



Phân tích khám phá về bệnh đái tháo đường (Diabetes Mellitus)



Thành viên nhóm

DANH SÁCH

- Bùi Đức Chiến - 3122410039
- Trần Khải An - 3122410005
- Từ Nhật Anh - 3122410012
- Trần Thị Kiều Diễm - 3122410049



I. Giới thiệu đề tài

- Bệnh tiểu đường (Diabetes Mellitus) là bệnh mạn tính nguy hiểm, ảnh hưởng >537 triệu người trên thế giới (2021, WHO).
- Type 2 Diabetes chiếm đa số, gắn với lối sống, béo phì, di truyền.
- VIỆC Chẩn đoán sớm giúp giảm biến chứng, chi phí y tế, nâng cao chất lượng sống.
- Pima Indians: cộng đồng có tỷ lệ mắc tiểu đường rất cao → dataset lý tưởng cho Việc nghiên cứu.

Mục tiêu báo cáo

- Phân tích dữ liệu khám phá (EDA).
- Xác định đặc trưng quan trọng liên quan đến nguy cơ tiểu đường.
- Đề xuất hướng phát triển mô hình dự đoán.

II. CƠ SỞ LÝ THUYẾT

II. 1. GIỚI THIỆU VỀ BỆNH TIỂU ĐƯỜNG

Type 1:

- Bệnh tự miễn, hệ miễn dịch phá hủy tế bào beta tuyến tụy.
- Cơ thể gần như không còn insulin, người bệnh cần tiêm insulin suốt đời.
- Thường gặp ở trẻ em, thanh thiếu niên.

Type 2:

- Phổ biến nhất (90–95% trường hợp).
- Cơ thể có insulin nhưng tế bào không đáp ứng (kháng insulin).
- Nguy cơ tăng do: béo phì, lối sống ít vận động, tuổi tác, di truyền.
- Diễn biến âm thầm, thường chỉ phát hiện khi có biến chứng.

Tiểu đường thai kỳ (GDM):

- Xuất hiện ở phụ nữ mang thai (thường quý 2–3).
- Thường biến mất sau sinh nhưng tăng nguy cơ mắc type 2 cho cả mẹ và con.

→ Trong nghiên cứu này, chúng ta tập trung vào tiểu đường type 2 vì có mối liên hệ chặt chẽ với lối sống, tuổi tác và di truyền.

II. 2. CỘNG ĐỒNG PIMA INDIANS

Đặc điểm cộng đồng

- Người bản địa ở Arizona (Mỹ) và Mexico
- Tỷ lệ mắc tiểu đường type 2 cao nhất thế giới
- Có giai đoạn > 50% người trưởng thành mắc bệnh
- Lối sống thay đổi:
- Truyền thống: săn bắt, nông nghiệp
- Hiện đại: ăn tinh bột tinh chế, ít vận động
- Kết hợp di truyền + môi trường → tăng nguy cơ

Ý nghĩa nghiên cứu

- Y học & dịch tễ học: Hiểu cơ chế phát sinh tiểu đường type 2
- Khoa học dữ liệu: Bộ dữ liệu chuẩn cho học máy (classification)
- Ứng dụng thực tiễn: Mô hình dự đoán, hỗ trợ sàng lọc & chẩn đoán sớm

II. 3. Bộ dữ liệu Pima Indians Diabetes

Mô tả :

- 768 quan sát – phụ nữ Pima ≥ 21 tuổi
- 8 đặc trưng đầu vào + 1 biến mục tiêu
- 65% không mắc bệnh, 35% mắc bệnh

Các biến đầu vào
(Predictor Variables):

1. Pregnancies
2. Glucose
3. BloodPressure
4. SkinThickness
5. Insulin
6. BMI
7. DiabetesPedigreeFunction (DPF)
8. Age

II. 3. Bộ dữ liệu Pima Indians Diabetes

Mô tả :

- 768 quan sát – phụ nữ Pima ≥ 21 tuổi
- 8 đặc trưng đầu vào + 1 biến mục tiêu
- 65% không mắc bệnh, 35% mắc bệnh

Ý nghĩa dữ liệu:

- Dataset kết hợp cả yếu tố sinh lý (Glucose, Insulin, BloodPressure, BMI), nhân khẩu học (Age, Pregnancies) và di truyền (DPF).
- Đây là một trong những dataset hiếm hoi có sự cân bằng giữa dữ liệu y tế và nhân khẩu học, giúp khai thác đa chiều trong phân tích dữ liệu và mô hình học máy.

II. 4. Ý nghĩa của các đặc trưng trong dữ liệu

- **Pregnancies**: số lần mang thai → phản ánh sức khỏe sinh sản & nguy cơ tiểu đường thai kỳ.
- **Glucose**: nồng độ glucose trong máu sau 2 giờ test dung nạp → chỉ số quan trọng nhất để chẩn đoán tiểu đường.
- **BloodPressure**: huyết áp tâm trương (mmHg) → cao huyết áp thường đi kèm với tiểu đường type 2.
- **SkinThickness**: độ dày nếp gấp da (mm) → gián tiếp phản ánh lượng mỡ dưới da.
- **Insulin**: nồng độ insulin huyết thanh sau 2 giờ (mu U/ml) → đánh giá chức năng tuyến tụy và kháng insulin.
- **BMI**: chỉ số khối cơ thể = cân nặng / (chiều cao)² → yếu tố nguy cơ quan trọng.
- **DiabetesPedigreeFunction (DPF)**: chỉ số về di truyền → đánh giá mức độ “tiền sử gia đình” liên quan đến tiểu đường.
- **Age**: tuổi bệnh nhân → nguy cơ tăng theo tuổi, đặc biệt sau 40.

→ TÓM LẠI: 8 BIẾN NÀY KẾT HỢP VỚI NHAU CUNG CẤP GÓC NHÌN NHÂN KHẨU HỌC - LỐI SỐNG - SINH LÝ - DI TRUYỀN, TẠO THÀNH BỘ DỮ LIỆU PHONG PHÚ ĐỂ PHÂN TÍCH VÀ HUẤN LUYỆN MÔ HÌNH DỰ ĐOÁN TIỂU ĐƯỜNG.

II.5. CÁC NGHIÊN CỨU LIÊN QUAN

1988 – Smith et al.

- Tiên phong ứng dụng học máy (ADAP, Logistic Regression, Perceptron) trên dữ liệu Pima Indians.
- Dự đoán nguy cơ mắc tiểu đường trong 5 năm (phân loại nhị phân).

1990s – Classification System

- Xây dựng hệ thống phân loại chi tiết: Normal, IGT, DM, GDM, PrevAGT, PotAGT.
- DM phân nhỏ: Type 1 (IDDM) / Type 2 (NIDDM, obese – non-obese).
- Ý nghĩa: chuẩn hóa chẩn đoán, theo dõi tiến triển bệnh, hỗ trợ decision support system.

1999 – WHO Report

- Chuẩn hóa tiêu chuẩn chẩn đoán tiểu đường toàn cầu.
- Giảm ngưỡng FPG từ 7.8 → 7.0 mmol/L.
- Phân loại: Type 1, Type 2, Gestational, Other.
- Bổ sung nhóm trung gian (IFG, IGT) & Hội chứng chuyển hóa.

Kết luận: Vừa cung cấp **cơ sở lý thuyết và y học**, vừa gợi mở hướng ứng dụng **khoa học dữ liệu và học máy** trong việc phân tích và dự đoán bệnh tiểu đường.

III. DỮ LIỆU VÀ PHƯƠNG PHÁP

III.1. DỮ LIỆU

III.2. VẤN ĐỀ DỮ LIỆU VÀ TIỀN
XỬ LÝ SƠ BỘ

III.3. PHƯƠNG PHÁP PHÂN
TÍCH DỮ LIỆU (EDA)

III.4. CÔNG CỤ VÀ MÔI
TRƯỜNG PHÂN TÍCH

III.1. DỮ LIỆU

Dữ liệu nghiên cứu

- Nguồn dữ liệu: Pima Indians Diabetes Dataset – từ Viện NIDDK (Mỹ)
- Mục tiêu: Dự đoán nguy cơ mắc **bệnh tiểu đường loại II**
- Đối tượng: Phụ nữ người Pima (Arizona, Mỹ), ≥ 21 tuổi
- Số mẫu: 768 mẫu quan sát
- Biến số: 9 biến (8 đặc trưng + 1 biến mục tiêu)

Phân bố nhãn mục tiêu (Outcome) cho thấy dữ liệu có sự mất cân bằng nhẹ:

Outcome	Số lượng	Tỷ lệ (%)
0 = Không mắc	500	65,1%
1 = Mắc	268	34,9%,

III.2. VẤN ĐỀ DỮ LIỆU VÀ TIỀN XỬ LÝ SƠ BỘ

Mặc dù bộ dữ liệu Pima Indians Diabetes có chất lượng tốt, nhưng quá trình khám phá dữ liệu cho thấy một số **bất thường cần xử lý trước khi phân tích sâu**.

Vấn đề phát hiện:

- Một số biến có giá trị 0 không hợp lý về mặt sinh lý → phản ánh dữ liệu thiếu (missing values)

Biến bị ảnh hưởng:

- Glucose, BloodPressure, BMI → Có một số giá trị 0
- Insulin, SkinThickness → Tỷ lệ 0 rất cao

Ý nghĩa:

- Các biến này đều quan trọng về y học, cần xử lý cẩn thận để đảm bảo độ tin cậy khi phân tích/mô hình hóa.

III.3. PHƯƠNG PHÁP PHÂN TÍCH DỮ LIỆU (EDA)

EDA giúp hiểu dữ liệu, phát hiện bất thường và rút ra nhận định ban đầu. Các bước thực hiện:

1. Thống kê mô tả

- Dùng: count, mean, std, min-max, các phân vị (25%, 50%, 75%)
- Mục tiêu: Hiểu phân phối & phát hiện giá trị bất thường

2. Kiểm tra dữ liệu thiếu

- Xác định các giá trị 0 không hợp lý (Glucose, BMI, v.v.)
- Gợi ý xử lý: điền giá trị trung vị (median) hoặc nội suy

3. Phân tích đơn biến

- Dùng histogram, boxplot
- Mục tiêu: Nhận biết skewness, outliers

4. Phân tích hai biến

- So sánh từng đặc trưng với Outcome
- Dùng: histogram theo nhóm, bar chart (ví dụ theo BMI, tuổi)
- Mục tiêu: Tìm đặc trưng ảnh hưởng đến bệnh tiểu đường

III.3. PHƯƠNG PHÁP PHÂN TÍCH DỮ LIỆU (EDA)

5. Phân tích đa biến

- Tính ma trận tương quan → heatmap, pairplot
- Mục tiêu: Nhận diện đặc trưng quan trọng cho mô hình

6. Phân tích ngoại lệ (Outliers)

- Dùng boxplot, phương pháp IQR
- Đánh giá mức độ ảnh hưởng đến phân tích

7. Tổng hợp kết quả

- Rút ra insights chính: biến quan trọng, biến ít giá trị, vấn đề cần xử lý thêm
- Là cơ sở cho bước xây dựng mô hình dự đoán

📌 Kết luận:

EDA là bước nền tảng, giúp hiểu rõ dữ liệu và định hướng chiến lược tiền xử lý & mô hình hóa hiệu quả.

III.4. CÔNG CỤ VÀ MÔI TRƯỜNG PHÂN TÍCH

Môi trường & Công cụ Phân tích

Môi trường phân tích: Google Colab

- Nền tảng chạy Python trực tuyến, miễn phí
- Không cần cài đặt, dễ chia sẻ, hỗ trợ GPU khi cần

Ngôn ngữ lập trình: Python 3.x

- Cú pháp đơn giản, phổ biến trong phân tích dữ liệu & học máy

Thư viện sử dụng

Thư viện	Mục đích chính
pandas	Quản lý dữ liệu bảng (DataFrame), thống kê
numpy	Tính toán số học, mảng và ma trận
matplotlib & seaborn	Vẽ biểu đồ: histogram, boxplot, heatmap, pairplot
scikit-learn (sklearn)	Tiền xử lý, chia dữ liệu, phân tích mối quan hệ

III.4. CÔNG CỤ VÀ MÔI TRƯỜNG PHÂN TÍCH

Thư viện	Chức năng
joblib	Lưu trữ đối tượng (model, pipeline, v.v.)
IPython.display	Hiển thị bảng và hình ảnh trực quan
warnings	Ẩn cảnh báo không cần thiết
os, sys	Quản lý đường dẫn & thao tác file

Quy trình phân tích dữ liệu

1. Đọc dữ liệu với pandas từ file CSV
2. Khám phá sơ bộ: kích thước, kiểu dữ liệu, dữ liệu bất thường
3. Thống kê mô tả & trực quan hóa bằng matplotlib, seaborn
4. Tiền xử lý: xử lý giá trị 0, chuẩn hóa dữ liệu
5. Phân tích mối quan hệ: dùng heatmap, histogram, violin plot, pairplot

IV. PHÂN TÍCH DỮ LIỆU (EDA)

IV.1. THỐNG KÊ MÔ TẢ

IV.2. KIỂM TRA DỮ LIỆU THIẾU VÀ BẤT HỢP LÝ

IV.3. PHÂN TÍCH PHÂN PHỐI CÁC BIỂN

IV.4. SO SÁNH NHÓM MẮC BỆNH VÀ KHÔNG MẮC BỆNH

IV.5. PHÂN TÍCH ĐA BIỂN VÀ TƯƠNG QUAN

IV.6. OUTLIERS (GIÁ TRỊ NGOẠI LAI)

IV.7. KẾT QUẢ TỔNG HỢP TỪ EDA

IV.8. XỬ LÝ DỮ LIỆU THIẾU VÀ BẤT HỢP LÝ

IV.1. THỐNG KÊ MÔ TẢ

IV.1.1. KÍCH THƯỚC VÀ KIỂU DỮ LIỆU

IV.1.2. 5 DÒNG ĐẦU VÀ 5 DÒNG CUỐI

IV.1.3. THỐNG KÊ CƠ BẢN (SUMMARY STATISTICS)

IV.1.4. PHÂN BỐ BIẾN MỤC TIÊU (OUTCOME)

