BÁO CÁO KẾT QUẢ BÀI TẬP

Phân tích dữ liệu về đái tháo đường ở người Ấn Độ

Giảng viên hướng dẫn: TS. Đỗ Như Tài

Sinh viên thực hiện: Huỳnh Nhật Minh

Đại học Sài Gòn

huynhnhatminh20102005@gmail.com

273 An Dương Vương

Ngày 29 tháng 9 năm 2025

Presentation Overview

1. Định nghĩa vấn đề
   - Mô tả bài toán và thông tin dữ liệu
2. Phân tích dữ liệu
   - Thống kê mô tả
   - Hiển thị dữ liệu
3. Chuẩn bị dữ liệu
   - Làm sạch và Biến đổi dữ liệu
   - Chuẩn hóa dữ liệu
   - Chia dữ liệu thực nghiệm

Mô tả dữ liệu

Mô tả: Bộ dữ liệu Pima Indians Diabetes gồm 768 bản ghi của phụ nữ gốc Pima từ 21 tuổi trở lên.
Mỗi bản ghi chứa 8 thông số y tế liên quan đến nguy cơ mắc bệnh tiểu đường, cùng với nhãn chẩn đoán cho biết người đó có mắc bệnh hay không.
Dữ liệu vào:

- Pregnancies – Số lần mang thai
- Glucose – Nồng độ glucose trong huyết tương (mg/dL)
- BloodPressure – Huyết áp tâm trương (mmHg)
- SkinThickness – Độ dày nếp gấp da (mm)
- Insulin – Nồng độ insulin trong huyết thanh (μU/mL)
- BMI – Chỉ số khối cơ thể (kg/m^2)
- DiabetesPedigreeFunction – Hệ số di truyền tiểu đường
- Age – Tuổi (năm)

Kết quả:

- Outcome – 0: Không mắc bệnh tiểu đường, 1: Mắc bệnh tiểu đường

Thông tin dữ liệu

Nhận xét:

- Dữ liệu có 8 thuộc tính đầu vào: Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, Age.
- Các thuộc tính này là các chỉ số y tế, đơn vị đo tùy từng loại (mg/dL, mmHg, chỉ số BMI, ...).
- Tổng số dòng dữ liệu: 768 dòng.
- Thuộc tính mục tiêu (nhãn phân lớp): cột Outcome
  - 0 = Không mắc tiểu đường
  - 1 = Mắc tiểu đường

Thông tin dữ liệu

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
763	10	101	76	48	180	32.9	0.171	63	0
764	2	122	70	27	0	36.8	0.340	27	0
765	5	121	72	23	112	26.2	0.245	30	0
766	1	126	60	0	0	30.1	0.349	47	1
767	1	93	70	31	0	30.4	0.315	23	0

Kiểm tra tính toàn vẹn dữ liệu

Nhận xét:

- Dữ liệu không có giá trị rỗng (Null, NaN).
- Số dòng bị trùng bằng 0 $\Rightarrow$ không tồn tại bản ghi trùng lặp trong tập Pima Indians Diabetes.

Thống kê mô tả dữ liệu

	count	mean	std	min	25%	50%	75%	max
Pregnancies	768.0	3.845052	3.369578	0.000	1.00000	3.0000	6.00000	17.00
Glucose	768.0	120.894531	31.972618	0.000	99.00000	117.0000	140.25000	199.00
BloodPressure	768.0	69.105469	19.355807	0.000	62.00000	72.0000	80.00000	122.00
SkinThickness	768.0	20.536458	15.952218	0.000	0.00000	23.0000	32.00000	99.00
Insulin	768.0	79.799479	115.244002	0.000	0.00000	30.5000	127.25000	846.00
BMI	768.0	31.992578	7.884160	0.000	27.30000	32.0000	36.60000	67.10
DiabetesPedigreeFunction	768.0	0.471876	0.331329	0.078	0.24375	0.3725	0.62625	2.42
Age	768.0	33.240885	11.760232	21.000	24.00000	29.0000	41.00000	81.00
Outcome	768.0	0.348958	0.476951	0.000	0.00000	0.0000	1.00000	1.00

Thống kê mô tả dữ liệu

Bộ dữ liệu gồm 8 thuộc tính đầu vào với đặc điểm và khoảng giá trị:

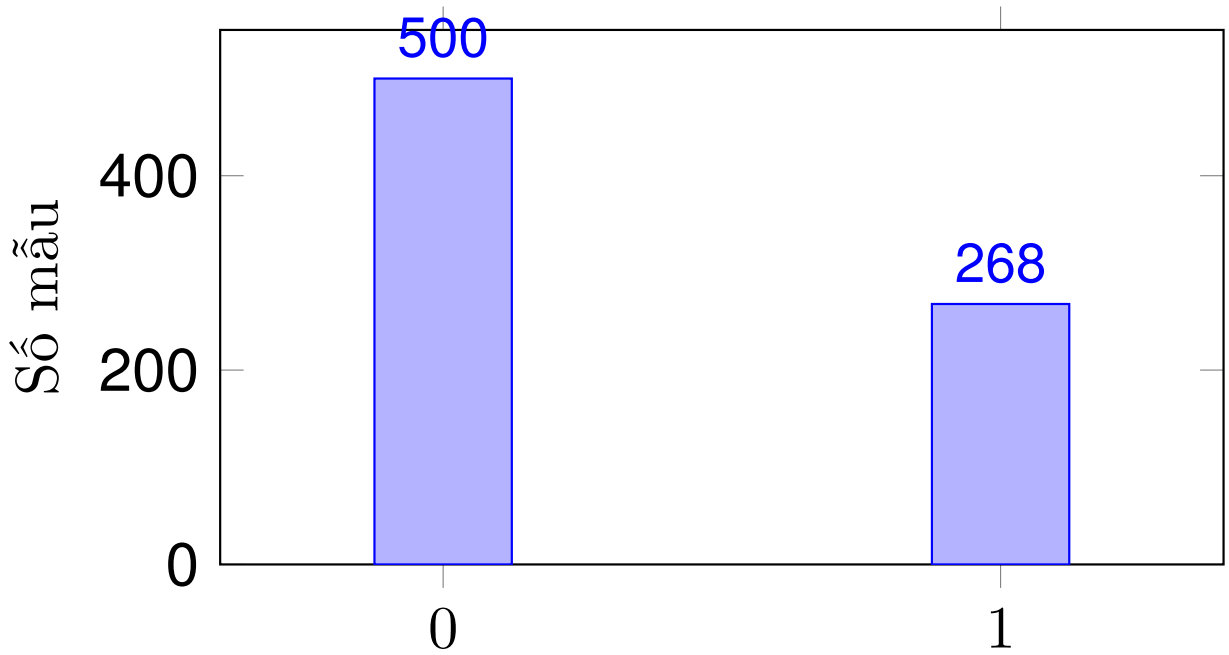
- Pregnancies: 0 – 17 (số lần mang thai)
- Glucose: 0 – 199 mg/dL (đường huyết)
- BloodPressure: 0 – 122 mmHg (huyết áp tâm trương)
- SkinThickness: 0 – 99 mm (độ dày da gấp)
- Insulin: 0 – 846 μU/mL (nồng độ insulin)
- BMI: 0 – 67.1 (chỉ số khối cơ thể)
- DiabetesPedigreeFunction: 0.078 – 2.42 (hệ số di truyền)
- Age: 21 – 81 (tuổi)

Thuộc tính mục tiêu: Outcome

- 0 = Không mắc tiểu đường
- 1 = Mắc tiểu đường

Phân bố dữ liệu (Distribution)

- Dữ liệu cần phân loại thành 2 lớp kết quả:
  - 0 = Không mắc tiểu đường
  - 1 = Mắc tiểu đường
- Các lớp không cân bằng:
  - Lớp 0: 500 mẫu
  - Lớp 1: 268 mẫu



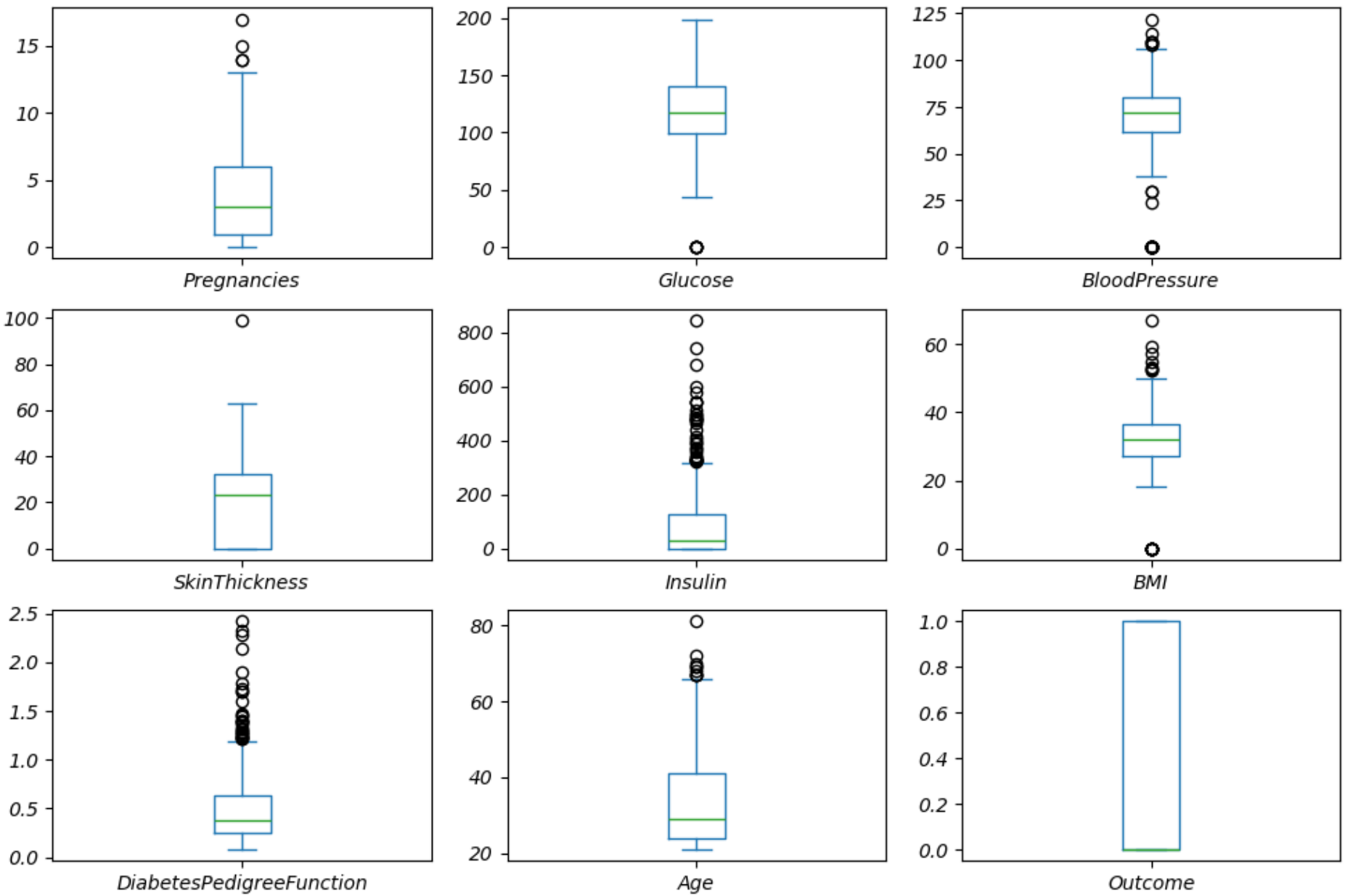
Mối tương quan giữa các thuộc tính

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
Pregnancies	1.000000	0.129459	0.141282	-0.081672	-0.073535	0.017683	-0.033523	0.544341	0.221898
Glucose	0.129459	1.000000	0.152590	0.057328	0.331357	0.221071	0.137337	0.263514	0.466581
BloodPressure	0.141282	0.152590	1.000000	0.207371	0.088933	0.281805	0.041265	0.239528	0.065068
SkinThickness	-0.081672	0.057328	0.207371	1.000000	0.436783	0.392573	0.183928	-0.113970	0.074752
Insulin	-0.073535	0.331357	0.088933	0.436783	1.000000	0.197859	0.185071	-0.042163	0.130548
BMI	0.017683	0.221071	0.281805	0.392573	0.197859	1.000000	0.140647	0.036242	0.292695
DiabetesPedigreeFunction	-0.033523	0.137337	0.041265	0.183928	0.185071	0.140647	1.000000	0.033561	0.173844
Age	0.544341	0.263514	0.239528	-0.113970	-0.042163	0.036242	0.033561	1.000000	0.238356
Outcome	0.221898	0.466581	0.065068	0.074752	0.130548	0.292695	0.173844	0.238356	1.000000

Mối tương quan giữa các thuộc tính

- Các cặp thuộc tính có độ tương quan Pearson cao trong bộ Pima Indians Diabetes:
  - (SkinThickness, Insulin) = 0.437
  - (BMI, SkinThickness) = 0.393
  - (Insulin, Glucose) = 0.331
  - (BMI, BloodPressure) = 0.282
- Biến mục tiêu Outcome tương quan mạnh nhất với:
  - Glucose = 0.467
  - BMI = 0.293
  - Age = 0.238

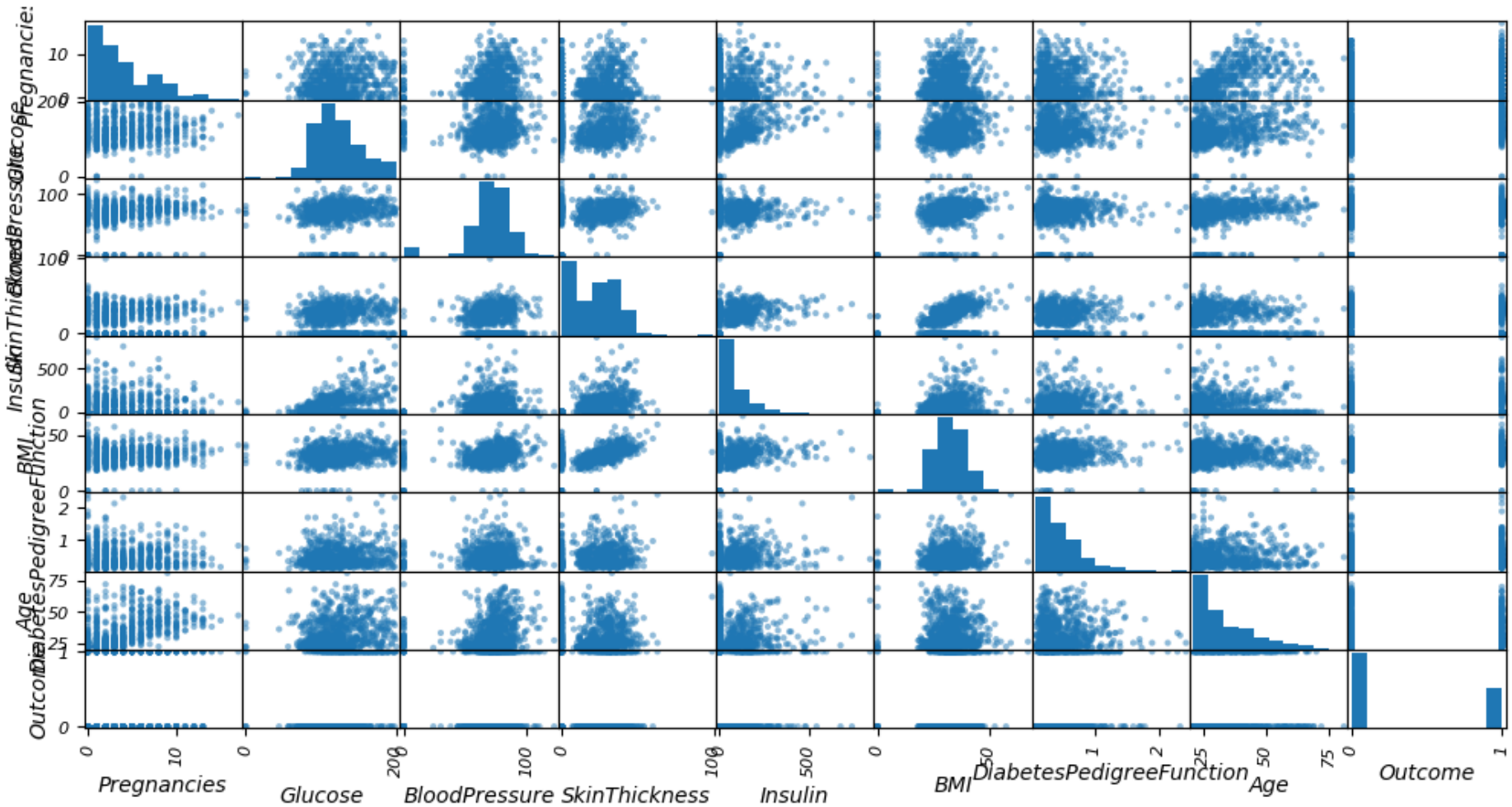
Hiển thị trên từng tính chất đơn



Hiển thị trên từng tính chất đơn

- Độ trải rộng:
  - Các biến số có phạm vi rất khác nhau.
  - Glucose, Insulin có phạm vi giá trị lớn hơn nhiều so với Outcome hoặc DiabetesPedigreeFunction.
- Độ lệch:
  - Glucose, Insulin, BMI lệch về phía giá trị lớn, nhiều outliers phía trên.
  - Pregnancies, SkinThickness lệch về phía giá trị nhỏ.
- Phân bố:
  - BloodPressure, Age có phân bố cân bằng, ít lệch rõ rệt.
  - Outcome, DiabetesPedigreeFunction phạm vi hẹp, phân bố không đều, tập trung nhiều ở mức thấp.

Hiển thị nhiều tính chất (Multivariate Plots)



Hiển thị nhiều tính chất (Multivariate Plots)

- Các cặp tính chất có độ tương đồng cao:
  - (Pregnancies, Age) = 0.544
  - (Glucose, Insulin) = 0.331
  - (SkinThickness, Insulin) = 0.437
  - (SkinThickness, BMI) = 0.393

Làm sạch dữ liệu (Data Cleaning)

Các bước thực hiện:

- Tạo bảng dữ liệu làm sạch: chuẩn hóa cấu trúc dữ liệu, giữ nguyên 768 bản ghi hợp lệ.
- Xóa dữ liệu trùng nhau: kiểm tra trùng lặp $\Rightarrow$ không có bản ghi trùng.
- Xử lý giá trị rỗng / không hợp lệ:
  - Không tồn tại giá trị Null hoặc NaN.
  - Các giá trị bất thường (0 ở Glucose, BloodPressure, BMI, Insulin) được ghi nhận để xử lý ở bước tiền xử lý tiếp theo.

Kết quả: Bộ dữ liệu sau làm sạch đảm bảo không trùng lặp, không Null, không NaN.

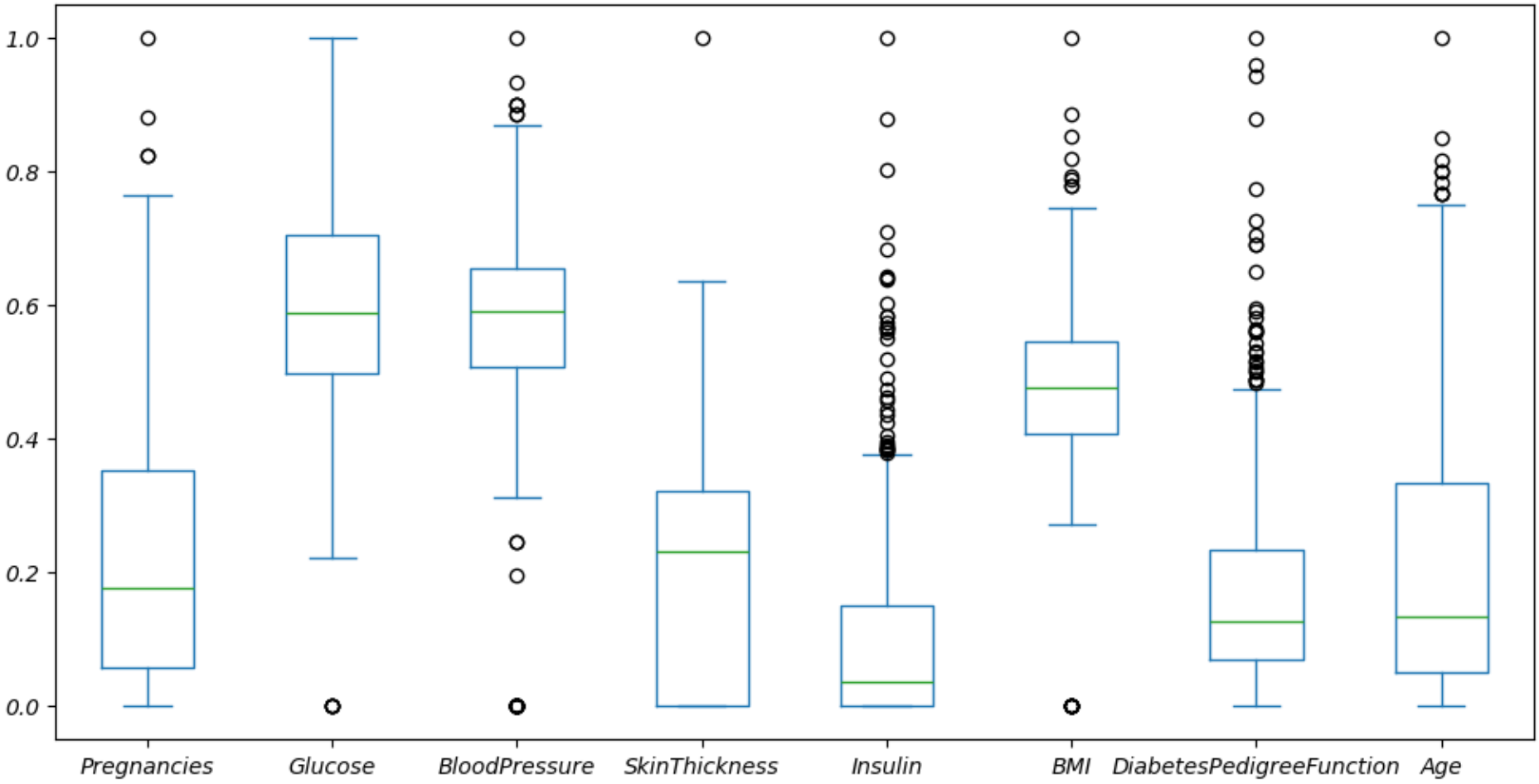
Biến đổi dữ liệu (Data Transformation)

Chuyển đổi dữ liệu danh mục (Category) thành dữ liệu số:

- Trong bộ dữ liệu Pima Indians Diabetes:
  - Cột Outcome đã ở dạng số
    - 0 = Không mắc tiểu đường
    - 1 = Mắc tiểu đường
  - $\Rightarrow$ Không cần sử dụng LabelEncoder như ví dụ Iris.
- Tất cả các cột còn lại (Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, Age) đều là dữ liệu số.
- $\Rightarrow$ Không cần chuyển đổi dữ liệu danh mục sang dạng One-Hot.

Lưu ý: Một số thuật toán và hàm mất mát CategoryEntropy yêu cầu dữ liệu phân lớp dạng One-Hot. Tuy nhiên, với bộ dữ liệu này, bước biến đổi không cần thiết.

Min-Max Normalization

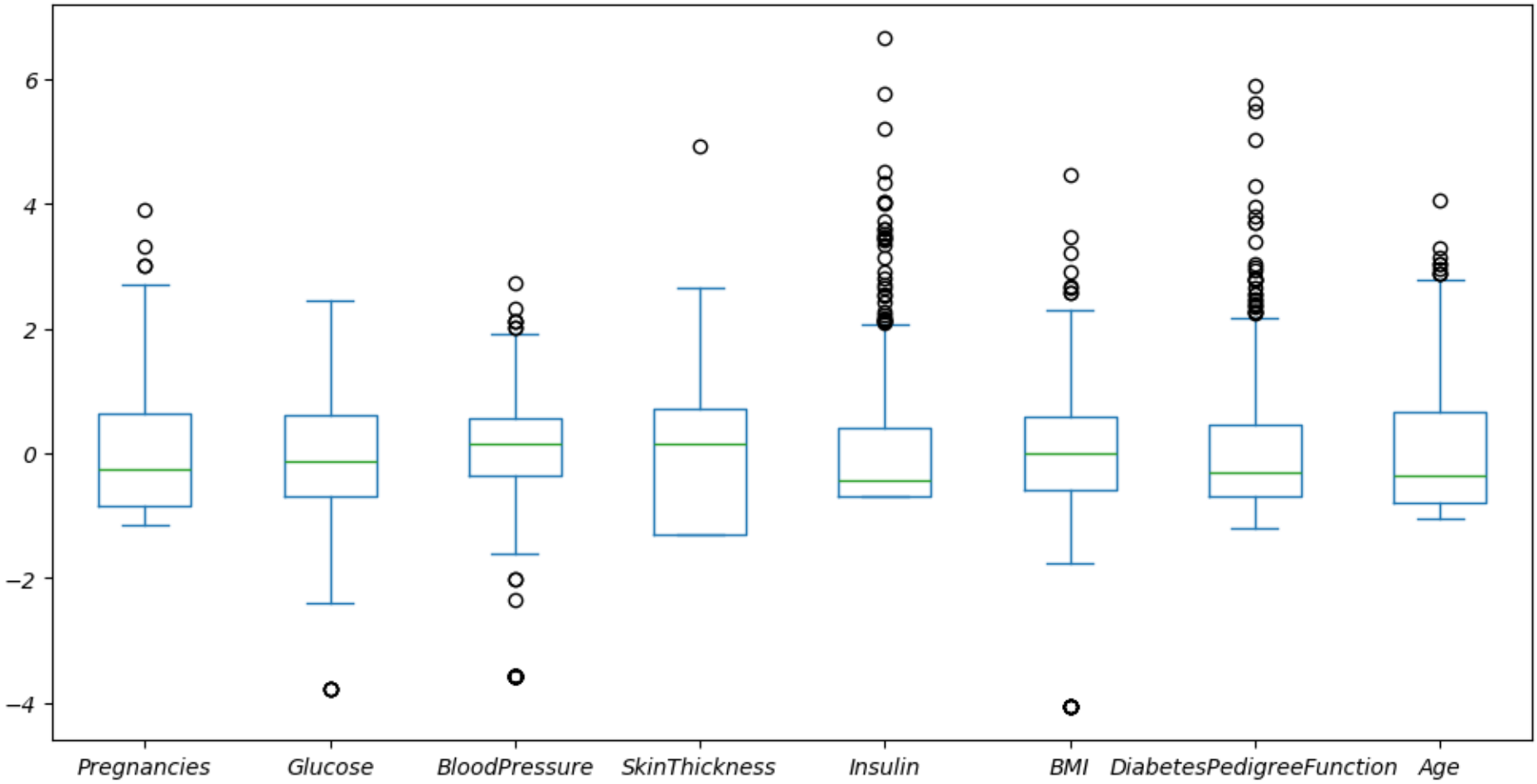


Min-Max Normalization

Đặc điểm dữ liệu sau khi chuẩn hóa:

- Tất cả các đặc trưng đầu vào (Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, Age) đã được đưa về khoảng $[0, 1]$.
- Insulin và SkinThickness có nhiều giá trị ngoại lai (outliers) nằm sát 0 hoặc gần 1.
- Glucose và BloodPressure phân bố tập trung quanh giá trị chuẩn hóa 0.5–0.6.
- Các cột còn lại trải rộng nhưng phần lớn giá trị nằm ở nửa dưới (0–0.4).

Standard Normalization



Standard Normalization

Đặc điểm dữ liệu sau khi chuẩn hóa:

- Tất cả các đặc trưng đầu vào đã được chuẩn hóa theo phân phối chuẩn với trung bình ≈ 0 và độ lệch chuẩn ≈ 1.
- Các cột đều có tâm phân bố quanh 0, độ trải rộng chủ yếu trong khoảng từ -2 đến +2.
- Insulin và SkinThickness vẫn xuất hiện nhiều giá trị ngoại lai (outliers) lớn hơn 4–6.
- Các cột khác như Glucose, BMI, Age phân bố khá cân đối quanh 0.

Chia dữ liệu thực nghiệm

Các bước thực hiện:

- Chuyển đổi dữ liệu sang dạng `numpy` với:
  - Input: *X_data*
  - Output: *y_data*
- Chia dữ liệu thành tập Train/Test với tỷ lệ 70/30.
- Kết quả: Train Ratio = 0.69921875

Tập Train:

- Shape = (537, 8)
- Ví dụ Input:

$$\begin{bmatrix} 5 & 121 & 72 & 23 & 112 & 26.2 & 0.245 & 30 \\ 1 & 130 & 60 & 23 & 170 & 28.6 & 0.692 & 21 \\ 4 & 110 & 92 & 0 & 0 & 37.6 & 0.191 & 30 \\ 7 & 107 & 74 & 0 & 0 & 29.6 & 0.254 & 31 \\ 9 & 164 & 84 & 21 & 0 & 30.8 & 0.831 & 32 \end{bmatrix}$$

- Output: [0, 0, 0, 1, 1]

Cảm ơn Thầy và các bạn lắng nghe!