Big Data Course

Capstone Project   
Final Report

ⓒ2025 SAMSUNG. All rights reserved.

Samsung Electronics Corporate Citizenship Office holds the copyright of this document.

This document is a literary property protected by copyright law so reprint and reproduction without permission are prohibited.

To use this document other than the curriculum of Samsung Innovation Campus, you must receive written consent from copyright holder.

| XÂY DỰNG MÔ HÌNH DỰ ĐOÁN BỆNH ĐỘT QUỴ DỰA TRÊN TẬP DỮ LIỆU Y SINH HỌC |
| --- |

Date (30/07/2025)

BIGDATA\_NHOM6

Phạm Nhật Hào

Phùng Đức Hiệu

Trần Minh Nhật

Trần Quốc Nhân

Trần Thị Phương Uyên

Content

1. Introduction

1.1. Background Information

1.2. Motivation and Objective

1.3. Members and Role Assignments

1.4. Schedule and Milestones

2. Project Execution

2.1. Simulated Scenario Description

2.2. Datasets Selection and Description

2.3. Data Ingestion Pipeline

2.4. Data Transformation Processing

2.5. Data Query and Insight

3. Results

3.1. Data Ingestion Scripts and Code

3.2. Data Transformation Scripts and Code

3.3. Description and Sample of Transformed Datasets

3.4. Data Visualization of Query Results

4. Projected Impact

4.1. Accomplishments and Benefits

4.2. Future Improvements

5. Team Member Review and Comment

6. Instructor Review and Comment

# 1. Introduction

## **1.1. Background Information**

**1.1.1 Bối cảnh bệnh đột quỵ hiện nay?**

Đột quỵ não là nguyên nhân gây tử vong đứng hàng thứ ba sau bệnh tim mạch và ung thư. Đột quỵ có nguy cơ tử vong cao chiếm 75,2% 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. Người sống sót có thể bị mất thị lực, khả năng nói, liệt và lú lẫn.

**1.1.2 Bệnh đột quỵ là gì?**

Đột quỵ (Stroke) còn gọi là tai biến mạch máu não thường xảy ra đột ngột khi nguồn máu cung cấp cho não bị tắc nghẽn, gián đoạn hoặc suy giảm. Khi đó, não bị thiếu oxy, dinh dưỡng và các tế bào não bắt đầu chết trong vòng vài phút. Người bị đột quỵ có nguy cơ tử vong cao nếu không được phát hiện và cấp cứu kịp thời. Có hai loại đột quỵ chính: đột quỵ do thiếu máu cục bộ (chiếm 87% các trường hợp) và đột quỵ do xuất huyết (chiếm 13% các trường hợp).

## **1.2. Motivation and Objective**

Động lực:

* Phát hiện sớm nguy cơ đột quỵ: Đột quỵ là căn bệnh thường xảy ra đột ngột và khó lường với những hậu quả nghiêm trọng mang đến. Việc dự đoán sớm nguy cơ đột quỵ dựa trên các chỉ số y sinh học có thể giúp cá nhân và đội ngũ y tế can thiệp kịp thời trước khi tình trạng trở nên nghiêm trọng. Phát hiện sớm có thể giúp bệnh nhân có cơ hội điều trị tích cực, hạn chế tổn thương não không thể phục hồi và giảm nguy cơ tử vong.
* Tối ưu hóa nguồn lực y tế: Bệnh viện và cơ sở y tế có thể tối ưu hóa công tác chuẩn bị như phân bổ nhân lực, thiết bị, giường bệnh và tài nguyên phục hồi. Điều này đặc biệt quan trọng trong bối cảnh dân số già hóa và tỉ lệ người mắc bệnh mạn tính ngày càng cao.
* Nâng cao ý thức sức khỏe: Cảnh báo nguy cơ đột quỵ từ dữ liệu y sinh học giúp tăng cường nhận thức của cộng đồng về các yếu tố nguy cơ như tăng huyết áp, hút thuốc, lười vận động,… Từ đó, người dân sẽ chủ động hơn trong việc kiểm tra sức khỏe định kỳ và xây dựng lối sống lành mạnh, phòng ngừa bệnh.
* Hỗ trợ bác sĩ trong quản lý nguy cơ và dự phòng bệnh: Hỗ trợ bác sĩ đưa ra các tư vấn, phác đồ điều trị dự phòng phù hợp với từng bệnh nhân, tiết kiệm thời gian và nâng cao hiệu quả điều trị dài hạn.

Mục tiêu:

* Thu thập dữ liệu đầu vào: tuổi, giới tính, tiền xử bệnh nền (tim mạch, cao huyết áp), tình trạng hôn nhân, BMI, mức đường huyết trung bình trong máu, tình trạng hút thuốc, loại công việc, nơi cư trú.
* Xây dựng mô hình dự đoán nguy cơ mắc bệnh đột quỵ: Ứng dụng các thuật toán học máy (Random Forest, Support Vector Machine, XGBoost,...) để phân tích tập dữ liệu y sinh học và xác định mức độ rủi ro mắc bệnh.
* Thẩm định mức độ hiệu quả của mô hình dự đoán: Sử dụng các thước đo hiệu suất như độ nhạy, độ đặc hiệu, độ chính xác và đường cong ROC để đánh giá khách quan mức độ tin cậy và hiệu quả của mô hình trước khi ứng dụng vào thực tế.
* Tích hợp mô hình vào ứng dụng thực tiễn: Xây dựng mô hình theo hướng có thể triển khai vào nền tảng web giúp nâng cao khả năng cảnh báo sớm cho bác sĩ và người dân.
* Tối ưu và cập nhật mô hình theo dữ liệu thực tế mới: Tiến hành cải tiến mô hình định kỳ bằng cách cập nhật thêm dữ liệu mới, tinh chỉnh thuật toán và điều chỉnh theo xu hướng sức khỏe cộng đồng nhằm giữ vững độ chính xác và độ thích nghi theo thời gian.

## **1.3. Members and Role Assignments**

1. Phạm Nhật Hào: Huấn luyện mô hình, xây dựng giao diện

2. Trần Đức Hiệu: Tiền xử dữ liệu, viết báo cáo

3. Trần Minh Nhật: Đánh giá mô hình, xây dựng giao diện

4. Trần Quốc Nhân: Thu thập dữ liệu, Xây dựng giao diện

5. Trần Thị Phương Uyên (Nhóm Trưởng): Thu thập dữ liệu, viết báo cáo

## **1.4. Schedule and Milestones**

| Công việc | Lịch trình thực hiện |
| --- | --- |
| Thu thập dữ liệu | 30/06/2025 - 09/7/2025 |
| Tiền xử dữ liệu | 10/07/2025 - 11/07/2025 |
| Huấn luyện mô hình | 13/07/2025 - 15/07/2025 |
| Đánh giá mô hình và lựa chọn thuật toán phù hợp | 16/07/2025 - 17/07/2025 |
| Xây dựng giao diện | 18/07/2025 - 23/07/2025 |
| Viết báo cáo + Thuyết trình thử | 24/07/2025 - 30/07/2025 |

# 2. Project Execution

## **2.1. Simulated Scenario Description**

Đề tài mô phỏng một kịch bản ứng dụng trí tuệ nhân tạo trong lĩnh vực y tế, với mục tiêu xây dựng mô hình dự đoán nguy cơ đột quỵ dựa trên các thông tin y sinh học của bệnh nhân.

Dữ liệu sau khi thu thập được tiến hành tiền xử lý và làm sạch để phù hợp cho việc huấn luyện mô hình học máy. Mô hình được xây dựng nhằm hỗ trợ đội ngũ y tế trong việc phát hiện sớm nhóm bệnh nhân có nguy cơ cao mắc bệnh đột quỵ, từ đó có biện pháp can thiệp kịp thời.

Kịch bản giả lập được triển khai trong bối cảnh hệ thống y tế tại các cơ sở khám chữa bệnh, nơi bác sĩ có thể sử dụng phần mềm hỗ trợ chẩn đoán để nhập thông tin bệnh nhân và nhận lại dự đoán nguy cơ đột quỵ.

## **2.2. Datasets Selection and Description**

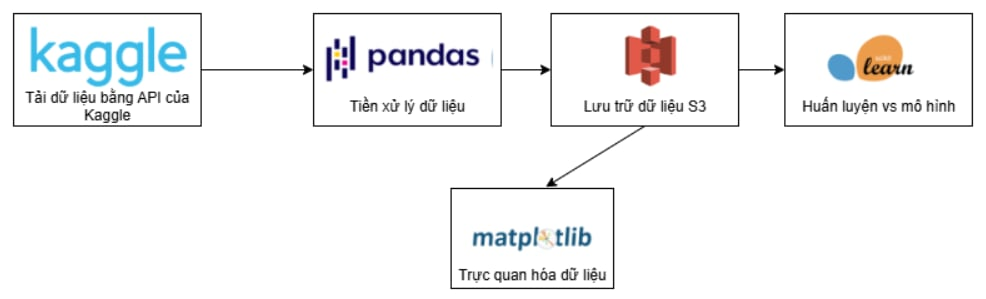
Để phục vụ cho việc xây dựng mô hình dự đoán bệnh đột quỵ, nhóm đã lựa chọn tập dữ liệu *“Stroke Prediction Dataset”* được công bố trên nền tảng Kaggle bởi người dùng có tên *fedesoriano*. Đây là một tập dữ liệu mở, được sử dụng rộng rãi trong các bài toán học máy liên quan đến y tế, đặc biệt là dự đoán nguy cơ bệnh đột quỵ dựa trên dữ liệu hồ sơ bệnh nhân. Lý do nhóm chọn tập dữ liệu này là do dữ liệu đã được làm sạch ở mức cơ bản, định dạng chuẩn dưới dạng bảng (.csv), các thuộc tính rõ ràng, phù hợp với mục tiêu nghiên cứu về dự đoán đột quỵ và trên hết, tập dữ liệu được tác giả cho phép sử dụng miễn phí với mục đích học tập.

