ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

PHÂN TÍCH VÀ DỰ ĐOÁN ĐỘT QUỴ Ở NGƯỜI

Nhóm 20			
Sinh viên thực hiện:			
STT	Họ tên	MSSV	Ngành
1	Nguyễn Thái Nhật An	22520024	KHDL
2	Trần Trí Đức	22520276	KHDL

TP. HỒ CHÍ MINH – 12/2024

1. GIỚI THIỆU

Đồ án này tập trung vào việc phân tích và xử lý dữ liệu liên quan đến nguy cơ đột quỵ ở bệnh nhân dựa[1] trên phân tích dữ liệu y tế. Nhóm đã áp dụng nhiều phương pháp và công cụ bao gồm Python với các thư viện như Pandas, Matplotlib, Seaborn,… để tiến hành xử lý dữ liệu, loại bỏ các giá trị khuyết, thăm dò các đặc trưng để phục vụ cho mục đích phân tích và xây dựng mô hình dự đoán đột quỵ dựa trên các thuộc tính quan trọng. Nhiều mô hình học máy khác nhau từ thư viện sklearn cũng đã được thí nghiệm và mô hình RandomForest đạt kết quả tốt nhất với AUC = 0.92 trên bộ dữ liệu đã thu thập.

Nhóm chúng em cam kết minh bạch về nguồn gốc và xử lý của dữ liệu để đảm bảo tính minh bạch và tin cậy của đề tài. Bộ dữ liệu được sử dụng trong đề tài này được nhóm tham khảo tại Kaggle, không phụ thuộc vào bất kỳ nguồn dữ liệu nào khác.

2. MÔ TẢ BỘ DỮ LIỆU

Bộ dữ liệu bao gồm các thông số đầu vào của bệnh nhân như giới tính, độ tuổi, các bệnh lý nền (như cao huyết áp, tiểu đường) và tình trạng hút thuốc.

Bộ dữ liệu phân tích được tham khảo tại Kaggle với đường liên kết sau: https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset/data.[2]

Bộ dữ liệu gồm 12016 dòng và 12 cột, trong đó:

- Biến phân loại: 5 biến (các biến dạng chuỗi ký tự).
- Biến số: 7 biến (các biến dạng số học).

Thống kê các cột dữ liệu:

Bảng 1. Chi tiết các thuộc tính trong bộ dữ liệu

Tên cột	Kiểu dữ liệu	Số lượng khuyết
Id	Int64	0
Gender	Object	0
Age	Float64	0
Hypertension	Int64	0
Heart_disease	Int64	0
Ever_married	Object	0
Work_type	Object	0
Residence-type	Object	0
Avg-glucose-level	Float64	0
Bmi	Float64	201

