

PHÂN TÍCH KHÁM PHÁ TẬP DỮ LIỆU CHẨN ĐOÁN BỆNH TIỂU ĐƯỜNG

Trần Hồ Minh Hải
Trương Văn Thiện
Phan Đức Nhân
Võ Gia Kiệt



VẤN ĐỀ THẢO LUẬN

- Danh mục (Agenda)
- Bản đồ nhiệt tương quan (Correlation heatmap)
- Kết luận (Conclusion)



DANH MỤC (AGENDA)

- Tóm tắt dữ liệu (Data Summary)
- Phân tích đơn biến (Univariate analysis)
- Xử lý dữ liệu bị mất (Missing values proccessing)
- Phân tích đa biến (Mulvariate analysis)
- Phân tích theo Nhóm Tuổi (Age Group wise analysis)
- Phân tích theo Mức độ Béo phì (BMI Category analysis)
- Phân tích theo Ngưỡng Đường huyết (Glucose Threshold analysis)
- Phân tích theo Tiền sử Gia đình (Pedigree Function analysis)



PHÂN CÔNG NHIỆM VỤ

- Người 1: Trần Hồ Minh Hải

Tóm tắt dữ liệu (Data Summary)

Phân tích đơn biến (Univariate analysis)

Bản đồ nhiệt tương quan (Correlation heatmap)

Làm powerpoint

- Người 2: Võ Gia Kiệt

Xử lý dữ liệu bị mất (Missing values processing)

Phân tích đa biến (Multivariate analysis)

Kết luận (Conclusion)

- Người 3: Phan Đức Nhân

Phân tích theo Nhóm Tuổi (Age Group wise analysis)

Phân tích theo Mức độ Béo phì (BMI Category analysis)

- Người 4: Trương Văn Thiện

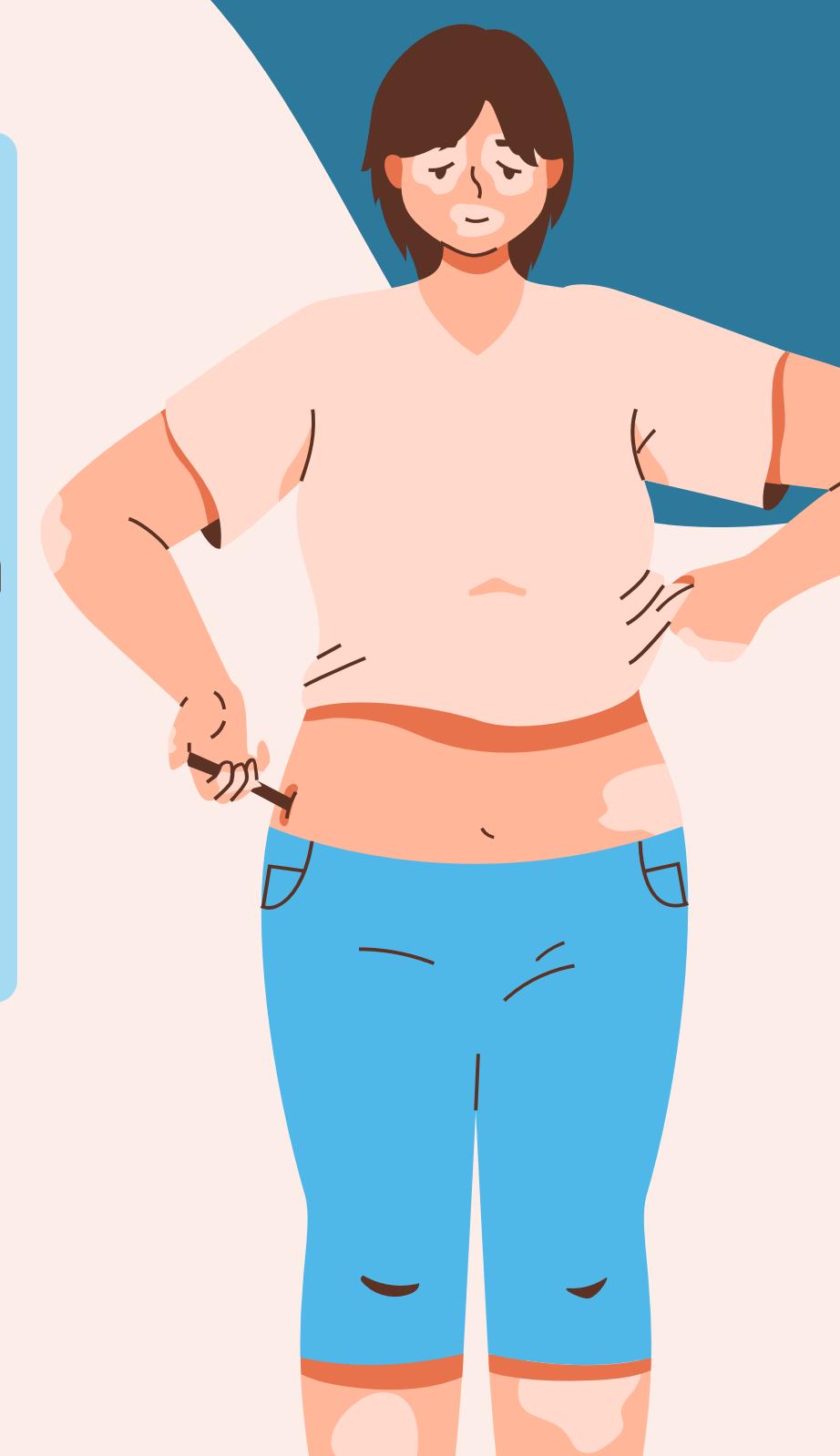
Phân tích theo Ngưỡng Đường huyết (Glucose Threshold analysis)

Phân tích theo Tiền sử Gia đình (Pedigree Function analysis)



TÓM TẮT DỮ LIỆU (DATA SUMMARY)

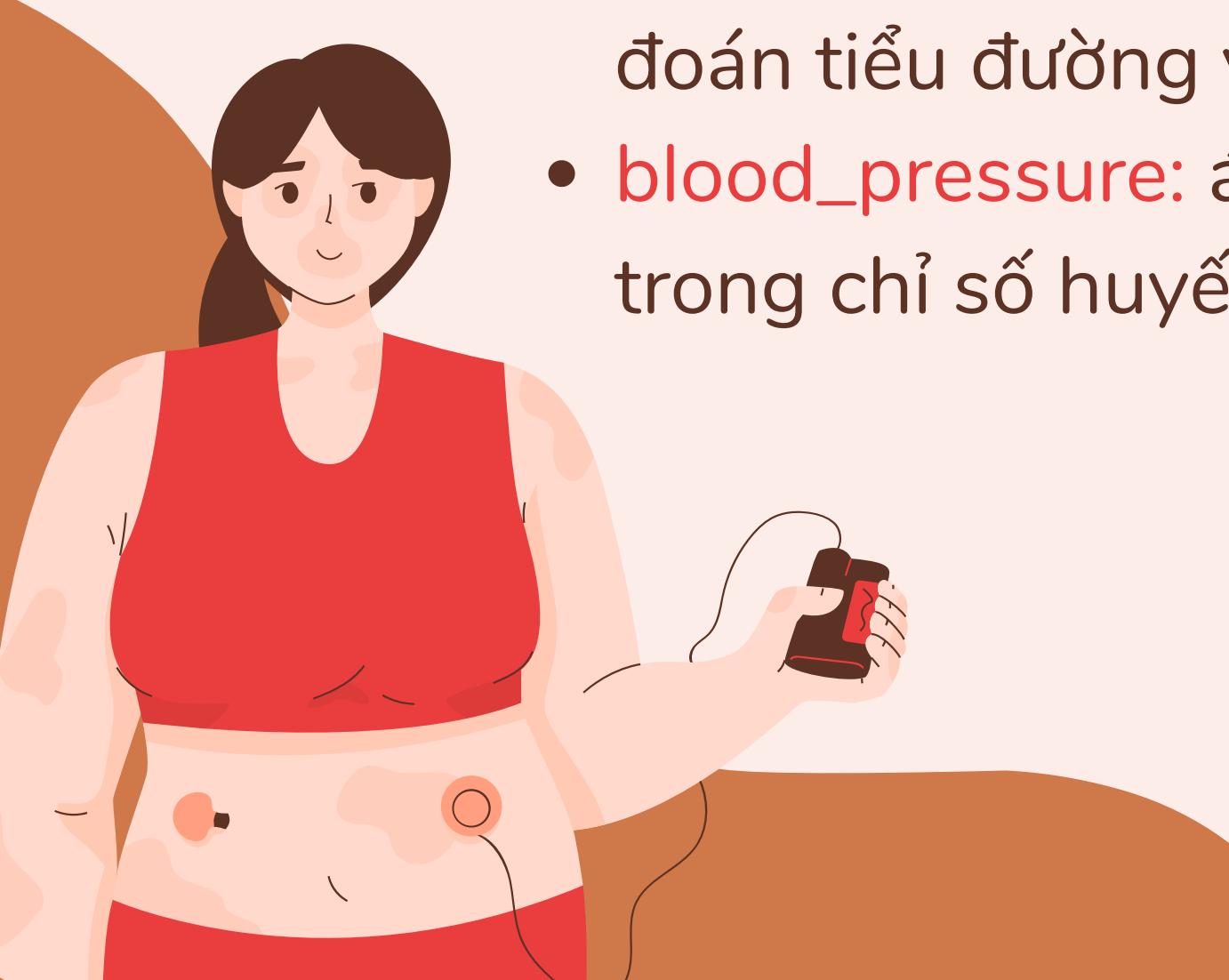
TÓM TẮT DỮ LIỆU (DATA SUMMARY)



Tập dữ liệu cho trước chứa các cột biến khác nhau đóng vai trò quan trọng trong việc dự đoán bệnh tiểu đường. Các biến đó bao gồm:

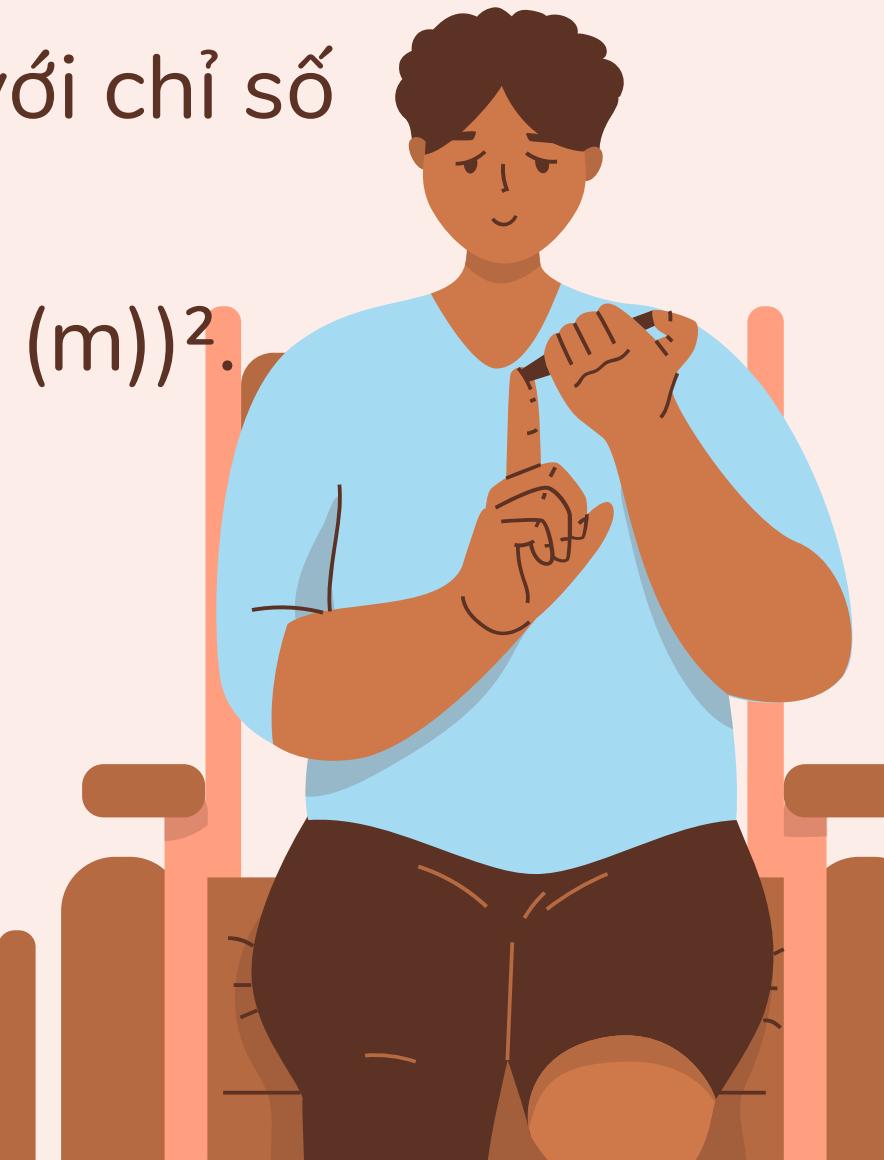
TÓM TẮT DỮ LIỆU (DATA SUMMARY)

- **pregnancies**: số lần một người phụ nữ từng mang thai.
- **glucose**: chỉ số đường huyết được đo 2 giờ sau khi bệnh nhân uống 75g dung dịch glucose. Đây là tiêu chuẩn vàng để chẩn đoán tiểu đường và tiền tiểu đường.
- **blood_pressure**: áp lực trong động mạch khi tim giãn ra (là số dưới trong chỉ số huyết áp, ví dụ 120/80).



TÓM TẮT DỮ LIỆU (DATA SUMMARY)

- **skin_thickness**: một phương pháp đo lượng mỡ dự trữ trong cơ thể. Người ta dùng một thước kẹp đặc biệt để kẹp và đo độ dày của một nếp da ở mặt sau cánh tay.
- **insulin**: lượng insulin trong máu được đo cùng thời điểm với chỉ số đường huyết sau 2 giờ.
- **bmi (body mass index)**: $BMI = \frac{\text{Cân nặng (kg)}}{(\text{Chiều cao (m)})^2}$. Đây là chỉ số đánh giá thể trạng gầy/béo phì biến.

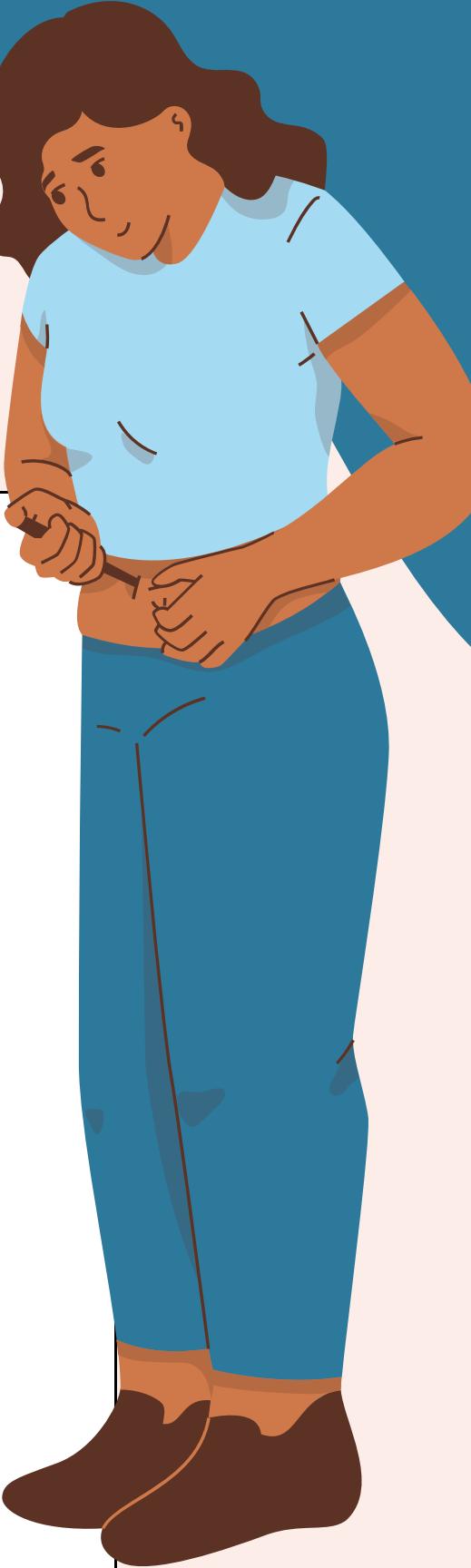
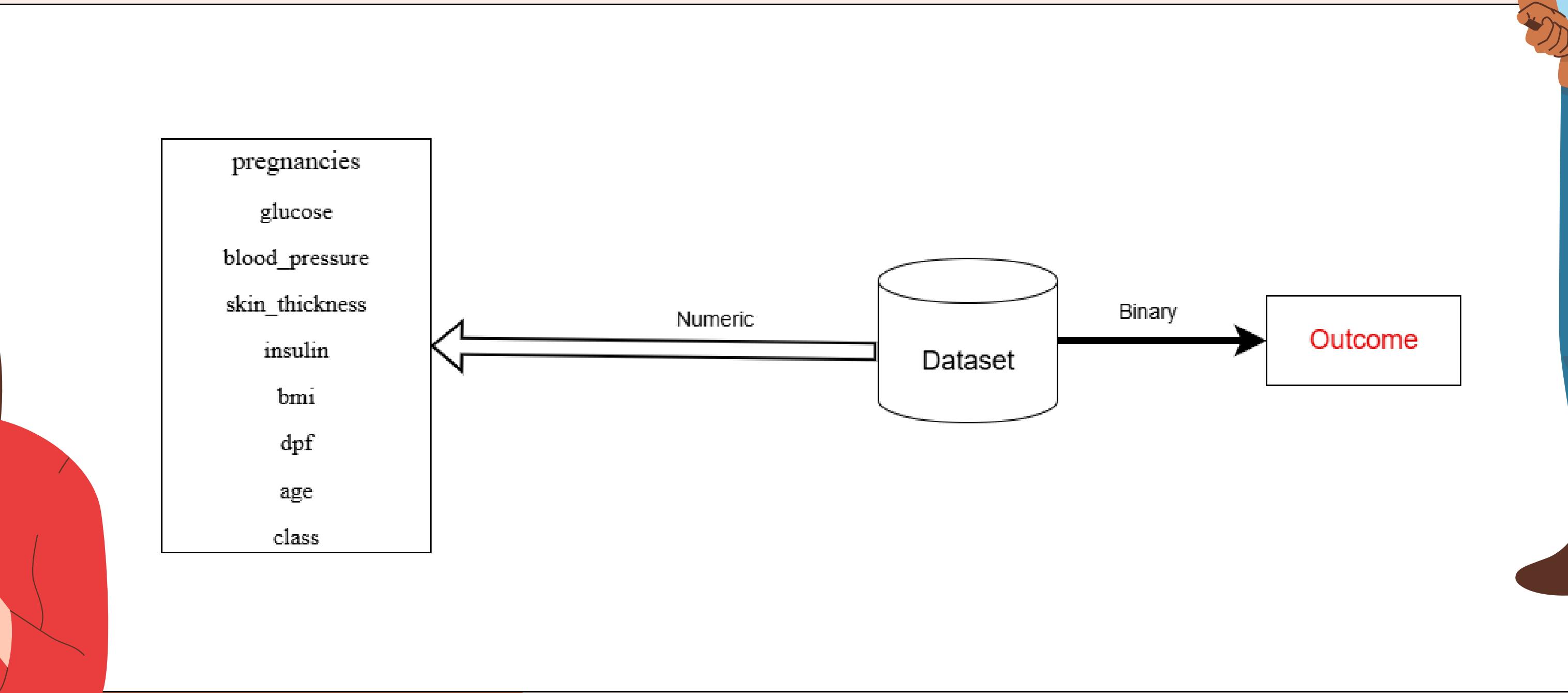


TÓM TẮT DỮ LIỆU (DATA SUMMARY)

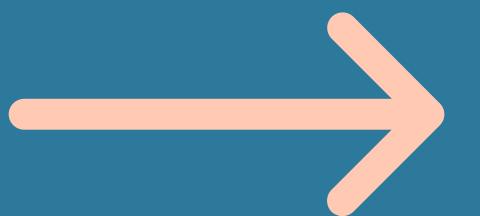
- **dpf** (diabetes pedigree function): chỉ số lượng hóa tiền sử gia đình mắc bệnh tiểu đường. Nó được tính toán dựa trên mối quan hệ huyết thống và số lượng thành viên trong gia đình mắc bệnh.
- **age**: Tuổi của một người tại thời điểm khảo sát.
- **outcome**: kết quả chẩn đoán mà mô hình máy học cần dự đoán (0 – Âm tính với bệnh tiểu đường; 1 – Dương tính với bệnh tiểu đường).



TÓM TẮT DỮ LIỆU (DATA SUMMARY)



PHÂN TÍCH ĐƠN BIẾN (UNIVARIATE ANALYSIS)



PHÂN TÍCH ĐƠN BIẾN (UNIVARIATE ANALYSIS)

1. HISTOGRAM HOẶC KDE PLOT CHO CÁC BIẾN SỐ (GLUCOSE, BMI, AGE, INSULIN...).
2. BOX PLOT CHO TẤT CẢ CÁC BIẾN SỐ ĐỂ PHÁT HIỆN OUTLIER.
3. COUNT PLOT (BAR CHART) CHO BIẾN MỤC TIÊU OUTCOME.

PHÂN TÍCH ĐƠN BIẾN (UNIVARIATE ANALYSIS)

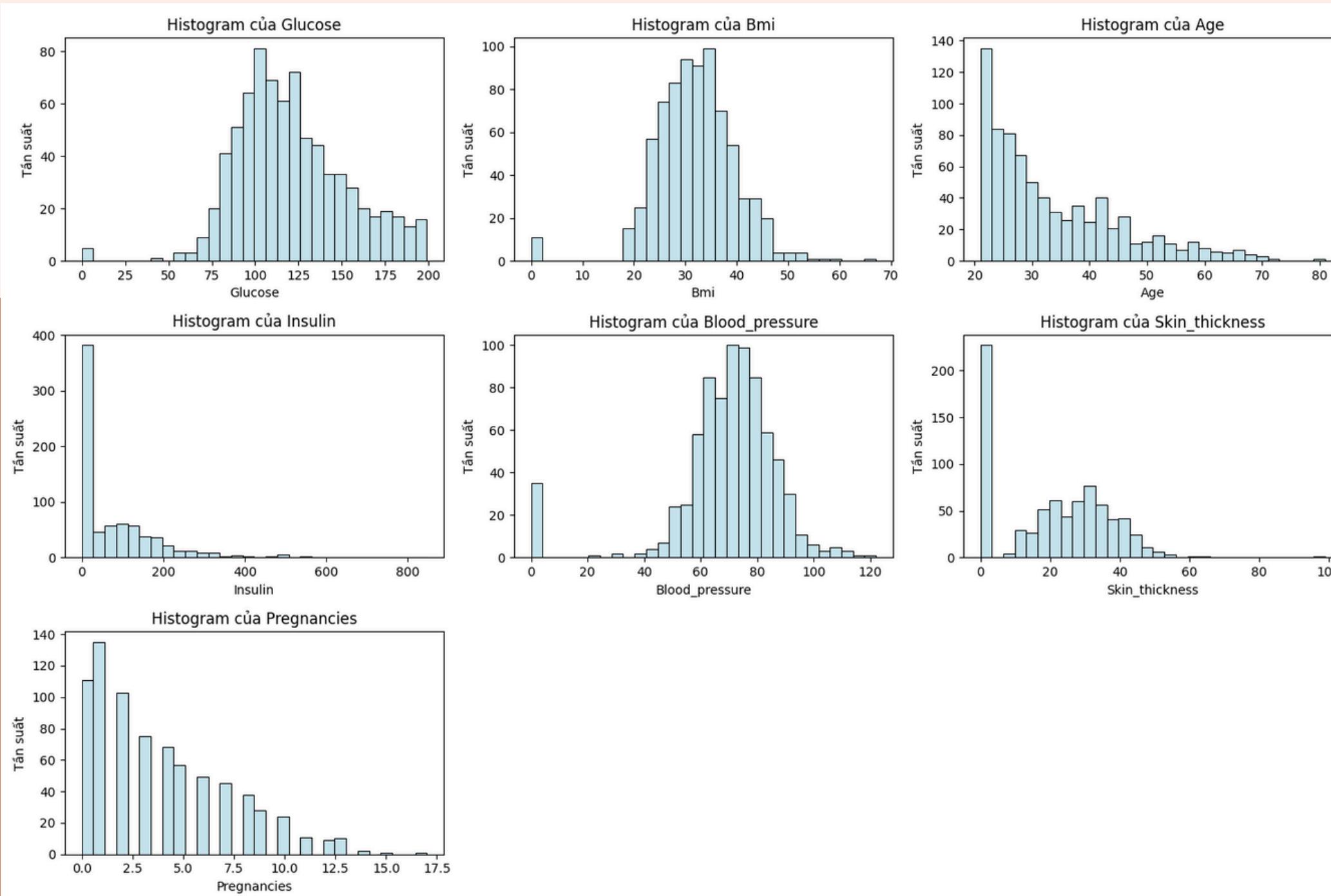


Chart 1. Histogram của các thuộc tính

Từ histogram ta có thể nhận xét được:

- Các thuộc tính có phân phối gần chuẩn với tần suất cao ở vùng trung tâm: **glucose, blood_pressure, BMI** (peak ở trung tâm).
- Các thuộc tính có phân phối lệch phải với tần suất cao ở giá trị thấp: **age, insulin, skin_thickness, pregnancies** (peak ở giá trị thấp).
- Đặc biệt: **glucose, bmi, blood_pressure, insulin và skin_thickness** có nhiều giá trị bằng 0 (có thể là dữ liệu missing).

PHÂN TÍCH ĐƠN BIẾN (UNIVARIATE ANALYSIS)

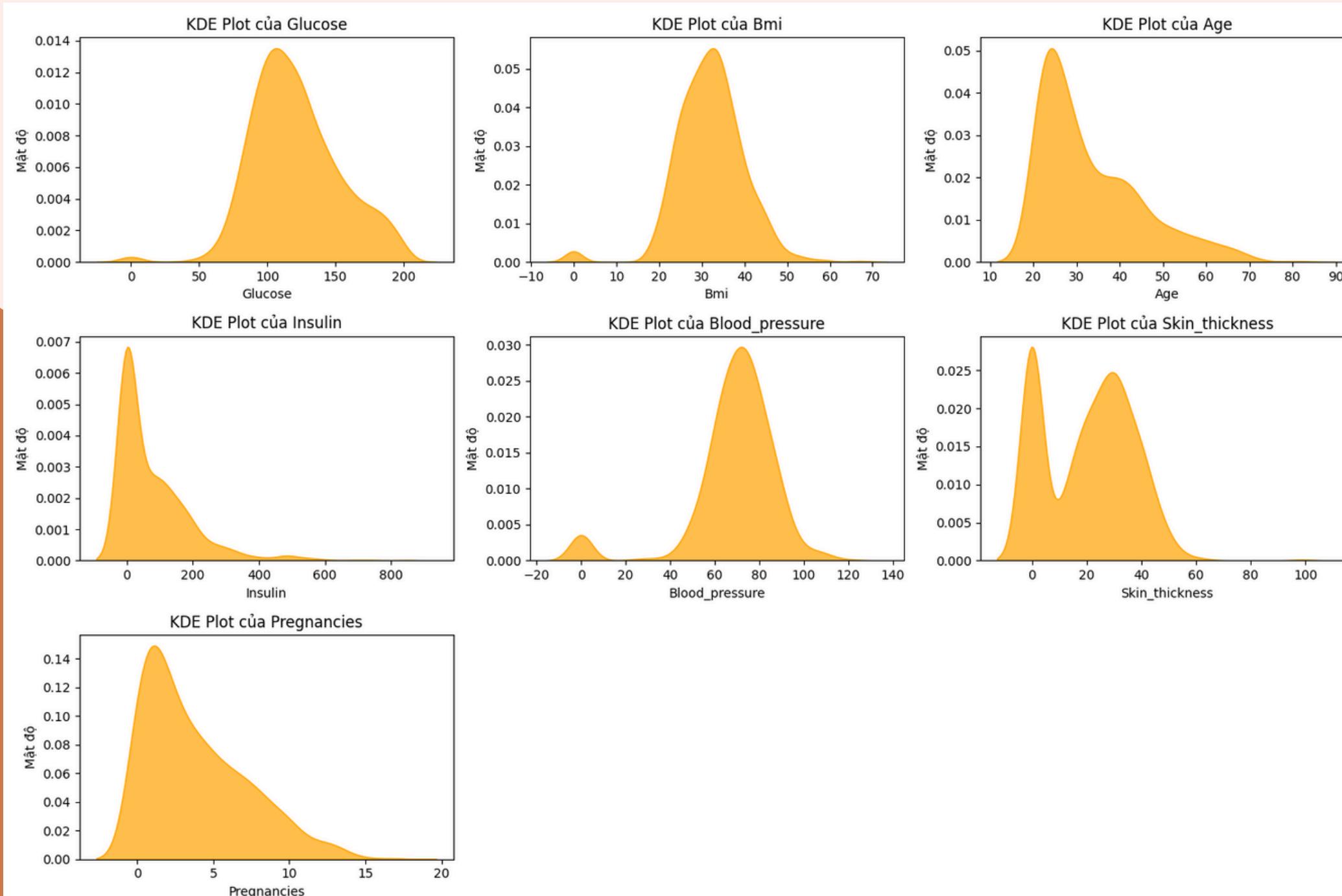


Chart 2. KDE Plot của các thuộc tính

Từ KDE Plot ta có nhận xét:

1. Glucose:

Phân phối gần chuẩn, đỉnh quanh 100-120 mg/dL

Ít outliers, phân phối tập trung → chất lượng data tốt

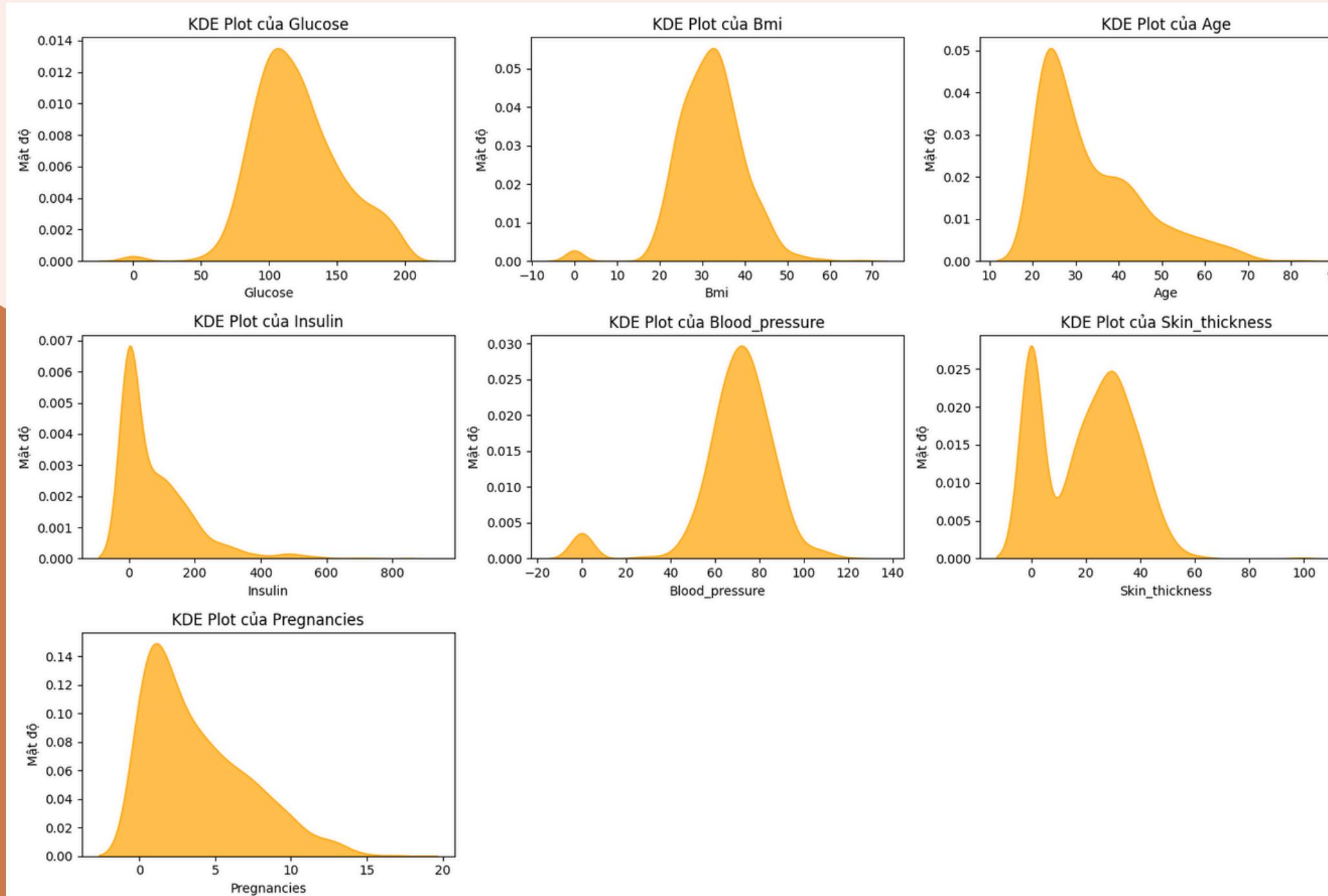
2. BMI:

Phân phối chuẩn lệch phải nhẹ, đỉnh quanh 30-35

Không có giá trị âm (KDE mở rộng do bandwidth)

BMI trung bình khá cao (hơn 30) → nhiều đối tượng thừa cân

PHÂN TÍCH ĐƠN BIẾN (UNIVARIATE ANALYSIS)



3. Age:

Phân phối lệch phải rõ rệt
Đa số bệnh nhân trẻ tuổi (20-35 tuổi)
Phù hợp với đối tượng nghiên cứu (phụ nữ trong độ tuổi sinh sản)

4. Insulin:

Phân phối lệch phải mạnh, peak ở giá trị thấp
Có một số outliers ở giá trị cao (>400)
Phản ánh đúng đặc điểm insulin: nhiều người có insulin thấp, ít người có insulin rất cao

Chart 2. KDE Plot của các thuộc tính

PHÂN TÍCH ĐƠN BIẾN (UNIVARIATE ANALYSIS)

