Cộng hòa xã hội chủ nghĩa Việt Nam

Độc lập - Tự do - Hạnh phúc

BÁO CÁO FINAL PROJECT

Đề tài: Phân tích dữ liệu bệnh tim mạch

Người hướng dẫn : Trịnh Bá Tú

Môn học : Python for Machine Learning

Nhóm : 3

Sinh viên thực hiện : Tô Minh Đức

Nguyễn Minh Đăng

Nguyễn Trường Giang

Nguyễn Văn Thi

Nguyễn Thị Thu Trang

Hà Nội, 2025

Mục lục

Danh sách hình vẽ

Danh sách bảng

1. Mô tả dữ liệu

1.1. Giới thiệu đề tài

Bệnh tim mạch từ lâu đã được ghi nhận là nguyên nhân hàng đầu gây tử vong trên toàn cầu, chiếm tỷ lệ lớn trong tổng số ca tử vong mỗi năm, theo Tổ chức Y tế Thế giới (WHO). Không chỉ gây ảnh hưởng nghiêm trọng đến sức khỏe cá nhân, bệnh còn kéo theo hệ lụy về chi phí y tế, năng suất lao động và chất lượng sống của cộng đồng. Trong bối cảnh đó, việc phát hiện sớm các yếu tố nguy cơ là yếu tố then chốt trong công tác phòng ngừa và điều trị hiệu quả bệnh lý tim mạch.

Đề tài này tập trung vào việc phân tích một tập dữ liệu thực tế về bệnh tim mạch, với mục tiêu hiểu rõ hơn về mối quan hệ giữa các chỉ số sức khỏe cá nhân (như tuổi tác, huyết áp, cân nặng, mức cholesterol, thói quen sinh hoạt, v.v.) và nguy cơ mắc bệnh tim. Thông qua việc áp dụng các kỹ thuật phân tích dữ liệu (EDA), trực quan hóa, nhóm thực hiện đề tài không chỉ khám phá được các yếu tố quan trọng ảnh hưởng đến tình trạng tim mạch, mà còn định hướng rõ ràng cho việc ứng dụng các mô hình học máy sau này – nhằm xây dựng hệ thống phân loại nguy cơ mắc bệnh một cách chính xác và hiệu quả hơn.

1.2. Đặc điểm của bộ dữ liệu

Bộ dữ liệu được sử dụng trong đề tài có nguồn gốc từ nền tảng Kaggle, với tên gọi "Cardiovascular Disease Dataset". Đây là một tập dữ liệu lớn, bao gồm 70.000 bản ghi đại diện cho các bệnh nhân đã được theo dõi về tình trạng sức khỏe tim mạch. Mỗi bản ghi chứa thông tin của một cá nhân với 13 thuộc tính, trong đó có cả thuộc tính dạng số (numerical) như tuổi, chiều cao, cân nặng, huyết áp,... và thuộc tính phân loại (categorical) như giới tính, tình trạng hút thuốc, mức độ hoạt động thể chất, mức cholesterol, glucose và nhãn bệnh (cardio).

Tập dữ liệu này phù hợp để khai thác trong các bài toán phân tích dữ liệu y sinh học và đặc biệt hữu ích trong việc xây dựng các mô hình dự đoán nguy cơ mắc bệnh tim mạch dựa trên hồ sơ sức khỏe cá nhân.

1.3. Định hướng thực hiện

- **Phân tích dữ liệu tổng quan:** Bắt đầu bằng việc khám phá dữ liệu thô nhằm hiểu rõ cấu trúc tập dữ liệu, kiểm tra sự đầy đủ và hợp lệ của các trường thông tin. Đồng thời xác định các biến dạng số (numerical) và phân loại (categorical) để có cách tiếp cận phân tích phù hợp.

- **Thống kê mô tả và trực quan hóa:** Áp dụng các phương pháp thống kê mô tả để nắm bắt đặc điểm phân bố của từng thuộc tính như độ lệch chuẩn, giá trị trung bình, trung vị, v.v. Bên cạnh đó, sử dụng các kỹ thuật trực quan hóa dữ liệu (biểu đồ histogram, boxplot, heatmap, scatter plot...) để phát hiện xu hướng, mẫu hình hoặc sự bất thường tiềm ẩn trong dữ liệu.
- **Phân tích mối quan hệ giữa các yếu tố:** Đánh giá mối liên hệ giữa các dữ liệu dạng số như tuổi, huyết áp (ap_hi, ap_lo), cân nặng và chiều cao với các dữ liệu dạng danh mục như giới tính, mức độ cholesterol, glucose trong máu và các thói quen sinh hoạt hút thuốc (smoke), uống rượu (alco), hoạt động thể chất (active).
- **Phân tích mối quan hệ giữa các yếu tố sức khỏe và bệnh tim:** Đánh giá mức độ liên hệ giữa các chỉ số sinh học như tuổi tác (age), huyết áp (ap_hi, ap_lo), cân nặng, glucose, cholesterol, thói quen hút thuốc (smoke), uống rượu (alco), hoạt động thể chất (active) với khả năng mắc bệnh tim (cardio).
- **Nhận xét và kết luận:** Dựa vào các phân tích từ đó tóm tắt các phát hiện, mối liên hệ và những đặc trưng quan trọng của dữ liệu có thể ảnh hưởng lớn đến mô hình học máy.
- **Định hướng phát triển:** Nêu bật những bài học rút ra trong việc phân tích dữ liệu y tế và định hướng phát triển mô hình dự đoán nguy cơ bệnh tim trong các bước tiếp theo.

2. Phân tích khám phá dữ liệu (EDA)

2.1. Ý nghĩa các cột dữ liệu

Trước khi tiến hành phân tích sâu, việc tìm hiểu tổng quan về cấu trúc và đặc điểm của bộ dữ liệu là bước quan trọng nhằm đảm bảo tính chính xác và hiệu quả trong các bước xử lý tiếp theo. Phần này sẽ trình bày về số lượng bản ghi, số lượng thuộc tính, kiểu dữ liệu, phân loại biến (số hay phân loại), cũng như kiểm tra chất lượng dữ liệu ban đầu như giá trị thiếu (null) và trùng lặp (duplicate). Qua đó, nhóm thực hiện có cái nhìn tổng thể về dữ liệu đầu vào, từ đó xây dựng chiến lược phân tích phù hợp cho đề tài.

	id	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco	active	cardio
0	0	18393	2	168	62.0	110	80	1	1	0	0	1	0
1	1	20228	1	156	85.0	140	90	3	1	0	0	1	1
2	2	18857	1	165	64.0	130	70	3	1	0	0	0	1
3	3	17623	2	169	82.0	150	100	1	1	0	0	1	1
4	4	17474	1	156	56.0	100	60	1	1	0	0	0	0

Hình 1: Tổng quan về giá trị các cột của dataset

Bảng 1: Ý nghĩa các cột trong bộ dữ liệu

Cột	**Ý nghĩa**
`age`	Tuổi bệnh nhân (đơn vị ngày)
`gender`	Giới tính (Nam/Nữ)
`height`	Chiều cao (cm)
`weight`	Cân nặng (kg)
`ap_hi`	Huyết áp tâm thu (mmHg)
`ap_lo`	Huyết áp tâm trương (mmHg)
`cholesterol`	Mức cholesterol trong máu (bình thường / cao / rất cao)
`gluc`	Mức đường huyết (bình thường / cao / rất cao)
`smoke`	Có hút thuốc không (có/không)
`alco`	Có uống rượu không (có/không)
`active`	Có hoạt động thể chất không (có/không)
`cardio`	Nhãn: có/không mắc bệnh tim mạch

Dựa trên dữ liệu trong tập dataset, ta có thể chia các thuộc tính thành hai nhóm chính: dạng số (numerical) và dạng phân loại (categorical). Cụ thể, các cột thuộc nhóm numerical bao gồm: id, age, height, weight, ap_hi và ap_lo. Trong đó, cột id là định danh duy nhất cho mỗi bệnh nhân và không mang ý nghĩa phân tích thống kê, do đó thường được loại bỏ trong các bước xử lý tiếp theo. Các cột còn lại trong nhóm này đại diện cho các chỉ số sức khỏe có giá trị liên tục hoặc rời rạc và sẽ được sử dụng trong phân tích mô tả và trực quan hóa.

Nhóm các cột categorical gồm các biến phân loại như: gender, cholesterol, gluc, smoke, alco, active. Các biến này phần lớn phản ánh các yếu tố nhân khẩu học, thói quen sinh hoạt hoặc kết quả chẩn đoán, thường được mã hóa dưới dạng nhãn rời rạc.

Biến cardio là biến mục tiêu (target), biểu thị việc một bệnh nhân có mắc bệnh tim mạch hay không. Những cột này sẽ được sử dụng chủ yếu trong phân tích tần suất, kiểm

định thống kê và làm biến đầu vào cho mô hình phân loại trong các bước tiếp theo.

Đáng chú ý, toàn bộ tập dữ liệu gồm 70.000 dòng không chứa bất kỳ giá trị thiếu (null) hay bản ghi trùng lặp (duplicate) nào, cho thấy chất lượng dữ liệu đầu vào khá tốt và không cần thực hiện các bước xử lý thiếu dữ liệu hay loại bỏ trùng lặp. Điều này giúp tiết kiệm đáng kể công sức trong khâu tiền xử lý và cho phép tập trung vào việc phân tích và xây dựng mô hình.

