**TRƯỜNG ĐẠI HỌC GIAO THÔNG VẬN TẢI**

**KHOA CÔNG NGHỆ THÔNG TIN**

---------o0o---------



BÁO CÁO BÀI TẬP LỚN

KHAI PHÁ DỮ LIỆU

**Sử dụng giải thuật K-Nearest xử lý dữ liệu**

**của bệnh ung thư vú**

**Giảng viên hướng dẫn: Nguyễn Quốc Tuấn**

**Sinh viên thực hiện:**

|  |  |
| --- | --- |
| Nguyễn Thị Ngũ | 191200758 |
| Bùi Văn Hùng | 191202050 |
| Ngô Trung Hiếu | 191213136 |
| Đào Nhật Tân | 191204022 |
| Đoàn Huy Tuấn Hải | 191210360 |

*Hà Nội, Ngày 26 Tháng 10 Năm 2022*

---------------o0o---------------

**Mục lục**

[LỜI MỞ ĐẦU 3](#_Toc120467766)

[PHẦN I: GIỚI THIỆU 4](#_Toc120467767)

[PHẦN II: PHÂN LỚP TẬP DỮ LIỆU BREAST CANCER WISCONSIN BẰNG THUẬT TOÁN 6](#_Toc120467768)

[**1. Giới thiệu về thuật toán K Nearest Neighbor** 6](#_Toc120467769)

[**2. Công cụ sử dụng để phân tích** 6](#_Toc120467770)

[**3. Tiến hành khai phá** 7](#_Toc120467771)

[3.1 Sơ đồ luồng phân tích 7](#_Toc120467772)

[3.2 Nhập các gói pakage từ thư viện anacoda 7](#_Toc120467773)

[3.3 Nhập và làm sạch dữ liệu 8](#_Toc120467774)

[3.4 Chuẩn hóa dữ liệu 8](#_Toc120467775)

[PHẦN III: Tổng kết 11](#_Toc120467776)

[1. Kết luận 11](#_Toc120467777)

[2. Tư liệu tham khảo 11](#_Toc120467778)

# **LỜI MỞ ĐẦU**

Sự phát triển của công nghệ thông tin và việc ứng dụng công nghệ thông tin  
trong nhiều lĩnh vực của đời sống, kinh tế xã hội trong nhiều năm qua cũng đồng  
nghĩa với lượng dữ liệu đã được các cơ quan thu thập và lưu trữ ngày một tích luỹ  
nhiều lên. Đi cùng với việc này là việc bùng nổ dữ liệu (big data) là một vấn đề hiện nay. Đây có thể được coi là vùng đất mới cho các nhà phát triển công nghệ. Tương lai không xa nữa thì việc khai thác dữ liệu sẽ là điều không thể thiếu trong công việc của từng nhà phát triển.

Vì vậy việc khai thác những tinh hoa từ bộ dữ liệu khổng lồ sẽ là miếng bánh mà được các nhà phát triển nhắm tới. Dữ liệu có thể được khai thác theo nhiều cách khác nhau, theo nhiều mục đích khác nhau để phục vụ cho nhiều vấn đề thiết yếu của xã hội, người dùng.. Mặt khác, trong môi trường cạnh tranh, người ta ngày càng có nhiều thông tin với tốc độ nhanh để trợ giúp việc ra quyết định và ngày càng có nhiều câu hỏi mang tính chất định tính cần phải trả lời dựa trên một khối lượng dữ liệu khổng lồ đã có. Với những lý do như vậy, các phương pháp quản trị và khai  
thác cơ sở dữ liệu truyền thống ngày càng không đáp ứng được thực tế đã làm phát  
triển một khuynh hướng kỹ thuật mới đó là Kỹ thuật phát hiện tri thức và khai  
phá dữ liệu (KDD - Knowledge Discovery and Data Mining).

Trong quá trình tìm hiểu và thực hiện, vì thời gian có hạn nên sẽ không tránh được những sai sót. Chúng em rất mong nhận được những ý kiến đóng góp của thầy để đề tài của chúng em được hoàn thiện hơn.

Chúng em xin chân thành cảm ơn!

