**BỘ GIÁO DỤC VÀ ĐÀO TẠO**Shape, square

Description automatically generated

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

**KHOA CÔNG NGHỆ THÔNG TIN**

**🙠🙟🕮🙝🙢**



**ĐỒ ÁN CUỐI KÌ**

**PHÂN TÍCH DỮ LIỆU VÀ XÂY DỰNG DASHBOARD TỪ TẬP DỮ LIỆU VỀ ĐỘT QUỴ BẰNG STREAMLIT**

**Môn: Tương tác dữ liệu**

**Thực hiện**: Nhóm 9. Thứ 4, tiết 1, 4.

**GVHD**: Lê Minh Tân

**Thành phố Hồ Chí Minh, tháng 05 năm 2023**

**DANH SÁCH NHÓM LÀM ĐỒ ÁN CUỐI KỲ  
MÔN TƯƠNG TÁC DỮ LIỆU  
HỌC KỲ II NĂM HỌC 2022-2023**

**1.** **Mã lớp môn học**: IDVI333677\_22\_2\_02 (Thứ 4. tiết 1, 4)  
**2. Giảng viên hướng dẫn**: Lê Minh Tân  
**3. Tên đề tài:** Phân tích dữ liệu và xây dựng dashboard từ tập dữ liệu về đột quỵ bằng Streamlit  
**4. Danh sách nhóm viết tiểu luận cuối kỳ:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Mã số sinh viên** | **Họ và tên** | **Phân công** | **% Hoàn tất** | **Điểm** |
| 20133084 | Nguyễn Trị Quốc | Làm sạch dữ liệu, phân tích dữ liệu & vẽ biểu đồ tròn, biểu đồ cột. | 100% |  |
| 20133105 | Mai Thành Trung | Lọc dữ liệu, vẽ biểu đồ hộp, biểu đồ nhiệt, biểu đồ điểm. | 100% |  |
| 20133092 | Nguyễn Minh Tiến | Phân tích dữ liệu & vẽ biểu đồ điểm, đường xu hướng. | 100% |  |

- Tỷ lệ % = 100%- Trưởng nhóm: Nguyễn Trị Quốc

**5. Bảng đánh giá**

|  |  |  |  |
| --- | --- | --- | --- |
| STT | Nội dung | **Điểm** | Cơ sở đánh giá |
| 1 | Hình thức: Bìa, lời cảm ơn, lời mở đầu, mục lục, canh lề, đúng font, trang đánh giá & chấm điểm, chú thích, danh mục tài liệu tham khảo, ít lỗi chính tả,... | …/2 |  |
| 2 | Trình bày đầy đủ nội dung: Giới thiệu dữ liệu, cách xử lý dữ liệu, giải thích đồ thị/biểu đồ được chọn và cách vẽ, công nghệ/phần mềm sử dụng, kết luận.  *Mục này chỉ xét tính đầy đủ nội dung. Chất lượng sẽ xét theo các mục dưới.* | …/2 |  |
| 3 | Chạy được dashboard phù hợp nội dung báo cáo và yêu cầu đề ra (lập trình desktop app/web app/Zeppelin). | …/1 |  |
| 4 | Chất lượng dashboard: Ít nhất 01 biểu đồ/đồ thị có thể tương tác bằng cách trỏ chuột (hiện chú thích hoặc đổi màu). | .../0.5 |  |
| 5 | Chất lượng dashboard: Ít nhất 01 biểu đồ/đồ thị có thể thay đổi biểu đồ bằng trường nhập dữ liệu như: textbox, combobox,... | .../1 |  |
| 6 | Chất lượng dashboard: Dữ liệu cập nhật lại chậm nhất là 1 phút. Thời gian vẽ lại đồ thị/biểu đồ không quá 5 giây.  *Sinh viên demo bằng cách thay đổi dữ liệu. Đợi tối đa 1 phút để xem kết quả bên dashboard.* | .../0.5 |  |
| 7 | Dữ liệu: Dùng ít nhất 02 bảng dữ liệu có ý nghĩa khác nhau. *Sinh viên thể hiện trên code.* | .../0.5 |  |
| 8 | Dữ liệu: Có thực hiện ít nhất hai loại thao tác dữ liệu: xóa cột, nhóm (group), nối (concat), kết hợp (join), lọc (filter), thay thế cột bằng tính toán dữ liệu. *Sinh viên thể hiện trên code.* | .../0.75 |  |
| 9 | Biểu đồ: Có chú thích, tên biểu đồ, giá trị trên trục tùy theo loại biểu đồ. | .../0.25 |  |
| 10 | Biểu đồ: Có ít nhất 02 biểu đồ/đồ thị cơ bản có ý nghĩa: Tròn, cột, đường, miền, điểm (scatter).  *Sinh viên chỉ các biểu đồ/đồ thị trên dashboard.* | .../0.75 |  |
| 11 | Biểu đồ: Có ít nhất 01 biểu đồ/đồ thị nâng cao có ý nghĩa: Histogram, biểu đồ hộp, biểu đồ nến, bản đồ (nhiệt, địa lý), mạng, lưới, tọa độ song song.  *Sinh viên chỉ các biểu đồ/đồ thị trên dashboard.* | .../0.5 |  |
| 12 | Đường thống kê: Có ít nhất 1 đường thống kê: Đường trung bình, đường xu hướng, đường mục tiêu. | .../0.25 |  |
|  | Tổng: | .../10 |  |
|  | **Trưởng nhóm ký, ghi họ tên** **Giảng viên ký, ghi họ tên** | | |

***Nhận xét của giáo viên:***……………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………….………………………………………………………

*Tháng 05 năm 2023*  
**Giáo viên chấm điểm**

# **LỜI CẢM ƠN**

Lời đầu tiên, nhóm em xin được gởi lời cảm ơn đặc biệt đến Thầy Lê Minh Tân - Giảng viên phụ trách môn Tương tác dữ liệu – trường đại học Sư Phạm Kỹ Thuật Tp.Hồ Chí Minh.

Trong thời gian nhóm em làm đồ án, tụi em đã nhận được nhiều sự giúp đỡ từ thầy. Thầy đã cung cấp đầy đủ kiến thức, chỉ bảo và đóng góp những ý kiến quý báu giúp tụi em có thể hoàn thành được đồ án môn học của mình một cách tốt nhất.

Xuất phát từ mục đích học tập, tìm hiểu xâu hơn các kiến thức về tương tác dữ liệu, cũng như tìm hiểu kỹ về quy trình nghiệp vụ của lên ý tưởng, xây dựng mô hình dự đoán, thống kê dữ liệu. Nhóm chúng em đã thực hiện đồ án “Phân tích dữ liệu và xây dựng dashboard từ tập dữ liệu về đột quỵ bằng Streamlit”. Trong quá trình thực hiện đồ án, dựa trên kiến thức được Thầy cung cấp qua các buổi học lý thuyết cũng như thực hành trên lớp, kết hợp với việc tự tìm hiểu những công cụ và kiến thức mới, nhóm đã cố gắng thực hiện đồ án một cách tốt nhất. Tuy nhiên, đồ án còn chưa được hoàn thiện và có nhiều sai sót.

Nhóm rất mong nhận được sự góp ý từ thầy nhằm rút ra những kinh nghiệm quý báu và hoàn thiện vốn kiến thức để nhóm có thể hoàn thành những đồ án, dự án khác trong tương lai.

Nhóm chúng em xin chân thành cảm ơn quý thầy!

# **LỜI MỞ ĐẦU**

Đột quỵ là một trong những nguyên nhân hàng đầu gây tử vong và tàn tật trên toàn cầu. Tuy nhiên, nếu được phát hiện sớm và điều trị kịp thời, tỷ lệ sống sót và hồi phục của bệnh nhân đột quỵ có thể được cải thiện đáng kể. Trong nỗ lực giảm thiểu tác động của đột quỵ, nghiên cứu về các yếu tố nguy cơ và phát hiện sớm đột quỵ đã trở thành một chủ đề quan tâm lớn trong ngành y tế.

Tập dữ liệu dự đoán đột quỵ được công bố trên Kaggle bao gồm 5110 bản ghi về các yếu tố nguy cơ và kết quả của bệnh nhân trong quá trình dự đoán đột quỵ. Trong bài tiểu luận này, nhóm em sẽ phân tích tập dữ liệu này và tìm hiểu các yếu tố nguy cơ có liên quan đến đột quỵ, cũng như đưa ra các dự đoán về khả năng bệnh nhân gặp phải đột quỵ trong tương lai.

**Mục Lục**

[**LỜI CẢM ƠN** 4](#_heading=h.30j0zll)

[**LỜI MỞ ĐẦU** 5](#_heading=h.1fob9te)

[**A. PHẦN MỞ ĐẦU** 8](#_heading=h.3znysh7)

[**1. Lý do chọn đề tài** 8](#_heading=h.2et92p0)

[**2. Mục đích của đề tài** 8](#_heading=h.tyjcwt)

[**2.1. Mục đích** 8](#_heading=h.3dy6vkm)

[**2.2. Mục tiêu** 8](#_heading=h.1t3h5sf)

[**3. Phương pháp nghiên cứu** 8](#_heading=h.4d34og8)

[**3.1. Về mặt lý thuyết** 8](#_heading=h.2s8eyo1)

[**3.2. Đối tượng nghiên cứu** 8](#_heading=h.17dp8vu)

[**3.3. Phạm vi nghiên cứu** 9](#_heading=h.3rdcrjn)

[**4. Kết quả dự kiến đạt được** 9](#_heading=h.26in1rg)

[**B. PHẦN NỘI DUNG** 9](#_heading=h.lnxbz9)

[**I.  Giới thiệu về tập dữ liệu** 9](#_heading=h.35nkun2)

[**1. Nguồn gốc** 9](#_heading=h.1ksv4uv)

[**2. Mô tả tập dữ liệu** 9](#_heading=h.44sinio)

[**II. Giới thiệu về công nghệ sử dụng** 10](#_heading=h.2jxsxqh)

[**1. Python** 10](#_heading=h.z337ya)

[**2. Streamlit** 10](#_heading=h.3j2qqm3)

[**3. Một số thư viện** 10](#_heading=h.1y810tw)

[**III.   Cách xử lý dữ liệu** 11](#_heading=h.4i7ojhp)

[**1. Đọc dữ liệu vào chương trình** 11](#_heading=h.2xcytpi)

[**2. Thực hiện yêu cầu tốc độ cập nhật dữ liệu của chương trình** 11](#_heading=h.1ci93xb)

[**3. Xem thông tin từng cột** 12](#_heading=h.3whwml4)

[**4. Lược bỏ các giá trị Other trong cột Gender** 12](#_heading=h.2bn6wsx)

[**5. Bổ sung các giá trị bị thiếu trong tập dữ liệu** 12](#_heading=h.qsh70q)

[**6. Từ tập dữ liệu tạo ra hai bảng dữ liệu đột quỵ và không đột quỵ** 12](#_heading=h.3as4poj)

[**7. Xử lý dữ liệu trong lúc vẽ đồ thị** 13](#_heading=h.1pxezwc)

[**IV. Vẽ biểu đồ** 13](#_heading=h.49x2ik5)

[**1. Biểu đồ tròn biểu thị biểu diễn theo phần trăm so sánh giữa những người đột quỵ và không đột quỵ theo các thuộc tính** 13](#_heading=h.2p2csry)

[**2. Biểu đồ cột thể hiện số lượng người đột quỵ và không đột quỵ theo các thuộc tính** 14](#_heading=h.147n2zr)

[**3. Biểu đồ nhiệt thể hiện sự tương quan giữa các thuộc tính** 15](#_heading=h.3o7alnk)

[**4. Biểu đồ nến thể hiện sự phân bố của đột quỵ dựa trên giới tính và độ tuổi** 16](#_heading=h.23ckvvd)

[**5. Biểu đồ điểm thể hiện sự đột quỵ giữa độ tuổi và lượng đường trong máu** 17](#_heading=h.ihv636)

[**C. PHẦN KẾT LUẬN** 18](#_heading=h.32hioqz)

[**1. Kết luận** 18](#_heading=h.1hmsyys)

[**2. Kết quả đạt được** 18](#_heading=h.41mghml)

[**3. Hạn chế** 18](#_heading=h.2grqrue)

[**4. Tài liệu tham khảo** 18](#_heading=h.vx1227)

# **A. PHẦN MỞ ĐẦU**

## **1. Lý do chọn đề tài**

Đề tài về dự đoán đột quỵ là một chủ đề quan trọng trong lĩnh vực y tế, vì đột quỵ là một trong những nguyên nhân hàng đầu của tử vong và tàn tật trên toàn thế giới. Vì vậy, việc phát triển các công cụ dự đoán đột quỵ có thể giúp chẩn đoán và điều trị sớm, từ đó cải thiện kết quả điều trị và tăng cơ hội sống sót cho bệnh nhân.

Dữ liệu về dự đoán đột quỵ có thể được thu thập từ nhiều nguồn, và Kaggle là một trong những nguồn phổ biến để tìm kiếm các bộ dữ liệu cần thiết. Bộ dữ liệu Stroke Prediction trên Kaggle chứa thông tin của hơn 5000 bệnh nhân về các yếu tố nguy cơ đột quỵ như độ tuổi, giới tính, hút thuốc, tiểu đường, cao huyết áp, các bệnh lý tim mạch, hành động thể chất và cân nặng. Với các thông tin này, chúng ta có thể áp dụng các kỹ thuật học máy để xây dựng mô hình dự đoán đột quỵ.

Vì vậy, chọn đề tài sử dụng bộ dữ liệu Stroke Prediction trên Kaggle là một sự lựa chọn tốt để thực hiện các nghiên cứu và phát triển các mô hình dự đoán đột quỵ.

## **2. Mục đích của đề tài**

### **2.1. Mục đích**

Để nghiên cứu về các yếu tố ảnh hưởng đến nguy cơ đột quỵ và dự đoán khả năng mắc bệnh này dựa trên các thông tin y tế và lối sống của bệnh nhân

### **2.2. Mục tiêu**

Cung cấp cho người dùng các công cụ trực quan để hiểu rõ hơn về các yếu tố ảnh hưởng đến nguy cơ mắc bệnh đột quỵ và cách phòng ngừa bệnh này thông qua việc thể hiện dữ liệu trực quan, bao gồm biểu đồ, đồ thị và các công cụ phân tích dữ liệu khác.

Dữ liệu này cũng có thể giúp các nhà nghiên cứu và chuyên gia y tế tìm ra các mối liên hệ giữa các yếu tố nguy cơ khác nhau và đưa ra các khuyến nghị về phòng ngừa và điều trị bệnh đột quỵ.

## **3. Phương pháp nghiên cứu**

### **3.1. Về mặt lý thuyết**

* Tìm hiểu về cách thức xử lý dữ liệu và trực quan hóa dữ liệu.
* Tìm hiểu về các công cụ, ngôn ngữ hỗ trợ trong việc phân tích và trực quan
* Lý thuyết về lập trình python.
* Tham khảo tài liệu liên quan đến xử lý dữ liệu đồ thị Matplotlib, Seaborn,…

### **3.2. Đối tượng nghiên cứu**

Bao gồm ngôn ngữ python và các module hỗ trợ việc trực quan hóa dữ liệu như matplotlib và seaborn,...

Tập dữ liệu “Stroke Prediction Dataset” được lấy từ Kaggle.

### **3.3. Phạm vi nghiên cứu**

Phạm vi nghiên cứu bao quát trong môn Tương tác dữ liệu trực quan.

## **4. Kết quả dự kiến đạt được**

Dự đoán chính xác hơn nguy cơ mắc bệnh đột quỵ dựa trên các yếu tố nguy cơ, giúp đưa ra các quyết định chẩn đoán và điều trị đúng đắn.

Hiểu rõ hơn về các yếu tố ảnh hưởng đến bệnh đột quỵ và cách phòng ngừa bệnh này, giúp người dùng cải thiện chế độ ăn uống và lối sống để giảm nguy cơ mắc bệnh đột quỵ.

Tăng sự nhận thức của công chúng về bệnh đột quỵ và cách phòng ngừa bệnh này, giúp giảm tỷ lệ mắc bệnh và cải thiện sức khỏe cộng đồng.

Phát triển các công cụ trực quan và tiện ích hỗ trợ quyết định trong lĩnh vực y tế, giúp cho quy trình chẩn đoán và điều trị bệnh đột quỵ trở nên nhanh chóng và hiệu quả hơn.

# **B. PHẦN NỘI DUNG**

## **I.  Giới thiệu về tập dữ liệu**

### **1. Nguồn gốc**

Bộ dữ liệu Stroke Prediction được tạo ra bởi Feliciano Ordóñez và chia sẻ trên Kaggle, một trong những nền tảng trực tuyến phổ biến nhất để tìm kiếm và chia sẻ các bộ dữ liệu và mô hình học máy.

Bộ dữ liệu này được thu thập bởi Mỹ đại học California, San Francisco (UCSF), trong khuôn khổ một cuộc nghiên cứu về đột quỵ tại một trung tâm y tế ở Venezuela. Dữ liệu được thu thập từ 2015 đến 2016 và bao gồm thông tin về hơn 5000 bệnh nhân, bao gồm cả những người có đột quỵ và những người không có.

Dữ liệu bao gồm các biến liên quan đến các yếu tố nguy cơ đột quỵ như độ tuổi, giới tính, hút thuốc, tiểu đường, cao huyết áp, các bệnh lý tim mạch, hành động thể chất và cân nặng. Bộ dữ liệu cũng bao gồm thông tin về kết quả của các xét nghiệm y tế và các bước chẩn đoán được đưa ra cho các bệnh nhân.

Bộ dữ liệu này được chia sẻ công khai trên Kaggle để cộng đồng nghiên cứu và phát triển học máy có thể sử dụng nó để xây dựng các mô hình dự đoán đột quỵ và phát triển các công cụ hữu ích trong việc chẩn đoán và điều trị bệnh.

### **2. Mô tả tập dữ liệu**

Bộ dữ liệu Stroke Prediction được chia sẻ trên Kaggle bao gồm thông tin về hơn 5000 bệnh nhân, trong đó có hơn 200 bệnh nhân đã mắc đột quỵ và hơn 4800 bệnh nhân khác không mắc bệnh này. Dữ liệu được thu thập tại một trung tâm y tế ở Venezuela trong khoảng thời gian từ 2015 đến 2016.

Bộ dữ liệu này bao gồm các biến đầu vào như độ tuổi, giới tính, hút thuốc, tiểu đường, cao huyết áp, các bệnh lý tim mạch, hành động thể chất và cân nặng. Ngoài ra, dữ liệu còn bao gồm các kết quả của các xét nghiệm y tế và các bước chẩn đoán được đưa ra cho các bệnh nhân.

Cụ thể, bộ dữ liệu bao gồm các thuộc tính sau:

|  |  |
| --- | --- |
| Id | Số hiệu bệnh nhân |
| gender | Giới tính của bệnh nhân (Male/Female/Other) |
| Age | Độ tuổi của bệnh nhân |
| hypertension | Có bị cao huyết áp hay không (0: Không, 1: Có) |
| heart\_disease | Có bị bệnh lý tim mạch hay không (0: Không, 1: Có) |
| ever\_married | Đã kết hôn hay chưa (Yes/No) |
| work\_type | Loại công việc của bệnh nhân (Private/Self-employed/Govt\_job/Children/Never\_worked) |
| Residence\_type | Loại địa chỉ của bệnh nhân (Urban/Rural) |
| avg\_glucose\_level | Trung bình đường huyết của bệnh nhân |
| Bmi | Chỉ số khối cơ thể của bệnh nhân |
| smoking\_status | Tình trạng hút thuốc của bệnh nhân (formerly smoked/never smoked/smokes/Unknown) |
| Stroke | Có mắc đột quỵ hay không (0: Không, 1: Có) |

## **II. Giới thiệu về công nghệ sử dụng**

### **1. Python**

Python là một ngôn ngữ lập trình được sử dụng rộng rãi trong các ứng dụng web, phát triển phần mềm, khoa học dữ liệu và máy học (ML). Các nhà phát triển sử dụng Python vì nó hiệu quả, dễ học và có thể chạy trên nhiều nền tảng khác nhau. Phần mềm Python được tải xuống miễn phí, tích hợp tốt với tất cả các loại hệ thống và tăng tốc độ phát triển.

### **2. Streamlit**

Streamlit là một framework mã nguồn mở cho phép người dùng xây dựng các ứng dụng web trực quan và tương tác dữ liệu dễ dàng bằng Python. Streamlit cung cấp các công cụ và thư viện để tạo ra các ứng dụng web chất lượng cao mà không cần có kiến thức về phát triển web, giúp cho việc phát triển các ứng dụng web trở nên nhanh chóng và dễ dàng hơn.

### **3. Một số thư viện**

Pandas là một thư viện mã nguồn mở của Python, được sử dụng rộng rãi trong việc phân tích và xử lý dữ liệu.

NumPy là một thư viện mã nguồn mở của Python, được sử dụng rộng rãi trong việc tính toán khoa học và số học.

Matplotlib là một thư viện mã nguồn mở của Python, được sử dụng rộng rãi trong việc tạo và hiển thị các biểu đồ và đồ thị trực quan.

Plotly là một thư viện trực quan hóa dữ liệu được sử dụng rộng rãi trong ngành khoa học dữ liệu và phân tích dữ liệu.

Time là một trong những thư viện cơ bản trong Python, cung cấp các hàm xử lý thời gian và ngày tháng.

## **III.   Cách xử lý dữ liệu**

### **1. Đọc dữ liệu vào chương trình**

Sau khi tải tập dữ liệu từ Kaggle, nhóm sử dụng hàm read\_csv() từ thư viện pandas để đọc dự liệu.

A picture containing text, font, screenshot, graphics

Description automatically generated

### **2. Thực hiện yêu cầu tốc độ cập nhật dữ liệu của chương trình**

Vì nhóm sử dụng streamlit để tạo 1 dashboard trên trang web nên nhóm sử dụng vòng lặp vô hạn “while True:” để lặp liên tục.

A screen shot of a computer program

Description automatically generated with low confidence

Ở cuối vòng lặp nhóm để thời gian nghỉ là 60 và hàm reload lại trang web

A screenshot of a computer program

Description automatically generated with low confidence

Mỗi khi reload lại trang thì trang web sẽ đọc lại dữ liệu từ file .csv để có thể tự cập nhật lại dữ liệu mỗi 1 phút.

### **3. Xem thông tin từng cột**

Sử dụng hàm info để xem thông tin từng cột

A screenshot of a computer program

Description automatically generated with medium confidence

### **4. Lược bỏ các giá trị Other trong cột Gender**

A picture containing text, screenshot, font, line

Description automatically generated

Để có tập dữ liệu dễ dàng vẽ biểu đồ so sánh được tỉ lệ tương quan giữa nam và nữ

**5. Bổ sung các giá trị bị thiếu trong tập dữ liệu**

A picture containing text, screenshot, font

Description automatically generated

Trong cột bmi có những giá trị bị thiếu, nhóm đã bổ sung và thay thế những giá trị đó bằng giá trị mean của chúng.

**6. Từ tập dữ liệu tạo ra hai bảng dữ liệu đột quỵ và không đột quỵ**

A picture containing text, screenshot, font

Description automatically generated

Để có thể dễ dàng sử dụng để vẽ biểu đồ cũng như so sánh thì nhóm đã chia thành 2 tập dữ liệu, một là dữ liệu những người bị đột quỵ và hai là tập dữ liệu những người không bị đột quỵ

**7. Xử lý dữ liệu trong lúc vẽ đồ thị**

Ngoài việc xử lý ban đầu thì nhóm còn phải xử lý dữ liệu cho từng biểu đồ riêng biệt để có thể dễ dàng trực quan.

A screen shot of a computer code

Description automatically generated with low confidence

## **IV. Vẽ biểu đồ**

### **1. Biểu đồ tròn biểu thị biểu diễn theo phần trăm so sánh giữa những người đột quỵ và không đột quỵ theo các thuộc tính**

**1.1. Cách vẽ**

Nhóm sử dụng thư viện Matplotlib để vẽ biểu đồ tròn (Pie chart) cho các thuộc tính trong tập dữ liệu. Biểu đồ tròn hiển thị phần trăm giá trị của từng giá trị riêng lẻ trong thuộc tính được chọn bởi người dùng thông qua hộp chọn (selectbox).

Hàm piedist được định nghĩa để tính toán số lượng các giá trị riêng lẻ của thuộc tính và vẽ biểu đồ tròn. Nếu thuộc tính được chọn là "hypertension" hoặc "heart\_disease", hàm sẽ thay thế giá trị 0 và 1 bằng 'No' và 'Yes' tương ứng để có nhãn trục thích hợp. Hàm sử dụng labels để lưu trữ các giá trị riêng lẻ của thuộc tính và values để lưu trữ số lượng các giá trị tương ứng. Sau đó, hàm sẽ sử dụng plt.pie để vẽ biểu đồ tròn, autopct='%1.1f%%' để hiển thị phần trăm giá trị và plt.legend để hiển thị nhãn các giá trị trên biểu đồ.

A picture containing text, screenshot, font

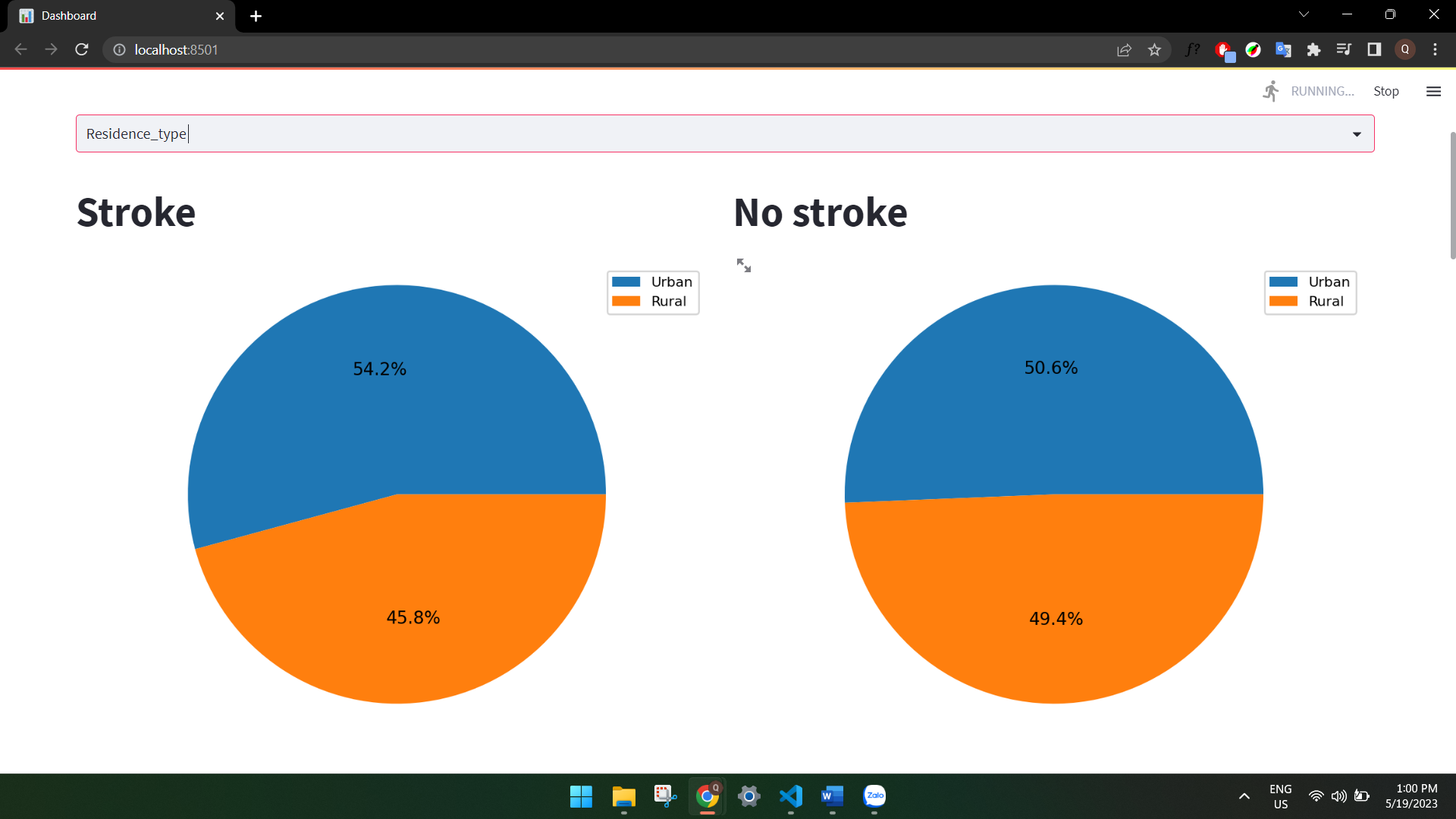
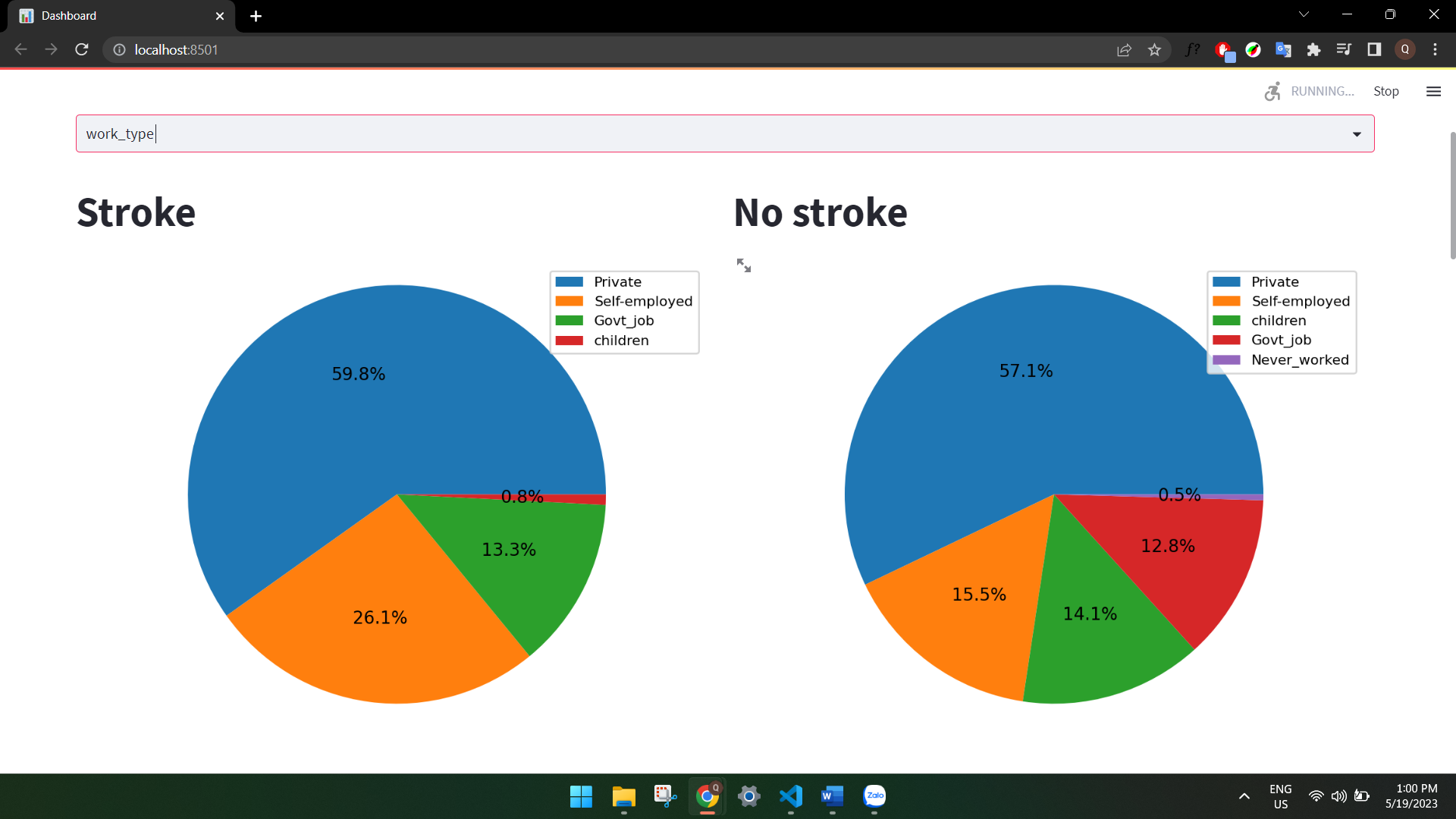
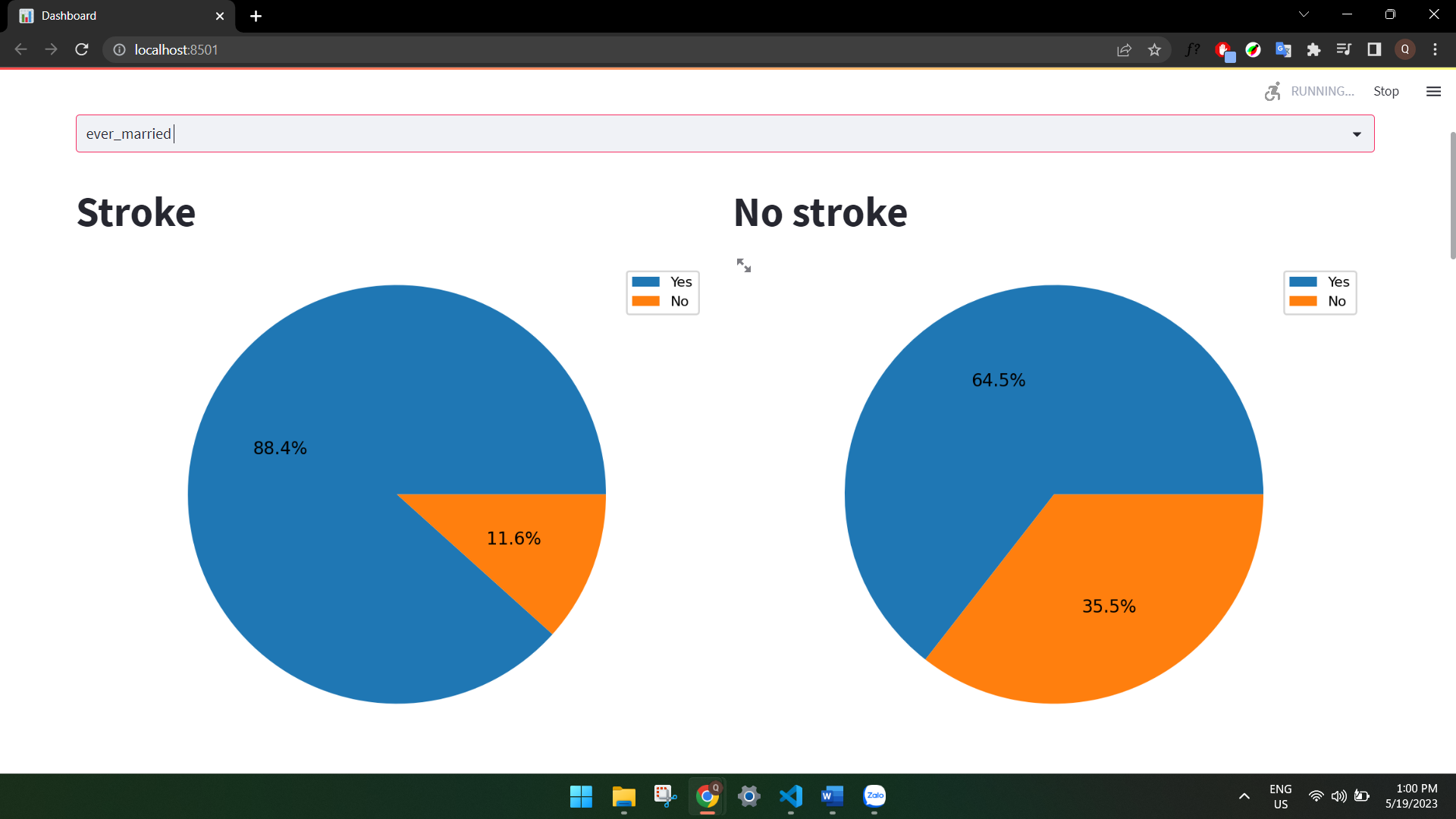
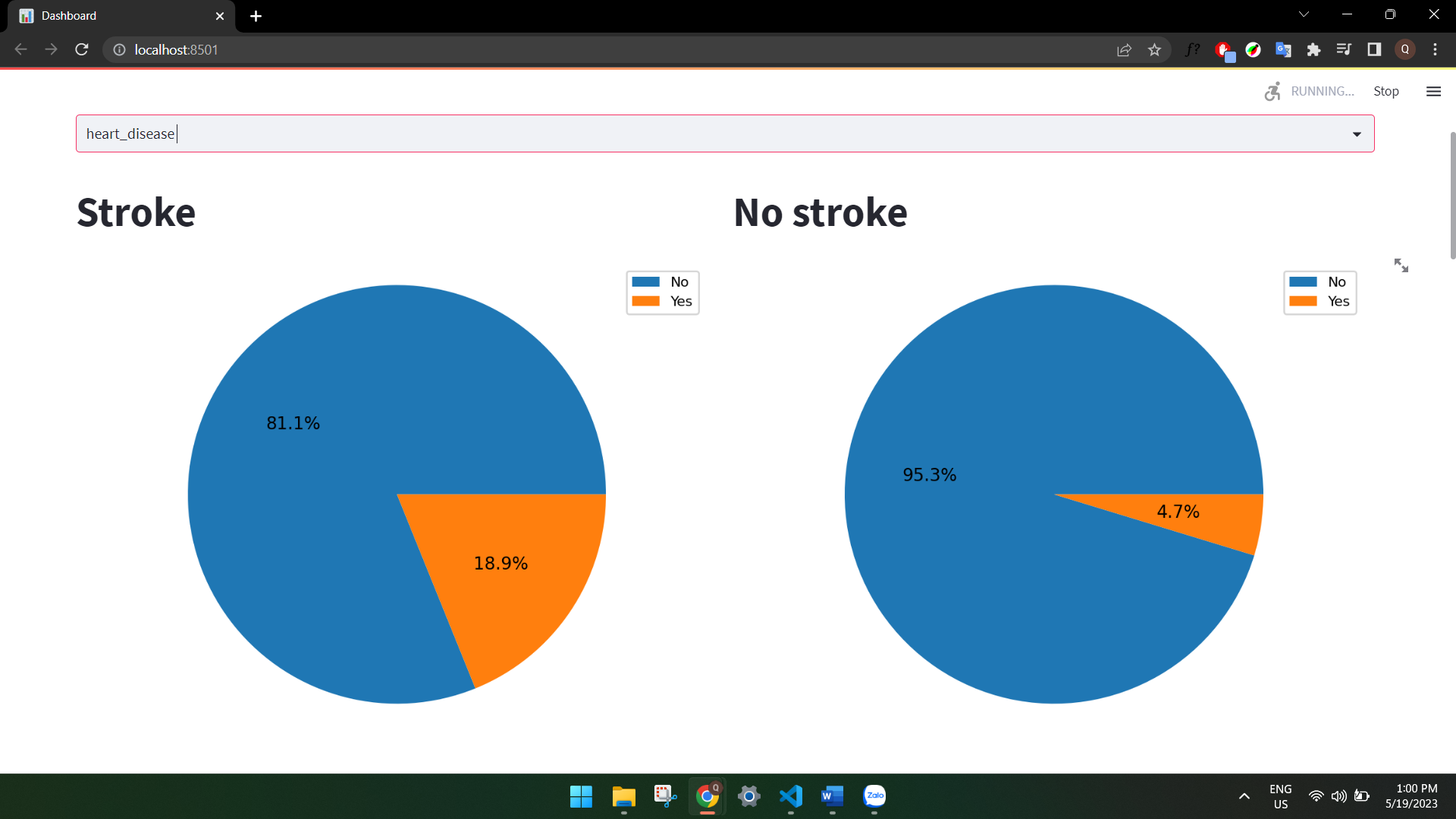
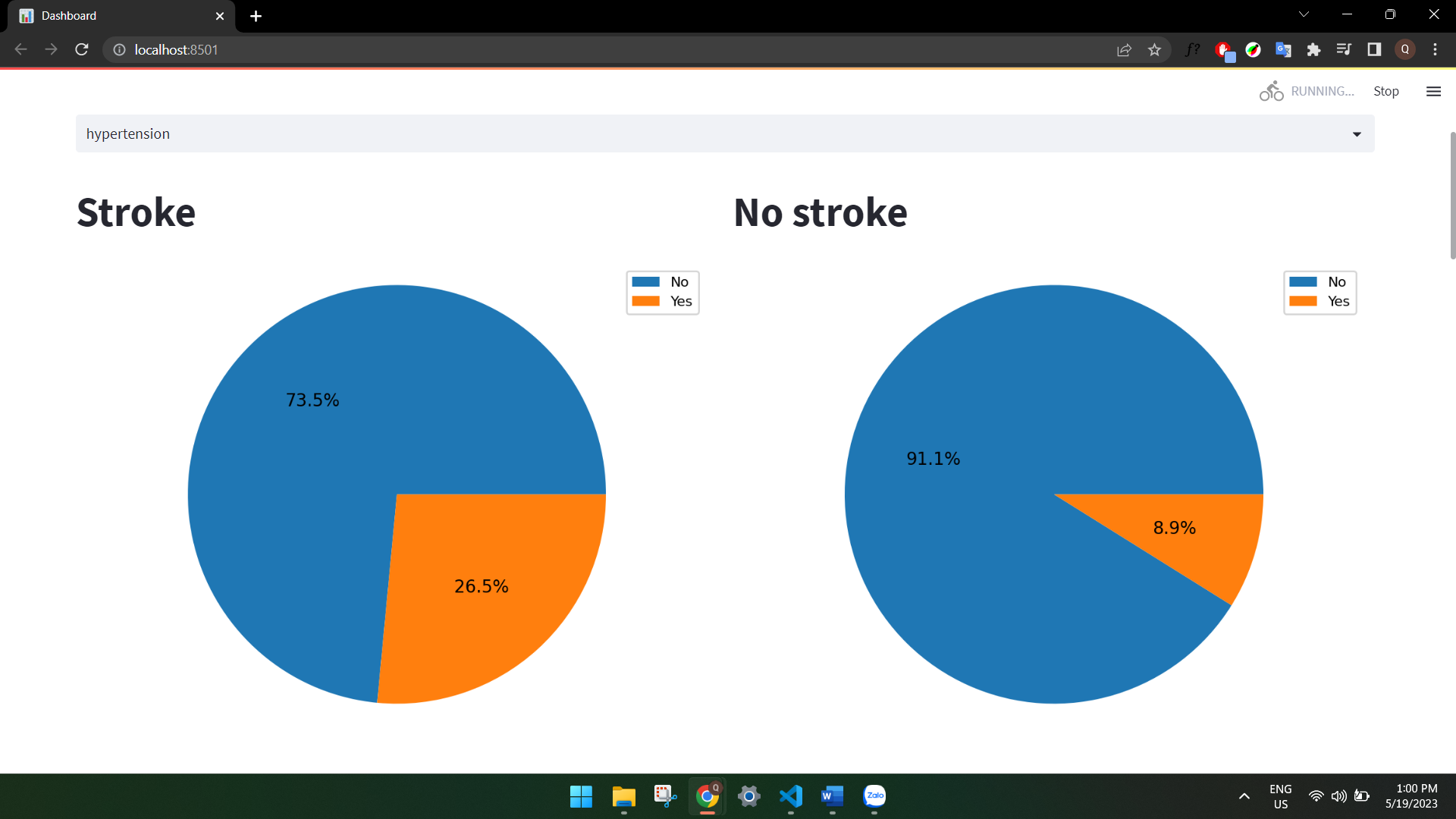
Description automatically generated

**1.2. Kết quả**

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated



### **2. Biểu đồ cột thể hiện số lượng người đột quỵ và không đột quỵ theo các thuộc tính**

**2.1. Cách vẽ**

Nhóm sử dụng thư viện Plotly để vẽ biểu đồ cột (bar chart) dựa trên dữ liệu được chọn từ dropdown menu. Trước khi vẽ, hàm replace được sử dụng để thay đổi giá trị 0 và 1 trong dữ liệu thành 'No' và 'Yes'. Biểu đồ cột được vẽ dựa trên dữ liệu sau khi thay đổi này. Biểu đồ cột này sẽ hiển thị số lượng người trong từng nhóm của thuộc tính được chọn, và nhóm này sẽ được phân biệt bằng màu sắc dựa trên có mắc đột quỵ hay không (stroke).

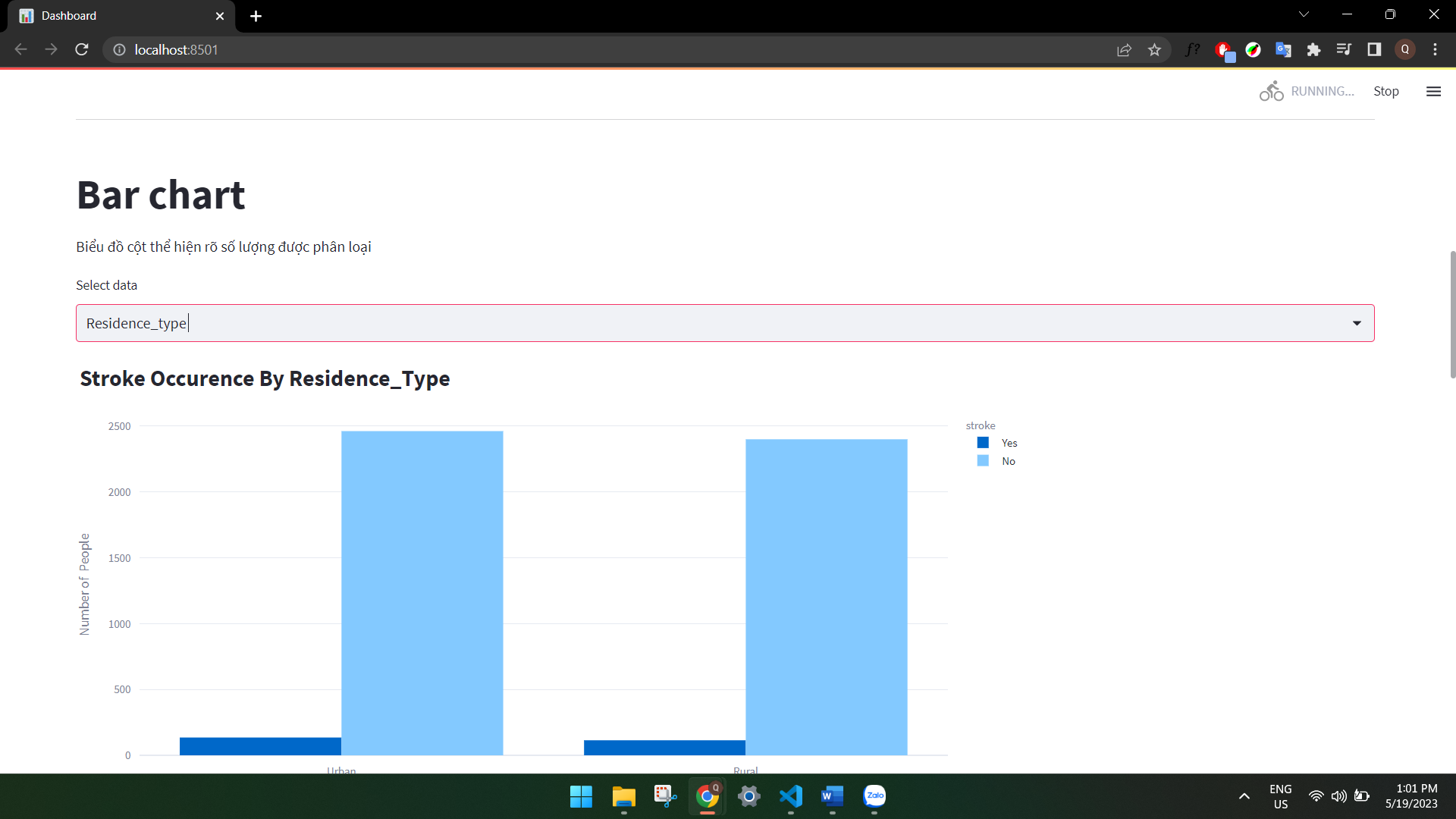
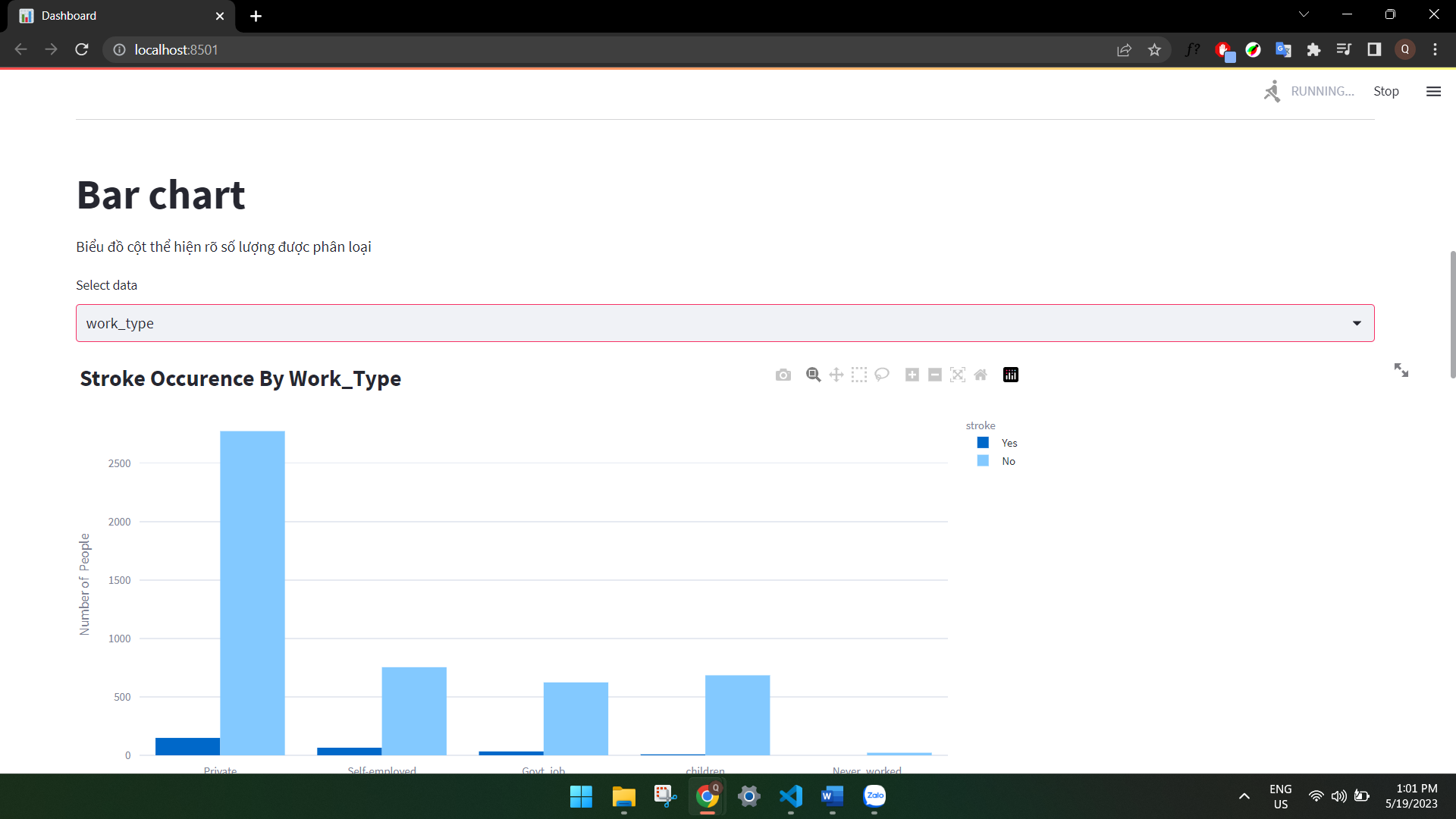
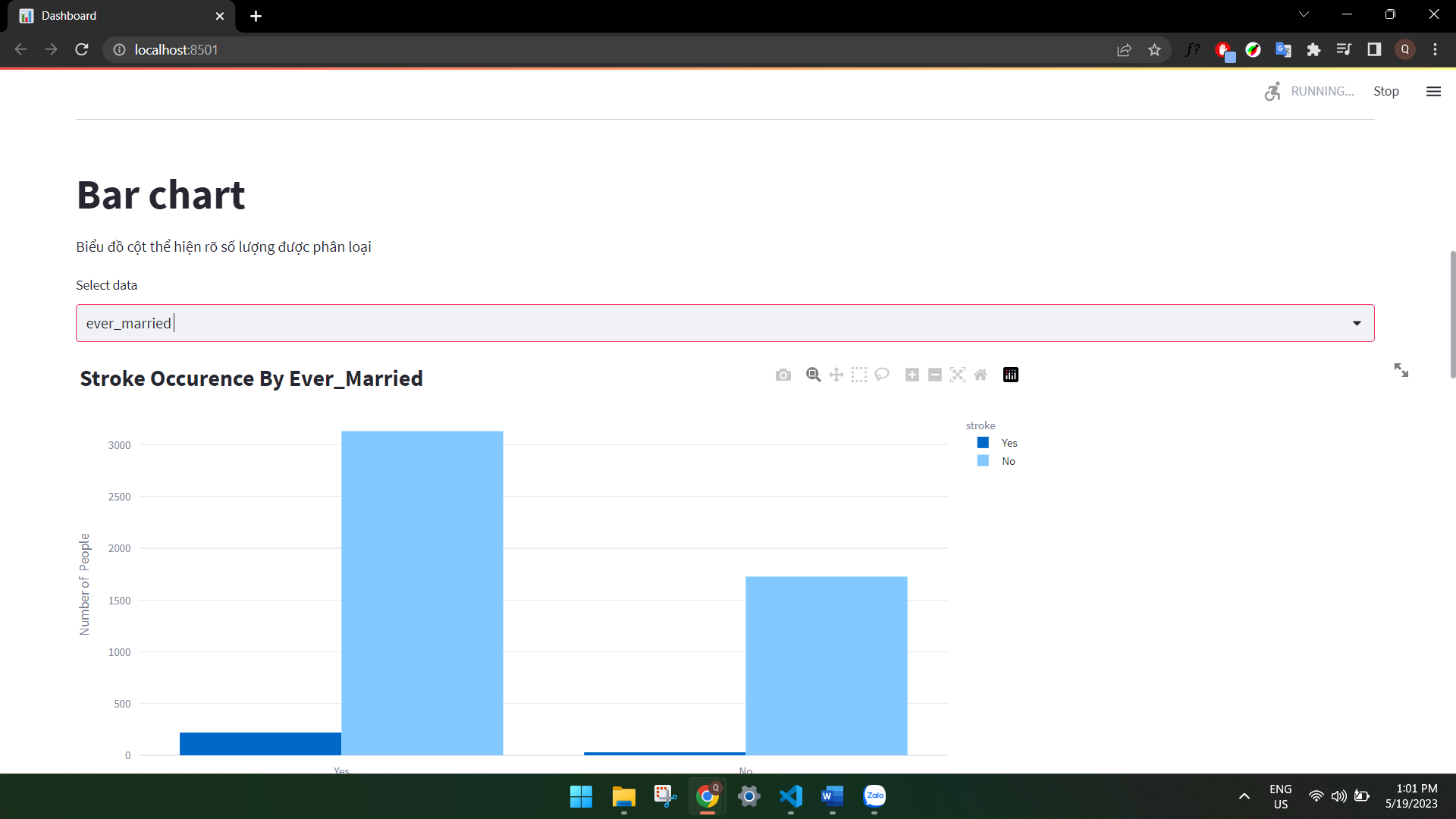
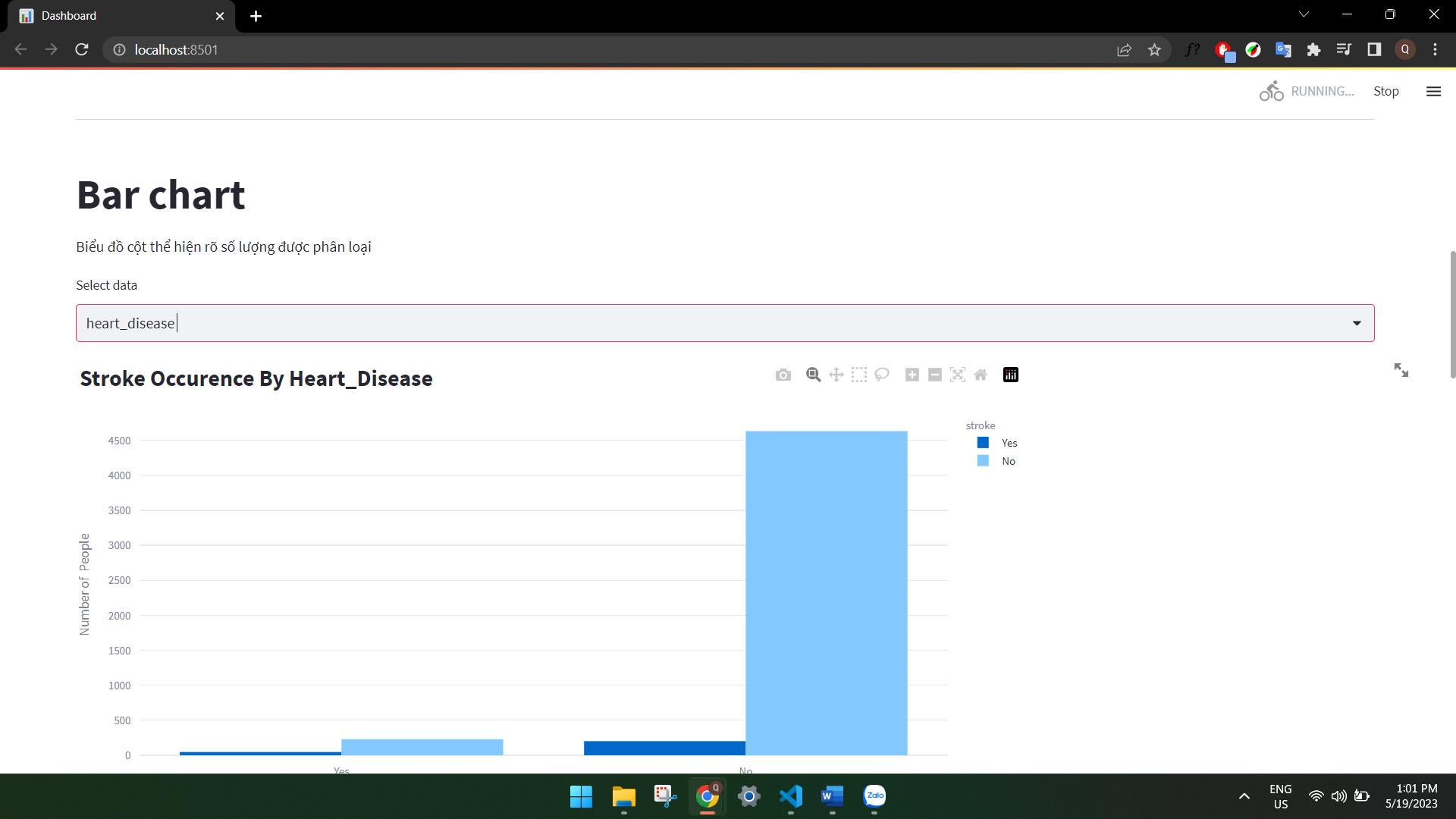
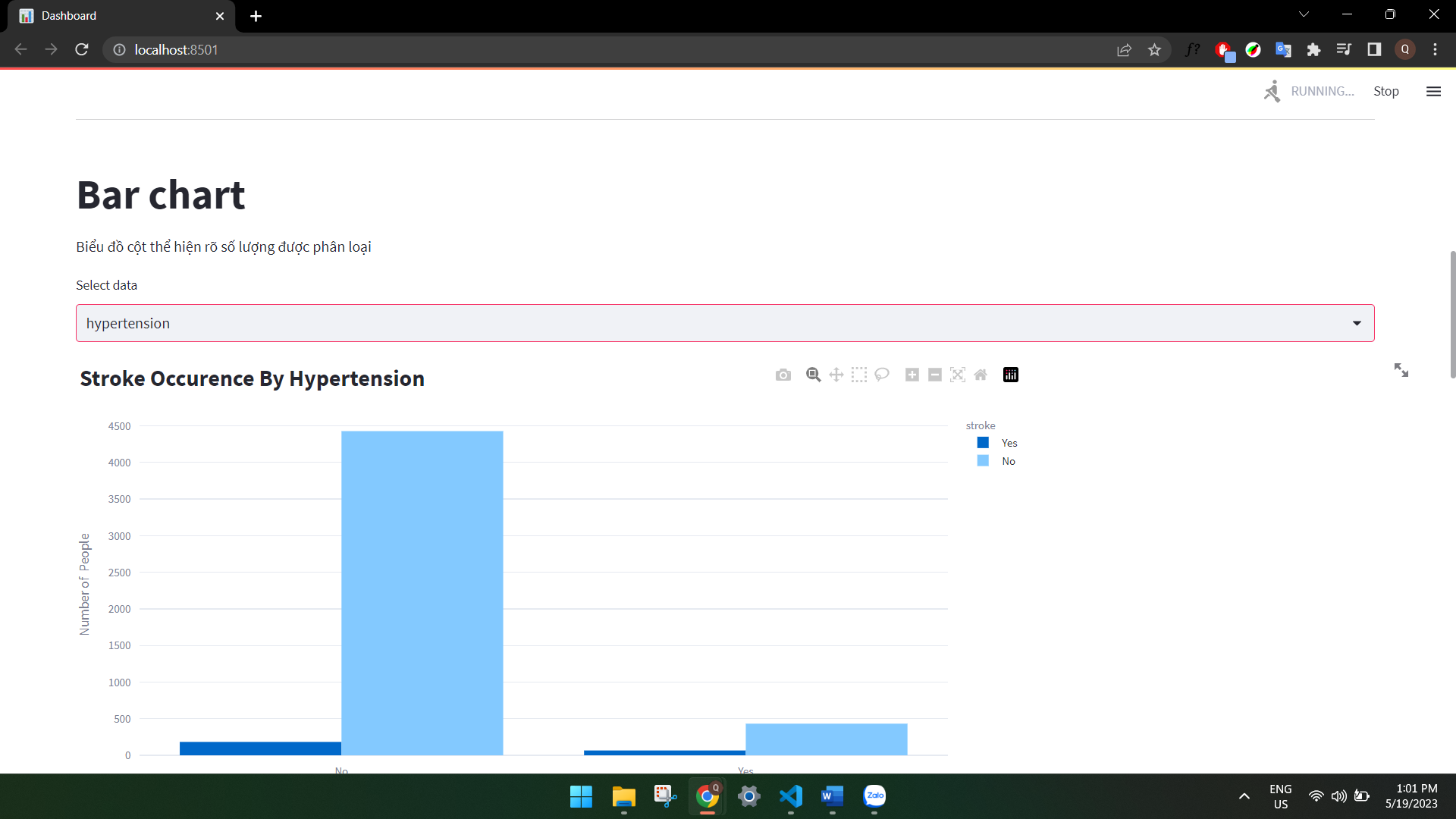
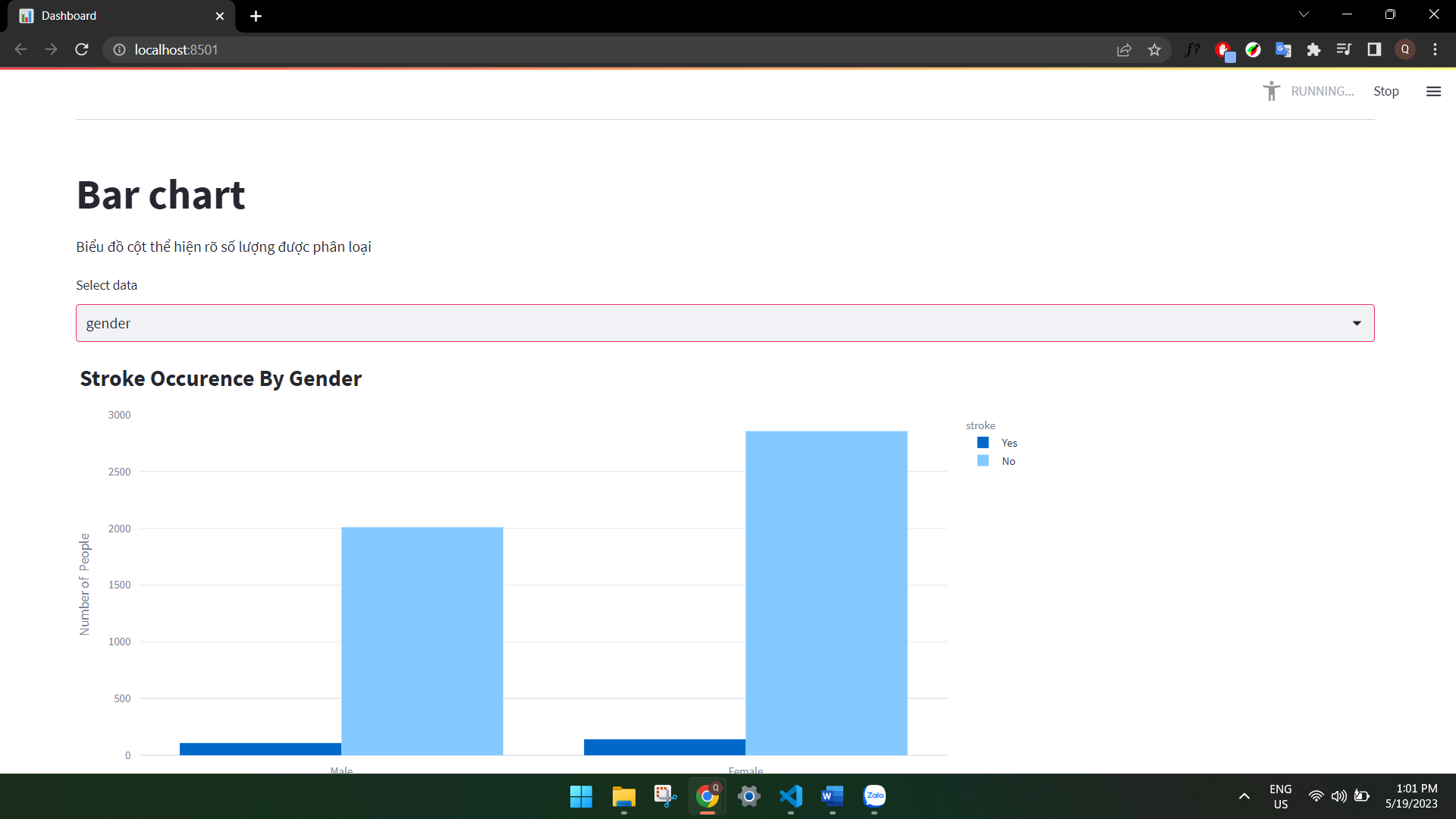
A picture containing screenshot, line, font

Description automatically generated

**2.2. Kết quả**

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated

### **3. Biểu đồ nhiệt thể hiện sự tương quan giữa các thuộc tính**

**3.1. Cách vẽ**

Đoạn code này sử dụng thư viện Plotly để vẽ biểu đồ heatmap (imager plot) dựa trên ma trận tương quan giữa các thuộc tính của tập dữ liệu. Trước khi vẽ, cột "id" được loại bỏ khỏi tập dữ liệu. Biểu đồ sử dụng hàm imshow để tạo ra heatmap với các giá trị tương quan được trình bày bằng màu sắc khác nhau. Giá trị tương quan càng lớn thì màu sắc sẽ càng đậm, và càng nhỏ thì màu sắc sẽ càng nhạt. Cuối cùng, kích thước của biểu đồ được cập nhật và được vẽ bằng hàm plotly\_chart của Streamlit để hiển thị trên giao diện web.

A picture containing text, screenshot, font

Description automatically generated

**3.2. Kết quả**A screenshot of a computer

Description automatically generated

### **4. Biểu đồ nến thể hiện sự phân bố của đột quỵ dựa trên giới tính và độ tuổi**

**4.1. Cách vẽ**

Mã này được sử dụng để vẽ biểu đồ boxplot với dữ liệu được lấy từ cột "gender" và "age" của tập dữ liệu. Biểu đồ được phân tách theo cột "stroke", với mỗi giá trị trong "stroke" được đại diện bằng màu sắc khác nhau. Cụ thể, giá trị "stroke" bằng 0 được đại diện bằng màu xanh lá cây và giá trị "stroke" bằng 1 được đại diện bằng màu đỏ.

Để làm điều này, chúng ta sử dụng plotly express, một thư viện dựa trên plotly, để tạo biểu đồ boxplot. Sau đó, chúng ta sử dụng phương thức "update\_layout" để cài đặt chiều rộng của biểu đồ là 1000. Cuối cùng, biểu đồ được hiển thị bằng cách sử dụng "st.plotly\_chart".

A picture containing text, font, screenshot

Description automatically generated

**4.2. Kết quả**

A screenshot of a computer

Description automatically generated

### **5. Biểu đồ điểm thể hiện sự đột quỵ giữa độ tuổi và lượng đường trong máu**

**5.1.Cách vẽ**

Nhóm sử dụng thư viện Plotly để tạo biểu đồ Scatter Plot với trục x là Age, trục y là Average Glucose Level, và facet\_col là giới tính. Tính toán và hiển thị trendline sử dụng phương pháp least squares (OLS). Cập nhật nhãn trục và giá trị giới tính để thể hiện đầy đủ ý nghĩa của dữ liệu. Sử dụng hàm plotly\_chart của Streamlit để hiển thị biểu đồ trên trang web Streamlit.

**5.2.Kết quả**

A picture containing text, line, diagram, font

Description automatically generated

# **C. PHẦN KẾT LUẬN**

## **1. Kết luận**

Qua quá trình xử lý và trực quan hóa dữ liệu thì nhóm đã vẽ được những biểu đồ khác nhau biểu thị nhiều mối liên quan giữa bệnh đột quỵ và giới tính, lượng đường trong máu cũng như độ tuổi. Từ đó dự đoán được những đặc điểm của người dễ dàng mắc đột quỵ để phòng tránh.

## **2. Kết quả đạt được**

Nắm được lý thuyết về cách sử dụng các công nghệ và ngôn ngữ như python, streamlit,...

Vẽ được sơ đồ dùng để phân tích và trực quan như biểu đồ cột, tròn,..., tạo được dashboard chứa các biểu đồ có thể tương tác

## **3. Hạn chế**

Hạn chế trong việc dữ liệu cung cấp, hạn chế về kinh nghiệm,..

Chưa thực sự đào sâu vào phân tích và áp dụng các thuật toán machine learning để dự đoán.

## **4. Tài liệu tham khảo**

1. Công nghệ vẽ dashboard streamlit: <https://docs.streamlit.io/>

2. Tập dữ liệu để phân tích: https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset