Tarea del 5 de Mayo 2020 Tanagra. Clustering Jerárquico.

Parte 1

1. **Aplique el algoritmo HAC 1 con las variables de input: Activo Corriente/Pasivo Corriente, Activo Corriente/ Ventas Netas.**

    a. **Observe el dendrograma. ¿cuántos cluster conviene generar? _____**

    b. **¿Cuántos cluster se presentan? ____ ¿Cuántos elementos tiene cada cluster? _____**

    c. **En View Dataset, ¿qué representa la última columna? _____**

    d. **Genere un diagrama de dispersión para ambas variables, e indique en el mismo los clusters definidos.**

2. **Aplique el algoritmo K-means con las opciones por defecto. ¿Cuántos clusters se generaron?__ ¿Cuál es la cantidad máxima de iteraciones?____ ¿Se presenta el dendrograma?_____ ¿Cuál es el motivo?**

3. **Aplique el algoritmo K-means para generar 4 clusters. ¿Cuántos elementos tiene cada cluster?_____ __**

4. **Genere un diagrama de dispersión para representar indistintamente los clusters generados por con HAC1 o con K-means. ¿Algún algoritmo de clustering le parece preferible? Fundamente.**

Parte 2

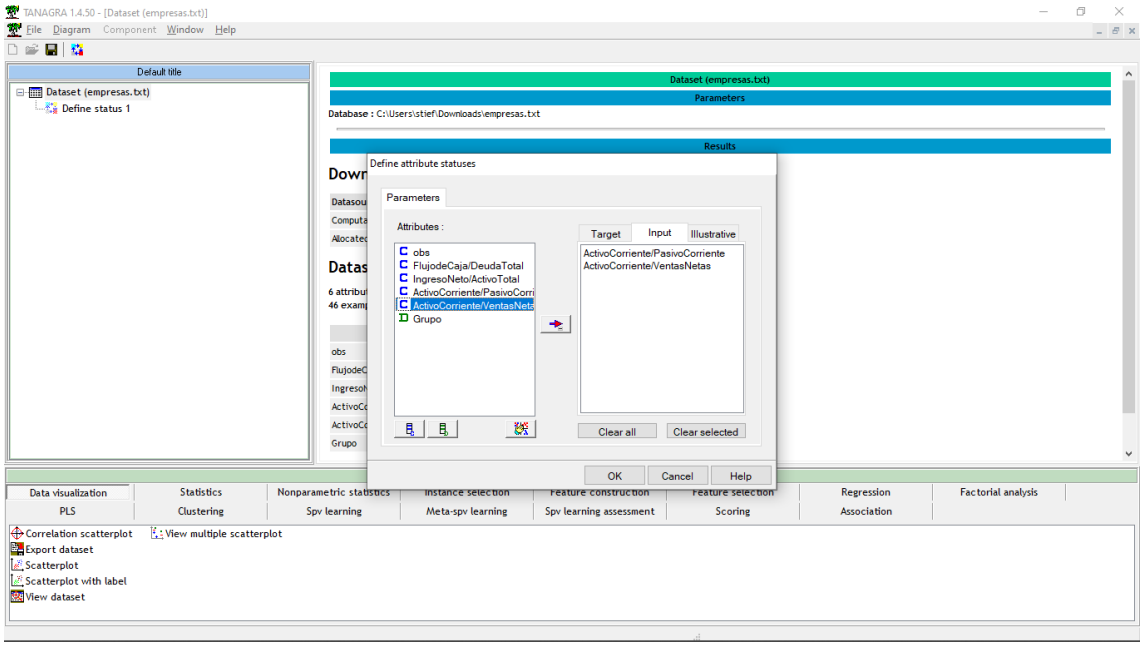
5. **Repita lo realizado en la Parte 1 trabajando con las 4 variables cuantitativas como input.**

1.

Para comenzar abrimos Tanagra. En “File → New File” buscamos el archivo de trabajo empresas.txt y lo cargamos. Pulsamos “Ok”.

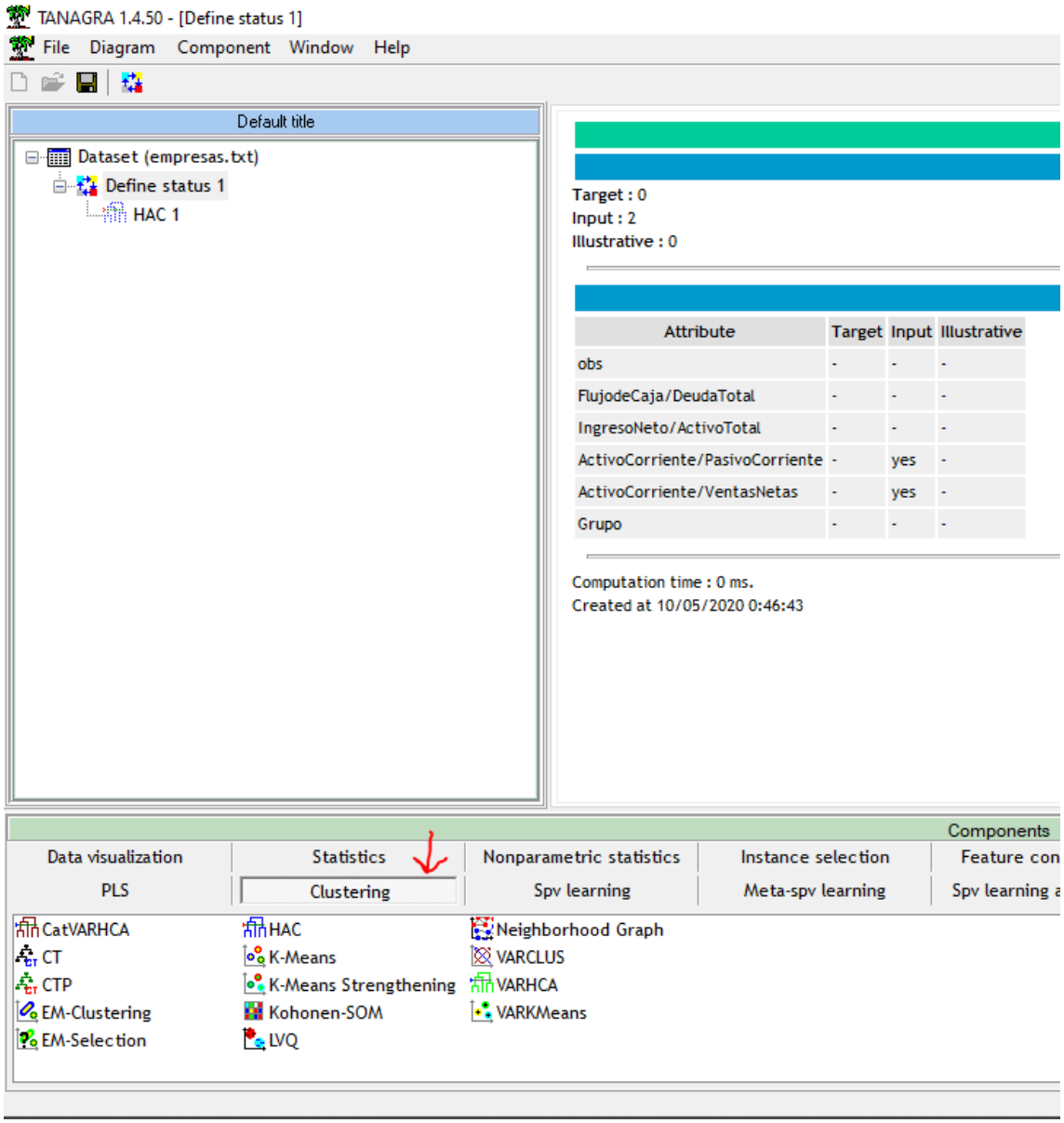
Vamos a aplicar el algoritmo de clustering jerarquico aglomerativo . Como entradas vamos a tomar ActivoCorriente/PasivoCorriente y ActivoCorriente/VentasNetas.

Por lo tanto procedemos a crear un “Define Statues” con dichas entradas:

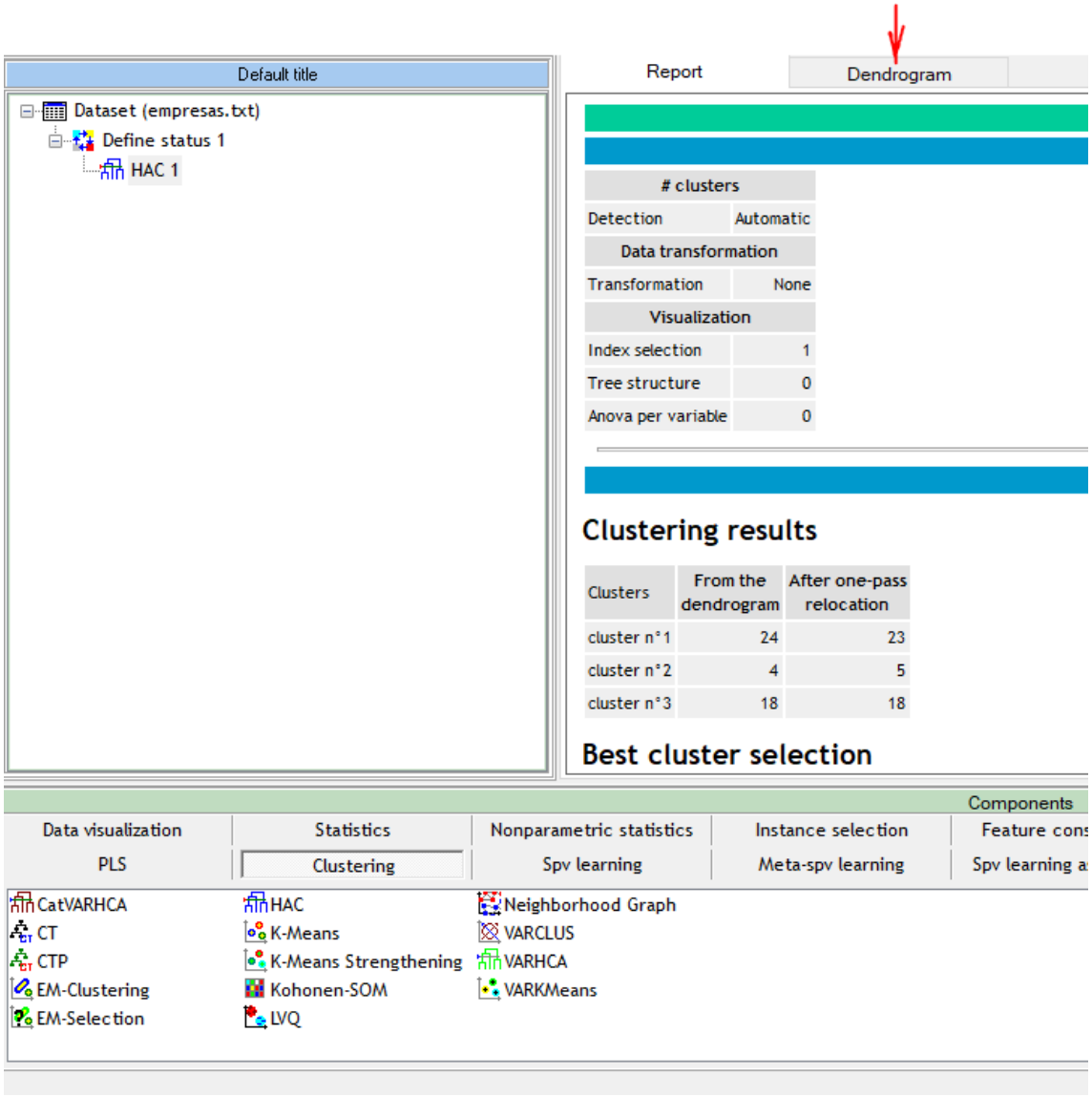


Como el clustering es una técnica de aprendizaje no supervisada, no hay atributo objetivo (Target).

Ahora vamos a ir a la pestaña “Clustering” y vamos a buscar el algoritmo HAC, el cual seleccionaremos y arrastraremos hasta nuestro “Define Statues”.

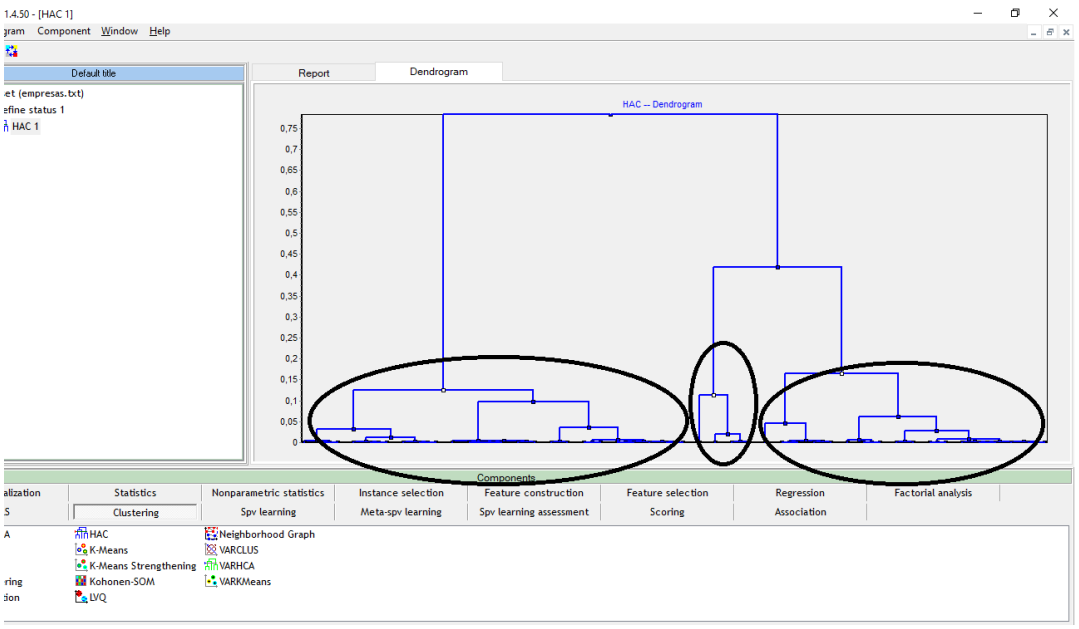


Ejecutamos el algoritmo y observamos lo siguiente:



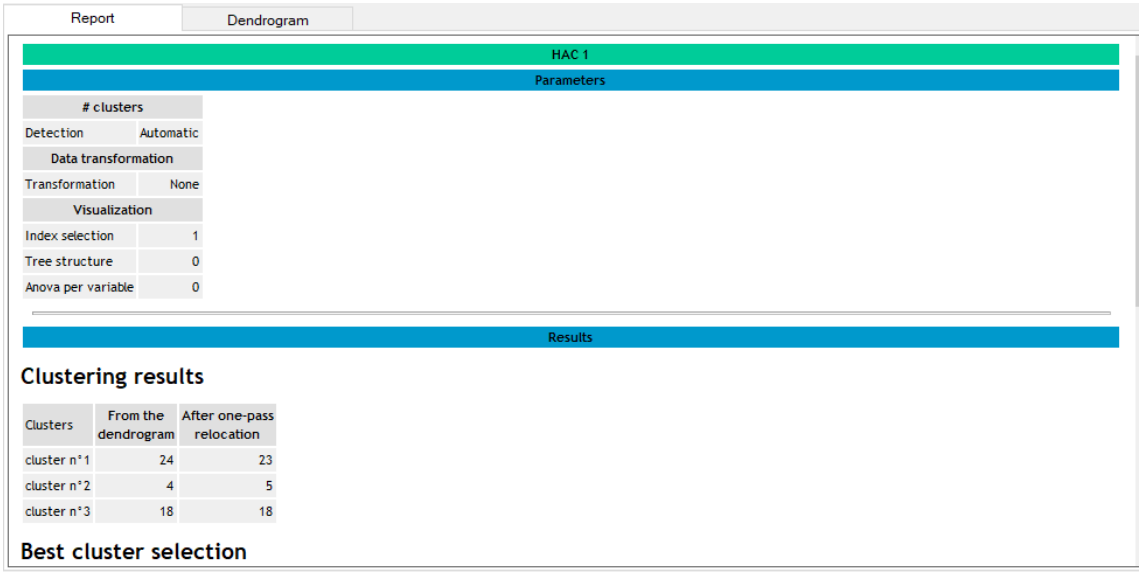
Si pulsamos en “Dendogram” observamos una acumulación de puntos en la parte inferior de la gráfica. Aquí se observa gráficamente que datos tienen mayor parecido entre ellos. El eje de la ordenada nos muestra la distancia entre datos. Por tanto, cuanto más cercano estén los puntos al eje de las abcisas, más parecidos serán entre ellos los datos.

Viendo dónde se concentran más los datos, podemos agrupar los mismo en tres “clusters”:



Podemos agruparlos en 3 “clusters” porque entre el “cluster” de la izquierda y el del centro, la distancia sería demasiado grande para considerarlo un solo “cluster” en vez de dos.

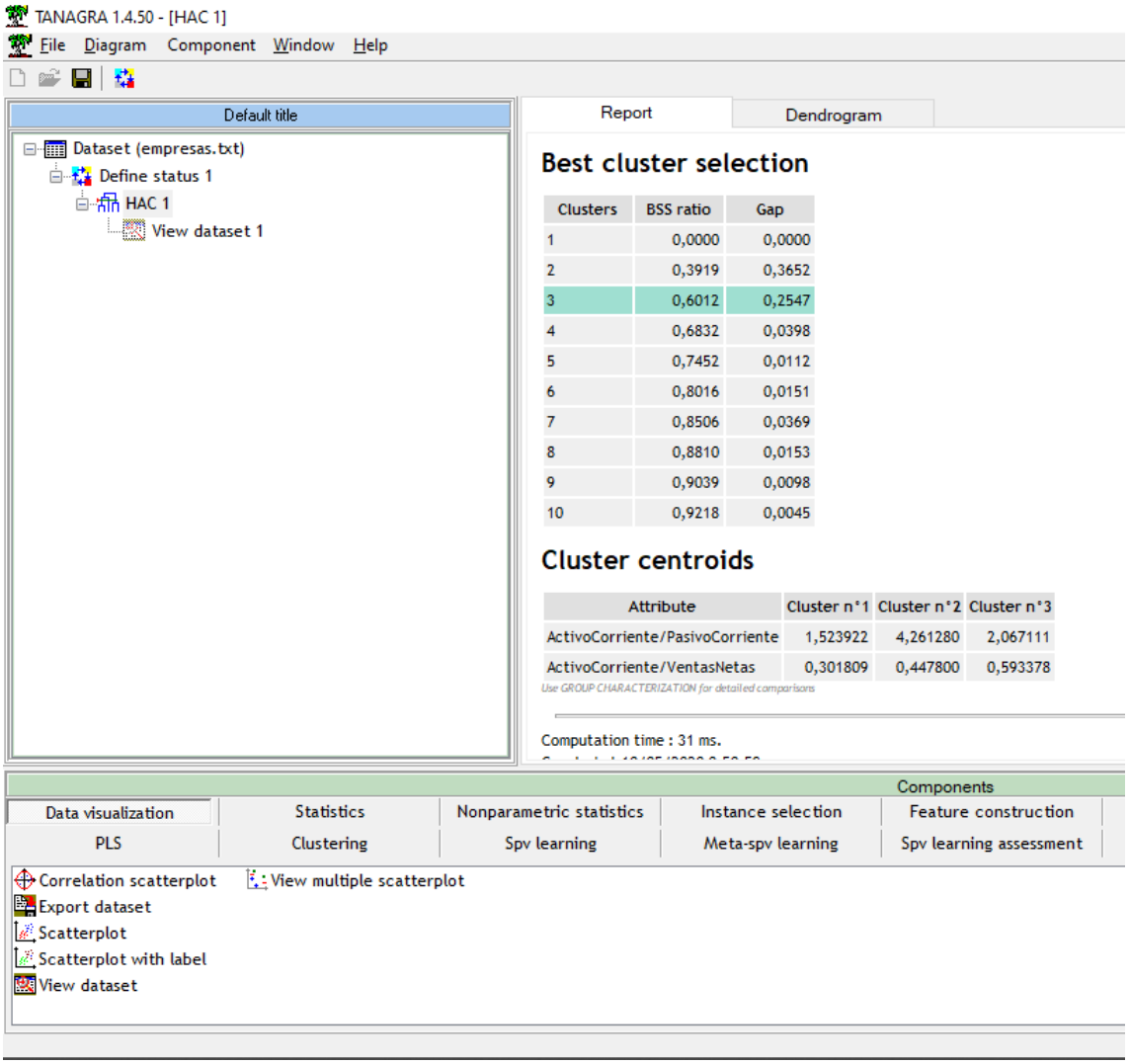
Para comprobar analíticamente el número de “clusters” reales, nos vamos a la pestaña “Report”, que está justo a la izquierda de la pestaña “Dendrogram”. Aquí nos vienen todos los datos generados por el algoritmo HAC:



Aquí comprobamos que realmente se forman tres “clusters”. El algoritmo HAC detecta el mismo el número de “clusters” que se pueden formar.

