I. Dữ liệu:

1. Cách tải:

```
import pandas as pd
from sklearn.datasets import load_diabetes
# Load diabetes dataset
bunch = load_diabetes(as_frame=True)
df = pd.concat([bunch.data, bunch.target.rename("target")], axis=1)
# Export to CSV
csv_path = "du_lieu_diabetes.csv"
df.to_csv(csv_path, index=False)
print(f"Dataset đã được lưu vào {csv_path}")
```

2. Mô tả

Bộ dữ liệu **Diabetes** trong scikit-learn dùng để mô phỏng mô hình hồi quy dự đoán diễn biến (progression) bệnh tiểu đường một năm sau lần đo ban đầu. Cụ thể:

- **Nguồn gốc và mục đích**
    - Được tổng hợp từ nghiên cứu y khoa (Efron et al., 1983), thường dùng làm ví dụ điển hình cho bài toán hồi quy tuyến tính đa biến.
    - Mục tiêu: dự đoán chỉ số diễn tiến bệnh tiểu đường (“disease progression”) sau 1 năm dựa trên các biến lâm sàng lấy ở thời điểm ban đầu.
- **Kích thước**
    - **442** mẫu quan sát.
    - **10** biến độc lập (predictors) và 1 biến phụ thuộc (target).
- **Các biến độc lập (X)**
    - **age**: tuổi (đã được chuẩn hóa đều về trung bình 0, độ lệch chuẩn 1)
    - **sex**: giới tính (đã mã hóa và chuẩn hóa)
    - **bmi**: chỉ số khối cơ thể, Body Mass Index
    - **bp**: huyết áp trung bình (mean arterial pressure)
    - **s1**: mức độ một loại cholesterol (tổng)
    - **s2**: mức độ lipoprotein mật độ thấp (LDL)
    - **s3**: mức độ lipoprotein mật độ cao (HDL)
      s4: mức độ đường huyết (thiết lập ban đầu)
    - **s5**: mức độ insulin huyết thanh
    - **s6**: một chỉ số phosphatase kiềm

- Các biến s1–s6 đều là những chỉ số sinh hoá trong máu, đã được chuẩn hóa (standardized) để thuận tiện cho mô hình hóa.

- **Biến phụ thuộc (y)**
    - **target**: chỉ số định lượng diễn tiến bệnh tiểu đường sau 1 năm. Giá trị càng cao thể hiện bệnh tiến triển nặng hơn.
- **Tại sao dùng dataset này?**
    - Số chiều vừa phải (10 biến) để minh hoạ MLR mà không quá phức tạp.
    - Biến mục tiêu liên tục, phù hợp với hồi quy tuyến tính.
    - Các biến đã được chuẩn hóa sẵn, giúp tập trung vào khái niệm ước lượng hệ số mà không lo chuẩn hoá dữ liệu.
- **Lưu ý khi áp dụng**
    - Dữ liệu đã chuẩn hóa: nếu dùng trực tiếp hệ số hồi quy, bạn cần nhớ trừ bù giá trị gốc (nghĩa là kết quả của mô hình áp lên dữ liệu chuẩn hóa).
    - Kiểm tra giả định: phân phối sai số, đa cộng tuyến, heteroscedasticity… để đảm bảo ước lượng OLS là hợp lý.

Tóm lại, bộ dữ liệu Diabetes là một "chuẩn" cho bài toán hồi quy tuyến tính đa biến: có đủ biến sinh lý/chẩn đoán, kích thước mẫu vừa phải, và đã được xử lý chuẩn hoá, rất thích hợp để minh hoạ lý thuyết và thực nghiệm MLR.