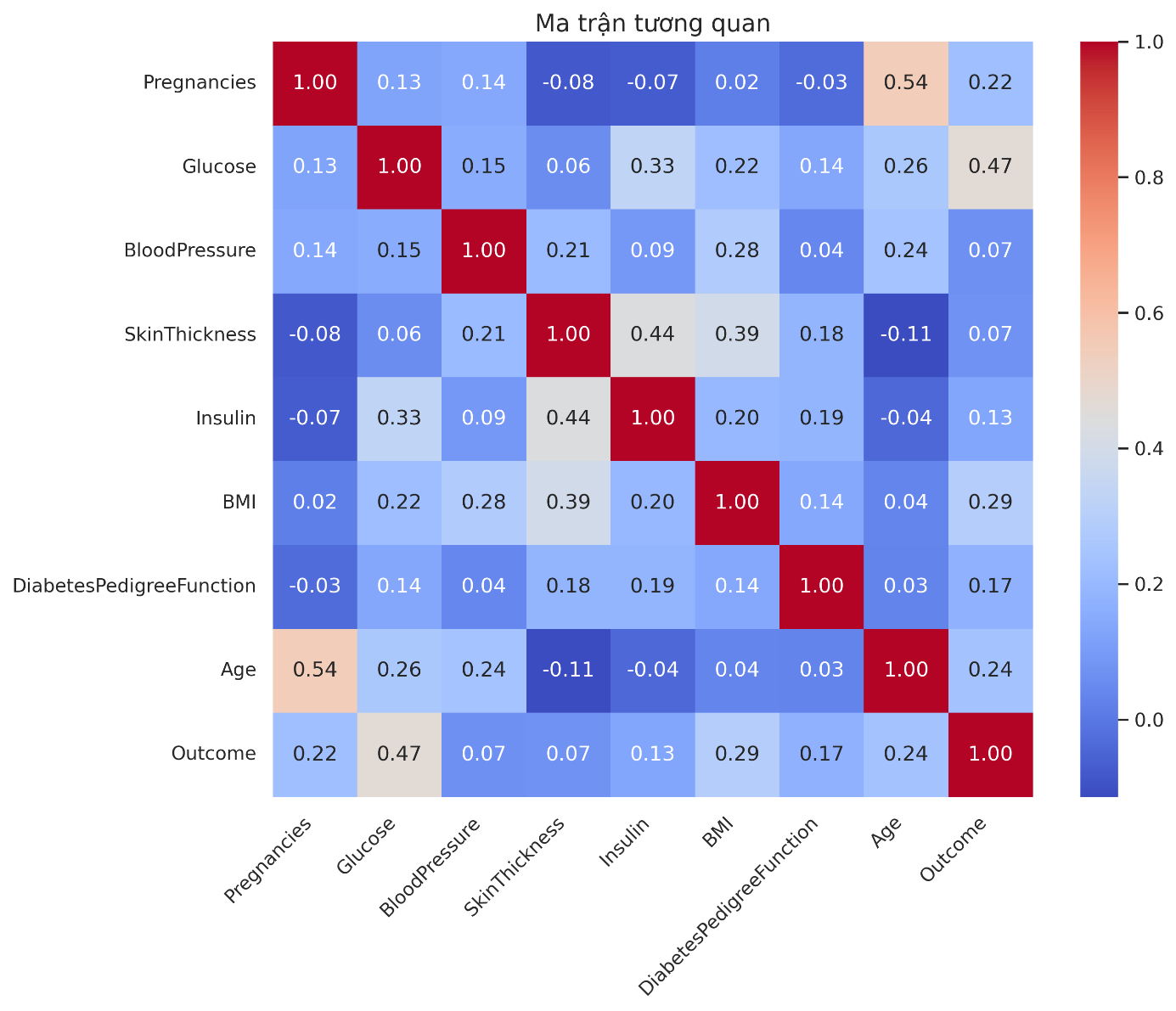
**5. PHÂN TÍCH TƯƠNG QUAN VÀ MỐI QUAN HỆ**

