Ejercicio 2 - DBSCAN

PASO 1 DATA PREP

1. Insertar el dataset “Iris” en un nuevo proceso
2. Solamente utilizaremos dos de los atributos: A3 (petal length) y A4 (petal width) para poder visualizar mejor los clusters y comprender el modelo. Agregar un operador entonces para filtrar sólo estos atributos

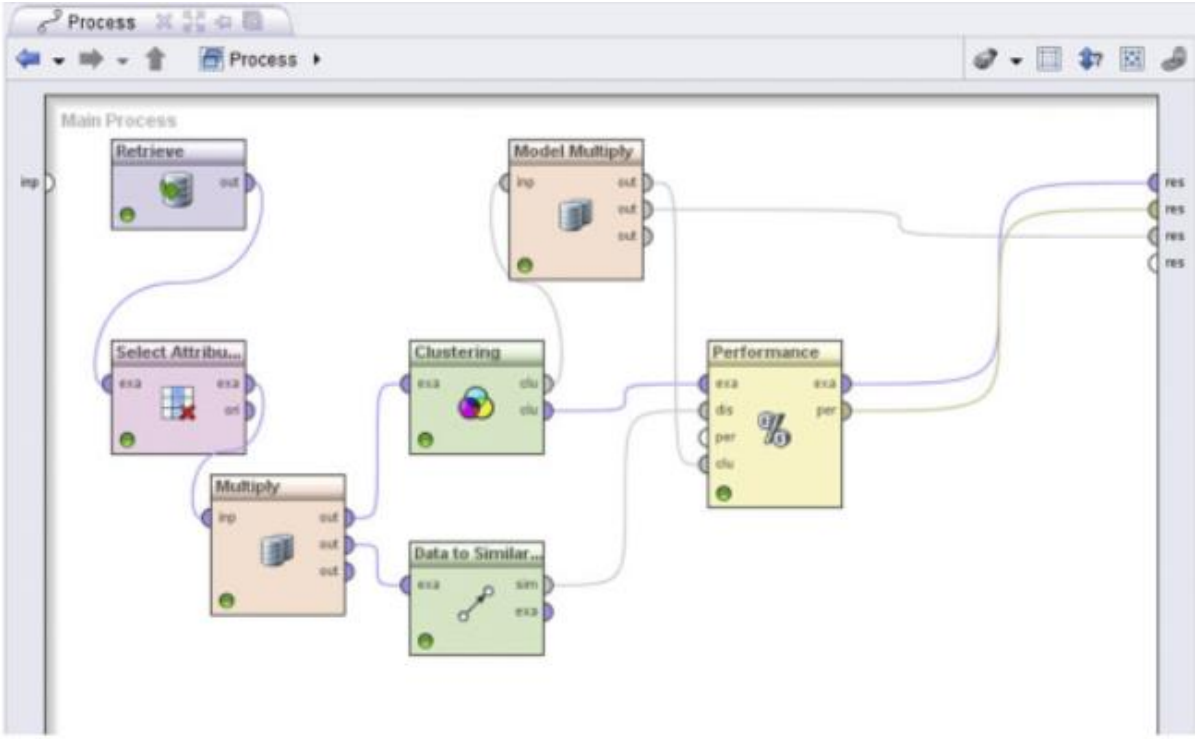
PASO 2 – OPERADOR Y PARÁMETROS

3. Agregara un operador DBSCAN
4. Parámetros a configurar:
    a. Epsilon – tamaño del grupo de alta densidad, por defecto 1
    b. MinPoints – cantidad mínima de ejemplos dentro del grupo de épsilon para configurar un cluster
    c. Medida de distancia. Analizar y documentar las medidas disponibles. ¿En qué casos o tipos de problemas conviene aplicar cada una? Haz un breve resumen y reporta en la tarea correspondiente.
    d. “Add cluster as atributes” – recomendado para el análisis posterior

PASO 3 – EVALUACION

Al igual que en k-means, podemos evaluar la efectividad de los grupos de clustering utilizando la media de las distancias dentro de los clusters

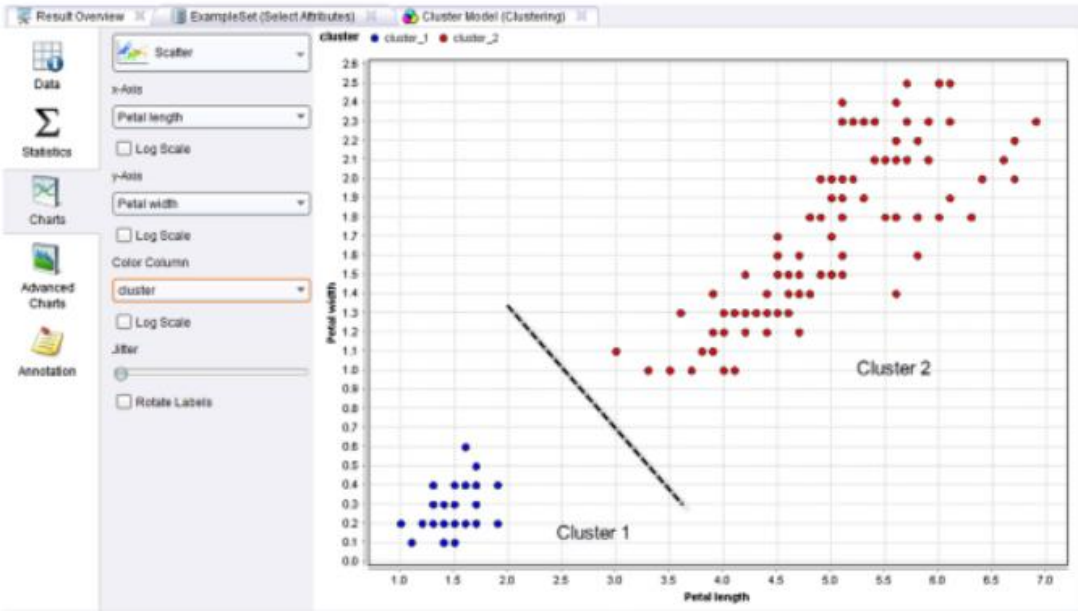
- Agregar un operador “Cluster density performance”. Analizar los parámetros disponibles.
- El operador de performance también espera un operador “Similartiy Measure” para facilitar los cálculos. Un vector de medida de similaridad es una medida de distancia de cada ejemplo con respecto al otro ejemplo. La medida de similaridad puede ser calculada utilizando un operador “Data to similarity” sobre el dataset de ejemplo.



PASO 4 – EJECUCION E INTERPRETACIÓN

Después de conectar las salidas del operador de performance a los puertos de resultados, se puede ejecutar el modelo, y se pueden observar los siguientes resultados:

- **Model**: la salida del modelo de cluster. Observa y nota que contiene
  - Información sobre la cantidad de clusters encontrados en el dataset
  - Objetos de datos identificados como puntos de ruido (cluster 0). Si no se encuentran puntos de ruido, entonces el cluster 0 estará vacío.
  - Utilizando el Folder View y el Graph view, visualizar estos clusters y su contenido (ejemplos correspondientes)
- **Clustered example set**: el dataset de ejemplo ahora tiene otro atributo: la etiqueta de clustering, que puede ser usada para ulterior análisis y visualización. Hacer una vista de scatterplot para este dataset , configurar los ejes x e y con los atributos originales (petal length y petal width). Configurar "Color Column" con el nuevo atributo de etiqueta de cluster. En el gráfico vemos cómo el algoritmo encontró dos clusters en el dataset.



Los objetos de datos correspondientes a la especie *setosa* tienen áreas de alta densidad bien diferenciadas. Sin embargo, no hay un área de baja densidad bien definida para particionar los grupos de *versicolor* y *virginica*. Por ello estos dos clusters naturales aparecen combinados en un nuevo cluster artificial.

Los parámetros Epsilon y MinPoints pueden ser ajustados para encontrar diferentes resultados de clustering.

- **Vector de performance**. La pestaña del vector de performance muestra la distancia media dentro de cada cluster y la media de todos los clusters. La distancia media es la distancia entre todos los puntos de datos dividida entre la cantidad de puntos de datos. Utilizando estas medidas, evalúa diferentes ejecuciones del modelo configurando diferentes valores para los parámetros básicos. Compila una tabla comparativa de resultados y remítela junto con el proceso completo a la tarea correspondiente del ejercicio.