Universidad de Córdoba

Escuela Politécnica Superior de Córdoba

Grado de Ingeniería Informática - Mención en computación

Tercer curso - Segundo cuatrimestre - 2020/2021

Introducción al Aprendizaje Automático

Práctica 2: Exploración de datos con Pandas

Profesor: *Nicolás Emilio García Pedrajas*

Autor: *Ventura Lucena Martínez*

Córdoba, 30 de junio de 2021

Índice

1 Ejercicio 1 **1**

2 Ejercicio 2 **1**

3 Ejercicio 3 **3**
3.1 Normalización 3
3.2 Estandarización 5

4 Ejercicio 4 **7**
4.1 *weather.arff* 7
4.2 *glass.arff* 9
4.3 *iris.arff* 10

5 Ejercicio 5 **12**

6 Ejercicio 6 **13**

7 Ejercicio 7 **15**

8 Ejercicio 8 **17**

9 Ejercicio 9 **19**

Índice de figuras

1 Histograma del conjunto de datos *weather.arff* . . . 1
2 Histograma del conjunto de datos *iris.arff* . . . 2
3 Histograma del conjunto de datos *glass.arff* . . . 2
4 Histograma normalizado del conjunto de datos *weather.arff* . . . 3
5 Histograma normalizado del conjunto de datos *iris.arff* . . . 4
6 Histograma normalizado del conjunto de datos *glass.arff* . . . 4
7 Histograma estandarizado del conjunto de datos *weather.arff* . . . 5
8 Histograma estandarizado del conjunto de datos *iris.arff* . . . 6
9 Histograma estandarizado del conjunto de datos *glass.arff* . . . 6
10 PCA *weather.arff* . . . 7
11 PCA *glass.arff* . . . 9
12 PCA *iris.arff* . . . 10
13 Diagrama de dispersión - Librería *Seaborn*. . . . 12
14 Diagrama de dispersión - Librería *Pandas*. . . . 13
15 Diagrama de dispersión normalizado. . . . 14
16 Diagrama de dispersión estandarizado. . . . 15
17 PCA sobre el diagrama de dispersión *weather.arff* . . . 16
18 PCA sobre el diagrama de dispersión *glass.arff* . . . 16
19 PCA sobre el diagrama de dispersión *iris.arff* . . . 17
20 Diagrama de correlaciones *weather.arff* . . . 18
21 Diagrama de correlaciones *glass.arff* . . . 18
22 Diagrama de correlaciones *iris.arff* . . . 19
23 Representación en coordenadas paralelas *weather.arff* . . . 20
24 Representación en coordenadas paralelas *glass.arff* . . . 20
25 Representación en coordenadas paralelas *iris.arff* . . . 21

Listings

1 **Función de normalización** 3
2 **Función de estandarización** 5
3 **Resultados datos originales *weather.arff*** 7
4 **Resultados datos PCA *weather.arff*** 8
5 **Resultados datos originales *glass.arff*** 9
6 **Resultados datos PCA *glass.arff*** 10
7 **Resultados datos originales *iris.arff*** 11
8 **Resultados datos PCA *iris.arff*** 11

1 Ejercicio 1

Obtenga tres ejemplos de ficheros de datos en formato CSV, ARFF u otro cualquiera de:

- **Weka datasets.**
- **UCI MLR.**

Los datasets utilizados se encuentran en los siguientes directorios:

- ../weather.arff.
- ../iris.arff.
- ../glass.arff.

2 Ejercicio 2

Usando Pandas, cargue los ficheros y evalúe qué información puede obtener del histograma de atributos.

Los resultados obtenidos en los siguientes histogramas de atributos hacen referencia a las variables númericas de su respectivo conjunto de datos. Para ello, se ha hecho uso de la función *histogram* de la librería de Pandas:

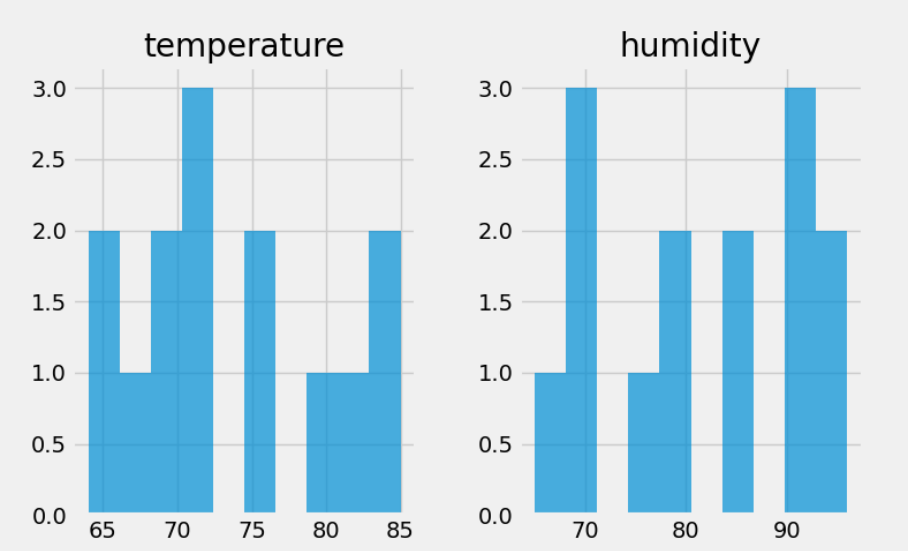

Figura 1: Histograma del conjunto de datos *weather.arff*

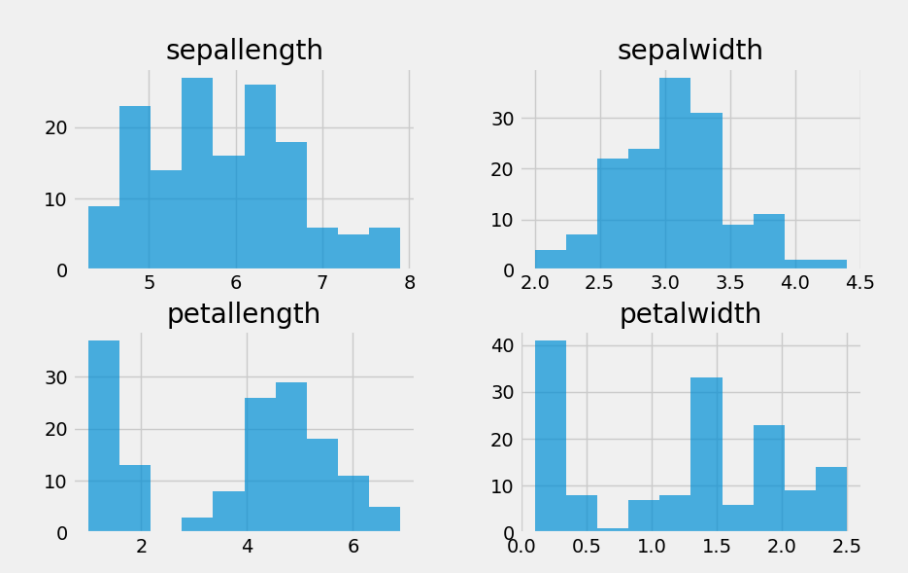

Figura 2: Histograma del conjunto de datos *iris.arff*

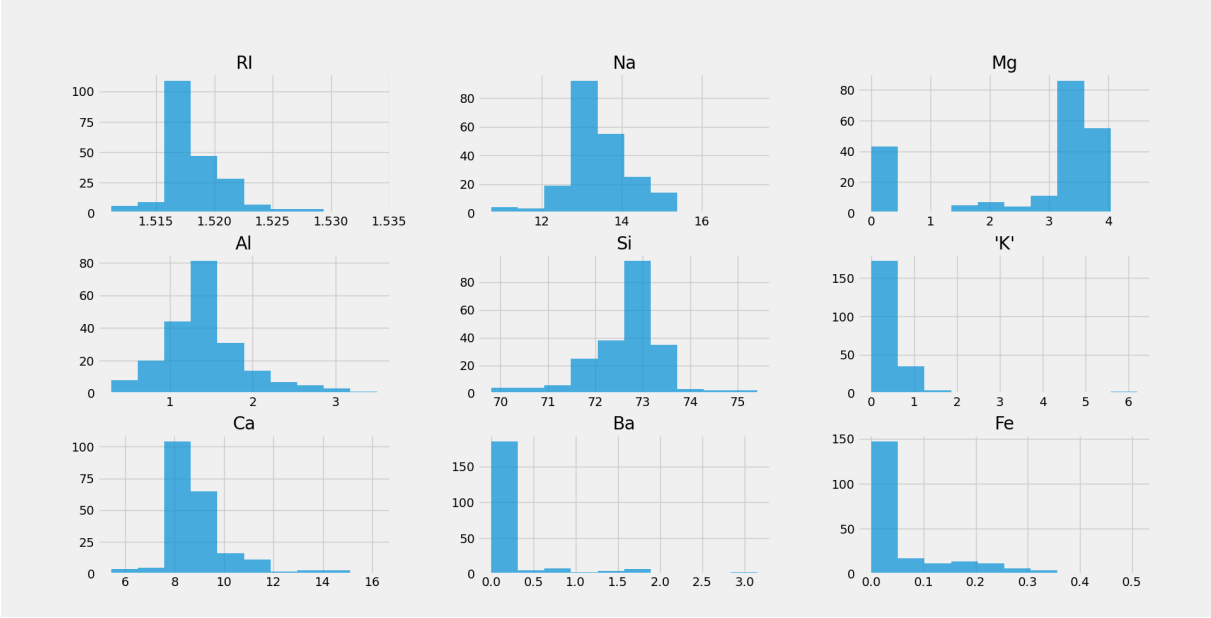

Figura 3: Histograma del conjunto de datos *glass.arff*

3 Ejercicio 3

Estudie el efecto de la normalización (reescalar en el intervalo [0, 1]) y la estandarización ($\mu = 0$, $\sigma = 1$) sobre el histograma.

3.1 Normalización

Para realizar la normalización de un conjunto de datos seleccionado se ha implementado la siguiente funcionalidad:

Listing 1: **Función de normalización**

```
def normalizeData(file_name, df):
    # Normalizes dataset.
    df = (df - df.min()) / (df.max() - df.min())
    printNormalizedData(file_name, df)
```

Una vez implementada, los histogramas normalizados obtenidos son los siguientes:

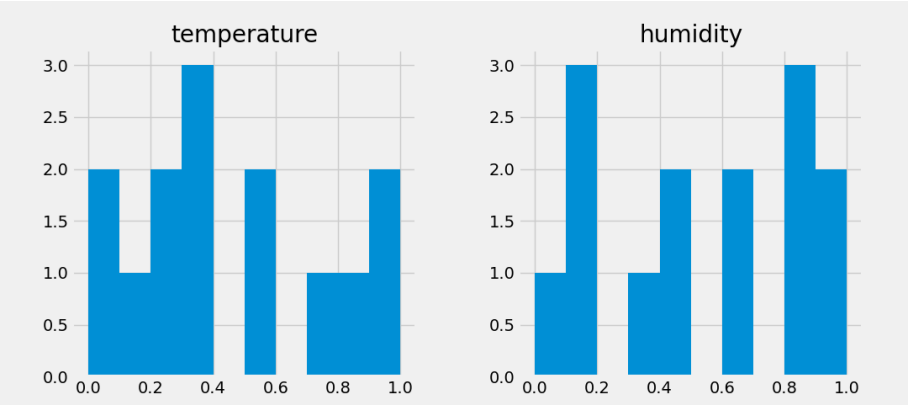

Figura 4: Histograma normalizado del conjunto de datos *weather.arff*

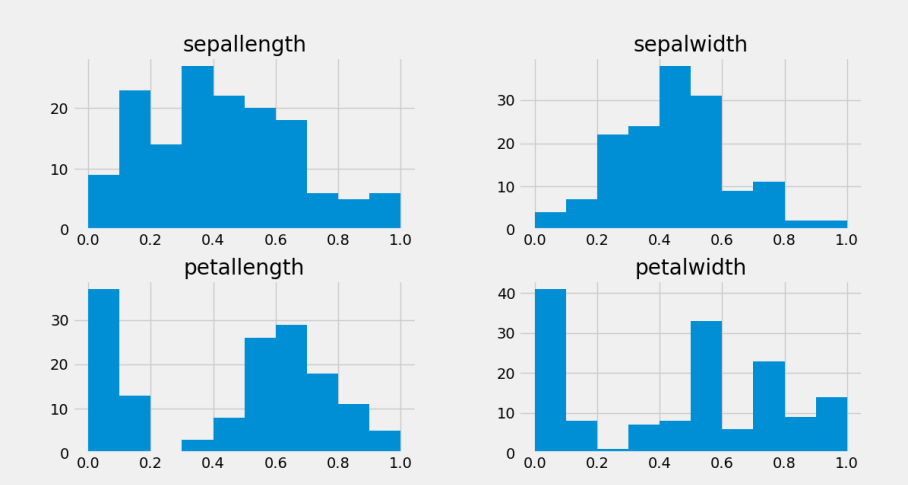

Figura 5: Histograma normalizado del conjunto de datos *iris.arff*

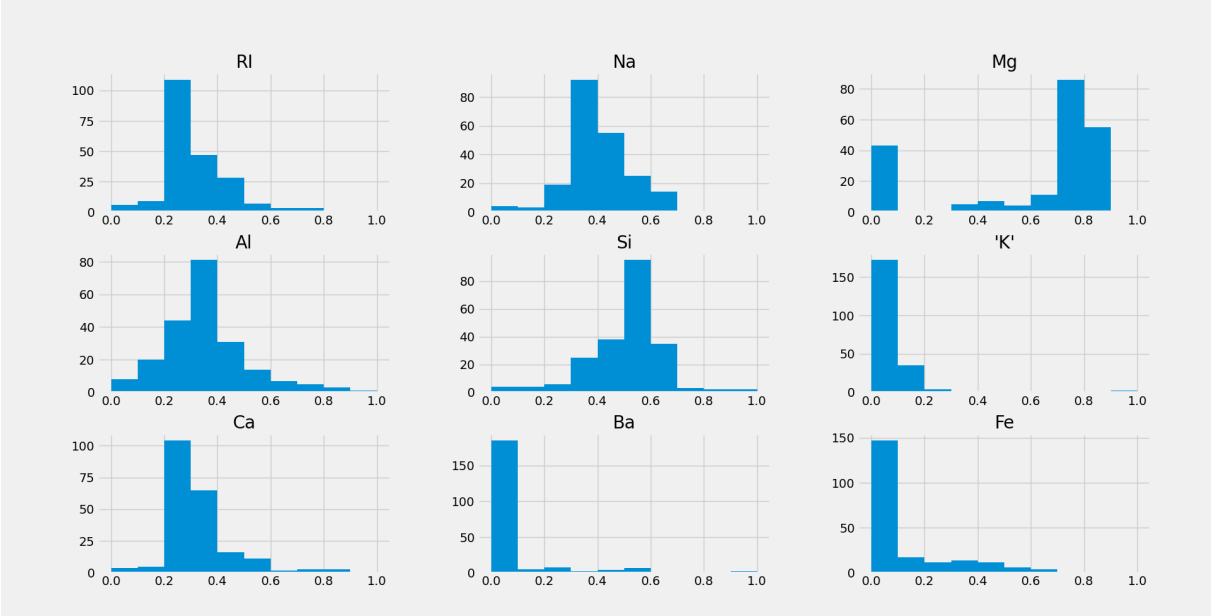

Figura 6: Histograma normalizado del conjunto de datos *glass.arff*

3.2 Estandarización

Para realizar la estandarización de un conjunto de datos seleccionado se ha implementado la siguiente funcionalidad, muy parecida a la implementada en la sección anterior:

Listing 2: **Función de estandarización**

```
def standardizeData(file_name, df):
    # Normalizes dataset.
    df = (df - df.mean()) / df.std()
    printStandardizedData(file_name, df)
```

Una vez implementada, los histogramas estandarizados obtenidos son los siguientes:

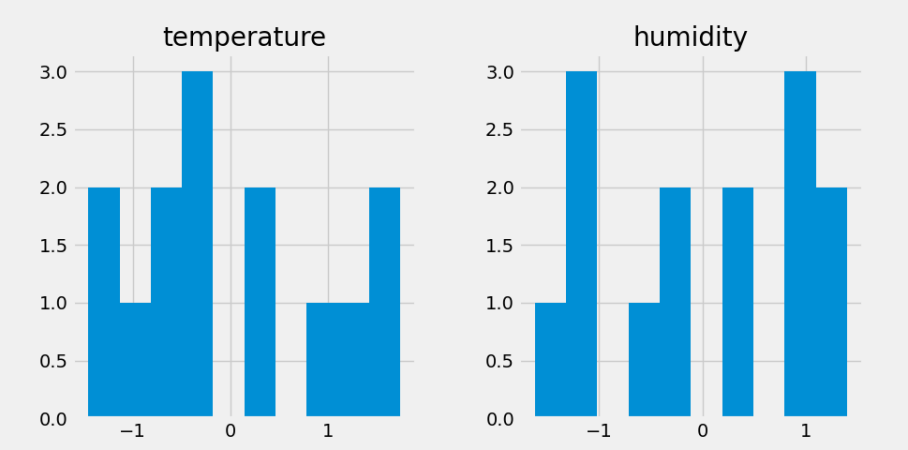

Figura 7: Histograma estandarizado del conjunto de datos *weather.arff*

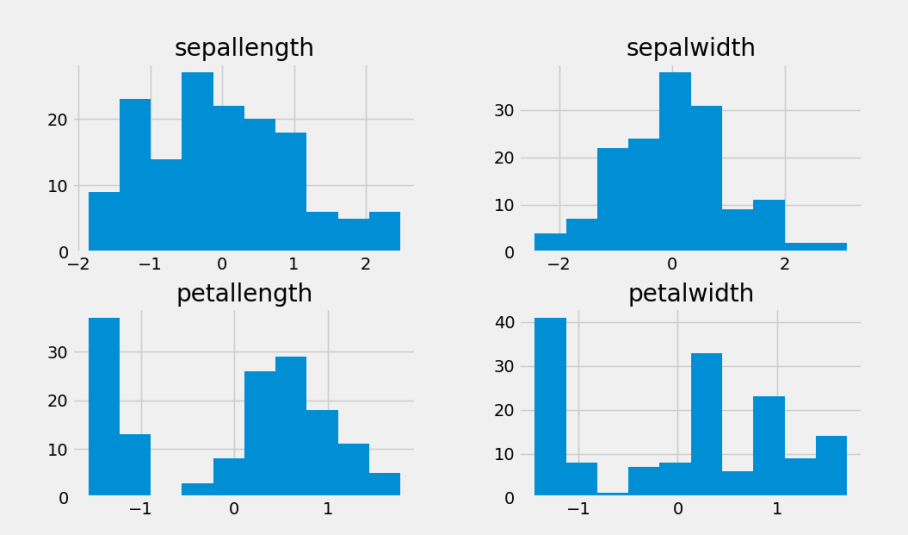

Figura 8: Histograma estandarizado del conjunto de datos *iris.arff*

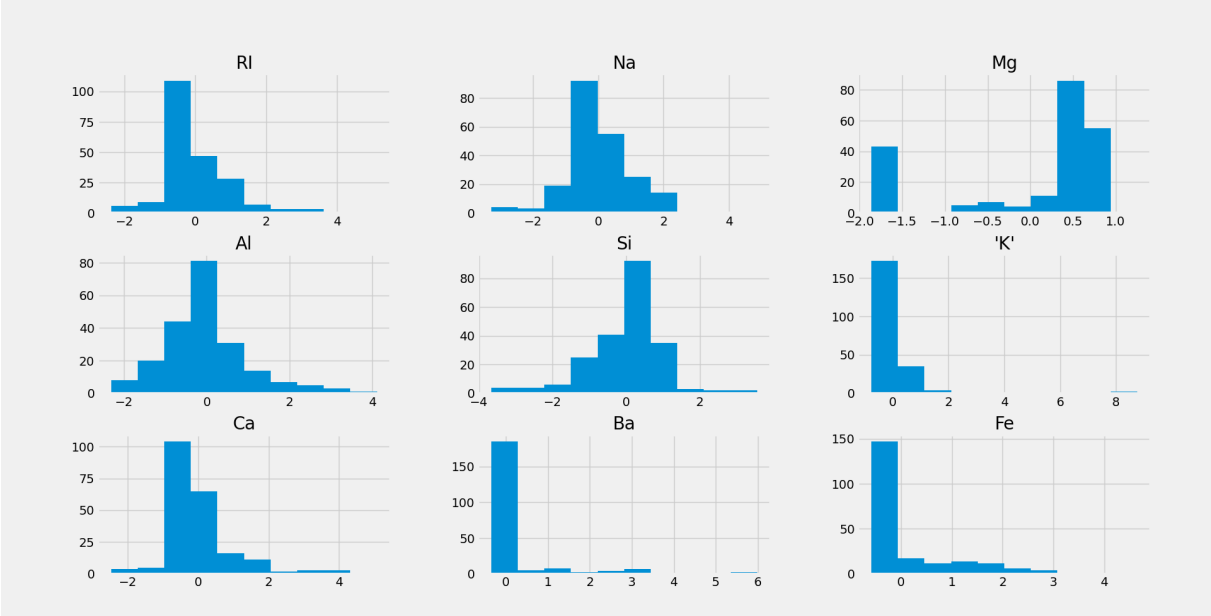

Figura 9: Histograma estandarizado del conjunto de datos *glass.arff*

4 Ejercicio 4

Estudie el efecto del análisis en componentes principales sobre el histograma.

La aplicación de este efecto reduce la dimensionalidad lineal usando la descomposición de valor singular de los datos para proyectarlos en un espacio dimensional menor. Aplicándolo a los 3 conjuntos de datos obtenemos lo siguiente:

4.1 *weather.arff*

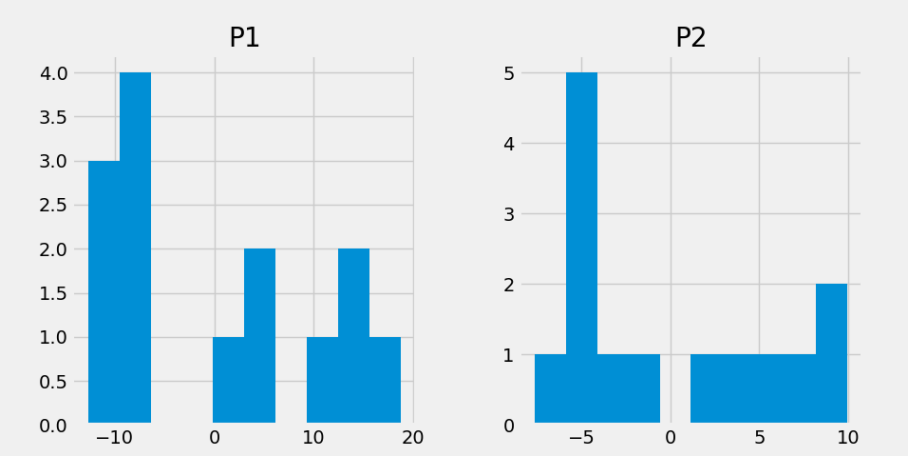

Figura 10: PCA *weather.arff*

Listing 3: **Resultados datos originales *weather.arff***

```
    temperature  humidity
0          85.0      85.0
1          80.0      90.0
2          83.0      86.0
3          70.0      96.0
4          68.0      80.0
5          65.0      70.0
6          64.0      65.0
7          72.0      95.0
8          69.0      70.0
9          75.0      80.0
10         75.0      70.0
11         72.0      90.0
12         81.0      75.0
13         71.0      91.0

        temperature   humidity
count     14.000000  14.000000
mean      73.571429  81.642857
```

```
std        6.571667   10.285218
min       64.000000   65.000000
25%       69.250000   71.250000
50%       72.000000   82.500000
75%       78.750000   90.000000
max       85.000000   96.000000
```

Listing 4: **Resultados datos PCA *weather.arff***

```
            P1         P2
0    -6.571975   9.934375
1    -9.878989   3.684282
2    -6.939070   7.728645
3   -12.670176  -7.638671
4     3.209811  -4.841168
5    13.649841  -4.765222
6    18.722702  -4.249394
7   -12.303081  -5.432942
8    12.472610  -0.942380
9     1.149656   1.848807
10   10.706763   4.791885
11   -7.524527  -3.961403
12    4.162363   9.054610
13   -8.185930  -5.211422
```

```
                    P1             P2
count   1.400000e+01   1.400000e+01
mean   -3.552714e-15   8.881784e-16
std     1.059924e+01   6.052154e+00
min    -1.267018e+01  -7.638671e+00
25%    -8.020579e+00  -4.822182e+00
50%    -2.711159e+00  -2.451891e+00
75%     9.070663e+00   4.514984e+00
max     1.872270e+01   9.934375e+00
```

4.2 *glass.arff*

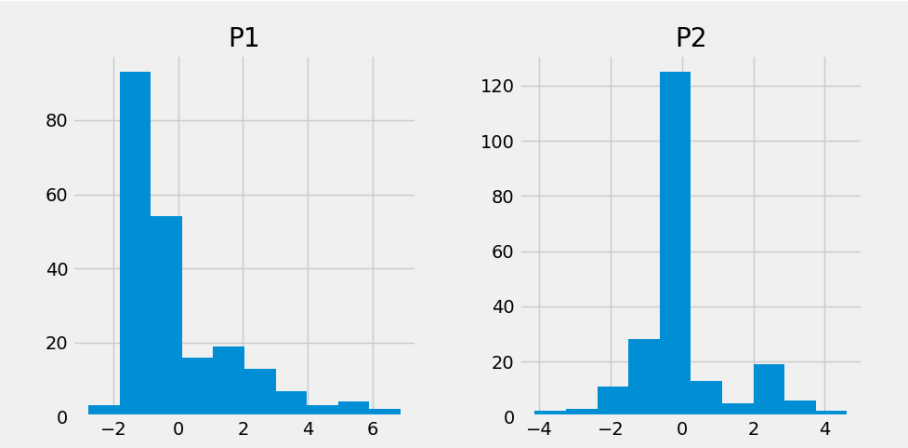

Figura 11: PCA *glass.arff*

Listing 5: **Resultados datos originales *glass.arff***

```
           RI      Na    Mg    Al     Si    'K'     Ca   Ba    Fe
0     1.51793   12.79  3.50  1.12  73.03   0.64   8.77  0.0  0.00
1     1.51643   12.16  3.52  1.35  72.89   0.57   8.53  0.0  0.00
2     1.51793   13.21  3.48  1.41  72.64   0.59   8.43  0.0  0.00
3     1.51299   14.40  1.74  1.54  74.55   0.00   7.59  0.0  0.00
4     1.53393   12.30  0.00  1.00  70.16   0.12  16.19  0.0  0.24
..        ...     ...   ...   ...    ...    ...    ...  ...   ...
209   1.51610   13.42  3.40  1.22  72.69   0.59   8.32  0.0  0.00
210   1.51592   12.86  3.52  2.12  72.66   0.69   7.97  0.0  0.00
211   1.51613   13.92  3.52  1.25  72.88   0.37   7.94  0.0  0.14
212   1.51689   12.67  2.88  1.71  73.21   0.73   8.54  0.0  0.00
213   1.51852   14.09  2.19  1.66  72.67   0.00   9.32  0.0  0.00

[214 rows x 9 columns]

                RI          Na          Mg          Al          Si          'K'
                            Ca          Ba          Fe
count   214.000000  214.000000  214.000000  214.000000  214.000000  214.000000
    214.000000  214.000000  214.000000
mean      1.518365   13.407850    2.684533    1.444907   72.650935    0.497056
       8.956963    0.175047    0.057009
std       0.003037    0.816604    1.442408    0.499270    0.774546    0.652192
       1.423153    0.497219    0.097439
min       1.511150   10.730000    0.000000    0.290000   69.810000    0.000000
       5.430000    0.000000    0.000000
25%       1.516522   12.907500    2.115000    1.190000   72.280000    0.122500
       8.240000    0.000000    0.000000
50%       1.517680   13.300000    3.480000    1.360000   72.790000    0.555000
       8.600000    0.000000    0.000000
```

```
75%       1.519157    13.825000     3.600000     1.630000    73.087500     0.610000
       9.172500     0.000000     0.100000
max       1.533930    17.380000     4.490000     3.500000    75.410000     6.210000
    16.190000     3.150000     0.510000
```

Listing 6: **Resultados datos PCA *glass.arff***

```
            P1         P2
0    -0.770340  -0.607790
1    -0.941573  -0.696785
2    -0.959380  -0.225572
3    -0.213312   2.055071
4     6.854614  -4.124768
..         ...        ...
209  -0.981839  -0.070013
210  -1.282118   0.094865
211  -1.301039   0.318026
212  -0.461254   0.019891
213   0.651698   0.286295

[214 rows x 2 columns]
```

```
                  P1            P2
count   2.140000e+02  2.140000e+02
mean   -5.063447e-16  9.462835e-16
std     1.732631e+00  1.288089e+00
min    -2.771320e+00 -4.124768e+00
25%    -1.168368e+00 -5.347394e-01
50%    -7.680799e-01 -2.527875e-01
75%     6.341453e-01  9.433407e-02
max     6.854614e+00  4.624431e+00
```

4.3 *iris.arff*

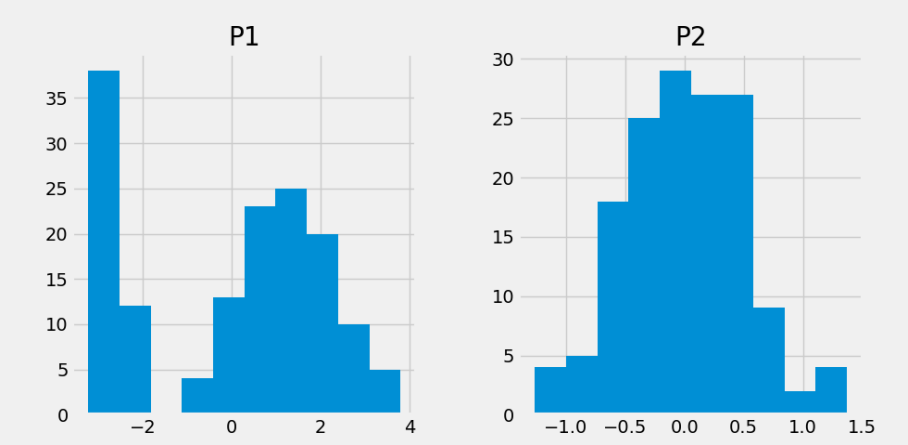

Figura 12: PCA *iris.arff*

Listing 7: **Resultados datos originales *iris.arff***

```
     sepallength  sepalwidth  petallength  petalwidth
0            5.1         3.5          1.4         0.2
1            4.9         3.0          1.4         0.2
2            4.7         3.2          1.3         0.2
3            4.6         3.1          1.5         0.2
4            5.0         3.6          1.4         0.2
..           ...         ...          ...         ...
145          6.7         3.0          5.2         2.3
146          6.3         2.5          5.0         1.9
147          6.5         3.0          5.2         2.0
148          6.2         3.4          5.4         2.3
149          5.9         3.0          5.1         1.8

[150 rows x 4 columns]

        sepallength  sepalwidth  petallength  petalwidth
count    150.000000  150.000000   150.000000  150.000000
mean       5.843333    3.054000     3.758667    1.198667
std        0.828066    0.433594     1.764420    0.763161
min        4.300000    2.000000     1.000000    0.100000
25%        5.100000    2.800000     1.600000    0.300000
50%        5.800000    3.000000     4.350000    1.300000
75%        6.400000    3.300000     5.100000    1.800000
max        7.900000    4.400000     6.900000    2.500000
```

Listing 8: **Resultados datos PCA *iris.arff***

```
           P1        P2
0   -2.684207  0.326607
1   -2.715391 -0.169557
2   -2.889820 -0.137346
3   -2.746437 -0.311124
4   -2.728593  0.333925
..        ...       ...
145  1.944017  0.187415
146  1.525664 -0.375021
147  1.764046  0.078519
148  1.901629  0.115877
149  1.389666 -0.282887

[150 rows x 2 columns]

                  P1            P2
count  1.500000e+02  1.500000e+02
mean   2.842171e-16 -5.210647e-16
std    2.055442e+00  4.921825e-01
min   -3.225200e+00 -1.262492e+00
25%   -2.530159e+00 -3.235986e-01
50%    5.533290e-01 -3.251102e-02
75%    1.549463e+00  3.288601e-01
max    3.794687e+00  1.370524e+00
```

5 Ejercicio 5

Usando la visualización del diagrama de dispersión (scatter plot) estudie qué información puede obtener de dicha representación gráfica.

Los datos obtenidos para los diagramas de dispersión sobre el conjunto de datos *iris.arff* son los siguientes:

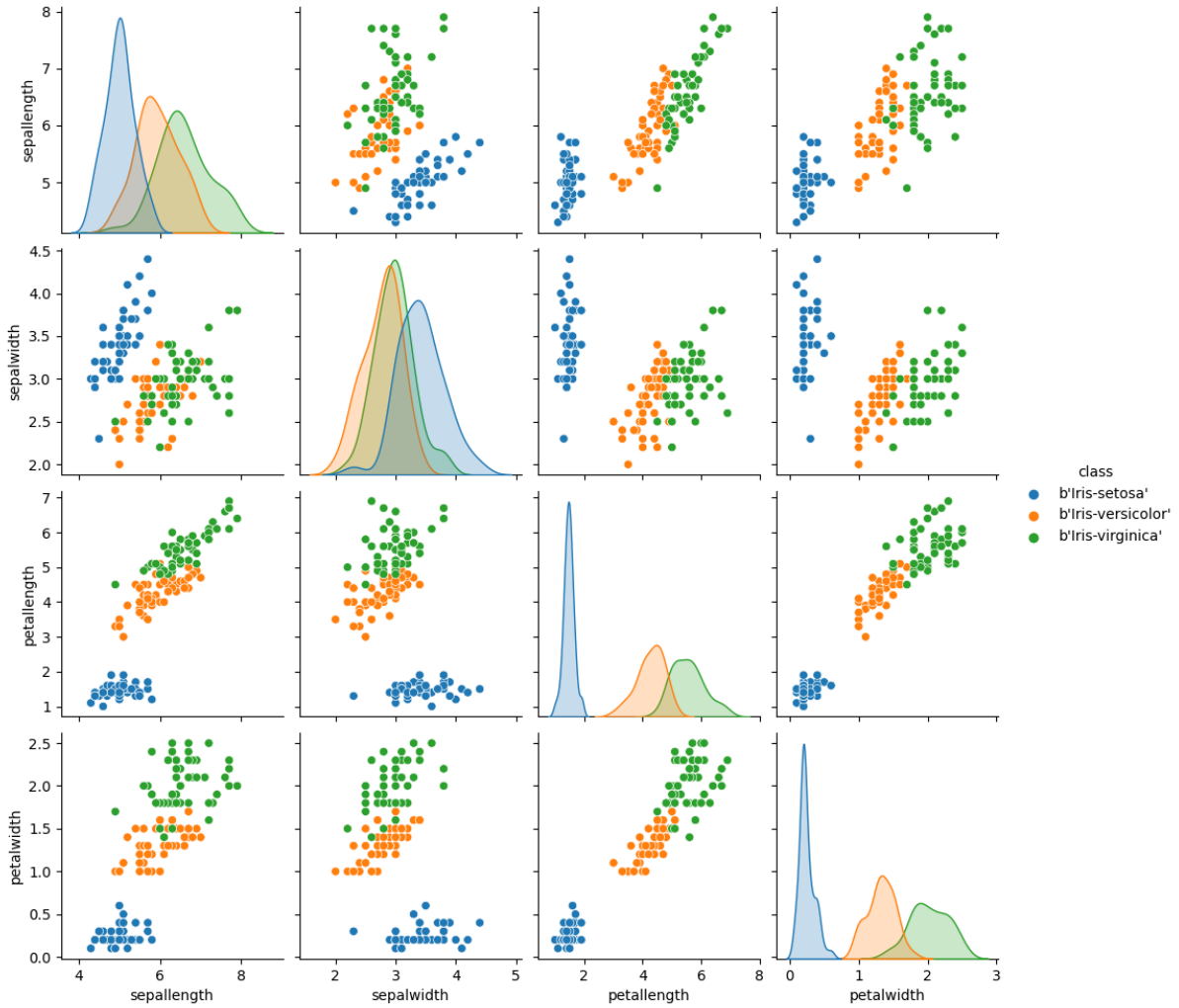

Figura 13: Diagrama de dispersión - Librería *Seaborn*.

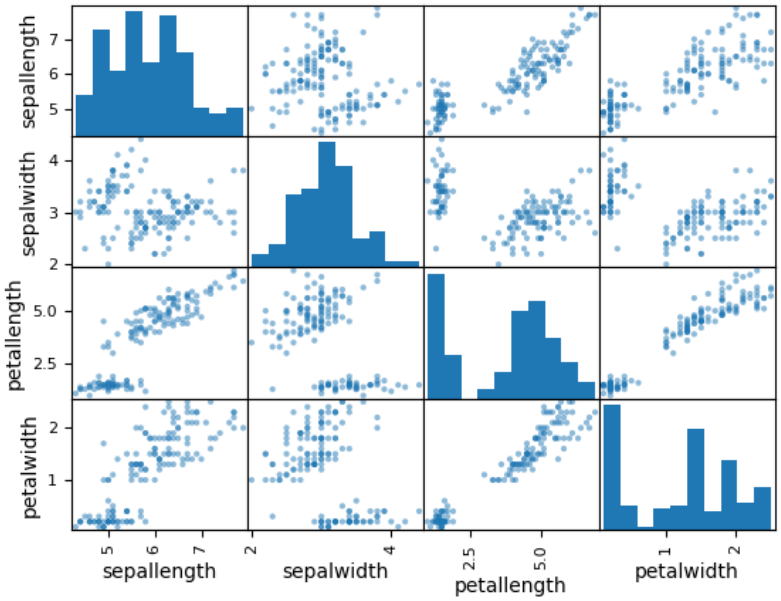

Figura 14: Diagrama de dispersión - Librería *Pandas*.

6 Ejercicio 6

Estudie el efecto de la normalización y la estandarización sobre el diagrama de dispersión.

De igual manera que en el ejercicio 3 visto con anterioridad, se ha procedido de igual manera al realizar la normalización y la estandarización:

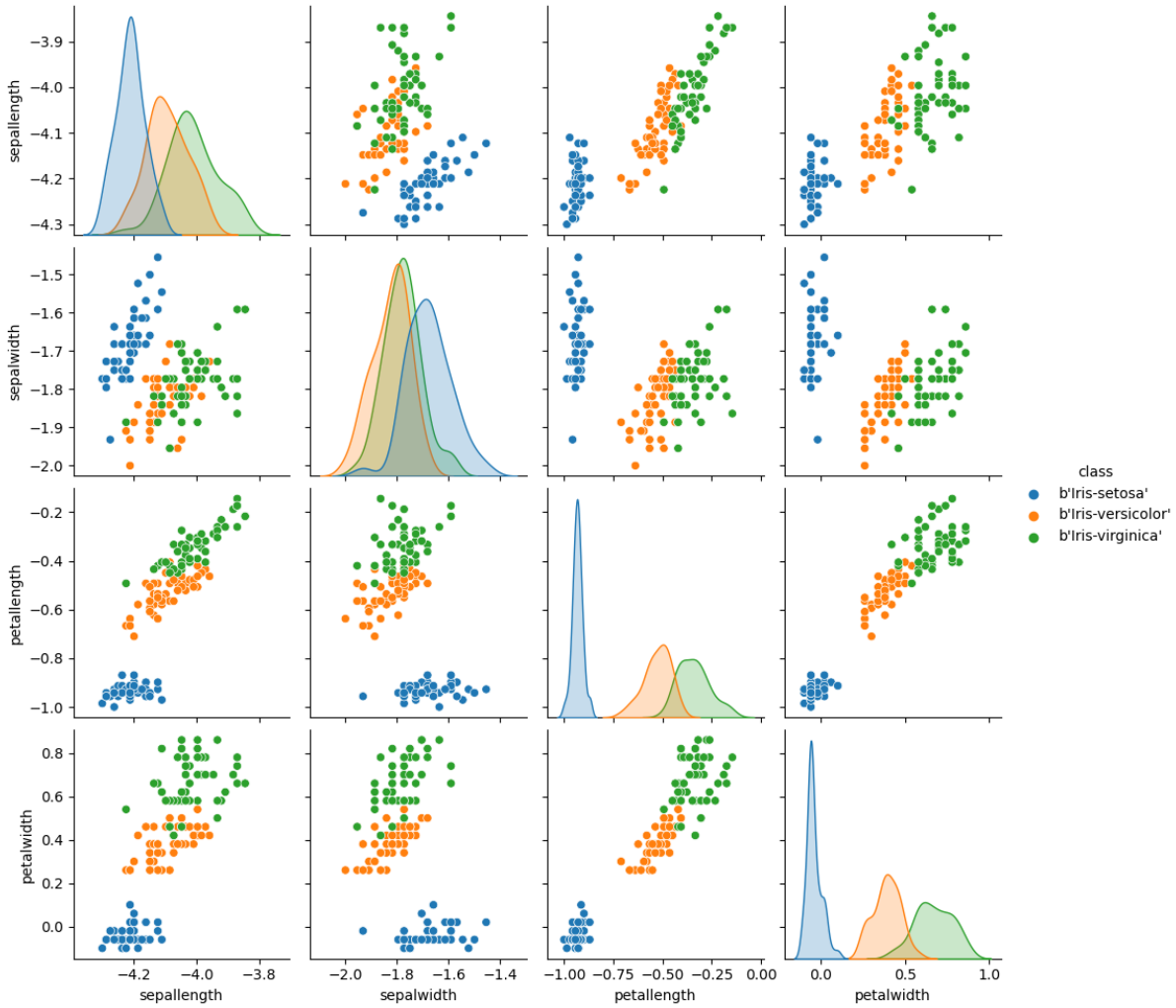

Figura 15: Diagrama de dispersión normalizado.

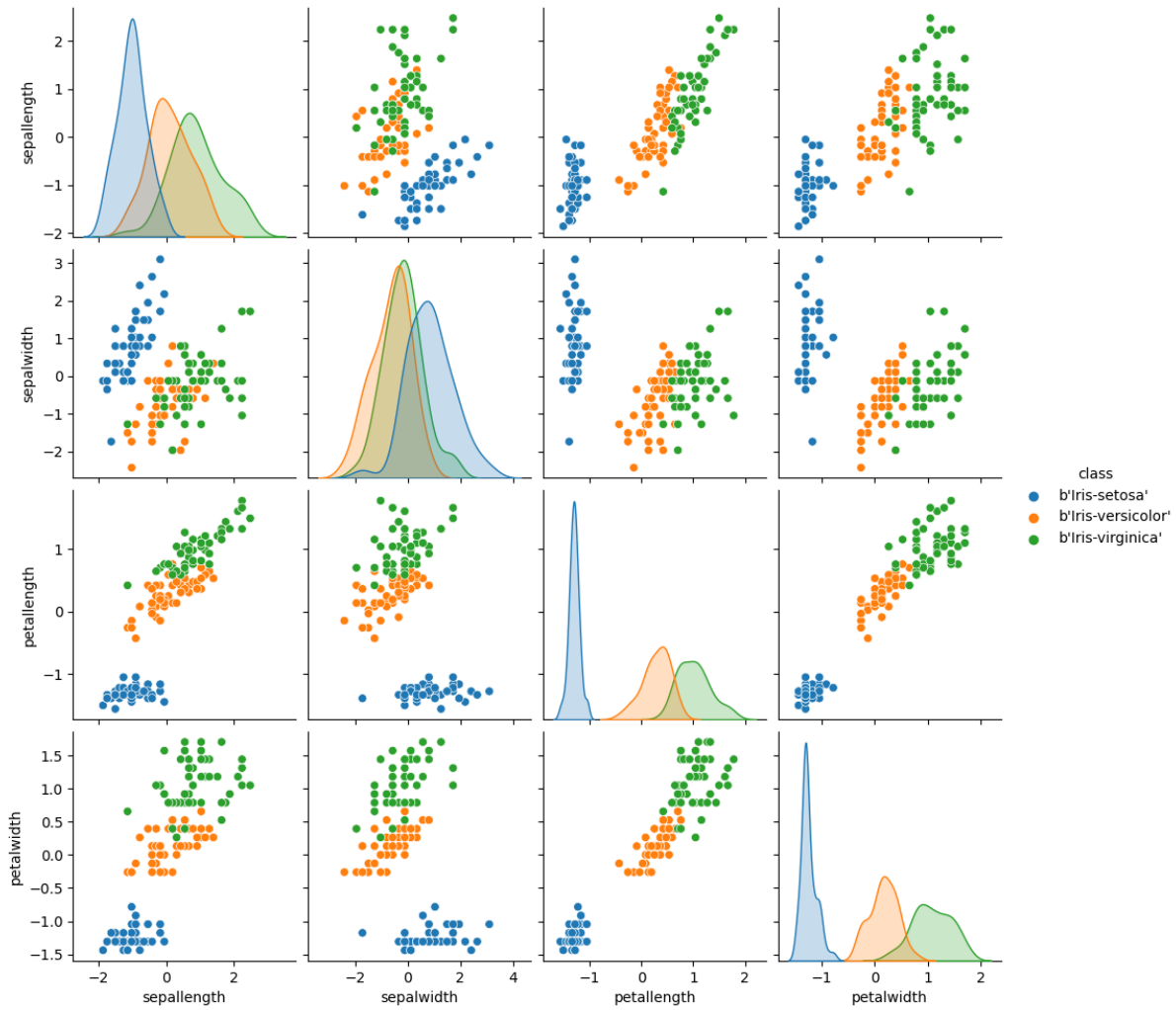

Figura 16: Diagrama de dispersión estandarizado.

7 Ejercicio 7

Estudie el efecto del análisis en componentes principales sobre el diagrama de dispersión.

Se realizará de manera análoga al ejercicio 4, enfocándonos esta vez únicamente en los grafos. Para consultar los datos, se puede ejecutar el código fuente referente al ejercicio.

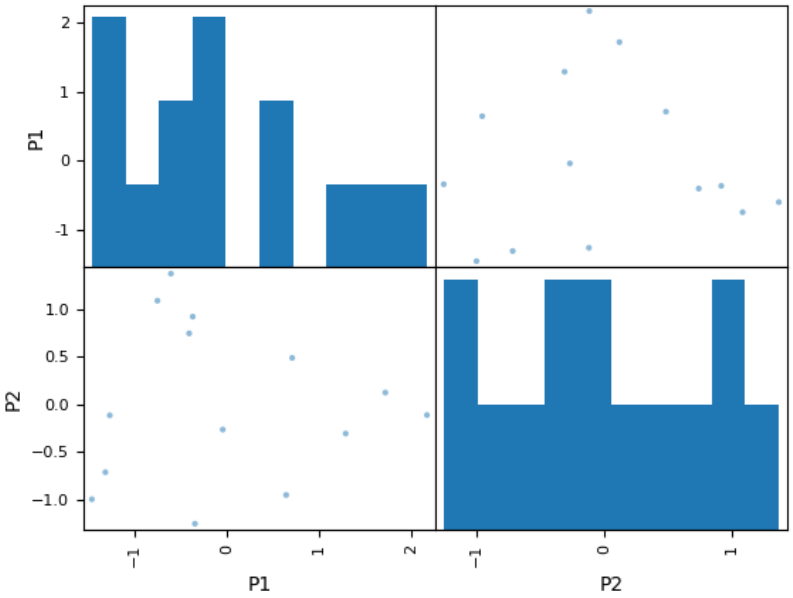

Figura 17: PCA sobre el diagrama de dispersión *weather.arff*

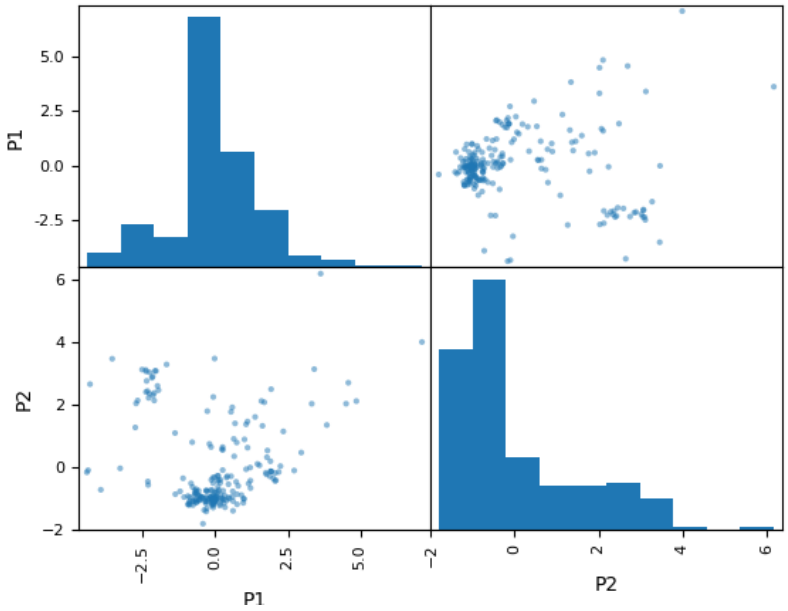

Figura 18: PCA sobre el diagrama de dispersión *glass.arff*

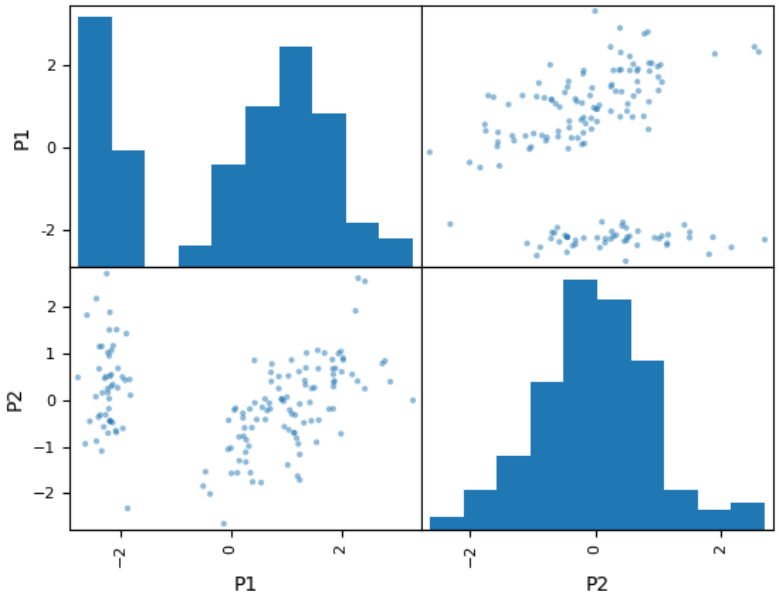

Figura 19: PCA sobre el diagrama de dispersión *iris.arff*

8 Ejercicio 8

Estudie el diagrama de correlaciones de los tres conjuntos e indique qué información relativa a las diferentes clases puede obtener.

El presente digrama representa el grado de relación entre las columnas del dataframe. Se grafica un *heatmap* en el que, cuanto más oscuro es el cuadrado de la matriz, menor es la relación entre atributos.

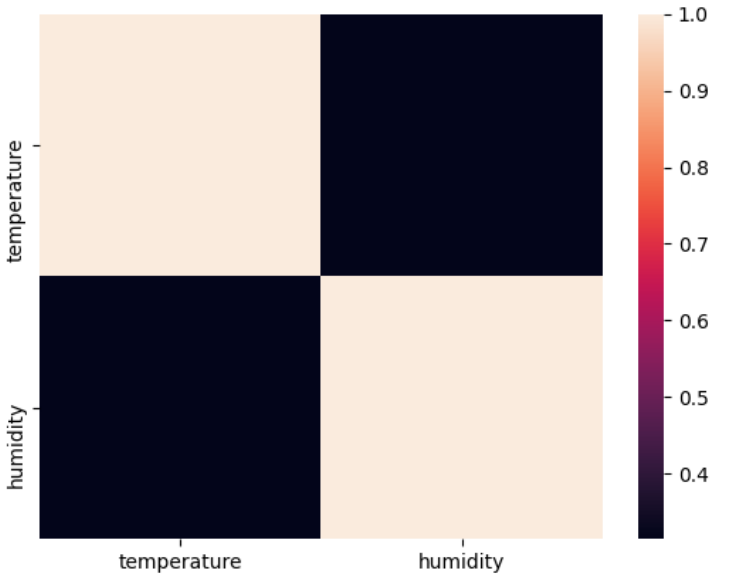

Figura 20: Diagrama de correlaciones *weather.arff*

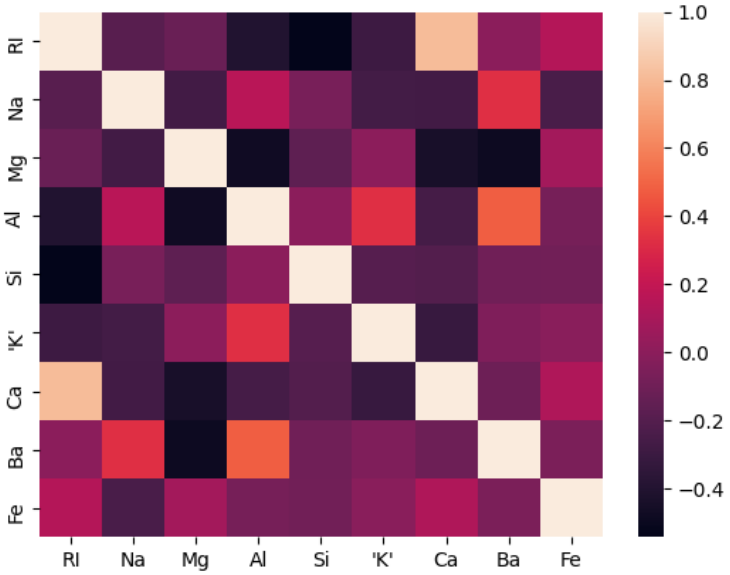

Figura 21: Diagrama de correlaciones *glass.arff*

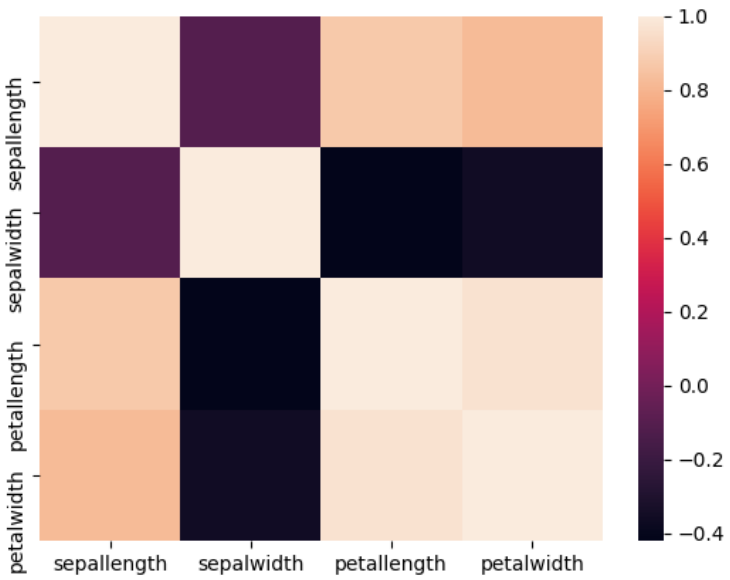

Figura 22: Diagrama de correlaciones *iris.arff*

9 Ejercicio 9

Estudie la representación en coordenadas paralelas de los tres conjuntos e indique qué información relativa a las diferentes clases puede obtener.

La representación en coordenadas paralelas de los conjutos de datos utilizados son las siguientes:

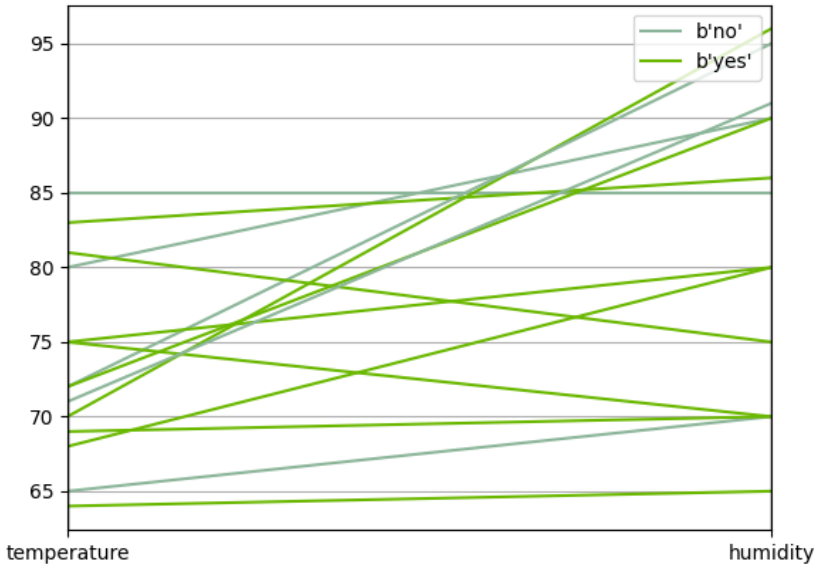

Figura 23: Representación en coordenadas paralelas *weather.arff*

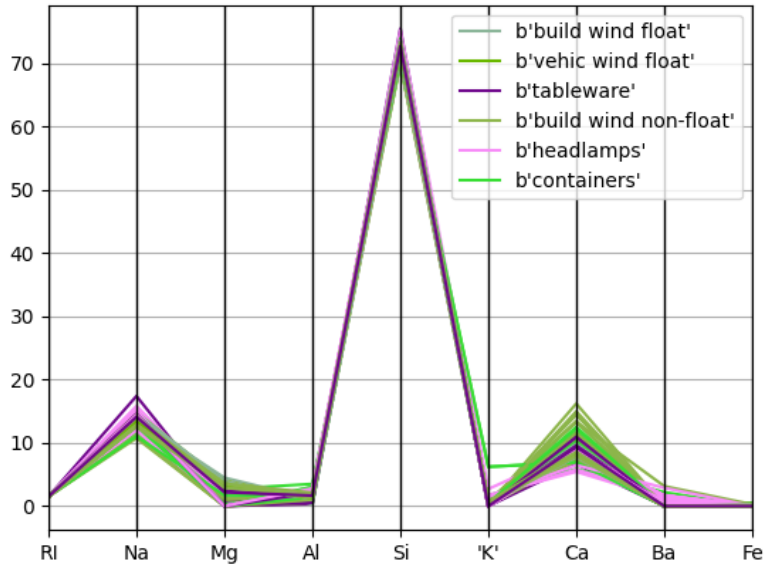

Figura 24: Representación en coordenadas paralelas *glass.arff*

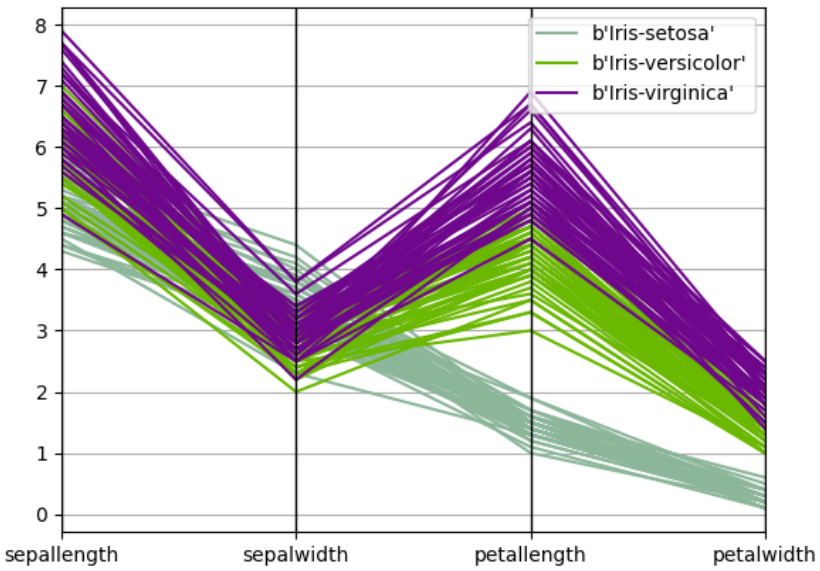

Figura 25: Representación en coordenadas paralelas *iris.arff*

Referencias

[1] Moodle Universidad de Córdoba - Enunciado práctica 2.

[2] Moodle Universidad de Córdoba - Introducción a Pandas.

[3] Moodle Universidad de Córdoba - Introducción a Matplotlib.