# Práctica 2. Limpieza y validación de los datos

Miembros del grupo: Víctor Camí Núñez

Índice

2. Análisis descriptivo del fichero
3. Objetivos a cumplir
4. Procesamiento del fichero
5. Aplicación de los modelos
6. Conclusiones
7. Recursos

# 1. Introducción

En primer lugar se va a realizar un análisis descriptivo de la estructura del fichero winequality-red.csv que contiene los datos del dataset sin entrar en detalles de su contenido.

En segundo lugar se va proceder a explicar los objetivos y preguntas que se quieren contestar empleando dicho dataset. Es importante definir claramente cuáles son las respuestas que queremos responder mediante este dataset dado que esto condicionará el procesado y limpieza de los datos.

En tercer se describirán los pasos realizados para el procesado y limpieza de los datos del dataset necesarios para poder contestar a las preguntas y objetivos planteados.

En cuarto lugar se explicará la implementación del modelo empleado para

Por último se presentarán las conclusiones extraídas en la realización de esta práctica.

## 2. Análisis descriptivo del fichero

El fichero se compone de 1599 registros con datos y un registro extra que corresponde a la cabecera del fichero. El formato del fichero es csv cuyo delimitador es la coma y todos los campos que contiene son numéricos.

El fichero se compone de 12 campos y sus descripciones (copiadas de las descripciones que se encuentran en el enlace del que se ha obtenido el dataset) son:

- fixed acidity: most acids involved with wine or fixed or nonvolatile (do not evaporate readily)
- volatile acidity: the amount of acetic acid in wine, which at too high of levels can lead to an unpleasant, vinegar taste
- citric acid: found in small quantities, citric acid can add 'freshness' and flavor to wines
- residual sugar: the amount of sugar remaining after fermentation stops, it's rare to find wines with less than 1 gram/liter and wines with greater than 45 grams/liter are considered sweet
- chlorides: the amount of salt in the wine
- free sulfur dioxide: the free form of SO2 exists in equilibrium between molecular SO2 (as a dissolved gas) and bisulfite ion; it prevents microbial growth and the oxidation of wine
- total sulfur dioxide: amount of free and bound forms of S02; in low concentrations, SO2 is mostly undetectable in wine, but at free SO2 concentrations over 50 ppm, SO2 becomes evident in the nose and taste of wine
- density: the density of water is close to that of water depending on the percent alcohol and sugar content
- pH: describes how acidic or basic a wine is on a scale from 0 (very acidic) to 14 (very basic); most wines are between 3-4 on the pH scale
- sulphates: a wine additive which can contribute to sulfur dioxide gas (S02) levels, wich acts as an antimicrobial and antioxidant
- alcohol: the percent alcohol content of the wine
- quality: output variable (based on sensory data, score between 0 and 10). Los registros se dividen de la siguiente forma en cuanto al valor que toman:
    - 0: aparece 0 veces
    - 1: aparece 0 veces
    - 2: aparece cero veces
    - 3: aparece 10 veces
    - 4: aparece 52 veces
    - 5: aparece 680 veces
    - 6: aparece 637 veces
    - 7: aparece 198 veces
    - 8: aparece 18 veces
    - 9: aparece 0 veces
    - 10: aparece 0 veces

## 3. Objetivos a cumplir

Como ya se ha comentado, este dataset contiene la composición química de diferentes vinos así como una nota del 1 al 10 que representa su calidad.

El hecho que tengamos el campo quality nos permitiría emplear dicho dataset para generar un pequeño programa para la creación de un modelo que clasificase de forma automática los vinos en base a su composición química.

Cabe resaltar, que en la página web de la que se ha obtenido el dataset se comenta que por motivos de confidencialidad se han eliminado el tipo de uva, la marca del vino y el precio de venta pero estos indicadores han influenciado en la obtención de una nota u otra en el campo quality del dataset. Esto puede tener un impacto en el modelo pero dado que no hay forma de recuperar los campos eliminados por temas de confidencialidad se hará la hipótesis que dichos campos no han afectado de forma significativa al campo quality.

Con esto, los objetivos de esta práctica serán:

- Analizar el dataset para detectar registros incorrectos.
- Seleccionar aquellos campos que se consideren más relevantes para la generación del modelo de clasificación.
- Se procederá a la limpieza y procesado de los datos que permitan su uso en el modelo de clasificación.
- Se creará un modelo de clasificación sencillo a base de reglas de clasificación que permitirán asociar una calidad a los vinos en base de sus composiciones químicas.

## 4. Procesamiento del fichero

Selección de los campos

De todos los campos que contiene el dataset, a priori no se va a descartar ninguno de los campos dado que todos contienen información sobre la composición química de los vinos a excepción del campo quality que tampoco se va a descartar dados los objetivos a cumplir en esta práctica y que se han expuesto en el apartado anterior.

En el transcurso de esta práctica se han descartado y modificado campos del dataset pero esto ya se comentará en los apartados correspondientes.

Ceros y elementos nulos

En primer lugar se va a analizar y a procesar aquellos registros que contengan elementos nulos o ceros en alguno de sus campos.

Empezaremos mirando el número de ceros que contiene cada uno de los campos que conforman el dataset:

```
colSums(data_0 == 0)
```

```
> #miramos el número de ceros de cada campo
> colSums(data_0 == 0)
       fixed.acidity     volatile.acidity          citric.acid       residual.sugar            chlorides  free.sulfur.dioxide
                   0                    0                  132                    0                    0                    0
total.sulfur.dioxide              density                   pH            sulphates              alcohol              quality
                   0                    0                    0                    0                    0                    0
> |
```

Figura 1. Número de ceros de los diferentes campos que componen el dataset

Como se puede observar en la Figura 1, el único campo que contiene ceros es el citric acid. Recordemos que los campos del dataset, menos quality, dan información sobre composiciones química. El hecho de que una composición química será cero significaría que no contiene dicha substancia, también podría ser un indicativo de que la concentración es tan baja que no se puede medir con precisión, si este fuese el caso, descartaría dichos registros por considerar que están incompletos.

No obstante, si miramos la descripción del campo citric acid del apartado 2 dice que en pequeñas cantidades añade frescura al vino, esta descripción me da a entender que dicho elemento es un añadido extra al vino y que no se encuentra de forma natural en el vino, es decir, el hecho de que sea cero significa que no se le ha añadido citric acid. Por lo tanto dichos registros con ceros son correctos y no se van a eliminar.

Una vez se ha determinado cual va a ser el tratamiento que se les aplicará a los campos con ceros, vamos a realizar un ejercicio similar pero para encontrar campos que contengan valores nulos.

```
sapply(data_0,function(x) sum(is.na(x)))
```

```
> #miramos el número de nulos de cada campo
> sapply(data_0,function(x) sum(is.na(x)))
       fixed.acidity     volatile.acidity          citric.acid       residual.sugar            chlorides  free.sulfur.dioxide
                   0                    0                    0                    0                    0                    0
total.sulfur.dioxide              density                   pH            sulphates              alcohol              quality
                   0                    0                    0                    0                    0                    0
> |
```

Figura 2. Número de nulos de los diferentes campos que componen el dataset

Como se puede observar en la Figura 2, ningún campo contiene valores nulos. En el caso de que hubiese habido valores nulos en uno o más campos de un registro, hubiese descartado dicho registró por estar incompleto. Si se quisiese rellenar estos valores nulos se podrían emplear el promedio de dicho campo para el resto de vinos de la misma calidad asumiendo que vinos de la misma calidad tienen valores parecidos en un mismo campo.

