# Wine Quality Data Classification and Clustering

Victoria Lester
Chicago, USA
VictoriaDLester@LewisU.edu

Johannah Cushing
Saco, USA
JohannaKCushing@LewisU.edu

***Abstract*— This paper serves as a brief analysis of the use of physiochemical attributes to identify similarities between wine samples and the ability of the attributes to predict quality. The Wine Quality Data Set [1] was analyzed using kNN and clustering algorithms using Orange and Weka software. Some attributes were eliminates as not being significantly predictive of quality for each part of the set (red and white wine), and different attributes were found to be predictive for red compared to white wine. kNN algorithms were found to be very predictive of quality for both the red and white wine. Clustering algorithms were explored as a means to categorize wines, but a more detailed analysis would be need to determine a "best" cluster result.**

*Keywords—wine, wine quality, clustering, classification, Orange, kNN, Kmeans, Canopy, EM Clustering, WEKA*

## Introduction

The purpose of this paper is to investigate the use of physiochemical data in assessing wine quality. The Wine Quality Data Set [1], made available through UC Irvine Machine Learning Repository[2], contains both objective physiochemical attributes and sensory attributes for white and red wines. The sensory attribute, "quality", is the mean of three quality ratings provided by wine experts and serves as the class when performing regression and classification tasks. For classification algorithms to accurately identify the subjective "quality" class based off the objective physiochemical attributes, one can infer the wines that fall within the same "quality" class have similar physiochemical attributes. As such, the analysis of this data was approached with two questions in mind:

1. How accurately can classification algorithms identify the quality of wine samples when a set number of "quality" rankings have already been identified?
2. Can a clustering algorithm, which groups data points by their similarities to one another rather than their similarities to a known sample, group wines in a similar manner without the expert determined "quality" class?

To address these questions the Wine Quality Data Set was analyzed in Orange and Weka using a combination of classification and clustering algorithms.

## I. METHODOLOGY

### A. *Description and Features of the Wine Quality Data Set*

The Wine Quality Data Set is composed of two data sets, one for red wine and one for white wine, from "vino verde" wine samples from Portugal. The red wine data set contains 1599 instances and the white wine data set contains 4898 instances. In additional to the "quality" sensory attribute, the data sets utilize the same objective attributes:

1. Fixed Acidity
2. Volatile Acidity
3. Citric Acid
4. Residual Sugar
5. Chlorides
6. Free Sulfur Dioxide
7. Total Sulfur Dioxide
8. Density
9. Ph
10. Sulphates
11. Alcohol

These objective attributes were obtained through physiochemical chemical testing, as opposed to the subjective "quality" attribute which was determined by calculating the mean of three ratings provided by wine experts. All attributes, including "quality", are continuous numerical values. It should be noted that while the quality rankings fall on a scale of 1-10, the samples fall primarily within the middle of the range rather than at the two extremes (excellent or poor). UC Irvine Machine Learning Repository's recommends utilizing the Wine Quality Data Set for regression or classification tasks.

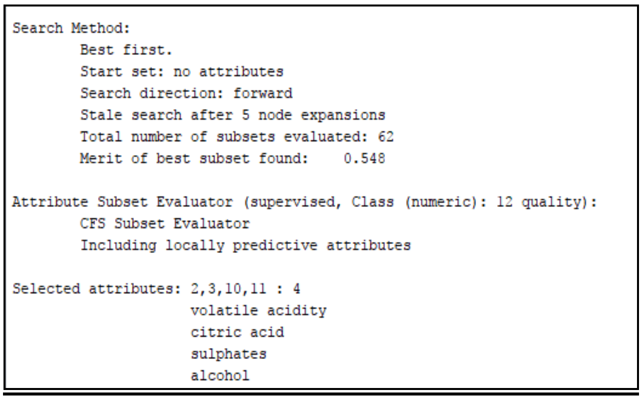
### B. *Data Preperation*

As the Wine Quality Data Set contained no missing or invalid values, it was unnecessary to clean the data prior to importing it into Weka and Orange. On the UC Irvine Machine Learning Repository page for the Wine Quality Data Set it is noted that all 11 attributes may not be necessary for the classification of the wine. As such, analysis was performed individually on each sub-set of data to identify the relative importance for each attribute of the data sets. The analysis was performed in both Orange and Weka, which produced different results due to variations in the software's ranking algorithms.

#### 1) *Attribute Ranking*

##### a) *Ranking in Weka*

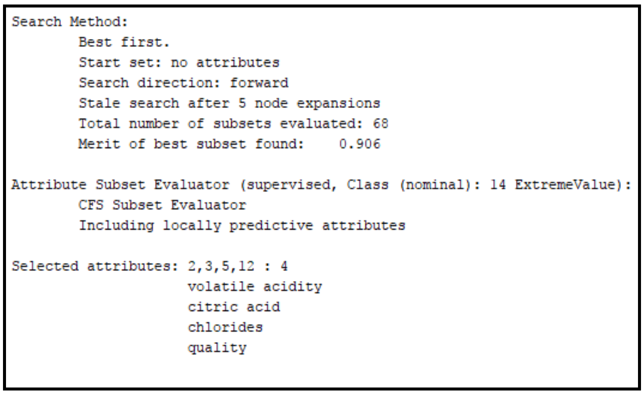
The red wine data was analyzed first, and the following attributes were retained for further analysis: volatile acidity, citric acid, sulphates, and alcohol.



```
Search Method:
        Best first.
        Start set: no attributes
        Search direction: forward
        Stale search after 5 node expansions
        Total number of subsets evaluated: 62
        Merit of best subset found:    0.548

Attribute Subset Evaluator (supervised, Class (numeric): 12 quality):
        CFS Subset Evaluator
        Including locally predictive attributes

Selected attributes: 2,3,10,11 : 4
                     volatile acidity
                     citric acid
                     sulphates
                     alcohol
```

*Fig. 1. CFSSubsetEval of red wine data*

When CFSSubsetEval was performed on the white wine, the following attributes were retained for further analysis: volatile acidity, citric acid, chlorides, and alcohol.

