# RMIT University Vietnam - SGS Campus

COSC 2789 - Practical Data Science

Assignment II

# Classification of Wines based on their Chemical and Physical Attributes

by

Anh Nguyen

s3616128@rmit.edu.vn

September 13th, 2019

## TABLE OF CONTENTS

## I. Abstract

The primary goal of this report is to classify three types of wine by their chemical and physical properties.

This report uses the physical and chemical properties to see if three types of wine produced by three different winemakers in the same part of Italy can be differentiated solely by these properties. It was found that by limiting the number of properties to be considered to 7 which has a classification error rate of 0.018.

To arrive at this result, two classifiers were used: k Nearest Neighbors (kNN) and decision tree. The three types of wines were reliably differentiated by the chemical and physical properties, and the decision tree classifier performed suitably well with the data.

## II. Introduction

Wine has been a popular alcoholic drink around the world for many years. There is an incredibly diverse selection of wines thanks to its complex chemical makeup. This report uses thirteen chemical and physical properties to see if three types of wine produced by three different winemakers in the same region can be discerned by these properties.

The thirteen properties are malic acid, alcohol, ash, alkalinity of ash, magnesium, total phenols, flavonoids, non-flavonoid phenols, proanthocyanins, color intensity, hue, OD280/OD315 of diluted wine, and proline.

// TODO FINISH THIS

The classification tasks used two different algorithms: K nearest neighbors (kNN) and decision tree. First, the model was trained by decision tree algorithms. Several parameters were adjusted to prevent overfitting in the model.

Next, the kNN classifier was used. In order to reduce the classification error rate, the value of k was chosen by finding the value which performed best in several metrics.

## III. Methodology

This project analyzes a wine dataset public available through the University of California Irvine Machine Learning repository (Dua, D. and Graff, C., 2019)[5]. The data has been cleaned by the source and contains no missing values. The data consists of 178 wine samples, all from the same geographic region in Italy, but from 3 different labels. The objective is to classify which samples belong to which labels based on their chemical attributes.

The dataset has 14 variables with *Label* being the label's identifier, and the others are chemical attributes. The chemical attributes are all numerical and continuous. All variables are listed below:

- Label: wine labels {values: 1,2,3}

- Alcohol: alcoholic content in the wine measured in ABV (alcohol by column) { $0 < R < 100$} (R = real numbers)
- Malic acid: an acid that contributes greatly to the taste of wine ("Malic Acid in Wine")[1] {$R \geq 0$}
- Ash: a collection of substances, which remain after ashing of evaporation residue (Heidger, 2015)[2] {$R \geq 0$}
- Alkalinity of ash: a chemical property of ash {$R \geq 0$}
- Magnesium: a mineral in wine {$R \geq 0$}
- Total phenols: a class of molecules that account for the taste, smell, medical benefits and diversity of wine{$R \geq 0$}
- Flavonoids: a type of phenol that contributes to the taste of wine {$R \geq 0$}
- Non-flavonoid phenols: another type of phenol {$R \geq 0$}
- Proanthocyanins: a type of flavonoids {$R \geq 0$}
- Color intensity: an attribute of color {$R \geq 0$}
- Hue: another attribute of color {$R \geq 0$}
- OD280/OD315 of diluted wine: protein content measurements {$R \geq 0$}
- Proline: an amino acid that is abundant in grapes {$R \geq 0$}

The data was split 3 different ways with random state 20 (this is set this way to create reproducible results):

- 50% for training, 50% for testing
- 60% for training, 40% for testing
- 80% for training, 20% for testing

Each dataset was fitted into DecisionTreeClassifier and KNeighborsClassifier which were provided by the *sklearn* library.

Each model's performance was measured based on 4 metrics:

- Recall Score, F1 Score, Precision Score ( the higher the better)
- Classification Error Rate ( the lower the better)

The kNN model works by associating an unknown wine sample with the samples with the most similar attributes. Two parameters of this classifier were focused on:

- k: number of neighbors {values: from 1 to 10}.
- Weights: weight function used in prediction {values: 'uniform' and 'distance'}
  - 'uniform': all points in a neighborhood are weighted equally
  - 'distance': weights points by the inverse of their distance.

As for the decision tree classifier, it works by classifying the unknown wine sample by making decisions based on all the attributes. Each decision creates sub-nodes to increase the homogeneity of the resulting sub-nodes. To avoid overfitting - a situation when a model has too many parameters relative to the number of observations-, a constrains on the number of attributes used by the tree was placed.

