Tarea 2: Big Data y Machine Learning

Daniel Redel

5/20/2021

A. Modelo k-means con 2 Clusters (k=2)

1.En primer lugar, se hace necesario normalizar las variables, ya que cada variable opera bajo escalas distintas. Usaremos dos métodos de normalización y veremos cómo cambian los resultados finales según el método que se use:

$$Norm_1 = \frac{x - min(x)}{max(x) - min(x)}$$

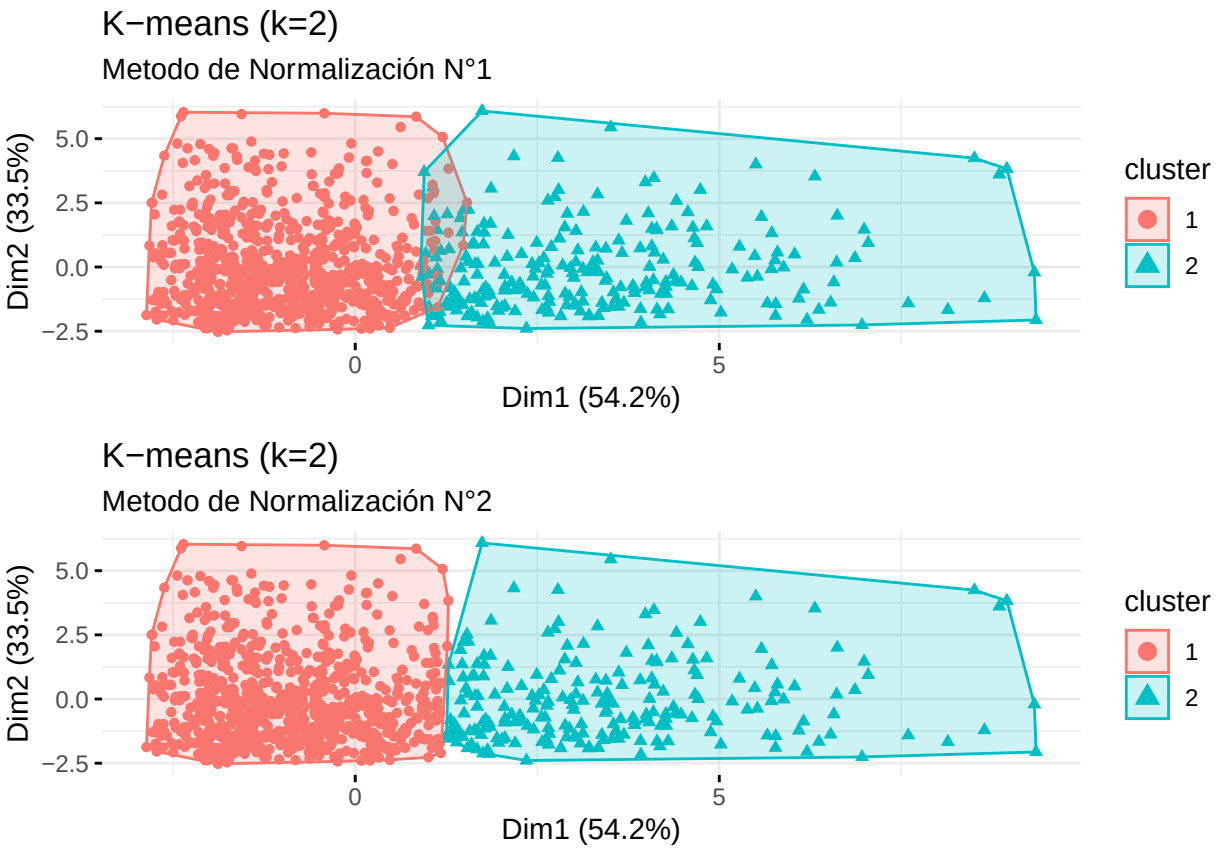
$$Norm_2 = \frac{x - mean(x)}{sd(x)}$$

2. Encontramos los dos cluster a partir del método aprendizaje no-supervisado k-means. La siguiente tabla reporta las características principales de cada cluster para cada método de normalización:

cluster1	age	amount	duration
1	35.7	2158.8	15.7
2	35.1	6852.6	37.8

cluster2	age	amount	duration
1	35.7	2155.6	16.2
2	35.1	7273.4	37.9

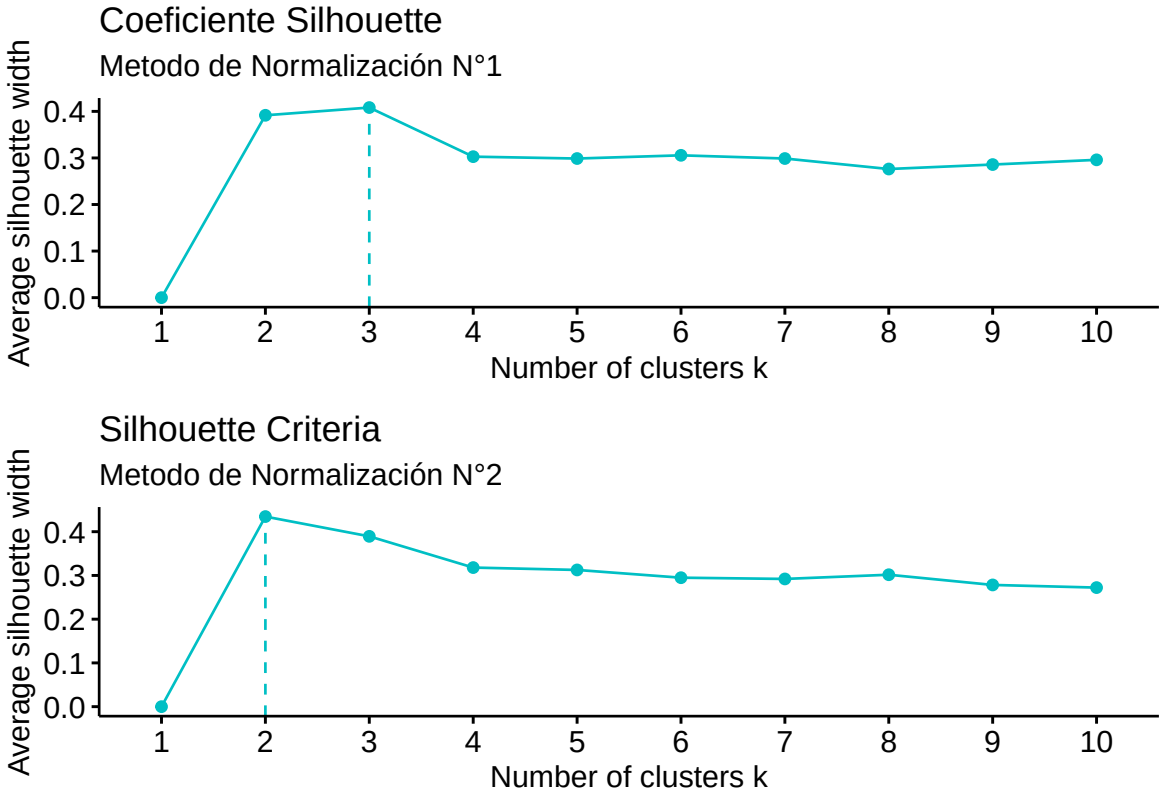
3. Gráficamente, el método k-means estima los siguientes grupos:



Claramente, el método 2 de normalización genera menor solapamiento entre los clusters.

B. Coeficiente de Silhouette

1. Evaluamos la eficiencia de los clusters encontrados usando el coeficiente de silhouette:



Bajo el método 1 de normalización de variables, el coeficiente de silhouette alcanza 0.339. Luego, bajo este criterio, el número de cluster óptimo es 3, ya que el coeficiente máximo es 0.357. En cambio, con el método 2 de normalización, el número óptimo de cluster es 2 (coeficiente es igual a 0.375).