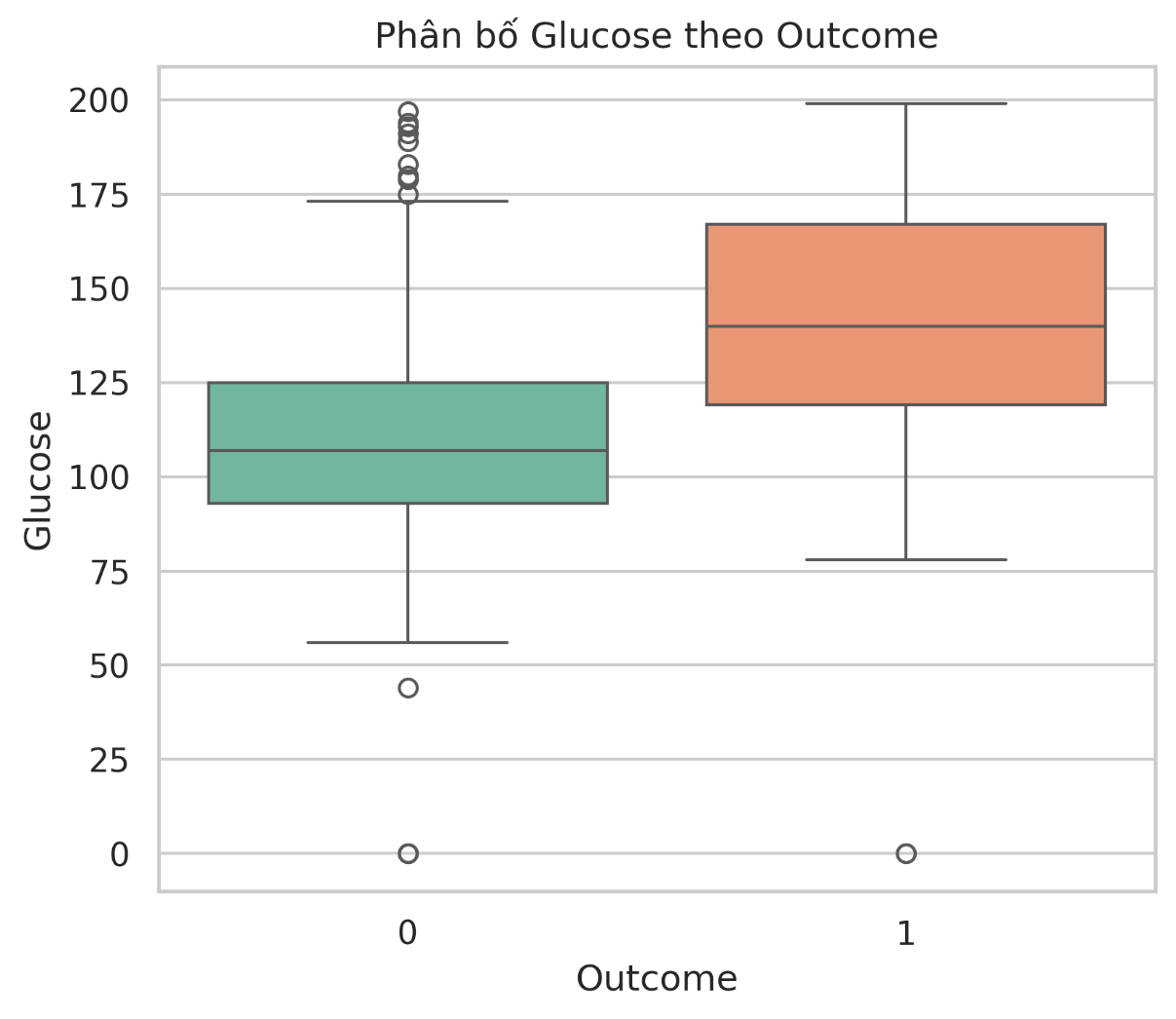
**5.1 Ma trận tương quan**



Hình 1: Ma trận tương quan

Glucose và Outcome có tương quan dương khá mạnh → Chỉ số đường huyết cao liên quan đến tiểu đường.

**5.2 Phân tích cặp biến quan trọng**



Hình 2: Sự phân bố của Glucose theo Outcome

Mối tương quan mạnh nhất với r = 0.47.

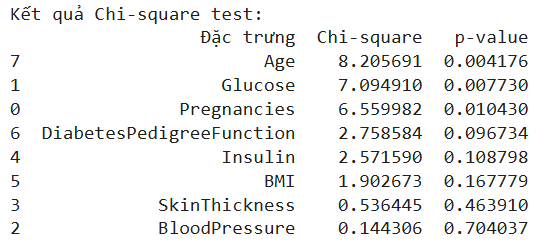
Nhóm mắc bệnh có glucose trung bình = 141mg/dL so với 109mg/dL.

