Final Project

Charaf Lachouri, Mohamed Tounkara, Rohan Thaliachery, Tatiana Uklist

2022-11-30

Introduction

Wine making, also known as vinification, is the process of producing wine, starting from the selection of fruit (typically grapes), its fermentation, and the bottling of the finished product. This art form of a process stretches over millennium with the first documented instances being around since between 5000 - 5400 BC. The process has been perfected and celebrated over the years while at the same time, the finished product itself has been beloved and held both religiously and socially sacred ever since. And as the process has grown, the classification of wine has evolved as well.

Today, the 100 point scale is what is widely used with top rated wines usually rating above 90. The finalized ranking is typically the average of all the points given to that wine. The tasters are usually looking at taste and physical features such as color, sugar-level, growing method, and climate that the fruits are grown in. The tastings are usually done blindly in order to prevent any bias towards brands, vineyards or winemakers. And while the taste features are important, the taster's own personal preference will always bias their ranking. That got our group thinking of alternative ways that the wine could in theory be ranked. We were curious about what the chemical composition did to the wine classification.

Using multinomial regression analysis, we will create a model to predict the wine class. We will train our multinomial regression analysis model over a hundred iterations. After training the data, we will split the data into thirty percent for training and seventy percent for testing the data in order to predict the wine classes.

Data set information

Source:

Original Owners:

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.

Donor:

Stefan Aeberhard, email: stefan '@' coral.cs.jcu.edu.au

These data are the results of a chemical analysis of wines grown in the same region in Italy but derived from three different cultivars. The analysis determined the quantities of 13 constituents found in each of the three types of wines.

I think that the initial data set had around 30 variables, but for some reason I only have the 13 dimensional version. I had a list of what the 30 or so variables were, but a.) I lost it, and b.), I would not know which 13 variables are included in the set.

The attributes are (dontated by Riccardo Leardi, riclea '@' anchem.unige.it) 1) Alcohol 2) Malic acid 3) Ash 4) Alcalinity of ash 5) Magnesium 6) Total phenols 7) Flavanoids 8) Nonflavanoid phenols 9) Proanthocyanins 10) Color intensity 11) Hue 12) OD280/OD315 of diluted wines 13) Proline

1. Loading Packages and Libraries

```r
library(tidyverse) # To structure, manipulate and visualize data.
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.4.0      v purrr   0.3.5
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.5.0
## v readr   2.1.3      v forcats 0.5.2

## Warning: package 'ggplot2' was built under R version 4.2.2

## Warning: package 'tidyr' was built under R version 4.2.2

## Warning: package 'readr' was built under R version 4.2.2

## Warning: package 'purrr' was built under R version 4.2.2

## Warning: package 'dplyr' was built under R version 4.2.2

## Warning: package 'stringr' was built under R version 4.2.2

## Warning: package 'forcats' was built under R version 4.2.2

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```r
library(car) # To test, transform and visualize data.
```

```
## Loading required package: carData
##
## Attaching package: 'car'
##
## The following object is masked from 'package:dplyr':
##
##     recode
##
## The following object is masked from 'package:purrr':
##
##     some
```

```r
library(MASS) # To do data transformation.
```

```
##
## Attaching package: 'MASS'
##
## The following object is masked from 'package:dplyr':
##
##     select
```

```r
library(ggplot2) # To do data visualization.
library(KODAMA) # To do unsupervised features prediction.
```

```
## Warning: package 'KODAMA' was built under R version 4.2.2

## Loading required package: minerva

## Warning: package 'minerva' was built under R version 4.2.2

## Loading required package: Rtsne

## Warning: package 'Rtsne' was built under R version 4.2.2
```

```r
library(dplyr) # To do data manipulation
library(nnet) # To do neural network classification
```

2. Loading the data and Eploratory Analysis

```r
wine <- read.csv("Wine_Dataset.csv")
attach(wine)
head(wine, 10)
```

```
##    Classes Alcohol Malic.acid  Ash Alcalinity.of.ash Magnesium Total.phenols
## 1        1   14.23       1.71 2.43              15.6       127          2.80
## 2        1   13.20       1.78 2.14              11.2       100          2.65
## 3        1   13.16       2.36 2.67              18.6       101          2.80
## 4        1   14.37       1.95 2.50              16.8       113          3.85
## 5        1   13.24       2.59 2.87              21.0       118          2.80
## 6        1   14.20       1.76 2.45              15.2       112          3.27
## 7        1   14.39       1.87 2.45              14.6        96          2.50
## 8        1   14.06       2.15 2.61              17.6       121          2.60
## 9        1   14.83       1.64 2.17              14.0        97          2.80
## 10       1   13.86       1.35 2.27              16.0        98          2.98
##    Flavanoids Nonflavanoid.phenols Proanthocyanins Color.intensity  Hue
## 1        3.06                 0.28            2.29            5.64 1.04
## 2        2.76                 0.26            1.28            4.38 1.05
## 3        3.24                 0.30            2.81            5.68 1.03
## 4        3.49                 0.24            2.18            7.80 0.86
## 5        2.69                 0.39            1.82            4.32 1.04
## 6        3.39                 0.34            1.97            6.75 1.05
## 7        2.52                 0.30            1.98            5.25 1.02
## 8        2.51                 0.31            1.25            5.05 1.06
## 9        2.98                 0.29            1.98            5.20 1.08
## 10       3.15                 0.22            1.85            7.22 1.01
##    OD280.OD315.of.diluted.wines Proline
## 1                          3.92    1065
## 2                          3.40    1050
## 3                          3.17    1185
```

```
## 4                              3.45    1480
## 5                              2.93     735
## 6                              2.85    1450
## 7                              3.58    1290
## 8                              3.58    1295
## 9                              2.85    1045
## 10                             3.55    1045
```

