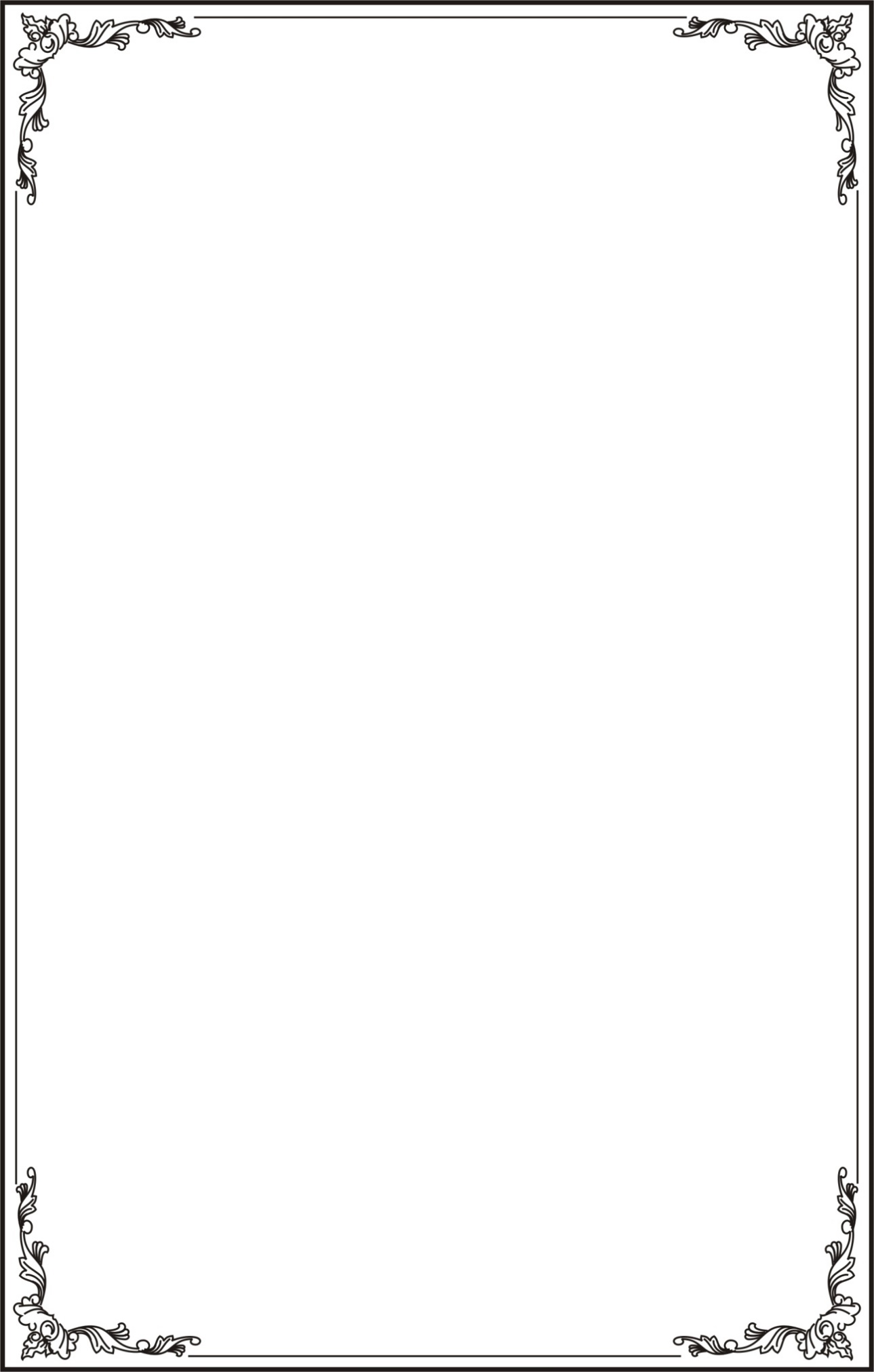
**BỘ GIÁO DỤC VÀ ĐÀO TẠO**

**TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT TP.HCM**

**KHOA CÔNG NGHỆ THÔNG TIN**

****

**BÁO CÁO CUỐI KỲ**

**ĐỀ TÀI: DỰ ĐOÁN NGUY CƠ ĐỘT QUỴ SỬ DỤNG**

**PHÂN TÍCH DỮ LIỆU LÂM SÀN**

**GVHD: ThS. Lê Minh Tân**

**Mã LHP: INDS331085\_23\_2\_01**

**Nhóm sinh viên thực hiện:**

Nguyễn Thi Phú - 21110600

Nguyễn Quang Vinh - 21110727

Lê Phúc Hậu - 21110879

*Thành phố Hồ Chí Minh, tháng 05 năm 2024*

**NHẬN XÉT CỦA GIẢNG VIÊN HƯỚNG DẪN**

**Điểm:**

**MỤC LỤC**

[**NỘI DUNG** 1](#_Toc165319216)

[**1.** **Giới Thiệu** 1](#_Toc165319217)

[**1.1.** **Mô tả:** 1](#_Toc165319218)

[**1.2.** **Phạm vi dự án:** 1](#_Toc165319219)

[**1.3.** **Instruments** 1](#_Toc165319220)

[**1.4.** **Protocols** 2](#_Toc165319221)

[**1.5.** **Các yếu tố ảnh hưởng đến sự hữu ích của dữ liệu** 3](#_Toc165319222)

[**2.** **Tổng quan bộ dữ liệu** 5](#_Toc165319223)

[**3.** **Phân tích mối tương quan giữa các đặc trưng của bộ dữ liệu:** 7](#_Toc165319224)

[**3.1.** **Mối tương quan giữa BMI và chỉ số đường huyết trung bình:** 7](#_Toc165319225)

[**3.2.** **Mối tương quan giữa bệnh tim với một số đặc trưng khác:** 10](#_Toc165319226)

[*3.2.1.* *Bệnh tim và chỉ số BMI:* 10](#_Toc165319227)

[*3.2.2.* *Bệnh tim và giới tính, tuổi:* 11](#_Toc165319228)

[*3.2.3.* *Bệnh tim và tình trạng hút thuốc:* 12](#_Toc165319229)

[**3.3.** **Mối tương quan giữa cao huyết áp và một số đặc trưng khác:** 15](#_Toc165319230)

[*3.3.1.* *Cao huyết áp và tuổi:* 15](#_Toc165319231)

[*3.3.2.* *Cao huyết áp và bệnh tim mạch:* 17](#_Toc165319232)

[**3.4.** **Mối tương quan giữa đột quỵ và các đặc trưng khác:** 19](#_Toc165319233)

[*3.4.1.* *Đột quỵ và giới tính:* 19](#_Toc165319234)

[*3.4.2.* *Đột quỵ và tuổi tác:* 20](#_Toc165319235)

[*3.4.3.* *Đột quỵ và tình trạng hôn nhân:* 22](#_Toc165319236)

[*3.4.4.* *Đột quỵ và nơi ở:* 23](#_Toc165319237)

[*3.4.5.* *Đột quỵ và loại việc làm:* 25](#_Toc165319238)

[*3.4.6.* *Đột quỵ và tình trạng hút thuốc:* 29](#_Toc165319239)

[*3.4.7.* *Đột quỵ và cao huyết áp:* 32](#_Toc165319240)

[*3.4.8.* *Đột quỵ và bệnh tim mạch:* 33](#_Toc165319241)

[*3.4.9.* *Đột quỵ và chỉ số BMI:* 35](#_Toc165319242)

[*3.4.10.* *Đột quỵ và chỉ số đường huyết trung bình:* 37](#_Toc165319243)

[**4.** **Xây dựng mô hình dự đoán bệnh đột quỵ:** 40](#_Toc165319244)

[**4.1** **Tiền xử lí dữ liệu** 40](#_Toc165319245)

[**4.2** **Huấn luyện** 40](#_Toc165319246)

[**4.3** **Đánh giá** 41](#_Toc165319247)

**DANH MỤC HÌNH ẢNH**

[Hình 1: Biểu đồ histogram thể hiện phân bố chỉ số tuổi tác của bộ dữ liệu 5](#_Toc165319459)

[Hình 2: Biểu đồ hộp phân bố của chỉ số đường huyết trung bình của bộ dữ liệu 6](#_Toc165319460)

[Hình 3: Biểu đồ hộp phân bố của chỉ số BMI của bộ dữ liệu 6](#_Toc165319461)

[Hình 4: Bảng tổng quan bộ dữ kiệu 6](#_Toc165319462)

[Hình 5: Phân bố của BMI và chỉ số đường huyết trung bình của nhóm có chỉ số đường huyết bình thường 7](#_Toc165319463)

[Hình 6: Phân bố của BMI và chỉ số đường huyết trung bình của nhóm tiền tiểu đường 8](#_Toc165319464)

[Hình 7: Phân bố của BMI và chỉ số đường huyết trung bình của nhóm tiểu đường 8](#_Toc165319465)

[Hình 8: Kết quả của kiểm định least squares để xác định mối quan hệ giữa chỉ số BMI và chỉ số đường huyết trung bình 9](#_Toc165319466)

[Hình 9: Hàm tuyến tính của chỉ số đường huyết trung bình theo BMI 9](#_Toc165319467)

[Hình 10: Biểu đồ hộp phân bố chỉ số BMI của nhóm bị bệnh tim 10](#_Toc165319468)

[Hình 11: Biểu đồ hộp phân bố chỉ số BMI của nhóm bị bệnh tim 10](#_Toc165319469)

[Hình 12: Kết quả của kiểm định t – test về mối tương quan giữa bệnh tim và chỉ số BMI 11](#_Toc165319470)

[Hình 13: Biểu đồ cột thể hiện tỉ lệ nam nữ mắc bệnh tim mạch theo từng khoảng tuổi 12](#_Toc165319471)

[Hình 14: Biểu đồ cột thể hiện mối tương quan giữa bệnh tim và hút thuốc 13](#_Toc165319472)

[Hình 15: Kết quả kiểm định chi – square về mối tương quan giữa bệnh tim mạch và tình trạng hút thuốc 14](#_Toc165319473)

[Hình 16: Biểu đồ hộp thể hiện phân bố tuổi của nhóm không bị cao huyết áp 15](#_Toc165319474)

[Hình 17: Biểu đồ hộp thể hiện phân bố tuổi của nhóm bị cao huyết áp 15](#_Toc165319475)

[Hình 18: Kết quả của kiểm định chi – square về mối tương quan giữa tuổi và cao huyết áp 16](#_Toc165319476)

[Hình 19: Biểu đồ cột thể hiện mối tương quan giữa bệnh tim và cao huyết áp 17](#_Toc165319477)

[Hình 20: Kết quả của kiểm định chi – square về mối tương quan giữa bệnh tim và cao huyết áp 18](#_Toc165319478)

[Hình 21: Biểu đồ cột thể hiện mối tương quan giữa giới tính và bệnh đột quỵ 19](#_Toc165319479)

[Hình 22: Kết quả của kiểm định chi – square về mối tương quan giữa giới tính và bệnh đột quỵ 20](#_Toc165319480)

[Hình 23: Biểu đồ thể hiện mối tương quan giữa tuổi và đột quỵ 21](#_Toc165319481)

[Hình 24: Kết quả của kiểm định chi – square về mối tương quan giữa tuổi và đột quỵ 22](#_Toc165319482)

[Hình 25: Biểu đồ cột thể hiện mối tương quan giữa đột quỵ và tình trạng hôn nhân theo từng khoảng tuổi 22](#_Toc165319483)

[Hình 26: Kết quả của kiểm định chi – square về mối tương quan giữa tình trạng hôn nhân và đột quỵ 23](#_Toc165319484)

[Hình 27: Biểu đồ cột thể hiện mối tương quan giữa đột quỵ và nơi ở 24](#_Toc165319485)

[Hình 28: Kết quả của kiểm định chi – square về mối tương quan giữa đột quỵ và nơi ở 25](#_Toc165319486)

[Hình 29: Biểu đồ cột thể hiện phân bố các loại công việc của nhóm bị đột quỵ 25](#_Toc165319487)

[Hình 30: Biểu đồ đường thể hiện số lượng của nhóm việc làm trẻ em bị đột quỵ theo từng mốc tuổi 26](#_Toc165319488)

[Hình 31: Biểu đồ đường thể hiện số lượng của nhóm việc làm riêng tư bị đột quỵ theo từng mốc tuổi 26](#_Toc165319489)

[Hình 32: Biểu đồ đường thể hiện số lượng của nhóm việc làm tự kinh doanh bị đột quỵ theo từng mốc tuổi 27](#_Toc165319490)

[Hình 33: Biểu đồ đường thể hiện số lượng của nhóm việc làm nhân viên chính phủ bị đột quỵ theo từng mốc tuổi 27](#_Toc165319491)

[Hình 34: Kết quả của kiểm định chi – square về mối tương quan giữa đột quỵ và loại việc làm 29](#_Toc165319492)

[Hình 35: Biểu đồ cột thể hiện mối tương quan giữa đột quỵ và tình trạng hút thuốc 30](#_Toc165319493)

[Hình 36: Kết quả của kiểm định chi – square về mối tương quan giữa đột quỵ và tình trạng hút thuốc 31](#_Toc165319494)

[Hình 37: Biểu đồ cột thể hiện mối tương quan giữa đột quỵ và cao huyết áp 32](#_Toc165319495)

[Hình 38: Kết quả kiểm định chi – square về mối tương quan giữa đột quỵ và cao huyết áp 33](#_Toc165319496)

[Hình 39: Biểu đồ cột thể hiện mối tương quan giữa đột quỵ và bệnh tim mạch 34](#_Toc165319497)

[Hình 40: Kết quả kiểm định chi – square về mối tương quan giữa đột quỵ và bệnh tim mạch 34](#_Toc165319498)

[Hình 41: Biểu đồ hộp thể hiện phân bố chỉ số BMI của nhóm không bị bệnh đột quỵ 35](#_Toc165319499)

[Hình 42: Biểu đồ hộp thể hiện phân bố chỉ số BMI của nhóm bị bệnh đột quỵ 35](#_Toc165319500)

[Hình 43: Kết quả kiểm định Chi – square về mối tương quan giữa đột quỵ và chỉ số BMI 36](#_Toc165319501)

[Hình 44: Biểu đồ hộp thể hiện phân bố chỉ số đường huyết trung bình của nhóm không bị bệnh đột quỵ 37](#_Toc165319502)

[Hình 45: Biểu đồ hộp thể hiện phân bố chỉ số đường huyết trung bình của nhóm bị bệnh đột quỵ 37](#_Toc165319503)

[Hình 46: Kết quả kiểm định Chi-square về mối tương quan giữa đột quỵ và chỉ số đường huyết trung bình 38](#_Toc165319504)

[Hình 47: Sơ đồ của thuật toán học máy Stacking 40](#_Toc165319505)

[Hình 48: Biểu đồ đánh giá mô hình 41](#_Toc165319506)

# **NỘI DUNG**

1. **Giới Thiệu**
2. **Mô tả:**

Dự án sẽ sử dụng các tập dữ liệu để dự đoán, phân tích khả năng mắc đột quỵ của các tình nguyện viên được khảo sát dựa trên các thông số đầu vào như giới tính, tuổi tác, chỉ số sinh hóa và tình trạng sức khỏe của cơ thể.

1. **Phạm vi dự án:**

Địa điểm: Dự án có thể thực hiện ở bất kỳ địa điểm nào có sẵn dữ liệu lâm sàn về bệnh đột quỵ. Một số nơi như bệnh viện, trung tâm y tế hoặc các cơ sở chăm sóc sức khỏe khác là những nơi có dữ liệu lâm sàn về bệnh đột quỵ và có thể triển khai dự án.

Thời gian: Thu thập dữ liệu trong 1 – 2 năm và 2 tháng để phân tích dữ liệu

Đối tượng: Những người đang điều trị và những người không mắc bệnh. Dựa trên các chỉ số lâm sàn của 2 nhóm này từ đó tìm ra những điểm khác biệt và mối liên hệ.

* P: những người bị bệnh đột quỵ và những người không bị bệnh đột quỵ tại Việt Nam
* A: các đối tượng bị bệnh đột quỵ và những người không bị bệnh đột quỵ ở thành phố Hồ Chí Minh
* S: Những bệnh nhân đang điều trị bệnh đột quỵ hoặc theo dõi sức khỏe tại các bệnh viện ở thành phố Hồ Chí Minh.

1. **Instruments**

Khảo sát trực tiếp:

* Bộ câu hỏi về các triệu chứng
* Yêu cầu: Câu hỏi dạng có hoặc không (Yes-No Question)
* Hồ sơ bệnh án
* Yêu cầu: Cần có các mục cơ bản: họ tên, địa chỉ, giới tính, SĐT, tiền sử bệnh lý, danh sách các thuốc đang sử dụng (nếu có), kết quả đo lường về huyết áp, đường huyết, nồng độ cồn, lipid, cân nặng, chiều cao,...
* Cuff và Sphygmomanometer
* Yêu cầu: Cuff bao phủ ít nhất 80% chiều dài của cánh tay và phải bao phủ toàn bộ vùng xung quanh.
* Thương hiệu: Omron
* Ống nghe và đồng hồ đo nhịp tim (Stethoscope - Watch)
* Yêu cầu: Ống nghe có chất lượng cao, với độ dẻo và khả năng truyền âm tốt.
* Thương hiệu: Littmann 3M
* Thiết bị đo SPO2 (Pulse Oximeter)
* Yêu cầu: Không
* Thương hiệu: Omron
* Đường kế (Glucometer)
* Yêu cầu: Không
* Thương hiệu: Accu-Check - Roche
* Thước dây dùng để đo chiều cao
* Yêu cầu: Độ chia nhỏ nhất của thước là 1mm
* Thương hiệu: Không yêu cầu
* Cân điện tử
* Yêu cầu: Thiết bị đạt độ chính xác cấp III theo tiêu chuẩn OIML. Bên cạnh đó giới hạn đo phải đạt tới 300kg.
* Thương hiệu: Cân bàn điện tử JWI 700W Jadever

Khảo sát trực tuyến (online)

* Dùng Google Form để khảo sát

1. **Protocols**

* Đo huyết áp
* Bước 1: Người bệnh ngồi yên, lưng thẳng, chân không chạm đất. Tay được đặt trên bàn ở tư thế thoải mái.
* Bước 2: Người đo sẽ đeo cái cuff quanh cánh tay trên của người bệnh, đảm bảo cuff bao phủ ít nhất 80% chiều dài cánh tay.
* Bước 3: Bóp nút xả hơi trên bóp huyết áp để xả hết không khí trong cuff.
* Bước 4: Bật nút bơm để bơm hơi dần vào cuff cho đến khi cuff bắt đầu chèn ép động mạch cánh tay.
* Bước 5: Nghe âm thanh tiếng máu chảy qua động mạch bằng ống nghe và đọc chỉ số huyết áp cao nhất.
* Bước 6: Nút xả hơi cho hơi trong cuff thoát ra từ từ cho đến khi không còn nghe âm thanh và đọc chỉ số huyết áp thấp nhất.
* Bước 7: Ghi lại kết quả 2 lần đo và trung bình chúng
* Đối với đo nhịp tim:
* Bước 1: Người bệnh nằm ngửa, chân duỗi thẳng.
* Bước 2: Người đo đặt ống nghe lên vị trí tim, và định vị được âm thanh truyền nhất.
* Bước 3: Người đo nhìn vào đồng hồ đo nhịp và đếm số nhịp trong 15 giây.
* Bước 4: Nhân kết quả đếm nhịp với 4 để tính toán nhịp tim trong 1 phút (nhịp/phút).
* Bước 5: Đo 2 lần và tính trung bình các kết quả.
* Đối với đo đường huyết:
* Bước 1: Lấy máy đo đường huyết và que thử.
* Bước 2: Lấy một giọt máu từ ngón tay cái bằng lấy máu
* Bước 3: Đặt máu lên đầu que thử.
* Bước 4: Bật máy và đọc kết quả trên màn hình.
* Bước 5: Ghi lại kết quả.
* Đối với đo chiều cao: Sử dụng thước dây có độ chia nhỏ nhất 1mm, đặt thước dây lên đỉnh đầu và đọc kết quả tại điểm gặp nhau của đầu đo và đầu thước.
* Đối với việc đo cân nặng: Sử dụng cân điện tử JWI 700W Jadever, yêu cầu người được đo đứng trên cân và đọc kết quả trên màn hình hiển thị.
* Đối với việc đo nồng độ cồn: Sử dụng thiết bị đo nồng độ cồn Alcohol Tester AT6000, người được kiểm tra thở vào ống hút hoặc cảm biến và đọc kết quả trên màn hình hiển thị.

Lưu ý:

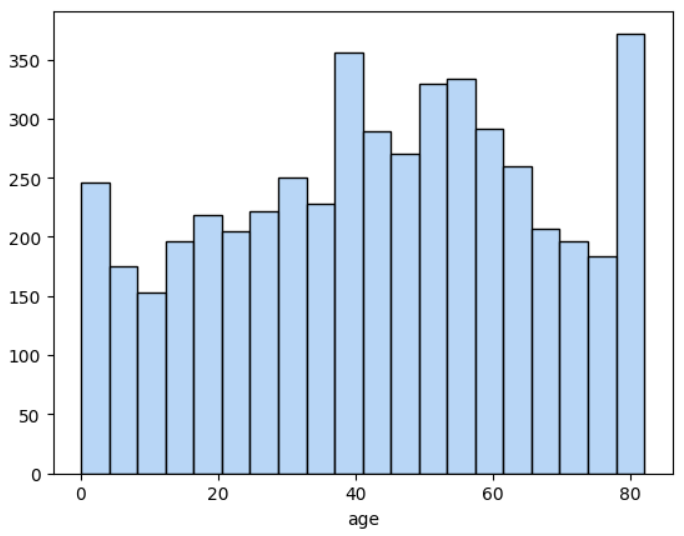
* Phải đảm bảo vệ sinh đối với các thiết bị đo và cần được thay, vệ sinh thiết bị để đảm bảo vệ sinh, độ chính xác, phải cài đặt lại thiết bị sau mỗi lần sử dụng.
* Thực hiện đo, sử dụng thiết bị theo hướng dẫn của nhà sản xuất và tuân thủ các quy định về an toàn và sử dụng thiết bị.
* Tất cả những người thực hiện đo các thông sô trên đều phải được đào tạo và có chứng chỉ y tế theo đúng quy định của pháp luật.

1. **Các yếu tố ảnh hưởng đến sự hữu ích của dữ liệu**

* Dữ liệu bị thiếu:
* Người được khảo sát ngại trả lời các câu hỏi về tiền sử bệnh, tình trạng hôn nhân, các chỉ số cơ thể,...
* Người được khảo sát thiếu kiên nhẫn khi thực hiện quá nhiều bước đo để lấy các thông tin về chỉ số sinh hóa.
* Người khảo sát không rõ về tiền sử bệnh của người thân.
* Dữ liệu không đúng đắn
* Khi khảo sát online qua form người khảo sát có thể điền các dữ liệu không được đo một cách chính xác, hoặc cố tình điền sai
* Người được khảo sát cố tình trả lời sai để che giấu các thông tin cá nhân quan trọng như SĐT, địa chỉ nhà, công việc hiện tại,...
* Người khảo sát không muốn người khác biết bản thân đang stress, hay mắc các bệnh trong người
* Dữ liệu mang tính chủ quan
* Tình nguyên viên được khảo sát chỉ dựa trên quan điểm cá nhân, chứ không có bất kì đánh giá khách quan nào của chuyên gia để trả lời các câu hỏi sức khỏe.
* Câu hỏi mang tính chất hàn lâm, thuộc lĩnh vực Y tế
* Nếu trả lời qua form khảo sát trực tuyến, người được khảo sát sẽ bối rối, lúng túng khi gặp các câu hỏi về chỉ số đường huyết, huyết áp, nhịp tim,…

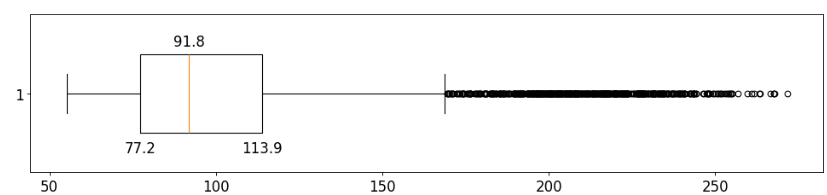
1. **Tổng quan bộ dữ liệu**

* Tên: Brain Stroke Dataset
* Tác giả: Jillani Soft Tech
* Data Scientist and Machine Learning Engr at Blocbelt Lahore, Punjab, Pakistan
* Link:<https://www.kaggle.com/datasets/jillanisofttech/brain-stroke-dataset>
* Gồm 4.981 records và 11 fields:
* gender: Giới tính (Male: nam (2074) , Female: nữ (2907))
* age: Tuổi



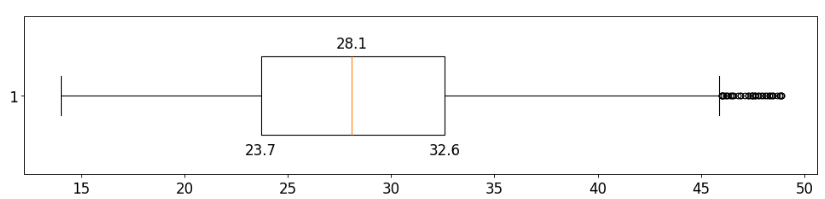
Hình 1: Biểu đồ histogram thể hiện phân bố chỉ số tuổi tác của bộ dữ liệu

* hypertension: Cao huyết áp (0: không bị cao huyết áp (4502), 1: có cao huyết áp (479))
* heart\_disease: Bệnh tim (0: không bị bệnh tim (4706), 1: có bị bệnh tim (275))
* ever\_married: Đã lập gia đình chưa (No: chưa có gia đình (1701), Yes: đã có gia đình (3280))
* work\_type: Loại công việc (Self-employed: nghề tự do (804), Private: không tiết lộ (2860), Government Job: cán bộ Nhà nước (644), Children: trẻ em (673))
* Residence\_type: Nơi ở (Urban: thành thị (2532) hoặc Rural: nông thôn (2449))
* avg\_glucose\_level: Chỉ số mức đường huyết trung bình



