Các vấn đề thảo luận (Points to discuss)

Danh mục (Agenda)

Tóm tắt dữ liệu (Data Summary)

Phân tích đơn biến (Univariate analysis)

Phân tích nhị biến (Bivariate analysis)

Phân tích đa biến (Mulvariate analysis)

Phân tích theo Nhóm Tuổi (Age Group wise analysis)

Phân tích theo Mức độ Béo phì (BMI Category analysis)

Phân tích theo Ngưỡng Đường huyết (Glucose Threshold analysis)

Phân tích theo Tiền sử Gia đình (Pedigree Function analysis)

Bản đồ nhiệt tương quan (Correlation heatmap)

Kết luận (Conclusion)

Danh mục (Agenda)

Tóm tắt dữ liệu (Data Summary)

Phân tích đơn biến (Univariate analysis)

Phân tích nhị biến (Bivariate analysis)

Phân tích đa biến (Mulvariate analysis)

Phân tích theo Nhóm Tuổi (Age Group wise analysis)

Phân tích theo Mức độ Béo phì (BMI Category analysis)

Phân tích theo Ngưỡng Đường huyết (Glucose Threshold analysis)

Phân tích theo Tiền sử Gia đình (Pedigree Function analysis)

Tóm tắt dữ liệu (Data Summary)

Tập dữ liệu cho trước chứa các cột biến khác nhau đóng vai trò quan trọng trong việc dự đoán bệnh tiểu đường. Các biến đó bao gồm:

pregnancies: số lần một người phụ nữ từng mang thai.

glucose: chỉ số đường huyết được đo **2 giờ sau** khi bệnh nhân uống 75g dung dịch glucose. Đây là tiêu chuẩn vàng để chẩn đoán tiểu đường và tiền tiểu đường.

blood_pressure: áp lực trong động mạch khi tim giãn ra (là số dưới trong chỉ số huyết áp, ví dụ 120/**80**).

skin_thickness: một phương pháp đo lượng mỡ dự trữ trong cơ thể. Người ta dùng một thước kẹp đặc biệt để kẹp và đo độ dày của một nếp da ở mặt sau cánh tay.

insulin: lượng insulin trong máu được đo cùng thời điểm với chỉ số đường huyết sau 2 giờ.

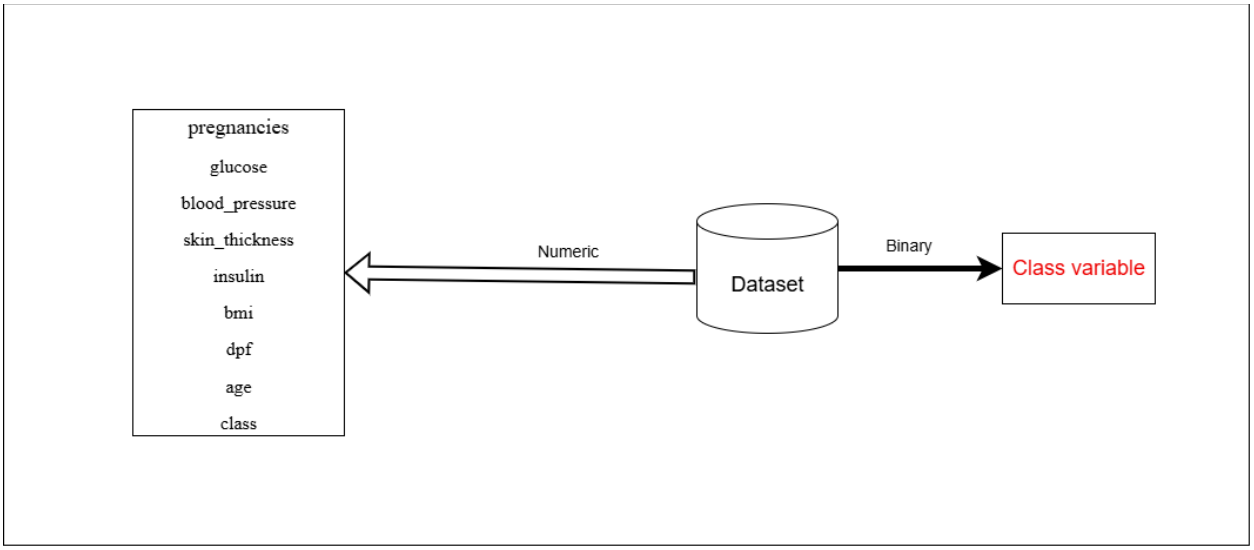
bmi (bode mass index): BMI = Cân nặng (kg) / (Chiều cao (m))^2. Đây là chỉ số đánh giá thể trạng gầy/béo phổ biến.

dpf (diabetes pedigree function): chỉ số lượng hóa tiền sử gia đình mắc bệnh tiểu đường. Nó được tính toán dựa trên mối quan hệ huyết thống và số lượng thành viên trong gia đình mắc bệnh.

age: Tuổi của một người tại thời điểm khảo sát.

class: kết quả chẩn đoán mà mô hình máy học cần dự đoán (0 – Âm tính với bệnh tiểu đường; 1 – Dương tính với bệnh tiểu đường)

Tóm tắt dữ liệu (Data Summary)



Phân tích đơn biến (Univariate analysis)

1. **Histogram hoặc KDE Plot** cho tất cả các biến số (Glucose, BMI, Age, Insulin...).
2. **Box Plot** cho tất cả các biến số để phát hiện outlier.
3. **Count Plot (Bar Chart)** cho biến mục tiêu Class.

