

## TAREA NÚMERO 5

1. En este ejercicio usaremos la tabla de datos `EjemploAlgoritmosRecomendación.csv`, la cual contiene los promedios de evaluación de 100 personas que adquirieron los mismos productos o muy similares en la tienda AMAZON. La idea consiste en recomendar a un cliente los productos que ha comprado otra persona que pertenece al mismo clúster.
  - a) Ejecute el método  $k$ -medias con `iter.max = 200`, `nstart = 100` para  $k = 4$ , luego desde `RStudio` verifique el Teorema de Fisher para este ejemplo.
  - b) Ejecute el método  $k$ -medias con `iter.max = 200`, `nstart = 100`, para esto encuentre valor de  $k$  usando los métodos Gap Statistic, wss y Average Silhouette usando la función `fviz_nbclust`, luego interprete los resultados usando interpretación Horizontal-Vertical y gráficos tipo radar plot.
  - c) Si se tienen 7 clústeres usando el método de  $k$ -medias ¿Qué productos recomendaría a Teresa, a Leo y a Justin?, es decir, ¿los productos que compra cuál otro cliente? Usando distancia euclídea ¿cuál es la mejor recomendación de compra que le podemos hacer a Teresa, a Leo y a Justin?
2. El conjunto de datos `DatosBeijing.csv` contiene datos por hora de la concentración de la partícula PM2.5 en la ciudad de Beijing, también incluye datos meteorológicos del Aeropuerto Internacional de Beijing. Contiene un `Id` y 12 variables que se explican seguidamente:
  - `Id`: Número de fila.
  - `Anno`: Año de datos en esta fila.
  - `Mes`: Mes de datos en esta fila.
  - `Dia`: Día de datos en esta fila.
  - `Hora`: Hora de datos en esta fila
  - `ConcetracionParticula_pm2.5`: Concentración de PM2.5.
  - `PuntoRocio`: Punto de rocío.
  - `Temperatura`: Temperatura.
  - `Presion`: Presión (hPa).
  - `DireccionViento`: Dirección del viento combinado.
  - `VelocidadViento`: Velocidad del viento acumulada.
  - `HorasNieve`: Horas acumuladas de nieve.
  - `HorasLluvia`: Horas acumuladas de lluvia.

Efectúe un análisis de  $k$ -medias siguiendo los siguientes pasos:

- a) Cargue la tabla de datos y ejecute un `str(...)`, `summary(...)` y un `dim(...)`, verifique la correcta lectura de los datos.

- b) Elimine las filas con NA usando el comando `na.omit(...)`. ¿Cuántas filas de eliminaron?
- c) Elimine de la tabla de datos la variable `DireccionViento`. ¿Por qué se debe eliminar? ¿Qué otra alternativa se tiene en lugar de eliminarla?
- d) ¿Qué pasa si ejecutamos un clustering jerárquico con `hclust(...)`. ¿Por qué sucede esto?
- e) Ejecute un  $k$ -medias con  $k = 3$ , `iter.max=1000` y `nstart=50`.
- f) Dé una interpretación de los resultados usando un gráfico tipo radar.
- g) Construya el Codo de Jambu usando `iter.max=100` y `nstart=5`, ¿cuántos conglomerados (clústeres) sugiere el codo? Utilice también el método `silhouette` de la función `fviz_nbclust`, ¿cuántos conglomerados (clústeres) sugiere este método?

**Entregables:** Incluya en documento autoreproducible (HTML) todas las instrucciones y códigos R utilizados en cada ejercicio, incluya los resultados de los cálculos, los gráficos generados y las respuestas a las preguntas. El ejercicio 5 lo pueden hacer a mano.