TRƯỜNG ĐẠI HỌC SÀI GÒN

KHOA CÔNG NGHỆ THÔNG TIN

----------ஜ📖ரு----------

BÁO CÁO

Môn học: Nhập môn máy học

Tên đề tài: Phân tích khám phá dữ liệu

Giảng viên hướng dẫn: Đỗ Như Tài

Thành viên:

3123410387 – Nguyễn Hữu Tri

3123410274 – Lư Hồng Phúc

Thành phố Hồ Chí Minh, 29 Tháng 09 Năm 2025

TRƯỜNG ĐẠI HỌC SÀI GÒN

KHOA CÔNG NGHỆ THÔNG TIN

----------ꕥ📖ꕥ----------

BÁO CÁO

Môn học: Nhập môn máy học

Tên đề tài: Phân tích khám phá dữ liệu

Giảng viên hướng dẫn: Đỗ Như Tài

Thành viên:

3123410387 – Nguyễn Hữu Tri

3123410274 – Lư Hồng Phúc

Thành phố Hồ Chí Minh, 29 Tháng 09 Năm 202

LỜI MỞ ĐẦU

Bệnh tiểu đường, đặc biệt là type 2, đã và đang trở thành một trong những thách thức sức khỏe cộng đồng lớn nhất của thế kỷ 21. Được mệnh danh là "kẻ giết người thầm lặng", căn bệnh này tiến triển âm thầm nhưng để lại những hậu quả nặng nề, từ các biến chứng tim mạch, suy thận, mù lòa đến tổn thương thần kinh không thể phục hồi. Gánh nặng mà bệnh tiểu đường đặt lên vai người bệnh, gia đình và toàn xã hội là vô cùng lớn, không chỉ về mặt chi phí y tế mà còn về chất lượng cuộc sống bị suy giảm nghiêm trọng. Trong bối cảnh đó, việc phát hiện sớm và dự báo chính xác nguy cơ mắc bệnh không còn là một lựa chọn, mà là một yêu cầu cấp thiết để có thể can thiệp kịp thời, làm chậm tiến trình của bệnh và giảm thiểu các biến chứng nguy hiểm.

Trước thách thức này, sự giao thoa giữa y học và khoa học dữ liệu mở ra một hướng đi đầy hứa hẹn. Các kỹ thuật học máy và trí tuệ nhân tạo đang dần chứng tỏ vai trò không thể thiếu trong việc xây dựng các hệ thống sàng lọc thông minh, có khả năng nhận diện các mẫu tiềm ẩn trong dữ liệu y tế để đưa ra những dự báo sớm với độ chính xác cao. Để minh họa cho tiềm năng to lớn này, báo cáo tập trung vào việc phân tích sâu một trong những bộ dữ liệu kinh điển và có giá trị lịch sử nhất trong lĩnh vực y tế: bộ dữ liệu Pima Indians Diabetes. Được thu thập từ một quần thể có tỷ lệ mắc bệnh tiểu đường cao kỷ lục, bộ dữ liệu này không chỉ là một tập hợp các con số mà còn là một "phòng thí nghiệm" lý tưởng để khám phá các yếu tố nguy cơ và kiểm nghiệm các phương pháp dự báo.

Trong phạm vi khảo sát đề tài này, nhóm chúng tôi sẽ thực hiện một hành trình phân tích toàn diện, bắt đầu từ việc khám phá và làm sạch dữ liệu, giải quyết các thách thức thực tế như dữ liệu thiếu và mất cân bằng. Thông qua các phương pháp thống kê và trực quan hóa, chúng tôi sẽ làm sáng tỏ mối liên hệ phức tạp giữa các chỉ số sinh học và nguy cơ mắc bệnh, đối chiếu các phát hiện với nền tảng kiến thức y khoa từ các tổ chức uy tín như Tổ chức Y tế Thế giới (WHO). Báo cáo không chỉ dừng lại ở việc phân tích mà còn đặt nền móng vững chắc cho việc xây dựng và đánh giá các mô hình dự báo trong tương lai, qua đó hy vọng đóng góp một phần vào nỗ lực chung nhằm ứng dụng công nghệ để cải thiện sức khỏe cộng đồng.

LỜI CẢM ƠN

Để hoàn thành bài báo cáo nhỏ này, bên cạnh sự nỗ lực của các thành viên trong nhóm, chúng em đã nhận được rất nhiều sự quan tâm, giúp đỡ và động viên từ giảng viên và bạn bè.

Trước hết, chúng em xin gửi lời cảm ơn chân thành và sâu sắc nhất đến Giảng viên hướng dẫn, thầy Đỗ Như Tài. Trong suốt quá trình học tập và hoàn thành báo cáo này, thầy đã tận tình chỉ bảo, truyền đạt cho chúng em những kiến thức nền tảng quý báu, định hướng phương pháp nghiên cứu khoa học và luôn đưa ra những góp ý xác đáng để bài báo cáo được hoàn thiện hơn. Sự nhiệt tình và tâm huyết của thầy là nguồn động lực lớn lao giúp chúng em vượt qua những khó khăn và hoàn thành tốt nhiệm vụ của mình.

Chúng em cũng xin trân trọng cảm ơn Ban Giám hiệu Trường Đại học Sài Gòn và quý thầy cô trong Khoa Công nghệ Thông tin đã tạo mọi điều kiện thuận lợi về cơ sở vật chất và môi trường học tập để chúng em có thể hoàn thiện báo cáo này.

Cuối cùng, xin cảm ơn tất cả các thành viên trong nhóm đã cùng nhau đoàn kết, hợp tác, nỗ lực và hỗ trợ lẫn nhau trong suốt quá trình thực hiện.

Mặc dù đã rất cố gắng, nhưng do kiến thức và kinh nghiệm còn hạn chế, bài báo cáo chắc chắn không thể tránh khỏi những thiếu sót. Chúng em rất mong nhận được những ý kiến đóng góp quý báu từ thầy để đề tài được hoàn thiện hơn nữa.

BẢNG PHÂN CÔNG CÔNG VIỆC

MSSV	Họ tên	Công việc
3123410274	**Lư Hồng Phúc**	**Viết code khảo sát dữ liệu**
3123410387	**Nguyễn Hữu Tri**	**Viết báo cáo khảo sát dữ liệu**

TÓM TẮT

Báo cáo này trình bày một phân tích dữ liệu khám phá toàn diện trên bộ dữ liệu Pima Indians Diabetes nhằm xác định các yếu tố nguy cơ chính của bệnh tiểu đường type 2. Nghiên cứu tập trung vào việc kết nối các phát hiện thống kê với kiến thức y khoa nền tảng để xây dựng một bức tranh đa chiều về bệnh. Báo cáo cũng nhấn mạnh tầm quan trọng của việc xử lý các vấn đề chất lượng dữ liệu như giá trị thiếu ẩn và mất cân bằng lớp, đồng thời đặt nền móng cho việc xây dựng các mô hình học máy dự báo trong tương lai.

- **Lý do nghiên cứu:** Bệnh tiểu đường là một thách thức sức khỏe toàn cầu với những biến chứng nguy hiểm, đòi hỏi phải có các phương pháp phát hiện sớm và hiệu quả. Sự phát triển của khoa học dữ liệu và học máy mang lại tiềm năng to lớn trong việc xây dựng các hệ thống sàng lọc thông minh, chi phí thấp để hỗ trợ chẩn đoán và can thiệp y tế kịp thời.
- **Mục tiêu nghiên cứu:**
  - Phân tích sâu bộ dữ liệu Pima Indians Diabetes để xác định các yếu tố nguy cơ chính liên quan đến bệnh tiểu đường.
  - Xây dựng một bài báo mạch lạc, kết nối các phát hiện thống kê với kiến thức y khoa từ các tài liệu của WHO và NDDG.
  - Đánh giá và đề xuất các phương pháp xử lý các thách thức về chất lượng dữ liệu như giá trị thiếu và mất cân bằng lớp.
  - Đặt nền móng cho việc xây dựng các mô hình học máy có khả năng dự báo chính xác.
- **Phương pháp nghiên cứu:**
  - **Dữ liệu:** Sử dụng bộ dữ liệu Pima Indians Diabetes gồm 768 mẫu và 8 đặc trưng đầu vào.
  - **Tiền xử lý:** Phát hiện và xử lý các giá trị `0` bất hợp lý trong các biến lâm sàng (như `Glucose`, `BMI`, `Insulin`) bằng cách thay thế chúng bằng giá trị trung vị (median) của mỗi cột để bảo toàn dữ liệu.
  - **Phân tích:** Áp dụng phương pháp Phân tích Dữ liệu Khám phá (EDA), bao gồm thống kê mô tả, phân tích đơn biến (kiểm tra phân phối), phân tích song biến (so sánh giữa hai nhóm có và không có bệnh), và phân tích đa biến (ma trận tương quan) để tìm ra các mối liên hệ và các yếu tố quan trọng.
- **Kết quả chính:**
  - Các yếu tố dự báo quan trọng nhất cho bệnh tiểu đường là nồng độ `Glucose` cao, chỉ số `BMI` cao, và `Age` (tuổi) lớn.
  - Các yếu tố khác có ảnh hưởng đáng kể bao gồm `DiabetesPedigreeFunction` (yếu tố di truyền) và `Pregnancies` (số lần mang thai).
  - Dữ liệu tồn tại vấn đề chất lượng nghiêm trọng: nhiều giá trị thiếu được mã hóa bằng `0` và sự mất cân bằng lớp rõ rệt (65.1% không mắc bệnh so với 34.9% mắc bệnh).
  - Phân tích tương quan cho thấy mối liên hệ giữa các biến, chẳng hạn như `Age` và `Pregnancies`, và khả năng đa cộng tuyến giữa các biến liên quan đến mỡ cơ thể (`BMI`, `SkinThickness`).

- **Kết luận chính:**
  - Phân tích dữ liệu đã xác nhận và định lượng thành công các yếu tố nguy cơ y khoa đã biết của bệnh tiểu đường type 2.
  - Việc xử lý cẩn thận các vấn đề về chất lượng dữ liệu là bước cực kỳ quan trọng để đảm bảo tính hợp lệ của phân tích và là tiền đề cho việc xây dựng mô hình hiệu quả.
  - Các kết quả từ EDA cung cấp một nền tảng vững chắc và những hiểu biết sâu sắc, định hướng cho việc lựa chọn đặc trưng và xây dựng các mô hình học máy tiên tiến để dự báo sớm nguy cơ mắc bệnh.

