Link dữ liệu: <https://archive.ics.uci.edu/ml/datasets/wine+quality>

# Mô tả dữ liệu

Dữ liệu này mô hình hóa sở thích rượu vang bằng cách khai thác dữ liệu từ các đặc tính hóa lý.

1. Tên tiêu đề: Wine Quality
2. Nguồn: P. Cortez, A. Cerdeira, F. Almeida, T. Matos và J. Reis. Mô hình hóa sở thích rượu vang bằng cách khai thác dữ liệu từ các đặc tính hóa lý. Trong Hệ thống Hỗ trợ Quyết định, Elsevier, 47 (4): 547-553, 2009
3. Số Instances: red wine - 1599; white wine - 4898
4. Số thuộc tính: 11
5. Thông tin thuộc tính:

*Input variables*

* fixed acidity
* volatile acidity
* citric acid
* residual sugar
* chlorides
* free sulfur dioxide
* total sulfur dioxide
* density
* pH
* sulphates
* alcohol

*Output variable*

* + quality (score between 0 and 10)

6. Missing Attribute Values: Không

7. Đặc điểm tập dữ liệu: đa biến

8. Bài toán: Phân loại, hồi quy

# Khám phá dữ liệu

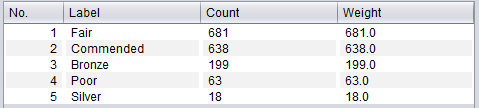
Trên tập dữ liệu winequality-red

| Số mẫu | 1599 |
| --- | --- |
| Thuộc tính | 12 |
| Kiểu dữ liệu | số thực |
| Số lớp | 8 |
| Missing value | 0 |
| Bài toán | Phân loại, hồi quy |

|  | min | max | min | std |
| --- | --- | --- | --- | --- |
| fixed acidity | 4,6 | 15,9 | 8,32 | 1,741 |
| volatile acidity | 0,12 | 1,58 | 0,528 | 0 |
| citric acid | 0 | 1 | 0,271 | 0 |
| residual sugar | 0,9 | 15,5 | 2,539 | 1,41 |
| chlorides | 0,012 | 611 | 0,087 | 0 |
| free sulfur dioxide | 1 | 72 | 15,875 | 10,46 |
| total sulfur dioxide | 6 | 289 | 46,468 | 32,895 |
| density | 0,99 | 1 | 0,997 | 0 |
| pH | 2,74 | 4,01 | 3,311 | 0,154 |
| sulphates | 0,33 | 2 | 0,658 | 0,17 |
| alcohol | 8,4 | 14,9 | 10,423 | 1,066 |
| quality | 3 | 8 | 5,636 | 0,808 |

# Phân lớp trên tập Winequality-red

## Chuyển label quality từ số sang text

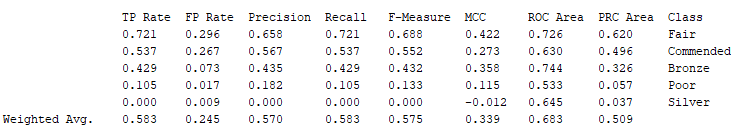


## Phân lớp dữ liệu với DT(J48)

| Thời gian build model | 0.27 seconds | |
| --- | --- | --- |
| Tổng số bản ghi | 1599 | |
| Tổng số lớp | 544 | |
| Số mẫu phân lớp đúng | 317 | |
| Số mẫu phân lớp sai | 227 | |
| Tỷ lệ phân lớp đúng | 58.2721 % | |
| Tỷ lệ phân lớp sai | 41.7279 % | |
| Độ sâu cây | 457 | |
| Tỷ lệ huấn luyện/kiểm thử là 66/34 | | |
| Ma trận nhầm lẫn |  | |
|
|
|
|
|

| Số mẫu phân lớp đúng vào các lớp | | Số mẫu phân sai vào các lớp | |
| --- | --- | --- | --- |
| 173 mẫu vào đúng lớp a Fair | | 67 mẫu sai vào lớp a | |
| 115 mẫu vào đúng lớp b Commended | | 99 mẫu sai vào lớp b | |
| 27 mẫu vào đúng lớp c Bronze | | 36 mẫu sai vào lớp c | |
| 2 mẫu vào đúng lớp d Poor | | 17 mẫu sai vào lớp d | |
|  | | 8 mẫu sai vào lớp e | |
|  | |  | |

Độ chính xác của class

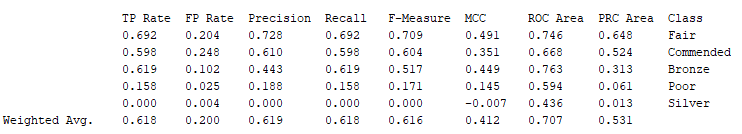


## Phân lớp dữ liệu với K-NN(IBK với K=1)

| Thời gian build model | 0 seconds | |
| --- | --- | --- |
| Tổng số bản ghi | 1599 | |
| Tổng số lớp | 544 | |
| Số mẫu phân lớp đúng | 336 | |
| Số mẫu phân lớp sai | 208 | |
| Tỷ lệ phân lớp đúng | 61.7647 % | |
| Tỷ lệ phân lớp sai | 38.2353 % | |
| Tỷ lệ huấn luyện/kiểm thử là 66/34 | | |
| Ma trận nhầm lẫn |  | |
|
|
|
|
|

| Số mẫu phân lớp đúng vào các lớp | | Số mẫu phân sai vào các lớp | |
| --- | --- | --- | --- |
| 166 mẫu vào đúng lớp a Fair | | 74 mẫu sai vào lớp a | |
| 128 mẫu vào đúng lớp b Commended | | 86 mẫu sai vào lớp b | |
| 39 mẫu vào đúng lớp c Bronze | | 24 mẫu sai vào lớp c | |
| 3 mẫu vào đúng lớp d Poor | | 16 mẫu sai vào lớp d | |
|  | | 8 mẫu sai vào lớp e | |
|  | |  | |

Độ chính xác của class



## Phân lớp dữ liệu với SVM(SMO)

**Sử dụng 10 kernel**

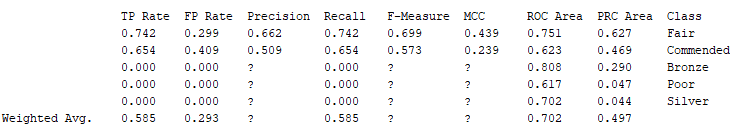
**Sử dụng hàm nhân tuyến tính**

**Khi sử dụng Cross-Validation ta lấy C = 1**

| Test mode | split 66.0% train, remainder test | | 30-fold cross-validation |
| --- | --- | --- | --- |
| Time taken to build model | 0.26 seconds | | 0.13 seconds |
| Classifier for classes | Fair, Commended | | |
| Number of kernel evaluations | 60910 (69.046% cached) | | |
| Classifier for classes | Fair, Bronze | | |
| Number of kernel evaluations | 13733 (68.947% cached) | | |
| Classifier for classes | Fair, Poor | | |
| Number of kernel evaluations | 71732 (65.436% cached) | | |
| Classifier for classes | Fair, Silver | | |
| Number of kernel evaluations | 6827 (69.366% cached) | | |
| Classifier for classes | Commended, Bronze | | |
| Number of kernel evaluations | 31695 (75.579% cached) | | |
| Classifier for classes | Commended, Poor | | |
| Number of kernel evaluations | 34982 (76.379% cached) | | |
| Classifier for classes | Commended, Silver | | |
| Number of kernel evaluations | 14986 (80.873% cached) | | |
| Classifier for classes | Bronze, Poor | | |
| Number of kernel evaluations | 3796 (68.594% cached) | | |
| Classifier for classes | Bronze, Silver | | |
| Number of kernel evaluations | 6300 (77.694% cached) | | |
| Classifier for classes | Poor, Silver | | |
| Number of kernel evaluations | 533 (72.667% cached) | | |
| Correctly Classified Instances | 58.4559 % | | 58.349 % |
| Incorrectly Classified Instances | 41.5441 % | | 41.651 % |
| Confusion Matrix |  | |  |
|
|
|
|
|

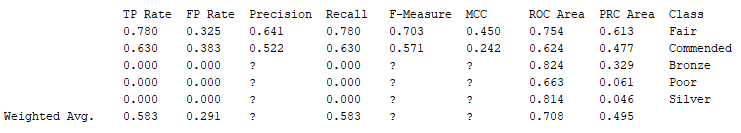
| Train/Test: 66/34 | | | |
| --- | --- | --- | --- |
| Số mẫu phân lớp đúng vào các lớp | | Số mẫu phân sai vào các lớp | |
| 178 mẫu vào đúng lớp a Fair | | 62 mẫu sai vào lớp a | |
| 140 mẫu vào đúng lớp b Commended | | 74 mẫu sai vào lớp b | |
|  | | 63 mẫu sai vào lớp c | |
|  | | 19 mẫu sai vào lớp d | |
|  | | 8 mẫu sai vào lớp e | |
|  | |  | |

Độ chính xác của class



| 30-fold cross-validation | | | |
| --- | --- | --- | --- |
| Số mẫu phân lớp đúng vào các lớp | | Số mẫu phân sai vào các lớp | |
| 531 mẫu vào đúng lớp a Fair | | 150 mẫu sai vào lớp a | |
| 402 mẫu vào đúng lớp b Commended | | 236 mẫu sai vào lớp b | |
|  | | 199 mẫu sai vào lớp c | |
|  | | 63 mẫu sai vào lớp d | |
|  | | 18 mẫu sai vào lớp e | |
|  | |  | |

Độ chính xác của class



## Nhận xét và đánh giá

Kết quả trên ta thấy phân lớp dựa trên KNN cho kết quả chính xác nhiều nhất