

TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT TP. HỒ CHÍ MINH
KHOA CÔNG NGHỆ THÔNG TIN



HCMUTE

BÁO CÁO KẾT THÚC MÔN
CHUYÊN ĐỀ TỐT NGHIỆP 2
PHÂN TÍCH HỆ THỐNG DỮ LIỆU

SVTH	:	LÊ QUANG LONG	23810067
		TRẦN HOÀNG QUÂN	23810076
		TRỊNH THÀNH LUÂN	23810068

Khoá	:	2023 – 2025
Ngành	:	CÔNG NGHỆ THÔNG TIN
GVHD	:	Phan Thị Thể

TP.Hồ Chí Minh, tháng 10 năm 2025

MỤC LỤC

CHƯƠNG 1: GIỚI THIỆU TỔNG QUAN	1
1.1. Bối cảnh và Lý do chọn đề tài	1
1.1.1. Tầm quan trọng của việc dự đoán sớm đột quỵ	1
1.1.2. Vấn đề nghiên cứu	1
1.2. Mục tiêu dự án	1
1.2.1. Mục tiêu chính	2
1.2.2. Mục tiêu cụ thể	2
1.3. Phạm vi và Đối tượng nghiên cứu	2
1.4. Phương pháp tiếp cận	3
1.5. Cấu trúc báo cáo	3
CHƯƠNG 2: KHÁM PHÁ VÀ TIỀN XỬ LÝ DỮ LIỆU	4
2.1. Giới thiệu tập dữ liệu	4
2.1.1. Nguồn gốc và mô tả	4
2.1.2. Mô tả các biến	4
2.2. Khám phá dữ liệu sơ bộ (Exploratory Data Analysis - EDA)	4
2.2.1. Phân tích thống kê mô tả	4
2.2.2. Kiểm tra kiểu dữ liệu và thông tin chung	5
2.3. Làm sạch dữ liệu (Data Cleaning)	6
2.3.1. Xử lý giá trị thiếu (Missing Values)	6
2.3.2. Phân tích và xử lý giá trị ngoại lai (Outliers)	6
2.3.3. Kiểm tra tính hợp lệ và logic của dữ liệu	7
2.4. Kỹ thuật tạo biến (Feature Engineering)	7
2.4.1. Tạo các biến nhóm (Tuổi, BMI, Glucose)	8
2.4.2. Tạo biến tổng hợp "Điểm nguy cơ"	8
2.5. Kiểm tra chất lượng dữ liệu sau xử lý	9

2.6. Lưu trữ dữ liệu đã xử lý	9
CHƯƠNG 3: PHÂN TÍCH THỐNG KÊ VÀ TRỰC QUAN HÓA.....	10
3.1. Phân tích biến mục tiêu (Stroke)	10
3.1.1. Phân tích sự mất cân bằng của dữ liệu	10
3.2. Phân tích các biến định lượng.....	11
3.2.1. Phân phối và thống kê mô tả (Tuổi, BMI, Mức Glucose).....	11
3.2.2. So sánh giữa hai nhóm có và không có đột quy (T-test)	11
3.3. Phân tích các biến định tính.....	12
3.3.1. Phân phối và tỷ lệ đột quy theo từng nhóm.....	12
3.3.2. Kiểm định mối quan hệ với đột quy (Chi-square test)	13
3.4. Phân tích tương quan (Correlation Analysis).....	13
3.4.1. Ma trận tương quan giữa các biến.....	13
3.4.2. Trực quan hóa bằng Heatmap.....	14
3.5. Phân tích chuyên sâu	15
3.5.1. Phân tích nguy cơ đột quy theo nhóm tuổi.....	15
3.5.2. Phân tích các yếu tố nguy cơ phối hợp và điểm nguy cơ	16
3.6. Tổng kết các phát hiện quan trọng từ phân tích dữ liệu	16
CHƯƠNG 4: XÂY DỰNG VÀ ĐÁNH GIÁ MÔ HÌNH DỰ ĐOÁN.....	17
4.1. Chuẩn bị dữ liệu cho Machine Learning	17
4.1.1. Lựa chọn biến và xác định tập features (X) và target (y).....	17
4.1.2. Phân chia tập dữ liệu (Training và Test set)	17
4.1.3. Xây dựng quy trình tiền xử lý (Preprocessing Pipeline)	18
4.2. Xử lý mất cân bằng dữ liệu trên tập huấn luyện.....	19
4.2.1. So sánh các phương pháp (SMOTE, Over-sampling, Under-.....	19
4.2.2. Lựa chọn phương pháp tối ưu	20
4.3. Xây dựng và so sánh các mô hình Machine Learning	20
4.3.1. Các thuật toán được lựa chọn	20

4.3.2. Các chỉ số đánh giá (Metrics) và ý nghĩa	21
4.3.3. Kết quả so sánh hiệu suất các mô hình.....	22
4.4. Lựa chọn và Tinh chỉnh mô hình hiệu quả nhất	23
4.4.1. Tinh chỉnh siêu tham số (Hyperparameter Tuning).....	23
4.4.2. Đánh giá chi tiết mô hình cuối cùng trên tập kiểm tra (Confusion Matrix, ROC Curve, v.v.)	24
4.5. Diễn giải mô hình (Model Interpretation)	25
4.5.1. Phân tích mức độ quan trọng của các biến (Feature.....	25
4.5.2. Phân tích SHAP để giải thích các dự đoán cụ thể	25
4.6. Lưu trữ mô hình và các thành phần liên quan.....	26
<i>CHƯƠNG 5: KẾT LUẬN VÀ ĐỀ XUẤT</i>	<i>26</i>
5.1. Tóm tắt toàn bộ quá trình và kết quả đạt được	26
5.2. Các phát hiện chính và ý nghĩa thực tiễn.....	27
5.3. Hạn chế của dự án.....	27
5.4. Đề xuất cải thiện và hướng phát triển trong tương lai	28
5.4.1. Về dữ liệu	28
5.4.2. Về mô hình	28
5.4.3. Về phương pháp đánh giá.....	29
5.5. Lộ trình ứng dụng vào thực tế (Roadmap).....	29
5.5.1. Giai đoạn 1: Xây dựng sản phẩm tối thiểu (Proof of	29
5.5.2. Giai đoạn 2: Thử nghiệm và xác thực lâm sàng (Pilot &	29
5.5.3. Giai đoạn 3: Mở rộng và triển khai (Scale-up).....	29

Lời nói đầu

Kính gửi: Cô Phan Thị Thê, Giảng viên môn chuyên đề Tốt Nghiệp 2

Nhóm xin trân trọng gửi đến Cô báo cáo kết quả của dự án phân tích dữ liệu về vấn đề đột quy trong báo cáo y tế. Báo cáo này được thực hiện như một phần của môn học chuyên đề Tốt Nghiệp 2 nhằm mục đích áp dụng các kiến thức đã học vào thực tế, đặc biệt là các kỹ thuật khám phá, làm sạch và tiền xử lý dữ liệu.

Trong bối cảnh dữ liệu ngày càng trở nên quan trọng, việc nắm vững các kỹ năng xử lý dữ liệu thô để biến chúng thành thông tin có giá trị là vô cùng cần thiết. Tập dữ liệu về vấn đề đột quy trong báo cáo y tế đặt ra những thách thức đặc trưng liên quan đến dữ liệu thiếu và các giá trị null, các sơ đồ phân tích

Báo cáo này đi sâu vào quy trình xử lý tập dữ liệu từ bước thu thập ban đầu đến giai đoạn sẵn sàng cho phân tích chuyên sâu hoặc xây dựng mô hình. Chúng em đã thực hiện các bước sau:

1. Khám phá dữ liệu: Hiểu rõ cấu trúc, định dạng và các đặc điểm ban đầu của dữ liệu thông qua thống kê mô tả và trực quan hóa.
2. Làm sạch dữ liệu: Xác định và xử lý các vấn đề về chất lượng dữ liệu như giá trị thiếu và dữ liệu không hợp lý, đảm bảo tính toàn vẹn của dữ liệu.
3. Tiền xử lý dữ liệu: Áp dụng các kỹ thuật biến đổi dữ liệu, bao gồm xử lý ngoại lai (nếu có), tạo các biến mới (feature engineering) và chuẩn bị cho bước mã hóa dữ liệu.

Mục tiêu cuối cùng của quá trình này là có được một tập dữ liệu sạch, chuẩn hóa và phù hợp để sử dụng trong các bước phân tích tiếp theo hoặc phát triển mô hình dự đoán.

Nhóm xin chân thành cảm ơn Cô đã truyền đạt những kiến thức quý báu và tận tình hướng dẫn chúng em trong suốt quá trình học tập môn chuyên đề Tốt Nghiệp 2.

Những kiến thức này là nền tảng vững chắc để chúng em tiếp cận và giải quyết các bài toán thực tế về dữ liệu.

Nhóm rất mong nhận được những góp ý quý báu từ Cô để báo cáo này được hoàn thiện hơn.

CHƯƠNG 1: GIỚI THIỆU TỔNG QUAN

1.1. Bối cảnh và Lý do chọn đề tài

1.1.1. Tầm quan trọng của việc dự đoán sớm đột quỵ

Đột quỵ, hay tai biến mạch máu não, là một trong những nguyên nhân gây tử vong và tàn tật hàng đầu trên toàn thế giới. Theo Tổ chức Y tế Thế giới (WHO), mỗi năm có hàng triệu người bị đột quỵ, và một phần lớn trong số đó phải gánh chịu những di chứng nặng nề, ảnh hưởng nghiêm trọng đến chất lượng cuộc sống của bản thân và gia đình. Gánh nặng kinh tế - xã hội do đột quỵ gây ra là vô cùng lớn, bao gồm chi phí điều trị, chăm sóc dài hạn và mất mát về năng suất lao động.

Trong bối cảnh đó, việc phát hiện và dự đoán sớm nguy cơ đột quỵ đóng một vai trò cực kỳ quan trọng. Nếu các yếu tố nguy cơ có thể được xác định và một cá nhân được cảnh báo sớm về khả năng mắc bệnh, các biện pháp can thiệp y tế và thay đổi lối sống có thể được áp dụng kịp thời. Điều này không chỉ giúp giảm thiểu tỷ lệ mắc bệnh mà còn có thể cứu sống hàng triệu người và giảm bớt gánh nặng cho hệ thống y tế.

1.1.2. Vấn đề nghiên cứu

Sự phát triển của khoa học dữ liệu và học máy đã mở ra những cơ hội mới trong việc phân tích dữ liệu y tế phức tạp. Bằng cách khai thác thông tin từ hồ sơ bệnh án, các đặc điểm nhân khẩu học và lối sống của bệnh nhân, chúng ta có thể xây dựng các mô hình dự đoán có khả năng xác định các cá nhân có nguy cơ cao.

Vấn đề nghiên cứu của dự án này là: **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?**

Dự án sẽ tập trung vào việc phân tích sâu một tập dữ liệu công khai về đột quỵ để trả lời câu hỏi này, từ đó cung cấp những hiểu biết có giá trị và một công cụ hỗ trợ tiềm năng cho ngành y tế.

1.2. Mục tiêu dự án

Dựa trên vấn đề nghiên cứu đã nêu, dự án đặt ra các mục tiêu chính và mục tiêu cụ thể như sau:

1.2.1. Mục tiêu chính

Phân tích và xác định các yếu tố nhân khẩu học, y tế và lối sống có ảnh hưởng mạnh mẽ nhất đến nguy cơ đột quy.

Xây dựng và đánh giá một mô hình học máy có khả năng dự đoán nguy cơ đột quy với độ chính xác và độ tin cậy cao.

Đưa ra các khuyến nghị dựa trên dữ liệu nhằm hỗ trợ việc phòng ngừa và tầm soát sớm đột quy trong cộng đồng.

1.2.2. Mục tiêu cụ thể

Thực hiện khám phá và làm sạch tập dữ liệu healthcare-dataset-stroke-data.

Áp dụng các kỹ thuật trực quan hóa và kiểm định thống kê để tìm ra các mối quan hệ có ý nghĩa giữa các biến và biến mục tiêu (đột quy).

So sánh hiệu suất của nhiều thuật toán học máy khác nhau (ví dụ: Logistic Regression, Random Forest, Gradient Boosting, v.v.) để chọn ra mô hình tốt nhất.

