ỦY BAN NHÂN DÂN TP. HỒ CHÍ MINH

TRƯỜNG ĐẠI HỌC SÀI GÒN

-

**TÊN ĐỀ TÀI**

BÁO CÁO MÔN KHAI PHÁ DỮ LIỆU

TÊN SINH VIÊN: VŨ VĂN THIÊN

MÃ SỐ SINH VIÊN:3118410406

TÊN LỚP: DCT1186

TP. HỒ CHÍ MINH, THÁNG 5 NĂM 2022

# CHƯƠNG I. GIỚI THIỆU

## Bối cảnh của bài toán

Thuật toán k-mean là một trong những thuật toán phân cụm lâu đời nhất và được sử dụng phổ biến nhất. Đó là một điểm khởi đầu tuyệt vời cho những người đam mê Machine Learn(ML), vì sự đơn giản trong việc thực hiện nó.

**Rượu vang** là một loại thức uống có cồn được lên men từ nho. Sự cân bằng hóa học tự nhiên cho phép nho lên men không cần thêm các loại đường, axit, enzym, nước hoặc chất dinh dưỡng khác Men tiêu thụ đường trong nho và chuyển đổi chúng thành rượu và carbon dioxide. Giống nho khác nhau và chủng nấm men khác nhau tạo thành các dạng khác nhau của rượu vang. Các dạng rượu vang nổi tiếng là kết quả của sự tương tác rất phức tạp giữa phát triển sinh hóa của nho, các phản ứng liên quan đến quá trình lên men, cùng với sự can thiệp của con người trong quá trình tổng thể. Là một loại nước uống làm cho con người say, cũng giống như tất cả các loại đồ uống có cồn,thường được sử dụng vì các hiệu ứng say của nó trong suốt lịch sử cho đến ngày hôm nay. Mức độ ảnh hưởng đến thần kinh của rượu vang được thể hiện qua mức độ cồn trong máu.

Để phù hợp cho việc phát triển sản xuất và bảo quản rượu được hiệu quả hơn, cũng như phân ra nhóm rượu để kinh doanh mà ta sẽ tiến hành phân tích đánh giá để chia nhóm rượu cho chúng. Mục tiêu bài toán nhằm xác định nhóm loại rượu cho danh sách rượu được cung cấp.

### Lịch sử.

Thuật ngữ "k-mean" được James MacQueen sử dụng lần đầu tiên vào năm 1967 như một phần trong bài viết của ông về "Một số phương pháp phân loại và phân tích các quan sát đa biến". Thuật toán tiêu chuẩn cũng được sử dụng trong Bell Labs như một phần của kỹ thuật điều chế mã xung vào năm 1957. Nó cũng được xuất bản vào năm 1965 bởi EW Forgy và thường được gọi là phương pháp Lloyd-Forgy.

### Sử dụng K-Means.

* Thuật toán K-means được sử dụng rất nhiều trong thực tế ở các lĩnh vực khác nhau:
* Xử lý ảnh: Phân nhóm các chữ số viết tay, tách vật thể (image segmentation), nén ảnh/dữ liệu (image compression),…
* Các lĩnh vực khác như: phân cụm tài liệu, xác định các khu vực dễ phạm tội, phân khúc khách hàng, phát hiện gian lận bảo hiểm, phân tích dữ liệu giao thông công cộng, phân cụm cảnh báo CNTT,…

### Một số bài toán tiêu biểu sử dụng K-means

* Phân loại tài liệu
* Tối ưu hóa cửa hàng giao hàng
* Xác định địa phương tội phạm
* Phân khúc khách hàng
* Phân tích thống kê Fantasy League
* Phát hiện gian lận
* Tội phạm mạng
* Phân tích chi tiết cuộc gọi

## 2. Định nghĩa bài toán

### Giới thiệu về K-Means

* K-means là một thuật toán phân cụm đơn giản thuộc loại học không giám sát (tức là dữ liệu không có nhãn hay còn hiểu là dữ liệu hoàn toàn chưa biết trước được) và được sử dụng để giải quyết bài toán phân cụm. Ý tưởng của thuật toán phân cụm k-means là phân chia 1 bộ dữ liệu thành các cụm khác nhau. Trong đó số lượng cụm được cho trước là k. Công việc phân cụm được xác lập dựa trên nguyên lý: Các điểm dữ liệu trong cùng 1 cụm thì phải có cùng 1 số tính chất nhất định. Tức là giữa các điểm trong cùng 1 cụm phải có sự liên quan lẫn nhau. Đối với máy tính thì các điểm trong 1 cụm đó sẽ là các điểm dữ liệu gần nhau.
* Thuật toán phân cụm k-means thường được sử dụng trong các ứng dụng cỗ máy tìm kiếm, phân đoạn khách hàng, thống kê dữ liệu,…

### Thuật toán k-means là gì?

* Thuật toán phân cụm k-means là một phương pháp được sử dụng trong phân tích tính chất cụm của dữ liệu. Nó đặc biệt được sử dụng nhiều trong khai phá dữ liệu và thống kê. Nó phân vùng dữ liệu thành k cụm khác nhau. Giải thuật này giúp chúng ta xác định được dữ liệu của chúng ta nó thực sự thuộc về nhóm nào.

### Ý tưởng của thuật toán k-means

* Ý tưởng của thuật toán phân cụm k-means là phân chia 1 bộ dữ liệu thành các cụm khác nhau. Trong đó số lượng cụm được cho trước là k. Công việc phân cụm được xác lập dựa trên nguyên lý: Các điểm dữ liệu trong cùng 1 cụm thì phải có cùng 1 số tính chất nhất định. Tức là giữa các điểm trong cùng 1 cụm phải có sự liên quan lẫn nhau. Đối với máy tính thì các điểm trong 1 cụm đó sẽ là các điểm dữ liệu gần nhau.

1. Khởi tạo K điểm dữ liệu trong bộ dữ liệu và tạm thời coi nó là tâm của các cụm dữ liệu của chúng ta.
2. Với mỗi điểm dữ liệu trong bộ dữ liệu, tâm cụm của nó sẽ được xác định là 1 trong K tâm cụm gần nó nhất.
3. Sau khi tất cả các điểm dữ liệu đã có tâm, tính toán lại vị trí của tâm cụm để đảm bảo tâm của cụm nằm ở chính giữa cụm.
4. Bước 2 và bước 3 sẽ được lặp đi lặp lại cho tới khi vị trí của tâm cụm không thay đổi hoặc tâm của tất cả các điểm dữ liệu không thay đổi.

Trên thực tế, sẽ có 1 vài lưu ý cần phải giải quyết khi áp dụng thuật toán k-means

## 3. Các giải pháp hiện tại/thủ công để giải quyết bài toán

### Bằng mắt nhìn.

Rượu vang cơ bản có 3 màu là đỏ, trắng và hồng, và dựa vào đó chia ra được 3 loại khác nhau

do xưa người ta phân loại nho chỉ có hai màu là đậm và nhạt. "Đậm" ý chỉ vang đỏ, được làm từ những trái nho có vỏ màu tối. "Nhạt" ý chỉ vang trắng, được làm từ những trái nho có vỏ màu sáng. Khái niệm màu sắc khá là đơn giản cho đến khi người ta làm vang trắng từ nho đỏ (không cho ngâm vỏ nho nữa), hai màu sắc đậm và nhạt trở thành khác niệm không chính xác.Sau khi bỏ hai khái niệm trên, người ta dùng cụm từ mới đó là: Red wine (vang đỏ), White wine (vang trắng), Rose wine (vang hồng)

### Nếm thử mùi vị

Theo độ đường, rượu vang chia thành vang ngọt, vang hơi ngọt (Medium Dry) và vang không ngọt. Có nhiều cách gọi vang ngọt khác nhau: Sweet Wine (tiếng Anh), Doux Vin (tiếng Pháp), Dole Wine (tiếng Ý). Còn vang không ngọt có thể gọi: Dry Wine, Brut Vin (tiếng Pháp). Loại vang siêu không ngọt thì gọi là Extra Dry hoặc Extra Bruf.

Vang chia theo độ cồn gồm vang nhẹ, vang trung bình và vang nặng. Nếu bạn thấy một chai vang có độ cồn thấp quá (từ 8 – 10,5% Alc) thì hãy gọi nó là Light Body; độ cồn trung bình (từ 11 – 12,5% Alc) thì gọi là Medium Body; còn độ cồn cao (từ 13 – 15% Alc) thì người ta gọi nó là Full Body.

Dựa vào dữ kiện trên, nếu có chút kiến thức về vang thì có thể nhận ra được và phân loại rượu.

### Ngửi mùi rượu

Từ loại nho cho đến men và kỹ thuật ngâm ủ của nhà sản xuất đã tạo ra thức uống rượu vang là tổng hòa của từ 200 – 250 hợp chất. Chúng tạo thành nhiều tầng hương vị khác nhau bao gồm:

Tầng hương 1: hương trực tiếp của nho

Tầng hương 2: mùi hương được tạo ra từ quá trình lên men

Tầng hương 3: trùm hương tổng thể được tạo ra từ quá trình lão hóa rượu

### Độ bọt của rượu

Dựa vào độ bọt thì chỉ ra được 2 loại là có bọt và không bọt

## 4. Vai trò của khai phá dữ liệu trong việc giải quyết bài toán trên

Vấn đề cốt lõi của khai phá dữ liệu là mô hình thống kê có thể được áp dụng cho hồi quy tuyến tính hoặc logistic. Kết hợp với phân tích dự đoán, từ đó có thể phát hiện ra một loạt các xu hướng, sự bất thường và các vấn đề trước đây mà các công ty có thể sử dụng để kinh doanh tốt hơn.

Không phải ai cũng có đủ kiến thức để phân biệt được loại rượu vang và chất lượng của nó hoặc đôi khi thiếu những vật dụng cần thiết để kiểm tra. Khai phá dữ liệu giúp phát hiện nhanh và gần chính xác loại của các mẫu rượu được cung cấp nhằm giảm chi phí và thời gian cho việc phân loại.

Tuy còn hạn chế nhưng thuật toán K-Means là nền tảng, chỉ ra một hướng đi về

khai phá dữ liệu bằng cách gom cụm và đạt được hiệu quả cao. Không những vậy,

K-Means còn là khởi đầu cho nhiều thuật toán gom cụm ra đời tiếp theo với mục

đích đạt được hiệu quả tối ưu nhất.

