

# 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!

# 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 |

# MỤC LỤC

[LỜI CẢM ƠN 1](#_heading=h.30j0zll)

[PHÂN CÔNG CÔNG VIỆC 2](#_heading=h.u03pprdrw9xk)

[MỤC LỤC 3](#_heading=h.os4kp4kuwm50)

[I. GIỚI THIỆU 5](#_heading=h.ua56i0g34v20)

[1. Cập nhật về Chẩn đoán và Phân loại Bệnh Đái tháo đường 5](#_heading=h.ud2dnqgopcnv)

[2. Định nghĩa và tiêu chí chẩn đoán bệnh đái tháo đường và các thể rối loạn dung nạp glucose khác 5](#_heading=h.g69lb3xqu84m)

[2.1 Định nghĩa 5](#_heading=h.i06fafv1j621)

[2.2 Chẩn đoán và tiêu chí chẩn đoán 6](#_heading=h.jjrdn4wo9pl0)

[2.3 Tiêu chí chẩn đoán 7](#_heading=h.vv2z9wpsuthl)

[3. Phân loại 8](#_heading=h.656h8qcw42j0)

[3.1 Các phân loại trước đây 8](#_heading=h.3t747j5omxqp)

[3.2 Phân loại cập nhật 9](#_heading=h.9hd0o3xzrkdi)

[3.3 Thuật ngữ 9](#_heading=h.tjjfps4tl029)

[4. Phân giai đoạn lâm sàng của bệnh đái tháo đường và các loại dung nạp glucose khác 10](#_heading=h.q9v6zp8eympo)

[5. Nguyên nhân bệnh 11](#_heading=h.xl4l6oh39y4l)

[6. Đái tháo đường thai kỳ và tăng đường huyết thai kỳ 12](#_heading=h.yazawu66dkyi)

[II. ĐỊNH NGHĨA VẤN ĐỀ (DEFINE PROBLEM) 12](#_heading=h.2mocjcycxqnx)

[III. CHUẨN BỊ VẤN ĐỀ (PREPARE PROBLEM) 13](#_heading=h.b2e9p9q2tsck)

[IV. PHÂN TÍCH 14](#_heading=h.ljbasi8muu8p)

[1. EDA không dùng biểu đồ (Non-graphie EDA) 14](#_heading=h.9t7utas2xgc2)

[1.1 Dataset 14](#_heading=h.5mt06wsad10)

[1.2 In 5 dòng data đầu 15](#_heading=h.vlulr19azqcn)

[1.3 Kiểm tra thông tin khái quát 15](#_heading=h.58n6hafekji3)

[1.4 Kiểm tra giá trị Null/NaN cho từng cột 16](#_heading=h.gdxo6sowxae2)

[1.5 Thống kê mô tả cho từng cột 16](#_heading=h.g3vqrf94p12c)

[1.6 Mối tương quan giữa các cột 16](#_heading=h.gzg6hhixo64k)

[1.7 Phân phối độ tuổi 18](#_heading=h.vnqpb7ri6xqy)

[2. EDA trực quan bằng sơ đồ (Graphical EDA) 18](#_heading=h.xoyeki98hkwh)

[2.1 Phân phối độ tuổi bằng sơ đồ 18](#_heading=h.6i1d1g32uizt)

[A. Seaborn 19](#_heading=h.tuumyzjxaa2u)

[B. Pandas 19](#_heading=h.q2qai333uur5)

[C. Matplotlib 20](#_heading=h.xvmrrodjgi6p)

[2.2 Kết luận về độ tuổi 20](#_heading=h.ovz81ehun4xs)

[3. Biểu đồ so sánh (Số tuổi và số lần mang thai) 21](#_heading=h.nf2tjudfn06e)

[3.1 Biểu đồ phân tán sử dụng pandas 21](#_heading=h.385of2o8pv3p)

[3.2 Biểu đồ phân tán sử dụng matplotlib 21](#_heading=h.wgnnmkk1tach)

[3.3 Biểu đồ phân tán sử dụng seaborn 23](#_heading=h.qgajiwemkgkx)

[4. Biểu đồ trực quan hóa ma trận tương quan (Correlogram) 24](#_heading=h.ohej8729mhi7)

[5. Heatmap 26](#_heading=h.h5blzlt8vkto)

[5.1 Biểu đồ hộp (Box plot) 27](#_heading=h.hlyb92746vuh)

[5.2 Biểu đồ hình đàn vĩ cầm (Violin plot) 28](#_heading=h.94tz9givbapc)

[III. KẾT LUẬN 29](#_heading=h.2srdqcqlhx5)

[TÀI LIỆU THAM KHẢO 30](#_heading=h.bd6z3290hgfj)

# I. GIỚI THIỆU

## 1. Cập nhật về Chẩn đoán và Phân loại Bệnh Đái tháo đường

* Những năm 1970: WHO và NDDG đã giới thiệu các tiêu chí chẩn đoán mới. Giải quyết sự nhầm lẫn trong thuật ngữ và tiêu chuẩn xét nghiệm
* Năm 1985: WHO sửa đổi tiêu chí để phù hợp hơn với NDDG
* 1996–1997: WHO và ADA triệu tập các nhóm chuyên gia nhằm cập nhật phân loại, tinh chỉnh tiêu chí và bao gồm định nghĩa Hội chứng chuyển hóa

Kết quả:

- WHO và ADA đạt được những kết luận tương đồng

- WHO phát hành báo cáo sau khi thu thập ý kiến phản hồi quốc tế

## 2. Định nghĩa và tiêu chí chẩn đoán bệnh đái tháo đường và các thể rối loạn dung nạp glucose khác

### 2.1 Định nghĩa

Một rối loạn chuyển hóa do nhiều nguyên nhân, đặc trưng bởi tình trạng tăng đường huyết mạn tính do rối loạn tiết insulin, rối loạn tác động của insulin, hoặc cả hai

* Tác động chuyển hóa: Rối loạn chuyển hóa carbohydrate, chất béo và protein.
* Triệu chứng (có thể rõ rệt hoặc âm thầm): Khát nước, tiểu nhiều, nhìn mờ, sụt cân. Trường hợp nặng: nhiễm toan ceton hoặc tình trạng tăng áp lực thẩm thấu → hôn mê hoặc tử vong nếu không được điều trị
* Nguy cơ chẩn đoán muộn: Tăng đường huyết có thể gây tổn thương trước khi các triệu chứng xuất hiện.
* Biến chứng lâu dài: Mắt:
* Bệnh võng mạc → mù lòa.
* Thận: Bệnh thận do đái tháo đường → suy thận.
* Thần kinh: Bệnh thần kinh → loét, cắt cụt chi, khớp Charcot, rối loạn thần kinh tự động và chức năng tình dục
* Mạch máu: Tăng nguy cơ bệnh tim mạch, mạch máu não và mạch ngoại biên
* Cơ chế bệnh sinh:
* Hủy hoại tế bào beta → thiếu insulin
* Đề kháng insulin → giảm tác dụng của insulin trên các mô đích