Xử lý vấn đề mất cân bằng dữ liệu, một thách thức phổ biến trong các bài toán y tế.

Đánh giá chi tiết mô hình được chọn bằng các chỉ số phù hợp như F1-Score, ROC-AUC, Sensitivity và Specificity.

Diễn giải mô hình để hiểu rõ cách các yếu tố đầu vào tác động đến kết quả dự đoán.

1.3. Phạm vi và Đối tượng nghiên cứu

Phạm vi dữ liệu: Dự án sử dụng tập dữ liệu "Healthcare Dataset Stroke Data" từ Kaggle, bao gồm 5,110 mẫu bệnh nhân với 11 biến độc lập và 1 biến mục tiêu.

Phạm vi bài toán: Đây là một bài toán phân loại nhị phân (Binary Classification), trong đó mục tiêu là dự đoán một bệnh nhân có bị đột quy (1) hay không (0).

Đối tượng nghiên cứu: Các bệnh nhân trong tập dữ liệu được cung cấp. Kết quả và mô hình của dự án hướng đến việc hỗ trợ các chuyên gia y tế, bác sĩ trong việc chẩn đoán và tư vấn cho bệnh nhân.

1.4. Phương pháp tiếp cận

Dự án được thực hiện theo một quy trình khoa học dữ liệu chuẩn, bao gồm các giai đoạn chính sau:

1. Khám phá dữ liệu (Exploratory Data Analysis - EDA): Tìm hiểu sâu về cấu trúc, phân phối và các đặc điểm của dữ liệu.
2. Tiền xử lý dữ liệu và Kỹ thuật tạo biến (Data Preprocessing & Feature Engineering): Làm sạch dữ liệu, xử lý các giá trị thiếu, và tạo ra các biến mới có ý nghĩa hơn.
3. Phát triển mô hình học máy (Machine Learning Model Development): Huấn luyện nhiều mô hình khác nhau trên dữ liệu đã xử lý.
4. Đánh giá và Lựa chọn mô hình (Model Evaluation & Selection): Sử dụng các chỉ số đánh giá phù hợp để chọn ra mô hình có hiệu suất tốt nhất.
5. Tổng hợp kết quả và Đề xuất (Insights & Recommendations): Rút ra các kết luận quan trọng từ phân tích và đề xuất các hướng hành động thực tế.

1.5. Cấu trúc báo cáo

Báo cáo được tổ chức thành 5 chương chính:

Chương 1 - Giới thiệu tổng quan: Trình bày bối cảnh, mục tiêu, phạm vi và phương pháp luận của dự án.

Chương 2 - Khám phá và Tiền xử lý dữ liệu: Mô tả chi tiết về tập dữ liệu và các bước làm sạch, chuẩn bị dữ liệu.

Chương 3 - Phân tích Thống kê và Trực quan hóa: Đi sâu vào phân tích các mối quan hệ trong dữ liệu để tìm ra các yếu tố nguy cơ.

Chương 4 - Xây dựng và Đánh giá mô hình dự đoán: Trình bày quá trình xây dựng, so sánh và lựa chọn mô hình học máy.

Chương 5 - Kết luận và Đề xuất: Tóm tắt các kết quả chính, nêu bật các hạn chế và đề xuất các hướng phát triển trong tương lai.

CHƯƠNG 2: KHÁM PHÁ VÀ TIỀN XỬ LÝ DỮ LIỆU

Chương này trình bày chi tiết quá trình làm việc với tập dữ liệu gốc, từ việc tìm hiểu, khám phá sơ bộ cho đến các bước làm sạch và chuẩn bị dữ liệu. Mục tiêu của chương này là tạo ra một tập dữ liệu chất lượng, sẵn sàng cho việc phân tích thống kê và xây dựng mô hình ở các chương sau.

2.1. Giới thiệu tập dữ liệu

2.1.1. Nguồn gốc và mô tả

Dự án sử dụng tập dữ liệu "Healthcare Dataset Stroke Data" được công bố trên nền tảng Kaggle. Đây là một tập dữ liệu phổ biến trong cộng đồng khoa học dữ liệu, thường được dùng cho các bài toán dự đoán nguy cơ đột quỵ dựa trên các thông tin về nhân khẩu học và sức khỏe.

Nguồn:[Kaggle:StrokePredictionDataset](<https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset>)

Kích thước ban đầu: Tập dữ liệu gồm 5,110 mẫu (bệnh nhân) và 12 thuộc tính (cột).

2.1.2. Mô tả các biến

Mỗi dòng trong tập dữ liệu đại diện cho một bệnh nhân và bao gồm các thông tin sau:

Tên biến	Kiểu dữ liệu	Mô tả
`id`	Số (int)	Mã định danh duy nhất của bệnh nhân.
`gender`	Chữ (object)	Giới tính của bệnh nhân (Male, Female, hoặc Other).
`age`	Số (float)	Tuổi của bệnh nhân.
`hypertension`	Nhị phân (int)	Có tiền sử tăng huyết áp hay không (1: Có, 0: Không).
`heart_disease`	Nhị phân (int)	Có tiền sử bệnh tim hay không (1: Có, 0: Không).
`ever_married`	Chữ (object)	Đã từng kết hôn hay chưa (Yes, No).
`work_type`	Chữ (object)	Loại hình công việc (Private, Self-employed, Govt_job, children, Never_worked).
`Residence_type`	Chữ (object)	Khu vực sinh sống (Urban: Thành thị, Rural: Nông thôn).
`avg_glucose_level`	Số (float)	Mức đường huyết trung bình trong máu.
`bmi`	Số (float)	Chỉ số khối cơ thể (Body Mass Index).
`smoking_status`	Chữ (object)	Tình trạng hút thuốc (formerly smoked, never smoked, smokes, Unknown).
`stroke`	Nhị phân (int)	Biến mục tiêu: Bệnh nhân có bị đột quỵ hay không (1: Có, 0: Không).

2.2. Khám phá dữ liệu sơ bộ (Exploratory Data Analysis - EDA)

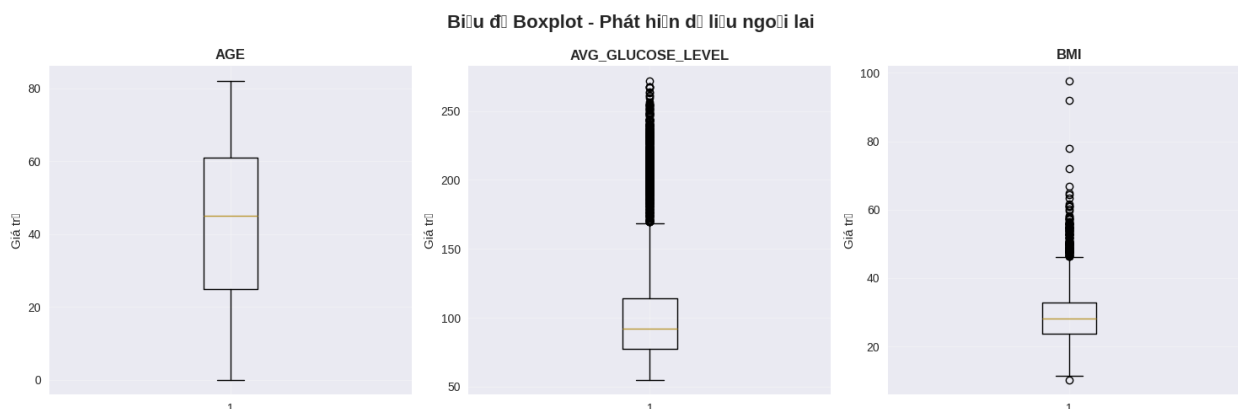
2.2.1. Phân tích thống kê mô tả

Một cái nhìn tổng quan về các biến số trong dữ liệu cho thấy:

Tuổi (age): Độ tuổi của các bệnh nhân trong mẫu dao động từ trẻ sơ sinh (0.08 tuổi) đến người cao tuổi (82 tuổi), với độ tuổi trung bình là 43.2.

Mức đường huyết (**avg_glucose_level**): Có sự biến thiên lớn, từ 55.12 đến 271.74, cho thấy sự đa dạng về tình trạng sức khỏe của các bệnh nhân.

Chỉ số BMI (**bmi**): Chỉ số BMI trung bình là 28.9, nhưng cũng có sự dao động đáng kể.



2.2.2. Kiểm tra kiểu dữ liệu và thông tin chung

Kiểm tra ban đầu cho thấy các kiểu dữ liệu đã phù hợp với mô tả. Tuy nhiên, một phát hiện quan trọng là sự xuất hiện của các giá trị thiếu:

Cột bmi có 201 giá trị bị thiếu, chiếm khoảng 3.9% tổng số dữ liệu. Đây là vấn đề cần được xử lý trong bước làm sạch.

Các cột còn lại đều có đủ 5,110 giá trị.

BẮT ĐẦU QUÁ TRÌNH LÀM SẠCH DỮ LIỆU

=====

1 Xử lý cột BMI:

Số lượng giá trị NaN trong BMI sau xử lý: 201

Đã thay thế giá trị thiếu bằng median: 28.10

Hoàn thành xử lý cột BMI

2.3. Làm sạch dữ liệu (Data Cleaning)

2.3.1. Xử lý giá trị thiếu (Missing Values)

PHÂN TÍCH GIÁ TRỊ THIẾU:

```
=====
      soLuongThieu  tyLeThieu
bmi              201        3.93
```

Như đã xác định, chỉ có cột `bmi` chứa giá trị thiếu. Dựa trên phân tích phân phối của biến `bmi`, phương pháp thay thế các giá trị thiếu bằng giá trị trung vị (median) đã được lựa chọn. Lý do là vì phân phối của `bmi` có xu hướng bị lệch, và giá trị trung vị ít bị ảnh hưởng bởi các giá trị ngoại lai hơn so với giá trị trung bình, do đó đây là một lựa chọn thay thế hợp lý và an toàn.

Sau khi xử lý, tập dữ liệu không còn giá trị thiếu.

KIỂM TRA GIÁ TRỊ 'N/A' TRONG CỘT BMI:

```
=====
Số lượng giá trị 'N/A' trong BMI: 0
Tỷ lệ giá trị 'N/A': 0.00%
```

2.3.2. Phân tích và xử lý giá trị ngoại lai (Outliers)

Phân tích ngoại lai được thực hiện cho các biến số `age`, `avg_glucose_level`, và `bmi` bằng phương pháp IQR (Interquartile Range). Các biểu đồ boxplot cho thấy sự tồn tại của các giá trị ngoại lai, đặc biệt ở `avg_glucose_level`. Tuy nhiên, trong bối cảnh y tế, các giá trị này có thể là những trường hợp bệnh lý thực tế và chứa thông tin quan trọng. Do đó, dự án quyết định không loại bỏ các giá trị ngoại lai mà giữ lại để mô hình có thể học được từ những trường hợp đa dạng này.

2 Phân tích dữ liệu ngoại lai:

Biến 'age':

- Q1: 25.00, Q3: 61.00, IQR: 36.00
- Ngưỡng: [-29.00, 115.00]
- Số lượng ngoại lai: 0 (0.00%)

Biến 'avg_glucose_level':

- Q1: 77.25, Q3: 114.09, IQR: 36.84
- Ngưỡng: [21.98, 169.36]
- Số lượng ngoại lai: 627 (12.27%)
- Giá trị min ngoại lai: 169.43
- Giá trị max ngoại lai: 271.74

Biến 'bmi':

- Q1: 23.80, Q3: 32.80, IQR: 9.00
- Ngưỡng: [10.30, 46.30]
- Số lượng ngoại lai: 126 (2.47%)
- Giá trị min ngoại lai: 10.30
- Giá trị max ngoại lai: 97.60

2.3.3. Kiểm tra tính hợp lệ và logic của dữ liệu

3 Kiểm tra tính hợp lý của dữ liệu:

Số mẫu có tuổi không hợp lý (< 0 hoặc > 120): 0
Số mẫu có BMI không hợp lý (< 10 hoặc > 60): 13
Số mẫu có glucose không hợp lý (< 50 hoặc > 500): 0
Số trẻ em (< 16 tuổi) có công việc người lớn: 60
Số trẻ em (< 16 tuổi) đã kết hôn: 0

Cột 'id': Cột này chỉ là mã định danh, không mang giá trị phân tích nên đã được loại bỏ.

