Práctica 2: Exploración de datos

Asignatura: Introducción a la Minería de Datos, 4º Grado de Ingeniería Informática Escuela Politécnica Superior de Córdoba - Universidad de Córdoba 2020 - 2021

Trabajo realizado por:

-Antonio Gómez Giménez (32730338G)

i72gogia@uco.es

Índice:

Ejercicio 1

Para el primer ejercicio se han escogido las siguientes bases de datos, las cuales tienen un formato arff:

-Iris. Para iris los atributos y las clases son las siguientes:
Attribute Information:
% 1. sepal length in cm
% 2. sepal width in cm
% 3. petal length in cm
% 4. petal width in cm
% 5. class:
% -- Iris Setosa
% -- Iris Versicolour
% -- Iris Virginica

-Glass. Para el caso de la base de datos Glass los atributos y clases son las siguientes:
Attribute Information:
% 1. Id number: 1 to 214
% 2. RI: refractive index
% 3. Na: Sodium (unit measurement: weight percent in corresponding oxide, as are attributes 4-10)
% 4. Mg: Magnesium
% 5. Al: Aluminum
% 6. Si: Silicon
% 7. K: Potassium
% 8. Ca: Calcium
% 9. Ba: Barium
% 10. Fe: Iron
% 11. Type of glass: (class attribute)
% -- 1 building_windows_float_processed
% -- 2 building_windows_non_float_processed
% -- 3 vehicle_windows_float_processed
% -- 4 vehicle_windows_non_float_processed (none in this database)
% -- 5 containers
% -- 6 tableware
% -- 7 headlamps

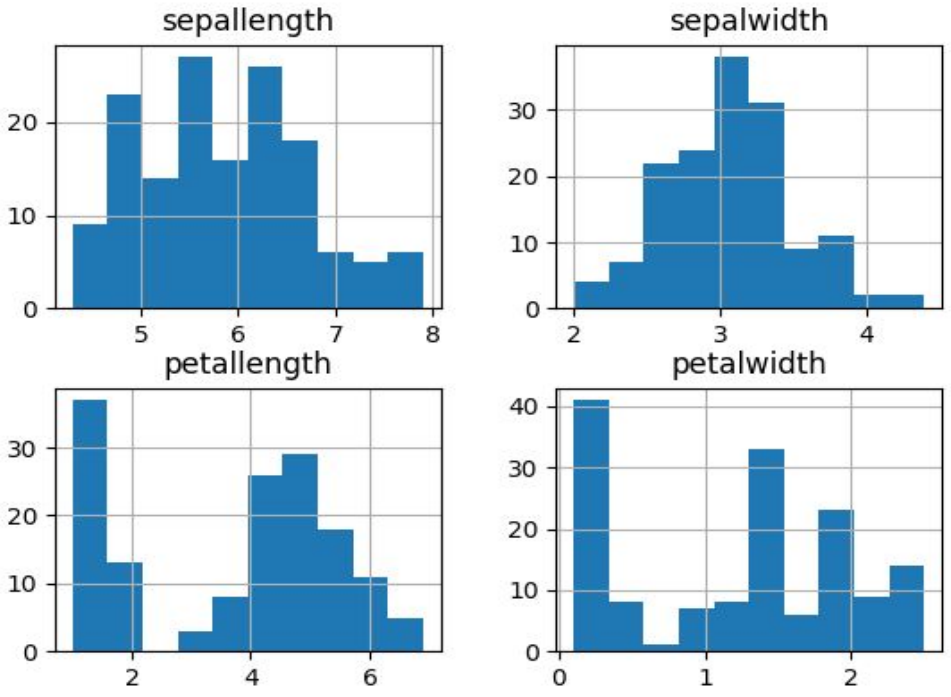
-Diabetes. Por último, para la clase Diabetes, los atributos y las clases son las siguientes:
Attribute Information:
% 7. For Each Attribute: (all numeric-valued)
% 1. Number of times pregnant
% 2. Plasma glucose concentration a 2 hours in an oral glucose tolerance test
% 3. Diastolic blood pressure (mm Hg)

% 4. Triceps skin fold thickness (mm)
% 5. 2-Hour serum insulin (mu U/ml)
% 6. Body mass index (weight in kg/(height in m)^2)
% 7. Diabetes pedigree function
% 8. Age (years)
% 9. Class variable (0 or 1)

Ejercicio 2

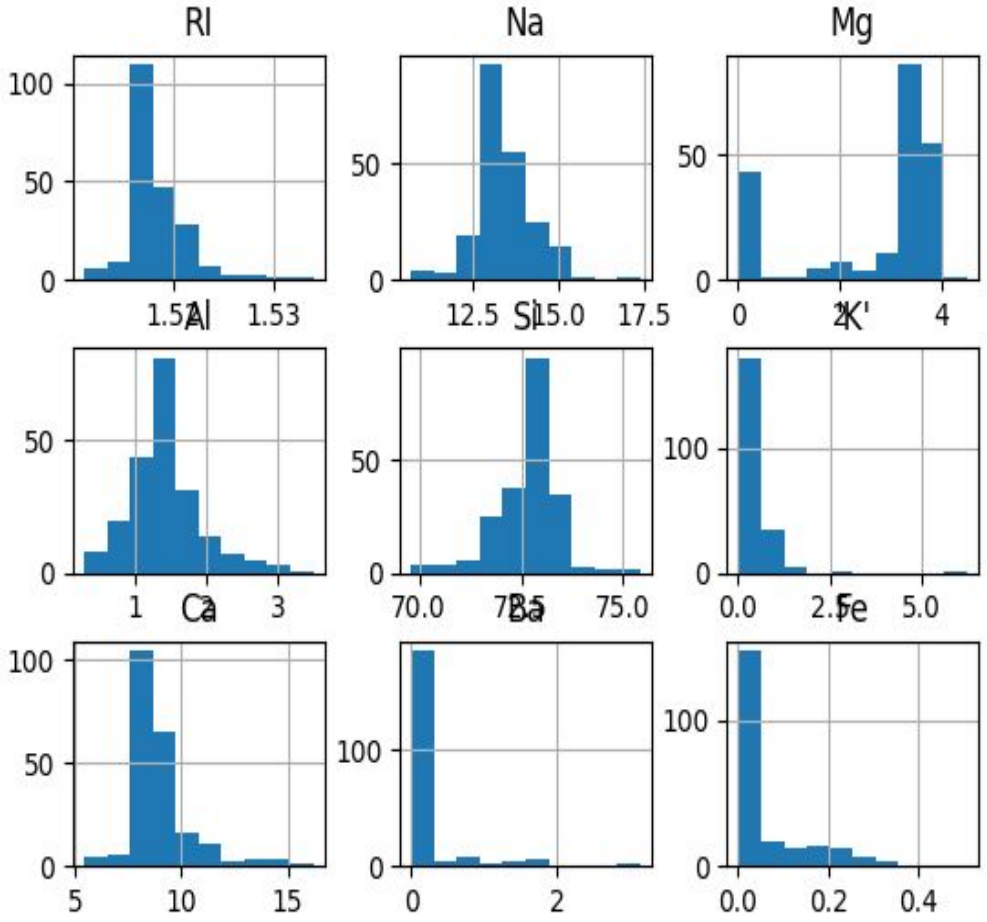
Para el ejercicio dos hemos obtenido los histogramas de todos los atributos para cada base de datos:

Iris:



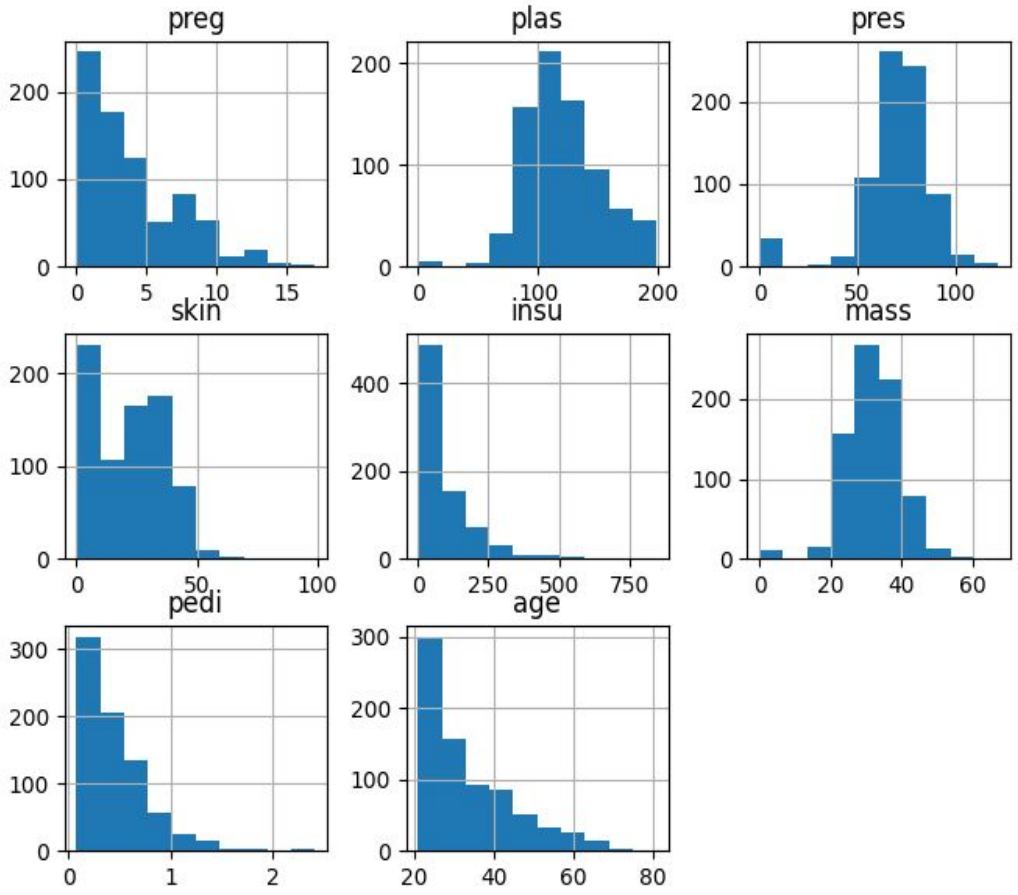
Estos histogramas nos dan una relación entre valores y cantidad, por ejemplo, para el ancho de pétalo, podemos observar que para el valor de dos, es decir, un ancho de pétalo de 2 cm podemos encontrar alrededor de 20 patrones. De esta forma podemos hacernos una breve idea de la cantidad de patrones respecto a qué valores tenemos. De esta forma, podremos intentar que la base de datos sea más homogénea añadiendo si es posible más patrones a aquellos que estén más descompensados o reduciendo el número de los mismo para el caso contrario.

Glass:



Para esta base de datos podemos ver la relación entre la cantidad de cierto material y la aparición de esa cantidad en los atributos, como podemos observar, tanto el hierro como el barium, tienen muchos patrones con valores bajos y muy pocos con valores altos, esto puede llegar a ser un problema a la hora de entrenar por ejemplo un clasificador, por ello será necesario estandarizar y normalizar.

Diabetes:

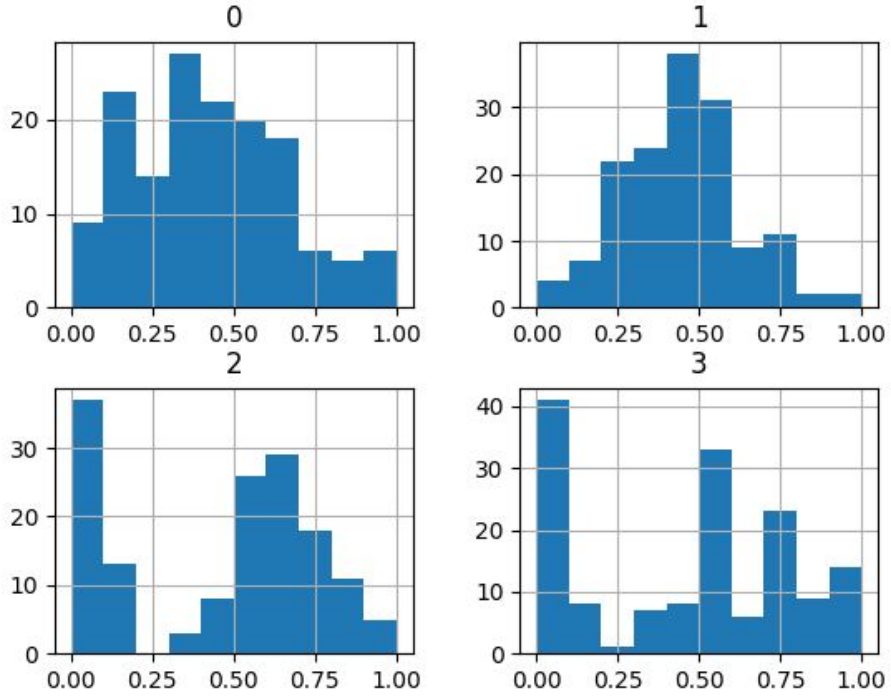


En la base de datos de diabetes, realizamos los histogramas de cada atributo como en las bases de datos anteriores, para cada atributo conseguimos la relación de la cantidad de ese atributo, por ejemplo en el atributo años, sería la edad, donde tenemos muchos patrones de 20 años y pocos patrones de 80. Es decir, con los histogramas, obtenemos la cantidad de veces que aparece cierto valor de un atributo en nuestros patrones de la base datos.

