Trường đại học Sài Gòn

A blue button with white text

AI-generated content may be incorrect.Khoa Công Nghệ Thông Tin

BÁO CÁO TIỂU LUẬN PHÂN TÍCH DỮ LIỆU

PHÂN TÍCH CHẤT LƯỢNG RƯỢU VANG

Giảng viên hướng dẫn: Đỗ Như Tài

Thông tin thành viên nhóm:

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 GIẢNG VIÊN

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

**LỜI MỞ ĐẦU**

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!

## Tóm tắt

Báo cáo này trình bày quá trình phân tích khám phá dữ liệu (Exploratory Data Analysis) và trực quan hóa cho bộ dữ liệu **Wine Quality (Red Wine)** từ UCI. Mục tiêu là khám phá phân bố các đặc trưng hoá học, hiểu mối quan hệ giữa các biến và chất lượng rượu, đồng thời rút ra những nhận xét ban đầu để làm nền tảng cho bài toán phân loại chất lượng.

## 1. Giới thiệu

Bộ dữ liệu Wine Quality chứa các đặc trưng hoá học đo đạc trên các mẫu rượu vang đỏ, kèm theo điểm chất lượng (quality) do chuyên gia đánh giá. Phân tích trực quan giúp ta nắm được đặc điểm dữ liệu, phát hiện ngoại lệ, và tìm những biến khả dĩ ảnh hưởng tới chất lượng.

### Mục tiêu báo cáo

* Mô tả dữ liệu và kiểm tra chất lượng dữ liệu.
* Trực quan hoá phân bố các thuộc tính (univariate).
* Khám phá mối quan hệ giữa các biến (bivariate & multivariate).
* Trình bày kết luận chính và đề xuất bước tiếp theo cho mô hình hóa.

## 2. Nguồn dữ liệu và thông tin chung

* **Nguồn:https://www.kaggle.com/datasets/uciml/red-wine-quality-cortez-et-al-2009/data**
* **Số mẫu:** 1599
* **Số thuộc tính:** 12 (11 đặc trưng hoá học + 1 mục tiêu quality)

### Danh sách thuộc tính

* fixed acidity: Axit cố định
* volatile acidity: Axit bay hơi
* citric acid: Axit citric
* residual sugar: Lượng đường còn lại
* chlorides: Hàm lượng muối
* free sulfur dioxide:  Lượng SO₂ tự do
* total sulfur dioxide: Tổng lượng SO₂
* density: Mật độ
* pH: Độ pH
* sulphates: Hàm lượng sunfat
* alcohol: Nồng độ cồn
* quality (mục tiêu): Chất lượng rượu

## 3. Phân tích dữ liệu

1. **Nạp dữ liệu và kiểm tra thông tin cơ bản (head, info, describe).**

* Nạp dữ liệu

A computer screen shot of a black screen

AI-generated content may be incorrect.

*Hình 1: Nạp dữ liệu từ file*

* Kiểm tra thông tin

A screenshot of a computer

AI-generated content may be incorrect.

*Hình 2: Kiểm tra thông tin file*

* Lấy 10 dữ liệu đầu tiên trong bảng data

A screenshot of a computer

AI-generated content may be incorrect.

*Hình 3: Lấy 10 dữ liệu đầu tiên của data*

* Lấy 10 dữ liệu cuối cùng trong bảng data

A screenshot of a graph

AI-generated content may be incorrect.

*Hình 4: Lấy 10 dữ liệu cuối cùng của data*

* Thống kê mô tả tóm tắt dữ liệu trong bảng data

A screenshot of a computer screen

AI-generated content may be incorrect.

