ỦY BAN NHÂN DÂN THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC SÀI GÒN
KHOA CÔNG NGHỆ THÔNG TIN

TIỂU LUẬN

TÊN HỌC PHẦN: PHÂN TÍCH DỮ LIỆU

ĐỀ TÀI: ỨNG DỤNG THUẬT TOÁN K-MEANS CLUSTERING ĐỂ PHÂN CỤM RƯỢU VANG ĐỎ

Giảng viên hướng dẫn : Trịnh Tấn Đạt

Lớp : DCT121C2

Sinh viên thực hiện : Nguyễn Hoàng Bảo Huy - 312141086
Huỳnh Lê Trung Hiếu - 3121411070
Nguyễn Hữu Đức - 3121411058

Thành phố Hồ Chí Minh, tháng 05 năm 2024

Lời cảm ơn

Chúng em xin gửi lời cảm ơn chân thành đến thầy về môn học Phân tích dữ liệu. Thầy đã tận tâm hướng dẫn và truyền đạt kiến thức với sự kiên nhẫn và sự chỉ dẫn tỉ mỉ của Thầy trong suốt quá trình học tập. Những bài giảng, bài tập và phản hồi của Thầy đã giúp em hiểu sâu hơn về môn học phân tích dữ liệu và áp dụng kiến thức vào dự án này. Dù thời gian học môn này không dài nhưng em tin rằng kiến thức mà em thu được sẽ tiếp tục hỗ trợ em trong hành trình nghề nghiệp và học tập của mình. Em rất biết ơn vì sự hỗ trợ của Thầy.

Với lòng biết ơn chân thành, em xin gửi tới thầy Trịnh Tấn Đạt lời chúc sức khỏe và thành công.

Mục lục

Danh mục hình ảnh

Lời nói đầu

Công nghệ ngày càng phổ biến và không ai có thể phủ nhận được tầm quan trọng và những hiệu quả mà nó đem lại cho cuộc sống chúng ta. Bất kỳ trong lĩnh vực nào, sự góp mặt của trí tuệ nhân tạo sẽ giúp con người làm việc và hoàn thành tốt công việc hơn. Và gần đây, một thuật ngữ "machine learning" rất được nhiều người quan tâm. Thay vì phải code phần mềm với cách thức thủ công theo một bộ hướng dẫn cụ thể nhằm hoàn thành một nhiệm vụ đề ra thì máy sẽ tự "học hỏi" bằng cách sử dụng một lượng lớn dữ liệu cùng những thuật toán cho phép nó thực hiện các tác vụ.

Đây là một lĩnh vực khoa học tuy không mới, nhưng cho thấy lĩnh vực trí tuệ nhân tạo đang ngày càng phát triển và có thể tiến xa hơn trong tương lai. Đồng thời nó được xem là một lĩnh vực "nóng" và dành rất nhiều mối quan tâm để phát triển nó một cách mạnh mẽ, bùng nổ hơn.

Hiện nay, việc quan tâm machine learning càng ngày càng tăng lên là vì nhờ có machine learning giúp gia tăng dung lượng lưu trữ các loại dữ liệu sẵn, việc xử lý tính toán có chi phí thấp và hiệu quả hơn rất nhiều.

Những điều trên được hiểu là nó có thể thực hiện tự động, nhanh chóng để tạo ra những mô hình cho phép phân tích các dữ liệu có quy mô lớn hơn và phức tạp hơn đồng thời đưa ra những kết quả một cách nhanh và chính xác hơn. Chính sự hiệu quả trong công việc và các lợi ích vượt bậc mà nó đem lại cho chúng ta khiến machine learning ngày càng được chú trọng và quan tâm nhiều hơn. Vì vậy chúng em đã chọn đề tài "Ứng dụng thuật toán K-Means Clustering để phân cụmrượu vang đỏ.

Chúng em xin chân thành gửi lời cảm ơn tới thầy Trịnh Tấn Đạt đã tận tình giảng dạy, truyền đạt cho chúng em những kiến thức cũng như kinh nghiệm quý báu trong suốt quá trình học. Thầy đã giúp đỡ, trực tiếp hướng dẫn trong suốt quá trình học tập của chúng em.

Chương 1. Mở đầu

1.1 Giới thiệu

1.1.1 Nhu cầu thực tế

Trong thời đại ngày nay, với sự phát triển của xã hội, nhu cầu tìm hiểu về chất lượng của rượu vang đang trở nên ngày càng quan trọng hơn. Đối với nhiều người, rượu vang không chỉ là một loại đồ uống mà còn là một phần của văn hóa ẩm thực và cảm nhận sự thăng hoa của cuộc sống.

Tuy nhiên, với sự đa dạng của loại rượu vang trên thị trường, việc đánh giá chất lượng và phân biệt giữa các loại rượu vang trở nên khó khăn đối với người tiêu dùng. Điều này dẫn đến nhu cầu cần có một công cụ hỗ trợ, và trong trường hợp này, đó là một hệ thống thông minh có khả năng phân tích chất lượng của rượu vang.

Hệ thống này có thể hoạt động dựa trên dữ liệu thu thập được từ nhiều nguồn khác nhau như thông tin về nơi sản xuất, loại giống nho, điều kiện thời tiết, và các chỉ số hóa học của rượu vang. Sử dụng thuật toán học máy, hệ thống có thể phân tích các yếu tố này để đưa ra đánh giá về chất lượng của từng loại rượu vang.

Với hệ thống này, người tiêu dùng sẽ không cần phải dựa vào kiến thức chuyên môn sâu rộng về rượu vang để đưa ra quyết định mua hàng. Thay vào đó, họ có thể dễ dàng truy cập vào thông tin phân tích về chất lượng của rượu vang mà họ quan tâm, giúp họ lựa chọn sản phẩm phù hợp với sở thích và nhu cầu của mình một cách tự tin và hiệu quả hơn.

1.1.2 Bài toán phân tích chất lượng rượu vang đỏ

Bài toán phân loại chất lượng của rượu vang đỏ đưa ra các tập dữ liệu là các thông số của từng mẫu rượu vang đỏ để phân loại chất lượng của chúng

- Giá trị input: thông tin, đặc tính của mỗi mẫu rượu
- Giá trị output: tên của cụm chúng được phân vào

1.1.3 Yêu cầu

- Lấy dữ liệu mô tả đặc tính rượu
- Trích chọn đặc trưng từ tập dữ liệu lấy được
- Xử lý, làm sạch dữ liệu
- Tiến trình phân cụm
- Dữ liệu hóa đồ thị

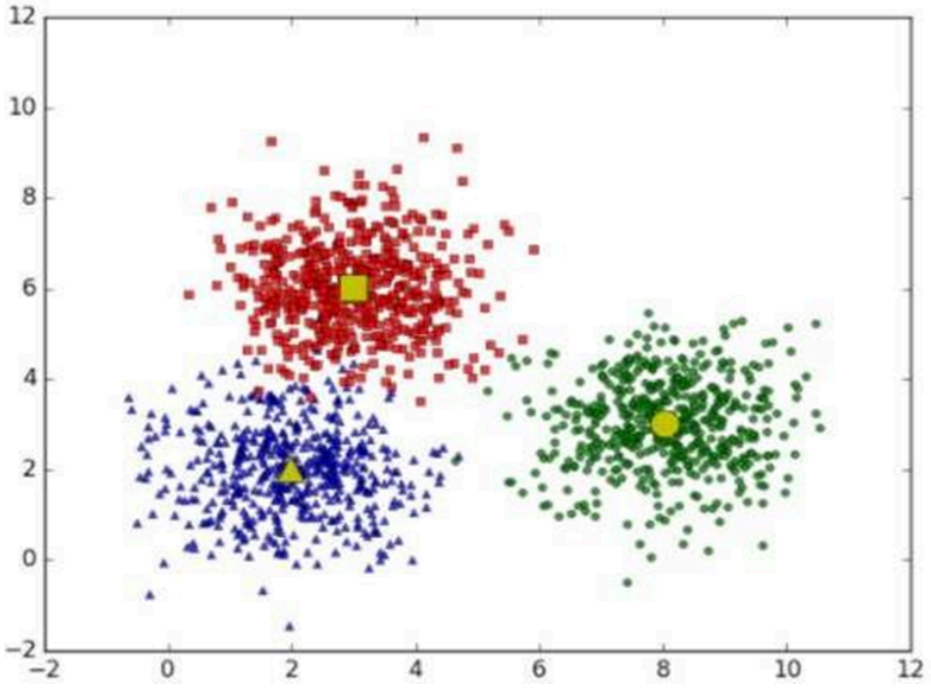
1.1.4 Lý do chọn đề tài

1. Tính thị trường: Rượu vang là một trong những sản phẩm được tiêu thụ rộng rãi trên toàn cầu, với nhiều loại và mức giá khác nhau. Do đó, việc phân tích và đánh giá chất lượng của rượu vang không chỉ mang lại lợi ích cho người tiêu dùng mà còn là một thị trường tiềm năng cho các nhà sản xuất và nhà phân phối.
2. Nhu cầu người tiêu dùng: Ngày nay, người tiêu dùng ngày càng chú trọng đến việc lựa chọn sản phẩm chất lượng và phù hợp với nhu cầu cá nhân của họ. Việc có một hệ thống phân tích chất lượng rượu vang sẽ giúp họ đưa ra quyết định mua hàng thông minh và tự tin hơn.
3. Ứng dụng của công nghệ: Việc sử dụng công nghệ thông tin và học máy để phân tích dữ liệu về rượu vang là một xu hướng phát triển mới trong lĩnh vực này. Bằng cách kết hợp sức mạnh của máy tính và dữ liệu lớn, chúng ta có thể tạo ra các công cụ thông minh giúp cải thiện trải nghiệm mua sắm của người tiêu dùng.
4. Tầm quan trọng của chất lượng: Chất lượng của rượu vang không chỉ ảnh hưởng đến trải nghiệm thưởng thức của người tiêu dùng mà còn có thể ảnh hưởng đến sức khỏe của họ. Do đó, việc có công cụ phân tích chất lượng đáng tin cậy sẽ giúp người tiêu dùng tránh được các sản phẩm kém chất lượng và nguy hại.

Chương 2. Thuật toán K-Means Clustering trong bài toán phân cụm

2.1 Tổng quan về thuật toán K-Means Clustering

Với thuật toán K-Means Clustering, chúng ta không biết nhãn (label) của từng điểm dữ liệu. Mục đích là làm thế nào để phân dữ liệu thành các cụm (cluster) khác nhau sao cho dữ liệu trong cùng một cụm có tính chất giống nhau. Ý tưởng đơn giản nhất về cluster (cụm) là tập hợp các điểm ở gần nhau trong một không gian nào đó (không gian này có thể có rất nhiều chiều trong trường hợp thông tin về một điểm dữ liệu là rất lớn). Hình bên dưới là một ví dụ về 3 cụm dữ liệu (viết gọn là cluster).



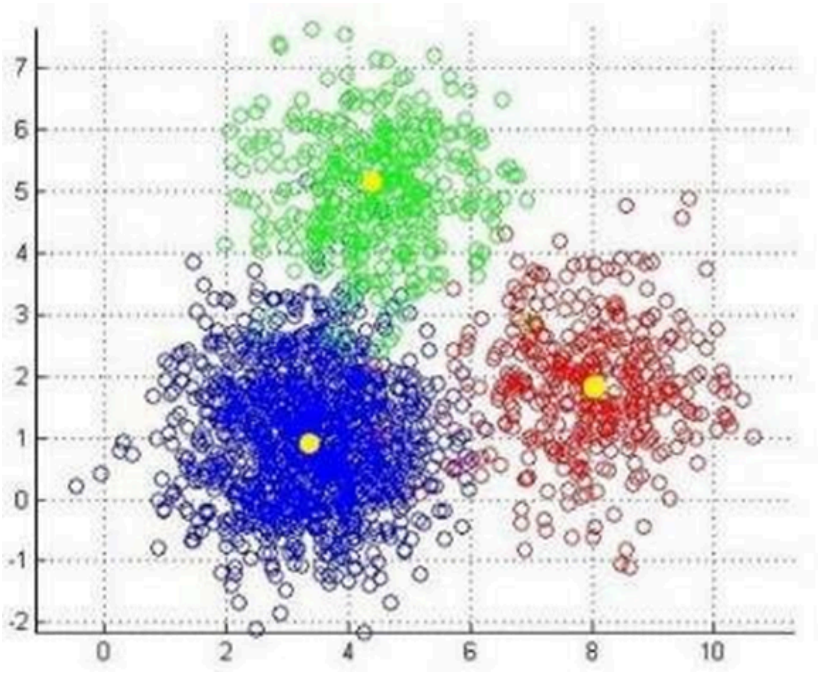
Hình 2.1. Bài toán với 3 clusters

Giả sử mỗi cluster có một điểm đại diện (center) màu vàng. Và những điểm xung quanh mỗi center thuộc vào cùng nhóm với center đó. Một cách đơn giản nhất, xét một điểm bất kỳ, ta xét xem điểm đó gần với center nào nhất thì nó thuộc về cùng nhóm với center đó.

2.2 Thuật toán K-Means Clustering

2.2.1 Mô hình toán học

Ta gọi điểm tại vị trí trung bình của tất cả các điểm dữ liệu trong một cụm là trung tâm cụm. Như vậy, nếu có K cụm thì sẽ có K trung tâm cụm và mỗi trung tâm cụm sẽ nằm gần các điểm dữ liệu trong cụm tương ứng hơn các trung tâm cụm khác. Trong hình dưới đây, K = 3 và ta có 3 trung tâm cụm là các điểm màu vàng.



Hình 2.2. Mô hình dữ liệu phân cụm

Để phân cụm dữ liệu bằng K-Means Clustering, trước hết ta chọn K là số cụm để phân chia và chọn ngẫu nhiên K trong số m dữ liệu ban đầu làm trung tâm cụm μ1, μ2, …, μK. Sau đó, với điểm dữ liệu x(i) ta sẽ gán nó cho cụm c(i) là cụm có trung tâm cụm gần nó nhất

$$c^{(i)} = argmin_k ||x^{(i)} - \mu_k||^2 \qquad (2.1)$$

Khi tất cả các điểm dữ liệu đã được gán về các cụm, bước tiếp theo là tính toán lại vị trí các trung tâm cụm bằng trung bình tọa độ các điểm dữ liệu trong cụm đó.

$$\mu_k = \frac{1}{n}(x^{(k_1)} + x^{(k_2)} + \ldots + x^{(k_n)}) \quad (2.2)$$

Với k1, k2, …, kn là chỉ số các dữ liệu thuộc cụm thứ k. Các bước trên được lặp lại cho tới khi vị trí các trung tâm cụm không đổi sau một bước lặp nào đó.

2.2.2 Độ chính xác của thuật toán

Hàm mất mát của thuật toán K-Means Clustering đặc trưng cho độ chính xác của nó sẽ càng lớn khi khoảng cách từ mỗi điểm dữ liệu tới trung tâm cụm càng lớn.

$$J(c^{(1)}, \ldots, c^{(m)}, \mu_1, \ldots, \mu_K) = \frac{1}{m}\Sigma_{i=1}^{m}||x^{(i)} - \mu_{c^{(i)}}||^2 \quad (2.3)$$

2.2.3 Nghiệm của thuật toán K-Means Clustering

Trong các bước của thuật toán, thực chất bước gán các điểm dữ liệu về trung tâm cụm gần nhất và bước thay đổi trung tâm cụm về vị trí trung bình của các điểm dữ liệu trong cụm đều nhằm mục đích giảm hàm mất mát. Thuật toán kết thúc khi vị trí các trung tâm cụm không đổi sau một bước lặp nào đó. Khi đó hàm mất mát đạt giá trị nhỏ nhất.

Khi K càng nhỏ so với m, thuật toán càng dễ đi đến kết quả chưa phải tối ưu. Điều này phụ thuộc vào cách chọn K trung tâm cụm ban đầu.

Để khắc phục điều này, ta cần lặp lại thuật toán nhiều lần và chọn phương án có giá trị hàm mất mát nhỏ nhất.

2.2.4 Tóm tắt thuật toán

Đầu vào: Dữ liệu X và số lượng cluster cần tìm K.
Đầu ra: Các center M và label vector cho từng điểm dữ liệu Y.

1. Chọn KK điểm bất kỳ làm các center ban đầu.
2. Phân mỗi điểm dữ liệu vào cluster có center gần nó nhất.
3. Nếu việc gán dữ liệu vào từng cluster ở bước 2 không thay đổi so với vòng lặp trước nó thì ta dừng thuật toán.

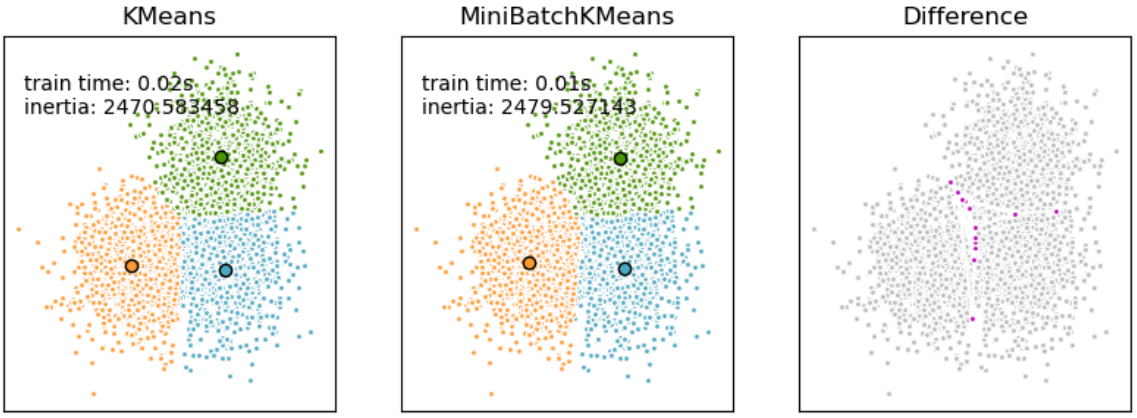
4. Cập nhật center cho từng cluster bằng cách lấy trung bình cộng của tất các các điểm dữ liệu đã được gán vào cluster đó sau bước 2.
5. Quay lại bước 2. [1] [2]

2.3 Thuật toán Mini-Batch K-Means

Mini-Batch K-Means là một biến thể của thuật toán K-Means sử dụng các mini-batch để giảm thời gian tính toán, trong khi vẫn cố gắng tối ưu hóa cùng một hàm mục tiêu. Các mini-batch là các tập con của dữ liệu đầu vào, được lấy mẫu ngẫu nhiên trong mỗi vòng lặp huấn luyện. Những mini-batch này giảm đáng kể lượng tính toán cần thiết để hội tụ đến một giải pháp cục bộ. Khác với các thuật toán khác giảm thời gian hội tụ của k-means, mini-batch k-means cho ra kết quả thường chỉ kém hơn một chút so với thuật toán tiêu chuẩn.

Thuật toán lặp lại giữa hai bước chính, tương tự như k-means thông thường. Trong bước đầu tiên, các mẫu được lấy ngẫu nhiên từ tập dữ liệu để tạo thành một mini-batch. Sau đó, chúng được gán vào trọng tâm gần nhất. Trong bước thứ hai, các trọng tâm được cập nhật. Khác với k-means, việc cập nhật này được thực hiện trên cơ sở từng mẫu. Đối với mỗi mẫu trong mini-batch, trọng tâm được gán được cập nhật bằng cách lấy trung bình trọng bình quân của mẫu và tất cả các mẫu trước đó được gán vào trọng tâm đó. Phương pháp này giúp giảm tốc độ thay đổi của trọng tâm theo thời gian. Các bước này được thực hiện cho đến khi hội tụ hoặc đạt đến số lần lặp đã xác định trước. [3]

MiniBatchKMeans hội tụ nhanh hơn so với KMeans, nhưng chất lượng kết quả bị giảm đi một chút. Trong thực tế, sự khác biệt về chất lượng này thường không đáng kể, như được thể hiện trong *Hình 2.3* bên dưới đây.