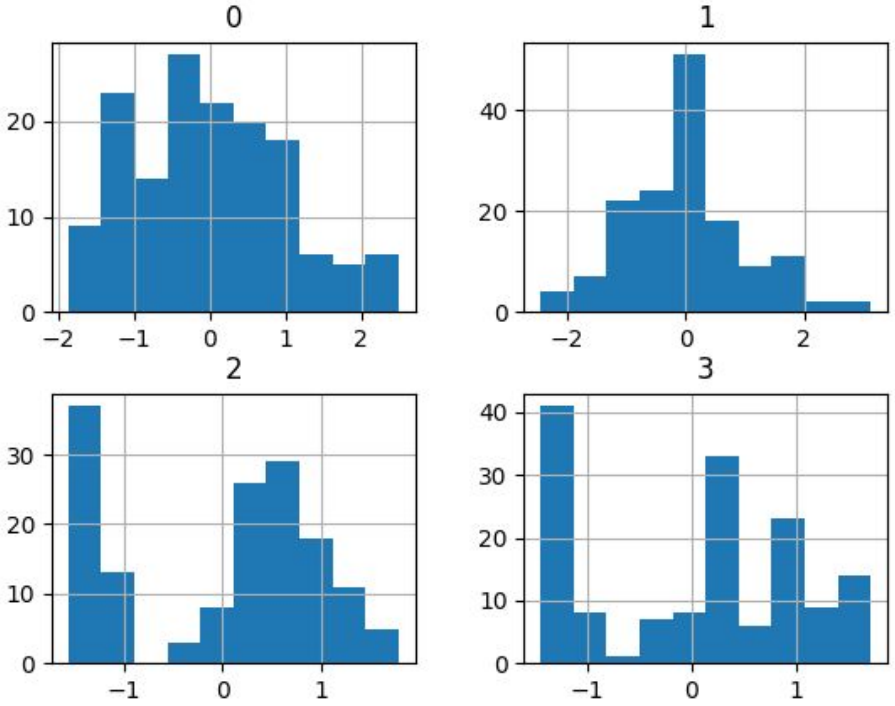
Ejercicio 3

Para este ejercicio, se ha realizado la estandarización y la normalización en las tres bases de datos, los resultados obtenidos son los siguientes, los valores numéricos encima de los histogramas, son el nombre de los atributos que hemos visto anteriormente:

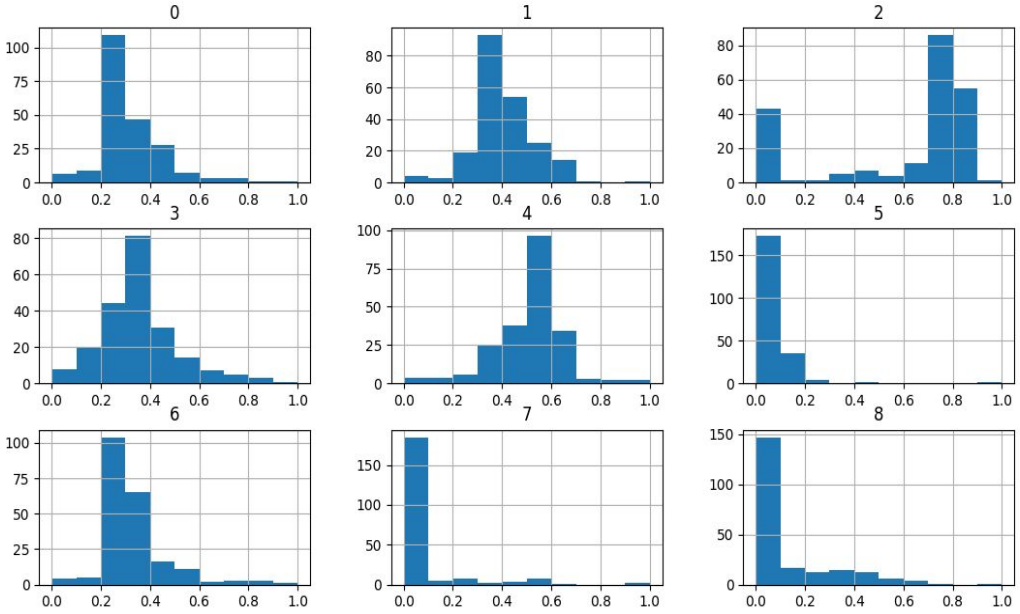
Iris (normalizado):



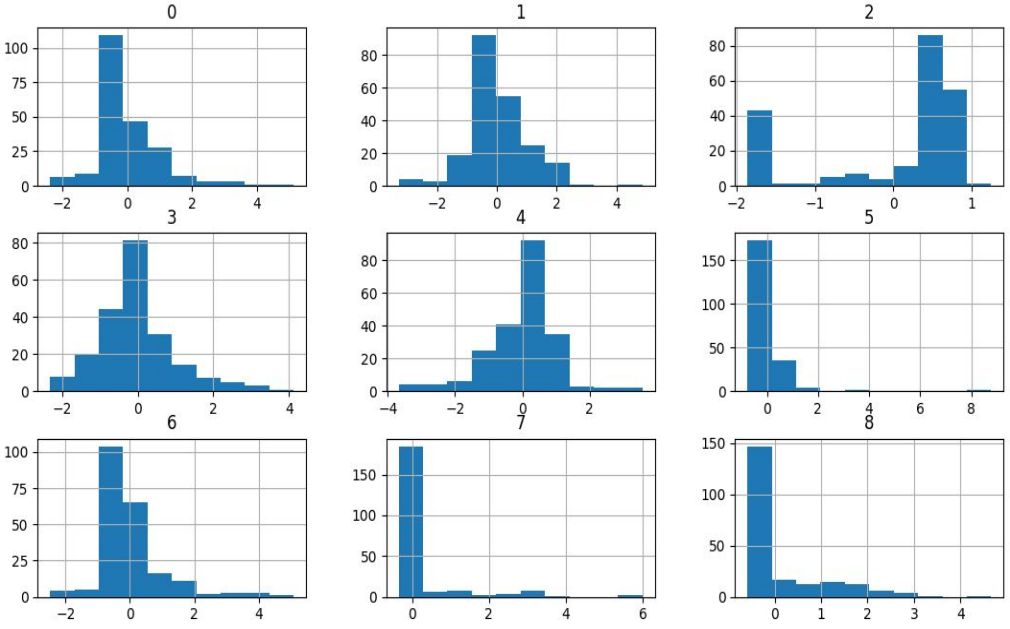
Iris (Estandarizado):



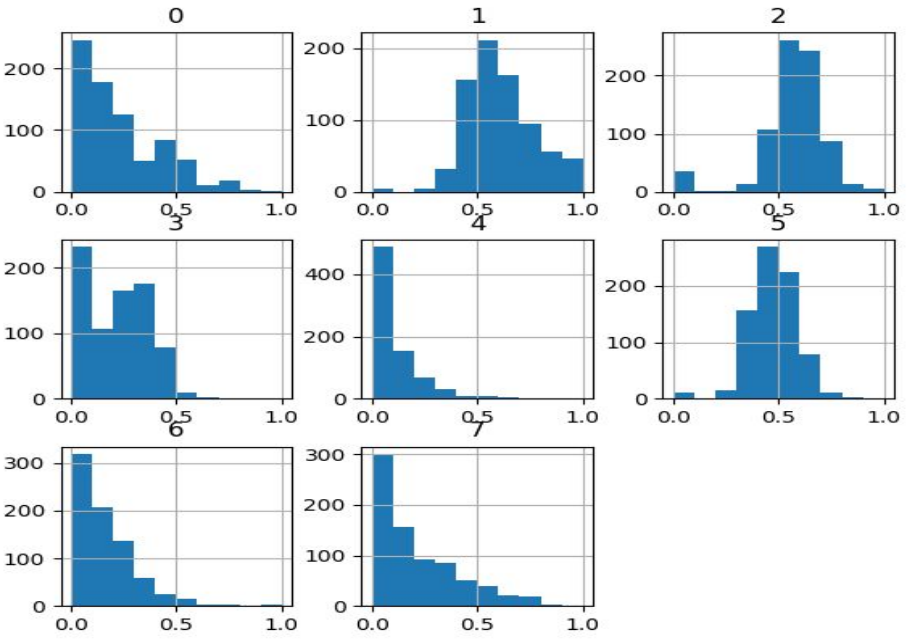
Glass (normalizado):