IV.1.1. KÍCH THƯỚC VÀ KIỂU DỮ LIỆU

- Dataset gồm 768 dòng và 9 cột (8 biến đầu vào và 1 biến mục tiêu).
- Các biến đều là numeric (kiểu số nguyên hoặc số thực), thuận tiện cho việc phân tích thống kê và trực quan hóa.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 768 entries, 0 to 767
Data columns (total 9 columns):
 #   Column           Non-Null Count  Dtype  
--- 
 0   Pregnancies      768 non-null    int64  
 1   Glucose          768 non-null    int64  
 2   BloodPressure    768 non-null    int64  
 3   SkinThickness    768 non-null    int64  
 4   Insulin          768 non-null    int64  
 5   BMI              768 non-null    float64 
 6   DPF              768 non-null    float64 
 7   Age              768 non-null    int64  
 8   Outcome          768 non-null    int64  
dtypes: float64(2), int64(7)
memory usage: 54.1 KB
```

IV.1.2. 5 DÒNG ĐẦU VÀ 5 DÒNG CUỐI

- 5 dòng đầu tiên và 5 dòng cuối cùng:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome	
0	6	148	72	35	0	33.6		0.627	50	1
1	1	85	66	29	0	26.6		0.351	31	0
2	8	183	64	0	0	23.3		0.672	32	1
3	1	89	66	23	94	28.1		0.167	21	0
4	0	137	40	35	168	43.1		2.288	33	1
	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome	
763	10	101	76	48	180	32.9		0.171	63	0
764	2	122	70	27	0	36.8		0.340	27	0
765	5	121	72	23	112	26.2		0.245	30	0
766	1	126	60	0	0	30.1		0.349	47	1
767	1	93	70	31	0	30.4		0.315	23	0

IV.1.3. THỐNG KÊ CƠ BẢN (SUMMARY STATISTICS)

Thống kê mô tả dữ liệu với `df.describe()`

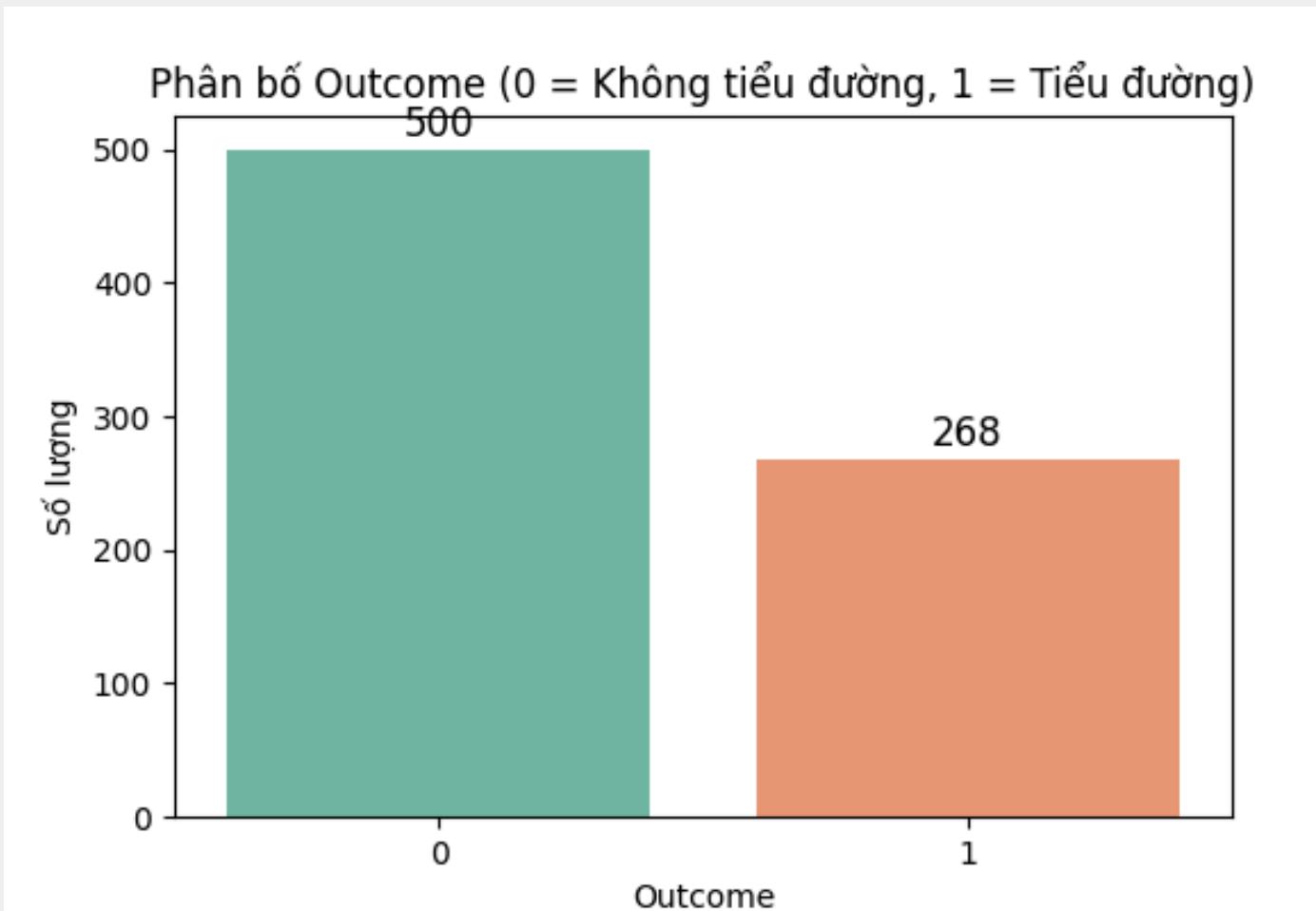
	count	mean	std	min	25%	50%	75%	max
Pregnancies	768.0	3.845052	3.369578	0.000	1.00000	3.0000	6.00000	17.00
Glucose	768.0	120.894531	31.972618	0.000	99.00000	117.0000	140.25000	199.00
BloodPressure	768.0	69.105469	19.355807	0.000	62.00000	72.0000	80.00000	122.00
SkinThickness	768.0	20.536458	15.952218	0.000	0.00000	23.0000	32.00000	99.00
Insulin	768.0	79.799479	115.244002	0.000	0.00000	30.5000	127.25000	846.00
BMI	768.0	31.992578	7.884160	0.000	27.30000	32.0000	36.60000	67.10
DiabetesPedigreeFunction	768.0	0.471876	0.331329	0.078	0.24375	0.3725	0.62625	2.42
Age	768.0	33.240885	11.760232	21.000	24.00000	29.0000	41.00000	81.00
Outcome	768.0	0.348958	0.476951	0.000	0.00000	0.0000	1.00000	1.00

Nhận xét:

- Các biến Glucose, BloodPressure, SkinThickness, Insulin, BMI có giá trị 0 bất hợp lý → cần xử lý.
- Có outliers đặc biệt ở Insulin (846) → cần kiểm tra bằng boxplot/IQR.
- Các biến Glucose, BMI, Age có nhiều ý nghĩa y khoa → nên tập trung khi phân tích.

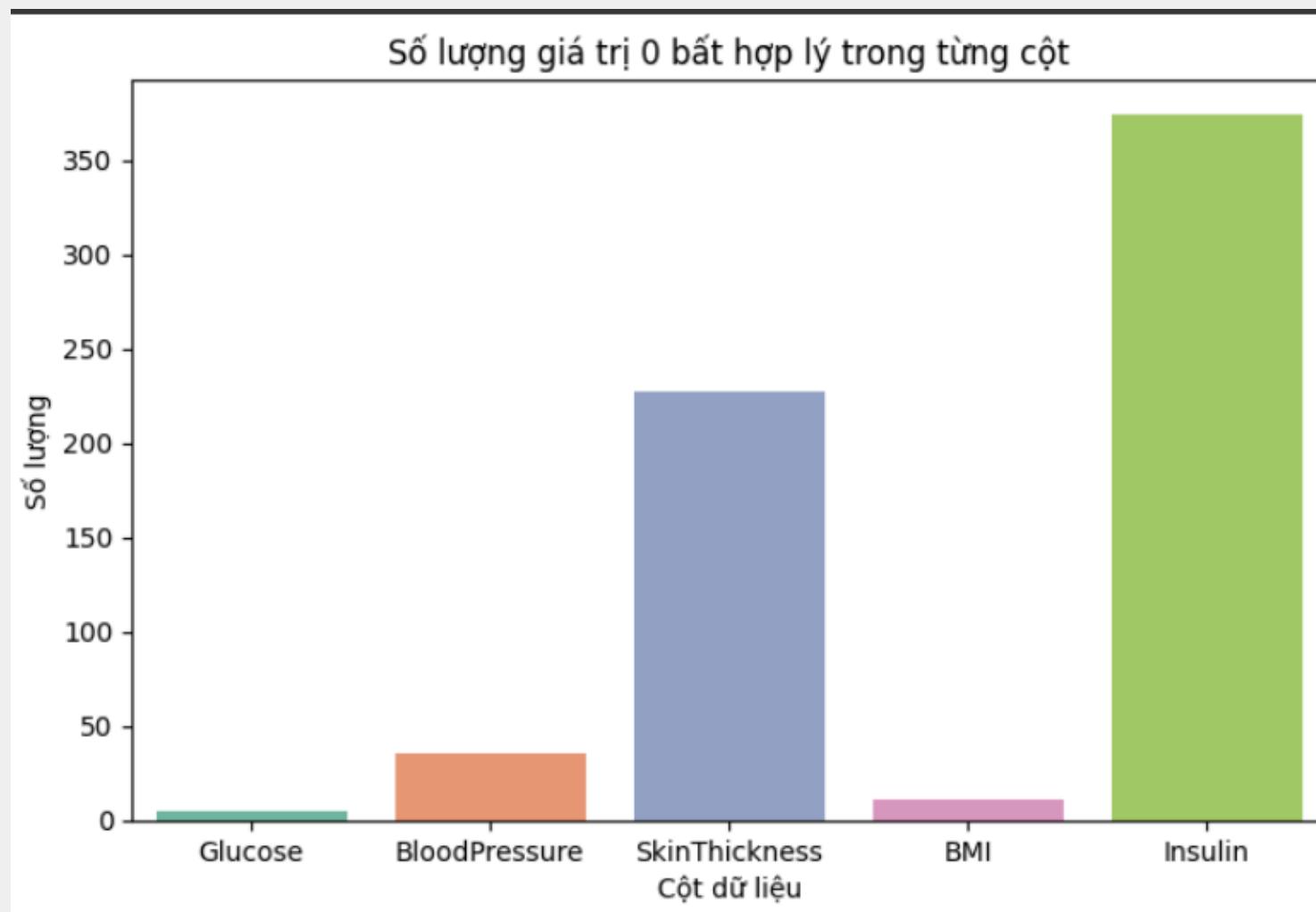
IV.1.4. PHÂN BỐ BIẾN MỤC TIÊU (OUTCOME)

- Outcome = 0: 500 trường hợp (65.1%).
- Outcome = 1: 268 trường hợp (34.9%).
- Dataset có hiện tượng mất cân bằng lớp nhẹ (class imbalance).



IV.2. KIỂM TRA DỮ LIỆU THIẾU VÀ BẤT HỢP LÝ

- Không có giá trị NaN hay dòng trùng lặp → dữ liệu được lưu trữ khá tốt về mặt kỹ thuật.
- Tuy nhiên, phát hiện nhiều giá trị bằng 0 không hợp lý về mặt sinh lý, thực



BẢNG THỐNG KÊ SỐ LƯỢNG GIÁ TRỊ 0 BẤT HỢP LÝ TRONG TỪNG BIẾN

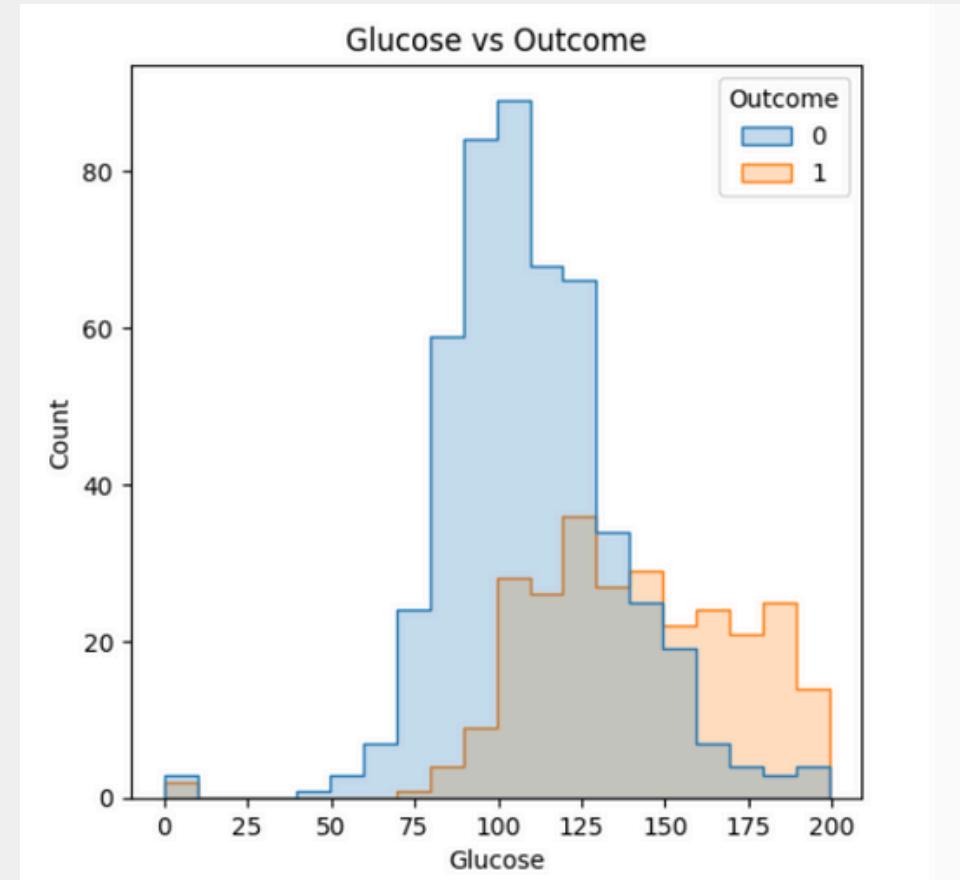
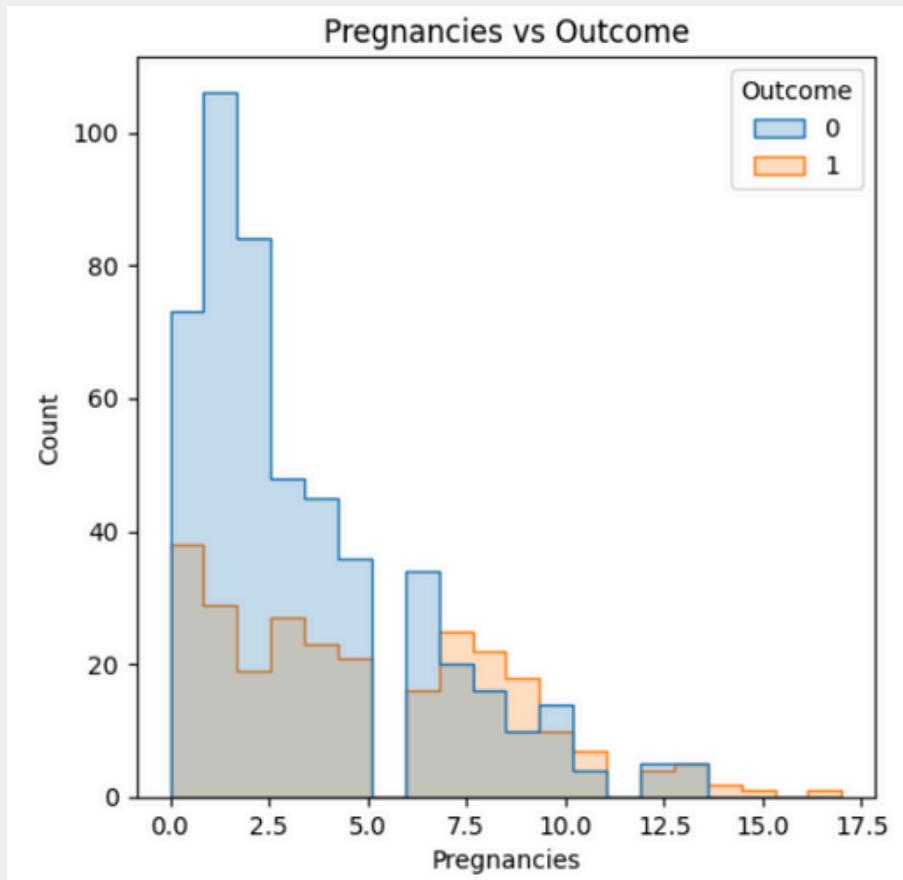
Biến	Số lượng giá trị 0	Tỷ lệ (%)	Nhận xét
Glucose	5	0,7%	Rất hiếm, coi như thiếu
BloodPressure	35	4,6%	Một số trường hợp bất hợp lý
SkinThickness	227	29,6%	Thiếu dữ liệu nhiều
Insulin	374	48,7%	Gần một nửa dữ liệu bị thiếu
BMI	11	1,4%	Một số giá trị bất hợp lý

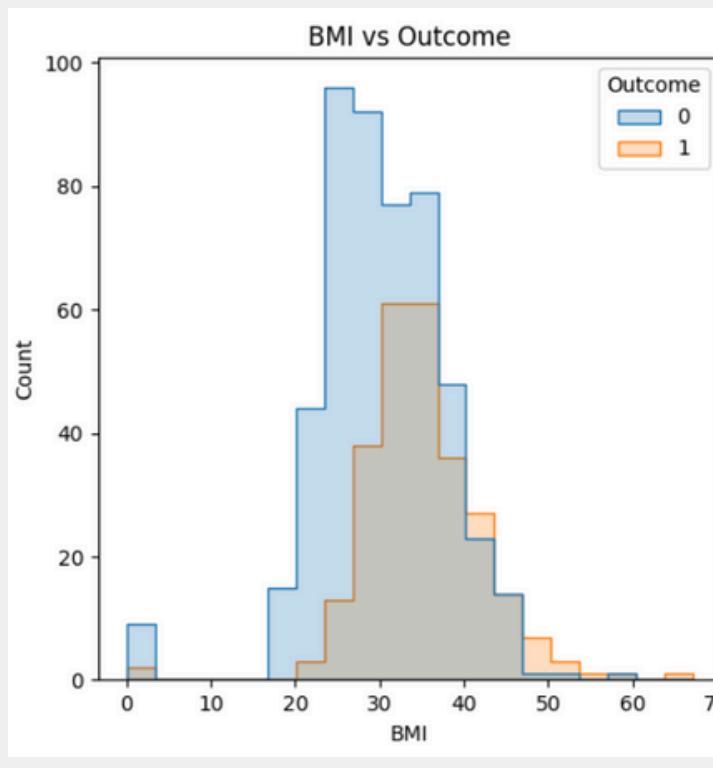
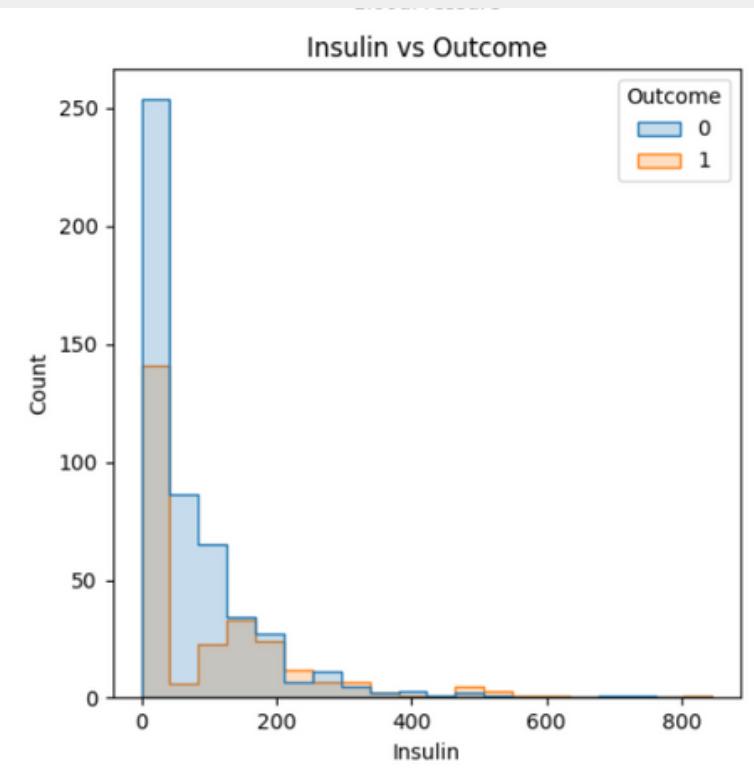
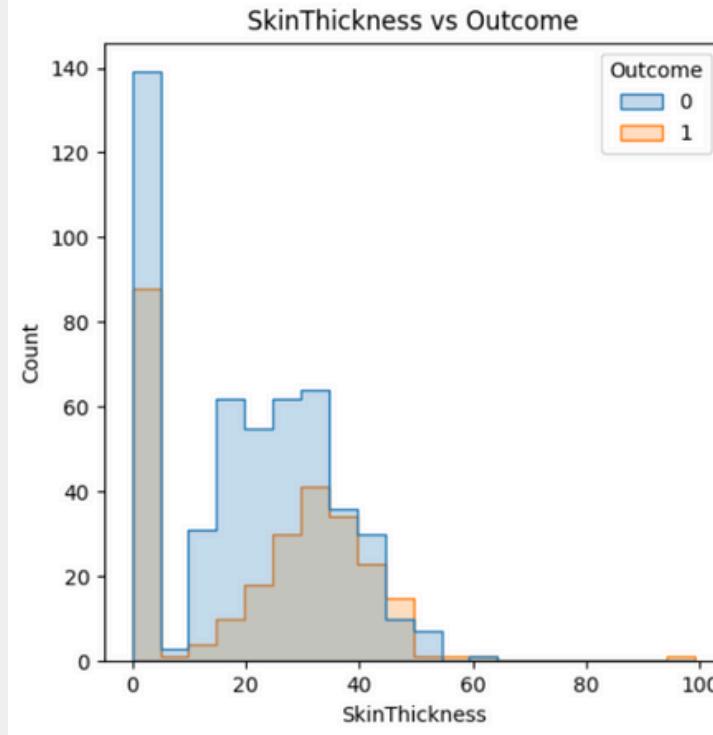
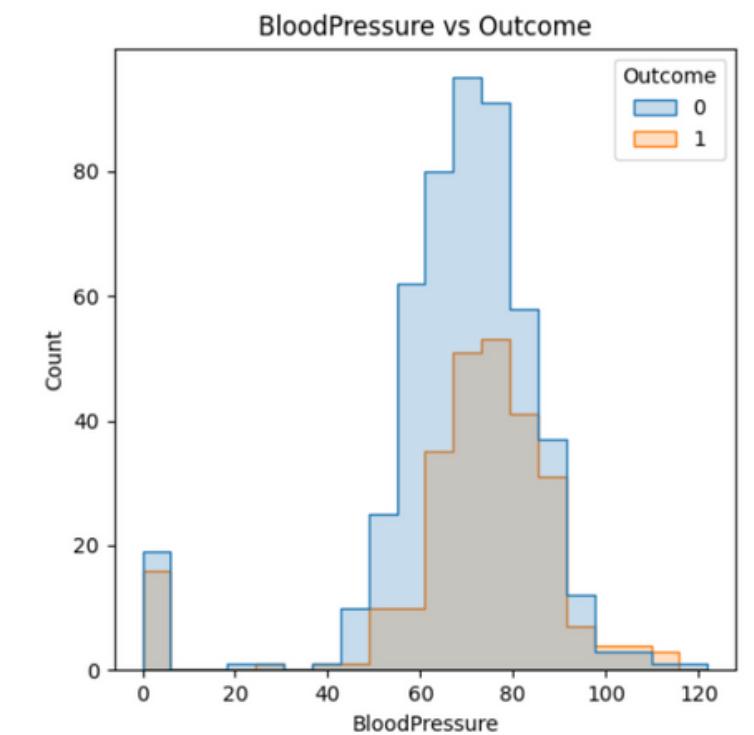
Hướng xử lý đề xuất:

- Glucose, BloodPressure, BMI:
 → Loại bỏ các dòng có giá trị 0 (ít, không ảnh hưởng lớn) → chỉ chiếm 6%
- SkinThickness, Insulin:
 → Impute (ước lượng) giá trị thiếu bằng mô hình hồi quy tuyến tính
 ! Tránh loại bỏ do mất dữ liệu lớn → ảnh hưởng phân tích

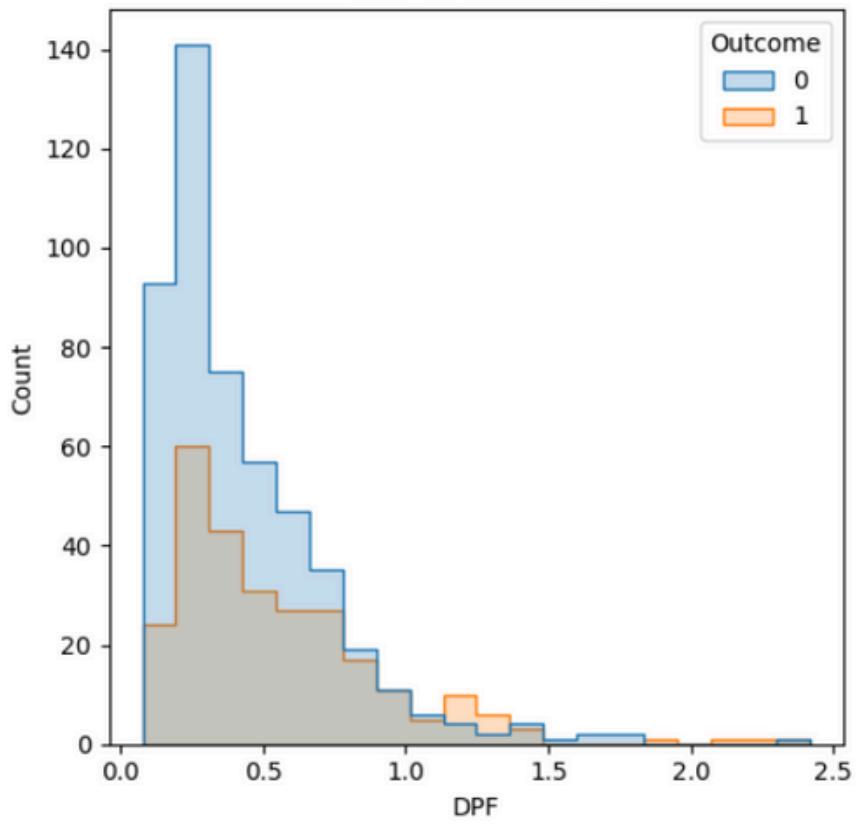
IV.3. PHÂN TÍCH PHÂN PHỐI CÁC BIẾN

Sử dụng Histogram để quan sát phân bố tần suất.

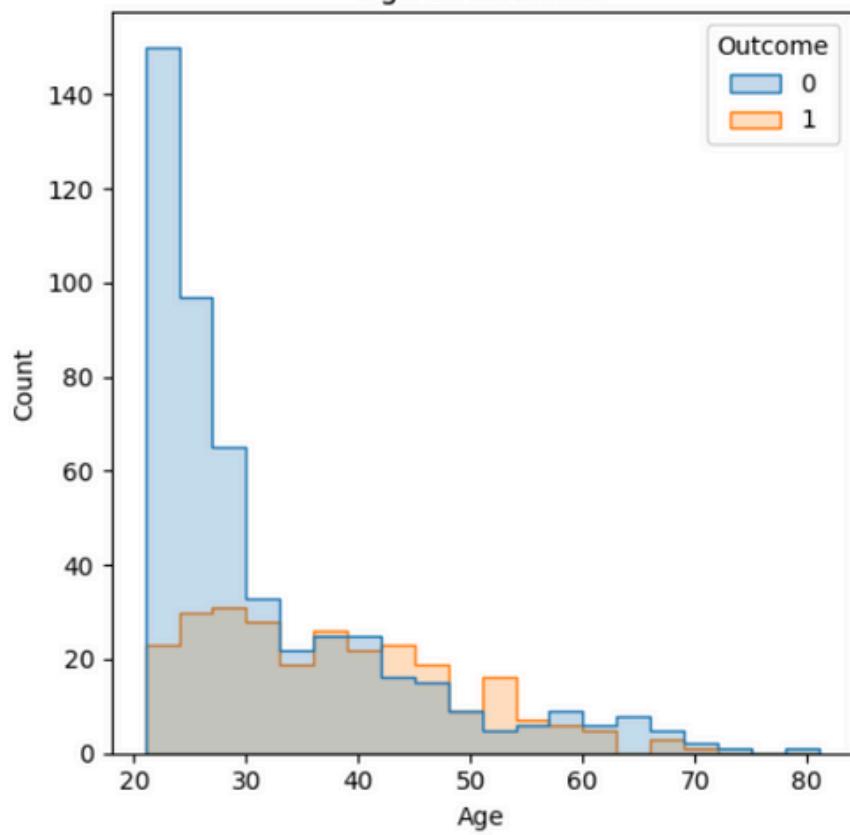


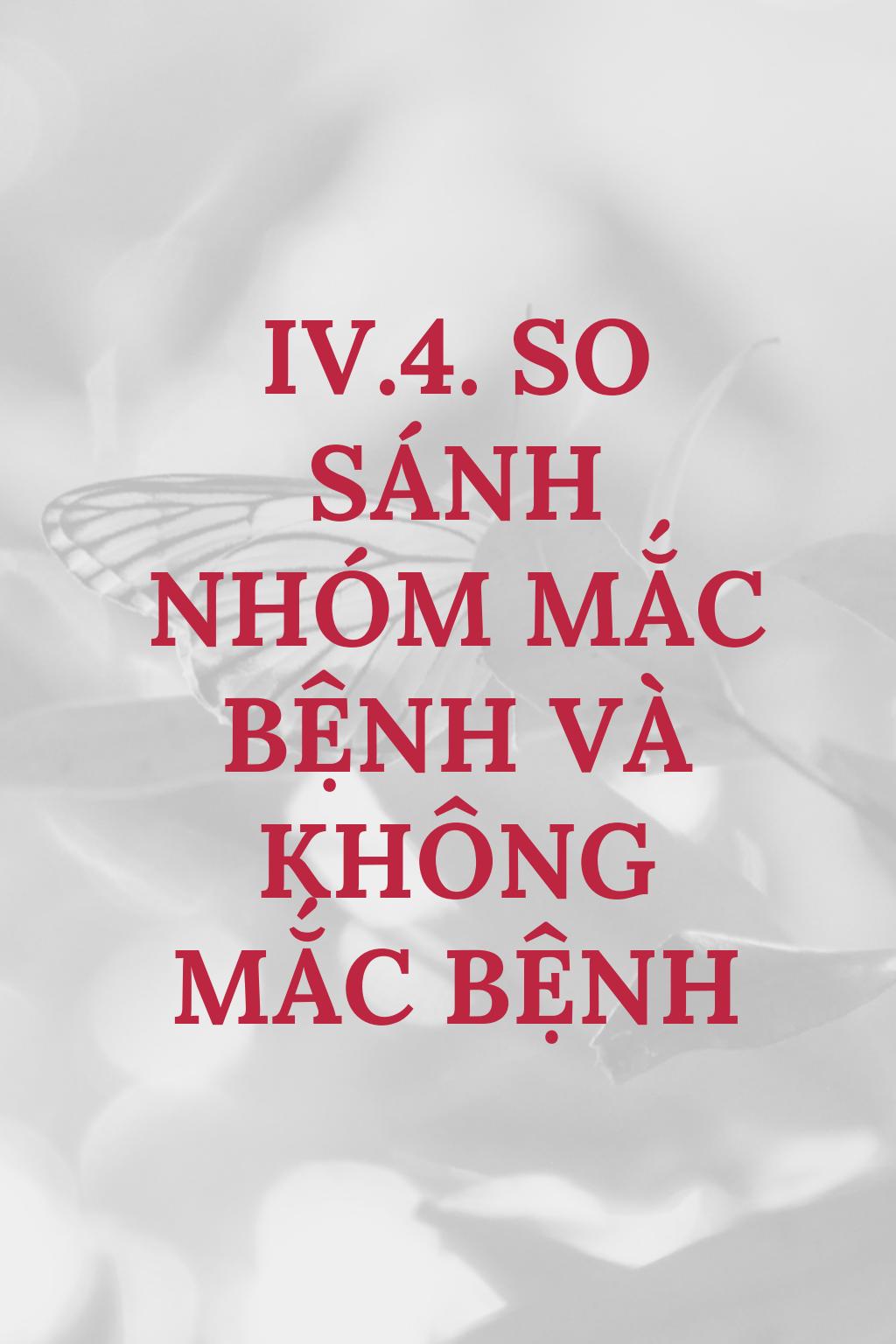


DPF vs Outcome



Age vs Outcome





IV.4. SO SÁNH NHÓM MẮC BỆNH VÀ KHÔNG MẮC BỆNH

IV.4.1 THỐNG KÊ MÔ TẢ
PHÂN TÁCH THEO
OUTCOME

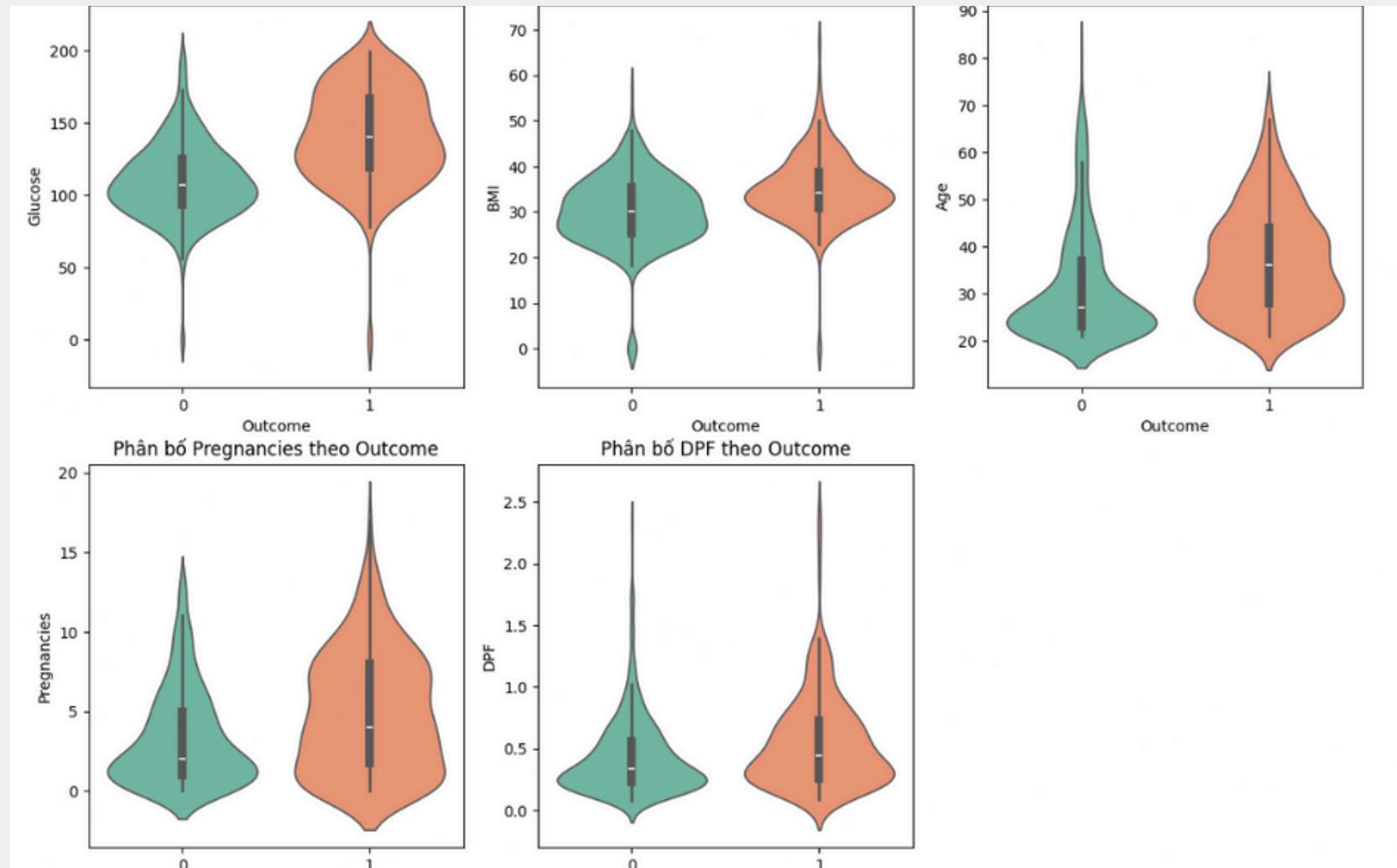
IV.4.2 VIOLIN PLOT 5
THUỘC TÍNH

IV.4.1 THỐNG KÊ MÔ TẢ PHÂN TÁCH THEO OUTCOME

- Các biến Glucose, BMI, Age, Pregnancies, DPF có sự khác biệt rõ rệt giữa hai nhóm → nhiều khả năng là các predictor mạnh.
- Insulin có dữ liệu chưa sạch (median = 0 ở nhóm tiểu đường) → cần cẩn trọng.
- BloodPressure và SkinThickness ít khác biệt hơn, có thể ảnh hưởng nhưng không mạnh.

		Outcome	0	1
Pregnancies	mean	3.298000	4.865672	
	median	2.000000	4.000000	
	std	3.017185	3.741239	
Glucose	mean	109.980000	141.257463	
	median	107.000000	140.000000	
	std	26.141200	31.939622	
BloodPressure	mean	68.184000	70.824627	
	median	70.000000	74.000000	
	std	18.063075	21.491812	
SkinThickness	mean	19.664000	22.164179	
	median	21.000000	27.000000	
	std	14.889947	17.679711	
Insulin	mean	68.792000	100.335821	
	median	39.000000	0.000000	
	std	98.865289	138.689125	
BMI	mean	30.304200	35.142537	
	median	30.050000	34.250000	
	std	7.689855	7.262967	
DiabetesPedigreeFunction	mean	0.429734	0.550500	
	median	0.336000	0.449000	
	std	0.299085	0.372354	
Age	mean	31.190000	37.067164	
	median	27.000000	36.000000	
	std	11.667655	10.968254	

IV.4.2 VIOLIN PLOT 5 THUỘC TÍNH



- Glucose là đặc trưng mạnh nhất để phân biệt hai nhóm.
- BMI, Age, Pregnancies cũng thể hiện sự khác biệt, đóng vai trò là yếu tố nguy cơ bổ sung.
- DPF ít có sự khác biệt giữa hai nhóm.

IV.5. PHÂN TÍCH ĐA BIỂN VÀ TƯƠNG QUAN

**IV.5.1 HEATMAP TƯƠNG
QUAN GIỮA CÁC BIỂN**

**IV.5.2 PAIRPLOT (SCATTER
MATRIX)**

**IV.5.3 NHẬN XÉT TỔNG
QUAN:**

IV.5.1 HEATMAP TƯƠNG QUAN GIỮA CÁC BIẾN

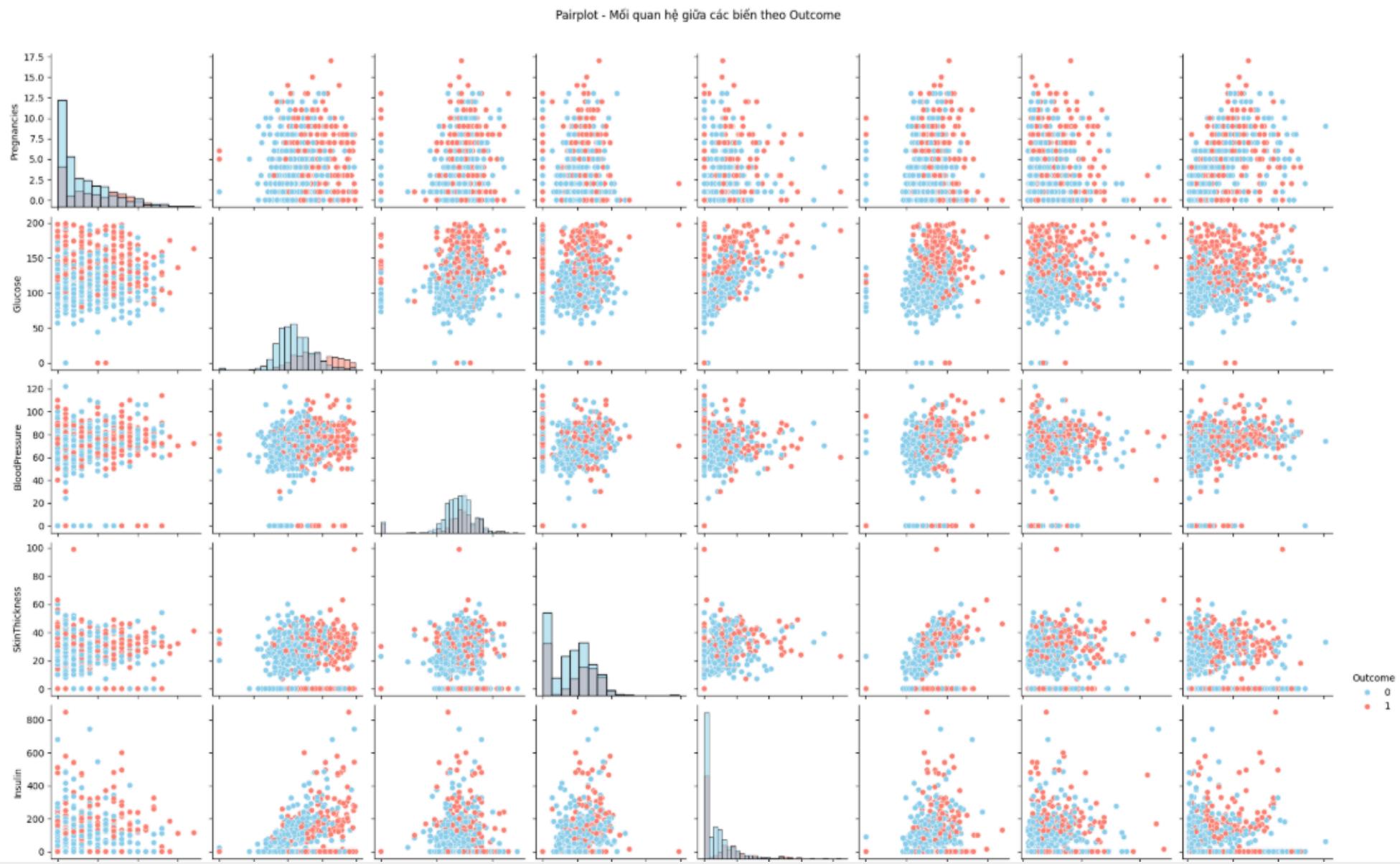