Hình 2.3. So sánh sự khác biệt giữa K-means và Mini-Batch K-means

Chương 3. Phương pháp đề xuất

3.1 Các bước tiền xử lý dữ liệu

DataFrame này có tổng cộng 12 cột dữ liệu với chi tiết như sau:

Bảng 1. Cấu trúc dataframe

Column	Non-Null Count	DType
fixed acidity	1599 non-null	float64
volatile acidity	1599 non-null	float64
citric acid	1599 non-null	float64
residual sugar	1599 non-null	float64
chlorides	1599 non-null	float64
free sulfur dioxide	1599 non-null	float64
total sulfur dioxide	1599 non-null	float64
density	1599 non-null	float64
pH	1599 non-null	float64
sulphates	1599 non-null	float64
alcohol	1599 non-null	float64
quality	1599 non-null	int64

-

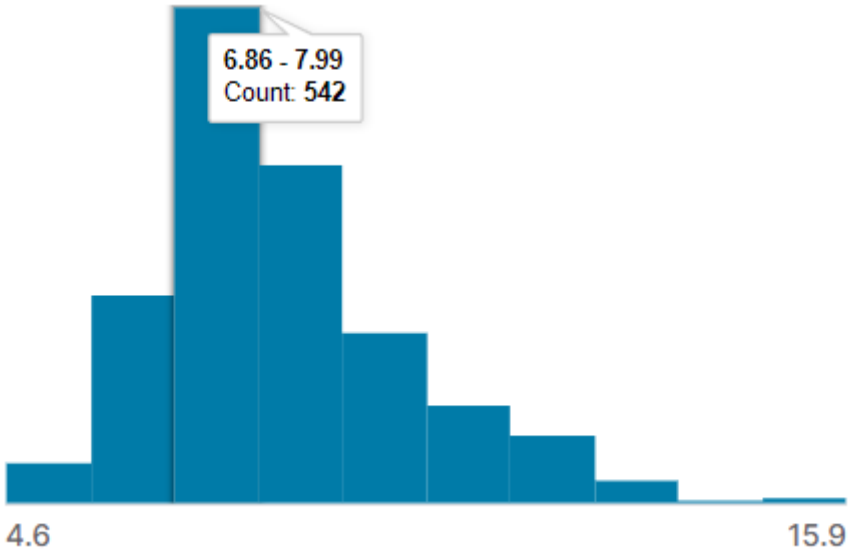
- Trong thông tin mô tả DataFrame, "Non-Null Count" là số lượng giá trị không bị thiếu trong mỗi cột. Trong trường hợp của DataFrame bạn cung cấp, số này là 1599 cho mỗi cột, điều này ngụ ý rằng không có giá trị null (hoặc thiếu) nào trong dữ liệu của các cột đó. Điều này cho thấy dữ liệu của DataFrame đã được

chuẩn bị sẵn sàng cho các phân tích hoặc xử lý tiếp theo mà không cần lo lắng về dữ liệu bị thiếu.

- Dtype là kiểu dữ liệu từ cột trong DataFrame

Fixed acidity

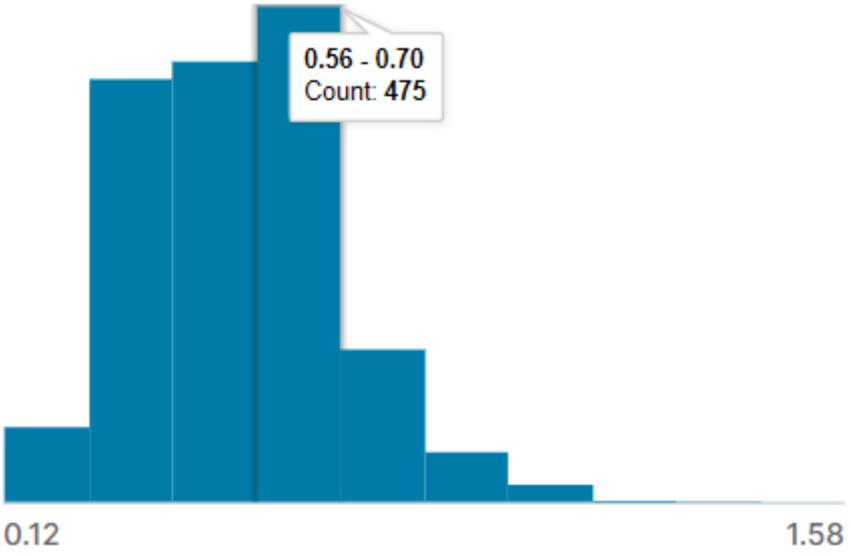
Mật độ biểu đồ Fixed acidity ở hình 3.1 phân bố nhiều ở khoảng fixed acidity 6.86 - 7.99 và ít nhất ở 14.77 - 15.9



Hình 3.1. Biểu đồ dữ liệu của fixed acidity

volatile acidity

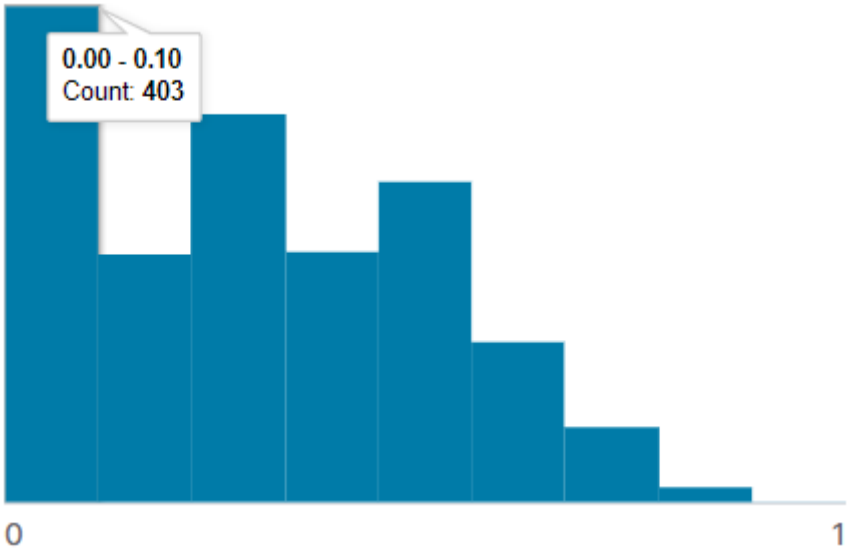
Mật độ biểu đồ volatile acidity ở hình 3.2 phân bố nhiều nhất ở khoảng 0.56 - 0.7 và ít nhất ở 1.14 - 1.29



Hình 3.2. Biểu đồ dữ liệu của volatile acidity

citric acid

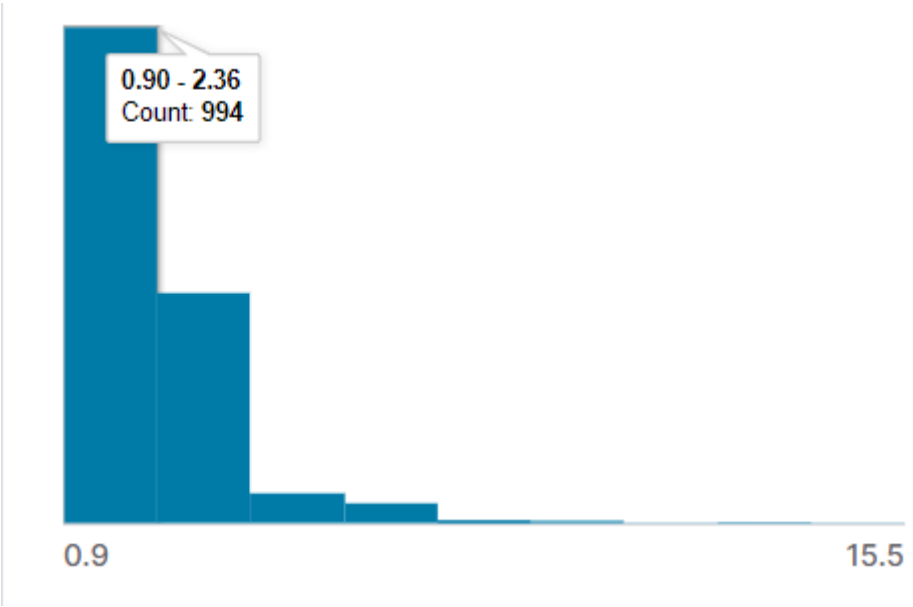
Mật độ biểu đồ citric acid hình 3.3 phân bố nhiều nhất ở 0 - 0.1 và ít nhất ở 0.7 - 0.8



Hình 3.3. Biểu đồ dữ liệu của citric acid

residual sugar

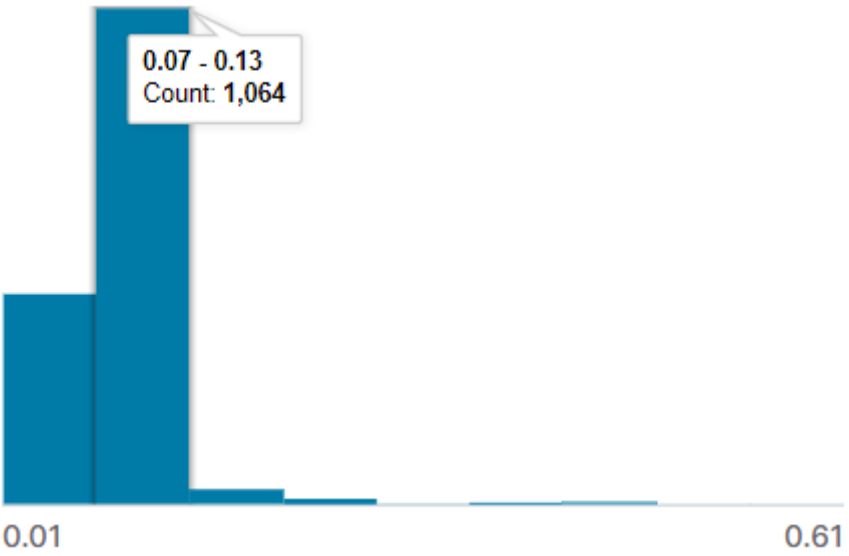
Mật độ biểu đồ residual sugar ở hình 3.4 phân bố nhiều nhất ở 0.9 - 2.36 và ít nhất ở 8.2 - 9.66



