UT5 _ TA2

Ejercicio 1

Clustering aglomerativo, recibe como parámetros la técnica que usa para medir la distancia entre los clusters, los tipos de medidas usadas para determinar las distancias.

Realiza agrupación aglomerativa, una estrategia ascendente de agrupación jerárquica. Admite como estrategias el enlace único, completo y promedio. No proporciona una sola partición del conjunto de datos, sino una jerarquía de grupos que se fusionan a diferentes distancias. En un dendrograma, el eje y marca la distancia a la que se fusionan los grupos, mientras que los objetos se colocan a lo largo del eje x para que los grupos no se mezclen.

Además de elegir funciones de distancia, debe decidirse el criterio de vinculación las ofrecidas por rapid miner y más populares son : agrupación de enlace único (el mínimo de distancias de objeto), agrupación de enlace completo (el máximo de distancias de objeto) o agrupamiento de enlace promedio (el promedio de las distancias).

Toma los parámetros de **técnica a utilizar para medir distancia** (único, completo, promedio)**, tipos de medidas** (mixtas, nominales, numéricas), **tipo de kernel** (cuando se seleccionan medidas numéricas de distancia euclideana)

El **Clustering top-down** agrupa de arriba hacia abajo, partiendo de todo el data set y dividiendo iterativamente el mismo en clusters. Es una estrategia de agrupación jerárquica. Tiene dentro un subproceso, precisa un operador de agrupación plana (por ejemplo k-means).

El operador recibe como parámetro la máxima profundidad y el máximo tamaño de hoja, permitiendo a su vez habilitar la opción de crear una etiqueta para el cluster asignado.

k-means, agglomerative clustering y top-down clustering.

Evaluar el rendimiento de cada uno de estos modelos (justificar la elección de los parámetros correspondientes en cada caso).

En el Clustering agglomerative usando el tipo de distancia promedio y la distancia euclideana dado que se considera óptima para variables numéricas.

En top down Clustering se utiliza una profundidad máxima de 5 y mínimo tamaño de hoja de 5 dado que no queremos que alcance una complejidad demasiado alta y por el número de casos tampoco podemos tener hojas demasiado aglomerativas.

Para k-means se utilizan 5 clusters

Hierarchical Cluster Model Top down

Cantidad de clusters :59

Número de items :250

Hierarchical Cluster Model Agglomerative

Cantidad de clusters:499

Número de items:250

Modificando las distancias el número de clusters no se modifica.

En cuanto a k-means con k=5 y 10 o 100 corridas se obtienen 4 cluster, con una distancia promedio entre cluster de -13.214

Con k=7 se obtienen 7 clusters con distancia promedio de -7,086.