Valores estadísticos, identificación y tratamiento de los valores extremos

Para tener una visión general de los principales valores estadísticos (máximos, mínimos, valores medios, cuartiles…). Además son de utilidad a la hora de ver visualmente cuanto se alejan los valores extremos del promedio, así como aportar otra información relevante de cara a realizar el análisis.

Por ejemplo, el campo quality está definido en el rango 0-10 mientras que si observamos la figura 3 podemos observar que el rango real del dataset para dicho campo es 3-8, es decir, no tenemos vinos de todas las calidades.

```
summary(data_0)
```

```
> #valores estadisticos del dataset
> summary(data_0)
 fixed.acidity   volatile.acidity  citric.acid    residual.sugar     chlorides       free.sulfur.dioxide total.sulfur.dioxide
 Min.    : 4.60   Min.    :0.1200   Min.    :0.000   Min.    : 0.900   Min.    :0.01200   Min.    : 1.00       Min.    :  6.00
 1st Qu.: 7.10   1st Qu.:0.3900   1st Qu.:0.090   1st Qu.: 1.900   1st Qu.:0.07000   1st Qu.: 7.00       1st Qu.: 22.00
 Median : 7.90   Median :0.5200   Median :0.260   Median : 2.200   Median :0.07900   Median :14.00       Median : 38.00
 Mean    : 8.32   Mean    :0.5278   Mean    :0.271   Mean    : 2.539   Mean    :0.08747   Mean    :15.87       Mean    : 46.47
 3rd Qu.: 9.20   3rd Qu.:0.6400   3rd Qu.:0.420   3rd Qu.: 2.600   3rd Qu.:0.09000   3rd Qu.:21.00       3rd Qu.: 62.00
 Max.    :15.90   Max.    :1.5800   Max.    :1.000   Max.    :15.500   Max.    :0.61100   Max.    :72.00       Max.    :289.00
    density             pH          sulphates         alcohol          quality
 Min.    :0.9901   Min.    :2.740   Min.    :0.3300   Min.    : 8.40   Min.    :3.000
 1st Qu.:0.9956   1st Qu.:3.210   1st Qu.:0.5500   1st Qu.: 9.50   1st Qu.:5.000
 Median :0.9968   Median :3.310   Median :0.6200   Median :10.20   Median :6.000
 Mean    :0.9967   Mean    :3.311   Mean    :0.6581   Mean    :10.42   Mean    :5.636
 3rd Qu.:0.9978   3rd Qu.:3.400   3rd Qu.:0.7300   3rd Qu.:11.10   3rd Qu.:6.000
 Max.    :1.0037   Max.    :4.010   Max.    :2.0000   Max.    :14.90   Max.    :8.000
>
```

Figura 3. Principales indicadores estadísticos del dataset

En vista de los resultados de la figura3 y viendo los datos del dataset, podemos descartar el campo density para realizar el modelo de clasificación dado que apenas hay variación en dicho campo entre un vino y otro. No obstante, no se eliminará hasta el final dado que puede haber correlaciones de este campo con otros y es necesario identificarlas.

La identificación de los valores extremos se puede hacer tanto visualmente empleando gráficos de cajas o empleando formular estadísticas que te devuelven un el valor extremo más pequeño. En esta práctica se emplearán graficas de cajas y luego se empelará una sentencia de R que devuelve los valores extremos.

Este tratamiento se va a realizar a todos los campos incluido el quality aunque no tiene ningún sentido buscar valores extremos en un campo con un rango de valores delimitado que además se va a emplear para obtener un modelo de clasificación. Se realizará simplemente vara ver los diferentes cuartiles y ver su distribución que ya se puede intuir viendo los datos estadísticos de dicho campo en la Figura 3.

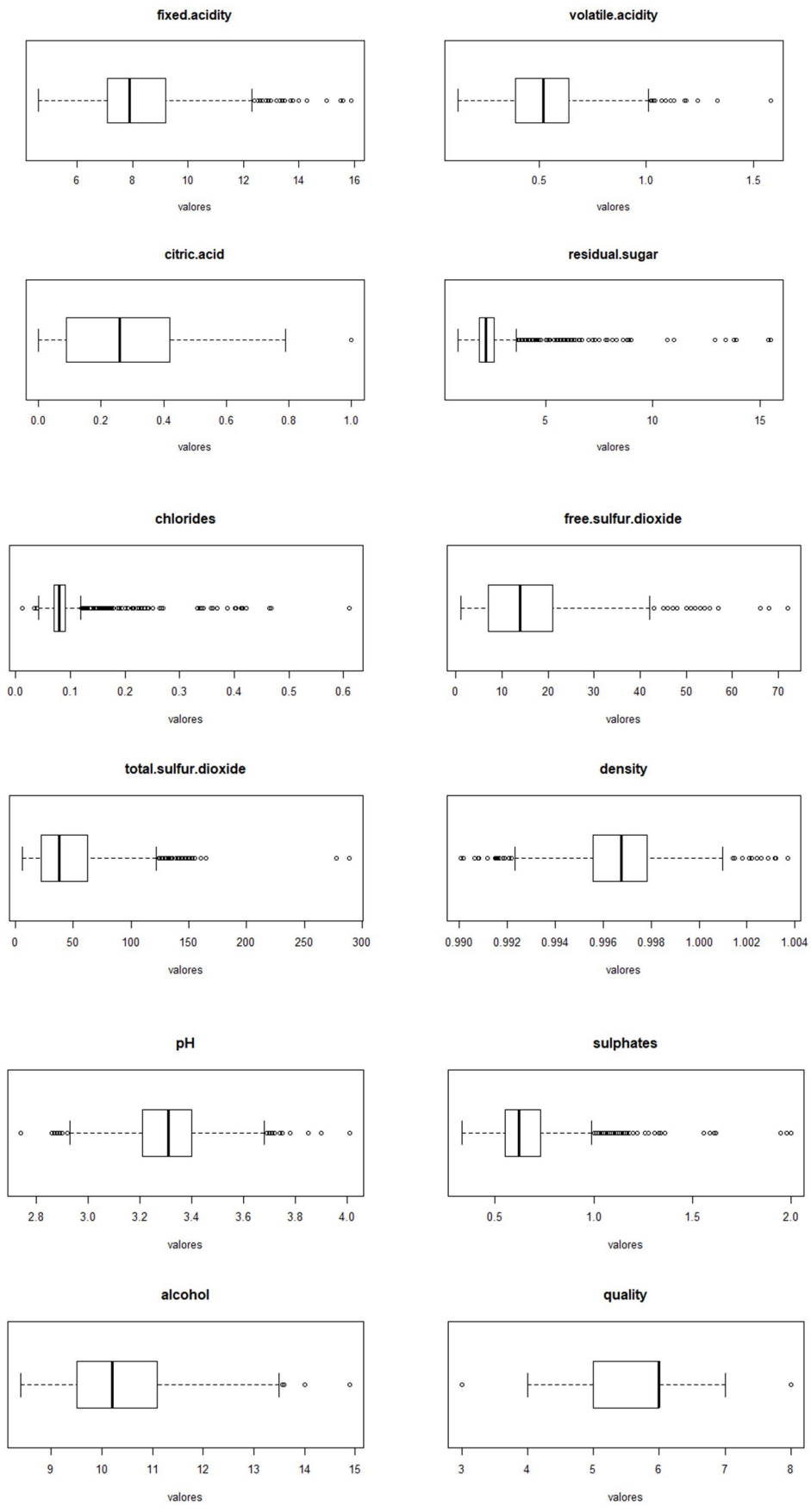


Figura 4. Graficos de cajas de los difererentes campos que componen el dataset.

