Phân tích dữ liệu về bệnh đái tháo đường

Áp dụng báo cáo WHO cho phân tích khám phá dữ liệu

Nguyễn Hoàng Phúc

Mục lục

Giới thiệu

Mục tiêu: Sử dụng báo cáo của Tổ chức Y tế Thế giới (WHO) về "Định nghĩa, Chẩn đoán và Phân loại Bệnh Đái tháo đường" làm nền tảng lý thuyết.

Mục đích: Thực hiện phân tích khám phá dữ liệu (EDA) trên tập dữ liệu Pima Indians Diabetes để:

- Khám phá các yếu tố nguy cơ chính liên quan đến bệnh đái tháo đường.
- Trực quan hóa mối quan hệ giữa các yếu tố này và kết quả chẩn đoán.
- Xác nhận tính nhất quán giữa dữ liệu thực tế và các tiêu chí chẩn đoán của WHO.

Data Summary

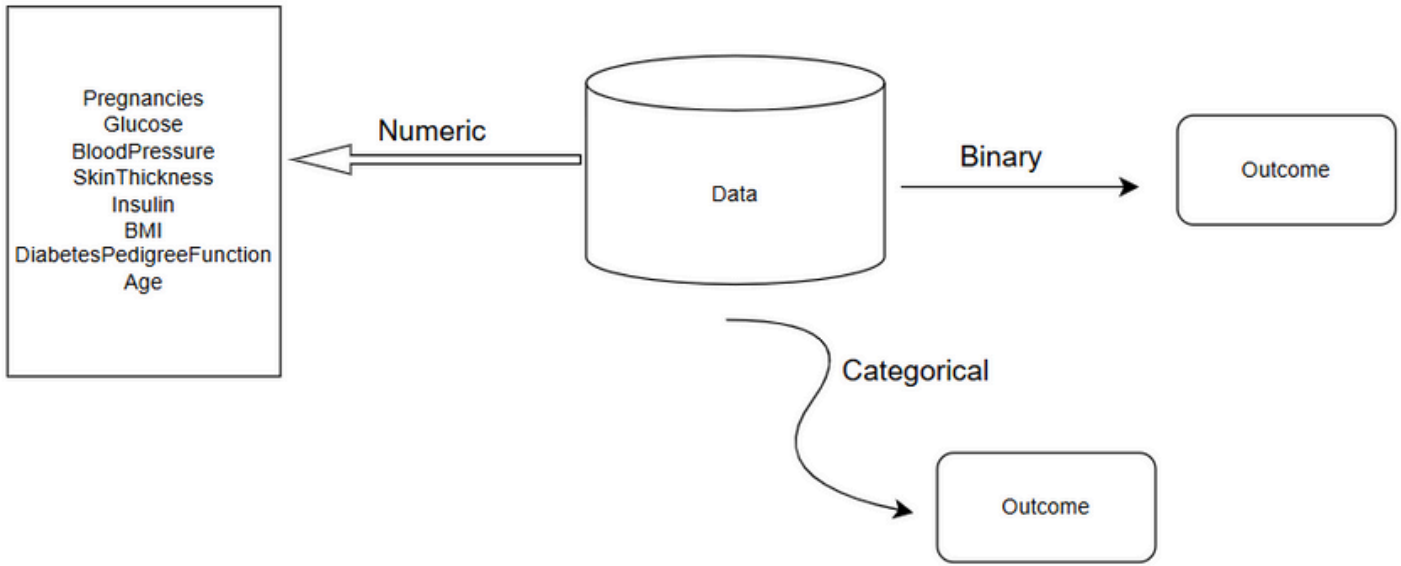
Bộ dữ liệu Pima Indians Diabetes từ một nghiên cứu y tế được thu thập bởi Viện Tiểu đường và Các bệnh Tiêu hóa và Thận Quốc gia.

- Pregnancies: Số lần mang thai của bệnh nhân (số nguyên).
- Glucose: Nồng độ glucose trong huyết tương.
- BloodPressure: Huyết áp tâm trương.
- SkinThickness: Độ dày nếp gấp da cơ tam đầu.
- Insulin: Nồng độ insulin trong huyết thanh.

Data Summary(contd..)

- BMI: Chỉ số khối cơ thể.
- DiabetesPedigreeFunction: Chức năng phả hệ tiểu đường, một chỉ số đánh giá khả năng di truyền của bệnh.
- Age: Tuổi của bệnh nhân.
- Outcome: Biến mục tiêu (0 = Không mắc bệnh, 1 = Mắc bệnh tiểu đường).

Data Summary

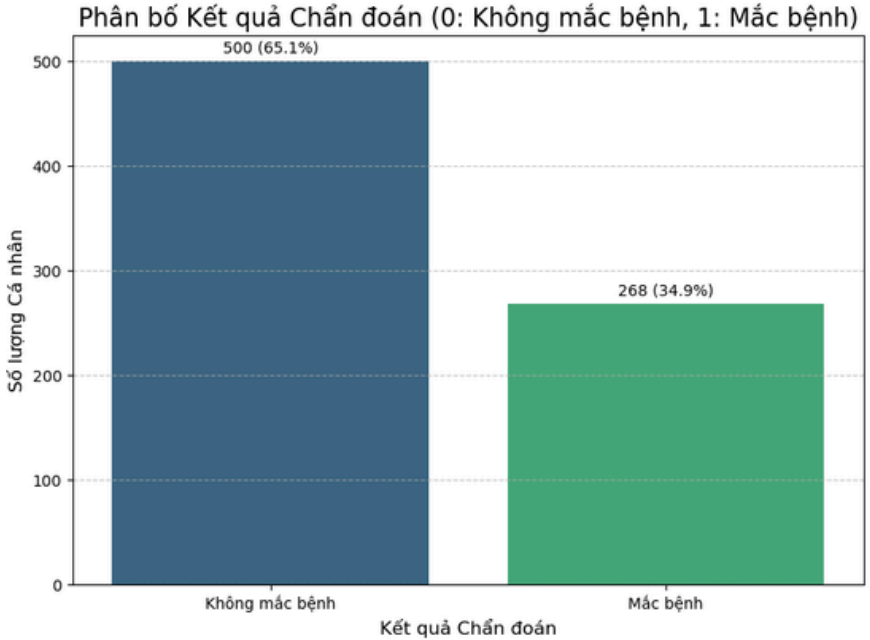


EDA - Tổng quan & Tỷ lệ Mắc bệnh

Tổng quan về EDA và Tỷ lệ bệnh Đái tháo đường

Biểu đồ thanh Outcome

- Hiển thị số lượng/tỷ lệ phần trăm người mắc bệnh (Outcome=1) và không mắc bệnh (Outcome=0).
- Trong tổng số 768 cá nhân, khoảng [~65%] không mắc bệnh và [~35%] được chẩn đoán mắc bệnh đái tháo đường. Điều này cho thấy sự mất cân bằng trong dữ liệu và một vấn đề sức khỏe đáng kể trong nhóm dân số nghiên cứu.

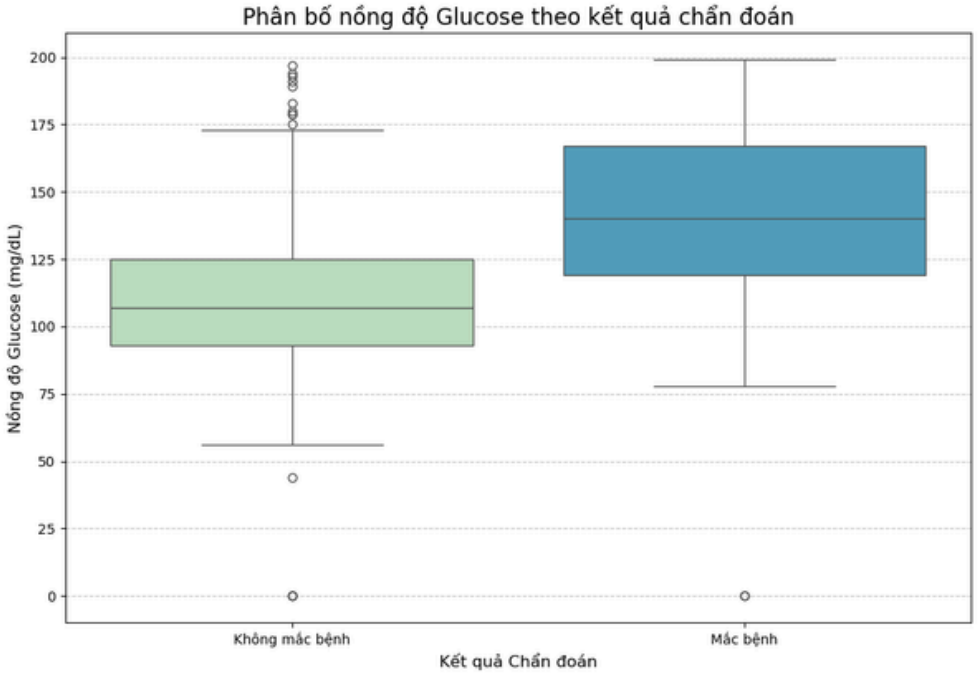


EDA - Phân tích Đơn biến (Phân bố các yếu tố nguy cơ)

Phân bố của các yếu tố nguy cơ chính

- Glucose có phân bố hơi lệch phải.
- Phần lớn cá nhân có nồng độ glucose bình thường hoặc thấp, nhưng có một số ít cá nhân có nồng độ rất cao, có thể là dấu hiệu của bệnh tiểu đường.

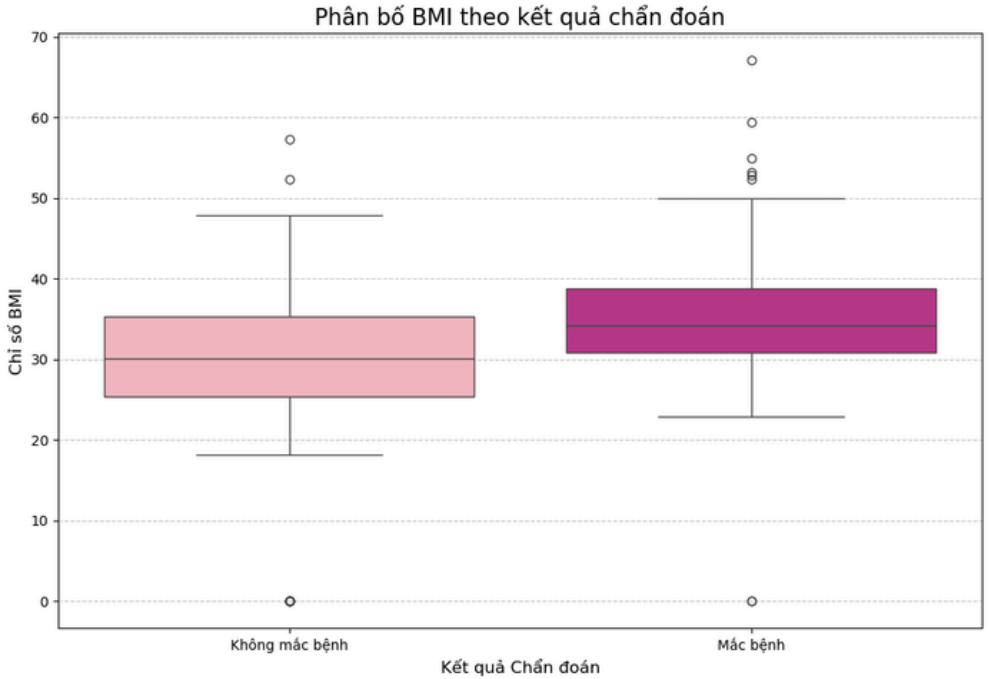
Biểu đồ Histogram/KDE Plot cho Glucose



EDA - Phân tích Đơn biến (Phân bố các yếu tố nguy cơ)

- BMI có phân bố lệch phải.
- Trung vị BMI cao hơn ở nhóm mắc bệnh
- có sự khác biệt về trung vị, nhưng hai boxplot chồng lấp lên nhau một cách đáng kể
- Phần lớn người không mắc bệnh nằm trong khoảng BMI từ khoảng 25 đến 35, tương ứng với mức thừa cân hoặc béo phì độ 1.
- Phân bố BMI của nhóm này tập trung ở mức cao hơn, với phần lớn dữ liệu nằm trong khoảng từ 30 trở lên, thường là béo phì.

Biểu đồ Histogram/KDE Plot cho BMI

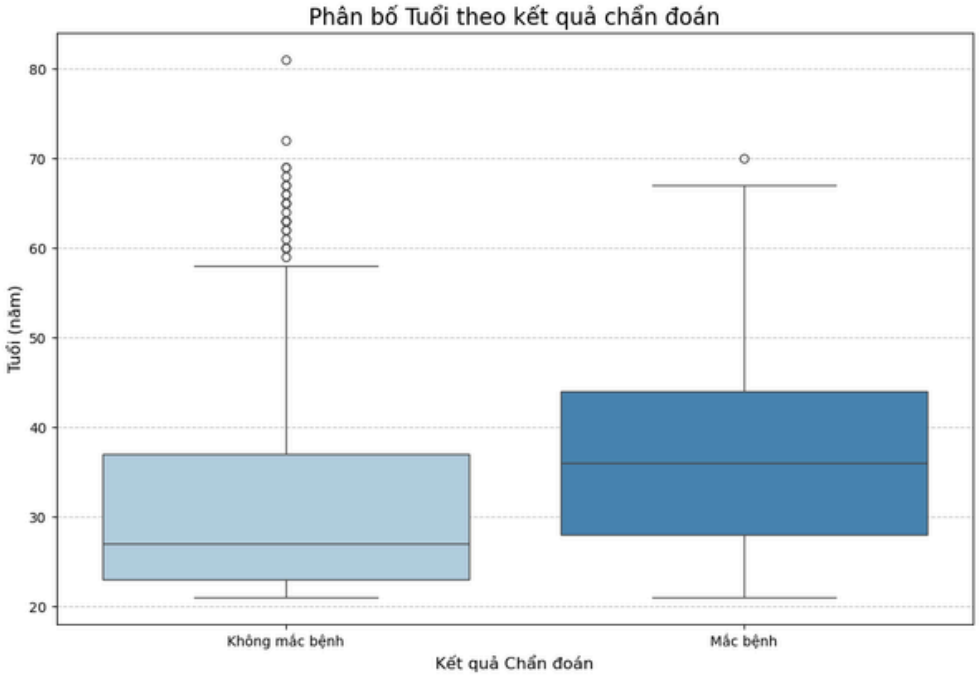


EDA - Phân tích Đơn biến (Phân bố các yếu tố nguy cơ)

- Đa phần nhóm mắt bệnh nằm trong khoảng từ ~28-~45 tuổi.
- Người bệnh tập trung ở độ tuổi trung niên và người cao tuổi.
- Người không mắt bệnh thường tập trung ở độ tuổi khoảng 25-30 tuổi.

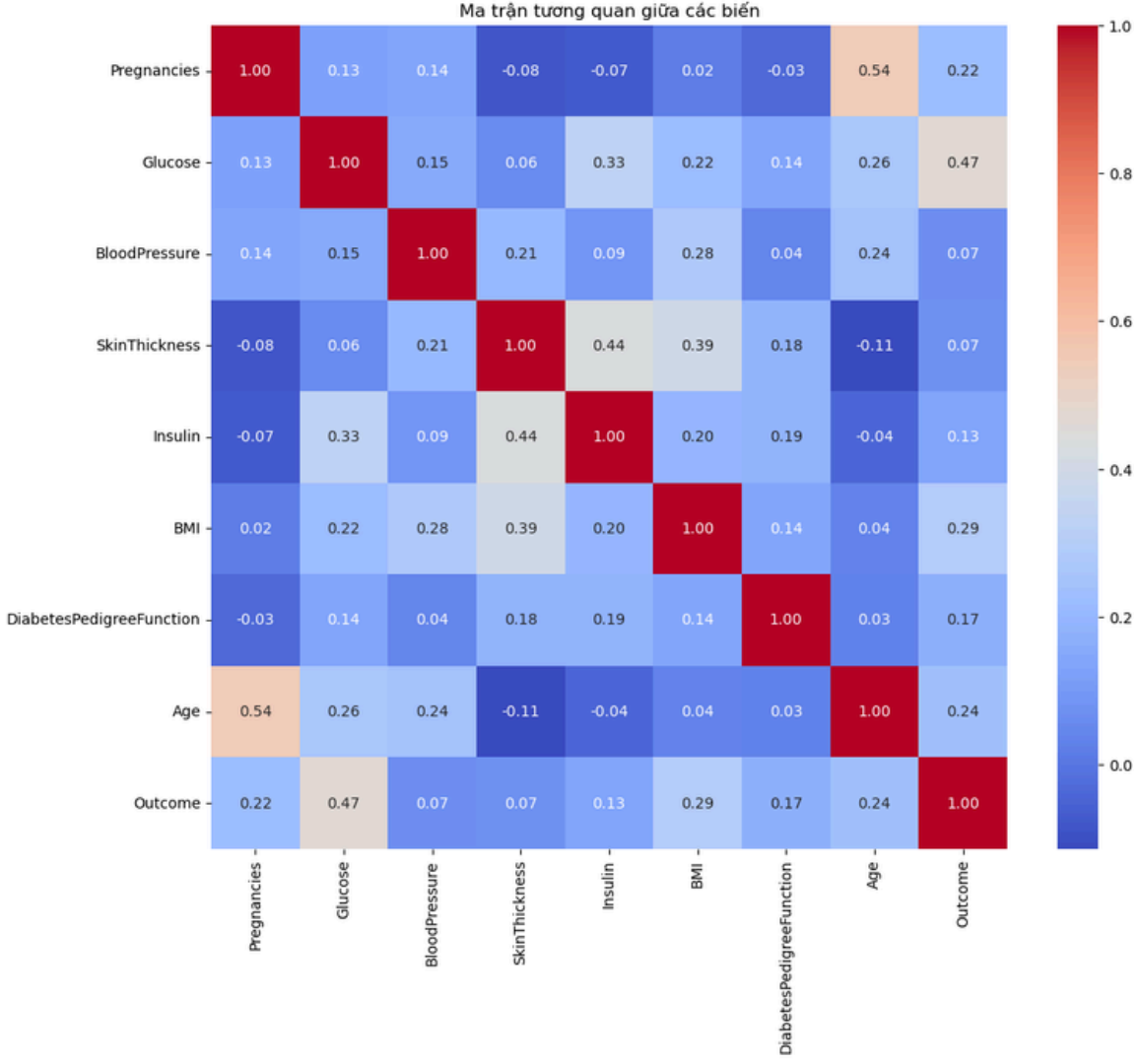
- hai boxplot vẫn có sự chồng chéo đáng kể. Điều này có nghĩa là có những người trẻ tuổi vẫn mắc bệnh, và những người lớn tuổi vẫn không mắc bệnh.

Biểu đồ Histogram/KDE Plot cho Age



EDA - Phân tích Mối tương quan

Biểu đồ Heatmap của ma trận tương quan.



Glucose, BMI và Age là những yếu tố nguy cơ chính đối với bệnh tiểu đường, với Glucose nổi bật là biến dự đoán quan trọng nhất. Điều này nhất quán với các phân tích boxplot trước đó và kiến thức y tế từ báo cáo WHO.

Mối tương quan của các biến với Outcome (Hàng/Cột cuối cùng)

- **Glucose (0.47):** Đây là biến có tương quan dương mạnh nhất với Outcome. Điều này có nghĩa là khi nồng độ Glucose trong huyết tương tăng, khả năng mắc bệnh tiểu đường cũng tăng lên một cách rõ rệt. Đây là yếu tố dự đoán mạnh nhất cho bệnh tiểu đường trong bộ dữ liệu này, hoàn toàn phù hợp với các tiêu chuẩn chẩn đoán y tế.
- **BMI (0.29):** Có mối tương quan dương khá mạnh với Outcome. BMI cao hơn liên quan đến nguy cơ mắc bệnh tiểu đường cao hơn.
- **Age (0.24)**: Có mối tương quan dương vừa phải với Outcome. Tuổi tác càng cao, nguy cơ mắc bệnh tiểu đường càng lớn.
- **Pregnancies (0.22):** Có mối tương quan dương vừa phải với Outcome. Điều này có thể cho thấy số lần mang thai cao hơn có liên quan đến nguy cơ tiểu đường thai kỳ hoặc tăng nguy cơ tiểu đường Type 2 sau này.

Mối tương quan của các biến với Outcome (Hàng/Cột cuối cùng)

- **DiabetesPedigreeFunction (0.17):** Có mối tương quan dương nhẹ với Outcome. Yếu tố phả hệ, đại diện cho tiền sử gia đình, cũng đóng góp vào nguy cơ mắc bệnh.
- **Insulin (0.13):** Có mối tương quan dương nhẹ với Outcome. Mặc dù insulin cao có thể là một phản ứng của cơ thể với glucose cao, nhưng trong một số trường hợp (kháng insulin), nó cũng có thể liên quan đến bệnh tiểu đường.
- **BloodPressure (0.07):** Có mối tương quan dương rất yếu với Outcome. Huyết áp có vẻ không phải là một yếu tố dự đoán mạnh mẽ trực tiếp trong bộ dữ liệu này, mặc dù trong thực tế, cao huyết áp thường đi kèm với tiểu đường.
- **SkinThickness (0.07):** Có mối tương quan dương rất yếu với Outcome. Tương tự như huyết áp, độ dày da không phải là một yếu tố dự đoán trực tiếp mạnh mẽ.

Mối tương quan của các biến với Outcome (Hàng/Cột cuối cùng)

- **Glucose và Insulin (0.54):** Có mối tương quan dương mạnh nhất trong số các biến độc lập. Điều này là hợp lý về mặt sinh lý: khi nồng độ glucose trong máu tăng cao, tuyến tụy sẽ giải phóng nhiều insulin hơn để cố gắng hạ thấp glucose.
- **Pregnancies và Age (0.54):** Có mối tương quan dương khá mạnh. Điều này là tự nhiên vì phụ nữ lớn tuổi hơn thường có số lần mang thai cao hơn.
- **SkinThickness và BMI (0.39):** Có mối tương quan dương đáng kể. Cả hai biến này đều là các chỉ số liên quan đến lượng mỡ cơ thể, vì vậy việc chúng có tương quan với nhau là điều dễ hiểu.
- **BloodPressure và Age (0.29):** Có mối tương quan dương nhẹ, huyết áp có xu hướng tăng theo tuổi tác.
- Các mối tương quan khác thường yếu hoặc không đáng kể, cho thấy chúng cung cấp thông tin tương đối độc lập với nhau.

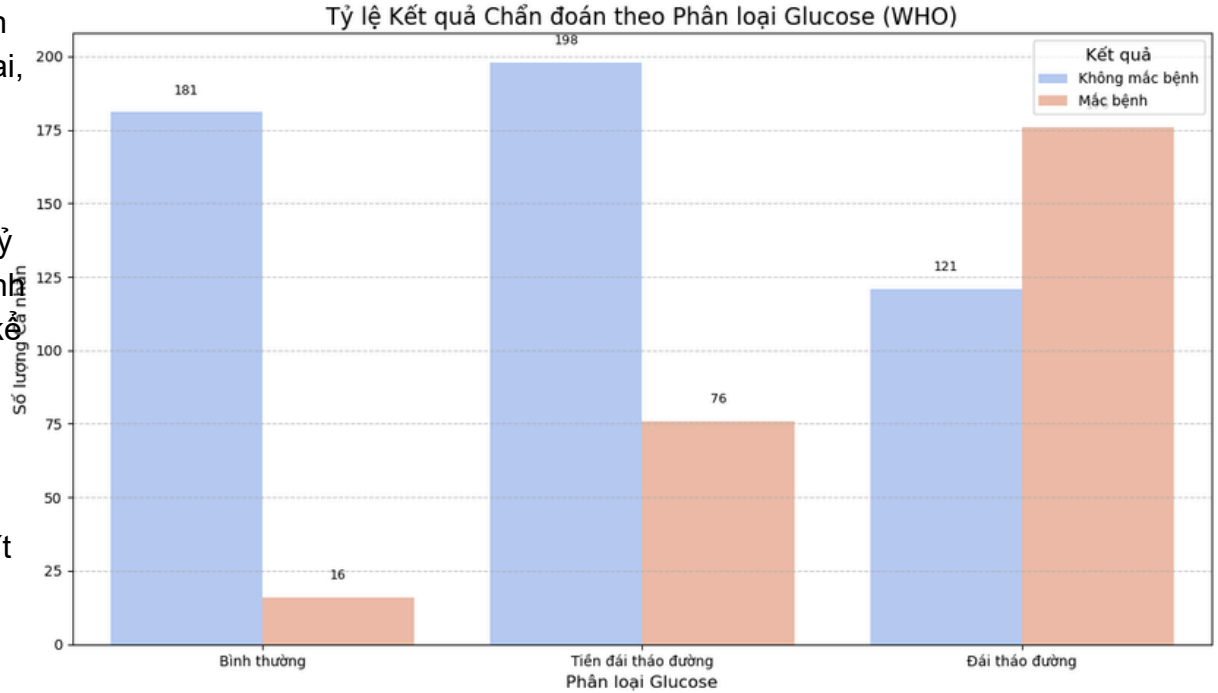
Phân tích theo tiêu chuẩn WHO (Glucose)

"Bình thường" (<100 mg/dL): Đa số cá nhân trong nhóm này không mắc bệnh, nhưng vẫn có một số ít mắc bệnh (có thể là chẩn đoán sai, sai sót dữ liệu, hoặc các yếu tố khác không phải glucose).

"Tiền đái tháo đường" (100-125 mg/dL): Tỷ lệ mắc bệnh tăng lên đáng kể so với nhóm bình thường. Biểu đồ sẽ cho thấy một phần đáng kể trong nhóm này đã mắc bệnh.

"Đái tháo đường" (>=126 mg/dL): Tỷ lệ mắc bệnh trong nhóm này là áp đảo, với rất ít cá nhân được xếp loại là không mắc bệnh.

Biểu đồ Countplot Glucose_Classification vs Outcome



Kết luận

- Các yếu tố như Glucose, BMI, Tuổi và Số lần mang thai được xác định là những yếu tố dự báo mạnh mẽ và đáng tin cậy cho bệnh đái tháo đường.

- Phân tích khám phá dữ liệu trên tập Pima Indians Diabetes hoàn toàn phù hợp với các định nghĩa và tiêu chí chẩn đoán của WHO.

- Cần xử lý các giá trị 0 không hợp lý trong một số biến trước khi xây dựng mô hình.

- Bộ dữ liệu này là nguồn tài nguyên có giá trị để phát triển các mô hình học máy nhằm dự đoán sớm nguy cơ mắc bệnh đái tháo đường.

Thank You