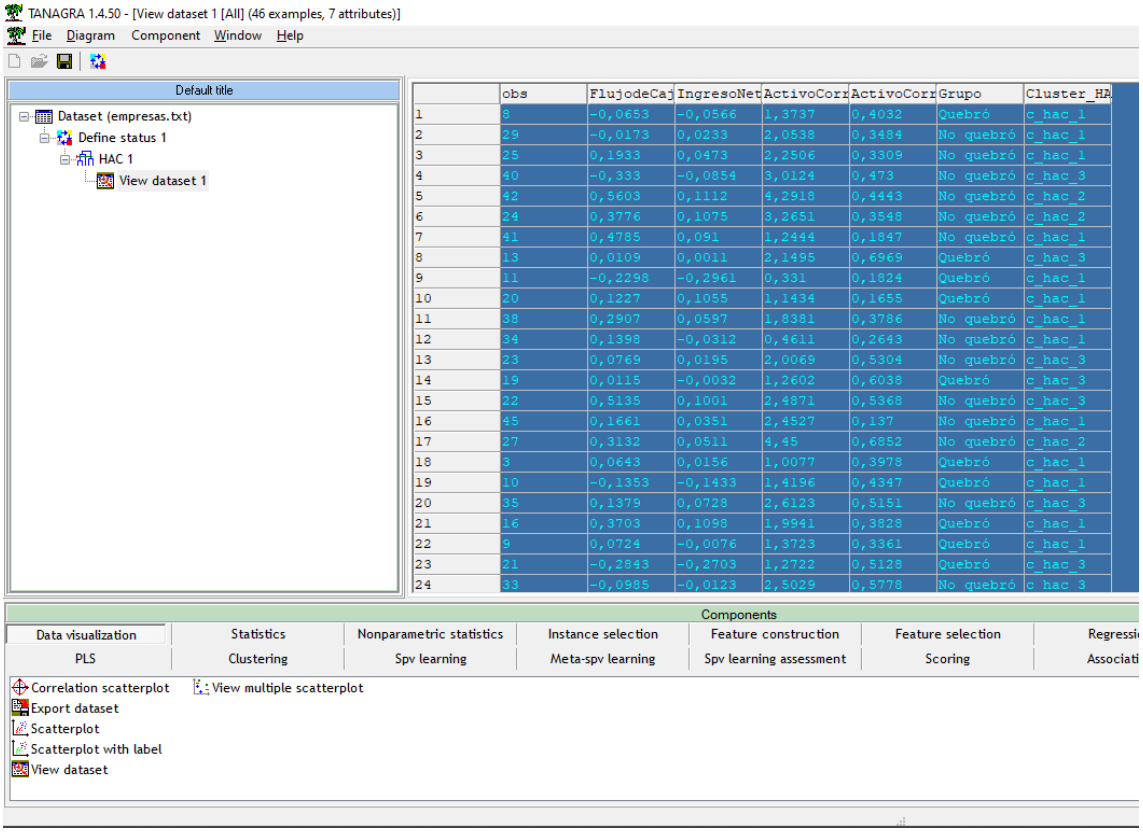
El “cluster” 1 tiene 24 elementos, el “cluster” 2 tiene 4 elementos y el “cluster” 3 tiene 18 elementos.

En esta misma pantalla tenemos datos de interés como el centro de los “clusters”, que son datos importantes ya que nos muestran el dato más representativo.



Para ver que se sitúa en la última columna nos vamos a la pestaña que dice “Data visualization” y seleccionamos “View Dataset”. Lo arrastramos hasta debajo de “HAC 1”.

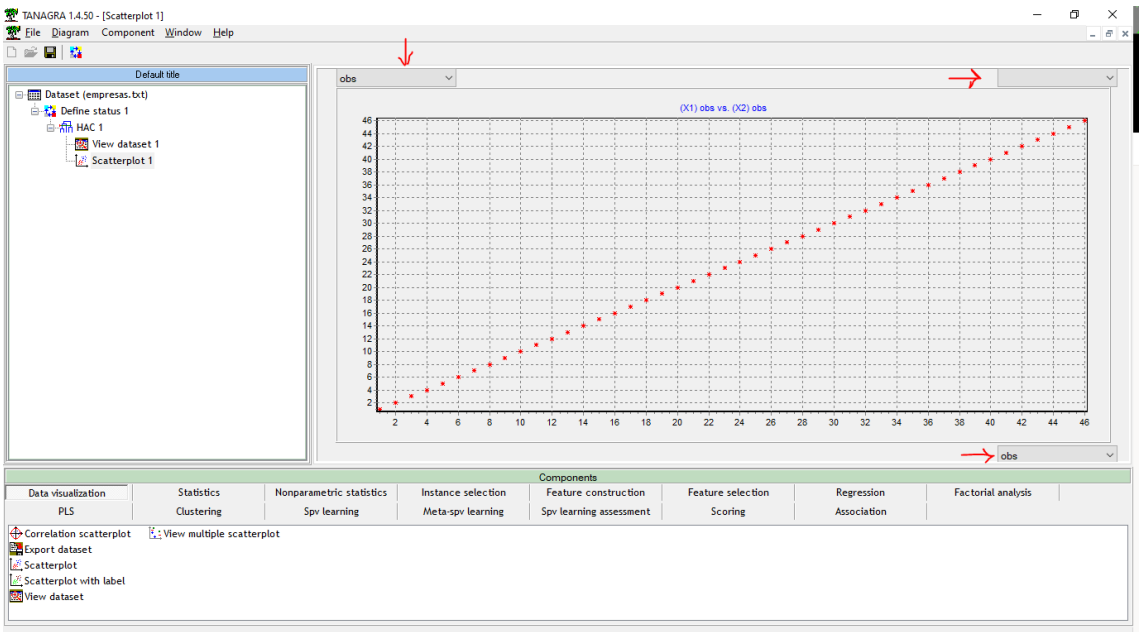
Si lo ejecutamos, observamos que la última columna nos dice en que “cluster” se encuentra el dato de esa fila (cada fila es una empresa):



	obs	FlujodeCaj	IngresoNet	ActivoCorr	ActivoCorr	Grupo	Cluster_HA
1	8	-0,0653	-0,0566	1,3737	0,4032	Quebró	c_hac_1
2	29	-0,0173	0,0233	2,0538	0,3484	No quebró	c_hac_1
3	25	0,1933	0,0473	2,2506	0,3309	No quebró	c_hac_1
4	40	-0,333	-0,0854	3,0124	0,473	No quebró	c_hac_3
5	42	0,5603	0,1112	4,2918	0,4443	No quebró	c_hac_2
6	24	0,3776	0,1075	3,2651	0,3548	No quebró	c_hac_2
7	41	0,4785	0,091	1,2444	0,1847	No quebró	c_hac_1
8	13	0,0109	0,0011	2,1495	0,6969	Quebró	c_hac_3
9	11	-0,2298	-0,2961	0,331	0,1824	Quebró	c_hac_1
10	20	0,1227	0,1055	1,1434	0,1655	Quebró	c_hac_1
11	38	0,2907	0,0597	1,8381	0,3786	No quebró	c_hac_1
12	34	0,1398	-0,0312	0,4611	0,2643	No quebró	c_hac_1
13	23	0,0769	0,0195	2,0069	0,5304	No quebró	c_hac_3
14	19	0,0115	-0,0032	1,2602	0,6038	Quebró	c_hac_3
15	22	0,5135	0,1001	2,4871	0,5368	No quebró	c_hac_3
16	45	0,1661	0,0351	2,4527	0,137	No quebró	c_hac_1
17	27	0,3132	0,0511	4,45	0,6852	No quebró	c_hac_2
18	3	0,0643	0,0156	1,0077	0,3978	Quebró	c_hac_1
19	10	-0,1353	-0,1433	1,4196	0,4347	Quebró	c_hac_1
20	35	0,1379	0,0728	2,6123	0,5151	No quebró	c_hac_3
21	16	0,3703	0,1098	1,9941	0,3828	Quebró	c_hac_1
22	9	0,0724	-0,0076	1,3723	0,3361	Quebró	c_hac_1
23	21	-0,2843	-0,2703	1,2722	0,5128	Quebró	c_hac_3
24	33	-0,0985	-0,0123	2,5029	0,5778	No quebró	c_hac_3

