**TRƯỜNG ĐẠI HỌC CÔNG THƯƠNG**

**THÀNH PHỐ HỒ CHÍ MINH**

⎯⎯⎯℘\*\*\*℘⎯⎯⎯



**TIỂU LUẬN**

**HỌC MÁY**

**DỰ ĐOÁN BỆNH ĐỘT QUỴ (STROKE)**

**NHÓM THỰC HIỆN:** NHÓM 17

**THÀNH VIÊN NHÓM:**

TRẦN CHÍ KHANH - 2001210862(NT)

NGUYỄN TẤN PHÁT – 2001210345

NGUYỄN THẢO TRÂM – 2001210216

**LỚP: 12DHTH16**

**GIẢNG VIÊN:**TRẦN ĐÌNH TOÀN

**Bản phân Công**

|  |  |  |
| --- | --- | --- |
| **Mã số sinh viên** | **Họ và tên** | **Công Việc** |
| **2001210862(NT)** | **Trần Chí Khanh** | Tạo model |
| **2001210345** | **Nguyễn Tấn Phát** | Tạo ứng dụng trên web |
| **2001210216** | **Nguyễn Thảo Trâm** | Tìm hiểu thuật toán |

**Lời mở đầu**

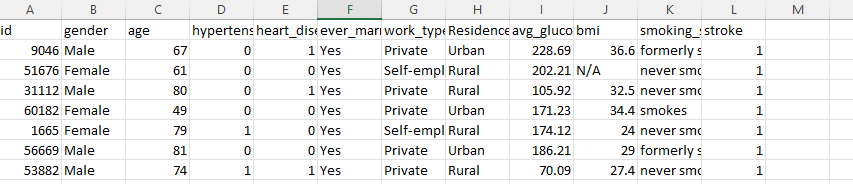
Trong thời đại số hóa ngày nay, dữ liệu ngày càng trở thành nguồn tài nguyên quý báu nhất, từ thông tin cá nhân trên mạng xã hội đến dữ liệu sản xuất trong các nhà máy. Tuy nhiên, dữ liệu này thường rất lớn và phức tạp, gần như không thể được hiểu và khai thác một cách truyền thống.

Ở trong tình hình đó, Máy Học (Machine Learning) đã trở thành một công cụ mạnh mẽ để khám phá, phân tích và rút ra những thông tin quý giá từ dữ liệu. Máy Học không chỉ là một lĩnh vực nghiên cứu, mà còn là một công cụ hữu ích trong nhiều ứng dụng thực tiễn, từ dự đoán thị trường tài chính, xử lý ngôn ngữ tự nhiên đến tự động lái xe.

**Mục lục**

**I. Xử lí bộ dữ liệu**

**1.Phân tích dữ liệu bệnh đột quỵ**

****

Bộ dữ liệu có tổng là 5110 dòng.

**Bộ dữ liệu gồm 12 cột bao gồm:** id , giới tính, tuổi, tăng huyết áp,bệnh tim, tình trạng hôn nhân, công việc, Nơi cư trú, lượng đường tb, chỉ số khối cơ thể, đã từng hút thuốc, đột quỵ.

Trong giới tính thì 2 giá trị nam, nữ.

Tuổi.

Có bệnh tim hay không: 0, 1.

Tình trạng hôn nhân đã kêt hôn và chưa kết hồn

Công việc : tư nhân, tự kinh doanh , nhà nước , trẻ em , chưa từng làm.

Nơi cư trú: thành thị hay nông thôn.

Lượng đường trung bình trong máu.

Chỉ số cơ thể.

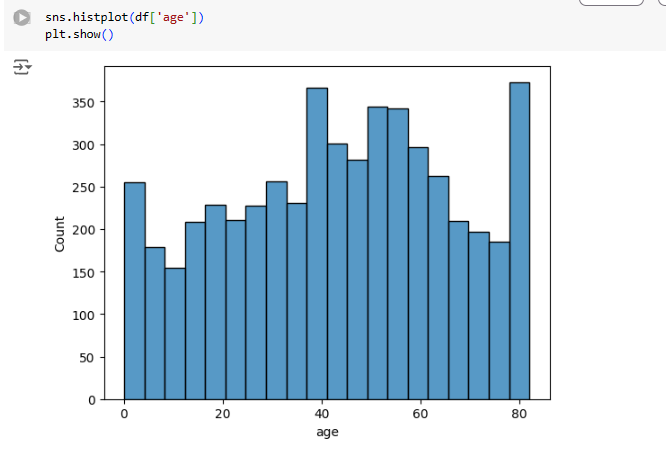
Tình trạng hút thuốc : không biết , không bao giờ hút , trước đây hút, đang hút.

Kết quả có bệnh đột quỵ hay không: 0, 1.