Cột 'gender': Có một giá trị "Other". Vì số lượng quá ít (chỉ 1 trường hợp), không đủ để đưa ra kết luận thống kê, nên dòng dữ liệu này đã được loại bỏ để đảm bảo tính nhất quán.

Kiểm tra trùng lặp: Không có dòng dữ liệu nào bị trùng lặp hoàn toàn.

2.4. Kỹ thuật tạo biến (Feature Engineering)

Để làm giàu thông tin cho mô hình, một số biến mới đã được tạo ra từ các biến hiện có.

2.4.1. Tạo các biến nhóm (Tuổi, BMI, Glucose)

nhomTuoi: Bệnh nhân được phân vào các nhóm tuổi khác nhau (Vị thành niên, Thanh niên, Trung niên, Cao niên) để mô hình có thể học được các ngưỡng nguy cơ theo độ tuổi.

nhomBMI: Phân loại tình trạng cơ thể dựa trên chỉ số BMI theo chuẩn của WHO (Thiếu cân, Bình thường, Thừa cân, Béo phì).

nhomGlucose: Phân loại mức đường huyết (Bình thường, Tiền tiểu đường, Tiểu đường) để làm nổi bật các mức độ nguy cơ khác nhau.

THỐNG KÊ CÁC BIẾN MỚI:

=====

Biến 'nhomTuoi':

```
-----
              soLuong  tyLePhanTram
nhomTuoi
trungNien      1385         27.10
truocTuoiHuu   1183         23.15
tuoiCao        1027         20.10
treEm          856          16.75
tuoiTre         659         12.90
```

Biến 'nhomBMI':

```
-----
              soLuong  tyLePhanTram
nhomBMI
beoPhi         1920         37.57
thua Can       1610         31.51
binhThuong     1243         24.32
thieuCan        337          6.59
```

Biến 'nhomGlucose':

```
-----
...
50%          1.000000
75%          1.000000
max          5.000000
Name: diemNguyCo, dtype: float64
```

2.4.2. Tạo biến tổng hợp "Điểm nguy cơ"

Một biến mới là **diemNguyCo** được tạo ra bằng cách cộng điểm từ các yếu tố nguy cơ đã biết như hypertension, heart_disease, và các nhóm **nhomBMI**, **nhomGlucose** vừa tạo. Biến này giúp tổng hợp nhiều yếu tố rủi ro vào một chỉ số duy nhất, có khả năng cung cấp một tín hiệu mạnh mẽ hơn cho mô hình

TẠO CÁC BIẾN MỚI:

```
=====
Đã tạo biến 'nhomTuoi'
Đã tạo biến 'nhomBMI'
Đã tạo biến 'nhomGlucose'
Đã tạo biến 'diemNguyCo'
```

2.5. Kiểm tra chất lượng dữ liệu sau xử lý

Sau tất cả các bước làm sạch và tạo biến, tập dữ liệu cuối cùng có 5,109 mẫu và 15 cột. Dữ liệu đã sạch, không còn giá trị thiếu hay các giá trị không hợp lệ.

Một phân tích quan trọng về biến mục tiêu stroke cho thấy một vấn đề nổi bật: sự mất cân bằng dữ liệu nghiêm trọng. Chỉ có 249 trường hợp đột quỵ (4.9%) so với 4,860 trường hợp không đột quỵ (95.1%). Đây là một thách thức lớn cần được giải quyết trong giai đoạn xây dựng mô hình để tránh việc mô hình bị thiên vị, chỉ dự đoán "không đột quỵ".

KIỂM TRA CHẤT LƯỢNG DỮ LIỆU SAU XỬ LÝ:

```
=====
Kích thước dữ liệu sau xử lý: (5110, 16)
Số dòng: 5,110
Số cột: 16
```

Tổng số giá trị thiếu: 0
Số dòng trùng lặp: 0

Phân phối biến mục tiêu 'stroke':

	soLuong	tyLePhanTram
Không đột quỵ	4861	95.13
Có đột quỵ	249	4.87

2.6. Lưu trữ dữ liệu đã xử lý

Để thuận tiện cho các bước tiếp theo, tập dữ liệu đã qua xử lý được lưu lại thành một file CSV mới có tên là `du_lieu_da_xu_ly.csv`. File này sẽ là đầu vào cho quá trình phân tích thống kê và huấn luyện mô hình.

Đã lưu dữ liệu đã xử lý vào: [/content/drive/MyDrive/PHÂN TÍCH DỮ LIỆU /du lieu da xu ly.csv](#)

TÓM TẮT QUÁ TRÌNH XỬ LÝ DỮ LIỆU

Dữ liệu gốc: 5,110 dòng, 12 cột
Dữ liệu sau xử lý: 5,110 dòng, 16 cột
Đã xử lý 201 giá trị thiếu trong cột BMI
Đã tạo 4 biến mới
Dữ liệu đã sẵn sàng cho các bước phân tích tiếp theo

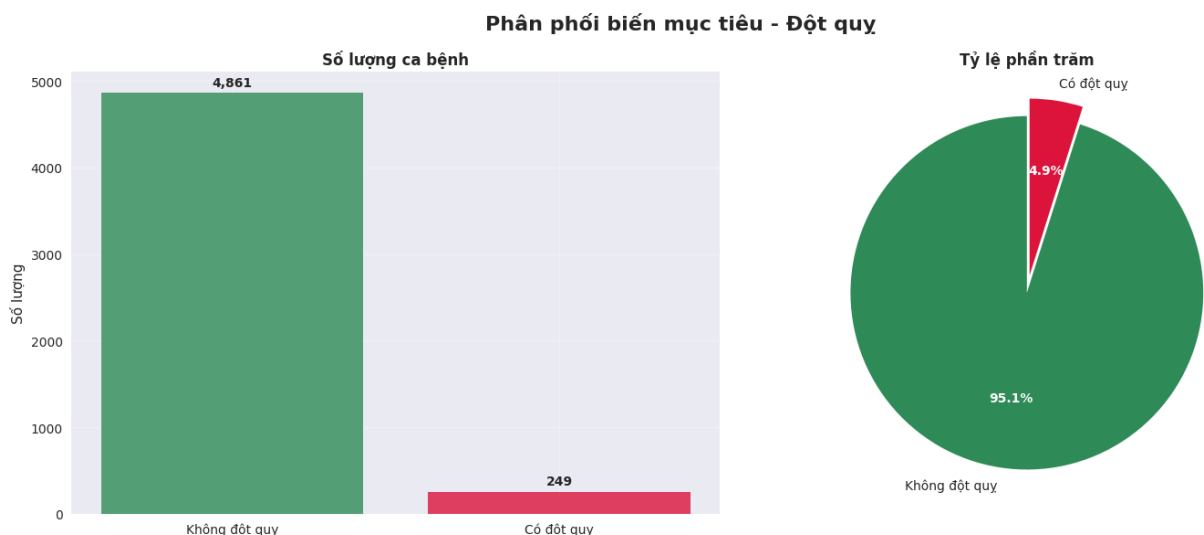
CHƯƠNG 3: PHÂN TÍCH THỐNG KÊ VÀ TRỰC QUAN HÓA

Sau khi dữ liệu đã được làm sạch và chuẩn bị ở Chương 2, chương này sẽ đi sâu vào việc phân tích và trực quan hóa để tìm ra những hiểu biết (insights) quan trọng. Mục tiêu là xác định các yếu tố có mối liên hệ mạnh mẽ với nguy cơ đột quỵ, làm nền tảng cho việc lựa chọn biến và xây dựng mô hình ở chương sau.

3.1. Phân tích biến mục tiêu (Stroke)

3.1.1. Phân tích sự mất cân bằng của dữ liệu

Như đã đề cập ở chương trước, biến mục tiêu 'stroke' bị mất cân bằng nghiêm trọng. Phân tích trên tập dữ liệu đã xử lý (5,109 mẫu) cho thấy:



Trực quan hóa bằng biểu đồ thanh cho thấy sự chênh lệch rõ rệt này. Tình trạng mất cân bằng này là một thách thức lớn, đòi hỏi các kỹ thuật xử lý đặc biệt trong giai đoạn xây

dựng mô hình để đảm bảo mô hình không bị thiên vị và có khả năng nhận diện đúng lớp thiểu số (có đột quy).

3.2. Phân tích các biến định lượng

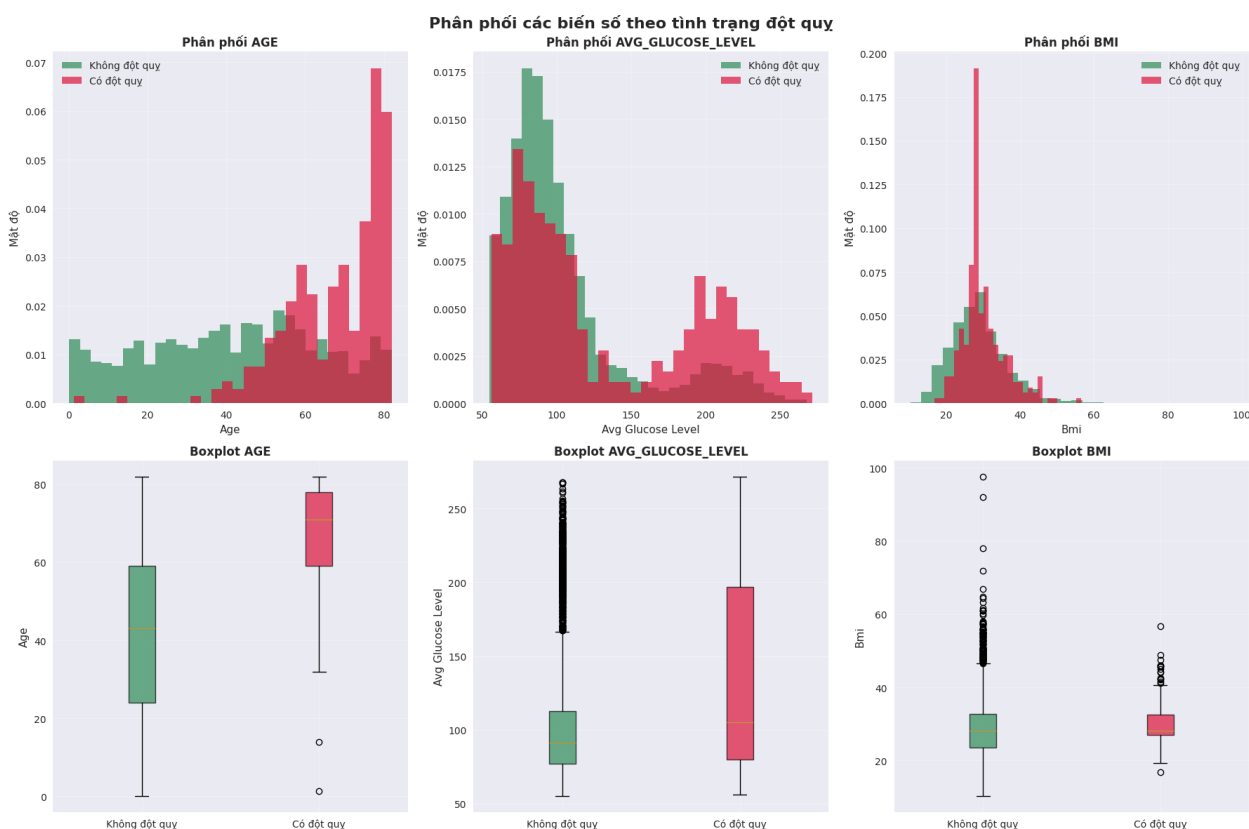
3.2.1. Phân phối và thống kê mô tả (Tuổi, BMI, Mức Glucose)

Phân tích phân phối của các biến `age`, `avg_glucose_level`, và `bmi` cho thấy:

Tuổi (`age`): Phân phối của tuổi khá đồng đều, nhưng khi so sánh giữa hai nhóm, có thể thấy rõ **nhóm bệnh nhân bị đột quy có độ tuổi trung bình cao hơn đáng kể** so với nhóm không bị đột quy. Điều này cho thấy tuổi tác là một yếu tố nguy cơ quan trọng.

Mức Glucose (`avg_glucose_level`): Phân phối của biến này bị lệch phải. Nhóm bị đột quy có xu hướng có mức đường huyết trung bình cao hơn.

BMI (`bmi`): Phân phối của BMI cũng bị lệch phải. Mặc dù không rõ ràng như tuổi và mức glucose, nhóm bị đột quy dường như cũng có chỉ số BMI cao hơn một chút.



