**KHOA CÔNG NGHỆ THÔNG TIN TRƯỜNG ĐẠI HỌC SÀI GÒN**

**  
  
BÁO CÁO KẾT QUẢ PHÂN TÍCH DỮ LIỆU**

Pima Indians Diabetes - Phân tích EDA

**Thành viên nhóm**

Võ Hoàng Thông - 3123410363;

Phan Thanh Thịnh - 3123410360;

Lê Văn Thông - 3123410362

**GVHD: TS. Đỗ Như Tài**

Mục Lục

1.Định Nghĩa Vấn Đề (Define Problem) ..................................................... 3

2.Chuẩn Bị Vấn Đề (Prepare Problem) ..................................................... 5

2.1. Khai Báo Thư Viện (Load Libraries) ............................................. 5

2.2. Nạp Dữ Liệu (Load Dataset) ....................................................... 6

3.Phân Tích Dữ Liệu (Analyze Data) ....................................................... 7

3.1. Thống Kê Mô Tả (Descriptive Statistics) ...................................... 7

3.1.1. Hiển Thị Một Số Thông Tin Về Dữ Liệu ............................... 7

3.1.2. Kiểm Tra Tính Toàn Vẹn Của Dữ Liệu .................................. 8

3.1.3. Các Tính Chất Thống Kê Trên Dữ Liệu Số ........................... 9

3.1.4. Tần Số Xuất Hiện (Distribution) ....................................... 10

3.1.5. Mối Tương Quan Giữa Các Tính Chất (Correlations) .......... 11

3.2. Hiển Thị Dữ Liệu (Visualize Data) ............................................. 12

3.2.1. Hiển Thị Trên Từng Tính Chất Đơn (Univariate Plots) ........ 12

3.2.2. Hiển Thị Nhiều Tính Chất (Multivariate Plots) .................. 13

4.Chuẩn Bị Dữ Liệu (Prepare Data) ..................................................... 14

4.1. Làm Sạch Dữ Liệu (Data Cleaning) .......................................... 14

4.2. Biến Đổi Dữ Liệu (Data Transforms) ........................................ 15

4.2.1. Chuẩn Hóa Dữ Liệu (Data Normalize) .............................. 15

4.2.2. Chuẩn Hóa Dữ Liệu (Standard Data) ................................ 16

**1. Định Nghĩa Vấn Đề (Define Problem)**

**-Mô Tả**

Bộ dữ liệu Pima Indians Diabetes được thu thập bởi Viện Quốc Gia về Bệnh Tiểu Đường, Tiêu Hóa và Thận (National Institute of Diabetes and Digestive and Kidney Diseases). Tập dữ liệu này bao gồm thông tin y tế của 768 bệnh nhân nữ từ 21 tuổi trở lên, tất cả đều có nguồn gốc từ cộng đồng người Pima Indian ở Arizona, Mỹ.

Mục tiêu chính của bài toán là xây dựng một mô hình dự đoán khả năng mắc bệnh tiểu đường (diabetes) dựa trên các thuộc tính y tế được cung cấp. Đây là một bài toán phân lớp nhị phân (binary classification), nơi chúng ta cần phân loại bệnh nhân thành hai nhóm: mắc bệnh tiểu đường hoặc không mắc bệnh.

Bộ dữ liệu này đặc biệt hữu ích trong lĩnh vực y tế dự đoán, giúp phát hiện sớm bệnh tiểu đường – một căn bệnh mãn tính phổ biến, có thể dẫn đến các biến chứng nghiêm trọng nếu không được chẩn đoán kịp thời. Việc sử dụng dữ liệu từ một cộng đồng cụ thể (Pima Indians) cũng nhấn mạnh vào yếu tố di truyền và lối sống ảnh hưởng đến nguy cơ mắc bệnh.

**Dữ Liệu Vào (Input Attributes)**

Bộ dữ liệu bao gồm 8 thuộc tính đầu vào (features), tất cả đều là các chỉ số y tế đo lường được:

**Number of times pregnant (Số lần mang thai)**: Số lần bệnh nhân đã mang thai trong đời. Giá trị này có thể phản ánh lịch sử sinh sản và có liên quan đến nguy cơ tiểu đường thai kỳ.

