Wine quality analysis by data mining from physicochemical properties

Yijing Tan

Abstract: In this study, we propose a data mining approach to understand how different physiochemical properties affect wine quality and to predict wine taste preferences that is based on easily available analytical tests at the certification step. Six regression techniques were applied, under a computationally efficient procedure that performs simultaneous variable and model selection. The support vector machine achieved promising results, outperforming than the others. Such a model is useful to support the oenologist wine tasting evaluations and improve wine production. Furthermore, similar techniques can help in target marketing by modeling consumer tastes from niche markets.

1 Introduction

It is difficult to determine an empirical relation between the subjective quality of a wine and its chemical composition. Winemakers want to know what they can do to their processes to optimize the quality of their wine. In addition to testing new algorithms and variations of algorithms for prediction purposes, we are also interested in analyzing the data itself. For example, do some ingredients have a stronger impact on the perceived wine quality that others? Should winemakers be focusing more on certain ingredients than others?

The aim of this project is to predict the quality of wine given a set of features as inputs. The dataset used is Wine Quality Data set from Kaggle.com. Input variables are fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol. And the output variable is quality. We are dealing only with red wine. We have quality being one of these values: [3, 4, 5, 6, 7, 8]. The higher the value the better the quality. In this project we will treat each class of the wine separately and their aim is to be able and find decision boundaries that work well for new unseen data. These are the classifiers.

In this paper we are explaining the steps we followed to build our models for predicting the quality of red wine in a simple non-technical way.

2 Materials and methods

2.1 Data set:

Dataset/Source: Kaggle https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009
Structured/Unstructured data: Structured Data in CSV format.

2.2 Dataset Description and features:

In the data set, there 1599 different wine as row data and 12 features as columns. Furthermore, there are no null values to deal with, all values are numeric, input values are float and only output value is integer.

In more detail,

• Fixed acidity: a measurement of the total concentration of titratable acids and free hydrogen ions present in the wine. Theoretically, having a low acidity will result in a flat and boring wine while having too much acid can lead to tartness or even a sour wine. These acids either occur naturally in the grapes or are created through the fermentation process.
• Volatile acidity: a measure of steam distillable acids present in a wine. In theory, our palates are quite sensitive to the presence of volatile acids and for that reason a good wine should keep their concentrations as low as possible.
• Citric acid: one of the many acids that are measured to obtained fixed acidity.
• Residual sugar: measurement of any natural grape sugars that are leftover after fermentation ceases. In theory residual sugar can help wines age well.
• Chlorides: the amount of salt in the wine.
• Free sulfur dioxide: the free form of SO2 exists in equilibrium between molecular SO2 (as a dissolved gas) and bisulfite ion; it prevents microbial growth and the oxidation of wine.
• Total sulfur dioxide: amount of free and bound forms of SO2; in low concentrations, SO2 is mostly undetectable in wine, but at free SO2 concentrations over 50 ppm, SO2 becomes evident in the nose and taste of wine.
• Density: measure of density of wine.
• pH: value for pH.
• Sulfates: a wine additive which can contribute to sulfur dioxide gas (SO2) levels, which acts as an

antimicrobial and antioxidant.
• Alcohol: the percentage of alcohol present in the wine.
• Quality: Wine experts graded the wine quality between 0 (very bad) and 10 (very excellent). The eventual quality score is the median of at least three evaluations made by the same wine experts.

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
count	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000
mean	8.319637	0.527821	0.270976	2.538806	0.087467	15.874922	46.467792	0.996747	3.311113	0.658149	10.422983	5.636023
std	1.741096	0.179060	0.194801	1.409928	0.047065	10.460157	32.895324	0.001887	0.154386	0.169507	1.065668	0.807569
min	4.600000	0.120000	0.000000	0.900000	0.012000	1.000000	6.000000	0.990070	2.740000	0.330000	8.400000	3.000000
25%	7.100000	0.390000	0.090000	1.900000	0.070000	7.000000	22.000000	0.995600	3.210000	0.550000	9.500000	5.000000
50%	7.900000	0.520000	0.260000	2.200000	0.079000	14.000000	38.000000	0.996750	3.310000	0.620000	10.200000	6.000000
75%	9.200000	0.640000	0.420000	2.600000	0.090000	21.000000	62.000000	0.997835	3.400000	0.730000	11.100000	6.000000
max	15.900000	1.580000	1.000000	15.500000	0.611000	72.000000	289.000000	1.003690	4.010000	2.000000	14.900000	8.000000

TABLE I SUMMARY OF FEATURES AND RESPONSE VARIABLE

The measures of central tendency and variability or distribution are some commonly used measures to define the data set. The measures used to define the central tendency are mean, median and mode. The standard deviations are the minimum and maximum values of variables. The table above is a summary of some statistical measures for each numeric predictor of the dataset: **count** indicates the number of records for each attribute that corresponds to the number of wines. **mean** indicates the average value around which each group of attributes is attested. **std** indicates the standard deviation of each attribute group, from which we can guess the degree of data dispersion around the average. **max** and min indicate the attribute that I assume the highest and lowest value for each attribute group.

The histogram shown in Figure 1, shows that the values for the response variable are not uniformly distributed.

