

STROKE PREDICTION USING MACHINE LEARNING TECHNIQUES

Khổng Đức Quang - MSSV: 20225072

Ngày 4 tháng 12 năm 2024

1 Tóm tắt

Đột quỵ là một **tình trạng y tế khẩn cấp** xảy ra khi dòng máu đến một phần não bị gián đoạn do xuất huyết hoặc tắc nghẽn bởi cục máu đông. Đây là nguyên nhân tử vong đứng thứ hai trên toàn cầu, với khoảng **5,5 triệu ca tử vong** mỗi năm. Theo thống kê, mỗi năm có hơn **15 triệu người** trên thế giới bị ảnh hưởng bởi đột quỵ, và *trung bình cứ mỗi 4 phút lại có một trường hợp tử vong* do tình trạng này.

Phần lớn các trường hợp đột quỵ có liên quan chặt chẽ đến lối sống không lành mạnh, dẫn đến ước tính khoảng 80% các trường hợp có thể được phòng ngừa. Do đó, việc xây dựng các mô hình dự đoán nguy cơ đột quỵ có ý nghĩa quan trọng trong việc ngăn chặn các tổn thương nghiêm trọng và giảm thiểu tỷ lệ tử vong liên quan. Công tác dự đoán không chỉ giúp cảnh báo sớm mà còn hỗ trợ trong việc đưa ra các biện pháp can thiệp kịp thời nhằm bảo vệ sức khỏe cộng đồng.

2 Mục tiêu dự án

Mục tiêu của dự án này là dự đoán khả năng xảy ra đột quỵ não bằng cách ứng dụng các kỹ thuật học máy. Bằng cách phân tích dữ liệu y tế, em sẽ huấn luyện một số mô hình học máy nhằm nhận diện các mẫu và yếu tố rủi ro liên quan đến đột quỵ. Điều này sẽ hỗ trợ phát hiện sớm, cung cấp những thông tin quan trọng giúp đưa ra các biện pháp phòng ngừa và can thiệp kịp thời.

Mục lục

1	Tóm tắt	1
2	Mục tiêu dự án	1
3	Mô tả bộ dữ liệu	4
4	Xử lý dữ liệu	5
5	Phân tích dữ liệu	6
6	Xem xét sự mất cân bằng dữ liệu	12

Danh sách hình vẽ

1	Kiểm tra giá trị NULL	5
2	Correlation Heatmap	6
3	Biểu đồ gender	7
4	Biểu đồ residencetype_type	7
5	Biểu đồ ever_married	8
6	Biểu đồ work_type	8
7	Biểu đồ smoking_status	9
8	Biểu đồ hypertension	9
9	Biểu đồ heart_disease	10
10	Biểu đồstroke liên hệ age và stroke	10
11	Biểu đồ liên hệ average_glucose_level	11
12	Biểu đồ liên hệ bmi và stroke	11
13	Biểu đồ so sánh số lượng người đột quỵ và không đột quỵ	12
14	Phân bố tỷ lệ người có và không đột quỵ	13
15	Xử lý dữ liệu mất cân bằng	14

3 Mô tả bộ dữ liệu

Bộ dữ liệu được em thu thập từ trang Web Kaggle để ước tính xem bệnh nhân có khả năng đột quỵ hay không. Đây là bộ dữ liệu về thông tin của 5110 người bao gồm 11 thuộc tính và 1 cột stroke(output) là có khả năng đột quỵ hay không. (FEDESORIANO 2021).

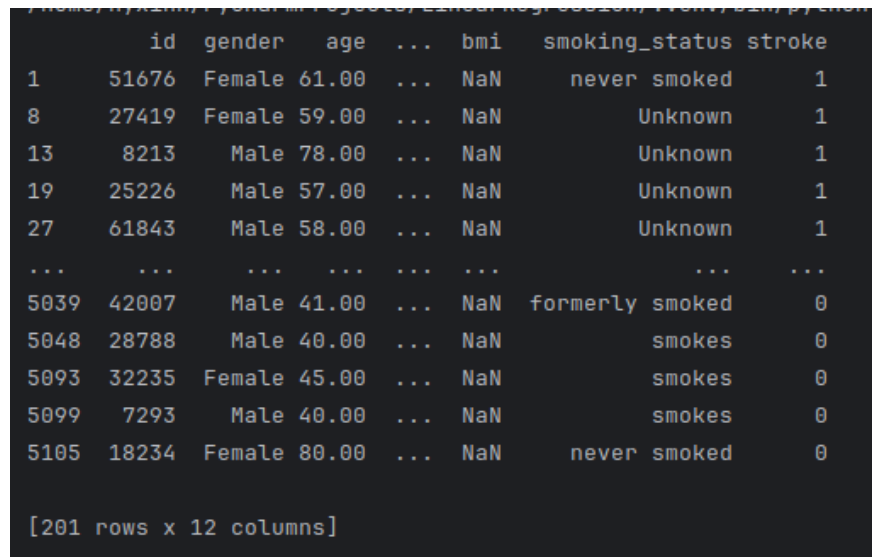
Dưới đây là danh sách các thuộc tính:

1. **Id (Integer Feature)**: Đây là dữ liệu kiểu số nhằm. Tuy nhiên thuộc tính này không làm ảnh hưởng tới Output nên em sẽ không phân tích thêm.
2. **Gender (Nominal Feature)**: Đây là thuộc tính kiểu chữ bao gồm các giá trị: Male, Female, Other. Thuộc tính "Gender"(giới tính) có thể ảnh hưởng tới khả năng đột quỵ do sự khác biệt sinh học, nội tiết tố, và yếu tố hành vi giữa nam và nữ.
3. **Age (Integer Feature)**: Dữ liệu kiểu số. Thuộc tính này là một yếu tố quan trọng ảnh hưởng đến khả năng đột quỵ.
4. **Hypertension (Integer Feature)**: Với hai giá trị khác nhau là: 0, 1. Đây là yếu tố nguy cơ lớn nhất đối với đột quỵ. Nó có ảnh hưởng sâu sắc đến khả năng đột quỵ.
5. **Heart Disease (Integer Feature)**: Với hai giá trị khác nhau là: 0, 1. Là một yếu tố có nguy cơ đáng kể gây đột quỵ.
6. **Ever married (Boolean Feature)**: Với hai giá trị khác nhau là: True, False. Có thể liên quan đến nguy cơ đột quỵ qua các yếu tố gián tiếp, chẳng hạn như lối sống, sức khỏe tâm lý, và sự hỗ trợ xã hội.
7. **Work type (Nominal Feature)**: Có 5 loại giá trị khác nhau: Private, Self-employed, children, Govt-job, Never-worked. Và ảnh hưởng đáng kể đến nguy cơ đột quỵ thông qua các yếu tố như mức độ căng thẳng, hoạt động thể chất, và tiếp xúc với các yếu tố nguy cơ môi trường.
8. **Residence type (Nominal Feature)**: Có hai giá trị khác nhau là: Urban, Rural. Ảnh hưởng đến nguy cơ đột quỵ thông qua các yếu tố môi trường, điều kiện sống, và khả năng tiếp cận dịch vụ chăm sóc sức khỏe.
9. **Avg glucose level (Float Feature)**: Là một yếu tố quan trọng liên quan đến nguy cơ đột quỵ, đặc biệt thông qua mối liên hệ với bệnh tiểu đường và rối loạn chuyển hóa. Mức đường huyết bất thường, cả cao lẫn thấp, đều có thể làm tăng nguy cơ đột quỵ.
10. **BMI (Float Feature)**: Là một chỉ số quan trọng để đánh giá mức độ béo phì hoặc thừa cân của một người, và nó có mối liên hệ mạnh mẽ với nguy cơ đột quỵ.
11. **Smoking status(Nominal Feature)**: Với 4 giá trị khác nhau: Never smoked, Unknown, formerly smoked, smokes. Là một yếu tố nguy cơ quan trọng đối với nhiều bệnh lý, bao gồm đột quỵ. Hút thuốc lá ảnh hưởng đến sức khỏe của tim và mạch máu, gây ra những tác động tiêu cực trực tiếp làm tăng nguy cơ đột quỵ.

4 Xử lý dữ liệu

Trong quá trình tiền xử lý số liệu, em nhận thấy có các vấn đề sau cần xử lý:

1. Đầu tiên em **loại bỏ đi cột ID** do cột này không ảnh hưởng tới khả năng đột quỵ.
2. **Xử lý các giá trị NULL**



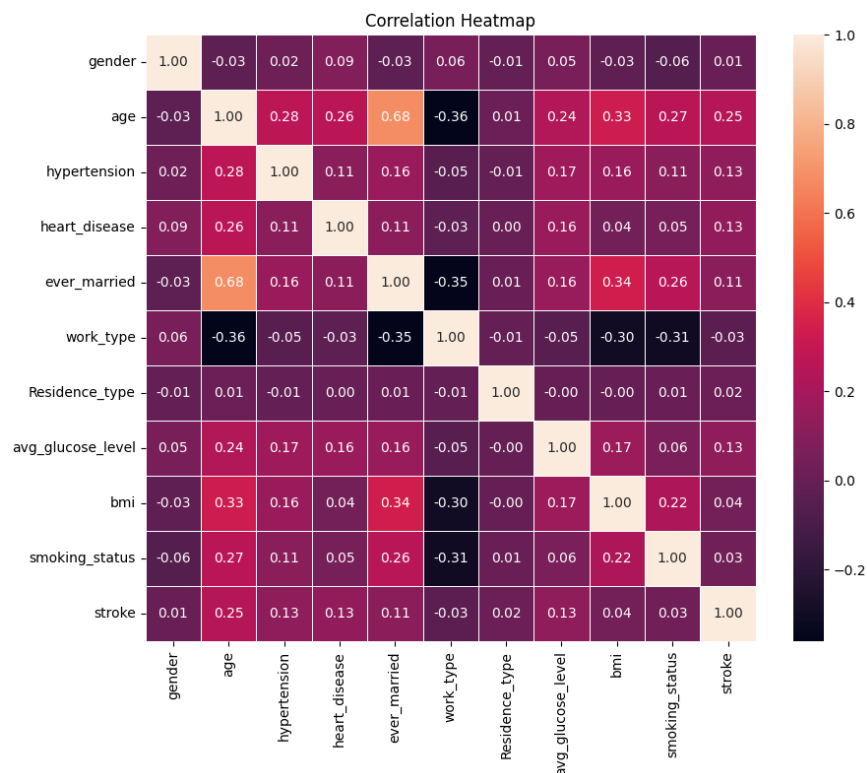
	id	gender	age	...	bmi	smoking_status	stroke
1	51676	Female	61.00	...	NaN	never smoked	1
8	27419	Female	59.00	...	NaN	Unknown	1
13	8213	Male	78.00	...	NaN	Unknown	1
19	25226	Male	57.00	...	NaN	Unknown	1
27	61843	Male	58.00	...	NaN	Unknown	1
...
5039	42007	Male	41.00	...	NaN	formerly smoked	0
5048	28788	Male	40.00	...	NaN	smokes	0
5093	32235	Female	45.00	...	NaN	smokes	0
5099	7293	Male	40.00	...	NaN	smokes	0
5105	18234	Female	80.00	...	NaN	never smoked	0
[201 rows x 12 columns]							

Hình 1: Kiểm tra giá trị NULL

Sau khi kiểm tra em nhận thấy cột thuộc tính "bmi" có 201 giá trị NULL. Để xử lý vấn đề này có ba phương pháp phổ biến: Xóa các dòng chứa giá trị NULL, Điền giá trị trung bình hoặc trung vị, Điền giá trị dựa trên mô hình (Model-based Imputation). Sau khi tìm hiểu ưu và nhược điểm của từng phương pháp em quyết định lựa chọn cách điền giá trị trung bình là lựa chọn vừa đơn giản vừa hiệu quả, giúp giữ được lại toàn bộ dữ liệu mà không làm giảm kích thước bộ dữ liệu.

3. Bước tiếp theo là **chuyển đổi dữ liệu kiểu Categorical thành dữ liệu kiểu Numerical**. Phương pháp Hash e Encoding và One-hot Encoding được sử dụng để thực hiện các bước chuyển đổi này.
4. **Phân chia dữ liệu**: Em chia bộ dữ liệu thành hai tập là: tập huấn luyện(80%), tập kiểm tra(20%) bằng cách sử dụng hàm `train_test_split` được cung cấp trong thư viện `scikit-learn`
5. **Feature scaling (chuẩn hóa đặc trưng)** là bước cuối cùng

5 Phân tích dữ liệu



Hình 2: Correlation Heatmap

Nhận xét:

- Một số tương quan dương mạnh:

+) Age và ever_married (0.68) có tương quan dương mạnh nhất

+) Age cho thấy mối tương quan tích cực vừa phải với BMI (0.33), Hypertension (0.28) và smoking_status (0.27)

- Một số tương quan yếu đáng chú ý:

+) Work_type có một số tương quan yếu với: age (-0.36), ever_married (-0.35), bmi (-0.30), smoking_status (-0.31)

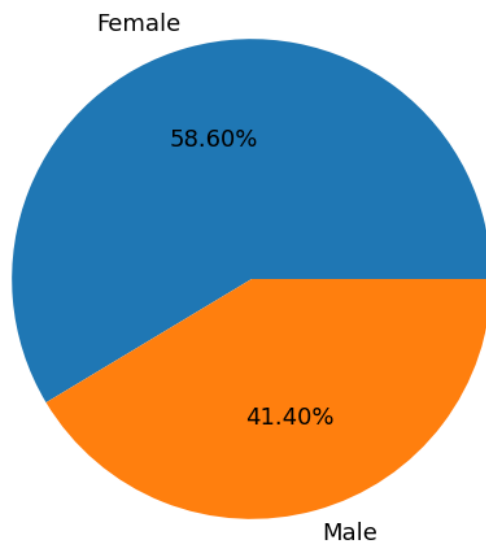
- Tương quan đột quỵ:

+) Stroke có tương quan yếu hoặc trung bình yếu với hầu hết các thuộc tính trong bảng.

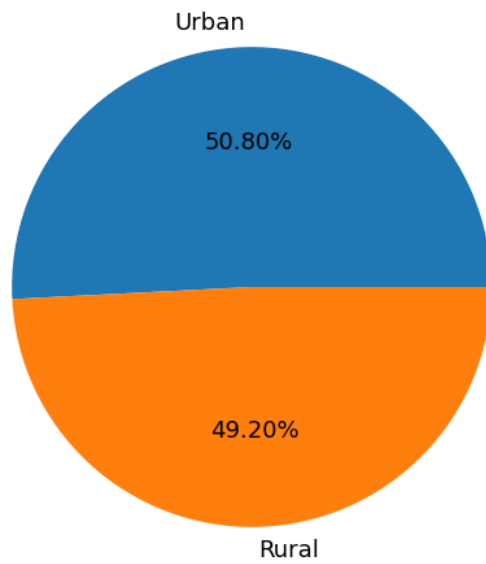
+) Age có tương quan cao nhất nhưng vẫn là tương quan yếu với Stroke (0.25), tương tự là heart_disease và hypertension, cả hai đều là 0.13

- Hầu hết các tương quan đều là yếu và rất yếu, gần giá trị 0

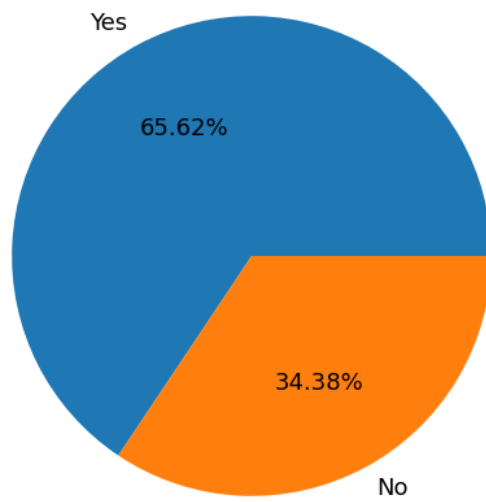
Tóm lại, age là yếu tố có ảnh hưởng lớn nhất đến khả năng đột quỵ.



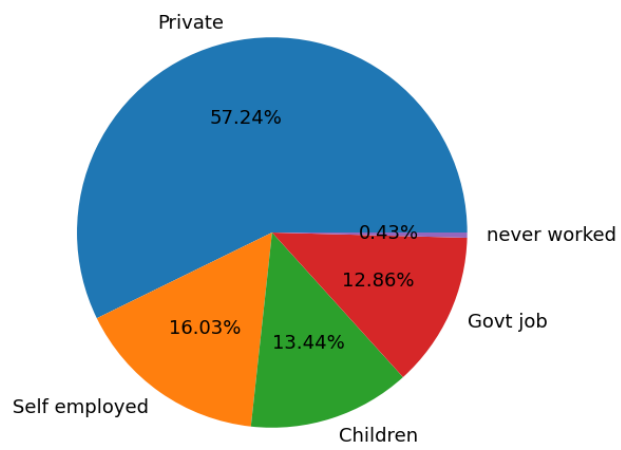
Hình 3: Biểu đồ gender



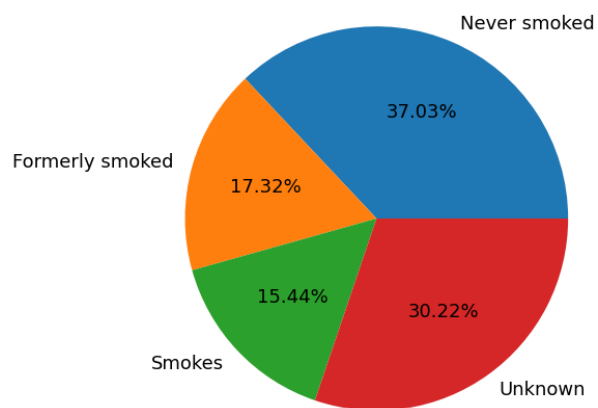
Hình 4: Biểu đồ residencetype_type



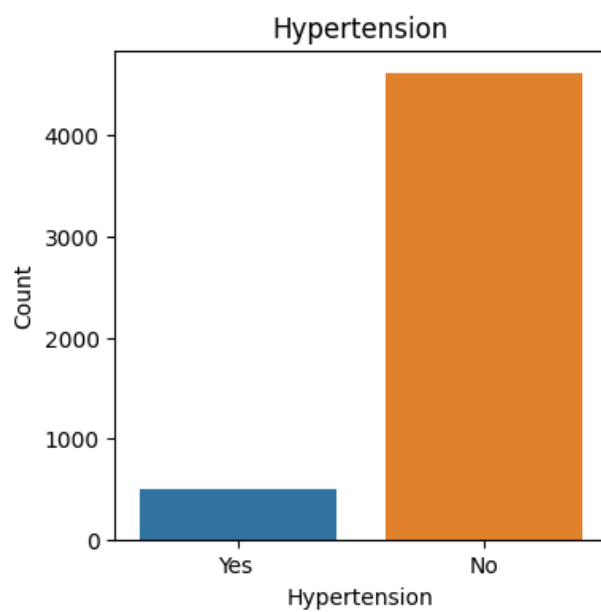
Hình 5: Biểu đồ ever_married



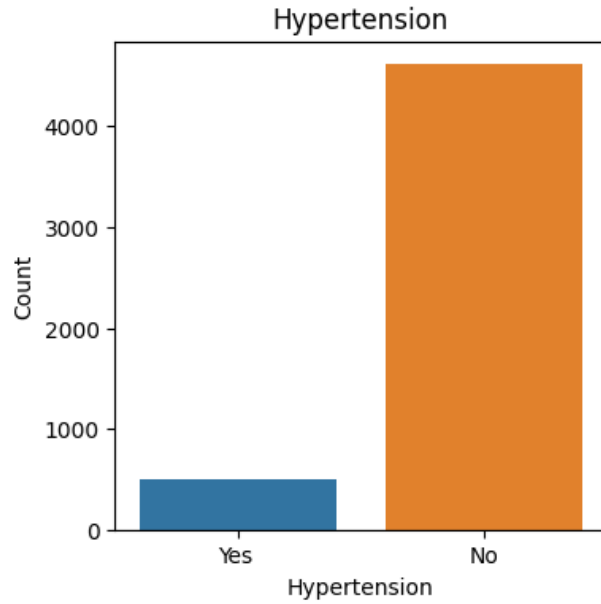
Hình 6: Biểu đồ work_type



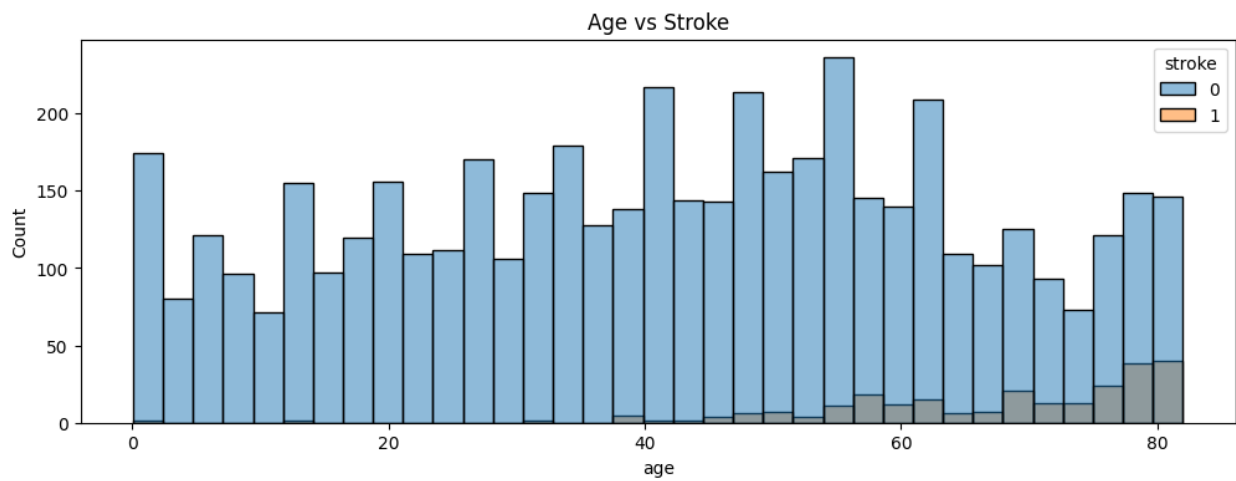
Hình 7: Biểu đồ smoking_status



Hình 8: Biểu đồ hypertension



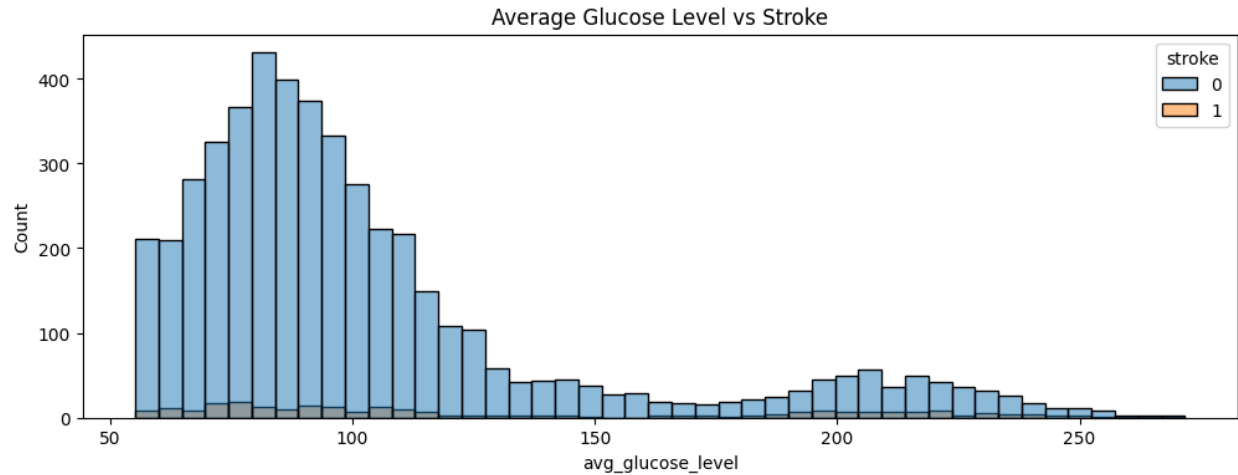
Hình 9: Biểu đồ heart_disease



Hình 10: Biểu đồstroke liên hệ age và stroke

Age và Stroke:

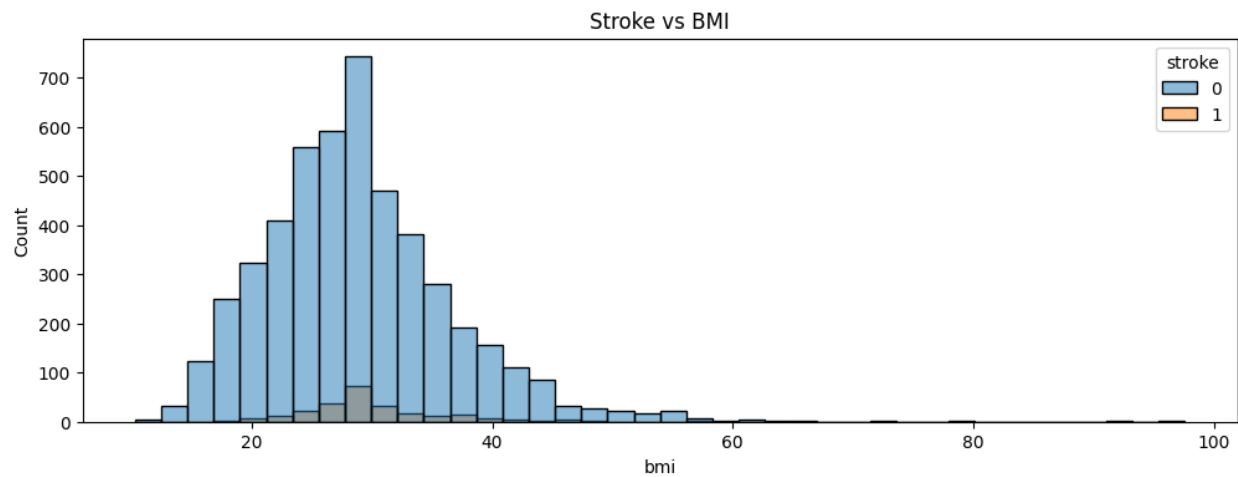
1. Các trường hợp đột quỵ tăng dần sau 40 tuổi.
2. Tỷ lệ đột quỵ cao nhất nằm trong độ tuổi từ 60-80 tuổi.
3. Có rất ít trường hợp đột quỵ dưới 40 tuổi.
4. Phân bố cho thấy tuổi tác là yếu tố có ảnh hưởng đáng kể đối với đột quỵ.



Hình 11: Biểu đồ liên hệ average_glucose_level

Average_glucose_level và StrokeS:

1. Các trường hợp đột quỵ có xu hướng xuất hiện nhiều hơn ở các mức đường huyết trung bình cao hơn, đặc biệt từ 120 mg/dL trở lên.
2. Trong phạm vi đường huyết từ 180 mg/dL đến 250 mg/dL, tỷ lệ người bị đột quỵ tăng lên đáng kể, dù số lượng tổng thể thấp hơn.
3. Ở mức đường huyết dưới 100 mg/dL, tỷ lệ đột quỵ thấp hơn rõ rệt.



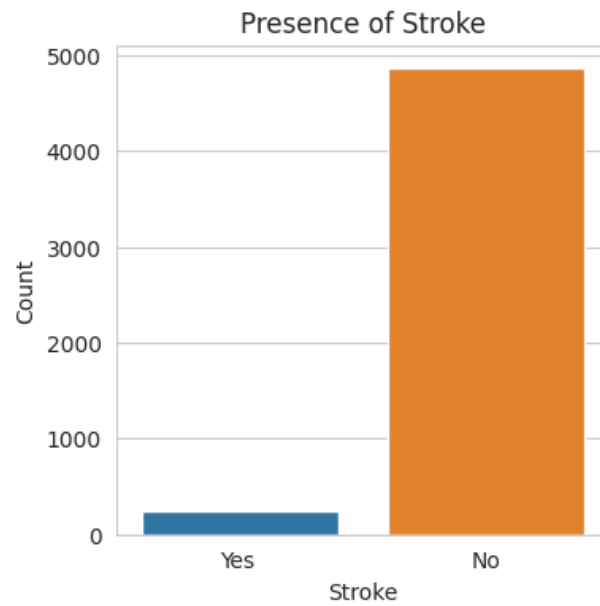
Hình 12: Biểu đồ liên hệ bmi và stroke

BMI và Stroke:

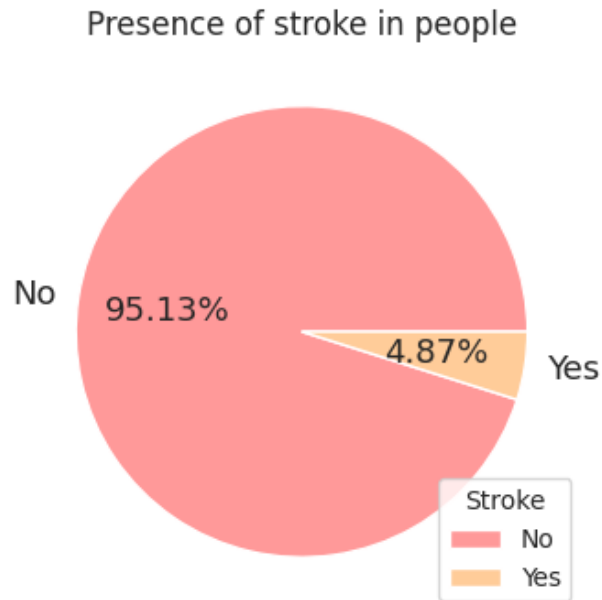
1. Phần lớn mọi người có chỉ số BMI từ 20-40.
2. Phân bố đỉnh nằm trong khoảng 25-35.

3. Các trường hợp đột quỵ (màu xám) xuất hiện thường xuyên hơn ở nhóm thừa cân và béo phì ($BMI > 25$).
4. Dữ liệu cho thấy tình trạng thừa cân hoặc béo phì có thể làm tăng nguy cơ đột quỵ, mặc dù mối quan hệ này có vẻ ít nghiêm trọng hơn so với tuổi tác hoặc lượng đường trong máu.

6 Xem xét sự mất cân bằng dữ liệu



Hình 13: Biểu đồ so sánh số lượng người đột quỵ và không đột quỵ



Hình 14: Phân bố tỷ lệ người có và không đột quỵ

Nhận xét:

1. Mất cân bằng dữ liệu nghiêm trọng. Tỷ lệ người không bị đột quỵ (No) *chiếm 95.13%* tổng số dữ liệu. Trong khi đó, tỷ lệ người bị đột quỵ (Yes) *chỉ chiếm 4.87%*. Điều này cho thấy dữ liệu mất cân bằng nghiêm trọng giữa hai nhóm.
2. Việc mất cân bằng dữ liệu này có thể dẫn đến việc mô hình học máy thiên vị, tập trung vào việc dự đoán chính xác lớp chiếm đa số (No) mà bỏ qua lớp thiểu số (Yes). **Từ đó dẫn đến kết quả** mô hình có thể đạt độ chính xác cao nhưng không phát hiện chính xác các trường hợp bị đột quỵ, làm tăng tỷ lệ False Negative (bỏ sót những người thực sự bị đột quỵ).

Giải pháp:

Để giải quyết vấn đề trên cũng như cải thiện khả năng dự đoán của mô hình, em sẽ áp dụng kỹ thuật xử lý dữ liệu mất cân bằng đó là Oversampling: Tăng số lượng mẫu cho lớp thiểu số bằng cách sử dụng SMOTE.

```
[133] ✓ 0.0s Python
data1 = data.copy()
X = data.drop('stroke', axis = 1)
Y = data['stroke']
Y = pd.DataFrame(Y)

[148] ✓ 0.0s Python
smote = SMOTE(random_state = 10)
X1, Y1 = smote.fit_resample(X, Y)

[149] ✓ 0.0s Python
Y1 = pd.DataFrame(Y1)
X1 = pd.DataFrame(X1)
Y1.value_counts()

... stroke
0      4860
1      4860
Name: count, dtype: int64
```

Hình 15: Xử lý dữ liệu mất cân bằng

Sau khi sử dụng kỹ thuật SMOTE để xử lý sự mất cân bằng của dữ liệu thì số trường hợp đột quỵ và không đột quỵ đã bằng nhau như kết quả trên hình.