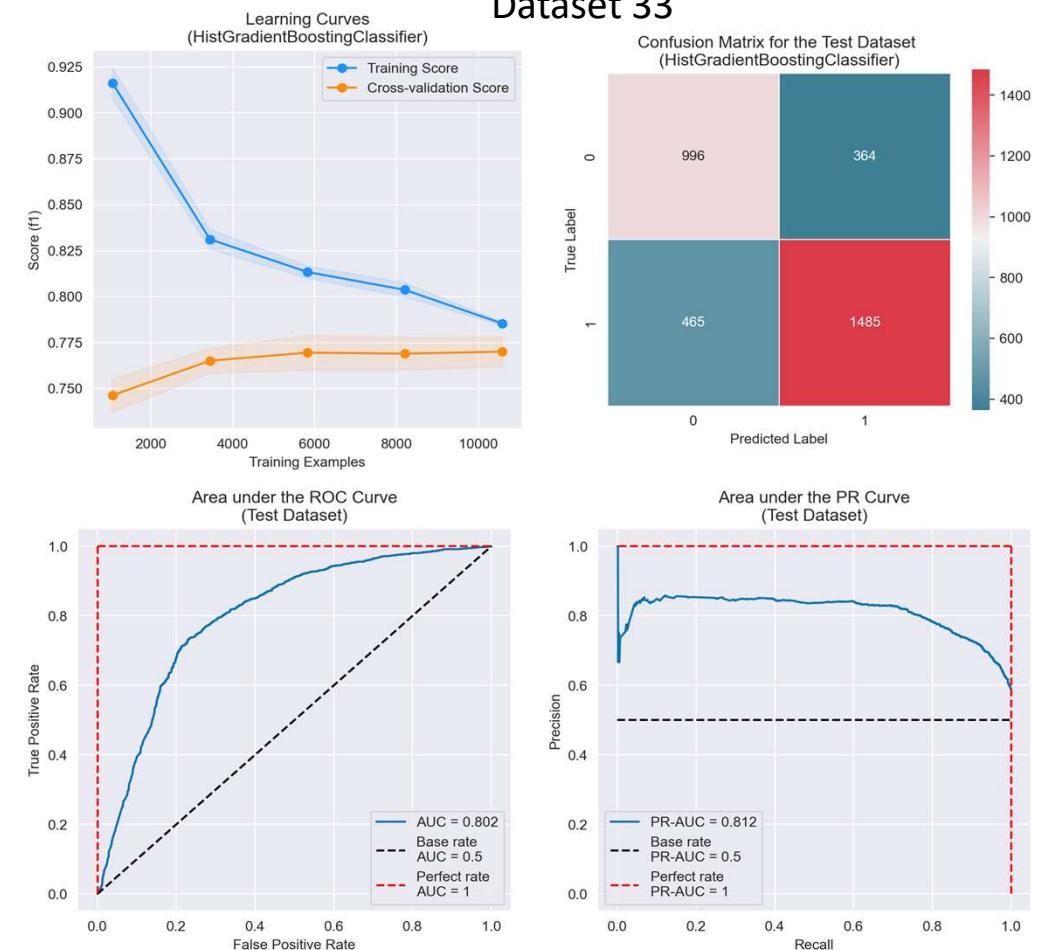
3. Lựa chọn thuộc tính và xây dựng mô hình Modeling: Model Building – Default Params

Metric	Accuracy	Precision	Recall/Sensitivity	f1-score	Specificity
Train_dataset_31	0.773	0.838	0.770	0.803	0.776
Train_dataset_32	0.810	0.852	0.823	0.837	0.789
Train_dataset_33	0.754	0.817	0.750	0.782	0.759
Train_dataset_34	0.753	0.805	0.755	0.780	0.750
Test_dataset_31	0.736	0.810	0.733	0.769	0.741
Test_dataset_32	0.726	0.784	0.746	0.764	0.697
Test_dataset_33	0.750	0.803	0.762	0.782	0.732
Test_dataset_34	0.734	0.789	0.737	0.762	0.730
Diff_dataset_31	0.037	0.028	0.037	0.034	0.035
Diff_dataset_32	0.084	0.068	0.077	0.073	0.092
Diff_dataset_33	0.004	0.014	0.012	0.000	0.027
Diff_dataset_34	0.019	0.016	0.018	0.018	0.020

➔ Dataset 33
➔ HalvingGridSearchCV(

```
factor=2,
scoring='f1',
cv=5,
verbose=False,
random_state=1981,
n_jobs=-1
```

