Final Assignment

STROKE PREDICTION

Lương Quang Trường - Nhóm 4 - DA23

TABLE OF CONTENT

GIỚI THIỆU

- Giới thiệu bộ dữ liệu
- Đặt vấn đề

PHÂN TÍCH

- Làm sạch dữ liệu
- Kiểm định mô hình

KẾT LUẬN

GIỚI THIỆU

- Dữ liệu:
  - Bộ dữ liệu chứa thông tin về các bệnh nhân đi khám tim.
  - Có các thông tin nhân khẩu học cũng như là tiền sử bệnh, của bệnh nhân khám.
- Đặt vấn đề:
  - Đột quỵ là nguyên nhân gây chết nhiều thứ 2 (theo WHO), khoảng 11% số người tử vong.Dựa vào các thông tin thu thập được về các bệnh nhân để xây dựng mô hình dự báo các bệnh nhân có khả năng gặp đột quỵ.
- Nguồn: https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset

	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
0	9046	Male	67.0	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1
1	51676	Female	61.0	0	0	Yes	Self-employed	Rural	202.21	NaN	never smoked	1
2	31112	Male	80.0	0	1	Yes	Private	Rural	105.92	32.5	never smoked	1
3	60182	Female	49.0	0	0	Yes	Private	Urban	171.23	34.4	smokes	1
4	1665	Female	79.0	1	0	Yes	Self-employed	Rural	174.12	24.0	never smoked	1
...	...	...	...	...	...	...	...	...	...	...	...	...
5105	18234	Female	80.0	1	0	Yes	Private	Urban	83.75	NaN	never smoked	0
5106	44873	Female	81.0	0	0	Yes	Self-employed	Urban	125.20	40.0	never smoked	0
5107	19723	Female	35.0	0	0	Yes	Self-employed	Rural	82.99	30.6	never smoked	0
5108	37544	Male	51.0	0	0	Yes	Private	Rural	166.29	25.6	formerly smoked	0
5109	44679	Female	44.0	0	0	Yes	Govt_job	Urban	85.28	26.2	Unknown	0

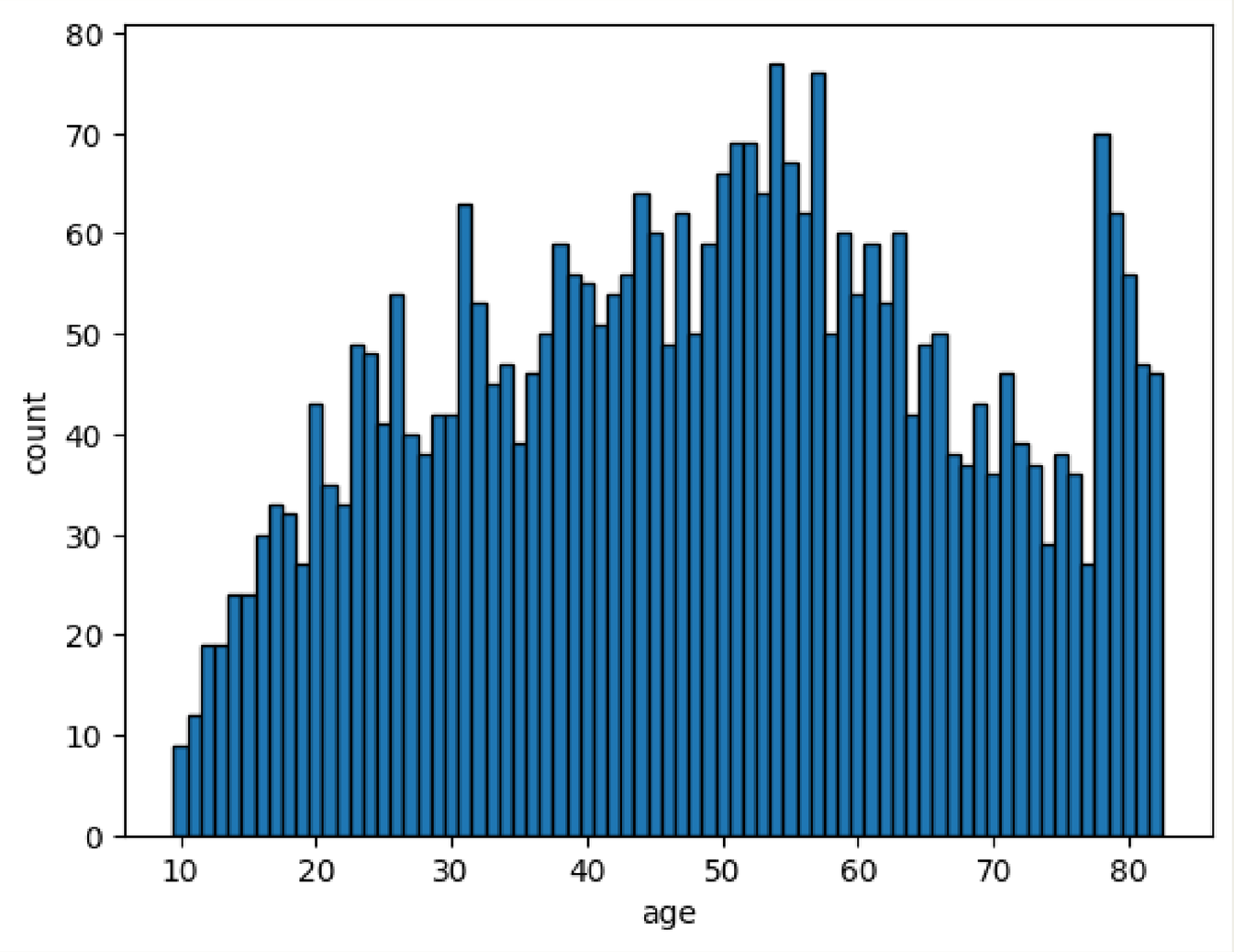
PHÂN TÍCH

Các biến của bộ dữ liệu trước khi làm sạch (5110 quan sát)

- **id**
- **gender: "Male", "Female" or "Other"**
- **age**
- **hypertension: 0/1**
- **heart_disease: 0/1**
- **ever_married: "No"/"Yes"**
- **work_type: "children", "Govt_job", "Never_worked", "Private" or "Self-employed"**
- **residence_type: "Rural"/"Urban"**
- **avg_glucose_level:**
- **bmi:**
- **smoking_status: "formerly smoked"/ "never smoked"/"smokes"/"Unknown"**
- **stroke -> biến phân loại**

PHÂN TÍCH

- Độ tuổi của những quan sát chủ yếu tập trung ở khoảng từ 50-60 tuổi.
- Đặc biệt nhiều những quan sát ở khoảng gần 80 tuổi.



C L E A N I N G

- Dựa vào biến bmi, tạo cột body_type dựa trên các khoảng bmi

```
# thêm cột body_type:
body_type = []
for row in df['bmi']:
    if row < 18:
        body_type.append("underweight")
    elif row >= 18 and row < 25:
        body_type.append("healthy")
    elif row >= 25 and row < 30:
        body_type.append("overweight")
    elif row >= 30:
        body_type.append("obese")
```

✓ 0.0s

- Sử dụng các LabelEncoder để mã hóa các cột string.

```
non_objects = ["age", "hypertension", "heart_disease", "avg_glucose_level", "stroke"]
objects = ["gender", "ever_married", "work_type", "Residence_type", "smoking_status", "body_type"]
```

✓ 0.0s

- Sử dụng MinMaxScaler để chuẩn hóa các cột số

- Drop cột id, các dòng có NaN

C L E A N I N G

- Biến gender có thể là "Female", "Male", "Other", tuy nhiên do Other chỉ có 1 row nên chúng ta sẽ drop.

```
df["gender"].value_counts()
```
✓ 0.0s