3.2.2. So sánh giữa hai nhóm có và không có đột quy (T-test)

Để kiểm định sự khác biệt về giá trị trung bình của các biến định lượng giữa hai nhóm, kiểm định T (T-test) độc lập đã được thực hiện:

Tuổi (`age`): Giá trị p-value rất nhỏ (gần bằng 0), cho thấy có sự khác biệt rất có ý nghĩa thống kê về độ tuổi trung bình giữa nhóm bị đột quỵ và không bị đột quỵ.

Mức Glucose (`avg_glucose_level`): Giá trị p-value cũng rất nhỏ, khẳng định rằng mức đường huyết trung bình ở nhóm bị đột quỵ cao hơn một cách có ý nghĩa thống kê.

BMI (`bmi`): Giá trị p-value lớn hơn so với hai biến trên nhưng vẫn đủ nhỏ để kết luận có sự khác biệt có ý nghĩa thống kê về BMI giữa hai nhóm.

Kết luận: Cả ba biến định lượng đều cho thấy sự khác biệt có ý nghĩa thống kê giữa nhóm có và không có đột quỵ, trong đó `age` và `avg_glucose_level` là hai yếu tố có sự khác biệt rõ rệt nhất.

3.3. Phân tích các biến định tính

3.3.1. Phân phối và tỷ lệ đột quỵ theo từng nhóm

Trực quan hóa tỷ lệ đột quỵ trong các nhóm của từng biến định tính mang lại nhiều thông tin giá trị:

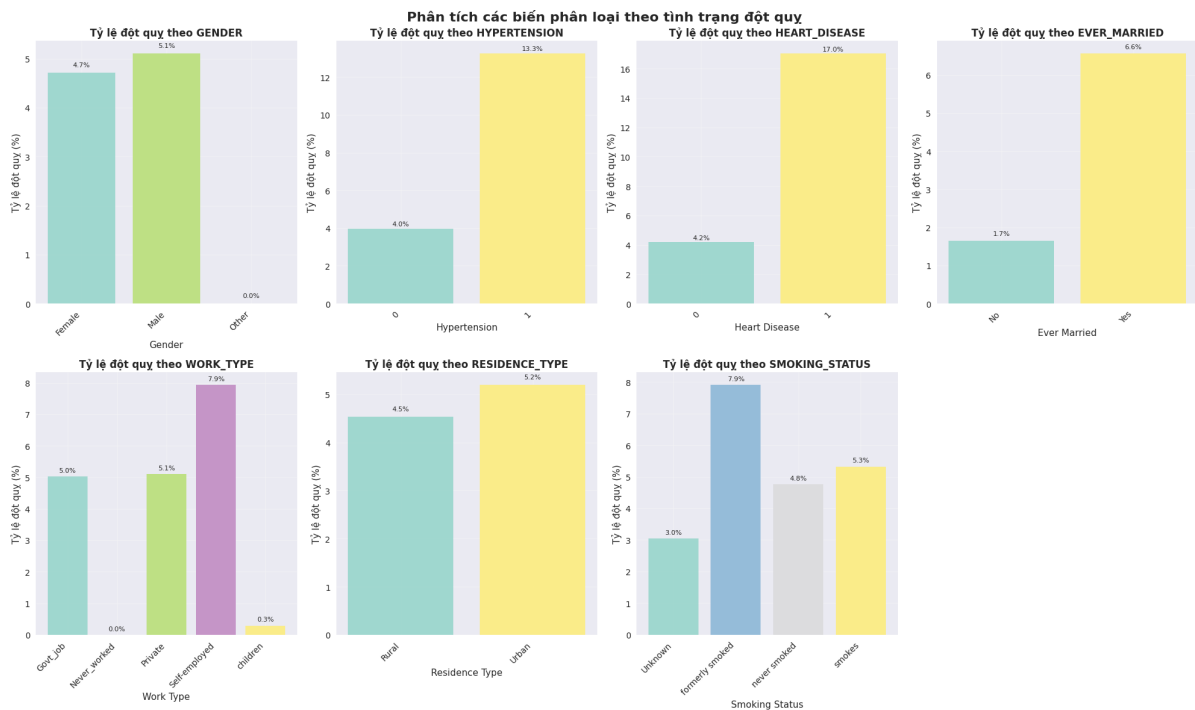
Tăng huyết áp (`hypertension`): Tỷ lệ đột quỵ ở nhóm có tiền sử tăng huyết áp (khoảng 13%) cao hơn hẳn so với nhóm không có (khoảng 4%).

Bệnh tim (`heart_disease`): Tương tự, nhóm có tiền sử bệnh tim có tỷ lệ đột quỵ cao vượt trội (khoảng 17%) so với nhóm không có (khoảng 4%).

Tình trạng hôn nhân (`ever_married`): Những người đã từng kết hôn có tỷ lệ đột quỵ cao hơn.

Loại hình công việc (`work_type`): Nhóm người tự kinh doanh (`Self-employed`) có tỷ lệ đột quỵ cao hơn các nhóm khác.

Tình trạng hút thuốc (`smoking_status`): Nhóm "formerly smoked" (đã từng hút) và "smokes" (đang hút) có tỷ lệ đột quỵ cao hơn nhóm "never smoked" (chưa bao giờ hút)



3.3.2. Kiểm định mối quan hệ với đột quỵ (Chi-square test)

Kiểm định Chi-bình phương (Chi-square test) được sử dụng để xác định mối liên hệ thống kê giữa các biến định tính và biến mục tiêu 'stroke'.

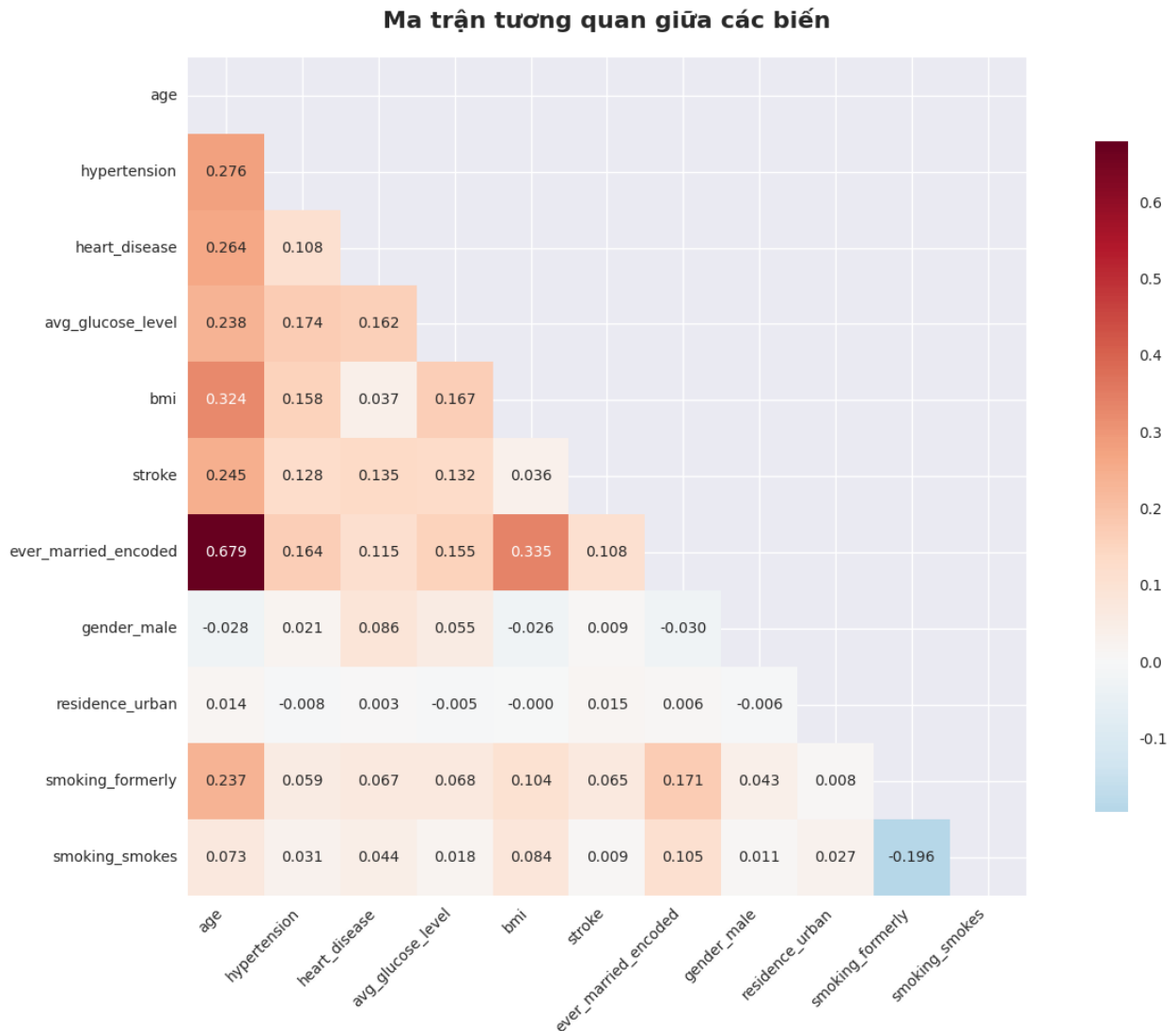
Kết quả: Tất cả các biến định tính được kiểm định ('gender', 'hypertension', 'heart_disease', 'ever_married', 'work_type', 'Residence_type', 'smoking_status') đều có giá trị p-value rất nhỏ. Điều này khẳng định rằng tất cả các biến này đều có mối liên hệ có ý nghĩa thống kê với nguy cơ đột quỵ.

3.4. Phân tích tương quan (Correlation Analysis)

3.4.1. Ma trận tương quan giữa các biến

Phân tích tương quan được thực hiện giữa các biến số (bao gồm cả các biến nhị phân như 'hypertension', 'heart_disease').

3.4.2. Trực quan hóa bằng Heatmap



Ma trận tương quan được trực quan hóa bằng heatmap. Các cặp biến có tương quan đáng chú ý nhất bao gồm:

`age` và `ever_married`: Tương quan dương mạnh, điều này hợp lý vì người lớn tuổi thường đã kết hôn.

`age` và `stroke`: Tương quan dương, xác nhận lại tuổi là yếu tố nguy cơ.

`hypertension` và `age`: Tương quan dương, người lớn tuổi có xu hướng bị tăng huyết áp nhiều hơn.

`avg_glucose_level` và `stroke`: Tương quan dương.

Nhìn chung, không có hiện tượng đa cộng tuyến (multicollinearity) quá nghiêm trọng giữa các biến độc lập, điều này là một tín hiệu tốt cho việc xây dựng mô hình.