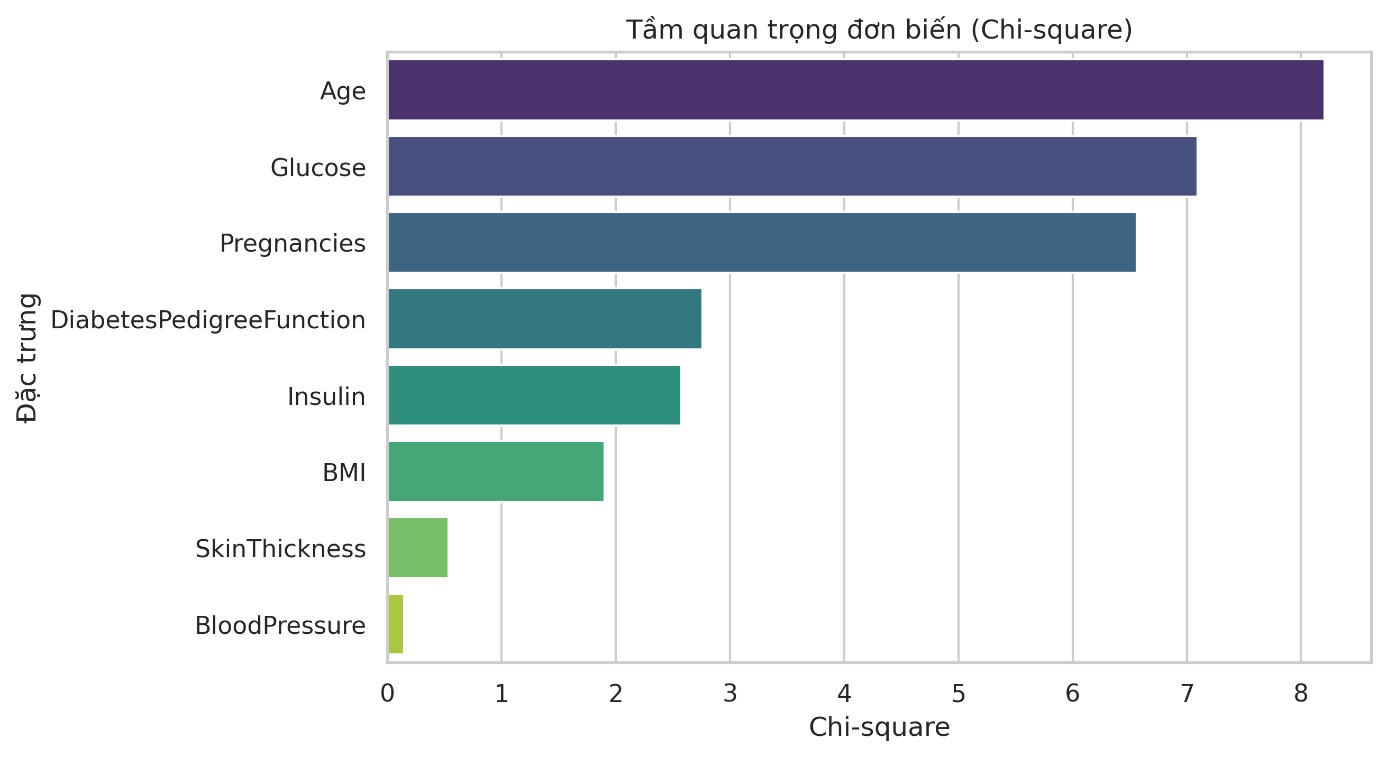
Nguy cơ tiểu đường tăng mạnh khi Glucose > 126mg/dL (theo ngưỡng chẩn đoán lâm sàng).