MỤC LỤC

I: Giới thiệu sơ lược về đề tài báo cáo

Bệnh tiểu đường (Diabetes Mellitus) là một rối loạn chuyển hóa mạn tính, đặc trưng bởi tình trạng tăng đường huyết, gây ra bởi sự thiếu hụt insulin, kháng insulin, hoặc cả hai. Đây là một vấn đề sức khỏe cộng đồng toàn cầu với tỷ lệ mắc bệnh ngày càng tăng, trở thành một mối đe dọa lớn trên toàn thế giới. Tiểu đường type 2 chiếm đa số các trường hợp, thường liên quan đến các yếu tố lối sống và di truyền. Hậu quả của việc chẩn đoán muộn là vô cùng nghiêm trọng, có thể dẫn đến các biến chứng nặng nề về tim mạch, suy thận, mù lòa và tổn thương thần kinh, làm suy giảm chất lượng cuộc sống và tăng gánh nặng kinh tế. Do đó, việc phát hiện sớm và dự báo chính xác nguy cơ mắc bệnh đóng vai trò cực kỳ quan trọng, cho phép can thiệp y tế kịp thời và thay đổi lối sống, từ đó giảm thiểu gánh nặng bệnh tật cho cả cá nhân và toàn xã hội.

Nghiên cứu này được thực hiện với mục tiêu phân tích sâu tập dữ liệu Pima Indians Diabetes để xác định các yếu tố nguy cơ chính, xây dựng một "câu chuyện dữ liệu" mạch lạc, và đặt nền móng cho việc xây dựng các mô hình học máy có khả năng dự báo chính xác, làm cơ sở cho các hệ thống hỗ trợ ra quyết định lâm sàng trong tương lai.

II: Cơ sở lý thuyết/ Nghiên cứu liên quan

II.1. Giới thiệu về bệnh tiểu đường

Bệnh tiểu đường là một nhóm các rối loạn chuyển hóa đặc trưng bởi tình trạng tăng đường huyết kéo dài. Các dạng chính bao gồm:

- **Tiểu đường Type 1:** Gây ra bởi sự phá hủy các tế bào beta của tuyến tụy, thường do một quá trình tự miễn, dẫn đến thiếu hụt insulin tuyệt đối.
- **Tiểu đường Type 2:** Là dạng phổ biến nhất, đặc trưng bởi sự kết hợp giữa kháng insulin (cơ thể không sử dụng insulin hiệu quả) và thiếu hụt insulin tương đối. Đây là loại tiểu đường được tập trung nghiên cứu trong tập dữ liệu này.
- **Tiểu đường thai kỳ (Gestational Diabetes):** Xuất hiện trong quá trình mang thai.

Ngoài ra, y học hiện đại còn công nhận các trạng thái tiền tiểu đường như **Suy giảm Dung nạp Glucose (Impaired Glucose Tolerance - IGT)**, là tình trạng có nguy cơ cao tiến triển thành tiểu đường type 2. Các yếu tố nguy cơ thường tập hợp lại thành

Hội chứng Chuyển hóa (Metabolic Syndrome), bao gồm béo phì, tăng huyết áp, rối loạn lipid máu và tăng đường huyết, làm tăng đáng kể nguy cơ mắc bệnh tim mạch và tiểu đường

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

Người Pima bản địa ở Arizona (Mỹ) là một quần thể có tỷ lệ mắc bệnh tiểu đường type 2 cao bất thường, khiến họ trở thành đối tượng quan trọng cho các nghiên cứu về di truyền và các yếu tố nguy cơ của bệnh. Dữ liệu được thu thập bởi Viện Quốc gia về Bệnh Tiểu đường, Tiêu hóa và Thận của Hoa Kỳ (NIDDK) từ những năm 1965, cung cấp một nguồn tài nguyên quý giá cho việc nghiên cứu bệnh tiểu đường theo chiều dọc.

II.3. Bộ dữ liệu Pima Indians Diabetes

Đây là một trong những tập dữ liệu kinh điển trong lĩnh vực học máy y tế, được sử dụng lần đầu trong một nghiên cứu tiên phong năm 1988 để thử nghiệm thuật toán học ADAP.

- **Số mẫu:** 768 bệnh nhân nữ.
- **Đối tượng:** Phụ nữ từ 21 tuổi trở lên thuộc tộc người Pima.
- **Số thuộc tính:** 8 đặc trưng đầu vào và 1 biến mục tiêu (`Outcome`).
- **Biến mục tiêu:** `Outcome` là biến nhị phân, với `1` là mắc bệnh tiểu đường và `0` là không mắc bệnh.

II.4. Ý nghĩa của các đặc trưng trong dữ liệu

1. **Pregnancies (Số lần mang thai):** Biến này ghi nhận số lần mang thai của bệnh nhân. Về mặt y học, thai kỳ là một giai đoạn gây ra những thay đổi chuyển hóa và nội tiết tố đáng kể. Một số phụ nữ phát triển tình trạng "tiểu đường thai kỳ" (Gestational Diabetes), một dạng không dung nạp glucose được phát hiện lần đầu trong thai kỳ. Mặc dù tình trạng này thường hết sau khi sinh, những phụ nữ đã từng bị tiểu đường thai kỳ có nguy cơ cao hơn đáng kể trong việc phát triển bệnh tiểu đường type 2 sau này trong đời. Do đó, số lần mang thai có thể là một chỉ số gián tiếp về những "thử thách" chuyển hóa mà cơ thể đã trải qua.

2. **Glucose (Nồng độ glucose huyết tương sau 2 giờ trong nghiệm pháp dung nạp glucose đường uống - OGTT):** Đây là một trong những chỉ số quan trọng nhất để chẩn đoán bệnh tiểu đường. Nghiệm pháp OGTT đánh giá khả năng của cơ thể trong việc chuyển hóa một lượng đường chuẩn sau một thời gian nhịn ăn. Nồng độ glucose cao sau 2 giờ cho thấy cơ thể không thể sử dụng hoặc lưu trữ glucose một cách hiệu quả, đây là dấu hiệu của tình trạng kháng insulin hoặc thiếu hụt insulin. Các tổ chức y tế như WHO và NDDG đã đặt ra các ngưỡng cụ thể cho chỉ số này để chẩn đoán bệnh tiểu đường (ví dụ, ≥200 mg/dl). Do đó, biến này là một yếu tố dự báo trực tiếp và mạnh mẽ.

3. **BloodPressure (Huyết áp tâm trương - mm Hg):** Huyết áp cao (tăng huyết áp) là một yếu tố nguy cơ tim mạch phổ biến và là một thành phần cốt lõi của "Hội chứng Chuyển hóa". Hội chứng này là một tập hợp các tình trạng (bao gồm béo phì trung tâm, rối loạn lipid máu và tăng đường huyết) làm tăng đáng kể nguy cơ mắc bệnh tim mạch và tiểu đường type 2. Mối liên hệ giữa tăng huyết áp và tiểu đường rất phức tạp, có thể cùng chia sẻ các cơ chế bệnh sinh chung như kháng insulin và rối loạn chức năng nội mô

4. **SkinThickness (Độ dày nếp gấp da cơ tam đầu - mm):** Chỉ số này được sử dụng như một thước đo gián tiếp về lượng mỡ dưới da của cơ thể. Lượng mỡ cơ thể dư thừa, đặc biệt là béo phì, là một trong những yếu tố nguy cơ hàng đầu gây ra kháng insulin, tiền đề chính của bệnh tiểu đường type 2. Mặc dù không chính xác bằng các phương pháp đo lường thành phần cơ thể khác, độ dày nếp

gấp da là một phương pháp đơn giản và không xâm lấn để ước tính tình trạng mỡ của cơ thể.

5. **Insulin (Nồng độ insulin huyết thanh sau 2 giờ - mu U/ml):** Insulin là hormone do tuyến tụy tiết ra để giúp các tế bào hấp thụ glucose từ máu. Trong bối cảnh của OGTT, nồng độ insulin sau 2 giờ phản ánh phản ứng của tuyến tụy đối với một lượng đường lớn. Ở những người khỏe mạnh, insulin sẽ tăng lên để xử lý glucose và sau đó giảm xuống. Tuy nhiên, ở giai đoạn đầu của bệnh tiểu đường type 2, cơ thể phát triển tình trạng kháng insulin, khiến tuyến tụy phải làm việc quá sức và tiết ra lượng insulin cao hơn bình thường (tăng insulin máu) để duy trì mức đường huyết ổn định. Theo thời gian, các tế bào beta của tuyến tụy có thể bị suy kiệt, dẫn đến giảm sản xuất insulin. Do đó, một giá trị insulin cao bất thường có thể là dấu hiệu của kháng insulin.

6. **BMI (Chỉ số khối cơ thể - kg/m²):** BMI là một chỉ số đơn giản về mối quan hệ giữa cân nặng và chiều cao, được sử dụng rộng rãi để phân loại tình trạng thiếu cân, bình thường, thừa cân và béo phì. Béo phì (BMI cao) là một yếu tố nguy cơ đã được chứng minh rõ ràng cho bệnh tiểu đường type 2. Các mô mỡ dư thừa, đặc biệt là mỡ nội tạng, giải phóng các axit béo tự do và các chất gây viêm, góp phần gây ra tình trạng kháng insulin toàn thân.

7. **DiabetesPedigreeFunction (Chỉ số nguy cơ di truyền):** Biến này là một thước đo tổng hợp, được thiết kế để định lượng ảnh hưởng di truyền dựa trên tiền sử bệnh tiểu đường trong gia đình của bệnh nhân. Nó tính đến cả số lượng người thân mắc bệnh và mức độ quan hệ di truyền (ví dụ: cha mẹ, anh chị em so với anh em họ). Bệnh tiểu đường type 2 có một thành phần di truyền mạnh mẽ; có tiền sử gia đình mắc bệnh làm tăng đáng kể nguy cơ của một cá nhân. Biến này cố gắng nắm bắt yếu tố nguy cơ phức tạp đó bằng một con số duy nhất.

8. **Age (Tuổi của bệnh nhân - năm):** Tuổi tác là một yếu tố nguy cơ không thể thay đổi đối với nhiều bệnh mạn tính, bao gồm cả tiểu đường type 2. Nguy cơ mắc bệnh tăng lên theo tuổi. Điều này có thể do sự tích tụ của các yếu tố lối sống không lành mạnh theo thời gian, sự gia tăng tự nhiên của kháng insulin liên quan đến tuổi tác, và sự suy giảm chức năng của các tế bào beta sản xuất insulin trong tuyến tụy.

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"

Báo cáo của WHO năm 1999 là một tài liệu nền tảng, cập nhật các tiêu chí chẩn đoán và hệ thống phân loại bệnh tiểu đường. Một trong những thay đổi quan trọng nhất là việc hạ ngưỡng chẩn đoán cho **Glucose huyết tương lúc đói (FPG)** từ ≥140 mg/dl xuống ≥126 mg/dl (7.0 mmol/L). Sự thay đổi này phản ánh một triết lý mới: chẩn đoán dựa trên nguy cơ phát triển biến chứng vi mạch (như bệnh võng mạc) thay vì chỉ dựa trên các triệu chứng lâm sàng rõ rệt. Báo cáo cũng chính thức hóa các khái niệm về IGT và Hội chứng Chuyển hóa, nhấn mạnh tầm quan trọng của việc xác định các trạng thái nguy cơ.

II.5.2. Nghiên cứu 2 (Paper 2): Smith et al. (1988) - "Using the ADAP learning algorithm to forecast the onset of diabetes mellitus"

Đây là một nghiên cứu tiên phong, áp dụng một dạng sơ khai của mạng nơ-ron (thuật toán ADAP) để dự báo sự khởi phát của bệnh tiểu đường trên chính tập dữ liệu Pima Indians. Nghiên cứu đã sử dụng 8 biến đầu vào tương tự và đạt được độ nhạy (sensitivity) và độ đặc hiệu (specificity) cùng ở mức 76%. Kết quả này không chỉ chứng minh tiềm năng của học máy trong y tế từ rất sớm mà còn cung cấp một đường cơ sở (baseline) quan trọng để so sánh với các thuật toán hiện đại.

II.5.3. Nghiên cứu 3: Hệ thống phân loại bệnh tiểu đường (NDDG, 1979)

Trước báo cáo của WHO, Nhóm Dữ liệu Bệnh tiểu đường Quốc gia Hoa Kỳ (NDDG) đã công bố một hệ thống phân loại vào năm 1979, mang lại trật tự cho một lĩnh vực vốn "hỗn loạn" về danh pháp. NDDG đã đề xuất các thuật ngữ **IDDM (Tiểu đường phụ thuộc insulin)** và **NIDDM (Tiểu đường không phụ thuộc insulin)**, đặt nền móng cho việc phân loại dựa trên đặc điểm lâm sàng và điều trị. Họ cũng đưa ra các tiêu chí chẩn đoán cụ thể, bao gồm ngưỡng FPG ≥140 mg/dl và các giá trị cho OGTT, vốn là cơ sở để WHO sau này điều chỉnh.

II.5.4. Tổng Kết Nghiên cứu

Các tài liệu này cung cấp một khuôn khổ y khoa và lịch sử vững chắc cho việc phân tích tập dữ liệu Pima. Báo cáo của NDDG và WHO định nghĩa các tiêu chí chẩn đoán và các yếu tố nguy cơ, giúp chúng ta diễn giải các kết quả thống kê một cách có ý nghĩa. Nghiên cứu của Smith et al. cho thấy giá trị lịch sử của tập dữ liệu và đặt ra một tiêu chuẩn hiệu suất để các mô hình hiện đại hướng tới.

III. Dữ liệu và phương pháp

III.1. Dữ liệu

Nghiên cứu sử dụng bộ dữ liệu Pima Indians Diabetes, bao gồm 768 mẫu và 9 thuộc tính (8 đặc trưng, 1 nhãn).

III.2. Vấn đề dữ liệu và tiền xử lý sơ bộ

- **Giá trị thiếu ẩn:** Nhiều biến số lâm sàng như `Glucose`, `BloodPressure`, `SkinThickness`, `Insulin`, và `BMI` chứa các giá trị `0`, vốn không hợp lý về mặt sinh học và được coi là giá trị thiếu.
- **Mất cân bằng lớp:** Dữ liệu bị mất cân bằng, với số lượng bệnh nhân không mắc bệnh (lớp `0`) nhiều gần gấp đôi số bệnh nhân mắc bệnh (lớp `1`). Điều này có thể làm cho mô hình học máy thiên vị về phía lớp đa số.
- **Giá trị ngoại lệ (Outliers):** Một số biến, đặc biệt là `Insulin`, có các giá trị rất cao, có thể ảnh hưởng đến các mô hình nhạy cảm với ngoại lệ.

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

Phân tích dữ liệu khám phá được thực hiện để hiểu sâu hơn về dữ liệu thông qua:

- **Thống kê mô tả:** Tính toán các chỉ số trung tâm và độ phân tán.
- **Phân tích đơn biến:** Kiểm tra phân phối của từng biến số bằng biểu đồ histogram và density plot.
- **Phân tích song biến:** So sánh sự khác biệt của các biến số giữa hai nhóm `Outcome` bằng biểu đồ hộp (boxplot) và violin plot.
- **Phân tích đa biến:** Đánh giá mối tương quan tuyến tính giữa các biến bằng ma trận tương quan (heatmap).

III.4. Công cụ và môi trường phân tích

Phân tích được thực hiện trong môi trường Python, sử dụng các thư viện phổ biến như Pandas để xử lý dữ liệu, Matplotlib và Seaborn để trực quan hóa.

IV. Phân tích dữ liệu (EDA)

Đây là phần trọng tâm của báo cáo, trình bày chi tiết quá trình khám phá dữ liệu thô (raw data) của tập dữ liệu Pima Indians Diabetes. Mục tiêu của phần này là để hiểu sâu sắc về cấu trúc, đặc điểm phân phối, mối quan hệ giữa các biến, đồng thời xác định các vấn đề tiềm ẩn về chất lượng dữ liệu như giá trị thiếu, mẫu bất thường và các giá trị ngoại lệ. Các kết quả từ phân tích này sẽ là cơ sở vững chắc cho các bước tiền xử lý và xây dựng mô hình học máy sau này.

IV.1. Thống kê mô tả

Bước đầu tiên trong mọi quy trình phân tích dữ liệu là kiểm tra các đặc điểm tổng quan để có cái nhìn ban đầu về dữ liệu.

IV.1.1. Kích thước và kiểu dữ liệu

- **Kích thước (Shape):** Tập dữ liệu bao gồm **768 mẫu (dòng)**, tương ứng với 768 bệnh nhân, và **9 thuộc tính (cột)**, trong đó có 8 đặc trưng đầu vào và 1 biến mục tiêu `Outcome`.
- **Kiểu dữ liệu (Data Types):** Tất cả 9 cột đều có kiểu dữ liệu số (`int64` hoặc `float64`). Điều này cho thấy dữ liệu đã được mã hóa sẵn sàng cho việc phân tích và không yêu cầu các bước chuyển đổi kiểu dữ liệu phức tạp như xử lý biến chuỗi.

IV.1.2. Hiển thị một số thông tin về dữ liệu

- Kiểu dữ liệu của từng cột
    - Pregnancies: int64

- Glucose: int64
- BloodPressure: int64
- SkinThickness: int64
- Insulin: int64
- BMI: float64
- DiabetesPedigreeFunction float64
- Age: int64
- Outcome: int64

- 5 dòng đầu

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	Diabetes Pedigree Function	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1

- 5 dòng cuối

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	Diabetes Pedigree Function	Age	Outcome
763	10	101	76	48	180	32.9	0.171	63	0
764	2	122	70	27	0	36.8	0.340	27	0
765	5	121	72	23	112	26.2	0.245	30	0
766	1	126	60	0	0	30.1	0.349	47	1
767	1	93	70	31	0	30.4	0.315	23	0

Nhận xét

- Dữ liệu có 8 tính chất để phân lớp: Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, Age.
- Giá trị cho 8 tính chất được tính bằng:
  - mm: SkinThickness.
  - mg/dL: Glucose.
  - mmHg: BloodPressure.
  - μU/mL: Insulin.
  - kg/m²: BMI.
  - Năm: Age.
  - Không đơn vị: Pregnancies, DiabetesPedigreeFunction.
- Tổng số dòng dữ liệu là 768 dòng
- Dữ liệu để phân lớp ở cột Outcome (nếu outcome=0 -> không có bệnh tiểu đường, Outcome=1 -> có bệnh tiểu đường)

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

Bảng thống kê mô tả (`describe()`) cung cấp các chỉ số thống kê cốt lõi cho từng biến số ở trạng thái dữ liệu thô, trước khi xử lý.

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	Diabetes Pedigree Function	Age	Outcome
count	768.00	768.00	768.00	768.00	768.00	768.00	768.00	768.00	768.00
mean	3.85	120.89	69.11	20.54	79.80	31.99	0.47	33.24	0.35
std	3.37	31.97	19.36	15.95	115.24	7.88	0.33	11.76	0.48
min	0.00	0.00	0.00	0.00	0.00	0.00	0.08	21.00	0.00
25%	1.00	99.00	62.00	0.00	0.00	27.30	0.24	24.00	0.00
50%	3.00	117.00	72.00	23.00	30.50	32.00	0.37	29.00	0.00
75%	6.00	140.25	80.00	32.00	127.25	36.60	0.63	41.00	1.00
max	17.00	199.00	122.00	99.00	846.00	67.10	2.42	81.00	1.00

Nhận xét:

- Phát hiện quan trọng nhất từ bảng này là sự hiện diện của giá trị tối thiểu (`min`) bằng **0** ở các cột `Glucose`, `BloodPressure`, `SkinThickness`, `Insulin`, và `BMI`. Về mặt y sinh, các chỉ số này không thể bằng 0 ở một người còn sống. Điều này cho thấy một vấn đề nghiêm trọng về chất lượng dữ liệu: các giá trị thiếu đã được mã hóa một cách ngầm định bằng số 0.
- Độ lệch chuẩn (`std`) của `Insulin` (115.24) rất lớn so với giá trị trung bình (79.80), và giá trị `max` (846) cao hơn rất nhiều so với phân vị thứ 75 (127.25), cho thấy sự tồn tại của các giá trị ngoại lệ và phân phối lệch.

IV.1.4. Phân bố biến mục tiêu (Outcome)

	Outcome
0	500
1	268

- **Lớp 0 (Không bệnh):** 500 mẫu, chiếm **65.1%** tổng số.
- **Lớp 1 (Có bệnh):** 268 mẫu, chiếm **34.9%** tổng số.

Nhận xét: Dữ liệu thể hiện sự **mất cân bằng lớp (class imbalance)** rõ rệt, với số lượng bệnh nhân không mắc bệnh nhiều gần gấp đôi số bệnh nhân mắc bệnh. Đây là một yếu tố quan trọng cần được xử lý trong giai đoạn mô hình hóa, vì các thuật toán học máy có xu hướng thiên vị về phía lớp đa số, dẫn đến hiệu suất kém trong việc dự đoán lớp thiểu số (lớp quan trọng hơn trong chẩn đoán y khoa).

IV.1.5 Mối tương quan giữa các tính chất (Correlations)

Sự tương quan (correlation) đề cập đến mối quan hệ giữa hai biến và cách chúng có thể có hoặc không cùng nhau thay đổi.

Phương pháp phổ biến nhất để tính toán tương quan là Pearson's Correlation Coeficient, giả định có một phân phối chuẩn của các thuộc tính liên quan. Tương quan -1 hoặc 1 cho thấy mối tương quan âm hoặc dương đầy đủ tương ứng. Trong khi giá trị 0 hiển thị không tương quan ở tất cả.

$$r = \frac{\sum_{i=1}^{n}(x_i - \hat{x})(y_i - \hat{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \hat{x})^2 \sum_{i=1}^{n}(y_i - \hat{y})^2}}$$

Một số thuật toán học máy như hồi quy tuyến tính và logistic có hiệu suất kém nếu có các thuộc tính tương quan cao trong tập dữ liệu của bạn.

Như vậy, thật sự cần thiết để xem xét tất cả các mối tương quan theo cặp của các thuộc tính trong tập dữ liệu.

	Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes Pedigree Function	Age	Outcome
Pregnancies	1.000000	0.127911	0.208522	0.082989	0.056027	0.021565	-0.033523	0.544341	0.221898
Glucose	0.127911	1.000000	0.218367	0.192991	0.420157	0.230941	0.137060	0.266534	0.492928
Blood Pressure	0.208522	0.218367	1.000000	0.192816	0.072517	0.281268	-0.002763	0.324595	0.166074
Skin Thickness	0.082989	0.192991	0.192816	1.000000	0.158139	0.542398	0.100966	0.127872	0.215299
Insulin	0.056027	0.420157	0.072517	0.158139	1.000000	0.166586	0.098634	0.136734	0.214411
BMI	0.021565	0.230941	0.281268	0.542398	0.166586	1.000000	0.153400	0.025519	0.311924
Diabetes Pedigree Function	-0.033523	0.137060	-0.002763	0.100966	0.098634	0.153400	1.000000	0.033561	0.173844
Age	0.544341	0.266534	0.324595	0.127872	0.136734	0.025519	0.033561	1.000000	0.238356
Outcome	0.221898	0.492928	0.166074	0.215299	0.214411	0.311924	0.173844	0.238356	1.000000

IV.1.6 Xác định ngoại lệ

Định nghĩa: Ngoại lệ (outlier) là các giá trị nằm xa một cách đáng kể so với phần lớn các điểm dữ liệu khác trong cùng một thuộc tính. Các giá trị này có thể xuất phát từ lỗi trong quá trình đo lường, sai sót khi nhập liệu, hoặc chúng có thể là đại diện cho các hiện tượng hiếm nhưng có thật (ví dụ: chỉ số `Insulin` rất cao, lên tới 846 µU/mL trong bộ dữ liệu Diabetes).

Phương pháp xác định: Để phát hiện ngoại lệ một cách hệ thống, phương pháp **Khoảng tứ phân vị (Interquartile Range - IQR)** đã được sử dụng. Một điểm dữ liệu được xem là ngoại lệ nếu nó nằm ngoài khoảng được xác định bởi công thức sau:

$$[Q_1 - 1.5 \times IQR, Q_3 + 1.5 \times IQR]$$

Trong đó:

- Q1 là tứ phân vị thứ nhất (phân vị 25%).
- Q3 là tứ phân vị thứ ba (phân vị 75%).
- IQR=Q3−Q1 là khoảng cách giữa hai tứ phân vị.

Kết quả: Áp dụng phương pháp IQR cho các thuộc tính trong bộ dữ liệu (trừ cột `Outcome`), kết quả phát hiện ngoại lệ được tổng hợp trong bảng dưới đây:

Thuộc tính	Số lượng Ngoại lệ	Giới hạn Dưới	Giới hạn Trên
Pregnancies	4	-6.50	13.50
Glucose	0	39.00	201.00
BloodPressure	14	40.00	104.00
SkinThickness	87	14.50	42.50
Insulin	164	70.43	206.62
BMI	8	13.85	50.25
Diabetes Pedigree Function	29	-0.33	1.20
Age	9	-1.50	66.50

Nhận xét và Đánh giá

- **`Insulin` và `SkinThickness`:** Hai thuộc tính này có số lượng ngoại lệ lớn nhất (lần lượt là 164 và 87). Điều này chủ yếu do phân bố dữ liệu của chúng bị lệch phải mạnh (ví dụ, độ lệch chuẩn của `Insulin` là 118.78). Đáng chú ý, giới hạn dưới được tính toán cho `Insulin` là giá trị âm, điều này không hợp lý trong thực tế và cho thấy sự ảnh hưởng của các giá trị rất lớn đến việc xác định khoảng IQR.

- **`DiabetesPedigreeFunction`:** Thuộc tính này cũng có một lượng ngoại lệ đáng kể (29 giá trị), phân bố cũng bị lệch phải.

- **`Glucose`, `BMI`, và `BloodPressure`:** Mặc dù có ít ngoại lệ hơn, các giá trị ngoại lệ trong các cột này cần được xem xét cẩn thận vì chúng có độ tương quan cao với biến mục tiêu `Outcome` (tương quan tương ứng là 0.467 và 0.293 đối với `Glucose` và `BMI`).

- **Tác động của ngoại lệ:**

- **Phân tích đa biến:** Sự tồn tại của các ngoại lệ, đặc biệt trong `Glucose` và `BMI`, có thể làm méo mó mối quan hệ tuyến tính trên các biểu đồ phân tán (scatter plot) và làm sai lệch nhận định về độ tương quan trên biểu đồ nhiệt (heatmap).

- **Hiệu suất mô hình:** Các ngoại lệ này có khả năng làm tăng phương sai của mô hình học máy. Khi thử nghiệm với mô hình LightGBM, sự hiện diện của chúng đã góp phần làm giảm hiệu suất, thể hiện qua chỉ số F1-score chỉ đạt 0.6275, đặc biệt là ảnh hưởng đến khả năng dự đoán đúng cho lớp thiểu số (`Outcome=1`).

IV.2. Kiểm tra dữ liệu thiếu và bất hợp lý

Dữ liệu không chứa giá trị `Null` hay `NaN` tường minh, và không có dòng nào bị trùng lặp. Tuy nhiên, như đã phát hiện ở trên, các giá trị `0` trong một số cột là các mẫu bất thường, được xem như giá trị thiếu ẩn.

Bảng Thống Kê Số Lượng Giá Trị 0 Bất Hợp Lý Trong Từng Biến

Biến số	Số lượng giá trị 0	Tỷ lệ (%)
Insulin	374	48.7%
SkinThickness	227	29.6%
BloodPressure	35	4.6%
BMI	11	1.4%
Glucose	5	0.7%

- **Nhận xét:** Vấn đề chất lượng dữ liệu là rất nghiêm trọng. Đặc biệt, gần một nửa số bệnh nhân (`48.7%`) thiếu thông tin về `Insulin` và gần một phần ba (`29.6%`) thiếu thông tin về `SkinThickness`. Tổng cộng có **376 trên 768 bệnh nhân (khoảng 49%)** có ít nhất một giá trị bất thường này, cho thấy việc loại bỏ các hàng này sẽ làm mất đi một lượng lớn thông tin quý giá và không phải là một giải pháp khả thi.

IV.3. Phân tích Khám phá Dữ liệu Đơn biến (Univariate EDA)

Phân tích đơn biến được thực hiện để tìm hiểu đặc điểm của từng thuộc tính riêng lẻ trong bộ dữ liệu. Quá trình này bao gồm cả phân tích thống kê mô tả và trực quan hóa dữ liệu.

- **Tiền xử lý:** Trước khi phân tích, các giá trị 0 trong các cột `Glucose`, `BloodPressure`, `SkinThickness`, `Insulin`, và `BMI` được thay thế bằng giá trị trung bình của cột tương ứng. Đây là một bước xử lý cần thiết vì giá trị 0 trong các thuộc tính này thường biểu thị dữ liệu bị thiếu

IV.3.1 Phân tích Thống kê Mô tả (Không sử dụng Đồ thị)

Bảng thống kê mô tả cung cấp cái nhìn tổng quan về các đặc trưng số học của dữ liệu:

Đặc trưng	Trung bình	Độ lệch chuẩn	Tối thiểu	Tối đa
Pregnancies	3.85	3.37	0.00	17.00
Glucose	121.69	30.44	44.00	199.00
BloodPressure	72.41	12.10	24.00	122.00
SkinThickness	29.15	8.79	7.00	99.00
Insulin	155.55	85.02	14.00	846.00
BMI	32.46	6.88	18.20	67.10
Diabetes Pedigree Function	0.47	0.33	0.08	2.42
Age	33.24	11.76	21.00	81.00

Phân bố của biến mục tiêu `Outcome`:

- **Số lượng:** 500 mẫu thuộc lớp `0` (Không bị tiểu đường) và 268 mẫu thuộc lớp `1` (Bị tiểu đường).
- **Tỷ lệ:** Lớp `0` chiếm 65.1%, trong khi lớp `1` chiếm 34.9%.

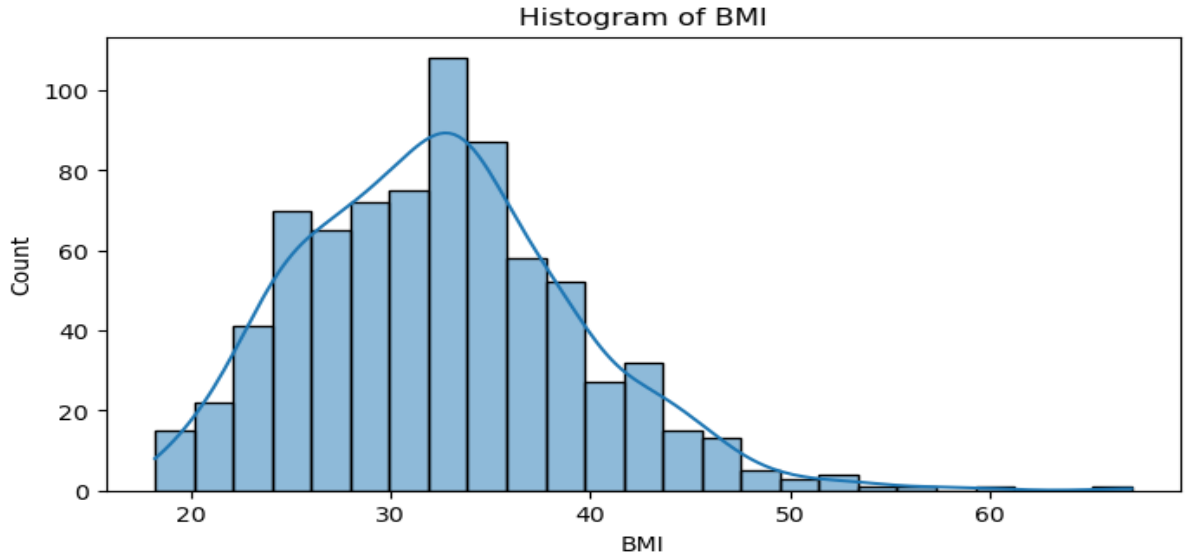
Nhận xét:

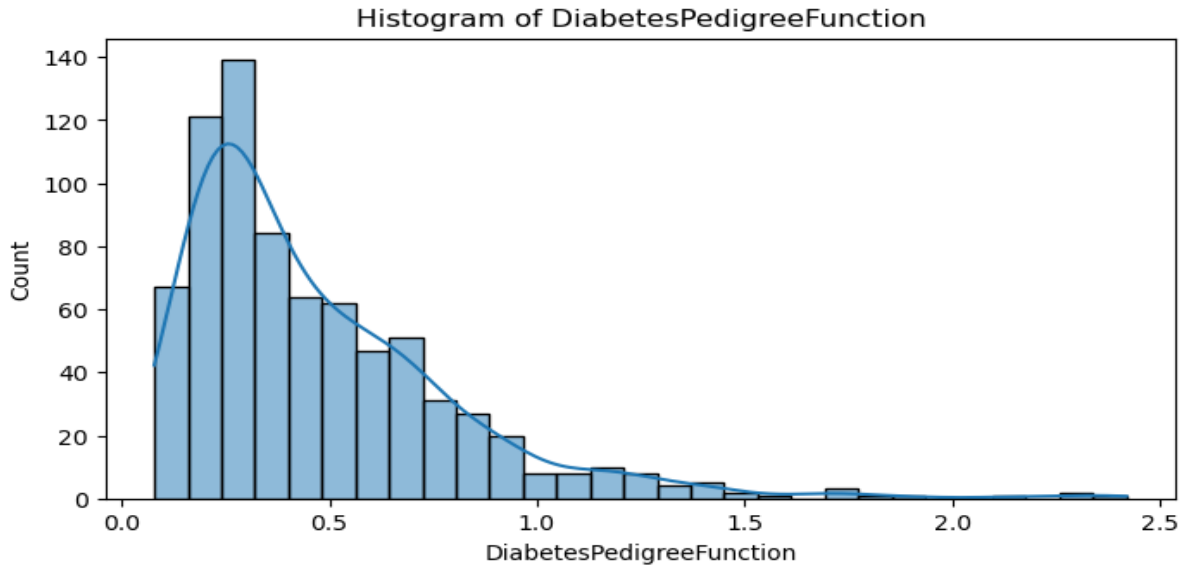
- **Độ lệch và Biến động:** Thuộc tính `Insulin` có độ biến động rất cao (độ lệch chuẩn là 85.02 so với trung bình 155.55) và giá trị tối đa là 846, cho thấy phân bố lệch phải mạnh. Ngược lại, `Glucose` có phân bố gần đối xứng hơn (trung bình 121.69 ≈ trung vị 117.0).
- **Phạm vi Dữ liệu:** `Pregnancies` có giá trị tối đa là 17, đây là một con số cao nhưng vẫn có thể xảy ra trong thực tế.
- **Mất cân bằng Lớp:** Dữ liệu cho thấy sự mất cân bằng rõ rệt giữa hai lớp `Outcome`. Lớp `0` chiếm đa số, điều này có thể khiến mô hình học máy (như LightGBM) có xu hướng dự đoán thiên về lớp này, dẫn đến hiệu suất thấp (chỉ số F1-score thấp) đối với lớp `1`.

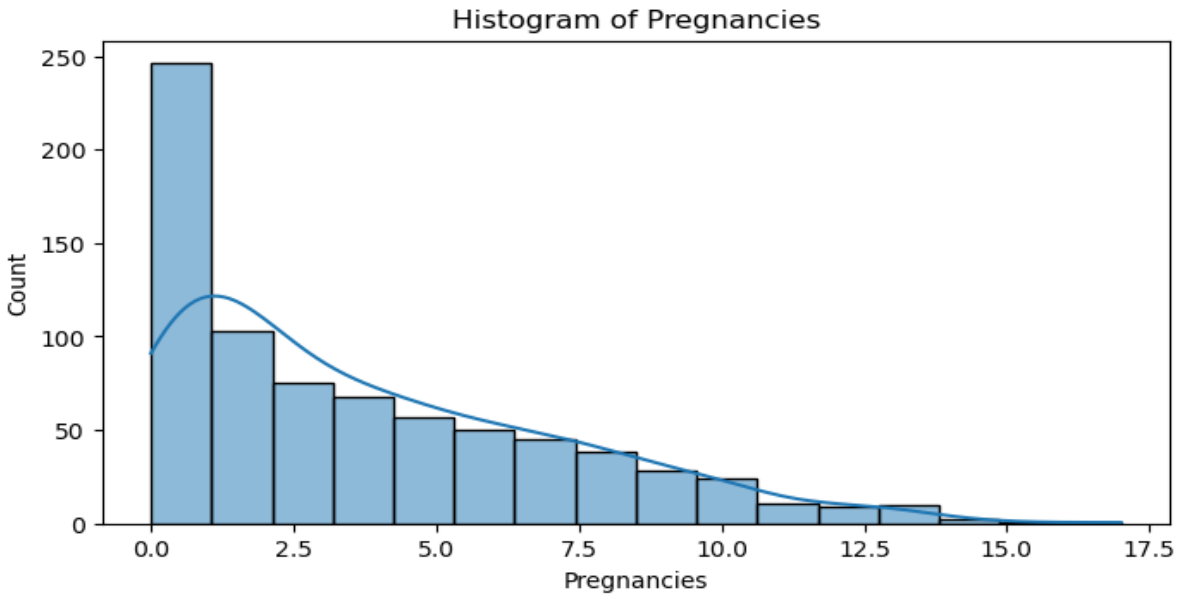
IV.3.2. Phân tích Trực quan hóa (Sử dụng Đồ thị)

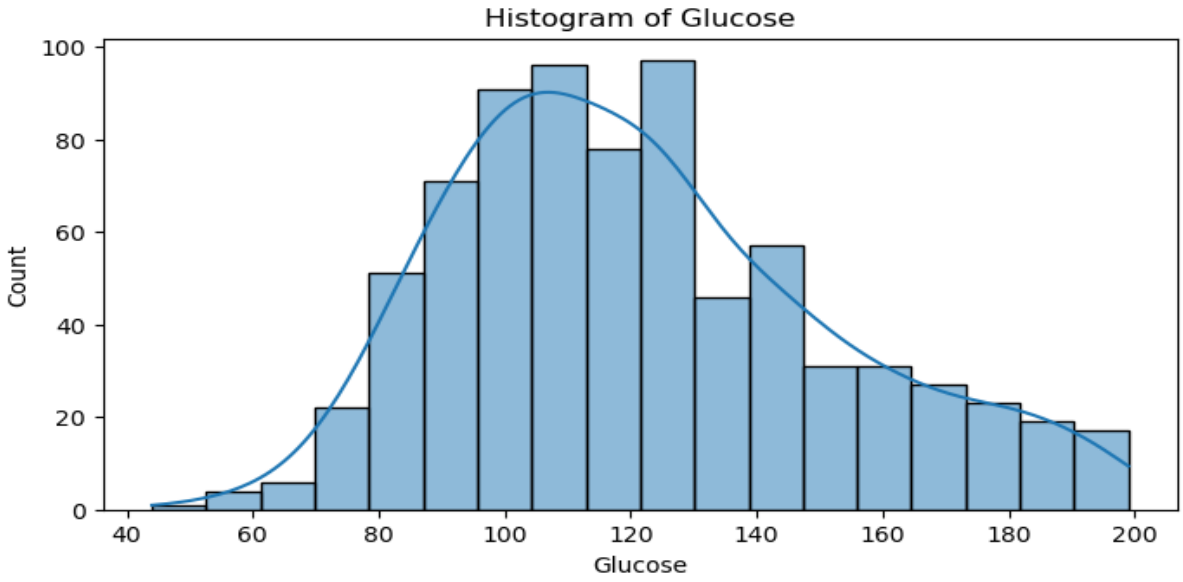
a. Phân bố của các Đặc trưng (Histogram)

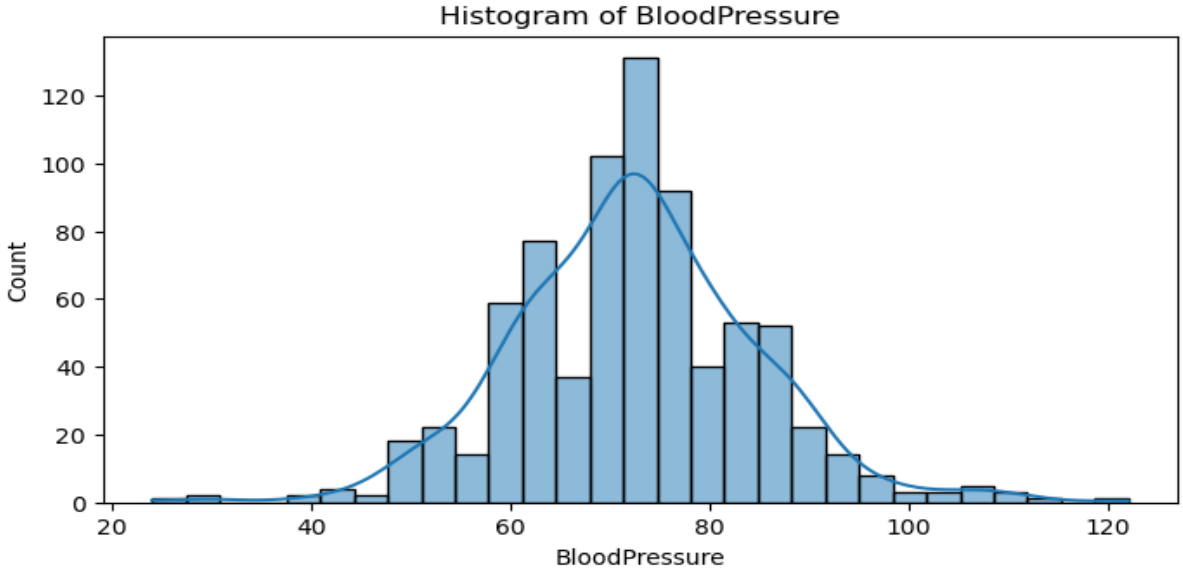
Biểu đồ histogram được sử dụng để kiểm tra hình dạng phân bố của từng thuộc tính.