# **PHẦN I: GIỚI THIỆU**

 Ung thư vú là nguyên nhân thứ hai gây tử vong do ung thư ở phụ nữ tại Hoa Kỳ. Mặc dù trong thập kỉ qua, số người tử vong do bệnh này đã giảm đáng kể. Tuy nhiên điều quan trọng là phải tiếp tục thực hiện các tiến bộ trong cách thức chẩn đoán để phát hiện bệnh sớm giúp nâng cơ hội sống sót cho các bệnh nhân. Trong vài thập kỷ qua, nhận thức của công chúng và nghiên cứu khoa học liên quan đến ung thư vú tăng lên đáng kể. Thật không may, ung thư vú tiếp tục là nguyên nhân thứ hai gây tử vong ở phụ nữ ở Hoa Kỳ, chỉ đứng sau ung thư phổi. Nghiên cứu gần đây ước tính khoảng 15% bệnh nhân ung thư mới được chẩn đoán sẽ chết trong năm 2015.

Mục tiêu của nhóm chúng em là sử dụng thuật toán K-nearest neighbors trong kỹ thuật phân lớp dữ liệu để xác định mô hình chẩn đoán nhất chính xác sự hiện diện của một khối u ác tính trong tập dữ liệu này.

Bộ dữ liệu này được lấy từ cơ sở dữ liệu ung thư vú Wisconsin 1991, được biên soạn bởi Tiến sĩ William H. Wolberg và được lưu trữ tại trang web UCI Machine Learning Repository Bộ dữ liệu chứa 699 mẫu bệnh phẩm được lấy từ các tế bào ung thư vú của bệnh nhân.

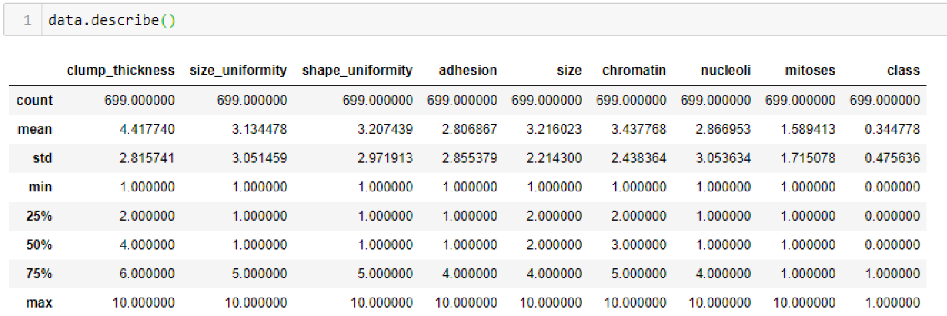
Có 11 thuộc tính cho mỗi quan sát được miêu tả trong bảng sau:

|  |  |
| --- | --- |
| Thuộc tính | Thang đo |
| 1. Sample code number | ID |
| 1. Clump Thickness | 1-10 |
| 1. Uniformity of Cell Size | 1-10 |
| 1. Uniformity of Cell Shape | 1-10 |
| 1. Marginal Adhesion | 1-10 |
| 1. Single Epithelial Cell Size | 1-10 |
| 1. Bare Nuclei | 1-10 |
| 1. Bland Chromatin | 1-10 |
| 1. Normal Nucleoli | 1-10 |
| 1. Mitoses | 1-10 |
| 1. Class | 2 – cho bệnh lành tính  4 – cho bệnh ác tính |

*Bảng 1: Mô tả tập dữ liệu Breast cancer wisconsin (Original)*

Biến ID thể hiện thông tin của quan sát Biến Class thể hiện chẩn đoán xem khối u lành tính (65,52% trường hợp) ác tính (34,48% trường hợp).

Các biến đầu vào còn lại được đo trên thang đo thứ tự (1-10), với giá trị 1 cho biết trạng thái bình thường và giá trị 10 cho biết trạng thái bất thường cao. Các giá trị bị thiếu : Có 16 trường hợp chứa một giá trị thuộc tính bị thiếu được biểu thị bằng "?".



*Bảng 2: Khám phá bộ dữ liệu bằng explore*

Count: Số lượng quan sát

Mean: Giá trị trung bình

STD: Độ lệch chuẩn

Min: thang đo thấp nhất

Max: Thang đo cao nhất

# **PHẦN II: PHÂN LỚP TẬP DỮ LIỆU BREAST CANCER WISCONSIN BẰNG THUẬT TOÁN**

**K – NEAREST NEIGHBOR**

* + - 1. **Giới thiệu về thuật toán K Nearest Neighbor**

K-nearest neighbor là một trong những thuật toán supervised-learning đơn giản nhất (mà hiệu quả trong một vài trường hợp) trong Machine Learning. Khi training, thuật toán này không học một điều gì từ dữ liệu training (đây cũng là lý do thuật toán này được xếp vào loại lazy learning), mọi tính toán được thực hiện khi nó cần dự đoán kết quả của dữ liệu mới. K-nearest neighbor có thể áp dụng được vào cả hai loại của bài toán Supervised learning là Classification và Regression.

