Proyecto Vinos

Claudia Meneses

31/3/2019

Introducción

La cata de vinos y su clasificación es tan antigua como su producción. Por ejemplo, Aristóteles propuso una degustación sensorial definida por los cuatro elementos (aire, agua, fuego y tierra) profundizada por la noble romana Lucrecia. Más adelante, en el siglo XIV se formalizó su metodología. En la actualidad, los catadores de vinos modernos y profesionales (como sommeliers o compradores para minoristas) utilizan una terminología especializada en constante evolución que se utiliza para describir la gama de sabores, aromas y características generales percibidos de un vino.

Los ácidos en el vino son un componente importante tanto en la vinificación como en el producto terminado del vino. Están presentes tanto en las uvas como en el vino, y tienen influencias directas en el color, el equilibrio y el sabor del vino, así como el crecimiento y la vitalidad de la levadura durante la fermentación y la protección del vino de las bacterias. La medida de la cantidad de acidez en el vino se conoce como “acidez valorable” o “acidez total” y se refiere a la prueba que produce el total de todos los ácidos presentes, mientras que la concentración de la acidez se mide según el pH. En general, cuanto más bajo es el pH, mayor es la acidez en el vino. Sin embargo, no existe una conexión directa entre la acidez total y el pH (es posible encontrar vinos con un pH alto para el vino y una alta acidez).

Se encuentran tres ácidos primarios en las uvas de vino: tartárico, málico y cítrico. Durante el curso de la vinificación y en los vinos terminados, los ácidos acético, butírico, láctico y succínico pueden desempeñar un papel importante. La mayoría de los ácidos relacionados con el vino son ácidos fijos, con la notable excepción del ácido acético, que se encuentra principalmente en el vinagre, que es volátil y puede contribuir a la falla del vino conocida como acidez volátil. A veces, se utilizan ácidos adicionales, como los ácidos ascórbico, sórbico y sulfúrico, en la vinificación.

A continuación, se presentan los datos de 178 vinos con 13 variables de sus propiedades químicas. El propósito del análisis es presentar un resumen de los datos con información relevante, después clasificarlos y por último, tratar de predecir su clasificación.

Las variables químicas que se consideran son:

1. **Tipo** : Clasificación dada que se intentará predecir con base en una muestra.
2. **Ácido málico (Malic_acid)** : El ácido málico, junto con el ácido tartárico, es uno de los principales ácidos orgánicos que se encuentran en las uvas para vino. Se encuentra en casi todas las plantas de frutas y bayas, pero con mayor frecuencia se asocia con manzanas verdes (verdes), el sabor que más fácilmente proyecta en el vino. Su concentración varía según la variedad de uva, con algunas variedades, como Barbera, Carignan y Sylvaner, que se encuentran naturalmente en niveles altos. La pérdida respiratoria de ácido málico es más pronunciada en climas más cálidos. Cuando todo el ácido málico se consume en la uva, se considera “demasiado maduro” o senescente. En general, los vinos tintos se someten más a menudo a través de la fermentación maloláctica o MLF que los blancos, lo que significa una mayor probabilidad de encontrar ácido málico en los vinos blancos.
3. **Ceniza (Ash)** : La ceniza es una suma de sustancias, que permanecen después de la incineración del residuo de evaporación, por lo que los elementos principales son los metales alcalinos y los metales alcalinotérreos (por ejemplo, potasio, calcio, magnesio, sodio) y fósforo no metálico. Elementos traza como manganeso, zinc, cobre y hierro también están representados. El contenido de ceniza de los productos puede determinarse gravimétricamente o puede calcularse mediante el análisis de cationes y aniones. El contenido de ceniza en el vino se encuentra normalmente entre 1,3 y 3,5 mg / ly está fuertemente influenciado por el balance de agua de la vid. Todas las sustancias relevantes se absorben durante la maduración de la uva a través del suelo, donde los minerales son la mayor parte de la ceniza. Debido a la relación entre las cenizas y el contenido de minerales y oligoelementos, es posible evaluar la calidad del vino con este parámetro. La relación entre la ceniza y la calidad del producto conduce a regulaciones legales para el contenido mínimo de ceniza, por ejemplo, para el vinagre, por lo que el análisis de la ceniza también es importante para la evaluación de las especificaciones del producto.

4. **Alcalinidad de la ceniza (Ash_Alcalinity)** : La alcalinidad de la ceniza se define como la suma de cationes, aparte del ion amonio, combinada con los ácidos orgánicos en el vino. La alcalinidad de las cenizas se expresará en miliequivalentes por litro o en gramos por litro de carbonato de potasio.
5. **Magnesio (Magnesium)** : El magnesio es el elemento químico de símbolo Mg y número atómico 12. Es el séptimo elemento en abundancia en el orden del 2 % de la corteza terrestre y el tercero más abundante disuelto en el agua de mar. El ion magnesio es esencial para todas las células vivas. El metal puro no se encuentra en la naturaleza.
6. **Contenido fenólico (Total_Phenoles)** : El contenido fenólico en el vino se refiere a los compuestos fenólicos (fenol natural y polifenoles) en el vino, que incluyen un gran grupo de varios cientos de compuestos químicos que afectan el sabor, el color y la sensación en boca del vino. Estos compuestos incluyen ácidos fenólicos, estilbenoides, flavonoles, dihidroflavonoles, antocianinas, monómeros de flavanol (catequinas) y polímeros de flavanol (proantocianidinas). Este gran grupo de fenoles naturales se puede separar ampliamente en dos categorías, flavonoides y no flavonoides. Los flavonoides incluyen las antocianinas y los taninos que contribuyen al color y la sensación en boca del vino. Los no flavonoides incluyen los estilbenoides como el resveratrol y los ácidos fenólicos como los ácidos benzoico, cafeico y cinnámico. El benzaldehído (vainillina) y el ácido benzoico (ácidos vanílico y gálico) son los compuestos fenólicos que se saben más en los vinos. Las antocianinas son responsables de la pigmentación del vino tinto y están presentes en proporción al color del vino.
7. **Flavanoides** : Es un grupo de compuestos fenólicos que incluyen las antocianinas y los taninos que contribuyen al color y la sensación en boca del vino.
8. **Fenoles no flavanoides (Non_Flavanoides_Phenoles)** : Es un grupo de compuestos fenólicos que incluyen los estilbenoides como el resveratrol y los ácidos fenólicos como los ácidos benzoico, cafeico y cinnámico.
9. **Proantocianidinas (Proanthocyanins)** : Las proantocianidinas son una clase de polifenoles que se encuentran en una variedad de plantas. Químicamente, son flavonoides oligoméricos. Muchos son oligómeros de catequina y epicatequina y sus ésteres de ácido gálico. Las proantocianidinas juegan un papel importante en el vino con la capacidad de unir proteínas salivales, estos taninos condensados influyen fuertemente en la percepción de la astringencia del vino. Estos compuestos suelen estar presentes en niveles de 300 mg / L1 en el vino tinto, aunque el procesamiento enológico puede afectar las concentraciones finales.
10. **Intensidad de color (Color_intensity)** : Variable que mide que tan intenso es el color del vino.
11. **Tono (Hue)** : Variable que mide el tono del vino. El tono en el vino se ve parcialmente afectado por el nivel de pH del vino.
12. **OD280.OD315** : Concentración de esas proteínas en vinos diluidos.
13. **Prolina (Proline)** : La prolina (símbolo Pro o P) es un aminoácido proteinogénico que se utiliza en la biosíntesis de proteínas.
14. **Variable 14 (var 14)** : Sin información.

