UNIDAD TEMÁTICA 5: Aprendizaje No Supervisado, Clustering

Trabajo de Aplicación 1 – k-means

ESCENARIO

La Dra. Martínez es directora de programa para una gran empresa de seguros médicos. Recientemente ha visto una cantidad de investigaciones que ponen gran énfasis en la influencia del peso, sexo y colesterol en el desarrollo de enfermedad coronaria.

A efectos de decidir qué tipos de medidas preventivas recomendar a sus pacientes, piensa si no habrá grupos naturales de individuos que tengan más riesgo por alto colesterol o peso, y en caso de que estos grupos existan, dónde estarían las fronteras que los separan.

Datos

La Dra. Martínez ha obtenido una base de datos que contiene información de historias clínicas de 547 pacientes, de la cual ha extraído los tres atributos que está considerando.

Los atributos del dataset son:

- **Peso**: el peso en libras del paciente
- **Colesterol** : último nivel de colesterol registrado de la persona
- **Sexo:** 0 para femenino, 1 para masculino

Utilizaremos este dataset para construir un modelo de clustering que permita comprender cómo los pacientes se pueden agrupar en base a su peso, sexo y niveles de colesterol.

Recordemos que las medias son particularmente susceptibles a la influencia indebida de outliers extremos, por lo que es importante analizar la existencia de datos inconsistentes al utilizar la técnica de k-means.

Ejercicio 1

Preparación de los datos

1. Importa en RM el dataset de entrenamiento ("***k-means-cardio.csv***").
   a) Verifica que la primera fila se configura como nombres de los atributos.
   b) Revisa los nombres y tipos de los atributos.

2. Crea un nuevo proceso en blanco y arrastra el dataset al mismo. Conéctalo a la salida
   a. Ejecuta el proceso y analiza los datos
   b. Analiza las estadísticas de los atributos
   c. Analiza si hay inconsistencias en los datos que puedan afectar al algoritmo

Modelado

1. Agrega un operador "k-Means" al proceso, y conecta sus puertos al dataset y a la salida.
2. La Dra. Martínez, como habíamos indicado al principio, ya ha visualizado que podría haber al menos 4 grupos potencialmente diferentes. Por lo tanto, configura el operador para que k=4.
3. Puedes cambiar el valor de "max runs" si lo deseas, aunque el valor por defecto debería ser apropiado en primera instancia.

4. Ejecuta el modelo
5. Observa el reporte inicial que indica la cantidad de ejemplos que caen en cada grupo
6. Crea diferentes gráficos para visualizar la información resultante. Captura estos gráficos y repórtalos en la tarea correspondiente para este ejercicio.

Evaluación

Utilizando el operador k-means de RapidMiner hemos identificado 4 clusters para la Dra. Martínez, y podemos ahora evaluar su utilidad para responder la pregunta de la Dra.

En la ventana de resultados del modelo, vemos varias opciones para analizar los clusters. Observa la tabla de centroides, y los valores que allí aparecen.

Vemos que el cluster 3 presenta los mayores promedios para peso y colesterol.

Siendo que, para el atributo sexo, 0 indica femenino y 1 masculino, el valor de 0.591 en el cluster 3 indica que hay más hombres que mujeres representados en este cluster.

Sabiendo que el alto colesterol y peso son dos claros indicadores de riesgo de enfermedad coronaria, la Dra. Martínez podría comenzar con los integrantes del cluster 3 para aplicar sus nuevos programas.

Ahora bien, ¿cuáles son los pacientes que integran este cluster 3? Selecciona la vista “Folder View” y observa la información disponible.

Despliegue

1. A efectos de seleccionar solamente los datos de los pacientes que resultan en el cluster 3, agrega un operador “filter examples” al proceso, luego del operador “k-nn”. Configura el parámetro “string =cluster_3”.
2. Ejecuta el modelo y observa los resultados.
3. Con esta información la Dra. Martínez podría ya realizar consultas directas sobre la base de datos de la mutualista, por ejemplo,

```
SELECT First_Name, Last_Name, Policy_Num, Address, Phone_Num
FROM PolicyHolders_view
WHERE Weight >= 167
AND Cholesterol >= 204;
```

Este query le daría los datos de todos los pacientes incluidos en el cluster 3

¿cómo serían las consultas SQL para los otros 3 clusters?

(reporta todos los resultados, proceso y queries en la tarea correspondiente)

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

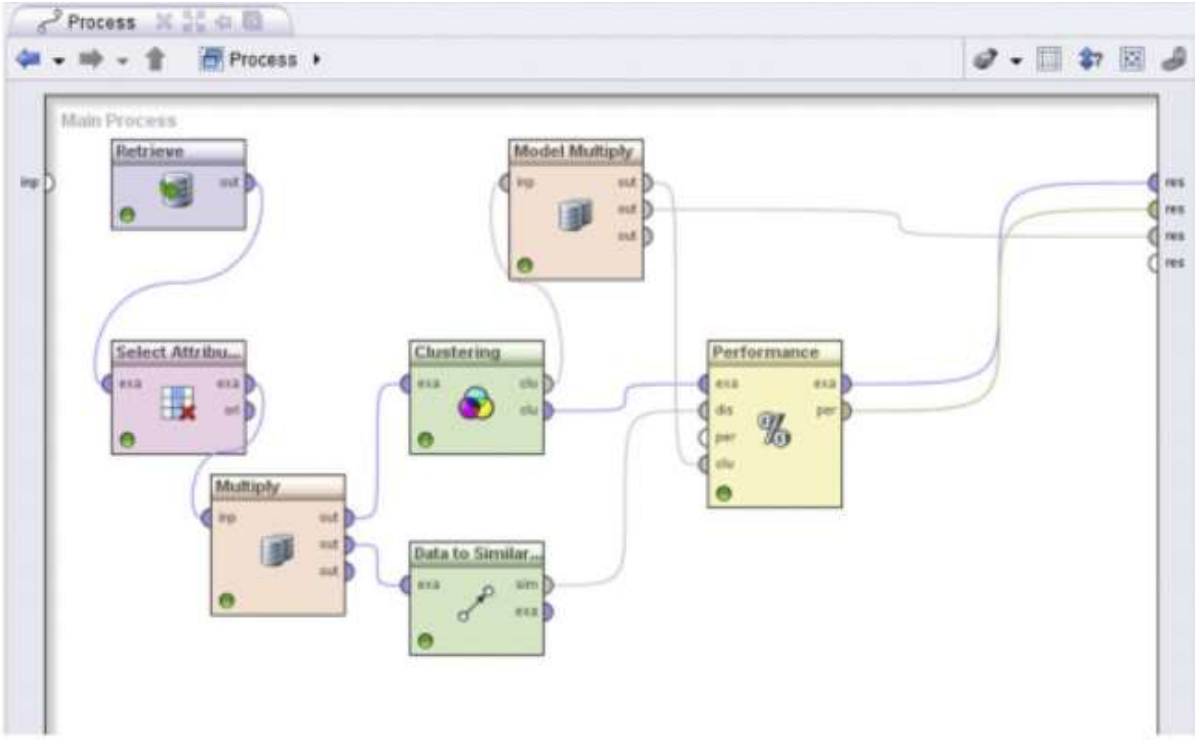
PASO 2 – OPERADOR Y PARÁMETROS

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

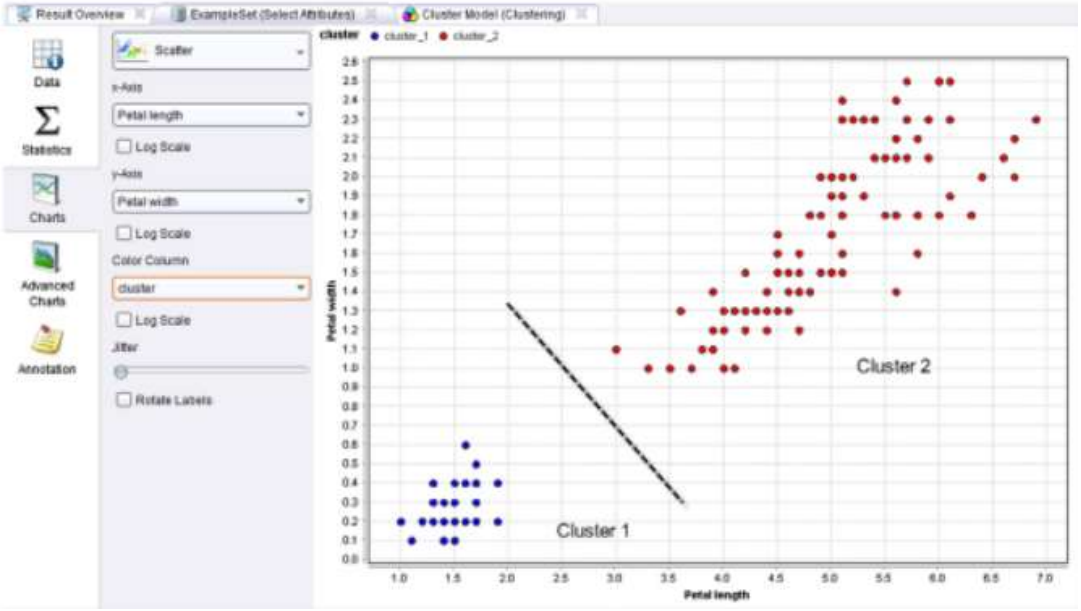
- Agregar un operador “Cluster density performance”. Analizar los parámetros disponibles.
- El operador de performance también espera un operador “Similartiy Measure” para facilitar los cálculos. Un vector de medida de similaridad es una medida de distancia de cada ejemplo con respecto al otro ejemplo. 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.