# TRƯỜNG ĐẠI HỌC SÀI GÒN

KHOA CÔNG NGHỆ THÔNG TIN





BÁO CÁO TIỂU LUẬN PHÂN TÍCH DỮ LIỆU

**PHÂN TÍCH KHÁM PHÁ VỀ BỆNH ĐÁI THÁO ĐƯỜNG**

**Giảng viên hướng dẫn: Đỗ Như Tài**

**Thông tin thành viên:**

Lê Nguyễn Quốc Hưng 3121410240

Trần Bá Tiên 3122410410

Dương Huỳnh Gia 3121410167

Bùi Thành Công 3122410042

***TP. Hồ Chí Minh, tháng 9/2025***

NHẬN XÉT VÀ ĐÁNH GIÁ CỦA GIẢNG VIÊN

…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..……………………………………….……………………..…………………………………..

# LỜI CẢM ƠN

Chúng em xin gửi lời cảm ơn chân thành và sự tri ân sâu sắc đối với các thầy cô của trường Đại Học Sài Gòn, đặc biệt là các thầy cô ở khoa Công Nghệ Thông Tin đã tạo điều kiện cho em tiếp cận và tìm hiểu để hoàn thành đồ án môn học lần này.

Và chúng em cũng xin chân thành đặc biệt cảm ơn thầy Đỗ Như Tài, thầy là giảng viên giảng dạy đã nhiệt tình hướng dẫn giúp chúng em hoàn thành được đồ án lần này.

Trong quá trình nghiên cứu và làm bài báo cáo đồ án, kiến thức cũng như kinh nghiệm thực tế còn nhiều hạn chế nên bài báo cáo không thể tránh khỏi những thiếu sót, chúng em rất mong nhận được ý kiến đóng góp của thầy, cô để em học hỏi được nhiều kỹ năng, kinh nghiệm và sẽ hoàn thành tốt hơn cho những bài báo cáo sắp tới.

Chúng em xin chân thành cảm ơn!

# MỤC LỤC

[LỜI CẢM ƠN](#_heading=h.30j0zll) 1

[MỤC LỤC](#_heading=h.os4kp4kuwm50) 2

[I. GIỚI THIỆU](#_heading=h.ua56i0g34v20) 3

[II. ĐỊNH NGHĨA VẤN ĐỀ (DEFINE PROBLEM)](#_heading=h.2mocjcycxqnx) 3

[III. CHUẨN BỊ VẤN ĐỀ (PREPARE PROBLEM)](#_heading=h.b2e9p9q2tsck) 3

[IV. PHÂN TÍCH](#_heading=h.ljbasi8muu8p) 4

[1. EDA không dùng biểu đồ (Non-graphie EDA)](#_heading=h.9t7utas2xgc2) 4

[1.1 Dataset](#_heading=h.5mt06wsad10) 4

[1.2 In 5 dòng data đầu](#_heading=h.vlulr19azqcn) 4

[1.3 Kiểm tra thông tin khái quát](#_heading=h.58n6hafekji3) 5

[1.4 Kiểm tra giá trị Null/NaN cho từng cột](#_heading=h.gdxo6sowxae2) 5

[1.5 Thống kê mô tả cho từng cột](#_heading=h.g3vqrf94p12c) 6

[1.6 Mối tương quan giữa các cột](#_heading=h.gzg6hhixo64k) 6

[1.7 Phân phối độ tuổi](#_heading=h.vnqpb7ri6xqy) 7

[2. EDA trực quan bằng sơ đồ (Graphical EDA)](#_heading=h.xoyeki98hkwh) 8

[2.1 Phân phối độ tuổi bằng sơ đồ](#_heading=h.6i1d1g32uizt) 8

[A. Seaborn](#_heading=h.tuumyzjxaa2u) 8

[B. Pandas](#_heading=h.q2qai333uur5) 8

[C. Matplotlib](#_heading=h.xvmrrodjgi6p) 9

[2.2 Kết luận về độ tuổi](#_heading=h.ovz81ehun4xs) 9

[3. Biểu đồ so sánh (Số tuổi và số lần mang thai)](#_heading=h.nf2tjudfn06e) 10

[3.1 Biểu đồ phân tán sử dụng pandas](#_heading=h.385of2o8pv3p) 10

[3.2 Biểu đồ phân tán sử dụng matplotlib](#_heading=h.wgnnmkk1tach) 10

[3.3 Biểu đồ phân tán sử dụng seaborn](#_heading=h.qgajiwemkgkx) 11

[4. Biểu đồ trực quan hóa ma trận tương quan (Correlogram)](#_heading=h.ohej8729mhi7) 12

[5. Heatmap](#_heading=h.h5blzlt8vkto) 15

[5.1 Biểu đồ hộp (Box plot)](#_heading=h.hlyb92746vuh) 15

[5.2 Biểu đồ hình đàn vĩ cầm (Violin plot)](#_heading=h.94tz9givbapc) 16

[III. KẾT LUẬN](#_heading=h.2srdqcqlhx5) 17

[TÀI LIỆU THAM KHẢO](#_heading=h.bd6z3290hgfj) 18

[PHÂN CÔNG CÔNG VIỆC](#_heading=h.u03pprdrw9xk) 19

# I. GIỚI THIỆU

Bạn đã bao giờ tự hỏi những yếu tố nào có thể ảnh hưởng lớn đến khả năng mắc bệnh đái tháo đường? Liệu chỉ số đường huyết, chỉ số khối cơ thể (BMI) hay số lần mang thai có phải là những biến quan trọng nhất? Hoặc làm thế nào để dự đoán nguy cơ mắc bệnh từ thông tin y tế cơ bản của một người? Bộ dữ liệu về bệnh đái tháo đường này có thể giúp bạn khám phá những câu hỏi đó!