### 2.2 Chẩn đoán và tiêu chí chẩn đoán

* Chẩn đoán phải chắc chắn: Do hậu quả kéo dài suốt đời, các bác sĩ lâm sàng cần phải chắc chắn trong việc chẩn đoán.
* Có triệu chứng và không triệu chứng:
* Triệu chứng rõ rệt + tăng đường huyết rõ ràng → chẩn đoán gần như chắc chắn
* Bệnh nhân không triệu chứng: Không bao giờ chẩn đoán chỉ dựa trên một giá trị glucose bất thường duy nhất
* Cần xác nhận: Với người không có triệu chứng: Cần ít nhất một xét nghiệm khác trong ngưỡng đái tháo đường (đường huyết lúc đói, ngẫu nhiên hoặc nghiệm pháp dung nạp glucose OGTT)
* Tăng đường huyết tạm thời ≠ đái tháo đường: Tăng đường huyết liên quan đến stress (ví dụ: nhiễm trùng, chấn thương) có thể chỉ là tạm thời và không đủ để chẩn đoán
* Theo dõi liên tục: Nếu chưa thể chẩn đoán chắc chắn, cần xét nghiệm định kỳ và theo dõi các yếu tố nguy cơ
* Các yếu tố cần cân nhắc thêm: Chủng tộc, tiền sử gia đình, tuổi, béo phì, bệnh lý kèm theo cần được xem xét trong đánh giá lâm sàng
* Hemoglobin glycated (HbA1c): Phản ánh mức đường huyết trung bình trong vài tuần: Đầy hứa hẹn, nhưng chưa được khuyến cáo sử dụng chính thức do vấn đề chuẩn hóa và giới hạn về khả năng tiếp cận

\*Bệnh đái tháo đường ở trẻ em

* Biểu hiện điển hình:
* Triệu chứng nghiêm trọng, đường huyết rất cao
* Thường gặp tình trạng glucose niệu (glycosuria) và xeton niệu (ketonuria)
* Chẩn đoán thường nhanh chóng và rõ ràng
* Điều trị ngay lập tức:
* Liệu pháp insulin thường được bắt đầu ngay
* Có thể cứu sống trong các trường hợp cấp tính
* Không cần làm nghiệm pháp dung nạp glucose (OGTT) nếu biểu hiện rõ ràng
* Trường hợp nhẹ hơn:
* Một số trẻ em/thanh thiếu niên có triệu chứng nhẹ hơn
* Có thể cần làm xét nghiệm đường huyết lúc đói và/hoặc OGTT để xác nhận chẩn đoán

### 2.3 Tiêu chí chẩn đoán

**Biểu hiện lâm sàng và xét nghiệm**

Triệu chứng phổ biến:

– Khát nước nhiều, tiểu nhiều

– Sụt cân không rõ nguyên nhân

– Nhiễm trùng tái phát

– Buồn ngủ, hôn mê

Chẩn đoán được xác nhận khi có một kết quả đo đường huyết vượt ngưỡng chẩn đoán.

Nghiệm pháp dung nạp glucose (OGTT) 75g chỉ cần thực hiện khi đường huyết lúc đói hoặc đường huyết ngẫu nhiên nằm ở ngưỡng ranh giới.

**Giá trị ngưỡng chẩn đoán mới**

* Đường huyết lúc đói trong huyết tương: giảm từ 7,8 xuống còn 7,0 mmol/L (126 mg/dL).
* Đường huyết lúc đói trong máu toàn phần: giảm từ 6,7 xuống còn 6,1 mmol/L (110 mg/dL).
* Lý do: Để phù hợp hơn với giá trị đường huyết sau 2 giờ nạp glucose và có mối liên hệ chặt chẽ hơn với các biến chứng.

**Các nghiên cứu dịch tễ học**

* Phân loại dựa trên đường huyết lúc đói và/hoặc sau 2 giờ nạp glucose.
* Nghiệm pháp dung nạp glucose 2 giờ (OGTT) được ưu tiên sử dụng trong các nghiên cứu dân số.
* Một số nhóm dân số (ví dụ: người cao tuổi, người gầy, người châu Á) có thể bị bỏ sót nếu chỉ dùng xét nghiệm đường huyết lúc đói.

**Chẩn đoán cá nhân**

* Không nên dựa vào chỉ một xét nghiệm đường huyết duy nhất.
* Cần có triệu chứng và/hoặc xác nhận bằng các xét nghiệm lặp lại.
* Lý tưởng là sử dụng cả giá trị đường huyết lúc đói và sau 2 giờ OGTT để phát hiện sớm những người có nguy cơ cao.
* Điều này khác với phương pháp của ADA, ưu tiên sử dụng đường huyết lúc đói.

### 3. Phân loại

#### 3.1 Các phân loại trước đây

1980 WHO giới thiệu:

* IDDM (Bệnh Đái tháo đường phụ thuộc insulin – Loại 1)
* NIDDM (Bệnh Đái tháo đường không phụ thuộc insulin – Loại 2)

1985 WHO:

* Loại bỏ nhãn “Loại 1/2”
* Thêm MRDM (Đái tháo đường liên quan đến suy dinh dưỡng)
* Bao gồm IGT (Rối loạn dung nạp glucose), GDM (Đái tháo đường thai kỳ) và các loại khác

Ứng dụng

* Các giai đoạn phản ánh mức độ nghiêm trọng của tăng đường huyết và tiến triển của bệnh.
* Nguyên nhân có thể được xác định sớm (ví dụ: các dấu ấn tự miễn trong trường hợp đường huyết bình thường).
* Nguyên nhân của loại 2 chưa được xác định rõ ràng nhưng dự kiến sẽ được cải thiện.
* Một số bệnh nhân có thể đảo ngược hoặc kiểm soát bệnh đái tháo đường bằng thay đổi lối sống hoặc thuốc uống.
* Những bệnh nhân khác cần insulin, tùy thuộc vào chức năng tế bào beta.
* Tình trạng chuyển hóa có thể cải thiện, xấu đi hoặc duy trì ổn định theo thời gian.

#### 3.2 Phân loại cập nhật

* Đặc điểm chính
* Kết hợp các giai đoạn lâm sàng và các loại nguyên nhân.
* Nhận diện đái tháo đường là một bệnh tiến triển; bệnh nhân có thể chuyển đổi giữa các giai đoạn.
* Có thể phân giai đoạn lâm sàng ngay cả khi chưa biết rõ nguyên nhân.

#### 3.3 Thuật ngữ

Những thuật ngữ không nên dùng: IDDM” và “NIDDM” gây nhầm lẫn giữa chẩn đoán và điều trị.

