TRƯỜNG ĐẠI HỌC KỸ THUẬT CÔNG NGHIỆP

**KHOA ĐIỆN TỬ**

**Bộ môn : Công nghệ thông tin**

**BÀI TẬP LỚN**

MÔN HỌC

**KHOA HỌC DỮ LIỆU**

**Sinh viên :** Xiêm Văn Hoàn

**Lớp :** 55KMT

**Giáo viên hướng dẫn :** Nguyễn Văn Huy

**Thái Nguyên – 2023**

|  |  |
| --- | --- |
| **TRƯỜNG ĐHKTCN** | **CỘNG HOÀ XÃ HỘI CHỦ NGHĨA VIỆT NAM** |
| **KHOA ĐIỆN TỬ** | ***Độc lập - Tự do - Hạnh phúc*** |

**BÀI TẬP LỚN**

**MÔN HỌC:** KHOA HỌC DỮ LIỆU

BỘ MÔN : CÔNG NGHỆ THÔNG TIN

*Sinh viên: Xiêm Văn Hoàn*  *Lớp*: *55KMT*

*Ngành: Kỹ thuật máy tính*

*Ngày giao đề: 30/01/2023 Ngày hoàn thành: 21/02/2023*

1.Tên đề tài : Dự đoán nguy cơ mắc bệnh Ung Thư Phổi

2. Yêu cầu

|  |  |
| --- | --- |
| TRƯỞNG BỘ MÔN | GIÁO VIÊN HƯỚNG DẪN |
| *(Ký và ghi rõ họ tên)* | *(Ký và ghi rõ họ tên)* |

**NHẬN XÉT CỦA GIÁO VIÊN HƯỚNG DẪN**

### Thái Nguyên, ngày….tháng…..năm 20....

## GIÁO VIÊN HƯỚNG DẪN

*(Ký ghi rõ họ tên)*

**NHẬN XÉT CỦA GIÁO VIÊN CHẤM**

### Thái Nguyên, ngày….tháng…..năm 20....

## GIÁO VIÊN CHẤM

*(Ký ghi rõ họ tên)*

MỤC LỤC

[TRƯỜNG ĐẠI HỌC KỸ THUẬT CÔNG NGHIỆP 1](#_Toc136405418)

[**KHOA ĐIỆN TỬ** 1](#_Toc136405419)

[Chương I : Khảo Sát Hệ Thống 6](#_Toc136405426)

[1.1. Đặt vấn đề 6](#_Toc136405427)

[1.2. Mục tiêu 7](#_Toc136405428)

[1.3. Ứng dụng 8](#_Toc136405429)

[1.4. Công nghệ hiện nay 9](#_Toc136405430)

[1.5. Phương pháp thực hiện 11](#_Toc136405431)

[Chương II : Cơ Sở Lý Thuyết 12](#_Toc136405432)

[2.1. Xử lí dữ liệu 12](#_Toc136405433)

[2.2. Model 14](#_Toc136405434)

[Chương III : Xây Dựng Chương Trình 17](#_Toc136405435)

[3.1. Xử lý code / data 17](#_Toc136405436)

[3.2. Huấn luyện model 19](#_Toc136405437)

[3.3. Demo 20](#_Toc136405438)

[Chương 4 : Kết Luận 21](#_Toc136405439)

[4.1. Chương trình đã làm được gì ? 21](#_Toc136405440)

[4.2. Ưu điểm và nhược điểm 23](#_Toc136405441)

[4.3. Phát triển ứng dụng 24](#_Toc136405442)

[TÀI LIỆU THAM KHẢO 25](#_Toc136405443)

**Lời mở đầu**

Ung thư phổi là một trong những loại ung thư gây tử vong hàng đầu trên toàn cầu. Với sự gia tăng đáng kể về số ca mắc mới và tình trạng diễn biến không lường trước, nghiên cứu về nguy cơ mắc bệnh ung thư phổi đóng vai trò quan trọng trong việc hiểu và giảm thiểu tác động của căn bệnh này đối với sức khỏe cộng đồng.

Tiểu luận này nhằm mục đích tìm hiểu và phân tích các yếu tố nguy cơ gây mắc bệnh ung thư phổi. Tôi sẽ đi sâu vào nghiên cứu các yếu tố môi trường, hành vi sinh hoạt và di truyền có thể góp phần vào sự phát triển và phổ biến căn bệnh này. Ngoài ra, tôi cũng sẽ tìm hiểu về những biện pháp phòng ngừa và kiểm soát để giảm thiểu rủi ro mắc bệnh ung thư phổi trong cộng đồng.

Hy vọng rằng thông qua tiểu luận này, chúng ta có thể nâng cao nhận thức và sự hiểu biết về nguy cơ mắc bệnh ung thư phổi, từ đó đưa ra các giải pháp và chính sách hiệu quả để bảo vệ sức khỏe của mọi người. Chúng tôi hy vọng rằng nội dung của tiểu luận sẽ cung cấp cho bạn cái nhìn tổng quan và cung cấp cơ sở cho các nghiên cứu tiếp theo về nguy cơ mắc bệnh ung thư phổi.

Tôi xin chân thành cảm ơn sự quan tâm và đóng góp của tất cả những người đã hỗ trợ chúng tôi trong quá trình thực hiện tiểu luận này. Xin cảm ơn thầy Nguyễn Văn Huy và các bạn k55kmt đã hỗ trợ trong quá trình làm bài .

Sinh viên

Xiêm Văn Hoàn

Chương I : Khảo Sát Hệ Thống

* 1. Đặt vấn đề

Vấn đề về bệnh ung thư phổi là một chủ đề quan trọng và có tác động đáng kể đến sức khỏe cộng đồng. Dưới đây là một số ý đề bạn có thể đặt ra khi nghiên cứu về bệnh ung thư phổi:

Tăng cường nhận thức về bệnh ung thư phổi: Nghiên cứu về mức độ nhận thức của công chúng về bệnh ung thư phổi, những khái niệm sai lầm phổ biến, và cách tăng cường thông tin và giáo dục để tăng sự nhận thức về căn bệnh này.

Yếu tố nguy cơ: Tìm hiểu các yếu tố nguy cơ gây mắc bệnh ung thư phổi, bao gồm hút thuốc lá, tiếp xúc với chất gây ô nhiễm không khí, di truyền, lối sống và hành vi sinh hoạt không lành mạnh.

Điều tra về chẩn đoán sớm và phát hiện: Nghiên cứu về phương pháp chẩn đoán sớm và phát hiện bệnh ung thư phổi, như kiểm tra sàng lọc, chẩn đoán hình ảnh, xét nghiệm máu và phương pháp mới tiềm năng để tìm ra bệnh ung thư phổi trong giai đoạn đầu.

Điều trị và quản lý bệnh ung thư phổi: Nghiên cứu về các phương pháp điều trị hiện có, bao gồm phẫu thuật, hóa trị, và xạ trị, cũng như phát triển các phương pháp mới, như liệu pháp di truyền và immunotherapy, để cải thiện kết quả điều trị và chất lượng cuộc sống cho bệnh nhân ung thư phổi.

Tác động tâm lý và xã hội: Nghiên cứu về tác động tâm lý và xã hội của bệnh ung thư phổi đối với bệnh nhân và gia đình, bao gồm cả tác động tâm lý sau chẩn đoán, tình trạng tâm lý và chất lượng cuộc sống, và hỗ trợ cộng đồng và các dịch vụ hỗ trợ cho bệnh nhân ung thư phổi.

Phòng ngừa và giảm nguy cơ: Nghiên cứu về các biện pháp phòng ngừa và giảm nguy cơ mắc bệnh ung thư phổi, bao gồm chính sách cấm hút thuốc lá, giảm ô nhiễm không khí, tư vấn về lối sống lành mạnh, và tăng cường giáo dục về yếu tố nguy cơ và phòng ngừa.

Đây chỉ là một số ý tưởng về vấn đề bệnh ung thư phổi mà bạn có thể sử dụng trong đề tài của mình. Tùy thuộc vào phạm vi và mục tiêu nghiên cứu của bạn, bạn có thể tập trung vào một hoặc nhiều khía cạnh cụ thể để khám phá sâu hơn.

Để giải quyết vấn đề này, một giải pháp là xây dựng một hệ thống dự đoán nguy cơ mắc bệnh ung thư phổi từ dữ liệu cho trước. Hệ thống này sẽ sử dụng các thuật toán máy học để phân tích dữ liệu và dự đoán nguy cơ mắc bệnh ung thư phổi dựa trên những yếu tố ảnh hưởng như tuổi , giới tính , rủi ro di chuyền, mức hút thuốc lá thụ động …

* 1. Mục tiêu

Tạo ra hệ thống có khả năng dự đoán nguy cơ mắc bệnh từ những dữ liệu .

Mục tiêu nghiên cứu về bệnh ung thư phổi có thể được đặt ra như sau:

Nâng cao nhận thức và kiến thức: Mục tiêu này tập trung vào tăng cường nhận thức và kiến thức về bệnh ung thư phổi trong cộng đồng. Nghiên cứu có thể tìm hiểu về mức độ nhận thức hiện tại, nhận biết các khái niệm sai lầm, và phát triển các chiến lược giáo dục và thông tin để nâng cao nhận thức và kiến thức về bệnh ung thư phổi.

Phân tích yếu tố nguy cơ: Mục tiêu này nhằm nghiên cứu và phân tích các yếu tố nguy cơ gây mắc bệnh ung thư phổi, bao gồm hút thuốc lá, ô nhiễm không khí, di truyền, lối sống không lành mạnh và các yếu tố khác. Mục tiêu là hiểu rõ hơn về tác động của những yếu tố này và đánh giá nguy cơ mắc bệnh ung thư phổi của các cá nhân và cộng đồng.

Cải thiện chẩn đoán sớm và phát hiện: Mục tiêu này tập trung vào nghiên cứu và phát triển các phương pháp chẩn đoán sớm và phát hiện bệnh ung thư phổi. Điều này có thể bao gồm nghiên cứu về kiểm tra sàng lọc, kỹ thuật hình ảnh, xét nghiệm máu và sự phát triển của các phương pháp mới để phát hiện bệnh ung thư phổi ở giai đoạn đầu.Nghiên cứu và cải thiện điều trị: Mục tiêu này nhằm nghiên cứu các phương pháp điều trị hiện có và phát triển các phương pháp mới để cải thiện điều trị bệnh ung thư phổi. Điều này có thể bao gồm nghiên cứu về phẫu thuật, hóa trị, xạ trị, liệu pháp di truyền và immunotherapy để tìm ra những phương pháp hiệu quả hơn trong việc kiểm soát và điều trị bệnh ung thư phổi.

Tăng cường hỗ trợ tâm lý và xã hội: Mục tiêu này tập trung vào nghiên cứu và cải thiện các dịch vụ hỗ trợ tâm lý và xã hội cho bệnh nhân ung thư phổi và gia đình. Điều này có thể bao gồm nghiên cứu về tác động tâm lý sau chẩn đoán, tình trạng tâm lý và chất lượng cuộc sống của bệnh nhân, và phát triển các dịch vụ hỗ trợ và chăm sóc tăng cường cho những người bị ảnh hưởng bởi bệnh ung thư phổi.

Mỗi mục tiêu trên đều có vai trò quan trọng trong việc nghiên cứu và giải quyết vấn đề bệnh ung thư phổi. Tuy nhiên, do thời gian có hạn và trình độ chưa được cao nên chưa thể khai thác hết về nguy cơ mắc bệnh ung thư phổi.

* 1. Ứng dụng

Có nhiều ứng dụng tiềm năng trong việc áp dụng các kết quả nghiên cứu về bệnh ung thư phổi. Dưới đây là một số ứng dụng có thể được xem xét:

Giáo dục và tư vấn: Các kết quả nghiên cứu về nguy cơ, phòng ngừa và chẩn đoán bệnh ung thư phổi có thể được ứng dụng trong việc giáo dục và tư vấn cộng đồng. Các chương trình giáo dục và chiến dịch tư vấn có thể giúp tăng cường nhận thức và kiến thức về bệnh ung thư phổi, khuyến khích lối sống lành mạnh và cung cấp thông tin về kiểm tra sàng lọc và chẩn đoán sớm.

Chẩn đoán và điều trị: Các kết quả nghiên cứu về phương pháp chẩn đoán sớm và điều trị bệnh ung thư phổi có thể được áp dụng trong thực tế y tế. Các công nghệ và phương pháp mới có thể cung cấp cho các chuyên gia y tế công cụ và kiến thức để phát hiện sớm và điều trị bệnh ung thư phổi một cách hiệu quả hơn.

Phát triển chính sách và quy định: Các kết quả nghiên cứu về yếu tố nguy cơ và tác động của bệnh ung thư phổi có thể hỗ trợ trong việc phát triển chính sách và quy định liên quan đến kiểm soát và phòng ngừa bệnh ung thư phổi. Điều này có thể bao gồm việc áp dụng các biện pháp giảm hút thuốc lá, cải thiện chất lượng không khí và thúc đẩy lối sống lành mạnh trong cộng đồng.

Nâng cao hỗ trợ tâm lý và xã hội: Các kết quả nghiên cứu về tác động tâm lý và xã hội của bệnh ung thư phổi có thể được áp dụng trong việc cung cấp dịch vụ hỗ trợ tâm lý và xã hội tốt hơn cho bệnh nhân và gia đình. Điều này có thể bao gồm việc phát triển các chương trình hỗ trợ tâm lý, nhóm hỗ trợ và các dịch vụ hỗ trợ xã hội.

* 1. Công nghệ hiện nay

Công nghệ hiện nay đóng vai trò quan trọng trong nghiên cứu và ứng dụng liên quan đến bệnh ung thư phổi. Dưới đây là một số công nghệ được áp dụng trong lĩnh vực này:

Hình ảnh y tế tiên tiến: Các công nghệ hình ảnh y tế như máy CT (Computed Tomography), MRI (Magnetic Resonance Imaging) và PET (Positron Emission Tomography) được sử dụng để xác định và đánh giá tổn thương phổi và khối u. Các kỹ thuật này cung cấp hình ảnh chi tiết về vị trí, kích thước và tính chất của khối u, giúp trong quá trình chẩn đoán và lựa chọn phương pháp điều trị.

Di truyền học và genomics: Công nghệ di truyền học và genomics đang được áp dụng để nghiên cứu sự biến đổi gen và tìm hiểu về cơ chế di truyền của ung thư phổi. Việc hiểu rõ hơn về các biến thể gen liên quan đến bệnh ung thư phổi có thể giúp xác định nhóm rủi ro cao và phát triển phác đồ điều trị cá nhân hóa.

Kỹ thuật tế bào học và tế bào đột biến: Các công nghệ tế bào học hiện đại như phân tích dòng tế bào (flow cytometry), phân loại tế bào đột biến (cytogenetics) và phân tích tế bào đơn lẻ (single-cell analysis) được sử dụng để nghiên cứu sự phát triển, biến đổi và kháng thuốc của tế bào ung thư phổi.

Kỹ thuật tiếp cận chích thuốc và chẩn đoán không xâm lấn: Các kỹ thuật tiếp cận chích thuốc như biopsi thông qua máy nội soi (endobronchial ultrasound-guided biopsy) và chẩn đoán không xâm lấn như xét nghiệm máu chứa tế bào ung thư (liquid biopsy) đang được phát triển để thu thập mẫu và xác định tế bào ung thư phổi một cách không gây đau và tác động ít đến cơ thể.

Hệ thống thông tin y tế và trí tuệ nhân tạo: Công nghệ thông tin y tế (Health Information Technology) và trí tuệ nhân tạo (Artificial Intelligence) được áp dụng để phân tích dữ liệu y tế lớn, hỗ trợ trong việc chẩn đoán, đánh giá và dự đoán kết quả điều trị của bệnh ung thư phổi.

Kỹ thuật phẫu thuật tiên tiến: Trong lĩnh vực phẫu thuật, các công nghệ tiên tiến như robot hỗ trợ phẫu thuật (robotic-assisted surgery), hệ thống hình ảnh bên trong cơ thể (intraoperative imaging) và máy hút khói phẫu thuật (surgical smoke evacuation) đang được sử dụng để tăng tính chính xác và an toàn trong quá trình loại bỏ khối u và điều trị ung thư phổi.

Các công nghệ trên đây đang đóng góp quan trọng trong việc nghiên cứu, chẩn đoán, điều trị và quản lý bệnh ung thư phổi, mang lại hi vọng trong việc cải thiện chẩn đoán sớm, tăng cường hiệu quả điều trị và nâng cao chất lượng cuộc sống của bệnh nhân.

* 1. Phương pháp thực hiện

Phương pháp nghiên cứu trong đề tài này là sử dụng mô hình học máy Linear Regression để dự đoán nguy cơ mắc bệnh ung thư phổi dựa trên các thông tin về tuổi, giới tính , mức độ sử dụng rượu , hút thuốc lá thụ động, rủi do di truyền…. Dữ liệu được thu thập từ một số trang web bệnh ung thư phổi , được tiền xử lý và lưu trữ trong một tập tin CSV. Sau đó, các bước tiền xử lý dữ liệu được thực hiện để chuẩn bị dữ liệu cho việc huấn luyện mô hình. Mô hình Linear Regression được huấn luyện trên tập dữ liệu đã được xử lý và sau đó được lưu trữ thành một tệp pickle để sử dụng trong ứng dụng web Flask. Các yêu cầu của người dùng sẽ được gửi đến ứng dụng web, mô hình sẽ được sử dụng để dự đoán giá trị của chiếc xe dựa trên các thông tin được cung cấp và kết quả sẽ được trả về cho người dùng.

Chương II : Cơ Sở Lý Thuyết

2.1. Xử lí dữ liệu

* Thuộc tính:

"Age" (Tuổi): Đây là thuộc tính đại diện cho tuổi của bệnh nhân. Kiểu dữ liệu: Số.

"Gender" (Giới tính): Đây là thuộc tính đại diện cho giới tính của bệnh nhân. Kiểu dữ liệu: Danh mục (có thể là "Nam" hoặc "Nữ").

"Air Pollution" (Ô nhiễm không khí): Đây là thuộc tính đại diện cho mức độ tiếp xúc với ô nhiễm không khí của bệnh nhân. Kiểu dữ liệu: Danh mục.

"Alcohol use" (Sử dụng rượu): Đây là thuộc tính đại diện cho mức độ sử dụng rượu của bệnh nhân. Kiểu dữ liệu: Danh mục.

"Dust Allergy" (Dị ứng bụi): Đây là thuộc tính đại diện cho mức độ dị ứng bụi của bệnh nhân. Kiểu dữ liệu: Danh mục.

"Occupational Hazards" (Nguy cơ nghề nghiệp): Đây là thuộc tính đại diện cho mức độ nguy hiểm nghề nghiệp của bệnh nhân. Kiểu dữ liệu: Danh mục.

"Genetic Risk" (Rủi ro di truyền): Đây là thuộc tính đại diện cho mức độ rủi ro di truyền của bệnh nhân. Kiểu dữ liệu: Danh mục.

"Chronic Lung Disease" (Bệnh phổi mãn tính): Đây là thuộc tính đại diện cho mức độ bệnh phổi mãn tính của bệnh nhân. Kiểu dữ liệu: Danh mục.

"Balanced Diet" (Chế độ ăn uống cân bằng): Đây là thuộc tính đại diện cho mức độ cân bằng chế độ ăn uống của bệnh nhân. Kiểu dữ liệu: Danh mục.

"Obesity" (Béo phì): Đây là thuộc tính đại diện cho mức độ béo phì của bệnh nhân. Kiểu dữ liệu: Danh mục.

"Smoking" (Hút thuốc): Đây là thuộc tính đại diện cho mức độ hút thuốc của bệnh nhân. Kiểu dữ liệu: Danh mục.

"Passive Smoker" (Hút thuốc lá thụ động): Đây là thuộc tính đại diện cho mức độ hút thuốc lá thụ động của bệnh nhân. Kiểu dữ liệu: Danh mục.

"Chest Pain" (Đau ngực): Đây là thuộc tính đại diện cho mức độ đau ngực của bệnh nhân. Kiểu dữ liệu: Danh mục.

"Coughing of Blood" (Ho có máu): Đây là thuộc tính đại diện cho mức độ ho có máu của bệnh nhân. Kiểu dữ liệu: Danh mục.

"Fatigue" (Mệt mỏi): Đây là thuộc tính đại diện cho mức độ mệt mỏi của bệnh nhân. Kiểu dữ liệu: Danh mục.

"Weight Loss" (Giảm cân): Đây là thuộc tính đại diện cho mức độ giảm cân của bệnh nhân. Kiểu dữ liệu: Danh mục.

"Shortness of Breath" (Khó thở): Đây là thuộc tính đại diện cho mức độ khó thở của bệnh nhân. Kiểu dữ liệu: Danh mục.

"Wheezing" (Ngưng thở): Đây là thuộc tính đại diện cho mức độ ngưng thở của bệnh nhân. Kiểu dữ liệu: Danh mục.

"Swallowing Difficulty" (Khó nuốt): Đây là thuộc tính đại diện cho mức độ khó nuốt của bệnh nhân. Kiểu dữ liệu: Danh mục.

"Clubbing of Finger Nails" (Móng tay tròn): Đây là thuộc tính đại diện cho mức độ móng tay tròn của bệnh nhân. Kiểu dữ liệu: Danh mục.

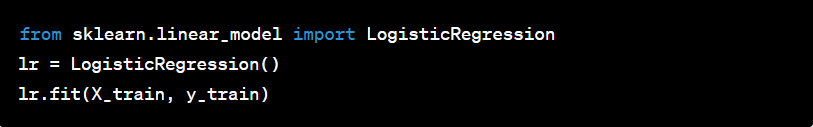
"Frequent Cold" (Cảm cúm thường xuyên): Đây là thuộc tính đại diện cho mức độ mắc cảm cúm thường xuyên của bệnh nhân. Kiểu dữ liệu: Danh mục.

"Dry Cough" (Ho khô): Đây là thuộc tính đại diện cho mức độ ho khô của bệnh nhân. Kiểu dữ liệu: Danh mục.

"Snoring" (Ngáy): Đây là thuộc tính đại diện cho mức độ ngáy của bệnh nhân. Kiểu dữ liệu: Danh mục.

2.2. Model

Bạn có thể thấy dòng mã sau:



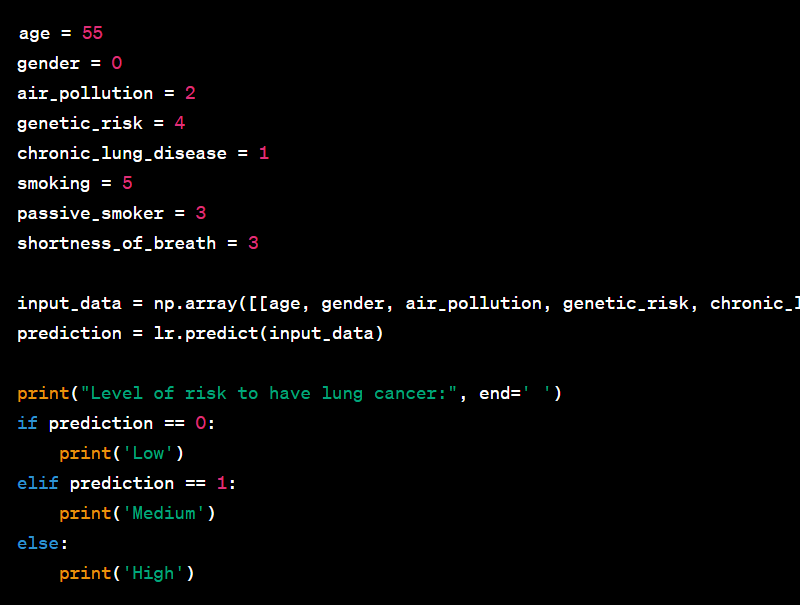
Trong đoạn mã trên, mô hình lr được khởi tạo bằng cách gọi LogisticRegression(), sau đó được huấn luyện bằng cách gọi fit(X\_train, y\_train). Điều này có nghĩa là mô hình đã được huấn luyện trên dữ liệu X\_train và y\_train.

Sau khi mô hình đã được huấn luyện, chúng ta có thể sử dụng nó để dự đoán mức độ nguy cơ bằng cách gọi lr.predict(input\_data). Đoạn mã dưới đây thực hiện dự đoán trên các dữ liệu đầu vào cụ thể:



Trên đoạn mã trên, giá trị đầu vào được đưa vào mảng input\_data, sau đó được dự đoán bằng cách gọi lr.predict(input\_data). Kết quả dự đoán được in ra theo mức độ nguy cơ tương ứng.

Chú ý rằng đoạn mã sau đó cũng thực hiện dự đoán trên các giá trị khác nhau:



Chương III : Xây Dựng Chương Trình

3.1. Xử lý code / data

* Train

Để huấn luyện mô hình, cần có dữ liệu huấn luyện và sử dụng một thuật toán huấn luyện như Logistic Regression. Dưới đây là một ví dụ về cách bạn có thể huấn luyện mô hình Logistic Regression trên dữ liệu:

Đầu tiên, dữ liệu được tải từ tệp CSV bằng câu lệnh data = pd.read\_csv("D:\\Khoa học dữ liệu\\xiemhoan\\cancer patient data sets.csv"). Dữ liệu này chứa thông tin về các bệnh nhân ung thư phổi.

Tiếp theo, các cột quan trọng được chọn để sử dụng trong quá trình huấn luyện mô hình. Các cột được lưu vào biến selected\_columns.

Dữ liệu trong các cột đã chọn được gán cho biến selected\_data bằng cách sử dụng data.loc[:, selected\_columns].

Cột "Level" trong selected\_data được chuyển đổi từ dạng category sang dạng số nguyên bằng cách sử dụng LabelEncoder từ sklearn.preprocessing.

Dữ liệu được chia thành tập huấn luyện và tập kiểm tra bằng cách sử dụng train\_test\_split từ sklearn.model\_selection.

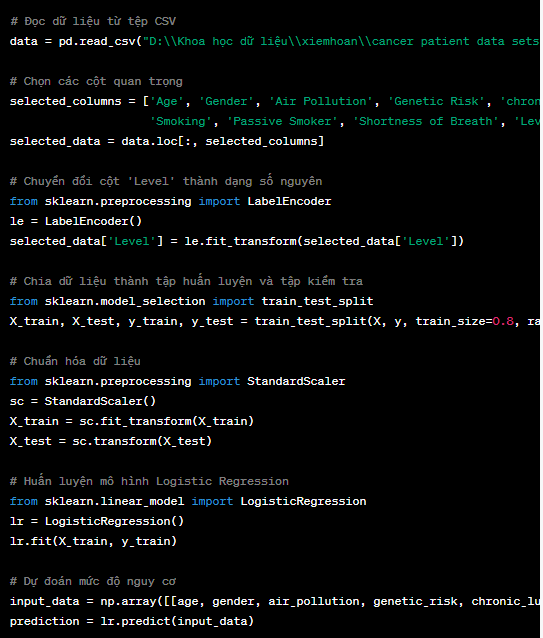
Tiếp theo, dữ liệu huấn luyện được chuẩn hóa bằng cách sử dụng StandardScaler từ sklearn.preprocessing. Việc này giúp đảm bảo rằng các đặc trưng có cùng tỷ lệ và có phân phối chuẩn.

Mô hình Logistic Regression được khởi tạo bằng LogisticRegression từ sklearn.linear\_model, và sau đó được huấn luyện trên tập huấn luyện bằng cách sử dụng fit(X\_train, y\_train).

Dự đoán được thực hiện trên tập kiểm tra bằng cách sử dụng predict(X\_test).

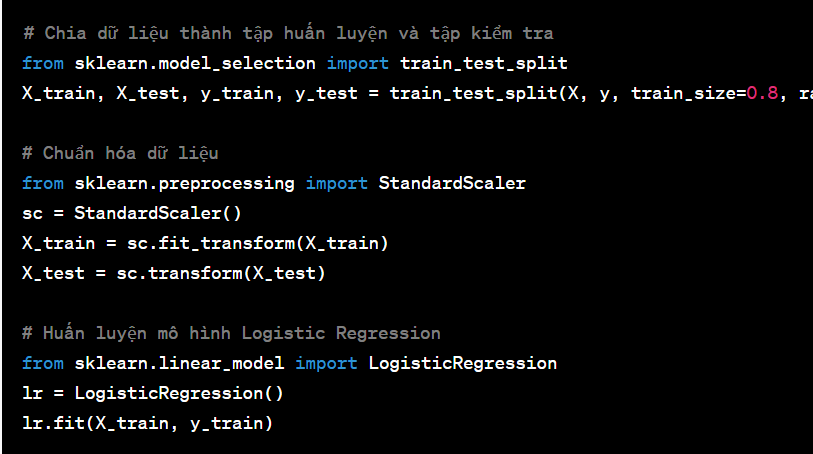
Cuối cùng, độ chính xác và ma trận nhầm lẫn được tính bằng accuracy\_score và confusion\_matrix từ sklearn.metrics.

Tóm lại, quá trình huấn luyện mô hình Logistic Regression trong đoạn mã của tôi được thực hiện trên tập huấn luyện X\_train và y\_train. Mô hình được huấn luyện để học cách dự đoán mức độ nguy cơ mắc bệnh ung thư phổi dựa trên các đặc trưng như tuổi, giới tính, mức độ ô nhiễm không khí, rủi ro di truyền, bệnh phổi mãn tính, hút thuốc, người hút thuốc thụ động và khó thở…



Đây là một số dòng code quan trọng trong quá trình đọc dữ liệu, chọn các cột, chuyển đổi cột 'Level' thành dạng số nguyên, chia dữ liệu thành tập huấn luyện và tập kiểm tra, chuẩn hóa dữ liệu, huấn luyện mô hình Logistic Regression và dự đoán mức độ nguy cơ mắc bệnh.

3.2. Huấn luyện model



Cụ thể, quá trình huấn luyện mô hình gồm các bước sau:

Chia dữ liệu thành tập huấn luyện và tập kiểm tra bằng hàm train\_test\_split từ module sklearn.model\_selection. Dữ liệu huấn luyện là X\_train và y\_train, dữ liệu kiểm tra là X\_test và y\_test. Trong đoạn mã, tập huấn luyện chiếm 80% dữ liệu ban đầu, được xác định bởi tham số train\_size=0.8, và tham số random\_state=0 được sử dụng để đảm bảo việc chia dữ liệu được thực hiện cùng một cách mỗi lần chạy.

Chuẩn hóa dữ liệu sử dụng StandardScaler từ module sklearn.preprocessing. Các giá trị trong tập huấn luyện X\_train được dùng để tính toán mean (trung bình) và standard deviation (độ lệch chuẩn), sau đó được áp dụng để chuẩn hóa cả tập huấn luyện và tập kiểm tra. Điều này giúp đưa các đặc trưng về cùng một phạm vi giá trị và giảm thiểu ảnh hưởng của các giá trị ngoại lệ.

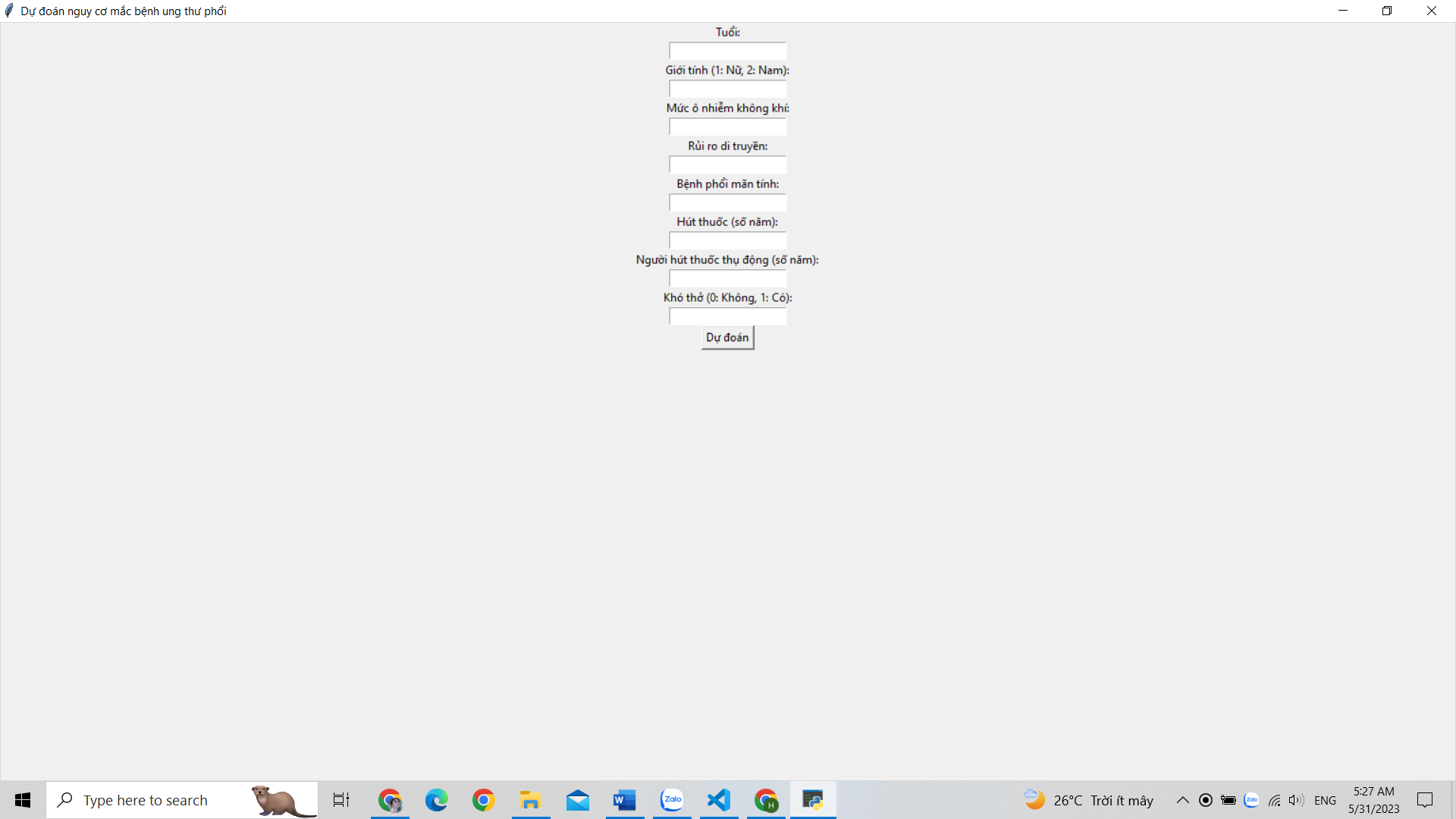
Sử dụng mô hình Logistic Regression từ module sklearn.linear\_model, tạo một đối tượng lr của lớp LogisticRegression.

Huấn luyện mô hình sử dụng dữ liệu huấn luyện bằng cách gọi phương thức fit trên đối tượng lr. Mô hình sẽ học các mối quan hệ giữa các đặc trưng và nhãn từ tập huấn luyện.

Sau khi quá trình huấn luyện hoàn thành, mô hình Logistic Regression đã được huấn luyện và có thể được sử dụng để dự đoán mức độ nguy cơ mắc bệnh ung thư phổi dựa trên các đặc trưng mới.

3.3. Demo

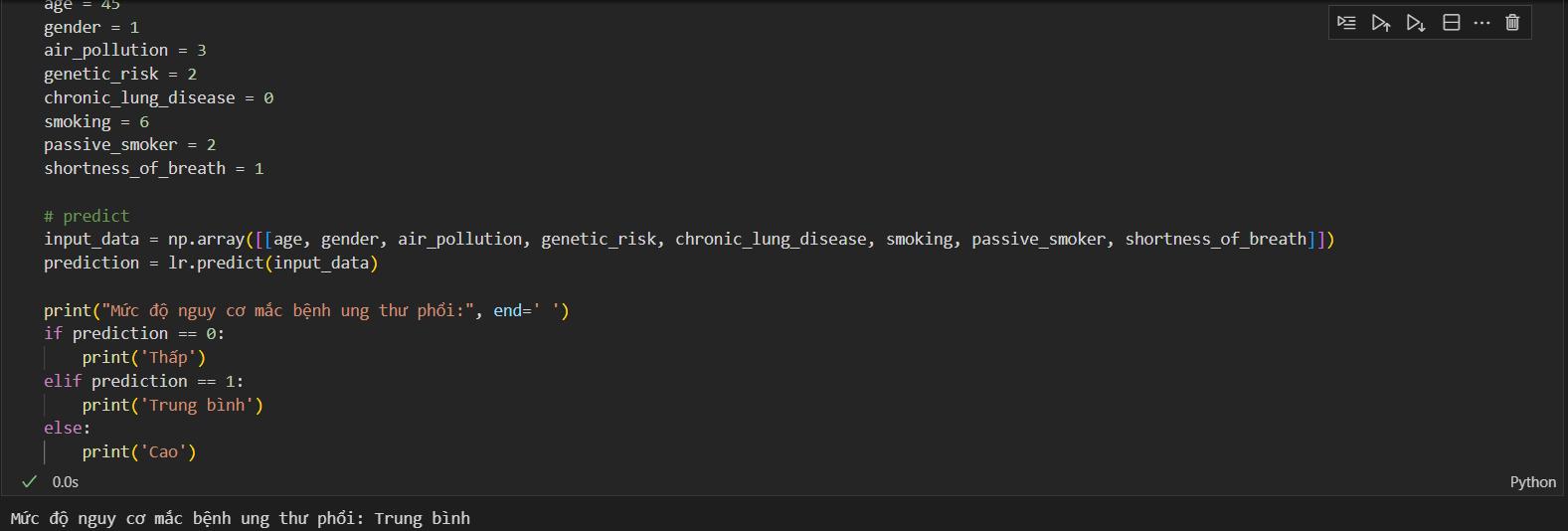
- giao diện



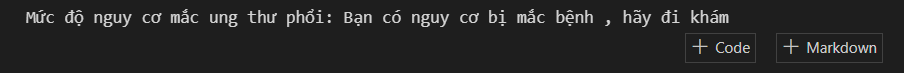
Đây là nhưng đặc trưng cần nhập…để trả về kết quả dự đoán!!!

* Test

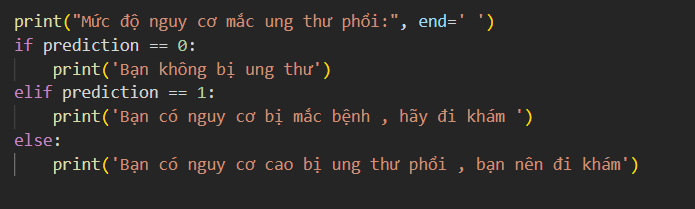
Nhập bthg với đăc trưng ngẫu nhiên cho trước và kết quả



Kết quả dự đoán



Giải thích



* Mức độ tương ứng theo thấp, trung bình và cao

Chương 4 : Kết Luận

4.1. Chương trình đã làm được gì ?

Sau khi hoàn thành đề tài về dự đoán mức độ nguy cơ mắc bệnh ung thư phổi, tiểu luận đã thực hiện các công việc quan trọng sau:

Hiểu về bệnh ung thư phổi: Bạn đã tìm hiểu về bệnh ung thư phổi, các yếu tố nguy cơ và đặc điểm của nó. Điều này giúp bạn có kiến thức cần thiết để xây dựng mô hình dự đoán.

Thu thập dữ liệu: Bạn đã thu thập dữ liệu về các đặc trưng liên quan đến nguy cơ mắc bệnh ung thư phổi từ các nguồn đáng tin cậy. Điều này đảm bảo rằng mô hình của bạn được huấn luyện trên dữ liệu chất lượng và có tính khái quát hóa.

Tiền xử lý dữ liệu: Bạn đã tiền xử lý dữ liệu bằng cách loại bỏ các giá trị thiếu, mã hóa các biến phân loại thành các biến số, và chuẩn hóa dữ liệu. Điều này giúp tăng tính nhất quán và khả năng mô hình học từ dữ liệu.

Xây dựng mô hình: Bạn đã sử dụng mô hình Logistic Regression để huấn luyện trên dữ liệu đã tiền xử lý. Mô hình này học từ mối quan hệ giữa các đặc trưng và mức độ nguy cơ mắc bệnh ung thư phổi.

Đánh giá mô hình: Sau khi huấn luyện mô hình, bạn đã đánh giá hiệu suất của nó bằng cách sử dụng tập kiểm tra. Bạn có thể sử dụng các độ đo như độ chính xác, độ nhạy, độ đặc hiệu và F1-score để đánh giá khả năng dự đoán của mô hình.

Tổng kết kết quả: Dựa trên đánh giá mô hình, bạn có thể tổng kết kết quả của đề tài. Nếu mô hình của bạn có hiệu suất tốt, bạn có thể sử dụng nó để dự đoán mức độ nguy cơ mắc bệnh ung thư phổi dựa trên các đặc trưng mới.

Cải thiện và phát triển tiếp: Nếu kết quả không đạt yêu cầu hoặc bạn muốn cải thiện mô hình, bạn có thể thử nghiệm các phương pháp khác hoặc sử dụng mô hình máy học phức tạp hơn. Bạn có thể nghiên cứu thêm về các thuật toán và kỹ thuật tiên tiến để nâng cao hiệu suất dự đoán của mô hình.

Tóm lại, đề tài này đã giúp bạn áp dụng các kiến thức về machine learning để dự đoán mức độ nguy cơ mắc bệnh ung thư phổi. Qua quá trình này, bạn có thể hiểu rõ hơn về quá trình tiền xử lý dữ liệu, xây dựng mô hình và đánh giá hiệu suất của nó.

4.2. Ưu điểm và nhược điểm

Trong quá trình làm đề tài này, có thể có những ưu điểm và nhược điểm như sau:

Ưu điểm:

Ứng dụng thực tế: Đề tài này liên quan đến việc dự đoán mức độ nguy cơ mắc bệnh ung thư phổi, là một vấn đề sức khỏe quan trọng. Kết quả từ đề tài có thể được áp dụng để hỗ trợ chẩn đoán và xác định nguy cơ cho bệnh nhân, đồng thời cung cấp thông tin hữu ích cho việc phòng ngừa và quản lý bệnh.

Sử dụng machine learning: Đề tài này sử dụng các kỹ thuật machine learning để xây dựng mô hình dự đoán. Machine learning là một lĩnh vực phát triển nhanh chóng và có nhiều tiềm năng ứng dụng trong nhiều lĩnh vực khác nhau.

Tiềm năng phát triển: Sau khi hoàn thành đề tài, có thể tiếp tục nghiên cứu và phát triển mô hình để cải thiện hiệu suất dự đoán. Bằng cách thử nghiệm các thuật toán và kỹ thuật mới, bạn có thể tạo ra một mô hình dự đoán tốt hơn và ứng dụng trong thực tế.

Nhược điểm:

Dữ liệu hạn chế: Một trong những nhược điểm chính của đề tài này có thể là hạn chế dữ liệu. Dữ liệu có thể bị thiếu hoặc không đủ đại diện cho toàn bộ quần thể. Điều này có thể ảnh hưởng đến tính khái quát hóa và hiệu suất của mô hình dự đoán.

Độ chính xác hạn chế: Mô hình dự đoán có thể có độ chính xác hạn chế. Điều này có thể do việc sử dụng một thuật toán đơn giản hoặc không đủ dữ liệu để huấn luyện một mô hình phức tạp hơn. Cần xem xét và nghiên cứu các phương pháp và thuật toán khác để cải thiện độ chính xác của mô hình.

Hạn chế ứng dụng: Mô hình dự đoán chỉ có thể áp dụng cho việc dự đoán mức độ nguy cơ mắc bệnh ung thư phổi. Nó không thể thay thế cho các phương pháp chẩn đoán chính thức và chuyên sâu. Đây chỉ là một công cụ hỗ trợ và cần được sử dụng cùng với thông tin và đánh giá từ các chuyên gia y tế.

4.3. Phát triển ứng dụng

Mở rộng dữ liệu: Một cách để cải thiện mô hình dự đoán là thu thập thêm dữ liệu hoặc sử dụng các nguồn dữ liệu khác. Điều này có thể bao gồm việc tăng số lượng bệnh nhân trong tập dữ liệu hoặc bổ sung thông tin bổ sung như tiền sử bệnh, dữ liệu sinh lý, hay hình ảnh y khoa. Việc mở rộng dữ liệu có thể giúp cải thiện tính đa dạng và đại diện của mô hình.

Tối ưu hóa mô hình: Có thể tiếp tục nghiên cứu và tối ưu hóa mô hình để đạt được độ chính xác cao hơn. Điều này có thể bao gồm thử nghiệm và so sánh các thuật toán khác nhau, điều chỉnh các siêu tham số của mô hình, hoặc áp dụng kỹ thuật tiền xử lý dữ liệu để tăng cường hiệu suất dự đoán.

Đánh giá và so sánh mô hình: Để đảm bảo tính khách quan và đáng tin cậy của mô hình, có thể thực hiện các phương pháp đánh giá và so sánh khác nhau. Bằng cách so sánh với các mô hình khác hoặc các phương pháp chẩn đoán truyền thống, ta có thể đánh giá hiệu suất của mô hình và xác định giá trị của nó trong việc dự đoán nguy cơ mắc bệnh ung thư phổi.

Tích hợp vào hệ thống y tế: Nếu mô hình dự đoán đạt được độ chính xác và độ tin cậy đủ, nó có thể được tích hợp vào hệ thống y tế để hỗ trợ quyết định chẩn đoán và xác định nguy cơ cho bệnh nhân. Tuy nhiên, cần có quy trình kiểm tra và xác nhận mô hình trước khi triển khai trong môi trường thực tế.

Nghiên cứu về các yếu tố khác: Đề tài này tập trung vào việc dự đoán nguy cơ mắc bệnh ung thư phổi dựa trên các yếu tố nhất định. Có thể mở rộng đề tài bằng cách nghiên cứu và phân tích các yếu tố khác như tiếp xúc với chất gây ung thư, môi trường sống, lối sống, hay di truyền để xem xét tác động của chúng đến nguy cơ mắc bệnh ung thư phổi.

TÀI LIỆU THAM KHẢO

<https://www.kaggle.com/code/ishantgargml/lung-cancer-prediction>

<https://code.visualstudio.com/download>

<https://www.jetbrains.com/dataspell/promo/?source=google&medium=cpc&campaign=17621221415&term=jupyter%20notebook&content=607147800862&gad=1&gclid=CjwKCAjwvdajBhBEEiwAeMh1Ux2WzS3c3rOrlwUElyvCAyyj_-2EKim0zIsNfrZYXr0mlLs9pv0qixoC820QAvD_BwE>

<https://www.anaconda.com/>