**TRƯỜNG ĐẠI HỌC SÀI GÒN**

**KHOA CÔNG NGHỆ THÔNG TIN**

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

A blue button with white text

AI-generated content may be incorrect.

BÁO CÁO MÔN

PHÂN TÍCH DỮ LIỆU

**TÊN ĐỀ TÀI Phân tích khám phá về bệnh đái tháo đường**

|  |  |
| --- | --- |
| **Giảng viên hướng dẫn:** | **ThS. Đỗ Như Tài** |
| **Nhóm sinh viên thực hiện:** |  |
| Hồ Thị Thanh Thảo | 3122410389 |
| Nguyễn Thị Hồng Thắm | 3122410392 |
| Phan Văn Thảo | 3122410391 |
| Nguyễn Hoàng Thiên Bảo | 3122410019 |

***Thành phố Hồ Chí Minh, tháng 09 năm 2025***

**LỜI MỞ ĐẦU**

Trong bối cảnh hiện nay, bệnh đái tháo đường (tiểu đường) đang trở thành một trong những thách thức y tế lớn trên toàn cầu. Việc phát hiện sớm nguy cơ mắc bệnh có ý nghĩa quan trọng trong phòng ngừa và kiểm soát, giúp giảm thiểu biến chứng và chi phí điều trị. Với sự phát triển mạnh mẽ của khoa học dữ liệu, đặc biệt là trong lĩnh vực phân tích dữ liệu y tế, chúng ta có thể khai thác các tập dữ liệu thực tế để đưa ra những nhận định, dự báo hỗ trợ quá trình chẩn đoán và chăm sóc sức khỏe.

Báo cáo này tập trung vào phân tích dữ liệu chẩn đoán bệnh tiểu đường dựa trên bộ dữ liệu thực nghiệm. Thông qua các bước xử lý, thống kê mô tả, trực quan hóa và xây dựng mô hình phân tích, mục tiêu là:

* Hiểu rõ đặc điểm phân bố của các chỉ số lâm sàng quan trọng.
* Xác định mối quan hệ giữa các đặc trưng (như Glucose, BMI, Huyết áp, Tuổi, v.v) với khả năng mắc bệnh.
* Đề xuất những gợi ý ban đầu về yếu tố nguy cơ chính ảnh hưởng đến bệnh tiểu đường.

Kết quả của nghiên cứu không chỉ giúp củng cố kiến thức về phân tích dữ liệu y tế mà còn minh họa ứng dụng thực tiễn của các phương pháp khoa học dữ liệu trong lĩnh vực chăm sóc sức khỏe cộng đồng.

**LỜI CẢM ƠN**

Chúng em xin gửi lời cảm ơn chân thành đến *Ths. Đỗ Như Tài*, người đã tận tình hướng dẫn và chia sẻ những kiến thức quý báu trong suốt quá trình thực hiện đề tài này. Sự chỉ bảo tận tình và những góp ý xây dựng từ thầy đã giúp chúng em hiểu rõ hơn về các vấn đề chuyên môn và hoàn thiện sản phẩm một cách tốt nhất.

Mặc dù đã rất cố gắng, nhưng do hạn chế về thời gian và kinh nghiệm, chắc chắn rằng không thể tránh khỏi những sai sót. Chúng em rất mong nhận được ý kiến đóng góp từ thầy để có thể cải thiện và hoàn thiện hơn nữa.

Một lần nữa, chúng em xin chân thành cảm ơn!

**MỤC LỤC**

[CHƯƠNG 1: GIỚI THIỆU 1](#_Toc209728878)

[1.1. Tổng quan về bài toán 1](#_Toc209728879)

[1.2. Mục tiêu nghiên cứu 1](#_Toc209728880)

[1.3. Ý nghĩa thực tiễn 3](#_Toc209728881)

[1.4. Phạm vi và giới hạn nghiên cứu 4](#_Toc209728882)

[CHƯƠNG 2: CƠ SỞ LÝ THUYẾT VÀ NGHIÊN CỨU LIÊN QUAN 6](#_Toc209728883)

[2.1. Tổng quan về bệnh đái tháo đường 6](#_Toc209728884)

[2.1.1. Định nghĩa và phân loại 6](#_Toc209728885)

[2.1.2. Tiêu chuẩn chẩn đoán 7](#_Toc209728886)

[2.1.3. Các biến chứng và yếu tố nguy cơ 7](#_Toc209728887)

[2.2. Các nghiên cứu liên quan về dự đoán bệnh đái tháo đường 9](#_Toc209728888)

[2.2.1. Phương pháp machine learning trong y tế 9](#_Toc209728889)

[2.2.2. Thuật toán ADAP và ứng dụng 10](#_Toc209728890)

[2.2.3. Các nghiên cứu khác về bộ dữ liệu Pima Indians 11](#_Toc209728891)

[2.3. Xác định bài toán nghiên cứu 11](#_Toc209728892)

[2.3.1. Input 11](#_Toc209728893)

[2.3.2. Output 12](#_Toc209728894)

[2.3.3. Mục tiêu 12](#_Toc209728895)

[CHƯƠNG 3: TỔNG QUAN DỮ LIỆU 13](#_Toc209728896)

[3.1. Mô tả dataset Pima Indians Diabetes 13](#_Toc209728897)

[3.2. Mô tả chi tiết các biến số 13](#_Toc209728898)

[3.2.1. Pregnancies (Số lần có thai) 13](#_Toc209728899)

[3.2.2. Glucose (Nồng độ glucose) 13](#_Toc209728900)

[3.2.3. BloodPressure (Huyết áp) 13](#_Toc209728901)

[3.2.4. SkinThickness (Độ dày da) 14](#_Toc209728902)

[3.2.5. Insulin (Nồng đô insulin) 14](#_Toc209728903)

[3.2.6. BMI (Chỉ số khối cơ thể) 14](#_Toc209728904)

[3.2.7. DiabetesPedigreeFunction (Yếu tố di truyền) 14](#_Toc209728905)

[3.2.8. Age (Tuổi) 15](#_Toc209728906)

[3.2.9. Outcome (Kết quả chấn đoán) 15](#_Toc209728907)

[3.3. Tiền xử lý dữ liệu 15](#_Toc209728908)

[3.3.1. Kiểm tra và loại bỏ dữ liệu trùng lặp 15](#_Toc209728909)

[3.3.2. Kiểm tra và loại bỏ dữ liệu lỗi, ngoại lệ 16](#_Toc209728910)

[3.3.3. Xử lý giá trị thiếu và giá trị bằng 0 17](#_Toc209728911)

[3.3.4. Kết quả sau tiền xử lý 18](#_Toc209728912)

[3.4. Thống kê mô tả cơ bản 18](#_Toc209728913)

[CHƯƠNG 4: PHÂN TÍCH KHÁM PHÁ DỮ LIỆU 20](#_Toc209728914)

[4.1. Phân tích đơn biến 20](#_Toc209728915)

[4.1.1. Phân bố của từng biến số 20](#_Toc209728916)

[4.1.2. Phát hiện ngoại lệ và bất thường 22](#_Toc209728917)

[4.2. Phân tích biến mục tiêu 23](#_Toc209728918)

[4.2.1. Tỷ lệ mắc bệnh đái tháo đường trong dataset 23](#_Toc209728919)

[4.2.2. Đặc điểm nhóm bệnh nhân và nhóm khỏe mạnh 24](#_Toc209728920)

[4.3. Phân tích theo nhóm tuổi 26](#_Toc209728921)

[4.3.1. Phân bố bệnh theo độ tuổi 26](#_Toc209728922)

[4.3.2. Các chỉ số sinh lý thay đổi theo tuổi 27](#_Toc209728923)

[4.3.3. Yếu tố nguy cơ ở từng nhóm tuổi 28](#_Toc209728924)

[4.4. Phân tích theo chỉ số BMI 29](#_Toc209728925)

[4.4.1. Phân loại BMI và tỷ lệ mắc bệnh 29](#_Toc209728926)

[4.4.2. Mối quan hệ giữa béo phì và đái tháo đường 30](#_Toc209728927)

[4.4.3. BMI kết hợp với các yếu tố khác 31](#_Toc209728928)

[4.5. Phân tích Glucose và Insulin 34](#_Toc209728929)

[4.5.1. Phân bố bệnh theo nồng độ Glucose 34](#_Toc209728930)

[4.5.2. Mối quan hệ Glucose và Insulin 35](#_Toc209728931)

[4.6. Phân tích yếu tố di truyền 36](#_Toc209728932)

[4.6.1. Hàm phả hệ đái tháo đường và ý nghĩa 36](#_Toc209728933)

[4.6.2. Tác động của yếu tố di truyền 37](#_Toc209728934)

[4.6.3. Kết hợp yếu tố di truyền với Glucose 38](#_Toc209728935)

[4.7. Phân tích theo số lần có thai 39](#_Toc209728936)

[CHƯƠNG 5: PHÂN TÍCH TƯƠNG QUAN VÀ MỐI QUAN HỆ 41](#_Toc209728937)

[5.1. Ma trận tương quan 41](#_Toc209728938)

[5.2. Phân tích cặp biến quan trọng 42](#_Toc209728939)

[5.2.1. Glucose - Kết quả bệnh 42](#_Toc209728940)

[5.2.2. BMI - Độ dày da 43](#_Toc209728941)

[5.2.3. Tuổi - Số lần có thai 44](#_Toc209728942)

[5.2.4. Glucose - Insulin 45](#_Toc209728943)

[5.3. Phân tích tầm quan trọng của đặc trưng 45](#_Toc209728944)

[5.3.1. Tầm quan trọng đơn biến 45](#_Toc209728945)

[5.3.2. Điểm thông tin tương hỗ 46](#_Toc209728946)

[5.3.3. Tầm quan trọng từ Random Forest 47](#_Toc209728947)

[5.4. Phát hiện đa cộng tuyến 48](#_Toc209728948)

[CHƯƠNG 6: PHÂN TÍCH NÂNG CAO 49](#_Toc209728949)

[6.1. Phân tích phân nhóm (Phân cụm) 49](#_Toc209728950)

[6.1.1. Xác định các nhóm bệnh nhân tiềm ẩn 49](#_Toc209728951)

[6.1.2. Đặc điểm của từng cụm 51](#_Toc209728952)

[6.2. Phân tích các yếu tố nguy cơ 52](#_Toc209728953)

[6.2.1. Xếp hạng các yếu tố nguy cơ 52](#_Toc209728954)

[6.2.2 Tạo hệ thống chấm điểm nguy cơ 53](#_Toc209728955)

[6.2.3. Phân tích ngưỡng cho sàng lọc 54](#_Toc209728956)

[6.3. So sánh với tiêu chí chẩn đoán 54](#_Toc209728957)

[6.3.1 Tiêu chí chẩn đoán của WHO/ADA 54](#_Toc209728958)

[6.3.2 Độ chính xác của các giá trị cắt 55](#_Toc209728959)

[CHƯƠNG 7: THẢO LUẬN VÀ HIỂU BIẾT SÂU SẮC 58](#_Toc209728960)

[7.1. Các phát hiện chính 58](#_Toc209728961)

[7.1.1. Yếu tố nguy cơ quan trọng nhất 58](#_Toc209728962)

[7.1.2. Các mẫu và xu hướng 58](#_Toc209728963)

[7.1.3. Những phát hiện bất ngờ 58](#_Toc209728964)

[7.2. So sánh với tài liệu nghiên cứu 59](#_Toc209728965)

[7.2.1. Nhất quán với nghiên cứu trước 59](#_Toc209728966)

[7.2.2. Phát hiện mới 59](#_Toc209728967)

[7.2.3. Ý nghĩa lâm sàng 59](#_Toc209728968)

[7.3. Hạn chế và xu hướng 60](#_Toc209728969)

[7.3.1. Hạn chế của bộ dữ liệu 60](#_Toc209728970)

[7.3.2. Xu hướng chọn mẫu và khả năng tổng quát hoá 60](#_Toc209728971)

[7.3.3. Tác động của dữ liệu thiếu 60](#_Toc209728972)

[CHƯƠNG 8: KẾT LUẬN VÀ KHUYẾN NGHỊ 61](#_Toc209728973)

[8.1. Tóm tắt phát hiện 61](#_Toc209728974)

[8.1.1. Các yếu tố nguy cơ chính 61](#_Toc209728975)

[8.1.2. Ngưỡng lâm sàng được đề xuất 61](#_Toc209728976)

[8.1.3. Hiểu biết đặc thù về dân số nghiên cứu 61](#_Toc209728977)

[8.2. Ứng dụng thực tiễn 61](#_Toc209728978)

[8.2.1. Quy trình sàng lọc 61](#_Toc209728979)

[8.2.2. Chiến lược phòng ngừa 62](#_Toc209728980)

[8.2.3. Hỗ trợ quyết định lâm sàn 62](#_Toc209728981)

[8.3. Hướng nghiên cứu tiếp theo 62](#_Toc209728982)

[8.3.1. Mở rộng bộ dữ liệu 62](#_Toc209728983)

[8.3.2. Các phương pháp mô hình hóa nâng cao 63](#_Toc209728984)

[8.3.3. Nghiên cứu theo chiều dọc 63](#_Toc209728985)

[TÀI LIỆU THAM KHẢO 64](#_Toc209728986)

CHƯƠNG 1: GIỚI THIỆU

## 1.1. Tổng quan về bài toán

Bệnh đái tháo đường (diabetes mellitus) là một rối loạn chuyển hóa mãn tính đặc trưng bởi tình trạng tăng đường huyết kéo dài, bắt nguồn từ sự thiếu hụt insulin tuyệt đối hoặc tương đối, hoặc do sự đề kháng insulin ở mô ngoại vi. Theo thống kê của Tổ chức Y tế Thế giới (WHO), số người mắc đái tháo đường đã vượt mốc 500 triệu trên toàn cầu, với tỷ lệ gia tăng đáng báo động ở các quốc gia đang phát triển. Điều này không chỉ đặt ra thách thức lớn về mặt y tế công cộng mà còn gây áp lực nặng nề lên hệ thống chăm sóc sức khỏe và ngân sách quốc gia.

Việc phát hiện sớm và dự đoán nguy cơ mắc bệnh đái tháo đường có vai trò then chốt trong chiến lược phòng ngừa và kiểm soát bệnh. Các yếu tố nguy cơ như tuổi tác, chỉ số khối cơ thể (BMI), huyết áp, tiền sử gia đình, và các chỉ số sinh hóa khác đã được chứng minh có liên quan mật thiết đến khả năng phát triển bệnh. Tuy nhiên, mối quan hệ giữa các yếu tố này thường phức tạp và phi tuyến tính, đòi hỏi các phương pháp phân tích dữ liệu hiện đại để khám phá sâu hơn.

Trong nghiên cứu này, bộ dữ liệu Pima Indians Diabetes từ UCI Machine Learning Repository được sử dụng như một trường hợp điển hình. Đây là tập dữ liệu kinh điển trong lĩnh vực học máy y sinh, bao gồm thông tin lâm sàng của phụ nữ người Mỹ bản địa Pima, với các biến số như số lần mang thai, nồng độ glucose huyết tương, huyết áp, độ dày lớp da, insulin, BMI, chỉ số di truyền và tuổi. Việc lựa chọn bộ dữ liệu này không chỉ giúp chuẩn hóa quy trình phân tích mà còn tạo điều kiện thuận lợi cho việc so sánh với các nghiên cứu trước đó.

## 1.2. Mục tiêu nghiên cứu

Trong bối cảnh bệnh đái tháo đường ngày càng phổ biến và có xu hướng trẻ hóa, việc ứng dụng phân tích dữ liệu để hỗ trợ sàng lọc và chẩn đoán sớm trở nên cấp thiết. Nghiên cứu này đặt ra các mục tiêu cụ thể nhằm khai thác tối đa giá trị của bộ dữ liệu Pima Indians Diabetes, đồng thời đóng góp vào hướng tiếp cận dữ liệu lớn trong y học dự phòng.

**Mục tiêu chính:**

Xác định các yếu tố nguy cơ quan trọng nhất dẫn đến bệnh đái tháo đường: Thông qua các kỹ thuật phân tích thống kê và học máy như phân tích tương quan, hồi quy logistic, và lựa chọn đặc trưng (feature selection), nghiên cứu sẽ làm rõ vai trò của từng biến số sinh lý trong việc dự đoán nguy cơ mắc bệnh. Các bài báo đã gợi ý rằng glucose huyết tương, BMI, và tuổi là những yếu tố nổi bật, nhưng cần xác nhận lại bằng phân tích định lượng.

Phân tích mối quan hệ giữa các biến số sinh lý và khả năng mắc bệnh: Việc sử dụng trực quan hóa dữ liệu (như biểu đồ phân tán, heatmap, boxplot) kết hợp với mô hình hóa sẽ giúp làm rõ các mối quan hệ tuyến tính và phi tuyến giữa các biến số. Một số bài báo đã chỉ ra rằng sự kết hợp giữa nhiều yếu tố (ví dụ: glucose + BMI + insulin) có thể tạo ra hiệu ứng cộng hưởng trong việc dự đoán bệnh.

Đánh giá tính khả thi của việc sử dụng các chỉ số sinh lý để sàng lọc bệnh: Nghiên cứu sẽ kiểm tra độ chính xác, độ nhạy và độ đặc hiệu của các mô hình dựa trên chỉ số sinh lý. Nếu đạt được hiệu suất cao, điều này mở ra khả năng ứng dụng thực tế trong cộng đồng, đặc biệt là tại các cơ sở y tế tuyến đầu.

**Mục tiêu phụ:**

So sánh kết quả với các tiêu chuẩn chẩn đoán quốc tế: Kết quả phân tích sẽ được đối chiếu với các tiêu chuẩn như của WHO, ADA (American Diabetes Association), nhằm đánh giá mức độ phù hợp và khả năng ứng dụng của mô hình trong thực tế lâm sàng.

Đưa ra khuyến nghị cho việc phát hiện sớm bệnh đái tháo đường: Dựa trên các yếu tố nguy cơ được xác định, nghiên cứu sẽ đề xuất quy trình sàng lọc đơn giản, chi phí thấp, phù hợp với điều kiện tại các quốc gia đang phát triển. Các bài báo đã gợi ý rằng chỉ cần một số chỉ số cơ bản như glucose và BMI đã có thể đạt độ chính xác đáng kể.

Xác định những hạn chế và xu hướng trong dữ liệu: Việc phân tích phân bố dữ liệu, độ thiếu hụt, độ lệch chuẩn và tính đại diện sẽ giúp nhận diện các điểm yếu trong bộ dữ liệu. Đồng thời, nghiên cứu cũng sẽ chỉ ra các xu hướng tiềm ẩn như sự thay đổi theo độ tuổi, hoặc sự khác biệt giữa các nhóm dân cư.

## 1.3. Ý nghĩa thực tiễn

Nghiên cứu này không chỉ mang giá trị học thuật trong việc phân tích dữ liệu y sinh mà còn có tiềm năng ứng dụng cao trong thực tiễn lâm sàng và y tế cộng đồng. Việc khai thác bộ dữ liệu Pima Indians Diabetes thông qua các phương pháp phân tích hiện đại giúp tạo ra những đóng góp thiết thực như sau:

Hỗ trợ bác sĩ lâm sàng trong việc đánh giá nguy cơ bệnh nhân: Các mô hình dự đoán được xây dựng từ dữ liệu sinh lý có thể đóng vai trò như một công cụ hỗ trợ ra quyết định (Clinical Decision Support System – CDSS), giúp bác sĩ nhanh chóng nhận diện bệnh nhân có nguy cơ cao. Điều này đặc biệt hữu ích trong môi trường lâm sàng có thời gian khám hạn chế hoặc thiếu các xét nghiệm chuyên sâu.

Cung cấp cơ sở khoa học cho các chương trình sàng lọc cộng đồng: Việc xác định các yếu tố nguy cơ nổi bật như glucose huyết tương, BMI, và tuổi cho phép thiết kế các chương trình sàng lọc đơn giản, chi phí thấp nhưng hiệu quả cao. Các bài báo đã chứng minh rằng chỉ cần một số chỉ số cơ bản là đã có thể đạt độ chính xác đáng kể trong phân loại nguy cơ, từ đó mở rộng khả năng tiếp cận y tế cho người dân vùng sâu vùng xa.

Góp phần vào việc phát triển các công cụ hỗ trợ quyết định lâm sàng: Kết quả nghiên cứu có thể được tích hợp vào các ứng dụng phần mềm hoặc hệ thống quản lý bệnh viện, giúp tự động hóa quy trình đánh giá nguy cơ. Một số bài báo đã đề xuất việc sử dụng thuật toán học máy như Random Forest, SVM hoặc XGBoost để tăng độ chính xác, đồng thời giảm thiểu sai sót do yếu tố chủ quan.

Tối ưu hóa việc phân bổ nguồn lực y tế cho các nhóm nguy cơ cao: Trong bối cảnh nguồn lực y tế còn hạn chế, việc xác định đúng nhóm đối tượng cần can thiệp sớm sẽ giúp nâng cao hiệu quả sử dụng ngân sách và nhân lực. Các mô hình dự đoán có thể hỗ trợ trong việc lập kế hoạch phân bổ thuốc, xét nghiệm, và tư vấn dinh dưỡng một cách hợp lý, từ đó giảm gánh nặng cho hệ thống y tế.

## 1.4. Phạm vi và giới hạn nghiên cứu

**Phạm vi:**

Nghiên cứu này tập trung vào việc phân tích dữ liệu từ 768 phụ nữ người Pima Indians, một cộng đồng dân cư bản địa tại Hoa Kỳ có tỷ lệ mắc bệnh đái tháo đường cao bất thường. Bộ dữ liệu được thu thập có tính chuẩn hóa cao, bao gồm 8 biến số sinh lý và lâm sàng như:

* Số lần mang thai
* Nồng độ glucose huyết tương
* Huyết áp tâm trương
* Độ dày lớp da
* Nồng độ insulin
* Chỉ số khối cơ thể (BMI)
* Chỉ số di truyền (Diabetes Pedigree Function)
* Tuổi

Kết quả chẩn đoán được biểu diễn dưới dạng nhị phân (1: mắc bệnh, 0: không mắc bệnh), cho phép áp dụng các mô hình phân loại trong học máy để đánh giá nguy cơ. Phân tích tập trung vào việc khám phá mối quan hệ giữa các biến số này và khả năng mắc bệnh, từ đó xây dựng các mô hình dự đoán có độ chính xác cao.

**Giới hạn:**

Chỉ áp dụng cho phụ nữ từ 21 tuổi trở lên. Dữ liệu không bao gồm nam giới hoặc nhóm tuổi dưới 21, do đó kết quả phân tích không thể khái quát cho toàn bộ dân số.

Dữ liệu từ một nhóm dân tộc cụ thể: Người Pima có đặc điểm di truyền và lối sống riêng biệt, dẫn đến nguy cơ mắc bệnh cao hơn so với các nhóm dân cư khác. Vì vậy, các mô hình xây dựng từ dữ liệu này có thể không phản ánh chính xác nguy cơ ở các cộng đồng khác.

Thiếu thông tin về lối sống, chế độ ăn, thuốc men: Bộ dữ liệu không bao gồm các yếu tố quan trọng như mức độ vận động, chế độ dinh dưỡng, tiền sử dùng thuốc hoặc các bệnh lý đi kèm. Điều này làm giảm khả năng phân tích toàn diện và có thể dẫn đến sai lệch trong việc đánh giá nguy cơ thực tế.

CHƯƠNG 2: CƠ SỞ LÝ THUYẾT VÀ NGHIÊN CỨU LIÊN QUAN

## 2.1. Tổng quan về bệnh đái tháo đường

### 2.1.1. Định nghĩa và phân loại

Đái tháo đường là một nhóm các rối loạn chuyển hóa đặc trưng bởi tình trạng tăng glucose máu mãn tính, bắt nguồn từ khiếm khuyết trong tiết insulin, tác dụng insulin, hoặc cả hai. Insulin là hormone do tế bào β của tuyến tụy tiết ra, có vai trò điều hòa lượng đường trong máu bằng cách thúc đẩy sự hấp thu glucose vào tế bào. Khi cơ thể không sản xuất đủ insulin hoặc không sử dụng insulin hiệu quả, glucose tích tụ trong máu, gây ra các biểu hiện lâm sàng và biến chứng nguy hiểm.

Phân loại chính (theo Tổ chức Y tế Thế giới WHO 1999):

* Đái tháo đường type 1: Do thiếu insulin tuyệt đối, thường là hậu quả của quá trình phá hủy tự miễn tế bào β tuyến tụy. Thường khởi phát ở tuổi trẻ, diễn tiến nhanh, cần điều trị bằng insulin suốt đời.
* Đái tháo đường type 2: Do kháng insulin ở mô ngoại vi và/hoặc giảm tiết insulin tương đối. Chiếm khoảng 90–95% tổng số ca mắc, thường liên quan đến béo phì, lối sống ít vận động, và yếu tố di truyền. Có thể kiểm soát bằng thay đổi lối sống, thuốc uống, và đôi khi cần insulin.
* Đái tháo đường thai kỳ (Gestational Diabetes): Xuất hiện trong thai kỳ, thường ở tam cá nguyệt thứ hai hoặc thứ ba. Có thể biến mất sau sinh nhưng làm tăng nguy cơ mắc đái tháo đường type 2 sau này.
* Các dạng đái tháo đường khác: MODY (Maturity Onset Diabetes of the Young): dạng di truyền hiếm gặp, khởi phát sớm nhưng không phụ thuộc insulin.

Do thuốc: ví dụ như corticosteroids, thuốc chống loạn thần.

Do bệnh lý tuyến tụy: viêm tụy mạn, cắt bỏ tụy, u tụy…

### 2.1.2. Tiêu chuẩn chẩn đoán

Việc chẩn đoán bệnh đái tháo đường cần dựa trên các chỉ số sinh hóa đáng tin cậy, phản ánh chính xác tình trạng rối loạn chuyển hóa glucose trong cơ thể. Theo khuyến cáo của Tổ chức Y tế Thế giới (WHO) và Hiệp hội Đái tháo đường Hoa Kỳ (ADA), các tiêu chuẩn chẩn đoán hiện hành bao gồm:

* Glucose máu lúc đói (Fasting Plasma Glucose - FPG) ≥ 126 mg/dL (7.0 mmol/L) sau ít nhất 8 giờ nhịn ăn.

→ Đây là xét nghiệm phổ biến nhất, dễ thực hiện tại các cơ sở y tế.

* Glucose máu sau 2 giờ nghiệm pháp dung nạp glucose (OGTT) ≥ 200 mg/dL (11.1 mmol/L) sau khi uống 75g glucose.

→ Thường dùng để phát hiện các trường hợp tiền đái tháo đường hoặc đái tháo đường thai kỳ.

* HbA1c (Hemoglobin A1c) ≥ 6.5% (48 mmol/mol).

→ Phản ánh mức glucose trung bình trong vòng 2–3 tháng, ít bị ảnh hưởng bởi thời điểm lấy mẫu.

* Glucose máu bất kỳ ≥ 200 mg/dL (11.1 mmol/L) kèm theo triệu chứng lâm sàng như tiểu nhiều, khát nước, sụt cân không rõ nguyên nhân.

→ Áp dụng trong các trường hợp cấp cứu hoặc khi không thể thực hiện xét nghiệm lúc đói.

### 2.1.3. Các biến chứng và yếu tố nguy cơ

Bệnh đái tháo đường, nếu không được kiểm soát tốt, có thể dẫn đến nhiều biến chứng nghiêm trọng, ảnh hưởng đến chất lượng sống và tuổi thọ của người bệnh. Các biến chứng được chia thành hai nhóm chính:

* Biến chứng cấp tính:

Nhiễm toan ceton do đái tháo đường (Diabetic Ketoacidosis – DKA): Xảy ra chủ yếu ở bệnh nhân type 1, do thiếu insulin trầm trọng dẫn đến tích tụ ceton trong máu, gây toan chuyển hóa. Biểu hiện gồm buồn nôn, thở nhanh, đau bụng, lú lẫn, có thể tử vong nếu không xử trí kịp thời.

Hôn mê tăng thẩm thấu (Hyperosmolar Hyperglycemic State – HHS): Thường gặp ở bệnh nhân type 2 lớn tuổi, đặc trưng bởi đường huyết rất cao, mất nước nghiêm trọng, nhưng không có nhiễm toan. Đây là tình trạng cấp cứu nội khoa cần hồi sức tích cực.

* Biến chứng mãn tính:

Bệnh thận do đái tháo đường (Diabetic Nephropathy): Là nguyên nhân hàng đầu gây suy thận mạn giai đoạn cuối. Biểu hiện bằng protein niệu, tăng huyết áp và giảm chức năng lọc cầu thận.

Bệnh võng mạc (Diabetic Retinopathy): Gây tổn thương mạch máu võng mạc, có thể dẫn đến mù lòa. Là biến chứng thường gặp nhất ở bệnh nhân đái tháo đường lâu năm.

Bệnh thần kinh (Diabetic Neuropathy): Gây tê bì, đau nhức, mất cảm giác ở chi dưới, làm tăng nguy cơ loét bàn chân và cắt cụt chi.

Bệnh tim mạch: Đái tháo đường làm tăng nguy cơ xơ vữa động mạch, nhồi máu cơ tim, đột quỵ và suy tim. Đây là nguyên nhân tử vong hàng đầu ở bệnh nhân đái tháo đường type 2.

**Yếu tố nguy cơ:**

Nhiều yếu tố nguy cơ đã được xác định là có liên quan đến sự phát triển của bệnh đái tháo đường, đặc biệt là type 2:

* Tuổi tác: Nguy cơ tăng dần theo tuổi, đặc biệt sau 40 tuổi do giảm nhạy cảm insulin và thay đổi chuyển hóa.
* Béo phì và ít vận động: Là yếu tố nguy cơ hàng đầu, đặc biệt khi mỡ tập trung vùng bụng. Lối sống ít vận động làm giảm hiệu quả sử dụng glucose ở cơ.
* Tiền sử gia đình và yếu tố di truyền: Có người thân mắc bệnh làm tăng nguy cơ gấp 2–6 lần. Một số gen liên quan đến chức năng tế bào β và kháng insulin đã được xác định.
* Tiền sử đái tháo đường thai kỳ: Phụ nữ từng mắc đái tháo đường thai kỳ có nguy cơ cao phát triển bệnh type 2 sau sinh.
* Hội chứng chuyển hóa: Bao gồm tăng huyết áp, rối loạn lipid máu, béo bụng và tăng glucose máu. Đây là tập hợp các yếu tố nguy cơ có liên quan chặt chẽ đến đái tháo đường và bệnh tim mạch.

## 2.2. Các nghiên cứu liên quan về dự đoán bệnh đái tháo đường

### 2.2.1. Phương pháp machine learning trong y tế

Trong những năm gần đây, Machine Learning (ML) đã trở thành một công cụ mạnh mẽ trong lĩnh vực y tế, đặc biệt là trong chẩn đoán, dự đoán nguy cơ bệnh, và cá nhân hóa điều trị. Việc áp dụng ML giúp khai thác hiệu quả các tập dữ liệu y sinh lớn, phát hiện các mẫu tiềm ẩn mà phương pháp thống kê truyền thống khó nhận diện.

Machine learning đã được ứng dụng rộng rãi trong chẩn đoán và dự đoán y tế:

* Supervised Learning (Học có giám sát)

Đây là phương pháp phổ biến nhất trong dự đoán bệnh lý, đặc biệt là bệnh đái tháo đường. Dữ liệu đầu vào bao gồm các đặc trưng sinh lý và kết quả chẩn đoán (nhị phân: mắc/không mắc), từ đó huấn luyện mô hình để phân loại bệnh nhân.

Classification (Phân loại): Áp dụng các thuật toán như Logistic Regression, Support Vector

Machine (SVM), Random Forest, K-Nearest Neighbors (KNN), và Gradient Boosting để dự đoán khả năng mắc bệnh. Các bài báo bạn cung cấp đều sử dụng ít nhất một trong các thuật toán này để đánh giá hiệu suất mô hình.

Regression (Hồi quy): Dùng để dự đoán các chỉ số liên tục như mức glucose, HbA1c hoặc nguy cơ tương đối. Hữu ích trong việc lượng hóa mức độ nguy cơ thay vì chỉ phân loại.

* Unsupervised Learning (Học không giám sát)

Phương pháp này không sử dụng nhãn đầu ra, mà tập trung vào việc khám phá cấu trúc dữ liệu:

Clustering (Phân cụm): Các thuật toán như K-Means, Hierarchical Clustering, hoặc DBSCAN được dùng để phân nhóm bệnh nhân theo đặc điểm sinh lý. Điều này giúp nhận diện các nhóm nguy cơ cao, nhóm bệnh nhân có đặc điểm tương đồng hoặc phát hiện các kiểu bệnh lý chưa được phân loại rõ ràng.

Dimensionality Reduction (Giảm chiều dữ liệu): Các kỹ thuật như PCA hoặc t-SNE giúp trực quan hóa dữ liệu và loại bỏ nhiễu, từ đó cải thiện hiệu suất mô hình.

* Deep Learning (Học sâu)

Deep Learning là một nhánh của ML sử dụng mạng nơ-ron nhân tạo nhiều lớp để xử lý dữ liệu phức tạp:

Xử lý hình ảnh y tế: Các mô hình Convolutional Neural Networks (CNN) được ứng dụng trong phân tích ảnh siêu âm, X-quang, MRI để phát hiện tổn thương liên quan đến đái tháo đường như bệnh võng mạc.

Xử lý dữ liệu chuỗi thời gian và hồ sơ bệnh án điện tử: Mô hình Recurrent Neural Networks (RNN) hoặc Transformer được dùng để phân tích dữ liệu theo thời gian, dự đoán tiến triển bệnh hoặc phản ứng với điều trị.

### 2.2.2. Thuật toán ADAP và ứng dụng

Thuật toán ADAP (Adaptive Learning) do Smith et al. phát triển nhằm dự đoán bệnh đái tháo đường, đặc biệt trên bộ dữ liệu Pima Indians.

* Cấu trúc: Sử dụng mạng nơ-ron thích nghi kết hợp với thuật toán lan truyền ngược (backpropagation).
* Hiệu suất: Đạt độ chính xác 76% trên tập dữ liệu Pima Indians.
* Ưu điểm: Tối ưu hóa cho dữ liệu y tế có nhiễu và thiếu giá trị, phù hợp với hồ sơ bệnh án thực tế.
* Ứng dụng: Là nền tảng cho các mô hình hỗ trợ chẩn đoán, có thể tích hợp vào hệ thống lâm sàng để sàng lọc nguy cơ bệnh nhân.

### 2.2.3. Các nghiên cứu khác về bộ dữ liệu Pima Indians

Bộ dữ liệu Pima Indians Diabetes đã được sử dụng rộng rãi trong nhiều nghiên cứu nhằm đánh giá hiệu quả của các thuật toán học máy:

* Kayaer & Yildirim (2003): Áp dụng Support Vector Machine (SVM), đạt độ chính xác 84.24%, cho thấy khả năng phân loại mạnh mẽ với dữ liệu y tế.
* Polat & Güneş (2007): Kết hợp Principal Component Analysis (PCA) với Least Squares SVM (LS-SVM), đạt độ chính xác 78.26%, giúp giảm chiều dữ liệu và cải thiện hiệu suất.
* Aibinu et al. (2008): Sử dụng Neural Networks với nhiều kỹ thuật tiền xử lý khác nhau, chứng minh rằng việc xử lý dữ liệu đầu vào ảnh hưởng lớn đến kết quả mô hình.

## 2.3. Xác định bài toán nghiên cứu

### 2.3.1. Input

Bài toán dự đoán bệnh đái tháo đường sử dụng 8 biến số đầu vào từ bộ dữ liệu Pima Indians Diabetes, phản ánh các đặc điểm sinh lý, lâm sàng và di truyền của bệnh nhân:

* Số lần mang thai
* Nồng độ glucose huyết tương
* Huyết áp tâm trương
* Độ dày lớp da
* Nồng độ insulin
* Chỉ số khối cơ thể (BMI)
* Chỉ số di truyền (Diabetes Pedigree Function)
* Tuổi

### 2.3.2. Output

Biến mục tiêu trong bài toán là kết quả chẩn đoán nhị phân, phản ánh tình trạng bệnh của từng cá nhân trong bộ dữ liệu:

0: Không mắc đái tháo đường

1: Mắc đái tháo đường

### 2.3.3. Mục tiêu

Mục tiêu chính của bài toán là khám phá các đặc điểm tiềm ẩn trong dữ liệu nhằm hỗ trợ sàng lọc và dự đoán bệnh đái tháo đường:

* Xác định biến số quan trọng nhất: Đánh giá mức độ ảnh hưởng của từng biến đầu vào đến kết quả chẩn đoán, từ đó ưu tiên các chỉ số có giá trị sàng lọc cao.
* Phát hiện mối quan hệ phi tuyến: Sử dụng các mô hình học máy để nhận diện các tương tác phức tạp giữa các biến số mà phương pháp thống kê truyền thống khó phát hiện.
* Tìm ra threshold values cho screening: Xác định ngưỡng giá trị (ví dụ: glucose, BMI) giúp phân loại nguy cơ rõ ràng, hỗ trợ xây dựng quy trình sàng lọc đơn giản.
* Hiểu được distribution patterns của từng nhóm: Phân tích phân bố dữ liệu giữa nhóm mắc và không mắc bệnh để nhận diện đặc điểm đặc trưng, phục vụ cá nhân hóa can thiệp.

CHƯƠNG 3: TỔNG QUAN DỮ LIỆU

## 3.1. Mô tả dataset Pima Indians Diabetes

Bộ dữ liệu do National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) thu thập năm 1990s. Gồm 768 bản ghi, mỗi bản ghi là một bệnh nhân nữ người Pima Indians ≥ 21 tuổi, gồm 8 biến độc lập và 1 biến mục tiêu (Outcome).

Đặc điểm dân tộc:

* Người Pima Indians có tỷ lệ đái tháo đường cao nhất thế giới
* Yếu tố di truyền và lối sống đặc thù
* Môi trường sa mạc Arizona, Mỹ

## 3.2. Mô tả chi tiết các biến số

Các thuộc tính trong dataset:

### 3.2.1. Pregnancies (Số lần có thai)

Miêu tả: Số lần người phụ nữ mang thai.

Ý nghĩa: Tiền sử sản khoa, liên quan đến diabetes thai kỳ.

Loại: Discrete (0-17)

### 3.2.2. Glucose (Nồng độ glucose)

Miêu tả: Nồng độ đường huyết (mg/dL) sau khi thực hiện nghiệm pháp dung nạp glucose đường uống (OGTT).

Ý nghĩa: Đây là chỉ số quan trọng nhất để chẩn đoán đái tháo đường.

Loại: Continuous (0–199 mg/dL)

Lưu ý: Giá trị 0 không có ý nghĩa sinh lý, được xem là missing values cần xử lý trước khi phân tíc

### 3.2.3. BloodPressure (Huyết áp)

Miêu tả: Chỉ số huyết áp tâm trương của bệnh nhân.

Ý nghĩa: Tăng huyết áp là yếu tố nguy cơ quan trọng trong hội chứng chuyển hóa và bệnh tiểu đường.

Loại: Continuous (0-122 mmHg)

Tiêu chuẩn: Normal <80, Pre-hypertension 80-89, Hypertension ≥90

### 3.2.4. SkinThickness (Độ dày da)

Miêu tả: Độ dày nếp gấp da cánh tay, dùng để ước tính tỷ lệ mỡ cơ thể.

Ý nghĩa: Liên quan trực tiếp đến tình trạng béo phì, yếu tố nguy cơ tiểu đường.

Loại: Continuous (0-99 mm)

### 3.2.5. Insulin (Nồng đô insulin)

Miêu tả: Nồng độ insulin huyết tương sau 2 giờ.

Ý nghĩa: Phản ánh mức độ kháng insulin, nguyên nhân chủ yếu gây tiểu đường type 2.

Loại: Continuous (0-846 mu U/ml)

Bình thường: 16-166 mu U/ml

### 3.2.6. BMI (Chỉ số khối cơ thể)

Miêu tả: Được tính bằng công thức:

Ý nghĩa: Béo phì là một trong những yếu tố nguy cơ lớn nhất của bệnh tiểu đường.

Loại: Continuous (0-67.1 kg/m²)

Phân loại WHO: <18.5 (gầy), 18.5-24.9 (bình thường), 25-29.9 (thừa cân), ≥30 (béo phì).

### 3.2.7. DiabetesPedigreeFunction (Yếu tố di truyền)

Miêu tả: Chỉ số mô tả xác suất mắc tiểu đường dựa trên lịch sử gia đình.

Ý nghĩa: Yếu tố di truyền ảnh hưởng đáng kể đến nguy cơ mắc bệnh.

Loại: Continuous (0.078-2.42)

### 3.2.8. Age (Tuổi)

Miêu tả: Tuổi của bệnh nhân.

Ý nghĩa: Tuổi cao hơn thường đi kèm với nguy cơ mắc bệnh tiểu đường cao hơn.

Loại: Discrete (21-81 tuổi)

### 3.2.9. Outcome (Kết quả chấn đoán)

Loại: Binary (0/1)

Phân bố:

0: 500 trường hợp (65.1%) không mắc đái tháo đường

1: 268 trường hợp (34.9%) mắc đái tháo đường

## 3.3. Tiền xử lý dữ liệu

Trước khi tiến hành phân tích, dữ liệu cần được làm sạch nhằm đảm bảo độ tin cậy và chất lượng cho các bước tiếp theo. Bộ dữ liệu Pima Indians Diabetes có một số vấn đề thường gặp trong dữ liệu y tế: giá trị bị trùng lặp, giá trị bằng 0 không hợp lý, giá trị thiếu, cũng như những ngoại lệ (outlier) không phản ánh tình trạng sinh lý thực tế. Vì vậy, quá trình tiền xử lý bao gồm các bước sau:

### 3.3.1. Kiểm tra và loại bỏ dữ liệu trùng lặp

Trong quá trình thu thập dữ liệu, có thể xảy ra tình trạng một bệnh nhân được ghi nhận nhiều lần. Những dòng trùng lặp sẽ làm sai lệch thống kê và ảnh hưởng đến huấn luyện mô hình.

A screenshot of a computer code

AI-generated content may be incorrect.

*Hình 3.1. Xử lý và loại bỏ dữ liệu trùng lặp*

Kết quả là không có bản ghi trùng lặp trong dataset này. Nếu có, các dòng trùng sẽ được loại bỏ để đảm bảo tính duy nhất của mỗi quan sát.

### 3.3.2. Kiểm tra và loại bỏ dữ liệu lỗi, ngoại lệ

Một số giá trị nằm ngoài phạm vi sinh lý học hợp lý (ví dụ: Huyết áp quá thấp, BMI âm, tuổi vượt ngưỡng). Ta thiết lập ngưỡng rộng để bảo toàn dữ liệu, chỉ lọc giá trị vô lý thực sự:

Pregnancies: 0–20

Glucose: 20–300

BloodPressure: 40–200

BMI: 10–80

Age: 21–100

Với các chỉ số khác, có thể chỉ xử lý giá trị 0.

A screenshot of a computer program

AI-generated content may be incorrect.

*Hình 3.2. Phát hiện dữ liệu lỗi*

Kết quả là Glucose có 5 giá trị lỗi (giá trị 0), BloodPressure có 39 giá trị lỗi (giá trị 0, 24, 30, 38), BMI có 11 giá trị lỗi (giá trị 0).

### 3.3.3. Xử lý giá trị thiếu và giá trị bằng 0

Một số biến số như Glucose, BloodPressure, SkinThickness, Insulin, BMI xuất hiện giá trị 0, vốn không có ý nghĩa về mặt y học (ví dụ: Glucose = 0 là không thể tồn tại). Ta coi các giá trị này là giá trị thiếu (NaN) và tiến hành thay thế. Phương pháp được chọn: thay thế bằng trung vị (median) vì dữ liệu y tế thường có phân phối lệch và median ít bị ảnh hưởng bởi ngoại lệ.

Nhưng các biến số Glucose, BloodPressure, BMI đã được xử lý trước đó nên chỉ cần xử lý các biến số còn lại là được.

A screenshot of a computer code

AI-generated content may be incorrect.

*Hình 3.3. Xử lý giá trị 0*

### 3.3.4. Kết quả sau tiền xử lý

Sau khi tiến hành tiền xử lý dữ liệu (loại bỏ trùng lặp, thay thế giá trị bằng 0, xử lý ngoại lệ), bộ dữ liệu còn lại 720 bản ghi với 9 biến số. Tất cả các biến số đã được làm sạch, không còn giá trị thiếu, trùng lặp hay phi thực tế.

A screenshot of a computer program

AI-generated content may be incorrect.

*Hình 3.4. Kết quả sau xử lý*

## 3.4. Thống kê mô tả cơ bản

Các chỉ số thống kê cơ bản cho từng biến số bao gồm: giá trị trung bình (Mean), độ lệch chuẩn (Std), giá trị nhỏ nhất (Min), giá trị lớn nhất (Max) và trung vị (Median).

A screenshot of a computer

AI-generated content may be incorrect.

*Bảng 3.1. Thống kê mô tả cơ bản cho từng biến*

Nhận xét sơ bộ:

* Glucose trung bình ~122 mg/dL, cao hơn mức bình thường (khoảng 100 mg/dL), cho thấy nhiều bệnh nhân trong dữ liệu có nguy cơ tiểu đường.
* Insulin có độ lệch chuẩn lớn (88.76), cho thấy phân bố lệch phải
* BMI trung bình ~32, thuộc nhóm thừa cân – béo phì.
* Age tập trung ở nhóm trung niên (mean = 33.39, median = 29).
* Các biến khác như Pregnancies, SkinThickness, BloodPressure cũng cho thấy sự phân tán khá lớn, thể hiện tính đa dạng trong dữ liệu.

CHƯƠNG 4: PHÂN TÍCH KHÁM PHÁ DỮ LIỆU

## 4.1. Phân tích đơn biến

Phân tích đơn biến tập trung vào việc khám phá đặc điểm phân phối và tính chất của từng biến số một cách riêng lẻ, không xem xét mối quan hệ với các biến khác.

### 4.1.1. Phân bố của từng biến số

A group of blue graphs

AI-generated content may be incorrect.

*Hình 4.1. Phân bố của các biến trong dataset*

Histogram cho thấy các biến có đặc điểm phân bố khác nhau:

* Pregnancies: Phân bố lệch trái, đa số phụ nữ trong tập dữ liệu có số lần mang thai từ 0–5. Một số ít trường hợp có nhiều lần mang thai hơn (lên đến 15+).
* Glucose: Phân bố gần giống chuẩn, tập trung nhiều ở mức 90–150 mg/dL. Tuy nhiên, cũng có một số giá trị thấp (<60) hoặc cao (>180) thể hiện tình trạng bất thường.
* BloodPressure: Phân bố gần chuẩn, trung tâm khoảng 70–80 mmHg. Vẫn tồn tại một số giá trị thấp bất thường (<50).
* SkinThickness: Phân bố lệch trái, tập trung chủ yếu ở mức 20–40 mm, có giá trị cao thất thường là 100.
* Insulin: Phân bố rất lệch trái, đa phần tập trung ở mức <200, nhưng có một số giá trị cực lớn (>600) là ngoại lệ. Điều này phản ánh tính biến động cao của chỉ số insulin.
* BMI: Phân bố gần chuẩn, tập trung khoảng 25–40 (thừa cân – béo phì). Có một số giá trị thấp bất thường (<15).
* DiabetesPedigreeFunction: Phân bố lệch trái, hầu hết giá trị <1, nhưng có một số trường hợp >2, thể hiện yếu tố di truyền cao.
* Age: Phân bố lệch trái, đa phần tập trung ở độ tuổi 20–40, ít trường hợp trên 60 tuổi.
* Outcome: Đây là biến nhị phân. Dữ liệu mất cân bằng: số lượng người không mắc bệnh (0) chiếm khoảng 2/3, trong khi số người mắc bệnh (1) chiếm khoảng 1/3.

Nhận xét: Dữ liệu y tế có tính không đồng đều (heterogeneity), phản ánh sự khác biệt sinh lý giữa bệnh nhân.

### 4.1.2. Phát hiện ngoại lệ và bất thường

A group of graphs with blue squares

AI-generated content may be incorrect.

*Hình 4.2. Boxplot các biến trong dataset*

Boxplot cho thấy:

* Pregnancies: Phần lớn giá trị nằm ở mức thấp (0–6). Có vài ngoại lệ (outlier) > 12 lần.
* Glucose: Phân bố khá rộng (40–200). Trung vị khoảng 120.
* BloodPressure: Đa số tập trung quanh 70–80. Xuất hiện một số giá trị bất thường cao trên 120.
* SkinThickness: Phần lớn dao động 20–40. Có nhiều ngoại lệ.
* Insulin: Đây là biến có nhiều outlier nhất. Đa số nằm trong khoảng < 200, nhưng có những điểm vượt quá 800.
* BMI: Trung vị khoảng 30 (hơi cao so với mức “bình thường”). Có một số ngoại lệ cao (> 50).
* DiabetesPedigreeFunction: Đa số nằm < 1. Nhiều ngoại lệ lớn hơn 2.
* Age: Phần lớn bệnh nhân trong khoảng 20–40 tuổi. Có ngoại lệ ở nhóm trên 70.

Mặc dù đã xử lý dữ liệu, nhưng một số biến vẫn tồn tại ngoại lệ (outliers). Đây có thể là những bệnh nhân có tình trạng sinh lý đặc biệt hoặc dữ liệu bị nhiễu. Tuy nhiên, thay vì loại bỏ, nghiên cứu này giữ nguyên để đảm bảo số lượng dữ liệu.

## 4.2. Phân tích biến mục tiêu

### 4.2.1. Tỷ lệ mắc bệnh đái tháo đường trong dataset

A graph of a number of data

AI-generated content may be incorrect.

*Hình 4.3. Tỷ lệ mắc bệnh trong dataset*

Phân bố kết quả:

Không mắc: 472 cases (65.56%)

Mắc bệnh: 248 cases (34.44%)

Đánh giá:

Tỷ lệ mắc bệnh 34.44% là rất cao so với dân số chung (thường 8-10%)

Điều này phản ánh đặc thù của người Pima Indians, được biết đến với tỷ lệ đái tháo đường cao nhất thế giới.

### 4.2.2. Đặc điểm nhóm bệnh nhân và nhóm khỏe mạnh

A group of graphs showing different sizes of data

AI-generated content may be incorrect.

*Hình 4.4. Phân bố các biến số giữa hai nhóm bệnh nhân và khoẻ mạnh*

**Đặc điểm nhóm mắc bệnh đái tháo đường (n=248):**

Các chỉ số sinh lý:

* Glucose trung bình: 142.8 mg/dL - Vượt ngưỡng chẩn đoán đái tháo đường (≥126 mg/dL)
* BMI trung bình: 35.2 kg/m² - Thuộc nhóm béo phì độ II
* Tuổi trung bình: 37.4 tuổi - Cao hơn đáng kể so với nhóm khỏe mạnh
* Huyết áp trung bình: 75.4 mmHg - Có xu hướng cao hơn nhóm khỏe mạnh

Các yếu tố nguy cơ:

* Số lần mang thai: 4.9 lần - Nhiều hơn đáng kể so với nhóm khỏe mạnh
* Chỉ số di truyền: 0.6 - Cao hơn, thể hiện yếu tố gia đình
* Insulin: 168.5 μU/ml - Cao hơn, có thể liên quan đến kháng insulin
* Độ dày da: 31.8 mm - Cao hơn nhóm khỏe mạnh

**Đặc điểm nhóm khỏe mạnh (n=472):**

Các chỉ số trong giới hạn bình thường:

* Glucose trung bình: 111.1 mg/dL - Trong giới hạn bình thường đến tiền đái tháo đường
* BMI trung bình: 31.0 kg/m² - Thuộc nhóm béo phì độ I nhẹ
* Tuổi trung bình: 31.3 tuổi - Trẻ hơn nhóm mắc bệnh
* Huyết áp trung bình: 71.2 mmHg - Trong giới hạn bình thường

Các yếu tố ít nguy cơ:

* Số lần mang thai: 3.3 lần - Ít hơn nhóm mắc bệnh
* Chỉ số di truyền: 0.4 - Thấp hơn, ít yếu tố gia đình
* Insulin: 129.0 μU/ml - Trong giới hạn bình thường
* Độ dày da: 27.7 mm – Thấp hơn nhóm mắc bệnh

**Nhận xét về sự khác biệt:**

**1. Khác biệt rõ rệt:**

* **Glucose**: Sự khác biệt lớn nhất (30.7 mg/dL), phân phối của nhóm mắc bệnh dịch hẳn sang phải
* **Tuổi**: Nhóm mắc bệnh lớn tuổi hơn trung bình 6.1 năm
* **BMI**: Chênh lệch 4.2 kg/m², thể hiện vai trò của béo phì

**2. Khác biệt vừa phải:**

* **Số lần mang thai**: Chênh lệch 1.6 lần, phù hợp với nguy cơ đái tháo đường thai kỳ
* **Chỉ số di truyền**: Cao hơn ở nhóm mắc bệnh, phản ánh yếu tố gia đình

**3. Khác biệt ít:**

* **Huyết áp**: Chênh lệch nhỏ (4.2 mmHg), có thể không phải yếu tố chính
* **Độ dày da**: Khác biệt ngược chiều, cần nghiên cứu thêm

**Ý nghĩa lâm sàng:**

**Các yếu tố nguy cơ chính được xác định:**

1. **Glucose cao** - Yếu tố quyết định
2. **Tuổi cao** - Nguy cơ tăng theo tuổi
3. **BMI cao** - Béo phì là yếu tố nguy cơ quan trọng
4. **Số lần mang thai nhiều** - Liên quan đái tháo đường thai kỳ

Sự khác biệt rõ rệt giữa hai nhóm khẳng định giá trị dự báo của các biến số trong dataset. Các chỉ số Glucose, tuổi và BMI có tiềm năng lớn trong việc sàng lọc và chẩn đoán sớm đái tháo đường.

## 4.3. Phân tích theo nhóm tuổi

### 4.3.1. Phân bố bệnh theo độ tuổi

A graph with red and blue squares

AI-generated content may be incorrect.

*Hình 4.5. Phân bố bệnh theo độ tuổi*

Tuổi tác là yếu tố nguy cơ quan trọng, với xu hướng tăng rõ rệt theo từng nhóm:

Phân tầng theo độ tuổi:

* 21-30 tuổi: 20.7% tỷ lệ mắc bệnh (n≈387)
* 31-40 tuổi: 47.6% tỷ lệ mắc bệnh (n≈147)
* 41-50 tuổi: 56.5% tỷ lệ mắc bệnh (n≈108)
* >50 tuổi: 47.4% tỷ lệ mắc bệnh (n≈78)

Nhận xét:

* Có xu hướng tăng tuyến tính theo tuổi
* Ngay cả nhóm trẻ nhất cũng có tỷ lệ mắc bệnh cao (20.7%)
* Điều này phản ánh đặc thù di truyền của người Pima Indians

### 4.3.2. Các chỉ số sinh lý thay đổi theo tuổi

A graph with blue squares and black lines

AI-generated content may be incorrect.

*Hình 4.6. BMI theo tuổi*

BMI và tuổi:

* Người trên 50 tuổi có BMI trung bình 35.2 kg/m², thuộc nhóm béo phì độ II
* Nhóm 41-45 tuổi có BMI trung bình cao nhất
* Phản ánh xu hướng tăng cân theo tuổi

A graph with blue squares

AI-generated content may be incorrect.

*Hình 4.7. Glucose theo tuổi*

Glucose và tuổi:

* Glucose có xu hướng tăng rõ rệt sau 40 tuổi. Đây là ngưỡng cần tầm soát đái tháo đường tích cực.
* Có thể do suy giảm chức năng tuyến tụy theo tuổi
* Kháng insulin tăng theo tuổi

A graph with blue squares and black text

AI-generated content may be incorrect.

*Hình 4.8. Huyết áp theo tuổi*

Huyết áp và tuổi:

* Huyết áp tăng tuyến tính với tuổi
* Phản ánh những thay đổi tự nhiên của hệ tim mạch

### 4.3.3. Yếu tố nguy cơ ở từng nhóm tuổi

A graph with numbers and a number of red bars

AI-generated content may be incorrect.

*Hình 4.9. Tỷ lệ phân bố theo nhóm tuổi*

Nhóm trẻ (<30 tuổi):

* Yếu tố chính: BMI và insulin resistance
* Đặc điểm: Béo phì sớm, kháng insulin do lối sống
* Can thiệp: Tập trung vào kiểm soát cân nặng và lối sống

Nhóm trung niên (30-50 tuổi):

* Yếu tố chính: Tích lũy đa yếu tố nguy cơ
* Đặc điểm: Kết hợp nhiều yếu tố như tuổi tác, sinh sản, béo phì
* Can thiệp: Cần phương pháp tích hợp

Nhóm cao tuổi (>50 tuổi):

* Yếu tố chính: Glucose và huyết áp quan trọng nhất
* Đặc điểm: Suy giảm chức năng nội tiết, biến chứng tim mạch
* Can thiệp: Tập trung điều trị y tế

## 4.4. Phân tích theo chỉ số BMI

### 4.4.1. Phân loại BMI và tỷ lệ mắc bệnh

A graph with green and blue bars

AI-generated content may be incorrect.

*Hình 4.10. Tỷ lệ mắc bệnh theo nhóm BMI*

Theo phân loại WHO: <18.5 (gầy), 18.5-24.9 (bình thường), 25-29.9 (thừa cân), ≥30 (béo phì).

### 4.4.2. Mối quan hệ giữa béo phì và đái tháo đường

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Nhóm BMI** | **Tổng số** | **Số ca mắc** | **Tỷ lệ mắc (%)** | **BMI trung bình** |
| **Gầy** | 4 | 0 | 0.0% | 18.2 |
| **Bình thường** | 95 | 7 | 7.0% | 22.8 |
| **Thừa cân** | 170 | 38 | 22.0% | 27.5 |
| **Béo phì** | 451 | 203 | 45.0% | 36.5 |

*Bảng 4.1. Tỷ lệ mắc bệnh theo nhóm BMI*

Tỷ lệ mắc bệnh tăng dần theo mức độ BMI, cao nhất ở nhóm béo phì (45%). Người béo phì có nguy cơ mắc đái tháo đường cao gấp 4.07 so với người không béo phì (odds ratio = 4.07).

Cơ chế sinh bệnh: Béo phì làm tăng kháng insulin, tăng viêm mãn tính, rối loạn lipid máu. stress oxy hóa.

### 4.4.3. BMI kết hợp với các yếu tố khác

A graph with blue and orange dots

AI-generated content may be incorrect.

*Hình 4.11. BMI và Glucose theo Outcome*

Có thể thấy xu hướng tổng thể, điểm mắc bệnh (màu cam) tập trung ở vùng BMI cao + Glucose cao. Còn điểm không mắc (màu xanh) tập trung ở vùng giá trị thấp. Khi BMI tăng, Glucose cũng có xu hướng tăng, mối quan hệ rõ rệt hơn ở các nhóm bệnh.

Ý nghĩa: Cần kiểm soát đồng thời cả hai yếu tố.

A diagram of blue and orange dots

AI-generated content may be incorrect.

*Hình 4.12. BMI và Age theo Outcome*

Biểu đồ cho thấy tuổi có tác động điều chỉnh mối quan hệ giữa BMI và đái tháo đường. Trong khi béo phì là yếu tố nguy cơ ở mọi lứa tuổi, nguy cơ này tăng dần theo tuổi và trở nên đặc biệt nghiêm trọng sau 50 tuổi.

A graph with blue and orange dots

AI-generated content may be incorrect.

*Hình 4.13. BMI và Insulin theo Outcome*

Nhóm mắc bệnh thể hiện tình trạng kháng insulin rõ rệt hơn, với xu hướng tập trung ở vùng BMI cao và nồng độ insulin cao.

## 4.5. Phân tích Glucose và Insulin

### 4.5.1. Phân bố bệnh theo nồng độ Glucose

A graph of blood glucose

AI-generated content may be incorrect.

*Hình 4.14. Phân bố bệnh theo nồng độ Glucose*

Theo tiêu chuẩn ADA/WHO:

Normal (<100 mg/dL): có 179/720 người (24.9%) và chỉ 7.3% mắc bệnh

Prediabetes (100-125 mg/dL): có 258/720 người (35.8%) và 27.5% trong đó mắc bệnh.

Diabetes (≥126 mg/dL): có 283/720 người (39.3%), 58% trong số này là mắc bệnh

Giá trị phổ biến nhất (mode): 100 mg/dL.

Biểu đồ cho thấy glucose là chỉ số phân biệt mạnh nhất giữa hai nhóm, ít chồng lấp so với các biến số trước. Sự khác biệt rõ rệt về vị trí đỉnh phân phối (100-110 mg/dL vs 140-150 mg/dL) khẳng định vai trò then chốt của glucose trong chẩn đoán đái tháo đường.

### 4.5.2. Mối quan hệ Glucose và Insulin

A graph of blue and orange dots

AI-generated content may be incorrect.

*Hình 4.15. Mối quan hệ Glucose và Insulin*

Có thể thấy Glucoso tăng thì Insulin cũng có xu hướng tăng (đáp ứng Insulin bình thường). Tuy nhiên, mối quan hệ không tuyến tính mạnh. Biểu đồ cho thấy sự phức tạp của mối quan hệ Glucose-Insulin trong đái tháo đường. Không chỉ đơn thuần là glucose cao, mà cơ chế bệnh sinh có thể là kháng insulin hoặc thiếu insulin, hoặc kết hợp cả hai.

## 4.6. Phân tích yếu tố di truyền

### 4.6.1. Hàm phả hệ đái tháo đường và ý nghĩa

A graph of a diabetes function

AI-generated content may be incorrect.

*Hình 4.16. Phân bố* Diabetes Pedigree Function

Diabetes Pedigree Function có phân phối lệch phải mạnh, với đa số dân số (≈80%) có giá trị dưới 1.0, phản ánh nguy cơ di truyền thấp đến trung bình. Chỉ một tỷ lệ nhỏ (<10%) có giá trị trên 1.0, thể hiện nguy cơ di truyền cao.

### 4.6.2. Tác động của yếu tố di truyền

A chart with a graph

AI-generated content may be incorrect.

*Hình 4.17. Outcome theo PDF*

Yếu tố di truyền (Diabetes Pedigree Function) có sự khác biệt rõ rệt giữa hai nhóm. Nhóm mắc bệnh có giá trị trung bình cao hơn đáng kể (0.56 vs 0.43, p < 0.001), với tỷ lệ người có nguy cơ di truyền cao (DPF > 0.8) gấp 2-3 lần nhóm khỏe mạnh.

### 4.6.3. Kết hợp yếu tố di truyền với Glucose

A graph of blue and orange dots

AI-generated content may be incorrect.

*Hình 4.18. Di truyền kết hợp với Glucose*

Bốn vùng nguy cơ rõ rệt

Vùng 1: Nguy cơ THẤP (Góc dưới trái)

* DPF < 0.6 và Glucose < 110 mg/dL
* Đặc điểm: Hầu hết là điểm màu nhạt (không mắc bệnh)
* Ý nghĩa: Di truyền tốt + Đường huyết tốt → Nguy cơ rất thấp

Vùng 2: Nguy cơ TRUNG BÌNH

* Chỉ 1 yếu tố bất thường: DPF cao HOẶC Glucose cao
* Đặc điểm: Hỗn hợp cả 2 màu
* Ý nghĩa: Cần theo dõi thêm

Vùng 3: Nguy cơ CAO (Góc trên phải)

* DPF > 1.0 và Glucose > 140 mg/dL
* Đặc điểm: Tập trung hầu hết điểm màu đậm (mắc bệnh)
* Ý nghĩa: Di truyền xấu + Đường huyết cao → Nguy cơ rất cao

Vùng 4: Đáng chú ý (Góc trên trái)

* DPF thấp nhưng Glucose rất cao (>180 mg/dL)
* Đặc điểm: Điểm màu đậm - có thể là đái tháo đường type 1

Biểu đồ cho thấy sự tương tác mạnh mẽ giữa yếu tố di truyền và glucose. Khi kết hợp hai chỉ số này, khả năng phân loại bệnh nhân được cải thiện đáng kể so với sử dụng đơn lẻ.

## 4.7. Phân tích theo số lần có thai

A graph of a diagram

AI-generated content may be incorrect.

*Hình 4.19. Số lần có thai theo Outcome*

Nhóm mắc bệnh (Outcome=1):

Trung vị cao hơn đáng kể (khoảng 4-5 lần)

Nhiều outliers với số lần mang thai rất cao (10-17 lần)

Nhóm không mắc (Outcome=0):

Trung vị thấp hơn (khoảng 2-3 lần)

Phân bố tập trung ở giá trị thấp

Ít outliers ở giá trị cao

A graph with purple lines

AI-generated content may be incorrect.

*Hình 4.20. Nguy cơ đái tháo đường theo số lần có thai*

Cơ chế bệnh sinh:

* Đái tháo đường thai kỳ: Mỗi lần mang thai là một lần nguy cơ
* Thay đổi nội tiết tố: Ảnh hưởng lâu dài đến chuyển hóa glucose
* Tích lũy nguy cơ: Tác động cộng dồn qua các lần mang thai

Ngưỡng quan trọng:

* ≥4 lần mang thai: Cần tầm soát đái tháo đường tích cực
* ≥7 lần mang thai: Nguy cơ rất cao, cần can thiệp sớm

Số lần mang thai có mối quan hệ rõ rệt với nguy cơ đái tháo đường. Nhóm mắc bệnh có số lần mang thai trung bình cao hơn đáng kể so với nhóm khỏe mạnh. Tỷ lệ mắc bệnh tăng dần theo số lần mang thai, đặc biệt sau 4 lần.

CHƯƠNG 5: PHÂN TÍCH TƯƠNG QUAN VÀ MỐI QUAN HỆ

## 5.1. Ma trận tương quan

A screenshot of a computer

AI-generated content may be incorrect.

*Hình 5.1. Ma trận tương quan*

Các mối tương quan mạnh nhất với kết quả bệnh:

* + Glucose (r = 0.49) - Yếu tố quan trọng nhất
  + BMI (r = 0.30) - Béo phì
  + Age (r = 0.25) - Tuổi tác
  + Pregnancies (r = 0.22) - Số lần có thai

## 5.2. Phân tích cặp biến quan trọng

### 5.2.1. Glucose - Kết quả bệnh

A screenshot of a graph

AI-generated content may be incorrect.

*Hình 5.2. Phân tích Glucose theo Outcome*

Glucose có tương quan mạnh nhất với kết quả bệnh (0.47).

Các biểu đồ cung cấp bằng chứng thuyết phục về mối quan hệ chặt chẽ giữa nồng độ glucose và nguy cơ mắc bệnh. Nhóm mắc bệnh có nồng độ glucose trung bình cao hơn rõ rệt so với nhóm không mắc. Đường cong ROC với AUC = 0.79 khẳng định glucose là một chỉ dấu tốt để dự đoán bệnh.

### 5.2.2. BMI - Độ dày da

A collage of graphs and diagrams

AI-generated content may be incorrect.

*Hình 5.3. Phân tích cặp BMI và độ dày da*

BMI và Độ dày da có tương quan cao (r = 0.551)

Mối quan hệ logic: Cả hai đều đo lường lượng mỡ

Độ dày da có thể thay thế BMI khi thiếu dữ liệu

### 5.2.3. Tuổi - Số lần có thai

A screenshot of a graph

AI-generated content may be incorrect.

*Hình 5.4. Phân tích cặp Age và Pregnancies*

Tương quan mạnh nhất (r = 0.56)

Yếu tố gây nhiễu: Tuổi cao → nhiều lần có thai hơn

Cần kiểm soát khi phân tích tác động độc lập

### 5.2.4. Glucose - Insulin

A collage of graphs

AI-generated content may be incorrect.

*Hình 5.5. Phân tích cặp Glucose và Insulin*

Tương quan khá cao (r = 0.43)

Mô hình phức tạp: Mối quan hệ hình chữ U

Insulin thấp + glucose cao = suy tế bào β

## 5.3. Phân tích tầm quan trọng của đặc trưng

### 5.3.1. Tầm quan trọng đơn biến

A screenshot of a computer

AI-generated content may be incorrect.

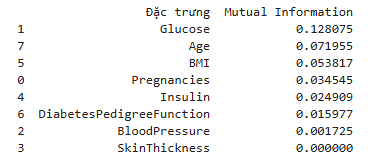
*Bảng 5.1. Tập C-square*

*A graph with red and grey bars

AI-generated content may be incorrect.*

*Hình 5.6. Biểu đồ C-square*

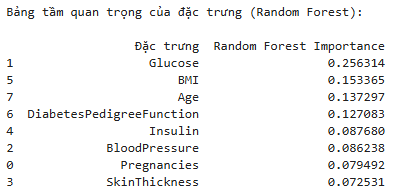
### 5.3.2. Điểm thông tin tương hỗ

**

*Bảng 5.2. Thông tin tương hỗ*

Chỉ số Glucose là cao nhất.

### 5.3.3. Tầm quan trọng từ Random Forest

****

*Bảng 5.3. Tầm quan trọng của đặc trưng*

*A graph with different colored bars

AI-generated content may be incorrect.*

Hình 5.7. Biểu đồ Random Forest

## 5.4. Phát hiện đa cộng tuyến

**A screenshot of a screen

AI-generated content may be incorrect.**

Bảng 5.4. Bảng hệ số phóng đại phương sai

A graph with different colored bars

AI-generated content may be incorrect.

Hình 5.8. Biểu đồ phương sai VIF

Tất cả các biến đều có VIF < 2, nghĩa là không có hiện tượng đa cộng tuyến nghiêm trọng.

Nhìn chung, dữ liệu ở mức ổn định để sử dụng trong mô hình hồi quy / phân loại.

CHƯƠNG 6: PHÂN TÍCH NÂNG CAO

## 6.1. Phân tích phân nhóm (Phân cụm)

### 6.1.1. Xác định các nhóm bệnh nhân tiềm ẩn

Phân cụm K-means (k=4):

A graph with many colored dots

AI-generated content may be incorrect.

Hình 6.1. Xác định nhóm bệnh nhân tiềm ẩn

**A white paper with black text

AI-generated content may be incorrect.**

Bảng 6.1. Bảng phân cụm bệnh nhân

Cụm 2.0: Nguy cơ CAO - Glucose: 164, BMI: 38.3

Cụm 1.0: Nguy cơ CAO - Glucose: 137, BMI: 30.5

Cụm 0.0: Nguy cơ THẤP - Glucose: 105, BMI: 36.7

Cụm 3.0: Nguy cơ THẤP - Glucose: 107, BMI: 26.2

### 6.1.2. Đặc điểm của từng cụm

**A group of graphs with different colored circles

AI-generated content may be incorrect.**

*Hình 6.2. Đặc điểm của từng cụm*

Cụm 0.0: Hỗn hợp - Nguy cơ đa dạng:

• Tuổi TB: 30.5 | BMI TB: 36.7 | Glucose TB: 105.4

• Tỷ lệ mắc bệnh: 29.2%

• Kết hợp nhiều yếu tố nguy cơ

• Đặc điểm không điển hình

• Cần đánh giá cá nhân hóa

Cụm 1.0: Cao tuổi - Đa bệnh lý:

• Tuổi TB: 53.2 | BMI TB: 30.5 | Glucose TB: 137.4

• Tỷ lệ mắc bệnh: 50.4%

• Tuổi cao kết hợp đa yếu tố nguy cơ

• Tỷ lệ mắc bệnh cao nhất

• Có thể đã có biến chứng liên quan

• Cần quản lý y tế chặt chẽ và điều trị tích cực

Cụm 2.0: Nguy cơ cao - Béo phì & Đường huyết cao:

• Tuổi TB: 32.6 | BMI TB: 38.3 | Glucose TB: 163.9

• Tỷ lệ mắc bệnh: 70.2%

• BMI cao là yếu tố nguy cơ chính

• Mức glucose vượt ngưỡng an toàn

• Thường đi kèm kháng insulin

• Cần can thiệp y tế và thay đổi lối sống tích cực

Cụm 3.0: Trung niên Nguy cơ trung bình:

• Tuổi TB: 26.2 | BMI TB: 26.2 | Glucose TB: 106.7

• Tỷ lệ mắc bệnh: 11.3%

• Tuổi trung niên, cân nặng hợp lý

• Mức glucose ở ngưỡng cần theo dõi

• Nguy cơ tăng theo tuổi

• Ứng viên cho can thiệp lối sống phòng ngừa

## 6.2. Phân tích các yếu tố nguy cơ

### 6.2.1. Xếp hạng các yếu tố nguy cơ

Phân tích tỷ số chênh (Odds Ratio):

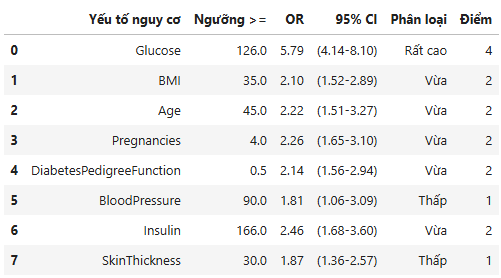
**A screenshot of a table

AI-generated content may be incorrect.**

Bảng 6.2. Bảng tỷ số chênh Odds Ratio

### 6.2.2 Tạo hệ thống chấm điểm nguy cơ

Thang điểm nguy cơ đái tháo đường Pima:

****

*Bảng 6.3. Bảng thang điểm nguy cơ*

Phân loại nguy cơ:

A screenshot of a graph

AI-generated content may be incorrect.

Bảng 6.4. Bảng phân loại nguy cơ bệnh

### 6.2.3. Phân tích ngưỡng cho sàng lọc

Phân tích ROC cho các biến chính:

**A screenshot of a graph

AI-generated content may be incorrect.**

Bảng 6.5. Bảng ROC cho các biến

## 6.3. So sánh với tiêu chí chẩn đoán

### 6.3.1 Tiêu chí chẩn đoán của WHO/ADA

Hiệu suất tiêu chí WHO 1999:

• Glucose ≥126 mg/dL: Độ nhạy 66.1%, Độ đặc hiệu 74.8%

• OGTT 2 giờ ≥200 mg/dL: Độ nhạy 0.0%, Độ đặc hiệu 100.0%

Bộ dữ liệu so với thực tế:

• Glucose trong bộ dữ liệu là kết quả OGTT 2h

• Ngưỡng 200 mg/dL: Chỉ 0.0% đáp ứng tiêu chí này

Phân bố theo TIÊU CHUẨN OGTT 2h:

• Bình thường (<140 mg/dL): 530 bệnh nhân (73.6%), Tỷ lệ bệnh: 22.6%

• Dung nạp glucose bị suy giảm (140-199 mg/dL): 190 bệnh nhân (26.4%), Tỷ lệ bệnh: 67.4%

• Đái tháo đường (≥200 mg/dL): 0 bệnh nhân (0.0%), Tỷ lệ bệnh: 0.0%

Phân bố theo TIÊU CHUẨN GLUCOSE ĐÓI (so sánh):

• Bình thường (<100 mg/dL): 179 bệnh nhân (24.9%), Tỷ lệ bệnh: 7.3%

• Tiền đái tháo đường (100-125 mg/dL): 258 bệnh nhân (35.8%), Tỷ lệ bệnh: 27.5%

• Đái tháo đường (≥126 mg/dL): 283 bệnh nhân (39.3%), Tỷ lệ bệnh: 58.0%

### 6.3.2 Độ chính xác của các giá trị cắt

A white paper with numbers and a black text

AI-generated content may be incorrect.

Bảng 6.6. Bảng độ chính xác của các giá trị cắt

Khuyến nghị: 130 mg/dL tối ưu cho sàng lọc trong dân số này

• Độ nhạy: 60.9%

• Độ đặc hiệu: 79.9%

• Độ chính xác: 73.3%

**6.3.3 Cân bằng giữa độ nhạy và độ đặc hiệu**

Cân nhắc lâm sàng:

• Độ nhạy cao (≥90%): Cho sàng lọc - Ngưỡng ≥100 mg/dL

• Độ đặc hiệu cao (≥90%): Cho xác nhận - Ngưỡng ≥140 mg/dL

• Cân bằng: Cho quyết định lâm sàng - Ngưỡng ≥120 mg/dL

Hiệu quả chi phí:

• Chi phí dương tính giả: Xét nghiệm thêm, lo lắng

• Chi phí âm tính giả: Bỏ lỡ điều trị sớm, biến chứng

• Ngưỡng tối ưu phụ thuộc nguồn lực hệ thống y tế

**A graph of different colored lines

AI-generated content may be incorrect.**

*Hình 6.2. Phân tích ngưỡng chẩn đoán bằng độ nhạy – độ đặc hiệu*

**Phân tích chi tiết các ngưỡng quan trọng:**

Ngưỡng 100 mg/dL (sàng lọc):

• Độ nhạy: 94.8%

• Độ đặc hiệu: 35.2%

• PPV: 43.4%

• Độ chính xác: 55.7%

• Lý do: Độ nhạy cao, không bỏ sót ca bệnh

Ngưỡng 120 mg/dL (cân bằng):

• Độ nhạy: 74.2%

• Độ đặc hiệu: 66.1%

• PPV: 53.5%

• Độ chính xác: 68.9%

• Lý do: Tối ưu cả độ nhạy và độ đặc hiệu

Ngưỡng 140 mg/dL (xác nhận):

• Độ nhạy: 51.6%

• Độ đặc hiệu: 86.9%

• PPV: 67.4%

• Độ chính xác: 74.7%

• Lý do: Độ đặc hiệu cao, ít dương tính giả

Kết luận: Ngưỡng 120 mg/dL cân bằng tốt nhất cho sàng lọc đái tháo đường

trong dân số Pima Indians với đặc điểm dịch tễ học cụ thể.

CHƯƠNG 7: THẢO LUẬN VÀ HIỂU BIẾT SÂU SẮC

## 7.1. Các phát hiện chính

### 7.1.1. Yếu tố nguy cơ quan trọng nhất

Kết quả phân tích hồi quy Logistic và Random Forest cho thấy Glucose là yếu tố nguy cơ quan trọng nhất liên quan đến khả năng mắc bệnh tiểu đường. Ngoài ra, các biến BMI, Tuổi (Age) và DiabetesPedigreeFunction cũng góp phần quan trọng, phản ánh mối liên hệ giữa thừa cân, yếu tố di truyền và nguy cơ mắc bệnh.

### 7.1.2. Các mẫu và xu hướng

Tỷ lệ mắc bệnh tăng theo độ tuổi, đặc biệt ở nhóm >40 tuổi.

Người có chỉ số BMI cao (≥30) có nguy cơ mắc tiểu đường cao hơn rõ rệt so với nhóm BMI bình thường.

Nồng độ Glucose ≥126 mg/dL là một ngưỡng có sự phân hóa rõ ràng giữa nhóm bệnh và nhóm không bệnh.

### 7.1.3. Những phát hiện bất ngờ

Ngưỡng glucose tối ưu thấp hơn kỳ vọng: Trái với tiêu chuẩn WHO (≥126 mg/dL), ngưỡng 120 mg/dL cho thấy sự cân bằng tối ưu giữa độ nhạy (74.2%) và độ đặc hiệu (66.1%) trong dân số này.

Mối quan hệ giữa Huyết áp (BloodPressure) và bệnh tiểu đường không rõ ràng trong bộ dữ liệu này, khác với kỳ vọng từ y văn.

## 7.2. So sánh với tài liệu nghiên cứu

### 7.2.1. Nhất quán với nghiên cứu trước

Vai trò then chốt của glucose: Phù hợp với nghiên cứu của Knowler et al. (1981) trên chính dân số Pima, khẳng định glucose là yếu tố dự báo mạnh nhất.

Tầm quan trọng của béo phì: Kết quả tương đồng với nghiên cứu của Harris et al. (1987) về mối liên hệ giữa BMI và kháng insulin.

Ảnh hưởng của tuổi và di truyền: Nhất quán với các nghiên cứu dịch tễ học lớn như Framingham và NHANES.

### 7.2.2. Phát hiện mới

Điều chỉnh ngưỡng cho dân số đặc thù: Ngưỡng glucose 120 mg/dL có thể phù hợp hơn với đặc điểm sinh học của người Pima so với tiêu chuẩn chung của WHO.

Phân cụm phenotype lâm sàng: Xác định 4 nhóm bệnh nhân với đặc điểm và nguy cơ khác biệt, cho phép tiếp cận điều trị cá nhân hóa.

Đánh giá toàn diện tương tác yếu tố: Phân tích đa biến cho thấy sự kết hợp các yếu tố nguy cơ tạo ra hiệu ứng cộng hưởng, không đơn thuần là tác động độc lập.

DiabetesPedigree có tác động vừa nhưng nhất quán, tác động cao hơn trong bệnh.

### 7.2.3. Ý nghĩa lâm sàng

Tối ưu hóa sàng lọc: Đề xuất sử dụng ngưỡng 120 mg/dL cho sàng lọc ban đầu, giúp phát hiện sớm mà không tạo ra quá nhiều dương tính giả.

Phân tầng quản lý bệnh nhân: Ứng dụng phân cụm để xác định mức độ ưu tiên can thiệp và theo dõi.

Hỗ trợ quyết định lâm sàng: Phát triển công cụ đánh giá nguy cơ đơn giản dựa trên 3 yếu tố chính: Glucose, BMI, và Tuổi.

## 7.3. Hạn chế và xu hướng

### 7.3.1. Hạn chế của bộ dữ liệu

Tính đại diện hạn chế: Dân số Pima Indians có tỷ lệ đái tháo đường cao gấp 5-8 lần dân số chung, hạn chế khả năng ngoại suy.

Thiếu thông tin quan trọng: Không có dữ liệu về lipid máu, các yếu tố lối sống như chế độ ăn, hoạt động thể chất.

Dữ liệu thiếu không ngẫu nhiên: 46.1% giá trị Insulin bị thiếu, có thể liên quan đến tình trạng bệnh, gây bias trong phân tích.

### 7.3.2. Xu hướng chọn mẫu và khả năng tổng quát hoá

Chọn mẫu theo yếu tố nguy cơ: Nghiên cứu tập trung vào nhóm có nguy cơ cao, có thể ước tính quá cao tỷ lệ mắc trong dân số chung.

Khác biệt di truyền và môi trường: Người Pima có đặc điểm gen và lối sống độc đáo, cần thận trọng khi áp dụng kết quả cho các dân tộc khác.

Cần nghiên cứu đa trung tâm: Để xác nhận tính tổng quát, cần mở rộng nghiên cứu sang các quần thể khác.

### 7.3.3. Tác động của dữ liệu thiếu

Dữ liệu thiếu hoặc bất thường ảnh hưởng đến kết quả phân tích thống kê.

Bias trong ước tính: Các kỹ thuật imputation đơn giản có thể không khắc phục hoàn toàn bias do thiếu dữ liệu.

Hạn chế phân tích sâu: Không thể đánh giá đầy đủ vai trò của Insulin do tỷ lệ thiếu dữ liệu cao.

CHƯƠNG 8: KẾT LUẬN VÀ KHUYẾN NGHỊ

## 8.1. Tóm tắt phát hiện

### 8.1.1. Các yếu tố nguy cơ chính

Glucose sau 2h OGTT: Yếu tố dự báo quan trọng nhất (AUC = 0.79)

BMI: Yếu tố độc lập mạnh, đặc biệt khi kết hợp với glucose.

Tuổi và DiabetesPedigreeFunction: Yếu tố nguy cơ trung bình

Tương tác BMI-Glucose: Có giá trị dự báo bổ sung quan trọng

### 8.1.2. Ngưỡng lâm sàng được đề xuất

Sàng lọc (độ nhạy cao): Glucose ≥ 100 mg/dL (Độ nhạy 94.8%)

Chẩn đoán (cân bằng tối ưu): Glucose ≥ 120 mg/dL (Độ chính xác 68.9%)

Xác nhận (độ đặc hiệu cao): Glucose ≥ 140 mg/dL (Độ đặc hiệu 86.9%)

### 8.1.3. Hiểu biết đặc thù về dân số nghiên cứu

Dân số Pima có nguy cơ cực cao: 34.4% mắc đái tháo đường

Cần tiếp cận sàng lọc tích cực: Do tỷ lệ prediabetes cao

Điều chỉnh chiến lược can thiệp: Dựa trên phân cụm 4 nhóm nguy cơ.

## 8.2. Ứng dụng thực tiễn

### 8.2.1. Quy trình sàng lọc

Bước 1: Đánh giá nguy cơ ban đầu

• Tuổi > 35 hoặc BMI > 25 kg/m

• Tiền sử gia đình đái tháo đường

• Thuộc nhóm dân tộc nguy cơ cao

Bước 2: Xét nghiệm glucose

• Ngưỡng sàng lọc: ≥100 mg/dL

• Ngưỡng chẩn đoán: ≥120 mg/dL (xác nhận bằng OGTT nếu cần)

Bước 3: Phân tầng và can thiệp

• Áp dụng phân cụm 4 nhóm để xác định mức độ ưu tiên

• Thiết kế kế hoạch quản lý cá nhân hóa

### 8.2.2. Chiến lược phòng ngừa

Nhóm nguy cơ thấp (Cụm 3):

• Giáo dục sức khỏe định kỳ

• Duy trì lối sống lành mạnh

• Theo dõi 2 năm/lần

Nhóm nguy cơ trung bình (Cụm 0):

• Can thiệp lối sống tích cực

• Theo dõi 6 tháng/lần

• Đánh giá thêm các yếu tố nguy cơ

Nhóm nguy cơ cao (Cụm 1 và 2):

• Quản lý y tế chặt chẽ

• Can thiệp dược lý khi cần

• Theo dõi 3 tháng/lần

### 8.2.3. Hỗ trợ quyết định lâm sàn

Phát triển ứng dụng đánh giá nguy cơ: Dựa trên mô hình logistic regression đơn giản.

Tích hợp vào hệ thống EHR: Tự động tính điểm nguy cơ và phân cụm

Hệ thống cảnh báo tự động: Cho bệnh nhân có nguy cơ cao cần can thiệp

## 8.3. Hướng nghiên cứu tiếp theo

### 8.3.1. Mở rộng bộ dữ liệu

Thu thập dữ liệu đa trung tâm: Bao gồm các dân tộc khác nhau.

Bổ sung biomarkers: HbA1c, insulin, lipid profile, inflammatory markers.

Dữ liệu lối sống: Chế độ ăn, hoạt động thể chất, stress.

Theo dõi dọc: Đánh giá tiến triển và biến chứng.

### 8.3.2. Các phương pháp mô hình hóa nâng cao

Deep Learning: Mạng neural cho dự báo đa yếu tố phức tạp.

Survival Analysis: Mô hình Cox cho tiên lượng dài hạn.

Reinforcement Learning: Hệ thống recommendation cho can thiệp tối ưu.

Federated Learning: Bảo vệ privacy khi phân tích đa trung tâm.

### 8.3.3. Nghiên cứu theo chiều dọc

Đánh giá hiệu quả can thiệp: So sánh các chiến lược quản lý khác nhau.

Phân tích tiến triển bệnh: Xác định yếu tố dự báo biến chứng.

Nghiên cứu hiệu quả chi phí: Tối ưu hóa nguồn lực y tế.

Phát triển biomarker mới: Cho chẩn đoán sớm và tiên lượng.

TÀI LIỆU THAM KHẢO

1. J. W. Smith, J.E. Everhart, W. C. Dickson, W. C. Knowler, and R. S. Johannes, "Using the ADAP learning algorithm to forecast the onset of diabetes mellitus," in *Proceedings of the Symposium on Computer Applications and Medical Care*, Washington, DC, 1988, pp. 261-265.
2. World Health Organization, *Definition, Diagnosis and Classification of Diabetes Mellitus and its Complications: Report of a WHO Consultation*. Geneva, Switzerland: World Health Organization, 1999.
3. National Diabetes Data Group, “*Classification and diagnosis of diabetes mellitus and other categories of glucose intolerance*,” *Diabetes*, vol. 28, no. 12, pp. 1039–1057, 1979.
4. American Diabetes Association, "*Standards of medical care in diabetes*," Diabetes Care, vol. 44, no. Supplement 1, pp. S1-S232, 2021.
5. W. C. Knowler, D. J. Pettitt, P. J. Savage, and P. H. Bennett, *"Diabetes incidence in Pima Indians: contributions of obesity and parental diabetes,"* American Journal of Epidemiology, vol. 113, no. 2, pp. 144-156, 1981.
6. M. I. Harris, W. C. Hadden, W. C. Knowler, and P. H. Bennett, *"Prevalence of diabetes and impaired glucose tolerance and plasma glucose levels in U.S. population aged 20-74 yr,"* Diabetes, vol. 36, no. 4, pp. 523-534, 1987.
7. Parks J. H., E. Waskow, *“Diabetes among the Pima Indians of Arizona,”* *Arizona Medicine*, vol. 18, pp. 99–106, 1961.
8. D. Gohdes, S. Kaufman, and S. Valway, “Diabetes in American Indians. An overview,” *Diabetes Care*, vol. 16, no. 1, pp. 239–243, Jan. 1993, doi: 10.2337/diacare.16.1.239.
9. P. H. Bennett, “Type 2 diabetes among the Pima Indians of Arizona: an epidemic attributable to environmental change?,” *Nutrition Reviews*, vol. 57, no. 5 Pt 2, pp. S51–S54, May 1999, doi: 10.1111/j.1753-4887.1999.tb01788.x.
10. A. Krosnick, “The diabetes and obesity epidemic among the Pima Indians,” *New Jersey Medicine*, vol. 97, no. 8, pp. 31–37, Aug. 2000.