UNIVERSIDAD DE LIMA

# ANÁLISIS CLUSTER

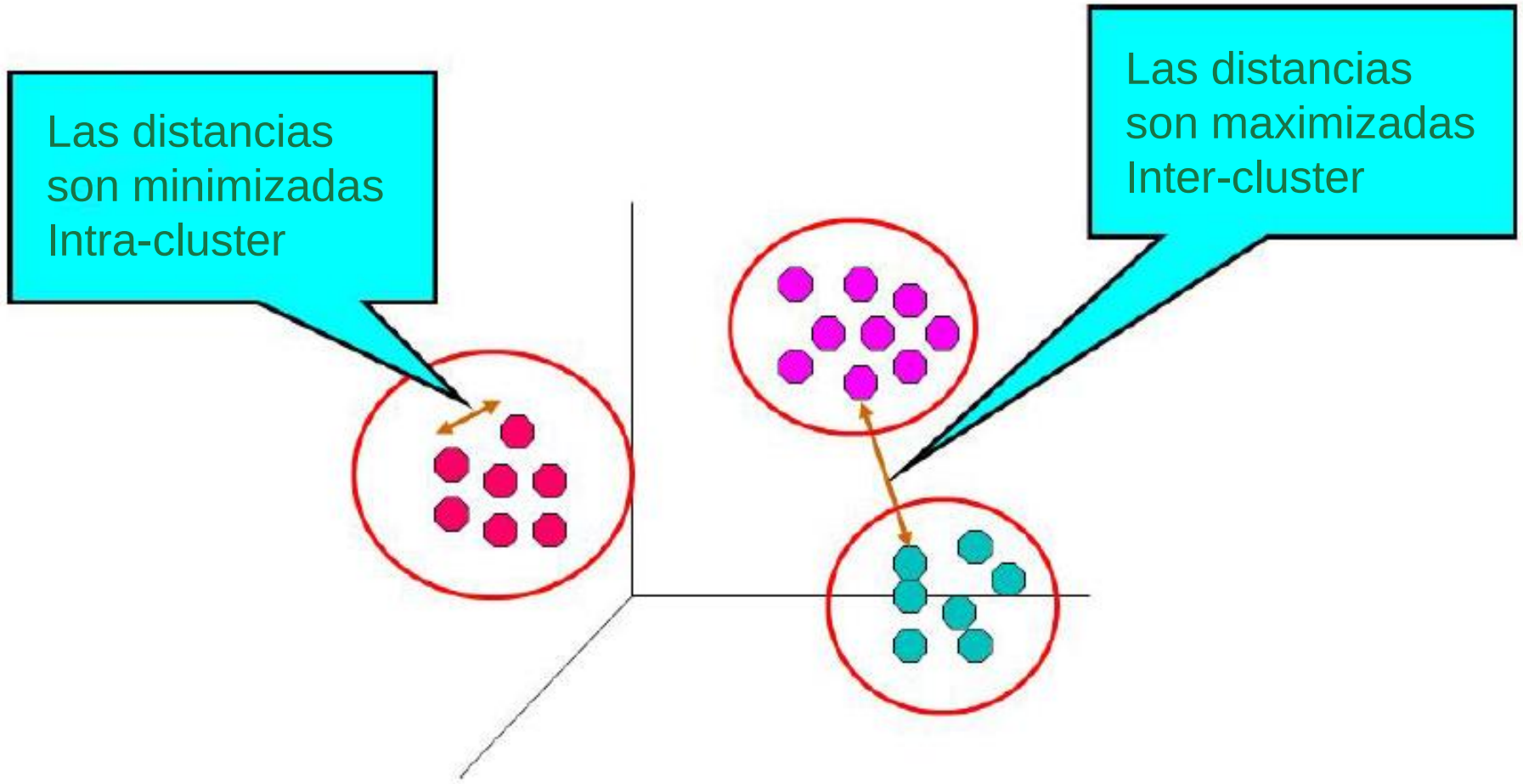
**Mg Sc. Meza Rodríguez, Aldo Richard**

**armeza@ulima.edu.pe**

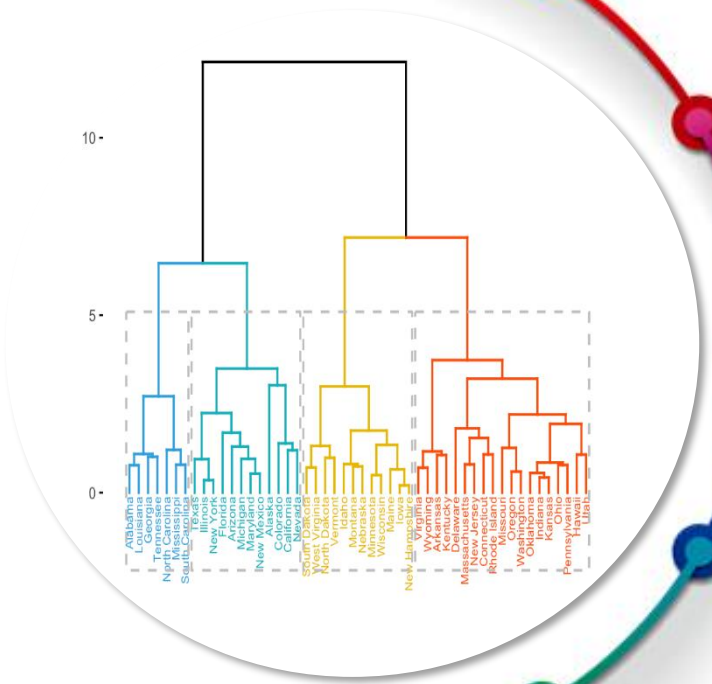
# El análisis Clúster

Es un método multivariado cuyo propósito es el de agrupar objetos o individuos, de tal manera que los objetos que se clasifican y pertenecen a un grupo o cluster son muy similares (homogeneidad dentro del grupo o cluster), mientras que son muy disímiles aquellos objetos que pertenecen a grupos diferentes (heterogeneidad entre grupos o cluster).

