PRINCIPAL COMPONENT ANALYSIS AND AN APPLICATION IN WINE DATA ANALYSIS

Nong Quynh Van
TNU - University of Education

ARTICLE INFO

Received: **08/5/2024**
Revised: **17/6/2024**
Published: **18/6/2024**

KEYWORDS

Principal component analysis
Dimensionality reduction
Multivariate data analysis
Correlation
Wine data

ABSTRACT

Analyzing the relationship between chemical properties and wine quality will contribute to better wine's price valuation. In this study, we use an availabe online dataset on white wines that produced from the "Vinho Verde" region in Portugal, which contains 11 variables measuring the chemical properties of white wines and 1 variable represents the quality of wine (ranked from 0 to 10 according to the sensory assessment of the wine tasters). The goal is to find chemical property variables that influence the wine quality classification. To achieve the goal, we propose to use principal component analysis in correlation and classification assessment. The results showed that PCA performs relatively well in classifying white wine based on its chemical properties. Chemical compositions such as residual sugar, total sulfur dioxide, density, alcohol, pH, and fixed acidity play an important role in the flavor of white wine. On the other hand, it was found that Alcohol and pH contributed highly to quality of the wine.

PHÂN TÍCH THÀNH PHẦN CHÍNH VÀ ỨNG DỤNG TRONG PHÂN TÍCH DỮ LIỆU RƯỢU VANG

Nông Quỳnh Vân
Trường Đại học Sư phạm - ĐH Thái Nguyên

THÔNG TIN BÀI BÁO

Ngày nhận bài: **08/5/2024**
Ngày hoàn thiện: **17/6/2024**
Ngày đăng: **18/6/2024**

TỪ KHÓA

Phân tích thành phần chính
Giảm số chiều
Phân tích đa biến
Tương quan
Dữ liệu rượu vang

TÓM TẮT

Phân tích mối quan hệ giữa các đặc tính hóa học và chất lượng rượu vang sẽ giúp chuyên gia định giá rượu vang được tốt hơn. Trong nghiên cứu này, chúng tôi sử dụng bộ dữ liệu về rượu vang trắng được sản xuất tại khu vực "Vinho Verde" ở Bồ Đào Nha, trong đó có 11 biến đo lường các đặc tính hóa học của rượu vang trắng và 1 biến thể hiện chất lượng của rượu vang (được xếp hạng từ 0 đến 10 theo đánh giá cảm quan của người nếm rượu). Mục tiêu là tìm các biến đặc tính hóa học có ảnh hưởng đến phân loại chất lượng rượu vang. Để đạt được mục tiêu, chúng tôi đề xuất sử dụng phương pháp phân tích thành phần chính trong đánh giá mối tương quan và phân loại. Kết quả cho thấy PCA hoạt động tương đối tốt trong phân loại rượu vang trắng dựa trên các đặc tính hóa học của nó. Các đặc tính hóa học như lượng đường dư, tổng lưu huỳnh dioxit, tỷ trọng, nồng độ cồn, pH, độ axit cố định thể hiện cho mùi vị đặc trưng của rượu vang trắng. Đồng thời kết quả phân tích cũng chỉ ra rượu vang trắng chất lượng cao sẽ có nồng độ cồn và độ pH lớn hơn rượu vang trắng chất lượng thấp.

DOI: https://doi.org/10.34238/tnu-jst.10324

Email: vannq@tnue.edu.vn

1. Giới thiệu

Ngành công nghiệp rượu cho thấy xã hội ngày càng quan tâm đến chất lượng và giá cả của rượu vang. Giá rượu phụ thuộc vào các đánh giá cảm quan của những người nếm rượu, ý kiến của họ tác động mạnh đến sự thay đổi giá rượu. Định giá rượu vang phụ thuộc vào một yếu tố biến động như vậy ở một mức độ nào đó. Một yếu tố quan trọng khác trong chứng nhận và đánh giá chất lượng rượu là các xét nghiệm hóa lý dựa trên cơ sở phòng thí nghiệm và có tính đến các yếu tố như độ axit, độ pH, đường và các đặc tính hóa học khác. Đối với thị trường rượu vang, sẽ rất đáng quan tâm nếu ta tìm được mối liên quan giữa chất lượng rượu theo đánh giá cảm quan của con người và các đặc tính hóa học của rượu để quá trình chứng nhận, đánh giá chất lượng được đảm bảo cũng như được kiểm soát nhiều hơn.

