ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC BÁCH KHOA
KHOA KHOA HỌC VÀ KỸ THUẬT MÁY TÍNH

HỌC MÁY - CO3117

Bài tập lớn 1

Chẩn đoán bệnh tim mạch

Heart Disease Prediction

Giảng viên hướng dẫn: TS. Lê Thành Sách
Lớp: TN01
Nhóm: DNA05

THÀNH PHỐ HỒ CHÍ MINH , THÁNG 9 - 2025

BÁO CÁO KẾT QUẢ LÀM VIỆC NHÓM

STT	MSSV	Họ và Tên	Nhiệm vụ phụ trách	Mức độ hoàn thành
1	2310167	Tăng Hồng Ái	Phụ trách tiền xử lý dữ liệu (xử lý thiếu, mã hóa, chuẩn hóa), tham gia viết báo cáo phần tiền xử lý	100%
2	2310510	Phạm Khánh Duy	Tìm kiếm và lựa chọn dataset phù hợp; thực hiện phân tích EDA (khám phá dữ liệu, trực quan hóa); hỗ trợ viết báo cáo	100%
3	2312506	Nguyễn Trần Yến Nhi	Tìm tham số tối ưu hoá; xây dựng và huấn luyện mô hình; đánh giá kết quả; đóng góp viết báo cáo	100%

Mục lục

1 EDA

1.1 Tổng quan dữ liệu

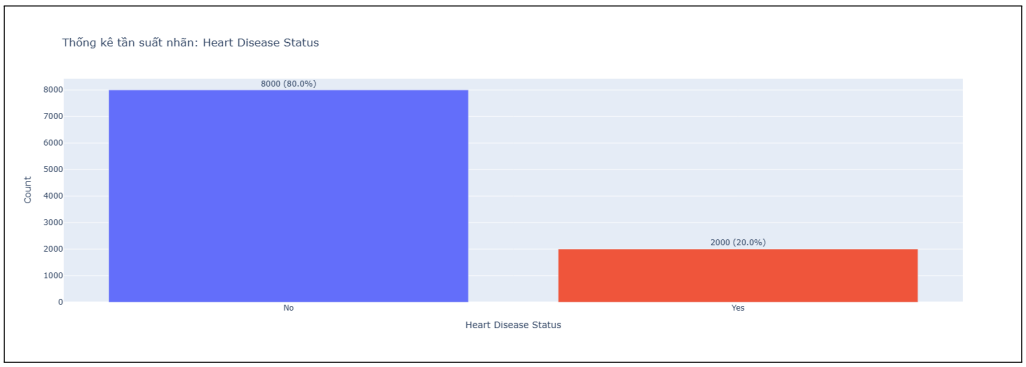
Bộ dữ liệu `heart_disease.csv` có 10.000 mẫu, mỗi mẫu đại diện cho một người. Dữ liệu ghi lại các thông tin về sức khỏe, chỉ số sinh học và thói quen sinh hoạt của mỗi người, kèm theo thông tin cho biết người đó có mắc bệnh tim mạch hay không. Dưới đây là một số thuộc tính quan trọng trong bộ dữ liệu:

- Age (float): tuổi của cá nhân.
- Gender (categorical: Male/Female): giới tính của cá nhân.
- Blood Pressure (float): chỉ số huyết áp (mmHg).
- Cholesterol Level (float): nồng độ cholesterol trong máu (mg/dL).
- Exercise Habits (categorical: High/Low): mức độ tập luyện thể dục.
- Smoking (categorical: Yes/No): tình trạng hút thuốc.
- Family Heart Disease (categorical: Yes/No): tiền sử gia đình mắc bệnh tim.
- Diabetes (categorical: Yes/No): tình trạng mắc bệnh tiểu đường.
- BMI (float): chỉ số khối cơ thể (Body Mass Index).
- High Blood Pressure (categorical: Yes/No): tình trạng tăng huyết áp.
- High LDL Cholesterol (categorical: Yes/No): mức LDL ("cholesterol xấu") cao bất thường.
- Alcohol Consumption (categorical: None/Low/Medium/High): mức độ tiêu thụ rượu bia.
- Stress Level (categorical: Low/Medium/High): mức độ căng thẳng.
- Sleep Hours (float): số giờ ngủ trung bình mỗi ngày.
- Sugar Consumption (categorical: Low/Medium/High): mức độ tiêu thụ đường.
- Triglyceride Level (float): nồng độ triglyceride (mỡ) trong máu (mg/dL).
- Fasting Blood Sugar (float): đường huyết lúc đói.
- CRP Level (float): nồng độ protein phản ứng C (C-reactive protein) – chỉ số viêm.
- Homocysteine Level (float): nồng độ axit amin homocysteine trong máu.
- Heart Disease Status (categorical: Yes/No): tình trạng bệnh tim.

Feature	Number of Missing Values
Alcohol Consumption	32
Diabetes	30
Sugar Consumption	30
Cholesterol Level	30
Age	29
Triglyceride Level	26
CRP Level	26
High LDL Cholesterol	26
High Blood Pressure	26
Low HDL Cholesterol	25
Sleep Hours	25
Exercise Habits	25
Smoking	25
Fasting Blood Sugar	22
BMI	22
Stress Level	22
Family Heart Disease	21
Homocysteine Level	20
Blood Pressure	19
Gender	19
Heart Disease Status	0

Bảng 1: Thống kê số lượng Missing Values cho từng biến.

Một số biến trong dataset có giá trị missing chiếm tỉ lệ vừa phải (5%), trong khi biến mục tiêu `Heart Disease Status` không có missing. Những giá trị thiếu này có thể được xử lý bằng phương pháp imputation ở bước tiền xử lý tiếp theo mà không ảnh hưởng đáng kể đến chất lượng dữ liệu.



Hình 1.1: Thống kê tần suất nhãn

Trong 10.000 mẫu của biến mục tiêu `Heart Disease Status`, nhãn No chiếm 8.000 mẫu, nhãn Yes chiếm 2.000 mẫu, cho thấy dữ liệu bị mất cân đối với tỷ lệ 4:1. Để cải thiện hiệu quả dự đoán cho lớp thiểu số (Yes), sẽ áp dụng kỹ thuật **SMOTE (Synthetic Minority Oversampling Technique)** để tạo mẫu tổng hợp cho tập huấn luyện, cân bằng tỷ lệ giữa hai lớp trước khi huấn luyện mô hình.

1.2 Thống kê mô tả cho Numeric Columns

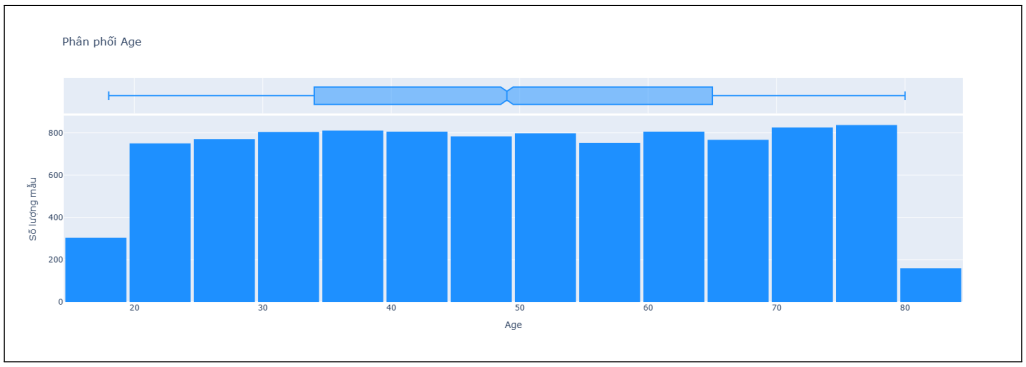
	count	mean	std	min	25%	50%	75%	max
Age	9971.000000	49.296259	18.193970	18.000000	34.000000	49.000000	65.000000	80.000000
Blood Pressure	9981.000000	149.757740	17.572969	120.000000	134.000000	150.000000	165.000000	180.000000
Cholesterol Level	9970.000000	225.425577	43.575809	150.000000	187.000000	226.000000	263.000000	300.000000
BMI	9978.000000	29.077269	6.307098	18.002837	23.658075	29.079492	34.520015	39.996954
Sleep Hours	9975.000000	6.991329	1.753195	4.000605	5.449866	7.003252	8.531577	9.999952
Triglyceride Level	9974.000000	250.734409	87.067226	100.000000	176.000000	250.000000	326.000000	400.000000
Fasting Blood Sugar	9978.000000	120.142213	23.584011	80.000000	99.000000	120.000000	141.000000	160.000000
CRP Level	9974.000000	7.472201	4.340248	0.003647	3.674126	7.472164	11.255592	14.997087
Homocysteine Level	9980.000000	12.456271	4.323426	5.000236	8.723334	12.409395	16.140564	19.999037

Hình 1.2: Thống kê Numeric Columns

Bảng thống kê cho thấy các biến số sức khỏe có phân bố đa dạng và trải dài trên nhiều khoảng giá trị.

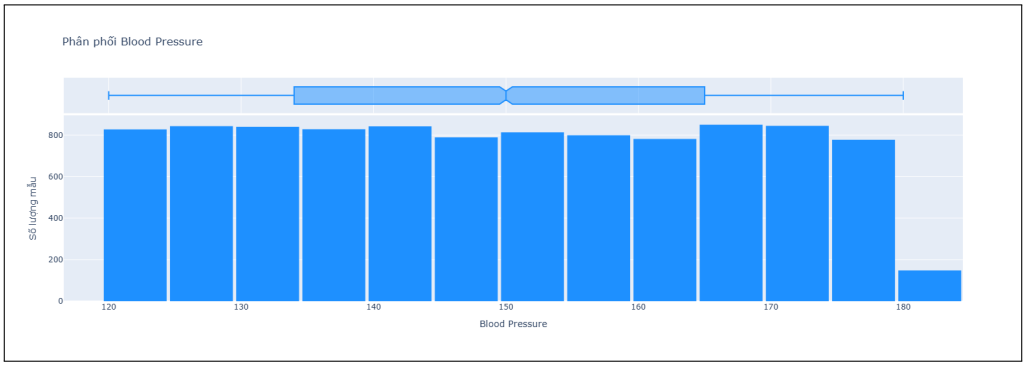
- Độ tuổi của mẫu khảo sát dao động từ 18 đến 80, tập trung nhiều nhất quanh mức trung bình gần 49 tuổi.
- Các chỉ số quan trọng như huyết áp, cholesterol, chỉ số BMI và triglyceride đều có độ lệch chuẩn khá lớn, phản ánh sự khác biệt rõ rệt giữa các cá nhân.
- Một số biến có khoảng biến thiên hẹp hơn như số giờ ngủ (dao động 4–10 giờ) và mức Homocysteine.
- Đáng chú ý, các chỉ số liên quan đến nguy cơ tim mạch (cholesterol, triglyceride, đường huyết, CRP) có giá trị tối đa cao, cho thấy sự hiện diện của các cá thể có tình trạng sức khỏe bất thường trong dữ liệu.

Nhìn chung, tập dữ liệu này phản ánh rõ sự đa dạng về đặc điểm sinh học và thói quen của người tham gia, đồng thời cung cấp nền tảng hữu ích để phân tích các yếu tố chẩn đoán bệnh tim mạch hiệu quả.



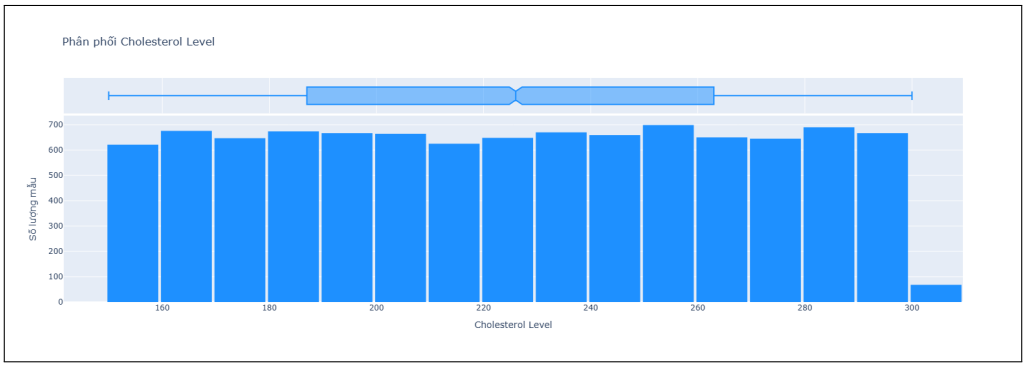
Hình 1.3: Phân phối Age

Tuổi trung bình của mẫu khảo sát là 49.3 tuổi, với phạm vi phân bố rộng từ 18 đến 80 tuổi. Điều này cho thấy dữ liệu bao quát nhiều nhóm tuổi khác nhau, từ người trẻ đến người cao tuổi, giúp phản ánh đa dạng đặc điểm của bộ dữ liệu.



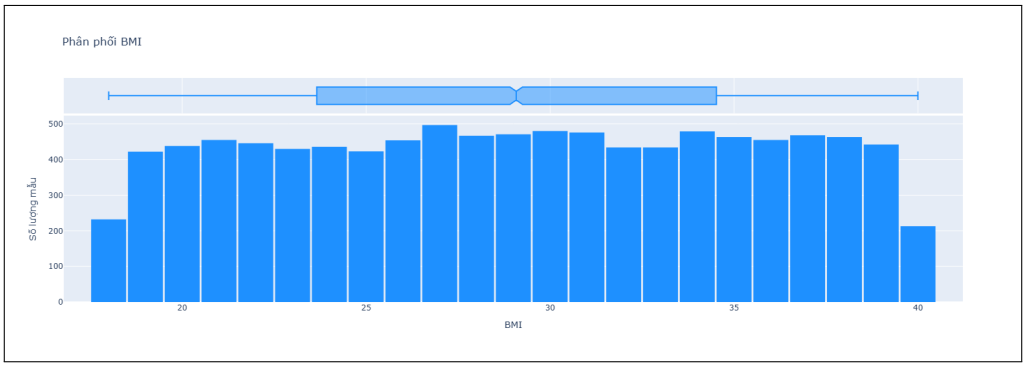
Hình 1.4: Phân phối Blood Pressure

Huyết áp trung bình 149.8 mmHg, cao hơn ngưỡng 140 mmHg, cho thấy phần lớn đối tượng thuộc nhóm tăng huyết áp độ 1. Độ lệch chuẩn 17.6 mmHg phản ánh sự phân tán đáng kể, với giá trị dao động từ 120 mmHg (cao bình thường) đến 180 mmHg (tăng huyết áp nặng). Điều này cho thấy dữ liệu khá đồng đều, cũng có sự hiện diện rõ rệt của những người mắc bệnh tăng huyết áp.



