Tarea 5 Mayo 2020 Knime: Clustering Jeràrquico Parte 2

1. **Describa los datos iris. Puede usar como fuente https://archive.ics.uci.edu/ml/datasets/iris u otras de su preferencia.**
2. **Si no lo tiene instalado: Descargar e instalar KNIME Konstanz Information Miner (https://www.knime.org/downloads/overview)**
3. **Crear un workspace (File -> New)**
4. **Incluir en el workspace los nodos:**

- **IO ArffReader,**
- **Column Filter: incluir dos atributos: . petal length, . petal width**
- **Hierarchical Clustering: 3 clusters, distancia euclideana, linkage type: average**
- **Color Manager: verde, rojo azul para cluster 0,1,2**
- **ScatterPlot para mostrar x: petal length, . y: petal width**
- **En el diagrama de dispersión observa un 'buen' clustering? Fundamente su respuesta.**

5 . Repita el punto 4 pero utilizando las variables sepal length, . sepal width

1.

Estos datos se refieren a los tipos de flor que nos podemos encontrar según los atributos de la misma. Los datos iris se usan mucho para el testeo de procesimientos de clasificación.

En función de 4 atributos de las flores, podemos saber cual de los tres tipos de iris es cada flor. Si abrimos el enlace https://archive.ics.uci.edu/ml/datasets/iris vemos lo siguiente:

Tenemos 4 atributos numéricos y 150 casos. Estos 4 atributos numéricos corresponden a 3 tipos de plantas. Se representa las plantas por el largo y ancho del cáliz, y por el largo y ancho del pétalo. Con estos 4 atributos podremos clasificar las flores en Iris Setosa, Iris Versicolour e Iris Virginia.

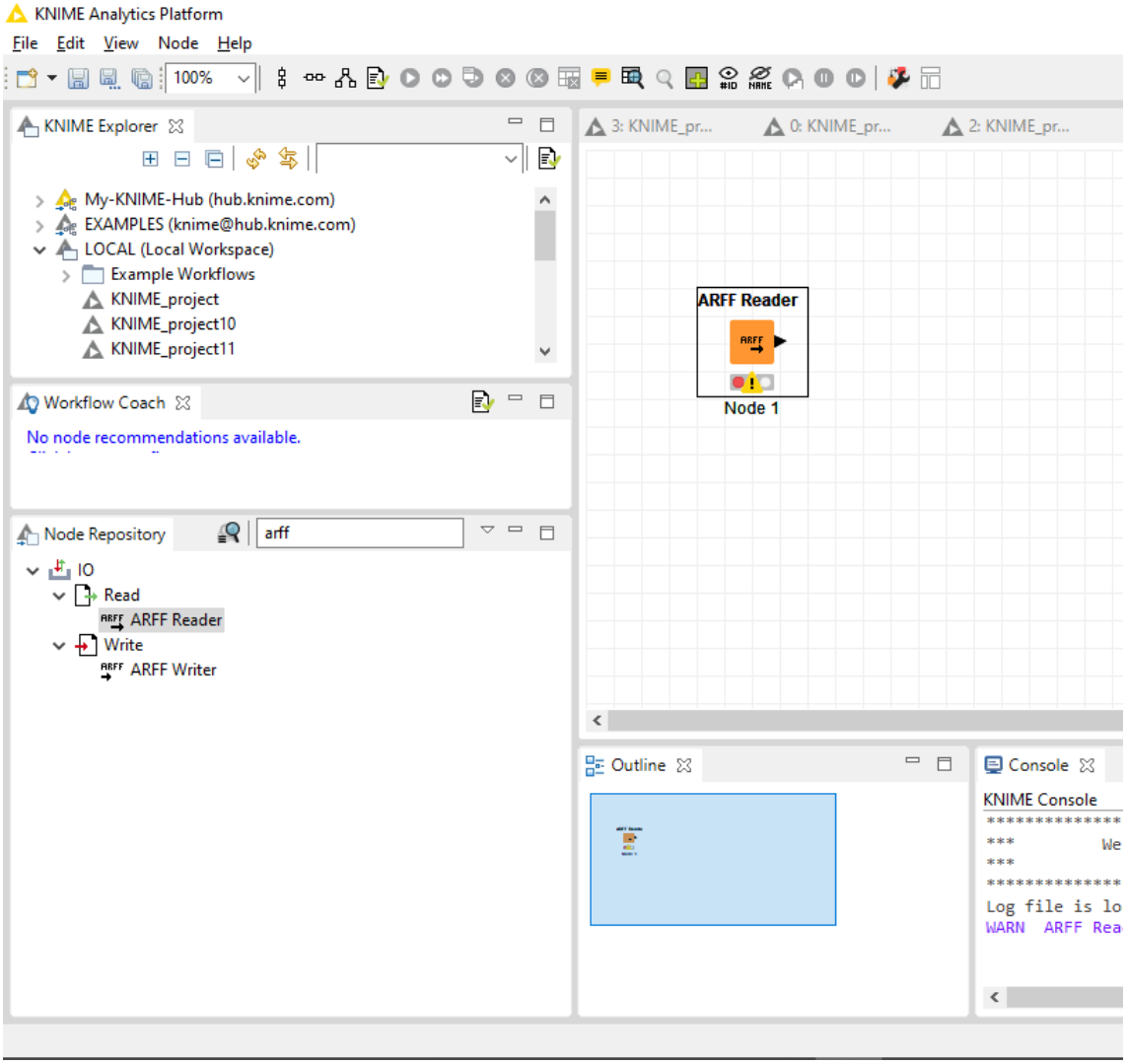
2, 3 y 4.

