

**TRƯỜNG ĐẠI HỌC SÀI GÒN**

**KHOA CÔNG NGHỆ THÔNG TIN**

----------

****

----------

**BÁO CÁO**

**Môn học: Phân tích dữ liệu**

**Tên đề tài: Phân tích khám phá về bệnh đái tháo đường (Diabetes Mellitus)**

**Giảng viên hướng dẫn:** Đỗ Như Tài

**Thành viên:**

3122410039 - Bùi Đức Chiến

3122410012 - Từ Nhật Anh

3122410005 - Trần Khải An

3122410049 - Trần Thị Kiều Diễm

*Thành phố Hồ Chí Minh – Tháng 9/2025*

# **LỜI MỞ ĐẦU**

Trong những thập kỷ gần đây, bệnh đái tháo đường (diabetes mellitus), đặc biệt là tiểu đường type 2, đã trở thành một vấn đề y tế công cộng toàn cầu, với tỷ lệ người mắc bệnh tăng nhanh tại nhiều quốc gia. Tiểu đường không chỉ ảnh hưởng đến chất lượng cuộc sống của người bệnh mà còn làm gia tăng nguy cơ mắc các biến chứng nghiêm trọng về tim mạch, thận, mắt và thần kinh, gây gánh nặng lớn cho hệ thống y tế. Việc phát hiện sớm và dự đoán nguy cơ mắc tiểu đường có ý nghĩa quan trọng trong việc phòng ngừa, kiểm soát và điều trị bệnh.

Bộ dữ liệu Pima Indians Diabetes, được xây dựng từ nghiên cứu trên cộng đồng người Pima tại bang Arizona (Hoa Kỳ), là một trong những bộ dữ liệu kinh điển trong nghiên cứu dịch tễ học và học máy. Bộ dữ liệu này bao gồm các đặc trưng lâm sàng và nhân khẩu học cơ bản, qua đó phản ánh mối quan hệ giữa các yếu tố nguy cơ và khả năng mắc bệnh tiểu đường type 2.

Trong phạm vi đề tài này, chúng tôi thực hiện phân tích khám phá dữ liệu (Exploratory Data Analysis – EDA) trên bộ dữ liệu Pima Indians Diabetes nhằm: (i) tìm hiểu đặc điểm và chất lượng dữ liệu, (ii) phát hiện các vấn đề dữ liệu còn tồn tại, (iii) phân tích sự phân bố và mối quan hệ giữa các biến, và (iv) rút ra ý nghĩa y học cũng như định hướng ứng dụng trong học máy. Nghiên cứu không chỉ giúp làm rõ vai trò của các yếu tố nguy cơ kinh điển như Glucose, BMI, tuổi, số lần mang thai, mà còn khẳng định tầm quan trọng của EDA trong xử lý và khai thác dữ liệu y tế**.**

# **LỜI CẢM ƠN**

Trước tiên, nhóm nghiên cứu xin bày tỏ lòng biết ơn sâu sắc đến thầy Đỗ Như Tài, người đã tận tình hướng dẫn, truyền đạt kiến thức, định hướng phương pháp và hỗ trợ chúng tôi trong suốt quá trình thực hiện đề tài. Sự tận tâm và những góp ý quý báu của thầy là nền tảng quan trọng giúp nhóm hoàn thiện báo cáo này.

Chúng tôi cũng xin chân thành cảm ơn các tác giả và tổ chức đã xây dựng và công bố bộ dữ liệu Pima Indians Diabetes, tạo điều kiện cho cộng đồng nghiên cứu có một nguồn dữ liệu giá trị để học tập, phân tích và phát triển các mô hình ứng dụng trong y học.

Cuối cùng, xin gửi lời cảm ơn đến tất cả các thành viên trong nhóm đã cùng nhau hợp tác, chia sẻ ý tưởng, phân công công việc hợp lý và nỗ lực hết mình để hoàn thành đề tài.

Mặc dù đã cố gắng hết sức, song do hạn chế về kiến thức và thời gian, báo cáo chắc chắn khó tránh khỏi thiếu sót. Nhóm rất mong nhận được sự góp ý từ thầy cô và bạn bè để nghiên cứu ngày càng hoàn thiện hơn.

# **BẢNG PHÂN CÔNG**

|  |  |  |
| --- | --- | --- |
| **MSSV** | **Họ Tên** | **Công việc** |
| 3122410039 | Bùi Đức Chiến - Nhóm trưởng | Phân tích khám phá dữ liệu, Tìm hiểu Paper3, PPT |
| 312241005 | Trần Khải An | Phân tích khám phá dữ liệu, Tìm hiểu Paper2, PPT |
| 3122410012 | Từ Nhật Anh | Phân tích khám phá dữ liệu, Tìm hiểu Paper1, PPT |
| 3122410049 | Trần Thị Kiều Diễm | Phân tích khám phá dữ liệu, Tổng hợp Paper |

**MỤC LỤC**

[LỜI MỞ ĐẦU 2](#_Toc4992)

[LỜI CẢM ƠN 3](#_Toc15811)

[BẢNG PHÂN CÔNG 4](#_Toc12591)

[I. Giới thiệu sơ lược về đề tài báo cáo 7](#_Toc6527)

[II. Cơ sở lý thuyết / Nghiên cứu liên quan 8](#_Toc3400)

[II.1. Giới thiệu về bệnh tiểu đường 8](#_Toc3699)

[II.2. Cộng đồng Pima Indians và ý nghĩa nghiên cứu 9](#_Toc3409)

[II.3. Bộ dữ liệu Pima Indians Diabetes 10](#_Toc456)

[II.4. Ý nghĩa của các đặc trưng trong dữ liệu 11](#_Toc10405)

[II.5. Các nghiên cứu liên quan 13](#_Toc24820)

[II.5.1 Nghiên cứu 1 (Paper 1): WHO (1999) – “Definition, Diagnosis and Classification of Diabetes Mellitus and its Complications” 13](#_Toc17472)

[II.5.2 Nghiên cứu 2: Smith et al. (1988) – “Using the ADAP learning algorithm to forecast the onset of diabetes mellitus” 15](#_Toc31972)

[II.5.3 Nghiên cứu 3: Hệ thống phân loại bệnh tiểu đường (Classification System for Diabetes and Glucose Tolerance Abnormalities) 17](#_Toc21968)

[II.5.4 Tổng Kết Nghiên cứu 19](#_Toc819)

[III. Dữ liệu và phương pháp 20](#_Toc20614)

[III.1. Dữ liệu 20](#_Toc11848)

[III.2. Vấn đề dữ liệu và tiền xử lý sơ bộ 21](#_Toc12605)

[III.3. Phương pháp phân tích dữ liệu (EDA) 22](#_Toc19295)

[III.4. Công cụ và môi trường phân tích 24](#_Toc27879)

[IV. Phân tích dữ liệu (EDA) 26](#_Toc28091)

[IV.1. Thống kê mô tả 26](#_Toc23398)

[IV.1.1. Kích thước và kiểu dữ liệu 26](#_Toc8349)

[IV.1.2. 5 dòng đầu và 5 dòng cuối 26](#_Toc19009)

[IV.1.3. Thống kê cơ bản (Summary Statistics) 27](#_Toc13741)

[IV.1.4. Phân bố biến mục tiêu (Outcome) 27](#_Toc32203)

[IV.2. Kiểm tra dữ liệu thiếu và bất hợp lý 28](#_Toc31741)

[Bảng Thống Kê Số Lượng Giá Trị 0 Bất Hợp Lý Trong Từng Biến 29](#_Toc23735)

[IV.3. Phân tích phân phối các biến (Univariate Analysis) 31](#_Toc19403)

[IV.3.1. Pregnancies (Số lần mang thai) 31](#_Toc8503)

[IV.3.2. Glucose (Nồng độ glucose huyết) 32](#_Toc30124)

[IV.3.3. BloodPressure (Huyết áp tâm trương) 33](#_Toc1038)

[IV.3.4. SkinThickness (Độ dày nếp gấp da) 34](#_Toc22480)

[IV.3.5. Insulin (Nồng độ insulin sau 2 giờ) 35](#_Toc14660)

[IV.3.6. BMI (Body Mass Index) 36](#_Toc12908)

[IV.3.7. DiabetesPedigreeFunction (Chỉ số nguy cơ di truyền) 37](#_Toc4951)

[IV.3.8. Age (Tuổi bệnh nhân) 38](#_Toc23710)

[IIV.4. So sánh nhóm mắc và không mắc (Bivariate Analysis) 39](#_Toc2337)

[IV.4.1 Thống kê mô tả phân tách theo Outcome 39](#_Toc17730)

[IV.4.2 Violin plot 5 thuộc tính 40](#_Toc31695)

[IV.5. Phân tích đa biến và tương quan 41](#_Toc18317)

[IV.5.1 Heatmap tương quan giữa các biến 41](#_Toc17811)

[IV.5.2 Pairplot (Scatter matrix) 42](#_Toc22894)

[IV.5.3 Nhận xét tổng quan: 43](#_Toc7939)

[IV.6. Outliers (giá trị ngoại lai) 44](#_Toc2563)

[IV.6.1 Boxplot được sử dụng để phát hiện giá trị ngoại lai trong các biến 44](#_Toc6202)

[IV.7. Kết quả tổng hợp từ EDA 45](#_Toc14496)

[IV.7.1. Chất lượng dữ liệu 45](#_Toc3128)

[IV.7.2. Phân bố dữ liệu (Univariate Analysis) 45](#_Toc28997)

[IV.7.3. So sánh theo Outcome (Bivariate Analysis) 45](#_Toc26646)

[IV.7.4. Phân tích đa biến (Multivariate Analysis) 46](#_Toc361)

[IV.7.5. Tổng hợp 46](#_Toc11529)

[IV.8. Xử lý dữ liệu thiếu và bất hợp lý (Data Cleaning) 47](#_Toc12428)

[IV.8.1. Nguyên tắc xử lý 47](#_Toc2890)

[IV.8.2. Các bước xử lý chi tiết 47](#_Toc8688)

[IV.8.3. Nhận xét sau khi hoàn thành xử lý 48](#_Toc9554)

[V. Thảo luận kết quả 49](#_Toc20828)

[V.1. Ý nghĩa y học và thực tiễn 49](#_Toc17951)

[V.1.1. Ý nghĩa y học 49](#_Toc27366)

[V.1.2. Ý nghĩa thực tiễn 49](#_Toc27213)

[V.2. Hạn chế của dữ liệu 51](#_Toc32273)

[V.2.1. Giới hạn về đối tượng nghiên cứu 51](#_Toc15179)

[V.2.2. Vấn đề về dữ liệu thiếu và bất hợp lý 51](#_Toc20156)

[V.2.3. Hạn chế về số lượng biến 51](#_Toc21755)

[V.2.4. Hạn chế về tính thời gian (temporal aspect) 51](#_Toc23227)

[V.2.5. Mất cân bằng lớp (class imbalance) 52](#_Toc12660)

[V.3. Định hướng ứng dụng trong học máy 53](#_Toc8467)

[V.3.1. Bài toán phân loại nhị phân (Binary Classification) 53](#_Toc13254)

[V.3.2. Phát hiện đặc trưng quan trọng (Feature Importance) 53](#_Toc24619)

[V.3.3. Xử lý dữ liệu mất cân bằng (Imbalanced Learning) 53](#_Toc20976)

[V.3.4. Dự báo nguy cơ và hệ thống hỗ trợ quyết định (Decision Support Systems) 54](#_Toc27252)

[V.3.5. Định hướng mở rộng 54](#_Toc8036)

[VI. Kết luận 55](#_Toc27982)

[VI.1. Kết quả chính 55](#_Toc17615)

[VI.2. Ý nghĩa y học và thực tiễn 56](#_Toc9281)

[VI.3. Hạn chế của dữ liệu 56](#_Toc5886)

[VI.4. Định hướng nghiên cứu và ứng dụng 56](#_Toc698)

[VI.5 Kết luận cuối cùng 56](#_Toc19995)

[TÀI LIỆU THAM KHẢO 57](#_Toc28907)

# 

# **I. Giới thiệu sơ lược về đề tài báo cáo**

Bệnh tiểu đường (Diabetes Mellitus) là một trong những căn bệnh mạn tính phổ biến và nguy hiểm nhất trên thế giới. Theo thống kê của Tổ chức Y tế Thế giới (WHO), có hơn 537 triệu người (2021) mắc tiểu đường, và con số này dự kiến còn tiếp tục gia tăng. Trong đó, **tiểu đường loại II (Type 2 Diabetes)** chiếm tỷ lệ lớn, thường xuất hiện ở người trưởng thành, liên quan mật thiết đến **thói quen ăn uống, mức độ béo phì và di truyền**.

