BÁO CÁO PHÂN TÍCH KHÁM PHÁ DỮ LIỆU (EDA)

I. Tổng quan Dữ liệu

Tập dữ liệu được phân tích là pima-indians-diabetes.csv, chứa 768 mục nhập (entries). Tập dữ liệu này bao gồm 9 cột, với tên được gán là: **'Pregnancies'** (Số lần mang thai), **'Glucose'** (Đường huyết), **'BloodPressure'** (Huyết áp), **'SkinThickness'** (Độ dày da), **'Insulin'**, **'BMI'** (Chỉ số khối cơ thể), **'DiabetesPedigreeFunction'** (Chức năng phả hệ tiểu đường), **'Age'** (Tuổi), và **'Outcome'** (Kết quả - biến mục tiêu).

Ban đầu, dữ liệu bao gồm 2 cột thuộc kiểu float64 và 7 cột thuộc kiểu int64.

II. Khảo sát và Làm sạch Dữ liệu

1. Vấn đề về giá trị 0

Mặc dù việc kiểm tra ban đầu cho thấy tất cả các cột đều có 768 giá trị không thiếu (Non-Null Count), nhưng thống kê mô tả ban đầu (df.describe()) cho thấy các giá trị tối thiểu (min) bằng 0 ở các biến sau, điều này được nhận xét là bất hợp lý hoặc thiếu dữ liệu ghi nhận:

• **Glucose** (Đường huyết)

• **BloodPressure** (Huyết áp)

• **SkinThickness** (Độ dày da)

• **Insulin**

• **BMI** (Chỉ số khối cơ thể)

• *Lưu ý:* Giá trị min=0 cho **Pregnancies** được coi là hợp lệ (chưa mang thai lần nào).

2. Quy trình Xử lý Giá trị Bất hợp lý

Để xử lý các giá trị 0 bất hợp lý này, các bước sau đã được thực hiện:

1. Các cột chứa giá trị 0 bất hợp lý (Glucose, BloodPressure, SkinThickness, Insulin, BMI) được chuyển về kiểu float.

2. Giá trị 0 trong các cột này được thay thế bằng giá trị thiếu (NaN).

3. Các giá trị NaN sau đó được điền bằng **giá trị trung bình (mean)** của từng cột tương ứng.

3. Tình trạng Dữ liệu sau Làm sạch

Sau khi làm sạch, dữ liệu **đã không còn giá trị NaN**. Thống kê mô tả sau khi xử lý cho thấy các giá trị trung bình (mean) của các biến đã thay đổi đáng kể so với trước khi làm sạch:

Biến	Giá trị Trung bình (Mean) sau làm sạch	Giá trị Tối thiểu (Min) sau làm sạch
Glucose	121.686763	44.000000
BloodPressure	72.405184	24.000000
SkinThickness	29.153420	7.000000
Insulin	155.548223	14.000000
BMI	32.457464	18.200000

III. Phân tích Đơn biến và Trực quan hóa

1. Phân phối các biến (Histograms và Skewness)

Phân tích histogram và độ lệch (skewness) cung cấp các nhận xét sau về phân phối của các biến:

• **Phân phối lệch phải (Right-skewed):**

◦ **Insulin** có độ lệch phải cao nhất (3.019).

◦ **DiabetesPedigreeFunction** có độ lệch cao thứ hai (1.920).

◦ Các biến **Age** (1.13), **Pregnancies** (0.90), và **BMI** (0.60) cũng có xu hướng lệch phải.
Điều này cho thấy phần lớn bệnh nhân là người trẻ tuổi và có số lần mang thai ít.

• **Phân phối gần chuẩn: Glucose** và **BloodPressure** có phân phối gần giống với phân phối chuẩn.

2. Phân tích Biến Mục tiêu (Outcome)

Biến **Outcome** cho biết bệnh nhân có bị tiểu đường (1) hay không (0).

• Kết quả đếm cho thấy **số lượng người không bị bệnh nhiều hơn gần gấp đôi số người bị bệnh**.

• Điều này dẫn đến nhận xét rằng dữ liệu đang ở trạng thái **mất cân bằng**.

IV. Phân tích Đa biến

1. Ma trận Tương quan (Correlation Matrix)

Các mối tương quan dương đáng chú ý nhất giữa các cặp biến là:

Cặp biến	Hệ số tương quan	Nhận xét
BMI và **SkinThickness**	0.57	Hợp lý, chỉ số khối cơ thể thường đi đôi với độ dày lớp mỡ dưới da.
Age và **Pregnancies**	0.54	Hợp lý, tuổi càng cao thì số lần mang thai có thể càng nhiều.
Glucose và **Outcome**	0.49	Cho thấy mức đường huyết cao có liên quan chặt chẽ đến việc bị tiểu đường.
Glucose và **Insulin**	0.42	Mối quan hệ sinh lý học đã được biết đến.

Các mối tương quan khác được nhận xét là khá yếu.

2. Pairplot

Phân tích biểu đồ cặp (Pairplot) tập trung vào các biến quan trọng ('Glucose', 'BMI', 'Age', 'Pregnancies') so với 'Outcome' cho thấy:

• **Phân tách:** Phân phối của người bị bệnh (màu cam) thường dịch về phía bên phải (giá trị cao hơn) đối với các biến Glucose, BMI, Age, và Pregnancies.

• **Glucose:** Biểu đồ của Glucose so với các biến khác cho thấy sự phân tách rõ ràng hơn giữa hai nhóm Outcome.

V. Phân tích Ngoại lệ và Phân phối theo Outcome

Phân tích biểu đồ hộp (Boxplot) các biến theo Outcome cho thấy:

• **Glucose:** Nhóm người bị bệnh (Outcome=1) có mức đường huyết **trung vị cao hơn hẳn** so với nhóm không bị bệnh.

• **Age & Pregnancies:** Trung vị của tuổi và số lần mang thai đều cao hơn ở nhóm bị bệnh. Điều này chỉ ra rằng bệnh tiểu đường phổ biến hơn ở những người lớn tuổi và đã mang thai nhiều lần.

• **BMI & SkinThickness:** Cả hai chỉ số này đều có trung vị cao hơn ở nhóm người bị bệnh. Cả hai biến này đều có nhiều giá trị ngoại lệ (outliers) ở cả hai nhóm.

• **Insulin:** Biến này có **số lượng lớn các giá trị ngoại lệ**, phản ánh sự biến thiên rất lớn trong nồng độ insulin giữa các cá nhân, đặc biệt là ở những người mắc bệnh.

VI. Kết luận Chung

Dựa trên toàn bộ quá trình Phân tích Khám phá Dữ liệu, kết luận được đưa ra là:

Các biến Glucose, BMI, và Age dường như là những yếu tố dự báo quan trọng cho việc chẩn đoán bệnh tiểu đường.