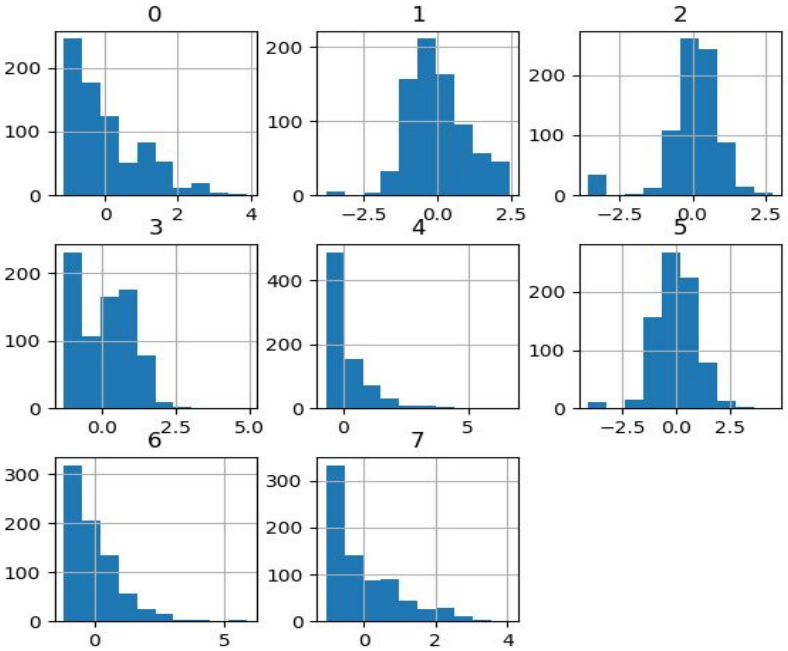
Glass (Estandarizado):



Diabetes (normalizado):



Diabetes (Estandarizado):

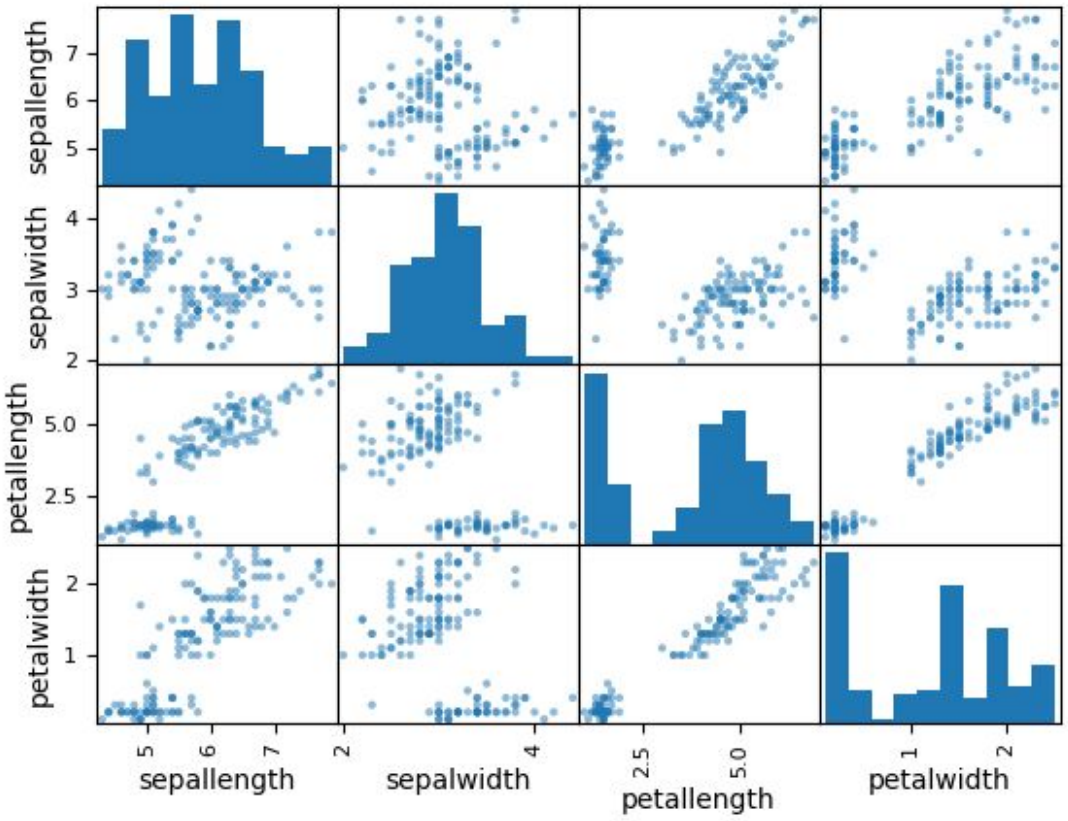


Para normalizar, normalizamos entre 0 y 1 todos los valores y para estandarizar, se resta la media y se divide entre la desviación típica.
Con lo realizado anteriormente conseguimos a la hora de estandarizar, dar la misma importancia a todos los datos (le afecta bastante los outliers) y respecto a la normalización, para evitar que aquellos valores que aparezcan demasiadas veces puedan opacar a los valores que aparecen menos, se normaliza entre 1(valores más altos) y 0 (valores más bajos), de esta forma, al realizar operaciones entre valores, los resultados serán más coherentes (sobre todo si tenemos algún outlier).

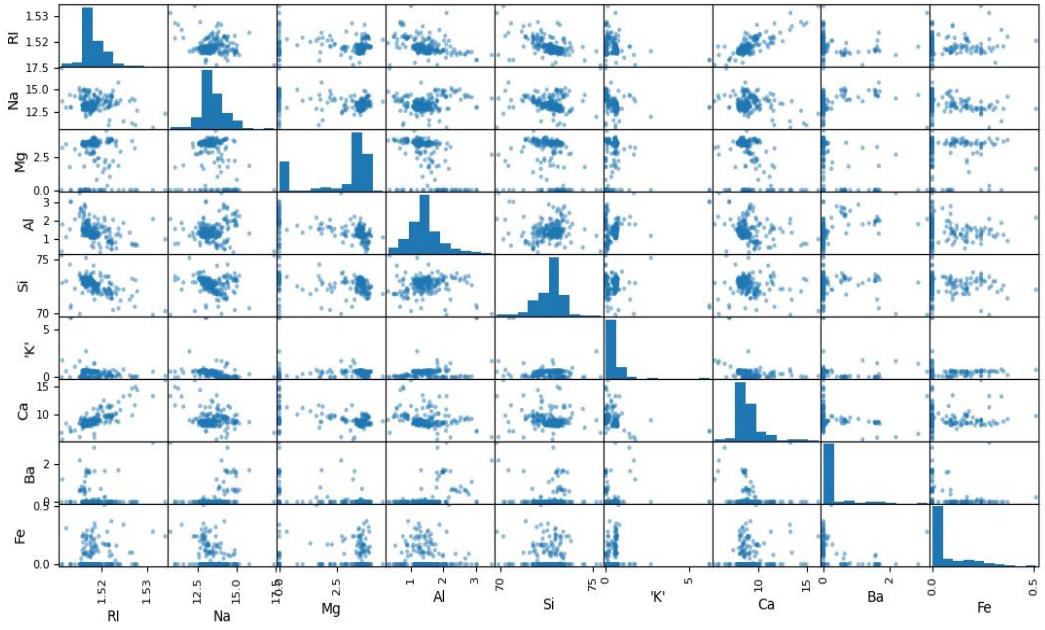
Ejercicio 5

Para este ejercicio, se han obtenido los diagramas de dispersión para cada base de datos:

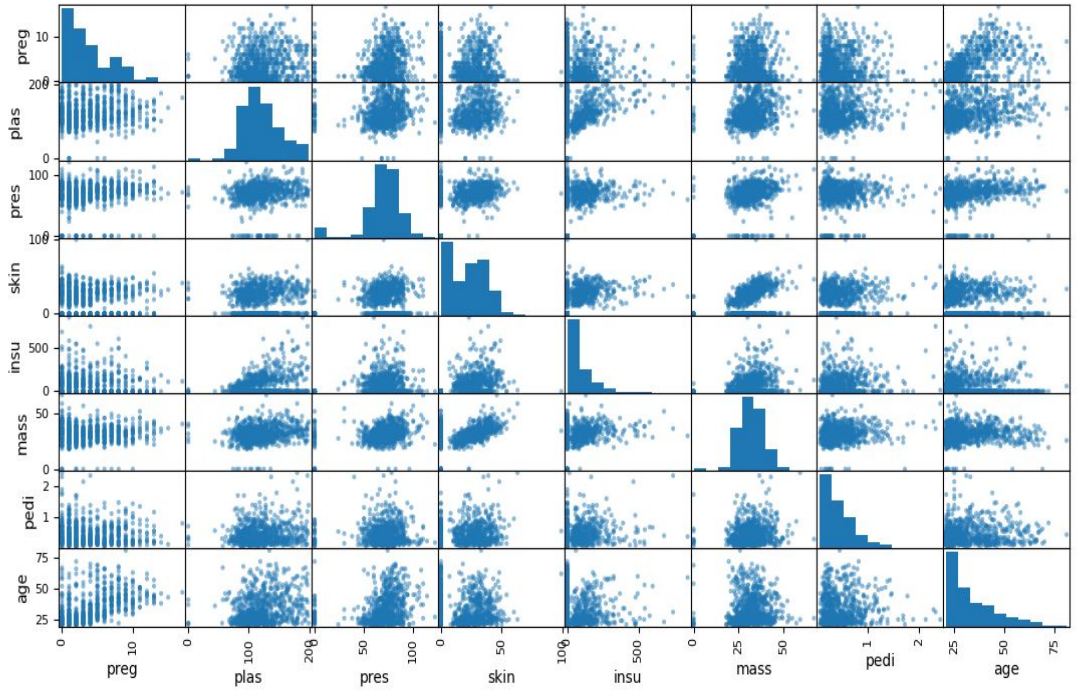
Iris:



Glass:



Diabetes:



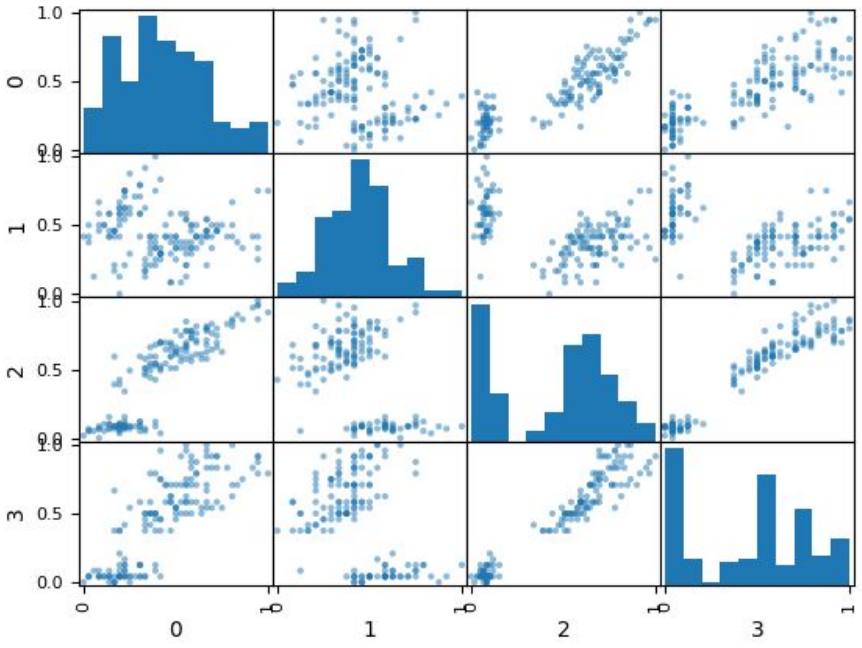
Las bases de datos Glass y Diabetes, es complicado apreciarlas bien ya que tienen un mayor número de datos.
Respecto a la clase Iris podemos observar las relaciones entre todos los atributos, siendo la diagonal, los histogramas, anteriormente vistos. Cada punto representa un patrón, el cual se sitúa en dicho espacio respecto a sus dos atributos. Ese patrón, dependiendo del color, pertenece a una clase, en este caso, dependiendo del tipo de azul pertenece a Iris Setosa, Iris Versicolour o Iris Virginica. Por ejemplo, en el caso del atributo petal length con petal width, podemos observar que separa bastante bien las tres clases, ocurre algo similar con sepal width y petal width, aunque en este caso sí que hay ciertos patrones muy juntos entre las clases versicolour y virginica.

Tanto en Glass como en diabetes nos encontramos con una nube de puntos al situar los patrones con dos atributos, donde todas las clases se encuentran juntas, por ello mismo es muy difícil ver qué atributos son los que mejor separan las clases, probablemente sea necesario el uso de varios atributos para poder separar las clases ya que añadiendo más dimensionalidad al problema (añadir más atributos), será más probable que las clases sean más linealmente separables.

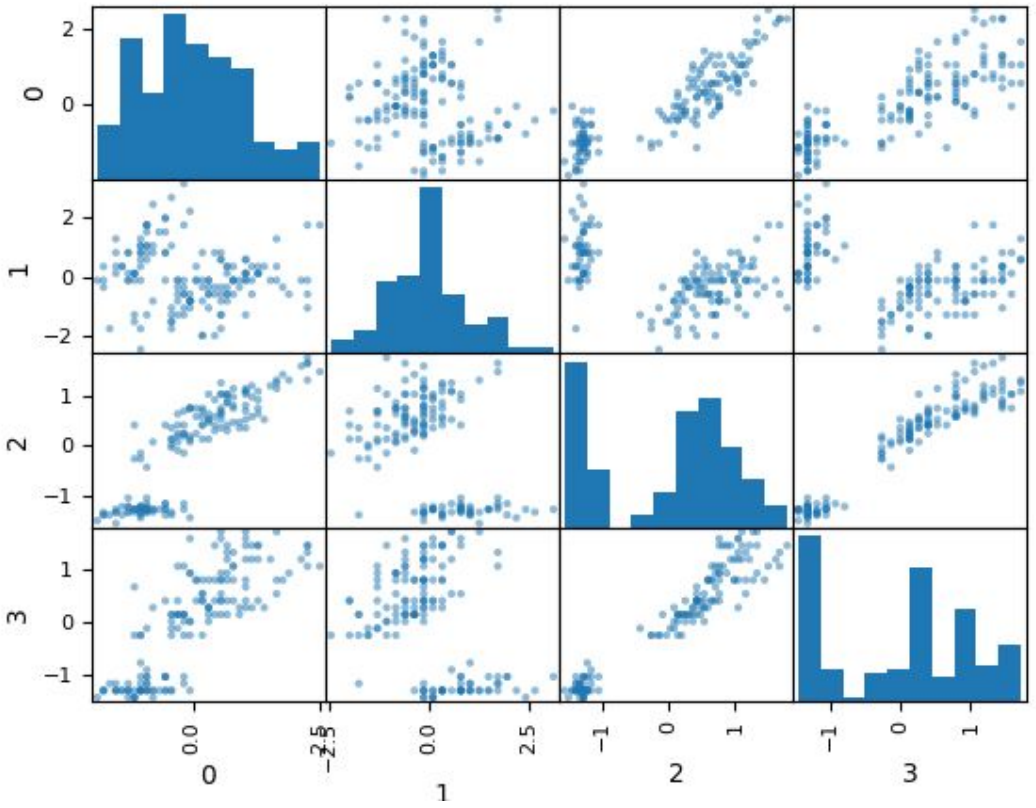
Ejercicio 6

Este ejercicio es similar al anterior pero se realiza con las bases de datos normalizadas y estandarizadas.

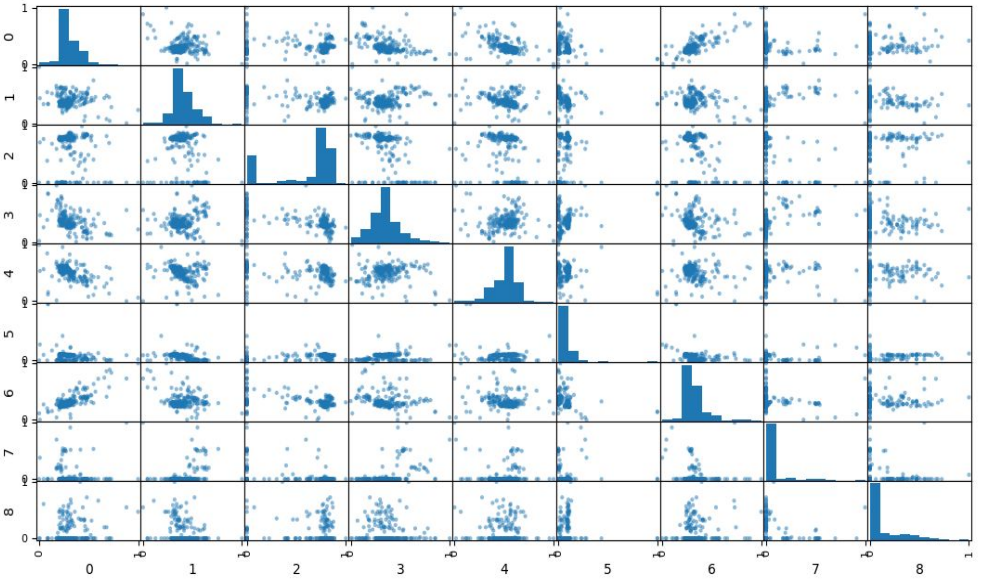
Iris (normalizado):



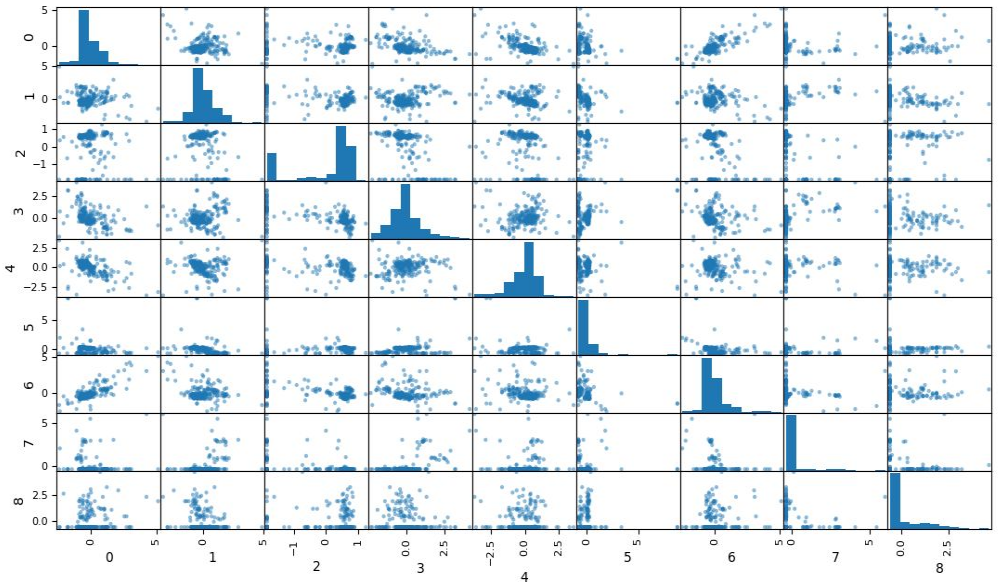
Iris (Estandarizado):



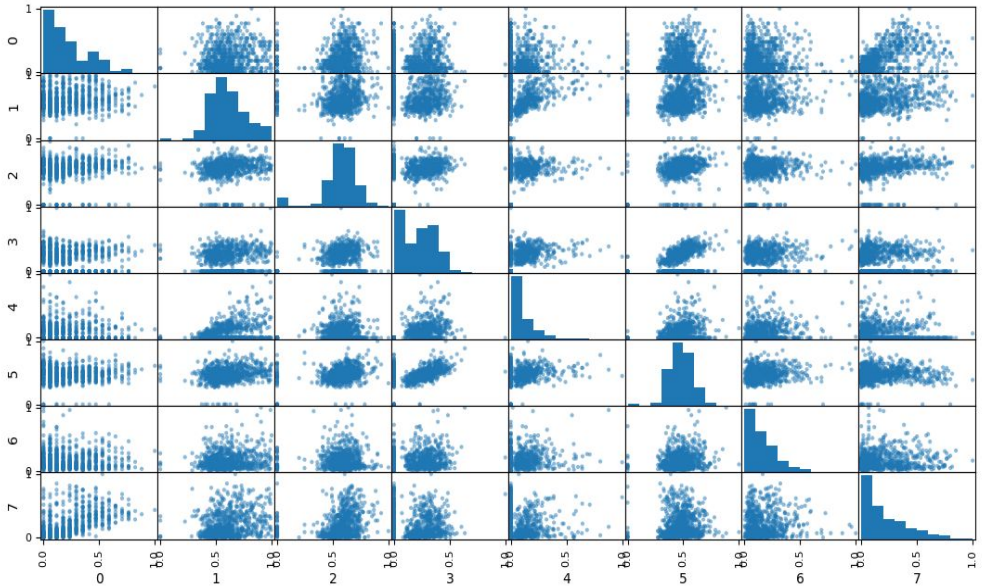
Glass (normalizado):



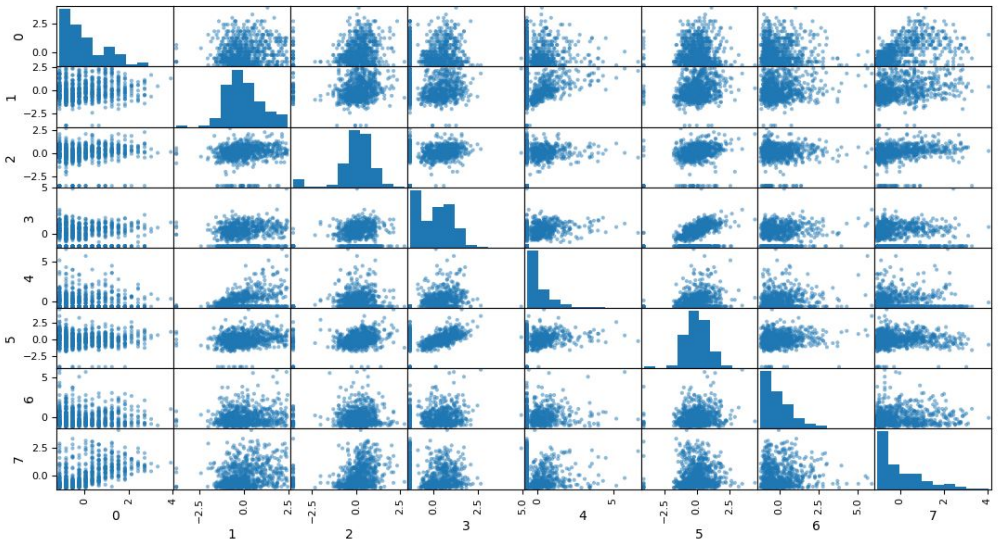
Glass (Estandarizado):



Diabetes (normalizado):



Diabetes (Estandarizado):



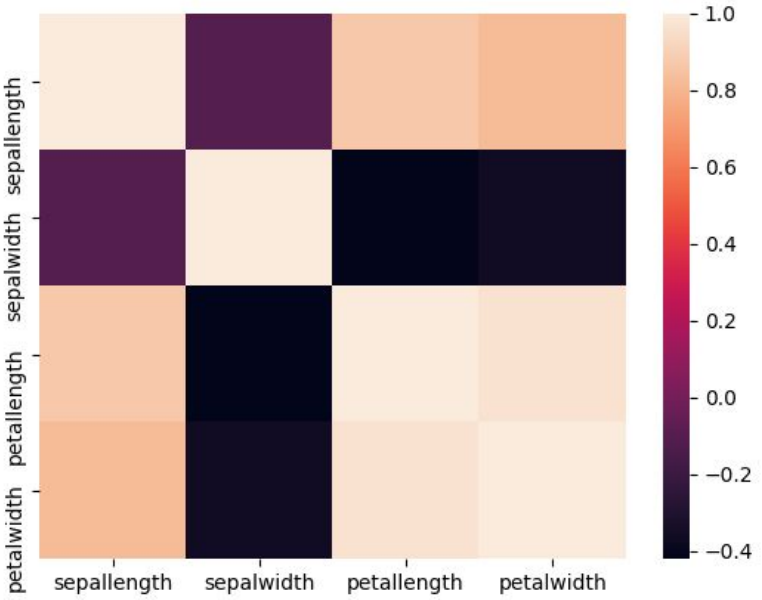
Respecto a la clase Iris no podemos observar ningún cambio significativo, las relaciones son como en el caso anterior, donde la base no se encuentra estandarizada y normalizada. El único cambio que podemos observar es en los valores tomados por los patrones que en el caso de normalizar va entre 0 y 1 y en estandarizar va variando dependiendo de la media entre la desviación típica, aun así la posición del patrón es la misma.

Respecto a Glass y Diabetes ocurre algo similar a la clase Iris, no se pueden percibir cambios notables a parte de los mencionados con la clase iris.
El único cambio se puede observar en los valores de los diagramas que varían dependiendo si es estandarización o normalización.

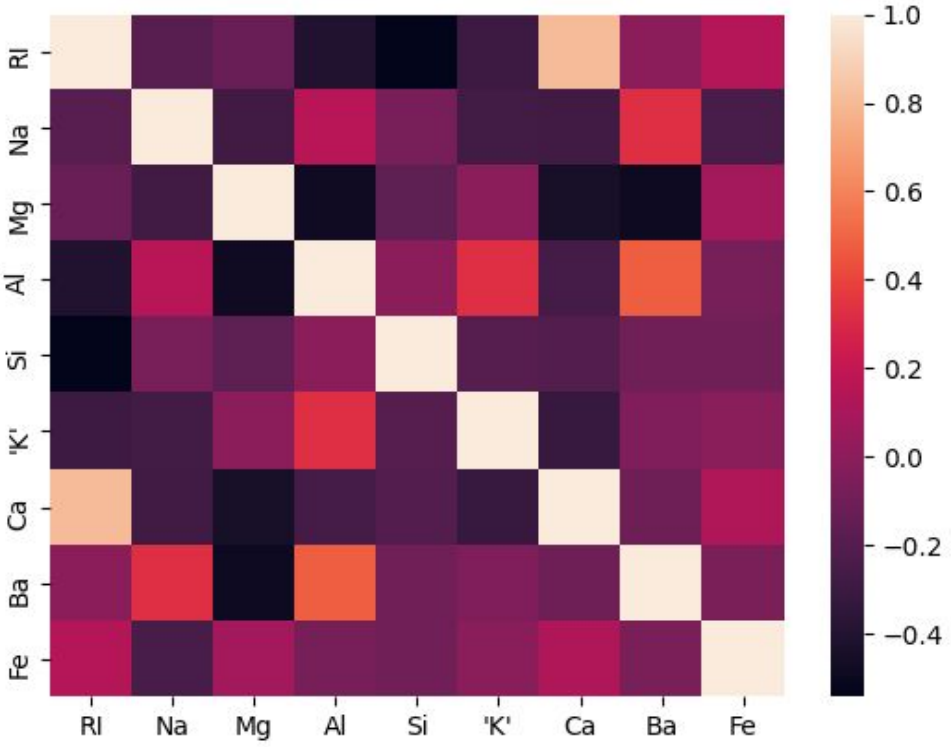
Ejercicio 8

En este ejercicio obtenemos el diagrama de correlaciones de las tres bases de datos, iris, glass y diabetes.

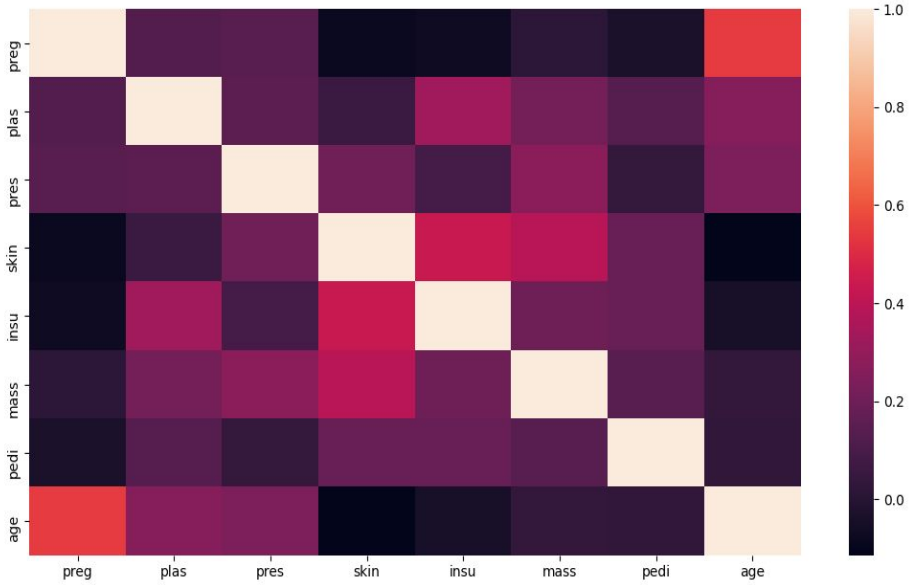
Iris:


sepallength
sepalwidth
petallength
petalwidth
sepallength
sepalwidth
petallength
petalwidth
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4

Glass:


RI
Na
Mg
Al
Si
'K'
Ca
Ba
Fe
RI
Na
Mg
Al
Si
'K'
Ca
Ba
Fe
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4

Diabetes:



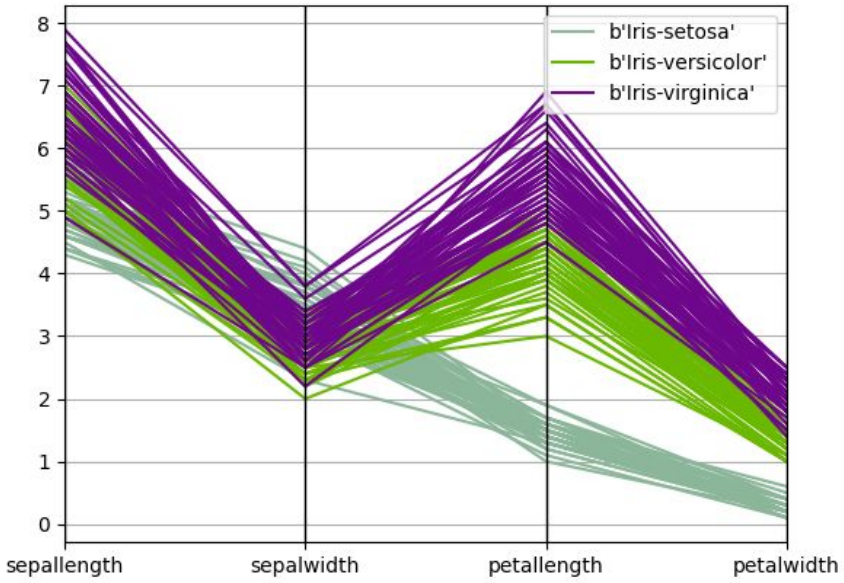
En general, las conclusiones obtenidas son similares al scatter plot. La diagonal de la matriz en las tres bases de datos es blanca, esto quiere decir que cada atributo es muy correlado consigo mismo, esto es lógico ya los vectores son los mismo por ello la correlación es 1.

Respecto a la base de datos iris, como vimos con el scatter plot, petal length con petal width, decíamos que separa bastante bien las tres clases, ocurre algo similar con sepal width y petal width. Si comprobamos esto con la matriz de correlaciones lo podemos confirmar, ya que entre estos atributos, podemos ver que tienen una correlación cercana a 1.
También podemos observar que el atributo sepal width, sea cual sea el otro atributo (excepto el mismo), tiene poca correlación de tal forma que no permite separar bien las clases, podríamos considerar este atributo como poco útil a la hora de separar las clases entre sí.
Respecto a las bases de datos Glass y Diabetes confirmamos que hay pocas correlaciones, por ello en el scatter plot era tan difícil separar las clases con dos atributos ya que hay poca correlación entre los pares de atributos y se forman una única nube de puntos.
Cabe mencionar, que en la base de datos Glass, podemos observar que existe una correlación que no habíamos apreciado en el scatter plot y es que el atributo de calcio e Índice de reflectancia tienen una correlación, esto era muy difícil verlo en el scatter plot por que hay varias clases. Por tanto, esta combinación de atributos permite separar las clases de la base de datos Glass.

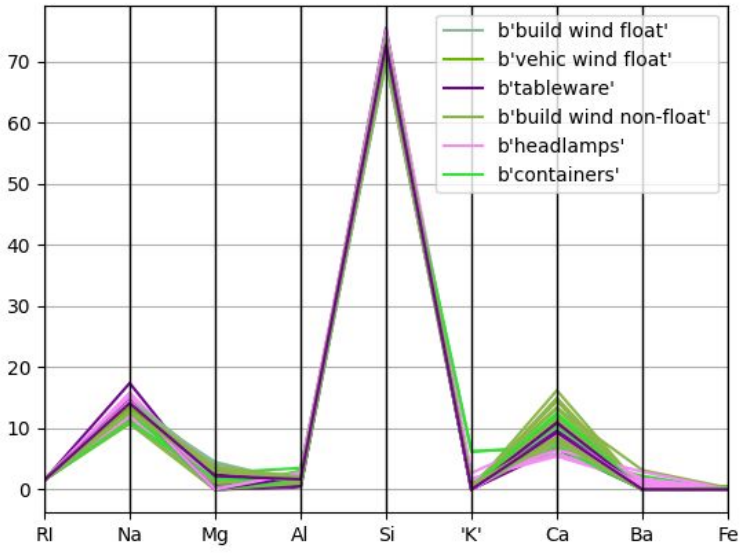
Ejercicio 9

Por último, en este ejercicio obtenemos las coordenadas paralelas de las tres bases de datos, iris, glass y diabetes.

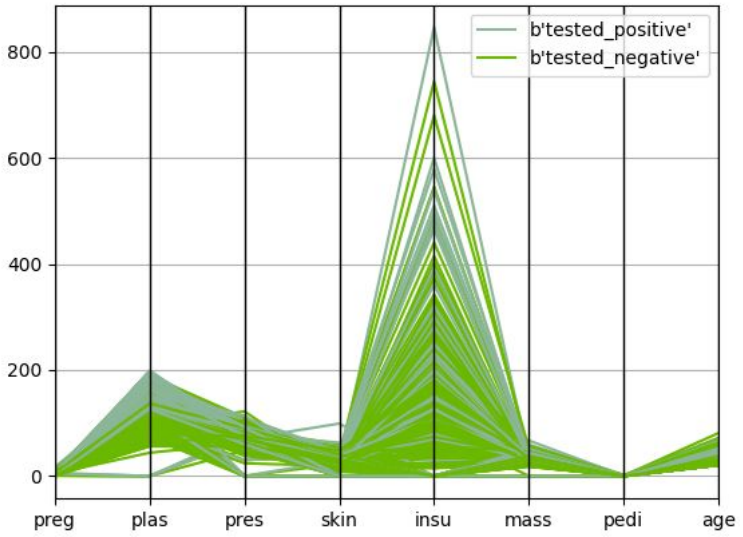
Iris:



Glass:



Diabetes:



Si observamos la base de datos iris podemos observar que sepalwidth como variable única no es capaz de separar la clases, ya que con este tipo de gráfico mostramos para cada atributo respecto a la columna de cm para todos los patrones de cada clase y en este caso no hay diferencia entre cm de los tres tipos de clase, sin embargo, si nos fijamos en el atributo petal length podemos observar que este atributo es capaz de separar las clases por tanto podemos afirmar que va a ser un problema sencillo de resolver.

Respecto a las bases de datos Glass y Diabetes a todos los atributos le ocurre algo similar a lo que hemos visto en el caso anterior con el atributo sepalwidth de la clase iris. No es posible separar correctamente las clases.
No por esto, quiere decir que el problema sea imposible de resolver, ya que puede ser que una combinación de atributos permita separar bien las clases, como hemos visto en el caso de los atributos de calcio e Índice de reflectancia los cuales tienen una correlación, permitiendo separar las clases entre sí.