BITMONEY

# HERRAMIENTA: CONGLOMERADO CLUSTER

BITMONEY

# Autores

**M.I.C. Carlos Abraham Carballo Monsivais**

**I.S.C. Leticia Edith Trujillo Ballesteros**

**C.L.M.A. Sacbe García García**

BITMONEY

Hackathon Blockchain 2020

Elaboración 2020, Primera edición

BITMONEY

# HERRAMIENTA. Análisis de conglomerados

## Contenido


4. Análisis de Conglomerados ..... 4
4.1. Introducción ..... 4
4.1.1. Antecedentes ..... 6
4.1.2. Aplicaciones ..... 6
4.2. Análisis de Conglomerados ..... 7
4.2.1. Esquema general de aplicación ..... 7
4.2.2. Objetivo y tipos de Análisis de Conglomerados ..... 7
4.2.3. Clasificación de técnicas ..... 8
4.2.4. Pasos para su aplicación ..... 9

BITMONEY

# 4. Análisis de Conglomerados

## 4.1. Introducción

Cuando se aborda el problema de clasificación de datos dado un determinado grupo de mediciones realizadas a observaciones o individuos bajo estudio existen dos perspectivas diferentes:

1) Se desea investigar si existen algunos grupos naturales o clases de individuos a partir de la estructura de las mediciones reportadas, o
2) Clasificar a los individuos según un conjunto de grupos existentes y ya definidos previamente.

El análisis de conglomerados es un término genérico que engloba una amplia gama de métodos numéricos para examinar datos multivariados respecto al primer caso, es decir, de descubrir grupos de observaciones que son homogéneos (comparten características similares de acuerdo a las variables bajo estudio) y están separados de otros grupos.

El análisis de conglomerados se aplica en muchos campos, como las ciencias naturales, las ciencias médicas, la economía, el marketing, etc. En medicina, por ejemplo, descubrir que una muestra de pacientes con mediciones en una variedad de características y síntomas en realidad consiste en un pequeño número de grupos dentro de los cuales estas características son relativamente similares, y entre los cuales son diferentes, podría tener implicaciones importantes tanto en términos de tratamiento futuro y para investigar la etiología de una condición. En marketing, es útil construir y describir los diferentes segmentos de un mercado a partir de una encuesta sobre consumidores potenciales. Por otro lado, una compañía de seguros podría estar interesada en la distinción entre clases de clientes potenciales para que pueda obtener precios óptimos por sus servicios. Más recientemente, las técnicas de análisis de conglomerados se han aplicado a al análisis de imágenes y búsqueda de patrones.

En psicología, el análisis de conglomerados se usa para encontrar tipos de personalidades en el

base de cuestionarios En la arqueología, se aplica para clasificar los objetos de arte en diferentes períodos de tiempo. Las técnicas de agrupamiento esencialmente intentan formalizar lo que los observadores humanos hacen bien en dos o tres dimensiones.

Los individuos que pertenecen a un determinado grupo (conglomerado o clúster) deben ser lo más homogéneos posible entre sí y las diferencias entre los diversos grupos lo más grandes posible. El análisis de conglomerados se puede dividir en dos pasos fundamentales:

1) Elección de una medida de proximidad: *Uno verifica cada par de observaciones (objetos) por la similitud de sus valores. Una medida de similitud (proximidad) se define para medir la "cercanía" de los objetos. Cuanto más "cerca" están, más homogéneos son.*
2) Elección del algoritmo de creación de grupos: *A partir de la proximidad, se establece una estrategia para asignar los objetos o individuos a los grupos de modo que las diferencias entre los grupos se vuelvan grandes y las observaciones en un grupo se vuelvan lo más parecidas posible*.

BITMONEY

De forma general existen dos tipos de procedimientos

a) Métodos jerárquicos
b) Métodos no jerárquicos

***Ejemplo: Ir de compras.***

Un reconocido Centro Comercial de la ciudad que alberga a más de 50 negocios reporta tener una baja afluencia de personas en el último semestre en comparación con el año pasado. Ante esta situación, se programó una reunión con los responsables de los negocios para fijar una estrategia que aumente la afluencia.

Se tiene por ***objetivo:*** Aumentar la frecuencia que las personas van al centro comercial a realizar sus compras.

Por lo tanto, se desea realizar un estudio sobre las motivaciones que llevan a las personas a ir de compras habitualmente y posteriormente establecer agrupaciones con motivaciones afines y establecer estrategias enfocadas a cada grupo identificado.

Algunas de las motivaciones pueden ser:

- Es un acto divertido en sí mismo.
- Intento ir poco porque compro compulsivamente y es malo para mi.
- Voy con frecuencia porque aprovecho para cenar fuera con mi pareja.
- Me encanta la aventura de encontrar productos a buen precio.
- No me atrae especialmente, voy por obligación o por necesidad.
- Puedes ahorrar mucho dinero si vas a comprar con frecuencia y estás informado.
- …

Así, es de interés formar grupos de personas con motivaciones similares. Para formar dichos grupos, se recomienda aplicar *un* ***Análisis de Conglomerados.***

*Análisis de conglomerados o cluster*: técnica multivariada que permite **CLASIFICAR** una muestra de observaciones (individuos u objetos) en un número pequeño de *grupos* o *clústers* *mutuamente excluyentes* basado en las similitudes que hay entre las observaciones (individuos u objetos de interés). Esto se realiza agrupando a los individuos que son similares siguiendo algún criterio apropiado.

Se ***DESCONOCE***, al inicio del estudio, el número de grupos y las características de las personas que conforman cada uno de ellos.

BITMONEY

**Clasificar** = Asignar un nuevo objeto u observación en su lugar correspondiente dentro de un conjunto de categorías establecidas.

### 4.1.1. Antecedentes

- ✓ Desde inicios de la civilización la *clasificación* es una actividad básica.
- ✓ Primeros trabajos de clasificación se dan en la Biología al establecer una clasificación de las especies del reino animal (carne roja y los que no).
- ✓ El Análisis de Conglomerados se formalizó con la publicación de "*Principios de Taxonomía Numérica*" en 1963.
- ✓ El rápido crecimiento en años recientes se debe a:

1. *El desarrollo de las computadoras*
2. *La importancia fundamental de la clasificación en todos los campos (la ciencia)*

### 4.1.2. Aplicaciones

Los roles más comunes que el Análisis de Conglomerados desempeña son:

- ***REDUCCIÓN DE DATOS***: Al hacer frente a un gran volumen de observaciones sin sentido y no agrupadas es necesario contar con procedimientos objetivos para reducir la información mediante agrupaciones o clústeres.
- ***GENERACION DE HIPOTESIS***: Es probable que se desee desarrollar alguna hipótesis acerca de la naturaleza de un dato o característica de los individuos u objetos bajo estudio.

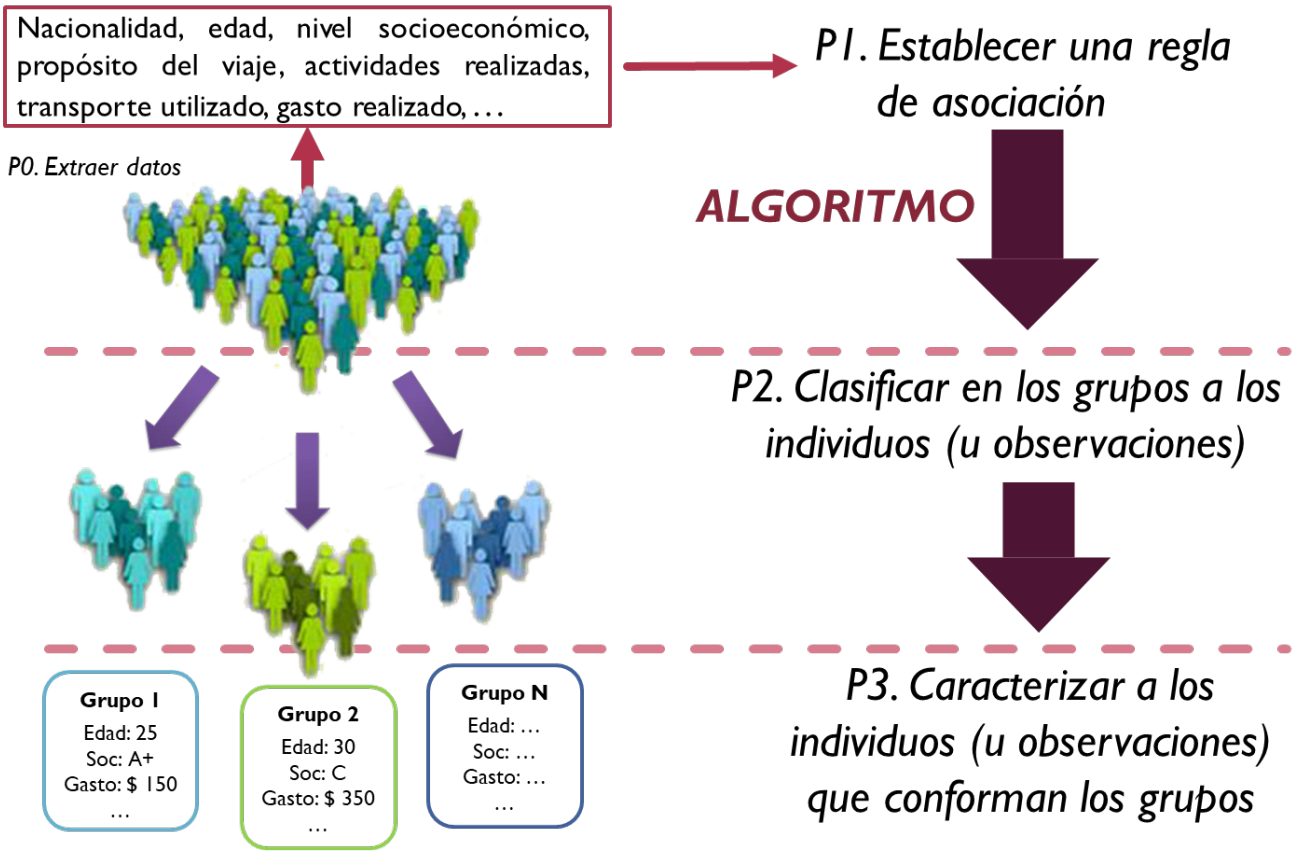
*Ejemplo*: Existe la creencia que considerando el ingreso de cada persona es posible establecer segmentos o grupos lógicos de consumidores de bebidas dietéticas.