## 5. Kết quả dự đoán và ứng dụng của khai phá dữ liệu

1. **Đầu ra dữ liệu**

Với khái niệm thuật toán K-Means ta được biết dữ liệu đầu vào chưa được xác định trước “học không giám sát(tức là dữ liệu không có nhãn)” và sau khi biến đổi , phân tích vô số lần ta được đầu ra là dữ liệu đã được phân cụm thành các cụm có tính chất riêng và các thành phần có nét tương đồng và đặc trưng, từ dữ liệu chưa biết trước để thành nhóm các dữ liệu có điểm chung, nhờ vào đó việc xử lý và truy cập thông tin từ dữ liệu dễ dàng hơn.

Với bài toán này đầu ra cho biết được các mẫu rượu được phân ra theo nhóm và có điểm tương đồng, giúp cho bảo quản và phát triển loại rượu được nhanh chóng tiện lợi hơn.

1. **Ứng dụng của khai phá dữ liệu**

Có nhiều ứng dụng của Data Mining thường thấy như:

* Phân tích thị trường và chứng khoán
* Phát hiện gian lận
* Quản lý rủi ro và phân tích doanh nghiệp
* Phân tích giá trị trọn đời của khách hàng

Với mục đích học thuật bài toán được áp dụng với dữ liệu của các loại rượu vì thế việc chuyển đổi dữ liệu từ rượu để sang bài toán với tính chất học thuật sẽ mang lại hiệu quả không cao, bài toán được phát triển với tính chất nghiên cứu chứ chưa áp dụng được nhiều.

## CHƯƠNG II. MÔ TẢ DỮ LIỆU

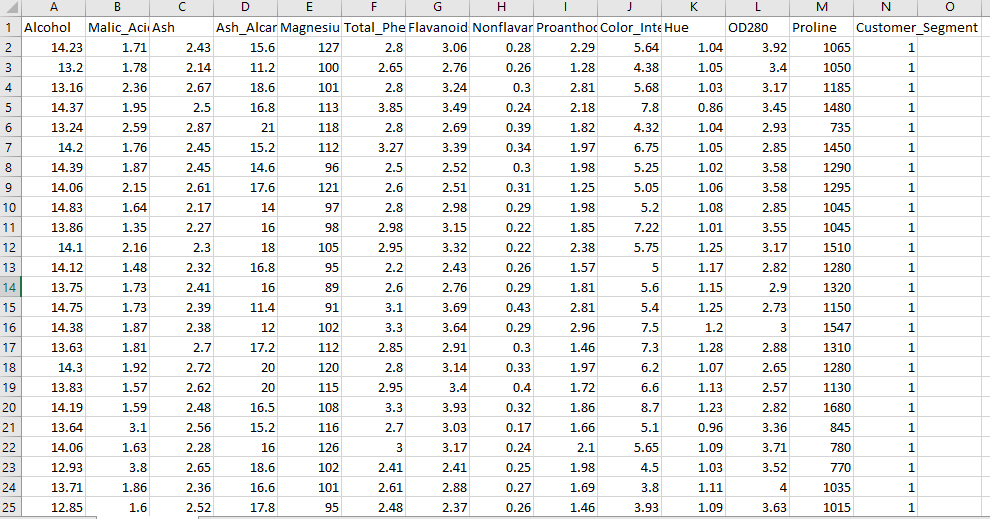
## Kích thước, chiều và trích dẫn nguồn dữ liệu

Để xử lý bài toán K-means tốt nhất và hiệu quả, dữ liệu dùng cho xử lý bài toán cần phải đủ lớn để máy có thể học được bài toán và xử lý cũng như vẽ biểu đồ cho bài toán được chuẩn xác nhất.

Dữ liệu có thể từ vài dòng hoặc có thể lên đến hàng chục ngàn dòng dữ liệu tuỳ vào quy mô bài toán

Ở đây dữ liệu các chất có trong rượu được lấy mẫu và đo từ 179 mẫu rượu khác nhau

được lấy từ <https://archive.ics.uci.edu/ml/machine-learning-databases/wine/>



các thành phần cơ bản tiêu biểu như Alcohol(nồng độ cồn), Malic\_Acid, Hue, độ đậm màu…

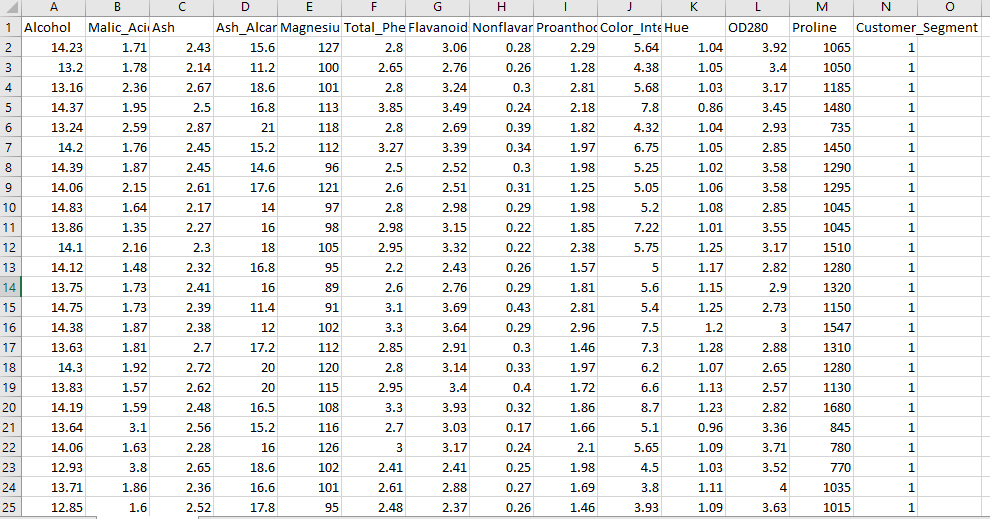
* Dữ liệu cung cấp cần được chính xác và đủ lớn để đảm bảo bài toán không có sai sót hay sự ngụy tạo dữ liệu dẫn đến đầu ra sai lệch. Vì thế nguồn dữ liệu mà ta khai thác hoặc cung cấp phải chuẩn xác để phục vụ cho nghiên cứu dữ liệu, ta có thể thu thập dữ liệu từ khảo sát, tìm kiếm trên internet hoặc báo cáo từ nguồn tin cậy.
* Một số nguồn cung cấp dữ liệu uy tín trên internet :
* [Kaggle: Your Home for Data Science](https://www.kaggle.com/): Kaggle được thành lập với hoạt động chủ yếu là một cộng đồng trực tuyến và dành cho những nhà khoa học dữ liệu cùng mọi đối tượng có thể thực hành học máy. Trong đó có thể hiểu về khoa học dữ liệu chính là một lĩnh vực liên ngành và có sử dụng đến các phương pháp, các quy trình hay thuật toán cùng với hệ thống khoa học công nghệ nhất định nhằm mang đến những kiến thức, những hiểu biết cần thiết có liên quan đến vấn đề cấu trúc và phi cấu trúc. Nguồn data cung cấp từ đây cũng được thu thập từ nhiều nguồn, và đương nhiên cũng đảo bảo đúng đắn.
* [UCI Machine Learning Repository: Data Sets](https://archive.ics.uci.edu/ml/datasets.php):là một trong những nguồn tập dữ liệu lâu đời nhất trên web và là điểm dừng đầu tiên tuyệt vời khi tìm kiếm các tập dữ liệu thú vị. Mặc dù tập dữ liệu do người dùng đóng góp và do đó có các mức độ sạch khác nhau, nhưng đại đa số đều tốt.
* Ngoài ra còn 1 số nguồn khác, tuỳ vào bài toán mà ta sẽ tìm kiếm data ở các kho khác nhau.

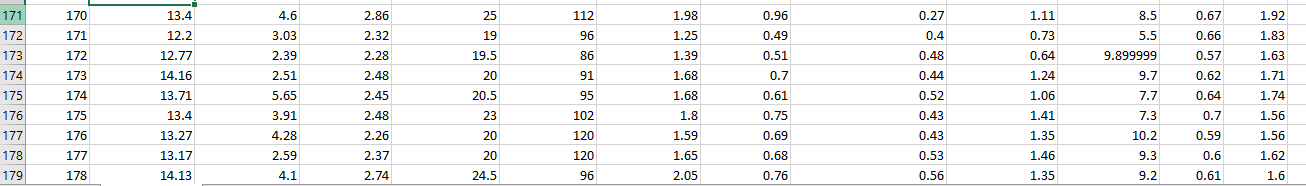
## Các kiểu dữ liệu

* Có rất nhiều kiểu dữ liệu khác nhau, tuỳ vào bài toán mà ta giải quyết, có 3 kiểu dữ liệu phổ biến:
* Dữ liệu dạng số.
* Dữ liệu dạng chữ.
* Kiểu logic.
* Giải quyết bằng giải thuật K-means theo tổng quát thì dữ liệu mà ta tính toán phải là dạng số để có thể phân cụm và vẽ biểu đồ vì thế ta sẽ phải có bước mã hoá nó thành số, hoặc có thể là màu sắc, hình ảnh để có sự phân biệt.
* Để sử dụng được dữ liệu thì phải được qua chuẩn hóa dữ liệu.

