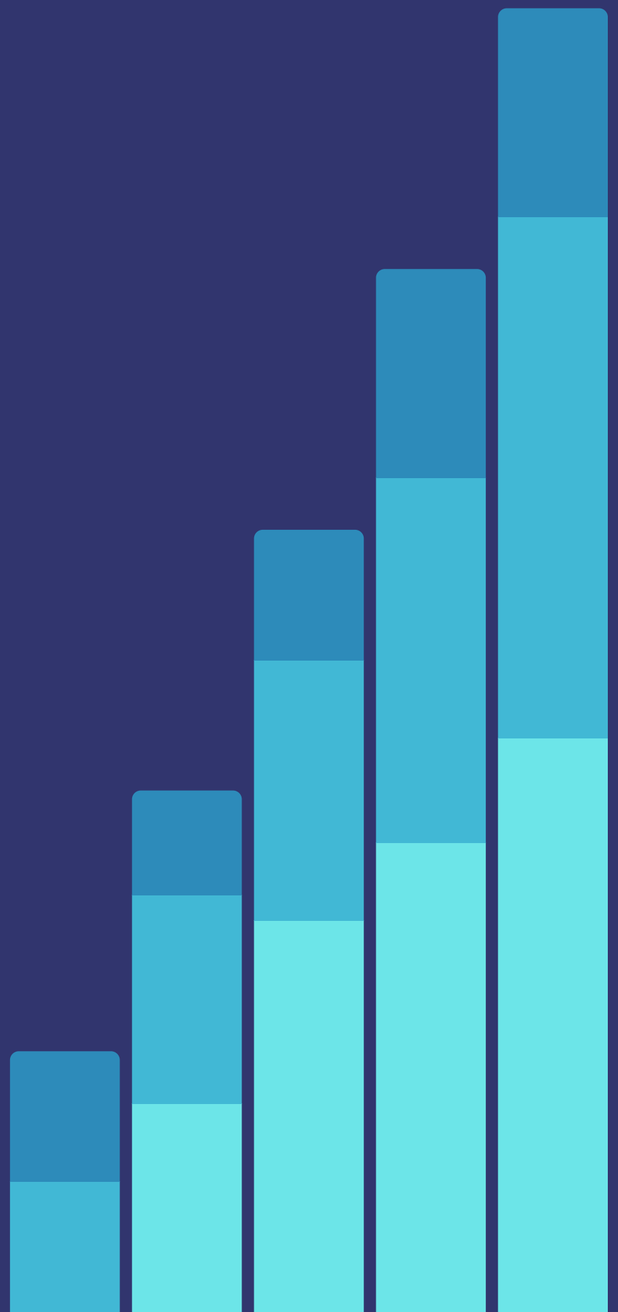
HỌC PHẦN: PHÂN TÍCH DỮ LIỆU

GIẢNG VIÊN HƯỚNG DẪN: THS. ĐỖ NHƯ TÀI

PHÂN TÍCH KHÁM PHÁ VỀ BỆNH ĐÁI THÁO ĐƯỜNG

NGUYỄN QUỐC KHÁNH - 3122410180
TRẦN ĐĂNG KHOA - 3122410186
LÊ THỊ UYÊN NHI - 3122410280



1. Mô tả bài toán
2. Xử lý dữ liệu
3. Phân tích đơn biến
4. Phân tích đa biến
5. Một số phát hiện chính & Kết luận

1. MÔ TẢ BÀI TOÁN

- Định nghĩa:

Đái tháo đường (Diabetes Mellitus) là một nhóm bệnh rối loạn chuyển hóa đặc trưng bởi tình trạng tăng glucose máu mạn tính do hậu quả của sự thiếu hụt tiết insulin, giảm tác dụng của insulin, hoặc cả hai.

Bệnh đái tháo đường type 2 là bệnh mạn tính phổ biến, gây nhiều biến chứng nguy hiểm.

Một số tài liệu nghiên cứu liên quan

1. Classification and Diagnosis of Diabetes Mellitus and Other Categories of Glucose Intolerance - NATIONAL DIABETES DATA GROUP (1979)

Phân loại làm 4 nhóm chính:

- Bệnh tiểu đường phụ thuộc insulin (IDDM, Type I)
- Bệnh tiểu đường không phụ thuộc insulin (NIDDM, Type II)
- Tiểu đường thai kỳ (Gestational Diabetes Mellitus)
- Các loại tiểu đường khác: Tiểu đường thứ phát do các bệnh hoặc tình trạng khác gây ra.

Đã giới thiệu thuật ngữ "Không dung nạp glucose" (IGT) để thay thế các thuật ngữ cũ, gây nhầm lẫn.

Một số tài liệu nghiên cứu liên quan

1. Classification and Diagnosis of Diabetes Mellitus and Other Categories of Glucose Intolerance - NATIONAL DIABETES DATA GROUP (1979)

Các tiêu chí chẩn đoán

- Glucose lúc đói (FPG): ≥ 140 mg/dL (7.8 mmol/L).
- Glucose sau 2 giờ làm OGTT: ≥ 200 mg/dL (11.1 mmol/L).
- Glucose ngẫu nhiên: ≥ 200 mg/dL ở những người có triệu chứng điển hình.

Đây là tiêu chuẩn trong gần hai thập kỷ, thay thế những tiêu chuẩn không thống nhất trước đó.

Một số tài liệu nghiên cứu liên quan

2. Definition, Diagnosis and Classification of Diabetes Mellitus and its Complications - Report of a WHO Consultation

Sự thay đổi về phân loại:

- Đái tháo đường Type 1: Phá hủy tế bào beta.
- Đái tháo đường Type 2: Tình trạng đề kháng insulin hoặc thiếu hụt insulin tương đối.
- Đái tháo đường thai kỳ.
- Các loại đái tháo đường khác: Gồm các nhóm bệnh tiểu đường do nguyên nhân đặc biệt như bệnh tụy ngoại tiết, do thuốc, do gen, v.v.

Nhấn mạnh vào cơ chế bệnh sinh thay vì biểu hiện lâm sàng ban đầu.

Một số tài liệu nghiên cứu liên quan

2. Definition, Diagnosis and Classification of Diabetes Mellitus and its Complications - Report of a WHO Consultation

Sự thay đổi về tiêu chí:

- Hạ thấp ngưỡng FPG: 126 mg/dL (7.0 mmol/L).
- Lý do: Các nghiên cứu cho thấy nguy cơ biến chứng đã xuất hiện ở những mức đường huyết thấp hơn. Việc hạ ngưỡng giúp phát hiện bệnh sớm hơn.
- Giới thiệu thuật ngữ mới: Bổ sung danh mục "Rối loạn đường huyết lúc đói" (IFG).

Một số tài liệu nghiên cứu liên quan

3. Using the ADAP Learning Algorithm to Forecast the Onset of Diabetes Mellitus

ADAP là viết tắt của Adaptive Perceptron, một thuật toán học máy ban đầu thuộc họ mạng nơ-ron nhân tạo.

Cơ chế hoạt động: Thuật toán này "học" từ dữ liệu bằng cách điều chỉnh các trọng số để tìm ra các mô hình (patterns) và đưa ra dự đoán. Nó có khả năng xử lý nhiều biến số cùng lúc để phân loại các trường hợp.

Một số tài liệu nghiên cứu liên quan

3. Using the ADAP Learning Algorithm to Forecast the Onset of Diabetes Mellitus

- Bộ dữ liệu: Nghiên cứu sử dụng Bộ dữ liệu tiểu đường của người da đỏ Pima (Pima Indians Diabetes Database), một tập hợp dữ liệu đặc trưng về một nhóm dân số có nguy cơ mắc bệnh tiểu đường cao.

- Mục tiêu của thuật toán: Dựa trên các biến số này, thuật toán sẽ dự đoán "Outcome" – liệu bệnh nhân có mắc bệnh tiểu đường hay không.

1. MÔ TẢ BÀI TOÁN

Giới thiệu bộ dữ liệu

- Tên: Pima Indians Diabetes Database
- Nguồn: Viện Quốc gia về Tiểu đường, Tiêu hoá và Thận (NIDDK, Mỹ)
- Đối tượng: 768 phụ nữ Pima (Arizona, USA), tuổi ≥ 21

Cấu trúc dữ liệu

- Số mẫu (instances): 768
- Số đặc trưng (features): 8 + 1 biến mục tiêu
- Biến mục tiêu (Outcome):
  - 0 = Không mắc tiểu đường (500 mẫu)
  - 1 = Mắc tiểu đường (268 mẫu)

1. MÔ TẢ BÀI TOÁN

Input

1. Pregnancies – số lần mang thai
2. Glucose – nồng độ glucose sau 2h OGTT
3. BloodPressure – huyết áp tâm trương (mmHg)
4. SkinThickness – độ dày nếp gấp da (mm)
5. Insulin – insulin huyết thanh sau 2h (mu U/ml)
6. BMI – chỉ số khối cơ thể
7. DiabetesPedigreeFunction – hàm phả hệ tiểu đường (yếu tố di truyền)
8. Age – tuổi

Output

Biến mục tiêu ‘Outcome’: 0 = Không mắc tiểu đường / 1 = Mắc tiểu đường

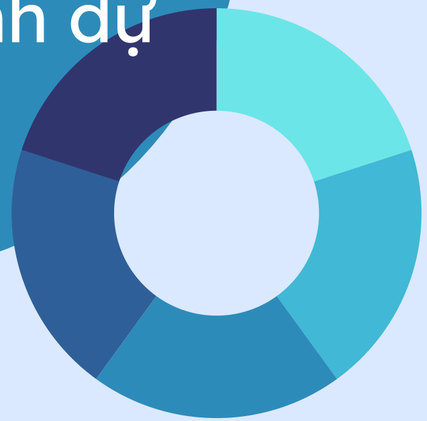
1. MÔ TẢ BÀI TOÁN

Mục tiêu

Xác định khả năng mắc bệnh tiểu đường type 2 dựa trên dữ liệu đầu vào.

Tìm ra các yếu tố quan trọng nhất ảnh hưởng đến nguy cơ mắc bệnh.

Chuẩn bị dữ liệu và nền tảng phân tích cho bước xây dựng mô hình dự đoán.



Xử lý dữ liệu bị thiếu | Xử lý dữ liệu trùng lặp

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

```
Pregnancies                 0
Glucose                     0
BloodPressure               0
SkinThickness               0
Insulin                     0
BMI                         0
DiabetesPedigreeFunction    0
Age                         0
Outcome                     0
dtype: int64
```

- **Nhận xét:** Không có giá trị nào Null

```
data.duplicated().sum()
```

```
0
```

- **Nhận xét:** Không có giá trị nào trùng lặp

2. XỬ LÝ DỮ LIỆU

```
(data==0).sum()
```

```
Pregnancies                 111
Glucose                       5
BloodPressure                35
SkinThickness               227
Insulin                     374
BMI                          11
DiabetesPedigreeFunction      0
Age                           0
Outcome                     500
dtype: int64
```

Kiểm tra dữ liệu có giá trị 0

```python
from sklearn.impute import SimpleImputer
# Thay thế giá trị 0 bằng NaN để xử lý
for column in ['Glucose', 'BloodPressure', 'SkinThickness', 'Insulin', 'BMI']:
    data[column] = data[column].replace(0, np.nan)

# Sử dụng SimpleImputer để điền giá trị thiếu bằng trung bình
imputer = SimpleImputer(strategy='mean')
data[['Glucose', 'BloodPressure', 'SkinThickness', 'Insulin', 'BMI']] = imputer.fit_transform(data[['Glucose', 'BloodP
```

Xử lý dữ liệu có giá trị 0

2. XỬ LÝ DỮ LIỆU

Mô tả thống kê dữ liệu

Phạm vi giá trị của các cột:

	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	121.686763	72.405184	29.153420	155.548223	32.457464	0.471876	33.240885	0.348958
std	3.369578	30.435949	12.096346	8.790942	85.021108	6.875151	0.331329	11.760232	0.476951
min	0.000000	44.000000	24.000000	7.000000	14.000000	18.200000	0.078000	21.000000	0.000000
25%	1.000000	99.750000	64.000000	25.000000	121.500000	27.500000	0.243750	24.000000	0.000000
50%	3.000000	117.000000	72.202592	29.153420	155.548223	32.400000	0.372500	29.000000	0.000000
75%	6.000000	140.250000	80.000000	32.000000	155.548223	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

Chuẩn hóa dữ liệu

MinMaxScaler là một công cụ trong thư viện Scikit-learn giúp đưa các đặc trưng về cùng 1 thang đo.

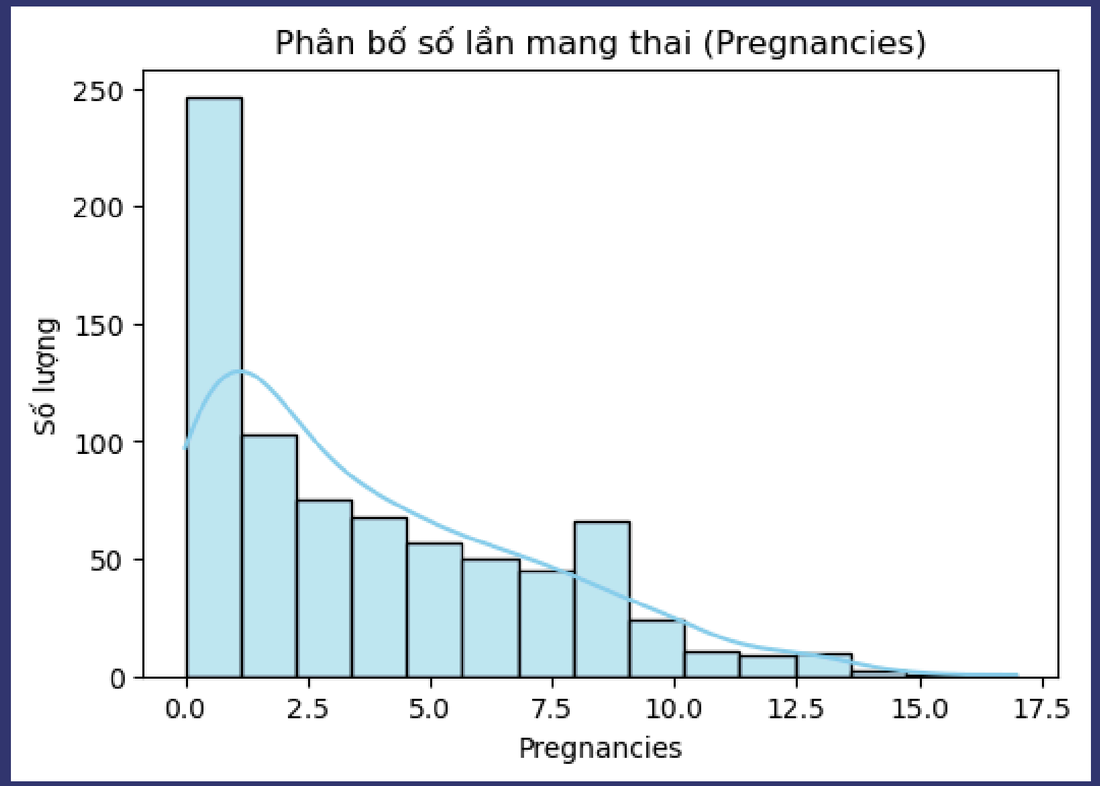
$$x' = \frac{x - x_{min}}{x_{max} - x_{min}}$$

