TRƯỜNG ĐẠI HỌC ĐIỆN LỰC

**KHOA CÔNG NGHỆ THÔNG TIN**



**BÁO CÁO CHUYÊN ĐỀ HỌC PHẦN**

**KHAI PHÁ DỮ LIỆU**

**ĐỀ TÀI:**

**KHAI PHÁ DỮ LIỆU SỬ DỤNG THUẬT TOÁN PHÂN LỚP NHỊ PHÂN DỰ ĐOÁN KHẢ NĂNG MẮC BỆNH UNG THƯ VÚ VÀ BỆNH TIM**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  | | --- | --- | | **Sinh viên thực hiện** | **: TRỊNH NHẬT HUY**  **NGUYỄN TUẤN ANH** | | **Giảng viên hướng dẫn** | **: NGUYỄN THỊ THANH TÂN** | | | **Ngành** | **: CÔNG NGHỆ THÔNG TIN** | | | **Lớp** | **: D13CNPM4** | | | **Khóa** | **: 2018-2023** | | |  |

***Hà Nội, tháng 6 năm 2021***

**PHIẾU CHẤM ĐIỂM**

Sinh viên thực hiện:

|  |  |  |
| --- | --- | --- |
| **Họ và tên** | **Chữ ký** | **Ghi chú** |
| TRỊNH NHẬT HUY |  | Bài toán dự đoán khả năng mắc bệnh ung thư vú |
| NGUYỄN TUẤN ANH |  | Bài toán dự đoán khả năng mắc bệnh tim |
| TRẦN HẢI LONG |  |  |

Giảng viên chấm:

|  |  |  |
| --- | --- | --- |
| **Họ và tên** | **Chữ ký** | **Ghi chú** |
| Giảng viên chấm 1 : |  |  |
| Giảng viên chấm 2 : |  |  |

**MỤC LỤC**

[CHƯƠNG 1: TỔNG QUAN VỀ KHAI PHÁ DỮ LIỆU](#_Toc67813584)

[1.1. Khái niệm 5](#_Toc67813585)

[1.2. Các chức năng chính 5](#_Toc67813586)

[1.3. Ứng dụng 6](#_Toc67813587)

[CHƯƠNG 2: PHÂN LỚP DỮ LIỆU VỚI MÔ HÌNH SVM 7](#_Toc67813588)

[2.1. Khái niệm phân lớp dữ liệu 7](#_Toc67813589)

[2.2 Mô hình SVM 10](#_Toc67813592)

[CHƯƠNG 3: ỨNG DỤNG PHƯƠNG PHÁP PHÂN LỚP NHỊ PHÂN DỰ ĐOÁN KHẢ NĂNG MẮC BỆNH ÚNG THƯ VÚ VÀ BỆNH TIM 12](#_Toc67813595)

[3.1. Giới thiệu bài toán 12](#_Toc67813596)

[3.1.1. Bài toán dự đoán khả năng mắc bệnh ung thư vú 12](#_Toc67813597)

[3.1.2. Bài toán dự đoán khả năng mắc bệnh tim 13](#_Toc67813599)

[3.2 Môi trường thử nghiệm 13](#_Toc67813600)

[3.2.1. Giới thiệu python 13](#_Toc67813601)

[3.2.2.Đặc điểm của Python: 14](#_Toc67813602)

[3.2.3.Cài đặt Python 14](#_Toc67813603)

[Cài đặt bộ thư viện : Numpy , Pip install pandas , Pip install sklearn 14](#_Toc67813604)

[3.3 Xây dựng bộ dữ liệu 15](#_Toc67813605)

[3.3.1 Bộ dữ liệu dự đoán khả năng mắc bệnh ung thư vú 15](#_Toc67813606)

[3.3.2 Bộ dữ liệu dự đoán khả năng mắc bệnh tim 20](#_Toc67813607)

[3.4 Đánh Giá 24](#_Toc67813608)

CHƯƠNG 4 : [KẾT LUẬN 25](#_Toc67813609)

**LỜI NÓI ĐẦU**

Thế giới đang chứng kiến sự thay đổi nhanh chóng trong cuộc cách mạng công nghiệp 4.0. Nếu chúng ta nắm bắt được lợi thế và có sự chuẩn bị chu đáo trước sự thay đổi này thì sự phát triển của mỗi doanh nghiệp, mỗi nền kinh tế sẽ có lợi thế bứt phá lớn. Ngược lại, nếu các doanh nghiệp chuẩn bị chưa chuẩn bị tốt sẽ dẫn đến sự tụt hậu, ảnh hưởng đến sự phát triển chung của nền kinh tế quốc gia. Như chúng em đã đề cập ở trên, cuộc cách mạng công nghiệp 4.0 là cơ hội lớn giúp Việt Nam tiếp cận thành tựu công nghệ mới, tạo sự bứt phá cho nền kinh tế - xã hội cũng như hướng tới sự phát triển bền vững.

Mặc dù đã rất cố gắng hoàn thiện báo cáo với tất cả sự nỗ lực, tuy nhiên do bước đầu tìm hiểu và xây dựng báo cáo môn học trong thời gian có hạn, với lượng kiến thức còn hạn chế, nhiều bỡ ngỡ, nên báo cáo “Khai phá dữ liệu sử dụng thuật toán phân lớp nhị phân dự đoán khả năng mắc bệnh ung thư vú và bệnh tim” chắc chắn sẽ không thể tránh khỏi những thiếu sót. Chúng em rất mong nhận được sự quan tâm, thông cảm và những đóng góp quý báu của các thầy cô và các bạn để báo cáo này ngày càng hoàn thiện hơn.

Tuy vậy chúng em rất mong nhận được sự ủng hộ của quý thầy, cô giáo để chúng em đạt được kết quả tốt nhất. Chúng em xin chân thành cảm ơn thầy giáo Vũ Văn Định đã hướng dẫn, giúp đỡ chúng em trong quá trình học tập. Thầy đã giúp trang bị cho chúng em kiến thức môn học và hơn cả là động lực để tiếp tục trên con đường học tập.

