**BỘ CÔNG THƯƠNG**

**TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP THỰC PHẨM TP.HỒ CHÍ MINH**

**KHOA CÔNG NGHỆ THÔNG TIN**

-----o0o----

****

**BÀI TẬP TIỂU LUẬN HỌC PHẦN:**

**THỰC HÀNH HỌC MÁY**

**TÊN ĐỀ TÀI:**

**ỨNG DỤNG THUẬT TOÁN SVM**

**ĐỂ DỰ ĐOÁN BỆNH THẬN MÃN TÍNH**

**Thành phố Hồ Chí Minh, tháng 6 năm 2023**

**BỘ CÔNG THƯƠNG**

**TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP THỰC PHẨM TP.HỒ CHÍ MINH**

**KHOA CÔNG NGHỆ THÔNG TIN**

-----o0o----

****

**TÊN ĐỀ TÀI:**

**ỨNG DỤNG THUẬT TOÁN SVM**

**ĐỂ DỰ ĐOÁN BỆNH THẬN MÃN TÍNH**

|  |
| --- |
| **Nhóm: 2**  **Giảng Viên Hướng Dẫn: Trần Đình Toàn**  **Trưởng nhóm:** **Phạm Tiến Phúc - 2001207043**  **Thành viên:**  **Đỗ Chí Hùng - 2001207048**  **Lưu Thiên Phong - 2001200101** |

**Thành phố Hồ Chí Minh, tháng 6 năm 2023**

# Bảng phân công

|  |  |  |  |
| --- | --- | --- | --- |
| **MSSV** | **Họ Tên** | **Công việc** | **Đánh giá** |
| 2001207043 | Phạm Tiến Phúc | Viết code xây dựng ứng dụng Desktop Application và trên nền tảng web  Làm word với ppt  Viết code trực quan hóa dữ liệu, chuẩn hóa dữ liệu và kết nối CSDL | 100% |
| 2001207048 | Đỗ Chí Hùng | Tìm hiểu thuật toán SVM, bệnh thận, xây dựng CSDL | 100% |
| 2001200101 | Lưu Thiên Phong | Tìm hiểu thuật toán SVM, bệnh thận, xây dựng CSDL | 100% |

# Mục lục

[Bảng phân công 3](#_Toc151)

[Mục lục 4](#_Toc21948)

[Lời nói đầu 6](#_Toc15250)

[CHƯƠNG I: TỔNG QUAN VỀ KHAI THÁC DỮ LIỆU 7](#_Toc13168)

[1. Khái niệm cơ bản 7](#_Toc1056)

[2. Khai thác dữ liệu 7](#_Toc3046)

[3. Kỹ thuật phân lớp 7](#_Toc24323)

[4. Kỹ thuật phân cụm 8](#_Toc24203)

[CHƯƠNG II: THUẬT TOÁN SVM 9](#_Toc3270)

[1. Giới thiệu SVM 9](#_Toc30029)

[2. Vì sao chọn SVM 9](#_Toc13569)

[3. Phương trình của SVM 9](#_Toc20797)

[4. Sai số dự đoán 11](#_Toc21745)

[5. Hàm mất mát 12](#_Toc16717)

[6. Hàm đánh giá 13](#_Toc21445)

[CHƯƠNG III: THUẬT TOÁN SVM VÀ BÀI TOÁN DỰ ĐOÁN BỆNH THẬN MÃN TÍNH 15](#_Toc7146)

[1. Phân tích bài toán 15](#_Toc29253)

[2. Xây dựng cơ sở dữ liệu 15](#_Toc6788)

[a. Bộ dữ liệu huấn luyện 15](#_Toc14529)

[b. Trong bộ dữ liệu có các thuộc tính gồm 15](#_Toc3656)

[3. Cài đặt thuật toán 17](#_Toc23916)

[4. Cài đặt ứng dụng 29](#_Toc8073)

[Kết luận 32](#_Toc11911)

[Tài liệu tham khảo 34](#_Toc5561)

# Lời nói đầu

Trong thời kỳ xã hội phát triển hiện đại như ngày nay, việc ứng dụng những thành tựu khoa học công nghệ thông tin với tất cả các lĩnh vực khác nhau được phát triển không ngừng, lượng cở sở dữ liệu cũng như thông tin được thu thập và lưu trữ ngày càng trở nên nhiều lên. Con người chúng ta cũng vì thế mà cần có thông tin với tốc độ xử lý nhanh chống nhằm để đưa ra quyết định một cách chính xác từ những cơ sở dữ liệu vốn đã được thu thập và lưu trữ. Trong đó, việc quản lý và khai thác cở sở dữ liệu theo cách truyền thông thì ngày càng không được đáp ứng với như cầu của xã hội ngày nay. Vì vậy, việc nghiên cứu và đưa ra phương pháp theo khuynh hướng kỹ thuật mới là vô cũng cần thiết nhằm đáp ứng như cầu đối với xã hội ngày nay.

Việc thu thập và khai thác cơ sở dữ liệu đã và đang được nghiên cứu và phát triển, nhằm để ứng dụng trong mọi lĩnh vực khác nhau trên thế giới. Đối với đất nước Việt Nam thì phương pháp này đang được nghiên cứu và dần có thể đưa vào ứng dụng. Hiện nay tất cả mọi người không ngừng tìm tòi và phát triển các ứng dụng kỹ thuật mới để thực hiện việc khai thác dữ liệu một cách nhanh chống và mong muốn đạt được kết quả tốt nhất.

Đối với bài toán dự đoán được “bệnh thận mãn tính” thì nhóm em sẽ trình bày về một kỹ thuật trong khai thác các dữ liệu để phân lớp các dữ liệu bệnh nói chung và bênh viêm gan nói riêng.

# CHƯƠNG I: TỔNG QUAN VỀ KHAI THÁC DỮ LIỆU

## Khái niệm cơ bản

Khai thác dữ liệu là kỹ thuật có sự hỗ trợ của máy tính được sử dụng trong hoạt động phân tích để xử lý và khám phá các tập dữ liệu lớn. Nhờ có công cụ và phương pháp khai thác dữ liệu, các tổ chức có thể khám phá những mẫu hình và mối quan hệ ẩn trong dữ liệu của họ. Quá trình khai thác dữ liệu chuyển đổi dữ liệu thô thành kiến thức thực tế. Các công ty áp dụng kiến thức này để giải quyết vấn đề, phân tích tác động trong tương lai từ quyết định kinh doanh và tăng biên lợi nhuận của họ.

## Khai thác dữ liệu

“Khai thác dữ liệu” là một thuật ngữ dùng sai do mục tiêu của quy trình khai thác dữ liệu không phải là trích xuất hoặc khai thác chính dữ liệu đó. Thay vào đó, một lượng lớn dữ liệu đã có sẵn và quy trình khai thác dữ liệu sẽ trích xuất ý nghĩa hoặc kiến thức có giá trị từ dữ liệu đó. Bên dưới là nội dung phác thảo quy trình thu thập, lưu trữ, phân tích và khai thác dữ liệu điển hình.

Thu thập dữ liệu là quá trình ghi lại dữ liệu từ nhiều nguồn khác nhau như phản hồi của khách hàng, thanh toán và đơn đặt hàng.

Lưu kho dữ liệu là quy trình lưu trữ dữ liệu đó trong một cơ sở dữ liệu lớn hoặc [kho dữ liệu](https://aws.amazon.com/data-warehouse/).

Phân tích dữ liệu sẽ xử lý, lưu trữ và phân tích dữ liệu sâu hơn bằng các phần mềm và thuật toán phức tạp.

Khai thác dữ liệu là một nhánh của phân tích dữ liệu hoặc chiến lược phân tích được sử dụng để tìm các kiểu mẫu ẩn hoặc chưa biết trước đó trong dữ liệu.

## Kỹ thuật phân lớp

Trong kỹ thuật phân lớp gồm có các thuật toán

Phân lớp bằng cây quyết định (giải thuật ID3, J48) phân lớp dữ liệu dựa trên việc lập nên cây quyết định, nhìn vào cây quyết định có thể ra quyết định dữ liệu thuộc phân lớp nào.

Phân lớp dựa trên xác suất (Naive Bayesman) dựa trên việc giả định các thuộc tỉnh độc lập mạnh với nhau qua việc sử dụng định lý Baves. Phân lớp dựa trên khoảng cách (giải thuật K – láng giềng): làm như làng giềng làm, dữ liệu sẽ được phân vào lớp của k đối tượng gần với dữ liệu đỏ nhất.

Phần lớp bảng SVM phân lớp dữ liệu dựa trên việc tìm ra một siêu phẳng “tốt nhất” để tách các lớp dữ liệu trên không gian nhiều chiều hơn.

## Kỹ thuật phân cụm

Phân cụm dữ liệu là cách phân bố các đối tượng dữ liệu vào các nhóm cụm sao cho các đối tượng trong một cụm thì giống nhau hơn các phần tử khác cụm.

Phân cụm bằng phương pháp K-mean tìm ra tâm của các cụm mà khoảng cách của tâm đó đến các đối tượng, dữ liệu khác là ngăn + Phân cụm trên đỗ thị

Ngoài ra, khai phá dữ liệu có rất nhiều kỹ thuật, nhưng đây là những kỹ thuật cơ bản và đơn giản trong khai phá dữ liệu mà chúng em được tìm hiểu.

# CHƯƠNG II: THUẬT TOÁN SVM

## Giới thiệu SVM

Support Vector Machine (SVM):là một thuật toán học máy có giám sát được sử dụng phổ biến trong các bài toán phần lớp (classification) và hồi quy (Regression).

Thuật toán SVM (Support Vector Machine) là một trong những thuật toán phân loại trong học máy. Nó được sử dụng rộng rãi trong việc giải quyết các bài toán phân loại dữ liệu có tính phi tuyến tính.

Thuật toán SVM hoạt động dựa trên việc tìm ra một đường phân chia tối ưu giữa các lớp dữ liệu.