Bộ dữ liệu bao gồm thông tin y tế của các bệnh nhân nữ thuộc cộng đồng người Pima, với các đặc trưng như nồng độ glucose, huyết áp, độ dày lớp da, mức insulin, chỉ số BMI, yếu tố di truyền và độ tuổi. Mỗi bản ghi đều được gắn nhãn cho biết người đó có mắc bệnh đái tháo đường hay không. Tất cả thông tin cá nhân đã được loại bỏ để đảm bảo tính ẩn danh.

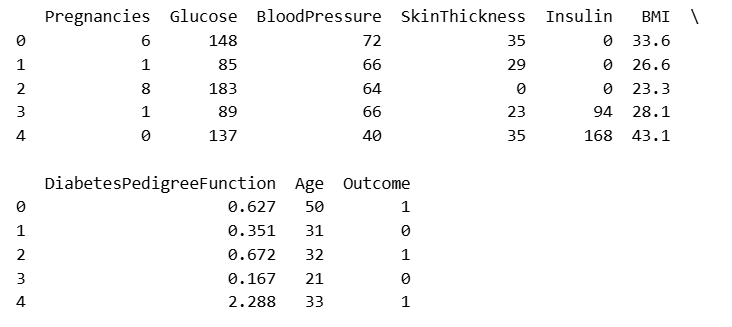
Hãy cùng phân tích và khám phá dữ liệu để tìm ra những yếu tố quan trọng quyết định đến khả năng mắc bệnh, từ đó hiểu rõ hơn về nguy cơ và hỗ trợ các mô hình dự đoán y tế.

# II. ĐỊNH NGHĨA VẤN ĐỀ (DEFINE PROBLEM)

* Mô tả:
* Bộ dữ liệu bao gồm thông tin y tế của nhiều bệnh nhân nữ từ 21 tuổi trở lên
* Các chỉ số liên quan đến sức khỏe được ghi nhận để phục vụ việc dự đoán khả năng mắc bệnh đái tháo đường
* Dữ liệu vào:
* Pregnancies: số lần mang thai của bệnh nhân.
* Glucose: nồng độ glucose trong huyết tương sau khi nhịn đói.
* BloodPressure: huyết áp tâm trương (mm Hg).
* SkinThickness: độ dày của lớp da gấp (mm) – dùng để ước tính lượng mỡ cơ thể.
* Insulin: nồng độ insulin trong huyết thanh (mu U/ml).
* BMI: chỉ số khối cơ thể (kg/m²).
* DiabetesPedigreeFunction: chỉ số dự đoán nguy cơ mắc bệnh đái tháo đường dựa trên yếu tố di truyền.
* Age: tuổi của bệnh nhân.
* Kết quả: Outcome: biến mục tiêu (label):(1: bệnh; 0: không bệnh)

# III. CHUẨN BỊ VẤN ĐỀ (PREPARE PROBLEM)

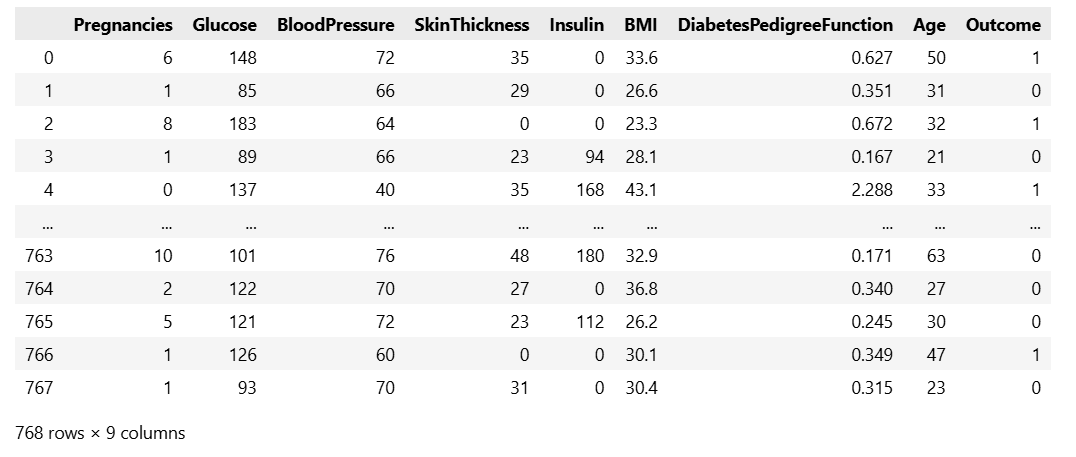
* Bộ dữ liệu được lấy từ Kaggle [1].