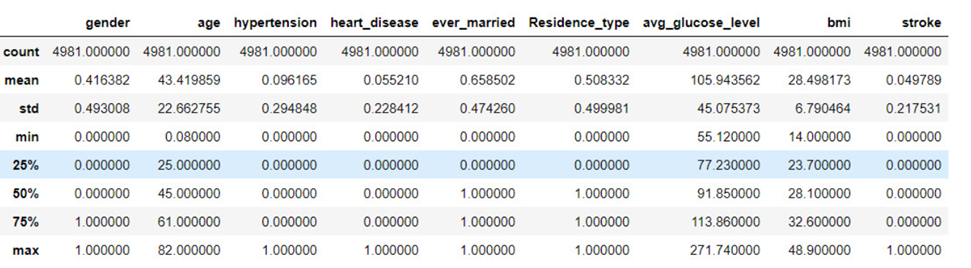
Hình 2: Biểu đồ hộp phân bố của chỉ số đường huyết trung bình của bộ dữ liệu

* BMI: Chỉ số thể trọng = cân nặng/chiều cao2



Hình 3: Biểu đồ hộp phân bố của chỉ số BMI của bộ dữ liệu

* smoking\_status: Tình trạng hút thuốc (smokes: hút gần đây (776), formerly smoked: từng hút (867), never smoked: chưa từng hút (1838), Unknown: không biết hoặc bất khả dụng (1500))
* stroke: Đột quỵ (1: đột quỵ (248), 0: không bị đột quỵ (4733))
  + Nhận xét: Tất cả các cột đều đủ dữ liệu.
  + Thống kê chất lượng dữ liệu: Tốt



Hình 4: Bảng tổng quan bộ dữ kiệu

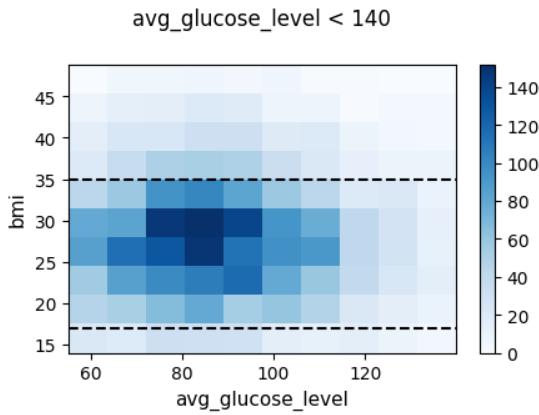
1. **Phân tích mối tương quan giữa các đặc trưng của bộ dữ liệu:**
2. **Mối tương quan giữa BMI và chỉ số đường huyết trung bình:**

Chỉ số BMI theo phân loại quốc tế (WHO) thì dưới 18,5 là nhẹ cân, 18,5-24,9 là bình thường, 25-29,9 là thừa cân và từ 30 trở lên là béo phì.

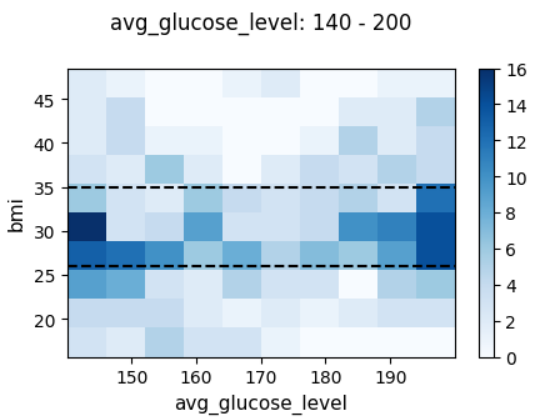
Béo phì có thể làm cho các tế bào của cơ thể đề kháng với insulin. Insulin là một loại hormone mang đường từ máu đến các tế bào trong cơ thể của bạn, nơi nó được sử dụng để tạo năng lượng. Nếu bạn đề kháng với insulin, đường không thể được các tế bào hấp thụ, dẫn đến lượng đường trong máu cao. [1]

Theo Thư viện Y khoa Quốc gia Hoa Kỳ, nghiên cứu về mối tương quan giữa chỉ số BMI ở bệnh nhân tiểu đường và không tiểu đường đã chỉ ra rằng những người mắc bệnh tiểu đường có chỉ số BMI cao hơn so với nhóm không mắc bệnh. [2]

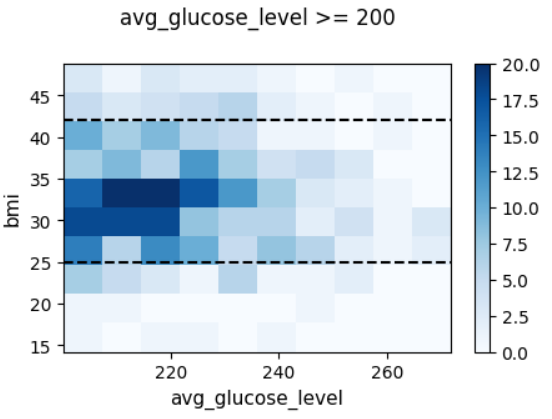
Kiểm định tính xác thực, ta chia bộ dữ liệu thành 3 nhóm dựa trên chỉ số đường huyết trung bình: bình thường (đường huyết < 140), tiền tiểu đường (140 ≤ đường huyết < 200) và tiểu đường (đường huyết ≥ 200). Quan sát những thay đổi về chỉ số BMI giữa 3 nhóm này.



Hình 5: Phân bố của BMI và chỉ số đường huyết trung bình của nhóm có chỉ số đường huyết bình thường



Hình 6: Phân bố của BMI và chỉ số đường huyết trung bình của nhóm tiền tiểu đường



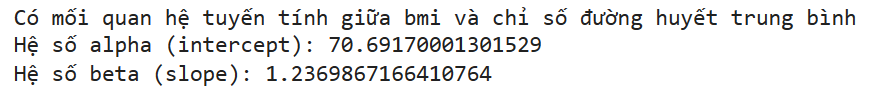
Hình 7: Phân bố của BMI và chỉ số đường huyết trung bình của nhóm tiểu đường

Quan sát 3 biểu đồ:

* Biểu đồ chỉ số đường huyết bình thường: chỉ số BMI tập trung nhiều trong khoảng 17 đến 35.
* Biểu đồ chỉ số đường huyết từ 140 đến 200: chỉ số BMI tập trung nhiều trong khoảng 25 đến 35. Phân bố của chỉ số BMI dưới 25 (BMI của nhóm bình thường và nhóm nhẹ cân) thấp hơn so với nhóm đường huyết bình thường.
* Biểu đồ chỉ số đường huyết từ 200 trở lên: chỉ số BMI tập trung nhiều trong khoảng 25 đến 42. Phân bố của chỉ số BMI dưới 25 chiếm số lượng cực thấp. Đa phần nhóm này có chỉ số BMI thuộc nhóm béo phì và số ít thuộc nhóm thừa cân

Kiểm định bằng phương pháp Least Squares Estimate:

* H0 : “Không có mối quan hệ tuyến tính giữa BMI và chỉ số đường huyết trung bình.”
* H1 : “Có mối quan hệ tuyến tính giữa BMI và chỉ số đường huyết trung bình.”



Hình 8: Kết quả của kiểm định least squares để xác định mối quan hệ giữa chỉ số BMI và chỉ số đường huyết trung bình

* Do β ≠ 0, nên ta bác bỏ giả thuyết H0
* BMI và chỉ số đường huyết trung bình có mối quan hệ tuyến tính, tuân theo:



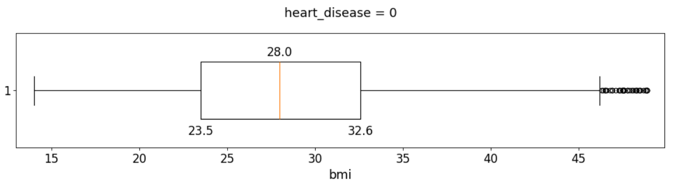
Hình 9: Hàm tuyến tính của chỉ số đường huyết trung bình theo BMI

Kết luận: Có mối quan hệ tuyến tính giữa chỉ số BMI và chỉ số đường huyết trung bình. Người có chỉ số BMI thuộc nhóm thừa cân và béo phì có nguy cơ cao tăng chỉ số đường huyết cao. Điều này phù hợp với kết quả nghiên cứu của y khoa.

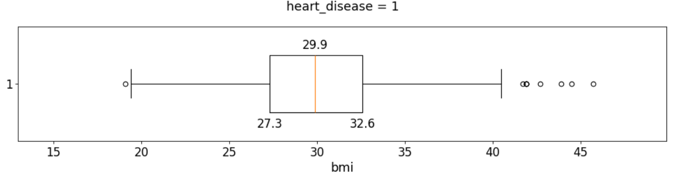
1. **Mối tương quan giữa bệnh tim với một số đặc trưng khác:**
2. *Bệnh tim và chỉ số BMI:*

Theo Thư viện Y khoa Quốc gia Hoa Kỳ, béo phì làm tăng đáng kể nguy cơ mắc bệnh tim mạch so với nhóm BMI thông thường, thừa cân liên quan đến đến nguy cơ phát triển bệnh tim mạch đáng kể ở độ tuổi sớm hơn. [3]

Kiểm định tính xác thực: Chia bộ dữ liệu thành với nhóm không mắc bệnh tim và nhóm mắc bệnh tim quan sát sự khác biệt về phân bố BMI giữa hai nhóm



Hình 10: Biểu đồ hộp phân bố chỉ số BMI của nhóm bị bệnh tim



Hình 11: Biểu đồ hộp phân bố chỉ số BMI của nhóm bị bệnh tim

Quan sát 2 biểu đồ:

* Biểu đồ hộp về chỉ số BMI của người không bị bệnh tim: Q1 = 23.5, Q2 = 28.0, Q3 = 32.6.
* Biểu đồ hộp về chỉ số BMI của người bị bệnh tim: Q1 = 27.3, Q2 = 29.9, Q3 = 32.6. Có hơn 75% người thuộc nhóm thừa cân và béo phì (BMI ≥ 25). Có đến 50% người bị béo phì (BMI từ 30 trở lên).
* Có thể thấy, giá trị Q3 giữa hai biểu đồ không có sự khác biệt giá trị Q1, Q2 giữa hai nhóm lại cho thấy sự khác biệt. Ở nhóm bị bệnh tim thì những giá trị này cao hơn nhiều so với nhóm bị người không bị bệnh tim.

Kiểm định t – test với mức ý nghĩa 5%:

* H0 : “Không có mối tương quan giữa chỉ số BMI và bệnh tim.”
* H1 : “Có mối tương quan giữa chỉ số BMI và bệnh tim.”

A black text on a white background

Description automatically generated

Hình 12: Kết quả của kiểm định t – test về mối tương quan giữa bệnh tim và chỉ số BMI

* Do p-value bé hơn mức ý nghĩa, nên ta bác bỏ giả thuyết H0
* Với mức ý nghĩa 5%, ta thấy có mối tương quan giữa chỉ số BMI và bệnh tim mạch.

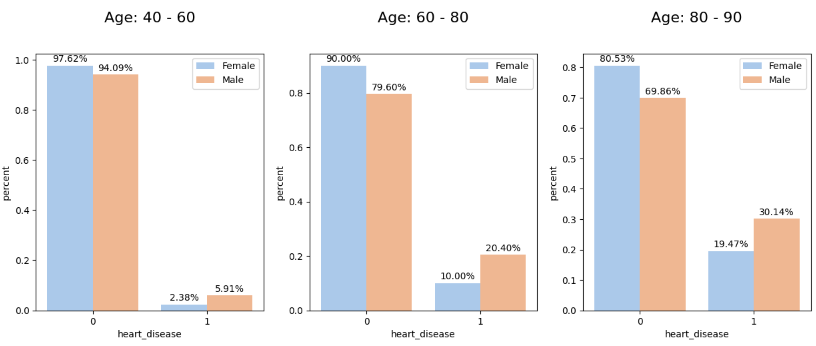
Kết luận: Có mối tương quan giữa BMI và bệnh tim mạch. Nhóm bị bệnh tim có tỉ lệ người bị thừa cân và béo phì cao hơn so với nhóm không bị bệnh tim. Điều này phù hợp với kết quả nghiên cứu của y khoa.

1. *Bệnh tim và giới tính, tuổi:*

Theo Thư viện Y khoa Quốc gia Hoa Kỳ:

* Nguy cơ xảy ra các biến cố tim mạch thường gia tăng nhanh hơn khi tuổi đời càng cao [4].
* Nam giới có nguy cơ mắc các bệnh tim mạch cao hơn so với nữ giới. Tuy nhiên sau thời kỳ mãn kinh (khoảng ngoài 50 tuổi), nguy cơ mắc bệnh tim mạch ở nữ giới sẽ gia tăng nhanh hơn. [5]

Kiểm định tính xác thực: Ta chọn mốc tuổi dễ bị bệnh tim là từ 40 tuổi, chia bộ dữ liệu theo 3 mốc tuổi: từ 40 - 60 tuổi, từ 60 - 80 tuổi và ngoài 80 tuổi. Quan sát tỉ lệ nam nữ mắc bệnh tim mạch.