Những thuật ngữ được khuyến cáo sử dụng

* Đái tháo đường loại 1:
* Hủy hoại tế bào beta do tự miễn hoặc không rõ nguyên nhân
* Dễ dẫn đến nhiễm toan ceton
* Loại trừ các nguyên nhân cụ thể đã biết (ví dụ: xơ nang)
* Có thể phát hiện trước khi bệnh đái tháo đường hoàn chỉnh
* Đái tháo đường loại 2:
* Rối loạn tiết insulin kết hợp với đề kháng insulin
* Dạng phổ biến ở người lớn
* Một số kiểu hình gầy (ví dụ: nhóm dân số Ấn Độ)

Các nhóm phân loại bị loại bỏ:

* Đái tháo đường liên quan đến suy dinh dưỡng (MRDM) đã bị loại bỏ
* PDPD = đái tháo đường biến đổi do suy dinh dưỡng (cần nghiên cứu thêm)
* FCPD = hiện được gọi là bệnh tụy ngoại tiết

Các phân loại mới:

* Rối loạn dung nạp glucose (IGT) = giai đoạn rối loạn điều hòa, chưa phải đái tháo đường
* Rối loạn đường huyết lúc đói (IFG) = đường huyết lúc đói cao hơn bình thường nhưng chưa đủ tiêu chuẩn đái tháo đường
* Đái tháo đường thai kỳ (GDM) bao gồm cả GIGT và GDM trước đây

### 4. Phân giai đoạn lâm sàng của bệnh đái tháo đường và các loại dung nạp glucose khác

1. **Phân chia bệnh đái tháo đường theo nhu cầu insulin**

* Cần insulin để duy trì sự sống:
* Trước đây gọi là “IDDM”
* Thiếu C-peptide, thiếu insulin tuyệt đối
* Cần insulin để kiểm soát bệnh:
* Vẫn còn một phần insulin nội sinh
* Insulin cần để kiểm soát chuyển hóa, không phải để duy trì sự sống
* Không cần insulin:
* Kiểm soát được bằng lối sống hoặc thuốc uống

1. **Rối loạn điều hòa glucose (IGT & IFG)**

* Trạng thái chuyển tiếp về chuyển hóa giữa bình thường và đái tháo đường
* IGT: Rối loạn dung nạp glucose sau ăn
* IFG: Đường huyết lúc đói tăng nhưng chưa đạt ngưỡng đái tháo đường
* Cả hai đều làm tăng nguy cơ đái tháo đường và bệnh tim mạch
* IGT thường liên quan đến Hội chứng chuyển hóa
* Khuyến cáo làm nghiệm pháp dung nạp glucose (OGTT) cho người có IFG để loại trừ đái tháo đường

1. **Đường huyết bình thường**

* Định nghĩa: Đường huyết lúc đói < 6,1 mmol/L (110 mg/dL) được coi là bình thường.
* Ngoại lệ: Một số người có đường huyết lúc đói bình thường nhưng vẫn có rối loạn dung nạp glucose (IGT).
* Nguy cơ: Giá trị cao hơn mức bình thường làm tăng nguy cơ biến chứng mạch máu.
* Tiến triển bệnh: Những thay đổi bệnh lý có thể bắt đầu ngay cả khi người bệnh đang ở mức đường huyết bình thường.

### 5. Nguyên nhân bệnh

1. **Đái tháo đường tuýp 1**

* Đặc trưng: Do quá trình phá huỷ tế bào beta tuyến tụy, dẫn đến thiếu insulin trầm trọng → cần insulin để sống sót, nếu không dễ bị nhiễm toan ceton, hôn mê và tử vong.
* Diễn tiến: Người bệnh có thể bình thường về chuyển hoá trước khi bệnh biểu hiện, nhưng sự phá huỷ tế bào beta đã xảy ra và có thể phát hiện.
* Chẩn đoán: Thường dựa vào sự hiện diện của các tự kháng thể (anti-GAD, đảo tụy, insulin kháng thể), cho thấy cơ chế tự miễn gây phá hủy tế bào beta.
* Trường hợp đặc biệt: Ở một số người (nhất là không phải người da trắng), không phát hiện bằng chứng rối loạn tự miễn → được gọi là Type 1 vô căn (idiopathic).
* Phân loại căn nguyên: Có thể xác định khi xét nghiệm kháng thể phù hợp, nhưng không phải lúc nào cũng có điều kiện thực hiện. Nếu có xét nghiệm, việc phân loại bệnh nhân cần phản ánh điều này.

1. **Đái tháo đường tuýp 2**

* Là dạng phổ biến nhất.
* Đặc trưng bởi rối loạn tác động của insulin và bài tiết insulin (có thể một hoặc cả hai).
* Thường cả hai rối loạn này cùng tồn tại khi bệnh biểu hiện lâm sàng.
* Nguyên nhân cụ thể gây ra những bất thường này hiện chưa được biết rõ.

1. **Các thể đặc hiệu khác**

* Ít gặp hơn, do các khiếm khuyết hoặc tiến trình bệnh lý đặc thù.
* Có thể xác định rõ nguyên nhân.
* Ví dụ: tiểu đường do xơ hoá tụy (fibrocalceous pancreatopathy) – trước đây từng được xếp vào nhóm tiểu đường liên quan đến suy dinh dưỡng

### 6. Đái tháo đường thai kỳ và tăng đường huyết thai kỳ

* Định nghĩa: Đái tháo đường thai kỳ là tình trạng rối loạn dung nạp carbohydrate gây tăng đường huyết với mức độ khác nhau, khởi phát hoặc được phát hiện lần đầu trong thai kỳ.
* Lưu ý: Không loại trừ khả năng rối loạn dung nạp glucose đã tồn tại trước đó nhưng chưa được chẩn đoán. Định nghĩa này áp dụng bất kể có dùng insulin hay tình trạng còn tồn tại sau sinh hay không.
* Phân biệt: Phụ nữ có tiểu đường từ trước khi mang thai → không gọi là tiểu đường thai kỳ mà là “tiểu đường có sẵn khi mang thai” và cần được điều trị trước, trong và sau thai kỳ.
* Sinh lý thai kỳ:
* Trong giai đoạn sớm (tam cá nguyệt 1 và nửa đầu tam cá nguyệt 2), đường huyết lúc đói và sau ăn thường thấp hơn phụ nữ không mang thai.
* Đường huyết tăng bất thường ở giai đoạn này có thể phản ánh tiểu đường có sẵn trước đó.
* Chưa có tiêu chuẩn rõ ràng để chẩn đoán “cao bất thường” ở giai đoạn này.

# II. ĐỊNH NGHĨA VẤN ĐỀ (DEFINE PROBLEM)

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!

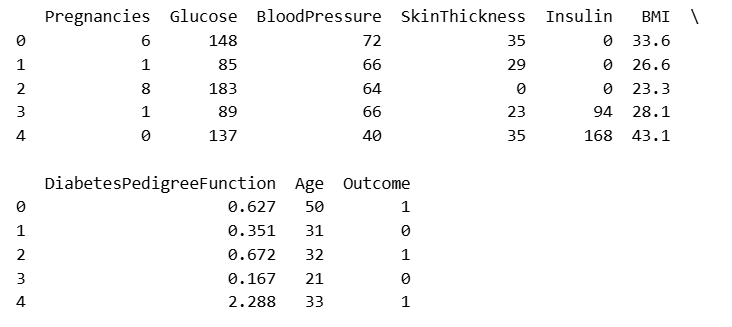
Bộ dữ liệu bao gồm thông tin y tế của các bệnh nhân nữ thuộc cộng đồng người Pima, với các đặc trưng như nồng độ glucose, huyết áp, độ dày lớp da, mức insulin, chỉ số BMI, yếu tố di truyền và độ tuổi. Mỗi bản ghi đều được gắn nhãn cho biết người đó có mắc bệnh đái tháo đường hay không. Tất cả thông tin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

* 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].

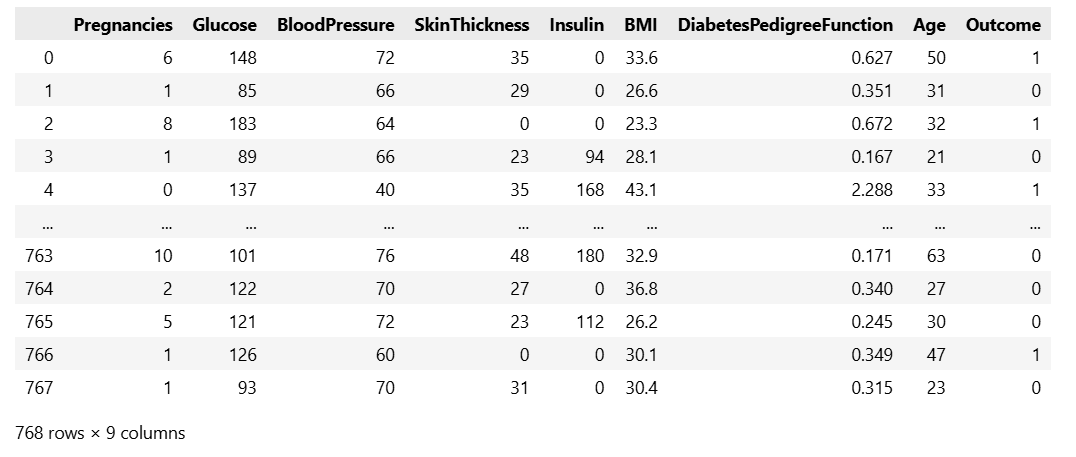
****

*Hình III.1: Bộ dữ liệu*

# IV. PHÂN TÍCH

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

### 1.1 Dataset



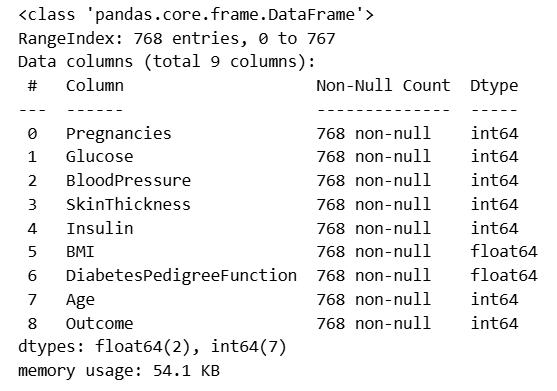
*Hình IV.1.1.1: Hình dataset*

### 1.2 In 5 dòng data đầu



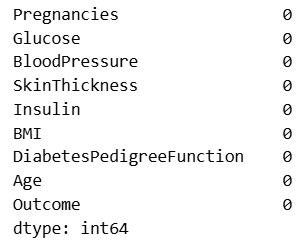
*Hình IV.1.1.2: Hình 5 dòng đầu dữ liệu*

### 1.3 Kiểm tra thông tin khái quát



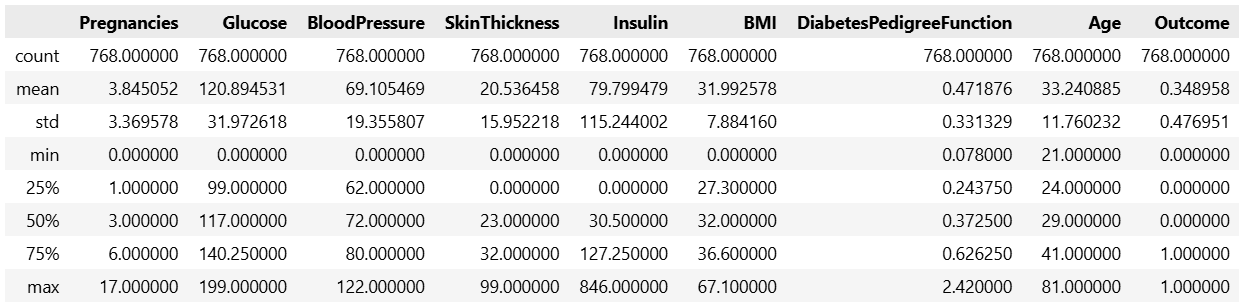
*Hình IV.1.1.3: Hình thông tin khái quát*

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



*Hình IV.1.1.4: Hình kiểm tra giá trị Null/NaN*

### 1.5 Thống kê mô tả cho từng cột

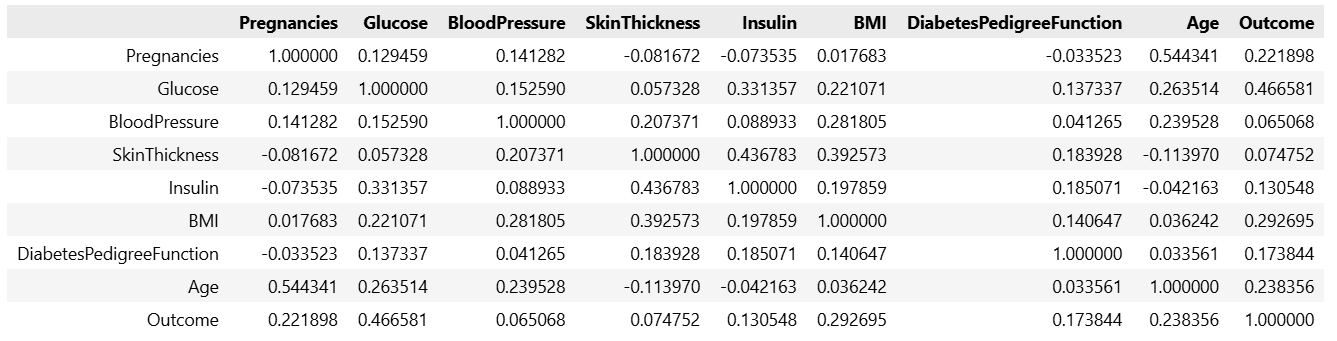


*Hình IV.1.1.5: Hình thống kê mô tả từng cột*

* 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

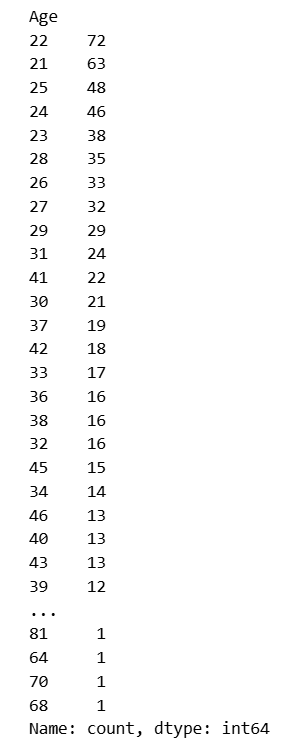
* Glucose Outcome 0.466 Tương quan dương khá mạnh → Glucose cao thường liên quan đến nguy cơ tiểu đường
* BMI Outcome 0.19 Tương quan dương vừa phải → Chỉ số BMI cao có thể liên quan đến tiểu đường
* Age Pregnancies 0.544 Tương quan dương mạnh → Người lớn tuổi thường có nhiều lần mang thai hơn
* SkinThickness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