5 dòng đầu của dữ liệu đã chuẩn hóa:

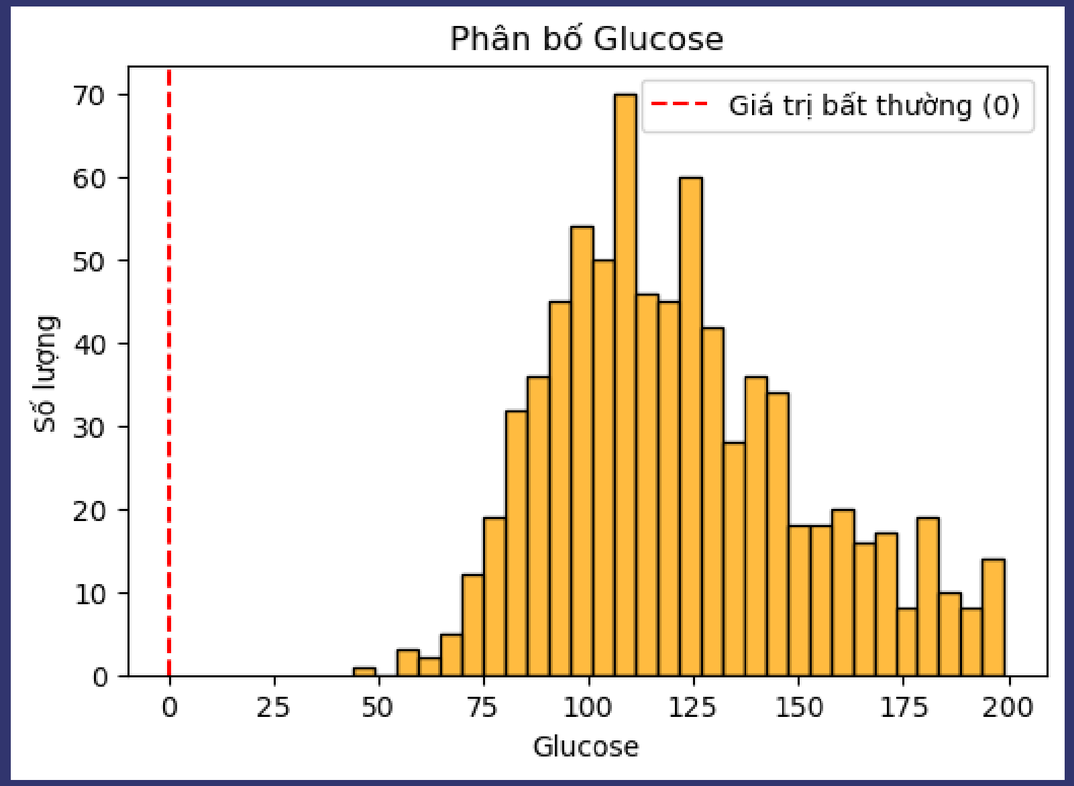
	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI
0	0.352941	0.670968	0.489796	0.304348	0.170130	0.314928
...						
1		0.116567	0.166667	0		
2		0.253629	0.183333	1		
3		0.038002	0.000000	0		
4		0.943638	0.200000	1		

3. PHÂN TÍCH ĐƠN BIẾN

Pregnancies

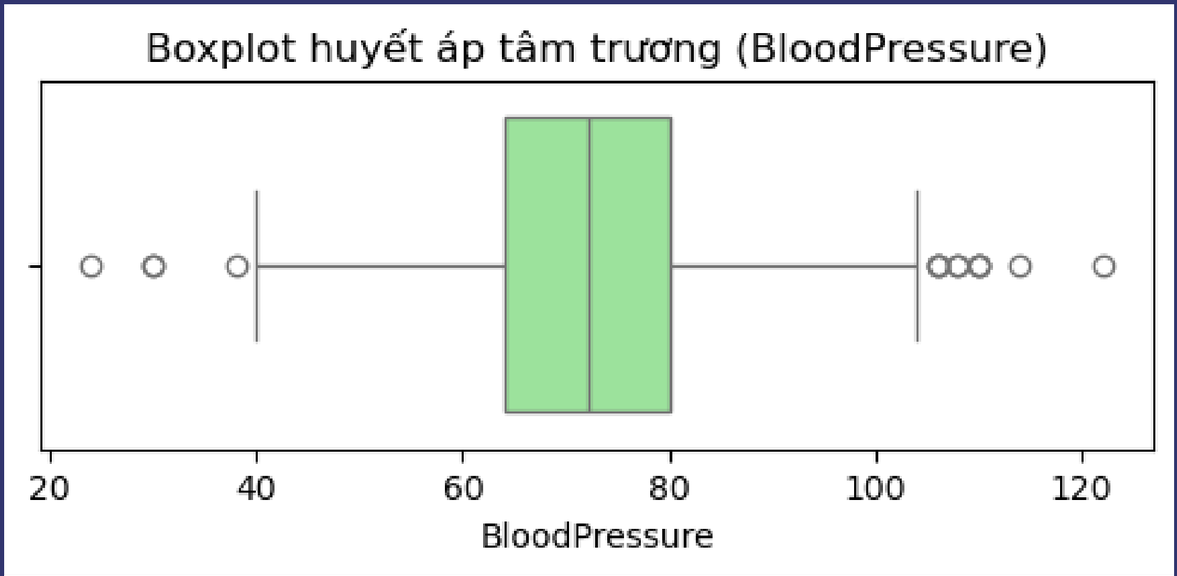


Glucose

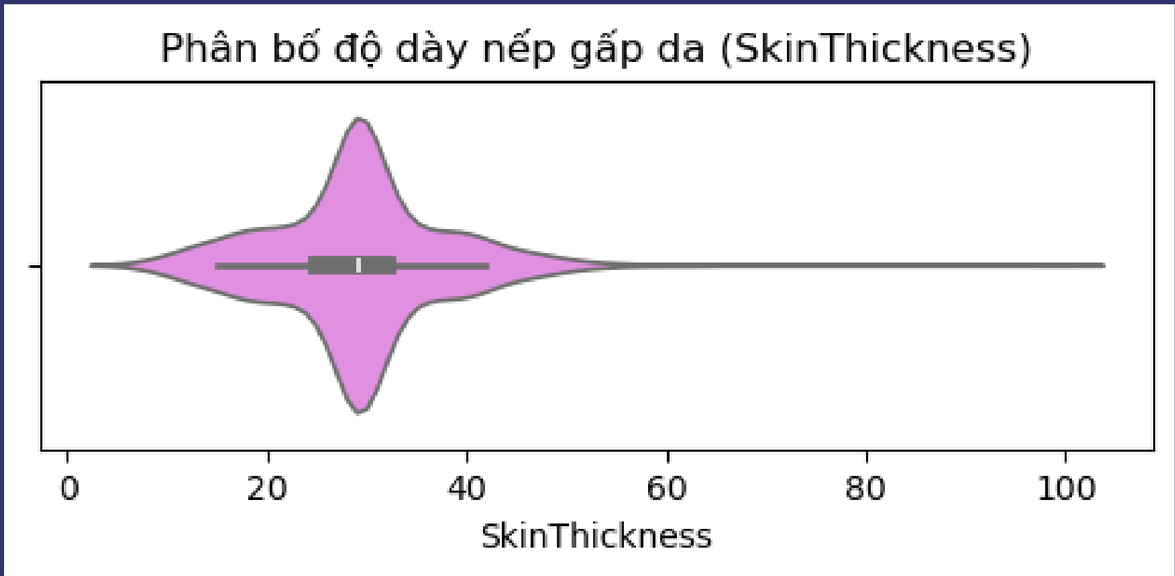


3. PHÂN TÍCH ĐƠN BIẾN

BloodPressure

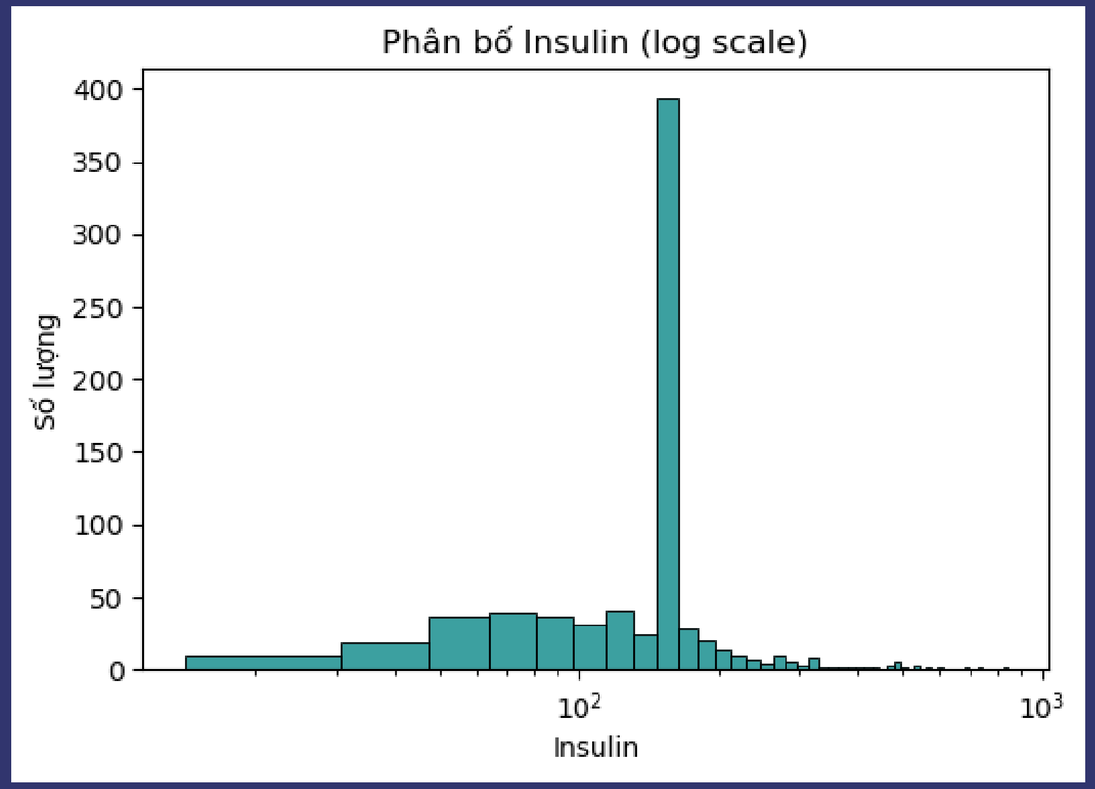


SkinThickness

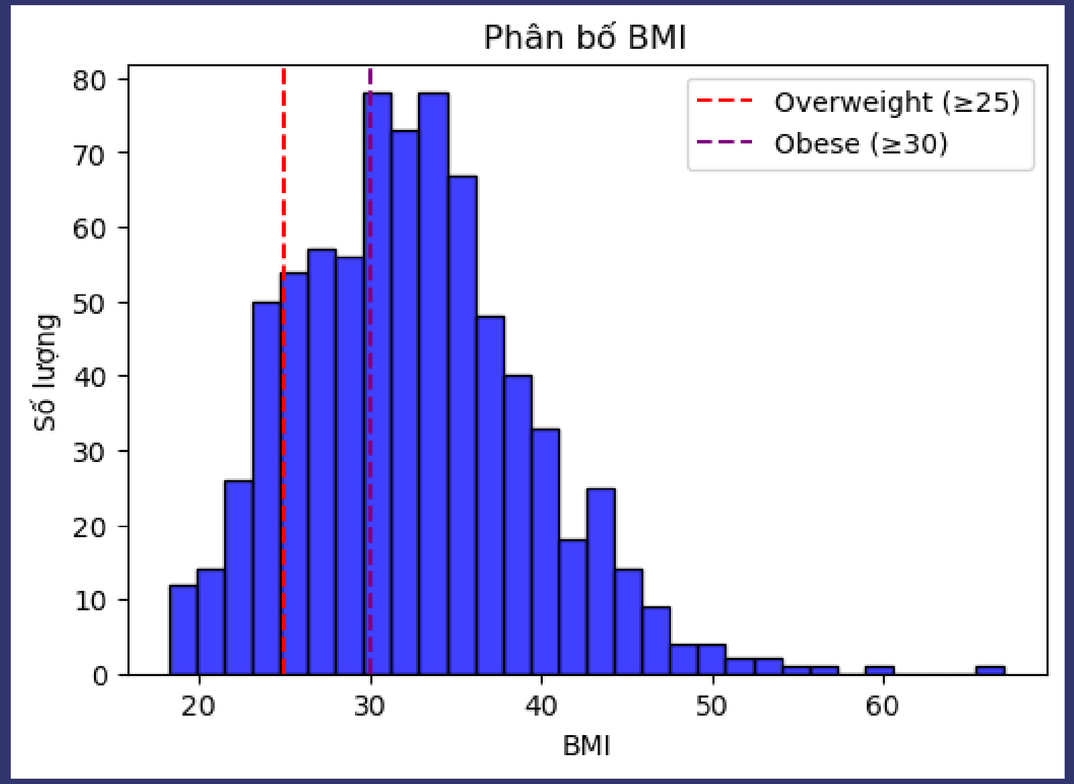


3. PHÂN TÍCH ĐƠN BIẾN

Insulin

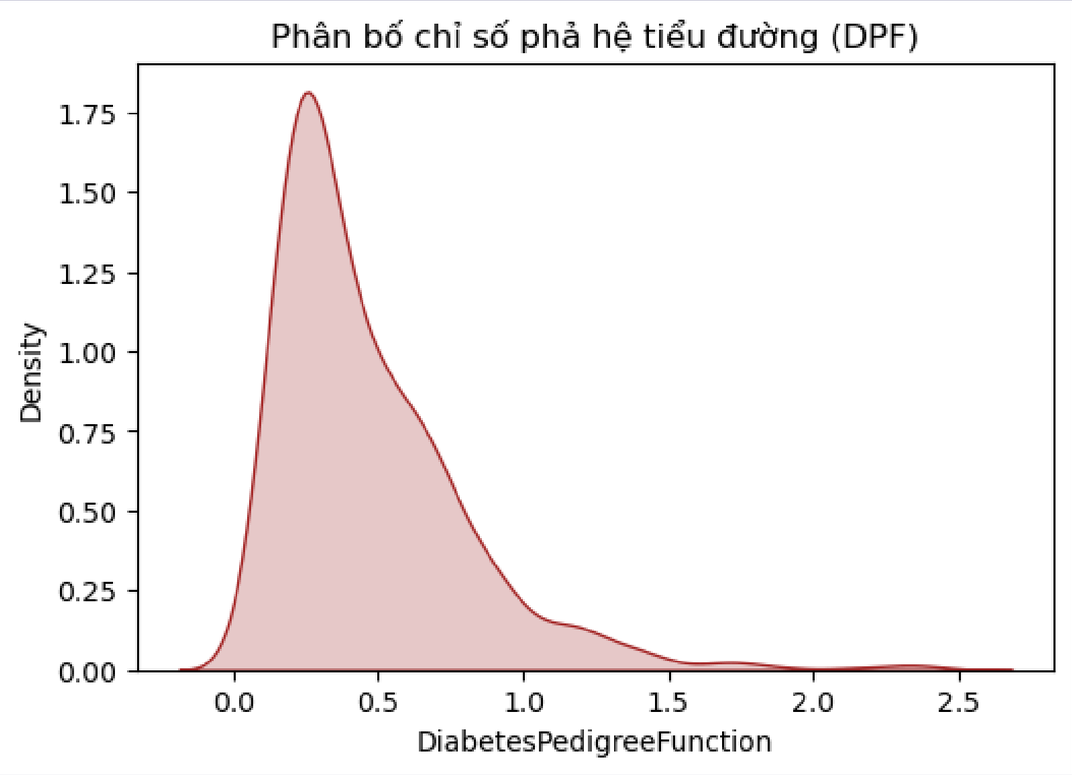


BMI

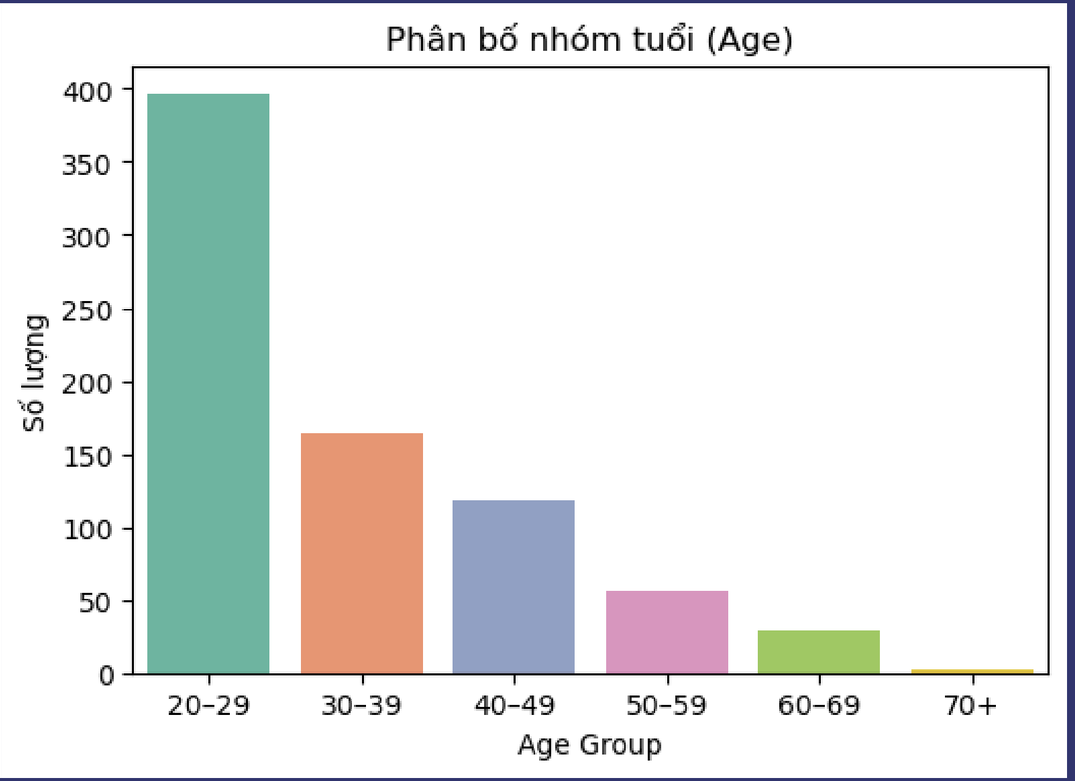


3. PHÂN TÍCH ĐƠN BIẾN

DiabetesPedigreeFunction

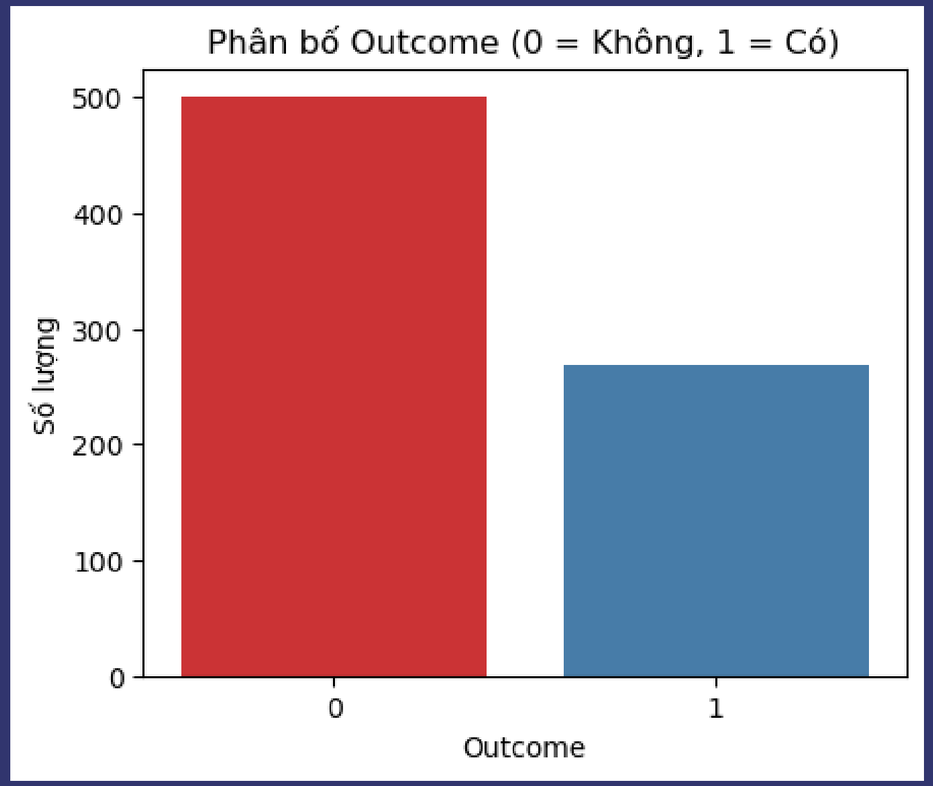


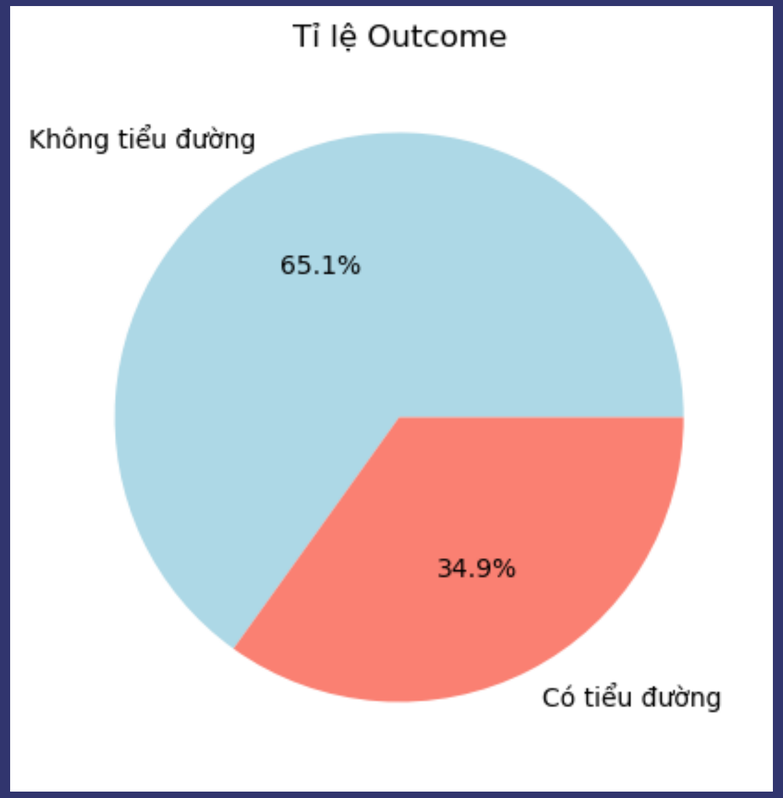
Age



