Práctica 2. Limpieza y validación de los datos

Miembros del grupo: Víctor Camí Núñez

Índice

1. Introducción

Este trabajo se engloba en la realización de la segunda práctica de la asignatura de *Tipología y ciclo de vida de los datos*.

De entre todos los datasets que se encuentran en el repositorio de datos online *kaggle*, se ha escogido uno de los datasets propuestos en el enunciado de esta segunda práctica, winequality-red.csv.

Dicho dataset contiene la información relativa a la composición química de un conjunto de vinos y a cada uno de estos vinos se les ha asignado una calidad numérica dentro de una escala del 0 al 10 siendo el 0 un indicador de la calidad más baja y el 10 un indicador de la calidad más alta.

El principal motivo por el cual se ha escogido este dataset, es el hecho que asocia una calidad a cada vino en base a su composición química. Esto hace que una vez procesado y limpiado el dataset, se pueda generar un sencillo modelo de clasificación para clasificar cada vino en una calidad u otra dependiendo de su composición química.

Con lo anteriormente dicho, y para facilitar la estructuración de la resolución de esta práctica, se ha dividido este documento en una serie de apartados:

- En primer lugar se va a realizar un análisis descriptivo de la estructura del fichero winequality-red.csv que contiene los datos del dataset para ver los datos de los que disponemos y hacer un primer análisis estadístico de los datos que contiene.
- 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 preguntas que queremos responder mediante este dataset dado que esto condicionará el procesado y limpieza de los datos. Aunque ya se ha comentado que se quiere hacer un pequeño modelo de clasificación, en este apartado se entrará en más detalle.
- En tercer lugar 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. Entre estas acciones destacan: la detección de valores cero o nulos, la detección y tratamiento de valores extremos, ver el tipo de distribución que siguen los datos y la contribución de los diferentes campos del dataset sobre el campo de la calidad de los vinos.
- En cuarto lugar se explicará la implementación del modelo de clasificación empleado para determinar la calidad del vino en base a su composición química.
- Por último se presentarán las conclusiones extraídas en la realización de esta práctica.

2. Análisis 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.

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). En el dataset no tenemos una muestra uniforme de vinos en cuanto a sus calidades. La mayoría de los vinos del dataset son de calidad 5, 6. 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

Tipología de los datos

Referente al tipo de datos que conforman el dataset, todos los campos son numéricos tal y como puede verse en las siguiente imagen.

```
> sapply(data_0, class)
       fixed.acidity     volatile.acidity          citric.acid       residual.sugar            chlorides  free.sulfur.dioxide
           "numeric"            "numeric"            "numeric"            "numeric"            "numeric"            "numeric"
total.sulfur.dioxide              density                   pH            sulphates              alcohol              quality
           "numeric"            "numeric"            "numeric"            "numeric"            "numeric"            "integer"
> |
```

Valores estadísticos

Para tener una visión general de los datos se pueden calculas los principales valores estadísticos (máximos, mínimos, valores medios, cuartiles…) de los diferentes campos del dataset. Además son de utilidad a la hora de ver visualmente cuanto se alejan los valores extremos (que se obtendrán en el apartado 4) del promedio, así como aportar otra información relevante de cara a realizar el análisis.

```
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 1. Principales indicadores estadísticos del dataset

Por ejemplo en la figura 1 vemos que el campo citric acid contiene ceros dado que el valor mínimo que toma ese campo es cero, esto se puede deber a un error en la toma de los datos o ser correcto. En cualquier caso el tratamiento de los valore nulos y ceros se comenta en el siguiente apartado.

Otro ejemplo sería el campo quality, para el cual podemos ver que está definido en el rango 0-10 mientras que si observamos la figura 1 podemos observar que el rango real de dicho campo en el del dataset es 3-8, es decir, no tenemos vinos de todas las calidades y esto condicionará la forma en que se realizará el modelo de clasificación.

3. Objetivos a cumplir

Como ya se ha comentado, este dataset contiene la composición química de diferentes vinos así como una nota del 0 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 clasificación 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, lo que se quiere obtener es el mejor modelo de clasificación posible en función de los datos que contiene el dataset.

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.

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:

```
> #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 2. Número de ceros de los diferentes campos que componen el dataset

Como se puede observar en la Figura 2, 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.

```
> #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
> |
```

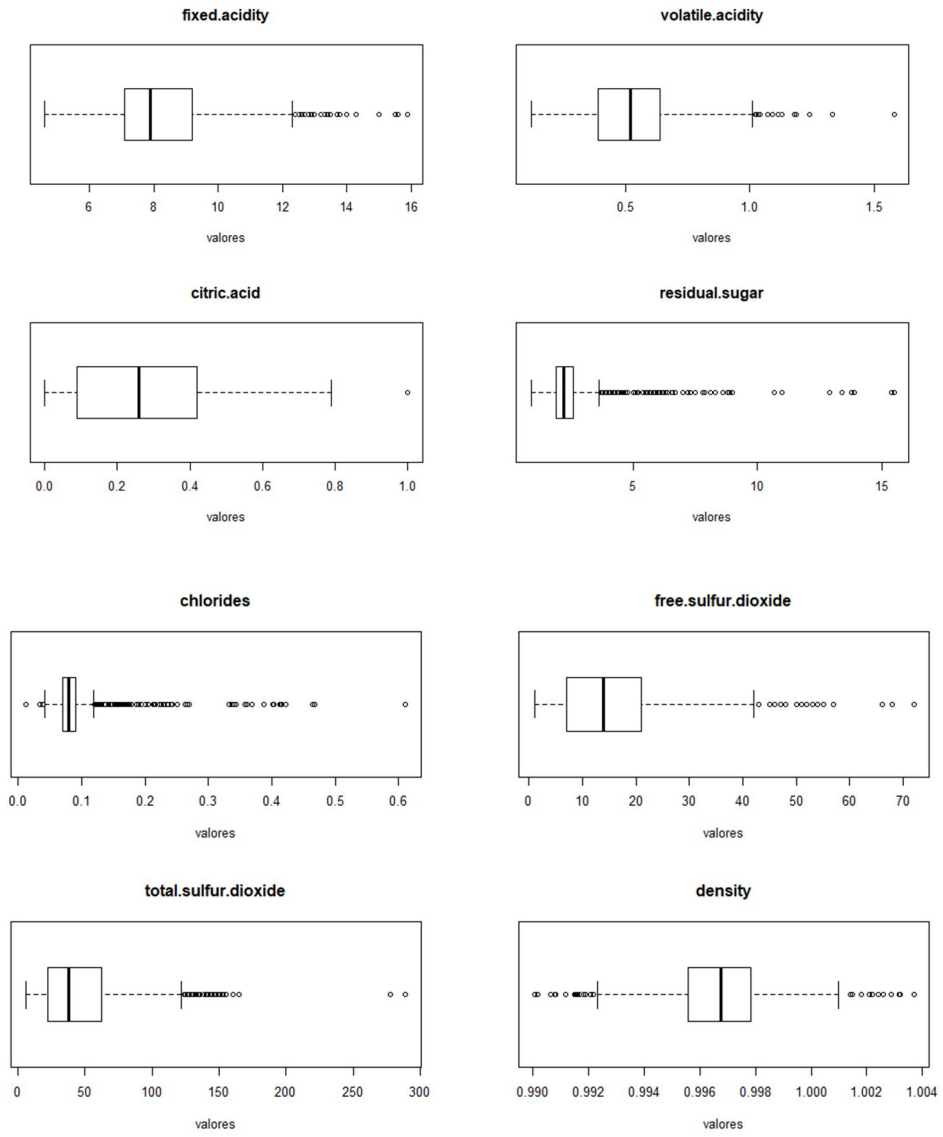
Figura 3. Número de nulos de los diferentes campos que componen el dataset

Como se puede observar en la Figura 3, 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.

Identificación y tratamiento de los valores extremos

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 para ver los diferentes cuartiles y ver su distribución que ya se puede intuir viendo los datos estadísticos de dicho campo en la Figura 1.



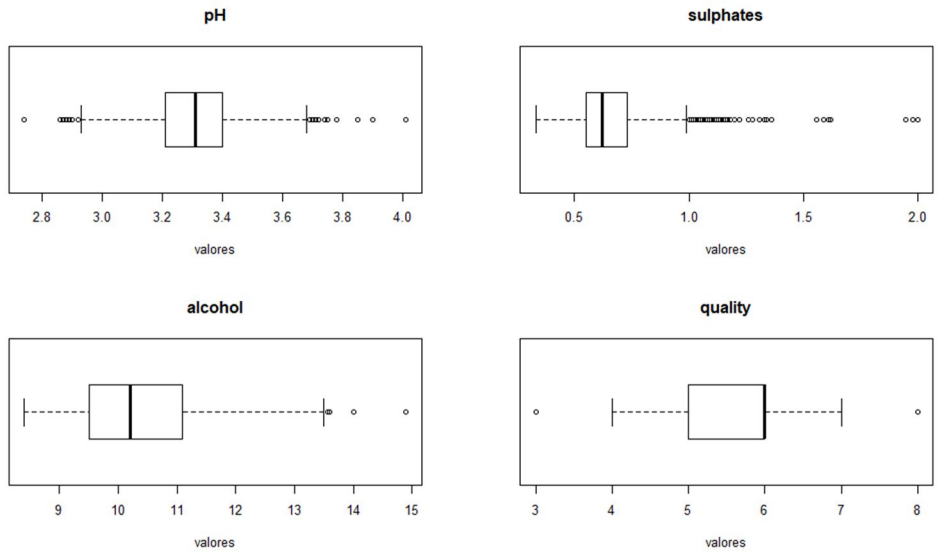


Figura 4. Gráficos de cajas de los diferentes campos que componen el dataset.

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(https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009), 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.

Detección de distribuciones normales y homogeneidad de la varianza

En primer lugar se va a proceder a comprobar si los distintos campos siguen una distribución de población normal o no dado que esto da información relevante sobre los campos además de condicionar el tipo de pruebas a realizar.

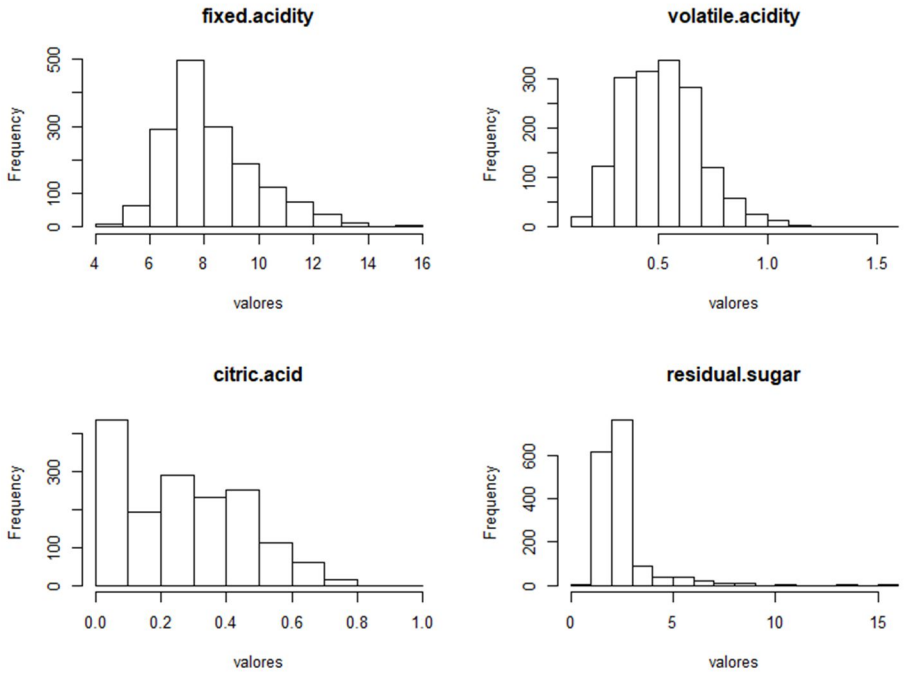
Para comprobar la normalidad emplearemos el método Anderson-Darling con $\alpha = 0.05$. Para cada uno de los campos se calculará el p-value mediante este método y si p-value > α significará que dicho campo sigue una distribución normal y en caso contrario no la seguirá. Además de esto, se generarán los histogramas de cada variable para ver visualmente como se distribuyen los datos de cada uno de los campos.

