Ejercicio 2 Desempeño de las golfistas profesionales en la gira LPGA de 2008

Alondra Sánchez Molina

Introducción

El conjunto de datos se puede obtener del archivo lpga.csv y tiene los siguientes atributos:

i. Golfer: nombre del jugador
ii. Distancia de conducción promedio
iii. Porcentaje de calle (Fairway)
iv. Verdes en regulación: en porcentaje
v. Promedio de putts por ronda
vi. Intentos de arena por ronda
vii. Ahorro de arena: en porcentaje
viii. Ganancias totales por ronda
ix. Log: calculado como (Total de victorias/ronda)
x. Rondas totales
xi. Id: identificación única que representa a cada jugador

Usa la agrupación en este conjunto de datos para averiguar qué jugadores tienen un rendimiento similar en la misma temporada.

Preparación de los datos

Primeramente, se carga el archivo; y se visualiza con el fin de tener un primer acercamiento a estos.

Posteriormente, se utiliza la función sum() de R, para observar si existen datos sin valor en el dataset, en este caso, se obtiene que no.

```
> sum(is.na(dt_lgpa))
[1] 0
```

Al visualizar el dataset, se observa que es necesario remover la columna de los nombres de los golfistas, es por ello que se opta por colocar estos como el nombre de las filas.

```
> rownames(dt_lgpa) <- dt_lgpa$ï..Name
> dt_lgpa <- dt_lgpa[, -c(colnames(dt_lgpa) %in% ("ï..Name"))]
```

Así mismo, se nota que la última columna del dataset es el id, es por ello que se elimina dicha columna.

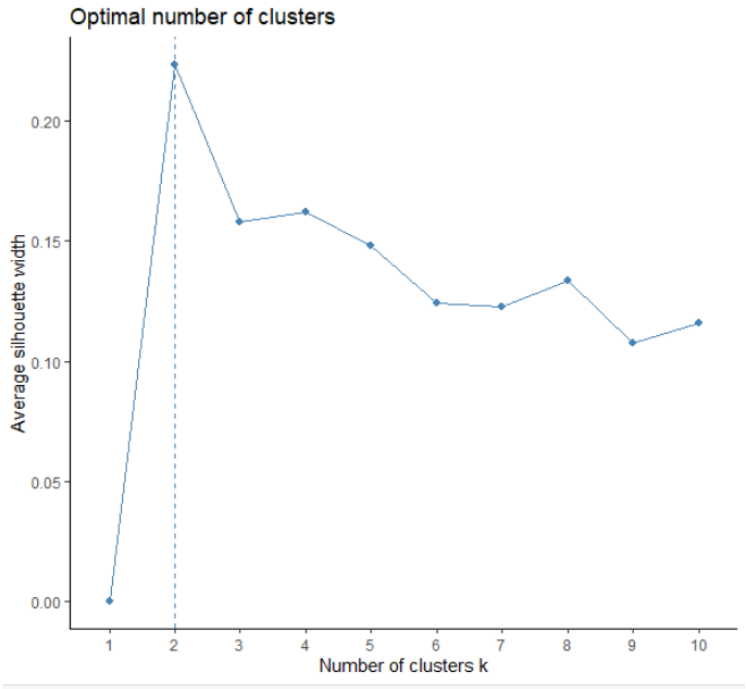
dt_lgpa <- dt_lgpa[, -c(10)]

Finalmente, es necesario escalar los datos con la finalidad de normalizarlos.

data <- scale(dt_lgpa)

Estimación de k

Es necesario elegir el número adecuado de ellos, para estimar el valor k más óptimo se utiliza la función *fviz_nbclust()*. El método a utilizar fue silhouette; el cual dibuja la silueta de los grupos promedio de acuerdo con el número de grupos.



Algoritmo CLARA

Justificación

Se utilizó dicho algoritmo, ya que al examinar el dataset, se puede notar que son más datos, es por ello que se decidió usar la versión de K-Medoids que implementa CLARA, ya que funciona dividiendo el conjunto de datos en varios subconjuntos con tamaño fijo.

CLARA

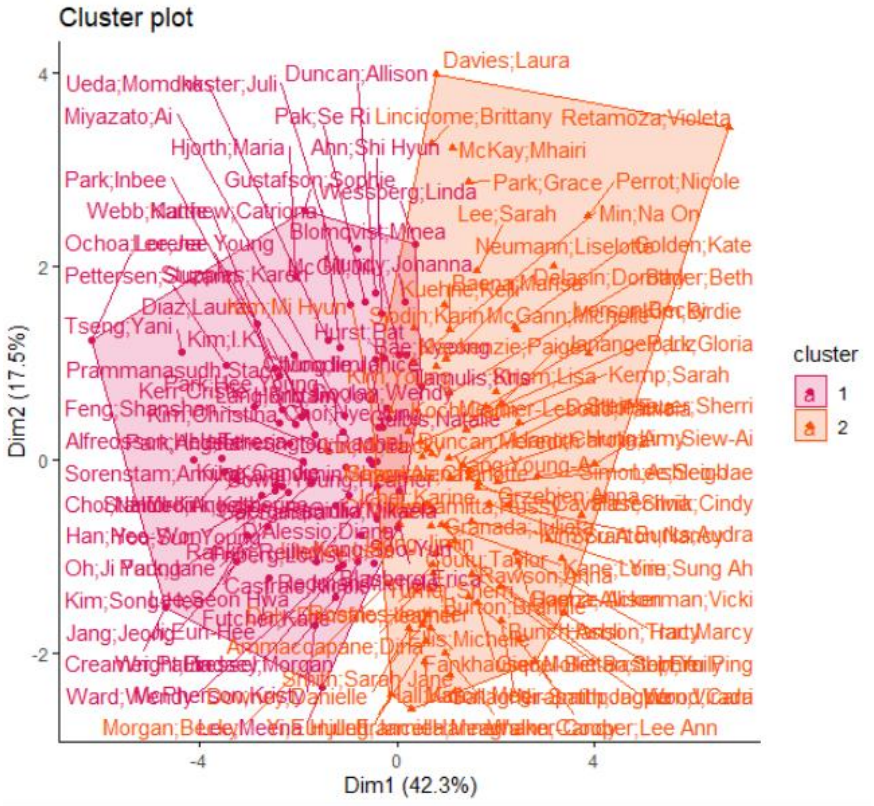
Se utiliza la función clara(), pasando nuestros datos y el número de clusters, en este caso 2. Dicha función al imprimirla nos retorna la información de los medoids de cada cluster. Así como el vector de cluster generado.

```
> clara.res <- clara(data, 2, samples = 50, pamLike = TRUE)
> print(clara.res)
Call:    clara(x = data, k = 2, samples = 50, pamLike = TRUE)
Medoids:
                  Avg.Drive    Fairway     Greens  Avg.Putts       Sand Sand.Saves   Winnings  Win.Round     Rounds
Prammanasudh;Stacy  0.3697398 -0.3252168  0.5464808 -0.7861759 -0.7342167  0.1403620  0.2419801  0.7181244  1.0686381
Hung;Amy           -0.6263573  0.3329390 -0.4847995  0.3913666  0.8917019 -0.1534625 -0.5994113 -0.5982817 -0.4036443
Objective function:      2.501972
Clustering vector:       Named int [1:157] 1 1 2 2 1 2 2 1 1 1 2 2 2 1 2 2 1 1 ...
 - attr(*, "names")= chr [1:157] "Ahn;Shi Hyun" "Alfredsson;Helen" "Ammaccapane;Dina" "Bader;Beth" "Bae;Kyeong" "Baena;Marisa" '
mily" ...
Cluster sizes:           74 83
Best sample:
  [1] Bae;Kyeong              Baena;Marisa            Bowie Young;Heather     Burks;Audra             Castrale;Nicole
  [6] Cavalleri;Silvia        Cho;Irene               Chung;Ilmi              Davies;Laura            Diaz;Laura
 [11] Dunn;Moira              Gulyanamitta;Russy      Hart;Marcy              Hong;Jin Joo            Hung;Amy
 [16] Hurst;Pat               Icher;Karine            Janangelo;Liz           Kemp;Sarah              Kim;Young
 [21] Lang;Brittany           Lee;Jee Young           Lee;Meena               Lee;Sarah               Lin;Yu Ping
 [26] Lindley;Leta            Lucidi;Becky            Mackenzie;Paige         McPherson;Kristy        Meunier-Lebouc;Patricia
 [31] Miyazato;Ai             Park;Gloria             Park;Inbee              Perrot;Nicole           Prammanasudh;Stacy
 [36] Rankin;Reilley          Redman;Michele          Scranton;Nancy          Strom;Lisa              Walker-Cooper;Lee Ann
 [41] Wessberg;Linda          Yang;Young-A            Yim;Sung Ah             Yoo;Sun Young

Available components:
 [1] "sample"     "medoids"    "i.med"      "clustering" "objective"  "clusinfo"   "diss"       "call"       "silinfo"
[10] "data"
```

Visualización

Con el plotear de los datos, podremos analizar de una manera visual cuáles países pertenecen a cada grupo.



Interpretación

El generar clusters, no simplemente es plotear los agrupamientos generados, es inverosímil analizarlos. Para ello, se utiliza la función aggregate(), para obtener los valores promedios de cada cluster.

```
> aggregate(dt_lgpa, by=list(cluster=clara.res$cluster), mean)
  cluster Avg.Drive  Fairway   Greens Avg.Putts      Sand Sand.Saves Winnings Win.Round   Rounds
1       1  251.0405 67.17838 64.82432  28.55122 0.8904054   38.49459 8280.649  8.775870 70.97297
2       2  242.8506 67.93373 61.23614  29.76807 1.0427711   37.47108 1817.614  7.277914 48.90361
```

Conclusiones

Se observa que los golfistas se dividen en dos grupos, los datos en ambos no se encuentran tan alejados, pero se observa que la distancia de conducción promedio, es mayor en el grupo 1, sin embargo, el porcentaje de calle es muy similar en ambos grupos. Los golfistas del grupo 1, han ganado más veces que los pertenecientes al cluster número 2, aunque esto tal vez se deba a que los golfistas que integran el cluster 1, han participado en más rondas.