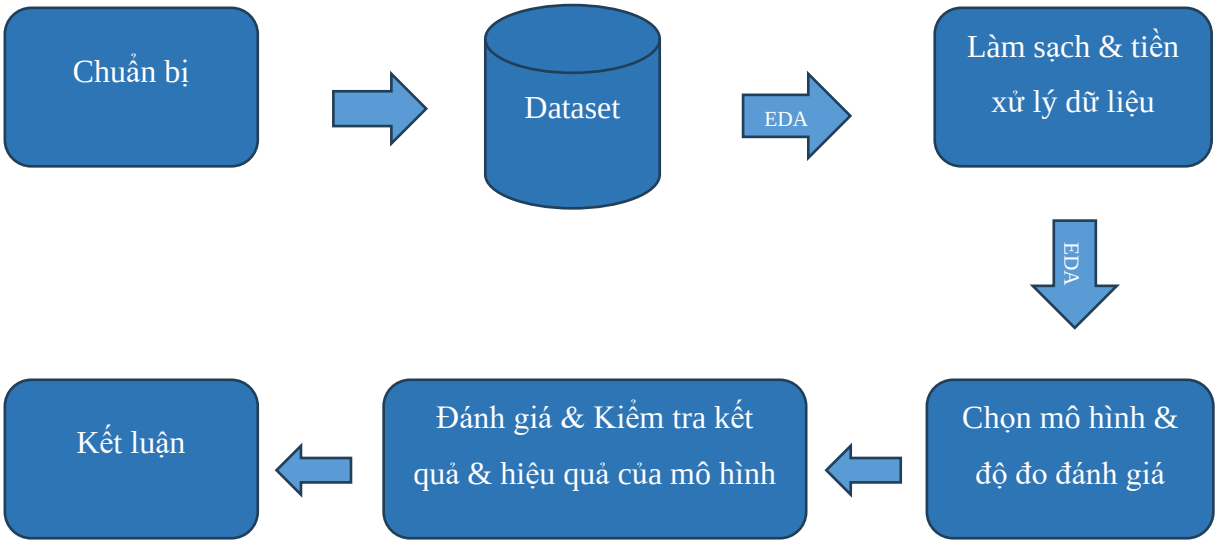
Smoking-status	Object	0
stroke	Int64	0

Bộ dữ liệu gồm 12 thuộc tính:

- **ID**: mã định danh.
- **Gender** (giới tính): "Male" (nam), "Female" (nữ) or "Other" (khác).
- **Age**: độ tuổi của bệnh nhân.
- **Hypertension** (bệnh cao huyết áp): “0” (nếu bệnh nhân không có bệnh cao huyết áp, “1” (nếu bệnh nhân có bệnh cao huyết áp).
- **Heart_disease** (bệnh tim): “0” (nếu bệnh nhân không có bệnh tim, “1” (nếu bệnh nhân có bệnh tim).
- **Ever_married** (tình trạng hôn nhân): "No" or "Yes".
- **Work_type** (loại công việc): "children" (trẻ em), "Govt_job" (công việc nhà nước), "Never_worked" (chưa từng làm việc), "Private" (riêng tư) or "Self-employed" (tự làm chủ).
- **Residence_type** (loại hình cư trú): "Rural" (nông thôn) or "Urban" (thành thị).
- **Avg_glucose_level** (Mức đường huyết trung bình): lượng đường huyết trung bình trong máu.
- **Bmi** (Body Mass Index): chỉ số khối cơ thể.
- **Smoking_status** (tình trạng hút thuốc): "formerly smoked" (đã từng hút thuốc), "never smoked" (chưa từng hút thuốc), "smokes" (đang hút thuốc) or "Unknown" (không có thông tin).

3. PHƯƠNG PHÁP PHÂN TÍCH

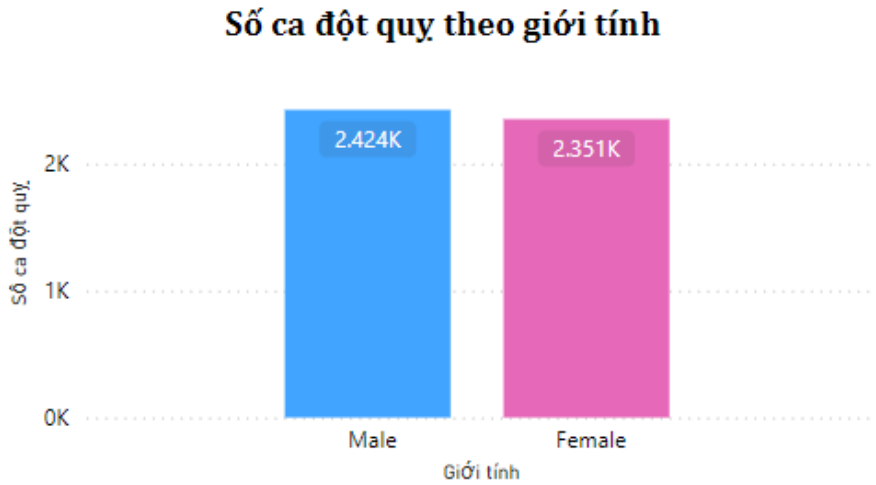
Hình 1. Quy trình phân tích dữ liệu



4. PHÂN TÍCH SƠ BỘ

Số ca đột quỵ theo giới tính:

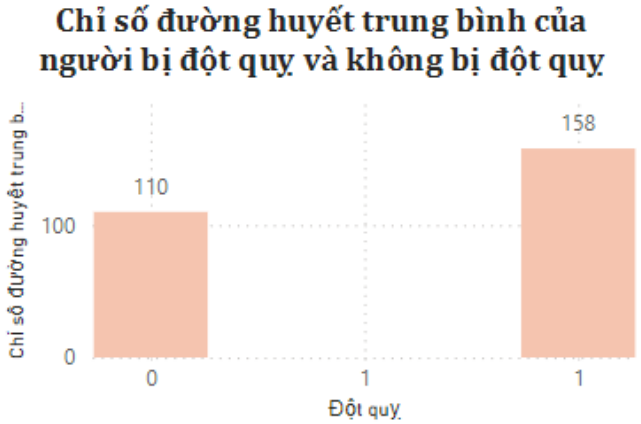
- Nam và nữ có số ca đột quỵ tương đương nhau, với một chút chênh lệch nhỏ nghiêng về phía nam giới.



Hình 2. Số ca đột quỵ theo giới tính

Chỉ số đường huyết:

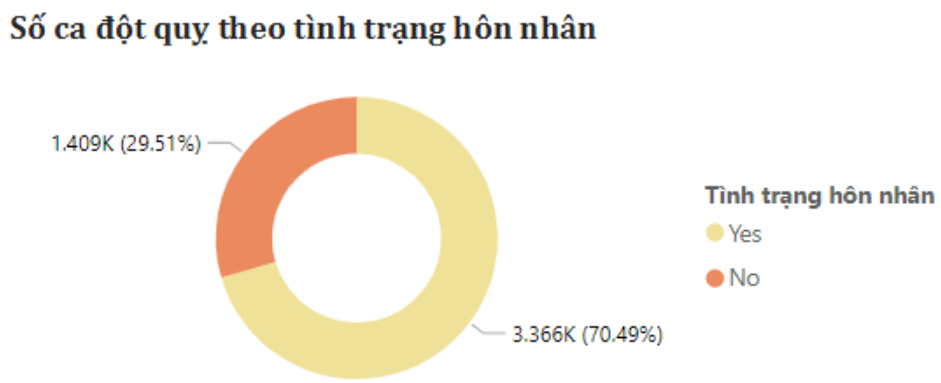
- Chỉ số đường huyết trung bình của nhóm bị đột quỵ cũng cao hơn so với nhóm không bị đột quỵ (158 so với 110), cho thấy chỉ số đường huyết cao có ảnh hưởng đến đột quỵ.



Hình 3. Chỉ số đường huyết trung bình của người bị đột quỵ và không bị đột quỵ.

Tình trạng hôn nhân:

- Người đã kết hôn chiếm 70.49% trong số các ca đột quỵ, cho thấy yếu tố này có thể liên quan đến nguy cơ mắc đột quỵ.
- Người chưa kết hôn chiếm ít hơn (29.51%) trong số các ca đột quỵ, vì có phần lớn những người trẻ tuổi chưa lập gia đình nên có ít nguy cơ mắc đột quỵ.