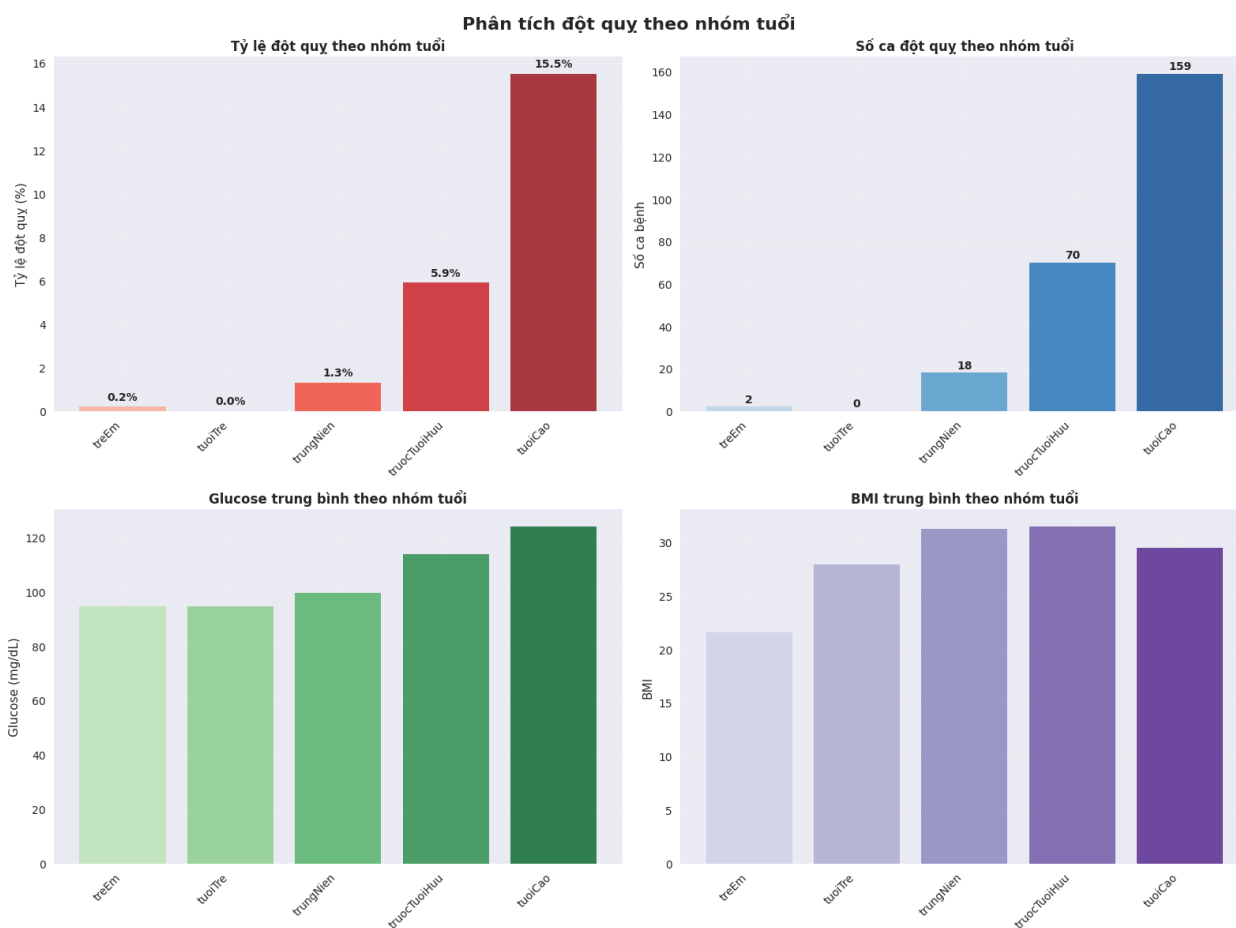
🔍 PHÂN TÍCH CÁC TƯƠNG QUAN MẠNH VỚI ĐỘT QUỴ:

- 📈 age: 0.2453 (Thuận, Trung bình)
- 📈 heart_disease: 0.1349 (Thuận, Trung bình)
- 📈 avg_glucose_level: 0.1319 (Thuận, Trung bình)
- 📈 hypertension: 0.1279 (Thuận, Trung bình)
- 📈 ever_married_encoded: 0.1083 (Thuận, Trung bình)

3.5. Phân tích chuyên sâu

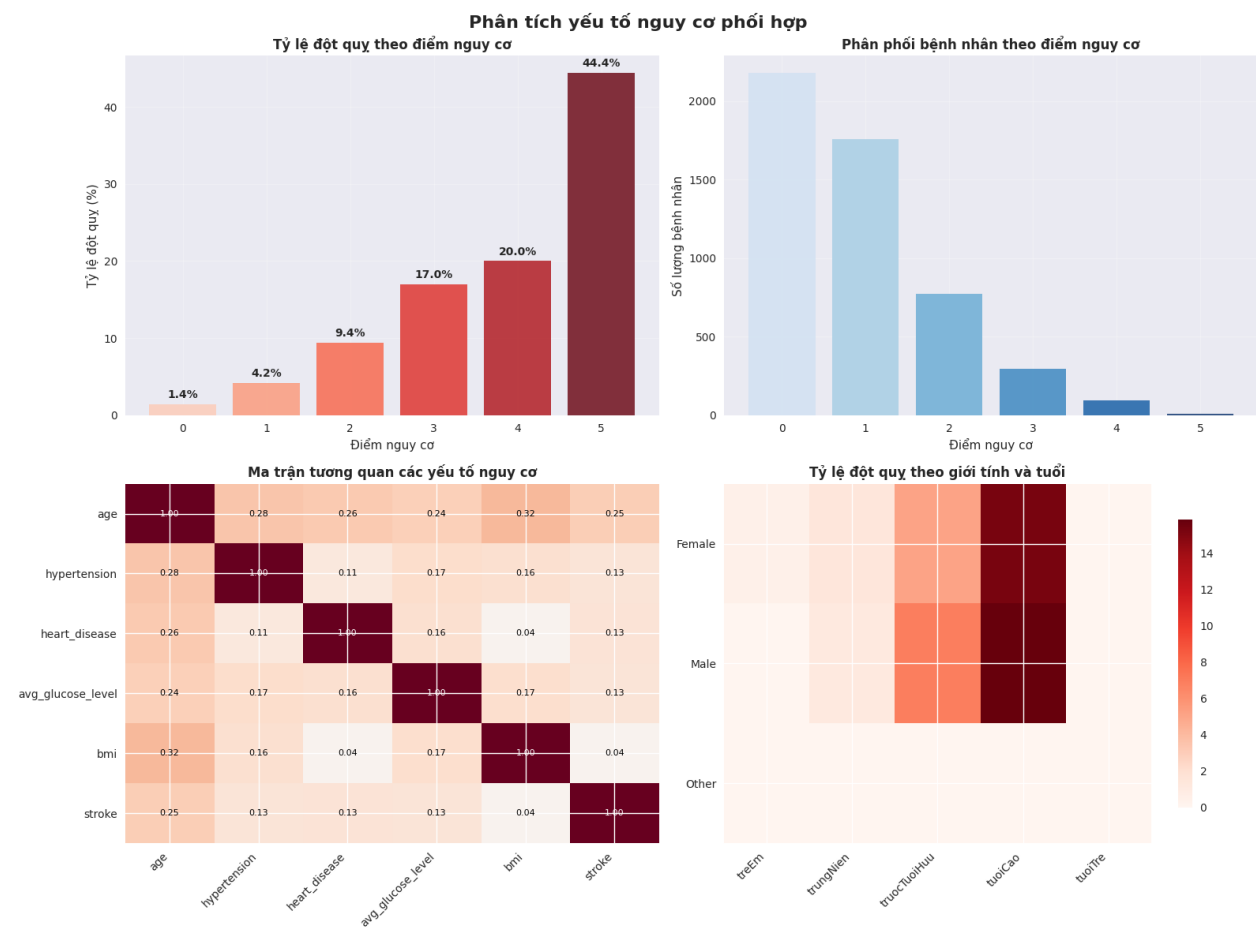
3.5.1. Phân tích nguy cơ đột quỵ theo nhóm tuổi

Khi kết hợp phân tích `age` và `stroke`, biểu đồ phân phối cho thấy rõ ràng rằng tần suất đột quỵ tăng mạnh ở các nhóm tuổi cao, đặc biệt là từ 60 tuổi trở lên. Điều này nhấn mạnh tuổi tác là một trong những yếu tố dự báo quan trọng nhất.



3.5.2. Phân tích các yếu tố nguy cơ phối hợp và điểm nguy cơ

Phân tích biến `diemNguyCo` (tạo ở Chương 2) cho thấy những bệnh nhân có điểm nguy cơ cao hơn cũng có tỷ lệ đột quỵ cao hơn một cách rõ rệt. Điều này chứng tỏ việc kết hợp các yếu tố rủi ro vào một biến tổng hợp là một kỹ thuật tạo biến hiệu quả, giúp nắm bắt tốt hơn nguy cơ tổng thể của một bệnh nhân



3.6. Tổng kết các phát hiện quan trọng từ phân tích dữ liệu

Qua quá trình phân tích, các yếu tố sau đây được xác định là có mối liên hệ mạnh mẽ và có ý nghĩa thống kê với nguy cơ đột quỵ:

1. Các yếu tố nguy cơ hàng đầu:

Tuổi (`age`): Yếu tố có ảnh hưởng rõ rệt nhất, tuổi càng cao nguy cơ càng lớn.

Tiền sử bệnh lý: `hypertension` (tăng huyết áp) và `heart_disease` (bệnh tim) làm tăng đáng kể nguy cơ đột quỵ.

Mức đường huyết (`avg_glucose_level`): Mức đường huyết cao là một chỉ báo nguy cơ quan trọng.

2. Các yếu tố ảnh hưởng khác:

`ever_married`, `work_type`, `smoking_status`, và `bmi` cũng cho thấy mối liên hệ có ý nghĩa, mặc dù mức độ ảnh hưởng có thể không mạnh bằng các yếu tố trên.

3. Vấn đề cần lưu ý cho mô hình hóa:

Sự mất cân bằng nghiêm trọng của biến mục tiêu `stroke` là thách thức chính cần được giải quyết.

Những phát hiện này sẽ là kim chỉ nam cho việc lựa chọn các biến đầu vào và thiết kế các chiến lược mô hình hóa trong chương tiếp theo.

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

4.1. Chuẩn bị dữ liệu cho Machine Learning

4.1.1. Lựa chọn biến và xác định tập features (X) và target (y)

Tập Features (X): Bao gồm tất cả các biến độc lập trong tập dữ liệu đã qua xử lý (`du_lieu_da_xu_ly.csv`), ngoại trừ các biến nhóm đã được tạo (`nhomTuoi`, `nhomBMI`, `nhomGlucose`) để tránh đa cộng tuyến và dư thừa thông tin.

Biến Mục tiêu (y): Là cột `stroke`, với giá trị 1 (có đột quỵ) và 0 (không đột quỵ).

✅ Đã tải dữ liệu từ file đã xử lý

📊 Kích thước dữ liệu: (5110, 16)

🎯 Phân phối target:

stroke

0 4861

1 249

Name: count, dtype: int64

📊 Tỷ lệ mất cân bằng: 19.5:1


4.1.2. Phân chia tập dữ liệu (Training và Test set)





Tập dữ liệu được phân chia thành hai tập:


Tập huấn luyện (Training set): Chiếm 80% dữ liệu, được sử dụng để huấn luyện các mô hình.

Tập kiểm tra (Test set): Chiếm 20% dữ liệu còn lại, được giữ riêng và chỉ sử dụng một lần duy nhất để đánh giá hiệu suất cuối cùng của mô hình tốt nhất. Điều này đảm bảo kết quả đánh giá là khách quan.

Việc phân chia được thực hiện theo phương pháp `stratified sampling` (lấy mẫu phân tầng) dựa trên biến mục tiêu `y` để đảm bảo tỷ lệ các lớp (có/không đột quỵ) trong cả hai tập huấn luyện và kiểm tra là tương đương nhau.

 CHIA DỮ LIỆU TRAIN/TEST:

 Training set: 4088 samples
 Test set: 1022 samples
 Train target distribution: {0: 3889, 1: 199}
 Test target distribution: {0: 972, 1: 50}

 Sau preprocessing:


- Training features shape: (4088, 26)
- Test features shape: (1022, 26)
- Tổng số features sau xử lý: 26

4.1.3. Xây dựng quy trình tiền xử lý (Preprocessing Pipeline)




Để đảm bảo các bước tiền xử lý được áp dụng một cách nhất quán và tự động cho cả dữ liệu huấn luyện và kiểm tra, một `Pipeline` của Scikit-learn đã được xây dựng. Quy trình này bao gồm hai bước chính:




1. Mã hóa biến định tính (Categorical Encoding): Các biến có kiểu dữ liệu `object` (ví dụ: `gender`, `work_type`) được chuyển đổi thành dạng số bằng phương pháp One-Hot Encoding. Phương pháp này tạo ra các cột nhị phân mới cho mỗi giá trị của biến, phù hợp cho các thuật toán tuyến tính và dựa trên cây.

2. Co giãn dữ liệu (Data Scaling): Các biến số (ví dụ: `age`, `avg_glucose_level`, `bmi`) được chuẩn hóa bằng StandardScaler. Kỹ thuật này biến đổi dữ liệu sao cho có giá trị trung bình bằng 0 và độ lệch chuẩn bằng 1, giúp các thuật toán nhạy cảm với thang đo (như Logistic Regression, SVM) hoạt động hiệu quả hơn.