El Análisis de Conglomerados ha sido aplicado en:

- ❑ ***Biología***: Taxonomía
- ❑ ***Psicología***: Q-análisis
- ❑ ***Ingeniería***: Reconocimiento de patrones, caracterización de proveedores
- ❑ ***Salud***: Zona de atención en desastres
- ❑ ***Economía***: Perfiles sociodemográficos
- ❑ ***Mercadotecnia***: Segmentación de mercado

## 4.2. Análisis de Conglomerados

### 4.2.1. Esquema general de aplicación




### 4.2.2. Objetivo y tipos de Análisis de Conglomerados

*Objetivo:* Considerando los datos de la MUESTRA sobre los objetos bajo estudio (personas, productos, ideas, etc), CLASIFICAR en un CONJUNTO REDUCIDO DE GRUPOS (conglomerados, clústers o clases) dichos objetos.

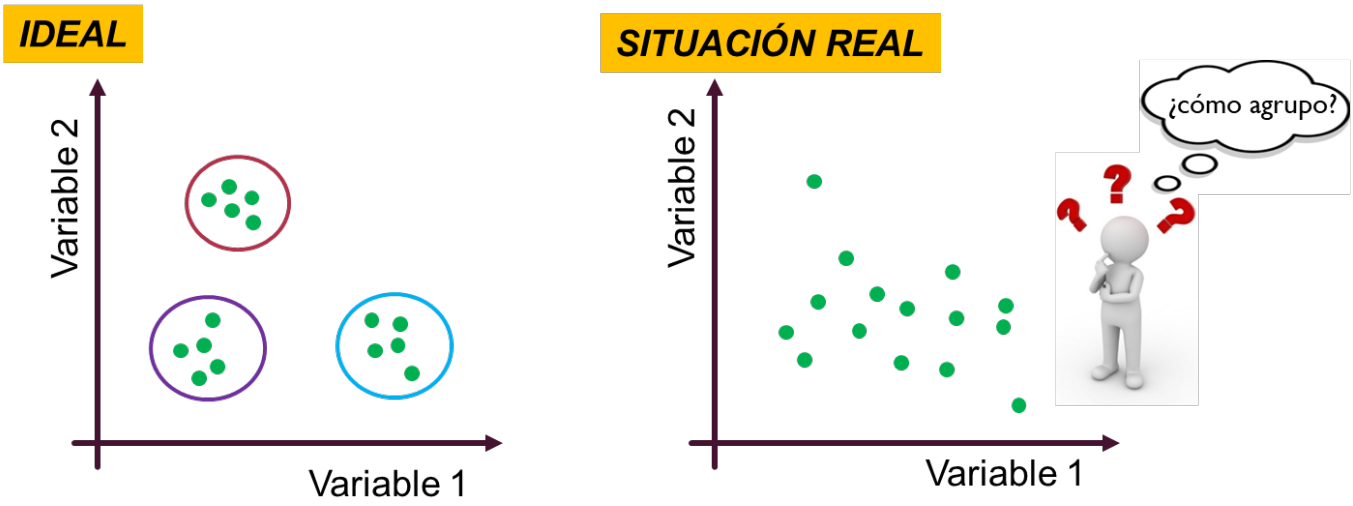
Al Análisis de Conglomerados (AC) es comparable con el *Análisis Factorial* en el sentido de que busca *evaluar la estructura* de los datos de la muestra. Pero difiere en:

- *El AF busca la agrupación de variables, y*
- *El AF hace la agrupación basado en patrones de variación (correlaciones) en datos mientras que el A. Conglomerados agrupa mediante proximidades (distancia).*

BITMONEY

Las agrupaciones resultantes deben exhibir una *alta homogeneidad interna* (dentro del grupo los individuos tienen características muy similares) y una *muy alta heterogeneidad externa* (entre individuos de grupos diferentes se aprecian diferencias significativas).

***Postulado***: Dada una colección de ***n*** objetos, individuos, animales, etc., los cuales se describen por un conjunto de ***p*** características o variables; producir una ***DIVISIÓN ÚTIL*** en cierto número de clases. El número de clases y sus características serán determinadas por el analista.




*¿Si se agrega otra(s) variable(s) será más evidente el agrupamiento?*

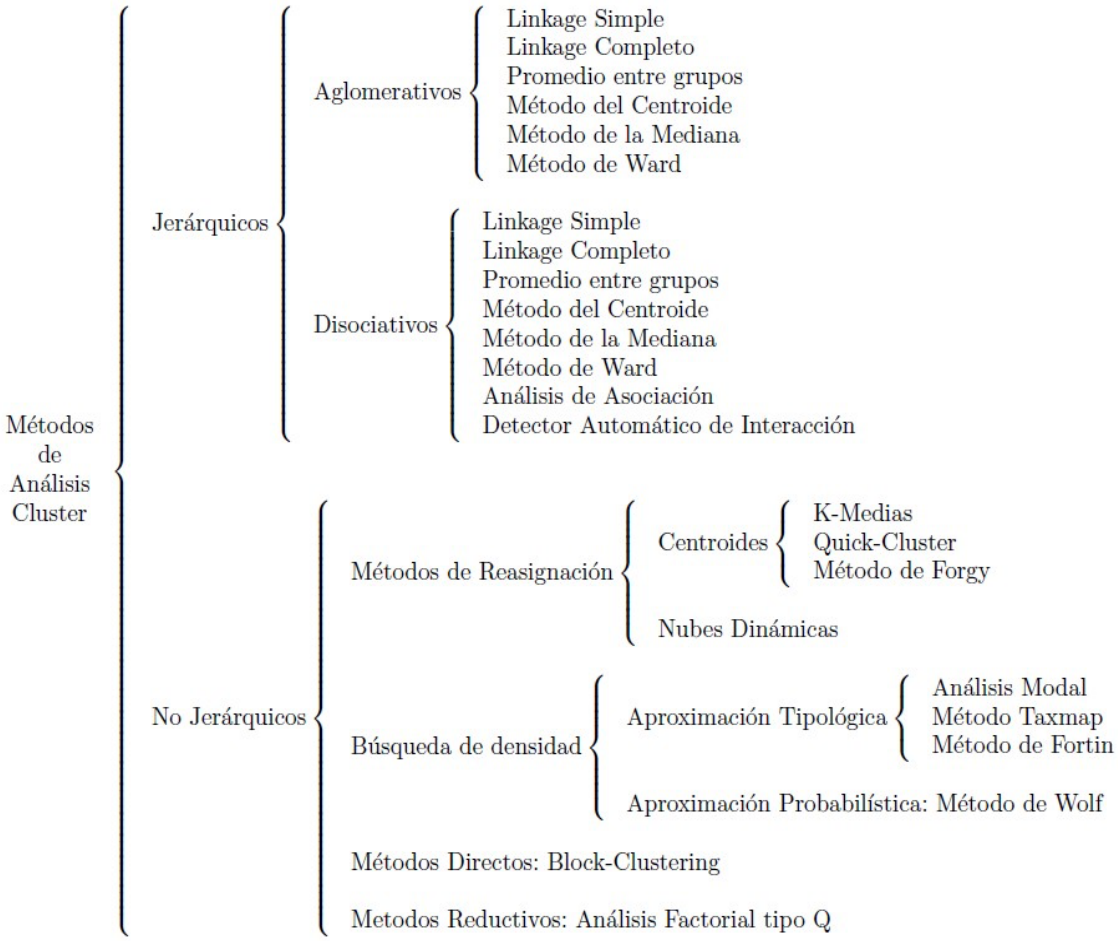
4.2.3. Clasificación de técnicas

Todos los métodos comparten dos pasos principales:

1. *La selección de la medida de proximidad (similitud o distancia)*
2. *La selección del algoritmo de agrupamiento*

Se distinguen dos grandes categorías:

- ***Jerárquicos***: Agrupan clústers para formar uno nuevo o bien separar alguno existente de tal forma que minimice alguna función de distancia o maximice alguna medida de similitud.
- ***No jerárquicos*** (*partitivos o de optimización*): Realizan una sola partición de los individuos en *K* grupos. Se debe especificar *a priori* los grupos que deben ser formados. Difiere de la categoría anterior ya que trabaja con la matriz de los datos originales y no requiere una conversión en matriz de distancia o similitud.




### 4.2.4. Pasos para su aplicación

#### 4.2.4.1. Métodos jerárquicos

***A) Selección de variables***

- Selección de un conjunto concreto de características usadas para describir a cada individuo que sirva de marco de referencia para establecer las agrupaciones. Refleja la opinión del investigador acerca del propósito de la clasificación.

- Una gran cantidad de variables puede ocasionar problemas y dificultar la identificación de la estructura de los grupos. Es posible utilizar previamente un Análisis de Componentes Principales o Análisis de Factores para reducir la dimensionalidad.

- El *tipo de variable* y las *unidades* de medición influyen en la forma de tratar esos datos para generar las agrupaciones. Es recomendable trabajar con *datos transformados*, para eliminar el efecto de la escala, además facilitar la interpretación.

***B) Escoger la medida de asociación***

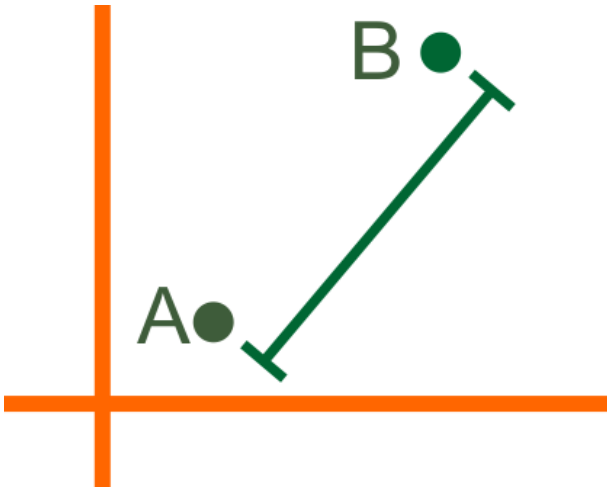
Medir la *proximidad* o *similitud* de los objetos en estudio puede expresarse en forma de ***DISTANCIA***. Aquel par de observaciones que tengan más características en común, tendrán una distancia más corta. Una distancia más grande indica poca similitud entre el par de objetos considerando las variables seleccionadas previamente.

