

STROKE PREDICTION USING MACHINE LEARNING TECHNIQUES

Khổng Đức Quang - MSSV: 20225072

Ngày 3 tháng 12 năm 2024

TÓM TẮT

Con người ngày nay bị ảnh hưởng bởi nhiều loại bệnh tật do tác động của tình trạng môi trường và sự lựa chọn lối sống của con người. Việc phát hiện và lựa chọn các giải pháp phòng tránh bệnh là rất quan trọng để ngăn ngừa chúng tiến triển tới giai đoạn cuối. Đột quỵ là một trong những nguyên nhân gây tử vong hàng đầu và là gánh nặng tài chính cho người bệnh.

Chính vì lý do đó trên mà em quyết định lựa chọn chủ đề này cho học phần Project I. Để xử lý bài toán, em sử dụng bộ dữ liệu "Stroke Prediction Dataset" có sẵn trên Kaggle. Bộ dữ liệu bao gồm 5110 hàng và 12 cột thuộc tính, được xử lý trước để phù hợp với dự đoán

Danh sách các từ viết tắt

AI	Artificial Intelligence
DT	Decision Tree
KNN	K-Nearest Neighbor
ML	Machine Learning
RF	Random Forest
SVM	Super Vector Machine

1 Mô tả bộ dữ liệu

Bộ dữ liệu được em thu thập từ trang Web Kaggle để ước tính xem bệnh nhân có khả năng đột quỵ hay không. Đây là bộ dữ liệu về thông tin của 5110 người bao gồm 11 thuộc tính và 1 cột stroke(output) là có khả năng đột quỵ hay không. (FEDESORIANO 2021).

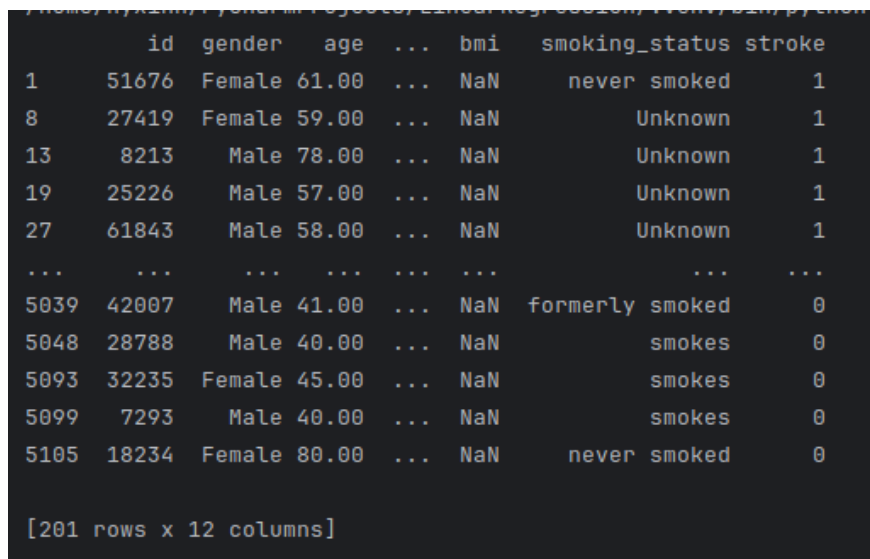
Dưới đây là danh sách các thuộc tính:

1. Id (Integer Feature): Đây là dữ liệu kiểu số nhằm. Tuy nhiên thuộc tính này không làm ảnh hưởng tới Output nên em sẽ không phân tích thêm.
2. Gender (Nominal Feature): Đây là thuộc tính kiểu chữ bao gồm các giá trị: Male, Female, Other. Thuộc tính "Gender"(giới tính) có thể ảnh hưởng tới khả năng đột quỵ do sự khác biệt sinh học, nội tiết tố, và yếu tố hành vi giữa nam và nữ.
3. Age (Integer Feature): Dữ liệu kiểu số. Thuộc tính này là một yếu tố quan trọng ảnh hưởng đến khả năng đột quỵ.
4. Hypertension (Integer Feature): Với hai giá trị khác nhau là: 0, 1. Đây là yếu tố nguy cơ lớn nhất đối với đột quỵ. Nó có ảnh hưởng sâu sắc đến khả năng đột quỵ.
5. Heart Disease (Integer Feature): Với hai giá trị khác nhau là: 0, 1. Là một yếu tố có nguy cơ đáng kể gây đột quỵ.
6. Ever married (Boolean Feature): Với hai giá trị khác nhau là: True, False. Có thể liên quan đến nguy cơ đột quỵ qua các yếu tố gián tiếp, chẳng hạn như lối sống, sức khỏe tâm lý, và sự hỗ trợ xã hội.
7. Work type (Nominal Feature): Có 5 loại giá trị khác nhau: Private, Self-employed, children, Govt-job, Never-worked. Và ảnh hưởng đáng kể đến nguy cơ đột quỵ thông qua các yếu tố như mức độ căng thẳng, hoạt động thể chất, và tiếp xúc với các yếu tố nguy cơ môi trường.
8. Residence type (Nominal Feature): Có hai giá trị khác nhau là: Urban, Rural. Ảnh hưởng đến nguy cơ đột quỵ thông qua các yếu tố môi trường, điều kiện sống, và khả năng tiếp cận dịch vụ chăm sóc sức khỏe.
9. Avg glucose level (Float Feature): Là một yếu tố quan trọng liên quan đến nguy cơ đột quỵ, đặc biệt thông qua mối liên hệ với bệnh tiểu đường và rối loạn chuyển hóa. Mức đường huyết bất thường, cả cao lẫn thấp, đều có thể làm tăng nguy cơ đột quỵ.
10. BMI (Float Feature): Là một chỉ số quan trọng để đánh giá mức độ béo phì hoặc thừa cân của một người, và nó có mối liên hệ mạnh mẽ với nguy cơ đột quỵ.
11. Smoking status(Nominal Feature): Với 4 giá trị khác nhau: Never smoked, Unknown, formerly smoked, smokes. Là một yếu tố nguy cơ quan trọng đối với nhiều bệnh lý, bao gồm đột quỵ. Hút thuốc lá ảnh hưởng đến sức khỏe của tim và mạch máu, gây ra những tác động tiêu cực trực tiếp làm tăng nguy cơ đột quỵ.

2 Xử lý dữ liệu

Trong quá trình tiền xử lý số liệu, em nhận thấy có các vấn đề sau cần xử lý:

1. Đầu tiên em **loại bỏ đi cột ID** do cột này không ảnh hưởng tới khả năng đột quỵ.
2. **Xử lý các giá trị NULL**



	id	gender	age	...	bmi	smoking_status	stroke
1	51676	Female	61.00	...	NaN	never smoked	1
8	27419	Female	59.00	...	NaN	Unknown	1
13	8213	Male	78.00	...	NaN	Unknown	1
19	25226	Male	57.00	...	NaN	Unknown	1
27	61843	Male	58.00	...	NaN	Unknown	1
...
5039	42007	Male	41.00	...	NaN	formerly smoked	0
5048	28788	Male	40.00	...	NaN	smokes	0
5093	32235	Female	45.00	...	NaN	smokes	0
5099	7293	Male	40.00	...	NaN	smokes	0
5105	18234	Female	80.00	...	NaN	never smoked	0
[201 rows x 12 columns]							

Hình 1: Kiểm tra giá trị NULL

Sau khi kiểm tra em nhận thấy cột thuộc tính "bmi" có 201 giá trị NULL. Để xử lý vấn đề này có ba phương pháp phổ biến: Xóa các dòng chứa giá trị NULL, Điền giá trị trung bình hoặc trung vị, Điền giá trị dựa trên mô hình (Model-based Imputation). Sau khi tìm hiểu ưu và nhược điểm của từng phương pháp em quyết định lựa chọn cách điền giá trị trung bình là lựa chọn vừa đơn giản vừa hiệu quả, giúp giữ được lại toàn bộ dữ liệu mà không làm giảm kích thước bộ dữ liệu.

3. Bước tiếp theo là **chuyển đổi dữ liệu kiểu Categorical thành dữ liệu kiểu Numerical**. Phương pháp Hash e Encoding và One-hot Encoding được sử dụng để thực hiện các bước chuyển đổi này.
4. **Phân chia dữ liệu**: Em chia bộ dữ liệu thành hai tập là: tập huấn luyện(80%), tập kiểm tra(20%) bằng cách sử dụng hàm `train_test_split` được cung cấp trong thư viện `scikit-learn`
5. **Feature scaling (chuẩn hóa đặc trưng)** là bước cuối cùng