# GIỚI THIỆU

### 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.

Phân tích khám phá dữ liệu (Exploratory Data Analysis - EDA) được triển khai nhằm:

Hiểu rõ phân bố và mối quan hệ giữa các biến số

Phát hiện các mẫu tiềm ẩn, xu hướng và bất thường trong dữ liệu

Định hướng cho việc xây dựng mô hình dự đoán hiệu quả hơn

### 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

### 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.

### 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:

Mặc dù bộ dữ liệu Pima Indians Diabetes là một nguồn tài nguyên quý giá, nghiên cứu vẫn tồn tại một số giới hạn cần được nhận diện rõ ràng:

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. Điều này hạn chế khả năng áp dụng mô hình trong các chương trình sàng lọc toàn diện.

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, đặc biệt là ở châu Á hoặc châu Phi.

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.

# CƠ SỞ LÝ THUYẾT VÀ NGHIÊN CỨU LIÊN QUAN

### Tổng quan về bệnh đái tháo đường

Bệnh đái tháo đường (Diabetes Mellitus) là một trong những bệnh lý chuyển hóa phổ biến nhất hiện nay, với tỷ lệ mắc ngày càng gia tăng trên toàn cầu. Đây là một tình trạng rối loạn nội tiết kéo dài, ảnh hưởng đến khả năng điều hòa glucose trong máu, từ đó dẫn đến nhiều biến chứng nghiêm trọng như bệnh tim mạch, tổn thương thần kinh, suy thận và mù lòa.

Theo các nghiên cứu dịch tễ học, đái tháo đường không chỉ là gánh nặng cá nhân mà còn là thách thức lớn đối với hệ thống y tế và nền kinh tế quốc gia. Việc hiểu rõ bản chất bệnh học và phân loại chính xác là nền tảng cho các chiến lược phòng ngừa, chẩn đoán và điều trị hiệu quả.

* + 1. Định nghĩa và phân loại

Định nghĩa: 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 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…

* + 1. 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.

* + 1. 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.

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

* + 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.

* + 1. 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.

* + 1. 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.

### Xác định bài toán nghiên cứu

* + 1. Input (Đầu vào): Các chỉ số sinh lý và lâm sàng

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:

Đặc điểm nhân khẩu học:

Tuổi

Số lần mang thai

Chỉ số anthropometric:

BMI (Body Mass Index)

Độ dày lớp da dưới cánh tay (triceps skin fold thickness)

Chỉ số sinh hóa:

Nồng độ glucose huyết tương lúc đói

Nồng độ insulin

Chỉ số sinh lý:

Huyết áp tâm trương

Yếu tố di truyền:

Diabetes Pedigree Function – chỉ số phản ánh nguy cơ di truyền mắc bệnh trong gia đình

* + 1. Output (Đầu ra): Dự đoán khả năng mắc đái tháo đường

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

Mục tiêu của mô hình học máy là phân loại chính xác từng trường hợp dựa trên các chỉ số đầu vào, từ đó hỗ trợ sàng lọc, can thiệp sớm và ra quyết định lâm sàng.

* + 1. Mục tiêu: Phát hiện các yếu tố nguy cơ và patterns

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.

# TỔNG QUAN DỮ LIỆU

### Mô tả dataset Pima Indians Diabetes

Nguồn gốc: National Institute of Diabetes and Digestive and Kidney Diseases

Đối tượng: Phụ nữ người Pima Indians ≥ 21 tuổi

Thời gian thu thập: 1990s

Kích thước: 768 quan sát × 9 biến (8 features + 1 target)

Đặc điểm dân tộc học:

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

### Các biến số trong dataset

* + 1. Pregnancies (Số lần có thai)

Loại: Discrete (0-17)

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

Phân bố: Right-skewed, median = 3

* + 1. Glucose (Nồng độ glucose)

Loại: Continuous (0–199 mg/dL)

Ý nghĩa: Glucose huyết tương sau 2 giờ nghiệm pháp dung nạp glucose (OGTT)

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

* + 1. BloodPressure (Huyết áp)

Loại: Continuous (0-122 mmHg)

Ý nghĩa: Huyết áp tâm trương

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

* + 1. SkinThickness (Độ dày da)

Loại: Continuous (0-99 mm)

Ý nghĩa: Độ dày nếp gấp da triceps

Mục đích: Ước lượng tỷ lệ mỡ cơ thể

* + 1. Insulin (Nồng độ insulin)

Loại: Continuous (0-846 mu U/ml)

Ý nghĩa: Insulin huyết thanh sau 2h OGTT

Bình thường: 16-166 mu U/ml

* + 1. BMI (Chỉ số khối cơ thể)

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)

* + 1. DiabetesPedigreeFunction (Yếu tố di truyền)

Loại: Continuous (0.078-2.42)

Ý nghĩa: Hàm ước lượng khả năng mắc bệnh dựa trên gia đình

Tính toán: Dựa trên tuổi và mối quan hệ của người thân bị bệnh

* + 1. Age (Tuổi)

Loại: Discrete (21-81 tuổi)

Phân bố: Right-skewed, median = 29

Nhóm nguy cơ: ≥45 tuổi theo ADA

* + 1. Outcome (Kết quả chẩn 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

### Thống kê mô tả cơ bản

| **Biến số** | **Mean** | **Std** | **Min** | **Max** | **Median** |
| --- | --- | --- | --- | --- | --- |
| Pregnancies | 3.8 | 3.4 | 0 | 17 | 3 |
| Glucose | 120.9 | 31.9 | 0 | 199 | 117 |
| BloodPressure | 69.1 | 19.4 | 0 | 122 | 72 |
| SkinThickness | 20.5 | 16.0 | 0 | 99 | 23 |
| Insulin | 79.8 | 115.2 | 0 | 846 | 30.5 |
| BMI | 31.9 | 7.9 | 0 | 67.1 | 32.3 |
| DiabetesPedigree | 0.47 | 0.33 | 0.08 | 2.42 | 0.37 |
| Age | 33.2 | 11.8 | 21 | 81 | 29 |
|  |  |  |  |  |  |

* 1. Kiểm tra chất lượng dữ liệu

Missing values (thể hiện bằng 0):

Glucose: 5 cases (0.65%)

BloodPressure: 35 cases (4.56%)

SkinThickness: 227 cases (29.56%)

Insulin: 374 cases (48.70%)

BMI: 11 cases (1.43%)

Outliers:

Insulin: Một số giá trị cực cao (>400)

BMI: Một số giá trị cực cao (>50)

Age: Phân bố không đều, tập trung ở nhóm trẻ