Antes de definir medidas de distancia, recuerde que muchas de las técnicas analíticas son particularmente sensitivas a los *outliers*. Por lo tanto, existen algunos chequeos preliminares para *outliers* y errores de dedo como el gráfico de dispersión, el diagrama de cajas, …

Matemáticamente se da el nombre de distancia entre dos puntos (A, B), a toda medida que verifique los axiomas siguientes:

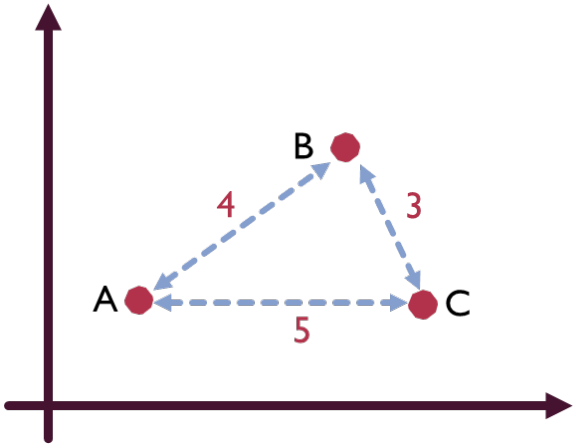
1. La distancia del punto A hacia B es positiva, $d(A, B) \geq 0$
2. La distancia hacia un mismo punto es cero, $d(A, A) = 0$
3. La distancia del punto A hacia B es la misma que si se parte del punto B hacia el punto A, $d(A, B) = d(B, A)$
4. La distancia del punto A hacia el punto B es menor o igual que ir del punto A al punto B pasando por un punto C,

   $d(A, B) \leq d(A, C) + d(C, B)$



BITMONEY

Ejemplo:




| | **Distancia (A, B)** | **Distancia (B, C)** |
|---|---|---|
| 1. | $d(A, B) = 4 \therefore d(A, B) \geq 0$ ✓ | $d(B,C) = 3 \therefore d(A, B) \geq 0$ ✓ |
| 2. | $d(A, A) = 0$ y $d(B, B) = 0$ ✓ | $d(B, B) = 0$ y $d(C, C) = 0$ ✓ |
| 3. | $d(A, B) = 4$ y $d(B, A) = 4$ ✓ | $d(B, C) = 3$ y $d(C, B) = 3$ ✓ |
| 4. | $d(A, B) = 4, d(A, C) = 5, d(C, B) = 3$ | $d(B, C) = 3, d(B, A) = 4, d(A, C) = 5$ ✓ |

$\underbrace{d(A, B)}_{4} \leq \underbrace{d(A, C) + d(C, B)}_{8}$ ✓ $\underbrace{d(B, C)}_{3} \leq \underbrace{d(B, A) + d(A, C)}_{9}$ ✓

Existen diferentes medidas de asociación (formas de medir distancia) para variables cuantitativas:

- **Distancia Euclidiana.** Este tipo de distancia es probablemente el más usado. Se calcula así:

$$d(x, y) = \sqrt{\sum_i (x_i - y_i)^2}$$

BITMONEY

- **Distancia Euclidiana Cuadrada.** Uno puede desear elevar al cuadrado la Distancia Euclidiana Standard para ponderar progresivamente más objetos que están más lejos. Esta distancia se calcula así:

$$d(x,\ y) = \sum_i (x_i - y_i)^2$$

- **Distancia City-block (Manhattan).** Esta distancia es simplemente el promedio de las diferencias a lo largo de las dimensiones. En la mayoría de los casos, se obtienen resultados similares que con el método de la distancia Euclidiana. La distancia city-block se calcula así:

$$d(x,\ y) = \sum_i |x_i - y_i|$$

- **Distancia de Chebychev.** Esta distancia puede ser apropiada en casos cuando uno quiere definir si son diferentes en alguna dimensión. La distancia de Chebychev se calcula así:

$$d(x,\ y) = \max|x_i - y_i|$$

- **Distancia potencia.** Algunas veces uno puede desear incrementar o disminuir progresivamente el peso que se coloca en las dimensiones en los cuales los respectivos objetos son muy diferentes. Esto se puede lograr vía la *distancia potencia*. La distancia se calcula así:

$$d(x,\ y) = \sqrt[r]{\sum_i |x_i - y_i|^p}$$

- **Porcentaje de desacuerdo.** Esta distancia es particularmente útil si los datos incluidos en el análisis son de naturaleza categóricos. Esta distancia se calcula de la siguiente manera:

  El propósito es construir una *matriz de las distancias iniciales* a partir de cada par de individuos de la muestra considerando las variables seleccionadas para el estudio.

$$d(x,\ y) = \frac{(número\ de\ x_i \neq y_i)}{i}$$

BITMONEY

**EJEMPLO:** Obtener la matriz de distancia inicial usando primero la distancia euclidiana y en un segundo intento usando la distancia de Manhattan para la tabla de datos siguientes que representa 7 casos y 5 variables medidas:

| **Original Data** | | | | | |
|---|---|---|---|---|---|
| **Case** | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ |
| 1 | 7 | 10 | 9 | 7 | 10 |
| 2 | 9 | 9 | 8 | 9 | 9 |
| 3 | 5 | 5 | 6 | 7 | 7 |
| 4 | 6 | 6 | 3 | 3 | 4 |
| 5 | 1 | 2 | 2 | 1 | 2 |
| 6 | 4 | 3 | 2 | 3 | 3 |
| 7 | 2 | 4 | 5 | 2 | 5 |

*Distancia Euclidiana:*

$$d(x,\ y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + (x_3 - y_3)^2 + (x_4 - y_4)^2 + (x_5 - y_5)^2}$$

$$d(2,\ 1) = \sqrt{(9-7)^2 + (9-10)^2 + (8-9)^2 + (9-7)^2 + (9-10)^2} = \mathbf{3.32}$$

$$d(7,\ 2) = \sqrt{(2-9)^2 + (4-9)^2 + (5-8)^2 + (2-9)^2 + (5-9)^2} = \mathbf{12.17}$$

$$d(5,\ 4) = \sqrt{(1-6)^2 + (2-6)^2 + (2-3)^2 + (1-3)^2 + (2-4)^2} = \mathbf{7.07}$$

| Caso | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| 1 | 0.00 | 3.32 | 6.86 | 10.25 | 15.78 | 13.11 | 11.27 |
| 2 | 3.32 | 0.00 | 6.63 | 10.20 | 16.19 | 13.00 | 12.17 |
| 3 | 6.86 | 6.63 | 0.00 | 6.00 | 10.10 | 7.28 | 6.32 |
| 4 | 10.25 | 10.20 | 6.00 | 0.00 | 7.07 | 3.87 | 5.10 |
| 5 | 15.78 | 16.19 | 10.10 | 7.07 | 0.00 | 3.87 | 4.90 |
| 6 | 13.11 | 13.00 | 7.28 | 3.87 | 3.87 | 0.00 | 4.36 |
| 7 | 11.27 | 12.17 | 6.32 | 5.10 | 4.90 | 4.36 | 0.00 |

El resultado es una matriz simétrica que contiene todas las distancias de todos los posibles pares de observaciones. De la matriz anterior se observa que los caso 1 y 2 son el par con mayor similitud ya que tiene la distancia más corta. La pareja con mayor disimilitud se haya con los casos 2 y 5 ya que su distancia es la mayor.

Al observar el caso 7, se puede concluir que el orden de similitud con respecto a los demás casos es el siguiente: 6, 5, 4, 3, 1, 2.

*Distancia de Manhattan:*

$$d(x,\ y) = |x_1 - y_1| + |x_2 - y_2| + |x_3 - y_3| + |x_4 - y_4| + |x_5 - y_5|$$

$$d(2,\ 1) = |9-7| + |9-10| + |8-9| + |9-7| + |9-10| = \mathbf{7}$$

BITMONEY

$$d(7,\ 2) = |2-9| + |4-9| + |5-8| + |2-9| + |5-10| = \mathbf{26}$$

$$d(5,\ 4) = |1-6| + |2-6| + |2-3| + |1-3| + |2-4| = \mathbf{14}$$

Matriz de distancias resultante

| Caso | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| 1 | 0.00 | 7.00 | 13.00 | 21.00 | 35.00 | 28.00 | 25.00 |
| 2 | 7.00 | 0.00 | 14.00 | 22.00 | 36.00 | 29.00 | 26.00 |
| 3 | 13.00 | 14.00 | 0.00 | 12.00 | 22.00 | 15.00 | 12.00 |
| 4 | 21.00 | 22.00 | 12.00 | 0.00 | 14.00 | 7.00 | 10.00 |
| 5 | 35.00 | 36.00 | 22.00 | 14.00 | 0.00 | 7.00 | 10.00 |
| 6 | 28.00 | 29.00 | 15.00 | 7.00 | 7.00 | 0.00 | 9.00 |
| 7 | 25.00 | 26.00 | 12.00 | 10.00 | 10.00 | 9.00 | 0.00 |

*Para cualitativas dicotómicas (dos categorías):*

A partir de la tabla de contingencia de 2 x 2

| | | **Variable $X_i$** | | |
|---|---|---|---|---|
| | | Presente (1) | Ausencia (0) | Total |
| **Variable $X_j$** | Presente (1) | a | b | a+b |
| | Ausencia (0) | c | d | c+d |
| | Total | a+c | b+d | N=a+b+c+d |

**a** = # de individuos que toman el valor de 1 en cada variable de forma simultánea.

**b** = # de individuos que toman el valor de 0 en $X_i$ y 1 en $X_j$.

**c** = # de individuos que toman el valor de 1 en $X_i$ y 0 en $X_j$.

**d** = # de individuos que toman el valor de 0 en cada variable de forma simultánea.

BITMONEY

Así, para dichas variables categóricas se tienen las siguientes medidas de similitud:

| | Similitud o similaridad | Disimilaridad |
|---|---|---|
| Russel y Rao | $RR = \frac{a}{N}$ | $\frac{N-a}{N}$ |
| Parejas simples | $PS = \frac{a+d}{N}$ | $\frac{N-(a+d)}{N}$ |
| Jaccard | $J = \frac{a}{a+b+c}$ | $\frac{b+c}{a+b+c}$ |
| Dice y Sorensen | $D = \frac{2a}{2a+b+c}$ | $\frac{b+c}{2a+b+c}$ |