Hình 13: Biểu đồ cột thể hiện tỉ lệ nam nữ mắc bệnh tim mạch theo từng khoảng tuổi

Quan sát biểu đồ:

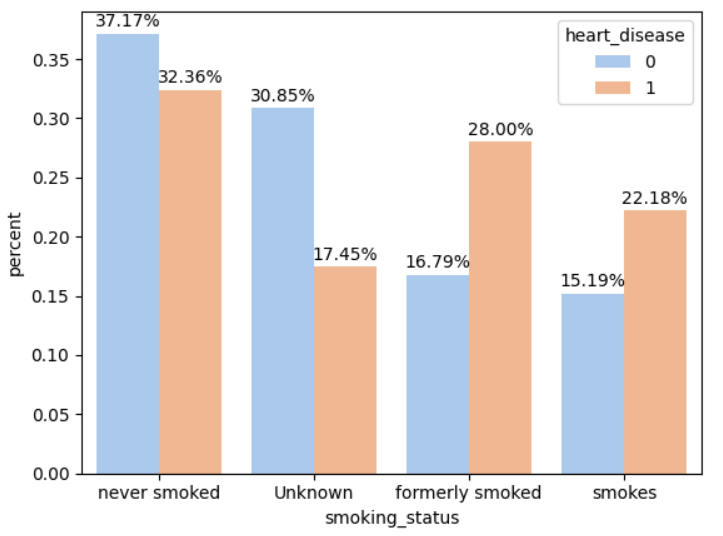
* Nhìn chung tỷ lệ nam, nữ mắc bệnh tim mạch đều tăng dần theo độ tuổi. Tại cả 3 biểu đồ, tỷ nam mắc bệnh tim nhiều hơn nữ.
* Mốc tuổi 40 - 60, nam mắc bệnh tim mạch cao gấp 2.5 lần nữ giới. Mốc tuổi 60 - 80, nam mắc bệnh tim cao gấp 2 lần nữ. Mốc tuổi ngoài 80, nam giới mắc bệnh tim cao gấp 1.5 lần nữ giới. Có thể thấy tuy nam có tỷ lệ mắc bệnh tim mạch cao hơn phụ nữ, nhưng ở độ tuổi ngày càng cao, khoảng cách chênh lệch giữa nam nữ mắc bệnh tim mạch dần dần được thu hẹp. Điều này chứng minh phụ nữ sau thời kỳ mãn kinh (khoảng ngoài 50 tuổi) có tỉ lệ mắc bệnh tim mạch tăng lên nhanh chóng.

Kết luận: Có mối tương quan giữa độ tuổi và bệnh tim mạch. Độ tuổi càng cao thì nguy cơ mắc bệnh tim càng gia tăng, nam giới có tỉ lệ mắc bệnh tim mạch cao hơn nữ giới, ở độ tuổi sau mãn kinh tỷ lệ nữ giới mắc bệnh tim mạch gia tăng nhanh chóng. Điều này phù hợp với nghiên cứu của y khoa

1. *Bệnh tim và tình trạng hút thuốc:*

Theo nghiên cứu của Thư viện Y khoa Quốc gia Hoa Kỳ: hút thuốc lá là một trong những yếu tố nguy cơ hàng đầu mắc các bệnh tim mạch mãn tính. Hút thuốc làm tổn thương tim và mạch máu rất nhanh, nhưng tổn thương này sẽ được phục hồi nhanh chóng đối với hầu hết những người hút thuốc ngừng hút thuốc. [6]

Kiểm định tính xác thực: chia bộ dữ liệu thành 2 nhóm, nhóm bị bệnh tim và nhóm không bị bệnh tim, sau đó theo dõi sự phân bố của thuộc tính tình trạng hút thuốc trong 2 nhóm trên.



Hình 14: Biểu đồ cột thể hiện mối tương quan giữa bệnh tim và hút thuốc

Quan sát biểu đồ:

* Hút thuốc trong nhóm không bị bệnh tim chiếm 15.19% và 22.18% ở nhóm bị bệnh tim. Con số chênh lệch đáng kể khi ở nhóm bị bệnh tim gấp 1.46 lần nhóm không bị bệnh tim. Chứng tỏ hút thuốc có nguy cơ bệnh tim cao.
* Đã từng hút thuốc: chiếm 16.79% trong nhóm không bị bệnh tim và 28% trong nhóm bị bệnh tim. Có thể thấy đã từng hút thuốc trong nhóm bị bệnh tim cao gấp 1.67 nhóm không bị bệnh tim. Việc đã từng hút thuốc cũng có nguy cơ gây bệnh tim.
* Không hút thuốc: chiếm 37.17% trong nhóm không bị bệnh tim và 32.36% trong nhóm bị bệnh tim. Không hút thuốc làm giảm khả năng bị bệnh tim một chút.

Kiểm định chi – square về mối tương quan giữa tuổi và bệnh tim:

* H0 : “Không có mối tương quan giữa các nhóm tuổi và bệnh tim mạch.”
* H1 : “Có mối tương quan giữa các nhóm tuổi và bệnh tim mạch.”

|  |
| --- |
|  |
|  |

A number and numbers on a white background

Description automatically generated

Hình 15: Kết quả kiểm định chi – square về mối tương quan giữa bệnh tim mạch và tình trạng hút thuốc

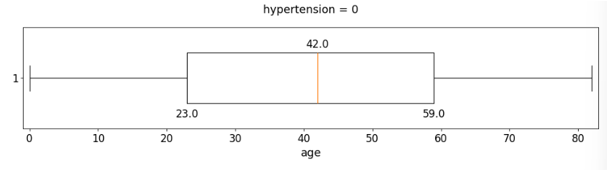
Kết luận:

* Việc hút thuốc làm tăng khả năng bị bệnh tim, điều này đúng với nghiên cứu của y khoa.
* Nhưng với nhóm đã từng hút thuốc lại có tỉ lệ mắc bệnh tim mạch cao hơn so với nhóm hút thuốc. Điều này không phù hợp với nghiên cứu của y khoa. Điều này có khả năng còn phụ thuộc vào thời gian hút thuốc trước đó và thời gian bỏ hút thuốc, nhưng các dữ liệu này không được thể hiện trong bộ dữ liệu nên không thể đưa ra kết luận chính xác.

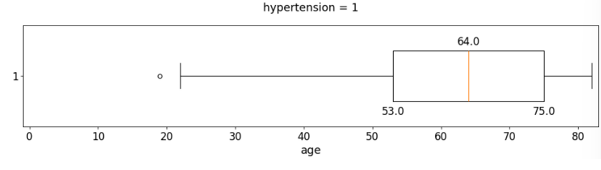
1. **Mối tương quan giữa cao huyết áp và một số đặc trưng khác:**
2. *Cao huyết áp và tuổi:*

Theo nghiên cứu của Thư viện Y khoa Quốc gia Hoa Kỳ, cao huyết áp có tỉ lệ mắc tăng đáng kể khi tuổi ngày càng tăng. Điều này là do khi con người về già, thành động mạch bị lão hóa, giảm tính đàn hồi và trở nên cứng hơn, tăng tích lũy mỡ dẫn đến xơ mỡ động mạch, làm tăng khả năng bị cao huyết áp. [7]

Kiểm định tính xác thực: Chia bộ dữ liệu thành 2 nhóm bị cao huyết áp và không bị cao huyết áp, theo dõi sự phân bố của độ tuổi giữa hai nhóm này.



Hình 16: Biểu đồ hộp thể hiện phân bố tuổi của nhóm không bị cao huyết áp



Hình 17: Biểu đồ hộp thể hiện phân bố tuổi của nhóm bị cao huyết áp

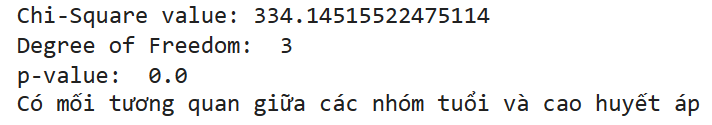
Quan sát biểu đồ:

* Nhóm không bị bệnh tim: chỉ số tuổi tác Q1 = 23, Q2 = 42, Q3 = 59
* Nhóm bị bệnh tim: chỉ số tuổi tác Q1 = 53, Q2 = 64, Q3 = 75. Có đến trên 75% người bị bệnh tim từ 53 tuổi trở lên, đây là người ở độ tuổi trung niên và người già
* Các giá trị Q1, Q2, Q3 của nhóm bị bệnh tim cao hơn rất nhiều so với nhóm không bị bệnh tim. Điều này cho thấy đa phần những người bị cao huyết áp thuộc nhóm người cao tuổi.

Kiểm định chi – square với mức ý nghĩa 5%

* H0 : “Không có mối tương quan giữa các nhóm tuổi và bệnh cao huyết áp.”
* H1 : “Có mối tương quan giữa các nhóm tuổi và bệnh cao huyết áp.”

|  |
| --- |
|  |
|  |



Hình 18: Kết quả của kiểm định chi – square về mối tương quan giữa tuổi và cao huyết áp

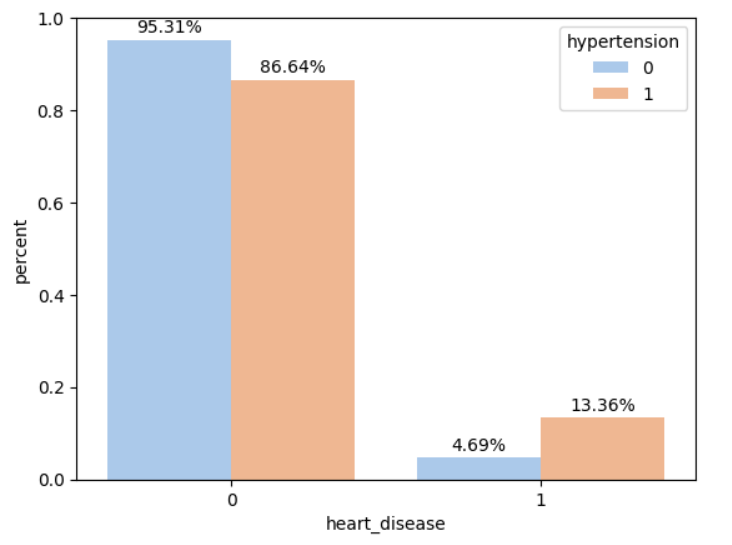
* Do p-value bé hơn mức ý nghĩa, nên ta bác bỏ giả thuyết H0
* Với mức ý nghĩa 5%, ta thấy có mối tương quan giữa các nhóm tuổi và bệnh cao huyết áp.

Kết luận: tập dữ liệu thể hiện người cao tuổi có nguy cơ cao bị cao huyết áp, điều này phù hợp với nghiên cứu của y khoa.

1. *Cao huyết áp và bệnh tim mạch:*

Theo nghiên cứu của Theo Thư viện Y khoa Quốc gia Hoa Kỳ, người bị cao huyết áp có nguy cơ mắc bệnh tim mạch cao [8]

Kiểm định tính xác thực: Chia bộ dữ liệu thành 2 nhóm bị cao huyết áp và không bị cao huyết áp, quan sát sự phân bố của đặc trưng bệnh tim giữa hai nhóm này



Hình 19: Biểu đồ cột thể hiện mối tương quan giữa bệnh tim và cao huyết áp

Quan sát biểu đồ: Ở nhóm người không bị cao huyết áp có 4.69% bị bệnh tim. Ở nhóm người bị cao huyết áp có 13.36% bị bệnh tim. Hai con số này giữa hai nhóm có sự chênh lệch đáng kể

Kiểm định chi – square với mức ý nghĩa 5%:

* H0 : “Không có mối tương quan giữa bệnh tim mạch và bệnh cao huyết áp.”
* H1 : “Có mối tương quan giữa bệnh tim mạch và bệnh cao huyết áp.”

|  |
| --- |
|  |
|  |

A close up of numbers

Description automatically generated

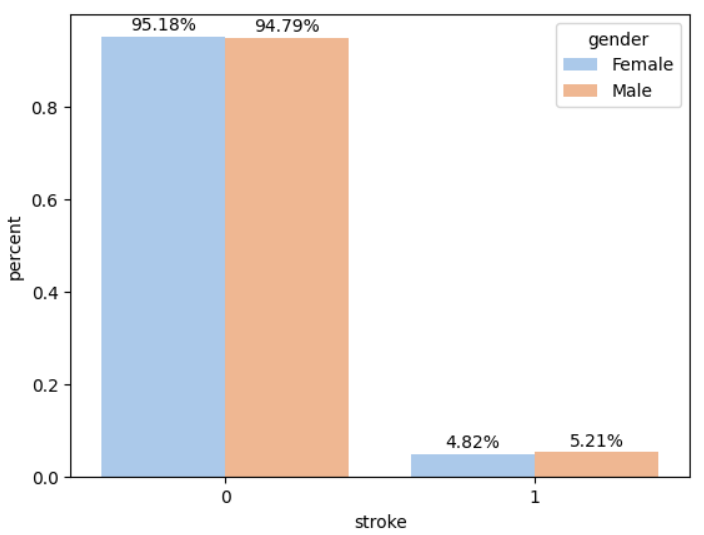
Hình 20: Kết quả của kiểm định chi – square về mối tương quan giữa bệnh tim và cao huyết áp

* Do p-value bé hơn mức ý nghĩa, nên ta bác bỏ giả thuyết H0
* Với mức ý nghĩa 5%, ta thấy có mối tương quan giữa bệnh tim mạch và bệnh cao huyết áp.

Kết luận: Có mối tương quan giữa cao huyết áp và bệnh tim mạch. Kết quả quan sát từ biểu đồ nhóm người bị cao huyết áp có nguy cơ bị tim mạch cao hơn. Điều này phù hợp với kết quả nghiên cứu của y khoa

1. **Mối tương quan giữa đột quỵ và các đặc trưng khác:**
2. *Đột quỵ và giới tính:*

Theo nghiên cứu của Thư viện Y khoa Quốc gia Hoa Kỳ, nam giới có nguy cơ bị đột quỵ nhiều hơn phụ nữ. [9]



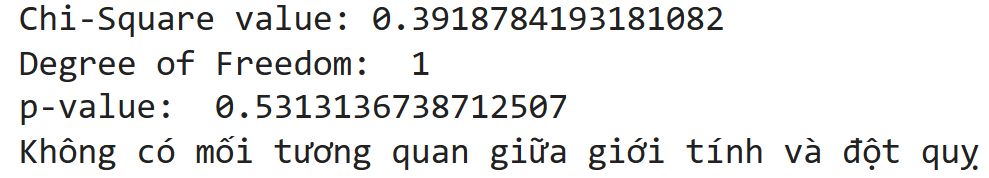
Hình 21: Biểu đồ cột thể hiện mối tương quan giữa giới tính và bệnh đột quỵ

Quan sát biểu đồ: tỉ lệ bị đột quỵ giữa nam và nữ gần như không có quá nhiều sự khác biệt

Kiểm định Chi-square với mức ý nghĩa 5%:

* H0 : “Không có mối tương quan giữa giới tính và bệnh đột quỵ.”
* H1 : “Có mối tương quan giữa giới tính và bệnh đột quỵ.”

|  |
| --- |
|  |
|  |



Hình 22: Kết quả của kiểm định chi – square về mối tương quan giữa giới tính và bệnh đột quỵ

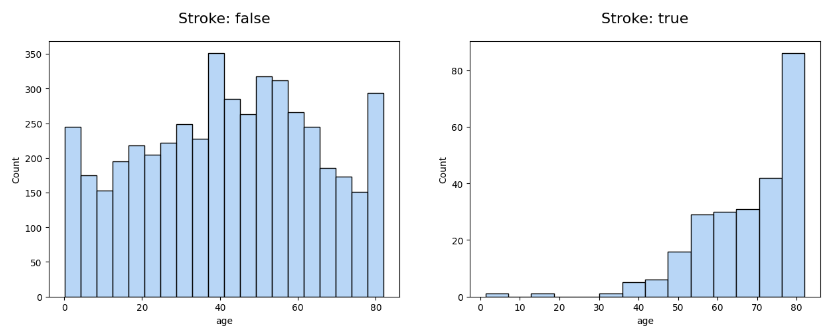
* Do p-value lớn hơn mức ý nghĩa, nên ta không thể bác bỏ giả thuyết H0
* Với mức ý nghĩa 5%, ta thấy không có mối tương quan giữa giới tính và bệnh đột quỵ.

Kết luận: giới tính không có sự ảnh hưởng đến đột quỵ. Điều này không phù hợp với kết quả nghiên cứu

1. *Đột quỵ và tuổi tác:*

Theo nghiên cứu của Thư viện Y khoa Quốc gia Hoa Kỳ: Lão hóa có liên quan đến nhiều thay đổi đáng chú ý ở động mạch não trong và ngoài sọ người. Đột quỵ có thể xảy ra ở mọi lứa tuổi nhưng càng cao tuổi thì nguy cơ bị đột quỵ càng tăng và tăng rõ rệt sau 50 tuổi. [10]

Kiểm định tính xác thực: chia bộ dữ liệu thành 2 phần, bị đột quỵ và không bị đột quỵ sau đó quan sát sự phân bố độ tuổi giữa 2 nhóm



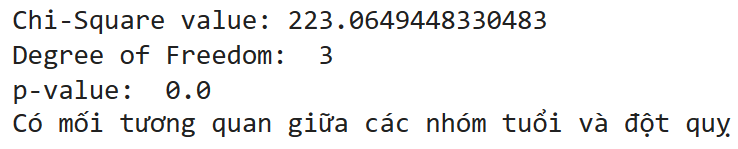
Hình 23: Biểu đồ thể hiện mối tương quan giữa tuổi và đột quỵ

Quan sát biểu đồ: Nhóm không bị đột quỵ, phân phối giữa các nhóm tuổi có độ dàn trải. Nhóm bị đột quỵ: Có 2 giá trị ngoại lai trong khoảng tuổi dưới 20 tuổi, 2 giá trị ngoại lai này có số lượng không đáng kể. Dữ liệu của biểu đồ có phân phối bị nghiêng về phía bên phải. Số lượng người mắc đột quỵ phân bố tăng dần theo số tuổi, trước 30 tuổi có rất ít trường hợp đột quỵ; sau 30 tuổi tỉ lệ bị đột quỵ tăng dần qua các năm và tăng mạnh nhất sau 50 tuổi.

Kiểm định chi – square với mức ý nghĩa 5%:

* H0 : “Không có mối tương quan giữa độ tuổi và bệnh đột quỵ.”
* H1 : “Có mối tương quan giữa độ tuổi và bệnh đột quỵ.”

|  |
| --- |
|  |
|  |

****

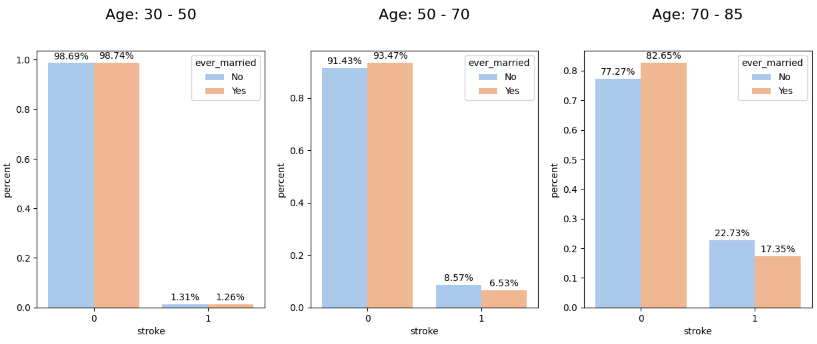
Hình 24: Kết quả của kiểm định chi – square về mối tương quan giữa tuổi và đột quỵ

* Do p-value bé hơn mức ý nghĩa, nên ta bác bỏ giả thuyết H0
* Với mức ý nghĩa 5%, ta thấy có mối tương quan giữa tuổi và bệnh đột quỵ.

Kết luận: càng lớn tuổi nguy cơ mắc đột quỵ càng cao, điều này phù hợp với kết quả nghiên cứu của y khoa.

1. *Đột quỵ và tình trạng hôn nhân:*

Ở bộ dữ liệu này độ tuổi bắt đầu xuất hiện người bị đột quỵ là từ 30 trở đi. Với từng khoảng tuổi từ 30 trở đi, chia thành 2 nhóm đã kết hôn và chưa kết hôn, sau đó quan sát sự phân bố của đột quỵ trên 2 nhóm này

****

Hình 25: Biểu đồ cột thể hiện mối tương quan giữa đột quỵ và tình trạng hôn nhân theo từng khoảng tuổi

Quan sát biểu đồ: Ở mốc tuổi từ 30 - 50, tỉ lệ bị đột quỵ giữa người chưa kết hôn và người đã kết hôn là không đáng kể. Mốc tuổi từ 50 - 70, người chưa kết hôn có tỉ lệ bị đột quỵ cao hơn 1.3 lần so với người đã kết hôn. Mốc tuổi 70 - 80, người chưa kết hôn có tỉ lệ bị đột quỵ cao gấp 1,31 lần so với người đã kết hôn.

Kiểm định chi-square với mức ý nghĩa 5%:

* H0 : “Không có mối tương quan giữa việc lập gia đình và bệnh đột quỵ.”
* H1 : “Có mối tương quan giữa việc lập gia đình và bệnh đột quỵ.”

|  |
| --- |
|  |
|  |

A close up of numbers

Description automatically generated

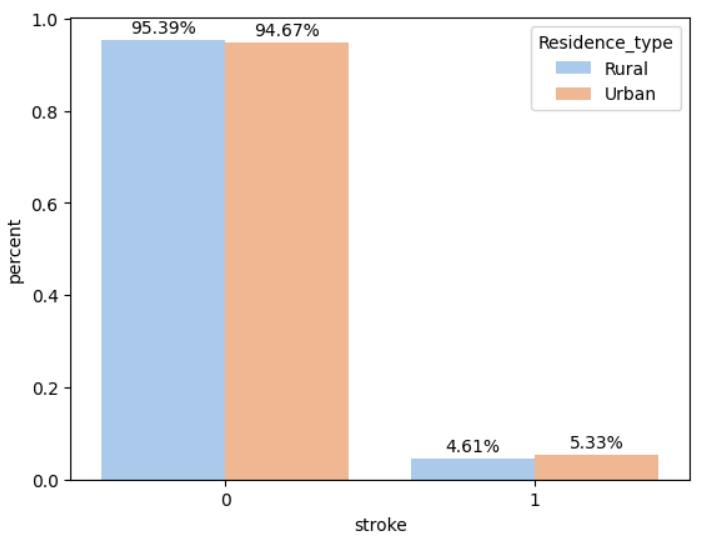
Hình 26: Kết quả của kiểm định chi – square về mối tương quan giữa tình trạng hôn nhân và đột quỵ

* Do p-value bé hơn mức ý nghĩa, nên ta bác bỏ giả thuyết H0
* Với mức ý nghĩa 5%, ta thấy có mối tương quan giữa việc lập gia đình và bệnh đột quỵ.

Kết luận: Kết hôn và đột quỵ có mối tương quan với nhau. Nhìn chung từ mốc 50 tuổi trở đi, nhóm người không kết hôn có khả năng tăng nguy cơ đột quỵ cao hơn một chút so với nhóm đã kết hôn

1. *Đột quỵ và nơi ở:*

Chia bộ dữ liệu thành thành 2 nhóm ở thành thị và nông thôn, quan sát tỉ lệ đột quỵ ở hai nhóm.



Hình 27: Biểu đồ cột thể hiện mối tương quan giữa đột quỵ và nơi ở

Tỉ lệ đột quỵ của thành thị là 5.33%, cao gấp 1.15 lần so với tỉ lệ đột quỵ của nông thôn. Sự khác biệt này là không quá rõ ràng.

Kiểm định Chi-square với mức ý nghĩa 5%:

* H0 : “Không có mối tương quan giữa nơi ở và bệnh đột quỵ.”
* H1 : “Có mối tương quan giữa nơi ở và bệnh đột quỵ.”

|  |
| --- |
|  |
|  |



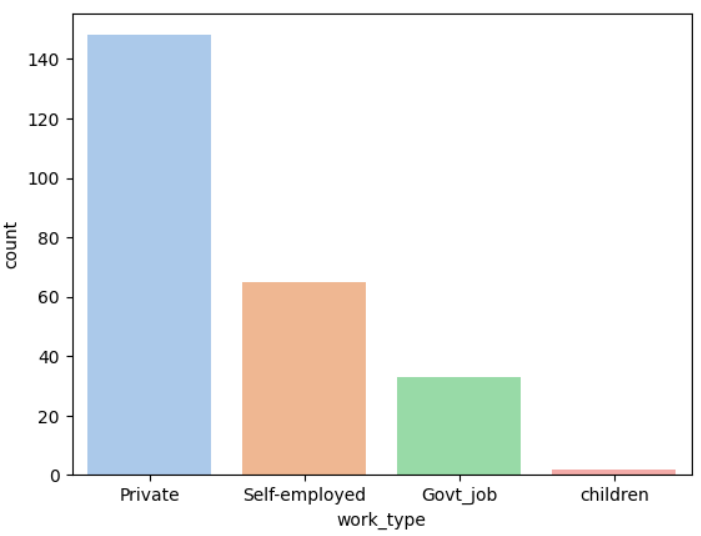
Hình 28: Kết quả của kiểm định chi – square về mối tương quan giữa đột quỵ và nơi ở

* Do p-value lớn hơn mức ý nghĩa, nên ta không thể bác bỏ giả thuyết H0
* Với mức ý nghĩa 5%, ta thấy không có mối tương quan giữa nơi ở và bệnh đột quỵ.

Kết luận: Nơi ở không có mối tương quan với bệnh đột quỵ, sự khác biệt giữa tỉ lệ người thành thị và nông thôn mắc bệnh đột quỵ là không đáng kể

1. *Đột quỵ và loại việc làm:*

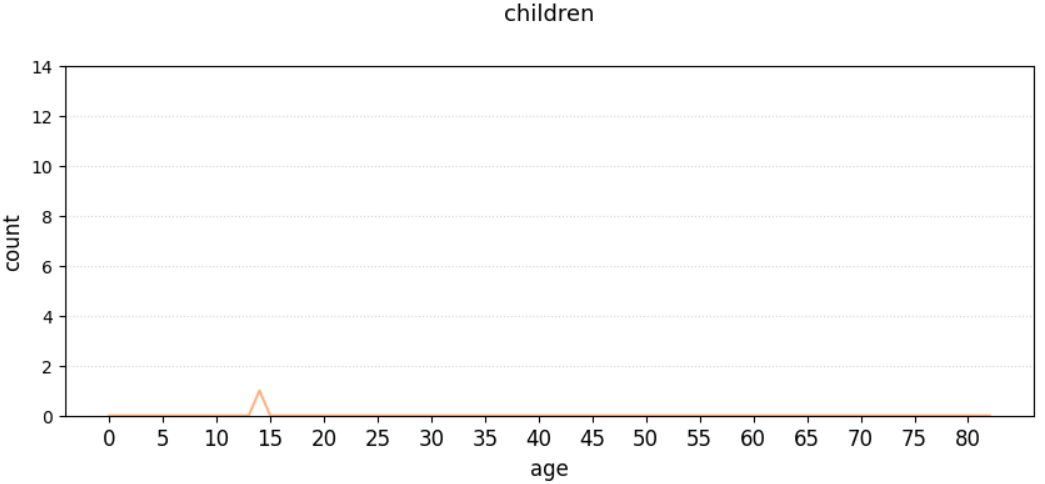
Quan sát sự phân bố của loại việc làm trong nhóm người bị đột quỵ.



