Autores: María Inmaculada Rodríguez García, María Gema Carrasco García, Javier González Enrique, Juan Jesús Ruiz Aguilar, Ignacio J. Turias Domínguez. Universidad de Cádiz

Módulo 3. Análisis de datos y *machine lerning* en MATLAB y Phyton

3.1. Estadística descriptiva.

Ejercicio 85_01. Exploración de un *dataset* en Python

```
# Una vez que cargamos el dataset, una de las primeras acciones es mostrar las
primeras filas para tener una primera idea de cómo son los datos. También es
conveniente ver las dimensiones del dataset que hemos cargado, así como los tipos de
dato de cada variable.
# Importamos las librerías necesarias
import pandas as pd
nomFichero = 'winequality-red.csv'
# No es necesario añadir nombres de columnas, ya que el csv los trae.
# Se usa ; como separador, y no la coma por defecto.
datos = pd.read_csv(nomFichero, sep=";")
print("Mostramos las 10 primeras filas del dataset:\n")
print(datos.head(10))
print("\nDimensiones del dataset:\n")
print(datos.shape)
print("\nTipo de datos de cada variable (columna):\n")
print(datos.dtypes)
```

```
Mostramos las 10 primeras filas del dataset:
   fixed acidity       volatile acidity   citric acidity     residual sugar chlorides\

0          7.4                  0.70                0.00                1.9      0.076
1          7.8                  0.88                0.00                2.6      0.098
2          7.8                  0.76                0.04                2.3      0.092
3         11.2                  0.28                0.56                1.9      0.075
4          7.4                  0.70                0.00                1.9      0.076
5          7.4                  0.66                0.00                1.8      0.075
6          7.9                  0.60                0.06                1.6      0.069
7          7.3                  0.65                0.00                1.2      0.065
8          7.8                  0.58                0.02                2.0      0.073
9          7.5                  0.50                0.36                6.1      0.071

free sulfur dioxide        total sulfur dioxide     density         ph      sulphates
0               11.0                        34.0      0.9978       3.51       0.56
1               25.0                        67.0      0.9968       3.20       0.68
2               15.0                        54.0      0.9970       3.26       0.65
3               17.0                        60.0      0.9980       3.16       0.58
4               11.0                        34.0      0.9978       3.51       0.56
5               13.0                        40.0      0.9978       3.51       0.56
6               15.0                        59.0      0.9964       3.30       0.46
7               15.0                        21.0      0.9946       3.39       0.47
8                9.0                        18.0      0.9968       3.36       0.57
9               17.0                       102.0      0.9978       3.35        0.8

   alcohol   quality
0      9.4         5
1      9.8         5
2      9.8         5
3      9.8         6
4      9.4         5
```

```
5      9.4         5
6      9.4         5
7     10.0         7
8      9.5         7
9     10.5         5

Dimensiones del dataset:
(1599, 12)

Tipo de datos de cada variable (columna):

fixed acidity             float64
volatile acidity          float64
citric acidity            float64
residual sugar            float64
chlorides                 float64
free sulfur dioxide       float64
total sulfur dioxide      float64
density                   float64
ph                        float64
sulphates                 float64
alcohol                   float64
quality                   int64
dtype: object
```

Ejercicio 85_02. Sumario estadístico MATLAB

```
% Importar datos desde un archivo CSV en MATLAB
nomFichero = 'winequality-red.csv';
datos = readtable(nomFichero, 'Delimiter', ',');
% Configurar el formato de impresión
format shortG; % Mostrar tres decimales
% Mostrar el sumario estadístico
disp('Sumario estadístico:');
disp(summary(datos));


Sumario estadístico:
                  Var1: [1×1 struct]
          fixedAcidity: [1×1 struct]
       volatileAcidity: [1×1 struct]
         citricAcidity: [1×1 struct]
         residualSugar: [1×1 struct]
             chlorides: [1×1 struct]
     freeSulfurDioxide: [1×1 struct]
    totalSulfurDioxide: [1×1 struct]
               density: [1×1 struct]
                    ph: [1×1 struct]
             sulphates: [1×1 struct]
               alcohol: [1×1 struct]
               quality: [1×1 struct]
```

Autores: María Inmaculada Rodríguez García, María Gema Carrasco García, Javier González Enrique, Juan Jesús Ruiz Aguilar, Ignacio J. Turias Domínguez. Universidad de Cádiz

Ejercicio 86_01. Sumario estadístico Python

```python
# Mediante pd.describe() obtenemos una estadística descriptiva que resume la
tendencia central, la dispersión y la forma de la distribución de un conjunto de
datos, excluyendo los valores de NaN.
import pandas as pd
nomFichero = 'winequality-red.csv'
# No es necesario añadir nombres de columnas, ya que el csv los trae.
# Se usa ; como separador, y no la coma por defecto.
datos = pd.read_csv(nomFichero, sep=";")
# Mostrar sumario con tres decimales
pd.set_option('precision', 3)
# Para mostrar la información de todas las columnas
pd.set_option('max_columns', None)
print("Sumario estadístico: ")
print(datos.describe)
```

```
Sumario estadístico:
<bound method NDFrame.describe of
   fixed acidity       volatile acidity   citric acidity     residual sugar chlorides\

0          7.4                  0.700                0.00                1.9      0.076
1          7.8                  0.880                0.00                2.6      0.098
2          7.8                  0.760                0.04                2.3      0.092
3         11.2                  0.280                0.56                1.9      0.075
4          7.4                  0.700                0.00                1.9      0.076
…            …                      …                   …                  …          …

1594       6.2                  0.600                0.08                2.0      0.090
1595       5.9                  0.550                0.10                2.2      0.062
1596       6.3                  0.510                0.13                2.3      0.076
1597       5.9                  0.640                0.12                2.0      0.075
1598       6.0                  0.310                0.47                3.6      0.067

free sulfur dioxide        total sulfur dioxide     density          ph      sulphates
0                 11.0                        34.0      0.998        3.51       0.56
1                 25.0                        67.0      0.997        3.20       0.68
2                 15.0                        54.0      0.997        3.26       0.65
3                 17.0                        60.0       0.998       3.16       0.58
4                 11.0                        34.0      0.998        3.51       0.56
…                    …                           …         …           …          …
1594              32.0                        44.0      0.995        3.30       0.58
1595              39.0                        51.0      0.995        3.39       0.76
1596              29.0                        40.0      0.996        3.36       0.75
1597              32.0                        44.0      0.995        3.35       0.71
1598              18.0                       42.0       0.995       3.39        0.66

   alcohol   quality
0      9.4         5
1      9.8         5
2      9.8         5
3      9.8         6
4      9.4         5
…      …           …
1594    10.5        5
```

```
1595    11.2       6
1596    11.0       6
1597    10.2       5
1598    11.0       6

 [1599 rows x 12 columns]
```

Ejercicio 86_02. Exploración de un dataset en MATLAB

```
% Importar datos desde un archivo CSV en MATLAB
nomFichero = 'winequality-red.csv';
datos = readtable(nomFichero, 'Delimiter', ',');
% Mostrar las 10 primeras filas del dataset
disp('Mostramos las 10 primeras filas del dataset:');
disp(head(datos, 10));
% Mostrar las dimensiones del dataset
disp('Dimensiones del dataset:');
disp(size(datos));
% Mostrar el tipo de datos de cada variable (columna)
disp('Tipo de datos de cada variable (columna):');
disp(class(datos));
```

```
Mostramos las 10 primeras filas del dataset:

   fixedAcidity    volatileAcidity    citricAcidity    residualSugar        chlorides
   ____________    _______________    _____________    _____________        _________

          7.4              0.70              0.00             1.9             0.076
          7.8              0.88              0.00             2.6             0.098
          7.8              0.76              0.04             2.3             0.092
         11.2              0.28              0.56             1.9             0.075
          7.4              0.70              0.00             1.9             0.076
          7.4              0.66              0.00             1.8             0.075
          7.9              0.60              0.06             1.6             0.069
          7.3              0.65              0.00             1.2             0.065
          7.8              0.58              0.02             2.0             0.073
          7.5              0.50              0.36             6.1             0.071


  freeSulfurDioxide    totalSulfurDioxide       density        ph      sulphates
  _________________    __________________       _______      ____      _________

                 11                    34        0.9978      3.51       0.56
                 25                    67        0.9968      3.20       0.68
                 15                    54        0.997       3.26       0.65
                 17                    60        0.998       3.16       0.58
                 11                    34        0.9978      3.51       0.56
                 13                    40        0.9978      3.51       0.56
                 15                    59        0.9964      3.30       0.46
                 15                    21        0.9946      3.39       0.47
                  9                    18        0.9968      3.36       0.57
                 17                   102        0.9978      3.35        0.8

     alcohol    quality
     _______    _______
```

```
         9.4         5
         9.8         5
         9.8         5
         9.8         6
         9.4         5
         9.4         5
         9.4         5
        10.0         7
         9.5         7
        10.5         5

Dimensiones del dataset:
    1599     12
Tipo de datos de cada variable (columna):
table
```

Ejercicio 87_01. Distribución de clases en problemas de clasificación en Python

```
# En los problemas de clasificación es muy importante conocer cómo se distribuyen las clases. <br>
# Si hay una gran diferencia entre los registros que corresponde a cada clase, puede ser necesario un tratamiento posterior.
import pandas as pd
nomFichero = 'winequality-red.csv'
# No es necesario añadir nombres de columnas, ya que el csv los trae.
# Se usa ; como separador, y no la coma por defecto.
datos = pd.read_csv(nomFichero, sep=";")
# La variable quality es de tipo categórico, con valores de 1 a 5
conteo_por_clases = datos.grouphy('quality').size()
print("Nº de registros por cada clase en la variable quality: ")
print(conteo_por_clases)
```

```
Nº de registros por cada clase en la variable quality:
quality
3    10
4    53
5    681
6    638
7    199
8    18
dtype: int64
```

Ejercicio 87_02. Distribución de clases en problemas de clasificación en MATLAB

```
% Importar datos desde un archivo CSV en MATLAB
nomFichero = 'winequality-red.csv';
datos = readtable(nomFichero, 'Delimiter', ',');
% Contar el número de registros por cada clase en la variable 'quality'
conteo_por_clases = varfun(@length, datos, 'GroupingVariables', 'quality');
% Mostrar el número de registros por cada clase en la variable 'quality'
disp('Número de registros por cada clase en la variable quality:');
disp(conteo_por_clases);
```

Autores: María Inmaculada Rodríguez García, María Gema Carrasco García, Javier González Enrique, Juan Jesús Ruiz Aguilar, Ignacio J. Turias Domínguez. Universidad de Cádiz

```
Número de registros por cada clase en la variable quality:
quality    GroupCount    length_fixedAcidity    length_volatileAcidity
_______    __________    ___________________    ______________________
3           10                 10                     10
4           53                 53                     53
5          681                681                    681
6          638                638                    638
7          199                199                    199
8           18                 18                     18
length_citricAcid    length_residualSugar    length_chlorides
_________________    ____________________    ________________
10                    10                     10
53                    53                     53
681                   681                    681
638                   638                    638
199                   199                    199
18                    18                     18
length_freeSulfurDioxide    length_totalSulfurDioxide    length_density
________________________    _________________________    ______________
10                           10                          10
53                           53                          53
681                          681                         681
638                          638                         638
199                          199                         199
18                           18                    18
length_pH     length_sulphates      length_alcohol
_________     ________________      ______________
 10             10                  10
 53             53                  53
 681            681                 681
 638            638                 638
 199            199                 199
 18             18                  18
```