# Análisis Clúster



# Características y requisitos




El AC, es un método de interdependencia y como una técnica exploratoria

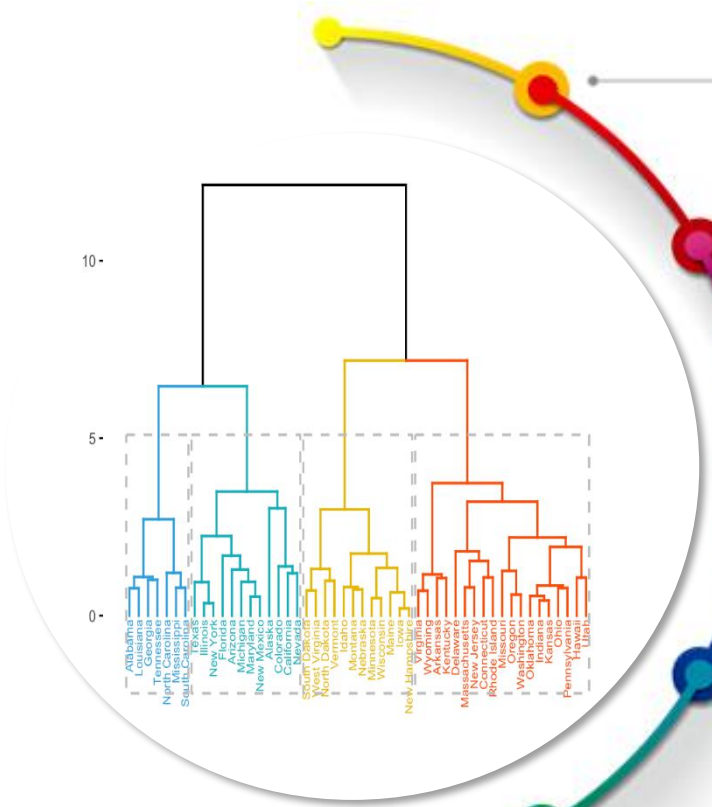
El AC, se utiliza para agrupar unidades elementales y también para agrupar variables

El AC, utiliza métricas e índices de similitud como una medida de distancia.

Es importante identificar cuantos conglomerados formar, que tipos de variables utilizar

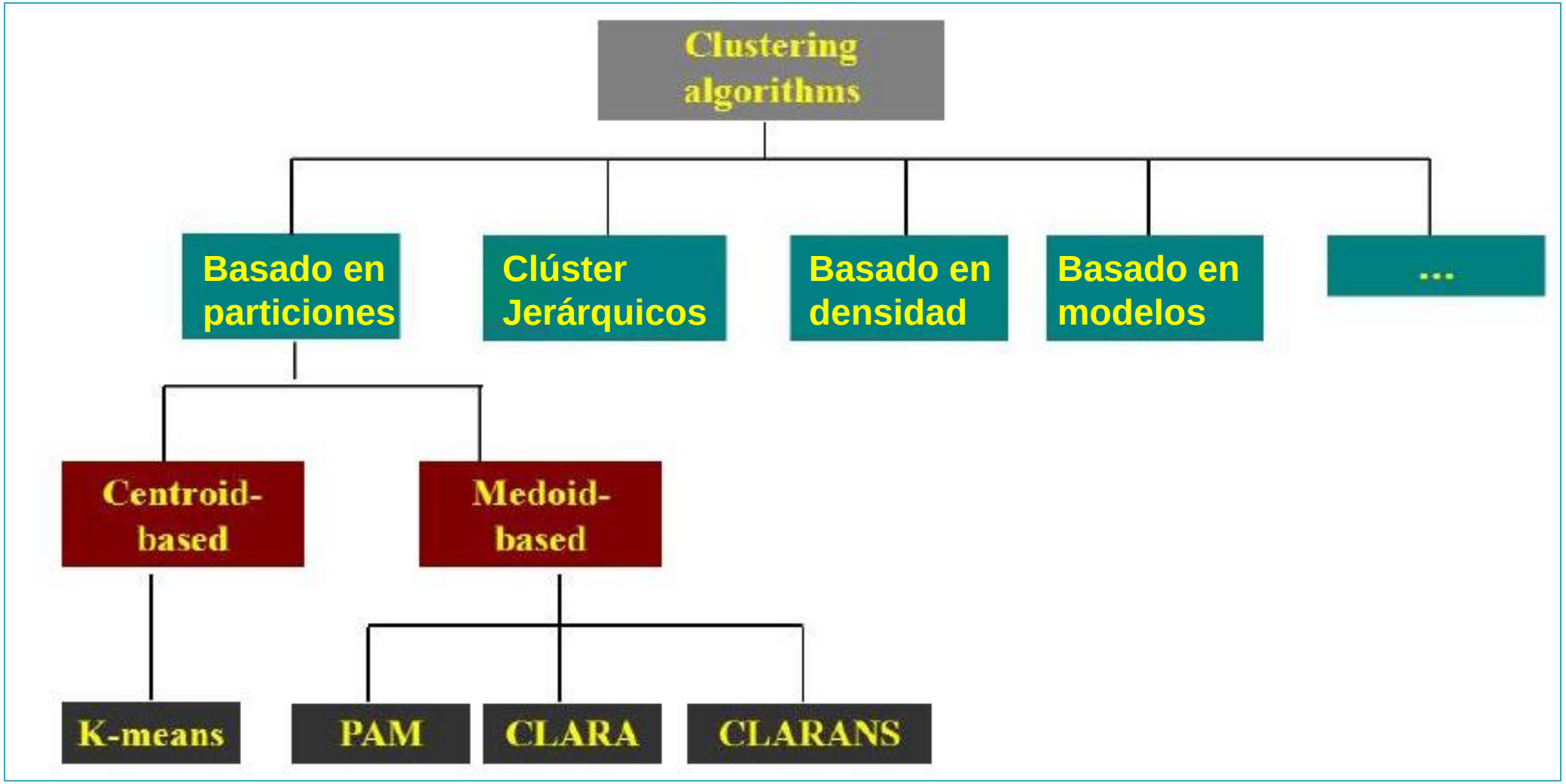
Es necesario identificar que método, algoritmo o enlace a trabajar.

# Características y requisitos




- Es importante que haya ausencia de valores atípicos puesto que estos son perjudiciales para las distancias.
- Las variables no deben estar correlacionadas entre si.
- El número de variables no debe ser muy elevado.
- El número de observaciones no debe ser muy grande, de ser así, es preferible utilizar métodos de particionamiento
- Las variables no deben estar en unidades diferentes, de ser así, será necesario estandarizar las variables.

# Clasificación de las técnicas Clúster



**Medidas de distancia**

A mayor distancia los individuos están más alejados. La distancia entre dos individuos i y j: $X_{ik}$ y $X_{jk}$

**Euclidiana.** Depende de la unidades de medida, aplicándose a variables estandarizadas. No considera la posible correlación entre las variables.

$$d_{ij} = \left[ \sum_{k=1}^{p} (x_{ik} - x_{jk})^2 \right]^{\frac{1}{2}}$$

**Mahalanobis**. Considera la correlación entre las variables, utilizando las covariancias como ponderación.

$$d_{ij}^2 = (x_i - x_j)' S^{-1} (x_i - x_j)$$

**Pearson**.

$$d_{ij} = \left[ \sum_{k=1}^{p} \frac{(x_{ik} - x_{jk})^2}{S_k^2} \right]^{\frac{1}{2}}$$

# Medidas de distancia

## ¿Cómo medir la similitud?

| Individuo | $X_1$ | $X_2$ |
|---|---|---|
| A | 2 | 4 |
| B | 4 | 6 |
| C | 9 | 10 |

¿Quién se parece más al individuo A? ¿B ó C?

$$Euclidean\ Distance = d = \sqrt{\sum_{i=1}^{N}(Xi - Yi)^2}$$

$$d(A,B) = \sqrt{(2-4)^2+(4-6)^2} = 2.82$$

$$d(A,C) = \sqrt{(2-9)^2+(4-10)^2} = 9{,}22$$

$$d(B,C) = \sqrt{(4-9)^2+(6-10)^2} = 6{,}40$$

**Matriz de Distancia**

| | A | B | C |
|---|---|---|---|
| A | 0 | 2.82 | 9.22 |
| B | 2.82 | 0 | 6.40 |
| C | 9.22 | 6.40 | 0 |

# Determinación del número de cluster

El número de cluster depende de la distancia que se haga el corte (menor distancia mayor número de grupos).

- **Método Gráfico**. Se utiliza el Dendograma, que muestra el agrupamiento de los objetos en la conformación de cada cluster.
- **Pseudo Estadística F (Tipo F de Beale)**. Obteniendo y Evaluando ANVAs de cada variable, esperando que resulte significativa la F.
- **Pseudo Estadística de $T^2$ de Hotelling**. Permite probar la igualdad de los vectores de medias de las variables de dos grupos. Se espera que la prueba sea significativa, indicando que los vectores de medias son diferentes (cluster diferentes).

# Interpretación de los clusters

- Consiste en interpretar la composición de cada grupo formado, para ello se debe considerar las características de los elementos que lo componen y analizando si poseen o representan determinadas características en mayor medida que otras.

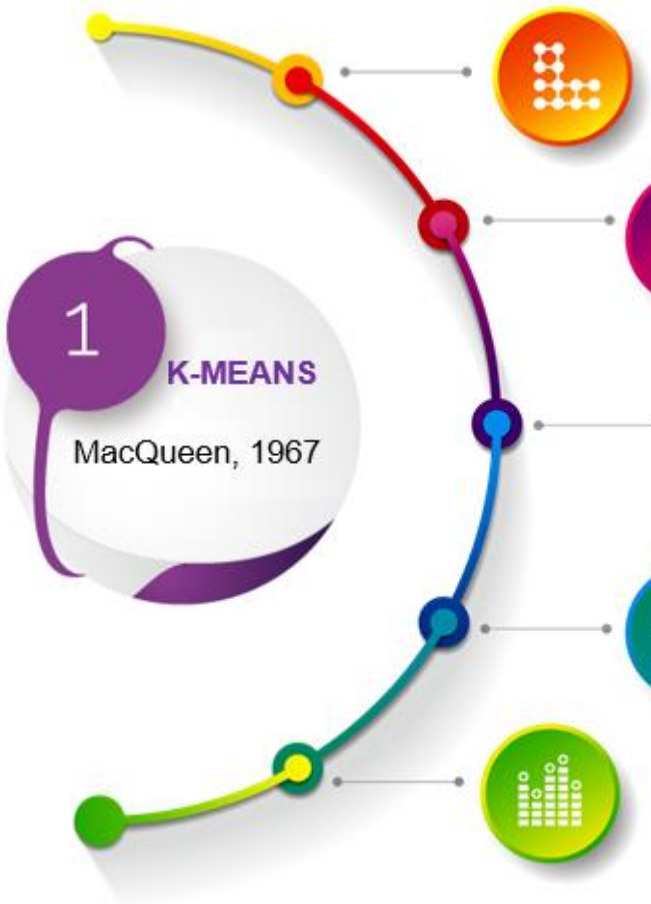
- Una forma es obteniendo estadísticas descriptivas de cada grupo. Así mismo, es posible utilizar otras variables de análisis que no se han considerado en el proceso de clasificación de tal manera que permitan interpretar y extraer conclusiones de cada grupo, a fin de asignarles algún calificativo o nombre genérico.