Los valores extremos de los diferentes campos y los comentarios respecto a los valores que toman( si son necesarios) son:

- fixed acidity

```
12.8 12.8 15.0 15.0 12.5 13.3 13.4 12.4 12.5 13.8 13.5 12.6 12.5 12.8 12.8 14.0 13.7 13.7 12.7 12.5
12.8 12.6 15.6 12.5 13.0 12.5 13.3 12.4 12.5 12.9 14.3 12.4 15.5 15.5 15.6 13.0 12.7 13.0 12.7 12.4
12.7 13.2 13.2 13.2 15.9 13.3 12.9 12.6 12.6
```

- volatile acidity

```
1.130 1.020 1.070 1.330 1.330 1.040 1.090 1.040 1.240 1.185 1.020 1.035 1.025 1.115 1.020 1.020 1.580
1.180 1.040
```

- citric acid

```
1
```

- residual sugar. En la descripción de este campo se dice que es raro encontrar vinos con menos de 1g/l y mayor a 45g/l. En vista de los datos, el valor mínimo es 0.9g/l y el mayor 15 por lo que estamos en el rango esperado pese a tener algún valor por debajo de 1g/l.

```
 6.10  6.10  3.80  3.90  4.40 10.70  5.50  5.90  5.90  3.80  5.10  4.65  4.65  5.50  5.50  5.50  5.50
 7.30  7.20  3.80  5.60  4.00  4.00  4.00  4.00  7.00  4.00  4.00  6.40  5.60  5.60 11.00 11.00  4.50
 4.80  5.80  5.80  3.80  4.40  6.20  4.20  7.90  7.90  3.70  4.50  6.70  6.60  3.70  5.20 15.50  4.10
 8.30  6.55  6.55  4.60  6.10  4.30  5.80  5.15  6.30  4.20  4.20  4.60  4.20  4.60  4.30  4.30  7.90
 4.60  5.10  5.60  5.60  6.00  8.60  7.50  4.40  4.25  6.00  3.90  4.20  4.00  4.00  4.00  6.60  6.00
 6.00  3.80  9.00  4.60  8.80  8.80  5.00  3.80  4.10  5.90  4.10  6.20  8.90  4.00  3.90  4.00  8.10
 8.10  6.40  6.40  8.30  8.30  4.70  5.50  5.50  4.30  5.50  3.70  6.20  5.60  7.80  4.60  5.80  4.10
12.90  4.30 13.40  4.80  6.30  4.50  4.50  4.30  4.30  3.90  3.80  5.40  3.80  6.10  3.90  5.10  5.10
 3.90 15.40 15.40  4.80  5.20  5.20  3.75 13.80 13.80  5.70  4.30  4.10  4.10  4.40  3.70  6.70 13.90
 5.10  7.80
```

- chlorides

```
0.176 0.170 0.368 0.341 0.172 0.332 0.464 0.401 0.467 0.122 0.178 0.146 0.236 0.610 0.360 0.270 0.039
0.337 0.263 0.611 0.358 0.343 0.186 0.213 0.214 0.121 0.122 0.122 0.128 0.120 0.159 0.124 0.122 0.122
0.174 0.121 0.127 0.413 0.152 0.152 0.125 0.122 0.200 0.171 0.226 0.226 0.250 0.148 0.122 0.124 0.124
0.143 0.222 0.039 0.157 0.422 0.034 0.387 0.415 0.157 0.157 0.243 0.241 0.190 0.132 0.126 0.038 0.165
0.145 0.147 0.012 0.012 0.039 0.194 0.132 0.161 0.120 0.120 0.123 0.123 0.414 0.216 0.171 0.178 0.369
0.166 0.166 0.136 0.132 0.132 0.123 0.123 0.123 0.403 0.137 0.414 0.166 0.168 0.415 0.153 0.415 0.267
0.123 0.214 0.214 0.169 0.205 0.205 0.039 0.235 0.230 0.038
```

- free sulfur dioxide

```
52 51 50 68 68 43 47 54 46 45 53 52 51 45 57 50 45 48 43 48 72 43 51 51 52 55 55 48 48 66
```

- total sulfur dioxide

```
145 148 136 125 140 136 133 153 134 141 129 128 129 128 143 144 127 126 145 144 135 165 124 124 134 124
129 151 133 142 149 147 145 148 155 151 152 125 127 139 143 144 130 278 289 135 160 141 141 133 147 147
131 131 131
```

- density

```
0.99160 0.99160 1.00140 1.00150 1.00150 1.00180 0.99120 1.00220 1.00220 1.00140 1.00140 1.00140 1.00140
1.00320 1.00260 1.00140 1.00315 1.00315 1.00315 1.00210 1.00210 0.99170 0.99220 1.00260 0.99210 0.99154
0.99064 0.99064 1.00289 0.99162 0.99007 0.99007 0.99020 0.99220 0.99150 0.99157 0.99080 0.99084 0.99191
1.00369 1.00369 1.00242 0.99182 1.00242 0.99182
```

- pH. Respecto a este campo, en su descripción se dice que los valores típicos para el vino son entre 3 – 4 de pH. Como puede observar en la subfigura correspondiente al ph de la figura 4, los valores que toma dicho campo están en el rango esperado o muy cerca de este.

```
3.90 3.75 3.85 2.74 3.69 3.69 2.88 2.86 3.74 2.92 2.92 2.92 3.72 2.87 2.89 2.89 2.92 3.90 3.71 3.69
3.69 3.71 3.71 2.89 2.89 3.78 3.70 3.78 4.01 2.90 4.01 3.71 2.88 3.72 3.72
```

- sulphates

```
1.56 1.28 1.08 1.20 1.12 1.28 1.14 1.95 1.22 1.95 1.98 1.31 2.00 1.08 1.59 1.02 1.03 1.61 1.09 1.26
1.08 1.00 1.36 1.18 1.13 1.04 1.11 1.13 1.07 1.06 1.06 1.05 1.06 1.04 1.05 1.02 1.14 1.02 1.36 1.36
1.05 1.17 1.62 1.06 1.18 1.07 1.34 1.16 1.10 1.15 1.17 1.17 1.33 1.18 1.17 1.03 1.17 1.10 1.01
```

- alcohol

```
14.00000 14.00000 14.00000 14.00000 14.90000 14.00000 13.60000 13.60000 13.60000 14.00000 14.00000
13.56667 13.60000
```

- quality

```
8 8 8 8 8 3 8 8 8 3 8 3 8 3 3 8 8 8 8 8 3 3 8 8 3 3 3 8
```

En vista de los valores extremos obtenidos y del papel que juegan dentro del dataset, no se va a realizar ningún tratamiento ni se van a eliminar los registros que contienen datos extremos por los siguientes motivos:

- No tengo conocimientos referentes a la composición química de los diferentes vinos, más allá de lo que se comenta en las diferentes descripciones de los campos, con lo cual no tengo ningún criterio en que basarme para determinar si estos valores extremos son correctos o no y en el caso de que fuesen incorrectos como corregirlos.
- Pese a que hay valores extremos cuyo valor es bastante superior al valor medio, visualmente no se aprecia ningún valor extremo que realmente este aislado del resto. Es cierto, que hay valores extremos dos órdenes de magnitud superior (ejemplo total sulfur dioxide) pero viendo el resto de valores que toma dicho campo no me parecen significativamente distintos del resto.
- Estos valores extremos pueden haber sido los causantes de que un vino se clasificase en una categoría muy alta o por el contrario en una categoría muy baja. Es decir, si se eliminasen podrían afectar de forma negativa al modelo de clasificación que se quiere generar.

Correlaciones

De cara a determinar si alguno de los campos del dataset es redundante es necesario calcular la matriz de correlaciones entre los diferentes elementos. Esta matriz no incluirá el campo quality dado que lo que nos interesa es determinar las correlaciones entre los diferentes campos que incluyen la composición química.

```
data_1 <- data_0[-12]

cor(data_1)
```

```
> cor(data_1)
                      fixed.acidity volatile.acidity citric.acid residual.sugar    chlorides free.sulfur.dioxide
fixed.acidity            1.00000000     -0.256130895  0.67170343    0.114776724  0.093705186        -0.153794193
volatile.acidity        -0.25613089      1.000000000 -0.55249568    0.001917882  0.061297772        -0.010503827
citric.acid              0.67170343     -0.552495685  1.00000000    0.143577162  0.203822914        -0.060978129
residual.sugar           0.11477672      0.001917882  0.14357716    1.000000000  0.055609535         0.187048995
chlorides                0.09370519      0.061297772  0.20382291    0.055609535  1.000000000         0.005562147
free.sulfur.dioxide     -0.15379419     -0.010503827 -0.06097813    0.187048995  0.005562147         1.000000000
total.sulfur.dioxide    -0.11318144      0.076470005  0.03553302    0.203027882  0.047400468         0.667666450
density                  0.66804729      0.022026232  0.36494718    0.355283371  0.200632327        -0.021945831
pH                      -0.68297819      0.234937294 -0.54190414   -0.085652422 -0.265026131         0.070377499
sulphates                0.18300566     -0.260986685  0.31277004    0.005527121  0.371260481         0.051657572
alcohol                 -0.06166827     -0.202288027  0.10990325    0.042075437 -0.221140545        -0.069408354
                      total.sulfur.dioxide     density          pH    sulphates     alcohol
fixed.acidity               -0.11318144  0.66804729 -0.68297819  0.183005664 -0.06166827
volatile.acidity             0.07647000  0.02202623  0.23493729 -0.260986685 -0.20228803
citric.acid                  0.03553302  0.36494718 -0.54190414  0.312770044  0.10990325
residual.sugar               0.20302788  0.35528337 -0.08565242  0.005527121  0.04207544
chlorides                    0.04740047  0.20063233 -0.26502613  0.371260481 -0.22114054
free.sulfur.dioxide          0.66766645 -0.02194583  0.07037750  0.051657572 -0.06940835
total.sulfur.dioxide         1.00000000  0.07126948 -0.06649456  0.042946836 -0.20565394
density                      0.07126948  1.00000000 -0.34169933  0.148506412 -0.49617977
pH                          -0.06649456 -0.34169933  1.00000000 -0.196647602  0.20563251
sulphates                    0.04294684  0.14850641 -0.19664760  1.000000000  0.09359475
alcohol                     -0.20565394 -0.49617977  0.20563251  0.093594750  1.00000000
>
```

Figura 4. Matriz de correlaciones entre los diferentes elementos del dataset. Recuadradas en rojo se encuentran los valores con una correlación significativa.

En la figura 4 se pueden observar las correlaciones entre los diferentes campos y se han resaltado en rojo aquellas correlaciones que destacaban con respecto al resto. En vista de los datos obtenidos, se ha cogido como criterio para establecer una correlación significativa que el valor fuese menor a -0.49 o mayor a 0.49.

Vamos a analizar las diferentes correlaciones significativas que se han encontrado para determinar si dicha correlación permite o no eliminar campos redundantes.

- Citric acid – fixed acidity. Que haya una correlación significativa entre estos dos campos dado que el ácido cítrico contribuye al grado de acidez del vino. Como puede

observarse en la figura 5, a medida que aumenta citric acid, también la hace fixed acidity.

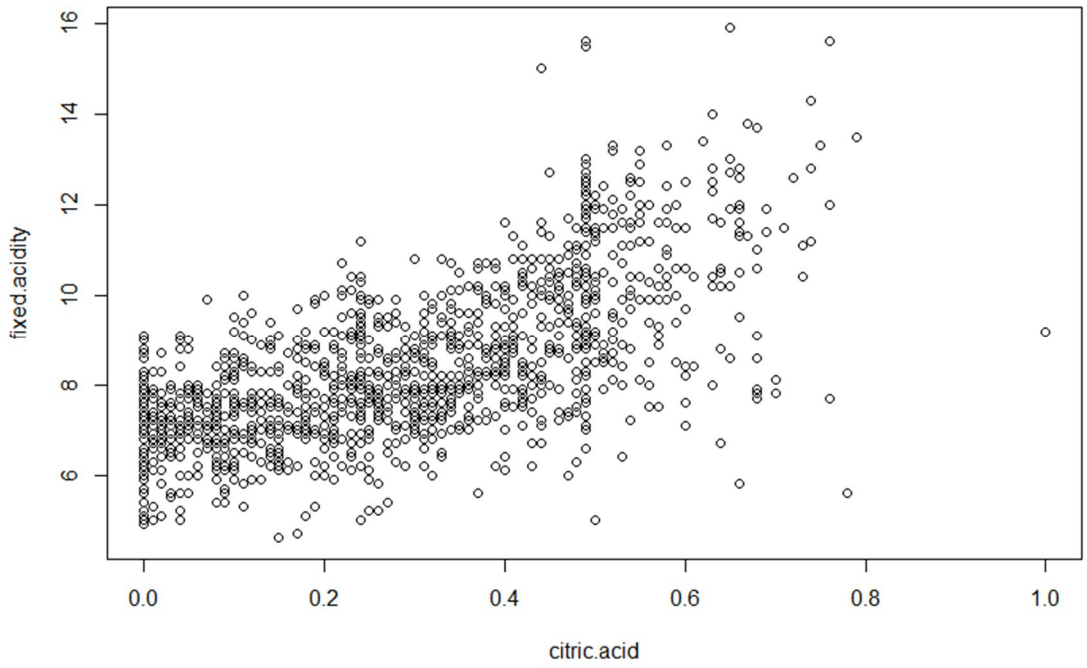


Figura 5. Fixed acidity en función de citric acid.

- Citric acid – ph. De forma similar a lo anterior, la cantidad de ácido cítrico afecta a la ph del vino. Como puede observarse en la figura 6, a medida que aumenta citric acid, baja el ph, esto es, el vino se vuelve más ácido.

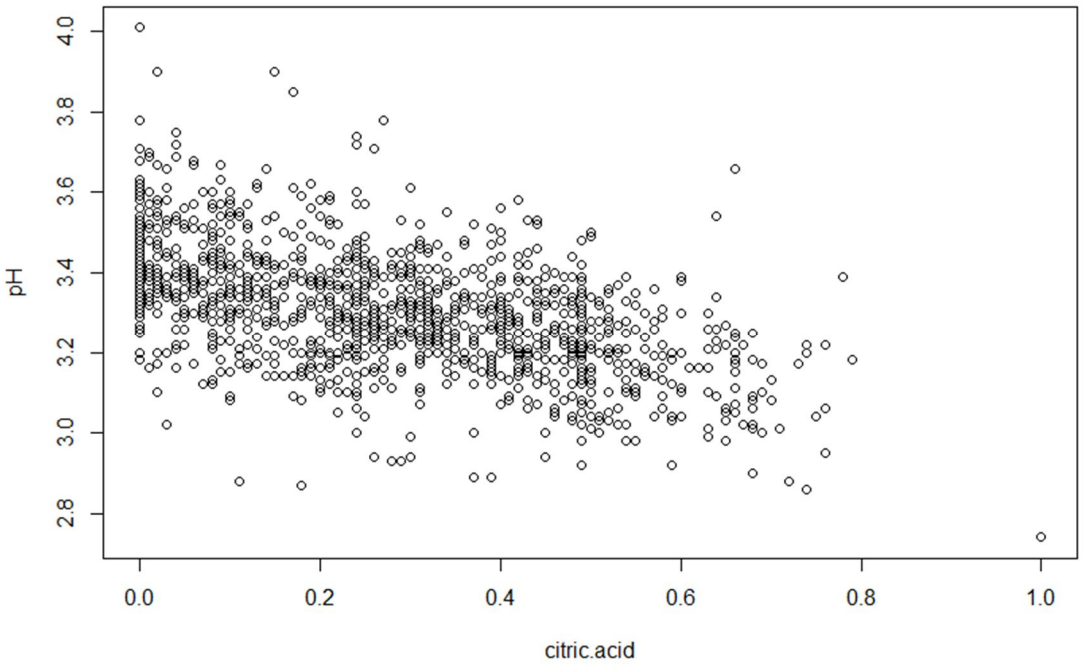


Figura 6. pH en función de citric acid.

- Citric acid – volatile acidity. La acidez volátil mide la cantidad de ácido acético (vinagre) en el vino y no veo a priori ninguna relación entre el ácido cítrico y el ácido acético a no ser que el ácido cítrico se emplee para contrarrestar un exceso de ácido acético pero no tengo nada en que basarme para realizar dicha afirmación. Como puede observarse en la figura 7, a medida que aumenta citric acid, baja la volatile acidity.

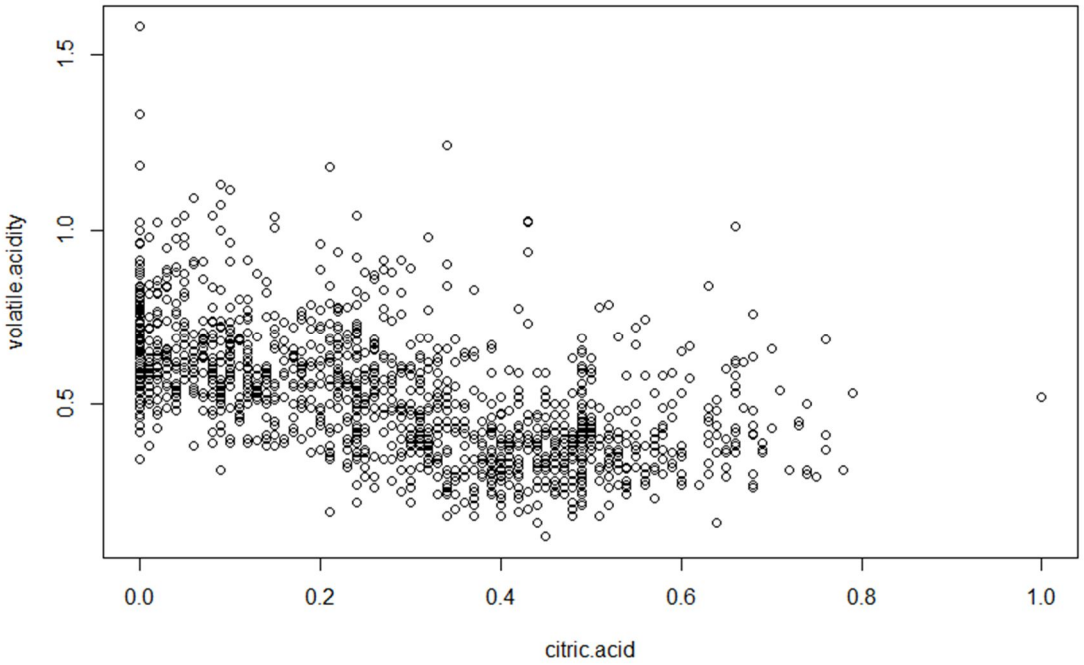


Figura 7. Volatile acidity en función de citric acid.

- Citric acid – density. Por lo que deduzco de la descripción del campo citric acid, este es su añadido por lo tanto al añadir ácido cítrico aumento la densidad. Como puede observarse en la figura 8, a medida que aumenta citric acid, también aumenta la density.

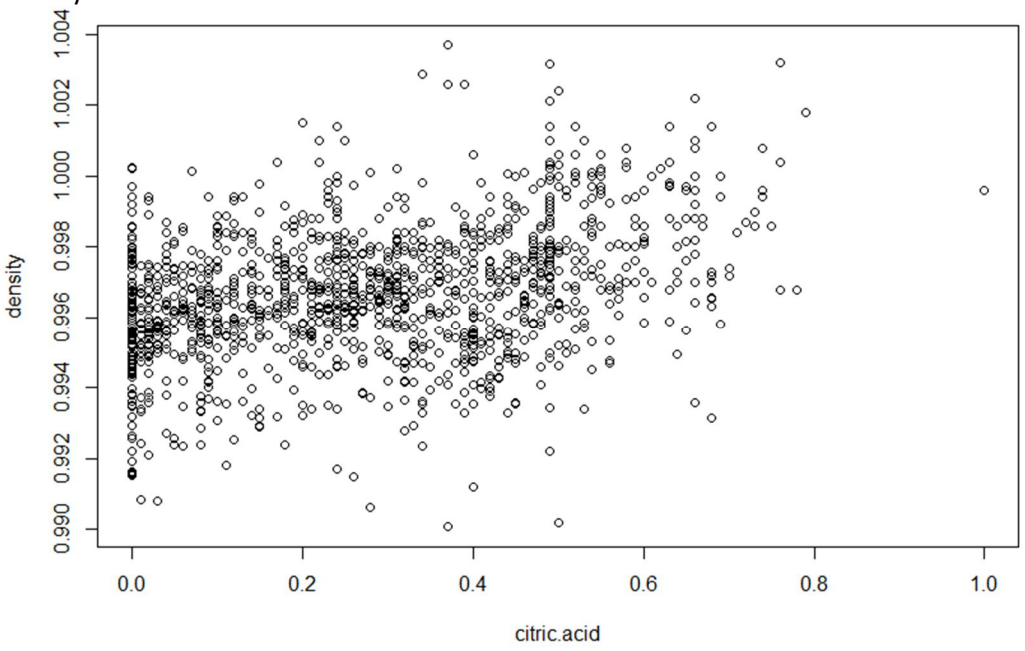


Figura 8. Density en función de citric acid.

- Total sulfur dioxide – free sulfur dioxide. Estos dos campos están relacionados, deduzco por los nombre que el total = libre + no libres por lo que dentro del total se encuentra ya la información de los libre. En la figura 9 no se puede ver ninguna relación aparente entre ambos campos pese a que la correlación es significativa.

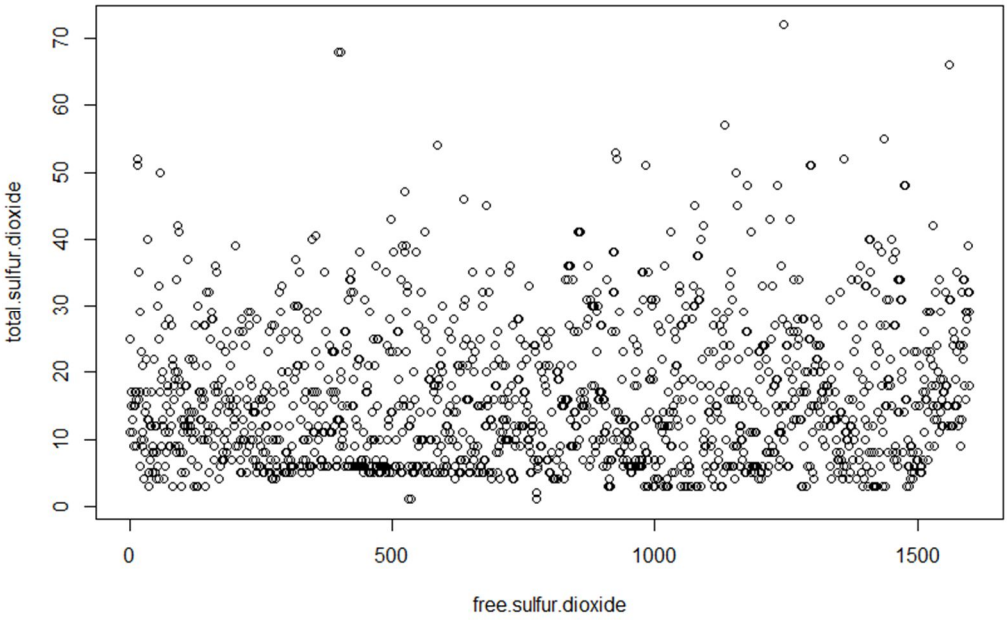


Figura 9. Total sulfur dioxide en función de free sulfur dioxide.

- Alcohol – density. Está relación también es 'directa', la cantidad de alcohol (azúcar al fin y al cabo) afecta a la densidad. Tal y como puede verse en la figura 10, si al alcohol aumenta, la densidad disminuye.

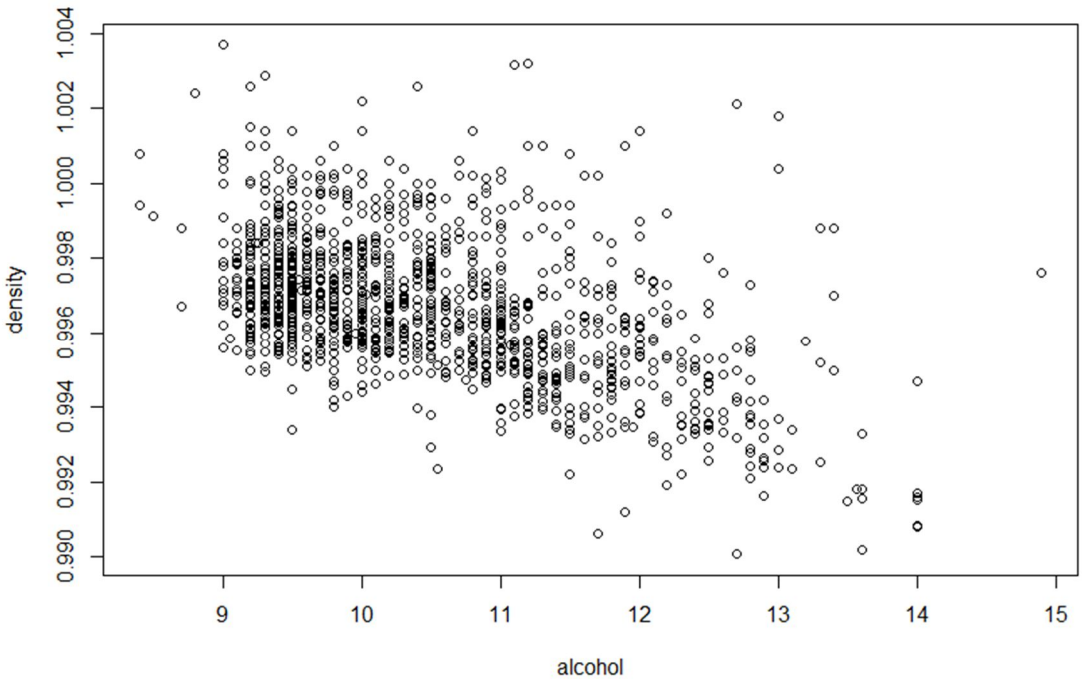


Figura 10. Density en función de alcohol.

En vista de los resultados obtenidos, se puede decir que hay campos que están significativamente correlacionados entre ellos pero ninguno de ellos podría considerarse redundante. Además, con las diferentes representaciones se ha podido ver que pese a que se sigue una tendencia, no se puede determinar una regresión que nos pase de un campo a otro.

La única excepción en cuanto a redundancias es el campo Total sulfur dioxide que contiene la información de free sulfur dioxide y de los 'bound sulfur dioxide', tal y como pone en la descripción del campo. Por ello se va a generar una nueva columna que sea 'bound sulfur dioxide' y se eliminará la columna Total sulfur dioxide por contener datos redundantes.

```
#añadimos la columna de bound sulfur dioxide y eliminamos Total sulfur dioxide

bound <-  data_0$total.sulfur.dioxide - data_0$free.sulfur.dioxide

data_0 $total.sulfur.dioxide <- bound

colnames(data_0)[7] <- "bound.sulfur.dioxide"
```

| | fixed.acidity | volatile.acidity | citric.acid | residual.sugar | chlorides | free.sulfur.dioxide | bound.sulfur.dioxide | density | pH |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 7.4 | 0.700 | 0.00 | 1.90 | 0.076 | 11 | 23 | 0.9978 | 3.51 |
| 2 | 7.8 | 0.880 | 0.00 | 2.60 | 0.098 | 25 | 42 | 0.9968 | 3.20 |
| 3 | 7.8 | 0.760 | 0.04 | 2.30 | 0.092 | 15 | 39 | 0.9970 | 3.26 |
| 4 | 11.2 | 0.280 | 0.56 | 1.90 | 0.075 | 17 | 43 | 0.9980 | 3.16 |
| 5 | 7.4 | 0.700 | 0.00 | 1.90 | 0.076 | 11 | 23 | 0.9978 | 3.51 |

Figura 11. Imagen en la que se ve el nuevo campo  bound sulfur dioxide

<u>Discretización de los datos</u>

El último tratamiento que se le va a realizar a los datos es discretizar los diferentes campos numéricos por rangos y a cada rango se le asociará una variable categórica. Esto se va a realizar dado que queremos procesar los datos para obtener un modelo de clasificación y estos suelen presentar mejores resultados con variables categóricas antes que numéricas. Además, si se empleasen variables numéricas, dado que estas fluctúan bastante, el modelo que se obtenido tendría muchas reglas.