Đã có nhiều công trình nghiên cứu về phân loại rượu vang theo các yếu tố trồng nho như giống nho, vùng địa lý, loại đất [1], [2]. Hoặc các nghiên cứu về phân loại rượu vang trắng và rượu vang đỏ dựa trên các yếu tố hóa học với các mô hình phân tích như nhận dạng mẫu, mạng nơ ron nhân tạo, phân tích phương sai, phân tích phân biệt... [3], [4]. Tuy nhiên nghiên cứu về mối tương quan giữa các đặc tính hóa học của rượu vang và chất lượng của rượu vang (được đánh giá theo cảm quan của các chuyên gia thẩm rượu) vẫn còn hạn chế. Bài báo này đề xuất sử dụng phương pháp phân tích thành phần chính, một phương pháp thống kê nhiều chiều rất hữu ích trong phân tích dữ liệu lớn, để khám phá kiến thức về mối quan hệ giữa chất lượng và các yếu tố hóa học của tập dữ liệu rượu vang trắng được trích xuất từ [5].

Ý tưởng của phân tích thành phần chính (PCA) là giảm số chiều của tập dữ liệu, đồng thời duy trì càng nhiều "tính biến thiên" (tức là thông tin thống kê) càng tốt. Để làm được điều này ta cần tìm các biến mới không tương quan với nhau, là các tổ hợp tuyến tính của các biến trong tập dữ liệu gốc, giúp tối đa hóa phương sai và hiệp phương sai. Việc tìm các biến mới - gọi là các thành phần chính (PC) như vậy, sẽ dẫn đến việc giải bài toán giá trị riêng/vector riêng. Tài liệu sớm nhất về PCA có từ Pearson [6] và Hotelling [7], nhưng phải đến khi máy tính điện tử trở nên phổ biến rộng rãi thì việc sử dụng PCA trên các tập dữ liệu lớn mới được vận dụng nhiều và khả thi về mặt tính toán. Kể từ đó nhiều biến thể của PCA dành cho các loại dữ liệu đặc biệt đã được nghiên cứu và phát triển trong nhiều lĩnh vực khác nhau. Một số vận dụng của PCA trong các lĩnh vực khoa học liên ngành như khí tượng học, khí hậu học, sinh vật học, tin học (giải quyết bài toán nhận dạng khuôn mặt) có thể xem trong các tài liệu [8] – [12].

Nội dung tiếp theo của bài báo được cấu trúc như sau: phần 2 giới thiệu qua về ý tưởng và mô hình của PCA, phần 3 trình bày cụ thể kết quả phân tích dữ liệu rượu vang dựa trên mô hình PCA và cuối cùng phần 4 đưa ra kết luận của bài báo.

2. Phương pháp phân tích thành phần chính

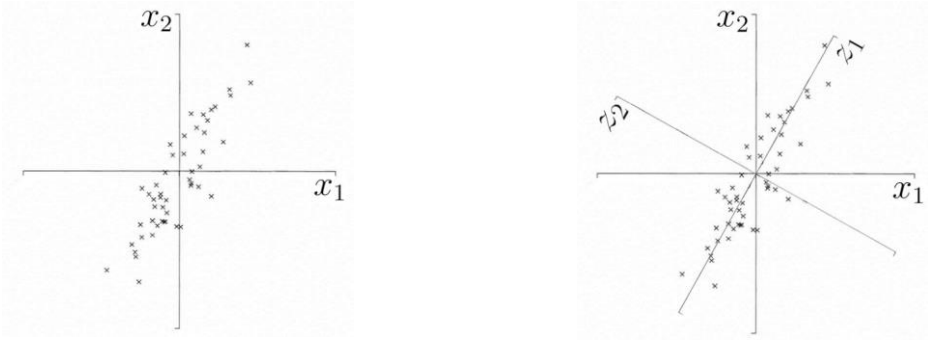
Để có thể hình dung được mục tiêu và nội dung của PCA, chúng ta cùng xét một ví dụ trong không gian hai chiều như sau.

2.1. Ví dụ trong không gian 2 chiều

Xét tập dữ liệu gồm 50 điểm trong không gian hai chiều tương ứng với 50 phép đo trên một cặp biến x_1, x_2. Ví dụ, x_1, x_2 có thể biểu thị nhiệt độ đất và nhiệt độ không khí trong 50 ngày khác nhau tại một địa điểm cụ thể, xem Hình 1(a).

Giả sử ta muốn giảm số chiều của tập dữ liệu này. Việc thay thế cặp biến bằng từng biến thành phần x_1 hoặc x_2 sẽ không hợp lý vì nếu bỏ qua một trong hai biến sẽ làm mất đi một tỷ lệ không hề nhỏ của lượng biến thiên dữ liệu (thông tin của dữ liệu) ban đầu. Tuy nhiên, ta nhận thấy các điểm nằm rải rác khá gần nhau trên một đường thẳng. Điều này ngụ ý rằng có một hàm tuyến tính của x_1, x_2 giải thích mức độ biến thiên của tập dữ liệu tốt hơn hơn đáng kể so với khi chỉ xét riêng x_1 hoặc x_2. Hàm tuyến tính tối đa hóa phương sai này là trục z_1, được gọi là thành phần chính thứ nhất (PC1). Hình 1(b) thể hiện đồ thị của 50 quan sát đối với hai trục z_1 và z_2,

thành phần chính thứ hai (PC2). Từ đồ thị ta có thể thấy rằng gần như tất cả các điểm dữ liệu đều phân tán theo trục z_1, nghĩa là độ biến thiên của tập dữ liệu gốc được giữ nguyên trong chiều mới được xác định bởi z_1.



(a) *(b)*

Hình 1. *(a) Biểu đồ phân tán dữ liệu. (b) Mô tả hình học PCA trong không gian 2 chiều*

Ta có nhận xét sau, Hình 1(b) giống với Hình 1(a) ngoại trừ các trục đã được xoay, sao cho trục z_1 đi qua giữa các điểm dữ liệu và trục thứ hai, z_2 vuông góc với trục đầu tiên. Như vậy, mục tiêu của PCA là sử dụng một phép biến đổi trực giao để biến đổi một tập dữ liệu từ không gian nhiều chiều sang không gian mới ít chiều hơn nhằm thể hiện tối ưu sự biến thiên của tập dữ liệu. Trong không gian mới các trục tọa độ mới được xây dựng sao cho chúng trực giao đôi một với nhau.

2.2. Mô hình toán học của PCA

Cho một mẫu gồm n quan sát của một vector có p biến, $X = (x_1,\ x_2, \dots, x_p)$. Hay nói cách khác, X là một ma trận dữ liệu cỡ $n \times p$, trong đó cột thứ j biểu thị các giá trị quan sát x_j của biến thứ j. Ta định nghĩa các thành phần chính như sau:

Thành phần chính thứ nhất là tổ hợp tuyến tính của các biến $x_1,\ x_2, \dots, x_p$,

$$z_1 = a_1^T X = \sum_{i=1}^{p} a_{i1} x_i \ ,$$

trong đó, vector $a_1 = (a_{11}, a_{21}, \dots, a_{p1})$ được xác định sao cho phương sai của z_1, $Var(z_1)$, là lớn nhất. Tương tự, ***thành phần chính thứ k*** được định nghĩa là $z_k = a_k^T X = \sum_{i=1}^{p} a_{ik} x_i$, trong đó vector $a_k = (a_{1k}, a_{2k}, \dots, a_{pk})$ được xác định sao cho $Var(z_k)$ lớn nhất với các điều kiện ràng buộc $cov(z_k, z_l) = 0$ với mọi $k > l \geq 1$ và $a_k^T a_k = 1$.