Heatmap tương quan giữa các biến

Pregnancies -

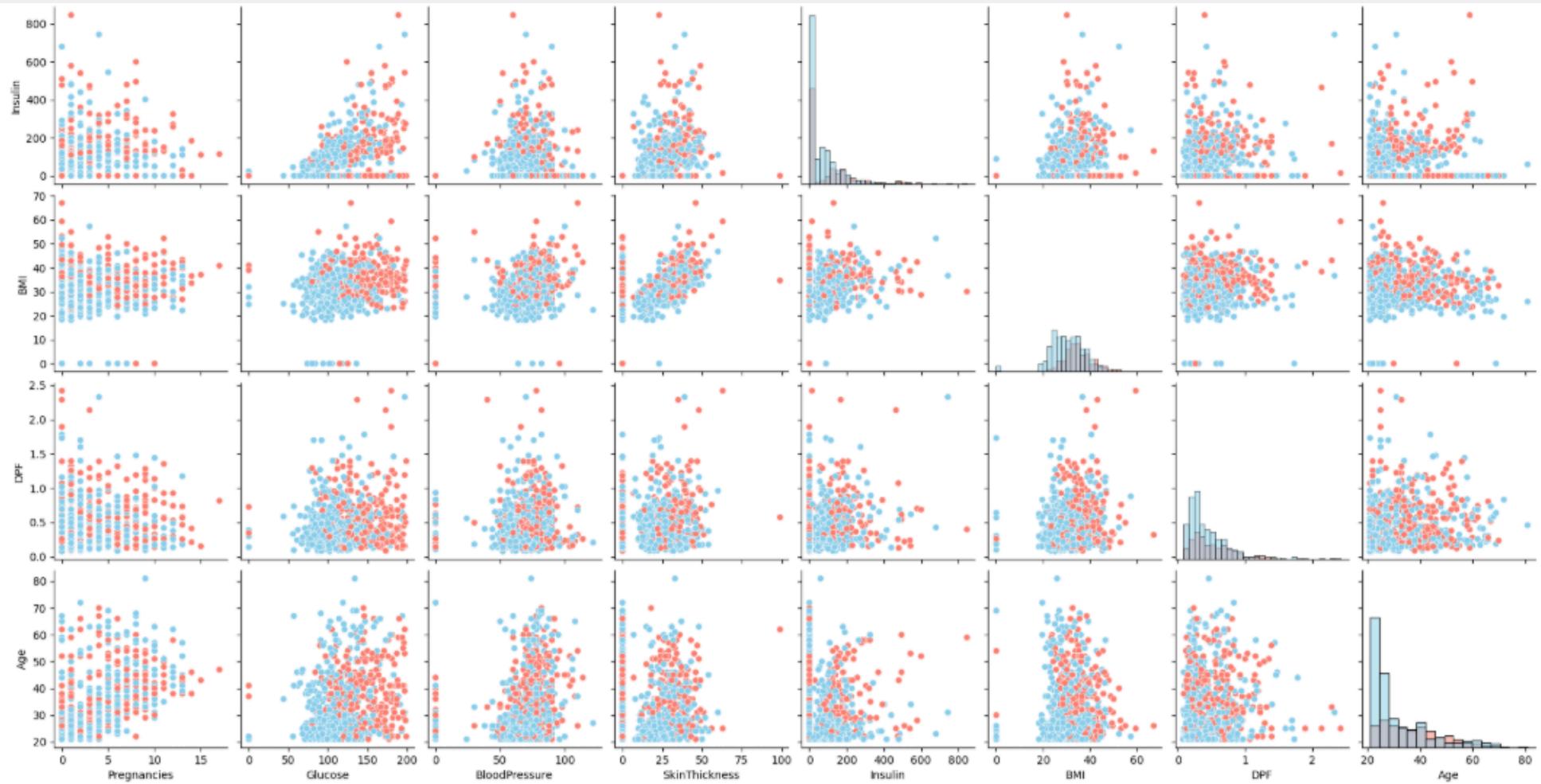
	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DPF	Age	Outcome
Pregnancies	-	0.13							
Glucose	0.13	-	0.15						
BloodPressure	0.14	0.15	-						
SkinThickness	-0.08	0.06	0.21	-					
Insulin	-0.07	0.33	0.09	0.44	-				
BMI	0.02	0.22	0.28	0.39	0.20	-			
DPF	-0.03	0.14	0.04	0.18	0.19	0.14	-		
Age	0.54	0.26	0.24	-0.11	-0.04	0.04	0.03	-	
Outcome	0.22	0.47	0.07	0.07	0.13	0.29	0.17	0.24	-



IV.5.2 PAIRPLOT (SCATTER MATRIX)



IV.5.2 PAIRPLOT (SCATTER MATRIX)

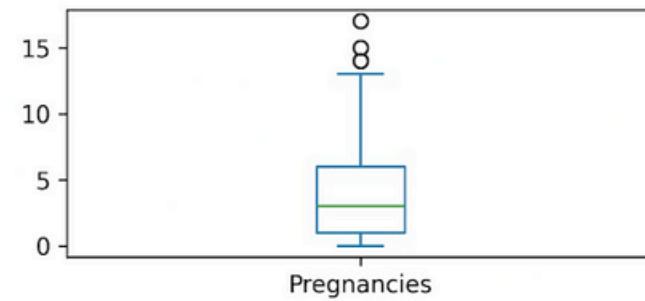


IV.5.3 NHẬN XÉT TỔNG QUAN:

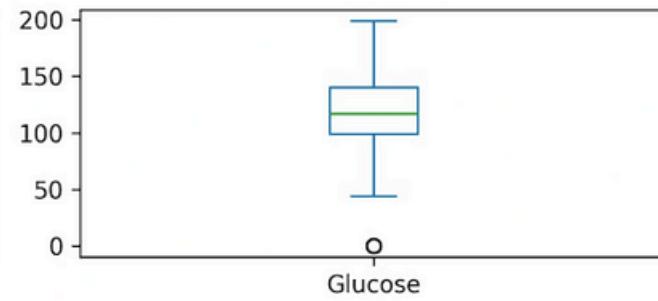
- Glucose (0.47) và BMI (0.29):
 - + Hai biến này cho thấy sự phân tách khá rõ giữa hai nhóm.
 - + **Outcome có tương quan cao nhất với Glucose (0.47)** → xác nhận tầm quan trọng.
 - + Nhóm Outcome = 1 (bệnh) thường có Glucose và BMI cao hơn.
 - + Điều này khẳng định lại kết quả từ heatmap và histogram: Glucose và BMI là hai đặc trưng quan trọng nhất để phân loại.
- Age (0.24) và Pregnancies (0.22):
 - + Có xu hướng cao hơn ở nhóm bệnh, nhưng sự chồng lấn vẫn nhiều.
 - + Không thể phân biệt rõ rệt chỉ dựa vào hai biến này.
- BloodPressure, SkinThickness, Insulin tương quan thấp (<0.15):
 - + Phân bố giữa hai nhóm gần như trùng nhau, hỗn loạn và khó tách biệt.

IV.6 BOXPLOT ĐƯỢC SỬ DỤNG ĐỂ PHÁT HIỆN GIÁ TRỊ NGOẠI LAI TRONG CÁC BIẾN

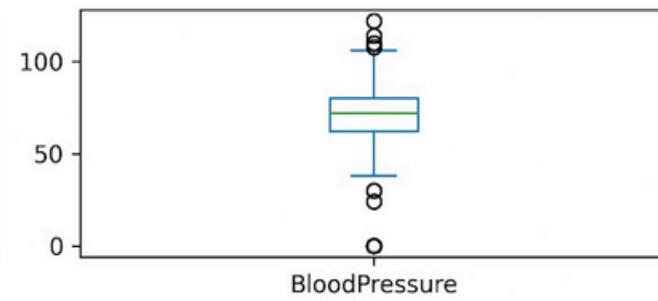
Boxplot cho từng đặc trưng



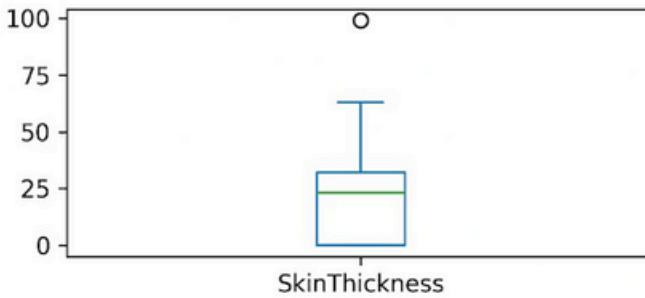
Pregnancies



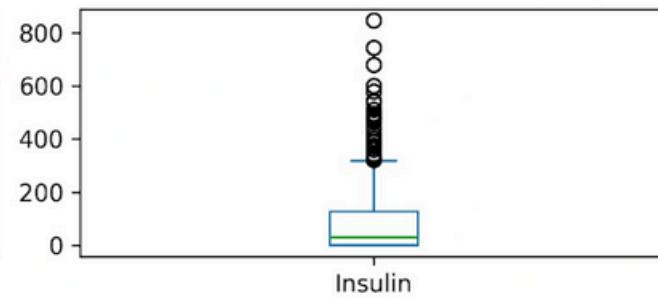
Glucose



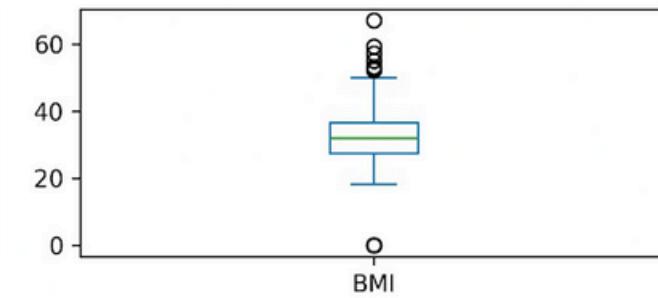
BloodPressure



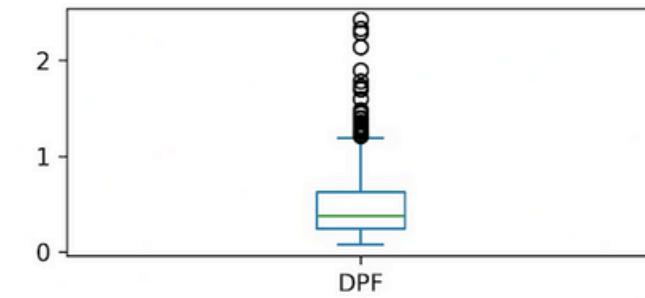
SkinThickness



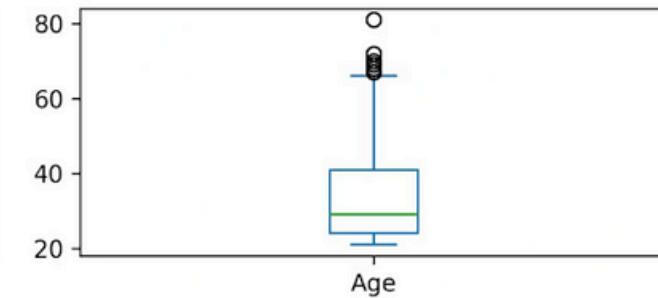
Insulin



BMI



DPF



Age

NHẬN XÉT THEO BIẾN:

- Pregnancies: một số trường hợp ≥ 15 lần mang thai \rightarrow hiếm gặp, có thể coi là ngoại lệ.
- Insulin: nhiều giá trị cực trị $> 500 \mu\text{U/mL}$ \rightarrow bất thường, có thể do sai số đo hoặc đặc thù sinh lý.
- BMI: ít ngoại lệ, phân phối tập trung hơn.
- Glucose: một số giá trị rất cao (~ 200), nhưng không quá nhiều.
- BloodPressure: có giá trị bằng 0 (không hợp lý).
- SkinThickness: nhiều giá trị 0, coi là thiếu dữ liệu.
- Age: phân phối tự nhiên, ít outliers.
- DiabetesPedigreeFunction(DPF): có vài giá trị cao > 2.0 .

IV.7. KẾT QUẢ TỔNG HỢP TỪ EDA

CHẤT LƯỢNG DỮ LIỆU

- Dataset: 768 dòng, 9 cột (8 biến đầu vào + 1 biến mục tiêu).
- Không có NaN, Null, không trùng lặp.
- Nhiều giá trị 0 bất hợp lý ở các biến sinh lý (Glucose, BloodPressure, SkinThickness, Insulin, BMI).
 - Insulin: ~48.7% bằng 0.
 - SkinThickness: ~30% bằng 0.
- Đây được xem là missing values → cần xử lý trước khi huấn luyện.

PHÂN BỐ DỮ LIỆU (UNIVARIATE)

- **Glucose**: lệch phải, cao hơn ở nhóm mắc bệnh → quan trọng nhất.
- **BMI**: đa số thừa cân/béo phì, đặc biệt nhóm mắc bệnh.