Hình 3.4. Biểu đồ dữ liệu của residual sugar

chlorides

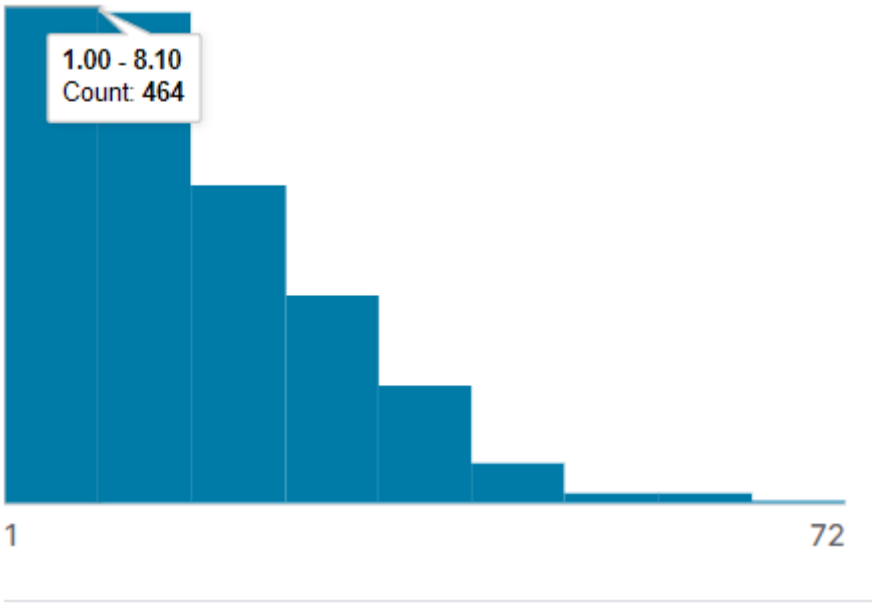
Mật độ biểu đồ chlorides ở hình 3.5 phân bố nhiều nhất ở khoảng 0.07 - 0.13 và ít nhất 0.37 - 0.43



Hình 3.5. Biểu đồ dữ liệu của chlorides

free sulfur dioxide

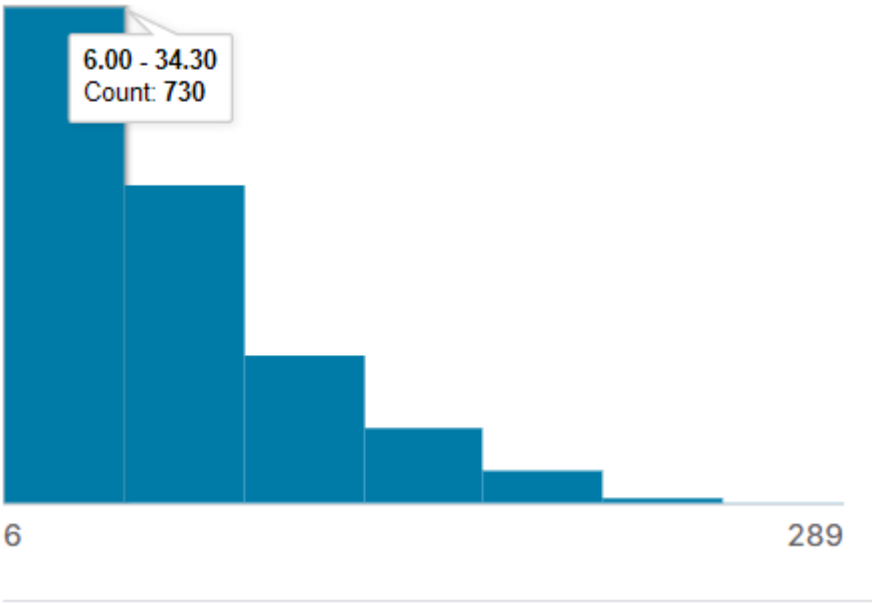
Mật độ biểu đồ free sulfur dioxide ở hình 3.6 phân bố nhiều nhất ở 1.00 - 8.1 và ít nhất ở 64.9 - 72



Hình 3.6. Biểu đồ dữ liệu của free sulfur dioxide

total sulfur dioxide

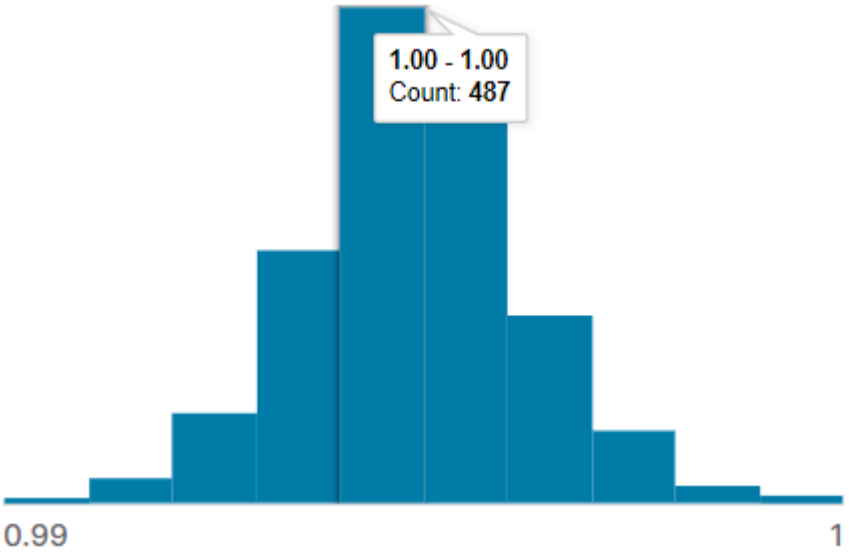
Mật độ biểu đồ total sulfur dioxide ở hình 3.7 phân bố nhiều nhất ở khoảng 6.0 - 34.3
và ít nhất ở 147.5 - 175.8



Hình 3.7. Biểu đồ dữ liệu của free total sulfur dioxide

density

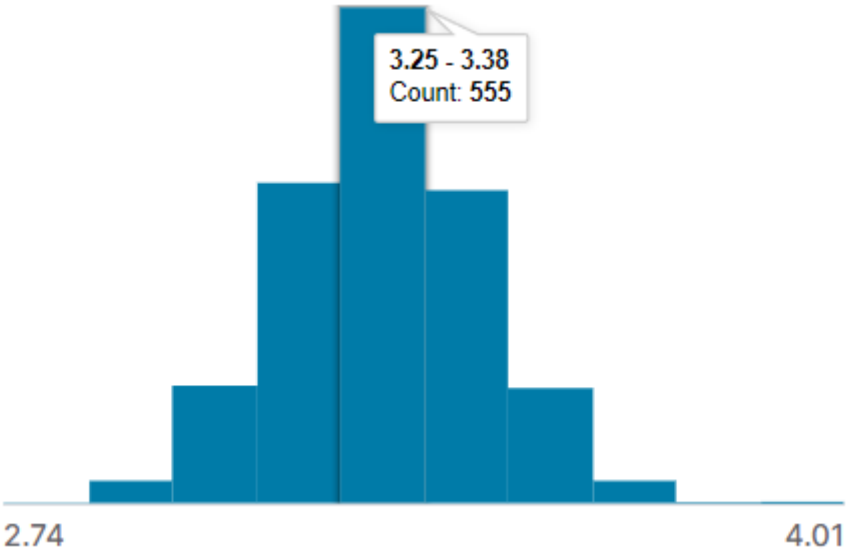
Mật độ biểu đồ density ở hình 3.8 phân bố nhiều nhất ở khoảng 1 và ít nhất ở khoảng 0.99



Hình 3.8. Biểu đồ dữ liệu của density

pH

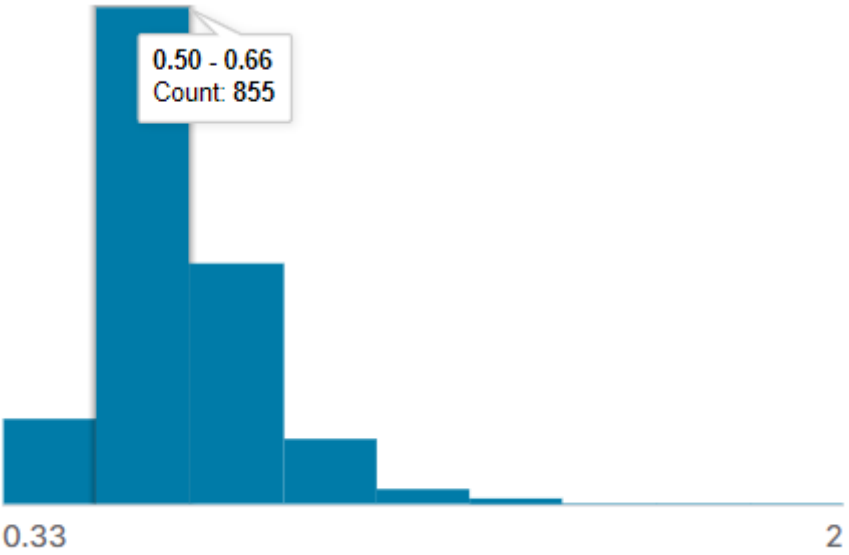
Mật độ biểu đồ pH ở hình 3.9 phân bố nhiều nhất ở khoảng 3.25 - 3.38 và ít nhất ở khoảng 3.36 - 3.76



Hình 3.9. Biểu đồ dữ liệu của pH

sulphates

Mật độ biểu đồ sulphates ở 3.10 phân bố nhiều nhất ở khoảng 0.5 - 0.66 và ít nhất ở khoảng 1.17 - 1.33



Hình 3.10. Biểu đồ dữ liệu của sulphates

```
------------------------------
       fixed acidity  volatile acidity  citric acid  residual sugar  \
count    1599.000000       1599.000000  1599.000000     1599.000000
mean        8.319637          0.527821     0.270976        2.538806
std         1.741096          0.179060     0.194801        1.409928
min         4.600000          0.120000     0.000000        0.900000
25%         7.100000          0.390000     0.090000        1.900000
50%         7.900000          0.520000     0.260000        2.200000
75%         9.200000          0.640000     0.420000        2.600000
max        15.900000          1.580000     1.000000       15.500000
```

Hình 3.11. Bản tóm tắt dữ liệu

```
         chlorides  free sulfur dioxide  total sulfur dioxide      density  \
count  1599.000000          1599.000000           1599.000000  1599.000000
mean      0.087467            15.874922             46.467792     0.996747
std       0.047065            10.460157             32.895324     0.001887
min       0.012000             1.000000              6.000000     0.990070
25%       0.070000             7.000000             22.000000     0.995600
50%       0.079000            14.000000             38.000000     0.996750
75%       0.090000            21.000000             62.000000     0.997835
max       0.611000            72.000000            289.000000     1.003690
```

Hình 3.12. Bản tóm tắt dữ liệu

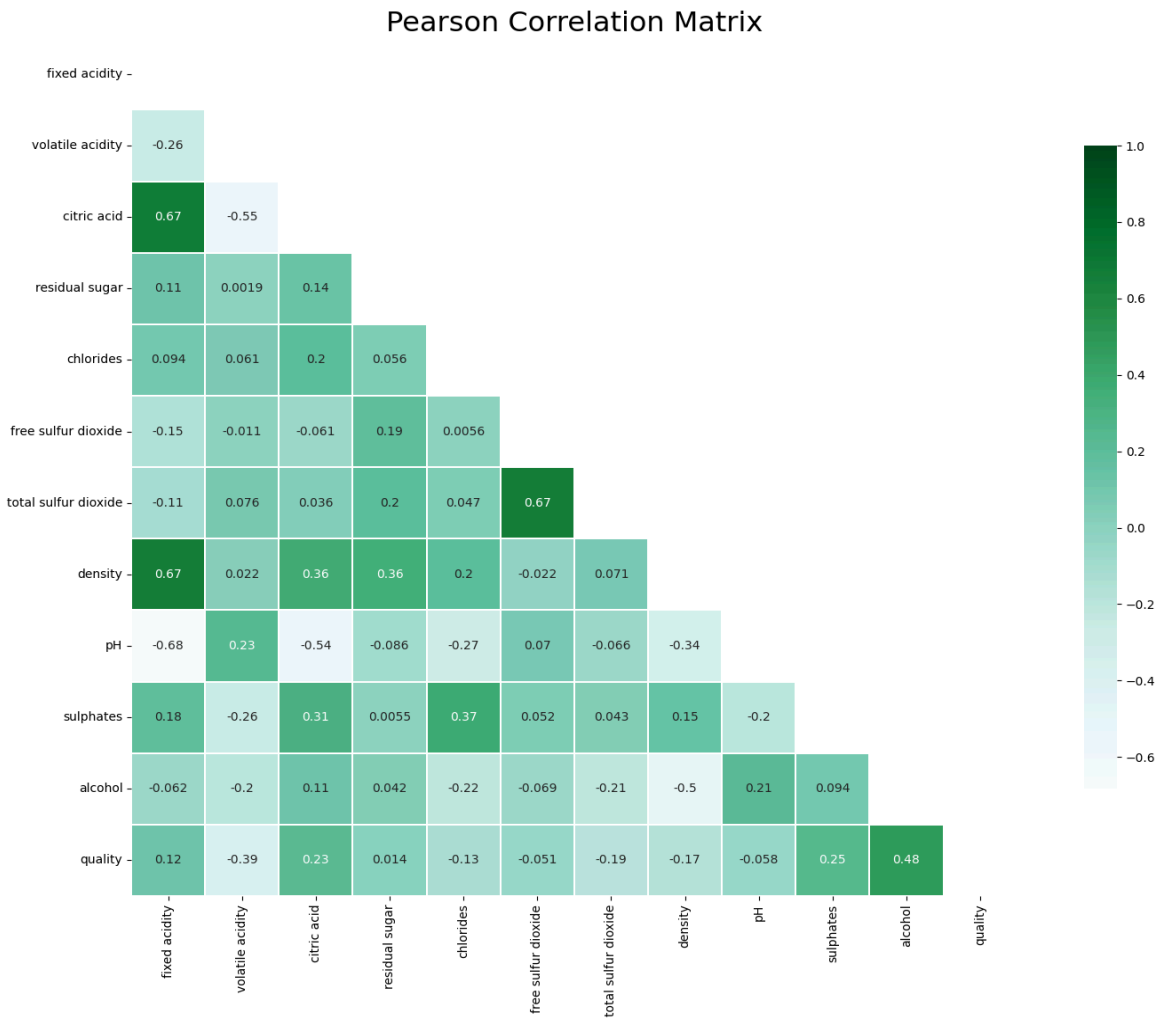
```
                pH    sulphates      alcohol      quality
count  1599.000000  1599.000000  1599.000000  1599.000000
mean      3.311113     0.658149    10.422983     5.636023
std       0.154386     0.169507     1.065668     0.807569
min       2.740000     0.330000     8.400000     3.000000
25%       3.210000     0.550000     9.500000     5.000000
50%       3.310000     0.620000    10.200000     6.000000
75%       3.400000     0.730000    11.100000     6.000000
max       4.010000     2.000000    14.900000     8.000000
```

Hình 3.13. Bản tóm tắt dữ liệu

Các bản tóm tắt dữ liệu trong *Hình 3.1, 3.12, 3.13* cho ta biết cái nhìn tổng quan về dữ liệu

Cụ thể:

1. Count: Số lượng mẫu không bị thiếu dữ liệu (không bị giá trị null) trong cột.
2. Mean: Giá trị trung bình của các mẫu trong cột.
3. std: Độ lệch chuẩn, chỉ sự biến động của dữ liệu xung quanh giá trị trung bình. Giá trị này càng lớn cho thấy dữ liệu càng phân tán.
4. Min: Giá trị nhỏ nhất trong cột
5. 25% (Q1): Phần vị 25%, là giá trị mà 25% các mẫu nhỏ hơn và 75% các mẫu lớn hơn.
6. 50% (Q2): Phần vị 50%, tương đương với giá trị trung vị (median), là giá trị giữa của tập dữ liệu khi được sắp xếp theo thứ tự tăng dần.
7. 75% (Q3): Phần vị 75%, là giá trị mà 75% các mẫu nhỏ hơn và 25% các mẫu lớn hơn.
8. max: Giá trị lớn nhất trong cột.



Hình 3.14. Ma trận tương quan các dữ liệu

Trong *Hình 3.14.* mỗi ô trong ma trận hiển thị giá trị tương quan giữa hai biến tương ứng. Giá trị này dao động từ -1 đến 1, với ý nghĩa:

- -1: Mối tương quan nghịch hoàn toàn: Khi một biến tăng, biến kia giảm và ngược lại.
- 0: Không có mối tương quan.
- 1: Mối tương quan thuận hoàn toàn: Khi một biến tăng, biến kia cũng tăng và ngược lại.

Màu sắc trong ô thể hiện mức độ tương quan:

- Xanh lá cây: Mối tương quan thuận.
- Đỏ: Mối tương quan nghịch.
- Trắng: Không có mối tương quan.

Phân tích biểu đồ:

Fixed acidity:

- Có mối tương quan thuận mạnh với citric acid (0.67): Rượu vang có độ chua cố định cao cũng có xu hướng có hàm lượng citric acid cao.
- Có mối tương quan thuận mạnh với mật độ (0.67): Rượu vang có độ chua cố định cao cũng có xu hướng có mật độ cao hơn.
- Có mối tương quan nghịch mạnh với pH (-0.68): Rượu vang có Fixed acidity cao có xu hướng có độ pH thấp hơn (tính acid cao hơn).

Volatile acidity:

- Có mối tương quan nghịch mạnh với citric acid (-0.55): Rượu vang có volatile acidity cao có xu hướng có hàm lượng citric acid cao.
- Có mối tương quan nghịch vừa phải với sulphates (-0.26): Rượu vang có volatile acidity cao có xu hướng có hàm lượng sulphates thấp.

Citric acid:

- Có mối tương quan thuận mạnh với mật độ (0.36): Rượu vang có hàm lượng Citric acid cao có xu hướng có mật độ cao hơn.
- Có mối tương quan thuận vừa phải với sulphates (0.31): Rượu vang có hàm lượng Citric acid cao có xu hướng có hàm lượng sulphates cao hơn.
- Có mối tương quan nghịch mạnh với pH (-0.54): Rượu vang có hàm lượng Citric acid cao có xu hướng có độ pH thấp hơn (tính acid cao hơn).

Residual sugar:

- Có mối tương quan nghịch mạnh với alcohol (-0.22): Rượu vang có lượng đường dư cao có xu hướng có độ cồn thấp.
- Có mối tương quan nghịch vừa phải với chất lượng (-0.13): Rượu vang có lượng residual sugar cao có xu hướng có chất lượng thấp hơn.

Chlorides:

- Có mối tương quan thuận vừa phải với mật độ (0.36): Rượu vang có hàm lượng chlorides cao có xu hướng có mật độ cao hơn.
- Có mối tương quan thuận vừa phải với sulphates (0.37): Rượu vang có hàm lượng Chlorides cao có xu hướng có hàm lượng sulphates cao hơn.

3.2 Input, output của mô hình

Input:

- **Dữ liệu:** Dữ liệu chứa thông tin về 1599 mẫu rượu vang đỏ, bao gồm 11 thuộc tính như chất lượng rượu vang (quality), độ axit (fixed acidity), độ pH, lượng đường dư (residual sugar), lượng cồn (alcohol), v.v.
- **Số lượng cụm K:** Số lượng cụm K thích hợp để mô hình K-means Clustering phân chia dữ liệu.

Output:

- **Nhãn cụm (label):** Cho mỗi điểm dữ liệu trong dataset, mô hình K-means Clustering sẽ gán một nhãn cụm cho điểm dữ liệu đó. Nhãn cụm này biểu thị rằng điểm dữ liệu đó thuộc về cụm nào.
- **Vị trí trung tâm của các cụm:** Mô hình K-means Clustering sẽ xác định vị trí trung tâm của mỗi cụm. Vị trí trung tâm này là điểm dữ liệu trung bình của tất cả các điểm dữ liệu đã được gán cho cụm đó.
- **Biểu đồ phân cụm:** Các biểu đồ phân cụm để trực quan hóa cách các điểm dữ liệu được phân chia thành các cụm.

3.3 Đề xuất mô hình sử dụng

K-means Clustering là một thuật toán phân cụm không giám sát phổ biến được sử dụng để nhóm các điểm dữ liệu có cùng đặc điểm.

K-means Clustering được lựa chọn cho bài toán phân cụm rượu vang vì những lý do sau:

- **Đơn giản:** Dễ hiểu và dễ triển khai.
- **Hiệu quả:** Có thể xử lý được lượng dữ liệu lớn một cách hiệu quả.
- **Linh hoạt:** Được áp dụng cho nhiều loại dữ liệu khác nhau, bao gồm cả dữ liệu rượu vang.
- **Hiệu quả:** Giúp phát hiện các mẫu và nhóm ẩn trong dữ liệu rượu vang, có thể được sử dụng để phân loại rượu vang hoặc để hiểu rõ hơn về đặc điểm của rượu vang.

Chương 4. Thực nghiệm và đánh giá kết quả

4.1 Mô tả data, thống kê mô tả cho data, kết quả các bước tiền xử lý data

4.1.1 Mô tả data

Bộ dữ liệu này có thể dùng trong các nhiệm vụ phân loại hoặc dự đoán. Các loại rượu được sắp xếp theo thứ tự và không cân bằng (ví dụ: Có nhiều loại rượu bình thường hơn nhiều so với loại xuất sắc hoặc kém).
Input (dựa trên thử nghiệm hóa lý)
1 – Fixed acidity: Hầu hết các axit có liên quan đến rượu vang đều là axit cố định hoặc không bay hơi (không bay hơi nhanh chóng).

2 – Volatile acidity: Lượng axit axetic trong rượu vang, ở mức cao quá có thể dẫn đến hương vị gắt, giống như giấm, không dễ chịu.

3 – Axit citric: Trong một lượng nhỏ, axit citric có thể thêm vào 'sự tươi mới' và hương vị cho rượu vang.

4 – Citric acid: Lượng đường còn lại sau khi quá trình lên men kết thúc, thường thấy rất hiếm rượu có ít hơn 1 gram/lít

5 – Chlorides: Lượng muối có trong rượu

6 – Free sulfur dioxide tự do: Dạng tự do của SO2 tồn tại trong cân bằng giữa SO2 phân tử (dưới dạng khí tan) và ion bisulfit; ngăn chặn sự oxy hóa và phát triển của vi khuẩn

7 – Total sulfur dioxide: Số lượng dạng tự do và kết hợp của SO2; ở nồng độ thấp, SO2 chủ yếu không thể được phát hiện trong rượu

8 – Density: Tỉ trọng của nước gần bằng với nước tùy thuộc vào tỷ lệ phần trăm cồn và đường trong rượu.

9 – pH: Mô tả mức độ axit hoặc cơ bản của một loại rượu trên thang đo từ 0 (rất axit) đến 14 (rất cơ bản); hầu hết các loại rượu có giá trị từ 3-4.

10 – Sulphates: Một phụ gia cho rượu vang có thể góp phần vào việc tạo ra mức độ khí lưu huỳnh dioxit (SO2), có tác dụng như một chất kháng vi khuẩn và chống oxy hóa.

11 – Alcohol: Tỉ lệ cồn trong rượu

12 – Quality : Điểm từ 1 đến 10

4.1.2 Thống kê mô tả cho data

Từ mục 3.1 tiền xử lý dữ liệu trên có kết quả thống kê mô tả cho thấy:

- Đa số các thuộc tính có giá trị trung bình nằm trong khoảng hợp lý.
- Một số thuộc tính có độ lệch chuẩn cao, cho thấy sự phân bố dữ liệu không đồng đều.
- Không có giá trị thiếu cho bất kỳ thuộc tính nào.

4.1.3 Kết quả các bước tiền xử lý dữ liệu

Dữ liệu đã được chuẩn hóa bằng StandardScaler. Việc này giúp đảm bảo rằng tất cả các thuộc tính có cùng đơn vị đo lường và thang giá trị tương đương. Điều này cải thiện hiệu quả của thuật toán K-means Clustering, vốn nhạy cảm với sự khác biệt về đơn vị đo lường.

Ví dụ: trước khi chuẩn hóa, giá trị của thuộc tính "alcohol" có thể dao động từ 8.4 đến 14.9, trong khi giá trị của thuộc tính "volatile acidity" có thể dao động từ 0.12 đến 1.58. Do đó, khoảng cách giữa các điểm dữ liệu dựa trên "alcohol" có thể lớn hơn nhiều so với khoảng cách giữa các điểm dữ liệu dựa trên "volatile acidity", dẫn đến việc K-means Clustering tập trung vào "alcohol" nhiều hơn các thuộc tính khác.

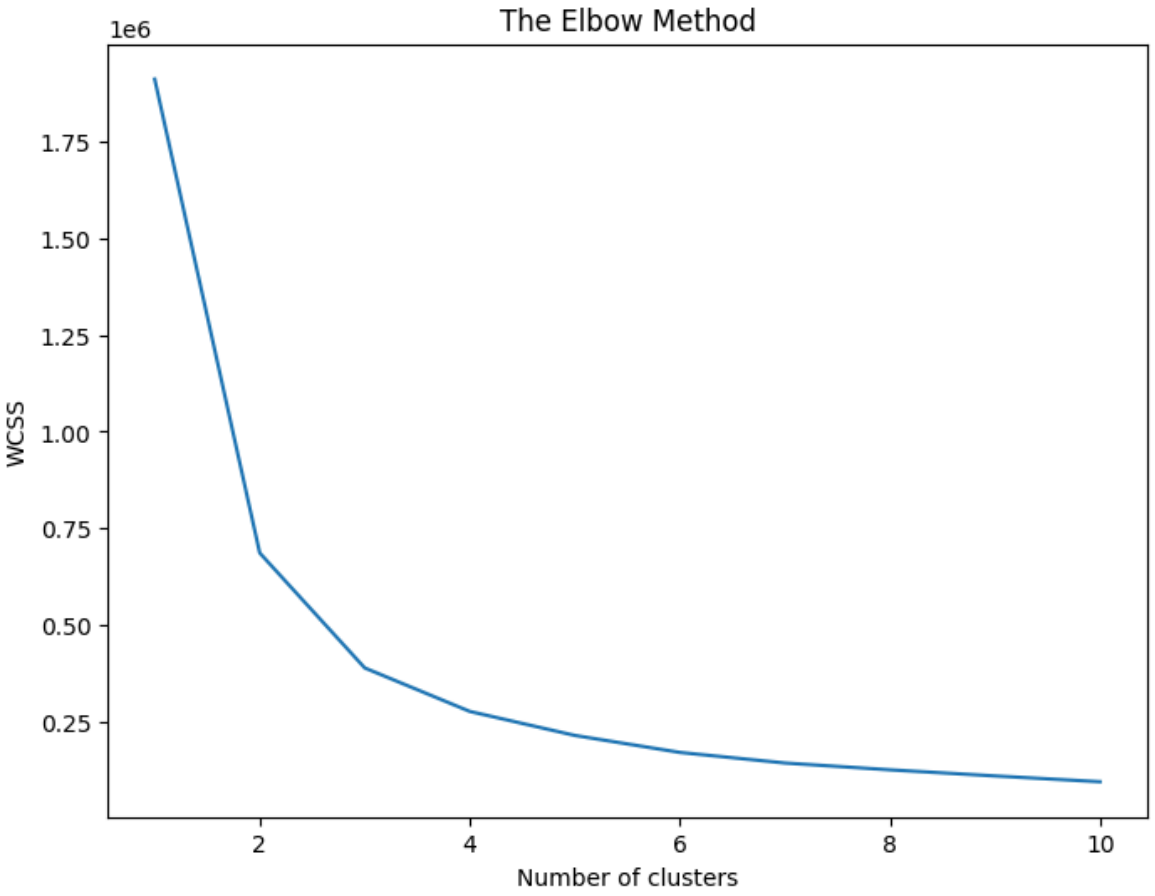
Sau khi chuẩn hóa, tất cả các thuộc tính đều có giá trị trung bình bằng 0 và độ lệch chuẩn bằng 1. Điều này đảm bảo rằng tất cả các thuộc tính có cùng tầm quan trọng trong quá trình phân cụm.

Kết quả:

Việc tiền xử lý dữ liệu đã giúp cải thiện hiệu quả của thuật toán K-means Clustering. Dữ liệu sau khi được tiền xử lý có tính phân bố đồng đều hơn và ít nhiễu hơn, dẫn đến kết quả phân cụm chính xác và đáng tin cậy hơn.

4.2 Thông số cho mô hình, độ đo đánh giá

4.2.1 Thông số cho mô hình



Hình 4.1. Elbow Method

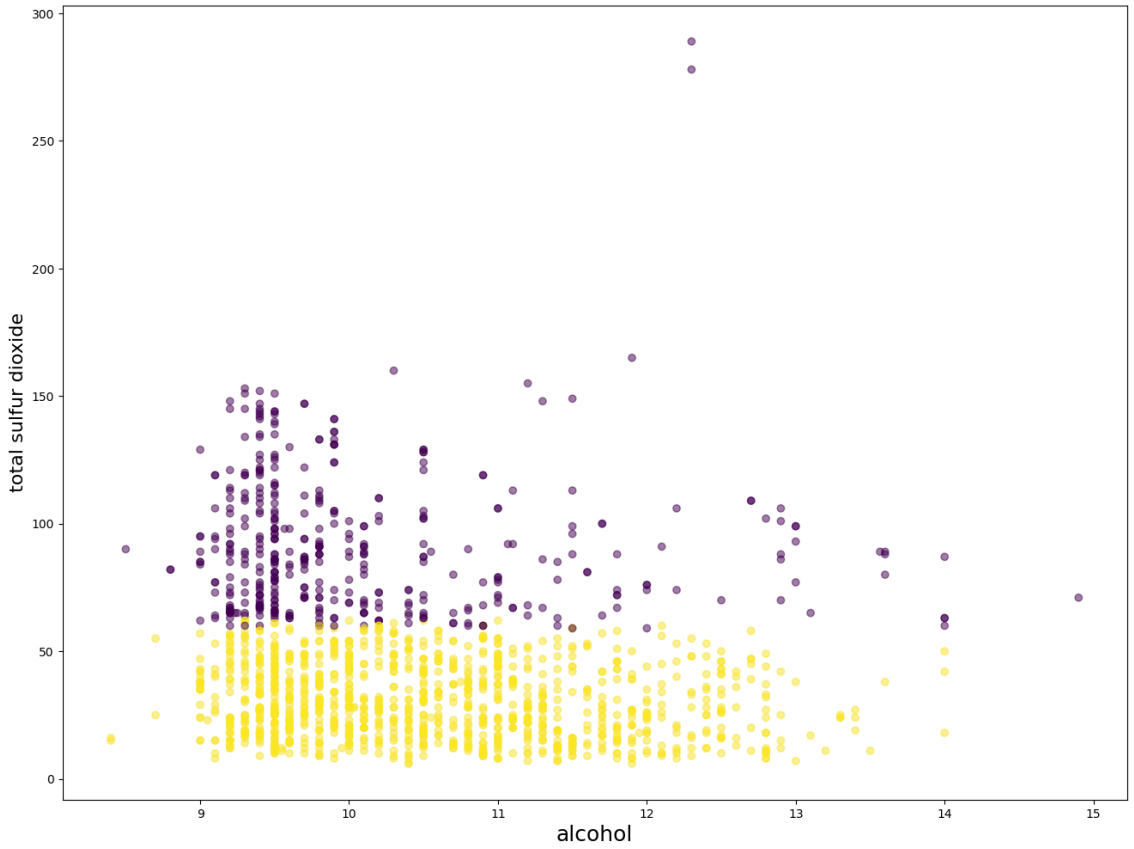
- Số lượng cụm (K): 2 (được xác định bằng *Hình 4.1*)
- Thuật toán khởi tạo: k-means++
- Số lần khởi tạo: 12
- Số lần lặp tối đa: 300

4.2.2 Độ đo đánh giá

- WCSS (Within-Cluster Sum of Squared Errors)**:** Đo lường tổng phương sai bình phương của các điểm dữ liệu trong mỗi cụm. Giá trị WCSS thấp hơn cho thấy sự phân cụm chặt chẽ hơn.

4.3 Đánh giá kết quả: biểu diễn dạng đồ thị, các thông số đánh giá mô hình

- Biểu đồ 2 chiều với số cụm (K) = 2:

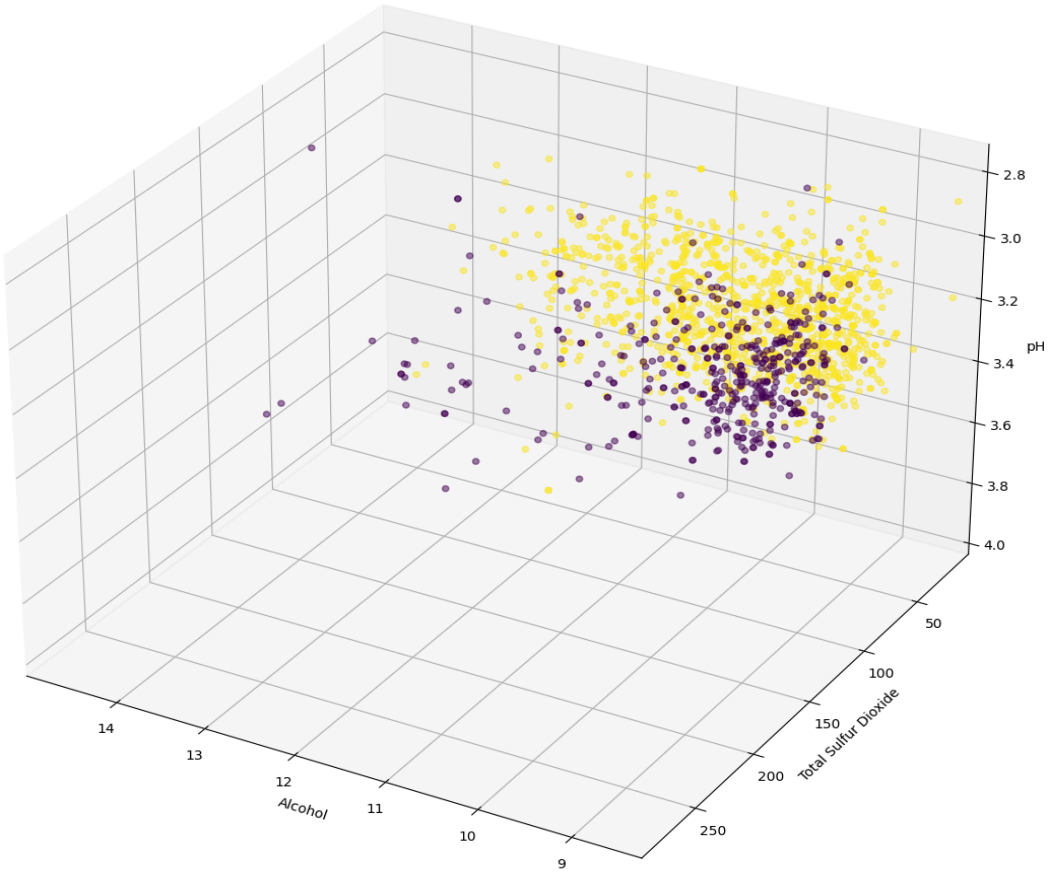


Hình 4.2. Biểu đồ phân cụm 2 chiều với k = 2

Dựa vào *Hình 4.2.* bên trên, kết quả K-means Clustering được biểu diễn bằng biểu đồ phân cụm 2D, với các điểm dữ liệu được tô màu theo nhãn cụm (0 hoặc 1). Biểu đồ cho thấy sự phân bố rõ ràng của các điểm dữ liệu trong hai cụm:

- **Cụm 0 (màu vàng):** Có mật độ cao hơn và tập trung ở khu vực phía dưới bên trái của biểu đồ. Cụm này có thể bao gồm các loại rượu vang có độ cồn cao, độ axit thấp.
- **Cụm 1 (màu tím):** Có mật độ thưa hơn và phân bố rải rác hơn trên biểu đồ. Cụm này có thể bao gồm các loại rượu vang có độ cồn thấp, độ axit cao.

- Biểu đồ 3 chiều với số cụm (K) = 2:

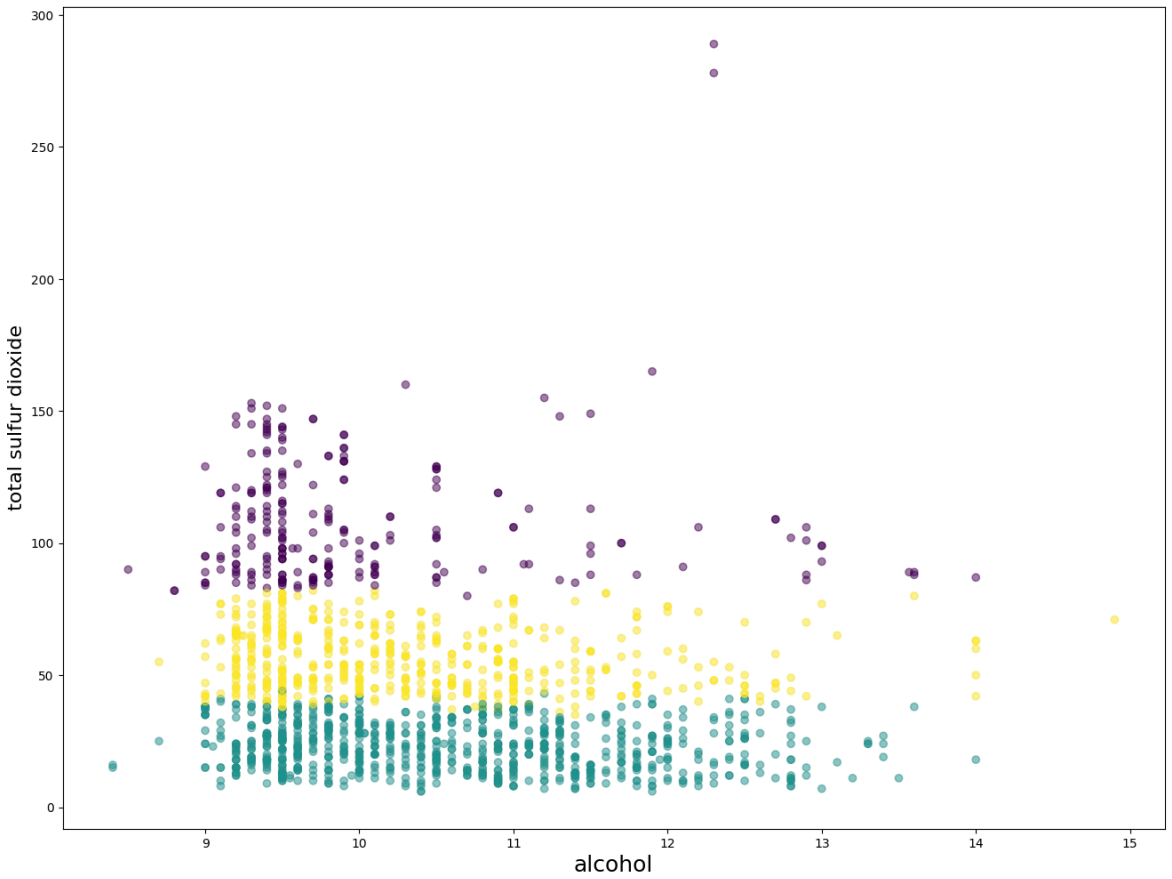


Hình 4.3. Biểu đồ phân cụm 3 chiều với k = 2

Biểu đồ phân cụm 3D ở *Hình 4.3.* thể hiện kết quả phân chia dữ liệu rượu vang thành hai cụm (màu vàng và màu tím) dựa trên ba thuộc tính: alcohol, total sulfur dioxide và pH.

- **Cụm 0 (màu vàng):** Tập trung ở khu vực phía dưới bên trái biểu đồ, bao gồm các điểm dữ liệu có hàm lượng cồn cao, lượng SO2 tổng thể thấp và độ pH cao.
- **Cụm 1 (màu tím):** Phân bố rải rác hơn ở khu vực trung tâm và phía trên biểu đồ, bao gồm các điểm dữ liệu có hàm lượng cồn thấp, lượng SO2 tổng thể cao và độ pH thấp.

- Biểu đồ 2 chiều với số cụm (K) = 3:

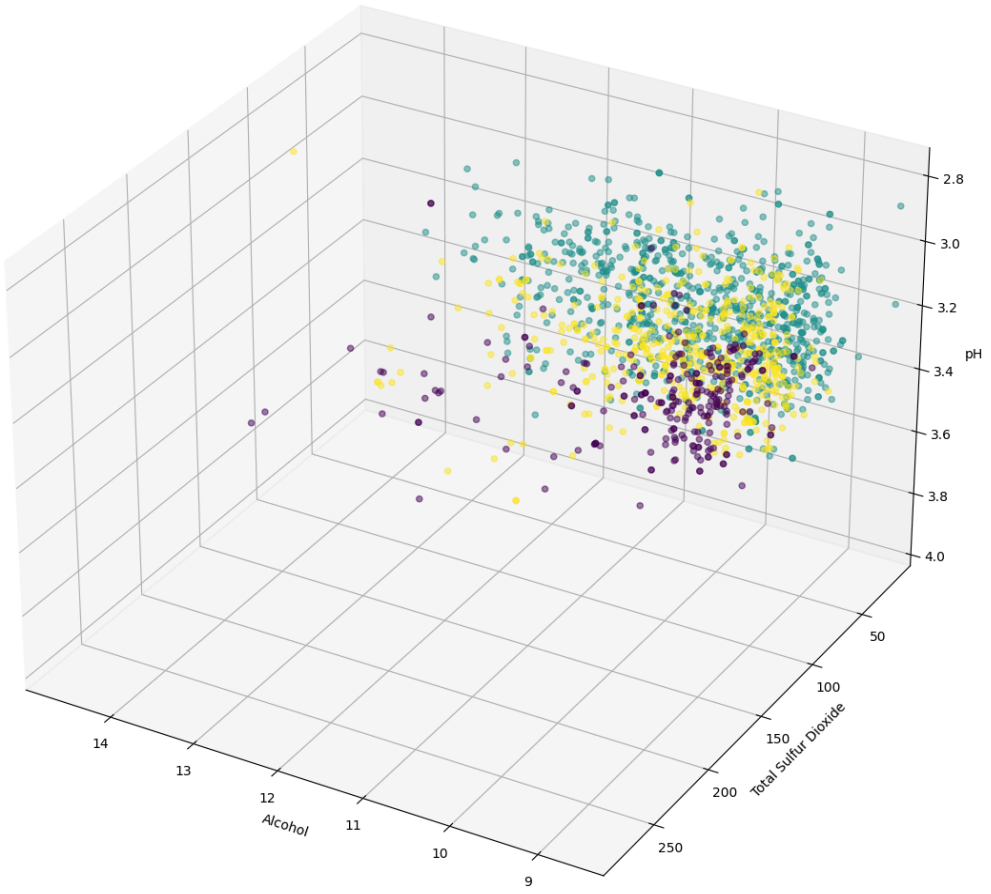


Hình 4.4. Biểu đồ phân cụm 2 chiều với k = 3

Dựa vào *Hình 4.4.* trên, kết quả K-means Clustering được biểu diễn bằng biểu đồ phân cụm 2D, với các điểm dữ liệu được tô màu theo nhãn cụm (0,1 hoặc 2). Biểu đồ cho thấy sự phân bố rõ ràng của các điểm dữ liệu trong ba cụm:

- **Cụm 0 (màu xanh lá):** có xu hướng nồng độ cồn cao và tổng lượng SO2 thấp.
- **Cụm 1 (màu tím):** có xu hướng nồng độ cồn thấp và tổng lượng SO2 cao.
- **Cụm 2 (màu vàng):** nằm ở vị trí trung gian giữa hai cụm còn lại, với nồng độ cồn và tổng lượng SO2 dao động trong khoảng trung bình.

- Biểu đồ 3 chiều với số cụm (K) = 3:



Hình 4.5. Biểu đồ phân cụm 3 chiều với k = 3

Biểu đồ phân thành ba cụm (màu xanh lá, màu vàng và màu tím) dựa trên ba thuộc tính: alcohol, total sulfur dioxide và pH ở *Hình 4.5*.

- **Cụm 0 (màu xanh lá):** có xu hướng nồng độ cồn thấp, tổng lượng SO2 thấp và độ pH cao.
- **Cụm 1 (màu tím):** có xu hướng nồng độ cồn cao, tổng lượng SO2 cao và độ pH thấp.
- **Cụm 2 (màu vàng):** nằm ở vị trí trung gian giữa hai cụm còn lại, với nồng độ cồn, tổng lượng SO2 và độ pH dao động trong khoảng trung bình.

Chương 5. Kết luận và đề xuất

5.1 Kết luận về kết quả

Kết quả phân tích:

- Việc tạo ra nhóm thứ ba với đặc điểm trung bình không mang lại lợi ích thực tế rõ ràng trong việc phân loại và lựa chọn rượu vang. Nên số cụm (K) = 2 vẫn là tối ưu nhất theo phương pháp Elbow

- **Phân cụm:** Mô hình K-means đã phân chia hiệu quả dữ liệu thành hai cụm riêng biệt.

- **Số lượng cụm:** Dựa trên phương pháp Elbow, k = 2 được xác định là số cụm tối ưu để phân loại dữ liệu.

- **Đặc điểm cụm:** Phân tích thống kê cho thấy sự khác biệt rõ ràng giữa hai cụm về nồng độ cồn, tổng lượng SO2 và độ pH. Cụm 0 có xu hướng nồng độ cồn cao, tổng lượng SO2 cao và độ pH thấp, trong khi Cụm 1 có xu hướng nồng độ cồn thấp, tổng lượng SO2 thấp và độ pH cao.

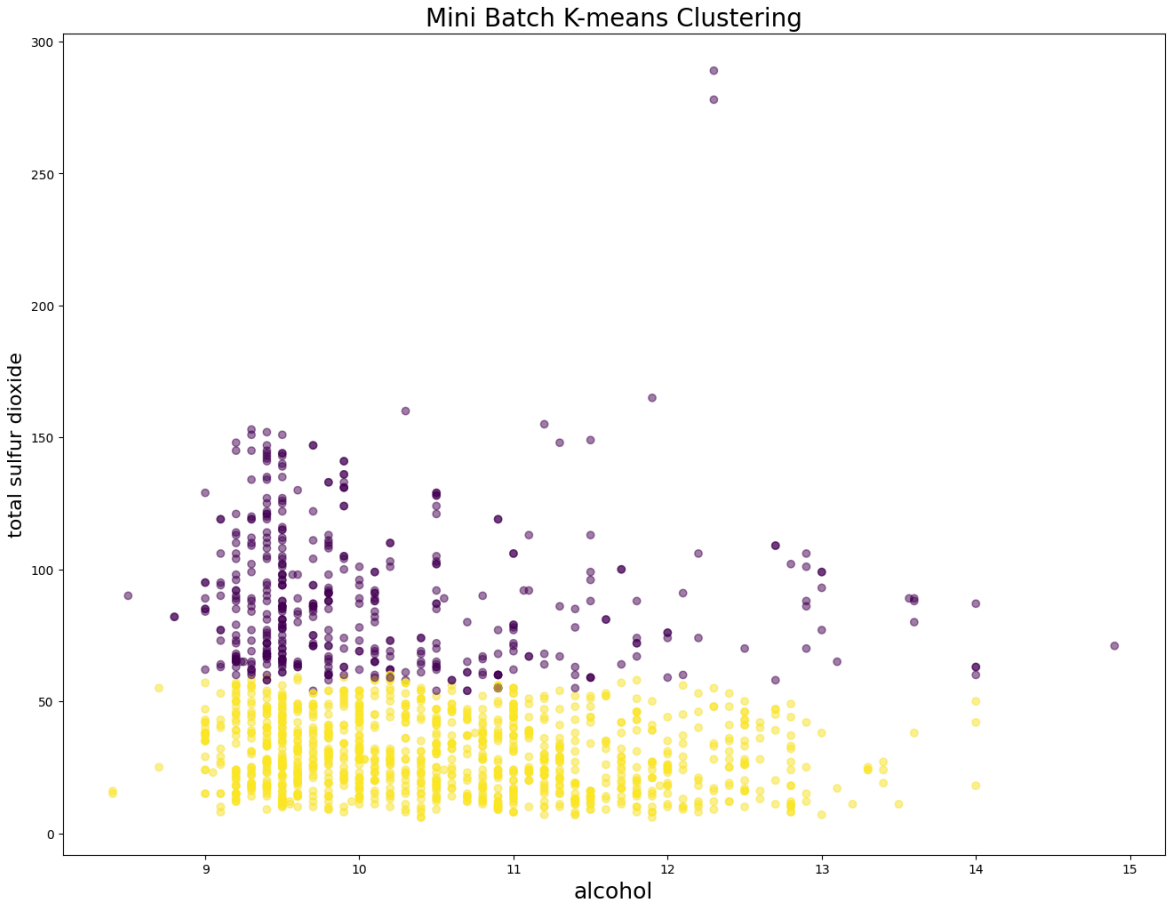
5.2 Các đề xuất cải tiến mô hình

Đề xuất: Sử dụng Mini-batch K-means.

Mini-batch K-means là một biến thể của K-means tiêu chuẩn, hoạt động bằng cách xử lý dữ liệu theo từng lô nhỏ (mini batch) thay vì xử lý toàn bộ dữ liệu cùng một lúc. Có ưu và khuyết điểm như sau: [4]

1. Ưu điểm của Mini-batch K-means:
    - Tốc độ tính toán nhanh hơn: Do xử lý theo lô nhỏ, thuật toán Mini-batch K-means thường nhanh hơn so với K-means truyền thống.
    - Khả năng xử lý dữ liệu lớn: Mini-batch K-means phù hợp với các tập dữ liệu lớn mà K-means truyền thống có thể gặp khó khăn.
2. Khuyết điểm của Mini-batch K-means:
    - Kết quả có thể khác nhỏ so với K-means truyền thống: Do xử lý theo lô nhỏ, Mini-batch K-means có thể cho kết quả gần giống, nhưng không hoàn toàn giống với K-means truyền thống.

- Biểu đồ phân cụm 2 chiều của Mini-batch K-means với k = 2:



Hình 5.1. Biểu đồ phân cụm 2 chiều của Mini-batch K-means với k = 2

Chúng ta gần như khó mà thấy được sự khác nhau giữa Hình 5.1. trên và Hình 4.2. nếu không quan sát kỹ càng. Vậy chúng ta có thể thấy được về mặt kết quả thì Mini-batch K-means không khác với K-means tiêu chuẩn là bao, tiếp theo chúng ta sẽ so sánh về tốc độ thực hiện:

```
[1 0 1 ... 1 1 1]
K-means elapsed time: 0.571988582611084
Mini Batch K-means elapsed time: 0.12208342552185059
```

Hình 5.2. Thời gian chạy giữa K-means và Mini-batch K-means

Như *Hình 5.2.* bên trên, kết quả in ra cho thấy thời gian chạy của Mini-batch K-means nhanh hơn đáng kể so với K-means tiêu chuẩn. Điều này minh họa cho ưu điểm về tốc độ tính toán của Mini-batch K-means.

TÀI LIỆU THAM KHẢO

[1] DAT HOANG'S BLOG. (n.d.). *Thuật Toán K-Means Clustering. k-means-clustering*. Retrieved May 10, 2024, from https://dathoangblog.com/2019/01/k-means-clustering.html

[2] Machinelearningcoban. (2017, 1 1). *K-means Clustering. Machine Learning cơ bản*. Retrieved May 10, 2024, from https://machinelearningcoban.com/2017/01/01/kmeans/

[3] Scikit-learn. (n.d.). *Comparison of the K-Means and MiniBatchKMeans clustering algorithms. Scikit-learn*. Retrieved May 10, 2024, from https://scikit-learn.org/stable/auto_examples/cluster/plot_mini_batch_kmeans.html

[4] Scikit-learn. (n.d.). *Mini Batch K-Means. Clustering*. Retrieved May 10, 2024, from https://scikit-learn.org/stable/modules/clustering.html#mini-batch-kmeans