Việc chẩn đoán sớm bệnh tiểu đường có vai trò quan trọng trong **giảm biến chứng**, **giảm chi phí y tế** và **nâng cao chất lượng cuộc sống**. Do đó, phân tích dữ liệu và ứng dụng học máy trong y tế là xu hướng cần thiết.

**Pima Indians Diabetes Dataset** là một tập dữ liệu y tế nổi tiếng, được thu thập từ phụ nữ người Pima tại bang Arizona (Mỹ). Cộng đồng này có tỷ lệ mắc bệnh tiểu đường cao bất thường, khiến dữ liệu của họ trở thành đối tượng nghiên cứu lý tưởng cho các mô hình dự đoán.

**Mục tiêu của báo cáo**:

1. Thực hiện phân tích dữ liệu khám phá (EDA) trên bộ dữ liệu Pima Indians Diabetes.
2. Xác định các đặc trưng quan trọng liên quan đến nguy cơ mắc bệnh.
3. Rút ra những nhận định và định hướng cho việc xây dựng mô hình dự đoán.

# 

# 

# **II. Cơ sở lý thuyết / Nghiên cứu liên quan**

## **II.1. Giới thiệu về bệnh tiểu đường**

Bệnh tiểu đường (Diabetes Mellitus) là một nhóm bệnh rối loạn chuyển hóa đặc trưng bởi tình trạng **tăng đường huyết mạn tính**, xảy ra do cơ thể **không sản xuất đủ insulin** hoặc **không sử dụng insulin một cách hiệu quả**. Insulin là hormone do tuyến tụy tiết ra, giúp vận chuyển glucose từ máu vào tế bào để cung cấp năng lượng.

Theo **Tổ chức Y tế Thế giới (WHO)**, bệnh tiểu đường đang là một trong những thách thức y tế lớn nhất toàn cầu. Năm 2021, có khoảng **537 triệu người trưởng thành (20–79 tuổi)** mắc bệnh, con số này dự kiến sẽ tăng lên **643 triệu vào năm 2030**. Đây là nguyên nhân hàng đầu gây ra nhiều biến chứng nghiêm trọng như: **bệnh tim mạch, suy thận, mù lòa và hoại tử chi**, làm giảm tuổi thọ và tăng gánh nặng kinh tế – xã hội.

Các loại bệnh tiểu đường chính:

**1. Tiểu đường type 1**

* Là bệnh **tự miễn dịch**, hệ miễn dịch nhầm lẫn và phá hủy các tế bào beta tuyến tụy – nơi sản xuất insulin.
* Hậu quả: cơ thể **gần như không còn insulin**, người bệnh buộc phải tiêm insulin suốt đời.
* Thường xuất hiện ở trẻ em, thanh thiếu niên hoặc người trưởng thành trẻ tuổi, chiếm khoảng **5–10%** số ca bệnh.

**2. Tiểu đường type 2**

* Là dạng phổ biến nhất (**90–95%** các trường hợp).
* Cơ thể vẫn sản xuất insulin nhưng các tế bào không đáp ứng tốt với insulin (**kháng insulin**), và theo thời gian tuyến tụy có thể suy giảm khả năng tiết insulin.
* Yếu tố nguy cơ: **béo phì, lối sống ít vận động, chế độ dinh dưỡng không lành mạnh, tuổi tác cao, tiền sử gia đình**.
* Diễn biến âm thầm, nhiều trường hợp chỉ được phát hiện khi đã có biến chứng.

**3. Tiểu đường thai kỳ (Gestational Diabetes)**

* Xuất hiện ở phụ nữ mang thai, khi cơ thể không đáp ứng đủ với insulin do sự thay đổi hormone.
* Thường được phát hiện vào tam cá nguyệt thứ hai hoặc thứ ba.
* Có thể biến mất sau sinh, nhưng làm tăng nguy cơ phát triển **type 2** ở cả mẹ và con trong tương lai.

Trong nghiên cứu này, chúng ta tập trung vào **tiểu đường type 2**, do có mối liên hệ chặt chẽ với **tuổi tác, lối sống, mức độ béo phì và yếu tố di truyền** – cũng chính là các yếu tố được phản ánh trong bộ dữ liệu **Pima Indians Diabetes**.

## **II.2. Cộng đồng Pima Indians và ý nghĩa nghiên cứu**

Cộng đồng **Pima Indians** là một nhóm người bản địa sinh sống chủ yếu tại bang **Arizona (Hoa Kỳ)** và một phần tại **Sonora (Mexico)**. Trong nhiều thập kỷ, cộng đồng này được quan tâm đặc biệt trong nghiên cứu y tế, do có **tỷ lệ mắc bệnh tiểu đường type 2 cao nhất thế giới**.

Một số đặc điểm nổi bật của cộng đồng Pima Indians:

* Tỷ lệ béo phì và tiểu đường type 2 cực kỳ cao, có giai đoạn lên tới **hơn 50% người trưởng thành** mắc bệnh.
* Sự thay đổi lối sống từ truyền thống (săn bắn, nông nghiệp) sang hiện đại (chế độ ăn giàu tinh bột tinh chế, ít vận động) được coi là nguyên nhân chính thúc đẩy dịch bệnh này.
* Nhiều nghiên cứu dịch tễ học và di truyền học đã chỉ ra rằng **yếu tố di truyền kết hợp với môi trường sống** là nguyên nhân cốt lõi làm gia tăng nguy cơ tiểu đường ở nhóm dân cư này.

**Ý nghĩa nghiên cứu trên cộng đồng Pima Indians:**

1. **Y học và dịch tễ học:** Giúp hiểu rõ cơ chế phát sinh tiểu đường type 2, đặc biệt là mối quan hệ giữa di truyền – lối sống – dinh dưỡng.
2. **Khoa học dữ liệu và học máy:** Bộ dữ liệu Pima Indians Diabetes được coi là một trong những **dataset chuẩn** để kiểm chứng các thuật toán học máy trong phân loại nhị phân (classification problem).
3. **Ứng dụng thực tiễn:** Kết quả nghiên cứu có thể được ứng dụng để xây dựng mô hình dự đoán nguy cơ tiểu đường, từ đó hỗ trợ **chẩn đoán sớm, sàng lọc và dự phòng** trong cộng đồng có nguy cơ cao.

## **II.3. Bộ dữ liệu Pima Indians Diabetes**

Bộ dữ liệu **Pima Indians Diabetes** được xây dựng từ một nghiên cứu dịch tễ học lớn do **National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK)** thực hiện. Dataset này hiện nay được sử dụng rộng rãi trong cộng đồng khoa học dữ liệu và học máy như một bộ dữ liệu mẫu cho các bài toán **phân loại nhị phân (binary classification)**.

**Đặc điểm của bộ dữ liệu:**

* **Số lượng mẫu:** 768 dòng dữ liệu (ứng với 768 bệnh nhân nữ).
* **Đối tượng:** Tất cả bệnh nhân đều là nữ, **≥ 21 tuổi**, thuộc cộng đồng Pima Indians sinh sống ở Arizona (Hoa Kỳ).
* **Biến mục tiêu (Target):**
  + **Outcome** = 1: dương tính với tiểu đường.
  + **Outcome** = 0: âm tính với tiểu đường.

→ Cho thấy **tỉ lệ mất cân bằng lớp** (imbalanced data), cần được chú ý khi xây dựng mô hình.

**Các biến đầu vào (Predictor Variables):**

1. **Pregnancies** – Số lần mang thai.
2. **Glucose** – Nồng độ glucose trong huyết tương sau 2 giờ (mg/dL) trong nghiệm pháp dung nạp glucose.
3. **BloodPressure** – Huyết áp tâm trương (mm Hg).
4. **SkinThickness** – Độ dày nếp gấp da (mm).
5. **Insulin** – Nồng độ insulin huyết thanh sau 2 giờ (µU/mL).
6. **BMI** – Chỉ số khối cơ thể, tính theo công thức: BMI = Cân nặng(kg) /Chiều cao (m)
7. **DiabetesPedigreeFunction (DPF)** – Hàm số thể hiện **yếu tố di truyền** nguy cơ mắc tiểu đường.
8. **Age** – Tuổi của bệnh nhân (năm).

**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**

Mỗi biến trong bộ dữ liệu Pima Indians Diabetes đều được chọn lọc dựa trên các yếu tố y học có liên quan đến nguy cơ mắc tiểu đường type 2. Ý nghĩa cụ thể của từng đặc trưng như sau:

1. **Pregnancies (Số lần mang thai):**
   * Phụ nữ từng mang thai nhiều lần có nguy cơ cao phát triển **tiểu đường thai kỳ**.
   * Tiểu đường thai kỳ có thể biến mất sau sinh, nhưng làm tăng khả năng phát triển **tiểu đường type 2** trong tương lai.
   * Do đó, số lần mang thai được xem là **chỉ số nhân khẩu học** liên quan đến nguy cơ mắc bệnh.
2. **Glucose (Nồng độ glucose huyết tương sau 2 giờ):**
   * Đây là một trong những chỉ số chẩn đoán **trực tiếp nhất** của bệnh tiểu đường.
   * Theo **WHO**, nếu nồng độ glucose sau 2 giờ ≥ 200 mg/dL trong nghiệm pháp dung nạp glucose, bệnh nhân được coi là mắc tiểu đường.
   * Vì vậy, biến này có mối liên hệ chặt chẽ nhất với biến mục tiêu (Outcome).
3. **BloodPressure (Huyết áp tâm trương):**
   * Bệnh nhân tiểu đường thường có nguy cơ cao bị **tăng huyết áp**.
   * Tăng huyết áp kết hợp với tăng đường huyết làm gia tăng biến chứng tim mạch.
   * Tuy nhiên, biến này có thể bị ảnh hưởng bởi nhiều yếu tố khác (lối sống, stress), nên mức tương quan với Outcome thường thấp hơn Glucose và BMI.
4. **SkinThickness (Độ dày nếp gấp da):**
   * Được dùng như một chỉ số gián tiếp để đo **mức độ mỡ dưới da**.
   * Thừa mỡ cơ thể liên quan chặt chẽ đến tình trạng kháng insulin và nguy cơ tiểu đường type 2.
   * Tuy nhiên, trong dataset có nhiều giá trị bằng 0 không thực tế → cần xử lý trước khi phân tích.
5. **Insulin (Nồng độ insulin sau 2 giờ):**
   * Là chỉ số quan trọng để đánh giá **khả năng tiết insulin của tuyến tụy**.
   * Nồng độ insulin bất thường có thể phản ánh tình trạng **kháng insulin**, đặc trưng của tiểu đường type 2.
   * Trong dataset, biến này chứa nhiều giá trị 0, phản ánh dữ liệu thiếu chứ không phải giá trị thực.
6. **BMI (Body Mass Index)**
   * Thừa cân, béo phì (BMI ≥ 25 hoặc ≥ 30) là một trong những yếu tố nguy cơ cao nhất của tiểu đường type 2.
   * Chỉ số này phản ánh trực tiếp **mức độ tích tụ mỡ trong cơ thể** và mối quan hệ với đề kháng insulin.
   * Do đó, BMI thường có mức tương quan cao thứ hai với Outcome (sau Glucose).
7. **DiabetesPedigreeFunction (DPF – Chỉ số nguy cơ di truyền)**
   * Là một chỉ số tổng hợp phản ánh **tiền sử gia đình** và yếu tố di truyền liên quan đến tiểu đường.
   * Người có cha/mẹ hoặc anh/chị em mắc tiểu đường có nguy cơ cao hơn.
   * Biến này cung cấp thông tin độc lập, không đo được qua các chỉ số sinh lý.
8. **Age (Tuổi)**
   * Nguy cơ mắc tiểu đường type 2 tăng theo tuổi, đặc biệt từ 40 trở lên.
   * Đây là một yếu tố nhân khẩu học quan trọng, thường được kết hợp với BMI và Glucose trong các nghiên cứu lâm sàng.

## 

## **II.5. Các nghiên cứu liên quan**

### **II.5.1 Nghiên cứu 1 (Paper 1): WHO (1999) – “Definition, Diagnosis and Classification of Diabetes Mellitus and its Complications”**

* **Bài toán và giải pháp:**Trước năm 1999, nhiều tổ chức y tế như **NDDG (1979), WHO (1980/1985), ADA (1997)** đưa ra các tiêu chuẩn chẩn đoán khác nhau, dẫn đến sự **không thống nhất** trong nghiên cứu dịch tễ và điều trị lâm sàng. WHO (1999) đã đưa ra một bộ tiêu chuẩn chuẩn hóa toàn cầu:
  + Chuẩn hóa định nghĩa: tiểu đường là rối loạn chuyển hóa do thiếu hoặc kháng insulin, dẫn đến tăng glucose máu mạn tính và biến chứng.
  + Điều chỉnh tiêu chuẩn chẩn đoán: hạ ngưỡng **FPG** từ 7.8 xuống **7.0 mmol/L (126 mg/dL)**.
  + Phân loại bệnh dựa trên cơ chế bệnh sinh: **Type 1, Type 2, Gestational, Other specific types**.
  + Bổ sung nhóm trung gian: **IFG (Impaired Fasting Glycaemia)**, **IGT (Impaired Glucose Tolerance)**.
  + Đưa vào khái niệm **Hội chứng chuyển hóa (Metabolic Syndrome)**.
* **Input:**
  + Xét nghiệm máu: **FPG, OGTT 2h**, và (chưa chuẩn hóa) **HbA1c**.
  + Dữ liệu lâm sàng: triệu chứng (khát, tiểu nhiều, sụt cân), tiền sử gia đình, béo phì, tuổi, thai kỳ.
* **Output:**
  + Phân loại bệnh nhân: **Normoglycaemia (bình thường), IFG/IGT (tiền tiểu đường), Diabetes Mellitus (Type 1, Type 2, Gestational, Other)**.
* **Mục tiêu:**
  + Chuẩn hóa tiêu chuẩn chẩn đoán trên toàn cầu.
  + Phát hiện sớm bệnh nhân nguy cơ cao để dự phòng biến chứng.
  + Làm cơ sở cho nghiên cứu dịch tễ và so sánh quốc tế.
  + Hỗ trợ lâm sàng trong điều trị và phân loại bệnh.
* **Nghiên cứu liên quan và nền tảng:**
  + **NDDG (1979):** Đưa ra chuẩn đầu tiên nhưng ngưỡng quá cao → bỏ sót nhiều ca bệnh.
  + **WHO (1980, 1985):** Tiêu chuẩn quốc tế đầu tiên nhưng còn thiếu nhất quán.
  + **ADA (1997):** Đề xuất hạ FPG xuống 7.0 mmol/L, trở thành nền tảng cho WHO 1999.
  + **DECODE Study Group (1998):** Khẳng định OGTT đặc biệt quan trọng ở người già và châu Á.
  + **Ramachandran et al. (1998):** Nghiên cứu ở Ấn Độ chứng minh nguy cơ tim mạch tăng ngay cả khi glucose thấp hơn chuẩn cũ.

**Kết luận:** WHO (1999) là bước ngoặt lớn, giúp thống nhất tiêu chuẩn chẩn đoán tiểu đường, đồng thời nhấn mạnh tầm quan trọng của **phát hiện sớm tiền tiểu đường (IFG/IGT)** trong phòng chống bệnh.

### **II.5.2 Nghiên cứu 2: Smith et al. (1988) – “Using the ADAP learning algorithm to forecast the onset of diabetes mellitus”**

* **Bối cảnh và mục tiêu:**Tiểu đường type 2 (NIDDM) là bệnh mạn tính phổ biến với tỷ lệ cao bất thường trong cộng đồng Pima Indians gần Phoenix, Arizona (Mỹ). Kể từ năm 1965, cộng đồng này được NIDDK nghiên cứu sâu rộng.  
   Paper này nhằm dự đoán nguy cơ phát triển tiểu đường trong vòng 5 năm dựa trên dữ liệu nhân khẩu học, lâm sàng và sinh học, đồng thời thử nghiệm các mô hình học máy.
* **Input (biến đầu vào): 8 biến y tế quan trọng:**
  + Pregnancies – số lần mang thai.
  + Glucose – nồng độ glucose huyết tương sau 2 giờ (OGTT).
  + BloodPressure – huyết áp tâm trương.
  + SkinThickness – độ dày nếp gấp da (mm).
  + Insulin – nồng độ insulin huyết thanh sau 2 giờ.
  + BMI – chỉ số khối cơ thể.
  + Diabetes Pedigree Function (DPF) – nguy cơ di truyền.
  + Age – tuổi bệnh nhân.
* **Output (biến mục tiêu):**
  + 0 → Không mắc tiểu đường trong vòng 5 năm.
  + 1 → Mắc tiểu đường trong vòng 5 năm.  
     → Đây là bài toán phân loại nhị phân.
* **Mục tiêu và phương pháp:**
  + Xây dựng mô hình dự đoán nguy cơ tiểu đường dựa trên 8 đặc trưng.
  + So sánh ba phương pháp:
    - ADAP algorithm (Adaptive Learning, dạng neural network thời kỳ đầu).
    - Logistic Regression.
    - Linear Perceptron.
  + **Các chỉ số đánh giá:**
    - Sensitivity (Độ nhạy): TP / (TP + FN).
    - Specificity (Độ đặc hiệu): TN / (TN + FP).
    - ROC curve & AUC: đo lường hiệu quả mô hình.
* **Kết quả:**
  + ADAP đạt 76% sensitivity và 76% specificity trên tập kiểm tra (192 ca bệnh).
  + Cho thấy khả năng dự báo khá tốt, nhưng không vượt trội so với Logistic Regression.
  + Nghiên cứu mở ra tiềm năng ứng dụng mô hình học máy trong y học, dù vẫn còn hạn chế về dữ liệu và hiệu quả.

**Kết luận:** Paper này là một trong những nghiên cứu sớm nhất ứng dụng thuật toán học máy (neural network) cho dự đoán bệnh tiểu đường, đặt nền móng cho nhiều nghiên cứu hiện đại sau này.

### **II.5.3 Nghiên cứu 3: Hệ thống phân loại bệnh tiểu đường (Classification System for Diabetes and Glucose Tolerance Abnormalities)**

* **Mục tiêu:**Paper này tập trung xây dựng một hệ thống phân loại (classification system) để hỗ trợ chẩn đoán và theo dõi bệnh nhân. Các mục tiêu chính:  
  + Chẩn đoán ban đầu: Phân loại bệnh nhân vào các nhóm: DM (Diabetes Mellitus), IGT (Impaired Glucose Tolerance), GDM (Gestational Diabetes Mellitus), PrevAGT (previous abnormal glucose tolerance), PotAGT (potential abnormal glucose tolerance), Normal.
  + Xác định subtype: Phân biệt bệnh nhân DM thuộc IDDM (Type 1) hay NIDDM (Type 2, obese / non-obese).
  + Chuẩn hóa quy trình lâm sàng: Giảm thiểu nguy cơ nhầm lẫn, bỏ sót hoặc chẩn đoán sai trong thực hành y khoa.
* **Input (dữ liệu đầu vào):**Bao gồm nhiều nhóm thông tin:
  + Xét nghiệm sinh hóa:
    - FPG (Fasting Plasma Glucose).
    - OGTT (Oral Glucose Tolerance Test): đo glucose tại các thời điểm 0, 30, 60, 90, 120 phút (và 180 phút nếu nghi GDM).
  + Thông tin lâm sàng:
    - Triệu chứng kinh điển: khát nhiều, tiểu nhiều, sụt cân, nhiễm toan ceton.
    - Tiền sử: mang thai, từng có GDM, từng rối loạn dung nạp glucose.
  + Thông tin nhân khẩu học: tuổi, giới, cân nặng, chiều cao (BMI), chủng tộc, tiền sử gia đình.
  + Thông tin bổ sung (nếu có): xét nghiệm HLA, islet cell antibody, insulin, C-peptide để phân biệt type 1 và type 2; thông tin về thuốc hoặc bệnh nền.
* **Output (nhãn phân loại):**
  + Chính:
    - Normal (bình thường).
    - IGT (rối loạn dung nạp glucose).
    - DM (thoả tiêu chuẩn tiểu đường).
    - GDM (tiểu đường thai kỳ).
    - PrevAGT (từng bất thường nhưng hiện bình thường).
    - PotAGT (nguy cơ cao, chưa bất thường).
  + Phụ:
    - Nếu DM → phân loại tiếp thành IDDM (Type 1) hoặc NIDDM (Type 2, obese / non-obese).
* **Quy tắc quyết định (Decision Rules):**
  + DM: FPG ≥ 140 mg/dL (ít nhất 2 lần) hoặc OGTT 2h ≥ 200 mg/dL, hoặc triệu chứng + glucose bất kỳ ≥ 200 mg/dL.
  + IGT: FPG < 140 mg/dL, OGTT 2h từ 140–199 mg/dL.
  + GDM: OGTT 100 g: ≥2 giá trị vượt ngưỡng (F ≥ 105, 1h ≥ 190, 2h ≥ 165, 3h ≥ 145 mg/dL).
  + PrevAGT / PotAGT: dựa trên tiền sử hoặc yếu tố nguy cơ.
* **Ý nghĩa nghiên cứu:**
  + Lâm sàng: Giúp phân loại bệnh nhân chính xác hơn, theo dõi tiến triển từ IGT → DM, và hỗ trợ quản lý bệnh nhân nguy cơ cao.
  + Nghiên cứu: Cung cấp dữ liệu chuẩn để đánh giá tỷ lệ mắc bệnh trong các nhóm và so sánh đặc điểm dịch tễ học.
  + Kỹ thuật: Có thể triển khai thành một Decision Support System hoặc áp dụng mô hình học máy để tự động hóa quá trình phân loại bệnh nhân.

**Kết luận:** Paper này nhấn mạnh tầm quan trọng của một hệ thống phân loại chi tiết, không chỉ dừng lại ở “có/không” mắc tiểu đường mà còn mở rộng sang tiền tiểu đường, tiểu đường thai kỳ, nguy cơ tiềm tàng, qua đó tạo nền tảng cho dự đoán và can thiệp sớm.

### **II.5.4 Tổng Kết Nghiên cứu**

Qua ba nghiên cứu tiêu biểu có thể thấy rằng:

* **WHO (1999)** đóng vai trò nền tảng khi đưa ra bộ tiêu chuẩn chẩn đoán và phân loại tiểu đường được sử dụng rộng rãi trên toàn cầu. Đây là cơ sở khoa học quan trọng để thống nhất chẩn đoán lâm sàng và nghiên cứu dịch tễ học.
* **Smith et al. (1988)** là một trong những nghiên cứu tiên phong ứng dụng **thuật toán học máy** (ADAP, Logistic Regression, Perceptron) nhằm dự đoán nguy cơ mắc tiểu đường dựa trên dữ liệu lâm sàng và nhân khẩu học. Nghiên cứu này đặt nền móng cho hướng đi mới trong việc sử dụng mô hình tính toán để hỗ trợ y học.
* **Hệ thống phân loại lâm sàng (Classification System)** cho thấy tầm quan trọng của việc phân loại chi tiết bệnh nhân không chỉ thành “có/không mắc tiểu đường”, mà còn nhận diện các nhóm trung gian như **IGT, GDM, PrevAGT, PotAGT**. Điều này mở ra cơ hội phát hiện sớm và theo dõi tiến triển bệnh.

Như vậy, các nghiên cứu trước đây 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.

Trong báo cáo này, chúng tôi tập trung vào **khai phá dữ liệu Pima Indians Diabetes** thông qua phương pháp **phân tích dữ liệu khám phá (Exploratory Data Analysis – EDA)**. Đây sẽ là bước tiền đề quan trọng trước khi tiến tới xây dựng và đánh giá các mô hình dự đoán ở giai đoạn tiếp theo.

# 

# **III. Dữ liệu và phương pháp**

## **III.1. Dữ liệu**

Nghiên cứu sử dụng bộ dữ liệu **Pima Indians Diabetes Dataset**, được công bố bởi *National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK)*. Bộ dữ liệu này được dùng rộng rãi trong lĩnh vực khoa học dữ liệu và y học để nghiên cứu khả năng dự đoán bệnh tiểu đường loại II.

* **Đối tượng**: phụ nữ người Pima (Arizona, Hoa Kỳ), độ tuổi ≥ 21.
* **Quy mô**: 768 quan sát (dòng dữ liệu).
* **Biến số**: 9 biến (8 đặc trưng + 1 biến nhãn 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% |

Điều này cho thấy nhóm không mắc bệnh chiếm ưu thế, có thể ảnh hưởng đến kết quả huấn luyện mô hình nếu không xử lý thích hợp.

## 

## **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**.

Trong dữ liệu y tế, một số biến **không thể có giá trị bằng 0** về mặt sinh lý, nhưng trong dataset lại xuất hiện nhiều giá trị 0. Điều này phản ánh dữ liệu bị **thiếu thông tin (missing values được mã hóa bằng 0)** thay vì giá trị hợp lệ.

Các biến bị ảnh hưởng gồm:

* **Glucose** (nồng độ glucose huyết tương).
* **BloodPressure** (huyết áp tâm trương).
* **SkinThickness** (độ dày nếp gấp da).
* **Insulin** (nồng độ insulin huyết thanh).
* **BMI** (chỉ số khối cơ thể).