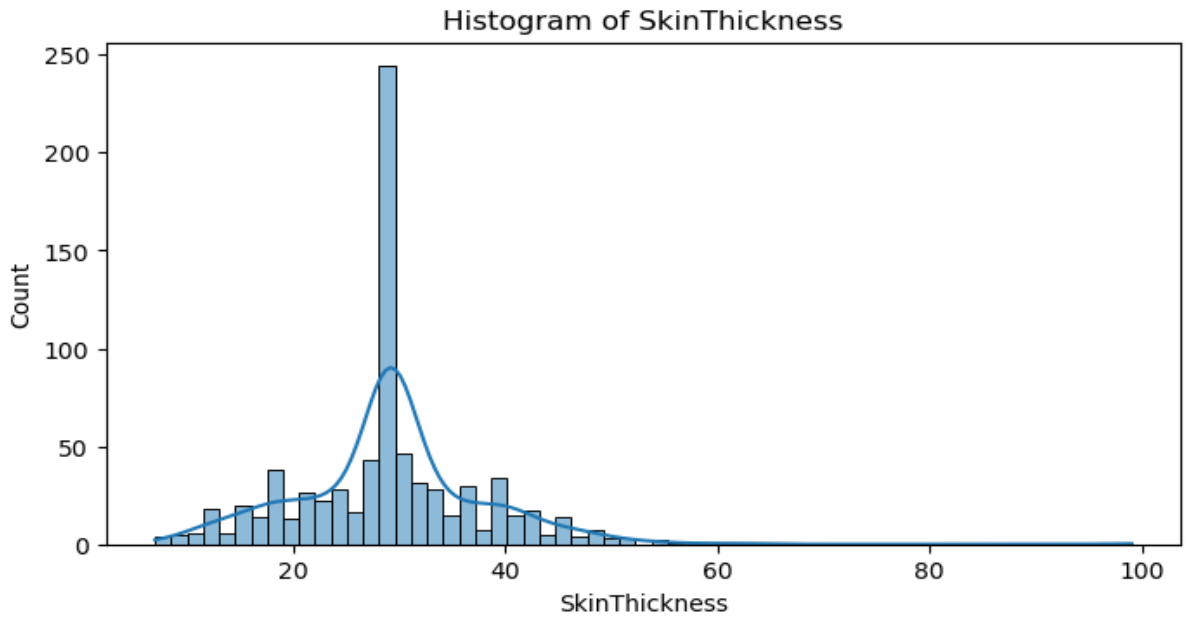


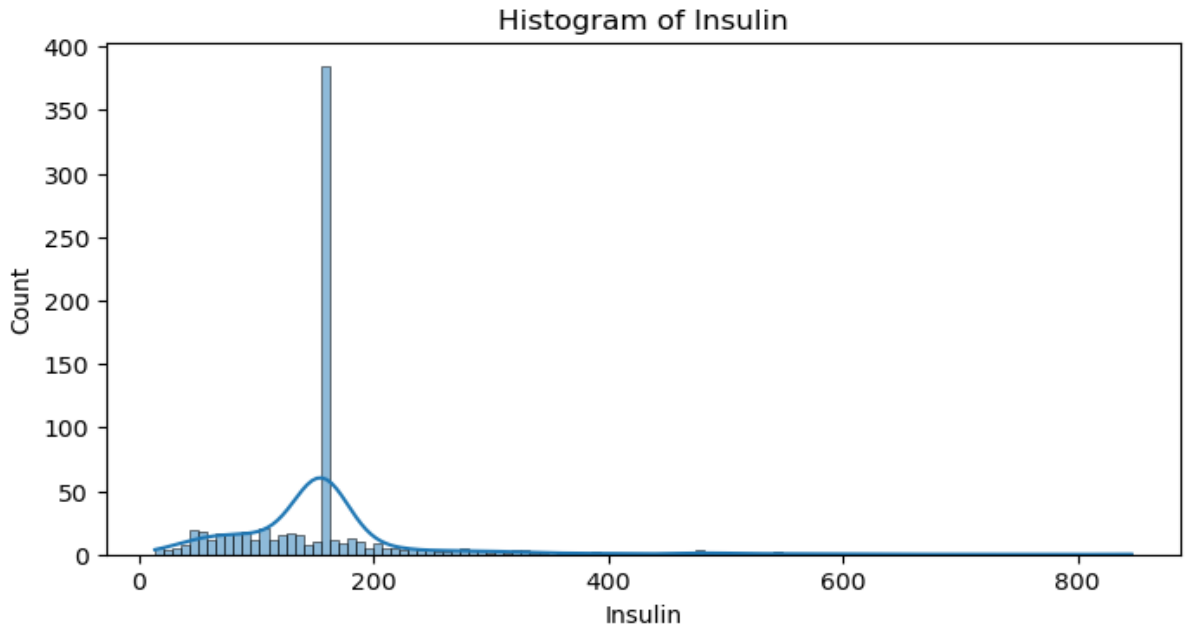



Histogram of Pregnancies
Count
Pregnancies
0
50
100
150
200
250
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5


Histogram of Glucose
Count
Glucose
0
20
40
60
80
100
40
60
80
100
120
140
160
180
200


Histogram of BloodPressure
Count
BloodPressure
0
20
40
60
80
100
120
20
40
60
80
100
120



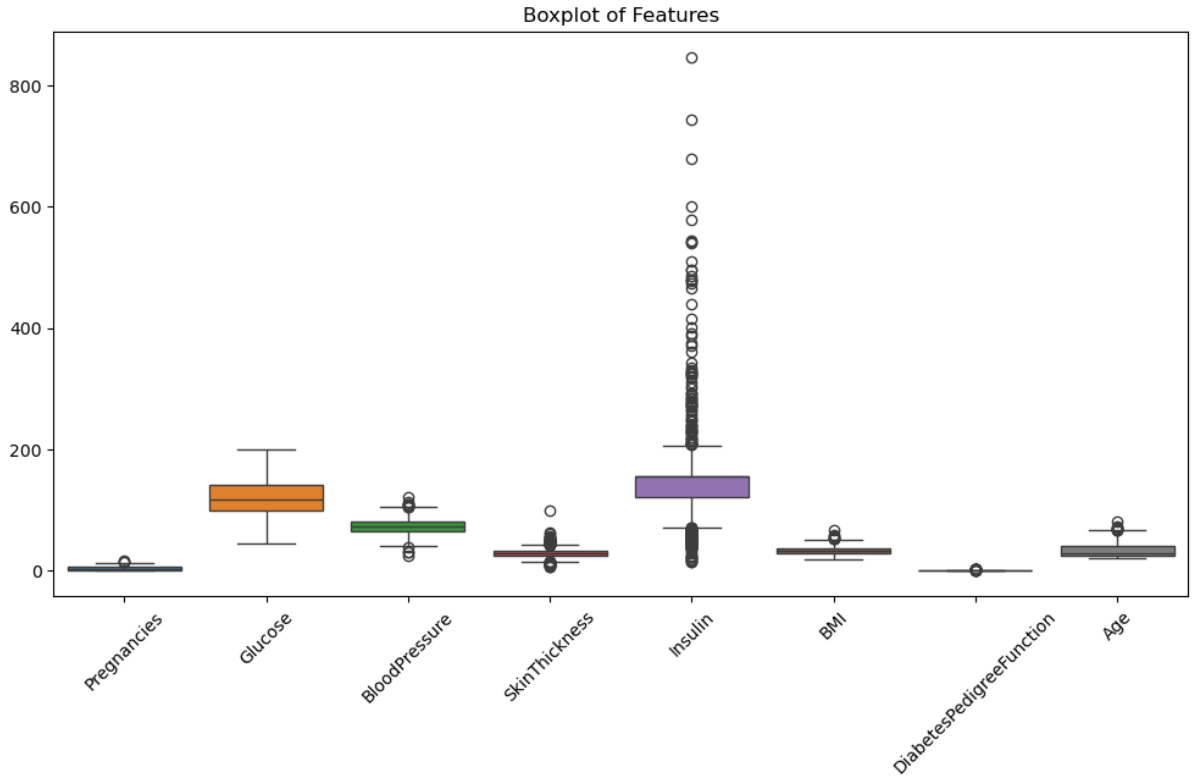


Nhận xét:

- **Phân bố gần chuẩn:** `Glucose`, `BloodPressure` và `BMI` có dạng phân bố gần giống hình chuông (phân bố chuẩn), mặc dù có một chút lệch nhẹ.
- **Phân bố lệch phải:** `Pregnancies`, `SkinThickness`, `Insulin`, `DiabetesPedigreeFunction` và `Age` đều cho thấy sự lệch phải rõ rệt. Đặc biệt, `Insulin` có một đuôi dài về phía bên phải, xác nhận sự tồn tại của các giá trị rất cao.
- **Ý nghĩa:** Việc xác định các phân bố lệch này rất quan trọng. Đối với các mô hình nhạy cảm với giả định về phân bố chuẩn (như hồi quy tuyến tính), các phép biến đổi (ví dụ: log-transform) có thể cần được áp dụng để cải thiện hiệu suất.

b. Phân tích Ngoại lệ (Boxplot)

Biểu đồ hộp trực quan hóa sự phân tán của dữ liệu và giúp xác định các giá trị ngoại lệ.

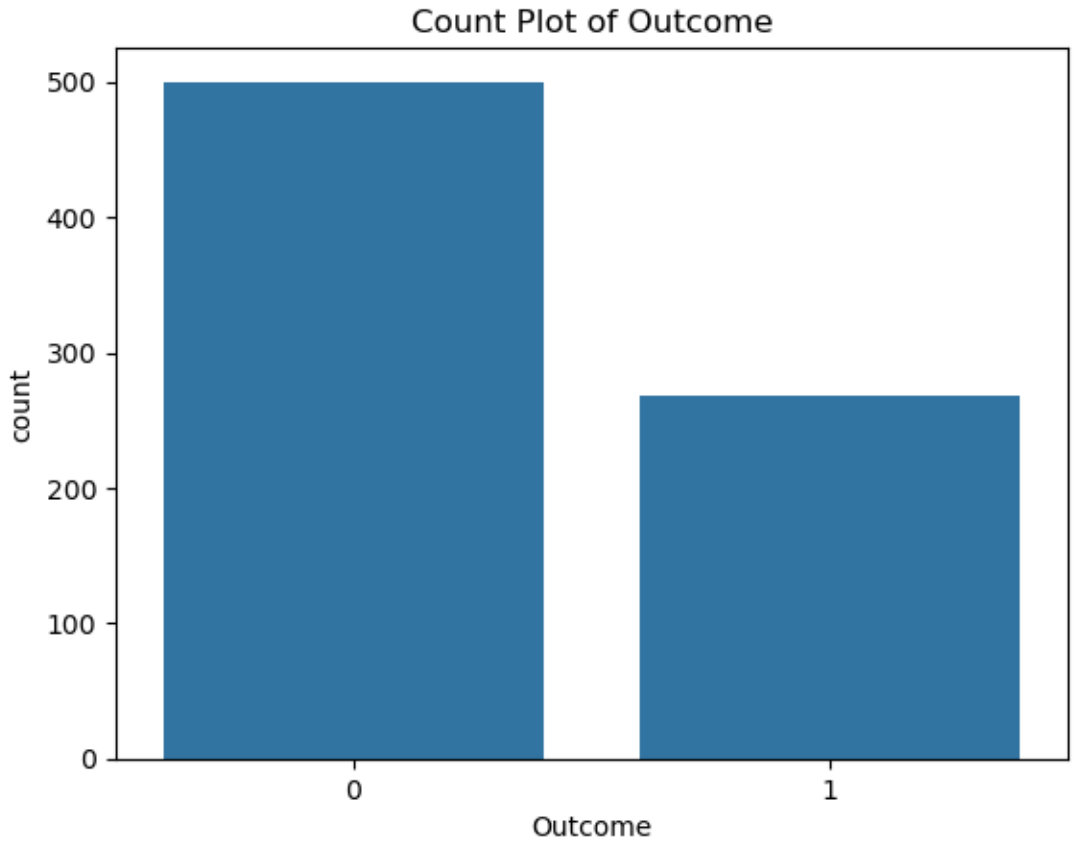


Nhận xét:

- `Insulin` là thuộc tính có nhiều ngoại lệ nhất, với các giá trị vượt xa ngưỡng trên (trên 300 µU/mL).
- `BMI`, `SkinThickness`, `BloodPressure`, và `Age` cũng có một số giá trị ngoại lệ. Ví dụ, `BMI` có các điểm dữ liệu trên 50 kg/m², tương ứng với mức độ béo phì nghiêm trọng.
- **Ý nghĩa:** Các ngoại lệ này có thể ảnh hưởng tiêu cực đến quá trình huấn luyện mô hình, làm tăng phương sai và giảm độ chính xác. Cần có các kỹ thuật xử lý ngoại lệ (như capping hoặc loại bỏ) để cải thiện hiệu suất của mô hình LightGBM.

c. Phân bố của Biến mục tiêu (Count Plot)

Biểu đồ đếm được dùng để xác nhận lại sự mất cân bằng của biến `Outcome`.