## Thống kê mô tả về dữ liệu

Dữ liệu của các mẫu rượu được cung cấp:



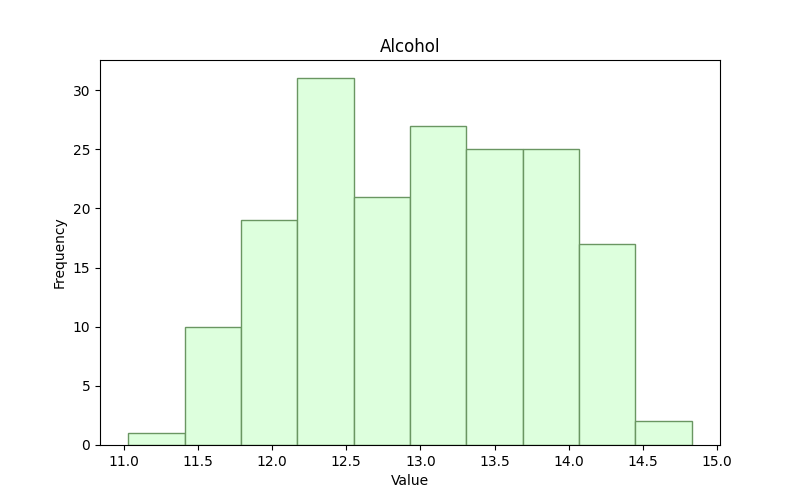


Trong này cung cấp cho ta thông tin về nồng độ :

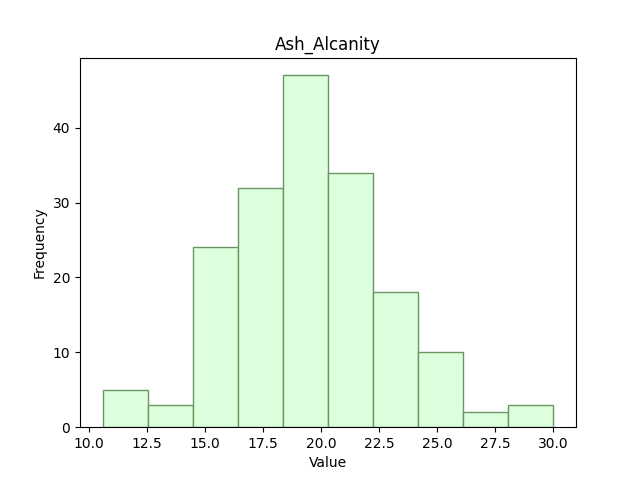
* **Alcohol**: hay còn gọi là Ethanol, hình thành khi trái cây hoặc ngũ cốc được lên men để sản xuất thức uống có cồn.  
   khoảng giá trị từ [11.3;14.83]
* **Malic\_acid**: là một axit dicarboxylic được tạo ra bởi tất cả các sinh vật sống, góp phần tạo nên vị chua của trái cây  
   khoảng giá trị từ [0.74;5.8]
* **Ash**: Hàm lượng tro khoảng giá trị từ [1.36;3.23]
* **Ash\_Alcanity**: (ash alkalinity) độ kiềm của tro khoảng giá trị từ [10.6;30]
* **Magnesium**:Magiê là đồng yếu tố trong hơn 300 hệ thống enzym điều chỉnh các phản ứng sinh hóa đa dạng trong cơ thể, bao gồm tổng hợp protein, chức năng cơ và thần kinh, kiểm soát đường huyết và điều hòa huyết áp  
   khoảng giá trị từ [70;162]
* **Total\_Phenols**: là các hợp chất phenolic là thành phần thực vật quan trọng có đặc tính oxy hóa khử chịu trách nhiệm cho hoạt động chống oxy hóa  
   khoảng giá trị từ [0.98;3.88]
* **Flavanoids** : là chất chuyển hóa thứ cấp có hoạt tính chống oxy hóa, hiệu lực phụ thuộc vào số lượng và vị trí của các nhóm OH tự do   
   khoảng giá trị từ [0.34;5.08]
* **Nonflavonoid Phenols**:Các phenol không phải flavonoid bao gồm một số phân lớp có tầm quan trọng đối với rượu vang, đặc biệt là hydroxycinnamates, stilbene và axit benzoic. Các hydroxycinnamat là các axit phenolic bao gồm một liên kết đôi liên hợp giữa vòng phenolic và nhóm cacboxylat. Ba axit thường được tìm thấy là axit coumaric, caffeic và ferulic   
   khoảng giá trị từ [0.13;0.66]
* **Proanthocyanidins**: Proanthocyanidins là một loại polyphenol được tìm thấy trong nhiều loại thực vật, chẳng hạn như nam việt quất, việt quất và hạt nho. Về mặt hóa học, chúng là flavonoid oligomeric.  
   khoảng giá trị từ [0.41;3.58]
* **Color\_Intensity**: Độ sậm màu của rượu   
   khoảng giá trị từ [1.28;13]
* **Hue**: Tông màu khoảng giá trị từ [0.48;1.71]
* **OD280**:Nguồn Protein khoảng giá trị từ [1.27;4]
* **Proline:** là một axit hữu cơ được phân loại là axit amin tạo protein  
   Khoảng giá trị từ [278;1680]