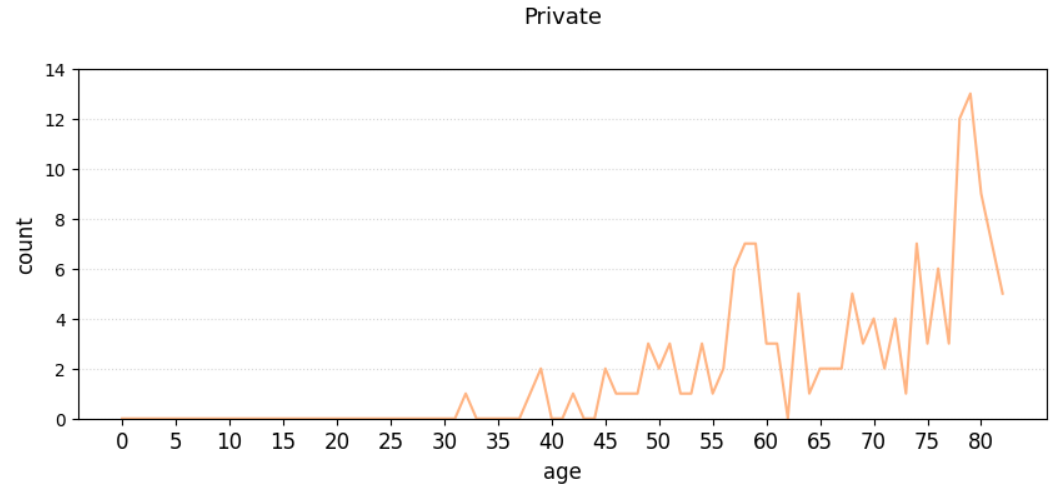
Hình 29: Biểu đồ cột thể hiện phân bố các loại công việc của nhóm bị đột quỵ

Quan sát biểu đồ trên ta thấy cao nhất là nhóm việc làm riêng tư, thấp nhất là nhóm trẻ em.

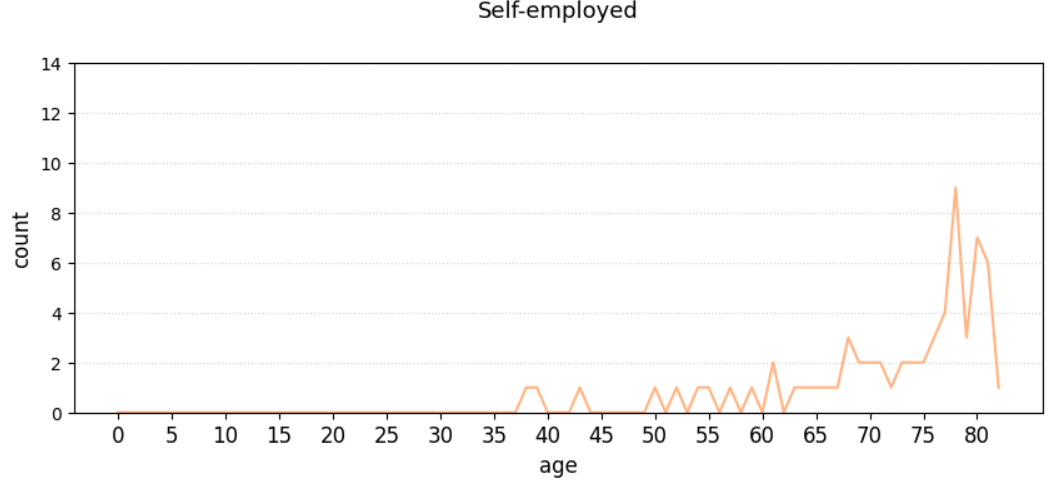
Để có thể quan sát rõ hơn sự khác biệt giữa các nhóm việc làm, ta sẽ quan sát thêm thuộc tính tuổi. Với những người bị đột quỵ, thực hiện chia thành các nhóm theo từng loại công việc, tiến hành quan sát sự phân bố theo độ tuổi.



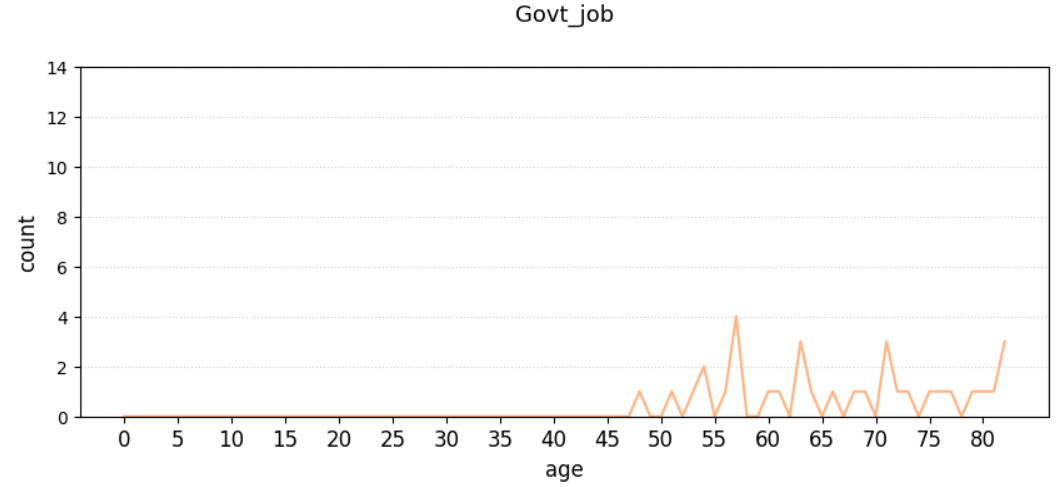
Hình 30: Biểu đồ đường thể hiện số lượng của nhóm việc làm trẻ em bị đột quỵ theo từng mốc tuổi



Hình 31: Biểu đồ đường thể hiện số lượng của nhóm việc làm riêng tư bị đột quỵ theo từng mốc tuổi



Hình 32: Biểu đồ đường thể hiện số lượng của nhóm việc làm tự kinh doanh bị đột quỵ theo từng mốc tuổi



Hình 33: Biểu đồ đường thể hiện số lượng của nhóm việc làm nhân viên chính phủ bị đột quỵ theo từng mốc tuổi

Quan sát biểu đồ:

* Một lần nữa chúng ta thấy , với độ tuổi ngày càng tăng thì nguy cơ bị đột quỵ càng cao.
* Nhóm trẻ em: chỉ có một trường hợp bị đột quỵ, điều này là dễ hiểu vì trẻ em nằm trong độ tuổi dưới 16, mà ta đã biết ở độ tuổi càng cao thì nguy cơ bị đột quỵ càng tăng, điều này lý giải tại sao nhóm trẻ em có ít nguy cơ bị đột quỵ.
* Nhóm nhân viên chính phủ: Nhóm này có nguy cơ bị đột quỵ ít hơn hai nhóm riêng tư và tự kinh doanh. Nhóm này bắt đầu xuất hiện người bị đột quỵ từ 48 tuổi và dao động về số người đột quỵ các năm sau đó là không quá lớn
* Nhóm nghề tự do: Nhóm này bắt đầu xuất hiện người bị đột quỵ từ mốc 38 tuổi, dao động về số người đột quỵ tăng mạnh trong đoạn 75 đến 80 tuổi. Ở độ tuổi 78 tuổi, số người đột quỵ lên cao nhất đạt 9 người.
* Nhóm riêng tư: bắt đầu xuất hiện người đột quỵ sau 32 tuổi và tăng mạnh khi tiến gần đến 80 tuổi. Ở tuổi 79, số người bị đột quỵ đạt cao nhất lên đến 13 người.
* Ta nhận thấy ngoài nhóm trẻ em ra, ở 3 loại việc làm còn lại, số tuổi càng cao thì số người bị đột quỵ càng tăng. Điều này có ảnh hưởng trực tiếp đến sự phân bố số người đột quỵ giữa các nhóm. Ta biết khi ở độ tuổi càng cao thì sẽ tiến dần đến độ tuổi nghỉ hưu, lúc này người nghỉ hưu thường sẽ làm các công việc trong nhóm nghề tự do và riêng tư dẫn đến nhóm này tăng mạnh số người đột quỵ. Hoặc cũng có khả năng người bị bệnh đột quỵ thì có tình trạng sức khỏe không đáp ứng được các yêu cầu làm việc trong nhóm nhân viên chính phủ. Hay nói cách khác, nhóm nhân viên chính phủ có số lượng người đột quỵ thấp hơn 2 nhóm tự kinh doanh và riêng tư là do để có thể vào làm nhân viên chính phủ phải đạt đủ tiêu chuẩn sức khỏe, nên nhóm này có nguy cơ bị đột quỵ thấp hơn.

Kiểm định Chi-square với mức ý nghĩa 5%

* H0 : “Không có mối tương quan giữa công việc và bệnh đột quỵ.”
* H1 : “Có mối tương quan giữa công việc và bệnh đột quỵ.”

|  |
| --- |
|  |
|  |

A number and numbers on a white background

Description automatically generated

Hình 34: Kết quả của kiểm định chi – square về mối tương quan giữa đột quỵ và loại việc làm

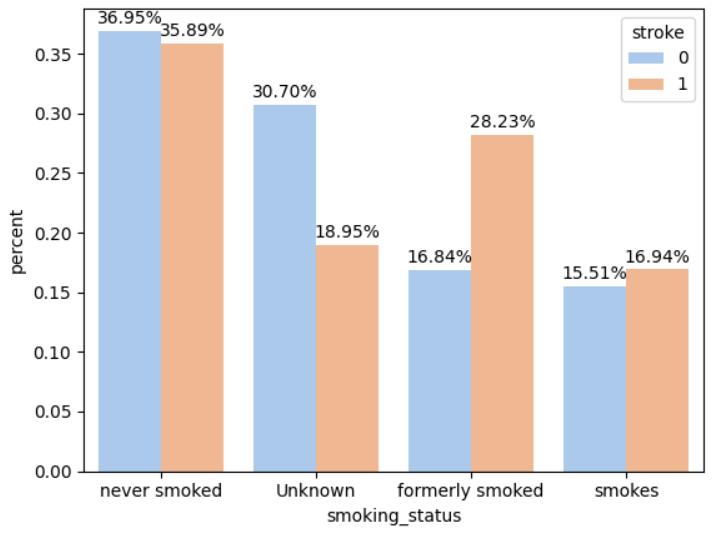
* Do p-value bé hơn mức ý nghĩa, nên ta bác bỏ giả thuyết H0
* Với mức ý nghĩa 5%, ta thấy có mối tương quan giữa loại công việc và bệnh đột quỵ.

Kết luận: Có mối tương quan giữa loại việc làm và bệnh đột quỵ. Nhóm việc làm riêng tư và nghề tự do có khả năng đột quỵ cao, nhóm nhân viên chính phủ có khả năng bị đột quỵ thấp hơn, và rất hiếm có người đột quỵ ở nhóm trẻ em.

1. *Đột quỵ và tình trạng hút thuốc:*

Theo nghiên cứu của Thư viện Y khoa Quốc gia Hoa Kỳ, những người hút thuốc có nguy cơ đột quỵ cao hơn so với những người không hút thuốc. Người hút thuốc lá trong một khoảng thời gian dài dẫn đến tăng huyết áp, từ đó khiến các mạch máu đến não bị tổn thương ở thành mạch. Lúc này, các chất béo, canxi, chất lắng đọng sẽ dễ dàng bám vào thành mạch và tạo thành các mảng xơ vữa, cục máu đông gây tắc mạch, hẹp mạch. Về lâu dài, khi lượng máu không cung cấp đủ cho não sẽ gây ra đột quỵ. [11]

Chia bộ dữ liệu thành hai 2 nhóm đột quỵ và không đột quỵ, quan sát sự phân bố giữa các loại tình trạng hút thuốc giữa 2 nhóm trên.



Hình 35: Biểu đồ cột thể hiện mối tương quan giữa đột quỵ và tình trạng hút thuốc

Quan sát biểu đồ:

* Nhóm không hút thuốc: tỉ lệ bị đột quỵ ít hơn. Sự khác biệt giữa tỉ lệ bị đột quỵ và không bị đột quỵ là không quá lớn.
* Nhóm hút thuốc: tỉ lệ bị đột quỵ cao hơn so với không bị đột quỵ, tuy nhiên sự khác biệt là không quá đáng kể.
* Nhóm không biết: tỉ lệ bị không bị đột quỵ là 30.7%, cao hơn 1.62 lần so với bị đột quỵ.
* Nhóm đã từng hút thuốc: tỉ lệ bị đột quỵ là 28.23%, cao hơn 1.67 lần so với không bị đột quỵ.

