**BỘ GIÁO DỤC VÀ ĐÀO TẠO**

**TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT TP. HCM**

**KHOA CÔNG NGHỆ THÔNG TIN**

----🙦🙦🕮🙤🙤----

****

**Báo Cáo Cuối Kỳ**

**Môn: LẬP TRÌNH R CHO PHÂN TÍCH**

**Đề tài**

**PHÂN TÍCH VÀ ĐÁNH GIÁ CHẤT LƯỢNG RƯỢU**

**GVHD:** Ths.Quách Đình Hoàng

**SVTH:**

Nguyễn Thị Mỹ Linh 19133032

Lê Thị Thanh Phương 19133046

Nguyễn Duy Phước 19133003

Võ Thị Ngọc Thắm 19133051

**Tên học phần:** 211RPAN233577

**Học kỳ: I Năm học: 2021 2022**

**Tp. Hồ Chí Minh, tháng 12 năm 2021**

# 1. Tóm tắt (abstract)

Các yếu tố lí hóa ảnh hưởng tới chất lượng rượu, các chất lí hóa khác nhau thì chất lượng đánh giá khác nhau. Do đó điều quan trọng là phải biết điều chỉnh hàm lượng của từng chất để tạo nên rượu được đánh giá là tốt. Dự án này sử dụng một bộ dữ liệu nguồn mở, bao gồm 13 cột và 6497 dòng trong tập dữ liệu của kaggle. Bằng cách sử dụng các tiền xử lí, chọn ra những biến có ảnh hưởng lớn tới chất lượng rượu. Bên cạnh đó, trực quan hóa dữ liệu cũng giúp cho dữ liệu được nhìn thấy dễ dàng và dễ hiểu hơn. Có nhiều thuật toán khác nhau để dự đoán chất lượng rượu, trong bài báo cáo này nhóm lựa chọn sử dụng 3 thuật toán: k-Nearest Neighbors, Linear Regression, Decsion. Kết quả của dự án này có độ chính xác là ….

## Các câu hỏi đặt ra cho bài toán

Các yếu tố ảnh hưởng đến chất lượng rượu

Dự đoán chất lượng rượu

## Kết quả

Model K-nn là model tốt nhất đề đánh giá được chất lượng rượu trong 3 model mà nhóm tìm hiểu

# 2. Giới thiệu (introduction)

Đối với việc dự đoán chất lượng rượu là vô cùng quan trọng trong sản xuất kinh doanh để giảm chi phí các lần thử và nâng cao hiệu quả kinh doanh. Machine learning giúp cho việc dự đoán này trở nên dễ dàng hơn. Mục đích chính của đề tài này là dự đoán một cách chính xác chất lượng rượu có thể

## 2.1. Input

Tập dữ liệu winequalityN.csv

## 2.2. Output

Chất lượng rượu được đánh giá

# 3. Dữ liệu (data)

## 3.1. Giới thiệu tập dữ liệu

Tập dữ liệu lấy trên trang kaggle

Link dataset: <https://www.kaggle.com/rajyellow46/wine-quality>

Bộ dữ liệu có liên quan đến các biến thể màu đỏ và trắng của rượu vang "Vinho Verde" của Bồ Đào Nha với các thuộc tính lí hóa của rượu và mức đánh giá chất lượng.

Tập dữ liệu có 13 cột và 6497 dòng

Giải thích các thuộc tính:

"type": loại rượu (rượu trắng hoặc đỏ)

"fixed.acidity": độ axit cố định

"volatile.acidity": độ axit dễ bay hơi

"citric.acid": axit xitric

"residual.sugar": đường dư

"chlorides": clorua

"free.sulfur.dioxide": lưu huỳnh đioxit tự do

"total.sulfur.dioxide": tổng lưu huỳnh đioxit

"density": tỉ trọng

"pH" : độ pH

"sulphates": sunfat

"alcohol": độ công

"quality": chất lượng (điểm từ 0 đến 10)

