1. **Tóm tắt nhanh các paper (dựa trên trích xuất trang đầu + kiến thức phổ biến về Pima dataset)**

**paper1.pdf (preview):**

Là tài liệu liên quan đến tiêu chuẩn chẩn đoán / mô tả y tế (WHO style). Nội dung đầu nói về phân loại/định nghĩa các trạng thái rối loạn đường huyết, có thể dùng làm tham chiếu y khoa cho việc chuẩn hóa định nghĩa “diabetes” / “glucose intolerance”.

**paper2.pdf (preview):**

Tài liệu có đề cập “Using the ADAP Learning Algorithm to Forecast the Onset …” — tức là đề xuất/ứng dụng một thuật toán học máy (ADAP learning) để dự báo khởi phát bệnh (onset). So sánh với logistic regression và perceptron — cho thấy ý tưởng dùng các mô hình phân loại tuyến tính / mạng nơ-ron nhỏ cho bài toán này.

**paper3.pdf (preview):**

Về “Classification and Diagnosis of Diabetes Mellitus” — là bài tổng quan/phân loại và chẩn đoán, thảo luận các lớp glucose và điều trị. Có thể là review hoặc trình bày các phương pháp phân loại y khoa.

1. **Input/Output/Mục tiêu**

**Bối cảnh:**  
Dữ liệu Pima Indians Diabetes dùng để dự đoán xem một cá nhân có bị tiểu đường (diabetes) hay không dựa trên các chỉ số sức khỏe.

* **Input (Đầu vào)**: mỗi record (hàng) gồm các đặc trưng (features):

1. Pregnancies — số lần mang thai
2. Glucose — nồng độ glucose (mg/dL)
3. BloodPressure — huyết áp (mm Hg)
4. SkinThickness — độ dày nếp da (mm)
5. Insulin — insulin (mu U/ml)
6. BMI — chỉ số khối cơ thể (kg/m²)
7. DiabetesPedigreeFunction — hàm di truyền tiểu đường (dùng để biểu diễn tiền sử gia đình)
8. Age — tuổi (years)

* **Output (Đầu ra):**
* Outcome — 0 (không bị tiểu đường), 1 (bị tiểu đường). Đây là bài toán phân loại nhị phân.
* **Mục tiêu (Goal):**
* Xây dựng mô hình học máy phân loại (ví dụ: logistic regression, decision tree, random forest, SVM, neural network) để dự đoán Outcome với độ chính xác / AUC / F1 tốt.
* Trước khi training, cần hiểu và tiền xử lý dữ liệu (EDA), giải quyết giá trị thiếu, cân bằng lớp, chọn đặc trưng.

1. **Phân tích khám phá dữ liệu**

**Thông số chung**

* Kích thước dataset (sau đọc đúng header): 768 hàng × 9 cột.
* Phân bố output (Outcome):
  + 0 (không tiểu đường): 500 mẫu
  + 1 (có tiểu đường): 268 mẫu  
    => Có mild class imbalance (khoảng 65% không bệnh, 35% bệnh).

**Vấn đề missing values**

* Trong dataset, một số cột dùng giá trị 0 để biểu thị thiếu/không đo được, cụ thể:
  + Pregnancies: 111 giá trị = 0 (lưu ý: đây có thể đúng vì có người chưa mang thai)
  + Glucose: 5 giá trị = 0 (không hợp lý — đánh dấu missing)
  + BloodPressure: 35 giá trị = 0 (không hợp lý)
  + SkinThickness: 227 giá trị = 0 (không hợp lý)
  + Insulin: 374 giá trị = 0 (không hợp lý)
  + BMI: 11 giá trị = 0 (không hợp lý)

**Cách xử lý đã dùng:**

* Chuyển các giá trị 0 trong 5 cột (Glucose, BloodPressure, SkinThickness, Insulin, BMI) thành NaN.
* Impute bằng median của cột (chiến lược đơn giản, ít bị ảnh hưởng bởi outlier).

Các thống kê quan trọng (sau impute)

* Trung bình Glucose:
  + Outcome 0: ~110.68
  + Outcome 1: ~142.13  
    => Ngưỡng glucose trung bình cao hơn đáng kể ở nhóm positive.
* Trung bình BMI:
  + Outcome 0: ~30.89
  + Outcome 1: ~35.38
* Tuổi (Age) trung bình:
  + Outcome 0: ~31.19
  + Outcome 1: ~37.07

**Correlation với Outcome (liệt kê top)**

Số liệu correlation (giá trị tuyệt đối Pearson với Outcome), từ cao xuống thấp:

1. Glucose: 0.493
2. BMI: 0.312
3. Age: 0.238
4. Pregnancies: 0.222
5. SkinThickness, Insulin, DiabetesPedigreeFunction, BloodPressure lần lượt giảm dần.

Insight: Glucose là biến có liên quan mạnh nhất tới Outcome — hợp lý y học. BMI và tuổi cũng là yếu tố đáng chú ý.