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

- Biến smoking_status có nhiều value "Unknown" nên drop các dòng có chứa "Unknown"

```
df["smoking_status"].value_counts()
```
✓ 0.0s

```
never smoked       1892
Unknown            1544
formerly smoked     885
smokes              789
Name: smoking_status, dtype: int64
```

- Khi xét đến biến stroke, ta có thể thấy số lượng 0 và 1 chênh lệch nhau nhiều -> imbalanced dataset -> downsample

```
df["stroke"].value_counts(normalize = True)
```
✓ 0.0s

```
0    0.951272
1    0.048728
Name: stroke, dtype: float64
```

PHÂN TÍCH

Các biến của bộ dữ liệu trước khi làm sạch (5110 quan sát)

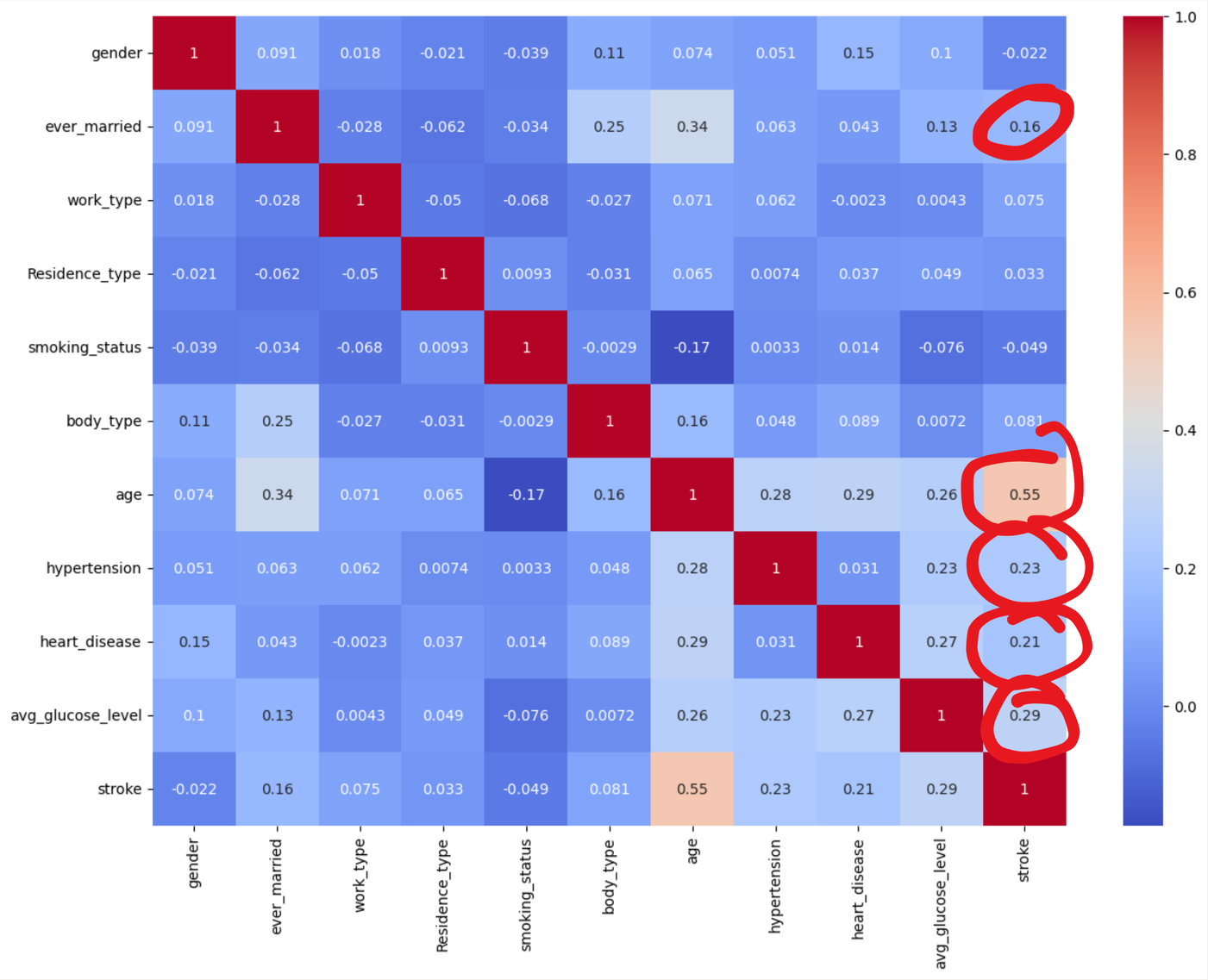
- **id**
- **gender: "Male", "Female" or "Other"**
- **age**
- **hypertension: 0/1**
- **heart_disease: 0/1**
- **ever_married: "No"/"Yes"**
- **work_type: "children", "Govt_jov", "Never_worked", "Private" or "Self-employed"**
- **residence_type: "Rural"/"Urban"**
- **avg_glucose_level:**
- **bmi:**
- **smoking_status: "formerly smoked"/ "never smoked"/"smokes"/"Unknown"**
- **stroke -> biến phân loại**

Các biến của bộ dữ liệu sau khi làm sạch (360 quan sát)

-
- **gender: 0/1**
- **age: normalized**
- **hypertension: 0/1**
- **heart_disease: 0/1**
- **ever_married: 0/1**
- **work_type: 0/1/2/3/4**
-
- **residence_type: 0/1**
- **avg_glucose_level: normalized**
- **body_type:**
- **smoking_status: 0/1/2**

- **stroke -> biến phân loại**

TƯƠNG QUAN VÀ CHỌN BIẾN



Dựa trên bảng tương quan, ta lựa chọn các biến:

- ever_married
- age
- hypertension
- heart_disease
- avg_glucose_level

để xây dựng mô hình phân loại

KIỂM ĐỊNH CÁC MÔ HÌNH

	precision	recall	f1-score	support
0	0.77	0.67	0.71	54
1	0.70	0.80	0.75	54
accuracy			0.73	108
macro avg	0.74	0.73	0.73	108
weighted avg	0.74	0.73	0.73	108

Logistic

	precision	recall	f1-score	support
0	0.82	0.67	0.73	54
1	0.72	0.85	0.78	54
accuracy			0.76	108
macro avg	0.77	0.76	0.76	108
weighted avg	0.77	0.76	0.76	108

SVM

	precision	recall	f1-score	support
0	0.77	0.74	0.75	54
1	0.75	0.78	0.76	54
accuracy			0.76	108
macro avg	0.76	0.76	0.76	108
weighted avg	0.76	0.76	0.76	108

Decision Tree (max_depth = 4)

	precision	recall	f1-score	support
0	0.79	0.69	0.73	54
1	0.72	0.81	0.77	54
accuracy			0.75	108
macro avg	0.75	0.75	0.75	108
weighted avg	0.75	0.75	0.75	108

Random Forest (n_estimators = 17)