***Ejemplo***: La siguiente tabla reúne la presencia/ausencia de 6 especímenes de bacterias en 7 lagos donde 1 indica presencia del espécimen y 0 indica ausencia.

| Lago | Especie 1 | Especie 2 | Especie 3 | Especie 4 | Especie 5 | Especie 6 |
|---|---|---|---|---|---|---|
| La Irene | 1 | 1 | 1 | 1 | 1 | 1 |
| Pedro Luro | 0 | 0 | 0 | 1 | 1 | 1 |
| Loncoche | 1 | 0 | 0 | 0 | 0 | 0 |
| Lefipan | 1 | 1 | 1 | 1 | 0 | 1 |
| Salamanca | 1 | 1 | 1 | 1 | 1 | 1 |
| Dorotea | 0 | 0 | 0 | 0 | 0 | 1 |
| Paso del sapo | 0 | 0 | 1 | 0 | 0 | 1 |

**Paso 1** Generar la tabla de contingencia de entre dos pares de lagos. Se elige los dos primeros registros de la tabla

| Lago | E 1 | E 2 | E 3 | E 4 | E 5 | E 6 |
|---|---|---|---|---|---|---|
| La Irene | 1 | 1 | 1 | 1 | 1 | 1 |
| Pedro Luro | 0 | 0 | 0 | 1 | 1 | 1 |

| | | La Irene | | |
|---|---|---|---|---|
| | | Presente (1) | Ausente (0) | Total |
| Pedro Luro | Presente (1) | 3 | 0 | 3 |
| | Ausente (0) | 3 | 0 | 3 |
| | Total | 6 | 0 | 6 |

**Paso 2** Seleccionar y calcular el indicador de medida de similitud. En este ejemplo se emplea el índice de Jaccard.

$$J = \frac{a}{a+b+c} = \frac{3}{3+0+3} = 0.500$$

BITMONEY

**Paso 3** Repetir paso 1 y 2 con todos los posibles emparejamientos.

| Lago | E 1 | E 2 | E 3 | E 4 | E 5 | E 6 |
|---|---|---|---|---|---|---|
| Salamanca | 1 | 1 | 1 | 1 | 1 | 1 |
| Dorotea | 0 | 0 | 0 | 0 | 0 | 1 |

Dorotea

| | Salamanca | | |
|---|---|---|---|
| | Presente (1) | Ausente (0) | Total |
| Presente (1) | 1 | 0 | 1 |
| Ausente (0) | 5 | 0 | 5 |
| Total | 6 | 0 | 6 |

$$J = \frac{a}{a+b+c} = \frac{1}{1+0+5} = 0.167$$

**Paso 4** Construir la matriz de similitud con todos los índices calculados

La matriz de distancias resultante

| | La Irene | Pedro Luro | Loncoche | Lefipan | Salamanca | Cerro Dorotea | Paso del sapo |
|---|---|---|---|---|---|---|---|
| La Irene | 1.000 | 0.500 | 0.167 | 0.833 | 1.000 | 0.167 | 0.333 |
| Pedro Luro | 0.500 | 1.000 | 0.000 | 0.333 | 0.500 | 0.333 | 0.250 |
| Loncoche | 0.167 | 0.000 | 1.000 | 0.200 | 0.167 | 0.000 | 0.000 |
| Lefipan | 0.833 | 0.333 | 0.200 | 1.000 | 0.833 | 0.200 | 0.400 |
| Salamanca | 1.000 | 0.500 | 0.167 | 0.833 | 1.000 | 0.167 | 0.333 |
| Cerro Dorotea | 0.167 | 0.333 | 0.000 | 0.200 | 0.167 | 1.000 | 0.500 |
| Paso del sapo | 0.333 | 0.250 | 0.000 | 0.400 | 0.333 | 0.500 | 1.000 |

***Transformación de valores:***

Es importante recordar que debido a que el análisis de conglomerados emplea medidas de distancia, estas son muy sensibles a las diferencias de escala o magnitudes de las variables.

La forma más común de transformación es la conversión de cada variable en su puntuación estándar (Z-score):

$$Z = \frac{x_i - \bar{x}}{S}$$

De esta forma se convierte los valores en puntuaciones estandarizadas con media = 0 y desviación estándar = 1, eliminado el sesgo originado por la diferencia de escalas.

Otras formas de transformación son:

a) **rango -1 a 1,** los valores originales son divididos entre el rango de cada variable o caso,

b) **rango 0 a 1**, se obtiene restando el mínimo y dividiendo por el rango de cada variable o caso,

BITMONEY

c) **magnitud máxima de 1**, se obtiene dividiendo los valores originales por el máximo de cada variable o caso, según corresponda al análisis,

d) **media de 1**, se dividen los valores originales entre la media de cada variable o caso,

**desviación típica 1,** se obtiene dividiendo los valores originales por la desviación típica de cada variable o caso.

*Encuesta en la ciudad de Aguascalientes entre la población económicamente inactiva*

| EDAD | ESTUDIANTES | HOGAR | JUBILADO | INCAP | OTRO | TOTAL |
|---|---|---|---|---|---|---|
| 12 - 14 | 39,949 | 5,988 | 6 | 78 | 3,059 | 49,080 |
| 15 - 19 | 28,636 | 16,253 | 22 | 178 | 5,398 | 50,487 |
| 20 - 24 | 6,580 | 19,582 | 27 | 225 | 2,896 | 29,310 |
| 25 - 29 | 834 | 19,180 | 32 | 196 | 1,667 | 21,909 |
| 30 - 34 | 188 | 17,461 | 57 | 174 | 1,198 | 19,078 |
| 35 - 39 | 68 | 14,446 | 109 | 143 | 1,002 | 15,768 |
| 40 - 44 | 44 | 11,698 | 184 | 135 | 800 | 12,861 |
| 45 - 49 | 29 | 9,725 | 339 | 158 | 794 | 11,045 |
| 50 - 54 | 14 | 7,915 | 579 | 149 | 735 | 9,392 |
| 55 - 59 | 10 | 6,773 | 820 | 145 | 742 | 8,490 |
| 60 - 64 | 13 | 6,152 | 1,311 | 213 | 819 | 8,508 |
| 65 Y MAS | 27 | 12,867 | 3,843 | 1,411 | 4,020 | 22,168 |

Como se aprecia en la tabla anterior, aquellas variables con una escala de medición mucho mayor tendrán una mayor influencia en el cálculo de las distancias entre pares de observaciones. Por lo tanto, es recomendable hacer una transformación de los datos.

***C) Elección de la técnica de agrupación***

La elección del método de agrupación será relativamente natural dependiendo de la naturaleza de los datos usados y de los objetivos perseguidos.

Se recomienda probar varias técnicas y contrastar los resultados obtenidos con cada una de ellas. Si los resultados finales son parecidos, las conclusiones son mucho más validas sobre la estructura natural de los datos. En caso contrario puede plantearse el hecho que tal vez los datos utilizados no obedezcan a una estructura bien definida.

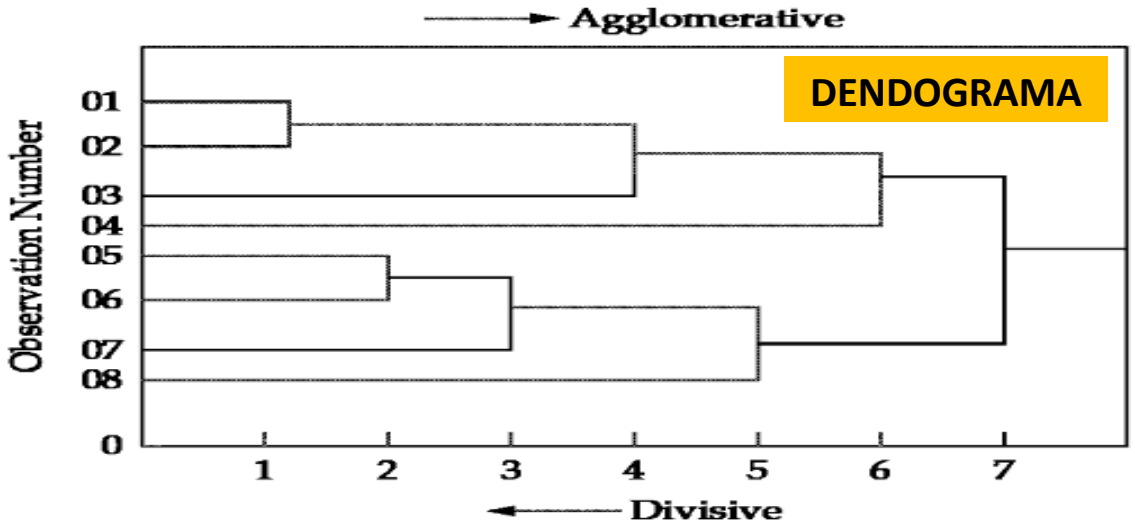
Los métodos jerárquicos permiten la construcción de un árbol de clasificación, que recibe el nombre de *DENDROGRAMA*. El dendrograma:

- Permite *visualizar el proceso de agrupamiento de los clústeres* en los distintos pasos.

BITMONEY

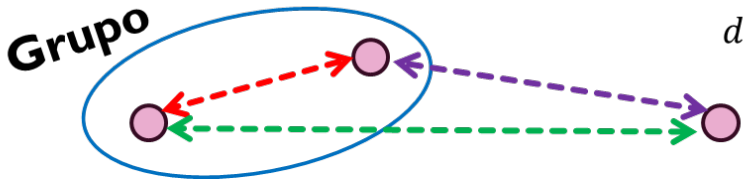
- Ayuda a *decidir el número de grupos* que representan mejor la estructura de los datos considerando la forma en que se van anidando los clústeres y la medida de similitud a la cual lo hacen.
- Permite al investigador "*seguir la pista*" de formación de los distintos clústeres, que van englobándose o anidándose, hasta resumirse en sólo uno.

  En biología a menudo es de mayor interés desvelar las distintas categorías en que van clasificándose los individuos estudiados, desde los grupos más particulares a los más generales.