**Plasma glucose concentration (Nồng độ glucose trong huyết tương sau 2 giờ)**: Đo lường mức đường huyết sau 2 giờ thử nghiệm dung nạp glucose (oral glucose tolerance test). Đây là chỉ số quan trọng nhất trong chẩn đoán tiểu đường.

**Diastolic blood pressure (Huyết áp tâm trương - mm Hg)**: Huyết áp khi tim nghỉ giữa các nhịp đập. Huyết áp cao thường liên quan đến bệnh tiểu đường và các vấn đề tim mạch.

**Triceps skin fold thickness (Độ dày nếp gấp da cánh tay - mm)**: Đo lường lượng mỡ dưới da, giúp ước lượng lượng mỡ cơ thể.

**2-Hour serum insulin (Insulin huyết thanh sau 2 giờ - mu U/ml)**: Mức insulin trong máu sau 2 giờ. Insulin cao có thể chỉ ra tình trạng kháng insulin, tiền đề của tiểu đường type 2.

**Body mass index (Chỉ số BMI = cân nặng/chiều cao²)**: Chỉ số khối cơ thể, dùng để đánh giá tình trạng thừa cân hoặc béo phì – yếu tố nguy cơ cao của tiểu đường.

**Diabetes pedigree function (Chỉ số phả hệ tiểu đường)**: Điểm số dựa trên lịch sử gia đình mắc bệnh tiểu đường, phản ánh yếu tố di truyền.

**Age (Tuổi - năm)**: Tuổi của bệnh nhân. Tuổi cao hơn thường liên quan đến nguy cơ mắc bệnh cao hơn.

Kết Quả (Output / Class Variable)

**0**: Không mắc tiểu đường (negative).

**1**: Mắc tiểu đường (positive).

Trong bộ dữ liệu, khoảng 65% trường hợp là negative (0), dẫn đến tình trạng mất cân bằng lớp (imbalanced classes), cần xử lý đặc biệt trong mô hình học máy để tránh thiên kiến.

**2. Chuẩn Bị Vấn Đề (Prepare Problem)**

**2.1. Khai Báo Thư Viện (Load Libraries)**

Để xử lý và phân tích dữ liệu, chúng tôi sử dụng các thư viện Python phổ biến trong khoa học dữ liệu:

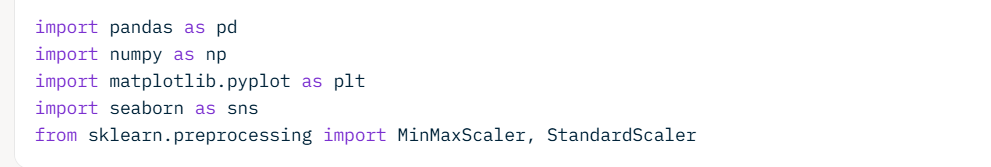
**Pandas**: Để đọc, thao tác và phân tích dữ liệu dạng bảng (DataFrame).

**NumPy**: Hỗ trợ tính toán số học và mảng đa chiều.

**Matplotlib và Seaborn**: Để vẽ biểu đồ và trực quan hóa dữ liệu (ví dụ: box plots, scatter plots, heatmap correlations).

**Scikit-learn**: Để chuẩn hóa dữ liệu, xử lý missing values và xây dựng mô hình (nếu cần mở rộng).

Ví dụ mã:

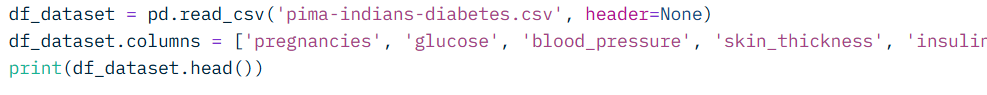


Những thư viện này giúp quá trình chuẩn bị dữ liệu trở nên hiệu quả và dễ dàng tái sử dụng.

**2.2. Nạp Dữ Liệu (Load Dataset)**

Dữ liệu được tải từ file CSV (pima-indians-diabetes.csv) với cấu trúc 9 cột: 8 thuộc tính đầu vào và 1 cột nhãn (class).

Ví dụ mã:



Bảng dữ liệu mẫu (5 dòng đầu tiên):

****

**3. Phân Tích Dữ Liệu (Analyze Data)**

**3.1. Thống Kê Mô Tả (Descriptive Statistics)**

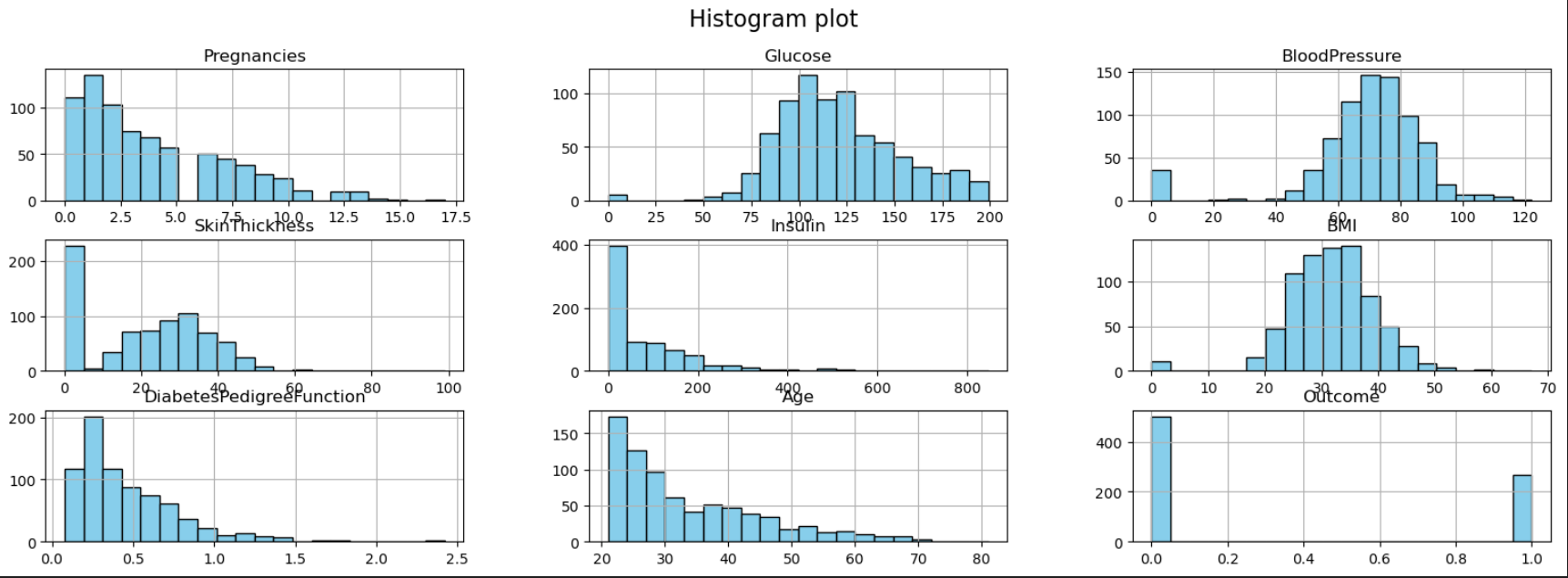
**3.1.1. Hiển Thị Một Số Thông Tin Về Dữ Liệu**

**Số dòng và cột**: 768 dòng, 9 cột (8 features + 1 class).

**Kiểu dữ liệu**: Tất cả các cột đều là số (int64 hoặc float64).

**5 dòng đầu và cuối**: Như bảng mẫu ở trên (đầu) và tương tự cho cuối (không liệt kê chi tiết để tránh dài dòng).

**Thông tin chung**: Dữ liệu bao gồm các chỉ số y tế với đơn vị khác nhau, nhãn class là binary (0/1).



**Nhận xét chi tiết**: Dữ liệu tập trung vào các yếu tố rủi ro tiểu đường ở phụ nữ Pima Indian. Các features có quy mô khác nhau (ví dụ: age từ 21-81, glucose từ 0-199), đòi hỏi chuẩn hóa sau này. Tổng số mẫu là 768, đủ lớn cho phân tích nhưng cần kiểm tra imbalanced classes.

**3.1.2. Kiểm Tra Tính Toàn Vẹn Của Dữ Liệu**

**Dữ liệu trùng lặp**: Không có (0 dòng).

**Giá trị Null/NaN**: Không có (False cho cả hai).

**Nhận xét chi tiết**: Dữ liệu khá sạch, không có missing values rõ ràng. Tuy nhiên, một số giá trị 0 ở các cột như glucose, blood\_pressure, bmi có thể là missing values ngầm (vì không thể bằng 0 về mặt sinh học). Chúng cần được xử lý ở phần cleaning.

**3.1.3. Các Tính Chất Thống Kê Trên Dữ Liệu Số**

Sử dụng df.describe() để tính:

|  | **pregnancies** | **glucose** | **blood\_pressure** | **skin\_thickness** | **insulin** | **bmi** | **diabetes\_pedigree** | **age** | **class** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| count | 768.000000 | 768.000000 | 768.000000 | 768.000000 | 768.000000 | 768.000000 | 768.000000 | 768.000000 | 768.000000 |
| mean | 3.845052 | 120.894531 | 69.105469 | 20.536458 | 79.799479 | 31.992578 | 0.471876 | 33.240885 | 0.348958 |
| std | 3.369578 | 31.972618 | 19.355807 | 15.952218 | 115.244002 | 7.884160 | 0.331329 | 11.760232 | 0.476951 |
| min | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.078000 | 21.000000 | 0.000000 |
| 25% | 1.000000 | 99.000000 | 62.000000 | 0.000000 | 0.000000 | 27.300000 | 0.243750 | 24.000000 | 0.000000 |
| 50% | 3.000000 | 117.000000 | 72.000000 | 23.000000 | 30.500000 | 32.000000 | 0.372500 | 29.000000 | 0.000000 |
| 75% | 6.000000 | 140.250000 | 80.000000 | 32.000000 | 127.250000 | 36.600000 | 0.626250 | 41.000000 | 1.000000 |
| max | 17.000000 | 199.000000 | 122.000000 | 99.000000 | 846.000000 | 67.100000 | 2.420000 | 81.000000 | 1.000000 |

**Nhận xét chi tiết**:

Nhiều min=0 ở glucose, blood\_pressure, skin\_thickness, insulin, bmi – đây là missing values ngầm, cần thay thế (ví dụ: bằng mean hoặc median).

Insulin có std lớn (115) và max=846, cho thấy phân bố lệch phải (right-skewed) với nhiều outliers.

Class mean ≈0.35: 35% positive, xác nhận imbalanced dataset. Blood\_pressure mean=69 mmHg, thấp hơn chuẩn (80 mmHg), có thể do đặc thù dân số.

3.1.4. Tần Số Xuất Hiện (Distribution) Trên Dữ Liệu Phân Lớp

Class 0 (không tiểu đường): 500 (65%).

Class 1 (có tiểu đường): 268 (35%).

**Nhận xét chi tiết**: Lớp 0 chiếm ưu thế, dẫn đến imbalanced problem. Trong học máy, cần kỹ thuật như oversampling (SMOTE), undersampling hoặc class weights để tránh mô hình thiên về lớp đa số.

**3.1.5. Mối Tương Quan Giữa Các Tính Chất (Correlations)**

Sử dụng Pearson correlation:

Heatmap cho thấy:

skin\_thickness ~ insulin (0.437): Da dày liên quan đến insulin cao.

skin\_thickness ~ bmi (0.393): BMI cao dẫn đến da dày.

pregnancies ~ age (0.544): Tuổi cao hơn, số lần mang thai nhiều hơn.

Tương quan với class: glucose (0.466) cao nhất, theo sau bmi (0.293), age (0.238), pregnancies (0.222).

**Nhận xét chi tiết**: Không có multicollinearity nghiêm trọng (correlation >0.9), nhưng cần theo dõi trong mô hình tuyến tính. Glucose là feature quan trọng nhất, phù hợp với kiến thức y tế về tiểu đường.

**3.2. Hiển Thị Dữ Liệu (Visualize Data)**

**3.2.1. Hiển Thị Trên Từng Tính Chất Đơn (Univariate Plots)**

**Box and Whisker Plots**: Cho thấy độ phân tán và outliers.

Insulin có nhiều outliers (>300, lên đến ~900).

Pregnancies, Insulin, Age lệch phải (skewed towards small values).

SkinThickness lệch trái (skewed towards large values).

Glucose, BloodPressure, BMI, DiabetesPedigreeFunction khá cân bằng.

**Nhận xét chi tiết**: Độ trải rộng khác nhau giữa features đòi hỏi chuẩn hóa. Outliers ở insulin có thể là dữ liệu thực (do kháng insulin) nhưng cần kiểm tra để tránh ảnh hưởng mô hình.

**3.2.2. Hiển Thị Nhiều Tính Chất (Multivariate Plots)**

**Scatter Plots và Pairplots**:

Pregnancies vs Age: Xu hướng dương (tuổi cao, mang thai nhiều).

Glucose vs Outcome: Phân tách rõ (glucose cao → tiểu đường).

BMI vs Outcome: BMI cao liên quan tiểu đường.

Insulin vs SkinThickness: Mối quan hệ tích cực.

**Nhận xét chi tiết**: Các biểu đồ xác nhận correlations từ phần trước. Chúng giúp xác định features quan trọng và phát hiện patterns, hỗ trợ feature selection.

**4. Chuẩn Bị Dữ Liệu (Prepare Data)**

**4.1. Làm Sạch Dữ Liệu (Data Cleaning)**

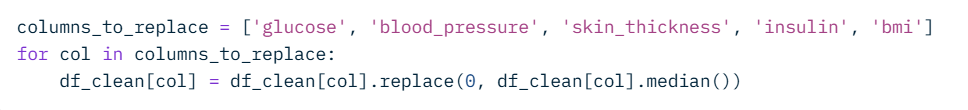
**Tạo bản sao**: df\_clean = df\_dataset.copy().

**Xóa trùng lặp**: Không có, nhưng kiểm tra bằng df\_clean.duplicated().

**Xử lý Null/NaN**: Không có, nhưng kiểm tra bằng isnull() và isna().

**Xử lý giá trị 0 không hợp lệ**: Thay thế 0 ở glucose, blood\_pressure, skin\_thickness, insulin, bmi bằng median (vì phân bố skewed).

Ví dụ mã:



Nhận xét chi tiết: Sau cleaning, dữ liệu sạch hơn, giảm bias từ missing values ngầm. Tổng số dòng vẫn 768.

**4.2. Biến Đổi Dữ Liệu (Data Transforms)**

**4.2.1. Chuẩn Hóa Dữ Liệu (Data Normalize)**

**Min-Max Normalization**: Đưa về [0,1] bằng MinMaxScaler. Ví dụ: X\_normalized = (X - min) / (max - min).

**Standard Normalization**: Đưa về mean=0, std=1 bằng StandardScaler. Ví dụ: X\_standard = (X - mean) / std.

**Nhận xét chi tiết**: Normalization giúp các features có quy mô tương đương, cải thiện hiệu suất mô hình như KNN, Neural Networks. StandardScaler phù hợp hơn cho dữ liệu Gaussian.

**4.2.2. Chuẩn Hóa Dữ Liệu (Standard Data)**

Tương tự phần trên, áp dụng StandardScaler cho toàn bộ features.

**Cuối cùng Kết Luận Và Lời Cảm Ơn**

Qua phân tích, bộ dữ liệu Pima Indians Diabetes cung cấp cái nhìn sâu sắc về các yếu tố rủi ro tiểu đường. Các bước chuẩn bị và phân tích đã làm rõ các vấn đề như missing values ngầm, imbalanced classes và correlations, sẵn sàng cho việc xây dựng mô hình học máy (ví dụ: Logistic Regression, Random Forest).

Cảm ơn thầy Đỗ Như Tài và các bạn đã hỗ trợ và lắng nghe báo cáo!