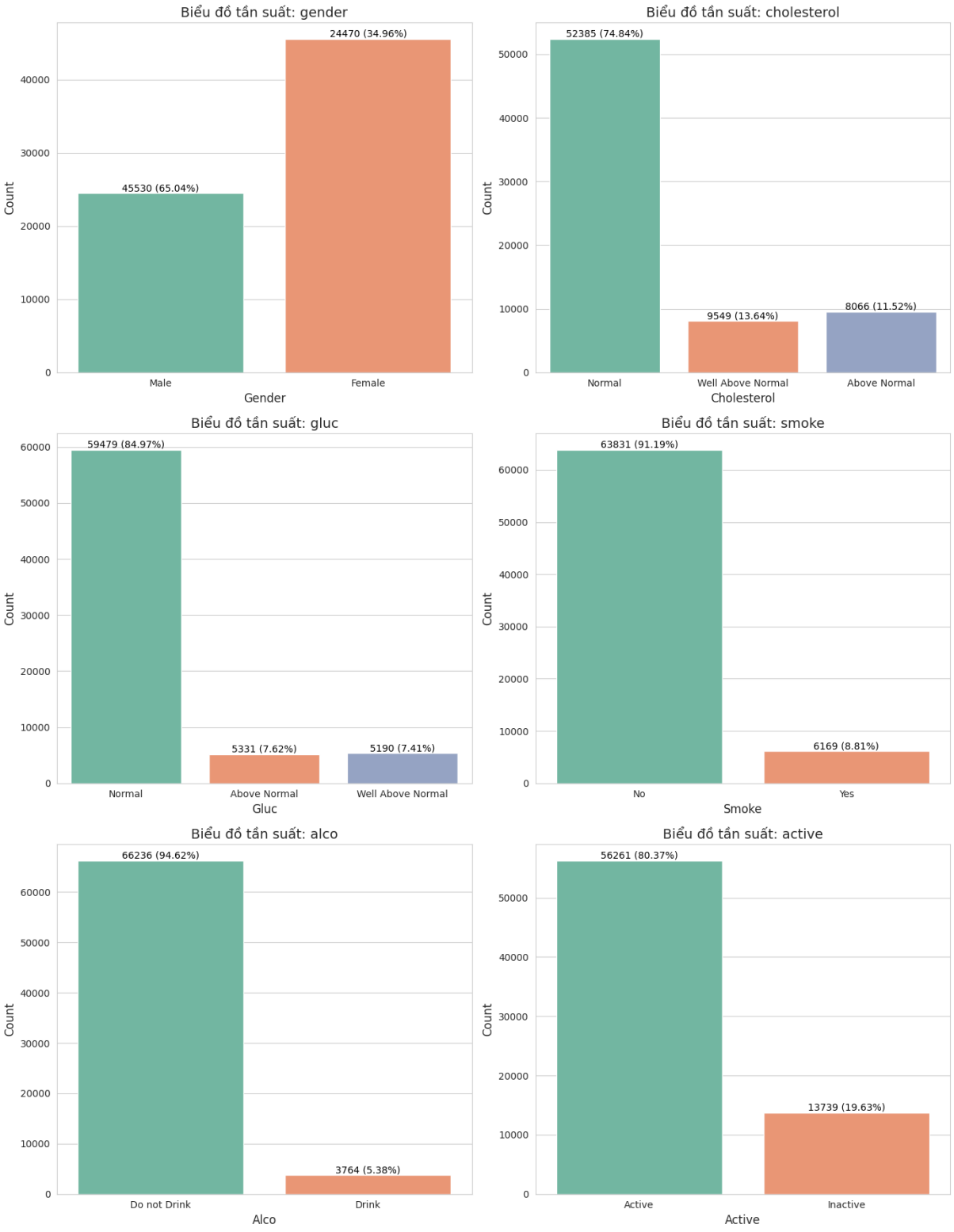
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 70000 entries, 0 to 69999
Data columns (total 13 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   id           70000 non-null  int64
 1   age          70000 non-null  int64
 2   gender       70000 non-null  int64
 3   height       70000 non-null  int64
 4   weight       70000 non-null  float64
 5   ap_hi        70000 non-null  int64
 6   ap_lo        70000 non-null  int64
 7   cholesterol  70000 non-null  int64
 8   gluc         70000 non-null  int64
 9   smoke        70000 non-null  int64
 10  alco         70000 non-null  int64
 11  active       70000 non-null  int64
 12  cardio       70000 non-null  int64
dtypes: float64(1), int64(12)
memory usage: 6.9 MB
```

Hình 2: Mô tả kiểu dữ liệu của các biến

2.2. Phân tích các thuộc tính phân loại (Categorical)

Trong tập dữ liệu, phân bố giới tính không đồng đều khi nữ giới chiếm khoảng 65%, trong khi nam giới chỉ chiếm 35%. Điều này phản ánh sự mất cân bằng giới tính rõ rệt ở đầu vào, có thể ảnh hưởng đến độ chính xác của các phân tích nếu không được xử lý phù hợp. Về chỉ số cholesterol, phần lớn bệnh nhân (75%) có mức cholesterol bình thường,

tuy nhiên có tới 25% rơi vào nhóm có mức cholesterol cao hoặc rất cao – đây là nhóm cần đặc biệt chú ý vì có nguy cơ cao mắc bệnh tim mạch. Tương tự, phần lớn bệnh nhân (85%) có chỉ số glucose trong máu ở mức bình thường, còn lại 15% thuộc nhóm trên mức bình thường và rất cao, cho thấy một tỷ lệ nhỏ nhưng đáng quan tâm về nguy cơ rối loạn chuyển hóa và tiểu đường.



Hình 3: Biểu đồ tần suất của các biến Categorical

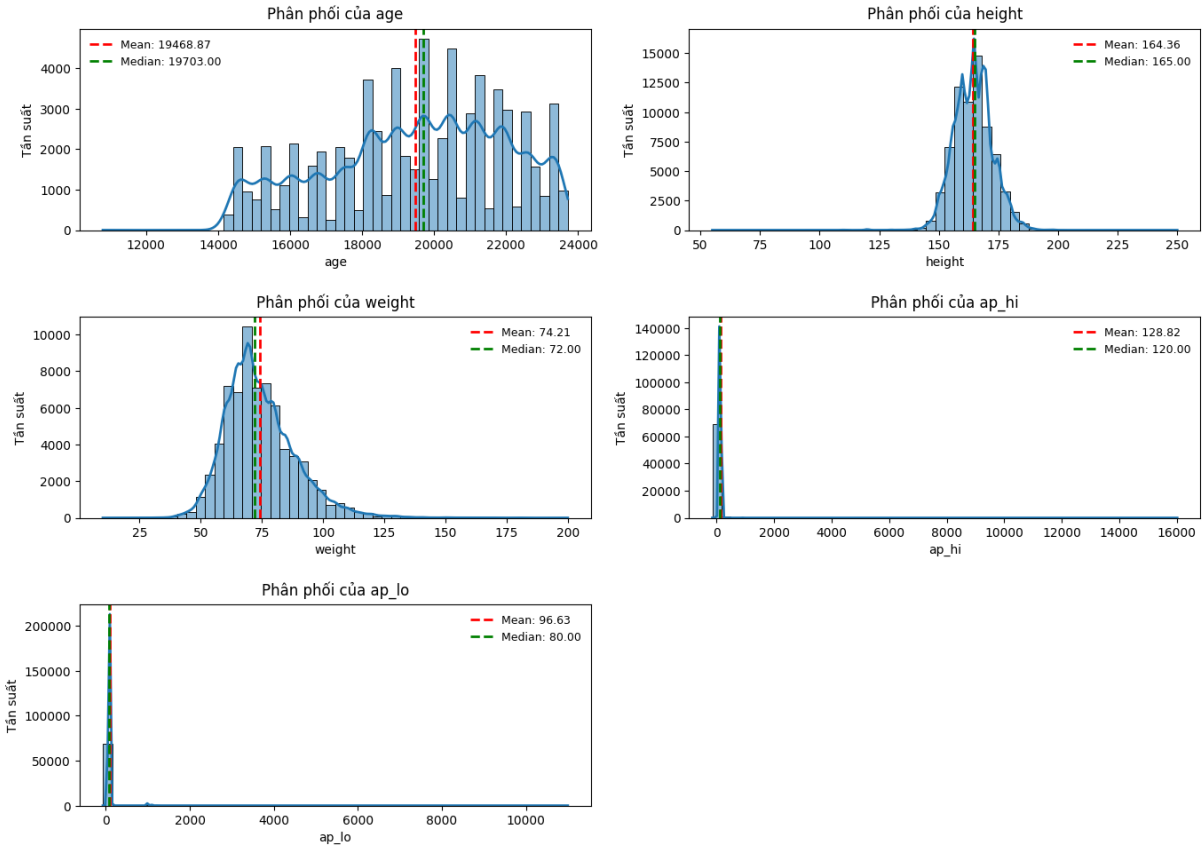
Tỷ lệ bệnh nhân hút thuốc và uống rượu trong tập dữ liệu tương đối thấp, lần lượt chỉ khoảng 9% và 5%. Con số này thấp hơn đáng kể so với thống kê ngoài cộng đồng, điều này có thể do thông tin chưa được khai báo đầy đủ hoặc dữ liệu chưa mang tính đại diện cao. Trong khi đó, có đến 80% bệnh nhân cho biết họ thường xuyên vận động thể chất, phản ánh xu hướng sống lành mạnh trong dữ liệu. Tuy vậy, vẫn còn khoảng 20% người thiếu vận động – đây là nhóm cần được lưu ý trong các phân tích liên quan đến nguy cơ tim mạch do lối sống ít vận động gây ra.

2.3. Phân tích biến số (Numerical)

Bảng thống kê mô tả các biến số (sử dụng hàm describe()) cho thấy dữ liệu gồm 70.000 bản ghi và không có giá trị thiếu đối với bất kỳ biến nào. Các đặc trưng số liên tục bao gồm: age, height, weight, ap_hi và ap_lo, được phân tích dưới đây dựa trên các giá trị thống kê như trung bình (mean), trung vị (median), độ lệch chuẩn (std), giá trị nhỏ nhất (min) và lớn nhất (max).

	id	age	height	weight	ap_hi	ap_lo
count	70000.000000	70000.000000	70000.000000	70000.000000	70000.000000	70000.000000
mean	49972.419900	19468.865814	164.359229	74.205690	128.817286	96.630414
std	28851.302323	2467.251667	8.210126	14.395757	154.011419	188.472530
min	0.000000	10798.000000	55.000000	10.000000	-150.000000	-70.000000
25%	25006.750000	17664.000000	159.000000	65.000000	120.000000	80.000000
50%	50001.500000	19703.000000	165.000000	72.000000	120.000000	80.000000
75%	74889.250000	21327.000000	170.000000	82.000000	140.000000	90.000000
max	99999.000000	23713.000000	250.000000	200.000000	16020.000000	11000.000000

Hình 4: Bảng thống kê mô tả biến số



Hình 5: Phân phối của age theo các thuộc tính số

Đối với biến age, đại diện cho tuổi bệnh nhân (tính bằng ngày), giá trị trung bình là 19.468 ngày (tương đương khoảng 53 tuổi), trong khi trung vị là 19.703 ngày. Hai giá trị này khá gần nhau, cho thấy phân phối của tuổi tương đối cân đối, chỉ hơi lệch trái nhẹ. Tuổi tối thiểu ghi nhận trong tập dữ liệu là khoảng 10.798 ngày (29 tuổi) và tối đa là 23.713 ngày (65 tuổi). Không có dấu hiệu rõ rệt về sự xuất hiện của outlier trong biến này và độ phân tán là hợp lý, tập trung chủ yếu vào nhóm tuổi trung niên đến cao tuổi.

Biến height (chiều cao) có giá trị trung bình khoảng 164.36 cm và trung vị 165 cm, gần như trùng khớp, cho thấy phân phối dữ liệu xấp xỉ chuẩn. Tuy nhiên vẫn tồn tại một số giá trị ngoại lệ ở cả hai phía, với chiều cao tối thiểu là 55 cm và tối đa lên đến 250 cm – đều là những con số không hợp lý nếu xét trong bối cảnh người trưởng thành, có thể xuất phát từ lỗi nhập liệu. Phần lớn bệnh nhân có chiều cao nằm trong khoảng 150–180 cm, là khoảng phổ biến ở người trưởng thành và hoàn toàn hợp lý về mặt sinh lý học.

Với biến weight (cân nặng), dữ liệu có phân phối lệch phải nhẹ, thể hiện qua việc trung bình là 74.21 kg, trong khi trung vị thấp hơn một chút – 72 kg. Đa số bệnh nhân có cân nặng nằm trong khoảng 50–100 kg, tuy nhiên vẫn có một số điểm outlier rõ rệt, bao gồm những bệnh nhân có cân nặng dưới 30 kg hoặc vượt quá 200 kg – những trường hợp rất hiếm gặp, thậm chí có thể là lỗi nhập liệu (ví dụ, nhập sai đơn vị hoặc thiếu số). Biểu đồ phân phối có phần đuôi kéo dài về bên phải, làm tăng độ lệch và ảnh hưởng đến