**5.3 Phân tích tầm quan trọng của đặc trưng**

**5.3.1 Tầm quan trọng đơn biến**

Bảng 1: Tập C-square

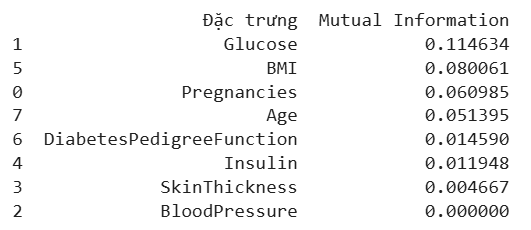




Hình 3: Biểu đồ C-square

**5.3.2 Điểm thông tin tương hỗ**

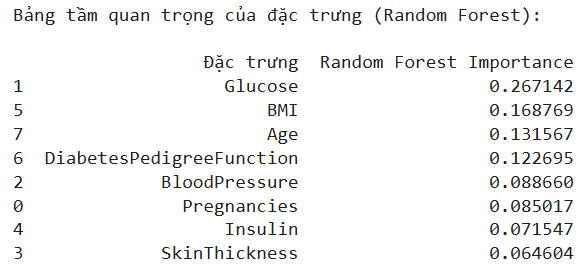
Bảng 2: Bảng thông tin tương hỗ

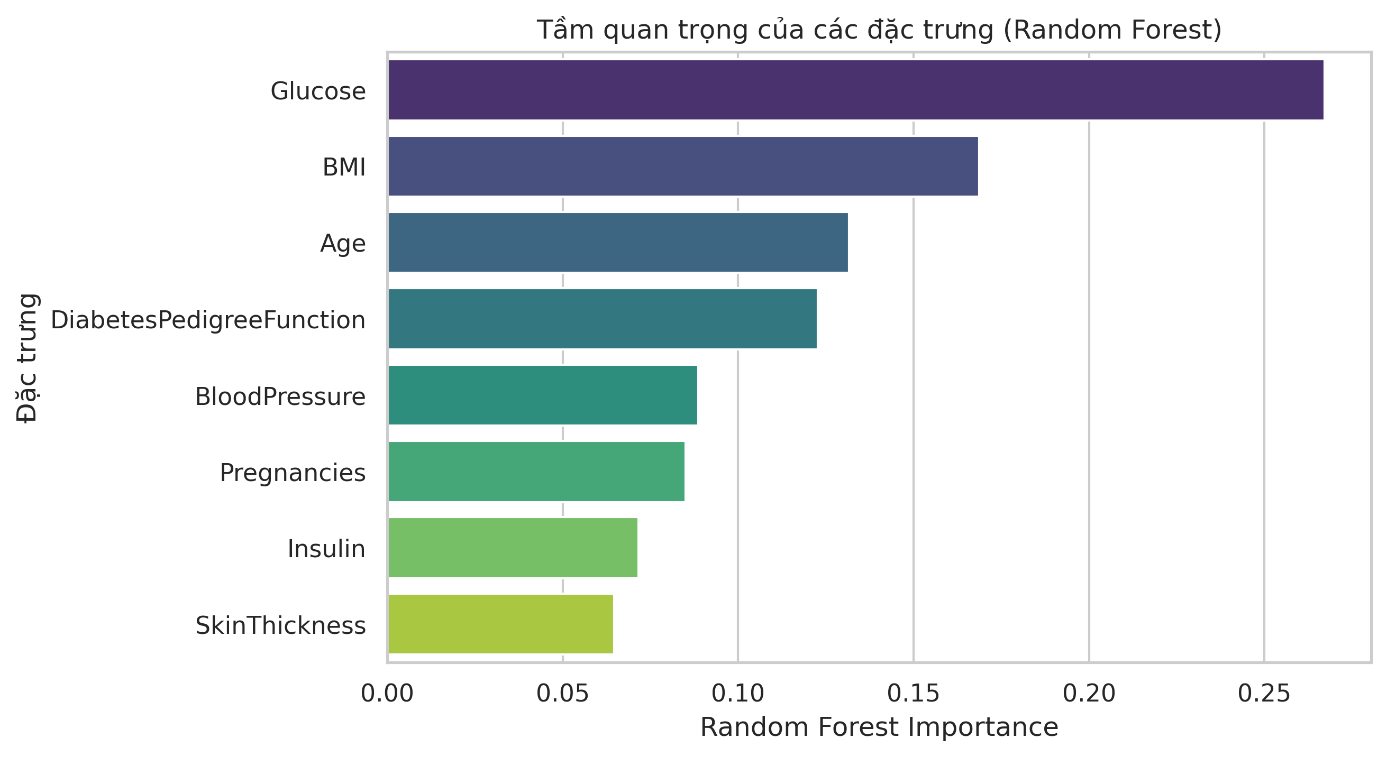


Chỉ số Glucose là cao nhất

**5.3.3 Tầm quan trọng từ Random Forest**

Bảng 3: Bảng tầm quan trọng của đặc trưng

****

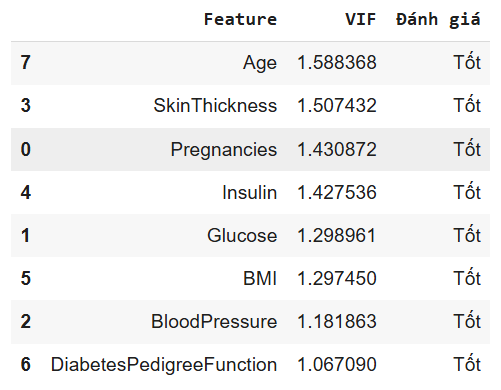


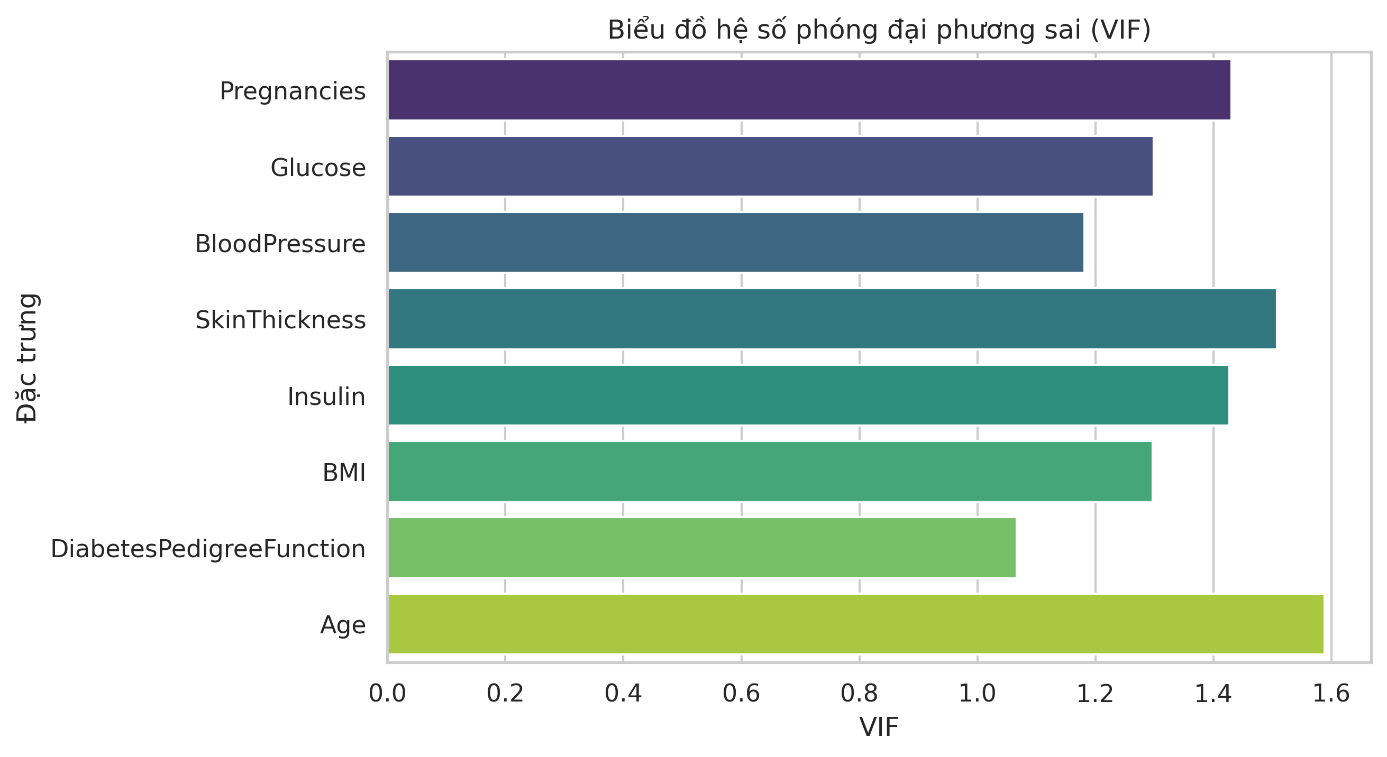
Hình 4: Biểu đồ Random Forest

**5.4 Phát hiện đa cộng tuyến**

**5.4.1 Hệ số phóng đại phương sai (VIF)**

Bảng 4: Bảng hệ số phóng đại phương sai

****



Hình 5: Biểu đồ phương sai VIF

**5.4.2 Phân tích chỉ số điều kiện**

Tất cả các biến đều có VIF < 2, nghĩa là không có hiện tượng đa cộng tuyến nghiêm trọng.

Nhìn chung, dữ liệu ở mức ổn định để sử dụng trong mô hình hồi quy / phân loại.

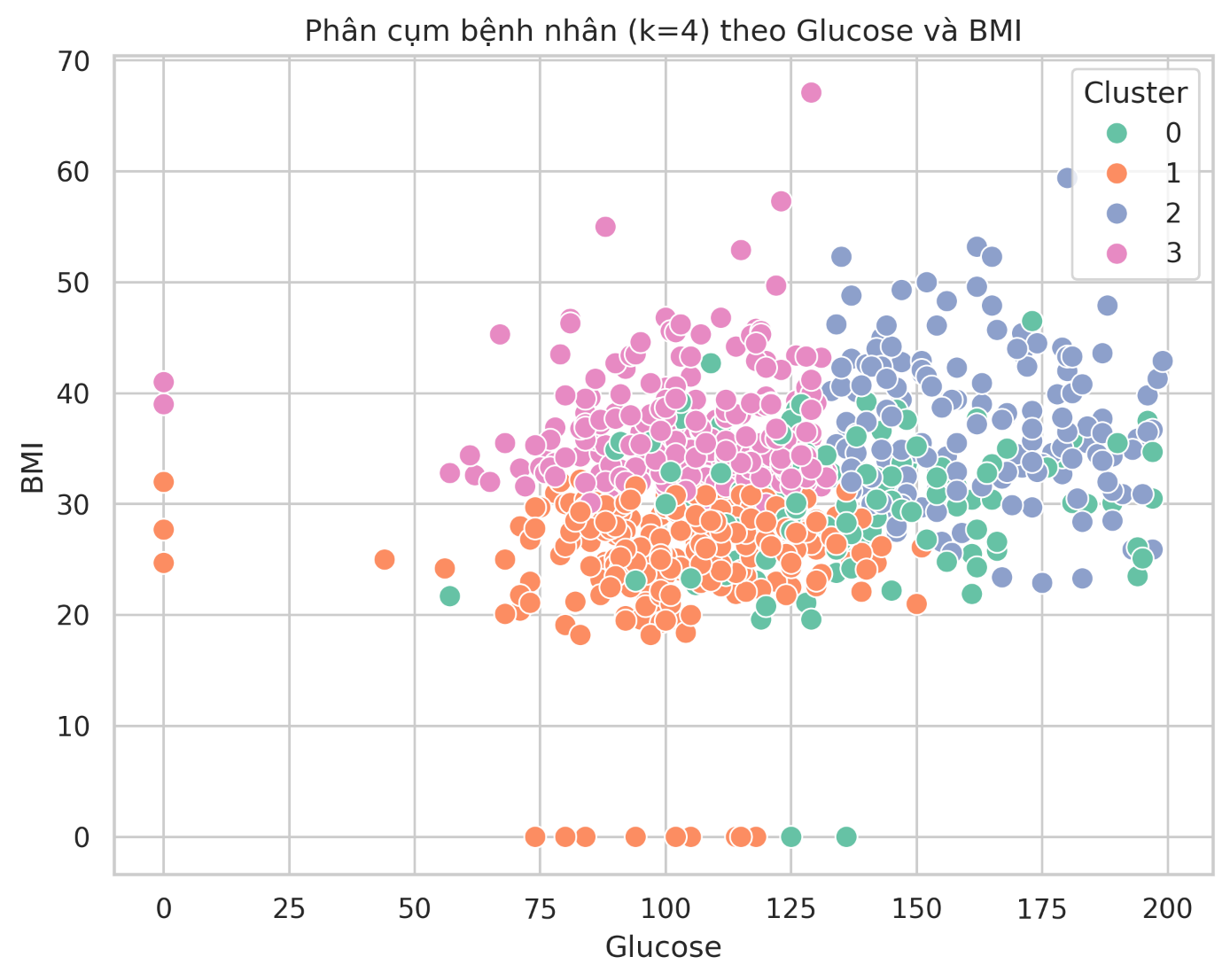
Do đó, có thể giữ lại toàn bộ 8 biến gốc trong phân tích tiếp theo.

**6. PHÂN TÍCH NÂNG CAO**

**6.1 PHÂN TÍCH PHÂN NHÓM (PHÂN CỤM)**

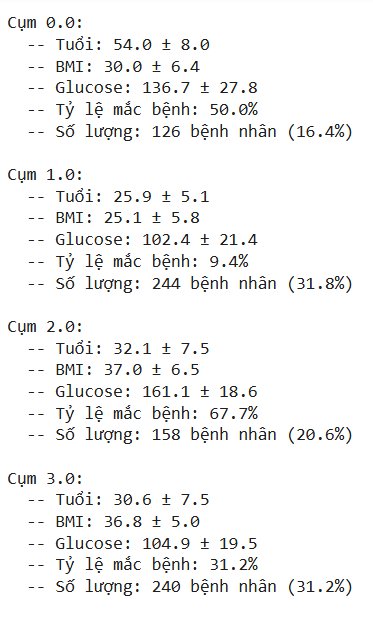
**6.1.1 Xác định các nhóm bệnh nhân tiềm ẩn**

