**TRƯỜNG ĐẠI HỌC SÀI GÒN**

**KHOA CÔNG NGHỆ THÔNG TIN**

**\*-\*-\***



**Bài báo cáo môn : Phân tích dữ liệu**

***Đề tài: Phân tích khám phá về bệnh đái tháo đường***

**Giảng viên : Đỗ Như Tài**

**Nhóm thực hiện:**

Nguyễn Thị Xuân Mai 3122410230

Trần Lê Đăng Khoa 3122410187

Lê Đức Nguyên Phú 3122410313

Mai Lê Mỹ Linh 3122410210

Thành phố Hồ Chí Minh - Tháng 10 năm 2025

**MỤC LỤC**

[1. Giới thiệu 1](#_Toc209707115)

[2. Tóm tắt nội dung của các Paper 2](#_Toc209707116)

[2.1. Paper 1 2](#_Toc209707117)

[a. Tóm tắt nội dung chính 2](#_Toc209707118)

[b. Phân loại bệnh 3](#_Toc209707119)

[b. Ý nghĩa bài toán 4](#_Toc209707120)

[d. Phân tích – khám phá dữ liệu 5](#_Toc209707121)

[2.2. Paper 2 5](#_Toc209707122)

[a. Tóm tắt nội dung chính 5](#_Toc209707123)

[b. Ý nghĩa & bối cảnh của bài toán 6](#_Toc209707124)

[c. Các nghiên cứu liên quan 6](#_Toc209707125)

[d. Phân tích – khám phá dữ liệu 7](#_Toc209707126)

[2.3. Paper 3 8](#_Toc209707127)

[a. Tóm tắt nội dung chính 8](#_Toc209707128)

[b. Ý nghĩa của bài toán 9](#_Toc209707129)

[c. Các nghiên cứu liên quan 9](#_Toc209707130)

[d. Phân tích – khám phá dữ liệu 10](#_Toc209707131)

[3. Mô tả tập dữ liệu 10](#_Toc209707132)

[3.1. Diabetes Dataset (AIM ’94) 10](#_Toc209707133)

[3.2. Pima Indians Diabetes Dataset (.names) 11](#_Toc209707134)

[3.3. Pima Indians Diabetes Dataset (.data.csv) 11](#_Toc209707135)

[4. Exploratory Data Analysis (EDA) in Python 11](#_Toc209707136)

[5. Kết Luận 24](#_Toc209707137)

**Bảng phân công công việc:**

|  |  |  |
| --- | --- | --- |
| **STT** | **Họ và tên** | **Nội dung công việc** |
| 1 | Mai Lê Mỹ Linh | Đọc hiểu và tóm tắt nội dung paper1, thiết kế slide thuyết trình |
| 2 | Trần Lê Đăng Khoa | Đọc hiểu và tóm tắt nội dung paper3, gõ code trên Google Colab |
| 3 | Lê Đức Nguyên Phú | Đọc hiểu và tóm tắt nội dung paper2, viết báo cáo Word, gõ code trên Google Colab |
| 4 | Nguyễn Thị Xuân Mai | Đọc và tóm tắt nội dung các file còn lại, gõ code trên Google Colab, viết báo cáo Word |

# 1. Giới thiệu

Bệnh đái tháo đường (tiểu đường) là một trong những vấn đề y tế toàn cầu nghiêm trọng nhất hiện nay, với tốc độ gia tăng nhanh chóng ở nhiều quốc gia, đặc biệt tại các nước đang phát triển. Đây là bệnh rối loạn chuyển hoá mạn tính, gây ảnh hưởng lớn đến sức khoẻ con người, làm tăng nguy cơ biến chứng tim mạch, suy thận, mù lòa và giảm chất lượng cuộc sống. Theo Tổ chức Y tế Thế giới (WHO), số ca mắc đái tháo đường đã tăng gấp hơn 4 lần trong 40 năm qua, cho thấy mức độ đáng báo động và nhu cầu cấp thiết trong việc dự phòng, phát hiện sớm và điều trị.

Trong bối cảnh khoa học dữ liệu ngày càng phát triển, việc ứng dụng các phương pháp phân tích dữ liệu vào lĩnh vực y tế mang lại giá trị to lớn. Các kỹ thuật phân tích khám phá dữ liệu (Exploratory Data Analysis - EDA) không chỉ giúp hiểu rõ đặc điểm của bộ dữ liệu, phát hiện xu hướng và mối liên hệ giữa các biến lâm sàng, mà còn tạo nền tảng quan trọng cho việc xây dựng các mô hình dự đoán nguy cơ mắc bệnh.

Đề tài *“Phân tích khám phá về bệnh đái tháo đường”* hướng đến việc sử dụng các công cụ và phương pháp phân tích dữ liệu để:

* Tìm hiểu phân bố và đặc điểm của các yếu tố nguy cơ (như chỉ số BMI, huyết áp, nồng độ glucose, tuổi…).
* Khảo sát mối quan hệ giữa các biến độc lập với tình trạng bệnh.
* Đánh giá sơ bộ khả năng dự báo của dữ liệu trong việc phân loại người có nguy cơ mắc đái tháo đường.

Kết quả của phân tích này không chỉ có ý nghĩa trong học thuật, mà còn mang giá trị thực tiễn, góp phần hỗ trợ công tác y tế cộng đồng trong việc nhận diện sớm các nhóm đối tượng nguy cơ, từ đó đưa ra giải pháp can thiệp phù hợp.

# 2. Tóm tắt nội dung của các Paper

## 2.1. Paper 1

### a. Tóm tắt nội dung chính

WHO, 1999 – “Định nghĩa, Chẩn đoán và Phân loại Bệnh Đái tháo đường và Biến chứng”. Tài liệu này đưa ra tiêu chuẩn quốc tế cho việc chẩn đoán, phân loại và đánh giá biến chứng của bệnh đái tháo đường. ĐTĐ là một trong những nguyên nhân hàng đầu gây tử vong và tàn phế, do đó việc chuẩn hóa chẩn đoán và phân loại có ý nghĩa to lớn trong y tế công cộng.

Theo WHO, đái tháo đường là một nhóm bệnh chuyển hóa đặc trưng bởi tình trạng tăng glucose huyết mạn tính, do khiếm khuyết bài tiết insulin, hoạt động insulin, hoặc cả hai. Tiêu chuẩn chẩn đoán bao gồm glucose huyết tương lúc đói (FPG), glucose huyết tương sau nghiệm pháp dung nạp glucose đường uống (OGTT), và HbA1c.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Phân loại | Glucose đói (mmol/L) | OGTT 2h (mmol/L) | HbA1c(%) | Đặc điểm |
| Bình thường | <6.1 (<110 mg/dl) | <7.8 | <6.0 | - |
| IFG | 6.1–6.9 | <7.8 | - | Thường gặp ở nam |
| IGT | <7.0 | 7.8–11.0 | - | Thường gặp ở nữ, người lớn tuổi |
| Đái tháo đường | ≥7.0 | ≥11.1 | ≥6.5 (\*) | Chuẩn WHO 1999 |
| GDM (thai kỳ) | ≥7.0 hoặc ≥11.1 | ≥11.1 | - | Xuất hiện tuần 24–28 |

**1. Tiểu đường thai kỳ (GDM)**

GDM được chẩn đoán lần đầu trong thai kỳ, thường xuất hiện từ tuần 24–28. GDM làm tăng nguy cơ cho cả mẹ và thai nhi, bao gồm nguy cơ sinh non, thai to và tiến triển thành ĐTĐ type 2 sau này.

**2. Hội chứng chuyển hóa**

Hội chứng chuyển hóa bao gồm tập hợp nhiều yếu tố nguy cơ: tăng glucose máu, tăng huyết áp, rối loạn lipid máu, và béo bụng. Đây là một trong những yếu tố góp phần thúc đẩy tiến triển ĐTĐ type 2 và bệnh tim mạch.

**3. Nghiệm pháp OGTT và phương pháp đo lường**

OGTT (Oral Glucose Tolerance Test) là phương pháp đo khả năng dung nạp glucose của cơ thể. Phương pháp này yêu cầu bệnh nhân uống 75g glucose sau khi nhịn đói và đo glucose huyết tương sau 2 giờ. Ngoài ra, WHO cũng khuyến nghị chuẩn hóa quy trình lấy mẫu và xét nghiệm để đảm bảo độ chính xác.

### b. Phân loại bệnh

WHO phân loại đái tháo đường thành nhiều nhóm dựa trên nguyên nhân và cơ chế bệnh sinh.

|  |  |
| --- | --- |
| Phân loại | Nguyên nhân |
| Type 1 | Tự miễn dịch; Không rõ nguyên nhân |
| Type 2 | Khiếm khuyết di truyền tế bào beta; Khiếm khuyết tác động insulin; Bệnh tụy ngoại tiết; Bệnh nội tiết; Do thuốc hoặc hóa chất; Nhiễm trùng; Miễn dịch hiếm; Hội chứng di truyền |
| GDM | Xuất hiện trong thai kỳ (24–28 tuần) |

*Bảng 2. Phân loại Nguyên nhân của các Rối loạn Đường huyết*

|  |  |
| --- | --- |
| Loại | Chi tiết |
| Khiếm khuyết di truyền tế bào beta | Nhiễm sắc thể 20, HNF4α (MODY1); Nhiễm sắc thể 7, glucokinase (MODY2); Nhiễm sắc thể 12, HNF1α (MODY3); Nhiễm sắc thể 13, IPF-1 (MODY4); Đột biến DNA ty thể 3243; Khác |

*Bảng 3. Các Loại Cụ thể Khác của Tiểu đường*

|  |
| --- |
| Thuốc / Hóa chất |
| Acid nicotinic |
| Glucocorticoid |
| Hormone tuyến giáp |
| Chất chủ vận alpha-adrenergic |
| Chất chủ vận beta-adrenergic |
| Thiazide |
| Dilantin |
| Pentamidine |
| Vacor |
| Liệu pháp interferon-alpha |
| Khác |

*Bảng 4. Tiểu đường do Thuốc hoặc Hóa chất Gây ra*

|  |  |
| --- | --- |
| Nhóm | Chi tiết |
| Hội chứng di truyền | Down; Friedreich; Huntington; Klinefelter; Lawrence-Moon-Biedl; Loạn dưỡng cơ; Porphyria; Prader-Willi; Turner; Wolfram; Khác |
| Khiếm khuyết tác động insulin | Đột biến gen thụ thể insulin |
| Bệnh tụy ngoại tiết | Viêm tụy mạn; Xơ nang; Ung thư tụy; Phẫu thuật cắt tụy |
| Bệnh nội tiết | Cường giáp; To đầu chi; Cushing; U tủy thượng thận |

*Bảng 5. Các Hội chứng Di truyền Đôi khi Liên quan đến Tiểu đường*

### b. Ý nghĩa bài toán

**Input**: Các biến lâm sàng và cận lâm sàng như:

* Glucose huyết tương lúc đói và sau 2h OGTT
* Huyết áp, BMI, tuổi, insulin huyết tương
* Lịch sử gia đình (diabetes pedigree function)
* Các yếu tố liên quan khác như thai kỳ (gestational diabetes) hoặc tình trạng rối loạn dung nạp glucose (IGT, IFG) .

**Output**: Một biến nhị phân cho biết **có hay không chẩn đoán đái tháo đường** theo tiêu chuẩn WHO:

* FPG ≥ 7.0 mmol/L (126 mg/dL) hoặc
* 2h OGTT ≥ 11.1 mmol/L (200 mg/dL) .

**Mục tiêu**:

* Dự đoán sớm khả năng mắc đái tháo đường dựa trên thông tin ban đầu, thay vì đợi OGTT.
* Ứng dụng vào **sàng lọc cộng đồng**, giảm chi phí xét nghiệm, phát hiện người nguy cơ cao để can thiệp kịp thời.
* Hỗ trợ nghiên cứu yếu tố nguy cơ và xu hướng dịch tễ (ví dụ tăng nhanh ở Pima Indians hoặc nhóm có béo phì, ít vận động) .

**c**. **Các nghiên cứu liên quan**

* WHO 1980, 1985, 1999: ban hành và điều chỉnh tiêu chuẩn chẩn đoán, phân loại Type 1, Type 2, GDM, IGT, IFG .
* ADA 1997: đề xuất chuẩn chẩn đoán dựa trên FPG, được so sánh và cập nhật trong WHO 1999 .
* Pima Indian studies (NIDDK, Phoenix): nghiên cứu dịch tễ cho thấy insulin resistance và di truyền góp phần lớn vào Type 2 .
* LADA, MODY, bệnh lý tụy, thuốc/hoá chất, nhiễm trùng: các phân nhóm đặc biệt được WHO ghi nhận, nhấn mạnh tính đa dạng nguyên nhân .
* Nghiên cứu áp dụng ML/AI: Paper2 với ADAP learning algorithm; Logistic Regression, Perceptron, SVM… đều thử nghiệm trên cùng bộ dữ liệu để so sánh hiệu quả.

### d. Phân tích – khám phá dữ liệu

Khi làm EDA với Pima dataset hoặc bộ dữ liệu tương tự, ta cần:

* Thống kê mô tả: Trung bình, độ lệch chuẩn, tỉ lệ giá trị bất thường (ví dụ glucose = 0).
* Phân bố biến: Histogram của glucose, BMI, insulin. Bar chart tỉ lệ chẩn đoán (Outcome).
* Quan hệ biến – nhãn: Scatter plot BMI vs Glucose 2h OGTT để thấy mối liên hệ . Boxplot glucose theo nhóm Outcome (0/1).
* Đa cộng tuyến: Heatmap correlation để phát hiện biến trùng lặp.
* Xử lý dữ liệu thiếu/ngoại lai: Ở Pima, giá trị 0 trong BMI/Glucose được coi như missing, cần thay thế hoặc loại bỏ.
* So sánh kịch bản: Dùng toàn bộ biến (có glucose 2h). Loại biến glucose 2h để mô phỏng screening thực tế (tránh concept leakage).

## 2.2. Paper 2

### a. Tóm tắt nội dung chính

Bài báo kiểm tra khả năng của một mô hình mạng nơ-ron cổ điển tên ADAP để dự báo khởi phát đái tháo đường typ 2 trong vòng 5 năm ở cộng đồng người Pima gần Phoenix, Arizona. Tác giả đánh giá bằng độ nhạy, độ đặc hiệu và đường ROC; điểm giao nhau của độ nhạy và độ đặc hiệu đạt 0.76. Nhóm tác giả cũng dự kiến so sánh cùng tập dữ liệu với hồi quy logistic và perceptron tuyến tính để đối chiếu kết quả.

Dữ liệu nghiên cứu được thu thập theo chu trình khám định kỳ 2 năm, có test dung nạp glucose đường uống (OGTT); chẩn đoán đái tháo đường dựa theo tiêu chí WHO: glucose huyết tương sau 2 giờ ≥ 200 mg/dL hoặc ghi nhận giá trị tương đương trong chăm sóc thường quy.

### b. Ý nghĩa & bối cảnh của bài toán

**Bối cảnh học máy:**

* *Đối tượng:* phụ nữ người Pima, chọn duy nhất 1 lần khám “chỉ mục” mỗi người, tuổi ≥ 21 tại thời điểm khám. Các ca được chọn sao cho: hoặc được chẩn đoán đái tháo đường trong 1-5 năm sau khám, hoặc vẫn không đái tháo đường ít nhất 5 năm sau. Các trường hợp phát bệnh trong < 1 năm sau khám bị loại để tránh ca “quá dễ” dự báo.
* *Quy mô:* 768 ca, chia ngẫu nhiên 576 train và 192 test.

**Đầu vào (Input features, 8 biến):**

1. Số lần mang thai
2. Glucose huyết tương 2 giờ sau OGTT
3. Huyết áp tâm trương
4. Độ dày nếp gấp da cánh tay (triceps)
5. Insulin huyết thanh sau 2 giờ
6. BMI
7. Chỉ số phả hệ đái tháo đường (DPF) - tổng hợp tiền sử gia đình và mức độ quan hệ huyết thống
8. Tuổi: Bài báo còn nêu công thức xây dựng DPF từ tình trạng bệnh của họ hàng, tuổi lúc mắc bệnh hoặc tuổi lần khám gần nhất còn không bệnh, và mức độ chia sẻ gen 0.5, 0.25, 0.125 tương ứng mức quan hệ.

**Đầu ra (Output):**

* Xác suất hoặc quyết định nhị phân: có/không phát bệnh đái tháo đường trong 5 năm kể từ thời điểm khám chỉ mục. Trong đánh giá cuối bài, nhóm dùng ngưỡng phân loại 0.448 để tối ưu cân bằng, đạt độ nhạy = độ đặc hiệu = 0.76 trên tập test và minh họa bằng ROC.

**Mục tiêu học máy:**

* Xây mô hình phân loại nhị phân nhằm dự báo sớm nguy cơ để sàng lọc và can thiệp. Mục tiêu định lượng: tối ưu đường ROC, lựa chọn ngưỡng theo mục tiêu lâm sàng - ví dụ ưu tiên độ nhạy nếu muốn giảm bỏ sót ca nguy cơ.

### c. Các nghiên cứu liên quan

* Bài báo đặt ADAP cạnh các phương pháp kinh điển như hồi quy logistic và perceptron tuyến tính trên cùng tập train-test để so sánh ROC, cho thấy cách tư duy “đối chiếu mô hình” đã có từ sớm.
* Nguồn gốc tiêu chí chẩn đoán và quan tâm đến ROC được neo vào tài liệu WHO và y văn ROC cổ điển mà bài báo trích dẫn.
* Về dịch tễ Pima, bài báo cũng trích các nghiên cứu trước đây thể hiện tỷ lệ và yếu tố nguy cơ trong quần thể này, nhấn mạnh vai trò của béo phì và tiền sử gia đình - cơ sở để đưa DPF vào mô hình.

### d. Phân tích – khám phá dữ liệu

1. **Kiểm tra cấu trúc và phân bố biến**

* Thống kê mô tả 8 biến đầu vào: trung vị, IQR, min-max; vẽ histogram cho các biến liên tục như glucose 2h, insulin 2h, BMI, tuổi; boxplot theo nhãn phát bệnh 5 năm. Mục tiêu: phát hiện lệch trái/phải, đuôi dài, ngoại lai. (Danh sách 8 biến trích từ paper.)
* Ma trận tương quan giữa các biến liên tục và cảnh báo đa cộng tuyến tiềm năng, ví dụ BMI với huyết áp, insulin với glucose 2h.
* Biến DPF: kiểm tra phân bố và mối liên hệ với nhãn; đọc lại định nghĩa để hiểu trực giác DPF tăng khi có nhiều thân nhân mắc sớm và quan hệ huyết thống gần.

1. **Chất lượng dữ liệu**

* Tỉ lệ thiếu vắng và quy tắc xử lý: insulin 2h và skinfold thường có tỷ lệ thiếu cao trong nhiều bản dữ liệu Pima; cân nhắc loại bỏ, nội suy hợp lý, hoặc dùng mô hình chịu thiếu.
* Kiểm tra giá trị bất thường - ví dụ giá trị 0 ở những biến sinh học không thể bằng 0 (huyết áp, insulin, skinfold). Với bản Pima phổ biến, 0 thường mã hóa “không đo được”. Nếu gặp, cần biến đổi về NA trước khi phân tích.

1. **Mục tiêu và sự sai lệch chọn mẫu**

* Nhãn mục tiêu là sự kiện phát bệnh trong 1-5 năm sau khám chỉ mục; các ca phát bệnh < 1 năm đã bị loại để tránh “dễ đoán”. Điều này ảnh hưởng phân bố xác suất nền và có thể làm mô hình ít gặp những ca “gần phát bệnh”, do đó cần ghi nhận khi diễn giải.
* Tỷ lệ dương - âm: tính tỉ lệ phát bệnh 5 năm trong tập dữ liệu, đối chiếu với tỷ lệ của cộng đồng nếu có, để phát hiện lệch chọn mẫu.

1. **Khám phá quan hệ Input – Output**

* Phân rã theo nhãn: so sánh phân bố glucose 2h, insulin 2h, BMI, tuổi giữa nhóm phát bệnh và không phát bệnh.
* Đường cong ROC sơ bộ với 1 biến: kiểm tra từng biến đơn lẻ, đặc biệt glucose 2h vì đây cũng là tiêu chí chẩn đoán WHO; lưu ý rủi ro “gần với tiêu chí gán nhãn” có thể làm biến này đóng vai trò mạnh một cách hiển nhiên. Paper xác nhận tiêu chí WHO đúng là ngưỡng 200 mg/dL sau 2h OGTT.

1. **Thiết lập baseline và ngưỡng phân loại**

* Baseline mô hình tuyến tính: logistic regression với 8 biến chuẩn hóa, thêm chuẩn hóa robust nếu nhiều ngoại lai.
* Đối chiếu với ADAP như trong paper: mô hình ADAP kích hoạt khoảng 8% đơn vị liên kết với ngưỡng kích hoạt 4 kết nối, dùng 100,000 “association units” - đây là đặc thù của ADAP và giải thích vì sao ADAP có khả năng nắm bắt tương tác.
* Lựa chọn ngưỡng: nếu mục tiêu sàng lọc, có thể ưu tiên độ nhạy hơn, chấp nhận tăng dương tính giả. Paper dùng ngưỡng 0.448 để cân bằng 0.76-0.76, có thể là mốc tham khảo ban đầu trước khi hiệu chỉnh theo chi phí lâm sàng.

1. **Diễn giải và rủi ro rò rỉ khái niệm**

* Vì nhãn dựa trên chuẩn WHO về glucose 2h ở các lần khám sau, còn biến glucose 2h tại thời điểm chỉ mục nằm trong Input, cần cẩn trọng khi suy diễn - đây không phải rò rỉ dữ liệu theo nghĩa dùng trực tiếp nhãn tương lai, nhưng là biến có liên hệ sinh lý - chẩn đoán rất mạnh với nhãn sau này. Nên báo cáo thêm kịch bản bỏ biến glucose 2h để đánh giá giá trị sàng lọc khi thiếu OGTT - cách mà nhiều nhóm thực hành khi triển khai trong thực tế sàng lọc ban đầu. (Tiêu chí WHO dẫn chiếu trong paper.)

1. **Mô tả ngắn gọn về ADAP trong paper**

* ADAP chia miền giá trị mỗi biến thành các “phân vùng - sensors”, kết nối ngẫu nhiên đến nhiều association units; một association unit “bật” nếu số kết nối khớp vượt ngưỡng; dự báo là tổng trọng số các đơn vị được kích hoạt. Trong nghiên cứu này, ADAP dùng 100,000 association units, điều chỉnh theo quy tắc giống Hebb, ngưỡng kích hoạt cấu hình sao cho 2-10% đơn vị được bật trên mỗi ca.
* *Quá trình học*: sau mỗi ca, ADAP phân bổ sai số lên các đơn vị đã kích hoạt; khi suy diễn thì không cập nhật nữa.

## 2.3. Paper 3

### a. Tóm tắt nội dung chính

Bài báo này là của National Diabetes Data Group (NIH, 1979), tập trung vào việc phân loại và chẩn đoán bệnh tiểu đường (diabetes mellitus) và các dạng rối loạn dung nạp glucose khác**.**

* Bài báo nhấn mạnh tiểu đường là một hội chứng không đồng nhất (heterogeneous syndrome),có nhiều nguyên nhân và biểu hiện khác nhau.
* Đề xuất hệ thống phân loại quốc tế, được các hiệp hội lớn (ADA, WHO, BDA, EASD…) công nhận, nhằm tạo sự thống nhất cho nghiên cứu lâm sàng và dịch tễ học.

Các nhóm chính:

* ***Type I (IDDM – Insulin Dependent Diabetes Mellitus)****:* Tiểu đường phụ thuộc insulin, dễ nhiễm toan ceton, liên quan đến yếu tố di truyền và miễn dịch.
* ***Type II (NIDDM – Non Insulin Dependent Diabetes Mellitus)****:* Không phụ thuộc insulin, liên quan chặt chẽ đến béo phì và kháng insulin.
* ***Diabetes* do các bệnh lý khác** (pancreatic disease, rối loạn hormone, thuốc, hội chứng di truyền…).
* ***Gestational Diabetes (GDM)****:* Tiểu đường thai kỳ.
* ***Impaired Glucose Tolerance (IGT)****:* Dung nạp glucose giảm.
* ***Previous abnormality / Potential abnormality****:* Những trường hợp từng rối loạn hoặc có nguy cơ cao bị tiểu đường.

### b. Ý nghĩa của bài toán

**Input:** Các chỉ số y khoa thu thập được (nồng độ glucose trong máu khi đói, sau nghiệm pháp dung nạp glucose, triệu chứng lâm sàng, tiền sử bệnh, yếu tố di truyền…).

**Output:** Xếp loại bệnh nhân vào các nhóm: Tiểu đường type I, type II, tiểu đường thai kỳ, dung nạp glucose giảm, hoặc bình thường.

Mục tiêu:

* + Xây dựng một hệ thống phân loại chuẩn hóa, giúp so sánh dữ liệu nghiên cứu trên toàn cầu.
  + Cải thiện chẩn đoán, phát hiện sớm và theo dõi bệnh.
  + Làm nền tảng để nghiên cứu về dịch tễ học, điều trị và phòng ngừa biến chứng tiểu đường.

### c. Các nghiên cứu liên quan

* Trước 1979, nhiều tiêu chuẩn chẩn đoán khác nhau được áp dụng: USPHS (Mỹ), WHO, Fajans & Conn, BDA (Anh),gây khó khăn khi so sánh dữ liệu dịch tễ.
* Các nghiên cứu dài hạn ở Mỹ, Anh, Ấn Độ, Pima Indian… cho thấy không có ranh giới rõ ràng giữa “bình thường” và “tiểu đường”, mà thường liên tục theo phân bố unimodal/bimodal.
* Nghiên cứu về di truyền (HLA trên nhiễm sắc thể 6, kháng thể tế bào đảo tụy) khẳng định type I có cơ chế tự miễn dịch.
* Các khảo sát về béo phì, kháng insulin, di truyền trội củng cố phân loại type II và MODY (maturity-onset diabetes of the young).