The dataset contains information about 178 unique wines divided into three categories which are represented by 1 to 3 numbers. The dependent variable here is Classes.

```
# Data Dimensions
dim(wine)
```

```
## [1] 178  14
```

3. Statistical Summary

In our dataset, the average alcohol percentage is 13%???

```
# Descriptions
summary(wine)
```

```
##     Classes         Alcohol        Malic.acid         Ash
##  Min.   :1.000   Min.   :11.03   Min.   :0.740   Min.   :1.360
##  1st Qu.:1.000   1st Qu.:12.36   1st Qu.:1.603   1st Qu.:2.210
##  Median :2.000   Median :13.05   Median :1.865   Median :2.360
##  Mean   :1.938   Mean   :13.00   Mean   :2.336   Mean   :2.367
##  3rd Qu.:3.000   3rd Qu.:13.68   3rd Qu.:3.083   3rd Qu.:2.558
##  Max.   :3.000   Max.   :14.83   Max.   :5.800   Max.   :3.230
##  Alcalinity.of.ash   Magnesium      Total.phenols     Flavanoids
##  Min.   :10.60     Min.   : 70.00   Min.   :0.980   Min.   :0.340
##  1st Qu.:17.20     1st Qu.: 88.00   1st Qu.:1.742   1st Qu.:1.205
##  Median :19.50     Median : 98.00   Median :2.355   Median :2.135
##  Mean   :19.49     Mean   : 99.74   Mean   :2.295   Mean   :2.029
##  3rd Qu.:21.50     3rd Qu.:107.00   3rd Qu.:2.800   3rd Qu.:2.875
##  Max.   :30.00     Max.   :162.00   Max.   :3.880   Max.   :5.080
##  Nonflavanoid.phenols Proanthocyanins Color.intensity       Hue
##  Min.   :0.1300       Min.   :0.410   Min.   : 1.280   Min.   :0.4800
##  1st Qu.:0.2700       1st Qu.:1.250   1st Qu.: 3.220   1st Qu.:0.7825
##  Median :0.3400       Median :1.555   Median : 4.690   Median :0.9650
##  Mean   :0.3619       Mean   :1.591   Mean   : 5.058   Mean   :0.9574
##  3rd Qu.:0.4375       3rd Qu.:1.950   3rd Qu.: 6.200   3rd Qu.:1.1200
##  Max.   :0.6600       Max.   :3.580   Max.   :13.000   Max.   :1.7100
##  OD280.OD315.of.diluted.wines    Proline
##  Min.   :1.270                Min.   : 278.0
##  1st Qu.:1.938                1st Qu.: 500.5
##  Median :2.780                Median : 673.5
##  Mean   :2.612                Mean   : 746.9
##  3rd Qu.:3.170                3rd Qu.: 985.0
##  Max.   :4.000                Max.   :1680.0
```

We have identify identify 3 classes which will be used to classify the wine based on several variables

```r
#Counts of classes in data
table(Classes)
```

```
## Classes
##  1  2  3
## 59 71 48
```

#Our dataset is structured around 2 types of data: 3 Integers (Classes, Magnesium and Proline) and 11 Numeric data

```r
# Checking the structure of wine dataset
str(wine)
```

```
## 'data.frame':    178 obs. of  14 variables:
##  $ Classes                     : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ Alcohol                     : num  14.2 13.2 13.2 14.4 13.2 ...
##  $ Malic.acid                  : num  1.71 1.78 2.36 1.95 2.59 1.76 1.87 2.15 1.64 1.35 ...
##  $ Ash                         : num  2.43 2.14 2.67 2.5 2.87 2.45 2.45 2.61 2.17 2.27 ...
##  $ Alcalinity.of.ash           : num  15.6 11.2 18.6 16.8 21 15.2 14.6 17.6 14 16 ...
##  $ Magnesium                   : int  127 100 101 113 118 112 96 121 97 98 ...
##  $ Total.phenols               : num  2.8 2.65 2.8 3.85 2.8 3.27 2.5 2.6 2.8 2.98 ...
##  $ Flavanoids                  : num  3.06 2.76 3.24 3.49 2.69 3.39 2.52 2.51 2.98 3.15 ...
##  $ Nonflavanoid.phenols        : num  0.28 0.26 0.3 0.24 0.39 0.34 0.3 0.31 0.29 0.22 ...
##  $ Proanthocyanins             : num  2.29 1.28 2.81 2.18 1.82 1.97 1.98 1.25 1.98 1.85 ...
##  $ Color.intensity             : num  5.64 4.38 5.68 7.8 4.32 6.75 5.25 5.05 5.2 7.22 ...
##  $ Hue                         : num  1.04 1.05 1.03 0.86 1.04 1.05 1.02 1.06 1.08 1.01 ...
##  $ OD280.OD315.of.diluted.wines: num  3.92 3.4 3.17 3.45 2.93 2.85 3.58 3.58 2.85 3.55 ...
##  $ Proline                     : int  1065 1050 1185 1480 735 1450 1290 1295 1045 1045 ...
```