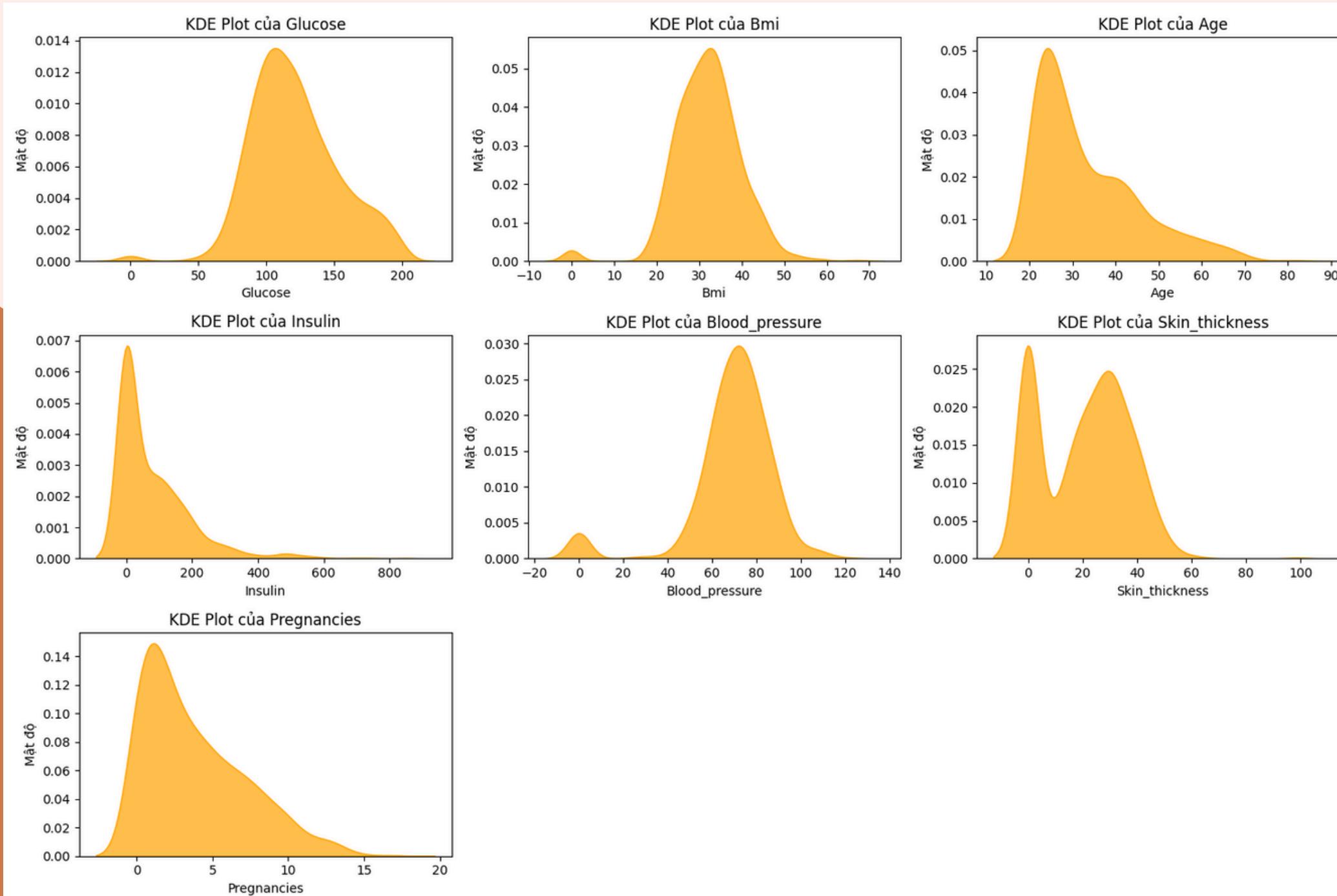


Chart 2. KDE Plot của các thuộc tính

5. Blood Pressure:

Phân phối gần chuẩn, đỉnh quanh 70-80 mmHg

Huyết áp trong ngưỡng bình thường

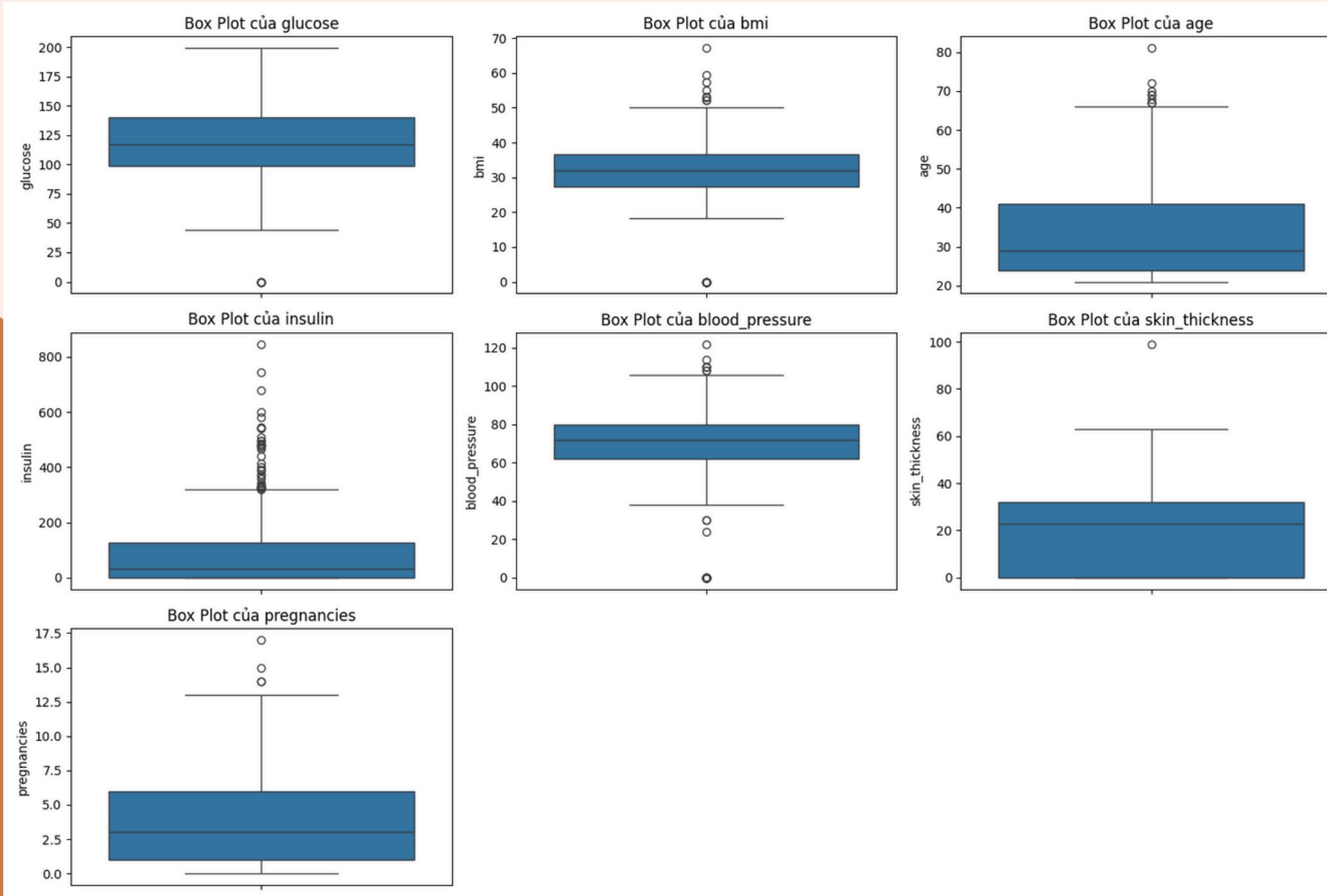
6. Skin Thickness:

Phân phối lệch phải, tập trung ở giá trị thấp

7. Pregnancies:

Phân phối lệch phải rõ rệt, peak ở giá trị thấp
Đa số có 0-2 lần mang thai

PHÂN TÍCH ĐƠN BIẾN (UNIVARIATE ANALYSIS)



1. Gulucose

Biến glucose có phân phối tương đối cân đối
Xuất hiện ít ngoại lai ở phía dưới
Hầu hết giá trị glucose tập trung trong khoảng 80-140 mg/dL

2. Insulin

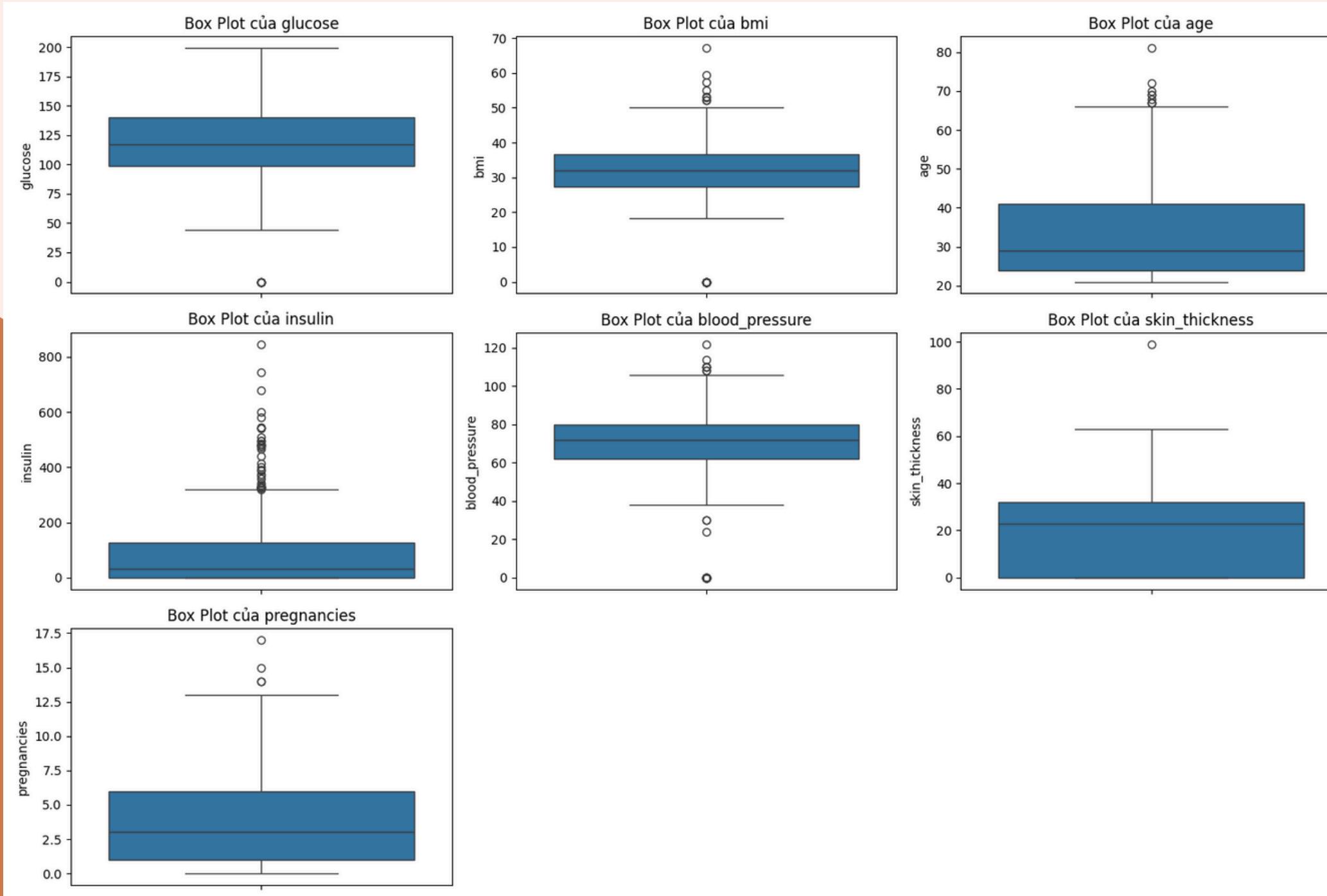
Phân phối lệch phải rõ rệt
Có nhiều điểm ngoại lai ở phía trên và một ít ở phía dưới
Đa số giá trị insulin dưới 200 μ U/ml

3. Blood pressure

Phân phối gần như đối xứng
Có nhiều điểm ngoại lai xuất hiện ở cả hai phía
Huyết áp trung bình khoảng 70-90 mmHg

Chart 3. Box Plot của các thuộc tính

PHÂN TÍCH ĐƠN BIẾN (UNIVARIATE ANALYSIS)



4. BMI

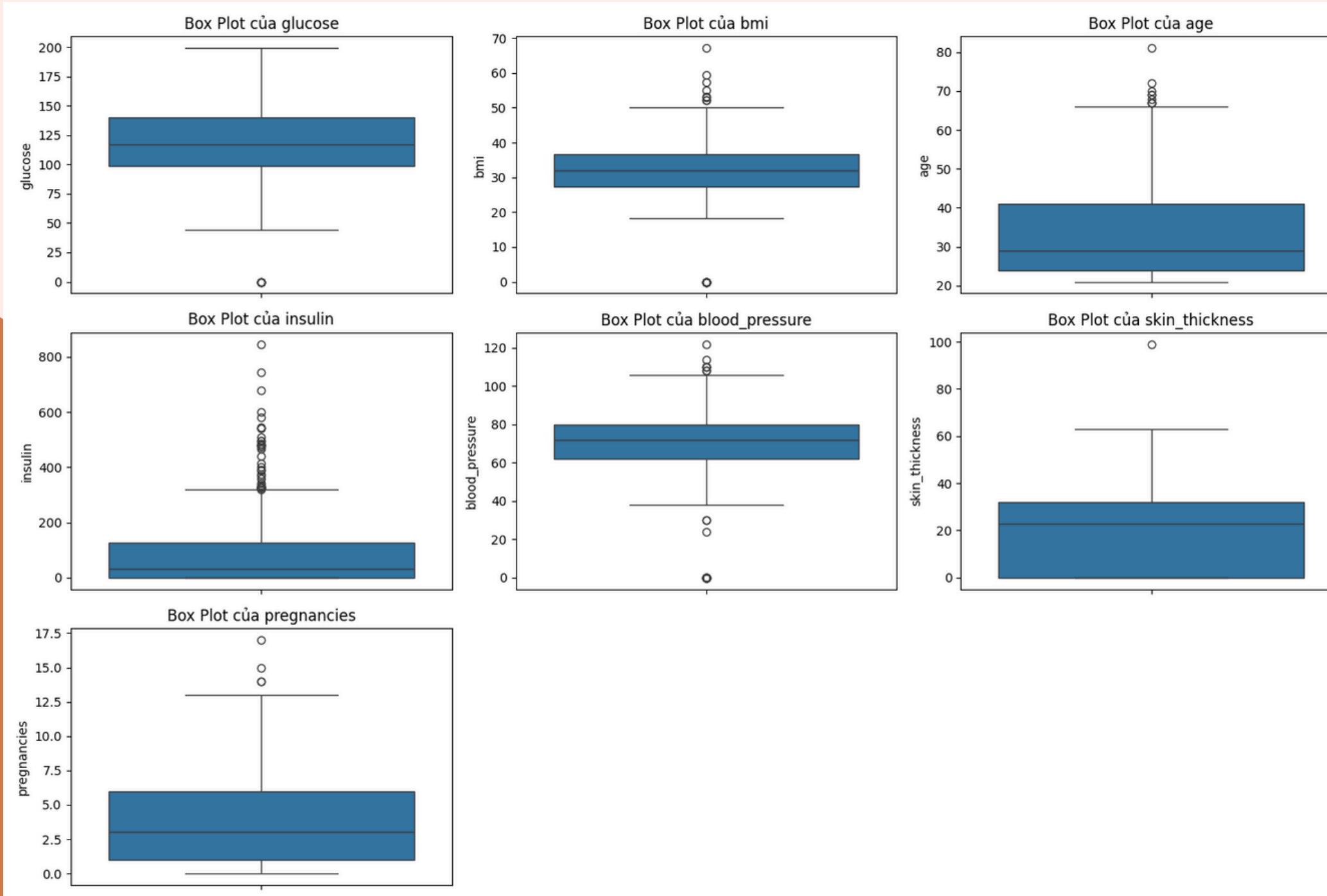
Phân phối lệch nhẹ về bên phải
Xuất hiện một số điểm ngoại lai ở giá trị cao
Chỉ số BMI chủ yếu trong khoảng $25-35 \text{ kg/m}^2$

5. Age

Phân phối lệch phải
Có điểm ngoại lai ở độ tuổi cao
Độ tuổi phổ biến từ 20-40 tuổi

Chart 3. Box Plot của các thuộc tính

PHÂN TÍCH ĐƠN BIẾN (UNIVARIATE ANALYSIS)



6. Skin thickness

Phân phối lệch phải

Ít điểm ngoại lai

Giá trị tập trung chủ yếu dưới 40 mm

7. Pregnancies

Phân phối lệch phải mạnh

Có ít điểm ngoại lai xuất hiện ở phía trên

Số lần mang thai chủ yếu từ 0-5

Chart 3. Box Plot của các thuộc tính

PHÂN TÍCH ĐƠN BIẾN (UNIVARIATE ANALYSIS)