Hình 4. Số ca đột quỵ theo tình trạng hôn nhân

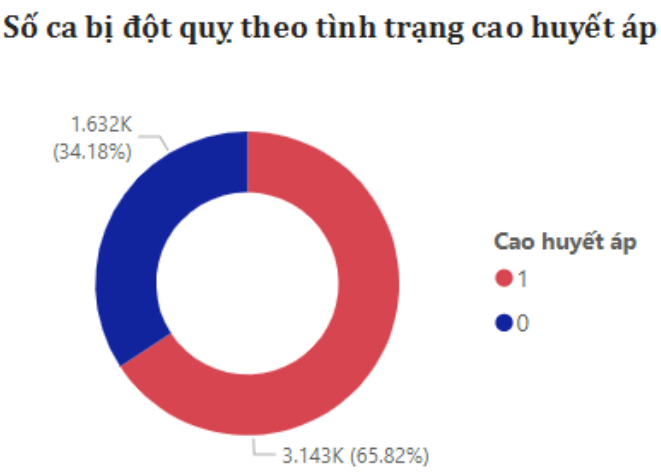
Tình trạng hút thuốc:

- Người đã từng hút thuốc (formerly smoked) chiếm tỷ lệ cao nhất (39.016%) trong số các ca đột quỵ, cho thấy hút thuốc trong quá khứ có thể là yếu tố nguy cơ đáng chú ý liên quan đến đột quỵ.
- Người chưa từng hút thuốc (never smoked) cũng chiếm tỷ lệ khá cao (35.539%), điều này có thể phản ánh rằng đột quỵ không chỉ phụ thuộc vào việc hút thuốc mà còn liên quan đến các yếu tố khác.
- Người đang hút thuốc (smokes) chiếm 24.461%, thấp hơn so với nhóm "đã từng hút thuốc". Tuy nhiên, vẫn cần lưu ý rằng hút thuốc hiện tại có thể làm gia tăng nguy cơ mắc đột quỵ theo thời gian.
- Unknown (không rõ) chỉ chiếm một phần rất nhỏ (0.984%), không ảnh hưởng đáng kể đến phân tích.

Hình 5. Số ca đột quỵ theo tình trạng hút thuốc

Cao huyết áp:

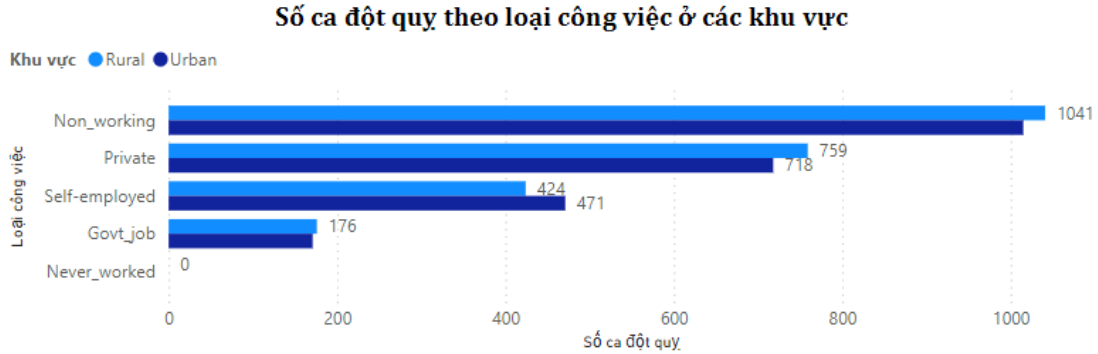
- Đa số trường hợp đột quỵ xảy ra ở nhóm người bị cao huyết áp, điều này nhấn mạnh rằng cao huyết áp là một yếu tố nguy cơ đáng kể gây ra đột quỵ.[3]



Hình 6. Số ca bị đột quỵ theo tình trạng cao huyết áp.