Phân tích đơn biến (Univariate analysis)

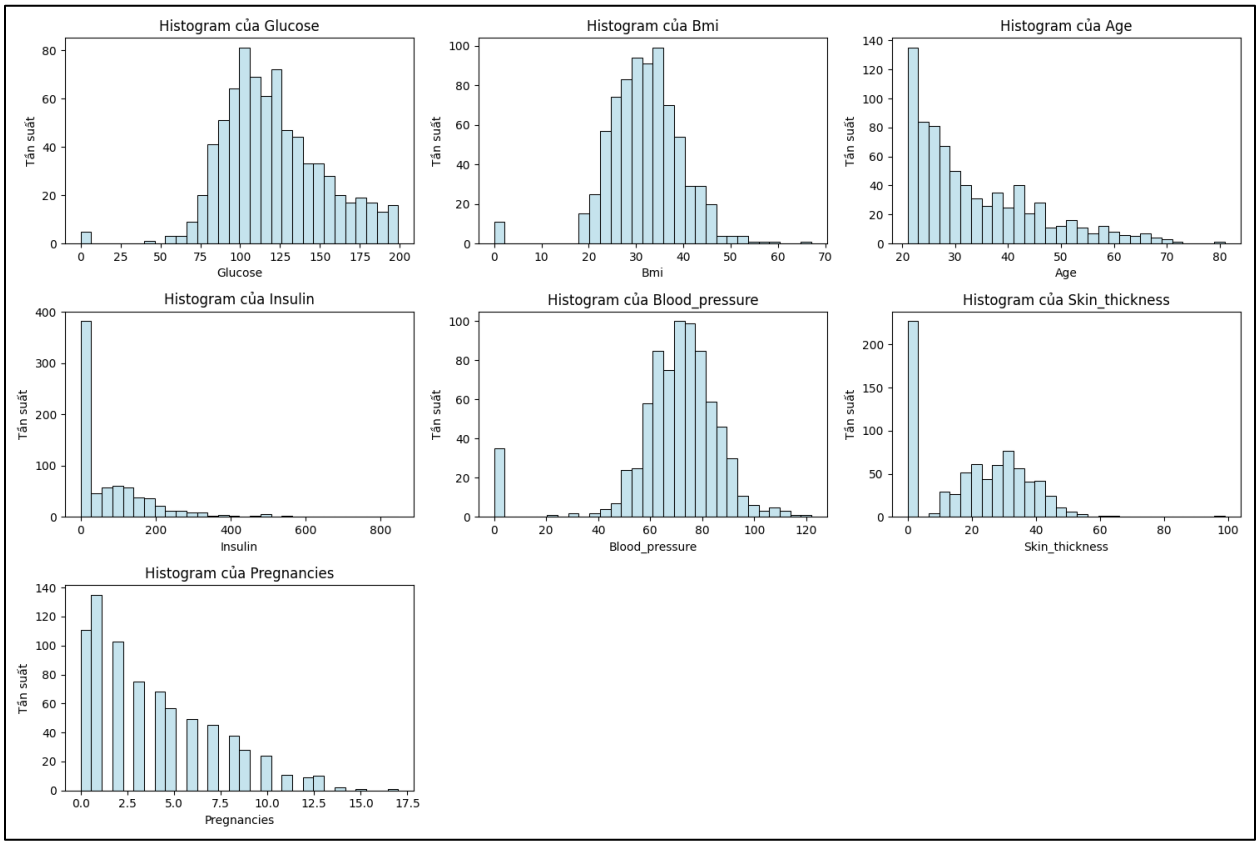


Chart 1. Histogram các thuộc tính

Từ histogram ta có thể nhận xét được:

- Các thuộc tính có phân phối gần chuẩn với tần suất cao ở vùng trung tâm: glucose, blood_pressure, BMI (peak ở trung tâm).

- Các thuộc tính có phân phối lệch phải với tần suất cao ở giá trị thấp: age, insulin, skin_thickness, pregnancies (peak ở giá trị thấp).

- Đặc biệt, insulin và skin_thickness có nhiều giá trị bằng 0.

Phân tích đơn biến (Univariate analysis)