### d. Phân tích – khám phá dữ liệu

Nếu xem dữ liệu lâm sàng và dịch tễ học (như trong các nghiên cứu được trích dẫn trong paper), các bước EDA có thể bao gồm:

1. Thống kê mô tả:
   * Tỷ lệ mắc bệnh theo độ tuổi, giới tính, chủng tộc.
   * Phân bố giá trị glucose lúc đói và sau nghiệm pháp dung nạp.
   * Tỷ lệ các loại tiểu đường (type I, type II, GDM, IGT).
2. Khám phá mối quan hệ:
   * Liên hệ giữa béo phì và nguy cơ NIDDM.
   * Ảnh hưởng của yếu tố di truyền (HLA, tiền sử gia đình) đến type I.
   * Liên hệ giữa tiểu đường thai kỳ và khả năng tiến triển thành type II sau sinh.
3. Trực quan hóa dữ liệu:
   * Vẽ histogram phân bố đường huyết lúc đói, 2h sau glucose.
   * Vẽ biểu đồ so sánh nguy cơ giữa các nhóm dân số.
   * Heatmap để biểu diễn mối quan hệ giữa BMI, glucose, huyết áp và nguy cơ tiểu đường.
4. Nhận định:
   * Không có “ngưỡng cứng” tuyệt đối, mà là sự liên tục giữa bình thường – IGT – tiểu đường.
   * Các yếu tố lối sống (ăn uống, hoạt động) và yếu tố di truyền đều đóng vai trò quan trọng.

# 3. Mô tả tập dữ liệu

Trong nghiên cứu này, chúng em tham khảo và sử dụng hai bộ dữ liệu kinh điển về bệnh đái tháo đường từ UCI Repository, bao gồm **Diabetes Dataset (AIM ’94)** và **Pima Indians Diabetes Dataset**.

## **3.1. Diabetes Dataset (AIM ’94)**

Bộ dữ liệu này có đặc điểm multivariate và time-series, thuộc lĩnh vực y tế và sức khỏe, với nhiệm vụ chính là phân loại (classification). Dữ liệu được thu thập từ hai nguồn: thiết bị điện tử tự động (có timestamp thực tế) và hồ sơ giấy (với mốc thời gian logic: sáng, trưa, tối, trước khi ngủ, quy đổi thành 08:00, 12:00, 18:00, 22:00). Cấu trúc dữ liệu bao gồm bốn trường: ngày, giờ, mã sự kiện (insulin, glucose, ăn uống, vận động, triệu chứng...) và giá trị đo lường. Một số mã tiêu biểu: liều insulin (33–35), đo glucose (48, 57–64), triệu chứng hạ đường huyết (65), lượng bữa ăn (66–68), hoạt động thể dục (69–71), sự kiện đặc biệt (72).

* **Input**: chuỗi bản ghi với 4 trường (ngày, giờ, code, giá trị).
* **Output**: chuỗi sự kiện y tế của bệnh nhân theo thời gian.
* **Mục tiêu**: phân tích tiến triển bệnh tiểu đường, dự đoán biến động glucose và hỗ trợ quyết định điều trị (liều insulin, chế độ ăn uống, hoạt động thể chất).

## 3.2. Pima Indians Diabetes Dataset (.names)

Bộ dữ liệu được thu thập bởi *National Institute of Diabetes and Digestive and Kidney Diseases* và đóng góp bởi Vincent Sigillito (Johns Hopkins University, Applied Physics Lab). Đây là bộ dữ liệu được sử dụng rộng rãi trong nghiên cứu dự báo nguy cơ đái tháo đường, tiêu biểu là nghiên cứu của Smith et al. (1988) với mô hình ADAP, đạt độ chính xác sensitivity/specificity khoảng 76%. Đối tượng là phụ nữ từ 21 tuổi trở lên, thuộc cộng đồng Pima Indian, sinh sống gần Phoenix, Arizona (USA). Bộ dữ liệu bao gồm 768 mẫu với 8 thuộc tính lâm sàng và 1 biến mục tiêu (nhị phân).

* **Thuộc tính**: số lần mang thai, nồng độ glucose (2h OGTT), huyết áp tâm trương, độ dày da cánh tay, insulin 2h, BMI, chỉ số di truyền (diabetes pedigree function), tuổi.
* **Biến mục tiêu (Outcome)**: 0 = không mắc tiểu đường, 1 = mắc tiểu đường.
* **Phân bố lớp**: 500 mẫu không tiểu đường, 268 mẫu có tiểu đường (tỉ lệ ≈ 34.9%).
* **Thống kê nhanh (Mean ± SD)**: Pregnancies (3.8 ± 3.4), Glucose (120.9 ± 32.0), Blood Pressure (69.1 ± 19.4), Skin fold (20.5 ± 16.0), Insulin (79.8 ± 115.2), BMI (32.0 ± 7.9), Pedigree (0.5 ± 0.3), Age (33.2 ± 11.8).

## 3.3. Pima Indians Diabetes Dataset (.data.csv)

File dữ liệu thực tế gồm 768 dòng và 9 cột, trong đó không có giá trị null. Do dữ liệu gốc không có tên cột, cần gắn nhãn theo chuẩn: *Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, Age, Outcome*. Kết quả phân bố class trùng khớp với phần mô tả: 500 mẫu không mắc bệnh và 268 mẫu có bệnh. Bộ dữ liệu này thường được sử dụng cho các mô hình phân loại nhị phân trong machine learning, nhằm dự đoán khả năng mắc bệnh đái tháo đường dựa trên các chỉ số y tế cơ bản.

# 4. Exploratory Data Analysis (EDA) in Python

Exploratory Data Analysis of the Pima Indians Diabetes dataset

The features present in the dataset are:

Pregnancies: Số lần mang thai

Glucose (plasma glucose concentration): Nồng độ glucose trong huyết tương

Blood Pressure (diastolic): Huyết áp tâm trương

Skin Thickness: Độ dày lớp da

Insulin

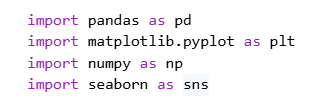
BMI (Body Mass Index)

Diabetes Pedigree Function (a function that scores likelihood of diabetes based on family history): Hàm số phả hệ bệnh dựa trên tiền sử gia đình

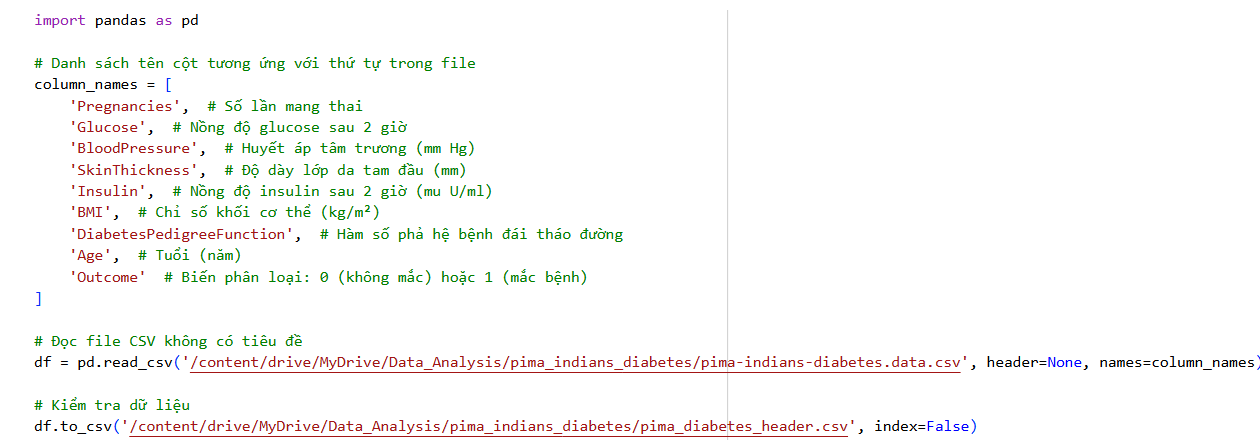
Age

The target "Outcome" has 2 classes: 0: Non-diabetes 1: diabetes

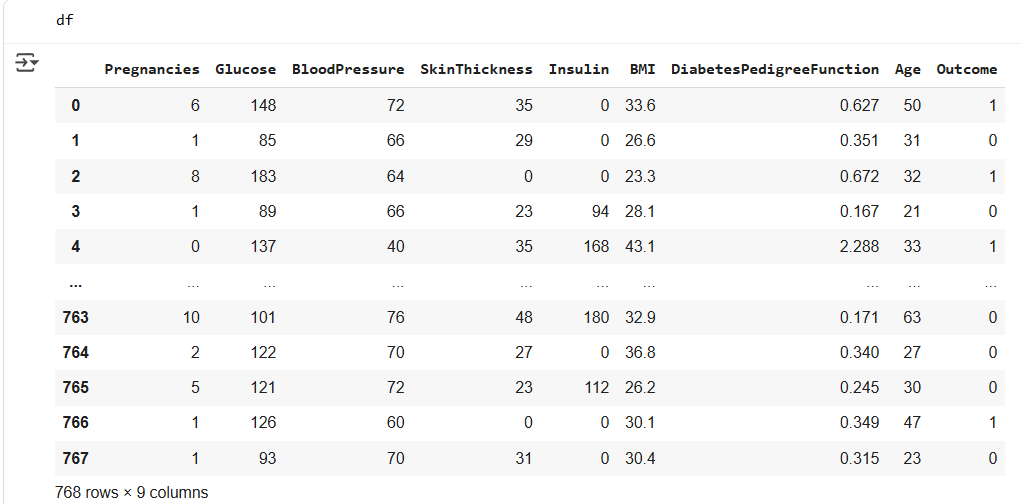
Initial EDA How: Using powerfull python moduls [pandas, matplotlib, and seaborn]

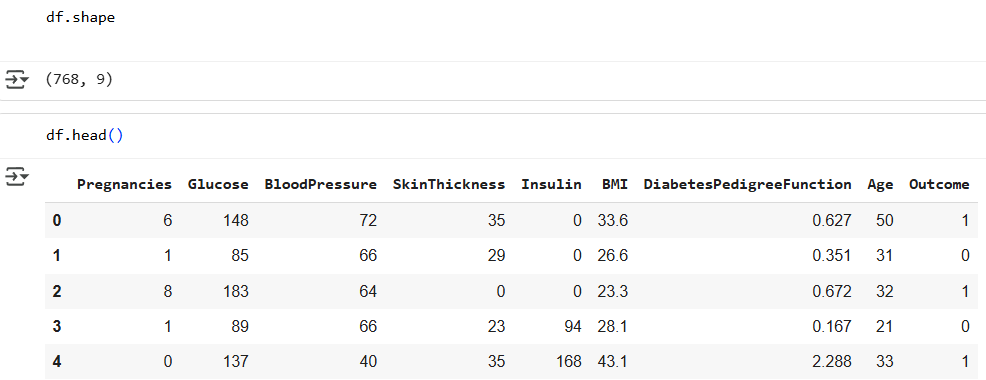


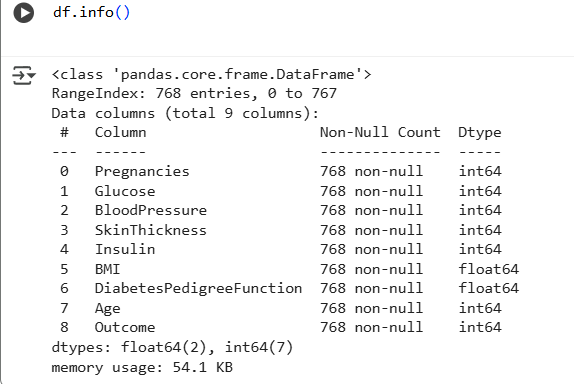
* Load the dataset:

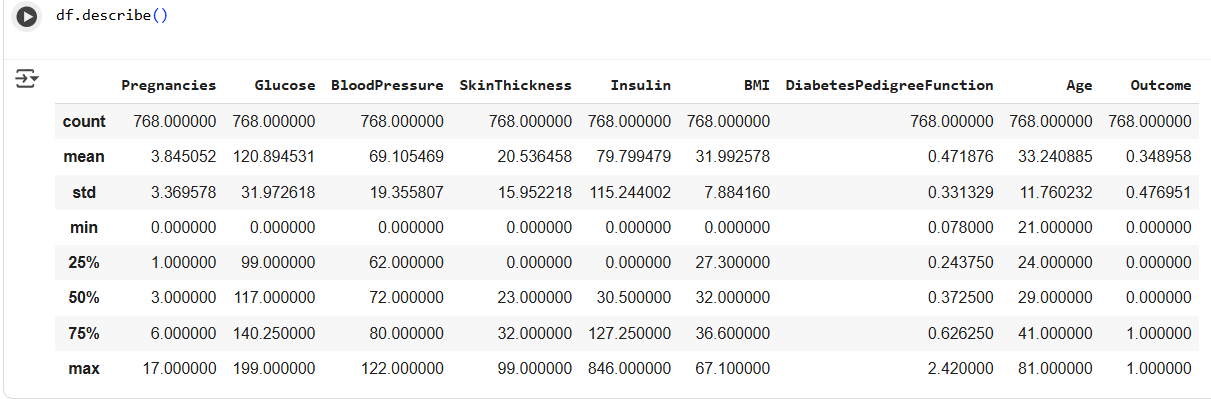


* A first look and non-graphical EDA

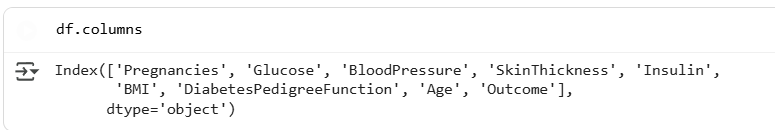


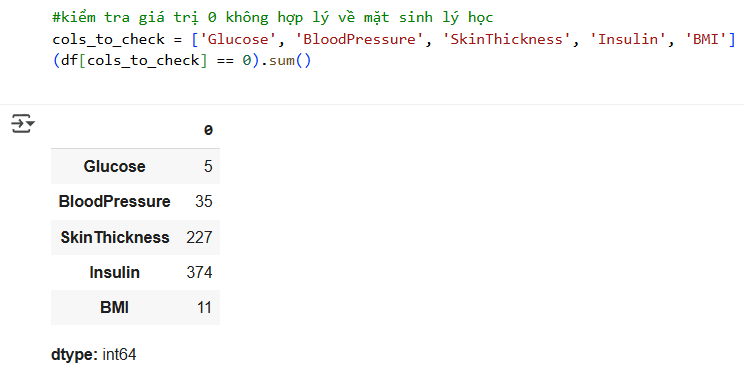




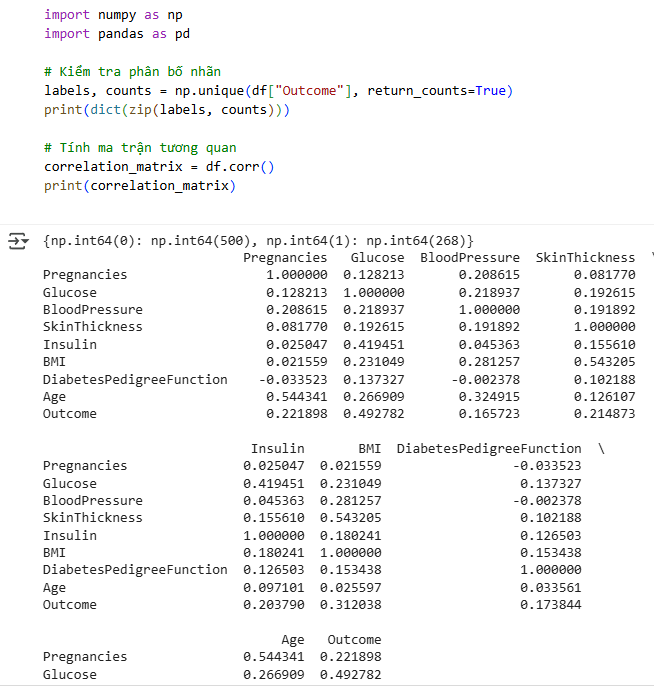


* Are the classes (species) balanced?





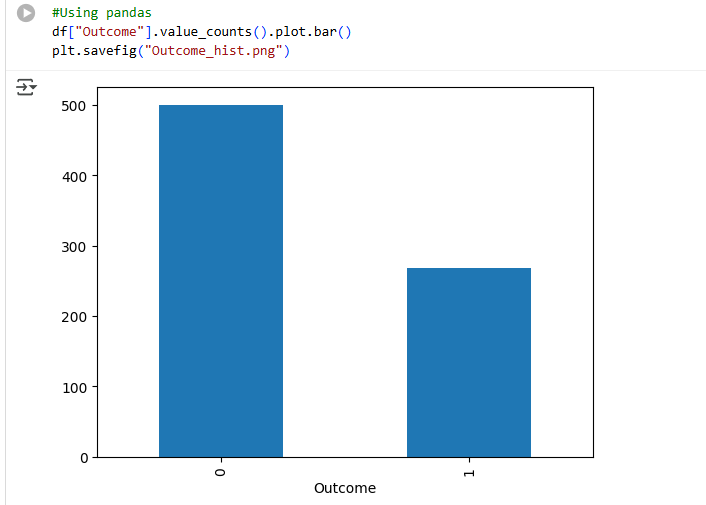
* Correlation Between Variables (sự tương quan giữa các biến)



* Graphical EDA
* Biểu đồ biểu diễn tuần suất giá trị Outcome

*#Using seaborn*

sns.countplot(x = df['Outcome']);



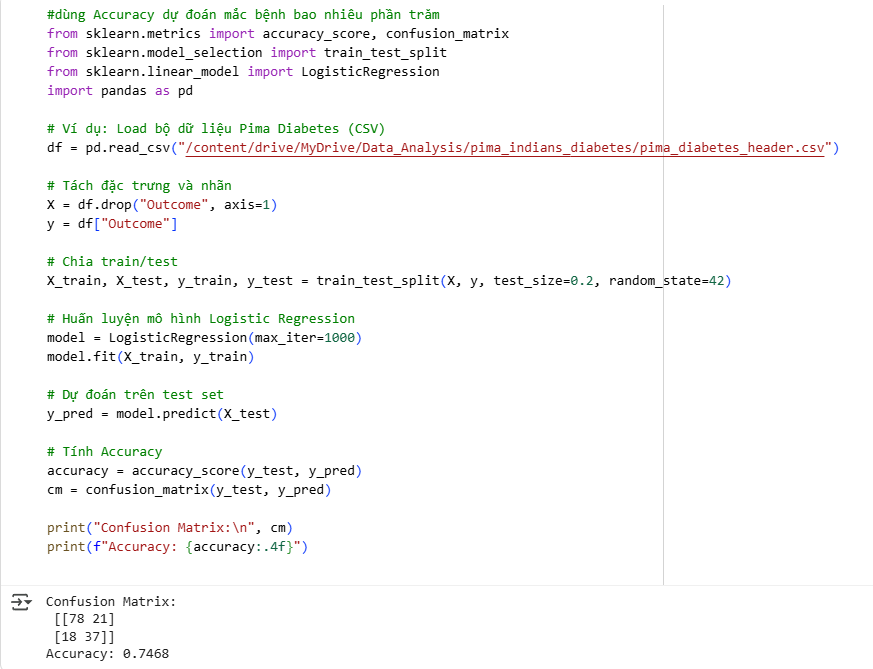
What we can conclude:

* Outcome distribution: The dataset consists of roughly 65% samples with Outcome = 0 (non-diabetic) and 35% with Outcome = 1 (diabetic). This distribution is not severely imbalanced, which means common evaluation metrics such as Accuracy, Precision, Recall, F1-score, and ROC AUC can be reliably applied without the need for data balancing techniques (e.g., oversampling or undersampling).
* Data preprocessing: Unrealistic zero values in medical features (Glucose, BloodPressure, SkinThickness, Insulin, BMI) were replaced with the median of each corresponding feature. This approach eliminates invalid values while preserving the natural distribution of the data.
* Significance: After replacement, the dataset is cleaner and more representative of real-world medical data, providing a more reliable basis for model training and improving the accuracy of diabetes prediction.

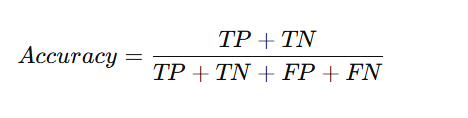
***Phân bố Outcome:*** *Tập dữ liệu gồm khoảng 65% mẫu với Outcome = 0 (không mắc tiểu đường) và 35% mẫu với Outcome = 1 (mắc tiểu đường). Phân bố này không quá mất cân bằng, do đó có thể sử dụng các thước đo đánh giá phổ biến như Accuracy, Precision, Recall, F1-score và ROC AUC một cách đáng tin cậy mà không cần đến các kỹ thuật cân bằng dữ liệu (ví dụ: oversampling hay undersampling).*

***Tiền xử lý dữ liệu:*** *Các giá trị bằng 0 không thực tế trong các đặc trưng y tế (Glucose, BloodPressure, SkinThickness, Insulin, BMI) đã được thay thế bằng giá trị trung vị của từng đặc trưng tương ứng. Cách tiếp cận này loại bỏ được những giá trị không hợp lệ đồng thời vẫn giữ được phân bố tự nhiên của dữ liệu.*

***Ý nghĩa:*** *Sau khi thay thế, tập dữ liệu trở nên sạch hơn và phản ánh sát hơn với dữ liệu y tế trong thực tế, từ đó tạo cơ sở đáng tin cậy hơn cho việc huấn luyện mô hình và nâng cao độ chính xác trong dự đoán bệnh tiểu đường.*



* Công thức Accuracy:

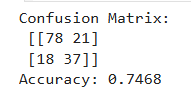


**TP (True Positive):** dự đoán đúng người bệnh tiểu đường.

**TN (True Negative):** dự đoán đúng người không mắc tiểu đường.

**FP (False Positive):** dự đoán nhầm là tiểu đường, nhưng thực tế không mắc.

**FN (False Negative):** dự đoán nhầm là không tiểu đường, nhưng thực tế mắc.

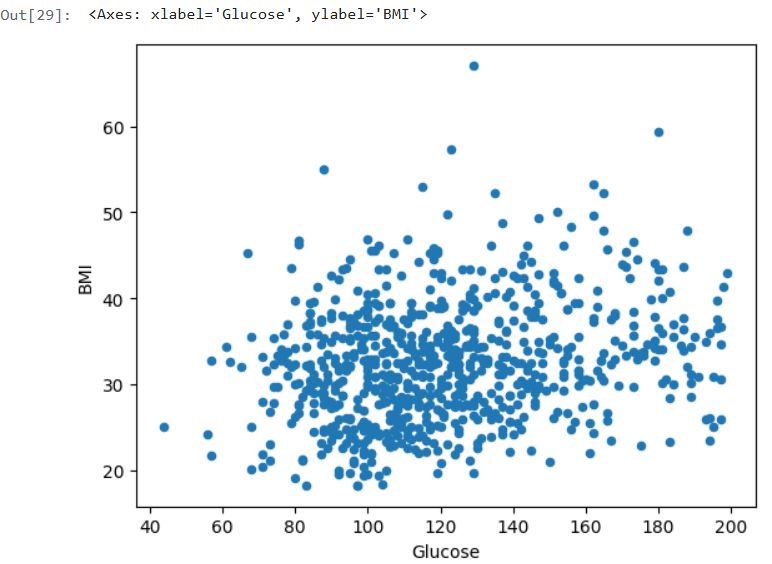
 Kết quả: mô hình dự đoán đúng 75% trên tập test.

* **Comparison graphs**

Scatterplots: biểu đồ phân tán

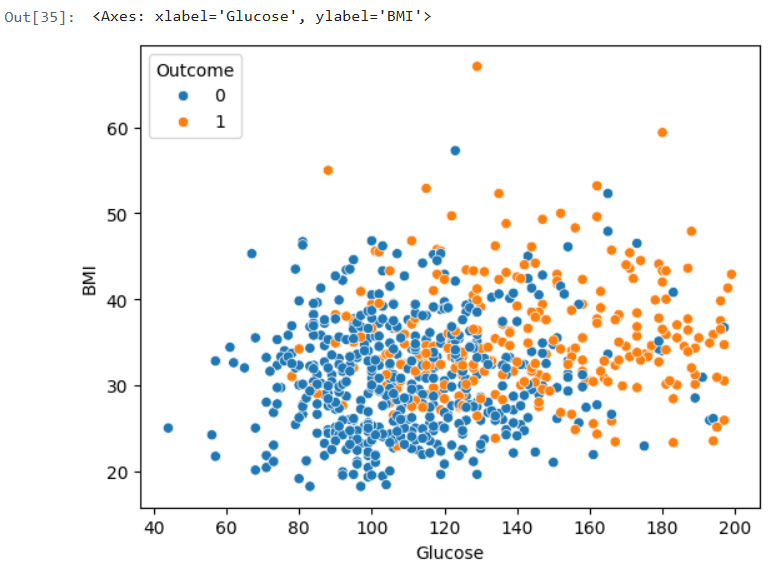
*#Scatterplot using pandas*

df.plot.scatter('Glucose','BMI')



Biểu đồ phân tán 2 gtri và phân biệt bằng Outcome

sns.scatterplot(x = 'Glucose', y = 'BMI', hue=df['Outcome'], data=df)



What we can conclude:

* Patients with diabetes (Outcome = 1) tend to have higher glucose levels, BMI, and age compared to non-diabetic patients.
* Non-diabetic cases (Outcome = 0) usually show lower glucose and BMI values, and often fall in the younger age range.
* Some features such as Insulin and SkinThickness show large variation but still indicate higher median values in diabetic cases.
* As seen above, different visualization tools could be used; from now on, only seaborn plots are demonstrated, while others are left for you to explore.

***Bệnh nhân mắc tiểu đường (Outcome = 1)*** *thường có mức đường huyết (glucose), chỉ số BMI và độ tuổi cao hơn so với những người không mắc tiểu đường.*

***Các trường hợp không mắc tiểu đường (Outcome = 0)*** *thường có giá trị glucose và BMI thấp hơn, đồng thời phần lớn rơi vào nhóm tuổi trẻ hơn.*

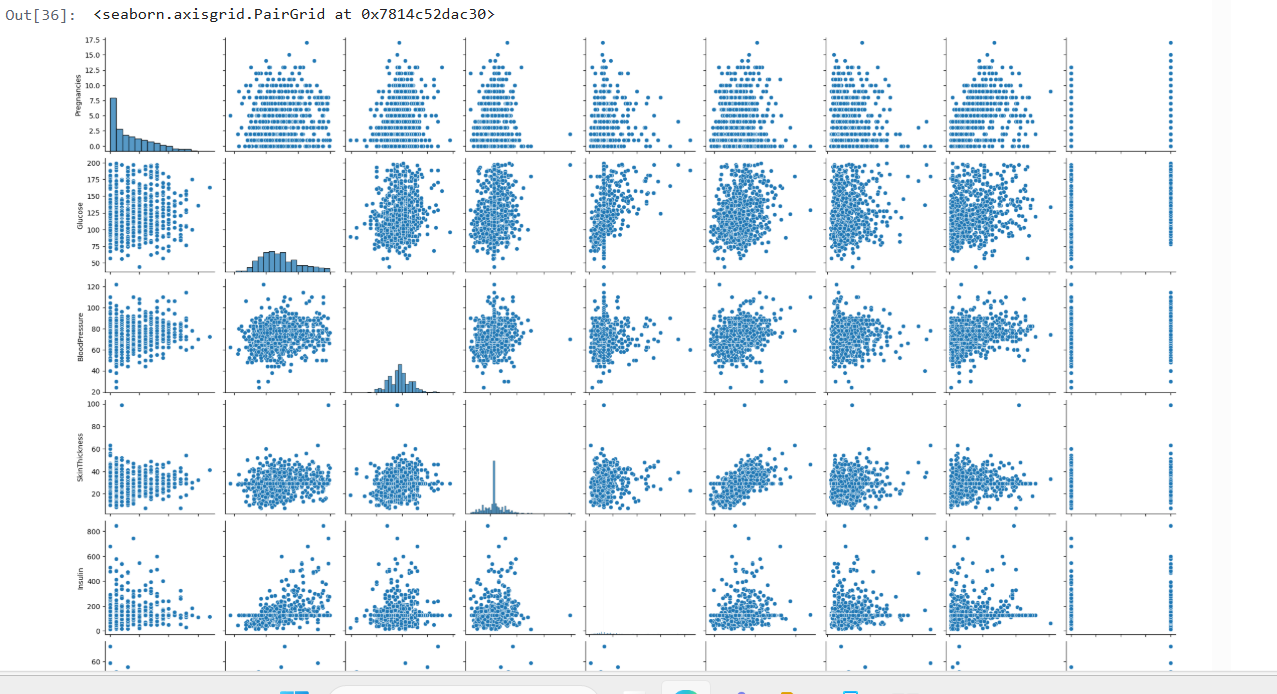
*Một số đặc trưng như* ***Insulin*** *và* ***SkinThickness*** *có độ biến thiên lớn, nhưng nhìn chung vẫn thể hiện giá trị trung vị cao hơn ở nhóm bệnh nhân tiểu đường.*

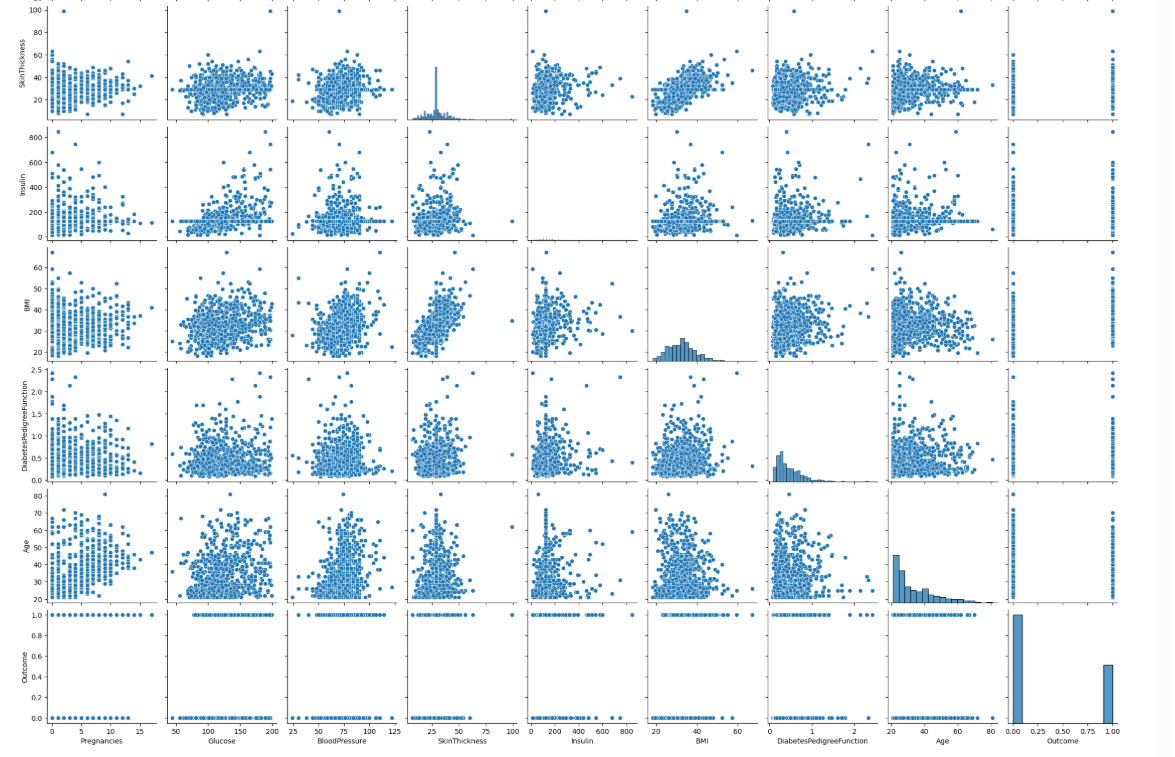
*Như đã thấy ở trên, có thể sử dụng nhiều công cụ trực quan hóa khác nhau; từ đây trở đi chỉ minh họa bằng* ***biểu đồ Seaborn****, còn các công cụ khác để bạn tự khám phá thêm.*

* **Correlogram**

Ma trận scatterplot giữa tất cả các cặp biến số

sns.pairplot(df)





What we can conclude:

* Two clusters can be observed when comparing glucose and BMI: diabetic patients generally lie in the higher range, while non-diabetic patients cluster in the lower range.
* Patients with diabetes (Outcome = 1) tend to have higher glucose levels, BMI, insulin, and age.
* Non-diabetic patients (Outcome = 0) mostly show lower glucose and BMI values, with relatively younger age.
* Diabetic cases exhibit clear differences in their characteristics compared to non-diabetic ones: they have higher glucose and BMI, while non-diabetic individuals show more normal values in these measures.
* Some features like BloodPressure and SkinThickness lie in between, showing average variations across both groups.

*Khi so sánh giữa* ***glucose và BMI****, có thể quan sát thấy hai cụm rõ rệt: bệnh nhân tiểu đường thường nằm ở khoảng giá trị cao hơn, trong khi những người không mắc tiểu đường tập trung ở khoảng thấp hơn.*

***Bệnh nhân tiểu đường (Outcome = 1)*** *có xu hướng có mức glucose, BMI, insulin và độ tuổi cao hơn.*

***Người không mắc tiểu đường (Outcome = 0)*** *phần lớn thể hiện giá trị glucose và BMI thấp hơn, đồng thời có độ tuổi tương đối trẻ hơn.*

*Các trường hợp tiểu đường cho thấy sự khác biệt rõ rệt về đặc trưng so với nhóm không tiểu đường: họ có glucose và BMI cao hơn, trong khi người không tiểu đường thể hiện các giá trị bình thường hơn ở những chỉ số này.*