Loại công việc và khu vực sinh sống:

- Nhóm **Non-working (không làm việc)** có số ca đột quỵ cao nhất, đặc biệt tại khu vực đô thị. Điều này có thể liên quan đến lối sống tĩnh tại, ít vận động.
- Nhóm **Private (làm việc tư nhân)** và **Self-employed (tự làm chủ)** cũng có số ca đột quỵ đáng kể, cho thấy những áp lực từ công việc có thể là yếu tố góp phần.
- Nhóm làm công việc nhà nước và chưa từng làm việc có số ca đột quỵ thấp nhất, có thể do lối sống ít căng thẳng hơn hoặc được hỗ trợ y tế tốt hơn.
- Ở khu vực vùng quê, nông thôn đa phần chiếm tỉ lệ mắc đột quỵ cao hơn so với khu vực thành thị có thể là vì chất lượng y tế kém hơn.

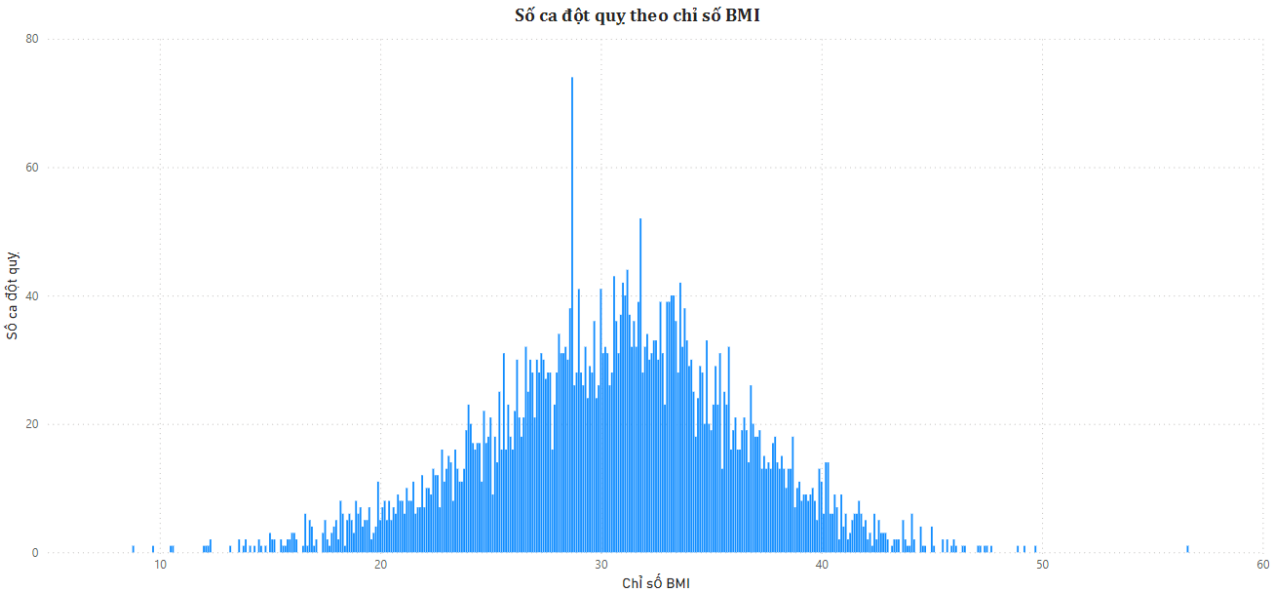


Hình 7. Số ca đột quỵ theo loại công việc ở các khu vực

BMI và lối sống:

- BMI < 20 có số ca đột quỵ rất ít, cho thấy nguy cơ thấp hơn so với nhóm đối tượng này.
- Số lượng ca đột quỵ tăng từ giá trị BMI thấp, đạt đỉnh tại khoảng BMI **30**.
- Sau BMI 30, số lượng ca đột quỵ giảm dần khi chỉ số BMI tăng cao hơn.
- BMI ở mức **30-35** thường được coi là thừa cân hoặc béo phì, dẫn đến nguy cơ cao hơn về bệnh tim mạch và đột quỵ.

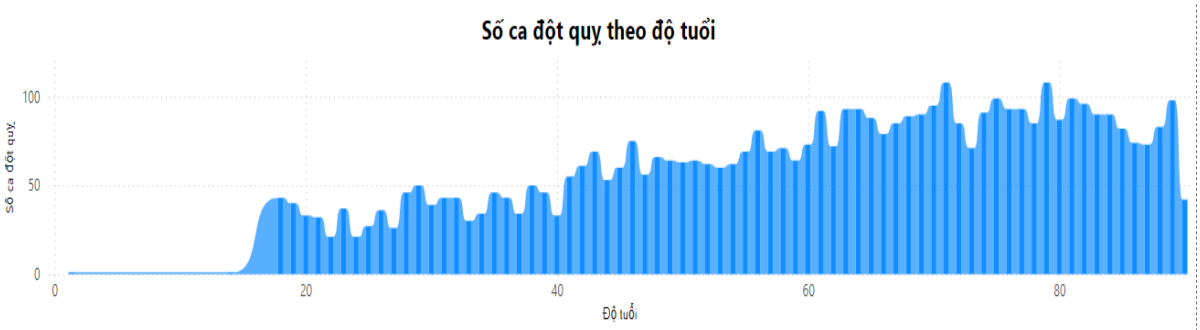
- Việc giảm số ca ở BMI cao hơn có thể phản ánh số lượng ít người thuộc nhóm này hoặc các yếu tố khác liên quan đến y tế.



Hình 8. Số ca đột quỵ theo chỉ số BMI

Độ tuổi:

- Dưới 20 tuổi: Nguy cơ đột quỵ thấp, số ca ghi nhận gần như bằng 0.
- Từ 40 đến 80 tuổi: Đây là giai đoạn nguy cơ cao nhất với số lượng ca đột quỵ nhiều hơn, liên quan đến các bệnh lý mãn tính như tăng huyết áp, tiểu đường, và các vấn đề tim mạch.
- Trên 80 tuổi: Dù nguy cơ vẫn cao, số ca ghi nhận giảm dần, có thể do dân số thuộc nhóm này ít hơn.
- Giai đoạn từ 40 đến 80 tuổi cần được chú ý trong việc phòng ngừa và quản lý sức khỏe.

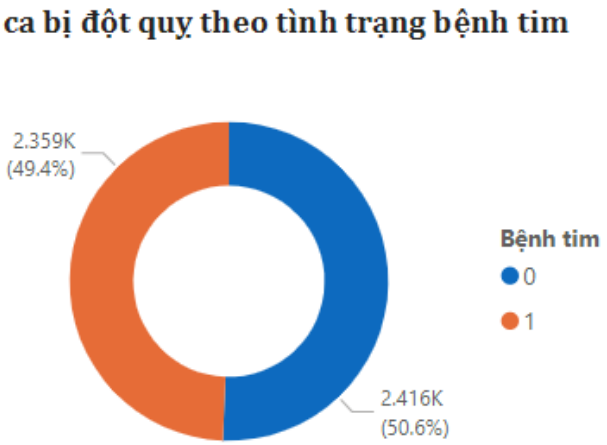


Hình 9. Số ca đột quỵ theo độ tuổi

Bệnh tim (Heart Disease):

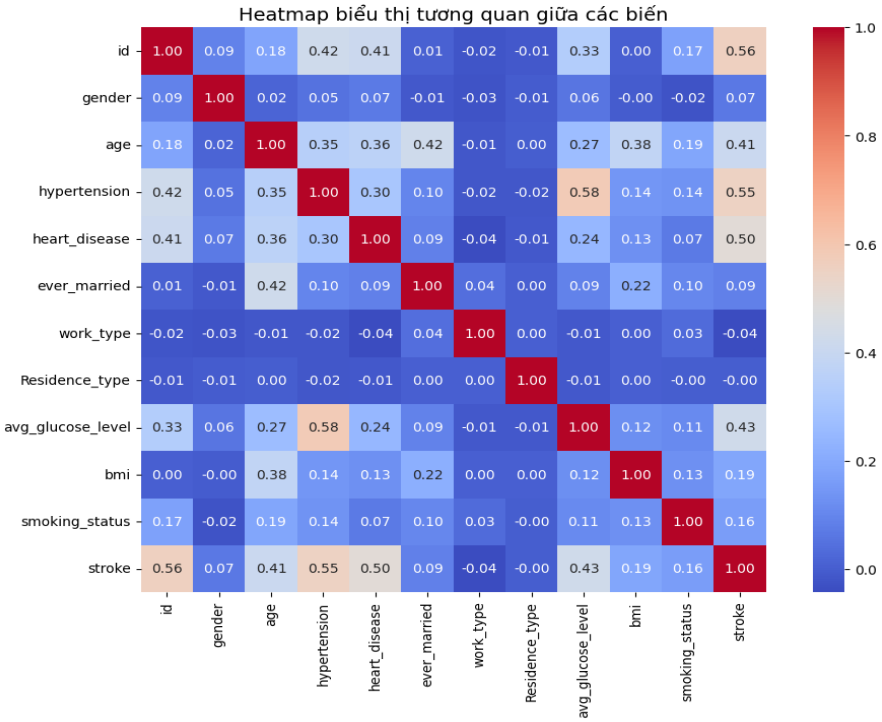
- Biểu đồ cho thấy số ca đột quỵ được phân bố gần như đồng đều giữa hai nhóm: có và không có bệnh tim.

– Điều này có thể ám chỉ rằng bệnh tim không phải là yếu tố duy nhất hoặc quan trọng nhất dẫn đến đột quỵ, và cần xem xét thêm các yếu tố nguy cơ khác như cao huyết áp, lối sống, hoặc di truyền.



Hình 10. Số ca bị đột quỵ theo tình trạng bệnh tim

Kết luận: Bằng số liệu và thông kê sơ bộ ta có thể dễ dàng thấy được các yếu tố ảnh hưởng đến nguy cơ đột quy đó là: Tình trạng hút thuốc, Huyết áp, Tuổi tác và BMI,…ta sẽ vẽ ma trận tương quan giữa các biến.



Hình 11. Heatmap biểu thị tương quan giữa các biến

5. KẾT QUẢ PHÂN TÍCH

Trong phần này, nhóm thực hiện dự đoán và phân tích tập dữ liệu liên quan đến đột quỵ thông qua các mô hình học máy. Quy trình bao gồm:

5.1. Tiền xử lý dữ liệu

Dữ liệu được mã hóa các cột phân loại như gender, work_type, smoking_status, Residence_type, và ever_married bằng LabelEncoder, nhằm chuyển đổi các giá trị phân loại thành số.

Các cột số như age, hypertension, heart_disease, avg_glucose_level, và bmi được chuẩn hóa sử dụng StandardScaler để đảm bảo dữ liệu có phân phối chuẩn, phù hợp với các thuật toán học máy.

Tập dữ liệu được chia thành tập huấn luyện và tập kiểm tra theo tỷ lệ 70:30, với train_test_split cùng tham số stratify để giữ cân bằng tỷ lệ các lớp mục tiêu.

5.2. Chọn mô hình

Nhóm đã triển khai hai nhóm mô hình:

Hồi quy Logistic:

- Áp dụng thuật toán Logistic Regression với 1000 vòng lặp tối đa để đảm bảo hội tụ. Mô hình này dự đoán xác suất đột quỵ dựa trên dữ liệu đã chuẩn hóa.[5]

Random Forest:

- Áp dụng mô hình RandomForestClassifier với 100 cây quyết định. Mô hình này sử dụng chiến lược tổng hợp kết quả từ nhiều cây để tăng độ chính xác và giảm lỗi dự đoán.[4]

Ngoài ra, nhóm cũng thực hiện phân cụm bằng thuật toán Linear Regression: Sử dụng mối quan hệ tuyến tính giữa các đặc trưng và kết quả dự đoán.[5]

5.3. Đánh giá

Chia dữ liệu thành tập huấn luyện và kiểm tra bằng train_test_split với tỷ lệ 70:30, đảm bảo cân bằng nhãn mục tiêu thông qua tham số stratify.

Đo lường hiệu suất mô hình bằng các chỉ số như AUC-ROC, Confusion Matrix, Classification Report. Ngoài ra nhóm còn sử dụng Feature Importance để đánh giá độ quan trọng của từng đặc trưng đo lường.[6]

5.4. Kết luận

Bảng 2. Kết quả độ đo của các mô hình

Mô hình	Precision	Recall	F1-score	Accuracy	Macro Avg	Weighted Avg
LR	0.85	0.87	0.86	0.83	0.82	0.82
LN	0.85	0.86	0.86	0.82	0.82	0.82
RF	0.85	0.89	0.87	0.84	0.84	0.84

6. KẾT LUẬN

Đồ án của nhóm tập trung vào nghiên cứu các yếu tố ngoại cảnh ảnh hưởng đến nguy cơ đột quỵ như BMI, tình trạng hút thuốc, tuổi tác,... [7] nhằm đưa ra những đánh giá khoa học. Tuy nhiên trong khuôn khổ của môn học thì việc nghiên cứu và đưa ra một mô hình hoàn chỉnh là rất khó khả thi, nên nhóm chỉ dừng lại ở việc đánh giá thông qua bộ dữ liệu từ Kaggle cùng với ba mô hình máy học cơ bản: Logistic Regression, Linear Regression, Random Forest. Nhóm tiến hành đánh giá các mô hình qua độ chính xác, F1-score, và ma trận nhầm lẫn. Kết quả cho thấy Random Forest có hiệu suất vượt trội hơn hai mô hình còn lại. Thông qua đồ án, nhóm rút ra được những bài học quan trọng trong phân tích dữ liệu và ứng dụng thuật toán.

TÀI LIỆU THAM KHẢO

[1] L. Feigin, B. Norrving, and G. A. Mensah, "Global Burden of Stroke," *Circulation Research*, vol. 120, no. 3, pp. 439–448, Feb. 2017. DOI: 10.1161/CIRCRESAHA.116.308413.

[2] Kaggle, "Stroke Prediction Dataset," [Online]. Available: https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset/data. [Accessed: Dec. 14, 2024].

[3] M. O'Donnell et al., "Glucose and risk of cardiovascular events and all-cause mortality," *The Lancet Diabetes & Endocrinology*, vol. 6, no. 6, pp. 476–485, 2018. DOI: 10.1016/S2213-8587(18)30037-2.

[4] M. O'Donnell et al., "Glucose and risk of cardiovascular events and all-cause mortality," *The Lancet Diabetes & Endocrinology*, vol. 6, no. 6, pp. 476–485, 2018. DOI: 10.1016/S2213-8587(18)30037-2.

[5] A. Géron, Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 2nd ed. Sebastopol, CA: O'Reilly Media, 2019.

[6] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning: With Applications in R*, 2nd ed. New York: Springer, 2021, ch. 4

[7] J. C. Kissela et al., "Stroke risk factors," *Stroke*, vol. 43, no. 5, pp. 1233–1238, May 2012. DOI: 10.1161/STROKEAHA.111.647620.

PHỤ LỤC PHÂN CÔNG NHIỆM VỤ

STT	Thành viên	Nhiệm vụ
1	Nguyễn Thái Nhật An	Thảo luận chọn đề tài Tiền xử lý dữ liệu Phân tích thăm dò Mô hình hóa và đánh giá Viết báo cáo
2	Trần Trí Đức	Thảo luận chọn đề tài Phân tích thăm dò dữ liệu Viết báo cáo Làm, chỉnh sửa slide