**Phân cụm K-means (k=4):**

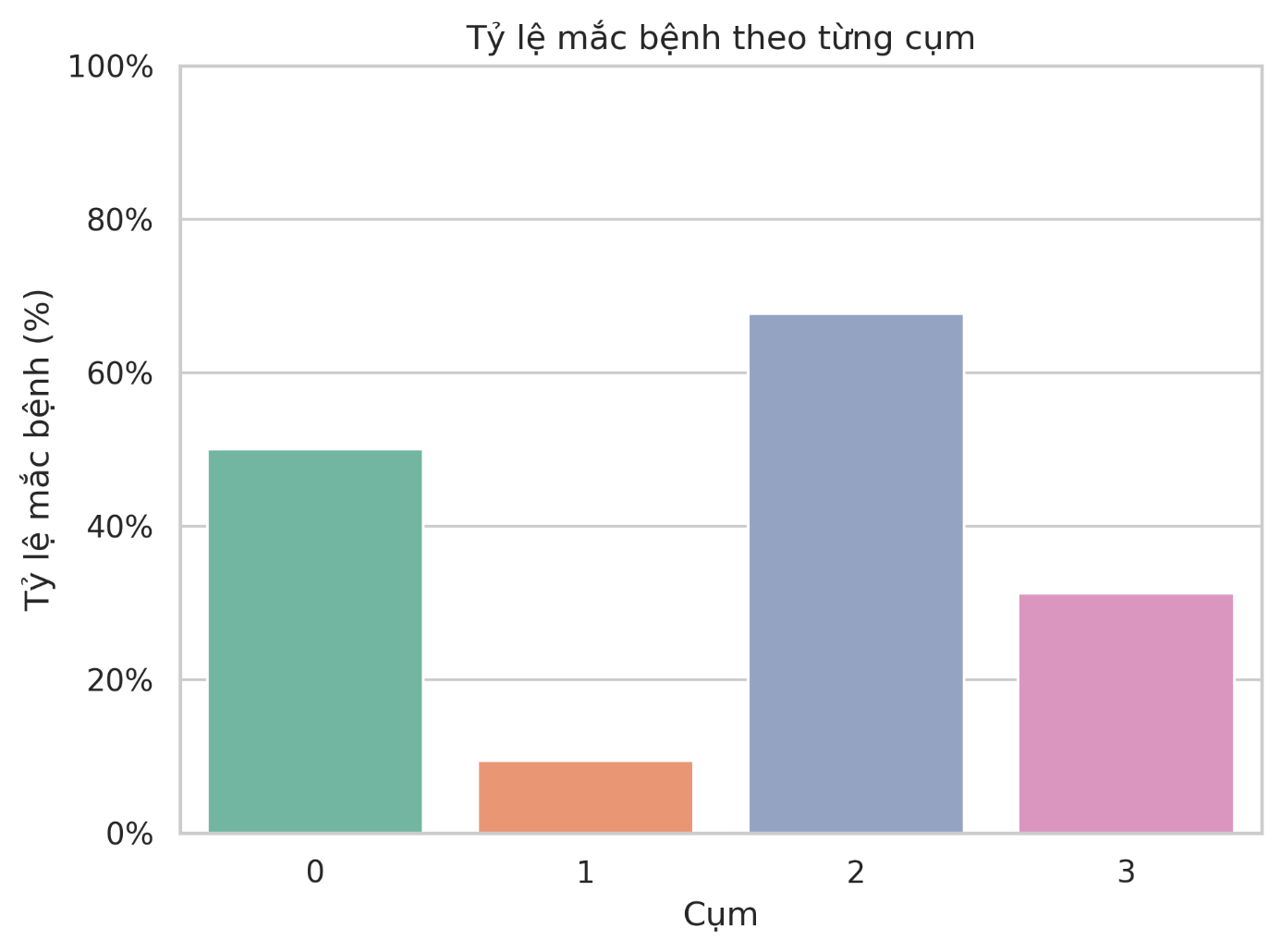


Hình 6: Xác định nhóm bệnh nhân tiềm ẩn

Bảng 5: Bảng phân cụm bệnh nhân

****

**6.1.2 Đặc điểm của từng cụm**

****

Hình 7: Thống kê tỷ lệ mắc bệnh của các cụm

**Cụm 0** **– "Cao tuổi, Nguy cơ trung bình":**

Chủ yếu là bệnh nhân lớn tuổi (tuổi trung bình ~54)

BMI ở mức thừa cân (~30)

Glucose cao (~137)

Tỷ lệ mắc bệnh ~50%

Cần theo dõi sức khỏe định kỳ, chú trọng quản lý đường huyết

**Cụm 1** **– "Trẻ & Khỏe mạnh":**

Nhóm trẻ nhất (tuổi ~26)

BMI thấp (~25)

Glucose bình thường (~102).

Tỷ lệ mắc bệnh thấp nhất ~9%

Chỉ cần duy trì lối sống lành mạnh

**Cụm 2 – "Nguy cơ cao, Béo phì & Tăng đường huyết":**

Tuổi trung bình ~32

BMI rất cao (~37 → béo phì nặng)

Glucose cao nhất (~161)

Tỷ lệ mắc bệnh cao nhất (~68%)

Cần can thiệp y tế và thay đổi lối sống mạnh mẽ

**Cụm 3 – "Béo phì, Nguy cơ tiềm ẩn":**

Tuổi trung bình ~31

BMI cao (~37 → béo phì)

Glucose ở mức bình thường (~105)

Tỷ lệ mắc bệnh trung bình (~31%)

Béo phì nhưng chưa biểu hiện rõrối loạn glucose, cần quản lý cân nặng, chế độ ăn và vận động để phòng ngừa.