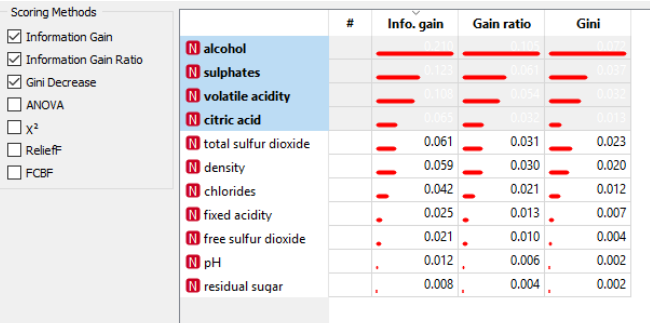


```
Search Method:
        Best first.
        Start set: no attributes
        Search direction: forward
        Stale search after 5 node expansions
        Total number of subsets evaluated: 68
        Merit of best subset found:    0.906

Attribute Subset Evaluator (supervised, Class (nominal): 14 ExtremeValue):
        CFS Subset Evaluator
        Including locally predictive attributes

Selected attributes: 2,3,5,12 : 4
                     volatile acidity
                     citric acid
                     chlorides
                     quality
```

*Fig. 2. CFSSubsetEval of white wine data*

### *b) Ranking in Orange*

The red wine data was analyzed using the “Rank” function in Orange. The Information Gain, Information Gain Ratio, and Gini Decrease scoring methods were used. The following attributes were retained for further analysis: volatile acidity, citric acid, sulphates, and alcohol. These attributes match those identified in Weka.

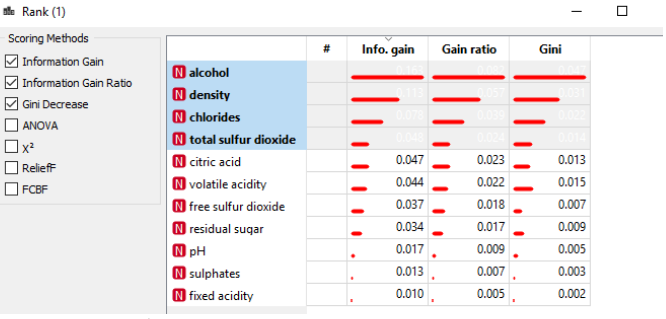


Scoring Methods
☑ Information Gain
☑ Information Gain Ratio
☑ Gini Decrease
☐ ANOVA
☐ $\chi^2$
☐ ReliefF
☐ FCBF

| | # | Info. gain | Gain ratio | Gini |
|---|---|---|---|---|
| N alcohol | | 0.210 | 0.105 | 0.072 |
| N sulphates | | 0.123 | 0.061 | 0.037 |
| N volatile acidity | | 0.108 | 0.054 | 0.032 |
| N citric acid | | 0.065 | 0.032 | 0.013 |
| N total sulfur dioxide | | 0.061 | 0.031 | 0.023 |
| N density | | 0.059 | 0.030 | 0.020 |
| N chlorides | | 0.042 | 0.021 | 0.012 |
| N fixed acidity | | 0.025 | 0.013 | 0.007 |
| N free sulfur dioxide | | 0.021 | 0.010 | 0.004 |
| N pH | | 0.012 | 0.006 | 0.002 |
| N residual sugar | | 0.008 | 0.004 | 0.002 |

*Fig.3. Ranking of red wine attributes in Orange*

The same scoring methods were used to rank the white wine attributes. It is important to note that Orange ranked the attributes differently than Weka. The following attributes were retained for further analysis: alcohol, density, chlorides, and alcohol.



Rank (1)

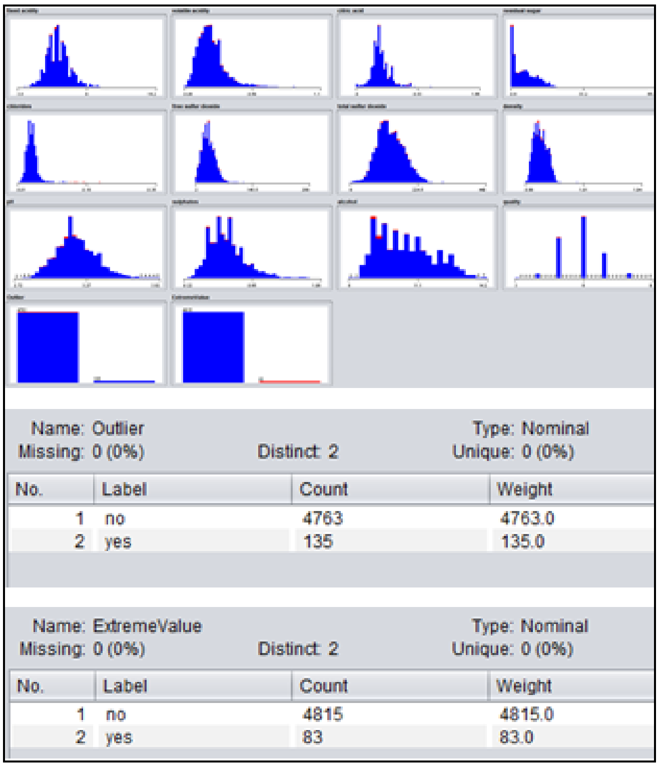
Scoring Methods
☑ Information Gain
☑ Information Gain Ratio
☑ Gini Decrease
☐ ANOVA
☐ $\chi^2$
☐ ReliefF
☐ FCBF

| | # | Info. gain | Gain ratio | Gini |
|---|---|---|---|---|
| N alcohol | | 0.163 | 0.082 | 0.047 |
| N density | | 0.113 | 0.057 | 0.031 |
| N chlorides | | 0.078 | 0.039 | 0.022 |
| N total sulfur dioxide | | 0.048 | 0.024 | 0.014 |
| N citric acid | | 0.047 | 0.023 | 0.013 |
| N volatile acidity | | 0.044 | 0.022 | 0.015 |
| N free sulfur dioxide | | 0.037 | 0.018 | 0.007 |
| N residual sugar | | 0.034 | 0.017 | 0.009 |
| N pH | | 0.017 | 0.009 | 0.005 |
| N sulphates | | 0.013 | 0.007 | 0.003 |
| N fixed acidity | | 0.010 | 0.005 | 0.002 |

*Fig.4. Ranking of white wine attributes in Orange*

### *2) Outliers*

Both data sets contained outliers, however, it was not deemed necessary to exclude the outliers as they are natural variations in the wine rather than erroneous data.



