PIMA INDIANS DIABETES ANALYSIS

Huỳnh Anh Nhựt
Nguyễn Tiến Minh

Members

Name	Job Description
Huỳnh Anh Nhựt	Phân tích biểu đồ và làm slide báo cáo
Nguyễn Tiến Minh	Code và phân tích tổng quan dữ liệu

Link GitHub: https://github.com/MinhNguyen-leo/SGU25_ML_GroupProject.git

Points to discuss

- Overview
- Data summary
- Visualizing and summarizing the distributions of the variables
- Univariate analysis
- Multivariate analysis
- Outliers analysis
- Conclusion

Overview

- Pima Indians Diabetes thường được dùng trong học máy để xây dựng mô hình dự đoán bệnh tiểu đường type 2.

- Mục tiêu: dựa trên một số đặc trưng y tế và nhân khẩu học của phụ nữ người da đỏ Pima (trên 21 tuổi), dự đoán xem họ có mắc tiểu đường hay không.

Data summary

Dữ liệu đầu vào

- - Pregnancies – Số lần mang thai
- - Glucose – Nồng độ đường huyết (mg/dL) sau 2 giờ test dung nạp glucose
- - BloodPressure – Huyết áp tâm trương (mm Hg)
- - SkinThickness – Độ dày nếp gấp da (mm)
- - Insulin – Nồng độ insulin huyết thanh (µU/ml)
- - BMI – Chỉ số khối cơ thể (kg/m²)
- - DiabetesPedigreeFunction – Chỉ số di truyền liên quan đến tiểu đường
- - Age – Tuổi (năm)

Dữ liệu đầu ra

- - 0: Không tiểu đường
- - 1: Tiểu đường

Visualizing and summarizing the distributions of the variables

```python
has_null = df.isnull().sum().any()
has_nan  = df.isna().sum().any()
n_duplicated = df.duplicated().sum()

print(f'Tính toàn vẹn dữ liệu:')
print(f'+ Có giá trị Null: {has_null}')
if has_null:
    display.display(df[df.isnull().any(axis=1)])

print(f'+ Có giá trị NaN: {has_nan}')
if has_nan:
    display.display(df[df.isna().any(axis=1)])

print(f'+ Số dòng trùng: {n_duplicated}')
if n_duplicated > 0:
    display.display(df[df.duplicated()])
```

[6] Python

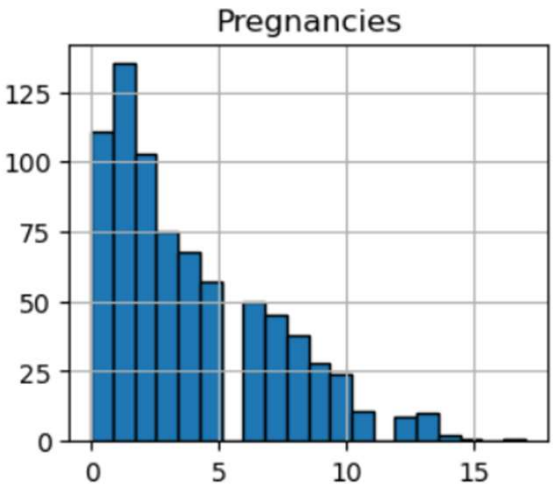
```
Tính toàn vẹn dữ liệu:
+ Có giá trị Null: False
+ Có giá trị NaN: False
+ Số dòng trùng: 0
```

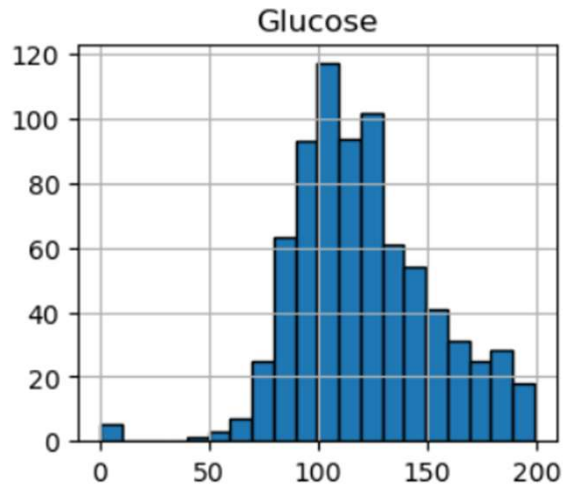
- Trong data không có các mẫu chứa giá trị Null hoặc NaN
- Không có dòng dữ liệu nào trùng lặp trong data
- Dữ liệu hoàn toàn sạch và có thể vào phân tích

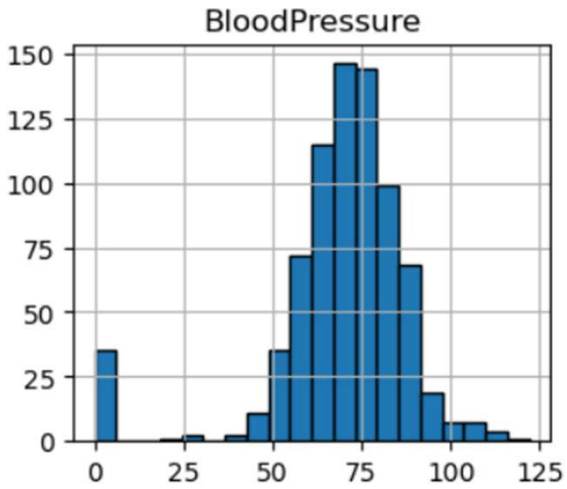
Đây là bảng thống kê của các cột trong dữ liệu

	count	mean	std	min	25%	50%	75%	max
Pregnancies	768.0	3.845052	3.369578	0.000	1.00000	3.0000	6.00000	17.00
Glucose	768.0	120.894531	31.972618	0.000	99.00000	117.0000	140.25000	199.00
BloodPressure	768.0	69.105469	19.355807	0.000	62.00000	72.0000	80.00000	122.00
SkinThickness	768.0	20.536458	15.952218	0.000	0.00000	23.0000	32.00000	99.00
Insulin	768.0	79.799479	115.244002	0.000	0.00000	30.5000	127.25000	846.00
BMI	768.0	31.992578	7.884160	0.000	27.30000	32.0000	36.60000	67.10
DiabetesPedigreeFunction	768.0	0.471876	0.331329	0.078	0.24375	0.3725	0.62625	2.42
Age	768.0	33.240885	11.760232	21.000	24.00000	29.0000	41.00000	81.00
Outcome	768.0	0.348958	0.476951	0.000	0.00000	0.0000	1.00000	1.00

Univariate analysis

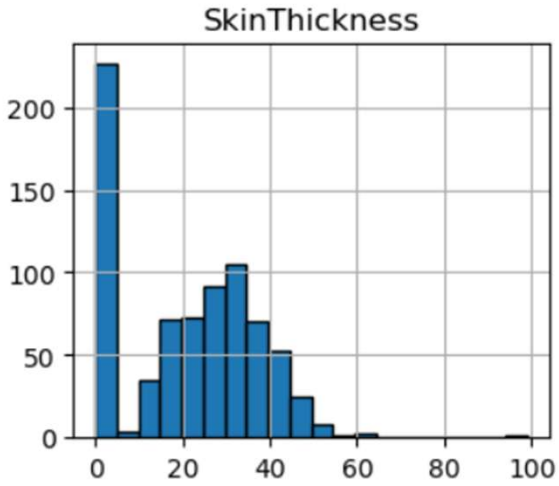


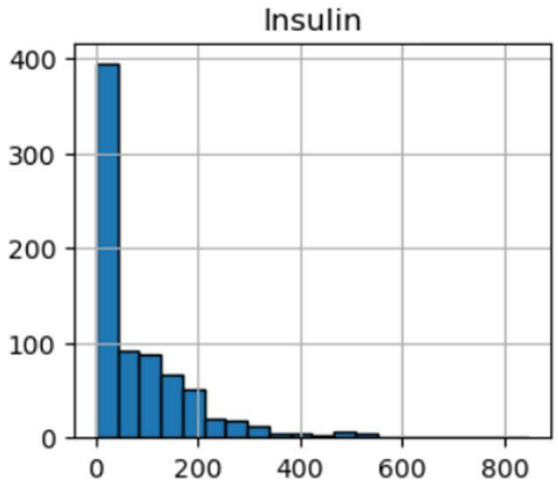


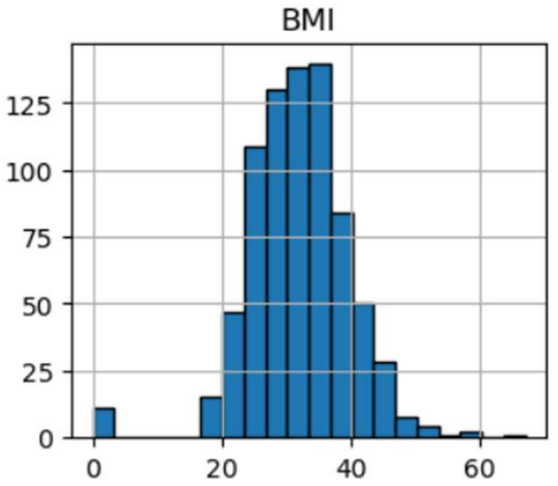


Nhìn vào biểu đồ ta có thể kết luận

- Đa số phụ nữ tham gia thí nghiệm có số lần mang thai =< 5 lần, số ít >= 10 lần
- Mức glucose tập trung ở mức từ 100-130, có vài giá trị bất thường mức glucose = 0 (Đây có thể là do lỗi đánh máy)
- Mức BloodPressure tập trung ở mức từ 60-80, có vài giá trị bất thưởng mức BloodPressure = 0 (Đây có thể là do lỗi đánh máy)

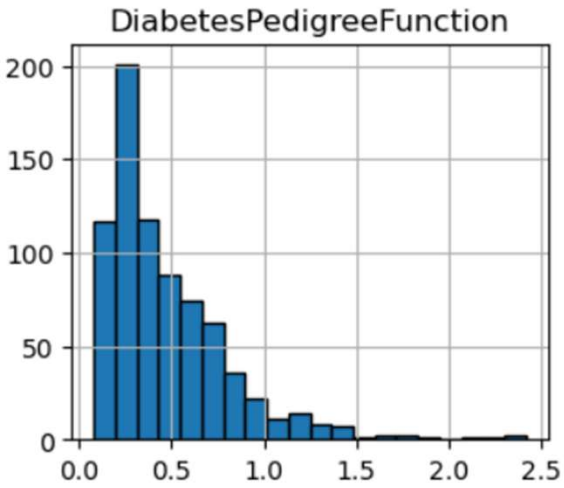


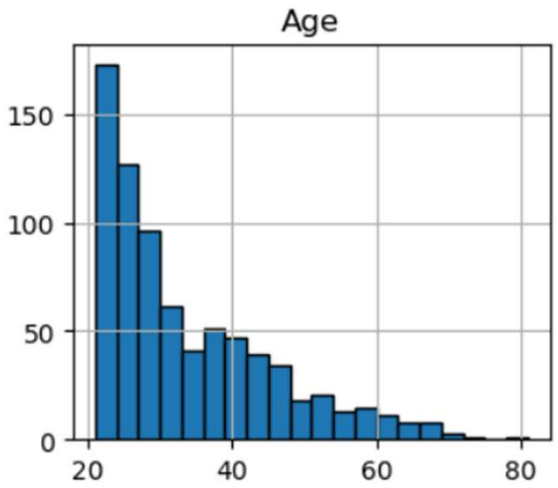


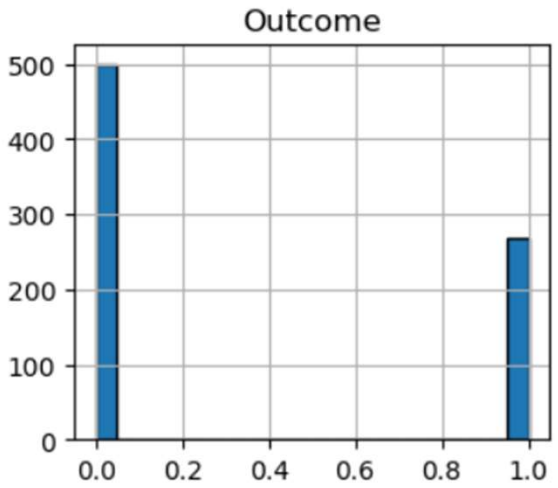


Nhìn vào biểu đồ ta có thể kết luận

- Mức SkinThickness tập trung ở mức từ 20-40, có vài giá trị bất thường mức SkinThickness = 0
-
- Có nhiều giá trị = 0 (Không thực tế), có vài giá trị mức insulin cực kì lớn
- Mức BMI tập trung ở mức từ 25-35, có nhiều giá trị bằng 0 (Không thực tế)



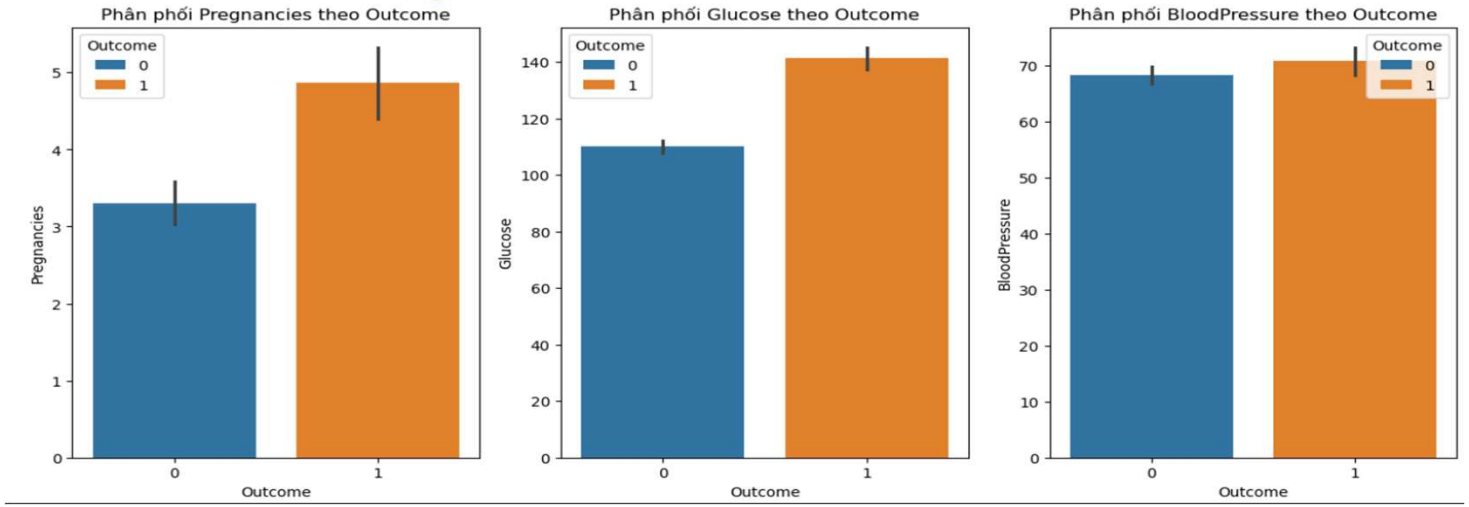




Nhìn vào biểu đồ ta có thể kết luận

- Hầu hết =< 1.0 và số ít >= 2
- Số tuổi trung bình từ 20-40 tuổi, số ít trên 60 tuổi
- Số người không bệnh nhiều hơn số người bị bệnh, dữ liệu không cân bằng có thể ảnh hưởng đến mô hình

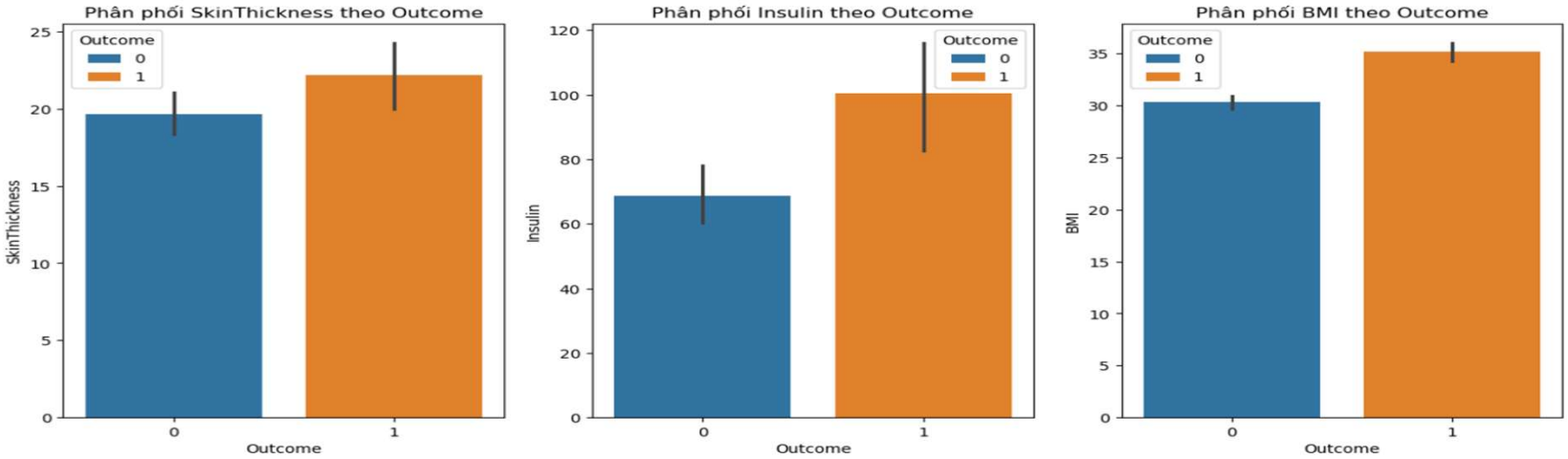
Multivariate analysis



Nhìn vào biểu đồ ta có thể kết luận

- Trung bình, những người không mắc tiểu đường có khoảng 3 lần mang thai, trong khi những người có tiểu đường thì trung bình gần 5 lần
- Đây là khác biệt rõ nhất. Người không tiểu đường có mức đường huyết trung bình khoảng 109 mg/dL, còn người có tiểu đường thì lên tới 141 mg/dL.
- Ở đây sự khác biệt không lớn: nhóm không bệnh trung bình khoảng 70 mmHg, còn nhóm có bệnh khoảng 72 mmHg.

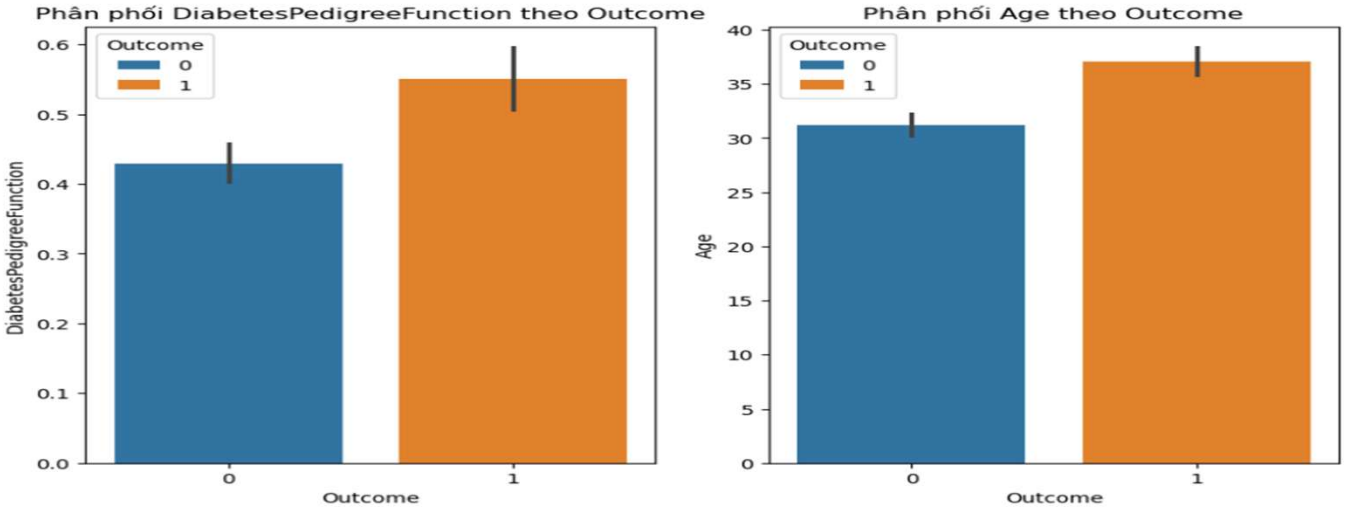
Multivariate analysis



Nhìn vào biểu đồ ta có thể kết luận

- Người không tiểu đường trung bình dày khoảng 20 mm, còn người có tiểu đường thì khoảng 22 mm. Có chênh lệch, nhưng không rõ ràng lắm.
- Nhóm không mắc bệnh có nồng độ insulin trung bình khoảng 69 µU/mL, trong khi nhóm mắc bệnh cao hơn, khoảng 100 µU/mL. Tuy vậy, dữ liệu insulin khá biến động, có thể có những giá trị ngoại lai.
- Người không mắc bệnh có BMI trung bình 30.3, trong khi người có bệnh lên tới 35.1. Điều này phù hợp với thực tế: béo phì là một yếu tố nguy cơ lớn của tiểu đường type 2.

Multivariate analysis

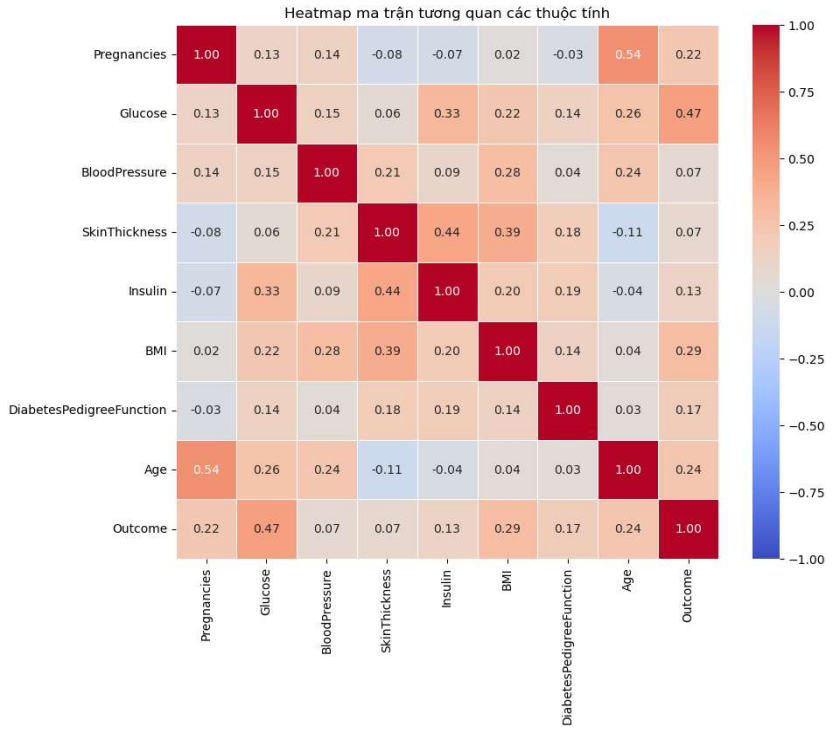


Nhìn vào biểu đồ ta có thể kết luận

- Người không mắc bệnh có chỉ số này trung bình khoảng 0.43, còn người có bệnh khoảng 0.55. Điều này cho thấy yếu tố digóp phần không nhỏ.

- Người không mắc bệnh có tuổi trung bình 31, trong khi người mắc bệnh trung bình 37 tuổi. Tuổi càng cao thì nguy cơ mắc tiểu đường càng lớn.

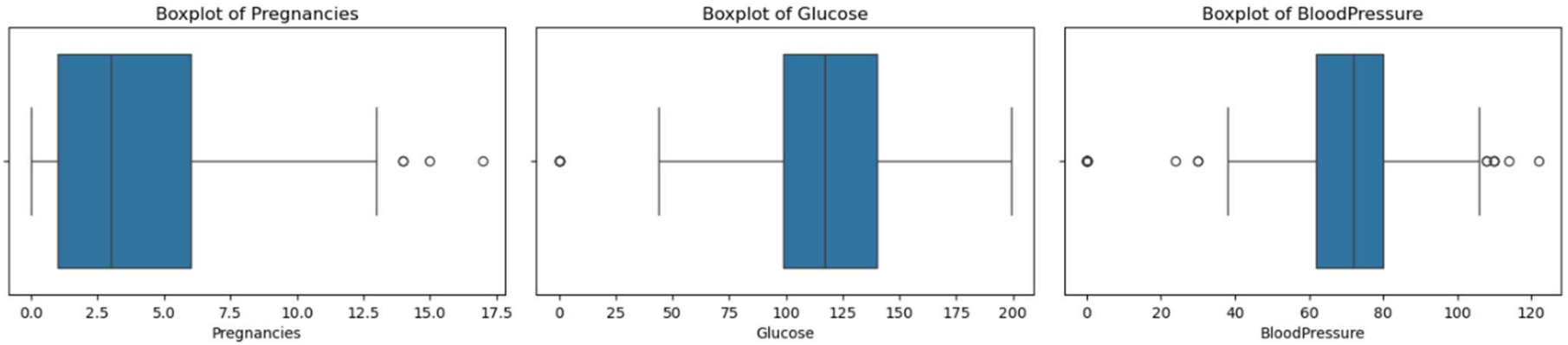
Heatmap



Nhìn vào biểu đồ ta có thể kết luận

- Số lần mang thai có tương quan với mức Outcome
- Mức Glucose tương quan mạnh với Outcome cho thấy Mức glucose càng cao thì có tỉ lệ bệnh cao hơn
- BloodPressure và SkinThickness = 0.07 hầu như không tương quan mạnh với Outcome
- Insulin = 0.13 tương quan khá yếu với Outcome nhưng lại tương quan mạnh với mức Glucose
- BMI = 0.29 cũng tương quan với Outcome
- Chỉ số DiabetesPedigreeFunction = 0.17 tương quan khá yếu với Outcome và cũng tương quan yếu với chỉ số khác
- Age = 0.24 tương quan với Outcome và cũng tương quan với 2 chỉ số glucose và BloodPressure, còn lại thì tương quan yếu

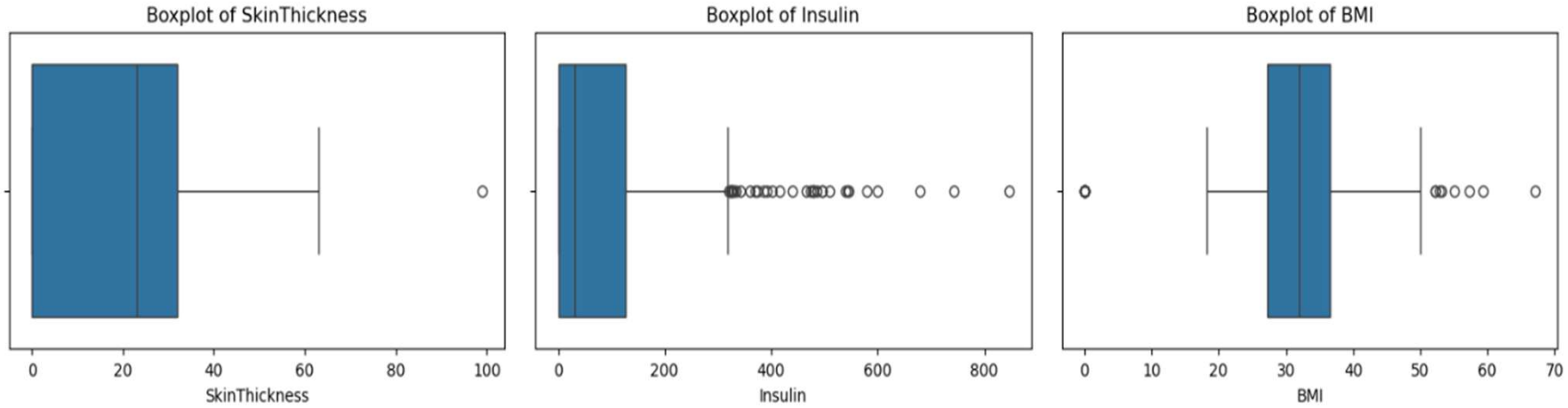
Outliers analysis



Nhìn vào biểu đồ ta có thể kết luận

- Số lần mang thai có vài mẫu >= 10 (Trên thực tế vẫn có thể xảy ra)
- Có vài mẫu glucose = 0 (Đây là giá trị missing), còn lại vẫn trong giới hạn thực tế
- Có vài mẫu BloodPressure = 0 (Có thể là giá trị missing), và vài ngoại lệ >= 120

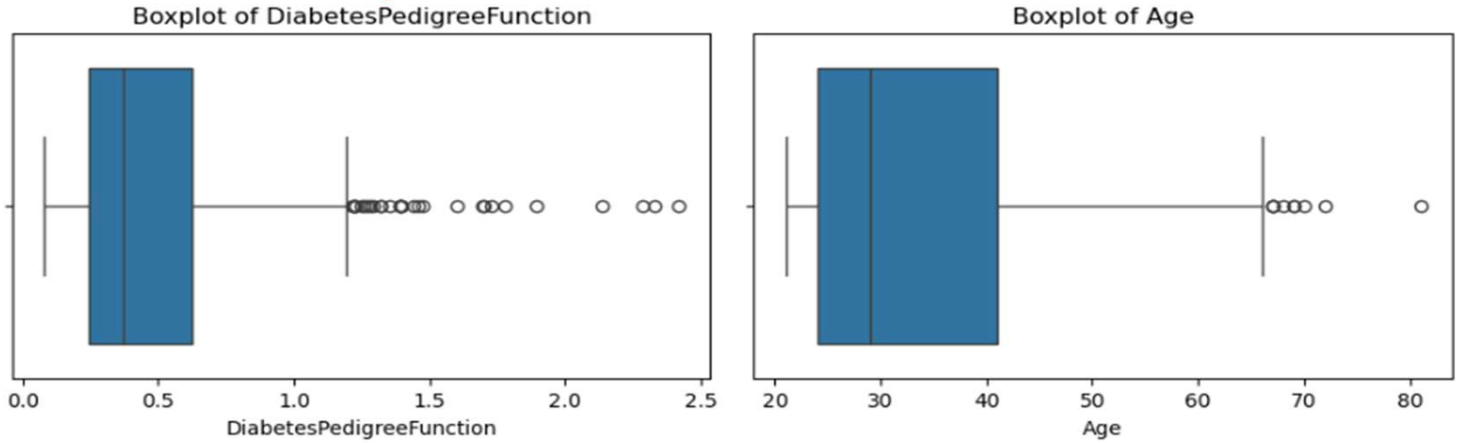
Outliers analysis



Nhìn vào biểu đồ ta có thể kết luận

- Có vài mẫu SkinThickness cực lớn (~100)
- Rất nhiều Insulin giá trị 0 và rất nhiều giá trị outlier ở mức >= 400
- Có một vài mẫu có giá trị BMI = 0 (Đây có thể là giá trị missing), có outliers ở mức 55–70

Outliers analysis



Nhìn vào biểu đồ ta có thể kết luận

- Có khá nhiều mẫu DiabetesPedigreeFunction có outliers ở mức >= 1.5
- Có vài mẫu >= 80 tuổi

Conclusion

- Dữ liệu có nhiều giá trị bất thường (0) ở các cột Glucose, BloodPressure, SkinThickness, Insulin, BMI – đây là các giá trị không thực tế và nên được coi là missing value để xử lý tiếp.
- Một số thuộc tính như Glucose, BMI, Age có sự khác biệt rõ rệt giữa hai nhóm Outcome (có và không mắc tiểu đường), cho thấy chúng là các đặc trưng quan trọng cho mô hình dự đoán.
- Các thuộc tính như BloodPressure, SkinThickness, Insulin, DiabetesPedigreeFunction có tương quan yếu với Outcome, có thể cân nhắc loại bỏ hoặc giảm trọng số khi xây dựng mô hình.
- Dữ liệu có nhiều outlier ở các cột Insulin, BMI, SkinThickness, DiabetesPedigreeFunction, cần xử lý để tránh ảnh hưởng đến mô hình.
- Tập dữ liệu bị mất cân bằng: số người không mắc bệnh nhiều hơn số người mắc bệnh, điều này có thể ảnh hưởng đến hiệu quả dự đoán và cần xem xét các kỹ thuật cân bằng dữ liệu.
- Không phát hiện giá trị null/NaN hoặc dòng trùng lặp, nhưng cần xử lý các giá trị 0 bất thường như đã nêu trên.

Thank You