En la tabla 1, se pueden ver los diferentes valores de p-value para cada uno de los campos.

Campo	p-value
fixed acidity	2.20E-16
volatile acidity	5.32E-14
citric acid	2.20E-16
residual sugar	2.20E-16
chlorides	2.20E-16
free sulfur dioxide	2.20E-16
total sulfur dioxide	2.20E-16
density	1.23E-09
pH	9.25E-05
sulphates	2.20E-16
alcohol	2.20E-16
quality	2.20E-16

Tabla 1. Valores de p-value para los diferentes campos. El valor más bajo que puede devolver es 2.20E-16, esto significa que para estos campos el valor de p-value era más pequeño pero el programa no es capaz de dar más resolución.

Como puede observarse en la tabla 1, todos los datos de p-value para los diferentes campos son menores a $\alpha = 0.05$ por lo que ninguno de los campos sigue una distribución normal.



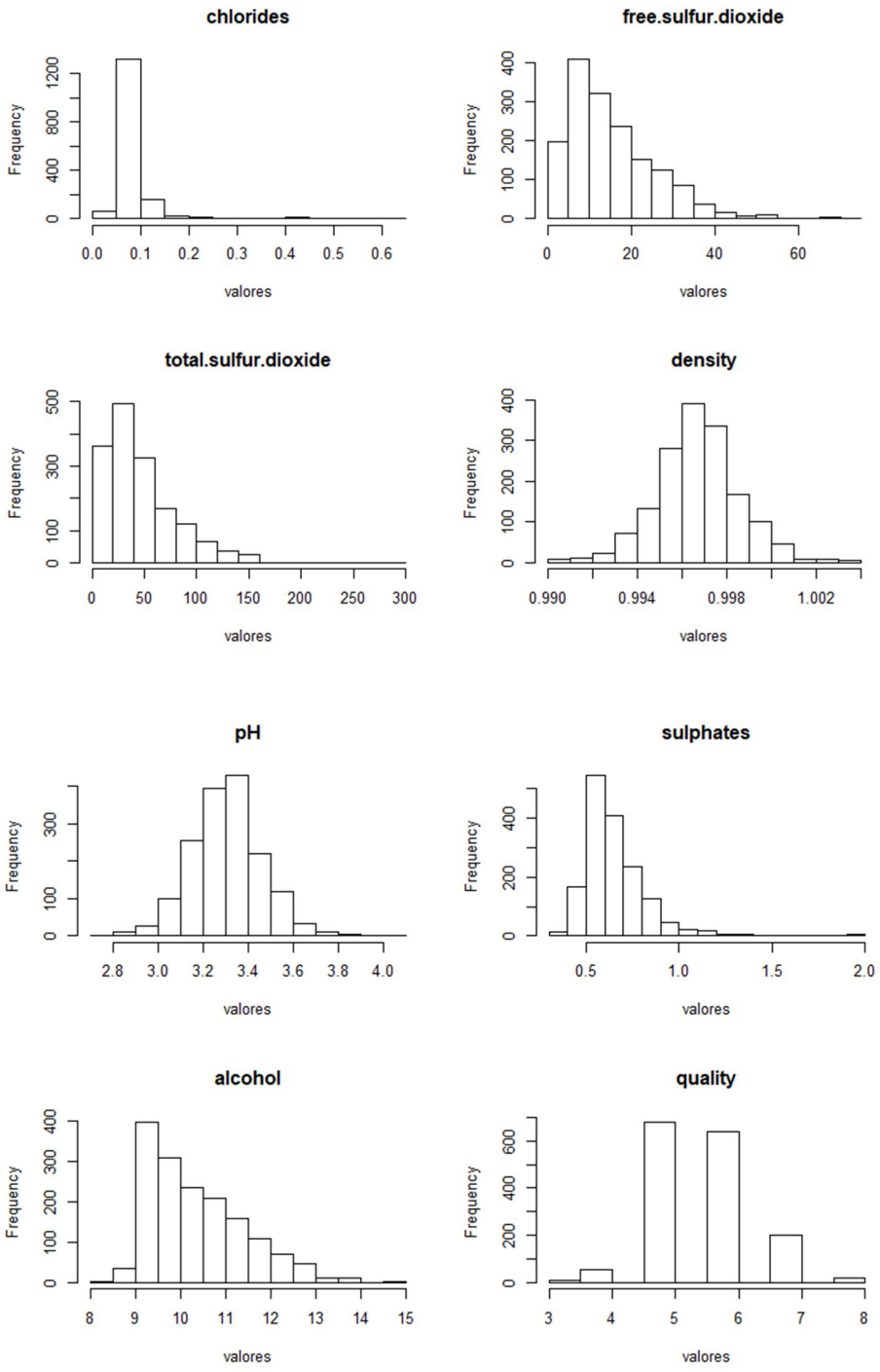


Figura 4. Histogramas de los diferentes campos que componen el dataset.

Si observamos los diferentes histogramas de los que se compone la figura 4, se puede observar que los campos no siguen una distribución normal. No obstante, los campos pH y Density, visualmente sí que recuerdan a una distribución normal de valores aunque el test realizado lo desmienta.

Para comprobar la homogeneidad de la varianza se empleará el método de Fligner-Killeen con $\alpha = 0.05$. Esto es, si el p-value > α, la variancia será homogénea. Dado que el campo por el que queremos crear el modelo de clasificación es quality, miraremos la homogeneidad de la

variancia de todos los campos con respecto al campo quality. Los valores de p-value se encuentran en la tabla 2.

Campo	p-value
fixed acidity	1.78E-06
volatile acidity	1.01E-05
citric acid	0.05307
residual sugar	0.1563
chlorides	0.01002
free sulfur dioxide	0.01476
total sulfur dioxide	2.20E-16
density	1.84E-09
pH	0.9408
sulphates	0.09398
alcohol	2.20E-16

Tabla 2. Valores de p-value para los diferentes campos referentes a la homogeneidad de la varianza. El valor más bajo que puede devolver es 2.20E-16, esto significa que para estos campos el valor de p-value era más pequeño pero el programa no es capaz de dar más resolución.

Observando los datos de la tabla 2, se puede comprobar que la varianza es homogénea o no dependiendo del campo. En el caso que no es homogénea significa que los vinos de los cuales provienen los datos corresponden a grupos diferentes. Pese a que hay campos con varianza homogénea a grandes rasgos se puede decir que los datos provienen de poblaciones distintas(diferentes tipos de uva, procedencia, composiciones químicas diferentes..).

Peso de los diferentes campos en la calidad del vino

Para comprobar el peso de cada uno de los campos en la calidad del vino se va a determinar si los valores de cada campo presentan diferencias significativas con respecto al campo quality mediante el método Krustal-Wallis con $\alpha = 0.05$. Esto significa que si al calcular el p-value de cada campo con respecto al campo quality, el p-value > α significará que no hay diferencias significativas y por ende que dicho campo no aporta información útil para la determinación de la calidad del vino y por lo tanto podemos eliminarlo de cara a realizar los modelos.

Campo	p-value
fixed acidity	1.15E-05
volatile acidity	2.20E-16
citric acid	2.20E-16
residual sugar	0.2339
chlorides	2.71E-14
free sulfur dioxide	8.84E-06
total sulfur dioxide	2.20E-16
density	9.58E-13
pH	0.000244
sulphates	2.20E-16
alcohol	2.20E-16

Tabla 3. Valores de p-value para los diferentes campos referentes a su peso con respecto al campo quality. El valor más bajo que puede devolver es 2.20E-16, esto significa que para estos campos el valor de p-value era más pequeño pero el programa no es capaz de dar más resolución.

Tal y como puede verse en la tabla 3, el campo residual sugar tiene un valor p-value > α. Esto significa que no hay diferencias significativas entre los valores de dicho campo para los diferentes tipos de calidad de vinos y por lo tanto que aporta poca información. Por ello dicho campo no se va a emplear para la creación de los diferentes modelos que se comentarán en el siguiente apartado.

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.
- 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. Para evitar esto, lo que se hará será discretizar el campo quality en tres grupos:
    - Calidad baja si quality < 5
    - Calidad normal si 5 <= quality <= 6
    - Calidad normal si quality >= 7

  Se ha escogido esta división dado que la calidad más baja es 3 y la más alta es 8. Además, para la determinación de los grupos se han teniendo en cuenta que los vinos malos y los vinos de calidad alta serán los menos abundantes y los normales los más abundantes. Esta última afirmación se sustenta de la descripción del dataset de https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009 en la cual se dice precisamente que los vinos más abundantes son los de calidad normal.

- 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.

Con esto en mente, se ha implementado el modelo y se ha obtenido un modelo formado por 30 reglas que 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, Predicho =p1)
        Predicho
ytest    Alta Baja Normal
  Alta     21    0     27
  Baja      1    0      9
  Normal   14    0    248
```

En la imagen anterior 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. El error de este modelo se ha calculado mediante validación cruzada y se ha obtenido los siguientes valores del error:

```
[1] 0.18125
[1] 0.11875
[1] 0.14375
[1] 0.1
[1] 0.16875
[1] 0.2075472
[1] 0.18125
[1] 0.16875
[1] 0.15
[1] 0.16875
> #el error final es la media de los errores
> error_total = error_total / 10.0
> print(error_total)
[1] 0.1588797
>
```

Por lo que el error final, es el promedio de los errores que es del 15.88%.

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 normal 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 ha establecido que ninguno de los campos seguía una distribución normal y la varianza respecto al campo quality no era homogénea para todos los campos. Luego se ha empleado un método para determinar que campos presentaban una diferencia estadísticamente significativa en cuanto a la determinación de la calidad del vino y se ha descartado el campo residual sugar por no aportar información estadísticamente significativa.
- Se ha discretizado el campo quality en tres categoría (baja, normal y alta) para clasificarlos en un grupo u otro los vinos en vez de determinar exactamente el valor de su calidad dado que había calidades que faltaban.
- 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 que ha sido del 15.88.
- 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. No obstante, dichos campos se utilizaron para determinar la calidad del vino, pese a ello se ha obtenido un modelo de clasificación en base a los datos que se tenían con un error razonable.

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
d) Libro manual: Dalgaard, Peter. Introductory statistics with R (Second Edition). New York : Springer, 2002. ISBN 038722632X
e) https://stackoverflow.com/questions/50991870/how-do-i-know-what-distribution-of-data-follows-in-r
f) http://www.sthda.com/english/wiki/kruskal-wallis-test-in-r
g) Casos de ejemplo propuestos por los docentes de la asignatura
h) https://rexplorations.wordpress.com/2015/08/11/normality-tests-in-r/
i) http://www.sthda.com/english/wiki/kruskal-wallis-test-in-r
j) https://www.statisticssolutions.com/the-assumption-of-homogeneity-of-variance/
k) https://www.graphpad.com/support/faq/what-to-do-when-data-fail-tests-for-homogeneity-of-variance/