Nhận xét:

- Biểu đồ xác nhận một cách trực quan rằng số lượng mẫu của lớp `0` (Không bị tiểu đường) cao hơn đáng kể so với lớp `1` (Bị tiểu đường).
- **Ý nghĩa:** Sự mất cân bằng này là một thách thức lớn. Để xây dựng một mô hình công bằng và hiệu quả, cần áp dụng các kỹ thuật như tái lấy mẫu (oversampling lớp thiểu số bằng SMOTE) hoặc sử dụng các hàm mất mát có trọng số (weighted loss functions) trong quá trình huấn luyện mô hình LightGBM.

IV.4. Kết quả tổng hợp từ EDA

Quá trình Phân tích Dữ liệu Khám phá (EDA) đã cung cấp một cái nhìn sâu sắc và toàn diện về bộ dữ liệu Pima Indians Diabetes. Các kết quả không chỉ xác nhận lại những giả thuyết y khoa mà còn định lượng rõ ràng các mối quan hệ và các vấn đề tiềm ẩn trong dữ liệu. Những phát hiện này là nền tảng cốt lõi, định hướng cho các bước tiền xử lý và xây dựng mô hình học máy sau này.

IV.4.1. Chất lượng dữ liệu

Phân tích ban đầu đã phát hiện hai vấn đề nghiêm trọng về chất lượng dữ liệu, đòi hỏi sự chú ý đặc biệt trong giai đoạn tiền xử lý để đảm bảo tính chính xác và độ tin cậy của mô hình.

- **Giá trị thiếu ẩn được mã hóa bằng 0:** Một vấn đề chất lượng dữ liệu nghiêm trọng đã được phát hiện là sự hiện diện của các giá trị 0 trong các cột mà về mặt y sinh không thể có giá trị này ở người sống, bao gồm `Glucose`, `BloodPressure`, `SkinThickness`, `Insulin`, và `BMI`. Đây thực chất là các giá trị thiếu được mã hóa ẩn. Tỷ lệ thiếu này là đặc biệt cao ở một số biến quan trọng:
    - **Insulin:** 374 mẫu (chiếm 48.7%).
    - **SkinThickness:** 227 mẫu (chiếm 29.6%).

      Việc gần một nửa số bệnh nhân thiếu dữ liệu về Insulin cho thấy việc loại bỏ các hàng này sẽ làm mất đi một lượng lớn thông tin và không phải là giải pháp khả thi.
- **Mất cân bằng lớp trong biến mục tiêu:** Dữ liệu thể hiện sự mất cân bằng lớp (class imbalance) rõ rệt. Trong tổng số 768 mẫu, có 500 mẫu thuộc lớp **0 (Không mắc bệnh)**, chiếm 65.1%, trong khi chỉ có 268 mẫu thuộc lớp **1 (Mắc bệnh)**, chiếm 34.9%. Sự chênh lệch này có thể khiến các thuật toán học máy có xu hướng "thiên vị" về phía lớp đa số, dẫn đến hiệu suất dự đoán kém trên lớp thiểu số, vốn là lớp quan trọng hơn trong bối cảnh chẩn đoán y khoa.

IV.4.2. Phân bố dữ liệu (Univariate Analysis)

Phân tích đơn biến, thông qua histogram và các chỉ số thống kê, cho thấy sự đa dạng trong phân bố của các đặc trưng, cung cấp manh mối về bản chất của từng biến số.

- **Phân bố gần chuẩn:** Các biến `Glucose`, `BloodPressure`, và `BMI` có dạng phân bố gần giống hình chuông (phân bố chuẩn). Điều này cho thấy các giá trị của chúng tập trung quanh giá trị trung bình và ít có sự xuất hiện của các giá trị cực đoan. Ví dụ, `Glucose` có đỉnh phân bố quanh mức 120 mg/dL, phù hợp với ngưỡng chẩn đoán của WHO.
- **Phân bố lệch phải:** Các biến `Insulin`, `Age`, `Pregnancies`, `SkinThickness`, và `DiabetesPedigreeFunction` đều cho thấy sự lệch phải rõ rệt. Điều này có nghĩa là phần lớn các giá trị tập trung ở mức thấp, nhưng có một "đuôi" dài các giá trị rất cao.
    - Đặc biệt, `Insulin` có độ lệch lớn nhất, với độ lệch chuẩn (85.02) cao và giá trị tối đa lên tới 846 μU/mL, xác nhận sự tồn tại của các giá trị ngoại lệ rất cao.

- Sự phân bố lệch này rất quan trọng, vì các mô hình nhạy cảm với giả định về phân bố chuẩn (ví dụ: hồi quy tuyến tính) có thể yêu cầu các phép biến đổi (như log-transform) để cải thiện hiệu suất.

IV.4.3. So sánh theo Outcome (Bivariate Analysis)

Khi so sánh phân bố của các đặc trưng giữa hai nhóm (mắc bệnh và không mắc bệnh), phân tích song biến đã làm nổi bật sự khác biệt rõ ràng, có ý nghĩa thống kê và phù hợp với kiến thức y khoa.

- Các biểu đồ hộp (boxplots) cho thấy giá trị trung vị (median) của

  `Glucose`, `BMI`, và `Age` ở nhóm mắc bệnh (`Outcome=1`) cao hơn một cách đáng kể so với nhóm không mắc bệnh (`Outcome=0`).

  - **Glucose:** Trung vị của nhóm mắc bệnh là khoảng 140 mg/dL, so với khoảng 110 mg/dL ở nhóm không bệnh. Mức 140 mg/dL là ngưỡng quan trọng trong các tiêu chí chẩn đoán của cả NDDG (1979) và WHO (1999), khẳng định đây là một chỉ báo mạnh mẽ.

  - **BMI:** Trung vị của nhóm mắc bệnh là khoảng 35 kg/m², so với khoảng 30 kg/m² ở nhóm không bệnh, cho thấy mối liên hệ mật thiết giữa béo phì và tiểu đường type 2, một yếu tố nguy cơ được nhấn mạnh trong các tài liệu y khoa.

  - **Age:** Trung vị của nhóm mắc bệnh là khoảng 36 tuổi, cao hơn so với 29 tuổi ở nhóm không bệnh, phản ánh nguy cơ mắc bệnh tăng theo tuổi tác.

- Mặc dù có sự khác biệt rõ ràng, vẫn có một vùng chồng lấn đáng kể giữa hai nhóm ở tất cả các biến. Điều này giải thích tại sao không thể dựa vào một ngưỡng duy nhất của một biến để chẩn đoán chính xác, và sự cần thiết của các mô hình đa biến phức tạp hơn để phân loại.

IV.4.4. Phân tích đa biến (Multivariate Analysis)

Phân tích đa biến, chủ yếu thông qua ma trận tương quan (heatmap) và biểu đồ phân tán (scatterplot), đã định lượng mối quan hệ giữa các biến với nhau và với biến mục tiêu `Outcome`.

- **Tương quan với Outcome:**

  - `Glucose` có tương quan dương mạnh nhất với `Outcome` ($r \approx 0.49$), xác nhận đây là yếu tố dự báo quan trọng hàng đầu.

  - `BMI` ($r \approx 0.31$) và `Age` ($r \approx 0.24$) cũng có tương quan dương ở mức độ vừa phải.

  - Các yếu tố này cũng là những biến đầu vào được sử dụng trong nghiên cứu tiên phong của Smith et al. (1988) trên cùng bộ dữ liệu này, cho thấy tầm quan trọng lịch sử của chúng.

- **Đa cộng tuyến (Multicollinearity):**
    - Phát hiện mối tương quan đáng chú ý giữa một số cặp biến độc lập, chẳng hạn như

      `Age` và `Pregnancies` ($r \approx 0.54$), điều này là hợp lý về mặt logic.
    - Ngoài ra, các biến liên quan đến mỡ cơ thể như `BMI` và `SkinThickness` cũng có tương quan với nhau ($r \approx 0.54$).
    - Mặc dù không có tương quan nào quá cao (> 0.8) để gây ra vấn đề đa cộng tuyến nghiêm trọng, việc nhận biết các mối quan hệ này là cần thiết khi lựa chọn và diễn giải các mô hình tuyến tính.

IV.4.5. Tổng hợp

Tóm lại, quá trình Phân tích Dữ liệu Khám phá đã hoàn thành xuất sắc các mục tiêu đề ra. Phân tích đã **xác định và định lượng thành công các yếu tố nguy cơ chính** của bệnh tiểu đường type 2, bao gồm nồng độ Glucose, chỉ số BMI và tuổi tác, hoàn toàn phù hợp với các tài liệu y khoa nền tảng từ WHO và NDDG. Quan trọng hơn, EDA đã **phát hiện các vấn đề nghiêm trọng về chất lượng dữ liệu**, bao gồm giá trị thiếu ẩn và mất cân bằng lớp, là những thách thức thực tế phải được xử lý cẩn thận. Cuối cùng, các kết quả này đã **cung cấp những định hướng chiến lược và rõ ràng** cho giai đoạn tiền xử lý dữ liệu (imputation, xử lý ngoại lệ, cân bằng lớp) và là cơ sở vững chắc cho việc lựa chọn đặc trưng và xây dựng các mô hình học máy hiệu quả trong tương lai.

V. Thảo luận kết quả

Phần này sẽ diễn giải sâu hơn về các kết quả thu được từ Phân tích Dữ liệu Khám phá (EDA), đặt chúng trong bối cảnh y khoa và thực tiễn, đồng thời thảo luận về những hạn chế của bộ dữ liệu và định hướng ứng dụng trong lĩnh vực học máy.

V.1 Ý nghĩa y học và thực tiễn

V.1.1 Ý nghĩa y học

Kết quả phân tích dữ liệu đã xác nhận và định lượng một cách rõ ràng các yếu tố nguy cơ cốt lõi của bệnh tiểu đường type 2, vốn đã được ghi nhận trong các tài liệu y khoa nền tảng của WHO và NDDG.

- **Glucose là yếu tố dự báo then chốt:** Phân tích cho thấy nồng độ glucose huyết tương sau 2 giờ (OGTT) có tương quan mạnh nhất với chẩn đoán bệnh tiểu đường ($r \approx 0.49$). Điều này hoàn toàn phù hợp với các tiêu chí chẩn đoán được thiết lập bởi WHO (1999) và NDDG (1979), trong đó ngưỡng glucose sau nghiệm pháp OGTT là một chỉ số quyết định. Kết quả này khẳng định rằng suy giảm khả năng chuyển hóa glucose là dấu hiệu sinh học trung tâm của bệnh.
- **Vai trò của Béo phì và Tuổi tác:** Chỉ số BMI và Tuổi tác (Age) là hai yếu tố có tương quan cao tiếp theo. Điều này củng cố kiến thức y khoa rằng béo phì là

nguyên nhân hàng đầu gây ra kháng insulin , và nguy cơ mắc bệnh tăng lên theo tuổi tác do sự suy giảm chức năng của tế bào beta và tích lũy các yếu tố nguy cơ khác. Các phát hiện này đồng nhất với mô tả về Hội chứng Chuyển hóa của WHO, trong đó béo phì và tăng đường huyết là các thành phần chính.

