

09.part2

August 19, 2021

Nhóm 09:

1. Nguyễn Văn Thé
2. Lê Trung Hiếu
3. Nguyễn Thị Kim Duyên

1 Giới thiệu

Theo Tổ chức Y tế thế giới (WHO), đột quy là lý do đứng thứ 2 gây ra tử vong và lý do đứng thứ 3 gây ra khuyết tật trên toàn cầu. Đột quy là hiện tượng chết đột ngột tế bào não do việc thiếu oxy khi lượng máu lên não bị mất đi do tắc nghẽn hoặc vỡ động mạch lên não. Đây cũng là nguyên nhân hàng đầu dẫn đến sa sút trí tuệ và trầm cảm.

Gần 800 nghìn người ở Mỹ trải qua 1 cơn đột quy mỗi năm, trong đó khoảng 3/4 là đột quy lần đầu tiên. 80% những cơn đột quy này có thể được ngăn chặn nên việc giáo dục nhận biết các triệu chứng đột quy là rất cần thiết.

Vậy nên mục đích của tiểu luận là xây dựng một mô hình dự đoán đột quy và đánh giá độ chính xác của mô hình. Trong bài tập lần này, nhóm phân tích dữ liệu nhiều biến sử dụng PCA và KMeans

2 Nguồn dữ liệu

Dữ liệu được thu thập của 5110 người tham gia vào nghiên cứu này với 2995 nam và 2115 nữ, được lấy từ trang [Kaggle](#). Để dự đoán 1 bệnh nhân liệu có bị đột quy hay không dựa theo các thông tin ở đây:

Các trường dữ liệu:

- id:
 - Là số nhận dạng (ID) của bệnh nhân
 - Là biến định lượng liên tục kiểu số
- gender:
 - Giới tính bệnh nhân
 - Là biến định tính rời rạc kiểu String
 - Nhận 1 trong các giá trị: “Male”, “Female” or “Other”
- age:

- Tuổi của bệnh nhân
 - Là biến định lượng liên tục kiểu số
- hypertension:
 - Chứng cao huyết áp của bệnh nhân
 - Là biến định tính rời rạc kiểu số
 - Nhận 1 trong 2 giá trị: 0 nếu bệnh nhân không bị cao huyết áp và 1 nếu bệnh nhân mắc cao huyết áp
- heart_disease:
 - Tình trạng đau tim của bệnh nhân
 - Là biến định tính rời rạc kiểu số
 - Nhận 1 trong 2 giá trị: 0 nếu bệnh nhân không bị đau tim, và 1 nếu bệnh nhân bị đau tim
- ever_married:
 - Tình trạng hôn nhân của bệnh nhân
 - Là biến định tính rời rạc kiểu String
 - Nhận 1 trong các giá trị: “No” nếu chưa kết hôn hoặc “Yes” nếu đã kết hôn
- work_type:
 - Loại nghề nghiệp của bệnh nhân
 - Là biến định tính rời rạc kiểu String
 - Nhận 1 trong các giá trị: “children”, “Govt_jov”, “Never_worked”, “Private” hoặc “Self-employed”
- Residence_type:
 - Nơi cư trú của bệnh nhân
 - Là biến định tính rời rạc kiểu String
 - Nhận 1 trong các giá trị: “Rural” hoặc “Urban”
- avg_glucose_level:
 - Lượng đường trung bình trong máu của bệnh nhân
 - Là biến định lượng liên tục kiểu số
- bmi:
 - Chỉ số đo lường cơ thể BMI của bệnh nhân
 - Là biến định lượng liên tục kiểu số
- smoking_status:
 - Tình trạng hút thuốc của bệnh nhân
 - Là biến định tính rời rạc kiểu String
 - Nhận 1 trong các giá trị: “formerly smoked”, “never smoked”, “smokes” or “Unknown”
- stroke:
 - Tình trạng đột quỵ của bệnh nhân
 - Là biến định tính rời rạc kiểu số

- Nhận 1 trong 2 giá trị: 0 nếu bệnh nhân không bị đột quỵ, và 1 nếu bệnh nhân bị đột quỵ.

3 Import các thư viện và dữ liệu

```

      id  gender   age  hypertension  heart_disease ever_married \
0    9046    Male  67.0           0            1        Yes
1   51676  Female  61.0           0            0        Yes
2   31112    Male  80.0           0            1        Yes
3   60182  Female  49.0           0            0        Yes
4   1665  Female  79.0           1            0        Yes

      work_type Residence_type  avg_glucose_level     bmi  smoking_status \
0          Private          Urban            228.69  36.6  formerly smoked
1  Self-employed          Rural            202.21   NaN  never smoked
2          Private          Rural            105.92  32.5  never smoked
3          Private          Urban            171.23  34.4       smokes
4  Self-employed          Rural            174.12  24.0  never smoked

      stroke
0      1
1      1
2      1
3      1
4      1

```

4 Data cleaning

```

(5110, 12)
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5110 entries, 0 to 5109
Data columns (total 12 columns):
 #  Column          Non-Null Count  Dtype  
---  --  
 0  id              5110 non-null   int64  
 1  gender          5110 non-null   object  
 2  age              5110 non-null   float64 
 3  hypertension     5110 non-null   int64  
 4  heart_disease    5110 non-null   int64  
 5  ever_married     5110 non-null   object  
 6  work_type         5110 non-null   object  
 7  Residence_type    5110 non-null   object  
 8  avg_glucose_level 5110 non-null   float64 
 9  bmi              4909 non-null   float64 
 10  smoking_status   5110 non-null   object  
 11  stroke           5110 non-null   int64  

```

```
dtypes: float64(3), int64(4), object(5)
```

```
memory usage: 479.2+ KB
```

```
None
```

```
0    4861
```

```
1    249
```

```
Name: stroke, dtype: int64
```

Có 201 người bị thiếu thông tin về chỉ số BMI. 1 cách đơn giản ta có thể xóa các dòng dữ liệu này tuy nhiên cần kiểm tra lại để chắc chắn thông tin của 201 người này không ảnh hưởng đến kết quả dự đoán sau này

	id	gender	age	hypertension	heart_disease	ever_married	\
1	51676	Female	61.00	0	0	Yes	
8	27419	Female	59.00	0	0	Yes	
13	8213	Male	78.00	0	1	Yes	
19	25226	Male	57.00	0	1	No	
27	61843	Male	58.00	0	0	Yes	
29	69160	Male	59.00	0	0	Yes	
43	1845	Female	63.00	0	0	Yes	
46	37937	Female	75.00	0	1	No	
50	18587	Female	76.00	0	0	No	
51	15102	Male	78.00	1	0	Yes	
54	8752	Female	63.00	0	0	Yes	
57	66400	Male	78.00	0	0	Yes	
64	7356	Male	75.00	0	0	Yes	
70	70676	Female	76.00	0	0	Yes	
78	45805	Female	51.00	0	0	Yes	
81	26015	Female	66.00	0	0	Yes	
84	70042	Male	58.00	0	0	Yes	
105	2346	Male	58.00	0	0	Yes	
112	36706	Female	76.00	0	0	Yes	
124	14164	Female	72.00	0	0	Yes	
126	3352	Male	78.00	1	0	Yes	
129	48796	Female	75.00	0	0	Yes	
133	31563	Female	38.00	0	0	Yes	
146	41241	Male	65.00	0	0	Yes	
150	11933	Female	79.00	0	0	Yes	
160	50931	Female	76.00	0	0	Yes	
161	16590	Male	71.00	0	1	Yes	
162	69768	Female	1.32	0	0	No	
167	43364	Male	79.00	1	0	Yes	
170	28939	Male	64.00	0	0	Yes	
171	60739	Female	79.00	1	1	No	
174	40899	Female	78.00	0	0	Yes	
178	33486	Female	80.00	0	0	Yes	
183	8003	Female	77.00	0	0	No	

189	66955	Male	61.00	0	1	Yes
198	18937	Male	79.00	0	0	Yes
200	54695	Male	74.00	0	0	Yes
218	25904	Female	76.00	1	1	Yes
227	39105	Male	74.00	0	0	Yes
247	34060	Male	71.00	1	0	Yes
		work_type	Residence_type	avg_glucose_level	bmi	smoking_status \
1	Self-employed	Rural	202.21	NaN	never smoked	
8	Private	Rural	76.15	NaN	Unknown	
13	Private	Urban	219.84	NaN	Unknown	
19	Govt_job	Urban	217.08	NaN	Unknown	
27	Private	Rural	189.84	NaN	Unknown	
29	Private	Rural	211.78	NaN	formerly smoked	
43	Private	Urban	90.90	NaN	formerly smoked	
46	Self-employed	Urban	109.78	NaN	Unknown	
50	Private	Urban	89.96	NaN	Unknown	
51	Private	Urban	75.32	NaN	formerly smoked	
54	Govt_job	Urban	197.54	NaN	never smoked	
57	Private	Urban	237.75	NaN	formerly smoked	
64	Private	Urban	104.72	NaN	Unknown	
70	Govt_job	Rural	62.57	NaN	formerly smoked	
78	Private	Urban	165.31	NaN	never smoked	
81	Self-employed	Urban	101.45	NaN	Unknown	
84	Private	Urban	71.20	NaN	Unknown	
105	Private	Urban	82.30	NaN	smokes	
112	Self-employed	Urban	106.41	NaN	formerly smoked	
124	Private	Urban	219.91	NaN	Unknown	
126	Self-employed	Urban	93.13	NaN	formerly smoked	
129	Govt_job	Urban	62.48	NaN	Unknown	
133	Private	Rural	101.45	NaN	formerly smoked	
146	Self-employed	Urban	68.43	NaN	formerly smoked	
150	Private	Rural	169.67	NaN	Unknown	
160	Private	Urban	57.92	NaN	formerly smoked	
161	Private	Urban	81.76	NaN	smokes	
162	children	Urban	70.37	NaN	Unknown	
167	Private	Rural	75.02	NaN	never smoked	
170	Self-employed	Rural	111.98	NaN	formerly smoked	
171	Self-employed	Rural	60.94	NaN	never smoked	
174	Self-employed	Rural	60.67	NaN	formerly smoked	
178	Govt_job	Urban	110.66	NaN	Unknown	
183	Private	Urban	81.32	NaN	Unknown	
189	Private	Urban	209.86	NaN	Unknown	
198	Private	Rural	114.77	NaN	formerly smoked	
200	Private	Urban	167.13	NaN	Unknown	
218	Self-employed	Urban	199.86	NaN	smokes	
227	Self-employed	Rural	60.98	NaN	never smoked	

247	Self-employed	Rural	87.80	NaN	Unknown
-----	---------------	-------	-------	-----	---------

stroke

1	1
8	1
13	1
19	1
27	1
29	1
43	1
46	1
50	1
51	1
54	1
57	1
64	1
70	1
78	1
81	1
84	1
105	1
112	1
124	1
126	1
129	1
133	1
146	1
150	1
160	1
161	1
162	1
167	1
170	1
171	1
174	1
178	1
183	1
189	1
198	1
200	1
218	1
227	1
247	1

Thấy rằng có 40 người thiếu thông tin BMI đã từng bị đột quỵ (trên tổng số 249 người bị đột quỵ) là một tỉ lệ lớn, nên để chắc chắn ta sẽ thay thế các dữ liệu BMI thiếu bằng trung bình của BMI các dữ liệu quan sát.

```

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5110 entries, 0 to 5109
Data columns (total 12 columns):
 #   Column            Non-Null Count  Dtype  
--- 
 0   id                5110 non-null    int64  
 1   gender             5110 non-null    object  
 2   age                5110 non-null    float64 
 3   hypertension        5110 non-null    int64  
 4   heart_disease      5110 non-null    int64  
 5   ever_married       5110 non-null    object  
 6   work_type          5110 non-null    object  
 7   Residence_type     5110 non-null    object  
 8   avg_glucose_level  5110 non-null    float64 
 9   bmi                5110 non-null    float64 
 10  smoking_status     5110 non-null    object  
 11  stroke              5110 non-null    int64  
dtypes: float64(3), int64(4), object(5)
memory usage: 479.2+ KB

```

Thống kê mô tả dữ liệu

	id	age	hypertension	heart_disease	\
count	5110.000000	5110.000000	5110.000000	5110.000000	
mean	36517.829354	43.226614	0.097456	0.054012	
std	21161.721625	22.612647	0.296607	0.226063	
min	67.000000	0.080000	0.000000	0.000000	
25%	17741.250000	25.000000	0.000000	0.000000	
50%	36932.000000	45.000000	0.000000	0.000000	
75%	54682.000000	61.000000	0.000000	0.000000	
max	72940.000000	82.000000	1.000000	1.000000	

	avg_glucose_level	bmi	stroke	
count	5110.000000	5110.000000	5110.000000	
mean	106.147677	28.893237	0.048728	
std	45.283560	7.698018	0.215320	
min	55.120000	10.300000	0.000000	
25%	77.245000	23.800000	0.000000	
50%	91.885000	28.400000	0.000000	
75%	114.090000	32.800000	0.000000	
max	271.740000	97.600000	1.000000	

4.1 ID

5110

Thấy rằng tổng số id bằng với số dòng dữ liệu, ở đây ta không cần định danh từng dòng dữ liệu nữa nên sẽ bỏ cột này đi.

(5110, 11)

4.2 Gender

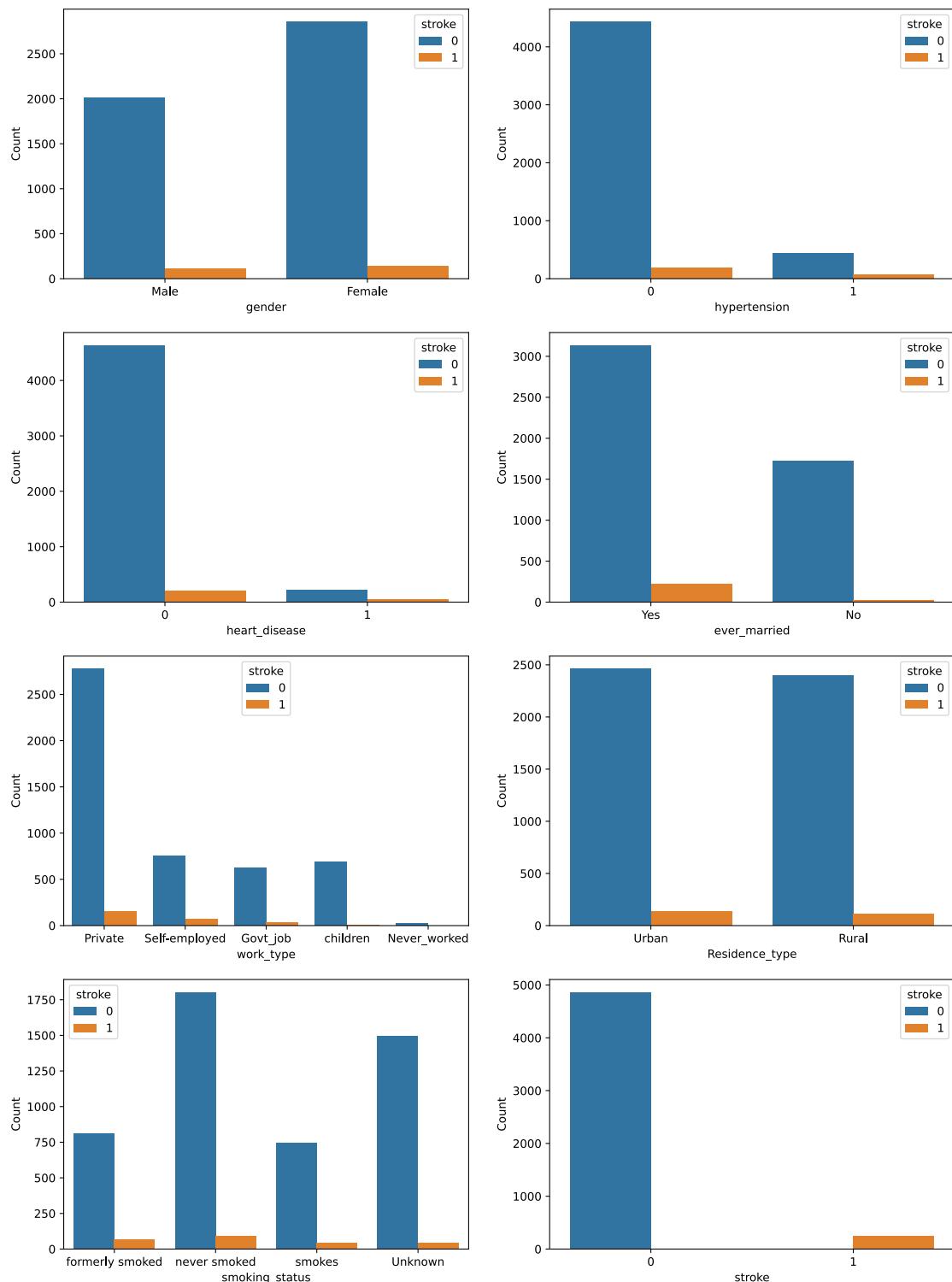
```
Female      2994  
Male       2115  
Other        1  
Name: gender, dtype: int64
```

Ta thấy giới tính “Other” ở đây chỉ có 1 dòng dữ liệu, nên sẽ sử dụng trung vị của Gender thay cho giới tính này.

```
Female      2995  
Male       2115  
Name: gender, dtype: int64
```

5 EDA

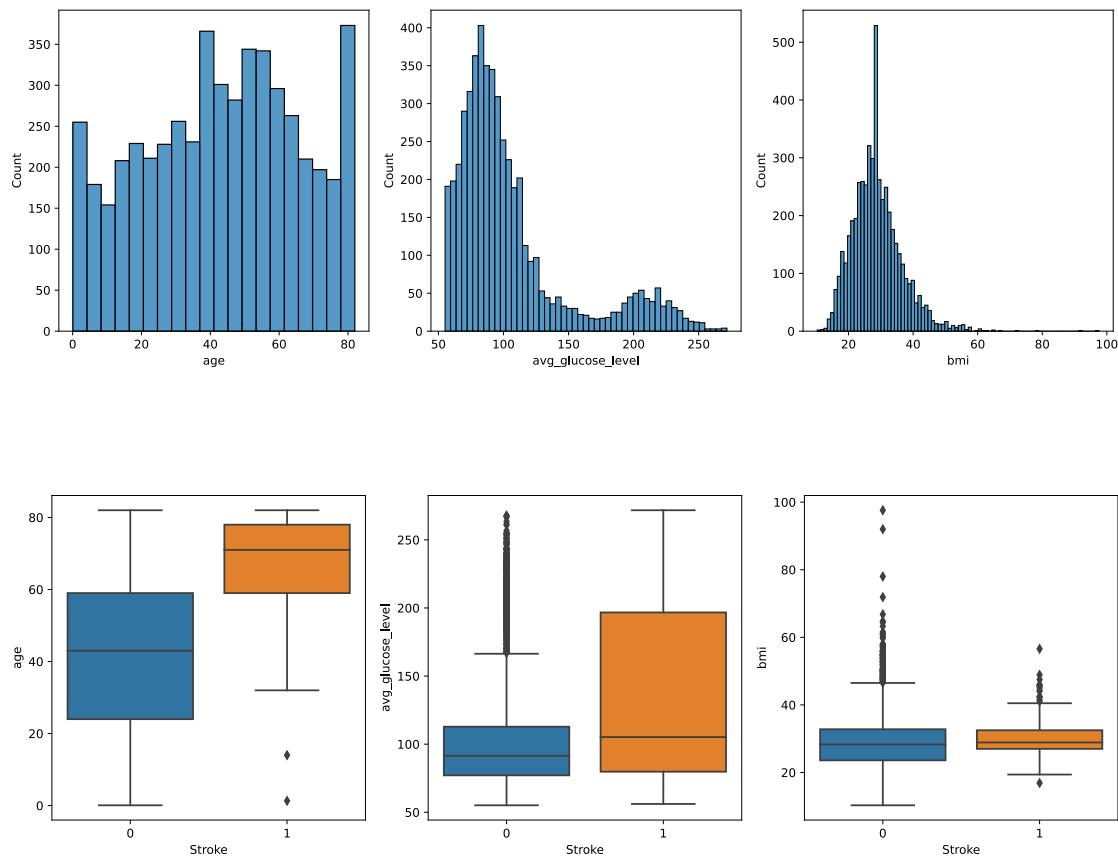
5.1 Phân tích biến định tính



Quan sát thấy:

- gender: Không có sự khác biệt nhiều giữa tỉ lệ người bị đột quỵ giữa 2 giới tính.
- hypertension: những người cao huyết áp có nguy cơ cao bị đột quỵ (trên biểu đồ do dữ liệu nhỏ nên đang thể hiện không rõ ràng - quan sát này dựa trên tìm hiểu cá nhân)
- heart_disease: những người được chẩn đoán mắc bệnh tim cũng có nguy cơ cao bị đột quỵ.
- ever_married: những người đã kết hôn có nguy cơ cao bị đột quỵ
- work_type: những người có kinh nghiệm làm việc và các công việc có liên quan đến nhà nước có nguy cơ cao bị đột quỵ, những người chưa đi làm hiếm khi bị đột quỵ.
- residence_type: không có mối liên quan rõ ràng giữa biến này với việc bị đột quỵ.
- smoking_status: Những người từng hút thuốc hoặc đang hút thuốc tăng nguy cơ bị đột quỵ.

5.2 Phân tích biến định lượng



Quan sát thấy:

- age: những người bị đột quỵ thường có độ tuổi trung bình cao hơn những người không bị. Những người ở độ tuổi >60 có nguy cơ bị đột quỵ cao hơn. Có 1 vài outlier là những người

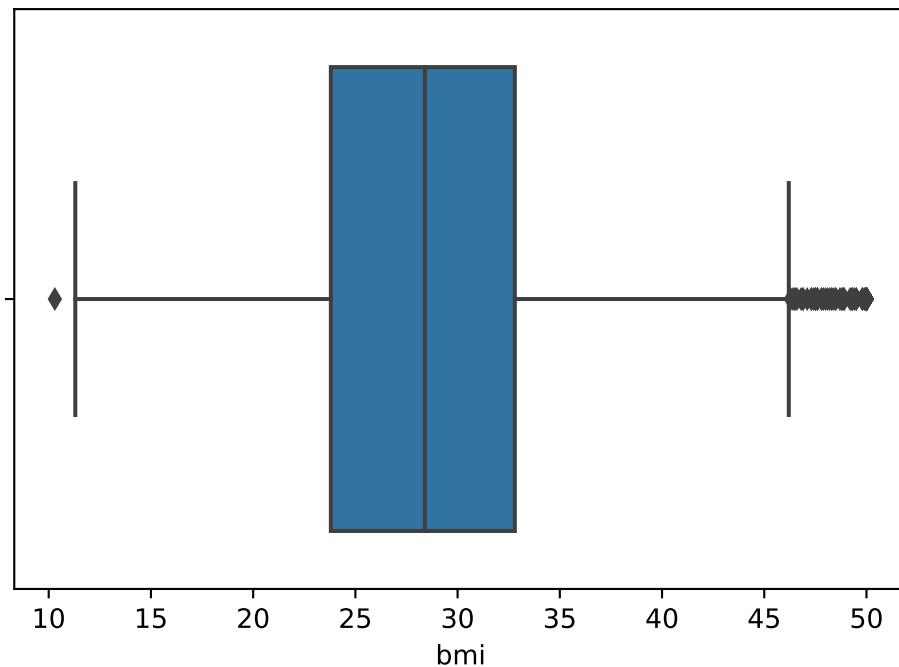
ở độ tuổi 20 bị đột quy, có thể coi là dữ liệu hợp lệ vì đột quy còn phụ thuộc vào lối sống và cách ăn uống. Những người không bị đột quy đều trong độ tuổi từ 20 - dưới 60.

- avg_glucose_level: những người bị đột quy có xu hướng trung bình đường huyết cao hơn những người không bị. Có khá nhiều outliers bệnh nhân không bị đột quy.
- bmi: chỉ số bmi không cho biết nhiều về khả năng bị đột quy. Tuy nhiên ta thấy có rất nhiều outlier ở biến này.

(79,)

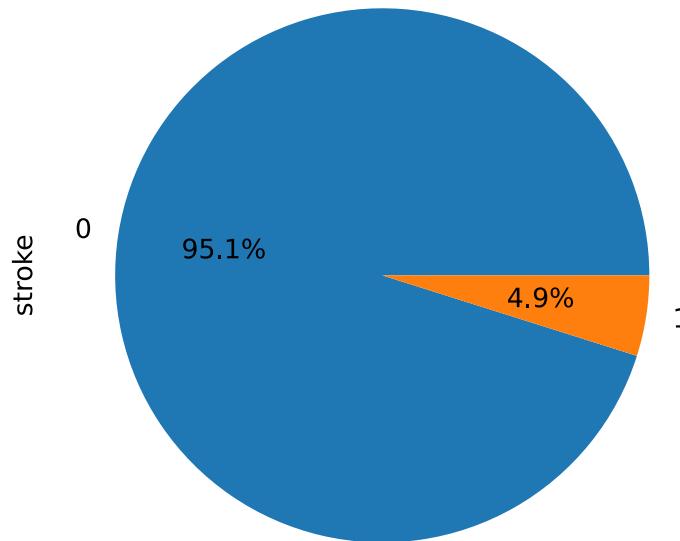
Có tất cả 79 outliers của bmi. Qua tìm hiểu thì thấy chỉ số BMI của những người rất béo phì là 50 -> ta sẽ thay thế các outlier bằng giá trị này để giảm thiểu số lượng outlier.

<AxesSubplot:xlabel='bmi'>



```
0      4861
1      249
Name: stroke, dtype: int64
```

Pie chart of stroke status



Ta thấy bộ dữ liệu có sự mất cân bằng khá lớn. Chỉ có 4.9% số quan sát trong mẫu đã từng bị đột quỵ.

5.3 Phân tích đa cộng tuyến

Việc phân tích tương quan chỉ chấp nhận các biến định lượng nên ta cần chuyển các biến định tính về các giá trị 0, 1. Ở đây ta dùng LabelEncoder từ thư viện sklearn.preprocessing để làm việc này.

	gender	age	hypertension	heart_disease	ever_married	work_type	\
0	1	67.0	0	1	1	2	
1	0	61.0	0	0	1	3	
2	1	80.0	0	1	1	2	
3	0	49.0	0	0	1	2	
4	0	79.0	1	0	1	3	

	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
0	1	228.69	36.600000	1	1
1	0	202.21	28.893237	2	1
2	0	105.92	32.500000	2	1
3	1	171.23	34.400000	3	1
4	0	174.12	24.000000	2	1

<AxesSubplot:>



Từ biểu đồ heatmap ở trên ta thấy có xuất hiện hiện tượng đa cộng tuyến giữa 1 vài biến. VD biến ever_married và age có tương quan lên đến 0.68. Giữa 2 biến này ta thấy age có tương quan với stroke cao hơn nên ta sẽ bỏ ever_married.

Đồng thời ta cũng thấy sự tương quan giữa biến phụ thuộc “stroke” với các biến “age”, “hypertension”, “heart_disease”, “avg_glucose_level”. Dưới đây là dữ liệu sau khi đã bỏ biến “ever_married”

	gender	age	hypertension	heart_disease	work_type	Residence_type	\
0	1	67.0		0	1	2	1
1	0	61.0		0	0	3	0
2	1	80.0		0	1	2	0
			avg_glucose_level	bmi	smoking_status	stroke	
0			228.69	36.600000		1	1
1			202.21	28.893237		2	1
2			105.92	32.500000		2	1

5.4 Normalize lại dữ liệu

Các biến được đo lường ở các tỉ lệ khác nhau sẽ không đóng góp như nhau vào việc fitting mô hình, và có thể tạo ra sai lệch. Để giải quyết vấn đề này, ta sử dụng các tiêu chuẩn ($\mu = 0, \sigma = 1$) để chuẩn

hóa lại các tỉ lệ này. Ở đây ta dùng thư viện StandardScaler() cho các biến “avg_glucose_level”, “bmi”, “age”

	avg_glucose_level	bmi	age	\	
0	2.706375	1.066746	1.051434		
1	2.121559	0.013363	0.786070		
2	-0.005028	0.506346	1.626390		
3	1.437358	0.766044	0.255342		
4	1.501184	-0.655458	1.582163		
gender	hypertension	heart_disease	work_type	Residence_type	\
0	1	0	1	2	1
1	0	0	0	3	0
2	1	0	1	2	0
3	0	0	0	2	1
4	0	1	0	3	0
smoking_status	stroke	avg_glucose_level	bmi	age	\
0	1	1	2.706375	1.066746	1.051434
1	2	1	2.121559	0.013363	0.786070
2	2	1	-0.005028	0.506346	1.626390
3	3	1	1.437358	0.766044	0.255342
4	2	1	1.501184	-0.655458	1.582163

6 Phân tích dữ liệu nhiều biến

6.1 Phân tích nhân tố PCA

Principal Component Analysis (PCA) là một kỹ thuật giảm chiều tuyến tính được sử dụng để trích xuất thông tin từ không gian nhiều chiều bằng cách chiếu vào không gian con ít chiều hơn. PCA cố gắng bảo toàn các phần cơ bản có nhiều biến đổi của dữ liệu và loại bỏ các thành phần không cơ bản với ít biến đổi hơn

Chiều của dữ liệu chính là các tính năng (features) đại diện của dữ liệu

PCA là kỹ thuật giảm chiều không giám sát, có thể phân cụm các điểm dữ liệu tương tự dựa trên mối tương quan tính năng giữa chúng mà không cần tới giám sát (hoặc nhãn) nào cần gán trước.

Người ta sử dụng PCA nhằm 2 mục đích: - Trực quan hóa dữ liệu: Khi làm việc với dữ liệu, 1 vấn đề đặt ra là khó khăn trong số lượng lớn dữ liệu và các tính năng định nghĩa dữ liệu. Để giải quyết vấn đề này, ta cần khám phá dữ liệu, tìm ra các tính năng tương quan với nhau thế nào hoặc hiểu phân bố của 1 vài tính năng. Khi trực quan hóa dữ liệu có nhiều biến (chiều) thì việc trực quan hóa này khá khó khăn hoặc hầu như không thể. Do đó PCA giúp ta làm việc này vì PCA chiều dữ liệu lên không gian ít chiều hơn, cho phép ta trực quan hóa dữ liệu lên không gian 2D hoặc 3D - Tăng tốc các thuật toán học máy: Vì ý tưởng chính của PCA để giảm chiều dữ liệu, ta có thể gia tăng tốc độ huấn luyện các thuật toán học máy và thời gian testing 1 cách đáng kể so với lúc dữ liệu nguyên bản có nhiều nhiều và tốc độ huấn luyện của các thuật toán lúc này là rất chậm

Các nhân tố chính (principal components) là chìa khóa của PCA, đại diện cho những gì bên dưới

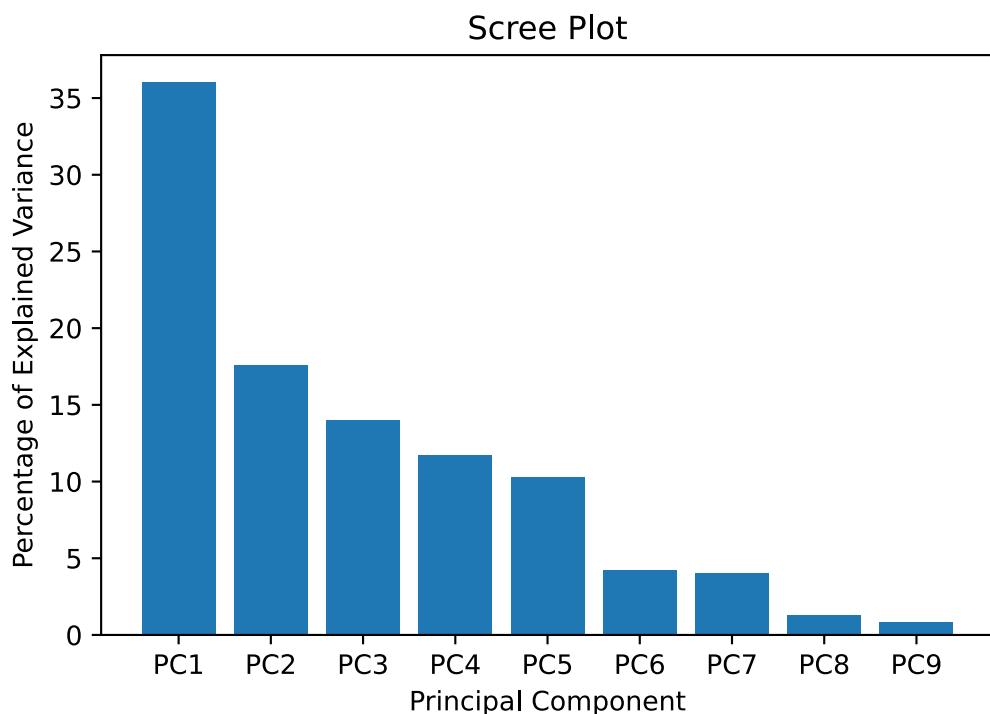
lớp dữ liệu của bạn. Khi dữ liệu được chiếu vào 1 chiều thấp hơn (giả sử 3 chiều) từ 1 không gian nhiều chiều hơn, thì 3 chiều này không gì khác ngoài 3 thành phần chính nắm (hoặc lưu giữ) hầu hết các phương sai (thông tin) của tập dữ liệu.

a. Sử dụng PCA để trực quan hóa dữ liệu:

```
Explained variation per principal component: [0.35955993 0.17614438 0.14043711
0.11736675 0.10348858 0.04188546
0.04023012 0.0131896 0.00769807]
```

Sau khi có các thành phần chính, ta có thể tìm explained_variance_ratio. Số này cung cấp cho chúng ta về tỉ lệ thông tin hoặc phương sai của mỗi thành phần chính giữ sau khi chiếu dữ liệu lên không gian con ít chiều hơn.

Từ số liệu trên, ta quan sát được principal component 1 chiếm tới 35,9% thông tin, trong khi principal component 2 chiếm 17,6% thông tin. Đồng thời, lưu ý rằng còn 46,5% thông tin đang bị mất nắm ở chiều dữ liệu khác



	gender	hypertension	heart_disease	work_type	Residence_type	\
0	-0.019819	0.047464	0.024368	-0.545629	0.004076	
1	0.049757	0.050066	0.038031	0.336096	-0.004251	
2	0.001022	0.016879	0.008088	0.472189	-0.000512	
3	-0.010044	0.008502	-0.020787	0.470637	-0.001463	
4	-0.011826	0.058605	0.054201	0.379038	0.012664	
5	0.113485	0.012236	0.005562	-0.002302	-0.993291	
6	0.990652	0.017888	0.045596	-0.019861	0.114396	

```

7 -0.021671      0.994834      0.028742 -0.025308      0.009433
8 -0.046359     -0.035825      0.995785 -0.012473     -0.000615

```

	smoking_status	avg_glucose_level	bmi	age
0	0.461444	0.225830	0.444392	0.487461
1	-0.392783	0.806992	0.166784	0.217395
2	0.791523	0.271302	-0.248364	-0.122013
3	0.057646	-0.299673	0.816647	-0.133618
4	-0.044564	-0.361608	-0.213273	0.819406
5	0.005519	-0.012191	-0.002076	0.011190
6	0.027257	-0.043988	0.006957	0.001302
7	-0.014102	-0.031983	-0.019404	-0.081736
8	0.000037	-0.022118	0.013682	-0.064393

Ta sắp xếp theo thứ tự giảm dần các tính năng (feature) của dữ liệu để độ ảnh hưởng tới biến đổi dữ liệu. Ta được bảng như sau:

```

work_type          -0.545629
age                0.487461
smoking_status    0.461444
bmi                0.444392
avg_glucose_level 0.225830
hypertension       0.047464
heart_disease     0.024368
gender             -0.019819
Residence_type    0.004076
dtype: float64

```

Trực quan 5110 mẫu trên trục tọa độ với principal component 1 và principal component 2, từ đó giúp ta nhìn được các mẫu phân bố thế nào giữa 2 classes

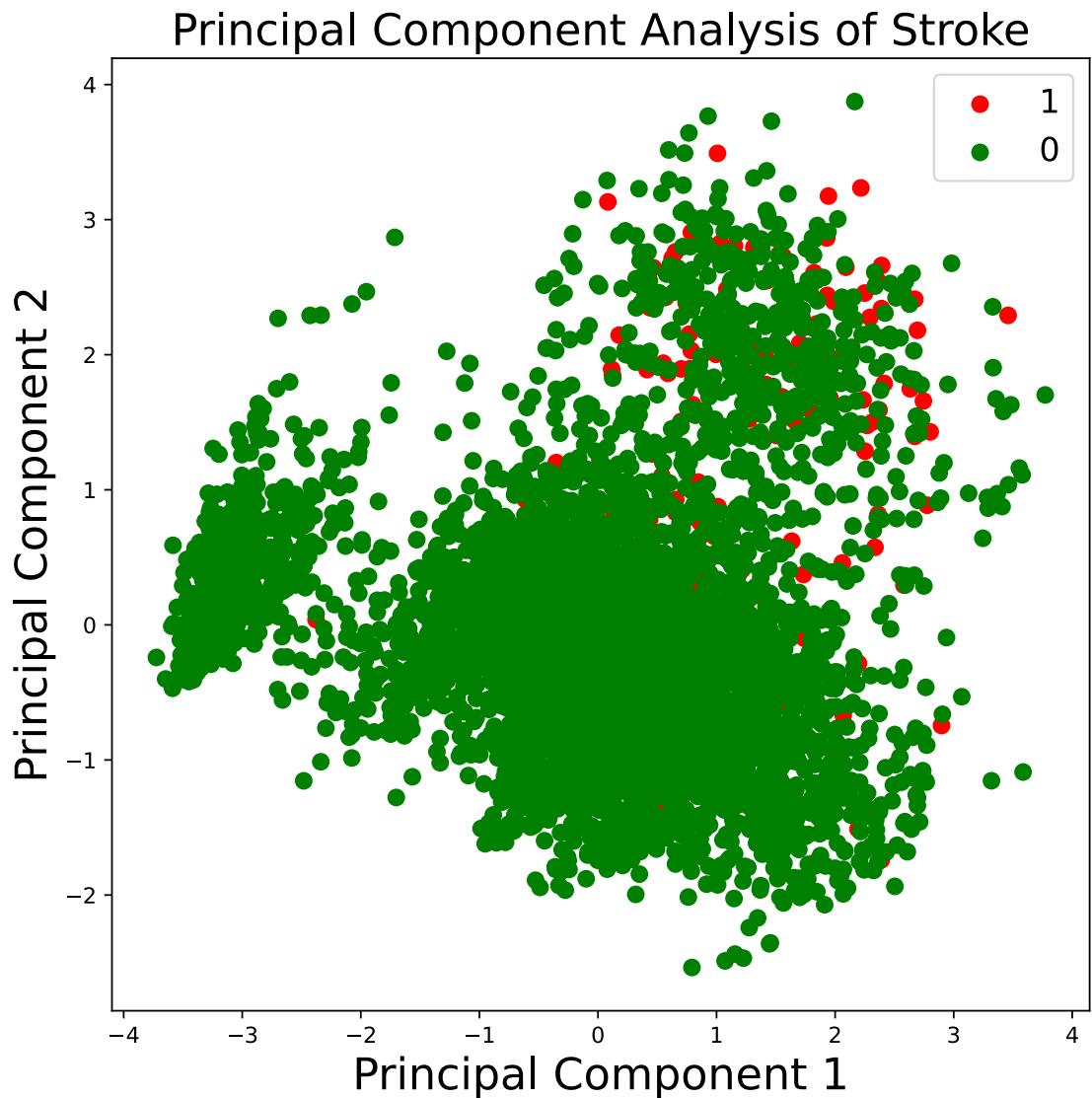
	principal component 1	principal component 2
5105	1.117793	-0.324046
5106	1.427624	0.963541
5107	-0.349658	-0.441069
5108	0.171409	1.189852
5109	0.306611	-0.641000

Explained variation per principal component: [0.35955993 0.17614438]

Từ số liệu trên, ta quan sát được principal component 1 chiếm tới 35.9% thông tin, trong khi principal component 2 chiếm 17.6% thông tin

<matplotlib.legend.Legend at 0x7fb40e269dc0>

<Figure size 432x288 with 0 Axes>



b. Tăng tốc các thuật toán học máy:

Trước tiên tạo 1 instance của mô hình PCA. Sau đó truyền vào bao nhiêu biến đổi mà bạn muốn PCA ghi lại. Ví dụ truyền vào tham số 0.95 cho mô hình PCA có nghĩa là PCA sẽ giữ 95% của phương sai và sau đó sẽ tính ra con số n_components là số chiều cần giảm tới để giữ được 95% phương sai đó.

```
Explained variation per principal component: [0.35955993 0.17614438 0.14043711
0.11736675 0.10348858 0.04188546
0.04023012]
```

	gender	hypertension	heart_disease	work_type	Residence_type	\
0	-0.019819	0.047464	0.024368	-0.545629	0.004076	
1	0.049757	0.050066	0.038031	0.336096	-0.004251	

```

2 0.001022      0.016879      0.008088      0.472189      -0.000512
3 -0.010044     0.008502     -0.020787      0.470637      -0.001463
4 -0.011826     0.058605      0.054201      0.379038      0.012664
5 0.113485      0.012236      0.005562     -0.002302     -0.993291
6 0.990652      0.017888      0.045596     -0.019861      0.114396

```

	smoking_status	avg_glucose_level	bmi	age
0	0.461444	0.225830	0.444392	0.487461
1	-0.392783	0.806992	0.166784	0.217395
2	0.791523	0.271302	-0.248364	-0.122013
3	0.057646	-0.299673	0.816647	-0.133618
4	-0.044564	-0.361608	-0.213273	0.819406
5	0.005519	-0.012191	-0.002076	0.011190
6	0.027257	-0.043988	0.006957	0.001302

7

Vậy chỉ cần 7 features trong top 10 features ở trên ta có thể giữ lại 95% lượng thông tin.

```

work_type          -0.545629
age                0.487461
smoking_status     0.461444
bmi                0.444392
avg_glucose_level 0.225830
hypertension        0.047464
heart_disease      0.024368
dtype: float64

```

6.2 Áp dụng KMeans để phân cụm dữ liệu

Tuy đã lựa chọn các biến dựa trên correlation ở trên nhưng để chắc chắn, ta sử dụng thư viện SelectKBest, f_classif để đánh lại điểm số mức độ ảnh hưởng của các biến đến biến phụ thuộc, sắp xếp theo thứ tự ảnh hưởng giảm dần.

	Attribute	Score
8	age	326.916568
2	heart_disease	94.698406
6	avg_glucose_level	90.503870
1	hypertension	84.953542
7	bmi	9.541558
3	work_type	5.340019
5	smoking_status	4.043033
4	Residence_type	1.220842
0	gender	0.424625

Ta thấy top 7 features quan trọng khi sử dụng thư viện giống với top 7 features sau khi thực hiện PCA. Tuy có khác nhau về thứ tự ảnh hưởng.

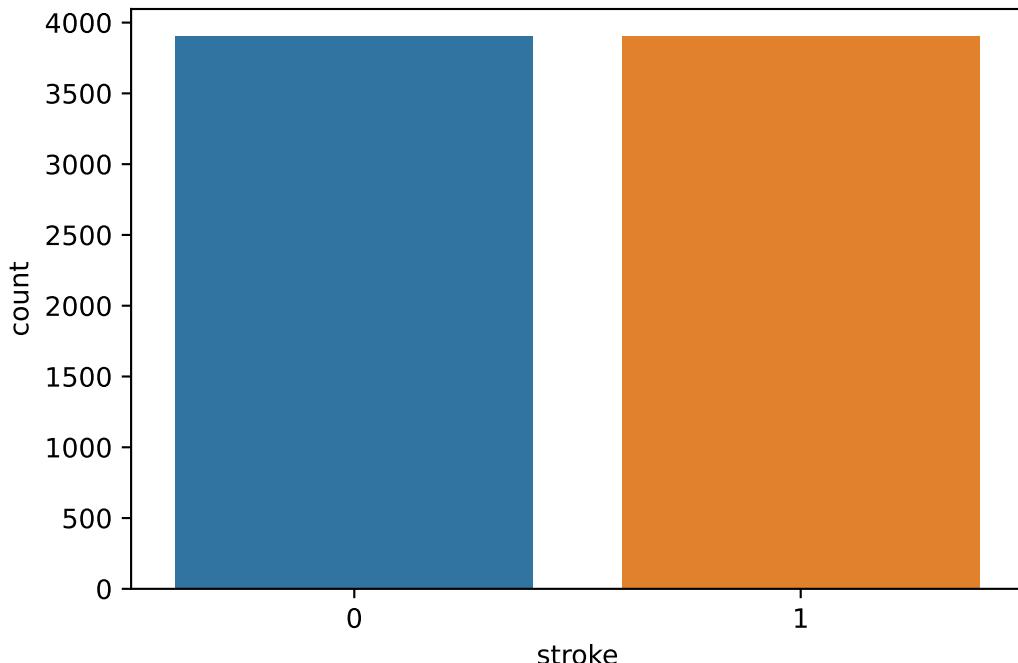
Chia tập dữ liệu

```
((4088, 7), (1022, 7), (4088,), (1022,))
```

Vì dữ liệu hiện tại mất cân bằng, nên ta sử dụng phương pháp **SMOTE** để cân bằng lại dữ liệu. Phương pháp này sẽ phân bổ lại dữ liệu với các bản ghi tương đồng với các bản ghi thuộc lớp có số lượng thiểu số.

```
((7802, 7), (7802,), (1920, 7), (1920,))
```

```
<AxesSubplot:xlabel='stroke', ylabel='count'>
```



Sau khi sử dụng thư viện SMOTE, ta thấy dữ liệu đã cân bằng trở lại.

```
hypertension  heart_disease  work_type  smoking_status  avg_glucose_level \
0            0              0           3                 1             0.143384
1            0              0           2                 0             -0.393728
2            0              0           2                 2             -1.029783
3            0              0           0                 1             -0.893296
4            0              0           2                 2             -1.027354

bmi          age
0 -0.040385  1.582163
1  1.025741  0.830297
2  0.670366 -0.983025
3  0.219312 -0.540751
4 -1.215858 -0.540751
```

6.3 Đánh giá

Để đánh giá được số cụm phù hợp với tập dữ liệu, ở đây ta chạy thuật toán thử nghiệm số lượng cụm từ 1-10, tâm cụm sinh ngẫu nhiên, sau đó so sánh SSE (sum-squared-error) (Phương pháp ELBOW).



Thư viện KneeLocator để tìm ra được điểm elbow point thay vì nhìn bằng mắt thường.

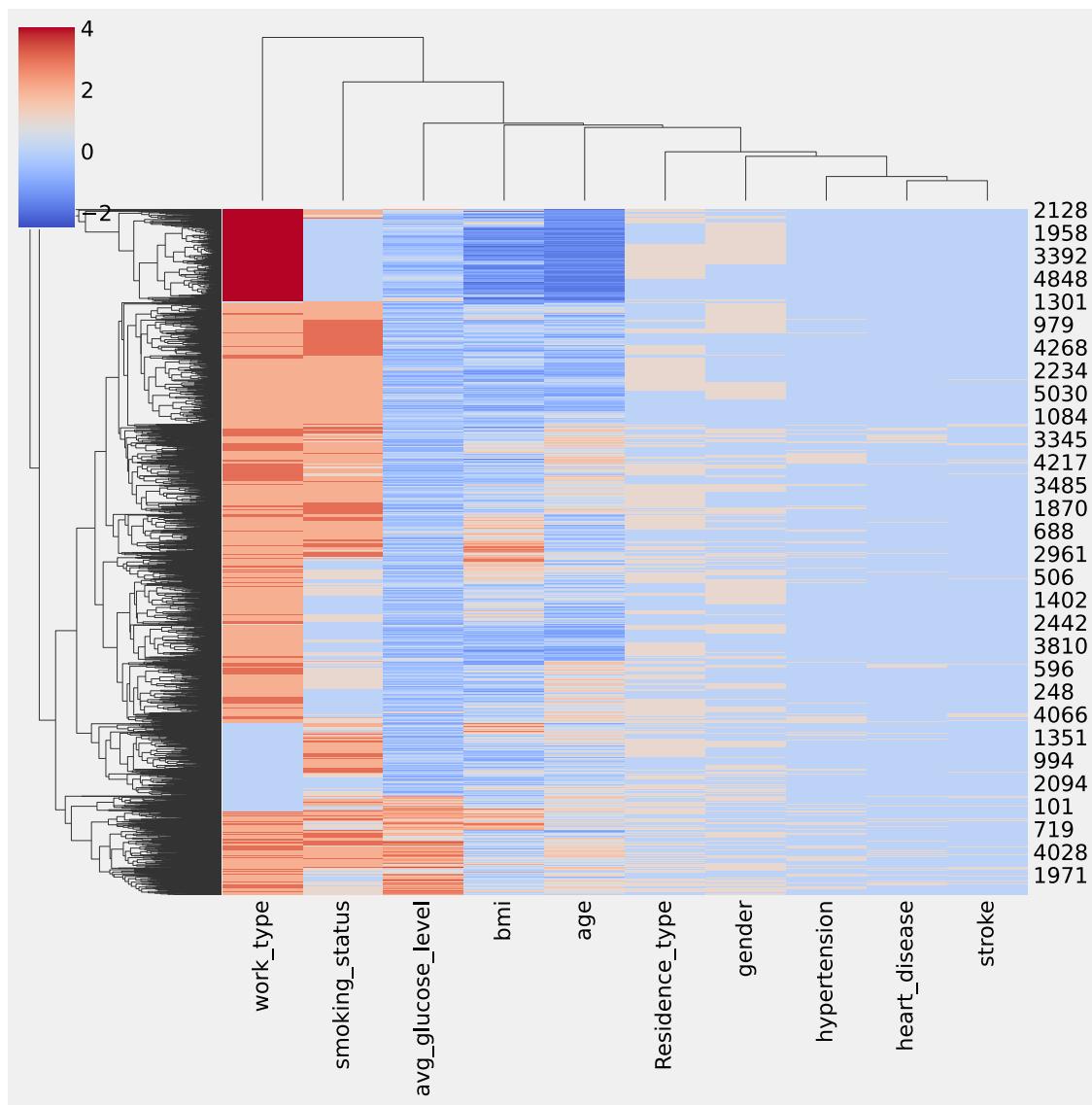
4

Hiện tại kết quả của nhóm sau khi chạy thuật toán đang trả ra số cụm lớn hơn số loại biến phụ thuộc thực tế. Nhóm đang tiếp tục tìm hiểu xem lý do.

Biểu đồ phân cụm phân cấp dữ liệu theo tương quan các biến.

```
/home/tm/.local/lib/python3.8/site-packages/seaborn/matrix.py:649: UserWarning:  
Clustering large matrix with scipy. Installing `fastcluster` may give better  
performance.  
warnings.warn(msg)
```

```
<seaborn.matrix.ClusterGrid at 0x7fb40e240df0>
```



<seaborn.matrix.ClusterGrid at 0x7fb3e2dc9040>