Nhận xét:

* **Insulin** và **SkinThickness** có tỷ lệ giá trị 0 rất cao, cho thấy dữ liệu bị thiếu đáng kể.
* **Glucose**, **BloodPressure**, và **BMI** cũng xuất hiện giá trị 0 nhưng ít hơn.
* Đây là các biến quan trọng về mặt y học → cần xử lý cẩn thận trước khi phân tích thống kê và mô hình hóa.

Hướng tiền xử lý sơ bộ:

* Các giá trị 0 bất hợp lý sẽ được xử lý bằng một trong các cách:
  + **Loại bỏ dòng dữ liệu** chứa giá trị thiếu (áp dụng nếu số lượng ít).
  + **Điền giá trị thay thế (imputation):** dùng **mô hình hồi quy tuyến tính**.

## 

## **III.3. Phương pháp phân tích dữ liệu (EDA)**

Phân tích dữ liệu khám phá (Exploratory Data Analysis – EDA) là bước quan trọng để hiểu rõ đặc trưng của dữ liệu, phát hiện bất thường và rút ra nhận định ban đầu. Trong nghiên cứu này, EDA được tiến hành theo các bước:

**1. Thống kê mô tả (Descriptive Statistics):**

* Sử dụng các chỉ số cơ bản như **số lượng (count), trung bình (mean), độ lệch chuẩn (std), giá trị nhỏ nhất, lớn nhất, các phân vị (25%, 50%, 75%)**.
* Mục tiêu: nắm bắt phân phối tổng quan của từng biến và phát hiện giá trị bất thường.

**2. Kiểm tra dữ liệu thiếu và bất hợp lý:**

* Xác định tỷ lệ giá trị **NaN** hoặc giá trị **0 không hợp lý** trong các biến y tế quan trọng (Glucose, BloodPressure, SkinThickness, Insulin, BMI).
* Đưa ra chiến lược xử lý: thay thế bằng median hoặc các phương pháp nội suy.

**3. Phân tích đơn biến (Univariate Analysis):**

* Sử dụng **histogram, boxplot** để phân tích phân phối từng biến.
* Mục tiêu: nhận diện độ lệch phân phối (skewness), outliers và tính chất dữ liệu.

**4. Phân tích hai biến (Bivariate Analysis):**

* So sánh từng đặc trưng với biến mục tiêu **Outcome (0/1)**.
* Sử dụng:
  + **Histogram so sánh phân phối giữa nhóm Outcome = 0 và Outcome = 1.**
  + **Bar chart** để thể hiện tần suất mắc bệnh theo từng nhóm giá trị (ví dụ: nhóm tuổi, nhóm BMI).
* Mục tiêu: nhận diện các biến có ảnh hưởng trực tiếp đến khả năng mắc tiểu đường.

**5. Phân tích đa biến (Multivariate Analysis):**

* Tính **ma trận tương quan (correlation matrix)** giữa các biến.
* Trực quan hóa bằng **heatmap** và **pairplot** để quan sát mối quan hệ giữa nhiều biến cùng lúc.
* Mục tiêu: xác định các đặc trưng quan trọng nhất trong dự đoán Outcome.

**6. Phân tích ngoại lệ (Outliers):**

* Sử dụng **boxplot và phương pháp IQR** để phát hiện giá trị ngoại lai.
* Đánh giá mức độ ảnh hưởng của outliers đến phân phối dữ liệu và khả năng phân tích.

**7. Tổng hợp kết quả từ EDA:**

* Đưa ra **các phát hiện chính** (insights) từ dữ liệu, bao gồm: đặc trưng quan trọng, đặc trưng ít giá trị, và vấn đề dữ liệu cần xử lý thêm.

Thông qua EDA, có thể xác định những biến quan trọng nhất cho bài toán dự đoán bệnh tiểu đường, đồng thời đưa ra các khuyến nghị cho bước tiền xử lý và xây dựng mô hình.

## 

## **III.4. Công cụ và môi trường phân tích**

Để thực hiện phân tích dữ liệu Pima Indians Diabetes, chúng tôi sử dụng môi trường và các công cụ sau:

**1. Môi trường phân tích**

* **Google Colab:**
  + Nền tảng điện toán đám mây miễn phí, cho phép chạy Python trực tiếp trên trình duyệt.
  + Ưu điểm: không cần cài đặt phức tạp, dễ chia sẻ và cộng tác nhóm.
  + Hỗ trợ GPU nếu cần thiết (trong trường hợp mở rộng sang mô hình học máy).

**2. Ngôn ngữ lập trình**

* **Python 3.x** – ngôn ngữ phổ biến trong phân tích dữ liệu và học máy nhờ cú pháp đơn giản, thư viện phong phú.

**3. Các thư viện chính sử dụng**

* **pandas:** xử lý và quản lý dữ liệu dạng bảng (DataFrame), tính toán thống kê cơ bản.
* **numpy:** hỗ trợ tính toán số học hiệu quả, xử lý ma trận và mảng.
* **matplotlib & seaborn:** trực quan hóa dữ liệu (biểu đồ histogram, violin plot, boxplot, heatmap, pairplot).
* **scikit-learn (sklearn):** tiền xử lý dữ liệu (chuẩn hóa, chia tập train/test), tính toán tương quan, cung cấp công cụ phân tích nâng cao.
* **joblib:** lưu trữ và tái sử dụng các đối tượng trong quá trình phân tích (nếu mở rộng sang mô hình).
* **IPython.display:** hiển thị bảng, kết quả và hình ảnh đẹp mắt trong notebook.

**4. Quản lý dữ liệu và cảnh báo**

* **warnings:** kiểm soát và ẩn các cảnh báo không cần thiết khi chạy notebook.
* **os, sys:** quản lý môi trường và đường dẫn dữ liệu (sử dụng khi cần thao tác file).

**5. Quy trình phân tích**

* Các công cụ trên được kết hợp theo quy trình:

1. **Đọc dữ liệu (CSV)** bằng pandas.
2. **Khám phá sơ bộ**: kiểm tra kích thước, kiểu dữ liệu, các giá trị bất thường.
3. **Thống kê mô tả** và **trực quan hóa** bằng matplotlib, seaborn.
4. **Tiền xử lý dữ liệu**: thay thế giá trị bất thường, chuẩn hóa dữ liệu.
5. **Phân tích mối quan hệ** giữa các biến qua heatmap, histogram, violin plot, pairplot.

# 

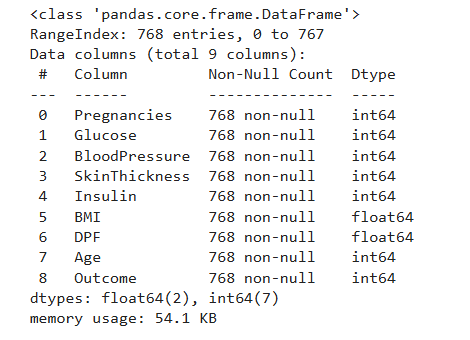
# 

# **IV. Phân tích dữ liệu (EDA)**

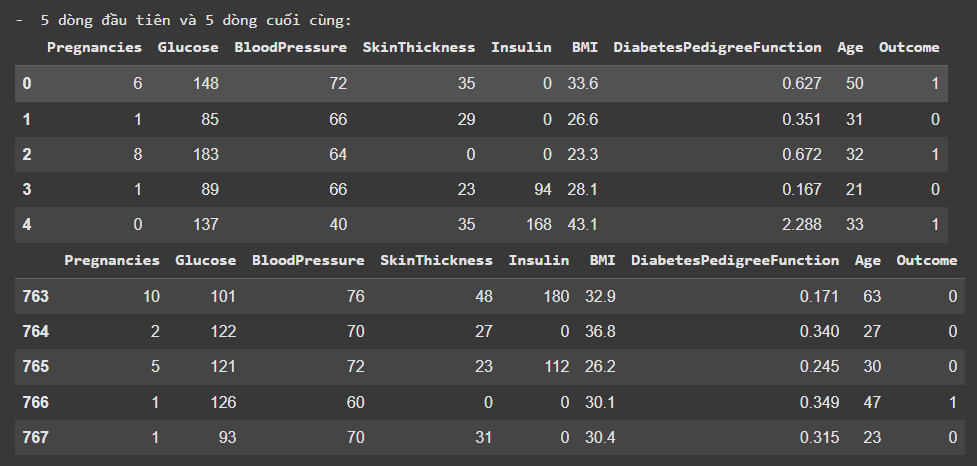
## **IV.1. Thống kê mô tả**

Trong bước đầu tiên của phân tích dữ liệu, chúng tôi tiến hành thống kê mô tả nhằm cung cấp cái nhìn tổng quan về phân bố các biến trong dataset.

### **IV.1.1. Kích thước và kiểu dữ liệu**

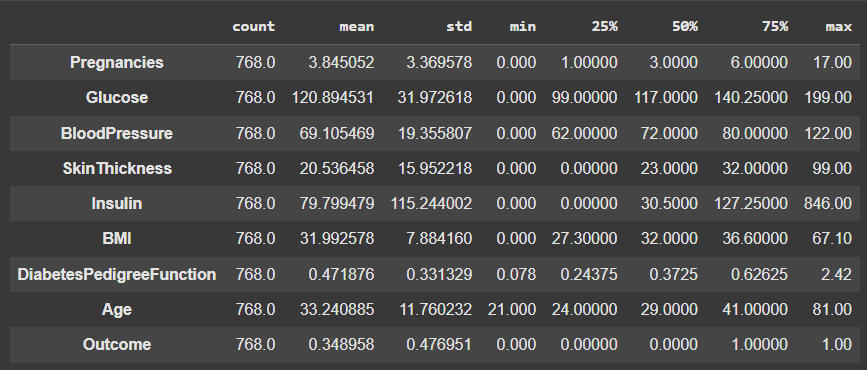
* Dataset gồm 768 dòng và 9 cột (8 biến đầu vào và 1 biến mục tiêu).
* Các biến đều là numeric (kiểu số nguyên hoặc số thực), thuận tiện cho việc phân tích thống kê và trực quan hóa. ****

### **IV.1.2. 5 dòng đầu và 5 dòng cuối**

* Giúp quan sát trực tiếp cấu trúc dữ liệu và giá trị điển hình trong dataset.****

### **IV.1.3. Thống kê cơ bản (Summary Statistics)**

Sử dụng df.describe() để tính toán:

* count, mean, std, min, max, 25%, 50%, 75%.
* Qua đó, phát hiện giá trị min bất hợp lý (nhiều giá trị bằng 0 trong các biến như Glucose, BloodPressure, SkinThickness, Insulin, BMI).  
  

**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) → nên 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**

Mặc dù bộ dữ liệu Pima Indians Diabetes không có giá trị **NaN** trực tiếp hay dòng dữ liệu trùng lặp, quá trình kiểm tra chi tiết cho thấy tồn tại nhiều giá trị **0 không hợp lý về mặt sinh lý**. Đây là một dạng **missing values được mã hóa bằng 0**, cần được nhận diện và xử lý trong các bước tiền xử lý.

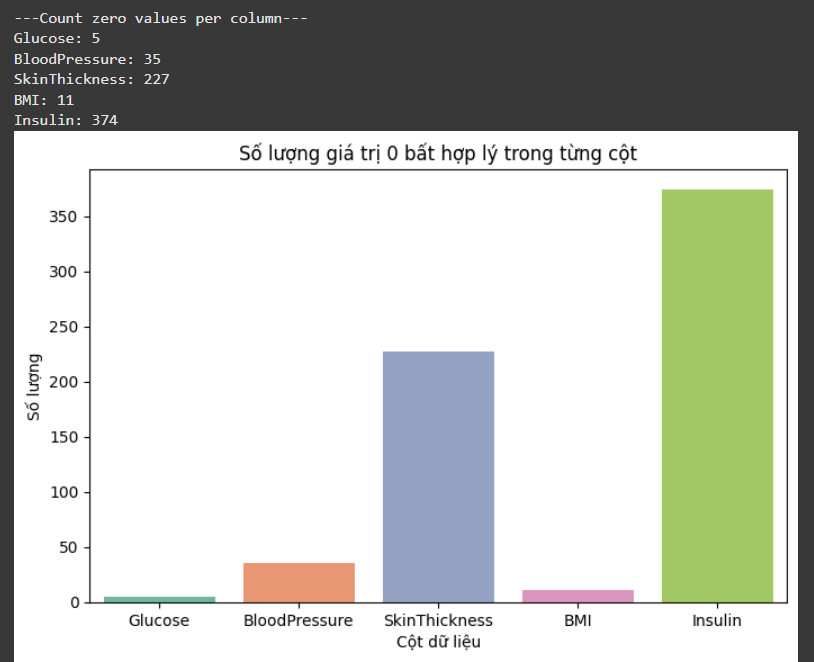
**1. Kiểm tra dữ liệu thiếu (Null/NaN) và trùng lặp**

* **Null/NaN:** không có.
* **Dòng trùng lặp:** không có.  
   → Dataset có chất lượng khá tốt về mặt kỹ thuật lưu trữ.

**2. Kiểm tra giá trị bất hợp lý (0 không thực tế)**

Các biến sinh lý sau đây xuất hiện giá trị **0**, vốn không hợp lý vì:

* **Glucose:** không thể bằng 0 ở người còn sống.
* **BloodPressure:** huyết áp tâm trương không thể bằng 0.
* **SkinThickness:** độ dày nếp gấp da không thể bằng 0 ở người bình thường.
* **Insulin:** giá trị 0 thường phản ánh dữ liệu bị thiếu, không phải sinh lý.
* **BMI:** chỉ số khối cơ thể không thể bằng 0.



### **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 |

**Nhận xét:**

Những giá trị này là bất hợp lý về mặt y học và cần được xem như dữ liệu thiếu. Việc xử lý là bước quan trọng trước khi xây dựng mô hình:

* Chúng ta có thể loại bỏ các hàng có giá trị bằng 0 ở các cột **Glucose, BloodPressure** hoặc **BMI** vì những hàng này chỉ chiếm khoảng **6%** dữ liệu.
* Trong khi đó, với cột **SkinThickness**, **Insulin** chúng ta nên ước lượng và điền giá trị (**impute**) thay vì loại bỏ, vì nếu bỏ đi thì sẽ mất khoảng **30%** dữ liệu, gây ảnh hưởng lớn đến phân tích.

## 

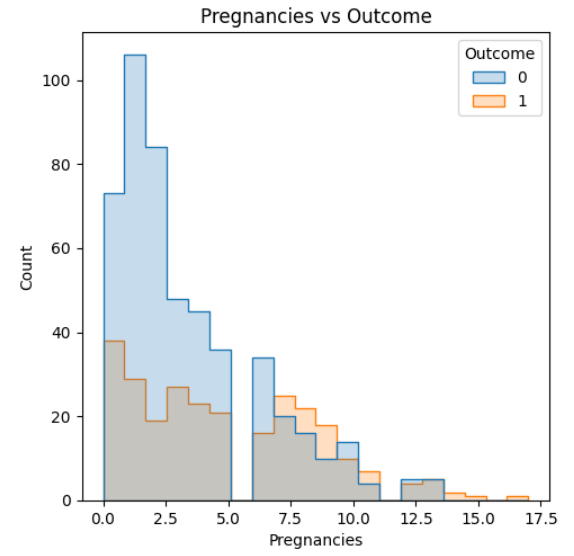
## **IV.3. Phân tích phân phối các biến (Univariate Analysis)**

Trong bước này, chúng tôi phân tích phân phối của từng biến độc lập nhằm:

* Hiểu rõ đặc điểm dữ liệu (trung tâm, phân tán, độ lệch).
* Phát hiện outliers.
* So sánh sơ bộ giữa nhóm mắc bệnh và không mắc bệnh.

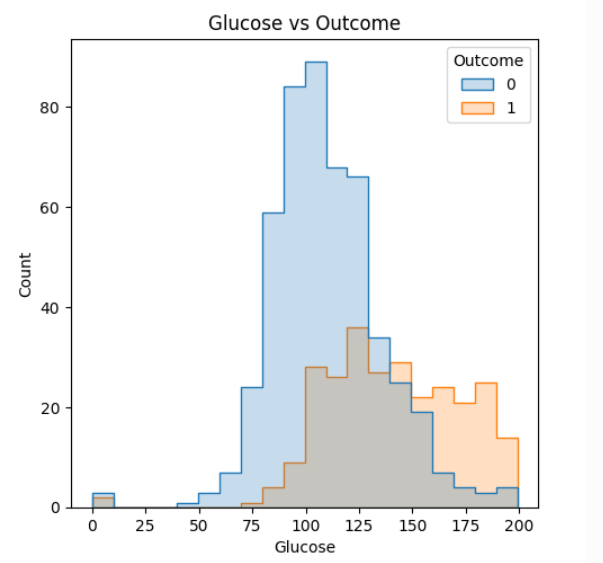
**Sử dụng Histogram để quan sát phân bố tần suất.**

### **IV.3.1. Pregnancies (Số lần mang thai)**

**Nhận xét:**

* Giá trị dao động từ **0 đến 17 lần**, trung bình khoảng **3.8 lần**.
* Phân bố lệch phải (nhiều phụ nữ có ≤ 3 lần mang thai, ít người có trên 10 lần).
* Nhóm mắc tiểu đường có xu hướng số lần mang thai cao hơn.

### **IV.3.2. Glucose (Nồng độ glucose huyết)**

**Nhận xét:**

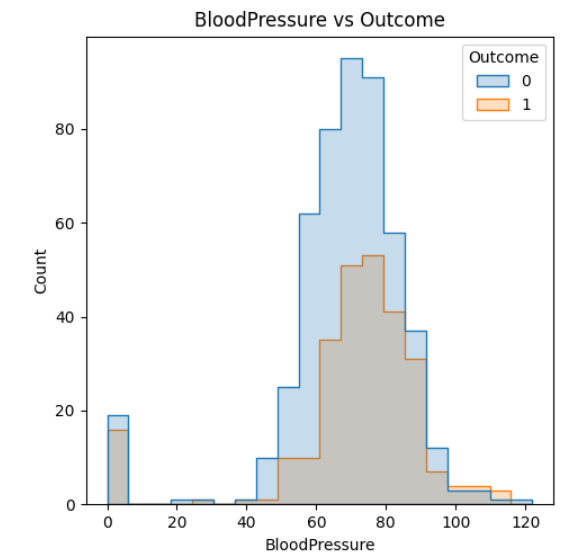
* Phân phối lệch phải, tập trung nhiều ở mức 100–140 mg/dL.
* Có nhiều giá trị trên 140 → vượt ngưỡng chẩn đoán tiền tiểu đường/tiểu đường.
* Một số giá trị bằng 0 (không hợp lý về mặt y học, coi là missing).
* Đây là biến có **sức phân biệt mạnh nhất** giữa hai nhóm Outcome.

### 

### 

### 

### **IV.3.3. BloodPressure (Huyết áp tâm trương)**

**Nhận xét:**

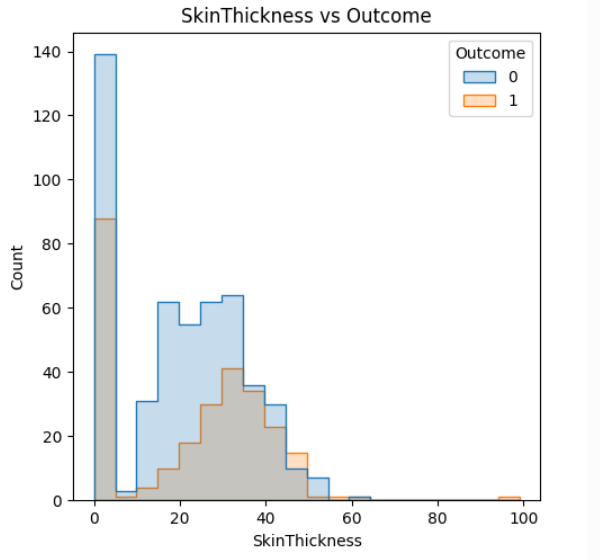
* Trung bình khoảng 70 mmHg, phù hợp với giá trị sinh lý.
* Một số giá trị bằng 0 → không hợp lý.
* Phân phối khá gần chuẩn, nhưng có đuôi dài về phía giá trị thấp.
* Sự khác biệt giữa nhóm mắc và không mắc không rõ rệt.

### 

### 

### 

### **IV.3.4. SkinThickness (Độ dày nếp gấp da)**

**Nhận xét:**

* Rất nhiều giá trị bằng 0 → dữ liệu thiếu (30%).
* Với các giá trị hợp lý, phân phối tập trung quanh 20–35 mm.
* Không có nhiều ngoại lệ, nhưng dữ liệu thiếu ảnh hưởng mạnh đến phân tích.

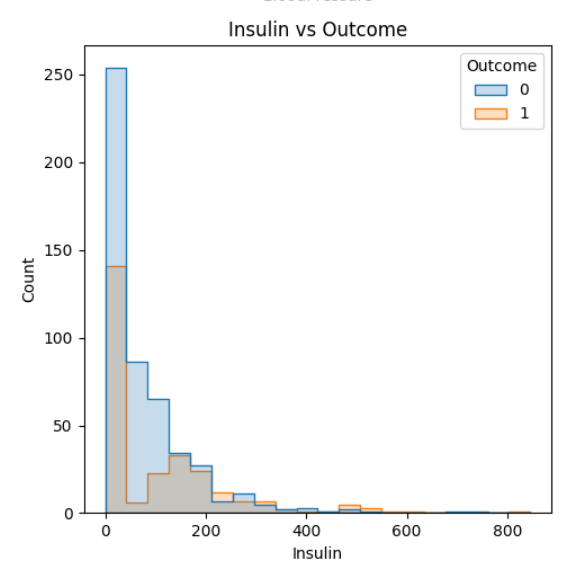
### 

### 

### 

### 

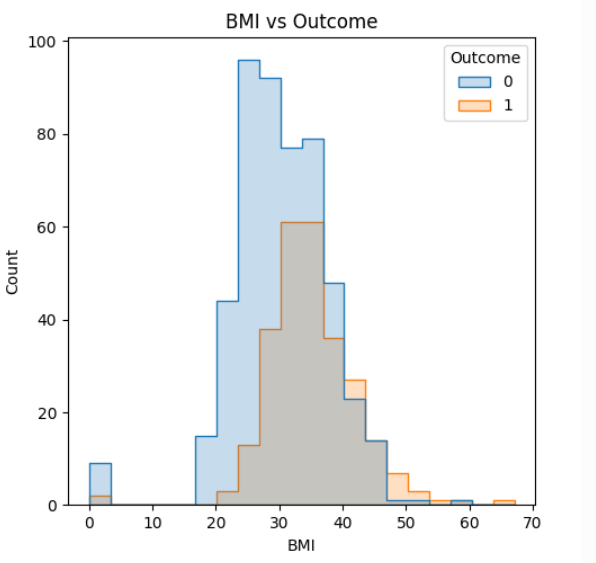
### **IV.3.5. Insulin (Nồng độ insulin sau 2 giờ)**

**Nhận xét:**

* Gần một nửa dữ liệu = 0 → thiếu nghiêm trọng.
* Phân phối rất lệch phải, có nhiều giá trị cực trị > 500 µU/mL.
* Do dữ liệu không sạch, biến này có độ tin cậy thấp nếu chưa xử lý.

### 

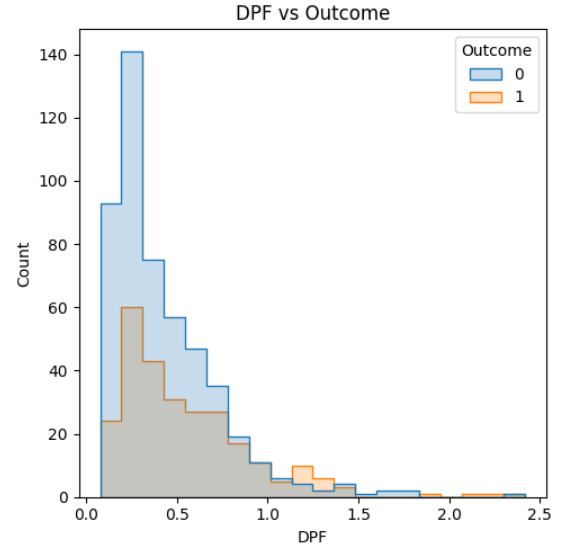
### **IV.3.6. BMI (Body Mass Index)**

**Nhận xét:**

* Trung bình khoảng 32, phần lớn dữ liệu nằm trong khoảng 25–40 (thừa cân – béo phì).
* Một số giá trị bằng 0 → bất hợp lý.
* Phân phối gần chuẩn, hơi lệch phải nhẹ.
* Đây là biến quan trọng vì béo phì liên quan trực tiếp đến tiểu đường.