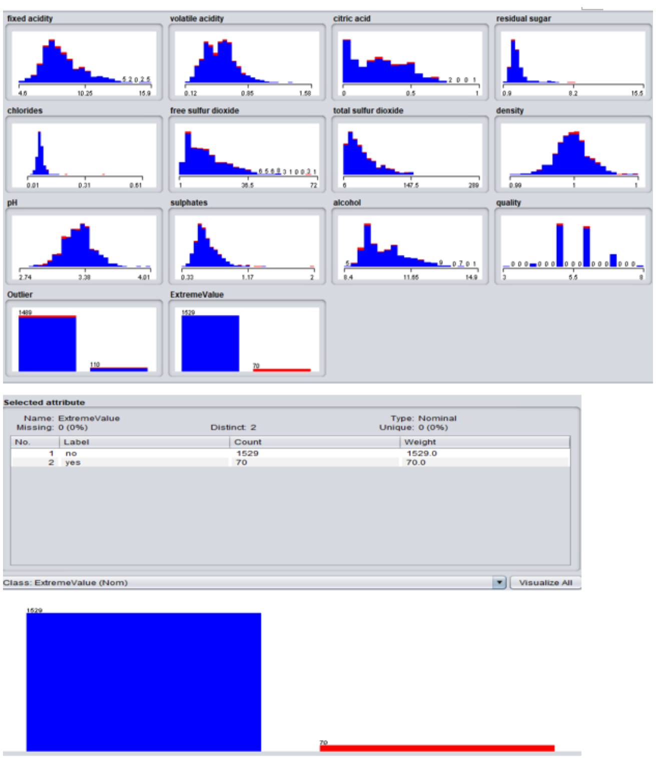
Name: Outlier    Type: Nominal
Missing: 0 (0%)    Distinct: 2    Unique: 0 (0%)

| No. | Label | Count | Weight |
|---|---|---|---|
| 1 | no | 4763 | 4763.0 |
| 2 | yes | 135 | 135.0 |

Name: ExtremeValue    Type: Nominal
Missing: 0 (0%)    Distinct: 2    Unique: 0 (0%)

| No. | Label | Count | Weight |
|---|---|---|---|
| 1 | no | 4815 | 4815.0 |
| 2 | yes | 83 | 83.0 |

*Fig. 5. Test run for outliers in red wine data using WEKA-Preprocess-Innerquartile Range*



Selected attribute

Name: ExtremeValue
Missing: 0 (0%)
Distinct: 2
Type: Nominal
Unique: 0 (0%)

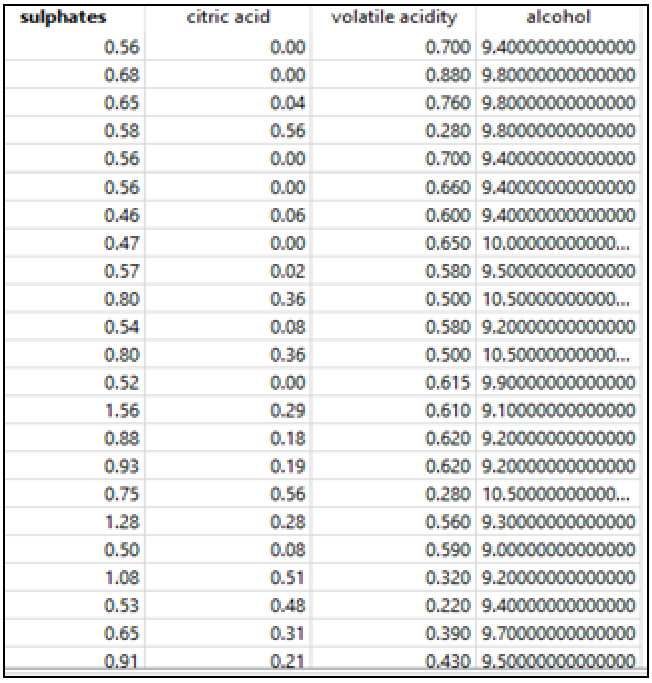
| No. | Label | Count | Weight |
|---|---|---|---|
| 1 | no | 1529 | 1529.0 |
| 2 | yes | 70 | 70.0 |

Class: ExtremeValue (Nom)

*Fig. 6. Test run for outliers in white wine data using WEKA-Preprocess-Innerquartile Range*

### 3) *Prepairing Data for Further Anlysis*

As the ranking system in Orange is better documented and more transparent than that of Weka, the attributes used in the further analysis of the data were those identified in Orange. The data sets were limited to the four attributes identified for clustering purposes and the same attributes with the additional of the "quality" classifier for classification.

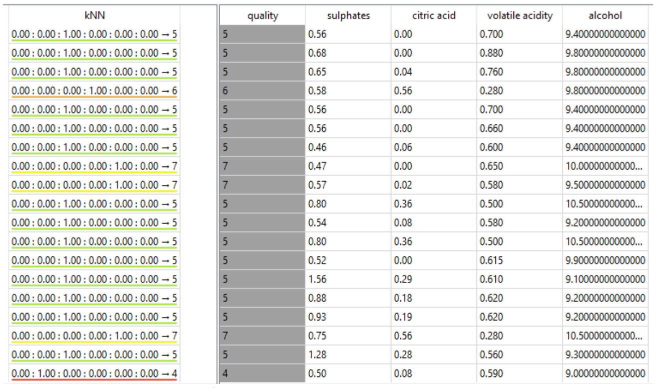


| sulphates | citric acid | volatile acidity | alcohol |
|---|---|---|---|
| 0.56 | 0.00 | 0.700 | 9.40000000000000 |
| 0.68 | 0.00 | 0.880 | 9.80000000000000 |
| 0.65 | 0.04 | 0.760 | 9.80000000000000 |
| 0.58 | 0.56 | 0.280 | 9.80000000000000 |
| 0.56 | 0.00 | 0.700 | 9.40000000000000 |
| 0.56 | 0.00 | 0.660 | 9.40000000000000 |
| 0.46 | 0.06 | 0.600 | 9.40000000000000 |
| 0.47 | 0.00 | 0.650 | 10.00000000000... |
| 0.57 | 0.02 | 0.580 | 9.50000000000000 |
| 0.80 | 0.36 | 0.500 | 10.50000000000... |
| 0.54 | 0.08 | 0.580 | 9.20000000000000 |
| 0.80 | 0.36 | 0.500 | 10.50000000000... |
| 0.52 | 0.00 | 0.615 | 9.90000000000000 |
| 1.56 | 0.29 | 0.610 | 9.10000000000000 |
| 0.88 | 0.18 | 0.620 | 9.20000000000000 |
| 0.93 | 0.19 | 0.620 | 9.20000000000000 |
| 0.75 | 0.56 | 0.280 | 10.50000000000... |
| 1.28 | 0.28 | 0.560 | 9.30000000000000 |
| 0.50 | 0.08 | 0.590 | 9.00000000000000 |
| 1.08 | 0.51 | 0.320 | 9.20000000000000 |
| 0.53 | 0.48 | 0.220 | 9.40000000000000 |
| 0.65 | 0.31 | 0.390 | 9.70000000000000 |
| 0.91 | 0.21 | 0.430 | 9.50000000000000 |

*Fig.7. Example of the red wine data set containing only the attributes identified as having the highest relative importance*

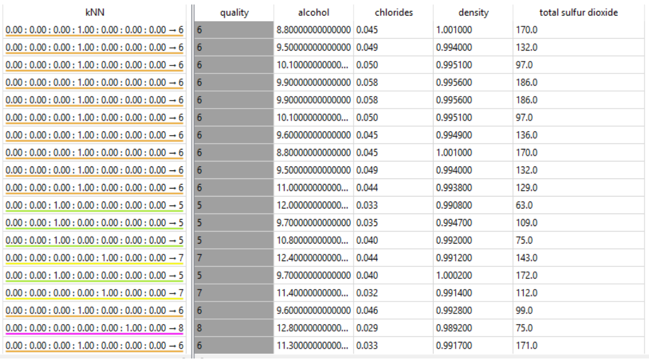
## II. FINDINGS

### A. *Classification Findings*



| kNN | quality | sulphates | citric acid | volatile acidity | alcohol |
|---|---|---|---|---|---|
| 0.00 : 0.00 : 1.00 : 0.00 : 0.00 : 0.00 → 5 | 5 | 0.56 | 0.00 | 0.700 | 9.4000000000000 |
| 0.00 : 0.00 : 1.00 : 0.00 : 0.00 : 0.00 → 5 | 5 | 0.68 | 0.00 | 0.880 | 9.8000000000000 |
| 0.00 : 0.00 : 1.00 : 0.00 : 0.00 : 0.00 → 5 | 5 | 0.65 | 0.04 | 0.760 | 9.8000000000000 |
| 0.00 : 0.00 : 0.00 : 1.00 : 0.00 : 0.00 → 6 | 6 | 0.58 | 0.56 | 0.280 | 9.8000000000000 |
| 0.00 : 0.00 : 1.00 : 0.00 : 0.00 : 0.00 → 5 | 5 | 0.56 | 0.00 | 0.700 | 9.4000000000000 |
| 0.00 : 0.00 : 1.00 : 0.00 : 0.00 : 0.00 → 5 | 5 | 0.56 | 0.00 | 0.660 | 9.4000000000000 |
| 0.00 : 0.00 : 1.00 : 0.00 : 0.00 : 0.00 → 5 | 5 | 0.46 | 0.06 | 0.600 | 9.4000000000000 |
| 0.00 : 0.00 : 0.00 : 0.00 : 1.00 : 0.00 → 7 | 7 | 0.47 | 0.00 | 0.650 | 10.0000000000... |
| 0.00 : 0.00 : 0.00 : 0.00 : 1.00 : 0.00 → 7 | 7 | 0.57 | 0.02 | 0.580 | 9.5000000000000 |
| 0.00 : 0.00 : 1.00 : 0.00 : 0.00 : 0.00 → 5 | 5 | 0.80 | 0.36 | 0.500 | 10.5000000000... |
| 0.00 : 0.00 : 1.00 : 0.00 : 0.00 : 0.00 → 5 | 5 | 0.54 | 0.08 | 0.580 | 9.2000000000000 |
| 0.00 : 0.00 : 1.00 : 0.00 : 0.00 : 0.00 → 5 | 5 | 0.80 | 0.36 | 0.500 | 10.5000000000... |
| 0.00 : 0.00 : 1.00 : 0.00 : 0.00 : 0.00 → 5 | 5 | 0.52 | 0.00 | 0.615 | 9.9000000000000 |
| 0.00 : 0.00 : 1.00 : 0.00 : 0.00 : 0.00 → 5 | 5 | 1.56 | 0.29 | 0.610 | 9.1000000000000 |
| 0.00 : 0.00 : 1.00 : 0.00 : 0.00 : 0.00 → 5 | 5 | 0.88 | 0.18 | 0.620 | 9.2000000000000 |
| 0.00 : 0.00 : 1.00 : 0.00 : 0.00 : 0.00 → 5 | 5 | 0.93 | 0.19 | 0.620 | 9.2000000000000 |
| 0.00 : 0.00 : 0.00 : 0.00 : 1.00 : 0.00 → 7 | 7 | 0.75 | 0.56 | 0.280 | 10.5000000000... |
| 0.00 : 0.00 : 1.00 : 0.00 : 0.00 : 0.00 → 5 | 5 | 1.28 | 0.28 | 0.560 | 9.3000000000000 |
| 0.00 : 1.00 : 0.00 : 0.00 : 0.00 : 0.00 → 4 | 4 | 0.50 | 0.08 | 0.590 | 9.0000000000000 |

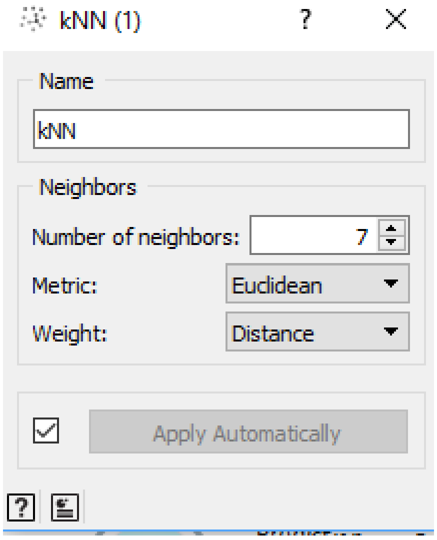
*Fig.8. kNN predictions for red wine data set analysis*



| kNN | quality | alcohol | chlorides | density | total sulfur dioxide |
|---|---|---|---|---|---|
| 0.00 : 0.00 : 0.00 : 1.00 : 0.00 : 0.00 : 0.00 → 6 | 6 | 8.80000000000000 | 0.045 | 1.001000 | 170.0 |
| 0.00 : 0.00 : 0.00 : 1.00 : 0.00 : 0.00 : 0.00 → 6 | 6 | 9.50000000000000 | 0.049 | 0.994000 | 132.0 |
| 0.00 : 0.00 : 0.00 : 1.00 : 0.00 : 0.00 : 0.00 → 6 | 6 | 10.10000000000... | 0.050 | 0.995100 | 97.0 |
| 0.00 : 0.00 : 0.00 : 1.00 : 0.00 : 0.00 : 0.00 → 6 | 6 | 9.90000000000000 | 0.058 | 0.995600 | 186.0 |
| 0.00 : 0.00 : 0.00 : 1.00 : 0.00 : 0.00 : 0.00 → 6 | 6 | 9.90000000000000 | 0.058 | 0.995600 | 186.0 |
| 0.00 : 0.00 : 0.00 : 1.00 : 0.00 : 0.00 : 0.00 → 6 | 6 | 10.10000000000... | 0.050 | 0.995100 | 97.0 |
| 0.00 : 0.00 : 0.00 : 1.00 : 0.00 : 0.00 : 0.00 → 6 | 6 | 9.60000000000000 | 0.045 | 0.994900 | 136.0 |
| 0.00 : 0.00 : 0.00 : 1.00 : 0.00 : 0.00 : 0.00 → 6 | 6 | 8.80000000000000 | 0.045 | 1.001000 | 170.0 |
| 0.00 : 0.00 : 0.00 : 1.00 : 0.00 : 0.00 : 0.00 → 6 | 6 | 9.50000000000000 | 0.049 | 0.994000 | 132.0 |
| 0.00 : 0.00 : 0.00 : 1.00 : 0.00 : 0.00 : 0.00 → 6 | 6 | 11.00000000000... | 0.044 | 0.993800 | 129.0 |
| 0.00 : 0.00 : 1.00 : 0.00 : 0.00 : 0.00 : 0.00 → 5 | 5 | 12.00000000000... | 0.033 | 0.990800 | 63.0 |
| 0.00 : 0.00 : 1.00 : 0.00 : 0.00 : 0.00 : 0.00 → 5 | 5 | 9.70000000000000 | 0.035 | 0.994700 | 109.0 |
| 0.00 : 0.00 : 1.00 : 0.00 : 0.00 : 0.00 : 0.00 → 5 | 5 | 10.80000000000... | 0.040 | 0.992000 | 75.0 |
| 0.00 : 0.00 : 0.00 : 0.00 : 1.00 : 0.00 : 0.00 → 7 | 7 | 12.40000000000... | 0.044 | 0.991200 | 143.0 |
| 0.00 : 0.00 : 1.00 : 0.00 : 0.00 : 0.00 : 0.00 → 5 | 5 | 9.70000000000000 | 0.040 | 1.000200 | 172.0 |
| 0.00 : 0.00 : 0.00 : 0.00 : 1.00 : 0.00 : 0.00 → 7 | 7 | 11.40000000000... | 0.032 | 0.991400 | 112.0 |
| 0.00 : 0.00 : 0.00 : 1.00 : 0.00 : 0.00 : 0.00 → 6 | 6 | 9.60000000000000 | 0.046 | 0.992800 | 99.0 |
| 0.00 : 0.00 : 0.00 : 0.00 : 0.00 : 1.00 : 0.00 → 8 | 8 | 12.80000000000... | 0.029 | 0.989200 | 75.0 |
| 0.00 : 0.00 : 0.00 : 1.00 : 0.00 : 0.00 : 0.00 → 6 | 6 | 11.30000000000... | 0.033 | 0.991700 | 171.0 |

*Fig.9. kNN predictions for red wine data set analysis*

For the kNN analysis, the number of neighbors for each data set were chosen based off the number of unique values in the "quality" attribute. For white wine this was 6 neighbors and for red wine it was 7.



Name

kNN

Neighbors

Number of neighbors: 7

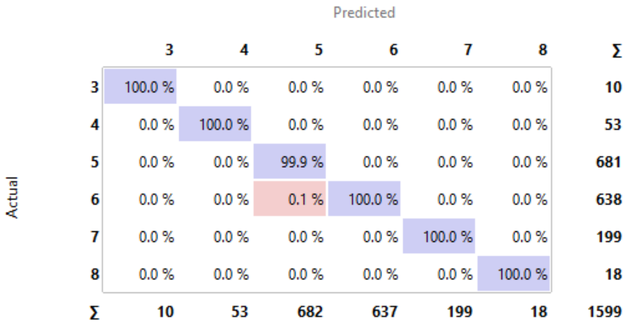
Metric: Euclidean

Weight: Distance

☑ Apply Automatically

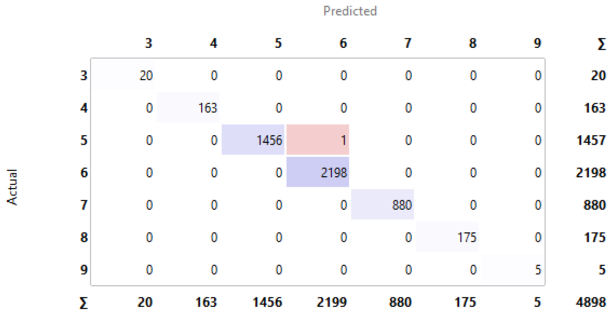
*Fig.10. kNN settings for red wine data set analysis*

The application of kNN to the Wine Data Set produced highly accurate results. When the white wine kNN results were tested the following results were produced: AUC .773, CA: .609, F1 .598, precision .597, and recall .609. For red wine the test of the KNN algorithm yielded the following results: .774 AUC, .538 CA, .532 F1, .532 precision and .538 recall. The classification matrix for both the analysis of the red wine and that of the white wine show only one misidentified instance each.



| Actual \ Predicted | 3 | 4 | 5 | 6 | 7 | 8 | Σ |
|---|---|---|---|---|---|---|---|
| 3 | 100.0 % | 0.0 % | 0.0 % | 0.0 % | 0.0 % | 0.0 % | 10 |
| 4 | 0.0 % | 100.0 % | 0.0 % | 0.0 % | 0.0 % | 0.0 % | 53 |
| 5 | 0.0 % | 0.0 % | 99.9 % | 0.0 % | 0.0 % | 0.0 % | 681 |
| 6 | 0.0 % | 0.0 % | 0.1 % | 100.0 % | 0.0 % | 0.0 % | 638 |
| 7 | 0.0 % | 0.0 % | 0.0 % | 0.0 % | 100.0 % | 0.0 % | 199 |
| 8 | 0.0 % | 0.0 % | 0.0 % | 0.0 % | 0.0 % | 100.0 % | 18 |
| Σ | 10 | 53 | 682 | 637 | 199 | 18 | 1599 |

*Fig.11. kNN confusion matrix for red wine data set*



| Actual \ Predicted | 3 | 4 | 5 | 6 | 7 | 8 | 9 | Σ |
|---|---|---|---|---|---|---|---|---|
| 3 | 20 | 0 | 0 | 0 | 0 | 0 | 0 | 20 |
| 4 | 0 | 163 | 0 | 0 | 0 | 0 | 0 | 163 |
| 5 | 0 | 0 | 1456 | 1 | 0 | 0 | 0 | 1457 |
| 6 | 0 | 0 | 0 | 2198 | 0 | 0 | 0 | 2198 |
| 7 | 0 | 0 | 0 | 0 | 880 | 0 | 0 | 880 |
| 8 | 0 | 0 | 0 | 0 | 0 | 175 | 0 | 175 |
| 9 | 0 | 0 | 0 | 0 | 0 | 0 | 5 | 5 |
| Σ | 20 | 163 | 1456 | 2199 | 880 | 175 | 5 | 4898 |

*Fig.12. kNN confusion matrix for white wine data set*

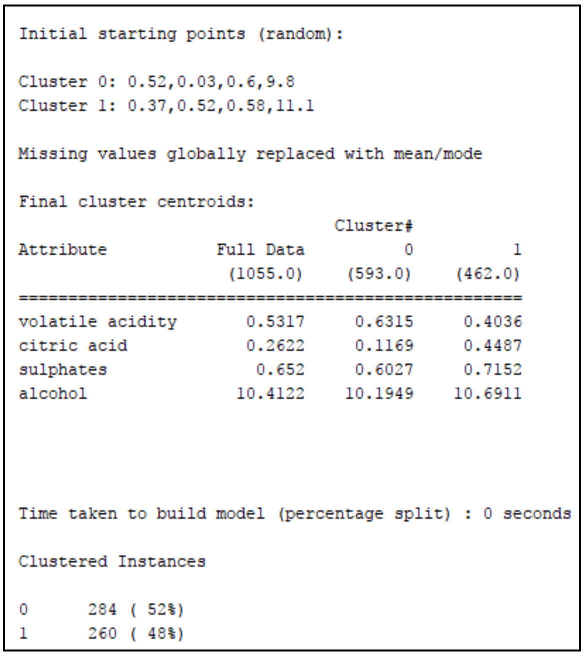
### B. *Clustering Findings*

The attempts to cluster were somewhat less successful. For red wine, KMeans showed two clusters with an almost equal split while EM produced 10 clusters. To compare the results from one to the other is somewhat difficult as they do not provide the same metrics (log likelihood, within cluster sum of squared errors, etc.). For white wine, KMeans produced two clusters with 43 and 57% in each cluster. The canopy and EM algorithms produced 5 and 10 clusters, respectively. As with the red wine set it is somewhat difficult to compare one algorithm to the other as they use different scoring metrics. Although an attempt was made to cluster each dataset with various algorithms, only a few with potential meaningful results are provided here. Further exploration would be needed to determine a “best” cluster result. The goal was to see if subtypes emerged. Thus, more than two clusters would be ideal. How many clusters would be ideal would depend on the potential use of the results (i.e. how specific wine recommendations would be).

It was interesting to note that selected attributes were not the same for white wine and red wine. Quality was found to be related to alcohol, sulphates, volatile acidity, and citric acid for the red wine data set. On the other hand, for the white wine set, quality was not found to be a selected attribute at all. The selected attributes were chlorides, total sulfur dioxide, density, and alcohol.

#### 1) *Red Wine*

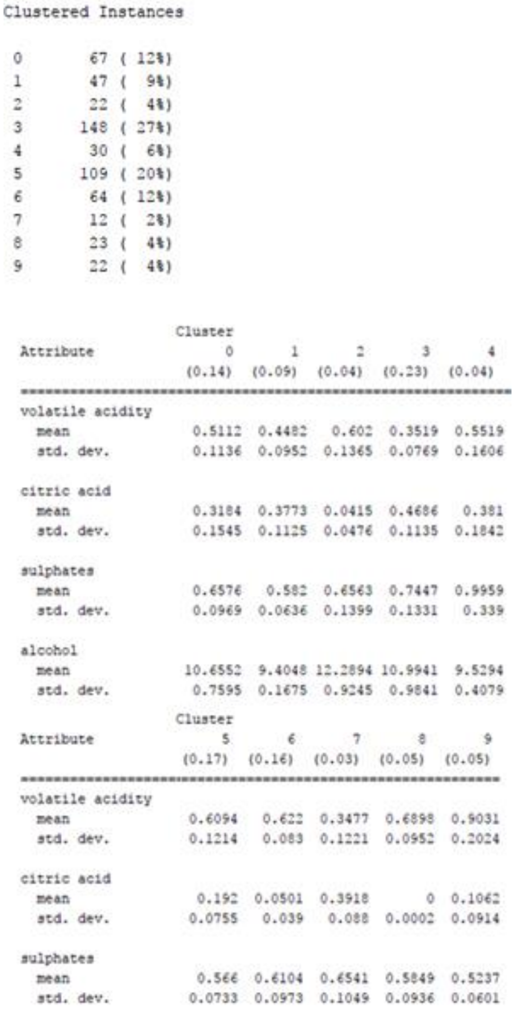
##### a) K-Means Clustering



```
Initial starting points (random):

Cluster 0: 0.52,0.03,0.6,9.8
Cluster 1: 0.37,0.52,0.58,11.1

Missing values globally replaced with mean/mode

Final cluster centroids:
                                Cluster#
Attribute           Full Data          0           1
                     (1055.0)     (593.0)     (462.0)
====================================================
volatile acidity       0.5317      0.6315      0.4036
citric acid            0.2622      0.1169      0.4487
sulphates               0.652      0.6027      0.7152
alcohol               10.4122     10.1949     10.6911




Time taken to build model (percentage split) : 0 seconds

Clustered Instances

0      284 ( 52%)
1      260 ( 48%)
```

*Fig.13. K-means clustering analysis (66% split)*

##### b) EM Clustering

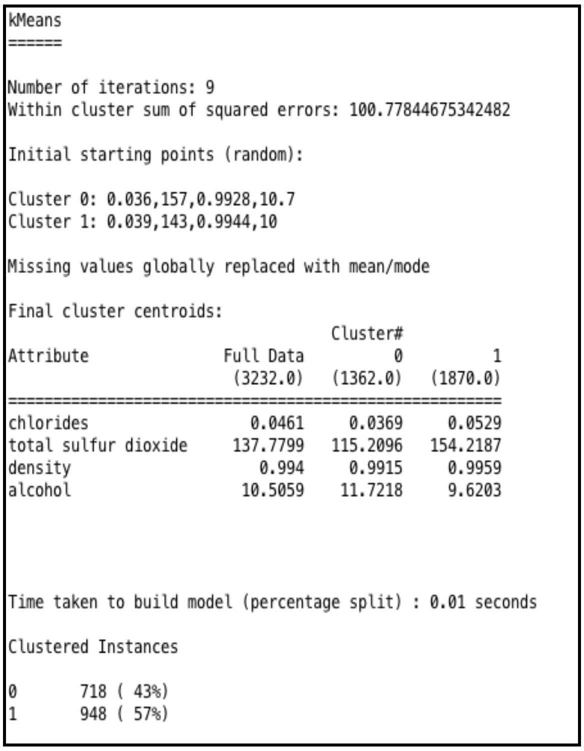


```
Clustered Instances

0       67 ( 12%)
1       47 (  9%)
2       22 (  4%)
3      148 ( 27%)
4       30 (  6%)
5      109 ( 20%)
6       64 ( 12%)
7       12 (  2%)
8       23 (  4%)
9       22 (  4%)


                       Cluster
Attribute                    0        1        2        3        4
                        (0.14)   (0.09)   (0.04)   (0.23)   (0.04)
==================================================================
volatile acidity
  mean                  0.5112   0.4482    0.602   0.3519   0.5519
  std. dev.             0.1136   0.0952   0.1365   0.0769   0.1606

citric acid
  mean                  0.3184   0.3773   0.0415   0.4686    0.381
  std. dev.             0.1545   0.1125   0.0476   0.1135   0.1842

sulphates
  mean                  0.6576    0.582   0.6563   0.7447   0.9959
  std. dev.             0.0969   0.0636   0.1399   0.1331    0.339

alcohol
  mean                 10.6552   9.4048  12.2894  10.9941   9.5294
  std. dev.             0.7595   0.1675   0.9245   0.9841   0.4079
                       Cluster
Attribute                    5        6        7        8        9
                        (0.17)   (0.16)   (0.03)   (0.05)   (0.05)
==================================================================
volatile acidity
  mean                  0.6094    0.622   0.3477   0.6898   0.9031
  std. dev.             0.1214    0.083   0.1221   0.0952   0.2024

citric acid
  mean                   0.192   0.0501   0.3918        0   0.1062
  std. dev.             0.0755    0.039    0.088   0.0002   0.0914

sulphates
  mean                   0.566   0.6104   0.6541   0.5849   0.5237
  std. dev.             0.0733   0.0973   0.1049   0.0936   0.0601
```

*Fig.14. EM clustering analysis (66% split)*

### 2) *White Wine*

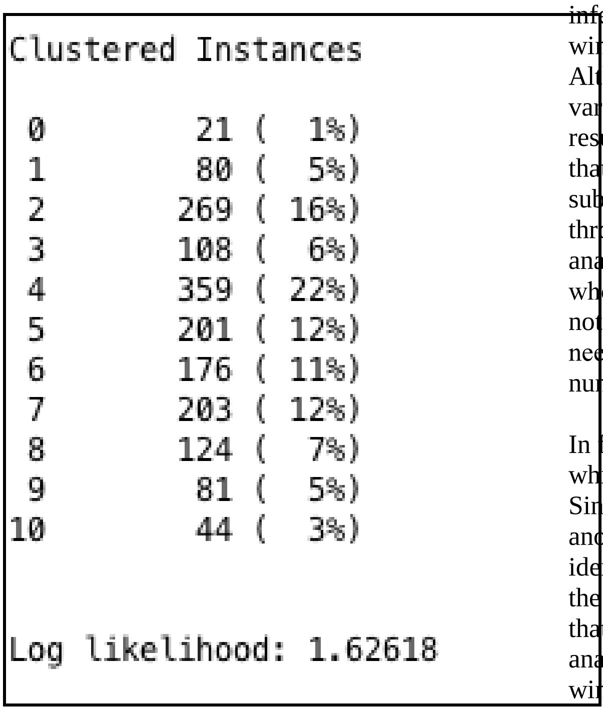
a) K-Means



```
kMeans
======

Number of iterations: 9
Within cluster sum of squared errors: 100.77844675342482

Initial starting points (random):

Cluster 0: 0.036,157,0.9928,10.7
Cluster 1: 0.039,143,0.9944,10

Missing values globally replaced with mean/mode

Final cluster centroids:
                                    Cluster#
Attribute              Full Data           0           1
                        (3232.0)    (1362.0)    (1870.0)
==========================================================
chlorides                 0.0461      0.0369      0.0529
total sulfur dioxide    137.7799    115.2096    154.2187
density                    0.994      0.9915      0.9959
alcohol                  10.5059     11.7218      9.6203




Time taken to build model (percentage split) : 0.01 seconds

Clustered Instances

0       718 ( 43%)
1       948 ( 57%)
```

*Fig.15. K-means clustering analysis (66% split)*

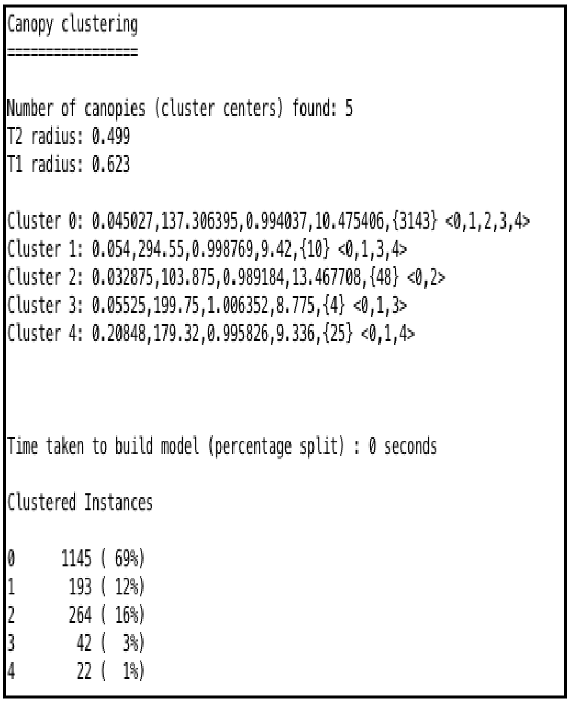
b) EM Clustering



```
Clustered Instances

 0         21 (  1%)
 1         80 (  5%)
 2        269 ( 16%)
 3        108 (  6%)
 4        359 ( 22%)
 5        201 ( 12%)
 6        176 ( 11%)
 7        203 ( 12%)
 8        124 (  7%)
 9         81 (  5%)
10         44 (  3%)


Log likelihood: 1.62618
```

*Fig.16. EM clustering analysis (66% training set)*

c) Canopy Clustering.



```
Canopy clustering
=================


Number of canopies (cluster centers) found: 5
T2 radius: 0.499
T1 radius: 0.623

Cluster 0: 0.045027,137.306395,0.994037,10.475406,{3143} <0,1,2,3,4>
Cluster 1: 0.054,294.55,0.998769,9.42,{10} <0,1,3,4>
Cluster 2: 0.032875,103.875,0.989184,13.467708,{48} <0,2>
Cluster 3: 0.05525,199.75,1.006352,8.775,{4} <0,1,3>
Cluster 4: 0.20848,179.32,0.995826,9.336,{25} <0,1,4>



Time taken to build model (percentage split) : 0 seconds

Clustered Instances

0      1145 ( 69%)
1       193 ( 12%)
2       264 ( 16%)
3        42 (  3%)
4        22 (  1%)
```

*Fig.17. Canopy clustering analysis (66% training set*

## III. CONCLUSION

When utilizing the four highest ranked attributes of each data set within the Wine Quality Data Set, highly accurate classification results can be produced. This supports the inference that wine of a similar quality, as determined by wine experts, contain similar physiochemical attributes. Although an attempt was made to cluster each dataset with various algorithms, only a few with potential meaningful results are provided here.  From these results, it does seem that clustering may be a viable option for identifying subtypes by utilizing only objective attributes determined through physiochemical chemical testing. This type of analysis could prove useful for recommendation systems when expert opinions (such as the “quality” attribute) are not available.  However, further exploration would be needed determine a “best” cluster result as the “ideal” number would depend on the potential use of the results.

In future studies association rules could be analyzed to see which are the stronger for white wine versus red wine. Since the algorithms cluster the wines by similarity to one another rather than the similarities to a wine with an identified ranking, it would also be interesting to compare the clusters that the algorithm produced to the groupings that are based off the expert rankings. The results of such analysis could support that the chemical components of the wine dictate its quality.

# REFERENCES

[1] P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553, 2009.*)*

[2] UCI Machine Learning Repository: Wine Quality Data Set. (n.d.). Retrieved from https://archive.ics.uci.edu/ml/datasets/Wine Quality

# APPENDIX I

We approached the Wine Data Quality Set with the idea of analyzing it in both Orange and Weka to see what we could learn from the data and what type of algorithms the data would work with. Due to technical difficulties, Johanna worked using the Weka program and Victoria worked using the Orange program. Except for the attribute ranking, our initial tests to determine data quality produced the same results in both programs. It is interesting to note that the attribute ranking produced the same top four attributes for red wine, but different attributes for white wine.
The breakdown of our contributions to the project are as following:

Johanna:

- Analyzed red and white wine data sets in Weka
- Prepared the initial outline of report
- Wrote analysis of results from clustering algorithms in Weka
- Wrote initial conclusion

Victoria:

- Analyzed red and white wine data sets in Orange
- Wrote analysis of results from clustering in Orange
- Formatted report
- Complete introduction and edited conclusion