Dataset 33

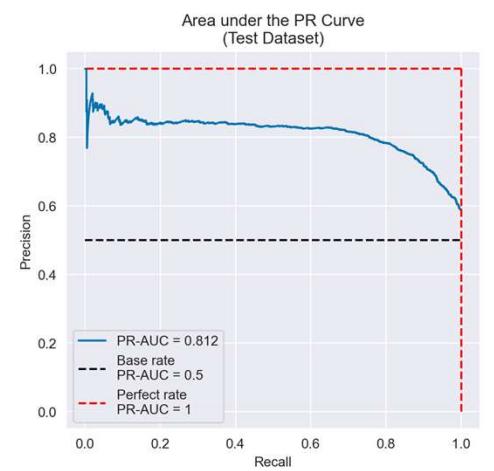
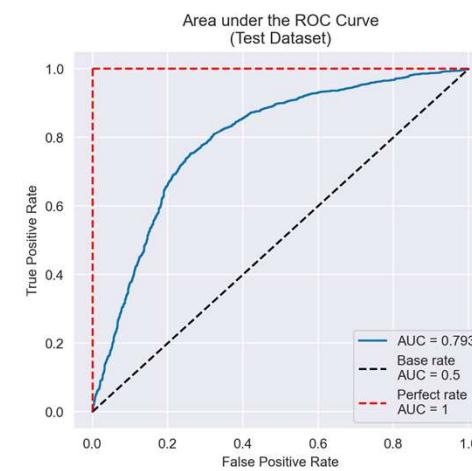
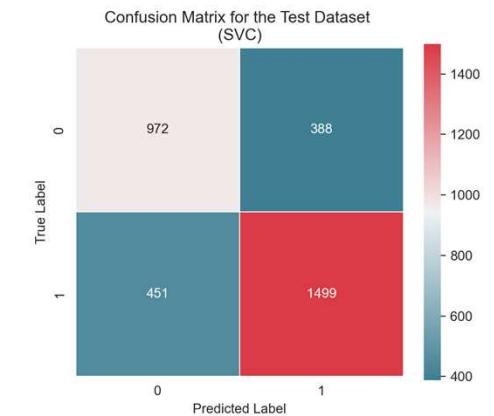
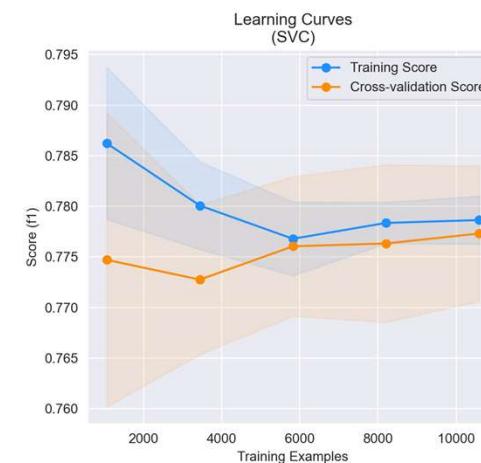


3. Lựa chọn thuộc tính và xây dựng mô hình

Improve Performance: Tuning Params on Dataset 33

```
sklsv.SVC(random_state=1981, probability=True  
          , C= 0.085  
          , gamma= 'scale'  
          , kernel= 'rbf'  
          , class_weight='balanced'  
          ),
```

SVC Performance			
	Train_dataset_33	Test_dataset_33	Diff_dataset_33
Metric			
Accuracy	0.745	0.747	-0.002
Precision	0.797	0.794	0.003
Recall/Sensitivity	0.760	0.769	0.009
f1-score	0.778	0.781	0.003
Specificity	0.723	0.715	0.008

