Đề tài: DỰ ĐOÁN NGUY CƠ ĐỘT QUỴ

Lớp: 23CL10DN2

Nhóm: 3

- Quang Long (Dev)
- Thành Luân (Docx)
- Hoàng Quân (Docx)

MỤC LỤC

CHƯƠNG 1: GIỚI THIỆU TỔNG QUAN

1.1. Bối cảnh và Lý do chọn đề tài

- **Tầm quan trọng:** Đột quỵ là một trong những nguyên nhân gây tử vong và tàn tật hàng đầu.
- **Vấn đề nghiên cứu:** "Làm thế nào để ứng dụng các kỹ thuật phân tích dữ liệu và học máy trên tập dữ liệu về sức khỏe để xác định các yếu tố nguy cơ chính và xây dựng một mô hình dự đoán chính xác khả năng bị đột quỵ của một cá nhân?"

```
import pandas as pd

# Load the dataset
df = pd.read_csv('healthcare-dataset-stroke-data.csv.xls')
df.info()
```

1.2. Mục tiêu dự án

- **Mục tiêu chính:**
  - Xác định yếu tố nguy cơ chính.
  - Xây dựng mô hình dự đoán chính xác.
  - Đưa ra khuyến nghị phòng ngừa.

- **Mục tiêu cụ thể:**
  - Khám phá và làm sạch dữ liệu.
  - Trực quan hóa và kiểm định thống kê.
  - So sánh và xử lý mất cân bằng.
  - Đánh giá chi tiết mô hình.

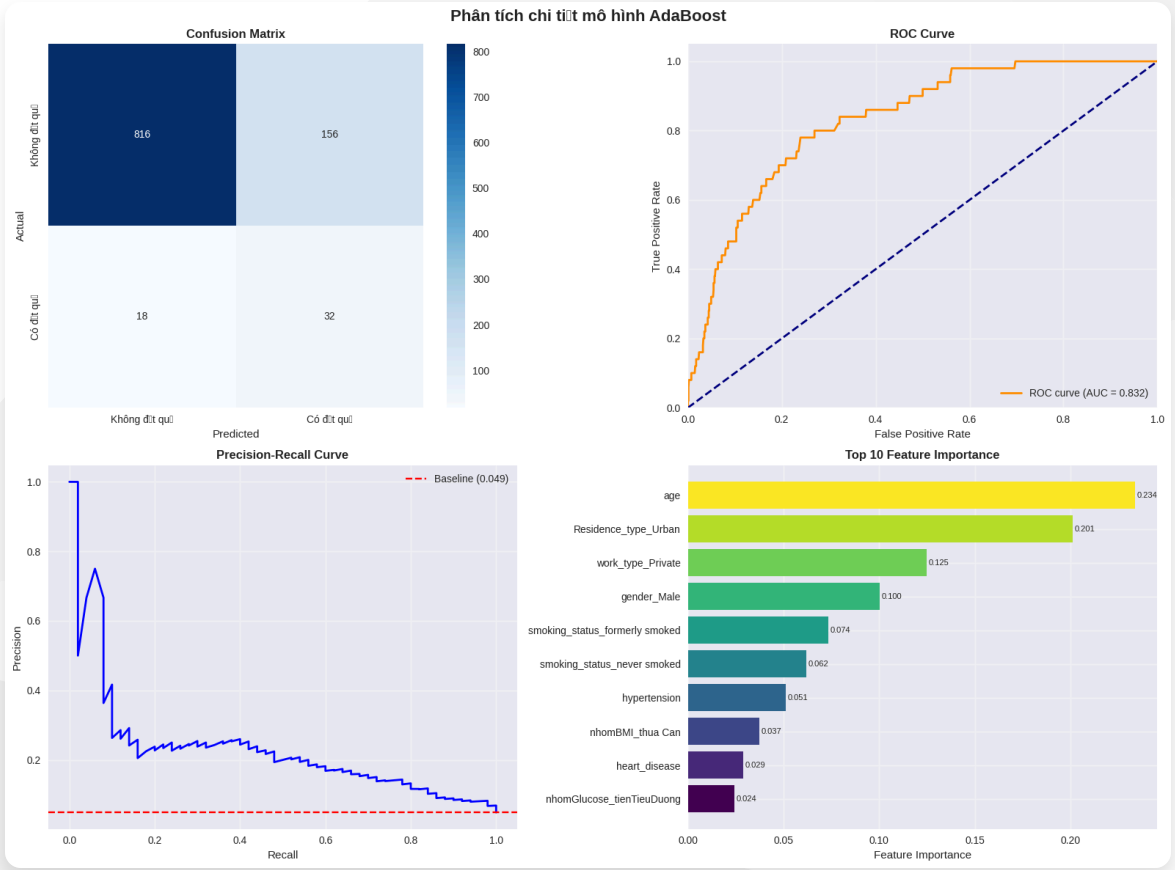
1.3. Phạm vi và Đối tượng nghiên cứu

- **Phạm vi dữ liệu:**
  - Kaggle: "Healthcare Dataset Stroke Data".
  - 5,110 mẫu, 11 biến độc lập.
- **Phạm vi bài toán:**
  - Phân loại nhị phân (Đột quỵ: 1/0).
- **Đối tượng:**
  - Bệnh nhân và chuyên gia y tế.

2.2. Phương pháp tiếp cận

Dự án được thực hiện theo một quy trình khoa học dữ liệu chuẩn:

1. **Khám phá dữ liệu (EDA)**
2. **Tiền xử lý và Kỹ thuật tạo biến**
3. **Phát triển mô hình**
4. **Đánh giá và Lựa chọn**
5. **Tổng hợp và Đề xuất**



CHƯƠNG 2: KHÁM PHÁ VÀ TIỀN XỬ LÝ DỮ LIỆU

2.1. Giới thiệu tập dữ liệu

- **Nguồn:** Kaggle: Stroke Prediction Dataset
- **Kích thước:** 5,110 mẫu, 12 thuộc tính.
- **Biến mục tiêu:** `stroke` (1: có, 0: không).

Biến	Mô tả
`age`	Tuổi
`gender`	Giới tính
`bmi`	Chỉ số khối cơ thể
`hypertension`	Tăng huyết áp
`heart_disease`	Bệnh tim mạch

2.2. Khám phá dữ liệu sơ bộ (EDA)

- **Phân tích thống kê mô tả:**
  - `age` : Dao động từ 0.08 đến 82 tuổi, trung bình 43.2.
  - `avg_glucose_level` : Biến thiên lớn, từ 55.12 đến 271.74.
  - `bmi` : Trung bình 28.9.
- **Kiểm tra dữ liệu:**
  - Cột `bmi` có 201 giá trị bị thiếu (chiếm 3.9%).

```
import pandas as pd

df = pd.read_csv('healthcare-dataset-stroke-data.csv')
df.head()
```

2.3. Làm sạch dữ liệu

- **Xử lý giá trị thiếu:** Thay thế giá trị thiếu trong `bmi` bằng giá trị trung vị (median) do phân phối của `bmi` bị lệch.
- **Xử lý giá trị ngoại lai:** Giữ lại các giá trị ngoại lai ở `avg_glucose_level` vì chúng có thể là những trường hợp bệnh lý thực tế và chứa thông tin quan trọng.
- **Kiểm tra tính hợp lệ:**
  - Loại bỏ cột `id` không mang giá trị phân tích.
  - Loại bỏ 1 dòng có `gender` là "Other" do số lượng quá ít.
  - Không có dòng dữ liệu nào bị trùng lặp.

```
# Xử lý giá trị thiếu
df['bmi'].fillna(df['bmi'].median(), inplace=True)

# Loại bỏ cột id và dòng có gender='Other'
df.drop('id', axis=1, inplace=True)
df = df[df['gender'] != 'Other']
```

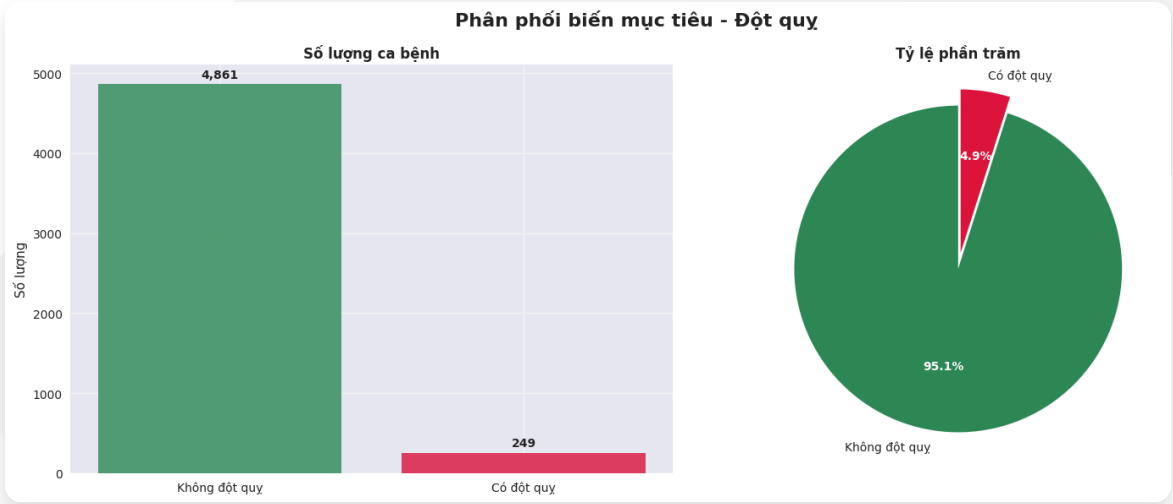
2.4. Kỹ thuật tạo biến (Feature Engineering)

- **Tạo các biến nhóm:**
  - `nhomTuoi`: Phân bệnh nhân vào các nhóm tuổi (Vị thành niên, Thanh niên, Trung niên, Cao niên).
  - `nhomBMI`: Phân loại tình trạng cơ thể dựa trên chỉ số BMI (Thiếu cân, Bình thường, Thừa cân, Béo phì).
  - `nhomGlucose`: Phân loại mức đường huyết (Bình thường, Tiền tiểu đường, Tiểu đường).
- **Tạo biến tổng hợp "Điểm nguy cơ":**
  - `diemNguyCo`: Tổng hợp các yếu tố rủi ro vào một chỉ số duy nhất.

```
# Tạo biến nhóm tuổi
df['nhomTuoi'] = pd.cut(df['age'], bins=[0, 18, 35, 55, 100], labels=['Vị thành niên', 'Thanh niên', 'Trung niên', 'Cao niên'])

# Tạo biến nhóm BMI
df['nhomBMI'] = pd.cut(df['bmi'], bins=[0, 18.5, 24.9, 29.9, 100], labels=['Thiếu cân', 'Bình thường', 'Thừa cân', 'Béo phì'])

# Tạo biến nhóm Glucose
df['nhomGlucose'] = pd.cut(df['avg_glucose_level'], bins=[0, 140, 200, 300], labels=['Bình thường', 'Tiền tiểu đường', 'Tiểu đường'])
```

2.5. Kiểm tra chất lượng dữ liệu sau xử lý

- **Kết quả:** Tập dữ liệu cuối cùng có 5,109 mẫu và 15 cột, đã sạch và sẵn sàng cho phân tích.
- **Vấn đề nổi bật:** Dữ liệu mất cân bằng nghiêm trọng.
  - **95.1%** không đột quỵ (4,860 trường hợp).
  - **4.9%** có đột quỵ (249 trường hợp).
- **Thách thức:** Cần xử lý để mô hình không bị thiên vị.

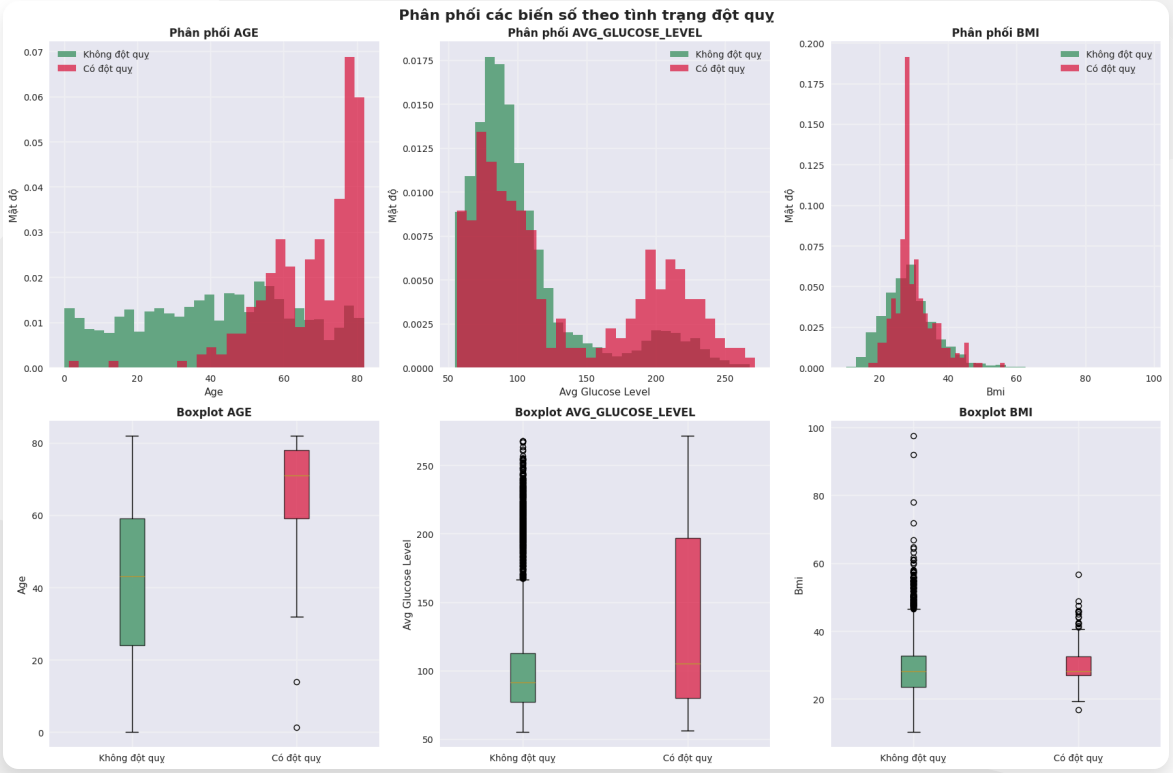


CHƯƠG 3: PHÂN TÍCH THỐNG KÊ VÀ TRỰC QUAN HÓA

3.1. Phân tích các biến định lượng

- **Phân phối:**
  - Nhóm bệnh nhân bị đột quỵ có `age` và `avg_glucose_level` trung bình cao hơn đáng kể.
- **Kiểm định T-test:**
  - `age`, `avg_glucose_level`, và `bmi` đều có sự khác biệt rất có ý nghĩa thống kê giữa nhóm bị đột quỵ và không bị đột quỵ (p-value < 0.05).

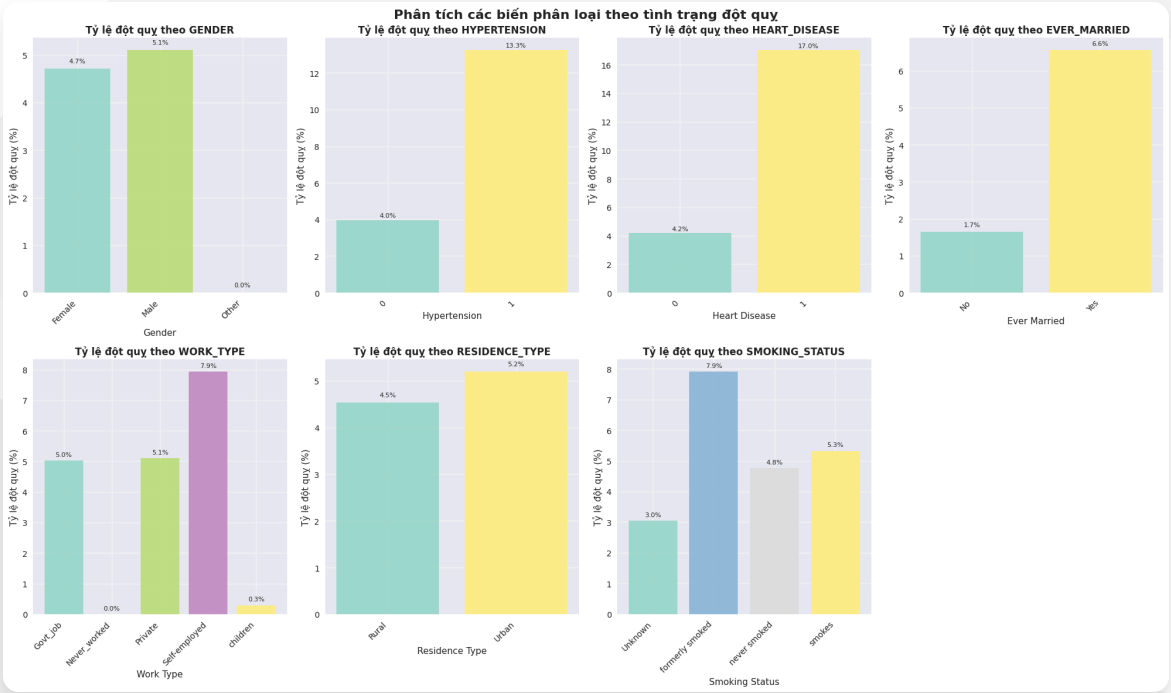
```
from scipy.stats import ttest_ind

# So sánh tuổi
group1 = df[df['stroke'] == 1]['age']
group2 = df[df['stroke'] == 0]['age']
stat, p = ttest_ind(group1, group2)
print(f'T-test for age: p-value={p:.3f}')
# T-test for age: p-value=0.000
```



3.2. Phân tích các biến định tính

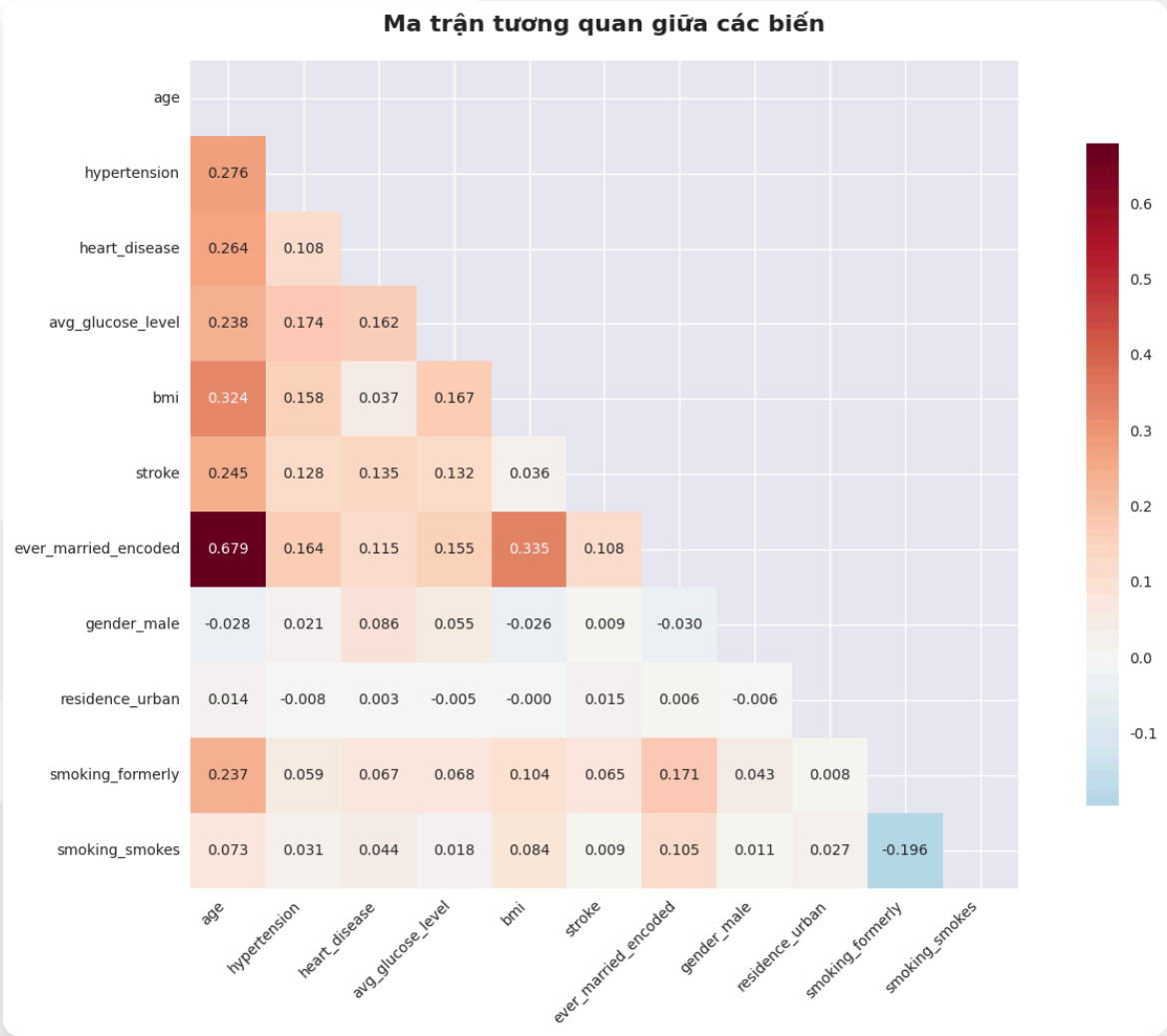
- **Tỷ lệ đột quỵ cao hơn ở các nhóm:**
  - Có tiền sử `hypertension` (13% vs 4%).
  - Có tiền sử `heart_disease` (17% vs 4%).
  - Đã `ever_married`.
  - `work_type` là `Self-employed`.
  - `smoking_status` là `formerly smoked` hoặc `smokes`.
- **Kiểm định Chi-square:** Tất cả các biến định tính đều có mối liên hệ có ý nghĩa thống kê với đột quỵ.

```
from scipy.stats import chi2_contingency
import pandas as pd

# Ví dụ với hypertension
contingency_table = pd.crosstab(df['hypertension'], df['stroke'])
chi2, p, _, _ = chi2_contingency(contingency_table)
print(f'Chi-square for hypertension: p-value={p:.3f}')
# Chi-square for hypertension: p-value=0.000
```



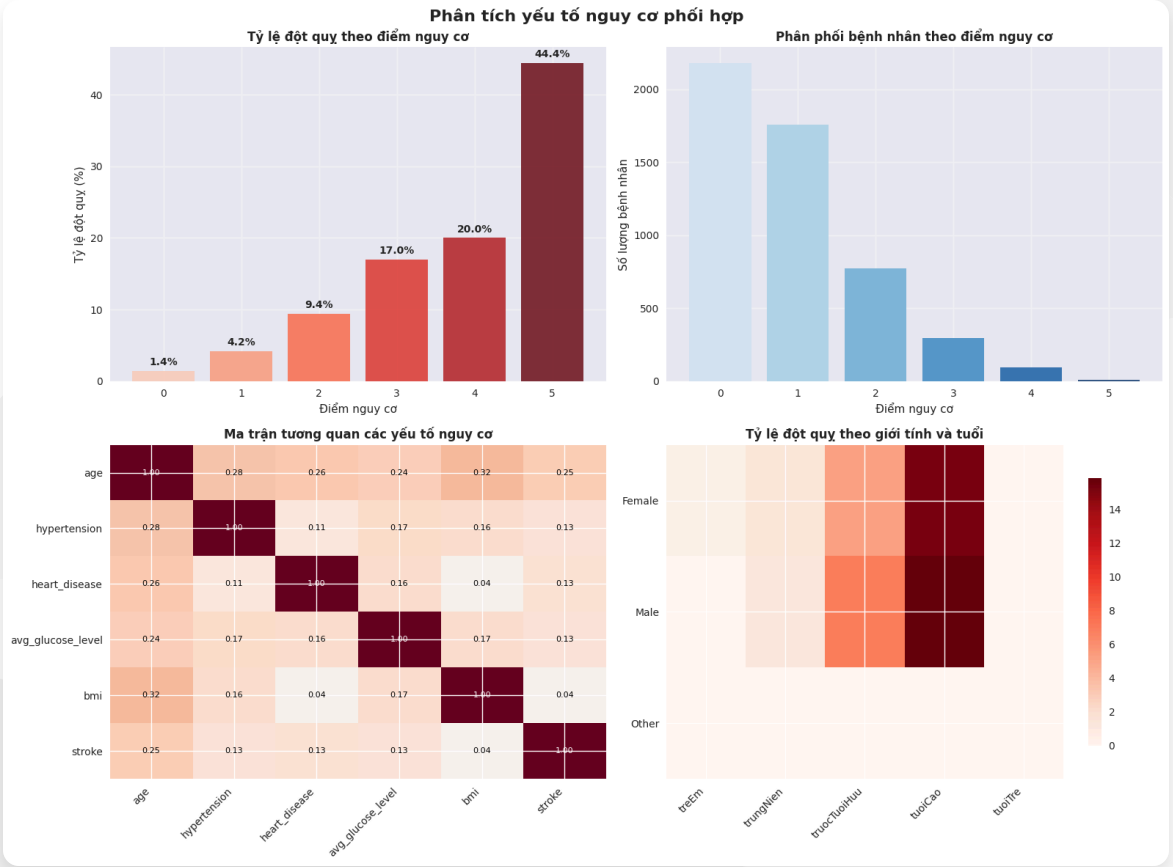
3.3. Phân tích tương quan

- **Ma trận tương quan:**
  - Tương quan dương mạnh nhất với `stroke` là `age`.
  - Các cặp biến có tương quan hợp lý: `age` và `ever_married`, `age` và `hypertension`.
- **Kết luận:** Không có hiện tượng đa cộng tuyến nghiêm trọng giữa các biến độc lập, tốt cho việc xây dựng mô hình.



3.1. Tổng kết các phát hiện quan trọng

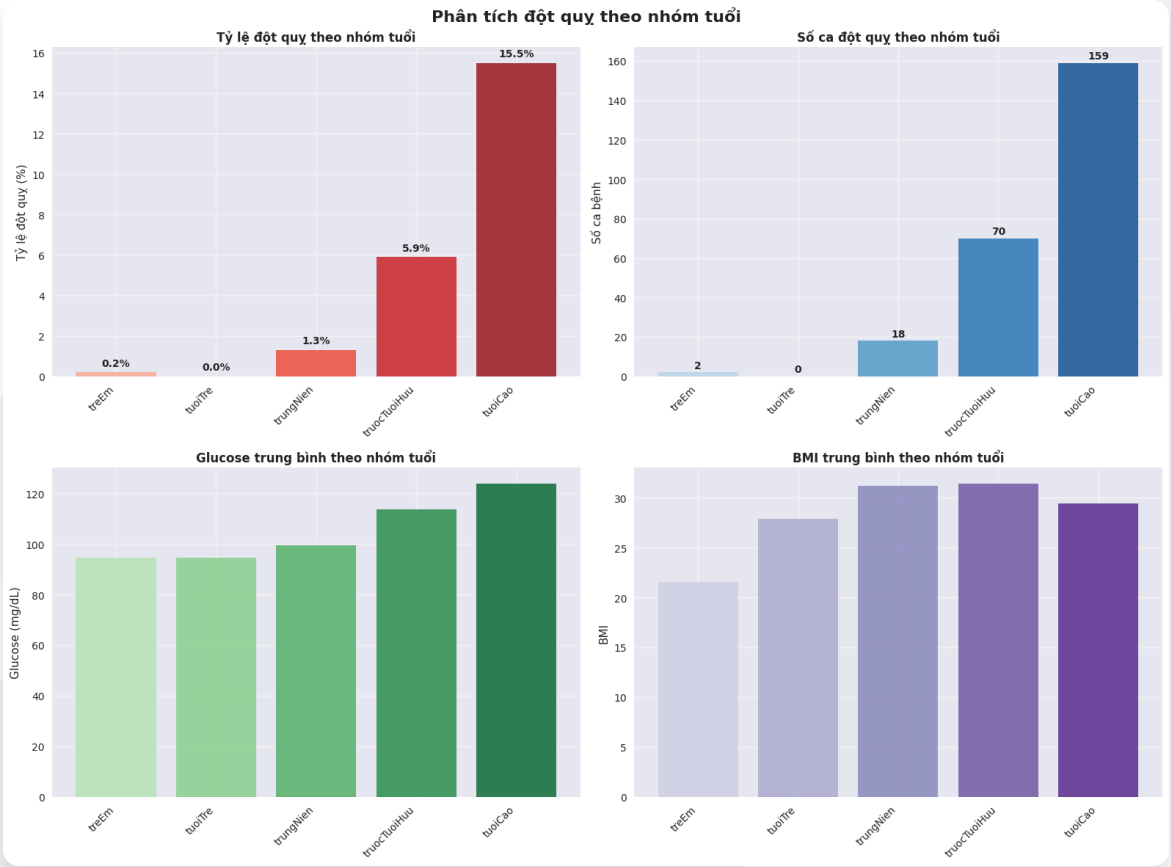
- **Yếu tố nguy cơ hàng đầu:**
  - `age`, `hypertension`, `heart_disease`, `avg_glucose_level`.
- **Các yếu tố ảnh hưởng khác:**
  - `ever_married`, `work_type`, `smoking_status`, `bmi`.
- **Vấn đề cần lưu ý:**
  - Mất cân bằng dữ liệu nghiêm trọng.



CHƯƠNG 4: XÂY DỰNG VÀ ĐÁNH GIÁ MÔ HÌNH DỰ ĐOÁN

4.1. Chuẩn bị dữ liệu cho Machine Learning

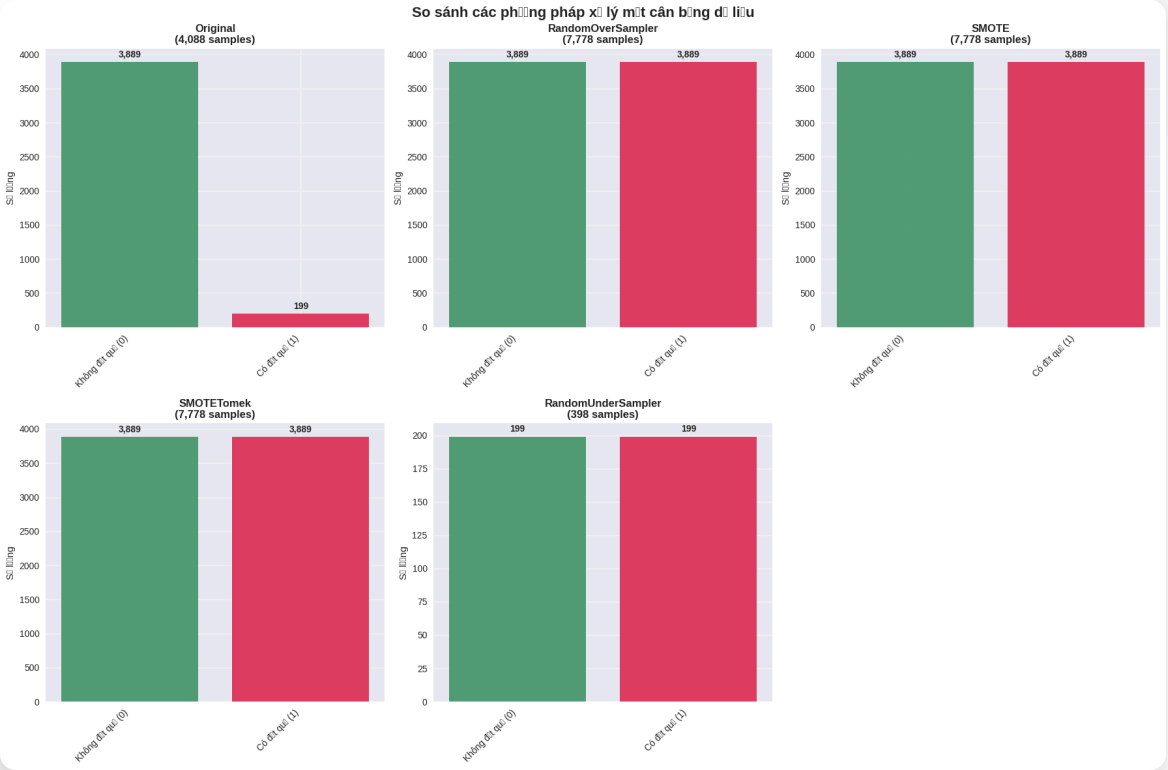
- **Lựa chọn biến:** Sử dụng tất cả các biến đã qua xử lý.
- **Phân chia:** 80% Training, 20% Test (`stratified`).
- **Pipeline:**
  - One-Hot Encoding (Biến định tính).
  - StandardScaler (Biến số).



4.2. Xử lý mất cân bằng dữ liệu

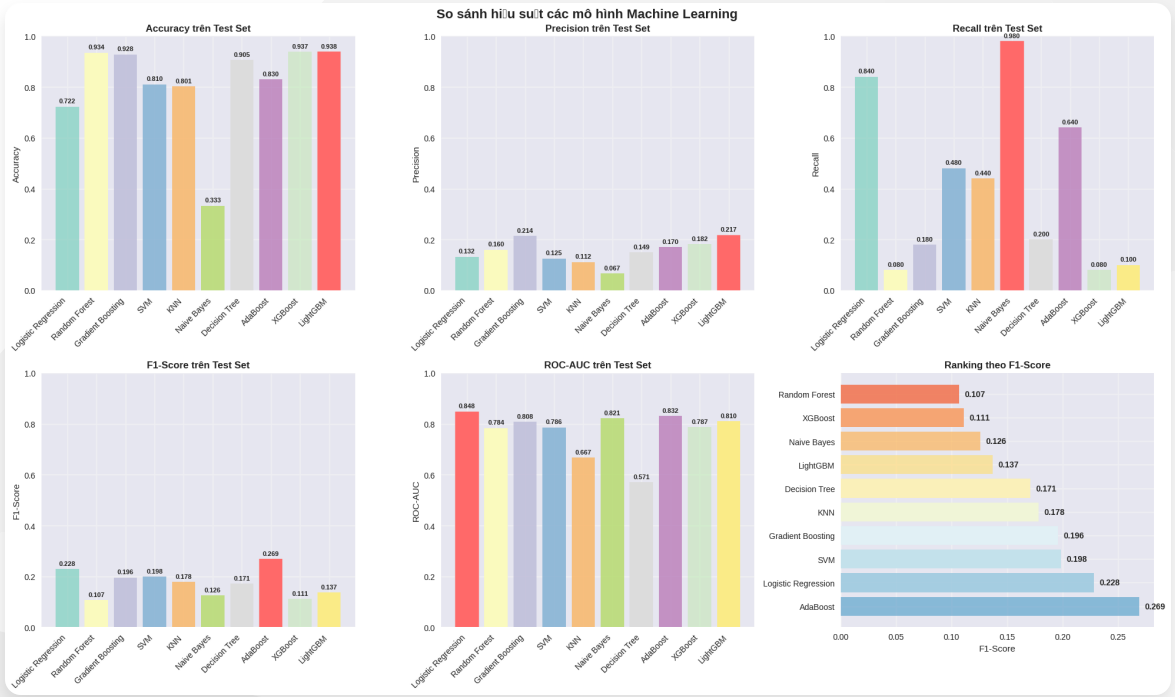
- **Phương pháp:** SMOTE (Synthetic Minority Over-sampling Technique).
- **Phạm vi áp dụng: Chỉ trên tập huấn luyện** để tránh rò rỉ dữ liệu.
- **Mục tiêu:** Tạo ra các mẫu tổng hợp cho lớp thiểu số ("có đột quỵ") để cân bằng dữ liệu huấn luyện.

```python
from imblearn.over_sampling import SMOTE
from sklearn.model_selection import train_test_split

# Phân chia dữ liệu
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42, stratify=y)

# Áp dụng SMOTE
smote = SMOTE(random_state=42)
X_train_smote, y_train_smote = smote.fit_resample(X_train, y_train)
```



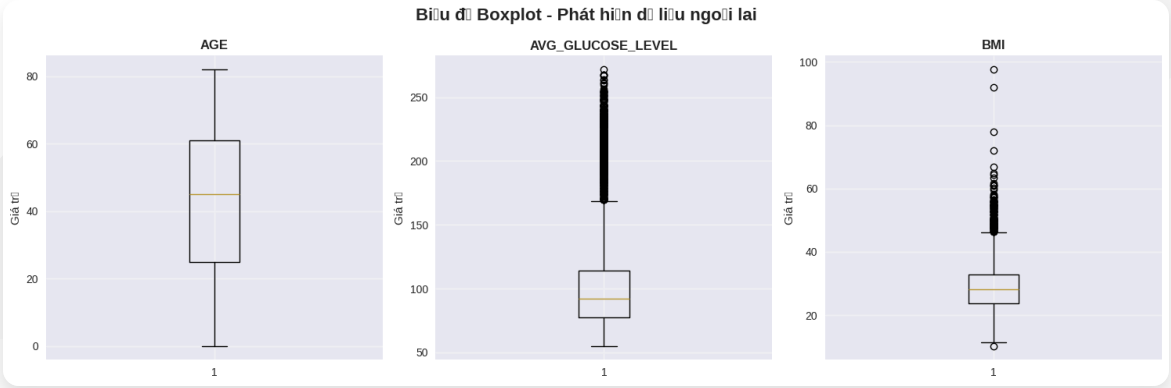
4.3. Xây dựng và so sánh các mô hình

- **Các thuật toán:** Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, XGBoost, LightGBM, SVM.
- **Chỉ số đánh giá:** F1-Score (quan trọng nhất), Recall, Precision, ROC-AUC.
- **Kết quả:** Các mô hình dựa trên cây (Random Forest, XGBoost, LightGBM) cho kết quả vượt trội. **LightGBM** nổi bật với F1-Score cao nhất.