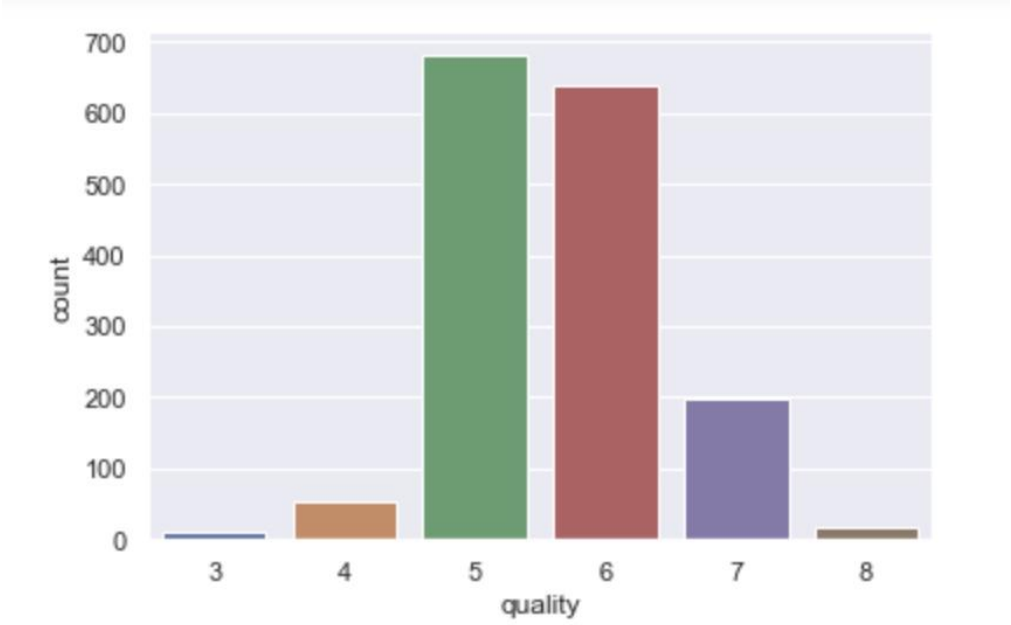


Fig. 1. Distribution of values for wine quality.

2.3 Data Visualization

This data set has many different features and it is important to understand relationship between these in order to analyze dataset better. For that reason, correlation map helps to understand these relations in a single representation. Correlation map is made by calculating the covariance of each features with respect to other. According to these information, it can be made a good analyze about dataset and columns.