Chúng em xin chân thành cảm ơn!

# CHƯƠNG 1: TỔNG QUAN VỀ KHAI PHÁ DỮ LIỆU

## 1.1. Khái niệm

**Khai phá dữ liệu** (Data Mining) Là quá trình tính toán để tìm ra các mẫu trong các bộ dữ liệu lớn liên quan đến các phương pháp tại giao điểm của máy học, thống kê và các hệ thống cơ sở dữ liệu. Đây là một lĩnh vực liên ngành của khoa học máy tính. Mục tiêu tổng thể của quá trình khai thác dữ liệu là trích xuất thông tin từ một bộ dữ liệu và chuyển nó thành một cấu trúc dễ hiểu để sử dụng tiếp. Ngoài bước phân tích thô, nó còn liên quan tới cơ sở dữ liệu và các khía cạnh quản lý dữ liệu, xử lý dữ liệu trước, suy xét mô hình và suy luận thống kê, các thước đo thú vị, các cân nhắc phức tạp, xuất kết quả về các cấu trúc được phát hiện, hiện hình hóa và cập nhật trực tuyến. Khai thác dữ liệu là bước phân tích của quá trình "khám phá kiến thức trong cơ sở dữ liệu".

## 1.2. Các chức năng chính

Data Mining được chia nhỏ thành một số hướng chính như sau:

* Mô tả khái niệm (concept description): thiên về mô tả, tổng hợp và tóm tắt khái niệm. Ví dụ: tóm tắt văn bản.
* Luật kết hợp (association rules): là dạng luật biểu diễn tri thứ ở dạng khá đơn giản. Ví dụ: “60 % nam giới vào siêu thị nếu mua bia thì có tới 80% trong số họ sẽ mua thêm thịt bò khô”. Luật kết hợp được ứng dụng nhiều trong lĩnh vực kính doanh, y học, tin-sinh, tài chính & thị trường chứng khoán...
* Phân lớp và dự đoán (classification & prediction): xếp một đối tượng vào một trong những lớp đã biết trước. Ví dụ: phân lớp vùng địa lý theo dữ liệu thời tiết. Hướng tiếp cận này thường sử dụng một số kỹ thuật của machine learning như cây quyết định (decision tree), mạng nơ ron nhân tạo (neural network) ... Người ta còn gọi phân lớp là học có giám sát (học có thầy).
* Phân cụm (clustering): xếp các đối tượng theo từng cụm (số lượng cũng như tên của cụm chưa được biết trước. Người ta còn gọi phân cụm là học không giám sát (học không thầy).
* Khai phá chuỗi (sequential/temporal patterns): tương tự như khai phá luật kết hợp nhưng có thêm tính thứ tự và tính thời gian. Hướng tiếp cận này được ứng dụng nhiều trong lĩnh vực tài chính và thị trường chứng khoán vì nó có tính dự báo cao.

## 1.3. Ứng dụng

Khai phá dữ liệu liên quan đến nhiều ngành nghề lĩnh vực thế nên nó có rất nhiều ứng dụng :

* Kinh tế - ứng dụng trong kinh doanh, tài chính, tiếp thị bán hàng, bảo hiểm, thương mại, ngân hàng... Đưa ra các bản báo cáo giàu thông tin, phân tích rủi ro trước khi đưa ra các chiến lược kinh doanh, sản xuất, phân loại khách hàng từ đó phân định ra thị trường.
* Khoa học: Thiên văn học - dự đoán đường đi các thiên thể, hành tinh...; Công nghệ sinh học – tìm ra các gen mới, cây con giống mới…
* Web: các công cụ tìm kiếm.
* Công nghệ sinh học và dược phẩm: Phân tích các dữ liệu di truyền
* Nhân sự: Chọn ứng cử viên khi tuyển dụng
* Bào hiểm tài chính và thị trường chứng khoán: phân tích tình hình tài chính và dự báo giá của các loại cổ phiếu trong thị trường chứng khoán.Danh mục vốn và giá , lãi xuất dữ liệu thẻ tín dụng,phát hiện gian lận...

# CHƯƠNG 2: PHÂN LỚP DỮ LIỆU VỚI MÔ HÌNH SVM

## 2.1. Khái niệm phân lớp dữ liệu

# Phân lớp là một hình thức học được giám sát tức là : tập dữ liệu huấn luyện (quan sát , thẩm định ) đi đôi với những nhãn chỉ định lớp quan sát , những dữ liệu mới được phân lớp dựa trên tập huấn luyện .

# Kỹ thuật phân lớp được tiến hành gồm 2 bước : Xây dựng mô hình và sử dụng mô hình.

* Xây dựng mô hình : là mô tả một tập những lớp được định nghĩa trước trong đó : mỗi bộ hoặc mẫu được gán thuộc về một lớp được định nghĩa trước như là được xát định bởi thuộc tính nhãn lớp , tập hợp của những bộ được sử dụng trong việc sử dụng mô hình được gọi là tập huấn luyện . Mô hình được biểu diễn là những luật phân lớp , cây quyết định và những công thức toán học .
* Sử dụng mô hình : Việc sử dụng mô hình phục vụ cho mục đích phân lớp dữ liệu trong tương lai hoặc phân lớp cho những đối tượng chưa biết đến . Trước khi sử dụng mô hình người ta thường phải đánh giá tính chính xác của mô hình trong đó : nhãn được biết của mẫu kiểm tra được so sánh với kết quả phân lớp của mô hình , độ chính xác là phần trăm của tập hợp mẫu kiểm tra mà phân loại đúng bởi mô hình , tập kiểm tra là độc lập với tập huấn luyện .

Các thuật toán phân lớp dữ liệu

* Phân lớp với cây quyết định (decision tree)
* Phân lớp với Naïve Bayesian
* Phân lớp với k phần tử gần nhất (k-nearest neighbor)
* Phân lớp với máy vector hỗ trợ (SVM)
* Phân lớp với mạng neural (neural network)
* Phân lớp dựa trên tiến hoá gen (genetic algorithms)
* Phân lớp với lý thuyết tập thô, tập mờ (rough sets)
* Phân lớp với lý thuyết tập mờ (fuzzy sets)

**Phân lớp nhị phân**(*Binary classification*) là nhiệm vụ phận loại các[*phần tử*](https://vi.wikipedia.org/w/index.php?title=Ph%E1%BA%A7n_t%E1%BB%AD&action=edit&redlink=1) của một [*tập hợp*](https://vi.wikipedia.org/wiki/T%E1%BA%ADp_h%E1%BB%A3p) các đối tượng ra thành 2 nhóm dựa trên cơ sở là chúng có một thuộc tính nào đó hay không (hay còn gọi là tiêu chí). Một số nhiệm vụ phân loại nhị phân điển hình:

* Kiểm tra y khoa xem một bệnh nhân có bệnh nào đó hay không (thuộc tính để phân loại là căn bệnh đó)
* Quản lý chất lượng trong nhà máy, ví dụ: xác định xem một sản phẩm làm ra là đủ tốt để bán chưa, hay nên loại bỏ nó (thuộc tính để phân loại là tính đủ tốt)
* Xác định xem một trang hay một bài báo có nên nằm trong tập kết quả của một truy vấn hay không (thuộc tính là độ liên quan của bài báo - thường là sự hiện diện của một số từ nào đó trong bài báo đó)

Đánh giá bộ phân lớp nhị phân: Để đánh giá độ hiệu quả của một xét nghiệm y khoa, người ta thường sử dụng các khái niệm [*độ nhạy*](https://vi.wikipedia.org/wiki/%C4%90%E1%BB%99_nh%E1%BA%A1y) và [*đặc trưng*](https://vi.wikipedia.org/w/index.php?title=%C4%90%E1%BA%B7c_tr%C6%B0ng&action=edit&redlink=1). Giả sử chúng ta xét nghiệm xem một vài người nào đó có bệnh hay không.

* Một số người có bệnh, và kết quả xét nghiệm là dương tính (*positive*). Họ được gọi là các *dương tính đúng*.
* Một số người có bệnh, nhưng kết quả xét nghiệm âm tính (*negative*). Họ được gọi là các *âm tính sai*.
* Một số không có bệnh, và kết quả xét nghiệm cũng là âm tính. Họ được gọi là các *âm tính đúng*.
* Một số không có bệnh, nhưng kết quả xét nghiệm lại là dương tính. Họ được gọi là các *dương tính sai*.

Tổng số người *dương tính đúng*, *âm tính đúng*, *dương tính sai*, *âm tính sai* chiếm 100% tổng số người được xét nghiệm.

* [**Độ nhạy**](https://vi.wikipedia.org/wiki/%C4%90%E1%BB%99_nh%E1%BA%A1y) (*sensitivity*) là tỉ lệ của số người bị bệnh được xác định đúng là có bệnh trên tổng số người bị bệnh, nghĩa là *(dương tính đúng)/(dương tính đúng + âm tính sai)*. Nó có thể được coi là "xác suất xét nghiệm cho kết quả dương tính khi người được xét nghiệm có bị bệnh". Độ nhạy càng cao, càng ít khả năng bệnh không được phát hiện (hoặc, trong trường hợp quản lý chất lượng ở nhà máy, càng ít sản phẩm lỗi được đưa ra thị trường).
* [**Đặc trưng**](https://vi.wikipedia.org/w/index.php?title=%C4%90%E1%BA%B7c_tr%C6%B0ng&action=edit&redlink=1) (*specificity*) là tỉ lệ của số người không bị bệnh có kết quả xét nghiệm âm tính trên tổng số người không có bệnh (thực), nghĩa là *(âm tính đúng)/(âm tính đúng + dương tính sai)*. Nó còn được coi là *xác suất xét nghiệm cho kết quả âm tính đối với người không có bệnh*. Độ đặc trưng càng cao, càng ít người mạnh khỏe được coi là bị bệnh (hoặc trong trường hợp nhà máy, càng ít tiền bị tốn phí do loại bỏ các sản phẩm chất lượng tốt thay vì đem bán chúng).
* Về mặt lý thuyết, *độ nhạy* và *đặc trưng* là độc lập, tức là cả hai đều có thể đạt đến 100%. Trong thực tế, chúng ta phải đánh đổi cái này để được cái kia - cái này tốt lên thì cái kia xấu đi, không thể đạt được cả hai.
* Một điểm cần chú ý nữa, là *độ nhạy* và *đặc trưng* là độc lập với tỉ lệ giữa số cá thể *âm tính* và số cá thể *dương tính*. Tuy nhiên, giá trị của chúng thì lại phụ thuộc vào tổng số cá thể kiểm tra (population). Ví dụ: kiểm tra có kết quả: *độ nhạy* 99%, *đặc trưng* 99%.
* Giả sử số người kiểm tra là 2000 người, trong đó 1000 có bệnh và 1000 khỏe mạnh. Như vậy, ta phát hiện đúng 990 người *dương tính đúng*, 990 người *âm tính đúng*, và 10 *âm tính sai*, 10 *dương tính sai*. Cuối cùng, tỉ lệ dự đoán trúng là 99% cho cả kết quả *dương tính* và *âm tính*. Như vậy, hệ thống này được coi là khá đáng tin cậy.
* Giả sử số người kiểm tra là 2000 người, trong đó chỉ có 100 là thực sự bị bệnh. Giả sử ta có 99 *dương tính đúng*, 1 *âm tính sai*, 1881 *âm tính đúng*, và 19 *dương tính sai*. Trong số 19+99 người xét nghiệm dương tính, chỉ có 99 người thực sự có bệnh.