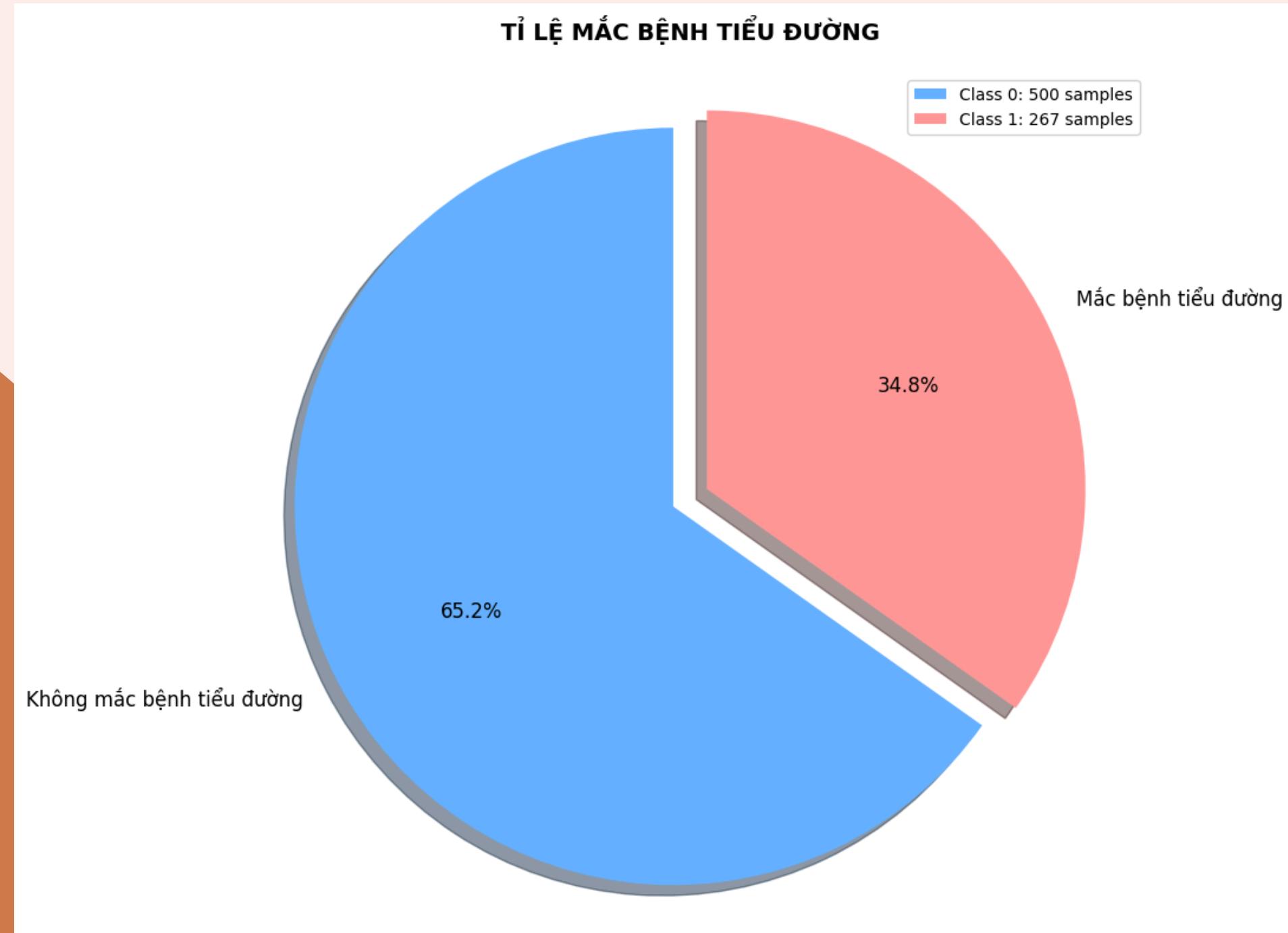
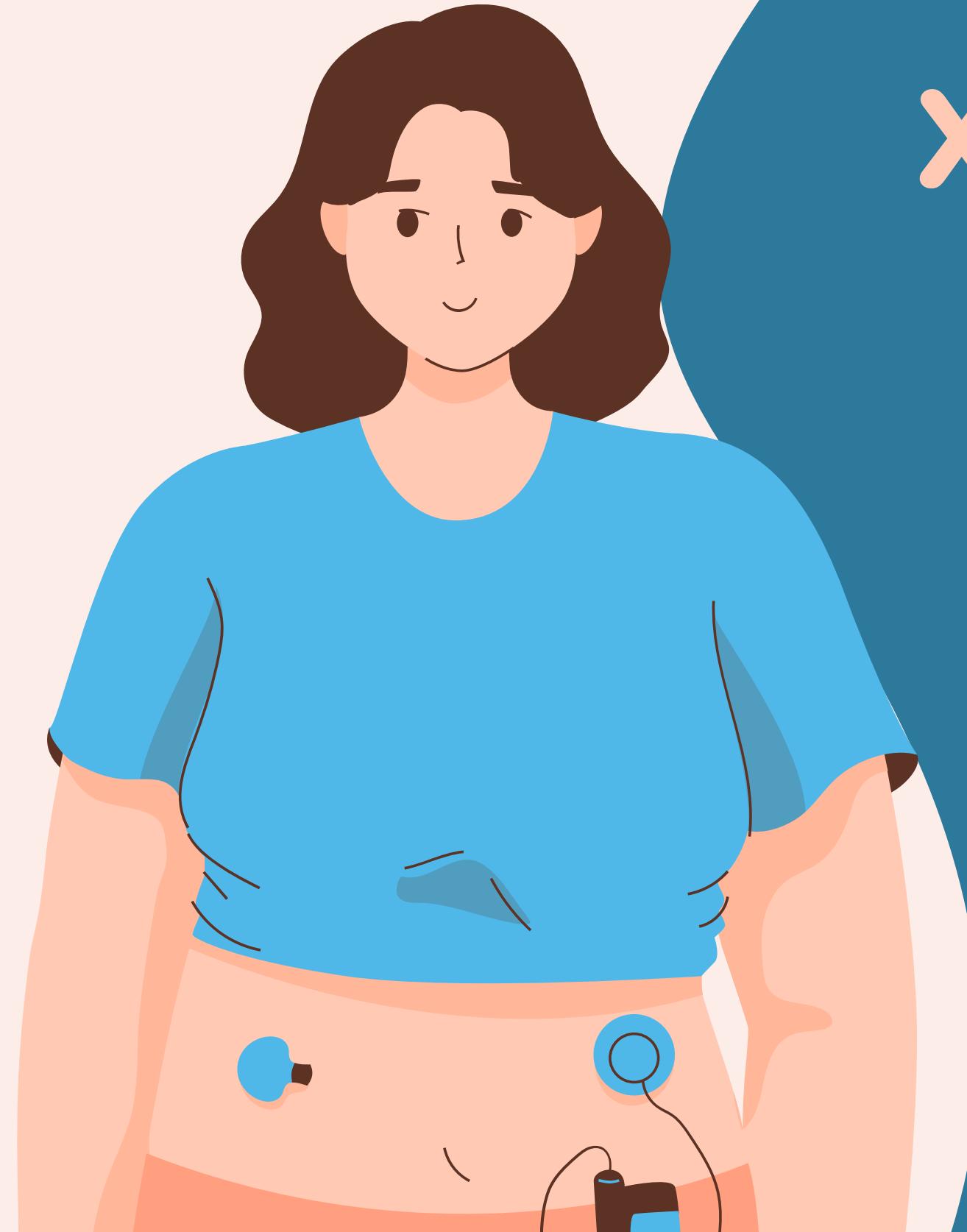


Chart 4. Count Plot của biến mục tiêu class

Từ biểu đồ tròn thể hiện tỉ lệ mắc bệnh tiểu đường trong tập dataset, ta có nhận xét sau: Biểu đồ cho thấy sự phân bố của biến mục tiêu với 65.2% mẫu không mắc bệnh tiểu đường và 34.8% mắc bệnh. Dataset có sự chênh lệch đáng kể giữa hai lớp.



XỬ LÝ DỮ LIỆU BỊ MẤT (MISSING VALUES PROCESSING)



XỬ LÝ DỮ LIỆU BỊ MẤT (MISSING VALUES PROCESSING)

Thông tin dữ liệu:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
count	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000
mean	3.845052	120.894531	69.105469	20.536458	79.799479	31.992578	0.471876	33.240885	0.348958
std	3.369578	31.972618	19.355807	15.952218	115.244002	7.884160	0.331329	11.760232	0.476951
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.078000	21.000000	0.000000
25%	1.000000	99.000000	62.000000	0.000000	0.000000	27.300000	0.243750	24.000000	0.000000
50%	3.000000	117.000000	72.000000	23.000000	30.500000	32.000000	0.372500	29.000000	0.000000
75%	6.000000	140.250000	80.000000	32.000000	127.250000	36.600000	0.626250	41.000000	1.000000
max	17.000000	199.000000	122.000000	99.000000	846.000000	67.100000	2.420000	81.000000	1.000000

Từ đó, ta có thể phát hiện ra những giá trị bất thường: Các giá trị min là 0 ở các cột Glucose, BloodPresure, SkinThickness, Insulin, BMI => Không hợp lý

XỬ LÝ DỮ LIỆU BỊ MẤT (MISSING VALUES PROCESSING)

Xử lí các giá trị bị thiếu thành NaN: các giá trị thiếu bằng cách điền trung bình (mean) cho các biến Glucose, BloodPressure và trung vị (median) cho các biến SkinThickness, Insulin, BMI.

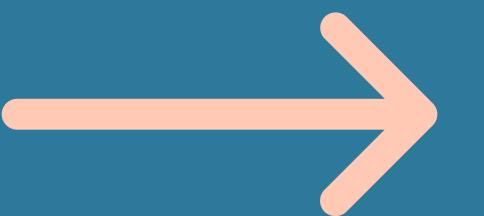


	0
Pregnancies	0
Glucose	0
BloodPressure	0
SkinThickness	0
Insulin	0
BMI	0
DiabetesPedigreeFunction	0
Age	0
Outcome	0

	0
Pregnancies	0
Glucose	5
BloodPressure	35
SkinThickness	227
Insulin	374
BMI	11
DiabetesPedigreeFunction	0
Age	0
Outcome	0

Ta có thể thấy được các giá trị bị thiếu sau khi xử lí thay thế các giá trị 0 thành NaN

PHÂN TÍCH ĐA BIẾN (MULTIVARIATE ANALYSIS)



PHÂN TÍCH ĐA BIẾN (MULTIVARIATE ANALYSIS)



Với sơ đồ trên ta có thể thấy các dữ liệu tương quan:

1. Biến quan trọng nhất (Phân biệt rõ Outcome)

Glucose: Outcome = 1 tập trung ở giá trị cao (>120)

BMI: Outcome = 1 tập trung nhiều ở BMI cao (>30)

Age: Outcome = 1 xuất hiện nhiều ở tuổi >40 .

2. Các cặp biến giúp tăng sức phân biệt

Glucose + BMI: Outcome = 1 rõ rệt ở vùng Glucose cao + BMI cao.

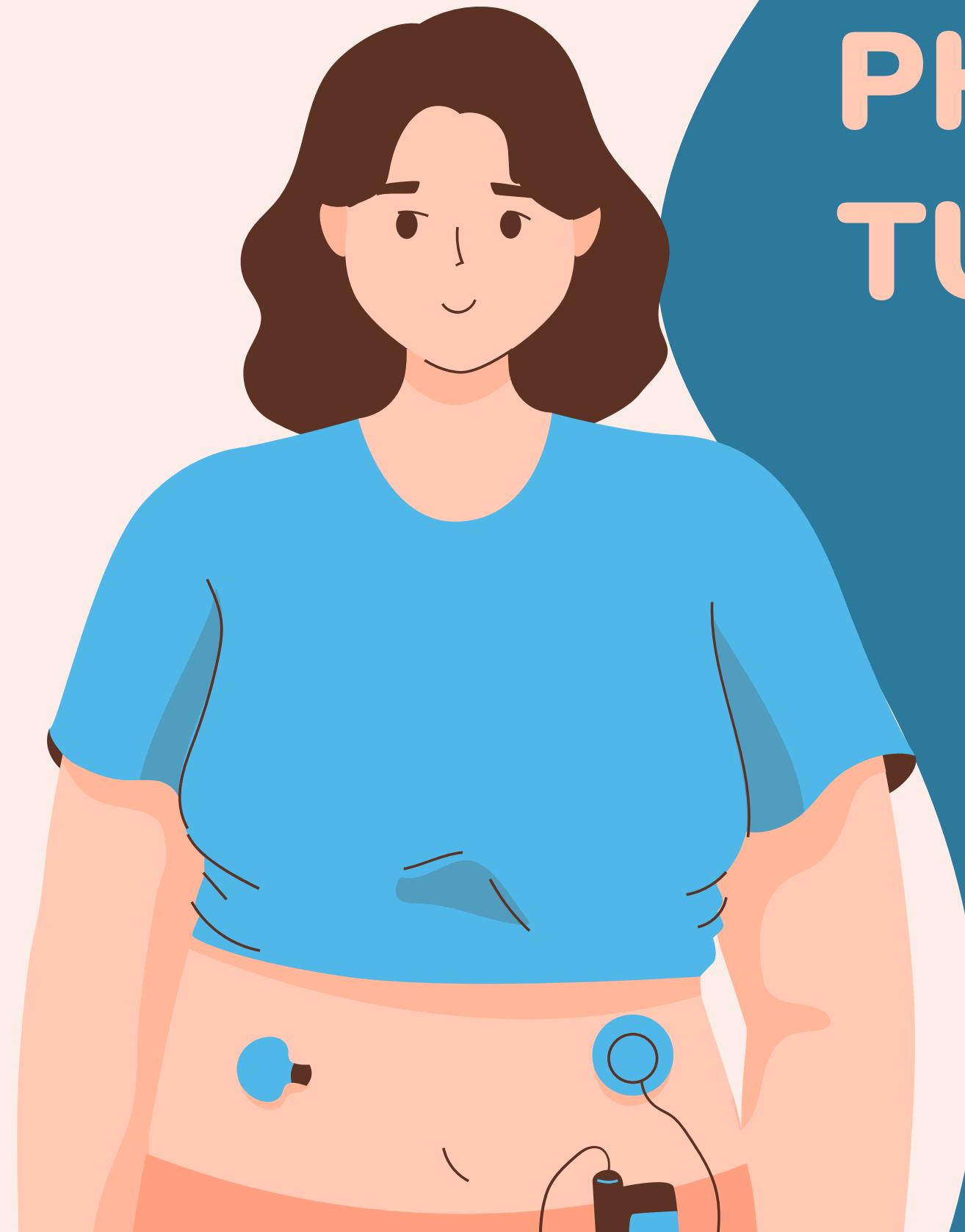
Glucose + Age: Outcome = 1 nhiều hơn ở Glucose cao và tuổi trung niên trở lên.

BMI + Age: Outcome = 1 thường ở tuổi cao và BMI cao.

PHÂN TÍCH ĐA BIẾN (MULTIVARIATE ANALYSIS)



Các biến có sức mạnh phân loại tốt nhất trong dataset này là Glucose, BMI, Age và Pregnancies, đặc biệt khi kết hợp theo cặp. Các biến như BloodPressure, SkinThickness, Insulin, DPF ít có khả năng phân biệt Outcome rõ rệt.



PHÂN TÍCH THEO NHÓM TUỔI (AGE GROUP WISE ANALYSIS)

PHÂN TÍCH THEO NHÓM TUỔI (AGE GROUP WISE ANALYSIS)

CÂU HỎI PHÂN TÍCH:

- TỶ LỆ MẮC BỆNH (OUTCOME=1) Ở NHÓM TUỔI NÀO LÀ CAO NHẤT? (KỲ VỌNG: TỶ LỆ TĂNG DẦN THEO TUỔI).
- Ở NHÓM TUỔI TRẺ (<30), NHỮNG NGƯỜI VẪN MẮC BỆNH CÓ ĐẶC ĐIỂM CHUNG GÌ? (GLUCOSE RẤT CAO? BMI RẤT CAO? TIỀN SỬ GIA ĐÌNH NĂNG?).

PHÂN TÍCH THEO NHÓM TUỔI (AGE GROUP WISE ANALYSIS)

Xem unique của thuộc tính age:

```
diabetes_df["age"].unique()
```

```
array([50, 31, 32, 21, 33, 30, 26, 29, 53, 54, 34, 57, 59, 51, 27, 41, 43,  
22, 38, 60, 28, 45, 35, 46, 56, 37, 48, 40, 25, 24, 58, 42, 44, 39,  
36, 23, 61, 69, 62, 55, 65, 47, 52, 66, 49, 63, 67, 72, 81, 64, 70,  
68])
```

PHÂN TÍCH THEO NHÓM TUỔI (AGE GROUP WISE ANALYSIS)

Chia biến Age thành các nhóm (vd: <30, 30-45, 45-60, >60)

```
bins=[0,29,45,60,100]
labels=["<30","30-45","45-60",">60"]
diabetes_df["age_group"]=pd.cut(diabetes_df["age"],bins=bins,labels=labels,
right=True)
diabetes_df.head()
```

	pregnancies	glucose	blood_pressure	skin_thickness	insulin	bmi	dpf	age	class	age_group
0	6	148.0	72.0	35.0	NaN	33.6	0.627	50	1	45-60
1	1	85.0	66.0	29.0	NaN	26.6	0.351	31	0	30-45
2	8	183.0	64.0	NaN	NaN	23.3	0.672	32	1	30-45
3	1	89.0	66.0	23.0	94.0	28.1	0.167	21	0	<30
4	0	137.0	40.0	35.0	168.0	43.1	2.288	33	1	30-45

PHÂN TÍCH THEO NHÓM TUỔI (AGE GROUP WISE ANALYSIS)