4. Data cleaning (remove noise and inconsistent data)

Using sum and is.na function we will check for any missing values in our dataset. If we find any missing values, we will remove it from our dataset by na.omit() function and check the dimension for data set.

```r
# Missing values ?
sum(is.na(wine))
```

```
## [1] 0
```

No missing values found. # Changing our response variable to a factor Changing our variables in factors helped us to identify the different types of classes. In our case classes are between 1 and 3

```r
Classes <- as.factor(Classes)
Classes
```

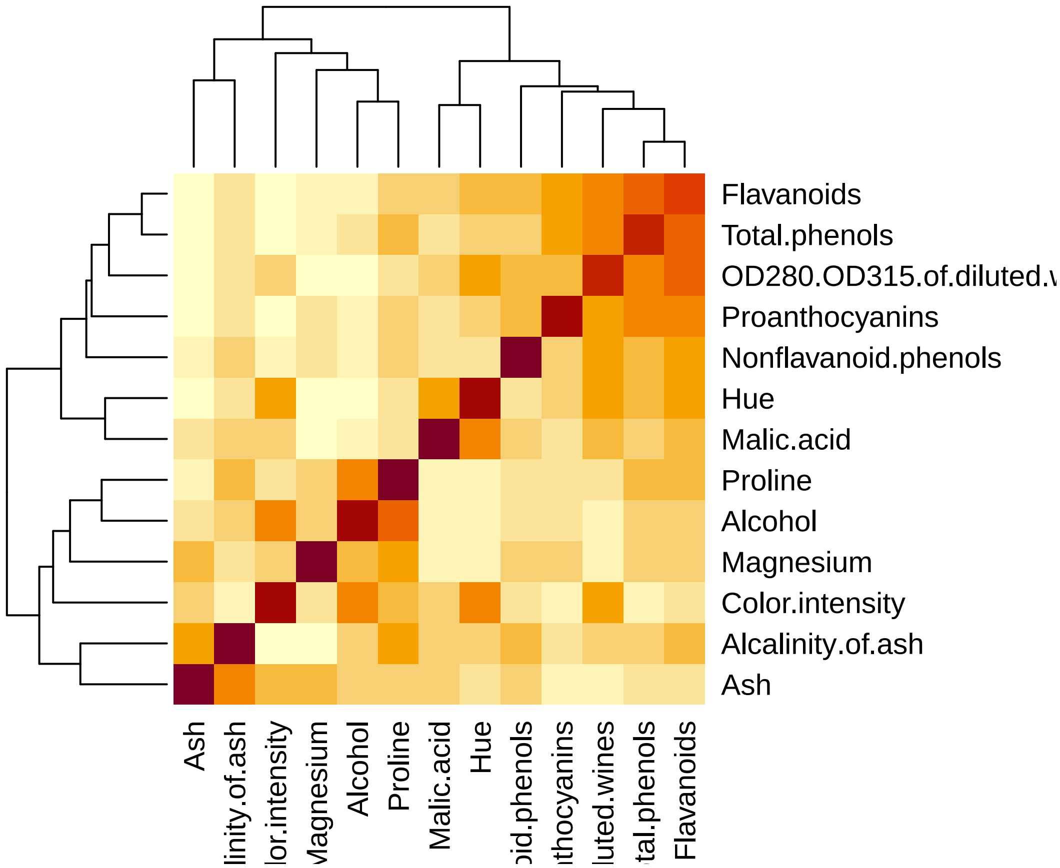
```
##   [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
##  [38] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
##  [75] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
## [112] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
## [149] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
## Levels: 1 2 3
```

```r
# Checking for correlation between the predictors
cor(wine[,-1])
```

```
##                                  Alcohol  Malic.acid          Ash
## Alcohol                       1.00000000  0.09439694  0.211544596
## Malic.acid                    0.09439694  1.00000000  0.164045470
## Ash                           0.21154460  0.16404547  1.000000000
## Alcalinity.of.ash            -0.31023514  0.28850040  0.443367187
## Magnesium                     0.27079823 -0.05457510  0.286586691
## Total.phenols                 0.28910112 -0.33516700  0.128979538
## Flavanoids                    0.23681493 -0.41100659  0.115077279
## Nonflavanoid.phenols         -0.15592947  0.29297713  0.186230446
## Proanthocyanins               0.13669791 -0.22074619  0.009651935
## Color.intensity               0.54636420  0.24898534  0.258887259
## Hue                          -0.07174720 -0.56129569 -0.074666889
## OD280.OD315.of.diluted.wines  0.07234319 -0.36871043  0.003911231
## Proline                       0.64372004 -0.19201056  0.223626264
##                              Alcalinity.of.ash   Magnesium Total.phenols
## Alcohol                            -0.31023514  0.27079823    0.28910112
## Malic.acid                          0.28850040 -0.05457510   -0.33516700
## Ash                                 0.44336719  0.28658669    0.12897954
## Alcalinity.of.ash                   1.00000000 -0.08333309   -0.32111332
## Magnesium                          -0.08333309  1.00000000    0.21440123
## Total.phenols                      -0.32111332  0.21440123    1.00000000
## Flavanoids                         -0.35136986  0.19578377    0.86456350
## Nonflavanoid.phenols                0.36192172 -0.25629405   -0.44993530
## Proanthocyanins                    -0.19732684  0.23644061    0.61241308
## Color.intensity                     0.01873198  0.19995001   -0.05513642
## Hue                                -0.27395522  0.05539820    0.43368134
## OD280.OD315.of.diluted.wines       -0.27676855  0.06600394    0.69994936
## Proline                            -0.44059693  0.39335085    0.49811488
##                              Flavanoids Nonflavanoid.phenols Proanthocyanins
## Alcohol                       0.2368149           -0.1559295     0.136697912
## Malic.acid                   -0.4110066            0.2929771    -0.220746187
## Ash                           0.1150773            0.1862304     0.009651935
## Alcalinity.of.ash            -0.3513699            0.3619217    -0.197326836
## Magnesium                     0.1957838           -0.2562940     0.236440610
## Total.phenols                 0.8645635           -0.4499353     0.612413084
## Flavanoids                    1.0000000           -0.5378996     0.652691769
## Nonflavanoid.phenols         -0.5378996            1.0000000    -0.365845099
## Proanthocyanins               0.6526918           -0.3658451     1.000000000
## Color.intensity              -0.1723794            0.1390570    -0.025249931
## Hue                           0.5434786           -0.2626396     0.295544253
## OD280.OD315.of.diluted.wines  0.7871939           -0.5032696     0.519067096
## Proline                       0.4941931           -0.3113852     0.330416700
##                              Color.intensity         Hue
## Alcohol                           0.54636420 -0.07174720
```

```
## Malic.acid                         0.24898534 -0.56129569
## Ash                                0.25888726 -0.07466689
## Alcalinity.of.ash                  0.01873198 -0.27395522
## Magnesium                          0.19995001  0.05539820
## Total.phenols                     -0.05513642  0.43368134
## Flavanoids                        -0.17237940  0.54347857
## Nonflavanoid.phenols               0.13905701 -0.26263963
## Proanthocyanins                   -0.02524993  0.29554425
## Color.intensity                    1.00000000 -0.52181319
## Hue                               -0.52181319  1.00000000
## OD280.OD315.of.diluted.wines      -0.42881494  0.56546829
## Proline                            0.31610011  0.23618345
##                              OD280.OD315.of.diluted.wines    Proline
## Alcohol                                       0.072343187  0.6437200
## Malic.acid                                   -0.368710428 -0.1920106
## Ash                                           0.003911231  0.2236263
## Alcalinity.of.ash                            -0.276768549 -0.4405969
## Magnesium                                     0.066003936  0.3933508
## Total.phenols                                 0.699949365  0.4981149
## Flavanoids                                    0.787193902  0.4941931
## Nonflavanoid.phenols                         -0.503269596 -0.3113852
## Proanthocyanins                               0.519067096  0.3304167
## Color.intensity                              -0.428814942  0.3161001
## Hue                                           0.565468293  0.2361834
## OD280.OD315.of.diluted.wines                  1.000000000  0.3127611
## Proline                                       0.312761075  1.0000000
```

```r
heatmap(abs(cor(wine[,-1])))
```



We have slightly correlated predictors : 1. "Alcohol" and "Proline" (0.64). 2. "Hue" and "Malic.acid" (-0.56), 3. "OD280.OD315.of.diluted.wines" and "Flavanoids" (0.79). 4. "OD280.OD315.of.diluted.wines" and "Total.phenols" (0.70).

Let's check the significance of each predictor !

```
# Multiple Linear Regression
fit = lm(Classes ~ Alcohol+Malic.acid+Ash+Alcalinity.of.ash+Magnesium+Total.phenols+Flavanoids+Nonflava
summary(fit)
```

```
##
## Call:
## lm(formula = Classes ~ Alcohol + Malic.acid + Ash + Alcalinity.of.ash +
##     Magnesium + Total.phenols + Flavanoids + Nonflavanoid.phenols +
##     Proanthocyanins + Color.intensity + Hue + OD280.OD315.of.diluted.wines +
##     Proline, data = wine)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.64129 -0.16074 -0.02535  0.15778  0.72912
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  4.4732853  0.4976137   8.989 5.79e-16 ***
## Alcohol                     -0.1170038  0.0369610  -3.166  0.00185 **
## Malic.acid                   0.0301710  0.0220400   1.369  0.17290
```

```
## Ash                           -0.1485522  0.1030816  -1.441  0.15146
## Alcalinity.of.ash              0.0398543  0.0085707   4.650 6.79e-06 ***
## Magnesium                     -0.0004898  0.0015948  -0.307  0.75916
## Total.phenols                  0.1443201  0.0636364   2.268  0.02464 *
## Flavanoids                    -0.3723914  0.0507762  -7.334 9.74e-12 ***
## Nonflavanoid.phenols          -0.3034743  0.2060150  -1.473  0.14265
## Proanthocyanins                0.0393565  0.0469782   0.838  0.40338
## Color.intensity                0.0756239  0.0143547   5.268 4.28e-07 ***
## Hue                           -0.1492451  0.1336834  -1.116  0.26588
## OD280.OD315.of.diluted.wines -0.2700542  0.0524220  -5.152 7.34e-07 ***
## Proline                       -0.0007011  0.0001021  -6.868 1.28e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2545 on 164 degrees of freedom
## Multiple R-squared:  0.9001, Adjusted R-squared:  0.8922
## F-statistic: 113.7 on 13 and 164 DF,  p-value: < 2.2e-16
```