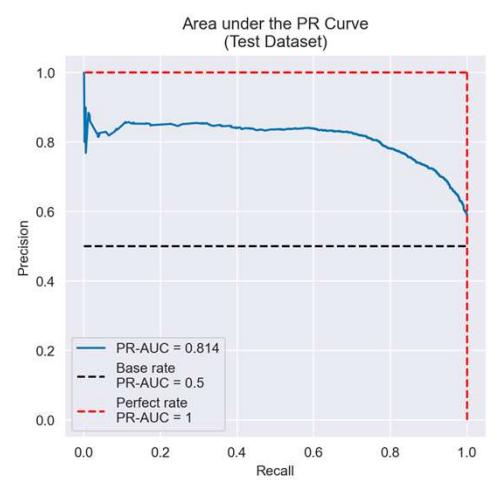
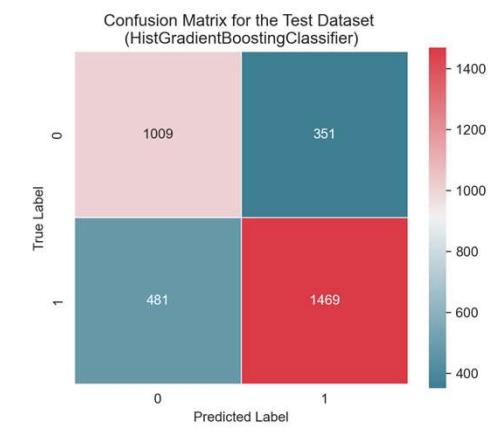
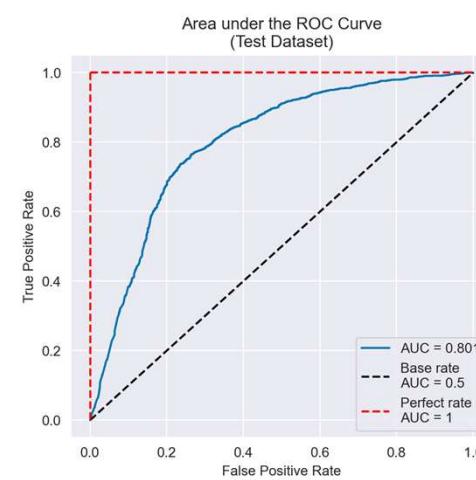
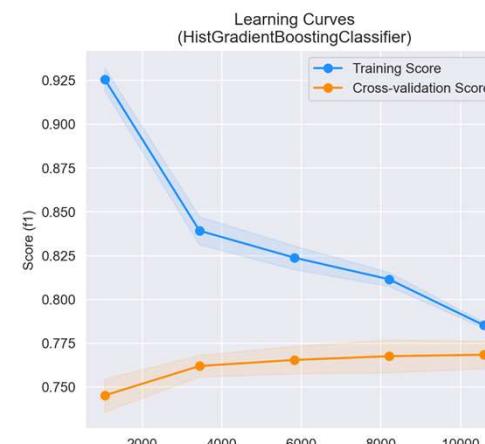


3. Lựa chọn thuộc tính và xây dựng mô hình

Improve Performance: Tuning Params on Dataset 33

```
HistGradientBoostingClassifier(  
    random_state=1981,  
    categorical_features=[  
        1, 3, 4,] # Columns nào là dạng categorial  
    , l2_regularization= 1.6  
    , learning_rate= 0.03  
    , max_depth=85  
    , max_iter= 600  
    , class_weight='balanced'  
)
```

	Train_dataset_33	Test_dataset_33	Diff_dataset_33
Metric			
Accuracy	0.751	0.749	0.002
Precision	0.818	0.807	0.011
Recall/Sensitivity	0.742	0.753	0.011
f1-score	0.778	0.779	0.001
Specificity	0.763	0.742	0.021

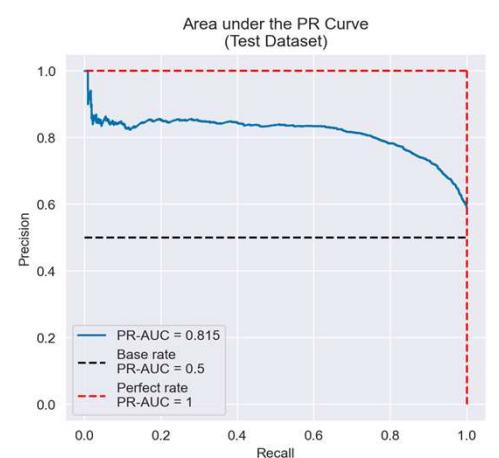
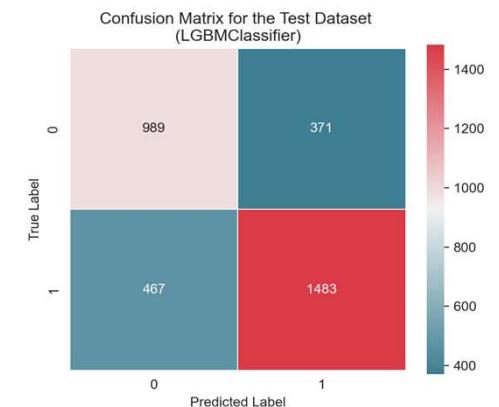
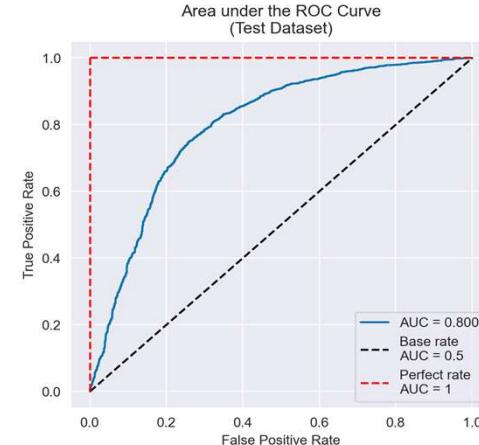
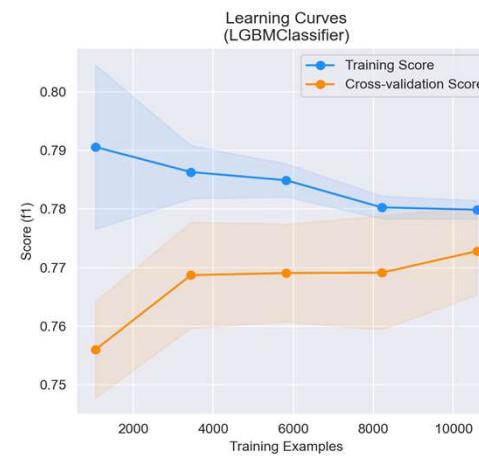


3. Lựa chọn thuộc tính và xây dựng mô hình

Improve Performance: Tuning Params on Dataset 33

```
LGBMClassifier(  
    random_state=1981,  
    learning_rate=0.05,  
    n_estimators=150,  
    num_leaves=100,  
    feature_fraction=0.5,  
    subsample=0.65,  
    max_depth=5,  
    min_child_samples=50,  
    reg_alpha=1.5,  
    reg_lambda=2,  
    class_weight='balanced'  
)
```

	Train_dataset_33	Test_dataset_33	Diff_dataset_33
Metric			
Accuracy	0.750	0.747	0.003
Precision	0.812	0.800	0.012
Recall/Sensitivity	0.749	0.761	0.012
f1-score	0.780	0.780	0.000
Specificity	0.752	0.727	0.025



3. Lựa chọn thuộc tính và xây dựng mô hình

Improve Performance: Model Evaluation on Dataset 33

Metric	Accuracy	Precision	Recall/Sensitivity	f1-score	Specificity
Model					
SVC	0.745	0.797	0.760	0.778	0.723
HGBC	0.751	0.818	0.742	0.778	0.763
LGBM	0.750	0.812	0.749	0.780	0.752

→ Dataset 33: “WHO mini”+ gender, smoke (“WHO full”)
X dataset: 16546 rows x 6 columns (['age_year', 'gender', 'ap_hi', 'cholesterol', 'smoke', 'bmi'])
Y dataset: 16546 rows x 1 columns (cardio)
Train dataset: 13236 rows. Percentage of each class: {1: '59.0%', 0: '41.0%'}

Metric	Accuracy	Precision	Recall/Sensitivity	f1-score	Specificity
Model					
SVC	0.747	0.794	0.769	0.781	0.715
HGBC	0.749	0.807	0.753	0.779	0.742
LGBM	0.747	0.800	0.761	0.780	0.727

→ Dataset 33: “WHO mini”+ gender, smoke (“WHO full”)
X dataset: 16546 rows x 6 columns (['age_year', 'gender', 'ap_hi', 'cholesterol', 'smoke', 'bmi'])
Y dataset: 16546 rows x 1 columns (cardio)
Test dataset: 3310 rows. Percentage of each class: {1: '59.0%', 0: '41.0%'}

Metric	Accuracy	Precision	Recall/Sensitivity	f1-score	Specificity
Model					
SVC	0.002	0.003	0.009	0.003	0.008
HGBC	0.002	0.011	0.011	0.001	0.021
LGBM	0.003	0.012	0.012	0.000	0.025

→ SVC ?

4. Tổng kết & Đề xuất



- # Tỷ lệ mắc bệnh tim mạch tăng theo tuổi tác (age), phụ nữ (gender) càng lớn tuổi càng có nguy cơ cao.
- # Huyết áp (ap_hi), cholesterol là những yếu tố nguy cơ quan trọng đối với bệnh tim mạch. Việc kiểm soát tốt các yếu tố này có thể giúp giảm nguy cơ mắc bệnh tim mạch.
- # Ngoài ra, cần duy trì lối sống lành mạnh, bao gồm: Không hút thuốc lá (smoke), giữ cân nặng hợp lý (BMI).



- # Model SVC có vẻ cho kết quả tốt nhất xét trên cả ba khía cạnh:
 - Hiệu quả trên tập train
 - Hiệu quả trên tập test
 - Mức độ goodfit



- # Thử nghiệm tuning thêm bằng Optuna đối với LGBM để tìm ra các tham số tốt hơn
- # Đánh giá model trên một tập test dataset độc lập