Chú ý rằng $Var(z_1) = a_1^T S a_1$, trong đó S là ma trận hiệp phương sai mẫu của các biến $X = (x_1,\ x_2, \dots, x_p)$. Như vậy để tìm thành phần chính thứ nhất, ta cần giải bài toán tối ưu sau:

$$a_1 = \arg\ \max_{a_1} a_1^T S a_1$$
$$\text{với ràng buộc } a_1^T a_1 = 1. \quad (1)$$

Bài toán (1) có thể được giải bằng phương pháp nhân tử Lagrange. Lagrangian của bài toán (1) là

$$\mathcal{L}(a_1, \lambda) = a_1^T S a_1 + \lambda(1 - a_1^T a_1).$$

Nghiệm của bài toán (1) sẽ thỏa mãn hệ phương trình

$$\frac{\partial \mathcal{L}}{\partial a_1} = S a_1 - \lambda a_1 = 0. \quad (2)$$

$$\frac{\partial \mathcal{L}}{\partial \lambda} = 1 - a_1^T a_1 = 0. \quad (3)$$

Từ (2) ta có $$(S - \lambda I_p) a_1 = 0. \quad (4)$$

Điều này suy ra a_1 là một vector riêng của S ứng với giá trị riêng $\lambda = \lambda_1$. Nhân cả hai vế của (4) với a_1^T ta có

$$a_1^T S a_1 = \lambda.$$

Nhận thấy vế trái của đẳng thức trên, $a_1^T S a_1$, chính là hàm mục tiêu trong (1). Vậy hàm mục tiêu đạt giá trị lớn nhất khi λ đạt giá trị lớn nhất. Suy ra $\lambda = \lambda_1$ chính là giá trị riêng lớn nhất của ma trận S. Thành phần chính thứ nhất thể hiện sự biến thiên của tập dữ liệu nhiều nhất.

Lập luận tương tự như vậy ta tìm được a_2 cũng là một vector riêng của S ứng với giá trị riêng $\lambda = \lambda_2$, là giá trị riêng lớn thứ hai của S.

Tổng quát, $Var(z_k) = a_k^T S a_k = \lambda_k$, tức là giá trị riêng lớn thứ k của S là phương sai của thành phần chính thứ k. Thành phần chính thứ k, z_k, thể hiện tỷ lệ lớn thứ k của lượng biến thiên dữ liệu ban đầu.

3. Ứng dụng PCA trong phân tích dữ liệu rượu vang

3.1. Mô tả dữ liệu

Bảng 1. *Ý nghĩa của các biến trong dữ liệu rượu vang trắng*

Biến	Ý nghĩa
1 - fixed acidity (tartaric acid - *g/dm³*)	**độ axit cố định** (Các axit tự nhiên có trong nho, chủ yếu bao gồm axit tartaric, malic, citric hoặc succinic được sử dụng để lên men rượu. Các axit này không dễ bay hơi).
2 - volatile acidity (acetic acid - *g/dm³*)	**tính axit dễ bay hơi** (lượng axit axetic trong rượu vang ở mức quá cao có thể dẫn đến rượu có vị giấm chua).
3 - citric acid (*g/dm³*)	**axit xitric** (được tìm thấy với số lượng nhỏ trong rượu vang, có tác dụng tăng thêm "độ tươi mới" và hương vị cho rượu vang).
4 - residual sugar (*g/dm³*)	**lượng đường dư** (là lượng đường còn lại sau khi quá trình lên men ngừng lên men. Hiếm có những loại rượu có lượng đường dư nhỏ hơn 1 gam/lít. Rượu có hàm lượng đường dư lớn hơn 45 gam/lít được coi là ngọt. Mặt khác, một loại rượu không có vị ngọt được coi là rượu khô).
5 – chlorides (sodium chloride - *g/dm³*)	**lượng muối** trong rượu.
6 - free sulfur dioxide (*mg/dm³*)	**lưu huỳnh đioxit tự do** (dạng SO_2 tự do tồn tại ở trạng thái cân bằng giữa phân tử SO_2 (như một chất khí hòa tan) và ion bisulfit; nó ngăn chặn sự phát triển của vi sinh vật và quá trình oxy hóa rượu vang).
7 - total sunfur dioxide (*mg/dm³*)	**tổng lưu huỳnh đioxit** (lượng SO_2 dạng tự do và liên kết; ở nồng độ thấp, SO_2 hầu như không thể phát hiện được trong rượu vang, nhưng ở nồng độ trên 50 ppm SO_2 trở nên rõ ràng trong mùi vị của rượu).
8 - density (*g/dm³*)	**tỷ trọng** (rất gần với tỷ trọng của nước, nó phụ thuộc vào nồng độ cồn và hàm lượng đường trong rượu).
9 - pH	**pH** (mô tả mức độ axit hoặc bazơ của rượu trên thang từ 0 (rất axit) đến 14 (rất bazơ). Hầu hết các loại rượu đều nằm trong khoảng 3-4 trên thang độ pH).
10 – sulphates (potassium sulphate - *g/dm³*)	**sulphates** (một phụ gia rượu vang hoạt động như một chất chống vi khuẩn và chống oxy hóa).
11 - alcohol (% by volume)	**độ cồn (**phần trăm) của rượu.
12 - quality (score between 0 and 10)	**chất lượng** (điểm từ 0 đến 10).