# Método de partición: Cluster No Jerárquico

Estos métodos son aplicados a volúmenes grandes de datos y para detectar datos atípicos (eligiendo un número alto de cluster, se identifica grupos con pocos datos).

- **Método de K-Medias:** Utiliza la técnica aglomerativa para la formación de cluster y la distancia euclidiana.

# El Algoritmo K-means




Es un algoritmo de agrupamiento por particiones el cual requiere especificar un número K de Clústers.

El algoritmo asignará a cada observación exactamente a un clúster y cada clúster tiene asociado un centroide (centro geométrico del clúster).

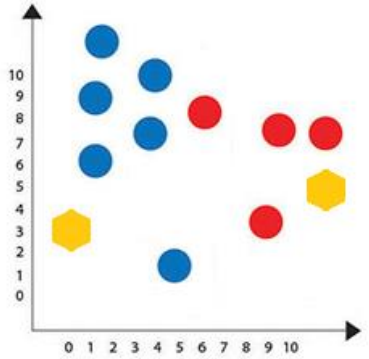
Los puntos se asignan al clúster cuyo centroide esté más cerca (utilizando cualquier métrica distancia)

Esta basado en la minimización de la suma de cuadrados dentro de cada clúster (variación). Es computacionalmente rápido.

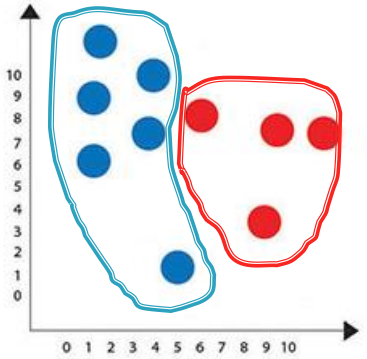
Iterativamente, se van actualizando los centroides en función de las asignaciones de puntos a clúster, hasta que los centroides dejen de cambiar.

# Procedimiento K-mens

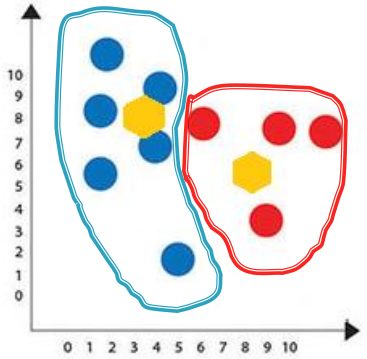
01 Selecciona aleatoriamento K=2 cluster como centroides



02 Asignar cada caso al cluster más cercano

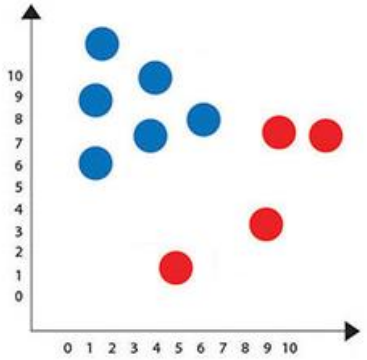


03 Actualizar los centroides

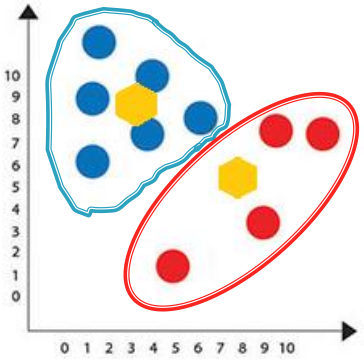


04 Resignar

05 Actualizar los centroides



06 Resignar cada observación



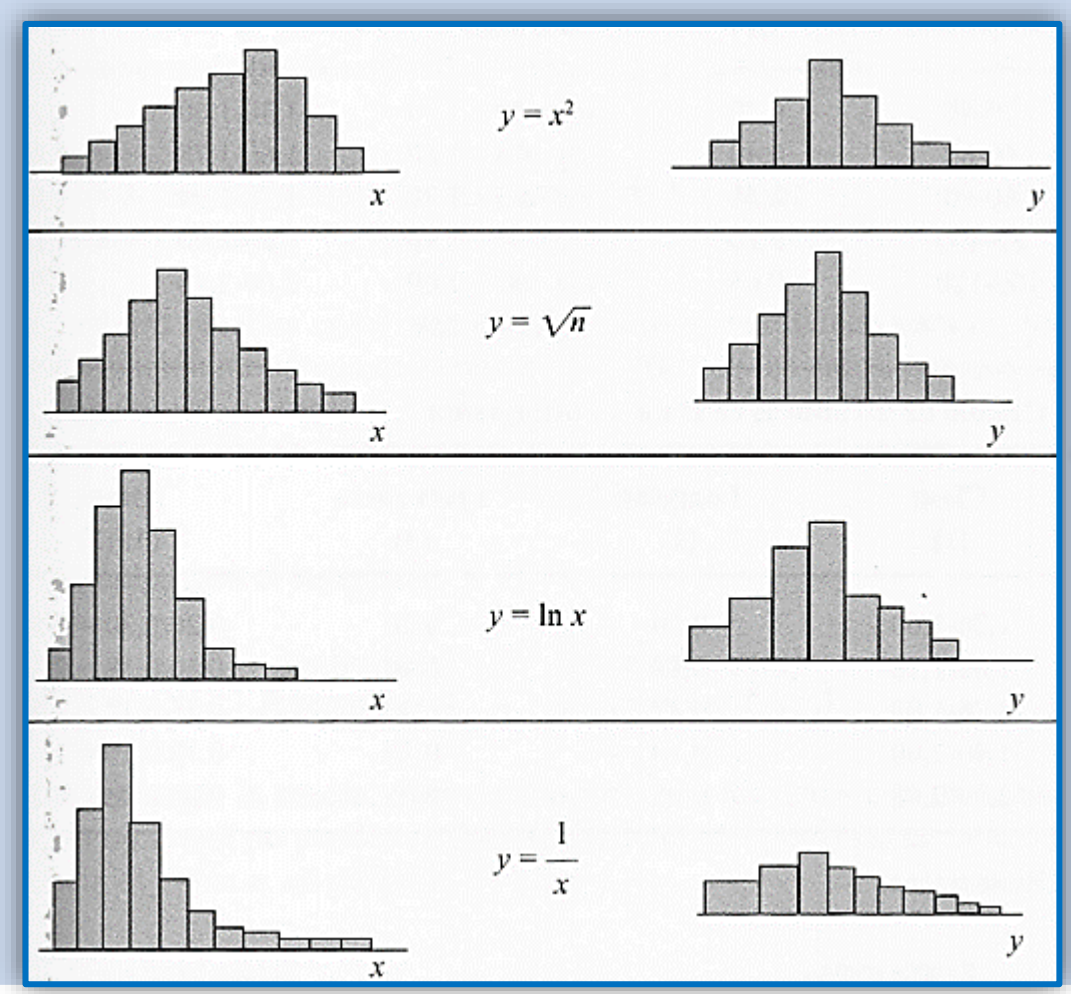
# El Algoritmo K-means

# Ventajas y Desventajas

- Es computacionalmente rápido.
- Simple, eficiente y general.
- El resultado puede variar en base a las semillas elegidas al inicio
- Es sensible a la elección de los centroides iniciales.
- No siempre puede solucionarse con múltiples inicializaciones.
- Es sensible a los “outliers”.
- No trata datos nominales (K-Modes)

# Transformaciones recomendadas

Distribución de variables sin transformar




Distribución de variables Transformadas

# APLICACIÓN DEL CLÚSTER BASADO EN PARTICIONES

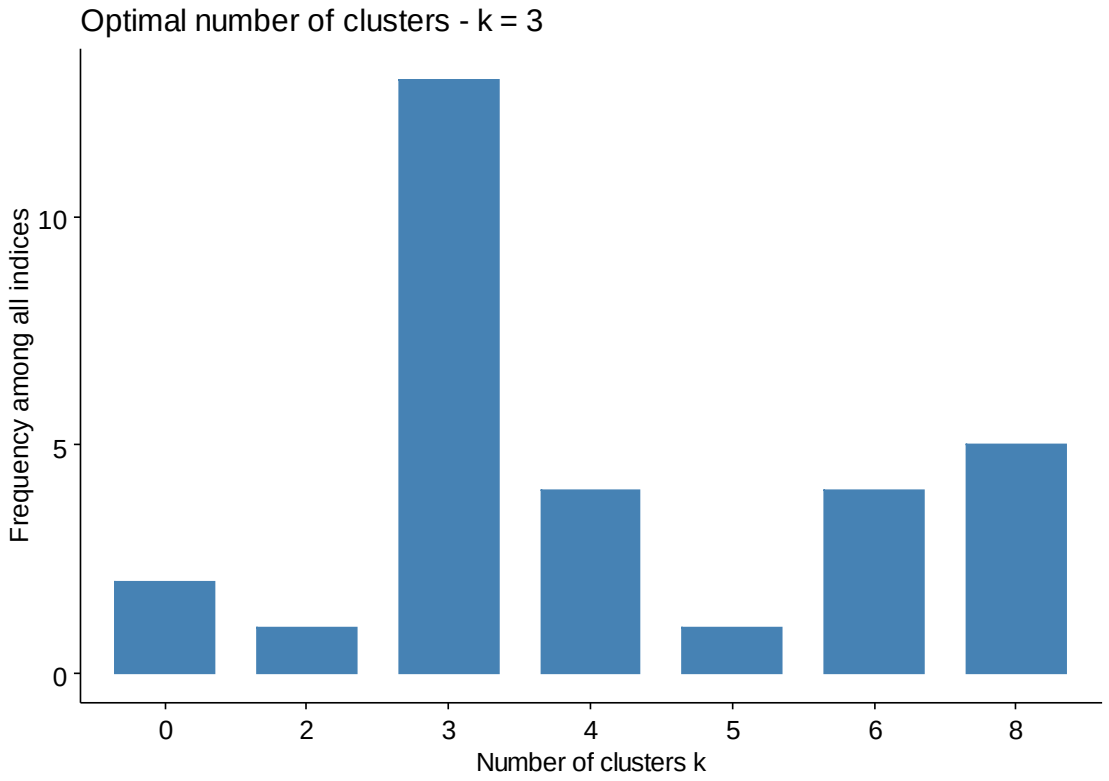
**DATA CALIDAD**

Es una base de datos que muestra el puntaje en calidad de atención que se le da a asesores comerciales después de su atención. Las variables y descripción son las siguientes:

| VARIABLE | DESCRIPCIÓN |
|---|---|
| ID | Código |
| Amab | Amabilidad de atención |
| Interes | Interés en el problema |
| Capa | Capaciad para resolver el problema |
| Clari | Claridad de información |
| Tiemp | Tiempo de atención |
| Soluc | Solución del problema |

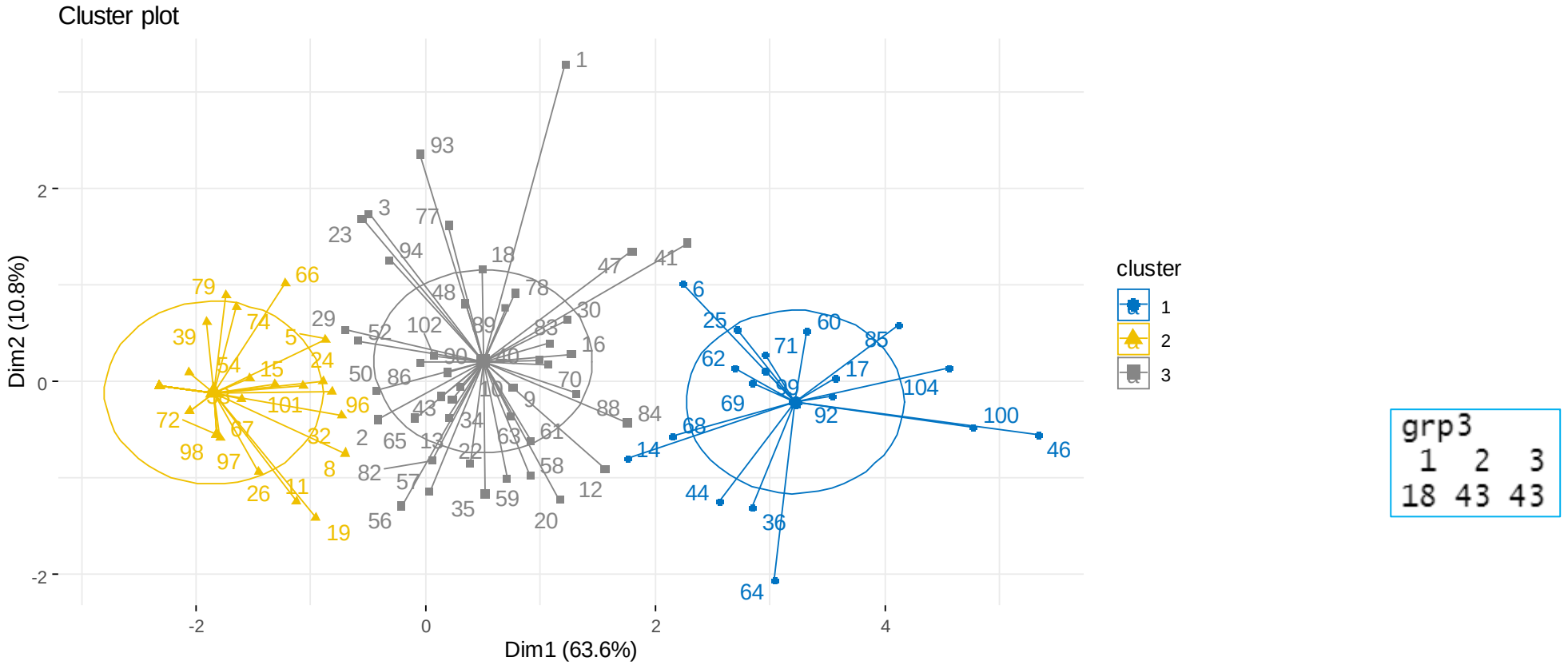
# Número de clusters

| | Number_clusters | Value_Index |
|---|---|---|
| KL | 3 | 5.237300e+00 |
| CH | 3 | 6.708370e+01 |
| Hartigan | 6 | 6.325700e+00 |
| CCC | 3 | -5.949000e-01 |
| Scott | 4 | 7.518820e+01 |
| Marriot | 6 | 8.686507e+09 |
| TrCovW | 4 | 4.369011e+02 |
| TraceW | 6 | 1.252310e+01 |
| Friedman | 8 | 2.305500e+00 |
| Rubin | 6 | -1.765000e-01 |
| Cindex | 8 | 3.062000e-01 |
| DB | 3 | 1.272400e+00 |
| Silhouette | 3 | 3.446000e-01 |
| Duda | 3 | 2.816600e+00 |
| PseudoT2 | 3 | -4.127790e+01 |
| Beale | 3 | -2.422300e+00 |
| Ratkowsky | 3 | 4.350000e-01 |
| Ball | 4 | 2.983490e+01 |
| PtBiserial | 5 | 6.012000e-01 |
| Gap | 3 | -1.630000e+00 |
| Frey | 2 | NA |
| McClain | 3 | 9.201000e-01 |
| Gamma | 8 | 8.451000e-01 |
| Gplus | 8 | 5.675880e+01 |
| Tau | 3 | 8.888531e+02 |
| Dunn | 4 | 2.036000e-01 |
| Hubert | 0 | 0.000000e+00 |
| SDindex | 3 | 1.561500e+00 |
| Dindex | 0 | 0.000000e+00 |
| SDbw | 8 | 3.911000e-01 |




La mayoría de reglas sugiere que el número de clusters ideal es de tamaño 3.

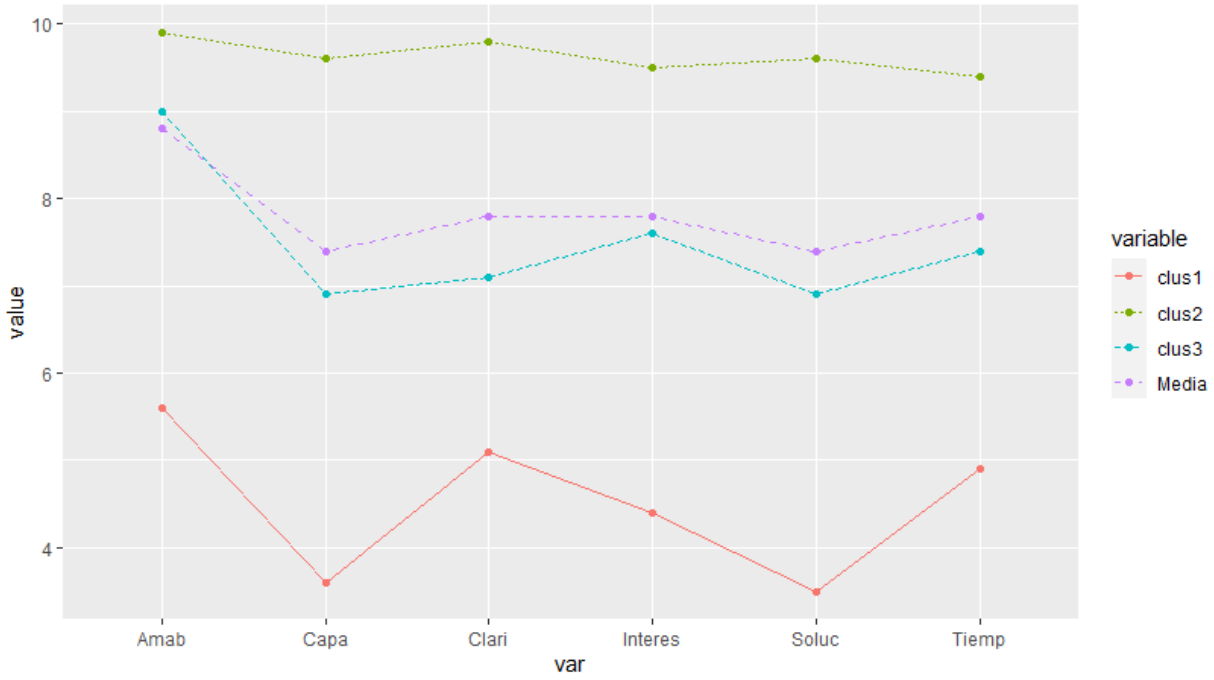
# Formación de los cluster

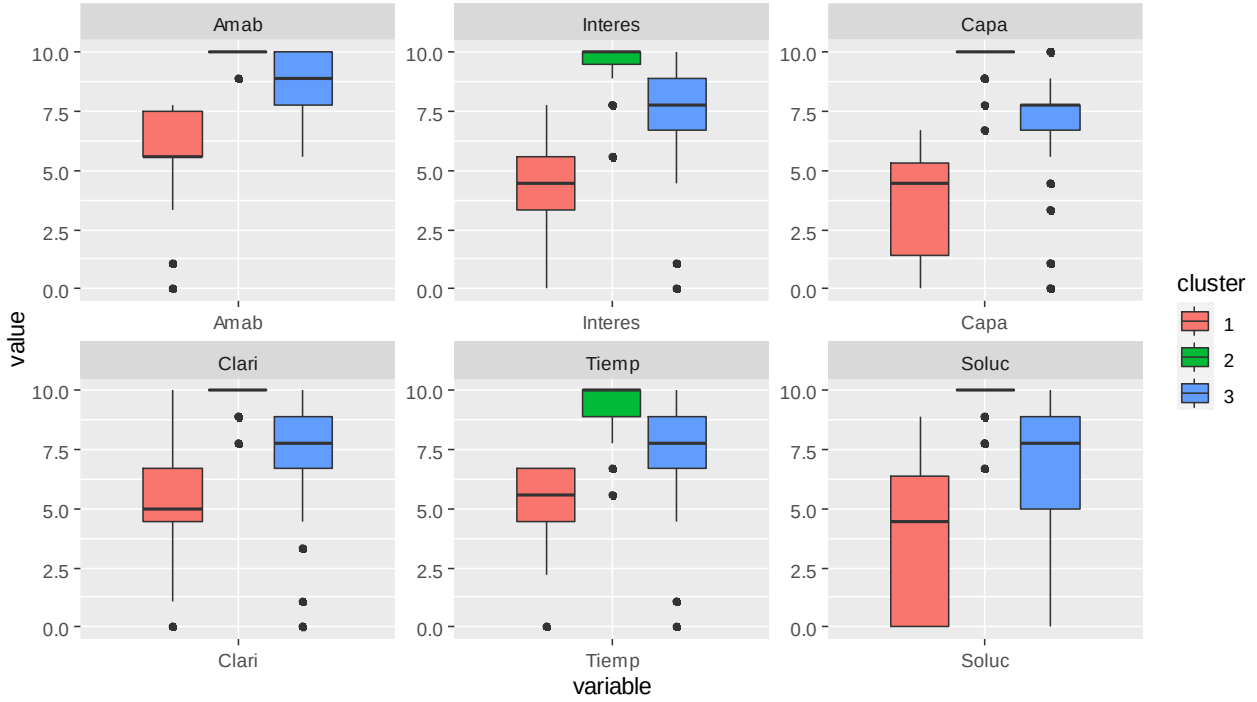



```
grp3
 1  2  3
18 43 43
```

En la gráfica cluster plot se observa la formación de los 3 cluster en dos dimensiones (el 63.6% de la variabilidad total se encuentra en al dimensión 1 y el 10.8% en la dimensión 2, el resto se podría representar en otra dimensión). Además se aprecia que los cluster 2 y 3 contiene la misma cantidad de individuos, mientras que el cluster 1 la menor cantidad.