****

# IV. PHÂN TÍCH

## 1. EDA không dùng biểu đồ (Non-graphie EDA)

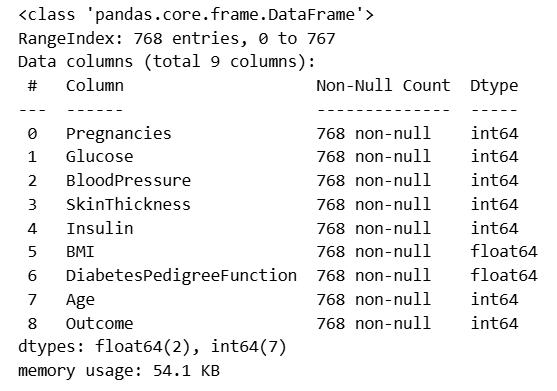
### 1.1 Dataset



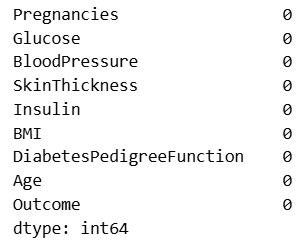
### 1.2 In 5 dòng data đầu



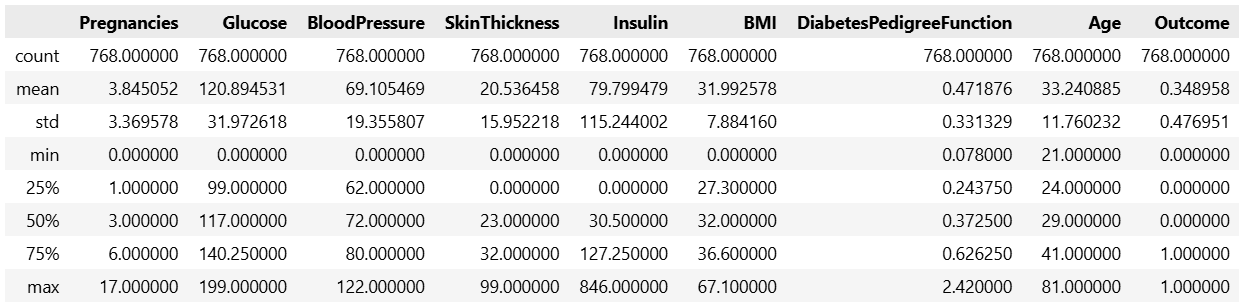
### 1.3 Kiểm tra thông tin khái quát



### 1.4 Kiểm tra giá trị Null/NaN cho từng cột



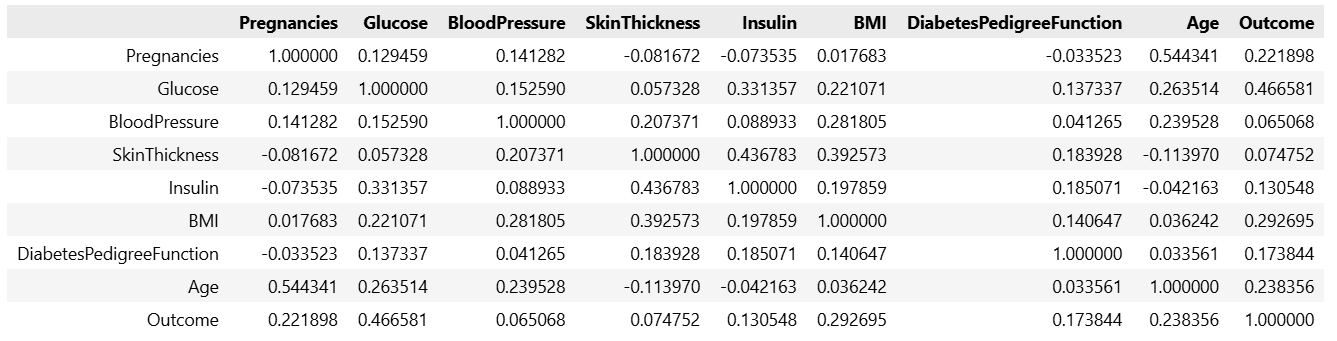
### 1.5 Thống kê mô tả cho từng cột



* Count: số lượng giá trị
* Mean: giá trị trung bình
* Std: độ lệch chuẩn (mức độ phân tán)
* Min, max: giá trị nhỏ nhất, lớn nhất
* 25%, 50%, 75%: các phân vị (Q1, median, Q3)

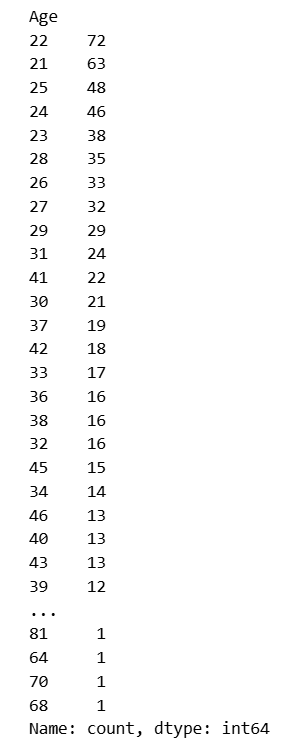
### 1.6 Mối tương quan giữa các cột

* Glucose Outcome 0.466 Tương quan dương khá mạnh → Glucose cao thường liên quan đến nguy cơ tiểu đường
* BMI Outcome 0.19 Tương quan dương vừa phải → Chỉ số BMI cao có thể liên quan đến tiểu đường
* Age Pregnancies 0.544 Tương quan dương mạnh → Người lớn tuổi thường có nhiều lần mang thai hơn
* SkinThickness BMI 0.392 Tương quan dương vừa phải → Da dày hơn thường đi kèm với BMI cao



* Gần 1 → quan hệ tuyến tính dương mạnh (ví dụ: Glucose ↔ Outcome).
* Gần -1 → quan hệ tuyến tính âm mạnh.
* Gần 0 → gần như không liên hệ.