Kiểm định chi-square với mức ý nghĩa 5%:

* H0 : “Không có mối tương quan giữa việc hút thuốc và bệnh đột quỵ.”
* H1 : “Có mối tương quan giữa việc hút thuốc và bệnh đột quỵ.”

|  |
| --- |
|  |
|  |

A close up of numbers

Description automatically generated

Hình 36: Kết quả của kiểm định chi – square về mối tương quan giữa đột quỵ và tình trạng hút thuốc

* Do p-value bé hơn mức ý nghĩa, nên ta bác bỏ giả thuyết H0
* Với mức ý nghĩa 5%, ta thấy có mối tương quan giữa việc hút thuốc và bệnh đột quỵ.

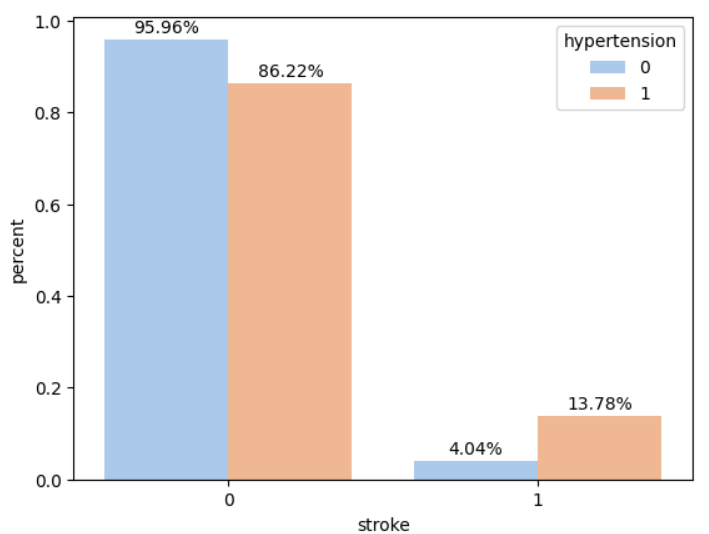
Kết luận:

* Đột quỵ và tình trạng hút thuốc có mối tương quan với nhau.
* Hút thuốc có khả năng làm tăng nguy cơ bị đột quỵ. Điều này phù hợp với nghiên cứu của y khoa.
* Ngừng hút thuốc lá làm tăng nguy cơ bị đột quỵ cao hơn so với nhóm hút thuốc. Điều này trái lại với nghiên cứu của y khoa, khi đáng lý ra việc bỏ hút thuốc mang lại các tác động tích cực cho sức khỏe và giảm các nguy cơ mắc đột quỵ hơn so với nhóm hút thuốc. Điều này còn phụ thuộc vào thời gian hút thuốc và thời gian bỏ hút thuốc. Nhưng các dữ liệu này không được thể hiện trong bộ dữ liệu nên không thể đưa ra kết luận chính xác.

1. *Đột quỵ và cao huyết áp:*

Theo nghiên cứu của Thư viện Y khoa Quốc gia Hoa Kỳ, cao huyết áp luôn gây những ảnh hưởng sâu sắc đến cấu trúc của mạch máu não. Tăng huyết áp thúc đẩy sự phát triển của mảng xơ vữa động mạch trên động mạch não và tiểu động mạch, có thể dẫn đến tắc động mạch và tổn thương do thiếu máu cục bộ. Đồng thời, tăng huyết áp cũng dẫn đến chứng xơ cứng mạch máu, dẫn đến tăng áp lực mạch, trở thành một yếu tố dự báo đột quỵ. [12]

Chia bộ dữ liệu thành 2 nhóm bị cao huyết áp và không bị cao huyết áp, quan sát sự phân bố của người bị đột quỵ và không đột quỵ trên 2 nhóm trên.



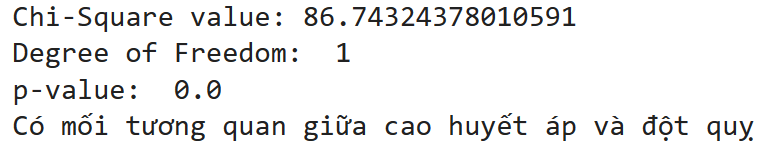
Hình 37: Biểu đồ cột thể hiện mối tương quan giữa đột quỵ và cao huyết áp

Quan sát biểu đồ: Tỉ lệ bị đột quỵ của nhóm bị cao huyết áp là 13,78%, cao gấp 3.4 lần so với tỉ lệ bị đột quỵ của nhóm không bị cao huyết áp

Kiểm định Chi-square với mức ý nghĩa 5%:

* H0 : “Không có mối tương quan giữa bệnh cao huyết áp và bệnh đột quỵ.”
* H1 : “Có mối tương quan giữa giới tính và bệnh cao huyết áp và bệnh đột quỵ.”

|  |
| --- |
|  |
|  |



Hình 38: Kết quả kiểm định chi – square về mối tương quan giữa đột quỵ và cao huyết áp

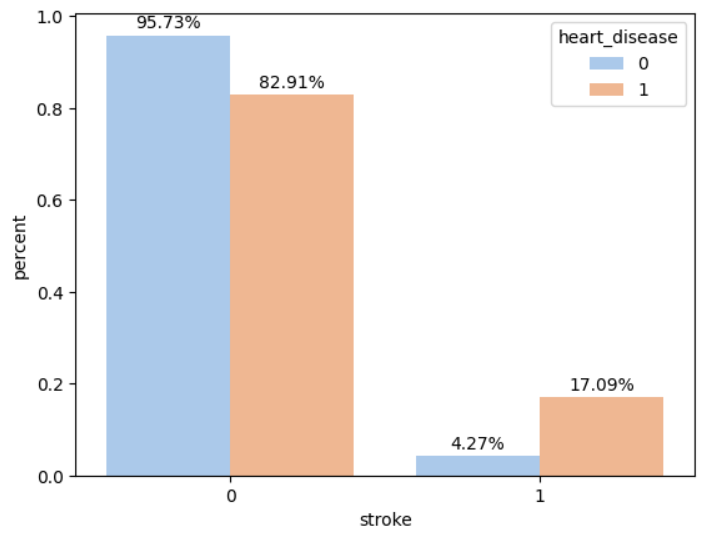
* Do p-value bé hơn mức ý nghĩa, nên ta bác bỏ giả thuyết H0
* Với mức ý nghĩa 5%, ta thấy có mối tương quan giữa bệnh cao huyết áp và bệnh đột quỵ.

Kết luận: Cao huyết áp có mối tương quan với đột quỵ, cao huyết áp làm tăng khả năng mắc đột quỵ, điều này đúng với nghiên cứu của y khoa

1. *Đột quỵ và bệnh tim mạch:*

Theo nghiên cứu của Thư viện Y khoa Quốc gia Hoa Kỳ, bị bệnh tim mạch làm tăng nguy cơ đột quỵ. Bệnh tim mạch có thể góp phần vào nguy cơ phát triển đột quỵ. Một số điều kiện tim mạch, chẳng hạn như bệnh động mạch vành, có thể dẫn đến sự tắc nghẽn của mạch máu. Khi một khu vực của não không nhận đủ máu, có thể xảy ra đột quỵ. [13]

Chia bộ dữ liệu thành 2 nhóm bị bệnh tim và không bị bệnh tim, quan sát sự phân bố của người bị đột quỵ và không bị đột quỵ trên hai nhóm này



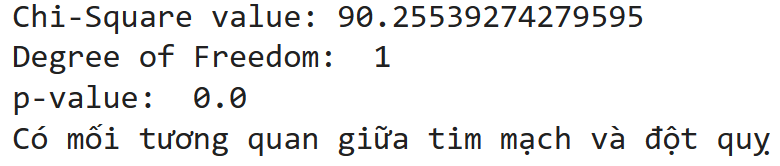
Hình 39: Biểu đồ cột thể hiện mối tương quan giữa đột quỵ và bệnh tim mạch

Quan sát biểu đồ: tỉ lệ người bị đột quỵ của nhóm bị bệnh tim là 17.09%, cao gấp 4 lần so với tỉ lệ bị đột quỵ của nhóm không bị bệnh tim.

Kiểm định Chi-square với mức ý nghĩa 5%:

* H0 : “Không có mối tương quan giữa bệnh tim mạch và bệnh đột quỵ.”
* H1 : “Có mối tương quan giữa bệnh tim mạch và bệnh đột quỵ.”

|  |
| --- |
|  |
|  |



Hình 40: Kết quả kiểm định chi – square về mối tương quan giữa đột quỵ và bệnh tim mạch

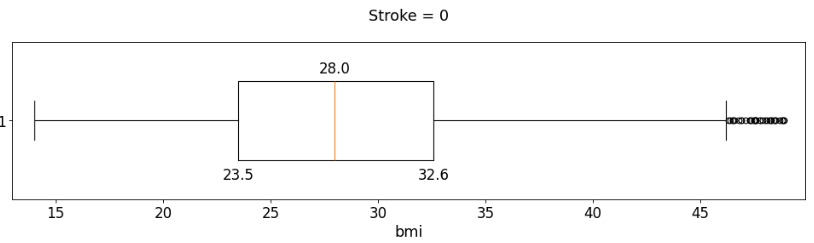
* Do p-value bé hơn mức ý nghĩa, nên ta bác bỏ giả thuyết H0
* Với mức ý nghĩa 5%, ta thấy có mối tương quan giữa bệnh tim mạch và bệnh đột quỵ.

Kết luận: Bệnh tim mạch có mối tương quan với đột quỵ. Bệnh tim có khả năng làm tăng nguy cơ bị đột quỵ. Điều này phù hợp với nghiên cứu của y khoa.

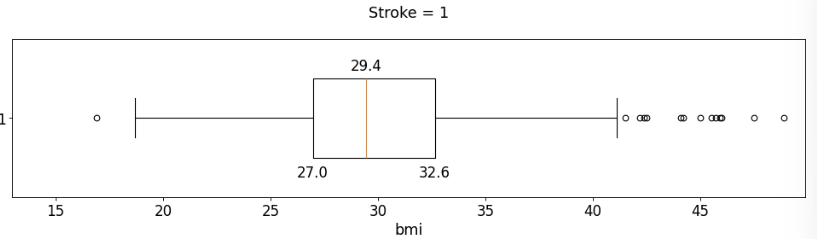
1. *Đột quỵ và chỉ số BMI:*

Theo nghiên cứu của Thư viện Y khoa Quốc gia Hoa Kỳ, người bị thừa cân và béo phì làm tăng nguy cơ mắc bệnh đột quỵ hơn so với người có chỉ số BMI bình thường. [14]

Chia bộ dữ liệu thành 2 nhóm đột quỵ và không bị đột quỵ, quan sát sự phân bố của chỉ số BMI.



Hình 41: Biểu đồ hộp thể hiện phân bố chỉ số BMI của nhóm không bị bệnh đột quỵ



Hình 42: Biểu đồ hộp thể hiện phân bố chỉ số BMI của nhóm bị bệnh đột quỵ

Quan sát biểu đồ:

* Chỉ số min của biểu đồ hộp BMI của nhóm bị đột quỵ xấp xỉ 18, cao hơn nhiều so với chỉ số min của nhóm bị đột quỵ. Tuy nhiên chỉ số max của nhóm bị đột quỵ lại thấp hơn so với chỉ số max của nhóm không bị đột quỵ.
* Nhóm không bị đột quỵ có chỉ số Q1 = 23.5 và Q2 = 28, trong khi nhóm bị đột quỵ chỉ số Q1 = 27 và Q2 = 29.4, cao hơn so với nhóm không bị đột quỵ. Nhóm bị đột quỵ có trên 75% thuộc nhóm thừa cân và béo phì (BMI ≥ 25) và có xấp xỉ 50% thuộc nhóm béo phì (BMI ≥ 30).

Kiểm định Chi-square với mức ý nghĩa 5%:

* H0 : “Không có mối tương quan giữa chỉ số BMI và bệnh đột quỵ.”
* H1 : “Có mối tương quan giữa chỉ số BMI và bệnh đột quỵ.”

|  |
| --- |
|  |
|  |

A number and numbers on a white background

Description automatically generated

Hình 43: Kết quả kiểm định Chi – square về mối tương quan giữa đột quỵ và chỉ số BMI

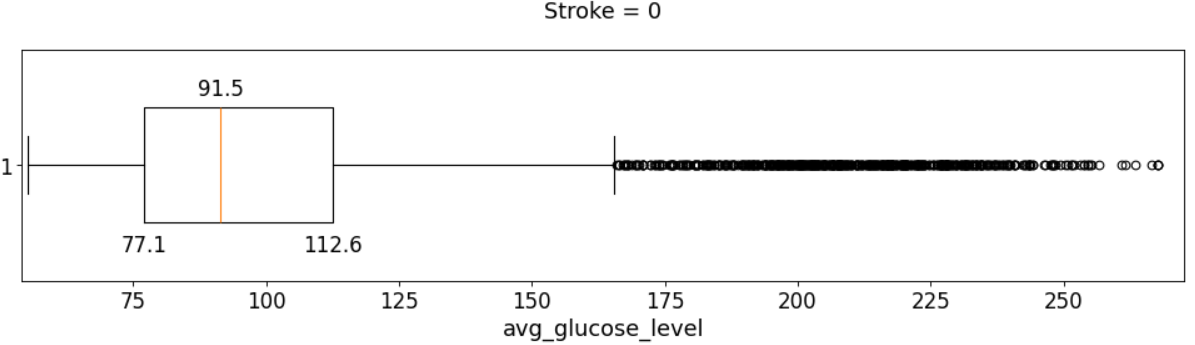
* Do p-value bé hơn mức ý nghĩa, nên ta bác bỏ giả thuyết H0
* Với mức ý nghĩa 5%, ta thấy có mối tương quan giữa bệnh đột quỵ và chỉ số BMI.

Kết luận: Nếu chỉ số BMI thuộc nhóm thừa cân và béo phì thì sẽ làm tăng nguy cơ mắc bệnh đột quỵ. Điều này phù hợp với nghiên cứu của y khoa.

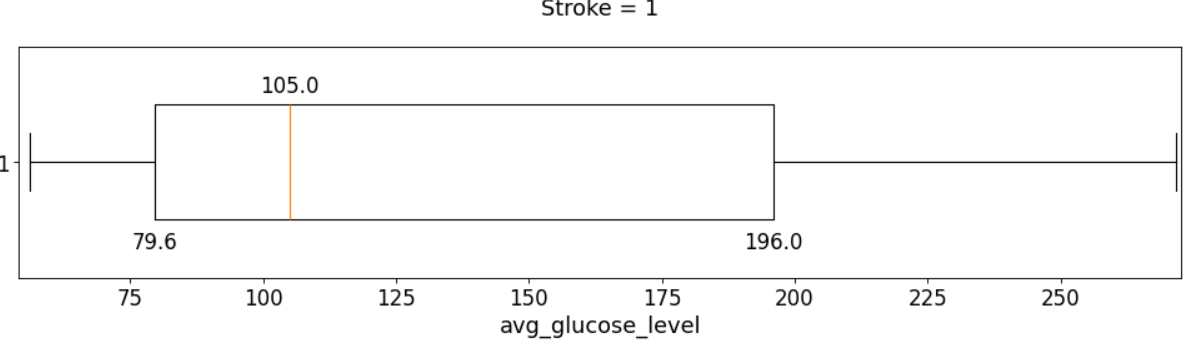
1. *Đột quỵ và chỉ số đường huyết trung bình:*

Theo nghiên cứu của Thư viện Y khoa Quốc gia Hoa Kỳ, người mắc bệnh tiểu đường thì có nguy cơ mắc bệnh đột quỵ cao gấp 1.5 - 2 lần so với người không mắc bệnh tiểu đường, nguy cơ tăng dần theo thời gian mắc bệnh tiểu đường [15]

Chia bộ dữ liệu thành 2 nhóm bị đột quỵ và không bị đột quỵ, quan sát sự phân bố của chỉ số đường huyết trung bình trên hai nhóm này



Hình 44: Biểu đồ hộp thể hiện phân bố chỉ số đường huyết trung bình của nhóm không bị bệnh đột quỵ



Hình 45: Biểu đồ hộp thể hiện phân bố chỉ số đường huyết trung bình của nhóm bị bệnh đột quỵ

Quan sát biểu đồ:

* Nhóm không bị đột quỵ: max của chỉ số đường huyết trung bình thấp hơn 175, có trên 75% thuộc nhóm đường huyết dành cho người bình thường (đường huyết < 140), chỉ có một phần nhỏ thuộc nhóm tiền tiểu đường (140 ≤ đường huyết < 200 ) và có rất nhiều ngoại lai nằm trên mốc 175. Nếu ta không quan tâm đến các điểm ngoại lệ của biểu đồ này thì không xuất hiện người bị béo phì thuộc nhóm này.
* Nhóm bị đột quỵ: chỉ số Q2, Q3 và max của biểu đồ cao hơn rất nhiều so với nhóm không bị đột quỵ. Có xấp xỉ 25% người bị béo phì trong nhóm này (đường huyết > 200).

Kiểm định Chi-square với mức ý nghĩa 5%:

* H0 : “Không có mối tương quan giữa lượng đường huyết trung bình và bệnh đột quỵ.”
* H1 : “Có mối tương quan giữa lượng đường huyết trung bình và bệnh đột quỵ.”

|  |
| --- |
|  |
|  |

A black text on a white background

Description automatically generated

Hình 46: Kết quả kiểm định Chi-square về mối tương quan giữa đột quỵ và chỉ số đường huyết trung bình

* Do p-value bé hơn mức ý nghĩa, nên ta bác bỏ giả thuyết H0
* Với mức ý nghĩa 5%, ta thấy có mối tương quan giữa chỉ số đường huyết trung bình và bệnh đột quỵ.

Kết luận: chỉ số đường huyết trung bình có ảnh hưởng nhiều đến đột quỵ. Nếu chỉ số đường huyết thuộc nhóm bị tiểu đường sẽ có nguy cơ cao mắc bệnh đột quỵ. Điều này phù hợp với nghiên cứu của y khoa

**Đánh giá độ tin cậy của bộ dữ liệu:** Dựa trên các phân tích về mối tương quan giữa các đặc trưng của bộ dữ liệu phía trên, bộ dữ liệu đã thể hiện được phần lớn các nghiên cứu của y khoa. Tuy nhiên vẫn còn một số ít các nghiên cứu của y khoa mà bộ dữ liệu không thể hiện được hoặc không đủ tiêu chí đánh giá do bộ dữ liệu không đủ các thông tin cần có để xem xét. Kết luận: Bộ dữ liệu có độ tin cậy cao.

1. **Xây dựng mô hình dự đoán bệnh đột quỵ:**
   1. **Tiền xử lí dữ liệu**

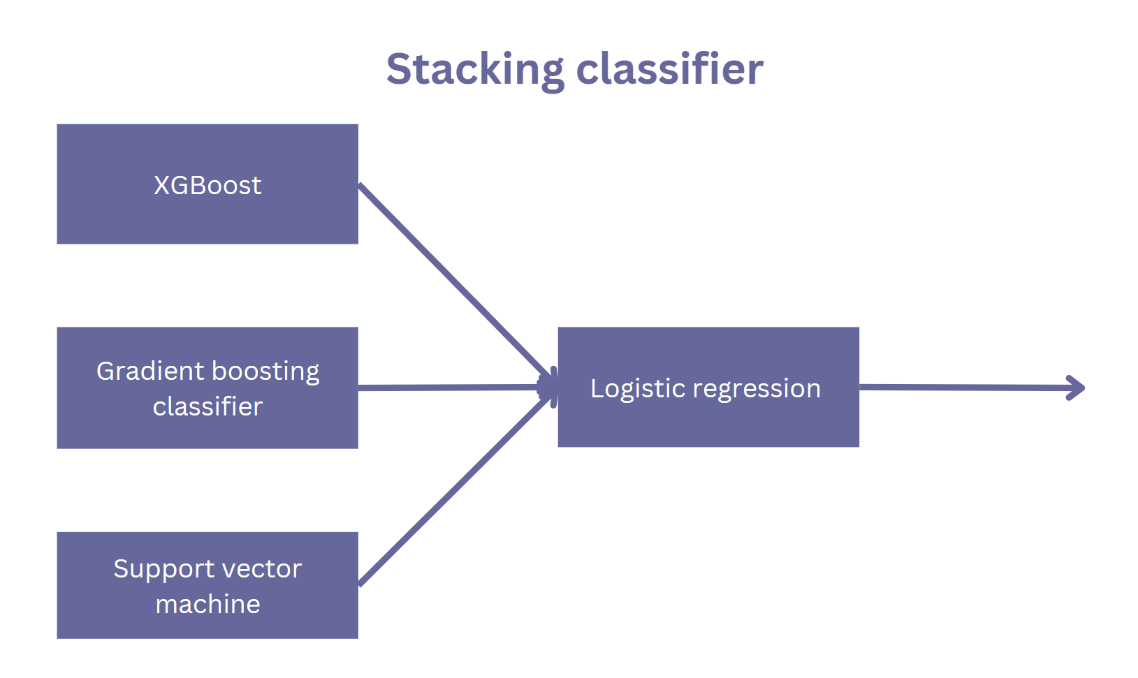
Mã hóa dữ liệu: Ta sẽ biến các trường dữ liệu là kiểu object thành kiểu dữ liệu số hết. Đối với trường dữ liệu smokin\_status có thư tự nên ta chuyển never smoke: 0, Unknown: 1,formerly smoked: 2,smokes: 3

Xử lí bộ dữ liệu mất cân bằng: Do tỉ lệ số người bị đột quỵ và không bị đột quỵ trên nhau quá lớn do đó ta sẽ sử dụng phương pháp SMOTE để xử lí dữ liệu mất cân bằng

Chuẩn hóa dữ liệu: Ta dùng MinMaxScaler để chuẩn hóa dữ liệu đối với các trường dữ liệu age, avg\_glucose\_level, BMI,smoking\_status\_encoded để giúp tăng tính ổn định khi xây dựng, huấn luyện mô hình

* 1. **Huấn luyện**

Ta dùng các mô hình học máy để huấn luyện mô hình dự đoán nguy cơ đột quỵ thông qua các dữ liệu lâm sàn

Các mô hình ta dùng để đánh giá bao gồm Logistic regression, Gradient boosting classifier, Support vector machine, XGBoost và cuối cùng là thuật toán học máy Stacking lấy XGBoost, Gradient boosting, Support vector machine, làm công cụ ước tính và dùng Logistic regression làm công cụ ước tính cuối cùng

Hình 47: Sơ đồ của thuật toán học máy Stacking

* 1. **Đánh giá**

Ta sẽ đánh giá độ chính xác, độ tốt của mô hình thông qua các thang đo accuracy, recall. Vì đây là bài toán dự đoán bệnh do đó ta sẽ ưu tiên thang đo recall hơn precision nên ta không sử dụng thang đo presicion để đánh giá ở đây

Hình 48: Biểu đồ đánh giá mô hình

Nhìn vào biểu đồ ta có thể thấy được Stacking đang hoạt động tốt nhất và nhỉnh hơn các mô hình, thuật toán khác một ít với accuracy là 0.963 và recall là 0.9409 cho nên ta sẽ dùng mô hình này để chuẩn đoán bệnh

1. **Kết luận:**

Biểu đồ dưới đây mô tả mối tương quan giữa các yếu tố với nhau, tuy nhiên ta không kết luận yếu tố này là nguyên nhân gây ra hay dẫn đến yếu tố khác, hay nói cách khác giữa hai yếu tố không mang tính nhân quả. Mà chỉ mang tính tương quan, tức là giá trị mô tả yếu tố này tăng thì giá trị mô tả yếu tố tương quan với yếu tố này cũng tăng và ngược lại. Vỉ dụ: Chỉ số BMI cao, thì khả năng mắc đột quỵ cũng cao, tuy nhiên BMI không phải là nguyên dân dẫn đến bệnh đột quỵ. Người có BMI cao, thì có khả năng có nhiều mỡ tích tụ, động mạch cũng có thể tắc nghẽn (đây mới chính là nguyên nhân dẫn tới đột quỵ theo nghiên cứu khoa học).

**A diagram of stroke

Description automatically generated**

**Kết luận:**

Ngoại trừ các yếu tố : loại công việc, tình trạng hôn nhân do chủ quan con người xen vào và tuổi – yếu tố do thời gian tác động, một quy luật tự nhiên của cuộc sống, thì lối sống và chế độ ăn là biến ẩn (lurking variable) ảnh hưởng lên tất cả các yếu tố trên. Do đó cần có chế độ ăn và lối sống lành mạnh, siêng tập thể dục thể thao, hạn chế ăn đồ dầu mỡ, cay nóng.

**A diagram of a blood sugar level

Description automatically generated**

**TÀI LIỆU THAM KHẢO**

|  |  |
| --- | --- |
| [1] | Obesity and insulin resistance. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC380258/. |
| [2] | Correlation of Body Mass Index (BMI) with Saliva and Blood Glucose Levels in Diabetic and Non-Diabetic Patients. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10485520/. |
| [3] | Association of Body Mass Index With Lifetime Risk of Cardiovascular Disease and Compression of Morbidity. https://pubmed.ncbi.nlm.nih.gov/29490333/. |
| [4] | Heart Health and Aging. https://www.nia.nih.gov/health/heart-health/heart-health-and-aging. |
| [5] | Gender differences in coronary heart disease. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3018605/. |
| [6] | Cardiovascular risk of smoking and benefits of smoking cessation. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7399440/. |
| [7] | Hypertension and Aging. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4768730/. |
| [8] | Hypertension and Coronary Heart Disease. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6169140/. |
| [9] | Stroke and gender. https://pubmed.ncbi.nlm.nih.gov/11252851/. |
| [10] | Aging and ischemic stroke. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6535078/. |
| [11] | The relationship between smoking and stroke, Smoking and stroke: the more you smoke the more you stroke. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2928253/. |
| [12] | Hypertension and Stroke: Update on Treatment. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6659031/. |
| [13] | Large Burden of Stroke Incidence in People with Cardiac Disease: A Linked Data Cohort Study. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9945299/. |
| [14] | Association of Body Mass Index with Ischemic and Hemorrhagic Stroke. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8308685/. |
| [15] | Diabetes and Stroke: What Are the Connections? https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9911852/. |