# GIỚI THIỆU

### Tổng quan về bài toán

Bệnh đái tháo đường (diabetes mellitus) là một trong những vấn đề sức khỏe cộng đồng lớn nhất toàn cầu, ảnh hưởng đến hàng trăm triệu người và gây ra chi phí y tế khổng lồ. Việc phát hiện sớm và dự đoán chính xác nguy cơ mắc bệnh có ý nghĩa quan trọng trong việc can thiệp kịp thời và giảm thiểu biến chứng.

Nghiên cứu này sử dụng bộ dữ liệu Pima Indians Diabetes từ UCI Machine Learning Repository để thực hiện phân tích khám phá dữ liệu (EDA) nhằm hiểu rõ hơn về các yếu tố nguy cơ và mô hình bệnh học của đái tháo đường.

### Mục tiêu nghiên cứu

Mục tiêu chính:

* Xác định các yếu tố nguy cơ quan trọng nhất dẫn đến bệnh đái tháo đường
* Phân tích mối quan hệ giữa các biến số sinh lý và khả năng mắc bệnh
* Đánh giá tính khả thi của việc sử dụng các chỉ số sinh lý để sàng lọc bệnh

Mục tiêu phụ:

* So sánh kết quả với các tiêu chuẩn chẩn đoán quốc tế
* Đưa ra khuyến nghị cho việc phát hiện sớm bệnh đái tháo đường
* Xác định những hạn chế và xu hướng trong dữ liệu

### Ý nghĩa thực tiễn

Kết quả nghiên cứu có thể:

* Hỗ trợ bác sĩ lâm sàng trong việc đánh giá nguy cơ bệnh nhân
* Cung cấp cơ sở khoa học cho các chương trình sàng lọc cộng đồng
* Góp phần vào việc phát triển các công cụ hỗ trợ quyết định lâm sàng
* Tối ưu hóa việc phân bổ nguồn lực y tế cho các nhóm nguy cơ cao

### Phạm vi và giới hạn nghiên cứu

Phạm vi:

* Phân tích dữ liệu từ 768 phụ nữ người Pima Indians
* 8 biến số sinh lý và lâm sàng
* Kết quả chẩn đoán nhị phân (mắc/không mắc)

Giới hạn:

* Chỉ áp dụng cho phụ nữ từ 21 tuổi trở lên
* Dữ liệu từ một nhóm dân tộc cụ thể
* Thiếu thông tin về lối sống, chế độ ăn, thuốc men

# CƠ SỞ LÝ THUYẾT VÀ NGHIÊN CỨU LIÊN QUAN

### Tổng quan về bệnh đái tháo đường

* + 1. Định nghĩa và phân loại

Định nghĩa: Đái tháo đường là một nhóm các rối loạn chuyển hóa đặc trưng bởi tăng glucose máu mãn tính do khiếm khuyết tiết insulin, tác dụng insulin, hoặc cả hai.

Phân loại chính (theo WHO 1999):

* + - * Type 1: Thiếu insulin tuyệt đối do phá hủy tế bào β tuyến tụy
      * Type 2: Kháng insulin và/hoặc giảm tiết insulin tương đối
      * Gestational diabetes: Đái tháo đường thai kỳ
      * Các dạng khác: MODY, do thuốc, do bệnh tuyến tụy
    1. Tiêu chuẩn chẩn đoán Theo WHO/ADA:
       - Glucose máu lúc đói ≥ 126 mg/dL (7.0 mmol/L)
       - Glucose máu sau 2h OGTT ≥ 200 mg/dL (11.1 mmol/L)
       - HbA1c ≥ 6.5% (48 mmol/mol)
       - Glucose máu bất kỳ ≥ 200 mg/dL với triệu chứng
    2. Các biến chứng và yếu tố nguy cơ

Biến chứng cấp tính: Nhiễm toan ceton, hôn mê tăng thẩm thấu Biến chứng mãn tính: Bệnh thận, võng mạc, thần kinh, tim mạch

Yếu tố nguy cơ:

* + - * Tuổi tác, béo phì, ít vận động
      * Tiền sử gia đình, yếu tố di truyền
      * Tiền sử đái tháo đường thai kỳ
      * Hội chứng chuyển hóa

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

* + 1. Phương pháp machine learning trong y tế

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

* + - * Supervised learning: Classification, regression cho chẩn đoán
      * Unsupervised learning: Clustering để phân nhóm bệnh nhân
      * Deep learning: Xử lý hình ảnh y tế, dữ liệu phức tạp
    1. Thuật toán ADAP và ứng dụng

Thuật toán ADAP (Adaptive Learning) được Smith et al. phát triển đặc biệt cho dự đoán đái tháo đường:

* + - * Sử dụng adaptive network với backpropagation
      * Đạt độ chính xác 76% trên dataset Pima Indians
      * Tối ưu hóa cho dữ liệu y tế với noise và missing values
    1. Các nghiên cứu khác về bộ dữ liệu Pima Indians
       - Nghiên cứu của Kayaer & Yildirim (2003): SVM đạt 84.24%
       - Polat & Güneş (2007): PCA + LS-SVM đạt 78.26%
       - Aibinu et al. (2008): Neural networks với các preprocessing khác nhau

### Xác định bài toán nghiên cứu

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

8 biến số đầu vào bao gồm thông tin về:

* Đặc điểm nhân khẩu học (tuổi, số lần có thai)
* Chỉ số anthropometric (BMI, độ dày da)
* Chỉ số sinh hóa (glucose, insulin)
* Chỉ số sinh lý (huyết áp)
* Yếu tố di truyền (diabetes pedigree function)
  + 1. Output (Đầu ra): Dự đoán khả năng mắc đái tháo đường

Biến mục tiêu là kết quả chẩn đoán nhị phân:

* + - * 0: Không mắc đái tháo đường
      * 1: Mắc đái tháo đường
    1. Mục tiêu: Phát hiện các yếu tố nguy cơ và patterns
       - Xác định biến số quan trọng nhất
       - Phát hiện mối quan hệ phi tuyến
       - Tìm ra threshold values cho screening
       - Hiểu được distribution patterns của từng nhóm

# 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
    2. Glucose (Nồng độ glucose)
       - Loại: Continuous (0-199 mg/dL)
       - Ý nghĩa: Glucose huyết tương sau 2h OGTT
       - Lưu ý: Giá trị 0 có thể là missing values
    3. BloodPressure (Huyết áp)
       - Loại: Continuous (0-122 mmHg)
       - Ý nghĩa: Huyết áp tâm trương
       - Tiêu chuẩn: Normal <80, Pre-hypertension 80-89, Hypertension ≥90
    4. 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
    5. 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
    6. 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)
    7. 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
    8. Age (Tuổi)
       - Loại: Discrete (21-81 tuổi)
       - Phân bố: Right-skewed, median = 29
       - Nhóm nguy cơ: ≥45 tuổi theo ADA
    9. Outcome (Kết quả chẩn đoán)
       - Loại: Binary (0/1)
       - Phân bố: 500 cases (65.1%) không mắc, 268 cases (34.9%) mắc bệnh

### Thống kê mô tả cơ bản

A screenshot of a computer

AI-generated content may be incorrect.

* 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