### 

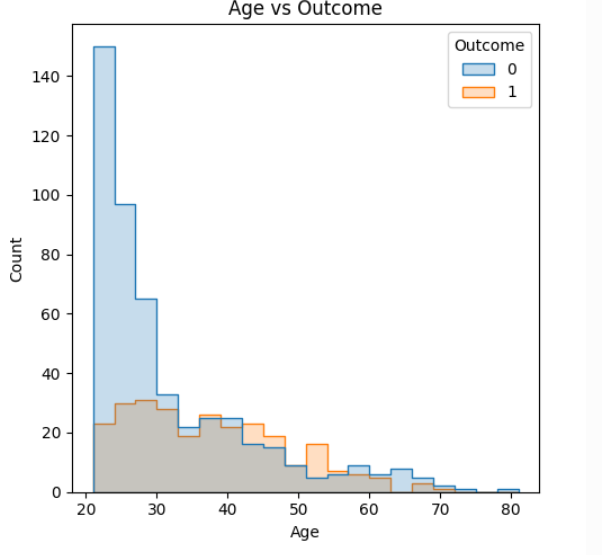
### **IV.3.7. DiabetesPedigreeFunction (Chỉ số nguy cơ di truyền)**

**Nhận xét:**

* Phân phối lệch phải mạnh.
* Phần lớn giá trị nằm < 0.5, nhưng có một số bệnh nhân > 2.0.
* Các giá trị cao cho thấy nguy cơ di truyền rất lớn.
* Đây là biến đặc thù của dataset, ít gặp trong dữ liệu lâm sàng thông thường.

### 

### **IV.3.8. Age (Tuổi bệnh nhân)**

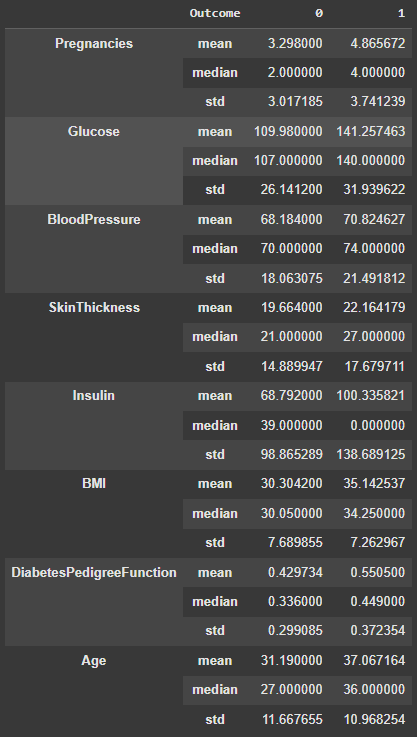


**Nhận xét:**

* Đa số bệnh nhân trong độ tuổi 20–40.
* Một số trường hợp 60–80 tuổi, số lượng ít.
* Phân phối lệch phải, phản ánh cấu trúc dân số trong dữ liệu.
* Tuổi càng cao thì nguy cơ mắc bệnh càng lớn.

## **IIV.4. So sánh nhóm mắc và không mắc (Bivariate Analysis)**

### **IV.4.1 Thống kê mô tả phân tách theo Outcome**



Bảng này cho thấy sự khác biệt trung bình (mean), trung vị (median), và độ lệch chuẩn (std) giữa nhóm tiểu đường và không tiểu đường.

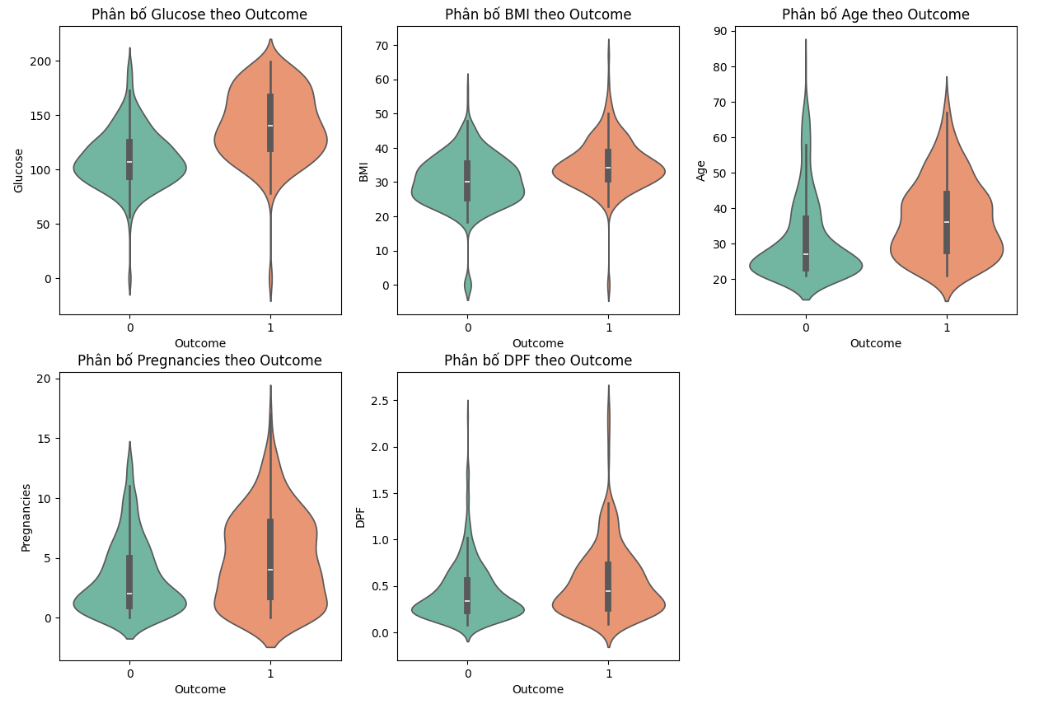
**Nhận xét:**

* 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.

### 

### **IV.4.2 Violin plot 5 thuộc tính**

So sánh phân bố của 5 thuộc tính có sự khác biệt rõ rệt hơn giữa nhóm tiểu đường và không tiểu đường.



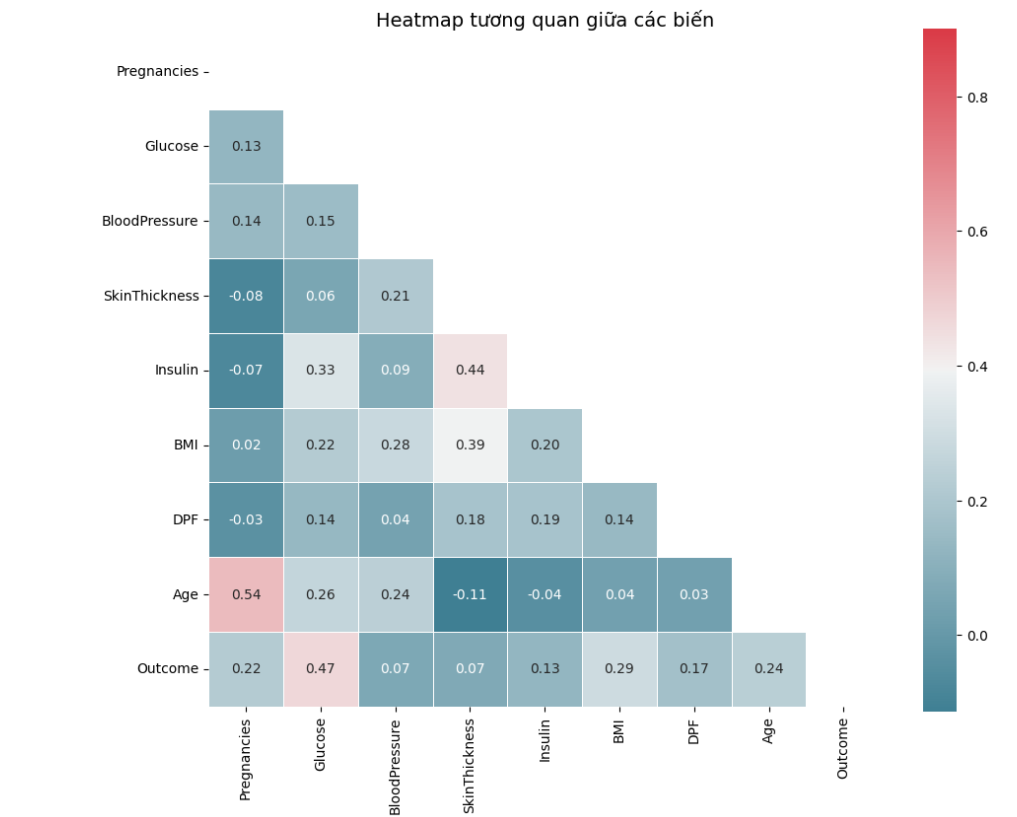
**Nhận xét:**

* 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.
* Kết quả này phù hợp với hiểu biết lâm sàng từ các paper (WHO 1999, Pima dataset).

## **IV.5. Phân tích đa biến và tương quan**

Để đánh giá mối quan hệ giữa các biến, tính hệ số tương quan Pearson và trực quan hóa bằng **heatmap**.

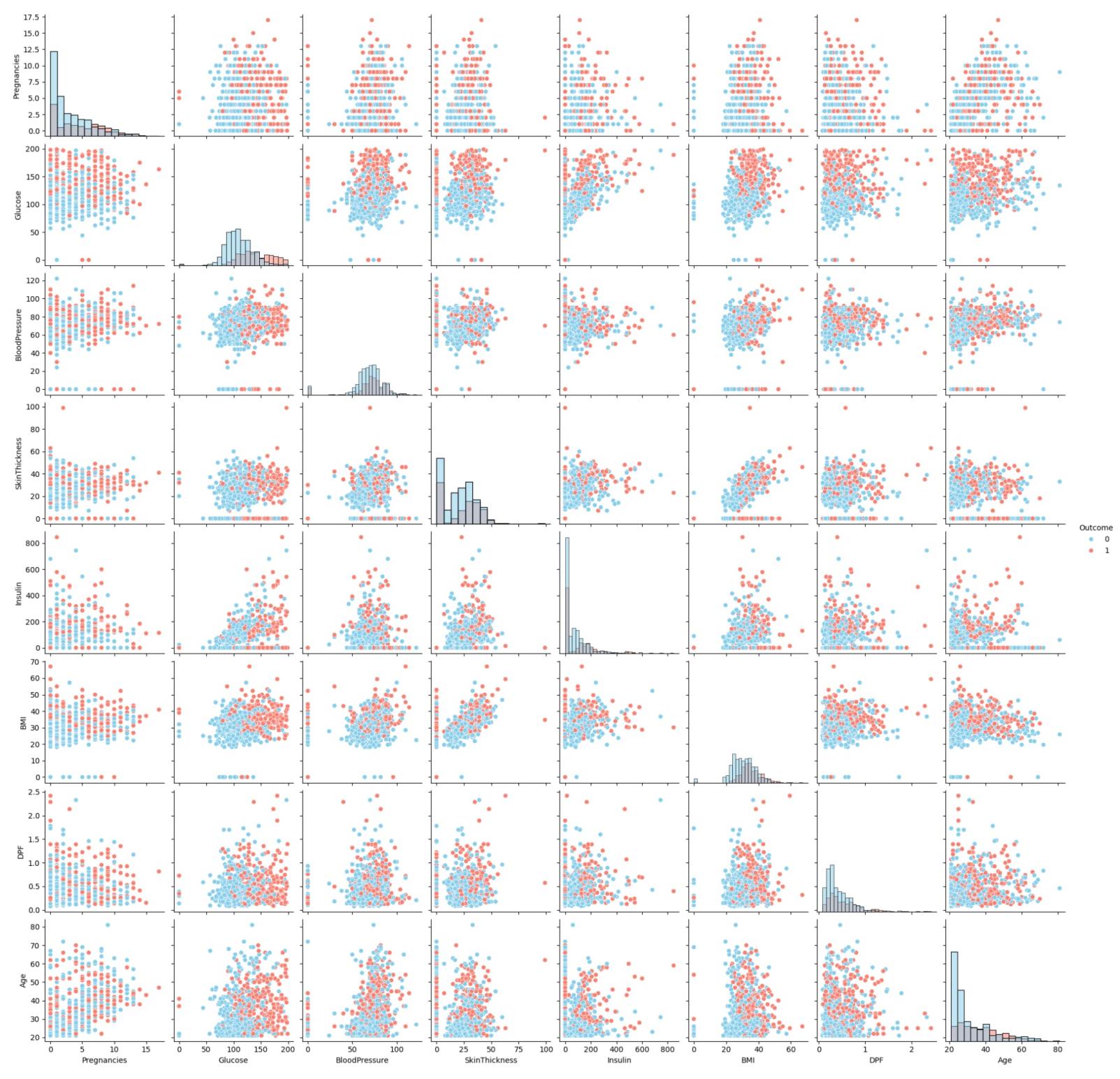
### **IV.5.1 Heatmap tương quan giữa các biến**



### 

### **IV.5.2 Pairplot (Scatter matrix)**

* Dùng pairplot để quan sát mối quan hệ 2 biến một cách trực quan, với màu sắc phân biệt Outcome.

****

### **IV.5.3 Nhận xét tổng quan:**

- Glucose và BMI:

+ Hai biến này cho thấy sự phân tách khá rõ giữa hai nhóm.

+ 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 và Pregnancies:

+ 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:

+ Phân bố giữa hai nhóm gần như trùng nhau, hỗn loạn và khó tách biệt.