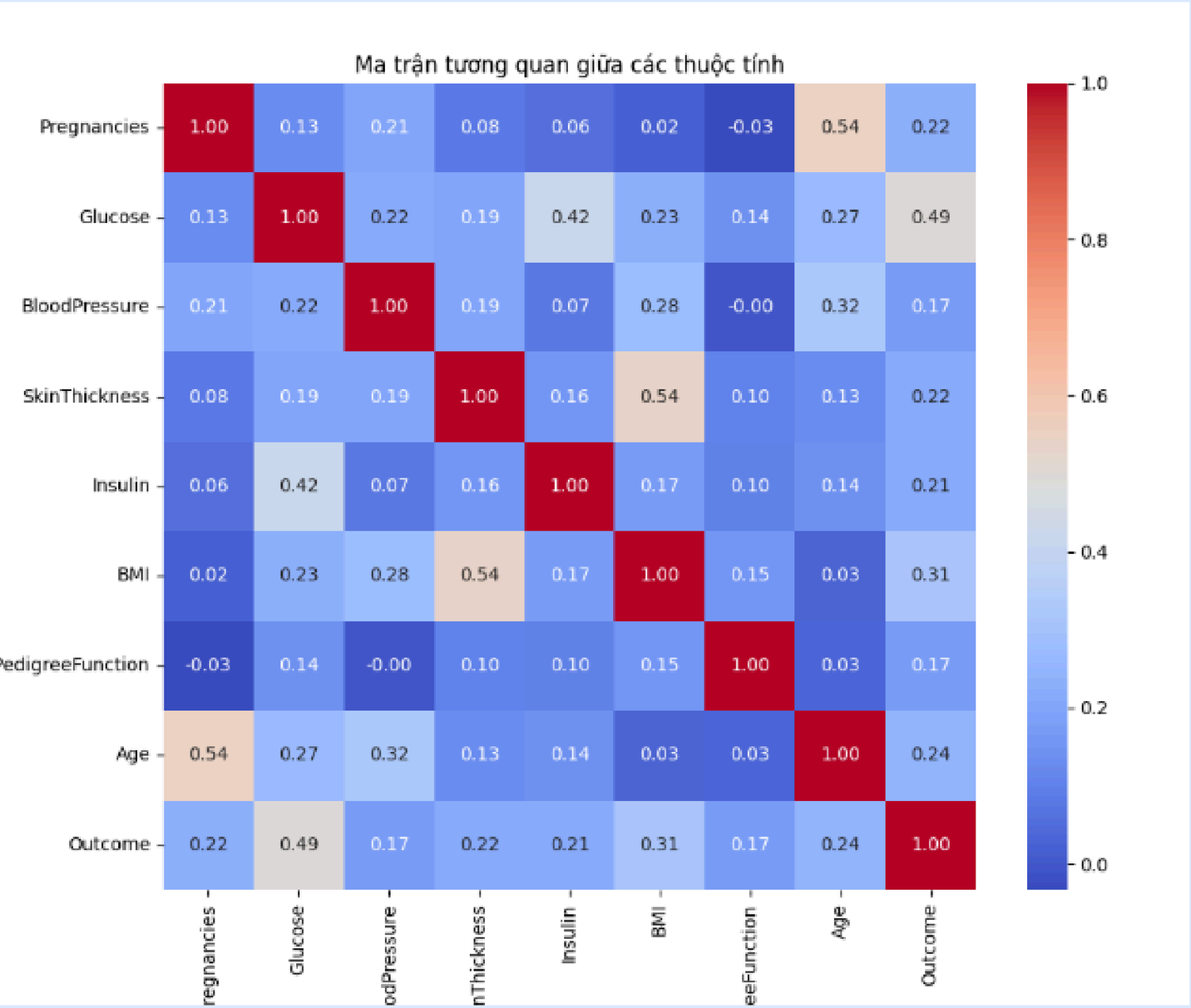
3. PHÂN TÍCH ĐƠN BIẾN

Outcome





4. PHÂN TÍCH ĐA BIẾN

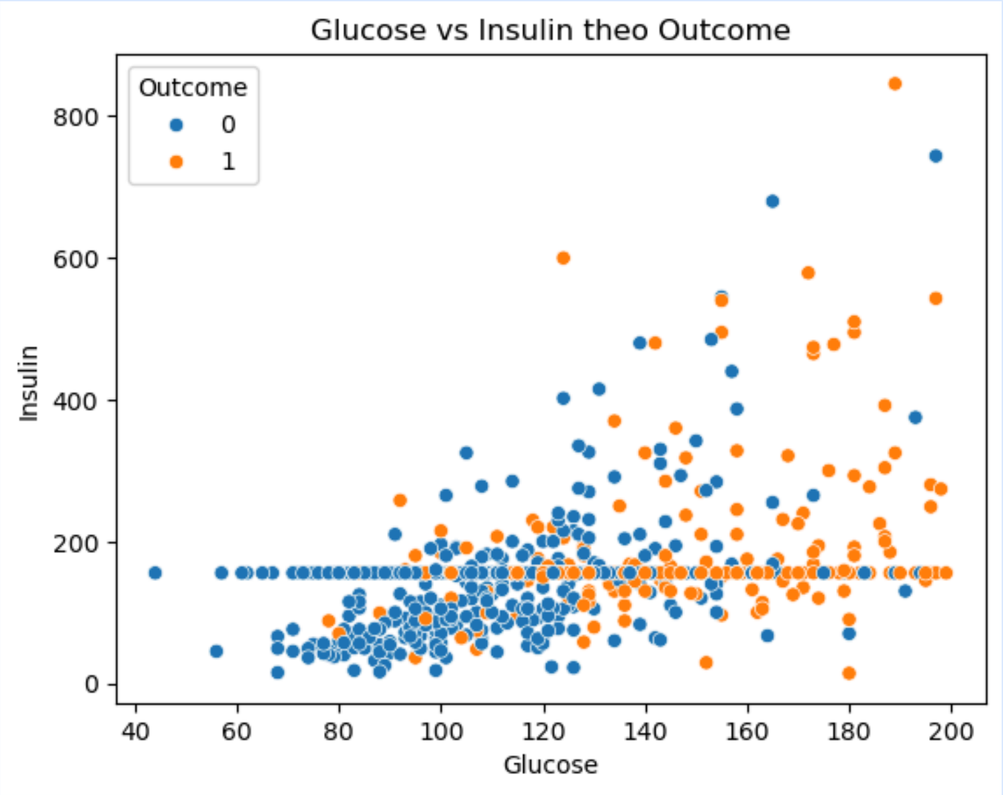


- Các biến tương quan với Outcome:
  - Glucose ↔ Outcome (0.49) → là yếu tố quyết định chính.
  - BMI (0.31), Age (0.24), Pregnancies (0.22) ↔ Outcome.

- Một số cặp biến có tương quan cao:
  - - Age ↔ Pregnancies (0.54)
  - - BMI ↔ SkinThickness (0.54)
  - - Glucose ↔ Insulin (0.42)

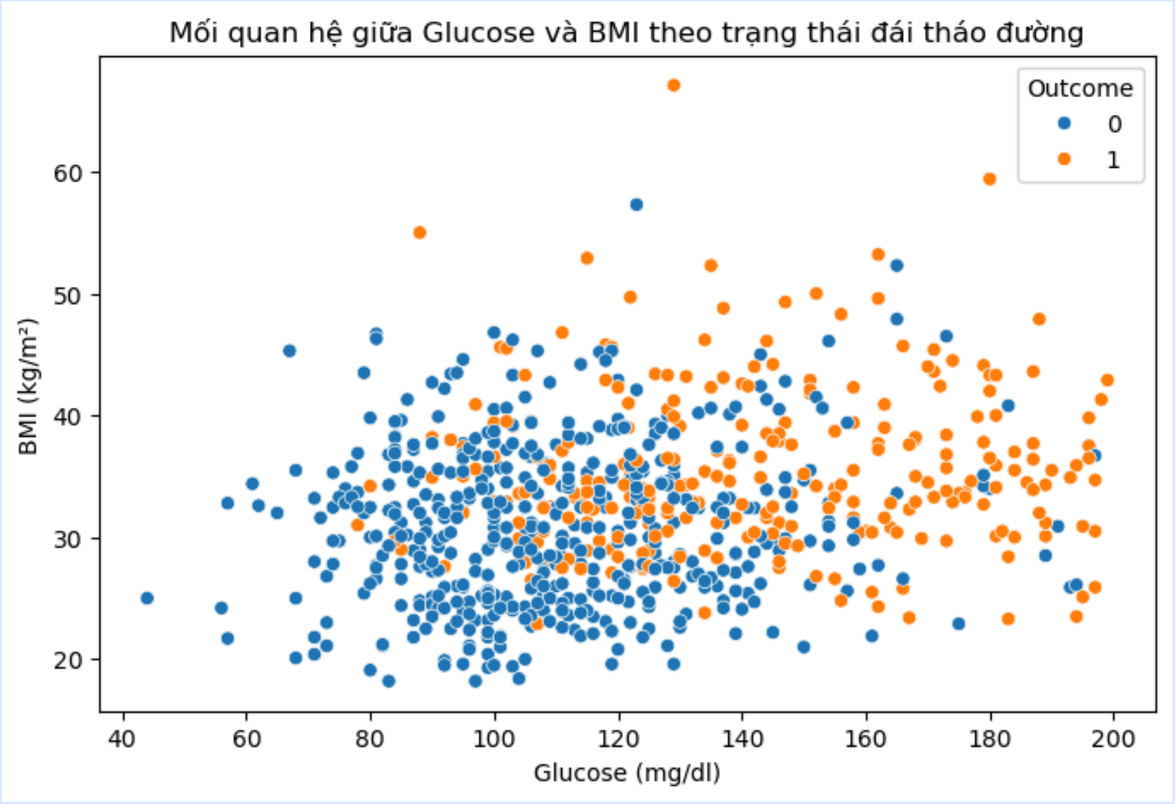
4. PHÂN TÍCH ĐA BIẾN

Glucose ↔ Insulin (0.42)



Glucose cao thường đi kèm Insulin cao.

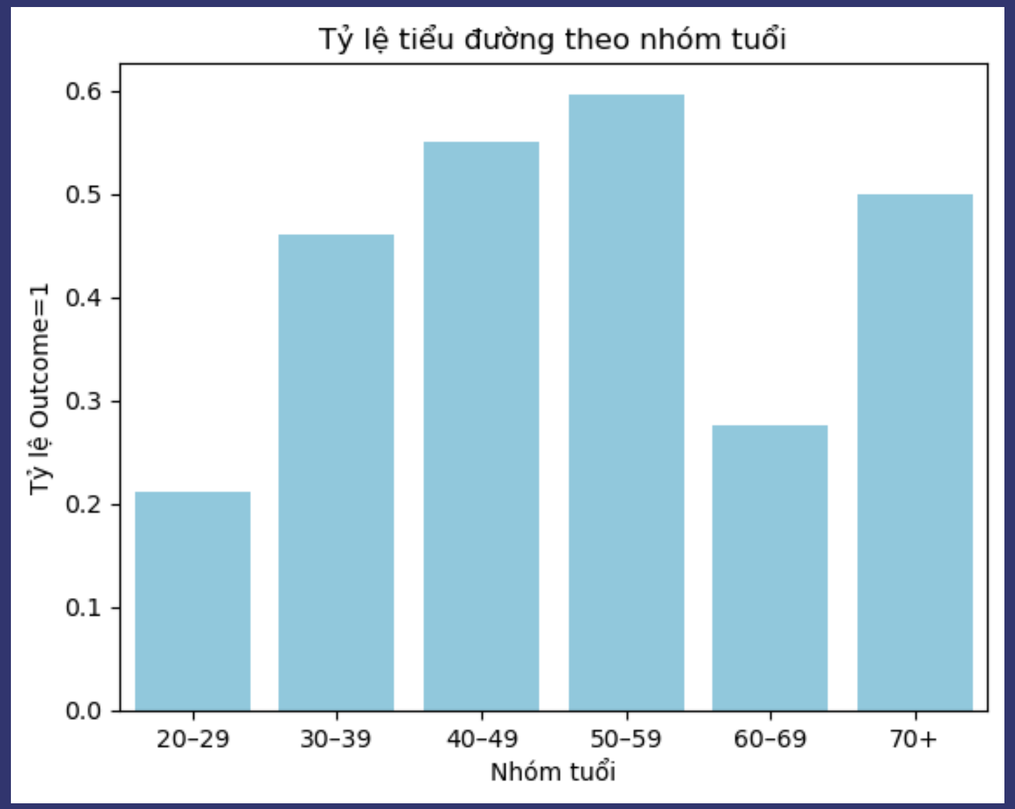
Glucose ↔ BMI (0.23)



Người có BMI cao thường có Glucose cao hơn.

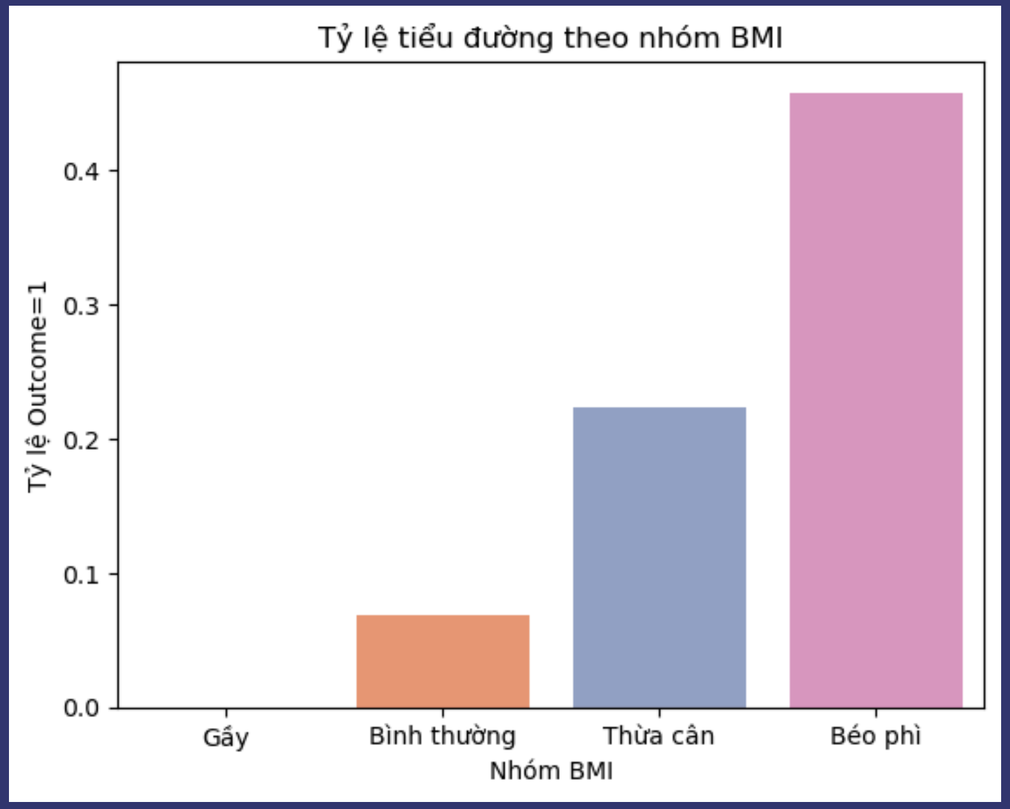
5. MỘT SỐ PHÁT HIỆN CHÍNH & KẾT LUẬN

Theo nhóm tuổi



Tỷ lệ mắc tiểu đường tăng dần theo tuổi, nhất là nhóm 40+.

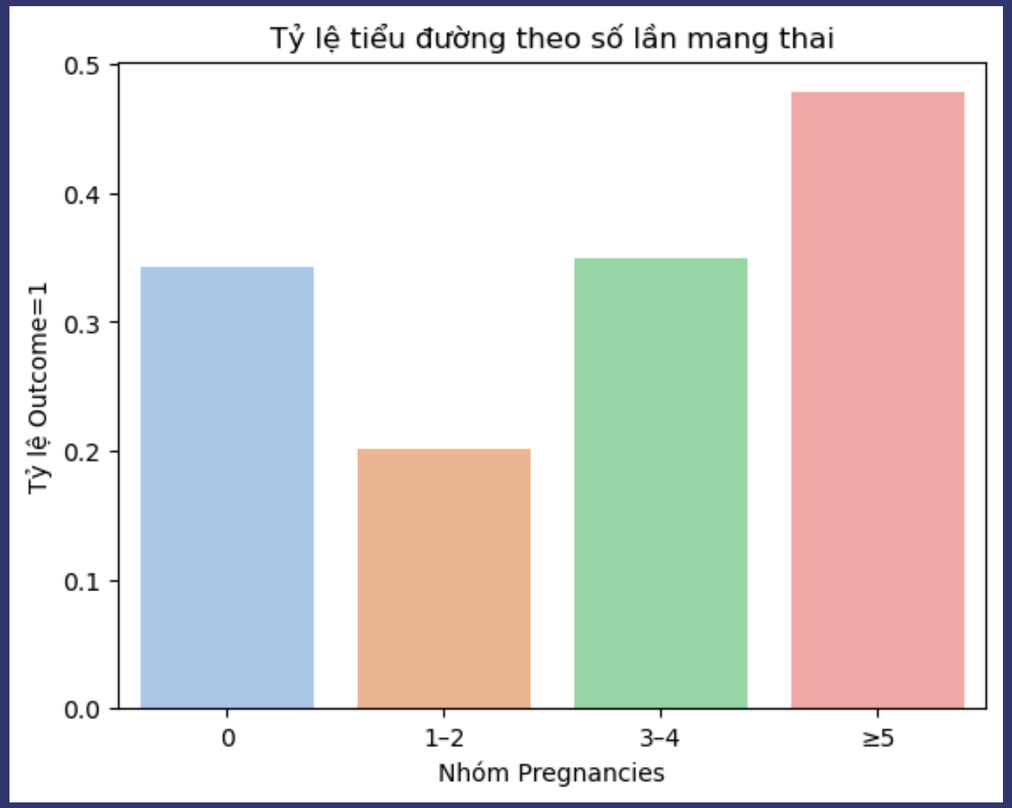
Theo nhóm BMI



Người thừa cân/béo phì có tỷ lệ tiểu đường cao hơn rõ rệt

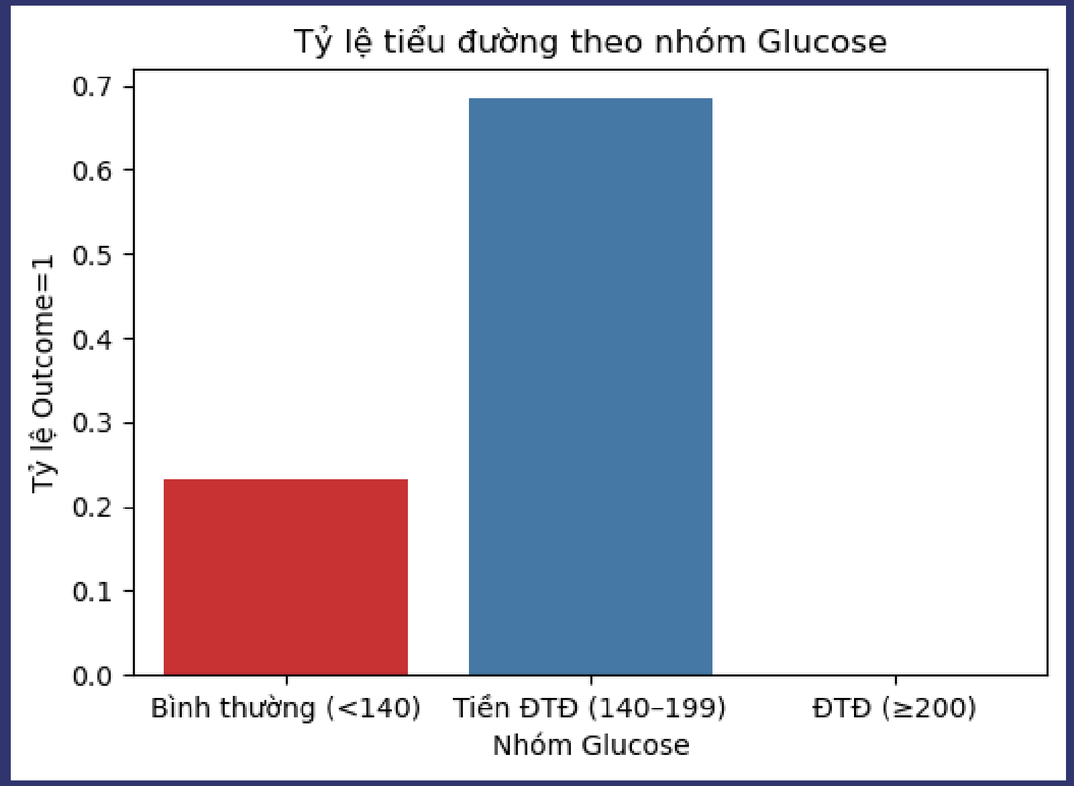
5. MỘT SỐ PHÁT HIỆN CHÍNH & KẾT LUẬN

Theo số lần mang thai



Nhóm có ≥5 lần mang thai có nguy cơ cao hơn.

Theo nhóm Glucoso



Người thừa cân/béo phì có tỷ lệ tiểu đường cao hơn rõ rệt

KẾT LUẬN

- Glucose: chỉ số y tế quyết định, phân nhóm Glucose dự đoán gần đúng tình trạng tiểu đường.
- BMI, Age, Pregnancies: các yếu tố nguy cơ bổ sung, làm tăng khả năng mắc bệnh.

Người trên 40 tuổi, BMI ≥ 30, Glucose ≥ 200 và có nhiều lần mang thai nằm trong nhóm nguy cơ rất cao → cần được quan tâm đặc biệt.