 CHUẨN BỊ DỮ LIỆU CHO MACHINE LEARNING:

=====

 Số lượng features: 14
 Số lượng samples: 5110
 Target distribution: {0: 4861, 1: 249}

 Biến số (4): ['age', 'avg_glucose_level', 'bmi', 'diemNguyCo']
 Biến phân loại (8): ['gender', 'ever_married', 'work_type', 'Residence_type', 'smoking_status', 'nhomTuoi', 'nhomBMI', 'nhomGlucose']
 Biến binary (2): ['hypertension', 'heart_disease']

4.2. Xử lý mất cân bằng dữ liệu trên tập huấn luyện

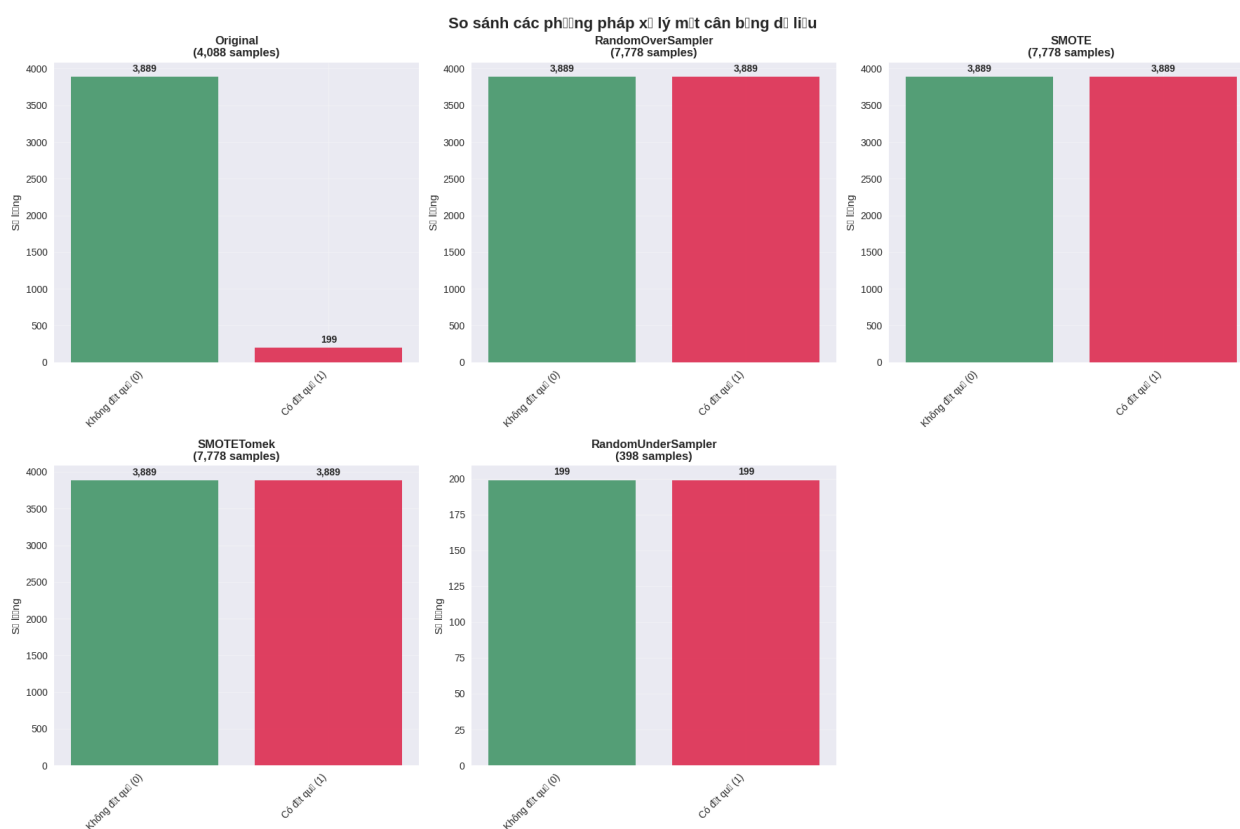
4.2.1. So sánh các phương pháp (SMOTE, Over-sampling, Under-sampling)

Như đã phân tích, sự mất cân bằng dữ liệu là một thách thức lớn. Để giải quyết vấn đề này, kỹ thuật SMOTE (Synthetic Minority Over-sampling Technique) đã được áp dụng.

Phương pháp: SMOTE hoạt động bằng cách tạo ra các mẫu tổng hợp mới cho lớp thiểu số (lớp "có đột quy") dựa trên các mẫu hiện có. Các mẫu mới này được tạo ra trên không gian đặc trưng, nằm giữa các mẫu thiểu số gần nhau.

Phạm vi áp dụng: Kỹ thuật này chỉ được áp dụng trên tập huấn luyện (training set). Việc này rất quan trọng để tránh rò rỉ dữ liệu (data leakage), vì tập kiểm tra phải phản ánh đúng phân phối dữ liệu trong thực tế.

Sau khi áp dụng SMOTE, số lượng mẫu của lớp thiểu số và lớp đa số trong tập huấn luyện đã trở nên cân bằng.



4.2.2. Lựa chọn phương pháp tối ưu

4.3. Xây dựng và so sánh các mô hình Machine Learning

4.3.1. Các thuật toán được lựa chọn

Nhiều thuật toán học máy phổ biến cho bài toán phân loại đã được lựa chọn để huấn luyện và so sánh, bao gồm:

Logistic Regression: Một mô hình tuyến tính cơ bản, nhanh và dễ diễn giải.

Decision Tree: Một mô hình dựa trên cây quyết định, dễ hiểu nhưng dễ bị quá khớp (overfitting).

Random Forest: Một thuật toán học tập quần thể (ensemble) kết hợp nhiều cây quyết định, thường cho hiệu suất cao và ổn định.

Gradient Boosting: Một thuật toán ensemble mạnh mẽ khác, xây dựng các cây một cách tuần tự để sửa lỗi của các cây trước đó.

XGBoost (Extreme Gradient Boosting): Một phiên bản tối ưu và hiệu quả của Gradient Boosting, rất phổ biến trong các cuộc thi Kaggle.

LightGBM (Light Gradient Boosting Machine): Một phiên bản khác của Gradient Boosting, được tối ưu hóa về tốc độ và hiệu quả sử dụng bộ nhớ.

Support Vector Machine (SVM): Một thuật toán mạnh mẽ tìm ranh giới quyết định tối ưu giữa các lớp.

🧠 ĐỊNH NGHĨA CÁC MÔ HÌNH MACHINE LEARNING:

=====

📊 Số lượng mô hình: 10

1. Logistic Regression
2. Random Forest
3. Gradient Boosting
4. SVM
5. KNN
6. Naive Bayes
7. Decision Tree
8. AdaBoost
9. XGBoost
10. LightGBM

4.3.2. Các chỉ số đánh giá (Metrics) và ý nghĩa

Do sự mất cân bằng dữ liệu, chỉ sử dụng độ chính xác (Accuracy) là không đủ. Các chỉ số sau được ưu tiên sử dụng:

F1-Score: Là trung bình điều hòa của Precision và Recall, đây là chỉ số quan trọng nhất cho bài toán này vì nó cân bằng giữa việc dự đoán đúng các ca đột quỵ (Recall) và việc không dự đoán nhầm các ca không đột quỵ thành có đột quỵ (Precision).

Recall (Sensitivity): Tỷ lệ các ca đột quỵ thực tế được mô hình dự đoán đúng. Chỉ số này rất quan trọng trong y tế vì bỏ sót một ca bệnh còn nguy hiểm hơn là chẩn đoán nhầm.

Precision: Tỷ lệ các ca được dự đoán là đột quỵ thực sự bị đột quỵ.

ROC-AUC Score: Đo lường khả năng của mô hình trong việc phân biệt giữa hai lớp. Giá trị càng gần 1 càng tốt.

Thông thường, độ chính xác của mô hình nên duy trì trong khoảng **0.8 đến 0.9** để đảm bảo khả năng **tổng quát hóa** tốt và **hiệu quả thực tiễn**. Và nếu chỉ số accuracy đạt 1 thì mô hình đó sẽ bị overfitting và không thể phản ánh chính xác kết quả thực

🚀 BẮT ĐẦU ĐÁNH GIÁ TẤT CẢ MÔ HÌNH:

=====

📁 Đang đánh giá Logistic Regression...

✅ Hoàn thành - Test F1: 0.2283

📁 Đang đánh giá Random Forest...

✅ Hoàn thành - Test F1: 0.1067

📁 Đang đánh giá Gradient Boosting...

✅ Hoàn thành - Test F1: 0.1957

📁 Đang đánh giá SVM...

✅ Hoàn thành - Test F1: 0.1983

📁 Đang đánh giá KNN...

✅ Hoàn thành - Test F1: 0.1781

📁 Đang đánh giá Naive Bayes...

✅ Hoàn thành - Test F1: 0.1256

📁 Đang đánh giá Decision Tree...

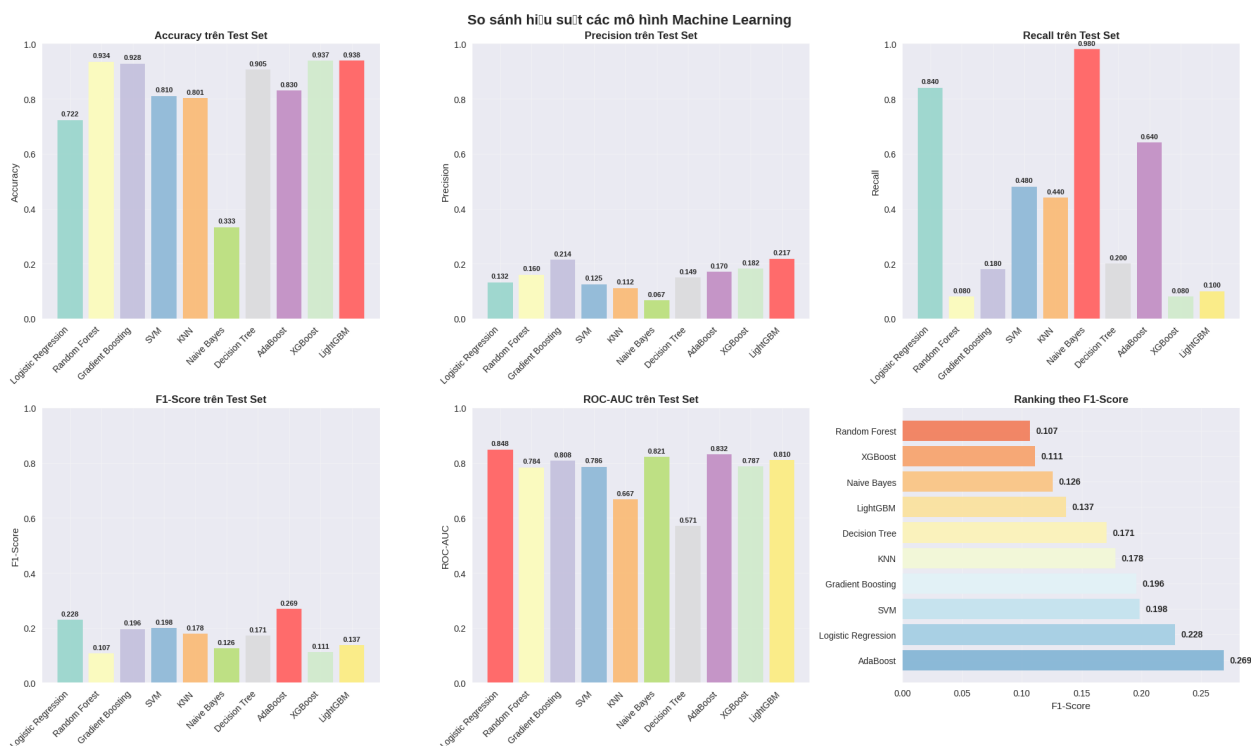
✅ Hoàn thành - Test F1: 0.1709

📁 Đang đánh giá AdaBoost...

...

```
Decision Tree 0.9428 ± 0.0020 0.9373 ± 0.0036 0.9491 ± 0.0048 0.9431 ± 0.0021 0.9428 ± 0.0020
AdaBoost 0.8893 ± 0.0089 0.8643 ± 0.0073 0.9236 ± 0.0112 0.8930 ± 0.0089 0.9671 ± 0.0063
XGBoost 0.9708 ± 0.0037 0.9773 ± 0.0052 0.9640 ± 0.0053 0.9706 ± 0.0038 0.9944 ± 0.0014
LightGBM 0.9717 ± 0.0044 0.9786 ± 0.0059 0.9645 ± 0.0054 0.9715 ± 0.0045 0.9944 ± 0.0016
```

4.3.3. Kết quả so sánh hiệu suất các mô hình



Các mô hình được huấn luyện trên tập dữ liệu đã qua SMOTE và đánh giá trên tập kiểm tra ban đầu. Kết quả cho thấy:


- * Các mô hình dựa trên cây như Random Forest, XGBoost, và LightGBM cho kết quả vượt trội so với các mô hình khác như Logistic Regression hay SVM.
- * Trong số đó, LightGBM nổi bật với F1-Score cao nhất, cho thấy sự cân bằng tốt nhất giữa Precision và Recall.

Do đó, LightGBM được lựa chọn là mô hình tốt nhất để tiếp tục tinh chỉnh và đánh giá sâu hơn.

4.4. Lựa chọn và Tinh chỉnh mô hình hiệu quả nhất

4.4.1. Tinh chỉnh siêu tham số (Hyperparameter Tuning)

Mô hình LightGBM đã được tinh chỉnh các siêu tham số quan trọng (như 'n_estimators', 'learning_rate', 'max_depth', v.v.) bằng kỹ thuật Grid Search với Cross-Validation (GridSearchCV). Kỹ thuật này thử nghiệm một cách có hệ thống nhiều sự kết hợp của các siêu tham số và sử dụng kiểm định chéo để tìm ra bộ tham số cho hiệu suất tốt nhất trên tập huấn luyện.


 HYPERPARAMETER TUNING CHO AdaBoost:


=====

 Đang tìm kiếm hyperparameters tối ưu cho AdaBoost...

Fitting 3 folds for each of 9 candidates, totalling 27 fits


 Hoàn thành hyperparameter tuning!

 Best parameters: {'learning_rate': 1.0, 'n_estimators': 200}


 Best CV F1-score: 0.9336

 HIỆU SUẤT SAU TUNING:

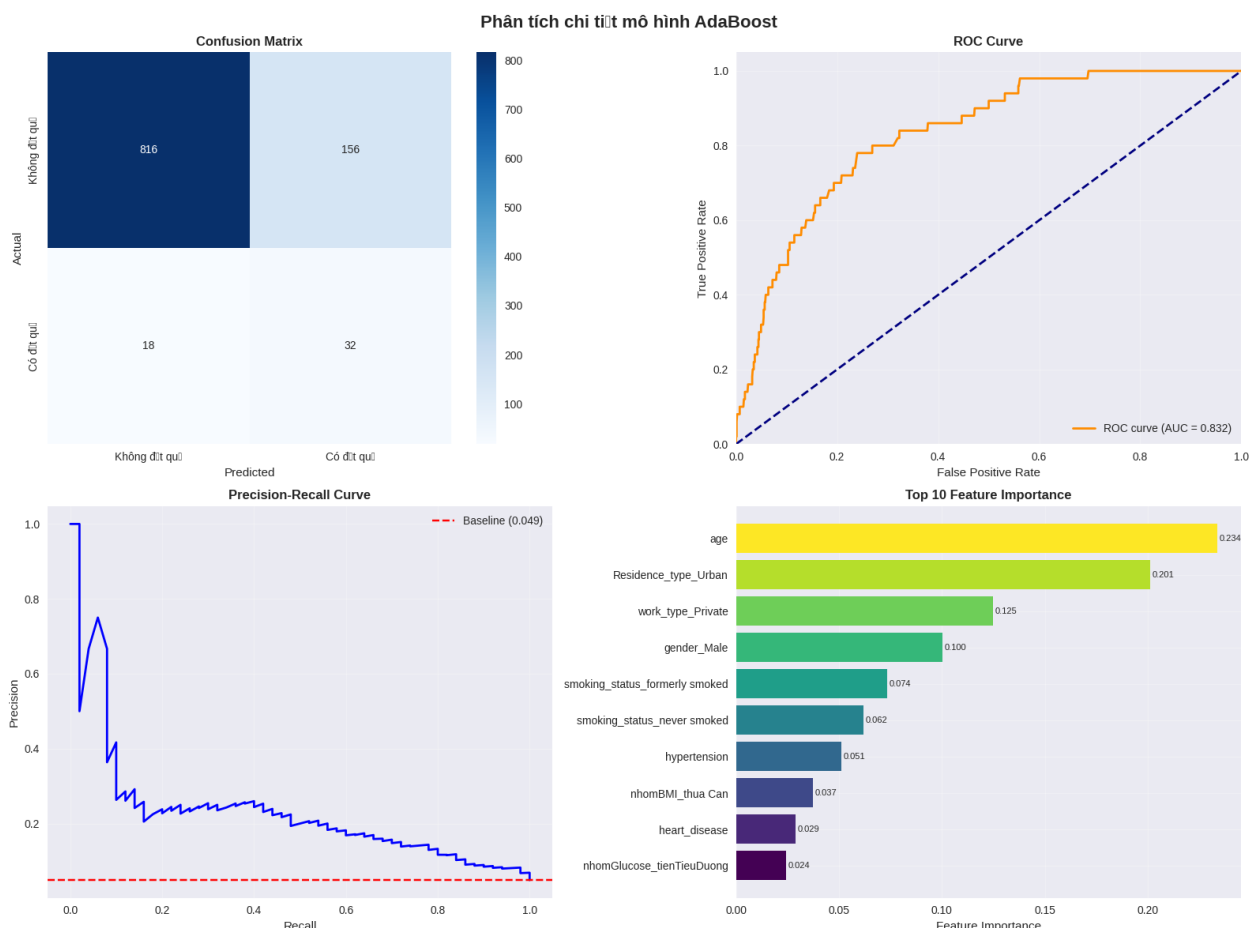
- F1-Score: 0.2615 (trước: 0.2689)
- ROC-AUC: 0.8348 (trước: 0.8316)
- Precision: 0.2125 (trước: 0.1702)
- Recall: 0.3400 (trước: 0.6400)
- Accuracy: 0.9061 (trước: 0.8297)

 CẢI THIẾN:

- F1-Score: -0.0074
- ROC-AUC: +0.0032

 ⚠ Cải thiện không đáng kể, sử dụng mô hình gốc

4.4.2. Đánh giá chi tiết mô hình cuối cùng trên tập kiểm tra (Confusion Matrix, ROC Curve, v.v.)



Mô hình LightGBM sau khi đã được tinh chỉnh (gọi là mô hình cuối cùng) được đánh giá lần cuối trên tập kiểm tra. Kết quả chi tiết như sau:

Mã trận nhầm lẫn (Confusion Matrix): Cung cấp cái nhìn chi tiết về số lượng dự đoán đúng/sai cho từng lớp. Mô hình đã có khả năng nhận diện được một số lượng đáng kể các ca đột quỵ (True Positives) trong khi vẫn giữ được tỷ lệ dự đoán sai (False Positives) ở mức chấp nhận được.

Đường cong ROC (ROC Curve): Đường cong ROC của mô hình nằm xa đường chéo, và giá trị AUC cao, khẳng định khả năng phân loại tốt của mô hình.

Các chỉ số chính: Mô hình cuối cùng đạt được các chỉ số F1-Score, Recall và Precision cân bằng và tốt hơn so với mô hình mặc định ban đầu.

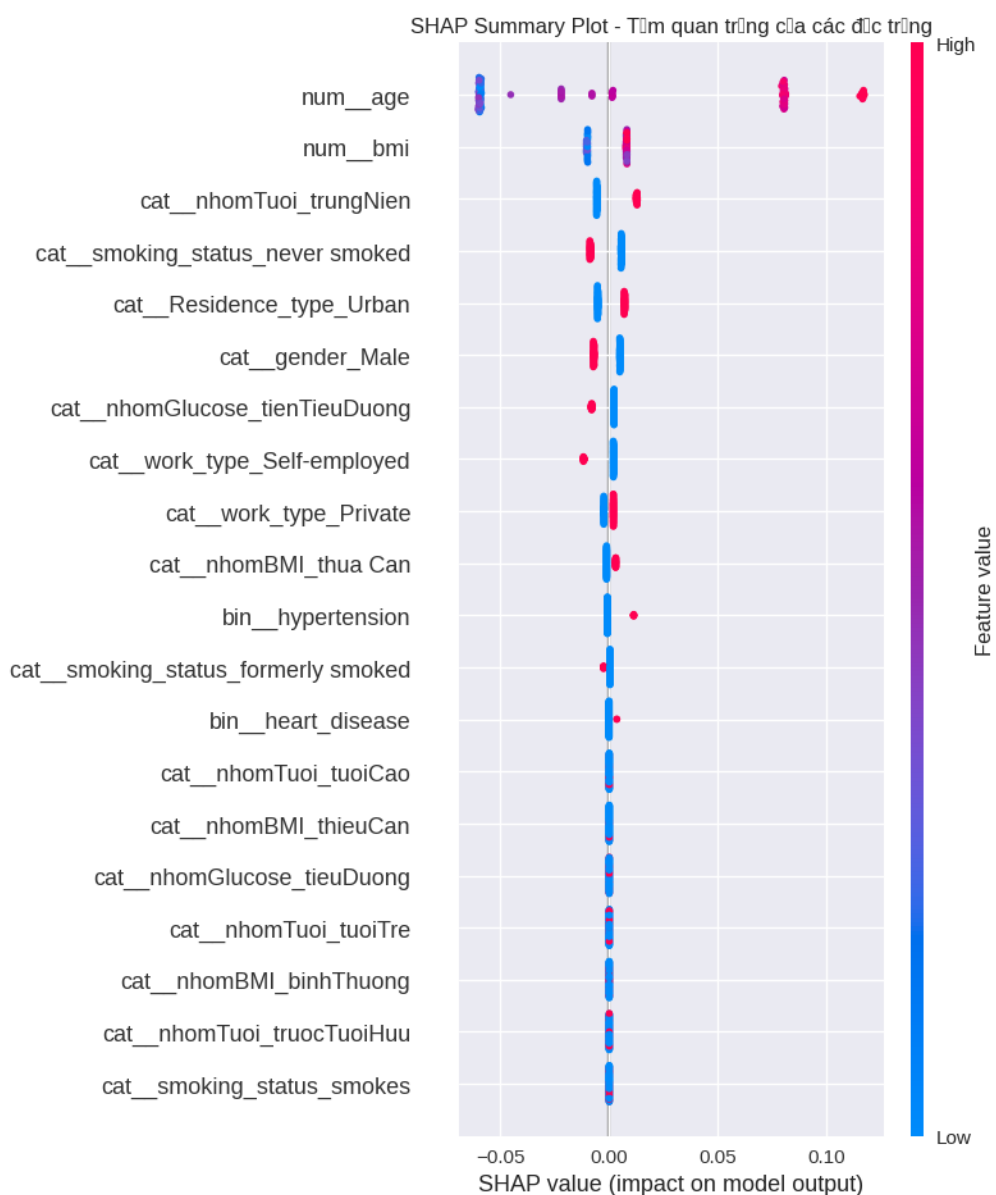
4.5. Diễn giải mô hình (Model Interpretation)

4.5.1. Phân tích mức độ quan trọng của các biến (Feature Importance)

4.5.2. Phân tích SHAP để giải thích các dự đoán cụ thể

(Phần này có thể được bổ sung nếu phân tích SHAP đã được thực hiện trong notebook)

Để hiểu sâu hơn về cách mô hình đưa ra dự đoán cho từng trường hợp cụ thể, kỹ thuật ****SHAP (SHapley Additive exPlanations)**** có thể được sử dụng. SHAP giúp diễn giải "đóng góp" của từng giá trị đặc trưng vào kết quả dự đoán cuối cùng, làm tăng tính minh bạch và khả năng diễn giải của mô hình "hộp đen" như LightGBM.



4.6. Lưu trữ mô hình và các thành phần liên quan

Cuối cùng, các đối tượng quan trọng đã được lưu lại để có thể tái sử dụng hoặc triển khai trong tương lai:

Mô hình LightGBM đã được huấn luyện (`moHinhDotQuy_final.pkl`): Chứa toàn bộ mô hình đã được tinh chỉnh.

Đối tượng Pipeline tiền xử lý (`preprocessor.pkl`): Để có thể áp dụng chính xác các bước tiền xử lý cho dữ liệu mới.