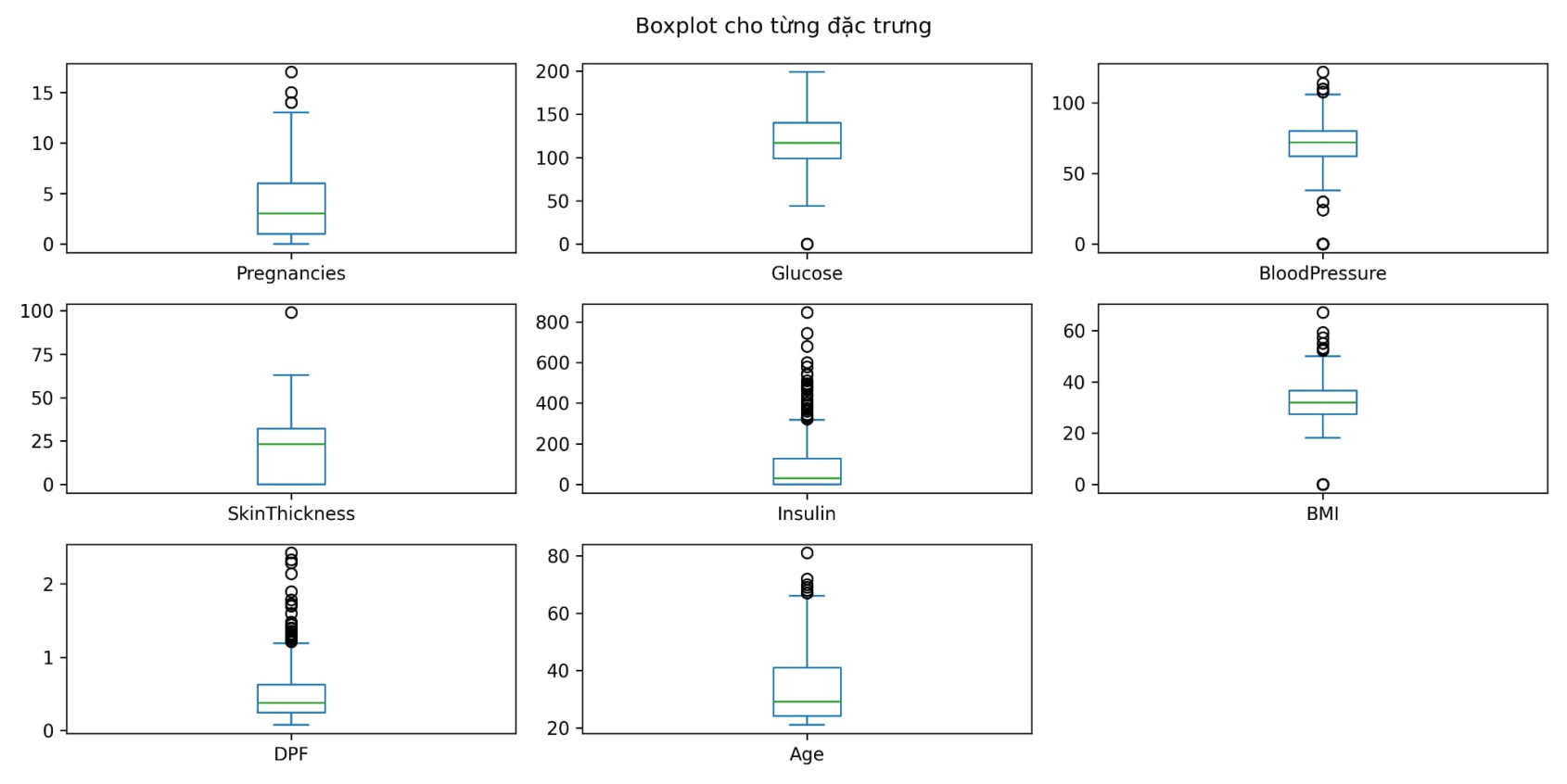
* **Outcome có tương quan cao nhất với Glucose (0.47)** → xác nhận tầm quan trọng.
* BMI (0.29), Age (0.24), Pregnancies (0.22) cũng có liên hệ với Outcome.
* Các biến khác (BloodPressure, SkinThickness, Insulin, DiabetesPedigreeFunction) có tương quan thấp (<0.15).
* Pregnancies và Age có tương quan dương với nhau, phù hợp logic (người lớn tuổi thường có nhiều lần mang thai hơn).

**Kết luận :** Glucose là đặc trưng quan trọng nhất, tiếp theo là BMI, Age và Pregnancies.

## 

## **IV.6. Outliers (giá trị ngoại lai)**

### **IV.6.1 Boxplot được sử dụng để phát hiện giá trị ngoại lai trong các biến**



**Nhận xét theo biến:**

* **Pregnancies**: một số trường hợp ≥15 lần mang thai → hiếm gặp, có thể coi là ngoại lệ.
* **Insulin**: nhiều giá trị cực trị > 500 µU/mL → 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.

**Kết luận:** Dataset chứa nhiều ngoại lệ, đặc biệt ở Insulin và Pregnancies. Cần xử lý trước khi xây dựng mô hình.

## **IV.7. Kết quả tổng hợp từ EDA**

Sau khi thực hiện các bước **phân tích khám phá dữ liệu (EDA)**, chúng tôi rút ra các kết luận chính như sau:

### **IV.7.1. Chất lượng dữ liệu**

* Bộ dữ liệu Pima Indians Diabetes gồm **768 dòng và 9 cột** (8 biến đầu vào và 1 biến mục tiêu – Outcome).
* Không có giá trị **NaN** hoặc dòng dữ liệu trùng lặp.
* Tuy nhiên, tồn tại nhiều giá trị **0 bất hợp lý** ở các biến sinh lý (Glucose, BloodPressure, SkinThickness, Insulin, BMI):
  + Ví dụ: **Insulin** có 48.7% giá trị bằng 0, **SkinThickness** có gần 30%.
  + Đây được xem là **missing values** được mã hóa bằng 0 → cần xử lý trước khi huấn luyện mô hình.

### **IV.7.2. Phân bố dữ liệu (Univariate Analysis)**

* **Glucose**: phân phối lệch phải, giá trị trung bình cao hơn ở nhóm mắc bệnh; biến quan trọng nhất.
* **BMI**: đa số bệnh nhân thừa cân/béo phì (BMI ≥ 30), đặc biệt phổ biến ở nhóm mắc bệnh.
* **Age**: bệnh nhân mắc bệnh tập trung nhiều ở nhóm **> 40 tuổi**.
* **Pregnancies**: số lần mang thai cao có liên quan đến khả năng mắc tiểu đường.
* **Diabetes Pedigree Function (DPF)**: phản ánh yếu tố di truyền, giá trị cao hơn ở nhóm mắc bệnh nhưng phân tách chưa rõ rệt.
* **BloodPressure, SkinThickness, Insulin**: phân phối không rõ ràng, có nhiều dữ liệu bất thường (0).

### **IV.7.3. So sánh theo Outcome (Bivariate Analysis)**

* **Violin plot** cho thấy sự khác biệt rõ rệt ở các biến: **Glucose, BMI, Age, Pregnancies, DPF** giữa hai nhóm Outcome.
* Các biến khác (BloodPressure, SkinThickness, Insulin) ít giá trị phân biệt khi so sánh độc lập.

### **IV.7.4. Phân tích đa biến (Multivariate Analysis)**

* **Heatmap (Correlation matrix)** chỉ ra rằng:
  + **Glucose (r ≈ 0.47)** có tương quan mạnh nhất với Outcome.
  + **BMI, Age, Pregnancies** cũng có tương quan dương, nhưng yếu hơn.
  + **BloodPressure, SkinThickness, Insulin** hầu như không tương quan với Outcome.
* **Pairplot** cho thấy sự phân tách rõ nhất ở cặp **Glucose – BMI** và **Age – Pregnancies**. Các cặp biến khác phân bố hỗn loạn, khó phân tách hai nhóm Outcome.

### **IV.7.5. Tổng hợp**

* **Biến quan trọng nhất:** Glucose, BMI, Age, Pregnancies.
* **Biến bổ trợ:** DPF.
* **Biến ít giá trị phân loại (và nhiều missing values):** BloodPressure, SkinThickness, Insulin.
* Dataset **không cân bằng tuyệt đối nhưng ở mức chấp nhận được** (65% không bệnh, 35% có bệnh). Tuy nhiên, cần cân nhắc kỹ trong huấn luyện mô hình để tránh bias.

**IV.7.6 Kết luận chung từ EDA:**

* Dữ liệu Pima Indians Diabetes cung cấp nhiều thông tin hữu ích để phân tích nguy cơ mắc bệnh tiểu đường.
* **Glucose và BMI** nổi bật là hai yếu tố then chốt trong phân loại bệnh nhân.
* Các vấn đề dữ liệu (giá trị 0 bất hợp lý) cần được xử lý cẩn thận trong bước **Chuẩn bị dữ liệu (Data Preparation)** để đảm bảo độ tin cậy cho các mô hình dự đoán.

# 

# **IV.8. Xử lý dữ liệu thiếu và bất hợp lý (Data Cleaning)**

Trong quá trình phân tích dữ liệu (EDA), chúng tôi đã phát hiện ra một số biến sinh lý có chứa nhiều giá trị 0 không hợp lý về mặt y học. Đây thực chất là các missing values được mã hóa bằng 0, cần xử lý trước khi tiếp tục phân tích và xây dựng mô hình.

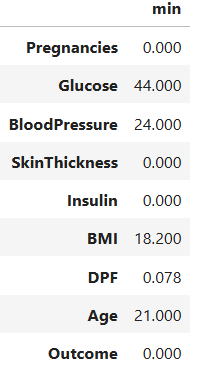
### **IV.8.1. Nguyên tắc xử lý**

* Các biến quan trọng trong chẩn đoán như Glucose, BloodPressure, BMI: do số lượng giá trị 0 ít và các giá trị này là hoàn toàn phi thực tế, chúng tôi chọn cách loại bỏ trực tiếp các dòng dữ liệu chứa 0 để tránh sai lệch.
* Các biến có tỷ lệ giá trị 0 cao như SkinThickness và Insulin: thay thế 0 bằng giá trị ước lượng thông qua mô hình hồi quy tuyến tính (Linear Regression Imputation). Phương pháp này tận dụng mối quan hệ với các biến khác (ví dụ Glucose, BMI, Age) để dự đoán giá trị thiếu.

### **IV.8.2. Các bước xử lý chi tiết**

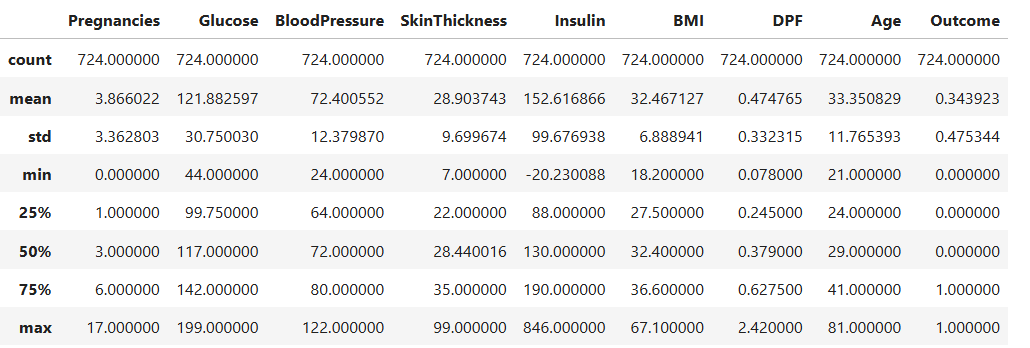
* Bước 1: Loại bỏ các dòng có giá trị 0 ở các biến:
  + Glucose
  + BloodPressure
  + BMI

Size of dataframe: (724, 9)



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 InsuliInsulin 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 phối rõ rệt hơn.

### **IV.8.3. 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.1.1. Ý nghĩa y học**

* **Phát hiện sớm nguy cơ tiểu đường:** Các đặc trưng như **Glucose, BMI, Age, Pregnancies** có mối quan hệ chặt chẽ với khả năng mắc bệnh tiểu đường type 2. Phân tích này giúp các bác sĩ nhận diện sớm nhóm bệnh nhân nguy cơ cao, ngay cả trước khi có triệu chứng rõ rệt.
* **Hỗ trợ chẩn đoán và điều trị:**
  + **Glucose** là chỉ số lâm sàng quan trọng nhất trong chẩn đoán tiểu đường theo tiêu chuẩn WHO và ADA.
  + **BMI** cho thấy mối liên hệ trực tiếp giữa béo phì và tiểu đường type 2, khẳng định vai trò của thay đổi lối sống (ăn uống, vận động) trong dự phòng.
  + **Pregnancies** nhấn mạnh nguy cơ tiểu đường thai kỳ (Gestational Diabetes) và mối liên hệ với bệnh tiểu đường type 2 sau này.
  + **Age** phản ánh rằng người trung niên và cao tuổi cần được sàng lọc thường xuyên hơn.
* **Ý nghĩa dịch tễ học:** Nghiên cứu trên cộng đồng **Pima Indians** cho thấy tiểu đường type 2 có tính **di truyền và môi trường** mạnh, giúp so sánh với các cộng đồng khác (ví dụ dân châu Á có ngưỡng glucose thấp hơn nhưng vẫn tăng nguy cơ).