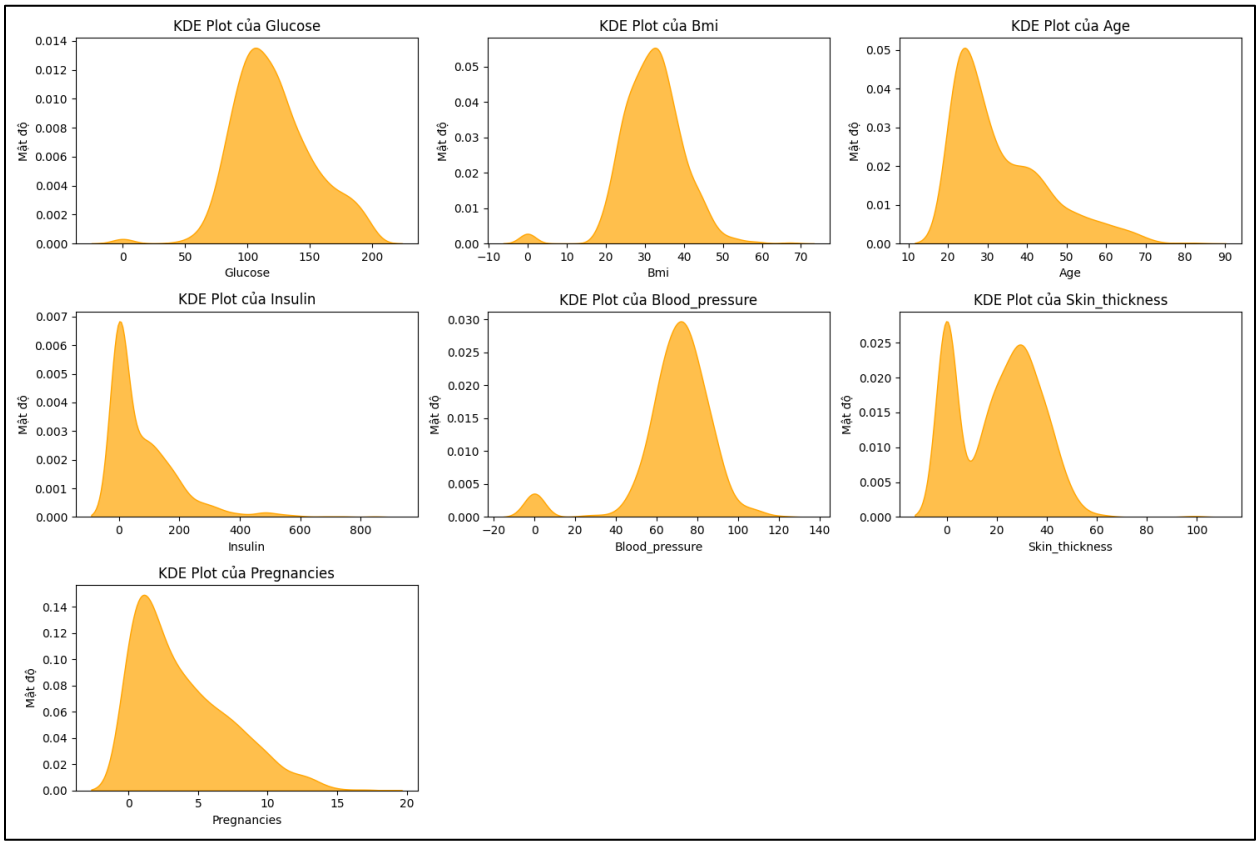


Chart 2. KDE Plot của các thuộc tính

Từ KDE Plot ta có nhận xét:

1. Glucose:

- Phân phối gần chuẩn, đỉnh quanh 100-120 mg/dL
- Ít outliers, phân phối tập trung → chất lượng data tốt

2. BMI:

- Phân phối chuẩn lệch phải nhẹ, đỉnh quanh 30-35
- Không có giá trị âm (KDE mở rộng do bandwidth)
- BMI trung bình khá cao (hơn 30) → nhiều đối tượng thừa cân

3. Age:

- **Phân phối lệch phải rõ rệt**
- Đa số bệnh nhân trẻ tuổi (20-35 tuổi)

- Phù hợp với đối tượng nghiên cứu (phụ nữ trong độ tuổi sinh sản)

4. Insulin:

- Phân phối lệch phải mạnh, peak ở giá trị thấp
- Có một số outliers ở giá trị cao (>400)
- Phản ánh đúng đặc điểm insulin: nhiều người có insulin thấp, ít người có insulin rất cao

5. Blood Pressure:

- Phân phối gần chuẩn, đỉnh quanh 70-80 mmHg
- Huyết áp trong ngưỡng bình thường

6. Skin Thickness:

- Phân phối lệch phải, tập trung ở giá trị thấp

7. Pregnancies:

- Phân phối lệch phải rõ rệt, peak ở giá trị thấp
- Đa số có 0-2 lần mang thai

Phân tích đơn biến (Univariate analysis)

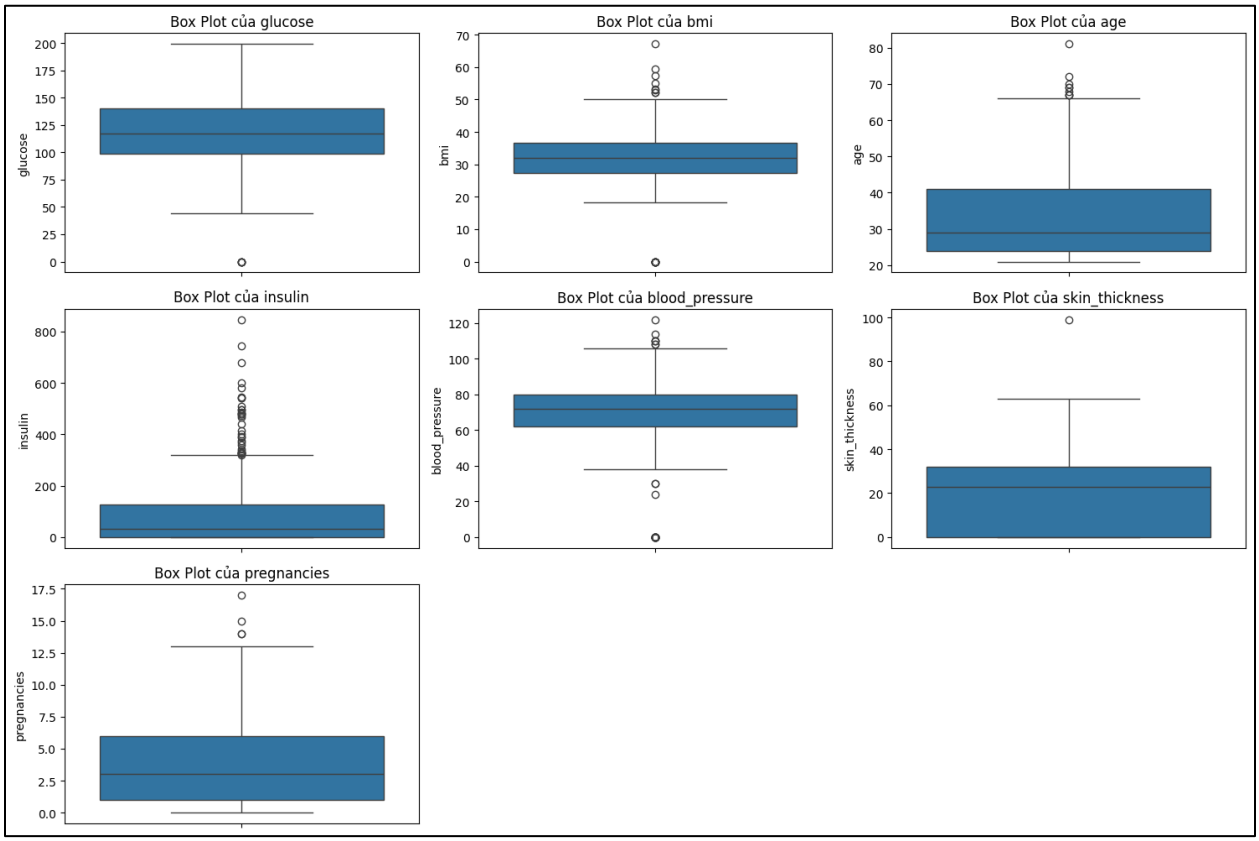


Chart 3. Box Plots của các thuộc tính

Từ các Box Plots ta có nhận xét:

1. Box Plot của Glucose

- **Phân tích:** Biến glucose có phân phối tương đối cân đối
- **Outlier:** Xuất hiện một điểm ngoại lai ở phía dưới
- **Nhận xét:** Hầu hết giá trị glucose tập trung trong khoảng 80-140 mg/dL

2. Box Plot của Insulin

- **Phân tích:** Phân phối lệch phải rõ rệt
- **Outlier:** Nhiều điểm ngoại lai ở phía trên
- **Nhận xét:** Đa số giá trị insulin dưới 200 µU/ml

3. Box Plot của Blood Pressure

- **Phân tích:** Phân phối gần như đối xứng

- **Outlier:** Ít điểm ngoại lai xuất hiện ở cả hai phía
- **Nhận xét:** Huyết áp trung bình khoảng 70-90 mmHg

4. Box Plot của BMI

- **Phân tích:** Phân phối lệch nhẹ về bên phải
- **Outlier:** Một số điểm ngoại lai ở giá trị cao
- **Nhận xét:** Chỉ số BMI chủ yếu trong khoảng 25-35 kg/m^2

5. Box Plot của Age

- **Phân tích:** Phân phối lệch phải
- **Outlier:** Có điểm ngoại lai ở độ tuổi cao
- **Nhận xét:** Độ tuổi phổ biến từ 20-40 tuổi

6. Box Plot của Skin Thickness

- **Phân tích:** Phân phối lệch phải
- **Outlier:** Ít điểm ngoại lai
- **Nhận xét:** Giá trị tập trung chủ yếu dưới 40 mm

7. Box Plot của Pregnancies

- **Phân tích:** Phân phối lệch phải mạnh
- **Outlier:** Ít điểm ngoại lai xuất hiện phía trên
- **Nhận xét:** Số lần mang thai chủ yếu từ 0-5

Phân tích đơn biến (Univariate analysis)

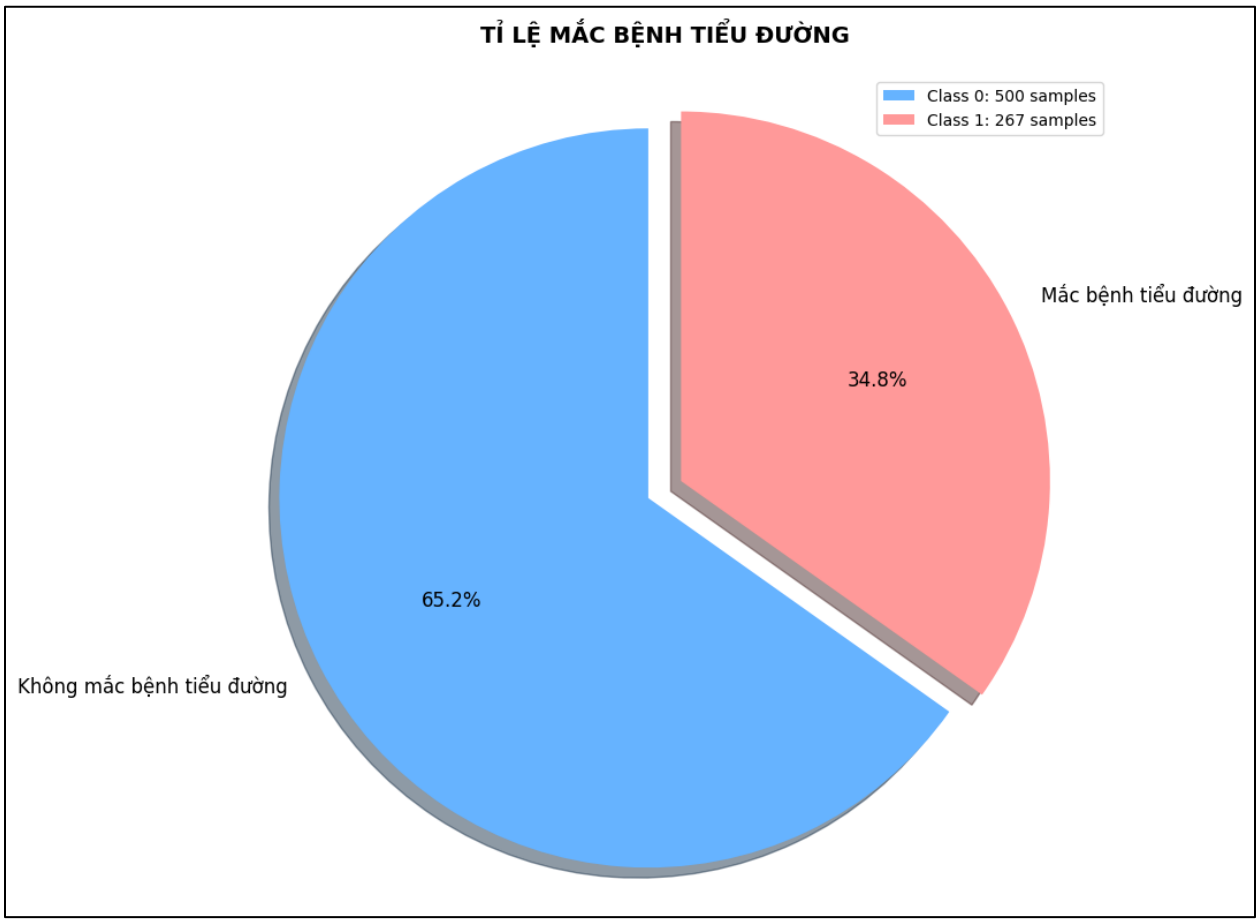


Chart 4. Tỉ lệ mắc bệnh tiểu đường

Từ biểu đồ tròn thể hiện tỉ lệ mắc bệnh tiểu đường trong tập dataset, ta có nhận xét sau:

Biểu đồ cho thấy sự phân bố của biến mục tiêu với 65.2% mẫu không mắc bệnh tiểu đường và 34.8% mắc bệnh. Dataset có sự chênh lệch đáng kể giữa hai lớp.

Phân tích nhị biến (Bivariate Analysis)

1. **Box Plot** của từng biến số, **nhóm theo** Outcome (Ví dụ: Box plot của Glucose, với 2 box: 1 box cho Outcome=0, 1 box cho Outcome=1).
2. **Heatmap** của ma trận tương quan, bao gồm cả biến Outcome.

Phân tích đa biến (Mulvariate Analysis)

1. **Scatter Plot** giữa 2 biến quan trọng nhất (ví dụ: **Glucose vs BMI**), **tô màu theo** Outcome.
2. **Pair Plot** (nếu số lượng biến không quá nhiều) để xem nhanh các mối quan hệ.

Phân tích theo Nhóm Tuổi (Age Group wise analysis)

Mục tiêu: Tuổi tác là yếu tố nguy cơ chính. Phân tích theo nhóm sẽ rõ hơn so với tuổi liên tục.

Cách làm: Chia biến Age thành các nhóm (vd: <30, 30-45, 45-60, >60).

Câu hỏi phân tích:

- **Tỷ lệ mắc bệnh (Outcome=1) ở nhóm tuổi nào là cao nhất?** (Kỳ vọng: tỷ lệ tăng dần theo tuổi).
- Ở nhóm tuổi trẻ (<30), những người vẫn mắc bệnh có đặc điểm chung gì? (Glucose rất cao? BMI rất cao? Tiền sử gia đình nặng?).

Phân tích theo Mức độ Béo phì (BMI Category analysis)

Mục tiêu: Đánh giá tác động của cân nặng một cách trực quan.

Cách làm: Chia BMI thành các hạng: Thiếu cân (<18.5), Bình thường (18.5-24.9), Thừa cân (25-29.9), Béo phì (>=30).

Câu hỏi phân tích:

- **Tỷ lệ mắc bệnh có tăng rõ rệt theo các cấp độ BMI không?**
- Trong nhóm BMI "Bình thường", có bao nhiêu % vẫn mắc bệnh? Đặc điểm của họ là gì? (Có thể do di truyền - DiabetesPedigreeFunction cao).

Phân tích theo Ngưỡng Đường huyết (Glucose Threshold analysis)

Mục tiêu: Hiểu rõ hơn về yếu tố quyết định trực tiếp nhất.

Cách làm: Phân loại Glucose theo ngưỡng chẩn đoán lâm sàng (dựa trên Bảng 1 bạn có): Bình thường, Tiền tiểu đường (IFG/IGT), Tiểu đường.

Câu hỏi phân tích:

- Trong số những người được mô hình dự đoán là mắc bệnh (Outcome=1), có bao nhiêu % thực sự đã ở ngưỡng đường huyết chẩn đoán tiểu đường?
- Có trường hợp nào Glucose ở mức "bình thường" hoặc "tiền tiểu đường" nhưng vẫn được chẩn đoán mắc bệnh (Outcome=1) không? Nếu có, các yếu tố khác (như Insulin, Age) của họ thế nào? → Đây là những ca "khó" mà mô hình cần học tốt.

Phân tích theo ngưỡng đường huyết(Glucose Threshold analysis)

Dữ liệu thống kê theo cột Glucose

Glucose_Level	Tổng mẫu	Số mắc bệnh	Tỷ lệ mắc bệnh
Bình thường(<140)	571	133	23.29%
Tiền tiểu đường(140- 199)	197	135	68.53%

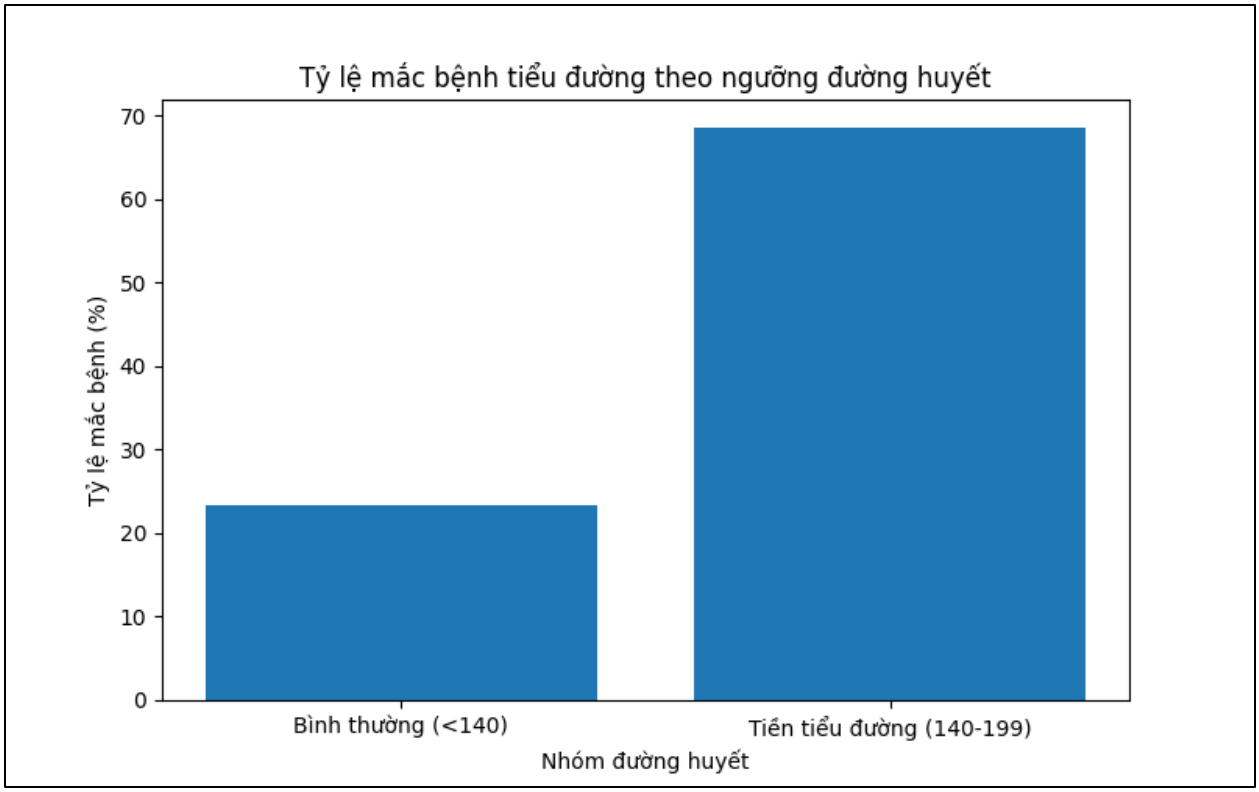


Chart 5. Tỷ lệ mắc bệnh tiểu đường theo ngưỡng đường huyết

Từ biểu đồ trên ta có thể thấy:

- Những nhóm người có nồng độ đường huyết trong máu cao(140-199) thì có nguy cơ mắc bệnh tiểu đường cao.
- Ngược lại những người có nồng độ đường huyết trong máu ở mức bình thường thì tỉ lệ mắc bệnh ít hơn.

Những người có nồng độ Glucose trong máu thấp mà vẫn bị thì thường lên quan đến các yếu tố Age , BMI,pdf, Pregnancies. Bảng thống kê :

```
Tổng mẫu trong nhóm Glucose < 140: 571
Outcome
0    438
1    133
Name: count, dtype: int64

So sánh từng biến (nhóm Glucose thấp):
                  feature  mean_out0  mean_out1           test  p_value
                      Age     30.395     35.361 Mann-Whitney U 0.000000
                      BMI     29.911     34.491 Mann-Whitney U 0.000000
              Pregnancies      3.205      4.835 Mann-Whitney U 0.000027
DiabetesPedigreeFunction      0.420      0.552 Mann-Whitney U 0.000093
           BloodPressure     67.153     68.662 Mann-Whitney U 0.018869
           SkinThickness     19.299     21.008 Mann-Whitney U 0.113893
                 Insulin     58.384     67.594 Mann-Whitney U 0.902003
```

Phân tích theo Tiền sử Gia đình (Pedigree Function analysis)

Mục tiêu: Định lượng ảnh hưởng của yếu tố di truyền.