## Vì sao chọn SVM

Rất hiệu quả cho việc giải quyết các bài toán có dữ liệu lớn

Giải quyết vấn đề Overfitting rất tốt

Là phương pháp phân lớp nhanh

Có hiệu suất tổng hợp tốt và hiệu suất tính toán cao

## Phương trình của SVM

Phương trình của SVM phụ thuộc vào loại hàm kernel được sử dụng. Để giải thích phương trình của SVM, chúng ta sẽ sử dụng ví dụ phân loại hai lớp dữ liệu với hàm kernel tuyến tính. Giả sử chúng ta có tập dữ liệu gồm N điểm, mỗi điểm được biểu diễn bởi một vector đặc trưng đầu vào x và một giá trị nhãn y (-1 hoặc 1) cho biết lớp của điểm đó.

Chúng ta cần tìm một siêu phẳng trong không gian đặc trưng (một siêu phẳng là một đường thẳng trong không gian hai chiều, một mặt phẳng trong không gian ba chiều và một siêu phẳng trong không gian có số chiều lớn hơn) để phân chia hai lớp sao cho khoảng cách từ siêu phẳng đến các điểm dữ liệu gần nhất là lớn nhất. Để đơn giản hóa, chúng ta giả sử rằng siêu phẳng này có phương trình dạng: w^T x + b = 0 trong đó w là vector trọng số và b là độ lệch (bias). Để tìm siêu phẳng tối ưu, chúng ta cần tìm vector trọng số w và độ lệch b sao cho khoảng cách từ siêu phẳng đến các điểm dữ liệu gần nhất là lớn nhất. Khoảng cách này được tính bằng cách đo đường vuông góc từ một điểm dữ liệu đến siêu phẳng: d = y(w^T x + b) / ||w|| trong đó y là nhãn của điểm dữ liệu, ||w|| là độ dài của vector trọng số.

Tối ưu hóa siêu phẳng SVM đồng nghĩa với việc tối đa hoá khoảng cách d, tức tìm giá trị w và b sao cho d lớn nhất. Tuy nhiên, để tính toán được khoảng cách d, ta cần thực hiện một số ràng buộc:

Tất cả các điểm của dữ liệu phải nằm đúng lớp của chúng, tức là y(w^T x + b) ≥ 1.

Khoảng cách từ siêu phẳng đến điểm dữ liệu gần nhất phải là lớn nhất, tức tối đa hoá ||w||.

Để giải quyết bài toán tối ưu hóa này, chúng ta có thể sử dụng phương pháp Lagrange để tính toán các hệ số alpha\_i của các điểm dữ liệu. Sau đó, siêu phẳng được xác định bởi các điểm dữ liệu có alpha\_i khác 0, được gọi là support vectors. Các điểm dữ liệu còn lại không cần thiết để xác định siêu phẳng.

Với hàm kernel tuyến tính, phương trình của siêu phẳng của SVM có dạng: f(x) = sign(w^T x + b) trong đó f(x) là hàm phân loại, sign là hàm signum (trả về giá trị -1 hoặc 1), w là vector trọng số và b là độ lệch (bias) được tính toán dựa trên các hệ số alpha\_i và nhãn của các điểm dữ liệu.

Với các loại hàm kernel khác nhau, phương trình của siêu phẳng SVM sẽ có dạng khác nhau, tuy nhiên cơ bản vẫn là tìm một siêu phẳng phân chia hai lớp sao cho margin là lớn nhất.

## Sai số dự đoán

Sai số dự đoán (prediction error) là sự khác biệt giữa giá trị dự đoán của mô hình và giá trị thực tế của dữ liệu.

Nó được đo bằng một số độ đo như độ chính xác (accuracy), độ phân loại sai (misclassification rate), độ F1 (F1-score), và AUC-ROC. Độ chính xác (accuracy) là độ đo phổ biến nhất để đo sai số dự đoán.

Nó được tính bằng tỷ lệ số lượng điểm được dự đoán đúng trên tổng số lượng điểm trong tập dữ liệu kiểm tra:

accuracy = (số lượng dự đoán đúng) / (tổng số lượng điểm)

Tuy nhiên, độ chính xác không phù hợp để đo lường hiệu suất của mô hình trong các trường hợp mà số lượng điểm của các lớp trong tập dữ liệu không cân bằng. Trong trường hợp này, độ phân loại sai (misclassification rate) đượcsử dụng để đo sai số dự đoán.

Độ phân loại sai là tỷ lệ giữa số lượng điểm được dự đoán sai và tổng số lượng điểm trong tập dữ liệu kiểm tra:

misclassification rate = (số lượng dự đoán sai) / (tổng số lượng điểm)

Độ F1 (F1-score) là một độ đo kết hợp giữa độ chính xác và độ phân loại sai, được tính bằng trung bình điều hòa của độ chính xác và độ phân loại sai:

F1-score = 2 \* (precision \* recall) / (precision + recall)

Trong đó precision (độ chắc chắn) là tỷ lệ số lượng điểm dự đoán đúng trong số các điểm được dự đoán là thuộc lớp dương, và recall (độ phủ) là tỷ lệ số lượng điểm dự đoán đúng trong số các điểm thực sự thuộc lớp dương.

Ngoài ra, AUC-ROC (Area Under the Receiver Operating Characteristic Curve) là một độ đo khác được sử dụng để đánh giá hiệu suất của mô hình phân loại. Nó đo lường khả năng của mô hình phân loại đưa ra các dự đoán chính xác cho cả hai lớp, và được tính bằng diện tích dưới đường cong ROC (Receiver Operating Characteristic) của mô hình. Đường cong ROC là biểu đồ trình bày tỷ lệ giữa true positive rate (tỷ lệ số lượng điểm thuộc lớp dương được dự đoán đúng) và false positive rate (tỷ lệ số lượng điểm thuộc lớp âm bị dự đoán là thuộc lớp dương) với các ngưỡng dự đoán khác nhau. AUC-ROC càng gần 1 thì mô hình phân loại càng tốt.

Sai số dự đoán (prediction error) là sự khác biệt giữa giá trị dự đoán của mô hình và giá trị thực tế của dữ liệu. Nó được đo bằng một số độ đo như độ chính xác (accuracy), độ phân loại sai (misclassification rate), độ F1 (F1-score), và AUC-ROC.

## Hàm mất mát

Hàm mất mát (loss function) là một hàm số đo lường mức độ sai lệch giữa giá trị dự đoán của mô hình và giá trị thực tế của dữ liệu. Mục đích của hàm mất mát là đưa ra một độ đo để đánh giá hiệu suất của mô hình và giúp tối ưu hóa các tham số của mô hình.

Ví dụ, trong bài toán phân loại nhị phân, hàm mất mát thường được sử dụng là hàm cross-entropy (hay negative log-likelihood), được tính bằng công thức sau đây:

L(y, f(x)) = -[y \* log(f(x)) + (1 - y) \* log(1 - f(x))]

Trong đó y là nhãn thực tế của dữ liệu (1 hoặc 0), f(x) là giá trị dự đoán của mô hình cho dữ liệu x. Hàm số này có ý nghĩa rằng, nếu giá trị dự đoán của mô hình gần vớigiá trị thực tế, thì giá trị của hàm mất mát sẽ gần bằng 0. Ngược lại, nếu giá trị dự đoán của mô hình rất khác so với giá trị thực tế, thì giá trị của hàm mất mát sẽ lớn.

Trong các bài toán khác, hàm mất mát có thể có dạng khác nhau tùy thuộc vào loại mô hình và bài toán. Ví dụ, trong bài toán hồi quy tuyến tính, hàm mất mát thường được sử dụng là Mean Squared Error (MSE), được tính bằng công thức sau:

L(y, f(x)) = (y - f(x))^2

Trong đó y là giá trị thực tế của biến phụ thuộc, f(x) là giá trị dự đoán của mô hình cho giá trị đầu vào x. Hàm số này có ý nghĩa rằng, nếu giá trị dự đoán của mô hình gần với giá trị thực tế, thì giá trị của hàm mất mát sẽ gần bằng 0. Ngược lại, nếu giá trị dự đoán của mô hình rất khác so với giá trị thực tế, thì giá trị của hàm mất mát sẽ lớn.

Hàm mất mát có vai trò quan trọng trong quá trình huấn luyện mô hình, bởi vì nó được sử dụng để tính toán gradient, tức là đạo hàm của hàm mất mát theo các tham số của mô hình. Gradient này được sử dụng để cập nhật các tham số của mô hình trong quá trình tối ưu hóa. Mục tiêu của quá trình tối ưu hóa là tìm ra các giá trị tham số của mô hình sao cho hàm mất mát là nhỏ nhất. Khi giá trị của hàm mất mát là nhỏ nhất, mô hình sẽ cho ra kết quả tốt nhất có thểđối với tập dữ liệu huấn luyện đã cho.

## Hàm đánh giá

Hàm đánh giá (evaluation function) là một hàm số được sử dụng để đánh giá hiệu suất của một mô hình học máy trên một tập dữ liệu kiểm tra. Mục đích của hàm đánh giá là đo lường khả năng của mô hình dự đoán các giá trị đầu ra chính xác trên dữ liệu mới.

Ví dụ, trong bài toán phân loại, hàm đánh giá thường được sử dụng là độ chính xác (accuracy), được tính bằng tỷ lệ số lượng điểm được dự đoán đúng trên tổng số lượng điểm trong tập dữ liệu kiểm tra. Các hàm đánh giá khác bao gồm độ phủ (recall), độ chắc chắn (precision), độ F1 (F1-score), AUC-ROC, và số khác.

Trong bài toán hồi quy, hàm đánh giá thường được sử dụng là Root Mean Squared Error (RMSE), tức căn bậc haicủa trung bình cộng của bình phương sai số giữa giá trị thực tế và giá trị dự đoán của mô hình trên tập dữ liệu kiểm tra. Các hàm đánh giá khác bao gồm Mean Absolute Error (MAE) và R-Squared (R^2).

Hàm đánh giá là một công cụ quan trọng để đánh giá hiệu suất của một mô hình học máy. Nó giúp cho người dùng có thể so sánh hiệu suất của các mô hình khác nhau trên cùng một tập dữ liệu kiểm tra. Tuy nhiên, khi lựa chọn hàm đánh giá, cần phải cân nhắc đến các yếu tố khác như tính cân bằng giữa độ chính xác và độ phủ, khả năng xử lý dữ liệu không cân bằng, và các yếu tố khác của tập dữ liệu và mô hình.

# CHƯƠNG III: THUẬT TOÁN SVM VÀ BÀI TOÁN DỰ ĐOÁN BỆNH THẬN MÃN TÍNH

## Phân tích bài toán

Suy thận mãn tính mô tả sự mất dần chức năng thận. Thận lọc các chất thải và dịch dư thừa từ máu và sau đó được bài tiết trong nước tiểu. Suy thận mãn tính thường thiệt hại thận, dịch và chất thải có thể tích tụ trong cơ thể ở mức độ nguy hiểm

Trong giai đoạn đầu suy thận mãn tính, có thể có vài dấu hiệu hoặc triệu chứng. Suy thận mãn tính có thể không rõ cho đến khi chức năng thận có ảnh hưởng đáng kể.

Điều trị suy thận mãn tính, còn gọi là bệnh thận mãn tính, tập trung vào việc làm chậm sự tiến triển của suy thận, thường là bằng cách kiểm soát các nguyên nhân. Suy thận mãn tính có thể tiến triển đến giai đoạn cuối bệnh thận, là tử vong nếu không lọc nhân tạo (lọc máu) hoặc ghép thận .

## Xây dựng cơ sở dữ liệu

### Bộ dữ liệu huấn luyện

Bộ dữ liệu để huấn luyện mô hình phân biệt xem bệnh nhân có bị thận mãn tính hay không.

Tập dữ liệu chứa các giá trị trong phòng thí nghiệm của người xét nghiệm và bệnh nhân mắc chứng thận mãn tính và các giá trị khác. Dữ liệu được lấy từ Kaggle của tác giá Tuấn Hiệp Nguyễn.

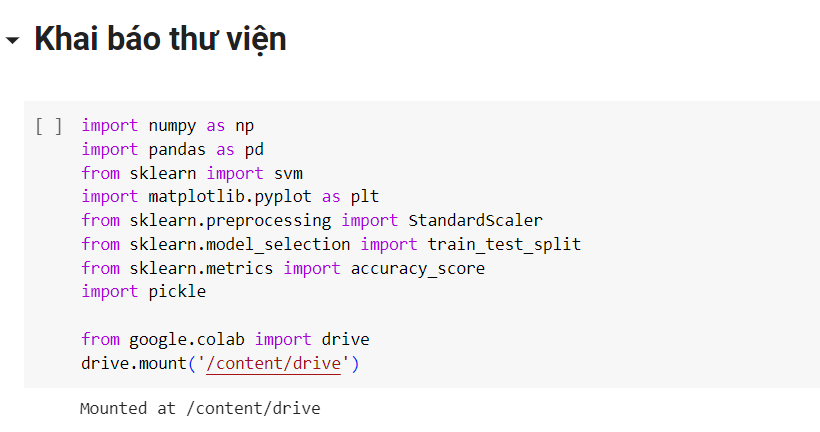
### Trong bộ dữ liệu có các thuộc tính gồm

* Bp: hay còn được hiểu là blood pressure - áp lực máu.
* Sg: hay còn được hiểu là Specific Gravity là chỉ số đo nồng độ chất rắn trong nước tiểu.
* Al: viết tắt của Albumin là một loại protein có trong huyết thanh.
* Su: viết tắt của Sugar là chỉ số thể hiện lượng đường trong máu
* Rbc: viết tắt của Red Blood Cells là chỉ số đo lượng tế bào máu đỏ trong một đơn vị máu.
* Pc: viết tắt của Platelet Count là chỉ số đo số lượng tiểu cầu trong một đơn vị máu.
* Pcc: viết tắt của Prothrombin Complex Concentrate là chỉ số đo nồng độ các yếu tố đông máu trong máu.
* Ba: viết tắt của Blood Urea Nitrogen là urea máu đo lượng urea trong máu để đánh giá chức năng thận.
* Bgr: viết tắt của Blood Glucose Random là đường huyết ngãu nhiên đo nồng độ đường trong máu để đánh giá chức năng đường huyết.
* Bu: viết tắt của Blood Uric Acid là axit uric máu đo lượng axit uric trong máu để đánh giá chức năng thận.
* Sc: viết tắt của Serum Creatinine - Creatinin huyết thanh.
* Sod: viết tắt của Sodium là dùng để đo lượng natri trong máu để đánh giá chức năng thận.
* Pot: viết tắt của Potassium là đo lượng Kali trong máu.
* Hemo: viết tắt của Hemoglobin là đo lượng hồng cầu trong máu.
* Pcv: viết tắt của Packed Cell Volume là tỷ lệ giữa thể tích hồng cầu và toàn bộ máu.
* Wc: viết tắt của White Blood Cell là đo lượng bạch cầu trong máu.
* Rc: viết tắt của Red Blood Cell Count là chỉ số lượng tế bào hồng cầu trên một microlit máu.
* Htn: viết tắt của Hypertension là để đánh giá mức độ tăng huyết áp.
* Dm: viết tắt của Diabetes Mellitus là bệnh tiểu đường
* Cad: viết tắt của Coronary Artery Disease là bệnh động mạch vành.
* Appet: viết tắt của Appetite là đánh giá sự thèm ăn
* Pe: viết tắt của Physical Examination là khám cơ thể
* Ane: viết tắt của Anemia là bệnh thiếu máu.
* Classification: đánh giá mức độ suy thận theo các giai đoạn từ 1 đến 5 với 1 là nhẹ nhất và 5 là nặng nhất.

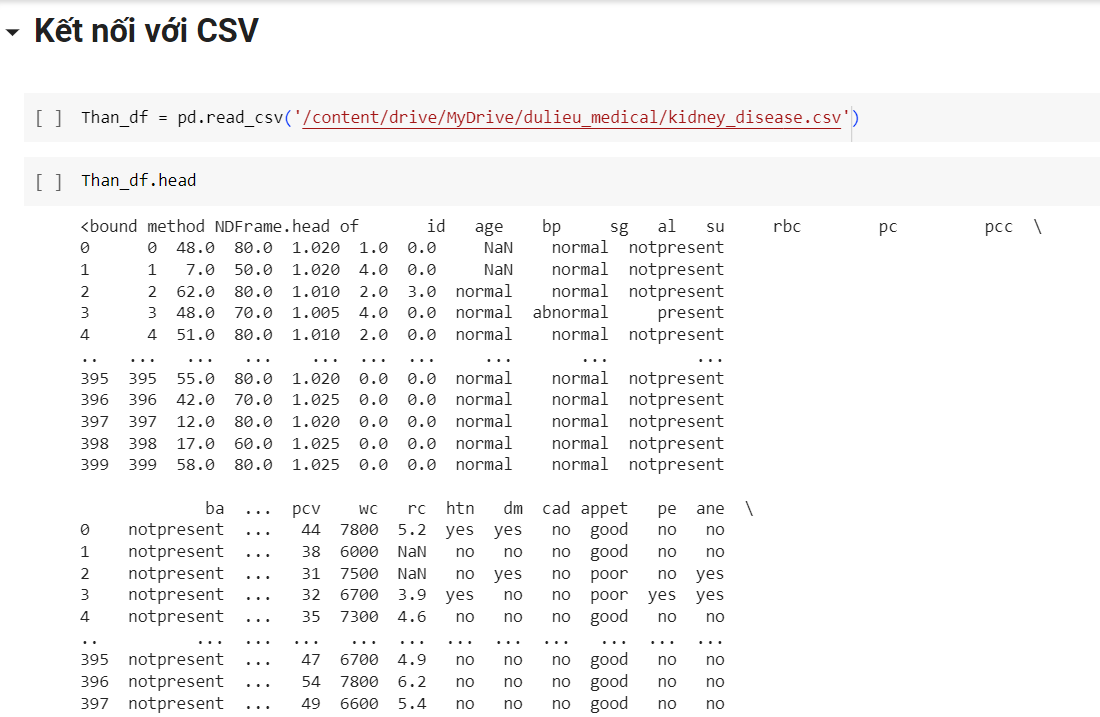
Thông qua việc phân tích các chỉ số này trong kết quả xét nghiệm, chúng ta có thể nhận được thông tin quan trọng về chức năng thận, sự cân bằng hóa chất trong cơ thể, và một số chỉ số liên quan đến bệnh thận mãn tính. Tuy nhiên, việc đánh giá bệnh thận mãn tính cần có sự phân tích kỹ lưỡng hơn, bao gồm việc sử dụng nhiều thông tin khác nhau và các xét nghiệm bổ sung.

