

# TRƯỜNG ĐẠI HỌC SÀI GÒN

KHOA CÔNG NGHỆ THÔNG TIN





BÁO CÁO TIỂU LUẬN PHÂN TÍCH DỮ LIỆU

**THỐNG KÊ MÔ TẢ TRÊN TẬP DỮ LIỆU VỀ PHÂN LOẠI CHẤT LƯỢNG RƯỢU ĐỎ**

**Giảng viên hướng dẫn: Đỗ Như Tài**

**Thông tin thành viên:**

Lê Nguyễn Quốc Hưng 3121410240

Trần Bá Tiên 3122410410

Dương Huỳnh Gia 3121410167

Bùi Thành Công 3122410042

***TP. Hồ Chí Minh, tháng 10/2025***

NHẬN XÉT VÀ ĐÁNH GIÁ CỦA GIẢNG VIÊN

…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..…………………………………..……………………………………….……………………..…………………………………..

# LỜI CẢM ƠN

Chúng em xin gửi lời cảm ơn chân thành và sự tri ân sâu sắc đối với các thầy cô của trường Đại Học Sài Gòn, đặc biệt là các thầy cô ở khoa Công Nghệ Thông Tin đã tạo điều kiện cho em tiếp cận và tìm hiểu để hoàn thành đồ án môn học lần này.

Và chúng em cũng xin chân thành đặc biệt cảm ơn thầy Đỗ Như Tài, thầy là giảng viên giảng dạy đã nhiệt tình hướng dẫn giúp chúng em hoàn thành được đồ án lần này.

Trong quá trình nghiên cứu và làm bài báo cáo đồ án, kiến thức cũng như kinh nghiệm thực tế còn nhiều hạn chế nên bài báo cáo không thể tránh khỏi những thiếu sót, chúng em rất mong nhận được ý kiến đóng góp của thầy, cô để em học hỏi được nhiều kỹ năng, kinh nghiệm và sẽ hoàn thành tốt hơn cho những bài báo cáo sắp tới.

Chúng em xin chân thành cảm ơn.

# MỤC LỤC

[LỜI CẢM ƠN 1](#_heading=h.30j0zll)

[MỤC LỤC 2](#_heading=h.os4kp4kuwm50)

[I. GIỚI THIỆU 3](#_heading=h.ua56i0g34v20)

[1. Giới thiệu về rượu vang đỏ (Red Wine) 3](#_heading=h.66s6y96bvof2)

[2. Giới thiệu tập dữ liệu 3](#_heading=h.m9i74nuvfbr4)

[2.1. Nguồn dữ liệu 3](#_heading=h.kf2cn9b51qbc)

[2.2. Giải thích các biến 3](#_heading=h.ou6kxl3s9gxr)

[II. ĐỊNH NGHĨA VẤN ĐỀ (DEFINE PROBLEM) 6](#_heading=h.2mocjcycxqnx)

[1. Nạp dữ liệu 7](#_heading=h.a7gvja14xqbd)

[2. Loại bỏ dữ liệu trùng lặp 7](#_heading=h.46cxqg7s07cq)

[3. Thay thế dữ liệu và thay đổi định dạng 7](#_heading=h.7fnzmansmfk2)

[4. Xử lý dữ liệu thiếu 8](#_heading=h.lpq3jgp4xo2z)

[5. Thống kê mô tả 8](#_heading=h.5rcr39u2d9pa)

[6. Thống kê theo nhóm chất lượng 9](#_heading=h.6ldclf76r6u4)

[7. Trực quan hóa 10](#_heading=h.fg3nwvxz51h1)

[III. KẾT LUẬN 10](#_heading=h.2srdqcqlhx5)

[TÀI LIỆU THAM KHẢO 12](#_heading=h.bd6z3290hgfj)

# I. GIỚI THIỆU

## 1. Giới thiệu về rượu vang đỏ (Red Wine)

Rượu vang đỏ là loại rượu được sản xuất chủ yếu từ các giống nho đỏ hoặc nho đen, với đặc điểm nổi bật là màu sắc đậm và hương vị phong phú. Màu đỏ của rượu đến từ vỏ nho – nơi chứa nhiều hợp chất anthocyanin và tannin, góp phần tạo nên vị chát đặc trưng và khả năng bảo quản lâu dài của rượu.

