Proposal

Nhóm 5

Phần 1: Giới thiệu

Dự liệu được thu thập tại UCI Machine Learning Repository.

Hai bộ dữ liệu có liên quan đến các 2 loại rượu vang màu đỏ và rượu vang màu trắng được phân tích từ công ty “Vinho Verde” của Bồ Đào Nha. Bộ dự liệu liên quan đến các vấn đề biến số hóa lý và đánh giá chất lượng rượu.

Data set gồm 12 thuộc tính của tập dữ liệu:

1 – type : loại rượu

2 – fixed.acidity : độ chua cố định

3 – volatile.acidity : tính axit dễ bay hơi

4 – citric.acid : axit citric

5 – residual.sugar : đường dư

6 – chlorides : clorua

7 – free.sulfur.dioxide : lưu huỳnh đioxit

8 – total.sulfur.dioxide : tổng lưu huỳnh điôxit

9 – density : mật độ rượu

10 – pH : pH

11 – sulphates : sulphates

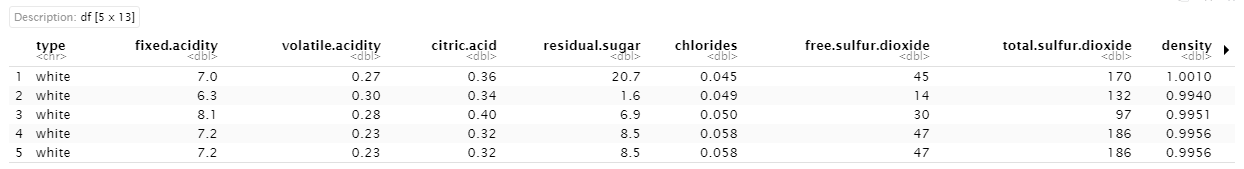
12 – alcohol : alcohol

13 - quality (score between 0 and 10) : đánh giá chất chượng rượu

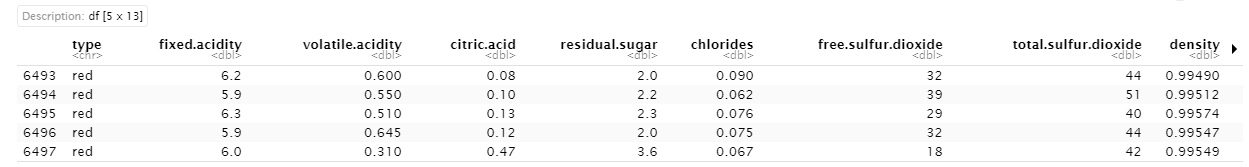
Phần 2: Dữ liệu

Dự liệu gồm có 13 cột và 6497 dòng.

* 5 dòng đầu tiên của dữ liệu:



* 5 dòng cuối cùng của dữ liêu



Phần 3: Kế hoạch phân tích dữ liệu

* Phân tích EDR mô tả một số thuộc tính của thuộc tính của tập dữ liệu
* Sử dụng mô hình Linear Regression, KNN, Decsion Tree để dự đoán các yếu tố ảnh hưởng đến yếu tố đánh giá chất lượng rượu.
* Đánh giá chất lượng mô hình dự đoán trên tập training và tập test.