BÁO CÁO PHÂN TÍCH BỘ DỮ LIỆU PIMA INDIAN DIABETES

Báo cáo phân tích quy trình tiền xử lý dữ liệu y tế sử dụng bộ dữ liệu Pima Indian Diabetes

Họ và tên người thực hiện: Trần Quốc Hoàng
Học phần: Khai phá dữ liệu

Ngày thực hiện: 29/09/2025

Tổng quan và mục tiêu phân tích

- **Mục tiêu:** Dự đoán khả năng một người Pima Indian (≥ 21 tuổi) bị tiểu đường dựa trên các chỉ số sức khỏe.
- **Bài toán phân loại nhị phân:** Có/Không mắc tiểu đường.
- **Ý nghĩa thực tiễn:** Hỗ trợ phát hiện sớm, ra quyết định lâm sàng.

Tại sao quan trọng?

Bộ dữ liệu Pima Indian Diabetes cung cấp cơ hội phân tích mối liên hệ giữa các yếu tố nguy cơ và bệnh tiểu đường, từ đó phát triển các công cụ dự đoán giúp can thiệp sớm và hiệu quả.

Mô tả bộ dữ liệu

- **Nguồn:** Pima Indians Diabetes (UCI Machine Learing Repository).
- **Kích thước:** 768 dòng, 9 cột.
- **Biến đầu vào:**

Tên biến	Mô tả
Pregnancies	Số lần mang thai
Glucose	Nồng độ glucose trong huyết tương sau 2h làm nghiệm pháp OGTT
BloodPressure	Huyết áp tâm trương (mm Hg)
SkinThickness	Độ dày nếp gấp da tam đầu (mm)
Insulin	Insulin huyết thanh 2 giờ (mu U/ml)
BMI	Chỉ số khối cơ thể (kg/(m^2))
DiabetesPedigreeFunction	Chức năng phả hệ tiểu đường (chỉ số di truyền)
Age	Tuổi (năm)

- **Biến mục tiêu:** Outcome (0: không tiểu đường, 1: có tiểu đường).
- **Dạng dữ liệu:** Numeric, không có missing value gốc nhưng có nhiều giá trị 0 bất hợp lý.

Phương pháp phân tích và công cụ sử dụng

- **Ngôn ngữ & Thư viện:** Python, pandas, numpy, matplotlib, seaborn, scikit-learn
- **Các bước chính:**

1. Đọc & khám phá dữ liệu

pd.read_csv, .describe(), .info(), value_counts()

2. Làm sạch & tiền xử lý giá trị thiếu

drop_duplicates(), replace(0, np.nan), SimpleImputer(strategy='median')

3. Chuẩn hóa dữ liệu

StandardScaler() - chuẩn hóa các biến đầu vào

4. Trực quan hóa & xuất báo cáo kết quả

sns.heatmap(), .hist(), boxplot, scatterplot, to_csv()

Tại sao sử dụng quy trình này?

Tiền xử lý dữ liệu chất lượng cao quyết định hiệu quả của mô hình dự đoán. Với dữ liệu y tế, việc xử lý giá trị thiếu và chuẩn hóa là cực kỳ quan trọng để đảm bảo mô hình có khả năng phát hiện các trường hợp mắc tiểu đường một cách chính xác.

Thống kê mô tả và khám phá dữ liệu

- **Phương pháp:** describe(), value_counts(), info() để kiểm tra tổng quan
- **Phân phối Outcome:**

- 0: 500 (65.1%) – Không tiểu đường
- 1: 268 (34.9%) – Tiểu đường

- **Thống kê các chỉ số quan trọng:**

- Glucose (trung bình): 120.89 mg/dL
- BMI (trung bình): 31.99 kg/m²

- **Ma trận tương quan các chỉ số quan trọng:**

Tương quan cao với Outcome			
Glucose (0.47)	BMI (0.29)	Age (0.24)	Insulin (0.22)

- **Phân phối đặc trưng:** Nhiều biến có phân phối lệch phải (skewed)

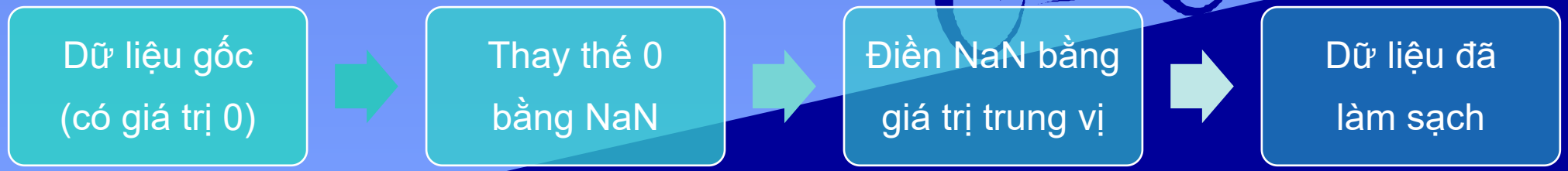
Insight: Glucose và BMI có tương quan cao nhất với nguy cơ mắc tiểu đường

Vấn đề chất lượng dữ liệu và cách xử lý

- **Phát hiện vấn đề:** Nhiều giá trị 0 bất hợp lý ở các cột chỉ số sinh học (không thể bằng 0 với các thông số như Glucose, BMI).
- **Xác định các giá trị cần xử lý:** Thay thế giá trị 0 bằng NaN ở 5 cột có giá trị sinh học không thể bằng 0.
- **Điền giá trị thiếu:** Sử dụng SimpleImputer với strategy='median' để thay thế các NaN bằng giá trị trung vị.
- **Kết quả:** Không còn missing values sau khi xử lý, dữ liệu sạch hơn và thực tế hơn.

