ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC BÁCH KHOA
KHOA KHOA HỌC & KỸ THUẬT MÁY TÍNH

Khai Phá Dữ Liệu

Xây dựng hệ thống dự đoán trụy tim

Giảng viên hướng dẫn: Đỗ Thanh Thái

Danh sách thành viên

Họ và tên	**MSSV**
Hồ Thanh Nhã	2212345
Trần Thế Nhân	2211091
Đặng Minh Nhật	2213111

Tp. Hồ Chí Minh, Tháng 9/2025

Work division table

Full name	Student ID	Assigned works	Completion rate
Hồ Thanh Nhã	2213111		100%
Đặng Minh Nhật	2212345		100%
Trần Thế Nhân	2211091		100%

Mục lục

1 Giới thiệu

1.1 Bối cảnh và động lực phát triển

Đột quỵ là một trong những nguyên nhân hàng đầu gây tử vong và tàn tật trên toàn thế giới. Theo Tổ chức Y tế Thế giới (WHO), mỗi năm có khoảng 15 triệu người bị đột quỵ, trong đó 5 triệu người tử vong và 5 triệu người bị tàn tật vĩnh viễn [1]. Những con số này cho thấy mức độ nghiêm trọng của đột quỵ đối với sức khỏe cộng đồng. Thực tế, đột quỵ được xếp hạng là nguyên nhân gây tử vong đứng thứ hai trên thế giới và là nguyên nhân hàng đầu gây tàn tật kéo dài [2] .

Tại Việt Nam, tỷ lệ mắc đột quỵ đang có xu hướng gia tăng, đặc biệt ở nhóm tuổi trung niên và cao tuổi. Theo số liệu từ Báo cáo Gánh nặng bệnh tật toàn cầu năm 2019, Việt Nam ghi nhận tới 135.999 ca tử vong do đột quỵ, đứng đầu trong nhóm các bệnh lý tim mạch. Tỷ lệ mắc mới ước tính khoảng 222 ca trên 100.000 dân mỗi năm, trong khi tỷ lệ hiện mắc lên đến 1.541 trên 100.000 dân, thuộc nhóm cao nhất khu vực Đông Nam Á. [3]

Phát hiện sớm nguy cơ đột quỵ đóng vai trò then chốt trong phòng ngừa và điều trị hiệu quả. Tuy nhiên, việc sàng lọc thủ công dựa trên kinh nghiệm lâm sàng thường tốn kém thời gian và nguồn lực y tế, đồng thời có thể bỏ sót các trường hợp nguy cơ cao. Trong bối cảnh đó, ứng dụng Machine Learning (ML) và Data Mining vào y tế đã mở ra hướng tiếp cận mới: xây dựng các mô hình dự đoán tự động dựa trên dữ liệu y tế và nhân khẩu học.

Với sự phát triển mạnh mẽ của công nghệ thông tin và khả năng thu thập dữ liệu sức khỏe điện tử (EHR - Electronic Health Records), chúng ta có cơ hội tiếp cận các tập dữ liệu lớn về lịch sử bệnh án, các chỉ số sinh học và yếu tố nguy cơ của bệnh nhân. Việc khai thác hiệu quả nguồn dữ liệu này thông qua các thuật toán ML có thể giúp xác định sớm những cá nhân có nguy cơ cao, từ đó hỗ trợ bác sĩ đưa ra quyết định can thiệp kịp thời.

Động lực chính của đề tài xuất phát từ:

- **Nhu cầu y tế cấp thiết**: Giảm gánh nặng bệnh tật và tử vong do đột quỵ thông qua phát hiện sớm
- **Cơ hội công nghệ**: Tận dụng sức mạnh của ML trong phân tích dữ liệu y tế quy mô lớn
- **Khả năng ứng dụng thực tế**: Xây dựng công cụ hỗ trợ quyết định lâm sàng có độ chính xác cao
- **Giá trị học thuật**: Khám phá các yếu tố nguy cơ quan trọng và mối quan hệ phi tuyến giữa chúng

1.2 Mục tiêu đề tài

Mục tiêu tổng quát của đề tài là xây dựng một hệ thống dự đoán nguy cơ đột quỵ chính xác và đáng tin cậy dựa trên các thuật toán Machine Learning, từ đó hỗ trợ công tác sàng lọc và phòng ngừa bệnh trong thực tiễn y tế.

1.2.1 Mục tiêu cụ thể

1. **Phân tích và làm sạch dữ liệu y tế**:
    - Thu thập và khảo sát bộ dữ liệu Healthcare Dataset Stroke Data từ Kaggle (5,110 bệnh nhân, 12 thuộc tính)

[1] Stroke, Cerebrovascular accident - https://www.emro.who.int/health-topics/stroke-cerebrovascular-accident/
[2] World Stroke Organization: Global Stroke Fact Sheet 2025 - https://pmc.ncbi.nlm.nih.gov/articles/PMC11786524/
[3] Việt Nam thuộc nhóm nước có tỷ lệ đột quỵ cao nhất Đông Nam Á - https://moh.gov.vn/su-kien-y-te-noi-bat/-/asset_publisher//8EeXRtRENhb6/content/viet-nam-thuoc-nhom-nuoc-co-ty-le-ot-quy-cao-nhat-ong-nam-a

- Xử lý giá trị thiếu (missing values), ngoại lai (outliers) và mất cân bằng dữ liệu (class imbalance)
- Thực hiện phân tích khám phá dữ liệu (EDA) để hiểu rõ đặc tính và phân phối của các biến

2. **Xây dựng quy trình tiền xử lý tự động**:
   - Thiết kế pipeline preprocessing chuẩn hóa sử dụng sklearn
   - Áp dụng kỹ thuật SMOTE (Synthetic Minority Oversampling Technique) để cân bằng dữ liệu huấn luyện
3. **Lựa chọn đặc trưng quan trọng**:
   - Sử dụng 4 phương pháp độc lập: Correlation Analysis, Mutual Information, Random Forest Importance, Statistical Tests
   - Xác định top 8 đặc trưng có tác động mạnh nhất đến nguy cơ đột quỵ
   - So sánh và xác thực kết quả với kiến thức y học hiện đại
4. **Huấn luyện và so sánh các mô hình ML**:
   - Triển khai 4 thuật toán: Logistic Regression, Random Forest, SVM (RBF kernel), K-Nearest Neighbors
   - Đánh giá hiệu năng dựa trên F1-Score, Recall, ROC-AUC (thay vì Accuracy do dữ liệu mất cân bằng)
   - Lựa chọn mô hình tối ưu ưu tiên Recall cao (giảm thiểu False Negatives - bỏ sót ca bệnh)
5. **Đánh giá và giải thích kết quả**:
   - Phân tích confusion matrix, ROC curves và các metrics chi tiết
   - Giải thích ý nghĩa y học của các kết quả dự đoán (false positives vs false negatives)
   - Đề xuất chiến lược triển khai mô hình trong môi trường lâm sàng

1.2.2 Chỉ số thành công

Đề tài được coi là thành công khi đạt được:

- **Recall $\geq$ 0.75**: Phát hiện ít nhất 75% các ca đột quỵ (minimize False Negatives)
- **F1-Score $\geq$ 0.20**: Cân bằng hợp lý giữa Precision và Recall
- **ROC-AUC $\geq$ 0.80**: Khả năng phân biệt tốt giữa hai lớp (stroke vs no stroke)
- **Reproducibility**: Kết quả ổn định và có thể tái tạo với random_state cố định

1.3 Đối tượng áp dụng

Hệ thống dự đoán nguy cơ đột quỵ được thiết kế nhằm phục vụ các đối tượng sau:

1.3.1 Đối tượng trực tiếp

1. **Bác sĩ lâm sàng**:
   - Bác sĩ đa khoa tại phòng khám, trung tâm y tế
   - Bác sĩ chuyên khoa tim mạch, thần kinh
   - Y sĩ và điều dưỡng thực hiện khám sàng lọc ban đầu
2. **Cơ sở y tế**:
   - Bệnh viện, phòng khám đa khoa
   - Trung tâm y tế dự phòng
   - Trạm y tế xã/phường (chăm sóc sức khỏe ban đầu)
3. **Người dân**:
   - Người lớn tuổi (từ 40 tuổi trở lên) cần kiểm tra sức khỏe định kỳ
   - Người có yếu tố nguy cơ cao: tăng huyết áp, bệnh tim, tiểu đường
   - Người có tiền sử gia đình mắc đột quỵ
   - Cộng đồng quan tâm đến phòng ngừa bệnh tật

1.3.2 Đối tượng gián tiếp

- **Nhà nghiên cứu y sinh**: Sử dụng mô hình và phương pháp làm tài liệu tham khảo
- **Sinh viên y khoa/công nghệ thông tin**: Học tập về ứng dụng ML trong y tế
- **Nhà quản lý y tế**: Đưa ra chính sách phòng ngừa dựa trên phân tích dữ liệu
- **Nhà phát triển phần mềm y tế**: Tích hợp mô hình vào hệ thống EMR/HIS

1.4 Phạm vi và ứng dụng

1.4.1 Phạm vi nghiên cứu

Phạm vi dữ liệu:

- **Kích thước mẫu**: 5,110 bệnh nhân
- **Nguồn dữ liệu**: Healthcare Dataset Stroke Data (Kaggle)
- **Các biến đầu vào**: 11 biến (1 biến ID bị loại bỏ)
  - Nhân khẩu học: Tuổi, giới tính, tình trạng hôn nhân, loại công việc, nơi cư trú
  - Y tế: Tăng huyết áp, bệnh tim, mức glucose trung bình, BMI (chỉ số khối cơ thể)
  - Lối sống: Tình trạng hút thuốc
- **Biến mục tiêu**: Stroke (0 = Không đột quỵ, 1 = Đột quỵ)
- **Class imbalance**: 95.13% No Stroke vs 4.87% Stroke

Phạm vi phương pháp:

- Preprocessing: Missing value imputation, outlier capping (IQR), scaling, encoding
- Class balancing: SMOTE oversampling (chỉ áp dụng trên training set)
- Feature selection: 4 phương pháp (Correlation, MI, RF Importance, Statistical Tests)
- Modeling: 4 thuật toán supervised learning (Logistic Regression, Random Forest, SVM, KNN)
- Evaluation: F1-Score, Recall, Precision, ROC-AUC, Confusion Matrix

Giới hạn nghiên cứu:

- Dữ liệu từ một nguồn duy nhất (có thể có bias về địa lý, dân tộc)
- Dữ liệu cắt ngang (cross-sectional), không theo dõi dọc (longitudinal)
- Thiếu một số biến y học quan trọng: lipid máu, hoạt động thể chất, chế độ ăn uống
- Không tích hợp dữ liệu hình ảnh y học (CT scan, MRI)