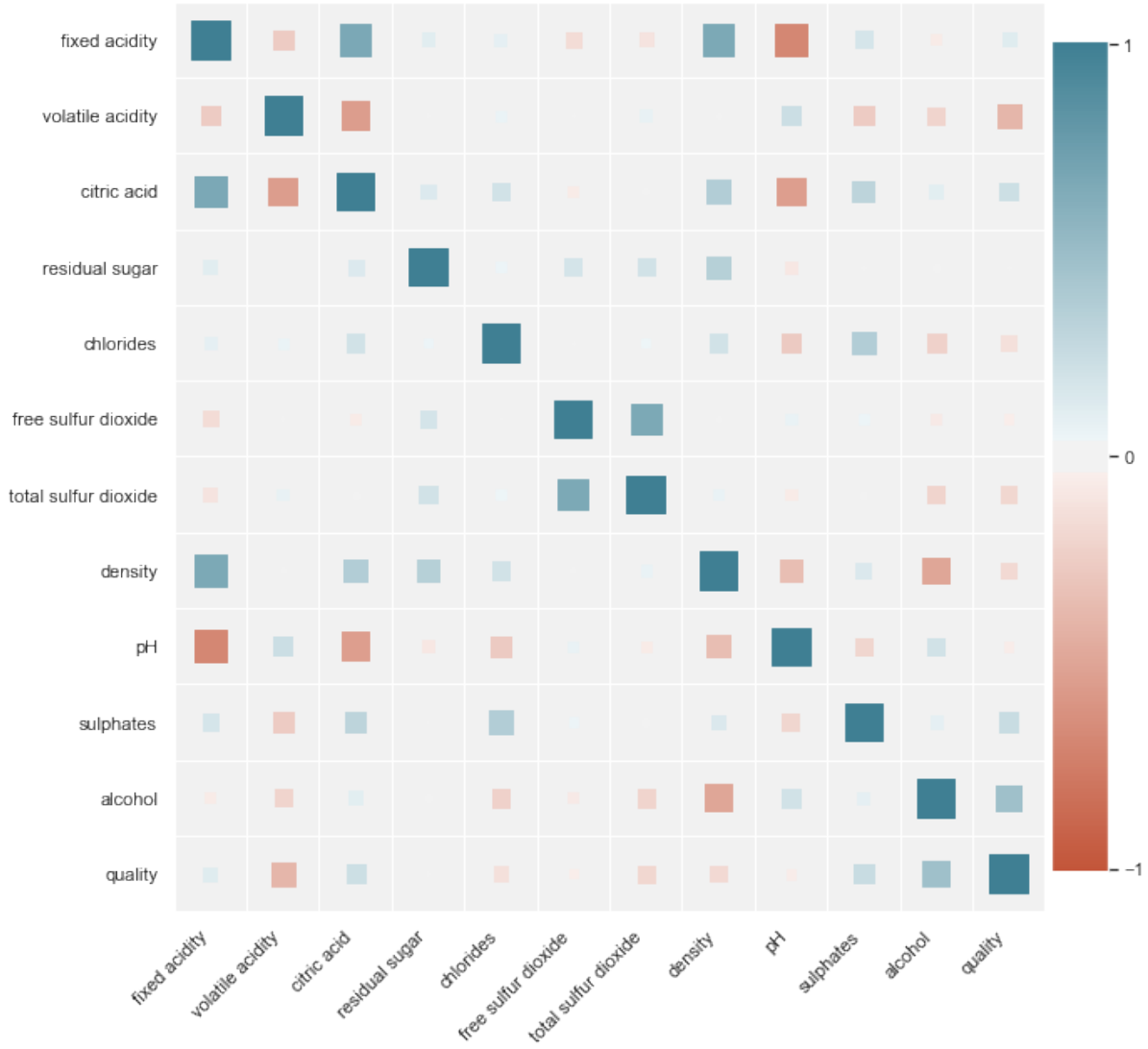


Fig. 2.Correlation map.

According to figure in above;

- Quality has a (+)positive relationship between alcohol
- Quality has a (-)negative weak relationship between volatile acidity
- Quality has almost no relationship between residual sugar, free sulfur dioxide, and pH.
- Alcohol has a (+)positive relationship between quality and weakly pH
- Alcohol has a (-)negative relationship between density
- Alcohol has almost no relationship between fixed acidity, residual sugar, free sulfur dioxide, sulphates
- Volatile acidity has a weak (+)positive relationship between pH.
- Volatile acidity has a strong (-)negative relationship between citric acid
  - Volatile acidity has weak (-)negative relationship between fixed acidity and sulphates
  - Volatile acidity has almost no relationship between residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density
  - Density has (+)positive relationship between fixed acidity
  - Density has (-)negative relationship between density

- Density has almost no relationship between volatile acidity, free sulfur dioxide, total sulfur dioxide
- Citric acid has (+)positive relationship between fixed acidity
- Citric acid has (-)negative relationship between volatile acidity, pH
- Citric acid has almost no relationship between residual sugar, free sulfur dioxide, total sulfur dioxide

2.4 Data processing methods

For making automated decisions on model selection we need to quantify the performance of our model and give it a score. For that reason, for the classifiers, we are using accuracy score which expresses how accurate the model was on predicting a certain class.

Splitting for Testing : Now that we've explored the data a bit, let's go ahead and split the data into training and testing sets. We are keeping 25% of our dataset to treat it as unseen data and be able and test the performance of our models. We are splitting our dataset in a way such that all of the wine qualities are represented proportionally equally in both training and testing dataset.
Other than that the selection is being done randomly with uniform distribution.
Various classification algorithms are used to fit the model. The algorithms used in this paper are as follows:

For classification:

Logistic Regression classifier
Random Forest classifier
Decision Tree classifier
QDA
K-Nearest Neighbors
Support Vector Machine classifier

1) Logistic Regression

The logistic regression is a predictive analysis of statistical method. Logistic regression is analyzing a dataset in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous variable (in which there are only two possible outcomes). It is used to predict a binary outcome (1 / 0, Yes / No, True / False) given a set of independent variables. Logistic regression is a classification algorithm used to assign observations to a discrete set of classes. For this data set, firstly using ordinal logistic regression then it is better to using binary logistic regression with modified dataset. Formulation of the logistic regression: $P = 1/(1 + e\text{^}(-(b\_0 + b\_1\, x))\,)$

Accuracy: Accuracy is one metric for evaluating classification models. For a given test data set, the ratio of the number of samples correctly classified by the classifier to the total number of samples.

Informally, accuracy is the fraction of predictions our model got right. Formally, accuracy has the following definition: Accuracy = Number of correct predictions/Total number of predictions.

- Logistic Regression gave us an accuracy of 57.25%

2) Random Forest

Random forest --> baging pf the decision tree Random forests construct many individual decision trees at training and it uses the simplicity of decision trees with flexibility resulting in improvement the accuracy. Predictions from all trees are pooled to make the final prediction; the mode of the classes for classification or the mean prediction for regression. As they use a collection of results to make a final decision. Random forest algorithm contains many variables, and many categorical variables with a large number of class labels. It gives results using data sets that show a loss or unbalanced distribution. When new trees are added into the random forest, algorithm updates itself with decreasing the loss by eliminating noises.

- Random Forest gave us an accuracy of 61.25%

3) Decision Tree

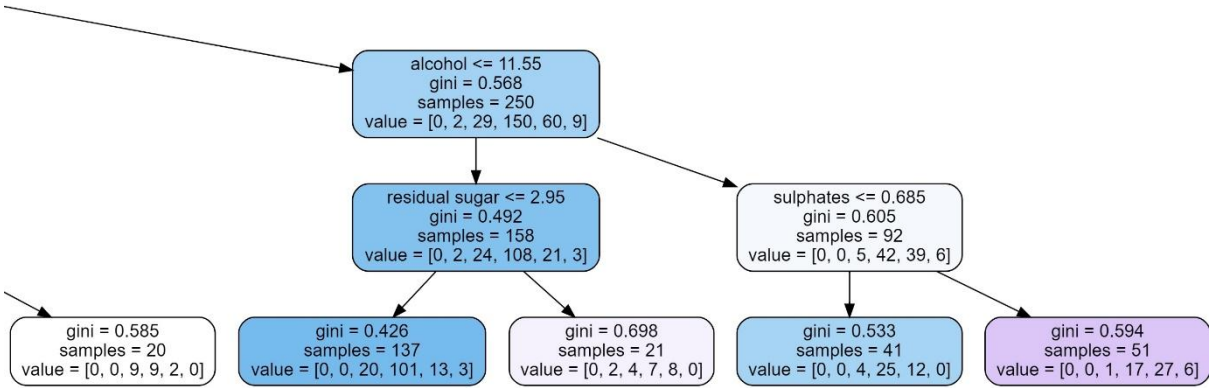
Tree models where the target variable can take a discrete set of values are called classification trees; in these tree structures, leaves represent class labels and branches represent conjunctions of features that lead to those class labels. The standard deviation is used to calculate the homogeneity of a numerical sample. After each standard deviation calculations, standard deviation reduction is used to classify dataset. The standard deviation reduction is based on the decrease in standard deviation after a dataset is split on an attribute. Constructing a decision tree is all about finding attribute that returns the highest standard deviation

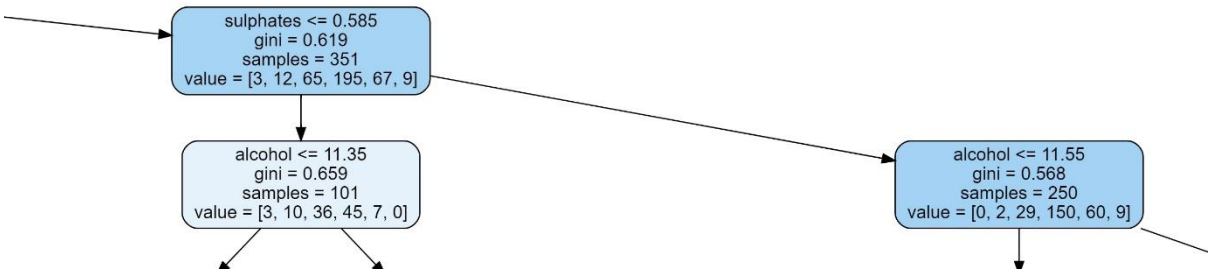
reduction. A decision tree is drawn upside down with its root at the top involves partitioning the data into subsets that contain instances with similar values (homogenous), then on the middle there are condition/internal node based on the tree split into branches/edges. The end of the branch that doesn't split anymore it is the decision/leaf tree, means that they are the last classification nodes(qualities). The base algorithm of the decision tree; recursive binary splitting. In this procedure, all the features are considered and different split points are tried and tested using a cost function. The split with the best cost (or lowest cost) is selected. The cost function is used to understand how model split and predict the split dataset classifications.

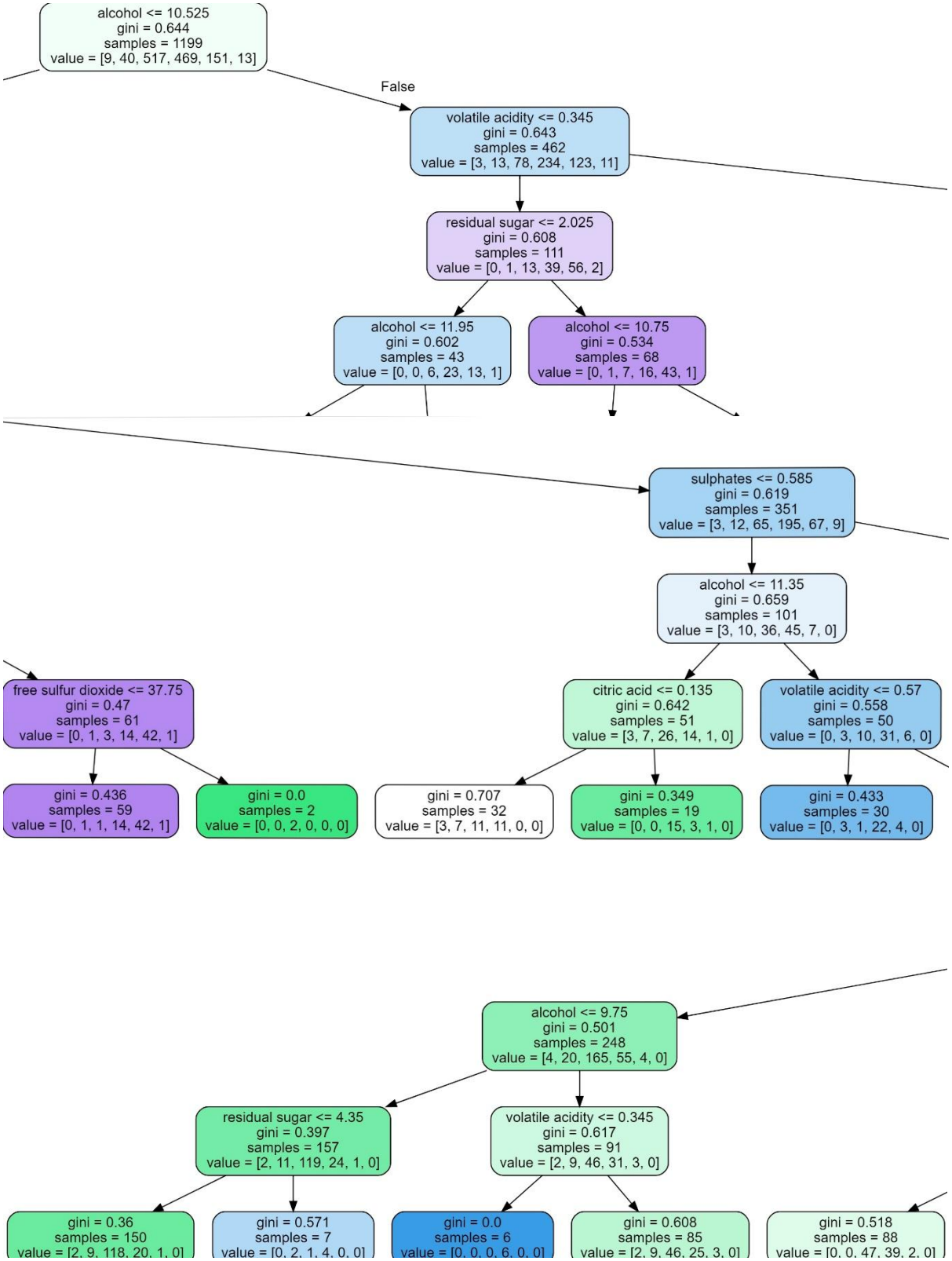
We initially used Decision Trees to classify the data based on the idea that, like regression, they would give us a good representation of which attributes were most important. Additionally, once built they offer a visual representation of the classifier. On the other hand, Decision Trees come with the difficulty of balancing generalization with over fitting. In general, these issues are solved by some combination of pre- and post-pruning. Some experimentation lead to the conclusion that for our data a tree depth of 5 gave the best results.

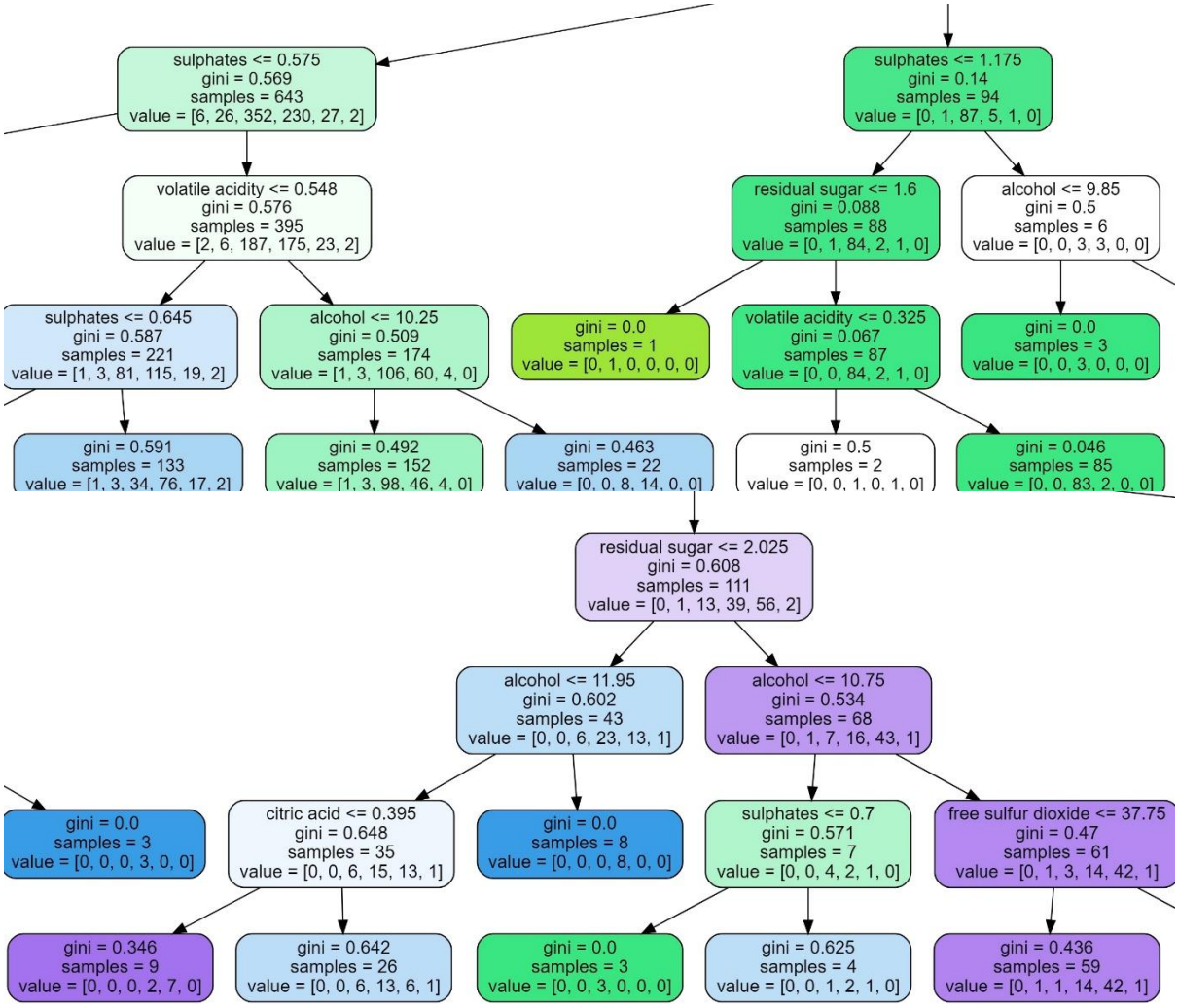
- Decision Tree gave us an accuracy of 57.49%

Fig. 3. Decision Tree Visualization.






alcohol <= 10.525
gini = 0.644
samples = 1199
value = [9, 40, 517, 469, 151, 13]
False
volatile acidity <= 0.345
gini = 0.643
samples = 462
value = [3, 13, 78, 234, 123, 11]
residual sugar <= 2.025
gini = 0.608
samples = 111
value = [0, 1, 13, 39, 56, 2]
alcohol <= 11.95
gini = 0.602
samples = 43
value = [0, 0, 6, 23, 13, 1]
alcohol <= 10.75
gini = 0.534
samples = 68
value = [0, 1, 7, 16, 43, 1]
sulphates <= 0.585
gini = 0.619
samples = 351
value = [3, 12, 65, 195, 67, 9]
alcohol <= 11.35
gini = 0.659
samples = 101
value = [3, 10, 36, 45, 7, 0]
free sulfur dioxide <= 37.75
gini = 0.47
samples = 61
value = [0, 1, 3, 14, 42, 1]
citric acid <= 0.135
gini = 0.642
samples = 51
value = [3, 7, 26, 14, 1, 0]
volatile acidity <= 0.57
gini = 0.558
samples = 50
value = [0, 3, 10, 31, 6, 0]
gini = 0.436
samples = 59
value = [0, 1, 1, 14, 42, 1]
gini = 0.0
samples = 2
value = [0, 0, 2, 0, 0, 0]
gini = 0.707
samples = 32
value = [3, 7, 11, 11, 0, 0]
gini = 0.349
samples = 19
value = [0, 0, 15, 3, 1, 0]
gini = 0.433
samples = 30
value = [0, 3, 1, 22, 4, 0]
alcohol <= 9.75
gini = 0.501
samples = 248
value = [4, 20, 165, 55, 4, 0]
residual sugar <= 4.35
gini = 0.397
samples = 157
value = [2, 11, 119, 24, 1, 0]
volatile acidity <= 0.345
gini = 0.617
samples = 91
value = [2, 9, 46, 31, 3, 0]
gini = 0.36
samples = 150
gini = 0.571
samples = 7
gini = 0.0
samples = 6
gini = 0.608
samples = 85
gini = 0.518
samples = 88



4) QDA

QDA is implemented in sklearn using the Quadratic Discriminant Analysis() function, which is again part of the discriminant analysis module.

Precision (also called positive predictive value) is the fraction of relevant instances among the retrieved instances, while Recall (also known as sensitivity) is the fraction of the total amount of relevant instances that were actually retrieved.

F1-Score is a measure of a test's accuracy. It considers both the precision p and the recall r of the test to compute the score: p is the number of correct positive results divided by the number of all positive results returned by the classifier, and r is the number of correct positive results divided by the number of all relevant samples (all samples that should have been identified as positive).

The support is the number of occurrences of each class.

- Quadratic Discriminant Analysis gave us an accuracy of 55.7%

Performance matrix of Quadratic Discriminant Analysis:

```
(array([4, 5, 6, 7, 8]), array([  5, 160, 185,  47,   3], dtyp
e=int64))
[[  0   0   0   0   0   0]
 [  0   1   1   3   0   0]
 [  1   9 107  43   0   0]
 [  0   3  53  98  30   1]
 [  0   0   3  23  17   4]
 [  0   0   0   2   1   0]]
              precision    recall  f1-score   support

           3      0.000     0.000     0.000         1
           4      0.200     0.077     0.111        13
           5      0.669     0.652     0.660       164
           6      0.530     0.580     0.554       169
           7      0.362     0.354     0.358        48
           8      0.000     0.000     0.000         5

    accuracy                          0.557       400
   macro avg      0.293     0.277     0.281       400
weighted avg      0.548     0.557     0.551       400
```

5) K-Nearest Neighbors

k Nearest Neighbor (or kNN) is a supervised machine learning algorithm useful for classification problems. It calculates the distance between the test data and the input and gives the prediction according. In k-NN classification, the output is a class membership. An object is classified by a plurality vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (k is a positive integer, typically small). If k = 1, then the object is simply assigned to the class of that single nearest neighbor.

- K-Nearest Neighbors gave us an accuracy of 59%

Performance matrix of K-Nearest Neighbors:

```
              precision    recall  f1-score   support

           3       0.00      0.00      0.00         1
           4       0.20      0.08      0.11        13
           5       0.64      0.70      0.67       164
           6       0.56      0.59      0.58       169
           7       0.54      0.40      0.46        48
           8       0.00      0.00      0.00         5

    accuracy                           0.59       400
   macro avg       0.32      0.29      0.30       400
weighted avg       0.57      0.59      0.58       400
```

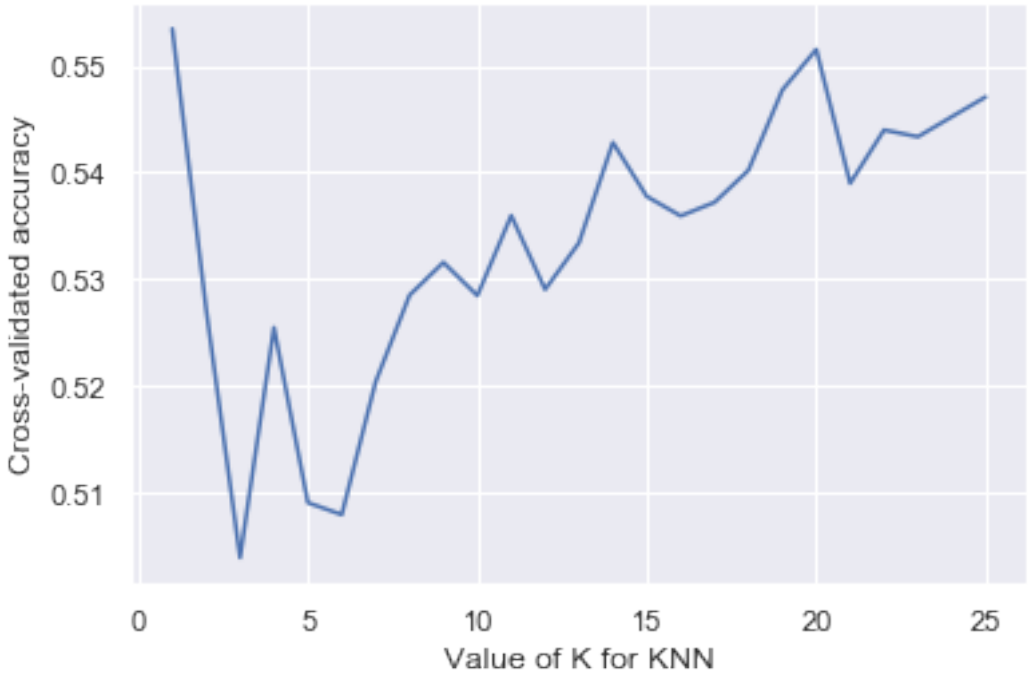


Fig. 4.KNN Cross-Validated accuracy.

6) SVM

Support Vector Machine is a discriminative classifier by a separating hyperplane and supervised learning technique for Machine Learning. Differently from the unsupervised learning algorithms, there is a dataset belongs to different classes(labels). Data is trained with those class labels and then it is predicted with test data set then calculate accuracy how the algorithm predicts test data correctly. In other words, separating dataset into labeled training and test(categorize) dataset with labeled data, it could be better with work with as binary classification. In more detail, SVM uses margin and hyperplane instead of line to separate data into two or more different class. In order to separate classes, it can be drawn many different lines but by choosing best line it is considered that margin should be maximum in between support vectors which are the closest points with different classes. While SVM algorithm is working, it follows two rules which are firstly classify correctly, then increase the margins in hyperplane.

- Support Vector Machine gave us an best accuracy of 66.64%

```
from sklearn.model_selection import GridSearchCV
parameters = [{'C': [1, 10, 100, 1000], 'kernel': ['linear
              {'C': [1, 10, 100, 1000], 'kernel': ['rbf'],
               'gamma': [0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.
grid_search = GridSearchCV(estimator = classifier,
                           param_grid = parameters,
                           scoring = 'accuracy',
                           cv = 10,)
grid_search.fit(x_tr, y_tr)
best_accuracy = grid_search.best_score_
best_parameters = grid_search.best_params_
#here is the best accuracy
best_accuracy
```

```
0.6663886572143453
```

For classifying the wine quality, we have implemented multiple algorithms, namely
1)Logistic Regression classifier
2)Random Forest
3)Decision Tree

4)QDA
5)K-Nearest Neighbors
6)Support Vector Machine

We were able to achieve maximum accuracy using Support Vector Machine of 66.64%. K-Nearest Neighbors gave us an accuracy of 59% ; Quadratic Discriminant Analysis gave us an accuracy of 55.7% ;Decision Tree gave us an accuracy of 57.49% ; Random Forest gave us an accuracy of 61.25% ; Logistic Regression gave us an accuracy of 57.25%.

3 Conclusion

This paper was motivated by the need for research that could improve the understanding of how the quality of the wine is influenced by its different physicochemical present in it. Out of the twelve attributes, the statistically significant attribute that influences the quality of the wine is an essential finding. By regression analysis, we come up with a model that highlights the significant attributes. The result of this analysis will be helpful in production and in quality prediction by studying the impact of those significant attributes in predicting the quality. There is space for further analysis to reveal the more interesting pattern and to employ rigor analytical tools to augment a more sophisticated model.