dựa vào thông tin đó ta sẽ phân cụm các loại rượu đó thành các cụm

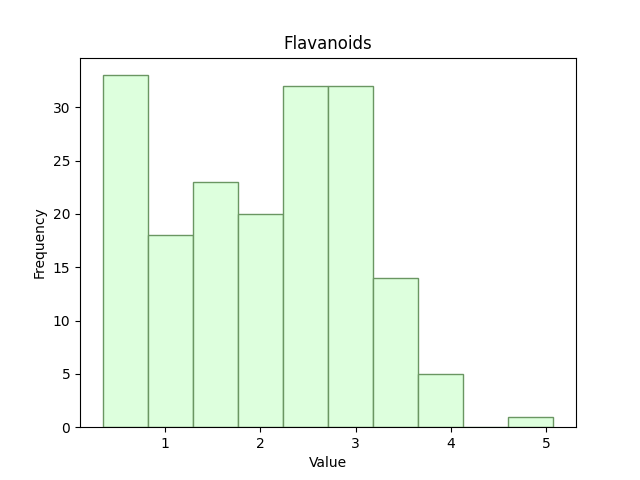
### Sơ đồ mô tả về dữ liệu



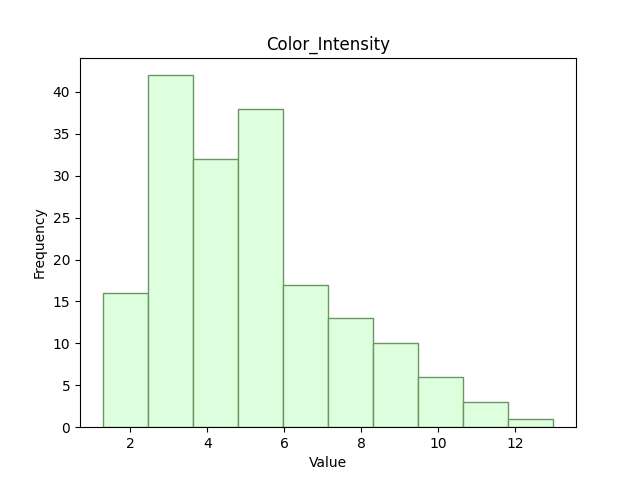
Thống kê nồng độ Cồn



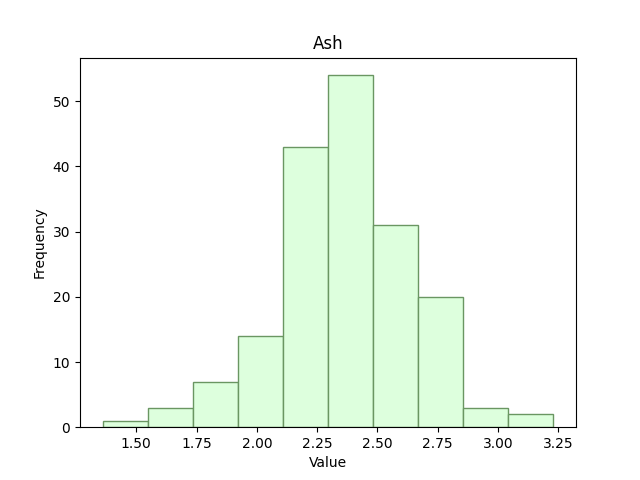
Hàm lượng Ash Alkalinity



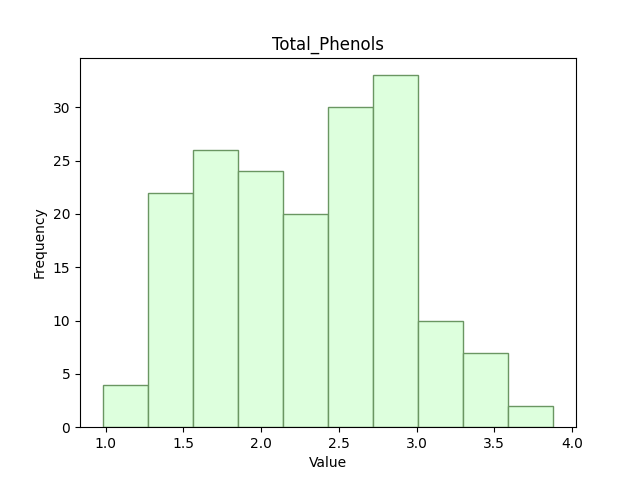
Thống kê Flavanoids



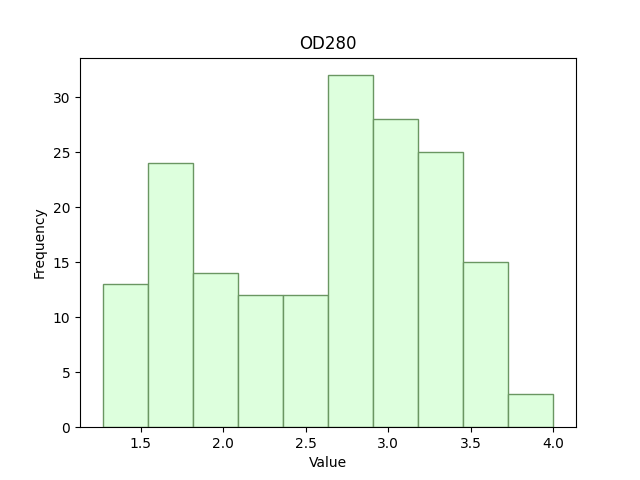
Độ sậm màu



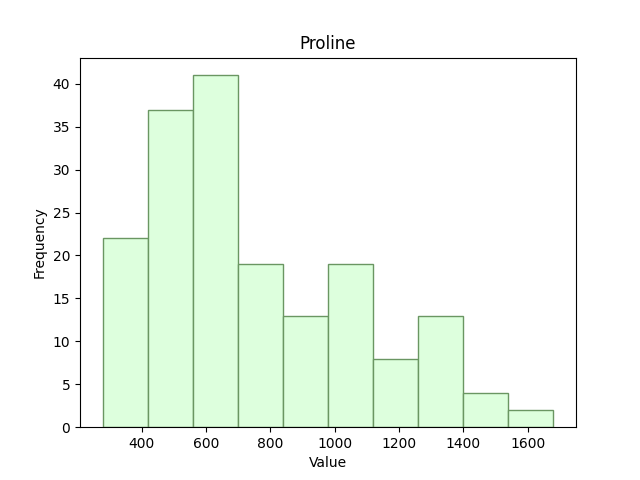
Hàm lượng tro



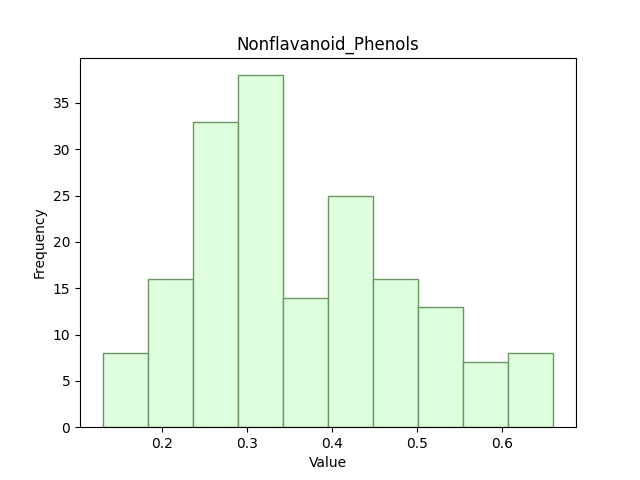
Total Phenols



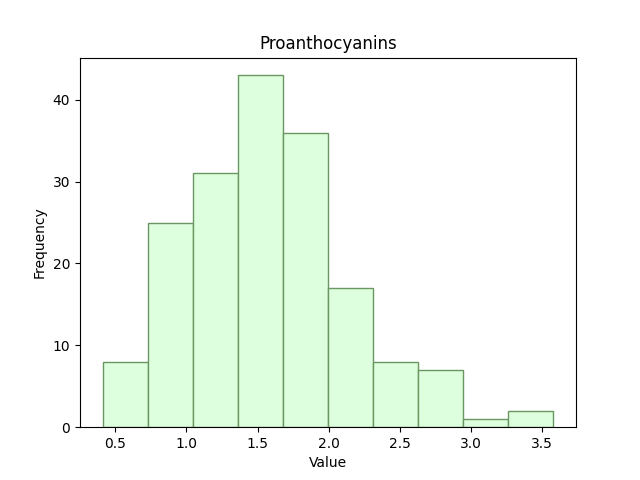
Hàm lượng Protein

****

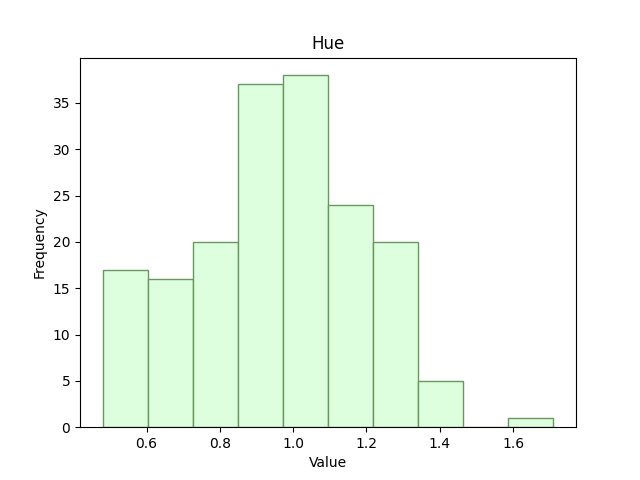
Axit amin

****

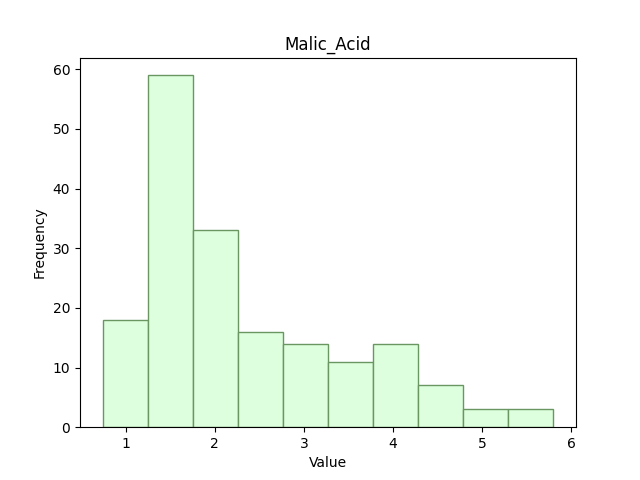
Nonflavanoid\_Phenols

****

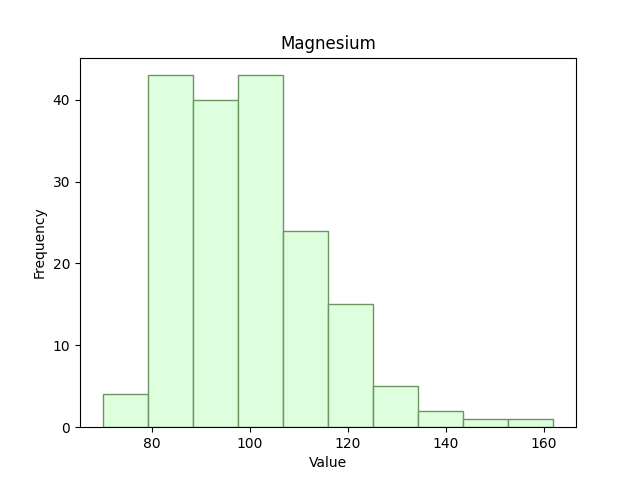
Proanthocyanidins



Tông màu của rượu



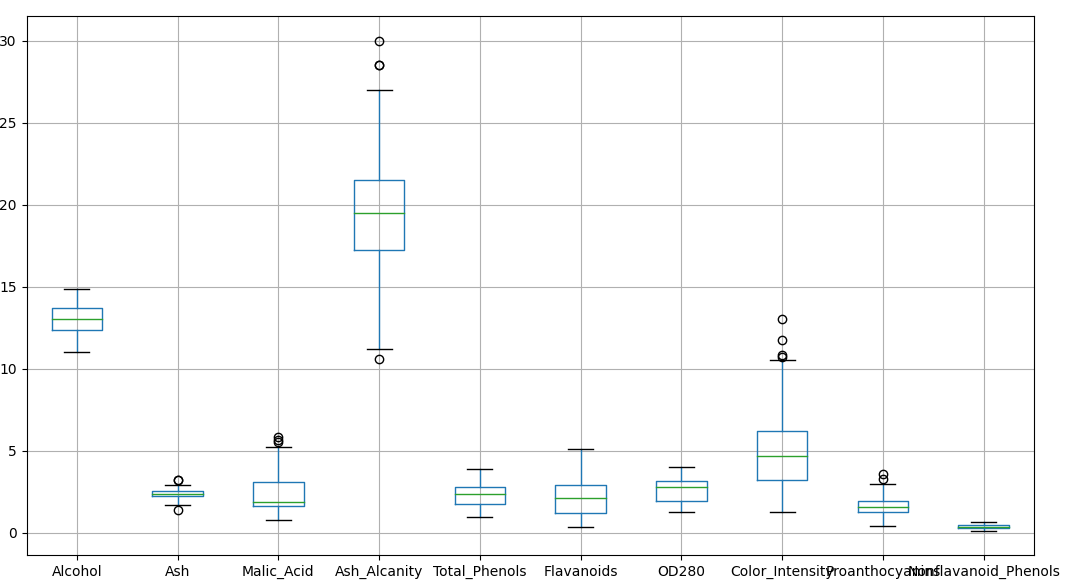
Axit Mallic



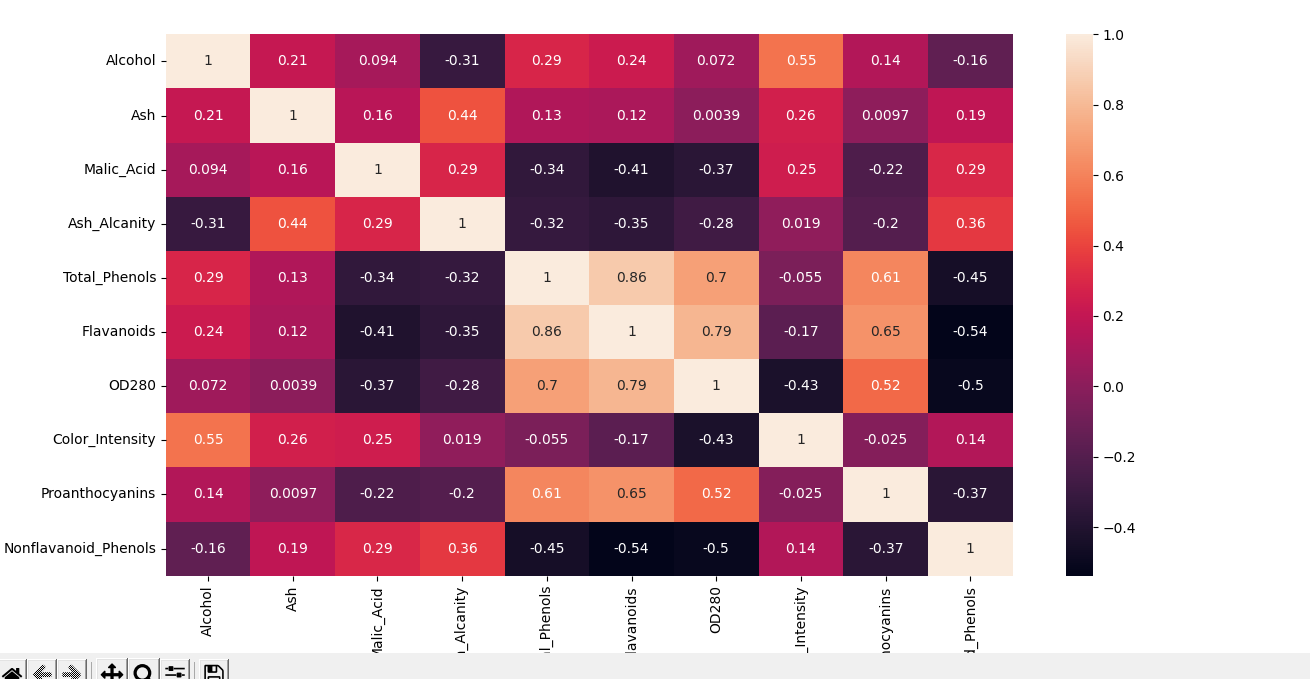
Magie

Sơ đồ Boxplot cho từng thuộc tính

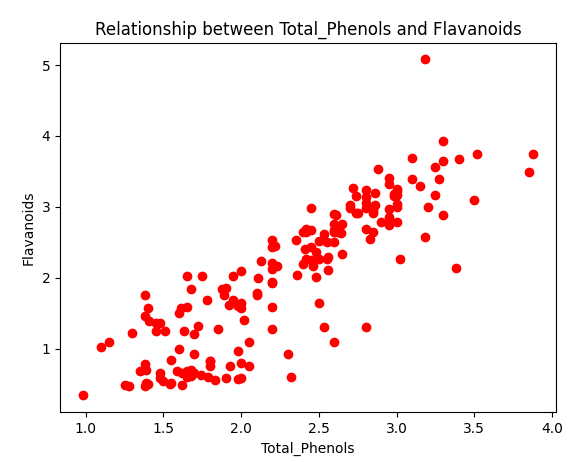
ở đây không liệt kê Proline và Magnesium vì nó khá lớn nên không vẽ chung được



Ma trận tương quan giữa các thuộc tính



Biểu đồ thể hiện sự tương quan giữa Total\_phenols và Flavonoids



ta thấy chúng tăng đều theo nhau xấp xỉ 1 đường thằng đi lên

## Những giả thiết khi thu thập dữ liệu

### Thu thập và sử dụng dữ liệu

Ở trên có 2 nguồn dữ liệu chính để khai thác, ta có thể dựa vào đó để thu thập dữ liệu, hoặc khảo sát từ thực tế để có được dữ liệu tốt nhất.