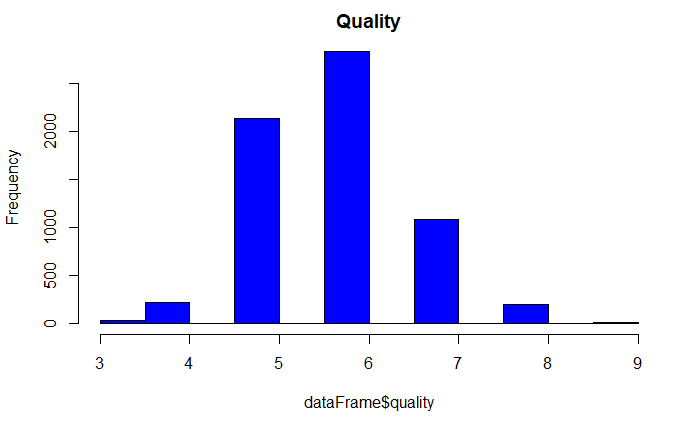
## 3.2. Tiền xử lí

Kiểm tra dữ liệu và thấy rằng có 38 giá trị NA, xử lí bằng cách loại bỏ

## 3.3. Tập dữ liệu

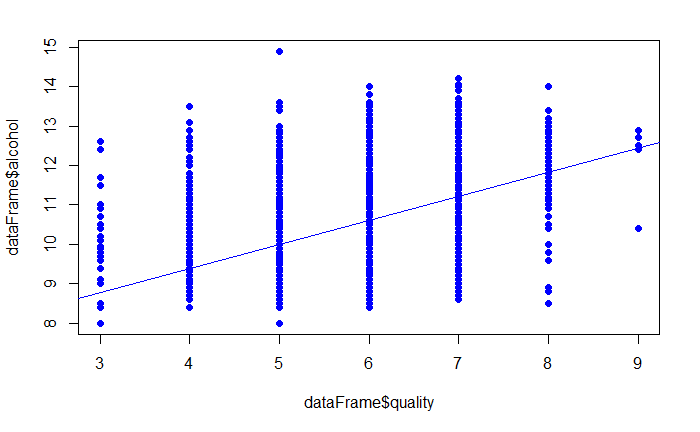
Chia tập dữ liệu thành hai phần: train (80%) và test(20%)

# 4. Trực quan hóa dữ liệu (data visulization)

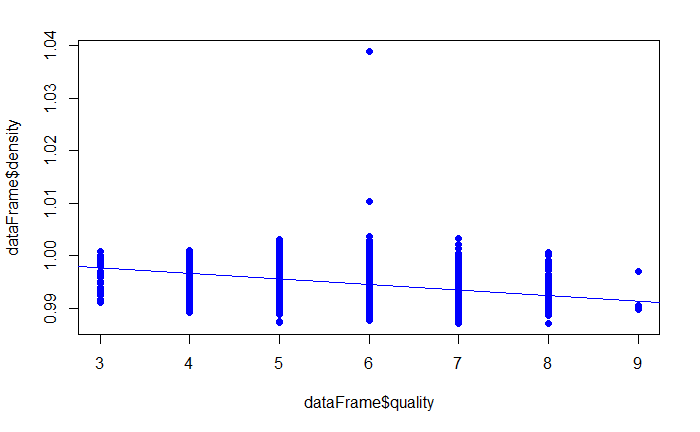


Chất lượng rượu tuân theo sự phân bố hình chuông khá bình thường. Hầu hết các loại rượu được đánh giá chất lượng ở mức trung bình.

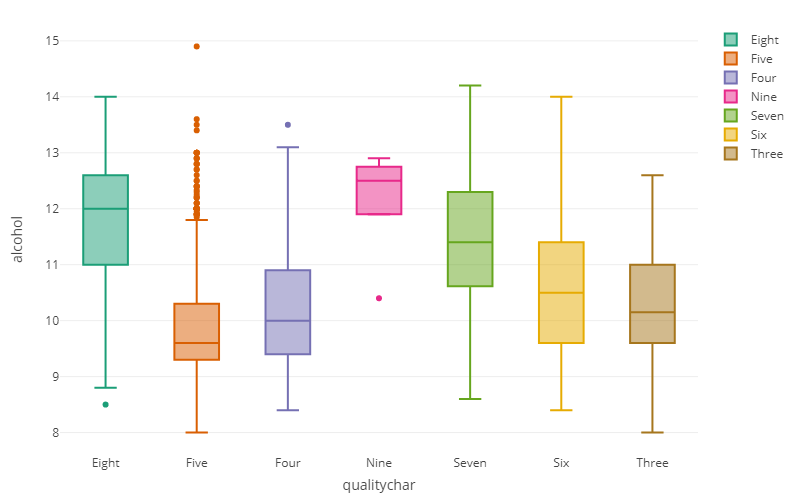
Ở đây sử dụng hai biến alcohol và densiy ( 2 biến này có ảnh hưởng nhiều tới biến quality) để phân tích mối quan với biến quality



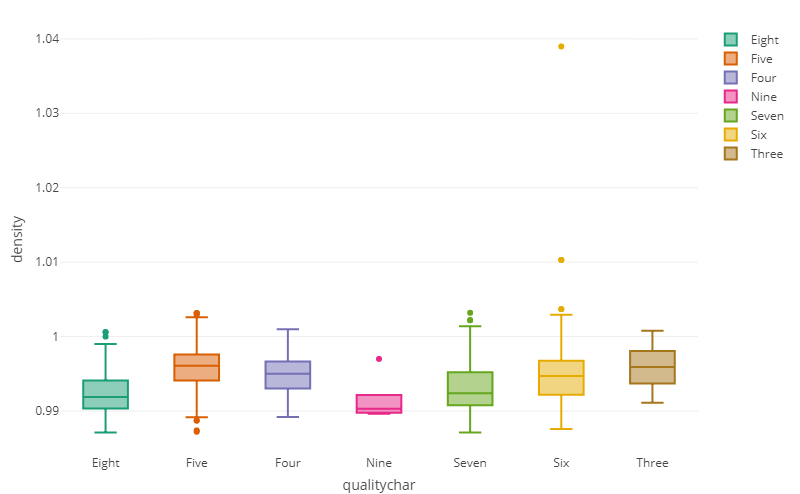
Biến quality và alcohol có mối quan hệ tỉ lệ thuận với nhau, nghĩa là khi nồng độ alcohol tăng thì quality tăng



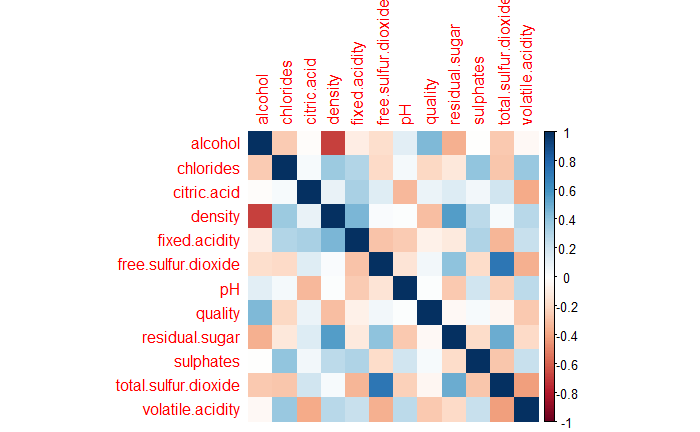
Biến quality và density có mối quan hệ tỉ lệ nghịch với nhau.



Nhìn vào hình này có thể thấy các giá trị min, max, median,Q1, Q3 của alcohol ở các mức chất lượng khác nhau. Có thể thấy các loại rượu có chất lượng trên mức trung bình thì có nồng độ alcohol cao hơn so với nồng độ alcohol của rượu có mức đánh giá dưới mức trung bình.



Ở biến density thì ngược lại với biến alcohol,các loại rượu có chất lượng trên mức trung bình thì có mức density thấp hơn so với mức density của rượu có mức đánh giá chất lượng dưới trung bình



Sử dụng biểu đồ carplot để đánh giá mối tương quan giữa các biến với nhau. Cường độ màu tỉ lệ với hệ số tương quan. Từ thang 0 đến 1 thể hiện mối quan hệ tích cực, ví dụ: giữa biến quality với alcohol thì mối tương quan của nó là 0.4443185. Từ thang 0 đến -1 thể hiện mối quan hệ tiêu cực, ví dụ: giữa biến quality với density thì mối tương quan của nó là -0.3058579

# 5. Mô hình hóa dữ liệu (data modeling)

## 5.1. K-Nearest Neighbors Regression

K-Nearest Neighbors (KNN) là một thuật toán đơn giản nhất trong nhóm thuật toán Học có giám sát. Ý tưởng của thuật toán này đó là tìm output của một dữ liệu mới dựa trên output của K điểm gần nhất xung quanh nó.

## 5.2. Linear Regression

Linear Regression là phương pháp tiếp cận dự đoán một biến phản hồi định lượng Y dựa trên cơ sở một hay nhiều biến dự đoán X.

Mô hình hồi quy tuyến tính có dạng:

= 𝑓(𝑥) = 𝜃0 + 𝜃1𝑥1 + 𝜃2𝑥2 + ⋯ + 𝜃𝑑𝑥𝑑 = 𝜃𝑇𝑥

Với 𝜃𝑇 = (𝜃0, 𝜃1, ⋯, 𝜃𝑑) ∈ 𝑅𝑑+1 được gọi là tham số của mô hình

- Các độ đo để đánh giá model:

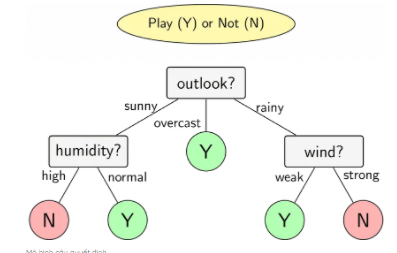
𝐿 = 𝑀𝐴𝐸 = ∑ |𝑓(𝑥𝑖 ) − 𝑦𝑖 |

R2 =

RSS = (yi − )2

TSS = (yi − )2 : total sum of squares

## 5.3. Decsion Tree Regression

Hồi quy cây quyết định quan sát các đặc điểm của một đối tượng và đào tạo một mô hình trong cấu trúc của cây để dự đoán dữ liệu trong tương lai nhằm tạo ra đầu ra liên tục có ý nghĩa. Đầu ra liên tục có nghĩa là đầu ra / kết quả không rời rạc, tức là nó không được biểu diễn chỉ bằng một tập hợp số hoặc giá trị rời rạc, đã biết.

# 6. Thực nghiệm, kết quả, và thảo luận (experiments, results, and discussion)

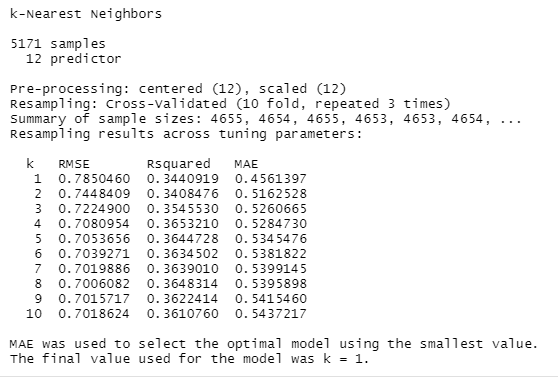
**Tách tập dữ liệu:** Sử dụng phương pháp createDataPartition và mục tiêu quality để chia dữ liệu thành hai tập train (80%) và test (20%)

**outcome:** quality

**intcome:** các biến còn lại trên tập dataset

## 6.1. K-Nearest Neighbors Regression

**KNN model**



Model trên tập train gồm 5171 mẫu 12 dự đoán

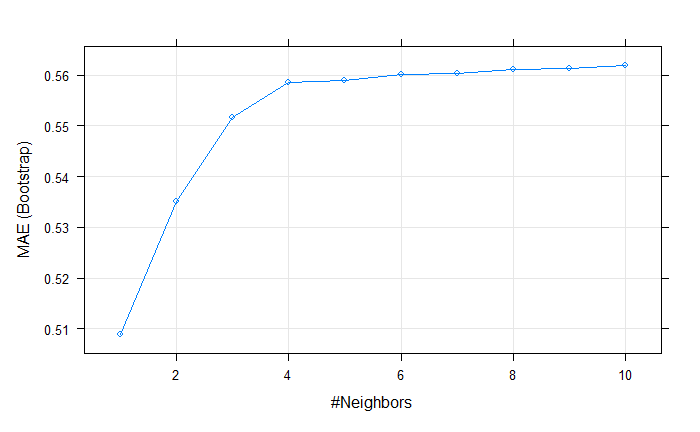
Mô hình sử dụng cross-validation chia làm tập train làm 10 fold

MAE được sử dụng để chọn mô hình tối ưu sử dụng giá trị nhỏ nhất

Giá trị cuối cùng được sử dụng cho mô hình là k = 1

Giá trị MAE nhỏ nhất là 0.4561397

**Kết quả trực quan**



**Đánh giá mô hình trên tập Test**

Tóm lược thực tế



Tóm lược dự đoán



Nhận xét: Dựa theo tóm lược về thực tế và dự đoán trên tập test ta thấy tuy các giá trị không có sự thay đổi lớn nhưng giá trị Max có sự thay đổi khá lớn. Do đó mô hình này không thật sự tốt trong đánh giá chất lượng rượu ngon hay dở.

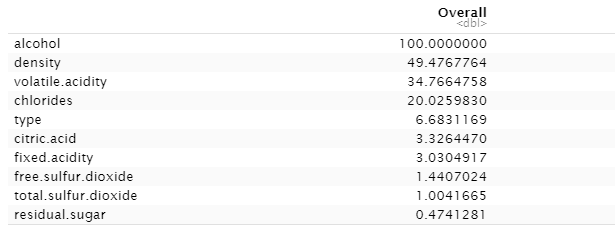
**Nhận xét các giá trị Rsquare và MAE**

Rsquare = 0.3681061

MAE = 0.4427245

Nhận xét: Kết quả R-squared (ghi lại tỷ lệ phần trăm của sự thay đổi trong biến Y (quality) được giải thích bởi các yếu tố dự đoán) có giá trị không lớn. Giá trị MAE (Trung bình của sai biệt tuyệt đối) không quá lớn và có thể chấp nhận được với MAE nằm trong thang điểm từ 0 đến 10

**Các biến quan trọng**



**Nhận xét**

Độ cồn là biến số quan trọng nhất có giá trị tổng thể là 100.00 và theo sau là tỷ trọng có giá trị tổng thể là 49.48, độ bay hơi. độ chua 34.77, clorua 20.02

## 6.2. Linear Regression

Đánh giá trên tập train:

Ảnh có chứa văn bản

Mô tả được tạo tự động

Mô hình linear regression trên tập dữ liệu training với biến phản hồi quality với các thuộc tính trong tập dữ liệu cho ra kết quả:

R-squared = 0.29

MAE = 0.57

Đánh giá mô hình:

Kết quả R-squared (ghi lại ly lệ phần trăm của sự thay đổi trong biến Y (qualty) được giải thích bởi các yếu tố dự đoán) có giá trị không lớn.

Ảnh có chứa văn bản, biên lai, ảnh chụp màn hình

Mô tả được tạo tự động

Mô hình linear rgression trên tập dữ liệu

R-squared = 0.25

MAE = 0.59

Đánh giá trên tập test:

|  |  |  |
| --- | --- | --- |
| # | R-squared | MAE |
| All | 0.3 | 0.57 |
| alcohol | 0.2 | 0.63 |
| alcohol + volatile.acidity | 0.27 | 0.59 |
| alcohol + volatile.acidity + density | 0.27 | 0.59 |
| alcohol + volatile.acidity + density + | 0.27 | 0.59 |

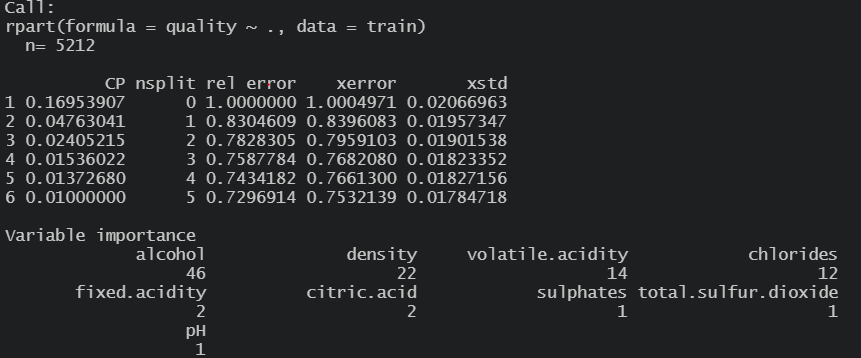
- Mô hình trên tập test không bị lệch so với tập training.

- Theo số liệu, trên bảng mối quan hệ tuyến tính của biến Y (quality – chất lượng rượu) với các yếu tố dự đoán trong tập dữ liệu được biết, mối quan hệ không có mối tương quan mạnh mẽ. Mô hình hồi quy tuyến tính không được đánh giá cao trong trường hợp này.

- Với mô hình này, các yếu ảnh hưởng đến giá trị biến chất lượng rượu: gồm biến alcohol, volatile.acidity là 2 biến thể hiện mối tượng quan hơn so với các biến khác.

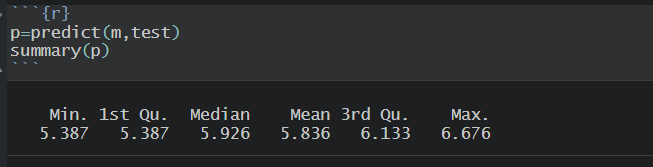
## 6.3. Decsion Tree Regression

**Model của Decision Tree**

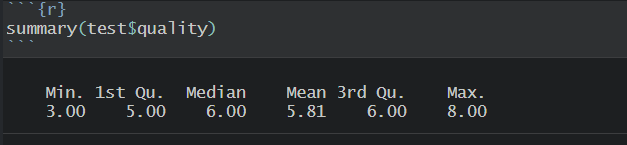


**Đánh giá mô hình trên tập Test**

Kết quả dự đoán

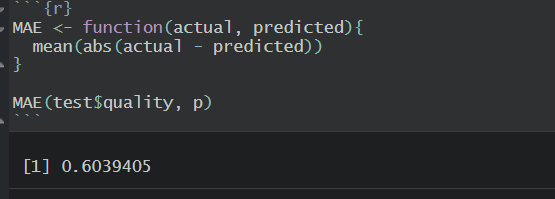


Tổng thể trên tập thực tế



Nhận xét: So sánh trên kết quả dự đoán so với tập quality thực tế ta thấy mô hình thực sự không tốt trong việc ước tính rượu dở và rượu thực sự ngon.

Đồng thời tính ra giá trị MAE (giá trị trung sai sô tuyệt đối ) = 0.6 với giá trị này chỉ tạm chấp nhận được vì khoảng giá trị của quality 0-10



# 7. Kết luận (conclision)

Kết quả mà nhóm làm được :

- Model K-nn là model tốt nhất đề đánh giá được chất lượng rượu trong 3 model mà nhóm tìm hiểu

- Giá trị MAE trong 3 model thì model K-nn có giá trị nhở nhất MAE=0.44

- Lí do chọn MAE để so sánh kết quả giữa 3 model vì MAE được biết đến là mạnh mẽ hơn đối với các yếu tố ngoại lai (outliers) so với MSE. Lý do chính bởi vì MSE sử dụng bình phương lỗi, các ngoại lai (những samples mà có lỗi cao hơn hẳn các samples khác) sẽ được chú ý và chiếm ưu thế hơn (do tính bình phương) trong việc đánh giá và điều này tác động đến các thông số của mô hình.

Nếu có thời gian thì nhóm sẽ :

- Tìm hiểu thêm nhiều model để dự đoán

- Tối hưu các model đồ đã làm

# 8. Phụ lục (appendices)

# 9. Đóng góp (contributions)

|  |  |
| --- | --- |
| Thành viên | Công việc |
| Ngọc Thắm | EDA |
| Mỹ Linh | k-Nearest Neighbors |
| Thanh Phương | Linear Regression |
| Duy Phước | Decsion Tree |

# 10. Tham khảo (references)

[https://rstudio-pubs-static.s3.amazonaws.com/227997\_869ca5f2dc144f7b85cdbc3f45a47bb6.html#step-3-model-training-regression-tree](https://rstudio-pubs-static.s3.amazonaws.com/227997_869ca5f2dc144f7b85cdbc3f45a47bb6.html" \l "step-3-model-training-regression-tree)