Mô tả thuộc tính dữ liệu:

1. id:mã định danh duy nhất của bệnh nhân.
2. gender: giới tính – "Male", "Female" hoặc "Other".
3. age: độ tuổi của bệnh nhân.
4. hypertension: 0 nếu bệnh nhân không bị cao huyết áp, 1 nếu có.
5. heart\_disease: 0 nếu bệnh nhân không mắc bệnh tim, 1 nếu có.
6. ever\_married: "No" hoặc "Yes" – thể hiện tình trạng hôn nhân của bệnh nhân.
7. word\_type: loại hình công việc – "children", "Govt\_jov", "Never\_worked", "Private" hoặc "Self-employed".
8. Residence\_type: nơi cư trú – "Rural" hoặc "Urban".
9. avg\_glucose\_level: mức đường huyết trung bình trong máu.
10. bmi: chỉ số khối cơ thể (BMI).
11. smoking\_status: tình trạng hút thuốc – "formerly smoked", "never smoked", "smokes" hoặc "Unknow"\*.
12. stroke: 1 nếu bệnh nhân đã từng bị đột quỵ, 0 nếu chưa từng bị.

\*Ghi chú: "Unknow" trong cột smoking\_status nghĩa là không có thông tin về tình trạng hút thuốc của bệnh nhân này.

## **2.3. Data Ingestion Pipeline**

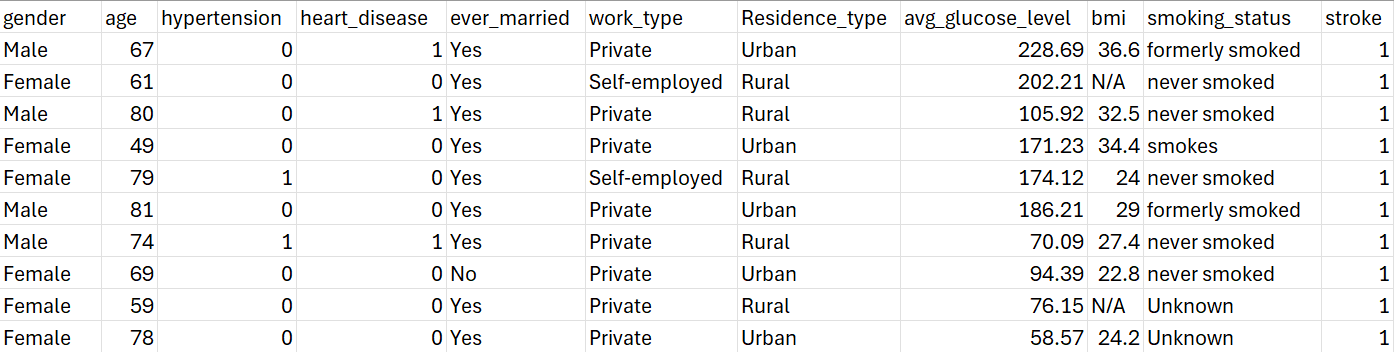


Sơ đồ pipeline này gồm có:

* Tải dữ liệu từ Kaggle. Ở bước này, dữ liệu được lấy trực tiếp trên website vẫn là dữ liệu thô chưa qua xử lý.
* Tiền xử lý dữ liệu với Pandas. Gồm có xử lý giá trị thiếu, mã hóa các biến phân loại, cân bằng dữ liệu.
* Lưu trữ dữ liệu trên S3 (Simple Storage Service). Dữ liệu sau biến đổi, xử lý sẽ được lưu trên đám mây để có thể lưu trữ dài hạn, có thể truy cập từ nhiều nơi, đảm bảo dữ liệu được lưu trữ an toàn.
* Huấn luyện và đánh giá mô hình học máy với Scikit-learn. Các mô hình học máy sẽ được sử dụng để huấn luyện cho mô hình, tạo ra mô hình dự đoán có hiệu quả.
* Trực quan hóa dữ liệu bằng matplotlib. Dữ liệu, kết quả có thể được trực quan hóa để hỗ trợ phân tích, giúp cung cấp cái nhìn trực quan để hiểu rõ hơn về dữ liệu và kết quả mô hình.

## **2.4. Data Transformation Processing**

Tập dữ liệu ban đầu được tải về từ nền tảng Kaggle, có định dạng .csv với tổng cộng 5110 bản ghi và 12 cột, đại diện cho dữ liệu y sinh học của bệnh nhân. Dưới đây là một vài bản ghi của dữ liệu thô:



*Hình 2.4.1 – Các bản ghi với dữ liệu thô.*

Để chuẩn bị cho quá trình huấn luyện mô hình, nhóm đã thực hiện các bước tiền xử lý và biến đổi dữ liệu như sau:

1. Xử lý giá trị thiếu

Cột bmi có vài bản ghi bị thiếu giá trị (N/A), nhóm đã thay thế bằng giá trị trung bình (mean) của toàn bộ cột bmi vào các ô bị thiếu.

1. Loại bỏ cột dữ liệu không cần thiết

Trong tập dữ liệu này, cột id không cần thiết cho việc huấn luyện mô hình. Vì vậy, nhóm đã loại bỏ cột id.

1. Mã hóa các biến phân loại

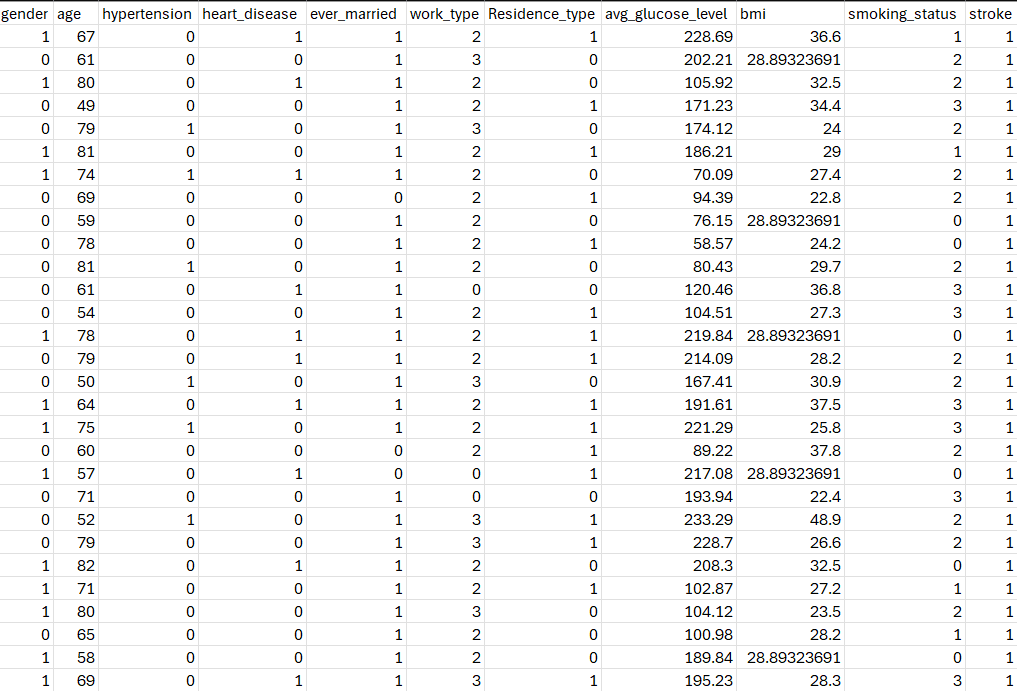
Các thuộc tính dạng chuỗi được mã hóa thành giá trị số bằng công cụ LabelEncoder từ thư viện sklearn.preprocessing:

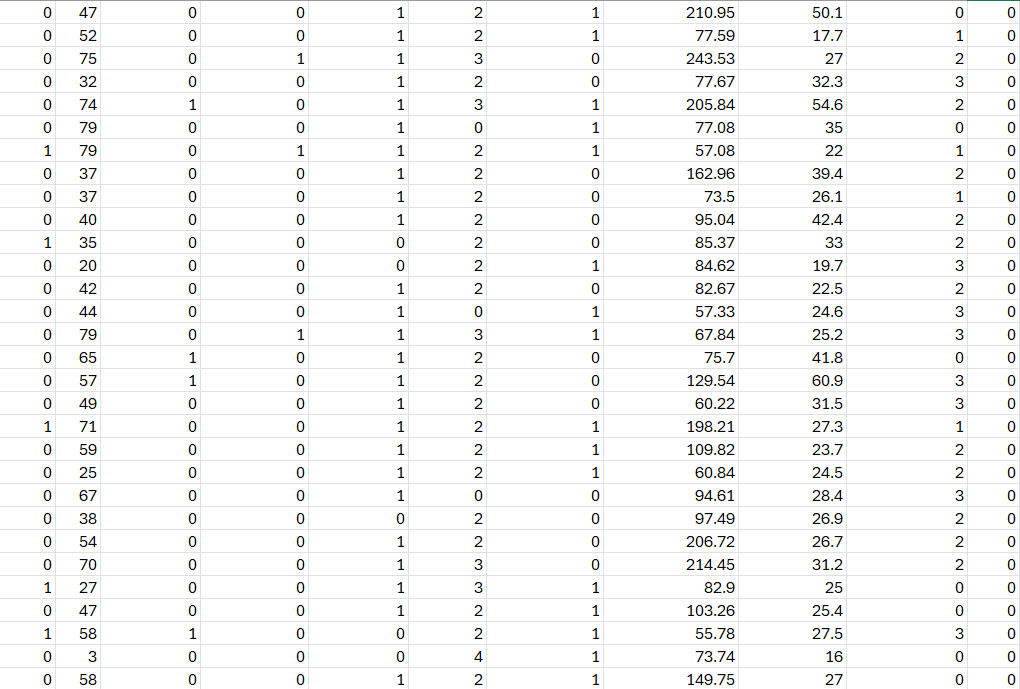
* gender( 0: Female; 1: Male; 2: Other )
* ever\_maried (0: no; 1: yes)
* work\_type (0: Govt\_job; 1: Never\_work; 2: private; 3: Self-employed; 4:children)
* Residence\_type (0: Rural; 1: Urban)
* smoking\_status: (0: Unknow; 1: formerly smoked; 2:never smoked; 3: smokes)

