

**STROKE PREDICTION**

**PROJECT MÔN HỌC**

Học máy và Khai phá dữ liệu

**NHÓM THỰC HIỆN**

Nhóm 9:

Nguyễn Văn Thế

Nguyễn Thị Kim Duyên

Lê Trung Hiếu

LINK PROJECT

[https://colab.research.google.com/drive/1asvqBesvonGmMEm7h40BIsmBnYek70r5#scrollTo=JVCyxkB2u7UF](https://colab.research.google.com/drive/1asvqBesvonGmMEm7h40BIsmBnYek70r5%23scrollTo=JVCyxkB2u7UF)

*Hà Nội, tháng 7 năm 2021*

Mục lục

[**Mở đầu** 3](#_Toc77635607)

[**I.** **Tổng quan dự án** 4](#_Toc77635608)

[***1.*** ***Phát biểu bài toán*** 4](#_Toc77635609)

[***2.*** ***Mô hình giải quyết bài toán*** 4](#_Toc77635610)

[**II.** **Phân tích dữ liệu** 5](#_Toc77635611)

[**1.** **Tập dữ liệu đầu vào** 5](#_Toc77635612)

[**1.1.** **Thu thập dữ liệu** 5](#_Toc77635613)

[**1.2.** **Các trường dữ liệu** 5](#_Toc77635614)

[**1.3.** **Mục đích phân tích dữ liệu** 7](#_Toc77635615)

[**2.** **Phân tích biến đơn** 7](#_Toc77635616)

[**2.1.** **Biến định lượng** 7](#_Toc77635617)

[**2.2.** **Biến định tính** 9](#_Toc77635618)

[**2.3.** **Các thông tin khác về tập dữ liệu** 10](#_Toc77635619)

[**3.** **Phân tích sự phụ thuộc giữa các biến** 10](#_Toc77635620)

[**3.1.** **Giữa các biến và biến target** 10](#_Toc77635621)

[**3.2.** **Giữa các biến** 11](#_Toc77635622)

**Mở đầu**

Thưa bạn đọc,

Trong thời gian học tập tại HUS lớp cao học chuyên ngành khoa học dữ liệu, nhóm nghiên cứu nhận thấy việc phân tích, khai phá dữ liệu và tự động hóa mô hình như học máy đã mang đến những kiến thức rất hữu ích, áp dụng rộng rãi được trong cuộc sống.

Khoa học công nghệ ngày càng phát triển, đặc biệt trong lĩnh vực học máy và khai phá dữ liệu. Điều này đã giúp con người càng ngày càng sáng tạo không ngừng, phát triển các sản phẩm công nghệ mang tính đột phá, ứng dụng thực tế trong đời sống con người và hỗ trợ con người một cách tối đa. Một trong những thành tựu nổi bật và hữu ích nhất là ứng dụng khoa học công nghệ trong y tế liên quan đến việc chuẩn đoán bệnh cho con người, từ đó đưa ra các phác đồ điều trị chính xác, giúp tăng cường chăm sóc sức khỏe con người.

Dự án này là một nỗ lực của nhóm trong việc nghiên cứu dữ liệu liên quan đến chuẩn đoán bệnh đột quỵ ở con người, tổng hợp các thao tác phân tích dữ liệu, cũng như là áp dụng các mô hình học máy để phân loại, xử lý dữ liệu, tìm ra các kết luận đúng đắn, đem lại một sản phẩm chính xác trong việc chuẩn đoán sức khỏe con người.

Hà Nội, tháng 7 năm 2021

Nhóm dự án Stroke Prediction Project

1. **Tổng quan dự án**
   1. ***Phát biểu bài toán***

Đột quỵ là căn bệnh nghiêm trọng ở con người, là sự tổn thương đến não xảy ra khi dòng máu cung cấp cho não bị gián đoạn hoặc giảm đáng kể. Não bị thiếu ô-xy và dinh dưỡng và các tế bào não bắt đầu chết trong vòng vài phút. Vì lý do đó, đột quỵ được coi là một tình huống cấp cứu y tế và cần có chẩn đoán và điều trị kịp thời.