1.4.2 Ứng dụng thực tiễn

1. **Hệ thống sàng lọc tự động**:
   - Tích hợp vào phần mềm quản lý bệnh viện (HIS - Hospital Information System)
   - Cảnh báo tự động khi bệnh nhân có nguy cơ cao ($> 50\%$)
   - Hỗ trợ phân loại ưu tiên cho khám chuyên sâu
2. **Công cụ hỗ trợ quyết định lâm sàng (CDSS)**:
   - Cung cấp điểm số nguy cơ (risk score) kèm giải thích các yếu tố đóng góp
   - Gợi ý xét nghiệm bổ sung dựa trên profile bệnh nhân
   - Đề xuất biện pháp can thiệp phòng ngừa (thuốc, thay đổi lối sống)
3. **Ứng dụng di động cho người dân**:
   - Tự đánh giá nguy cơ đột quỵ tại nhà
   - Nhận thông báo và khuyến nghị từ AI
   - Kết nối với bác sĩ khi phát hiện nguy cơ cao
4. **Nghiên cứu dịch tễ học**:
   - Phân tích yếu tố nguy cơ trên quy mô lớn
   - Xác định nhóm dân số có nguy cơ cao để có chính sách can thiệp
   - Đánh giá hiệu quả các chương trình phòng ngừa

1.5 Cấu trúc báo cáo

Báo cáo được tổ chức thành 6 chương chính như sau:

- **Chương 1: Giới thiệu**
- **Chương 2: Cơ sở lý thuyết**
- **Chương 3: Khảo sát và phân tích dữ liệu**
- **Chương 4: Xây dựng mô hình**
- **Chương 5: Kết quả và đánh giá**
- **Chương 6: Kết luận và hướng phát triển**
- **Phụ lục**

2 Cơ sở lý thuyết

2.1 Tổng quan về bài toán phân loại nhị phân

2.1.1 Định nghĩa bài toán

Bài toán phân loại nhị phân (Binary Classification) là một trong những bài toán cơ bản nhất trong học máy có giám sát, trong đó mục tiêu là phân loại các mẫu dữ liệu vào một trong hai lớp rời rạc. Trong dự án này, hai lớp được định nghĩa là:

- **Lớp 0 (Negative)**: Bệnh nhân không bị đột quỵ (No Stroke)
- **Lớp 1 (Positive)**: Bệnh nhân bị đột quỵ (Stroke)

Với dataset **Healthcare Stroke Data** gồm 5,110 bệnh nhân, bài toán có đặc điểm:

$$\text{Tỷ lệ lớp} = \frac{N_{\text{negative}}}{N_{\text{positive}}} = \frac{4,861}{249} \approx 19.5 : 1 \tag{1}$$

Đây là một bài toán **class imbalance nghiêm trọng**, với 95.13% mẫu thuộc lớp âm và chỉ 4.87% mẫu thuộc lớp dương.

2.1.2 Thách thức của class imbalance

Sự mất cân bằng lớp dữ liệu tạo ra các thách thức đặc biệt:

1. **Bias towards majority class**: Mô hình có xu hướng dự đoán tất cả mẫu thuộc lớp đa số để tối đa hóa accuracy.
2. **Poor minority class detection**: Khả năng phát hiện lớp thiểu số (stroke cases) bị hạn chế nghiêm trọng.
3. **Misleading accuracy**: Accuracy cao không đảm bảo hiệu suất tốt. Ví dụ: mô hình dự đoán tất cả là "No Stroke" vẫn đạt 95% accuracy nhưng hoàn toàn vô dụng trong thực tế y tế.

2.1.3 Metrics đánh giá phù hợp

Với bài toán imbalanced, các metrics quan trọng là:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

$$\text{Recall (Sensitivity)} = \frac{TP}{TP + FN} \tag{3}$$

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

$$\text{ROC-AUC} = \int_0^1 \text{TPR}(FPR^{-1}(x))\, dx \tag{5}$$

Trong đó:

- **TP** (True Positive): Số ca stroke được phát hiện đúng

- **FP** (False Positive): Số ca không stroke bị dự đoán nhầm là stroke
- **FN** (False Negative): Số ca stroke bị bỏ sót (nguy hiểm!)
- **TN** (True Negative): Số ca không stroke được phát hiện đúng

Trong ngữ cảnh y tế, **False Negative** (bỏ sót ca stroke) nguy hiểm hơn **False Positive** (cảnh báo nhầm), do đó **Recall** được ưu tiên cao hơn **Precision**.

2.2 Các kỹ thuật tiền xử lý dữ liệu

2.2.1 Xử lý giá trị thiếu (Missing Values)

Phương pháp Imputation

Trong dataset, cột `bmi` có 201 giá trị thiếu (3.93%). Các phương pháp imputation được sử dụng:

1. **Median Imputation** (cho biến số):

$$x_{\text{missing}} = \text{median}(\{x_i | x_i \text{ không thiếu}\}) \tag{6}$$

   Ưu điểm: Robust với outliers, không bị ảnh hưởng bởi giá trị cực trị.

2. **Mode Imputation** (cho biến phân loại):

$$x_{\text{missing}} = \text{mode}(\{x_i | x_i \text{ không thiếu}\}) \tag{7}$$

   Ưu điểm: Bảo toàn phân phối của biến categorical.

2.2.2 Xử lý outliers

Phương pháp IQR (Interquartile Range)

Outliers được xử lý bằng IQR-based capping:

$$Q_1 = 25\text{th percentile}, \quad Q_3 = 75\text{th percentile} \tag{8}$$

$$IQR = Q_3 - Q_1 \tag{9}$$

$$\text{Lower bound} = Q_1 - 1.5 \times IQR \tag{10}$$

$$\text{Upper bound} = Q_3 + 1.5 \times IQR \tag{11}$$

Giá trị ngoài khoảng [Lower, Upper] được cắt (clipping):

$$x_{\text{capped}} = \begin{cases} \text{Lower} & \text{if } x < \text{Lower} \\ \text{Upper} & \text{if } x > \text{Upper} \\ x & \text{otherwise} \end{cases} \tag{12}$$

Áp dụng cho: `bmi` (max 97.6 → outlier) và `avg_glucose_level`.

2.2.3 Feature Scaling

StandardScaler (Z-score Normalization)

Chuyển đổi features về phân phối chuẩn với mean = 0 và std = 1:

$$x_{\text{scaled}} = \frac{x - \mu}{\sigma} \tag{13}$$

trong đó:

- $\mu = \text{mean}(x)$
- $\sigma = \text{std}(x)$

Ưu điểm:

- Phù hợp với Logistic Regression, SVM
- Bảo toàn thông tin về outliers (sau khi capping)
- Robust với các thuật toán gradient-based

MinMaxScaler

Chuyển đổi features về khoảng [0, 1]:

$$x_{\text{scaled}} = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{14}$$

Ưu điểm:

- Bảo toàn zero entries trong sparse data
- Phù hợp với neural networks

2.2.4 Feature Encoding

One-Hot Encoding

Chuyển đổi biến phân loại thành binary vectors:

$$x \in \{\text{cat}_1, \text{cat}_2, \ldots, \text{cat}_n\} \rightarrow \mathbf{v} \in \{0,1\}^n \tag{15}$$

Ví dụ: `gender` $\in$ {Male, Female, Other} $\rightarrow$ 3 binary columns:

Original	gender_Male	gender_Female	gender_Other
Male	1	0	0
Female	0	1	0
Other	0	0	1

Bảng 2: *One-Hot Encoding cho biến gender*

2.3 Kỹ thuật cân bằng dữ liệu

2.3.1 SMOTE (Synthetic Minority Oversampling Technique)

Nguyên lý hoạt động

SMOTE tạo ra các mẫu tổng hợp (synthetic samples) cho lớp thiểu số bằng cách nội suy giữa các mẫu hiện có:

[1] each minority sample x_i Tìm k nearest neighbors trong lớp minority Chọn ngẫu nhiên một neighbor x_{nn} Tạo sample mới:

$$x_{\text{new}} = x_i + \lambda \times (x_{\text{nn}} - x_i) \tag{16}$$

với $\lambda \sim \text{Uniform}(0, 1)$

Algorithm 1: SMOTE Algorithm

Lưu ý quan trọng:

- SMOTE chỉ áp dụng trên **training set**
- Test set giữ nguyên phân phối gốc (realistic evaluation)
- Fit SMOTE sau khi split train/test để tránh data leakage

2.3.2 Stratified Sampling

Đảm bảo tỷ lệ lớp được bảo toàn khi split train/test:

$$\frac{N_{\text{positive}}^{\text{train}}}{N_{\text{total}}^{\text{train}}} = \frac{N_{\text{positive}}^{\text{test}}}{N_{\text{total}}^{\text{test}}} = \frac{N_{\text{positive}}}{N_{\text{total}}} \tag{17}$$

2.4 Các thuật toán Machine Learning

2.4.1 Logistic Regression

2.4.2 Random Forest

2.4.3 Support Vector Machine (SVM)

Support Vector Machine (SVM) là một mô hình học máy có giám sát mạnh mẽ, chủ yếu được sử dụng cho các bài toán phân loại, nhưng cũng có thể mở rộng cho bài toán hồi quy. Mục tiêu chính của SVM là tìm ra một *ranh giới quyết định* (*decision boundary*) tối ưu để phân tách các lớp dữ liệu khác nhau.

Nguyên lý hoạt động: Tìm siêu phẳng phân tách tối ưu

- **Siêu phẳng (Hyperplane):** Trong không gian n chiều, siêu phẳng là một mặt phẳng có số chiều $n - 1$.
  - Trong không gian 2D, siêu phẳng là một đường thẳng (1 chiều).
  - Trong không gian 3D, siêu phẳng là một mặt phẳng (2 chiều).
- **Mục tiêu:** SVM không chỉ tìm được siêu phẳng phân tách các lớp dữ liệu, mà còn *tối đa hóa khoảng cách* (*margin*) giữa siêu phẳng và các điểm dữ liệu gần nhất của từng lớp.
- **Margin và Support Vectors:** *Margin* là khoảng cách từ siêu phẳng đến các điểm dữ liệu gần nhất. Các điểm nằm trên biên của margin được gọi là *Support Vectors* — đây là các điểm dữ liệu "then chốt" xác định vị trí của siêu phẳng tối ưu. Khi các Support Vectors thay đổi, siêu phẳng cũng sẽ thay đổi tương ứng.

Ưu điểm:

- Hiệu quả trong không gian nhiều chiều, đặc biệt khi số lượng đặc trưng lớn hơn số mẫu.
- Khả năng tổng quát hóa cao nhờ nguyên lý "tối đa hóa margin", giúp mô hình hoạt động tốt ngay cả với dữ liệu huấn luyện nhỏ.
- Ổn định (Robust) với nhiễu do chỉ phụ thuộc vào một tập nhỏ các Support Vectors.

Hạn chế:

- Khó lựa chọn hàm kernel và tinh chỉnh các tham số (C, γ), đòi hỏi kiến thức chuyên sâu và thử nghiệm nhiều lần.
- Chi phí tính toán cao đối với bộ dữ liệu lớn, làm giảm tính khả thi khi triển khai thực tế.
- Khó diễn giải, đặc biệt với các kernel phi tuyến — mô hình hoạt động như một "hộp đen" khó hiểu.
- Hiệu suất giảm khi dữ liệu nhiễu hoặc mất cân bằng giữa các lớp.

2.4.4 K-Nearest Neighbors (KNN)

Cơ sở lý thuyết: K-Nearest Neighbors (KNN) là một thuật toán học máy giám sát (*supervised learning*) đơn giản nhưng hiệu quả, có thể được sử dụng cho cả bài toán phân loại (*classification*) và hồi quy (*regression*). Nguyên lý của KNN dựa trên giả định rằng các điểm dữ liệu có đặc trưng tương tự nhau sẽ có nhãn hoặc giá trị đầu ra gần nhau.

Nguyên lý hoạt động:

- Tính khoảng cách giữa mẫu cần dự đoán và toàn bộ các điểm dữ liệu trong tập huấn luyện (thường dùng khoảng cách Euclidean, Manhattan hoặc Minkowski).
- Chọn ra K điểm dữ liệu gần nhất — gọi là *neighbors*.
- Dự đoán:
  - **Phân loại:** Mẫu mới được gán vào lớp chiếm đa số trong K láng giềng gần nhất.
  - **Hồi quy:** Giá trị dự đoán là trung bình hoặc trung vị của giá trị đầu ra của K láng giềng gần nhất.

Công thức khoảng cách Euclidean phổ biến được sử dụng:

$$d(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

Trong đó, x và y là hai điểm dữ liệu trong không gian đặc trưng n chiều.

Lựa chọn tham số K:

- K nhỏ (ví dụ $K = 1$ hoặc $K = 3$): Mô hình nhạy cảm với nhiễu, dễ bị *overfitting*.
- K lớn (ví dụ $K = 15$ hoặc $K = 30$): Mô hình mượt hơn, ít nhiễu hơn nhưng có thể bị *underfitting*.
- Giá trị K tối ưu thường được xác định thông qua phương pháp *cross-validation*.

Ưu điểm:

- Đơn giản, dễ hiểu, không cần giai đoạn huấn luyện phức tạp.
- Linh hoạt: Áp dụng được cho cả phân loại và hồi quy.
- Không yêu cầu giả định về phân bố dữ liệu.

Hạn chế:

- Chi phí tính toán cao: Mỗi lần dự đoán phải tính khoảng cách với toàn bộ dữ liệu huấn luyện.
- Nhạy cảm với dữ liệu nhiễu và thang đo — cần chuẩn hóa dữ liệu trước khi áp dụng.
- Hiệu suất giảm khi dữ liệu có nhiều chiều (*curse of dimensionality*).

Ứng dụng:

- Nhận dạng chữ viết tay, nhận dạng khuôn mặt.
- Phân loại văn bản, hệ thống gợi ý sản phẩm.
- Dự đoán giá nhà hoặc nhiệt độ trong bài toán hồi quy.

2.5 Tóm tắt chương

Chương này đã trình bày các cơ sở lý thuyết quan trọng:

1. **Bài toán phân loại imbalanced**: Challenges và metrics phù hợp (F1, Recall, ROC-AUC)
2. **Preprocessing techniques**:
   - Missing value imputation (median/mode)
   - IQR-based outlier capping
   - Feature scaling (StandardScaler)
   - One-Hot encoding (12 $\rightarrow$ 21 features)
3. **SMOTE**: Balanced training set từ 4.82% $\rightarrow$ 50% positive class
4. **Algorithms**:
   - Logistic Regression
   - SVM
   - Random Forest
   - KNN

3 Khảo sát và phân tích dữ liệu - EDA

3.1 Tổng quan về dataset

3.1.1 Đặc điểm cơ bản

Dataset **Healthcare Stroke Data** được sử dụng trong dự án bao gồm thông tin y tế và nhân khẩu học của 5,110 bệnh nhân. Mục tiêu là dự đoán nguy cơ đột quỵ dựa trên 11 thuộc tính đầu vào.

3.1.2 Cấu trúc thuộc tính

Dataset bao gồm các loại thuộc tính sau:
1. Biến số liên tục (Numeric):

- `age`: Tuổi (0.08 - 82)
- `avg_glucose_level`: Mức glucose trung bình (55.12 - 271.74 mg/dL)
- `bmi`: Chỉ số khối cơ thể (10.3 - 97.6)

2. Biến nhị phân (Binary):

- `hypertension`: Tăng huyết áp (0: Không, 1: Có)
- `heart_disease`: Bệnh tim (0: Không, 1: Có)

3. Biến phân loại (Categorical):

- `gender`: Giới tính (Male, Female, Other)
- `ever_married`: Tình trạng hôn nhân (Yes, No)
- `work_type`: Loại công việc (Private, Self-employed, Govt_job, children, Never_worked)
- `Residence_type`: Nơi cư trú (Urban, Rural)
- `smoking_status`: Tình trạng hút thuốc (formerly smoked, never smoked, smokes, Unknown)

4. Biến mục tiêu (Target):

- `stroke`: Đột quỵ (0: Không, 1: Có)

3.2 Phân tích biến mục tiêu (Target Analysis)

3.2.1 Phân phối class

Một trong những vấn đề quan trọng nhất của dataset là **class imbalance nghiêm trọng**. Phân tích cho thấy:

$$P(\text{stroke} = 1) = \frac{249}{5110} = 0.0487 \approx 4.87\% \tag{18}$$

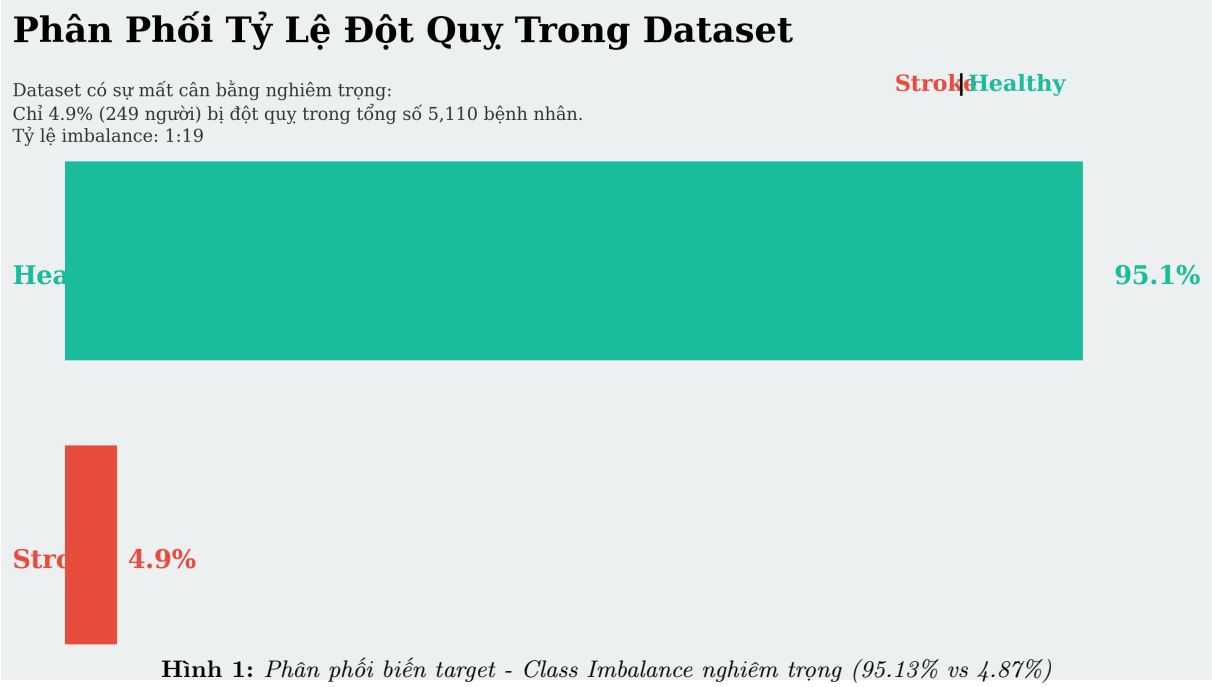
$$P(\text{stroke} = 0) = \frac{4861}{5110} = 0.9513 \approx 95.13\% \tag{19}$$

$$\text{Imbalance Ratio} = \frac{N_{\text{majority}}}{N_{\text{minority}}} = \frac{4861}{249} \approx 19.5 \tag{20}$$

Sự mất cân bằng này có ý nghĩa quan trọng:

- Models có xu hướng bias về lớp đa số (No Stroke)
- Accuracy không phải là metric đánh giá phù hợp
- Cần áp dụng techniques như SMOTE để cân bằng dữ liệu training
- Metrics như F1-Score, Recall, ROC-AUC quan trọng hơn

Visualization: Hình 1 minh họa phân phối không cân bằng của biến target.



Hình 1: *Phân phối biến target - Class Imbalance nghiêm trọng (95.13% vs 4.87%)*

3.3 Phân tích biến số (Numeric Analysis)

3.3.1 Thống kê mô tả

Ba biến số chính trong dataset được phân tích với các thống kê mô tả:

Bảng 3: *Thống kê mô tả các biến số*

Metric	age	avg_glucose_level	bmi
Count	5,110	5,110	4,909
Mean	43.23	106.15	28.89
Std Dev	22.61	45.28	7.85
Min	0.08	55.12	10.30
25%	25.00	77.24	23.50
Median	45.00	91.88	28.10
75%	61.00	114.09	33.10
Max	82.00	271.74	97.60

3.3.2 Phát hiện outliers

Phân tích box plots và histograms cho thấy:
1. Age (Tuổi):

- Phân phối tương đối đều, có một số trường hợp trẻ em (< 1 tuổi)
- Không có outliers nghiêm trọng
- Xu hướng: Nguy cơ đột quỵ tăng theo tuổi

2. Average Glucose Level:

- Phân phối lệch phải (right-skewed)
- Có outliers với giá trị cao (> 200 mg/dL)
- Cần áp dụng IQR capping trong preprocessing

3. BMI:

- Phân phối gần chuẩn (near-normal)
- Có outliers ở cả hai đầu (< 15 hoặc > 50)
- Missing values: 201 mẫu (3.93%) - cần imputation

Hình 2: *Phân tích biến số: Histograms và Box plots theo stroke status*

3.3.3 Correlation với target

Phân tích correlation Pearson với biến target cho thấy:

Bảng 4: *Correlation của biến số với stroke*

Feature	Pearson Correlation (absolute)
age	0.2453
avg_glucose_level	0.1319
bmi	0.0361

Nhận xét:

- **age** có correlation mạnh nhất với stroke (0.245)
- **avg_glucose_level** có correlation trung bình (0.132)
- **bmi** có correlation yếu (0.036)

3.4 Phân tích biến phân loại (Categorical Analysis)

3.4.1 Phân phối và stroke rate theo từng biến

1. Gender (Giới tính):

- Female: 2,994 mẫu (58.6%)
- Male: 2,115 mẫu (41.4%)

- Other: 1 mẫu (0.02%)
- Stroke rate: Female (4.71%), Male (5.11%), Other (0%)

2. Ever Married (Tình trạng hôn nhân):

- Yes: 3,353 mẫu (65.6%)
- No: 1,757 mẫu (34.4%)
- **Quan sát quan trọng**: Người đã kết hôn có stroke rate cao hơn (6.5% vs 1.3%)
- Có thể do correlation với tuổi (người đã kết hôn thường lớn tuổi hơn)

3. Work Type (Loại công việc):

- Private: 2,925 mẫu (57.2%)
- Self-employed: 819 mẫu (16.0%)
- children: 687 mẫu (13.4%)
- Govt_job: 657 mẫu (12.9%)
- Never_worked: 22 mẫu (0.4%)
- Stroke rate cao nhất: Self-employed (7.94%), thấp nhất: children (0.29%)

4. Residence Type (Nơi cư trú):

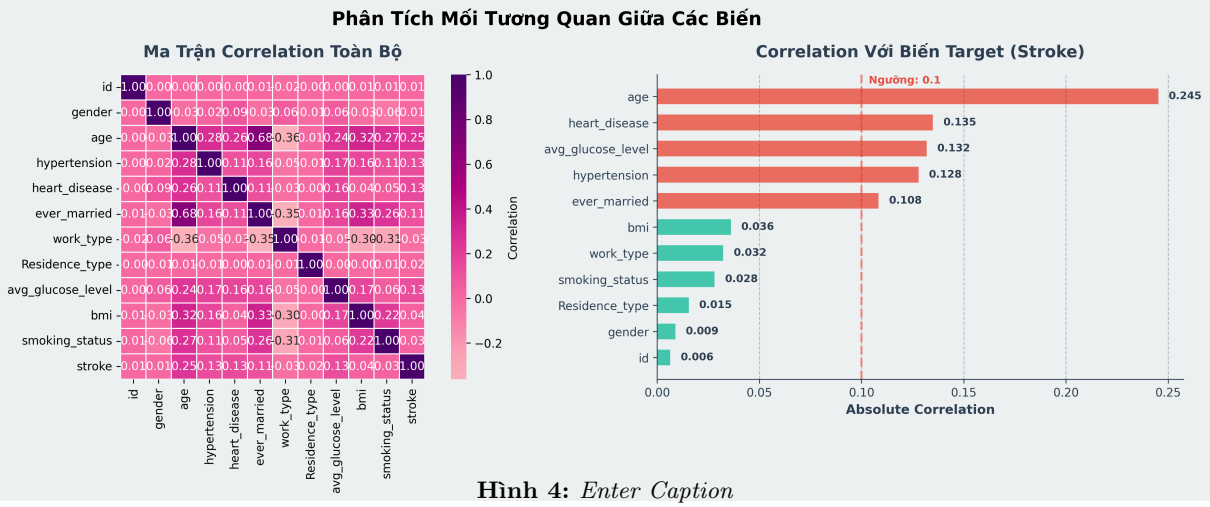
- Urban: 2,596 mẫu (50.8%)
- Rural: 2,514 mẫu (49.2%)
- Stroke rate: Urban (5.20%) vs Rural (4.53%) - Sự khác biệt không lớn

5. Smoking Status (Tình trạng hút thuốc):

- never smoked: 1,892 mẫu (37.0%)
- Unknown: 1,544 mẫu (30.2%)
- formerly smoked: 885 mẫu (17.3%)
- smokes: 789 mẫu (15.4%)
- Stroke rate cao nhất: formerly smoked (7.91%), Unknown có rate thấp nhất (3.04%)

Hình 3: *Phân tích biến phân loại: Count plots phân biệt theo stroke status*

3.5 Phân tích tương quan (Correlation Analysis)



Hình 4: *Enter Caption*

3.5.1 Ma trận correlation

Để phân tích correlation giữa các biến, các biến categorical được encode thành numeric codes. Ma trận correlation hoàn chỉnh cho thấy:

Bảng 5: *Top correlations với biến target stroke*

Feature	Correlation (absolute)
age	0.2453
heart_disease	0.1349
avg_glucose_level	0.1319
hypertension	0.1279
ever_married	0.1083
bmi	0.0361
work_type	0.0323
smoking_status	0.0281
Residence_type	0.0155
gender	0.0089

Insights quan trọng:

- **Age** là predictor mạnh nhất (r = 0.245)
- **Heart disease** và **Hypertension** có correlation trung bình (0.13-0.14)
- **Residence_type** và **Gender** có correlation rất yếu (< 0.02)
- Không có multicollinearity nghiêm trọng giữa các features

Hình 5: *Ma trận correlation heatmap - Phân tích mối quan hệ giữa các biến*

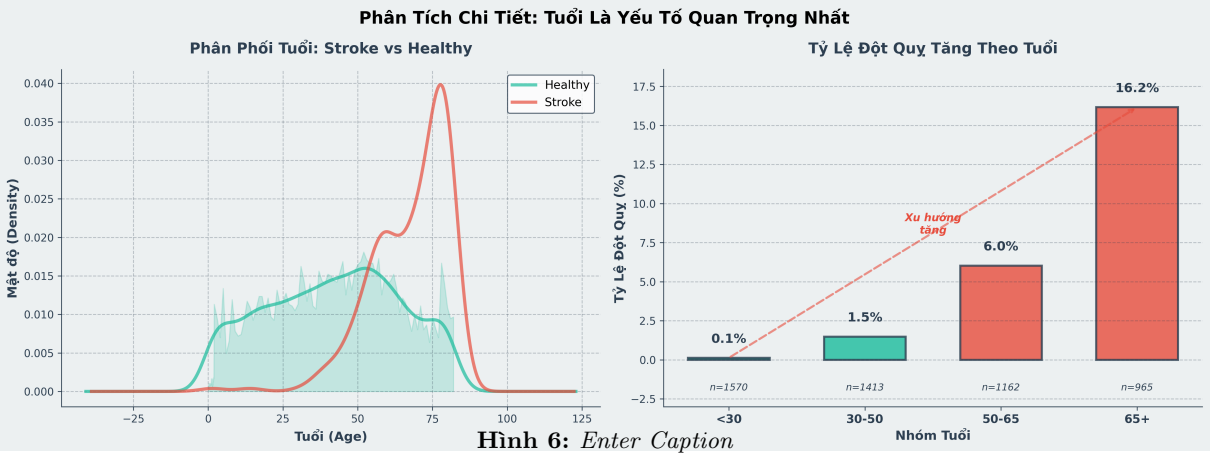
3.6 Phân tích chi tiết biến Age

Do **age** là biến có correlation cao nhất với stroke, một phân tích chuyên sâu được thực hiện:

3.6.1 Phân chia nhóm tuổi

Dữ liệu được chia thành 4 nhóm tuổi:

- $<$ **30**: Trẻ và người trẻ tuổi
- **30-50**: Trung niên
- **50-65**: Người lớn tuổi
- **65+**: Người cao tuổi



Hình 6: *Enter Caption*

3.6.2 Phát hiện chính

1. **Xu hướng tăng mạnh**: Tỷ lệ stroke tăng gấp 124 lần từ nhóm $<$ 30 (0.13%) đến nhóm 65+ (16.17%)
2. **Ngưỡng quan trọng**: Nguy cơ tăng đáng kể sau 50 tuổi

$$\text{Risk Ratio}_{65+/<30} = \frac{16.17\%}{0.13\%} \approx 124 \tag{21}$$

3. **Phân phối không đồng đều**: Nhóm $<$ 30 chiếm 30.7% mẫu nhưng chỉ có 0.8% ca stroke
4. **Nhóm cao rủi ro**: Người cao tuổi (65+) chiếm 18.9% mẫu nhưng có 62.7% tổng số ca stroke

Hình 7: *Phân tích chi tiết tuổi: Phân phối và tỷ lệ stroke theo nhóm tuổi*

3.7 Phát hiện và kết luận từ EDA

3.7.1 Phát hiện chính

1. **Class Imbalance nghiêm trọng**:
   - Tỷ lệ 19.5:1 (4,861 vs 249)
   - Cần SMOTE hoặc class_weight trong training

- Metrics: Ưu tiên F1-Score, Recall, ROC-AUC thay vì Accuracy

2. **Biến quan trọng nhất - Age**:
   - Correlation cao nhất (0.245)
   - Tỷ lệ stroke tăng 124 lần từ < 30 đến 65+
   - Có thể áp dụng feature engineering (age groups, polynomial features)
3. **Missing Values**:
   - Chỉ BMI có 201 missing values (3.93%)
   - Cần imputation strategy: median/mean/KNN imputation
4. **Outliers**:
   - avg_glucose_level: Nhiều giá trị cao > 200
   - BMI: Một số giá trị cực đoan (< 15 hoặc > 50)
   - Áp dụng IQR capping trong preprocessing
5. **Categorical Variables**:
   - ever_married có correlation với age và stroke (Yes: 6.56% vs No: 1.65%)
   - smoking_status: "formerly smoked" có stroke rate cao (7.91%)
   - Residence_type: Urban (5.20%) vs Rural (4.53%) - khác biệt nhỏ
   - Gender: Female (4.71%) vs Male (5.11%) - tương đương nhau
6. **Feature Engineering Opportunities**:
   - Age groups (binning)
   - Interaction features: age $\times$ hypertension, age $\times$ heart_disease
   - Polynomial features cho age

3.7.2 Hướng tiếp cận preprocessing

Dựa trên EDA, pipeline preprocessing cần bao gồm:

1. **Missing Value Handling**:
   - BMI: SimpleImputer với strategy='median'
2. **Outlier Treatment**:
   - IQR capping cho avg_glucose_level và BMI
   - Công thức: $[Q1 - 1.5 \times IQR, Q3 + 1.5 \times IQR]$
3. **Scaling**:
   - StandardScaler cho numeric features (age, avg_glucose_level, BMI)
   - Quan trọng cho SVM, KNN, Logistic Regression
4. **Encoding**:

- OneHotEncoder cho categorical variables
- handle_unknown='ignore' để xử lý unseen categories

5. **Balancing**:
   - SMOTE oversampling cho training set
   - Giữ nguyên test set distribution để đánh giá realistic

3.7.3 Kết luận

EDA cho thấy dataset có những đặc điểm:

- **Strengths**: Dữ liệu sạch với ít missing values, các features có ý nghĩa y tế rõ ràng
- **Challenges**: Class imbalance nghiêm trọng, outliers trong biến số, correlation yếu giữa hầu hết features với target
- **Key Insight**: Age là predictor mạnh nhất, cần đặc biệt chú ý trong modeling
- **Next Steps**: Áp dụng preprocessing pipeline toàn diện với SMOTE, scaling, encoding

Các visualizations từ EDA được sử dụng để hiểu sâu về dữ liệu và định hướng các bước tiền xử lý, feature selection, và model training tiếp theo.

4 Tiền xử lý dữ liệu

5 Xây dựng mô hình

5.1 Logistic Regression

5.2 Random Forest

5.3 KNN

5.4 SVM

6 Kết quả và đánh giá

7 Kết luận và hướng phát triển

8 Nguồn và tài liệu tham khảo