TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM
TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG
KHOA CÔNG NGHỆ THÔNG TIN

BÁO CÁO CUỐI KỲ

KHAI THÁC DỮ LIỆU VÀ KHAI PHÁ TRI THỨC

DỰ ĐOÁN NGUY CƠ ĐỘT QUỴ BẰNG PHÂN LOẠI NHỊ PHÂN

Người hướng dẫn: **ThS Hoàng Anh**

Người thực hiện: **NGUYỄN KHẮC ANH TÀI – 52100306**

LÊ TUẤN THÀNH – 52100312

TRẦN NAM ĐĂNG KHOA – 52100239

PHẠM HOÀNG TRUNG KIÊN - 52100904

TRẦN QUANG LUÂN - 52100254

NGUYỄN ĐỨC MINH - 52100977

Lớp **: 21050301**

Khoá **: 25**

THÀNH PHỐ HỒ CHÍ MINH, NĂM 2024

TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM
TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG
KHOA CÔNG NGHỆ THÔNG TIN

BÁO CÁO CUỐI KỲ

KHAI THÁC DỮ LIỆU VÀ KHAI PHÁ TRI THỨC

DỰ ĐOÁN NGUY CƠ ĐỘT QUỴ BẰNG PHÂN LOẠI NHỊ PHÂN

Người hướng dẫn: **ThS Hoàng Anh**

Người thực hiện: **NGUYỄN KHẮC ANH TÀI – 52100306**

LÊ TUẤN THÀNH – 52100312

TRẦN NAM ĐĂNG KHOA – 52100239

PHẠM HOÀNG TRUNG KIÊN - 52100904

TRẦN QUANG LUÂN - 52100254

NGUYỄN ĐỨC MINH - 52100977

Lớp **: 21050301**

Khoá **: 25**

THÀNH PHỐ HỒ CHÍ MINH, NĂM 2024

LỜI CẢM ƠN

Lời cảm ơn đầu tiên của chúng em được gửi đến các thầy cô giảng dạy trong trường ĐH Tôn Đức Thắng. Chúng em muốn gửi lời cảm ơn sâu sắc nhất đến những người đã truyền dạy cho chúng em những kiến thức quý báu, giúp chúng em phát triển và tiến bộ trong cuộc sống. Những giáo viên tận tâm và nhiệt tình đã luôn sẵn sàng giúp đỡ chúng em trong suốt quá trình học tập.

Đặc biệt, chúng em muốn gửi lời cảm ơn đến ThS Hoàng Anh, giảng viên bộ môn Khai Thác Dữ Liệu Và Khai Phá Tri Thức. Thầy đã truyền đạt những kiến thức bổ ích và kinh nghiệm thực tiễn giúp chúng em hiểu rõ hơn về đề tài và hoàn thành bài tiểu luận một cách tốt nhất.

Cuối cùng, chúng em xin chân thành cảm ơn các thầy cô trong trường ĐH Tôn Đức Thắng đã tạo điều kiện tốt nhất cho chúng em trong suốt thời gian học tập tại trường. Chúng em hy vọng sẽ có cơ hội được tiếp tục học tập và phát triển bản thân trong tương lai.

ĐỒ ÁN ĐƯỢC HOÀN THÀNH TẠI TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG

Tôi xin cam đoan đây là sản phẩm đồ án của riêng tôi / chúng tôi và được sự hướng dẫn của ThS Hoàng Anh. Các nội dung nghiên cứu, kết quả trong đề tài này là trung thực và chưa công bố dưới bất kỳ hình thức nào trước đây. Những số liệu trong các bảng biểu phục vụ cho việc phân tích, nhận xét, đánh giá được chính tác giả thu thập từ các nguồn khác nhau có ghi rõ trong phần tài liệu tham khảo.

Ngoài ra, trong đồ án còn sử dụng một số nhận xét, đánh giá cũng như số liệu của các tác giả khác, cơ quan tổ chức khác đều có trích dẫn và chú thích nguồn gốc.

Nếu phát hiện có bất kỳ sự gian lận nào tôi xin hoàn toàn chịu trách nhiệm về nội dung đồ án của mình. Trường đại học Tôn Đức Thắng không liên quan đến những vi phạm tác quyền, bản quyền do tôi gây ra trong quá trình thực hiện (nếu có).

TP. Hồ Chí Minh, ngày 16 tháng 5 năm 2024

Tác giả

(ký tên và ghi rõ họ tên)

Nguyễn Khắc Anh Tài

Lê Tuấn Thành

Trần Nam Đăng Khoa

Phạm Hoàng Trung Kiên

Trần Quang Luân

Nguyễn Đức Minh

PHẦN XÁC NHẬN VÀ ĐÁNH GIÁ CỦA GIẢNG VIÊN

Phần xác nhận của GV hướng dẫn

_____

_____

_____

_____

_____

_____

_____

Tp. Hồ Chí Minh, ngày tháng năm
(kí và ghi họ tên)

Phần đánh giá của GV chấm bài

_____

_____

_____

_____

_____

_____

_____

Tp. Hồ Chí Minh, ngày tháng năm
(kí và ghi họ tên)

TÓM TẮT

Báo cáo này tập trung vào nghiên cứu mô hình phân loại nhị phân để dự đoán nguy cơ đột quỵ dựa trên dữ liệu bảng. Đột quỵ, một trong những nguyên nhân hàng đầu gây tử vong trên toàn thế giới, đặt ra một thách thức lớn đối với ngành y tế và xã hội. Việc dự đoán nguy cơ đột quỵ có thể giúp trong việc sàng lọc và điều trị sớm, từ đó giảm thiểu nguy cơ và hậu quả nghiêm trọng của bệnh.

Chúng tôi sử dụng một tập dữ liệu chi tiết bao gồm thông tin về tuổi, giới tính, tiền sử bệnh lý như huyết áp cao và bệnh tim mạch, thông tin về hút thuốc, cũng như các chỉ số sinh học như mức đường huyết trung bình và chỉ số BMI. Mục tiêu của chúng tôi là xây dựng một mô hình phân loại hiệu quả để dự đoán xác suất một cá nhân có nguy cơ đột quỵ cao hay thấp dựa trên các yếu tố này.

Báo cáo này bao gồm một phân tích chi tiết về việc tiền xử lý dữ liệu, lựa chọn và đánh giá mô hình, cùng với việc giải thích sâu hơn về cách các đặc trưng ảnh hưởng đến dự đoán. Kết quả được thảo luận cung cấp một cái nhìn tổng quan về hiệu suất của mô hình và các yếu tố quan trọng trong việc dự đoán nguy cơ đột quỵ.

MỤC LỤC

DANH MỤC CÁC BẢNG BIỂU, HÌNH VẼ, ĐỒ THỊ

DANH MỤC HÌNH

DANH MỤC BẢNG

CHƯƠNG 1 – GIỚI THIỆU VÀ TỔNG QUAN VỀ ĐỀ TÀI

1.1. Giới thiệu về bệnh đột quỵ

1.1.1. Định nghĩa bệnh đột quỵ

Đột quỵ còn được gọi là tai biến mạch máu não. Đây là tình trạng não bộ bị tổn thương nghiêm trọng do quá trình cấp máu não bị gián đoạn hoặc giảm đáng kể khiến não bộ bị thiếu oxy, không đủ dinh dưỡng để nuôi các tế bào. Trong vòng vài phút nếu không được cung cấp đủ máu các tế bào não sẽ bắt đầu chết.

Do đó, người bị đột quỵ cần được cấp cứu ngay lập tức, thời gian kéo dài càng lâu, số lượng tế bào não chết càng nhiều sẽ ảnh hưởng lớn tới khả năng vận động và tư duy của cơ thể, thậm chí là tử vong. Hầu hết những người sống sót sau cơn đột quỵ đều có sức khỏe suy yếu hoặc mắc các di chứng như: tê liệt hoặc cử động yếu một phần cơ thể, mất ngôn ngữ, rối loạn cảm xúc, thị giác suy giảm...

Có 2 loại đột quỵ là đột quỵ do thiếu máu và đột quỵ do xuất huyết:

- Đột quỵ do thiếu máu cục bộ: Chiếm khoảng 85% tổng số các ca bị đột quỵ hiện nay. Đây là tình trạng đột quỵ do các cục máu đông làm tắc nghẽn động mạch, cản trở quá trình máu lưu thông lên não.
- Đột quỵ do xuất huyết: Đột quỵ do xuất huyết là tình trạng mạch máu đến não bị vỡ khiến máu chảy ồ ạt gây xuất huyết não. Nguyên nhân khiến mạch máu vỡ là do thành động mạch mỏng yếu hoặc xuất hiện các vết nứt, rò rỉ.

Ngoài ra, người bệnh có thể gặp phải cơn thiếu máu não thoáng qua. Đây là tình trạng đột quỵ nhỏ, dòng máu cung cấp cho não bộ bị giảm tạm thời. Người bệnh có những triệu chứng của đột quỵ nhưng chỉ diễn ra trong thời gian rất ngắn, thường kéo dài khoảng vài phút. Đây là dấu hiệu cảnh báo nguy cơ đột quỵ có thể xảy ra bất cứ lúc nào mà người bệnh cần lưu ý.

1.1.2. Dấu hiệu đột quỵ

Nhiều nước trên thế giới hiện đưa ra chữ "FAST" (2) để phổ cập các dấu hiệu của đột quỵ. "FAST" có nghĩa là nhanh (phản ứng tức thời), đồng thời là chữ viết tắt của Face (khuôn mặt), Arm (tay), Speech (lời nói) và Time (thời gian).

- Khuôn mặt: Dấu hiệu dễ nhìn thấy là mặt bệnh nhân bị méo. Nếu nghi ngờ hãy yêu cầu bệnh nhân cười vì méo có thể rõ hơn.
- Tay: Tay bị liệt, cũng có thể có diễn tiến từ từ như tê một bên tay, vẫn điều khiển được tay nhưng kém chính xác. Ngoài tay còn có một số dấu hiệu ở chân như nhấc chân không lên, đi rớt dép,….
- Lời nói: Rõ nhất là một số người đột quỵ bị "á khẩu" hay nói đớ.
- Thời gian: Đưa bệnh nhân bệnh viện khám ngay khi ghi nhận những dấu hiệu vừa kể.

Ngoài ra có những triệu chứng ở người bị đột quỵ có thể kể đến như:

- Lẫn lộn, sảng, hôn mê;
- Thị lực giảm sút, hoa mắt;
- Chóng mặt, người mất thăng bằng, không thể đứng vững;
- Đau đầu;
- Buồn nôn, nôn ói,….

1.1.3. Tầm quan trọng của việc dự đoán sớm bệnh đột quỵ

Dự đoán sớm đột quỵ có vai trò quan trọng trong việc cứu sống và giảm thiểu những tác động tiêu cực của bệnh. Nếu đột quỵ được nhận diện và điều trị kịp thời, có thể giảm thiểu tổn thương não và cải thiện cơ hội hồi phục hoàn toàn cho bệnh nhân. Việc dự đoán sớm cũng giúp nâng cao nhận thức của cộng đồng về các yếu tố nguy cơ và các biện pháp phòng ngừa, từ đó giảm tỷ lệ mắc bệnh và gánh nặng y tế lên hệ thống chăm sóc sức khỏe.

1.2. Mục tiêu bài báo cáo

1.2.1. Lý do chọn đề tài

Đột quỵ là một trong những nguyên nhân hàng đầu gây tử vong và tàn tật trên toàn thế giới. Với sự gia tăng của các yếu tố nguy cơ như lối sống ít vận động, chế độ ăn không lành mạnh và sự gia tăng của các bệnh mãn tính như tiểu đường và tăng huyết áp, số ca đột quỵ đang ngày càng tăng. Điều này đặt ra nhu cầu cấp thiết phải tìm ra các phương pháp dự đoán hiệu quả để can thiệp kịp thời. Đề tài này được chọn nhằm nghiên cứu và ứng dụng các phương pháp phân loại nhị phân để dự đoán nguy cơ đột quỵ, từ đó góp phần giảm thiểu gánh nặng bệnh tật cho xã hội.

1.2.2. Mục tiêu nghiên cứu và phạm vi

Mục Tiêu Nghiên Cứu: Mục tiêu chính của nghiên cứu này là ứng dụng các kỹ thuật phân loại nhị phân trong khai phá dữ liệu (data mining) để dự đoán nguy cơ đột quỵ dựa trên các yếu tố nguy cơ đã được xác định. Nghiên cứu sẽ tập trung vào việc xây dựng và đánh giá hiệu quả của các mô hình phân loại, từ đó lựa chọn mô hình tối ưu nhất.

Phạm Vi Nghiên Cứu: Nghiên cứu sẽ tập trung vào việc xử lý và phân tích dữ liệu dạng bảng (tabular data) từ các nguồn dữ liệu y tế công khai hoặc các cơ sở y tế. Các kỹ thuật phân loại nhị phân như Logistic Regression, Decision Trees, Random Forest, và SVM sẽ được so sánh và đánh giá.

1.2.3. Ý nghĩa của đề tài

Việc dự đoán nguy cơ đột quỵ đóng vai trò quan trọng trong việc sàng lọc, chẩn đoán sớm và điều trị hiệu quả bệnh này. Bằng cách sử dụng phân loại nhị phân để phân loại cá nhân vào nhóm có nguy cơ cao và nhóm có nguy cơ thấp, chúng ta có thể đưa ra các biện pháp phòng ngừa và can thiệp kịp thời, từ đó giảm thiểu tỷ lệ tử vong và tăng cường chất lượng cuộc sống.

CHƯƠNG 2 – MÔ TẢ VÀ KHÁM PHÁ BỘ DỮ LIỆU

2.1. Mô tả bộ dữ liệu

2.1.1. Ngữ cảnh

Theo Tổ chức Y tế Thế giới (WHO), đột quỵ là nguyên nhân gây tử vong đứng thứ 2 trên toàn cầu, chiếm khoảng 11% tổng số ca tử vong.

Tập dữ liệu này được sử dụng để dự đoán liệu một bệnh nhân có khả năng bị đột quỵ hay không dựa trên các thông số đầu vào như giới tính, tuổi tác, các bệnh khác nhau và tình trạng hút thuốc. Mỗi hàng trong dữ liệu cung cấp thông tin liên quan về bệnh nhân.

2.1.2. Các đặc trưng trong bộ dữ liệu

Bộ dữ liệu gồm 2 file train.csv và test.csv với các 12 đặc trưng được mô tả như sau:

1) *gender:* Giới tính của bệnh nhân, "Male", "Female" or "Other"
2) *age:* Tuổi bệnh nhân
3) *hypertension:* Chỉ số huyết áp, có giá trị là 0 nếu bệnh nhân không bị tăng huyết áp, 1 nếu bệnh nhân bị tăng huyết áp
4) *heart_disease:* Chỉ số về tim, có giá trị là 0 nếu bệnh nhân không mắc bệnh tim, 1 nếu bệnh nhân mắc bệnh tim
5) *ever_married:* Tình trạng kết hôn, "No" hoặc "Yes"
6) *work_type:* Loại công việc mà bệnh nhân đang làm, "children", "Govt_jov", "Never_worked", "Private" hoặc "Self-employed"
7) *Residence_type:* Khu vực cư trú, "Rural" hoặc "Urban", "Nông Thôn" hoặc "Thành Thị"
8) *avg_glucose_level:* Chỉ số lượng đường trung bình trong máu
9) *bmi:* Chỉ số BMI (Body Mass Index) được tính dựa trên tỉ lệ giữa cân nặng và chiều cao bình phương, nói lên tình trạng cân nặng hiện tại của

bệnh nhân. So với giá trị BMI tiêu chuẩn, chỉ số BMI cá nhân sẽ xác định một người đang thừa cân, thiếu cân hay có cân nặng cân đối.

10) *smoking_status:* Tình trạng hút thuốc "formerly smoked", "never smoked", "smokes" hoặc "Unknown"

11) *stroke:* Là đặc trưng mục tiêu của bài toán, có giá trị là 1 nếu bệnh nhân bị đột quỵ, và 0 nếu bệnh nhân không đột quỵ

2.2. Khám phá bộ dữ liệu

2.2.1. Tải bộ dữ liệu từ drive

Sử dụng thư viện Pandas trong Python để đọc các tệp CSV (Comma-Separated Values) và tạo DataFrame, một cấu trúc dữ liệu mạnh mẽ tương tự như bảng tính, cho dữ liệu huấn luyện và kiểm tra.

```
df_train = pd.read_csv('/content/drive/MyDrive/Data Mining/FinalProject/dataset/train.csv', index_col=0)
df_test = pd.read_csv('/content/drive/MyDrive/Data Mining/FinalProject/dataset/test.csv', index_col=0)
```

2.2.2. Các chỉ số và kiểu dữ liệu của các đặc trưng

Bộ dữ liệu dùng cho việc huấn luyện train.csv gồm 15304 dòng và 11 cột:

id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
0	Male	28.0	0	0	Yes	Private	Urban	79.53	31.1	never smoked	0
1	Male	33.0	0	0	Yes	Private	Rural	78.44	23.9	formerly smoked	0
2	Female	42.0	0	0	Yes	Private	Rural	103.00	40.3	Unknown	0
3	Male	56.0	0	0	Yes	Private	Urban	64.87	28.8	never smoked	0
4	Female	24.0	0	0	No	Private	Rural	73.36	28.8	never smoked	0

Hình 1. Bộ dữ liệu train.csv

Bộ dữ liệu dùng cho việc kiểm tra test.csv:

id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status
15304	Female	57.0	0	0	Yes	Private	Rural	82.54	33.4	Unknown
15305	Male	70.0	1	0	Yes	Private	Urban	72.06	28.5	Unknown
15306	Female	5.0	0	0	No	children	Urban	103.72	19.5	Unknown
15307	Female	56.0	0	0	Yes	Govt_job	Urban	69.24	41.4	smokes
15308	Male	32.0	0	0	Yes	Private	Rural	111.15	30.1	smokes

Hình 2. Bộ dữ liệu test.csv

Kiểu dữ liệu của các đặc trưng:

gender	object
age	float64
hypertension	int64
heart_disease	int64
ever_married	object
work_type	object
residence_type	object
avg_glucose_level	float64
bmi	float64
smoking_status	object
stroke	int64

Bảng 1: Kiểu dữ liệu của các đặc trưng

Trong 11 đặc trưng của bộ dữ liệu, có 6 đặc trưng số và 5 đặc trưng phân loại, bao gồm:

- 6 đặc trưng số: *age*, *hypertension*, *heart_disease*, *avg_glucose_level*, *bmi*, *stroke*.
- 5 đặc trưng phân loại: *gender*, *ever_married*, *work_type*, *Residence_type*, *smoking_status*.

Mô tả tổng quan về các đặc trưng số trong tập dữ liệu:

	age	hypertension	heart_disease	avg_glucose_level	bmi	stroke
count	15304.000000	15304.000000	15304.000000	15304.000000	15304.000000	15304.000000
mean	41.417708	0.049726	0.023327	89.039853	28.112721	0.041296
std	21.444673	0.217384	0.150946	25.476102	6.722315	0.198981
min	0.080000	0.000000	0.000000	55.220000	10.300000	0.000000
25%	26.000000	0.000000	0.000000	74.900000	23.500000	0.000000
50%	43.000000	0.000000	0.000000	85.120000	27.600000	0.000000
75%	57.000000	0.000000	0.000000	96.980000	32.000000	0.000000
max	82.000000	1.000000	1.000000	267.600000	80.100000	1.000000

Hình 3: Tổng quan về các đặc trưng số

CHƯƠNG 3 – TRỰC QUAN HÓA VÀ THỐNG KÊ CÁC THÔNG SỐ CỦA BỘ DỮ LIỆU

3.1. Kiểm tra các giá trị còn thiếu và giá trị trùng lặp trong dữ liệu

3.1.1. Kiểm tra giá trị còn thiếu

Sử dụng hàm *isnull()* để kiểm tra và đếm số lượng giá trị bị thiếu (missing values) trong từng cột của DataFrame:

```
1 df_train.isnull().sum()
```

```
gender              0
age                 0
hypertension        0
heart_disease       0
ever_married        0
work_type           0
Residence_type      0
avg_glucose_level   0
bmi                 0
smoking_status      0
stroke              0
dtype: int64
```

Không có giá trị bị thiếu trong tập dữ liệu này.

3.1.2. Kiểm tra giá trị trùng lặp

Sử dụng hàm *duplicated()* để kiểm tra và đếm số lượng giá trị bị thiếu (missing values) trong từng cột của DataFrame:

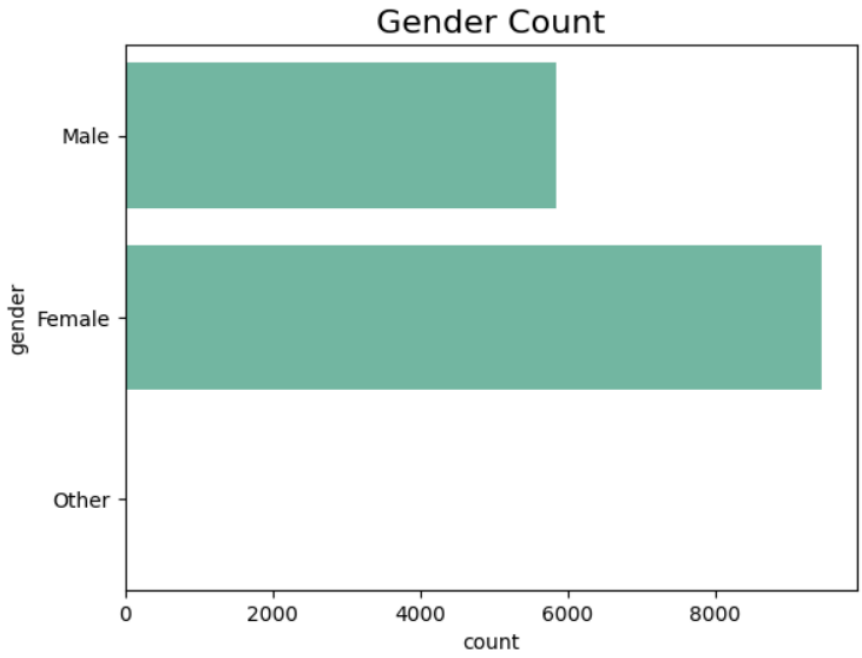
```
1 print("Số giá trị trùng lặp: ", df_train.duplicated().sum())
```

```
Số giá trị trùng lặp:  0
```

3.2. Trực quan hóa dữ liệu trên từng đặc trưng của bộ dữ liệu

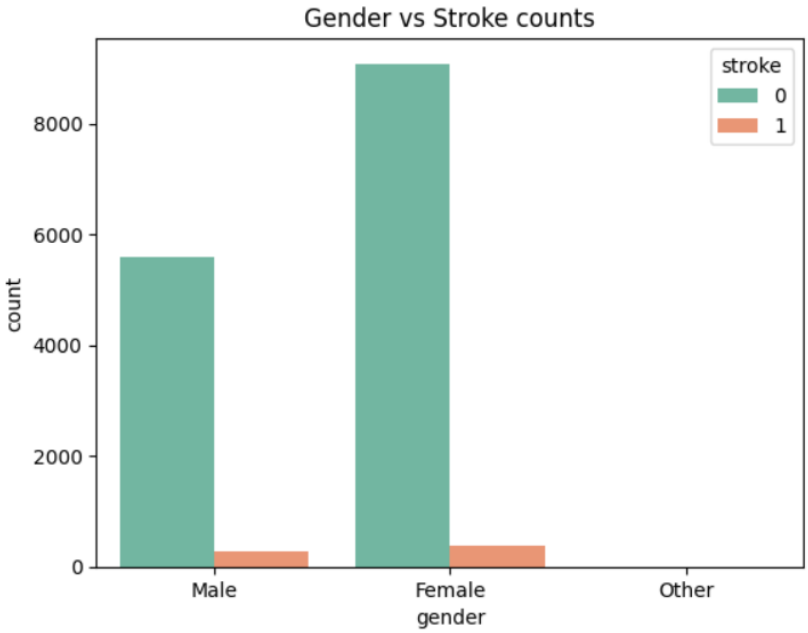
3.2.1. gender

- Đặc trưng có 3 giá trị : Male, Female, Other.



Hình 4: Biểu đồ thống kê giới tính

- Biểu đồ trên cho thấy trong tập dữ liệu thì giới tính nữ chiếm nhiều nhất, sau đó đến nam và giới tính khác gần như là không đáng kể.



Hình 5: Biểu đồ thống kê mối quan hệ giới tính và khả năng đột quỵ

- Bảng thống kê tỉ lệ đột quỵ theo giới tính:

stroke	0	1
gender		
Female	0.961042	0.038958
Male	0.954926	0.045074
Other	1.000000	0.000000
All	0.958704	0.041296

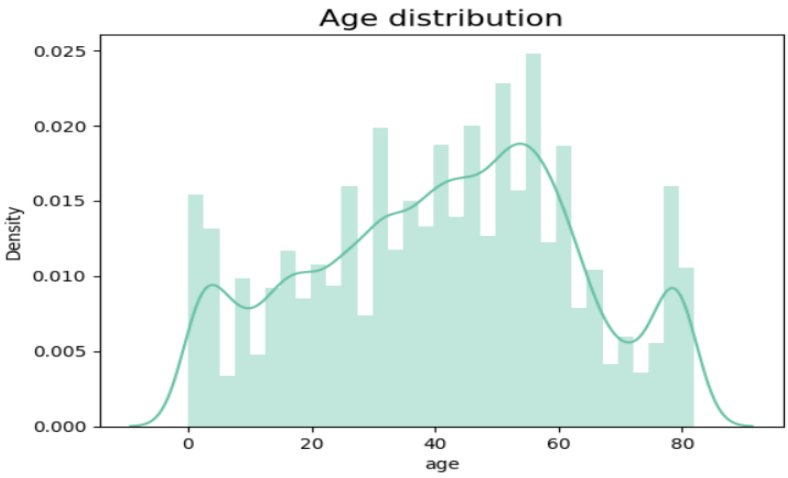
➢ Như chúng ta có thể thấy từ bảng và biểu đồ bên trên, nếu xét về giới tính. tỉ lệ 96% nữ và 95% nam không mắc bệnh đột quỵ, một tỷ lệ gần như là ngang nhau.

3.2.2. age

- Số tuổi nhỏ nhất torgn tập dữ liệu là 0.08 tuổi và lớn nhất là 82 tuổi:

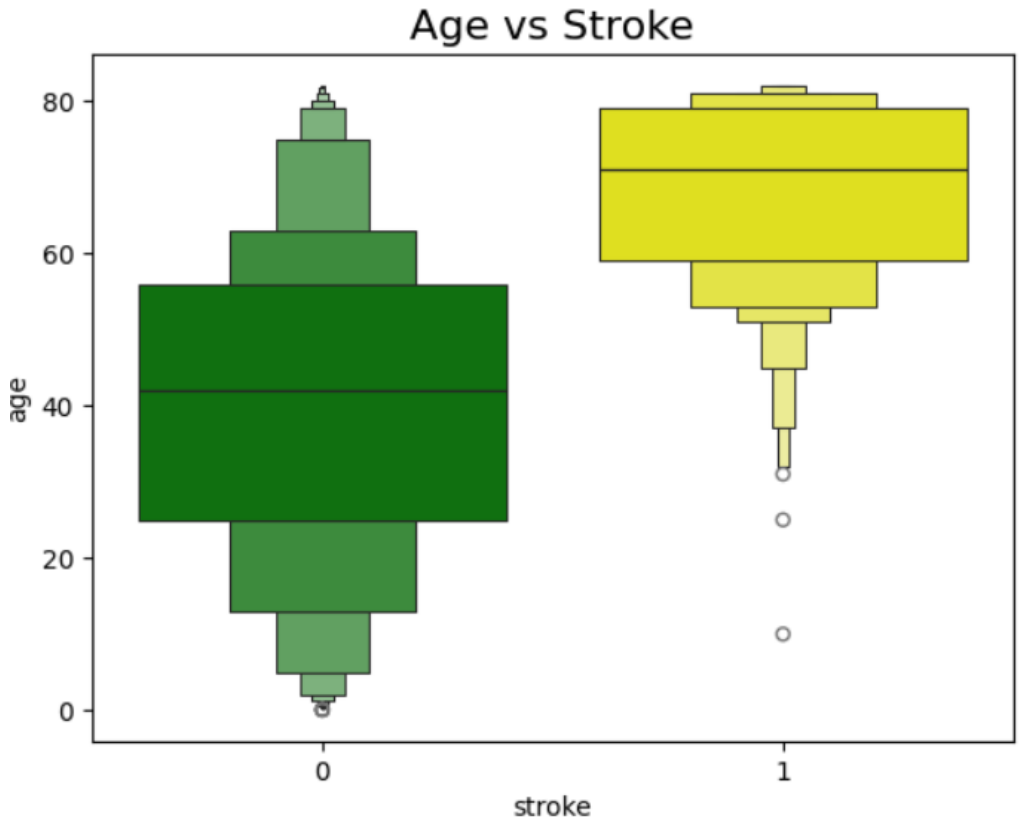
```
1 # age
2 df_train['age'].describe()
```

```
count    15304.000000
mean        41.417708
std         21.444673
min          0.080000
25%         26.000000
50%         43.000000
75%         57.000000
max         82.000000
Name: age, dtype: float64
```



Hình 6: Biểu đồ phân phối độ tuổi

- Độ tuổi trong tập dữ liệu phân phối chủ yếu từ 40 đến 60 tuổi

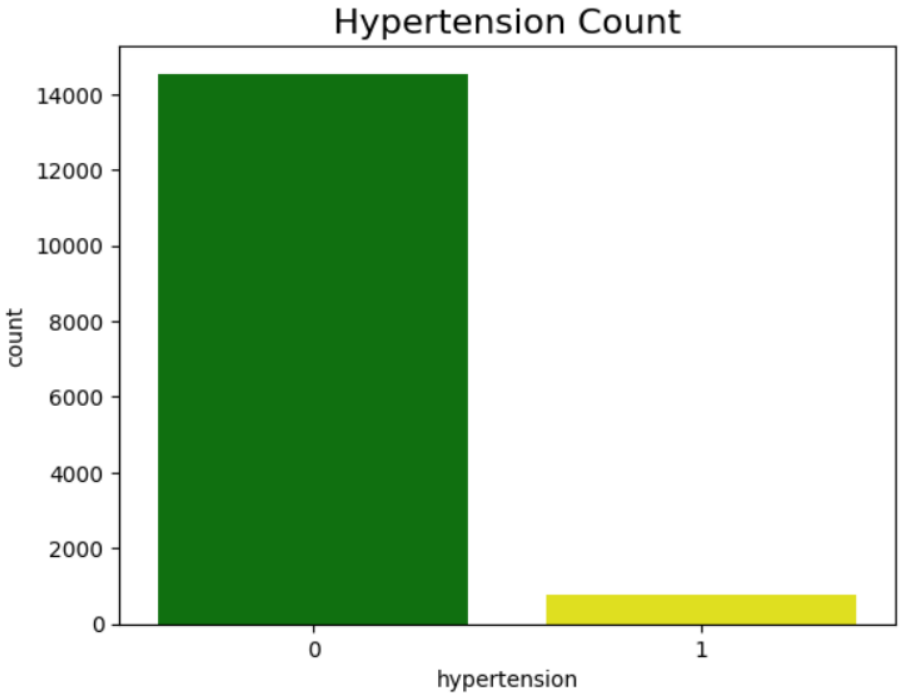


Hình 7: Biểu đồ thống kê mối quan hệ của độ tuổi và khả năng đột quỵ

➢ Chúng ta có thể thấy rõ rằng hầu hết các điểm dữ liệu bị đột quỵ đều có số tuổi trên 60 trong khi hầu hết các điểm dữ liệu không bị đột quỵ đều dưới 60. Vậy người lớn tuổi (trên 60) có nguy cơ mắc bệnh đột quỵ cao hơn so với người dưới 60 tuổi.

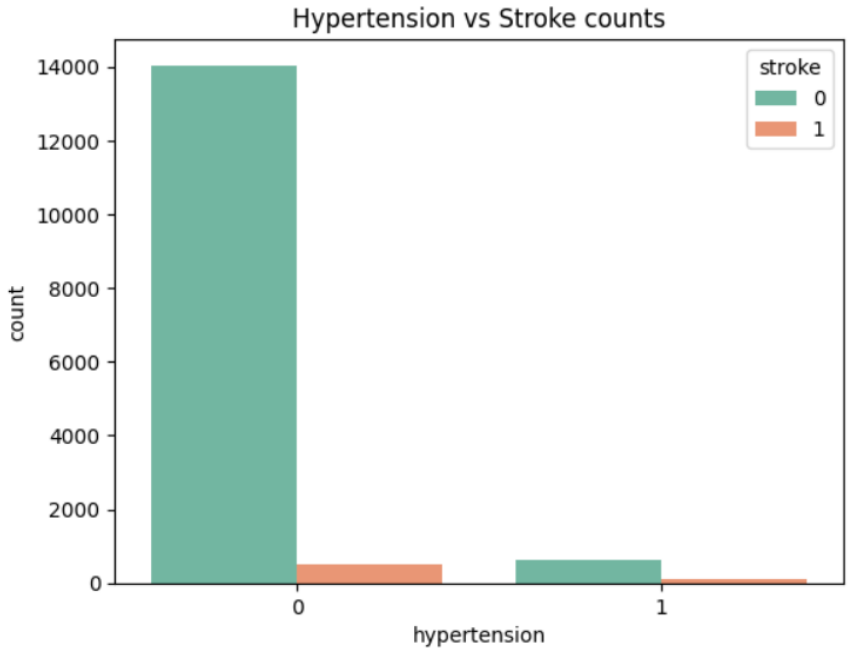
3.2.3. hypertension

- Chỉ có 2 giá trị cho thuộc tính hypertension(tăng huyết áp) đó là 0 và 1, 0 cho đối tượng không tăng huyết áp và 1 cho đối tượng bị tăng huyết áp.



Hình 8: Biểu đồ thống kê đặc trưng tăng huyết áp

- Số lượng dữ liệu không bị tăng huyết áp (14543) cao hơn nhiều so với lượng dữ liệu bị tăng huyết áp (761).



Hình 9: Biểu đồ thống kê mối quan hệ giữa tăng huyết áp và đột quỵ

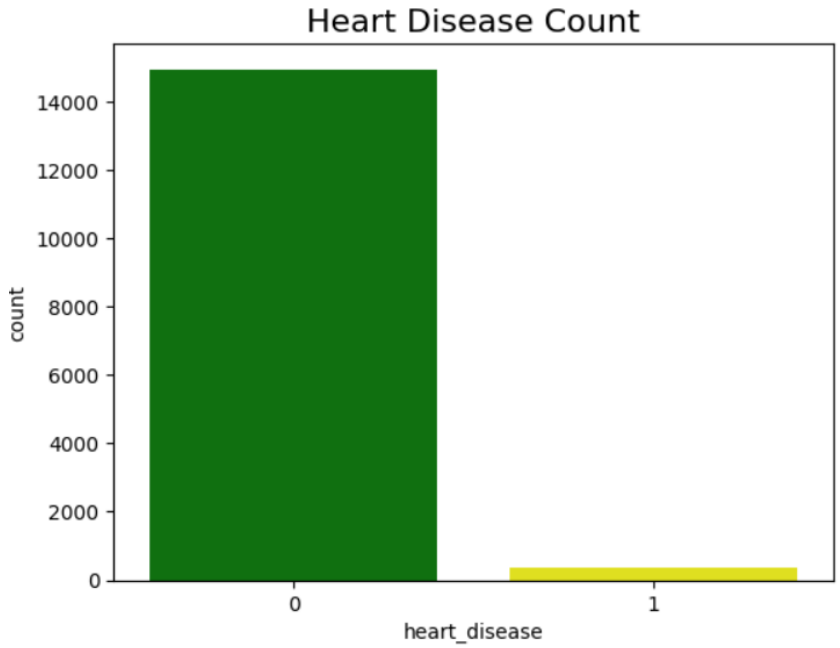
- Bảng thống kê tỉ lệ đột quỵ theo tình trạng tăng huyết áp:

stroke	0	1
hypertension		
0	0.965344	0.034656
1	0.831800	0.168200
All	0.958704	0.041296

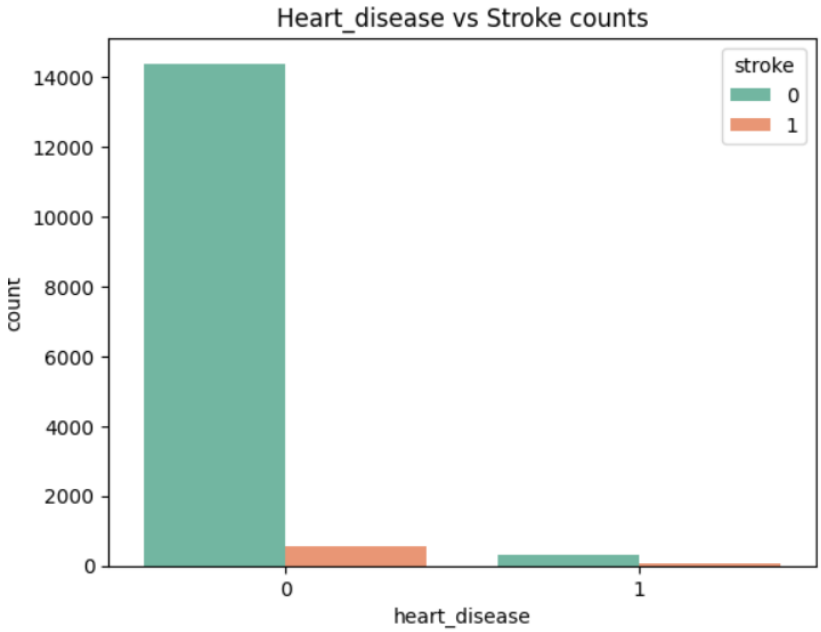
➢ Những người đang bị tăng huyết áp (hypertension) có nguy cơ bị đột quỵ cao hơn đáng kể so với những người không bị tăng huyết áp (hypertension).

3.2.4. heart_disease

- Có 2 giá trị cho thuộc tính heart_disease là 0 và 1, giá trị 0 cho biết bệnh nhân không mắc bệnh tim, giá trị 1 cho biết bệnh nhân đã mắc bệnh tim.



Hình 10: Biểu đồ thống kê đặc trưng bệnh tim



Hình 11: Biểu đồ thống kê mối quan hệ bệnh tim và đột quỵ

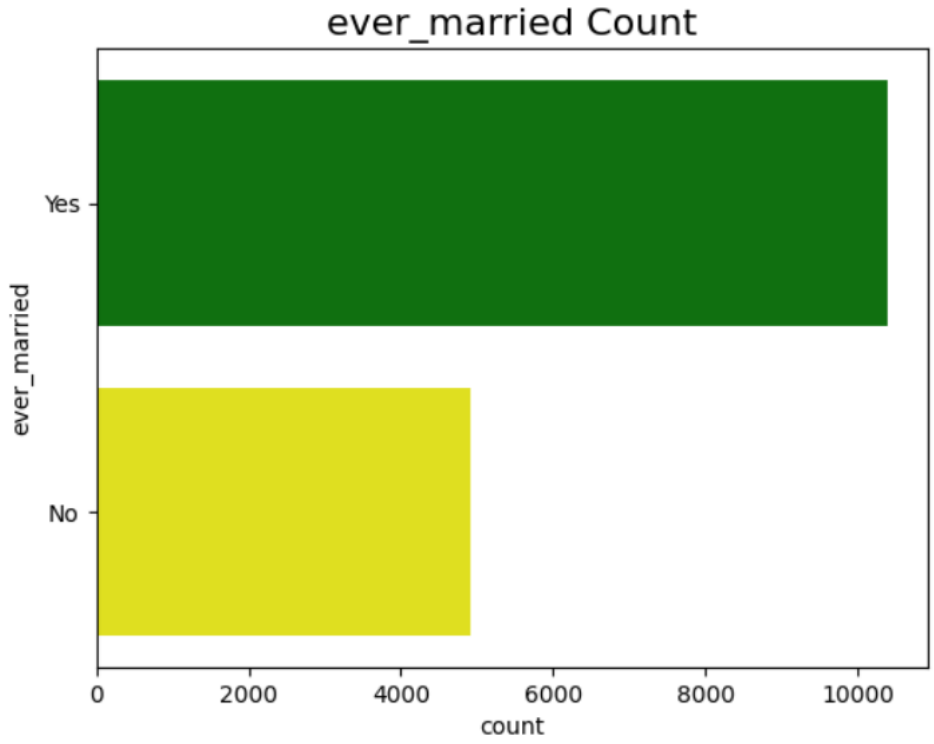
- Bảng thống kê tỉ lệ đột quỵ đối với tình trạng bệnh tim:

stroke	0	1
heart_disease		
0	0.962133	0.037867
1	0.815126	0.184874
All	0.958704	0.041296

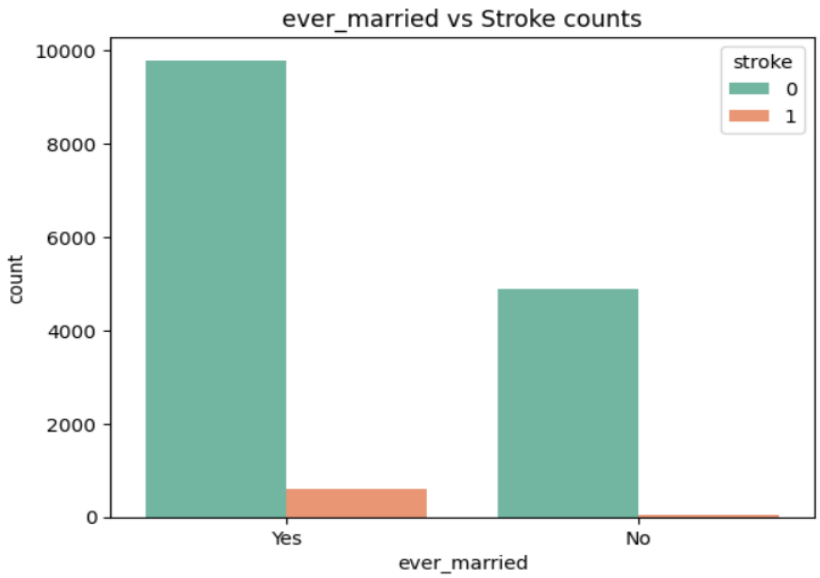
➢ Những người đang mắc bệnh tim (heart_disease) có nguy cơ bị đột quỵ cao đáng kể hơn những người không mắc bệnh tim (heart_disease).

3.2.5. ever_married

- Có 2 giá trị của thuộc tính ever_married là Yes và No. Yes là đã từng kết hôn và No là ngược lại.



Hình 12: Biểu đồ thống kê tình trạng hôn nhân



Hình 13: Biểu đồ thống kê mối quan hệ tình trạng hôn nhân và đột quỵ

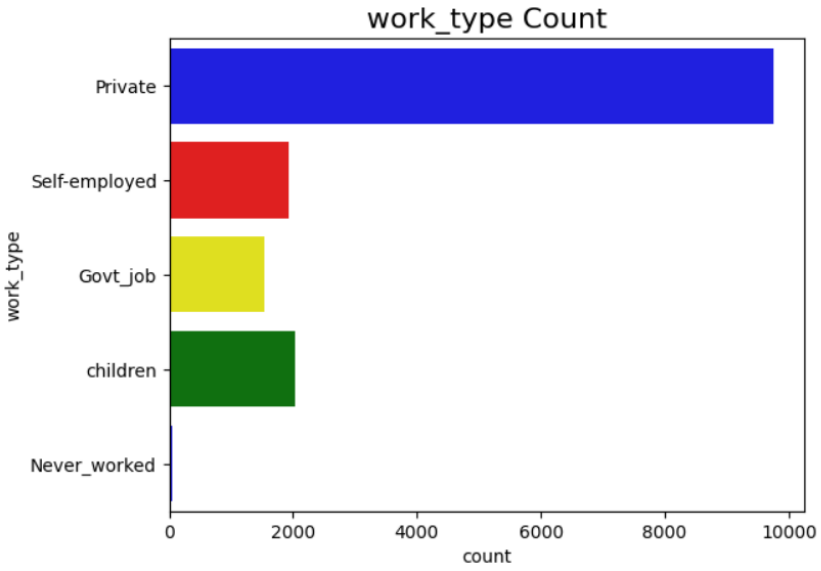
- Bảng thống kê tỉ lệ đột quỵ đối với tình trạng kết hôn:

stroke	0	1
ever_married		
No	0.991665	0.008335
Yes	0.943091	0.056909
All	0.958704	0.041296

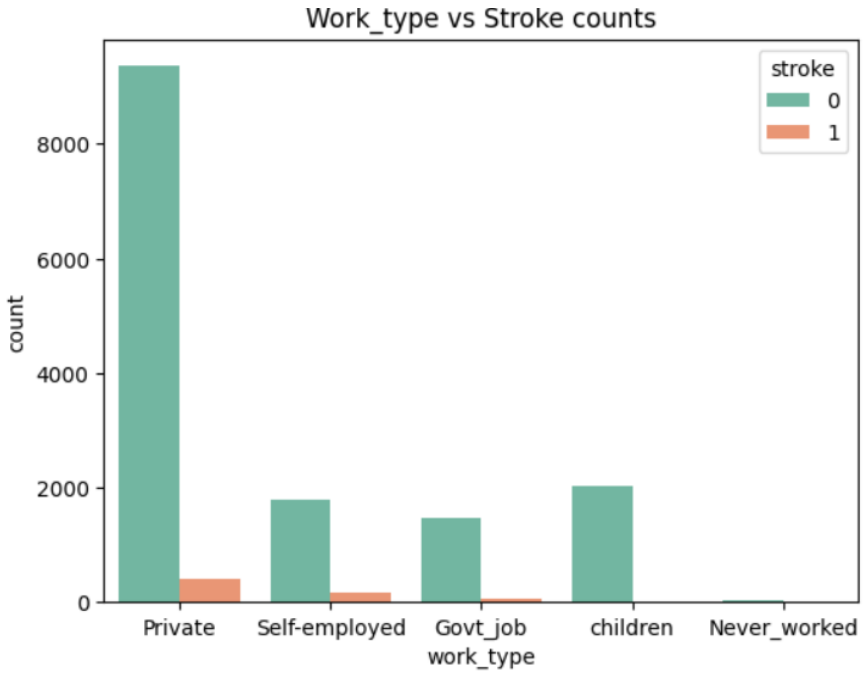
➢ Những người đã từng kết hôn có nguy cơ bị đột quỵ cao hơn những người chưa từng kết hôn.

3.2.6. work_type

- Thuộc tính work_type cho biết loại công việc mà bệnh nhân đang làm, "children", "Govt_jov", "Never_worked", "Private" hoặc "Self-employed".



Hình 14: Biểu đồ thống kê việc làm của bệnh nhân



Hình 15: Biểu đồ thống kê mối quan hệ giữa việc làm và đột quỵ

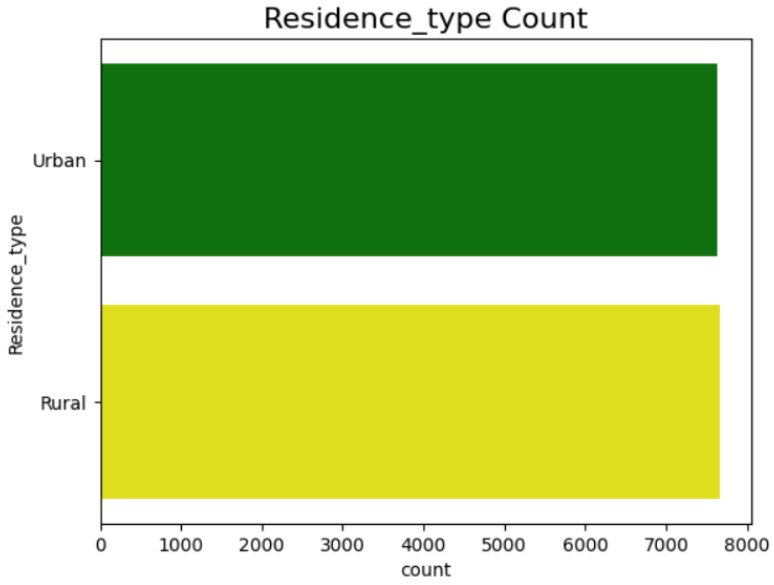
- Bảng thống kê tỉ lệ đột quỵ đối với việc làm của bệnh nhân:

stroke	0	1
work_type		
Govt_job	0.954990	0.045010
Never_worked	1.000000	0.000000
Private	0.958573	0.041427
Self-employed	0.918515	0.081485
children	0.999509	0.000491
All	0.958704	0.041296

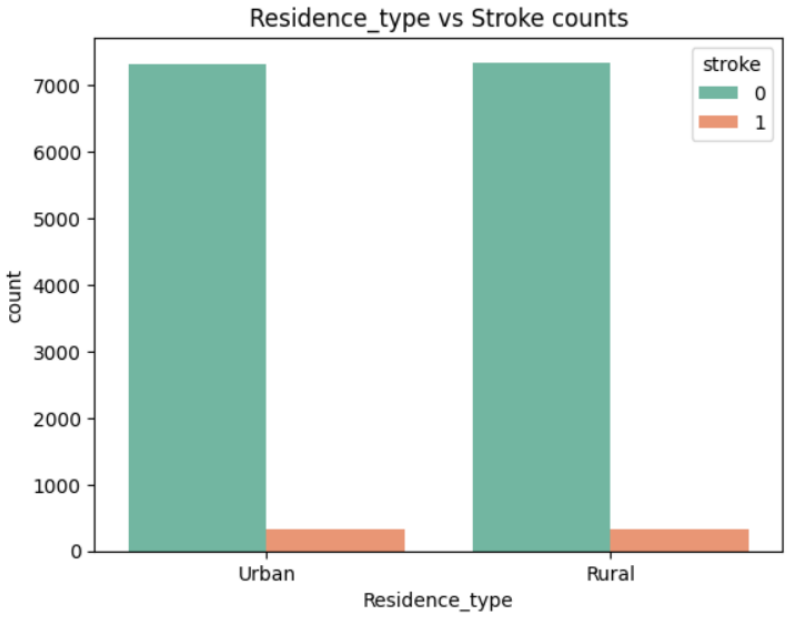
- Những người chưa bao giờ làm việc có nguy cơ bị đột quỵ thấp nhất, trong khi đó những người tự kinh doanh có nguy cơ bị đột quỵ cao nhất.

3.2.7. Residence_type

- Có 2 giá trị trong thuộc tính Residence_type, Urban và Rural tức là thành thị và nông thôn.



Hình 16: Biểu đồ thống kê loại hình cư trú



Hình 17: Biểu đồ thống kê mối quan hệ giữa nơi cư trú và đột quỵ

- Bảng thống kê tỉ lệ đột quỵ đối với nơi cư trú:

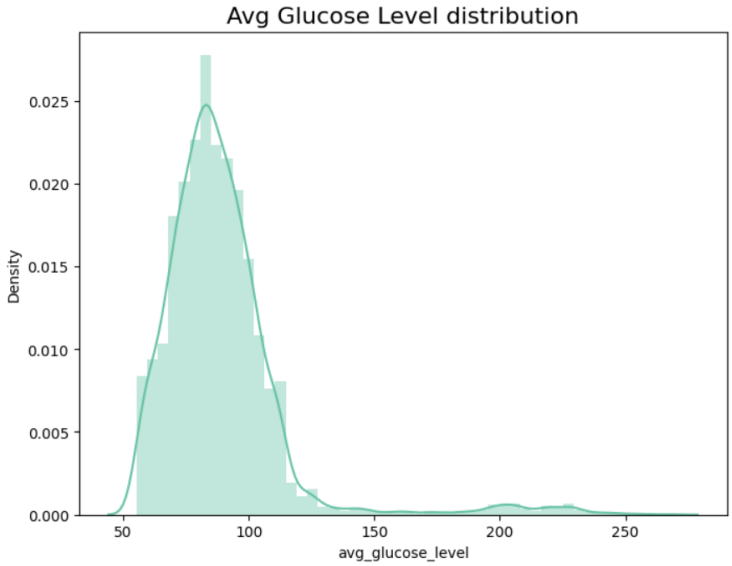
stroke	0	1
Residence_type		
Rural	0.958638	0.041362
Urban	0.958770	0.041230
All	0.958704	0.041296

➢ Tỷ lệ bị đột quỵ gần như là bằng nhau dù bệnh nhân ở nông thôn hay thành thị. Có thể xem xét loại bỏ thuộc tính này, vì tác động của các giá trị của thuộc tính đối với biến mục tiêu là như nhau.

3.2.8. avg_glucose_level

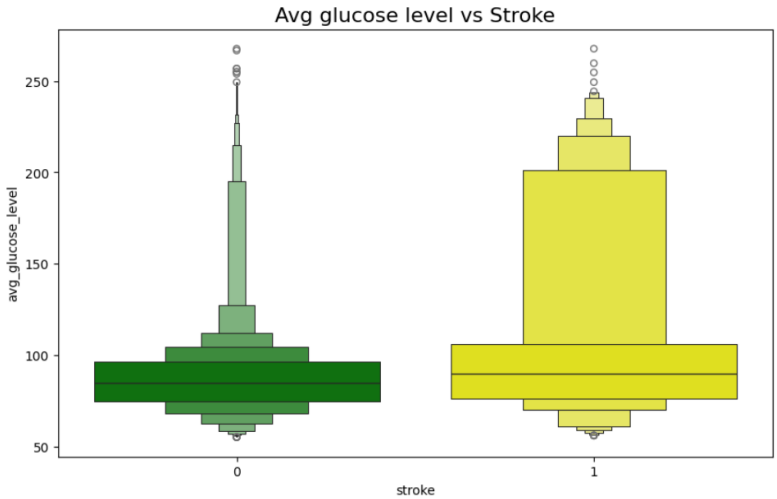
- Lượng avg_glucose_level trong tập dữ liệu nhỏ nhất là 55.22 và lớn nhất là 267.6:

```
count    15304.000000
mean        89.039853
std         25.476102
min         55.220000
25%         74.900000
50%         85.120000
75%         96.980000
max        267.600000
Name: avg_glucose_level, dtype: float64
```



Hình 18: Biểu đồ phân phối avg_glucose_level

- avg_glucose_level không được phân bổ đồng đều; phần lớn các điểm dữ liệu có trung bình là 60-120, nhưng rất ít trong số đó có mức 120-250.



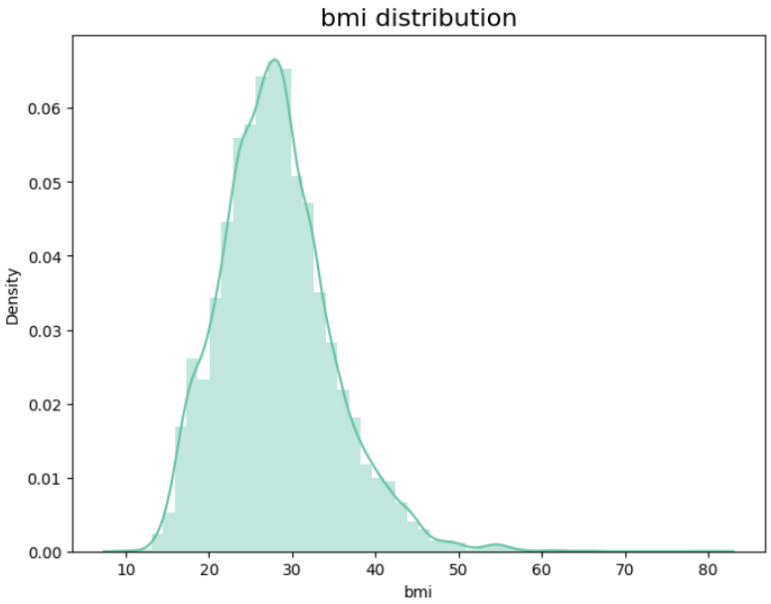
Hình 19: Biểu đồ thống kê mối quan hệ của lượng glucose và khả năng đột quỵ

➢ Lượng glucose trung bình càng cao thì nguy cơ bị đột quỵ càng cao.

3.2.9. bmi

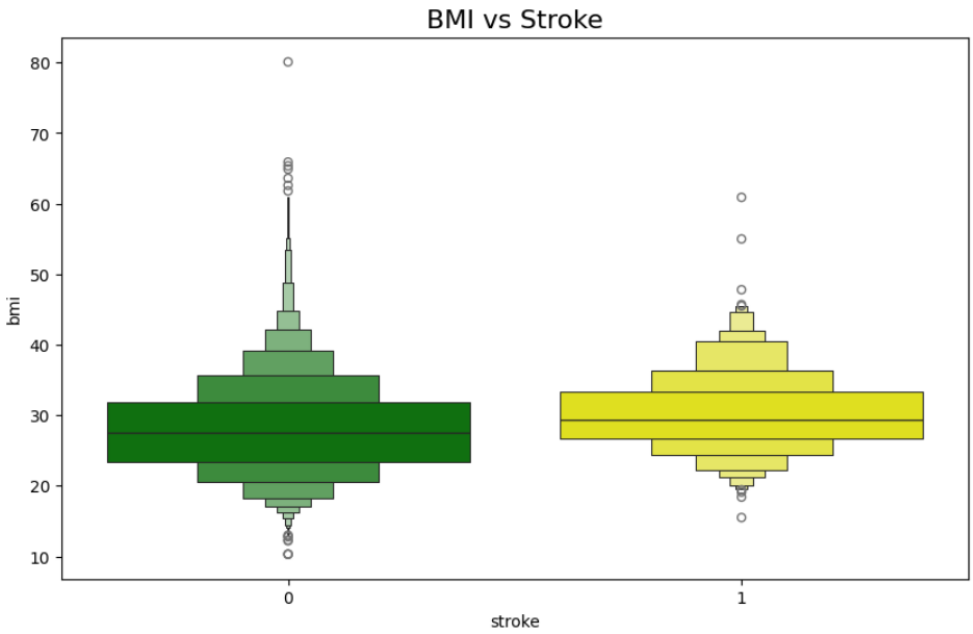
- Lượng bmi trong tập dữ liệu nhỏ nhất là 10.3 và lớn nhất là 80.1:

```
count    15304.000000
mean        28.112721
std          6.722315
min         10.300000
25%         23.500000
50%         27.600000
75%         32.000000
max         80.100000
Name: bmi, dtype: float64
```



Hình 20: Biểu đồ phân phối bmi

- Biểu đồ trực quan hóa thuộc tính bmi không được phân bổ đồng đều; phần lớn các quan sát có bmi là 25-35, phần ít trong số đó có chỉ số 50-80.

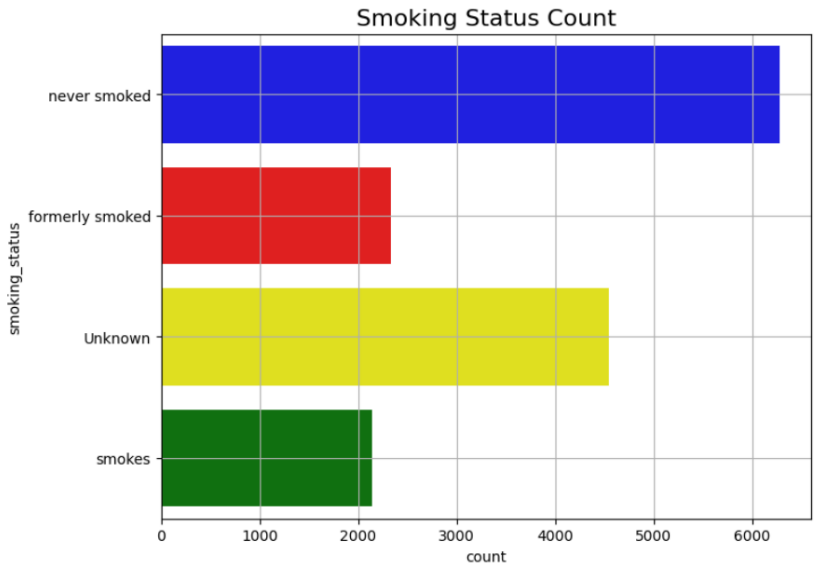


Hình 21: Biểu đồ thống kê mối quan hệ của lượng bmi và khả năng đột quỵ

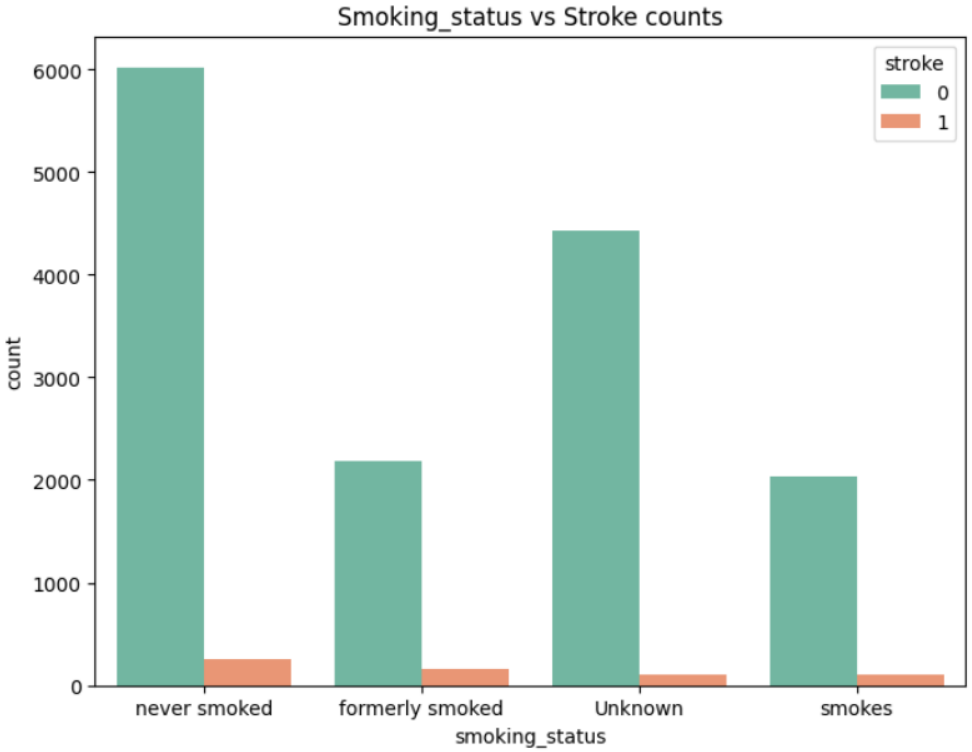
- Có thể thấy rằng chỉ số BMI của các bệnh nhân bị đột quỵ và không bị đột quỵ có phân bố tương đối tương đồng, tập trung ở khoảng 25 đến 35. Tuy nhiên, nhóm không bị đột quỵ có phân bố rộng hơn với nhiều giá trị ngoại lệ. Điều này gợi ý rằng BMI có thể không phải là yếu tố duy nhất hoặc quan trọng nhất liên quan đến nguy cơ đột quỵ, và cần thêm phân tích để xác định các yếu tố khác có liên quan.

3.2.10. smoking_status

- Có 4 giá trị thuộc đặc trưng smoking_status đó là 'never smoked', 'formerly smoked', 'Unknown' and 'smokes'.



Hình 22: Biểu đồ thống kê tình trạng hút thuốc của bệnh nhân



Hình 23: Biểu đồ thống kê mối quan hệ giữa tình trạng hút thuốc và đột quỵ

- Bảng thống kê tỉ lệ đột quỵ đối với việc hút thuốc:

stroke	0	1
smoking_status		
Unknown	0.976227	0.023773
formerly smoked	0.931964	0.068036
never smoked	0.959083	0.040917
smokes	0.949603	0.050397
All	0.958704	0.041296

➢ ‘Unknown’ & ‘never smoked’ có ít khả năng bị đột quỵ hơn, trong khi đó *formerly smoked* và *smokes* có khả năng cao hơn sẽ bị đột quỵ. Tóm lại, những người đã và đang hút thuốc có nguy cơ đột quỵ cao hơn những người chưa từng hút thuốc.

CHƯƠNG 4 – PHÂN TÍCH CÁC THUẬT TOÁN NHỊ PHÂN CÓ THỂ ÁP DỤNG VÀO BÀI TOÁN

4.1. Logistic Regression

Logistic Regression là một thuật toán phân loại nhị phân phổ biến, được sử dụng để dự đoán khả năng xảy ra của một sự kiện bằng cách mô hình hóa mối quan hệ giữa các đặc trưng đầu vào và một kết quả nhị phân.

Cách Hoạt Động của Hồi quy Logistic

1) Xác Suất Dự Đoán: Hồi quy logistic dự đoán xác suất một mẫu thuộc về lớp dương tính (ví dụ: có đột quỵ) dựa trên các đặc trưng đầu vào.
2) Hàm Sigmoid: Sử dụng hàm sigmoid để chuyển đổi đầu ra của mô hình hồi quy tuyến tính thành một giá trị xác suất nằm trong khoảng từ 0 đến 1. Hàm sigmoid có dạng chữ S, đảm bảo rằng đầu ra luôn nằm trong khoảng từ 0 đến 1.
3) Huấn Luyện Mô Hình: Mô hình học các hệ số cho các đặc trưng bằng cách tối ưu hóa để giảm thiểu sai số giữa dự đoán và nhãn thực tế, thông qua các thuật toán tối ưu hóa như Gradient Descent.
4) Dự Đoán:

- Sau khi huấn luyện, mô hình sẽ sử dụng các hệ số đã học để tính toán xác suất cho các mẫu mới.
- Mẫu mới được phân loại vào lớp dương tính nếu xác suất dự đoán lớn hơn một ngưỡng (thường là 0.5), ngược lại sẽ phân loại vào lớp âm tính.

Ưu Điểm của Hồi quy Logistic

- Đơn Giản và Dễ Hiểu: Dễ triển khai và trực quan, kết quả có thể giải thích một cách rõ ràng.
- Hiệu Suất Tính Toán: Tính toán hiệu quả và không yêu cầu nhiều tài nguyên.

- Khả Năng Giải Thích: Cung cấp các xác suất dự đoán và hệ số của mô hình có thể giải thích được.

Nhược Điểm của Hồi quy Logistic

- Giới Hạn trong Phân Loại Nhị Phân: Chủ yếu chỉ áp dụng cho các bài toán phân loại nhị phân. Đối với phân loại đa lớp cần mở rộng (multinomial logistic regression).
- Giả Định Tuyến Tính: Giả định rằng các đặc trưng có mối quan hệ tuyến tính với log-odds của kết quả.
- Nhạy Cảm với Ngoại Lai và Dữ Liệu Thiếu: Có thể bị ảnh hưởng bởi các điểm dữ liệu ngoại lai và thiếu dữ liệu.

4.2. k-Nearest Neighbors

K-Nearest Neighbors (K-NN) là một thuật toán học máy đơn giản và dễ hiểu được sử dụng cho cả bài toán phân loại và hồi quy. Nó hoạt động dựa trên nguyên tắc tìm kiếm các điểm dữ liệu gần nhất trong không gian đặc trưng để dự đoán nhãn hoặc giá trị của điểm dữ liệu mới.

Cách Hoạt Động của K-NN:

1) Xác định Số Lượng Hàng Xóm (K): Chọn một giá trị K, tức là số lượng hàng xóm gần nhất sẽ được xem xét để đưa ra dự đoán.
2) Tính Khoảng Cách: Tính khoảng cách giữa điểm dữ liệu cần dự đoán và tất cả các điểm dữ liệu trong tập huấn luyện. Khoảng cách thường được tính bằng Euclidean, nhưng cũng có thể sử dụng các khoảng cách khác như Manhattan, Minkowski, v.v.
3) Tìm K Hàng Xóm Gần Nhất: Chọn K điểm dữ liệu trong tập huấn luyện có khoảng cách gần nhất với điểm dữ liệu cần dự đoán.
4) Dự Đoán:

- Phân loại: Dự đoán nhãn của điểm dữ liệu mới bằng cách lấy nhãn phổ biến nhất trong số K hàng xóm (voting).

- Hồi quy: Dự đoán giá trị của điểm dữ liệu mới bằng cách lấy trung bình các giá trị của K hàng xóm.

Ưu Điểm của K-NN

- Đơn Giản và Dễ Hiểu: K-NN dễ triển khai và trực quan.
- Linh Hoạt: Có thể sử dụng cho cả bài toán phân loại và hồi quy.
- Không Cần Huấn Luyện: K-NN không cần giai đoạn huấn luyện mô hình, toàn bộ dữ liệu huấn luyện được sử dụng mỗi khi dự đoán.

Nhược Điểm của K-NN

- Tốc Độ Chậm với Dữ Liệu Lớn: Khi số lượng mẫu lớn, việc tính khoảng cách và tìm K hàng xóm gần nhất có thể rất chậm.
- Nhạy Cảm với Dữ Liệu Nhiễu và Ngoại Lai: Các điểm dữ liệu nhiễu hoặc ngoại lai có thể ảnh hưởng lớn đến dự đoán.
- Yêu Cầu Bộ Nhớ Lớn: Cần lưu trữ toàn bộ dữ liệu huấn luyện, có thể tiêu tốn nhiều bộ nhớ.
- Chọn K Tối Ưu: Việc chọn giá trị K tối ưu không phải lúc nào cũng rõ ràng và có thể cần thử nghiệm với các giá trị khác nhau.

Lựa chọn giá trị k phù hợp: Giá trị của k ảnh hưởng đáng kể đến hiệu suất của kNN. Giá trị k nhỏ có thể dẫn đến quá khớp (overfitting), trong khi giá trị k lớn có thể dẫn đến thiếu khớp (underfitting). Không có giải pháp chung, và giá trị k tối ưu thường cần được xác định thông qua thử nghiệm.

4.3. Support Vector Machine (SVM)

Support Vector Machine (SVM) là một thuật toán học máy mạnh mẽ được sử dụng cho cả bài toán phân loại và hồi quy. SVM đặc biệt hiệu quả trong các bài toán phân loại nhị phân, tìm ra một siêu phẳng tối ưu để phân tách các lớp dữ liệu khác nhau trong không gian đặc trưng.

Cách Hoạt Động của SVM

1) Siêu Phẳng Phân Tách (Hyperplane):

- SVM tìm kiếm một siêu phẳng (hyperplane) trong không gian nhiều chiều để phân tách các điểm dữ liệu của các lớp khác nhau một cách rõ ràng.
- Siêu phẳng tối ưu là siêu phẳng có khoảng cách lớn nhất đến các điểm dữ liệu gần nhất của cả hai lớp. Khoảng cách này được gọi là biên (margin).

2) Các Điểm Hỗ Trợ (Support Vectors):

- Các điểm dữ liệu gần nhất đến siêu phẳng được gọi là các vector hỗ trợ. Các điểm này xác định vị trí của siêu phẳng và biên.
- Chỉ các điểm hỗ trợ ảnh hưởng đến vị trí của siêu phẳng, các điểm dữ liệu khác không ảnh hưởng.

3) Biên Tối Đa (Maximum Margin): Mục tiêu của SVM là tối đa hóa biên giữa siêu phẳng và các vector hỗ trợ. Điều này giúp đảm bảo mô hình có khả năng tổng quát tốt hơn trên dữ liệu mới.

4) Hàm Kernel: Đối với dữ liệu không thể phân tách tuyến tính, SVM sử dụng các hàm kernel để ánh xạ dữ liệu sang một không gian đặc trưng cao hơn, nơi dữ liệu có thể phân tách tuyến tính. Các kernel phổ biến bao gồm:

- Linear Kernel: Dùng cho dữ liệu phân tách tuyến tính.
- Polynomial Kernel: Dùng cho dữ liệu phân tách phi tuyến.
- RBF (Radial Basis Function) Kernel: Dùng cho dữ liệu phức tạp và phi tuyến.

Ưu Điểm của SVM

- Hiệu Suất Cao: SVM thường đạt hiệu suất cao trong các bài toán phân loại với dữ liệu nhỏ và phức tạp.
- Khả Năng Tổng Quát Tốt: SVM tối đa hóa biên, giúp mô hình có khả năng tổng quát tốt hơn trên dữ liệu mới.

- Linh Hoạt với Các Kernel: Khả năng sử dụng các hàm kernel giúp SVM linh hoạt trong việc xử lý dữ liệu phi tuyến.

Nhược Điểm của SVM

- Tốc Độ Chậm với Dữ Liệu Lớn: SVM có thể chậm và tốn kém về mặt tính toán với dữ liệu lớn.
- Khó Chọn Kernel và Tham Số: Việc chọn hàm kernel phù hợp và điều chỉnh các siêu tham số (như C, gamma) có thể phức tạp.
- Không Hoạt Động Tốt với Dữ Liệu Nhiễu: SVM nhạy cảm với dữ liệu nhiễu và có thể bị ảnh hưởng bởi các điểm ngoại lai.

4.4. Decision Tree

Decision Tree là một thuật toán học máy sử dụng cấu trúc cây để đưa ra quyết định dựa trên các đặc trưng của dữ liệu. Đây là một phương pháp phân loại và hồi quy phổ biến, dễ hiểu và trực quan.

Cách Hoạt Động của Cây Quyết Định

1) Cấu trúc Cây:

- Cây quyết định bao gồm các nút (nodes) và các nhánh (branches).
- Nút gốc (Root Node): Nút đầu tiên đại diện cho toàn bộ dữ liệu.
- Nút bên trong (Internal Nodes): Mỗi nút đại diện cho một đặc trưng và một điều kiện phân chia dựa trên đặc trưng đó.
- Nút lá (Leaf Nodes): Nút cuối cùng đại diện cho nhãn (trong phân loại) hoặc giá trị (trong hồi quy) dự đoán.

2) Quá trình Phân chia:

Chọn Đặc trưng Phân chia: Chọn đặc trưng và ngưỡng phân chia sao cho làm giảm thiểu độ hỗn loạn (impurity) của các nút con. Các chỉ số phổ biến để đo độ hỗn loạn bao gồm:

- Gini Impurity: Đo lường độ hỗn loạn của phân bố xác suất tại nút.
- Entropy: Sử dụng trong Information Gain để chọn phân chia tối ưu.

- Mean Squared Error (MSE): Dùng trong cây hồi quy để đo lường độ hỗn loạn.

Phân chia Dữ liệu: Dữ liệu được chia nhỏ dựa trên đặc trưng và ngưỡng phân chia đã chọn, tạo ra các nhánh con.

3) Điều kiện Dừng: Quá trình phân chia tiếp tục cho đến khi đạt một trong các điều kiện dừng:

- Tất cả dữ liệu trong một nút thuộc về một lớp (trong phân loại) hoặc giá trị (trong hồi quy).
- Không còn đặc trưng nào để phân chia.
- Đạt đến độ sâu tối đa của cây (max depth).
- Số lượng mẫu trong một nút nhỏ hơn một ngưỡng (min samples split).

4) Dự đoán: Để dự đoán nhãn hoặc giá trị cho một mẫu mới, bắt đầu từ nút gốc và đi qua các nhánh của cây dựa trên giá trị của các đặc trưng cho đến khi đến một nút lá, nơi chứa nhãn hoặc giá trị dự đoán.

Ưu Điểm của Cây Quyết Định:

- Dễ Hiểu và Giải Thích: Kết quả dễ hiểu và trực quan, có thể biểu diễn dưới dạng sơ đồ cây.
- Không Yêu Cầu Tiền Xử Lý Nhiều: Ít yêu cầu chuẩn hóa dữ liệu và có thể xử lý cả dữ liệu số và danh mục.
- Xử Lý Tốt Các Quan Hệ Phi Tuyến: Có khả năng nắm bắt và xử lý các quan hệ phi tuyến trong dữ liệu.
- Khả năng Xử Lý Dữ Liệu Thiếu: Có thể xử lý các trường hợp dữ liệu thiếu mà không cần loại bỏ các mẫu.

Nhược Điểm của Cây Quyết Định:

- Dễ Bị Overfitting: Cây quyết định có thể học thuộc dữ liệu huấn luyện và kém hiệu quả trên dữ liệu mới. Cần áp dụng các kỹ thuật như pruning (cắt tỉa cây) để giảm thiểu overfitting.

- Nhạy Cảm với Dữ Liệu Nhiễu: Các thay đổi nhỏ trong dữ liệu có thể dẫn đến thay đổi lớn trong cấu trúc cây.
- Không Hiệu Quả với Dữ Liệu Rộng và Sâu: Với các tập dữ liệu lớn và phức tạp, cây quyết định có thể trở nên quá lớn và khó xử lý.

4.5. Random Forest

Rừng Ngẫu Nhiên (Random Forest) là một thuật toán học máy mạnh mẽ và linh hoạt, kết hợp nhiều cây quyết định để tạo ra một mô hình mạnh mẽ và chính xác hơn. Đây là một kỹ thuật ensemble learning, giúp cải thiện độ chính xác và giảm thiểu overfitting bằng cách sử dụng đa dạng các cây quyết định.

Cách Hoạt Động của Rừng Ngẫu Nhiên

1) Tạo Các Cây Quyết Định:

- Bagging (Bootstrap Aggregating): Tạo nhiều tập con từ tập dữ liệu gốc bằng cách lấy mẫu ngẫu nhiên có hoàn lại. Mỗi tập con này sẽ được sử dụng để huấn luyện một cây quyết định.
- Random Feature Selection: Tại mỗi nút phân chia, một tập con ngẫu nhiên của các đặc trưng được chọn. Cây quyết định sẽ chọn đặc trưng tốt nhất từ tập con này để phân chia. Điều này giúp các cây trong rừng đa dạng và không quá giống nhau.

2) Huấn Luyện Các Cây Quyết Định:

- Mỗi cây quyết định được huấn luyện trên một tập con khác nhau của dữ liệu và sử dụng tập con ngẫu nhiên của các đặc trưng tại mỗi nút phân chia.
- Các cây quyết định này có thể phát triển độc lập và song song.

3) Dự Đoán:

- Đối với bài toán phân loại: Mỗi cây quyết định đưa ra một dự đoán (nhãn) và rừng ngẫu nhiên đưa ra dự đoán cuối cùng bằng cách lấy nhãn phổ biến nhất (voting) từ các cây.

- Đối với bài toán hồi quy: Mỗi cây quyết định đưa ra một giá trị dự đoán và rừng ngẫu nhiên đưa ra dự đoán cuối cùng bằng cách lấy trung bình các giá trị dự đoán từ các cây.

Ưu Điểm của Rừng Ngẫu Nhiên

- Độ Chính Xác Cao: Kết hợp nhiều cây quyết định giúp cải thiện độ chính xác của mô hình so với việc sử dụng một cây duy nhất.
- Giảm Overfitting: Sự đa dạng của các cây quyết định giúp giảm thiểu nguy cơ overfitting và cải thiện khả năng tổng quát hóa.
- Khả Năng Xử Lý Tính Đa Dạng của Dữ Liệu: Random Forest có thể xử lý tốt các dữ liệu phức tạp và không tuyến tính.
- Kháng Nhiễu: Random Forest ít bị ảnh hưởng bởi các điểm dữ liệu nhiễu hoặc ngoại lai.

Nhược Điểm của Rừng Ngẫu Nhiên

- Tốc Độ và Tài Nguyên: Việc huấn luyện và dự đoán với nhiều cây quyết định có thể tốn kém về thời gian và tài nguyên, đặc biệt với các tập dữ liệu lớn.
- Giải Thích Khó Hơn: So với một cây quyết định đơn lẻ, kết quả của rừng ngẫu nhiên khó giải thích hơn vì nó là tổng hợp của nhiều cây.
- Bộ Nhớ: Random Forest cần nhiều bộ nhớ để lưu trữ tất cả các cây quyết định.

4.6.Naïve Bayes

Naive Bayes là một nhóm các thuật toán phân loại dựa trên định lý Bayes với giả định đơn giản rằng các đặc trưng (features) là độc lập với nhau, tức là sự xuất hiện của một đặc trưng không ảnh hưởng đến sự xuất hiện của đặc trưng khác. Mặc dù giả định này thường không đúng trong thực tế, nhưng Naive Bayes vẫn hoạt động tốt trong nhiều tình huống thực tế.

Cách hoạt động của Naïve Bayes:

1) Tính Xác Suất: Thuật toán tính xác suất của một mẫu dữ liệu thuộc về từng lớp trong bộ dữ liệu. Điều này được thực hiện dựa trên xác suất trước ($P(C)$) của từng lớp và xác suất của các đặc trưng ($P(X|C)$ trong mẫu dữ liệu đó thuộc về từng lớp.
2) Áp Dụng Giả Định Độc Lập: Naive Bayes giả định rằng các đặc trưng trong mẫu dữ liệu là độc lập với nhau. Mặc dù giả định này thường không đúng trong thực tế, nhưng nó giúp đơn giản hóa việc tính toán xác suất bằng cách giả sử không có sự tương tác giữa các đặc trưng.
3) Dự Đoán Lớp: Cuối cùng, Naive Bayes chọn lớp có xác suất cao nhất cho mẫu dữ liệu đó. Nó so sánh xác suất của mỗi lớp dựa trên các đặc trưng của mẫu dữ liệu và chọn lớp có xác suất cao nhất là lớp dự đoán cho mẫu đó.

Ưu Điểm của Naive Bayes

- Đơn Giản và Nhanh: Naive Bayes dễ hiểu, dễ triển khai và rất nhanh trong việc huấn luyện và dự đoán.
- Hiệu Suất Tốt với Dữ Liệu Lớn: Hiệu quả và có thể xử lý tốt với các tập dữ liệu lớn.
- Xử Lý Tốt Với Dữ Liệu Nhiễu: Không yêu cầu nhiều tài nguyên tính toán và thường có khả năng tổng quát hóa tốt trên dữ liệu nhiễu.
- Khả Năng Xử Lý Dữ Liệu Thiếu: Có khả năng xử lý các trường hợp dữ liệu bị thiếu một cách hiệu quả.

Nhược Điểm của Naive Bayes

- Giả Định Độc Lập Không Thực Tế: Giả định rằng các đặc trưng là độc lập không đúng trong hầu hết các trường hợp thực tế, có thể làm giảm độ chính xác của mô hình.
- Không Hiệu Quả với Các Tương Tác Phức Tạp: Không xử lý tốt các tương tác phức tạp giữa các đặc trưng.

- Cập Nhật Trực Tuyến Khó Khăn: Naive Bayes không dễ dàng cập nhật khi có thêm dữ liệu mới mà không huấn luyện lại từ đầu.

CHƯƠNG 5 – CHUẨN BỊ DỮ LIỆU

5.1. Xóa 2 đặc trưng Residence_type và bmi trong bộ dữ liệu

Loại bỏ 2 thuộc tính *Residence_type* và *bmi* vì chúng hầu như không có tác động đến biến mục tiêu và việc loại bỏ những thuộc tính như thế sẽ giúp cho việc huấn luyện mô hình đạt được hiệu quả cao.

```
# 1- Xóa 2 thuộc tính Residence_type và bmi trong bộ dữ liệu
df.drop(['Residence_type', 'bmi'], axis=1, inplace=True)
```

5.2. Mã hóa các categorical variable

Các thuộc tính phân loại trong tập dữ liệu bao gồm: *gender*, *ever_married*, *work_type*, *smoking_status*. Các giá trị chuỗi của các cột được chuyển đổi thành các giá trị số nguyên tương ứng. Quá trình này thường được thực hiện trước khi đưa dữ liệu vào mô hình học máy, vì hầu hết các mô hình yêu cầu dữ liệu đầu vào là dạng số.

```
# gender
df['gender'] = df['gender'].map({
    'Male': 0,
    'Female': 1,
    'Other': 2
}).astype('int')

# ever_married
df['ever_married'] = df['ever_married'].map({
    'Yes': 1,
    'No': 0
}).astype('int')

# work_type
df['work_type'] = df['work_type'].map({
    'Private':0,
    'Self-employed': 1,
    'Govt_job': 2,
    'children': 3,
    'Never_worked': 4
}).astype('int')

# smoking_status
df['smoking_status'] = df['smoking_status'].map({
    'never smoked' : 0,
    'formerly smoked': 1,
    'smokes': 2,
    'Unknown': 3
}).astype('int')
```

Sau khi mã hóa, ta nhận được kết quả sau:

```
gender                int64
age                 float64
hypertension          int64
heart_disease         int64
ever_married          int64
work_type             int64
avg_glucose_level   float64
smoking_status        int64
stroke              float64
dtype: object
```

Các cột này đã chứa giá trị số nguyên thay vì chuỗi, và dữ liệu đã sẵn sàng cho việc huấn luyện mô hình học máy.

5.3. Chuẩn hóa các numeric variables sử dụng MinMaxScaler

Min-Max Scaling là một kỹ thuật chuẩn hóa (scaling) dữ liệu phổ biến trong machine learning, được sử dụng để chuyển đổi các đặc trưng về cùng một phạm vi hoặc khoảng giá trị nhất định. Quá trình này giúp cải thiện hiệu suất của mô hình và làm cho các đặc trưng có ảnh hưởng tương đồng nhau đối với mô hình. Cụ thể, Min-Max Scaling chuyển đổi các giá trị của mỗi đặc trưng về một phạm vi nhất định, thường là từ 0 đến 1.

Các numeric variables trong tập dữ liệu là: *age, avg_glucoso_level.*

```
# 3- Chuẩn hóa các numeric variables (age và avg_glucose_level) sử dụng MinMaxScaler
from sklearn.preprocessing import MinMaxScaler
scaler = MinMaxScaler()
df[['age', 'avg_glucose_level']] = scaler.fit_transform(df[['age', 'avg_glucose_level']])
df.head(5)
```

	gender	age	hypertension	heart_disease	ever_married	work_type	avg_glucose_level	smoking_status	stroke
0	0	0.340820	0	0	1	0	0.112686	0	0.0
1	0	0.401855	0	0	1	0	0.107654	1	0.0
2	1	0.511719	0	0	1	0	0.221032	3	0.0
3	0	0.682617	0	0	1	0	0.045010	0	0.0
4	1	0.291992	0	0	0	0	0.084203	0	0.0

5.4. Chia tập dữ liệu thành hai tập dữ liệu huấn luyện và kiểm tra

5.4.1. Tập Huấn Luyện (Train Set)

Vai trò: Tập huấn luyện được sử dụng để huấn luyện mô hình machine learning bằng cách cung cấp cho mô hình một tập dữ liệu đã được gán nhãn.

Hoạt Động: Trong quá trình huấn luyện, mô hình học từ dữ liệu huấn luyện để tối ưu hóa các tham số và hàm mục tiêu (objective function), từ đó tạo ra một mô hình có khả năng dự đoán đúng nhãn của các dữ liệu mới một cách chính xác.

```
1 df_train
```

	gender	age	hypertension	heart_disease	ever_married	work_type	avg_glucose_level	smoking_status	stroke
0	0	0.340820	0	0	1	0	0.112686	0	0.0
1	0	0.401855	0	0	1	0	0.107654	1	0.0
2	1	0.511719	0	0	1	0	0.221032	3	0.0
3	0	0.682617	0	0	1	0	0.045010	0	0.0
4	1	0.291992	0	0	0	0	0.084203	0	0.0
...	...	...	...	...	...	...	...	...	...
15548	0	0.694824	0	0	1	0	0.137753	3	1.0
15549	1	0.169922	0	0	0	3	0.012972	3	1.0
15550	1	0.914551	0	0	1	1	0.109316	1	1.0
15551	0	0.865723	1	0	1	1	0.150863	3	1.0
15552	1	0.951172	0	0	1	0	0.109362	3	1.0

15553 rows × 9 columns

Hình 24: Tập huấn luyện (Train set)

5.4.2. Tập Kiểm Tra (Test Set)

Vai trò: Tập kiểm tra được sử dụng để đánh giá hiệu suất của mô hình trên dữ liệu mà nó chưa từng thấy trước đó.

Hoạt Động: Mô hình được đánh giá bằng cách dự đoán nhãn cho các mẫu dữ liệu trong tập kiểm tra, sau đó so sánh với nhãn thực tế để đo lường hiệu suất.

```
1 df_test
```

	gender	age	hypertension	heart_disease	ever_married	work_type	avg_glucose_level	smoking_status
15553	1	0.694824	0	0	1	0	0.126581	3
15554	0	0.853516	1	0	1	0	0.078201	3
15555	1	0.060059	0	0	0	3	0.224356	3
15556	1	0.682617	0	0	1	2	0.065183	2
15557	0	0.389648	0	0	1	0	0.258656	2
...	...	...	...	...	...	...	...	...
25752	1	0.328613	0	0	0	0	0.095328	0
25753	0	0.597168	0	0	1	0	0.220617	3
25754	1	0.035645	0	0	0	3	0.225833	3
25755	0	0.377441	0	0	1	0	0.125981	0
25756	1	0.023438	0	0	0	3	0.138491	3

10204 rows × 8 columns

Hình 25: Tập kiểm tra (Test set)

CHƯƠNG 6 – ĐÁNH GIÁ KẾT QUẢ CỦA CÁC MÔ HÌNH ĐƯỢC ÁP DỤNG VÀO BỘ DỮ LIỆU

6.1. Các phương pháp đánh giá kết quả

6.1.1. Accuracy

Accuracy là một trong những thước đo phổ biến nhất để đánh giá hiệu suất của một mô hình phân loại. Nó thể hiện tỷ lệ phần trăm của các dự đoán đúng trên tổng số mẫu dữ liệu.

Công thức tính Accuracy:

Accuracy = Số dự đoán đúng / Tổng số mẫu

Cách Tính Accuracy

1. Xác Định Dự Đoán Đúng:

- Dự đoán đúng là khi nhãn dự đoán (y) của mô hình khớp với nhãn thực tế (y)
- Nếu mô hình dự đoán đúng cho một mẫu dữ liệu, nó sẽ được tính vào số dự đoán đúng.

2. Tổng Số Mẫu:

- Tổng số mẫu là tổng số điểm dữ liệu mà mô hình đã được kiểm tra hoặc dự đoán.

3. Tính Toán:

- Đếm số lượng các dự đoán đúng.
- Chia số lượng dự đoán đúng cho tổng số mẫu.

6.1.2. F1-Score

F1 score là một thước đo hiệu suất của mô hình học máy, đặc biệt hữu ích trong các bài toán phân loại khi dữ liệu không cân bằng. Nó cung cấp một sự cân bằng giữa Precision và Recall, hai thước đo quan trọng khác trong đánh giá hiệu suất của mô hình.

Precision (Độ chính xác): Tỷ lệ mẫu dự đoán đúng là dương trên tổng số mẫu được dự đoán là dương.

Recall (Độ nhạy): Tỷ lệ mẫu thực sự là dương được dự đoán đúng trên tổng số mẫu thực sự là dương.

F1 score là trung bình điều hòa của Precision và Recall, được tính theo công thức:

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Ý Nghĩa của F1 Score:

- Cân Bằng Precision và Recall: F1 score cung cấp một số liệu tổng hợp cân bằng giữa Precision và Recall. Nó hữu ích khi bạn cần xem xét cả hai khía cạnh và không muốn bỏ qua một trong hai.
- Dữ Liệu Không Cân Bằng: F1 score đặc biệt hữu ích khi dữ liệu không cân bằng, nghĩa là số lượng mẫu giữa các lớp khác nhau đáng kể. Ví dụ, trong một bài toán phân loại bệnh hiếm gặp, số lượng mẫu dương (có bệnh) thường ít hơn nhiều so với số lượng mẫu âm (không bệnh).
- Xử Lý Trade-off: Khi Precision và Recall có sự mâu thuẫn (tăng cái này thì giảm cái kia), F1 score giúp đánh giá hiệu suất mô hình mà không thiên về một trong hai thước đo.

6.2. Đánh giá mô hình Logistic Regression

Kết quả đánh giá mô hình Logistic Regression:

```
Logistic Regression Accuracy: 0.9395692703310833
              precision    recall  f1-score   support

         0.0       0.94      0.99      0.97      2921
         1.0       0.53      0.09      0.15       190

    accuracy                           0.94      3111
   macro avg       0.74      0.54      0.56      3111
weighted avg       0.92      0.94      0.92      3111
```

Hình 26: Kết quả đánh giá mô hình Logistic Regression

6.3. Đánh giá mô hình k-Nearest Neighbors

Kết quả đánh giá mô hình k-Nearest Neighbors:

```
k-NN Accuracy: 0.9350691096110575
              precision    recall  f1-score   support

         0.0       0.94      0.99      0.97      2921
         1.0       0.38      0.11      0.17       190

    accuracy                           0.94      3111
   macro avg       0.66      0.55      0.57      3111
weighted avg       0.91      0.94      0.92      3111
```

Hình 27: Kết quả đánh giá mô hình k-Nearest Neighbors

6.4. Đánh giá mô hình Support Vector Machine (SVM)

Kết quả đánh giá mô hình Support Vector Machine (SVM):

```
SVM Accuracy: 0.9389263902282224
              precision    recall  f1-score   support

         0.0       0.94      1.00      0.97      2921
         1.0       0.00      0.00      0.00       190

    accuracy                           0.94      3111
   macro avg       0.47      0.50      0.48      3111
weighted avg       0.88      0.94      0.91      3111
```

Hình 28: Kết quả đánh giá mô hình Support Vector Machine (SVM)

6.5. Đánh giá mô hình Decision Trees

Kết quả đánh giá mô hình Decision Trees:

```
Decision Tree Accuracy: 0.9067823850851816
              precision    recall  f1-score   support

         0.0       0.95      0.95      0.95      2921
         1.0       0.26      0.28      0.27       190

    accuracy                           0.91      3111
   macro avg       0.61      0.61      0.61      3111
weighted avg       0.91      0.91      0.91      3111
```

Hình 29: Kết quả đánh giá mô hình Decision Trees

6.6. Đánh giá mô hình Random Forest

Kết quả đánh giá mô hình Random Forest:

```
Random Forest Accuracy: 0.9302475088396014
              precision    recall  f1-score   support

         0.0       0.95      0.98      0.96      2921
         1.0       0.33      0.14      0.19       190

    accuracy                           0.93      3111
   macro avg       0.64      0.56      0.58      3111
weighted avg       0.91      0.93      0.92      3111
```

Hình 30: Kết quả đánh giá mô hình Random Forest

6.7. Đánh giá mô hình Naïve Bayes

Kết quả đánh giá mô hình Naïve Bayes:

```
Naive Bayes Accuracy: 0.8993892639022822
              precision    recall  f1-score   support

         0.0       0.96      0.93      0.95      2921
         1.0       0.28      0.40      0.33       190

    accuracy                           0.90      3111
   macro avg       0.62      0.67      0.64      3111
weighted avg       0.92      0.90      0.91      3111
```

Hình 31: Kết quả đánh giá mô hình Naïve Bayes

6.8. Chọn mô hình có hiệu suất tốt nhất để dự đoán trên tập kiểm tra

Kết quả so sánh các mô hình bằng chỉ số Accuracy để chọn ra mô hình tốt nhất cho tập dữ liệu:

```
Logistic Regression Accuracy: 0.9395692703310833
k-NN Accuracy: 0.9350691096110575
SVM Accuracy: 0.9389263902282224
Decision Tree Accuracy: 0.9080681452909033
Random Forest Accuracy: 0.9292831886853102
XGBoost Accuracy: 0.9347476695596272
Neural Networks Accuracy: 0.9392478302796529
Naive Bayes Accuracy: 0.8993892639022822

Best model: LogisticRegression with accuracy: 0.9395692703310833
```

Hình 32: So sánh chỉ số Accuracy của các mô hình

Vậy, mô hình Logistic Regression có hiệu suất cao nhất trên tập dữ liệu huấn luyện, nên sẽ dùng mô hình Logistic Regression để dự đoán kết quả trên tập kiểm tra.

6.9. Viết kết quả dự đoán ra file submission.csv

Dùng mô hình Logistic Regression để dự đoán kết quả nguy cơ đột quỵ trên tập kiểm tra sau đó ghi ra một file submission.csv.

	id	stroke
0	15304	0.070772
1	15305	0.301295
2	15306	0.000844
3	15307	0.041666
4	15308	0.013048
5	15309	0.017934
6	15310	0.020196
7	15311	0.053370
8	15312	0.001315
9	15313	0.036862

Hình 33: File submission.csv dùng để ghi kết quả dự đoán đột quỵ

CHƯƠNG 7 – ỨNG DỤNG TRI THỨC VỀ DỰ ĐOÁN NGUY CƠ ĐỘT QUỴ

7.1. Ứng Dụng Trong Lĩnh Vực Y Tế

Hỗ Trợ Quyết Định Lâm Sàng: Mô hình dự đoán nguy cơ đột quỵ có thể được tích hợp vào hệ thống quản lý bệnh viện để hỗ trợ các bác sĩ trong việc ra quyết định lâm sàng. Khi một bệnh nhân nhập viện, hệ thống có thể tự động đánh giá nguy cơ đột quỵ và cảnh báo bác sĩ.

Chương Trình Sàng Lọc và Phòng Ngừa: Các tổ chức y tế có thể sử dụng mô hình để thiết lập các chương trình sàng lọc nguy cơ đột quỵ cho các nhóm dân số có nguy cơ cao, từ đó thực hiện các biện pháp phòng ngừa kịp thời như điều chỉnh lối sống, kiểm soát các yếu tố nguy cơ.

Giáo Dục Sức Khỏe: Dựa trên kết quả dự đoán, các chiến dịch giáo dục sức khỏe có thể được thiết kế để nâng cao nhận thức cộng đồng về nguy cơ đột quỵ và cách phòng ngừa.

7.2. Ứng Dụng Trong Các Nghiên Cứu Y Học

Nghiên Cứu Dịch Tễ Học: Các nhà nghiên cứu có thể sử dụng mô hình để phân tích dữ liệu và tìm ra các yếu tố nguy cơ mới liên quan đến đột quỵ, từ đó phát triển các phương pháp phòng ngừa và điều trị hiệu quả hơn.

Thử Nghiệm Lâm Sàng: Mô hình có thể giúp xác định các nhóm bệnh nhân phù hợp cho các thử nghiệm lâm sàng, đặc biệt là các thử nghiệm liên quan đến phòng ngừa và điều trị đột quỵ.

7.3. Ứng Dụng Trong Công Nghệ Sức Khỏe Cá Nhân

Ứng Dụng Di Động và Thiết Bị Đeo Thông Minh: Các ứng dụng sức khỏe trên điện thoại di động và thiết bị đeo thông minh có thể tích hợp mô hình dự đoán để cung cấp cảnh báo cá nhân về nguy cơ đột quỵ, giúp người dùng theo dõi và quản lý sức khỏe của họ một cách chủ động.

Giám Sát Sức Khỏe Từ Xa: Các hệ thống giám sát sức khỏe từ xa có thể sử dụng mô hình để theo dõi liên tục các chỉ số sức khỏe của bệnh nhân và đưa ra cảnh báo sớm về nguy cơ đột quỵ, đặc biệt hữu ích cho những người cao tuổi hoặc sống một mình.

BẢNG PHÂN CÔNG CÔNG VIỆC

Mssv	Họ và tên	Công việc	Mức độ hoàn thành
52100306	Nguyễn Khắc Anh Tài	Thuyết trình, trực quan hóa dữ liệu, ứng dụng tri thức vào cuộc sống, tổng hợp và format báo cáo.	100%
52100239	Trần Nam Đăng Khoa	Slide, trực quan hóa dữ liệu, thống kê thông số bộ dữ liệu, các phương pháp đánh giá hiệu suất mô hình.	100%
52100312	Lê Tuấn Thành	Slide, sơ lược đột quỵ, tổng quan về đề tài, chuẩn bị dữ liệu, tìm hiểu các thuật toán phù hợp.	100%
52100904	Phạm Hoàng Trung Kiên	Thu thập dữ liệu, mục tiêu nghiên cứu và phạm vi đề tài, mô tả và khám phá bộ dữ liệu, tiền xử lí dữ liệu.	100%
52100254	Trần Quang Luân	Phân tích, tìm hiểu các thuật toán phù hợp cho bài toán, đánh giá kết quả các mô hình.	100%
52100977	Nguyễn Đức Minh	Phân tích, tìm hiểu các thuật toán phù hợp cho bài toán, đánh giá kết quả các mô hình.	100%

Bảng 2: Bảng phân công công việc

TÀI LIỆU THAM KHẢO

Tiếng Việt

1. Vinmec, "*Đột quỵ: Nguyên nhân, dấu hiệu nhận biết, cách phòng tránh*".
2. aws.amazon.com (01/03/2023), *"Trực quan hóa dữ liệu là gì ?"*.
3. machinelearningcoban.com (08/01/2017), " *K-nearest neighbors*".
4. Machinelearningcoban.com (09/04/2017), *"Support Vector Machine"*.
5. machinelearningcoban.com (08/08/2017), "*Naive Bayes Classifier*".

Tiếng Anh

6. Kaggle, "*Binary Classification with a Tabular Stroke Prediction Dataset*".
7. Vijay Kanade (03/04/2022), "*What Is Logistic Regression? Equation, Assumptions, Types, and Best Practices*".
8. Anshul Saini (18/04/2024), "*What is Decision Tree? [A Step-by-Step Guide]*".
9. Sruthi E R (19/04/2024), "*Understand Random Forest Algorithms With Examples (Updated 2024)*".