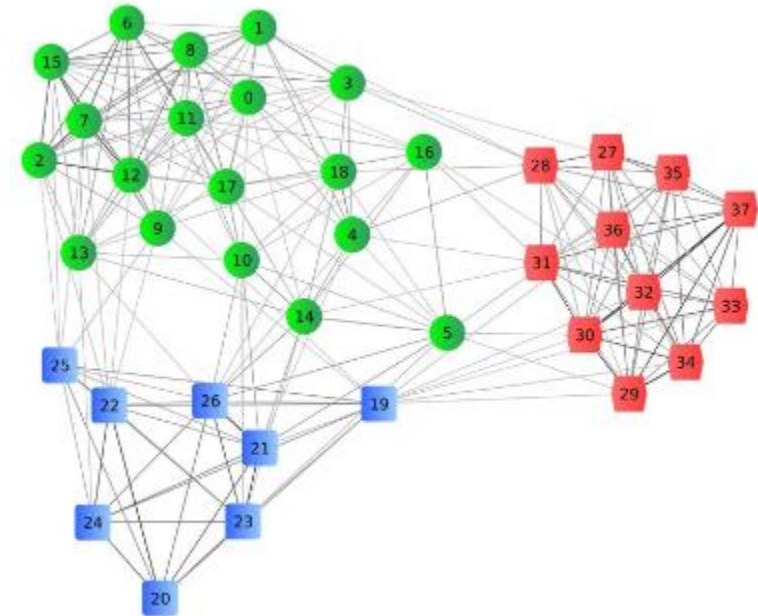




MÉTODOS PARA LA VALIDACIÓN DE CLUSTERING

Introducción

- Características de las técnicas de clustering
 - Operan de manera no supervisada.
 - Son sensibles a los parámetros de entrada
- Para evaluar el resultado del clustering se utilizan métricas de validación



Métricas de validación

- Como el objetivo del clustering es agrupar objetos similares en el mismo cluster y objetos diferentes en distintos clusters, las métricas de validación están basadas usualmente en los siguientes criterios:
 - *Cohesión*
 - *Separación*

Criterios de las métricas de validación

■ Cohesión

- *Se busca que los miembros de un mismo grupo se encuentren lo más cerca que sea posible unos de otros.*

■ Separación

- *Los clusters deben estar ampliamente separados entre ellos. Existen varios enfoques para medir esta distancia entre clusters: distancia entre los miembros más cercanos, distancia entre los miembros más distantes o la distancia entre los centroides.*

Métrica SSW (sum-of-squares within)

- Se usa para evaluar la Cohesión (distancia intra-cluster) de los clusters generados por el algoritmo de agrupamiento.

$$SSW = \sum_{i=1}^k \sum_{x \in C_i} dist^2(m_i, x)$$

- Siendo k el número de clusters, x un elemento del cluster C_i y m_i el centroide del cluster C_i

Métrica SSB (sum-of-squares between)

- Es una medida de separación utilizada para evaluar la distancia inter-cluster (Separación)

$$SSB = \sum_{j=1}^k n_j \text{dist}^2(m_j - \bar{x})$$

- Siendo k el número de clusters, n_j el número de elementos en el cluster C_j , m_j el centroide del cluster C_j y \bar{x} es la media del conjunto de datos completo.

Indice Silhouette

Dado un ejemplo x del conjunto de datos :

- **Cohesión $a(x)$** : distancia promedio de x a todos los demás ejemplos en el mismo cluster.



- **Separación $b(x)$** : distancia promedio de x a todos los demás ejemplos en el cluster más cercano.



Indice Silhouette de un elemento

- El índice silhouette para el ejemplo x está definido como

$$s(x) = \frac{b(x) - a(x)}{\max\{b(x), a(x)\}}$$

- El valor de $s(x)$ puede variar entre -1 y 1.

-1 = mal agrupamiento

0 = indiferente

1 = bueno

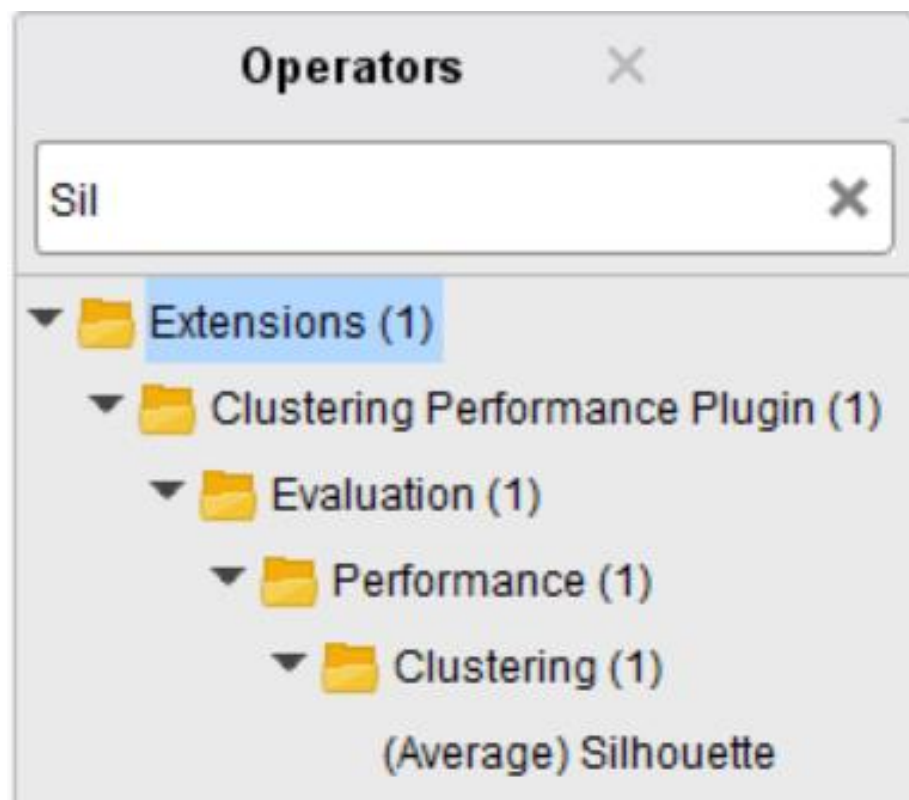
Indice Silhouette del agrupamiento

- El índice Silhouette para todo el agrupamiento es

$$SC = \frac{1}{N} \sum_{i=1}^N s(x_i)$$

- Será mejor cuanto mayor sea el valor del índice.

Silhouette en RapidMiner



Instale el operador Silhouette copiando el archivo

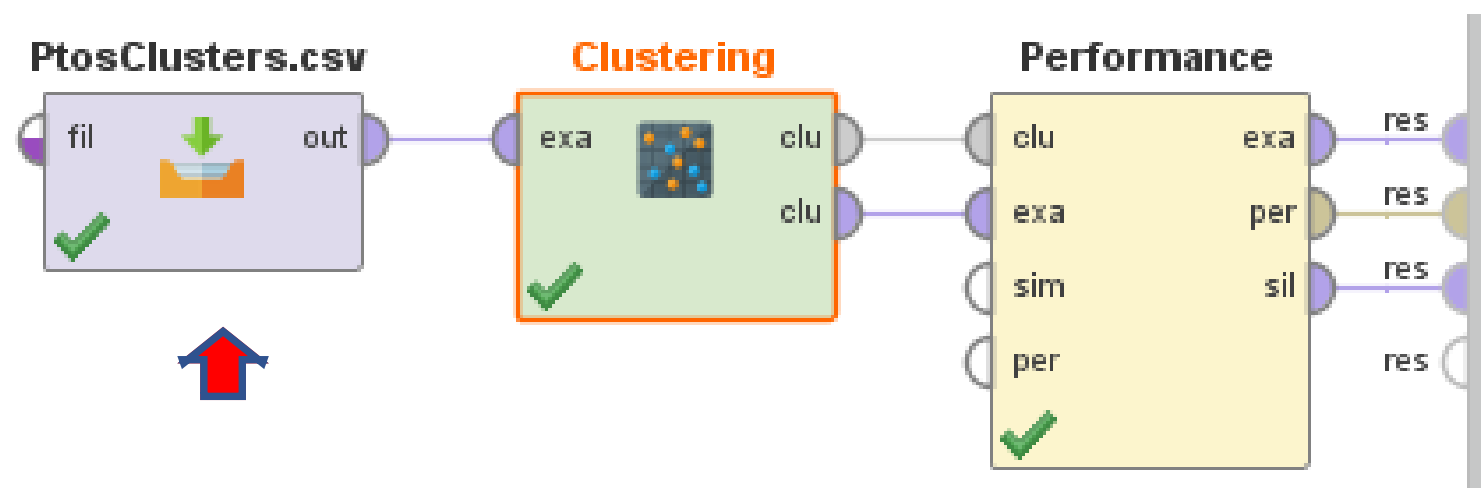
CPPlugin-0.3.jar

*en el directorio **lib/plugins** dentro del directorio de instalación de Rapidminer*

o en el directorio

C:\Users\Alumnos\.RapidMiner\extensions

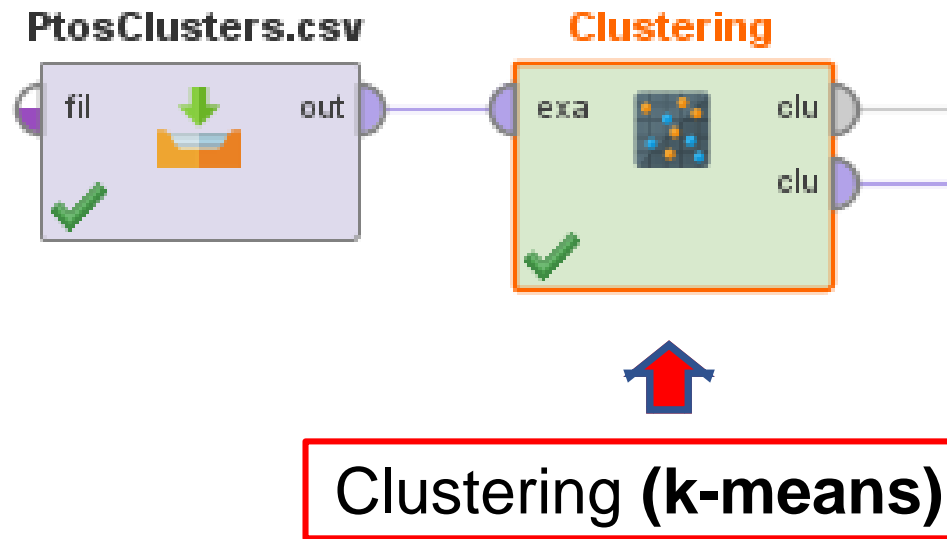
Silhouette en RapidMiner



attribute meta data information				
X1	<input type="checkbox"/>	column...	real	attribute
X2	<input checked="" type="checkbox"/>	column...	real	attribute
Clase	<input checked="" type="checkbox"/>	column...	polyno...	label

El agrupamiento se realizará sólo el atributo X2

Silhouette en RapidMiner



Parameters

Clustering (k-Means)

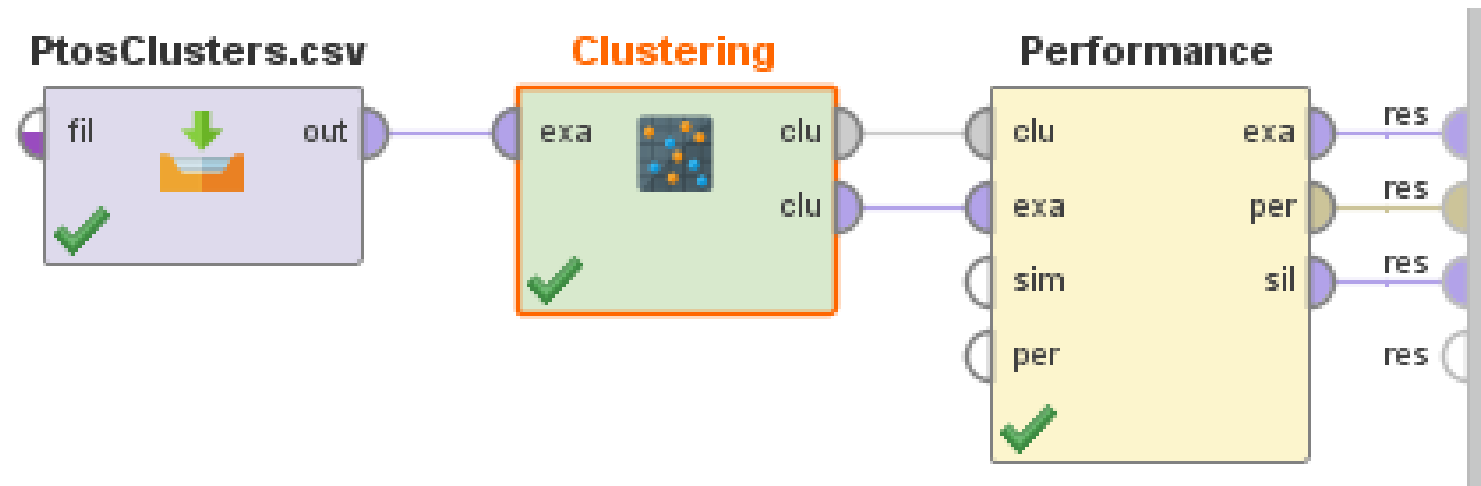
- ☒ add cluster attribute
- ☐ add as label
- ☐ remove unlabeled

k 5

max runs 10

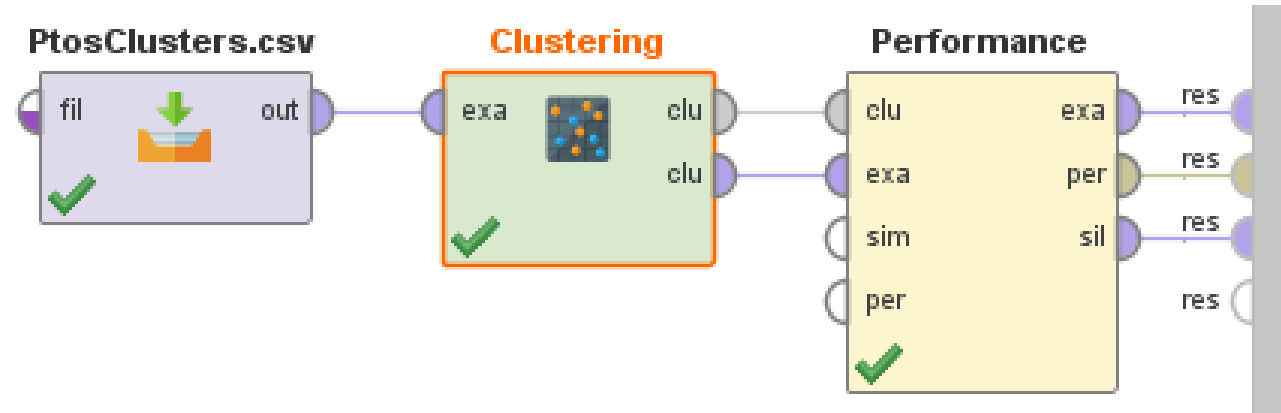
- ☒ determine good start values

Silhouette en RapidMiner

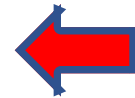


**Performance (Average)
Silhouette**

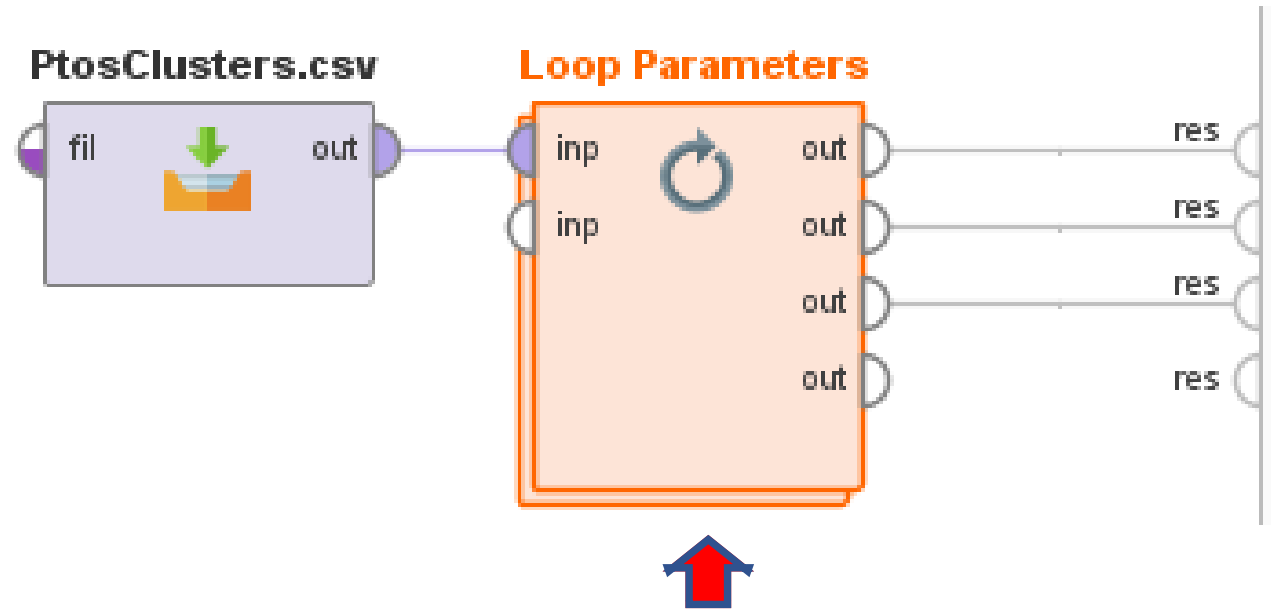
Ejercicio



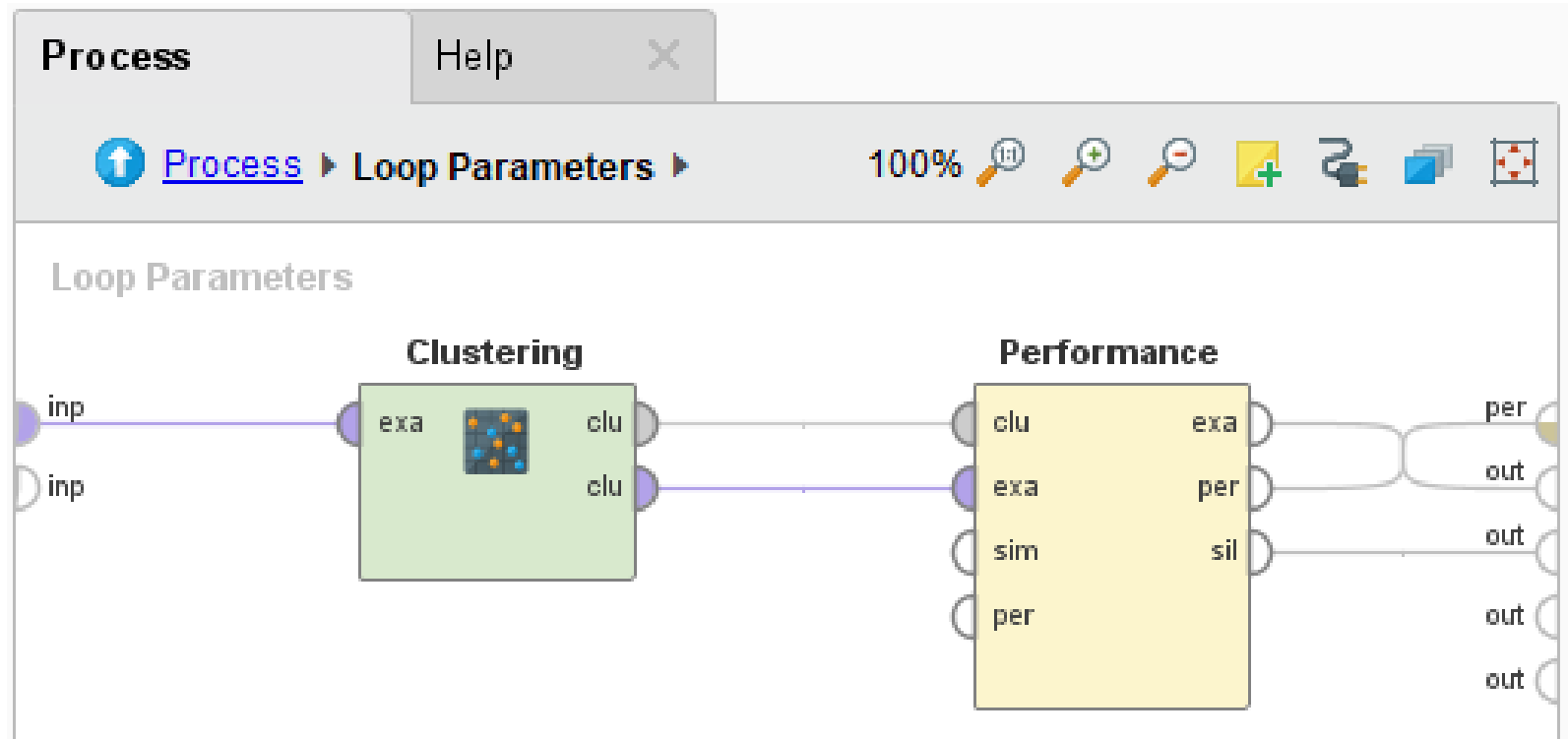
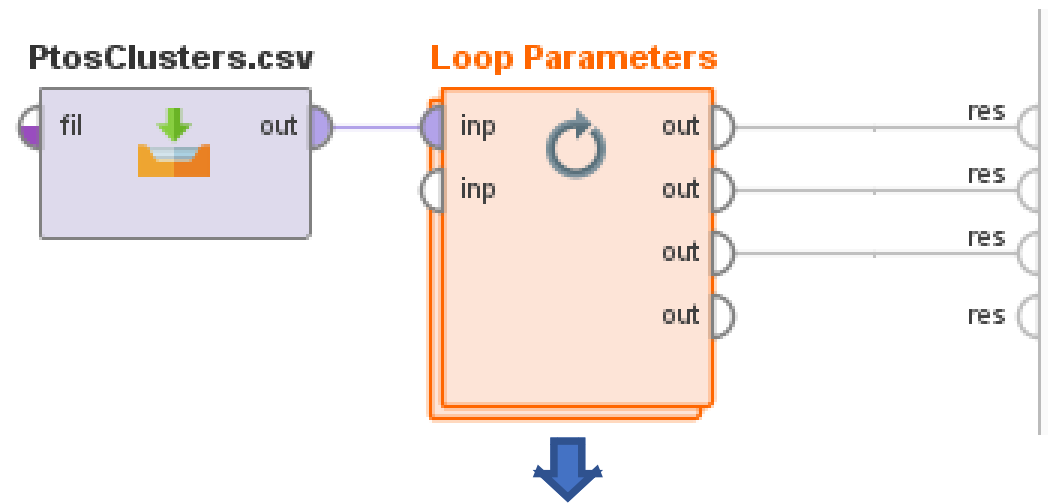
Cantidad de grupos (Valor de k)	Indice Silhouette
2	0.681
3	0.603
4	0.559
5	0.524



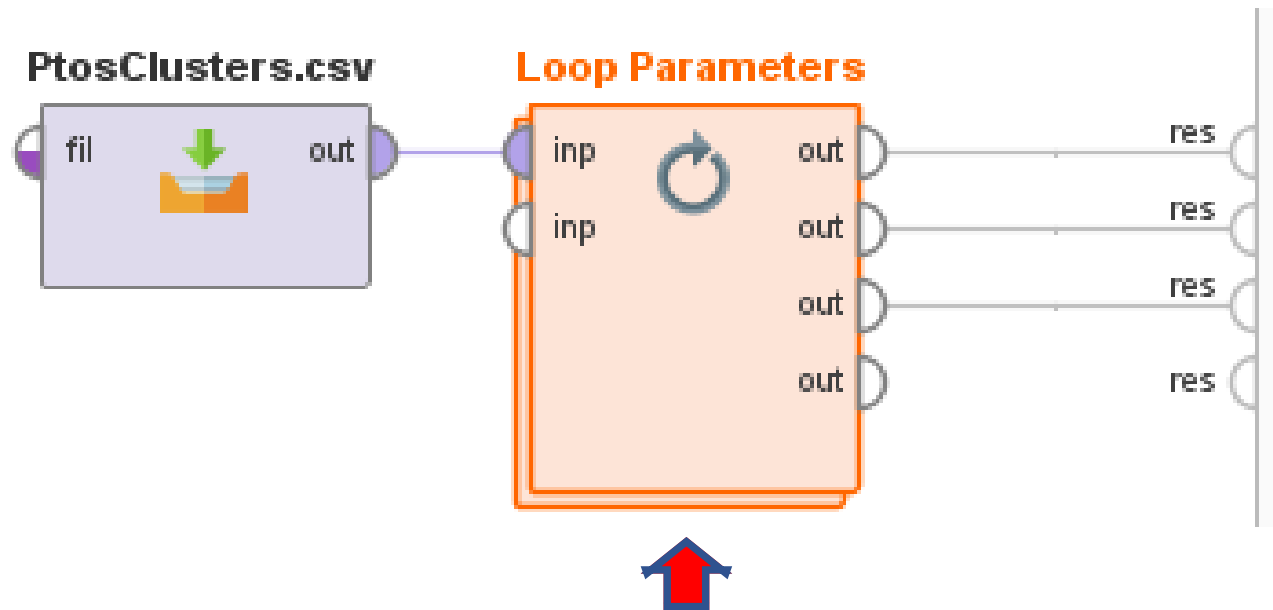
Operador *Loop Parameters*




- Este Operador itera sobre sus subprocesos para todas las combinaciones de parámetros definidos




Operator *Loop Parameters*



Parameters ✕

 **Loop Parameters**

 **Edit Parameter Settings...**

error handling fail on error

☒ log performance

☐ log all criteria

☐ synchronize

☒ enable parallel execution

Indice Davies-Bouldin

- Este índice compara a cada cluster con su vecino más cercano.
- Para medir la calidad del agrupamiento tiene en cuenta los siguientes dos conceptos
 - *dispersión de cada cluster.*
 - *distancia entre clusters.*

Indice Davies-Bouldin - Dispersión

- La **dispersión** S_i del cluster C_i se define como la distancia promedio entre los ejemplos que pertenecen al cluster y el centro del mismo.

$$S_i = \frac{1}{|C_i|} \sum_{x \in C_i} dist(x, m_i)$$



donde $|C_i|$ es la cantidad de elementos que pertenecen al cluster C_i , x es un elemento y m_i es el centroide de dicho cluster.

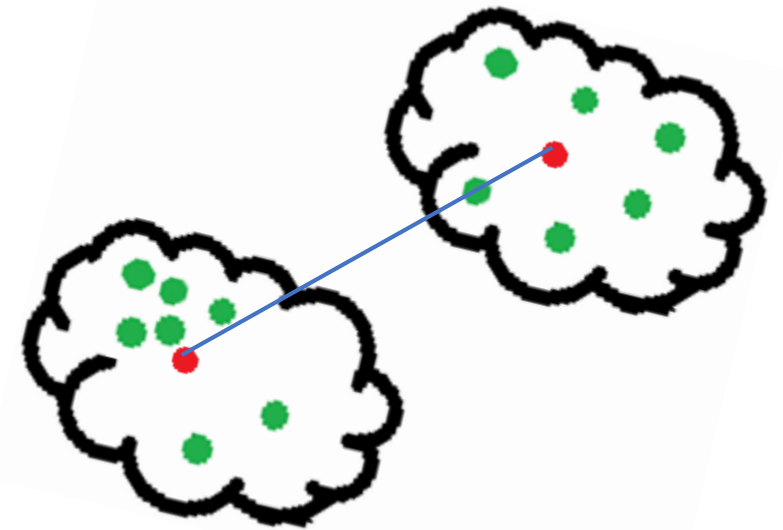
- Cuanto más cerca del centro estén los ejemplos, menor será la dispersión del cluster.

Indice Davies-Bouldin - Distancia

- La **distancia** entre los centros de los clusters C_i y C_j se calcula de la siguiente forma

$$D_{ij} = \text{dist}(m_i, m_j)$$

donde m_i y m_j son los centroides de los clusters C_i y C_j respectivamente.



Indice Davies-Bouldin – Comparación

- Dados dos clusters C_i y C_j , para calificar la manera en que los elementos quedan distribuidos se define la siguiente expresión

$$R_{ij} = \frac{S_i + S_j}{D_{ij}}$$

- Mientras más compactos sean los clusters C_i y C_j **menores** serán S_i y S_j , y mientras más separados estén, **mayor** será D_{ij} .
- Por lo tanto, cuanto más compactos sean los grupos y más separados se encuentren, **menor** será el valor de R_{ij}

Indice Davies-Bouldin

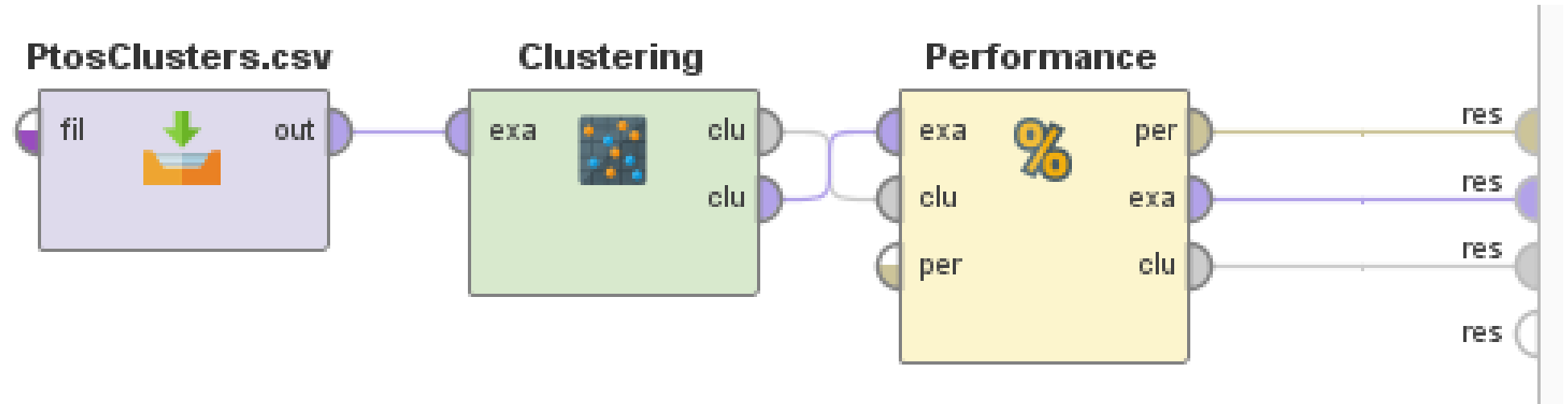
- El índice Davies-Bouldin se define como

$$DB = \frac{1}{k} \sum_{i=1}^k \max_{i \neq j} (R_{ij})$$

donde

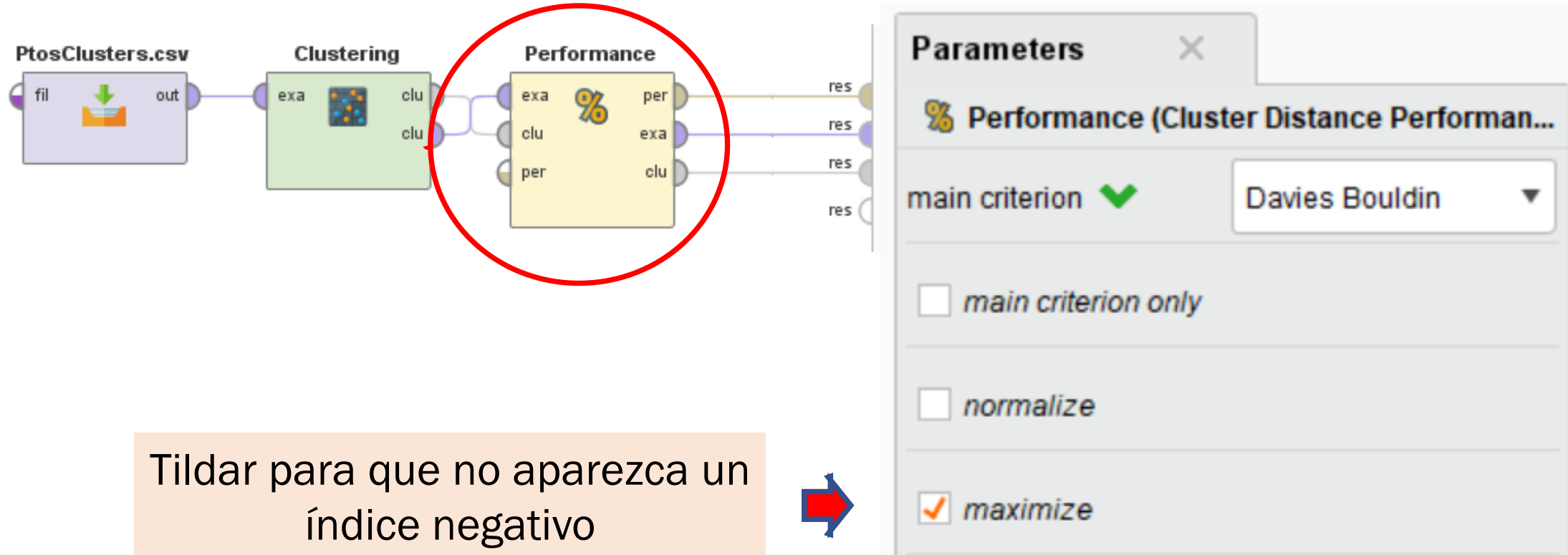
- k es el número de clusters
 - $\max_{i \neq j} (R_{ij})$ es el “peor caso” para el cluster C_i
- Será mejor cuanto menor sea el valor del índice Davies-Bouldin.
 - El valor del índice no está acotado. Puede tomar un valor arbitrario.

Davies-Bouldin en RapidMiner

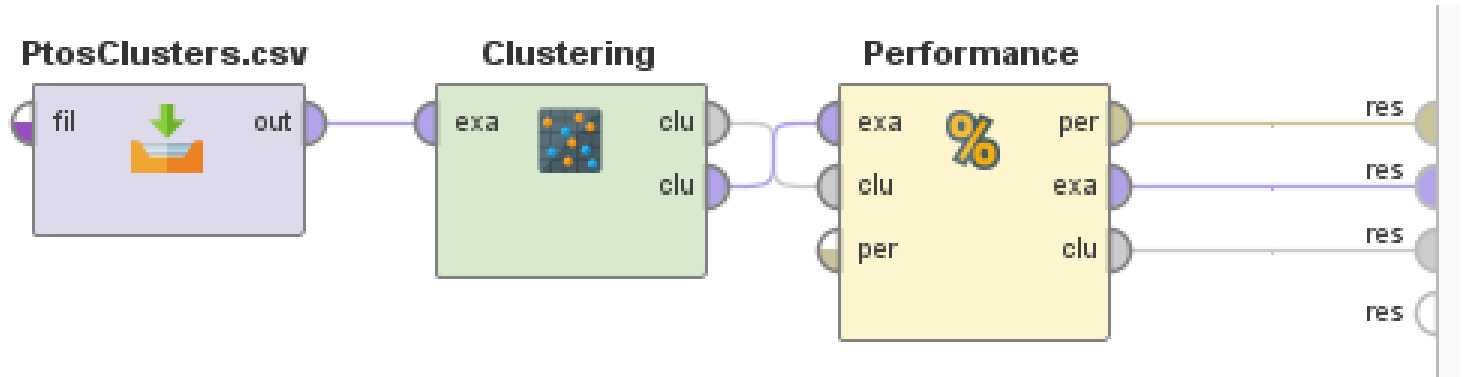


Operador **Cluster Distance**
Performance

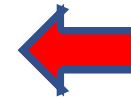
Operador Cluster Distance Performance



Ejercicio



Cantidad de grupos (Valor de k)	Indice Davies-Bouldin
2	0.418
3	0.500
4	0.556
5	0.581



Ejercicio



- Se busca agrupar los ejemplos del archivo **IRIS.CSV**
- Utilice los índices Silhouette y Davies-Bouldin para analizar los agrupamientos obtenidos con k-medias para $K=2$, $k=3$, $k=4$ y $k=5$.
- Realice un gráfico de coordenadas paralelas y analice los grupos obtenidos.

Ejercicio

- El archivo **SEMILLAS.csv** contiene información de granos que pertenecen a tres variedades diferentes de trigo: Kama, Rosa y Canadiense.
- Para cada grano se midieron las siguientes características:
 - *área A,*
 - *perímetro P,*
 - *compacidad $C = 4 * \pi * A / P^2$,*
 - *longitud del núcleo,*
 - *ancho del núcleo,*
 - *coeficiente de asimetría*
 - *longitud del surco del núcleo*
- Describa los tipos de semillas inspeccionados utilizando una técnica de clustering.