Quá trình sản xuất rượu vang đỏ bao gồm các bước: thu hoạch nho, ép và lên men cùng vỏ, ủ trong thùng gỗ sồi (oak barrel) và đóng chai. Các yếu tố như độ chua (acidity), hàm lượng đường, hàm lượng cồn, hàm lượng SO₂, và pH ảnh hưởng mạnh mẽ đến chất lượng cảm quan của rượu (mùi, vị, độ đậm và hậu vị).

Đánh giá chất lượng rượu vang thường được thực hiện bởi các chuyên gia nếm thử (wine tasters), dựa trên cảm giác vị giác và khứu giác, sau đó gán điểm từ 3 đến 8 (theo thang điểm trong tập dữ liệu Kaggle)

## 2. Giới thiệu tập dữ liệu

### 2.1. Nguồn dữ liệu

* Nguồn: [1]
* Số dòng (records): 1,599 mẫu rượu.
* Số biến (features): 12 biến độc lập + 1 biến mục tiêu (“quality”).
* Đơn vị đo: phần lớn ở dạng gram/lít (g/L) hoặc phần trăm (%).

### 2.2. Giải thích các biến

**1. Fixed acidity (Độ axit cố định):** Hầu hết các loại axit trong rượu là “cố định” hoặc “không bay hơi” (không dễ bay hơi). Độ axit là đặc tính được xác định bởi tổng lượng axit mà mẫu rượu chứa. Tổng độ axit có thể được định lượng theo cách tổng hợp (toàn phần) hoặc chia thành hai nhóm: độ axit cố định và độ axit bay hơi. Độ axit cố định tương ứng với nhóm axit hữu cơ có độ bay hơi thấp, như axit malic, lactic, tartaric hoặc citric, và là đặc điểm vốn có của mẫu rượu.

**2. Volatile acidity (Độ axit bay hơi):** Là lượng axit axetic (acetic acid) trong rượu. Nếu nồng độ quá cao, rượu có thể có mùi vị chua gắt giống giấm.  
 Độ axit bay hơi gồm nhóm axit hữu cơ chuỗi ngắn có thể tách ra khỏi mẫu rượu bằng phương pháp chưng cất, như axit formic, axetic, propionic và butyric.

**3. Citric acid (Axit citric):** Xuất hiện với lượng nhỏ trong rượu, axit citric giúp tăng độ “tươi” và hương vị. Đây là một axit hữu cơ yếu, không màu, xuất hiện tự nhiên trong các loại trái cây họ cam chanh. Trong sinh hóa, axit citric là một hợp chất trung gian trong chu trình axit citric – quá trình chuyển hóa cơ bản ở mọi sinh vật hiếu khí.

**4. Residual sugar (Đường dư):** Là lượng đường còn lại sau khi quá trình lên men kết thúc. Thông thường, rất hiếm khi rượu có ít hơn 1 gram đường mỗi lít. Đường dư là phần đường chưa được lên men trong rượu đã hoàn thành, đo bằng gram đường trên mỗi lít (g/L). Lượng đường dư ảnh hưởng đến độ ngọt của rượu, và trong Liên minh Châu Âu, mức đường dư này còn được quy định trong nhãn phân loại rượu.

**5. Chlorides (Hàm lượng muối):** Là lượng ion chloride (Cl⁻) trong rượu. Trong quá trình sản xuất rượu vang đỏ, việc chiết tách ion chloride từ vỏ nho khiến nồng độ chloride cao hơn. Dịch nho đỏ nên chứa không quá 356 mg/L ion chloride, để sau khi lên men, rượu thành phẩm không vượt mức hợp pháp tối đa 606 mg/L (vì 356 mg/L × 1.7 ≈ 606).

**6. Free sulfur dioxide (SO₂ tự do):** SO₂ tự do tồn tại trong trạng thái cân bằng giữa SO₂ phân tử (dạng khí hòa tan) và ion bisulfite. SO₂ tự do là phần sulfite sẵn sàng phản ứng, có đặc tính kháng khuẩn và chống oxy hóa. Ngược lại, SO₂ liên kết là phần đã phản ứng với các phân tử khác trong rượu. Tổng nồng độ SO₂ (free + bound) gọi là tổng lượng sulfur dioxide.

**7. Total sulfur dioxide (Tổng SO₂):** Là tổng của hai dạng SO₂ tự do và SO₂ liên kết. Ở nồng độ thấp, SO₂ hầu như không thể cảm nhận được, nhưng nó đóng vai trò quan trọng trong việc bảo quản rượu. Tổng lượng SO₂ thể hiện phần SO₂ tự do trong rượu cộng với phần đã liên kết với các hợp chất khác như aldehyde, sắc tố hoặc đường.

**8. Density (Mật độ):** Mật độ của rượu phụ thuộc vào hàm lượng cồn và đường. Dụng cụ đo mật độ thường dùng là hydrometer – một ống thủy tinh có đầu nặng, dùng để đo tỷ trọng chất lỏng so với nước tinh khiết (gọi là “specific gravity”). Người làm rượu sử dụng nó để theo dõi mật độ của nước nho, rượu đang lên men và rượu thành phẩm.

**9. pH:** Chỉ số đo mức độ axit hoặc bazơ của rượu, với thang đo từ 0 (rất axit) đến 14 (rất bazơ). Phần lớn rượu có pH trong khoảng 3–4.  
 Rượu có pH cao (>3.65) có thể gặp nhiều vấn đề trong quá trình lên men và bảo quản, do dễ bị vi sinh vật xâm nhập. Vì vậy, người làm rượu thường thêm SO₂ (dưới dạng kali metabisulfit) để ổn định rượu trong quá trình ủ.

**10. Sulphates (Sunphat):** Là chất phụ gia được thêm vào rượu, có thể tạo ra khí SO₂ giúp ức chế vi khuẩn. Sulphate có sẵn ở mức thấp tự nhiên trong mọi loại rượu, nhưng nhà sản xuất thường thêm vào để bảo quản và ngăn hư hỏng do vi sinh vật. Tuy nhiên, với một số người nhạy cảm, SO₂ có thể gây đau đầu hoặc nghẹt mũi sau khi uống rượu.

**11. Alcohol (Độ cồn):** Là phần trăm nồng độ cồn trong rượu, biểu thị mức độ rượu mạnh hay nhẹ.

**12. Quality (Chất lượng):** Là biến đầu ra của bộ dữ liệu, dựa trên đánh giá cảm quan (sensory score) của người thử rượu, với thang điểm từ 3 đến 8.

# II. ĐỊNH NGHĨA VẤN ĐỀ (DEFINE PROBLEM)

1. **Input:**

* **fixed acidity**: Tổng các acid không bay hơi (malic, lactic, tartaric, citric). Ảnh hưởng trực tiếp đến vị chua của rượu.
* **volatile acidity**: Lượng acid bay hơi (chủ yếu là acetic acid). Quá cao → vị giấm, mùi khó chịu.
* **citric acid**: Acid yếu tạo cảm giác “tươi” và “fresh” cho rượu. Có tự nhiên trong trái cây họ cam chanh.
* **residual sugar**: Lượng đường còn lại sau khi lên men (g/l). Quyết định độ ngọt của rượu.
* **chlorides**: Lượng muối trong rượu. Quá cao ảnh hưởng xấu đến hương vị và vượt giới hạn pháp lý.
* **free sulfur dioxide** (free SO₂): Dạng SO₂ tự do – có tác dụng chống oxi hóa và kháng khuẩn.
* **total sulfur dioxide** (total SO₂): Tổng SO₂ = tự do + liên kết. Ở mức thấp thì khó nhận biết, nhưng cao quá sẽ ảnh hưởng mùi vị.
* **density**: Tỉ trọng của rượu, phụ thuộc vào nồng độ cồn và đường. Đo bằng hydrometer.
* **pH**: Độ axit/bazơ của rượu (0–14). Rượu thường ở mức 3–4. pH cao → dễ hư, vi sinh phát triển.
* **sulphates**: Hợp chất chứa lưu huỳnh, giúp sinh ra SO₂, bảo quản rượu, chống vi khuẩn.
* **alcohol**: Tỷ lệ % cồn trong rượu – ảnh hưởng mạnh đến hương vị và “độ nặng”.

1. **Output:**

* **quality**: Điểm chất lượng rượu (từ 3–8) dựa trên đánh giá cảm quan của chuyên gia.

## 1. Nạp dữ liệu



*Hình II.1: Hình ảnh nạp data*

* Bộ dữ liệu có 1599 mẫu và 12 thuộc tính (11 biến đầu vào + 1 nhãn quality).
* Các thuộc tính mô tả đặc tính hóa học của rượu đỏ như độ chua, lượng đường, độ cồn, pH,...

## 2. Loại bỏ dữ liệu trùng lặp

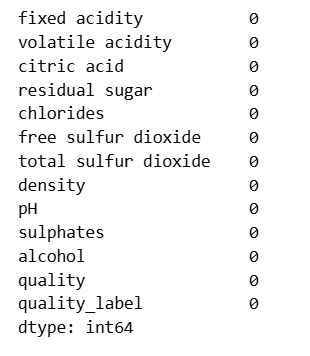
Nhận xét:

* Nếu có dòng trùng, việc loại bỏ giúp tránh sai lệch thống kê (double count).
* Sau khi xử lý, dữ liệu còn khoảng 1599 dòng, nghĩa là hầu như không có trùng lặp

## 3. Thay thế dữ liệu và thay đổi định dạng

* Việc phân nhóm quality\_label giúp dễ dàng phân tích rượu ngon (good) và rượu kém (bad).
* Khoảng 80% mẫu là rượu “bad” (chất lượng trung bình 5–6), số “good” ít hơn nhiều → dữ liệu mất cân bằng lớp.

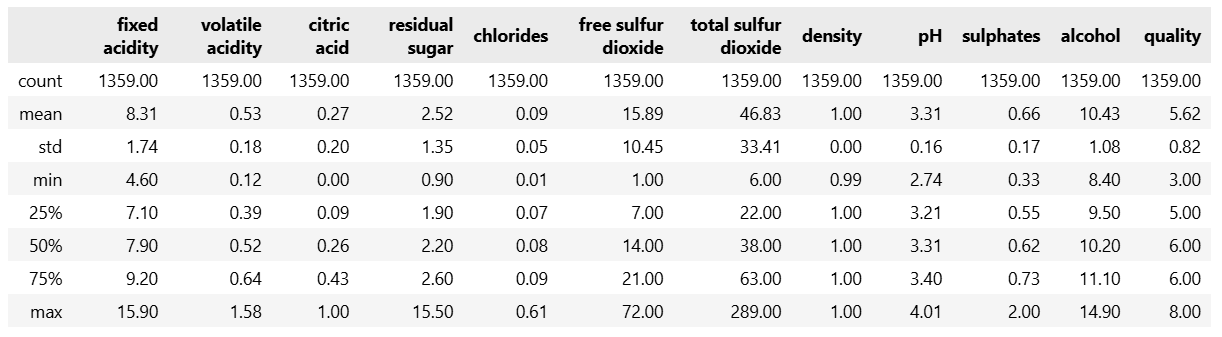
## 4. Xử lý dữ liệu thiếu



*Hình II.4: Hình ảnh xử lý dữ liệu thiếu*

* Kết quả cho thấy không có giá trị thiếu (NaN).
* Đây là một tập dữ liệu sạch, có thể tiến hành trực tiếp các bước EDA mà không cần điền giá trị.

## 5. Thống kê mô tả



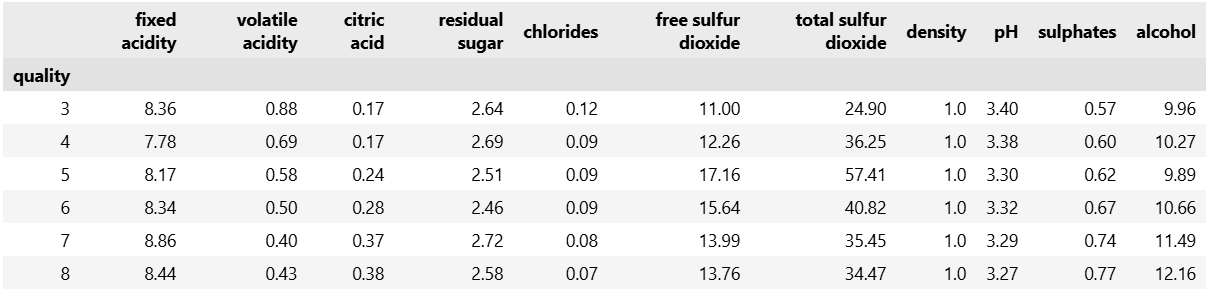
*Hình II.5: Hình ảnh thống kê mô tả*

* fixed acidity: Trung bình ~8.3 g/dm³. Độ chua ổn định, ít biến động.
* volatile acidity: Trung bình 0.53 g/dm³. Cao → có thể tạo mùi giấm nhẹ.
* citric acid: Rất thấp (0.27), góp phần tạo vị “tươi mát”.
* residual sugar: Khoảng 2.5 g/l → rượu khá khô (ít ngọt).
* chlorides:Rất nhỏ (0.08), phù hợp tiêu chuẩn rượu đỏ.
* free SO₂ / total SO₂: Lượng trung bình thấp, vừa đủ để chống oxi hóa.
* density: Khoảng 0.9968 g/cm³, thấp hơn nước → đúng với rượu có cồn.
* pH: Trung bình 3.31 → môi trường acid, ổn định cho bảo quản.
* sulphates: Khoảng 0.66 → giúp ổn định rượu và bảo quản.
* alcohol: Trung bình 10.42%, dao động 8.4–14.9%.
* quality: Dao động từ 3 đến 8, trung bình 5.64 → phần lớn rượu chất lượng trung bình.

Nhận xét tổng quát:

* Bộ dữ liệu ổn định, ít biến động cực đoan.
* Rượu trong tập dữ liệu đa số có nồng độ cồn thấp – trung bình, ít đường, acid ổn định → là loại rượu khô, vị hơi chua nhẹ.

## 6. Thống kê theo nhóm chất lượng

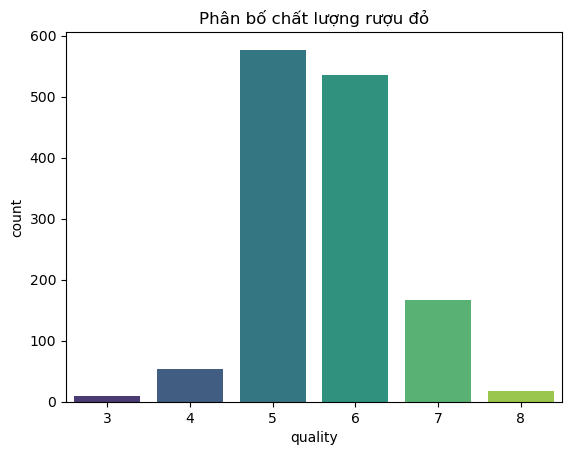


*Hình II.6: Hình ảnh thống kê theo nhóm chất lượng*

* Rượu có quality ≥7 có đặc điểm:
* Alcohol cao hơn (~12–13%)
* Volatile acidity thấp hơn (~0.35) → ít mùi giấm
* Citric acid cao hơn (~0.4) → vị tươi, cân bằng
* Rượu có quality = 5–>6 có:
* Cồn thấp (~9–10%)
* Độ acid bay hơi cao hơn
* Ít citric acid hơn

→ Alcohol và volatile acidity là hai yếu tố ảnh hưởng mạnh nhất đến chất lượng rượu.

## 7. Trực quan hóa



*Hình II.7: Hình ảnh trực quan hóa*

* Cột quality=5 và quality=6 chiếm đa số.
* Rượu có điểm 7–8 (tốt) ít, tạo nên dữ liệu bất cân đối

# III. KẾT LUẬN

* Độ cồn (alcohol) có mối quan hệ tích cực với chất lượng rượu — rượu có nồng độ cồn cao thường được đánh giá điểm chất lượng cao hơn.
* Độ axit bay hơi (volatile acidity) có mối quan hệ ngược chiều với chất lượng — khi hàm lượng axit bay hơi tăng, rượu dễ có mùi vị gắt, làm giảm chất lượng cảm quan.
* Axit citric và sunphat (sulphates) góp phần tích cực vào hương vị tươi và khả năng bảo quản của rượu.
* Các biến như pH, mật độ (density), đường dư (residual sugar) hay chlorides có sự dao động nhỏ, ảnh hưởng ít rõ rệt đến điểm chất lượng.

# 

# TÀI LIỆU THAM KHẢO

[1] “Red Wine Quality EDA & Classification,” Kaggle, UCI Machine Learning Repository. <https://www.kaggle.com/code/eisgandar/red-wine-quality-eda-classification>.

[Truy cập: 10-10-2025].