Análisis

Resumen estadístico del conjunto de datos

```
summary(datos)
```

```
##    Malic_acid          Ash        Ash_Alcalinity    Magnesium
##  Min.   :11.03   Min.   :0.740   Min.   :1.360   Min.   :10.60
##  1st Qu.:12.36   1st Qu.:1.603   1st Qu.:2.210   1st Qu.:17.20
##  Median :13.05   Median :1.865   Median :2.360   Median :19.50
##  Mean   :13.00   Mean   :2.336   Mean   :2.367   Mean   :19.49
##  3rd Qu.:13.68   3rd Qu.:3.083   3rd Qu.:2.558   3rd Qu.:21.50
##  Max.   :14.83   Max.   :5.800   Max.   :3.230   Max.   :30.00
##  Total_Phenoles    Flavanoides    Non_Flavanoides_Phenoles
##  Min.   : 70.00   Min.   :0.980   Min.   :0.340
##  1st Qu.: 88.00   1st Qu.:1.742   1st Qu.:1.205
##  Median : 98.00   Median :2.355   Median :2.135
##  Mean   : 99.74   Mean   :2.295   Mean   :2.029
##  3rd Qu.:107.00   3rd Qu.:2.800   3rd Qu.:2.875
##  Max.   :162.00   Max.   :3.880   Max.   :5.080
##  Proanthocyanins  Color_intensity      Hue           OD280.OD315
##  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
##     Proline          var14
##  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
```

Correlaciones entre las variables

Podemos ver que, una vez que los los *flavanoides* y los *fenoles no flavanoides* son parte del *contenido fenólico* (Total_Phenoles) estas variables están fuertemente relacionadas entre sí. También se puede observar que los *flavanoides* y los *fenoles no flavanoides* junto con la *prolina* juegan un papel importante en la *intensidad del color*.