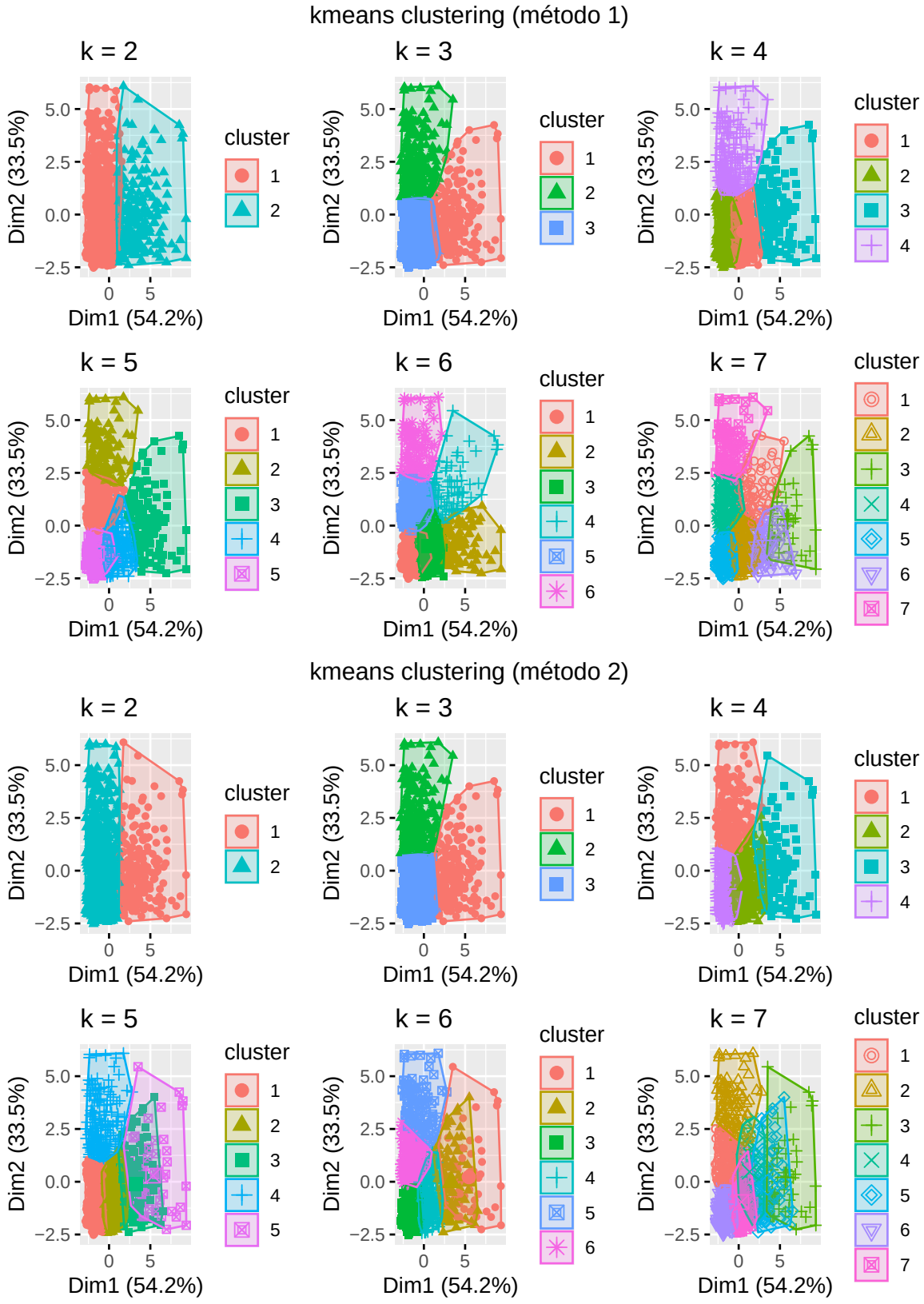
C. Calinski-Harabaz Score

1. La eficiencia de los cluster podemos evaluarla también bajo el criterio Calinski-Harabaz:

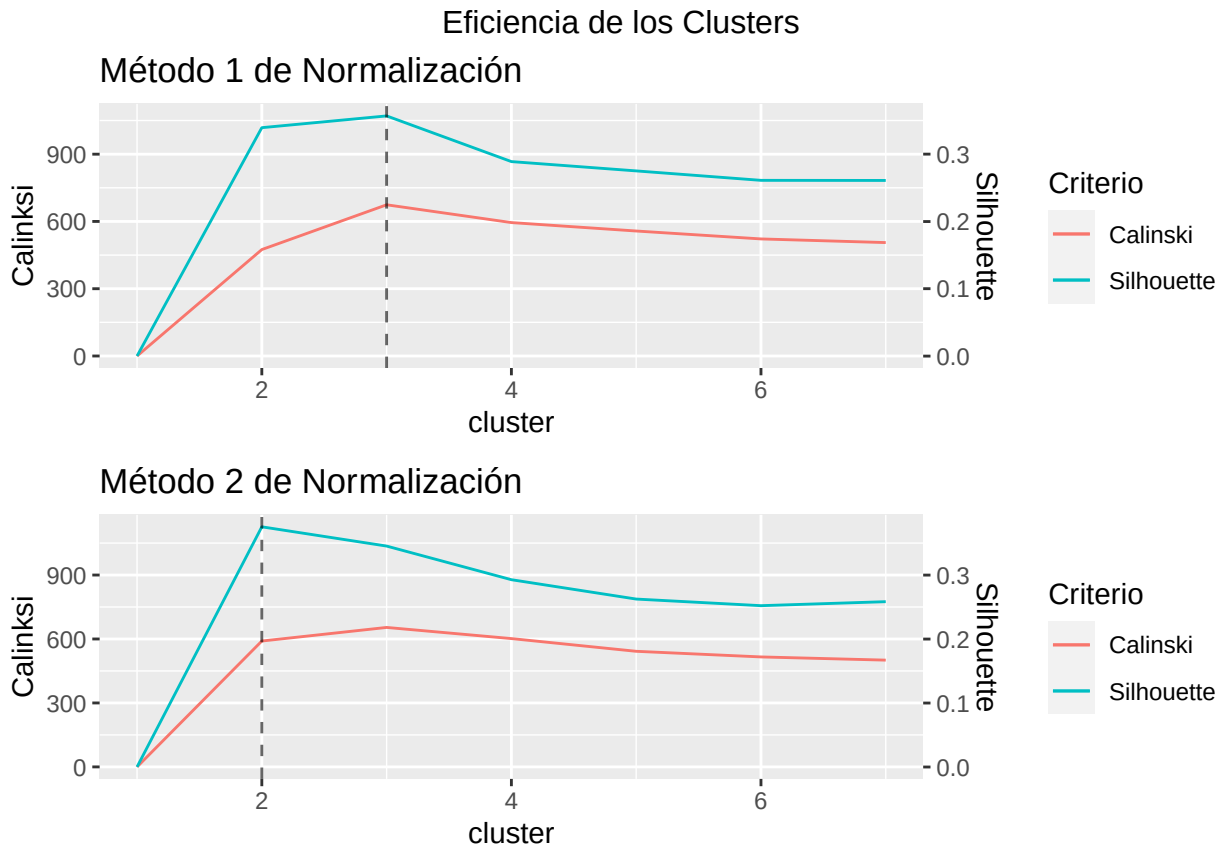
normalizacion	ch_score
método 1	474.6
método 2	590.4

D. Encontrando el cluster k óptimo

1. Estimaremos clusters hasta $k = 7$ y evaluaremos cuál es el óptimo a partir de los distintos criterios (Silhouette o Calinski-Harabaz). Para cada método de normalización, visualizamos los respectivos clusters:



2. Graficamos los criterios Silhouette y Calinski-Harabaz conjuntamente para definir el número de k clusters óptimo (según método de normalización):



Decimos entonces que, usando ambas métricas de eficiencia, el número de clusters óptimo es $k = 3$ si usamos el método de normalización 1 y $k = 2$ si usamos el método de normalización 2.