
UNIVERSIDAD DON BOSCO
VITAM IMPENDERE VERO

10 – Minería de Datos – Agrupamiento con K-means.

Objetivos:

- Genera reportes a partir de técnicas de Data Mining.
- Utiliza software de Data Mining.

Introducción

En la actualidad uno de los usos más importantes de las bases de datos es en la aplicación del uso de técnicas de minería de datos, con los cuales cualquier empresa o institución puede obtener resultados importantes para la toma de decisiones.

Como este es un tema muy amplio y complejo, lo que veremos es el uso de una aplicación RapidMiner, el cual proporciona un entorno muy bueno de pruebas y además tiene una versión de código abierto.

Ejemplo de Aplicación

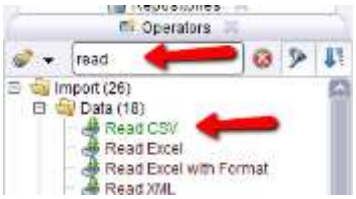
Realizaremos un ejemplo basado en local de renta video, analizando el tipo de películas que se rentan, la edad, ingresos y número de veces que se visita el establecimiento.

Para este proceso usaremos el archivo csv que se llama video.csv

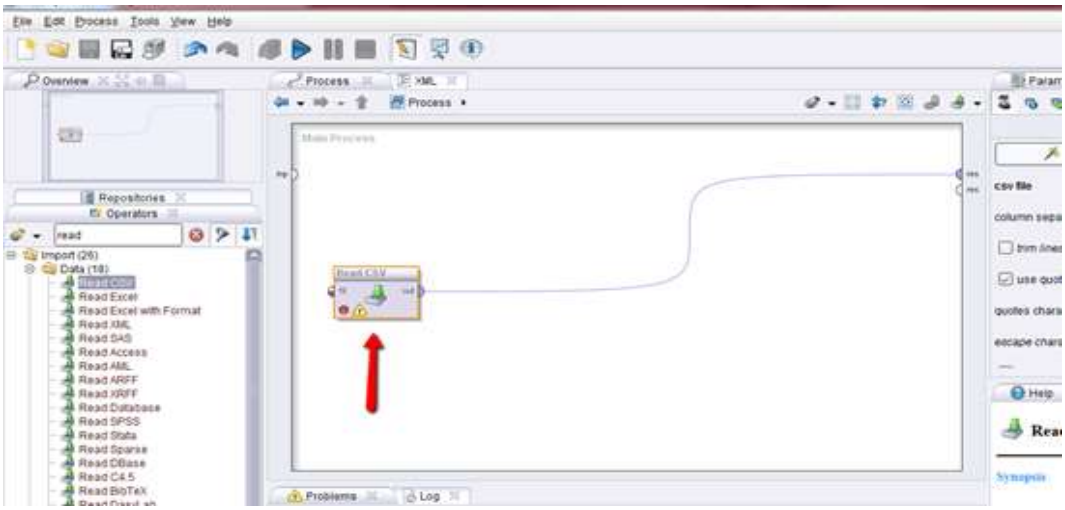
Al seleccionar esta opción la aplicación nos presentara un lienzo en el que podremos trabajar en base a una data, y a la unión de operadores, la aplicación aparecerá como la siguiente figura.

OPERADOR 1 – Read CSV

Ahora vamos seleccionar nuestro primer operador el cual será ***"Read CSV"***, para localizar este operador entre los más de 300 operadores, usaremos el buscador de operadores y digitaremos ***"read"*** y cual nos mostrara todas las coincidencias.



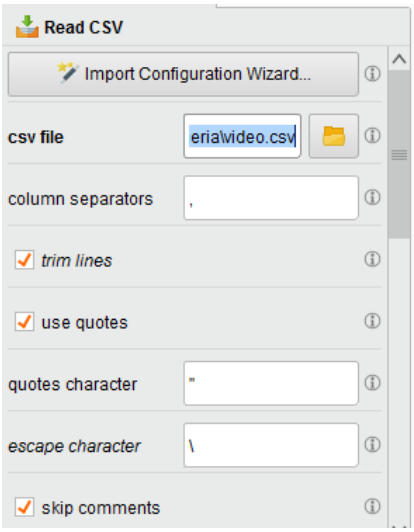
Para ocupar el operador tenemos dos formas de hacerlo, ya sea tomando y arrastrando el operador hacia el lienzo o dándole doble clic.



Una vez colocado el operador, lo primeros que haremos es borrar la línea que conecta a nuestro operador con el lado derecho del lienzo, para hacer esto, seleccionamos la línea y presionamos suprimir, ahora vamos a usar los datos de prueba, y ocuparemos la parte derecha de la aplicación que donde esta los parámetros de cada operador, y seleccionaremos la opción ***"Import Configuration Wizard"*** como lo muestra la siguiente figura.

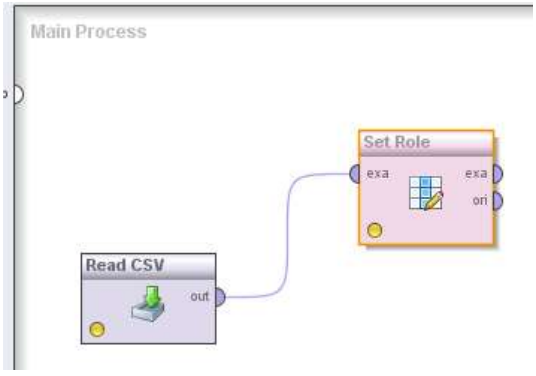
Recuerde en que carpeta descomprimió el archivos de datos, pues tendrá que llamar al archivo video.csv, realizare el mismo proceso que hicimos para las reglas de asociación.

Una vez importado los datos, me aparecerá en el lienzo el operador, solamente con un circulo amarrillo y debo de configurar la parte derecha con los datos tal y como aparecen en la siguiente figura.

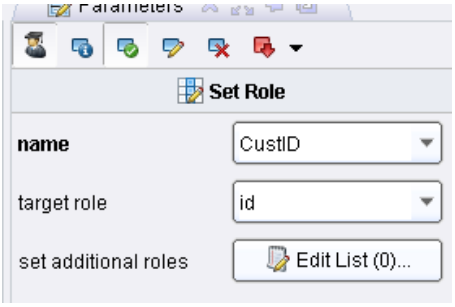


OPERADOR 2 – Set Role

Realizo el mismo proceso con el operador 1, lo busco y al encontrarlo, presiono doble clic o lo arrastro al lienzo, este proceso dejara unidos los dos operadores con una línea, y el modelo se verá como el siguiente, es importante conectar los dos cubos para realizar la configuración correcta del cubo Set Role.

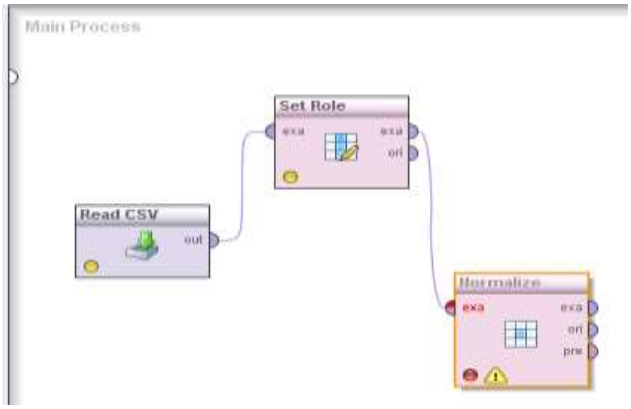


La forma en que lo configuraremos es en name ***CustID*** y en target role ***id***, los demás parámetros los dejaremos como están.

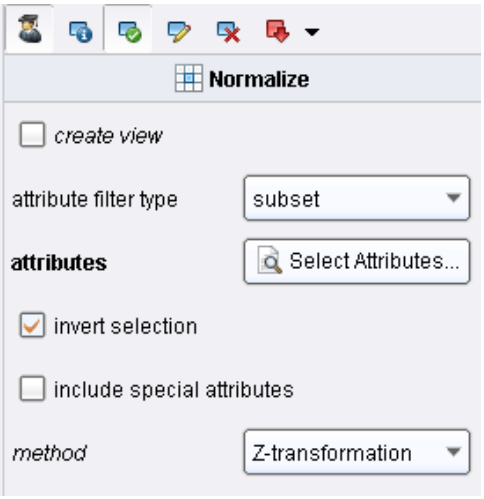


OPERADOR 3 – Normalize

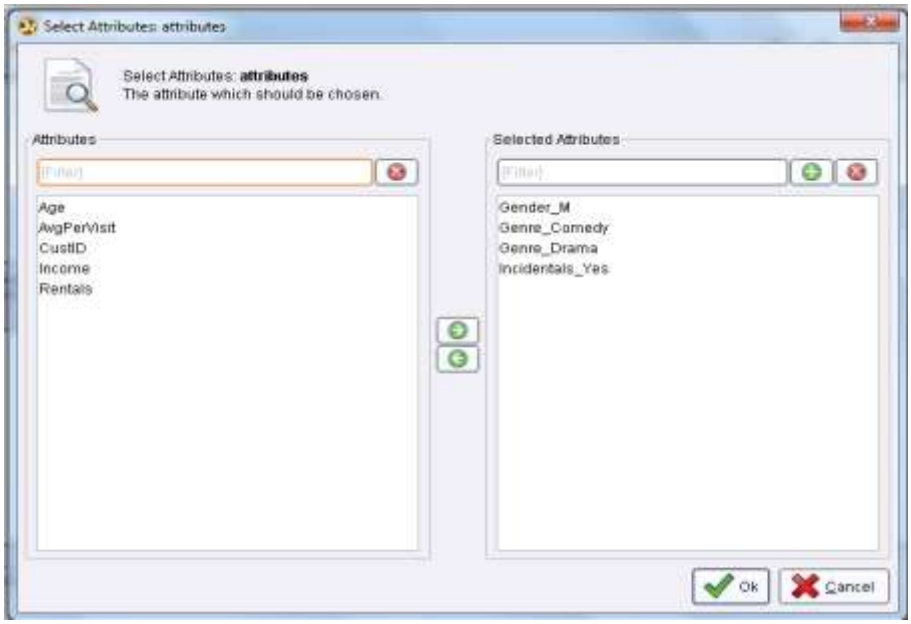
El siguiente cubo es ***Normalize***, la buscamos y luego la incorporamos al lienzo.



Ahora para la configurar el cubo, usaremos los siguientes parámetros para attribute filter type: ***subset***, y después configuraremos la parte que dice ***Select Attributes.***

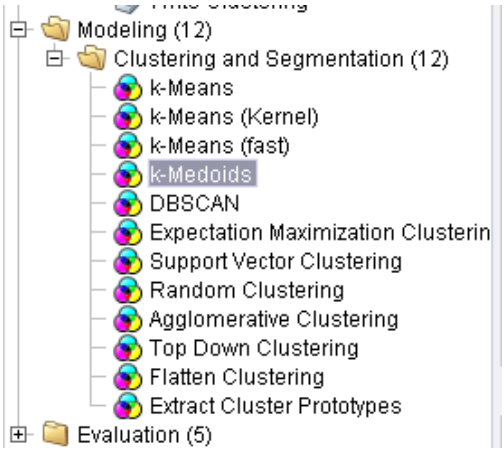


La selección de atributos será de la siguiente manera, es importante que se realiza de estar forma para los resultados esperados.

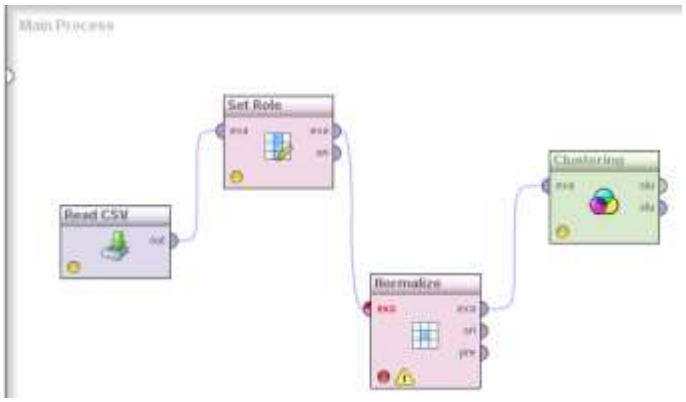


OPERADOR 4 – Clustering (k-Means)

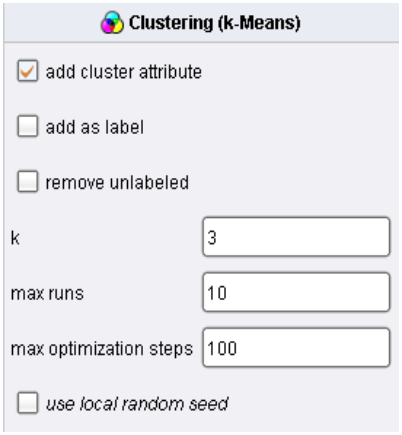
En este operador se define los cluster o grupos que deseamos hacer, tambien es dificil de encontrar, deben buscarlo en el rubro ***Modeling***, despues en sub rubro ***Clustering and Segmentation***, y por ultimo seleccionamos ***k-Medoids***, como lo muestra la siguiente figura.



Una vez seleccionado el cubo, el lienzo queda configurado de la siguiente manera.

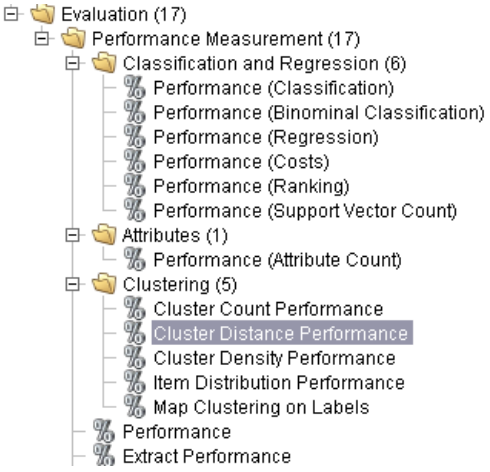


Ahora sigue la configuración del cubo de clusterin (k-Means), seleccionamos la opción: ***add cluster attribute***, y en la opción que dice ***k***, es donde seleccionamos la cantidad de grupos que deseamos crear, y también la opción más ***runs,*** la cual serían las iteraciones que deseamos que se realicen.

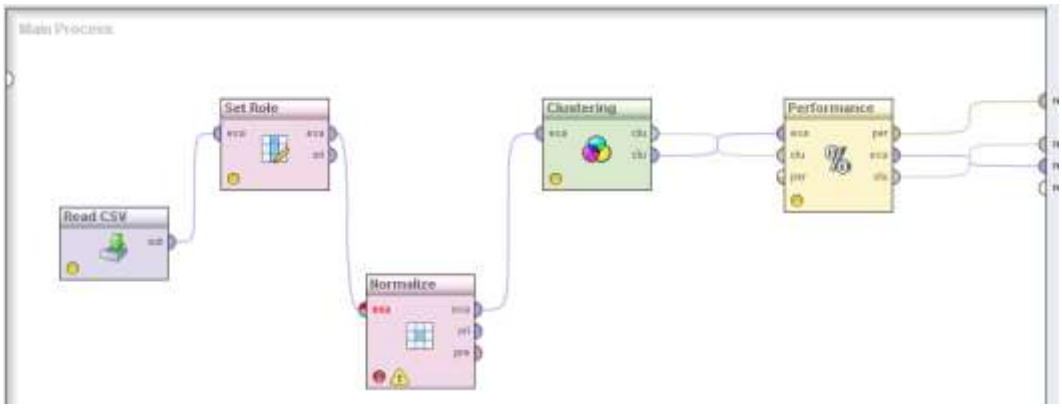


OPERADOR 5 – Performance

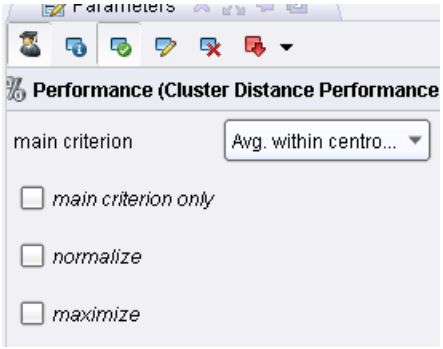
El quinto cuadro que utilizaremos sera el de ***Performance***, este es dificil de encontrar, y debe ser buscado en el rubro ***Clustering***, y despues ***Cluster Distance Performance.***



Una vez incorporado es cubo es importante enlazar sus tres salidas con la pared del lienzo, y es importante ver la forma en que están colocadas, para obtener los resultado correctos.

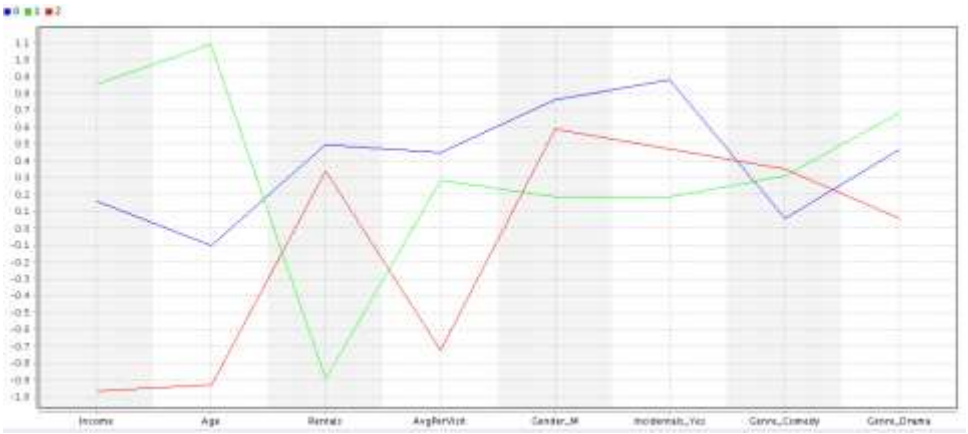


Este cubo solo tiene que ser configurado en su opción main Criterion: ***Avg. Withing.***



Con esto ya tenemos terminado el lienzo, ya podemos probar el modelo, recuerda que la prueba está configurada para 3 modelos pero podemos configurarla a nuestro parecer.

Ejemplo de resultado



Ejercicio:
Que el estudiante prepare información y utilizando el mismo procedimiento de la guía procese los datos.
La información debe ser de cualquier fuente real. (Datos abiertos por ejemplo).