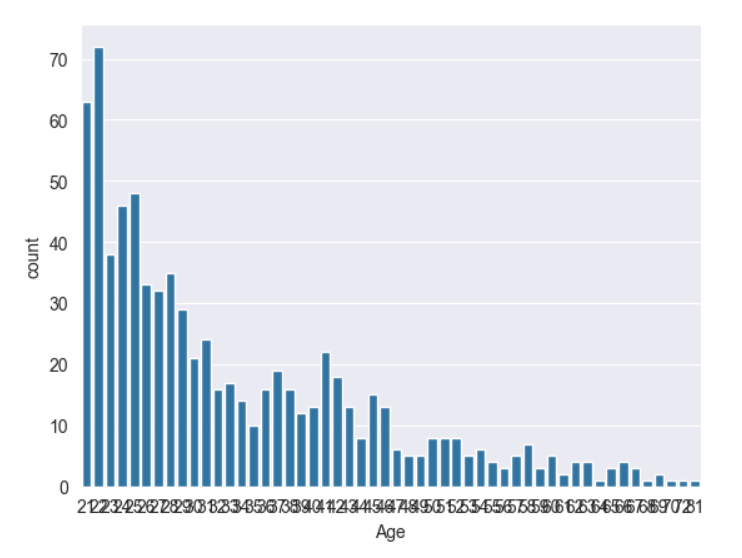
### 1.7 Phân phối độ tuổi

****

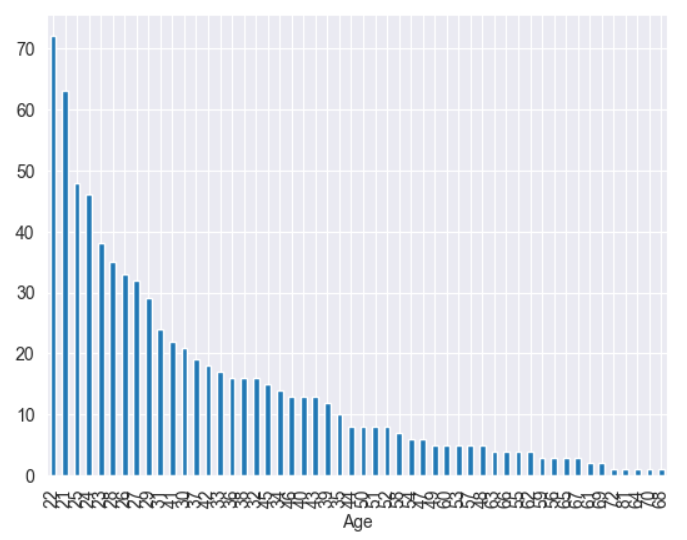
## 2. EDA trực quan bằng sơ đồ (Graphical EDA)