# Características de los clusters



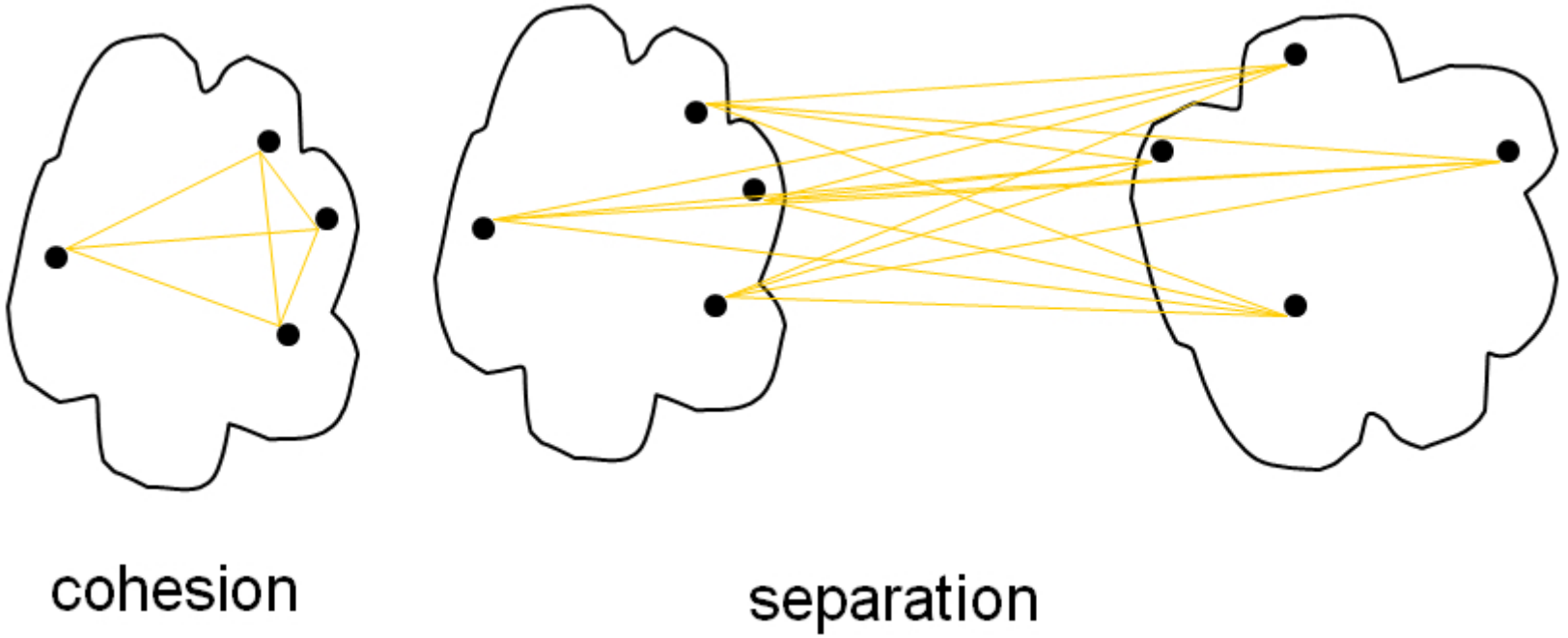




Se aprecia que los individuos que pertenecen al cluster 1 son los de menor rendimiento en todas las variables (a este grupo se les debería capacitar). El cluster 3 es un cluster intermedio, mientras que el cluster 2 son los de mejor rendimiento en todas las variables (a este grupo se les debe imitar).

# Validación de los clústeres

**Cohesión**
Un elemento de un clúster debe ser lo más cercano posible a los otros elementos del mismo clúster.

**Separación**
Los clústeres deben estar muy separados entre ellos.



# APLICACIÓN DEL CLÚSTER BASADO EN PARTICIONES

**DATA CALIDAD**

Es una base de datos que muestra el puntaje en calidad de atención que se le da a asesores comerciales después de su atención. Las variables y descripción son las siguientes:

| VARIABLE | DESCRIPCIÓN |
|---|---|
| ID | Código |
| Amab | Amabilidad de atención |
| Interes | Interés en el problema |
| Capa | Capaciad para resolver el problema |
| Clari | Claridad de información |
| Tiemp | Tiempo de atención |
| Soluc | Solución del problema |

# Validación para elegir el mejor método cluster

La validación se utiliza para diseñar el procedimiento de evaluación de bondad de los resultados del algoritmo de agrupamiento. Esto es importante para evitar encontrar patrones en datos aleatorios, así como en la situación en la que desea comparar diferentes algoritmos de agrupamiento.

## Medidas Internas

### Índice de Conectividad

Indica el grado de conexión de clústeres determinados por el vecino más cercano.

$$Conn = \sum_{i=1}^{N} \sum_{j=1}^{L} x_{i,nn_{i(j)}}, \quad 0 < Conn < \infty$$

### Índice de Dunn

Este índice compara las distancias inter grupos con el tamaño del grupo más disperso.

$$Dunn = \frac{\min_{1 \le p \le q \le k} dist(G_p, G_q)}{\max dist'(G_k)}, \quad 0 < Dunn < \infty$$

### Coeficiente de silueta

Mide qué tan bien se agrupa una observación y estima la distancia promedio entre los conglomerados.

# Validación para elegir el mejor método cluster

## Medidas de estabilidad

Una versión especial de medidas internas, que evalúa la consistencia de un resultado de agrupación comparándolo con los grupos obtenidos después de eliminar cada columna, uno a la vez.

Las medidas de estabilidad del clúster incluyen:

- **La proporción promedio de no superposición (APN)**
- **La distancia promedio (AD)**
- **La distancia promedio entre medias (ADM)**
- **La figura del mérito (FOM)**

Los valores de APN, ADM y FOM varían de 0 a 1, y el valor más pequeño corresponde a los resultados de agrupamiento altamente consistentes. AD tiene un valor entre 0 e infinito, y también se prefieren valores más pequeños.