MỤC LỤC

[I. TÓM TẮT 1](#_Toc209664358)

[II. MỤC TIÊU 1](#_Toc209664359)

[III. TÌNH HÌNH NGHIÊN CỨU VÀ CÔNG NGHỆ LIÊN QUAN 1](#_Toc209664360)

[IV. TÍNH CẤP THIẾT CỦA ĐỀ TÀI 1](#_Toc209664361)

[V. PHƯƠNG PHÁP NGHIÊN CỨU 2](#_Toc209664362)

[VI. KẾ HOẠCH THỰC HIỆN 2](#_Toc209664363)

[VII. DANH MỤC TÀI LIỆU THAM KHẢO 3](#_Toc209664364)

ĐỀ CƯƠNG ĐỒ ÁN 1

# I. TÓM TẮT

Đề tài tập trung nghiên cứu, phân tích và xây dựng một hệ thống hỗ trợ chẩn đoán đột quỵ dựa trên dữ liệu y tế. Sử dụng tập dữ liệu gồm 5.110 bệnh nhân với 12 thuộc tính lâm sàng và nhân khẩu học (tuổi, giới tính, chỉ số BMI, huyết áp, bệnh tim, mức glucose trung bình, tình trạng hôn nhân, loại công việc, nơi cư trú, thói quen hút thuốc…). Hệ thống ứng dụng các thuật toán Machine Learning nhằm phát hiện sớm nguy cơ đột quỵ, hỗ trợ bác sĩ ra quyết định, từ đó góp phần giảm tỷ lệ tử vong và di chứng.

# II. MỤC TIÊU

1. Xây dựng mô hình dự đoán nguy cơ đột quỵ dựa trên các đặc trưng y tế quan trọng.
2. Đánh giá và lựa chọn mô hình tối ưu bằng các chỉ số AUC, Precision, Recall, F1-score.
3. Xác định các yếu tố nguy cơ chính gây đột quỵ thông qua phân tích đặc trưng (feature importance).
4. Xây dựng hệ thống demo cho phép nhập dữ liệu bệnh nhân và đưa ra dự đoán nguy cơ đột quỵ.

# III. TÌNH HÌNH NGHIÊN CỨU VÀ CÔNG NGHỆ LIÊN QUAN

1. Trong nước: Một số nghiên cứu liên quan đến bệnh tim mạch và tiểu đường, nhưng hệ thống chuyên biệt cho chẩn đoán đột quỵ còn hạn chế.

2. Quốc tế: Nhiều nghiên cứu đã chứng minh hiệu quả của AI/ML trong y tế (Logistic Regression, Random Forest, Neural Networks) với AUC cao (>0.8).

3. Công nghệ sử dụng trong đề tài:

• Ngôn ngữ: Python.

• Thư viện: scikit-learn, pandas, matplotlib.

• Mô hình: Logistic Regression, Random Forest.

• Dữ liệu: Healthcare Stroke Dataset (5110 bản ghi, 249 ca đột quỵ).

# IV. TÍNH CẤP THIẾT CỦA ĐỀ TÀI

• Đột quỵ là nguyên nhân tử vong hàng đầu tại Việt Nam và thế giới.

• Dữ liệu y tế hiện có rất phong phú, tạo điều kiện ứng dụng AI để dự báo nguy cơ.

• Hệ thống có thể hỗ trợ bác sĩ đưa ra chẩn đoán nhanh chóng, giảm thiểu sai sót, nâng cao hiệu quả điều trị.

# V. PHƯƠNG PHÁP NGHIÊN CỨU

1. Cách tiếp cận

Nghiên cứu tiếp cận theo hướng **ứng dụng trí tuệ nhân tạo (AI)** và **học máy (Machine Learning)** trong y tế. Cụ thể, sử dụng dữ liệu bệnh án (thông tin nhân khẩu học, lối sống, chỉ số y tế) để huấn luyện mô hình dự đoán nguy cơ đột quỵ. Kết quả mô hình sẽ được tích hợp vào một hệ thống phần mềm hỗ trợ chẩn đoán, giúp bác sĩ đưa ra quyết định nhanh chóng và chính xác.

2. Phương pháp nghiên cứu, kỹ thuật sử dụng

* **Phương pháp nghiên cứu:**
* Phân tích dữ liệu thực nghiệm.
* Áp dụng các mô hình học máy có giám sát để dự đoán nhãn (stroke: có/không).
* So sánh, đánh giá hiệu quả mô hình qua các chỉ số (Precision, Recall, F1-score, ROC AUC).
* **Kỹ thuật sử dụng:**
* Tiền xử lý dữ liệu: xử lý giá trị thiếu, chuẩn hóa dữ liệu số, mã hóa biến phân loại.
* Thuật toán: Logistic Regression, Random Forest.
* Công cụ: Python (pandas, scikit-learn, matplotlib).
* Đánh giá mô hình: confusion matrix, classification report, ROC curve.

3. Các bước thực hiện

Bước 1: Thu thập dữ liệu  
• Sử dụng bộ dữ liệu về đột quỵ (stroke dataset) với hơn 5000 bản ghi.  
• Xác định biến mục tiêu: stroke (có/không).

Bước 2: Xử lý dữ liệu  
• Làm sạch dữ liệu: loại bỏ giá trị thiếu, dữ liệu bất thường.  
• Chuẩn hóa dữ liệu số và mã hóa dữ liệu phân loại.  
• Chia dữ liệu thành tập huấn luyện và tập kiểm tra.

Bước 3: Lựa chọn mô hình học máy  
• Logistic Regression, Random Forest, SVM, KNN.

Bước 4: Huấn luyện mô hình  
• Huấn luyện trên tập train.

Bước 5: Đánh giá mô hình  
• Chỉ số: Accuracy, Precision, Recall, F1-score, ROC AUC.  
• Vẽ confusion matrix và ROC curve.

Bước 6: Nâng cao độ chính xác  
• Dùng oversampling/undersampling.  
• So sánh nhiều mô hình để chọn mô hình tối ưu.

Bước 7: Thử nghiệm và ứng dụng  
• Kiểm thử với dữ liệu mới.  
• Đánh giá tính khả thi.

# VI. KẾ HOẠCH THỰC HIỆN

Bước 1: Thu thập dữ liệu

• Thời gian dự kiến: 1 tuần

• Sản phẩm dự kiến: Bộ dữ liệu bệnh nhân đã được tổng hợp và mô tả.

Bước 2: Xử lý và tiền xử lý dữ liệu

• Thời gian dự kiến: 2 tuần

• Sản phẩm dự kiến: Dữ liệu sạch, đã mã hóa và chuẩn hóa, sẵn sàng cho mô hình.

Bước 3: Lựa chọn và xây dựng mô hình

• Thời gian dự kiến: 1 tuần

• Sản phẩm dự kiến: Tập hợp các mô hình huấn luyện và báo cáo so sánh.

Bước 4: Huấn luyện và đánh giá mô hình

• Thời gian dự kiến: 2 tuần

• Sản phẩm dự kiến: Báo cáo kết quả đánh giá mô hình kèm trực quan hóa.

Bước 5: Nâng cao độ chính xác

• Thời gian dự kiến: 1 tuần

• Sản phẩm dự kiến: Mô hình tối ưu sau tinh chỉnh tham số và cân bằng dữ liệu.

Bước 6: Thử nghiệm và triển khai

• Thời gian dự kiến: 1 tuần

• Sản phẩm dự kiến: Ứng dụng thử nghiệm hỗ trợ chẩn đoán đột quỵ.

Bước 7: Hoàn thiện báo cáo và trình bày

• Thời gian dự kiến: 1 tuần

• Sản phẩm dự kiến: Báo cáo luận văn hoàn chỉnh và slide thuyết trình.

# VII. DANH MỤC TÀI LIỆU THAM KHẢO

1. Kaggle: https://www.kaggle.com/code/docxian/stroke-prediction