Se van a discretizar todos los campos menos dos:

- El campo quality no se va a discretizar dado que contiene el elemento por el cual queremos clasificar los datos.
- El campo density tampoco se va a discretizar dado que la oscilación de valores es muy pequeña. Tan pequeña es la variación de dicho campo se va a eliminar del dataset dada la poca información que aporta.

Se va a discretizar empleando el método k-menas, este método compara la distancia entre un valor y los valores medios de los intervalos adyacentes para decidir si un valor ha de ir en un intervalo o en otro. Esto puede hacer que los valores de las fronteras de los intervalos varíen ligeramente de una ejecución a otra.

Discretizaremos los datos en 5  categorías: Muy Baja, Baja, Normal, Alta, Muy Alta.  Los rangos para los diferentes campos son en el orden expuesto:

- fixed acidity

  `Levels: [4.6,6.58) [6.58,7.79) [7.79,9.24) [9.24,11.2) [11.2,15.9]`
- volatile acidity

Levels: [0.12,0.34) [0.34,0.476) [0.476,0.618) [0.618,0.809) [0.809,1.58]

- citric acid

  Levels: [0,0.0838) [0.0838,0.204) [0.204,0.339) [0.339,0.489) [0.489,1]

- residual sugar

  Levels: [0.9,2.62) [2.62,4.33) [4.33,6.97) [6.97,11.4) [11.4,15.5]

- chlorides

  Levels: [0.012,0.0695) [0.0695,0.0919) [0.0919,0.15) [0.15,0.304) [0.304,0.611]

- free sulfur dioxide

  Levels: [1,8.5) [8.5,14.5) [14.5,22.9) [22.9,35.9) [35.9,72]

- bound sulfur dioxide

  Levels: [3,15.6) [15.6,29.9) [29.9,52.9) [52.9,86.4) [86.4,252]

- pH

  Levels: [2.74,3.18) [3.18,3.32) [3.32,3.45) [3.45,3.62) [3.62,4.01]

- sulphates

  Levels: [0.33,0.538) [0.538,0.649) [0.649,0.799) [0.799,1.12) [1.12,2]

- alcohol

  Levels: [8.4,9.45) [9.45,9.98) [9.98,10.7) [10.7,11.8) [11.8,14.9]

| | fixed.acidity | volatile.acidity | citric.acid | residual.sugar | chlorides | free.sulfur.dioxide | bound.sulfur.dioxide | pH | sulphates | alcohol | quality |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Baja | Alta | Muy baja | Muy baja | Muy baja | Baja | Baja | Muy alta | Muy baja | Baja | 5 |
| 2 | Normal | Muy alta | Muy baja | Baja | Muy baja | Alta | Normal | Baja | Baja | Baja | 5 |
| 3 | Normal | Alta | Muy baja | Baja | Muy baja | Normal | Normal | Normal | Baja | Baja | 5 |
| 4 | Muy alta | Muy baja | Muy alta | Muy baja | Muy baja | Normal | Normal | Baja | Baja | Baja | 6 |
| 5 | Baja | Alta | Muy baja | Muy baja | Muy baja | Baja | Baja | Muy alta | Muy baja | Baja | 5 |
| 6 | Baja | Alta | Muy baja | Muy baja | Muy baja | Baja | Baja | Muy alta | Muy baja | Baja | 5 |
| 7 | Normal | Normal | Muy baja | Muy baja | Muy baja | Normal | Normal | Normal | Muy baja | Baja | 5 |
| 8 | Baja | Alta | Muy baja | Muy baja | Muy baja | Normal | Muy baja | Alta | Muy baja | Normal | 7 |
| 9 | Normal | Normal | Muy baja | Muy baja | Muy baja | Baja | Muy baja | Alta | Muy baja | Baja | 7 |
| 10 | Baja | Normal | Normal | Alta | Muy baja | Normal | Alta | Alta | Normal | Normal | 5 |
| 11 | Baja | Normal | Baja | Muy baja | Muy baja | Normal | Normal | Normal | Muy baja | Muy baja | 5 |
| 12 | Baja | Normal | Normal | Alta | Muy baja | Normal | Alta | Alta | Normal | Normal | 5 |
| 13 | Muy baja | Normal | Muy baja | Muy baja | Muy baja | Normal | Normal | Muy alta | Muy baja | Normal | 5 |
| 14 | Normal | Normal | Normal | Muy baja | Muy baja | Baja | Baja | Normal | Muy alta | Muy baja | 5 |
| 15 | Normal | Alta | Baja | Normal | Muy baja | Muy alta | Muy alta | Baja | Normal | Muy baja | 5 |
| 16 | Normal | Alta | Baja | Normal | Muy baja | Muy alta | Muy alta | Baja | Normal | Muy baja | 5 |
| 17 | Normal | Muy baja | Muy alta | Muy baja | Muy baja | Muy alta | Alta | Normal | Normal | Normal | 7 |
| 18 | Normal | Normal | Normal | Muy baja | Muy baja | Normal | Normal | Baja | Alta | Muy baja | 5 |
| 19 | Baja | Normal | Baja | Normal | Muy baja | Muy baja | Baja | Alta | Muy baja | Muy baja | 4 |
| 20 | Normal | Muy baja | Alta | Muy baja | Muy baja | Normal | Normal | Muy baja | Alta | Muy baja | 6 |
| 21 | Normal | Muy baja | Alta | Muy baja | Muy baja | Alta | Normal | Alta | Muy baja | Baja | 6 |

Figura 12. Imagen en la que se ve los diferentes campos discretizados.

<u>Exportación de los datos procesados</u>

Se crearán dos ficheros en los cuales se exportarán los datos procesados:

- datos-discretizados.csv que contendrá los datos discretizados
- datos-no-discretizados.csv que contendrá los datos no discretizados

Exportaremos los datos tanto discretizados como no discretizados para implementar el modelo con ambos ficheros y ver si el hecho de discretizar los datos mejora o no el modelo de clasificació.

## 5. Aplicación del modelo

Para aplicar el modelo de clasificación se han de tener en cuenta los siguientes factores:

- En el dataset no tenemos una muestra uniforme de vinos en cuanto a sus calidades. Como se ha visto en el apartado 2, la mayoría de los vinos del dataset son de calidad 5, 6, 7.
- No hay vinos de calidad 0, 1, 2, 9, 10. Por lo tanto el modelo que se genere no sabrá clasificar correctamente vinos de esas calidades y las clasificará en una calidad u otra en función de cómo se parezcan sus propiedades a la del resto de vinos.
- Se va a dividir el dataset en dos grupos: un primer grupo que lo conformarán el 80% de los datos que servirá para generar el modelo; un segundo grupo que lo conformarán el 20% restante de datos que servirá para comprobar la calidad del modelo. Además, estos dos grupos se generarán de forma aleatoria. Esto implica que cada vez que se ejecuta el programa se genera un modelo que puede diferir del modelo generado anteriormente.
- Para generar las reglas se empleará el algoritmo C5.0 implementado en la librearía library(C50)
- Para comprobar el error asociado al modelo se empelará la validación cruzada con 10 grupos. La validación cruzada consiste en dividir la muestra en 10 grupos, utilizar 9 para generar un modelo y el restante para hacer el test de calidad. Una vez se obtiene el error para dicha conjunto, se cambia de conjunto y se repite el proceso 10 veces. Por último el error del modelo es el promedio de los errores.
  Con este método todos los datos se emplean para generar el modelo y para testearlo y con esto se obtiene un error asociado al modelo más real.
- Dado el gran volumen de datos, no será posible generar el árbol de decisión gráficamente por lo que se genera un modelo de reglas de clasificación que se exportará a un fichero.

