TRƯỜNG ĐẠI HỌC THỦY LỢI

**KHOA CÔNG NGHỆ THÔNG TIN**



**BÁO CÁO BÀI TẬP LỚN MÔN HỌC**

**KHAI PHÁ DỮ LIỆU**

***Đề tài:***

***Phân tích so sánh biểu hiện protein của chuột trên bộ dữ liệu Mice Protein Expression Dataset***

**Nhóm sinh viên thực hiện:** Đặng Thị Ngọc Linh

Nguyễn Thành Trung

Phùng Khắc Dũng

**Giảng viên phụ trách môn học:** TS. Trần Mạnh Tuấn

*Hà Nội, tháng 12 năm 2022*

MỤC LỤC

[BẢNG PHÂN CÔNG CÔNG VIỆC 2](#_Toc121058269)

[LỜI MỞ ĐẦU 3](#_Toc121058270)

[**Chương 1. MÔ TẢ DỮ LIỆU** 4](#_Toc121058271)

[**1.** **Thông tin tập dữ liệu** 4](#_Toc121058272)

[**2.** **Thông tin thuộc tính** 5](#_Toc121058273)

[**CHƯƠNG 2. TIỀN XỬ LÝ DỮ LIỆU** 6](#_Toc121058274)

[**2.1.** **Dữ liệu thô** 7](#_Toc121058275)

[**2.2.** **Tiền xử lý** 7](#_Toc121058276)

[**2.2.1.** **Quan sát dữ liệu** 7](#_Toc121058277)

[**2.2.2.** **Làm sạch dữ liệu** 10](#_Toc121058278)

[**CHƯƠNG 3 KHAI PHÁ DỮ LIỆU** 13](#_Toc121058279)

[**3.1** **K-Nearest Neighbours (KNN)** 13](#_Toc121058280)

[**3.2.1** **Định nghĩa** 13](#_Toc121058281)

[**3.2.2** **Thuật toán KNN** 13](#_Toc121058282)

[**3.2.3**  **Tiến hành khai phá trên bộ dữ liệu bằng thuật toán KNN** 14](#_Toc121058283)

[**3.2.4** **Ưu điểm và nhược điểm** 17](#_Toc121058284)

[**3.2 Random Forest Classifier** 17](#_Toc121058285)

[**3.2.1** **Định nghĩa** 17](#_Toc121058286)

[**3.3.2** **Thuật toán** 18](#_Toc121058287)

[**3.3.3**  **Tiến hành khai phá dữ liệu** 18](#_Toc121058288)

[**3.3.4** **Ưu và nhược điểm** 19](#_Toc121058289)

[**CHƯƠNG 4. NHẬN XÉT ĐÁNH GIÁ, KẾT LUẬN VÀ TÀI LIỆU THAM KHẢO** 20](#_Toc121058290)

[**1.2.** **Kết luận** 20](#_Toc121058291)

[**1.3.** **Tài liệu tham khảo** 21](#_Toc121058292)

# BẢNG PHÂN CÔNG CÔNG VIỆC

|  |  |  |  |
| --- | --- | --- | --- |
| STT | Tên | MSV | Công việc |
| 1 | Đặng Thị Ngọc Linh (TN) | 1951060807 | * Phân công công việc * Viết báo cáo * KNN |
| 2 | Phùng Khắc Dũng | 1951060646 | * Viết báo cáo * Làm powerpoint * Tiền xử lý dữ liệu |
| 3 | Nguyễn Thành Trung | 1951060829 | * Viết báo cáo * Random forest |

# LỜI MỞ ĐẦU

Khai phá dữ liệu đã và đang được nghiên cứu, ứng dụng trong nhiều lĩnh vực khác nhau ở các nước trên thế giới. Ở Việt Nam, kỹ thuật này đang được nghiên cứu và dần đưa vào ứng dụng. Khai phá dữ liệu là một bước trong quy trình phát hiện tri thức. Hiện nay, mọi người không ngừng tìm tòi các kỹ thuật để thực hiện khai phá dữ liệu một cách nhanh nhất và có được kết quả tốt nhất.

Trong bài tập lớn này, chúng em tìm hiểu và trình bày về hai kỹ thuật trong khai phá dữ liệu để phân lớp dữ liệu với đề tài “Phân tích so sánh biểu hiện protein của chuột trên bộ dữ liệu Mice Protein Expression Dataset”.

Trong quá trình làm bài tập lớn này, chúng em xin gửi lời cảm ơn đến thấy giáo Trần Mạnh Tuấn. Thầy đã rất tận tình hướng dẫn chi tiết cho chúng em, những kiến thức thầy cung cấp rất hữu ích Chúng em rất mong nhận được những góp ý từ thầy.

**Chương 1. MÔ TẢ DỮ LIỆU**

1. **Thông tin tập dữ liệu**

- Tập dữ liệu về mức độ biểu hiện của 77 protein được đo trong vỏ não của 8 lớp chuột thí nghiệm và hội chứng down tiếp xúc với điều kiện sợ bối cảnh.

- Tập dữ liệu bao gồm các mức biểu hiện của 77 sửa đổi protein/protein tạo ra các tín hiệu có thể phát hiện được trong phần hạt nhân của vỏ não.

- Có 38 chuột kiểm soát và 34 chuột trisomic (hội chứng Down), với tổng số 72 chuột. Trong các thí nghiệm, 15 phép đo đã được đăng ký của mỗi protein trên mỗi mẫu / chuột. Do đó, đối với chuột điều khiển, có các phép đo 38x15, hoặc 570 và đối với chuột trisomic, có 34x15 hoặc 510 phép đo. Bộ dữ liệu chứa tổng số 1080 phép đo cho mỗi protein. Mỗi phép đo có thể được coi là một mẫu / chuột độc lập.

- Tám lớp chuột được mô tả dựa trên các đặc điểm như kiểu gen, hành vi và cách điều trị. Theo kiểu gen, chuột có thể được kiểm soát hoặc trisomic. Theo hành vi, một số con chuột đã được kích thích để học (sốc bối cảnh) và một số khác thì không (sốc bối cảnh) và để đánh giá hiệu quả của thuốc memantine trong việc phục hồi khả năng học ở chuột trisomic, một số chuột đã tiêm thuốc còn những người khác thì không.

1. **Thông tin thuộc tính**

Biến đầu vào:

* cột 1 : ID chuột
* cột 2 đến 78 Giá trị của mức biểu hiện của 77 protein; Tên của protein được theo sau bởi protein\_n chỉ ra rằng chúng được đo trong phần hạt nhân. giá trị trong khoảng từ 0 đến 2. Ví dụ: DYRK1A\_n
* cột 82: Lớp: c- CS-s, c-CS-m, c-SC-s, c-SC-m, t-CS-s, t-CS-m, t-SC-s, t-SC-m

Biến đầu ra: Là các lớp chuột đã được phân loại

c-CS-s: chuột điều khiển, được kích thích để học, được tiêm nước muối (9 con chuột).

c-CS-m: chuột điều khiển, được kích thích để học, được tiêm memantine (10 con chuột).

c-SC-s: chuột điều khiển , không được kích thích để học, tiêm nước muối (9 con chuột).

c-SC-m: kiểm soát chuột, không được kích thích để học, tiêm memantine (10 con chuột).

t-CS-s: chuột trisomy, được kích thích để học, được tiêm nước muối (7 con chuột).

t-CS-m: chuột trisomy, được kích thích để học, tiêm memantine (9 con chuột).

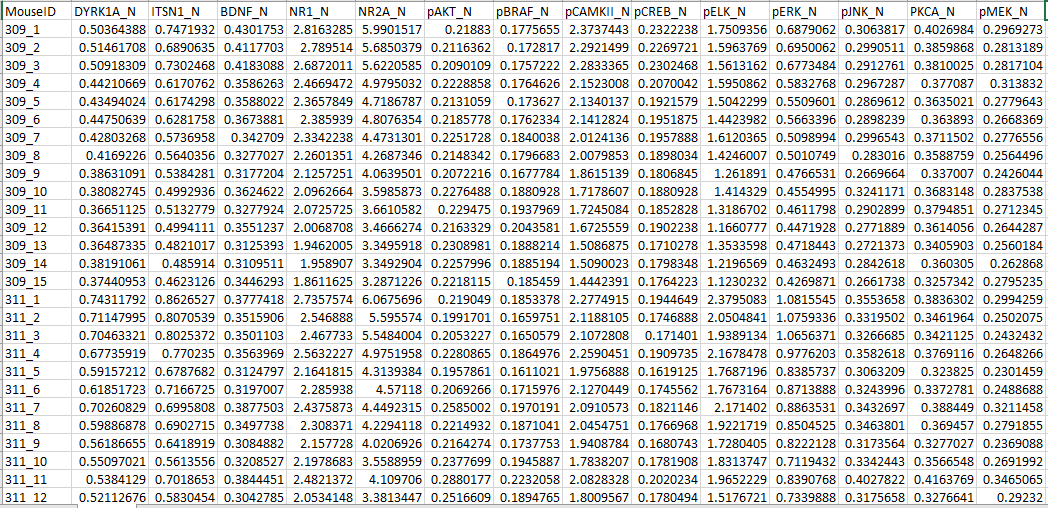
t-SC-s: chuột trisomy, không kích thích để học, tiêm nước muối (9 con chuột).

t-SC-m: chuột trisomy, không được kích thích để học, tiêm memantine (9 con chuột).

**CHƯƠNG 2. TIỀN XỬ LÝ DỮ LIỆU**

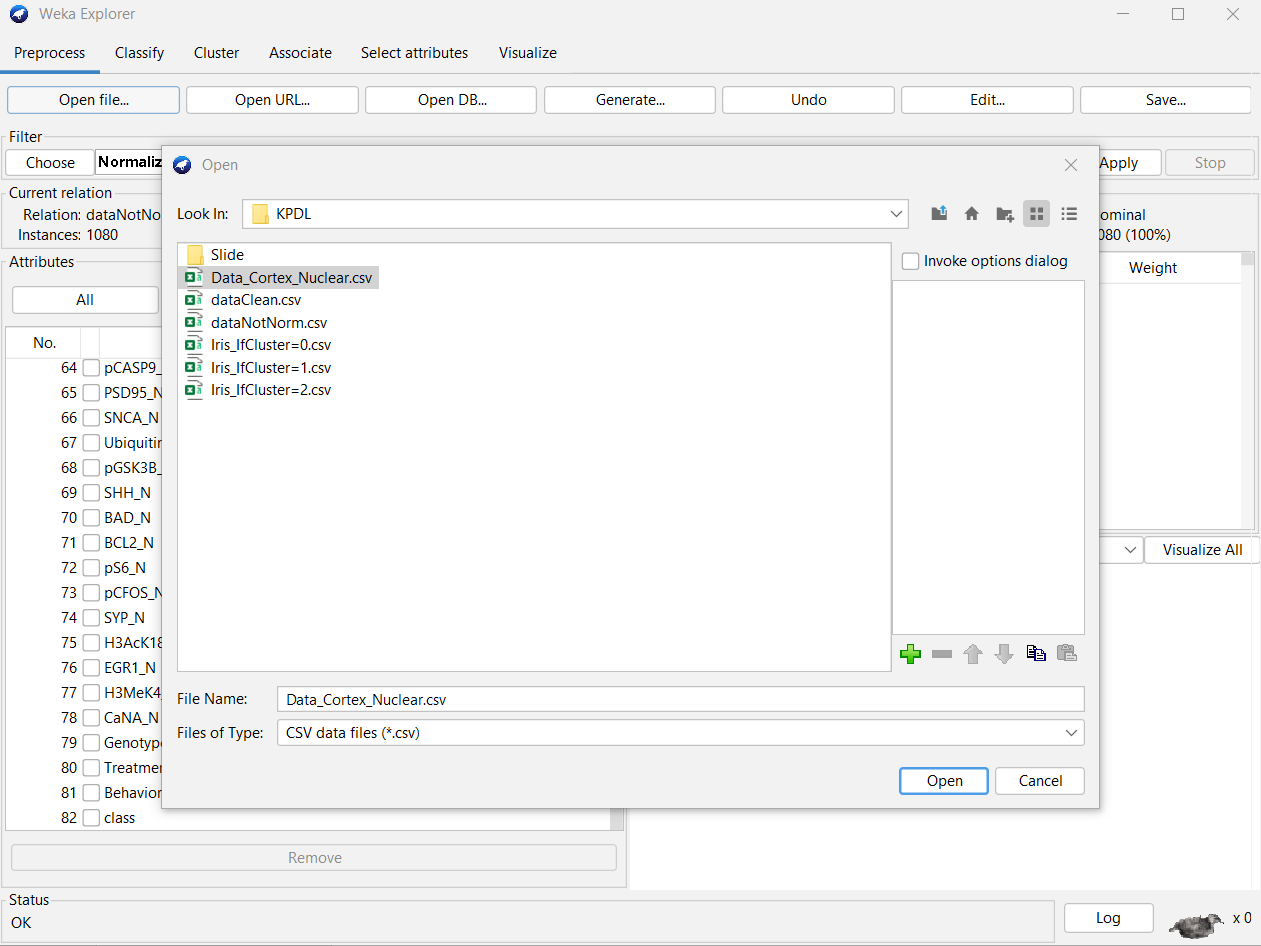
* 1. **Dữ liệu thô**

Dữ liệu ban đầu có 1082 bản ghi ở dạng csv. Giá trị được chia thành các cột.



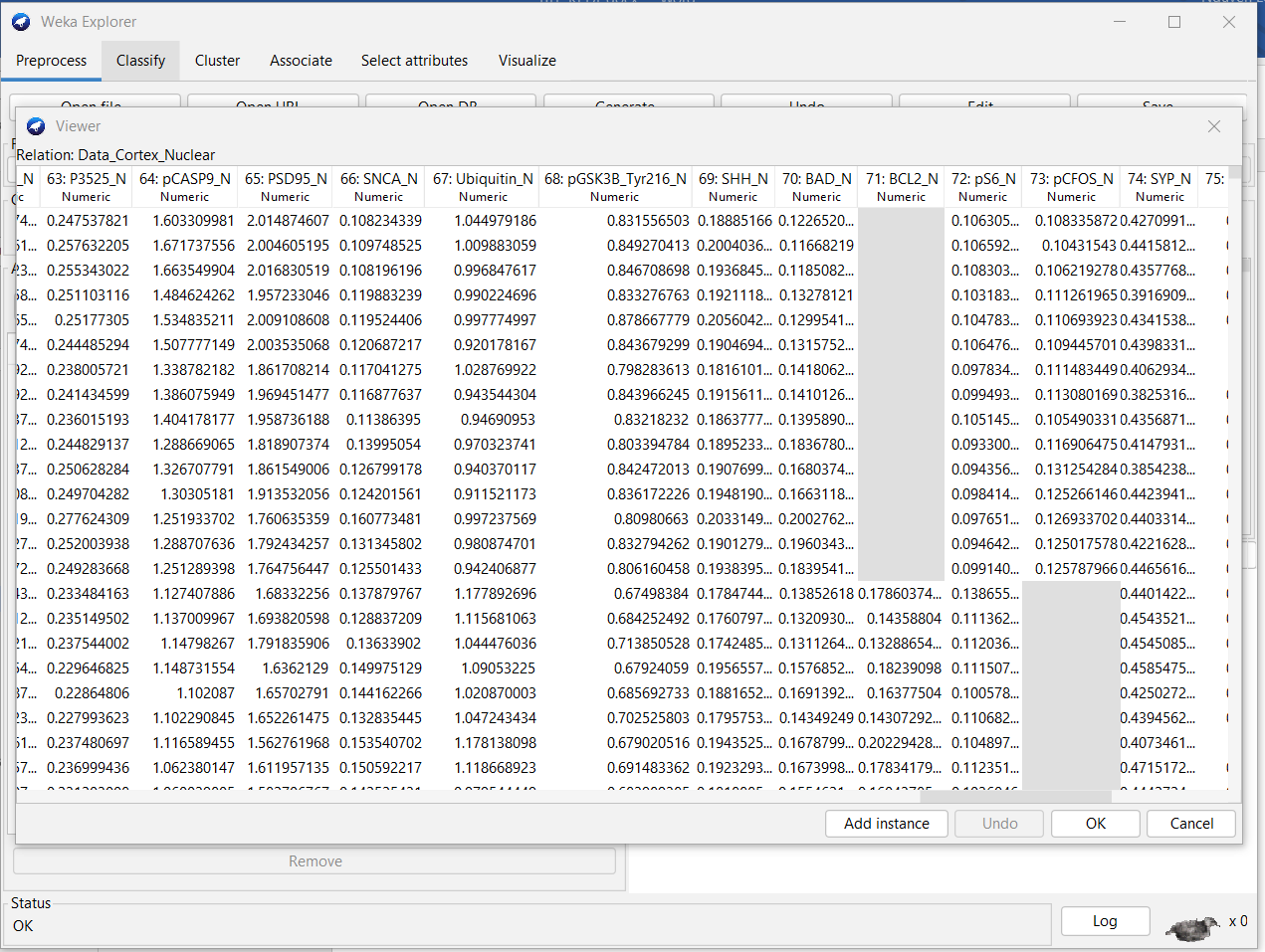
* 1. **Tiền xử lý** 
     1. **Quan sát dữ liệu**

Đầu tiên, ta dùng Weka để nạp dữ liệu Data\_Cortex\_Nuclear.csv vào chương trình như hình dưới. Ở tab Preprocessing, ta click vào nút Open files. Sau đó, tìm đến nơi lưu file Data\_Cortex\_Nuclear.csv và click Open (lưu ý: Files of Type ta chọn CSV data file (\*.csv)).

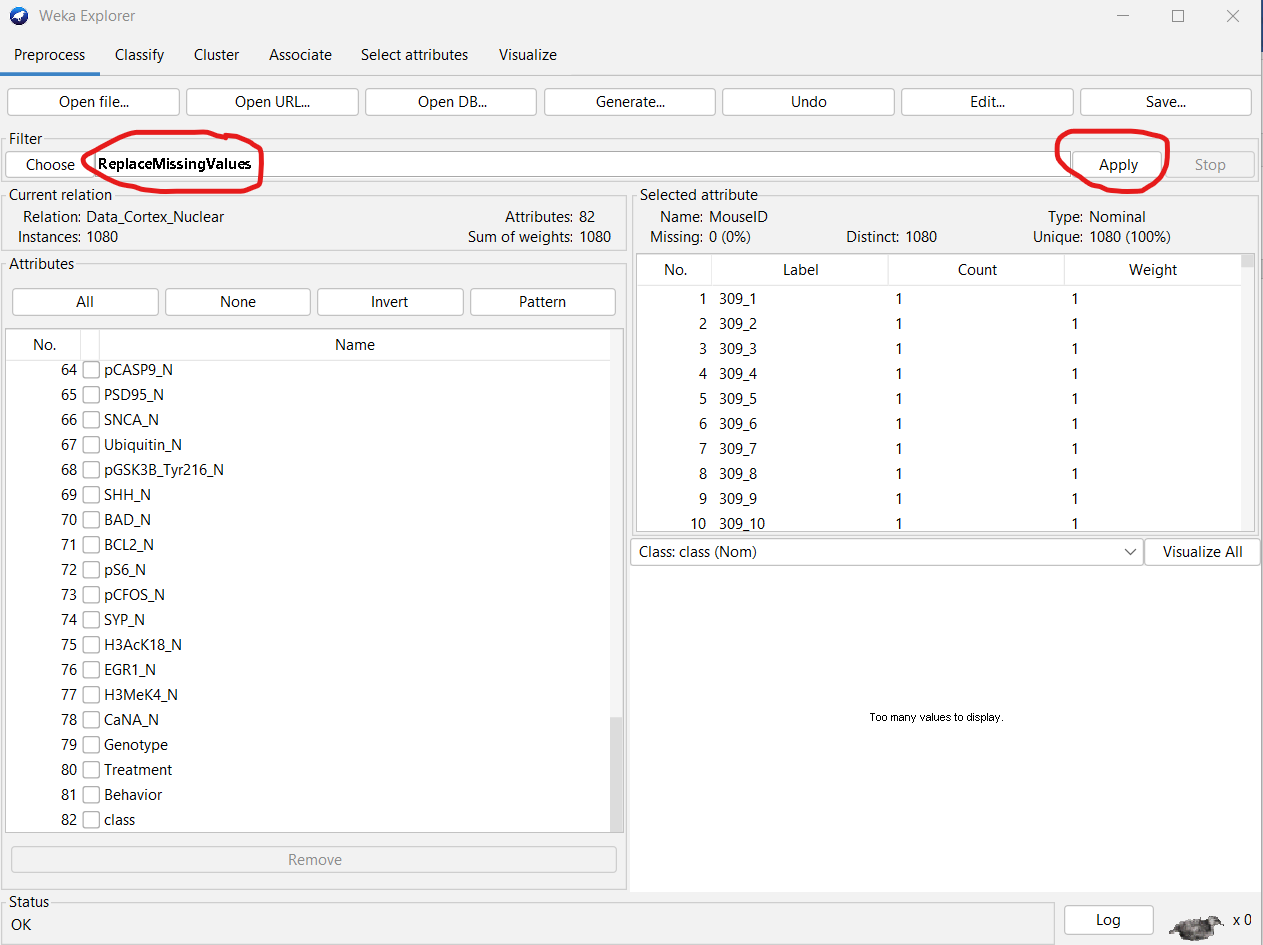


Trong tập dữ liệu tải về, ta quan sát thấy có 1080 mẫu dữ liệu, 82 thuộc tính, và trong đó có một sô thuộc tính bị thiếu về mặt dữ liệu.

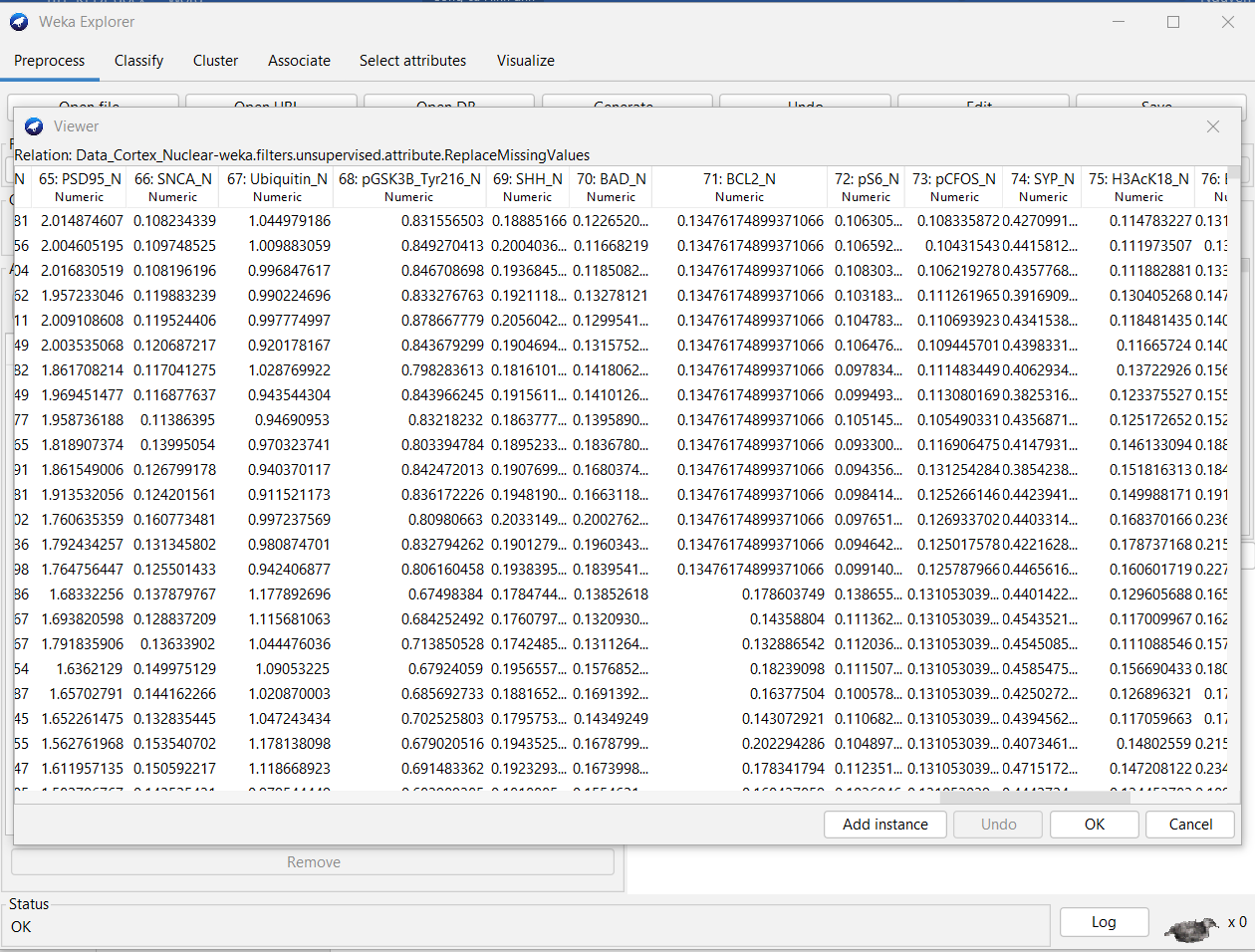
* Trước khi xử lý



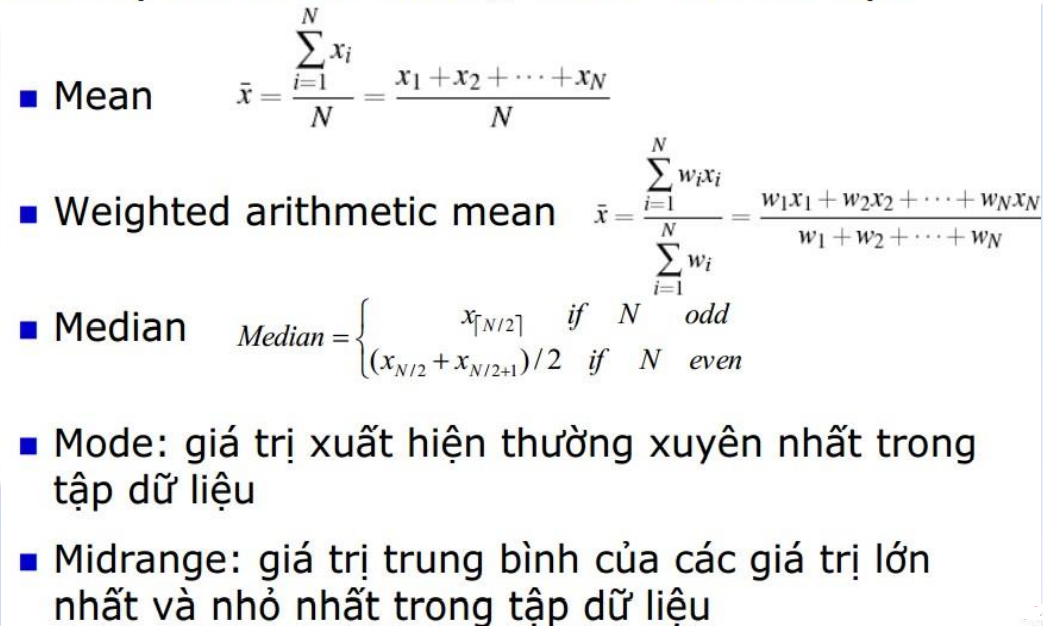
Từ đó sẽ sử dụng filter ReplaceMissingValues để tiến hành chuẩn hóa cũng như thay thế những dữ liệu bị lỗi hay bị thiếu bằng những giá trị gần nhất của nó.



Sau khi đã tiến hành ReplaceMissingValues:

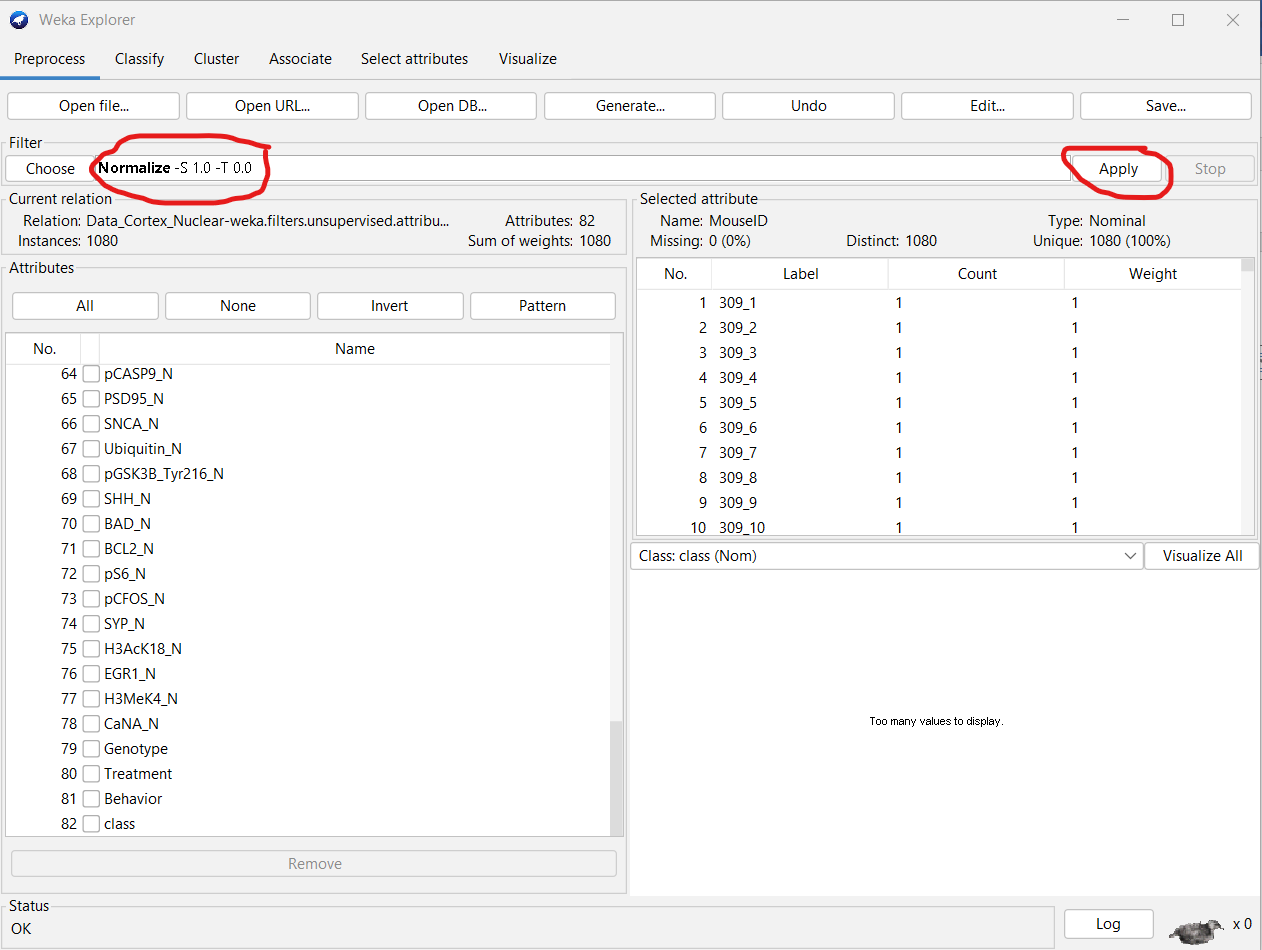


Phương pháp ReplaceMissingValues này dựa trên modes và means của tập dữ liệu và được tính theo công thức:

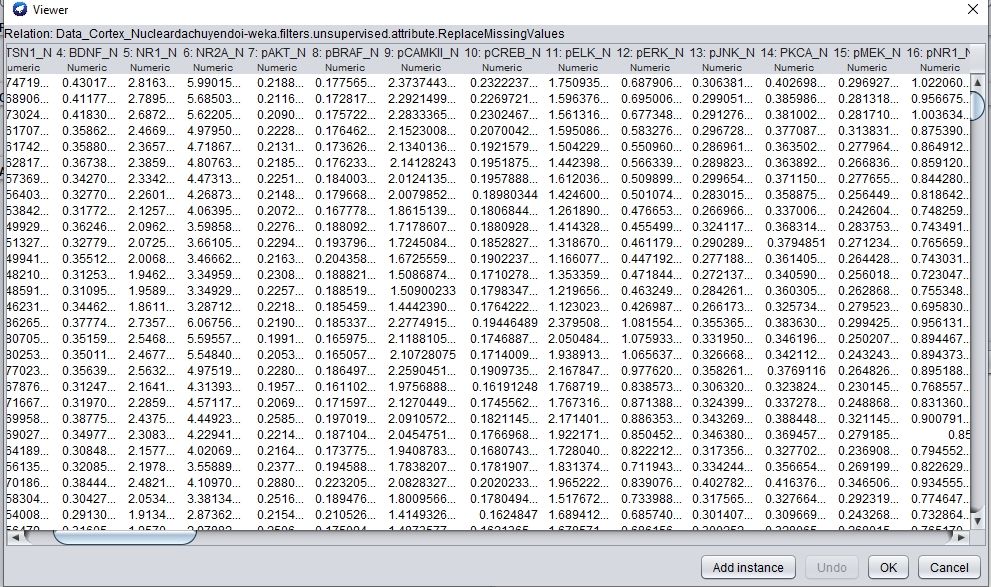


* + 1. **Làm sạch dữ liệu**

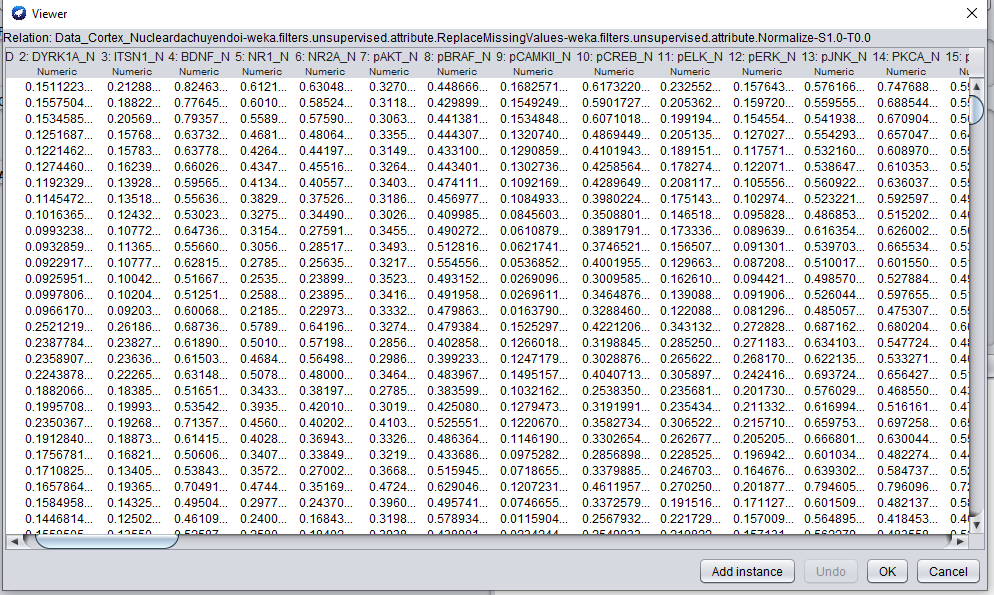
Trước tiên, ta chuẩn hóa các thuộc tính số về đoạn [0, 1] bằng bộ lọc Normalize.



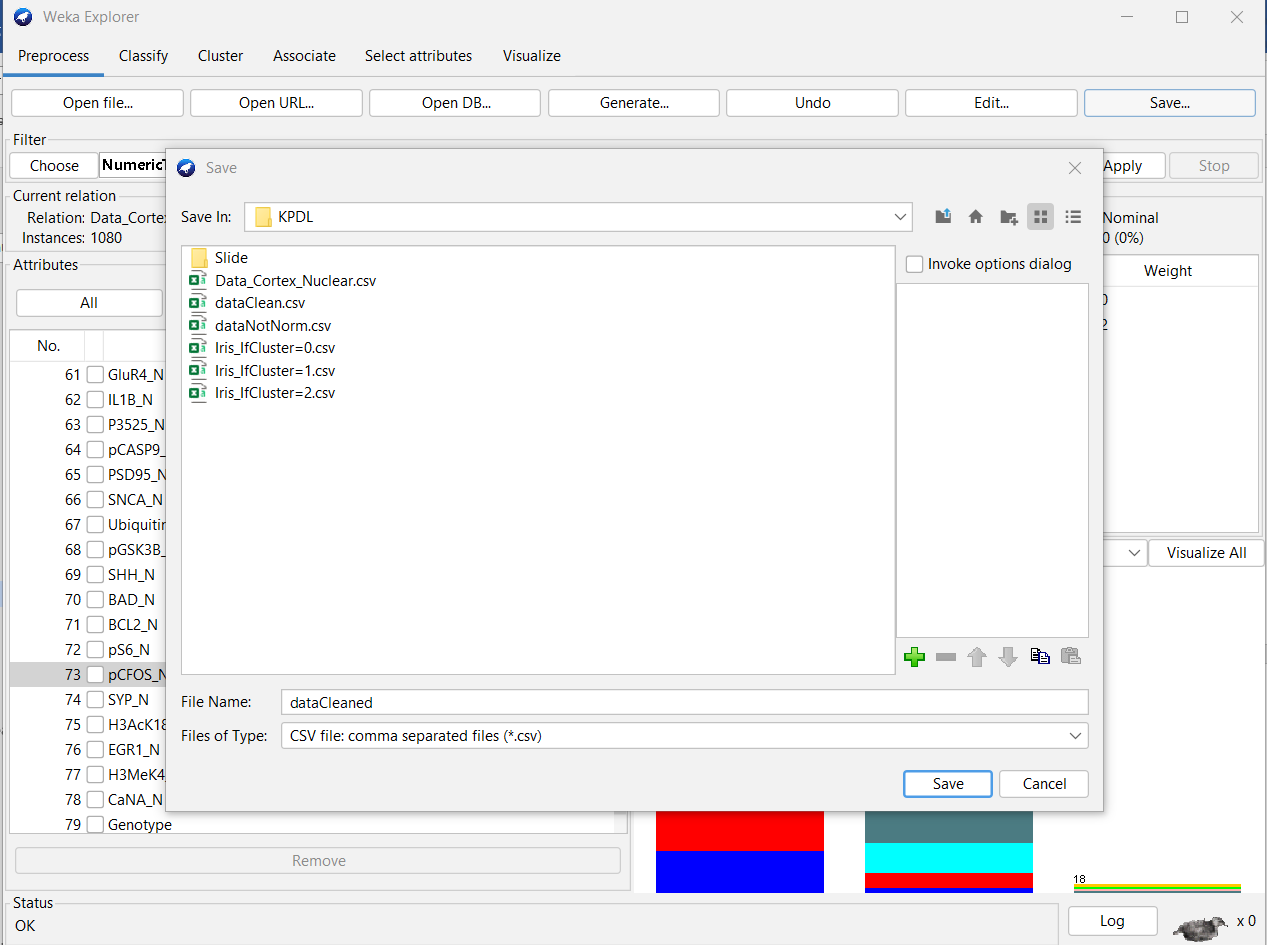
Trước khi xử lí Normalize:



Sau khi xử lí:



* Save data đã được làm sạch :



**CHƯƠNG 3 KHAI PHÁ DỮ LIỆU**

**3.1 K-Nearest Neighbours (KNN)**

**3.2.1 Định nghĩa**

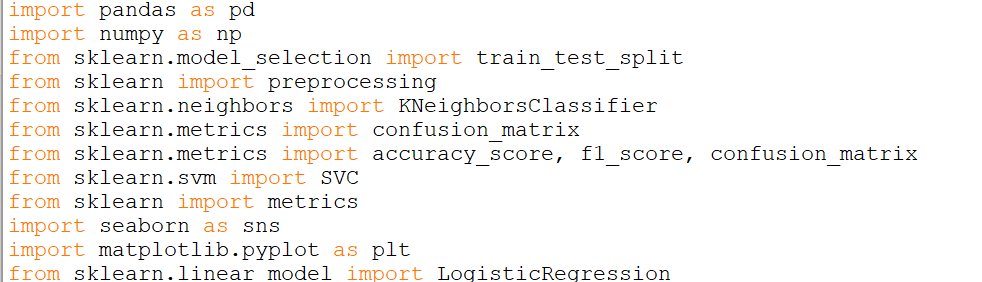
* K-nearest neighbor là một trong những thuật toán supervised-learning đơn giản nhất (mà hiệu quả trong một vài trường hợp) trong Machine Learning. Khi training, thuật toán này không học một điều gì từ dữ liệu training (đây cũng là lý do thuật toán này được xếp vào loại [lazy learning](https://en.wikipedia.org/wiki/Lazy_learning)), 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](https://machinelearningcoban.com/2016/12/27/categories/" \l "classification-phan-loai) và [Regression](https://machinelearningcoban.com/2016/12/27/categories/#regression-hoi-quy).
* 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.
* Trong bài toán Regresssion, đầu ra của một điểm dữ liệu sẽ bằng chính đầu ra của điểm dữ liệu đã biết gần nhất (trong trường hợp K=1), hoặc là trung bình có trọng số của đầu ra của những điểm gần nhất, hoặc bằng một mối quan hệ dựa trên khoảng cách tới các điểm gần nhất đó.
* Một cách ngắn gọn, KNN là thuật toán đi tìm đầu ra của một điểm dữ liệu mới bằng cách chỉ dựa trên thông tin của K điểm dữ liệu trong training set gần nó nhất (K-lân cận), không quan tâm đến việc có một vài điểm dữ liệu trong những điểm gần nhất này là nhiễu

**3.2.2 Thuật toán KNN**

* Bước 1: Xác định tham số K = số láng giềng gần nhất
* Bước 2: Tính khoảng cách đối tượng cần phân lớp với tất cả các đối tượng training data.
* Bước 3: Sắp xếp khoảng cách theo thứ tự tăng dần và xác định K láng giềng gần nhất với đối tượng cần phân lớp
* Bước 4: Lấy tất cả các lớp của K láng giềng gần nhất
* Bước 5: Dựa vào phần lớn lớp của K để xác định lớp cho đối tượng cần phân lớp .

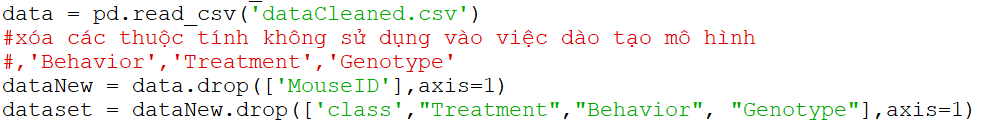
**3.2.3 Tiến hành khai phá trên bộ dữ liệu bằng thuật toán KNN**

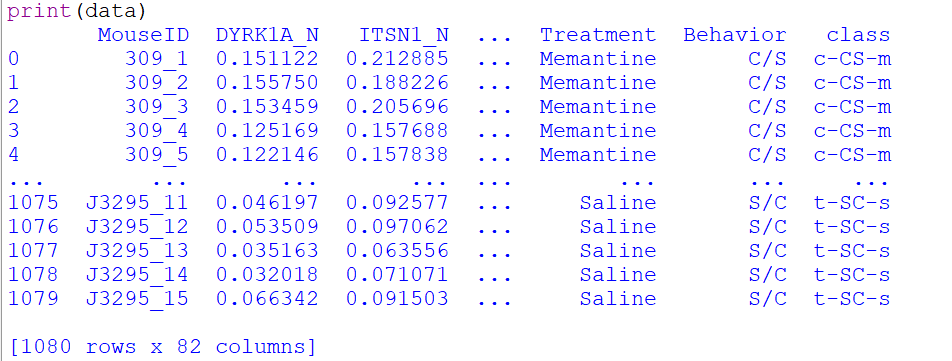
Import thư viện



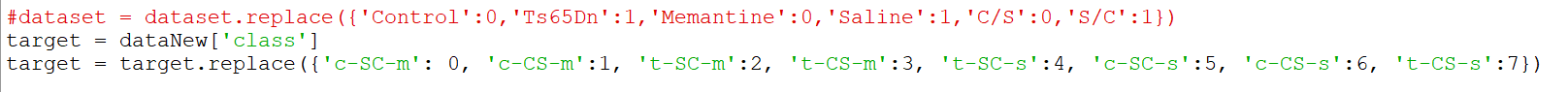
Đọc file dataCleaned.csv

Xóa các cột không sử dụng cho tranning model

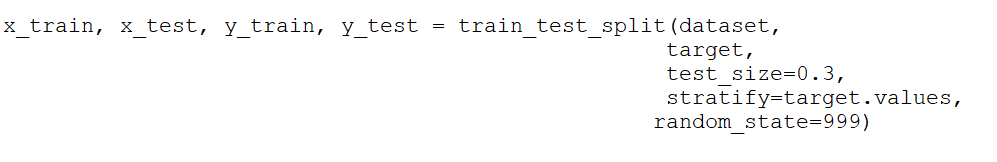


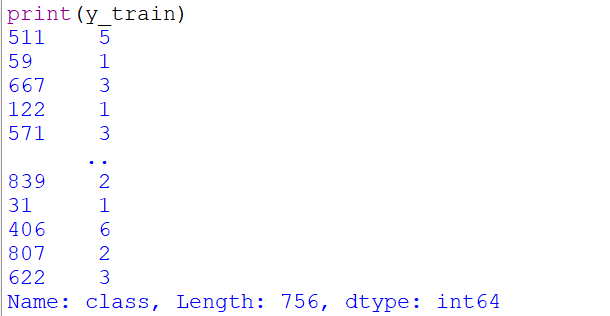
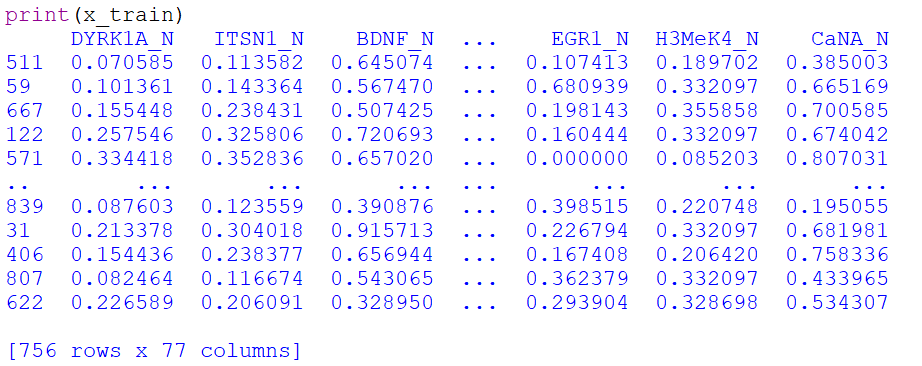


* Gán các nhãn c- CS-s, c-CS-m, c-SC-s, c-SC-m, t-CS-s, t-CS-m, t-SC-s, t-SC-m tương ứng với 0,1,2,3,4,5,6,7.

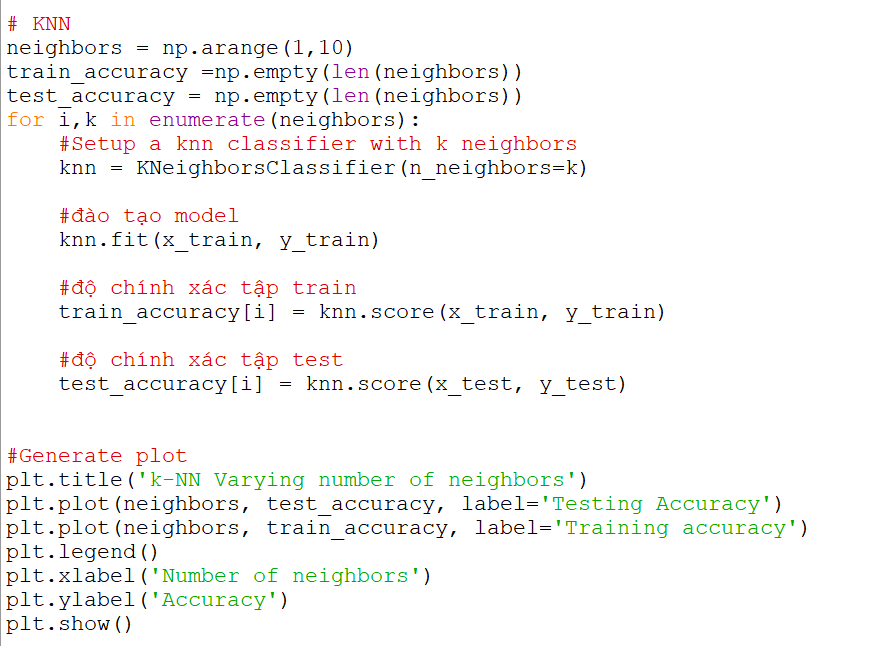


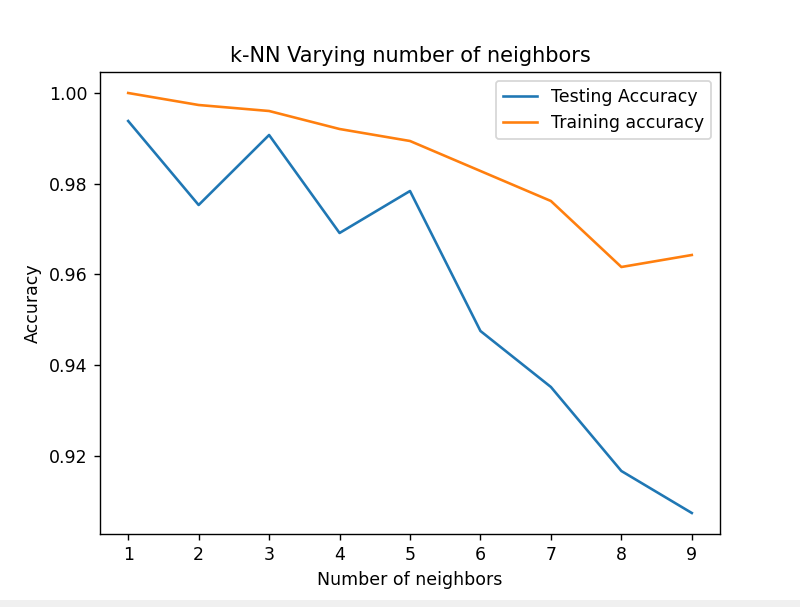
Tạo tập train (80%) và test (20%).



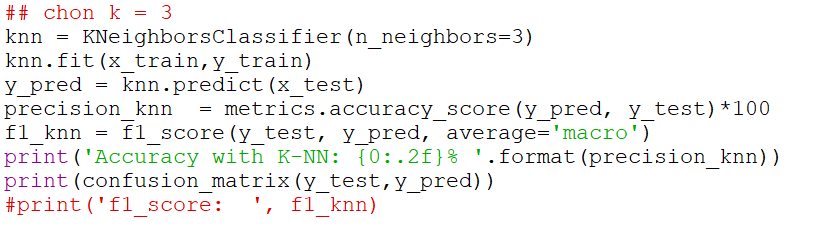


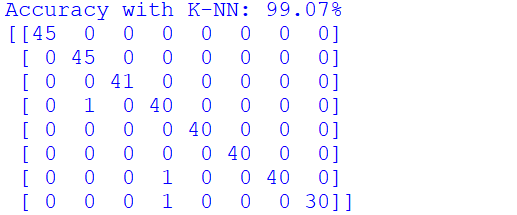
Đào tạo model với tập K từ 1 đến 10



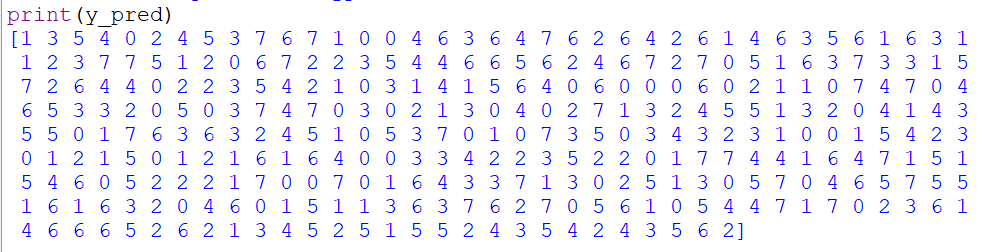


Tại K =3 độ chính xác của tập Test là cao nhất.





Kết quả dự đoán tập test:



**3.2.4 Ưu điểm và nhược điểm**

* *Ưu điểm:*
* Dễ sử dụng và cài đặt
* Việc dự đoán kết quả của dữ liệu mới dễ dàng
* Độ phức tạp tính toán nhỏ
* Độ chính xác cao (tương đối) - khá cao nhưng không cạnh tranh so với các mô hình học tập có giám sát tốt hơn
* *Nhược điểm:*
* Cần thời gian lưu training set , khi dữ liệu training và test tăng lên nhiều sẽ mất nhiều thời gian tính toán
* Tính toán tốn kém (vì thuật toán lưu trữ tất cả dữ liệu đào tạo )
* Yêu cầu bộ nhớ cao
* Giai đoạn dự đoán có thể chậm (với N lớn)

**3.2 Random Forest Classifier**

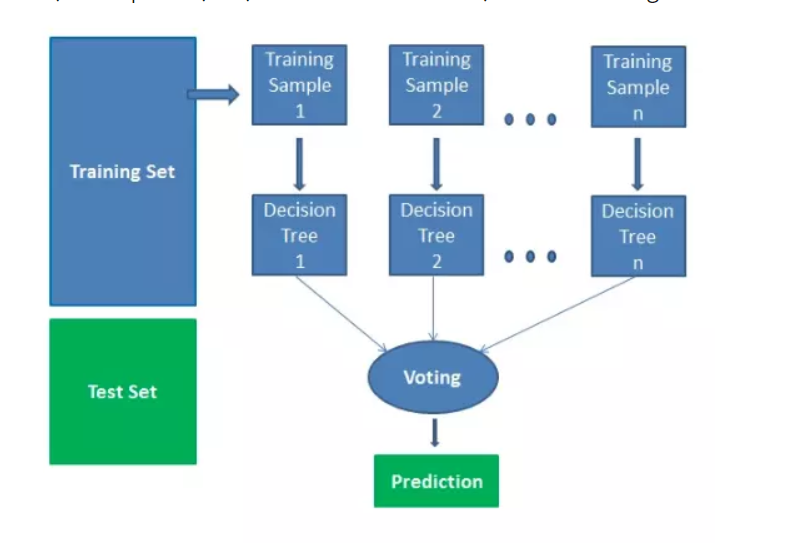
**3.2.1 Định nghĩa**

Random Forests là thuật toán học có giám sát (supervised learning). Nó có thể được sử dụng cho cả phân lớp và hồi quy. Nó cũng là thuật toán linh hoạt và dễ sử dụng nhất. Một khu rừng bao gồm cây cối. Người ta nói rằng càng có nhiều cây thì rừng càng mạnh. Random forests tạo ra cây quyết định trên các mẫu dữ liệu được chọn ngẫu nhiên, được dự đoán từ mỗi cây và chọn giải pháp tốt nhất bằng cách bỏ phiếu. Nó cũng cung cấp một chỉ báo khá tốt về tầm quan trọng của tính năng. Random forests có nhiều ứng dụng, chẳng hạn như công cụ đề xuất, phân loại hình ảnh và lựa chọn tính năng.

**3.3.2 Thuật toán**

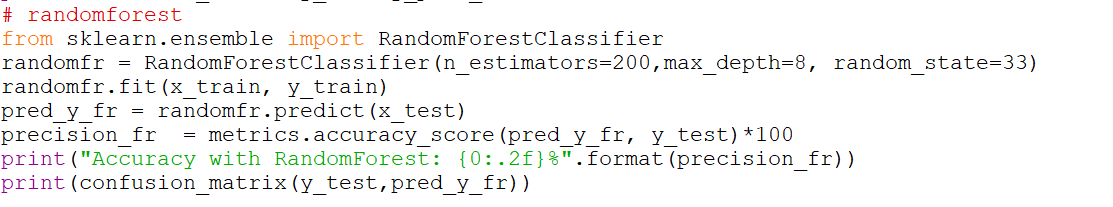
Nó hoạt động theo bốn bước:

1. Chọn các mẫu ngẫu nhiên từ tập dữ liệu đã cho.
2. Thiết lập cây quyết định cho từng mẫu và nhận kết quả dự đoán từ mỗi quyết định cây.
3. Hãy bỏ phiếu cho mỗi kết quả dự đoán.
4. Chọn kết quả được dự đoán nhiều nhất là dự đoán cuối cùng.

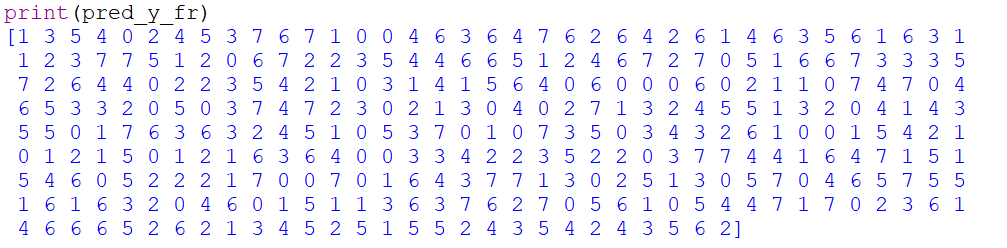


**3.3.3 Tiến hành khai phá dữ liệu**

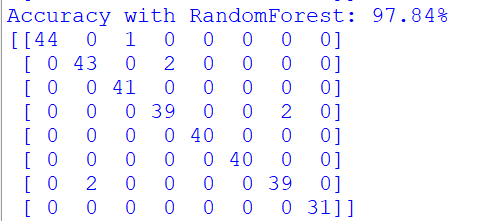
Đạo tạo mô hình :



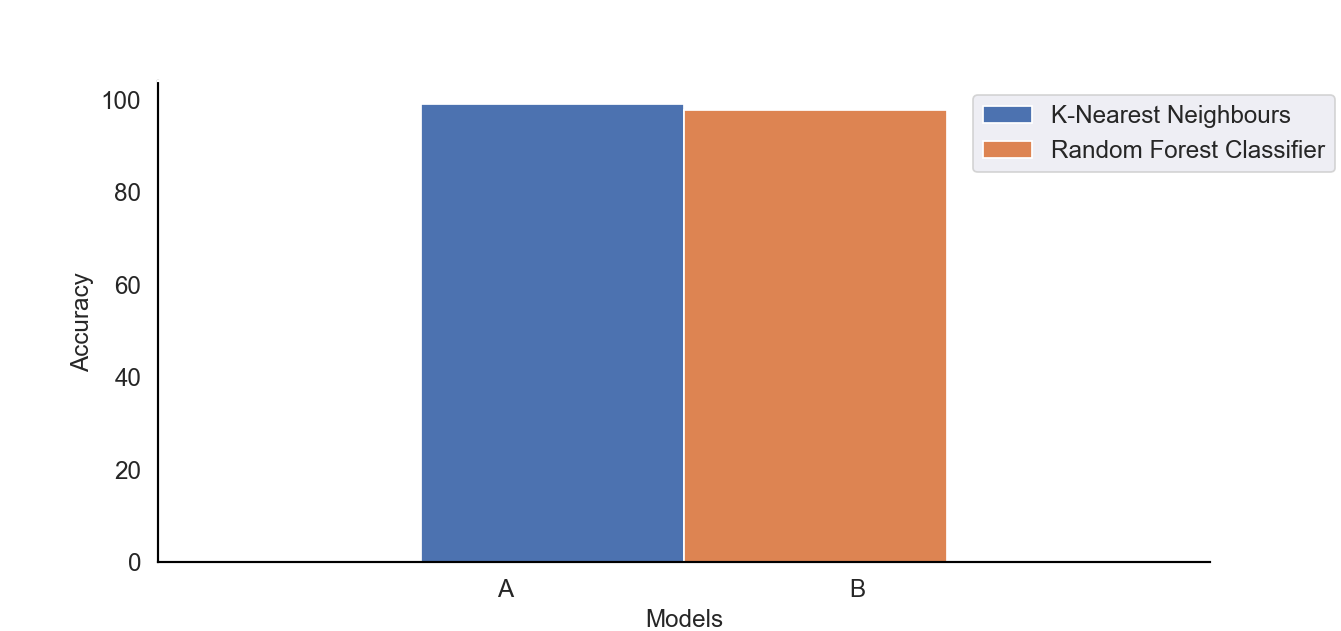
* Tập test sau khi dự đoán bằng Random Forest



* Độ chính xác và ma trận trận lỗi .



Biểu đồ thể hiện độ chính xác giữa KNN với RandomForest



**3.3.4 Ưu và nhược điểm**

Ưu điểm:

* Random Forest có khả năng thực hiện cả hai nhiệm vụ Phân loại và Hồi quy.
* Nó có khả năng xử lý các tập dữ liệu lớn với kích thước cao.
* Nó nâng cao độ chính xác của mô hình và ngăn chặn vấn đề trang bị quá mức.

Khuyết điểm:

* Mặc dù rừng ngẫu nhiên có thể được sử dụng cho cả nhiệm vụ phân loại và hồi quy, nó không phù hợp hơn cho các nhiệm vụ Hồi quy.

**CHƯƠNG 4. NHẬN XÉT ĐÁNH GIÁ, KẾT LUẬN VÀ TÀI LIỆU THAM KHẢO**

* 1. **Nhận xét đánh giá**

Từ 3 giải thuật phân cụm và phân lớp cùng với luật kết hợp đã đưa ra mô hình và các luật để chuẩn đoán về các lớp chuột thí nghiệm và hội chứng down tiếp xúc với điều kiện sợ bối cảnh. Sai số của mô hình là tương đối tốt nên có thể cho kết quả chuẩn đoán tương đối chính xác, có tính ứng dụng thực tế.

* 1. **Kết luận**

Phân cụm và phân lớp cùng với luật kết hợp là 3 lĩnh vực khá quan trọng trong khai phá dữ liệu xu hướng trong tương lai, nó được ứng dụng trong nhiều ngành như y tế, thương mại... Hoàn thành đề tài “*Phân tích so sánh biểu hiện protein của chuột trên bộ dữ liệu Mice Protein Expression Dataset*”.

**Nhóm em đã đạt được một số kết quả như sau:**

HÃ¬nh áº£nh cÃ³ liÃªn quanTìm hiểu tổng quan về khai phá dữ liệu, bài toán phân lớp, phân cụm, phương pháp luật kết hợp để từ đó xây dựng mô hình phân cụm và mô hình phân lớp, khai phá luật kết hợp hỗ trợ chẩn đoán gen chuột.

HÃ¬nh áº£nh cÃ³ liÃªn quanTiền xử lý dữ liệu bằng excel và weka. Xây dựng nên mô hình phân cụm và cây quyết định, khai phá luật kết hợp trên phần mềm weka.

HÃ¬nh áº£nh cÃ³ liÃªn quan Phân cụm dữ liệu

**Tuy nhiên bài tập nhóm vẫn còn một số hạn chế:**

HÃ¬nh áº£nh cÃ³ liÃªn quanViệc thu thập dữ liệu chưa đầy đủ,chi tiết với nhiều bệnh khác nhau mà chỉ tập trung vào bệnh Tăng huyết áp, do vậy các căn bệnh khác chưa hỗ trợ chẩn đoán được bệnh.

HÃ¬nh áº£nh cÃ³ liÃªn quanKết quả dự đoán tương đối cao nhưng vẫn chưa được tốt nhất.

**Hướng phát triển.**

HÃ¬nh áº£nh cÃ³ liÃªn quanXây dựng, cải tiến mô hình chẩn đoán bệnh với phương pháp học máy khác như KNN, Random Forest,…

HÃ¬nh áº£nh cÃ³ liÃªn quanÁp dụng mô hình học máy vào một số lĩnh vực khác và đi vào áp dụng thực.

Trong quá trình hoàn thành bài tập lớn, nhóm em đã cố gắng tìm hiểu và tham khảo các tài liệu liên quan. Tuy nhiên, thời gian có hạn nên chúng em sẽ không tránh khỏi những thiếu sót, rất mong nhận được sự đóng góp ý kiến của quý thầy cô và các bạn để báo cáo và kỹ năng của chúng em ngày được hoàn thiện hơn và có thế áp dụng được trong thực tiễn.

* 1. **Tài liệu tham khảo**

1. TS.Đặng Thị Thu Hiền, (2019), Bài giảng Khai Phá dữ liệu.
2. [https://ongxuanhong.wordpress.com/2015/08/20/tien-xu-ly-du-lieu-horse- colic-dataset/](https://ongxuanhong.wordpress.com/2015/08/20/tien-xu-ly-du-lieu-horse-colic-dataset/)
3. [https://www.slideshare.net/tenzou2411/tiu-lun-khai-ph-d-liu-s-dng- wekazphn-lp-trn-dataset-weather-arff](https://www.slideshare.net/tenzou2411/tiu-lun-khai-ph-d-liu-s-dng-wekazphn-lp-trn-dataset-weather-arff)
4. <https://cuongndh.blogspot.com/p/khai-pha-du-lieu.html>
5. <https://machinelearningcoban.com/2017/01/01/kmeans/>
6. <https://www.youtube.com/watch?v=W-3ZYGmLJ-4&t=270s&ab_channel=DataSciencewithSharan>
7. <https://sti.vista.gov.vn/tw/Lists/TaiLieuKHCN/Attachments/283129/CVv472S52018050.pdf>
8. <https://doc.edu.vn/tai-lieu/do-an-thuat-toan-phan-cum-du-lieu-mo-c-means-va-ung-dung-87358/>
9. https://machinelearningcoban.com/2017/01/01/kmeans/