Đột quỵ, thường gọi là “tai biến mạch máu não”, được ví như "sát thủ giấu mặt", là nguyên nhân tử vong đứng thứ 3 chỉ sau ung thư và bệnh tim mạch nhưng đứng hàng đầu về tỉ lệ tàn tật. Nếu trước đây đột quỵ thường gặp ở người lớn tuổi thì nay bệnh nhân của căn bệnh này dần dần trẻ hóa. Theo thống kê của Bộ y tế, tỉ lệ tử vong và tàn phế do đột quỵ ở Việt Nam chỉ tính trong năm 2019 là ~ 200,000 người bị đột quỵ; 11,000 người tử vong và hơn 100,000 bệnh nhân tàn phế. Xu hướng đột quỵ trước đây là tuổi thường gặp trên 60 tuổi, còn ngày nay căn bệnh này có xu hướng trẻ hóa, có lúc xảy ra với người trẻ dưới 16 tuổi. Tỷ lệ tái phát đột quỵ là 15-40%.

Việc xác định đúng nguyên nhân căn bản và vị trí của đột quỵ sẽ quyết định việc điều trị. Công nghệ y tế tiên tiến đã làm tăng rõ rệt khả năng chẩn đoán chính xác đột quỵ và đánh giá tổn thương đến não bộ. Tuy nhiên, không phải lúc nào cũng dễ nhận biết các cơn đột quỵ nhỏ vì triệu chứng có thể bị bỏ qua bởi bệnh nhân và gia đình do hiểu nhầm đó là do lão hóa, hoặc có thể bị nhầm lẫn với các triệu chứng của các bệnh thần kinh khác. Như đã lưu ý ở trên, bất kỳ dấu hiệu cảnh báo đột quỵ nào cần cũng cần phải đánh giá y tế ngay lập tức. Như là các chuyên gia y tế sẽ nói với bạn, “càng mất nhiều thời gian, tổn thương não càng nghiêm trọng hơn”.

Nhóm yếu tố nguy cơ hàng đầu gây đột quỵ não là tăng huyết áp, đái tháo đường, rối loạn lipid máu. Một số yếu tố khác là béo phì, hút thuốc lá, người bệnh tim mạch, nghiện rượu bia, từng có cơn thiếu máu não thoáng qua...Điều nguy hiểm là những thủ phạm chính gây đột quỵ như tăng huyết áp, đái tháo đường đòi hỏi phải dùng thuốc kiểm soát lâu dài, hầu như phải điều trị suốt đời.

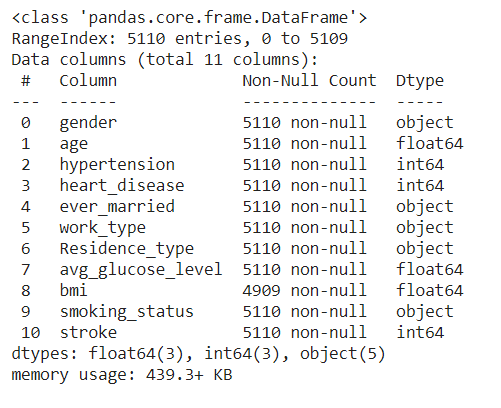
Từ tính chất nghiêm trọng của căn bệnh, nhóm nghiên cứu muốn áp dụng công nghệ trong việc phân tích khai phá dữ liệu và áp dụng các mô hình học máy để giúp dự đoán được căn bệnh đột quỵ này ở con người. Nhóm sử dụng bộ dữ liệu Stroke Prediction trên Kaggle làm bộ dữ liệu phân tích khai phá và học máy.

* 1. ***Mô hình giải quyết bài toán***
     + *Mô hình*

1. **Phân tích dữ liệu**
   1. **Tập dữ liệu đầu vào**
   2. **Thu thập dữ liệu**

* Bộ dữ liệu này được sử dụng để dự đoán liệu một bệnh nhân có khả năng bị đột quỵ hay không dựa trên các thông số đầu vào như giới tính, tuổi tác, các bệnh khác nhau và tình trạng hút thuốc. Mỗi hàng trong bộ dữ liệu cung cấp thông tin liên quan về bệnh nhân.
* Bộ dữ liệu được thu thập từ trang web: <https://www.kaggle.com/fedesoriano/stroke-prediction-dataset>
* Bộ dữ liệu gồm 5110 quan sát với 12 thuộc tính
  1. **Các trường dữ liệu**
* id:
  + Là số nhận dạng (ID) của bệnh nhân
  + Là biến định lượng liên tục kiểu số
* gender:
  + Giới tính bệnh nhân
  + Là biến định tính rời rạc kiểu String
  + Nhận 1 trong các giá trị: "Male", "Female" or "Other"
* age:
  + Tuổi của bệnh nhân
  + Là biến định lượng liên tục kiểu số
* hypertension:
  + Chứng cao huyết áp của bệnh nhân
  + Là biến định tính rời rạc kiểu số
  + Nhận 1 trong 2 giá trị: 0 nếu bệnh nhân không bị cao huyết áp và 1 nếu bệnh nhân mắc cao huyến áp
* heart\_disease:
  + Tình trạng đau tim của bệnh nhân
  + Là biến định tính rời rạc kiểu số
  + Nhận 1 trong 2 giá trị: 0 nếu bệnh nhân không bị đau tim, và 1 nếu bệnh nhân bị đau tim
* ever\_married:
  + Tình trạng hôn nhân của bệnh nhân
  + Là biến định tính rời rạc kiểu String
  + Nhận 1 trong các giá trị: "No" nếu chưa kết hôn hoặc "Yes" nếu đã kết hôn
* work\_type:
  + Loại nghề nghiệp của bệnh nhân
  + Là biến định tính rời rạc kiểu String
  + Nhận 1 trong các giá trị: "children", "Govt\_jov", "Never\_worked", "Private" hoặc "Self-employed"
* Residence\_type:
  + Nơi cư trú của bệnh nhân
  + Là biến định tính rời rạc kiểu String
  + Nhận 1 trong các giá trị: "Rural" hoặc "Urban"
* avg\_glucose\_level:
  + Lượng đường trung bình trong máu của bệnh nhân
  + Là biến định lượng liên tục kiểu số
* bmi:
  + Chỉ số đo lường cơ thể BMI của bệnh nhân
  + Là biến định lượng liên tục kiểu số
* smoking\_status:
  + Tình trạng hút thuốc của bệnh nhân
  + Là biến định tính rời rạc kiểu String
  + Nhận 1 trong các giá trị: "formerly smoked", "never smoked", "smokes" or "Unknown"
* stroke:
  + Tình trạng đột quỵ của bệnh nhân
  + Là biến định tính rời rạc kiểu số
  + Nhận 1 trong 2 giá trị: 0 nếu bệnh nhân không bị đột quỵ, và 1 nếu bệnh nhân bị đột quỵ

Sử dụng thư viện Pandas trong Python để tìm hiểu tập dữ liệu Stroke Prediction, ta thu được các thông tin tổng quan về các biến trong tập dữ liệu như sau:



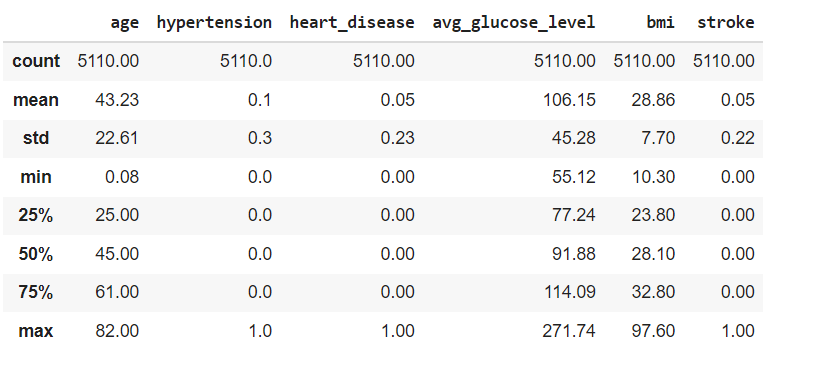
Cụ thể phần code minh họa cho báo cáo này nằm trong folder báo cáo chung, file 09.part1\_code.pdf.

* 1. **Mục đích phân tích dữ liệu**

Dự đoán đột quỵ ở con người với biến đích (mục tiêu phân tích, dự đoán) là biến "stroke"

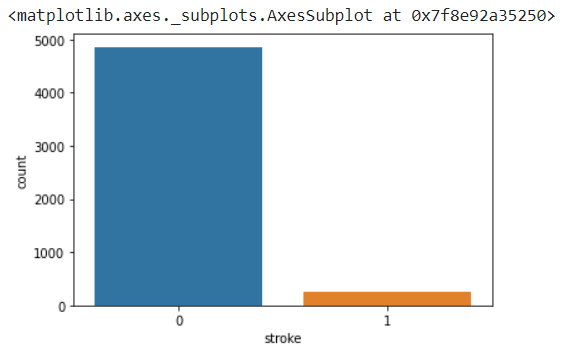
* + Tình trạng đột quỵ của bệnh nhân
  + Là biến định tính rời rạc kiểu số
  + Nhận 1 trong 2 giá trị: 0 nếu bệnh nhân không bị đột quỵ, và 1 nếu bệnh nhân bị đột quỵ
  1. **Phân tích biến đơn**
  2. **Biến định lượng**

Về các biến kiểu số (numerical data), tổng quan các thông số thống kê của biến như sau:

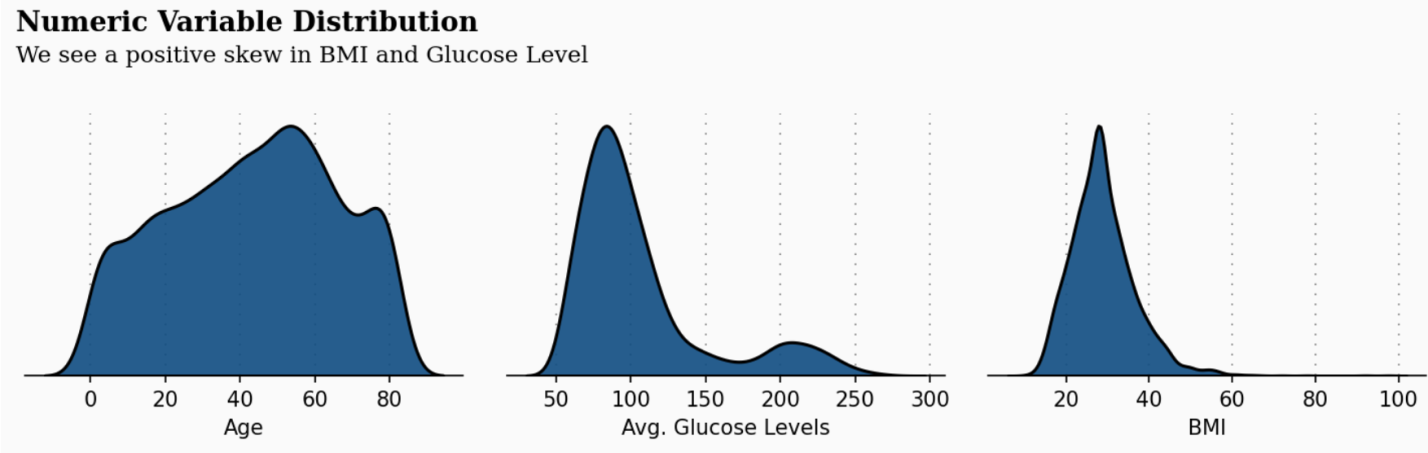


Với biến target là “stroke”, phân phối dữ liệu của biến như sau:

|  |  |
| --- | --- |
| Theo số lượng bản ghi trong data | Theo % bản ghi trong data |
|  |  |

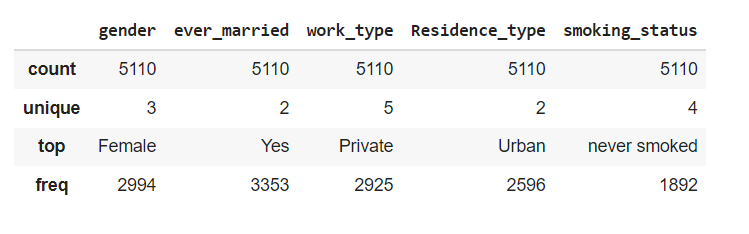


Với các biến định lượng là “age”, “avg\_glucose\_level”,”bmi” có phân phối:

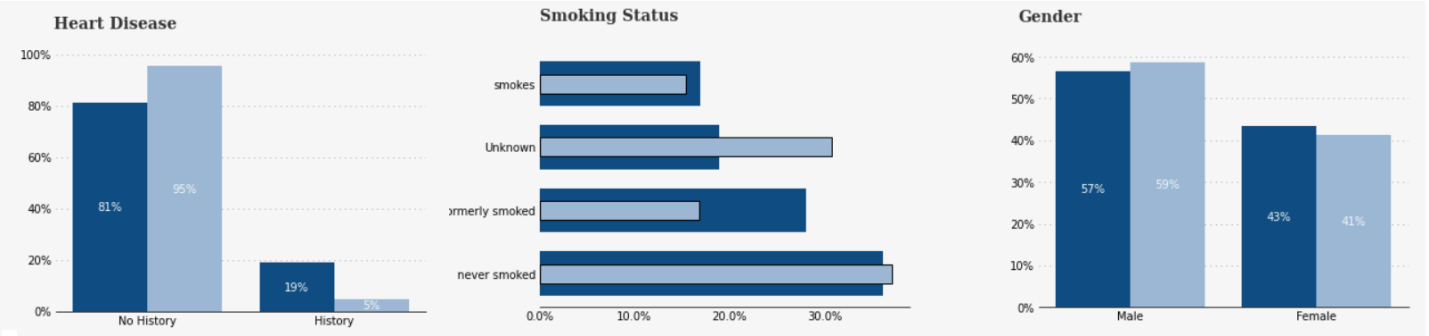


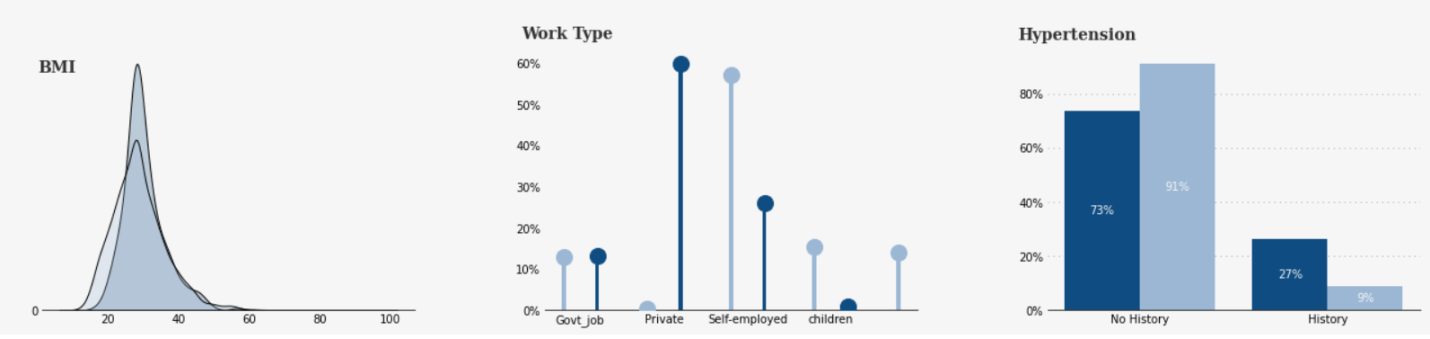
* 1. **Biến định tính**

Với các biến phân loại (categorical data), tổng quan các thông số thống kê của biến như sau:



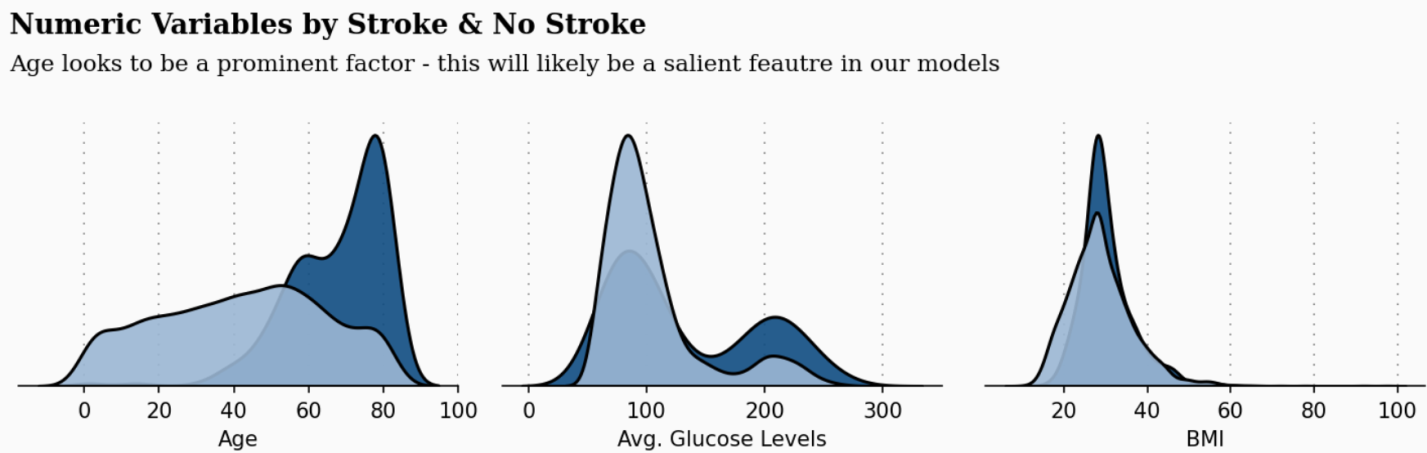
Biểu đồ thống kê của các biến định tính của bộ dữ liệu như sau:





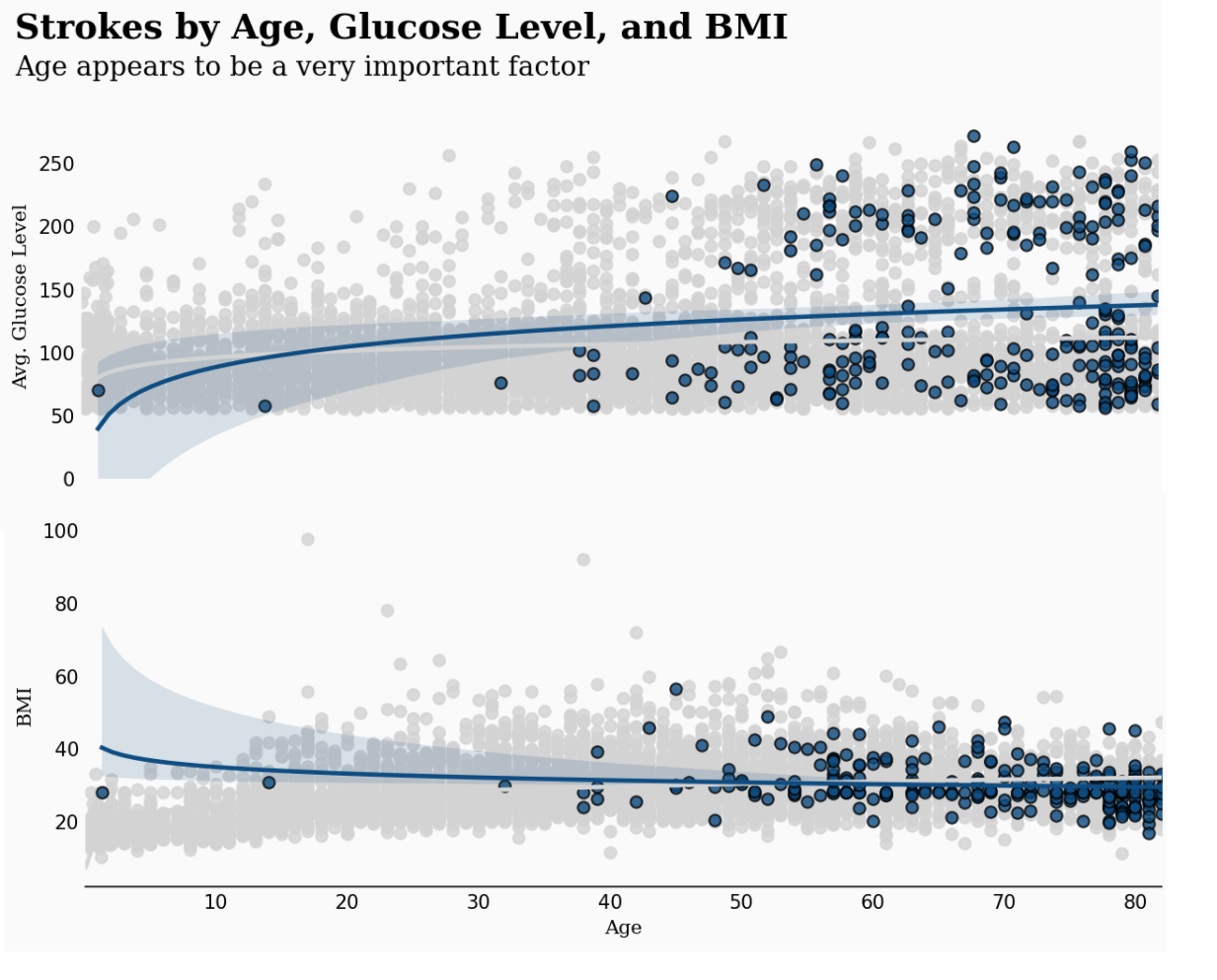
* 1. **Các thông tin khác về tập dữ liệu**
* Dữ liệu của biến “bmi” bị thiếu tại 201 bản ghi; xử lý bằng cách fill vào giá trị median của biến
* Về kiểu dữ liệu của biến, phân ra thành 3 loại:
  + Categorical Features: gender, ever\_married, work\_type, residence\_type, smoking\_status
  + Binary Numerical Features: hypertension, heart\_disease, stroke
  + Continous Numerical Feature: age, avg\_glucose\_leve, bmi
  1. **Phân tích sự phụ thuộc giữa các biến**
  2. **Giữa các biến và biến target**

Xem xét với các biến định lượng là “age”, “avg\_glucose\_level”,”bmi” sự khác nhau về phân bố ở nhóm những người bị đột quỵ và những người không bị đột quỵ:

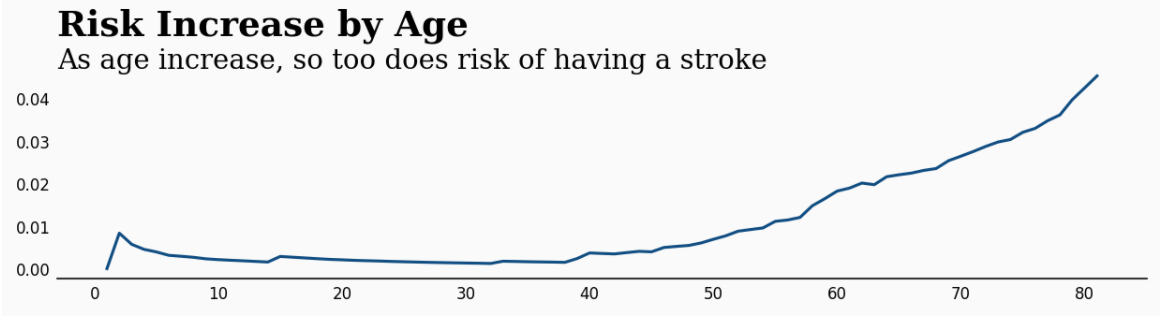


Nhìn ở biểu đồ trên, ta thấy rõ ràng biến “age”\_ tuổi tác là 1 vấn đề lớn đối với các bệnh nhân bị đột quỵ: ở nhóm người già nguy cơ bị đột quỵ cao hơn nhóm người trẻ. Còn biến “avg glucose levels” và “bmi” lại không có sự chênh lệch quá lớn ở 2 nhóm người này.

Biến “age” cũng có độ ảnh hưởng tới 2 biến định lượng còn lại là “avg glucose levels” và “bmi”, điều này thể hiện ở đồ thị dưới đây:

****

Biến “age” có tầm ảnh hưởng lớn tới biến target “stroke”, càng lớn tuổi càng tăng nguy cơ bị đột quỵ:



* 1. **Giữa các biến**

Độ tương tác giữa các biến trong bộ dữ liệu thể hiện ở biểu đồ heatmap sau