## Cài đặt thuật toán

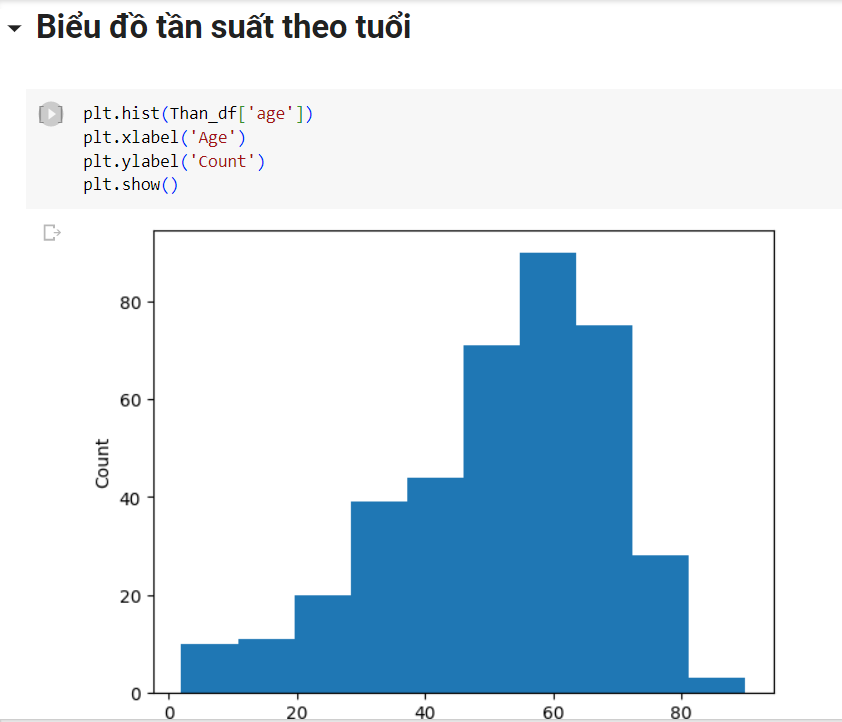
Thêm các thư viện cần thiết



Kết nối dataset và hiển thị lên

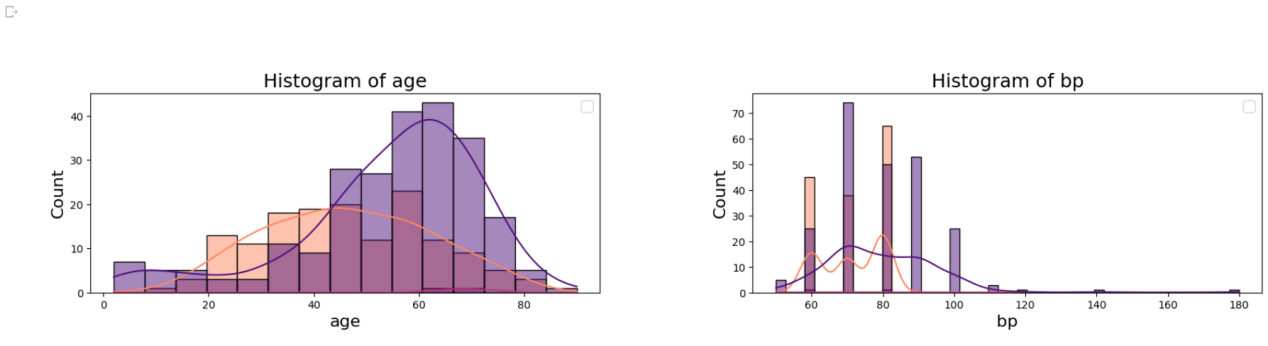


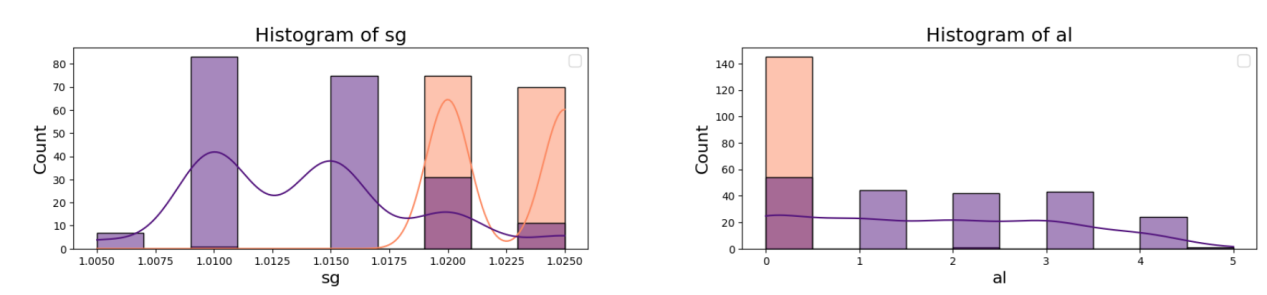
Biểu đồ tần suất theo tuổi

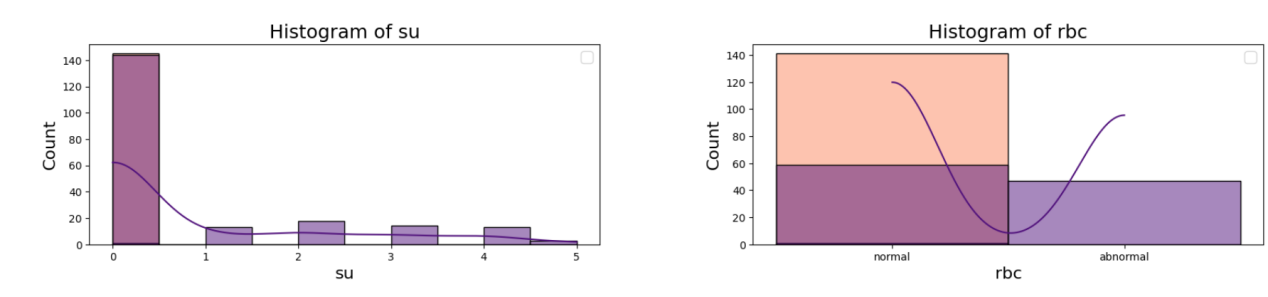


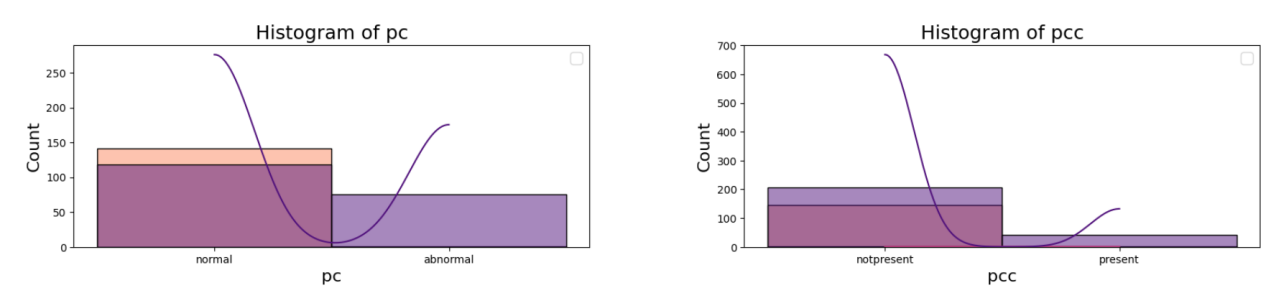
Phân phối kết quả xét nghiệm theo theo từng loại khác nhau

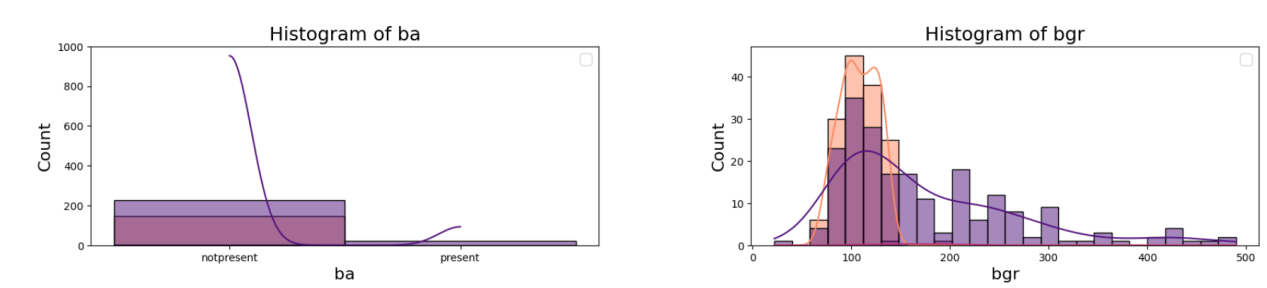


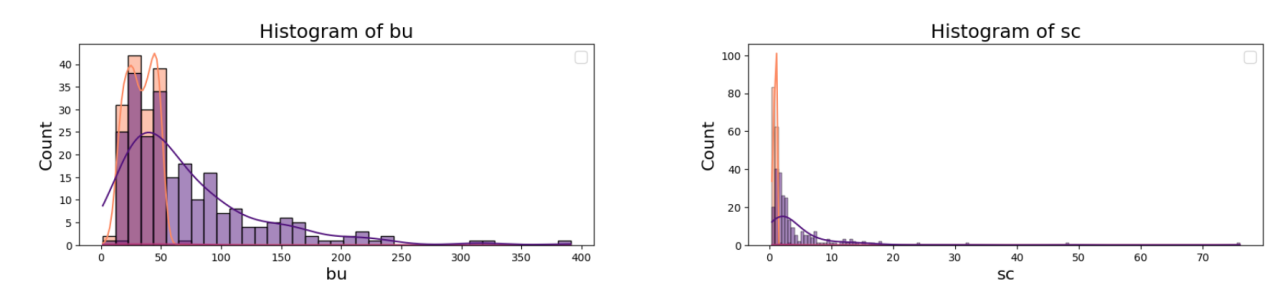


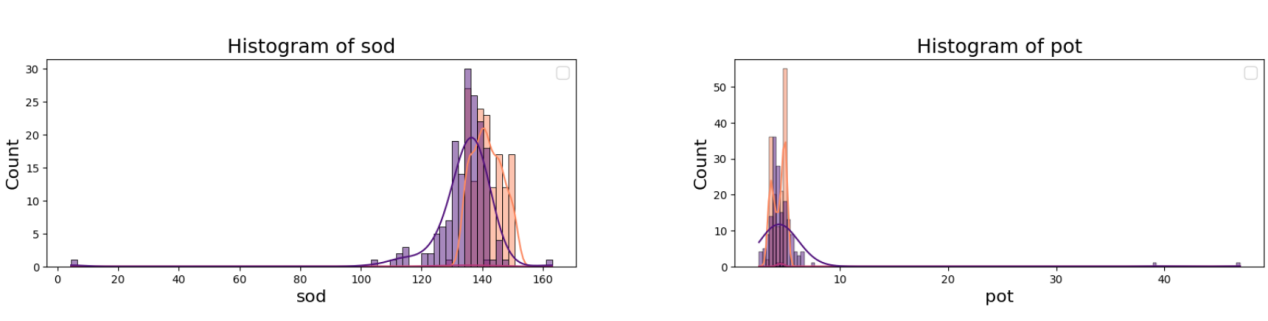


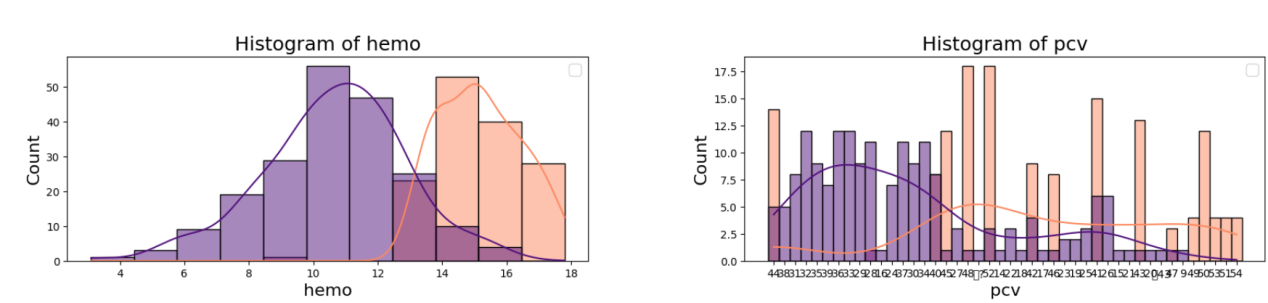


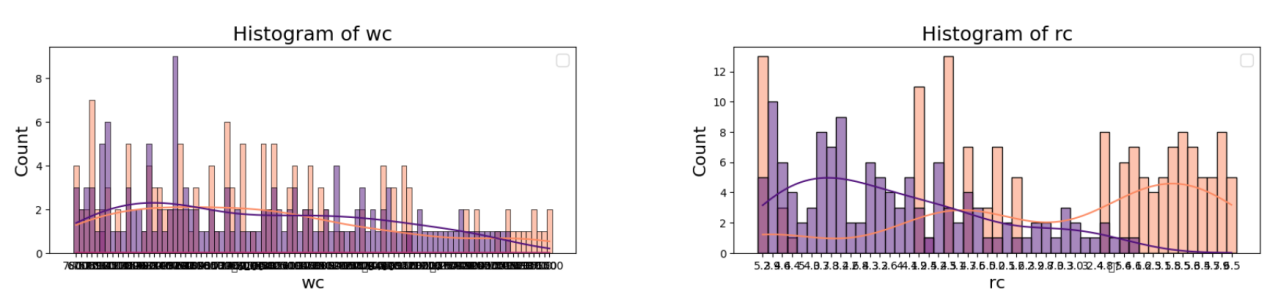


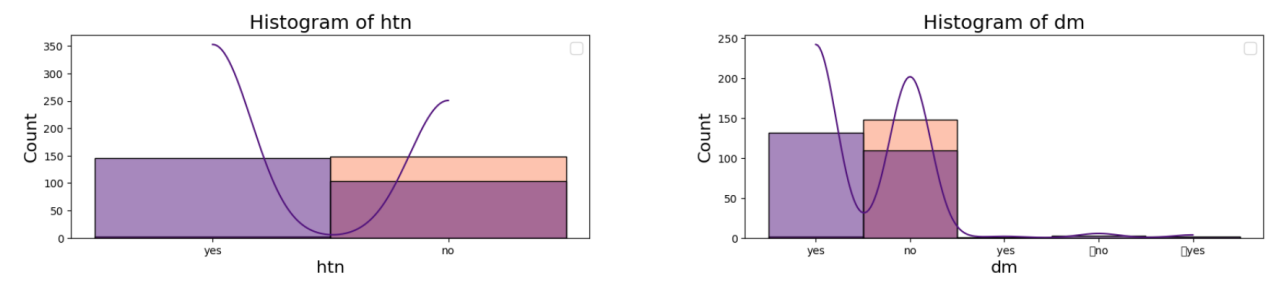


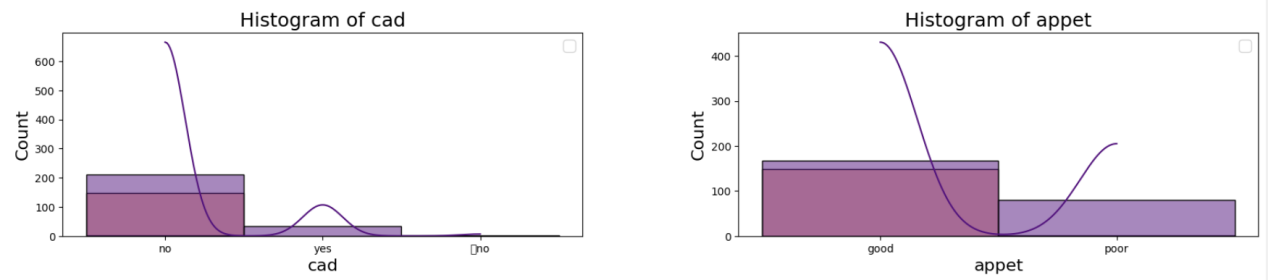


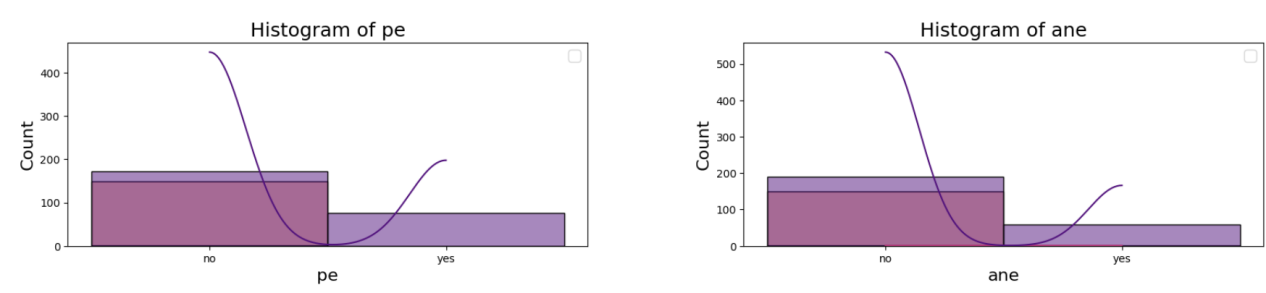




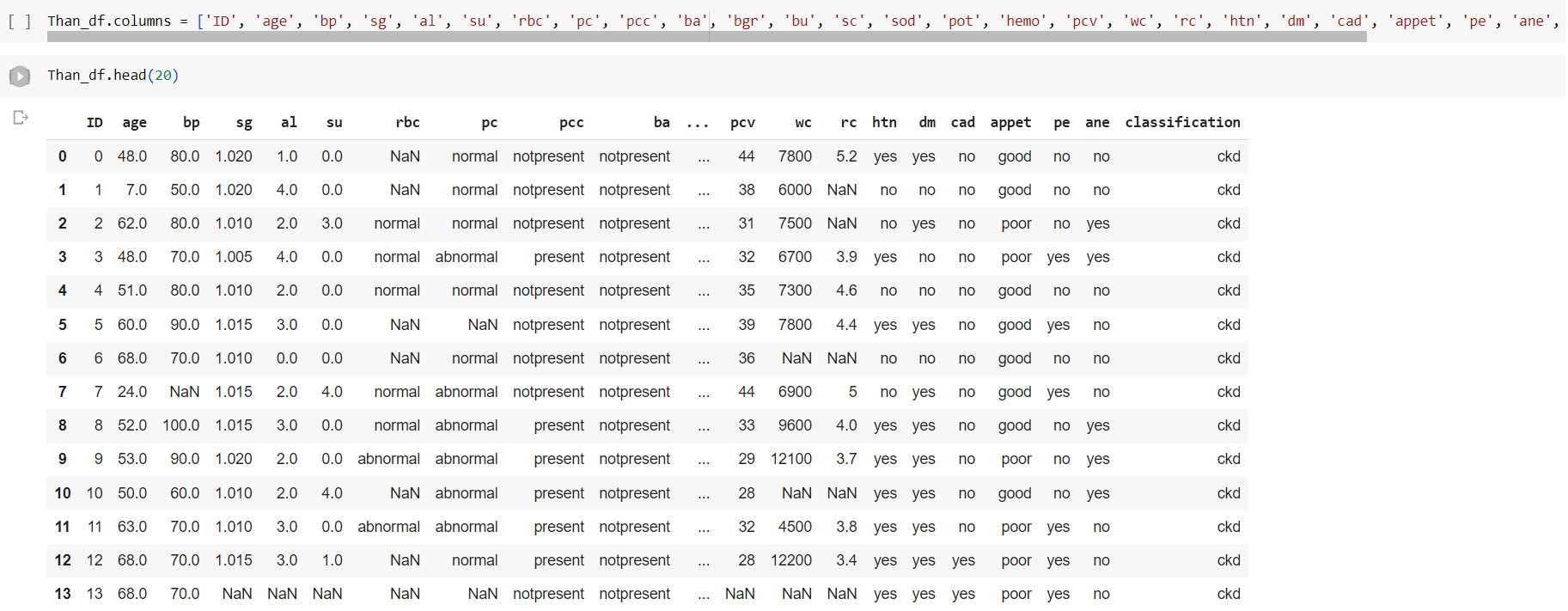








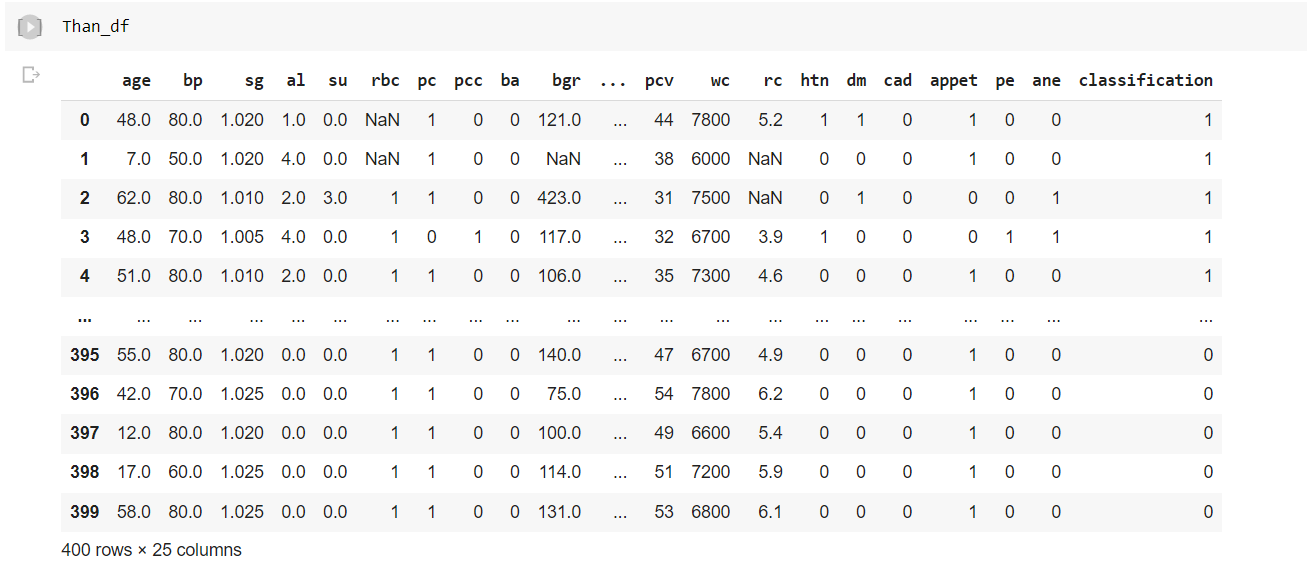
Hiển thị 20 dòng dữ liệu đầu tiên



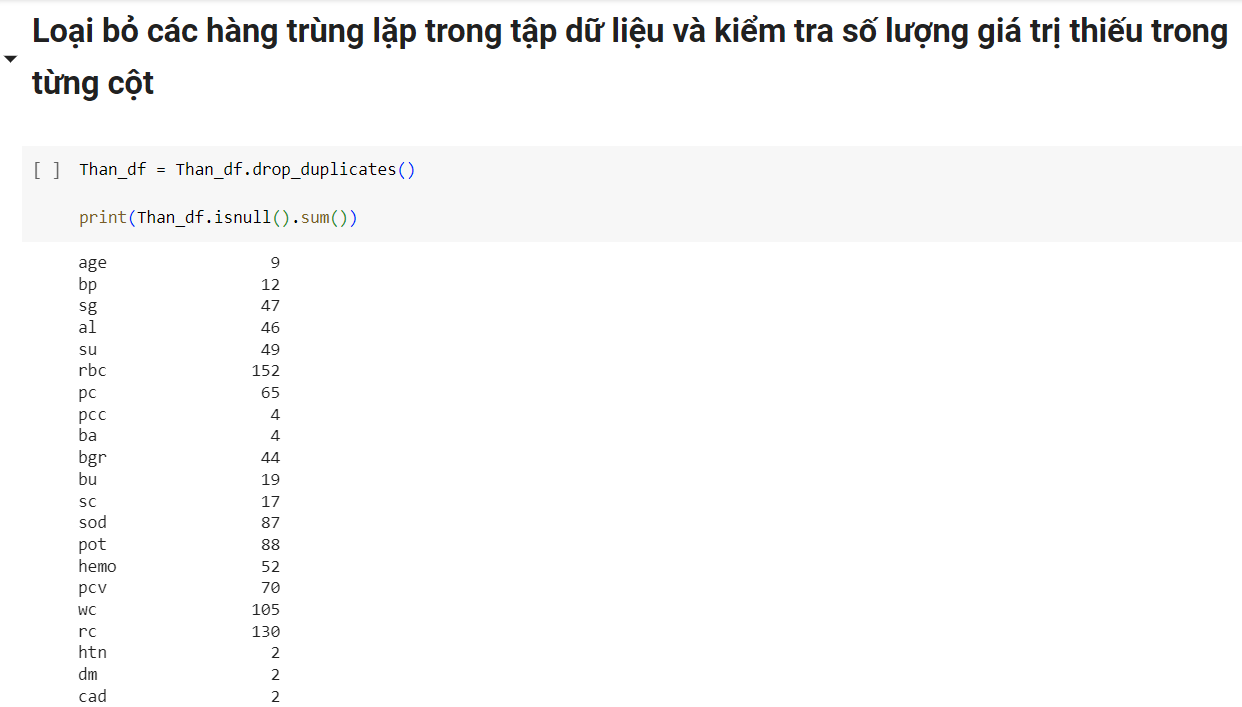
Bỏ cột ID và gán kiểu dữ liệu về 0 và 1

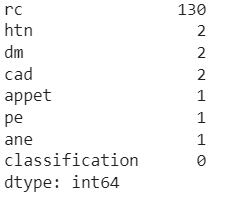


Dữ liệu sau khi gán



Loại bỏ các dữ liệu trùng và kiểm tra giá trị thiếu trong từng cột





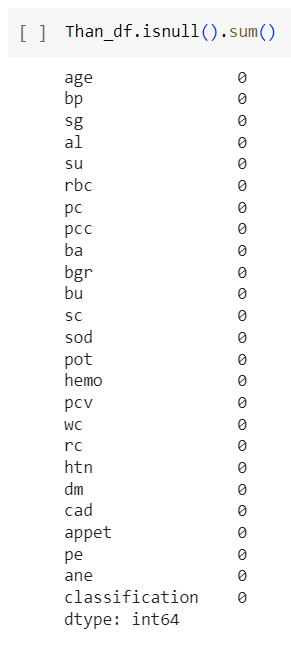
Chuyển đổi kiểu dữ liệu về dạng numeric



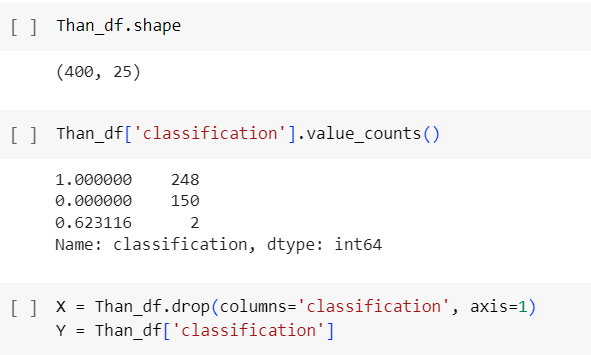
Điền các giá trị còn thiếu bằng giá trị trung bình



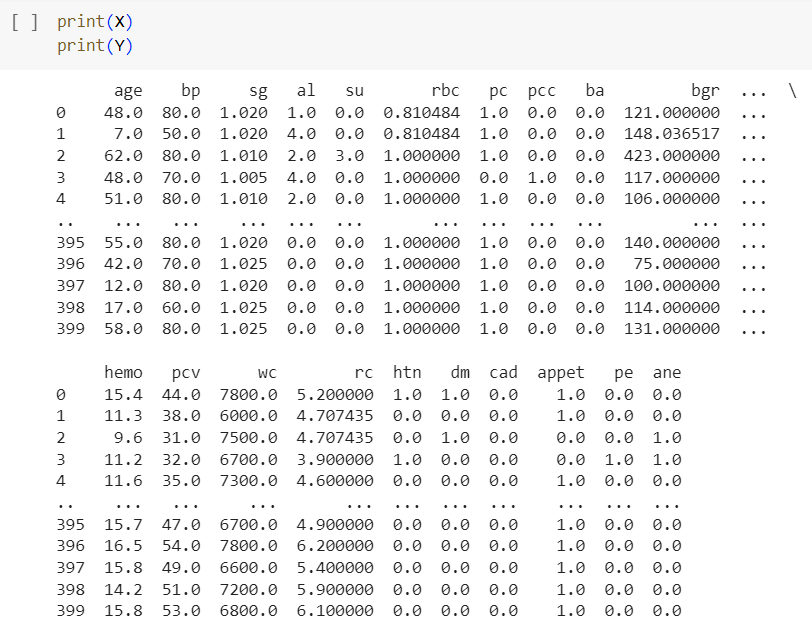
Kiểm tra xem còn giá trị null hay không

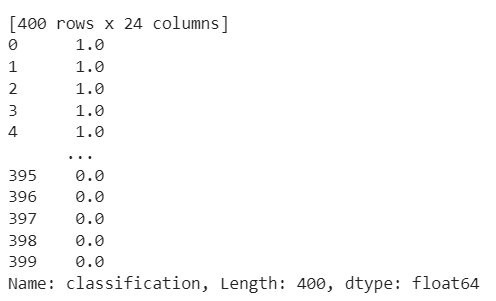


Xem thử bộ dữ liệu bao nhiêu cột và dòng, xem bộ dữ liệu thoe loại và loại bỏ classification và tạo biến mới X, Y

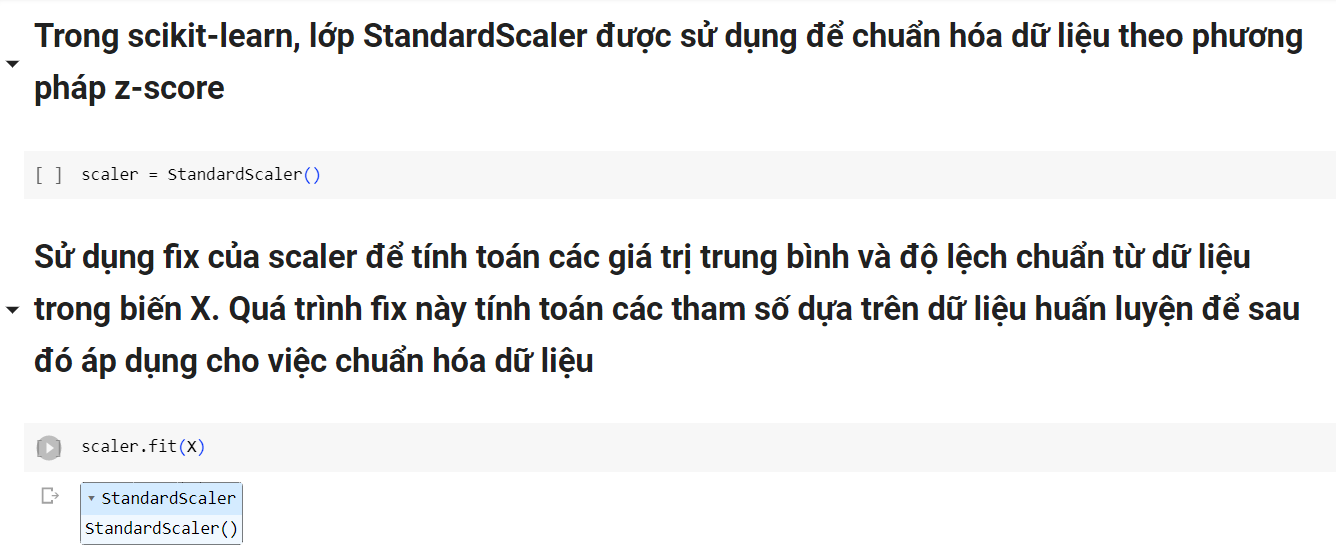


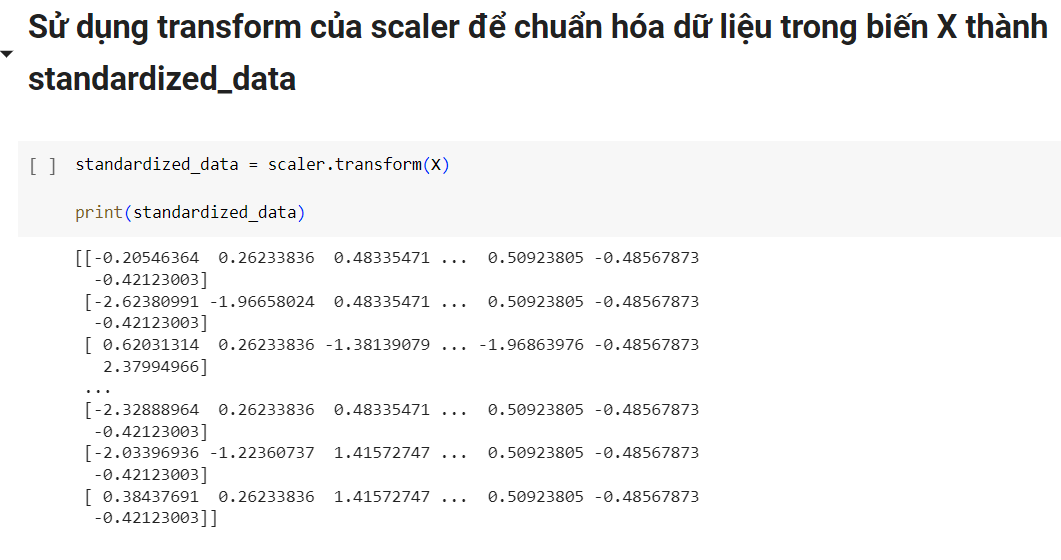
Xuất x và Y

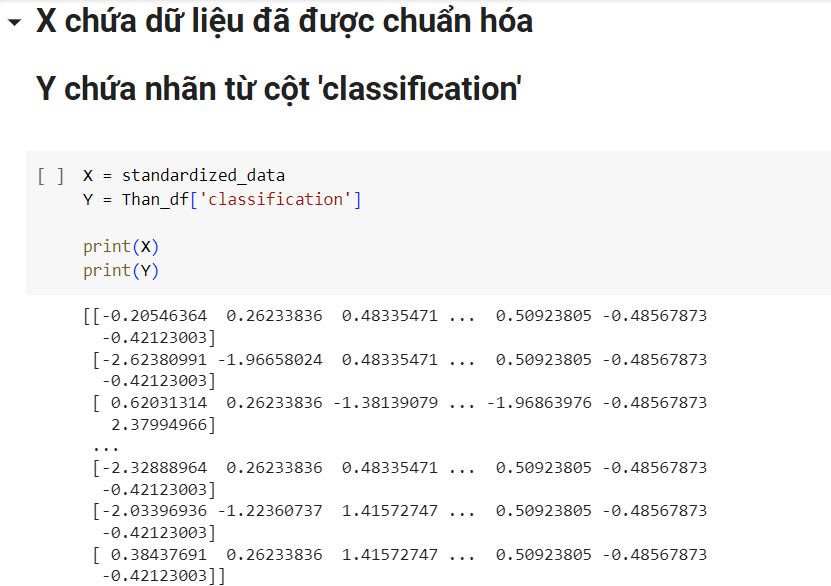


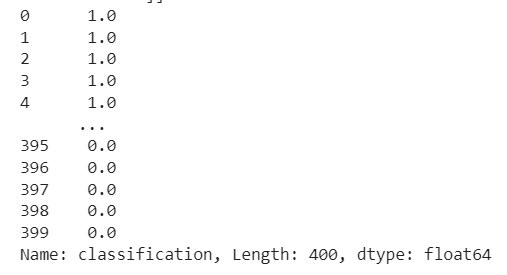


Chuẩn hóa dữ liệu







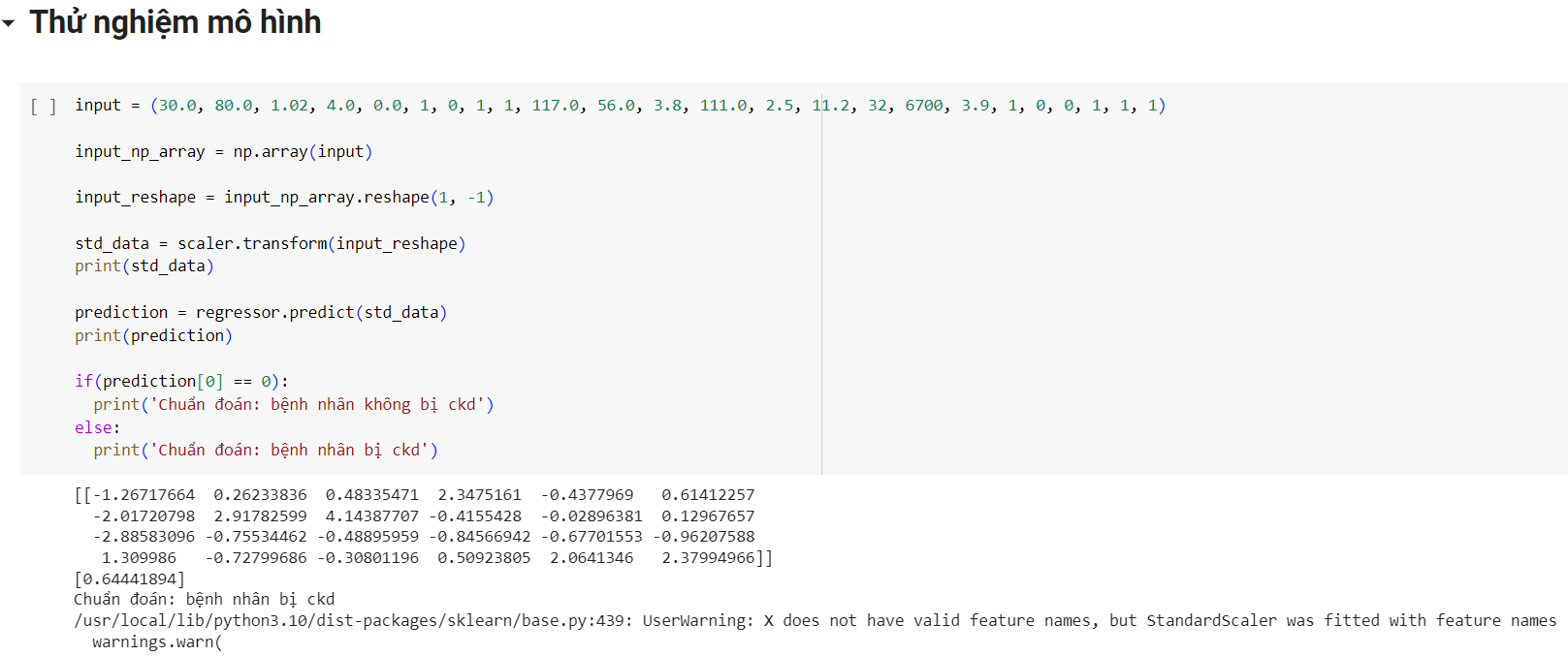


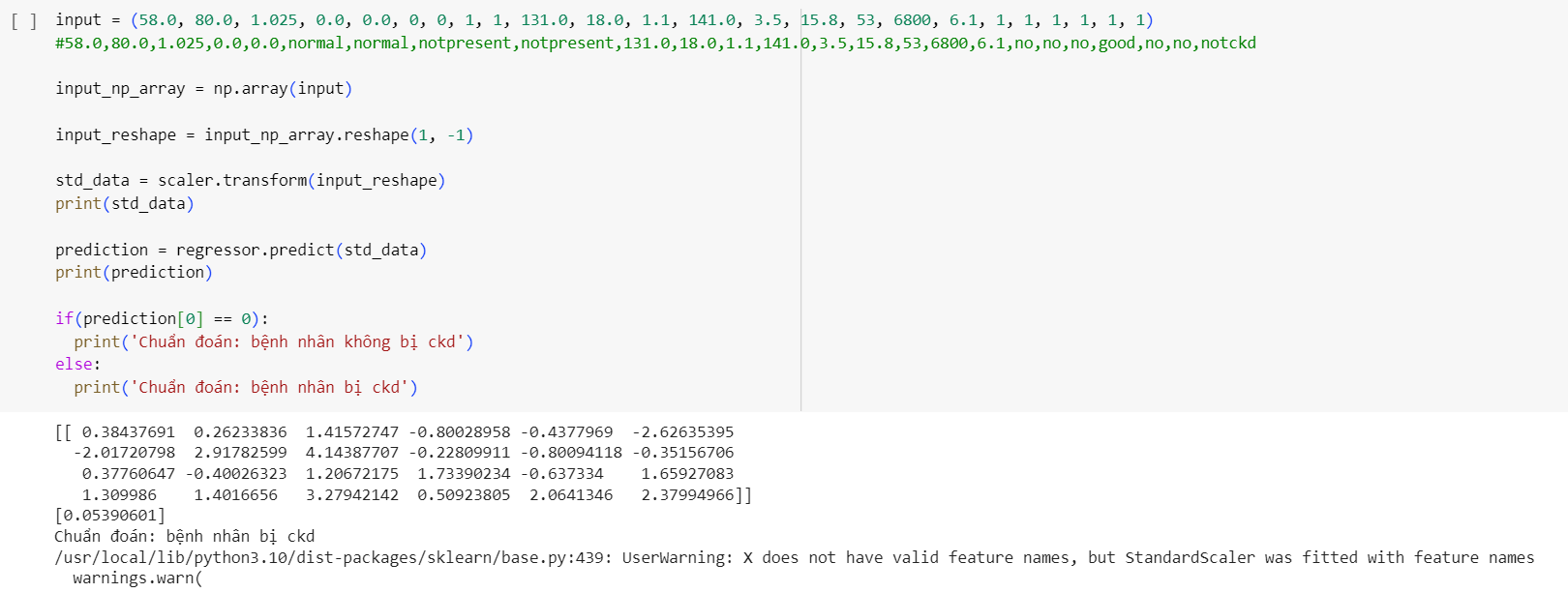
Chia bộ dữ liệu để huấn luyện





Thử nghiệm mô hình trên





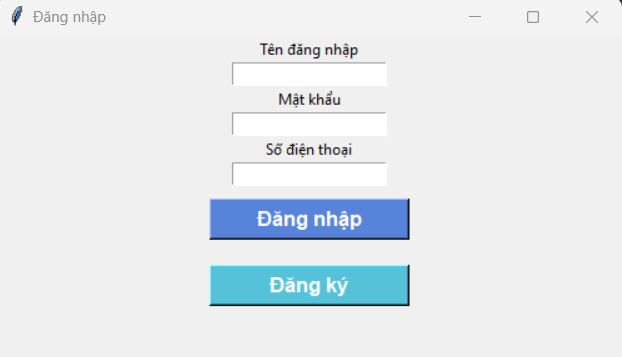
Lưu mô hình



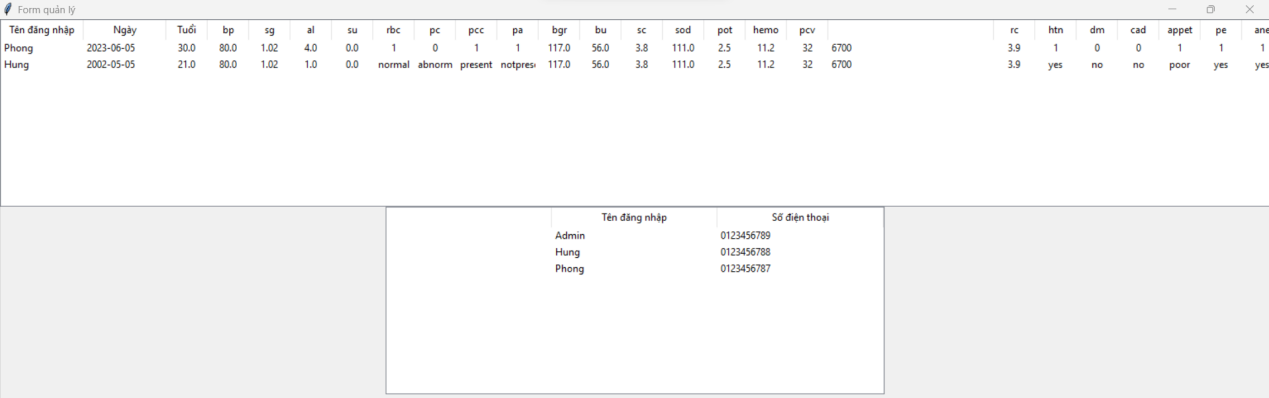
## Cài đặt ứng dụng

**Giao diện ứng dụng**

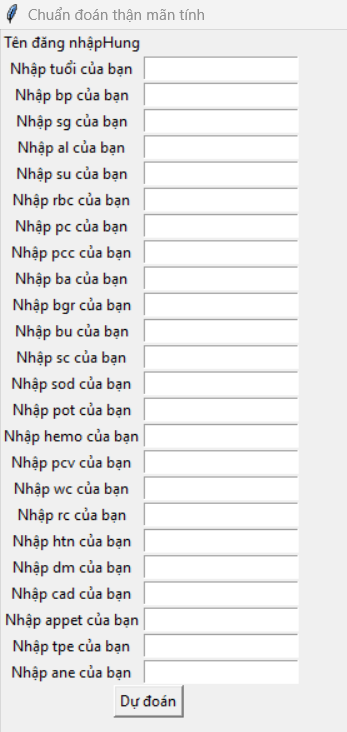
Phải chạy database trước khi chạy project



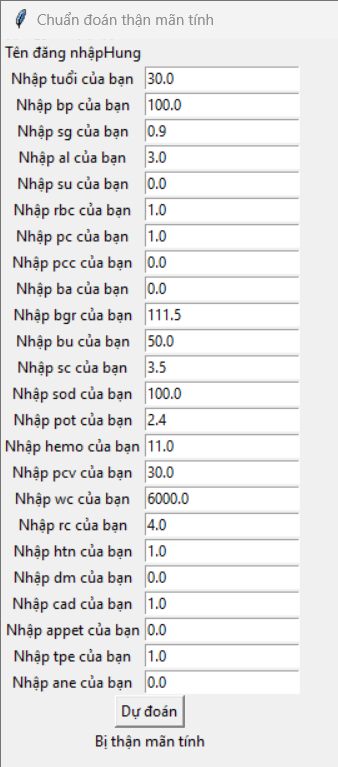
Khi đăng nhập với quyền admin



Khi đăng nhập với quyền bệnh nhân



Khi nhập liệu các chỉ số vào rồi bấm nút dự đoán nó sẽ xuất ra dự đoán có bị thận mãn tính hay không



# Kết luận

Trong đề tài này, chúng em đã xem xét việc sử dụng thuật toán SVM (Support Vector Machine) để chuẩn đoán thận mãn tính. SVM là một thuật toán học máy phổ biến được sử dụng rộng rãi trong lĩnh vực phân loại và chuẩn đoán. Chúng tôi đã trình bày quá trình xây dựng một mô hình SVM để dự đoán xác suất mắc thận mãn tính dựa trên các đặc trưng y tế của bệnh nhân.

Để thực hiện việc này, chúng tôi đã thu thập một tập dữ liệu gồm các bệnh nhân có và không có thận mãn tính, kèm theo các thông tin về các biến số y tế như tuổi, chỉ số chức năng thận và các chỉ số máu liên quan. Sau đó, chúng tôi đã tiến hành tiền xử lý dữ liệu, bao gồm việc chuẩn hóa các biến số và chọn lọc đặc trưng quan trọng.

Tiếp theo, chúng tôi đã xây dựng một mô hình SVM bằng cách huấn luyện trên tập dữ liệu đã được tiền xử lý. Chúng tôi đã sử dụng kỹ thuật tối ưu hóa thông qua tìm kiếm siêu tham số để tìm ra các thông số tối ưu cho mô hình SVM.

Kết quả đạt được từ mô hình SVM cho thấy hiệu suất đáng kể trong việc dự đoán thận mãn tính. Độ chính xác và độ nhạy của mô hình đều cao, cho thấy khả năng phân loại chính xác giữa các bệnh nhân có và không có thận mãn tính. Điều này có thể hỗ trợ các chuyên gia y tế trong việc đưa ra quyết định chẩn đoán và lựa chọn phương pháp điều trị phù hợp.

Tuy nhiên, cần lưu ý rằng mô hình SVM chỉ là một công cụ hỗ trợ và không thay thế cho quyết định của các chuyên gia y tế. Việc xác định thận mãn tính phụ thuộc vào nhiều yếu tố khác nhau, bao gồm kết quả xét nghiệm và triệu chứng lâm sàng. Do đó, mô hình SVM cần được sử dụng như một phần của quá trình chuẩn đoán toàn diện, kết hợp với thông tin từ các phương pháp khác để đưa ra quyết định chẩn đoán cuối cùng.

Tổng kết lại, việc sử dụng thuật toán SVM để chuẩn đoán thận mãn tính đã cho thấy tiềm năng trong việc hỗ trợ quyết định chẩn đoán và tăng cường khả năng dự đoán. Đây là một nghiên cứu đầy hứa hẹn và có thể đóng góp vào việc cải thiện việc phát hiện và điều trị thận mãn tính, từ đó nâng cao chất lượng chăm sóc sức khỏe cho bệnh nhân.

# Tài liệu tham khảo

1. C. Trung, "Giới thiệu về Support Vector Machine," 20 8 2020. [Online]. Available: <https://viblo.asia/p/gioi-thieu-ve-support-vector-machine-svm-6J3ZgPVElmB.>

***[Anna Malkina](https://www.msdmanuals.com/vi-vn/chuy%C3%AAn-gia/authors/malkina-anna)*** , MD, University of California, San Francisco “Bệnh thận mạn”, 10 2022.[Online]. Availabel: <https://www.msdmanuals.com/vi-vn/chuy%C3%AAn-gia/r%E1%BB%91i-lo%E1%BA%A1n-di-truy%E1%BB%81n/b%E1%BB%87nh-th%E1%BA%ADn-m%E1%BA%A1n/b%E1%BB%87nh-th%E1%BA%ADn-m%E1%BA%A1n>.