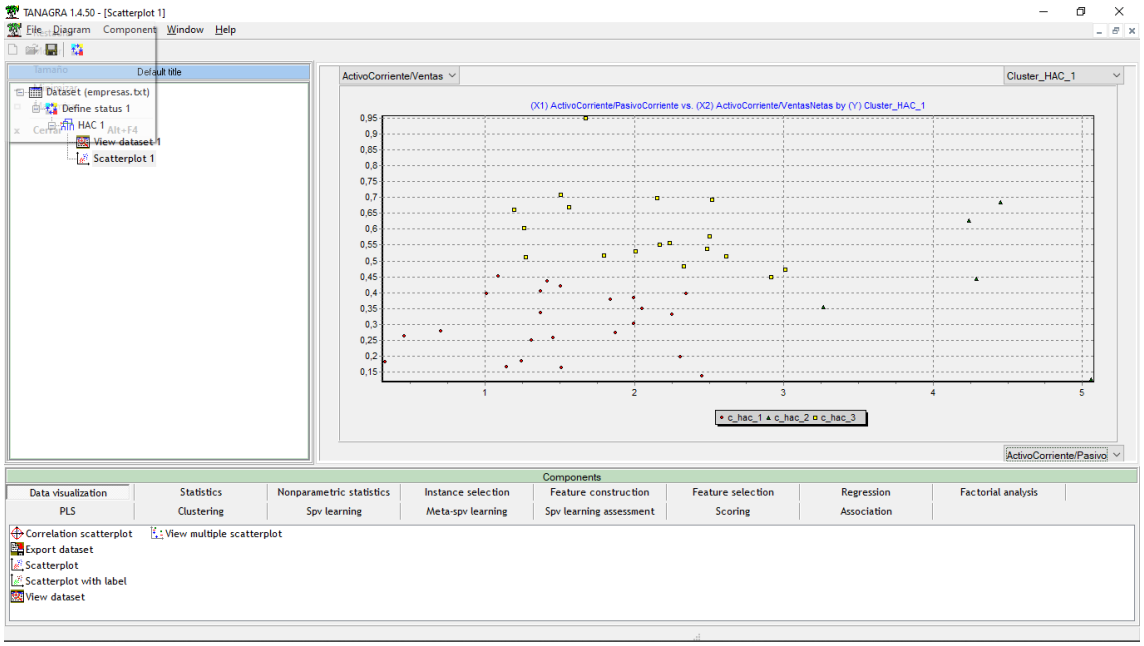
Para generar un diagrama de dispersión con esas variables, debemos ir a la pestaña “Data Visualization” y señalar la opción “Scatterplot”. Lo arrastramos hasta HAC 1.

Si lo ejecutamos observamos lo siguiente:



En cada uno de los ejes, pondremos los atributos que nosotros hemos considerado para generar el “clusters”.

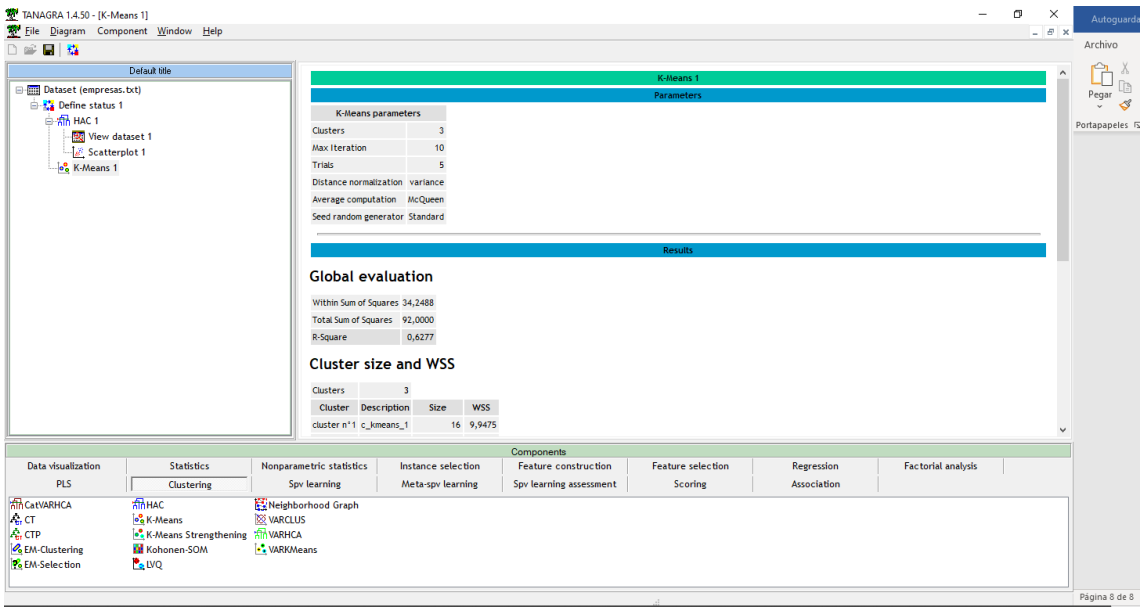
En el eje de las “x” pondremos ActivoCorriente/Pasivo corriente. En el eje de las “y” vamos a poner ActivoCorriente/VentasNetas. Por último, en la parte superior derecha seleccionaremos la opción “Cluster HAC” para generar puntos de distinto color dependiendo del “cluster” al que pertenezca.



Podemos observar que se han generado tres “clusters”. El primer “cluster” está representado por círculos rojos, el segundo con triángulos verdes y el tercero está representado por cuadrados amarillos, según reza la leyenda del gráfico.

2.

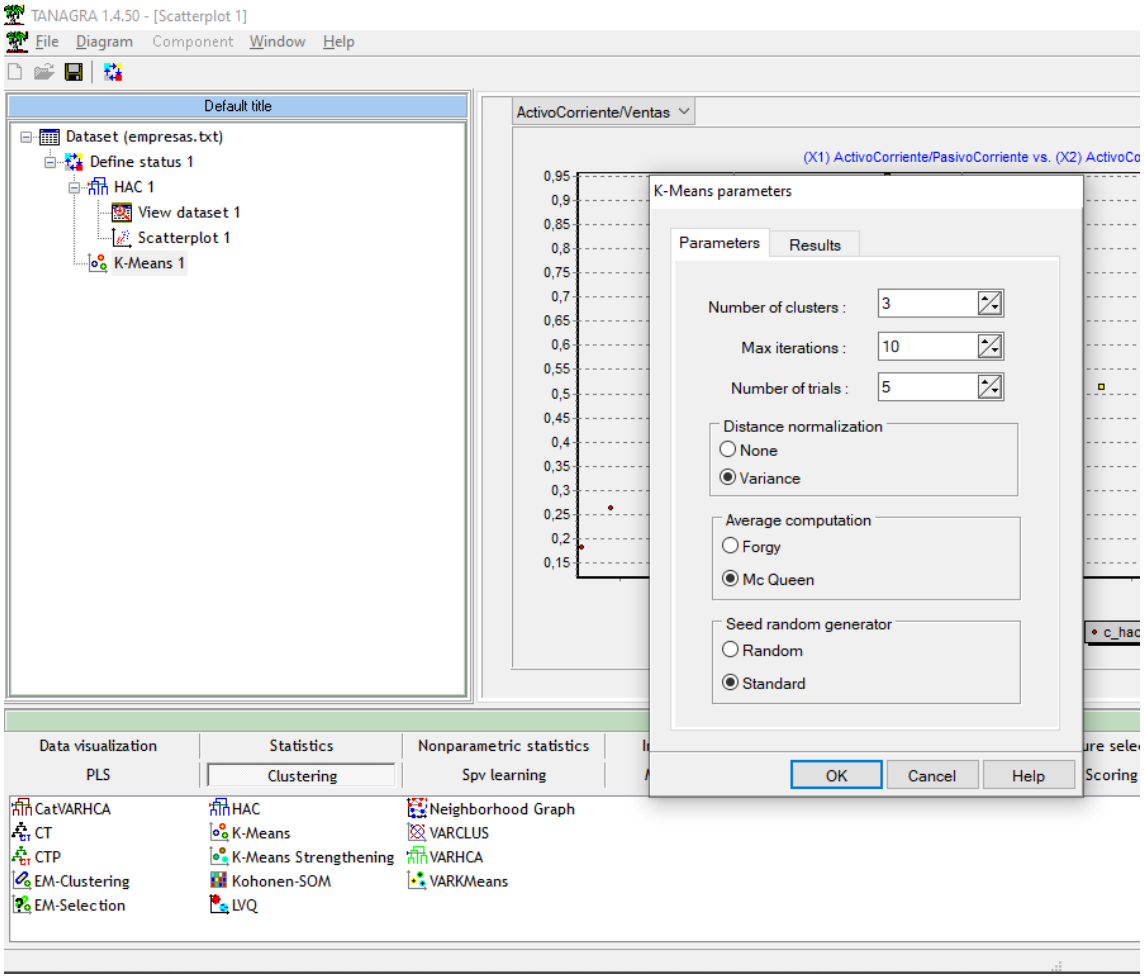
Ahora vamos a proceder a aplicar un algoritmo distinto de “clustering”, el K-Means. Para ello nos iremos a la pestaña “Clusterin” y seleccionaremos y arrastraremos hasta nuestro “Define Statues”, ejecutándolo con los parámetros por defecto.