Tính số lượng các nhóm tuổi tương ứng và bao nhiêu người trong số đó bị tiểu đường (positive)

```
age_group_stats =  
diabetes_df.groupby("age_group", observed=True)[ "outcome" ].agg(  
    total="count",  
    positive="sum"  
)  
age_group_stats
```

age_group	total	positive
<30	396	84
30-45	254	126
45-60	91	51
>60	27	7

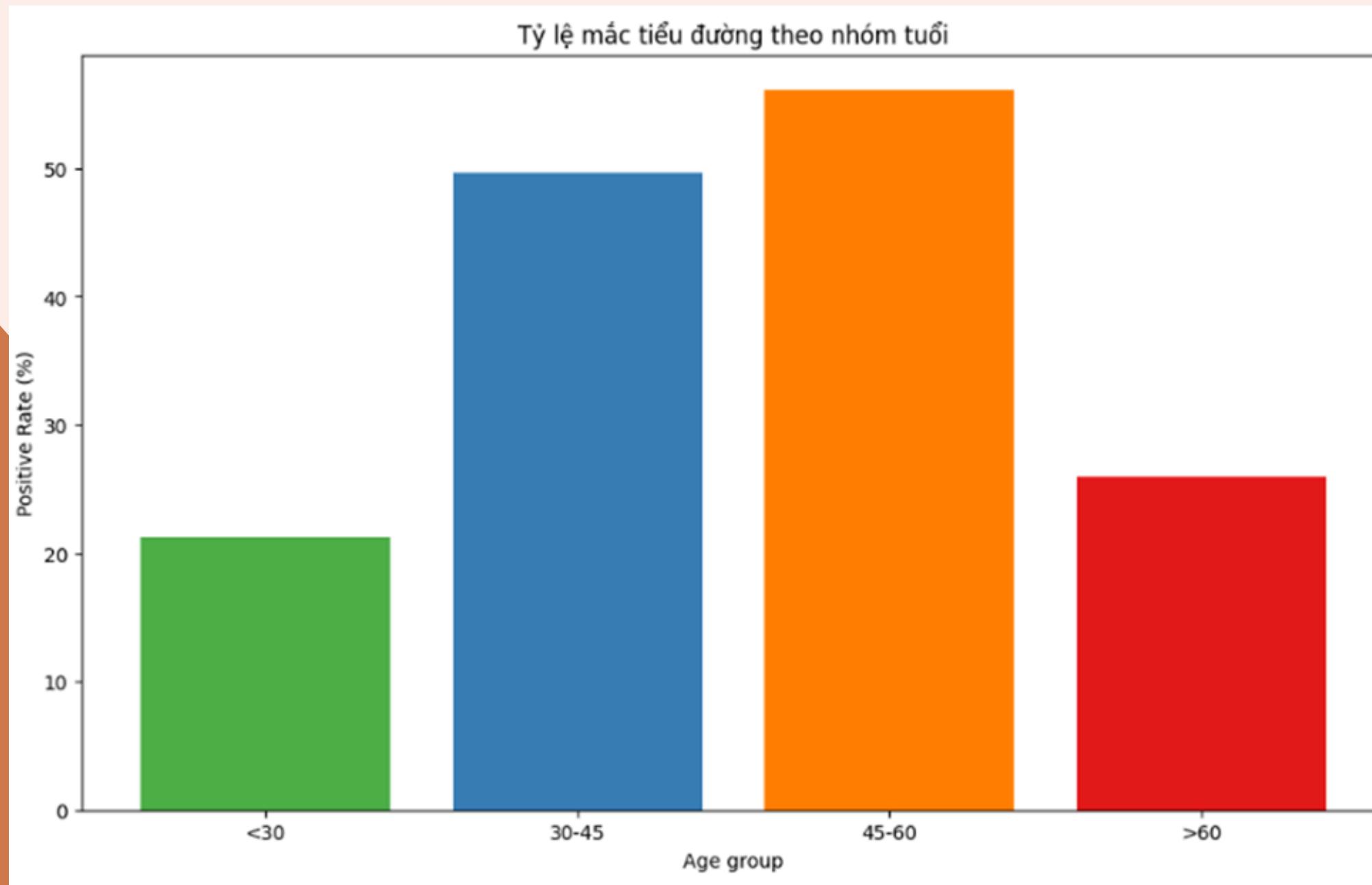
PHÂN TÍCH THEO NHÓM TUỔI (AGE GROUP WISE ANALYSIS)

Tính phần trăm người mắc bệnh tiểu đường theo nhóm tuổi:

```
age_group_stats["positive_rate"]=(age_group_stats["positive"]/age_group_stats["total"])*100  
age_group_stats
```

age_group	total	positive	positive_rate
<30	396	84	21.212121
30-45	254	126	49.606299
45-60	91	51	56.043956
>60	27	7	25.925926

PHÂN TÍCH THEO NHÓM TUỔI (AGE GROUP WISE ANALYSIS)



Nhận xét: Xu hướng chung: Tỷ lệ mắc tiểu đường tăng dần theo độ tuổi từ nhóm trẻ đến nhóm trung niên, sau đó giảm nhẹ ở nhóm cao tuổi.

TỈ LỆ NGƯỜI MẮC BỆNH TIỂU ĐƯỜNG THEO NHÓM TUỔI

PHÂN TÍCH THEO NHÓM TUỔI (AGE GROUP WISE ANALYSIS)

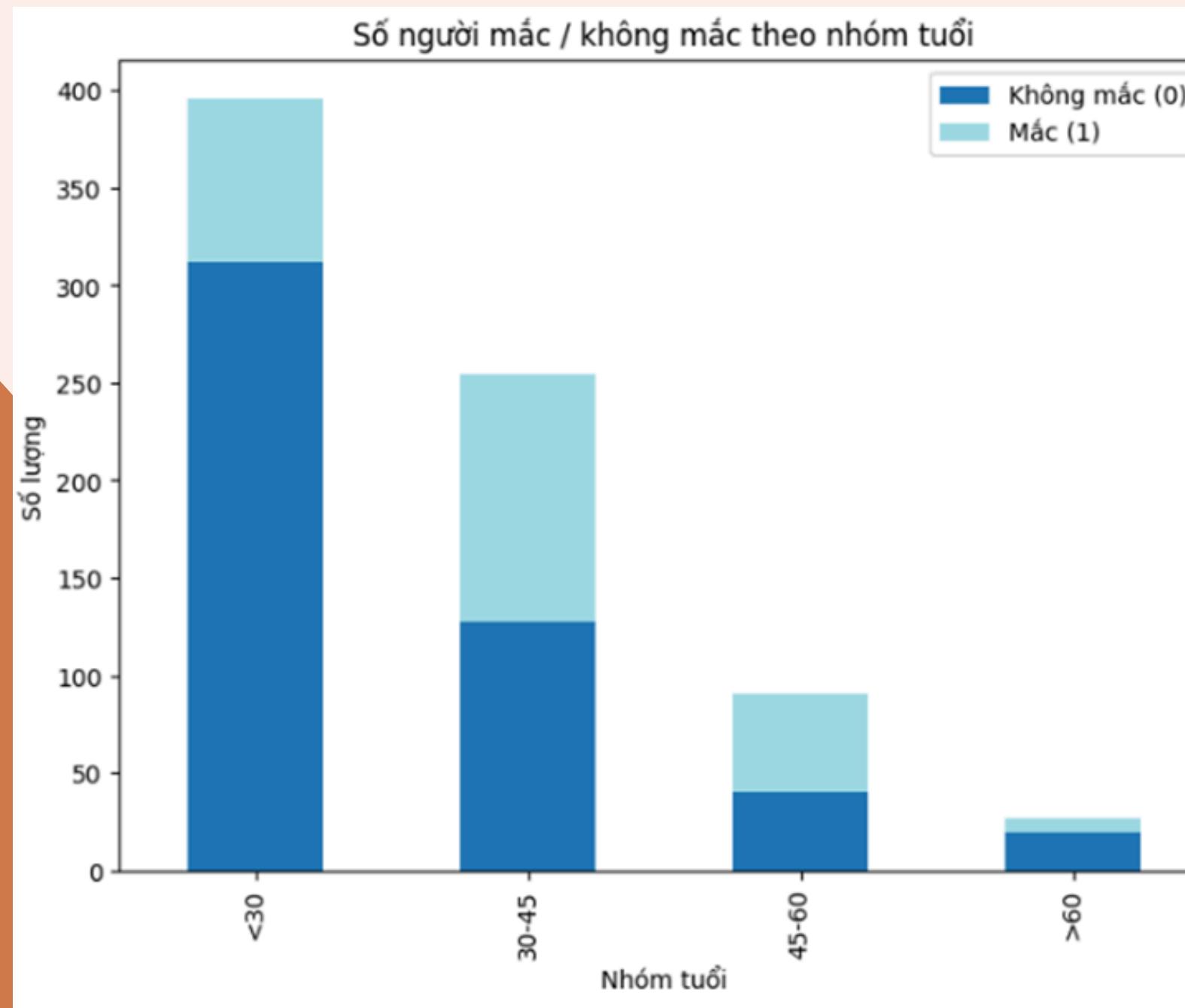
Nhóm < 30 tuổi: tỷ lệ mắc bệnh thấp (~21%), nhưng không bằng 0 → chứng tỏ vẫn có rủi ro ở tuổi trẻ, chủ yếu do glucose cao hoặc BMI cao.

Nhóm 30–45 tuổi: tỷ lệ mắc tăng mạnh (~50%), đồng thời đây cũng là nhóm có số ca mắc tuyệt đối cao nhất trong toàn bộ dataset.

Nhóm 45–60 tuổi: có tỷ lệ mắc bệnh cao nhất (~56%), phản ánh rõ rệt tác động của tuổi tác.

Nhóm >60 tuổi: tỷ lệ mắc giảm còn ~26%. Nguyên nhân có thể do kích thước mẫu nhỏ (ít người trên 60 trong dataset) → cần thận trọng khi diễn giải.

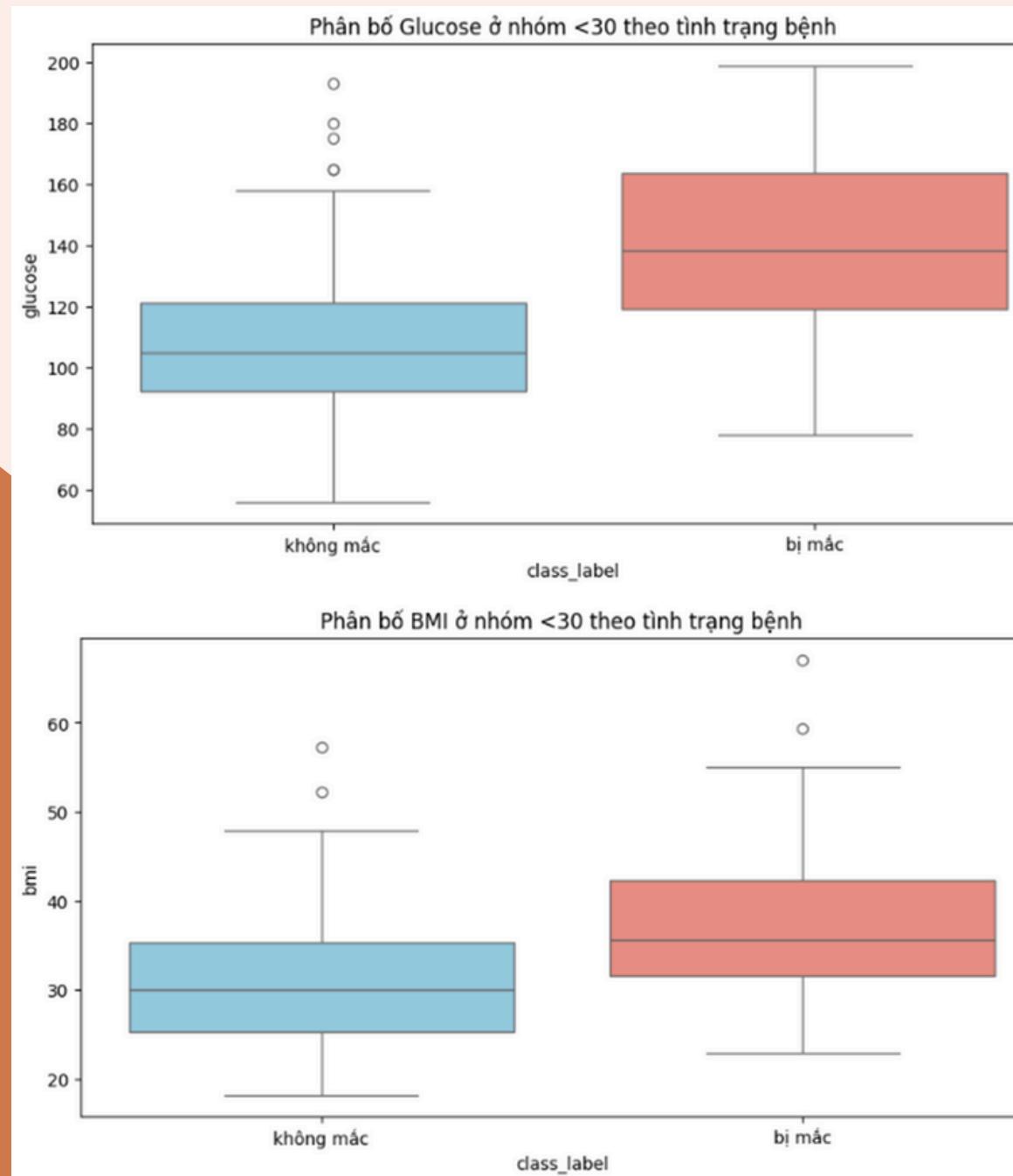
PHÂN TÍCH THEO NHÓM TUỔI (AGE GROUP WISE ANALYSIS)



Nhận xét:

- Nhóm 30–45 tuổi có 126 người mắc bệnh → là số ca mắc tuyệt đối cao nhất trong các nhóm tuổi của dataset
- Nhóm 45–60 mặc dù tỷ lệ % cao hơn (56%), nhưng số người mắc thực tế chỉ 51 → thấp hơn nhóm 30–45.

PHÂN TÍCH THEO NHÓM TUỔI (AGE GROUP WISE ANALYSIS)



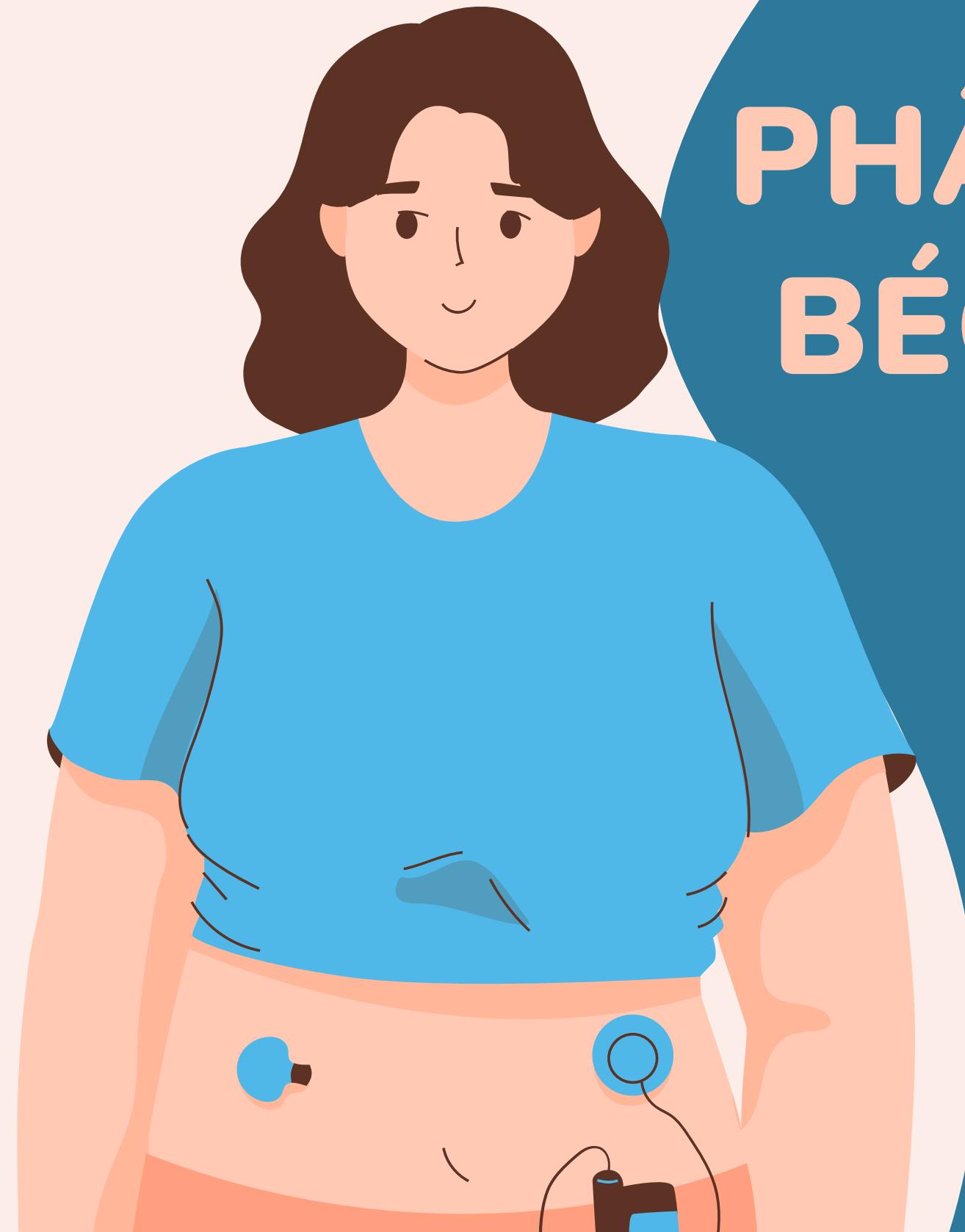
PHÂN BỐ GLUCOSE VÀ BMI THEO TÌNH TRẠNG
BỆNH Ở NHÓM TUỔI < 30

1. Biểu đồ Glucose

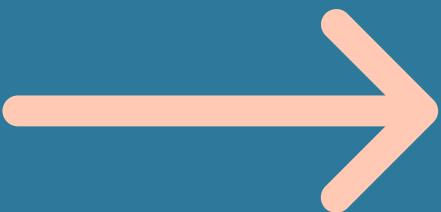
- Nhóm “bị mắc” có giá trị trung bình và trung vị (median) cao hơn hẳn nhóm “không mắc”.
- Khoảng giá trị (IQR) cũng lớn hơn, có nhiều outlier cao, nghĩa là một số người có glucose rất cao.
- Nhìn chung, glucose càng cao → khả năng bị tiểu đường càng lớn, điều này hợp lý về mặt y học.