Algunas de las técnicas de agrupación más empleadas son:

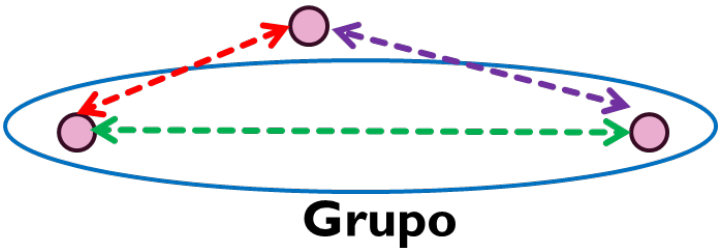
- **Vecino más cercano** (*linkage simple*): Agrupa a los individuos con la *distancia* o *similaridad* más próxima.




$$d(A,B) = \min_{i \in A, j \in B} d(i,j)$$

- **Vecino más lejano** (*linkage completo*): Agrupa a los individuos con la *distancia* o *similaridad* más lejana.

BITMONEY




$$d(A,B) = \max_{i \in A, j \in B} d(i,j)$$

**Ejemplo:** A partir de la siguiente matriz de distancia emplear las técnicas de agrupación anteriores y obtener el dendograma respectivo.

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{pmatrix} 0 & & & & \\ 2 & 0 & & & \\ 6 & 5 & 0 & & \\ 10 & 9 & 4 & 0 & \\ 9 & 8 & 5 & 3 & 0 \end{pmatrix} \end{array}$$

**Vecino más cercano**

**Paso 1.** Seleccionar la distancia más corta de la matriz para formar un grupo. De la matriz anterior se observa que es la *d*(2, 1) = 2, por lo tanto la observación 1 y 2 forman un clúster.

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{pmatrix} 0 & & & & \\ 2 & 0 & & & \\ 6 & 5 & 0 & & \\ 10 & 9 & 4 & 0 & \\ 9 & 8 & 5 & 3 & 0 \end{pmatrix} \end{array}$$

**Paso 2.** Calcular la nueva matriz de distancias. Las distancias de los individuos que no fueron agrupados no varían, solamente aquellas que tienen relación con los individuos que son agrupados.

$$d([1,2],3) = \min\big(d(1,3),\ d(2,3)\big) = \min(6,\ 5) = 5$$

$$d([1,2],4) = \min\big(d(1,4),\ d(2,4)\big) = \min(10,\ 9) = 9$$

$$d([1,2],5) = \min\big(d(1,5),\ d(2,5)\big) = \min(9,\ 8) = 8$$

BITMONEY

| | (1-2) | 3 | 4 | 5 |
|---|---|---|---|---|
| (1-2) | 0 | | | |
| 3 | | 0 | | |
| 4 | | 4 | 0 | |
| 5 | | 5 | 3 | 0 |

| | (1-2) | 3 | 4 | 5 |
|---|---|---|---|---|
| (1-2) | 0 | | | |
| 3 | 5 | 0 | | |
| 4 | 9 | 4 | 0 | |
| 5 | 8 | 5 | 3 | 0 |

**Paso 3**. Repetir los pasos 1 y 2 hasta agrupar a todos los individuos en un único grupo.

| | (1-2) | 3 | 4 | 5 |
|---|---|---|---|---|
| (1-2) | 0 | | | |
| 3 | 5 | 0 | | |
| 4 | 9 | 4 | 0 | |
| 5 | 8 | 5 | 3 | 0 |

| | (1-2) | 3 | (4-5) |
|---|---|---|---|
| (1-2) | 0 | | |
| 3 | 5 | 0 | |
| (4-5) | | | 0 |

$$d([1,2],[4,5]) = \min\big(d([1,2],4), d([1,2],5)\big) = \min(9,8) = 8$$
$$d(3,[4,5]) = \min\big(d(3,4), d(3,5)\big) = \min(4,5) = 4$$

| | (1-2) | 3 | (4-5) |
|---|---|---|---|
| (1-2) | 0 | | |
| 3 | 5 | 0 | |
| (4-5) | 8 | 4 | 0 |

| | (1-2) | 3 | (4-5) |
|---|---|---|---|
| (1-2) | 0 | | |
| 3 | 5 | 0 | |
| (4-5) | 8 | 4 | 0 |

| | (1-2) | (3-4-5) |
|---|---|---|
| (1-2) | 0 | |
| (3-4-5) | | 0 |

$$d([1,2],[3,4,5])$$
$$= \min\big(d([1,2],3), d([1,2],[4,5])\big)$$
$$= \min(5,8) = 5$$

| | (1-2) | (3-4-5) |
|---|---|---|
| (1-2) | 0 | |
| (3-4-5) | 5 | 0 |

**Vecino más cercano**

**Paso 1.** Seleccionar la distancia más corta de la matriz para formar un grupo. De la matriz anterior se observa que es la *d*(2, 1) = 2, por lo tanto la observación 1 y 2 forman un clúster.

BITMONEY

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{pmatrix} 0 & & & & \\ 2 & 0 & & & \\ 6 & 5 & 0 & & \\ 10 & 9 & 4 & 0 & \\ 9 & 8 & 5 & 3 & 0 \end{pmatrix} \end{array}$$

**Paso 2.** Calcular la nueva matriz de distancias. Al igual que en el método anterior, las distancias de los individuos que no fueron agrupados no varían, solamente aquellas que tienen relación con los individuos que son agrupados.

| | (1-2) | 3 | 4 | 5 |
|---|---|---|---|---|
| (1-2) | 0 | | | |
| 3 | | 0 | | |
| 4 | | 4 | 0 | |
| 5 | | 5 | 3 | 0 |

$$d([1,2],3) = \max\big(d(1,3), d(2,3)\big) = \max(6,5) = 6$$
$$d([1,2],4) = \max\big(d(1,4), d(2,4)\big) = \max(10,9) = 10$$
$$d([1,2],5) = \max\big(d(1,5), d(2,5)\big) = \max(9,8) = 9$$

| | (1-2) | 3 | 4 | 5 |
|---|---|---|---|---|
| (1-2) | 0 | | | |
| 3 | 6 | 0 | | |
| 4 | 10 | 4 | 0 | |
| 5 | 9 | 5 | 3 | 0 |

**Paso 3**. Repetir los pasos 1 y 2 hasta agrupar a todos los individuos en un único grupo.

| | (1-2) | 3 | 4 | 5 |
|---|---|---|---|---|
| (1-2) | 0 | | | |
| 3 | 6 | 0 | | |
| 4 | 10 | 4 | 0 | |
| 5 | 9 | 5 | 3 | 0 |

| | (1-2) | 3 | (4-5) |
|---|---|---|---|
| (1-2) | 0 | | |
| 3 | 6 | 0 | |
| (4-5) | | | 0 |

$$d([1,2],[4,5]) = \max\big(d([1,2],4), d([1,2],5)\big) = \max(10,9) = 10$$
$$d(3,[4,5]) = \max\big(d(3,4), d(3,5)\big) = \max(4,5) = 5$$

BITMONEY

| | (1-2) | 3 | (4-5) |
|---|---|---|---|
| (1-2) | 0 | | |
| 3 | 6 | 0 | |
| (4-5) | 10 | 5 | 0 |

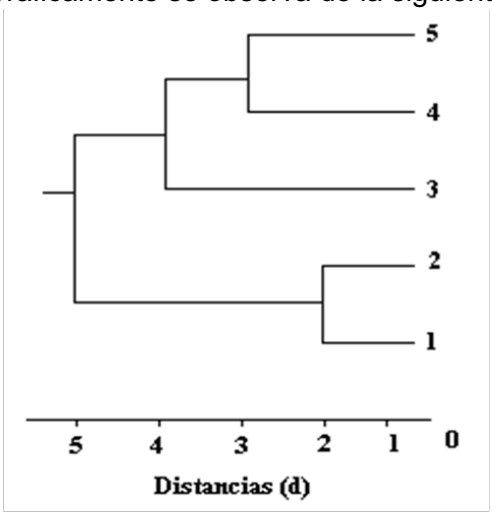
| | (1-2) | 3 | (4-5) |
|---|---|---|---|
| (1-2) | 0 | | |
| 3 | 6 | 0 | |
| (4-5) | 10 | 5 | 0 |

$$d([1,2],[3,4,5]) = \max\big(d([1,2],3), d([1,2],[4,5])\big)$$
$$= \max(6,10) = 10$$

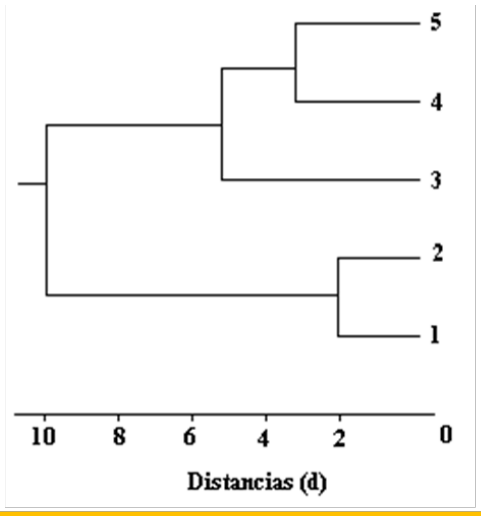
| | (1-2) | (3-4-5) |
|---|---|---|
| (1-2) | 0 | |
| (3-4-5) | | 0 |

| | (1-2) | (3-4-5) |
|---|---|---|
| (1-2) | 0 | |
| (3-4-5) | 10 | 0 |

Gráficamente se observa de la siguiente manera:




**Vecino más cercano**

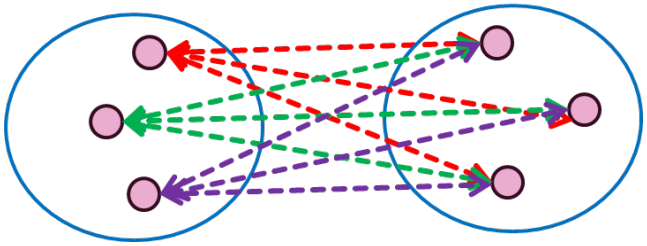



**Vecino más lejano**

La distancia (d) en ambos dendogramas indica el valor en que los individuos se fueron agrupando. En el caso del vecino más cercano la agrupación total ocurrió a una distancia de 5 unidades mientras que en el vecino más lejano se alcanzó hasta la distancia de 10 unidades.

BITMONEY

- **Agrupamiento promedio:** Promedio de las distancias entre todos los pares de individuos.



$$d(A,B) = \frac{1}{n_A n_B} \sum_{i \in A, j \in B} d(i,j)$$

$n_A$ = # individuos del grupo A
$n_B$ = # individuos del grupo B

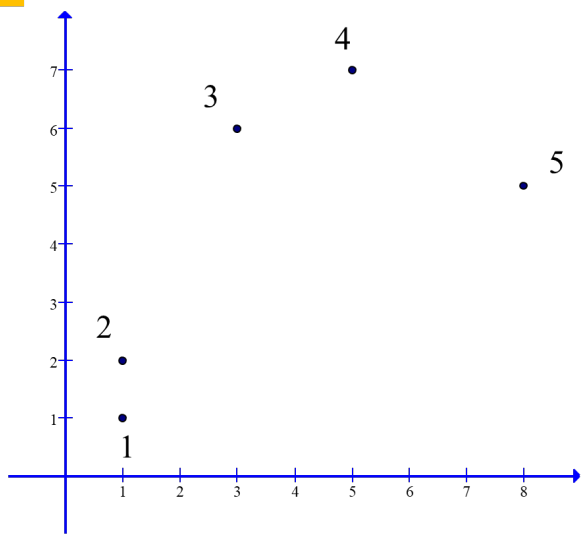
- **Centroide (centro de gravedad):** Con este método, una vez formados los grupos, son representados por su vector medio, y las distancias entre-grupos son ahora definidas en términos de distancias entre dos vectores medios.

$$d(A,\ B) = d(\bar{x}_A, \bar{x}_B)$$

$\bar{x}_A$ = centroide del grupo A

$\bar{x}_B$ = centroide del grupo B

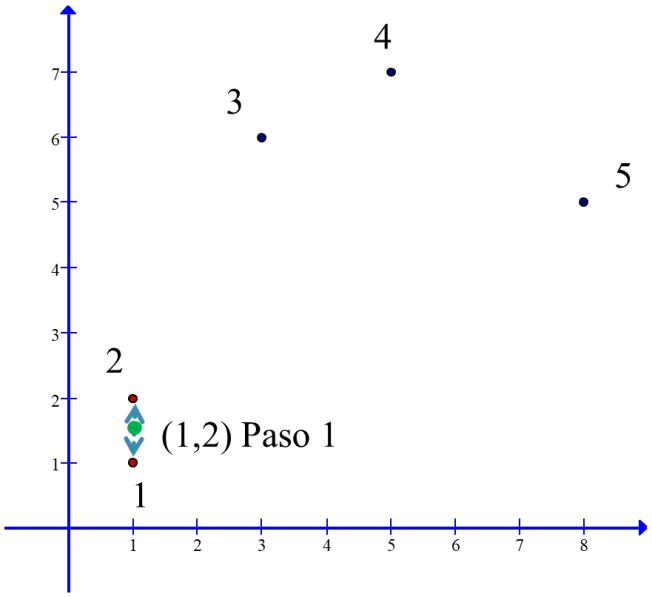
**Centroide** Calcular centroides iniciales



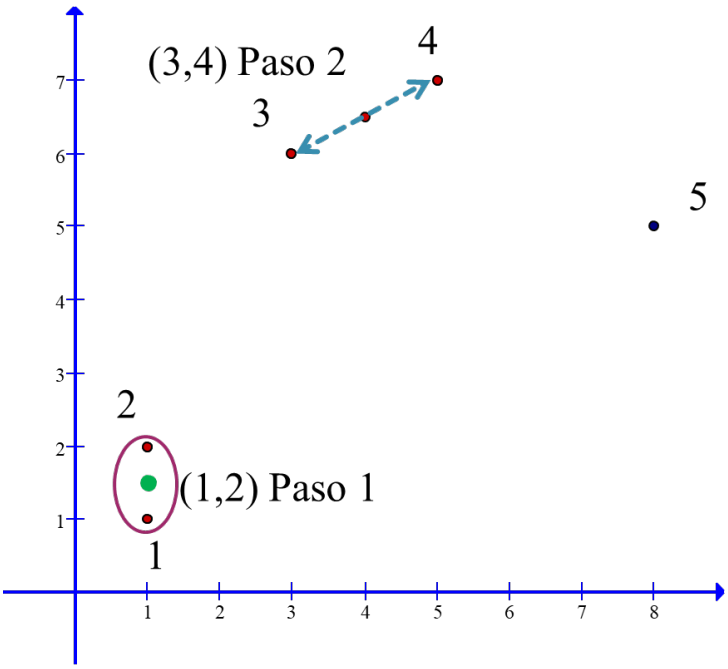

Calcular el centroide de la agrupación {1,2}, recalcular las distancias de los casos restantes hacia el nuevo centroide y seleccionar el de menor distancia. En el ejemplo {3,4}

BITMONEY

Seleccionar la pareja de menor distancia. En este ejemplo caso 1 y 2

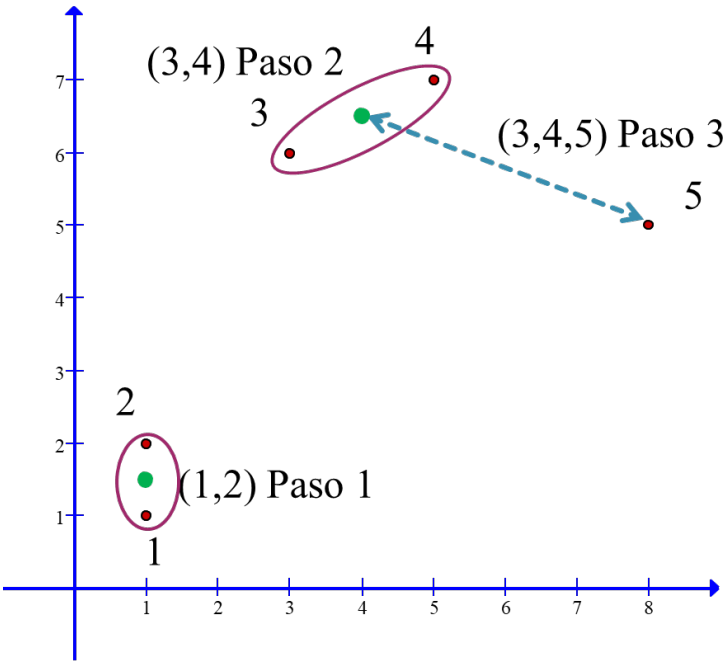


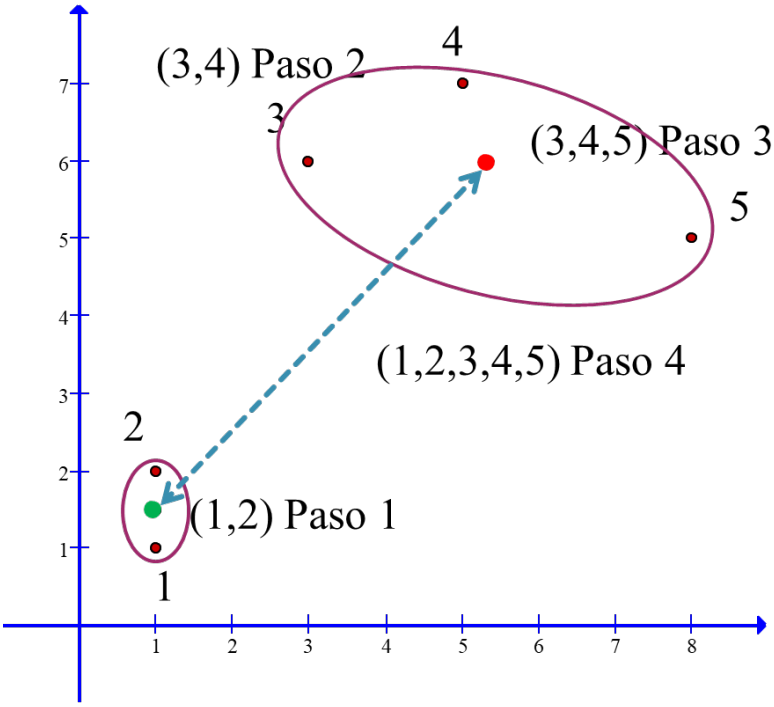

Calcular el centroide de la agrupación {1,2}, recalcular las distancias de los casos restantes hacia el nuevo centroide y seleccionar el de menor distancia. En el ejemplo {3,4}



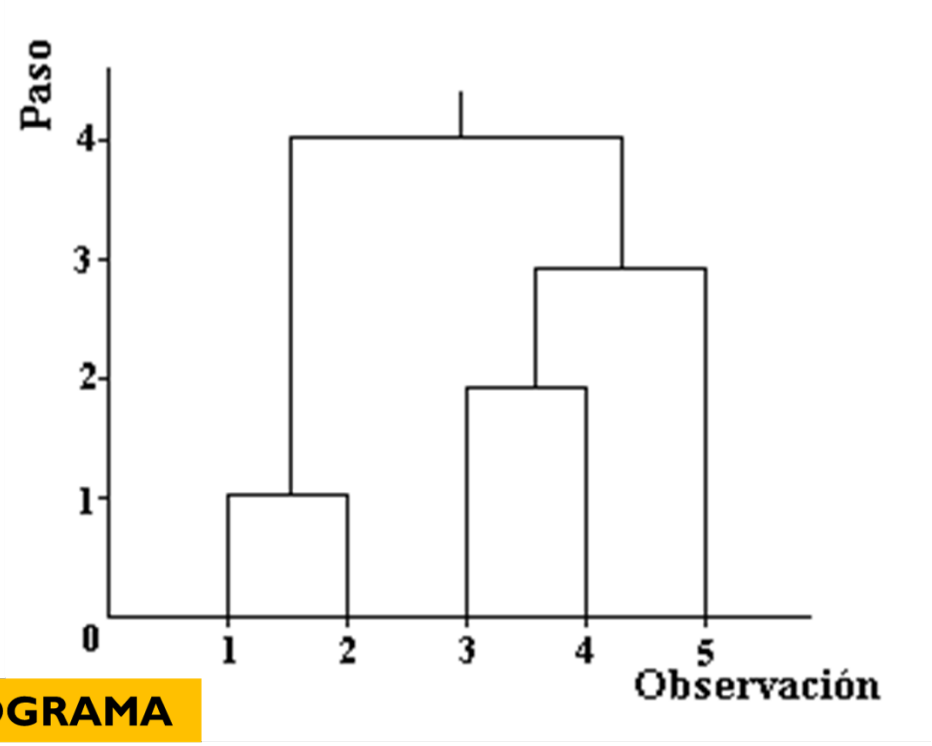

Repetir la mecánica anterior hasta agrupar todos los casos en un solo grupo.