1. Xử lý mất cân bằng dữ liệu

Áp dụng kỹ thuật SMOTE(Synthetic Minority Over-sampling Technique) để tăng cường dữ liệu cho lớp thiểu số (stroke = 1) bằng công cụ SMOTE từ thư viện **imblearn.**

1. Dữ liệu sau khi biến đổi



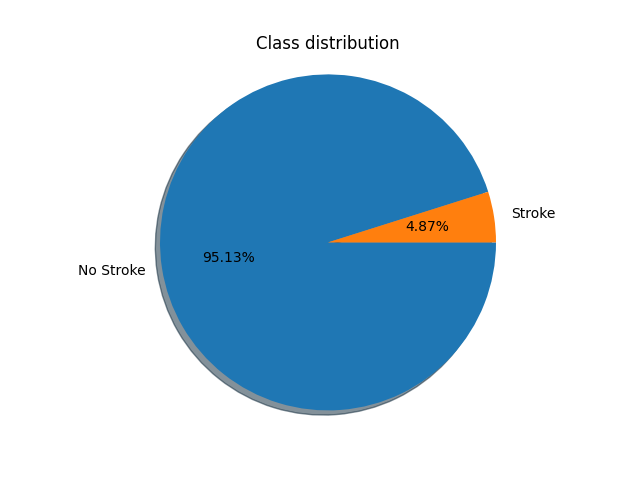


*Hình 2.4.2 – Dữ liệu sau khi tiến hành các bước biến đổi.*

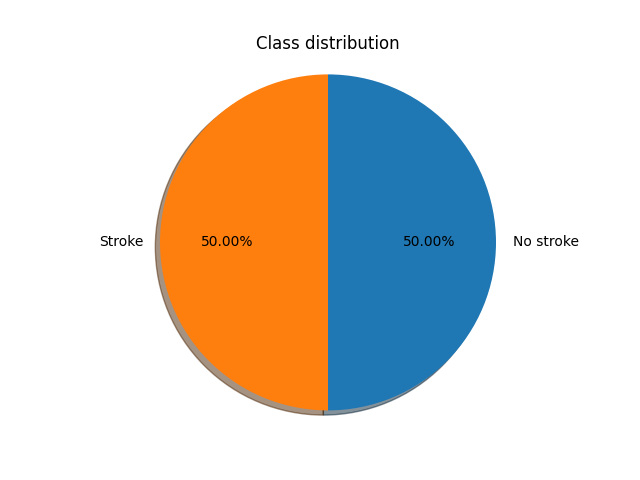
## **2.5. Data Query and Insight**

### 2.5.1 Data Query

Ban đầu, dữ liệu bị mất cân bằng nghiêm trọng, trong đó số người bị đột quỵ chỉ chiếm hơn 5% và điều này có thể gây sai lệch trong quá trình huấn luyện nếu không được xử lý. Sau quá trình xử lý và chuẩn hóa, tập dữ liệu đã ở trạng thái cân bằng, sẵn sàng được đưa vào huấn luyện. Có tổng cộng 9.722 bản ghi (sau khi áp dụng kỹ thuật SMOTE để cân bằng lớp) để sử dụng trong quá trình học máy.

**

*Hình 2.5.1 – Phân bố nhãn trước xử lý mất cân bằng bằng SMOTE*

**

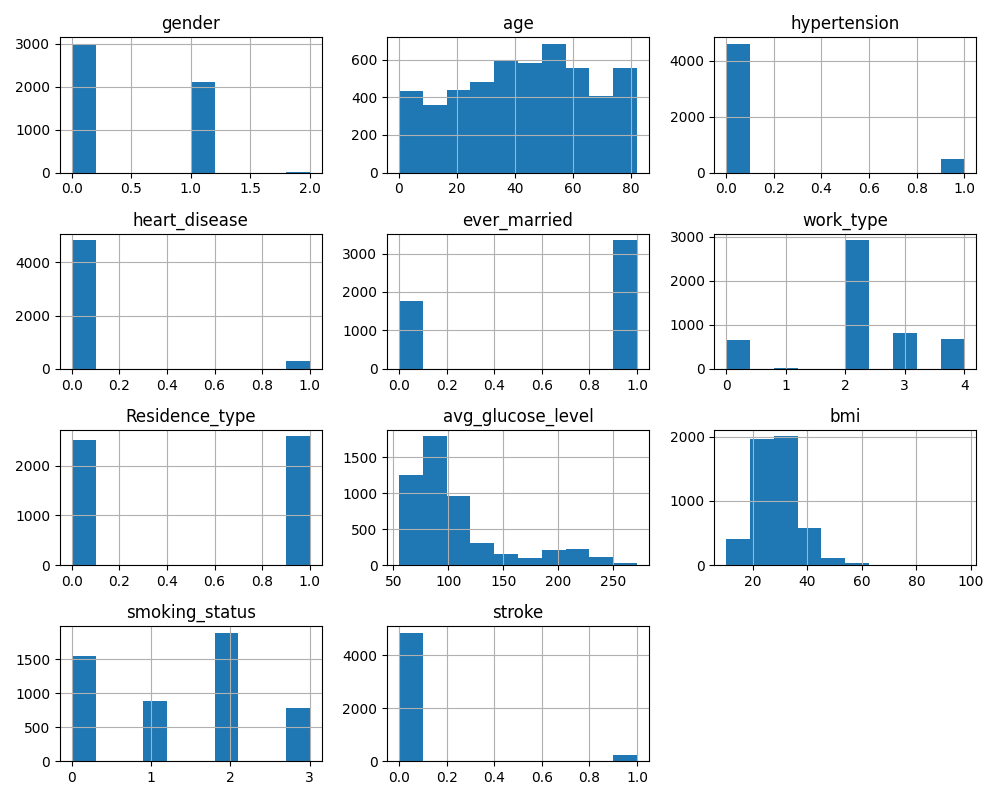
*Hình 2.5.2 – Phân bố nhãn sau xử lý bằng SMOTE*

Quan sát biểu đồ tương quan heatmap, có thể thấy một vài đặc trưng có mối liên hệ nhất định với biến mục tiêu stroke. Trong đó:

* Tuổi (age) có mối tương quan dương rõ ràng nhất với stroke so với các đặc trưng còn lại. Điều này phản ánh thực tế rằng người lớn tuổi thường có nguy cơ đột quỵ cao hơn. Đây là một đặc trưng đáng chú ý, cần được ưu tiên khi xây dựng mô hình dự đoán.
* Tăng huyết áp (hypertension) và bệnh tim (heart\_disease) cũng có mức tương quan dương với stroke. Dù không cao, nhưng vẫn thể hiện vai trò trong việc góp phần tăng nguy cơ đột quỵ. Hai đặc trưng này thường được xem là yếu tố nguy cơ điển hình trong y học.
* Mức đường huyết trung bình (avg\_glucose\_level) cũng có liên hệ với stroke. Mặc dù tương quan không mạnh, nhưng trong bối cảnh lâm sàng, chỉ số đường huyết cao thường liên quan đến bệnh lý tim mạch nên vẫn cần được cân nhắc.
* Các đặc trưng còn lại hầu như không cho thấy mối tương quan đáng kể với stroke trong biểu đồ này. Tuy nhiên, điều này không đồng nghĩa là chúng không có ý nghĩa, có thể các mối quan hệ phức tạp hơn mà biểu đồ tương quan tuyến tính không thể hiện được.



*Hình 2.5.3 – Phân bố dữ liệu đầu vào theo từng đặc trưng*

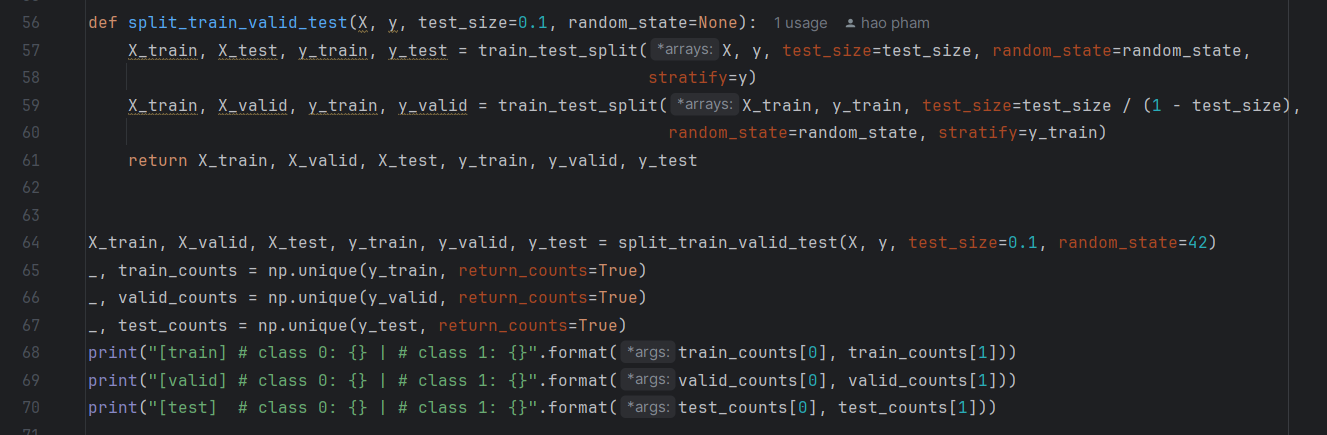


*Hình 2.5.4 – Mức độ tương quan giữa các thuộc tính trong dữ liệu*

### 2.5.2 Models

#### Phân chia tập dữ liệu

Chia bộ dữ liệu thành hai phần training data và test data nhằm đảm bảo tính công bằng trên mọi model:



#### Các mô hình học máy được áp dụng

* **Logistic Regression**

Logistic Regression là một thuật toán phân loại tuyến tính cơ bản thường được sử dụng trong các bài toán phân loại nhị phân. Mô hình dự đoán xác suất một đối tượng thuộc vào một trong hai nhóm, thông qua hàm sigmoid.

Trong dự án, biến mục tiêu **stroke** là một biến nhị phân (1: có bệnh, 0: không bệnh), do đó Logistic Regression là một lựa chọn phù hợp. Mô hình có ưu điểm là đơn giản, dễ cài đặt, tốc độ huấn luyện nhanh và có thể giải thích rõ ràng mối quan hệ giữa các đặc trưng đầu vào với đầu ra.