2. Biểu đồ BMI

- Nhóm “bị mắc” có BMI trung vị cao hơn nhóm “không mắc”.
- Khoảng IQR cũng rộng hơn, có một vài outlier cao (người rất béo).
- Điều này cho thấy BMI cao có liên quan tới nguy cơ tiểu đường ngay cả ở nhóm < 30 tuổi.



PHÂN TÍCH THEO MỨC ĐỘ BÉO PHÌ (BMI CATEGORY ANALYSIS)



PHÂN TÍCH THEO MỨC ĐỘ BÉO PHÌ (BMI CATEGORY ANALYSIS)

CÂU HỎI PHÂN TÍCH:

- TỶ LỆ MẮC BỆNH CÓ TĂNG RỘ RỆT THEO CÁC CẤP ĐỘ BMI KHÔNG?
- TRONG NHÓM BMI "BÌNH THƯỜNG", CÓ BAO NHIÊU % VẪN MẮC BỆNH? ĐẶC ĐIỂM CỦA HỌ LÀ GÌ? (CÓ THỂ DO DI TRUYỀN - DIABETES PREDIGREE FUNCTION CAO).

PHÂN TÍCH THEO MỨC ĐỘ BÉO PHÌ (BMI CATEGORY ANALYSIS)

Chia nhóm các mức độ béo phì:

```
bmi_bins = [0, 18.5, 24.9, 29.9, 100]
bmi_labels = ["Thiếu cân (<18.5)", "Bình thường (18.5-24.9)", "Thừa cân (25-29.9)", "Béo phì ( $\geq$ 30)"]

diabetes_df["bmi_group"] = pd.cut(diabetes_df["bmi"], bins=bmi_bins,
labels=bmi_labels, right=True)

# Quan sát thử
diabetes_df[["bmi", "bmi_group"]].head(10)
```

	bmi	bmi_group
0	33.6	Béo phì (\geq 30)
1	26.6	Thừa cân (25-29.9)
2	23.3	Bình thường (18.5-24.9)
3	28.1	Thừa cân (25-29.9)
4	43.1	Béo phì (\geq 30)
5	25.6	Thừa cân (25-29.9)
6	31.0	Béo phì (\geq 30)
7	35.3	Béo phì (\geq 30)
8	30.5	Béo phì (\geq 30)
9	NaN	NaN

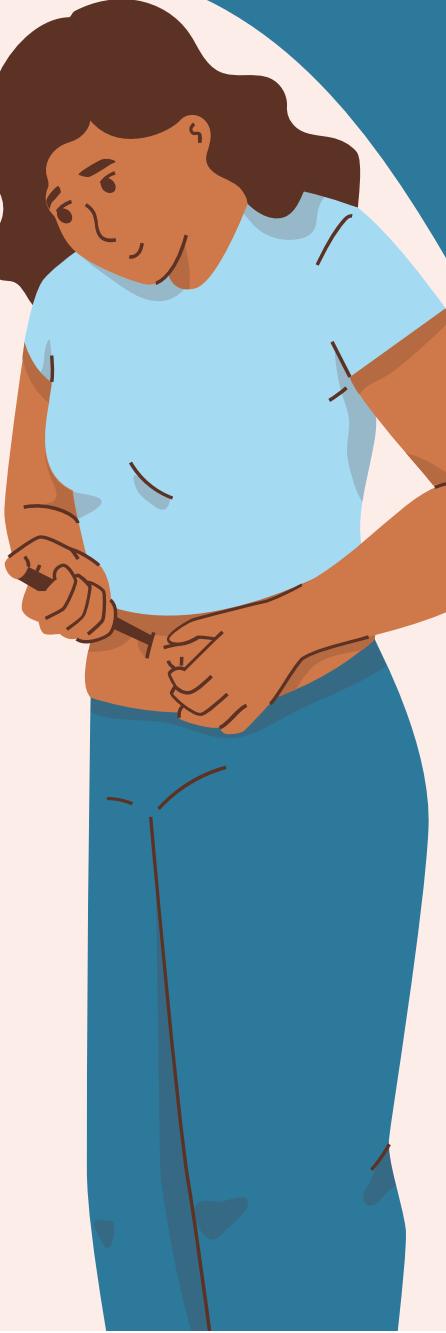


PHÂN TÍCH THEO MỨC ĐỘ BÉO PHÌ (BMI CATEGORY ANALYSIS)

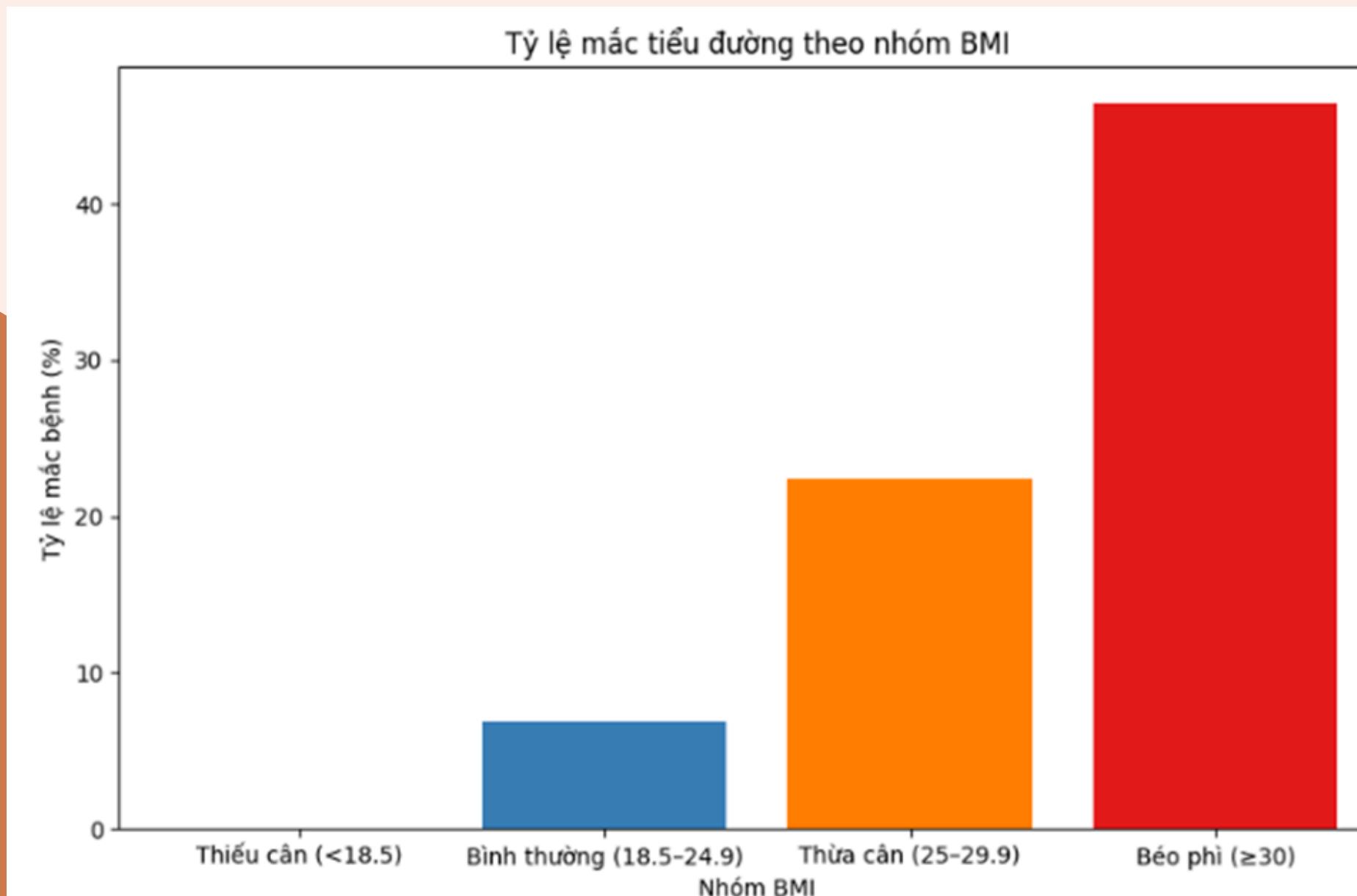
Tính tổng người theo nhóm và tỉ lệ mắc bệnh tương ứng

```
bmi_group_stats = diabetes_df.groupby("bmi_group",
observed=True)[ "outcome" ].agg(
    total="count",
    positive="sum"
)
bmi_group_stats[ "positive_rate" ] = (bmi_group_stats[ "positive" ] /
bmi_group_stats[ "total" ]) * 100
bmi_group_stats
```

bmi_group	total	positive	positive_rate
Thiếu cân (<18.5)	4	0	0.000000
Bình thường (18.5–24.9)	102	7	6.862745
Thừa cân (25–29.9)	179	40	22.346369
Béo phì (≥ 30)	472	219	46.398305



PHÂN TÍCH THEO MỨC ĐỘ BÉO PHÌ (BMI CATEGORY ANALYSIS)



Tỉ lệ mắc tiểu đường theo các nhóm BMI

Nhận xét:

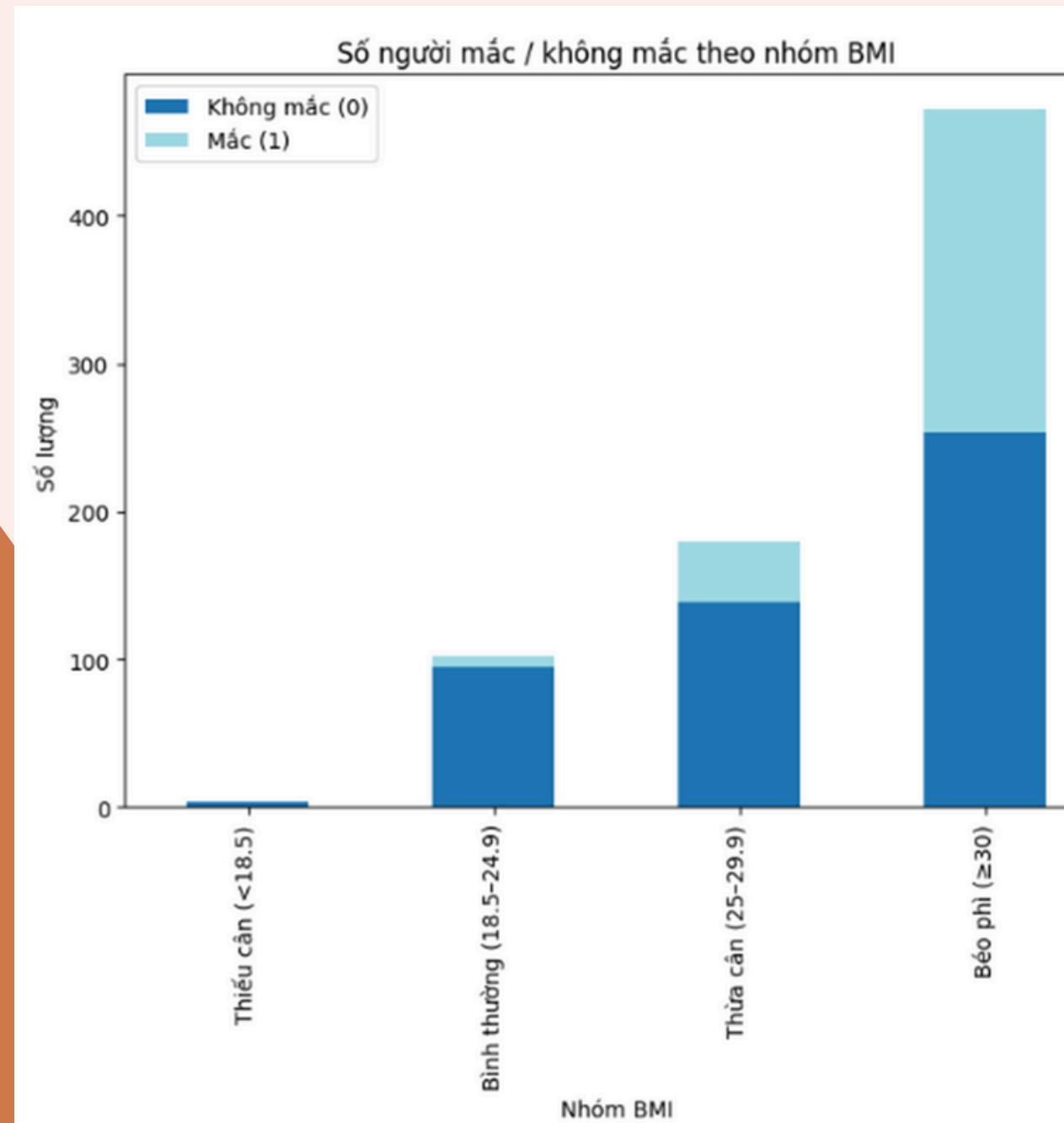
Tỷ lệ mắc tiểu đường tăng rõ rệt theo cấp độ BMI.

Nhóm Thiếu cân và Bình thường có tỷ lệ rất thấp (dưới 10%).

Nhóm Thừa cân tăng lên khoảng 20–25%.

Nhóm Béo phì (≥ 30) cao nhất, vượt 40%, cho thấy béo phì là yếu tố nguy cơ mạnh.

PHÂN TÍCH THEO MỨC ĐỘ BÉO PHÌ (BMI CATEGORY ANALYSIS)



Số người mắc / không mắc theo nhóm BMI

Nhận xét:

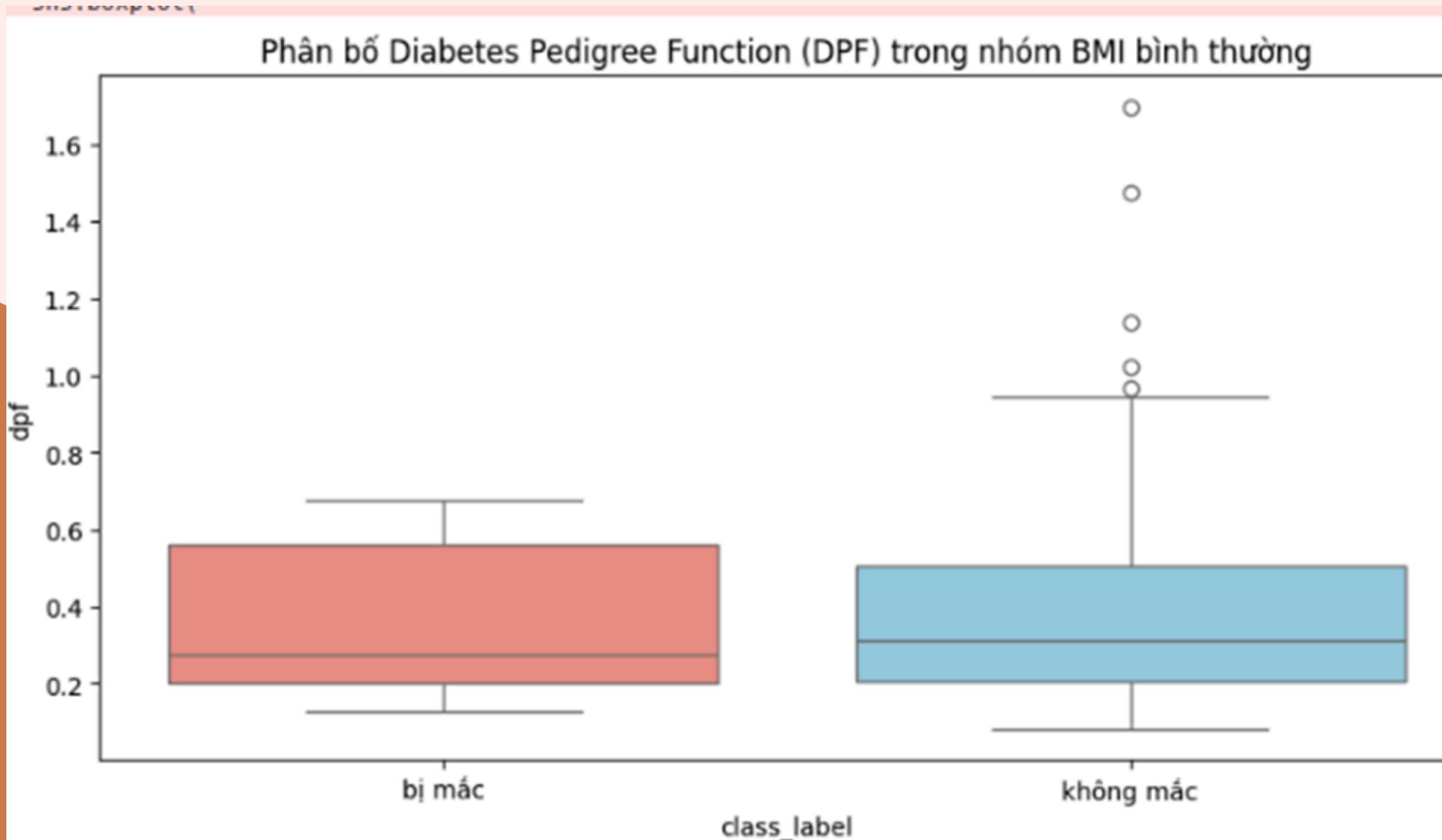
Nhóm Béo phì (≥ 30) có số lượng người nhiều nhất và cũng chiếm số ca mắc bệnh tuyệt đối cao nhất.

Nhóm Thừa cân có số ca mắc tăng lên nhưng vẫn thấp hơn nhiều so với nhóm béo phì.

Nhóm Bình thường vẫn có một tỷ lệ nhỏ bị mắc bệnh, gợi ý khả năng liên quan đến yếu tố di truyền hoặc glucose bất thường, không chỉ do cân nặng.

Nhóm Thiếu cân rất ít người và gần như không có ca mắc.

PHÂN TÍCH THEO MỨC ĐỘ BÉO PHÌ (BMI CATEGORY ANALYSIS)



Phân bố Diabetes Pedigree Function (DPF) trong nhóm BMI bình thường

Nhận xét :

Nhóm BMI Bình thường nhưng vẫn mắc tiểu đường (~15–20%).

Các bệnh nhân này thường có DPF cao, nghĩa là có yếu tố di truyền/tiền sử gia đình. Ngoài ra có thể kết hợp xem Glucose trong nhóm này, để thấy rằng tuy cân nặng bình thường nhưng lượng đường máu cao bất thường cũng là nguyên nhân.

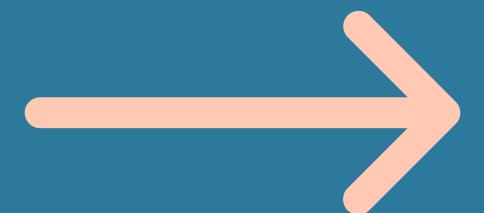
PHÂN TÍCH THEO MỨC ĐỘ BÉO PHÌ (BMI CATEGORY ANALYSIS)

Tỷ lệ mắc bệnh tăng dần theo BMI: Thiếu cân < Bình thường < Thừa cân < Béo phì.

Trong nhóm BMI bình thường, những người mắc bệnh chủ yếu do yếu tố di truyền (dpf cao) hoặc Glucose bất thường cao, không chỉ do cân nặng.



PHÂN TÍCH THEO NGƯỜNG ĐƯỜNG HUYẾT (GLUCOSE THRESHOLD ANALYSIS)



PHÂN TÍCH THEO NGƯỠNG ĐƯỜNG HUYẾT (GLUCOSE THRESHOLD ANALYSIS)

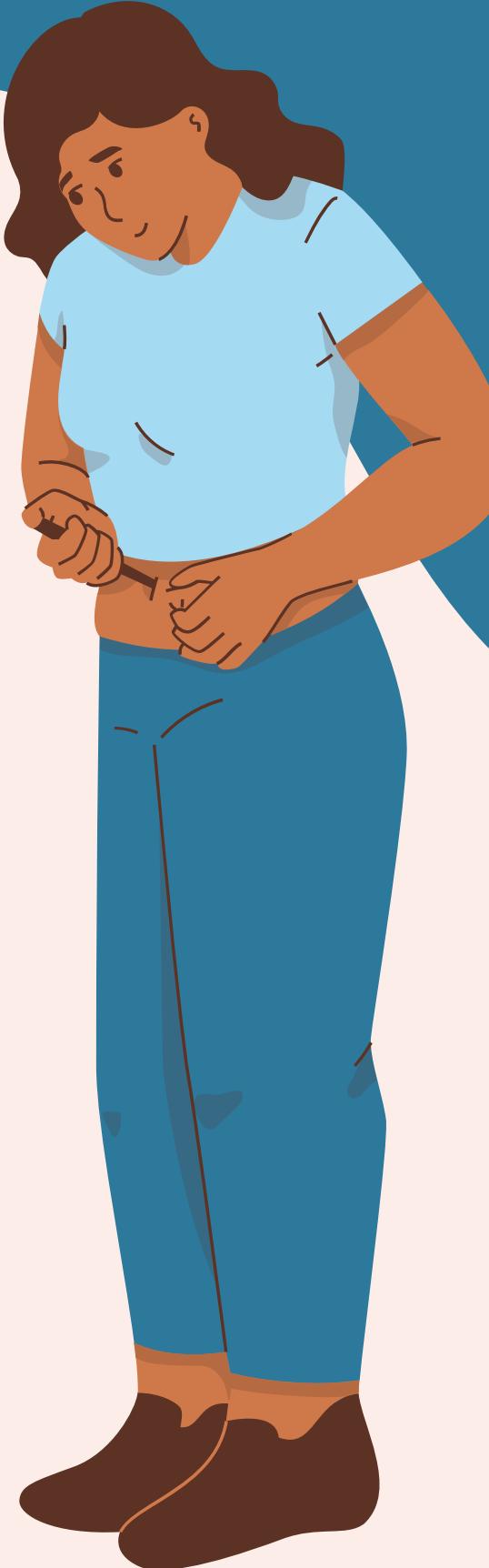
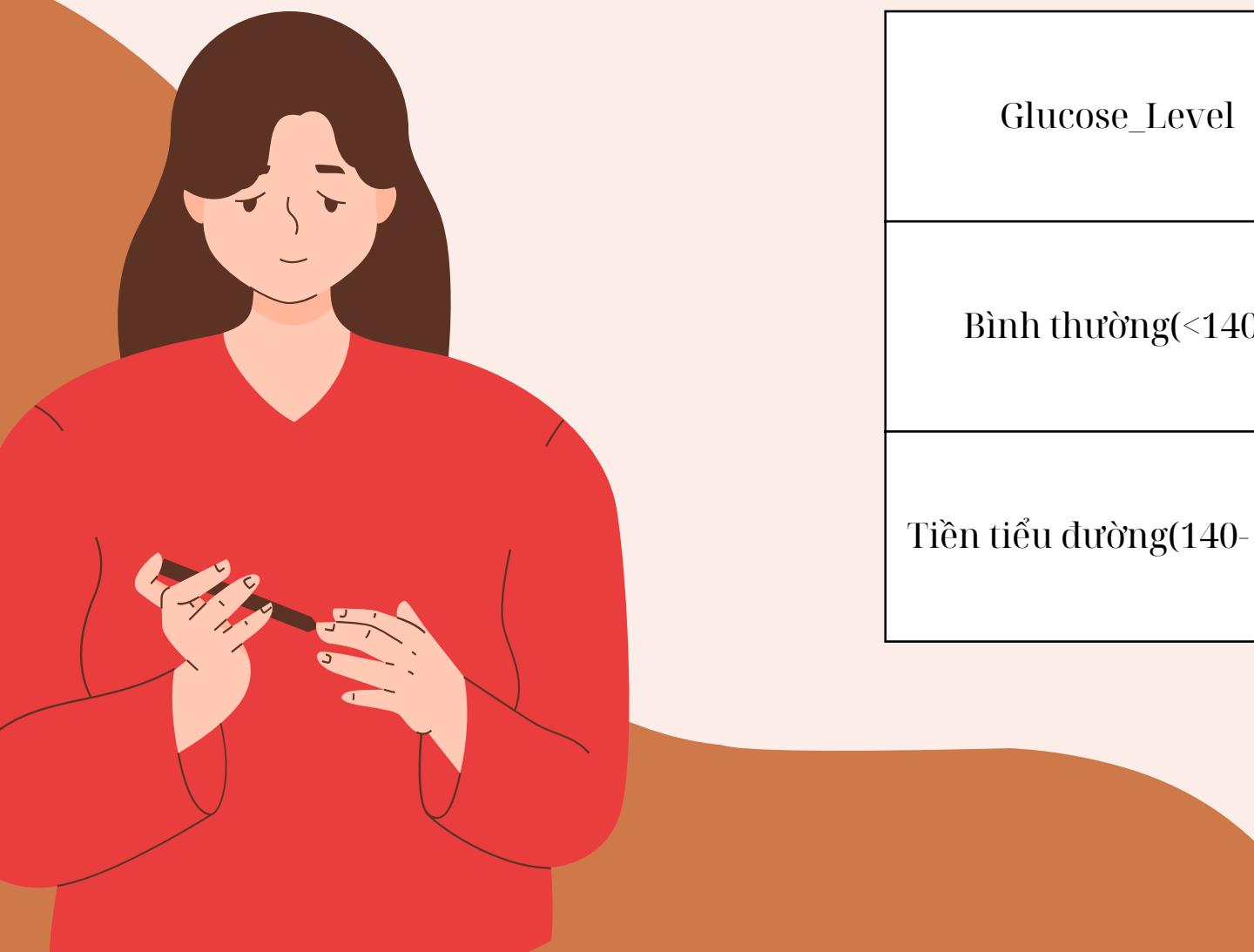
CÂU HỎI PHÂN TÍCH:

- TRONG SỐ NHỮNG NGƯỜI ĐƯỢC MÔ HÌNH DỰ ĐOÁN LÀ MẮC BỆNH (OUTCOME=1), CÓ BAO NHIÊU % THỰC SỰ ĐÃ Ở NGƯỠNG ĐƯỜNG HUYẾT CHẨN ĐOÁN TIỂU ĐƯỜNG?
- CÓ TRƯỜNG HỢP NÀO GLUCOSE Ở MỨC "BÌNH THƯỜNG" HOẶC "TIỀN TIỂU ĐƯỜNG" NHƯNG VẪN ĐƯỢC CHẨN ĐOÁN MẮC BỆNH (CLASS=1) KHÔNG? NẾU CÓ, CÁC YẾU TỐ KHÁC (NHƯ INSULIN, AGE) CỦA HỌ THẾ NÀO?

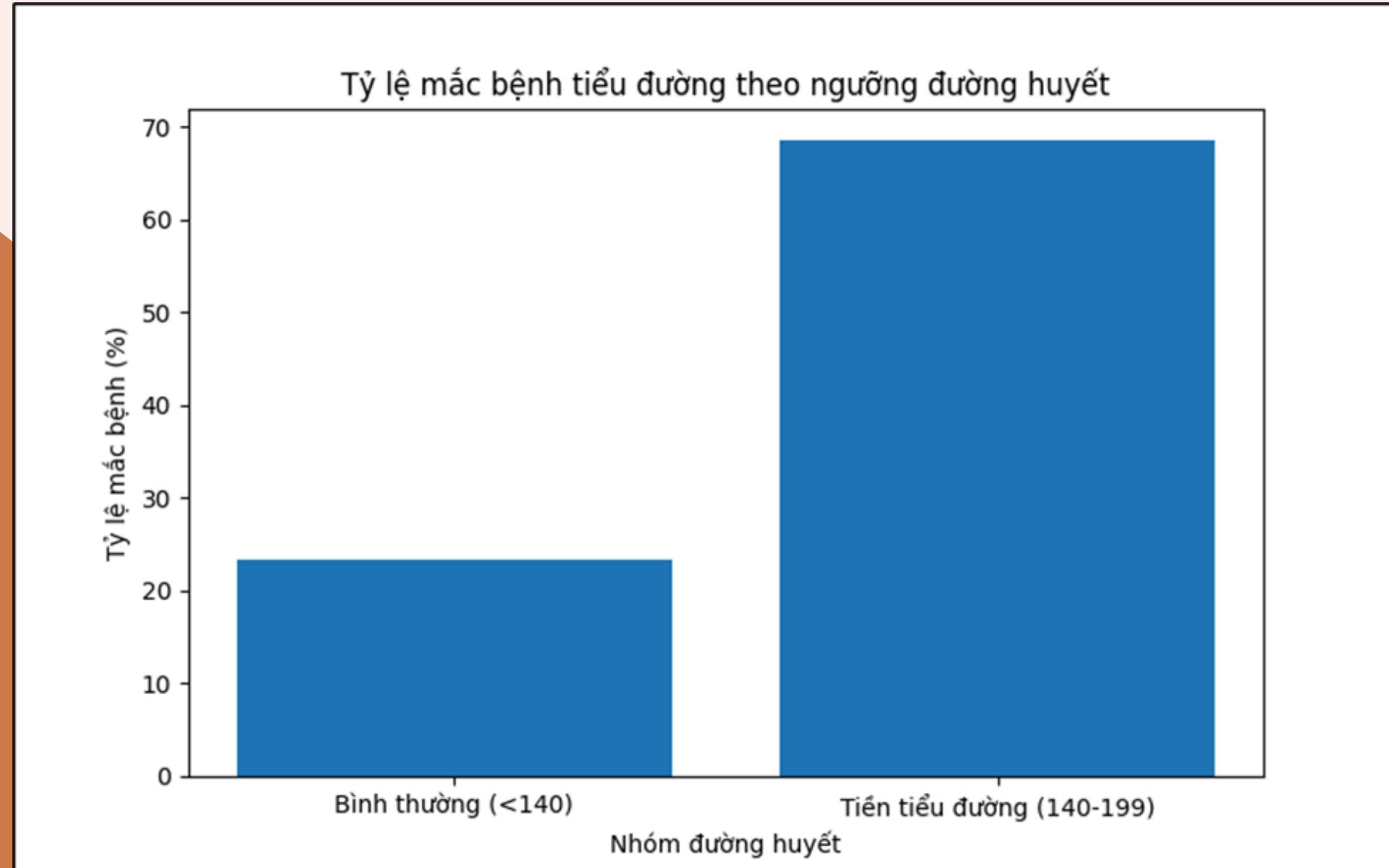
PHÂN TÍCH THEO NGƯỠNG ĐƯỜNG HUYẾT (GLUCOSE THRESHOLD ANALYSIS)

Dữ liệu thống kê theo cột Glucose:

Glucose_Level	Tổng mẫu	Số mắc bệnh	Tỷ lệ mắc bệnh
Bình thường(<140)	571	133	23.29%
Tiền tiểu đường(140- 199)	197	135	68.53%



PHÂN TÍCH THEO NGƯỠNG ĐƯỜNG HUYẾT (GLUCOSE THRESHOLD ANALYSIS)



Tỷ lệ mắc bệnh tiểu đường theo ngưỡng đường huyết

Từ biểu đồ trên ta có thể thấy:

- Những nhóm người có nồng độ đường huyết trong máu cao(140-199) thì có nguy cơ mắc bệnh tiểu đường cao.
- Ngược lại những người có nồng độ đường huyết trong máu ở mức bình thường thì tỉ lệ mắc bệnh ít hơn.

PHÂN TÍCH THEO NGƯỠNG ĐƯỜNG HUYẾT (GLUCOSE THRESHOLD ANALYSIS)

Những người có nồng độ Glucose trong máu thấp mà vẫn bị thì thường liên quan đến các yếu tố age , BMI, pdf, pregnancies. Dưới đây là bảng thống kê:

```
Tổng mẫu trong nhóm Glucose < 140: 571
```

```
Outcome
```

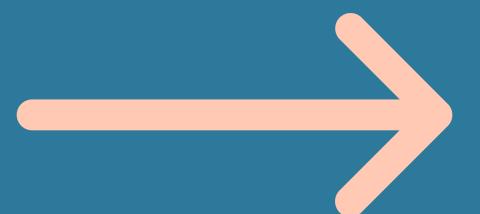
```
0    438  
1    133
```

```
Name: count, dtype: int64
```

```
So sánh từng biến (nhóm Glucose thấp):
```

feature	mean_out0	mean_out1	test	p_value
Age	30.395	35.361	Mann-Whitney U	0.000000
BMI	29.911	34.491	Mann-Whitney U	0.000000
Pregnancies	3.205	4.835	Mann-Whitney U	0.000027
DiabetesPedigreeFunction	0.420	0.552	Mann-Whitney U	0.000093
BloodPressure	67.153	68.662	Mann-Whitney U	0.018869
SkinThickness	19.299	21.008	Mann-Whitney U	0.113893
Insulin	58.384	67.594	Mann-Whitney U	0.902003

PHÂN TÍCH THEO TIỀN SỬ GIA ĐÌNH (PEDIGREE FUNCTION ANALYSIS)



PHÂN TÍCH THEO TIỀN SỬ GIA ĐÌNH (PEDIGREE FUNCTION ANALYSIS)

CÂU HỎI PHÂN TÍCH:

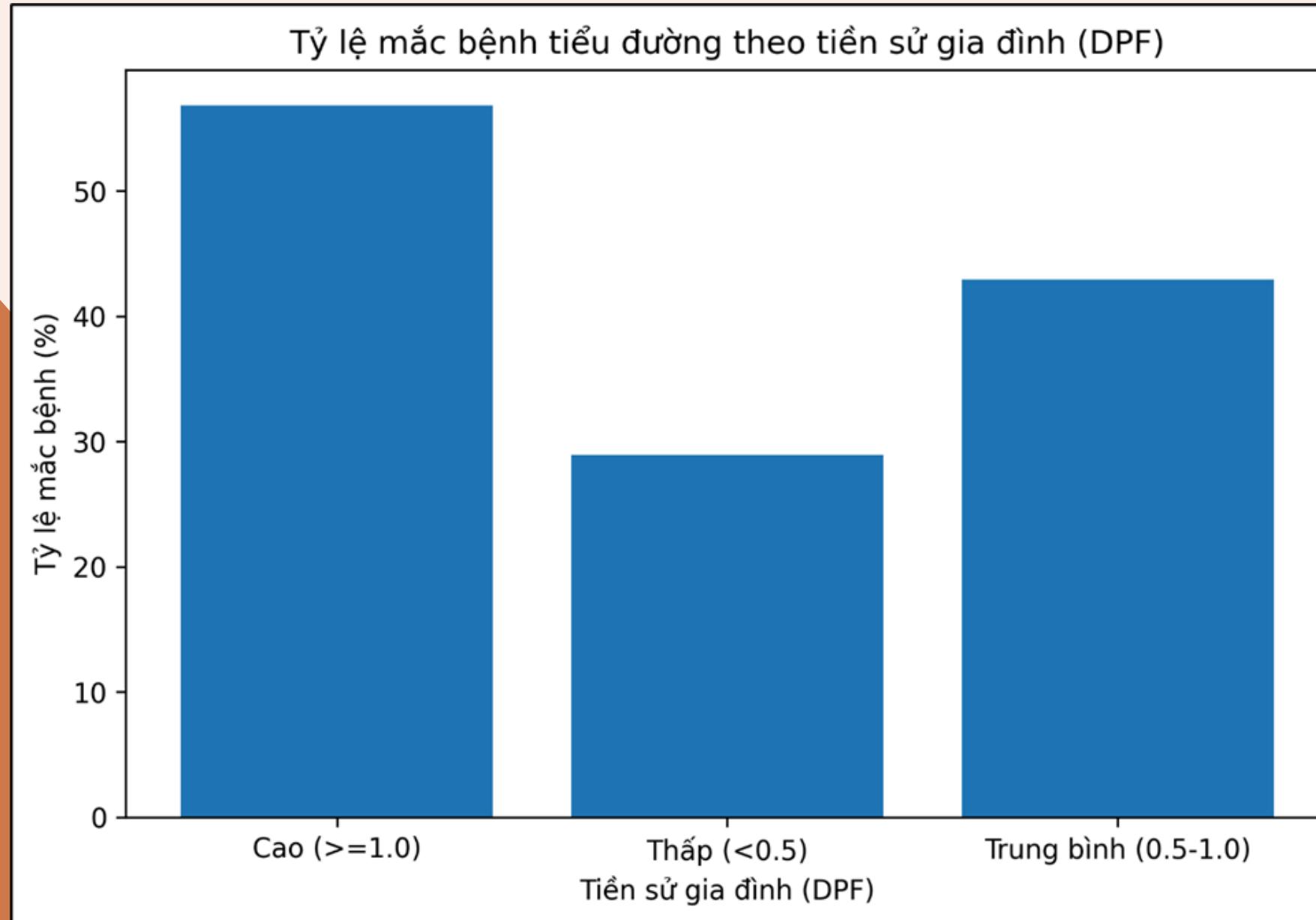
- VỚI NHỮNG NGƯỜI CÓ CÙNG MỨC BMI HOẶC TUỔI TÁC, NHÓM CÓ TIỀN SỬ GIA ĐÌNH (PEDIGREE) NẶNG HƠN CÓ TỶ LỆ MẮC BỆNH CAO HƠN KHÔNG?
- YẾU TỐ DI TRUYỀN CÓ MẠNH ĐẾN MỨC "LẤN ÁT" CÁC YẾU TỐ NGUY CƠ KHÁC KHÔNG? (VÍ DỤ: MỘT NGƯỜI TRẺ, GẦY NHƯNG CÓ PEDIGREE RẤT CAO THÌ NGUY CƠ THẾ NÀO?).

PHÂN TÍCH THEO TIỀN SỬ GIA ĐÌNH (PEDIGREE FUNCTION ANALYSIS)

Dữ liệu thống kê theo cột DiabetesPedigreeFunction(dpf)

Family_History	Tổng mẫu	Số mắc bệnh	Tỉ lệ mắc bệnh
Cao (≥ 1.0)	51	29	56.86%
Trung bình(0.5 - 1)	491	142	42.92%
Thấp (< 0.5)	226	97	28.92%

PHÂN TÍCH THEO TIỀN SỬ GIA ĐÌNH (PEDIGREE FUNCTION ANALYSIS)

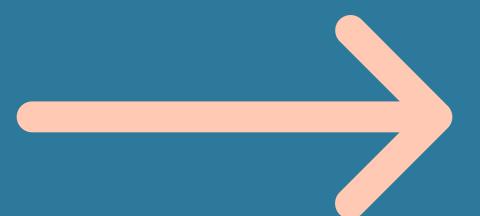


Tỷ lệ mắc bệnh tiểu đường theo tiền sử gia đình

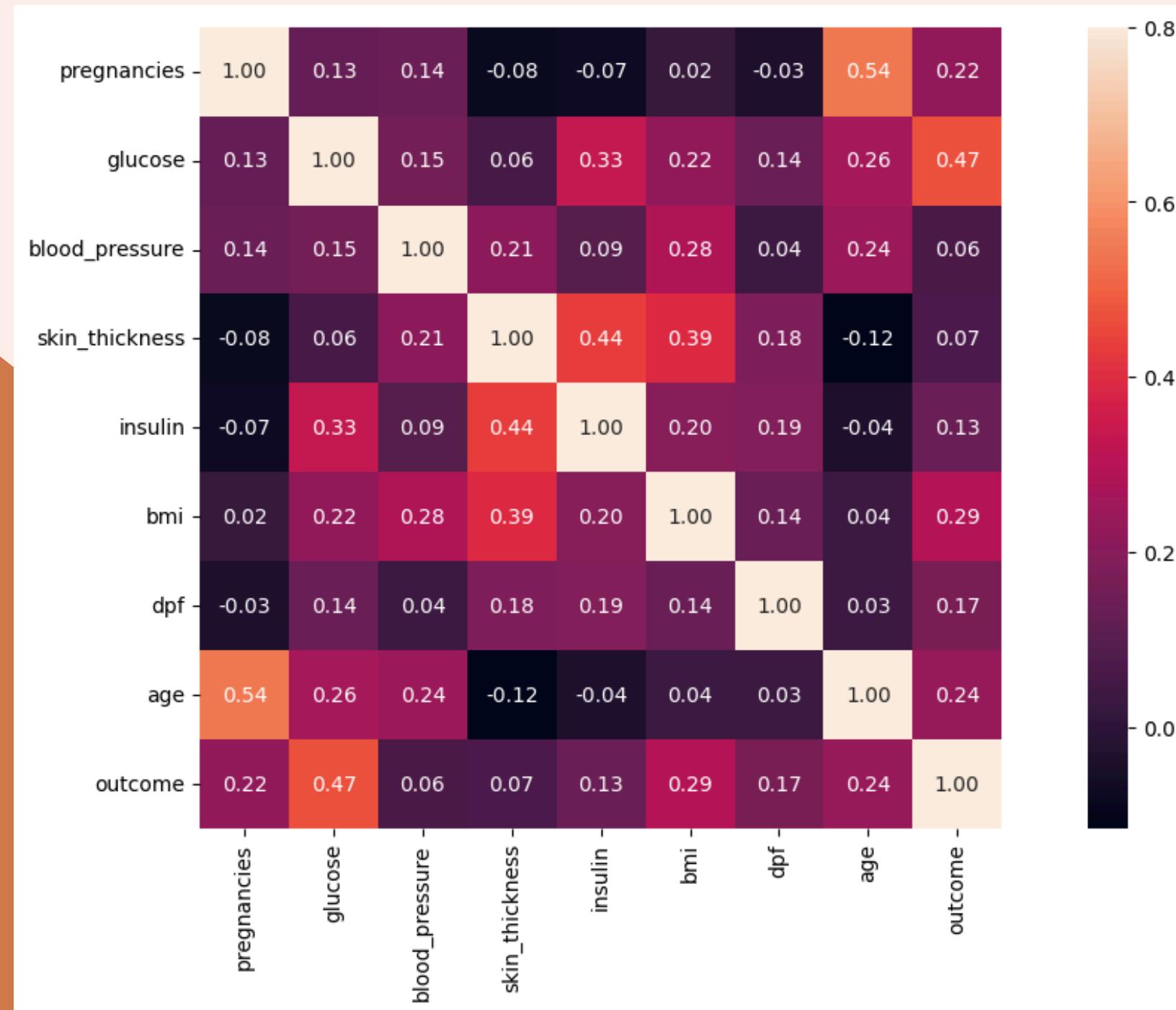
Từ biểu đồ ta có thể thấy:

- Những người có tiền sử gia đình mắc bệnh cao(>1.0) thì có tỉ lệ mắc bệnh cao
- Những người có tiền sử gia đình mắc bệnh ở mức trung bình thì có tỉ lệ mắc bệnh ở mức trung bình
- Những người có tiền sử gia đình mắc bệnh thấp thì tỉ lệ mắc bệnh thấp

BẢN ĐỒ NHIỆT TƯƠNG QUAN (CORRELATION HEATMAP)



BẢN ĐỒ NHIỆT TƯƠNG QUAN (CORRELATION HEATMAP)

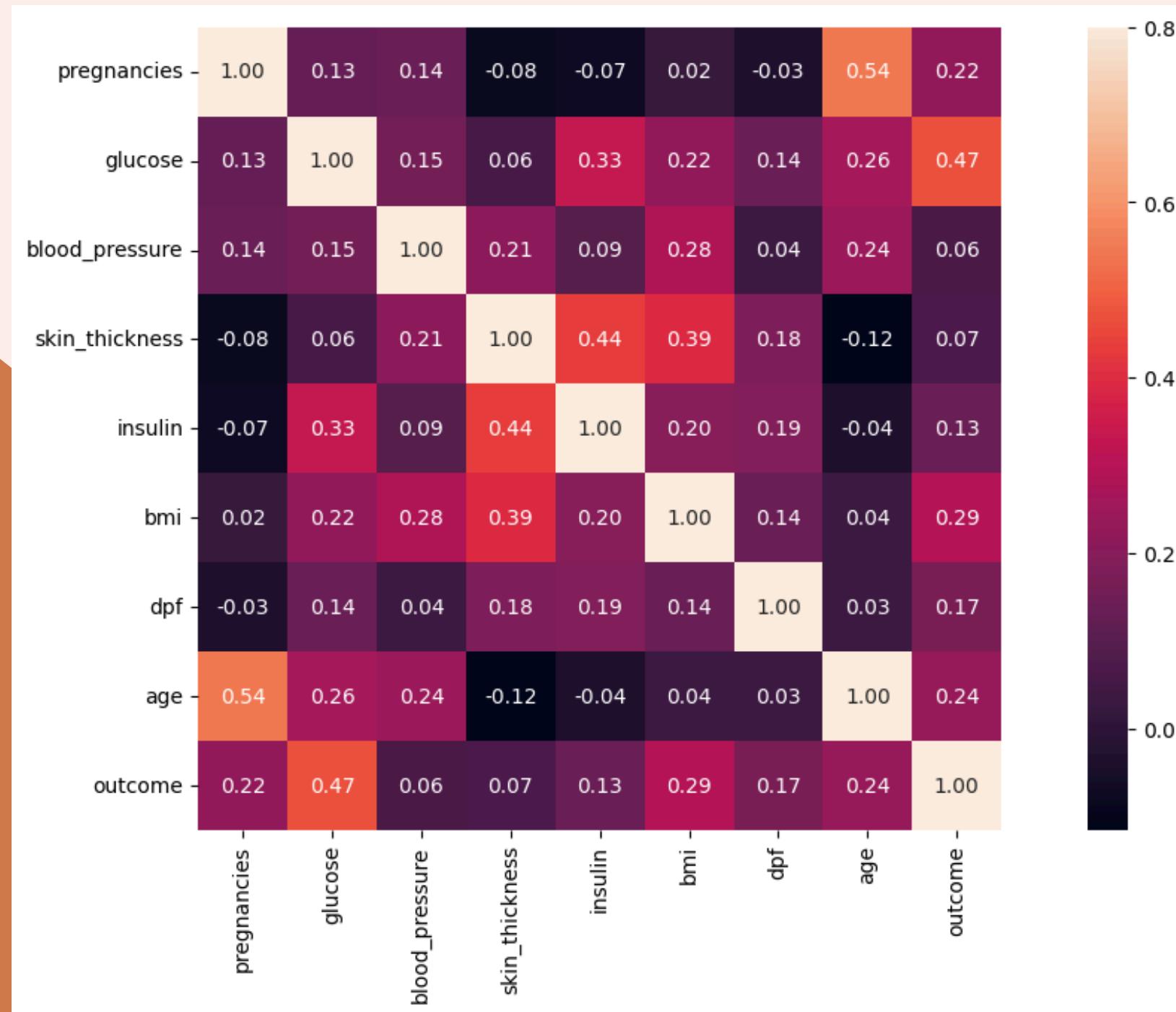


Phân tích chi tiết các tương quan:

Một số cặp tương quan mạnh ($|r| > 0.4$):

- **Glucose vs Outcome (0.47):** Đây là tương quan quan trọng nhất, phù hợp với y học vì glucose máu là chỉ số chính chẩn đoán tiểu đường
- **Pregnancies vs Age (0.54):** Hoàn toàn hợp lý - phụ nữ càng lớn tuổi càng có nhiều khả năng đã mang thai nhiều lần
- **Skin thickness vs Insulin (0.44):** Có thể phản ánh mối liên hệ giữa độ dày da (liên quan đến mỡ dưới da) và kháng insulin

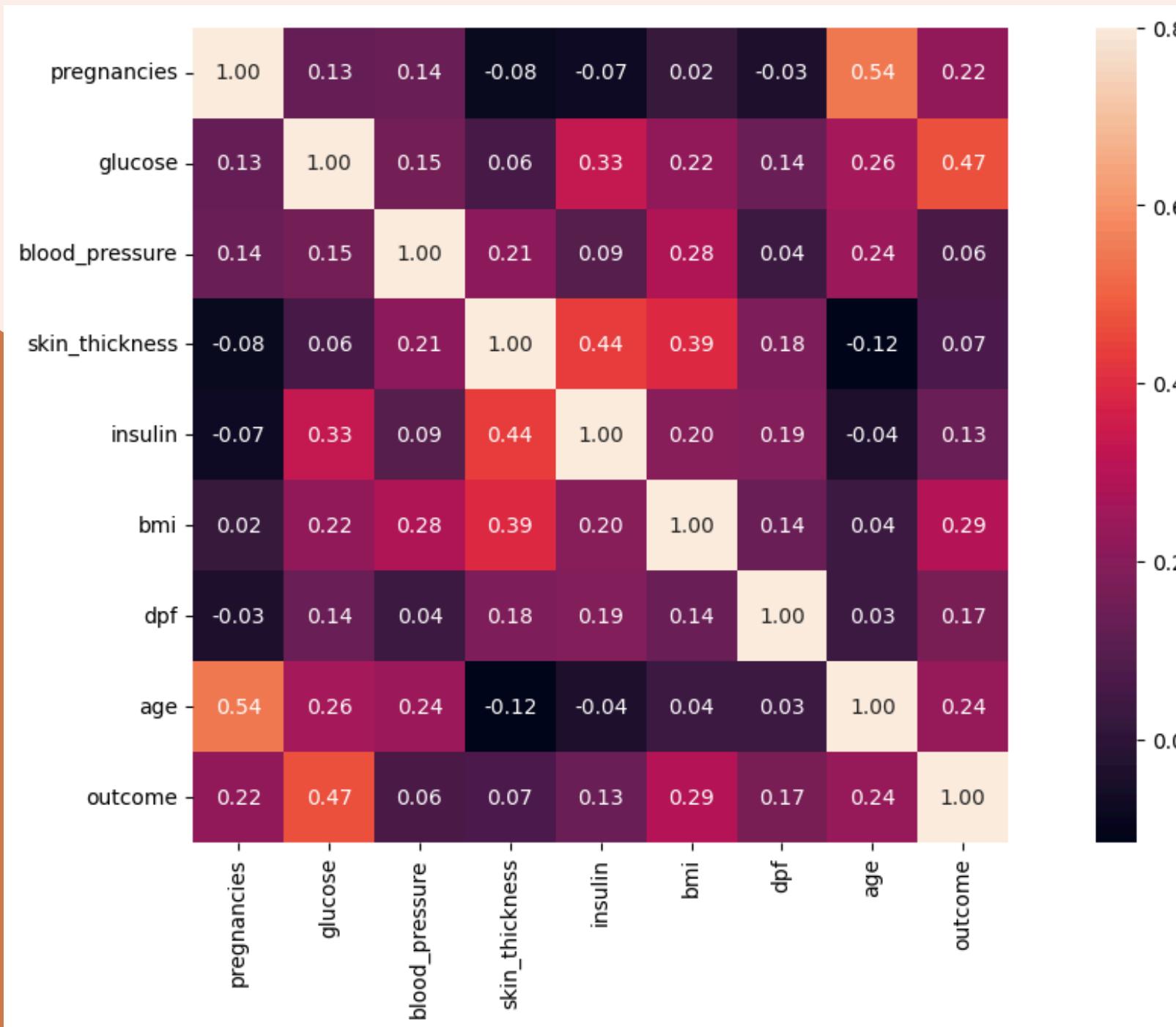
BẢN ĐỒ NHIỆT TƯƠNG QUAN (CORRELATION HEATMAP)



Một số cặp tương quan trung bình ($0.2 < |r| < 0.4$):

- Glucose vs Insulin (0.33): Insulin điều hòa glucose trong máu
- Glucose vs BMI (0.22): BMI cao thường đi kèm glucose cao
- Blood pressure vs BMI (0.28): Huyết áp và cân nặng có mối liên hệ
- Skin thickness vs BMI (0.39): Độ dày da liên quan đến chỉ số khối cơ thể

BẢN ĐỒ NHIỆT TƯƠNG QUAN (CORRELATION HEATMAP)



Một số cặp tương quan yếu ($|r| < 0.2$):
Hầu hết các cặp biến còn lại có tương quan rất yếu hoặc gần như không có

- Pregnancies vs BMI (0.02): gần như không tương quan
- Pregnancies vs DPF (-0.03): gần như không tương quan
- Blood pressure vs DPF (0.04): rất yếu
- Insulin vs Age (-0.04): rất yếu

KẾT LUẬN (CONCLUSION)

- Tỉ lệ người mắc bệnh tiểu đường là 34.8% và 65.2% mẫu không mắc bệnh tiểu đường.
- Tỉ lệ người mắc bệnh tiểu đường ở độ tuổi 30 - 45 (50%) và 45 - 60 (56%) đạt tỉ lệ cao hơn nhiều so với các nhóm tuổi khác.
- Tỉ lệ người mắc bệnh tiểu đường tăng dần theo BMI: Thiếu cân (0%) < Bình thường (6.86%) < Thừa cân (22.35%) < Béo phì (46.4%).
- Tỉ lệ người mắc bệnh tiểu đường tăng dần theo lượng Glucose trong máu: lượng Glucose < 140mmHg (23.29%); lượng Glucose từ 140 - 190mmHg (68.53%).
- Tỉ lệ người mắc bệnh tiểu đường tỉ lệ thuận với tiền sử gia đình mắc bệnh: thấp (28.92%) < trung bình (42.92%) < cao (56.86%).



THANK YOU