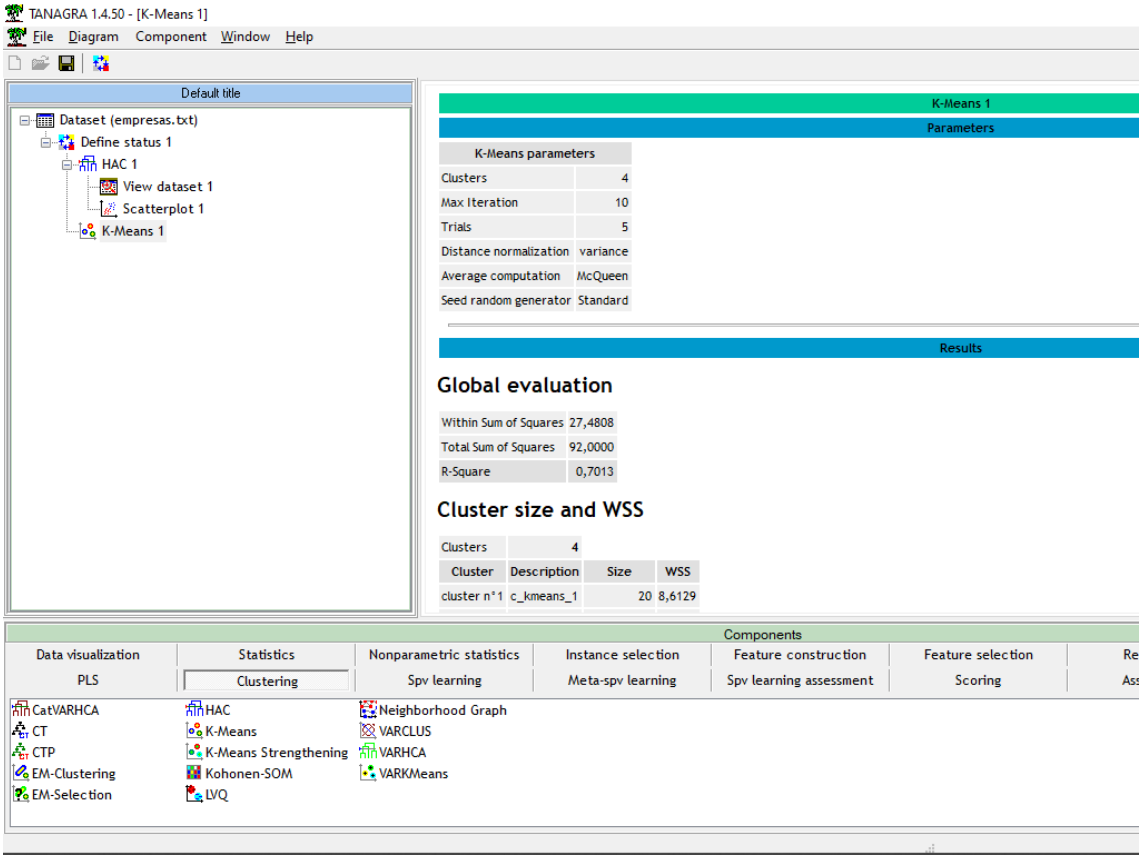
Se crean 3 “clusters”. La cantidad máxima de iteraciones son 10. Vemos que no se presenta el endrograma, ya que “K-Means” es un algoritmo de “clustering” no jerárquico y por lo tanto no hay un orden en las decisiones.

3.

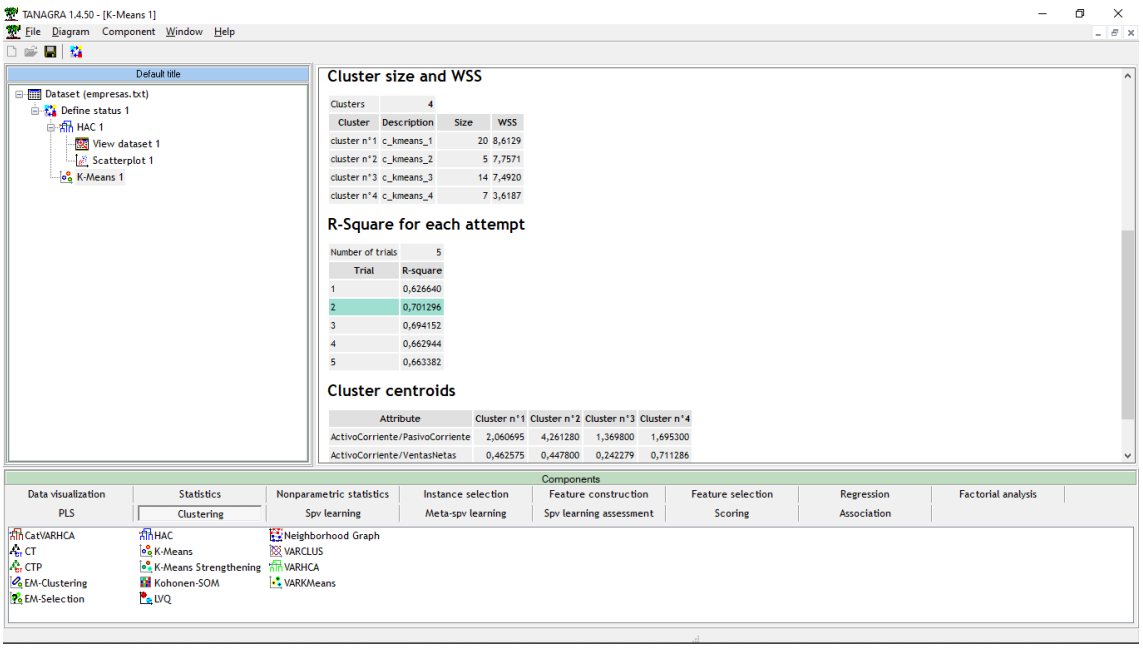
Vamos a generar ahora 4 “clusters” con la ayuda del algoritmo “K-Means”. Para ello, nos vamos encima de “K-Means” y hacemos click derecho y pulsamos “Parameters..”.



Donde dice “Number of clusters” vamos a indicar el número de “clusters” que queremos generar (4). Tras esto pulsamos “OK” y hacemos doble click en “K-Means” para ejecutar el algoritmo:

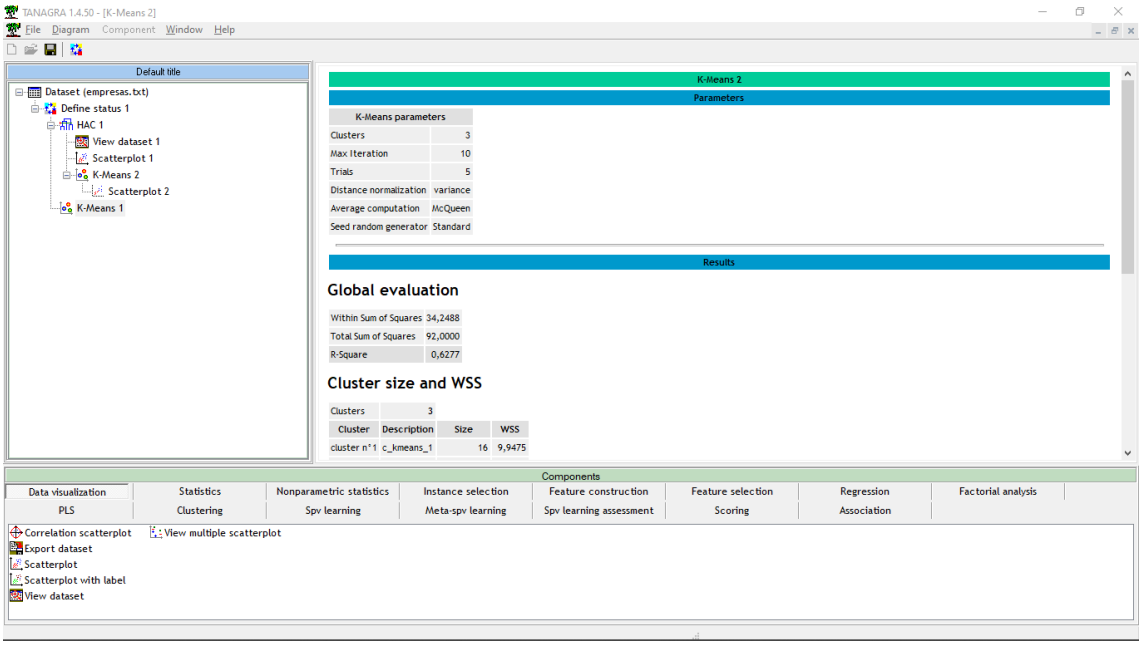


Vemos que se han generado los 4 “clusters” que esperábamos, y que el primero de ellos se compone de 20 elementos, el segundo de 5 elementos, el tercero de 14 elementos y el cuarto de 7 elementos.



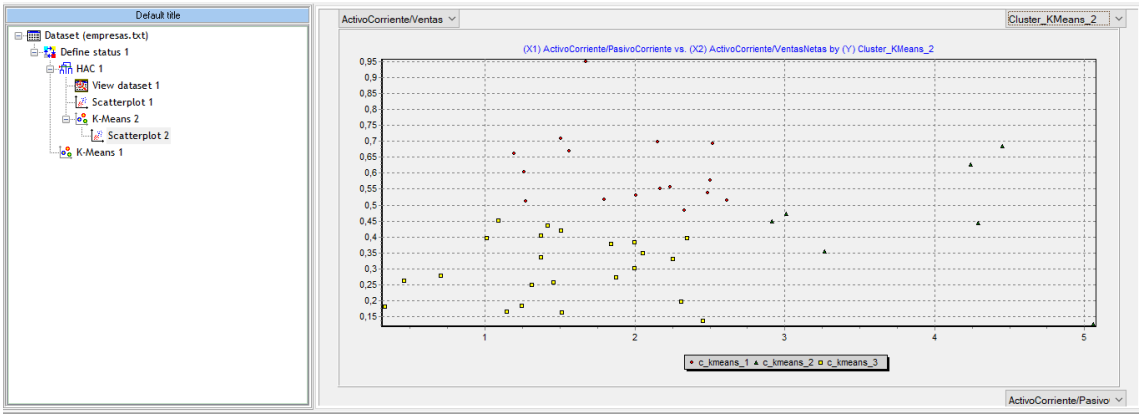
4.

Para generar de dispersión para representar indistintamente los “clusters” generados por “HAC1” o con “K-means”, debemos seleccionar en la pestaña “Clustering” el algoritmo “K-Means” y arrastrarlo hasta debajo de “HAC 1”. Tras esto lo ejecutamos haciendo doble click, y buscamos en la pestaña “Data View” el “Scatterplot”, el cual arrastramos debajo del nuevo “K-Means”.