* **Decision Tree**

Decision Tree là mô hình phân loại dựa trên cây nhị phân, trong đó mỗi nút chia nhánh tương ứng với một điều kiện trên đặc trưng. Mô hình này dễ hiểu, dễ triển khai và có khả năng mô tả rõ ràng quá trình ra quyết định của hệ thống.

Trong dự án này, Decision Tree cho phép trực quan hóa cách các yếu tố như tuổi, huyết áp, BMI,... ảnh hưởng đến kết quả chẩn đoán bệnh đột quỵ.

* **Random Forest**

Random Forest là một mô hình tổ hợp (ensemble learning) sử dụng nhiều cây quyết định, mỗi cây được huấn luyện trên một tập con dữ liệu và tập đặc trưng khác nhau. Kết quả cuối cùng được tổng hợp (thường là voting đa số), giúp giảm thiểu hiện tượng overfitting mà các cây đơn lẻ thường gặp.

Mô hình này cải thiện độ chính xác đáng kể và rất phù hợp cho bài toán phân loại nhị phân như trong dự án đột quỵ.

* **Gaussian Naive Bayes (GNB)**

Gaussian Naive Bayes là một thuật toán phân loại xác suất dựa trên định lý Bayes, giả định rằng các đặc trưng là độc lập với nhau và tuân theo phân phối chuẩn (Gaussian). Mặc dù giả định này hiếm khi đúng hoàn toàn trong thực tế, nhưng GNB thường cho kết quả tốt với dữ liệu có độ nhiễu cao hoặc kích thước mẫu nhỏ.

Trong dự án, GNB được dùng như một mô hình baseline đơn giản, cho phép đánh giá nhanh hiệu quả phân loại. Với tốc độ huấn luyện nhanh và yêu cầu tài nguyên thấp, GNB phù hợp cho việc thử nghiệm ban đầu và so sánh với các mô hình phức tạp hơn như Random Forest hay XGBoost.

* **XGBoost**

XGBoost là một phiên bản tối ưu hóa của Gradient Boosting, với khả năng kiểm soát overfitting tốt hơn nhờ cơ chế regularization, tốc độ huấn luyện nhanh và khả năng xử lý dữ liệu thiếu.

XGBoost thường được sử dụng trong các cuộc thi học máy nhờ hiệu suất cao và tính ổn định. Trong dự án, mô hình này giúp tăng độ chính xác so với GBT thuần túy.

* **LightGBM**

LightGBM là một thuật toán boosting sử dụng kỹ thuật lấy mẫu một phía dựa trên gradient (GOSS), cho phép xử lý tập dữ liệu lớn nhanh hơn so với các phương pháp boosting truyền thống.

Mô hình này phù hợp với dữ liệu y tế lớn như dự án của nhóm và có thể được dùng để thay thế hoặc so sánh với XGBoost.

* **Support Vector Machine (SVM)**  
   SVM là một thuật toán phân loại mạnh mẽ, hoạt động bằng cách tìm siêu phẳng (hyperplane) tối ưu phân tách các điểm dữ liệu thuộc các lớp khác nhau với khoảng cách (margin) lớn nhất. SVM đặc biệt hiệu quả trong các bài toán phân loại nhị phân và hoạt động tốt khi dữ liệu không tuyến tính nhờ sử dụng các kernel (nhân) như RBF, polynomial,...

Trong dự án, SVM được sử dụng nhằm kiểm tra khả năng phân loại của mô hình trên tập dữ liệu y tế. Nhờ tính linh hoạt của kernel, SVM có thể mô hình hóa các mối quan hệ phi tuyến giữa các đặc trưng và nhãn đích.

#### Lựa chọn kỹ thuật đánh giá:

Để đảm bảo đánh giá khách quan hiệu suất của các mô hình, nhóm đã áp dụng kỹ thuật **Train-test split** để chia dữ liệu thành hai tập:

**-Tập huấn luyện (Train set)**: dùng để huấn luyện mô hình.

**-Tập kiểm tra (Test set)**: dùng để đánh giá độ chính xác và khả năng tổng quát hóa của mô hình.

Tùy theo loại mô hình và bản chất bài toán, các chỉ số đánh giá được lựa chọn khác nhau như sau:

* **Với mô hình Linear Regression**

Mặc dù đây không phải là mô hình phù hợp nhất với bài toán phân loại nhị phân, Linear Regression vẫn được dùng để kiểm tra mối tương quan giữa các biến đầu vào và biến mục tiêu. Các chỉ số đánh giá bao gồm:

**-Mean Squared Error (MSE)**: Đo lường mức độ sai lệch trung bình bình phương giữa giá trị dự đoán và thực tế.

**-R-squared (R²)**: Đo phần trăm phương sai của biến mục tiêu được giải thích bởi mô hình hồi quy.

* **Với các mô hình phân loại nhị phân**

Áp dụng cho các mô hình như:

-Logistic Regression

-Decision Tree / Random Forest

-**Gaussian Naive Bayes (GNB)**

-XGBoost

-LightGBM

-Multilayer Perceptron (MLP)

Các chỉ số đánh giá được sử dụng gồm:

-Accuracy: Tỷ lệ dự đoán đúng trên toàn bộ tập dữ liệu.

-Precision: Tỷ lệ các dự đoán dương tính đúng (True Positive) trên tổng số dự đoán dương tính.

-Recall (Sensitivity): Tỷ lệ các dự đoán đúng trong số các mẫu thực sự dương tính.

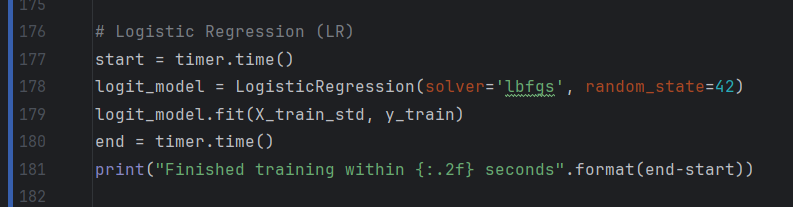
-F1-score: Trung bình điều hòa giữa precision và recall, đặc biệt hữu ích khi dữ liệu mất cân bằng giữa hai lớp.

-**Specificity**: Tỷ lệ dự đoán đúng các mẫu âm tính trong tổng số mẫu thực sự âm tính (True Negative rate), giúp đánh giá khả năng mô hình loại bỏ đúng các trường hợp không mắc bệnh.

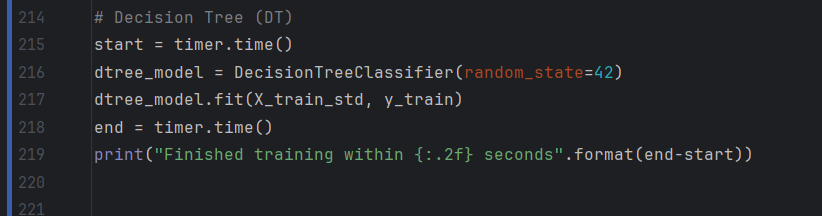
Việc sử dụng kết hợp nhiều chỉ số đánh giá giúp nhóm có cái nhìn toàn diện hơn về chất lượng của từng mô hình, từ đó lựa chọn mô hình tối ưu cho bài toán dự đoán bệnh tim mạch.

#### Huấn luyện mô hình:

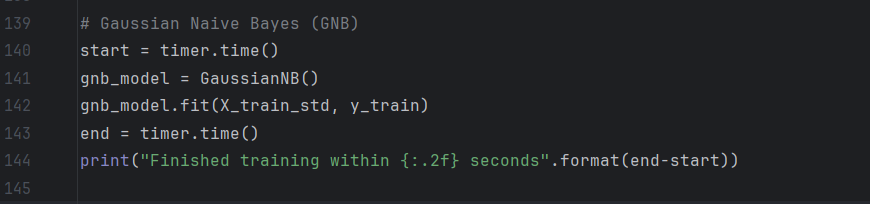
Logistic Regression:



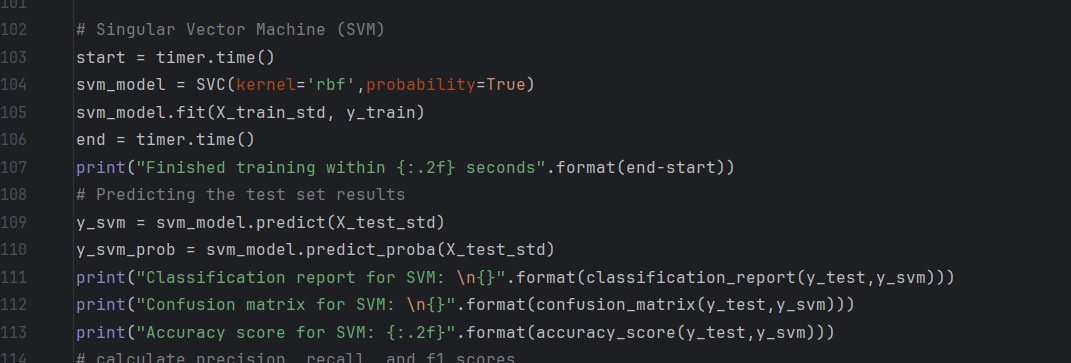
Decision Tree Classifier:



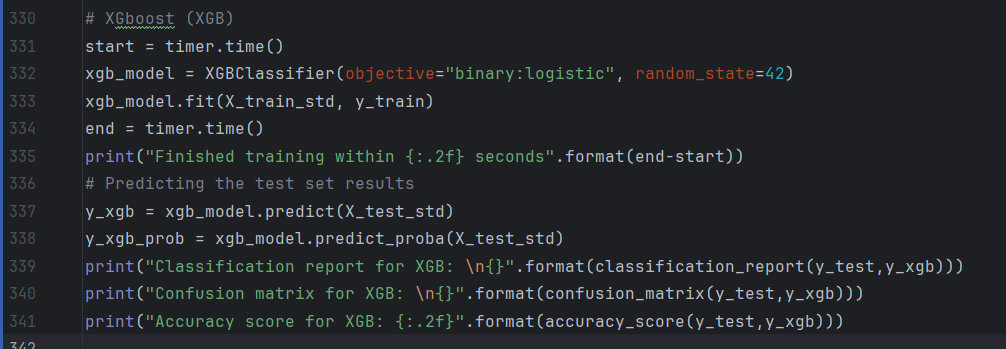
Gaussian Naive Bayes:



Linear SVM:



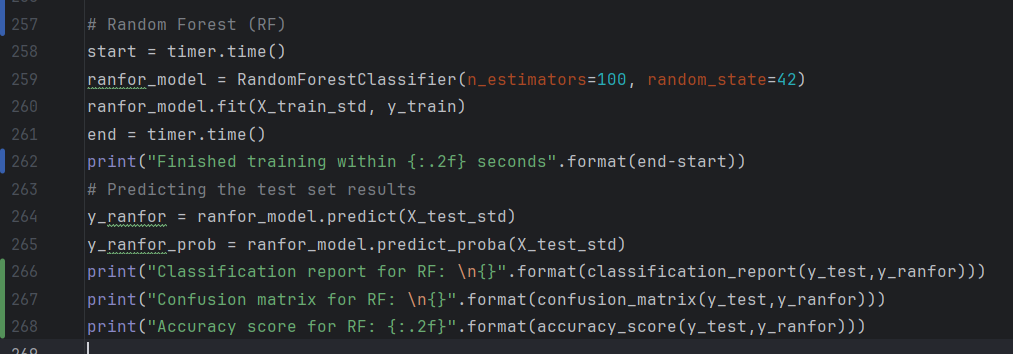
XGBoost:



LightGBM:



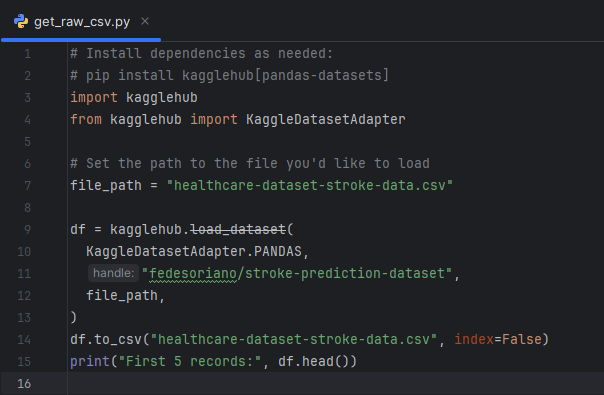
Random Forest:



# 3. Results

## 3.1. Data Ingestion Scripts and Code

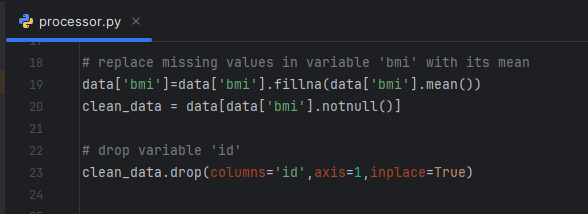
Để lấy được tập dữ liệu từ Kaggle, cần chỉ định tên file .csv (healthcare-dataset-stroke-data.csv) bên trong bộ dataset trên Kaggle. Tải dữ liệu với hàm load\_dataset() từ thư viện kagglehub, tham số KaggleDatasetAdapter.PANDAS sẽ giúp trả dữ liệu về dạng pandas.DataFrame. handle="fedesoriano/stroke-prediction-dataset" là ID của bộ dữ liệu trên Kaggle (gồm username và tên dataset). Sau khi tải, lưu dữ liệu từ biến df ra file healthcare-dataset-stroke-data.csv trong thư mục hiện tại. index=False sẽ không ghi cột chỉ mục vào file .csv



*Hình 3.1.1 – Code thực thi lấy dữ liệu từ Kaggle*

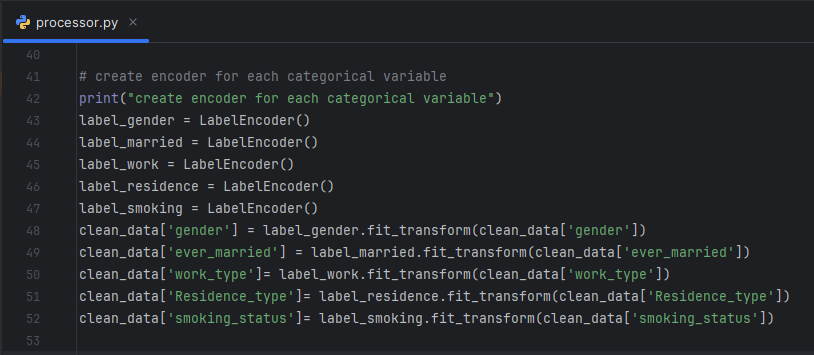
## 3.2. Data Transformation Scripts and Code

Thay các giá trị N/A trong cột bmi bằng giá trị trung bình (mean) của cột đó. Dữ liệu sau đó được lưu bằng biến clean\_data, rồi loại bỏ cột id. axis=1 chỉ định rằng cột sẽ bị xóa, inplace=True để thay đổi trực tiếp trên clean\_data mà không tạo bản sao mới.



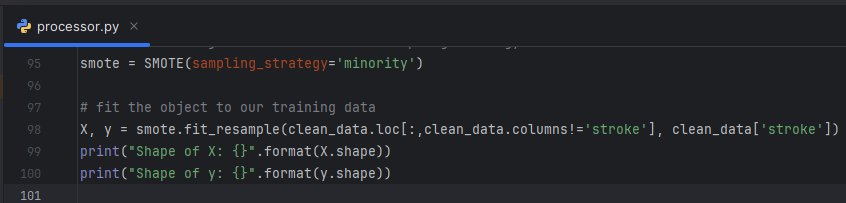
*Hình 3.2.1 – Xử lý giá trị thiếu và loại bỏ cột không cần thiết.*

Tạo các encoder riêng cho từng biến phân loại. Sau đó, thực hiện mã hóa từng biến phân loại bằng hàm fit\_transform. Hàm này sẽ lọc ra danh sách các nhãn duy nhất trong cột (fit), sau đó gán mỗi nhãn một số nguyên (transform).



*Hình 3.2.2 – Mã hóa các biến phân loại.*

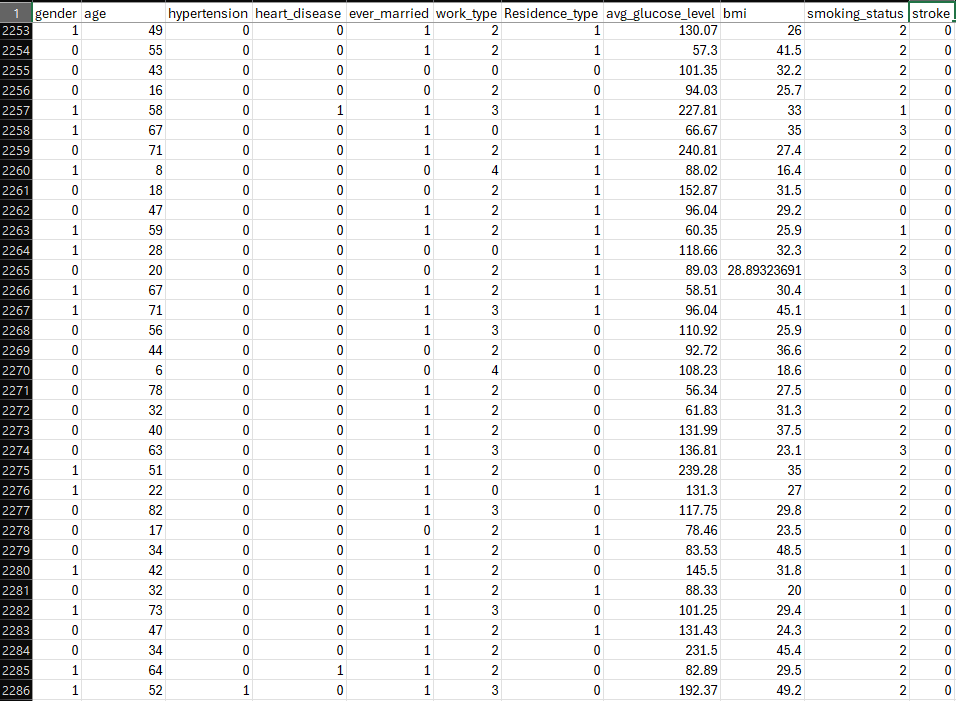
Tạo đối tượng smote từ thư viện, biến sampling\_strategy='minority' chỉ tạo thêm dữ liệu cho lớp thiểu số để cân bằng với lớp đa số.

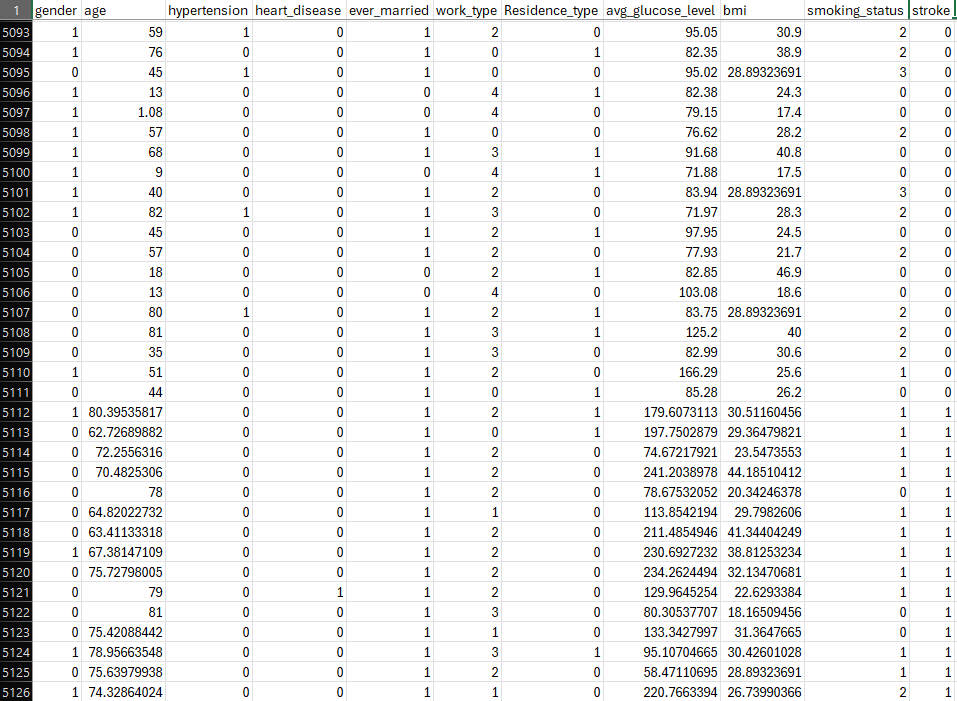


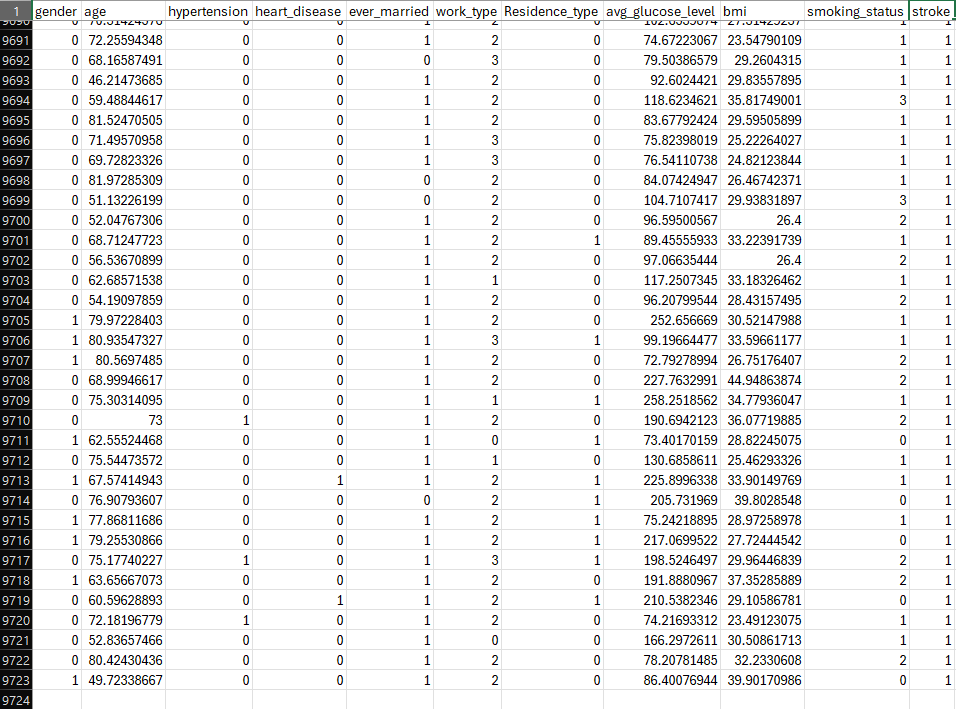
*Hình 3.2.3 – Xử lý mất cân bằng dữ liệu.*

## 3.3. Description and Sample of Transformed Datasets

Sau quá trình biến đổi, chất lượng dữ liệu đã được cải thiện đáng kể. Dữ liệu thiếu ở cột bmi đã được thay bằng giá trị trung bình của cột. Bên cạnh đó, các đặc trưng phân loại (gender, work\_type,..) cũng được mã hóa để phù hợp cho mô hình học máy. Tỷ lệ giữa 2 lớp stroke và no stroke cũng đã trở nên cân bằng hơn, giúp mô hình học tốt hơn thay vì bị thiên lệch về lớp chiếm đa số. Giờ đây, bộ dữ liệu đã sẵn sàng cho các thuật toán học máy.







*Hình 3.3.1 – Tập dữ liệu sau quá trình biến đổi và xử lý mất cân bằng.*

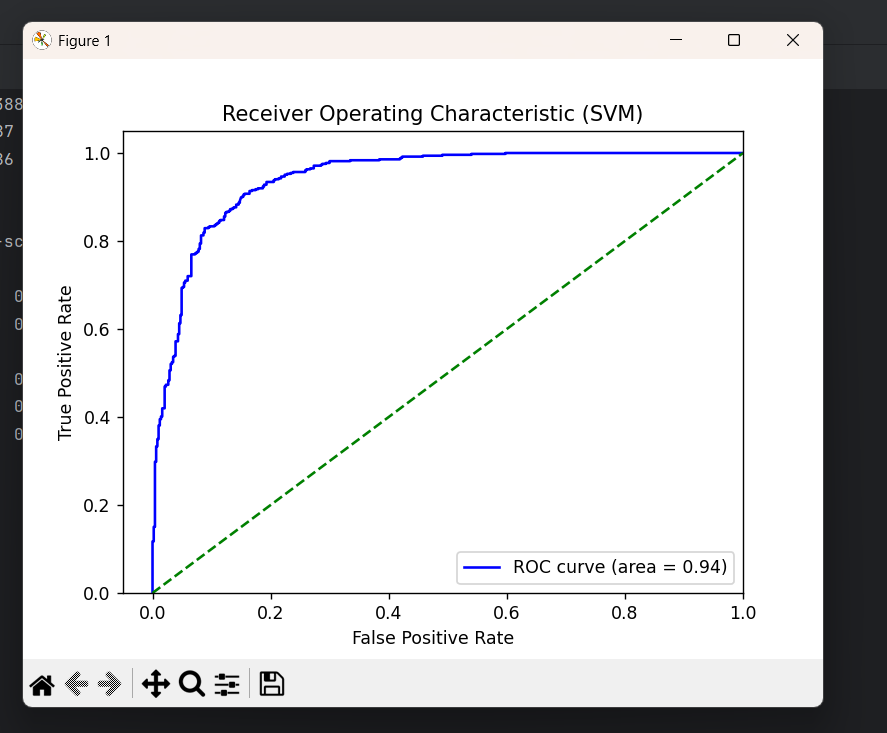
## 3.4. Data Visualization of Query Results

### 3.4.5 Lựa chọn mô hình:

Dựa vào kết quả huấn luyện mô hình:

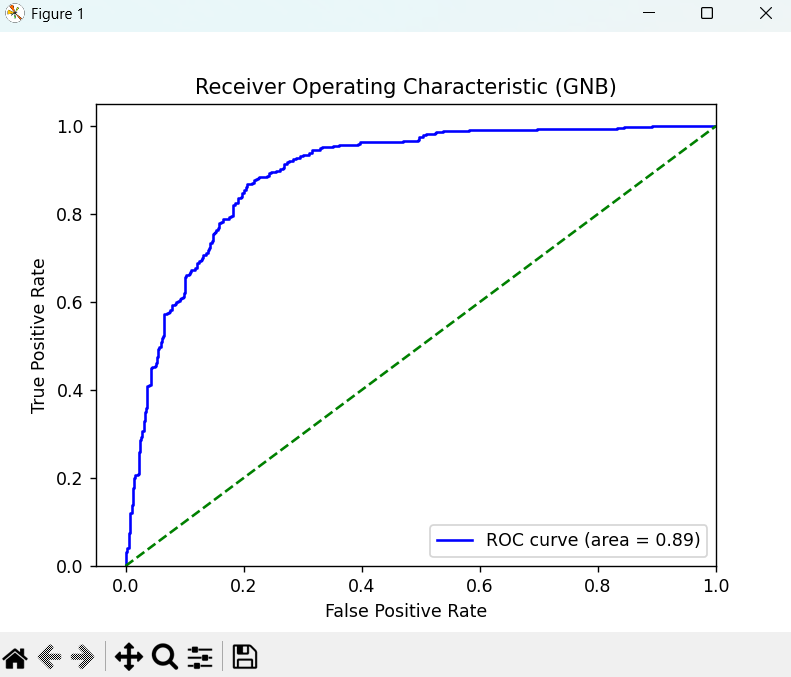
1. SVM

Diện tích dưới đường cong (AUC) đạt 0.94 chứng tỏ mô hình SVM có hiệu suất rất cao trong việc phân biệt giữa 2 lớp (dự đoán đúng người có nguy cơ đột quỵ và người không). Điều này nghĩa là mô hình có khả năng phân biệt tốt, ít nhầm lẫn giữa các lớp. Đường cong ROC nằm cao và lệch trái, thể hiện mô hình có tỉ lệ True Positive Rate cao trong khi giữ được False Positive Rate thấp, giúp phát hiện đúng phần lớn các trường hợp nguy cơ cao mà ít khi dự đoán nhầm. Với AUC = 0.94, SVM là mô hình rất tốt trong bài toán này.



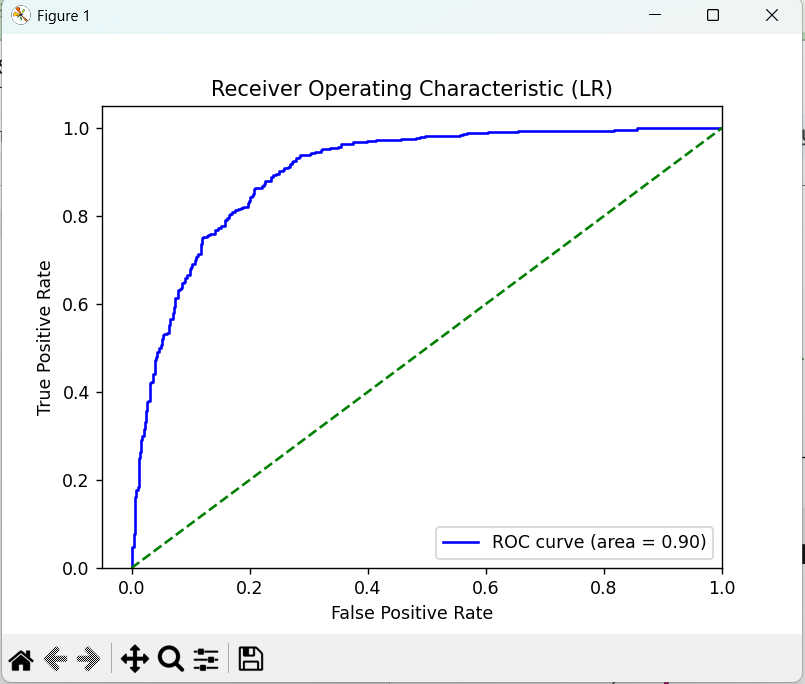
1. GNB

Giá trị AUC đạt 0.89 chứng tỏ GNB có hiệu quả tương đối tốt trong việc phân biệt hai lớp. Đây là con số khá tốt và cho thấy mô hình có khả năng nhận diện lớp mục tiêu đáng tín cậy. Tương tự mô hình SVM, đường cong ROC nằm cao nhưng hơi thoải, vì vậy mô hình có tỷ lệ dự đoán sai cao hơn SVM một chút.