### **V.1.2. Ý nghĩa thực tiễn**

* **Ứng dụng trong y tế dự phòng:** Phân tích dữ liệu cho phép xây dựng các công cụ sàng lọc (screening tools) để dự đoán nguy cơ tiểu đường dựa trên các đặc trưng đơn giản, chi phí thấp (BMI, tuổi, số lần mang thai, tiền sử gia đình).
* **Hỗ trợ ra quyết định lâm sàng:** Kết quả từ EDA có thể tích hợp vào **Clinical Decision Support Systems (CDSS)**, giúp bác sĩ đưa ra khuyến nghị chính xác hơn cho từng bệnh nhân.
* **Định hướng phát triển mô hình học máy:** Việc xác định các biến có giá trị phân loại cao (Glucose, BMI, Age, Pregnancies) tạo tiền đề cho việc xây dựng các mô hình dự đoán. Những mô hình này có thể được ứng dụng trong khám sàng lọc cộng đồng hoặc trong hệ thống y tế điện tử.
* **Tác động xã hội:** Hiểu rõ yếu tố nguy cơ giúp thiết kế **chính sách y tế cộng đồng**, ví dụ: chương trình giảm béo phì, tư vấn dinh dưỡng cho phụ nữ mang thai, khuyến nghị xét nghiệm glucose định kỳ ở người trung niên.

## 

## **V.2. Hạn chế của dữ liệu**

Mặc dù bộ dữ liệu Pima Indians Diabetes là một trong những dataset y tế được sử dụng rộng rãi nhất trong học máy, nó vẫn tồn tại nhiều hạn chế, cần được lưu ý khi diễn giải kết quả:

### **V.2.1. Giới hạn về đối tượng nghiên cứu**

* Tất cả bệnh nhân trong dataset đều là nữ giới ≥ 21 tuổi, thuộc cộng đồng Pima Indian tại Arizona, Mỹ.
* Điều này dẫn đến:
  + Không thể khái quát hóa kết quả cho nam giới hoặc các dân tộc khác.
  + Các yếu tố nhân khẩu học và môi trường sống của Pima Indians có thể rất khác với các cộng đồng khác (ví dụ: chế độ ăn, tỉ lệ béo phì cao bất thường).

### **V.2.2. Vấn đề về dữ liệu thiếu và bất hợp lý**

* Nhiều biến sinh lý có giá trị 0 không hợp lý, phản ánh dữ liệu bị thiếu nhưng được mã hóa sai:
  + Insulin: gần 50% giá trị = 0.
  + SkinThickness: gần 30% giá trị = 0.
* Điều này làm giảm chất lượng và độ tin cậy khi phân tích thống kê và huấn luyện mô hình.
* Cách xử lý missing values khác nhau có thể ảnh hưởng mạnh đến kết quả phân tích.

### **V.2.3. Hạn chế về số lượng biến**

* Dataset chỉ bao gồm 8 đặc trưng lâm sàng cơ bản, thiếu nhiều biến quan trọng trong chẩn đoán hiện đại, như:
  + HbA1c (đường huyết trung bình 3 tháng).
  + Cholesterol, lipid máu (liên quan hội chứng chuyển hóa).
  + Lối sống: chế độ ăn, vận động, hút thuốc.
  + Yếu tố di truyền nâng cao: marker sinh học, dữ liệu gene.
* Vì vậy, dataset mang tính giản lược, khó mô phỏng hoàn toàn thực tế y học.

### **V.2.4. Hạn chế về tính thời gian (temporal aspect)**

* Dữ liệu được thu thập tại một thời điểm nhất định (cross-sectional).
* Không có thông tin theo dõi dọc (longitudinal data) → khó phân tích tiến triển bệnh theo thời gian.
* Trong khi đó, tiểu đường type 2 là bệnh mạn tính phát triển dần, nên dữ liệu theo dõi lâu dài sẽ hữu ích hơn nhiều.

### **V.2.5. Mất cân bằng lớp (class imbalance)**

* Outcome có phân bố 500 (không bệnh) vs 268 (có bệnh) → chưa quá mất cân bằng, nhưng vẫn đủ để gây bias khi huấn luyện mô hình.
* Nếu không xử lý, mô hình có thể thiên về dự đoán "không bệnh".

## 

## 

## **V.3. Định hướng ứng dụng trong học máy**

Từ những kết quả phân tích và đặc điểm của bộ dữ liệu **Pima Indians Diabetes**, có thể xác định một số hướng ứng dụng trong lĩnh vực **học máy và trí tuệ nhân tạo (AI)**:

### **V.3.1. Bài toán phân loại nhị phân (Binary Classification)**

* **Mục tiêu:** dự đoán khả năng mắc tiểu đường (Outcome = 0 hoặc 1) từ 8 đặc trưng đầu vào.
* **Thuật toán áp dụng:**
  + Các mô hình truyền thống: Logistic Regression, KNN, Decision Tree, Random Forest.
  + Các mô hình mạnh hơn: Gradient Boosting (XGBoost, LightGBM, CatBoost).
  + Mạng nơ-ron (ANN) để mô phỏng phi tuyến tính giữa các đặc trưng.
* **Thước đo đánh giá:** Accuracy, Precision, Recall, F1-score, ROC-AUC.

### **V.3.2. Phát hiện đặc trưng quan trọng (Feature Importance)**

* Sử dụng các mô hình cây (Random Forest, XGBoost) hoặc phương pháp thống kê (ANOVA, Mutual Information) để xác định:
  + Đặc trưng nào quan trọng nhất trong dự đoán (thường là Glucose, BMI, Age, Pregnancies).
  + Đặc trưng nào ít ảnh hưởng → có thể giảm chiều dữ liệu (feature selection).

### **V.3.3. Xử lý dữ liệu mất cân bằng (Imbalanced Learning)**

* Vì dữ liệu có tỷ lệ 65% (không bệnh) vs 35% (bệnh), cần áp dụng:
  + **Oversampling** (SMOTE, ADASYN).
  + **Undersampling** (Random Undersampling).
  + **Class weights** trong các mô hình học máy.
* Điều này giúp mô hình cân bằng giữa việc phát hiện đúng bệnh nhân mắc bệnh và giảm false negative.

### **V.3.4. Dự báo nguy cơ và hệ thống hỗ trợ quyết định (Decision Support Systems)**

* Mô hình có thể tích hợp vào hệ thống y tế điện tử (EHR – Electronic Health Records) để **tự động cảnh báo** bệnh nhân nguy cơ cao.
* Ứng dụng trong **khám sàng lọc cộng đồng**: bác sĩ hoặc y tá có thể nhập thông tin cơ bản (tuổi, BMI, Glucose, số lần mang thai, tiền sử gia đình) để ước lượng nguy cơ mắc bệnh.

### **V.3.5. Định hướng mở rộng**

* **Kết hợp với dữ liệu khác:**
  + Thêm HbA1c, cholesterol, lipid máu, thói quen sinh hoạt để cải thiện mô hình.
* **Phân tích tiến triển bệnh (longitudinal analysis):**
  + Nếu có dữ liệu theo dõi dài hạn, có thể áp dụng các mô hình dự báo theo thời gian (time-series, survival analysis).
* **Ứng dụng explainable AI (XAI):**
  + Sử dụng SHAP, LIME để giải thích quyết định mô hình, giúp tăng độ tin cậy trong lâm sàng.

# 

# **VI. Kết luận**

Trong nghiên cứu này, chúng tôi đã tiến hành **phân tích khám phá dữ liệu (Exploratory Data Analysis – EDA)** trên bộ dữ liệu **Pima Indians Diabetes** nhằm tìm hiểu mối liên hệ giữa các yếu tố lâm sàng và khả năng mắc bệnh tiểu đường type 2.

### **VI.1. Kết quả chính**

* **Chất lượng dữ liệu:** Bộ dữ liệu không có giá trị NaN hoặc dòng trùng lặp, nhưng chứa nhiều giá trị 0 bất hợp lý ở các biến sinh lý quan trọng (Glucose, BloodPressure, BMI, SkinThickness, Insulin). Đây thực chất là missing values cần xử lý.
* **Phân tích đơn biến (Univariate Analysis):**
  + **Glucose** và **BMI** nổi bật với sự phân biệt rõ rệt giữa nhóm mắc bệnh và không mắc bệnh.
  + **Age** và **Pregnancies** cũng cho thấy mối liên hệ với Outcome.
  + **DPF** phản ánh yếu tố di truyền, có ảnh hưởng nhưng không mạnh.
  + Các biến **BloodPressure, SkinThickness, Insulin** ít có giá trị phân loại độc lập.
* **Phân tích hai biến (Bivariate Analysis):**
  + Violin plot cho thấy Glucose và BMI có sự khác biệt rõ ràng nhất giữa Outcome = 0 và Outcome = 1.
  + Age và Pregnancies cũng có sự khác biệt, nhưng mức độ tách biệt không mạnh bằng.
* **Phân tích đa biến (Multivariate Analysis):**
  + Heatmap chỉ ra Glucose có tương quan mạnh nhất với Outcome (r ≈ 0.47).
  + Pairplot cho thấy Glucose – BMI và Age – Pregnancies là những cặp đặc trưng có khả năng phân biệt tốt nhất.
* **Xử lý dữ liệu bất hợp lý:**
  + Loại bỏ dòng có giá trị 0 ở Glucose, BMI, BloodPressure.
  + Điền giá trị thiếu ở SkinThickness và Insulin bằng hồi quy tuyến tính.

### **VI.2. Ý nghĩa y học và thực tiễn**

* Xác nhận lại các yếu tố nguy cơ kinh điển của tiểu đường type 2: **Glucose cao, mức độ béo phì (BMI cao), tuổi cao, số lần mang thai nhiều**.
* Gợi ý áp dụng vào **khám sàng lọc cộng đồng**, giúp phát hiện sớm bệnh nhân nguy cơ cao.
* Tạo cơ sở dữ liệu để phát triển các hệ thống **hỗ trợ ra quyết định lâm sàng (Clinical Decision Support Systems)**.

### **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**

* **Nghiên cứu học máy:** so sánh nhiều thuật toán phân loại (Logistic Regression, Random Forest, XGBoost, Neural Networks).
* **Ứng dụng y tế:** xây dựng mô hình dự đoán nguy cơ tiểu đường cho khám sàng lọc và quản lý hồ sơ bệnh án điện tử.
* **Phát triển dữ liệu:** bổ sung đặc trưng mới (HbA1c, lipid máu, lối sống), thu thập dữ liệu đa dạng hơn về giới tính và dân tộc.
* **Explainable AI:** ứng dụng SHAP, LIME để giải thích mô hình, tăng tính minh bạch và tin cậy khi sử dụng trong lâm sàng.

### **VI.5 Kết luận cuối cùng**

Phân tích dữ liệu Pima Indians Diabetes cho thấy **Glucose, BMI, Age, Pregnancies và DPF** là những yếu tố then chốt trong việc dự đoán tiểu đường type 2. Kết quả nghiên cứu vừa củng cố bằng chứng y học, vừa mở ra hướng ứng dụng học máy trong **dự báo nguy cơ, y tế dự phòng và hỗ trợ quyết định lâm sàng**. Tuy còn nhiều hạn chế, nhưng bộ dữ liệu này vẫn là một nền tảng quan trọng trong giảng dạy, nghiên cứu và phát triển các ứng dụng AI trong y học.

# **TÀI LIỆU THAM KHẢO**

1. World Health Organization. (1999). *Definition, diagnosis and classification of diabetes mellitus and its complications: Report of a WHO consultation. Part 1: Diagnosis and classification of diabetes mellitus*. Geneva: World Health Organization.
2. Smith, J. W., Everhart, J. E., Dickson, W. C., Knowler, W. C., & Johannes, R. S. (1988). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In *Proceedings of the Annual Symposium on Computer Applications in Medical Care* (pp. 261–265). IEEE Computer Society Press.
3. Alberti, K. G. M. M., & Zimmet, P. Z. (1998). Definition, diagnosis and classification of diabetes mellitus and its complications. *Diabetic Medicine, 15*(7), 539–553.
4. Kaggle. (2017). *Pima Indians Diabetes Database*. Retrieved from <https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database>

# 