For this work, it was aimed that the analyzing which psychochemical are more related with wine quality and which approach is good for prediction of wine quality better. During this research, six important machine learning techniques was used;

1)Logistic Regression classifier
2)Random Forest
3)Decision Tree
4)QDA
5)K-Nearest Neighbors
6)Support Vector Machine

From all algorithms, it was obvious that for this dataset, SVM and then Random Forest algorithm gave the best model and accuracy means that those algorithms predict correctly test data. If someone wants to analyze similar data like that it is better to work SVM or Random Forest. They can better predict the quality of red wine. Hence, those algorithms variances are found better with high margin terminology, therefore with multiclass analysis, those algorithms will give the best accuracy.

And, after the analysis of this dataset, some features have more effect to deciding quality of the wine, through the correlation plot, we can see that

- Alcohol is the most important feature to decide the quality of the wine. If the alcohol percentage is high enough, it means that quality of the wine should be better
- Sulphates is another selecting criteria for good wines, with high percentage sulphates wine quality is increasing
- Citric Acid is another selecting criteria, it should be higher to decide more better wine
- Volatile Acidity should be less in the good wine
- Sulfur dioxide is another effect to decreasing wine quality and also it causes headache therefore if there is less sulfur dioxide in wine, it should be selected
- Chlorides value has very less effect to quality of the wine but again it is obvious more value of it causes bad quality of the wine

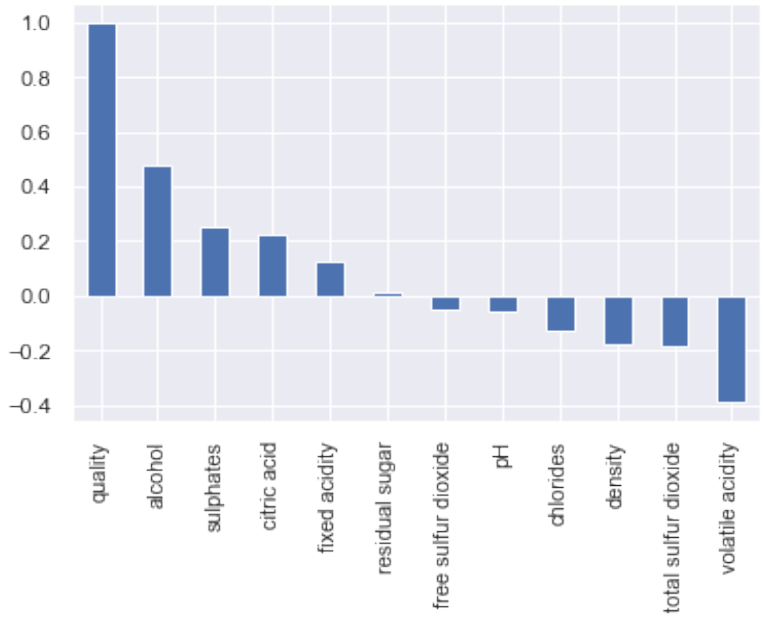


Fig. 5.Correlation plot.

References

[1] R. Gonzalez (2013, May, 8) “Wine tasting is bullshit. Here is why.”[Online] Available: http://io9.gizmodo.com/wine-tasting-isbullshit-heres-why-496098276

[2] F. Brochet, “Chemical Object Representation in the Field of Consciousness” Avilable: http://web.archive.org/web/20070928231853/http://www.academieamorim.com/us/laureat_2001/brochet.pdf

[3] G. Morrot, F. Brochet, D. Dubourdieu, “The Color of Odors” August 2001 Avilable: https://web.stanford.edu/class/linguist62n/morrot01colorofodors.pdf

[4] P. Cortez, A. Cerdeira, F, Almeida, T. Matos, and J. Reis, “Modeling wine preferences by data mining from physicochemical properties,” Decision Support Systems, vol. 47, iss. 4, pp. 547-553, Nov. 2009.

[5] A. Bednarova, D. Brodnjak-Voncina, R. Kranvogl, and T. Jug, “Prediction of wine sensoral quality by routinely measured chemical properties,” Nova Biotechnologica et Chimica, vol. 13, iss. 2, pp. 182196, Feb. 2015.

[6] Yunhui Zeng1, Yingxia Liu1 , Lubin Wu1, Hanjiang Dong1. “Evaluation and Analysis Model of Wine Quality Based on Mathematical Model ISSN 2330-2038 E-ISSN 2330-2046,Jinan University, Zhuhai, China.

[7] Paulo Cortez1, Juliana Teixeira1.“Modeling wine preferences by data mining from physicochemical properties”.

[8] Yesim Er*1 , Ayten Atasoy1. “The Classification of White Wine and Red Wine According to Their Physicochemical Qualities”,ISSN 2147-67992147-6799,3rd September 2016

[9] P. Cortez. (2010, Oct 2). Wine Quality Data Set [Online]. Available: http://archive.ics.uci.edu/ml/datasets/Wine+Quality