Cách làm: Chia DiabetesPedigreeFunction thành các nhóm: Thấp, Trung bình, Cao (dựa trên phân vùng hoặc giá trị cắt phù hợp).

Câu hỏi phân tích:

- **Với những người có cùng mức BMI hoặc tuổi tác, nhóm có tiền sử gia đình (Pedigree) nặng hơn có tỷ lệ mắc bệnh cao hơn không?**
- Yếu tố di truyền có mạnh đến mức "lấn át" các yếu tố nguy cơ khác không? (Ví dụ: Một người trẻ, gầy nhưng có pedigree rất cao thì nguy cơ thế nào?).

Phân tích theo Tiền sử Gia đình (Pedigree Function analysis)

Dữ liệu thống kê theo cột DiabetesPedigreeFunction(dpf)

Family_History	Tổng mẫu	Số mắc bệnh	Tỉ lệ mắc bệnh
Cao (>=1.0)	51	29	56.86%
Trung bình(0.5 - 1)	491	142	28.92%
Thấp (<0.5)	226	97	42.92%

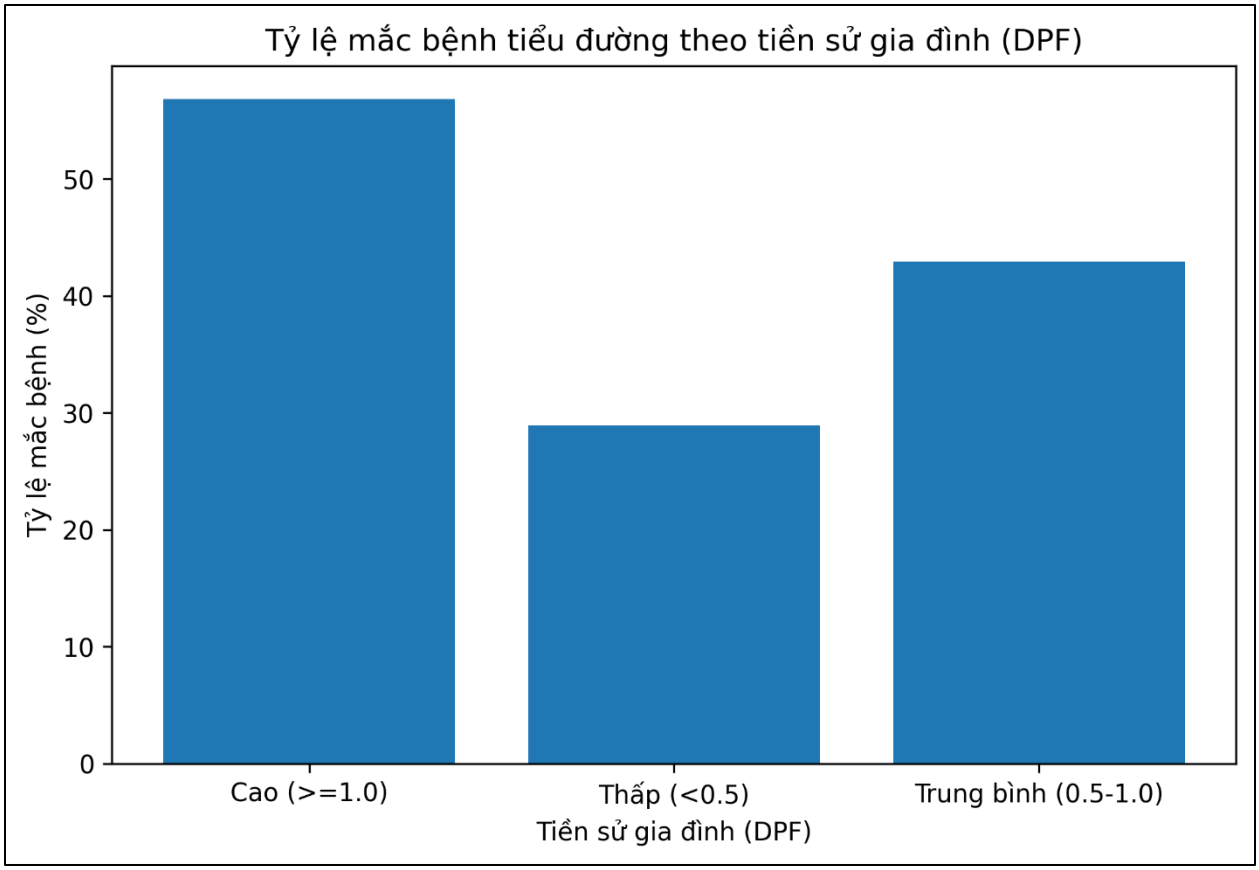


Chart 6. Tỷ lệ mắc bệnh tiểu đường theo tiền sử gia đình

Từ biểu đồ ta có thể thấy:

- Những người có tiền sử gia đình mắc bệnh cao(>1.0) thì có tỉ lệ mắc bệnh cao
- Những người có tiền sử gia đình mắc bệnh ở mức trung bình thì có tỉ lệ mắc bệnh ở mức trung bình
- Những người có tiền sử gia đình mắc bệnh thấp thì tỉ lệ mắc bệnh thấp

Bản đồ nhiệt tương quan (Correlation heatmap)

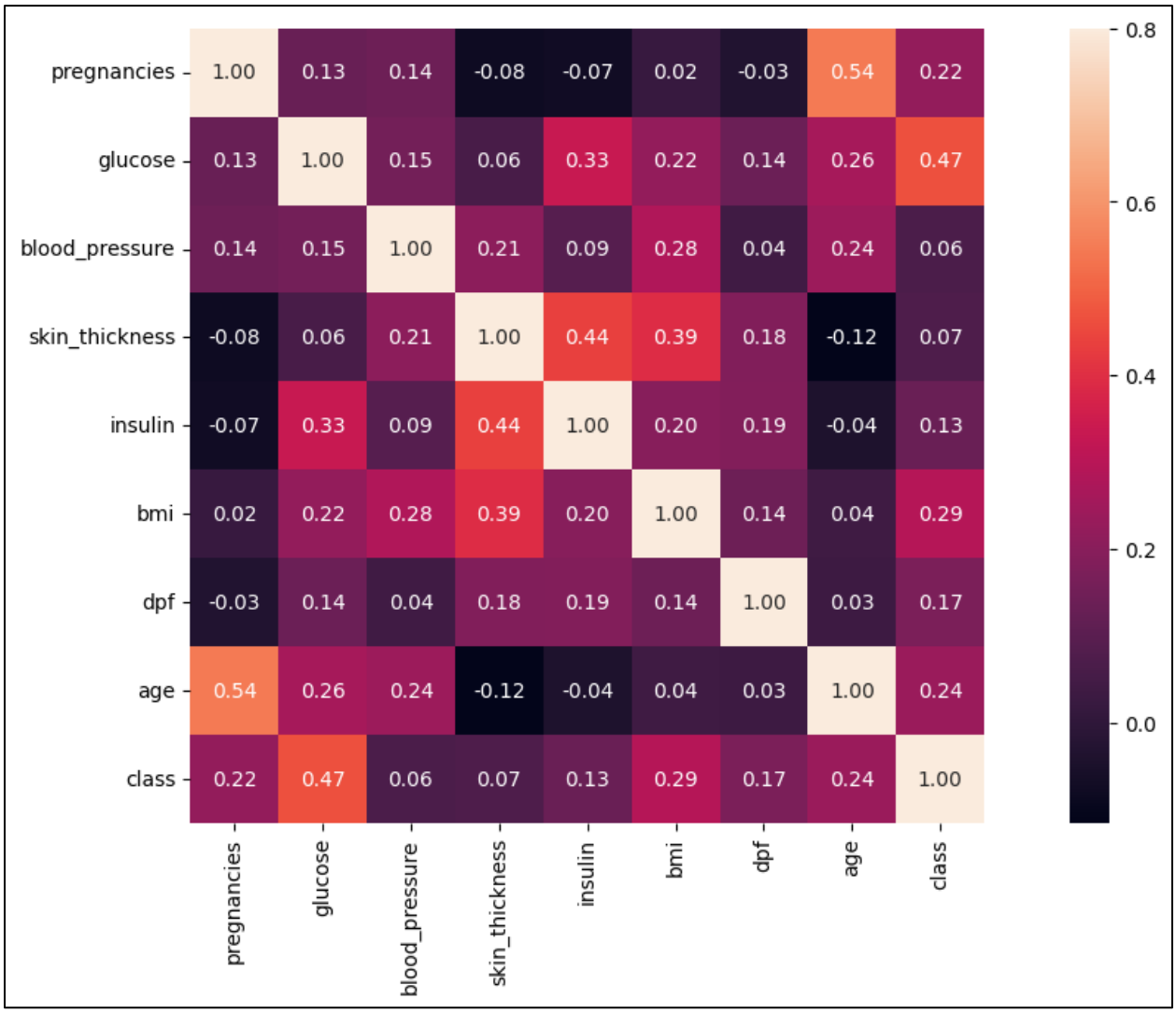


Chart 7. Biểu đồ nhiệt tương quan

Phân tích chi tiết các tương quan:

Một số cặp tương quan mạnh (|r| > 0.4):

- **Glucose vs Class (0.47)**: Đây là tương quan quan trọng nhất, phù hợp với y học vì glucose máu là chỉ số chính chẩn đoán tiểu đường
- **Pregnancies vs Age (0.54)**: Hoàn toàn hợp lý - phụ nữ càng lớn tuổi càng có nhiều khả năng đã mang thai nhiều lần
- **Skin thickness vs Insulin (0.44)**: Có thể phản ánh mối liên hệ giữa độ dày da (liên quan đến mỡ dưới da) và kháng insulin

Một số cặp tương quan trung bình (0.2 < |r| < 0.4):

- **Glucose vs Insulin (0.33)**: Insulin điều hòa glucose trong máu
- **Glucose vs BMI (0.22)**: BMI cao thường đi kèm glucose cao
- **Blood pressure vs BMI (0.28)**: Huyết áp và cân nặng có mối liên hệ
- **Skin thickness vs BMI (0.39)**: Độ dày da liên quan đến chỉ số khối cơ thể

Một số cặp tương quan yếu (|r| < 0.2):

Hầu hết các cặp biến còn lại có tương quan rất yếu hoặc gần như không có

- **Pregnancies vs BMI**: 0.02 (gần như không tương quan)
- **Pregnancies vs DPF**: -0.03 (gần như không tương quan)
- **Blood pressure vs DPF**: 0.04 (rất yếu)
- **Insulin vs Age**: -0.04 (rất yếu)

Kết luận (Conclusion)