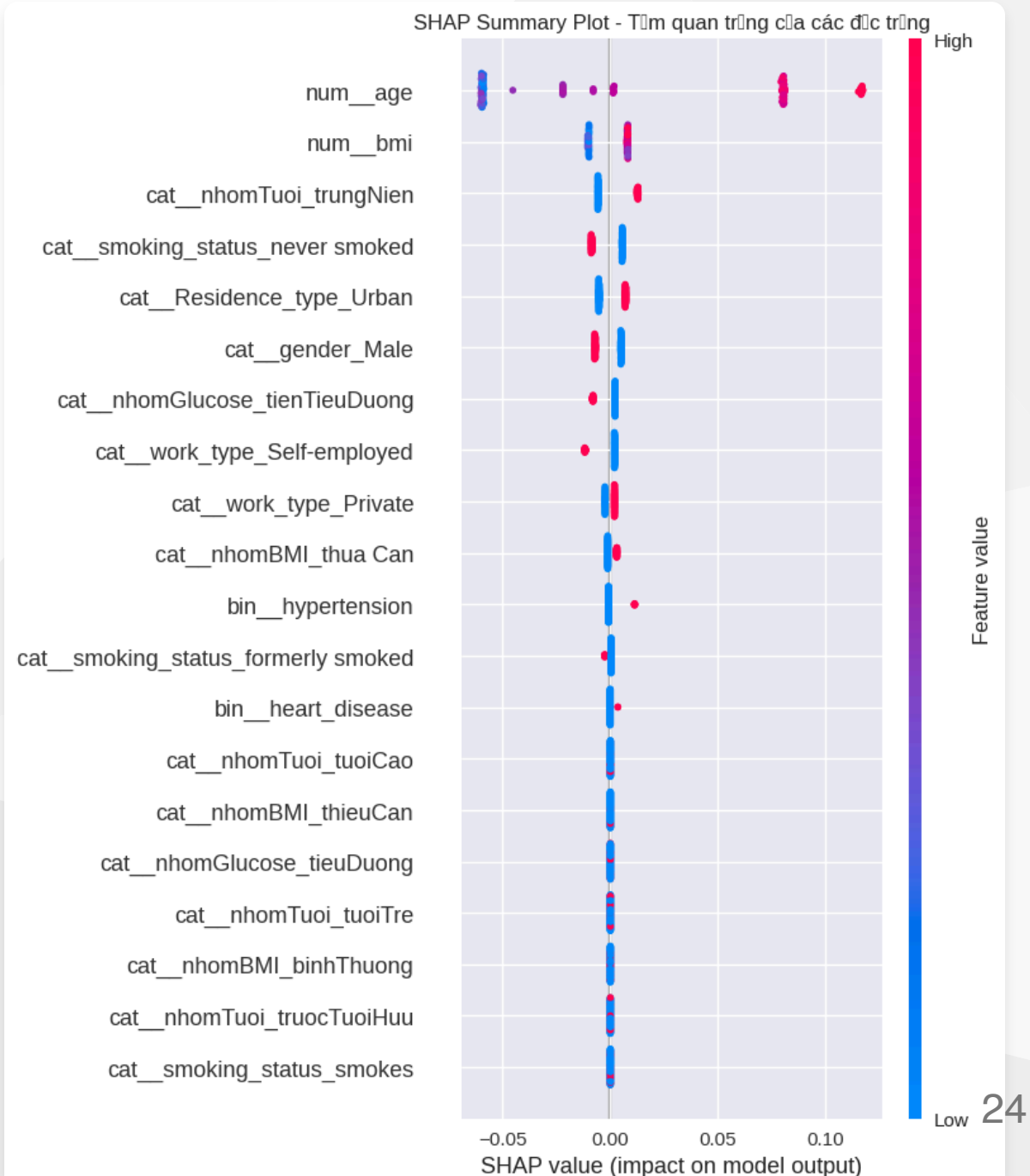
4.2. Lựa chọn và Tinh chỉnh mô hình

- **Mô hình tốt nhất:** LightGBM.
- **Tinh chỉnh:** GridSearchCV.
- **Đánh giá:**
  - Ma trận nhầm lẫn.
  - Đường cong ROC.



4.5. Diễn giải mô hình

- **Mức độ quan trọng của các biến (Feature Importance):**
  i. `age` (Tuổi)
  ii. `avg_glucose_level` (Mức đường huyết)
  iii. `bmi` (Chỉ số khối cơ thể)
  iv. `smoking_status` (Tình trạng hút thuốc)
- **Phân tích SHAP:** Giúp giải thích "đóng góp" của từng giá trị đặc trưng vào kết quả dự đoán cho từng trường hợp cụ thể, làm tăng tính minh bạch của mô hình.



CHƯƠNG 5: KẾT LUẬN VÀ ĐỀ XUẤT

5.1. Tóm tắt kết quả

- **Thành công:**
  - Quy trình phân tích hoàn chỉnh.
  - Xác định yếu tố nguy cơ chính.
  - Xây dựng mô hình LightGBM tốt.
- **Ý nghĩa thực tiễn:**
  - Tiềm năng xây dựng công cụ sàng lọc và cảnh báo sớm.

5.2. Hạn chế của dự án

- **Dữ liệu:**
  - Kích thước mẫu nhỏ.
  - Thiếu biến quan trọng.
- **Mô hình:**
  - SMOTE có thể không thực tế.
  - Cần kiểm định thêm.

5.3. Đề xuất và Hướng phát triển

- **Về dữ liệu:**
  - Thu thập thêm.
  - Dữ liệu theo thời gian.
- **Về ứng dụng:**
  - Xây dựng giao diện.
  - Thử nghiệm lâm sàng.
- **Về mô hình:**
  - Thử kỹ thuật khác.
  - Khám phá Deep Learning.
  - Mô hình diễn giải được.

Trân trọng cảm ơn!

Thông tin liên hệ:

- Gmail: long.lequang308@gmail.com

Tài liệu tham khảo:

1. World Health Organization - Stroke Guidelines
2. American Heart Association - Risk Factors
3. Scikit-learn Documentation
4. Pandas & NumPy Documentation
5. Seaborn & Matplotlib Documentation