Số lượng giá trị 0 bất hợp lý được xử lý:

Biến	Số giá trị 0	Ý nghĩa
Glucose	5	Nồng độ glucose không thể bằng 0 ở người sống
BloodPressure	35	Huyết áp 0 mmHg không thể tồn tại
SkinThickness	227	Độ dày da 0 mm không khả thi
Insulin	374	Số lượng giá trị 0 lớn nhất, cần điều tra kỹ
BMI	11	BMI bằng 0 là không thể



Phân tích trực quan dữ liệu

Heatmap ma trận tương quan

Hiển thị mối liên hệ giữa các biến và với Outcome

Phát hiện: Glucose và BMI có tương quan cao nhất với Outcome

Histogram Phân Bố

Trực quan phân phối của mỗi biến số trong dataset

Phát hiện: Một số biến như Insulin, Glucose có phân phối lệch phải

Boxplot Glucose vs Outcome

So sánh phân bố Glucose giữa 2 nhóm

Phát hiện: Nhóm tiểu đường có trung vị Glucose cao hơn đáng kể

Scatterplot BMI vs BP

BMI vs BloodPressure phân theo nhóm Outcome

Phát hiện: Nhóm tiểu đường thường có BMI và huyết áp cao hơn

Ý nghĩa phân tích trực quan:

Các biểu đồ giúp phát hiện các mối tương quan và đặc điểm phân biệt giữa nhóm có và không có tiểu đường, cung cấp insight cho việc xây dựng mô hình phân loại hiệu quả sau này.

Kết quả tiền xử lý dữ liệu

- **Đã xóa dòng trùng lặp:** Không phát hiện dòng trùng lặp trong bộ dữ liệu gốc.
- **Xử lý giá trị 0 bất hợp lý:** Chuyển thành NaN và điền đầy đủ bằng giá trị trung vị (median).
- **Chuẩn hoá dữ liệu:** Áp dụng StandardScaler để chuẩn hoá các biến đầu vào về Z-score.
- **Xuất kết quả:** Lưu dữ liệu đã xử lý thành file pima-indians-diabetes-processed.csv
- **Kết quả:** Dữ liệu sạch và chuẩn hóa, sẵn sàng cho việc xây dựng mô hình phân loại.

Ý nghĩa của việc chuẩn hóa

Chuẩn hóa dữ liệu giúp các thuật toán học máy hoạt động hiệu quả hơn, đặc biệt với các thuật toán nhạy cảm với sự khác biệt về thang đo giữa các biến như SVM, KNN và Neural Networks.

Kết quả và Insight chính

- **Làm sạch & chuẩn hóa dữ liệu** là bước cực kỳ quan trọng với dữ liệu y tế, đặc biệt khi có nhiều giá trị 0 bất hợp lý.
- **Tồn tại sự mất cân bằng nhãn Outcome** (65.1% không tiểu đường, 34.9% tiểu đường) - cần lưu ý khi xây dựng mô hình phân loại.
- **Một số biến đầu vào (Glucose, BMI)** có sự phân biệt rõ ràng giữa hai nhóm có/không tiểu đường - tiềm năng làm đặc trưng phân loại tốt.
- **Dataset đã sẵn sàng** cho các mô hình phân loại tiểu đường tiếp theo sau quá trình tiền xử lý toàn diện.

Insight Quan Trọng

Việc phát hiện và xử lý các giá trị 0 bất hợp lý trong các chỉ số y tế (Glucose, BloodPressure, SkinThickness, Insulin, BMI) là rất quan trọng trong tiền xử lý dữ liệu cho các bài toán phân loại y tế, giúp đảm bảo tính chính xác và hiệu quả của mô hình.

Khuyến nghị và bước tiếp theo

- **Xử lý mất cân bằng nhãn:** Áp dụng kỹ thuật SMOTE, class weights hoặc cân bằng lại tỷ lệ mẫu để cải thiện khả năng phân loại.
- **Mô hình phân loại:** Khám phá và so sánh hiệu năng các mô hình như Random Forest, SVM, Neural Networks, Logistic Regression và XGBoost.
- **Feature engineering:** Nghiên cứu thêm về tạo biến mới, lựa chọn thuộc tính quan trọng và tinh chỉnh hyperparameter.
- **Ứng dụng thực tiễn:** Phát triển công cụ hỗ trợ quyết định lâm sàng để phát hiện sớm nguy cơ tiểu đường cho người dân Pima Indian.

Lộ trình phát triển mô hình			
1 Tiền xử lý dữ liệu	2 Mô hình hóa	3 Đánh giá	4 Triển khai