Each model's performance was measured based on 4 metrics:

- Recall Score: the ability of the classifier to find all the positive samples.
- Precision Score: the ability of the classifier not to label as positive a sample that is negative.
- F1 Score: weighted average of recall and precision scores.
- Classification Error Rate: the rate that the model guesses wrong.

High recall, precision, F1 scores, and low error rate indicate a good model.

# IV. Results

1. Attribute Summary Statistics

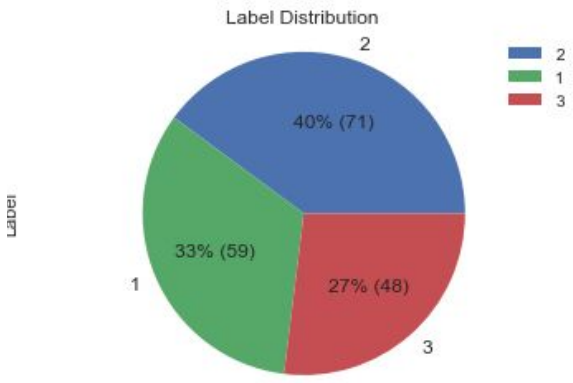



*Figure 1.1*

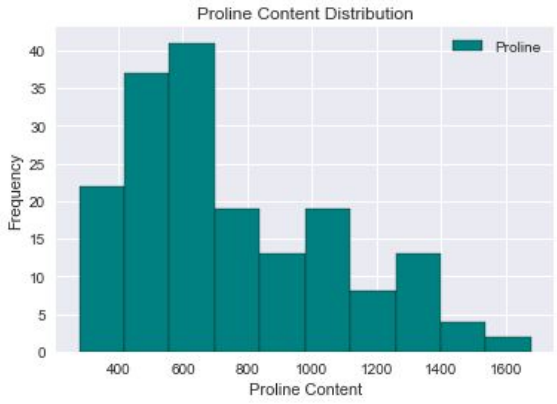



*Figure 1.2*

As mentioned in the introduction, of the 14 attributes in the wine dataset, *Label* is the only categorical (nominal) attribute. Thus, it is appropriate to use a pie chart to summarize this variable as shown in *Figure 1.1*.

The other 13 attributes are numerical attributes that could be summarized by using several different types of chart such as histogram, density chart, boxplot and so on. Figure 1.2 shows the distribution of Proline content. To view all the summary charts, please refer to the Jupyter notebook file that comes with this report.

2. Pair Value Relational Graphs

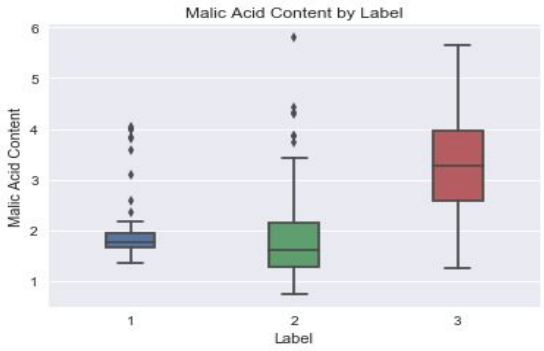



*Figure 2.1*

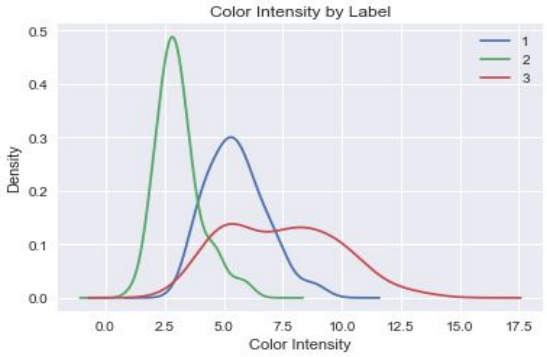



*Figure 2.2*

- Figure 2.1:
  - **Hypothesis:** malic acid contributes greatly to the taste of wine and its concentration depends on the ripeness of the grapes ("Malic Acid in Wine")[1]. Thus, different labels will attempt to differentiate from one another by adjusting the concentration of this chemical.
  - **Results:** different labels have different concentrations of malic acid. Label 3 has the highest range of malic acid. Whereas the other two labels' median concentrations are lower overall.
- Figure 2.2:
  - **Hypothesis:** the color of the wine is dependant on various chemicals. Each label will have its own method of making wine. Thus, each label's wines are different in color even if the difference is not visible to the eyes.
  - **Results:** each label has different peaks in color intensity with label 2 has the sharpest peak. The other two labels' peaks are duller which means their wines' color is less intense.

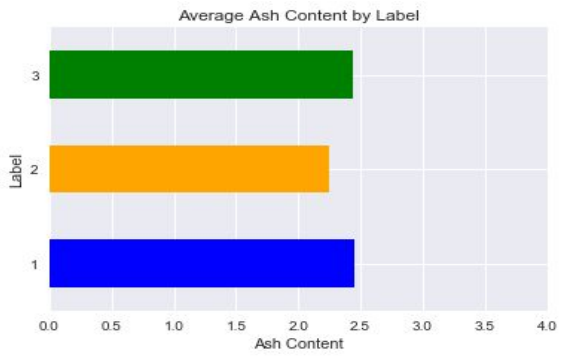



*Figure 2.3*

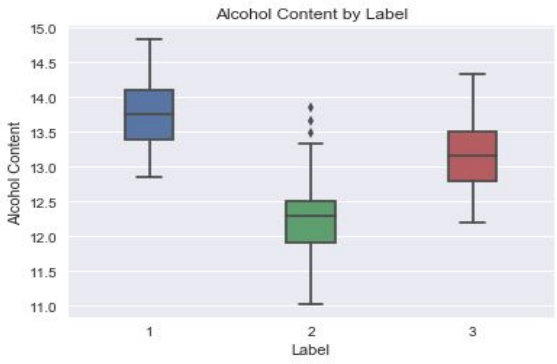



*Figure 2.4*

- Figure 2.3:

- ➢ **Hypothesis:** Normal ash concentration is between 1.3 and 3.5 mg/L (Heidger, 2015)[2]. There is a relation between ash and quality of the wine so labels will try to keep the ash content within range.
- ➢ **Results:** All three labels have normal concentrations of ash.

❖ Figure 2.4:

- ➢ **Hypothesis:** Most wines' alcohol percentage is between 11.5% - 13.5%.
- ➢ **Results:** all three labels' median alcohol percentage is within range.

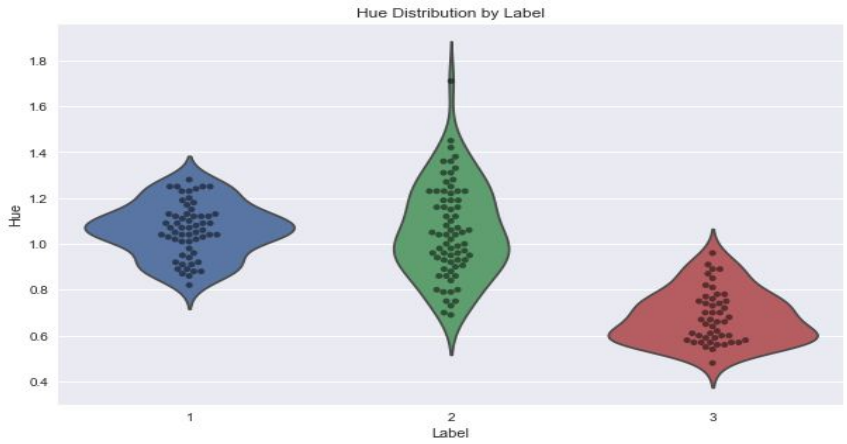



*Figure 2.5*

❖ Figure 2.5:

- ➢ **Hypothesis:** Similar to color intensity, hue - another component that makes up color - is dependant on the chemical makeup of the wine. Each label will have different production methods which lead to a slight difference in hue.
- ➢ **Results:** All three labels have a slight difference in hue. Label 2 has the widest range. Whereas, label 1 wines' hue is more evenly distributed. Label 3 has the lowest hue value.

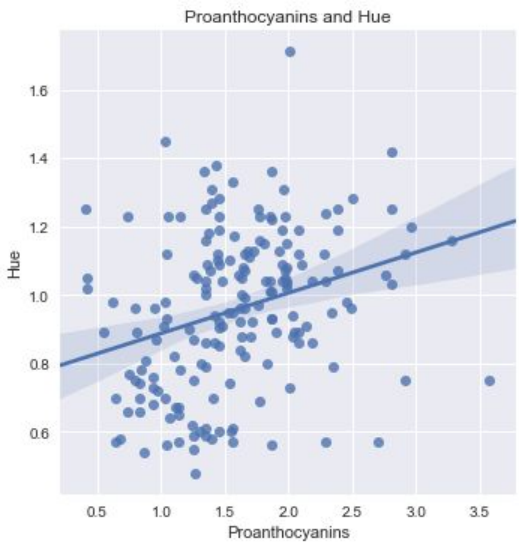



*Figure 2.6*

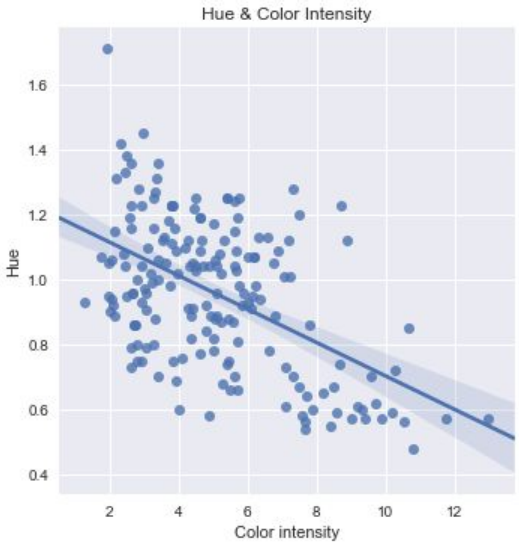
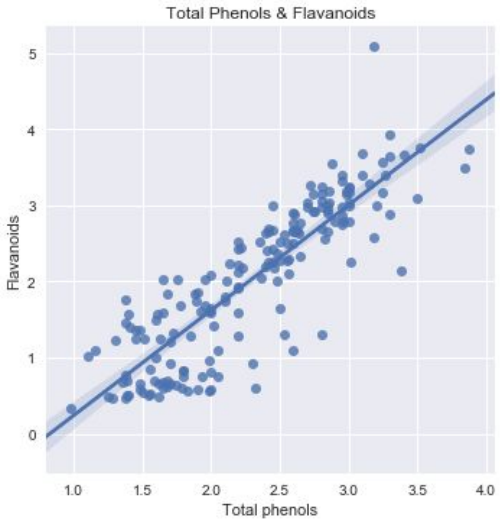



*Figure 2.7*

- Figure 2.6:
  - **Hypothesis:** Anthocyanins are said to be one of the main pigments in red wine**(**Liao, H. , Cai, Y. and Haslam, E,1992)[3]. Thus, it might have a correlation with hue value.
  - **Results:** There is a positive correlation between Proanthocyanins and hue. However, there are too many noises in the graph which indicate either a weak correlation or an inconclusive correlation.
- Figure 2.7:
  - **Hypothesis:** Hue and intensity are two properties of color. Thus, they must have a correlation with one another.
  - **Results:** Despite the noises, both values seem to have a negative correlation with one another.



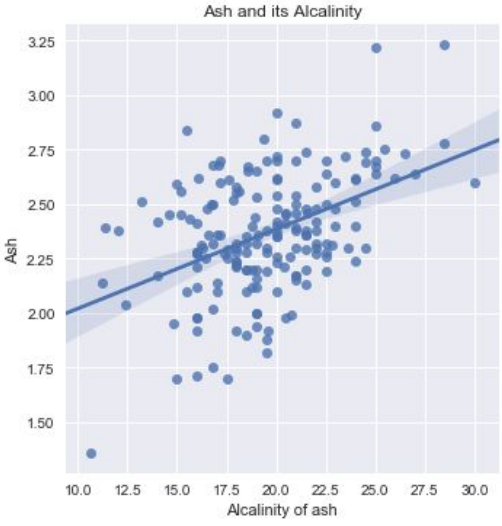



*Figure 2.8* *Figure 2.9*

- Figure 2.8:
  - **Hypothesis:** Flavonoids are phenolic. Thus, total phenols and flavonoids must have a relationship.
  - **Results:** there is a positive relationship between total phenols and flavonoids which confirms the hypothesis.
- Figure 2.9:
  - **Hypothesis:** ash and its Alkalinity must have a correlation with one another.
  - **Results:** both values have a positive correlation to one another as indicated in Figure 2.9.

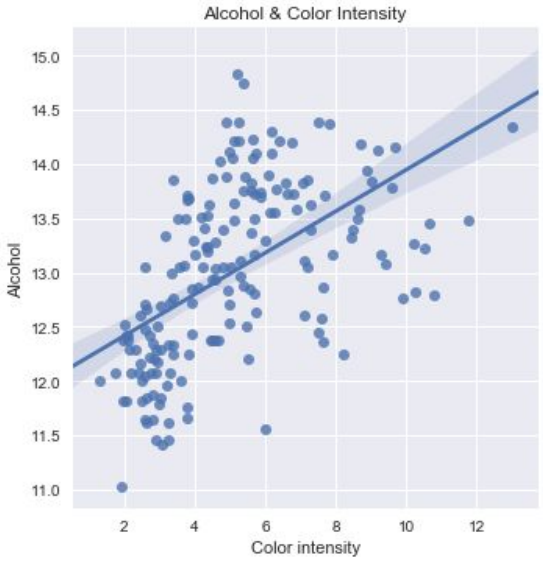



*Figure 2.10*

- Figure 2.10:
  - **Hypothesis:** red wine tends to have a higher alcohol percentage than white wine (Suckling, 2019)[4]. Thus, there might be a correlation between alcohol and color intensity.
  - **Results:** there is a positive correlation between alcohol and color intensity.

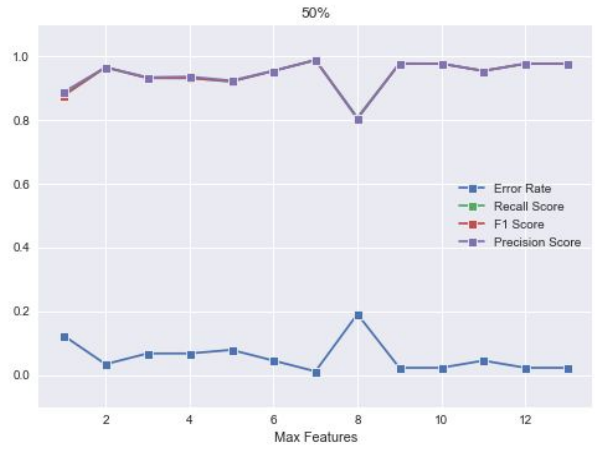
3. Models

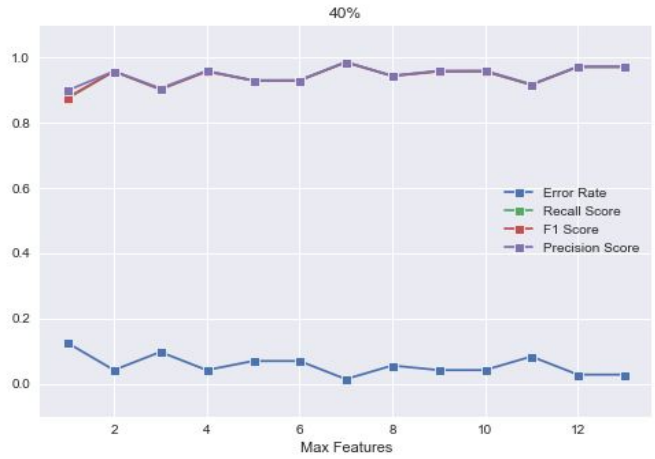
The data is prepared in 3 different ways:

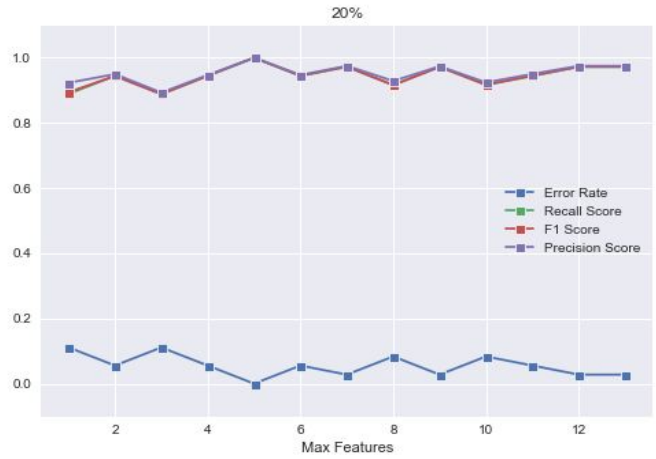
- 50% for training, 50% for testing
- 60% for training, 40% for testing
- 80% for training, 20% for testing

a. Decision Tree Model

**Decision tree classifier default settings with max_features = [1,3,5,7,9,11,13]**



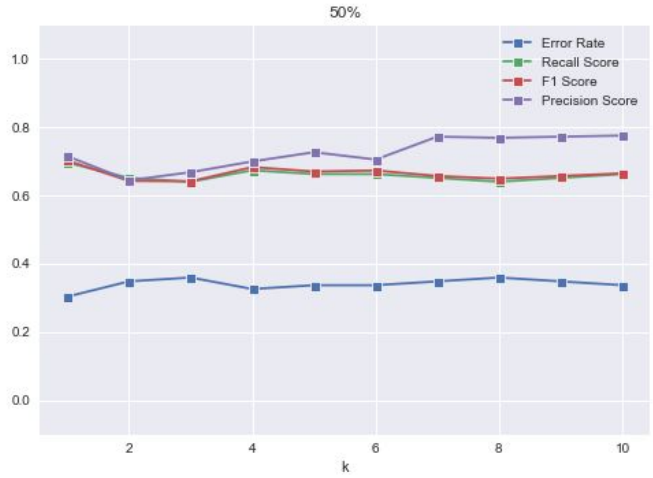



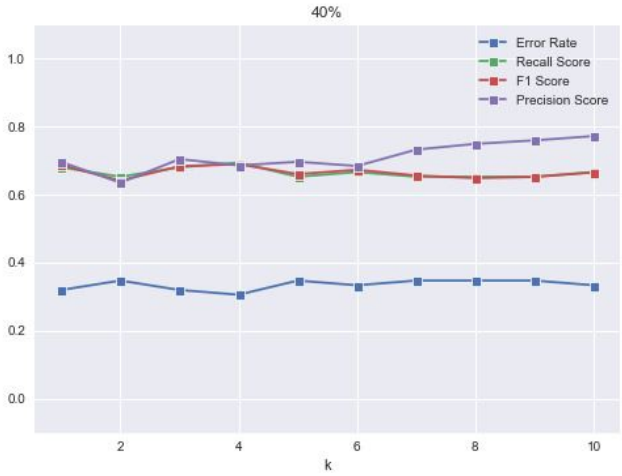

max_features=7 provides the best scores overall among three tests. The classification error rates at max_features=7 for 20%, 40%, 50% testing data are 0.028, 0.014, 0.011 respectively.

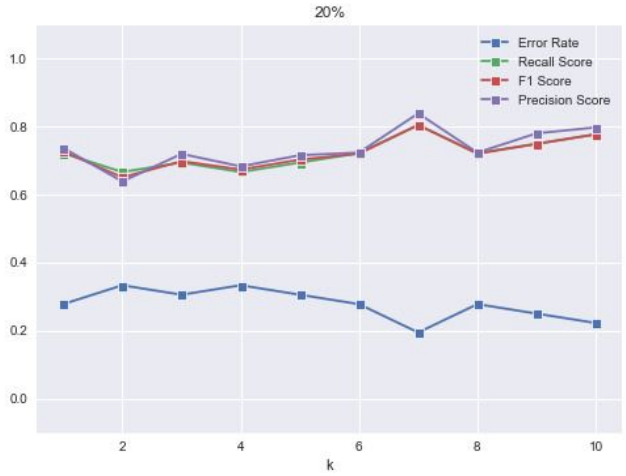
b. K Nearest Neighbors

**Default kNN classifier parameters**

k Nearest Neighbours classifier default settings with weights equal to ‘uniform’.



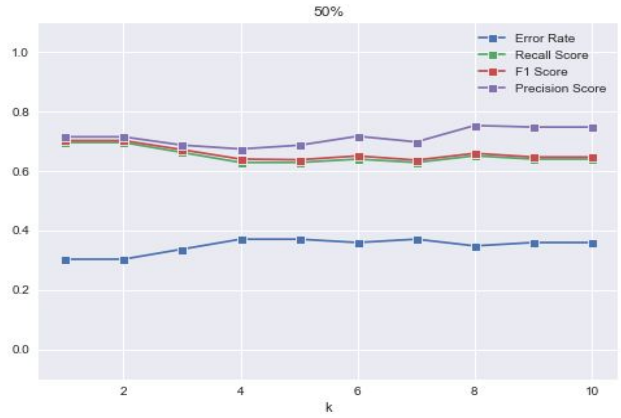



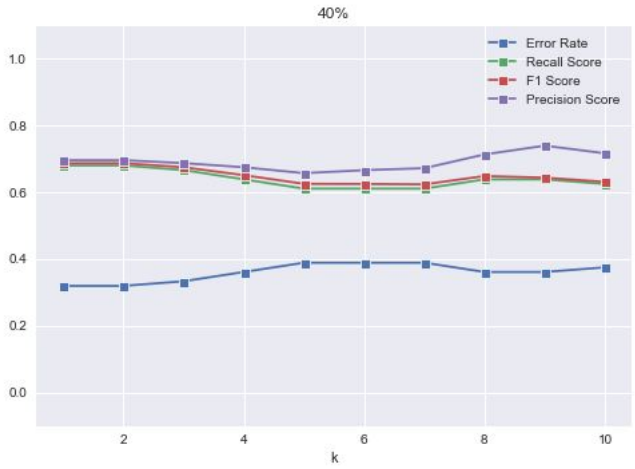



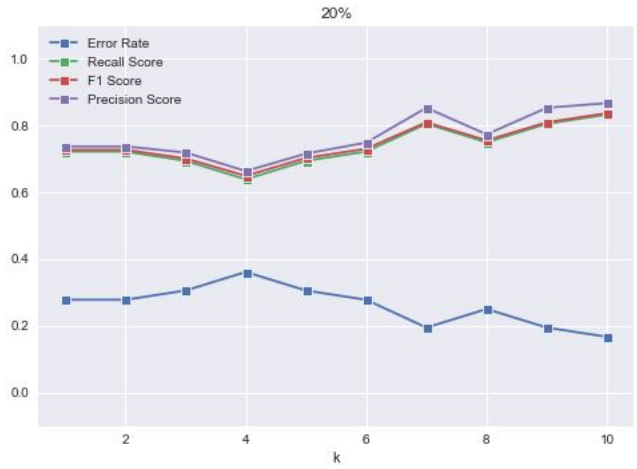

k=7 has the highest overall F1, recall, precision scores in all test cases. The classification error rates at k=7 for 20%, 40%, 50% testing data are 0.19, 0.35, 0.35 respectively.

**K Neighbors Weighted by Distance**

K Nearest Neighbors with weights equal to ‘distance’.




For 'distance' cases, the best k is k=9. The classification error rates at k=9 for 20%, 40%, 50% testing data are 0.19, 0.39, 0.37 respectively.

## V. Discussion

1. Data Exploration Tasks

// TODO FINISH THIS

2. Modeling Tasks

// TODO FINISH THIS

a. Decision Tree

b. k Nearest Neighbors

The uniformly weighted kNN classifier had an average error rate among three test sets of

## VI. Conclusion

The models were able to reliably differentiate the three wine types with a relatively low classification error rate from the chemical and physical attributes of the sample. The best of k for uniformly weighted kNN classifier is 7 which had an average error rate of 0.29. Surprisingly, with a different best k value (9), the average error rate for distance weighted kNN classifier is just slightly higher (0.32). If one must use the kNN classifier, change the *weights* parameter to 'uniform' and set k = 7 in order to produce the best results.

Compared to the classification error rate of the kNN classifier, the decision tree classifier has a much lower average error rate. At *max_features* = 7, among 3 test sets, the average error rate is as low as 0.018. Based on that fact, I recommend using the decision tree classifier in order to differentiate future wine samples.

## VII. References

[1] (n.d.). Malic Acid in Wine. Retrieved from http://www.calwineries.com/learn/wine-chemistry/wine-acids/malic-acid

[2] Heidger, M. (2015, November 23). Ash (gravimetric, calculated from minerals, the alkalinity of ash). Retrieved from https://www.institut-heidger.de/en/asche-summe-aller-mineralstoffe/

[3] Liao, H. , Cai, Y. and Haslam, E. (1992), Polyphenol interactions. Anthocyanins: Co‐pigmentation and color changes in red wines. J. Sci. Food Agric., 59: 299-305. doi:10.1002/jsfa.2740590305

[4] Suckling, J. (2019, April 30). Learn About Alcohol Content in Wine: Highest to Lowest ABV Wines - 2019. Retrieved from https://www.masterclass.com/articles/learn-about-alcohol-content-in-wine-highest-to-lowest-abv-wines#does-white-wine-or-red-wine-have-higher-alcohol-content

[5] Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.