*Hình 5: Mô tả dữ liệu*

**Độ axit cố định (fixed acidity):** giá trị trung bình là **8.31**, cao nhất là **15.9**.

**Độ axit bay hơi (volatile acidity):** giá trị trung bình là **0.52**, cao nhất là **1.58**.

**Axit citric (citric acid):** giá trị trung bình là **0.27**, cao nhất là **1.00**.

**Lượng đường dư (residual sugar):** giá trị trung bình là **2.53**, cao nhất là **15.5**.

**Hàm lượng chloride (chlorides):** giá trị trung bình là **0.08**, cao nhất là **0.61**.

**Lượng sulfur dioxide tự do (free sulfur dioxide):** giá trị trung bình là **15.87**, cao nhất là **72**.

**Tổng lượng sulfur dioxide (total sulfur dioxide):** giá trị trung bình là **46.46**, cao nhất là **289**.

**Khối lượng riêng (density):** giá trị trung bình là **0.99**, cao nhất là **1.00**.

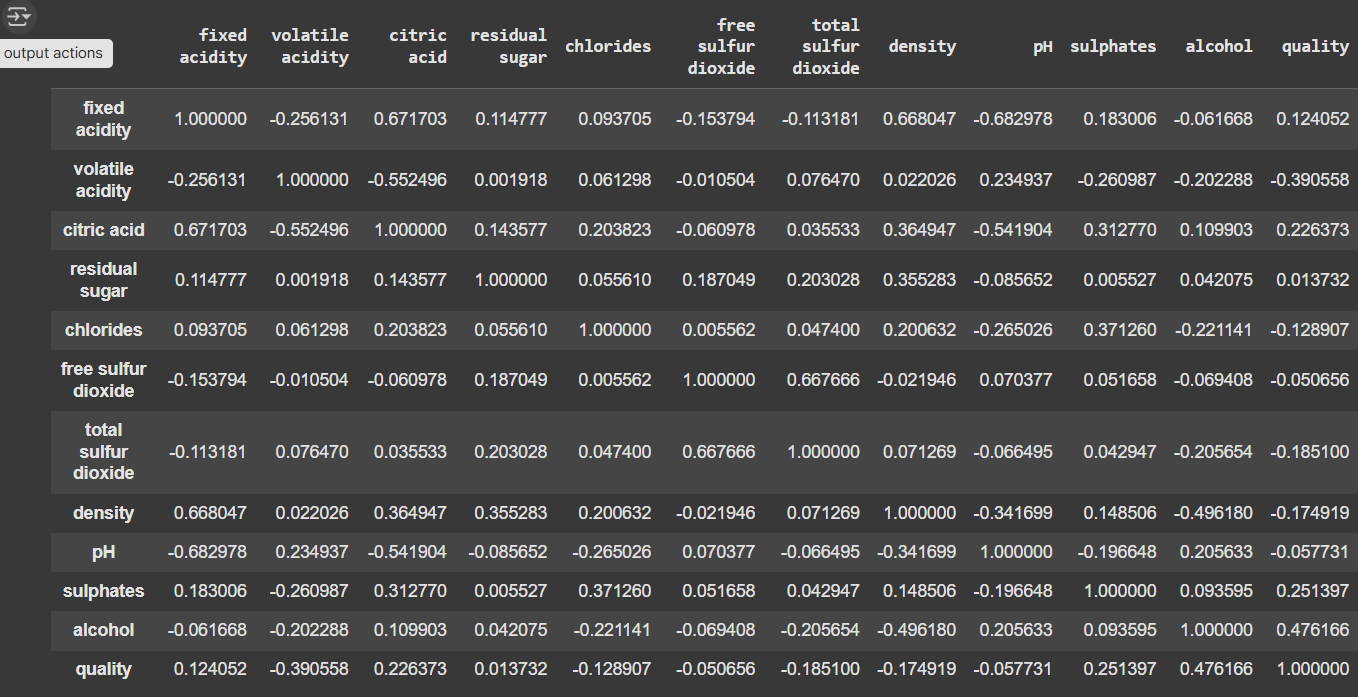
**Độ pH (pH):** giá trị trung bình là **3.31**, cao nhất là **4.01**.

**Hàm lượng sulphates (sulphates):** giá trị trung bình là **0.65**, cao nhất là **2.00**.

**Hàm lượng cồn (alcohol):** giá trị trung bình là **10.42**, cao nhất là **14.90**.

**Chất lượng rượu (quality):** giá trị trung bình là **5.63**, cao nhất là **8**.

* Tương quan dữ liệu:



*Hình 6: Mô tả sự tương quan dữ liệu*

* Loại bỏ giá trị lặp

A screenshot of a computer

AI-generated content may be incorrect.