Primero creamos un espacio de trabajo seleccionando File → New File → New KNIME Workflow.

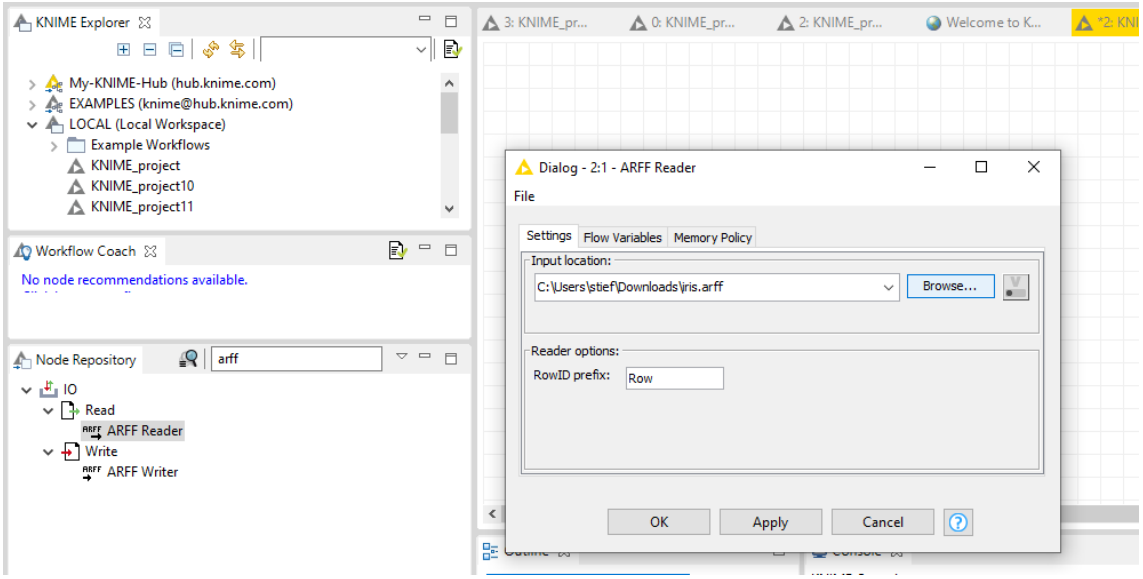
Una vez creado el espacio de trabajo, es hora de ir añadiendo los módulos.

Vamos a “Node Repository” y en la barra de búsqueda buscamos un nodo “ARFF Reader”, ya que el archivo que vamos a trabajar es de tipo .arff.

Seleccionamos el nodo y lo arrastramos hasta nuestro espacio de trabajo.

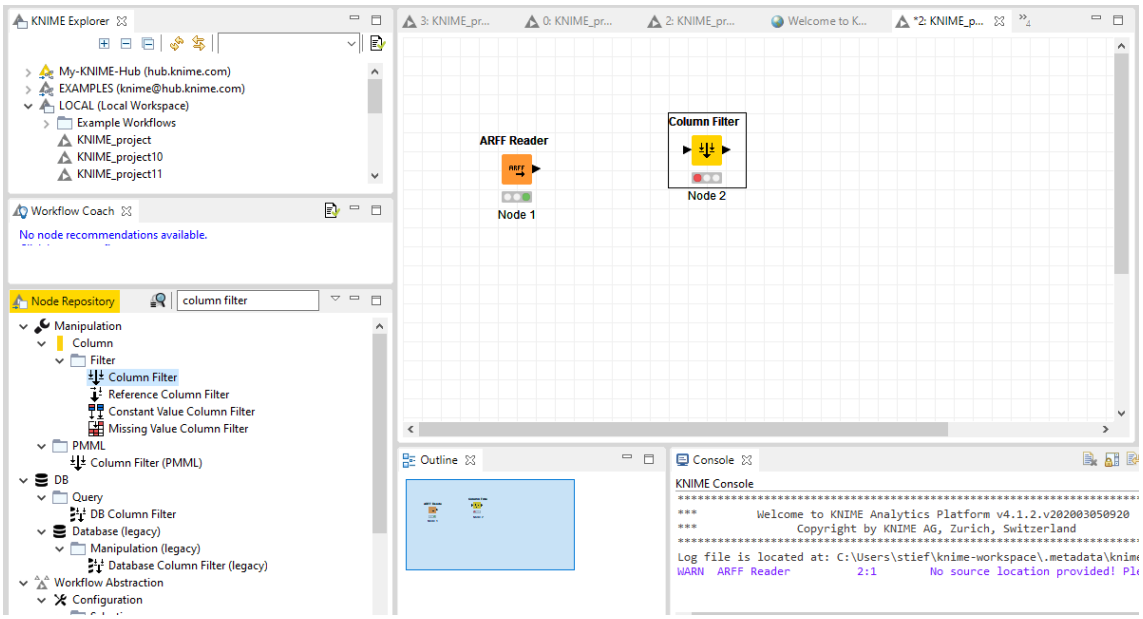


Ahora procedemos a configurarlo seleccionándolo y pulsando F6. En el menú de configuración del nodo buscamos el archivo que queremos trabajar, en este caso iris.arff



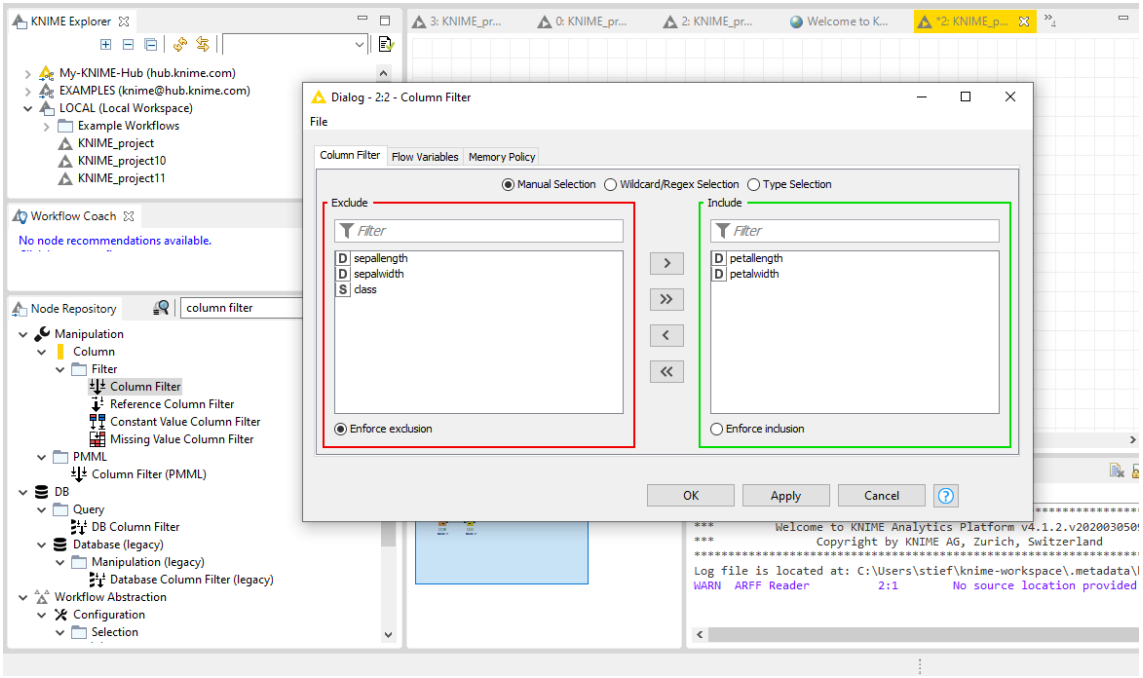
Con el nodo seleccionado, pulsamos F7 para ejecutarlo.

Como hemos visto, entre los datos tenemos 4 atributos numéricos, y uno que es categórico, que es el tipo de flor. Queremos realizar un filtrado de dichos datos y usar solamente dos atributos. Para ello vamos a ayudarnos del nodo “Column Filter”. Lo buscamos en “Node Repository” y lo añadimos al espacio de trabajo.



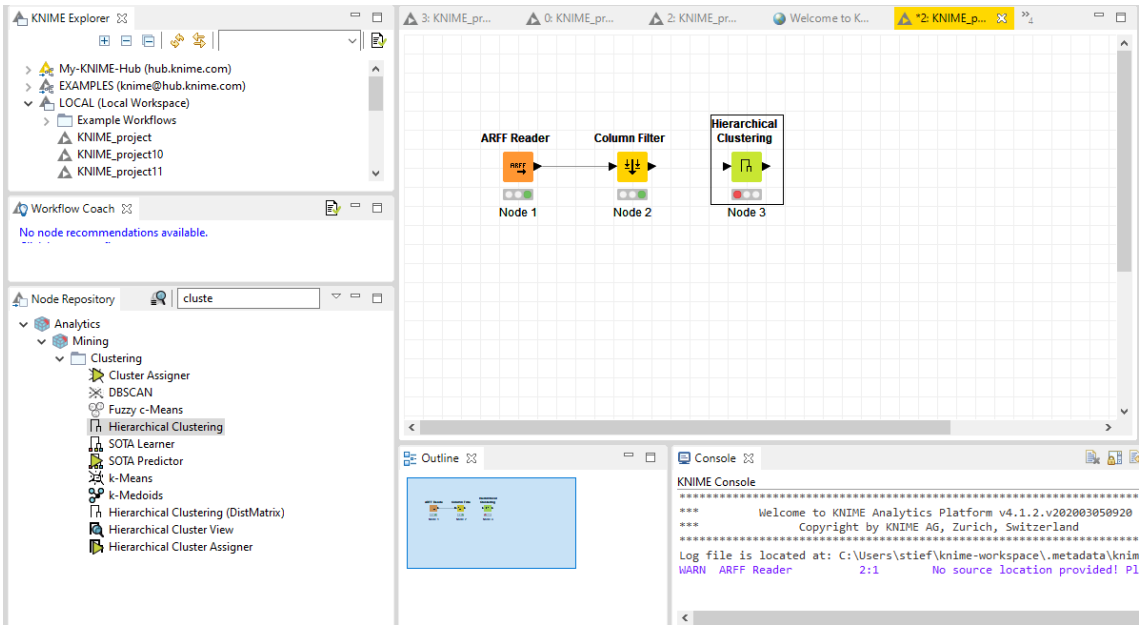
Para configurarlo primeros debemos unirlo con el nodo “ARFF Reader”, ya que si no, no tenemos datos que poder filtrar.

Seleccionamos el nodo y pulsamos F6. Ahora elegiremos los atributos que nos interesan, en este caso “petallength” y “petalwidth”, y el resto los excluimos.



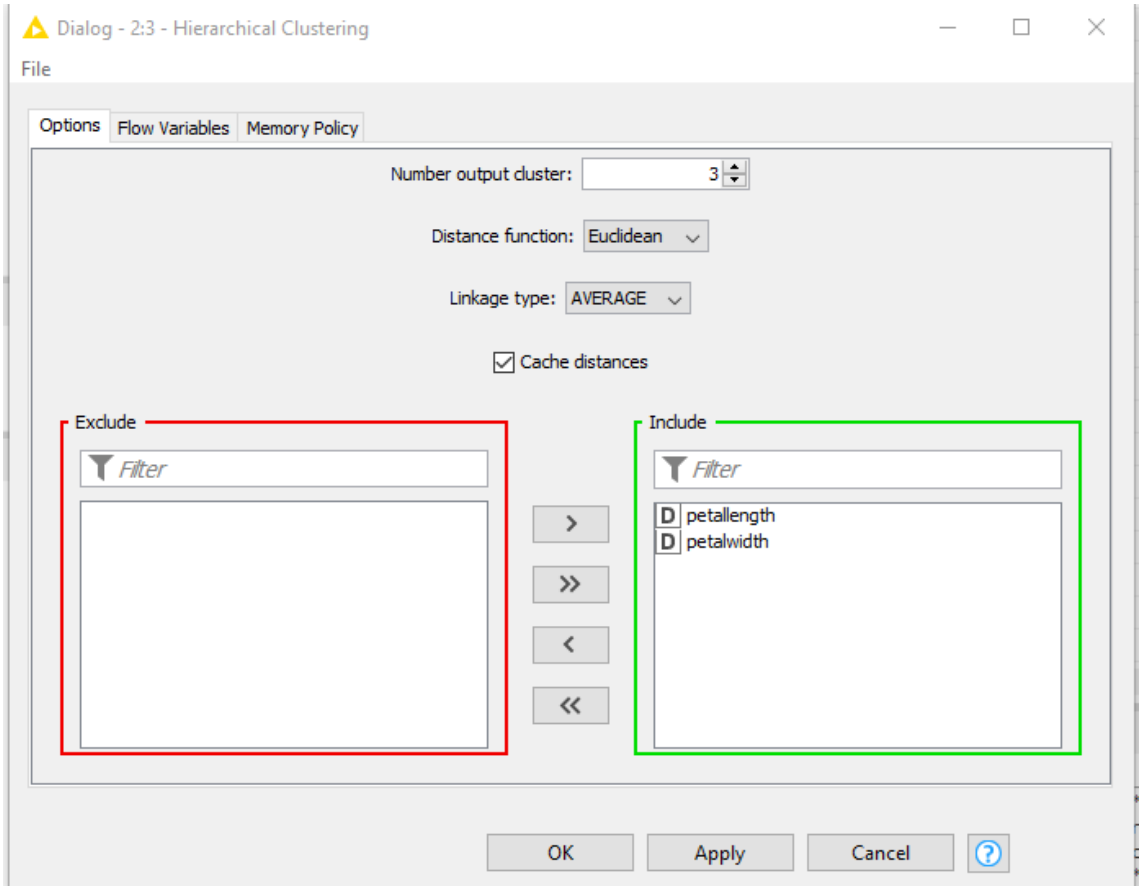
Para ejecutar el nodo lo seleccionamos y pulsamos F7.

Ahora debemos buscar en “Node Repository” el nodo “Hierarchical Clustering” y lo arrastramos al espacio de trabajo.



Lo conectamos con el nodo “columna filter” y pulsamos F6 para ejecutarlo.

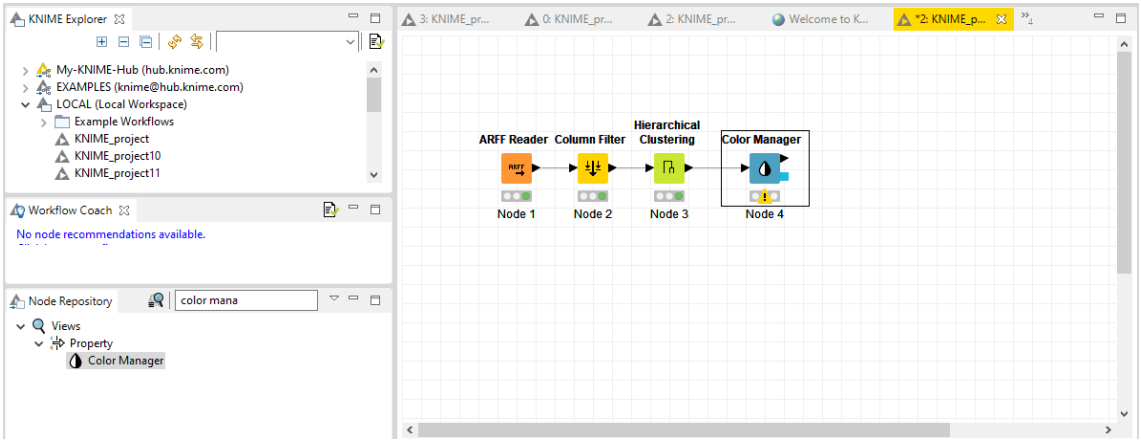
En el menú de configuración pondremos los siguientes parámetros:



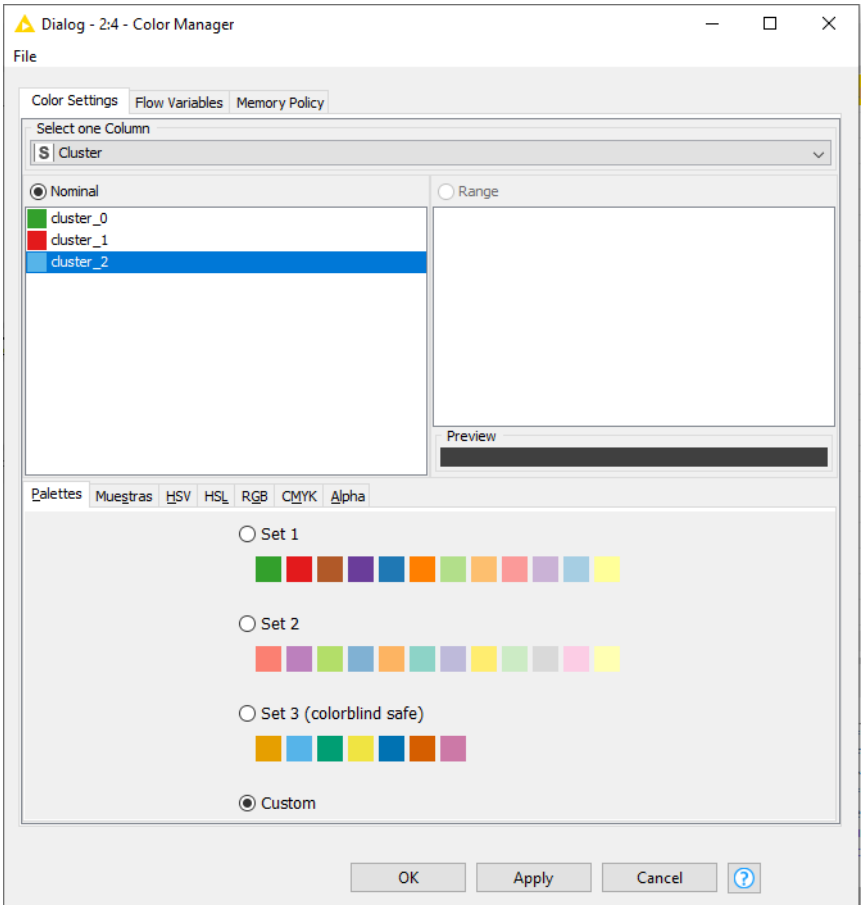
Luego procedemos a ejecutarlo pulsando F7.

Procedemos ahora a crear un nodo que nos permitirá gestionar los colores que queremos en nuestro cluster.

Para ello, en “Node Repository” buscaremos el nodo “Color Manager” y lo arrastraremos hasta el espacio de trabajo.

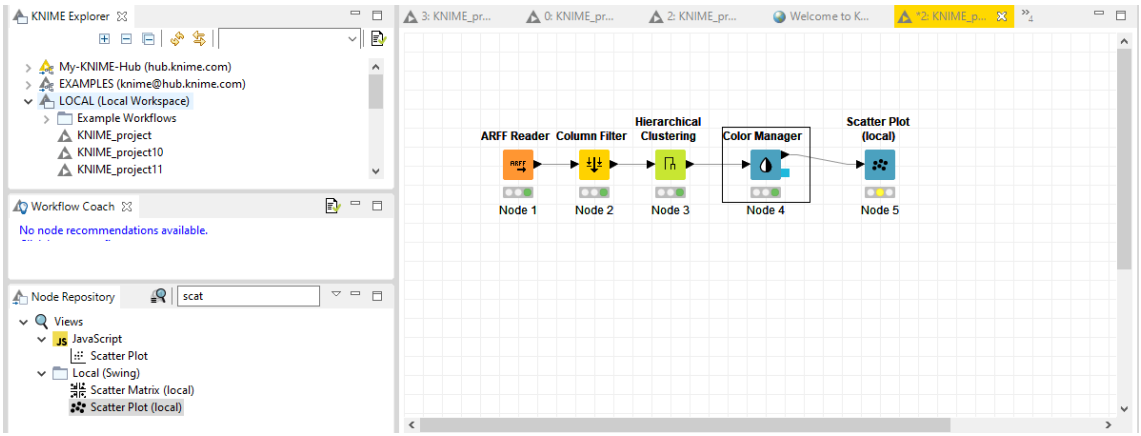


Lo seleccionamos y pulsamos F6 para empezar a configurarlo. Elegiremos el color verde para el cluster_0, el rojo para el 1 y el azul para el 2.

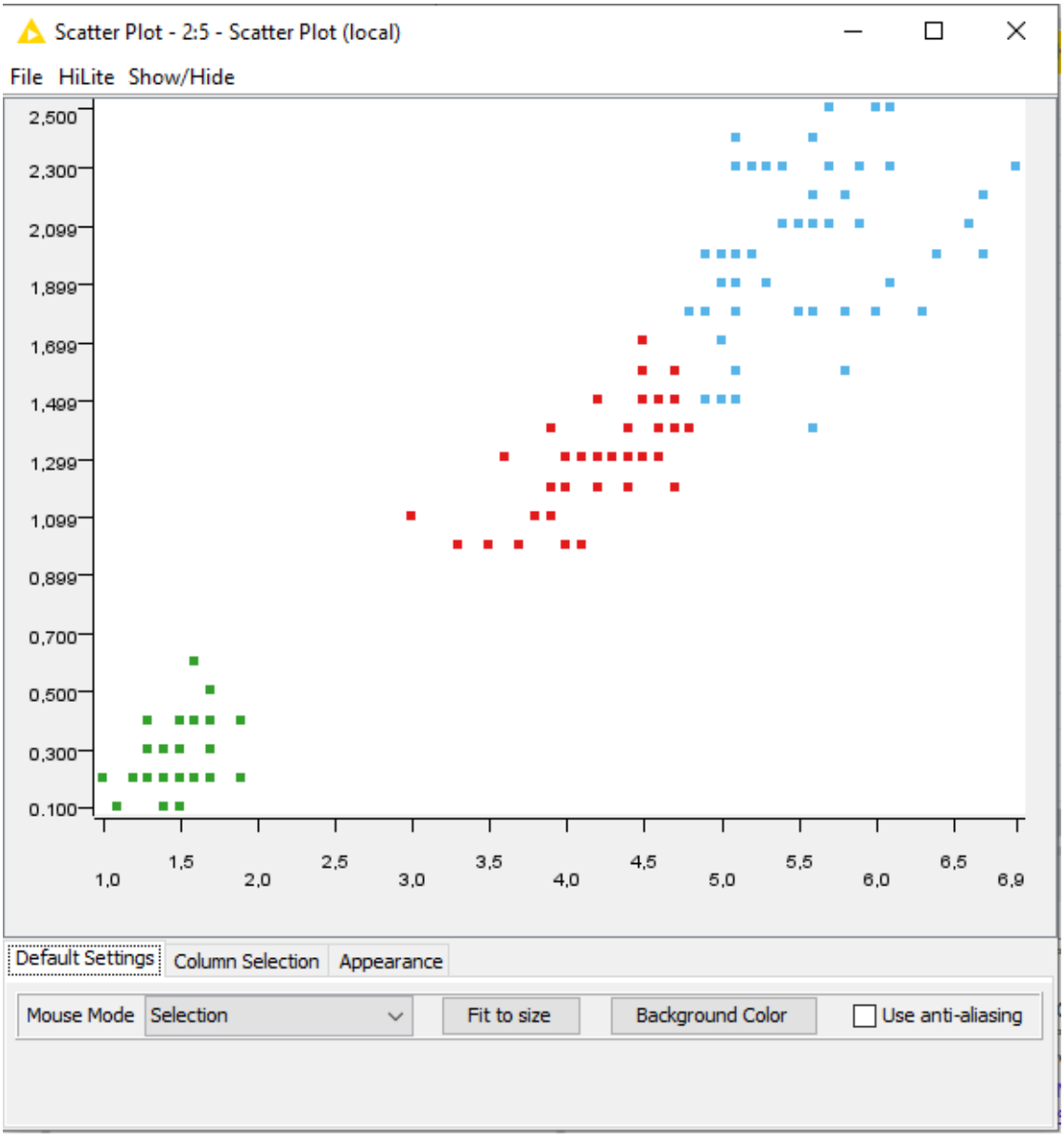


Aceptamos y pulsamos F7 para ejecutar el nodo.

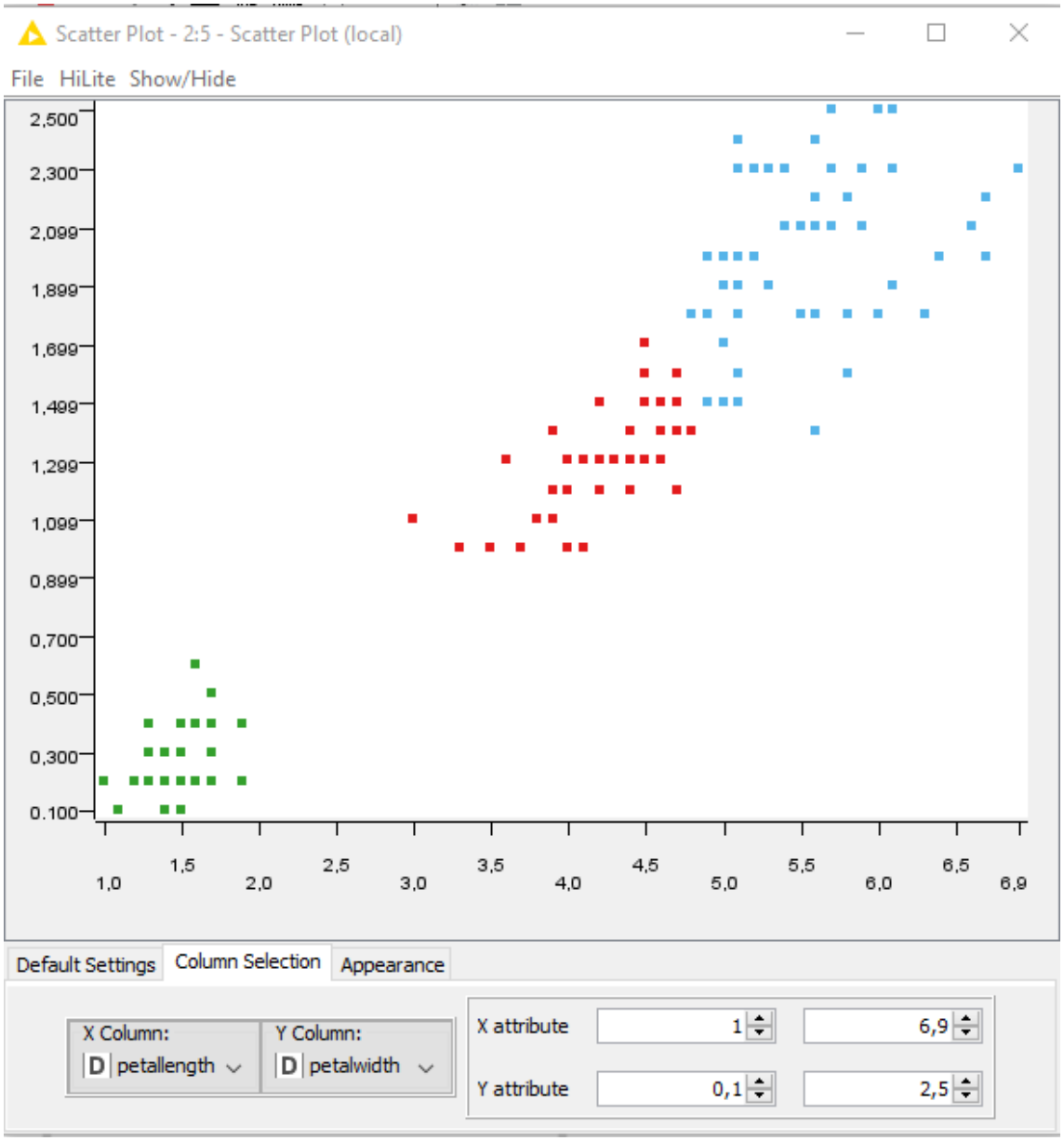
Por último, añadiremos el nodo “Scatterplot”.



Dejamos la configuración por defecto y ejecutamos pulsando F7. Haciendo click con el botón derecho seleccionamos la opción “View: Scatter Plot”.



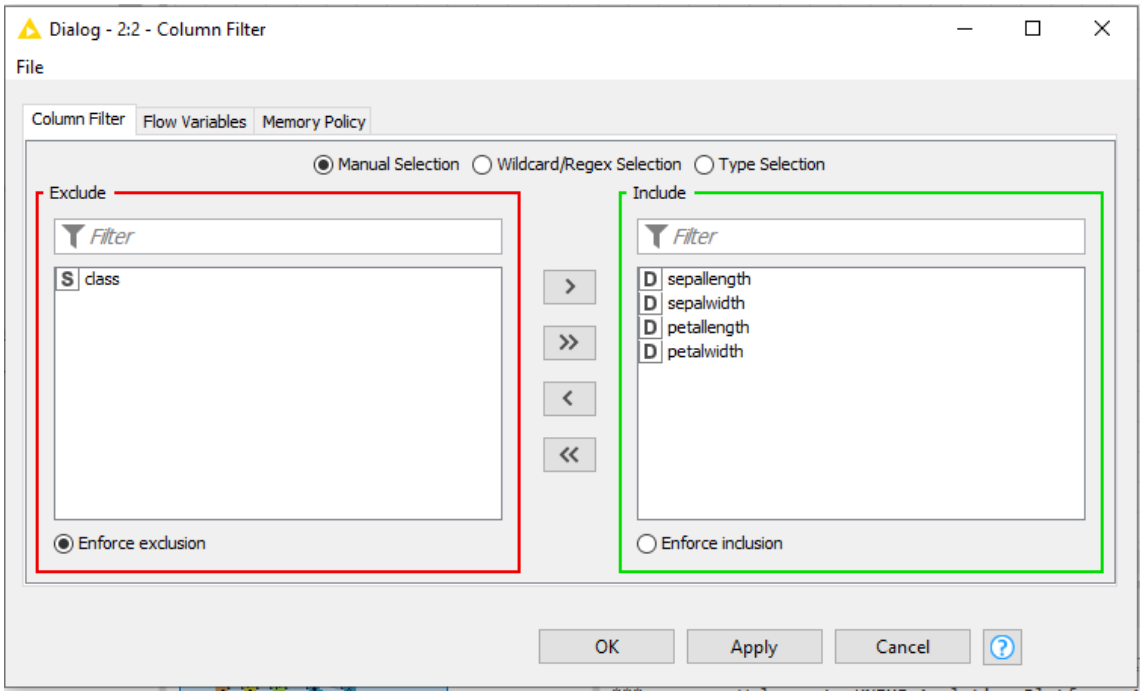
Ahora modificaremos ciertos parámetros de la gráfica. En primer lugar nos aseguramos que donde dice “Column Selection”, en el eje x se representa “petallengh” y en el eje y se representa “petalwidth”.



Podemos observar en este diagrama es que los grupos se generaron automáticamente, pero muestran que de alguna forma los que aparecen en el mismo cluster parecen estar juntos, que era lo que buscábamos al generar automáticamente los clusters. Por tanto se observa un buen clustering.

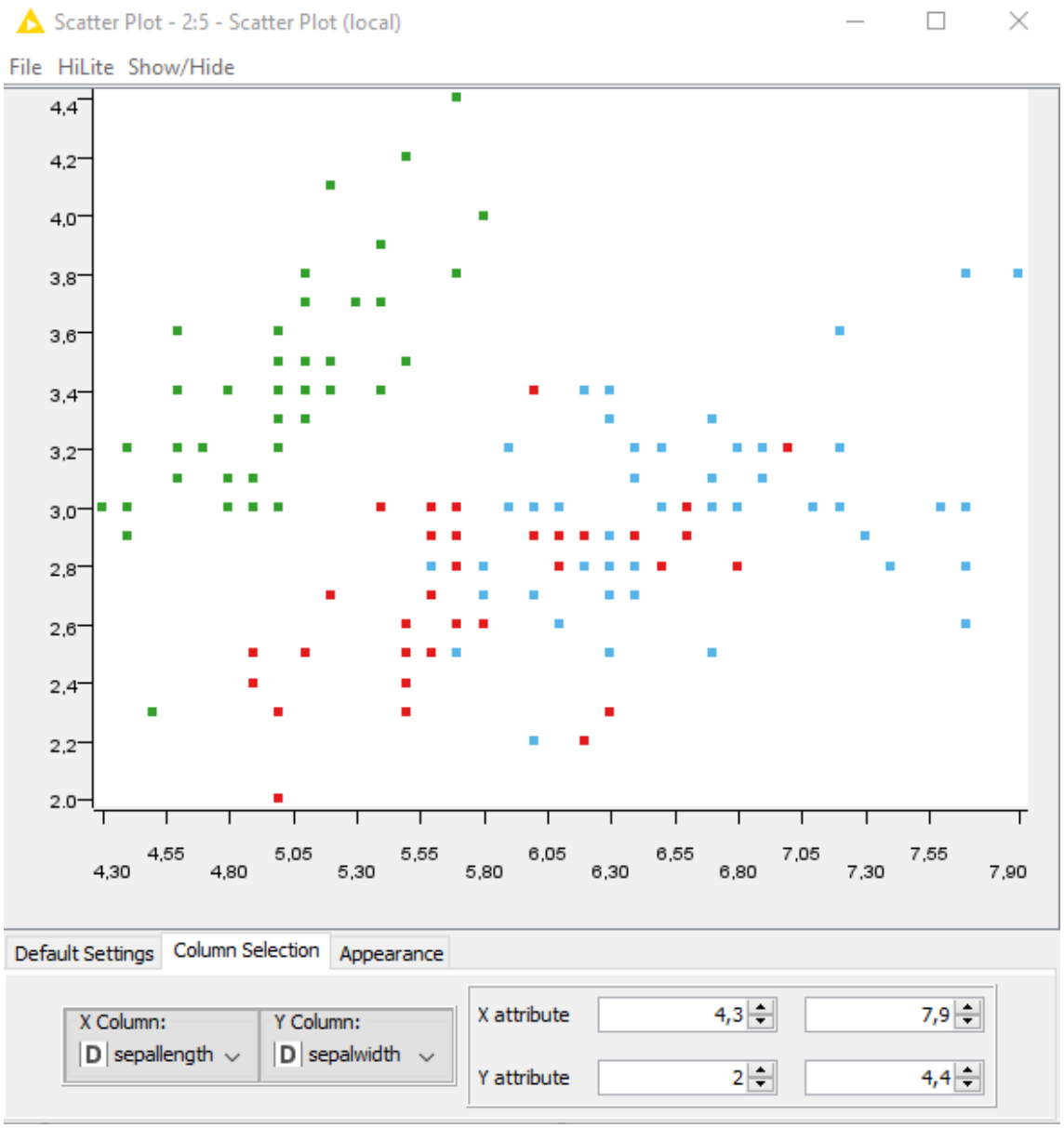
5.

En esta ocasión vamos a añadir los atributos “sepallength” y “sepalwidth”. Recordemos que estos atributos los habíamos filtrado anteriormente mediante el nodo “Column Filter”, por lo que para añadirlo de nuevo seleccionamos dicho nodo y pulsamos F6 para configurarlo. En esta pantalla podremos volver a incluir los atributos:



Una vez hecho esto pulsamos aceptar. Tendremos que volver a ejecutar los nodos, ya que ha cambiado la información. Una vez ejecutados todos, iremos al nodo “Scatter Plot” y haciendo click derecho elegiremos “View: Scatter Plot”.

En esta ocasión pondremos en el eje x el atributo “sepallength” y en el eje y el atributo “sepalwidth”. De esta manera obtenemos el siguiente diagrama de dispersión:



Podemos observar que en este caso no es un buen cluster ya que no están bien diferenciados entre ellos, si no que los puntos pertenecientes a dos clusters distintos se entremezclan.