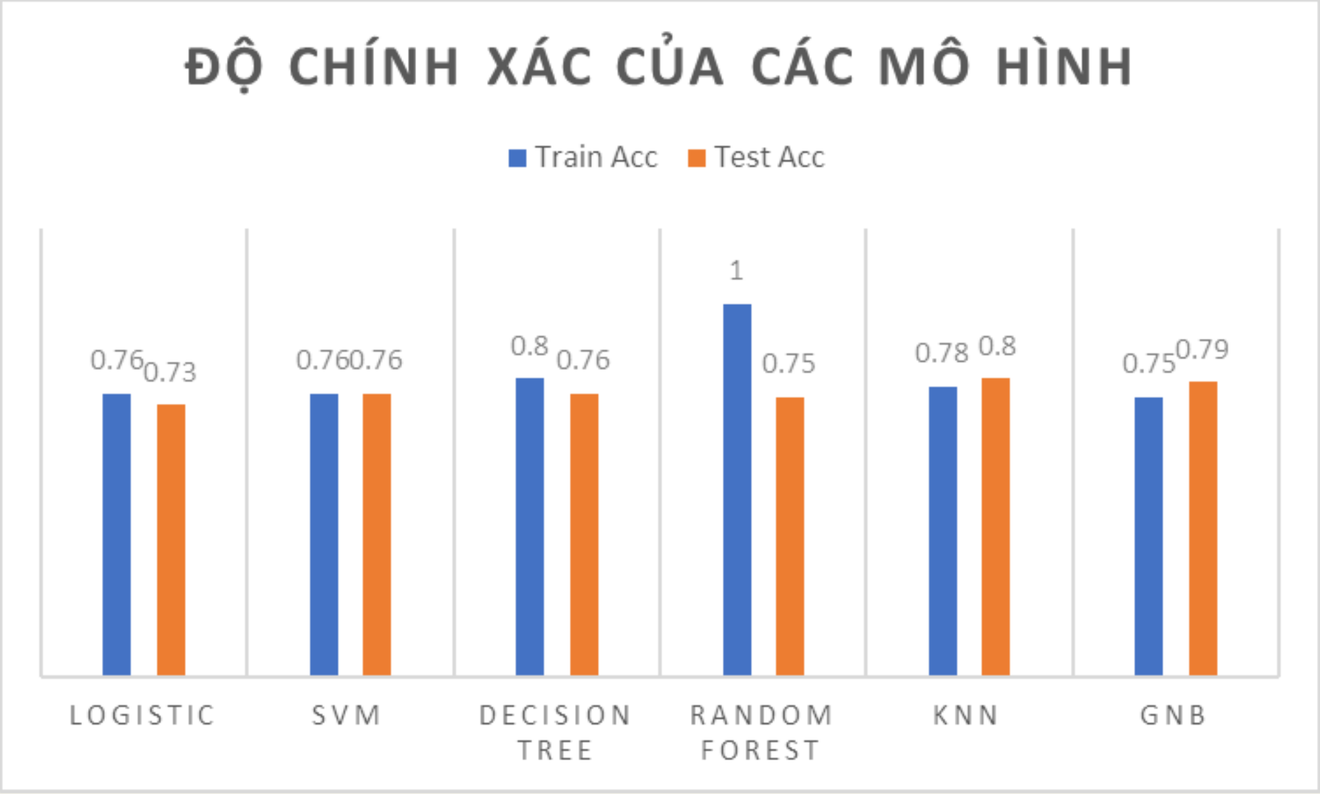
	precision	recall	f1-score	support
0	0.81	0.78	0.79	54
1	0.79	0.81	0.80	54
accuracy			0.80	108
macro avg	0.80	0.80	0.80	108
weighted avg	0.80	0.80	0.80	108

KNN (n_neighbors = 12)