In this case "TRUE" means that the P value<0.05 as a result it will show that there is significant relationship between the intercept and the those variables.

```r
# P-value of each coefficient less than 0.05
summary(fit)$coef[,4] < 0.05
```

```
##                  (Intercept)                      Alcohol
##                         TRUE                         TRUE
##                   Malic.acid                          Ash
##                        FALSE                        FALSE
##            Alcalinity.of.ash                    Magnesium
##                         TRUE                        FALSE
##                Total.phenols                   Flavanoids
##                         TRUE                         TRUE
##         Nonflavanoid.phenols              Proanthocyanins
##                        FALSE                        FALSE
##              Color.intensity                          Hue
##                         TRUE                        FALSE
## OD280.OD315.of.diluted.wines                      Proline
##                         TRUE                         TRUE
```

```r
# Variance Inflation Factor (VIF)
round(vif(fit),2)
```

```
##                      Alcohol                   Malic.acid
##                         2.46                         1.66
##                          Ash            Alcalinity.of.ash
##                         2.19                         2.24
##                    Magnesium                Total.phenols
##                         1.42                         4.33
##                   Flavanoids         Nonflavanoid.phenols
##                         7.03                         1.80
##              Proanthocyanins              Color.intensity
##                         1.98                         3.03
##                          Hue OD280.OD315.of.diluted.wines
```

```
##                           2.55                              3.79
##                        Proline
##                           2.82
```

From the heat map correlated predictors and the non-significant coefficients. We decided to remove the following independent variables : “Hue”, “Magnesium”, “Proanthocyanins” and “Ash”.

5. Splitting the data into train and test

To begin, we’ll create a fake indicator to indicate whether a row is in the training or testing data set. In an ideal world, we’d have 70% training data and 30% testing data, which would provide the highest level of accuracy.

```r
# Using sample_frac to create 30 - 70 slipt into test and train
train <- sample_frac(wine, 0.3)
sample_id <- as.numeric(rownames(train)) # rownames() returns character so as.numeric
test <- wine[-sample_id,]
head(test)
```

```
##    Classes Alcohol Malic.acid  Ash Alcalinity.of.ash Magnesium Total.phenols
## 54       1   13.77       1.90 2.68              17.1       115          3.00
## 55       1   13.74       1.67 2.25              16.4       118          2.60
## 56       1   13.56       1.73 2.46              20.5       116          2.96
## 57       1   14.22       1.70 2.30              16.3       118          3.20
## 58       1   13.29       1.97 2.68              16.8       102          3.00
## 59       1   13.72       1.43 2.50              16.7       108          3.40
##    Flavanoids Nonflavanoid.phenols Proanthocyanins Color.intensity  Hue
## 54       2.79                 0.39            1.68            6.30 1.13
## 55       2.90                 0.21            1.62            5.85 0.92
## 56       2.78                 0.20            2.45            6.25 0.98
## 57       3.00                 0.26            2.03            6.38 0.94
## 58       3.23                 0.31            1.66            6.00 1.07
## 59       3.67                 0.19            2.04            6.80 0.89
##    OD280.OD315.of.diluted.wines Proline
## 54                         2.93    1375
## 55                         3.20    1060
## 56                         3.03    1120
## 57                         3.31     970
## 58                         2.84    1270
## 59                         2.87    1285
```

We use mutinom() function from {nnet} package and relevel() function to set up the Classes baseline level. Multinomial regression is an extension of binomial logistic regression allows us to predict a categorical dependent variable which has more than two levels.

```r
# Setting up the baseline
train$Classes <- relevel(factor(train$Classes), ref = "3")
```

6. Training the multinomial model

```r
multinom.fit <- multinom (Classes ~ Alcohol+Malic.acid+Alcalinity.of.ash+Total.phenols+Flavanoids+Non
```

```
## # weights:  33 (20 variable)
## initial  value 58.226451
## iter  10 value 4.590928
## iter  20 value 0.087445
## iter  30 value 0.000261
## final  value 0.000068
## converged
```

```r
# Checking the model
summary(multinom.fit)
```

```
## Call:
## multinom(formula = Classes ~ Alcohol + Malic.acid + Alcalinity.of.ash +
##     Total.phenols + Flavanoids + Nonflavanoid.phenols + Color.intensity +
##     OD280.OD315.of.diluted.wines + Proline, data = train)
##
## Coefficients:
##   (Intercept)    Alcohol Malic.acid Alcalinity.of.ash Total.phenols Flavanoids
## 1   -73.45100 -10.065555 -13.381160        -0.4598077    -125.79314  110.63242
## 2    84.83466   1.366173   7.401313        -3.4035001      18.87283   17.37394
##   Nonflavanoid.phenols Color.intensity OD280.OD315.of.diluted.wines    Proline
## 1             50.17449        10.31050                    54.439161 0.14304140
## 2            -34.79025       -38.38645                     2.327358 0.08470754
##
## Std. Errors:
##   (Intercept)   Alcohol Malic.acid Alcalinity.of.ash Total.phenols Flavanoids
## 1    2.668735  22.39979   42.50948         134.06715      54.24256   49.69442
## 2   18.648924 278.27729  119.81909          92.20358      93.15999   79.24694
##   Nonflavanoid.phenols Color.intensity OD280.OD315.of.diluted.wines  Proline
## 1             6.606810        219.0967                     17.12509 2.670475
## 2             7.098462        114.9738                     72.48370 5.820107
##
## Residual Deviance: 0.0001356161
## AIC: 40.00014
```

The output of summary contains the table for coefficients and a table for standard error. Each row in the coefficient table corresponds to the model equation. The first row represents the coefficients for Class 2 wine in comparison to our baseline which is Class 3 wine and the second row represents the coefficients for Class 2 wine in comparison to our baseline which is Class 3 wine.

The output coefficients are represented in the log of odds.

This ratio of the probability of choosing Class 2 wine over the baseline that is Class 3 wine is referred to as relative risk (often described as odds). However, the output of the model is the log of odds. To get the relative risk IE odds ratio, we need to exponentiate the coefficients.

```r
## extracting coefficients from the model and exponentiate
exp(coef(multinom.fit))
```

```
##    (Intercept)      Alcohol   Malic.acid Alcalinity.of.ash Total.phenols
```

```
## 1 1.260772e-32 4.251918e-05 1.543961e-06        0.63140503  2.337395e-55
## 2 6.969896e+36 3.920318e+00 1.638135e+03        0.03325666  1.571686e+08
##     Flavanoids Nonflavanoid.phenols Color.intensity
## 1 1.114424e+48         6.173086e+21    3.004632e+04
## 2 3.510794e+07         7.776506e-16    2.132932e-17
##   OD280.OD315.of.diluted.wines  Proline
## 1                 4.391645e+23 1.153778
## 2                 1.025082e+01 1.088399
```

Here a value of 1 represents that there is no change. However, a value greater than 1 represents an increase and value less than 1 represents a decrease.

```r
head(probability.table <- fitted(multinom.fit))
```

```
##              3            1             2
## 1 1.000000e+00 9.313905e-34 5.413725e-110
## 2 1.723581e-69 1.000000e+00  2.176581e-33
## 3 1.000000e+00 5.565112e-66  2.410656e-34
## 4 2.275384e-64 1.000000e+00  1.186565e-35
## 5 2.714520e-37 6.073444e-72  1.000000e+00
## 6 1.000000e+00 6.791982e-12  1.374622e-93
```

The table above indicates that the probability of the 1st obs being Class 2 is 100 %, being Class 1 is 0 % and being Class 3 is 0 % and so on with other obs.

We will now check the model accuracy by building classification table. So let us first build the classification table for training data set and calculate the model accuracy.

7. The Prediction

```r
# Predicting the values for train dataset
train$precticed <- predict(multinom.fit, newdata = train, "class")

# Building classification table
ctable <- table(train$Classes, train$precticed)
ctable
```

```
##
##      3  1  2
##   3 15  0  0
##   1  0 14  0
##   2  0  0 24
```

```r
# Calculating accuracy - sum of diagonal elements divided by total obs
round((sum(diag(ctable))/sum(ctable))*100,2)
```

```
## [1] 100
```

Accuracy in training dataset is 100% which is perfect. We now repeat the above on the unseen dataset that tests dataset.

```
# Predicting the values for test dataset
test$precticed <- predict(multinom.fit, newdata = test, "class")

# Building classification table
ctable <- table(test$Classes, test$precticed)
ctable
```

```
##
##       3  1  2
##   1   0  6  0
##   2   3 12 56
##   3  47  0  1
```

Our model perfectly classified class 1 data points, misclassified 6 out of 71 data points on class 2 and misclassified 2 out of 48 data points on class 3.

```
accuracy <- round(mean(test$Classes == test$precticed)*100, 2)
accuracy
```

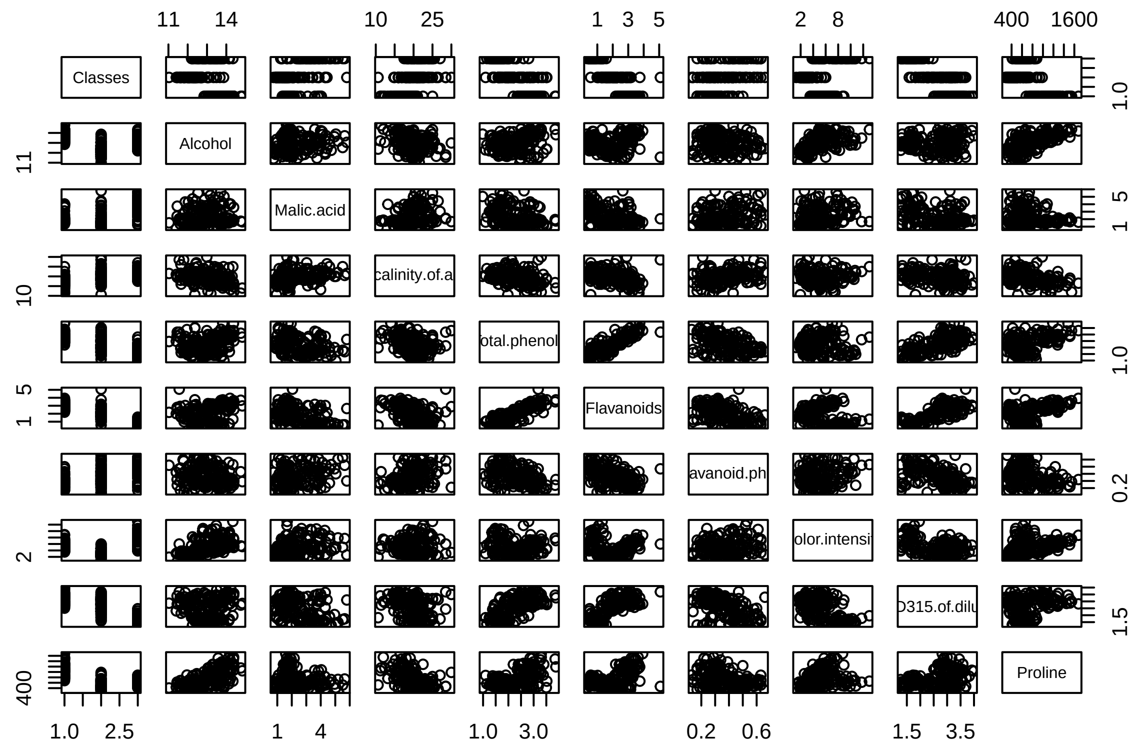
```
## [1] 87.2
```

Our Multinomial Logistic Regreesion model prediction accuracy is 90.4 % which is very good.

8. Improving the prediction accuracy

Let's see if we can improve the prediction accuracy of our model by transforming the predictor variables.

```
# Plotting the pairs plot of the data
pairs(Classes ~ Alcohol+Malic.acid+Alcalinity.of.ash+Total.phenols+Flavanoids+Nonflavanoid.phenols+Colo
```



Using powerTransform() to do a BoxCox on the predictor variables.

```
summary(powerTransform(cbind(Alcohol,Malic.acid,Alcalinity.of.ash,Total.phenols,Flavanoids,Nonflavanoid
```

```
## bcPower Transformations to Multinormality
##                              Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
## Alcohol                         1.6910        1.00      -0.2896       3.6717
## Malic.acid                     -0.2298        0.00      -0.5359       0.0763
## Alcalinity.of.ash               0.4992        1.00      -0.0676       1.0660
## Total.phenols                   0.8412        1.00       0.4554       1.2270
## Flavanoids                      0.7781        0.78       0.5799       0.9763
## Nonflavanoid.phenols            0.5078        0.50       0.1371       0.8785
## Color.intensity                 0.0087        0.00      -0.2327       0.2500
## OD280.OD315.of.diluted.wines    0.7613        1.00       0.3534       1.1693
## Proline                         0.2780        0.00      -0.0419       0.5979
##
## Likelihood ratio test that transformation parameters are equal to 0
##  (all log transformations)
##                                           LRT df       pval
## LR test, lambda = (0 0 0 0 0 0 0 0 0) 99.54269  9 < 2.22e-16
##
## Likelihood ratio test that no transformations are needed
##                                           LRT df       pval
## LR test, lambda = (1 1 1 1 1 1 1 1 1) 152.9881  9 < 2.22e-16
```

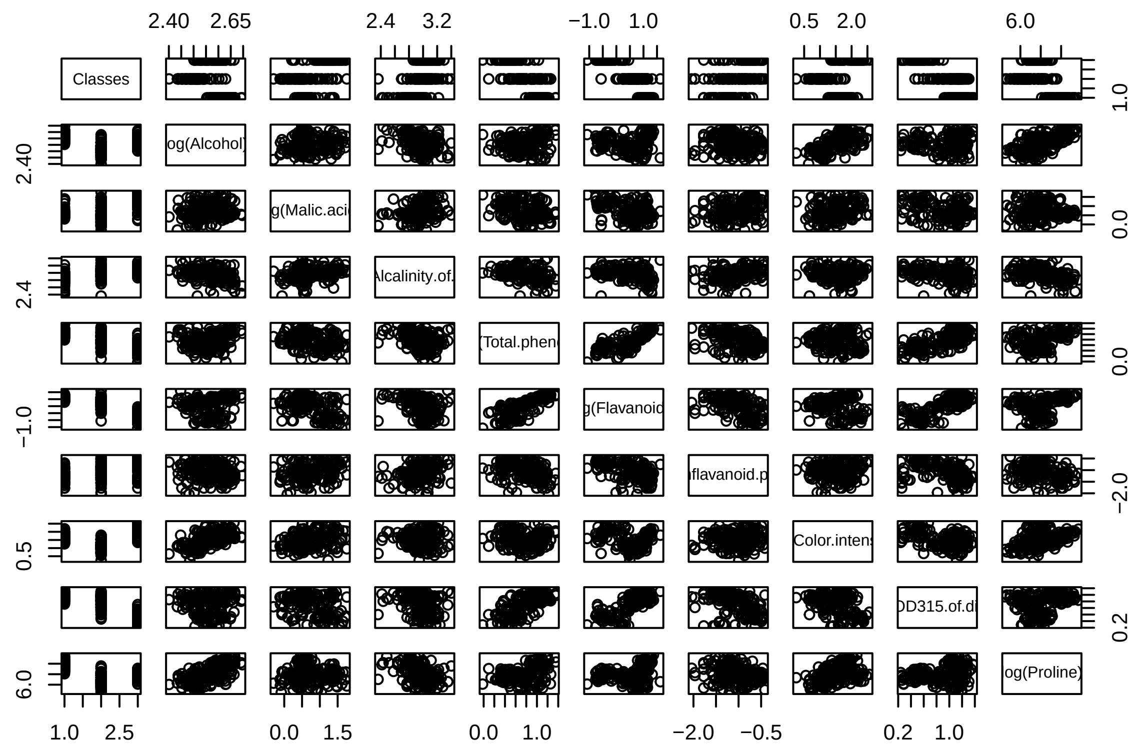
Most of the data is scruntched towards 0, So, Let’s Log transform all the predictors.

Now, an inverseResponsePlot:

```
multinom.fit_trns <- multinom (Classes ~ log(Alcohol)+log(Malic.acid)+log(Alcalinity.of.ash)+log(Total.p
```

```
## # weights:  33 (20 variable)
## initial  value 58.226451
## iter  10 value 0.990258
## iter  20 value 0.075524
## iter  30 value 0.024570
## iter  40 value 0.013693
## iter  50 value 0.003464
## iter  60 value 0.002057
## iter  70 value 0.001766
## iter  80 value 0.001021
## iter  90 value 0.000783
## iter 100 value 0.000706
## final  value 0.000706
## stopped after 100 iterations
```

```
pairs(Classes ~ log(Alcohol)+log(Malic.acid)+log(Alcalinity.of.ash)+log(Total.phenols)+log(Flavanoids)+l
```



well, we can see that we've gotten a slight improvement on couple predictors.

```r
summary(multinom.fit_trns)
```

```
## Call:
## multinom(formula = Classes ~ log(Alcohol) + log(Malic.acid) +
##     log(Alcalinity.of.ash) + log(Total.phenols) + log(Flavanoids) +
##     log(Nonflavanoid.phenols) + log(Color.intensity) + log(OD280.OD315.of.diluted.wines) +
##     log(Proline), data = train)
##
## Coefficients:
##   (Intercept) log(Alcohol) log(Malic.acid) log(Alcalinity.of.ash)
## 1   -8.099214    -7.386975      -0.5832012              -20.92072
## 2   14.134507    20.735869      -8.5330533               13.30889
##   log(Total.phenols) log(Flavanoids) log(Nonflavanoid.phenols)
## 1           10.51247        43.93748                 15.305104
## 2          -11.13368        -2.77007                 -5.718179
##   log(Color.intensity) log(OD280.OD315.of.diluted.wines) log(Proline)
## 1             9.147734                          30.40314    4.3414234
## 2           -71.279620                           4.94312    0.2077684
##
## Std. Errors:
##   (Intercept) log(Alcohol) log(Malic.acid) log(Alcalinity.of.ash)
## 1    4161.558    10023.745        1559.354               1531.782
## 2    5258.726     5467.375        1188.532               2263.408
##   log(Total.phenols) log(Flavanoids) log(Nonflavanoid.phenols)
## 1          11045.339       10049.059                  2435.800
## 2           6813.579        2577.903                  2356.943
##   log(Color.intensity) log(OD280.OD315.of.diluted.wines) log(Proline)
## 1             3449.120                          6184.116     3629.502
## 2             5378.155                          5142.385     3402.657
##
## Residual Deviance: 0.001411244
## AIC: 40.00141
```

9. The Prediction of the new model

```r
# Predicting the values for train dataset
train$precticed <- predict(multinom.fit_trns, newdata = train, "class")

# Building classification table
ctable <- table(train$Classes, train$precticed)
ctable
```

```
##
##      3  1  2
##   3 15  0  0
##   1  0 14  0
##   2  0  0 24
```

100% Training Prediction rate. Perfect !

```r
# Predicting the values for test dataset
test$precticed <- predict(multinom.fit_trns, newdata = test, "class")

# Building classification table
ctable <- table(test$Classes, test$precticed)
ctable
```

```
##
##       3  1  2
##    1  0  6  0
##    2  2 10 59
##    3 44  0  4
```

```r
accuracy <- round(mean(test$Classes == test$precticed)*100,)
accuracy
```

```
## [1] 87
```

The log transformation of the predictor variables did a good job on improving the prediction accuracy of our model, bringing it up from 90.3% to 97.6% which is an excellent accuracy rate.

Conclusion

The purpose of the project was develop a multinomial regression analysis model that would use the alcohol level, malic acid, alkalinity of ash, the total phenol's, the flavanoids, the nonflavoid phenols, the color intensity, the OD280 OD315 of diluted wine, hue and proline to predict the class of wine.

Before removing "Hue", "Magnesium", "Proanthocyanins" and "Ash", we found that our model was consistently misclassifying class two and three , which was surprising because we thought that the classification would be more evenly misclassified.

Once we selected our final predictors, we found that the model was able to predict the class of wine with a consistant accuracy between 80-90%, only missclassifying class three which is an improvement. Out of curiosity, we transformed the multinomial regression analysis model which ended up improving our accuracy to over 95%.

Limitations

Like all models, our model was not perfect and definitely had its limitations. The data was very limited so we were not able to show the accuracy between wines produced in different regions and if that had an impact. In the future, we would use more data to train and compare. And potentially add or replace different variables.