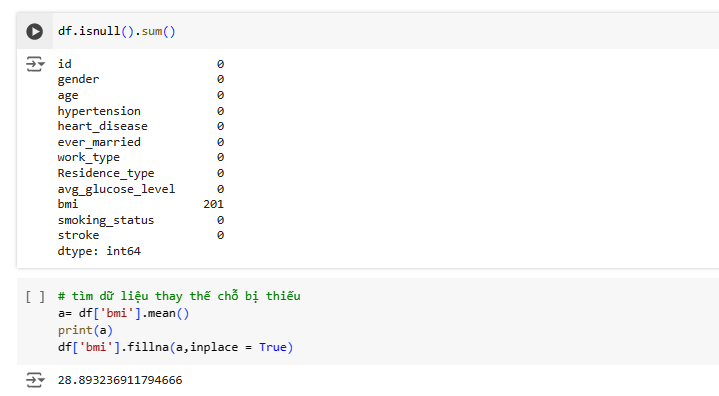
**2. Trực quan hóa dữ liệu**

**2.1 Trực quan bằng biểu đồ.**

Bộ dữ liệu có tổng là 5110 dòng.

  
bảng xem nhóm độ tuổi mắc bệnh đột quỵ cao.

**2.2 Xử lí các dữ liệu bị thiếu.**

Tiếp theo là tìm xem các dữ liệu bị thiếu và thay nó các dữ liệu đó. Dữ bị thiếu là chỉ số cơ thể bmi. Để Thuật toán có thể dự đoán tốt hơn và chính xác thì phải thay thế dữ liệu bị thiếu đó.

**2.3 Tiền xử lí dữ liệu**



Bước xử lí dữ liệu trước khi áp dụng thuật toán. Chuẩn hóa dữ liệu từ dạng chữ về dạng số để thuật toán dễ dàng xử lí và phân lớp.

**II. Thuật toán**

**1. Multinomial Naive Bayes là gì?**

Naive Bayes là một họ thuật toán xác suất dựa trên Định lý Bayes. Nó "ngây thơ" vì nó giả định tính năng độc lập, có nghĩa là sự hiện diện của một tính năng không ảnh hưởng đến sự hiện diện của một tính năng khác (điều này có thể không đúng trong thực tế).

**Đa thức Naive Bayes** là một bộ phân loại xác suất để tính toán phân phối xác suất của dữ liệu văn bản, điều này làm cho nó rất phù hợp với dữ liệu có các tính năng đại diện cho tần số rời rạc hoặc số lượng sự kiện trong các tác vụ xử lý ngôn ngữ tự nhiên (NLP) khác nhau.

**Phân bố đa thức**

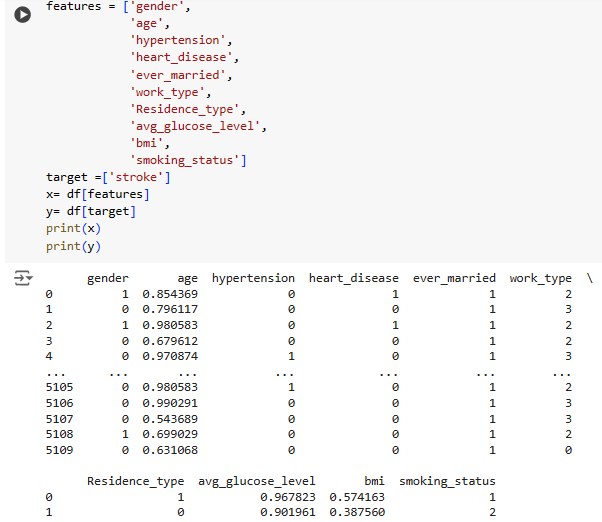
Thuật ngữ "đa thức" dùng để chỉ loại phân phối dữ liệu được giả định bởi mô hình. Các tính năng trong phân loại văn bản thường là số lượng từ hoặc tần số thuật ngữ. Phân bố đa thức được sử dụng để ước tính khả năng nhìn thấy một tập hợp số từ cụ thể trong tài liệu.