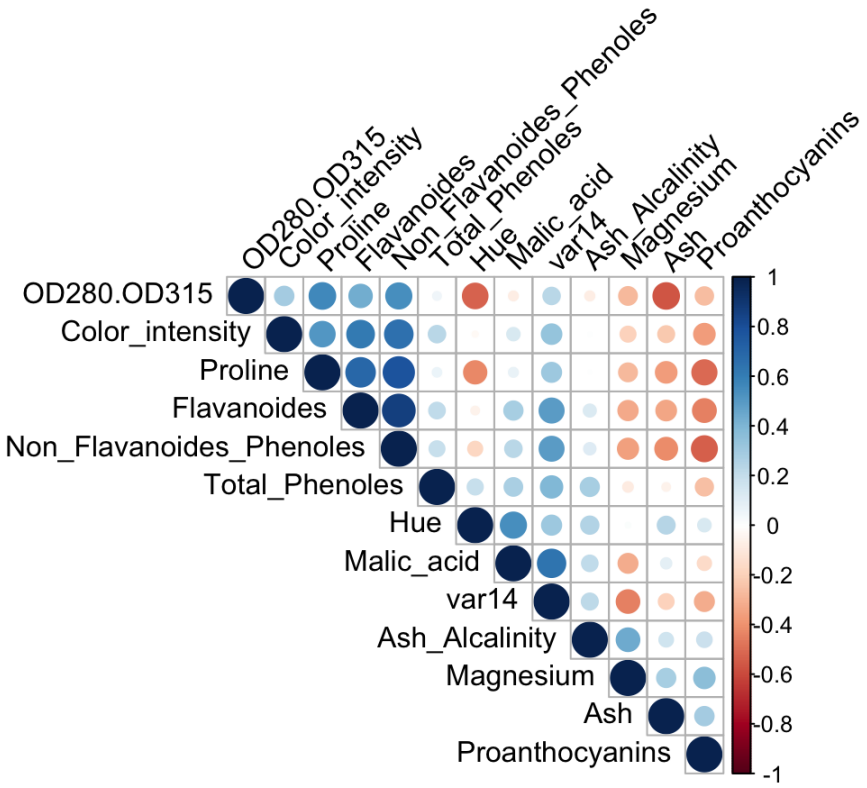
Las proteínas *OD280.OD315* parecen estar relacionadas (negativamente) con el *tono* (hue) y la *ceniza* (Ash).

```
cor_matrix <- cor(datos, method = "pearson", use = "complete.obs")
cor_matrix
```

```
##                          Malic_acid         Ash Ash_Alcalinity
## Malic_acid                1.00000000  0.09439694    0.211544596
## Ash                       0.09439694  1.00000000    0.164045470
## Ash_Alcalinity            0.21154460  0.16404547    1.000000000
## Magnesium                -0.31023514  0.28850040    0.443367187
## Total_Phenoles            0.27079823 -0.05457510    0.286586691
## Flavanoides               0.28910112 -0.33516700    0.128979538
## Non_Flavanoides_Phenoles  0.23681493 -0.41100659    0.115077279
## Proanthocyanins          -0.15592947  0.29297713    0.186230446
## Color_intensity           0.13669791 -0.22074619    0.009651935
## Hue                       0.54636420  0.24898534    0.258887259
## OD280.OD315              -0.07174720 -0.56129569   -0.074666889
## Proline                   0.07234319 -0.36871043    0.003911231
## var14                     0.64372004 -0.19201056    0.223626264
##                           Magnesium Total_Phenoles Flavanoides
## Malic_acid               -0.31023514     0.27079823  0.28910112
## Ash                       0.28850040    -0.05457510 -0.33516700
## Ash_Alcalinity            0.44336719     0.28658669  0.12897954
## Magnesium                 1.00000000    -0.08333309 -0.32111332
## Total_Phenoles           -0.08333309     1.00000000  0.21440123
## Flavanoides              -0.32111332     0.21440123  1.00000000
## Non_Flavanoides_Phenoles -0.35136986     0.19578377  0.86456350
## Proanthocyanins           0.36192172    -0.25629405 -0.44993530
## Color_intensity          -0.19732684     0.23644061  0.61241308
## Hue                       0.01873198     0.19995001 -0.05513642
## OD280.OD315              -0.27395522     0.05539820  0.43368134
## Proline                  -0.27676855     0.06600394  0.69994936
## var14                    -0.44059693     0.39335085  0.49811488
##                          Non_Flavanoides_Phenoles Proanthocyanins
## Malic_acid                              0.2368149      -0.1559295
## Ash                                    -0.4110066       0.2929771
## Ash_Alcalinity                          0.1150773       0.1862304
## Magnesium                              -0.3513699       0.3619217
## Total_Phenoles                          0.1957838      -0.2562940
## Flavanoides                             0.8645635      -0.4499353
## Non_Flavanoides_Phenoles                1.0000000      -0.5378996
## Proanthocyanins                        -0.5378996       1.0000000
## Color_intensity                         0.6526918      -0.3658451
## Hue                                    -0.1723794       0.1390570
## OD280.OD315                             0.5434786      -0.2626396
## Proline                                 0.7871939      -0.5032696
## var14                                   0.4941931      -0.3113852
##                          Color_intensity         Hue OD280.OD315
## Malic_acid                 0.136697912  0.54636420 -0.07174720
## Ash                       -0.220746187  0.24898534 -0.56129569
## Ash_Alcalinity             0.009651935  0.25888726 -0.07466689
## Magnesium                 -0.197326836  0.01873198 -0.27395522
## Total_Phenoles             0.236440610  0.19995001  0.05539820
## Flavanoides                0.612413084 -0.05513642  0.43368134
## Non_Flavanoides_Phenoles   0.652691769 -0.17237940  0.54347857
## Proanthocyanins           -0.365845099  0.13905701 -0.26263963
## Color_intensity            1.000000000 -0.02524993  0.29554425
## Hue                       -0.025249931  1.00000000 -0.52181319
## OD280.OD315                0.295544253 -0.52181319  1.00000000
## Proline                    0.519067096 -0.42881494  0.56546829
## var14                      0.330416700  0.31610011  0.23618345
##                             Proline      var14
## Malic_acid                0.072343187  0.6437200
## Ash                      -0.368710428 -0.1920106
```

```
## Ash_Alcalinity             0.003911231  0.2236263
## Magnesium                 -0.276768549 -0.4405969
## Total_Phenoles             0.066003936  0.3933508
## Flavanoides                0.699949365  0.4981149
## Non_Flavanoides_Phenoles   0.787193902  0.4941931
## Proanthocyanins           -0.503269596 -0.3113852
## Color_intensity            0.519067096  0.3304167
## Hue                       -0.428814942  0.3161001
## OD280.OD315                0.565468293  0.2361834
## Proline                    1.000000000  0.3127611
## var14                      0.312761075  1.0000000
```

```
corrplot(cor_matrix, type = "upper", order = "hclust", tl.col = "black", tl.srt = 45)
```

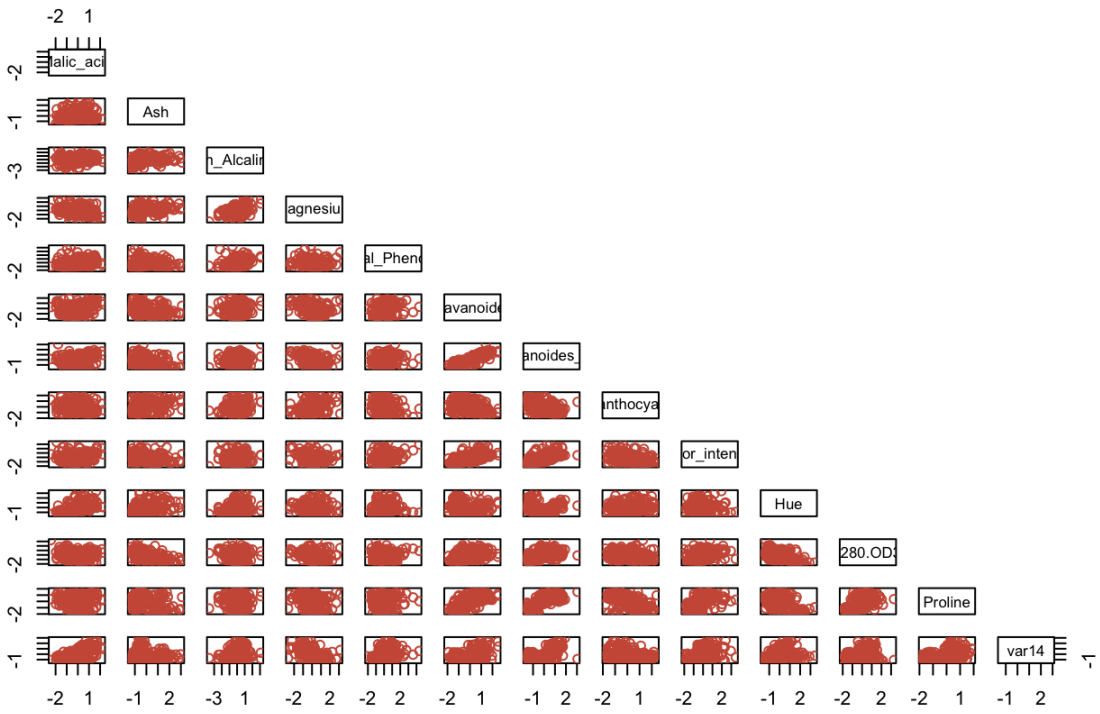


Estandarización de los datos

```
scaled_datos <- scale(datos)
```

Gráficas de dispersión de las variables

```
pairs(scaled_datos,main="Dispersión",col="coral3",upper.panel = NULL)
```



Análisis por Componentes Principales

```
acp<-prcomp(scaled_datos,center = TRUE, scale. = TRUE)
```

Proporción de la varianza y varianza acumulada

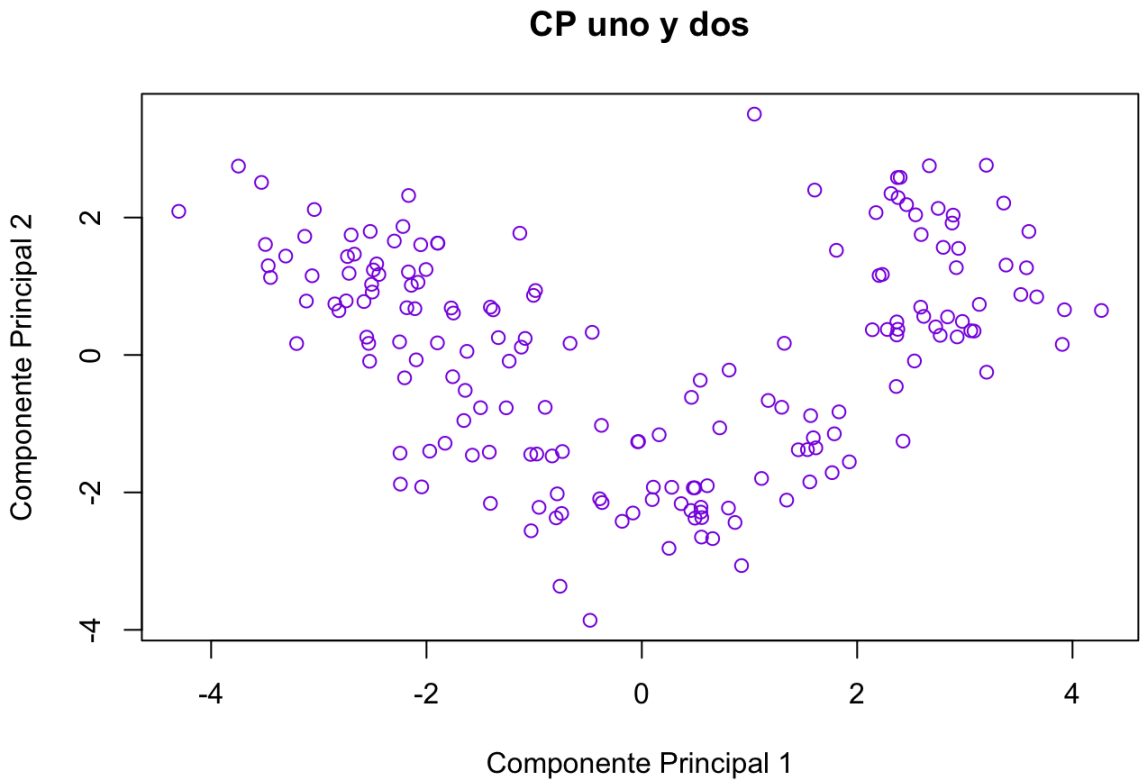
Con las primeras 2 componentes principales se obtine el 55% de la varianza, además necesitamos las primeras 5 componentes principales para obtener el 80% de la varianza.

```
summary(acp)
```

```
## Importance of components:
##                           PC1    PC2    PC3     PC4     PC5     PC6     PC7
## Standard deviation     2.169 1.5802 1.2025 0.95863 0.92370 0.80103 0.74231
## Proportion of Variance 0.362 0.1921 0.1112 0.07069 0.06563 0.04936 0.04239
## Cumulative Proportion  0.362 0.5541 0.6653 0.73599 0.80162 0.85098 0.89337
##                           PC8     PC9    PC10    PC11    PC12    PC13
## Standard deviation     0.59034 0.53748 0.5009 0.47517 0.41082 0.32152
## Proportion of Variance 0.02681 0.02222 0.0193 0.01737 0.01298 0.00795
## Cumulative Proportion  0.92018 0.94240 0.9617 0.97907 0.99205 1.00000
```

Graficando los datos (estadarizados) respecto a las 2 primeras componentes principales se observa que tienen forma “parabólica”.

```
CP1<-acp$x[,1]
CP2<-acp$x[,2]
plot(CP1,CP2,main="CP uno y dos",col="blueviolet", xlab = "Componente Principal 1", ylab = "Compone
nte Principal 2")
```



Representación bidimiensional de las 2 primeras componentes.

Se observa que las variables que más contribuyen a las 2 primeras componentes principales son:

1. Fenoles no flavanoides (Non_Flavanoides_Phenoles)
2. Flavanoides
3. Prolina (Proline)
4. Ácido málico (Malic_acid)
5. var14
6. Tono (Hue)

Las variables “mejor” representadas por la primera componente son:

1. Fenoles no flavanoides (Non_Flavanoides_Phenoles)
2. Magnesio (Magnesium)
3. Intensidad de color (Color_intensity)
4. Proantocianidinas (Proanthocyanins)

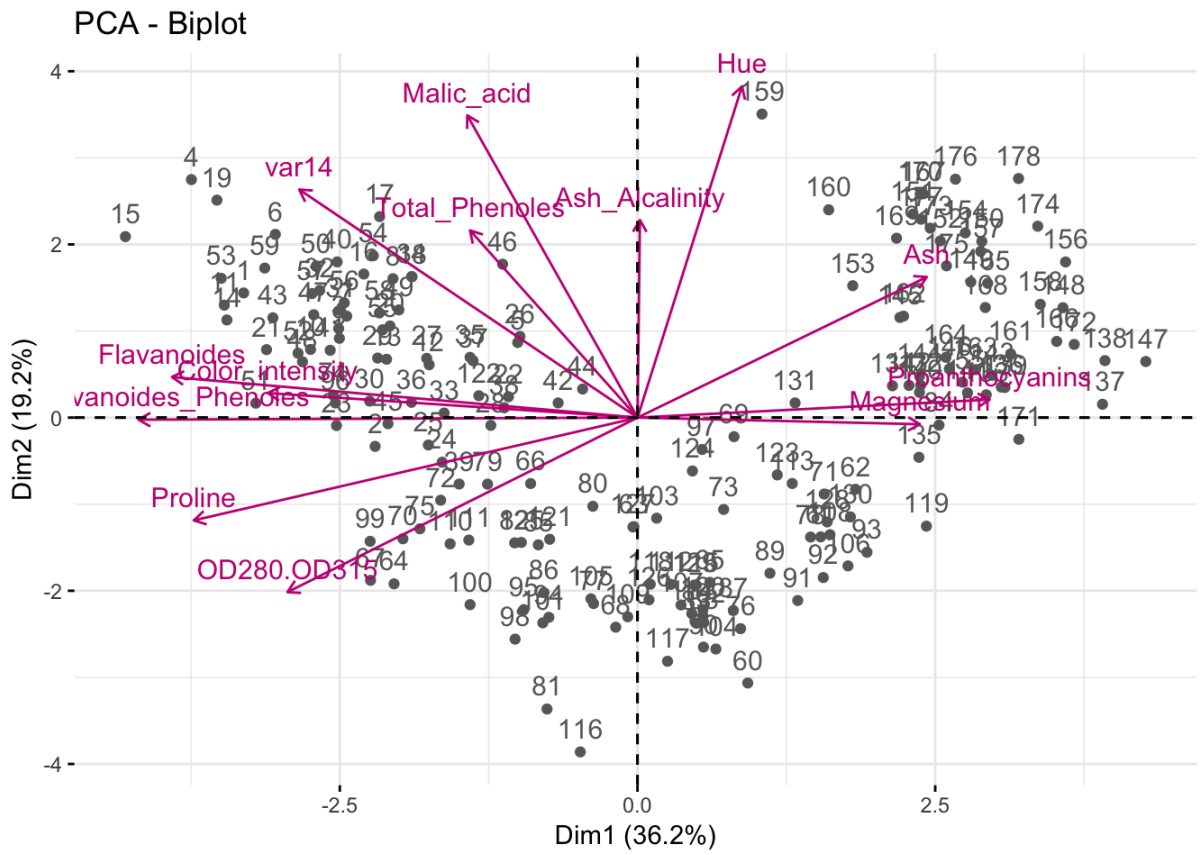
Las variables “mejor” representadas por la segunda componente son:

1. Ceniza alcalina (Ash_Alcalinity)
2. Tono (Huen)

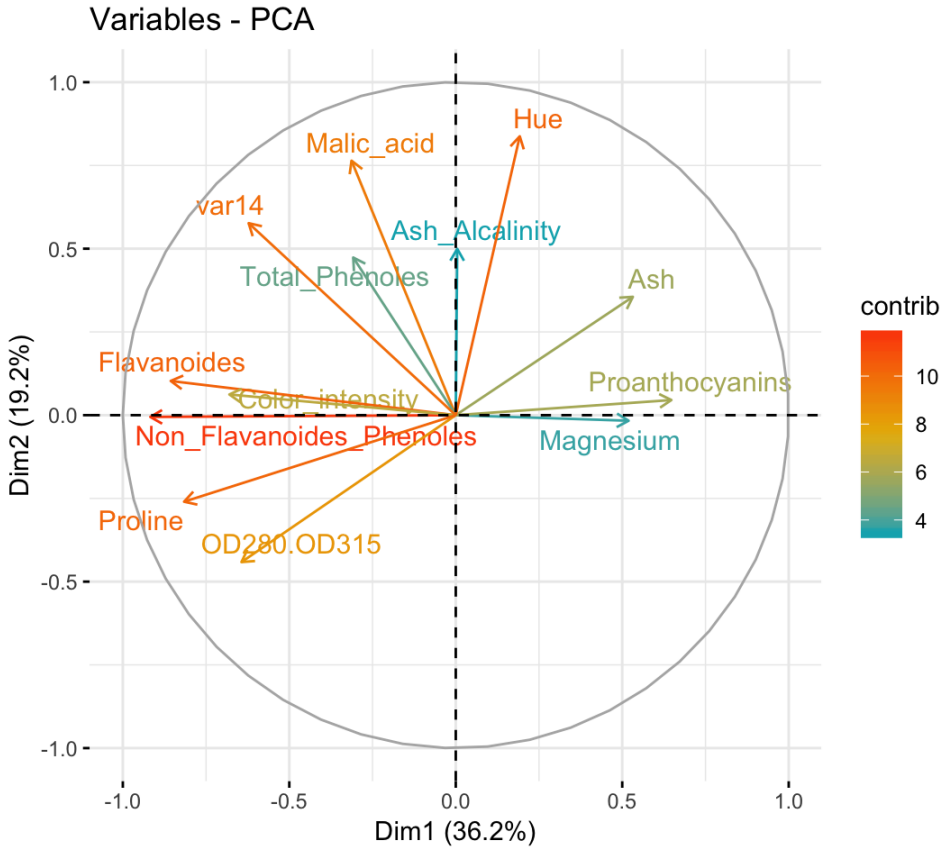
```
acp$rotation
```

```
##                                PC1          PC2         PC3         PC4
## Malic_acid               -0.144329395  0.483651548 -0.20738262  0.01785630
## Ash                       0.245187580  0.224930935  0.08901289 -0.53689028
## Ash_Alcalinity            0.002051061  0.316068814  0.62622390  0.21417556
## Magnesium                 0.239320405 -0.010590502  0.61208035 -0.06085941
## Total_Phenoles           -0.141992042  0.299634003  0.13075693  0.35179658
## Flavanoides              -0.394660845  0.065039512  0.14617896 -0.19806835
## Non_Flavanoides_Phenoles -0.422934297 -0.003359812  0.15068190 -0.15229479
## Proanthocyanins           0.298533103  0.028779488  0.17036816  0.20330102
## Color_intensity          -0.313429488  0.039301722  0.14945431 -0.39905653
## Hue                       0.088616705  0.529995672 -0.13730621 -0.06592568
## OD280.OD315              -0.296714564 -0.279235148  0.08522192  0.42777141
## Proline                  -0.376167411 -0.164496193  0.16600459 -0.18412074
## var14                    -0.286752227  0.364902832 -0.12674592  0.23207086
##                                 PC5          PC6         PC7          PC8
## Malic_acid               -0.26566365  0.21353865 -0.05639636  0.39613926
## Ash                       0.03521363  0.53681385  0.42052391  0.06582674
## Ash_Alcalinity           -0.14302547  0.15447466 -0.14917061 -0.17026002
## Magnesium                 0.06610294 -0.10082451 -0.28696914  0.42797018
## Total_Phenoles            0.72704851  0.03814394  0.32288330 -0.15636143
## Flavanoides              -0.14931841 -0.08412230 -0.02792498 -0.40593409
## Non_Flavanoides_Phenoles -0.10902584 -0.01892002 -0.06068521 -0.18724536
## Proanthocyanins          -0.50070298 -0.25859401  0.59544729 -0.23328465
## Color_intensity           0.13685982 -0.53379539  0.37213935  0.36822675
## Hue                      -0.07643678 -0.41864414 -0.22771214 -0.03379692
## OD280.OD315              -0.17361452  0.10598274  0.23207564  0.43662362
## Proline                  -0.10116099  0.26585107 -0.04476370 -0.07810789
## var14                    -0.15786880  0.11972557  0.07680450  0.12002267
##                                 PC9         PC10         PC11         PC12
## Malic_acid               -0.50861912  0.21160473  0.22591696 -0.26628645
## Ash                       0.07528304 -0.30907994 -0.07648554  0.12169604
## Ash_Alcalinity            0.30769445 -0.02712539  0.49869142 -0.04962237
## Magnesium                -0.20044931  0.05279942 -0.47931378 -0.05574287
## Total_Phenoles           -0.27140257  0.06787022 -0.07128891  0.06222011
## Flavanoides              -0.28603452 -0.32013135 -0.30434119 -0.30388245
## Non_Flavanoides_Phenoles -0.04957849 -0.16315051  0.02569409 -0.04289883
## Proanthocyanins          -0.19550132  0.21553507 -0.11689586  0.04235219
## Color_intensity           0.20914487  0.13418390  0.23736257 -0.09555303
## Hue                      -0.05621752 -0.29077518 -0.03183880  0.60422163
## OD280.OD315              -0.08582839 -0.52239889  0.04821201  0.25921400
## Proline                  -0.13722690  0.52370587 -0.04642330  0.60095872
## var14                     0.57578611  0.16211600 -0.53926983 -0.07940162
##                                PC13
## Malic_acid                0.01496997
## Ash                       0.02596375
## Ash_Alcalinity           -0.14121803
## Magnesium                 0.09168285
## Total_Phenoles            0.05677422
## Flavanoides              -0.46390791
## Non_Flavanoides_Phenoles  0.83225706
## Proanthocyanins           0.11403985
## Color_intensity          -0.11691707
## Hue                      -0.01199280
## OD280.OD315              -0.08988884
## Proline                  -0.15671813
## var14                     0.01444734
```

```
fviz_pca_biplot(acp, repel = FALSE, col.var = "mediumvioletred", col.ind = "#696969")
```




```
fviz_pca_var(acp, col.var = "contrib", gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),repel = TRUE )
```



Cluster análisis

K-means

Las observaciones se dividen en grupos K (en este caso k = 3)y se reorganizan para formar los grupos más cohesivos posibles de acuerdo con un criterio dado. El algoritmo K-means es el siguiente:

1. Selecciona K centroides (K filas elegidas al azar)
2. Asigna cada punto de datos a su centroide más cercano
3. Recalcula los centroides como el promedio de todos los puntos de datos en un grupo (es decir, los centroides son vectores medios de longitud p, donde p es el número de variables)
4. Asigna puntos de datos a sus centroides más cercanos
5. Continúa con los pasos 3 y 4 hasta que las observaciones no se reasignen o se alcance el número máximo de iteraciones (R usa 10 como valor predeterminado).

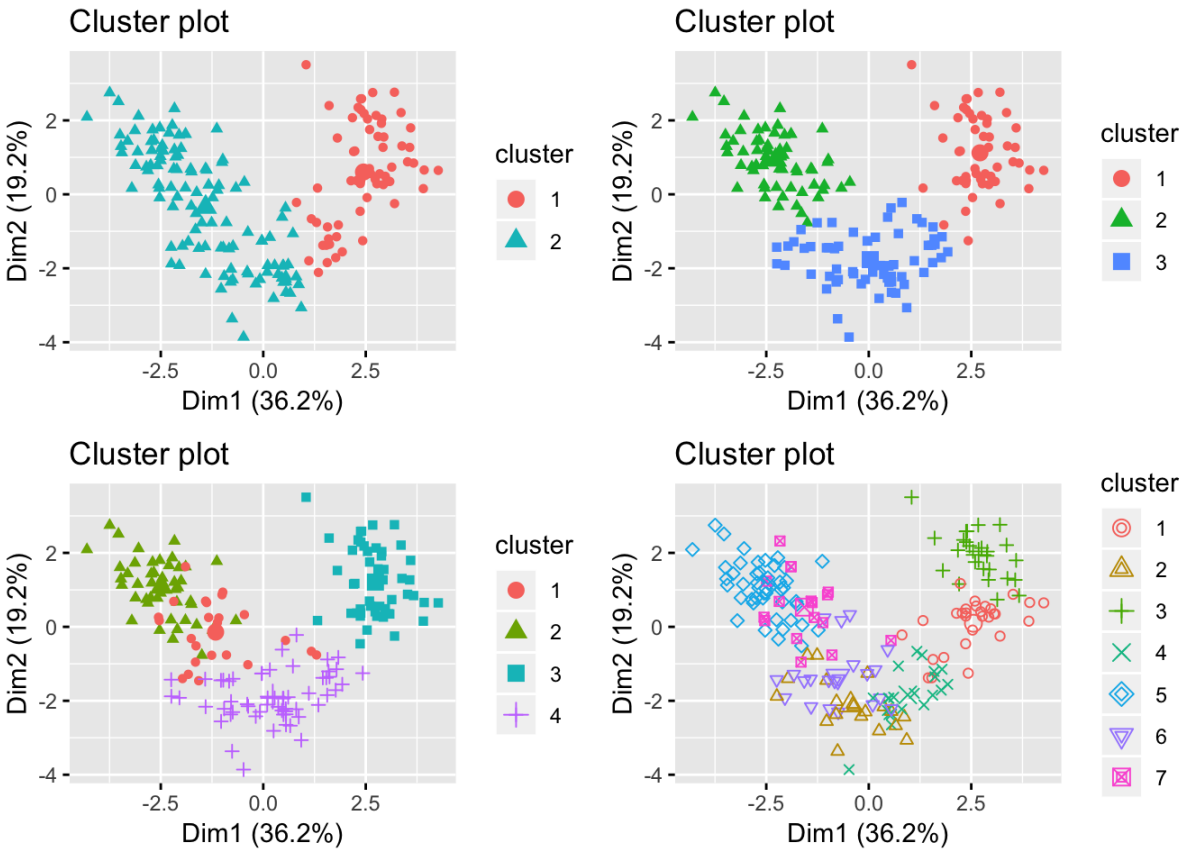
Nota : Las siguientes gráficas están respecto a las 2 primeras componentes.

```
km <- kmeans(scaled_datos, 3)
fviz_cluster(km, datos, ellipse.type = "norm")
```

Pruebo para distintas valores de k (k=2, k=3, k=4, k=7)

Observo que para k = 2 los clusters son muy claros (parece una agrupación intuitiva), de igual forma para k =3. Para k = 4 y k = 7 los clusters están menos claros.

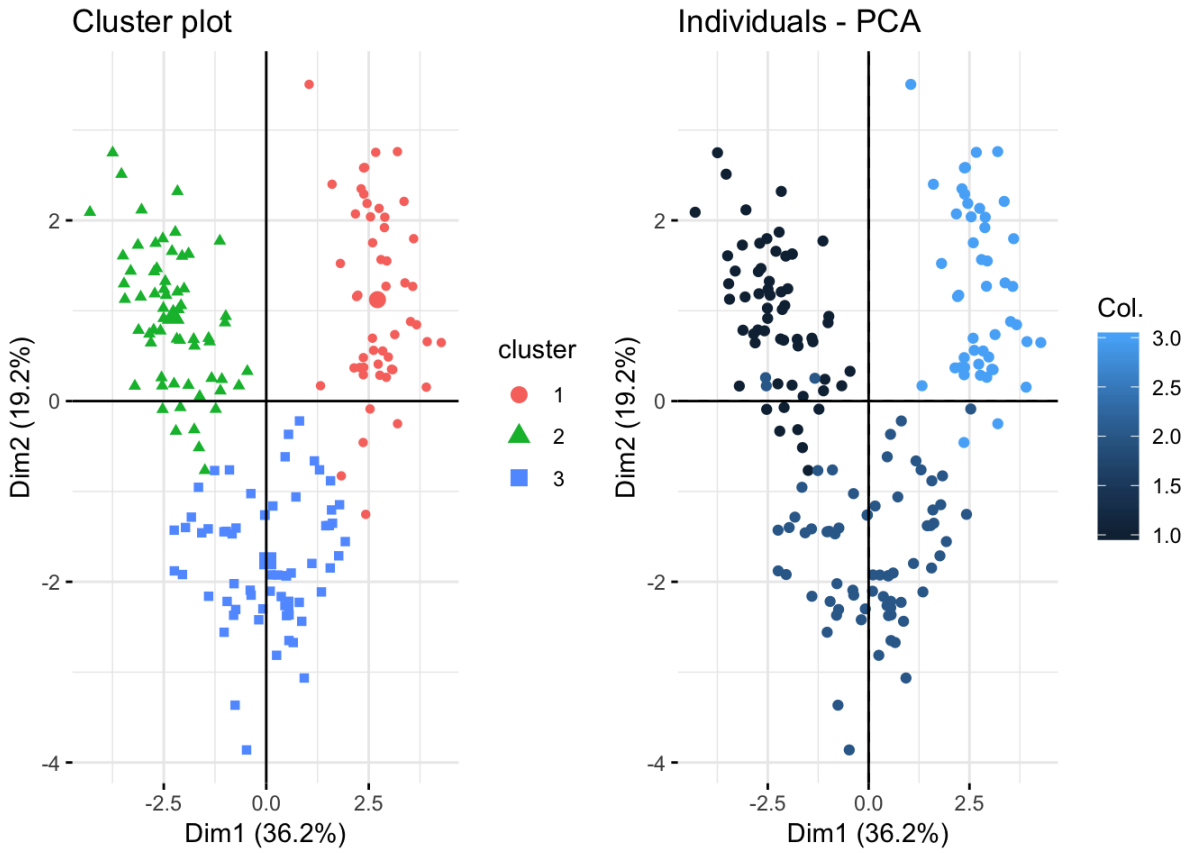
```
kmacp2 <- kmeans(scaled_datos ,2)
k2 <- fviz_cluster(kmacp2, scaled_datos, geom = "point", ellipse = FALSE)
kmacp3 <- kmeans(scaled_datos , 3)
k3 <- fviz_cluster(kmacp3, scaled_datos, geom = "point", ellipse = FALSE)
kmacp4 <- kmeans(scaled_datos ,4)
k4 <- fviz_cluster(kmacp4, scaled_datos, geom = "point", ellipse = FALSE)
kmacp7 <- kmeans(scaled_datos ,7)
k7 <- fviz_cluster(kmacp7, scaled_datos, geom = "point", ellipse = FALSE)
grid.arrange(k2, k3, k4,k7, nrow = 2)
```



Comparo la clasificación dada con la obtenida con k-means clustering

Observo que los clusters dados por el método de k-means (k = 3) se parecen bastante a la clasificación dada, con algunas excepciones en las “fronteras” de los clusters.

```
p1 <- fviz_cluster(kmacp3, scaled_datos, geom = "point", ellipse = FALSE,  ggtitle = "k-means clust
ering", ggtheme = theme_minimal()) + geom_hline(yintercept = 0) + geom_vline(xintercept = 0 )
p2 <- fviz_pca_ind(acp, ggtitle="Clasificación dada", repel = FALSE, col.ind = data$Tipo, geom = "p
oint", ggtheme = theme_minimal())  + geom_hline(yintercept = 0) + geom_vline(xintercept = 0 )
grid.arrange(p1, p2, nrow = 1)
```



KNN

El algoritmo KNN o k-vecinos más cercanos es uno de los algoritmos de aprendizaje automático más simples y es un ejemplo de aprendizaje basado en instancias, donde los datos nuevos se clasifican según las instancias etiquetadas y almacenadas. La distancia (generalmente la distancia euclidiana) entre los datos almacenados y la nueva instancia se calcula por medio de algún tipo de medida de similitud. Luego, este valor de similitud se usa para realizar un modelado predictivo. El modelado predictivo es una clasificación, la asignación de una etiqueta o una clase a la nueva instancia, o la regresión, la asignación de un valor a la nueva instancia.

A continuación se usará el algoritmo KNN para predecir los tipos de vinos.

Primero genero número aleatorios para tomar una muestra de aprox 80% de los datos.

```
data <- data.frame(data)
ran <- sample(1:nrow(data), 0.8 * nrow(data))
```

Normalizo los datos debido a que algunas columnas cuentan con escalas mayores a otras, el modelo considera que todas las columnas tienen la misma influencia para la clasificación, por lo que, para evitar que la magnitud de cualquier atributo influya más que los otros se normalizan los datos.

```
nor <-function(x) { (x - min(x) )/( max(x) - min(x))   }
data_norm <- as.data.frame(lapply(data[,c(2,3,4,5,6,7,8,9,10,11,12,13,14)], nor))
summary(data_norm)
```

```
##    Malic_acid          Ash         Ash_Alcalinity     Magnesium
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.3507   1st Qu.:0.1705   1st Qu.:0.4545   1st Qu.:0.3402
##  Median :0.5316   Median :0.2223   Median :0.5348   Median :0.4588
##  Mean   :0.5186   Mean   :0.3155   Mean   :0.5382   Mean   :0.4585
##  3rd Qu.:0.6967   3rd Qu.:0.4629   3rd Qu.:0.6404   3rd Qu.:0.5619
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##  Total_Phenoles    Flavanoides     Non_Flavanoides_Phenoles
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.1957   1st Qu.:0.2629   1st Qu.:0.1825
##  Median :0.3043   Median :0.4741   Median :0.3787
##  Mean   :0.3233   Mean   :0.4535   Mean   :0.3564
##  3rd Qu.:0.4022   3rd Qu.:0.6276   3rd Qu.:0.5348
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##  Proanthocyanins  Color_intensity       Hue          OD280.OD315
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.2642   1st Qu.:0.2650   1st Qu.:0.1655   1st Qu.:0.2459
##  Median :0.3962   Median :0.3612   Median :0.2910   Median :0.3943
##  Mean   :0.4375   Mean   :0.3725   Mean   :0.3224   Mean   :0.3882
##  3rd Qu.:0.5802   3rd Qu.:0.4858   3rd Qu.:0.4198   3rd Qu.:0.5203
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##     Proline           var14
##  Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.2445   1st Qu.:0.1587
##  Median :0.5531   Median :0.2821
##  Mean   :0.4915   Mean   :0.3344
##  3rd Qu.:0.6960   3rd Qu.:0.5043
##  Max.   :1.0000   Max.   :1.0000
```

Selecciono los datos que se usarán para “aprender”" y los que se usarán para “predecir”

```
data_train <- data_norm[ran,]
data_test <- data_norm[-ran,]
```

Quito la primer columna (Tipo) de los datos que se usarán para “aprender” porque serán el argumento de la función knn (de la librería “class”) que implementa el algoritmo.

```
data_target_category <- data[ran,1]
data_test_category <- data[-ran,1]
```

Implemento el algoritmo usando k = 13.

```
pr <- knn(data_train,data_test,cl=data_target_category,k=13)
```

Reviso en como fue la clasificación por categoria.

```
tab <- table(pr, data_test_category)
tab
```

```
##    data_test_category
## pr   1  2  3
##   1 10  0  0
##   2  0 12  0
##   3  0  1 13
```

Checo que tan acertado es el modelo dividiento las predicciones correctas entre el número total de predicciones

```
accuracy <- function(x){sum(diag(x)/(sum(rowSums(x)))) * 100}
accuracy(tab)
```

```
## [1] 97.22222
```

Conclusión

Con base en el análisis anterior, se puede concluir:

1. La varianza acumulada de las primeras dos compenentes (55%) no es muy alta, sin embargo funcionan muy acertadamente para predecir los tipos de vinos dados con los métodos k-means y kNN.
2. Usando el método k-means la clasificación obtenida de los tipos de vino es muy parecida a la clasificación dada.
3. Usando el método kNN para predecir los tipos de vinos se obtiene una clasificción muy acertada (95% - 100%) por lo que el método funciona bastante bien.