Với KNN, trong bài toán Classification, label của một điểm dữ liệu mới (hay kết quả của câu hỏi trong bài thi) được suy ra trực tiếp từ K điểm dữ liệu gần nhất trong training set. Label của một test data có thể được quyết định bằng major voting (bầu chọn theo số phiếu) giữa các điểm gần nhất, hoặc nó có thể được suy ra bằng cách đánh trọng số khác nhau cho mỗi trong các điểm gần nhất đó rồi suy ra label.

Đánh giá về thuật toán K-nearest neighbor

Ưu điểm:

• Dễ sử dụng và cài đặt

• Xử lý tốt với dữ liệu bị nhiễu

Khuyết điểm:

• Cần lưu tất cả các mẫu để dự đoán (Gây tốn bộ nhớ Ram )

• Cần nhiều thời gian để xác định lớp cho một mẫu mới (cần tính và so sánh khoảng cách đến tất cả các mẫu huấn luyện) 4 / 8

• Phụ thuộc vào giá trị k do người dùng lựa chọn. Nếu k quá nhỏ, nhạy cảm với nhiễu. Nếu k quá lớn, vùng lân cận có thể chứa các điểm của lớp khác.

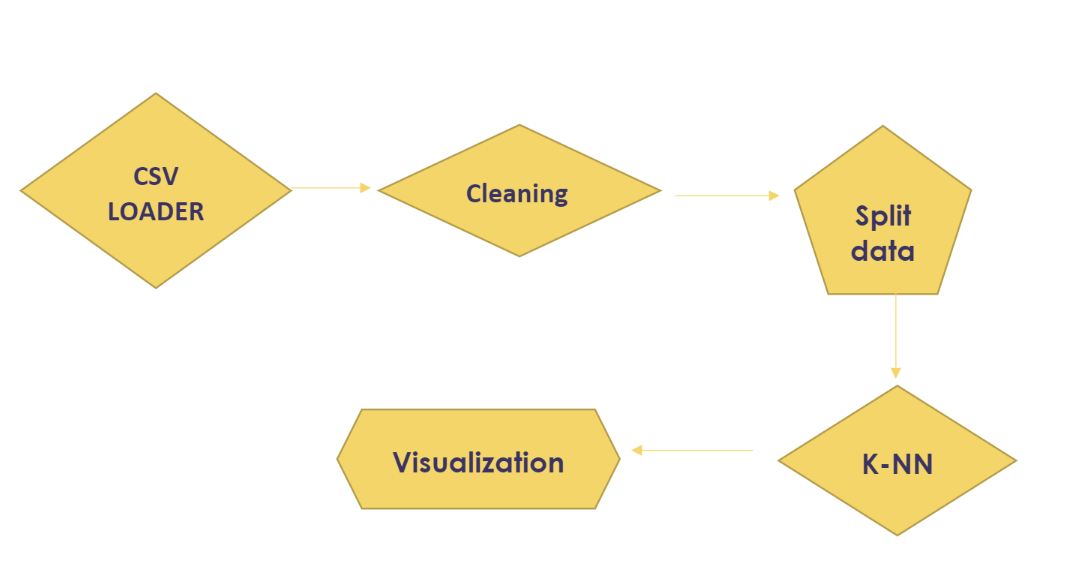
* + - 1. **Công cụ sử dụng để phân tích**

Python là một ngôn ngữ lập trình thông dịch do Guido van Rossum tạo ra năm 1990. Python hoàn toàn tạo kiểu động và dùng cơ chế cấp phát bộ nhớ tự động; do vậy nó tương tự như Perl, Ruby, Scheme, Smalltalk, và Tcl. Python được phát triển trong một dự án mã mở, do tổ chức phi lợi nhuận Python Software Foundation quản lý.

Jupyter notebook là một IDE chạy trên nền web rất phổ biến dùng cho lập trình python.

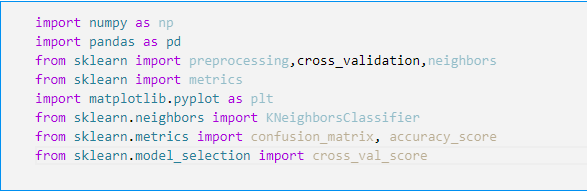
## **3. Tiến hành khai phá**

### 3.1 Sơ đồ luồng phân tích



*Hình 3: Sơ đồ phân tích*

## 3.2 Nhập các gói pakage từ thư viện anacoda



**Numpy:** là pakage chủ yếu cho việc tính toán khoa học trên Python. Vì Numpy hỗ trợ mạnh mẽ việc tính toán với matrix, vector là các hàm đại số tuyến tính cơ bản nên nó được sử dụng nhiều trong việc implement các thuật toán Machine Learning

**Pandas:** Pakage dùng để tính toán dạng data frame

**Matplotlib:** Pakage vẽ biểu đồ

### https://lh6.googleusercontent.com/bHBHiIlvJLtoq078kJ0kO2TEfICA1h_sQRq47ycVKEV88bNBoeA3hB3MEzdYq9Z5kg3Xo_0tusmeatl3cBG2uHVeMMaANxYzwIOXoGHr9xC8G3nOcHI2oeWrydi_PN-uwD2hpI5qpXJZTY88_GCwryeK47p7IvvxPhFkWhDiRuUUBaJqnR6U9vBTmS-IvQ 3.3 Nhập và làm sạch dữ liệu

Download bộ dữ liệu và định dạng thuộc tính trong excel. Lưu file dưới dạng CSV.

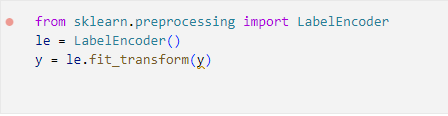
Xóa thuốc tính “id”. Thuộc tính “id” không có ý nghĩa trong tính toán và nó làm ảnh hưởng đến kết quả tính toán.

Đặt tên biến của dữ liệu là “data”

Có 16 giá trị không xác định, ta tiến hành xóa 16 giá trị này đi vì số lượng giá trị lỗi khá nhỏ với tổng thể là 699 giá trị, nó không gây ảnh hưởng đến quá nhiều đến việc tính toán

### 3.4 Chuẩn hóa dữ liệu

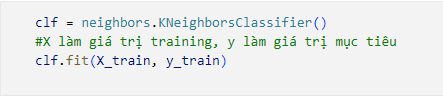
Để thuân tiện cho việc tính toán, ta chuyển dữ liệu sang dạng số nguyên



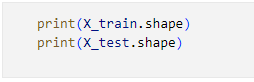
Cắt lớp dữ liệu thành 2 phần train và test

Tỷ lệ train/ test = 0.2

https://lh3.googleusercontent.com/9WHM1ZPuqSdpXSrzFdyO-PYWeNyv7EuojnhsoPqnzVvTVTtdTKyvLPPjWP5aQARx30PeOuA1PkIns4q-_LGQo1xPwXNs_vGtS7TjeKrgGESXHYFSlCBk7JNdd9lSckbsGI5HzVPd3dWnDKdGFz_akiqmkXWw8hV-pfqLBcrZeESE6o-3PlRRtAE12Beu0w



Xem kích thước của tập tin sau khi cắt:



Kết quả nhận được:

https://lh5.googleusercontent.com/oCegsEf2AxY3AzYH58kxiwn3TQkTVVk5b7P4Y-O7idZ5oFr7gys1_3y23UsoMBco4x8soICLga1i5Liua7AIciu0pIaaFItUUBdfqrWUz9VYs4Hsxg_wIfIdpOsoS2Vm36Al3DOPvGBdXdbJgHtHGB9rq22yW-9ZW9HmmMdLQJeG9PpKSik34TDyXUayaA

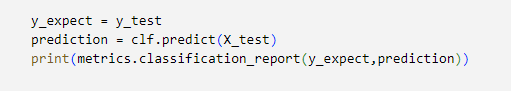
Kiểm tra độ chính xác của thuật toán:

https://lh3.googleusercontent.com/ELpKXijuhus_wiOp1wTZIoWTQx7F8iqY6tRv8TJaGQODwCrj62-JADDNCiiDqZ8Psh-0M4F9FCnn3umRqqPf4fwBXMeVxIAMlvuTyKivBYZJwh4e2XzTRYSZ4q0wUjHjzv5ZEwfbuqf_k9dq16FKQWzvhD97wC7CSt71vVntnCpnEl_ChNeb5dNkNHDlrw

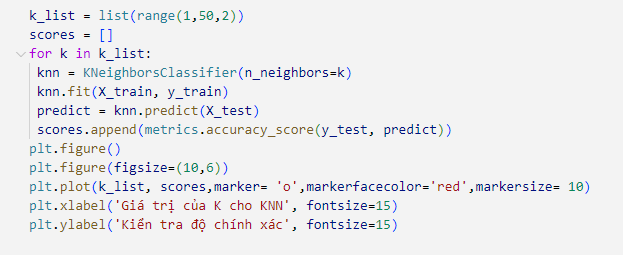
Kết quả nhận được:

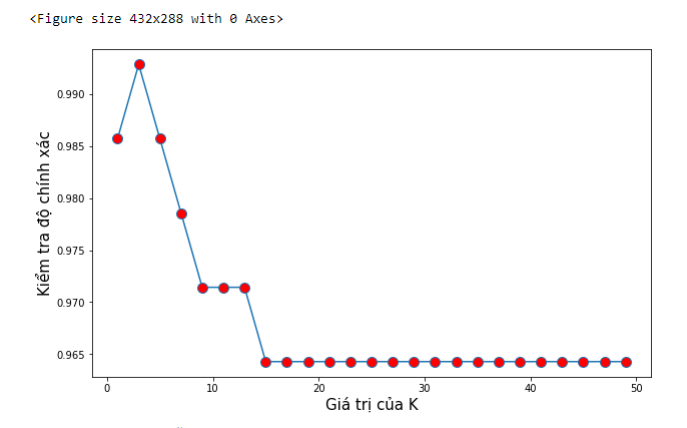
https://lh3.googleusercontent.com/yu5u8T0GsF9N8zoMBHDhMjBO2xHB5poLq0rvH8ztkqIYkwvSTLNeofCAtm1iRivuE3wc_bYLvxwqoz5snB3xcM6sGt8LPbi3MwvT0VY14xYpDHT7-MmdBqExnNXnXBd7b3sTUkSS0LJCBeWHZKm3vVxaV-gZelFSpiN033ZxhH44cqiD9jgqVF-77KN6sQ

Dự đoán nhãn lớp X cho dữ liệu được cung cấp:

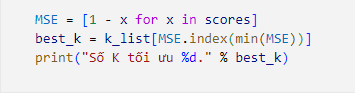


Đi tìm số K lân cận tối ưu:





Từ đồ thị ra có thể dễ dàng nhận thấy từ khoảng K (0;10) thuật toán cho độ chính xác cao nhất. Để tìm con số chính xác ta tiến hành kiểm tra tiếp:



Kết quả nhận được:

https://lh3.googleusercontent.com/cWvABucP3PZvT8XslOOFSXAzHKOx1Hlj5BL9fK9gj_fqjN5WBlhXJFbF-LEnezu1ZW5zN0jVmlYZio6MysgW5LGG0C19inozWtrRFaq7cvPJic-_TW0B7YdMADbZOwIlPvXjp5n9baPFjEYjKuk6JMwr6Hp-R5s4iFiGe2qrvx2LnLZXXWg9WY0SvoDu5A

# **PHẦN III: Tổng kết**

**Kết luận**

Vậy ta có thể kết luận rằng với k = 3 thì thuật toán KNN đạt độ chính xác cao nhất là Accuracy = 0.9857. Nghĩa là mô hình có thể dự đoán chính xác được mẫu tế bào ung thư của bệnh nhân trong bộ dữ liệu với độ chính xác tới 98.57%. Đây là con số khá cao và hoàn toàn chấp nhận được.

Khai phá dữ liệu bằng ngôn ngữ lập trình python là một lĩnh vực mới và khó. Trong quá trình thực hiện nhóm em gặp rất nhiều khó khăn trong việc tiếp cận vấn đề và sẽ có những sai sót. Chúng em hy vọng nhận được sự góp ý của thầy. Xin cảm ơn thầy!

## **2. Tư liệu tham khảo**

[1] Breast Cancer Wisconsin (Original) Data Set

[2] K-nearesr neighbors – machinelearningcoban.com

[3] Breast Cancer Wisconsin – [www.kaggle.com](http://www.kaggle.com)

[4] Ung thư vú from wikipedia.org

[5] K-nearest neighbors scikit-learn.org