CSE 5160 - 60

Erik Trujillo

Mario Valdez

Bat Erdene Narangerel

Machine Learning Group Project, Spring 2022

Group 8 Project Report

Data Analysis on Wine Data Set (classes version)

Background:

Wine is a popular alcoholic drink made in many different countries and also sold in many different places. Wine is typically made from grapes that have been fermented. Fermentation is the process of turning the sugars that can be found in yeast into ethanol and carbon dioxide. There are many different variants of wine such as red, white, amber, and sparkling wine. One of the fascinating facts about wine is that one of its oldest producers, Italy, happens to be the largest producer of wine today. As one can see, wine is one of the most popular alcoholic drinks in the world and despite its risks, many people around the globe enjoy it. However, the origin processes of creating wine itself can vary from wine type to wine type. In order to truly decipher the origin of different types of wine, wine experts use a variety of chemical analysis methods to determine the origin of the wine, which is of great economic interest to countries that import or export wine. In this particular project, our group is interested in using classification-related Machine Learning methods in conjunction with the dataset's chemical analysis to determine the origin of wines grown in the same region in Italy, but derived from three different cultivars.

Background Of Original Study:

If we wish to find the answer to this question, we decided the best course of action was to have a look at a Wine data set that is provided in the UCI machine learning repository. The data set comes from a paper called "*An Extendible Package for Data Exploration, Classification, and Correlation. Institute Institute of Pharmaceutical and Food Analysis and Technologies*" from Professor Forina and Brigata Salerno at the University of Genoa in the nation of Italy. This dataset is basically the results of a chemical analysis of wine grown in the same region in Italy but derived from 3 different cultivars (type of plant reproduces for specific traits). An analysis was done to determine the quantities of 13 constituents/attributes found in each of the three types of wine. The list of the 13 attributes will be provided below, and a significant side note is that the **first attribute is the class identifier. Moreover, all the attributes are continuous** and some of the attributes include the physical characteristics of the wine itself such as magnesium content, ash, flavanoids, the specific alcalinity of the ash, etc. were all collected.

Research Question:

With the following dataset above, what attributes are more commonly found in one wine type over the other? If enough clues are given with attributes, are we able to decipher what type of wine is respective to their class identifier? This dataset also has an emphasis on the "three types of wine" (which are in the 1st attribute class identifier (1-3)). What training method out of the models we used is best suited for the Wine classification dataset? For training methods that had parameters, what was the best optimal value?

Potential Solution:

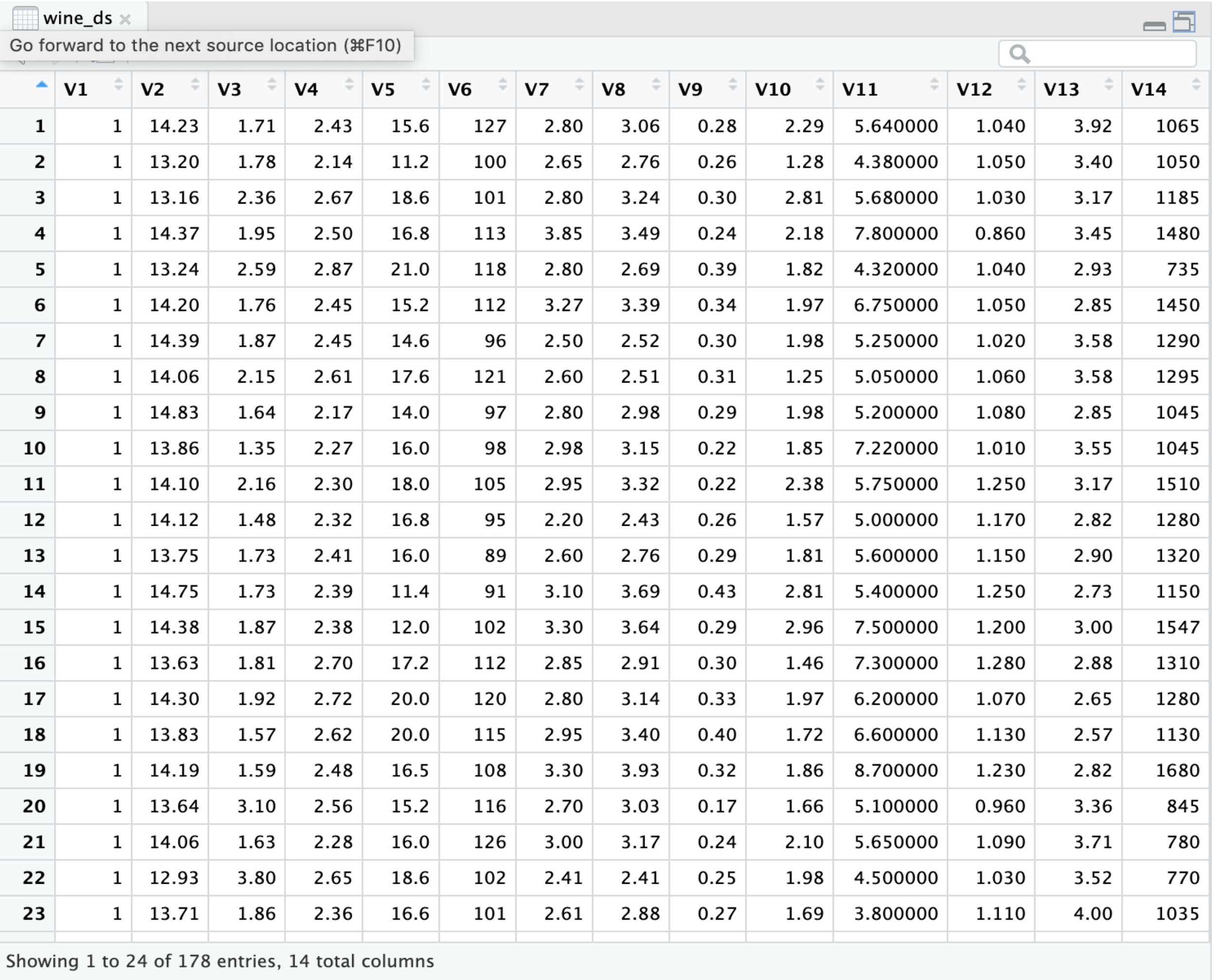
Something noteworthy is that some attributes are less prevalent in other wine types than others. For example, the median proline quantity is the smallest in class 2, while the highest is in class 1. The importance here is to use two or more machine learning techniques to find attributes that can help increase the chance of pinpointing each of the three wine types by how much quantity of constituents such as alcohol, magnesium, proline, ash, flavonoids are in either wine 1, wine 2, or wine 3. Due to the variation in 13 different quantities across all three wine types, the attributes that give us a high prediction rate may change from class identifier to class identifier. The available attributes are all either numeric or int, and thus the range is very much in the positives as “negative X quantities of alcohol” or “negative X quantities of proline” doesn’t really make much sense in the real world.

DATASET ATTRIBUTES TABLE

Attribute	Description	Domain	Note
class	There are 3 class identifiers. 1, 2, 3.	integer	Classes are referred to a specific type of wine that is not listed.
alcohol	the alcohol quantity in each wine	numeric	Range: Min: 11.03% || Max: 14.83%
malic_acid	the wine’s malic acid content	numeric	Range: Min: 0.74% || Max: 5.8%
ash	the wine’s ash quantity	numeric	Range: Min: 1.36% || Max: 3.23%
alcalinity_of_ash	the alcalinity of ash quantity in wine	numeric	Range: Min: 10.6% || Max: 30.00%
magnesium	the wine’s magnesium quantity	integer	Range: Min: 70mg || Max: 162mg
total_phenols	total phenols in the wine	numeric	Range: Min: 0.98% || Max: 3.88%
flavanoids	amount of flavanoids in the wine	numeric	Range: Min: 0.34% || Max: 5.08%
nonflavanoid_ phenols	amount of nonflavanoid phenols in wine	numeric	Range: Min: 0.13% || Max: 0.66%
proanthocyanins	proanthocyanins quantity in wine	numeric	Range: Min: 0.41% || Max: 3.58%
color_intensity	color intensity of	numeric	Range: Min: 1.28%(super mild) ||

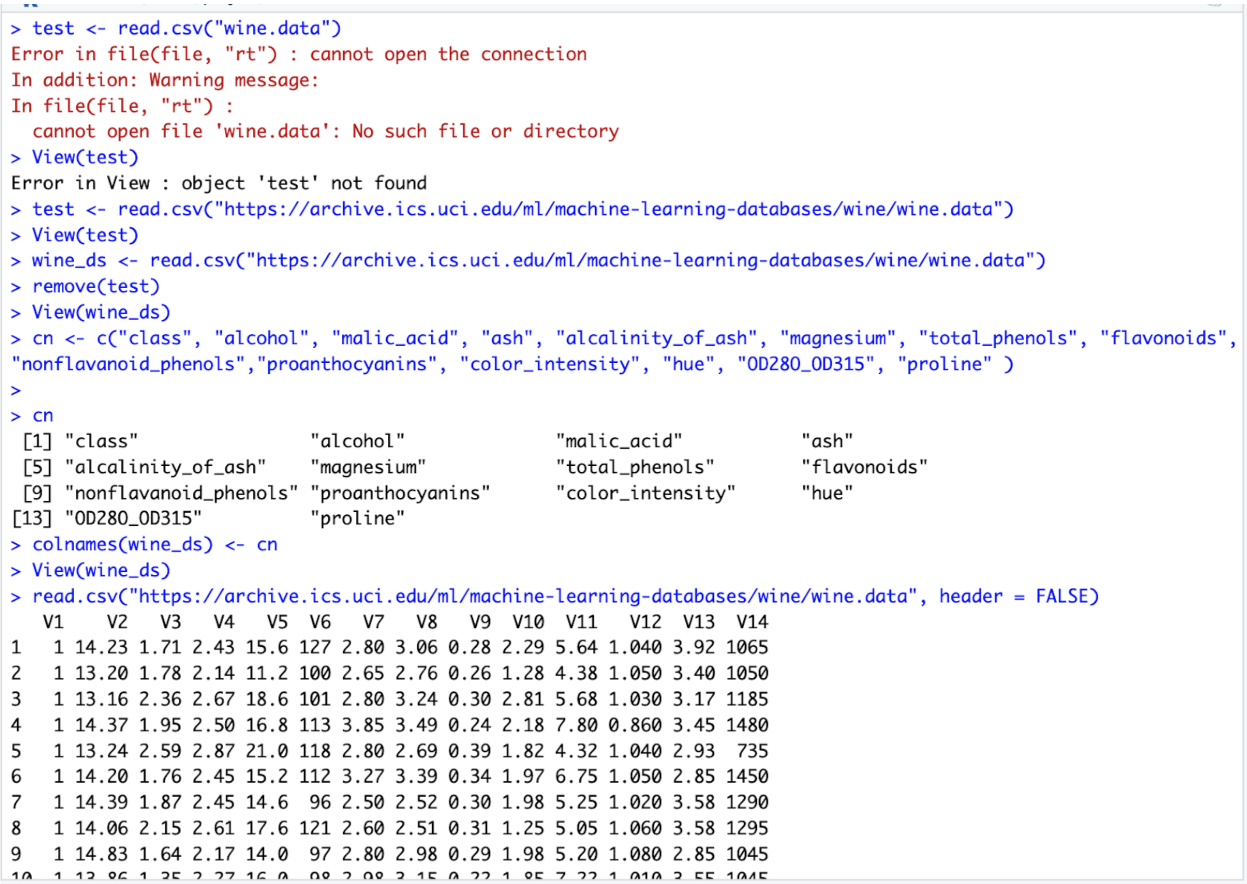
	the wine		Max: 13.00% (super intense)
hue	Hue of the wine (lowest- palest, highest - deepest)	numeric	Range: Min: 0.48% (lowest) || Max: 1.71% (deepest)
OD28O_OD315 (important note: the zero in “OD280” is an O in r code cause case sensitivity)	OD280/OD315 of diluted wines (protein concentration in the wine)	numeric	Range: Min: 1.27% || Max: 4.00%
proline	proline content of the wine	integer	Range: Min: 278mg || Max: 1680mg

Prologue to Data Analysis



	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14
1	1	14.23	1.71	2.43	15.6	127	2.80	3.06	0.28	2.29	5.640000	1.040	3.92	1065
2	1	13.20	1.78	2.14	11.2	100	2.65	2.76	0.26	1.28	4.380000	1.050	3.40	1050
3	1	13.16	2.36	2.67	18.6	101	2.80	3.24	0.30	2.81	5.680000	1.030	3.17	1185
4	1	14.37	1.95	2.50	16.8	113	3.85	3.49	0.24	2.18	7.800000	0.860	3.45	1480
5	1	13.24	2.59	2.87	21.0	118	2.80	2.69	0.39	1.82	4.320000	1.040	2.93	735
6	1	14.20	1.76	2.45	15.2	112	3.27	3.39	0.34	1.97	6.750000	1.050	2.85	1450
7	1	14.39	1.87	2.45	14.6	96	2.50	2.52	0.30	1.98	5.250000	1.020	3.58	1290
8	1	14.06	2.15	2.61	17.6	121	2.60	2.51	0.31	1.25	5.050000	1.060	3.58	1295
9	1	14.83	1.64	2.17	14.0	97	2.80	2.98	0.29	1.98	5.200000	1.080	2.85	1045
10	1	13.86	1.35	2.27	16.0	98	2.98	3.15	0.22	1.85	7.220000	1.010	3.55	1045
11	1	14.10	2.16	2.30	18.0	105	2.95	3.32	0.22	2.38	5.750000	1.250	3.17	1510
12	1	14.12	1.48	2.32	16.8	95	2.20	2.43	0.26	1.57	5.000000	1.170	2.82	1280
13	1	13.75	1.73	2.41	16.0	89	2.60	2.76	0.29	1.81	5.600000	1.150	2.90	1320
14	1	14.75	1.73	2.39	11.4	91	3.10	3.69	0.43	2.81	5.400000	1.250	2.73	1150
15	1	14.38	1.87	2.38	12.0	102	3.30	3.64	0.29	2.96	7.500000	1.200	3.00	1547
16	1	13.63	1.81	2.70	17.2	112	2.85	2.91	0.30	1.46	7.300000	1.280	2.88	1310
17	1	14.30	1.92	2.72	20.0	120	2.80	3.14	0.33	1.97	6.200000	1.070	2.65	1280
18	1	13.83	1.57	2.62	20.0	115	2.95	3.40	0.40	1.72	6.600000	1.130	2.57	1130
19	1	14.19	1.59	2.48	16.5	108	3.30	3.93	0.32	1.86	8.700000	1.230	2.82	1680
20	1	13.64	3.10	2.56	15.2	116	2.70	3.03	0.17	1.66	5.100000	0.960	3.36	845
21	1	14.06	1.63	2.28	16.0	126	3.00	3.17	0.24	2.10	5.650000	1.090	3.71	780
22	1	12.93	3.80	2.65	18.6	102	2.41	2.41	0.25	1.98	4.500000	1.030	3.52	770
23	1	13.71	1.86	2.36	16.6	101	2.61	2.88	0.27	1.69	3.800000	1.110	4.00	1035

Showing 1 to 24 of 178 entries, 14 total columns

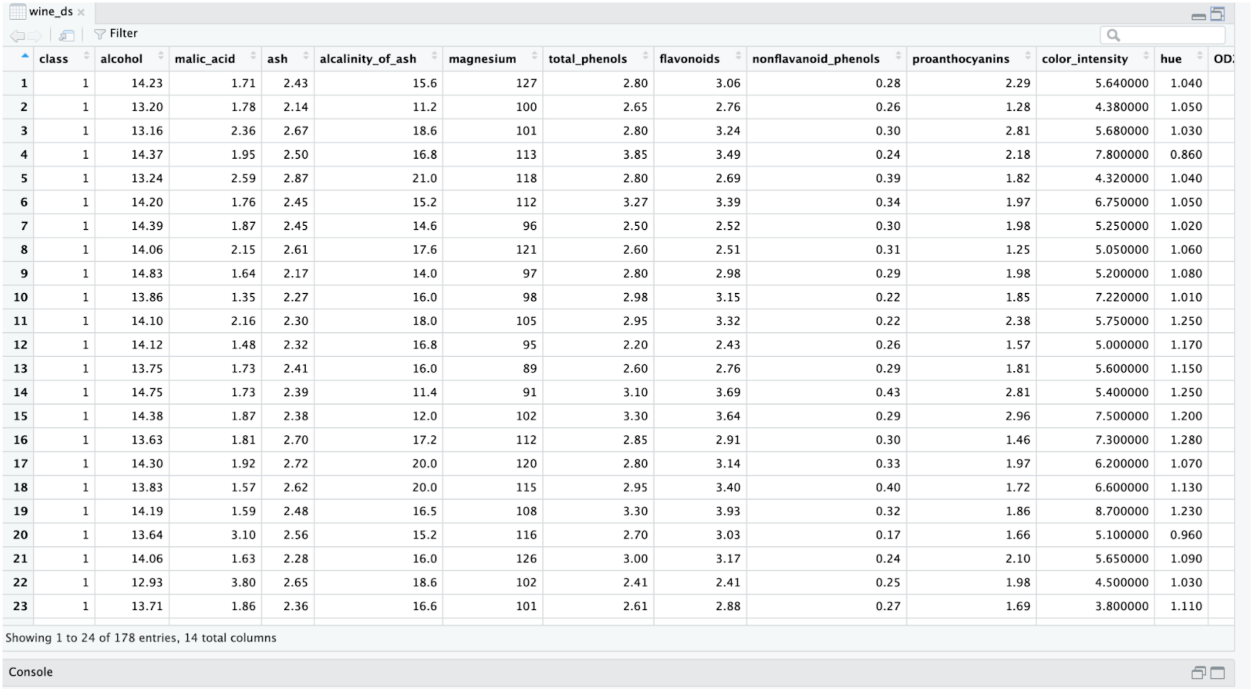


```
> test <- read.csv("wine.data")
Error in file(file, "rt") : cannot open the connection
In addition: Warning message:
In file(file, "rt") :
  cannot open file 'wine.data': No such file or directory
> View(test)
Error in View : object 'test' not found
> test <- read.csv("https://archive.ics.uci.edu/ml/machine-learning-databases/wine/wine.data")
> View(test)
> wine_ds <- read.csv("https://archive.ics.uci.edu/ml/machine-learning-databases/wine/wine.data")
> remove(test)
> View(wine_ds)
> cn <- c("class", "alcohol", "malic_acid", "ash", "alcalinity_of_ash", "magnesium", "total_phenols", "flavonoids",
"nonflavanoid_phenols","proanthocyanins", "color_intensity", "hue", "OD280_OD315", "proline" )
>
> cn
 [1] "class"                "alcohol"              "malic_acid"           "ash"
 [5] "alcalinity_of_ash"    "magnesium"            "total_phenols"        "flavonoids"
 [9] "nonflavanoid_phenols" "proanthocyanins"      "color_intensity"      "hue"
[13] "OD280_OD315"          "proline"
> colnames(wine_ds) <- cn
> View(wine_ds)
> read.csv("https://archive.ics.uci.edu/ml/machine-learning-databases/wine/wine.data", header = FALSE)
   V1    V2   V3   V4   V5  V6   V7   V8   V9  V10  V11   V12  V13  V14
1   1 14.23 1.71 2.43 15.6 127 2.80 3.06 0.28 2.29 5.64 1.040 3.92 1065
2   1 13.20 1.78 2.14 11.2 100 2.65 2.76 0.26 1.28 4.38 1.050 3.40 1050
3   1 13.16 2.36 2.67 18.6 101 2.80 3.24 0.30 2.81 5.68 1.030 3.17 1185
4   1 14.37 1.95 2.50 16.8 113 3.85 3.49 0.24 2.18 7.80 0.860 3.45 1480
5   1 13.24 2.59 2.87 21.0 118 2.80 2.69 0.39 1.82 4.32 1.040 2.93  735
6   1 14.20 1.76 2.45 15.2 112 3.27 3.39 0.34 1.97 6.75 1.050 2.85 1450
7   1 14.39 1.87 2.45 14.6  96 2.50 2.52 0.30 1.98 5.25 1.020 3.58 1290
8   1 14.06 2.15 2.61 17.6 121 2.60 2.51 0.31 1.25 5.05 1.060 3.58 1295
9   1 14.83 1.64 2.17 14.0  97 2.80 2.98 0.29 1.98 5.20 1.080 2.85 1045
10  1 13.86 1.35 2.27 16.0  98 2.98 3.15 0.22 1.85 7.22 1.010 3.55 1045
```

To start off our official data analysis, we need to give some context and background into what we are going to do first. For starters, we always start with wine_ds <- read.csv("https://archive.ics.uci.edu/ml/machine-learning-databases/wine/wine.data", header = FALSE).

Due to odd circumstances, we must include the header at the end to equal FALSE since for whatever reason leaving it out only gives us 177 entries instead of the desired 178 entries. This is important as the main numbers given to us by the UCI repository are “178 instances” and “13 attributes”. An immensely important side note is that the 1st attribute is the class identifier which is shown in column 1. It will go from one to three and is unique in its properties.

To recap, there are 178 instances, 13 attributes (columns), and the current UCI repository has attributes listed as V1, V2, V3, V4, … all the way to V14. Although the correct number of attributes is 13, it says 14 total columns due to the fact that the 1st attribute is the class identifier (1-3). The next step is to give it the proper names of the actual 13 attributes listed in the repository (see data attributes table above).

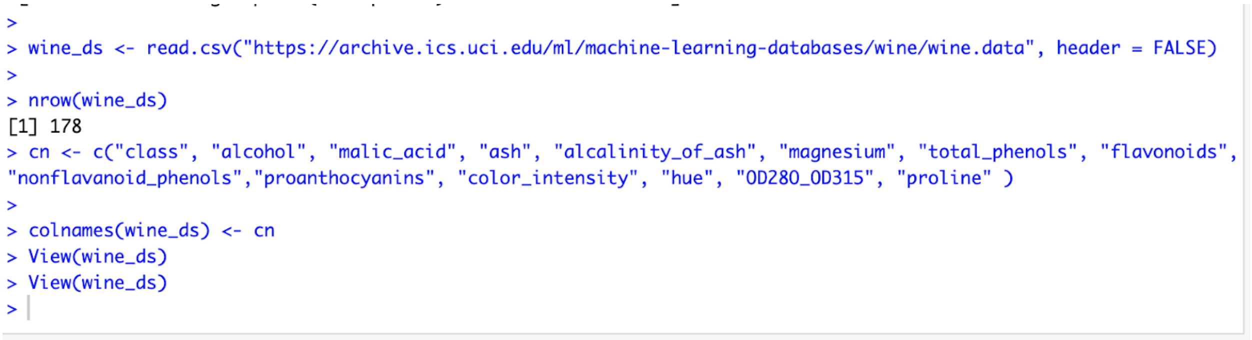


	class	alcohol	malic_acid	ash	alcalinity_of_ash	magnesium	total_phenols	flavonoids	nonflavanoid_phenols	proanthocyanins	color_intensity	hue	OD
1	1	14.23	1.71	2.43	15.6	127	2.80	3.06	0.28	2.29	5.640000	1.040	
2	1	13.20	1.78	2.14	11.2	100	2.65	2.76	0.26	1.28	4.380000	1.050	
3	1	13.16	2.36	2.67	18.6	101	2.80	3.24	0.30	2.81	5.680000	1.030	
4	1	14.37	1.95	2.50	16.8	113	3.85	3.49	0.24	2.18	7.800000	0.860	
5	1	13.24	2.59	2.87	21.0	118	2.80	2.69	0.39	1.82	4.320000	1.040	
6	1	14.20	1.76	2.45	15.2	112	3.27	3.39	0.34	1.97	6.750000	1.050	
7	1	14.39	1.87	2.45	14.6	96	2.50	2.52	0.30	1.98	5.250000	1.020	
8	1	14.06	2.15	2.61	17.6	121	2.60	2.51	0.31	1.25	5.050000	1.060	
9	1	14.83	1.64	2.17	14.0	97	2.80	2.98	0.29	1.98	5.200000	1.080	
10	1	13.86	1.35	2.27	16.0	98	2.98	3.15	0.22	1.85	7.220000	1.010	
11	1	14.10	2.16	2.30	18.0	105	2.95	3.32	0.22	2.38	5.750000	1.250	
12	1	14.12	1.48	2.32	16.8	95	2.20	2.43	0.26	1.57	5.000000	1.170	
13	1	13.75	1.73	2.41	16.0	89	2.60	2.76	0.29	1.81	5.600000	1.150	
14	1	14.75	1.73	2.39	11.4	91	3.10	3.69	0.43	2.81	5.400000	1.250	
15	1	14.38	1.87	2.38	12.0	102	3.30	3.64	0.29	2.96	7.500000	1.200	
16	1	13.63	1.81	2.70	17.2	112	2.85	2.91	0.30	1.46	7.300000	1.280	
17	1	14.30	1.92	2.72	20.0	120	2.80	3.14	0.33	1.97	6.200000	1.070	
18	1	13.83	1.57	2.62	20.0	115	2.95	3.40	0.40	1.72	6.600000	1.130	
19	1	14.19	1.59	2.48	16.5	108	3.30	3.93	0.32	1.86	8.700000	1.230	
20	1	13.64	3.10	2.56	15.2	116	2.70	3.03	0.17	1.66	5.100000	0.960	
21	1	14.06	1.63	2.28	16.0	126	3.00	3.17	0.24	2.10	5.650000	1.090	
22	1	12.93	3.80	2.65	18.6	102	2.41	2.41	0.25	1.98	4.500000	1.030	
23	1	13.71	1.86	2.36	16.6	101	2.61	2.88	0.27	1.69	3.800000	1.110	

Showing 1 to 24 of 178 entries, 14 total columns

After the proper corrections to the R code, we now have 178 instances, and 13 attributes, and the 13 attributes now have clearer names compared to V1, V2, etc. V1 was renamed class, V2 is now alcohol, V3 is now malic acid, and so on and so forth. This was one of the most important steps when starting off in Rstudio.

Here's how the correction was made:

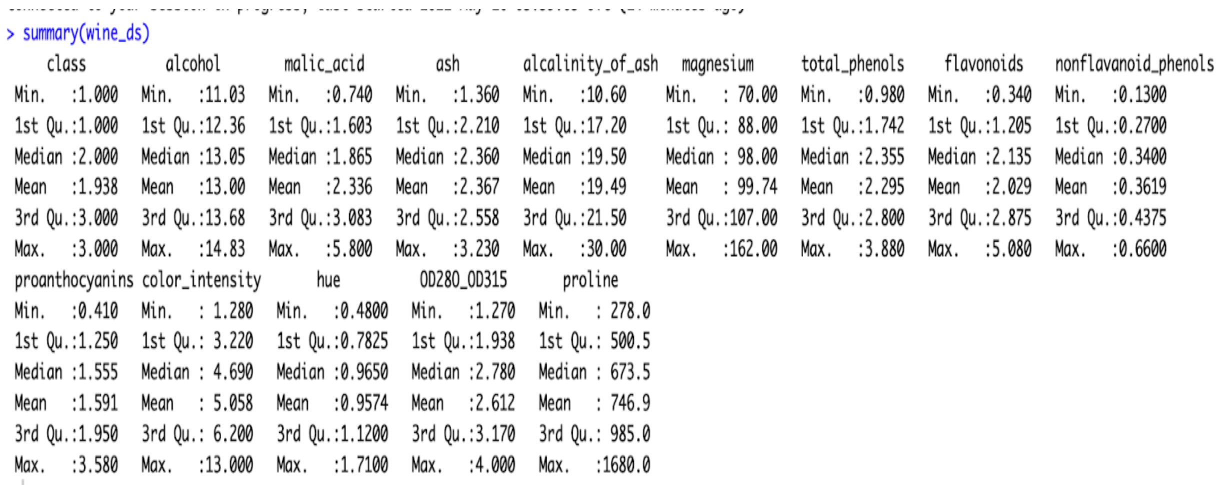


```
>
> wine_ds <- read.csv("https://archive.ics.uci.edu/ml/machine-learning-databases/wine/wine.data", header = FALSE)
>
> nrow(wine_ds)
[1] 178
> cn <- c("class", "alcohol", "malic_acid", "ash", "alcalinity_of_ash", "magnesium", "total_phenols", "flavonoids",
"nonflavanoid_phenols","proanthocyanins", "color_intensity", "hue", "OD280_OD315", "proline" )
>
> colnames(wine_ds) <- cn
> View(wine_ds)
> View(wine_ds)
>
```

With those minor fixes and details out of the way, we can now begin our official data analysis of the Wine dataset. To reiterate, the abstract problem is based on one key concept:

Using chemical analysis to determine the origin of wines….

>summary(wine_ds) / #Summary of Wine dataset



```
> summary(wine_ds)
     class          alcohol        malic_acid         ash        alcalinity_of_ash   magnesium      total_phenols     flavonoids    nonflavanoid_phenols
 Min.   :1.000   Min.   :11.03   Min.   :0.740   Min.   :1.360   Min.   :10.60     Min.   : 70.00   Min.   :0.980   Min.   :0.340   Min.   :0.1300
 1st Qu.:1.000   1st Qu.:12.36   1st Qu.:1.603   1st Qu.:2.210   1st Qu.:17.20     1st Qu.: 88.00   1st Qu.:1.742   1st Qu.:1.205   1st Qu.:0.2700
 Median :2.000   Median :13.05   Median :1.865   Median :2.360   Median :19.50     Median : 98.00   Median :2.355   Median :2.135   Median :0.3400
 Mean   :1.938   Mean   :13.00   Mean   :2.336   Mean   :2.367   Mean   :19.49     Mean   : 99.74   Mean   :2.295   Mean   :2.029   Mean   :0.3619
 3rd Qu.:3.000   3rd Qu.:13.68   3rd Qu.:3.083   3rd Qu.:2.558   3rd Qu.:21.50     3rd Qu.:107.00   3rd Qu.:2.800   3rd Qu.:2.875   3rd Qu.:0.4375
 Max.   :3.000   Max.   :14.83   Max.   :5.800   Max.   :3.230   Max.   :30.00     Max.   :162.00   Max.   :3.880   Max.   :5.080   Max.   :0.6600
 proanthocyanins color_intensity       hue            OD280_OD315       proline
 Min.   :0.410   Min.   : 1.280   Min.   :0.4800   Min.   :1.270   Min.   : 278.0
 1st Qu.:1.250   1st Qu.: 3.220   1st Qu.:0.7825   1st Qu.:1.938   1st Qu.: 500.5
 Median :1.555   Median : 4.690   Median :0.9650   Median :2.780   Median : 673.5
 Mean   :1.591   Mean   : 5.058   Mean   :0.9574   Mean   :2.612   Mean   : 746.9
 3rd Qu.:1.950   3rd Qu.: 6.200   3rd Qu.:1.1200   3rd Qu.:3.170   3rd Qu.: 985.0
 Max.   :3.580   Max.   :13.000   Max.   :1.7100   Max.   :4.000   Max.   :1680.0
```

Brief description: This is a summary of all the wine classification data, which gives us minimum, maximum, mean, median, etc. It has produced result summaries which are derived from the results of certain model fitting functions and will be critical when it comes time to finally begin our R code implementation

Data Analysis Methods:

Approach:

The approach to this Wine data set is of utmost importance and will take up the majority of our project report, presentation, and R language analysis. According to the UCI machine learning repository, this particular Wine dataset is only suitable for classification. For classification, we opted for numerical classification using the values in the attribute tables and focusing on the quantities of each constituent in each of the three types of wine. Models are trained for the dataset, subsets, class 1, class 2, and class 3. The dataset originally had 30 variables, but the UCI author claims to have lost them without elaboration. Therefore, the dataset only gives us thirteen to work with. No other supervised or unsupervised learning concept is mentioned other than classification, so therefore the bulk of our code will take a classification approach to this dataset and problem. For classification, we have three class identifiers for this particular dataset, and the abstract idea is to use chemical analysis to determine the origin of wines. More specifically, it tells us in wine. names (dataset description highlighted in yellow on the UCI page) that *"The analysis determined the quantities of 13 constituents found in each of the three types of wines"*

(UCI Machine Learning Repository). The 13 constituents obviously refer to the 13 attributes which are alcohol, malic acid, ash, hue, color intensity, etc. To conclude, an analysis workflow was created for classification to standardize the process and assist with rescaling.

Sampling:

After the prologue to data analysis and extensive research of classification, we can now begin our sampling. We utilized the concepts of cross-validation, dimensional analysis, and accuracy, to use as our main sampling approach. We also used Fisher's linear discriminant since we are about to see LDA and QDA used extensively in the project and will be elaborated on in the later sections. This sampling methodology was used for all the model training that took place during this group project as the UCI repository only lists classification and lacks other models such as those relating to regression. The main K-fold configuration was using K = 5 due to the limited class identifiers.

Data Preprocessing:

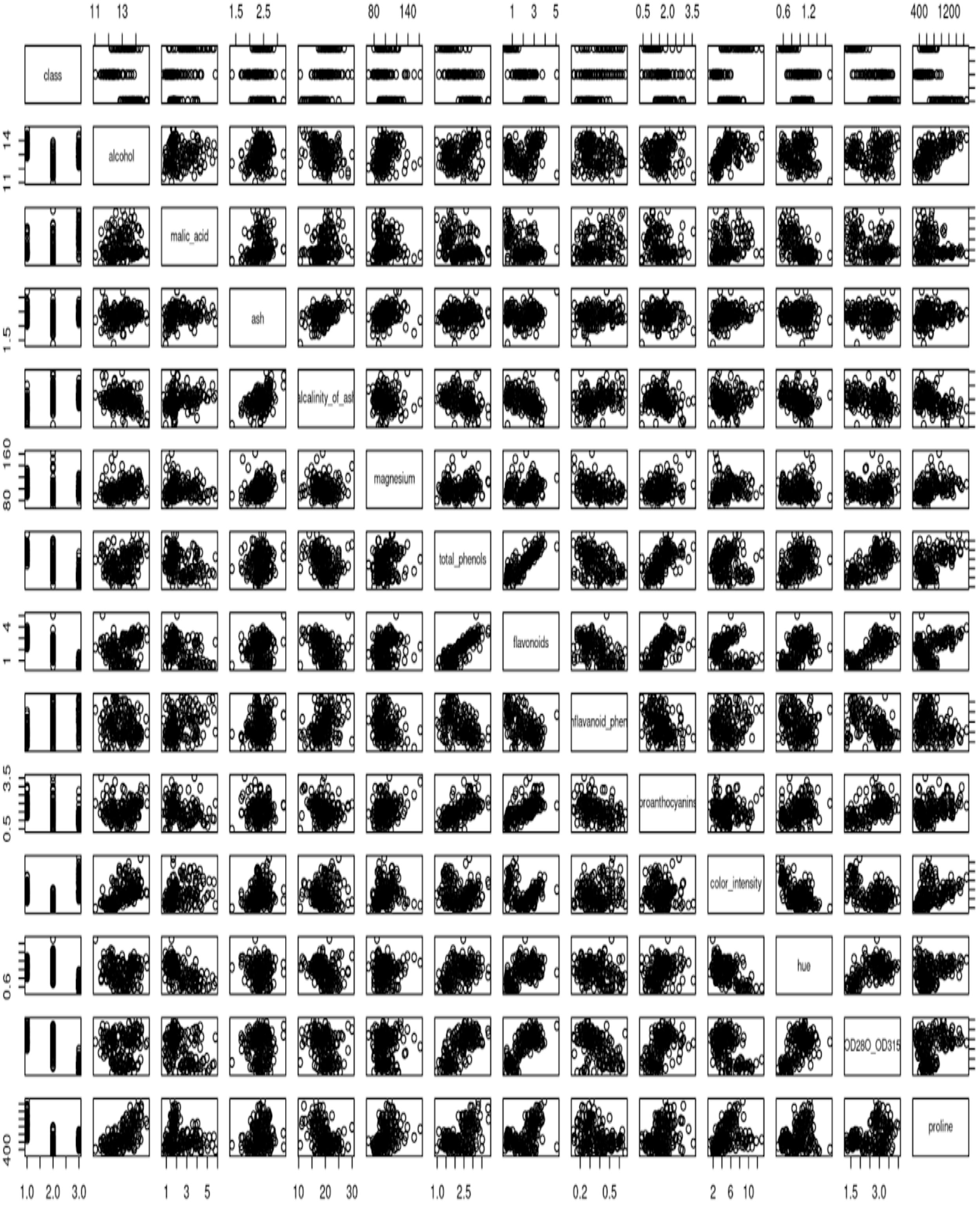
There are many different types of classification such as geographical classification, chronological classification, quantitative classification, etc. It is explicitly stated in the data set a description that the thirteen attributes are representing the results of a chemical analysis determining the quantities of 13 constituents(aka the attributes) found in each of the three types of wine (represented by the class identifier attribute (1-3)). Therefore, quantitative classification exclusively makes sense for this particular dataset in this particular project with a certain set of circumstances. Furthermore, our data preprocessing also involves understanding how many instances are found in each of the three classes. Class 1 will have 59 instances, Class 2 will have 71 instances, and finally Class shall have 48 instances which make up all the 178 instances in total.

The datasets are usually given as CSV files, .names files, or .data files. We can import the raw data into R using the command **wine_ds <-** read.csv("https://archive.ics.uci.edu/ml/machine-learning-databases/wine/wine.data", header = FALSE) with the converting of char to factor and rescaling or normalizing wherever needed. Since we have three methods, the pre-processing of data appropriately for each machine learning method is of utmost importance. When it came to accuracy, a table was shown for when the percentage is unscaled versus scaled as for research purposes. This is further explained in the comparison of results section later on. The attributes only have two domains which are integer and numeric, which is in line with the dataset description. For numeric attributes, everything is mostly in percentages and it is centered around zero. For int attributes, they are quite similar to numeric but they can only contain whole numbers within the context of the R language and are standardized to be a subset of “numeric”.

Training Models:

Since there is only classification listed for the UCI wine dataset, this is fairly straightforward. For classification, we chose the Quadratic Discriminant Analysis model, K-Nearest Neighbor model, and finally the Linear Discriminant Analysis model. That makes three machine learning methods instead of the two required, meaning we have already fulfilled & exceeded the expectations of that portion of the group project rubric.

Predictor Selection (insert graphs) plot command:



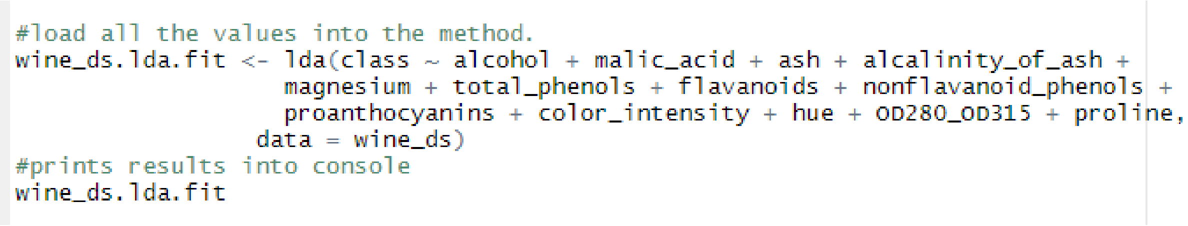
Classification Analysis Workflow:

For the classification analysis workflow, we display the prediction results in confusion matrices, accuracy percentages, and ROC curves. A summary of the analysis results will be close to the conclusion section at the bottom of this project report. Each training Models which have a parameter or parameters to specify which range of values works the best for each case. ROC curves are a fantastic tool for numeric classification, and will show the performance of a classification model at all classification thresholds. The first step is to attempt cross-validation to measure how accurately the predictive models can perform in practice, but also to protect against overfitting which can negatively affect our results or perhaps not give an accurate picture as we'd preferred. In addition, we can begin to plot the ROC curve with the preferred predictors (13 attributes, second step) and the best training model parameter. For comparison purposes, other ROC curves with other wine model parameters are shown on the side. Since there are many methods to determine the optimal threshold, one can employ a strategy to look for the closest point to the left top of the graph as covered in various CSE 5160 zoom lectures. A table/screenshot of the confusion matrix for the dataset will be present in the LDA, QDA, and KNN analysis sections which are up next.

Linear Discriminant Analysis (Erik's work, use as many pages as you need):

Classification:

Our overall goal with linear discriminant would be able to pinpoint which ingredient is common among the three classes. Since we need to see the commonality between the three classes, for the LDA function we are going to include all of the ingredients together.



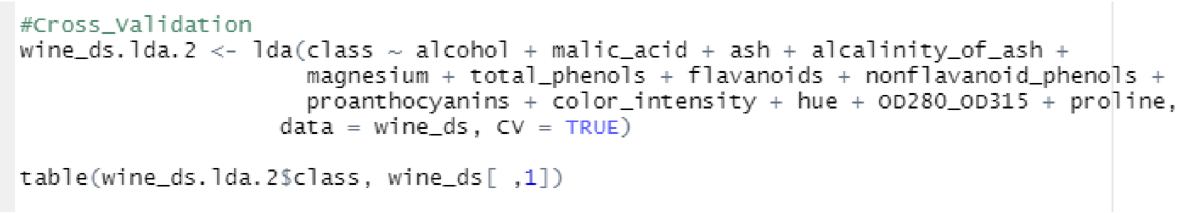
```
#load all the values into the method.
wine_ds.lda.fit <- lda(class ~ alcohol + malic_acid + ash + alcalinity_of_ash +
                       magnesium + total_phenols + flavanoids + nonflavanoid_phenols +
                       proanthocyanins + color_intensity + hue + OD280_OD315 + proline,
                     data = wine_ds)
#prints results into console
wine_ds.lda.fit
```

This prints out the following tables.

Group Means	Alcohol	Malic Acid	Ash	Alcalinity of Ash	Magnesium	Total Phenois	Flavanoids
1	13.66957	2.028913	2.432609	16.97826	105.41304	2.821087	2.9828261
2	12.27474	1.956842	2.267544	20.18772	96.07018	2.367895	2.1770175
3	13.17158	3.477368	2.437105	21.57895	99.13158	1.671842	0.7615789

Group Means	Nonflavanoid Phenois	Proanthocyanins	Color Intensity	Hue	OD280 OD315	Proline
1	0.2795652	1.906087	5.450652	1.0504348	3.166739	1087.8696
2	0.3612281	1.678421	3.110877	1.0492982	2.807368	522.9474
3	0.4576316	1.166316	7.590526	0.6718421	1.678421	625.3947

What exactly do these tables mean? They provide us the average amount that was put into each of these specific type of wines. The sections I have highlighted seem to have the same amount of values that was placed into each class. While they may be off by 0.5+-, we can leave that up to manufacturing error just look at the number as a whole. So we've narrowed down the certain ingredients that can be found among the three classes. Lets look at the cross validation. In order to do so, we utilize the same approach as before and just append a cv = true which turns out to be something like the following:



```
#Cross_validation
wine_ds.lda.2 <- lda(class ~ alcohol + malic_acid + ash + alcalinity_of_ash +
                     magnesium + total_phenols + flavanoids + nonflavanoid_phenols +
                     proanthocyanins + color_intensity + hue + OD280_OD315 + proline,
                   data = wine_ds, CV = TRUE)

table(wine_ds.lda.2$class, wine_ds[ ,1])
```

Which provides us the following table:

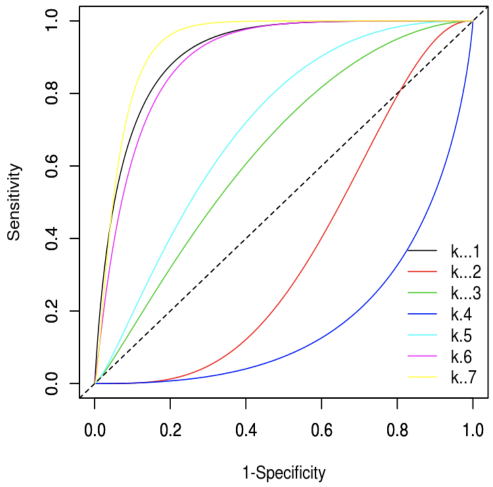
CV_Table	1	2	3
1	58	1	0
2	0	69	0
3	0	1	48

The boxes that were highlighted yellow provide to us that there was some misclassification happening. There is a possibility that class 2 has been misclassified as both class 1 and class 2. This could lead to some means being skewed which in turn could not show us what other ingredients are common to the three classes but with the three ingredients we are able to obtain through linear discriminant analysis.

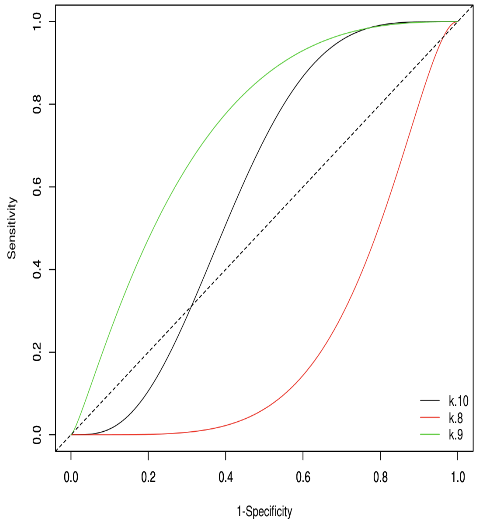
K-Nearest Neighbor - KNN (Mario's Work):

Classification:

ROC Curve: KNN on K = 1 to K = 7
with 0.05 type one error



ROC Curve: KNN on K = 8 to K = 10
with 0.10 type one error



Due to the way R studio cloud operates, it doesn't show in the ROC Curve, but when it says "k ..1, k…2, k.4, ,etc. " it's really just saying k = 1, k = 2, k= 3, and k =4 and so forth.

KNN: Confusion matrices

	Class 1	Class 2	Class 3
1	15	1	4
2	3	15	4
3	0	9	8

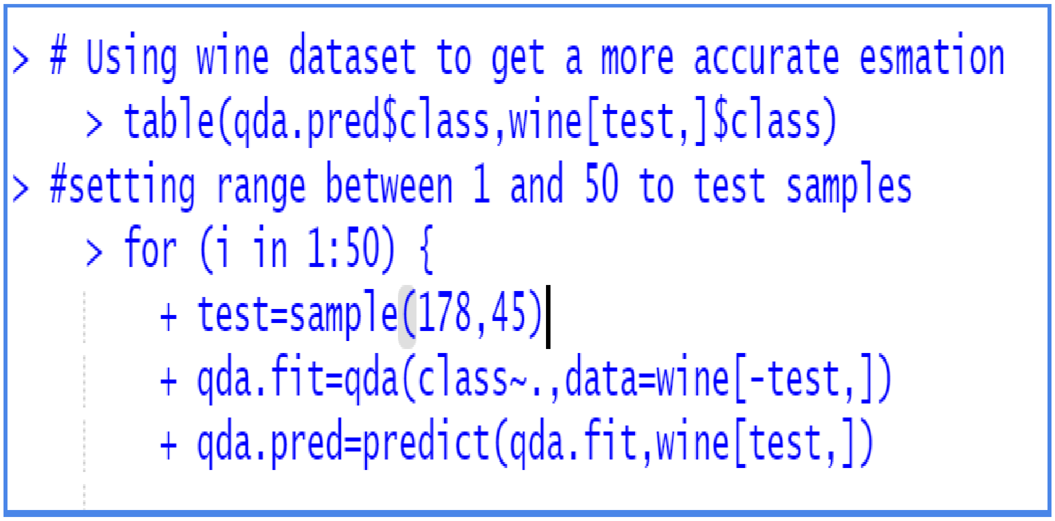
KNN Classification Accuracy

	Accuracy before data pre-processing	Accuracy after data pre-processing (convert char to factor, rescale or normalize)
K = 1	72.9%	94.9%
K = 2	70.0%	93.3%
K = 3	71.4%	94.4%
K = 4	67.6%	94.8%
K = 5	70.2%	95.1%
K = 6	70.5%	95.1%
K = 7	68.7%	94.61%

Quadratic Discriminant Analysis (Bat’s Work, use as many pages as you need):

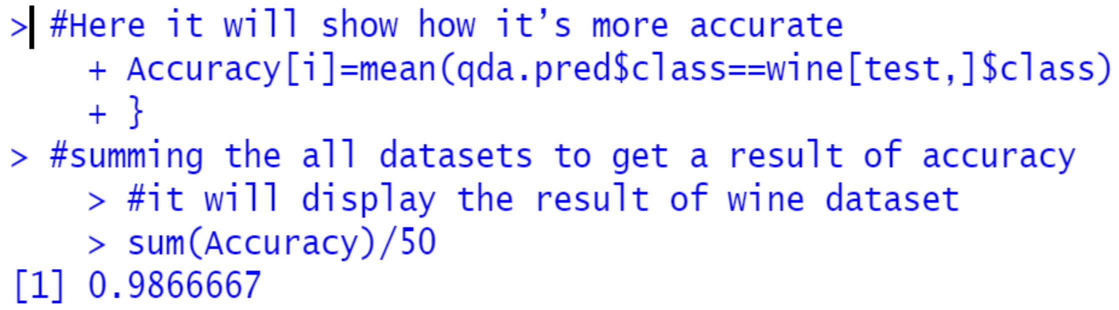
Classification:

Discriminant analysis is statistical technique used to classify observations into non-overlapping groups, based on scores on one or more quantitative predictor variables. For example, a doctor could perform a discriminant analysis to identify patients at high or low risk for stroke.



```
> # Using wine dataset to get a more accurate esmation
   > table(qda.pred$class,wine[test,]$class)
> #setting range between 1 and 50 to test samples
   > for (i in 1:50) {
     + test=sample(178,45)
     + qda.fit=qda(class~.,data=wine[-test,])
     + qda.pred=predict(qda.fit,wine[test,])
```

The output contains the group means. But it does not contain the coefficients of the linear discriminants, because the QDA classifier involves a quadratic, rather than a linear, function of the predictors. The predict() function works in exactly the same fashion as for LDA.



```
> #Here it will show how it’s more accurate
    + Accuracy[i]=mean(qda.pred$class==wine[test,]$class)
    + }
> #summing the all datasets to get a result of accuracy
    > #it will display the result of wine dataset
    > sum(Accuracy)/50
[1] 0.9866667
```

Summary & Comparison of Results (everyone contributes):

Classification:

Training Models	Accuracy Percentage Rate
Linear Discriminant Analysis	98.5%
Quadratic Discriminant Analysis	98.4%
KNN (k = 1)	94.9%
KNN (k = 3)	94.4%
KNN (k = 7)	94.61%

Conclusion (everyone contributes):

When referring to linear discriminate analysis, it can be seen from using the function that the three common ingredients between the three classes are to be: alcohol, ash, and proanthocyanins. There are other ingredients that are shared between only two classes such as magnesium, total phenols, flavonoids, hue, and proline. The overall goal was to determine the type of wine that these classes could possibly be. Proanthocyanins turn out to be grape seeds that can be found within red wine. The alcohol portion is a bit tricky to understand but in a sense, it uses yeast to break down the type of grape that will be used for the wine down to its sugars form which in turn produces both carbon dioxide and ethanol. There is something to note here. Both of these ingredients are not listed within the attributes which can either be one of two answers. They were included originally but were lost as the owner of the dataset had stated there were originally 30 and now they're only 13 present. The second reason would be that all three wines contained the same amount of carbon dioxide/ethanol that the author felt it was not needed to include. In the terms of KNN analysis, it has been discovered that KNN is the least accurate of the analysis, both from k = 1, k= 2, etc. Although in a close third, QDA and LDA rank ahead in terms of the most accurate training data. We find that because of our KNN analysis, pinpointing a certain amount of proline or magnesium will explicitly point towards class 1, class 2, or class 3 wines. This can be attributed to the fact that K-nearest neighbors are because the K in KNN has properties that if implemented correctly, can avoid the pitfalls from the curse of dimensionality

References/Works Cited:

- Wine Data Set
  https://archive.ics.uci.edu/ml/datasets/Wine

- Forina, M. et al, PARVUS - An Extendible Package for Data Exploration, Classification and Correlation. Institute of Pharmaceutical and Food Analysis and Technologies, Via Brigata Salerno, 16147 Genoa, Italy.

- Proline Content of Grapes and Wines - VITIS Journal of Grapevine Research
  https://ojs.openagrar.de/index.php/VITIS/article/view/7459#:~:text=The%20amount%20of%20proline%20in,higher%20percentages%20of%20residual%20proline