**Dự báo khởi phát Đái tháo đường trên dữ liệu Pima theo bài báo “Using the ADAP Learning Algorithm to Forecast the Onset of Diabetes Mellitus”**

**1. Đọc paper để hiểu bài toán**

Bài báo kiểm tra khả năng của một mô hình mạng nơ-ron cổ điển tên ADAP để dự báo khởi phát đái tháo đường typ 2 trong vòng 5 năm ở cộng đồng người Pima gần Phoenix, Arizona. Tác giả đánh giá bằng độ nhạy, độ đặc hiệu và đường ROC; điểm giao nhau của độ nhạy và độ đặc hiệu đạt **0.76**. Nhóm tác giả cũng dự kiến so sánh cùng tập dữ liệu với **hồi quy logistic** và **perceptron tuyến tính** để đối chiếu kết quả.

Dữ liệu nghiên cứu được thu thập theo chu trình khám định kỳ 2 năm, có test dung nạp glucose đường uống (OGTT); chẩn đoán đái tháo đường dựa theo tiêu chí WHO: glucose huyết tương sau 2 giờ ≥ 200 mg/dL hoặc ghi nhận giá trị tương đương trong chăm sóc thường quy.

**2. Xác định ý nghĩa bài toán: Input - Output - Mục tiêu**

**Bối cảnh học máy:**

* *Đối tượng:* phụ nữ người Pima, chọn duy nhất 1 lần khám “chỉ mục” mỗi người, tuổi ≥ 21 tại thời điểm khám. Các ca được chọn sao cho: hoặc được chẩn đoán đái tháo đường trong 1-5 năm sau khám, hoặc vẫn không đái tháo đường ít nhất 5 năm sau. Các trường hợp phát bệnh trong < 1 năm sau khám bị loại để tránh ca “quá dễ” dự báo.
* *Quy mô:* 768 ca, chia ngẫu nhiên 576 train và 192 test.

**Đầu vào (Input features, 8 biến):**

1. Số lần mang thai
2. Glucose huyết tương 2 giờ sau OGTT
3. Huyết áp tâm trương
4. Độ dày nếp gấp da cánh tay (triceps)
5. Insulin huyết thanh sau 2 giờ
6. BMI
7. Chỉ số phả hệ đái tháo đường (DPF) - tổng hợp tiền sử gia đình và mức độ quan hệ huyết thống
8. Tuổi: Bài báo còn nêu công thức xây dựng DPF từ tình trạng bệnh của họ hàng, tuổi lúc mắc bệnh hoặc tuổi lần khám gần nhất còn không bệnh, và mức độ chia sẻ gen 0.5, 0.25, 0.125 tương ứng mức quan hệ.

**Đầu ra (Output):**

* Xác suất hoặc quyết định nhị phân: có/không phát bệnh đái tháo đường trong 5 năm kể từ thời điểm khám chỉ mục. Trong đánh giá cuối bài, nhóm dùng ngưỡng phân loại 0.448 để tối ưu cân bằng, đạt độ nhạy = độ đặc hiệu = 0.76 trên tập test và minh họa bằng ROC.

**Mục tiêu học máy:**

* Xây mô hình phân loại nhị phân nhằm dự báo sớm nguy cơ để sàng lọc và can thiệp. Mục tiêu định lượng: tối ưu đường ROC, lựa chọn ngưỡng theo mục tiêu lâm sàng - ví dụ ưu tiên độ nhạy nếu muốn giảm bỏ sót ca nguy cơ.

**3. Các nghiên cứu liên quan**

* Bài báo đặt ADAP cạnh các phương pháp kinh điển như hồi quy logistic và perceptron tuyến tính trên cùng tập train-test để so sánh ROC, cho thấy cách tư duy “đối chiếu mô hình” đã có từ sớm.
* Nguồn gốc tiêu chí chẩn đoán và quan tâm đến ROC được neo vào tài liệu WHO và y văn ROC cổ điển mà bài báo trích dẫn.
* Về dịch tễ Pima, bài báo cũng trích các nghiên cứu trước đây thể hiện tỷ lệ và yếu tố nguy cơ trong quần thể này, nhấn mạnh vai trò của béo phì và tiền sử gia đình - cơ sở để đưa DPF vào mô hình.

**4. Phân tích khám phá dữ liệu (EDA) - đề xuất chi tiết**

Vì paper nêu rõ biến và quy tắc chọn mẫu, EDA nên bám sát các điểm sau để hiểu dữ liệu trước khi huấn luyện:

**4.1 Kiểm tra cấu trúc và phân bố biến**

* Thống kê mô tả 8 biến đầu vào: trung vị, IQR, min-max; vẽ histogram cho các biến liên tục như glucose 2h, insulin 2h, BMI, tuổi; boxplot theo nhãn phát bệnh 5 năm. Mục tiêu: phát hiện lệch trái/phải, đuôi dài, ngoại lai. (Danh sách 8 biến trích từ paper.)
* Ma trận tương quan giữa các biến liên tục và cảnh báo đa cộng tuyến tiềm năng, ví dụ BMI với huyết áp, insulin với glucose 2h.
* Biến DPF: kiểm tra phân bố và mối liên hệ với nhãn; đọc lại định nghĩa để hiểu trực giác DPF tăng khi có nhiều thân nhân mắc sớm và quan hệ huyết thống gần.

**4.2 Chất lượng dữ liệu**

* Tỉ lệ thiếu vắng và quy tắc xử lý: insulin 2h và skinfold thường có tỷ lệ thiếu cao trong nhiều bản dữ liệu Pima; cân nhắc loại bỏ, nội suy hợp lý, hoặc dùng mô hình chịu thiếu.
* Kiểm tra giá trị bất thường - ví dụ giá trị 0 ở những biến sinh học không thể bằng 0 (huyết áp, insulin, skinfold). Với bản Pima phổ biến, 0 thường mã hóa “không đo được”. Nếu gặp, cần biến đổi về NA trước khi phân tích.

**4.3 Mục tiêu và sự sai lệch chọn mẫu**

* Nhãn mục tiêu là sự kiện phát bệnh trong 1-5 năm sau khám chỉ mục; các ca phát bệnh < 1 năm đã bị loại để tránh “dễ đoán”. Điều này ảnh hưởng phân bố xác suất nền và có thể làm mô hình ít gặp những ca “gần phát bệnh”, do đó cần ghi nhận khi diễn giải.
* Tỷ lệ dương - âm: tính tỉ lệ phát bệnh 5 năm trong tập dữ liệu, đối chiếu với tỷ lệ của cộng đồng nếu có, để phát hiện lệch chọn mẫu.

**4.4 Khám phá quan hệ Input - Output**

* Phân rã theo nhãn: so sánh phân bố glucose 2h, insulin 2h, BMI, tuổi giữa nhóm phát bệnh và không phát bệnh.
* Đường cong ROC sơ bộ với 1 biến: kiểm tra từng biến đơn lẻ, đặc biệt glucose 2h vì đây cũng là tiêu chí chẩn đoán WHO; lưu ý rủi ro “gần với tiêu chí gán nhãn” có thể làm biến này đóng vai trò mạnh một cách hiển nhiên. Paper xác nhận tiêu chí WHO đúng là ngưỡng 200 mg/dL sau 2h OGTT.

**4.5 Thiết lập baseline và ngưỡng phân loại**

* Baseline mô hình tuyến tính: logistic regression với 8 biến chuẩn hóa, thêm chuẩn hóa robust nếu nhiều ngoại lai.
* Đối chiếu với ADAP như trong paper: mô hình ADAP kích hoạt khoảng 8% đơn vị liên kết với ngưỡng kích hoạt 4 kết nối, dùng 100,000 “association units” - đây là đặc thù của ADAP và giải thích vì sao ADAP có khả năng nắm bắt tương tác.
* Lựa chọn ngưỡng: nếu mục tiêu sàng lọc, có thể ưu tiên độ nhạy hơn, chấp nhận tăng dương tính giả. Paper dùng ngưỡng 0.448 để cân bằng 0.76-0.76, có thể là mốc tham khảo ban đầu trước khi hiệu chỉnh theo chi phí lâm sàng.

**4.6 Diễn giải và rủi ro rò rỉ khái niệm**

* Vì nhãn dựa trên chuẩn WHO về glucose 2h ở các lần khám sau, còn biến glucose 2h tại thời điểm chỉ mục nằm trong Input, cần cẩn trọng khi suy diễn - đây không phải rò rỉ dữ liệu theo nghĩa dùng trực tiếp nhãn tương lai, nhưng là biến có liên hệ sinh lý - chẩn đoán rất mạnh với nhãn sau này. Nên báo cáo thêm kịch bản bỏ biến glucose 2h để đánh giá giá trị sàng lọc khi thiếu OGTT - cách mà nhiều nhóm thực hành khi triển khai trong thực tế sàng lọc ban đầu. (Tiêu chí WHO dẫn chiếu trong paper.)

**5. Mô tả ngắn gọn về ADAP trong paper**

* ADAP chia miền giá trị mỗi biến thành các “phân vùng - sensors”, kết nối ngẫu nhiên đến nhiều association units; một association unit “bật” nếu số kết nối khớp vượt ngưỡng; dự báo là tổng trọng số các đơn vị được kích hoạt. Trong nghiên cứu này, ADAP dùng 100,000 association units, điều chỉnh theo quy tắc giống Hebb, ngưỡng kích hoạt cấu hình sao cho 2-10% đơn vị được bật trên mỗi ca.
* Quá trình học: sau mỗi ca, ADAP phân bổ sai số lên các đơn vị đã kích hoạt; khi suy diễn thì không cập nhật nữa.

**6. Kết luận**

* Bài toán là phân loại nhị phân dự báo 5 năm, với 8 biến lâm sàng - sinh hóa và một đặc trưng phả hệ tổng hợp (DPF). Đặt trong ngữ cảnh sàng lọc, cần báo cáo đầy đủ ROC, AUC, độ nhạy, độ đặc hiệu theo các ngưỡng khác nhau, và cân nhắc chi phí sai số. Paper cho thấy ở ngưỡng 0.448, mô hình đạt 0.76 - 0.76 trên tập test.
* Khi triển khai EDA và huấn luyện, nên chạy 2 kịch bản: dùng đủ 8 biến và kịch bản loại glucose 2h để đánh giá giá trị sàng lọc khi không có OGTT.
* So sánh baseline hiện đại như logistic regression, cây tăng cường, mạng nơ-ron MLP để đối chiếu với ADAP theo đúng tinh thần bài báo.