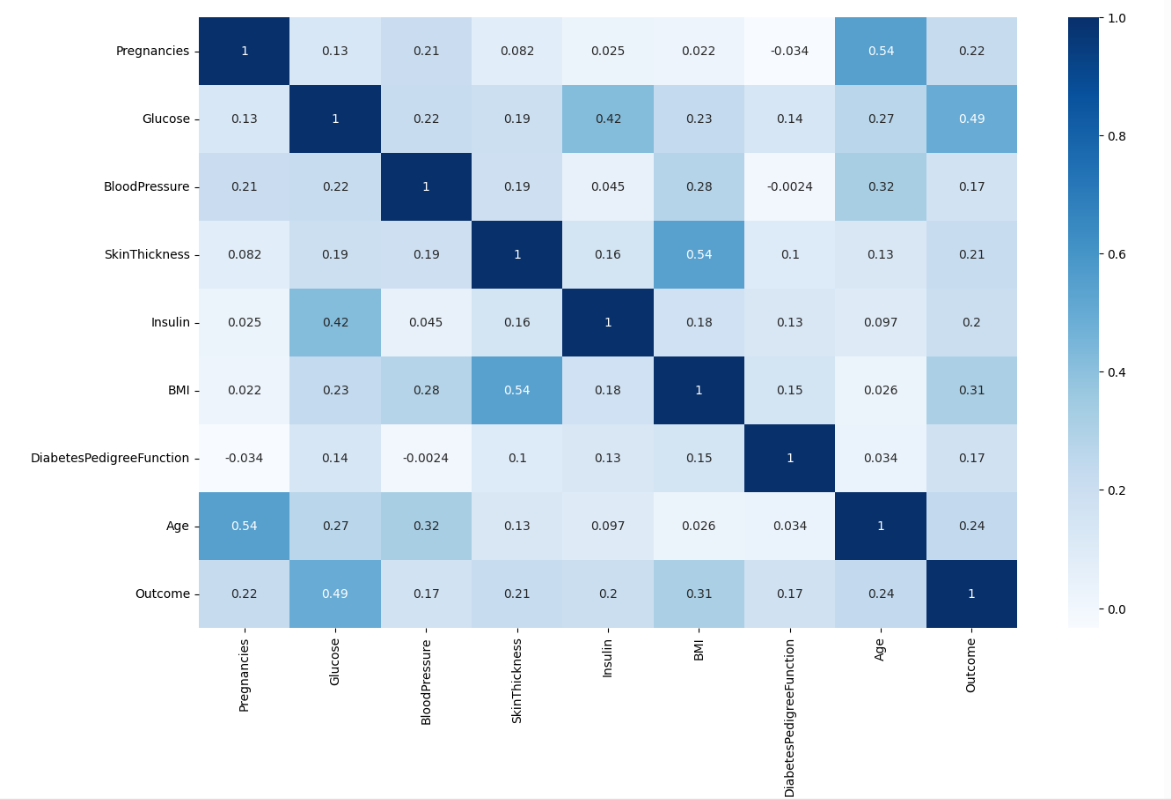
*Một số đặc trưng như* ***BloodPressure*** *và* ***SkinThickness*** *nằm ở mức trung gian, cho thấy sự biến thiên trung bình ở cả hai nhóm.*

* **Heatmap**

Biểu đồ nhiệt — thể hiện ma trận tương quan hoặc cường độ giá trị bằng màu sắc

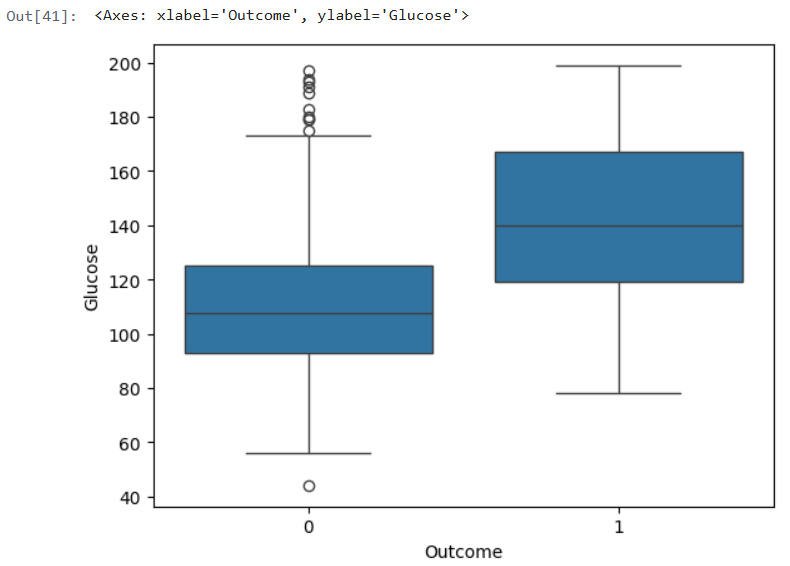
fig = plt.figure(figsize = (15,9))

sns.heatmap(df.corr(), cmap='Blues', annot = True);



Boxplot: biểu đồ hộp hiển thị phân bố giá trị *Glucose* cho từng nhóm Outcome.

sns.boxplot(x='Outcome', y='Glucose', data=df)

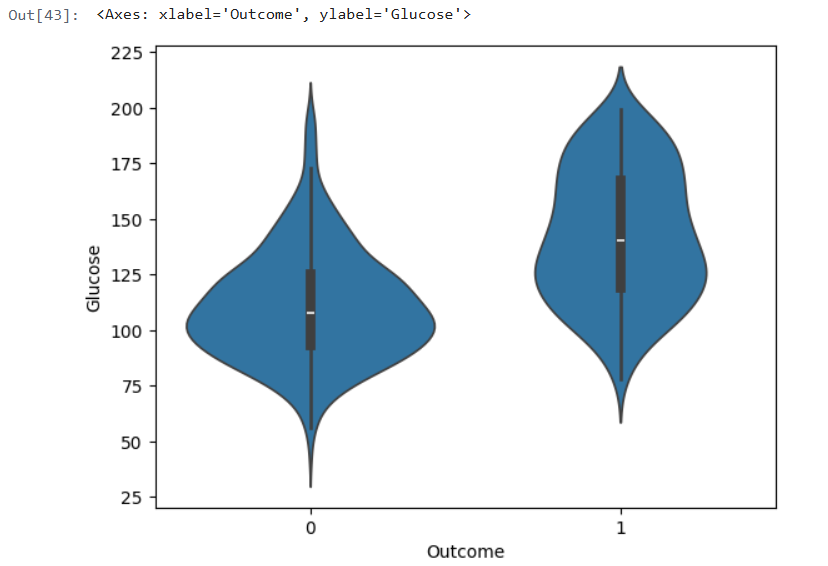


Violinplot: Biểu đồ kết hợp giữa **boxplot** và **biểu đồ mật độ (kernel density estimation – KDE)**.

Bên trong có thể có đường median và tứ phân vị giống boxplot

Hình dạng "cây đàn violin" hai bên biểu diễn mật độ phân bố của dữ liệu:

* Phần rộng hơn → nhiều dữ liệu tập trung.
* Phần hẹp hơn → ít dữ liệu.



# 5. Kết Luận

Ba tài liệu (WHO 1999, ADAP study, NDDG 1979) đã phác họa bức tranh toàn diện về định nghĩa, chẩn đoán, phân loại và dự báo bệnh đái tháo đường từ góc độ y học lâm sàng, dịch tễ học và ứng dụng công nghệ.

Thứ nhất, WHO 1999 đã chuẩn hóa tiêu chuẩn chẩn đoán và phân loại tiểu đường (Type 1, Type 2, GDM, IGT, IFG…) với những chỉ số then chốt như glucose huyết tương lúc đói, glucose sau OGTT 2 giờ và HbA1c. Đây là nền tảng để thống nhất trong lâm sàng và y tế công cộng, đồng thời chỉ ra sự đa dạng về nguyên nhân (tự miễn, di truyền, thuốc, bệnh tụy, nội tiết…).

Thứ hai, nghiên cứu ADAP (1980s–1990s) cho thấy khả năng ứng dụng các mô hình học máy cổ điển để dự báo sớm nguy cơ mắc đái tháo đường type 2. Với dữ liệu cộng đồng Pima Indians, ADAP đạt độ nhạy và đặc hiệu khá cân bằng (0.76), đồng thời đặt nền móng cho việc đối chiếu nhiều mô hình (logistic regression, perceptron) và khẳng định giá trị của ROC trong đánh giá y sinh học. Điều này mở ra triển vọng sử dụng AI/ML trong sàng lọc cộng đồng, phát hiện sớm và can thiệp kịp thời.

Thứ ba, NDDG 1979 đã khẳng định tiểu đường là một hội chứng không đồng nhất, nhấn mạnh sự cần thiết của một hệ thống phân loại quốc tế để tránh mâu thuẫn trong nghiên cứu. Bằng việc đưa ra phân loại Type I, Type II, GDM, IGT…, tài liệu này là nền tảng cho sự phát triển các tiêu chuẩn WHO và ADA sau này.

Từ ba hướng tiếp cận này, có thể rút ra một số điểm chung quan trọng:

* Tiểu đường là bệnh đa nguyên nhân, đa hình thái, đòi hỏi phân loại chi tiết để điều trị và dự phòng.
* Việc chuẩn hóa tiêu chuẩn chẩn đoán có ý nghĩa sống còn trong nghiên cứu và y tế cộng đồng.
* Các yếu tố nguy cơ di truyền, lối sống, béo phì, kháng insulin luôn được nhấn mạnh xuyên suốt.
* Ứng dụng mô hình dự báo giúp nâng cao hiệu quả sàng lọc, giảm chi phí xét nghiệm, phát hiện sớm và đưa ra các chiến lược phòng ngừa.

Như vậy, báo cáo đã cho thấy sự tiến hóa từ phân loại – chẩn đoán lâm sàng đến ứng dụng trí tuệ nhân tạo trong dự báo, phản ánh xu thế kết hợp y học truyền thống với khoa học dữ liệu hiện đại. Đây cũng chính là cơ sở quan trọng để nghiên cứu, phân tích và áp dụng vào thực tiễn quản lý bệnh đái tháo đường hiện nay.