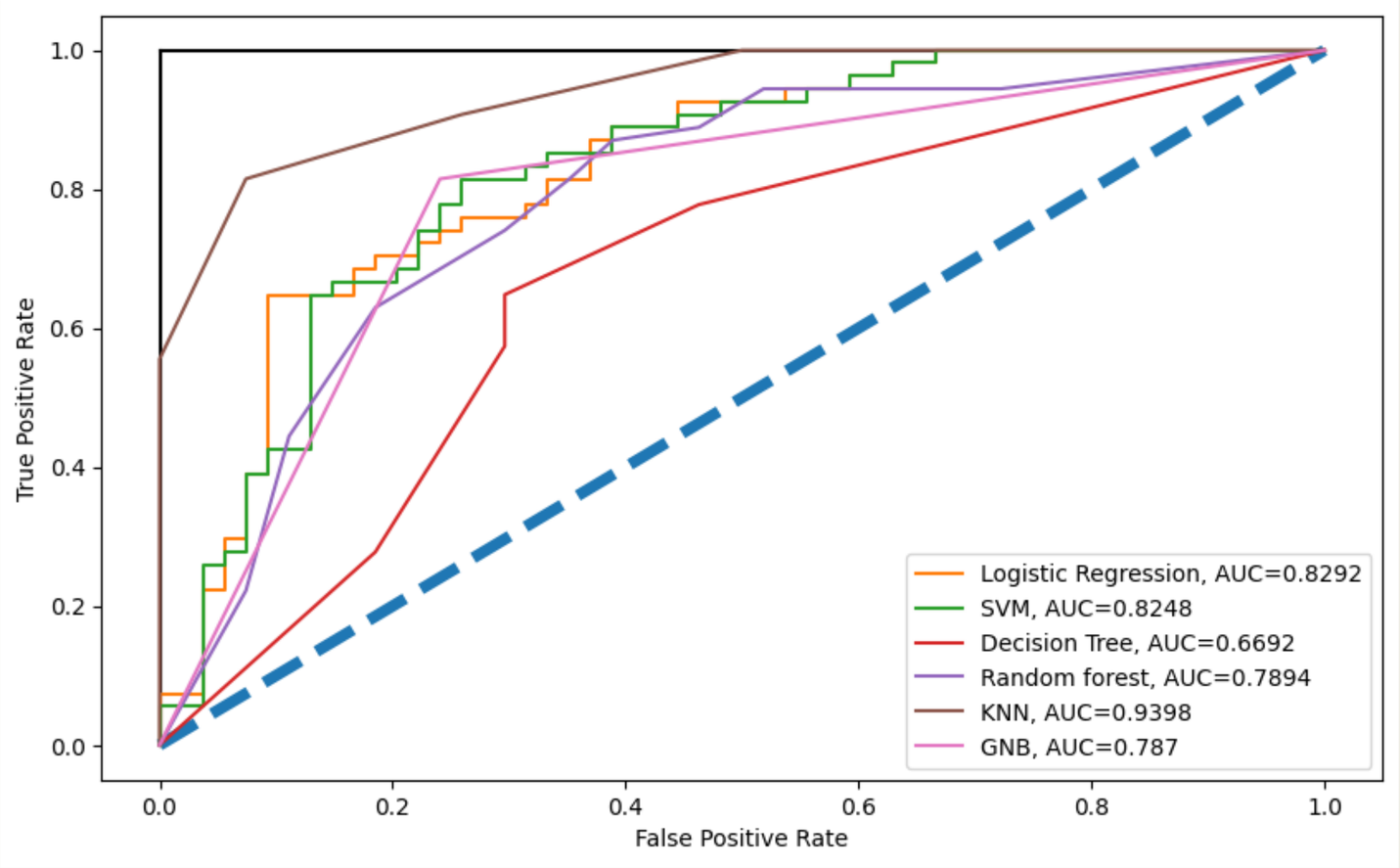
	precision	recall	f1-score	support
0	0.80	0.76	0.78	54
1	0.77	0.81	0.79	54
accuracy			0.79	108
macro avg	0.79	0.79	0.79	108
weighted avg	0.79	0.79	0.79	108

Gaussian Naive Bayes

KIỂM ĐỊNH CÁC MÔ HÌNH



KIỂM ĐỊNH CÁC MÔ HÌNH



Kết luận:

- **Các yếu tố dùng để dự báo đột quỵ bao gồm: tình trạng hôn nhân, tuổi, tiền sử bệnh tim mạch, huyết áp cao và lượng đường trong máu.**
- **Các mô hình phân loại dự báo có độ chính xác khá cao, nhưng tốt nhất là mô hình KNN.**

Thank you!