Tập dữ liệu được khám phá là tập dữ liệu về rượu vang trắng được sản xuất tại khu vực "Vinho Verde" ở Bồ Đào Nha [5]. Tập dữ liệu gồm 4898 mẫu. Mỗi mẫu được thu thập dựa trên 12 đặc tính khác nhau của các loại rượu, một trong số đó là chất lượng, dựa trên dữ liệu cảm quan và phần còn lại là các đặc tính hóa học của rượu bao gồm tỷ trọng, độ axit, nồng độ cồn,... Tất cả các đặc tính hóa học của rượu vang là các biến liên tục. Chất lượng là một biến thứ bậc được xác định bởi những người đánh giá cảm quan, họ (tối thiểu ba người) sẽ nếm thử rượu một cách độc lập và xếp loại rượu theo thang điểm chất lượng từ 0 (rất tệ) đến 10 (xuất sắc). Thứ hạng được ấn định cuối cùng là thứ hạng trung bình do những người nếm thử đưa ra.

Để hiểu rõ hơn về ý nghĩa của các biến, chúng tôi đã phân loại và mô tả cụ thể trong Bảng 1.

Mục tiêu của chúng tôi trong bài báo này là vận dụng PCA để phân tích mối tương quan giữa các biến đặc tính hóa học và biến chất lượng của rượu vang trắng.

3.2. Kết quả phân tích dữ liệu

Để hỗ trợ phân tích dữ liệu, chúng tôi sử dụng ngôn ngữ lập trình R đã tích hợp một số gói lệnh phù hợp với mô hình PCA. Kết quả phân tích được thể hiện trong Hình 2 và Hình 3.

PCA là một công cụ toán học giúp làm giảm chiều dữ liệu, làm tăng khả năng hiển thị hình ảnh của dữ liệu và vẫn giữ lại nhiều thông tin đã có trong dữ liệu gốc. Để làm điều này, PCA biến đổi một tập hợp các biến có thể tương quan thành một tập hợp các biến không tương quan mới, được gọi là các thành phần chính, PC. Bước đầu tiên của việc sử dụng PCA để phân tích dữ liệu là xây dựng ma trận tương quan của tập dữ liệu gồm 4898 mẫu rượu và 12 đặc tính hóa học (biến) tương ứng của chúng. Sau đó, các giá trị riêng và các vectơ riêng của ma trận tương quan được tính toán. Số lượng các thành phần chính có tầm quan trọng được xác định bằng cách sử dụng các giá trị riêng. Các vectơ riêng được xây dựng từ sự kết hợp tuyến tính của các đặc tính ban đầu trong tập dữ liệu. Các hệ số của mỗi đặc tính thể hiện mức độ đóng góp của các biến đó trong các thành phần chính. Cuối cùng, để hình dung sự biến đổi của các đặc tính hóa học trên các PC đã chọn và để đánh giá tầm quan trọng tương đối của từng biến, biểu đồ tương quan và biểu đồ phân loại được vẽ.

Các giá trị riêng thu được từ ma trận tương quan của tập dữ liệu được đưa ra trong Bảng 2. Trong bảng, các giá trị riêng được sắp xếp theo thứ tự từ cao đến thấp. Ứng với mỗi giá trị riêng là phần trăm phương sai tích lũy, thể hiện sự biến động của dữ liệu theo từng thành phần chính. 5 thành phần chính đầu tiên thể hiện 72,81% tổng biến thiên dữ liệu, hay nói cách khác có 72,81% thông tin của dữ liệu được giữ lại trong không gian mới. Một câu hỏi đặt ra là ta nên giữ lại bao nhiêu thành phần chính là tốt nhất?

Bảng 2. *Giá trị riêng của ma trận tương quan*

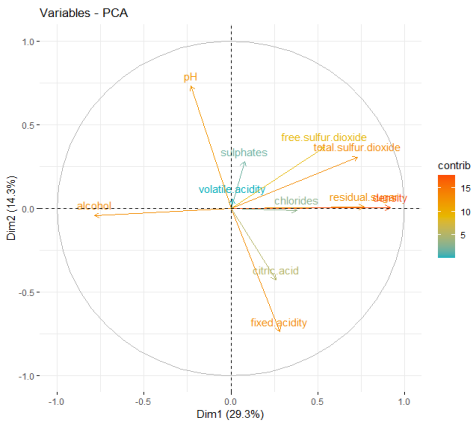
Thành phần chính	Giá trị riêng	Phần trăm phương sai	Phần trăm phương sai tích lũy
1	3,22260270	29,2963881	29,2963881
2	1,57239552	14,2945047	43,59089
3	1,22164229	11,1058390	54,69673
4	1,01930424	9,2664022	63,96313
5	0,97407794	8,8552540	72,81839
6	0,93747708	8,5225189	81,34091
7	0,72703917	6,6094470	87,95035
8	0,60090010	5,4627281	93,41308
9	0,41447764	3,7679785	97,18106
10	0,28941175	2,6310159	99,81208
11	0,02067159	0,1879236	100,00000

Thật không may, không có phương pháp khách quan nào được chấp nhận là tốt để quyết định nên lựa chọn bao nhiêu thành phần chính là đủ. Điều này sẽ phụ thuộc vào từng lĩnh vực ứng dụng cụ thể và từng tập dữ liệu cụ thể. Trong thực tế, chúng tôi có xu hướng xem xét một số thành phần chính đầu tiên để tìm ra các đặc tính thú vị trong dữ liệu. Trong phân tích trên, nếu chúng ta chọn chỉ giữ lại hai thành phần chính thì tổng biến thiên của tập dữ liệu gốc giảm xuống còn 43,58%. Một trong những mục tiêu của PCA là giảm chiều dữ liệu nhưng vẫn thể hiện được độ biến thiên tối đa của tập dữ liệu gốc, do đó chúng tôi quyết định rằng ma trận dữ liệu được mô tả tốt nhất bởi ba thành phần chính đầu tiên.

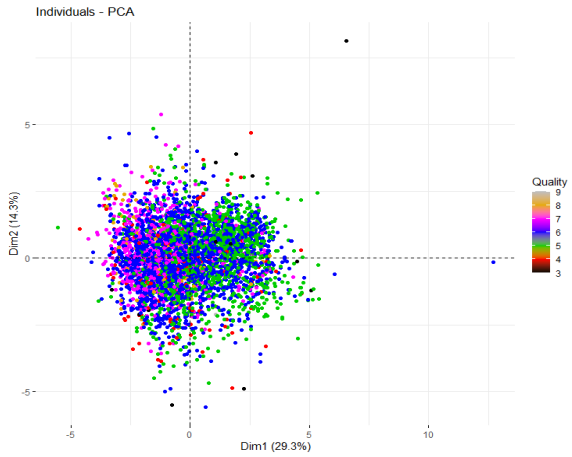
Các vectơ riêng của ma trận tương quan dữ liệu được trình bày trong Bảng 3. Các vectơ riêng được chia tỷ lệ từ +1 đến -1. Các biến có các giá trị được in đậm trong bảng mang ý nghĩa thống kê, nghĩa là thể hiện mức độ đóng góp cao hay thấp của các biến ban đầu lên thành phần chính của mô hình PCA.

Bảng 3. *Vector riêng ứng với 5 thành phần chính (PC) đầu tiên*

Đặc tính hóa học (biến)	PC1	PC2	PC3	PC4	PC5
Độ axit cố định	0,28170958	**-0,736891965**	0,1395388	0,02923923	0,24205194
Tính axit dễ bay hơi	0,01046093	0,056721611	**-0,6528882**	0,30296682	**0,62572265**
Axit xitric	0,25851826	**-0,427953994**	**0,5621494**	0,15358238	0,04774175
Lượng đường dư	**0,76728279**	0,009553841	-0,2338992	-0,27470451	0,01398569
Lượng muối trong rượu	0,38123927	-0,010949766	-0,1190317	**0,70132225**	-0,34192941
Lưu huỳnh dioxit tự do	**0,53907868**	**0,367844366**	0,3039784	-0,30077880	0,20309459
Tổng lưu huỳnh dioxit	**0,72996815**	0,306353708	0,1343161	-0,04899785	0,30007321
Tỷ trọng	**0,91839827**	0,005511851	-0,1403209	-0,02530880	-0,09247508
pH	-0,22889888	**0,730628512**	0,1376832	0,09687477	-0,12809803
Sulphates	0,07843082	0,279385644	**0,4777223**	**0,46577163**	**0,36220017**
Độ cồn của rượu	**-0,78534976**	-0,041782782	0,1170894	-0,12617547	**0,34224014**



Hình 2. *Đồ thị tương quan của dữ liệu rượu vang trắng*



Hình 3. *Đồ thị điểm theo chất lượng của dữ liệu rượu vang trắng*

Ví dụ, thành phần chính đầu tiên cung cấp các giá trị dương cao cho cả tỷ trọng, lượng đường dư, tổng lưu huỳnh dioxit và có giá trị âm cao cho lượng cồn trong rượu. Do đó, PC đầu tiên cung cấp thông tin về mùi vị đặc trưng của rượu vang trắng. Thực tế cho thấy, nồng độ cồn của rượu tương quan nghịch với tỷ trọng của rượu. Trong quá trình lên men, do đường chuyển hóa chậm thành rượu nên nếu đường giảm thì nồng độ rượu tăng và tỷ trọng chung giảm. Do đó rượu vang trắng có vị ngọt hơn so với rượu vang đỏ và nồng độ cồn sẽ thấp hơn. PC thứ hai liên quan đến pH, độ axit cố định, axit citric - các biến liên quan đến tính axit và bazơ trong rượu. Người ta tin rằng axit cần có trong rượu vang để tăng thêm mùi vị và để kết hợp được với nhiều loại thức ăn. Thật thú vị khi thấy rằng chất lượng và nồng độ axit trong rượu vang trắng có mối tương quan nghịch, chất lượng của rượu vang trắng càng giảm khi nồng độ axit tăng lên. PC thứ ba đưa ra mối quan hệ về tính axit dễ bay hơi và sulphates. Như vậy PC thứ ba đại diện cho tính chất chống oxi hóa và tính chống nhiễm khuẩn trong từng loại rượu vang trắng. Ba thành phần chính đầu tiên được mô tả ở trên cung cấp một số thông tin cơ bản về mùi vị đặc trưng của rượu vang trắng.

Hình 2 thể hiện mối tương quan giữa các đặc tính hóa học của rượu với hai thành phần chính đầu tiên trong mô hình PCA. Những biến nào có tương quan cao với PC1 và PC2 thì biến đó đóng vai trò quan trọng trong mô tả dữ liệu, ngược lại những biến có tương quan thấp thì có thể được loại bỏ ra khỏi mô hình. Trong hình 2, các biến nồng độ cồn, pH, lượng đường dư, axit cố định, tổng lưu huỳnh dioxit có ảnh hưởng lớn tới mô hình PCA. Còn các biến tính axit dễ bay hơi, lượng muối thì có ảnh hưởng nhỏ. Hình 2 cũng minh họa sự tương quan giữa các biến. Những biến có mối tương quan thuận thì được xếp thành một nhóm. Những biến có mối tương quan nghịch thì nằm ở các góc phần tư đối nhau. Chẳng hạn, nồng độ cồn và tỷ trọng là hai biến tương quan nghịch, pH tương quan nghịch với axit citric và axit cố định, tỷ trọng và lượng đường dư tương quan thuận.

Trong Hình 3, hầu hết các loại rượu vang trắng chất lượng cao (chất lượng 6, 7 và 8) nằm ở phía bên trái và rượu vang trắng chất lượng thấp (chất lượng 3, 4 và 5) tập trung chủ yếu ở phía bên phải của đồ thị. Do đó, chúng ta có thể kết luận rằng rượu vang trắng chất lượng cao có nồng độ cồn và độ pH lớn hơn rượu vang trắng chất lượng thấp.

4. Kết luận

Bài báo này trình bày ngắn gọn ý nghĩa và nội dung của mô hình PCA, một mô hình được sử dụng phổ biến với mục tiêu làm giảm số chiều của không gian chứa dữ liệu gốc. Ưu điểm của phương pháp PCA là trong không gian mới có số chiều nhỏ hơn không gian ban đầu, các dữ liệu khi được biểu diễn không bị trùng lặp, không bị mất thông tin quá nhiều. Đồng thời chúng ta còn khám phá thêm được những thông tin quý giá mới khi mà tại chiều không gian cũ những thông tin này bị che mất.

Với những phân tích trên, chúng tôi đề xuất sử dụng phương pháp PCA trong phân tích mối tương quan giữa các đặc tính hóa học và chất lượng của rượu vang trắng. Mục tiêu của chúng tôi là phân loại rượu vang trắng dựa trên các đặc tính hóa học của nó. Kết quả phân tích cho thấy các đặc tính hóa học bao gồm: lượng đường dư, tổng lưu huỳnh dioxit, tỷ trọng, nồng độ cồn, pH, độ axit cố định thể hiện cho mùi vị đặc trưng của rượu vang trắng. Đồng thời, rượu vang trắng chất lượng cao sẽ có nồng độ cồn và độ pH lớn hơn rượu vang trắng chất lượng thấp.

TÀI LIỆU THAM KHẢO/ REFERENCES

[1] E.P. Pérez - Álvarez *et al.,* "Classification of wines according to several factors by ICP-MS multi-element analysis," *Food Chemistry,* vol. 270, pp. 273–280, 2019.

[2] J. Spercova and M. Suchanek, "Multivariate classification of wines from different Bohemian regions (Czech Republic)," *Food Chemistry*, vol. 93, pp. 659–663, 2005.

[3] N. V. Bondarev, "Clustering and classification of Red Wines according to physical-chemical properties using data mining methods," *Russian Journal of General Chemistry*, vol. 93, pp. 2325–2339, 2023.

[4] W. O. Kwan and B. R. Kowalski, "Classification of Wines by applying pattern recognition to chemical composition data," *Journal of Food Science*, vol. 43, pp. 1320–1323, 1978.

[5] University of California at Irvine, "Wine Data Set," *UCI Machine Learning Repository*, 1991. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/wine+quality. [Accessed May 5, 2020].

[6] K. Pearson, "On lines and planes of closest fit to systems of points in space," *The Philosophical Magazine: A Journal of Theoretical Experimental and Applied Physics*, vol. 2, pp. 559–572, 1901.

[7] H. Hotelling, "Analysis of a complex of statistical variables into principal components," *Journal of Educational Psychology*, vol. 24, no. 6, pp. 417–441, 1933.

[8] F.I. Amy and P.P. Marvin, "Applications of Principal Component Analysis to Horticultural Research," *HortScience*, vol.26, no. 4, pp. 334–338, 1991.

[9] T. J. Ian, "Principal component analysis: A beginner's guide," *Weather,* vol. 45, no. 10, pp. 375–382, 1990.

[10] A. Herve and J. W. Lynne, "Principal component analysis," *WIREs Computational Satatistics*, vol. 2, pp. 433–459, 2010.

[11] K. P. Pramod *et al.*, "Image Processing using Principle Component Analysis," *International Journal of Computer Applications*, vol. 15, no. 4, pp. 37–40, 2011.

[12] C. Li, Y. Diao, H. Ma, and Y. Li, "A Statistical PCA Method for Face Recognition," *Intelligent Information Technology Application*, vol. 03, pp. 376–380, 2008.