Modelo con el fichero con los datos discretizados

Las reglas generadas con este modelo son y 78 se encuentran en el fichero reglas-discretizadas.txt.

Con el modelo generado, se ha intentado clasificar el conjunto de datos de testeo y se han obtenido los siguientes resultados.

```
> table(ytest, Predecido =p1)
     Predecido
ytest  3  4  5  6  7  8
    4  0  0 10  6  0  0
    5  0  1 99 34  3  0
    6  0  1 34 71 14  0
    7  0  0  8 19 13  0
    8  0  0  0  3  4  0
```

Figura 13. Tabla con las calidades reales (ytest) contra las predichas (Predecido) para el fichero con los datos discretizados.

En la figura 13 se muestra la tabla de los valores de las calidades reales del conjunto de prueba contra los valores predichos por el modelo. Todos los valores diferentes de cero fuera de la diagonal son valores incorrectamente clasificados mientras que los valores de la diagonal corresponden a los correctamente clasificados. Comentar también que tal y como puede verse el conjunto de prueba (ytest) no contenía ningún registro con calidad = 3, este se podía producir dado que sólo hay 10 registros en el fichero con calidad = 3.

El error de este modelo se ha calculado mediante validación cruzada y se ha obtenido los siguientes valores del error:

```
[1] 0.45
[1] 0.43125
[1] 0.43125
[1] 0.35
[1] 0.41875
[1] 0.4716981
[1] 0.36875
[1] 0.425
[1] 0.45
[1] 0.34375
> #el error final es la media de los errores
> error_total = error_total / 10.0
> print(error_total)
[1] 0.4140448
```

Figura 14. Los diferentes valores de los errores obtenidos mediante validación cruzada y el error final que es el promedio de los errores.

Por lo que el error final, es el promedio de los errores que es del 41.4%.

Modelo con el fichero con los datos sin discretizar

Las reglas generadas con este modelo son 72 y se encuentran en el fichero reglas-sin-discretizadar.txt.

Con el modelo generado, se ha intentado clasificar el conjunto de datos de testeo y se han obtenido los siguientes resultados.

| | Predecido | | | | | |
|---|---|---|---|---|---|---|
| ytest | 3 | 4 | 5 | 6 | 7 | 8 |
| 3 | 0 | 0 | 1 | 0 | 0 | 0 |
| 4 | 1 | 0 | 3 | 5 | 1 | 0 |
| 5 | 0 | 0 | 90 | 38 | 6 | 0 |
| 6 | 0 | 1 | 28 | 86 | 15 | 3 |
| 7 | 0 | 0 | 6 | 20 | 14 | 0 |
| 8 | 0 | 0 | 0 | 1 | 1 | 0 |

Figura 15. Tabla con las calidades reales (ytest) contra las predichas (Predecido) para el fichero con los datos sin discretizar.

La interpretación de la figura 15 es la misma que la figura 13 con la salvedad que en este caso el conjunto de pruebas sí que contenía registros con calidad = 3. El error de este modelo se ha calculado mediante validación cruzada y se ha obtenido los siguientes valores del error

```
[1] 0.33125
[1] 0.375
[1] 0.4
[1] 0.35
[1] 0.3625
[1] 0.4402516
[1] 0.43125
[1] 0.4
[1] 0.36875
[1] 0.41875
> #el error final es la media de los errores
> error_total = error_total / 10.0
> print(error_total)
[1] 0.3877752
```

Figura 16. Los diferentes valores de los errores obtenidos mediante validación cruzada y el error final que es el promedio de los errores.

Por lo que el error final de este modelo es el promedio de los errores que es del 38.7%.

## 6. Conclusiones

En esta práctica se ha analizado y procesado el fichero winequality-red.csv con el objetivo de generar un modelo de clasificación que permitiese clasificar de forma automática un vino en función de su composición química.

Para cumplir dicho objetivo se han realizado las siguientes acciones en cuanto a análisis y procesamiento de los datos:

- Un análisis descriptivo del fichero con el objetivo de saber que datos lo componen.
- Se han buscado campos erróneos, más concretamente, campos que fuesen nulos o contuviesen ceros.
  No se ha encontrado campos que tuviesen nulos pero sí que se ha encontrado que el campo citric acid contenía ceros. No obstante, era norma que dicho campo tuviese ceros dado que el citric acid es un añadido al vino y puede haber vino sin dicho elemento.
- Se han buscado los valores extremos tanto gráficamente como mediante una sentencia de R que devuelve los diferentes valores extremos para cada uno de los campos. Una vez obtenidos y analizados dichos valores extremos, se ha determinado que pese a ser valores extremos no se han eliminado dada mi falta de conocimiento en la materia de composiciones químicas de vino.
- Se han buscado correlaciones entre los diferentes campos para discernir si era posible eliminar campos que fuesen redundantes. Se ha visto que pese a que hay campos con correlaciones significativas, después de representarlo gráficamente no se podría realizar ningún tipo de regresión que permitiese pasar de una a otro.
- Por último, para intentar mejorar y simplificar el modelo de clasificación se han discretizado todos los campos numéricos en 5 categorías (muy baja, Baja, normal, alta muy alta) mediante el método de k-menas.

Una vez procesado el fichero, se ha creado un modelo de clasificación mediante el algoritmo C5.0 y se ha estimado su error mediante validación cruzada.

Se ha generado un modelo para los datos procesados discretizados y sin discretizar para poder comparar ambos modelos.

Estos modelos se componen de un conjunto de reglas que permiten clasificar los diferentes vinos en base a su composición química. El número de reglas no se ha visto modificado de forma significativa con los datos discretizado y sin discretizar.

Una vez realizado esto, se ha visto que el error asociado al modelo sin discretizar (38.7%) es ligeramente menor a la del modelo con los datos discretizados (41.4). Con esto se puede decir que el hecho de discretizar los diferentes campos no ha aportado nada al modelo de clasificación.

Por último quiero resaltar que hay datos relevantes como el tipo de uva, la marca y el precio de compra que por motivos de confidencialidad no se encontraban en el dataset original. Pero dichos campos se utilizaron para determinar la calidad del vino, es decir, es probable que este hecho haya tenido un impacto negativo en el modelo de clasificación. Pese a ello, ambos modelos generados tienen un acierto de alrededor del 60%

## 7. Recursos


a) https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009
b) Preparación de datos. Ramón Sangüesa i Solé. Editorial UOC PID_00165728
c) Clasificación: árboles de decisión. Ramón Sangüesa i Solé. Editorial UOC PID_00165729