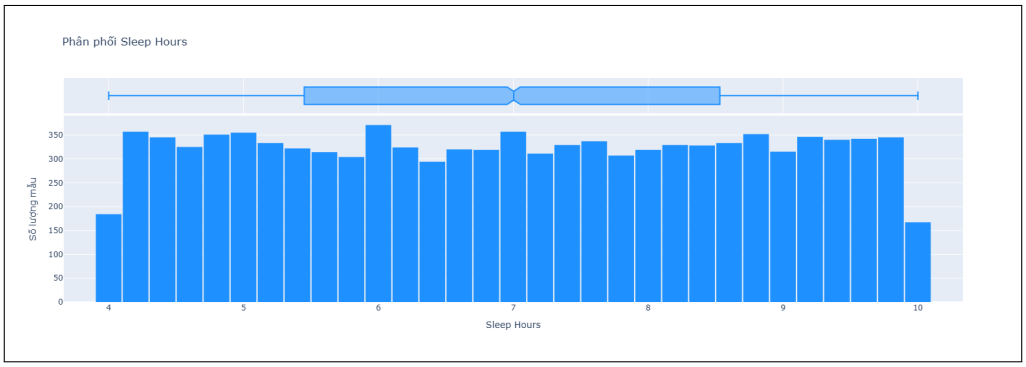
Hình 1.5: Phân phối Cholesterol Level

Trung bình cholesterol toàn phần của mẫu khảo sát là 225.4 mg/dL, cao hơn ngưỡng khuyến nghị (<200 mg/dL), cho thấy nhiều đối tượng có nguy cơ rối loạn lipid máu. Giá trị dao động từ 150 đến 300 mg/dL, với độ lệch chuẩn lớn phản ánh sự phân tán đáng kể. Điều này cho thấy dữ liệu khá đồng đều nhưng có sự hiện diện rõ rệt của những người có mức cholesterol cao bất thường.



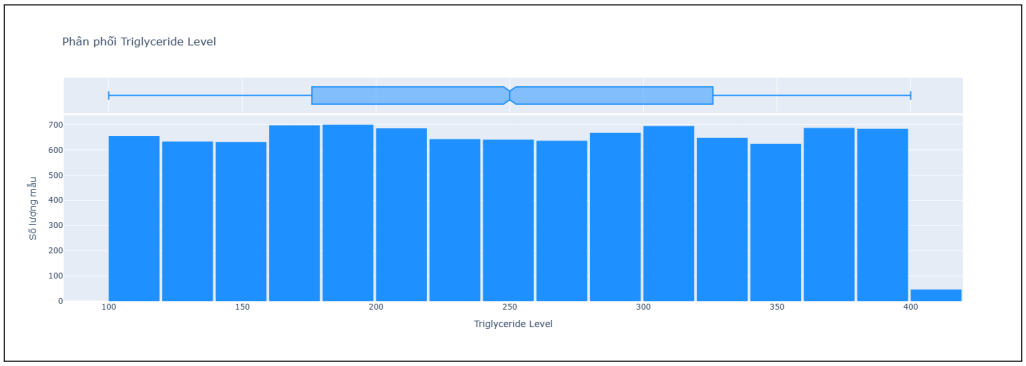
Hình 1.6: Phân phối BMI

Chỉ số khối cơ thể (BMI) trung bình của mẫu khảo sát là 29.1, cao hơn ngưỡng 25, cho thấy phần lớn đối tượng thuộc nhóm thừa cân hoặc béo phì. Giá trị dao động từ 18 (cận dưới mức bình thường) đến gần 40 (béo phì độ II), phản ánh sự hiện diện rõ rệt của những người có nguy cơ sức khỏe liên quan đến thừa cân, béo phì.



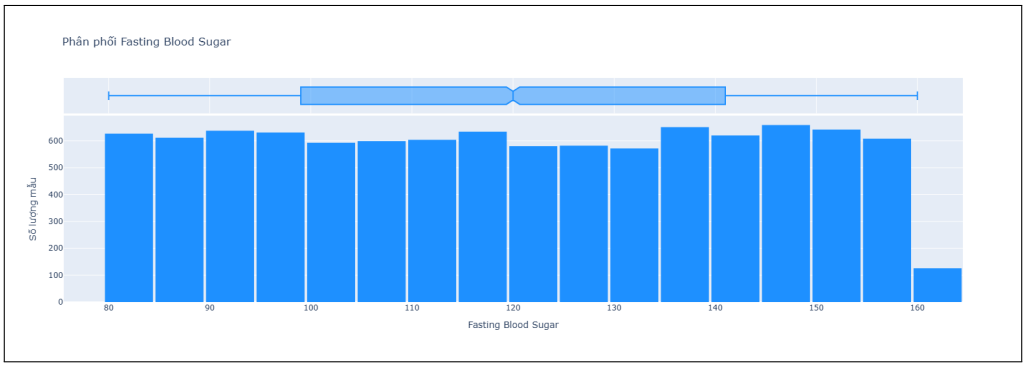
Hình 1.7: Phân phối Sleep Hours

Thời gian ngủ trung bình của mẫu khảo sát là 7.2 giờ/ngày, nằm trong khuyến nghị chung (7–8 giờ). Phần lớn giá trị dao động trong khoảng 5–9 giờ, cho thấy dữ liệu phân bố khá đồng đều và phản ánh thói quen ngủ tương đối ổn định của đa số đối tượng.



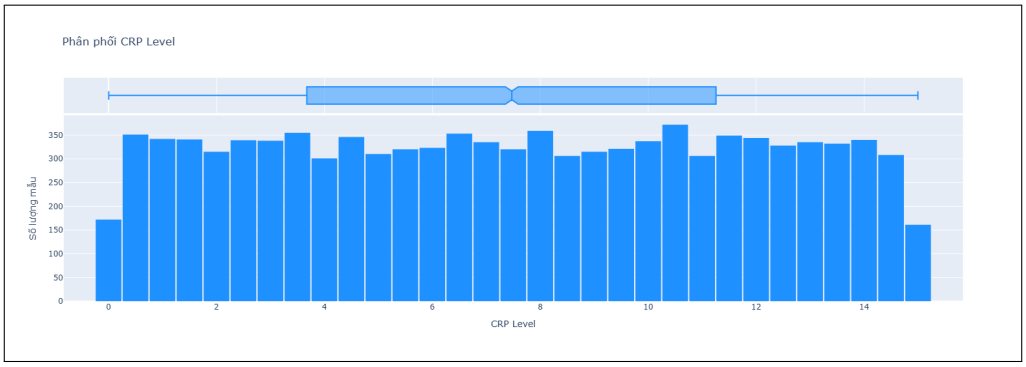
Hình 1.8: Phân phối Triglyceride Level

Trung bình của mẫu khảo sát là 250.7 mg/dL, cao hơn ngưỡng khuyến nghị thông thường, cho thấy nhiều đối tượng có mức chỉ số bất thường. Giá trị dao động rộng từ 100 đến 400 mg/dL, cùng với độ lệch chuẩn lớn, phản ánh sự biến thiên cao và sự hiện diện của những cá thể có giá trị vượt trội so với mức bình thường.



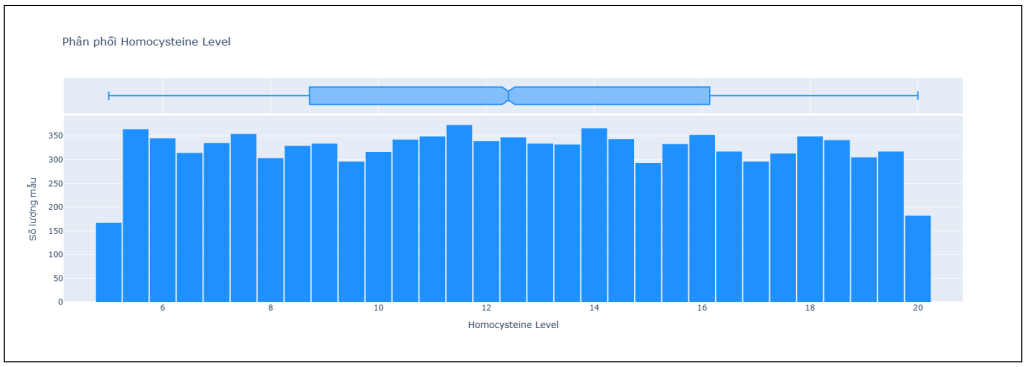
Hình 1.9: Phân phối Fasting Blood Sugar

Đường huyết trung bình của mẫu khảo sát là 120.1 mg/dL, cao hơn ngưỡng bình thường lúc đói (<100 mg/dL), cho thấy có nhiều trường hợp rơi vào nhóm tiền đái tháo đường. Giá trị dao động từ 80 đến 160 mg/dL, trong đó mức tối đa phản ánh sự hiện diện của một số đối tượng có đường huyết cao bất thường.



Hình 1.10: Phân phối CRP Level

Chỉ số CRP trung bình của mẫu khảo sát là 7.47, trong đó phần lớn giá trị ở mức thấp, phù hợp với tình trạng viêm bình thường hoặc nhẹ. Tuy nhiên, sự xuất hiện của một số giá trị cao cho thấy có những đối tượng gặp tình trạng viêm rõ rệt, phản ánh sự khác biệt về mức độ viêm trong bộ dữ liệu.



Hình 1.11: Phân phối Homocysteine Level

Chỉ số trung bình của mẫu khảo sát là 12.46, với giá trị dao động từ 5 đến gần 20. Phần lớn dữ liệu tập trung quanh mức trung bình, cho thấy phân bố khá cân đối và không có nhiều trường hợp cực đoan.

1.3 Thống kê mô tả Categorical Columns

	count	unique	top	freq
Gender	9981	2	Male	5003
Exercise Habits	9975	3	High	3372
Smoking	9975	2	Yes	5123
Family Heart Disease	9979	2	No	5004
Diabetes	9970	2	No	5018
High Blood Pressure	9974	2	Yes	5022
Low HDL Cholesterol	9975	2	Yes	5000
High LDL Cholesterol	9974	2	No	5036
Alcohol Consumption	9968	4	None	2554
Stress Level	9978	3	Medium	3387
Sugar Consumption	9970	3	Low	3390
Heart Disease Status	10000	2	No	8000

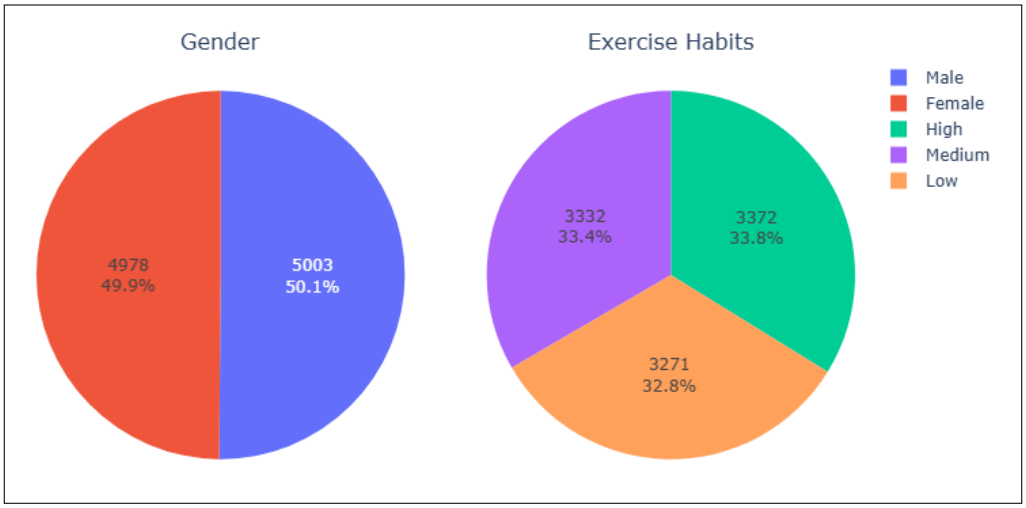
Hình 1.12: Thống kê Categorical Columns

Bảng thống kê các biến phân loại cho thấy dữ liệu được phân bố khá cân đối giữa các nhóm.

- Về giới tính, số lượng nam và nữ gần như tương đương.

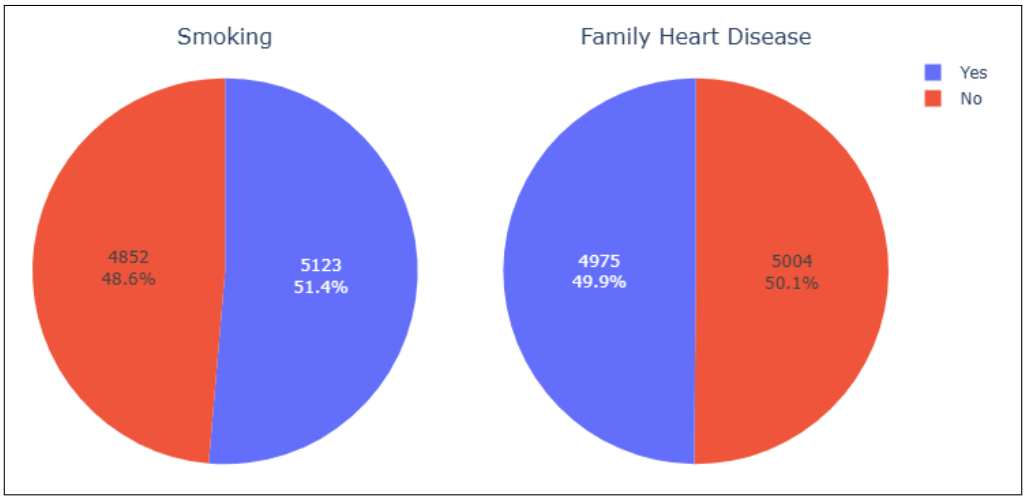
- Các yếu tố nguy cơ sức khỏe có tỷ lệ xuất hiện đáng kể, chẳng hạn như hơn một nửa số đối tượng có hút thuốc, khoảng một nửa có huyết áp cao hoặc mức HDL thấp.
- Đáng chú ý, tỷ lệ người có tiền sử bệnh tim trong gia đình hoặc mắc tiểu đường xấp xỉ 50%, cho thấy mức độ phổ biến của các bệnh lý nền.
- Đối với thói quen sinh hoạt, một bộ phận lớn đối tượng có mức vận động cao, song cũng có nhiều người tiêu thụ rượu bia và đường ở các mức độ khác nhau.
- Mức độ căng thẳng tập trung nhiều ở nhóm trung bình.

Đặc biệt, biến mục tiêu Heart Disease Status cho thấy 80% đối tượng không mắc bệnh tim, trong khi 20% còn lại có bệnh, phản ánh sự mất cân bằng trong bộ dữ liệu.



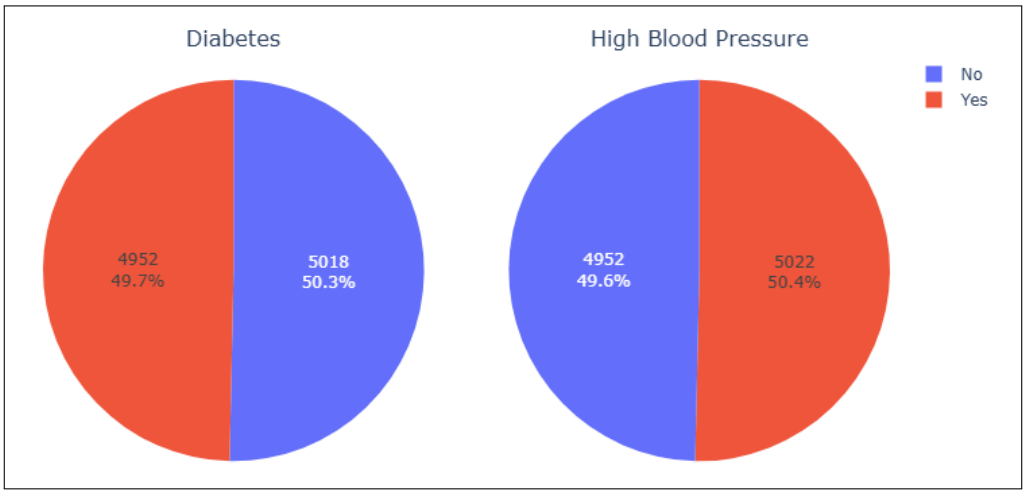
Hình 1.13: Cơ cấu Gender, Exercise Habits

Mẫu khảo sát có sự phân bố giới tính cân bằng, với số nam (5003) và nữ gần tương đương. Phần lớn đối tượng có thói quen tập luyện ở mức cao (3372 người), cho thấy hoạt động thể chất khá phổ biến.



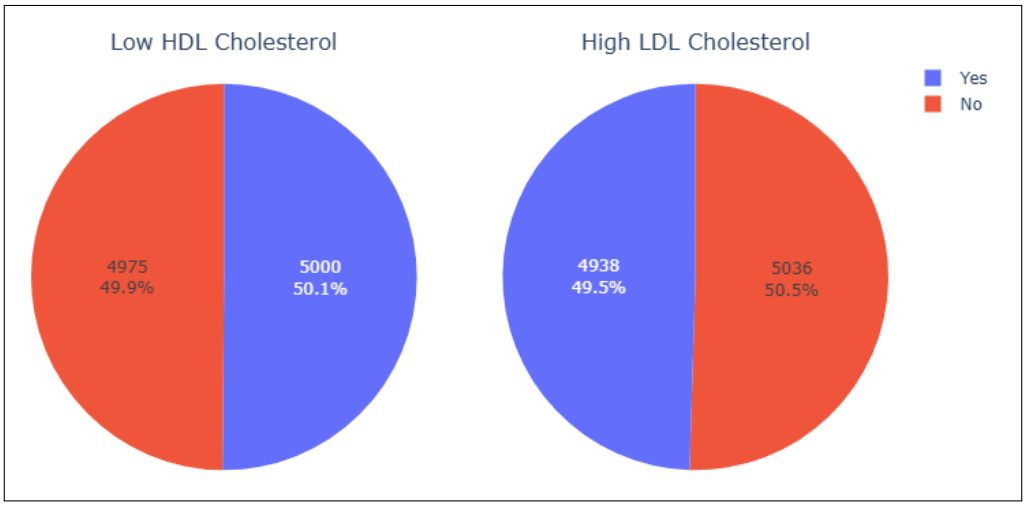
Hình 1.14: Cơ cấu Smoking, Family Heart Disease

Có hơn một nửa số người tham gia (5123 người) có hút thuốc, phản ánh tỷ lệ hút thuốc đồng đều trong bộ dữ liệu. Khoảng 50% đối tượng (5004 người) không có tiền sử bệnh tim trong gia đình, cho thấy sự phân bố gần như đồng đều giữa hai nhóm.



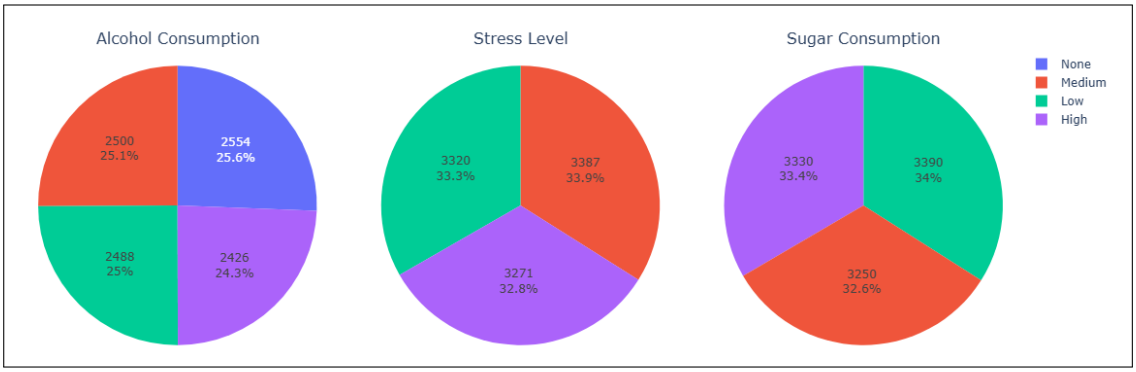
Hình 1.15: Cơ cấu Diabetes, High Blood Pressure

Phần lớn đối tượng không mắc tiểu đường (5018 người), tuy nhiên tỷ lệ mắc bệnh vẫn ở mức đáng kể. Khoảng một nửa số người tham gia (5022 người) được ghi nhận có huyết áp cao, tỷ lệ khá tương đồng ở hai nhóm dữ liệu.



Hình 1.16: Cơ cấu Low HDL Cholesterol, High LDL Cholesterol

Gần một nửa mẫu khảo sát (5000 người) có mức HDL thấp, phản ánh nguy cơ rối loạn lipid máu phổ biến. Số lượng đối tượng có LDL cao (4938 người) và bình thường gần như cân bằng, thể hiện sự đa dạng trong tình trạng mỡ máu.



Hình 1.17: Cơ cấu Alcohol Consumption, Stress Level, Sugar Consumption

Có sự phân bố đa dạng với 4 mức độ tiêu thụ, trong đó nhóm "không uống rượu bia" chiếm tỷ lệ lớn nhất (2554 người). Mức độ căng thẳng tập trung nhiều ở nhóm trung bình (3387 người), cho thấy phần lớn đối tượng duy trì mức stress vừa phải. Thói quen tiêu thụ đường phân bố khá đồng đều ở các mức, trong đó mức thấp chiếm tỷ lệ cao nhất (3390 người).

1.4 Tổng kết về tập dữ liệu

Tập dữ liệu bao gồm cả biến số định lượng và định tính, phản ánh đặc điểm sinh học, lối sống và tình trạng bệnh lý của đối tượng.

- Các chỉ số numeric như huyết áp, cholesterol, đường huyết, triglyceride và BMI đều cao hơn ngưỡng khuyến nghị, với độ lệch chuẩn lớn cho thấy sự biến thiên rõ rệt. Trong khi đó, thời gian ngủ và homocysteine phân bố đồng đều hơn.
- Ở nhóm categorical, dữ liệu phân bố cân đối giữa giới tính, bệnh nền và thói quen sinh hoạt, với tỷ lệ hút thuốc, rối loạn mỡ máu và huyết áp cao khá đáng kể.

Nhìn chung, các biến trong tập dữ liệu thể hiện sự đa dạng và phân bố khá đồng đều, phản ánh đầy đủ các đặc trưng sức khỏe và lối sống của đối tượng khảo sát. Tuy nhiên, biến mục tiêu lại có sự mất cân bằng đáng kể, tạo ra thách thức trong quá trình xử lý dữ liệu và huấn luyện mô hình dự đoán.

2 Tiền xử lý dữ liệu

2.1 Mã hóa dữ liệu (Encoding)

Dataset có nhiều feature dạng **categorical**, gồm các biến có thứ tự như `Exercise Habits`, `Stress Level`, `Sugar Consumption` và `Alcohol Consumption`, các biến nhị phân như `Smoking`, `Family Heart Disease`, `Diabetes`, `High Blood Pressure`, `Low HDL Cholesterol`, `High LDL Cholesterol` và biến `Gender`. Để mô hình có thể xử lý được các feature này, các bước mã hóa được thực hiện như sau:

1. **Định nghĩa thứ tự cho các feature ordinal**
   - Các biến `Exercise Habits`, `Stress Level`, `Sugar Consumption` có 3 mức: Low < Medium < High
   - Biến `Alcohol Consumption` có 4 mức: None < Low < Medium < High
   - Biến `Gender` có 2 mức: Male, Female
   - Các biến còn lại thuộc dạng nhị phân: No < Yes
2. **Áp dụng Ordinal Encoding cho các feature categorical**
   - Các giá trị của từng feature được chuyển thành số nguyên theo thứ tự đã định.
   - Các giá trị missing được bỏ qua trong bước này để không làm sai lệch dữ liệu.
3. **Mã hóa biến mục tiêu**
   - Biến `Heart Disease Status` được mã hóa thành 0 và 1, phục vụ cho các mô hình phân loại nhị phân.

Nhận xét: Mã hóa ordinal giúp giữ thông tin về thứ tự của các mức độ, thay vì chỉ chuyển thành các số ngẫu nhiên. Các feature nhị phân cũng được chuyển thành số để mô hình học máy có thể nhận dạng, đồng thời việc tách riêng mã hóa biến mục tiêu đảm bảo không ảnh hưởng đến các feature input. Kết quả là tất cả các biến **categorical** trong dataset đều được chuyển thành dạng numeric, sẵn sàng cho các bước tiền xử lý tiếp theo.

	Age	Gender	Blood Pressure	Cholesterol Level	Exercise Habits	Smoking	Family Heart Disease	Diabetes	BMI	High Blood Pressure	...	High LDL Cholesterol	Alcohol Consumption	Stress Level	Sleep Hours	Sugar Consumption	Triglyceride Level	Fasting Blood Sugar	CRP Level	Homocysteine Level	Heart Disease Status
0	56.0	0.0	153.0	155.0	2.0	1.0	1.0	0.0	24.991591	1.0	...	0.0	3.0	1.0	7.633228	1.0	342.0	NaN	12.969246	12.387250	0
1	69.0	1.0	146.0	286.0	2.0	0.0	1.0	1.0	25.221799	0.0	...	0.0	2.0	2.0	8.744034	1.0	133.0	157.0	9.355389	19.298875	0
2	46.0	0.0	126.0	216.0	0.0	0.0	0.0	0.0	29.855447	0.0	...	1.0	1.0	0.0	4.440440	0.0	393.0	92.0	12.709873	11.230926	0
3	32.0	1.0	122.0	293.0	2.0	1.0	1.0	0.0	24.130477	1.0	...	1.0	1.0	2.0	5.249405	2.0	293.0	94.0	12.509046	5.961958	0
4	60.0	0.0	166.0	242.0	0.0	1.0	1.0	1.0	20.486289	1.0	...	0.0	1.0	2.0	7.030971	2.0	263.0	154.0	10.381259	8.153887	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
9995	25.0	1.0	136.0	243.0	1.0	1.0	0.0	0.0	18.788791	1.0	...	1.0	2.0	2.0	6.834954	1.0	343.0	133.0	3.588814	19.132004	1
9996	38.0	0.0	172.0	154.0	1.0	0.0	0.0	0.0	31.856801	1.0	...	1.0	0.0	2.0	8.247784	0.0	377.0	83.0	2.658267	9.715709	1
9997	73.0	0.0	152.0	201.0	2.0	1.0	0.0	1.0	26.899911	0.0	...	1.0	0.0	0.0	4.436762	0.0	248.0	88.0	4.408867	9.492429	1
9998	23.0	0.0	142.0	299.0	0.0	1.0	0.0	1.0	34.964026	1.0	...	1.0	2.0	2.0	8.526329	1.0	113.0	153.0	7.215634	11.873486	1
9999	38.0	1.0	128.0	193.0	1.0	1.0	1.0	1.0	25.111295	0.0	...	1.0	3.0	1.0	5.659394	2.0	121.0	149.0	14.387810	6.208531	1

10000 rows × 21 columns

Hình 2.1: Thông tin dữ liệu sau khi encoding

2.2 Chia tập dữ liệu (Splitting)

Sau khi hoàn tất bước mã hóa biến categorical, dataset được tách thành **tập đặc trưng (X)** và **biến mục tiêu (y)**. Dữ liệu được chia thành **tập huấn luyện (train)** và **tập kiểm tra (test)** với tỷ lệ **80/20** (có thể tùy chỉnh tỉ lệ khác như 70/30, 60/40). Đồng thời, phương pháp **stratified split** được áp dụng, giữ nguyên tỷ lệ các nhãn trong biến mục tiêu ở cả tập train và test.

Nhận xét: Chia tập dữ liệu được thực hiện trước các bước tiền xử lý như xử lý **missing value**, **chuẩn hóa (scaling)** và **cân bằng dữ liệu (imbalance handling)**. Nhờ đó, tập test hoàn toàn mới, chưa bị ảnh hưởng bởi bất kỳ bước xử lý nào trên tập train, giúp đánh giá hiệu quả mô hình trên dữ liệu chưa thấy trước đó, đảm bảo không xảy ra data leakage.

2.3 Xử lý missing value với KNN Imputer

Để điền các giá trị missing trong dataset, nhóm áp dụng phương pháp **KNN Imputer**, sử dụng giá trị trung bình từ các k hàng lân cận gần nhất (có thể tùy chỉnh biến k). Trong bài toán này, k được chọn bằng 5. Các bước thực hiện bao gồm:

1. **Fit imputer trên tập huấn luyện (train):**
   Imputer được huấn luyện dựa trên các giá trị có sẵn của tập train để học mối quan hệ giữa các feature, sau đó các giá trị missing trong tập train được điền dựa trên trung bình của 5 hàng lân cận gần nhất.

2. **Áp dụng imputer cho tập kiểm tra (test):**
   Tập test được điền missing bằng cách sử dụng thông tin học được từ tập train, đảm bảo dữ liệu không đưa được từ test vào train.

3. **Ép các biến categorical về kiểu số nguyên:**
   Vì KNN Imputer trả về giá trị float, các biến categorical được làm tròn và chuyển về dạng integer để phù hợp với dữ liệu đã được mã hóa.

4. **Kiểm tra lại missing value:**
   Sau khi imputation, tập train và test không còn missing value, đảm bảo đã được lấp đầy để huấn luyện.

Nhận xét: Sử dụng KNN Imputer giúp giữ lại thông tin từ các feature liên quan, thay vì chỉ dùng median hoặc mean đơn giản. Việc tách riêng fit trên train và transform trên test đảm bảo tính khách quan và tránh rò rỉ dữ liệu.

2.4 Chuẩn hóa dữ liệu (Scaling)

Để các feature có cùng thang đo và giúp mô hình học máy hội tụ nhanh hơn, **Min-Max Scaling** được áp dụng với khoảng giá trị chuẩn hóa trong [a,b] (có thể tùy chỉnh). Trong bài toán này, **Min-Max Scaling** được áp dụng với khoảng giá trị [0,1]. Bao gồm các bước thực hiện:

1. **Fit và transform trên tập huấn luyện (train):**
   Bộ scaler được học dựa trên tập train, tính toán giá trị min và max của từng feature. Sau đó, tất cả các giá trị trong tập train được chuyển về khoảng [0, 1].

2. **Transform trên tập kiểm tra (test):**
   Tập test được chuẩn hóa dựa trên giá trị min và max học được từ tập train, đảm bảo không rò rỉ thông tin từ test vào train.

Nhận xét: Chuẩn hóa Min-Max giúp các feature có cùng thang đo, cải thiện hiệu suất và tốc độ huấn luyện của mô hình. Việc tách riêng fit trên train và transform trên test đảm bảo **tính khách quan** khi đánh giá mô hình.

2.5 Xử lý mất cân bằng dữ liệu với SMOTE

Dữ liệu biến mục tiêu ban đầu bị **imbalance** với tỷ lệ lớp **No:Yes = 4:1**. Để cải thiện hiệu quả dự đoán cho lớp thiểu số (**Yes**), phương pháp **SMOTE (Synthetic Minority Oversampling Technique)** được áp dụng chỉ trên **tập huấn luyện**. Bao gồm các bước thực hiện:

1. **Chọn sampling strategy và số lượng lân cận:**
   Tỷ lệ oversampling được chọn là 0.5 (có thể tùy chỉnh). Nghĩa là sau khi SMOTE, số lượng mẫu của lớp thiểu số bằng một nửa số mẫu lớp đa số. Sử dụng 5 neighbors (có thể tùy chỉnh số lượng neighbors) để tạo các mẫu tổng hợp mới dựa trên các điểm gần nhất.

2. **Fit và resample trên tập huấn luyện:**
   SMOTE học từ tập train và tạo các mẫu tổng hợp cho lớp thiểu số, giúp cân bằng dữ liệu. Tập train mới (`X_train_res`, `y_train_res`) có tỷ lệ lớp cân bằng hơn **No:Yes = 2:1**.

Nhận xét: Áp dụng SMOTE chỉ trên tập train đảm bảo **tính khách quan** của tập test. Dữ liệu cân bằng giúp mô hình học tốt hơn các lớp thiểu số, cải thiện **recall** và **F1-score** cho việc chẩn đoán người mắc bệnh tim.

3 Trích xuất và lựa chọn đặc trưng với PCA

Dataset sử dụng trong bài toán là **chẩn đoán mắc bệnh tim mạch**, chứa **20 features** với các thông tin liên quan đến lối sống, chỉ số sinh học và tiền sử bệnh. Do số lượng feature khá lớn và có thể có tương quan cao giữa các biến, việc áp dụng **Principal Component Analysis (PCA)** là cần thiết để giảm chiều dữ liệu và giữ lại các thông tin quan trọng. Bao gồm các bước thực hiện:

1. **Chuẩn hóa dữ liệu trước PCA:**
   Tất cả các feature đã được đưa về cùng thang đo [0, 1] nhờ Min-Max Scaling ở phần tiền xử lý, đảm bảo mỗi feature đóng góp tương đương vào PCA.

2. **Áp dụng PCA:**
   PCA được huấn luyện trên tập train để tìm các thành phần chính. Thay vì chỉ giữ một tỷ lệ phương sai cố định, các thử nghiệm được thực hiện với ba mức độ khác nhau: giữ 99%, 90%, 80%, 70% tổng phương sai. Việc này cho phép so sánh hiệu quả của các mô hình khi sử dụng số lượng components khác nhau, đồng thời đánh giá sự mất mát thông tin khi giảm chiều dữ liệu.

3. **Lựa chọn đặc trưng:**
   Các principal components được giữ lại ở từng mức phương sai sẽ là input cho các mô hình học máy, thay cho tập feature gốc. Việc này giúp giảm **overfitting**, tăng tốc độ huấn luyện và loại bỏ các feature có tương quan cao.

Nhận xét: Sử dụng PCA giúp giảm chiều dữ liệu một cách hiệu quả, giữ lại các thông tin quan trọng nhất. Trong dataset chẩn đoán tim mạch với nhiều feature, PCA là cần thiết để xử lý dữ liệu có tương quan cao. Việc thử nghiệm các mức phương sai khác nhau (99%, 90%, 80%, 70%) cho phép lựa chọn số lượng components tối ưu, cân bằng giữa việc giữ lại thông tin và giảm chiều dữ liệu. Kết hợp với các bước tiền xử lý như scaling và imputation, PCA tạo điều kiện cho các mô hình học máy học nhanh và chính xác hơn. Mặc dù PCA làm mất ý nghĩa trực tiếp của các feature gốc, lợi ích về hiệu suất và giảm noise thường lớn hơn nhược điểm này.

4 Học máy

4.1 Tập train - test

Dữ liệu tập huấn luyện sau các bước tiền xử lý được minh họa như bảng dưới đây, trong khi dữ liệu tập kiểm tra có dạng tương tự.

	Age	Gender	Blood Pressure	Cholesterol Level	Exercise Habits	Smoking	Family Heart Disease	Diabetes	BMI	High Blood Pressure
0	0.016129	0.0	0.916667	0.813333	1.0	0.0	1.0	0.0	0.494539	1.0
1	0.064516	1.0	0.050000	0.133333	1.0	0.0	0.0	0.0	0.557115	0.0
2	0.435484	1.0	0.633333	0.146667	0.5	1.0	1.0	1.0	0.725765	1.0
3	0.177419	0.0	0.383333	0.946667	0.0	1.0	0.0	1.0	0.162000	0.0
4	0.451613	1.0	0.250000	0.493333	0.0	0.0	1.0	0.0	0.466873	0.0

Low HDL Cholesterol	High LDL Cholesterol	Alcohol Consumption	Stress Level	Sleep Hours	Sugar Consumption	Triglyceride Level	Fasting Blood Sugar	CRP Level	Homocysteine Level
1.0	0.0	0.333333	0.0	0.889402	1.0	0.336667	0.9875	0.388340	0.864419
1.0	1.0	0.666667	1.0	0.765187	0.0	0.783333	0.5250	0.213807	0.177501
0.0	1.0	0.000000	0.5	0.824415	0.0	0.923333	0.3625	0.886494	0.982724
1.0	0.0	1.000000	1.0	0.975233	0.5	0.900000	0.7700	0.045874	0.516004
0.0	1.0	1.000000	1.0	0.660830	1.0	0.836667	0.8625	0.727305	0.879933

4.2 Huấn luyện và tìm tham số tối ưu cho từng mô hình

Trong bài toán học máy, việc lựa chọn *hyperparameters* phù hợp cho từng mô hình là bước quan trọng để cải thiện hiệu suất và khả năng tổng quát hóa của mô hình. Để tìm tham số tối ưu, sử dụng **GridSearchCV** từ thư viện `scikit-learn`. Bao gồm các bước thực hiện:

1. **Xác định mô hình và tham số cần tối ưu:**
    - Random Forest: số lượng cây (`n_estimators`), độ sâu tối đa của từng cây (`max_depth`), số lượng mẫu tối thiểu để tách nút (`min_samples_split`) và trọng số lớp (`class_weight`) để xử lý imbalance.
    - k-NN: số lượng lân cận (`n_neighbors`), kiểu trọng số (`weights`) áp dụng cho các điểm lân cận, khoảng cách Minkowski (`p`) và loại metric (`metric`) để đo khoảng cách giữa các điểm.
    - SVM: hệ số điều chuẩn (`C`) để cân bằng margin và sai số, loại `kernel` (ví dụ linear, rbf, poly) cùng với trọng số lớp (`class_weight`).
    - Decision Tree: tiêu chí tách nút (`criterion`), độ sâu tối đa (`max_depth`), số mẫu tối thiểu để tách nút (`min_samples_split`) và ở lá (`min_samples_leaf`), trọng số lớp (`class_weight`) và hệ số cắt tỉa (`ccp_alpha`).

- Naive Bayes: giá trị làm mượt phương sai (`var_smoothing`).

2. **Xây dựng lưới tham số (parameter grid):** Các giá trị thử nghiệm được xác định trước cho từng tham số và GridSearchCV sẽ thử tất cả tổ hợp khả thi.

3. **Sử dụng cross-validation để đánh giá hiệu suất:** Mỗi tổ hợp tham số được đánh giá bằng *K-fold cross-validation* (có thể tùy chọn k). Với dataset này, sử dụng k = 5 để thực hiện. Metric đánh giá có thể là `accuracy`, `f1`, `recall`,... tùy loại bài toán. Với dataset này, sử dụng Recall đánh giá để phù hợp với việc chẩn đoán bệnh tim.

4. **Chọn tham số tối ưu:** GridSearchCV trả về tổ hợp tham số cho hiệu suất tốt nhất trên tập validation và sử dụng tham số đó để đánh giá mô hình trên tập test.

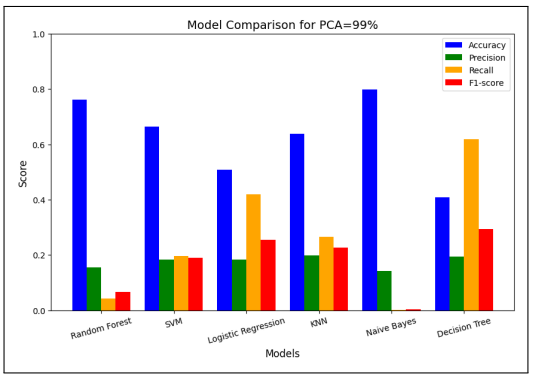
Sau khi huấn luyện và tìm tham số cho từng mô hình, bảng tổng kết tham số như sau:

Mô hình	**Tham số**	**Giá trị**
Random Forest	max_depth	20
	min_samples_split	2
	n_estimators	200
	class_weight	balanced
SVM	C	10
	kernel	rbf
	class_weight	balanced
Logistic Regression	C	0.1
	penalty	l1
	solver	liblinear
	class_weight	balanced
KNN	n_neighbors	3
	metric	minkowski
	p	1
	weights	distance
Naive Bayes	var_smoothing	1e-09
Decision Tree	criterion	gini
	max_depth	5
	min_samples_split	2
	ccp_alpha	0.0
	class_weight	balanced

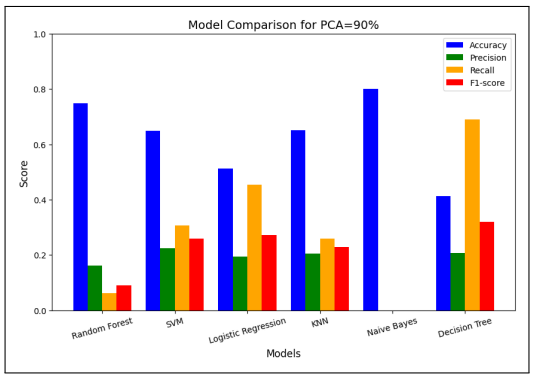
Bảng 2: Tham số tối ưu của các mô hình

5 Đánh giá mô hình

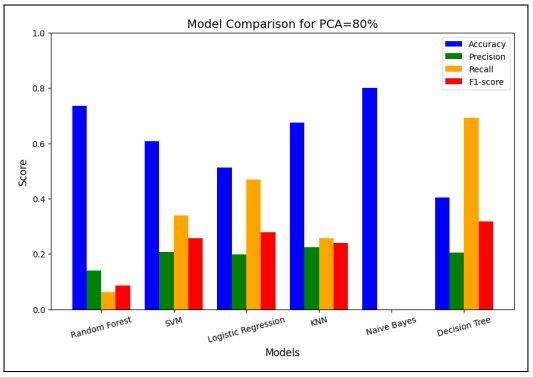
Sử dụng các tham số tối ưu của từng mô hình, tiến hành kiểm tra với các giá trị PCA đã được lựa chọn và đánh giá hiệu quả mô hình trên tập test dựa vào các chỉ số Accuracy, Precision, Recall và F1-score. Kết quả mô hình theo từng chỉ số với các PCA = 99%, 90%, 80%, 70% được trình bày dưới đây.



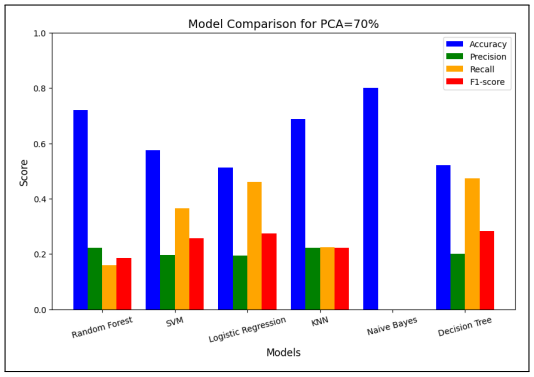
Hình 5.1: PCA = 99%



Hình 5.2: PCA = 90%



Hình 5.3: PCA = 80%



Hình 5.4: PCA = 70%

Nhận xét:

- Nhìn chung, khi giảm số thành phần PCA, hầu hết các mô hình như Random Forest, SVM, Logistic Regression và KNN đều cải thiện về Precision, Recall và F1-score, trong khi Accuracy gần như không thay đổi. Điều này cho thấy mặc dù dữ liệu bị mất cân bằng, nhưng khi tập trung vào các đặc trưng quan trọng, mô hình có khả năng học tập hiệu quả hơn.
- Với mô hình Naive Bayes, Accuracy đạt mức cao, nhưng các chỉ số Precision, Recall và F1-score lại rất thấp, bất kể có điều chỉnh PCA hay không. Điều này phản ánh rằng mô

hình không khai thác được đặc trưng dữ liệu, mà chủ yếu dựa vào việc dự đoán nhãn No (chiếm tỷ lệ lớn trong tập dữ liệu) để đạt Accuracy cao.

- Đối với mô hình Decision Tree, mặc dù không vượt trội về các chỉ số khác, nhưng Recall đạt mức nổi bật khoảng 69%, thể hiện khả năng học tập nhất định trong nhiệm vụ chẩn đoán bệnh tim mạch.

Lưu ý: Trong y tế, Recall là chỉ số đặc biệt quan trọng, vì bỏ sót bệnh (False Negative) có thể gây hậu quả nghiêm trọng nếu bệnh nhân không được chẩn đoán và điều trị kịp thời. Do đó, kết quả Recall của mô hình Decision Tree được xem là phù hợp với mục tiêu phân tích dữ liệu bệnh tim.

6 Kết luận

Kết quả mô hình có hiệu suất cao nhất theo từng chỉ số được trình bày trong bảng dưới đây.

=== Best Model per Metric ===

Metric	Best Model	Best PCA	Best Score
Accuracy	Naive Bayes	70	0.800000
Precision	KNN	80	0.225383
Recall	Decision Tree	80	0.692500
F1-score	Decision Tree	90	0.319444

Nhận xét: Kết quả cho thấy từng mô hình có ưu thế ở các chỉ số khác nhau.

- Naive Bayes đạt Accuracy cao nhất khi sử dụng PCA = 70%, tuy nhiên Precision, Recall và F1-score đều không nổi bật, phản ánh xu hướng dự đoán lệch về lớp chiếm đa số.
- KNN đạt Precision cao nhất khi sử dụng PCA = 80%, nhưng Recall và F1-score còn thấp.
- Mô hình Decision Tree thể hiện ưu thế rõ rệt về Recall và F1-score. Đặc biệt quan trọng trong bối cảnh y tế khi mục tiêu chính là hạn chế bỏ sót ca bệnh.

Như vậy, Decision Tree được xem là mô hình phù hợp nhất cho bài toán chẩn đoán bệnh tim trong tập dữ liệu này.

7 Phụ lục

Dataset: Heart Disease Dataset
Github: DNA05's github page
Colab Notebook: DNA05-BTL1

Tài liệu tham khảo

[1] Hossain Hedayati. Heart disease prediction with 83.8% accuracy. https://www.kaggle.com/code/hossainhedayati/heart-disease-prediction-with-83-8-accuracy/notebook, 2023. Accessed: 2025-09-18.

[2] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning: With Applications in R and Python*. Springer, 2nd edition, 2021.

[3] Kevin P. Murphy. *Machine Learning: A Probabilistic Perspective*. MIT Press, 2012.

[4] Scikit-learn developers. Classification metrics — scikit-learn documentation. https://scikit-learn.org/stable/modules/model_evaluation.html, 2025. Accessed: 2025-09-18.

[5] Scikit-learn developers. Preprocessing data — scikit-learn documentation. https://scikit-learn.org/stable/modules/preprocessing.html, 2025. Accessed: 2025-09-18.