*Hình 7: Xóa các giá trị lặp*

1. **Kiểm tra giá trị thiếu và kiểu dữ liệu.**

* Kiểm tra giá trị thiếu

A screen shot of a computer

AI-generated content may be incorrect.

*Hình 7: Kiểm tra dữ liệu lặp*

* Kiểm tra kiểu dữ liệu

A screen shot of a computer

AI-generated content may be incorrect.

*Hình 8: Mô tả kiểu dữ liệu*

1. **Trực quan hoá:**
   1. **Biểu đồ phân phối dữ liệu**

A screenshot of a graph

AI-generated content may be incorrect.

*Hình 9: Biểu đồ boxplot*

A group of blue graphs

AI-generated content may be incorrect.

*Hình 10: Biểu đồ Histogram*

| **Thuộc tính** | **Dạng phân bố** | **Nhận xét chính** |
| --- | --- | --- |
| **fixed acidity** | Lệch phải (Right-skewed) | Đa số mẫu có độ axit cố định trong khoảng 6–9 |
| **volatile acidity** | Gần lệch phải | Phần lớn có độ bay hơi thấp, ít mẫu có độ cao |
| **citric acid** | Lệch phải | Hầu hết giá trị nhỏ hơn 0.5, chỉ vài mẫu có acid cao |
| **residual sugar** | Lệch phải mạnh | Phần lớn rượu có lượng đường dư thấp, chỉ vài mẫu ngọt hơn |
| **Chlorides** | Lệch phải mạnh | Lượng muối (NaCl) thấp ở hầu hết mẫu |
| **free sulfur dioxide** | Lệch phải | Phần lớn giá trị nhỏ hơn 40 |
| **total sulfur dioxide** | Lệch phải | Tương tự, giá trị cao hiếm gặp |
| **Density** | Gần chuẩn | Dữ liệu khá đồng đều, ổn định |
| **pH** | Phân bố gần chuẩn | pH phổ biến trong khoảng 3.1–3.5 |
| **Sulphates** | Lệch phải | Đa số thấp, một vài giá trị cao gây ngoại lệ |
| **Alcohol** | Lệch trái nhẹ (Left-skewed) | Phần lớn rượu có nồng độ 9–11% |
| **Quality** | Gần chuẩn, hơi lệch trái | Chất lượng chủ yếu trong khoảng 5–6 điểm |

**3.2 Tương quan giữa các đặc trưng**

A chart of different colored squares

AI-generated content may be incorrect.

*Hình 10: Biểu đồ Heatmap mô tả tương quan*

**Các mối tương quan đáng chú ý**

* **Alcohol ↔ Quality (0.48)**: Đây là mối tương quan dương mạnh nhất với chất lượng. Có thể hiểu rằng rượu có nồng độ cồn cao thường được đánh giá chất lượng tốt hơn.
* **Density ↔ Alcohol (-0.69)**: Tương quan âm mạnh, cho thấy rượu có nồng độ cồn cao thường có mật độ thấp hơn.
* **Volatile Acidity ↔ Quality (-0.39)**: Tương quan âm khá rõ, nghĩa là độ axit bay hơi cao thường làm giảm chất lượng rượu.
* **Citric Acid ↔ Fixed Acidity (0.67)**: Hai loại axit này có mối liên hệ chặt chẽ, có thể cùng xuất hiện trong quá trình lên men.

**Các đặc trưng ít ảnh hưởng đến chất lượng**

* **Free Sulfur Dioxide ↔ Quality (0.01**): Gần như không có tương quan, cho thấy hai yếu tố này không ảnh hưởng rõ rệt đến đánh giá chất lượng rượu.
* **Alcohol ↔ Quality: Không có sự tương quan với nhau**

**3.2.1 Mối quan hệ giữa Density và Alcohol**

A chart with colored dots

AI-generated content may be incorrect.

*Hình 11: Biểu đồ scatter plot mô tả quan hệ giữa alcohol và quality*

* Xu hướng tăng: Khi chất lượng tăng từ 3 đến 8, nồng độ cồn cũng có xu hướng tăng từ khoảng 8 đến 15 độ. Điều này cho thấy có thể tồn tại mối liên hệ tích cực giữa độ cồn và chất lượng rượu.
* Phân bố điểm dữ liệu: Các điểm chất lượng thấp (màu đỏ, cam) thường tập trung ở vùng có độ cồn thấp, trong khi các điểm chất lượng cao (xanh dương đậm) lại xuất hiện nhiều ở vùng có độ cồn cao hơn.
* - Biến động trong từng nhóm chất lượng: Một số nhóm chất lượng (ví dụ chất lượng 6 hoặc 7) có độ cồn dao động khá rộng, cho thấy có thể có yếu tố khác ngoài alcohol ảnh hưởng đến chất lượng.

**3.2.2 Quan hệ Residual Sugar và Quality**

A graph with purple and white dots

AI-generated content may be incorrect.

*Hình 12: Biểu đồ scatter plot mô tả quan hệ giữa residual sugar và quality*

* **Phân bố đường dư:** Phần lớn mẫu rượu có lượng đường dư nằm trong khoảng từ 1 đến 5 g/L. Đây có thể là vùng phổ biến trong sản xuất rượu vang.
* **Chất lượng cao và đường thấp:** Các điểm có chất lượng 7 và 8 (màu xanh đậm) thường xuất hiện ở vùng có đường dư thấp, cho thấy có thể tồn tại mối liên hệ nghịch giữa đường dư và chất lượng.
* **Chất lượng trung bình chiếm đa số:** Các mẫu có chất lượng 4 đến 6 chiếm phần lớn, và có độ phân tán khá rộng về lượng đường dư.

**3.2.3 Quan hệ alcohol và pH**

A graph with a number of dots

AI-generated content may be incorrect.

*Hình 11: Biểu đồ scatter plot mô tả quan hệ giữa alcohol và pH*

Các điểm dữ liệu phân bố khá rải rác, không tạo thành một xu hướng tuyến tính hay cong rõ rệt. Cho thấy nồng độ cồn và độ pH không có mối liên hệ mạnh với nhau trong tập dữ liệu này.

**3.2.4 Phân bố nồng độ cồn theo chất lượng rượu**

A graph of different colored leaves

AI-generated content may be incorrect.

*Hình 12: Biểu đồ violin phân bố nồng độ cồn theo chất lượng rượu*

* **Chất lượng thấp (quality = 3, 4, 5):**
* Nồng độ cồn chủ yếu tập trung quanh 9–10 độ.
* Biểu đồ khá hẹp, cho thấy ít sự biến động về nồng độ cồn ở nhóm rượu chất lượng thấp.
* Các giá trị cực đại hiếm khi vượt quá 12 độ.
* **Chất lượng trung bình (quality = 6):**
* Phân bố bắt đầu mở rộng hơn, dao động từ khoảng 9 đến 13 độ.
* Trung vị nồng độ cồn cao hơn một chút so với các nhóm thấp (khoảng 10.5–11 độ).
* Đây là nhóm có phân bố khá cân bằng giữa nồng độ cồn thấp và cao.
* **Chất lượng cao (quality = 7, 8):**
* Nồng độ cồn dịch chuyển rõ rệt về phía cao (11–13.5 độ).
* Đặc biệt với quality = 8, phần lớn rượu có nồng độ cồn từ 11 đến 13.5 độ, trung vị gần 12 độ.
* Phân bố rộng hơn, cho thấy rượu chất lượng cao có sự đa dạng nhưng thiên về nồng độ cồn cao.
  1. **Phân bố và ngoại lệ các thuộc tính**

A graph with different colored squares and lines

AI-generated content may be incorrect.

*Hình 12: Mô hình boxplot phân bố và các ngoại lệ thuộc tính*

* **Total sulfur dioxide**: Có nhiều điểm nằm rất xa khỏi phần hộp, cho thấy một số mẫu rượu có lượng SO₂ cao bất thường.
* **Residual sugar**: Có một số mẫu có lượng đường dư rất cao, vượt xa phần lớn các mẫu khác.
* **Free sulfur dioxide** và **chlorides** cũng có một số ngoại lệ đáng chú ý.
  1. **Giá trị trung bình từng đặc trưng theo chất lượng rượu**

A chart with numbers and a number of different colored squares

AI-generated content may be incorrect.

*Hình 13: Mô hình heatmap mô tả mean từng đặc trưng theo chất lượng rượu*

* **Trục X (Chất lượng)**: thể hiện điểm chất lượng rượu (từ 3 → 8).
* **Trục Y (Đặc trưng)**: liệt kê các thuộc tính hóa học được đo trong mẫu rượu (như độ axit, nồng độ cồn, pH, v.v).
* **Màu sắc**: biểu diễn **mức độ lớn – nhỏ của giá trị trung bình**. Màu sáng hơn → giá trị cao hơn.
* Mỗi ô chứa **giá trị trung bình** của đặc trưng tương ứng với mức chất lượng rượu.

**Fixed acidity (Độ axit cố định):**

+ Giá trị dao động nhẹ quanh 8.0 – 8.8

+ Không có xu hướng tăng hoặc giảm rõ rệt theo chất lượng

➡️ **Không ảnh hưởng mạnh đến điểm chất lượng rượu.**

**Volatile acidity (Axit bay hơi)**

Giảm dần rõ rệt từ **0.88 (quality 3)** xuống **0.42 (quality 8)**.

➡️ Rượu **chất lượng cao có độ axit bay hơi thấp hơn**, nghĩa là ít chua, ít mùi gắt → dễ uống hơn.

**Citric acid (Axit citric)**

Tăng dần từ **0.17 → 0.39**

➡️ Rượu chất lượng cao **chứa nhiều axit citric hơn**, giúp cân bằng vị chua và làm hương vị tươi sáng hơn.

**Residual sugar (Lượng đường còn lại)**

Dao động nhẹ quanh 2.5–2.7, không có xu hướng rõ rệt.   
➡️ Hàm lượng đường **không ảnh hưởng đáng kể** đến chất lượng rượu.

**Chlorides (Lượng muối clorua)**

Giảm dần từ **0.12 → 0.07**.

➡️ Rượu **chất lượng cao có hàm lượng muối thấp hơn**, điều này tốt vì muối cao có thể làm vị rượu gắt hoặc đắng.

**Free sulfur dioxide (SO₂ tự do)**

Tăng nhẹ từ 11 → 16.9 ở mức 5, sau đó **giảm dần còn ~13** ở mức 8.

➡️ Có thể có ảnh hưởng **không tuyến tính**, nhưng nhìn chung rượu ngon có xu hướng **ít SO₂ hơn**, giúp hương vị tự nhiên hơn.

**Total sulfur dioxide (Tổng SO₂)**

Rất cao ở mức trung bình (56.5 tại quality = 5), rồi **giảm mạnh** ở mức cao hơn (33.4 tại quality = 8)

➡️ Rượu tốt hơn thường có ít SO₂, vì SO₂ cao làm giảm hương tự nhiên của rượu.

**Density (Mật độ)**

Gần như **không thay đổi (≈ 1.00)** → đặc trưng này **không phân biệt chất lượng rõ ràng**.

**pH**

Giảm dần từ 3.40 → 3.27.

➡️ Rượu chất lượng cao hơi có tính axit hơn (pH thấp hơn), giúp cân bằng vị rượu và tạo cảm giác tươi hơn.

**Sulphates (Muối sunfat)**

Tăng dần từ **0.57 → 0.77**.

➡️ Sulphates góp phần tạo cấu trúc và bảo quản rượu → giá trị cao hơn tương ứng rượu chất lượng tốt hơn.

**Alcohol (Nồng độ cồn)**

Tăng rõ rệt từ **9.96 → 12.09**.

➡️ Đây là một trong những yếu tố **quan trọng nhất ảnh hưởng đến chất lượng rượu** — rượu mạnh hơn thường có cấu trúc và hương vị đậm đà hơn

| **Xu hướng tăng theo chất lượng** | **Xu hướng giảm theo chất lượng** | **Không rõ xu hướng** |
| --- | --- | --- |
| Citric acid, Sulphates, Alcohol | Volatile acidity, Chlorides, Total SO₂, pH | Fixed acidity, Residual sugar, Density, Free SO₂ |

➡️ Tóm lại:

* **Rượu chất lượng cao** thường có:
  + **Nồng độ cồn cao hơn**
  + **Axit bay hơi thấp hơn**
  + **Axit citric và sulphates cao hơn**
  + **Ít muối và SO₂ hơn**
  + **pH thấp hơn (tính axit nhẹ hơn)**

Những yếu tố này góp phần làm rượu **cân bằng vị, ít chua gắt, đậm và thơm hơn.**

## Kết luận

* **Phân phối và ngoại lệ**
* Nhiều biến hóa học như **residual sugar**, **total sulfur dioxide**, **free sulfur dioxide**, **chlorides** có **giá trị ngoại lệ rõ rệt**, cho thấy sự đa dạng lớn giữa các mẫu rượu.
* Một số biến như **density**, **pH**, **fixed acidity** có phân bố tương đối ổn định, ít ngoại lệ.
* **Phân bố không chuẩn**
* Hầu hết các đặc trưng (ngoại trừ vài biến) **không tuân theo phân phối chuẩn**, thường lệch phải, có nhiều điểm cao. Điều này gợi ý cần thực hiện **biến đổi dữ liệu** (ví dụ log-transform) trước khi đưa vào mô hình.
* **Mối quan hệ giữa đặc trưng và chất lượng**
* Các biến nồng độ cồn (alcohol), citric acid, sulphates có xu hướng tăng khi chất lượng rượu tăng.
* Ngược lại, volatile acidity, chlorides, total sulfur dioxide có xu hướng giảm khi chất lượng rượu tăng.
* Một số biến như residual sugar, fixed acidity, density, free sulfur dioxide thể hiện xu hướng ít rõ ràng, có ảnh hưởng yếu hơn đến chất lượng.
* Đa cộng tuyến
* Một số biến có tương quan cao với nhau, điều này có thể gây ảnh hưởng tiêu cực khi xây dựng mô hình hồi quy hoặc phân loại (vấn đề multicollinearity). Cần cân nhắc loại bỏ hoặc gộp biến.
* **Ưu tiên biến mạnh ảnh hưởng**
* Alcohol và volatile acidity là hai biến nổi bật và dễ dàng nhận biết mối quan hệ rõ ràng với chất lượng rượu — có thể là những biến tiên đoán tốt trong mô hình phân loại.

Ý NGHĨA THỰC TẾ

* 1. **Đối với ngành sản xuất rượu**
* Kết quả giúp các nhà sản xuất tối ưu quy trình lên men để đạt được sản phẩm có chất lượng cao hơn.
* Có thể điều chỉnh hàm lượng cồn, độ chua, và SO₂ để cân bằng hương vị và đảm bảo tiêu chuẩn cảm quan.
  1. **Đối với nghiên cứu khoa học**
* Bộ dữ liệu cung cấp cơ sở để nghiên cứu mối liên hệ giữa hóa học và cảm quan, đóng góp cho lĩnh vực Data Science trong thực phẩm và đồ uống.
* Là bộ dữ liệu chuẩn (benchmark dataset) được sử dụng phổ biến trong dạy học và thử nghiệm mô hình học máy.
  1. **Đối với người tiêu dùng**
* Giúp hiểu rõ hơn về các yếu tố ảnh hưởng đến chất lượng rượu vang, hỗ trợ việc lựa chọn sản phẩm phù hợp hơn với sở thích và giá trị dinh dưỡng.
  1. **Đối với giáo dục và đào tạo**
* Là nguồn tài nguyên lý tưởng để thực hành phân tích dữ liệu, trực quan hóa, và xây dựng mô hình dự đoán cho sinh viên ngành Khoa học dữ liệu, Hóa học, và Công nghệ thực phẩm.