**2. Lợi ích của việc sử dụng Multinomial Naive Bayes**

Có một số lợi ích của việc sử dụng Multinomial Naive Bayes được thảo luận dưới đây:

1. **Hiệu quả:** Multinomial NB có hiệu quả tính toán và có thể xử lý các bộ dữ liệu lớn với nhiều tính năng khiến nó trở thành lựa chọn thiết thực cho các tác vụ phân loại văn bản như phát hiện spam, phân tích tình cảm và phân loại tài liệu trong đó các tính năng thường dựa trên số lượng.
2. **Dễ dàng thực hiện:** Multinomial NB rất đơn giản để thực hiện và yêu cầu điều chỉnh siêu tham số tối thiểu giúp người mới bắt đầu có thể truy cập và triển khai nhanh chóng.
3. **Các tính năng mạnh mẽ đến không liên quan:** Giả định "ngây thơ" về độc lập có điều kiện có thể làm cho Multinomial NB mạnh mẽ đối với các tính năng hoặc nhiễu không liên quan trong dữ liệu.
4. **Kết quả có thể diễn giải:** Các xác suất được tạo ra bởi MNB có thể cung cấp cái nhìn sâu sắc về khả năng của một tài liệu thuộc về một lớp cụ thể. Nó có thể phục vụ như một mô hình cơ sở vững chắc cho các nhiệm vụ phân loại văn bản, dựa vào đó các mô hình phức tạp hơn có thể được so sánh.

**III. Xây dựng mô hình**

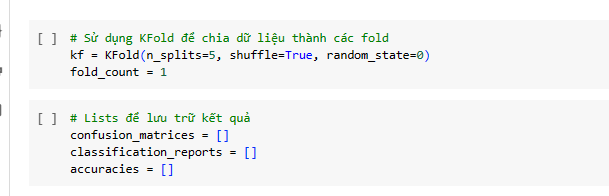
****

Thực hiện xây dựng bô mô hình kiểm thử.

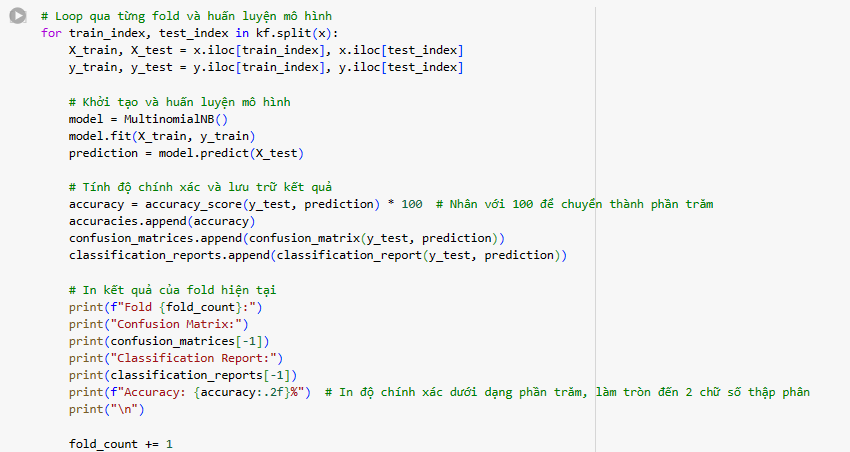
Xác định các thuộc tính độc lập và thuộc tính phụ thuộc:

Thuộc tính độc lập:tuổi, giới tính, cao huyết áp , bệnh tim, tình trạng hôn nhân, công việc, nơi cư trú, chỉ số đường trung bình trong máu, chỉ số cơ thể,

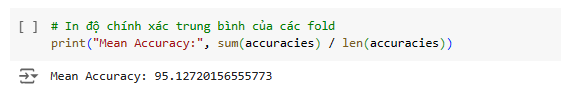
Thuộc tính phụ thuộc: Đột quỵ.



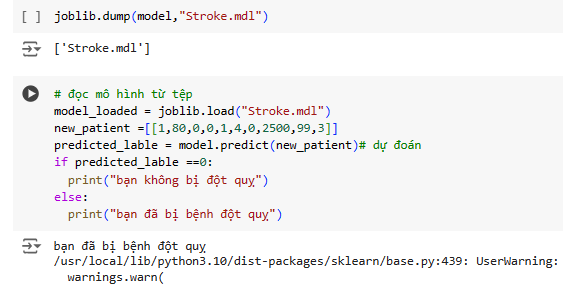
Sử dụng kfole để chia dữ liệu thành từng fold để quan sát. Tổng là 5 fold và random\_state= 0. Cần có list để lưu trữ kết quả của các thông số.



Tạo vòng lập for để xem thông tin ở từng fold nhầm đánh giá chính xác việc chia dữ liệu ở từng fold mà có thể tính toán độ chính xác trung bình của tổng các fold.



Trung bình các fold đạt 95.12 % là một số khá cao cũng vậy sẻ bắt đầu vào bước lưu model để đưa vào ứng dụng.



Lưu mô hình bằng thư viên joblib để lưu mô hình với đuôi chấm mdl. Sau cùng là kiểm tra lại mô hình lần cuối trước khi đưa vào trang web.

**Tài liệu tham khảo**

[Multinomial Naive Bayes - GeeksforGeeks](https://www.geeksforgeeks.org/multinomial-naive-bayes/)

[Stroke Prediction Dataset (kaggle.com)](https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset)