Ejercicio 1

Operadores de Clustering jerárquicos de RapidMiner

Aglomerativo

- Mode: cluster mode, linkage criteria. Se puede elegir entre SingleLink, CompleteLink, AverageLink
- Measure types: criterio para medir la distancia.

Divisivo

- Create Cluster label: especifica el cluster
- Max depth: máxima profundidad
- Max leaf size: máxima cantidad de hojas

Resultados para 4 clusters:

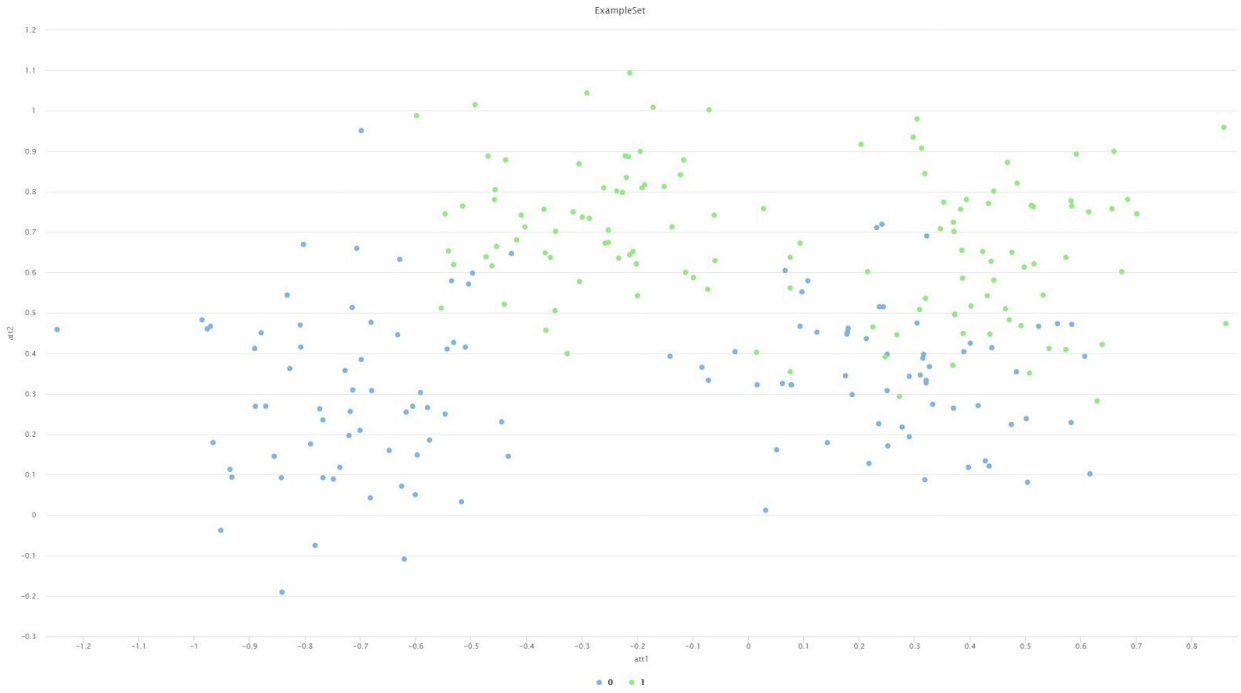
PerformanceVector (clustering aglomerativo):
Avg. within cluster distance: -163.520
Avg. within cluster distance for cluster 0: -165.506
Avg. within cluster distance for cluster 1: 0.000
Avg. within cluster distance for cluster 2: 0.000
Avg. within cluster distance for cluster 3: 0.000

PerformanceVector (clustering divisivo):
Avg. within cluster distance: -18.690
Avg. within cluster distance for cluster 0: -24.498
Avg. within cluster distance for cluster 1: -17.158
Avg. within cluster distance for cluster 2: -16.352
Avg. within cluster distance for cluster 3: -12.886

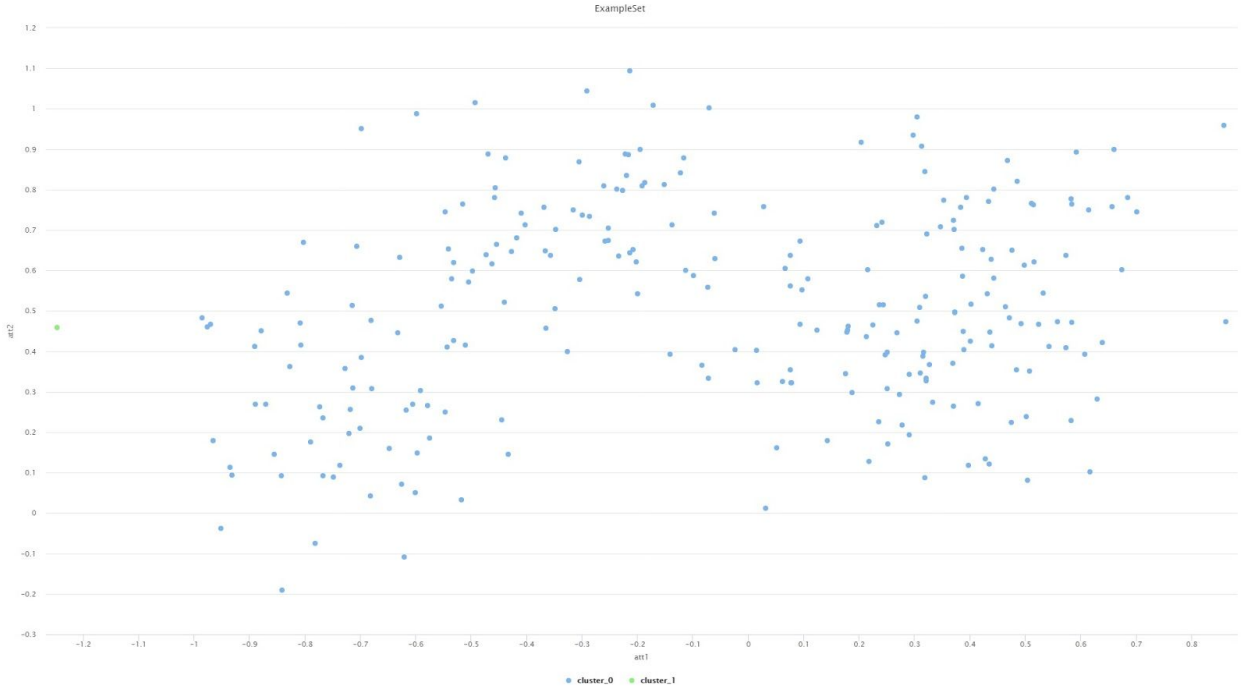
PerformanceVector (clustering k-means):
Avg. within cluster distance: -17.649
Avg. within cluster distance for cluster 0: -19.683
Avg. within cluster distance for cluster 1: -19.395
Avg. within cluster distance for cluster 2: -16.352
Avg. within cluster distance for cluster 3: -13.920

Para 2 clusters:

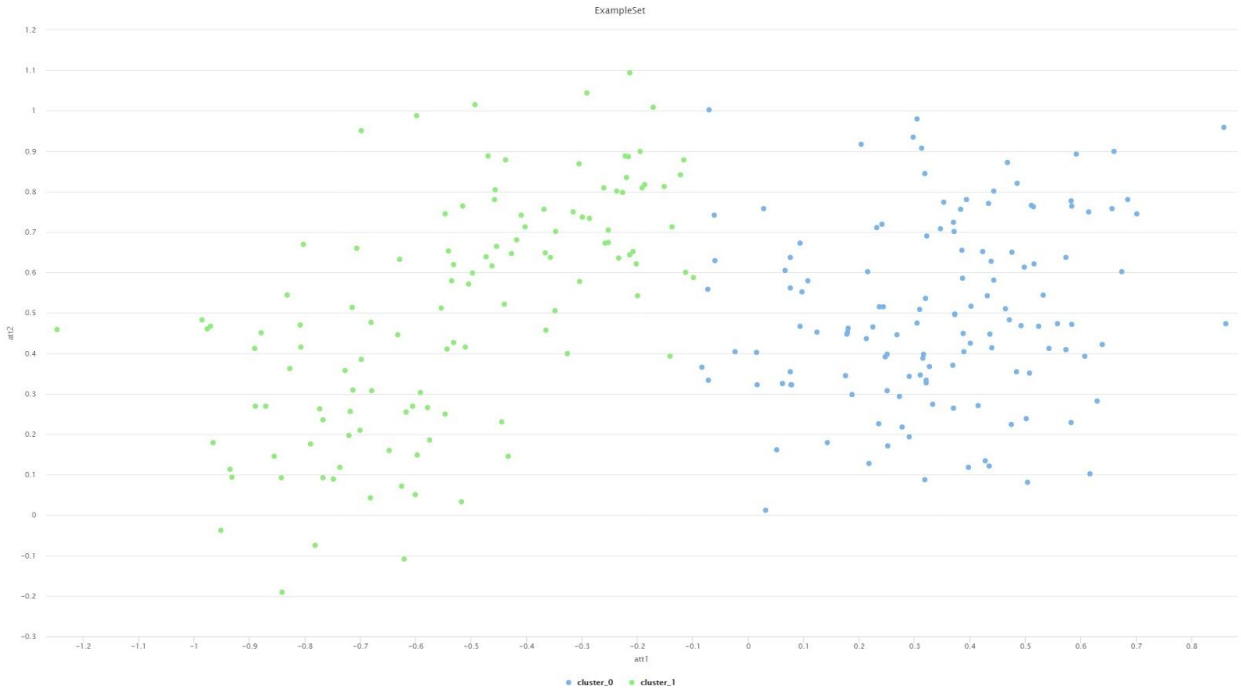
Dataset original:



Agglomerative clustering:



Top-down clustering:



K-means clustering:

Ejercicio 2

DataSet:

- Hay faltantes en varios atributos
- Los atributos Exam todos tienen el mismo rango (6-10)
- Los atributos EntranceExam y Average_Grade (los unicos otros atributos numericos) parecen tener distribuciones normales sin outliers
- Los valores faltantes en los atributos Exam fueron remplazados con 0, dado que si el valor falta puede significar que el estudiante no rindio ese examen. Como solo dos de los atributos Exam tienen un gran cantidad de faltantes, se podria tambien filtrar esos dos atributos completamente.
- Los valores faltantes en el atributo Region fueron reemplazados por la región mas frecuente.
- Dado que clustering requiere medir distancias entre ejemplos, tambien se aplica normalizacion para que los valores de todos los atributos esten en el mismo rango.

Parametros:

- Para el clustering k-means k=3 (dado que el atributo original Students_Success tiene 3 valores posibles) y max runs=10
- Para el aglomerativo de utilizo el modo SingleLink
- Para el top-down max depth=5 y max leaf size=1
- Para el DBSCAN epsilon=1.0 y min points=5
- En los flatten clustering se especifico 3 como numero de clusters

Performance:

K-means:
Avg. within cluster distance: -166.568
Avg. within cluster distance for cluster 0: -192.203
Avg. within cluster distance for cluster 1: -171.858
Avg. within cluster distance for cluster 2: -113.825

Agglomerative:
Avg. within cluster distance: -571.294
Avg. within cluster distance for cluster 0: -574.433
Avg. within cluster distance for cluster 1: 0.000
Avg. within cluster distance for cluster 2: 0.000

Top-down:
Avg. within cluster distance: -183.156
Avg. within cluster distance for cluster 0: -241.104

Avg. within cluster distance for cluster 1: -78.082
Avg. within cluster distance for cluster 2: -162.492

DBSCAN:
Avg. within cluster distance: -110.892
Avg. within cluster distance for cluster 0: -73.165
Avg. within cluster distance for cluster 1: -116.018
Avg. within cluster distance for cluster 2: -19.832
Avg. within cluster distance for cluster 3: -174.330
Avg. within cluster distance for cluster 4: -6.186
Avg. within cluster distance for cluster 5: -23.837
Avg. within cluster distance for cluster 6: -14.422
Avg. within cluster distance for cluster 7: -7.684
Avg. within cluster distance for cluster 8: -3.144
Avg. within cluster distance for cluster 9: -4.371