Để thu thập dữ liệu, trước tiên phải xác định quy mô bài toán, đối tượng cần thu thập, độ lớn của dữ liệu dựa vào đối tượng đó.

Có thể ta không cần phải sử dụng toàn bộ thuộc tính của dữ liệu để dùng cho bài toán, mà ta chọn lọc ra 1 vài thuộc tính cần thiết mà nó ảnh hưởng tới việc phân nhóm.

Ở đây các thành phần của rượu.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Alcohol | Malic\_Acid | Ash | Ash\_Alcanity | Magnesium | Total\_Phenols | Flavanoids | Nonflavanoid\_Phenols |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Proanthocyanins | Color\_Intensity | Hue | OD280 | Proline | Customer\_Segment |

ta không sử dụng thuộc tính Customer\_Segment

nó chỉ cho biết bộ phân khách hàng, không liên quan tới chất lượng , loại rượu

=> Không nhất thiết phải sử dụng toàn bộ nếu thuộc tính đó không thực sự ảnh hưởng tới phân cụm của bài toán.

# CHƯƠNG III. PHƯƠNG PHÁP KHAI PHÁ DỮ LIỆU

## Quy trình khai phá dữ liệu

* Quy trình khai phá dữ liệu thông thường gồm 10 bước:

***1)******Nghiên cứu lĩnh vực***

Ta cần nghiên cứu lĩnh vực cần sử dụng Data mining để xác định được những tri thức ta cần chắt lọc, từ đó định hướng để tránh tốn thời gian cho những tri thức không cần thiết.

***2)******Tạo tập tin dữ liệu đầu vào***

Ta xây dựng tập tin để lưu trữ các dữ liệu đầu vào để máy tính có thể lưu trữ và xử lý.

***3)******Tiền xử lý, làm sạch, mã hóa***

Ở bước này ta tiến hành bỏ bớt những dữ liệu rườm rà, không cần thiết, tinh chỉnh lại cấu trúc của dữ liệu và mã hóa chúng để tiện lợi cho quá trình xử lý.

***4) Rút gọn chiều***

Thông thường một tập dữ liệu có chiều khá lớn sẽ sinh ra một lượng dữ liệu khổng lồ, ví dụ với n chiều ta sẽ có 2^n tổ hợp. Do đó, đây là một bước quan trọng giúp giảm đáng kể hao tổn về tài nguyên trong quá trình xử lý tri thức. Thông thường ta sẽ dùng Rough set (<http://en.wikipedia.org/wiki/Rough_set>) để giảm số chiều.

***5) Chọn tác vụ khai thác dữ liệu***

Để đạt được mục đích ta cần, ta cần chọn được tác vụ khai thác dữ liệu sao cho phù hợp. Thông thường có các tác vụ sau:

\* Đặc trưng (feature)

\* Phân biệt (discrimination)

\* Kết hợp (association)

\* Phân lớp (classification)

\* Gom cụm (clusterity)

\* Xu thế (trend analysis)

\* Phân tích độ lệch

\* Phân tích hiếm

***6) Chọn các thuật giải Khai thác dữ liệu***

***7) Khai thác dữ liệu: Tìm kiếm tri thức***

Sau khi tiến hành các bước trên thì đây là bước chính của cả quá trình, ta sẽ tiến hành khai thác và tìm kiếm tri thức.

***8) Đánh giá mẫu tìm được***

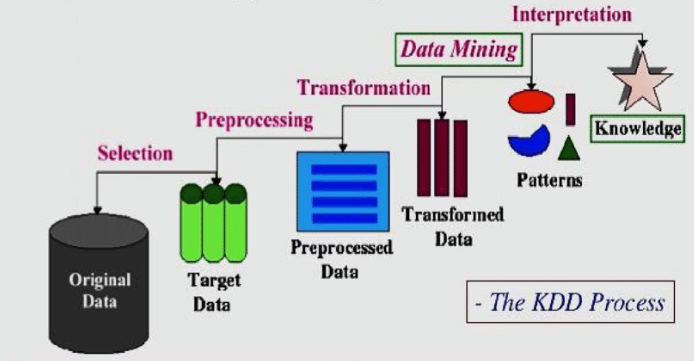
Ta cần đánh giá lại xem trong các tri thức tìm được, ta sẽ sử dụng được những tri thức nào, những tri thức nào dư thừa, không cần thiết.

***9) Biểu diễn tri thức***

Ta biểu diễn tri thức vừa thu thập được dưới dạng ngôn ngữ tự nhiên và hình thức sao cho người dùng có thể hiểu được những tri thức đó.

***10) Sử dụng các tri thức vừa khám phá***

Ta có thể tham khảo tiến trình KDD (Knowledge Discovery in Databases) để hiểu rõ hơn về Khai thác dữ liệu:



Knowledge Discovery in Databases

## Nguyên lý hoạt động của mô hình/thuật toán

Đây là cách thuật toán hoạt động:

Bước 1: Thuật toán tùy ý chọn k điểm làm trung tâm cụm ban đầu (phương tiện).

Bước 2: Mỗi điểm trong tập dữ liệu được gán cho cụm đã đóng, dựa trên khoảng cách Euclide giữa mỗi điểm và mỗi trung tâm cụm.

Bước 3: Mỗi trung tâm cụm được tính toán lại là mức trung bình của các điểm trong cụm đó.

Bước 4: Các bước 2 và 3 lặp lại cho đến khi các cụm hội tụ. Sự hội tụ có thể được xác định khác nhau tùy thuộc vào việc thực hiện, nhưng nó thường có nghĩa là không có sự quan sát nào thay đổi các cụm khi lặp lại bước 2 và 3 hoặc thay đổi không tạo ra sự khác biệt về vật liệu trong định nghĩa cụm.

## Cách cài đặt mô hình, các tham số (nếu có) của mô hình

link download python, phiên bản, hệ điều hành,IDE khuyến nghị,

các thư viện, phiên bản

## Tiêu chí đánh giá mô hình

### ****Hệ số Silhouette:****

Hệ số bóng hay điểm hình bóng là một số liệu được sử dụng để tính toán mức độ tốt của kỹ thuật phân nhóm. Giá trị của nó nằm trong khoảng từ -1 đến 1.

1: Các cụm có nghĩa là cách xa nhau và phân biệt rõ ràng.

0: Các cụm có nghĩa là không quan tâm, hoặc chúng ta có thể nói rằng khoảng cách giữa các cụm là không đáng kể.

-1: Các cụm phương tiện được gán sai cách.

Điểm Silhouette = **(b-a) / max (a, b)**

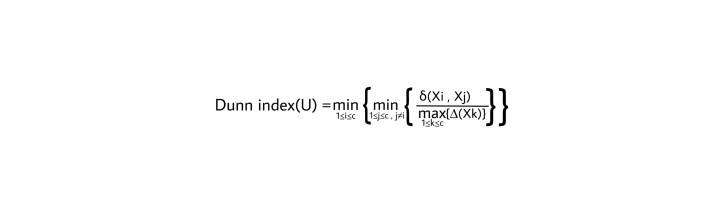
a = khoảng cách trung bình trong cụm tức là khoảng cách trung bình giữa mỗi điểm trong một cụm.

b = khoảng cách trung bình giữa các cụm tức là khoảng cách trung bình giữa tất cả các cụm.

* Điều này có nghĩ là: chỉ số Shlhouette càng lớn thì độ tối ưu càng cao

### Hệ số Dunn’s Index

Chỉ số Dunn (DI) (được giới thiệu bởi JC Dunn vào năm 1974), một số liệu để đánh giá các thuật toán phân cụm, là một sơ đồ tính điểm nội bộ, trong đó kết quả dựa trên chính dữ liệu tổng hợp. Giống như tất cả các chỉ số khác thuộc loại này, mục đích của chỉ mục Dunn này là xác định các tập hợp các cụm nhỏ gọn, với phương sai thấp giữa các thành viên của cụm và được phân tách rõ ràng, trong đó giá trị trung bình của các cụm khác nhau đủ xa so với bên trong phương sai của cụm. Chỉ số Dunn mô tả khoảng cách gần nhất tối thiểu giữa hai cụm bất kỳ chia cho khoảng cách tối đa giữa hai điểm xa nhất trong cụm. Chỉ số Dunn càng nhiều càng tốt. Chỉ số Dunn càng lớn thì chỉ số Dunn càng lớn nếu khoảng cách tối đa giữa hai điểm xa nhất trong bất kỳ cụm nào càng nhỏ. Giá trị của chỉ số Dunn càng cao thì sự tổng hợp càng tốt. Số cụm tối đa hóa chỉ số Dunn được coi là số cụm tối ưu k. Nó cũng có một số nhược điểm. Khi số lượng cụm và kích thước của dữ liệu tăng lên, chi phí tính toán cũng tăng theo.



# CHƯƠNG IV. THỰC NGHIỆM

## Yêu cầu về chương trình

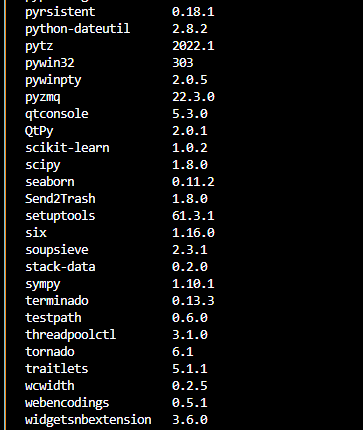
### Ngôn ngữ lập trình, phiên bản

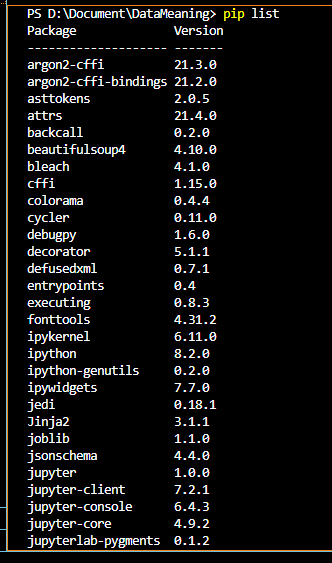
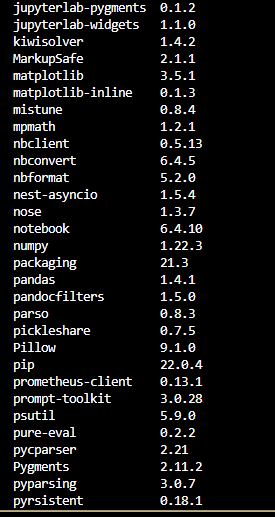
Ngôn ngữ lập trình: Python

Kiểm tra phiên bản Python với câu lệnh: **python -V**

****

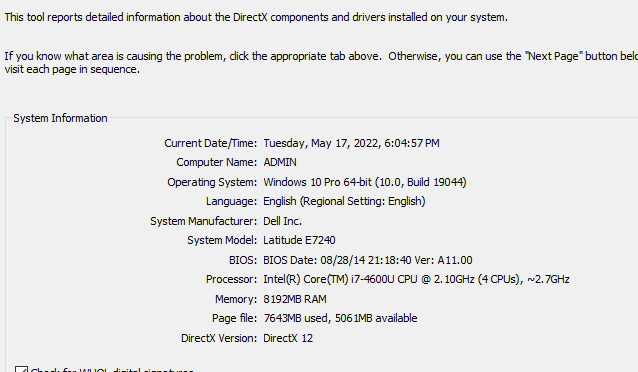
### Thư viện, phiên bản

****Kiểm tra lại các thư viện đã cài đặt cũng như version bằng lệnh **pip list**

****

### Cấu hình máy tính

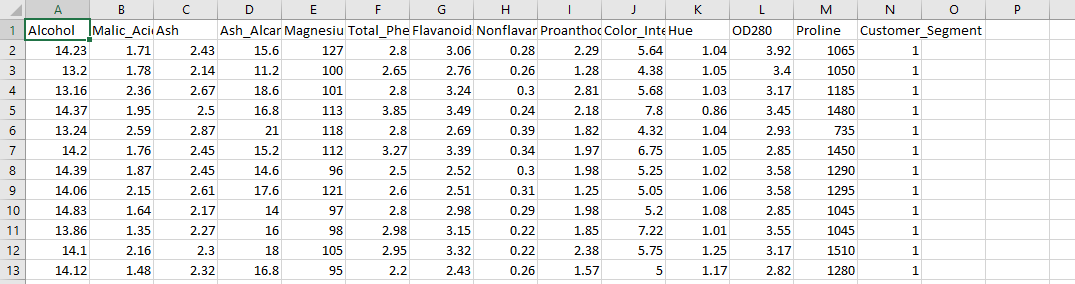
Kiểm tra cấu hình máy bằng lệnh dxdiax



## Mô tả chi tiết các bước khai phá dữ liệu

1. **Chuẩn bị**

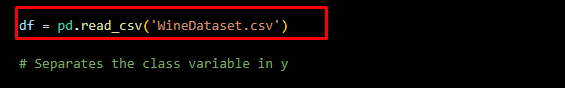
đầu tiên cần chuẩn bị file data(csv)



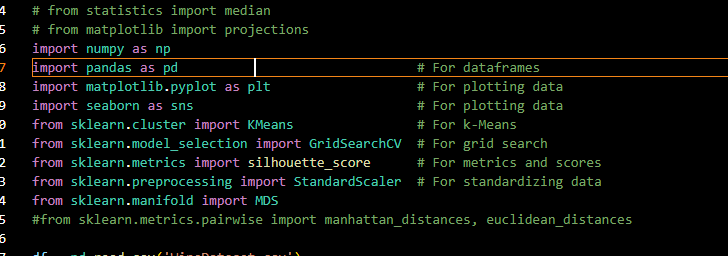
1. **Mô tả dữ liệu qua sơ đồ**

**Input**: file dữ liệu WineDataset.csv

**Output**: xuất ra thông tin file và dữ liệu

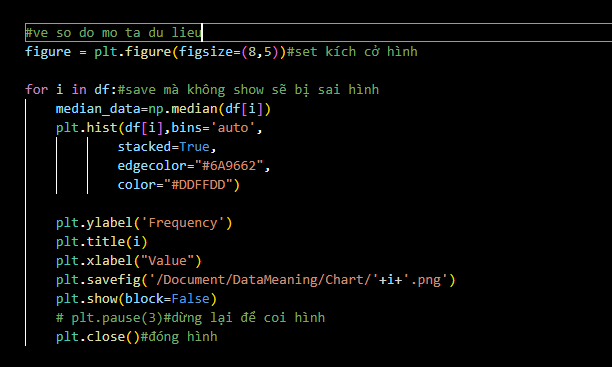
Đầu tiên ta đọc file CSV để lấy dữ liệu

Để sử dụng các hàm và biến số ta cần phải import các thư viện cần thiết vào

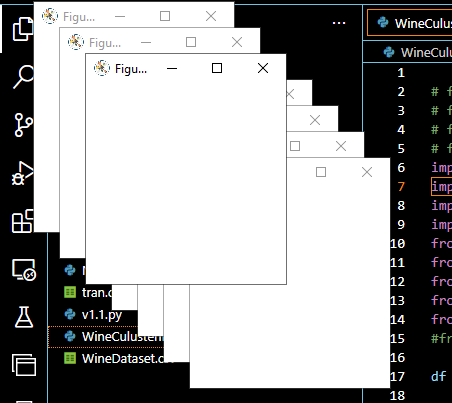


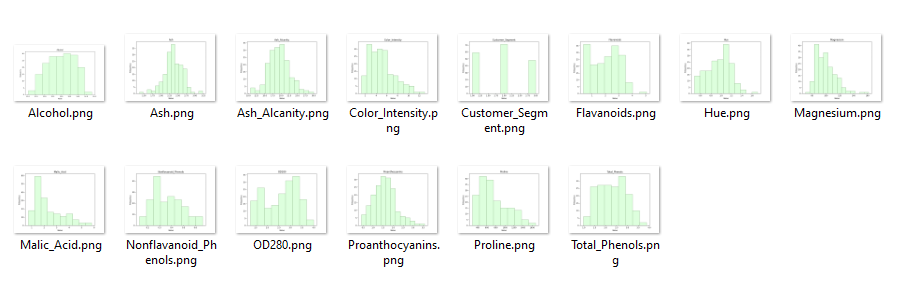
Ta có thể drop cột dữ liệu nếu không cần thiết

Tiếp đến ta vẽ sơ đồ mô tả dữ liệu trong file data ở trên vừa đọc được



Chạy code.



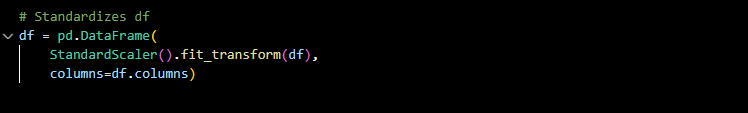
Kết quả sau khi vẽ xong

1. **Chuẩn hoá dữ liệu chuẩn bị cho bài toán**

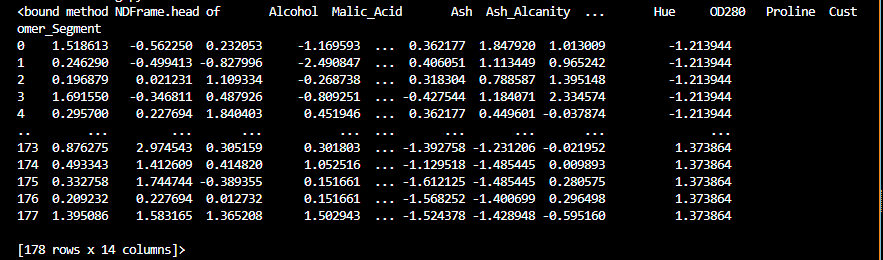
**Input**:dữ liệu từ file csv chưa qua chuẩn hoá

**Output**: dữ liệu đã qua chuẩn hoá, được lưu tạm vào biến X\_transform

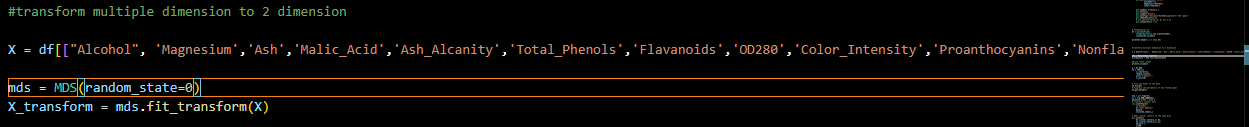
Tiếp theo tiến hành chuẩn hoá dữ liệu, vì dữ liệu không được thống nhất và chênh lệch , nên không thể biểu thị được sự tương quan giữa chúng nên phải qua chuẩn hoá



Hàm chuẩn hoá

Kết quả sau khi chuẩn hoá 

Do số lượng thuộc tính khá nhiều (13 column) nên không thể áp dụng để đưa vào không gian 2-3 chiều trên biểu đồ, vì thế ta phải biến đổi dữ liệu từ không gian n chiều về 2 chiều. sử dụng Hàm MDS để biến đổi thành dạng dữ liệu cho không gian 2 chiều.

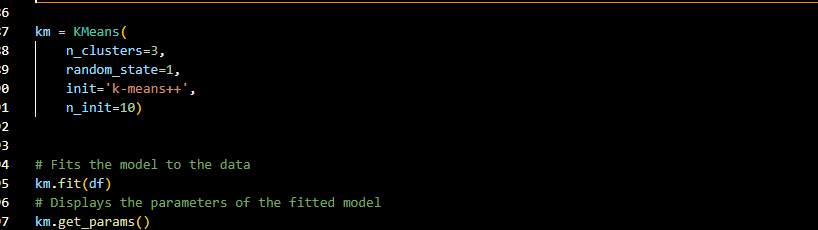


Kết quả sau khi transform 

1. **Chạy K-means**

**Input**:dữ liệu đã qua chuẩn hoá và 3 tâm khởi tạo tự chọn

**Output**:3 vị trí tâm và lựa chọn thuật toán k-means++



-Ta sẽ thiết lập một đối tượng KMeans với các thông số sau:

**n\_clusters**: Tổng số cụm cần tạo.

**random\_state**: Đặt thành một để tái tạo những kết quả này.

**init**: Cách khởi tạo các trung tâm k-means; ta sẽ sử dụng k-mean ++.

**n\_init**: Số lần k-mean sẽ được chạy; mô hình được trả về sẽ có giá trị quán tính nhỏ nhất.

-Một số thuộc tính của đối tượng KMeans,

cũng được sử dụng là:

**cluster**\_**centers**\_: Lưu trữ các trung tâm cụm được phát hiện.

**labels**\_: Nhãn của từng cá thể.

**inertia**: Tổng bình phương khoảng cách của mỗi đối tượng tính từ tâm tương ứng của nó.

**n\_iter**: Số lần lặp chạy để tìm các tâm.

ở đây ta thiết lập mặc định khởi đầu với 3 tâm do ta chọn

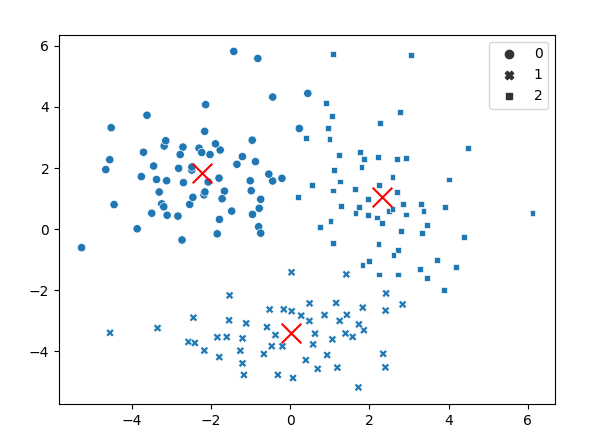
1. **Hiển thị biểu đồ và tâm cụm**

**Input**:dữ liệu qua chuẩn hoá

**Output**:hiển thị biểu đồ mô tả



Kết quả

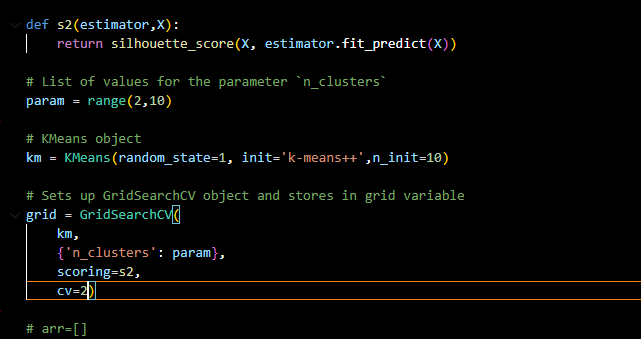


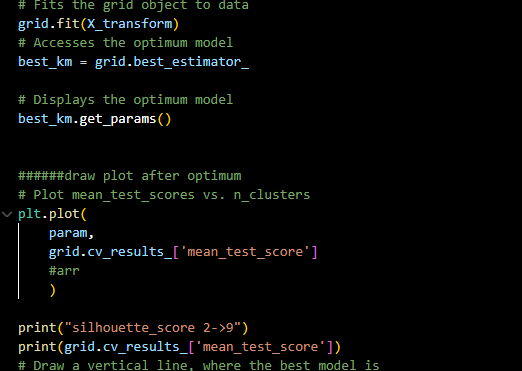
Ta thấy được 3 tâm là 3 dấu X đỏ trên đó và các dữ liệu tương ứng với 3 hình (vuông,tròn,x) sẽ thuộc về các cụm tương ứng

1. **Tối ưu bài toán**

Input:dữ liệu qua chuẩn hoá và 3 tâm hiện tại

Output:kết quả cho số lượng tâm tối ưu



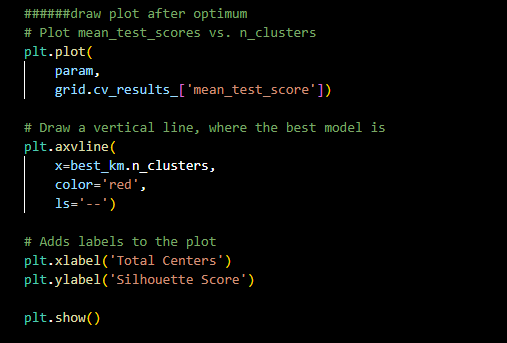


Thách thức chính trong k-means là tìm số lượng cụm tối ưu. Chúng ta có thể thiết lập một GridSearchCV đối tượng để tìm kiếm các tham số tối ưu. Đối với k-Mmeans, chúng ta yêu cầu công cụ ghi điểm tùy chỉnh tính toán giá trị hình bóng cho số lượng các cụm khác nhau được chỉ định bởi n\_clusters. Người ghi điểm tùy chỉnh được gọi s2() trong đoạn mã dưới đây, nơi nó sử dụng silhouette\_score()từ sklearn.metricsthư viện để tính điểm cho một ví dụ X.

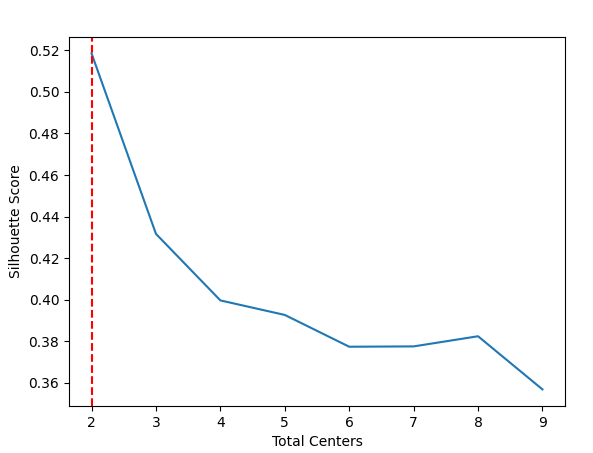
1. **Tìm số lượng cụm tối ưu nhất**

**Input**:số lượng tâm tối ưu

**Output**:hiển thị dạng biểu đồ



Tính toán và vẽ sơ đồ thể hiện số tâm tối ưu

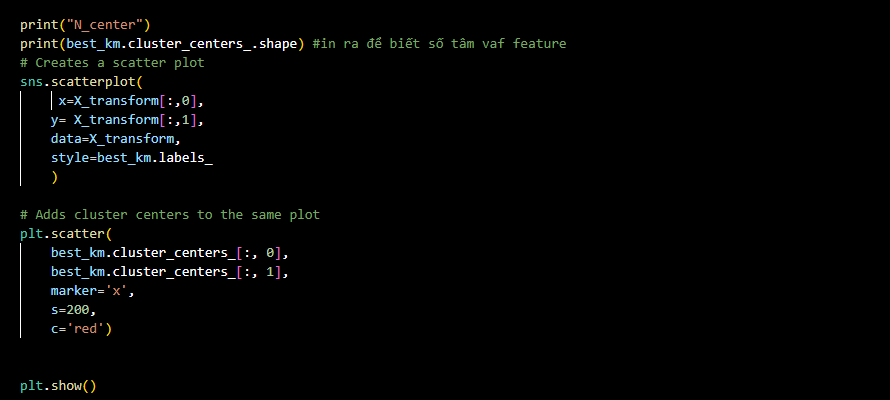


Như hình ta được tối ưu là 2 tâm cụm thay vì 3 tâm như ban đầu

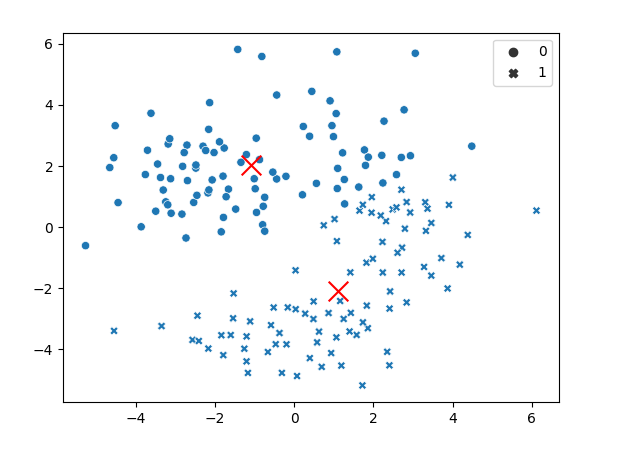
1. **Sơ đồ mô tả tâm cụm tối ưu**

Input:số lượng tâm tối ưu

Output:biểu đồ thể hiện phân cụm ở dạng tâm tối ưu nhất



Kết quả sau khi chạy

=> Ta có kết quả là 2 tâm với 2 hình dạng dữ liệu riêng biệt đặc trưng cho mỗi tâm

# CHƯƠNG V. KẾT QUẢ

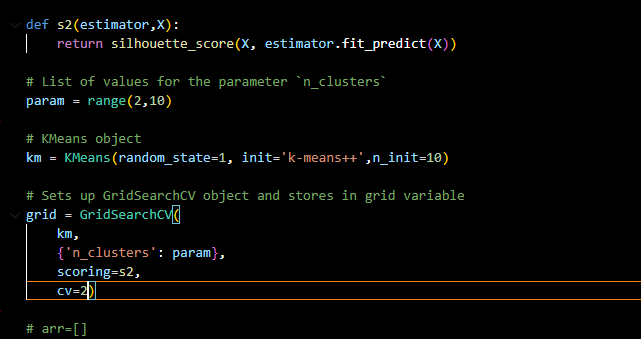
## Kết quả khai phá dữ liệu

### Dựa trên Tiêu chí đánh giá mô hình.

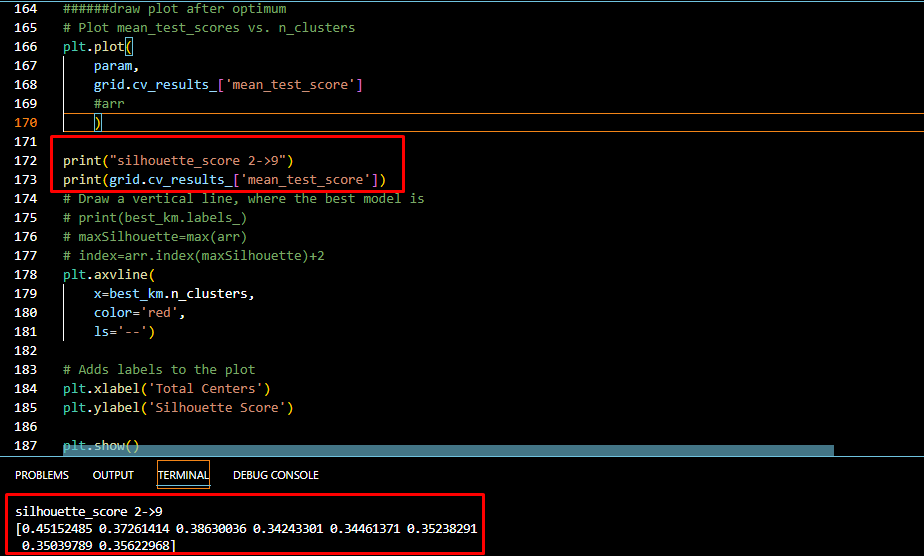
Ta có:

Điểm **Hệ số Silhouette:**

**ở đây có hàm tính toán hệ số Silhouette với random\_state=1**

****

Ta in ra chi tiết hệ số với từng số cụm để phân tích tối ưu



ở đây ta được

**silhouette\_score 2->9**

**[0.51851404 0.43170662 0.39970303 0.39271545 0.37742574 0.3775836**

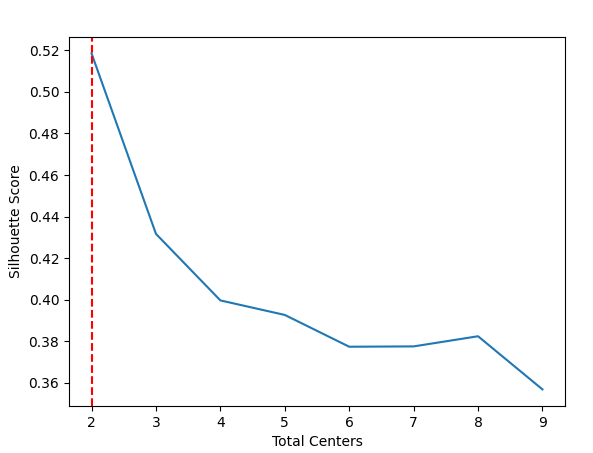
**0.38246797 0.35690198]**

Tương đương với

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **silhouette\_score** | **0.51851404** | **0.43170662** | **0.39970303** | **0.39271545** | **0.37742574** | **0.3775836** | **0.38246797** | **0.35690198** |
| Số tâm cụm | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |

Như vậy ta thấy được tại n=2 ta được điểm số cao nhất so với ban đầu ta chọn n=3

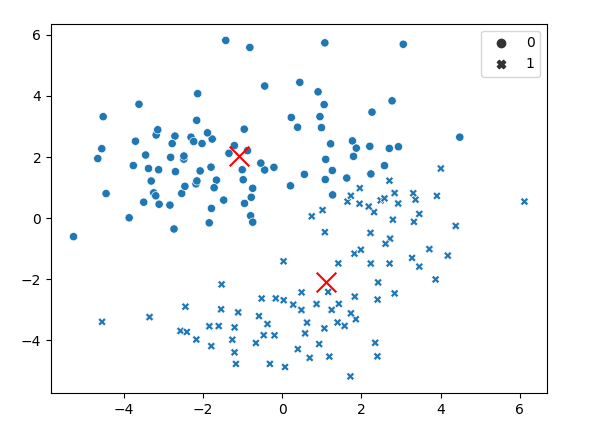
Vậy thuật toán chọn n=2 được coi là tối ưu nhất cho bài toán này

Sơ đồ mô tả

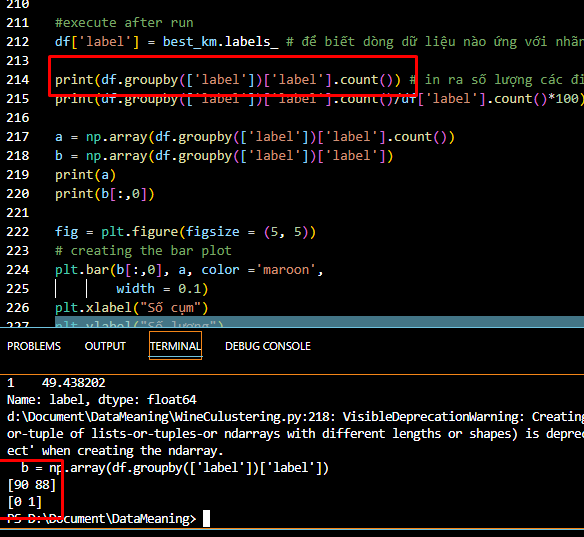
### Mô hình có chính xác hay không?

Sau khi được kết quả tối ưu ta thống kê phần tử của từng cụm dữ liệu

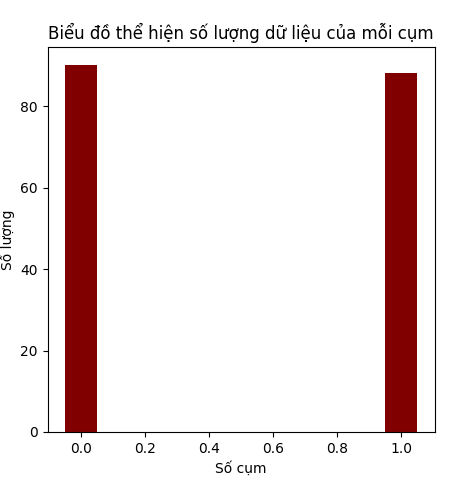
Kết quả sau khi chạy

Ta được 2 cụm như sau 

Sau khi có được số tâm cụm, ta kiểm tra thành viên trong tâm cụm đó



Như trên ta được tâm cụm (0) có 90 phần tử, tâm cụm (1) có 89 phần tử



## So sánh kết quả thực tế với kết quả dự đoán

### Chất lượng của dữ liệu

So sánh số liệu điểm số **silhouette** ta được tối ưu là 6 tâm cụm, với giá trị max là **0.51851404**

Như tiêu chí đánh giá với điểm số

**silhouette** =0 tương đương các điểm gần nhau, không có sự tách biệt rõ ràng

**silhouette** =1 có sự phân biệt rõ ràng, khoảng cách đủ lớn

**silhouette** =-1 đặt sai tâm cụm

vậy giá trị max là 0.51851404 so với 1 là đủ chấp nhận được

### Nguyên lý hoạt động của thuật toán có giải quyết được vấn đề hay không?

Bài toán trên thông qua thuật toán đã phân cụm được thành 2 cụm rượu vang, như ban đầu ta mong muốn phân cụm được thành 3 ,và kết quả ta về cho 2 phân cụm, vì thế bài toán cho a kết quả không được như mong muốn => kết quả so với dự đoán giảm còn 2 cụm.

Bài toán cho kết quả thực tế sai với dự đoán

# CHƯƠNG VI. THẢO LUẬN VÀ KẾT LUẬN

## Khả năng ứng dụng của giải pháp/mô hình

Về thực tế, mô hình khó áp dụng cho người tiêu dùng bình thường.Về người dùng bình thường, khi đi mua ở cửa hàng, siêu thị, chúng đã được dán nhãn, phân loại rất rõ ràng, chi 1 số ít không được dán nhãn và không phân loại được. Hơn thế nữa, có thể dùng các phương pháp thủ công cũng có thể đoán được phần nào về loại rượu. Vì thế, mô hình có thể được ứng dụng để phân rượu trong môi trường nghiên cứu và rèn luyện khai phá dữ liệu.

Kết luận: khả năng ứng dụng của mô hình không cao, không có tính ứng dụng.

## Ưu điểm – nhược điểm của giải pháp/mô hình

* Ưu điểm

+Tỉ lệ chính xác cao

+Tốc độ dự đoán nhanh

+Dữ liệu đều ở dạng số, dễ dàng chuyển hoá và chuẩn hoá

* Nhược điểm

+Dữ liệu đầu vào có nhiều thuộc tính, gây khó khăn cho quá trình nhập thông tin của rượu.

+Việc phân tích các chất trong rượu, cần phải có công cụ và kiến thức chuyên môn để nhận biết.

+Dữ liệu cần phải qua biến đổi toạ độ, gây khó khăn cho việc ước lượng và thể hiện quan hệ

+Chưa xử lý được những dữ liệu thực tế, những giá trị mới của thuộc tính nếu có sẽ gây ra sai sót trong khả năng phán đoán.

## Đề xuất

Những đề xuất để cải tiến mô hình:

- Chuyển đổi dữ liệu đầu vào thành hình ảnh để tiện nhập liệu và phù hợp với công nghệ hiện nay.

- Thay đổi khởi tạo của bài toán bằng các thuộc tính khác nhau và kiểm tra, so sánh kết quả để chọn giải pháp tối ưu

- Dữ liệu cần lớn hơn nữa để đo đọ chính xác được chuẩn hơn.

-Thường xuyên biểu thị bằng biểu đồ để thể hiện sự tương quan rõ hơn

# Tài liệu tham khảo

<https://machinelearningcoban.com/2017/01/01/kmeans/>

<https://nguyenvanhieu.vn/thuat-toan-phan-cum-k-means/#gioi-thieu-ve-k-means>

[Silhouette Coefficient. This is my first medium story, so… | by Ashutosh Bhardwaj | Towards Data Science](https://towardsdatascience.com/silhouette-coefficient-validating-clustering-techniques-e976bb81d10c)

[Scikit Learn - Introduction (tutorialspoint.com)](https://www.tutorialspoint.com/scikit_learn/scikit_learn_introduction.htm#:~:text=Scikit-learn%20%28Sklearn%29%20is%20the%20most%20useful%20and%20robust,dimensionality%20reduction%20via%20a%20consistence%20interface%20in%20Python.)

[scikit-learn: machine learning in Python — scikit-learn 1.1.0 documentation](https://scikit-learn.org/stable/index.html)

[Guide to Multidimensional Scaling in Python with Scikit-Learn (stackabuse.com)](https://stackabuse.com/guide-to-multidimensional-scaling-in-python-with-scikit-learn/?fbclid=IwAR1qSso12pVlIyMCJvkfjn0iQv5Hyo4IJ2qanq0iLGflzQVnsTvpXHfSRdc)

[Matplotlib Tutorial (w3schools.com)](https://www.w3schools.com/python/matplotlib_intro.asp)

[sklearn.metrics.silhouette\_score — scikit-learn 1.1.0 documentation](https://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html)

[Tutorial: Clustering wines with k-means | Kaggle](https://www.kaggle.com/code/xvivancos/tutorial-clustering-wines-with-k-means/report)

[UCI Machine Learning Repository: Wine Data Set](https://archive.ics.uci.edu/ml/datasets/Wine)