*Hình IV.1.1.6: Hình mối tương quan các cột*

* 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

****

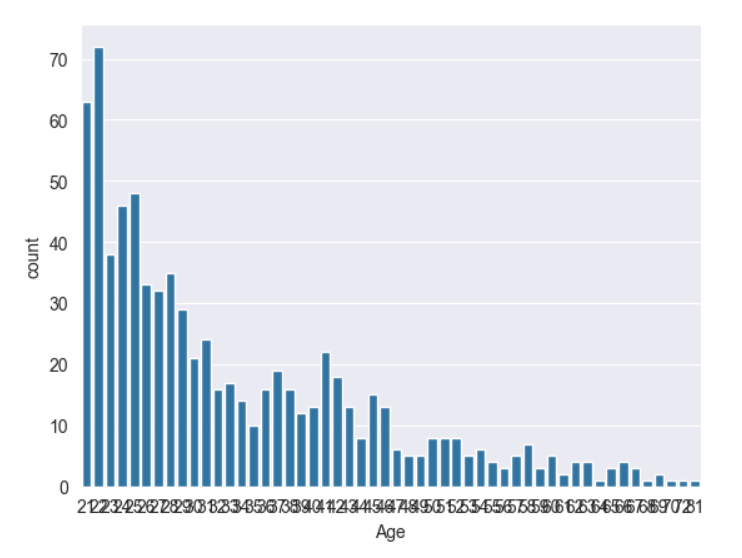
*Hình IV.1.1.7: Hình phân bố độ tuổi*

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

### 2.1 Phân phối độ tuổi bằng sơ đồ

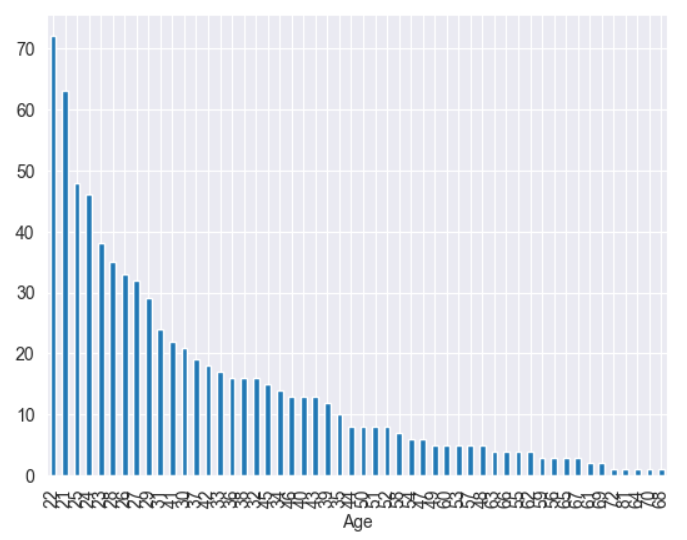
* Phân bố độ tuổi không cân bằng, đa số bệnh nhân từ 21–40 tuổi.
* Nhóm bệnh nhân trên 60 tuổi có rất ít mẫu → dữ liệu hạn chế ở nhóm này.
* Cần lưu ý khi huấn luyện mô hình vì dữ liệu thiên lệch theo tuổi.

#### A. Seaborn

****

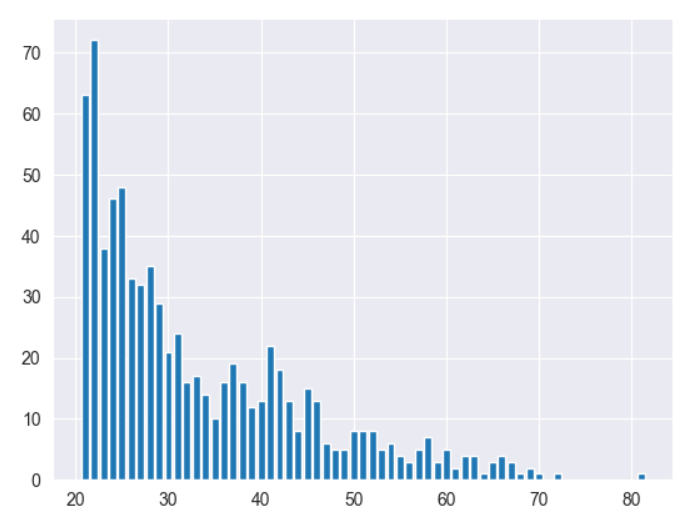
*Hình IV.2.a: Hình sơ đồ phân bố độ tuổi seaborn*

#### B. Pandas



*Hình IV.2.b: Hình sơ đồ phân bố độ tuổi pandas*

#### C. Matplotlib



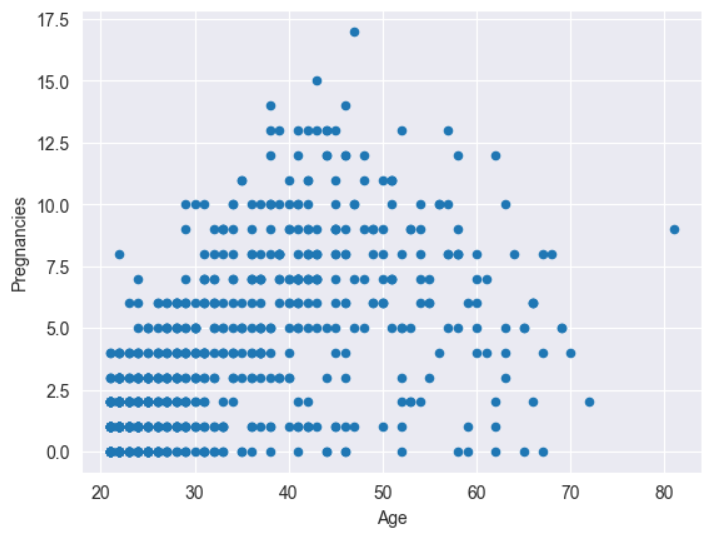
*Hình IV.2.c: Hình sơ đồ phân bố độ tuổi matplotlib*

### 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

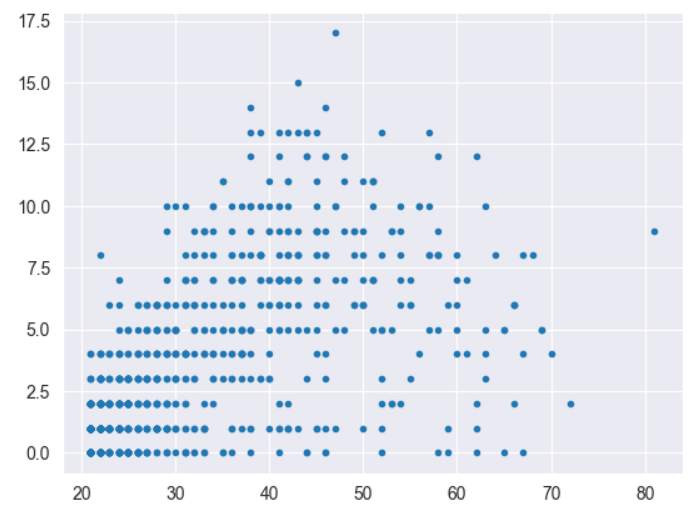


*Hình IV.3.3.1: Hình sơ đồ phân tán sử dụng pandas*

### 3.2 Biểu đồ phân tán sử dụng matplotlib

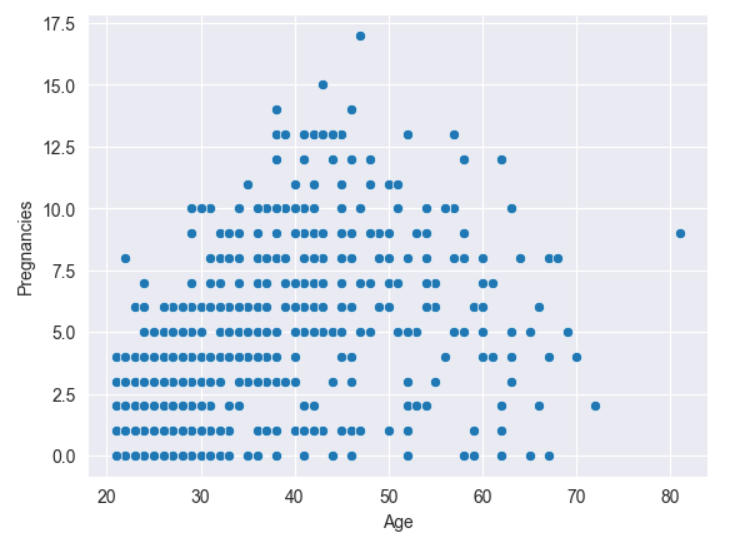


*Hình IV.3.3.2: Hình sơ đồ phân tán sử dụng matplotlib*



*Hình IV.3.3.3: Hình sơ đồ phân tán sử dụng mathplotlib*

### 3.3 Biểu đồ phân tán sử dụng seaborn



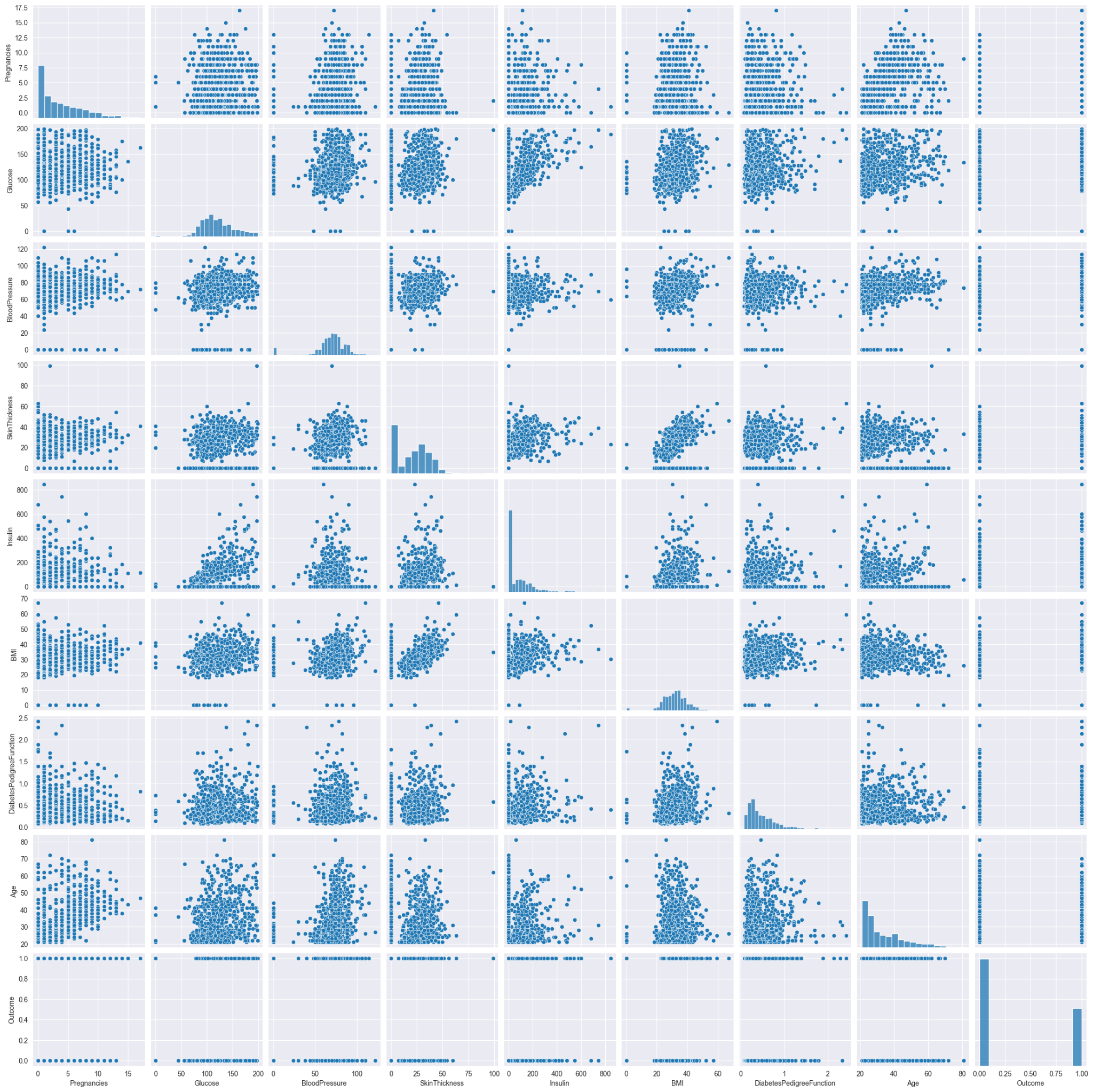
*Hình IV.3.3.4: Hình sơ đồ phân tán sử dụng seaborn*



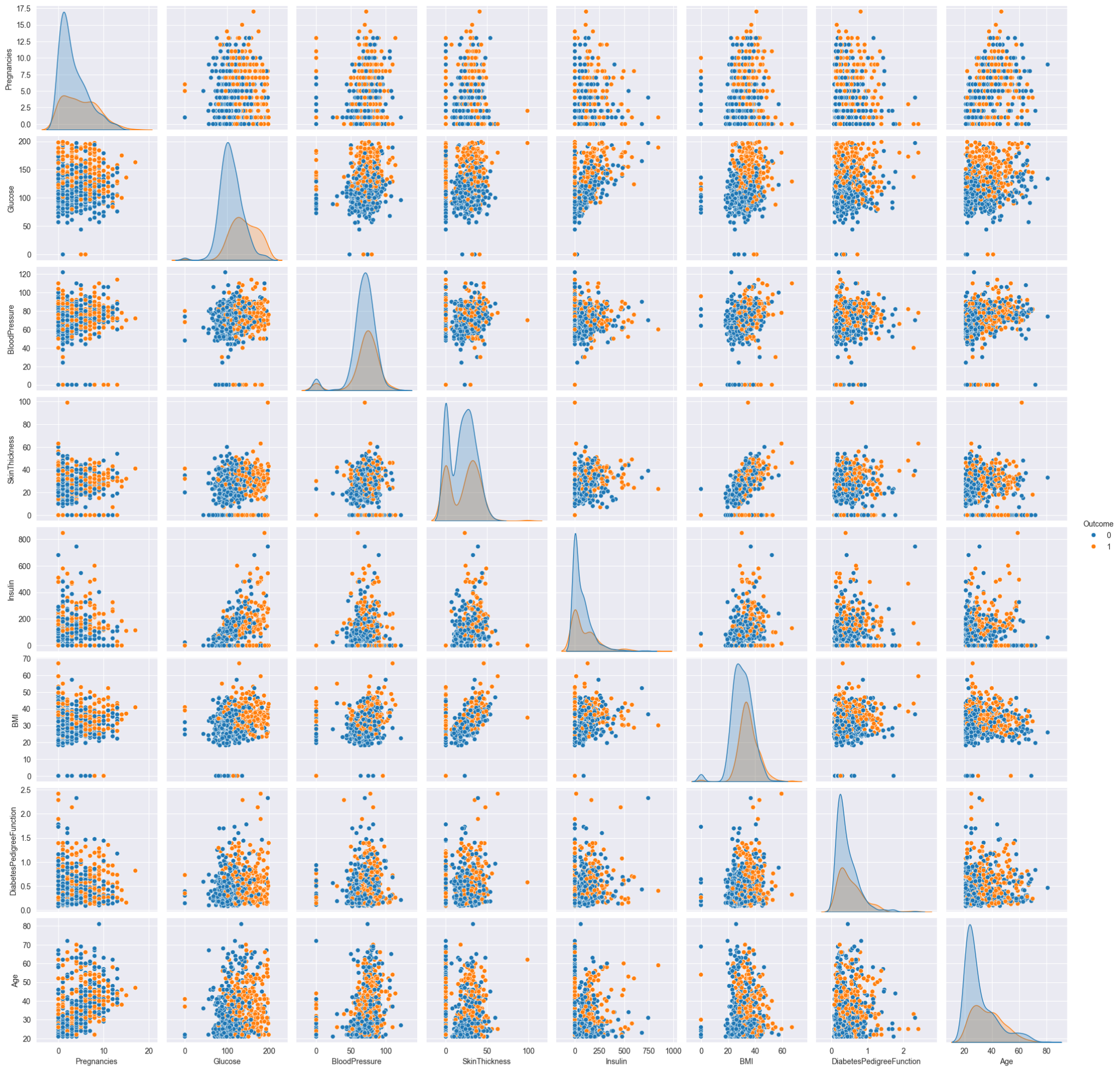
*Hình IV.3.3.5: Hình sơ đồ phân tán sử dụng seaborn*

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

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



*Hình IV.4.1: Hình biểu đồ correlogram*



*Hình IV.4.2: Hình biểu đồ correlogram*

- 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.

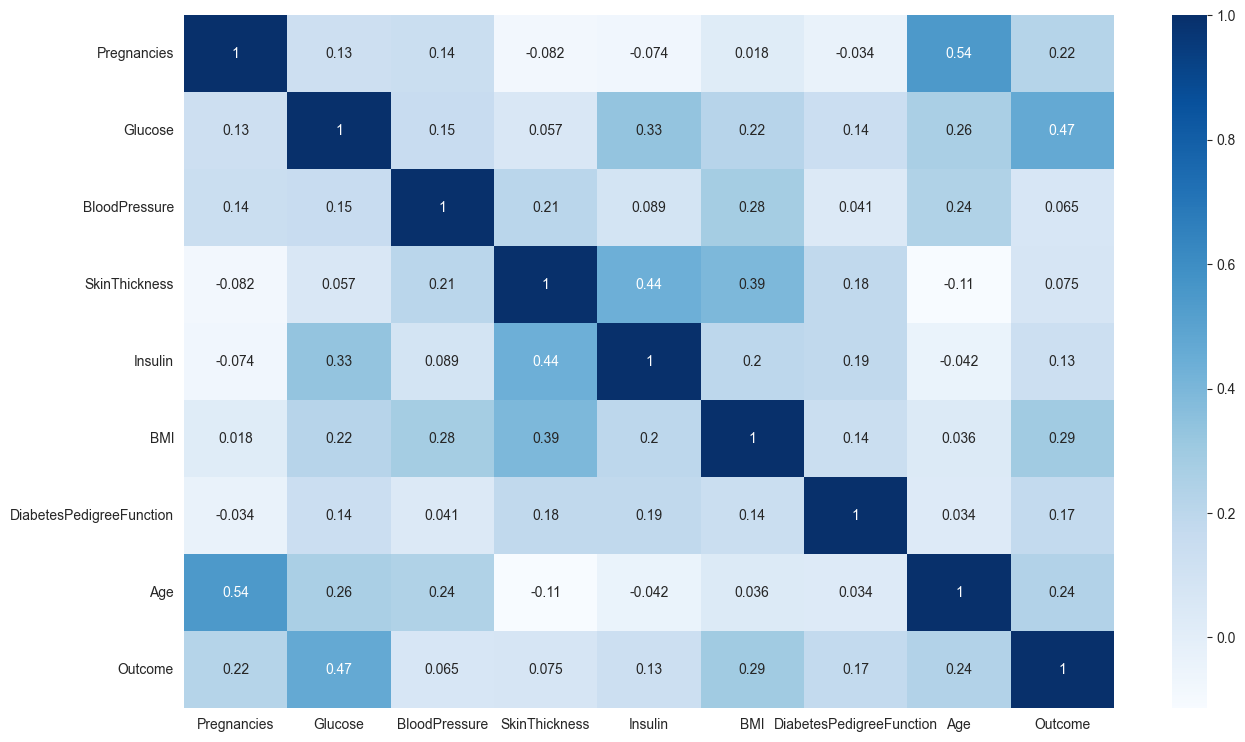
- Phân bố số lần mang thai (Pregnancies): Hầu hết các giá trị nằm trong khoảng 0–10 lần mang thai, nhưng có một số trường hợp ngoại lệ đến tận 17 lần.

- 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.

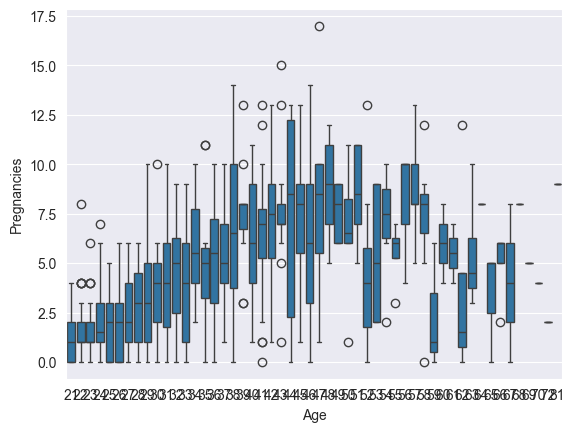
- Xu hướng tổng quát: Người nhiều lần mang thai thường rơi vào nhóm Outcome = 1 nhiều hơn so với nhóm ít lần mang thai. Tuy nhiên, cũng có khá nhiều người Outcome = 0 ngay cả khi mang thai nhiều lần → nghĩa là không có mối quan hệ tuyến tính rõ rệt, chỉ là có xu hướng.

## 5. Heatmap



*Hình IV.5.1: Hình biểu đồ heatmap*

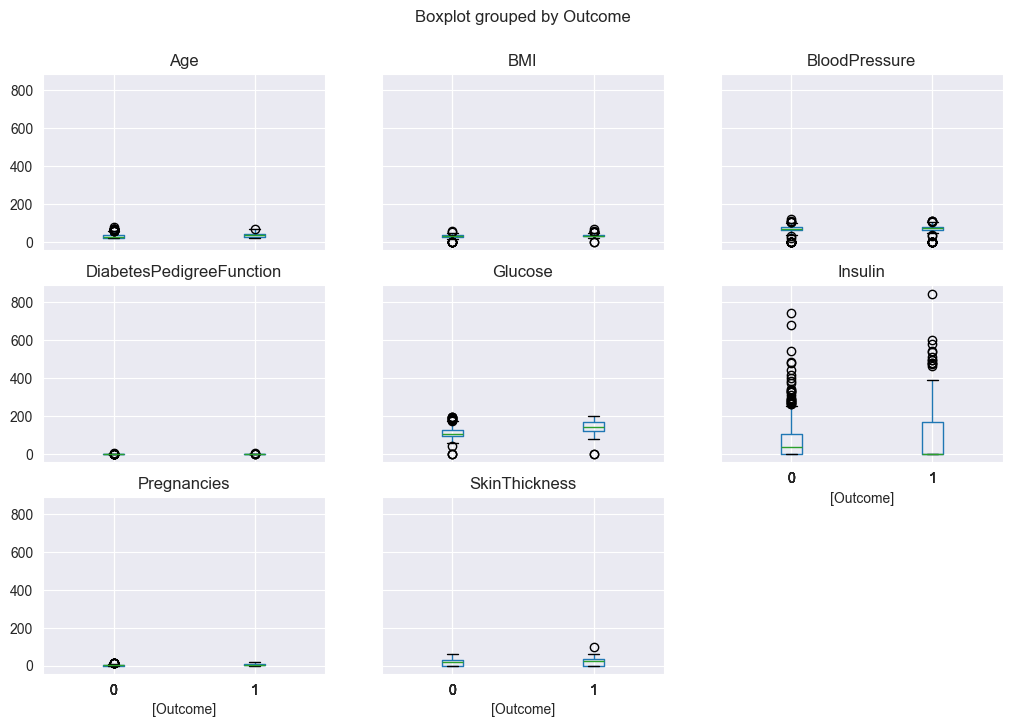
### 5.1 Biểu đồ hộp (Box plot)



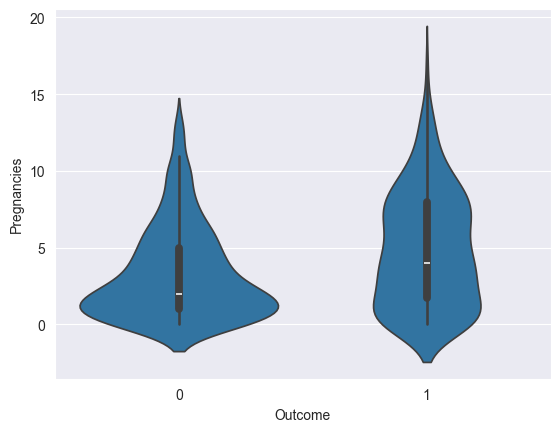
*Hình IV.5.2: Hình biểu đồ plotbox*

* Xu hướng chung: khi tuổi tăng thì số lần mang thai trung bình cũng cao hơn → phù hợp với thực tế sinh học.
* Độ phân tán: các cột boxplot càng cao ở nhóm tuổi lớn cho thấy sự khác biệt giữa phụ nữ cùng độ tuổi về số lần mang thai.
* Ngoại lệ (outliers): xuất hiện khá nhiều điểm bất thường, ví dụ có người trẻ nhưng số lần mang thai cao, hoặc ngược lại.

=> Kết luận: Pregnancies và Age có tương quan dương mạnh (cũng được xác nhận trong ma trận tương quan ~0.54).



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



*Hình IV.5.3: Hình biểu đồ violinplot*

* Violin plot cho thấy số lần mang thai ở nhóm mắc tiểu đường thường cao hơn nhóm không mắc.
* Nhóm không mắc tiểu đường tập trung ở Pregnancies thấp (0–2).
* Kết quả phù hợp với phân tích y khoa: nhiều lần mang thai có thể làm tăng nguy cơ mắc tiểu đường type 2.

# III. KẾT LUẬN

* ADAP đã chứng minh tiềm năng dự báo tiểu đường từ những năm 1980, mở đầu cho ứng dụng trí tuệ nhân tạo trong y tế.
* Pima Indians dataset trở thành một benchmark quan trọng trong nghiên cứu Machine Learning y tế.
* Hướng phát triển: Ứng dụng các mô hình hiện đại hơn (Deep Neural Network, XGBoost, AutoML). Cải thiện độ chính xác dự báo và tính ứng dụng thực tiễn.
* Ý nghĩa thực tế:
* Hỗ trợ bác sĩ sàng lọc sớm bệnh nhân có nguy cơ cao.
* Giúp bệnh nhân nhận thức và phòng ngừa tiểu đường từ sớm.
* Góp phần nâng cao chất lượng chăm sóc sức khỏe cộng đồng.

# 

# 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].