### 2.1 Phân phối độ tuổi bằng sơ đồ

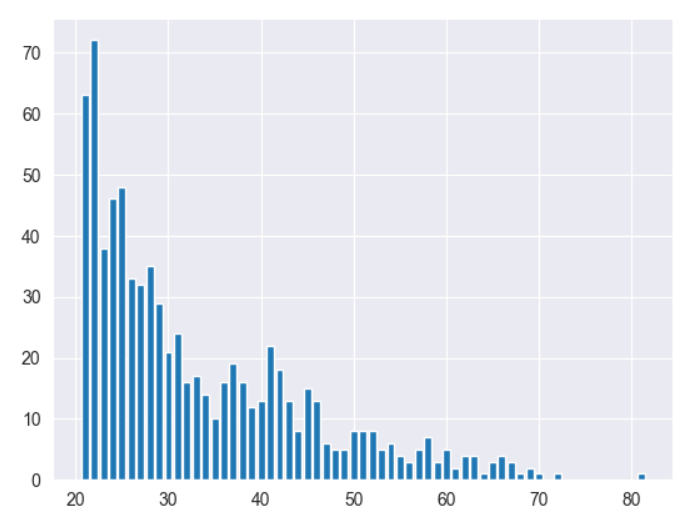
#### A. Seaborn

****

#### B. Pandas



#### C. Matplotlib

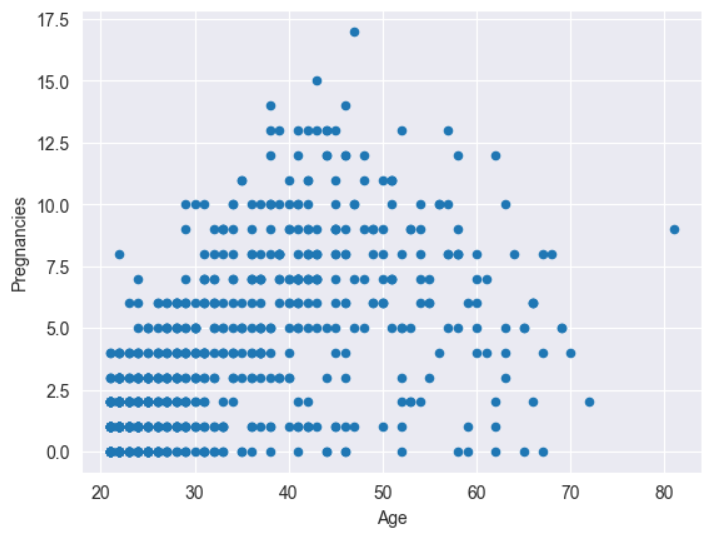


### 2.2 Kết luận về độ tuổi

* Độ tuổi không cân bằng
* Không phù hợp với biểu đồ tròn, biểu đồ cột vì dữ liệu lớn

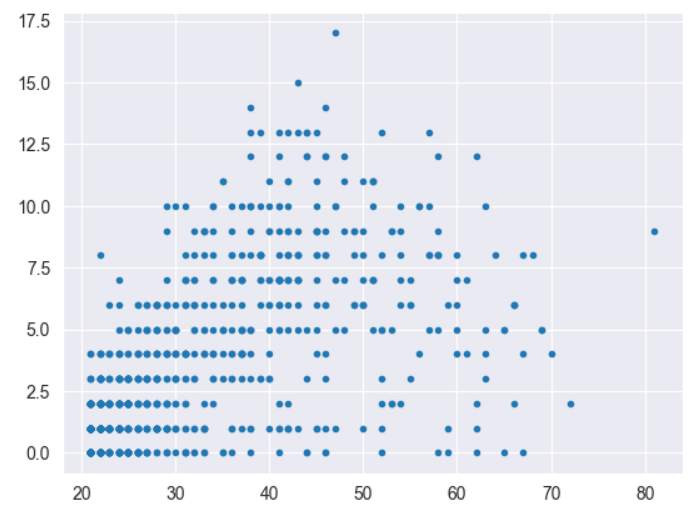
## 3. Biểu đồ so sánh (Số tuổi và số lần mang thai)

### 3.1 Biểu đồ phân tán sử dụng pandas

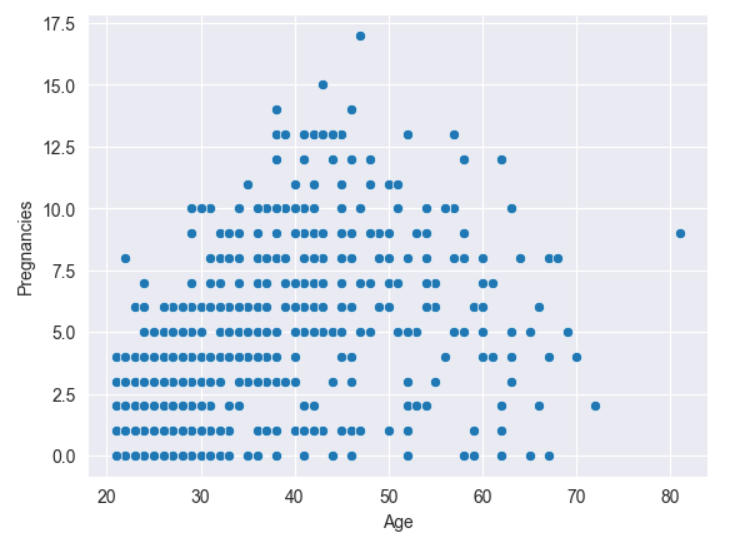


### 3.2 Biểu đồ phân tán sử dụng matplotlib





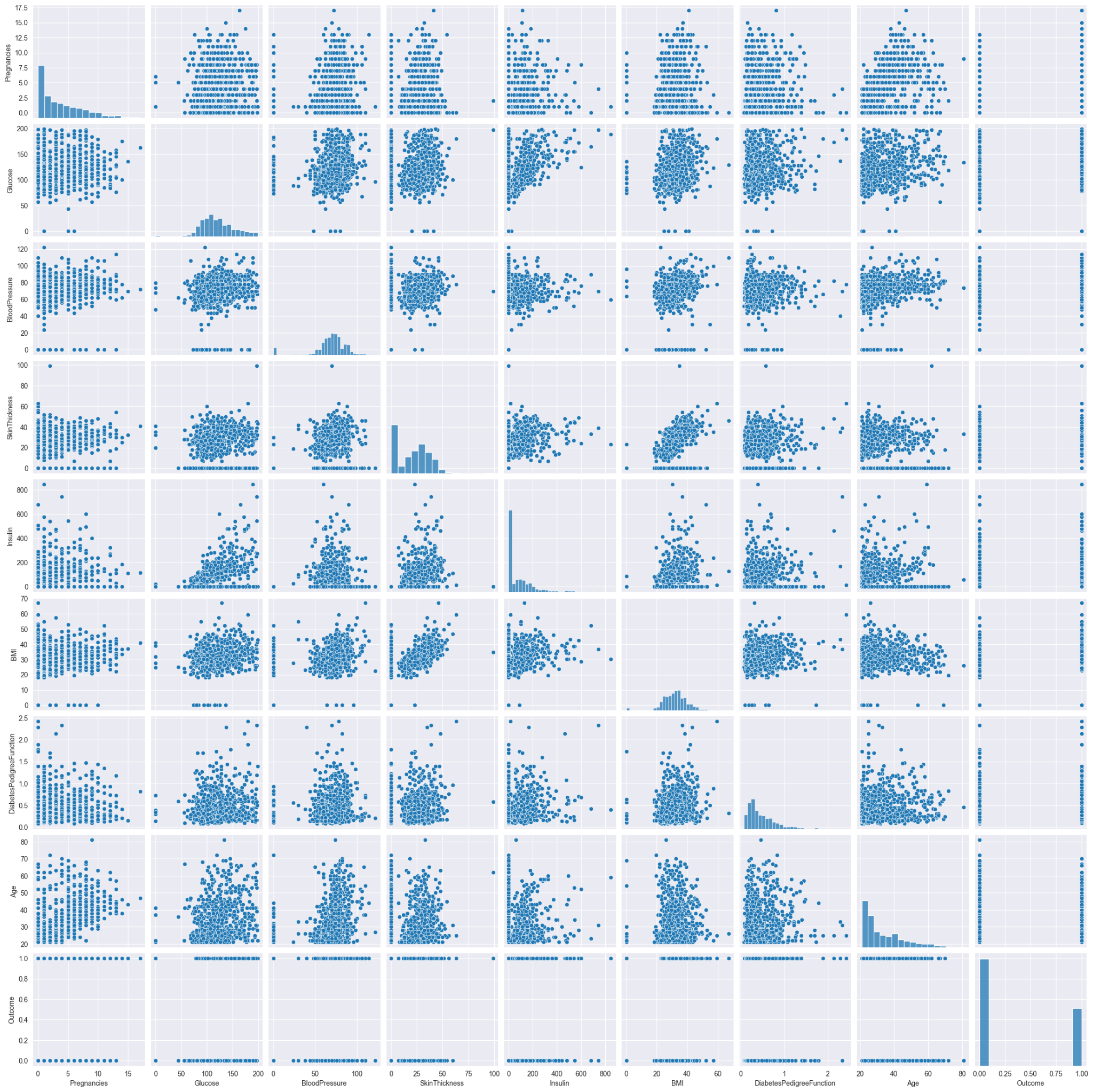
### 3.3 Biểu đồ phân tán sử dụng seaborn

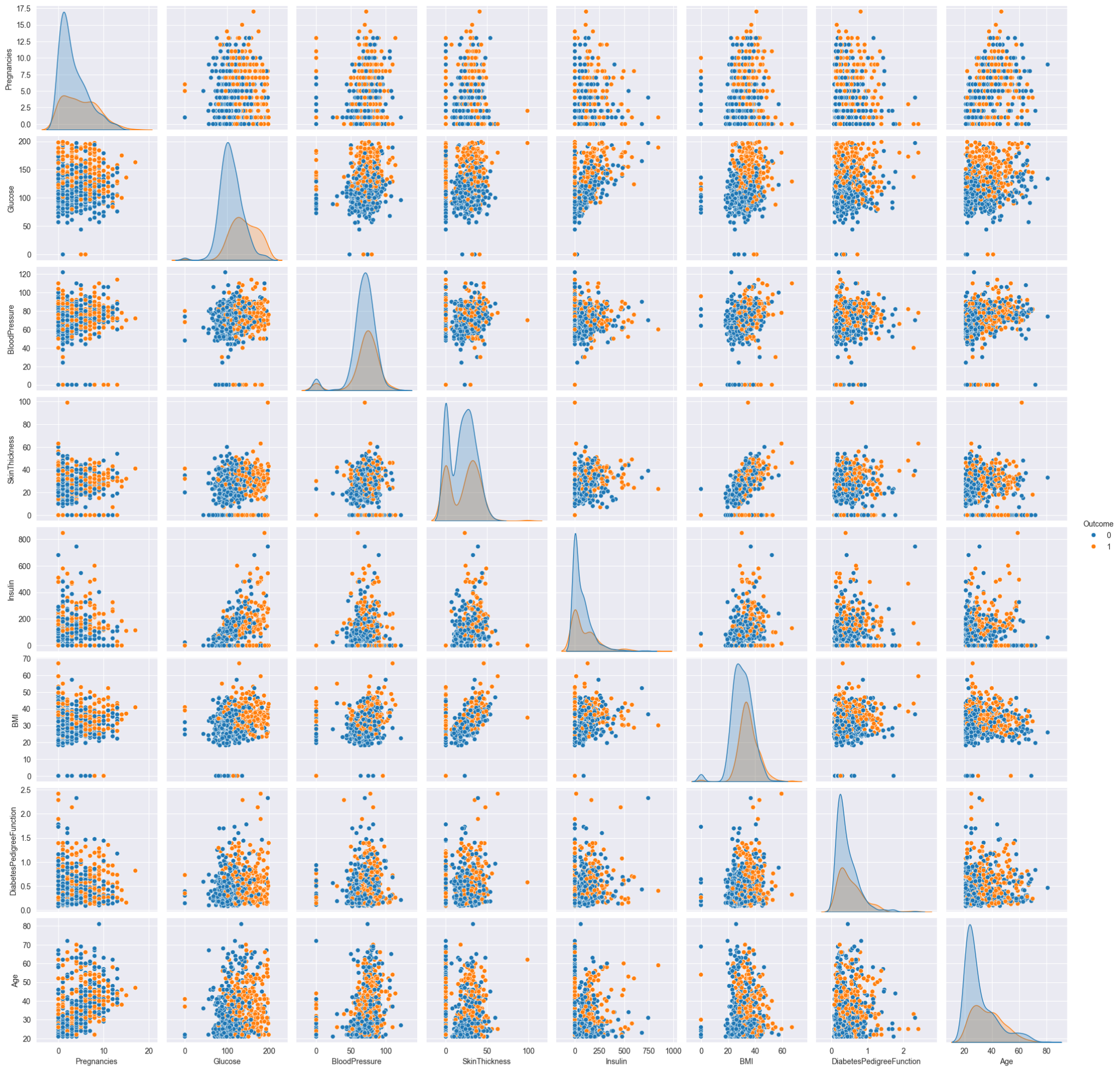




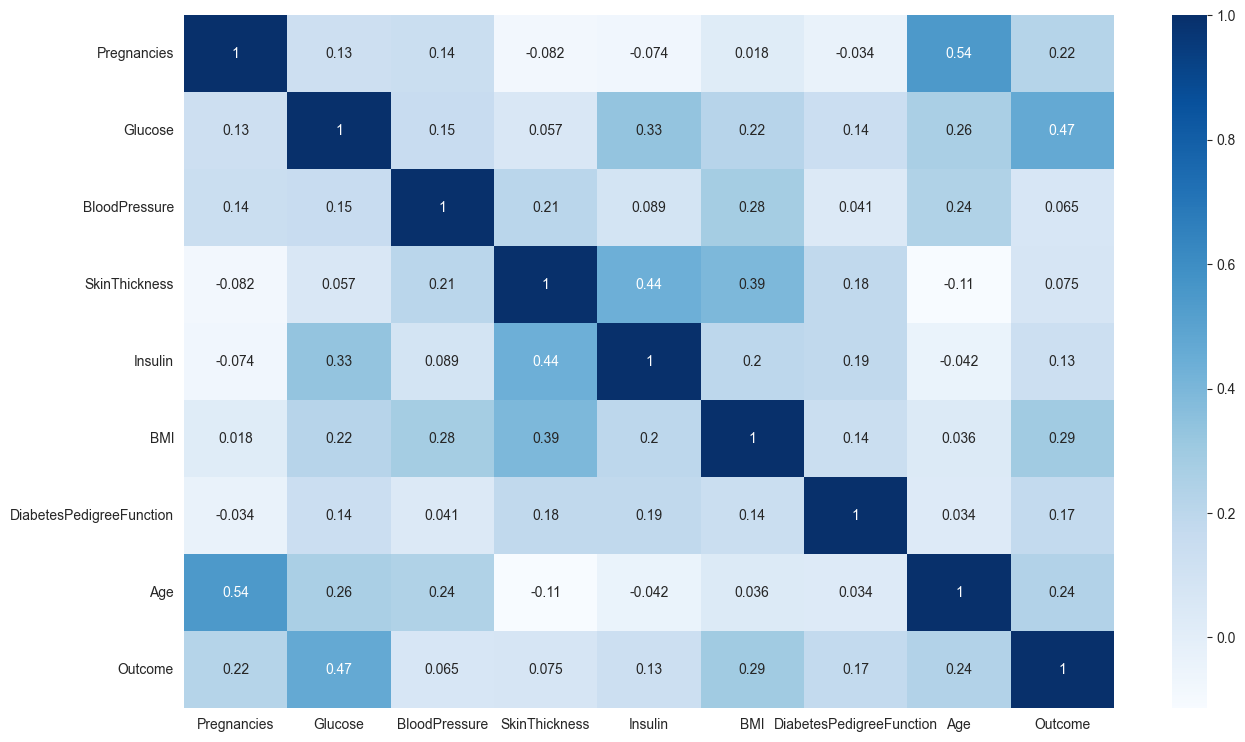
## 4. Biểu đồ trực quan hóa ma trận tương quan (Correlogram)

* Màu đậm → tương quan mạnh
* Màu nhạt → tương quan yếu.

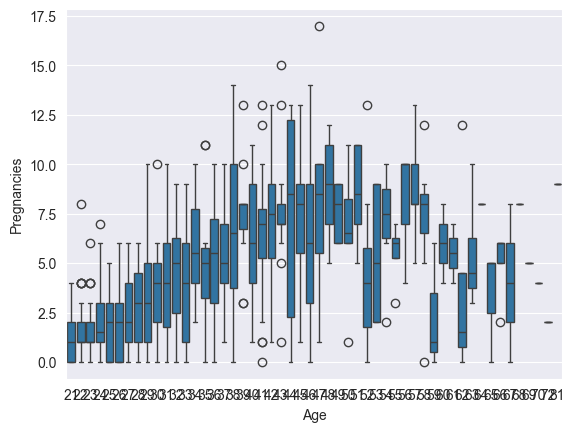




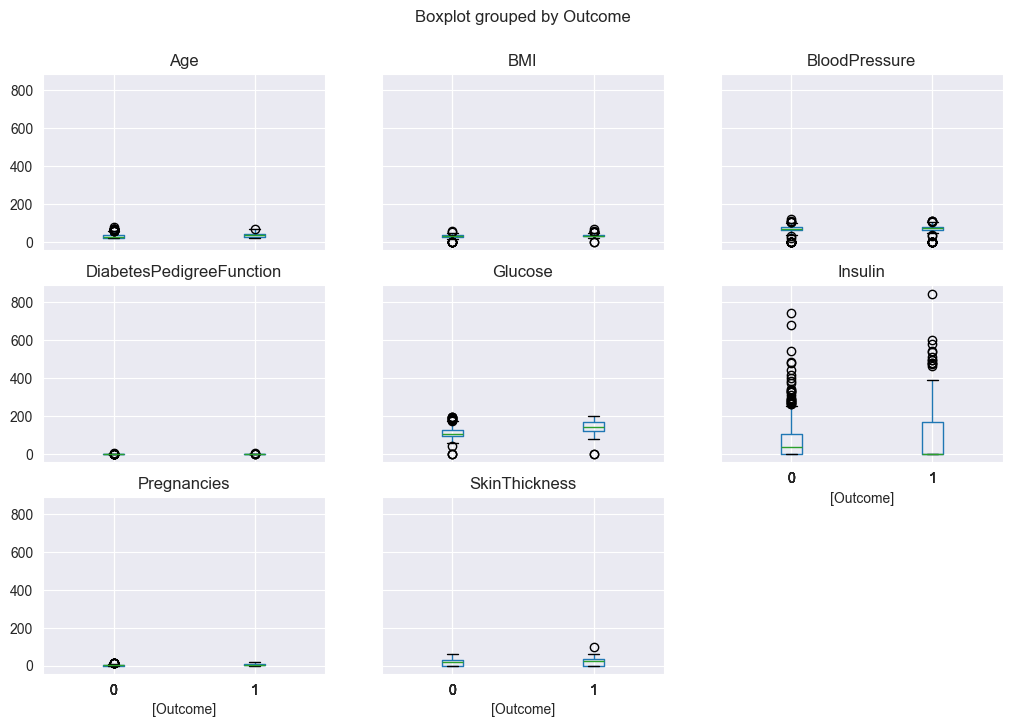
## 5. Heatmap



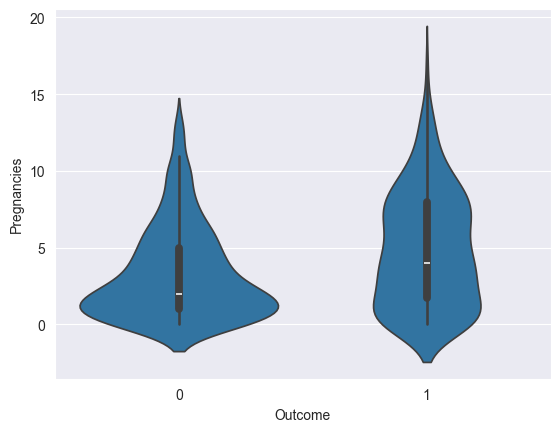
### 5.1 Biểu đồ hộp (Box plot)



### 



### 5.2 Biểu đồ hình đàn vĩ cầm (Violin plot)



# III. KẾT LUẬN

- Phân bố tuổi (Age): Dữ liệu tập trung chủ yếu trong khoảng 20–60 tuổi. Sau 60 tuổi số lượng điểm rất ít, nghĩa là mẫu khảo sát cho người lớn tuổi ít hơn.

- Phân bố số lần mang thai (Pregnancies): Hầu hết các giá trị nằm trong khoảng 0–10 lần mang thai, nhưng có một số trường hợp ngoại lệ đến tận 17 lần.

- So sánh Outcome (0 và 1):

* Outcome = 0 (màu xanh): có nhiều điểm hơn ở mức số lần mang thai thấp (0–5 lần).
* Outcome = 1 (màu cam): xuất hiện nhiều ở tuổi trung niên (30–50 tuổi) và thường có số lần mang thai từ 2–10.

- Xu hướng tổng quát: Người nhiều lần mang thai thường rơi vào nhóm Outcome = 1 nhiều hơn so với nhóm ít lần mang thai. Tuy nhiên, cũng có khá nhiều người Outcome = 0 ngay cả khi mang thai nhiều lần → nghĩa là không có mối quan hệ tuyến tính rõ rệt, chỉ là có xu hướng.

# 

# TÀI LIỆU THAM KHẢO

[1] "Pima Indians Diabetes Database," Kaggle, UCI Machine Learning Repository. https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database. [Truy cập: 24-09-2025].

# PHÂN CÔNG CÔNG VIỆC

|  |  |
| --- | --- |
| Lê Nguyễn Quốc Hưng | Code (xử lý dữ liệu, lập trình) |
| Trần Bá Tiên | Viết báo cáo word |
| Dương Huỳnh Gia | Thiết kế slide thuyết trình |
| Bùi Thành Công | Thiết kế slide thuyết trình |