- **Ảnh hưởng của Di truyền và Thai kỳ:** Biến `iabetesPedigreeFunction` và `Pregnancies` cũng cho thấy mối liên hệ có ý nghĩa. Điều này nhấn mạnh tầm quan trọng của yếu tố di truyền và những thay đổi chuyển hóa trong thai kỳ (tiểu đường thai kỳ) như những yếu tố nguy cơ tiềm tàng, một điểm đã được cả NDDG và WHO đề cập.

V.1.2 Ý nghĩa thực tiễn

Từ góc độ thực tiễn, các kết quả này mang lại những giá trị ứng dụng quan trọng trong y tế cộng đồng và lâm sàng.

- **Nền tảng cho các công cụ sàng lọc sớm:** Việc xác định được các chỉ số đơn giản nhưng có khả năng dự báo cao như Glucose, BMI, và Tuổi tác mở ra cơ hội xây dựng các hệ thống sàng lọc sớm, chi phí thấp. Một mô hình dự báo dựa trên các biến này có thể giúp xác định các cá nhân có nguy cơ cao để tư vấn can thiệp lối sống hoặc thực hiện các xét nghiệm chẩn đoán chuyên sâu hơn, giảm gánh nặng cho hệ thống y tế.

- **Định hướng cho chiến lược y tế công cộng:** Kết quả nhấn mạnh tầm quan trọng của việc kiểm soát cân nặng (BMI) trong phòng ngừa tiểu đường. Điều này cung cấp bằng chứng vững chắc để các nhà hoạch định chính sách y tế đẩy mạnh các chương trình giáo dục dinh dưỡng và khuyến khích hoạt động thể chất trong cộng đồng.

- **Hỗ trợ ra quyết định lâm sàng:** Các bác sĩ có thể sử dụng các yếu tố nguy cơ này để đánh giá tổng thể và cá nhân hóa kế hoạch chăm sóc cho bệnh nhân. Ví dụ, một phụ nữ có tiền sử mang thai nhiều lần, lớn tuổi và có chỉ số BMI cao sẽ cần được theo dõi chặt chẽ hơn.

V.2 Hạn chế của dữ liệu

Mặc dù mang lại nhiều hiểu biết giá trị, bộ dữ liệu Pima Indians Diabetes cũng tồn tại nhiều hạn chế cố hữu cần được xem xét cẩn thận khi diễn giải kết quả và xây dựng mô hình.

V.2.1 Giới hạn về đối tượng nghiên cứu

Bộ dữ liệu chỉ bao gồm các bệnh nhân nữ, từ 21 tuổi trở lên, và thuộc một dân tộc duy nhất là người Pima bản địa ở Arizona. Người Pima có tỷ lệ mắc bệnh tiểu đường cao bất thường do các yếu tố di truyền đặc thù. Do đó, các kết quả và mô hình được xây dựng từ dữ liệu này có thể không khái quát hóa tốt cho nam giới, các nhóm tuổi khác, hoặc các chủng tộc khác.

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

Đây là hạn chế nghiêm trọng nhất. Việc các giá trị thiếu được mã hóa bằng 0 trong các biến lâm sàng quan trọng như `Insulin` (48.7% thiếu) và `SkinThickness` (29.6% thiếu) đặt ra một thách thức lớn. Bất kỳ phương pháp xử lý nào (ví dụ: thay thế bằng giá trị trung bình/trung vị) cũng chỉ là ước tính và có thể đưa vào các sai lệch không mong muốn, ảnh hưởng đến độ chính xác của phân tích và mô hình.

V.2.3 Hạn chế về số lượng biến

Bộ dữ liệu chỉ có 8 đặc trưng đầu vào. Trong khi các biến này rất quan trọng, y học hiện đại đã xác định nhiều yếu tố nguy cơ khác không có trong dữ liệu, chẳng hạn như:

- **Các chỉ số lipid máu** (cholesterol, triglycerides).
- **Các yếu tố lối sống** (chế độ ăn uống, mức độ hoạt động thể chất).
- **Tiền sử hút thuốc lá, uống rượu bia.** Sự thiếu vắng các biến này làm giảm khả năng xây dựng một bức tranh toàn diện về nguy cơ của bệnh.

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

Đây là một bộ dữ liệu cắt ngang (cross-sectional), ghi lại thông tin tại một thời điểm duy nhất. Nó không cung cấp thông tin về diễn biến của các chỉ số theo thời gian. Một nghiên cứu theo chiều dọc (longitudinal) sẽ có giá trị hơn nhiều, cho phép theo dõi sự tiến triển từ trạng thái khỏe mạnh đến tiền tiểu đường và tiểu đường, như cách NDDG đã nhấn mạnh tầm quan trọng của các nghiên cứu dài hạn.

V.2.5 Mất cân bằng lớp (class imbalance)

Tỷ lệ mẫu không mắc bệnh (65.1%) cao gần gấp đôi mẫu mắc bệnh (34.9%). Vấn đề này có thể làm cho các mô hình học máy hoạt động kém hiệu quả trong việc nhận diện các trường hợp mắc bệnh (lớp thiểu số), vốn là mục tiêu quan trọng nhất trong chẩn đoán y khoa.

V.3. Định hướng ứng dụng trong học máy

Kết quả từ EDA không chỉ giúp hiểu dữ liệu mà còn trực tiếp định hướng cho việc xây dựng và triển khai các mô hình học máy.

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

Mục tiêu chính là xây dựng một mô hình phân loại nhị phân để dự đoán biến `Outcome` (0 hoặc 1) dựa trên 8 đặc trưng đầu vào. Các thuật toán phổ biến như Logistic Regression, Support Vector Machines (SVM), Random Forest, và các mô hình tăng cường độ dốc (Gradient Boosting) như LightGBM là những lựa chọn phù hợp.

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

EDA đã xác định `Glucose`, `BMI`, và `Age` là các yếu tố có khả năng dự báo cao nhất. Thông tin này rất hữu ích cho bước lựa chọn đặc trưng (feature selection). Nó cho phép tập trung vào các biến quan trọng nhất, có thể giúp đơn giản hóa mô hình, giảm thời gian huấn luyện và cải thiện khả năng diễn giải mà không làm giảm đáng kể hiệu suất.

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

Để giải quyết vấn đề mất cân bằng lớp, các kỹ thuật chuyên biệt cần được áp dụng. Có thể sử dụng các phương pháp tái lấy mẫu như **SMOTE (Synthetic Minority Over-sampling Technique)** để tạo thêm các mẫu tổng hợp cho lớp thiểu số, hoặc điều chỉnh trọng số của các lớp trong hàm mất mát (loss function) của mô hình để "trừng phạt" nặng hơn các lỗi dự đoán sai trên lớp thiểu số.

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

Mục tiêu cuối cùng không chỉ là một mô hình phân loại "có/không", mà là một hệ thống có thể cung cấp một **điểm số nguy cơ (risk score)**. Hệ thống này có thể hoạt động như một công cụ hỗ trợ quyết định cho các bác sĩ, giúp họ nhanh chóng xác định những bệnh nhân cần được quan tâm đặc biệt, tương tự như mục tiêu của nghiên cứu dùng thuật toán ADAP năm 1988.

V.3.5 Định hướng mở rộng

Trong tương lai, mô hình có thể được cải thiện bằng cách thu thập thêm dữ liệu từ các quần thể đa dạng hơn và bổ sung các biến số quan trọng khác (lipid máu, lối sống). Việc tích hợp các kỹ thuật học sâu (Deep Learning) cũng có thể giúp phát hiện các mối quan hệ phi tuyến tính phức tạp hơn trong dữ liệu.

VI. Kết luận

VI.1. Kết quả chính

Nghiên cứu đã thực hiện một phân tích dữ liệu khám phá toàn diện trên bộ dữ liệu Pima Indians Diabetes. Các kết quả chính đã xác định **nồng độ Glucose, chỉ số BMI, và Tuổi tác** là ba yếu tố dự báo quan trọng nhất cho bệnh tiểu đường type 2. Bên cạnh đó, nghiên cứu cũng đã phát hiện các vấn đề nghiêm trọng về chất lượng dữ liệu, bao gồm tỷ lệ cao các **giá trị thiếu ẩn** và sự **mất cân bằng lớp** rõ rệt trong biến mục tiêu.

VI.2 Ý nghĩa y học và thực tiễn

Các phát hiện từ EDA hoàn toàn phù hợp với các kiến thức y khoa đã được thiết lập, khẳng định vai trò trung tâm của tình trạng tăng đường huyết và béo phì trong sinh bệnh học của tiểu đường type 2. Về mặt thực tiễn, kết quả này cung cấp một nền tảng dựa trên bằng chứng để phát triển các công cụ sàng lọc sớm, chi phí thấp và định

hướng các chiến lược y tế công cộng tập trung vào việc kiểm soát cân nặng và theo dõi các chỉ số chuyển hóa.

VI.3 Hạn chế của dữ liệu

Nghiên cứu bị giới hạn bởi các đặc điểm của bộ dữ liệu, bao gồm tính không đại diện của quần thể nghiên cứu (chỉ phụ nữ Pima), vấn đề chất lượng dữ liệu (giá trị 0 bất hợp lý), thiếu các biến số lối sống quan trọng và bản chất cắt ngang của dữ liệu. Những hạn chế này cần được xem xét khi áp dụng các kết quả vào thực tế.

VI.4 Định hướng nghiên cứu và ứng dụng

Các kết quả phân tích cung cấp một định hướng rõ ràng cho các bước tiếp theo trong lĩnh vực học máy. Cần tập trung vào việc xây dựng các mô hình phân loại nhị phân, áp dụng các kỹ thuật xử lý dữ liệu mất cân bằng và lựa chọn đặc trưng dựa trên các yếu tố quan trọng đã được xác định. Mục tiêu dài hạn là phát triển một hệ thống hỗ trợ quyết định lâm sàng có khả năng dự báo nguy cơ, giúp cá nhân hóa việc chăm sóc sức khỏe.

VI.5 Kết luận cuối cùng

Báo cáo này đã minh họa thành công sức mạnh của Phân tích Dữ liệu Khám phá trong việc làm sáng tỏ các mối quan hệ phức tạp trong dữ liệu y tế. Bằng cách kết nối các phát hiện thống kê với kiến thức y khoa, chúng tôi không chỉ hiểu sâu hơn về các yếu tố nguy cơ của bệnh tiểu đường mà còn đặt ra một nền tảng vững chắc và các định hướng chiến lược cho việc xây dựng các mô hình dự báo chính xác và có ý nghĩa trong tương lai. Đây là một bước quan trọng trong nỗ lực ứng dụng khoa học dữ liệu để giải quyết những thách thức sức khỏe cộng đồng.

DANH MỤC TÀI LIỆU THAM KHẢO

1. **National Diabetes Data Group.** (1979). Classification and Diagnosis of Diabetes Mellitus and Other Categories of Glucose Intolerance. *Diabetes*, 28(12), 1039–1057.

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. *Proceedings of the Annual Symposium on Computer Application in Medical Care*, 261–265.

3. **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*. (No. WHO/NCD/NCS/99.2). World Health Organization.