

# Problemas de Clustering

1. Consideremos la tabla de datos `worldcup` del paquete `faraway` que nos información sobre los jugadores de Futbol que participaron en el Mundial de Futbol celebrado el año 2010 en Sudáfrica. Esta tabla de datos da información de 595 jugadores y tiene 7 variables:

- **Team:** el pais del jugador.
  - **Position:** la posición en que juega el jugador. Tiene 4 valores:
    - **Defender:** defensa.
    - **Forward:** delantero.
    - **Goalkeeper:** portero.
    - **Midfielder:** medio.
  - **Time:** tiempo jugado en minutos.
  - **Shots:** número de tiros que ha realizado el jugador.
  - **Passes:** número de pases del jugador.
  - **Tackles:** número de entradas del jugador.
  - **Saves:** número de paradas del jugador.
- a) Seleccionar una muestra de 25 jugadores usando la función `sample`. Escribir `set.seed(2020)` antes de elegir la muestra.
  - b) Aplicar el algoritmo k-means a la muestra anterior usando las variables cuantitativas para clasificar a los 25 jugadores en 4 grupos usando el algoritmo de MacQueen. Aplicar la función `kmeans` unas cuantas veces con el fin de que la suma de los cuadrados de todos los clusters sea mínima.
  - c) Queremos estudiar hasta qué punto la clasificación anterior coincide con la clasificación de los 25 jugadores según la posición que ocupan. Calcular la tabla bidimensional que dos dé el cluster a qué pertenece el jugador por un lado y la posición a la que juega. ¿Qué porcentaje de aciertos ha tenido el algoritmo k-means?

1. Consideremos la tabla de datos `worldcup` del paquete `faraway` que nos información sobre los jugadores de Futbol que participaron en el Mundial de Futbol celebrado el año 2010 en Sudáfrica. Esta tabla de datos da información de 595 jugadores y tiene 7 variables:

- **Team:** el pais del jugador.
- **Position:** la posición en que juega el jugador. Tiene 4 valores:
  - **Defender:** defensa.
  - **Forward:** delantero.
  - **Goalkeeper:** portero.
  - **Midfielder:** medio.
- **Time:** tiempo jugado en minutos.
- **Shots:** número de tiros que ha realizado el jugador.
- **Passes:** número de pases del jugador.
- **Tackles:** número de entradas del jugador.
- **Saves:** número de paradas del jugador.

- a) Seleccionar una muestra de 25 jugadores usando la función `sample`. Escribir `set.seed(2020)` antes de elegir la muestra.
- b) Calcular la matriz de distancias de los 25 jugadores anteriores usando la distancia euclídea entre las variables cuantitativas.
- c) Usando el método jerárquico aglomerativo del **enlace promedio** hallar el dendrograma para clasificar los 25 jugadores anteriores.
- d) Clasificar los 25 jugadores en 4 clusters a partir del dendrograma anterior.
- e) Queremos estudiar hasta qué punto la clasificación anterior coincide con la clasificación de los 25 jugadores según la posición que ocupan. Calcular la tabla bidimensional que nos dé el cluster a qué pertenece el jugador por un lado y la posición a la que juega. ¿Qué porcentaje de aciertos ha tenido el algoritmo aplicado?