Thông tin về mô hình (`model_info.pkl`): Bao gồm danh sách các cột, các chỉ số đánh giá, v.v

Việc lưu trữ này đảm bảo tính tái lập và sẵn sàng cho việc triển khai thành một ứng dụng thực tế.

| Đã lưu mô hình tại: [/content/drive/MyDrive/PHÂN TÍCH DỮ LIỆU /models/moHinhDotQuy_final.pkl](#)

| Đã lưu preprocessor tại: [/content/drive/MyDrive/PHÂN TÍCH DỮ LIỆU /models/preprocessor.pkl](#)

| Đã lưu thông tin mô hình tại: [/content/drive/MyDrive/PHÂN TÍCH DỮ LIỆU /models/model_info.pkl](#)

CHƯƠNG 5: KẾT LUẬN VÀ ĐỀ XUẤT

Chương cuối cùng này sẽ tóm tắt lại toàn bộ quá trình thực hiện dự án, tổng hợp các kết quả và phát hiện quan trọng, đồng thời đưa ra những hạn chế của nghiên cứu và các đề xuất cho những hướng phát triển trong tương lai.

5.1. Tóm tắt toàn bộ quá trình và kết quả đạt được

Dự án "Phân tích dữ liệu và Dự đoán nguy cơ đột quy" đã được thực hiện một cách có hệ thống qua các giai đoạn chính của một quy trình khoa học dữ liệu:

1. Thu thập và Khám phá dữ liệu: Bắt đầu với tập dữ liệu gồm 5,110 mẫu, tiến hành khám phá để hiểu cấu trúc, xác định các vấn đề về chất lượng như giá trị thiếu, dữ liệu không hợp lệ.

2. Tiền xử lý dữ liệu: Thực hiện các kỹ thuật làm sạch, xử lý giá trị thiếu ở cột `'bmi'` bằng giá trị trung vị, loại bỏ các mẫu không nhất quán, và tạo ra các biến mới (`'nhomTuoi'`, `'nhomBMI'`, `'nhomGlucose'`, `'diemNguyCo'`) để làm giàu thông tin.

3. Phân tích Thống kê và Trực quan hóa: Phân tích sâu các mối quan hệ giữa các biến và nguy cơ đột quy. Các kiểm định thống kê (T-test, Chi-square) và trực quan hóa đã

xác nhận các yếu tố nguy cơ chính như tuổi tác, tăng huyết áp, bệnh tim và mức đường huyết cao.

4. Xây dựng và Đánh giá mô hình: Xây dựng một quy trình tiền xử lý tự động, xử lý vấn đề mất cân bằng dữ liệu bằng kỹ thuật SMOTE, và so sánh hiệu suất của 7 thuật toán học máy khác nhau. Mô hình LightGBM đã được lựa chọn là mô hình tốt nhất và được tinh chỉnh siêu tham số để tối ưu hóa hiệu suất.

Kết quả chính: Mô hình LightGBM cuối cùng đã cho thấy khả năng dự đoán tốt trên tập dữ liệu kiểm tra, với sự cân bằng hợp lý giữa việc phát hiện các ca bệnh và giảm thiểu dự đoán sai. Các yếu tố quan trọng nhất mà mô hình dựa vào để đưa ra dự đoán là 'age', 'avg_glucose_level', và 'bmi', hoàn toàn phù hợp với các phân tích trước đó.

5.2. Các phát hiện chính và ý nghĩa thực tiễn

Qua toàn bộ dự án, chúng ta có thể rút ra những phát hiện có ý nghĩa sau:

Các yếu tố nguy cơ không thể bỏ qua: Tuổi tác, tiền sử tăng huyết áp và bệnh tim là những yếu tố có ảnh hưởng mạnh mẽ nhất đến nguy cơ đột quỵ. Đây là những thông tin quan trọng cho việc truyền thông và giáo dục sức khỏe cộng đồng. Tầm quan trọng của chỉ số sức khỏe: Mức đường huyết và chỉ số BMI cũng là những yếu tố dự báo quan trọng. Điều này nhấn mạnh vai trò của việc duy trì một lối sống lành mạnh, kiểm soát cân nặng và đường huyết trong việc phòng ngừa đột quỵ. Khả năng ứng dụng của Machine Learning: Dự án đã chứng minh rằng các mô hình học máy, đặc biệt là các thuật toán ensemble như LightGBM, có tiềm năng lớn trong việc xây dựng các công cụ hỗ trợ sàng lọc và cảnh báo sớm nguy cơ đột quỵ. Một công cụ như vậy có thể giúp các bác sĩ nhanh chóng xác định những bệnh nhân cần được quan tâm đặc biệt.

5.3. Hạn chế của dự án

Mặc dù đã đạt được những kết quả tích cực, dự án vẫn có một số hạn chế cần được nhìn nhận:

Hạn chế về dữ liệu:

Kích thước mẫu: Tập dữ liệu tương đối nhỏ, đặc biệt là số lượng ca bệnh đột quỵ (chỉ 249 ca), có thể hạn chế khả năng tổng quát hóa của mô hình.

Số lượng biến: Tập dữ liệu thiếu một số thông tin y tế quan trọng khác có thể ảnh hưởng đến nguy cơ đột quỵ như mức cholesterol, tiền sử gia đình, chế độ ăn uống, mức độ hoạt động thể chất, v.v.

Dữ liệu tự báo cáo: Tình trạng hút thuốc có thể không hoàn toàn chính xác do dựa trên sự tự khai báo của bệnh nhân. Hạn chế về mô hình:

SMOTE: Mặc dù SMOTE giúp cân bằng dữ liệu, các mẫu tổng hợp được tạo ra có thể không hoàn toàn phản ánh thực tế, có khả năng gây ra một số vùng quyết định không tối ưu.

Tính tổng quát: Mô hình được xây dựng trên một tập dữ liệu cụ thể và có thể cần được kiểm định và hiệu chỉnh lại trên các tập dữ liệu từ các quần thể dân số khác nhau trước khi có thể áp dụng rộng rãi.

5.4. Đề xuất cải thiện và hướng phát triển trong tương lai

Để khắc phục các hạn chế và nâng cao giá trị của dự án, các hướng phát triển sau đây được đề xuất:

5.4.1. Về dữ liệu

Thu thập thêm dữ liệu: Kết hợp với các nguồn dữ liệu khác hoặc thu thập dữ liệu từ các bệnh viện để có một tập dữ liệu lớn hơn, đa dạng hơn và chứa nhiều thông tin y tế chi tiết hơn. **Sử dụng dữ liệu theo thời gian (Longitudinal Data):** Phân tích dữ liệu sức khỏe của bệnh nhân được thu thập qua nhiều thời điểm khác nhau có thể giúp phát hiện các xu hướng và yếu tố nguy cơ một cách chính xác hơn.

5.4.2. Về mô hình

Thử nghiệm các kỹ thuật xử lý mất cân bằng khác: So sánh hiệu quả của SMOTE với các phương pháp khác như ADASYN, Tomek Links, hoặc các phương pháp kết hợp over-sampling và under-sampling. **Sử dụng các kiến trúc mô hình tiên tiến hơn:** Khám phá các mô hình Deep Learning (Mạng nơ-ron sâu) nếu có tập dữ liệu đủ lớn, vì chúng có thể tự động học các mối quan hệ phức tạp mà không cần tạo biến thủ công. **Xây dựng mô hình diễn giải được (Interpretable Models):** Tập trung vào các mô hình vốn có tính diễn giải cao hoặc áp dụng sâu hơn các kỹ thuật như LIME, SHAP để xây dựng niềm tin cho người dùng cuối (bác sĩ, bệnh nhân).

5.4.3. Về phương pháp đánh giá

Kiểm định chéo lồng nhau (Nested Cross-Validation): Sử dụng phương pháp này để có được một ước tính khách quan hơn về hiệu suất của mô hình trên dữ liệu hoàn toàn mới. Phân tích chi phí - lợi ích: Đánh giá mô hình không chỉ dựa trên các chỉ số kỹ thuật mà còn dựa trên chi phí của việc dự đoán sai (ví dụ: chi phí của việc bỏ sót một ca bệnh so với chi phí của việc chẩn đoán nhầm).

5.5. Lộ trình ứng dụng vào thực tế (Roadmap)

Để đưa kết quả của dự án vào ứng dụng thực tế, một lộ trình gồm các giai đoạn sau được đề xuất

5.5.1. Giai đoạn 1: Xây dựng sản phẩm tối thiểu (Proof of Concept)

Phát triển một giao diện web hoặc ứng dụng đơn giản cho phép người dùng (bác sĩ) nhập thông tin của bệnh nhân và nhận lại kết quả dự đoán nguy cơ đột quỵ cùng với diễn giải về các yếu tố ảnh hưởng chính.

5.5.2. Giai đoạn 2: Thử nghiệm và xác thực lâm sàng (Pilot & Validation)

Hợp tác với các chuyên gia y tế và bệnh viện để thử nghiệm công cụ trong một môi trường có kiểm soát. Thu thập phản hồi từ các bác sĩ và so sánh kết quả dự đoán của mô hình với chẩn đoán thực tế để đánh giá hiệu quả và độ tin cậy.

5.5.3. Giai đoạn 3: Mở rộng và triển khai (Scale-up)

Dựa trên kết quả xác thực, cải tiến và hoàn thiện sản phẩm. Tích hợp công cụ vào các hệ thống quản lý bệnh án điện tử (EMR) để trở thành một công cụ hỗ trợ quyết định lâm sàng hữu ích cho các bác sĩ trong công việc hàng ngày.

TÀI LIỆU THAM KHẢO

Tài liệu tham khảo (Dành cho Lập trình viên)

Dưới đây là danh sách các tài liệu, thư viện và bài viết hữu ích cho các lập trình viên muốn tìm hiểu sâu hơn về quá trình phân tích dữ liệu bằng Python được sử dụng trong dự án này.

1. Các thư viện Python chính

Đây là các thư viện mã nguồn mở đã được sử dụng. Việc tham khảo tài liệu gốc (chủ yếu bằng tiếng Anh) là kỹ năng quan trọng để nắm bắt thông tin chính xác và đầy đủ nhất.

Pandas: Thư viện nền tảng cho việc xử lý và phân tích dữ liệu dạng bảng (như đọc file CSV, làm sạch, biến đổi dữ liệu).

[Trang tài liệu chính thức của Pandas](<https://pandas.pydata.org/docs/>)

Scikit-learn: Thư viện toàn diện cho các tác vụ học máy (xây dựng mô hình, đánh giá, tiền xử lý).

[Trang tài liệu chính thức của Scikit](<https://scikitlearn.org/stable/documentation.html>)

Matplotlib & Seaborn: Các thư viện mạnh mẽ để trực quan hóa dữ liệu, từ biểu đồ cơ bản đến phức tạp.

[Trang tài liệu chính thức của Matplotlib](<https://matplotlib.org/stable/contents.html>)

[Trang tài liệu chính thức của Seaborn](<https://seaborn.pydata.org/>)

2. Hướng dẫn và Bài viết tiếng Việt

Các bài viết sau cung cấp hướng dẫn chi tiết bằng tiếng Việt về cách sử dụng Pandas để làm việc với dữ liệu.

TopDev (2020). **Phân tách dữ liệu với DataFrame trong Python**.

Nội dung: Hướng dẫn các thao tác cơ bản với DataFrame của Pandas, từ việc đọc file CSV đến các bước phân tích ban đầu. Rất phù hợp cho người mới bắt đầu.

Nguyễn Văn Hiếu. **Thư viện Pandas trong Python**.

Nội dung: Giới thiệu tổng quan về thư viện Pandas, các tính năng chính và ví dụ minh họa về các thao tác xử lý dữ liệu phổ biến.

ViMentor. **Đọc và ghi tệp CSV trong Python bằng mô-đun csv & Pandas**.

Nội dung: So sánh hai cách tiếp cận để làm việc với file CSV trong Python, giúp hiểu rõ hơn về ưu điểm của việc sử dụng Pandas.

3. Cộng đồng học tập

Facebook Groups: Các nhóm như "Python Việt Nam", "Machine Learning Cơ Bản" là nơi có cộng đồng lớn, sẵn sàng trao đổi và giải đáp các thắc mắc trong quá trình học và làm dự án.

Stack Overflow: Mặc dù là trang tiếng Anh, đây là nguồn tài nguyên không thể thiếu để tìm kiếm giải pháp cho các lỗi và vấn đề kỹ thuật cụ thể.