**2.2 Mô hình SVM**

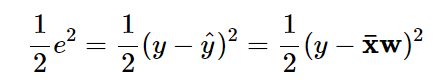
**SVM** (Support Vector Machine) là 1 thuật toán học máy thuộc nhóm Supervised Learning (học có giám sát) được sử dụng trong các bài toán phân lớp dữ liệu (classification) hay hồi qui (Regression).**SVM** là 1 thuật toán phân loại nhị phân, SVM nhận dữ liệu vào và phân loại chúng vào hai lớp khác nhau. Với 1 bộ các ví dụ luyện tập thuộc hai thể loại cho trước, thuật toán luyện tập SVM xây dựng 1 mô hình SVM để phân loại các ví dụ khác vào hai thể loại đó.

* Phương trình tuyến tính của SVM có dạng:



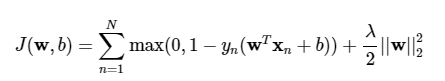
Trong đó w thuộc Rn là vector hệ số ứng với các chiều của vector  b là hệ số tự do trong không gian hai chiều thì được gọi là đường thẳng,không gian 3 chiều là mặt phẳng.

* Công thức của sai số dự đoán:



Trong đó, e là sai số dự đoán, y là giá trị thực và là giá trị dự đoán (hay còn gọi là y\_pred). Hàm bình phương để tránh phương trình có thể ra kết quả âm và vì e là sai số, nên giá trị này càng nhỏ càng tốt.

* Hàm mất mát

****

* Hàm đánh giá:
* Accuracy : (ACC)

Cách đơn giản và hay được sử dụng nhất là accuracy (độ chính xác). Cách đánh giá này đơn giản tính tỉ lệ giữa số điểm được dự đoán đúng và tổng số điểm trong tập dữ liệu kiểm thử: 

Trong đó TP, TN là dự đoán đúng

FP, FN là dự đoán sai

### Precision - bao nhiêu cái đúng được lấy ra

### Recall - bao nhiêu cái được lấy ra là đúng

# CHƯƠNG 3: ỨNG DỤNG PHƯƠNG PHÁP PHÂN LỚP NHỊ PHÂN DỰ ĐOÁN KHẢ NĂNG MẮC BỆNH ÚNG THƯ VÚ VÀ BỆNH TIM

## 3.1. Giới thiệu bài toán

### 3.1.1. Bài toán dự đoán khả năng mắc bệnh ung thư vú

# 3.1.1.1. Mô tả

Trên thế giới, ung thư vú là loại ung thư phổ biến nhất ở phụ nữ và cao thứ hai về tỷ lệ tử vong. Tỷ lệ tử vong thay đổi nhiều từ 25 – 35/100.000 dân tại Anh, Đan Mạch, Hà Lan, Hoa Kỳ và Canada đến 2 – 5/100.000 dân ở Nhật, Mehico, Venezuela.Tỷ lệ mắc bệnh hàng năm tăng cao ở các nước Bắc Mỹ và châu Âu, trong khi ở các nước châu Á và châu Phi có xu hướng thấp hớn. Ở Hoa Kỳ trong năm 1993, đã có 182.000 ca ung thư vú mới mắc ở phụ nữ. Tỷ lệ hằng năm theo ước tính 100,2 ca mới / 100.000 dân. Từ năm 1973, tỷ lệ đã tăng trung bình hằng năm 1,8%.Theo báo cáo của Tổ chức Y tế thế giới, ung thư vú đứng hàng thứ 2 tại các nước Đông Nam Á (sau ung thư phổi) và chiếm 20% trong tổng số các ung thư ở phụ nữ.

Bài toán Chẩn đoán ung thư vú được thực hiện khi phát hiện một khối u bất thường (từ việc tự kiểm tra hoặc chụp X-quang) hoặc một đốm nhỏ canxi được nhìn thấy (trên phim chụp X-quang). Sau khi phát hiện ra một khối u đáng ngờ, bác sĩ sẽ tiến hành chẩn đoán để xác định xem nó có phải là ung thư hay không

- Giá trị input: Thông tin, đặc tính của một người có có khả năng mắc bệnh hoặc không.

- Giá trị ouput: Kết quả người được chuẩn đoán có bị mắc bệnh ung thư vú hay không.

**3.1.1.2.** Yêu cầu bài toán

- Lấy dữ liệu về thông tin, chỉ số của người bệnh.

- Trích chọn đặc trưng từ tập dữ liệu lấy được.

- Huấn luyện tập dữ liệu.

- Chuẩn đoán khả năng mắc bệnh của người bệnh.

### 3.1.2. Bài toán dự đoán khả năng mắc bệnh tim

**3.1.2.1.** Mô tả

Tổ chức Y tế Thế giới đã ước tính có 12 triệu ca tử vong trên toàn thế giới, hàng năm do các bệnh về tim mạch. Một nửa số ca tử vong ở Hoa Kỳ và các nước phát triển khác là do bệnh tim mạch. Tiên lượng sớm các bệnh tim mạch có thể giúp đưa ra quyết định thay đổi lối sống ở những bệnh nhân có nguy cơ cao và do đó làm giảm các biến chứng. Nghiên cứu này dự định xác định chính xác các yếu tố nguy cơ / liên quan nhất của bệnh tim cũng như dự đoán nguy cơ tổng thể. Mục tiêu phân loại là để dự đoán liệu bệnh nhân có nguy cơ mắc bệnh mạch vành trong tương lai (CHD) 10 năm hay không. Bộ dữ liệu cung cấp thông tin của bệnh nhân. Nó bao gồm hơn 4.000 bản ghi và 15 thuộc tính.

* Input: Thông tin, số liệu sức khỏe của người chuẩn đoán.
* Output: Người chuẩn đoán có mắc bệnh tim hay không

**3.1.2.2.** Yêu cầu bài toán

- Lấy dữ liệu về thông tin, chỉ số của người bệnh.

- Trích chọn đặc trưng từ tập dữ liệu lấy được.

- Huấn luyện tập dữ liệu.

- Chuẩn đoán khả năng mắc bệnh của người bệnh.

## 3.2 Môi trường thử nghiệm

### **3.2.1.** Giới thiệu python

Python là ngôn ngữ lập trình được sử dụng rất phổ biến ngày nay để phát triển nhiều loại ứng dụng phần mềm khác nhau như các chương trình chạy trên desktop, server, lập trình các ứng dụng web... Ngoài ra Python cũng là ngôn ngữ ưa thích trong ngành khoa học về dữ liệu (data science) cũng như là ngôn ngữ phổ biến để xây dựng các chương trình trí tuệ nhân tạo trong đó bao gồm machine learning.

### **3.2.2**. Đặc điểm của Python:

* Python là ngôn ngữ dễ học: Ngôn ngữ Python có cú pháp đơn giản, rõ ràng, sử dụng một số lượng không nhiều các từ khoá, do đó Python được đánh giá là một ngôn ngữ lập trình thân thiện với người mới học.
* Python là ngôn ngữ dễ hiểu: Mã lệnh (source code hay đơn giản là code) viết bằng ngôn ngữ Python dễ đọc và dễ hiểu. Ngay cả trường hợp bạn chưa biết gì về Python bạn cũng có thể suy đoán được ý nghĩa của từng dòng lệnh trong source code.
* Python có tương thích cao (highly portable): Chương trình phần mềm viết bằng ngôn ngữ Python có thể được chạy trên nhiều nền tảng hệ điều hành khác nhau bao gồm Windows, Mac OSX và Linux.

### **3.2.3**. Cài đặt Python

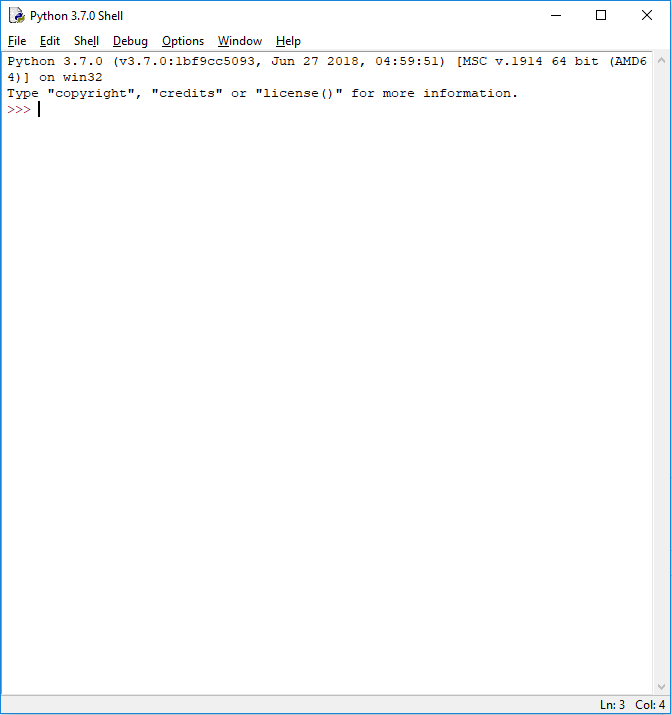
Cài đặt bộ thư viện : Numpy , Pip install pandas , Pip install sklearn

* + Cài đặt ngôn ngữ lập trình Python: Python 3.7.9

*(*[*https://www.python.org/downloads/windows/*](https://www.python.org/downloads/windows/)*)*

* + Môi trường lập trình Python: Python

*(https://www.python.com/)*



*Hình : Cài đặt Python*

## 3.3 Xây dựng bộ dữ liệu

### ***3.3.1*** *Bộ dữ liệu dự đoán khả năng mắc bệnh ung thư vú*

- Tập dữ liệu gồm 570 bệnh nhân với các chỉ số khối u bất thường khác nhau, , từ đó làm căn cứ chuẩn đoán bệnh nhân có bị mắc bệnh ung thư hay không.

- Đặt Y là khả năng mắc bệnh ung thư vú, với Y=0 là không mắc bệnh Y=1 là mắc bệnh.

- Bộ dữ liệu gồm 5 thuộc tính bao gồm:

+ Mean\_radius (Bán Kính Trung Bình)

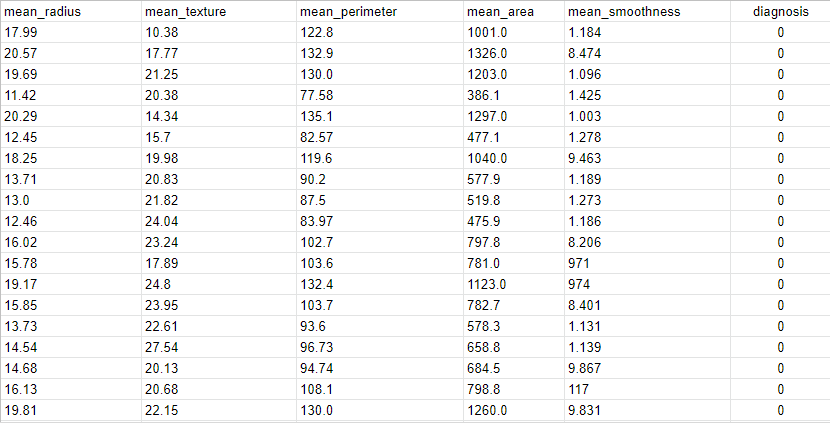
+ Mean\_texture (Kết cấu Trung Bình)

+ Mean\_perimeter (Chu vi Trung bình)

+ Mean\_area (Diện tích Trung bình)

+ Mean\_smoothness (Sai số Trung bình)

- Bộ dữ liệu chia thành 2 phần: Trong đó 80% dữ liệu được làm bộ dữ liệu huyến luyện mô hình ( tập train ), 20% còn lại làm bộ dữ liệu thử nghiệm độ chính xác của mô hình ( tập test ).

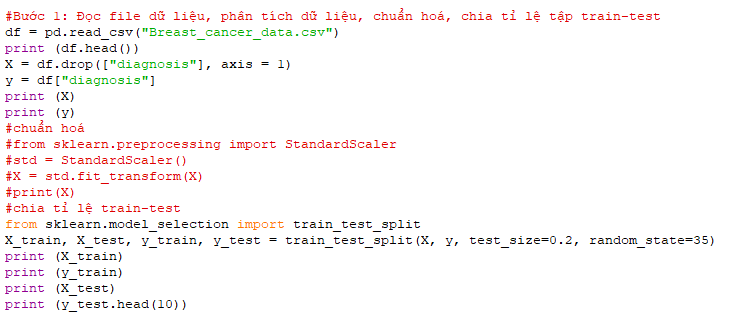
 *Hình :Dữ liệu dự đoán khả năng mắc bệnh ung thư vú của người bệnh*

**3.3.1.1.** Cài đặt thuật toán

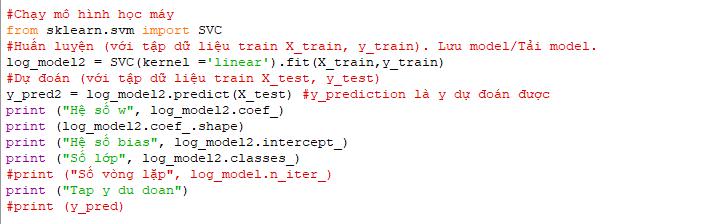
* Khai báo thư viện



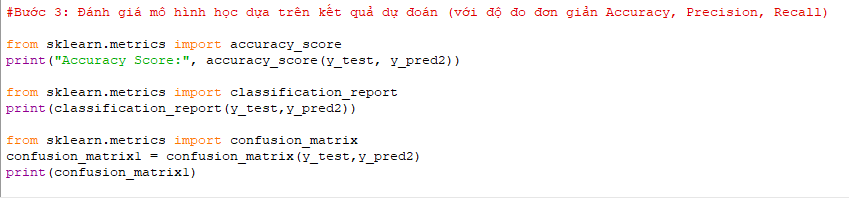
* Đọc file dữ liệu, phân tích dữ liệu, chuẩn hoá, chia tỉ lệ tập train-test



* Chạy mô hình học máy

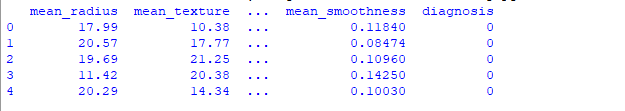


* Đánh giá mô hình dựa trên kết quả dự đoán (với độ đo đơn giản Accuracy, Precision, Recall)

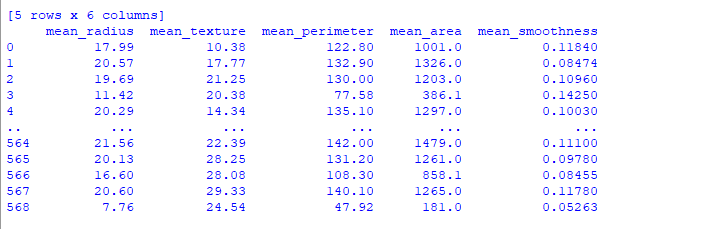


**3.3.1.2.** Kết quả

* In dữ liệu



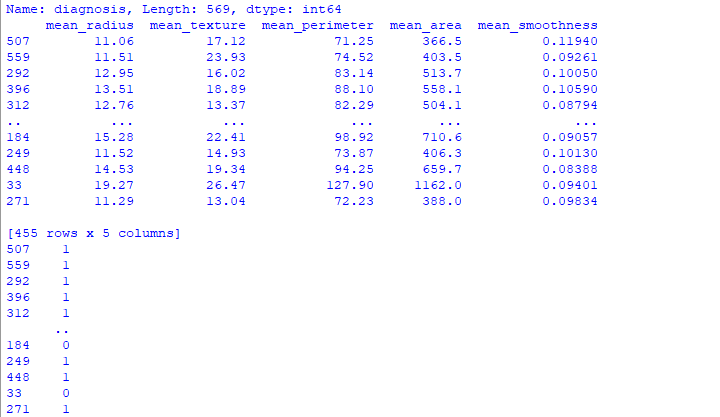
* + Dữ liệu đầu vào X:



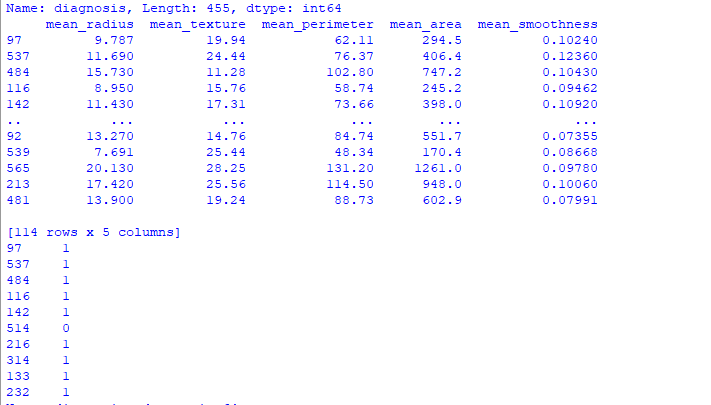
* + Dữ liệu đầu ra Y:



* X\_train,Y\_train,X\_test,Y\_test
  + X\_train,Y\_train



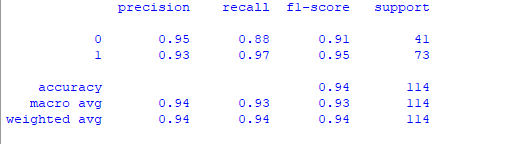
* + X\_test,Y\_test



* Đánh giá kết quả:
  + Sau khi thử nghiệm với bộ cơ sở dữ liệu thì kết quả phân lớp đạt 94%



* + Đối với Precision, Recall:



* + Ma trận dự đoán:



### **3.3.2** Bộ dữ liệu dự đoán khả năng mắc bệnh tim

Tập dữ liệu gồm 4239 bệnh nhân với các chỉ số mắc bệnh khác nhau, , từ đó làm căn cứ chuẩn đoán bệnh nhân có nguy cơ mắc bệnh cao, bệnh nhân có nguy cơ mắc bệnh thấp. Bệnh viện sẽ tổng hợp toàn bộ dữ liệu và phân loại bệnh nhân theo 2 trường hợp này.

- Đặt Y là khả năng mắc bệnh của bệnh nhân, với Y=0 là không mắc bệnh, ngược lại Y=1 là mắc bệnh.

- Bộ dữ liệu gồm 15 thuộc tính:

• Male (Giới tính): nam hoặc nữ

• Age (Tuổi): Tuổi của bệnh nhân

• Current Smoker (Người hút thuốc hiện tại)

• Cigs Per Day: số điếu thuốc mà người đó hút trung bình trong một ngày

• BP Meds: bệnh nhân có đang dùng thuốc huyết áp hay không

• Prevalent Stroke (Đột quỵ thịnh hành): bệnh nhân có bị đột quỵ trước đó hay không

• Prevalent Hyp (Hyp phổ biến): bệnh nhân có bị tăng huyết áp hay không

• Diabetes (Tiểu đường): bệnh nhân có bị tiểu đường hay không

• Tốt Chol: mức cholesterol toàn phần

• Sys HA: huyết áp tâm thu

• Dia BP: huyết áp tâm trương

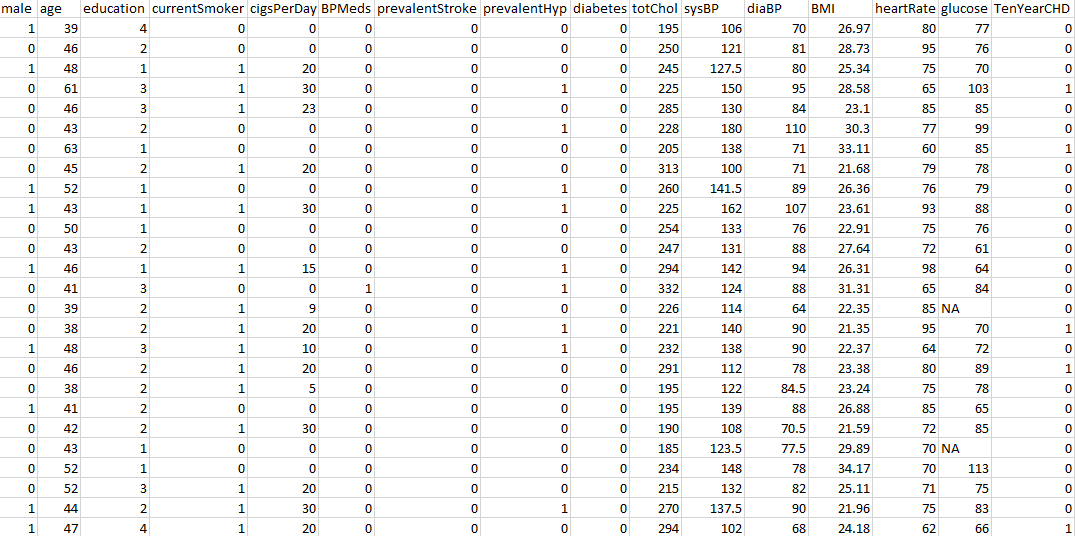
• BMI: Chỉ số khối cơ thể

• Heart Rate (Nhịp tim): nhịp

• Glucose: mức glucose

• Ten Year CHD : nguy cơ 10 năm của vành CHD bệnh tim (nhị phân: “1”, có nghĩa là “Có”, “0” có nghĩa là “Không”)

- Bộ dữ liệu chia thành 2 phần: Trong đó 80% dữ liệu được làm bộ dữ liệu huyến luyện mô hình ( tập train ), 20% còn lại làm bộ dữ liệu thử nghiệm độ chính xác của mô hình ( tập test )



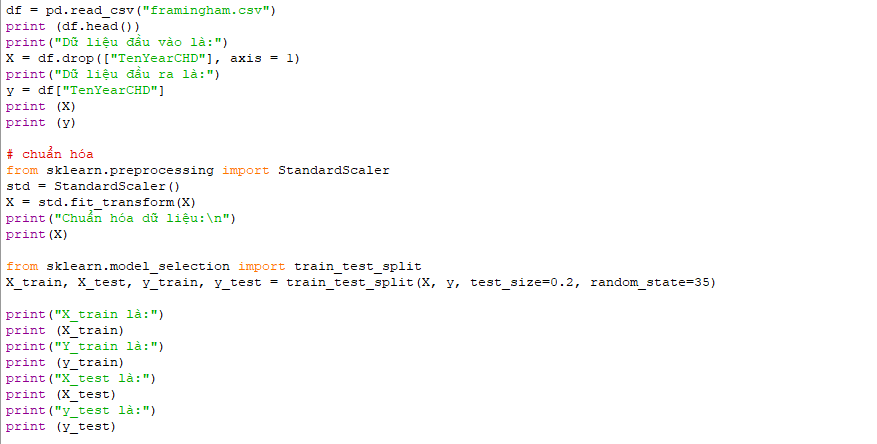
*Hình : Dữ liệu dự đoán khả năng mắc bệnh tim*

**3.3.2.1** Cài đặt thuật toán

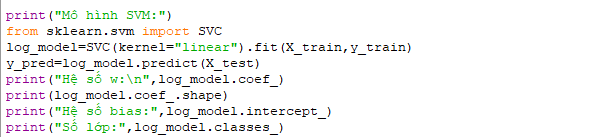
* Khai báo thư viện



* Đọc file dữ liệu, phân tích dữ liệu, chuẩn hoá, chia tỉ lệ tập train-test



* Chạy mô hình học máy

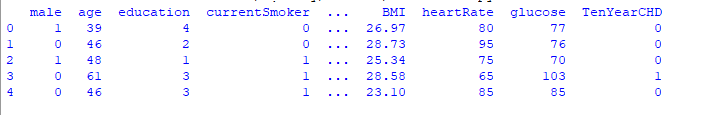


* Đánh giá mô hình dựa trên kết quả dự đoán (với độ đo đơn giản Accuracy, Precision, Recall)

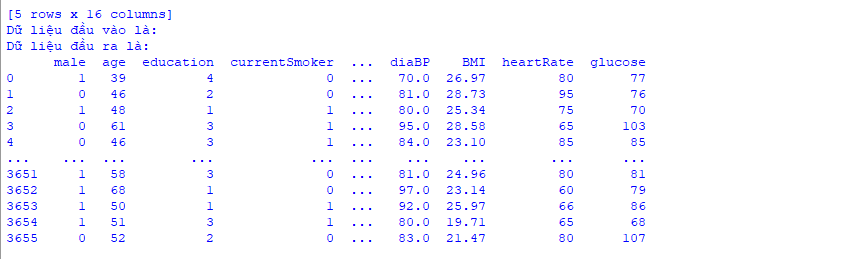


**3.3.2.2** Kết quả

* In dữ liệu



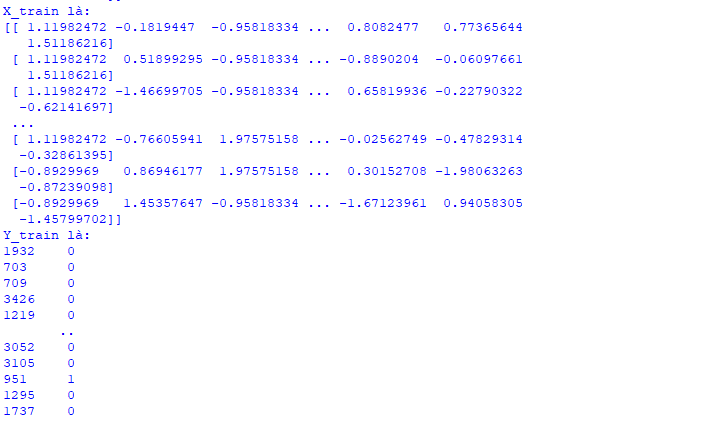
* + Dữ liệu đầu vào X:



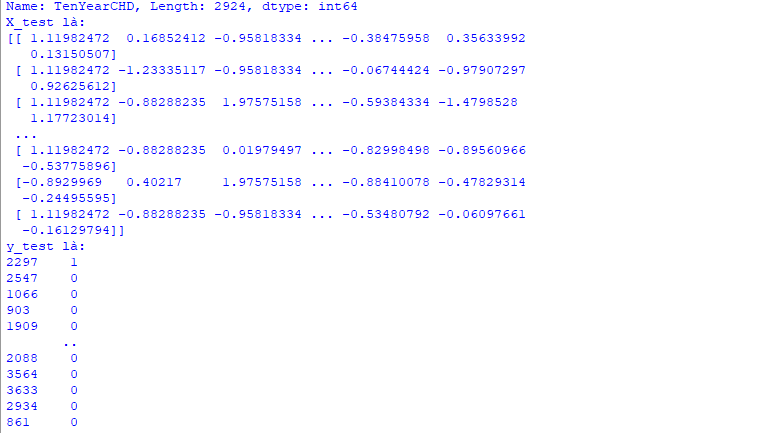
* + Dữ liệu đầu ra Y:



* X\_train,Y\_train,X\_test,Y\_test
  + X\_train,Y\_train



* + X\_test,Y\_test



* Đánh giá kết quả:
  + Sau khi thử nghiệm với bộ cơ sở dữ liệu thì kết quả phân lớp đạt 85%



## 3.4 Đánh Giá

Ưu điểm:

+ Cơ sở toán học chặt chẽ, đây là thuật toán mà ta luôn luôn tìm được lời giải tốt nhất cho thuật toán đó.

+ Sau khi học xong thì ta dự đoán rất nhanh, vì ta chỉ cần thay vào phương trình siêu phẳng và kiểm tra xem lớn hơn 0 hay nhỏ hơn 0 thì ta sẽ có kết quả.

Nhược điểm:

+ Để SVM hiệu quả thì dữ liệu cần phải được trích-chọn các thuộc tính phù hợp, SVM không thể lựa chọn thuộc tính được nên các bạn phải tự lựa chọn thuộc tính hoặc dùng các thuật toán khác.

# CHƯƠNG 4 : KẾT LUẬN

Trong thời gian và điều kiện nhất định, mặc dù đã rất cố gắng để hoàn thành đồ án với tất cả sự nỗ lực. Tuy nhiên, do bước đầu đi vào thực tế, tìm hiểu và xây dựng báo cáo trong thời gian có hạn, với lượng kiến thức còn hạn chế, nhiều bỡ ngỡ, nên bài làm của em không thể tránh khỏi những thiếu sót. Em rất mong nhận được sự quan tâm, thông cảm và những đóng góp quý báu của các thầy cô và các bạn để đồ án này ngày càng hoàn thiện hơn.

Em xin gửi lời cảm ơn đến giáo viên giảng dạy bộ môn **Khai phá dữ liệu** của cô Nguyễn Thị Thanh Tân. Cô đã tận tình chỉ dạy, quan sát và giúp đỡ em từ những ngày đầu cũng như trong quá trình nghiên cứu và học tập.

Em xin chân thành cảm ơn!