Phân tích dữ liệu khám phá

Dự Báo Bệnh Đái Tháo Đường

Nhiệm vụ từng thành viên

MSSv	Họ và Tên	Nhiệm vụ
3123410065	Phạm Minh Dương	Đọc paper 1-3, tổng hợp và soạn lại nội dung slide
3123410072	Phạm Tấn Đạt	Đọc paper 2, tóm tắt nội dung
3123410168	Nguyễn Đăng Khoa	Đọc paper 1, tóm tắt nội dung, edit nội dung sang PDF
3123410426	Đỗ Quốc Việt	Xây dựng code

Nội dung thảo luận

- **Khảo sát bài toán:** Tổng quan về bệnh tiểu đường và các tiêu chuẩn chẩn đoán liên quan.
- **Tóm tắt dữ liệu:** Mô tả các biến được sử dụng để dự báo.
- **Phân tích đơn biến:** Khám phá phân phối của từng yếu tố nguy cơ.
- **Phân tích đa biến:** Kiểm tra mối quan hệ giữa các yếu tố nguy cơ và tình trạng bệnh tiểu đường.
- **Xác định giá trị thiếu và ngoại lệ:** Đánh giá các vấn đề tiềm ẩn trong dữ liệu.
- **Phân tích tương quan:** Đánh giá mối tương quan giữa các biến.
- **Kết luận:** Tổng hợp các phát hiện chính từ phân tích.

1. Khảo sát bài toán

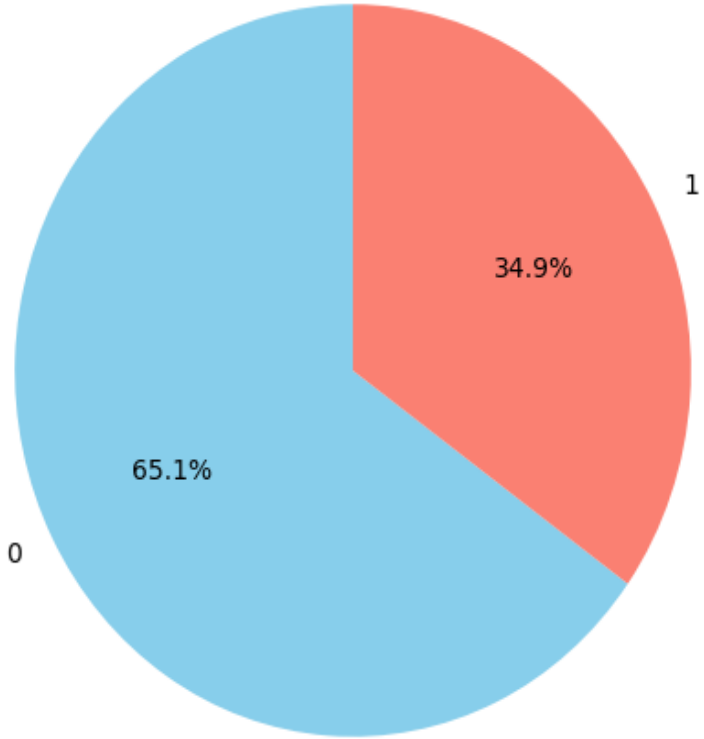
- **Bối cảnh:** Bệnh tiểu đường (Diabetes Mellitus) là một nhóm các rối loạn chuyển hóa đặc trưng bởi tình trạng tăng đường huyết mãn tính do khiếm khuyết trong việc tiết insulin, hoạt động của insulin, hoặc cả hai. Bài toán được khảo sát tập trung vào việc dự báo sự khởi phát của bệnh tiểu đường trong một quần thể có nguy cơ cao là người da đỏ Pima, như được mô tả trong nghiên cứu của Smith và cộng sự (1988).
- **Phân loại & Tiêu chí chẩn đoán:** Các hệ thống phân loại đã phát triển theo thời gian. Ban đầu, các thuật ngữ như tiểu đường phụ thuộc insulin (IDDM) và không phụ thuộc insulin (NIDDM) được sử dụng. Các khuyến nghị sau này của Tổ chức Y tế Thế giới (WHO) đã chuyển sang dùng
- **Type 1** và **Type 2** để phản ánh rõ hơn về căn nguyên bệnh. Tiêu chí chẩn đoán cũng đã thay đổi, đặc biệt là việc hạ thấp ngưỡng đường huyết lúc đói (Fasting Plasma Glucose) từ
- ≥140 mg/dl xuống ≥126 mg/dl để tăng độ nhạy trong chẩn đoán.
- **Mục tiêu bài toán:** Sử dụng các dữ liệu lâm sàng để xây dựng một mô hình có khả năng dự báo liệu một cá nhân có nguy cơ phát triển bệnh tiểu đường trong vòng 5 năm hay không. Việc này giúp xác định sớm các đối tượng cần can thiệp y tế để phòng ngừa hoặc làm chậm tiến triển của bệnh.

2. Phân tích khám phá dữ liệu

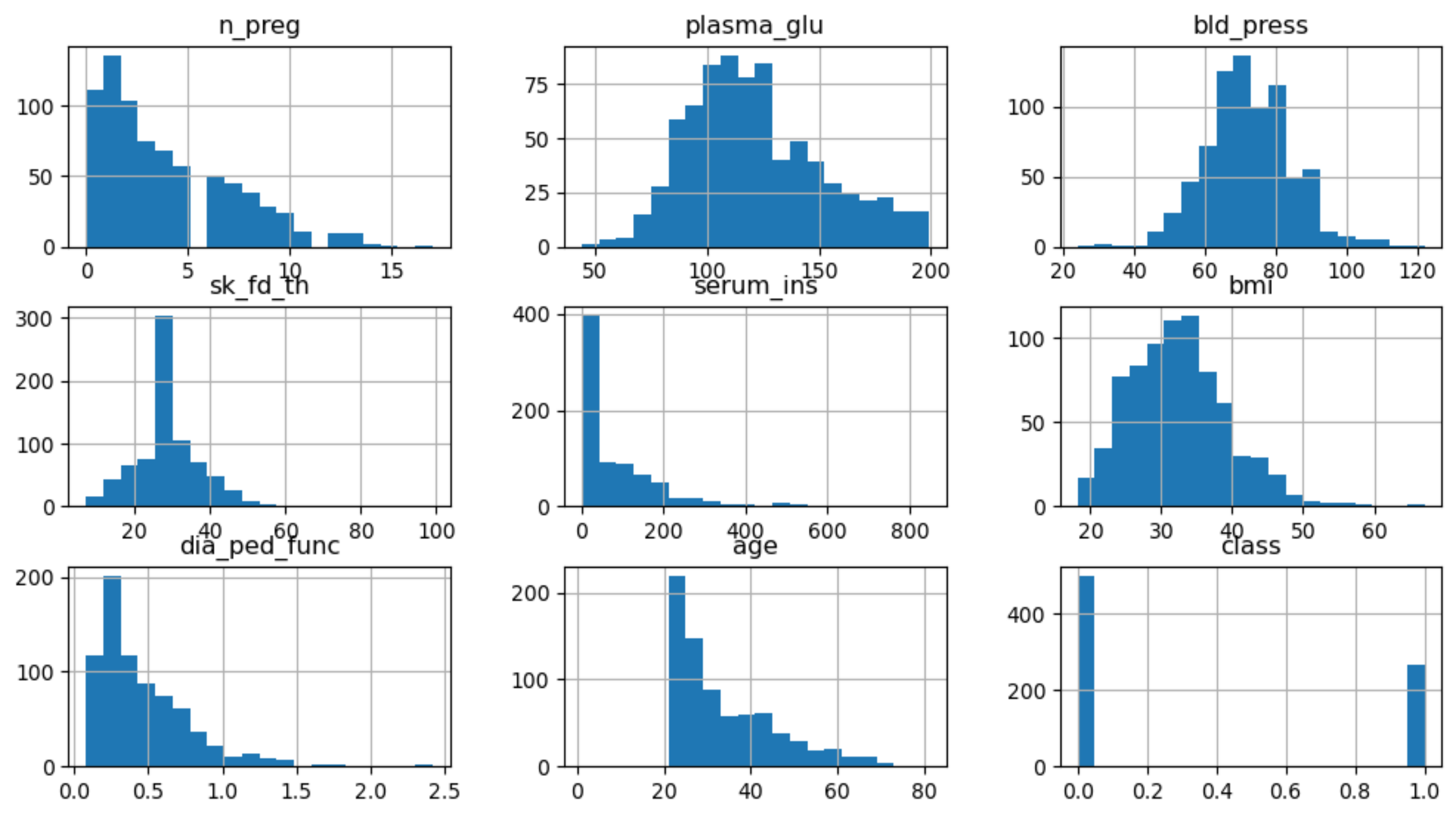
- Phân tích này sẽ khám phá bộ dữ liệu dựa trên 8 biến đầu vào được sử dụng trong nghiên cứu gốc của Smith và cộng sự. Biến mục tiêu là
- class (hoặc outcome), cho biết tình trạng mắc bệnh (1) hay không mắc bệnh (0).
- **Tính chất dữ liệu**
- Dữ liệu bao gồm các chỉ số sinh hóa và nhân khẩu học của bệnh nhân nữ thuộc bộ tộc Pima.
- **Biến số (Numeric):**
  - n_preg: Số lần mang thai
  - plasma_glu: Nồng độ glucose huyết tương sau 2 giờ
  - bld_press: Huyết áp tâm trương
  - sk_fd_th: Độ dày nếp gấp da
  - serum_ins: Nồng độ insulin huyết thanh sau 2 giờ
  - bmi: Chỉ số khối cơ thể
  - dia_ped_func: Hàm phả hệ tiểu đường
  - age: Tuổi
- **Biến mục tiêu (Binary):**
  - class: 1 (mắc bệnh), 0 (không mắc bệnh)

3.Phân tích đơn biến

- Phân tích này xem xét sự phân bổ của từng biến riêng lẻ.
- **Phân bổ lớp (Class Distribution):** Dữ liệu cho thấy sự mất cân bằng giữa hai lớp. Khoảng **65.1%** mẫu thuộc lớp không mắc bệnh (0) và **34.9%** thuộc lớp mắc bệnh (1). Điều này cần được lưu ý khi xây dựng và đánh giá mô hình.

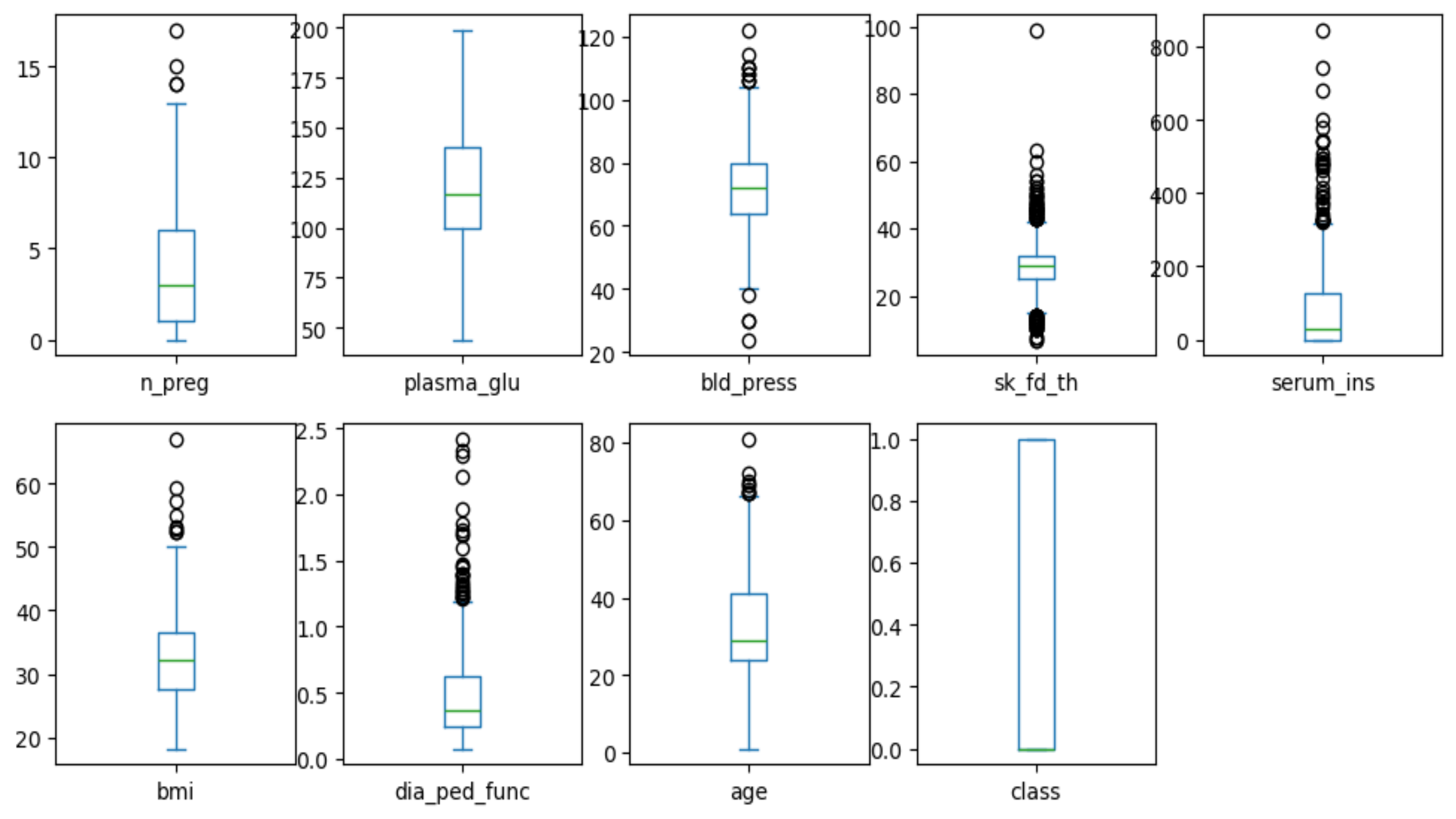


- **Phân phối các biến số:**
  - **n_preg (Số lần mang thai)** và **age (Tuổi)**: Cả hai đều có phân phối lệch phải, cho thấy phần lớn các cá nhân trong mẫu là người trẻ và có ít lần mang thai.
  - **plasma_glu (Glucose huyết tương)**, **bld_press (Huyết áp)** và **bmi (Chỉ số khối cơ thể)**: Các biến này có phân phối gần giống phân phối chuẩn, tập trung quanh giá trị trung bình. Đây là những chỉ số sức khỏe quan trọng.
  - **sk_fd_th (Độ dày nếp gấp da)** và **serum_ins (Insulin huyết thanh)**: Có phân phối lệch phải rất mạnh, với nhiều giá trị tập trung gần 0 và một số giá trị ngoại lệ rất cao.


n_preg
plasma_glu
bld_press
sk_fd_th
serum_ins
bmi
dia_ped_func
age
class

4. Phân tích đa biến

- Phân tích này khám phá mối quan hệ giữa các biến với nhau và với biến mục tiêu.
- **So sánh giữa nhóm Mắc bệnh (1) và Không mắc bệnh (0):**
  - **plasma_glu (Glucose huyết tương):** Đây là yếu tố khác biệt rõ ràng nhất. Nhóm mắc bệnh có phân phối glucose dịch chuyển hẳn về phía các giá trị cao hơn so với nhóm không mắc bệnh.
  - **bmi (Chỉ số khối cơ thể)** và **age (Tuổi)**: Nhóm mắc bệnh có xu hướng có chỉ số BMI và tuổi trung bình cao hơn.
  - **n_preg (Số lần mang thai)**: Phụ nữ mắc bệnh tiểu đường cũng có xu hướng có số lần mang thai nhiều hơn.
  - **bld_press (Huyết áp)** và **dia_ped_func (Hàm phả hệ)**: Cũng cho thấy sự khác biệt nhỏ, với giá trị trung bình cao hơn ở nhóm mắc bệnh.

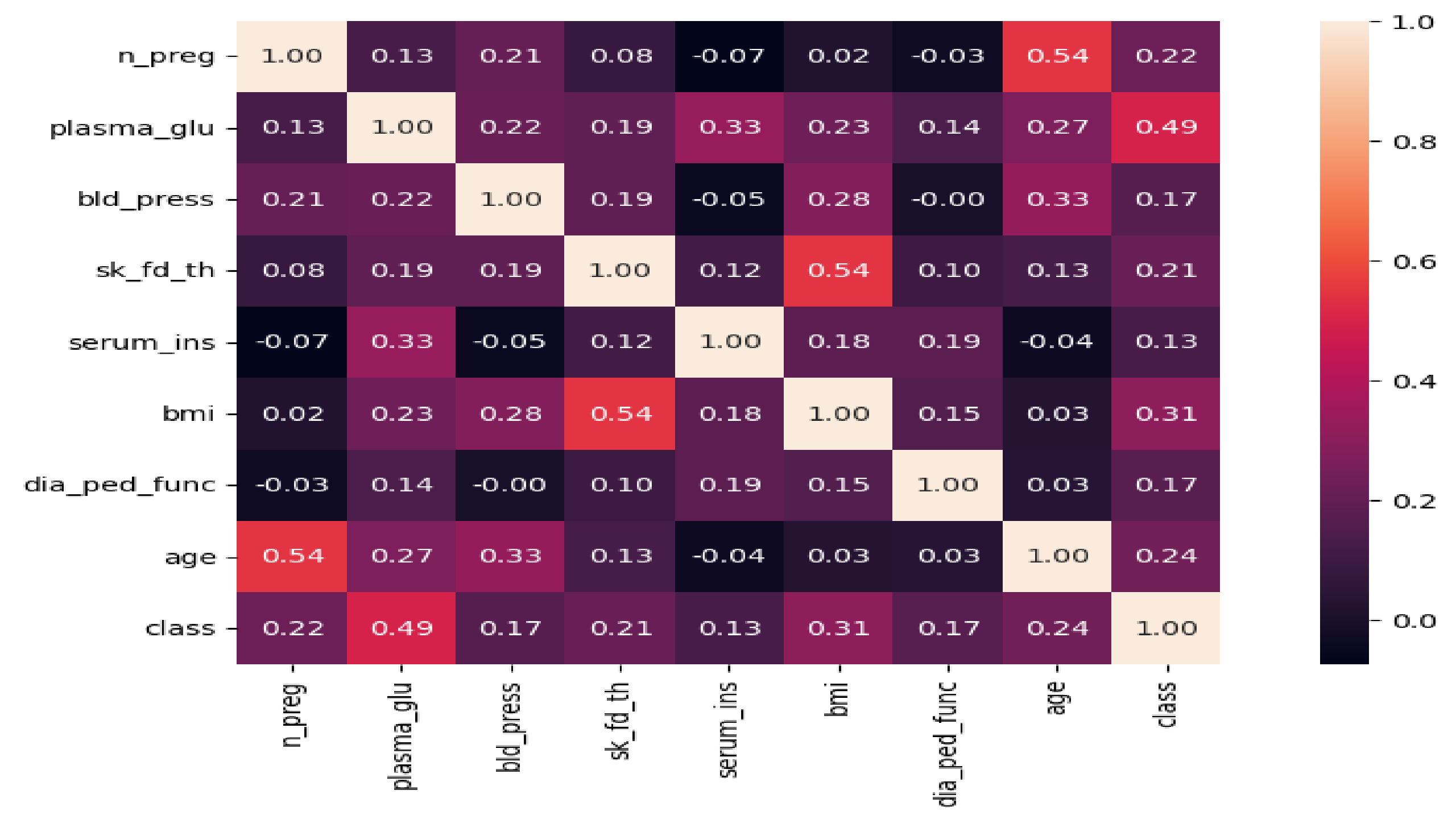

n_preg
plasma_glu
bld_press
sk_fd_th
serum_ins
bmi
dia_ped_func
age
class

5. Xác định giá trị thiếu và ngoại lệ

- **Giá trị thiếu (Missing Values):** Một số biến như bld_press, sk_fd_th, và bmi có giá trị tối thiểu là 0. Về mặt sinh lý, các giá trị này là không thể có (ví dụ: huyết áp bằng 0, độ dày nếp gấp da bằng 0). Đây rất có thể là các giá trị bị thiếu đã được mã hóa thành 0. Việc xử lý các giá trị này (ví dụ: thay thế bằng giá trị trung bình/trung vị hoặc loại bỏ) là cần thiết trước khi xây dựng mô hình.
- **Ngoại lệ (Outliers):**
  - Các biểu đồ hộp cho thấy sự hiện diện của nhiều giá trị ngoại lệ, đặc biệt ở các biến **serum_ins**, **sk_fd_th**, và **dia_ped_func**.
  - Trong lĩnh vực y tế, các giá trị ngoại lệ này có thể là các trường hợp bệnh lý thực sự và chứa thông tin quan trọng. Do đó, cần phải xem xét cẩn thận thay vì loại bỏ chúng một cách máy móc. Ví dụ, một mức insulin rất cao có thể là dấu hiệu của tình trạng kháng insulin nghiêm trọng.

6. Phân tích tương quan

- Ma trận tương quan cho thấy mối quan hệ tuyến tính giữa các cặp biến.
- **Tương quan với biến mục tiêu (class):**
  - **plasma_glu** có tương quan dương mạnh nhất với class (0.49), khẳng định vai trò quan trọng của nồng độ glucose trong việc chẩn đoán bệnh tiểu đường.
  - **bmi** (0.31), **age** (0.24), và **n_preg** (0.22) cũng có tương quan dương đáng kể với class.
- **Tương quan giữa các biến độc lập:**
  - **age** và **n_preg** có tương quan dương mạnh (0.54), điều này là hợp lý về mặt logic.
  - **bmi** và **sk_fd_th** cũng có tương quan dương mạnh (0.54), vì cả hai đều là chỉ số đo lường lượng mỡ trong cơ thể.
  - Mối tương quan giữa các biến khác hầu hết ở mức thấp đến trung bình, cho thấy chúng cung cấp những thông tin tương đối độc lập cho mô hình.


n_preg 1.00 0.13 0.21 0.08 -0.07 0.02 -0.03 0.54 0.22
plasma_glu 0.13 1.00 0.22 0.19 0.33 0.23 0.14 0.27 0.49
bld_press 0.21 0.22 1.00 0.19 -0.05 0.28 -0.00 0.33 0.17
sk_fd_th 0.08 0.19 0.19 1.00 0.12 0.54 0.10 0.13 0.21
serum_ins -0.07 0.33 -0.05 0.12 1.00 0.18 0.19 -0.04 0.13
bmi 0.02 0.23 0.28 0.54 0.18 1.00 0.15 0.03 0.31
dia_ped_func -0.03 0.14 -0.00 0.10 0.19 0.15 1.00 0.03 0.17
age 0.54 0.27 0.33 0.13 -0.04 0.03 0.03 1.00 0.24
class 0.22 0.49 0.17 0.21 0.13 0.31 0.17 0.24 1.00
n_preg
plasma_glu
bld_press
sk_fd_th
serum_ins
bmi
dia_ped_func
age
class
1.0
0.8
0.6
0.4
0.2
0.0

7. Kết luận

- Từ phân tích khám phá dữ liệu, có thể rút ra các kết luận sau:
- **Các yếu tố dự báo quan trọng: Nồng độ glucose huyết tương** là yếu tố dự báo mạnh mẽ và rõ ràng nhất. Bên cạnh đó, **chỉ số BMI, tuổi tác** và **số lần mang thai** cũng là những yếu tố nguy cơ đáng kể.
- **Vấn đề chất lượng dữ liệu:** Dữ liệu có chứa các giá trị "0" bất thường, có khả năng là giá trị thiếu được mã hóa. Việc xử lý các giá trị này là bước tiền xử lý quan trọng.
- **Đặc điểm của nhóm mắc bệnh:** Các bệnh nhân được chẩn đoán mắc bệnh tiểu đường có xu hướng lớn tuổi hơn, có chỉ số BMI cao hơn, nồng độ glucose cao hơn và có nhiều lần mang thai hơn so với nhóm không mắc bệnh.
- **Hướng đi tiếp theo:** Dựa trên những khám phá này, các bước tiếp theo bao gồm việc xử lý các giá trị thiếu một cách hợp lý, chuẩn hóa dữ liệu và xây dựng các mô hình học máy (như hồi quy logistic hoặc mạng nơ-ron như trong paper2) để dự báo nguy cơ mắc bệnh. Cần chú ý đến sự mất cân bằng của dữ liệu khi lựa chọn độ đo đánh giá mô hình.