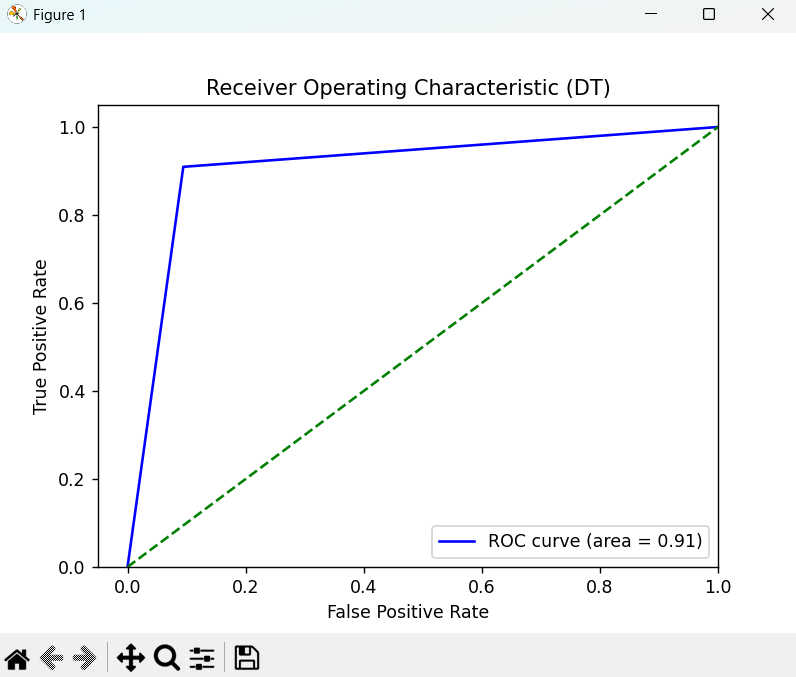
1. LR

Đường cong ROC nằm cao hơn rõ rệt so với đường chéo. Diện tích dưới đường cong (AUC) = 0.90, chứng minh rằng mô hình có khả năng phân biệt hai lớp (bị và không bị đột quỵ) rất tốt.



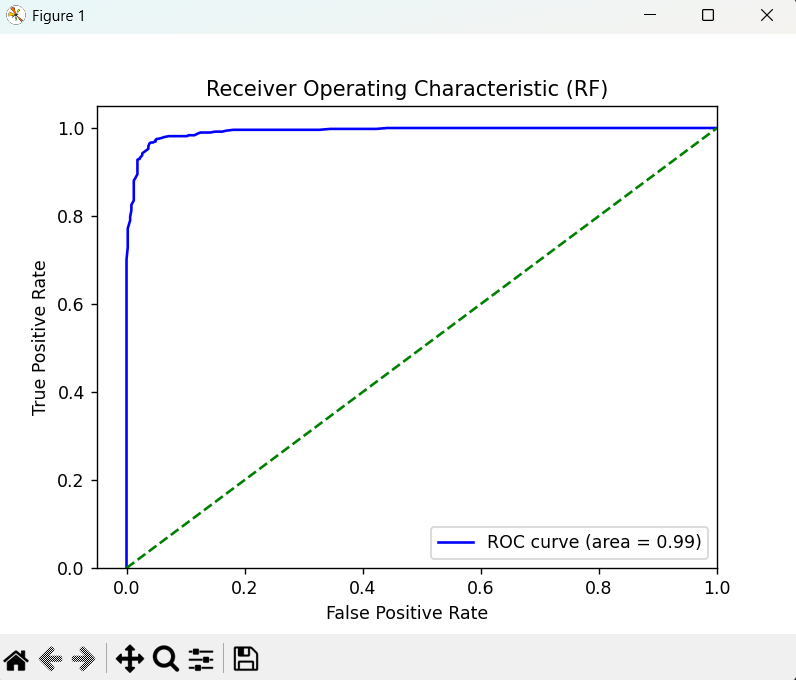
1. Decision Tree

Đường cong ROC trong mô hình này nhanh chóng đạt đến giá trị TPR rất cao ngay khi FPR còn thấp, điều này phản ánh khả năng nhận diện đúng các trường hợp dương tính rất tốt trong giai đoạn đầu. AUC = 0.91 là một con số ấn tượng, cho thấy mô hình có khả năng phân biệt mạnh mẽ giữa hai lớp. Đường cong không mượt như các mô hình tuyến tính do đặc trưng rời rạc của thuật toán Cây quyết định, nhưng vẫn đạt hiệu năng tổng thể cao.



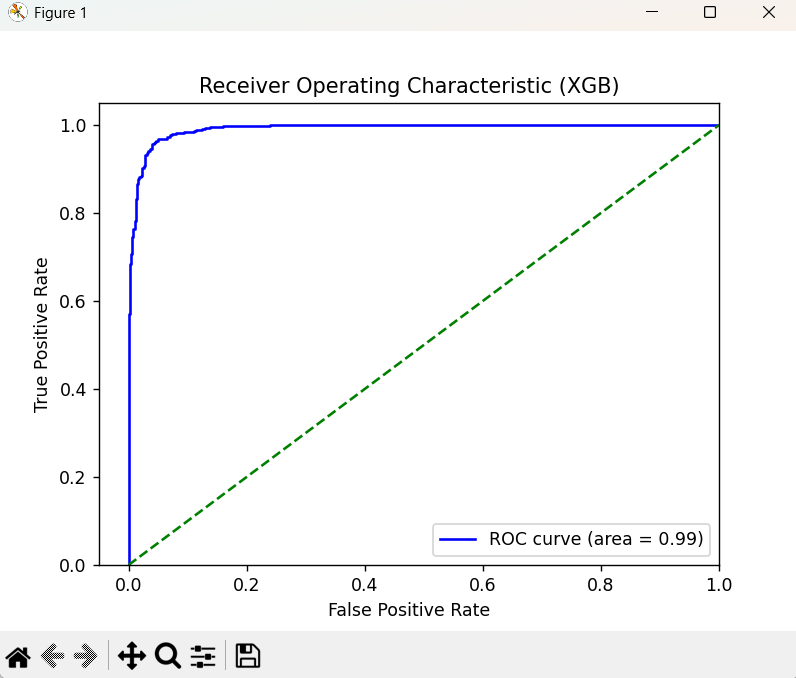
1. Random Forest

Đường ROC của mô hình nằm rất gần góc trên bên trái, cho thấy mô hình có khả năng phân loại rất tốt. Giá trị AUC là 0.99, gần như tuyệt đối, chứng tỏ mô hình có hiệu suất rất cao, khả năng dự đoán chính xác gần như hoàn hảo. Tuy nhiên, đây có thể là dấu hiệu của hiện tượng overfitting, cần kiểm tra thêm trên tập dữ liệu chưa thấy để xác định tính ổn định của mô hình.



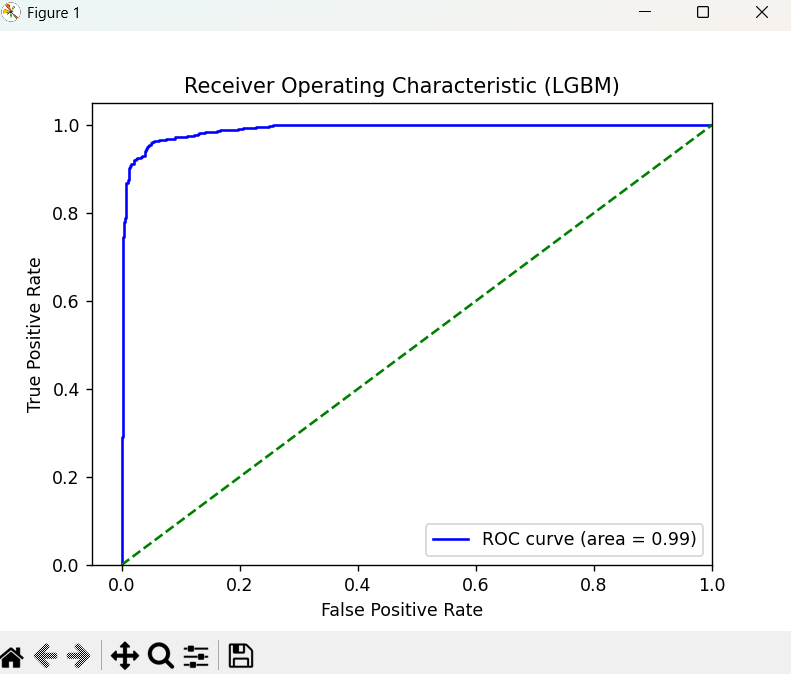
1. XGB

Tương tự như RF, mô hình này cũng có ROC = 0.99, rất cao. Vì thế, cần kiểm tra thêm các chỉ số khác như độ chính xác, độ nhạy và kiểm tra trên tập dữ liệu khác để đảm bảo không tính tổng quan và độ tin cậy, tránh overfỉtting.



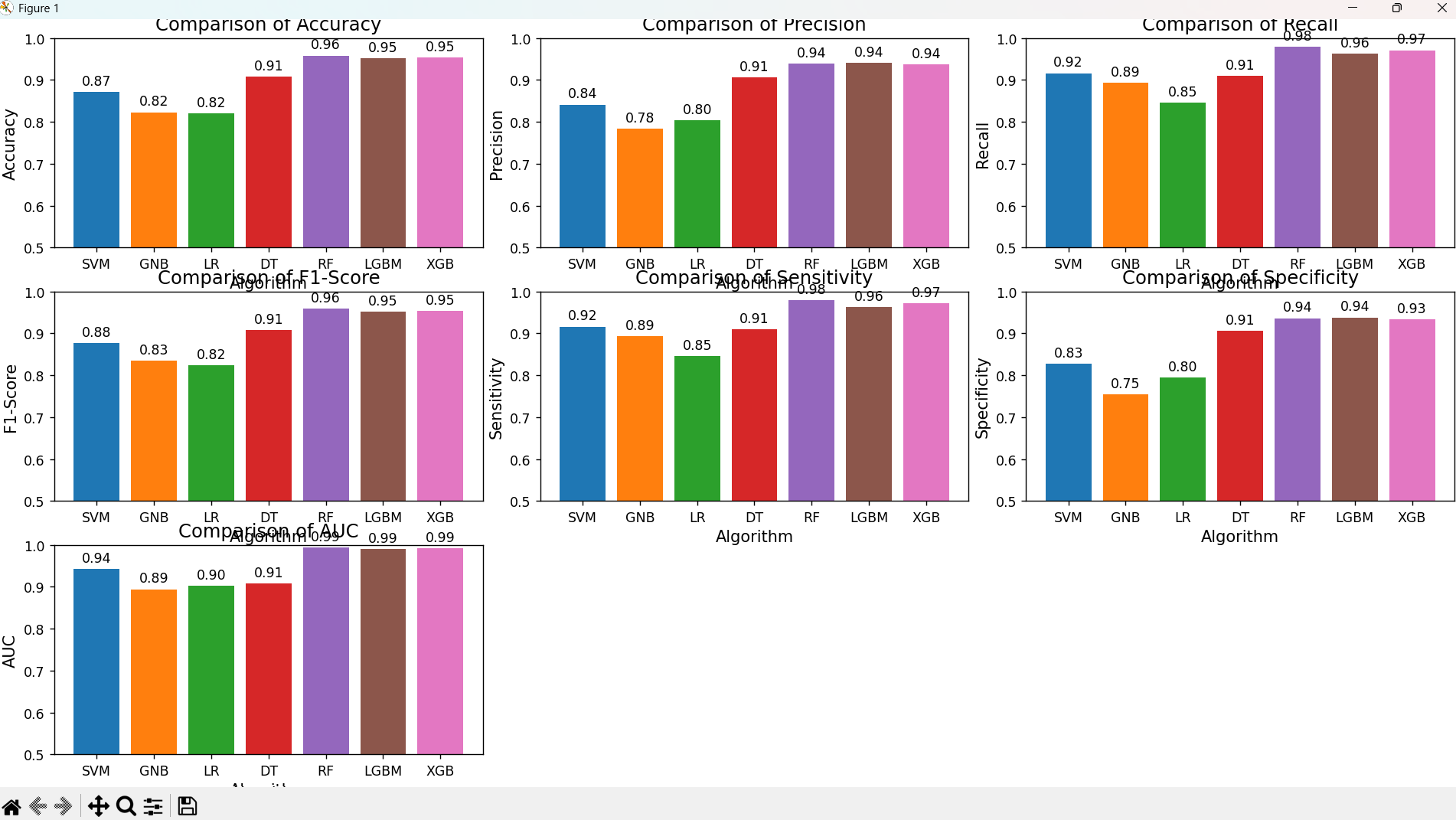
1. LGBM

Mô hình cũng đạt được kết quả huấn luyện rất tốt, thể hiện qua đường cong ROC và giá trị AUC rất cao, tương tự XGB và RF. Vì vậy, nên có hướng xử lý phù hợp để tránh overfitting.



1. Biểu đồ các chỉ số của 7 mô hình.

Dựa trên biểu đồ, ta thấy được 3 mô hình mạnh nhất là Random Forest (RF), LightGBM (LGBM),XGBoost (XGB). Cả ba mô hình đều cho kết quả vượt trội ở các chỉ số đánh giá (accuracy, precision, recall, specificity, F1-score, AUC). Đây là các mô hình nên ưu tiện lựa chọn cho bài toán phân loại này, vừa có độ chính xác, độ nhạy, độ đặc hiệu và khả năng phân biệt hai lớp rất cao.



| **Mô hình** | **Accuracy (%)** | **F1-score (%)** |
| --- | --- | --- |
| Logistic Regression | 82.0 | 82.40 |
| Decision Tree | 90.80 | 90.80 |
| Gradient Boosted Trees (GNB) | 82.30 | 83.50 |
| Linear SVM | 87.20 | 87.70 |
| LGBM | 95.10 | 95.10 |
| XGB | 95.30 | 95.40 |
| RD | 95.80 | 95.90 |

**Đánh giá:**

Mô hình tốt nhất (dựa trên AUC) là: RF, XGB, LGBM với AUC = 0.99.

Mô hình cho kết quả thấp nhất là: GNB vớiAUC = 0.89

**So sánh chi tiết các mô hình hàng đầu (Random Forest và XGBoost):**

**1. Hiệu suất:** Cả hai mô hình đều đạt AUC rất cao (≈ 0.99), cùng với F1-Score và Accuracy vượt trội.

**2. Sự khác biệt:**

- Random Forest dễ triển khai, ít cần tinh chỉnh siêu tham số, thích hợp với các dự án cần kết quả nhanh.

- XGBoost mạnh mẽ trong xử lý dữ liệu phức tạp và có khả năng chống overfitting tốt hơn.

**3. Lựa chọn:**

- Trong trường hợp này, Random Forest có thể được ưu tiên vì hiệu suất cực cao và đơn giản trong triển khai.

- Tuy nhiên, nếu muốn tinh chỉnh sâu và tối ưu thêm, XGBoost vẫn là lựa chọn rất cạnh tranh.

- Cả hai mô hình đều phù hợp để triển khai thực tế tùy theo yêu cầu về tốc độ, tài nguyên và độ phức tạp.



# 4. Projected Impact

## 4.1. Accomplishments and Benefits

Trình bày lợi ích (dựa vào mục tiêu) ít nhất 3 lợi ích thực tế

**Thành tựu đạt được**

* Xây dựng mô hình dự đoán nguy cơ đột quỵ bằng thuật toán học máy: Nhóm đã thử nghiệm và so sánh hiệu quả giữa các thuật toán như Logistic Regression, Random Forest, Support Vector Machine, XGBoost,... để lựa chọn mô hình tối ưu nhất cho tập dữ liệu thu thập được.
* Xử lý và chuẩn hóa dữ liệu đầu vào có độ tin cậy cao: Dữ liệu ban đầu bao gồm các thông tin y sinh học đa dạng như tuổi, giới tính, tiền xử bệnh nền (tim mạch, cao huyết áp), tình trạng hôn nhân, BMI, mức đường huyết trung bình trong máu, tình trạng hút thuốc, loại công việc, nơi cư trú,... đã được làm sạch, mã hóa, và chuyển đổi phù hợp với yêu cầu huấn luyện mô hình, đảm bảo tính chính xác và nhất quán.
* Đóng góp nguồn dữ liệu có giá trị cho nghiên cứu y tế ứng dụng AI: Bộ dữ liệu sau xử lý có thể tái sử dụng trong các đề tài khác về dự báo rủi ro sức khỏe, mang tính kế thừa và mở rộng cho nhiều hướng nghiên cứu liên ngành trong tương lai.

**Lợi ích mang lại**

* Đối với người dân, mô hình đóng vai trò như một công cụ cảnh báo sớm, giúp cá nhân hiểu rõ tình trạng sức khỏe của mình và có thể chủ động điều chỉnh thói quen sinh hoạt, hạn chế các yếu tố nguy cơ như tăng huyết áp, béo phì, hút thuốc,... từ đó giảm nguy cơ đột quỵ.
* Đối với tuyến y tế cơ sở, mô hình có thể hỗ trợ quá trình sàng lọc ban đầu, giúp phát hiện sớm các trường hợp cần theo dõi, đặc biệt tại những khu vực khó tiếp cận dịch vụ y tế chuyên sâu. Nhờ đó góp phần giảm gánh nặng cho bệnh viện tuyến trên.
* Đối với đội ngũ bác sĩ và nhà quản lý y tế, công cụ dự đoán có thể đóng vai trò như một kênh hỗ trợ ra quyết định – từ chẩn đoán ban đầu đến xây dựng phác đồ điều trị phù hợp, hoặc định hướng chính sách phòng chống đột quỵ dựa trên dữ liệu thực tế.
* Đối với lĩnh vực giáo dục và khoa học công nghệ, dự án là minh chứng cho việc ứng dụng thành công dữ liệu lớn và trí tuệ nhân tạo vào lĩnh vực y học, mở ra tiềm năng to lớn cho các nghiên cứu tiếp theo về bệnh lý thần kinh, tim mạch và chăm sóc sức khỏe cộng đồng.

## 4.2. Future Improvements

* Tăng cường thu thập thêm dữ liệu: Sử dụng các nguồn dữ liệu lớn hơn, bao gồm dữ liệu theo thời gian thực, dữ liệu từ nhiều quốc gia hoặc các bệnh viện khác nhau để tăng tính khái quát của mô hình.
* Thử nghiệm thêm những mô hình hiện đại hơn: Triển khai các mô hình mạnh hơn như XGBoost, LightGBM, hoặc các mô hình học sâu (Deep Learning) để tăng độ chính xác và khả năng phát hiện các mẫu phức tạp hơn trong dữ liệu.
* Kết hợp với các chỉ số sinh học thời gian thực (IoT): Hướng tới tích hợp với thiết bị theo dõi sức khỏe (smartwatch, cảm biến huyết áp...) để mô hình có thể dự đoán nguy cơ theo thời gian thực.

# 5. Team Member Review and Comment

|  |
| --- |

| NAME | REVIEW and COMMENT |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

# 6. Instructor Review and Comment

| CATEGORY | SCORE | REVIEW and COMMENT |
| --- | --- | --- |
| IDEA | \_\_/10 |  |
| APPLICATION | \_\_/30 |  |
| RESULT | \_\_/30 |  |
| PROJECT MANAGEMENT | \_\_/10 |  |
| PRESENTATION & REPORT | \_\_/20 |  |
| TOTAL | \_\_/100 |  |