ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

ĐÂY LÀ NHÓM PI

- *Nguyễn Phương Nam 23020406*
- *Nguyễn Đình Quyền 23020422*
- *Trần Doãn Thắng 23020438*

HEART ATTACK

BÀI TẬP LỚN MÔN LẬP TRÌNH XỬ LÝ DỮ LIỆU
Ngành: Trí tuệ nhân tạo

HÀ NỘI - 2024

ĐẠI HỌC QUỐC GIA HÀ NỘI

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

BÁO CÁO HỌC PHẦN LẬP TRÌNH XỬ LÝ DỮ LIỆU

ĐỀ TÀI: PHÂN TÍCH NGUY CƠ ĐAU TIM

Nhóm thực hiện:

- Nguyễn Phương Nam (MSSV: 23020406)
- Nguyễn Đình Quyền (MSSV: 23020422)
- Trần Doãn Thắng (MSSV: 23020438)

Github Repo: thangtrandoan/DataProcessing-Project

I/ *Đặt vấn đề*

1. Xác định vấn đề

Hiện nay, bệnh đau tim là một trong những nguyên nhân hàng đầu dẫn đến tử vong trên toàn cầu.
Việc hiểu rõ các yếu tố nguy cơ như tuổi tác, tiền sử gia đình, thói quen sinh hoạt, và chế độ ăn uống sẽ giúp tăng cường nhận thức cộng đồng và cải thiện chăm sóc y tế.

2. Mục tiêu dự án

- *Phân tích và hiểu dữ liệu liên quan đến nguy cơ đau tim.*
- *Xác định các yếu tố chính góp phần gia tăng nguy cơ.*
- *Đưa ra gợi ý thay đổi lối sống để giảm thiểu rủi ro.*
- *Xây dựng mô hình dự đoán nguy cơ đau tim dựa trên dữ liệu cá nhân.*

3. Các vấn đề đặt ra

- *Thói quen sinh hoạt nào có tác động lớn nhất đến nguy cơ đau tim?*
- *Liệu nguy cơ đau tim có khác biệt đáng kể giữa các nhóm dân cư dựa trên thu nhập hay quốc gia không?*
- *Tiền sử bệnh lý cá nhân và gia đình ảnh hưởng như thế nào đến nguy cơ đau tim?*

II/ *Mô tả dữ liệu*

Nguồn dữ liệu

Dữ liệu được lấy từ Heart Attack Risk Prediction Dataset

Mô tả dữ liệu

- ***Số bản ghi:*** *8,763.*
- ***Số cột:*** *26.*

- ***Giải thích các biến:***
  - ***Patient ID****: Mã định danh duy nhất của mỗi bệnh nhân. Không có ý nghĩa phân tích, thường được dùng để quản lý dữ liệu.*
  - ***Age****: Tuổi của bệnh nhân (tính bằng năm).*
  - ***Sex****: Giới tính của bệnh nhân (Male - Nam, Female - Nữ)*
  - ***Cholesterol****: Mức cholesterol của bệnh nhân (mg/dL).* ***Blood Pressure****: Huyết áp của bệnh nhân (mmHg, dạng "Systolic/Diastolic" (huyết áp tâm thu/ tâm trương)).*
  - ***Heart Rate****: Nhịp tim của bệnh nhân (nhịp/phút).*
  - ***Diabetes****: Có đang bị bệnh tiểu đường không (0: Không bị, 1: Có bị).*
  - ***Family History****: Tiền sử gia đình có mắc bệnh tim (0: Không, 1: Có). Yếu tố di truyền có thể làm tăng nguy cơ đau tim.*
  - ***Smoking****: Thói quen hút thuốc của bệnh nhân (0: Không hút thuốc, 1: Có hút thuốc).*
  - ***Obesity****: Tình trạng béo phì của bệnh nhân (0: Không béo phì, 1: Béo phì)..*
  - ***Alcohol Consumption****: Mức tiêu thụ rượu bia của bệnh nhân (0: Không uống, 1: Có uống).*
  - ***Exercise Hours Per Week****: Số giờ luyện tập thể dục của bệnh nhân trong một tuần (giờ).*
  - ***Diet****: Chế độ ăn uống của bệnh nhân (Healthy: Lành mạnh, Unhealthy: Không lành mạnh).*
  - ***Previous Heart Problems****: Tiền sử từng gặp vấn đề về tim (0: Không, 1: Có). Tiền sử bệnh tim làm tăng nguy cơ tái phát.*
  - ***Medication Use****: Sử dụng thuốc điều trị (0: Không sử dụng, 1: Có sử dụng).*
  - ***Stress Level****: Mức độ căng thẳng của bệnh nhân (thang đo từ 0 đến 10).*
  - ***Sedentary Hours Per Day****: Số giờ ngồi yên hoặc ít vận động mỗi ngày (giờ).*
  - ***Income****: Mức thu nhập hàng năm của bệnh nhân (USD).*
  - ***BMI****: Chỉ số khối cơ thể (Body Mass Index).*
  - ***Triglycerides****: Mức triglyceride trong máu của bệnh nhân (mg/dL).*
  - ***Physical Activity Days Per Week****: Số ngày bệnh nhân tham gia hoạt động thể chất trong một tuần.*
  - ***Sleep Hours Per Day****: Số giờ ngủ trung bình mỗi ngày của bệnh nhân.*
  - ***Country****: Quốc gia nơi bệnh nhân sinh sống.*
  - ***Continent****: Châu lục nơi bệnh nhân sinh sống (VD: Asia, Europe, North America).*
  - ***Hemisphere****: Bán cầu nơi bệnh nhân sinh sống (Northern Hemisphere: Bắc bán cầu, Southern Hemisphere: Nam bán cầu).*
  - ***Heart Attack Risk****: Nguy cơ đau tim (0: Nguy cơ thấp, 1: Nguy cơ cao). Đây là biến mục tiêu để phân tích và dự đoán.*

Đọc dữ liệu

```python
data = pd.read_csv('Dheart_attack_prediction_dataset.csv')
```

```
data.head()
```
Python

	Patient ID	Age	Sex	Cholesterol	Blood Pressure	Heart Rate	Diabetes	Family History	Smoking	Obesity	...	Sedentary Hours Per Day	Income
0	BMW7812	67	Male	208	158/88	72	0	0	1	0	...	6.615001	261404
1	CZE1114	21	Male	389	165/93	98	1	1	1	1	...	4.963459	285768
2	BNI9906	21	Female	324	174/99	72	1	0	0	0	...	9.463426	235282
3	JLN3497	84	Male	383	163/100	73	1	1	1	0	...	7.648981	125640
4	GFO8847	66	Male	318	91/88	93	1	1	1	1	...	1.514821	160555

5 rows × 26 columns

BMI	Triglycerides	Physical Activity Days Per Week	Sleep Hours Per Day	Country	Continent	Hemisphere	Heart Attack Risk
31.251233	286	0	6	Argentina	South America	Southern Hemisphere	0
27.194973	235	1	7	Canada	North America	Northern Hemisphere	0
28.176571	587	4	4	France	Europe	Northern Hemisphere	0
36.464704	378	3	4	Canada	North America	Northern Hemisphere	0
21.809144	231	1	5	Thailand	Asia	Northern Hemisphere	0

Các bước tiến hành

1. Xử lý dữ liệu (Data Cleaning)

1.1 Nhận xét qua về bộ dữ liệu

- **Kiểu dữ liệu:**
  - Kiểu float64: Exercise Hours Per Week, Sedentary Hours Per Day, BMI.
  - Kiểu object (str): Patient ID, Sex, Blood Pressure, Diet, Country, Continent, Hemisphere.
  - Kiểu int64: Các cột còn lại.

- **Số giá trị phân biệt của mỗi cột:**

```
data.nunique()
```

```
Patient ID                        8763
Age                                 73
Sex                                  2
Cholesterol                        281
Blood Pressure                    3915
Heart Rate                          71
Diabetes                             2
Family History                       2
Smoking                              2
Obesity                              2
Alcohol Consumption                  2
Exercise Hours Per Week           8763
Diet                                 3
Previous Heart Problems              2
Medication Use                       2
Stress Level                        10
Sedentary Hours Per Day           8763
Income                            8615
BMI                               8763
Triglycerides                      771
Physical Activity Days Per Week      8
Sleep Hours Per Day                  7
Country                             20
Continent                            6
Hemisphere                           2
Heart Attack Risk                    2
dtype: int64
```

*Đảm bảo có 8763 Patient ID phân biệt -> dữ liệu phân biệt hoàn toàn

```
data[data.duplicated()]
```

Patient ID	Age	Sex	Cholesterol	Blood Pressure	Heart Rate	Dia

0 rows × 26 columns

Không tồn tại giá trị trùng lặp.

- **Số giá trị bằng null:**

```
data.isnull().sum()
```

```
Patient ID                        0
Age                               0
Sex                               0
Cholesterol                       0
Blood Pressure                    0
Heart Rate                        0
Diabetes                          0
Family History                    0
Smoking                           0
Obesity                           0
Alcohol Consumption               0
Exercise Hours Per Week           0
Diet                              0
Previous Heart Problems           0
Medication Use                    0
Stress Level                      0
Sedentary Hours Per Day           0
Income                            0
BMI                               0
Triglycerides                     0
Physical Activity Days Per Week   0
Sleep Hours Per Day               0
Country                           0
Continent                         0
Hemisphere                        0
Heart Attack Risk                 0
dtype: int64
```

*Đảm bảo không tồn tại giá trị null, rỗng.

1.2 Loại bỏ dữ liệu trùng lặp, rỗng

- **Các hàng trùng lặp:**
  Không tồn tại các giá trị trùng lặp.

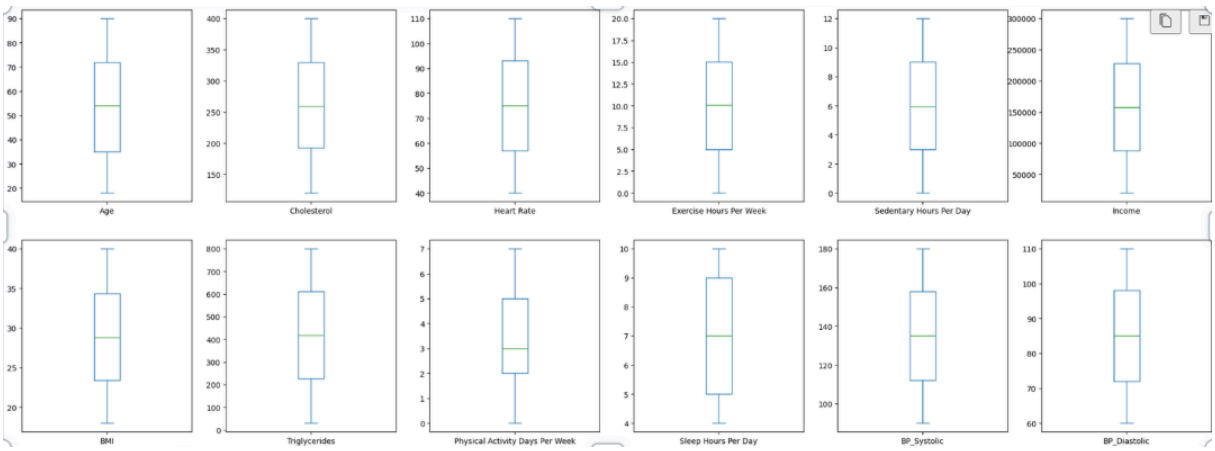
- **Về giá trị rỗng, null**
  Không tồn tại giá trị null.

1.3 Xử lý giá trị ngoại lai

- **Tiền xử** lý
  Ta sẽ tách cột Huyết áp (Blood Pressure) thành 2 cột là huyết áp tâm trương (BP Diastolic) và huyết áp tâm thu (SBP Systolic) kiểu int để xử lý ngoại lai như dạng số.

- **Xử lý ngoại lai dạng số**

  Tạo biểu đồ boxplot để phát hiện giá trị ngoại lai



  Xét thấy tất cả 12 cột kiểu số của bộ dữ liệu đều không tồn tại giá trị ngoại lai theo chuẩn IQR (giá trị nằm ngoài khoảng Q1 - 1,5*IQR và Q3 + 1,5*IQR là ngoại lai với IQR là trung vị, Q1 và Q3 là tứ phân vị thứ 1 và thứ 3)
  => Không có giá trị ngoại lai

- **Xử lý ngoại lai dạng object (str)**

  - Ở phần này, ta lập một dict các giá trị hợp lệ cho mỗi cột (ví dụ: cột Sex chỉ được chứa 2 giá trị là “Male” và “Female”)

  - Ta sẽ duyệt lần lượt các giá trị trong cột và kiểm tra xem giá trị có nằm trong dict hợp lệ hay không, nếu không thì thêm chỉ số vào 1 list các giá trị không hợp lệ và sẽ bị xóa ngay sau khi duyệt xong cột.

  - Quá trình:

```python
myCountries = []
with open("data\countries.txt", 'r') as file:
    for line in file:
        elements = line.strip()
        myCountries.append(elements)

lstKollekzone = {
 'Sex' : ['Male', 'Female'],
 'Diet' : ['Healthy', 'Average', 'Unhealthy'],
 'Continent' : ['South America', 'North America',
                'Asia', 'Europe', 'Africa', 'Australia'],
 'Country' : myCountries,
 'Hemisphere' : ['Northern Hemisphere', 'Southern Hemisphere']
}

for key in lstKollekzone:
    index = 0
    listDrop = []
    for x in data[key]:
        index += 1
        if x not in lstKollekzone[key]:
            listDrop.append(index)
            print(key, index)
    for i in listDrop:
        data.drop(data.iloc[index])
```

0.0s Python

nhận thấy không có bất kỳ giá trị vi phạm nào được in ra.

* Cuối cùng, toàn bộ dữ liệu đã được làm sạch và kiểm tra lại sẽ được xuất ra file ’data/exported_file.csv’

Kết luận: dữ liệu khá sạch, không có giá trị null, hàm trùng lặp, không có ngoại lai.

2. Khám phá dữ liệu (Data Exploration)

2.1 Phân tích thống kê cơ bản

Tính các giá trị trung bình, trung vị, và phân phối của các biến chính.

```
data.describe().T
```

	count	mean	std	min	25%	50%	75%	max
Age	8763.0	53.707977	21.249509	18.000000	35.000000	54.000000	72.000000	90.000000
Cholesterol	8763.0	259.877211	80.863276	120.000000	192.000000	259.000000	330.000000	400.000000
Heart Rate	8763.0	75.021682	20.550948	40.000000	57.000000	75.000000	93.000000	110.000000
Diabetes	8763.0	0.652288	0.476271	0.000000	0.000000	1.000000	1.000000	1.000000
Family History	8763.0	0.492982	0.499979	0.000000	0.000000	0.000000	1.000000	1.000000
Smoking	8763.0	0.896839	0.304186	0.000000	1.000000	1.000000	1.000000	1.000000
Obesity	8763.0	0.501426	0.500026	0.000000	0.000000	1.000000	1.000000	1.000000
Alcohol Consumption	8763.0	0.598083	0.490313	0.000000	0.000000	1.000000	1.000000	1.000000
Exercise Hours Per Week	8763.0	10.014284	5.783745	0.002442	4.981579	10.069559	15.050018	19.998709
Previous Heart Problems	8763.0	0.495835	0.500011	0.000000	0.000000	0.000000	1.000000	1.000000
Medication Use	8763.0	0.498345	0.500026	0.000000	0.000000	0.000000	1.000000	1.000000
Stress Level	8763.0	5.469702	2.859622	1.000000	3.000000	5.000000	8.000000	10.000000
Sedentary Hours Per Day	8763.0	5.993690	3.466359	0.001263	2.998794	5.933622	9.019124	11.999313
Income	8763.0	158263.181901	80575.190806	20062.000000	88310.000000	157866.000000	227749.000000	299954.000000
BMI	8763.0	28.891446	6.319181	18.002337	23.422985	28.768999	34.324594	39.997211
Triglycerides	8763.0	417.677051	223.748137	30.000000	225.500000	417.000000	612.000000	800.000000
Physical Activity Days Per Week	8763.0	3.489672	2.282687	0.000000	2.000000	3.000000	5.000000	7.000000
Sleep Hours Per Day	8763.0	7.023508	1.988473	4.000000	5.000000	7.000000	9.000000	10.000000
BP_Systolic	8763.0	135.075659	26.349976	90.000000	112.000000	135.000000	158.000000	180.000000
BP_Diastolic	8763.0	85.156111	14.676565	60.000000	72.000000	85.000000	98.000000	110.000000
Heart Attack Risk	8763.0	0.358211	0.479502	0.000000	0.000000	0.000000	1.000000	1.000000

- Độ tuổi tham gia khảo sát: từ 18 đến 90
- Nhận thấy trung vị của biến Heart Attack Risk là giá trị 0. Có thể kết luận có hơn một nửa số người tham gia khảo sát có nguy cơ mắc bệnh tim mạch**.** Tuy nhiên tứ phân vị thứ ba là 1 cho thấy tỷ lệ giữa giá trị 0 và 1 không lệch nhau quá nhiều.

2.2 Phân tích dữ liệu thăm dò

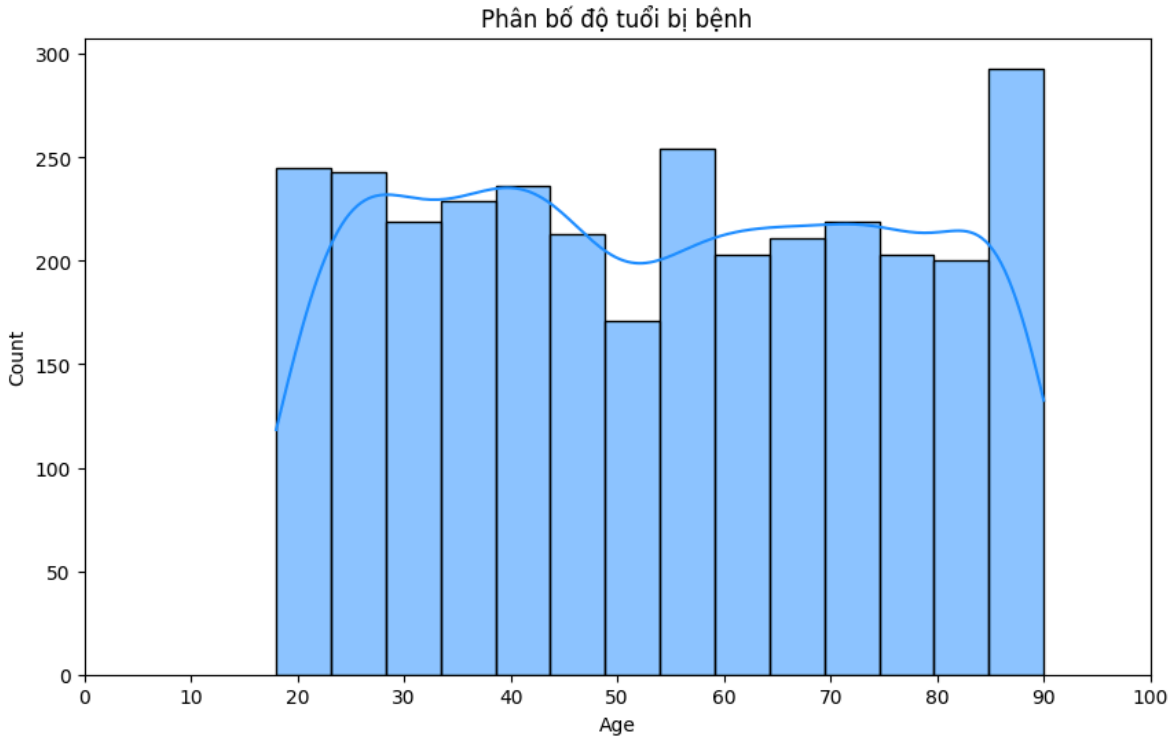
```
data.columns
```

Python

```
Index(['Patient ID', 'Age', 'Sex', 'Cholesterol', 'Blood Pressure',
       'Heart Rate', 'Diabetes', 'Family History', 'Smoking', 'Obesity',
       'Alcohol Consumption', 'Exercise Hours Per Week', 'Diet',
       'Previous Heart Problems', 'Medication Use', 'Stress Level',
       'Sedentary Hours Per Day', 'Income', 'BMI', 'Triglycerides',
       'Physical Activity Days Per Week', 'Sleep Hours Per Day', 'Country',
       'Continent', 'Hemisphere', 'Heart Attack Risk'],
      dtype='object')
```

Có tổng 26 cột, trong đó cột ‘Patient ID’ là mã bệnh nhân, không cần phân tích.

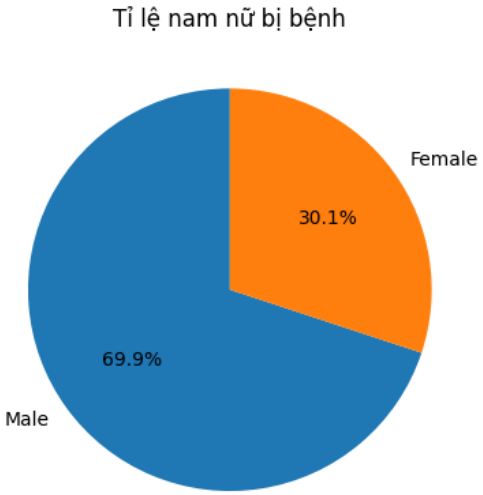
Trước hết ta sẽ vẽ đồ thị phân bố độ tuổi bị bệnh:



Nhận xét:

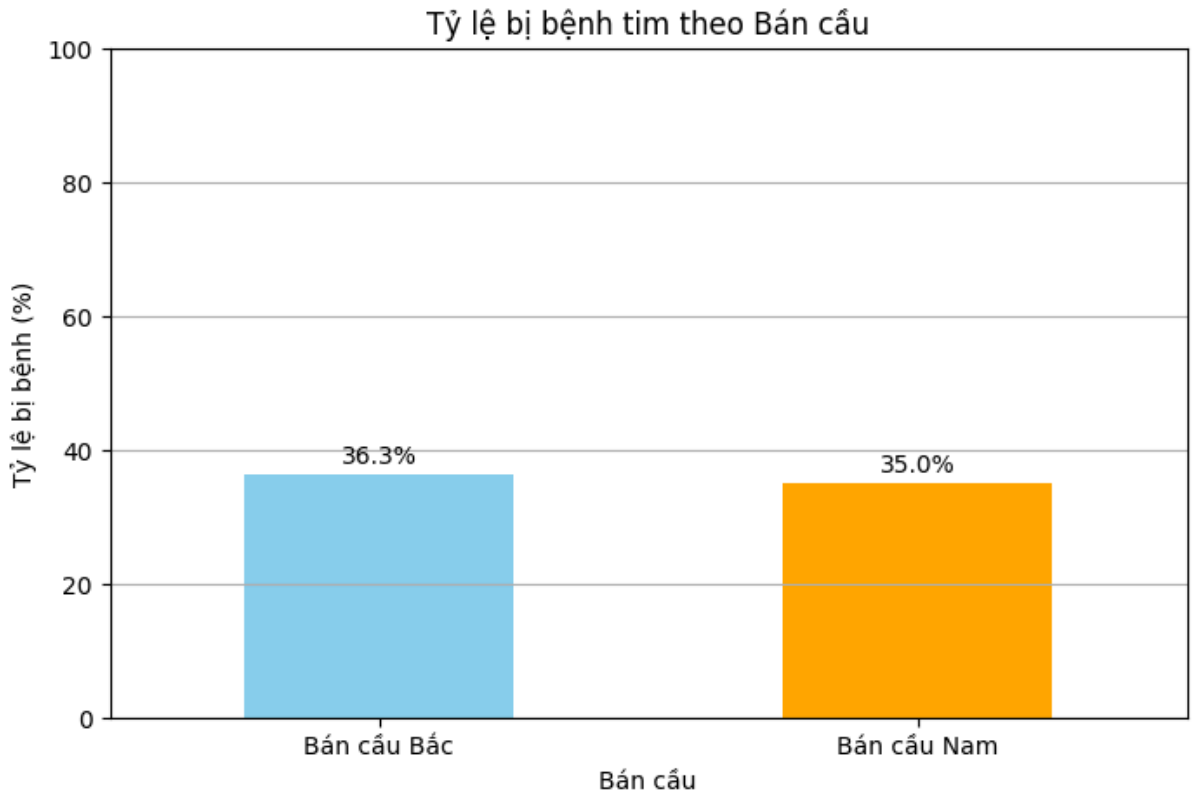
- Độ tuổi bị tim chiếm % cao nhất là 90 tuổi
- Người trẻ cũng có nguy cơ bị tim mạch. Thậm chí người khoảng 25-40 tuổi bị bệnh chiếm số lượng cao hơn so với người từ 60 đến 85 tuổi

Phân tích giới tính:



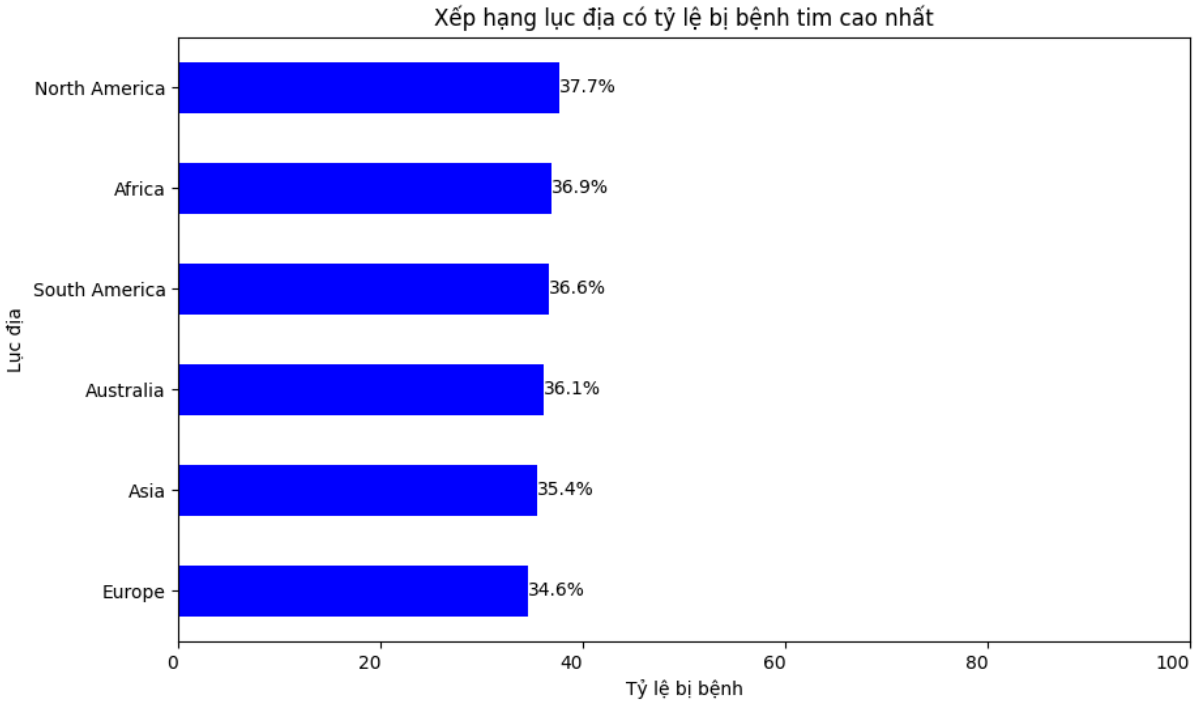
=> Trong số những người có nguy cơ cao bị bệnh tim, nam giới chiếm (69.9%), nhiều hơn so với nữ giới (30.1%)

Phân loại theo bán cầu:



Nhận xét: tỉ lệ bệnh nhân bị tim của 2 bán cầu là gần như nhau

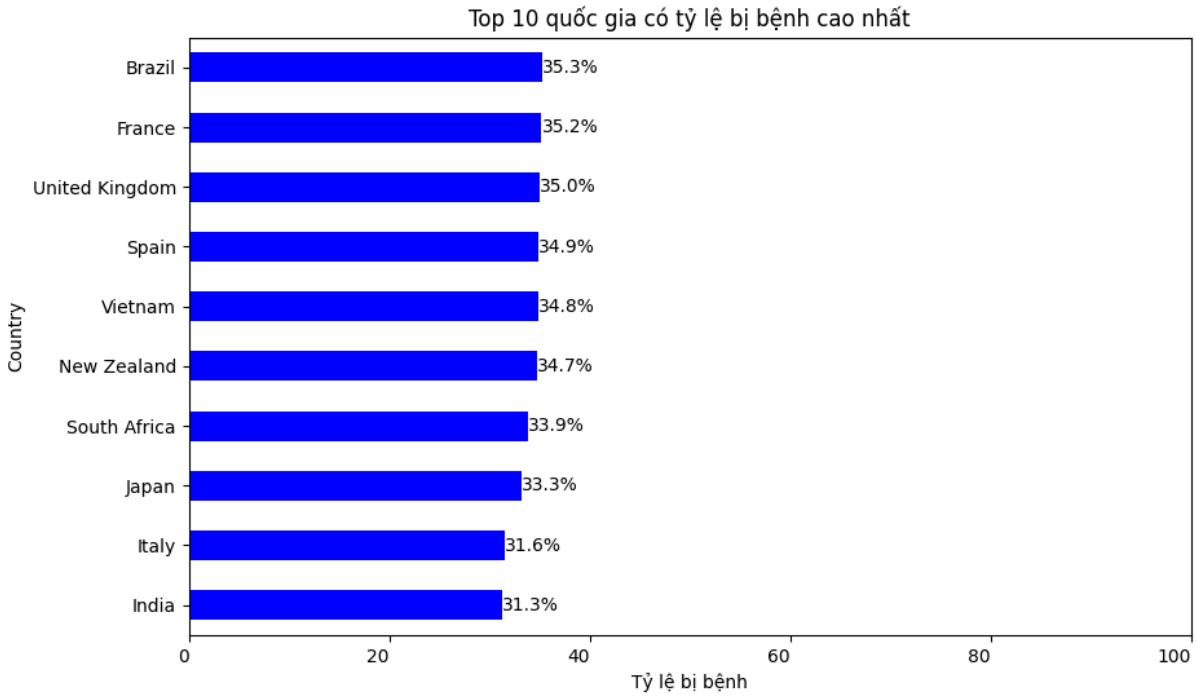
Phân loại theo Lục địa:



Nhận xét:

- Bắc Mĩ đứng đầu với 37.7%, cuối cùng là châu Âu (34.6%). Châu Á gần cuối với 35.4%.
- Tỷ lệ nguy cơ mắc bệnh tim khá gần nhau, dao động từ 34.6% - 37.7%.

Top 10 quốc gia có tỷ lệ bị bệnh cao nhất:



Nhận xét:

- Brazil top đầu (35.3%)
- Việt Nam top giữa (34.8%)
- Tỷ lệ mắc bệnh của các quốc gia top đầu không lệch nhau quá nhiều

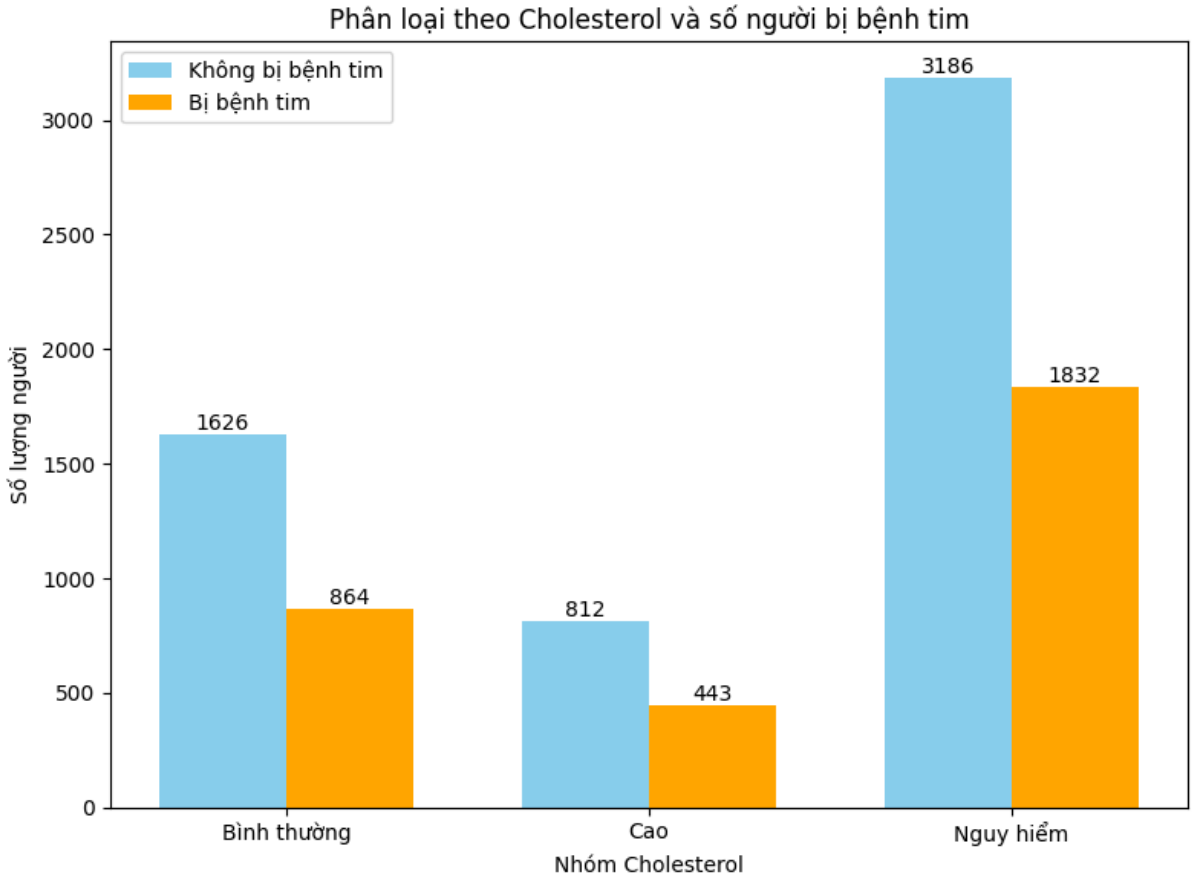
Phân loại theo Cholesterol:

Tham khảo từ nguồn : Kết quả xét nghiệm định lượng cholesterol như thế nào là bình thường? | Vinmec

Cholesterol từ 0 -> 200: Bình thường

Cholesterol từ 200 -> 240: Cao

Cholesterol từ 240 trở lên: Rủi ro sức khỏe



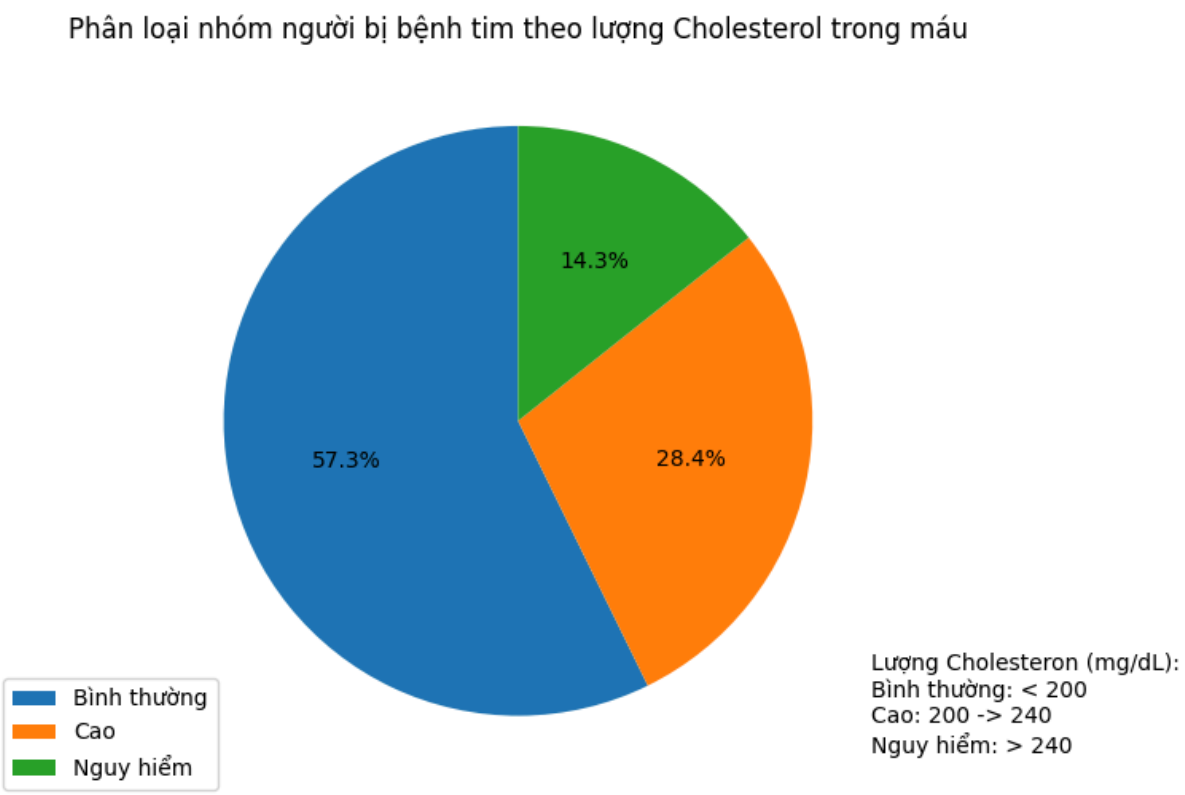
+ Số lượng người không bị bệnh tim luôn cao hơn số người bị bệnh tim ở cả ba nhóm Cholesterol. Điều này cho thấy rằng dù mức Cholesterol như thế nào, phần lớn vẫn không bị bệnh tim (ít nhất là trong mẫu đang được phân tích).
+ Nhóm Cholesterol "Nguy hiểm" có số lượng người (cả bị bệnh và không bị bệnh tim) cao nhất. Cụ thể, nhóm này có 3186 người không bị bệnh tim và 1832 người bị bệnh tim. Điều này cho thấy đa số những người tham gia khảo sát có mức cholesterol "nguy hiểm"
+ Tỷ lệ người bị bệnh tim có sự chênh lệch giữa các nhóm Cholesterol:

  Nhóm "Bình thường": 864 / (1626 + 864) ≈ 34.7%

  Nhóm "Cao": 443 / (812 + 443) ≈ 35.3%

  Nhóm "Nguy hiểm": 1832 / (3186 + 1832) ≈ 36.5%

  Tỷ lệ người bị bệnh tim tăng dần theo mức độ Cholesterol: Mặc dù mức tăng không quá lớn nhưng chúng ta có thể thấy một xu hướng: tỷ lệ người mắc bệnh tim tăng dần từ nhóm "Bình thường" (34.7%), lên "Cao" (35.3%), và sau đó là "Nguy hiểm" (36.5%).



+ Phần lớn những người bị bệnh tim trong mẫu nghiên cứu có mức Cholesterol "Bình thường" (57.3%). Điều này khá bất ngờ vì thường Cholesterol cao được coi là một yếu tố nguy cơ của bệnh tim.
+ Nhóm "Cao" chiếm 28.4% và nhóm "Có vấn đề" chiếm 14.3% trong số những người bị bệnh tim. Tổng cộng hai nhóm này chiếm 42.7% số người bị bệnh tim.

Kết luận:

- Biểu đồ cho thấy hơn một nửa số người bị bệnh tim trong mẫu nghiên cứu có mức Cholesterol "Bình thường" (dưới 200 mg/dL). Điều này ngụ ý rằng **Cholesterol thấp không đảm bảo không bị bệnh tim.**
- Gần một nửa (42.7%) số người bị bệnh tim có mức Cholesterol "Cao" hoặc "Có vấn đề". Điều này vẫn cho thấy Cholesterol cao có liên quan đến bệnh tim mạch.
- Cần lưu ý rằng đây chỉ là mối tương quan, không phải là mối quan hệ nhân quả. Không thể kết luận chắc chắn rằng Cholesterol cao trực tiếp gây ra bệnh tim dựa trên biểu đồ này. Cần thêm nhiều nghiên cứu và phân tích sâu hơn để khẳng định điều này. Có thể có nhiều yếu tố khác ảnh hưởng đến cả mức cholesterol và nguy cơ mắc bệnh tim.

Phân loại theo nhịp tim:

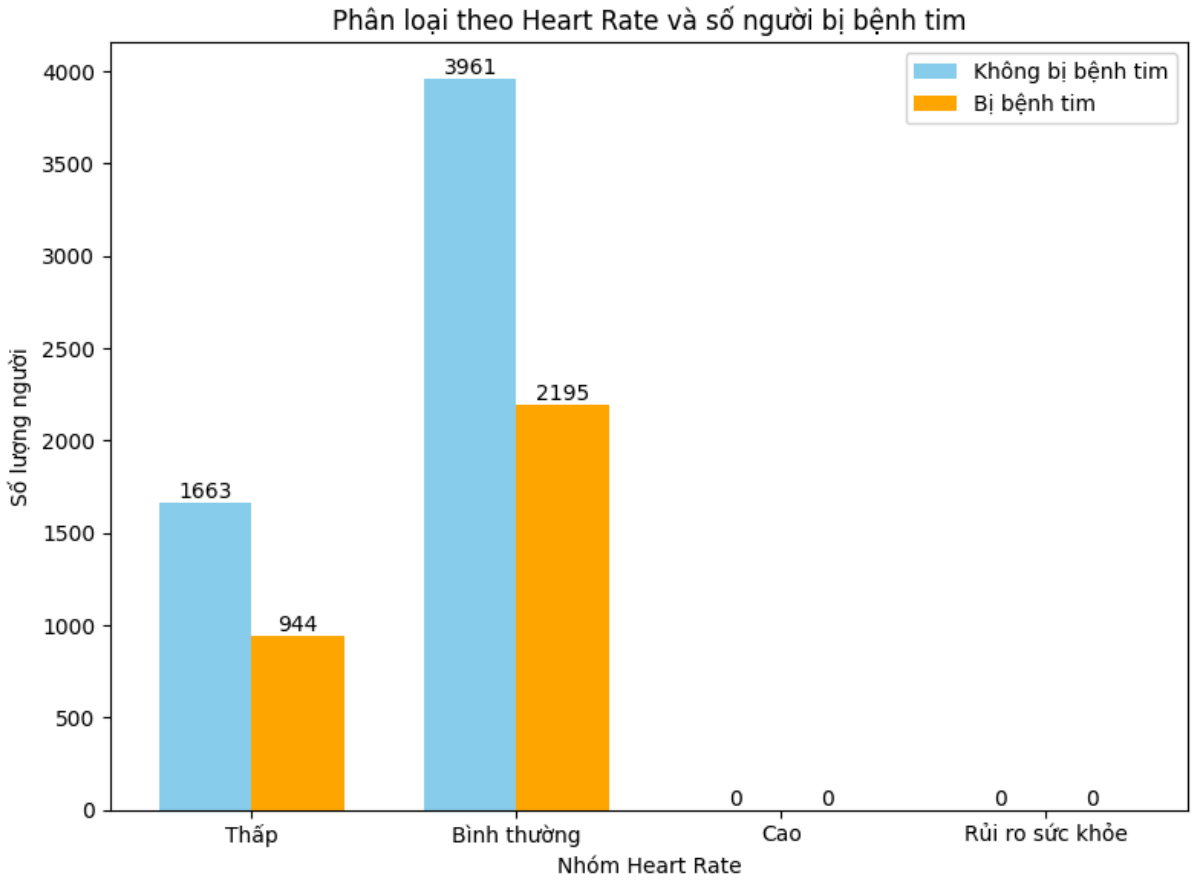
Nguồn : Nhịp tim chuẩn là bao nhiêu? | Vinmec

Nhịp tim từ 0 -> 60: Thấp

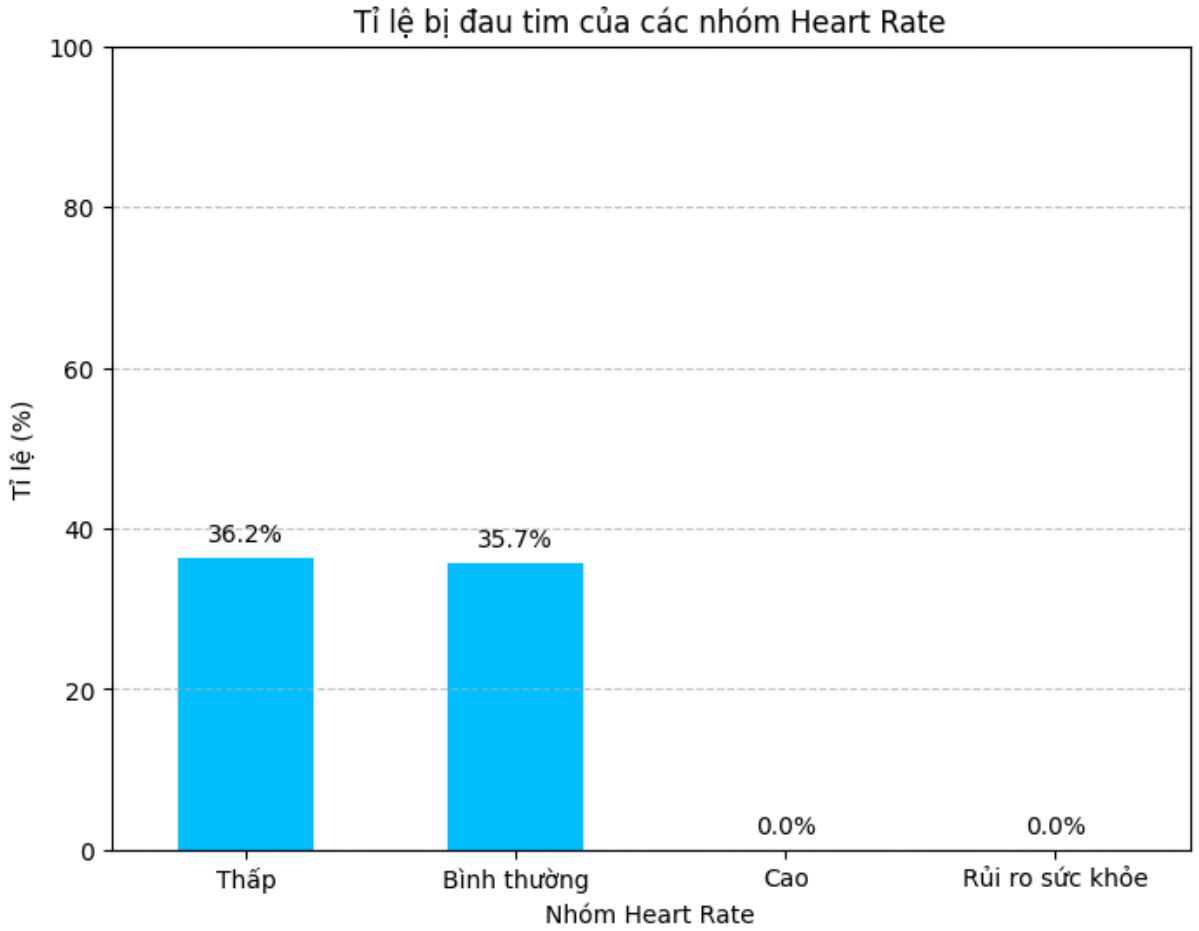
Nhịp tim từ 60 -> 125: Bình thường

Nhịp tim từ 125 -> 140: Cao

Nhịp tim từ 140 trở lên : Rủi ro sức khỏe



+ Nhóm Heart Rate "Bình thường" có số lượng người ở cả hai nhóm (bị bệnh tim và không bị bệnh tim) cao nhất. Cụ thể, nhóm này có 3961 người không bị bệnh tim và 2195 người bị bệnh tim.
+ Không có ai ở nhóm "Cao" và "Rủi ro sức khỏe" trong cả hai nhóm (bị bệnh tim và không bị bệnh tim).
+ Ở cả hai nhóm "Thấp" và "Bình thường", số người không bị bệnh tim đều cao hơn số người bị bệnh tim.

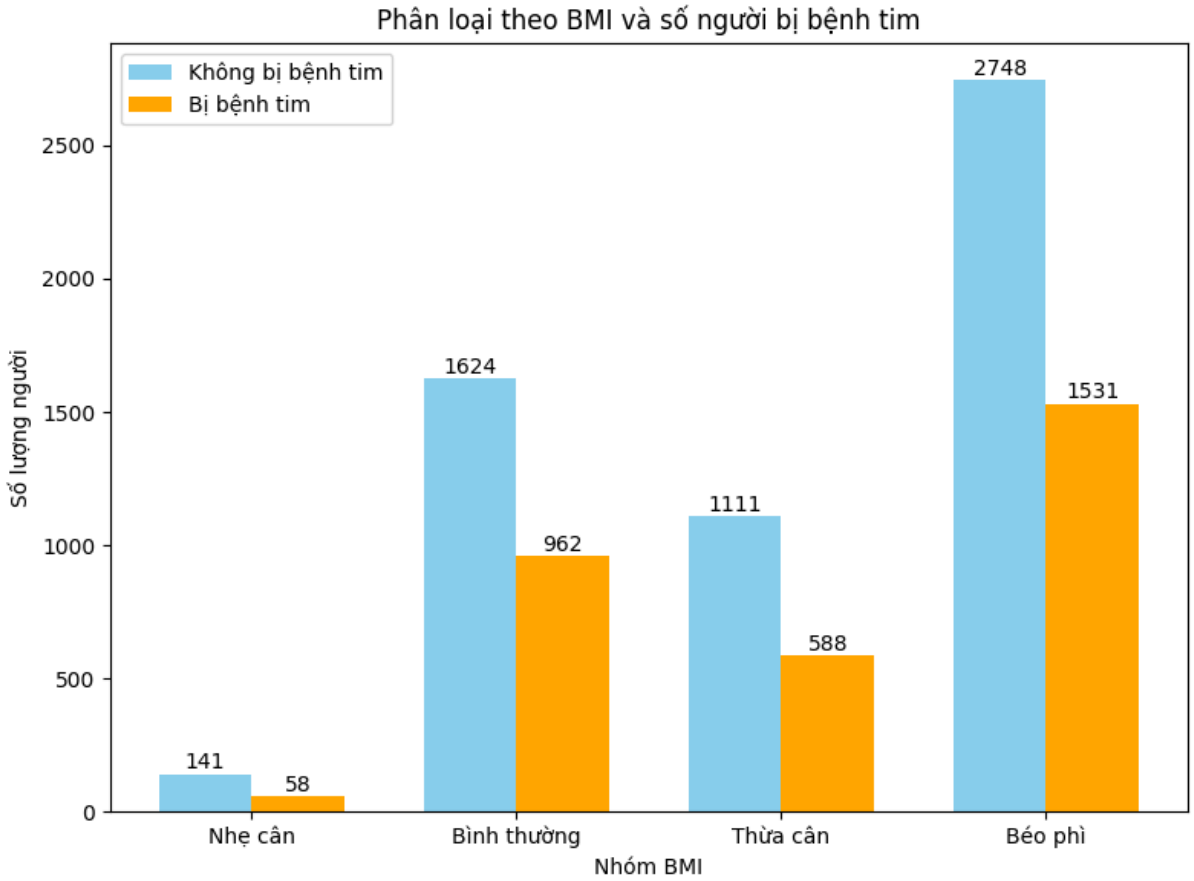


+ Tỉ lệ bị đau tim ở nhóm Heart Rate "Thấp" và "Bình thường" gần như tương đương nhau, lần lượt là 36.2% và 35.7%.
+ Không có trường hợp nào bị đau tim ở nhóm Heart Rate "Cao" và "Rủi ro sức khỏe". Điều này có nghĩa là trong mẫu nghiên cứu, những người có nhịp tim cao hoặc có rủi ro sức khỏe đều không bị đau tim.

Kết luận:

- Biểu đồ cho thấy tỷ lệ bị đau tim ở những người có nhịp tim "Thấp" và "Bình thường" là gần như nhau (khoảng 36%).
- Không có trường hợp đau tim nào được ghi nhận ở nhóm "Cao" và "Rủi ro sức khỏe" trong mẫu nghiên cứu. Điều này khá bất ngờ và có thể do cỡ mẫu nhỏ.
- Thiếu thông tin định nghĩa cụ thể về các nhóm Heart Rate gây khó khăn cho việc phân tích. Không rõ ngưỡng phân loại giữa các nhóm là như thế nào, và nhóm "Rủi ro sức khỏe" đề cập đến các yếu tố cụ thể nào.

Phân tích về BMI:



+ So sánh: Trong cả hai nhóm (bị bệnh tim và không bị bệnh tim), nhóm "Béo phì" luôn có số lượng người cao nhất. Tuy nhiên, tỷ lệ người bị bệnh tim trong nhóm "Béo phì" (1531/4279 = 35.8%) cao hơn so với tỷ lệ người không bị bệnh tim trong nhóm này (2748/4279 = 64.2%).
+ Ở tất cả các nhóm BMI, số người không bị bệnh tim đều cao hơn số người bị bệnh tim.
+ Nhóm "Béo phì" có số lượng người ở cả hai nhóm (bị bệnh tim và không bị bệnh tim) cao nhất. Nhóm "Nhẹ cân" có số lượng người ở cả hai nhóm thấp nhất. Nhóm này có 141 người không bị bệnh tim và 58 người bị bệnh tim.
+ Tỉ lệ người bị bệnh tim trong các nhóm:

  Nhẹ cân: 58 / (141 + 58) ≈ 29.1%

  Bình thường: 962 / (1624 + 962) ≈ 37.2%

  Thừa cân: 588 / (1111 + 588) ≈ 34.6%

  Béo phì: 1531 / (2748 + 1531) ≈ 35.8%

Kết luận:

- Đa số những người tham gia khảo sát có chỉ số BMI thuộc nhóm "Béo phì" và "Bình thường".
- Tỉ lệ bị bệnh tim cao nhất ở nhóm "Bình thường" (37.2%), theo sau là "Béo phì" (35.8%), "Thừa cân" (34.6%) và "Nhẹ cân" (29.1%). Nhóm “Bình thường” cao nhất là do mẫu khảo sát ít.

- Nhìn chung, biểu đồ cho thấy béo phì có thể là một yếu tố nguy cơ của bệnh tim

Phân tích huyết áp:

Tham khảo từ nguồn :[Chỉ số huyết áp bình thường và phân loại Tăng huyết áp của ESC 2018 | Vinmec]

- Huyết áp tối ưu: Huyết áp tâm thu < 120 mmHg và huyết áp tâm trương < 80 mmHg.
- Huyết áp bình thường: Huyết áp tâm thu 120-129 mmHg và/hoặc huyết áp tâm trương 80-84 mmHg.
- Huyết áp bình thường cao: Huyết áp tâm thu 130-139 mmHg và/hoặc huyết áp tâm trương 85-89 mmHg.
- Tăng huyết áp độ 1: Huyết áp tâm thu 140-159 mmHg và/hoặc huyết áp tâm trương 90-99 mmHg.
- Tăng huyết áp độ 2: Huyết áp tâm thu 160-179 mmHg và/hoặc huyết áp tâm trương 100-109 mmHg.
- Tăng huyết áp độ 3: Huyết áp tâm thu ≥ 180 mmHg và/hoặc huyết áp tâm trương ≥ 110 mmHg.
- Tăng huyết áp tâm thu đơn độc: Huyết áp tâm thu ≥ 140 mmHg và huyết áp tâm trương < 90mmHg.

+ Tăng huyết áp độ 1 (20.6%): Đây là nhóm chiếm tỷ lệ cao nhất, cho thấy rằng có nhiều bệnh nhân tim mạch đang ở giai đoạn đầu của bệnh tăng huyết áp.
+ Huyết áp bình thường (19.5%): Nhóm này chiếm tỷ lệ gần bằng với nhóm tăng

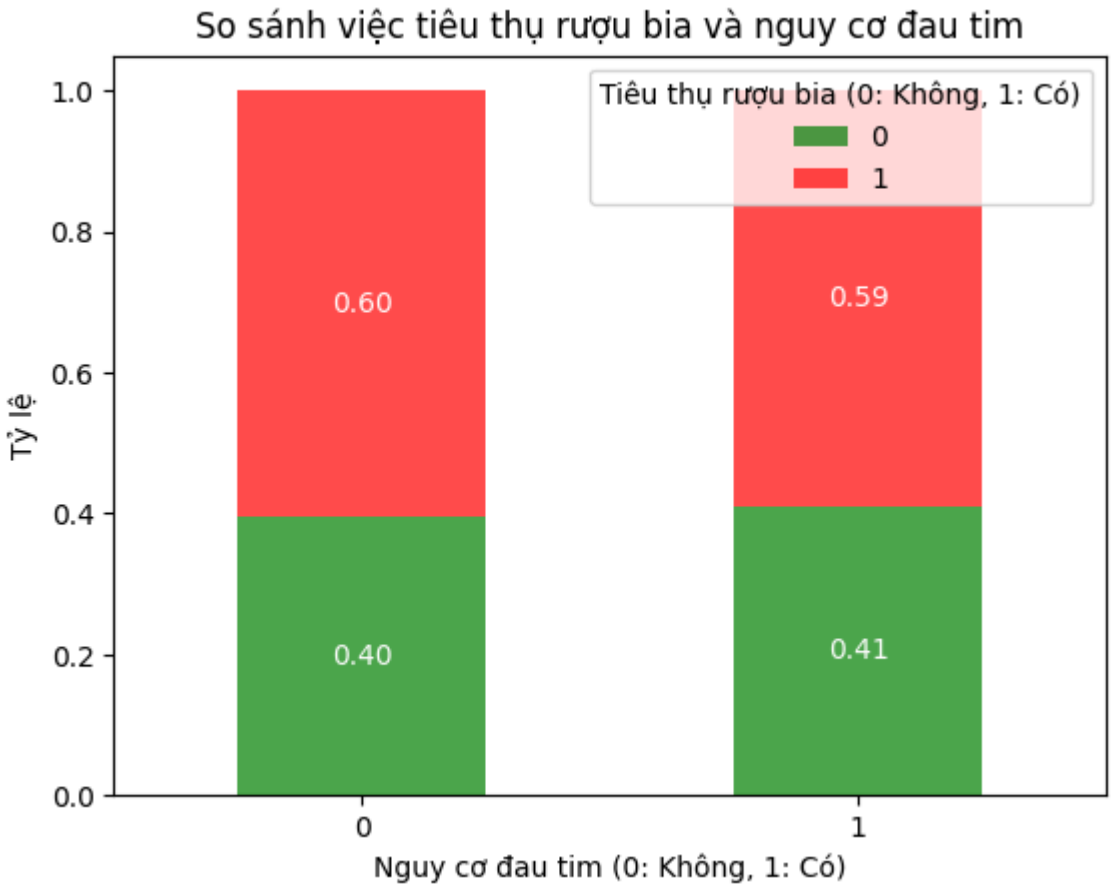
huyết áp độ 1, cho thấy rằng một bộ phận bệnh nhân tim mạch vẫn có huyết áp bình thường.

+ Tăng huyết áp tâm thu đơn độc (17.4%): Nhóm này cho thấy rằng có một lượng đáng kể bệnh nhân tim mạch bị tăng huyết áp tâm thu, trong khi huyết áp tâm trương vẫn bình thường.
+ Huyết áp tối ưu (13.3%): Đây là nhóm có huyết áp lý tưởng, cho thấy rằng một số ít bệnh nhân tim mạch vẫn duy trì được mức huyết áp tốt.
+ Tăng huyết áp độ 2 (11.6%): Nhóm này chiếm tỷ lệ thấp hơn so với tăng huyết áp độ 1, nhưng vẫn cho thấy rằng có một lượng bệnh nhân tim mạch bị tăng huyết áp ở mức độ trung bình.
+ Tăng huyết áp độ 3 (0.6%): Nhóm này chiếm tỷ lệ rất nhỏ, cho thấy rằng có rất ít bệnh nhân tim mạch bị tăng huyết áp ở mức độ nghiêm trọng.

Kết luận:

- Nhìn chung, phần lớn bệnh nhân bị tim trong mẫu khảo sát có vấn đề về huyết áp. Chỉ có 19.5% có huyết áp bình thường.
- Tỷ lệ bệnh nhân bị tăng huyết áp ở các mức độ khác nhau (độ 1, độ 2, độ 3, tâm thu đơn độc) chiếm phần lớn, lên tới gần 50% (20.6% + 17.4% + 11.6% + 0.6%). Điều này cho thấy mối liên hệ mật thiết giữa bệnh tim và tình trạng tăng huyết áp.

Phân tích lối sống

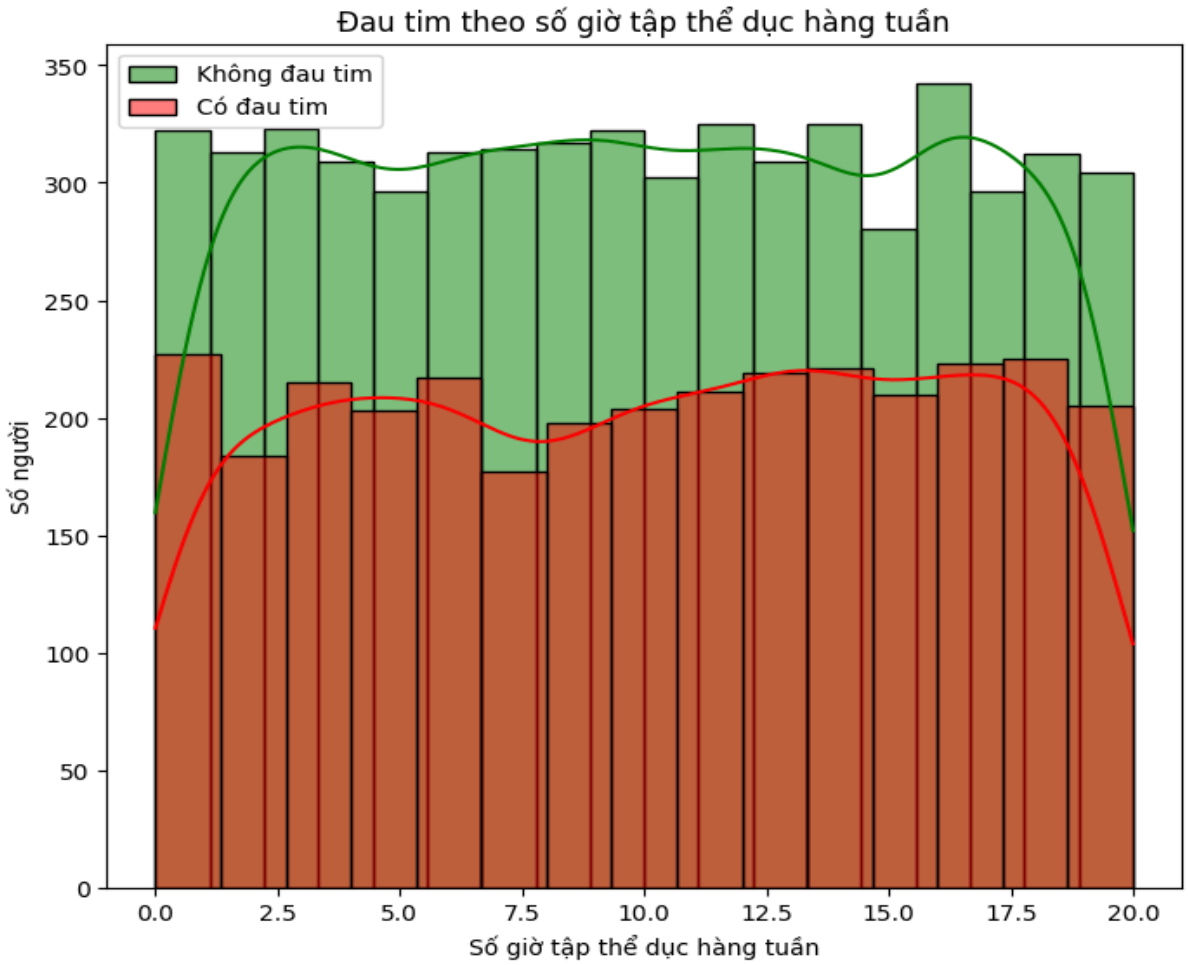


+ Người không có nguy cơ đau tim: Tỷ lệ tiêu thụ rượu bia ở nhóm này là 0.60 (60%), cao hơn một chút so với tỷ lệ không tiêu thụ rượu bia là 0.40 (40%).

+ Người có nguy cơ đau tim: Tỷ lệ tiêu thụ rượu bia ở nhóm này là 0.59 (59%), gần bằng với tỷ lệ không tiêu thụ rượu bia là 0.41 (41%).

Kết luận

- Có thể thấy, việc tiêu thụ rượu bia có gây gia tăng nhẹ tỉ lệ dẫn tới đau tim (tăng lên khoảng 1%).
- Dựa trên biểu đồ này, không có sự khác biệt đáng kể về tỷ lệ tiêu thụ rượu bia giữa nhóm người có nguy cơ đau tim và nhóm người không có nguy cơ đau tim.
- Biểu đồ cho thấy việc tiêu thụ rượu bia không có mối tương quan rõ ràng với nguy cơ đau tim trong mẫu dữ liệu này.



+ Số người không đau tim (màu xanh lá) luôn cao hơn số người có đau tim (màu đỏ) ở mọi mức thời gian tập thể dục.
+ Ở nhóm người không đau tim, nhóm người cao nhất có số giờ tập thể dục hàng tuần ở mức cao (16 giờ). Trong khi ở nhóm có nguy cơ đau tim, số giờ tập thể dục là 1 có số lượng người cao nhất.

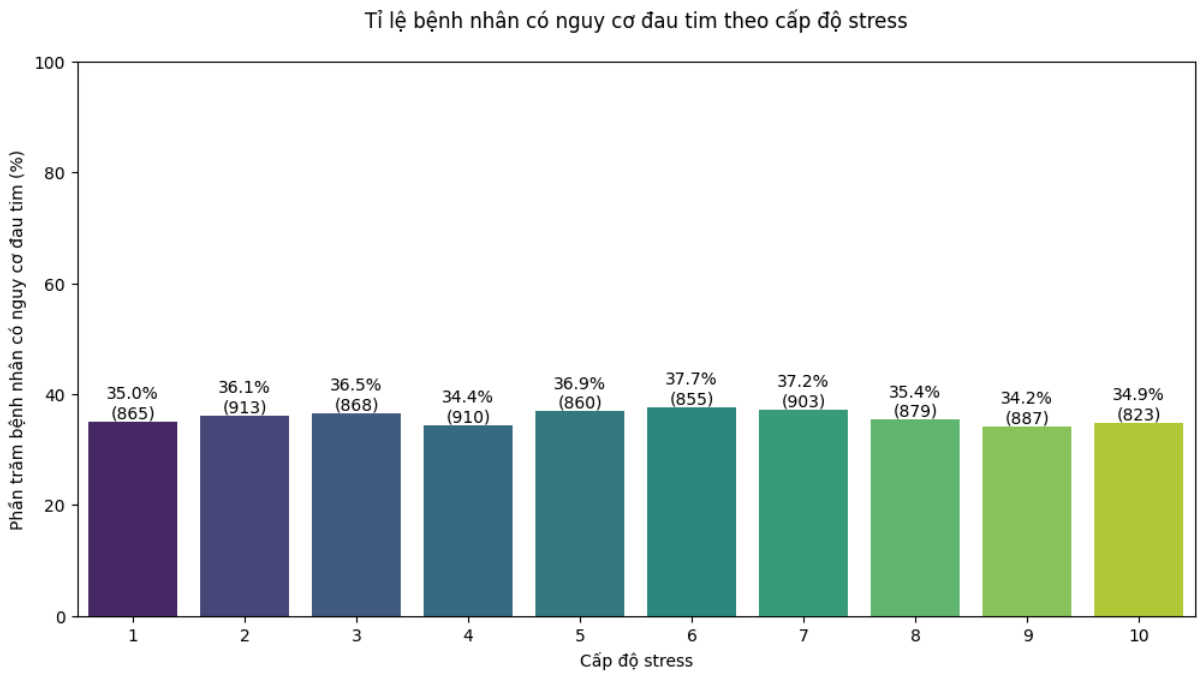
Kết luận:

- Biểu đồ cho thấy những người có tập thể dục có số lượng đông hơn những

người không tập thể dục.

- Nhóm không đau tim luôn chiếm ưu thế về số lượng ở mọi mức thời gian tập luyện.
- Số giờ tập thể dục hàng tuần có một phần ảnh hưởng đến nguy cơ bị bệnh tim mạch.

Kiểm tra Stress Level:

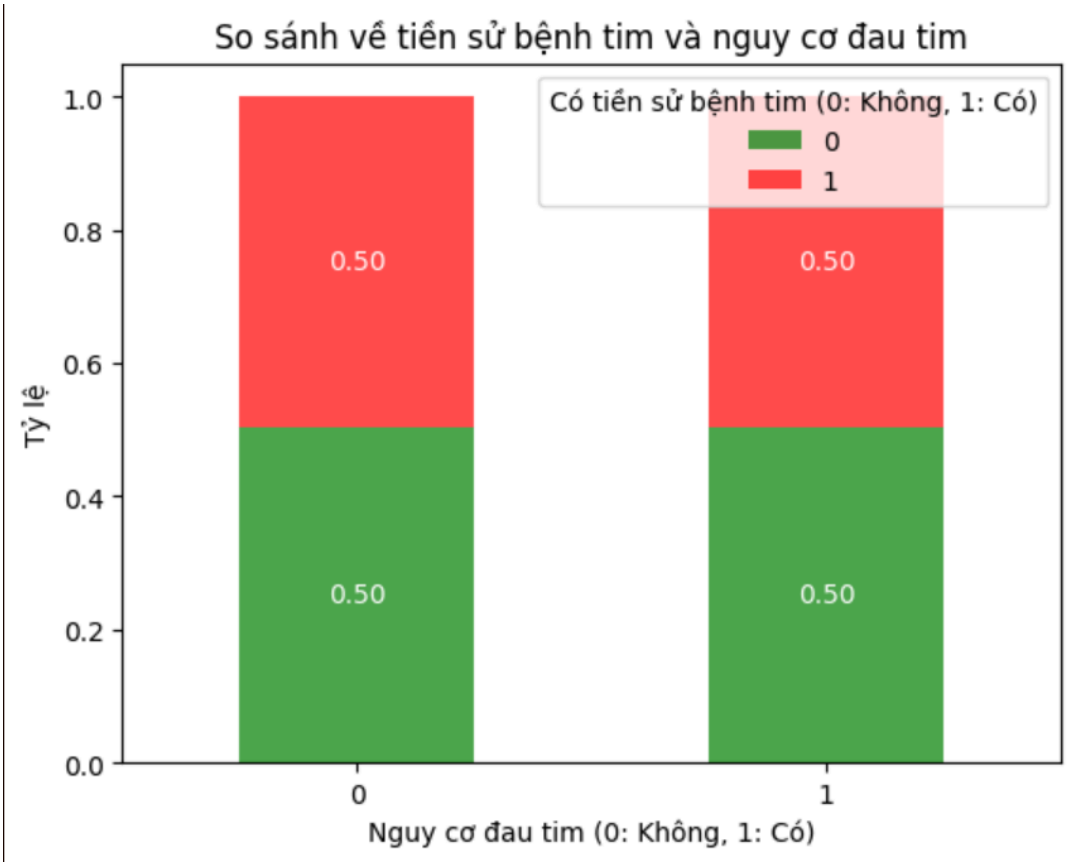


+ Biểu đồ cho thấy một xu hướng tương đối ổn định, không có sự tăng hay giảm rõ rệt theo cấp độ stress.
+ Cấp độ stress 6 có tỷ lệ bệnh nhân có nguy cơ đau tim cao nhất (37.7% với cỡ mẫu 855).
+ Cấp độ stress 9 có tỷ lệ bệnh nhân có nguy cơ đau tim thấp nhất (34.2% với cỡ mẫu 887).
+ Sự chênh lệch giữa tỷ lệ cao nhất và thấp nhất là 3.5% (37.7% - 34.2%)
  -> Không có sự chênh lệch quá lớn về tỷ lệ bệnh nhân có nguy cơ đau tim giữa các cấp độ stress..

Kết luận:

- Dựa trên biểu đồ này, không thể kết luận rằng cấp độ stress có ảnh hưởng trực tiếp và rõ ràng đến nguy cơ đau tim. Tỷ lệ bệnh nhân có nguy cơ đau tim gần như tương đương nhau ở các cấp độ stress khác nhau.
- Cần lưu ý rằng cỡ mẫu ở các cấp độ stress là khá lớn và đồng đều (dao động từ 823 đến 913), điều này làm tăng độ tin cậy của dữ liệu.

Dựa theo tiền sử bệnh lý



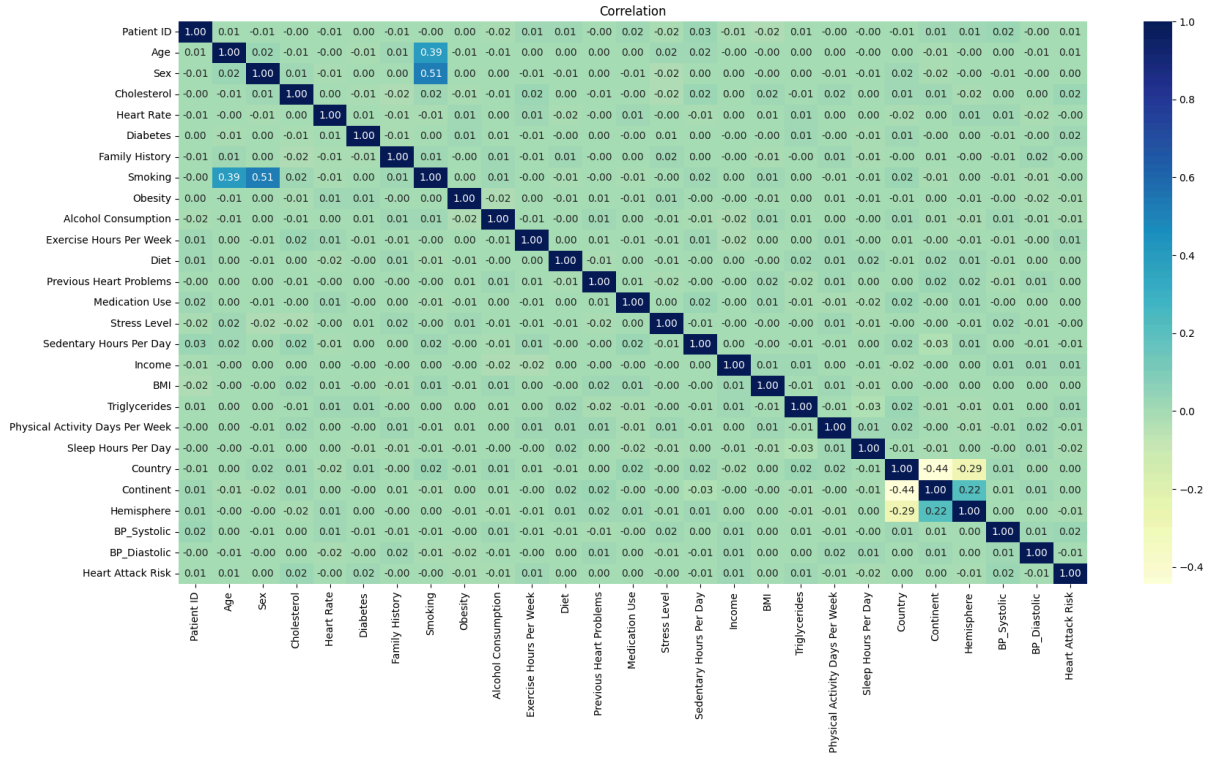
+ Tỷ lệ người không có tiền sử bệnh tim hoàn toàn giống nhau (50%) ở cả hai nhóm có và không có nguy cơ đau tim.
+ Tỷ lệ người có tiền sử bệnh tim cũng hoàn toàn giống nhau (50%) ở cả hai nhóm có và không có nguy cơ đau tim.

Kết luận:

- Dựa trên biểu đồ này, không có sự khác biệt về tỷ lệ tiền sử bệnh tim giữa nhóm người có nguy cơ đau tim và nhóm người không có nguy cơ đau tim.
- Biểu đồ cho thấy tiền sử bệnh tim không có mối tương quan với nguy cơ đau tim trong mẫu dữ liệu này.

2.3 Trực quan hóa dữ liệu

Vẽ biểu đồ heatmap để hiển thị mối quan hệ giữa các biến.



- Có sự tương quan nhẹ ở tình trạng hút thuốc với giới tính và độ tuổi: Độ tuổi càng cao thì tình trạng hút thuốc càng nhiều và nam hút thuốc nhiều hơn nữ. Ngoài ra còn có sự tương quan giữa 3 cột Country, Continent và Hemisphere.
- Các thuộc tính còn lại hầu không có sự tương quan -> độc lập

3. Mô hình dự đoán

3.1. Chuẩn bị dữ liệu:

- Mã hóa các cột về dạng số bằng LabelEncoder:

```
encoder = LabelEncoder()
for col in data.columns:
    if data[col].dtype == 'object':
        data[col] = encoder.fit_transform(data[[col]])
```

- Chia tập dữ liệu thành 2 phần: phần huấn luyện và phần kiểm thử bằng train_test_split() với tỉ lệ train/test là 80/20:

```
X = data.drop(['Patient ID', 'Heart Attack Risk'],
              axis=1,
              inplace=False)
y = data['Heart Attack Risk'].values
x_train, x_test, y_train, y_test = train_test_split(X, y, train_size=0.8, random_state=1)
```

- Chuẩn hóa dữ liệu bằng MinMaxScaler():

```
scaler = MinMaxScaler()
x_train = scaler.fit_transform(x_train)
x_test = scaler.transform(x_test)
```

3.2. Huấn luyện mô hình: Sử dụng 5 mô hình khác nhau:

+ Gaussian Naive Bayes:

```
Confusion matrix:
[[1132   10]
 [ 606    5]]
The accuracy of Gaussian Naive Bayes model is :  64.86023958927552 %
The precision of Gaussian Naive Bayes model is :  33.33333333333333 %
The recall of Gaussian Naive Bayes model is :  0.8183306055646482 %
The f1 score of Gaussian Naive Bayes model is :  1.5974440894568689 %
```

+ Support Vector Machine:

```
Confusion matrix:
[[1142    0]
 [ 611    0]]
The accuracy of Support Vector Machine model is :  65.14546491728466 %
The precision of Support Vector Machine model is :  100.0 %
The recall of Support Vector Machine model is :  0.0 %
The f1 score of Support Vector Machine model is :  0.0 %
```

+ Random Forest:

```
Confusion matrix:
[[1090   52]
 [ 582   29]]
The accuracy of Random Forest model is :  63.83342840844267 %
The precision of Random Forest model is :  35.80246913580247 %
The recall of Random Forest model is :  4.746317512274959 %
The f1 score of Random Forest model is :  8.38150289017341 %
```

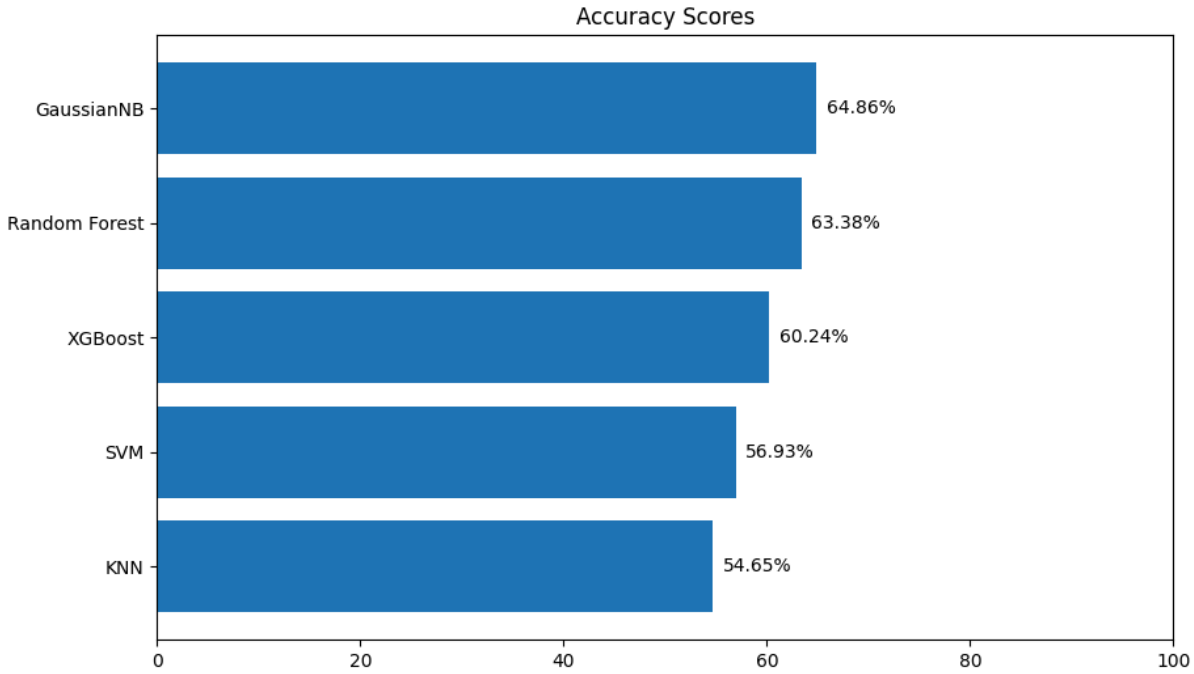
+ XGBoost:

```
Confusion matrix:
[[1117   25]
 [ 599   12]]
The accuracy of XGBoost model is :  64.40387906446092 %
The precision of XGBoost model is :  32.432432432432435 %
The recall of XGBoost model is :  1.9639934533551555 %
The f1 score of XGBoost model is :  3.7037037037037033 %
```

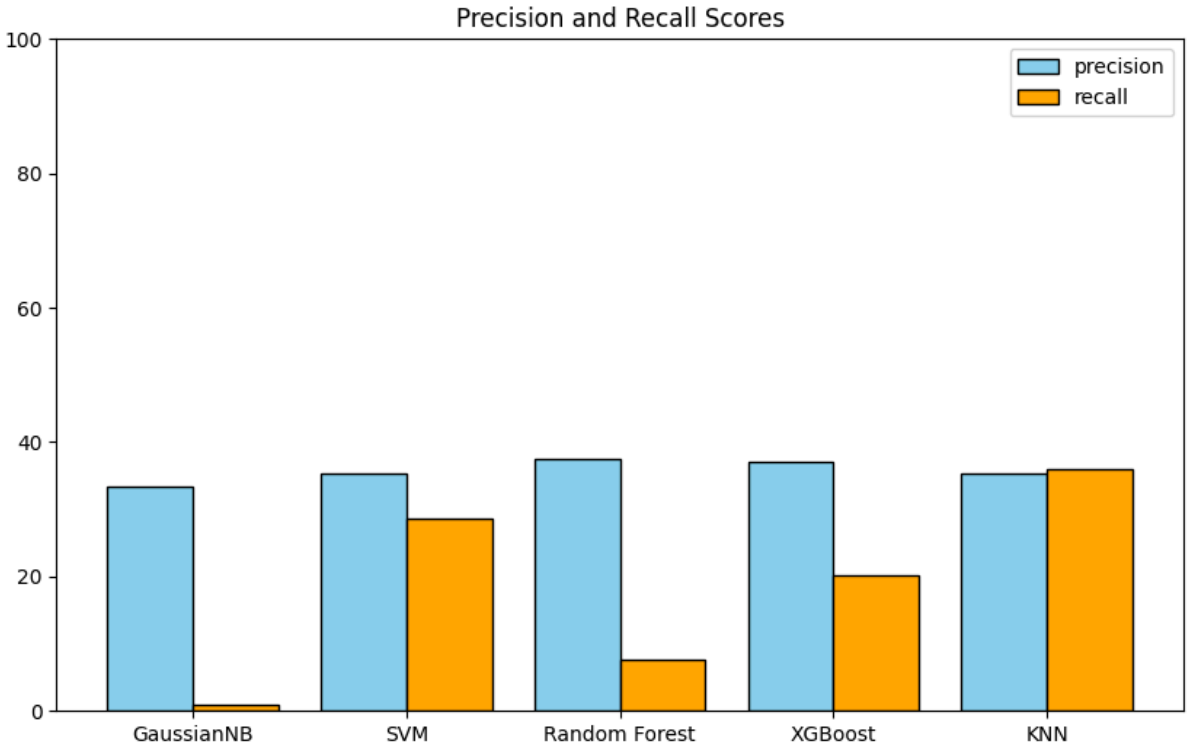
+ K Nearest Neighbor:

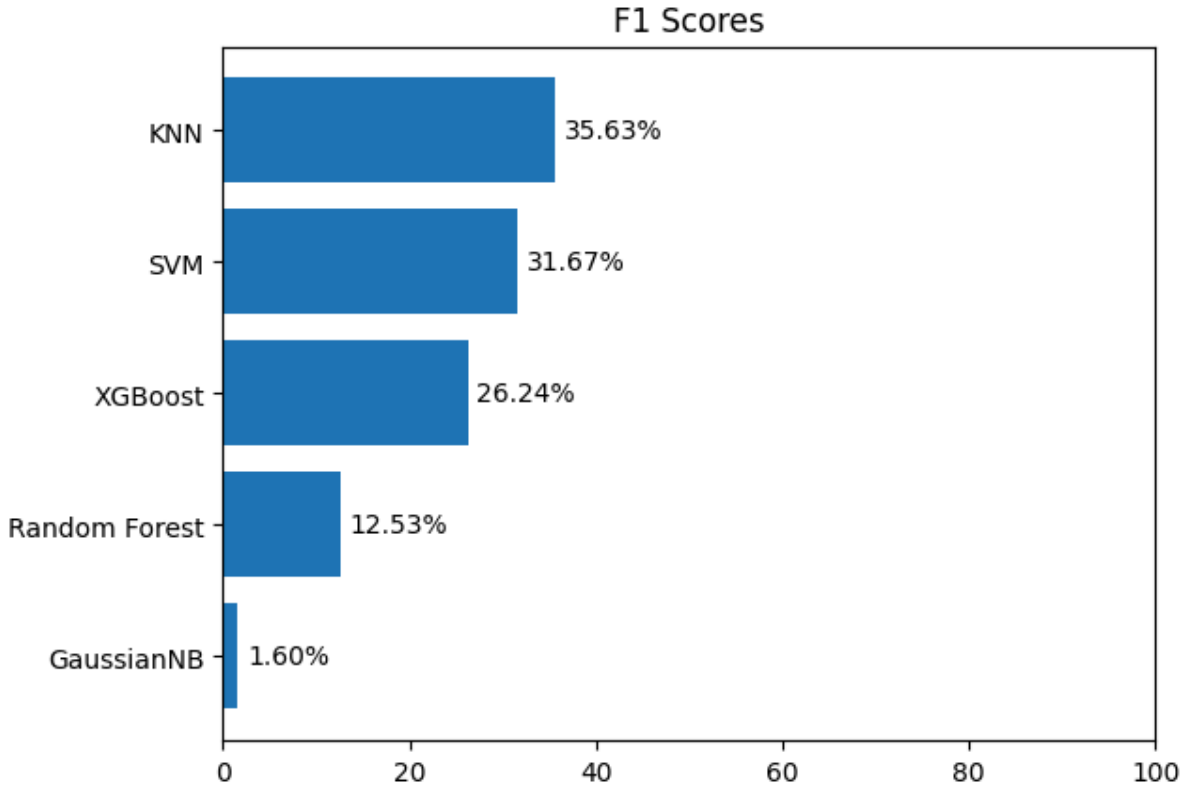
```
Confusion matrix:
[[1127   15]
 [ 606    5]]
The accuracy of K Nearest Neighbor model is :  64.5750142612664 %
The precision of K Nearest Neighbor model is :  25.0 %
The recall of K Nearest Neighbor model is :  0.8183306055646482 %
The f1 score of K Nearest Neighbor model is :  1.5847860538827259 %
```

- So sánh, đánh giá điểm số của các mô hình:



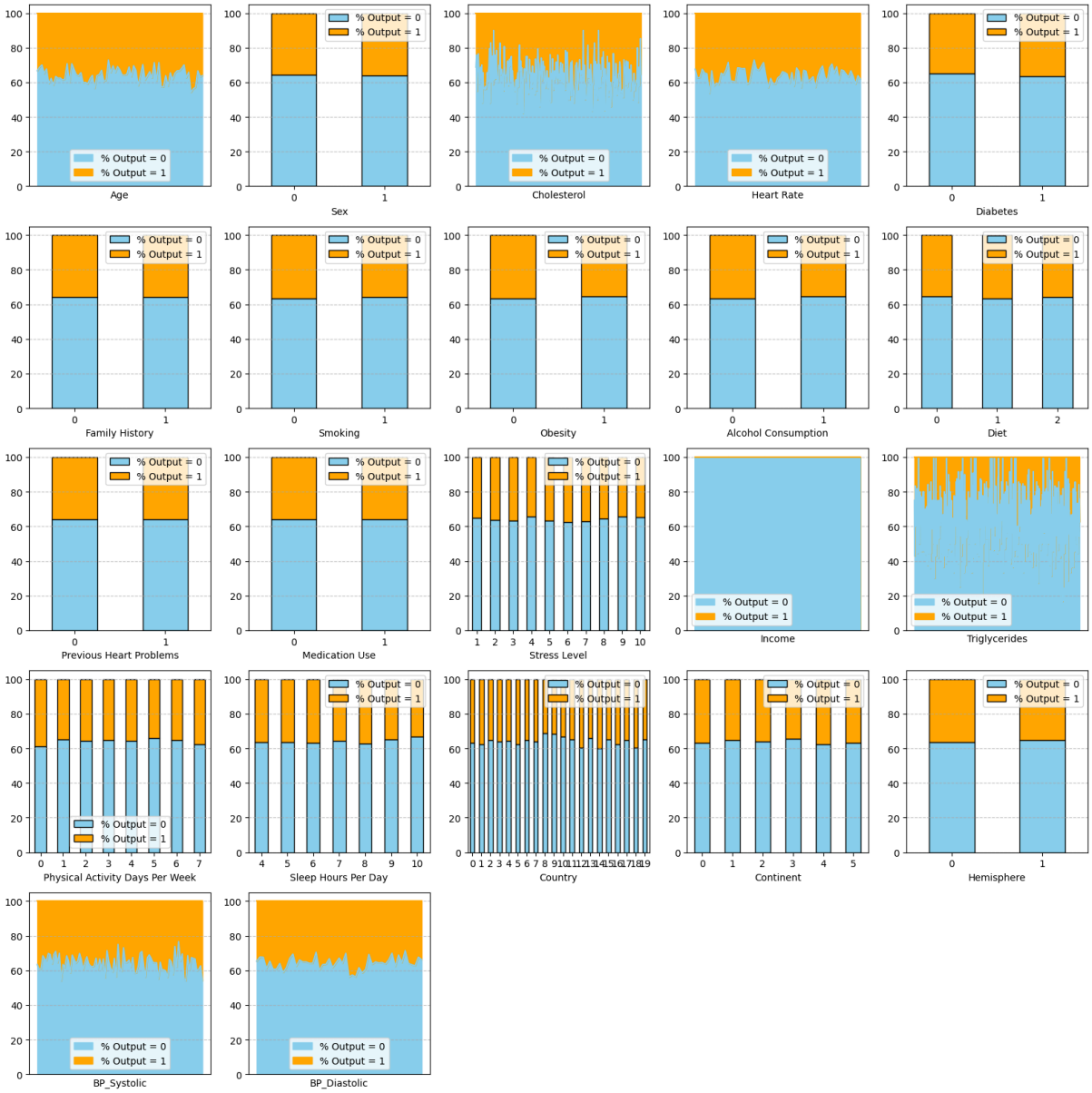
-> Các mô hình có độ chính xác khá thấp. GaussianNB là mô hình có độ chính xác cao nhất với accuracy score = 64.86%





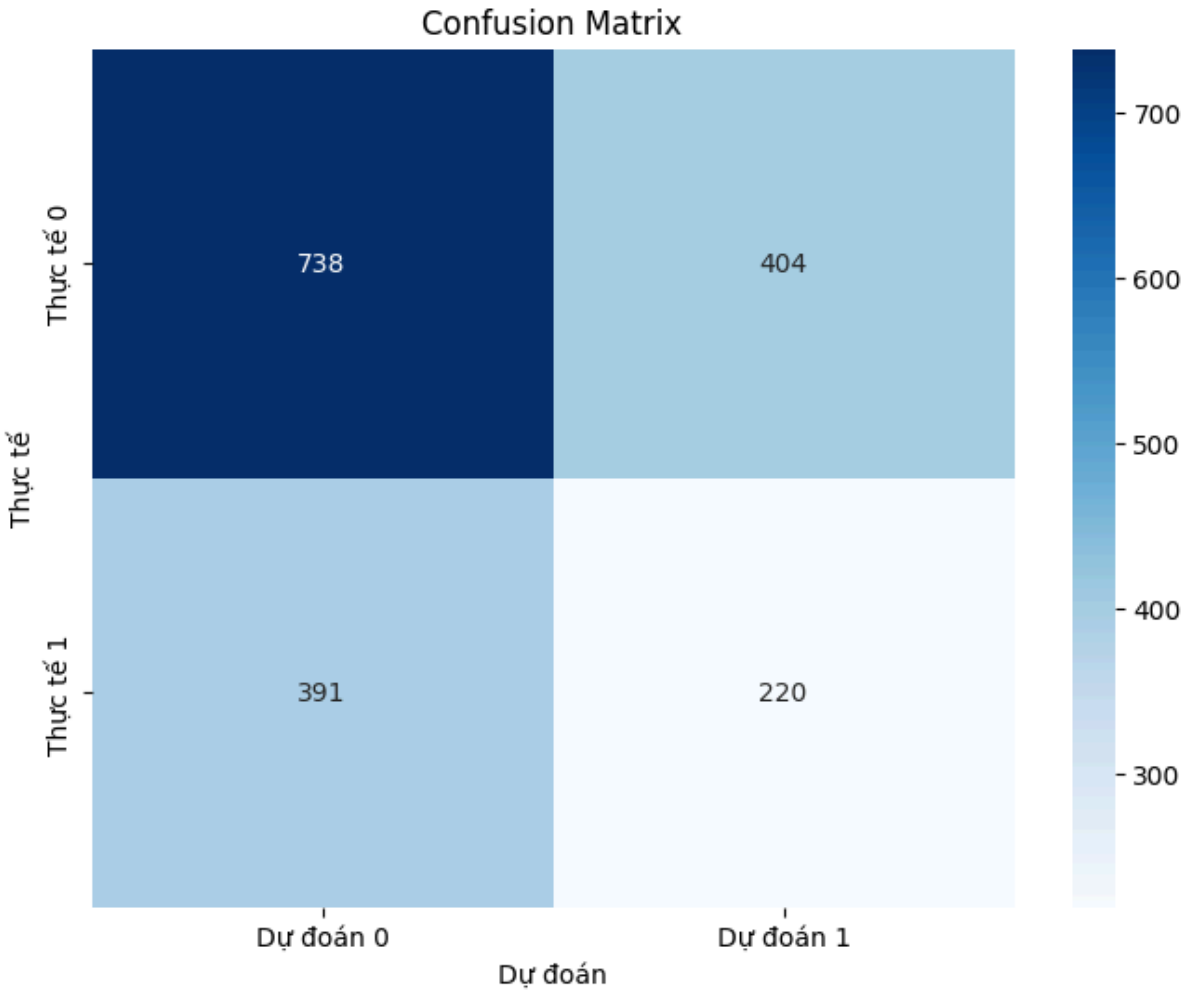
-> Điểm F1 có thứ hạng ngược lại so với độ chính xác.
-> KNN là mô hình có điểm F1 cao nhất là 35.64 điểm
-> Các mô hình có chỉ số khá thấp

- Tìm hiểu nguyên nhân:
  + Kiểm tra bộ data:

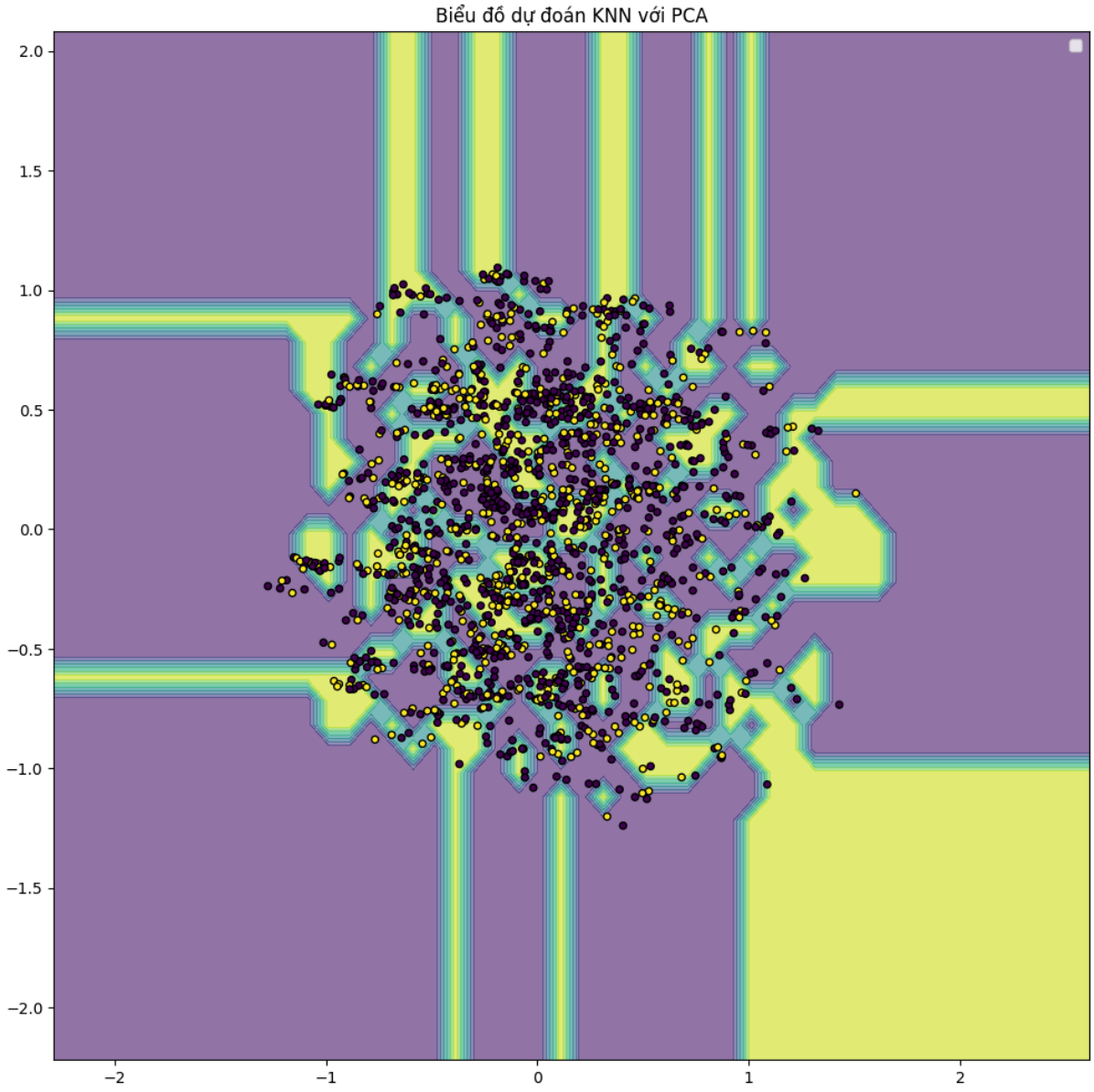


- Nhìn vào biểu đồ sự tương quan giữa từng thuộc tính với tỉ lệ output có thể thấy được nguyên nhân độ chính xác xấp xỉ 60% là do từng thuộc tính trong dataset đều có tỉ lệ output = 0 xấp xỉ 60% -> Nguyên nhân chính dẫn tới độ chính xác thấp là do bộ data

- Biểu đồ dự đoán
  + Lựa chọn mô hình: Khi dự đoán khả năng bị bệnh, kết quả False Positive (không bị bệnh nhưng kết quả dự đoán là có) dường như sẽ 'có lợi' hơn False Negative (bị bệnh nhưng dự đoán là không). Vì vậy khi lựa chọn mô hình, ta sẽ ưu tiên mô hình có FN thấp hơn, tức recall cao hơn.
    -> Dùng mô hình KNN để dự đoán
  + Biểu đồ nhiệt confusion matrix:



- Số người không bị bệnh và dự đoán đúng: 738
- Số người không bị bệnh và dự đoán sai: 404
- Số người bị bệnh và dự đoán sai: 391
- Số người bị bệnh và dự đoán đúng: 220
    + Biểu đồ dự đoán trực quan:



+ Sử dụng PCA (Principal Component Analysis) để giảm chiều tập các thuộc tính xuống 2 cột.
+ Các chấm tím thể hiện output = 0, tức là điểm dữ liệu người không bị bệnh tim; ngược lại các chấm vàng thể hiện điểm dữ liệu người bị bệnh tim.
+ Các vùng màu vàng là vùng mà mô hình dự đoán phân loại người bị bệnh tim; vùng màu tím là vùng dự đoán không bị bệnh tim.
+ Nhìn vào biểu đồ có thể thấy, bộ dữ liệu có tính phân loại khá thấp. Mô hình đã phân loại được vùng các điểm dữ liệu âm tính khá tốt, tuy nhiên các điểm dữ liệu dương tính chưa được chính xác. Nhìn chung, hiệu suất dự đoán của mô hình khá thấp.