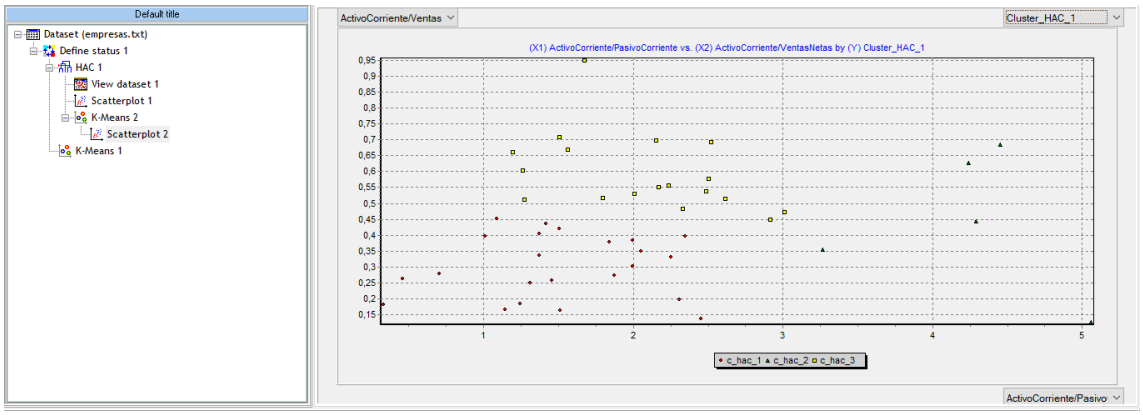


Hacemos doble click en el “Scatterplot” y procedemos a configurlo. En el eje de ordenadas pondremos el atributo “ActivoCorriente/VentasNetas”. En el eje de las ordenadas pondremos el atributo “ActivoCorriene/PasivoCorriente”, y vemos que en la esquina superior derecha podremos elegir con cual de los dos modelos representarlo, si con el “HAC” o con el “K-Means”.

Con “K-Means”:



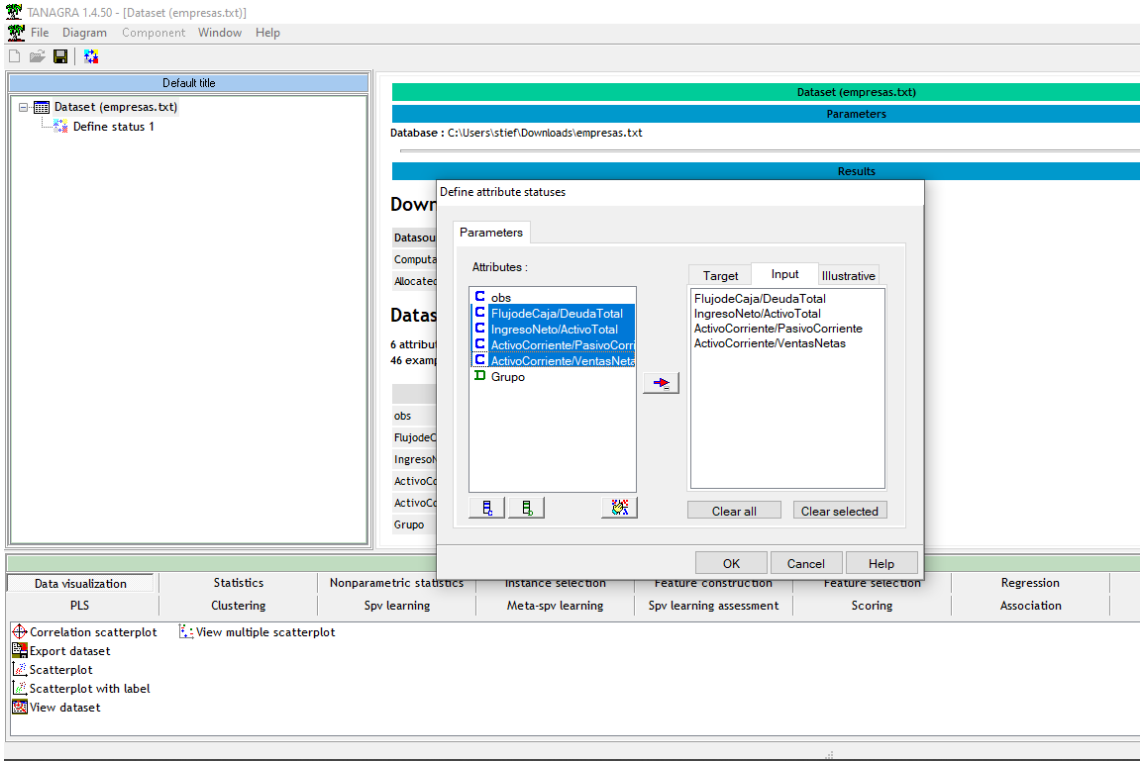
Con “HAC”:



PARTE 2.

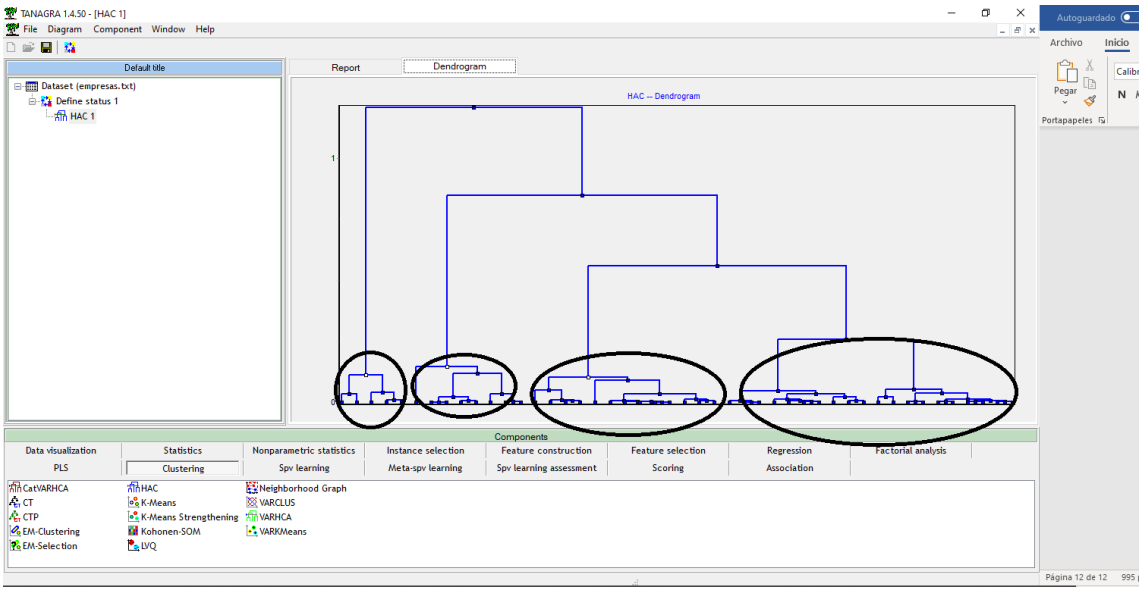
1.

En esta ocasión vamos a tener 4 variables como entradas. Por tanto abrimos el archivo empresas.txt, nos vamos a “Define Status” y añadimos como “Imput” las 4 varaibles:

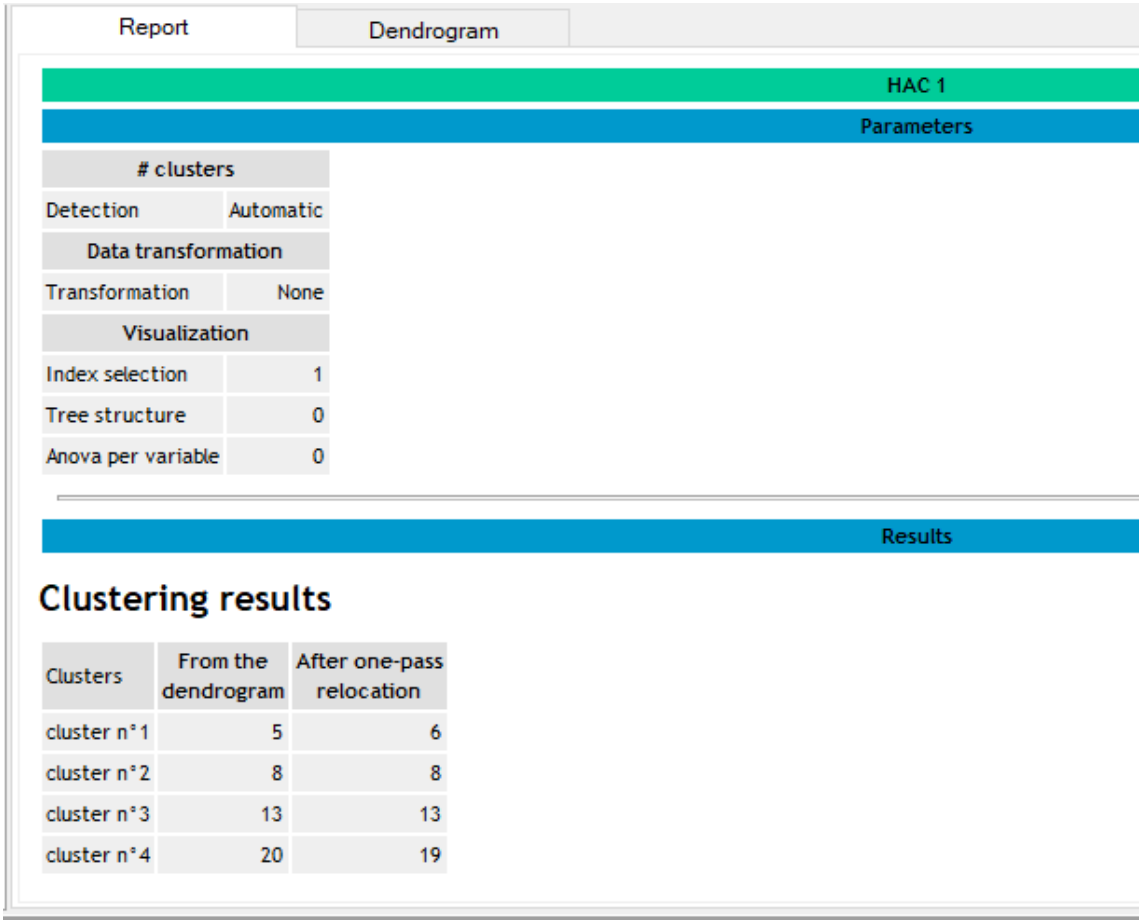


Tras esto vamos a la pestaña “Clustering” y seleccionamos el algoritmo HAC y arrastramos hasta nuestro “Define Status”. Ejecutamos el “HAC” y nos vamos a la pestaña “Dendogram”.

Aquí podemos observar que podríamos definir 4 “clusters”:



Si nos vamos a la pestaña “Report” podremos ver información útil. Podemos observar que realmente se han creado 4 “clusters”, el primero de ellos con 5 elementos, el segundo con 8 elementos, el tercero con 13 elementos y el cuarto con 20 elementos.



Report | Dendrogram

HAC 1

Parameters

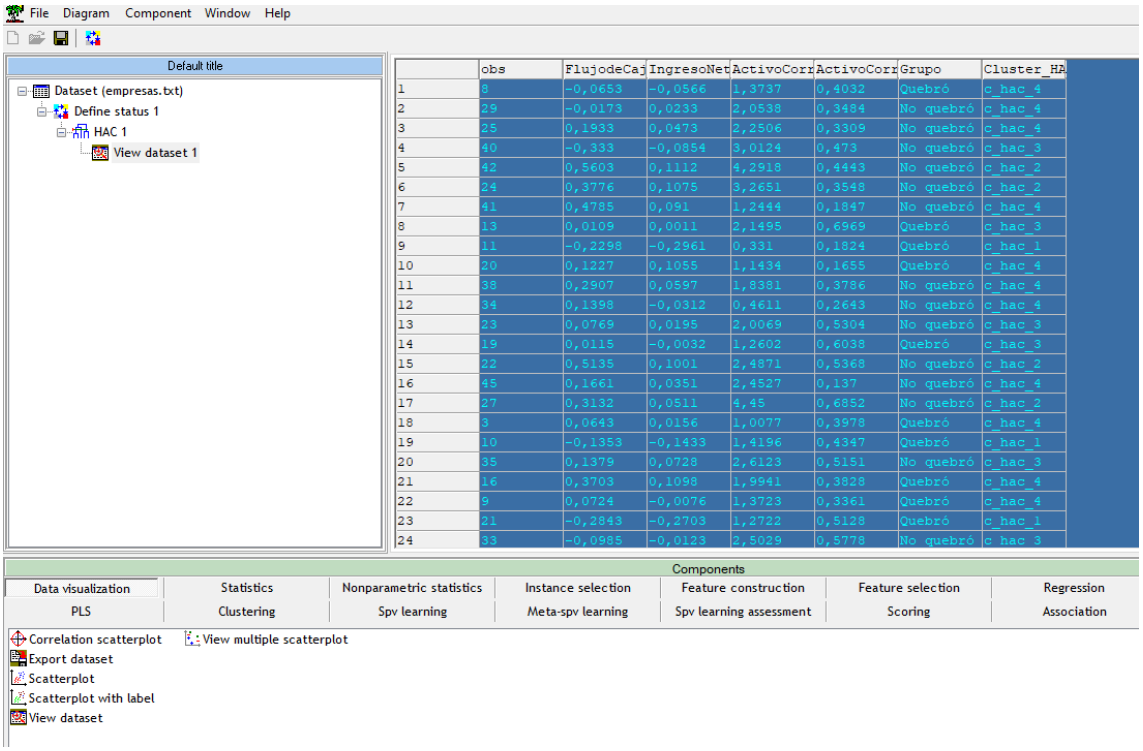
# clusters	
Detection	Automatic
Data transformation	
Transformation	None
Visualization	
Index selection	1
Tree structure	0
Anova per variable	0

Results

Clustering results

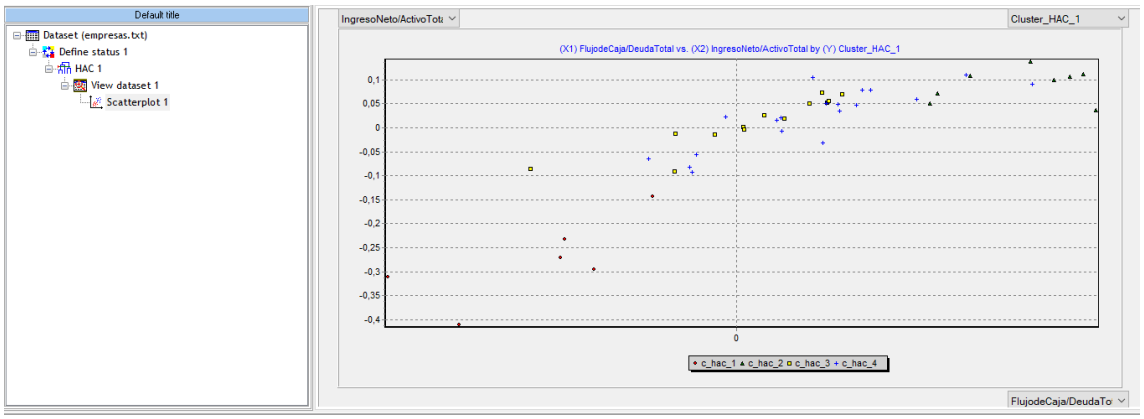
Clusters	From the dendrogram	After one-pass relocation
cluster n°1	5	6
cluster n°2	8	8
cluster n°3	13	13
cluster n°4	20	19

Ahora nos vamos a la pestaña “Data Visualization”, seleccionamos “View Dataset” y lo arrastramos hasta debajo de “HAC”. Lo ejecutamos y vemos que la última columnas nos indica a que “cluster” pertenece cada fila, siendo cada fila una empresa distinta.



	obs	FlujodeCaj	IngresoNet	ActivoCorr	ActivoCorr	Grupo	Cluster_HA
1	8	-0,0653	-0,0566	1,3737	0,4032	Quebró	c_hac_4
2	29	-0,0173	0,0233	2,0538	0,3484	No quebró	c_hac_4
3	25	0,1933	0,0473	2,2506	0,3309	No quebró	c_hac_4
4	40	-0,333	-0,0854	3,0124	0,473	No quebró	c_hac_3
5	42	0,5603	0,1112	4,2918	0,4443	No quebró	c_hac_2
6	24	0,3776	0,1075	3,2651	0,3548	No quebró	c_hac_2
7	41	0,4785	0,091	1,2444	0,1847	No quebró	c_hac_4
8	13	0,0109	0,0011	2,1495	0,6969	Quebró	c_hac_3
9	11	-0,2298	-0,2961	0,331	0,1824	Quebró	c_hac_1
10	20	0,1227	0,1055	1,1434	0,1655	Quebró	c_hac_4
11	38	0,2907	0,0597	1,8381	0,3786	No quebró	c_hac_4
12	34	0,1398	-0,0312	0,4611	0,2643	No quebró	c_hac_4
13	23	0,0769	0,0195	2,0069	0,5304	No quebró	c_hac_3
14	19	0,0115	-0,0032	1,2602	0,6038	Quebró	c_hac_3
15	22	0,5135	0,1001	2,4871	0,5368	No quebró	c_hac_2
16	45	0,1661	0,0351	2,4527	0,137	No quebró	c_hac_4
17	27	0,3132	0,0511	4,45	0,6852	No quebró	c_hac_2
18	3	0,0643	0,0156	1,0077	0,3978	Quebró	c_hac_4
19	10	-0,1353	-0,1433	1,4196	0,4347	Quebró	c_hac_1
20	35	0,1379	0,0728	2,6123	0,5151	No quebró	c_hac_3
21	16	0,3703	0,1098	1,9941	0,3828	Quebró	c_hac_4
22	9	0,0724	-0,0076	1,3723	0,3361	Quebró	c_hac_4
23	21	-0,2843	-0,2703	1,2722	0,5128	Quebró	c_hac_1
24	33	-0,0985	-0,0123	2,5029	0,5778	No quebró	c_hac_3

Para generar un gráfico de dispersión usaremos el “Scatterplot” que encontramos en la pestaña “Data Visualization”. Lo seleccionamos y arrastramos hacia debajo de “HAC”. Ejecutamos y en la gráfica podremos elegir que variables queremos enfrentar para obtener un gráfico de dispersión. Por ejemplo, si seleccionamos el eje de abcisas como FlujodeCaja/DeudaTotal y el eje de las ordenadas como IngresoNeto/ActivoTotal, obtendremos el siguiente gráfico de dispersión:

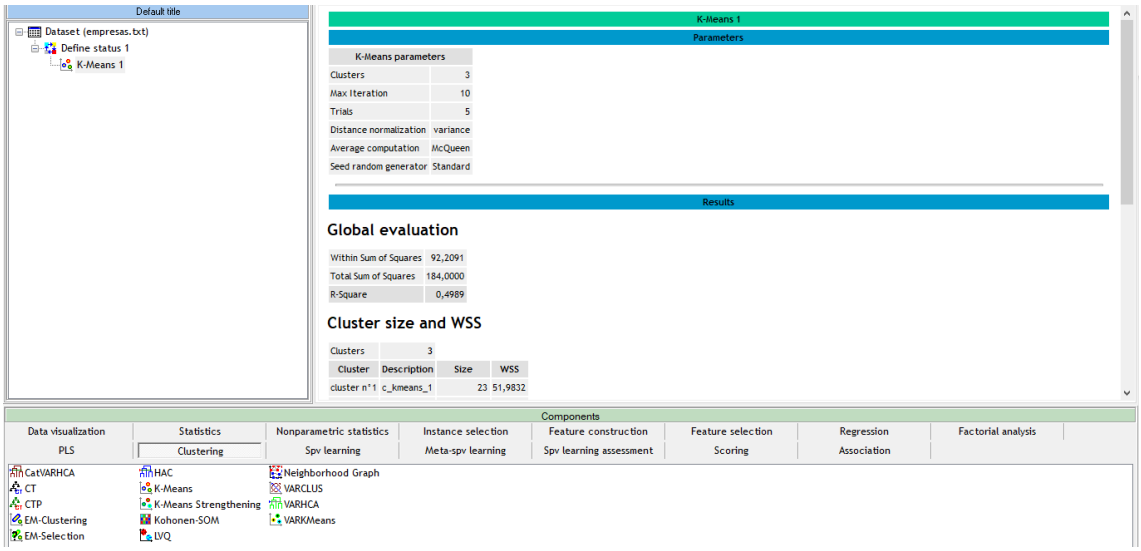


Observamos que el “cluster” uno se representa con círculos rojos, el dos con triángulos verdes, el tres con cuadrados amarillos y el cuarto con cruces azules.