**6.1.3 Giải thích lâm sàng**

**Cụm 0:** Cao tuổi, glucose tăng → nguy cơ trung bình.

**Cụm 1:** Khỏe mạnh → nguy cơ thấp.

**Cụm 2:** Béo phì + glucose tăng cao→ nguy cơ cao nhất, cần điều trị.

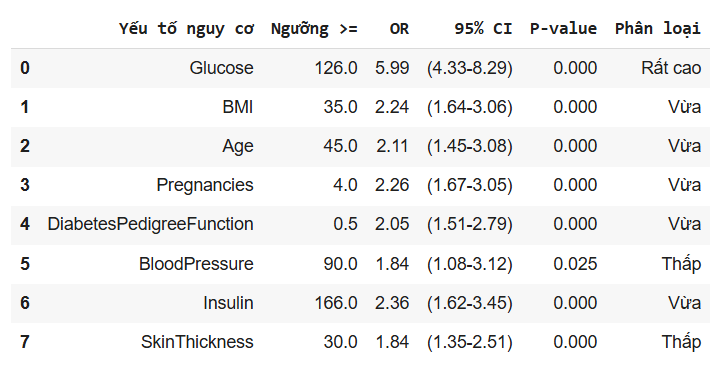
**Cụm 3:** Béo phì nhưng glucose bình thường → nguy cơ tiềm ẩn.

**6.2 PHÂN TÍCH YẾU TỐ NGUY CƠ**

**6.2.1 Xếp hạng các yếu tố nguy cơ**

**Phân tích tỷ số chênh (Odds Ratio):**

Bảng 6: Bảng tỷ số chênh Odds Ratio

****

**6.2.2 Tạo hệ thống chấm điểm nguy cơ**

**Thang điểm nguy cơ đái tháo đường Pima:**

Bảng 7: Bảng thang điểm nguy cơ

**A screenshot of a table

AI-generated content may be incorrect.**

**Phân loại nguy cơ:**

Bảng 8: Bảng phân loại nguy cơ bệnh

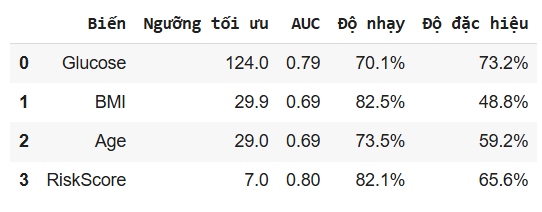
A screenshot of a computer

AI-generated content may be incorrect.

**6.2.3 Phân tích ngưỡng cho sàng lọc**

**Phân tích ROC cho các biến chính:**

Bảng 9: Bảng ROC cho các biến

****

**6.3 SO SÁNH VỚI TIÊU CHUẨN CHẨN ĐOÁN**

**6.3.1 Tiêu chí chẩn đoán của WHO/ADA**

**Hiệu suất tiêu chí WHO 1999:**

**Bộ dữ liệu so với thực tế:**

**6.3.2 Độ chính xác của các giá trị cắt**

**Ngưỡng điều chỉnh cho dân số Pima:**

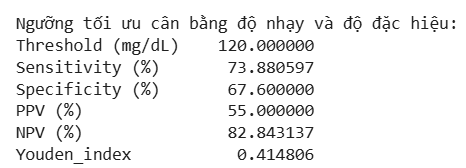
Bảng 10: Độ chính xác cho giá trị cắt

**A screenshot of a computer screen

AI-generated content may be incorrect.**

**6.3.3 Cân bằng giữa độ nhạy và độ đặc hiệu**

Bảng 11: Bảng ngưỡng tối ưu cân bằng

****

Sensitivity (%) → % bệnh nhân thật sự mắc bệnh được phát hiện.

Specificity (%) → % bệnh nhân không mắc bệnh được loại đúng.

PPV (%) → xác suất dương tính đúng khi xét nghiệm ≥ ngưỡng.

NPV (%) → xác suất âm tính đúng khi xét nghiệm < ngưỡng.

Youden\_index = Sensitivity + Specificity - 1 → tìm ngưỡng tối ưu cân bằng độ nhạy và độ đặc hiệu.

**Cân nhắc lâm sàng:**

Độ nhạy cao (≥90%): Cho sàng lọc - Ngưỡng ≥100 mg/dL

Độ đặc hiệu cao (≥90%): Cho xác nhận - Ngưỡng 150–160 mg/dL

Cân bằng: Cho quyết định lâm sàng - Ngưỡng 120–130 mg/dL