- **Age**: mắc bệnh nhiều ở >40 tuổi.
- **Pregnancies**: số lần mang thai cao liên quan đến bệnh.
- **DPF**: phản ánh di truyền, cao hơn ở nhóm mắc bệnh.
- **BloodPressure, SkinThickness, Insulin**: nhiều bất thường, ít giá trị phân biệt.

SO SÁNH THEO OUTCOME (BIVARIATE)

- Violin plot: sự khác biệt rõ ở Glucose, BMI, Age, Pregnancies, DPF.
- Các biến khác ít phân biệt.

PHÂN TÍCH ĐA BIẾN (MULTIVARIATE)

- Heatmap: Glucose ($r \approx 0.47$) tương quan mạnh nhất với Outcome.
- BMI, Age, Pregnancies: tương quan dương, yếu hơn.
- BloodPressure, SkinThickness, Insulin: gần như không tương quan.
- Pairplot: phân tách rõ ở cặp Glucose-BMI, Age-Pregnancies.

TỔNG HỢP & KẾT LUẬN

- **Biến quan trọng:** Glucose, BMI, Age, Pregnancies.
- Bổ trợ: DPF.
- Ít giá trị phân loại: BloodPressure, SkinThickness, Insulin.
- Dataset cân bằng tương đối (65% không bệnh, 35% bệnh).
- Glucose & BMI là yếu tố then chốt trong phân loại.
- Cần xử lý giá trị 0 bất hợp lý trước khi mô hình hóa.

IV.8. XỬ LÝ DỮ LIỆU THIẾU VÀ BẤT HỢP LÝ

Trong EDA phát hiện nhiều giá trị 0 phi thực tế ở các biến sinh lý
→ coi là missing values.

Nguyên tắc xử lý:

- Glucose, BloodPressure, BMI: số lượng 0 ít → loại bỏ trực tiếp các dòng chứa 0.
- SkinThickness, Insulin: tỷ lệ 0 cao → thay thế bằng giá trị dự đoán (Linear Regression Imputation) dựa trên Glucose, BMI, Age.

Size of dataframe: (724, 9)

	min
Pregnancies	0.000
Glucose	44.000
BloodPressure	24.000
SkinThickness	0.000
Insulin	0.000
BMI	18.200
DPF	0.078
Age	21.000
Outcome	0.000

IV.8.1 CÁC BƯỚC XỬ LÝ CHI TIẾT

Bước 1: Loại bỏ các dòng có giá trị 0 ở các biến:

- Glucose
- BloodPressure
- BMI

Nhận xét bước 1:

- Có 44 dòng có giá trị 0 ở cột Glucose, BloodPressure và BMI được loại bỏ.
- Giá trị min của 3 cột này đã trở nên hợp lý.

Bước 2: Thay thế giá trị 0 ở biến SkinThickness và Insulin bằng giá trị dự đoán từ mô hình hồi quy tuyến tính, sử dụng các biến liên quan như Glucose, BMI, Age và DiabetesPedigreeFunction

Nhận xét bước 2: Giá trị của Insulin và SkinThickness sau khi xử lý đã trở lên bình thường hơn, phân phôi rõ rệt hơn.

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DPF	Age	Outcome
count	724.000000	724.000000	724.000000	724.000000	724.000000	724.000000	724.000000	724.000000	724.000000
mean	3.866022	121.882597	72.400552	28.903743	152.616866	32.467127	0.474765	33.350829	0.343923
std	3.362803	30.750030	12.379870	9.699674	99.676938	6.888941	0.332315	11.765393	0.475344
min	0.000000	44.000000	24.000000	7.000000	-20.230088	18.200000	0.078000	21.000000	0.000000
25%	1.000000	99.750000	64.000000	22.000000	88.000000	27.500000	0.245000	24.000000	0.000000
50%	3.000000	117.000000	72.000000	28.440016	130.000000	32.400000	0.379000	29.000000	0.000000
75%	6.000000	142.000000	80.000000	35.000000	190.000000	36.600000	0.627500	41.000000	1.000000
max	17.000000	199.000000	122.000000	99.000000	846.000000	67.100000	2.420000	81.000000	1.000000

IV.8.2 NHẬN XÉT SAU KHI HOÀN THÀNH XỬ LÝ

- Dataset sau làm sạch không còn chứa các giá trị 0 bất hợp lý.
- Glucose, BloodPressure, BMI giữ lại tính chính xác và không còn vô lý do điền giá trị ảo.
- SkinThickness, Insulin đã được khôi phục bằng phương pháp hồi quy, phản ánh hợp lý hơn so với điền bằng trung bình hay trung vị.
- Số lượng dòng còn lại trong dataset vẫn đảm bảo đủ lớn để huấn luyện mô hình (trên 700 mẫu).

V. THẢO LUẬN KẾT QUẢ

V.1. Ý NGHĨA Y HỌC VÀ
THỰC TIỄN

V.2. HẠN CHẾ CỦA DỮ LIỆU

V.3. ĐỊNH HƯỚNG ỨNG
DỤNG TRONG HỌC MÁY

V.1. Ý NGHĨA Y HỌC VÀ THỰC TIỄN

Ý NGHĨA Y HỌC

- Phát hiện sớm nguy cơ: Glucose, BMI, Age, Pregnancies liên quan chặt chẽ đến tiểu đường type 2 → hỗ trợ sàng lọc sớm.
- Chẩn đoán & điều trị:
 - Glucose: chỉ số lâm sàng quan trọng nhất (WHO, ADA).
 - BMI: khẳng định vai trò của béo phì và lối sống.
 - Pregnancies: nhấn mạnh nguy cơ tiểu đường thai kỳ.
 - Age: trung niên & cao tuổi cần sàng lọc thường xuyên.
- Dịch tễ học: Pima Indians cho thấy ảnh hưởng di truyền & môi trường; có thể so sánh với cộng đồng khác.

V.1. Ý NGHĨA Y HỌC VÀ THỰC TIỄN

Ý NGHĨA THỰC TIỄN

- Y tế dự phòng: Xây dựng công cụ sàng lọc nguy cơ dựa trên đặc trưng đơn giản, chi phí thấp.
- Quyết định lâm sàng: Kết quả EDA hỗ trợ hệ thống CDSS đưa ra khuyến nghị cho từng bệnh nhân.
- Phát triển mô hình học máy: Glucose, BMI, Age, Pregnancies → nền tảng xây dựng mô hình dự đoán.
- Tác động xã hội: Hỗ trợ chính sách cộng đồng (giảm béo phì, dinh dưỡng cho phụ nữ mang thai, xét nghiệm định kỳ).

V.2. HẠN CHẾ CỦA DỮ LIỆU

- Giới hạn đối tượng:
 - Chỉ nữ ≥ 21 tuổi, cộng đồng Pima Indians (Arizona, Mỹ).
 - Không khái quát hóa cho nam giới, dân tộc khác.
 - Ảnh hưởng bởi yếu tố nhân khẩu học & môi trường đặc thù.
- Dữ liệu thiếu & bất hợp lý:
 - Nhiều giá trị 0 phi thực tế (Insulin ~50%, SkinThickness ~30%).
 - Giảm độ tin cậy, phụ thuộc vào cách xử lý missing values.

V.2. HẠN CHẾ CỦA DỮ LIỆU

- Số lượng biến hạn chế:
 - Chỉ có 8 đặc trưng cơ bản.
 - Thiếu biến quan trọng: HbA1c, cholesterol, lối sống, gene...
- Tính thời gian:
 - Dữ liệu cắt ngang (cross-sectional).
 - Không có theo dõi → khó phân tích tiến triển bệnh.
- Mất cân bằng lớp:
 - 500 không bệnh vs 268 có bệnh.
 - Có thể gây lệch, mô hình thiên về "không bệnh".

V.3. ĐỊNH HƯỚNG ỨNG DỤNG TRONG HỌC MÁY

- Bài toán phân loại nhị phân
 - Dự đoán khả năng mắc tiểu đường (Outcome = 0/1).
 - Thuật toán: Logistic Regression, KNN, Decision Tree, Random Forest, XGBoost, ANN...
 - Đánh giá: Accuracy, Precision, Recall, F1-score, ROC-AUC.
- Phát hiện đặc trưng quan trọng
 - Dùng Random Forest, XGBoost, ANOVA...
 - Các đặc trưng thường quan trọng: Glucose, BMI, Age, Pregnancies.
 - Giúp chọn lọc đặc trưng (Feature Selection).
- Xử lý dữ liệu mất cân bằng
 - 65% (không bệnh) vs 35% (bệnh).
 - Kỹ thuật: SMOTE/ADASYN, Undersampling, Class weights.
 - Mục tiêu: giảm bỏ sót bệnh nhân (false negative).

V.3. ĐỊNH HƯỚNG ỨNG DỤNG TRONG HỌC MÁY

- Hệ thống hỗ trợ quyết định
 - Tích hợp vào hồ sơ y tế điện tử (EHR).
 - Ứng dụng khám sàng lọc cộng đồng → cảnh báo nguy cơ cao.
- Định hướng mở rộng
 - Kết hợp dữ liệu bổ sung: HbA1c, cholesterol, thói quen sinh hoạt...
 - Phân tích tiến triển bệnh (longitudinal analysis).
 - Ứng dụng Explainable AI (SHAP, LIME) → tăng minh bạch, tin cậy lâm sàng.

VI. KẾT LUẬN

VI.1. KẾT QUẢ CHÍNH

VI.2. Ý NGHĨA Y HỌC VÀ THỰC
TIỄN

VI.3. HẠN CHẾ CỦA DỮ LIỆU

VI.4. ĐỊNH HƯỚNG NGHIÊN CỨU
VÀ ỨNG DỤNG

VI.5 KẾT LUẬN CUỐI CÙNG

VI.1. KẾT QUẢ CHÍNH

Chất lượng dữ liệu

- Không có NaN, không trùng lặp.
- Nhiều giá trị 0 bất hợp lý (Glucose, BloodPressure, BMI, SkinThickness, Insulin)
→ missing cần xử lý.

Phân tích đơn biến

- Glucose, BMI: phân biệt rõ rệt nhóm bệnh/không bệnh.
- Age, Pregnancies: liên quan đến Outcome.
- DPF: ảnh hưởng nhẹ.
- BloodPressure, SkinThickness, Insulin: giá trị phân loại kém.

Phân tích hai biến

- Violin plot: Glucose & BMI khác biệt rõ nhất.
- Age & Pregnancies: có khác biệt nhưng yếu hơn.

Phân tích đa biến

- Heatmap: Glucose tương quan mạnh nhất với Outcome ($r \approx 0.47$).
- Pairplot: cặp Glucose-BMI và Age-Pregnancies phân biệt tốt nhất.

Xử lý dữ liệu bất hợp lý

- Loại bỏ dòng có giá trị 0 ở Glucose, BMI, BloodPressure.
- Impute SkinThickness, Insulin bằng hồi quy tuyến tính.

VI.2. Ý NGHĨA Y HỌC VÀ THỰC TIỄN

Yếu tố nguy cơ kinh điển:

- Glucose cao
- BMI cao (béo phì)
- Tuổi cao
- Nhiều lần mang thai

Ứng dụng:

- Khám sàng lọc cộng đồng → phát hiện sớm nguy cơ.
- Nền tảng xây dựng Clinical Decision Support Systems (CDSS).

VI.3. HẠN CHẾ CỦA DỮ LIỆU

- Dataset chỉ gồm nữ giới Pima Indians ≥ 21 tuổi → hạn chế tính khái quát cho các cộng đồng khác.
- Tỷ lệ dữ liệu bị thiếu cao ở một số biến (đặc biệt là Insulin và SkinThickness).
- Thiếu nhiều đặc trưng y học hiện đại (HbA1c, cholesterol, lối sống, di truyền nâng cao).
- Không có dữ liệu theo dõi dài hạn (longitudinal).

VI.4. ĐỊNH HƯỚNG NGHIÊN CỨU VÀ ỨNG DỤNG

- Học máy: so sánh nhiều thuật toán (Logistic Regression, Random Forest, XGBoost, Neural Networks).
- Ứng dụng y tế: mô hình dự đoán nguy cơ → sàng lọc & quản lý hồ sơ bệnh án điện tử.
- Phát triển dữ liệu: bổ sung HbA1c, lipid máu, lối sống; mở rộng giới tính & dân tộc.
- Explainable AI: SHAP, LIME → giải thích mô hình, tăng minh bạch & tin cậy.

VI.5 KẾT LUẬN CUỐI CÙNG

- Các yếu tố then chốt: Glucose, BMI, Age, Pregnancies, DPF.
- Củng cố bằng chứng y học về nguy cơ tiểu đường type 2.
- Mở hướng ứng dụng học máy trong dự báo, y tế dự phòng & hỗ trợ quyết định lâm sàng.
- Dù còn hạn chế, dataset Pima Indians Diabetes vẫn là nền tảng quan trọng cho giảng dạy, nghiên cứu & phát triển AI trong y học.

THANK YOU
thank you
SO MUCH!
so much!