Procedemos ahora a buscar el algoritmo “K-Means” en la pestaña “Clustering”. Lo seleccionamos y arrastramos debajo de nuestro “Define Status”.

2.

Vamos a aplicar ahora el algoritmo “K-Means”. Lo buscamos en la pestaña “Clustering”, seleccionamos y arrastramos hasta nuestro “Define Status”. Vamos a dejar los parámetros por defecto, por lo que hacemos doble click para ejercutarlo.

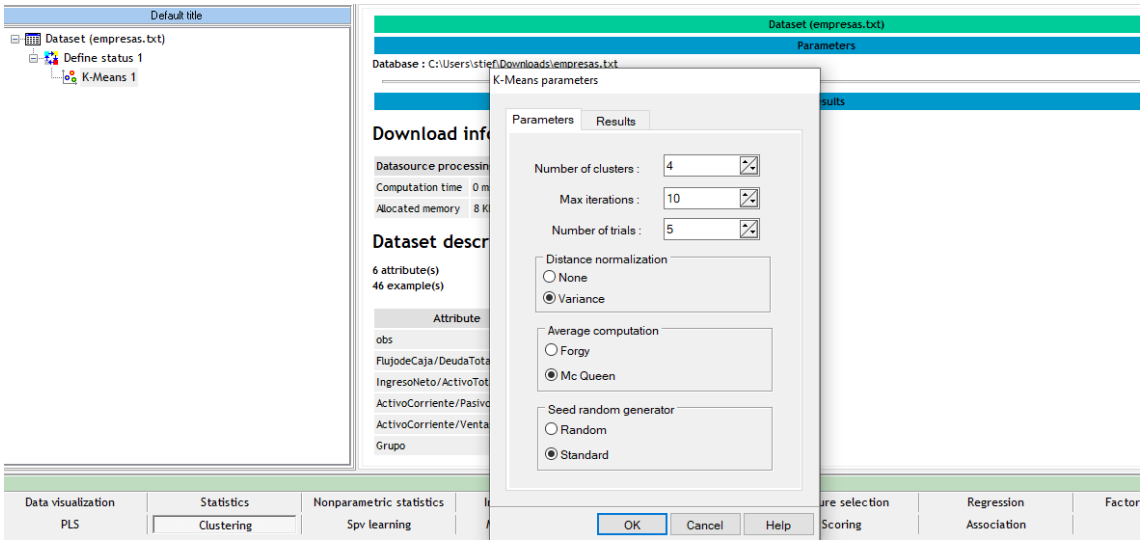


El número de “clusters” creados es 3. El máximo número de iteraciones 10.

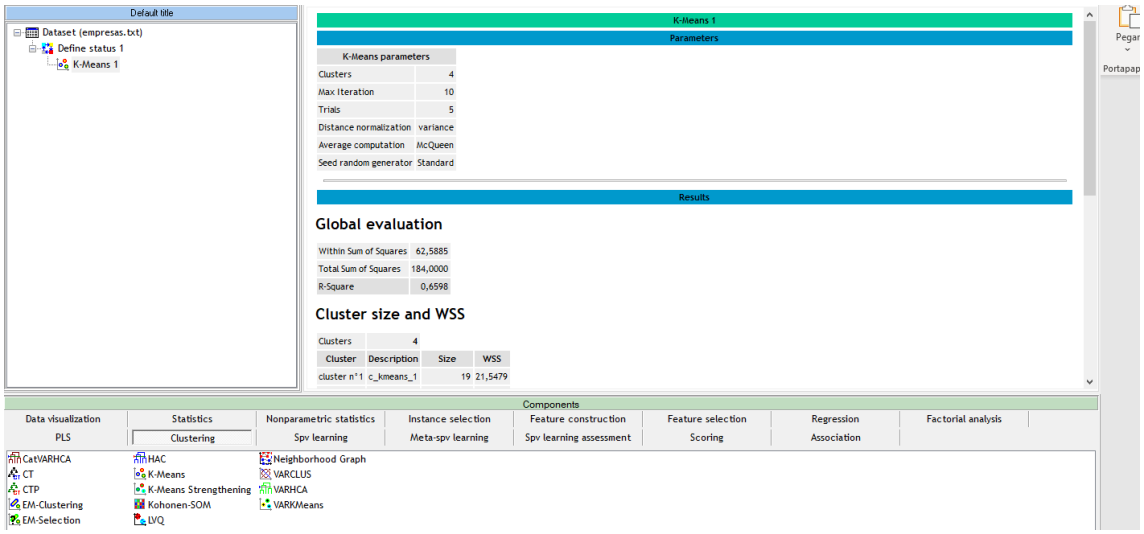
Vemos que no se presenta el endrograma, ya que “K-Means” es un algoritmo de “clustering” no jerárquico y por lo tanto no hay un orden en las decisiones.

3.

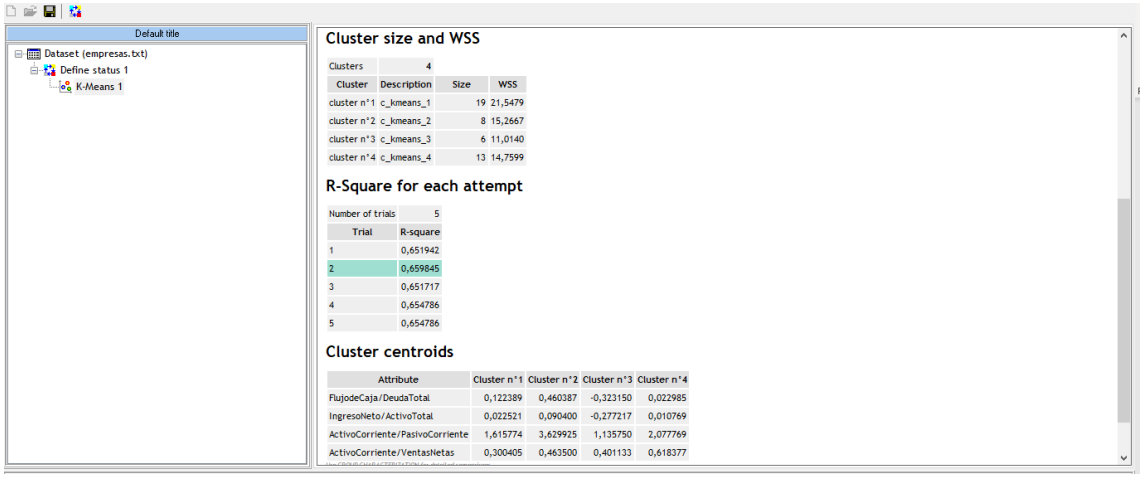
Queremos generar 4 “clusters”, por lo que hacemos click con el botón derecho sobre el algoritmo “K-Means” y seleccionamos la opción parámetros. Aquí debemos cambiar el número que indica la pestaña “Number of clusters” por 4.



Hacemos doble click para ejecutar el algoritmo y observamos la ventana emergente:



Podemos ver que se han generado 4 “clusters” tal y como habíamos determinado, y que el primero de ellos contiene 19 elementos, el segundo 8 elementos, el tercero 6 elementos y el cuarto 13 elementos.



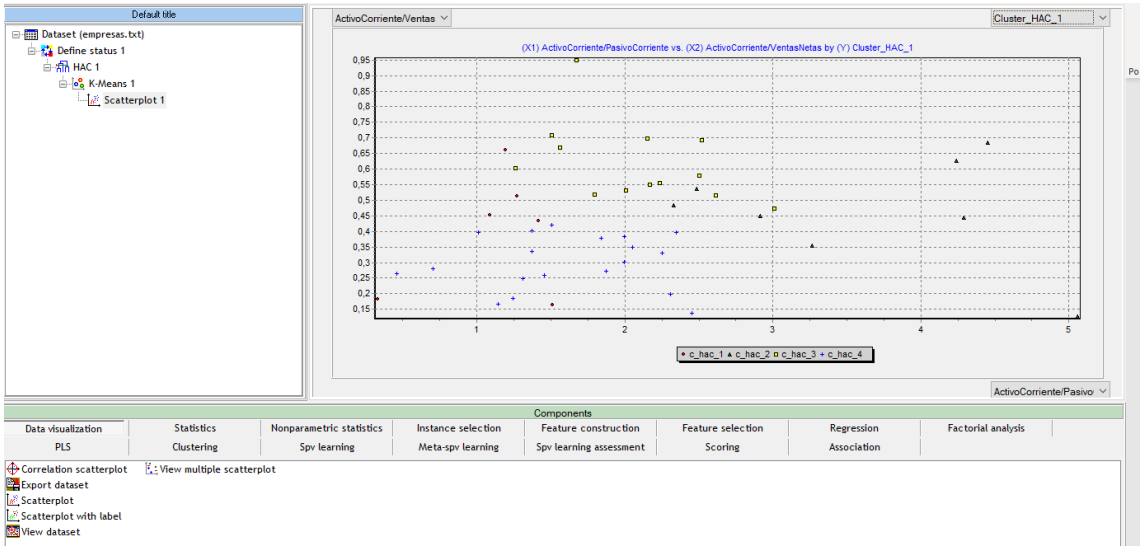
4.

Para poder mostrar el diagrama de dispersión de cualquiera de los dos algoritmos, estos tiene que estar en la misma “rama”.

Aplicamos al “Define Status”, que contiene las 4 variables como entradas, ambos algortimos de “clustering”: el algoritmo HAC y el algoritmo K-Means.

En el eje de ordenadas pondremos el atributo “ActivoCorriente/VentasNetas”. En el eje de las ordenadas pondremos el atributo “ActivoCorriene/PasivoCorriente”, y vemos que en la esquina superior derecha podremos elegir con cual de los dos modelos representarlo, si con el “HAC” o con el “K-Means”.

Con el algoritmo “HAC” el diagrama de dispersión sería el siguiente:



Con el algoritmo “K-Means” el diagrama de dispersión sería el siguiente: