Họ và tên: Lê Khánh Hoàng

MSSV: 3122410125

**Lab04: Phân tích khám phá dữ liệu**

**1. Đọc paper và khảo sát bài toán liên quan:**

Trong ba tài liệu được cung cấp, nội dung đều xoay quanh việc chuẩn hóa tiêu chuẩn chẩn đoán bệnh tiểu đường và thử nghiệm các phương pháp phân tích dữ liệu nhằm hỗ trợ dự đoán bệnh.

National Diabetes Data Group [1] đã đưa ra các tiêu chuẩn chẩn đoán bệnh tiểu đường dựa trên các chỉ số lâm sàng như nồng độ glucose huyết tương lúc đói, mức glucose sau nghiệm pháp dung nạp glucose. Đây là một mốc quan trọng, vì nó thống nhất cách xác định bệnh trong cộng đồng nghiên cứu và y tế, tạo tiền đề cho các công trình ứng dụng dữ liệu sau này.

Smith và cộng sự [2] đã giới thiệu *Adaptive Learning Algorithm (ADAP)* và áp dụng vào bộ dữ liệu *Pima Indians Diabetes* để dự đoán khả năng mắc bệnh. Nghiên cứu này là một trong những công trình tiên phong cho thấy tiềm năng của mạng nơ-ron nhân tạo trong việc hỗ trợ chẩn đoán y khoa. Thay vì chỉ dựa vào quy tắc lâm sàng, mô hình học máy có thể học được các quan hệ phức tạp trong dữ liệu.

WHO Expert Committee [3] cũng đưa ra khuyến nghị quốc tế về việc chuẩn hóa tiêu chuẩn chẩn đoán và tiền tiểu đường. Điều này khẳng định tầm quan trọng của dữ liệu y tế và sự cần thiết phải phát hiện sớm bệnh để có biện pháp can thiệp kịp thời.

**Kết luận khảo sát:**

Từ ba tài liệu, có thể thấy rằng bệnh tiểu đường là một thách thức y tế lớn, đòi hỏi sự kết hợp giữa tiêu chuẩn y khoa và các phương pháp phân tích dữ liệu tiên tiến. Bộ dữ liệu Pima Indians Diabetes trở thành một ví dụ điển hình và thường xuyên được sử dụng để thử nghiệm các thuật toán học máy. Đây cũng chính là cơ sở để tiến hành các bước phân tích dữ liệu trong nghiên cứu này.

**Liên hệ với yêu cầu bài tập:**

Qua khảo sát ba tài liệu [1–3], có thể nhận thấy xu hướng chung của cộng đồng khoa học là chuẩn hóa tiêu chuẩn y khoa kết hợp với ứng dụng phương pháp tính toán để chẩn đoán và dự đoán bệnh tiểu đường.

Trong đó, bộ dữ liệu Pima Indians Diabetes đóng vai trò quan trọng vì được sử dụng như một tập dữ liệu chuẩn trong nhiều nghiên cứu, đặc biệt là trong công trình của Smith và cộng sự [2]. Việc sử dụng tập dữ liệu này giúp đánh giá khả năng áp dụng của các mô hình học máy vào một vấn đề y tế có ý nghĩa thực tiễn.

Chính vì vậy, việc thực hiện phân tích khám phá dữ liệu (Exploratory Data Analysis – EDA) trên bộ dữ liệu Pima không chỉ giúp sinh viên rèn luyện kỹ năng xử lý và đánh giá dữ liệu, mà còn kết nối trực tiếp với các hướng nghiên cứu đã được trình bày trong các paper. Đây là bước tiền đề quan trọng trước khi xây dựng bất kỳ mô hình dự đoán nào, bởi chất lượng dữ liệu sẽ ảnh hưởng trực tiếp đến độ tin cậy của kết quả phân tích.

**2. Phân tích khám phá dữ liệu (Exploratory Data Analysis – EDA)**

**2.1. Xác định tính chất dữ liệu:**

Bộ dữ liệu Pima Indians Diabetes gồm 768 mẫu và 9 thuộc tính. Các thuộc tính bao gồm: số lần mang thai, nồng độ glucose, huyết áp, độ dày nếp gấp da, chỉ số insulin, chỉ số khối cơ thể (BMI), chỉ số nguy cơ di truyền (DPF), tuổi và biến mục tiêu (Outcome: 0 = không mắc tiểu đường, 1 = mắc tiểu đường).

Kết quả thống kê mô tả cho thấy dữ liệu có giá trị trung bình hợp lý đối với các chỉ số y tế, tuy nhiên tồn tại sự biến thiên khá lớn ở một số biến như Insulin và SkinThickness.

**2.2. Phân tích đơn biến:**

Khi quan sát phân bố của từng biến:

* Glucose có phân bố lệch phải, một số giá trị bằng 0 bất thường (không hợp lý về mặt y học).
* BloodPressure và SkinThickness cũng có nhiều giá trị bằng 0, cho thấy dữ liệu thiếu được biểu diễn dưới dạng số 0.
* Insulin có phân bố rất không đều, với nhiều giá trị bằng 0.
* BMI tập trung trong khoảng 20–40, nhưng vẫn có một số giá trị ngoại lệ rất cao.
* Age phân bố chủ yếu trong khoảng 20–50 tuổi, nhưng vẫn tồn tại một số trường hợp lớn hơn 80 tuổi.

Nhìn chung, dữ liệu chưa được làm sạch hoàn toàn và cần xử lý các giá trị bất hợp lý trước khi phân tích sâu hơn.

**2.3. Phân tích đa biến:**

Phân tích ma trận tương quan cho thấy:

* Glucose và BMI có mối liên hệ đáng kể với biến mục tiêu Outcome, gợi ý rằng đây là các yếu tố quan trọng để dự đoán khả năng mắc tiểu đường.
* Age cũng có ảnh hưởng nhất định, khi người lớn tuổi có nguy cơ cao hơn.
* Một số cặp biến khác như BloodPressure và SkinThickness có tương quan thấp, thể hiện tính độc lập tương đối.

Biểu đồ phân tán (scatter plot) giữa BMI và Glucose theo Outcome cho thấy người mắc tiểu đường thường có chỉ số BMI và mức glucose cao hơn so với người không mắc bệnh.

**2.4. Xác định giá trị thiếu và mẫu bất thường:**

Trong dữ liệu, nhiều biến quan trọng có số lượng giá trị bằng 0 cao bất thường:

* Glucose: 5 trường hợp.
* BloodPressure: 35 trường hợp.
* SkinThickness: 227 trường hợp.
* Insulin: 374 trường hợp.
* BMI: 11 trường hợp.

Các giá trị này không hợp lý về mặt sinh học, do đó có thể coi là dữ liệu thiếu và cần được xử lý (bằng cách loại bỏ, thay thế bằng giá trị trung bình/trung vị, hoặc áp dụng phương pháp suy đoán giá trị).

**2.5. Xác định ngoại lệ (Outliers):**

Thông qua biểu đồ boxplot, có thể nhận thấy:

* Một số giá trị cực đoan trong các biến Glucose, Insulin và BMI.
* Đặc biệt, biến Insulin có nhiều điểm nằm ngoài khoảng giá trị thông thường, cho thấy phân bố rất lệch.
* Biến Age cũng có một vài ngoại lệ ở mức trên 80 tuổi.

Việc phát hiện ngoại lệ là quan trọng, vì chúng có thể ảnh hưởng lớn đến mô hình dự đoán sau này.

**2.6. Kết luận sơ bộ:**

Từ phân tích khám phá dữ liệu, có thể rút ra một số nhận định chính:

* Bộ dữ liệu chứa nhiều giá trị 0 bất hợp lý, cần xử lý trước khi huấn luyện mô hình.
* Các biến Glucose, BMI và Age có vai trò quan trọng trong việc phân biệt nhóm mắc tiểu đường và nhóm không mắc.
* Dữ liệu có tồn tại các ngoại lệ, đặc biệt ở biến Insulin, cần cân nhắc loại bỏ hoặc chuẩn hóa.
* Bước EDA cho thấy dữ liệu mang tính thực tế cao và có thể được sử dụng làm cơ sở để xây dựng mô hình dự đoán, nhưng bắt buộc phải tiến hành tiền xử lý dữ liệu cẩn thận.

**3. Tài liệu tham khảo**

[1] National Diabetes Data Group, *Classification and Diagnosis of Diabetes Mellitus and Other Categories of Glucose Intolerance*, Diabetes, 28(12), 1039–1057, 1979.

[2] J. W. Smith, J. E. Everhart, W. C. Dickson, W. C. Knowler, R. S. Johannes, *Using the ADAP Learning Algorithm to Forecast the Onset of Diabetes Mellitus*, Proceedings of the Annual Symposium on Computer Application in Medical Care, 1988.

[3] World Health Organization, *WHO Expert Committee on Diabetes Mellitus: Second Report*, Technical Report Series 646, Geneva, 1980.