phân tích nếu không xử lý phù hợp.

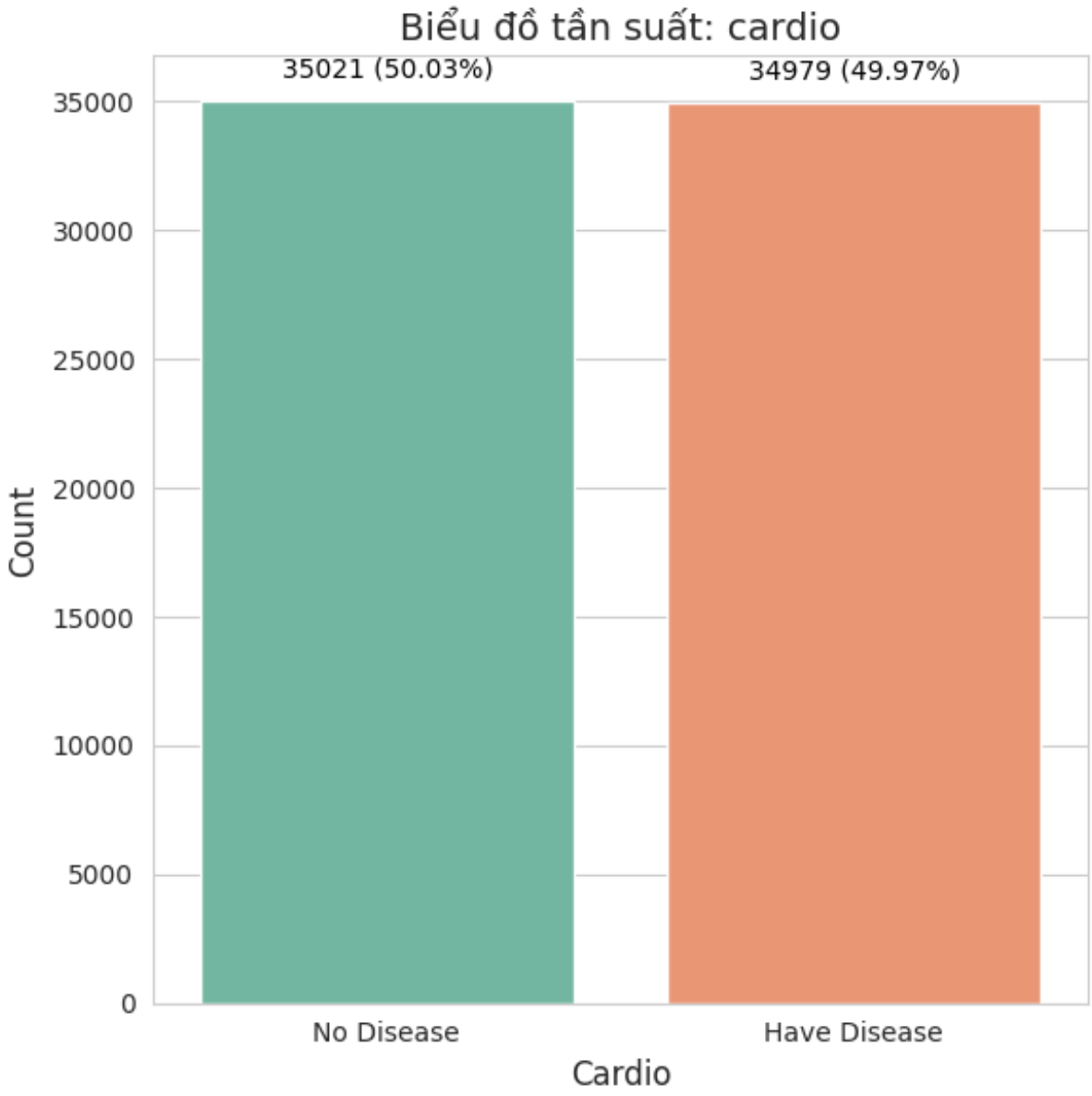
Hai biến đặc biệt cần chú ý là ap_hi (huyết áp tâm thu) và ap_lo (huyết áp tâm trương). Với ap_hi, giá trị trung bình là 128.82 mmHg, trong khi trung vị chỉ là 120 mmHg, cho thấy phân phối lệch phải mạnh. Đặc biệt, dữ liệu ghi nhận những giá trị bất thường như nhỏ hơn 0 và lớn hơn 16.000 mmHg, hoàn toàn không hợp lý về mặt y học, do đó được xác định là outlier nghiêm trọng. Phần lớn dữ liệu hợp lý nằm trong khoảng 100–150 mmHg, còn các giá trị quá lớn hoặc quá nhỏ nhiều khả năng là lỗi nhập liệu, ví dụ: nhầm đơn vị, thêm số 0, hoặc đảo ngược giữa ap_hi và ap_lo.

Biến ap_lo cũng có tình trạng tương tự, với giá trị trung bình 96.63 mmHg, trung vị 80 mmHg và xuất hiện các giá trị ngoài thực tế như -70 mmHg ở cực tiểu và 11.000 mmHg ở cực đại. Dải giá trị huyết áp tâm trương bình thường thường nằm trong khoảng 60–90 mmHg, nên các giá trị vượt ra ngoài đáng kể so với khoảng này cần được xem xét và xử lý kỹ lưỡng. Biểu đồ của ap_lo cũng có đuôi kéo dài và phân phối lệch phải, ảnh hưởng trực tiếp đến kết quả thống kê và mô hình hóa nếu không xử lý outlier.

Tóm lại, phần lớn các biến số trong tập dữ liệu có phân phối hợp lý, tuy nhiên tồn tại một số lượng đáng kể các giá trị bất thường, đặc biệt ở các biến liên quan đến huyết áp và cân nặng. Việc làm sạch dữ liệu – bao gồm việc loại bỏ hoặc cắt ngưỡng các giá trị không thực tế – là bắt buộc để đảm bảo chất lượng phân tích và độ chính xác của các mô hình học máy được xây dựng sau này.

2.4. Phân phối nhãn cardio

Dựa vào biểu đồ tần suất của biến mục tiêu `cardio`, có thể thấy dữ liệu được phân bố rất cân bằng giữa hai lớp: nhóm bệnh nhân được chẩn đoán mắc bệnh tim mạch (Have Disease) và nhóm không mắc bệnh (No Disease), với tỷ lệ lần lượt là 49.97% và 50.03%. Đây là một điểm mạnh đáng kể của tập dữ liệu, bởi trong thực tế, rất nhiều bài toán phân loại trong y học gặp phải vấn đề mất cân bằng nhãn nghiêm trọng, khiến mô hình học máy dễ bị thiên lệch, ưu tiên dự đoán theo lớp chiếm đa số.



Hình 6: Biểu đồ tần suất cardio

Với dữ liệu cân bằng như hiện tại, không cần áp dụng các kỹ thuật tái cân bằng (resampling) như oversampling (tăng số lượng mẫu của lớp thiểu số) hoặc undersampling (giảm mẫu từ lớp chiếm đa số), từ đó giúp giữ nguyên cấu trúc ban đầu của dữ liệu. Điều này cũng giúp giảm thiểu nguy cơ overfitting – một vấn đề phổ biến khi sử dụng các kỹ thuật tăng dữ liệu nhân tạo.

Hơn nữa, sự cân bằng nhãn tạo điều kiện thuận lợi để sử dụng các chỉ số đánh giá truyền thống như accuracy, precision, recall và F1-score mà không lo bị lệch trong việc đánh giá hiệu năng mô hình. Nhìn chung, đây là một điều kiện lý tưởng để xây dựng và huấn luyện các mô hình học máy một cách hiệu quả và đáng tin cậy.

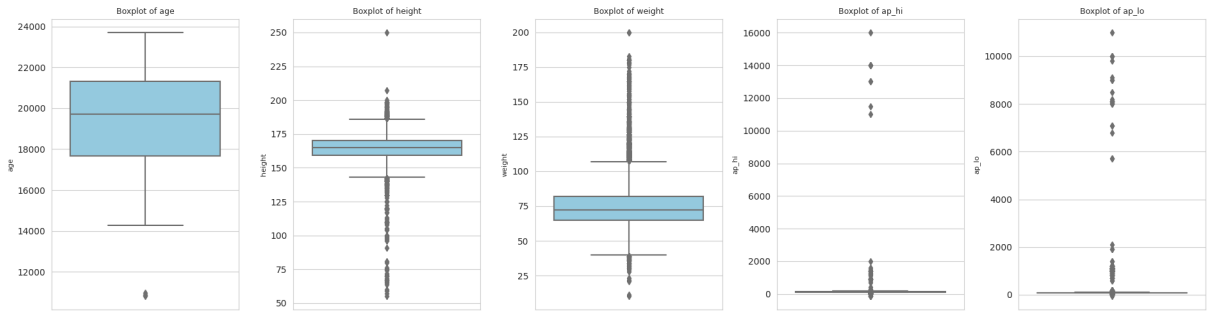
Một điểm đáng chú ý từ quá trình phân tích dữ liệu là mặc dù đa số bệnh nhân có các đặc trưng sức khỏe được đánh giá là "tốt" như: tỷ lệ không hút thuốc đạt khoảng

90%, không uống rượu khoảng 95%, mức đường huyết bình thường chiếm khoảng 85%, cholesterol trong ngưỡng an toàn khoảng 74% và mức độ vận động thể chất cao đạt khoảng 80%, thì tỷ lệ mắc bệnh tim mạch vẫn ở mức rất cao – gần 50%. Phát hiện này cho thấy rằng các đặc trưng riêng lẻ, dù tích cực, vẫn chưa đủ khả năng để phân biệt rõ ràng nguy cơ mắc bệnh tim mạch. Điều này đặt ra yêu cầu cần phải xem xét sự tương tác và mối quan hệ kết hợp giữa nhiều yếu tố khác nhau trong hồ sơ bệnh nhân.

Bên cạnh đó, các thói quen sinh hoạt (như hút thuốc, uống rượu, vận động) thường được tự khai báo bởi bệnh nhân. Điều này có thể dẫn đến sai lệch hoặc không chính xác, do nhiều người có xu hướng khai báo thấp hơn thực tế vì lý do xã hội hoặc cá nhân, từ đó ảnh hưởng trực tiếp đến khả năng học và dự đoán của mô hình.

2.5. Outliers

Khi tiến hành kiểm tra các đặc trưng dạng số trong tập dữ liệu, có thể nhận thấy biến age (tuổi) được ghi nhận theo đơn vị ngày, với phần lớn giá trị rơi vào khoảng hợp lý tương ứng với người trưởng thành. Tuy nhiên, vẫn xuất hiện một vài ngoại lệ thấp – khoảng 11.000 ngày (tương đương khoảng 30 tuổi), có thể phản ánh những sai sót nhỏ trong quá trình thu thập hoặc nhập liệu. Nhìn chung, phân phối của biến tuổi khá ổn định và không có sai lệch nghiêm trọng. Trong khi đó, hai biến height (chiều cao) và weight (cân nặng) lại cho thấy sự hiện diện rõ rệt của nhiều giá trị ngoại lệ ở cả hai phía. Đối với chiều cao so với tuổi của một người trưởng thành, một số điểm dữ liệu quá thấp (<120 cm) hoặc quá cao (>220 cm) so với thực tế y học bình thường. Cả chiều cao và cân nặng đều có khoảng IQR hẹp, cho thấy phần lớn dữ liệu tập trung xung quanh giá trị trung tâm, nhưng cũng khiến những outlier xuất hiện rõ ràng hơn. Đối với cân nặng, xuất hiện các giá trị thấp bất thường (dưới 30 kg) hoặc cao vượt mức (trên 200 kg), có khả năng đến từ lỗi nhập liệu, đơn vị không đồng nhất, hoặc các trường hợp bệnh lý đặc biệt.



Hình 7: Biểu đồ boxplot của các thuộc tính số

Đáng chú ý nhất trong số các đặc trưng dạng số là hai biến liên quan đến huyết áp: ap_hi (huyết áp tâm thu) và ap_lo (huyết áp tâm trương). Đây là những chỉ số y học rất

quan trọng trong đánh giá sức khỏe tim mạch, nhưng lại xuất hiện số lượng lớn giá trị ngoại lệ nghiêm trọng trong tập dữ liệu. Cụ thể, có nhiều trường hợp ap_hi vượt mức 1.000, thậm chí đạt tới hơn 16.000 mmHg – một giá trị hoàn toàn phi thực tế và không thể xảy ra trong thực tế lâm sàng. Tương tự, biến ap_lo cũng ghi nhận nhiều giá trị vượt quá 5.000 mmHg hoặc thậm chí mang giá trị âm, điều này hoàn toàn vô lý về mặt sinh lý học. Cả hai biến đều có khoảng IQR rất hẹp, cho thấy phần lớn dữ liệu nằm trong một khoảng giá trị bình thường, tuy nhiên sự tồn tại của những outlier cực đoan như vậy có thể ảnh hưởng nghiêm trọng đến cả mô tả thống kê và kết quả của các mô hình học máy nếu không được xử lý kỹ lưỡng. Việc giữ lại các điểm sai lệch này có thể khiến mô hình học sai xu hướng, giảm khả năng khái quát hóa, hoặc tạo ra các quyết định sai lệch trong bối cảnh ứng dụng y tế thực tế.

Để xác định các giá trị bất thường trong các đặc trưng dạng số của bộ dữ liệu, nhóm đã sử dụng phương pháp IQR (Interquartile Range) nhằm phát hiện các điểm dữ liệu nằm ngoài khoảng giá trị hợp lý. Kết quả cho thấy đặc trưng age có số lượng outlier rất thấp, chỉ 4 giá trị, phản ánh sự ổn định của phân phối tuổi trong tập dữ liệu. Tuy nhiên, các đặc trưng còn lại lại ghi nhận số lượng outlier đáng kể. Cụ thể, height (chiều cao), weight(cân nặng), ap_hi (huyết áp tâm thu), ap_lo (huyết áp tâm trương) lần lượt có 519, 1.819, 1.435 và 4.632 giá trị ngoại lệ – con số rất lớn, thể hiện rõ sự bất thường nghiêm trọng trong dữ liệu. Những giá trị không hợp lý như vậy không chỉ gây sai lệch trong phân tích thống kê mô tả mà còn ảnh hưởng tiêu cực đến hiệu quả của mô hình dự báo. Việc áp dụng IQR giúp nhận diện rõ các ngoại lệ trong dữ liệu, là cơ sở quan trọng để làm sạch dữ liệu và nâng cao độ tin cậy cho quá trình phân tích sau này.

STT	Thuộc tính	Số lượng outliers
0	age	4
1	height	519
2	weight	1819
3	ap_hi	1435
4	ap_lo	4632

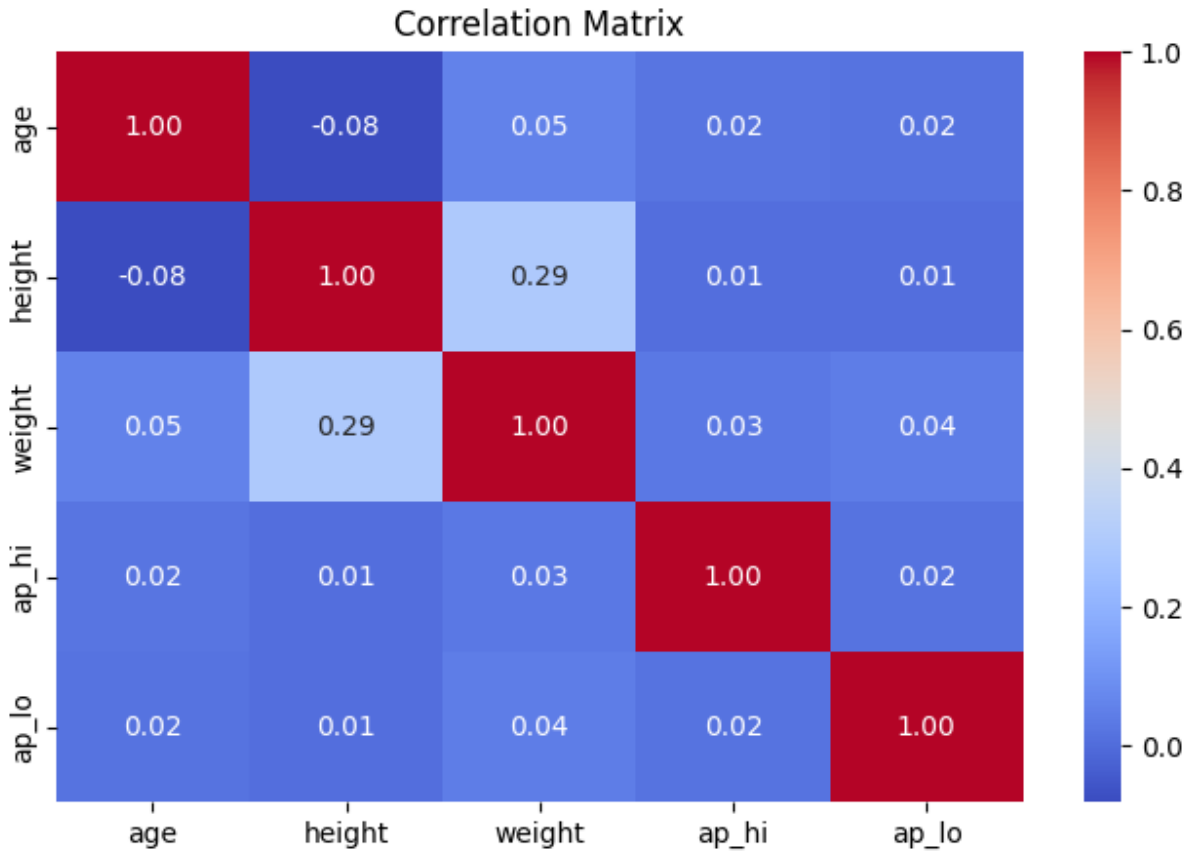
Bảng 2: Số lượng outlier theo từng thuộc tính dạng số

2.6. Phân tích mối quan hệ giữa các thuộc tính

2.6.1. Ma trận tương quan giữa các thuộc tính số

Cặp biến height và weight thể hiện tương quan dương nhẹ, cho thấy người có chiều cao lớn thường đi kèm với cân nặng cao hơn – đây là mối quan hệ sinh lý phổ biến và dễ

hiểu, phản ánh tính logic của dữ liệu. Trong khi đó, cặp biến ap_hi và ap_lo – đại diện cho huyết áp tâm thu và tâm trương – lại cho thấy tương quan yếu hơn kỳ vọng, mặc dù trong thực tế y học hai chỉ số này thường có mối liên hệ chặt chẽ và tăng đồng thời khi huyết áp tăng. Sự suy giảm tương quan trong dữ liệu hiện tại có thể được lý giải bởi sự tồn tại của outlier nghiêm trọng trong cả hai biến, gây sai lệch đáng kể trong việc tính toán hệ số tương quan thống kê.



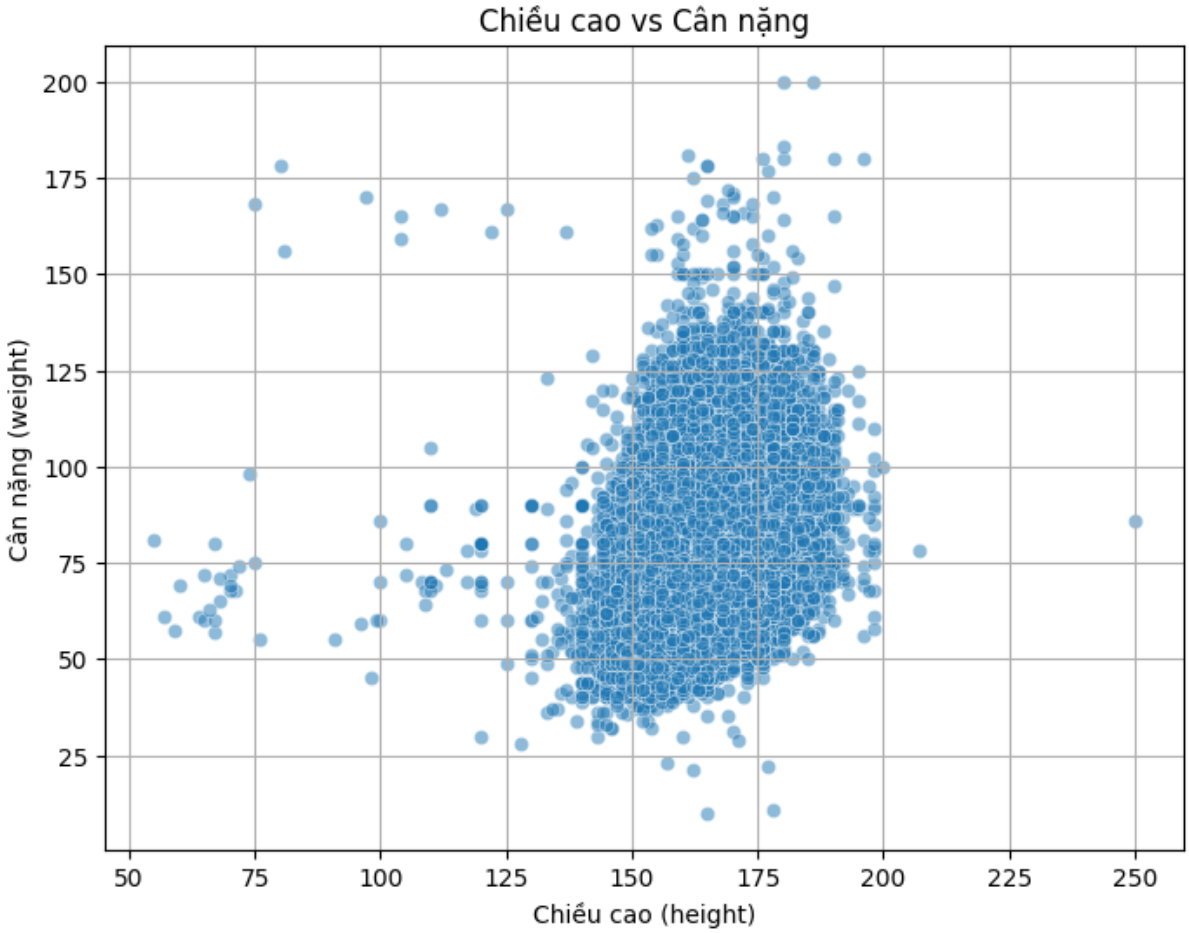
Hình 8: Ma trận tương quan giữa các thuộc tính số

Chính vì vậy, việc phân tích kỹ hơn mối quan hệ giữa các cặp biến quan trọng, đặc biệt là giữa chiều cao – cân nặng và huyết áp tâm thu – huyết áp tâm trương, là cần thiết để hiểu sâu hơn về cấu trúc dữ liệu và mối liên hệ giữa các yếu tố sinh lý. Phần tiếp theo sẽ tập trung trực quan hóa hai cặp biến này.

2.6.2. Mối quan hệ giữa chiều cao và cân nặng

Phân bố dữ liệu giữa chiều cao và cân nặng có dạng elip, cho thấy sự tồn tại của một tương quan dương nhẹ – người cao thường nặng hơn. Biểu đồ scatter cho thấy phần lớn các điểm dữ liệu tập trung trong khoảng chiều cao từ 150 đến 190 cm và cân nặng từ 40 đến 120 kg, phản ánh sự hợp lý về mặt sinh lý học. Tuy nhiên vẫn xuất hiện một số điểm nằm xa khỏi cụm chính, được xem là outlier, có thể ảnh hưởng đến phân tích thống kê

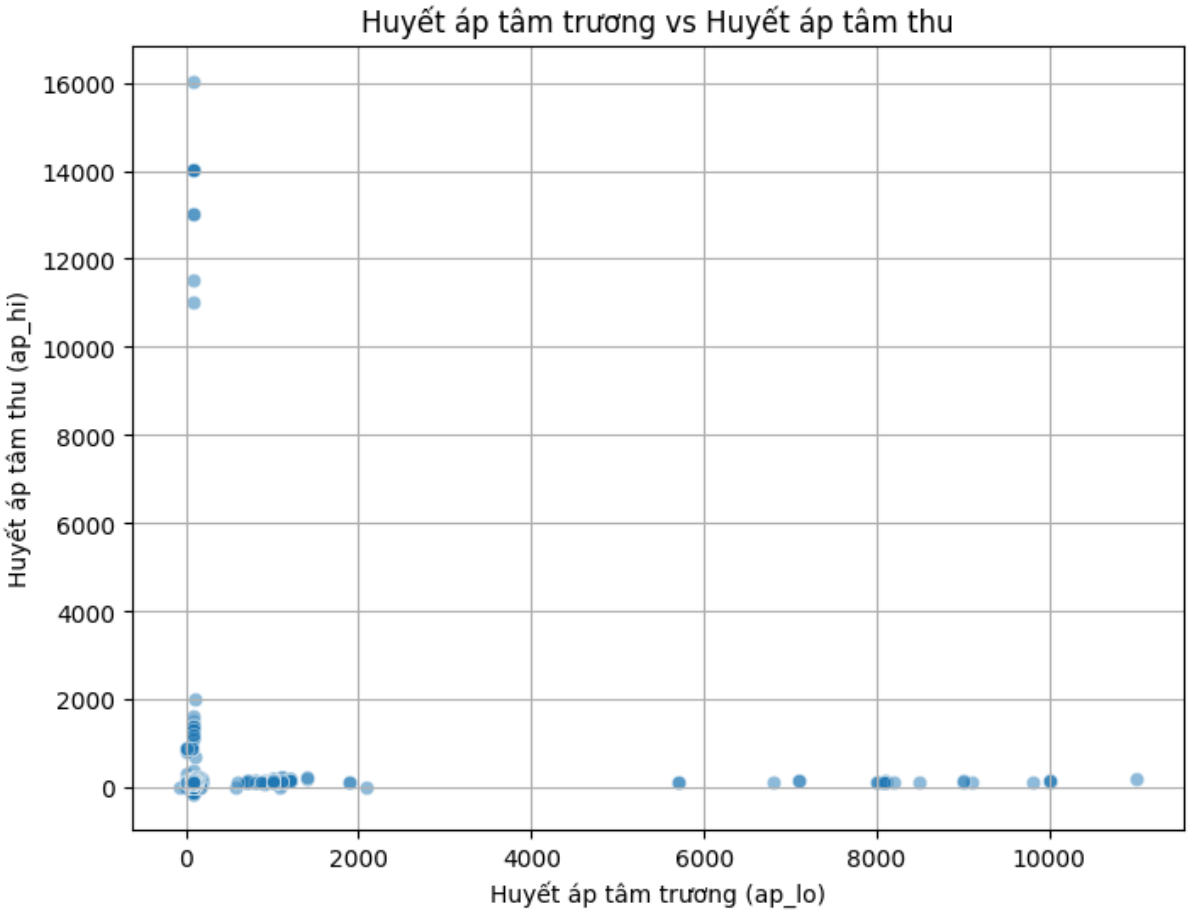
và mô hình hóa nếu không được xử lý. Nhìn chung, mối quan hệ giữa chiều cao và cân nặng trong tập dữ liệu là hợp lý và phù hợp với kỳ vọng thực tế.



Hình 9: Biểu đồ scatter chiều cao vs cân nặng

2.6.3. Mối quan hệ giữa huyết áp tâm thu và tâm trương

Biểu đồ scatter plot bị bóp méo nghiêm trọng do sự xuất hiện của quá nhiều giá trị ngoại lệ (outlier). Phần lớn các điểm dữ liệu tập trung tại một khu vực rất nhỏ gần gốc tọa độ, khiến cho biểu đồ không thể hiện rõ được xu hướng hay mối quan hệ giữa các biến. Do đó, để đánh giá chính xác tương quan thực sự, cần tiến hành lọc và loại bỏ các outlier trước, sau đó trực quan hóa lại biểu đồ scatter plot với dữ liệu đã được làm sạch.



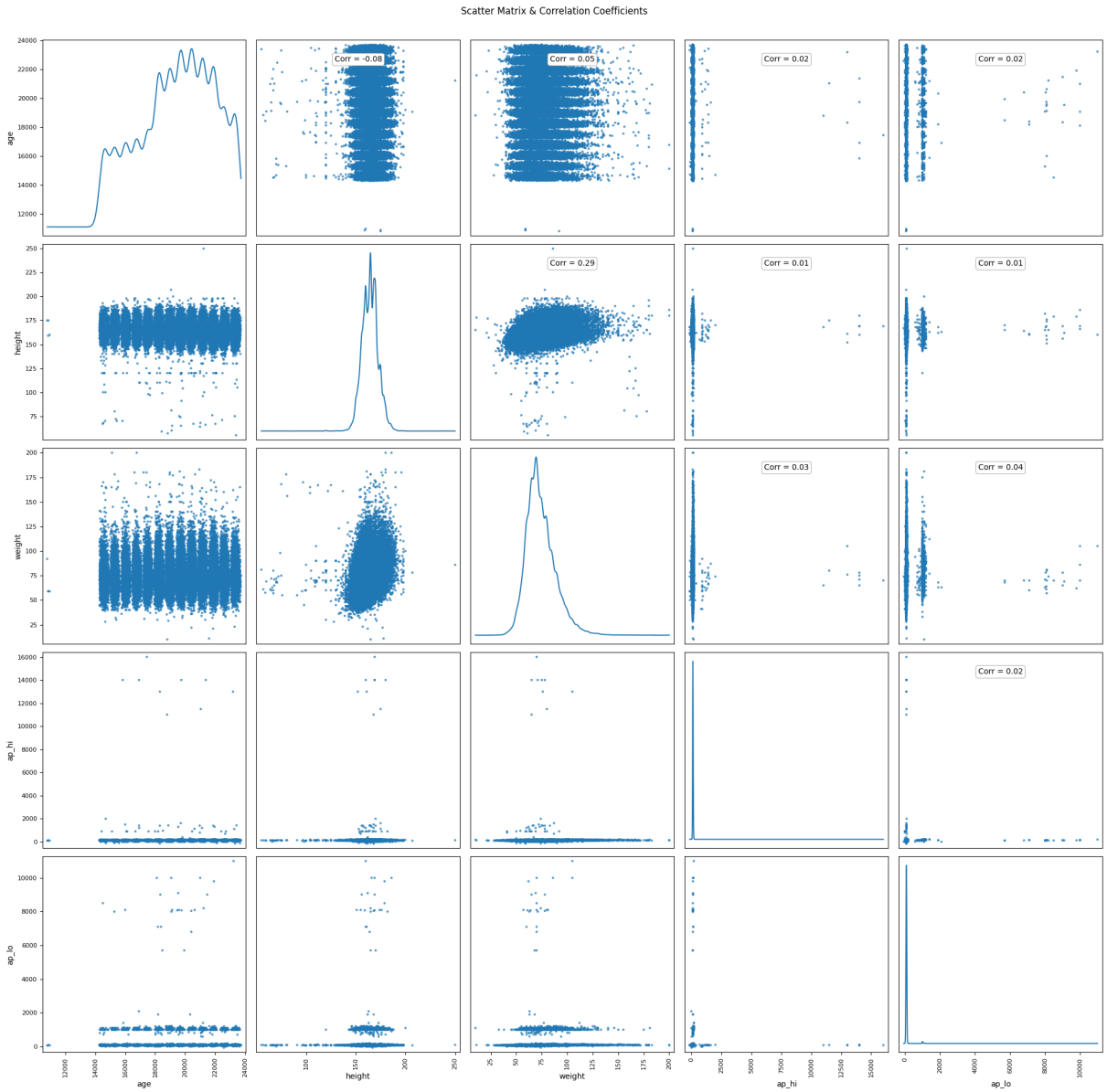
Hình 10: Biểu đồ scatter tâm trương vs tâm thu

2.6.4. Biểu đồ kết hợp hệ số tương quan giữa các cặp biến số

Biểu đồ Scatter Matrix thể hiện mối quan hệ giữa các biến số trong tập dữ liệu, đồng thời kết hợp với hệ số tương quan Pearson (corr) để đánh giá mức độ liên hệ tuyến tính giữa từng cặp biến. Trong đó, cặp biến height và weight có tương quan dương rõ ràng (hệ số tương quan 0.29), cho thấy người có chiều cao lớn hơn thường có cân nặng cao hơn. Ngược lại, cặp ap_hi và ap_lo – đại diện cho huyết áp tâm thu và tâm trương – lại có tương quan rất yếu (corr 0.02), chủ yếu do ảnh hưởng mạnh từ các giá trị ngoại lệ (outlier) bất thường.

Dựa trên biểu đồ phân tán và hệ số tương quan giữa tuổi (age) với các biến số dạng liên tục như chiều cao (height), cân nặng (weight), huyết áp tâm thu (ap_hi) và huyết áp tâm trương (ap_lo), có thể thấy rằng mối quan hệ tuyến tính giữa các biến này với tuổi là rất yếu. Cụ thể, hệ số tương quan giữa tuổi và chiều cao là -0.08 cho thấy xu hướng giảm nhẹ về chiều cao theo thời gian, tuy nhiên ảnh hưởng này không rõ rệt. Với cân nặng, hệ số tương quan là 0.05, tức là cân nặng có xu hướng tăng nhẹ theo tuổi nhưng không đáng kể. Hai biến huyết áp (ap_hi và ap_lo) đều có hệ số tương quan rất thấp, chỉ khoảng 0.02, cho thấy gần như không tồn tại mối liên hệ tuyến tính giữa tuổi và huyết

áp trong dữ liệu này. Như vậy, xét riêng từng biến, tuổi không cho thấy tác động rõ ràng, tuy nhiên cần lưu ý rằng trong các mô hình học máy về sau, ảnh hưởng của tuổi có thể xuất hiện dưới dạng tương tác phi tuyến hoặc kết hợp với các đặc trưng khác.

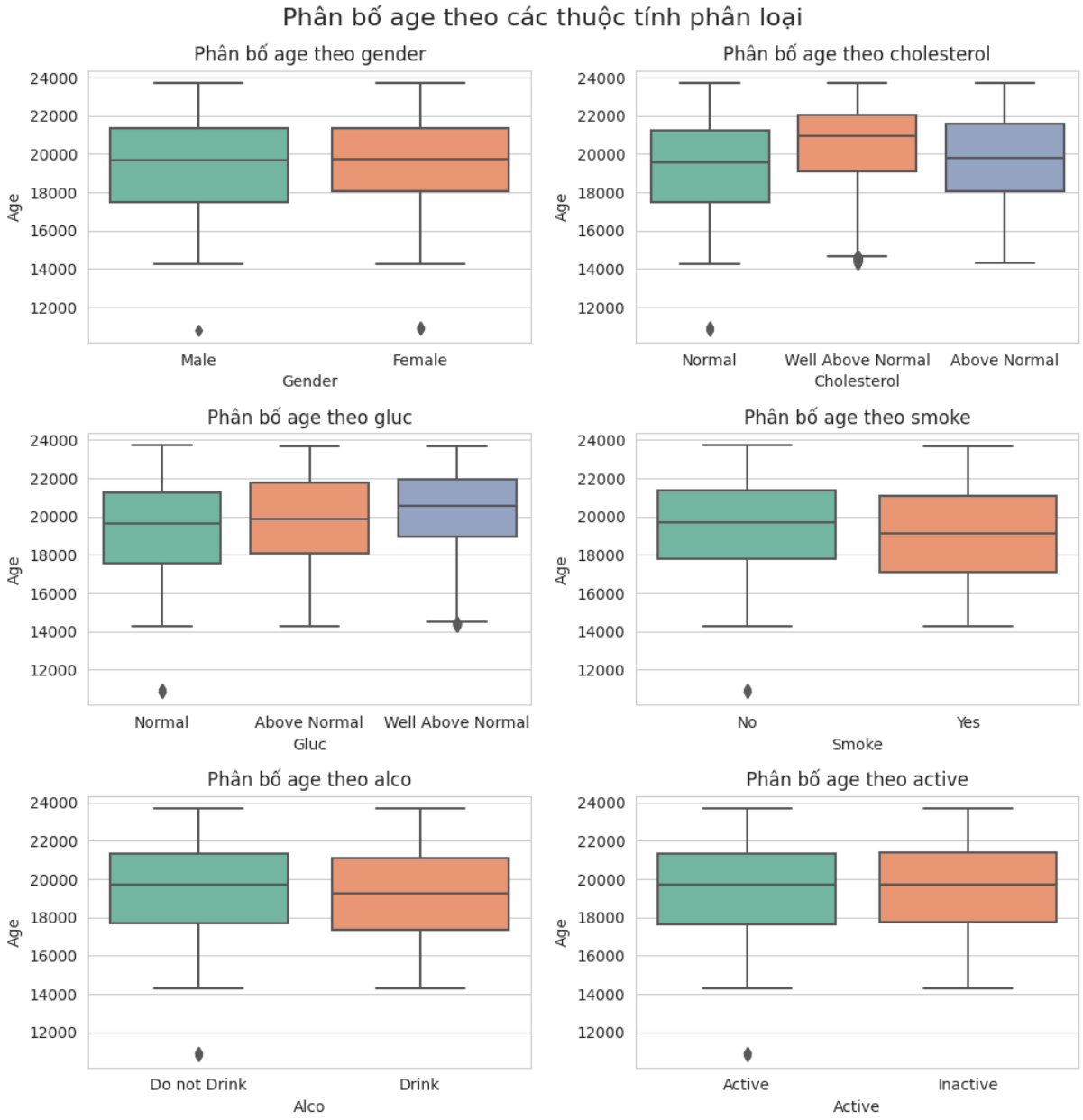


Hình 11: Biểu đồ kết hợp hệ số tương quan giữa các cặp biến số

2.6.5. Mối quan hệ giữa tuổi và các thuộc tính phân loại

Các biểu đồ phân phối theo từng đặc trưng phân loại cho thấy nhiều đặc điểm đáng chú ý trong dữ liệu. Về phân phối tuổi (age), phần lớn bệnh nhân tập trung trong khoảng từ 18.000 đến 22.000 ngày tuổi (tương đương 49–60 tuổi). Phụ nữ chiếm số lượng lớn hơn nam và có xu hướng cao tuổi hơn. Nhóm bệnh nhân có mức cholesterol và glucose cao thường nằm trong vùng tuổi cao hơn so với nhóm bình thường. Ngược lại, những người hút thuốc hoặc uống rượu có xu hướng trẻ hơn một chút. Đáng chú ý, những người

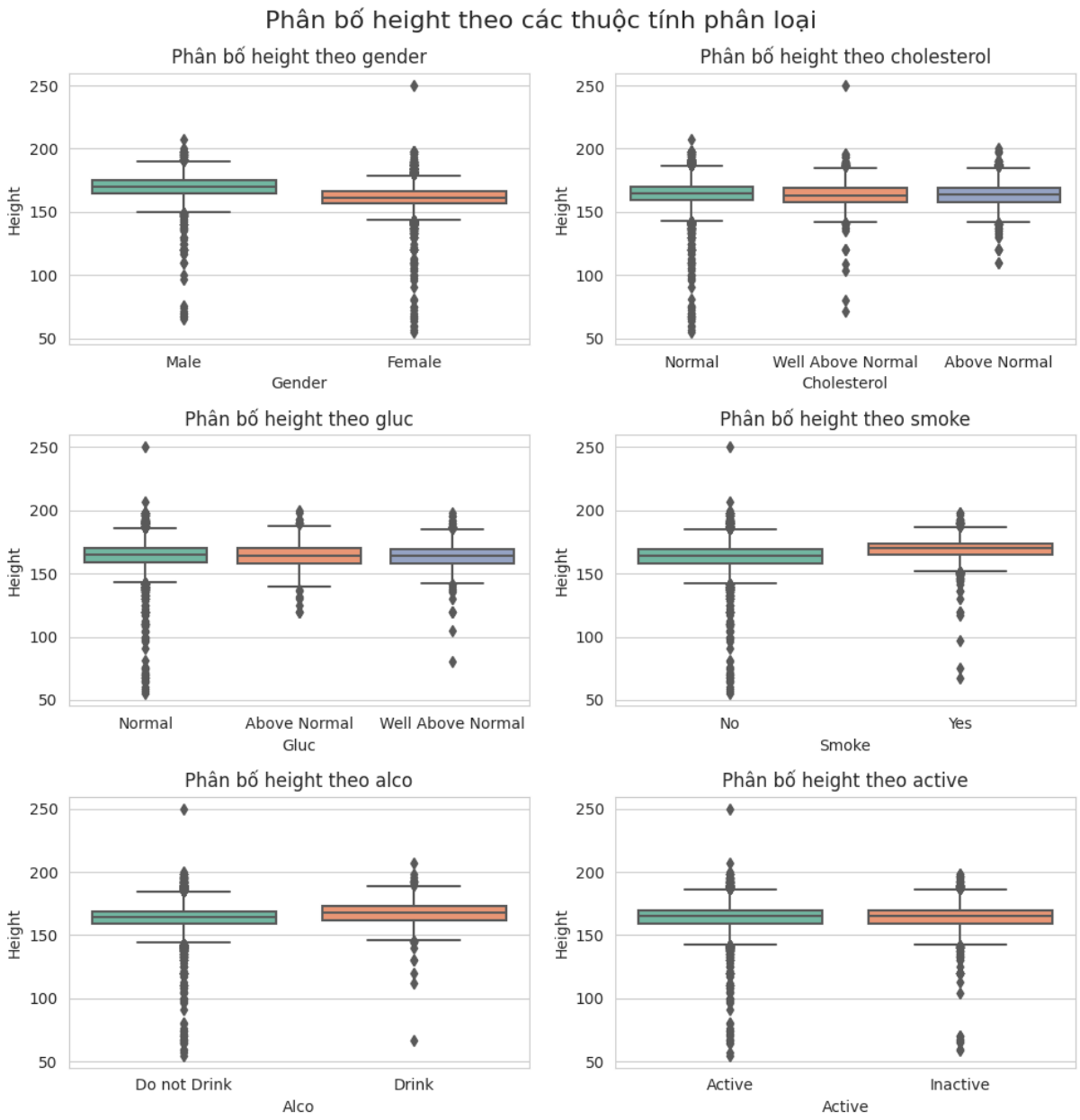
ít vận động thể chất (inactive) cũng tập trung nhiều ở nhóm tuổi cao, cho thấy tuổi tác đóng vai trò gia tăng nguy cơ mắc bệnh tim thông qua các yếu tố gián tiếp như rối loạn chuyển hóa và lối sống.



Hình 12: Biểu đồ phân bố tuổi và các thuộc tính phân loại

2.6.6. Mối quan hệ giữa chiều cao và các thuộc tính phân loại

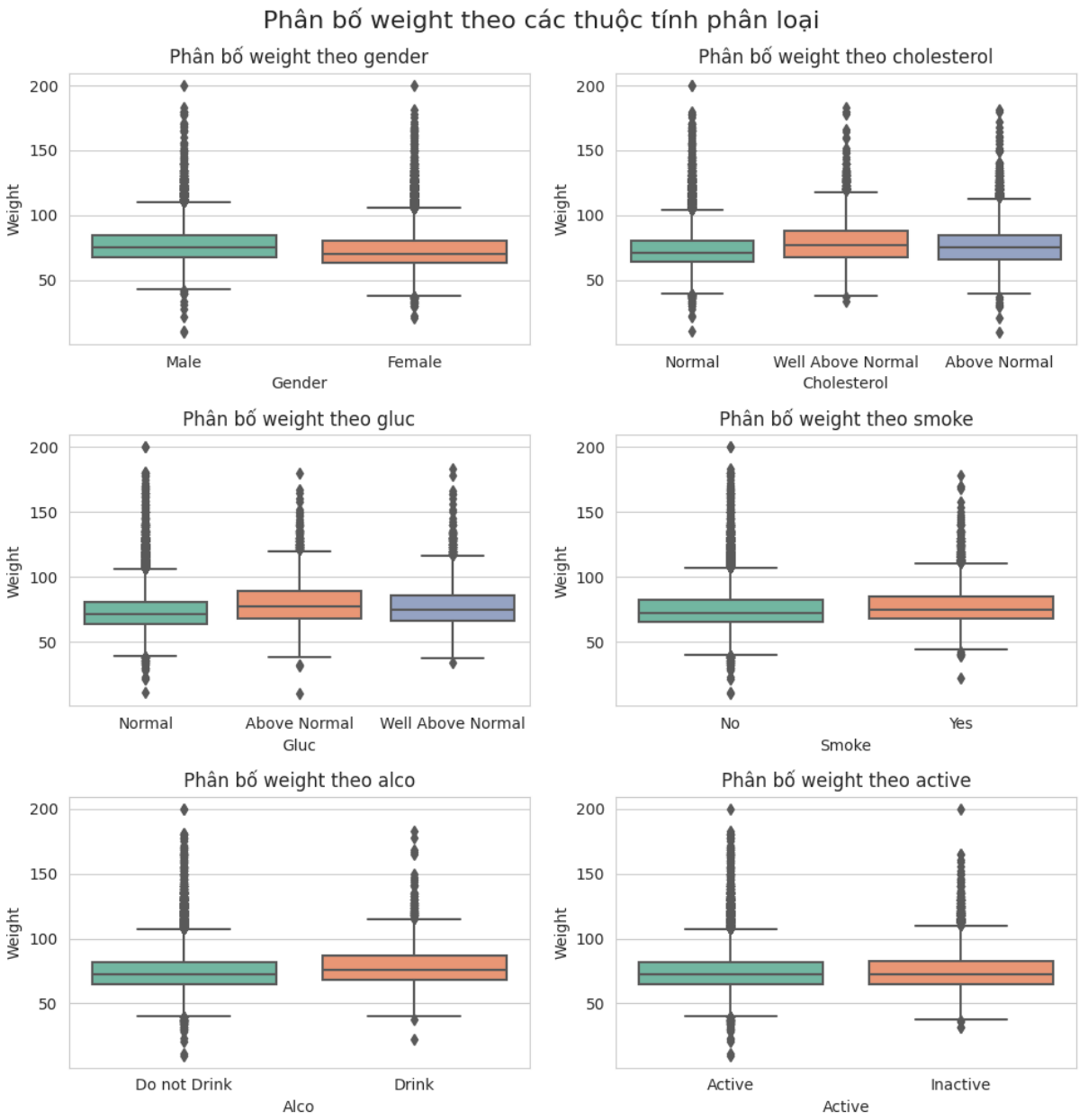
Với chiều cao (height), phân bố gần chuẩn, trong đó nam giới cao hơn nữ khoảng 9–10 cm. Tuy nhiên, chiều cao hầu như không có mối liên hệ rõ ràng với cholesterol, glucose, hút thuốc, rượu hay mức độ vận động. Do đó, chiều cao có thể không phải là một yếu tố dự báo trực tiếp nguy cơ tim mạch và thường chỉ được sử dụng để tính chỉ số BMI.



Hình 13: Biểu đồ phân bố chiều cao và các các thuộc tính phân loại

2.6.7. Mối quan hệ giữa cân nặng và các thuộc tính phân loại

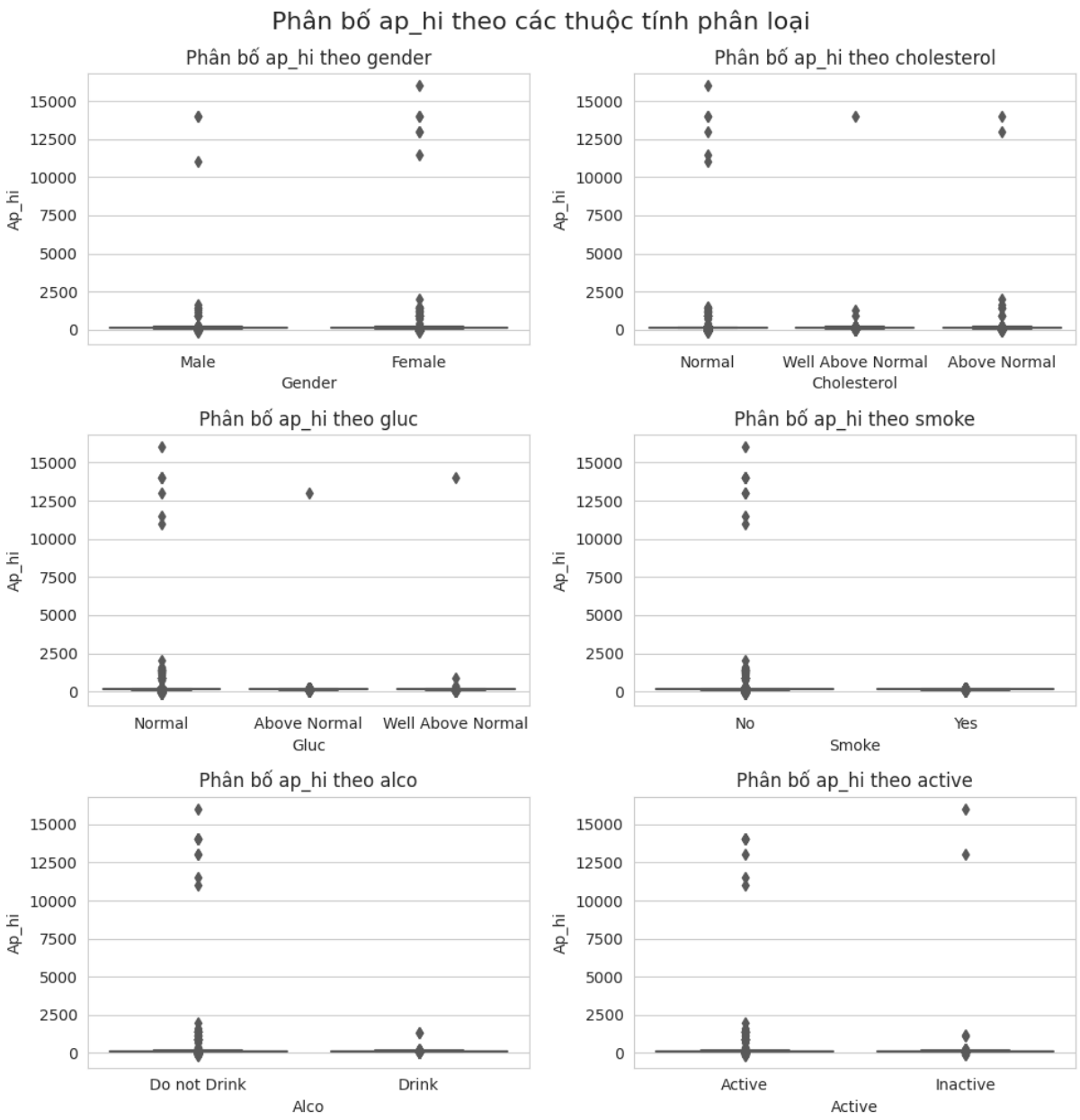
Phân tích phân phối cân nặng (weight) cho thấy biểu đồ lệch phải nhẹ, với đỉnh xung quanh 72 kg. Nam giới có xu hướng nặng hơn nữ. Cân nặng tăng dần theo mức cholesterol và glucose, đồng thời người ít vận động cũng có cân nặng cao hơn so với người vận động thường xuyên. Điều này cho thấy mối liên hệ rõ ràng giữa thừa cân và rối loạn chuyển hóa – một trục nguy cơ quan trọng dẫn đến bệnh tim mạch. Các yếu tố như hút thuốc hay uống rượu không cho thấy sự khác biệt rõ về cân nặng.



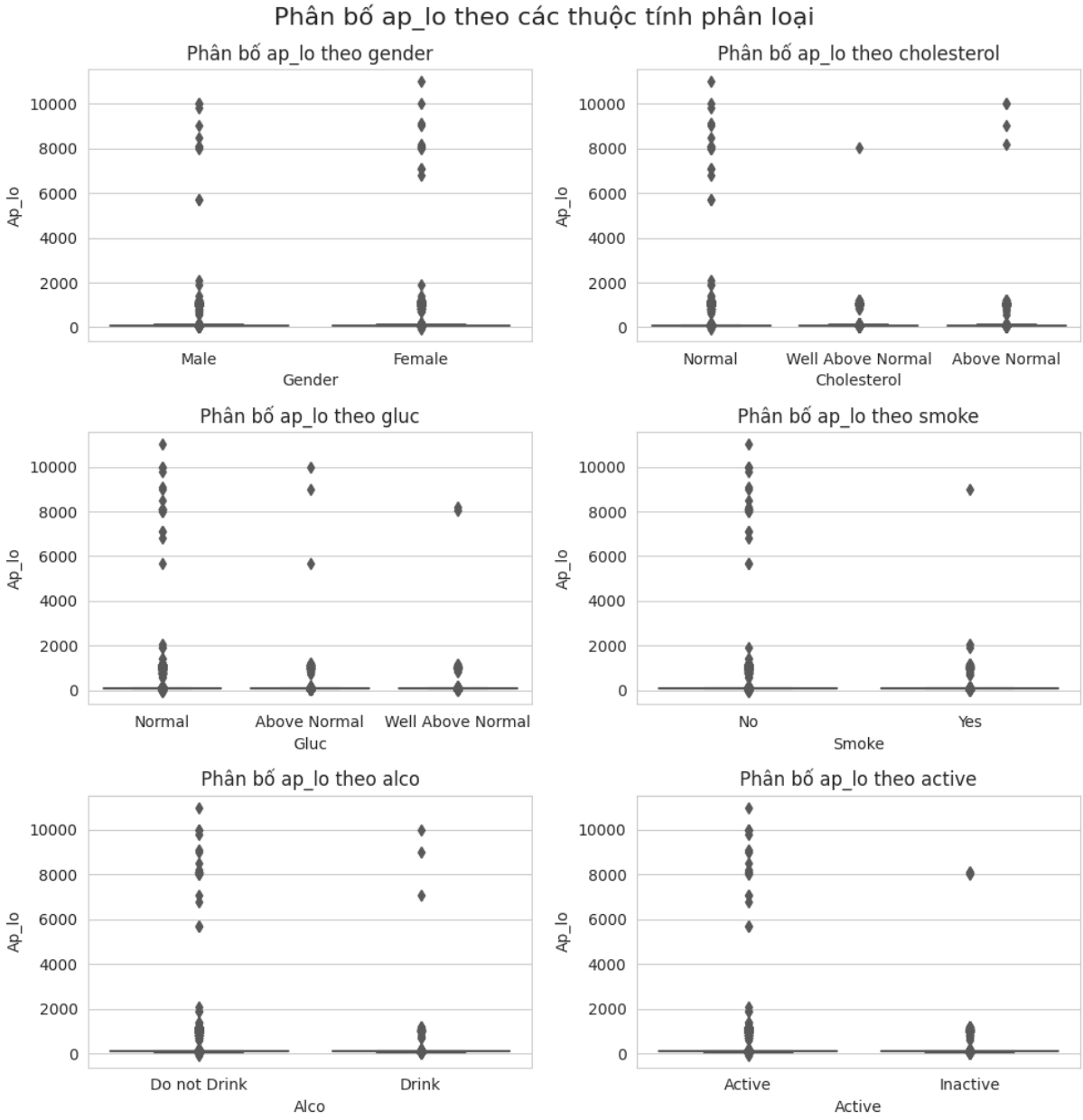
Hình 14: Biểu đồ phân bố cân nặng và các thuộc tính phân loại

2.6.8. Mối quan hệ giữa huyết áp tâm thu - tâm trương và các thuộc tính phân loại

Với huyết áp tâm thu (ap_hi) và tâm trương (ap_lo), biểu đồ cho thấy có sự xuất hiện của nhiều giá trị bất thường (outlier), với một số trường hợp vượt quá 10.000 hoặc thậm chí 16.000 mmHg – không hợp lý về mặt y học. Sau lớp giá trị bất thường này, phần lớn dữ liệu tập trung trong khoảng từ 90–200 mmHg (áp tâm thu) và 60–120 mmHg (áp tâm trương). Nhóm có cholesterol hoặc glucose cao thường có xu hướng huyết áp cao hơn, tuy nhiên các outlier đang làm biến dạng biểu đồ và ảnh hưởng đến khả năng phân tích chính xác. Do đó, cần loại bỏ hoặc cắt ngưỡng các giá trị không thực tế trước khi sử dụng hai biến này trong mô hình học máy.



Hình 15: Biểu đồ phân bố ap_hi và các thuộc tính phân loại

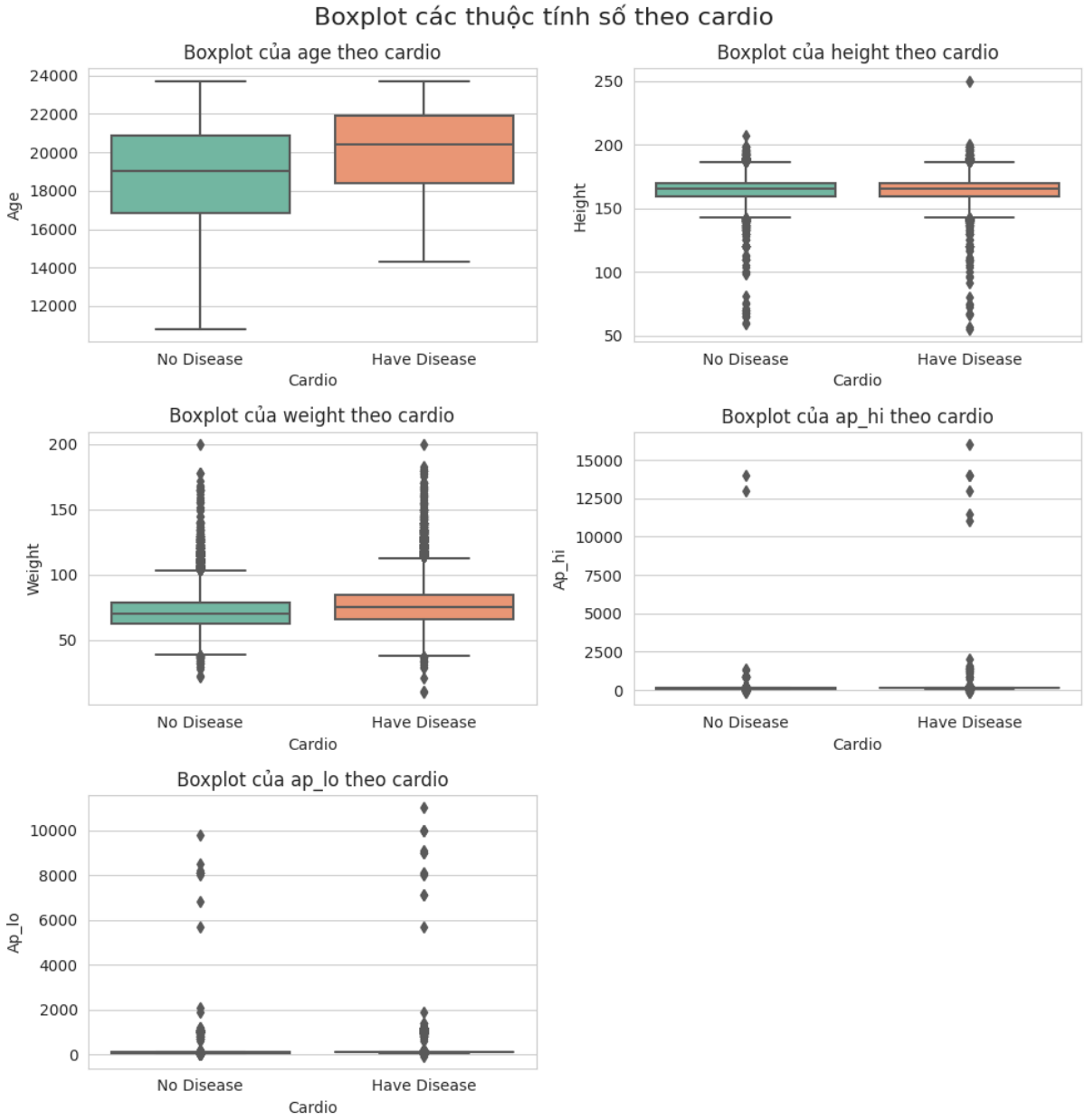


Hình 16: Biểu đồ phân bố ap_lo và các thuộc tính phân loại

2.7. Phân tích mối quan hệ giữa các thuộc tính và nhãn

2.7.1. Phân tích mối quan hệ giữa các thuộc tính số và nhãn

Các biểu đồ dạng boxplot trong hình thể hiện sự phân bố của các thuộc tính số bao gồm age, height, weight, ap_hi (huyết áp tâm thu) và ap_lo (huyết áp tâm trương) theo nhãn bệnh tim (cardio: 0 - không bệnh, 1 - có bệnh). Qua đó, ta có thể quan sát được mức độ khác biệt của từng đặc trưng giữa hai nhóm, từ đó rút ra các yếu tố có khả năng liên quan đến nguy cơ mắc bệnh tim mạch.



Hình 17: Biểu đồ boxplot các thuộc tính số theo cardio

Biểu đồ phân bố tuổi cho thấy sự tách biệt rõ ràng giữa hai nhóm. Nhóm có bệnh tim (màu cam) có xu hướng tập trung ở các mức tuổi cao hơn so với nhóm không có bệnh (màu xanh). Trong khi nhóm không bệnh phân bố chủ yếu ở khoảng từ 15000–21000 ngày tuổi (tương đương 41–58 tuổi), thì nhóm mắc bệnh lại nghiêng về phía 19000–23000 ngày tuổi (52–63 tuổi). Điều này phản ánh đúng mối liên hệ đã biết giữa tuổi tác và nguy cơ mắc bệnh tim mạch – người càng lớn tuổi thì nguy cơ mắc bệnh càng cao. Ngoài ra, sự phân bố có phần rộng hơn ở nhóm bệnh cũng cho thấy độ tuổi mắc bệnh khá đa dạng.

Chiều cao là một thuộc tính ít thể hiện sự khác biệt rõ ràng giữa hai nhóm. Cả hai đường phân phối gần như chồng khít lên nhau, tập trung chủ yếu ở khoảng 150–175 cm. Điều này cho thấy chiều cao không có mối liên hệ đáng kể với bệnh tim trong tập dữ liệu

này. Mặc dù có một số nghiên cứu cho rằng chiều cao có thể liên quan đến nguy cơ tim mạch (ví dụ người thấp có thể có nguy cơ cao hơn), nhưng trong tập dữ liệu này, yếu tố đó có vẻ không được phản ánh rõ.

Cân nặng là một đặc trưng thể hiện rõ sự phân biệt giữa hai nhóm. Biểu đồ cho thấy nhóm có bệnh tim có xu hướng nặng hơn so với nhóm không có bệnh. Đường phân phối của nhóm bệnh dịch sang phải và có phần trải rộng hơn, với nhiều bệnh nhân nặng trên 80 kg, trong khi nhóm không bệnh tập trung nhiều hơn ở khoảng 60–75 kg. Điều này hoàn toàn phù hợp với y văn, vì béo phì và thừa cân là các yếu tố nguy cơ đã được chứng minh đối với bệnh tim mạch. Mặc dù trong phạm vi dự án hiện tại chỉ dừng lại ở bước phân tích dữ liệu, nhưng kết quả quan sát cho thấy cân nặng là một yếu tố quan trọng, có thể được đưa vào mô hình học máy trong các bước nghiên cứu tiếp theo.

Phân phối của chỉ số huyết áp tâm thu cho thấy sự bất thường khá lớn trong dữ liệu. Cả hai nhóm đều xuất hiện các giá trị cực lớn (trên 8000), vượt xa ngưỡng sinh lý bình thường. Dù phần lớn dữ liệu tập trung quanh mức hợp lý (100–180 mmHg), các giá trị ngoại lai đã làm cho biểu đồ bị kéo dài về phía phải và che khuất sự khác biệt tiềm năng giữa hai nhóm. Nhìn kỹ hơn vào vùng giá trị hợp lý, có thể thấy nhóm mắc bệnh có xu hướng có huyết áp cao hơn một chút. Tuy nhiên, để kết luận rõ ràng hơn, cần phải loại bỏ hoặc xử lý các ngoại lệ trong dữ liệu, ví dụ thông qua việc cắt ngưỡng hoặc chuẩn hóa.

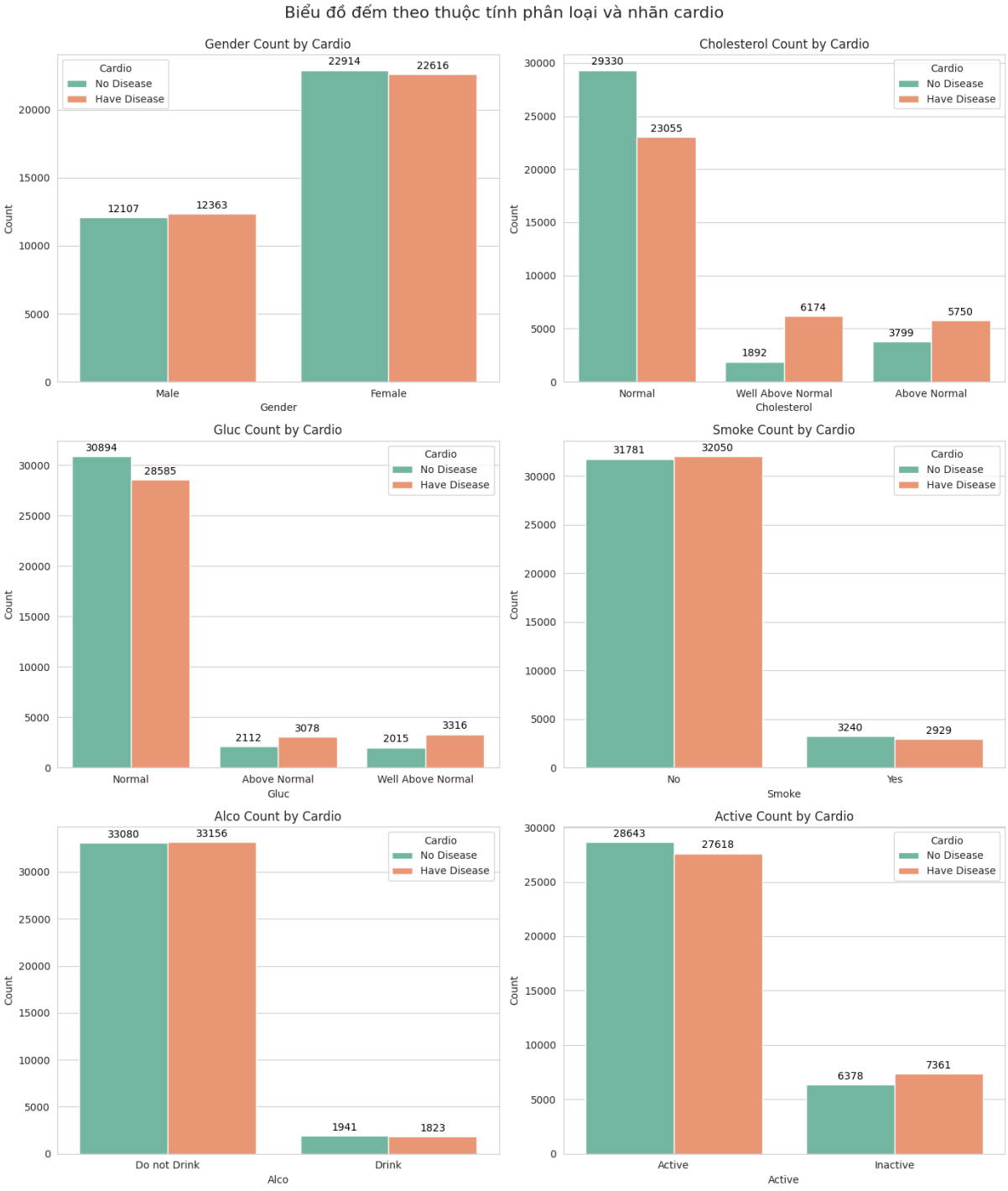
Tình trạng tương tự cũng xuất hiện với huyết áp tâm trương – rất nhiều giá trị bất hợp lý (ví dụ >10000) khiến biểu đồ bị kéo lệch nghiêm trọng. Nhìn vào phần dữ liệu có giá trị hợp lý (60–120 mmHg), sự khác biệt giữa hai nhóm là không rõ ràng, nhưng vẫn có xu hướng nhóm có bệnh có mức huyết áp cao hơn một chút. Tuy nhiên, cũng như với ap_hi, dữ liệu này cần được tiền xử lý để loại bỏ các điểm sai lệch trước khi đưa vào phân tích chuyên sâu hoặc mô hình học máy.

Qua các biểu đồ trên, có thể thấy rằng tuổi và cân nặng là hai đặc trưng số có mối liên hệ rõ ràng nhất với khả năng mắc bệnh tim mạch, cho thấy sự phân biệt giữa hai nhóm là rõ rệt và nhất quán. Ngược lại, chiều cao gần như không có ảnh hưởng đáng kể trong tập dữ liệu này.

Trong khi đó, hai chỉ số huyết áp (ap_hi và ap_lo) về mặt lý thuyết là các yếu tố y học quan trọng, nhưng dữ liệu hiện tại chứa quá nhiều giá trị bất thường và cần phải được xử lý sạch sẽ để đảm bảo phân tích đúng đắn. Nếu xử lý tốt, đây vẫn là những đặc trưng có tiềm năng cao trong việc dự đoán bệnh tim.

2.7.2. Phân tích mối quan hệ giữa các thuộc tính phân loại và nhãn

Các biểu đồ thể hiện số lượng bệnh nhân theo từng nhóm của các biến phân loại (categorical features), được chia thành hai nhóm: Không mắc bệnh tim (No Disease) và Mắc bệnh tim (Have Disease). Việc so sánh chiều cao của cột giữa hai nhóm trong mỗi biểu đồ giúp ta xác định những yếu tố nào có liên quan đáng kể đến nguy cơ mắc bệnh tim mạch trong tập dữ liệu này.



Hình 18: Biểu đồ đếm theo nhãn cardio và thuộc tính phân loại

Biểu đồ đầu tiên cho thấy sự phân bố giới tính gần như cân bằng giữa hai nhóm. Cả nam và nữ đều có tỷ lệ mắc bệnh tim khá tương đồng. Cụ thể, trong nhóm nam có 12,107 người không bệnh và 12,363 người bệnh; trong khi nhóm nữ có 22,914 người không bệnh và 22,616 người bệnh. Tỷ lệ gần như ngang bằng cho thấy giới tính không phải là yếu tố ảnh hưởng đáng kể đến khả năng mắc bệnh tim trong dữ liệu này.

Biểu đồ thứ hai cho thấy một mối liên hệ rõ ràng giữa cholesterol và nguy cơ mắc bệnh tim. Ở mức cholesterol bình thường, số người không bệnh chiếm đa số (29,330 người) so với người có bệnh (23,055). Tuy nhiên, ở các mức "trên chuẩn"(Above Normal) và "cao vượt chuẩn"(Well Above Normal), số người có bệnh tim lại chiếm ưu thế. Ví dụ, ở mức cholesterol cao vượt chuẩn, số người mắc bệnh là 6,174 so với chỉ 1,892 người không bệnh. Điều này khẳng định cholesterol là một yếu tố nguy cơ quan trọng đối với bệnh tim.

Kết quả tương tự được quan sát ở thuộc tính glucose. Nhóm có glucose bình thường chiếm đa số ở cả hai nhóm, nhưng tỷ lệ người mắc bệnh trong các nhóm glucose cao hơn mức bình thường (Above Normal và Well Above Normal) lại nhiều hơn rõ rệt so với nhóm không bệnh. Ví dụ, ở mức glucose “Well Above Normal”, số người mắc bệnh là 3,316, gần gấp đôi số người không bệnh (2,015). Điều này cho thấy mức đường huyết cao là một yếu tố đáng quan tâm trong đánh giá nguy cơ bệnh tim.

Ngược lại với cholesterol và glucose, hút thuốc không cho thấy sự khác biệt rõ ràng giữa hai nhóm. Số lượng người hút thuốc ở nhóm không bệnh (3,240) và nhóm bệnh (2,929) gần như bằng nhau. Tương tự, phần lớn bệnh nhân đều không hút thuốc. Điều này có thể do thiếu độ chính xác trong tự báo cáo hành vi hút thuốc, hoặc cỡ mẫu hút thuốc quá nhỏ, gây ra việc khó đánh giá mối liên hệ thực sự. Do đó, biến hút thuốc không thể hiện được mối liên hệ rõ ràng với bệnh tim trong tập dữ liệu này.

Phân bố giữa hai nhóm uống rượu và không uống rượu gần như giống nhau hoàn toàn ở cả hai nhóm bệnh và không bệnh. Điều này phản ánh rằng yếu tố “uống rượu” không không thể hiện được mối liên hệ rõ ràng với bệnh tim trong dữ liệu này. Ngoài ra, cũng có thể có vấn đề về độ tin cậy của dữ liệu (người dùng có thể khai sai hoặc thiếu thông tin thực tế về hành vi uống rượu).

Biểu đồ cuối cùng cho thấy sự khác biệt rõ ràng giữa người hoạt động và không hoạt động thể chất. Ở nhóm "Inactive", số người mắc bệnh tim (7,361) vượt đáng kể so với người không mắc bệnh (6,378). Trong khi đó, ở nhóm "Active", số người không mắc bệnh lại nhiều hơn (28,643 so với 27,618). Mặc dù sự khác biệt không lớn như ở cholesterol và glucose, nhưng đây vẫn là một yếu tố quan trọng cho thấy thiếu vận động thể chất có thể liên quan đến nguy cơ cao hơn bị bệnh tim.

Cholesterol và glucose là hai thuộc tính phân loại có mối liên hệ mạnh nhất với khả

năng mắc bệnh tim mạch. Những người có mức cholesterol hoặc glucose cao có tỷ lệ mắc bệnh tim cao hơn rõ rệt so với nhóm bình thường. Mức độ hoạt động thể chất cũng là một yếu tố quan trọng: người không vận động thường xuyên có tỷ lệ mắc bệnh cao hơn. Ngược lại, giới tính, hút thuốc và uống rượu không cho thấy sự khác biệt đáng kể, điều này có thể do dữ liệu không đủ đại diện hoặc chưa đủ chi tiết để phản ánh đúng ảnh hưởng của các yếu tố này.

Những biểu đồ trên cho thấy việc tập trung vào các yếu tố sinh học như cholesterol, glucose và hoạt động thể chất sẽ hữu ích hơn trong việc xây dựng mô hình dự đoán bệnh tim từ tập dữ liệu hiện tại. Các yếu tố hành vi như hút thuốc và uống rượu có thể cần được thu thập chi tiết hơn hoặc loại bỏ khỏi mô hình nếu không cải thiện hiệu suất phân loại.

3. Kết luận & Định hướng mô hình

3.1. Kết luận

Quá trình phân tích dữ liệu (Exploratory Data Analysis – EDA) đã mang lại nhiều phát hiện quan trọng, giúp nhóm hiểu sâu hơn về bản chất bộ dữ liệu và các yếu tố liên quan đến bệnh tim mạch. Bộ dữ liệu có quy mô lớn (70.000 bản ghi) và chất lượng tốt, không chứa giá trị thiếu (null) hay trùng lặp, tạo điều kiện thuận lợi cho việc thống kê và trực quan hóa. Tuy nhiên, một số biến dạng số như huyết áp, chiều cao, cân nặng lại xuất hiện nhiều giá trị ngoại lệ phi thực tế, ảnh hưởng không nhỏ đến tính đại diện của dữ liệu và gây méo mó khi phân tích các phân phối. Việc nhận diện các outlier này là bước quan trọng để đánh giá mức độ tin cậy của dữ liệu và xây dựng các tiền đề xử lý phù hợp trong các bước nâng cao sau này.

Thông qua phân tích mô tả và biểu đồ trực quan, nhóm nhận thấy có nhiều đặc trưng thể hiện mối quan hệ mạnh với khả năng mắc bệnh tim. Cụ thể, tuổi là yếu tố rõ ràng nhất: nhóm bệnh nhân có nhãn `cardio = 1` thường cao tuổi hơn hẳn nhóm còn lại, điều này phù hợp với hiểu biết y khoa về nguy cơ tim mạch tăng theo tuổi. Huyết áp tâm thu (`ap_hi`) và tâm trương (`ap_lo`) cũng cho thấy sự khác biệt đáng kể giữa hai nhóm nhãn, dù có nhiều nhiễu do outlier. Chỉ số `weight` thể hiện rằng bệnh nhân nặng cân thường có tỷ lệ mắc bệnh cao hơn, đặc biệt khi kết hợp với mức cholesterol và glucose tăng.

Bên cạnh đó, một số biến phân loại lại chưa thể hiện ảnh hưởng rõ ràng. Các yếu tố như hút thuốc, uống rượu và giới tính có phân bố lệch (đa số không hút, không uống, tỷ lệ nữ cao), gây khó khăn trong việc đánh giá tác động thực sự của những biến này đến

nguy cơ tim mạch. Điều này đặt ra nghi vấn về độ chính xác của dữ liệu tự báo cáo và cho thấy nhu cầu mở rộng mẫu hoặc tái thiết kế cách thu thập thông tin trong các nghiên cứu tương lai.

Tổng thể, việc EDA không chỉ giúp làm sáng tỏ bức tranh tổng quan về sức khỏe tim mạch trong tập dữ liệu mà còn là nền tảng quan trọng để định hướng xây dựng các mô hình học máy sau này. Những yếu tố như tuổi, huyết áp, cholesterol, glucose và hoạt động thể chất đều có ý nghĩa thống kê rõ rệt và giá trị thực tiễn cao. Qua đó, nhóm đã thành công trong việc xác định được tập biến tiềm năng, phát hiện các bất thường trong dữ liệu và phân tích được các xu hướng có ý nghĩa lâm sàng. Đây là bước chuẩn bị quan trọng nếu sau này mở rộng dự án sang hướng xây dựng hệ thống dự báo nguy cơ bệnh tim nhằm phục vụ chẩn đoán sớm và hỗ trợ cộng đồng.

3.2. Định hướng mô hình học máy

Mặc dù đề tài hiện tại tập trung vào phân tích khám phá dữ liệu (EDA), kết quả thu được đã mở ra nhiều hướng triển khai mô hình học máy trong tương lai. Cụ thể:

- Bài toán phân loại nhị phân (binary classification) với nhãn mục tiêu là `cardio` hoàn toàn phù hợp để áp dụng các mô hình học máy như **Logistic Regression**, **Random Forest**, hoặc **XGBoost**.
- Các đặc trưng đầu vào tiềm năng nên đưa vào mô hình gồm: `age`, `ap_hi`, `ap_lo`, `weight`, `cholesterol`, `gluc`, `active`.
- Cần loại bỏ hoặc xử lý các outlier nghiêm trọng trước khi huấn luyện để cải thiện độ chính xác và giảm ảnh hưởng tiêu cực đến mô hình.
- Do dữ liệu đã cân bằng giữa hai lớp nên chưa cần áp dụng các kỹ thuật xử lý mất cân bằng như oversampling hay weighting.
- Một số đặc trưng như `smoke`, `alco`, `gender` có thể được đưa vào thử nghiệm nhưng nên đánh giá lại sau bước *feature selection* do độ phân tán thấp và thiếu sự khác biệt rõ rệt giữa hai nhãn.
- Các bước cần thực hiện trước khi huấn luyện gồm: chuẩn hóa dữ liệu số (scaling), kiểm tra multicollinearity và áp dụng *cross-validation* để đánh giá tính ổn định của mô hình.

Trong tương lai, nhóm có thể mở rộng đề tài theo hướng xây dựng ứng dụng dự báo bệnh tim trực tuyến, tích hợp giao diện nhập liệu và mô hình học máy đã huấn luyện nhằm phục vụ mục đích sàng lọc nguy cơ trong cộng đồng.

Bên cạnh việc xây dựng các mô hình dự đoán nhãn bệnh tim, một hướng mở rộng quan trọng là **phân nhóm bệnh nhân theo đặc điểm sức khỏe**. Việc nhóm các bệnh nhân có thuộc tính tương đồng giúp hỗ trợ phân tích chuyên sâu, phát hiện các nhóm nguy cơ cao tiềm ẩn, hoặc gợi ý phác đồ điều trị phù hợp hơn cho từng nhóm.

Để thực hiện điều này, có thể sử dụng các mô hình học máy không giám sát như **KMeans** hoặc **DBSCAN** để phân cụm các bệnh nhân theo các thuộc tính như tuổi, huyết áp, `weight`, `cholesterol`, `glucose`, ... Kết quả phân cụm có thể hỗ trợ trong việc thiết kế các chương trình phòng ngừa và quản lý bệnh theo nhóm, từ đó cá nhân hóa việc chăm sóc sức khỏe và nâng cao hiệu quả can thiệp.

TÀI LIỆU THAM KHẢO

[1] Sulianova, Y. *Cardiovascular Disease Dataset*. Kaggle.
https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset

[2] Pandas Documentation.
https://pandas.pydata.org/docs/

[3] Matplotlib Documentation.
https://matplotlib.org/stable/contents.html

[4] Seaborn Documentation.
https://seaborn.pydata.org/

[5] Slide bài giảng và bài tập thực hành của môn học Python for Machine Learning.

[6] Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow* (2nd ed.). O'Reilly Media.

[7] Raschka, S., & Mirjalili, V. (2020). *Python Machine Learning* (3rd ed.). Packt Publishing.

[8] McKinney, W. (2018). *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython* (2nd ed.). O'Reilly Media.

[9] VanderPlas, J. (2016). *Python Data Science Handbook*. O'Reilly Media.

[10] Kuhn, M., & Johnson, K. (2013). *Applied Predictive Modeling*. Springer.

[11] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning (ISLR)*. Springer.

[12] Han, J., Kamber, M., & Pei, J. (2011). *Data Mining: Concepts and Techniques* (3rd ed.). Morgan Kaufmann.