BITMONEY


(3,4) Paso 2
4
3
(3,4,5) Paso 3
5
2
(1,2) Paso 1
1


(3,4) Paso 2
4
3
(3,4,5) Paso 3
5
(1,2,3,4,5) Paso 4
2
(1,2) Paso 1
1

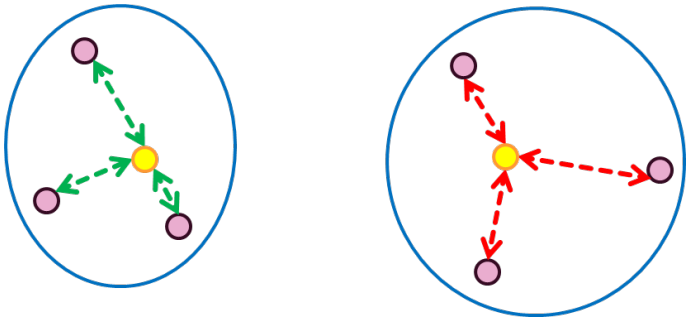
BITMONEY




**DENDOGRAMA**

- **Agrupamiento promedio:** Promedio de las distancias entre todos los pares de individuos.

**Ward:** procedimiento en el cual, en cada etapa, se unen los dos clústers para los cuales se tenga el menor incremento en el valor total de la suma de los cuadrados de las diferencias (*E*), dentro de cada clúster, de cada individuo al centroide del clúster. Es de los más empleados.



$$E = \sum_{k=1}^{h} E_k$$

$$E_k = \sum_{i=1}^{n_k} \sum_{j=1}^{n} \left(x_{ij}^k - m_j^k\right)^2$$

$x^k_{ij}$ = valor de la $j$-ésima variable sobre el $i$-ésimo clúster, suponiendo que posee $n_k$ individuos.
$m^k$ = centroide del clúster $k$ con componentes $m^k_j$
$E_k$ = suma del cuadrado de los errores del clúster (distancia euclidiana al cuadrado entre el individuo del clúster $k$ a su centroide)
$E$ = suma de los cuadrados de los errores para todas los $h$ clúster

BITMONEY

***D) Validación e interpretación de resultados***

Al plantear métodos jerárquicos se plantean los siguientes problemas:

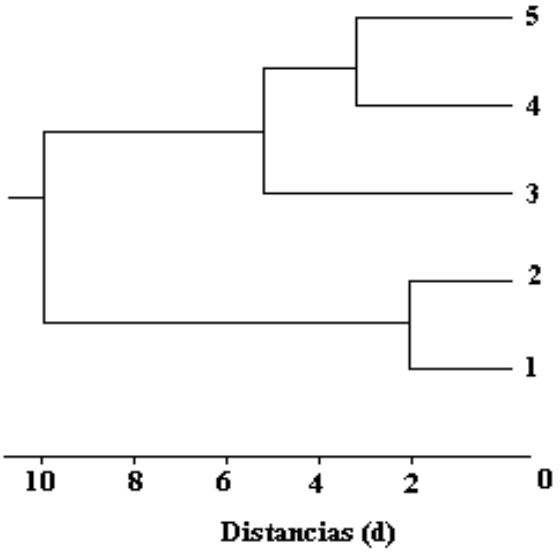
a) ¿Cuál es el número idóneo de clústers que mejor representa la estructura natural de los datos?

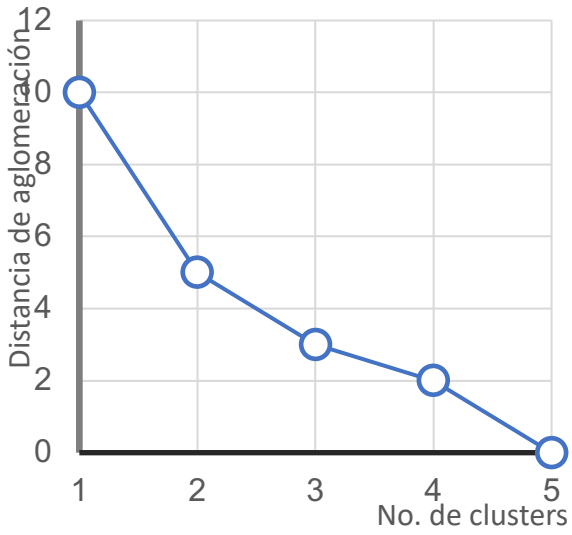
b) ¿En qué medida representa la estructura final las similitudes o diferencias entre los objetos?

Para contestar la primera pregunta, no existe una norma fija para establecer cuántos grupos pueden considerarse. Algunas estrategias empleadas son:

1) *Cortar el dendograma.* El dendrograma puede servir de ayuda visual para determinar dicho número dependiendo del coeficiente de proximidad usado. Este procedimiento generalmente presenta sesgos por la opinión y conocimiento que tiene el investigador sobre los datos.

2) Emplear un gráfico donde se represente la distancia de aglomeración y el número de grupos a esa distancia. En los primeros pasos la distancia es generalmente grande, mientras que los últimos es pequeño. El punto de corte será aquel en el que dejen de producirse saltos bruscos.



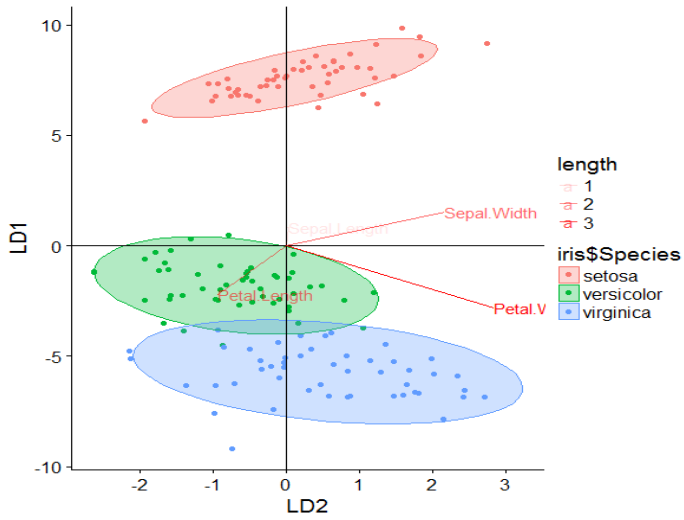

Si se utiliza el método del dendograma, bastara con trazar una línea que corte el árbol, si se desea cortar a una distancia de 6 (línea roja) se tienen dos clústers: {1, 2} y {3, 4, 5}. Pero si se decide cortar a una distancia de 4 (línea azul) se tienen 3 agrupaciones: {1, 2}, {3} y {4, 5}. En el caso del gráfico, se aprecia que al pasar de 1 a 2 se tiene un salto brusco pero ya no al pasar de 2 a 3 por lo que se considerarían dos grupos.

Para validar la pertinencia de la estructura seleccionada (número de agrupaciones elegidas) es necesario caracterizar cada grupo a partir de los individuos que fueron integrados.

Para ello se realiza un análisis descriptivo de cada uno de los grupos en las variables que fueron consideradas para realizar el análisis.

Otra opción es utilizar alguna técnica gráfica, como el grafico de dispersión o el grafico Biplot, representando a los individuos de cada agrupación

En caso de no poder diferenciar de forma adecuada a cada uno de los grupos, se debe replantear el número de grupos seleccionados e incluso si las variables medidas para realizar el análisis son las adecuadas.




*Ventajas y desventajas*:

***Ventajas:***

a) No requiere hacer inferencias previas sobre el número de clústers.

b) Permite representar la secuencia de agrupaciones en forma de árbol (dendograma).

***Desventajas:***

a) En ocasiones es alto el costo computacional.

b) Sensible respecto a las primeras agrupaciones a realizar.

Complicado de interpretar cuando el número de elementos a clasificar es grande.

### 4.2.4.2. Métodos no jerárquicos

Diseñados para clasificar individuos en *K* clústers. K se especifica *a priori* o bien se determina como una parte del proceso.

BITMONEY

El funcionamiento general de estos métodos es elegir una partición inicial de individuos y después intercambiar los miembros de estos clústers para obtener una mejor partición.

La mayoría de las aplicaciones adoptan *métodos heurísticos*:

- *k-medias*: Cada clúster está representado por el valor medio de los objetos del clúster.
- *k-medianas* o PAM (Partition around medoids): Cada clúster está representado por uno de los objetos situados cerca del centro del clúster.

***Elección de puntos semilla***

Es necesario establecer un conjunto de K semillas que puedan emplearse como núcleo de los clústers sobre los cuales el conjunto de individuos puede agruparse. Algunos procedimientos son:

Elegir los primeros K individuos del conjunto de datos (McQueen, 1967). Cuidar que los individuos hayan sido incluidos aleatoriamente.

Etiquetar los casos de 1 a m y elegir aquellos etiquetados como

$$\left[\frac{m}{k}\right], \left[\frac{2m}{k}\right], \ \ldots, \left[\frac{(k-1)m}{k}\right], m$$

Etiquetar los casos de 1 a m y elegir los casos correspondientes a K números aleatorios diferentes (McRae, 1971)

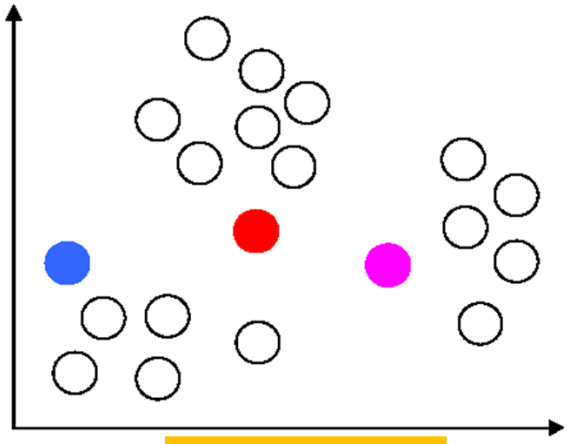
***K – Medias de McQueen***

Busca asignar cada individuo al clúster (de los *K* prefijados) con el centroide *más próximo*. El centroide es calculado a partir de los miembros del clúster tras cada asignación.

El algoritmo propuesto es el siguiente:

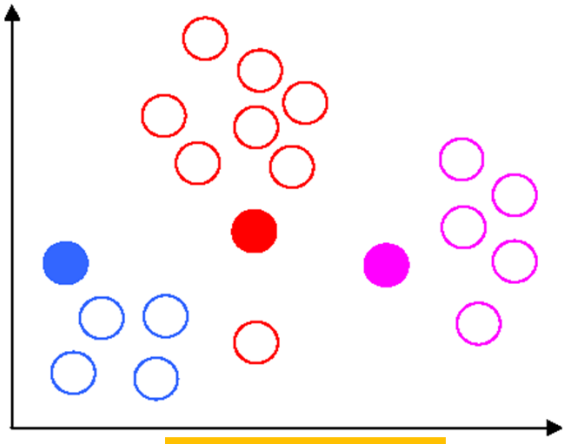
1. Tomar los *K* primeros casos como clústers unitarios.
2. Asignar cada uno de los *m-K* individuos restantes al clúster con el centroide más próximo. Después de cada asignación, recalcular el centroide del clúster obtenido.
3. Tras la asignación de todos los individuos en el paso anterior, tomar los centroides de los clústers existentes como puntos semilla fijos y hacer una pasada más sobre los datos asignados cada dato al punto semilla más cercano.

BITMONEY



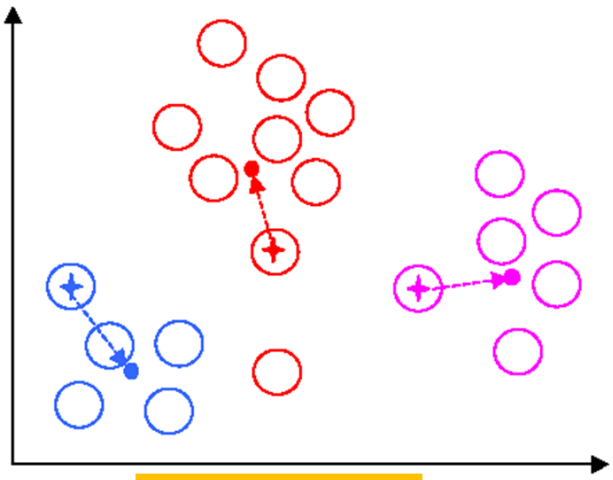
## PASO 1

Selección de las observaciones que serán las semillas para iniciar el algoritmo de agrupación



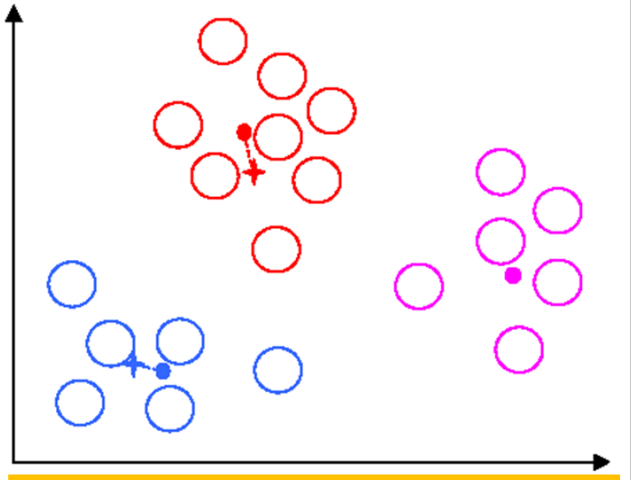
## PASO 2

Agrupación inicial de los casos restantes tomando como centroide las observaciones consideradas como semillas



## PASO 3

Calcular el centriode de cada cluster.

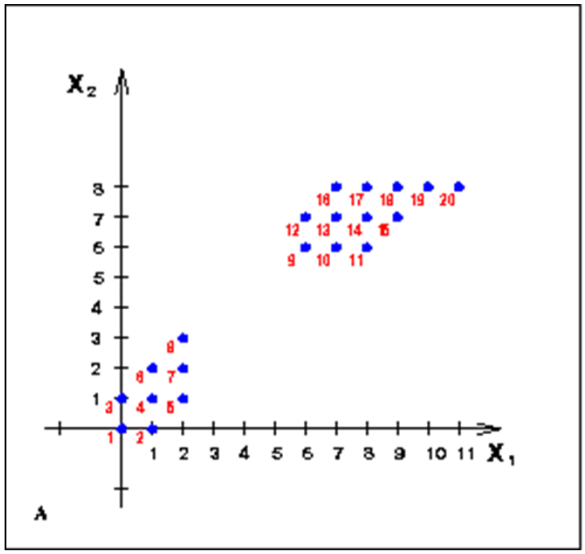


## Mejor resultado encontrado

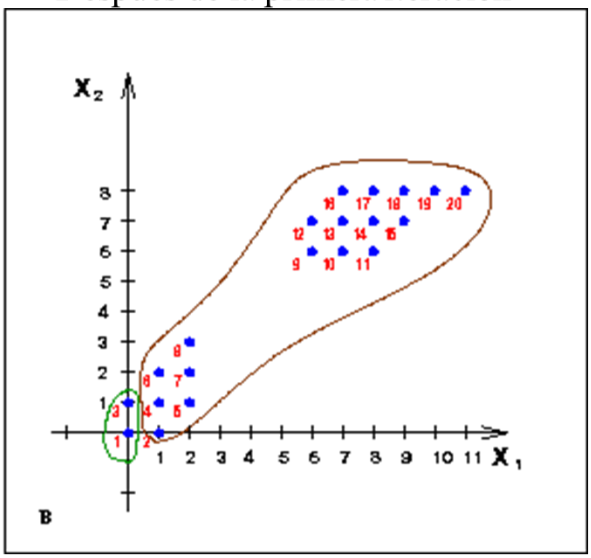
Se verifica si alguno de las observaciones esta más próxima a los nuevos centroides. Si alguna está más próxima se cambia de cluster y se recalcula el centroide. Repetir hasta que no haya cambios de cluster.

BITMONEY

Situación inicial




Después de la primera iteración



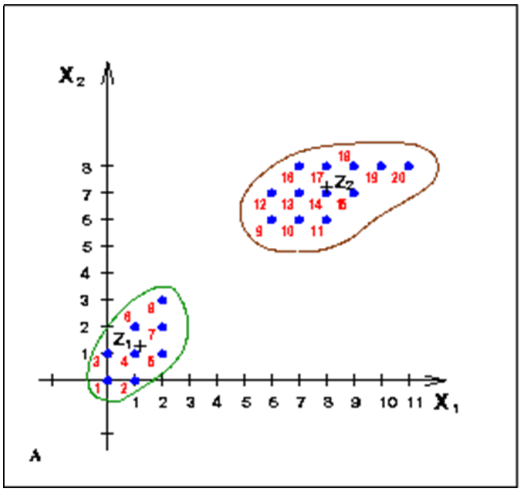

**Paso 1.** $S_1(0) = \{X_1\}$ $Z_1(0) = (0, 0)$
$S_2(0) = \{X_2\}$ $Z_2(0) = (1, 0)$

**Paso 2.** $S_1(1) = \{X_1, X_3\}$ $Z_1(1) = (0, 0.5)$
$S_2(1) = \{X_2,..., X_{20}\}$ $Z_2(1) = (5.8, 5.3)$

**Paso 3.** $Z_1(1) \neq Z_1(0)$ y $Z_2(1) \neq Z_2(0)$
Volver al paso 2

BITMONEY

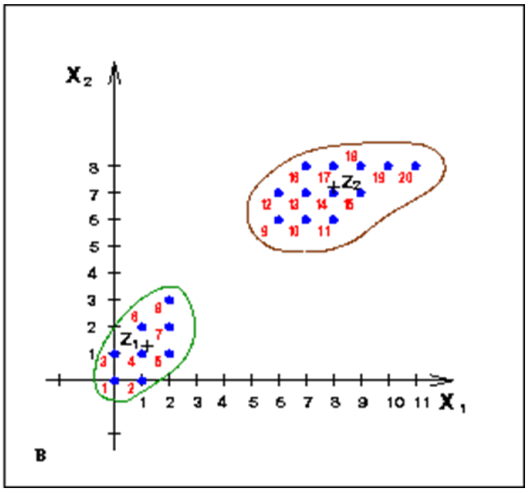
Segunda iteración




**Paso 2.** $S_1(2) = \{X_1,\ldots,X_8\}$ $\quad Z_1(2) = (1.1, 1.3)$
$S_2(2) = \{X_9,\ldots,X_{20}\}$ $\quad Z_2(2) = (8.0, 7.2)$

**Paso 3.** $Z_1(2) \neq Z_1(1)$ y $Z_2(2) \neq Z_2(1)$
Volver al paso 2

Tercera iteración




**Paso 2.** $S_1(3) = \{X_1,\ldots,X_8\}$ $\quad Z_1(3) = (1.1, 1.3)$
$S_2(3) = \{X_9,\ldots,X_{20}\}$ $\quad Z_2(3) = (8.0, 7.2)$

**Paso 3.** $Z_1(3) = Z_1(2)$ y $Z_2(3) = Z_2(2)$
FIN

**Ventajas:**

- Algoritmo es muy sencillo.
- Funciona bien para encontrar clústers con forma esférica.

**Inconvenientes:**

- El resultado final depende del valor de *K y* de la inicialización de los centros.
- Sólo para datos a los que se les puede aplicar la media.
- No adecuado para formas no convexas o clústers de diferentes tamaños.
- Sensible a ruidos y *outliers*.

***Otros métodos no jerarquicos***

- **K-modes**: Para datos cualitativos, reemplazando las medias por **modas**. Usando el total de *discordancias* entre dos objetos: *mientras más pequeño este número, más similar ambos objetos.*

$$d(X, Y) = \sum_{j=1}^{m} \delta(x_j, y_j)$$

BITMONEY

$$\delta(x_j,\ y_j) = \begin{cases} 0 & x_j = y_j \\ 1 & x_j \neq y_j \end{cases}$$

- **K-prototypes:** Integración de K-medias y K-modes para datos cualitativos y cuantitativos.
- **K-medianas**: Desarrollado por Kaufman y Rousseeuw en 1987. Soluciona la sensibilidad del K-medias frente a los *outliers*.

Se toma como punto de referencia el objeto situado en el centro del clúster, en vez de tomar el valor medio. Más robusto que K-medias. Su procesamiento es más costoso.