# Tóm tắt nội dung

## (file pima\_indians\_diabetes\_uci.htm)

**Diabetes Dataset (AIM ’94) – UCI Repository**

* **Đặc điểm**: Multivariate, Time-Series
* **Lĩnh vực**: Y tế & sức khỏe
* **Nhiệm vụ chính**: Phân loại (Classification)
* **Kích thước**: 1 instance, 20 features
* **Loại dữ liệu**: Categorical, Integer
* **Thiếu dữ liệu**: Không

**Thông tin dữ liệu**

* Nguồn thu thập:
  + **Thiết bị điện tử tự động** (có timestamp thực tế)
  + **Hồ sơ giấy** (chỉ theo mốc “logical time”: sáng, trưa, tối, trước khi ngủ → quy đổi thành 08:00, 12:00, 18:00, 22:00).
* **Cấu trúc file** (4 trường/record, tab-separated):
  + Ngày (MM-DD-YYYY)
  + Giờ (HH:MM)
  + Code (mã hoạt động/đo lường)
  + Giá trị (Value)
* **Các mã code tiêu biểu**:
  + 33–35: Liều insulin (Regular, NPH, UltraLente)
  + 48, 57–64: Đo glucose (trước/sau bữa ăn, không xác định)
  + 65: Triệu chứng hạ đường huyết
  + 66–68: Lượng bữa ăn (bình thường/nhiều/ít hơn)
  + 69–71: Hoạt động thể dục (bình thường/nhiều/ít hơn)
  + 72: Sự kiện đặc biệt

**Input (dữ liệu đầu vào)**

* 4 trường trong mỗi bản ghi:
  1. Ngày (MM-DD-YYYY)
  2. Giờ (HH:MM)
  3. Code (mã sự kiện: insulin, glucose, ăn uống, vận động, triệu chứng, …)
  4. Giá trị (Value: liều, mức glucose, …)

**Output (dữ liệu đầu ra)**

* Chuỗi sự kiện y tế của bệnh nhân tiểu đường theo thời gian (time-series).
* Có thể dùng để dự đoán/nhận diện các trạng thái: đường huyết, tác động của insulin, triệu chứng hạ đường huyết, ảnh hưởng bữa ăn và vận động.

**Mục tiêu**

* Phân tích hành vi và tiến triển bệnh tiểu đường theo thời gian.
* Xây dựng mô hình dự đoán biến động glucose và phản ứng của bệnh nhân.
* Hỗ trợ ra quyết định điều trị (insulin, ăn uống, hoạt động).

## (file pima\_indians\_diabetes.names)

**Nguồn gốc**

* Chủ sở hữu: National Institute of Diabetes and Digestive and Kidney Diseases
* Người đóng góp: Vincent Sigillito (Johns Hopkins University, Applied Physics Lab)
* Ngày nhận: 09/05/1990

**Mục đích & Sử dụng trước đây**

* Nghiên cứu: Smith et al. (1988) dùng thuật toán ADAP để dự đoán nguy cơ mắc tiểu đường theo tiêu chuẩn WHO.
* Độ chính xác: Sensitivity & Specificity ~ 76% trên tập kiểm thử.
* Đối tượng: Phụ nữ ≥ 21 tuổi, người Pima Indian, sống gần Phoenix (Arizona, USA).

**Thông tin dữ liệu**

* **Số mẫu (instances)**: 768
* **Số thuộc tính (attributes)**: 8 + 1 (class)
* **Kiểu dữ liệu**: Numeric (có missing values)
* **Phân bố lớp**:
  + Class 0 (không tiểu đường): 500
  + Class 1 (có tiểu đường): 268

**Thuộc tính**

1. Số lần mang thai
2. Glucose (2h OGTT)
3. Huyết áp tâm trương
4. Độ dày da cánh tay (mm)
5. Insulin 2h (mu U/ml)
6. BMI (kg/m²)
7. Chỉ số di truyền tiểu đường (Diabetes pedigree)
8. Tuổi (năm)
9. Biến mục tiêu: Có/Không tiểu đường (0/1)

**Thống kê nhanh (Mean ± SD)**

* Pregnancies: 3.8 ± 3.4
* Glucose: 120.9 ± 32.0
* Blood pressure: 69.1 ± 19.4
* Skin fold: 20.5 ± 16.0
* Insulin: 79.8 ± 115.2
* BMI: 32.0 ± 7.9
* Pedigree: 0.5 ± 0.3
* Age: 33.2 ± 11.8

**Input (dữ liệu đầu vào)**

* 8 thuộc tính lâm sàng của bệnh nhân (numeric):
  1. Số lần mang thai
  2. Nồng độ glucose (2h OGTT)
  3. Huyết áp tâm trương
  4. Độ dày da cánh tay
  5. Insulin 2h
  6. BMI
  7. Diabetes pedigree function (yếu tố di truyền)
  8. Tuổi

**Output (dữ liệu đầu ra)**

* Biến mục tiêu nhị phân (class):
  + 0 = không mắc tiểu đường
  + 1 = có tiểu đường

**Mục tiêu**

* Dự đoán khả năng mắc tiểu đường của bệnh nhân dựa trên các chỉ số y tế.
* Bài toán **Classification (phân loại nhị phân)**, được dùng rộng rãi để so sánh các thuật toán Machine Learning.

## (file pima\_indians\_diabetes.data.csv)

File dữ liệu chưa có tên cột

**Thống kê tổng quan:**

Số dòng (samples): 768

Số cột (features): 9

Không có giá trị null

**Tên cột(cần gắn vào)**

1. Pregnancies – số lần mang thai
2. Glucose – nồng độ glucose 2h OGTT
3. BloodPressure – huyết áp tâm trương
4. SkinThickness – độ dày da cánh tay
5. Insulin – insulin 2h
6. BMI – chỉ số khối cơ thể
7. DiabetesPedigreeFunction – chỉ số di truyền
8. Age – tuổi
9. Outcome – biến mục tiêu (0 = không tiểu đường, 1 = tiểu đường)

**Phân bố class (Outcome)**

* 0 (không tiểu đường): **500 mẫu**
* 1 (có tiểu đường): **268 mẫu**  
  → Tỉ lệ mắc bệnh ≈ **34.9%**

SLIDE

**🔹 AIM ’94 Diabetes Dataset (UCI)**

* **Đặc điểm:** Multivariate, Time-Series
* **Nhiệm vụ:** Classification
* **Kích thước:** 1 instance – 20 features
* **Loại dữ liệu:** Categorical, Integer – Không thiếu dữ liệu
* **Nguồn:** Ghi nhận từ thiết bị tự động (có timestamp) và hồ sơ giấy (mốc logical time)
* **Input:** Ngày, Giờ, Code (insulin, glucose, ăn uống, vận động…), Giá trị
* **Output:** Chuỗi sự kiện y tế theo thời gian
* **Mục tiêu:** Phân tích tiến triển bệnh, dự đoán biến động glucose, hỗ trợ quyết định điều trị

**🔹 Pima Indians Diabetes Dataset (UCI)**

* **Chủ sở hữu:** NIDDK – Đóng góp: Johns Hopkins APL
* **Số mẫu:** 768 phụ nữ Pima ≥ 21 tuổi
* **Số thuộc tính:** 8 numeric + 1 class (0/1)
* **Phân bố lớp:** 500 (không tiểu đường) – 268 (tiểu đường)
* **Input:** Pregnancies, Glucose, Blood Pressure, SkinThickness, Insulin, BMI, Pedigree, Age
* **Output:** Class nhị phân (0 = không, 1 = có tiểu đường)
* **Thống kê nhanh (Mean ± SD):**
  + Glucose: 120.9 ± 32.0
  + BMI: 32.0 ± 7.9
  + Age: 33.2 ± 11.8
* **Mục tiêu:** Classification – dự đoán nguy cơ mắc tiểu đường