BÁO CÁO PHÂN TÍCH KHÁM PHÁ DỮ LIỆU (EDA) - LAB 5: RƯỢU VANG ĐỎ

I. Tổng quan Dữ liệu và Quy trình

1. Tập dữ liệu

Tập dữ liệu được phân tích là winequality-red.csv.

• **Kích thước:** Tập dữ liệu gốc có **1599 mẫu** và **12 cột**.

• **Kiểu dữ liệu:** Tất cả dữ liệu đều là kiểu số, bao gồm 11 cột kiểu float64 và 1 cột kiểu int64.

• **Giá trị thiếu:** Ban đầu, dữ liệu **không có giá trị nào bị thiếu** (Non-Null Count là 1599 cho tất cả các cột).

• **Các thuộc tính** bao gồm: Nồng độ acid cố định (fixed acidity), Nồng độ acid dễ bay hơi (volatile acidity), Lượng acid citric (citric acid), Lượng đường tồn dư (residual sugar), Muối Cl- trong rượu (chlorides), Lượng SO2 tự do (free sulfur dioxide), Tổng SO2 (total sulfur dioxide), Tỉ trọng (density), pH, sulphates, Nồng độ cồn (alcohol), và Chất lượng (quality).

2. Thống kê Mô tả Sơ bộ

Phân tích thống kê mô tả cho thấy:

• **Số lượng mẫu** (Count) đều đủ 1599.

• Giá trị trung bình (mean) và độ lệch chuẩn (std) của biến total sulfur dioxide **rất cao**.

• Các biến residual sugar, chloride, và total sulfur dioxide có giá trị tối đa (max) cao hơn nhiều lần so với phân vị 75% (Q3). Điều này cho thấy phân phối của các biến này bị lệch.

II. Làm sạch Dữ liệu (Xử lý Trùng lặp)

1. Kiểm tra Dữ liệu Trùng lặp

Việc kiểm tra đã xác định được **240 hàng bị trùng lặp** trong tập dữ liệu gốc.

2. Xử lý Trùng lặp

• Các hàng trùng lặp đã được loại bỏ bằng cách tạo ra một DataFrame mới (df_cleaned).

• **Kích thước dữ liệu gốc** là (1599, 12).

• **Kích thước dữ liệu sau khi làm sạch** là (1359, 12).

• Tổng cộng **240 hàng trùng lặp** đã được loại bỏ.

III. Phân tích Trực quan hóa (Univariate Analysis)

1. Phân phối Chất lượng Rượu ()

Phân tích biểu đồ đếm (countplot) của biến quality cho thấy:

• Phân phối chất lượng rượu không thay đổi nhiều sau khi loại bỏ hàng trùng lặp.

• Phần lớn các mẫu rượu vang đỏ có chất lượng tập trung ở mức **5 và 6**.

• Dữ liệu vẫn tồn tại tình trạng **mất cân bằng** (do số lượng mẫu ở mức 5 và 6 áp đảo các mức còn lại).

2. Phân phối của các Thuộc tính khác (Histograms)

Phân tích Histogram cung cấp các nhận xét về độ lệch của các biến:

• **Phân phối lệch phải (Right-skewed):**
  ◦ volatile acidity
  ◦ residual sugar
  ◦ chlorides
  ◦ free sulfur dioxide
  ◦ total sulfur dioxide
  ◦ sulphates
  ◦ alcohol
• **Phân phối gần giống phân phối chuẩn:**
  ◦ density
  ◦ pH

IV. Phân tích Đa biến (Multivariate Analysis)

1. Ma trận Tương quan (Correlation Matrix)

Phân tích ma trận tương quan giữa tất cả các thuộc tính đã đưa ra những nhận định sau:

• Biến **alcohol có mối tương quan dương lớn nhất** với biến mục tiêu quality.

• Biến **volatile acidity có tương quan âm mạnh nhất** với biến mục tiêu quality.

2. Mối quan hệ giữa từng Thuộc tính và Chất lượng (Boxplots)

Phân tích biểu đồ hộp (Boxplot) từng thuộc tính so với quality (Chất lượng) cho thấy các xu hướng rõ ràng:

• **Xu hướng rõ ràng:**

  ◦ **alcohol** có xu hướng **tăng theo chất lượng rượu**.

  ◦ **volatile acidity** càng cao thì **chất lượng càng giảm**.

• **Xu hướng yếu hơn:**

  ◦ sulphates có xu hướng tăng nhẹ.

  ◦ citric acid có xu hướng tăng nhẹ.

• **Các thuộc tính còn lại không có xu hướng rõ ràng** so với chất lượng.

V. Tổng Kết

Phân tích trên dữ liệu rượu vang đỏ đã được làm sạch (loại bỏ 240 hàng trùng lặp) xác nhận các điểm chính sau:

1. Dữ liệu có chất lượng ban đầu tốt (không thiếu giá trị) nhưng tồn tại vấn đề trùng lặp đã được khắc phục.

2. Các yếu tố ảnh hưởng mạnh mẽ nhất đến chất lượng rượu vang đỏ là:

  ◦ **alcohol** (Nồng độ cồn).

  ◦ **volatile acidity** (Nồng độ acid dễ bay hơi).

Hai chỉ số này là những yếu tố dự đoán quan trọng nhất cho chất lượng rượu vang đỏ.

BÁO CÁO PHÂN TÍCH KHÁM PHÁ DỮ LIỆU (EDA) - LAB 5: RƯỢU VANG ĐỎ

I. Tổng quan Dữ liệu và Quy trình

1. Tập dữ liệu

Tập dữ liệu được phân tích là winequality-red.csv.

• **Kích thước:** Tập dữ liệu gốc có **1599 mẫu** và **12 cột**.

• **Kiểu dữ liệu:** Tất cả dữ liệu đều là kiểu số, bao gồm 11 cột kiểu float64 và 1 cột kiểu int64.

• **Giá trị thiếu:** Ban đầu, dữ liệu **không có giá trị nào bị thiếu** (Non-Null Count là 1599 cho tất cả các cột).

• **Các thuộc tính** bao gồm: Nồng độ acid cố định (fixed acidity), Nồng độ acid dễ bay hơi (volatile acidity), Lượng acid citric (citric acid), Lượng đường tồn dư (residual sugar), Muối Cl- trong rượu (chlorides), Lượng SO2 tự do (free sulfur dioxide), Tổng SO2 (total sulfur dioxide), Tỉ trọng (density), pH, sulphates, Nồng độ cồn (alcohol), và Chất lượng (quality).

2. Thống kê Mô tả Sơ bộ

Phân tích thống kê mô tả cho thấy:

• **Số lượng mẫu** (Count) đều đủ 1599.

• Giá trị trung bình (mean) và độ lệch chuẩn (std) của biến total sulfur dioxide **rất cao**.

• Các biến residual sugar, chloride, và total sulfur dioxide có giá trị tối đa (max) cao hơn nhiều lần so với phân vị 75% (Q3). Điều này cho thấy phân phối của các biến này bị lệch.

II. Làm sạch Dữ liệu (Xử lý Trùng lặp)

1. Kiểm tra Dữ liệu Trùng lặp

Việc kiểm tra đã xác định được **240 hàng bị trùng lặp** trong tập dữ liệu gốc.

2. Xử lý Trùng lặp

• Các hàng trùng lặp đã được loại bỏ bằng cách tạo ra một DataFrame mới (df_cleaned).

• **Kích thước dữ liệu gốc** là (1599, 12).

• **Kích thước dữ liệu sau khi làm sạch** là (1359, 12).

• Tổng cộng **240 hàng trùng lặp** đã được loại bỏ.

III. Phân tích Trực quan hóa (Univariate Analysis)

1. Phân phối Chất lượng Rượu ()

Phân tích biểu đồ đếm (countplot) của biến quality cho thấy:

• Phân phối chất lượng rượu không thay đổi nhiều sau khi loại bỏ hàng trùng lặp.

• Phần lớn các mẫu rượu vang đỏ có chất lượng tập trung ở mức **5 và 6**.

• Dữ liệu vẫn tồn tại tình trạng **mất cân bằng** (do số lượng mẫu ở mức 5 và 6 áp đảo các mức còn lại).

2. Phân phối của các Thuộc tính khác (Histograms)

Phân tích Histogram cung cấp các nhận xét về độ lệch của các biến:

• **Phân phối lệch phải (Right-skewed):**

  ◦ volatile acidity

  ◦ residual sugar

  ◦ chlorides

- free sulfur dioxide
- total sulfur dioxide
- sulphates
- alcohol

• **Phân phối gần giống phân phối chuẩn:**

- density
- pH

IV. Phân tích Đa biến (Multivariate Analysis)

1. Ma trận Tương quan (Correlation Matrix)

Phân tích ma trận tương quan giữa tất cả các thuộc tính đã đưa ra những nhận định sau:

• Biến **alcohol có mối tương quan dương lớn nhất** với biến mục tiêu quality.

• Biến **volatile acidity có tương quan âm mạnh nhất** với biến mục tiêu quality.

2. Mối quan hệ giữa từng Thuộc tính và Chất lượng (Boxplots)

Phân tích biểu đồ hộp (Boxplot) từng thuộc tính so với quality (Chất lượng) cho thấy các xu hướng rõ ràng:

• **Xu hướng rõ ràng:**

- **alcohol** có xu hướng **tăng theo chất lượng rượu**.
- **volatile acidity** càng cao thì **chất lượng càng giảm**.

• **Xu hướng yếu hơn:**

- sulphates có xu hướng tăng nhẹ.
- citric acid có xu hướng tăng nhẹ.

• **Các thuộc tính còn lại không có xu hướng rõ ràng** so với chất lượng.

V. Tổng Kết

Phân tích trên dữ liệu rượu vang đỏ đã được làm sạch (loại bỏ 240 hàng trùng lặp) xác nhận các điểm chính sau:

1. Dữ liệu có chất lượng ban đầu tốt (không thiếu giá trị) nhưng tồn tại vấn đề trùng lặp đã được khắc phục.

2. Các yếu tố ảnh hưởng mạnh mẽ nhất đến chất lượng rượu vang đỏ là:

- **alcohol** (Nồng độ cồn).
- **volatile acidity** (Nồng độ acid dễ bay hơi).

Hai chỉ số này là những yếu tố dự đoán quan trọng nhất cho chất lượng rượu vang đỏ.